EMC WHITE PAPER

EMC CLARiiON Storage System Fundamentals for Performance and Availability

Applied Best Practices

EMC
Corporate Headquarters
Hopkinton, MA 01748-9103
1-508-435-1000
www.EMC.com

Published January, 2010

EMC CLARiiON Storage System Fundamentals for Performance and Availability

P/N h1094.4

Contents

Figures

Tables

How to Use This Guide

This paper:

- Presents an overview of the CLARiiON storage system. You can read this as a tutorial to gain a basic understanding of its performance and availability features.
- Explains concepts that you need to understand when evaluating CLARiiON's performance and availability options.
- Answers many of the questions that new CLARiiON users frequently ask.
- Supplies the background information you need to understand the *EMC CLARiiON Best Practices for Performance and Availability: Release XX Firmware Update — Applied Best Practices* and *EMC CLARiiON Best Practices for Fibre Channel Storage: FLARE Release 26 Firmware Update* white papers. (These are also referred to as the Best Practices white papers.) If you are a new CLARiiON user, we urge you to read this paper before you read the Best Practices paper for your CLARiiON's firmware revision.
- Gives examples using CX4 and AX4 CLARiiONs. Some of the examples may not apply to legacy CLARiiONs (such as the CX3 or CX series).
- Mirrors the organization of the Best Practices white papers for easy cross-reference.
- Provides a glossary at the end of this paper defining many EMC-specific terms.

If you use this as a reference, you need to decide whether you need help with your CLARiiON's performance, availability, or both. Then find the appropriate sections and please read the entire section.

Audience

This document is for Information Technology (IT) personnel new to CLARiiON storage systems. It is an introductory paper for those who need help implementing CLARiiON CX4 and AX4 series storage system best practices. An understanding of the basics of data center hosts (servers), networks, and IT concepts is assumed. The beginning of the document is general in nature, and detail is gradually increased. These general sections may be skipped by experienced readers.

A note on terminology: In this document, the term *drive* refers to both mechanical hard drives and *Enterprise Flash Drives* (*EFDs*). EFDs are non-volatile memory-based drives that are sometimes referred to as solid state disks (SSDs) in the IT industry. Where the two drive types differ, they are discussed separately.

References

- *CX4 Model 960 Systems Hardware and Operational Overview*
- *EMC CLARiiON Best Practices for Performance and Availability: Release XX Firmware Update — Applied Best Practices*

- *EMC CLARiiON Best Practices for Fibre Channel Storage: FLARE Release 26 Firmware Update*
- *EMC Networked Storage Topology Guide*
- *Introduction to the EMC CLARiiON CX4 UltraFlex Series* white paper

Chapter 1 Storage Environment

This chapter presents these topics:

EMC® CLARiiON® storage systems reside in storage environments that include these components:

- CLARiiON storage systems
- Hosts
- Storage area networks (SANs)

In a storage environment, one or more storage systems (like the CLARiiON) are connected through SANs to one or more hosts. These hosts interface with *clients*, which can be human end users or other computer systems. Storage environments are usually complex. Each component has its own architecture that affects the performance and availability of itself *and* the entire storage environment.

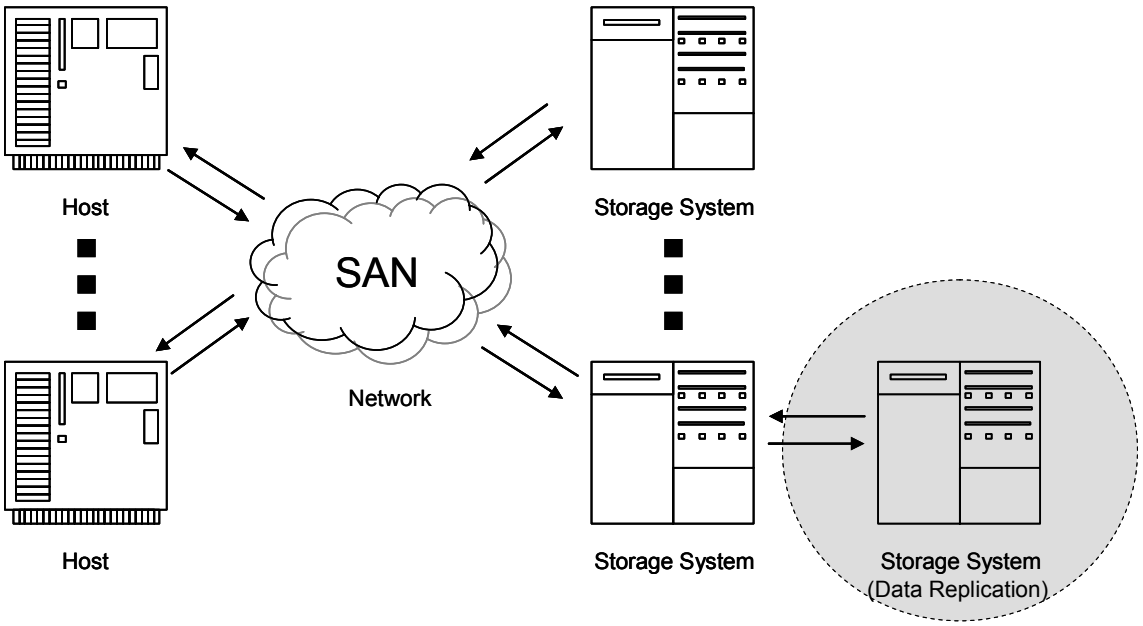

Figure 1 Components in a storage environment

The key measurements of a storage environment are its:

- Performance – how long it takes a storage environment to retrieve data for end users.
- Availability – a measure of a storage environment's ability to retrieve data for users, especially when a component in the environment fails.
- Capacity – the amount of data that can be stored in a storage system.

Storage systems

Generally, storage systems are categorized as:

- Entry level
- Midrange
- Enterprise

Entry-level systems typically serve the needs of small businesses, or as part of a larger distributed storage environment. Entry-level systems host a maximum of about 100 drives. The performance and availability for entry-level performance varies greatly between systems. CLARiiON AX4 storage systems are an example of an entry-level storage system.

Enterprise systems serve the central business needs of large corporations or public-sector organizations. Enterprise systems are designed and built for the highest possible performance and availability. Smaller enterprise storage systems may host about 500 drives, but 1,000 is more typical. Larger enterprise storage systems host more than 2,000 drives. The EMC Symmetrix® line of storage systems is an example of a large-enterprise storage system.

Between the entry-level and enterprise systems is the midrange. The CLARiiON CX4 is an example of a midrange storage system. The midrange spans a large amount of storage system performance and capacity. The smallest midrange systems are similar to the largest entry-level systems in performance and capacity, although midrange systems have more features and better availability. The smaller midrange systems are also more *scalable*, which means the performance and capacity of the system can easily be increased to meet growing needs. For example, in a CLARiiON CX4 it is easy to increase the number of drives in the system from 75 to 125.

Midrange systems are also more *extensible* than entry-level systems, which means additional hardware and software features can easily be added to address new requirements. For example, the CX4 can communicate over more than one type of SAN via Ethernet *and* Fibre Channel.

The largest midrange systems are similar in performance and capacity to the smaller enterprise systems. Large midrange systems typically have the advantage of being less expensive to purchase than enterprise systems.

The CLARiiON CX4 and AX4 are *block-level* storage systems; CLARiiON writes and reads blocks of data using *logical block addresses* (LBAs), which are translated into disk sector addresses on the drives. (See the CLARiiON Storage Objects chapter for details.) SAN storage environments use block-level storage to provide a higher level of performance compared to file-level storage.

The alternative architecture is a *file-level storage* system. Storage systems using file-level storage add a level of abstraction above the block-level access; the host's data is sent as file system extents, which must be mapped to logical disk blocks before they are stored on the hard drives. The term *network-attached storage* (NAS) refers to file-level storage. EMC Celerra® storage systems are an example of NAS systems.

General architecture

Midrange and entry-level CLARiiON storage systems like the AX4 and CX4 have a modular architecture that includes CPUs, memory, I/O subsystems, and storage. The main parts of the CLARiiON are the:

- Front-end – Consists of the front-end I/O ports that attach the storage system to the hosts directly or though storage area networks. Front-end ports can also be directly connected to other storage systems.

- Storage processors – Consists of CPUs that manage the storage system's functions and memory, including the important storage system cache.
- Back end – I/O buses connecting the storage processors to the drives.
- Drives – Mechanical hard drives and EFDs.

Traditionally, *drives* have described mechanical hard drives, although newer semiconductor-based storage technologies are also available. Hard drives have several types of physical attachments (SATA, SAS, Fibre Channel), a large range of capacities, and several rotational speeds. These characteristics have several effects on performance and availability.

Hosts

Hosts are server-class computers running applications. Hosts can connect to CLARiiON storage systems directly or over networks. Hosts are like the CLARiiON — they have CPU, memory, and I/O resources to manage. Hosts range in performance and availability from a high-performance *blade server* with many individual servers in a single enclosure designed for high availability, to a modest personal computer (PC) type desktop workstation. The way that hosts manage their CPU and I/O resources can have an effect on the storage system's performance and availability.

Hosts may operate singly, or be integrated into a *network* of hosts performing a related function. Networks of hosts are sometimes called *clusters*. A cluster is a group of host servers executing a common application in distributed fashion and presenting their clients with a single software system. Microsoft Cluster System (MSCS) is one example of this architecture.

Hosts can also reside in *virtual machines* (VMs). VMs are software applications that emulate a hardware environment. VMs are supported by a *hypervisor* layer that manages each VM's access to the underlying hardware platform. VMware ESX™ Server and Microsoft's Hyper-V™ are two examples of VM software. VMs can boost the usage rate of host hardware, but they can also pose performance challenges when multiple applications compete for the same host and storage system resources.

Hosts execute *system* and *application* software, both of which affect performance and availability of the CLARiiON storage system.

System software includes operating systems (*O/Ss*), such as Microsoft Windows Server 2008, HP-UX, and Linux.

Applications include programs such as Microsoft Exchange Server, Oracle 11i, and Microsoft SQL Server. Applications generate most of CLARiiON's workload, although system software generates part of the workload during startup and maintenance. Applications can be deployed onto storage systems in several ways:

- Dedicated: A single application's data has sole use of a storage system This creates a single storage system workload
- Shared: More than one application's data is stored on the storage system. This creates multiple workloads.
- Optimized: Application data is distributed across more than one storage system. This creates single or multiple workloads, depending on the scale of the application.

Storage area networks

The complexity of the SAN and its protocols has an affect on the CLARiiON's performance and availability. Storage networks are either Fibre Channel or Ethernet networks transporting SCSI protocol traffic (see the Protocols chapter). The CX4 series CLARiiONs can be connected to Fibre Channel or Ethernet networks. They can also be connected to both types of networks at once.

CLARiiONs can also connect directly to hosts (with no network devices between the CLARiiON and the host). This is known as *direct attach* (DAS). This connection is used primarily for local or remote data replication for use in backup or disaster recovery. Many smaller-capacity, entry-level CLARiiONs, such as the AX4-5, are configured in this fashion.

Fibre Channel has been the most common networking technology for SANs. It has high throughput and low latency. It is also highly reliable. However, general purpose Ethernet networks are widely used at storage system deployment sites. Ethernet's large installed base, low cost, and relative simplicity make it an attractive choice for SAN usage.

FC SANs connected to the CLARiiON use the *Fibre Channel* (FC) protocol, while Ethernet SANs use the *iSCSI* protocol.

The performance and availability of SANs are highly dependent on the *network topology*. Network topology is the physical and logical arrangement of the network's elements. Network elements are the components that make up the network such as high-capacity/ high-availability routers; fiber-optic or copper cabling; and Fibre Channel or IP switches. Readers interested in networking, addressing, network topologies, and how they relate to the CLARiiON and SANs, are encouraged to read the *EMC Networked Storage Topology Guide* available on Powerlink.

Chapter 2 Protocols

This chapter presents these topics:

A protocol is a set of rules that governs the communications between computers on a network. Several protocols describe how to physically transmit data and how to control, manage, and interpret the data transmitted between hosts and CLARiiON.

Protocol stack

Protocols set standards. Protocols also have a hierarchical structure. They layer on top of each other. This is sometimes called a *stack*. A brief discussion of the protocols only needs to discuss the following three stack levels:

- Physical and Data Link Layer Protocols – Describes the specification for a network's copper or optical Fibre cables and the operation of network devices.
- Network Layer Protocols – Data passed over the physical layer is sent using one or more *network protocols* that specify how data is sent and received.
- Application Layer – Describes how the data is interpreted. For example, it may be application data, or it may be a command controlling the communication. On the CLARiiON, these are read and write I/O requests.

For example, read and write I/O requests arrive at the CLARiiON over Fibre Channel or Ethernet networks. Read and write I/Os are SCSI protocol commands. Ethernet networks connect to CLARiiONs iSCSI front-end ports. The iSCSI protocol is one of several networking protocols used to deliver the SCSI protocol commands. Figure 2 Detailed iSCSI stack shows the details of the iSCSI stack.

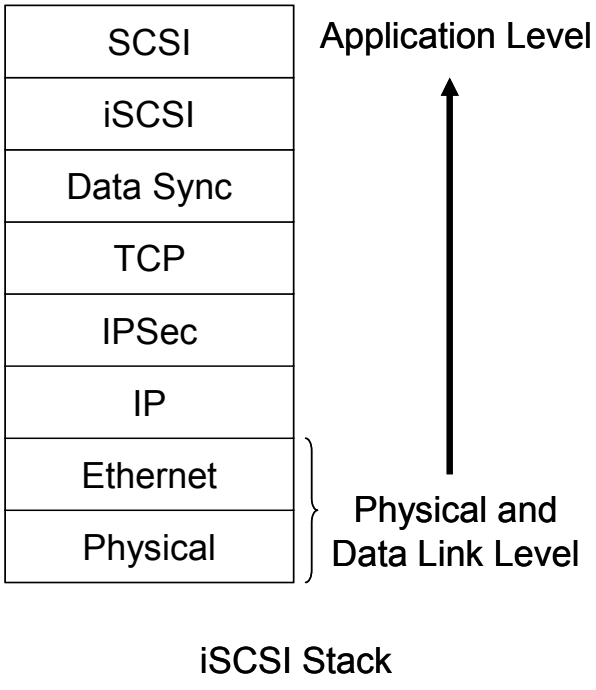

Figure 2 Detailed iSCSI stack

Physical and data link protocols

There are four main physical and data link protocols used to send and receive data to and the CLARiiON, and to send data within the CLARiiON series. They are:

- Fibre Channel
- Ethernet
- SAS
- SATA

These protocols are the physical layer protocols of the networks that connect the hosts to the CLARiiON and the CLARiiON's storage processors to its drives.

As previously mentioned, the CLARiiON's front-end ports are either Fibre Channel or Ethernet. The CX4 series CLARiiON's storage processors also connect to the storage system's drives by a Fibre Channel back-end bus. The AX4 series CLARiiON's storage processor(s) connect to their hard drives by a SAS interconnect.

Different protocols are used within the storage system. Drives connected to storage system processor buses use either Fibre Channel or SAS (AX4 series) attachments. The SATA protocol is also used. SATA drives can be used on either Fibre Channel or SAS bussed storage systems. Note that on the AX4 series there is a SAS point-to-point connection between the storage processor and the SATA or SAS hard drive.

Fibre Channel

The Fibre Channel protocol is standardized to the American National Standards Institute (ANSI) standard INCITS 387.

Fibre Channel is a layered protocol as described above. It is an *entire* stack. Each layer describes a different aspect of the protocol from the physical transmission of bits on the cable to the interpretation of high-level commands and data. Note that it is common to refer to one layer or all the Fibre Channel stack's layers simply as Fibre Channel.

There are two important Fibre Channel stack layers to understand:

- FC-PH
- FC-4: Protocol Mapping layer

The Fibre Channel protocol specifies the bus interface that device controllers use to communicate with one another

FC-PH

The FC-PH is the physical layer that actually consists of four layers, (FC-0 through FC-3). These layers constitute the Fibre Channel cabling, including the data link. Data links connect one location to another; digitally-encoded information is transferred over data links.

On the CLARiiON, Fibre Channel can provide front-end communication with hosts. In addition, the back end of CX4 series has a Fibre Channel architecture that provides the connection between

the storage processors and the storage-device controllers. The FC protocol has operational speeds of 1 Gb/s, 2 Gb/s, 4 Gb/s, 8 Gb/s, and 10 Gb/s.

FC-4

The FC-4 is the network layer in which SCSI protocol messages are encapsulated in Fibre Channel protocol packets for delivery. This layer is discussed below.

Ethernet

The Ethernet protocol is standardized as the Institute of Electrical and Electronics Engineers (IEEE) standard IEEE 802.3. Ethernet has a single physical layer.

Ethernet networks connect to the CLARiiON's front-end Ethernet ports. These Ethernet ports are referred to as *iSCSI* ports after the intermediate network protocol used in workload communications. The Ethernet protocol has operational speeds of: 10 Mb/s, 100 Mb/s, 1 Gb/s, and 10 Gb/s. The 1 Gb/s speed is referred to as *GigE*. The 10 Gb/s speed is *10 GigE*.

SAS

Serial Attached SCSI (SAS) is an ANSI standard (ANSI INCITS 376) for device attachment.

SAS is a block-level protocol. It provides a point-to-point serial interface in which device controllers may be directly linked to one another. In the CLARiiON's case it is a connection to drive controllers. SAS integrates two protocols: SCSI and Serial Advanced Technology Attachment (SATA). It combines the SCSI's device utility and reliability with SATA's serial architecture.

SAS is the "back-end" architecture of the AX series CLARiiONs. The SAS protocol has operational speeds of 3 Gb/s and 6 Gb/s (SAS-2 protocol).

SATA

Serial ATA (SATA) II is an ANSI standard (ANSI INCITS.397-2005) for drive attachment.

SATA is a block-level protocol. It provides a point-to-point serial interface in which device controllers may be directly linked to drives.

SATA is the attachment interface used on one type of drive on all CLARiiONs. The SATA protocol has an operational speed of 3 Gb/s and 6 Gb/s (SATA Gen-3).

Network protocols

There are several network protocols used in data communications in CLARiiON. They are:

- Fibre Channel
- TCP/IP
- iSCSI
- VLAN
- CHAP

These protocols describe how the data is transmitted and received by the CLARiiON. They also control and manage the connection between the sender and receiver.

Note that network protocols can be sent on different physical protocols. For example, Fibre Channel protocol formatted data can be sent on a TCP/IP (see below) network. This is called Fibre Channel over IP (FCIP or FC/IP).

Fibre Channel

This is the FC-4 layer Fibre Channel protocol. The Fibre Channel protocol carries embedded SCSI commands for reads and writes over Fibre Channel networks. It also includes control units. Control units are discrete messages that are separate from data bearing messages. These messages manage the connection over which the data is passed.

TCP/IP

TCP/IP is also referred to as the *Internet Protocol*. It is actually two standardized protocols that are "stacked" together. TCP (Transmission Control Protocol) is the higher-level protocol; it creates packets for data and messages that need to be transmitted over the internet. IP is the lower-level protocol; it is responsible for adding addresses to the packets to make sure that the packets reach the proper destination. IP can also be used with other high-level protocols.

The Internet Engineering Task Force (IETF) standardized TCP/IP, and documents them in a set of documents called RFCs, or Request for Comment:

- IPv4 is described in RFC 791.
- IPv6 is defined in RFC 2460 - FLARE 28 and later use this version of IP.
- TCP is described in RFC 793. – FLARE uses this version of TCP.

Like Fibre Channel, TCP/IP is a carrier protocol. It is used to encapsulate iSCSI protocol packets.

TCP/IP can also be used to deliver Fibre Channel packets on SANs. Fibre Channel over Ethernet (FCoE) is an encapsulation of Fibre Channel packets for delivery over Ethernet-based SANs. This allows Fibre Channel to use high-bandwidth and long-distance Ethernet networks while preserving the Fibre Channel protocol. It is also used on a CLARiiON's service and management Ethernet ports.

iSCSI

iSCSI is the Internet Small Computer System Interface protocol. iSCSI is layered on top of the TCP/IP protocol. iSCSI carries SCSI commands over Ethernet networks.

iSCSI is standardized in RFC 3720.

Note that the TCP/IP and iSCSI protocols are used on an Ethernet network at the same time. iSCSI would be used for SAN workload communications. TCP/IP would be used for other network communications. However it is recommended that SAN traffic be either physically or virtually separated from all network traffic.

iSCSI Virtual LANs (VLANs)

A storage network VLAN uses the iSCSI protocol to partition Ethernet networks to provide more resources.

VLAN operation is part of the TCP/IP suite of protocols, and is standardized under IEEE 802.1q.

VLANs create logical collision and broadcast domains. With the VLAN broadcast domains, all messages are not sent to all destinations. Traffic still physically co-exists on the same cabling, hubs, switches, and routers. However, broadcast and multicast traffic is limited by the creation of the VLAN domains. The VLANs bridging software defines which nodes are to be included in the broadcast.

As noted above, a dedicated network is recommended for iSCSI (storage) use. Physically isolating your storage environments network traffic produces higher network performance and greater reliability. However, sometimes a physically separate storage network is not practical, and the available Ethernet network may need serve both the SAN and the networking needs of the general users. If not using a dedicated Ethernet LAN, iSCSI traffic should be either separated onto its own physical LAN segments by firewalls that are not used for general LAN traffic, or use VLANs.

Restricting storage traffic to its own VLAN is not a proper security method. The confidentiality of the data cannot be protected by a VLAN alone. Proper security on a VLAN is possible only by implementing CHAP or iSCSI authentication.

CHAP

Security is provided on iSCSI networks though the use of Challenge Handshake Authentication Protocol (CHAP).

CHAP is a protocol that is used to authenticate the peer of a connection. RFC 1994 defines CHAP. It is based on symmetrical encryption, which is a shared key between the peers. On a SAN the peers are a host and the storage system. With CHAP both the host and storage system know the plaintext of the secret key. However the key is not sent over the network.

Application protocol (SCSI)

Small Computer System Interface (SCSI) is the application-layer protocol used to send commands to drives, including read and write I/Os. I/Os are sometimes referred to as *SCSI commands*. SCSI is the standard for connecting and transferring data between hosts and the CLARiiON's LUNs that are made up of its drives.

SCSI commands may contain data or control information. They are sent by host applications and arrive embedded in either iSCSI protocol or Fibre Channel protocol message units. Figure 3 Block protocol summary iSCSI layer shows the relationship between the major storage system protocols and the SCSI protocol.

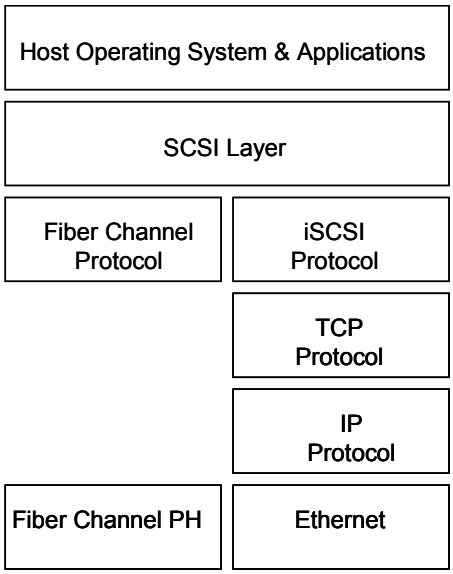

Figure 3 Block protocol summary iSCSI layer

Chapter 3 Performance and Availability Metrics

This chapter presents these topics:

Performance is the amount of work accomplished by a system compared to the time and resources used. The time and resources need to be measurable to judge performance. The definition of high performance is dependent on the production goal. The CLARiiON CX4 and AX4 series storage systems achieve high performance in a number of operational profiles, from high bandwidth, long latency to high rate, low latency.

Availability refers to the storage system‘s ability to provide users access to their applications and data even when a hardware fault is present (this is sometimes called a *degraded* state or mode). Midrange systems like the CLARiiON CX4 series are classified as *highly available* because they continue to provide access to data with any single failure.

Identifying the measurements appropriate to the resource is an important part of architecting or evaluating a storage system’s performance and availability. The most common terms used to measure a system’s performance and availability are:

- Bandwidth
- Throughput
- Response times
- Availability
- Recovery time objective (RTO)

The availability metric should not be confused with *general availability*, which is the combination of reliability and redundancy needed to avoid the not functioning system state.

Bandwidth

Bandwidth is the measurement of the amount of data that can be transferred along a channel per second. Storage-system bandwidths are measured in megabytes per second (MB/s) or gigabytes per second (GB/s). Communications network bandwidth is typically measured in bits per second (b/s).

Storage-system bandwidth usually describes sequential or large-block I/O performance, but it can describe any workload.

The storage system’s *deliverable* bandwidth is directly affected by the bandwidth of the attach layer, which is also referred to as the *attach-bandwidth.* Attach bandwidth is usually measured in megabits per second (Mb/s) or gigabits per second (Gb/s).

It is important to remember that storage-system bandwidth is measured in *bytes* per second, and attach-bandwidth is measured in *bits* per second. There are frequent conversions between *attach* bandwidth and *storage system drive’s* bandwidths. In addition, there is a difference between the *maximum specified* bandwidth (called *wire rate* or *wire speed*) and the *actual* bandwidth realized.

Due to the effects of packet overhead, and the network protocols, data always travels at less than the maximum speed of the hardware. For example, keeping in mind that a byte is eight bits, and the overhead of protocols, the maximum gigabit Ethernet (GigE) speed is equivalent to about 119 MB/s.

Throughput

Throughput is the number of I/O operations that are processed per second (*IOPS*) over a period of time. *This measurement is independent of request size.* Individual components, such as hard drives, up to entire storage systems can be rated for their throughput performance. Throughput is the measurement used to describe random and small-block I/O performance.

Throughput can be further divided into:

- Read IOPS: Read I/Os per second
- Write IOPS: Write I/Os per second
- Total IOPS: Average I/Os per second

This distinction is made because of the different amounts of storage system resources consumed in servicing a read operation versus a write. For a storage system I/O, it is very common to see the total IOPS stated with the ratio of read-to-write IOPS. An example is "10,000 IOPS, 70 percent reads."

Bandwidth and throughput on a storage device

The relationship between bandwidth and throughput on a storage device is important to understand. This relationship is related to I/O size. That is, as I/O operations (measured in bytes per second) go up in block size, the quantity of data per unit time transferred likewise goes up. However, with constant bandwidth, as the I/O size goes up, the I/O rate (throughput) goes down. The relationship can also be looked at the other way: As the I/O operations goes up in block size, the I/O rate goes up.

Figure 4 shows the relationship of throughput to bandwidth for a typical mechanical drive. On the left axis is throughput. On the right axis is bandwidth.

The red line is bandwidth with its MB/s axis of measurement on the right. The green line is throughput with its IOPS axis of measurement on the left.

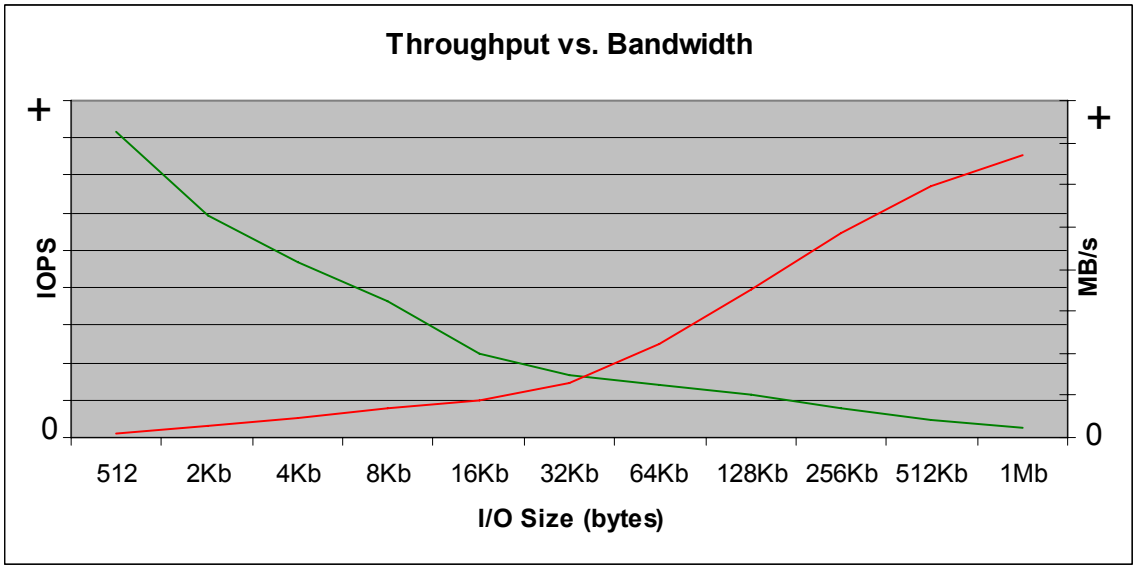

Figure 4 Mechanical hard drive throughput vs. bandwidth

From the graph, you can see that if the I/O size is a small 512 bytes, the bandwidth (red line) is low, but the throughput (green line) is high. Oppositely, as the I/O size increases, the bandwidth (red) goes up, and the throughput (green) goes down. This is shown most clearly by the far right side of the graph.

Note that any discussion of a device or interface being able to support "so many" GB/s of bandwidth, and "so many" IOPS of throughput is meaningless without knowing the size of the I/O. Knowing your I/O size is the first step in a bandwidth/throughput analysis.

Response time

Response time is a measurement of the elapsed time between the end of an inquiry or of a request for service and beginning of the request complete. There are two types of response time discussed: user and I/O.

User response time is their "user experience" with the system, of which the storage system is a part. I/O response time is a storage-system-only metric measuring how long it takes to satisfy a host read or write request.

User response time

Users expect a uniform, high level of performance across a variety of operating conditions. The primary metric for user performance is *user response time*. Sometimes this is called *client-side response time* or *host* response time. User response time is the amount of time (typically seconds) it takes a system to respond to a request. This may be a query request, report generation, or a simple terminal screen refresh. Often there is a *Quality of Service* (QoS) agreement to quantify the performance promised to the user by the system's architects and administrators.

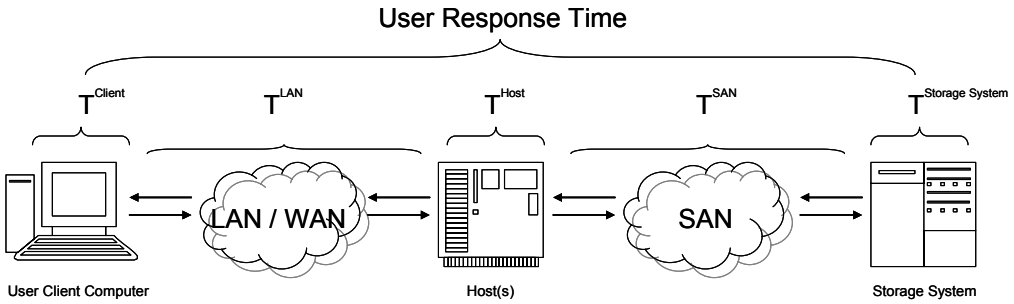

Figure 5 Conceptual view of user response time

Each component in the system shown above has its own response time ("T").

User response time is a particularly high-level measurement. The user response time is the sum of the response times both for transmit and receive of all the sub-systems within the system. These subsystems may include:

- The user's workstation (client)
- The network (local area network (LAN) or wide area network (WAN)) connecting the client computer to its server
- The host server or hosts

- The storage area network (SAN) connecting the servers to storage system
- The CLARiiON storage system itself

A single sub-system with a high response time can adversely affect overall user response time.

For example, it is possible for the CLARiiON to have a low I/O response time, and the intervening LAN to have a high response time due to network congestion. This results in an overall high user response time.

A high user response time is one of the first indications a bottleneck exists in a storage environment. However, it does not indicate where the bottleneck is located or its root cause.

I/O response time

I/O response time is a measurement of how long a single I/O takes to complete. It can apply to either a read or write from an individual drive such as a hard drive, or a read or write to a LUN made up of several drives. Always be clear on which object is being measured when stating response time.

I/O response time includes time spent waiting in queue as well as time spent servicing a request. Queues are temporary storage areas where I/Os wait for their turn at execution. Short service times and short queues result in short I/O response times. Long queues increase response time.

This measurement is typically in milliseconds (ms) although some devices can respond in microseconds (µs).

Availability

Availability measures the percentage of time the system is able to return data when requested by the client. Degraded performance is not included in this metric.

For example, *five 9s availability* means 99.999 percent availability. This percentage translates into a total data-not-available time of about five minutes and fifteen seconds per year.

RTO

The recovery time objective (*RTO*) is related to availability. RTO is the time required to return an application to a functioning state after a failure. The RTO metric determines the magnitude of a storage system's high-availability requirement. This metric may be described in intervals as short as milliseconds or as long as days

Chapter 4 Workload

This chapter presents these topics:

A *workload* is the number of work units assigned to a resource over a period of time. CLARiiON's workload is how many reads and write requests it receives and transmits each second. The overall system's applications—how they are used, how they are configured, and when they are used—greatly affect the storage system's workload.

A storage system may contain the data for one or more applications. Storage systems hosting a single application are dedicated, and have one workload. Many storage systems store data for more than one application. This results in two or more workloads. In addition a single application may change its workload profile as a result of the tasks it performs. More than one workload can result in complementary or conflicting storage system resource usage. Each application workload needs to be understood by itself before the overall system workload can be understood. The overall workload is the sum of the individual application workloads. Finally, there are applications whose data is distributed over more than one storage system. This case is treated in the same way as with multiple applications.

Readers may want to read the CLARiiON Storage Objects chapter before this one to familiarize themselves with some of the terms in this section.

Describing the workload

Workloads are described very specifically. You need to be familiar with the following characteristics of the workload's I/O to understand it:

- Type: Sequential or random
- Access: Reads versus writes
- Size: Large-block size or small-block size
- Flow: Steady or bursty
- Threading model: single versus multi-threaded

Types – sequential or random

Random refers to reads or writes from storage locations scattered across the drive. The reads and writes have non-consecutive addresses on the drive. A random-like I/O pattern can also be generated by an application performing I/O operations with multiple files. The metric typically used in describing random I/O is *throughput*. "Small" refers to the size of the I/O in bytes (see the I/O response time section next). Random I/O requires mechanical hard drives to seek between requests. Seeking lengthens an I/O request's response time.

Data locality effects random I/O performance. I/O locations are determined by their LBA addresses. The locality of reference is how widely the I/O locations are distributed on the hard drive. It is a description of how random the random I/O is. Application data sets with high locality have I/O requests that are located closely together on the disk. Data sets with low locality are widely distributed. The more widely distributed the I/Os are, the longer it takes for the disk to perform the I/O. This is because it must spend time repositioning the read/write head a greater distance between I/Os.

Note that EFDs do not need to seek before they access storage, because they are semiconductor-based drives. Data locality does not apply to this type of drive (see the CLARiiON Storage Objects chapter).

Sequential refers to reads or writes to addresses next to or located very close to each other on the drive. The metric describing sequential I/O is *bandwidth*. Read or write operations in sequential I/O are generally contiguous. Streams of sequential I/O can achieve better bandwidth than streams of random I/O because the drives read/write head does not need to be frequently re-positioned over large distances on the hard drive.

Note that random I/O can have either high or low locality. Sequential I/O by its nature has the highest locality.

Rarely does a single application have only one type of I/O. Most workloads are a combination of random and sequential I/O. This is commonly called *mixed* I/O. A mixed workload has different proportions of both I/O types. It is unusual for both I/O types to be performed in equal proportion. Typically, one I/O is the majority type. The ratios of random to sequential can also change over time. Working with workloads with both I/O types requires analysis and tradeoffs to ensure *both* bandwidth and throughput can be optimized.

Access – read versus write

Another aspect of I/O is whether an I/O access is a read or a write. It is important to know which access is in the majority, because the two access types use different amounts of storage-system resources. Very few workloads are all reads or all writes. It is more likely that a workload will be a mixture of reads and writes. The access type is described as the ratio of reads to writes, or a percentage relationship. For example, an on-line transaction processing (OLTP) workload might have a "ratio of 2:1 reads to writes" or "66 percent reads."

Reads consume fewer CLARiiON resources than writes. Reads that find their data in the CLARiiON's cache memory consume the least amount of resources and have the highest throughput (see the SP cache section). However, reads not found in cache have a much lower throughput and higher response time because the data has to be retrieved from storage.

Data changes with write I/Os. Many precautions are taken to ensure that the change is correct and completely executed to ensure the integrity of the data. Writes consume more CLARiiON resources than reads. The overhead of the data protection mechanisms affects the throughput of write operations. Generally, all writes need to be cached, mirrored, and acknowledged. This increases their response time.

Write caching may be disabled or bypassed by some writes and go directly to the drives. Writes that go directly to storage without caching have much higher host response time, which may reduce throughput. EFD-type drives are less affected by operating uncached than mechanical hard drives.

I/O request size – Large-block size or small-block size

Size refers to the size of the I/O in bytes, typically kilobytes, although megabytes is not uncommon. Small and large I/Os are frequently referred to as *small-block* and *large-block*, after the CLARiiON's block-level I/O architecture. The I/O size is determined by the application or the host operating system (*O/S*). Most workloads have more than one I/O size. In many cases, one I/O size makes up the majority of the application's I/O.

For the purposes of this paper, up to and including 16 KB I/Os are considered small, and equal to or greater than 64 KB I/Os are large. Note that I/Os as large as 32 MB are possible. The smallest possible I/O is 512 bytes, which is a single hard-drive sector.

The size of every I/O has a fixed and a variable effect on the throughput of the storage processor cache, back-end bus, and storage. It can also affect SAN bandwidth.

Large-block I/Os on a CLARiiON deliver better bandwidth than small-block I/Os; this is because data is transferred in fewer, larger transactions. If sufficient drives are addressed when executing large I/Os, the back-end bus speed may become the limiting performance factor (see the Back end section for a description). Bus speed can also become a limiting factor; in networked storage systems large I/Os take longer to transmit over the network. In addition, there is a small caching effect. Large block I/Os take longer to mirror in SP caches in addition to filling the caches.

Random-access I/O is typically small-block. For example OLTP workloads frequently have block-sizes equal to or less than 8 KB. Sequential access I/O is mainly large-block, although there are some large-block random workloads such as decision support system (DSS). Small-block sequential workloads are the typical workload for application logs.

Accounting for mixed I/O sizes requires an analysis to determine the most common I/O size. This information is important for sizing cache and RAID-group stripe elements. *Navisphere Analyzer* provides information on I/O sizes. Information about Navisphere® Analyzer and how to use it is available on Powerlink.

Flow – Steady or *bursty*

The rate at which I/O requests are received is important. I/O traffic flow to the storage system can be *steady* with requests having a high regularity, or *bursty,* with requests being sporadic and quickly increasing in volume. Bursty I/O is hard to plan for. Bursts, sometimes called *spikes,* require a margin of storage system performance to be held in reserve. The storage system needs to maintain a reserve capacity of resources, especially cache, to handle the "worst case" demand. Otherwise, user response times may suffer if spikes occur during busy periods.

For example, I/O patterns for OLTP systems may have busy periods where I/O rates peak. This may result from daily reports or from business cycles such as peak traffic periods.

It is prudent to monitor and analyze a workload's bursty behavior. Unanticipated bursty I/O can be an indication of an undiagnosed failure somewhere in the overall system. It can also be the result of an undiagnosed performance bottleneck somewhere in the system.

It is not uncommon for applications to have a random-access, bursty I/O pattern during business hours, and to have a sequential-access, steady I/O pattern at other times. The storage system may need to be reconfigured one or more times during the day to handle the application's different I/O patterns. An example is an OLTP system, whose behavior is bursty during normal operation and becomes steady during the nightly backup.

Threading and concurrency

The degree of concurrency of a workload is the average number of outstanding I/O requests made to the storage system at one time. The way those requests are dispatched depends on the threading model.

A *thread* is a thread of execution. Host-based applications create processes, which contain threads. Threads can be *synchronous* or *asynchronous*. A synchronous thread waits for its I/O to complete before continuing its execution. This wait is sometimes called *pending*. Asynchronous threads do not pend. They continue executing, and may issue additional I/O requests, handling each request as they complete, which may not be the order in which they were issued.

Single-threaded access means only one thread can perform I/O to storage (such as a LUN) at a time. The most extreme case is a single synchronous thread that is highly serialized. Historically, many large-block sequential workloads were single threaded and synchronous.

Asynchronous single threads can still achieve high rates of aggregate performance as the multiple I/O in their queues achieves concurrency.

Multi-threaded access means two or more threads perform I/O to storage at the same time. I/O from the application becomes parallelized. This results in a higher level of throughput. In the past, small-block random workloads were multithreaded. However, it is now common to find large-block sequential workloads that are multi-threaded.

Concurrency is a way to achieve high performance by engaging the multiple service centers, such as drives, on the storage system. However, at some point the service centers can become busy and I/O can start to queue. Queue lengths increase response time, but are usually how an application generates the highest throughput by reducing idle time.

Concurrency can also affect I/O when two types of access are combined on the same mechanical drives. (This is sometimes called *disk contention)*. When some threads perform random I/O while other threads perform sequential I/O, or different block size I/Os are executed by the threads on the same drives, the bandwidth (which is usually high) of sequential access is reduced and the bandwidth of all threads suffer longer response times. This does not apply to EFDs, because EFDs do not seek. However EFDs have a similar effect when mixing reads and writes, which reduces read performance.

I/O characterization and workload

Following are descriptions of six common workloads to demonstrate a wide range of I/O characteristics:

- On-line transaction processing (OLTP)
- Mail system
- File serving
- Decision support system (DSS)
- Backup-to-disk
- Rich media

OLTP workloads are found in database systems for transaction-oriented applications. Examples of these systems include sales ordering, inventory control, and electronic banking. This workload is random-access, majority-reads, small-block, and throughput intensive.

Mail systems are used for messaging and collaboration services application for all sized organizations. Mail systems are a popular application for storage systems. Microsoft Exchange Server is one example of a mail system. This workload is random access, mixed read and write, and throughput intensive.

File Serving workloads are found in storage systems used as a network attached storage (*NAS*) device. This workload requires the storage system to maintain a high number of IOPS. This workload is random access, majority read, and throughput intensive.

A DSS workload is the storage system's hosting of a transactional database with complex queries made to a large number of large tables with different data types. Historically this workload used to be sequential access, recent applications now have random access. In both cases the I/O is majority read, and bandwidth intensive.

The backup-to-disk workload is a backup process that uses disk-based storage (instead of magnetic tape) for archival storage. The storage system copies data, databases, or servers to create an archival duplicate of the source material. Staging primary storage backups to secondary disk-based storage takes less time than staging to tape-based storage. Disk-based backups are also quicker if data recovery is required. The backup data on disks is already online and available for a rapid data restore. This workload is large-block, sequential access, majority write, and bandwidth intensive.

The rich media workload is the storage system's usage in multimedia streaming. Multimedia streaming is presenting video and sound to many end-users. This workload is sequential per stream access, majority read, and bandwidth intensive. It's possible for this workload to be mostly random when a large number of streams are being serviced.

The following table summarizes the I/O patterns typically observed by these workloads.

Table 1 I/O characterization of workloads

WORKLOAD	I/O TYPE		ACCESS TYPE		I/O SIZE		I/O FLOW		DESCRIPTIVE METRIC	
	Random	*Sequential*	*Reads*	*Writes*	*Small*	*Large*	*Steady*	*Bursty*	*Throughput*	*Bandwidth*
OLTP	X		X	X	X			X	X	
Mail system	X		X	X	X	X		X	X	
File serving	X		X	X	X	X		X	X	
DSS	X	X	X			X	X			X
Backup-to-disk		X		X		X	X			X
Rich media		X	X			X	X			X

Note that some workloads have both types selected in a category. This indicates a mixed usage or different access types depending on the infrastructure. For example, in a mail-system workload the I/O size is not markedly small or large block. More information would typically be included a description to accurately describe a workload; for example an OLTP workload might be concisely described as, *"an application generating random I/O. The I/O creates a total of 20,000 IOPS of which 70 percent are reads. The I/O size is 16 KB."*

Know the workload

To implement best practices, you should understand your storage system's workload(s). This requires knowledge of the host applications. Please remember that **applying performance best practices has little effect when the workload's demands exceed the storage system's performance capabilities.**

Performance is an evaluated behavior. It is important to maintain historical records of system performance. **Having performance metrics *before* applying any best practices to judge results saves considerable time and labor.**

New workloads should be benchmarked in non-production environments before entering production. Contact an EMC professional for the proper techniques for estimating and simulating new system performance before going live.

Finally, be aware of any changes in the workload or overall system configuration so that you can understand the change's effect on overall performance. It is prudent to regularly use Navisphere Analyzer to monitor and analyze performance. **Regular monitoring with Analyzer provides the baseline performance metrics for historical comparison. This information can give early warning to unplanned changes in performance**

Chapter 5 Workload

This chapter presents these topics:

The following chapter provides a brief overview of the CLARiiON's software and several important features new users should become familiar with.

FLARE

The FLARE® Operating Environment controls the operation of the CX4 and AX4 series storage systems. FLARE manages all I/O functions of the storage system. FLARE is extensible, so it can supply various replication features depending on configuration.

It is important to know the operating environment revision of FLARE installed on your CLARiiON storage system. This revision number identifies the features available including the requirements of the operational environment and storage system hardware. Also, some default, maximum and minimum parameter settings of features are firmware revision dependent. Generally, CLARiiON CX series and CX3-series storage systems will be at the major FLARE revision 26. The CX4 series will be at revision 28 and later.

Navisphere

Navisphere is a set of applications used to configure, monitor, and manage CLARiiONs. Different Navisphere programs are useful in different storage environments.

Navisphere Manager

Navisphere Manager provides a browser-based interface to the CLARiiON for system management. Through Navisphere Manager, multiple CLARiiONs and hosts can be managed. Typical management tasks include:

- Provisioning (configuring) storage objects, including drives, RAID groups, and LUNs
- Managing the cache
- Monitoring errors
- Provisioning virtual LUNs

Navisphere CLI

The Navisphere Command Line Interface (*CLI*) provides a command-line interface to the CLARiiON. A small number of advanced tuning tasks are only possible through the *CLI*. The Secure CLI (naviseccli) provides a secure credential-based interface enforcing user roles and adding an audit capability.

The CLI can also be used for *scripting*. CLI scripts are executable files containing Navisphere command statements. Scripts can be used to automate storage system management tasks.

Navisphere Express

Navisphere Express is a task-oriented version of Navisphere Manager used to manage the entry-level AX series CLARiiON storage systems.

Navisphere Analyzer

Navisphere Analyzer is the CLARiiON performance analysis tool. The metrics it provides are an important part of CLARiiON performance and availability tuning. Analyzer allows a review (either real-time or off-line) of CLARiiON performance to identify performance bottlenecks. Analyzer also has a CLI interface.

Having recorded historical performance information on the CLARiiON's workload is the first step in performance and availability tuning.

Access Logix

Access Logix™ is a part of FLARE that provides LUN access control for hosts connected to the same storage system. The hosts may be running different operating systems or be virtual hosts.

Access Logix allows creation of *storage groups* to control access to both reading and writing data, configuration changes, and the management of storage system management resources by partitioning LUNs from access by all hosts. It can simplify the organization of information on a storage system, and provide data security by restricting access to information.

Replication layered applications

Replication layered applications, sometimes called *layered apps*, are storage system applications. They provide host-independent services important to overall system maintenance and availability. These applications are licensable software that must be purchased separately with the storage system.

Data Replication is an *availability* function that operates independently of the host(s). CLARiiON's data replication layered apps include:

- MirrorView™
- RecoverPoint
- SAN Copy™
- SnapView™

The important difference in these data replication apps is local versus remote replication. Local replication creates a copy in the local storage system. Remote replication creates a copy on another, sometimes geographically separate, storage system.

Best Practices for individual layered apps are found on Powerlink. For example, the Best Practices for SnapView is the *EMC NetWorker PowerSnap and SnapView for CLARiiON* white paper.

MirrorView

MirrorView/S and MirrorView/A are optional software applications supported by the CX4 series storage systems. Both provide remote replication of data,

MirrorView/S mirrors one or more LUNs *synchronously* (and consistently) from one storage system to another. The secondary data images remain in-sync with the primary during normal operation.

MirrorView/A mirrors one or more LUNs *asynchronously* (and consistently) from one storage system to another. The secondary images are synchronized with the primary periodically. The synchronization interval is adjustable to meet the deployment's architecture. This is typically a distance consideration.

RecoverPoint splitter

RecoverPoint is a data-protection application that provides local and remote data replication of LUNs. It works with any storage system, not just CLARiiON's. It runs on SAN network elements called RecoverPoint appliances, and on hosts. Its redirection feature (called a splitter) can also be installed to run directly on the CLARiiON.

SAN Copy

SAN Copy is a remote replication application. It supports the bulk transfer of data between or within storage systems. Both storage systems do not have to be CLARiiON's, or EMC systems. Using SAN Copy, data can be transferred from storage system to storage system without host involvement.

SnapView

SnapView is an application that allows the creation of either point-in-time copies of storage system data (called *Snapshots*) or full, local mirrors (called *Clones*). This feature is sometimes called *local replication*. These copies can be used for online backups or data replication. Copies can be accessed directly by other applications such as development testing. They can also be used to offload the backup activity from production hosts

Chapter 6 CLARiiON Physical Architecture

This chapter presents these topics:

The CLARiiON has a rack-mounted, modular physical architecture, made up of two or more enclosures. Each enclosure contains the hardware for performing one or more system functions. The CX4 CLARiiON uses the following enclosure types:

- Storage processor enclosure
- Standby power
- Disk array

Storage processor enclosures

CLARiiONs storage-processors are housed in storage processor enclosures. There are two types of storage processor enclosures, storage processor enclosures (SPEs) or disk processor enclosures (DPEs). Note the DPE is a type of storage processor enclosure, but never referred to as an SPE.

Storage processor enclosure-based (SPE) systems

A CLARiiON SPE enclosure provides connectivity, cache, and cooling, but no drives. This design offers the most flexible deployment options. Storage systems in the CLARiiON CX4 series are examples of SPE-based systems.

Disk processor enclosure-based (DPE) systems

The CLARiiON DPE enclosure provides connectivity, cache, and a number of disk drives in the same enclosure. These systems offer the highest density of storage since there is no separate SPE consuming rack space. Storage systems in the CLARiiON AX4 series are examples of DPE-based systems.

Standby power

CLARiiON CX4 series storage systems have redundant power supplies, redundant power distribution, and battery backup to ensure against internal and external power failures.

The following applies to Alternating Current (AC) supplied CLARiiON storage systems only. Direct Current (DC) supplied systems are available, but not discussed. Contact your EMC representative for details on DC power configured storage systems.

The CLARiiON employs *dual-shared power supplies*. That is, for each enclosure in the cabinet, two power supplies share power distribution to it. When one power supply fails, the surviving power supply supports the enclosure.

The CLARiiON cabinet comes equipped with power distribution units (PDUs). The enclosures in the cabinet plug directly in to the PDUs to receive power. Each enclosure is connected to a PDU on the left and a PDU on the right. This ensures redundancy in the event of a power failure. If power fails on either side, the opposite side can maintain power to the entire system.

The standby power supply (SPS) is a power monitoring and battery backup unit. Power monitoring determines if data center power quality variations or disturbances may adversely effect the storage system's operation and will trigger the battery backup function. The SPS provides protection against brief external power fluctuations and for complete power loss.

Some CLARiiON models can be configured for only one SPS. All CLARiiON models can be configured for two SPSs. Dual SPSs provided a higher degree of storage system availability than single SPS installations. CX4 series arrays require only one functional SPS to enable write caching. Maintaining an enable write cache while still guaranteeing data integrity is an important performance feature. Legacy CLARiiON models may require either one or two functional SPSs for write caching to be enabled.

The SPS does *not* function like an uninterruptable power supply (UPS). It is not designed to keep the storage system up and running for long periods in anticipation of power being restored. The SPS functions to protect cache and dump it to the vault when there is power loss.

SPS battery power

The different generations of CLARiiON storage systems use different battery types with different *Amp-hour* ratings. (Amp hours is the measurement of battery capacity.) These batteries are sized to support the connected SP and storage system components required to maintain the write cache long enough for it to be stored to the vault to complete. This process is sometimes called *a dump* to the vault.

SPS readiness testing

The CX4 CLARiiON's SPS system is tested for readiness in two ways: on storage system power up and periodically (weekly) during operation. Both tests verify configuration and correctness of power cabling, and that the SPS battery has a sufficient charge to support the connected SP and storage system components required to maintain the write cache long enough for the dump to the vault to complete. Except for the CX4-120 with one SPS, the weekly SPS test will not be conducted if it will cause the write cache to be disabled. The time of the weekly test is configurable through Navisphere. The interval is not configurable.

CX4 AC power failure behavior

The different generations of CLARiiON also behave differently to power failures. The CLARiiON CX4 series with its persistent cache continues operation though a wider range of failures than legacy CLARiiONs. (See the Vault and write cache availability section.) The CLARiiON CX4 series behavior in a AC power failure is dependent on the failure cause, and the number SPSs installed. The AC power failure scenarios are power lost to:

- Both SPSs
- One SPS in a dual SPS system
- Only SPS in a single SPS system
- The non-SPS-protected-SP in a single SPS system

AC power loss to both SPSs: This is the most serious condition. An example of this is a complete data center AC mains failure. When AC power is lost, all SPSs will report "Battery Online" to the FLARE operating environment. FLARE will regularly poll the SPS status for a short period to determine if the failure is an intermittent power loss or a complete outage. If all SPSs continue to report "Battery Online" through the polling period, the dump of the write cache to the CLARiiON vault is started. After the write cache is written to the vault, FLARE continues to check the SPS state. If "Battery Online" state persists, the SPs and Bus 0 Enclosure 0 (DAE 0) are powered off. When AC power returns, the SPs and Bus 0 Enclosure 0 are automatically powered up and operation is restored.

AC power loss one SPS in a dual SPS system: This condition may occur if one of the two data center AC mains providing power to the CLARiiON fails. As in the full AC power loss scenario, the SPS reports "Battery Online" to the connected SP and the peer SP, which will still be running on AC power from its SPS. In this case, the SPS with no AC power input is shut down by FLARE to prevent unnecessary battery discharge. Both SPs continue operation powered from the power supplies connected to the SPS receiving AC power. The CLARiiON write cache is still enabled.

AC power loss to the only SPS in a single SPS system: The SPS reports "Battery Online" to its connected SP and the peer storage processor, which will still be running on AC power. Note that the peer has no connected SPS. If the SPS reports "Battery Online" through the polling period, the write cache dump to the vault is started. After the write cache dump completes, the storage system continues to operate without write cache at a reduced level of performance until AC power to the SPS is restored and the SPS battery is fully recharged.

AC power loss to the non-SPS-protected-SP in a single SPS system: The SP reports AC power fail to FLARE, but the storage system continues to operate on the power from the peer SP's power supplies. Write cache remains enabled. Note that the CX4-120 with a single SPS installed is an exception. The power supplies for each SP are for its own SP. They are not redundant with the peer SP's unit(s). They are redundant to themselves through the second AC main.

Disk array enclosure

Disk array enclosures (DAEs) offer disk-drive expansion to the SPE or DPE. There are several types of DAEs. The CX4 series DAE hold up to15 drives. The AX4 DAE holds up to 12 drives. In addition, different types of DAEs appear in the legacy CX and CX3 series CLARiiONs.

DAEs and storage

DAEs provide redundant connections both to drives and storage processors. Drives (mechanical hard disks or EFDs) receive their I/O requests through *ports*. All CLARiiON drives are dual-ported. Each drive in the DAE is connected to each SP's back-end bus. The dual-ported storage can receive I/O requests from either or both SPs at any time.

CLARiiON CX series DAEs internal connections to drives and the CLARiiON's back-end bus are Fibre Channel-based. The AX series DAEs are SAS-based point-to-point connections.

Note that Bus 0 Enclosure 0 refers to the DAE housing the CLARiiON's system drives, which includes the vault.

For example, in the following figure, Bus 0 connects SP A and SP B to hard drive 0 in DAE0.

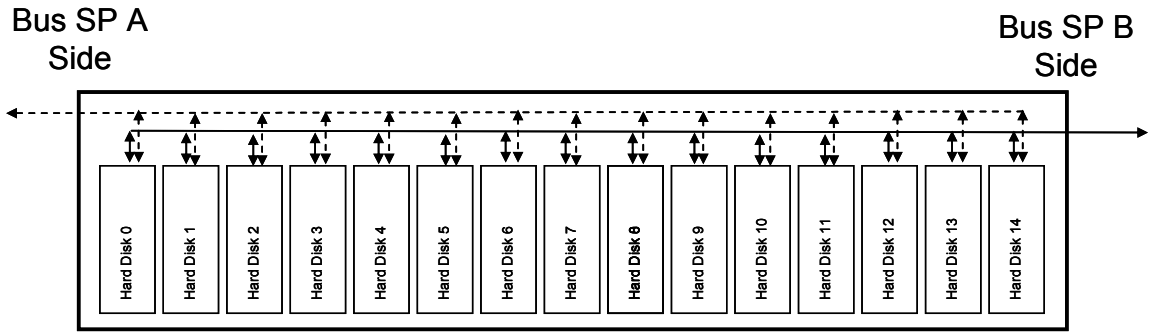

Figure 6 CX4 dual-ported storage connection to a back-end bus

A CX series DAE is supported by link controller cards (LCC). There are two LCCs per DAE. An LCC is the interface between the back-end buses, the DAE, and the drives. (See the Drives and DAEs section for additional information.)

Hardware documentation

To become more familiar with CLARiiON hardware we recommend reading the *Introduction to the EMC CLARiiON CX4 UltraFlex Series* white paper, and the Hardware and Operational Overview for your model CX4.

For example, the overview for the CLARiiON CX4-960 is the *CX4 Model 960 Systems Hardware and Operational Overview*. These documents are downloadable from Powerlink.

Chapter 7 CLARiiON Storage Objects

This chapter presents these topics:

CLARiiON has two types of storage objects:

- *Physical* storage objects – Drives such as mechanical hard drives and EFDs.
- *Logical* storage objects – Storage pools, RAID groups, LUNs, and metaLUNs.

Physical storage objects

Both mechanical and semi-conductor-based drives are used to store data and programs. Hosts read them to retrieve them, and write to them to create, modify, or delete them.

Basic hard drive terminology

The following section provides a brief review of the terms used to describe hard drives and their operation.

A hard drive contains one or more rotating magnetically-coated *platters*. The platters are connected together by a *spindle* that is turned by the drive motor. Sometimes hard drives are also referred to as *spindles*. Data is read from or written onto the platters by *read/write heads*. A drive has as many read/write heads as it has platters. The read/write heads are mounted on an *actuator* arm that positions the read/write heads over the location on the platter to be read or written. The movement of the arm is referred to as a *stroke*. A *full stroke* is a move from the outermost track to the innermost track adjacent to the spindle.

A *track* is a concentric ring-like region around the platter's spindle on which data is stored. A track is divided into *sectors*. A sector is the smallest individually-addressable unit of data. Sometimes sectors are referred to as *blocks*. A sector typically contains 512 bytes of data. CLARiiON sectors contain a total of 512 bytes of data and 8 additional bytes of metadata, totaling 520 bytes. This metadata is a data integrity feature of the CLARiiON. *Logical block addressing* (LBA) is a mapping technique for addressing a sector independent of its physical location on the drive's platters. Many CLARiiON configuration and tuning options are in *block-sized* units, which is 512 bytes.

Hard drives have physical *attachments* that use a specific *protocol*. The attachments are through the previously mentioned ports. All CLARiiON drives are either dual-ported or adapted to be dual-ported. Dual porting provides redundant connectivity. The CLARiiON product line supports hard drives with the Fibre Channel, SAS, and SATA protocol. The different protocols have different performance characteristics.

A hard drive is a computerized device. A microprocessor-based *disk controller* controls the operation of the drive. Among other things, it positions the read/write head in response to I/O requests, executes performance optimizations, data protection functions and diagnostics.

Categorizing mechanical hard drives

Mechanical hard drives are broadly categorized by their:

- Capacity - raw storage capacity
- Speed - measured in the spindle's revolutions per minute (rpm).
- Buffer Size
- Form factor – diameter of the storage media

- Type - attachment protocol.

For example, a drive might be referred to as a *1 TB, 7,200 rpm, 32 MB buffer, 3.5-inch, SATA hard drive.*

Capacity

Hard drives are frequently referred to by their capacity. Typically, capacity is in terms of either nominal gigabytes or terabytes of storage. Some examples are 500 GB or 1 TB.

The raw capacity of a drive is determined by the diameter of the platter, the number of platters within the drive, and the platter's areal density.

Areal density

Areal density is the number of bits that can be stored on an area of a drive's platter. With a high areal density, a large amount of capacity can be put on a single platter. Areal density is measured in Gb/in² or Gb/cm². A high areal density would be about 400 gigabits per square inch. With current technology, this allows for up to 500 GB to be stored on a single 3.5-inch platter. For example, one of the CLARiiON's largest SATA drives has five 3.5-inch 400 GB platters, which total to 2 TB of capacity.

Increasing areal density was necessary to increase drive capacity and performance while maintaining the form factor (3.5 or 2.5 inch) and decreasing drive complexity, power consumption, and cost per GB. With higher areal density and more sensitive read/write heads, fewer platters are needed to reach a given capacity. Fewer platters in turn reduces the number of actuators and drive heads. This makes the drives mechanically simpler. At the same time, it also facilitates improved performance by allowing the drive head to access more data over shorter physical distances on the platter's tracks.

Why capacity quotes can vary

Unfortunately, there is more than one way of quoting storage capacity. It can be:

- Capacity *after* the after the disk is formatted; this it its *usable* capacity, *or user-data* capacity.
- Capacity *before* the disk formatted; this is *nominal* capacity. Formatting the disk will use up some of this capacity.
- Decimal or binary – it can in *decimal base* (base 10) or *binary base* (base 2). This is explained below.

It is not uncommon for new users to wonder why their host's O/S reports that a 1 TB drive has a formatted capacity of 931 GB. The explanation is that hard-drive manufacturers report capacity in *decimal gigabytes*, while host O/Ss report capacity in *binary gigabytes*.

Binary vs. decimal capacity

A byte is always eight bits. However, a kilobyte can be *decimal* (1000 bytes) or *binary* (1024 bytes), depending on the source of the report. The difference between their capacity measurements increases as the unit of measure increases. Table 2 shows the difference between the common capacity measurements.

Table 2 Decimal versus binary capacity

Unit of measure	Binary bytes	Decimal bytes	Percentage difference
Kilobyte	1024	1000	2
Megabyte	1,048,576	1,000,000	5
Gigabyte	1,073,741,824	1,000,000,000	8
Terabyte	1,099,511,627,776	1,000,000,000,000	10

Note that in terabyte capacities, there is a 10 percent difference in the values. A computer O/S using a *binary* gigabyte to measure its available capacity will show a hard drive having a much lower capacity than the manufactures listed capacity, which is listed in *decimal* gigabytes.

The CLARiiON's Navisphere storage management application reports disk capacity in both binary gigabytes and decimal bytes to avoid confusion.

Drive formatting and FLARE private space

When a drive is formatted, its sector size is permanently set. The conventional standard is 512 bytes per sector. Drives on a CLARiiON storage system do not use conventional sectoring. An additional eight bytes per sector is used for storing sector-level data protection information. These additional bytes are used to verify and maintain data integrity. This 520-byte sectoring of a drive reduces the usable capacity of a hard drive by less than two percent. However, this small reduction in capacity has a large positive effect on availability.

Host LUN formatted capacity

Partitioning and formatting reduce the available capacity of a LUN This difference in capacity is due to *metadata*. Metadata is data about the data. It is used to support availability features such as data integrity, recovery, and information security (confidentiality). It also includes file system information

FLARE reserves a small amount of disk capacity for tracking tables and logs. This accounts for about 35 MB per drive.

Different file systems have different amounts of metadata overhead. For example, Linux file systems such as ext2 and ext3 have a smaller amount of metadata than Microsoft's NTFS.

Raw LUN storage

Partitioning and formatting a LUN reduces the available capacity before any user data is written to the drive. However, some applications do not require formatted drives. A raw device is a LUN without a file system, so it does not lose capacity to file-system overhead. Certain applications, such as Oracle, manage data on raw devices, which not only eliminates file system overhead but also improves transactional throughput.

Speed

Rotational speed categorizes hard drives into performance classes. Rotational speeds are measured in revolutions per minute (rpm). The most common rotational speeds for hard drives are:

- 15,000 rpm

- 10,000 rpm
- 7,200 rpm
- 5,400 rpm

The speed at which a disk drive spins (for example 15,000 rpm) is specific to each model disk drive. It does not change. The speed of other drives in the same enclosure or of the back-end bus has no effect on the drive speed.

The rotational speed can be abbreviated using the K-notation. For example, a 15,000 rpm hard drive can be described as being 15k rpm.

Note that rotational speed does *not* apply to EFD type drives.

Buffer size

Hard drives typically have an on-board buffer used by their microcontroller to optimize their operation. The buffer is a managed memory space allowing faster request response time than a disk access. Typical buffer sizes are 16 MB, and 32 MB, where the larger capacities can provide greater hard drive perfrmance.

The advantage of this buffer is in read buffering. The drive can prefetch reads into the buffer by itself. A read miss at the storage processor-level may be a read hit within the hard drive's buffer. This improves performance, expecially in file systems where there are gaps between sequentially arranged file sections drives.

Form factor

Form factor describes an IT industry standardized set of dimensions for a drive. A mechanical hard drive's *form factor* is determined by its platter size. The form factor quoted is the diameter of the hard drive's platters in inches. There are 3.5-, 2.5-, and 1.8-inch form factor mechanical hard drives commercially available. The 2.5-inch form factor is sometimes called *laptop format*. All CLARiiON storage systems currently use the 3.5-inch form factor.

Type

The CLARiiON CX4 and AX4 series support the following types of drives:

- Fibre Channel
- SAS
- SATA
- EFD

These drives have different characteristics that make them more or less appropriate for different workloads. Not all drive types are supported on all models of the CX4 and AX4.

In addition, on some legacy storage systems there are 7,200 rpm Fibre Channel interfaced drives with an ATA head-disk assembly. These drives are sometimes referred to as FATA (Fibre-ATA) or LCFC (Low Cost Fibre Channel). These drives are supported in the Fibre Channel enclosures (for example, DAE2, DAE2P) of the CLARiiON legacy CX300/500/700 series storage systems; they are not supported for use as vault drives.

Hard drive reliability classifications

Enterprise is the highest level of drive reliability. These drives are engineered and manufactured to provide reliable operation at 100% duty cycle over long periods of time. *Near-line* is a lesser, but still very high, standard of reliability. The lowest level is *consumer* or *desktop reliability*. Desktop hard drives are the products available at retail outlets. Desktop drives may be found in hosts, but never in CLARiiONs.

There are many differences between Enterprise and Near-line hard drives. For example, Enterprise drives are typically dual ported vs. single ported, they have faster speed motors (15k rpm vs. 7.2k rpm), their media is "characterized" to verify its magnetic properties, and they have greater structural resistance to vibration. Enterprise drives also have dual controllers: one for platter/head tracking and another for bus arbitration and data dequeue/enqueue and buffering.

The most important difference is drive reliability is terms of the drive error rates. This metric is reported in terms of their unrecoverable error rate (UER). The following table shows typical manufacturer reported mechanical hard drive UERs.

Table 3 Manufacturer-reported mechanical hard drive UERs

Typical manufacturer reported UER values		
Drive reliability rating	**UER (bits)**	**Bytes**
Consumer	<1:10^14	<12.5 TB
Near-line	1:10^15	125 TB
Enterprise	1:10^16	1.25 PB

For example, from the table with a near-line drive, an error will be observed about once, for every 125 TB read from the drive. Also, a 500 GB Enterprise drive would require about 2,500 full drive reads before statistically having a single error.

It is important to understand that a UER is a statistical measurement. That is, there is certain likelihood a particular near-line rated drive will never have a UER. Likewise there is the low probability that an Enterprise rated drive will have a UER after 125 TB of reads.

Fibre Channel hard drives

Fibre Channel hard drives are enterprise-class. They are among the highest-performance and highest-availability CLARiiON hard drives.

CLARiiON CX4 series Fibre Channel hard drives use the 4 Gb/s Fibre Channel protocol to attach directly to the CLARiiON's 4 Gb/s Fibre Channel back-end bus. This provides a theoretical maximum of 400 MB/s burst speed from a RAID group. Under ideal conditions, 90% of this maximum rate is possible as a sustained rate. Legacy CLARiiON hard drives may have 2 GB/s Fibre Channel attachments. The standard Fibre Channel FC-4 level protocol is used to run the SCSI storage protocol. Fibre Channel hard drives have Enterprise-level reliability.

The highest performance Fibre Channel hard drives are 15k rpm. CLARiiON also supports large capacity 10k rpm drives. Fibre Channel hard drives are not available on the entry-level AX4 series.

SAS hard drives

Serial Attach SCSI (SAS) hard drives have equivalent-performance, and availability to CLARiiON Fibre Channel hard drives. SAS hard drives currently use the 3 Gb/s SAS protocol attachment. This has a theoretical bandwidth of 300 MB/s. The 6 Gb/s SAS attachment is another SAS standard. This has a maximum bandwidth of 600 MB/s. SAS drives provide high performance storage for the entry-level AX series. They are not available on the CX4 series. SAS hard drives are Enterprise-level reliable hard drives.

The highest performance SAS hard drives are 15k rpm. In addition the AX series CLARiiON's support large capacity 10k rpm drives.

SATA hard drives

The SATA hard drives provide the most economical storage solutions capacity-wise on the CLARiiON CX4 and AX4 with modest performance.

SATA hard drives currently use the 3 Gb/s Serial ATA II protocol for attachment. The practical burst speed of this attachment is somewhat less than its 300 MB/s maximum. A 6 Gb/s SATA standard attachment also exists. This has a maximum burst speed of about 600 MB/s. 3 Gb/s SATA is compatible with all SAS standards. 6 Gb/s SATA is only upward-compatible with 6 Gb/s SAS. On the CX4 series, SATA drives are interfaced to the CX4's Fibre Channel bus. SATA drives in the AX4 series use the AX4-compatible SAS back-end interconnect for attachment. SATA hard drives are near-line reliability-level drives.

SATA drives are available in 7.2k rpm and 5.4k rpm speeds. SATA hard drives have the largest capacities available. They are very economical for workloads requiring large amounts of storage. The SATA hard drives with 5.4k rpm rotational speeds are the most energy-efficient CLARiiON mechanical hard drives. These 5.4k rpm hard drives are only offered as a set of 15 drives in a single DAE. They are not supported when mixed with other drives in an enclosure.

In addition, there are 10k rpm SATA drives. However, these are only available as replacements on legacy storage systems.

Enterprise Flash Drives (EFDs)

EFDs are semiconductor-based drive. They have no moving parts and have very high performance and reliability.

Metrics such as average service time for EFDs are extremely low when compared to mechanical hard drives. For example, an EFD can retrieve a random read in about 0.1 ms. (The fastest FC drive would take about 4 ms.)

EFD storage uses the Fibre Channel protocol to attach directly to the CLARiiON's 4 Gb/s Fibre Channel back-end bus. EFDs have enterprise-level reliability.

With the current EFD technology, read I/O is significantly faster than write I/O. This makes EFDs ideal for workloads with a significant amount of cache-miss reads. In addition, setting up multiple LUNs on EFD RAID groups owned separately by both storage processors is a useful way to take advantage of EFD random read performance. EFDs can also be an alternative to the use of short-stroked mechanical drives.

All SSDs, including CLARiiON EFDs, are subject to changes in performance as a result of its state and its load.

Externally, the smallest addressable unit of data for an EFD is a sector (512 bytes). However, internally the smallest unit of data that can be written is the page. A page is a multiple number of sectors: eight sectors (4 KB), 16 sectors (8 KB), or 32 sectors (16 KB) depending on the drive model. Empty, unused pages can be written to directly. Finally, an EFD's memory must first be erased before it can be written to. The unit of memory that is erased at a time is the block, which is a multiple number of pages.

Writing an entire block at once is the preferred operation. All the pages within a block must be addressed before a block is written. In addition, only blocks that have been erased may be written. A drive with a large amount of empty blocks has very high performance. While empty blocks exist, a new page of a block is used for every write. When all of a drive's empty blocks are consumed, its performance decreases. This is called the "re-write penalty" or "drive aging." Even a drive that outwardly shows available storage capacity may no longer have empty blocks—all blocks have a combination of valid data and pages that are invalid (due to deletion or overwrite).

Once a drive has no empty blocks, pages marked invalid are reused. Invalid pages are separated from in-use pages to create blocks with only in-use pages and new empty blocks with pages available for writing. The blocks with invalid pages are erased and become the target of all new writes. This separation involves a read-modify-write process.

To create new empty blocks internal to the drive, the contents of one or more blocks are read, in-use pages are packed onto blocks of wholly in-use pages. Packing the in-use pages together allows the creating of empty blocks from the surplus pages that were marked invalid. The empty blocks are erased and used for all new writes. These extra steps are performed by the drive's controller and take time. Every write request to the drive will be slowed by the need to perform block rewrites. Reads may be delayed by previously queued writes requiring a re-write, or by waiting for blocks that are locked while they are consolidated.

Tiered Storage

EMC developed a scheme to present the different types of drives based on performance, capacity, and cost. Called *Tiered Storage*, this method advocates using a mix of drive types on a storage system to satisfy the different performance, capacity, and availability requirements of one or more workloads.

A tiered provisioning strategy assumes that only a small percentage of data should be maintained on the highest cost, high performance drives. The rest is migrated to less expensive storage where it is quickly available when needed.

Moving downward through the tiers, the cost per GB of storage decreases. Performance and availability likewise decrease moving down through the tiers as the drives become slower. It takes longer to resume operations with slower drives in the event of a failure in the lower tiers.

The following figure shows the EMC Tiering scheme.

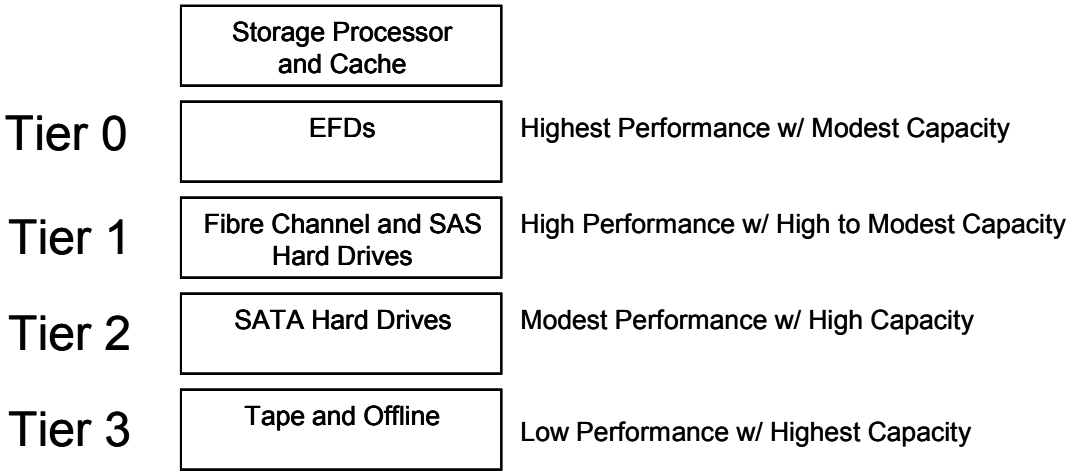

Figure 7 EMC Tiered Storage

The following table gives an example of how workloads may be distributed among tiers.

Table 4 Tiered storage workloads

Tiered Storage workloads	
Tier	**Candidate workload**
Tier 0	DSS OLTP
Tier 1	Batch Transactions CRM DSS ERP File Serving Mail Systems OLTP Software Development Web Server
Tier 2	Backup-to-Disk Data Archiving File Serving Regulatory Compliance
Tier 3	Offline Archival Backup

For example, to get the best system performance and capacity usage, a CLARiiON may use Tier 0 EFDs for high-usage: OLTP database tables, batch processing, or aspects of data warehousing; Tier 1 Fibre Channel hard drives for database logs and less highly used data; Tier 2 SATA hard drives for backup-to-disk usage; and Tier 3 Magnetic Tape cartridges for archival backup.

SMART

Self-Monitoring, Analysis, and Reporting Technology (SMART) is an industry-standardized, drive-based failure detection and monitoring system.

SMART is a built-in availability feature of drives used on the CLARiiON. SMART monitoring is performed independently by the drive's on-board disk controller. It uses the drive's onboard

sensors and the manufacture's diagnostic algorithms to continuously monitor and evaluate drive health.

Should certain attributes (such as read error rates or the number of uncorrectable sectors) exceed preset manufacturer thresholds, a drive's SMART feature signals the CLARiiON's system software of impending drive failure. The CLARiiON uses this information to initiate proactive sparing (see the Global hot sparing section).

Hard drive failure modes

Hard drives are highly reliable storage with mean time between failures (MTBFs) of hundreds of thousands, if not millions of hours. However, a hard drive is a complex mechanical device. Mechanical drives eventually wear out, they can also be damaged by shock, vibration, and excessive heat. EFDs are more reliable, because they have no moving parts. However, they have their own unique wearing profile.

A hard drive has several ways it can fail. The broad categories of failure are sometimes called *failure modes*. Following are the most common failure modes:

- Mechanical
- Media
- Electronic

Mechanical

Mechanical parts failures are usually related to the spindle motor or its bearings, including the read/write head actuator. An example of a mechanical failure is the spindle motor not being able to reach its full operational speed during a spin-up. Excessive heat or vibration is one cause of mechanical failures. Old age is another.

Read/Write head assembly failures can be mechanical or electronic failures. The infamous *head-crash*, where the read/write head scores the magnetic media of the platter, is a mechanical failure.

Rotational Vibration (RV) can contribute to failures in mechanical hard drives. RV is measured in Radians/sec^2. Vibration can come from inside the storage system, for example from neighboring hard drives and system fans. It can also come from outside the storage system, for example from nearby data center heavy equipment such as heating ventilation and air conditioning (HVAC) units. In addition, random and bursty workloads can cause drive vibration.

Some drives in multi-drive environments contain on-board optimizations including Rotational Vibration Feed Forward (RVFF) to attempt handling mechanical errors at the drive-level. RVFF detects and compensates for the ambient vibrations that may cause errors in positioning the drive heads.

Vibration of the hard drive can cause both mechanical and media failures. For example, vibration can increase seek times by delaying the time a hard drive head takes to settle, or cause the drive actuator to vibrate off track resulting in media errors being reported.

Mechanical failures are the most common failure mode. They are also a predicable error indicated by decreasing performance and increasing errors.

Media

Media failures occur due to read or write errors to a sector. A media failure can also relate to a read/write head failure.

A media failure is indicated by a read, write, or verification error occurring in a drive's sectors. When a read error occurs, the sector is remapped. In remapping, the sector's data is transferred to one of a set of sectors for this purpose. As a hard drive ages, the number of media errors increases. A large number of remapped sectors is a predictor of a drive failure.

Electronic

Electronic failures occur in the drive's circuit board when a discreet electronic component fails. There is also a possibility of a software error in the drive controller. A failure of the read/write head may be an electronic or mechanical failure. Heat and vibration are common causes of electronics failure. Old age is another. This failure mode is uncommon. It is also unpredictable.

Logical storage objects

Provisioning is the term used to describe the process of logically configuring the storage system to meet its workload's need for efficient resource usage, capacity, performance, and data security. Logical storage objects include RAID groups, LUNs, and metaLUNs.

RAID

Redundant Array of Independent Disks *(RAID*) is the key technology for performance and availability. An understanding of RAID technology is needed to decide how many disks to use in a RAID group, the capacity they will provide, and estimate their performance in operational and degraded states.

RAID levels provide data protection through redundancy. Do not confuse redundancy with data integrity. While both are related to availability, they are different.

Storage redundancy is the ability of a storage system to continue providing access to data after a hardware failure. Redundancy is achieved through the duplication of drives, power supplies, multiple bus access to drives, and RAID protection schemes.

Redundancy is not a substitute for data backup or point-in-time data replication. Erroneous data, user corrupted data, and data deletions are all redundantly stored on the drives. This "bad data" will still be available (or not in the case of a data deletion) in the event of a drive failure.

Data integrity is maintaining the accuracy of data while on or when going to or from the drive. Data integrity is handled on the CLARiiON through its automatic error detection and correction mechanisms (see the Availability section). It is possible for drives to corrupt data without detection; for example through a media error on the drive's platter. On the CLARiiON drives, data integrity is maintained by the drive's 520-bit sectoring. Additional capacity per sector is allocated to metadata, which is used for error detection and the correction of many sector-level data errors. Error detection and correction are performed both as a continuous background process by SNiiFFER and on data access.

It is important to note that RAID implementations may differ between storage system vendors. A storage system administrator or architect should not make assumptions about a RAID level's operational characteristics based on experience with another vendor's storage system.

RAID levels

The CLARiiON CX4 and AX4 series support six RAID levels. Each RAID level has different capacity, performance and availability characteristics. The two broad types of RAID are mirror and parity. A technique called striping applies to both RAID types. A striped non-redundant RAID-type (RAID 0) and individual disk storage type are separately discussed.

Mirror RAIDs

Mirror-type RAIDs are used when random write performance and availability is more important than cost of capacity. Mirror RAIDs create an exact copy of data on two or more drives to protect the data. The *primary* (drive) has the original copy of the data while the *mirror* (drive) has an exact copy of the data.

Parity RAIDs

Generally, parity-type RAIDs are used when cost of capacity is more important consideration than random I/O performance. (Parity RAID types give better aggregate bandwidth performance than mirroring.) Availability is somewhat less than mirroring and degraded performance is lower. (RAID 6 is the exception regarding availability; see below.) Parity is an algorithmic data protection technique where redundant data is stored to allow errors to be detected and possibly fixed. All parity RAID types are striped. A parity RAID stripes data across three or more disks. Depending on the RAID level the data is protected from the loss of either one or two drives in the RAID.

Striped RAID

A striped RAID is a RAID level that provides the capacity and performance benefits of multiple drives, but no data protection.

CLARiiON supported RAID Levels

The following table provides a brief summary of the RAID levels supported by the CLARiiON.

Table 5 RAID levels summary description

RAID level	Description	Minimum disks
0	Striped unprotected data. Data is striped over a set of hard drives. Provides the same individual access features as the RAID 5 type, but does not have parity information. A single disk failure in the RAID group results in the data on the RAID group being lost.	3
1	Mirror protected data. Provides data redundancy by mirroring its data onto another disk. This is data duplication. This RAID type provides high data availability at the greatest cost in disk space. This RAID level has a maximum size of two disks.	2
3	Striped, parity protected. Provides data redundancy using parity information that is stored on one physical disk in the RAID group.	3
5	Striped, distributed, parity protected data. Provides data redundancy using distributed parity information stored on each drive in the RAID group.	3
6	Striped, distributed, double parity protected data. Provides fault tolerance from two drive failures; meaning the storage system continues to operate with up to two failed drives. RAID 6 has the highest availability level.	4
1/0	Mirror protected data striped for performance. Commonly called *RAID ten*,	2

	sometimes “RAID one-zero.” Provides high performance with high data redundancy but high cost of storage capacity.	

The RAID level does *not* indicate the number of drives used to implement that RAID level. For example, a RAID 5 group does not necessarily have five drives in a stripe. This is a common, mistaken assumption made by newcomers to the technology.

RAID group composition

Some drives types (EFD, Fibre Channel, SAS, and SATA) cannot be bound together in RAID groups. For example, EFDs cannot be bound into RAID groups with mechanical drives. Navisphere prevents this from happening. However, any mechanical hard drive of the same type may be bound into a RAID group.

We highly recommend that RAID groups be created from the same type, speed (15k rpm, 10k rpm, and 7.2k rpm), and capacity drives. Having like drives in a RAID group ensures the highest usable capacity, high reliability, and consistent performance. For example, if you are mixing drives of different capacities in a RAID group, then the capacity of the smallest bound drive will be used from each disk in the group to make up the RAID group’s overall capacity. The full capacity of the RAID group's larger disks will be unused. Likewise, mixing higher rpm drives with lower rpm drives lowers the overall RAID group performance. I/O requests to the slower drives will take longer to complete than I/O to faster drives. This will increase the average RAID group response time and the response time of multi-drive I/O.

Minimum sized RAID groups

Table 5 gives the minimum drives for supported RAID levels.

The minimum RAID group size is two drives. Only mirror-type RAIDs can be configured with this minimum number of drives. Configuring a two-drive RAID 1/0 allows for additional drives to be added to the mirrored RAID group over time; in contrast configuring a 2-drive RAID 1 that has the same initial configuration, but does not allow adding additional drives directly. (LUN Migration or metaLUNs could be used to augment its capacity.)

Striping

Striping is the distribution of data across more than one drive to improve performance and increase available capacity.

Both bandwidth and throughput are increased through striping. This is because either single large or multiple small I/Os can be performed in *parallel* to the disks of the RAID group. Parallelism involves a number of disks servicing a host’s I/O request at the same time. More drives in a stripe can result in a higher degree of parallelism and a better transfer rate.

Likewise, capacity is increased by striping across multiple disks. Each additional drive added to the stripe adds to the total available storage capacity. This is also useful for data sets that require more capacity than a single drive offers.

The performance and capacity benefits of striping depend on the function of all the drives in the stripe. The distribution of data across the stripe creates interdependency between the drives. This interdependency decreases reliability. If one device in the stripe has a permanent data loss, all the data on all the devices is lost. The redundancy of data in parity and mirror-RAID levels mitigates this possibility. However, parity increases a RAID levels complexity, and in some cases decreases its performance. In general, striped RAID levels remain more vulnerable to data loss than mirror levels

Stripe elements

The stripe element is the amount of contiguous data stored on a single disk of the stripe. Stripe elements are measured in 512 byte blocks or in kilobytes (KB). The default CLARiiON stripe element size is 128 blocks, which is 64 KB. The stripe element size is *not* configurable through Navisphere. The default element size has been highly optimized for FLARE. Attempts to "tune" it are likely to reduce performance.

Stripe size

The *stripe size* is the amount of user data in a RAID group stripe. This does not include drives used for parity or mirroring. The stripe is measured in KB. It is calculated by multiplying the number of stripe disks by the stripe element size.

For example, an eight-disk RAID 1/0 has a stripe width of four and a stripe element size of 64 KB has a stripe size of 256 KB (4 * 64 KB). A five-disk RAID 5 (4+1) with a 64 KB stripe element size also has a stripe width of 256 KB drive.

CLARiiON RAID groups

On the CLARiiON, drives are collected together into related groups of a single RAID level, called a *RAID group*. The number of drives in the RAID group and its RAID level define its availability, capacity, and performance. Depending on the RAID level's minimum number of required disks (see Table 5), a group can contain as few as two and as many as 16 hard drives.

For example, the smallest RAID level 6 RAID group would contain four hard drives, the largest 16 disks.

Stripe width

Stripe width is simply the number of drives in a stripe. It applies to distributed parity schemes (RAID 5 and RAID 6) and is only really useful in that context. The width defines the number of drives that can be read from in parallel, assuming random I/O. So it is the number of drives in the stripe.

RAID group notation

RAID group notation uses two numbers in parentheses, separated by a plus (+) sign. For example: *(4+1)*. The sum of the two numbers is the total number of drives in the RAID group. The first number describes the number of drives in the group available for data storage. The second number is the number of drives or drive equivalents in capacity used for data redundancy (see the RAID group drive user or data capacity section).

The following are examples of possible RAID groups:

- A two-disk RAID 1 group: (1+1)
- A five-disk RAID 3 group: (4+1)
- A seven-disk RAID 5 group: (6+1)
- A ten-disk RAID 6 group: (8+2)
- A twelve-disk RAID 1/0 group: (6+6)

Note that RAID group notation does not distinguish between RAID types 5 and 3.

Create a storage pool

RAID groups are created using Navisphere, the CLARiiON's management software. A RAID group is considered a type of *storage pool*. There are also *thin* pools; thin pools are used in the CLARiiON's Virtual Provisioning technology. However, the RAID group is the fundamental drive level of organization.

The maximum number of RAID groups creatable per storage system is model dependent. Entry-level storage systems have fewer RAID groups than higher models.

Binding is the name of the process for creating a RAID group. Any drive of the same type may be bound into a RAID group. The speed and capacity of the drives in a RAID do not need to be the same. However, for performance and maintenance reasons it is prudent to use the same type, speed, and capacity together in RAID groups (see the Hard drive specifications section for details).

Note that in some cases, to configure RAID groups for maximum performance, the Navisphere CLI may be required.

RAID group and LUN capacity

The different RAID levels have different levels of capacity usage. When creating a RAID group, the eventual usable user data capacity depends on the RAID level chosen. The "Total capacity" reported by Navisphere for a RAID group represents all the allocatable storage on the RAID group. It is net of all disk-level formatting, private space usage and RAID protection.

In the LUN properties Navisphere will report "Raw" capacity, which is net of disk-level format, but includes RAID protection. Many service level agreements are described on the basis of raw capacity. Users can use this as a way of estimating total raw capacity used in the storage system.

Note that if an entire RAID group is allocated to a single LUN the "raw" capacity is much less than the capacity listed by the drive vendor, as discussed in the section above. This is due to the binary-to-decimal difference, CLARiiON 520-byte sector and FLARE metadata usage.

RAID group drive user or data capacity

The capacity available for data storage is called the *user capacity* or *data capacity*. The user capacity of drives placed in a RAID group is less than the raw capacity due to the capacity consumed for data protection by both redundancy and data integrity mechanisms.

Mirror RAID

With mirrored RAID levels, the user storage capacity of the drives at the RAID level is half the capacity of the drives. For example, the capacity available for data storage of four unformatted 500 GB hard drives (raw total of 1.834 TB) placed in RAID level 1/0 (2+2) together is an unformatted 917 GB.

Parity RAID

With parity RAID, the ratio of the capacity available for user data-to-data protection changes depending on the RAID level and the number of disks in the RAID group.

For parity RAID levels 3 and 5 one drive's worth of capacity is required for parity. For RAID 6, two drives' worth is used for parity. In RAID 5 and RAID 6 no single drive is dedicated to parity. In these RAID levels, a portion of each drive in the group is consumed by parity. In RAID 3 there is a single dedicated parity drive. This drive contains only parity data, and no user data.

Since the *number of drives' capacity used for parity is fixed,* for any size RAID group the *percentage* of storage capacity dedicated to parity decreases as the number of drives in the parity-RAID group increases. This makes parity-RAID groups with more drives very economical in cost per GB. Note that large RAID groups have a statistically greater chance of multi-drive failure and take longer to rebuild than smaller groups (see below).

For example, the raw capacity available for data storage of a RAID 5 (2+1) is 66 percent; 33 percent of total raw capacity dedicated to data protection and not available to store user data. By increasing this RAID group by two drives, to a RAID 5 (4+1), the percentage of the group's raw capacity dedicated to parity decreases to 20 percent.

Example RAID group capacities

The following table shows the user capacity for the most commonly provisioned RAID groups of the CX4, using the available hard drives. Note again that the user capacity takes into account the CLARiiON's 520-byte sectoring used to ensure data integrity (see the Capacity section), the RAID type's overhead used in redundancy, and FLARE metadata. The user capacity is in binary GB, *rounded* to the whole GB. Binary GB is the capacity reported by hosts. A LUN bound using a RAID group from the table will report that capacity to the host. A host file system formatting will further reduce the available user capacity.

Table 6 RAID group user capacity (GB)

RAID level and drives		300 GB FC	400 GB FC	450 GB FC	500 GB SATA	600 GB FC	750 GB SATA	1 TB SATA	2 TB SATA
RAID 1/0	3+3	805	1100	1208	1376	1610	2064	2751	5503
	4+4	1073	1467	1610	1834	2147	2751	3369	7337
RAID 5	4+1								
	8+1	2147	2934	3221	3669	4294	5503	7337	14675
RAID 6	8+2								
	10+2	2684	3668	4026	4586	5368	6879	9171	18343

RAID 0

RAID 0 is a special case. RAID 0 is the only striped RAID level. There is no data protection with RAID level 0 to consume capacity or inhibit performance. With RAID level 0, the raw storage capacity of the level's disks is the sum of the capacity of the raw drives. In addition, reads and writes to RAID 0 groups are faster than to either mirror or parity RAID levels. However, with no data protection, any drive failure within the group results in total data loss.

Disk

The CLARiiON supports an individual disk (disk) storage type. The disk type functions just like a standard single drive. Individual disk is not a RAID storage type. It does not have the data protection provided by parity or mirroring of data. It does not have the performance or capacity benefits provided by striping. This type can be used for temporary directories that are not critically important.

RAID group rebuilds

The reconstruction of a RAID-group drive's data in the event of a failure is called a *rebuild.* Sometimes this is referred to as *data recovery*. It takes time to recover from a drive error or failure. While a RAID group is rebuilding it is vulnerable to data loss. In addition, there may be some performance loss.

During a rebuild the data in a RAID group's LUNs is recovered (see the LUNs section). A rebuild replaces the failed drive of a LUN's RAID group with an operational drive. With drives in parity RAID groups, data from the failing drive is rebuilt from the group's parity onto the replacing drive. With drives in mirror RAID groups, data is copied from the healthy peer of the failed drive onto the replacing drive.

Rebuild time is an important availability consideration. The time it takes to rebuild is affected by several things. The most influential factors are:

- Drive capacity in use by LUNs
- Drive type (Fibre Chanel, SAS, SATA, EFD), and speed
- RAID type and size of the RAID group (Parity groups only)
- Workload
- Priority

Large drives with a lot of their capacity in use take longer to rebuild than smaller drives; there is more data to rebuild. Faster, higher performance drives such as EFDs rebuild more quickly than slower drives such as SATA hard drives. The RAID level chosen can also extend or shorten the process. Parity RAID types must read parity information from all drives in the group and compute the parity to rebuild; the larger the group, the more data that must be read. The background workload, and its requirement for storage system resources, can shorten or extend the process.

Finally, we need to consider *priority*. Rebuilds can take place at four possible priorities: As Soon As Possible (ASAP), High, Medium, and Low. ASAP executes a rebuild very quickly, but uses a lot of storage system resources. This reprioritization may adversely affect workload performance. High, Medium, and Low priorities are economical in their usage of resources. High priority, naturally, completes in a shorter interval than Low. The default rebuild priority is FLARE version dependent. For FLARE revision 28 and higher the default rebuild priority is High.

Rebuild time can become a major consideration when provisioning. During the rebuilding process, system performance is usually degraded. In addition, the data is vulnerable to additional drive failures during rebuild. It is prudent to make provisioning choices that help rebuilds complete as quickly as possible while maximizing data protection.

The resources most affected are the back-end's buses and the drives. An ASAP rebuild will noticeably slow host response times at the drive level, and in a parity rebuild, enough data may be put on the bus to reduce host bandwidth on that bus *even to groups not rebuilding.*

The effect of a rebuild on the back-end bus can be minimized by using more than one bus to support a parity RAID group (see the Performance effect of buses section). Spreading a parity RAID group's drives across additional buses likewise spreads the rebuild's I/O across the extra buses. This can result in faster rebuild times, making shorter the effect of the rebuild on the performance of application-generated workloads.

Disk power saving (Spin Down)

The CLARiiON supports an electrical-power saving feature called Spin Down. Spin Down allows a RAID group's mechanical disk drives to be configured to enter an electrical-power conserving state when not in use. This state is called *standby*. A RAID group enters standby when it does not receive I/O for a predetermined amount of time. Drives in standby are periodically tested to ensure their readiness. When a host I/O request is made to a LUN with drives in standby, the storage system will pend the I/O while it brings the RAID group's drives to readiness, and then the read or write request is satisfied.

In addition, drives not configured into RAID groups, such as hot spares and unbound drives, can be configured to be kept in standby until needed. Note that only drives qualified by EMC specifically for Spin Down may be spun down. Navisphere detects and presents candidate Spin Down drives to the user from the storage system's provisioned drivers.

LUNs

Logical Units (LUNs) are a logical construct overlaid onto RAID groups. Hosts see LUNs as physical disks. LUNs are frequently referred to as: disks, or *volumes*, or *partitions* depending on the context. LUNs hide the organization and composition of RAID groups from hosts. LUNs are created to allocate capacity, ensure performance, and for information security. Note that information security is *not* data protection. Information security is a confidentiality feature.

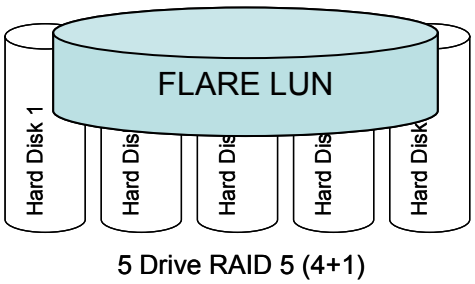

Figure 8 Logical Unit (LUN)

RAID groups and LUNs

The data capacity of a RAID group can be partitioned into one or more LUNs. The maximum number of LUNs per RAID group is FLARE version dependent. For FLARE revision 29 and later the maximum number of LUNs per RAID group is 512. A LUN can be of any size capacity-wise from one block to the maximum capacity of the RAID group. A LUN's capacity is taken equally from all disks of the underlying RAID group.

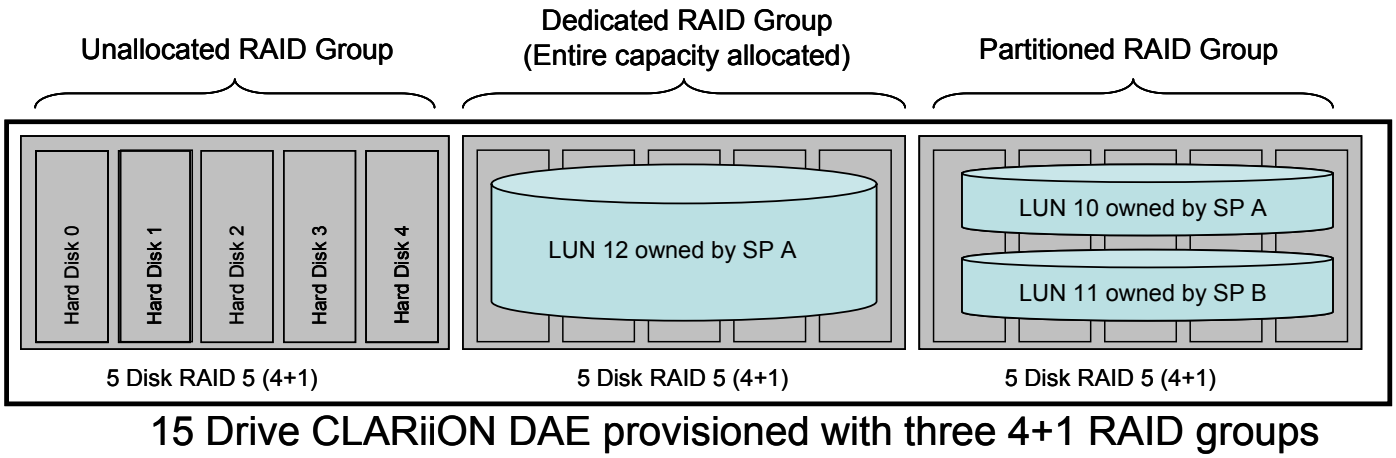

Figure 9 LUN conceptual diagram

Dedicated versus partitioned RAID groups

A RAID group with only one LUN, taking up all the available user capacity, is a *dedicated* RAID group. A RAID group having more than one LUN is a *partitioned* RAID group.

Due to the large capacities of individual drives, a single RAID group can have a very large capacity. LUNs can be used to partition a RAID group's available capacity into smaller parts. The capacity of most RAID groups is shared this way. LUNs on a partitioned RAID group consist of contiguous sets of RAID stripes. Sequential reads and writes and I/O operations using large I/O sizes cause all the drives of the underlying RAID group to work in parallel. This is a very efficient use of the storage system's resources resulting in high bandwidth.

Binding LUNs

The process of creating a LUN is called *binding*. LUNs are bound after RAID groups are created. LUNs are available for use immediately after they are created. However, the bind is not strictly complete until after all the bound storage has been prepared and verified. The preparation and verification occurs through background processing. Its affects are not observable. Depending on the LUN size, verification priority, and the storage system's workload, these two steps (preparation and verification) can vary considerably in duration.

During the preparation step, the storage allocated to the LUN is overwritten with binary zeroes. These zeroes erase any previous data from the storage and set up for the parity calculation. When zeroing is complete, parity and metadata is calculated for the LUN sectors. New systems received directly from EMC have storage "pre-zeroed." This eliminates this step for a drive's initial bind.

Verification involves a Background Verify (*BV*). A BV is a reading of the LUN's parity sectors and verification of their contents. A BV is executed by default after a bind. This default can be manually overridden in Navisphere to make the bind faster. A BV is also scheduled when a storage processor detects a difference between a stripe's parity and the hard drive's sector parity. This is an availability feature of the CLARiiON.

LUN ownership

LUNs are managed and accessed by a single storage processor. This is called *LUN ownership*. By default with FLARE revision 26 and later LUN ownership is automatically assigned within Navisphere to storage processors in a round-robin fashion when a LUN is bound.

Ownership can be manually changed through Navisphere. It may be necessary to change a LUN's ownership to its peer storage processor for performance or availability reasons. Changing ownership is called a *trespass*. For example, a trespass may be needed to balance storage processor usage within the storage system.

LUN Identification (ID)

A LUN is identified by either its LUN Name or its LUN ID.

The *LUN Name* is an identifier. On the current release of FLARE, a LUN can be assigned an identifying text string created by a user. Through Navisphere, the user can create a free form text field of up to 64-characters for host LUN identification. Navisphere only uses the LUN ID numbers (see below) that are assigned by the storage system; it does not use the LUN name, which is changeable. There is no check for duplicate LUN names and no restriction on duplicate names.

LUNs are internally identified on the CLARiiON by their *LUN ID*. The entire range of CLARiiON LUN IDs is from zero to a model-dependent maximum. The LUN IDs and the maximum number of LUNs are also FLARE revision dependent. The following table shows the values for the CX4 series under revisions 28.0 and later of FLARE.

Table 7 LUN IDs by CLARiiON CX4 model, FLARE rev. 28.0

	CX4-120	CX4-240	CX4-480	CX4-960
Max Total LUNs	1024	1024	4096	4096
Max LUN ID	2047	2047	8191	8191

There are two types of LUN IDs: user and private. The available LUN ID numbers (both user and private) always exceed the maximum total of LUNs that can be created on a CLARiiON. However, the sum of user and private LUN IDs in use cannot exceed the CLARiiON model's maximum total of LUNs. (There will always be extra LUN IDs.)

User LUNs created through Navisphere Manager are automatically assigned available user LUN IDs starting from 0 and incrementing by 1 for each LUN created. Users can also manually select an available unused user LUN ID number from the model's range of LUN IDs at the time of creation through Navisphere Manager or the Navisphere CLI.

Private LUNs support user-related LUN data objects such as metaLUN components and Hot Spare LUNs. Private LUNs are assigned LUN IDs from the highest available number in the LUN ID range at the time of creation. Users have no control over the assignment of private LUN IDs.

For example, the maximum number of LUNs on a CLARiiON CX4-120 with revision 28 and later of FLARE is 1,024. The range of all possible LUN IDs (user and Private) that can be assigned is 0 through 2047 (a total of 2,048) any of which are user-selectable, as long as it is available.

LUN expansion

Expansion is the process of adding capacity to a base LUN. However, a LUN cannot have a capacity larger then its underlying RAID group's capacity.

Unbinding

Unbinding a LUN releases all its component drives for reuse resulting in the previously stored data being lost.

RAID group defragmenting

A partitioned RAID group may become fragmented when one of its LUNs is unbound. LUN fragmentation occurs when gaps of unused capacity exist between LUNs in a RAID group. This leaves less contiguous space in the RAID group for new LUNs or for expanding existing LUNs. In addition, this gap or gaps increases the seek distance between LUNs within the RAID group. Through Navisphere, a RAID group can be defragmented to compress the gaps and recover the unused capacity within a RAID group. This makes all the RAID group's in-use capacity contiguous allowing for creation of new, larger LUNs.

Note that RAID group defragmentation is *not* file system defragmentation. (They are frequently confused.) RAID group defragmentation has *no* effect on the positioning of application data within a LUN. In addition, RAID 6 groups cannot be defragmented.

LUN queues

Access to a LUN is mediated through a queuing mechanism. Each arriving I/O for a LUN takes a queue position. I/Os are taken off the queue and processed by the storage processor. The number of queue entries available is called the *LUN queue depth*.

The maximum number of queue entries to a LUN depends on the number of user-data drives in the LUN. The larger the number of data drives in the RAID group the deeper the queue.

For example, a RAID 5 (4+1) based LUN requires 88 concurrent requests for its queue to be full. *Best Practices*[1] lists the calculation for determining LUN queue lengths.

If the queue depth is exceeded, the storage system returns a *queue full* (QFULL) status to the host in response to an I/O. The exact effect of a QFULL on a host is O/S dependent, but it has an adverse effect on performance.

---

[1] *EMC CLARiiON Performance and Availability: Release XX Firmware Update — Applied Best Practices* is available on Powerlink.

QFULLs are rare. Host Bus Adapter (HBA) queue depth settings usually eliminate the chance of them being generated. Note that QFULL can also be triggered at the I/O port-level (see the Front end section for details).

Storage groups

A storage group is an access control mechanism for LUNs. It segregates groups of LUNs from access by specific hosts.

Storage groups are set up using the Access Logix feature. When you configure a storage group, you identify a set of LUNs that will only be used by one or more hosts. The storage system then enforces access to the LUNs from the host. The LUNs are only presented to the hosts in the storage group, and the hosts can only see the LUNs in the group.

A large number of LUNs may be assigned to a storage group and a storage system may have many storage groups. However, the number of LUNs per storage group and the number of storage groups per storage system is CLARiiON model dependent. The *Best Practices* document provides current maximum configurations.

MetaLUNs

MetaLUNs are LUNs created from two or more LUNs. MetaLUNs are one solution to LUNs requiring capacity that exceeds the capacity of a single RAID group's. (Thin provisioning provides another.) The maximum metaLUN size supported by the CLARiiON is 16 exabytes. MetaLUNs also provide a path for later *growing* of a base LUN's capacity without the host adding an additional separate LUN. This reduces the number of LUNs needing to be managed on the host and the storage system. In addition, creating metaLUNs provides a way to increasing LUN performance through the addition of disk drives.

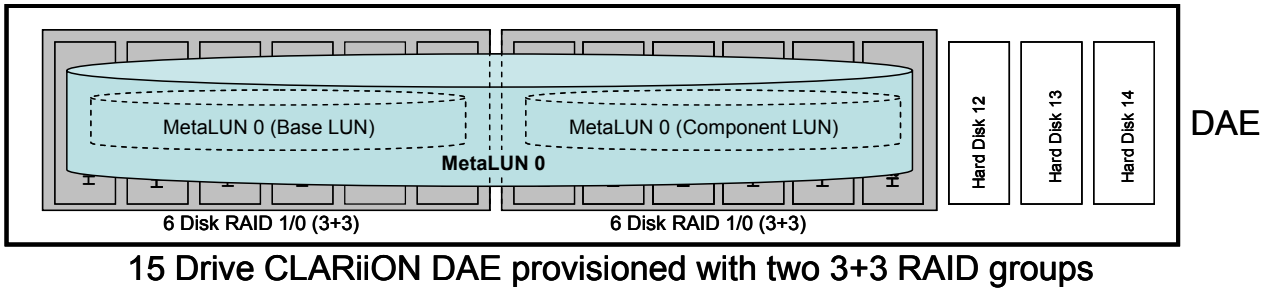

Figure 10 MetaLUN conceptual view

A metaLUN is made up of a *base* LUN and *component* LUNs. The base LUN is a special component LUN that provides the metaLUN's identity for addressing purposes. A metaLUN's capacity is the combined capacities of the base LUN and all of the component LUNs that make it up.

MetaLUN expansion

Expansion is the process of adding capacity to a base LUN or an existing metaLUN. In theory, it is possible to create a metaLUN that includes all the drives on a CLARiiON. However, this may not be practical due to host addressing limitations, provisioning restrictions, and prudent availability concerns.

Additional LUNs can be added to a metaLUN to further increase its capacity. A maximum of 32 FLARE LUNs can be included in a metaLUN component. During metaLUN expansion, the base

LUN (base component's) data is always accessible. The additional capacity is not available, until the expansion completes. MetaLUN expansion can be either *concatenated* or *striped*.

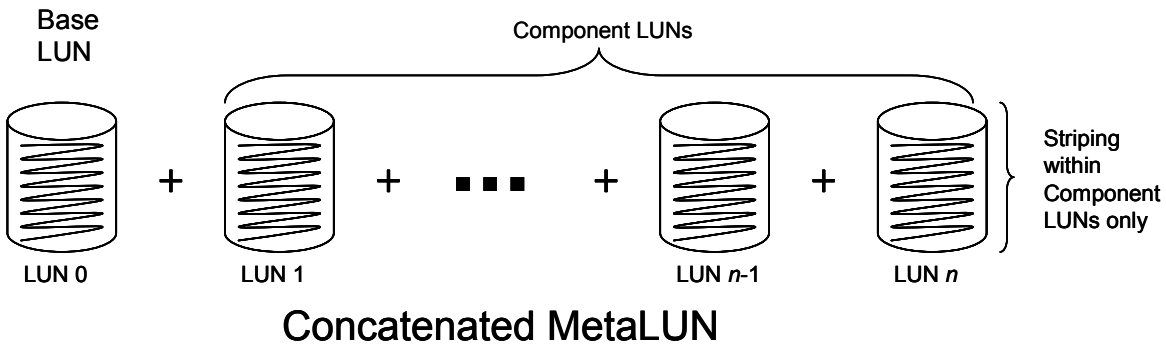

Figure 11 Concatenated metaLUN

Concatenated expansion simply adds additional capacity to the base LUN. Concatenated component LUNs do not have to be the same capacity as the base LUN. When concatenating LUNs, all component LUNs must have the same level of protection. The LUNs being concatenated must all be unprotected (like RAID 0), or all N+1 protected (like RAID 5 or RAID 3), or all N+2 protected (like RAID 6). That is, RAID 6 LUNs can only be concatenated with RAID 6 LUNs. Concatenated expansion is very quick but may not result in a performance benefit from the additional drives.

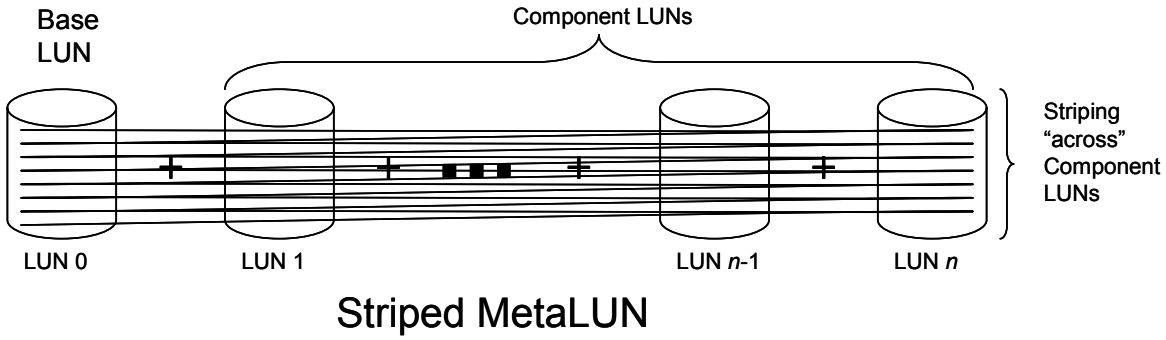

Figure 12 Striped metaLUN

Striped expansion restripes the base LUN's data across the base LUN and component LUNs being added. In striped expansion, all LUNs must be the same capacity and have the same underlying RAID group level. In addition, all LUN RAID group disks must be the same disk type. Performance is improved due to the increased number of drives being striped (see the Striping performance section). Striped expansion is the CLARiiON metaLUN default.

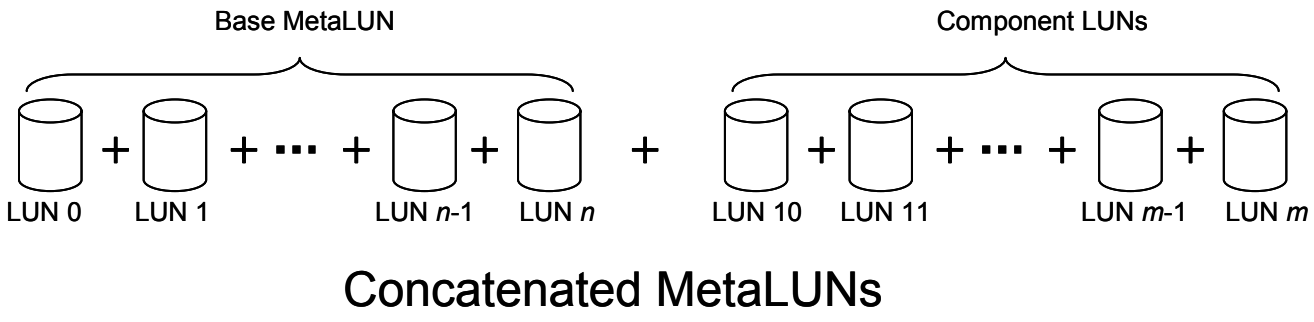

Figure 13 Concatenated components (metaLUNs)

All metaLUNs have a base component, called component 0. MetaLUNs can be expanded by concatenating additional LUNs on to the base component to make very large volumes. This creates a hierarchy of components. A metaLUN made up of two or more components up to a maximum of 16 components.

Unbinding

Unbinding a metaLUN releases all of its component LUNs, resulting in the previously stored data being lost.

Virtual Provisioning and thin LUNs

Virtual Provisioning™ provides for the thin provisioning of LUNs. Thin LUNs present more storage to an application than is physically available. The presentation of storage not physically available avoids overprovisioning of the hosts and the underutilization of the storage system's capacity. When a thin LUN eventually requires additional physical storage, capacity is nondisruptively and automatically added from a reserve storage pool that has previously been provisioned. Provisioning allocates drives to the pool. In addition, the storage pool's capacity can be nondisruptively and incrementally added to with no effect on the pool's thin LUNs.

The provisioning of thin LUNs is performed through Navisphere with commands also available through the Navisphere CLI.

Once created, thinly provisioned LUNs are largely automatic in their upkeep. This simplifies and reduces the administrative actions required to maintain the storage system. An initial investment in planning to correctly provision the storage pool will result in the best ongoing thin LUN performance, capacity utilization, and availability.

Conceptually, the thinly provisioned storage pool is a file system overlaid onto a traditional RAID group organization of hard disks. This file system imparts an overhead on thinly provisioned LUN's performance and capacity utilization. In addition, availability must be considered at the storage pool-level with thin provisioning. Always consider the number of the drives and capacity of the drives being used in the pool and the RAID group organization selected by thin provisioning. Apply the same availability discipline used with traditional LUNs to the automatically created thin provisioning RAID groups. (See the Storage Object Availability chapter.)

Thin provisioning provides scalable storage system capacity with low maintenance. Thin provisioning should not be used for workloads that require deterministic performance or the highest availability. In those cases traditional LUNs should be provisioned.

Chapter 8 Storage Object Performance

This chapter presents these topics:

Drive performance

The CLARiiON includes both hard-drive and non-volatile memory-based drives.

Mechanical hard drives are the traditional storage on CLARiiONs. Hard drive performance has traditionally been a product of: spinning platter rpms, seek times, and burst-transfer rates.

Solid-state drives, known in the IT industry as *SSD*s, are a semiconductor-based drive. On the CLARiiON these drives are referred to as *Enterprise Flash Drives* (EFDs). They have higher availability and performance in comparison to traditional hard drives. They have no moving parts to wear out. They do not seek, so I/Os are executed very quickly. However, with current technology these drives have only modest capacity; they execute reads more quickly than writes, and sequential I/O is not particularly efficient. EFDs also have unique provisioning needs.

Hard drive specifications

There are three important specifications for describing mechanical hard drive performance:

- Seek time - the time it takes to move read/write heads between tracks.
- Rotational latency - the time it takes the platter to move beneath the drive's read/write head.
- Transfer rate - the bandwidth of a hard drive.

Understanding the relationship between these specifications helps to classify a hard drive's performance. These values are important in determining two important hard drive performance metrics: throughput and response time.

Seek time

Seek time is the time it takes to move read/write heads between tracks. Seek time is measured in milliseconds. Seek times vary from drive to drive.

The average seek time is a frequently used metric; although both read and write seek times may also be specified. A typical seek time for a 15k rpm Fibre Channel hard drive is 3.5 ms to read and 4.0 ms to write; resulting in an average seek time of 3.8 ms. Seek time can also be measured track-to-track, or across the radius of the platter, a *full-stroke*.

Minimizing the amount of time a hard drive spends seeking improves performance. Sequential I/O has the lowest seek time, because the read/write heads track continuously on the platters with little or no seeking. Random I/O has longer seek times, because the read/write heads are constantly being positioned and repositioned on new locations on the platters.

Latency

The important factor relating to hard drive performance is *rotational latency.* Latency is the time it takes the platter to move beneath the drives read/write head. Latency is measured in milliseconds. Higher speed drives, where speed is measured in rpm, provide noticeably higher overall random-access throughput and slightly shorter response times than slower drives. Table 8 shows the relationship between average latency and rotational speeds.

Transfer rate

At the drive level, transfer rate is the bandwidth of a hard drive. Transfer rate is measured in MB/s. Transfer rate is further divided into internal and external rates.

The internal rate is how fast data can actually be read and written to the disk's platters. Transfer rate is higher on outer tracks than on inner tracks. For example, for sequential bandwidth applications using a 15k rpm Fibre Channel hard drive, a typical internal transfer rate is 50 MB/s for inner tracks to 100 MB/s for outer tracks. (See the section on Short stroking.)

External transfer rate is the rate at which data can be transferred between the drive's attachment connector and the HBA or NIC. On the CLARiiON, multiple drives of a RAID group share this bandwidth. Note that burst transfer rates given by manufacturers are not seen with a storage system where a shared bus is in use. The manufacturer burst rates assume an internal connection, which allows direct access to system memory for the drive. With storage system's bus implementation, the actual transfer rate is determined more by the back-end transfer protocol, arbitration times, and capacity of the bus.

Hard drive queues

A hard drive executes one I/O at a time. If more than one request for an I/O read or write is sent to a disk, the requests are temporarily held in a disk queue for later execution. If the disk does not have on-board queuing, the I/O requests are stored on the disk controller. Un-optimized disks execute requests stored on the queue in the order in which they were received.

Drive queues allow for efficient use of the drive. The queuing sets up for constant usage. A drive queue depth of at least two is needed for a drive to work efficiently. In addition, there is potential for optimizations to be performed on the queue's entries that can improve I/O performance.

When a disk reaches a queue size between six (for slower drives) and 20 (for faster drives) it is approaching its performance limit. This can occur because the bandwidth of the drive attachment typically exceeds the drive transfer rate. Some optimizations can be performed on queue entries as the queue lengthens. However, overall the disk queue depth is a large factor in drive response time.

Any time the disk spends seeking or waiting for a sector to spin under its head is wasted time because no data is being transferred. When data on successive sectors of a disk are accessed, no time is wasted seeking. The access pattern of reading or writing successive sectors on disk is called *sequential access* and the service time for sequential access is very low**.** The service time for a sequential read can be under 1 ms, so a much higher queue can be handled than for a random read, which has a service time of 4 to 6 ms.

The hard drive's disk controller may perform operations to improve response time. These optimizations try to maximize bandwidth by optimizing I/O to take advantage of the sequential nature of disk drive operation.

Command tag queuing (Fibre Channel and SAS drives) and *Native Command Queuing* (*NCQ*) on SATA drives use optimizations such as *The Elevator Algorithm*, to improve random I/O performance. CTQ and NCQ reorder I/O requests in the disk's queue to minimize the need to reposition the read/write head resulting in sequential access. Enterprise class drives perform these optimizations more efficiently than other classes of drives.

Calculating hard drive performance

There are four basic hard-disk performance metrics used to compare hard drives. These metrics are calculated from hard-drive specifications. The metrics are:

- Average service time
- Average response time
- Throughput
- Bandwidth

Average service time describes the average time it takes for a hard drive to serve-up data. Average response time describes how much data can be served-up to the host application. Bandwidth and throughput are two different descriptions of the quantity of data that can be read or written. They are dependent on the workload's I/O characteristics.

Average service time

Average service time measures the average time a disk can execute a single I/O. Service time is measured in milliseconds. The general formula for average service time is:

Average Service Time = Average Seek Time + Average Latency + Transfer Time

For practical purposes the transfer time is zero on the CLARiiON for small I/Os.

For example, to calculate the average service time for a 7.2k rpm hard drive with a read seek time of 8.2 ms and a write seek time of 9.2 ms.

- Write Seek Time: 9.2 ms
- Read Seek Time: 8.2 ms
- Average Latency: 4.2 ms (from Table 8)
- Transfer Time: 0 ms

Average Service Time = 12.9 ms = ((9.2 ms + 8.2 ms) / 2) + 4.2 ms + 0 ms

Note that seek time is a large component of average service time. In workloads characterized by random I/O, which is seek-intensive, this component of the measurement provides an important comparison of drive performance. In workloads characterized by sequential I/O, where seeks are minimized, the differences in the average latency values are more significant when comparing drive performance.

Average response time

Average response time is the duration from when a request is enqueued, to the moment the disk executes the request. The simplified calculation of average response time is as follows:

Response Time = (Queue Depth+ 1) * Average Service Time

For example, the response time of a drive with an average response time of six milliseconds, and a queue averaging six entries.

- Queue Depth: 6
- Service Time: 6 ms

Response Time = 42 ms. = (6+1) * 6 ms

Throughput and bandwidth

Throughput and bandwidth are two measures of how much data a hard drive can read and write under a specific I/O regime.

The throughput of a drive is measured in IOPS. The number of I/Os per second (IOPS) a drive is capable of is determined by the seek time, and the average latency. The drive's IOPS are affected by the ratio of reads to writes in the workload.

The general formula for calculating random I/O IOPS is:

IOPS = (1 IO /(Average Seek Time + Average Latency)) * 1024 ms/sec.

Bandwidth is the quantity of data that can transfer in a second with the supported number of IOPS. Bandwidth is given by multiplying the I/O size times the IOPS.

The general formula for calculating bandwidth with I/O size measured in KB is:

Bandwidth = (I/O Size / 1024 KB/MB) * IOPS

For example, the throughput and bandwidth for a 15k rpm hard drive with a 4.1 ms write seek time and a 3.6 ms read seek time in a workload with a 70:30 mix of read to write operations of 8 KB I/Os is:

- Write seek time: 4.1 ms
- Write ratio: 0.3
- Read seek time: 3.6 ms
- Read Ratio: 0.7
- Average Latency: 2.0 ms (from Table 8)
- I/O Size: 8 KB

Throughput = 168.0 IOPS = (1/((3.6 ms * 0.3)+(4.1 ms* 0.7)+2.0 ms))*1000 ms/sec

Bandwidth = 1.3 MB/s = (8.0 KB / 1024 KB/MB) * 168.0 IOPS

A common calculation error

Network and bus speed calculations can easily be in error. The correct rate is not always obvious. Network speed is always in decimal notation. Capacity notation is in binary. For example, a Gigabit Ethernet is 1000 Mb, not 1024 Mb.

MHz is always 1,000,000/sec, while MB/s is 1,048,576 bytes/s. 1 MB/s (megabyte per second) equals 8 Mb/s (megabits per second). To calculate the bandwidth, multiply the bus width (in bytes) by the decimal speed, then divide by the decimal unit.

For example, a 64-bit bus (eight bytes) running at 100 MHz would have a theoretical bandwidth of:

763 MB/s ((8*100,000,000)/1,048,576) *not* 800 MB/s

Hard drive speed and performance

The rotational speed of a hard drives has a large effect on performance. The speed of high-rpm hard drives affects the I/O latency and the efficiency of storage system cache usage.

Table 8 Spindle rpm to latency relationship

Spindle rpm	Average latency (ms)
5,400	5.6
7,200	4.2
10,000	3.0
15,000	2.0

The faster rotational speeds result in less latency with the I/Os. High-rpm drives have a direct positive effect on random-read performance and a lesser effect on random writes.

Sequential reads and writes do not benefit as highly as random I/O from the faster rotational speed. However, they do increase the sequential read/write bandwidth over hard drives

The rotational speed of a drive affects the rate of storage system cache flushing and filling. During cache flushing, data must be quickly written from the write cache to the drives (see the Write cache management section). Faster drives are particularly valuable under conditions when cache is being bypassed or cache is disabled. Likewise random reads that typically miss in cache, occur faster. The benefit is a higher overall I/O rate for the storage system.

Hard drive capacity utilization and performance

Drive capacity has an effect on performance that is related to seek time.

Larger-capacity drives of a given type and speed may offer better performance than smaller capacity drives. This is because on a large capacity drive, with the same amount of data, the track-to-track seek distance is less, because there is more data per track. Contiguous data on a large capacity drive will occupy fewer tracks that are closer together. This will take less time to access. On a smaller capacity drive that same amount of data will occupy a greater number of tracks. At worst case, the smaller drive may have to seek over the entire radius of the platter.

Mechanical hard drive performance comparison

The following table summarizes the characteristics of mechanical hard drives available with the CLARiiON CX4 and AX4 series.

Table 9 Mechanical hard drive performance factors

	Fibre Channel	SAS	SATA
Rotational Speed (rpm)	15k, 10k	15k, 10k	7.2k, 5.4k
Attachment Speed (Gb/s)	4	3	3

Typical Average Seek Time (ms)*	4.0	3.8	9.0

*Higher rpm hard drive listed.

The figure below shows the relative performance of a representative sample of the different types and speeds of hard drives on the basis of service time.

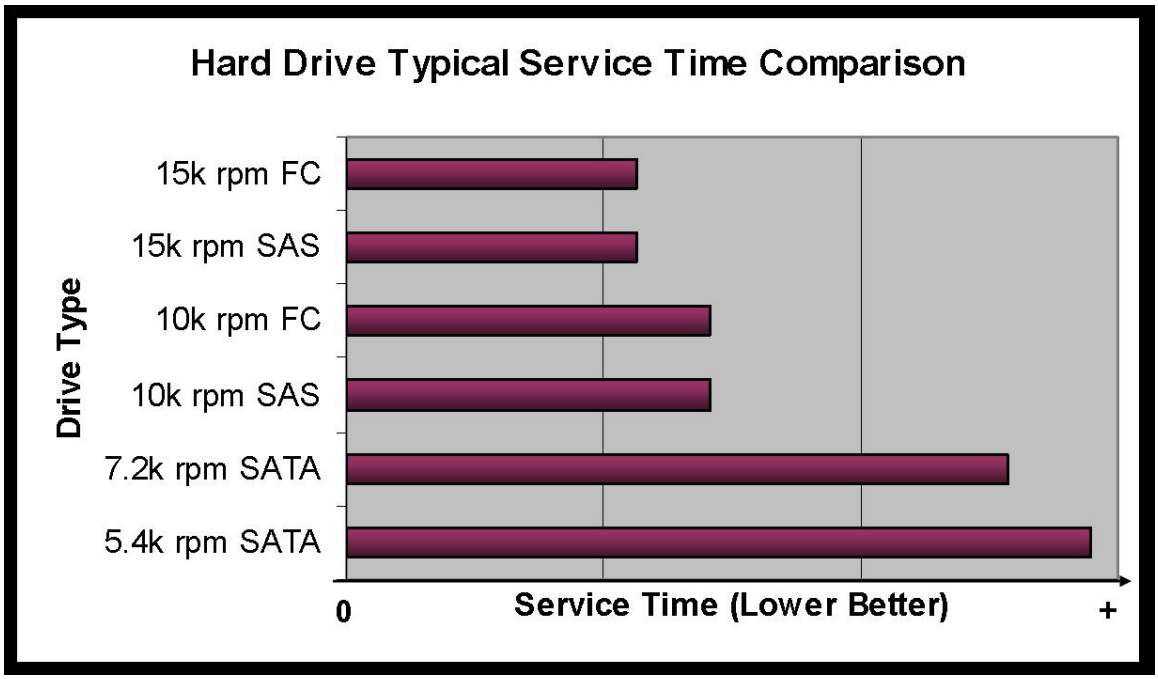

Figure 14 Mechanical hard drive typical service time comparison

From the figure, you can see that a hard drive's rotational speed has the largest effect on its performance. The 15k rpm Fibre Channel and SAS drives have the lowest service time, giving them the best performance. The 10k rpm Fibre Channel and SAS drives provide an intermediate level of performance. The lower rpm SATA drives have more modest performance. **Note that the difference in service times is an important consideration in random workloads, but less so in workloads characterized by sequential I/O.**

RAID group performance

The different RAID levels have different performance capabilities.

Striping is responsible for the largest part of a RAID level's performance. The remaining differences in RAID-level performance are only apparent under specific conditions involving the storage system's cache.

Striping performance

In I/O to a striped RAID level, the host I/O request is broken down into one or more I/O requests to the individual disks of the striped-RAID group based on I/O size and location. These I/Os are sometimes called *back- end I/O* or *disk I/O*. The characteristics of how a RAID implementation executes back-end IO has a big effect on performance.

RAID 1/0, 0, and all parity RAID levels are striped.

Generally, a striped RAID level has better *read* performance than an un-striped RAID level. This is because they use less back-end the bandwidth. Striped parity-type RAIDs can deliver higher write bandwidth than mirror RAID types.

Disk striping affects the amount of data that each disk transfers before seeking for the next I/O.

Striping has two beneficial effects: distribution of random I/O over more than one drive, and large-block parallel access.

In general for any given workload characterized by random I/O, more drives means better performance. Getting increased performance with sequential I/O on the CLARiiON is more complex. There is a dependency on the implementation details of CLARiiON RAID operation that must be considered.

High bandwidth, especially in random, large-block workloads, is dependent on the I/O size at the drive. RAID groups with fewer drives deliver higher *per drive* bandwidth than groups with a large number of drives. This is because larger I/Os can be sent to fewer drives. The highest bandwidth may be achieved by using: a larger number of RAID groups with a smaller number of drives per group, versus one large group. This also benefits by reducing contending threads. This is because there will be fewer LUNs on a small RAID group, for a given LUN size.

Stripe element size

The ideal stripe element size maximizes the amount of data each drive transfers in a single I/O request. The CLARiiON's stripe element size is fixed at a value that has been highly optimized for use with its hardware and software architecture. The stripe element size is 64 KB.

Stripe size

Stripe size is the number of drives in the stripe times the stripe element size.

Disk crossings (alignment)

I/Os that cannot be accommodated by a single stripe element cross to a second disk. These are called *disk crossing I/Os*. Disk crossings are not necessarily a sign of a problem. It is normal for I/O larger than the stripe element size to cross disks. In the case of I/O smaller than the stripe element size, disk crossing I/Os may have a modest adverse affect on performance.

In a disk crossing, an I/O is broken across two drives instead of more efficiently on only one. This is the most common misalignment case. The splitting of the I/O across two drives increases the load on the drives and the back end. This results in for a given I/O rate you having longer queues. Longer queues increases response times for any I/O a drive receives. Even if the disk operations are buffered by cache, the effect can be detrimental, as it is still adding to disk queues, which affect all disk-related performance. Random reads, which by nature require disk access, are also affected, both directly (waiting for two drives to return data) and indirectly (making the disks busier than they need to be).

Alignment is how the data is laid on the set of a RAID group's stripe elements. It ensures that I/O more neatly fits into stripe elements. File system metadata can cause misalignment. While an aligned file system would quickly service the I/O with a single drive a misaligned file system requires two for some portion of the I/Os. It is important to note that with its default 64 KB stripe element size, any I/O larger than 64 KB will involve a disk crossing.

LUNs can be aligned using file system utilities. Alignment is performed after the LUN is bound and presented to the host. The alignment process can only be performed on a LUN with no data. Alignment cannot be performed on previously partitioned LUNs.

Full stripe writes

The CLARiiON in its normal cached operation optimizes its disk I/O to perform full-stripe writes when possible. A full stripe write is sometimes called an *MR3 Write*. In a full stripe write either the I/O incoming is large enough to fill the stripe and aligned, or the storage processor accumulates write I/Os in write cache. Depending on the locality of the data, smaller I/O writes can be coalesced into fewer larger writes. Individual write requests can be sequenced until an entire stripe is cached before being written. This makes for an efficient use of the back-end bus and drives.

Note that if uncached I/O is performed such as when cache bypass is in effect, if the I/O completely fills the stripe or is an integer multiple of the stripe width, and is aligned, a full-stripe write is performed. However, when uncached it is difficult to control and ensure a full-stripe write. For this reason it is hard to get consistent high bandwidth with the cache disabled.

Cached performance

Under normal, fully cached operating conditions the performance of the different RAID levels on the CLARiiON is essentially the same. The CLARiiON's large physical cache and sophisticated caching mechanism result in optimally execute I/O performance. This caching results in a leveling of any performance differences between the RAID levels.

A small difference is seen when write cache is full, and flushing occurs (see the Memory section). Generally, RAID 1/0 will have a large advantage in flushing random writes from the write cache. RAID 3 and RAID 5 (even RAID 6) will have an advantage flushing large/sequential writes from the cache.

Uncached performance

Under circumstances of uncached write I/O, or the rare disabled write cache operation, the RAID level has a performance effect. Uncached I/O can occur when a cache is disabled by the user. Caching may be disabled on certain LUNS to give preference for cache usage to other LUNs. I/O will also bypass write cache based on the LUN Write-aside setting. LUNs made up of EFDs have write cache disabled by default. A disabled cache can also occur when the storage system is in a degraded mode. The high-availability cache of the CX4 series also makes this state very uncommon.

See the SP cache section for information on uncached I/O and disabled cache conditions.

RAID level performance differences

The following section describes the relative performance differences between the RAID levels based on I/O Type and Access Type. *Note that these differences would only be noticed under uncached, or cache full I/O conditions.*

RAID level performance: parity versus mirror

For individual I/Os, there is no read performance difference between parity and mirror RAID levels, while there is a write performance difference.

Back-end I/O operations

Read and write requests to LUNs are different from read and write I/O operations on the CLARiiON's back end. A single LUN read or write request results in one or more read or write I/O operations on the back-end bus and to the drives of a LUN's RAID group.

A CLARiiON read request generates a single back-end read (or more if larger than the stripe element size) I/O operation to storage.

One CLARiiON write request can generate one or more I/O operations. It is the number of I/Os generated by write requests that is the key difference in RAID level performance.

The number of back-end write I/Os depends on the RAID level and the operation of the cache. A CLARiiON full stripe write to a parity RAID level is a single back-end write I/O operation to each drive in the group (see the Full stripe writes section). A full stripe write to a mirror RAID is also a back-end write to each stripe, one to the primary stripe and another to the mirroring stripe. Because there are more drives to service a stripe of any arbitrary size, more data in is sent to the back end.

Random I/O is different. Assuming an I/O is smaller than a stripe element, mirror RAID writes generate two back-end write I/Os—one to each drive. A random write request to a parity RAID will generate two read and two write I/O operations.

For example, an uncached, RAID 5, random, 8K write request:

1. Reads destination data sector(s) from the drive.
2. Reads parity from drive holding parity for this stripe (new parity is calculated from this data).
3. Writes a new sector including data to destination drive.
4. Writes (new) parity to the drive holding parity for this stripe.

Mirror RAID levels (RAID 1 and RAID 1/0) perform better with small-block and random writes. For random writes, parity RAID levels (RAID 3, RAID 5, and RAID 6) require the storage processor to read, calculate, and write a parity sector for each write operation. One parity calculation is needed for RAID 5 and RAID 3. Two parity sector calculations are needed for RAID 6. The time and resources needed to perform this calculation adversely affect parity RAID level performance during random writes.

Parity RAID levels perform better with large-block and sequential writes. With sequential writes, when full-stripe writes are possible, parity RAIDs performing fewer writes perform better than mirror RAIDs. This is because less redundant data is written: a single parity element versus the two complete stripes of the mirror level RAID.

Mirror RAID levels (RAID 1 and RAID 1/0)

RAID 1

RAID 1 offers the best bandwidth on a per-drive basis. Random write performance is better than most RAID levels but limited by the maximum RAID group level size of two drives. Sequential read performance is about the same as a single drive.

In RAID 1, a read is one back-end operation and a write is two.

RAID 1 is not available on AX4-5 series storage systems.

RAID 1/0

RAID 1/0 receives a performance benefit from striping. RAID 1/0 includes some optimization of reads that takes advantage of two drives with identical data. It has very good random read and write performance. It also has good sequential read and write performance.

A RAID 1/0 group can contain between two and 16 drives. A two-drive RAID 1/0 has identical performance to a RAID 1 two-drive RAID group. However, a two-drive RAID 1/0 has the option of RAID group expansion to contain more drives. (A two-drive RAID 1 cannot be expanded.)

In RAID 1/0, a read is one back-end operation and a write is two.

Parity RAID levels (RAID 3, 5, and 6)

RAID 3

RAID 3 has good random read performance due to striping, but not as good as RAID 5. There are no reads from parity drive. Random write performance suffers due to parity computation, and the chance of bottlenecking at the single dedicated parity disk. Sequential read performance is likewise good due to striping. Sequential write performance is excellent as RAID 3 has the shortest simplest code path of all parity RAID types.

In RAID 3, a read is one back-end operation and a write is one for a full stripe and up to four for less than a stripe.

RAID 5

RAID 5 has excellent random read performance. Performance improves with increasing numbers of disks in the RAID group. Random write performance is fair due to the parity calculation. RAID 5 random write performance is better than RAID 3 because there is no dedicated parity disk to bottleneck. Sequential read performance is good. Sequential Write performance is good.

In RAID 5, a read is one back-end operation and a write is one for a cached full stripe and up to four for a single small I/O less than a stripe.

RAID 6

RAID 6 has similar performance to RAID 5. Where RAID 6 suffers in comparison is in the requirement for the additional parity calculation. It has the lowest random-write performance (equal user data drive count) of any RAID level. It has excellent random read performance. Sequential read performance is good. Performance improves with smaller stripe widths. Sequential write performance is fair.

In RAID 6, a read is one back-end operation and a write is one for a cached full stripe and six for a single small I/O less than a stripe.

RAID 0

RAID 0 is a special case. It offers all the performance advantages of a striped RAID. It also does not have the computational overhead of parity calculation. However, there is no data protection built into RAID 0.

RAID 0 has very good random read and random write performance, particularly with large stripes. Sequential reads and writes are also very good.

In RAID 0, a read is one back-end operation and a write is one for a full stripe and one for less than a stripe.

RAID 0 is not available on AX4-5 series storage systems.

RAID group performance calculation

The throughput or bandwidth performance of a RAID group depends on the provisioning of the RAID group and the I/O it is performing.

Throughput estimate

The available back-end IOPS and bandwidth of a RAID group depends on the RAID groups:

- Number of drives in the RAID group
- IOPS or MB/s capability per drive model

A gross performance estimate can be made by multiplying the IOPS or bandwidth times the number of hard drives in the RAID group.

For example, assume a single 15k rpm Fibre Channel hard drive is capable of about 185 IOPS. For a four-disk RAID 5 (3+1), data is spread across all four drives. This RAID group has a potential throughput of 740 IOPS (185*4).

For a mirror-type RAID, during read both stripes (primary and mirror) are read at the same time. During writes, both the primary and mirror are written. This requires the IOPS be multiplied by two. However, the net effect is that only the IOPS of the primary stripe need be used in the calculation.

You should *not* do final performance planning on the basis of this type of calculation. It is not possible to accurately calculate the RAID group's performance without knowing the characteristics of its I/O.

Throughput calculation

I/O characteristics have a significant effect on RAID group performance. A more valuable and accurate estimate of a RAID group's bandwidth, throughput, and response time need to account for the:

- I/O type
- Access type
- I/O size

The I/O type (random or sequential); the ratio or reads to writes; and the size of the I/O all affect the performance (see the I/O characterization and workload section for details).

The detailed method for RAID group performance calculation using RAID group characteristics and I/O types is given in the *EMC CLARiiON Best Practices for Performance and Availability: Release XX Firmware Update — Applied Best Practices* white paper along with the information to perform these calculations.

LUN performance

The performance of LUNs depends on the performance of the underlying RAID groups.

Creating LUNs on the largest practical RAID group or using metaLUNs is the easiest way to ensure high performance LUNs for small random IOPS: more drives give more performance. Increasing the capacity of the LUN within its current RAID group does not improve its performance.

In a partitioned RAID group, other LUNs will share the underlying RAID group's IOPS and bandwidth. That is, if separate applications use a LUN created on the same RAID group, their combined needs of the applications may exceed the RAID group's performance capabilities.

Create LUNs with application performance requirements in mind. Try to limit the LUNs per RAID group to the smallest possible number. *Contention* is when the I/O of two or more LUNs bound to the same RAID group interferes with each others' performance. Avoid *linked contention* and *drive contention*. Linked contention is when I/O to different LUNs forces the RAID group's drive heads into large movements going from LUN to LUN. Drive contention is when more than one I/O stream needs to access the same drive at the same time.

We recommend limiting the number of LUNs per RAID group because as the number of LUNs increases, it becomes more difficult to predict or determine if the I/O is complementary or contentious without time-consuming analysis and debug.

It is important to remember that more drives result in better performance than fewer drives.

Short stroking

Short stroking is a provisioning technique for increasing the throughput of a LUN on a RAID group made up of mechanical hard drives.

Conventional hard drives incur additional time to service I/O each time the read/write heads are required to move for a seek. Short stroking reduces these head repositioning delays by limiting the number of tracks being used on the drive. The outermost tracks are used on a drive first. The outermost tracks have more sectors per unit length than inner tracks. By limiting the number of tracks, we reduce service time for each I/O. That directly affects the response time of individual disks, and the RAID group as a whole. However, limiting the number of accessible tracks to achieve this performance gain reduces the usable capacity of the RAID group.

Small-block random I/O workloads with many threads benefit the most from short stroking, large-block sequential benefit the least. A short-stroked RAID group is set up by creating a single LUN on the RAID group. This positions the LUN on the *outermost tracks (fast tracks)* of the RAID group's drives. The smaller the capacity of the short-stroking LUN in relation to the RAID group's overall capacity. then the greater the performance increase. Ideally, only a single LUN should be bound to a RAID group when short stroking to eliminate the chance of LUN contention.

Chapter 9 CLARiiON Performance

This chapter presents these topics:

Optimal CLARiiON performance is achieved by balancing the workload between the CLARiiON's two storage processors. The workload balancing relates to the number of LUNs and hosts allocated to each storage processor. Balancing the workload requires an understanding of the following related system

- Front end - Consists of the front-end I/O ports that attach the storage system to the hosts.
- Storage processors (SPs) - Contain the storage system's CPUs and memory. Memory includes the important storage-system cache.
- Back end - Contains the I/O buses that connect the storage processors to the devices making up the LUNs.

Percentage utilization

Percentage utilization is a measurement of how busy a resource is. It is used in determining a system's performance capability. High utilization is not necessarily bad. Logical devices such as LUNs do not suffer physical limitations. Physical devices, such as storage processors, front-end ports, disks and buses do. For them, getting to high utilization may be a requirement for maximum throughput or IOPS. Conversely, if the goal is low response time, seek low utilization – as with physical devices, high utilization is accompanied by longer queues and the resulting higher response times.

From a planning perspective, systems with very high usage may be at the limits of their performance. This leaves them without any performance margins, should they enter degraded mode. Systems with low utilizations have more overhead for bursts or to make-up for failures.

Utilization is an average over a period of time. At any point a device is either busy or idle. The measurement is calculated by taking the time the device is busy, divided by total elapsed time.

For example, if a hard drive is receiving and servicing I/O requests for 300 ms during a 1-second interval, then its utilization is 300 ms / 1000 ms = 0.30, or 30 percent utilization.

A device or system receiving enough requests to never be idle has a utilization of 100 percent. One hundred percent utilization of a physical resource is not desirable. Unnoticed oversubscription becomes possible at that utilization. Should a device already be at 100 percent utilization and it receives 50 percent more requests, its utilization will *still* be 100 percent. Although, depending on the workload, a logical storage object such as a LUN at 100 percent may still have some available performance depending on the physical drives supporting it.

CLARiiON resource utilization

There are three physical resources whose percent utilization is particularly important to monitor on the CLARiiON. All are highly workload dependent. They break down into CPU, memory, and I/O utilization for the storage system. The resources are:

- Storage processor
- Cache
- Storage

Storage processor utilization measures how busy the CLARiiON's SP CPU is. For example, for performance *and* availability, SP utilization should be about 50 percent.

Cache utilization is a memory utilization metric. Cache utilization directly affects the storage system's ability to perform I/O in a timely fashion.

Storage utilization includes back-end bus and storage-device utilization. Back-end bus utilization cannot be directly measured without analysis tools; it must be inferred from disk utilization.

Front end

Each SP has its own front-end ports. The front-end ports communicate directly with hosts when directly connected, or through the network in a SAN environment.

CX4 front-end ports

The CX4 series comes with both Fibre Channel and Ethernet UltraFlex™ I/O modules in its baseline configuration. Ethernet ports are typically referred to as *iSCSI ports* after the iSCSI protocol used over the (IP-based) SANs they attach to.

The number and type of ports required depend on the application's workload and host connectivity. In some cases the front-end port requirements may be dictated by the throughput or bandwidth of the dominate workload.

I/O modules are always installed in pairs – one module in SP A and one module in SP B. Both SPs must have the same type of I/O modules in the same slots. Additional I/O modules can occupy any available slots.

AX4 front-end ports

AX4 series CLARiiONs may come configured with either Fibre Channel or Ethernet (iSCSI) ports, depending on the model purchased.

Legacy CLARiiONs

Legacy CLARiiONs such as the CX3 and CX series may come configured with either Fibre Channel or Ethernet ports, or both, depending on the model.

Review your model's documentation for information about its front-end port configuration options. Note that earlier CLARiiONs may not support the highest-speed versions of their port type. This may reduce available bandwidth.

Port location

Depending on the model of CLARiiON, I/O ports may be located within a disk processor enclosure (DPE) or a storage processor enclosure (SPE). This information is also found in the product documentation.

For example, all model CX4 UltraFlex front-end ports are located on SPEs. On the AX4-5, front-end ports are located on the DPE.

Management and service ports

CX4 series CLARiiONs have two additional Ethernet ports per SP that are not part of the front end. AX4s have a single Ethernet port. These are the *management* and *service* ports. These ports are used exclusively for access and management of the SP through Navisphere. They do not support the iSCSI protocol, and cannot be used for workload data communications.

Fibre Channel ports

The Fibre Channel ports communicate either through a Fibre Channel SAN to host(s), or directly to a host's Fibre Channel HBA. These ports can also be directly connected to the Fibre Channel ports of other storage systems.

The Fibre Channel front-end ports on the CX4 series are a 4 Gb/s or 8 Gb/s optical-interface standard. A port's speed automatically adjusts to a lower bandwidth switch or DAS connections with a resulting loss of bandwidth. This automatic adjustment is called *auto-negotiation*. In auto-negotiation, network peers exchange their link-level protocol capabilities to provide compatible service. Auto-negotiation is performed by all the CLARiiON's front-end ports by default.

Note that the Fibre Channel standard continues to evolve. The modularity of the UltraFlex port allows for the retrofitting of Fibre Channel ports to higher Fibre Channel connectivity as the standard develops.

iSCSI ports

The iSCSI ports communicate either through an Ethernet SAN to host(s), or directly to a host's iSCSI HBA. These ports can also be directly connected to the Ethernet ports of other storage systems.

The iSCSI ports communicate through a gigabit Ethernet (GigE) or 10 GigE LAN to hosts. Hosts must have either a Ethernet network interface card (NIC) or iSCSI HBA.

The GigE iSCSI front-end ports on the CX4 series are GigE-copper interfaces. CAT6 is recommended. The 10 GigE iSCSI interface is optically cabled. Ports automatically adjust to lower bandwidth switches or direct connections with a resulting loss of bandwidth. Some restrictions apply as to how low an iSCSI port can autonegotiate downward.

Note that the Ethernet (iSCSI) standard continues to evolve, particularly with regard to bandwidth. The modularity of the UltraFlex port allows for the retrofitting of higher-bandwidth Ethernet connections as the standard develops.

Front-end port performance

For small I/O size, the IOPS performance of Fibre Channel and iSCSI ports is actually very similar through most of their performance envelope. However, Fibre Channel ports provide higher bandwidth than iSCSI ports.

The difference in IOPS performance between the two port types depends on the protocol and how efficient the ports are at dequeueing I/O. Fibre Channel ports are more efficient at this than iSCSI.

However, unless the workload is bandwidth intensive or the I/O is large (> 64 KB), the performance differences between the Fibre Channel and iSCSI ports are minimal. Examples of bandwidth intensive workloads include rich media, backup-to-disk, and DSS.

Fan-in

Fan-in is how many hosts use a single port. The maximum number of hosts that can connect to a CLARiiON storage system, called *initiators,* is model dependent. If the maximum initiators-per-SP exceeds 256, then no single port should have more than 256 initiators. However, knowing how many hosts can attach to a port is an important consideration.

Initiators-per-port estimate

The initiator limit is not enforced at the per-port level. It is possible, although unlikely, to oversubscribe a port. Typically, a single host's I/O requirements are modest compared to a port's capabilities. Yet many hosts can be fanned into a single port.

The quick estimate of hosts into a port is performed as follows:

Host Port IOPS = Port IOPS / Host IOPS

Host Port Bandwidth = Port Bandwidth / Host Bandwidth

For example, assume a type of production host requires 250 IOPS of throughput and 10 MB/s of bandwidth. A single 4 Gb/s Fibre Channel port can handle 360 MB/s of bandwidth, and 50k IOPS. A single iSCSI port can handle 80 MB/s of bandwidth, and 10k IOPS.

How many of the described hosts can be fanned in to a single Fibre Channel port? How many servers can be fanned into an iSCSI port?

- Server IOPS: 250
- Server Bandwidth (MB/s): 10
- 1 GigE Port IOPS: 10,000
- 1 GigE Port Bandwidth (MB/s): 80
- 4 Gb/s Fibre Channel IOPS: 50,000
- 4 Gb/s Fibre Channel Bandwidth (MB/s): 360

Fibre Channel Hosts per Port IOPS = 200 = 50,000 IOPS / 250 IOPS

Fibre Hosts per Port Bandwidth = 36 = 360 MB/s / 10 MB/s

iSCSI Hosts per Port IOPS = 40 = 10,000 IOPS / 250 IOPS

iSCSI Hosts per Port Bandwidth = 8 = 80 MB/s / 10 MB/s

In both cases, bandwidth is the limiting factor. In this example, 36 hosts can be attached to a Fibre Channel port, or eight hosts can be attached to an iSCSI port.

Front-end port queues

Front-end port queues hold I/O that is being handled or awaiting dispatch to the storage processor.

The number of outstanding I/Os per port has a direct effect on its performance. I/O to a front-end port originating from several host HBAs can be heavy and be very bursty. To accommodate the I/O, the front-end ports are a queued device.

Each arriving I/O takes up a queue position. I/Os are taken off the queue and processed by the storage processor. The number of queue entries ready for use is called the *port queue depth*. Fibre Channel and iSCSI ports have the same queue depth.

If I/Os arrive at a port faster than they can be handled by the storage processor, the port queue fills up. If the queue depth is exceeded, the storage system returns *a queue full* (QFULL) status to the host in response to an I/O. The exact effect of a QFULL on a host depends on its O/S. However, the queue-full condition always adversely affects throughput.

To avoid the QFULL condition, it is prudent to distribute the host's LUN access across all the ready for use front-end ports as evenly as possible based on their throughput needs.

Another method for avoiding a QFULL condition is to limit the host-side I/O queue length. Most host operating systems or their I/O drivers have queue limiting capabilities. Please refer to your OS or driver documentation to limit your host-side I/O queues. Best practice guides provide recommendations for optimal queue limits.

Storage processors (SPs)

CLARiiON's SPs are special purpose processors optimized for I/O throughput. They have CPU, memory, and I/O resources to manage.

Every CLARiiON has two SPs; each SP has one or more CPUs. Higher-model CLARiiONs have more than one CPU. The SPs each have their own memory and I/O resources. In addition, they share memory resources with their peer SP. The most important use of memory on the CLARiiON is cache. Cache is memory used by the storage system to reduce the response time of read and write requests from hosts.

SP CPU

The CLARiiON CX4's CPU resources depend on the model. Every CLARiiON has two SPs. Each SP has its own CPU. The CLARiiON CX4 series CPUs are all multiprocessor-based because SPs have *multi-core* CPUs. A *core* is a processor unit on a CPU chip. Entry-level and midrange level have fewer cores. Enterprise-level CX4s have a larger number of cores. The CPUs on different models have different clock speeds; entry-level, and midrange models having lower-speeds than the Enterprise-level model. The cores on a CPU all run at the same speed.

CPU performance

An important SP performance metric is SP Utilization. This is a measure of the SP's CPU utilization.

Ideally, in a production environment with strict requirements for high availability and consistent performance, the SP CPU utilization should be around 50 percent. This way, if an failure triggers an SP failover, the peer SP can easily accommodate its own processing load and the failed-over load.

In Navisphere Analyzer this information is found in the SP Utilization screen. Note the SP Utilization is the average of all CPU or core utilizations. This does not reflect individual utilizations in multi-core SPs.

Memory

Each SP has its own memory. All CLARiiON SP memory performs error detecting and correcting to ensure the accuracy of its contents. The SPs of the different CLARiiON models have different amounts of installed memory. Entry-level CLARiiONs have less memory than higher-end models.

System memory is divided into SP memory and cache memory. Write cache memory is mirrored to the peer. The figure below shows a conceptual view of CLARiiON memory.

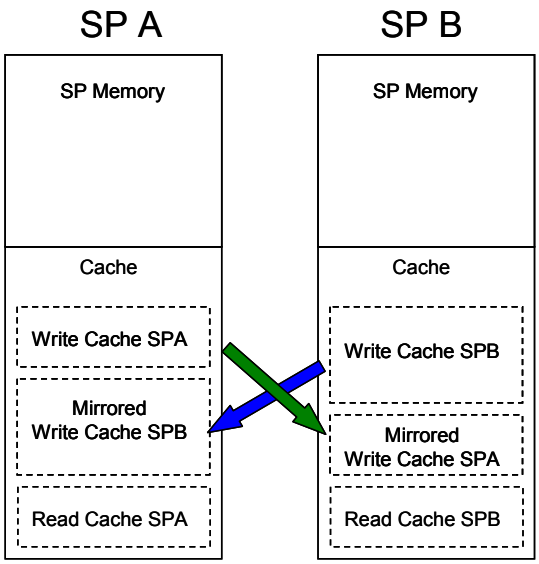

Figure 15 CLARiiON memory: conceptual view

SP memory

A certain portion of CLARiiON's memory is devoted to FLARE and Navisphere. Some is reserved for applications such as SnapView, MirrorView, and SAN Copy.

The amount of memory dedicated to SP memory and its applications depends on the model, and is equal to the physical amount of memory on the SP minus the maximum cache size per SP.

SP cache

CLARiiON's cache is a contiguous portion of the SP's memory that is separate from the SP memory used by FLARE.

The amount of cache memory available on a CLARiiON is model dependent. Entry-level models have less cache memory than higher models.

Cache itself is logically divided into two regions: *read cache* and *write cache*. Read cache is for data that is held in memory in anticipation of it being requested in a future read I/O. Write cache stores write request data waiting to be written to a drive.

The cache can be adjusted to adjust to some specific workloads or to spikes in demand.

In addition to an SP's own read and write cache, cache memory contains a mirror copy of its peer SP's write cache. This is an important availability feature. Note the amount of memory useable as write cache at any time on an SP depends on the write cache usage of its peer SP.

Cache can be re-allocated between reads and writes, though the cache must be disabled in order to do so. The write cache partitioning applies to both SPs. Read cache can be partitioned independently for each SP. The great majority of all cache memory is allocated to write cache, with the remainder read cache. This is because it is less likely for a read to find its I/O already in cache outside of workloads performing sequential read I/O. It is also important, *not* to run out of write cache.

Cache pages

All cache is organized into pages. Incoming I/O is placed in the page. A page is the smallest unit of allocation. If the I/O is smaller than the page, more than one I/O may be made to share the same page when I/O has contiguous LBA addresses. This is an efficient use of the cache pages. As the I/O is executed (sent to the drives and acknowledged), its page is reused for the next I/O. A write cache page whose I/O has not yet been written to storage is called a *dirty page*.

Read cache

When a host read request is received, the CLARiiON checks its read and write caches first. If the data not in either cache, it must be read from storage. After the data is read, it is held in the read cache for subsequent use.

Read requests that find their data in cache (a *cache hit*) consume the least amount of storage system resources; have the highest throughput, and the lowest response time. However, reads that are not found in cache (a *cache miss*) have a higher response time because the data has to be retrieved from a drive.

Read cache optimizations

With sequential read requests, the chances of a cache hit can be greatly increased through a technique called *prefetch* (this is also referred to as *read-ahead*). Here, when sequential access is detected by the SP, the read cache is filled ahead of actual host requests in anticipation of their being requested later.

Because of read-ahead, workloads characterized by sequential I/O can have a high percentage of cache hits. This greatly improves their response times. Workloads characterized by random I/O do not benefit from this technique.

Read cache hit ratio

The read cache hit rate (RCHR) is the percentage of read I/O requests found in the read cache. It is one of the metrics for determining if a read I/O workload is random or sequential.

If a workload's read I/Os are purely random, the RCHR can be up to 5 percent. If it is higher, there is probably either a mix of random and sequential I/O, or a multi-threaded sequential workload being measured. There is also the chance the same addresses are being repeatedly accessed, or there is high locality in the I/O. The metadata on the first 64 KB of a file system volume is an example of addresses that would receive repeated access.

For purely sequential I/O, the metric's value depends on the number of threads performing the I/O. With single-threaded sequential, the RCHR should be about 100 percent, but not less than 80 percent. This is an example of the prefetch working at peak efficiency.

Multiple-stream multi-threaded sequential access can reduce the effectiveness of the prefetch, which lowers the RCHR. Multiple-stream multi-threaded sequential I/O causes many read accesses to many different regions of a RAID group's drives. This slows the prefetch process lowering the RCHR. Modifying the prefetch parameters may increase an RCHR in the lower end of the range. However, an in-depth read cache usage analysis is needed to determine the cache parameter settings for a higher RCHR. The final possible reason for a low RCHR value is the workload is not truly sequential. The larger the random I/O component to the workload, then the lower the value of RCHR.

Write cache

The CLARiiON write cache is a mirrored write-back cache. This means for every write, the data is stored in cache, copied to the peer SP, and then the request is acknowledged to the host. In this process, write cache is duplicated *(mirrored*) between the storage system's SPs to ensure data protection through redundancy. In addition, requests are acknowledged before they are written to disk. Acknowledging the request from write cache decouples the response time of writes from disk speed and RAID effects, which improves performance.

The availability feature of mirroring write cache has a slight effect on performance. The mirroring delays the host I/O request's acknowledgment until it completes. The memory usable by an SP for write cache is also reduced by having to maintain a copy of its peer's cache. In bandwidth-intensive workloads, the mirroring process may limit overall storage system write bandwidth.

Flushing overview

Flushing is writing a write-cache page out to storage. The caching algorithms flush the pages in order to keep adequate margins of cache to handle bursty events. Ideally, the number and frequency of write I/Os from hosts are in balance with the storage system's ability to flush data to storage. This type of balance is a result of good performance planning.

Write cache optimizations

Caching of write I/Os is a more efficient use of memory than caching of read I/Os.

Sequential writes when detected are optimized. Several smaller I/O writes can be *coalesced* into fewer larger writes to storage. Ideally, they can be coalesced into one or more full-stripe writes. In addition, writes to the same location when found in cache are superseded with the last write. This is a very efficient use of the storage system's back-end resources.

Backfill is a write cache optimization where data will be read from storage so that a more efficient write I/O size can be achieved. For example, in a RAID 5 (4+1) RAID group-based LUN, if a 192 KB write is pending, the adjacent 64 KB from the destination LUN may be read to achieve a more ideal 256 KB full stripe write.

The CLARiiON storage system also has the option to write data directly to disk. The technique is called *write-aside*. It is sometimes referred to as *write-through*. Write-aside allows large I/Os to bypass the cache when cache is enabled. This prevents large-sized write I/Os from consuming too many cache pages. Losing these pages would adversely affect the workload's write-cache bandwidth.

Write cache management

Properly sizing cache to support its provisioned number of drives and ports is an important tuning task. In this tuning, the amount of cache is not as important as the storage system's ability to orderly flush write cache pages, and to perform prefetches to read cache.

Flushing details

There are three types of flushing:

- Idle flush
- High water flush
- Forced flush

Idle flushing is the execution of the pending I/Os to idle LUNs. A LUN is idle when it has no I/O for two seconds or longer. Idle flushing is a regularly occurring background process. As LUNs are found to go Idle their pages are flushed from cache to make room for the I/O of active LUNs. If no LUNs are idle or a LUN has previously been flushed down to no dirty pages, there is no idle flushing.

High water flushing activates when the percentage of dirty pages reaches a pre-set limit. This limit is called the *high water mark*. The high water mark is typically set to maintain a reserve capacity in the cache, which is intended to handle bursts of writes. When high water flushing is triggered, the SP increases the cache flushing rate to increase the flush rate over that of idle flushing. SP performance is minimally affected by high water flushing.

The write cache attempts to always have memory ready for incoming writes. Forced flushing is avoided, but will occasionally occur even in a well-designed system. When a write-I/O request is received and cache is already full, a forced flush is triggered to write the pages of the destination LUN receiving the current request, and clear pages for later requests. The new write request receives space on a newly cleared page. Forced flushing can be triggered many times in succession until the cache can provide enough pages for incoming write requests. Forced flushing when a single LUN to run out of its available pages occurs occasionally. On a well-tuned storage system, forced flushing for all LUNs should occur very rarely.

A force flush means the host is waiting for physical disk operations and full RAID operations, so response time goes up. During any heavy flushing storage system performance is adversely affected as the writing of the cache's contents to disk will contend with other I/O requests received.

Cache enable/disable

There are certain conditions where caching may not be desirable or needed. Under these conditions, cache can be disabled.

Cache is enabled or disabled at two levels: storage processor and LUN. The storage processor is the higher level. If cache is disabled at the storage processor, there is no caching at the LUN level. If cache is enabled at the storage processor level, individual LUNs may have their caches enabled or disabled.

How storage processor caching is performed is CLARiiON series dependent. Caching is performed differently on the legacy CX3 and CX series than on the CX4 series. The CX4 series

has the most robust caching. Storage processor caching on the CX4 series is automatically disabled only if the following conditions are not met:

- At least one SPS must be present, and it must be fully charged (write cache disabled only).
- At least one power supply must be present and functional in the DPE/SPE. (Some models can take three PS faults). In SPE models, both DAE 0 power supplies must be functional.

CLARiiON caching is controlled through Navisphere.

LUN caching

Caching on the CLARiiON is typically configured at the LUN level. Write cache pages are dynamically allocated from the available pages, based on the volume of I/O a LUN is experiencing. Busy LUNs will be allocated the most pages; LUNs with no activity will have no pages allocated. The caching algorithms moderate how many pages a LUN receives, when there are many busy LUNs.

LUNs can have any combination of read and write cache enabled or disabled. By default, a LUN's caching is enabled. EFD-based LUNs are different. They have cache disabled as a default.

LUN caching is typically manually disabled to preserve cache pages when there are conditions that might tax the storage processor's cache. For example when running multiple applications that generate large I/Os. Another example is to disable cache before running multiple full-SAN Copy sessions. (SAN Copy is a bulk data replication layered application; see the Replication layered applications section.)

Cache performance

The CLARiiON's cache is highly configurable. For most workloads, the default settings perform very well. However, for atypical workloads, or to achieve high levels of performance or availability, cache can be tuned through the adjustment of a just a few parameters:

- Cache page size (Global)
- Prefetch (Local)
- Watermarks (Global)
- Write-aside size (Local)

Note that some parameters are local and are apply to single LUNs and global apply to all LUNs.

Cache page size

Cache page size applies to both read and write cache. Cache pages have a fixed size in KB. They may be 2, 4, 8, or 16 KB. Both read and write cache use the same page sizes. An 8 KB page size is CLARiiON's default.

Ideally, the cache page size is the same as the most common I/O write request size received by the storage system. For systems performing mostly large I/O or sequential I/O, a larger cache page size improves overall system performance. However, where I/O request sizes are mixed, CLARiiON's default size is the best.

Prefetch

Prefetch applies only to Read Cache. CLARiiON offers constant and variable prefetch for read I/O requests.

Constant prefetch is useful for tightly controlling the amount of data prefetched to avoid filling read cache. With constant prefetch, the same number of KBs is prefetched ahead of read I/O requests when sequential access is detected. The caching algorithms flush the pages and maintain the efficiency of the storage system, along with keeping adequate margins of cache to guard against bursty behavior. Constant prefetch is most efficient when I/O sizes are uniform.

Variable prefetch works using a combination of the: Read I/O Request size, a Segment Multiplier, and a Prefetch Multiplier. Variable prefetch is CLARiiON's default.

- The Read I/O Request Size: is the size in KB of the read request from the host.
- Prefetch Multiplier: used to calculate the total size of the prefetch.
- Segment Multiplier: determines the size of I/Os to the back end.

If the Read Request size is smaller than a cache page, the Cache Page Size is used. The prefetch multiplier is used to calculate (Prefetch Multiplier * Read Request size) the total amount to prefetch in this request; the segment multiplier (Segment Multiplier * Read Request size) is then used to calculate the back-end IO size that will fulfill the total request

If the Segment Multiplier and the Prefetch Multiplier are equal, a single request to the back-end results. If the Segment Multiplier is smaller than the Prefetch Multiplier, multiple requests to storage result. The Segment Multiplier is never larger than the Prefetch Multiplier.

For example, assume:

- Cache Page Size: 4 KB
- Read I/O Request Size: 8 KB
- Prefetch Multiplier: 4
- Segment Multiplier: 2

The Read I/O Request Size is larger than the Cache Page Size. The Request Size is used. The total amount Prefetched will be 32 KB (4 * 8 KB). This 32 KB will be read from storage in two requests of 16 KB (2 * 8 KB) each.

For small I/O sizes, setting Segment Multiplier and Prefetch Multiplier to the same large number improves performance because the back end can work more efficiently. This setting is particularly efficient during sequential reads when the read cache is large enough to hold the result. Setting them the same for large I/Os can cause prefetching to consume valuable back-end bandwidth. In production environments, setting the segment multiplier to one-half or one-quarter of the Prefetch Multiplier breaks up the prefetch sequence allowing other I/Os on the storage system to execute. A smaller Segment Multiplier results in lower read throughput.

LUN prefetch settings

There are limiting parameters that can be applied to prefetch to prevent excessive storage requests that may adversely affect overall performance. These settings are made on a per-LUN basis. They are:

- Maximum prefetch (blocks)
- Prefetch disable size (blocks)
- Prefetch idle count.

Maximum Prefetch setting is the limit—in blocks—for data requested. This is the maximum number of blocks to bring into read cache ahead of actual requests. For example, setting the Maximum Prefetch to 4096 restricts prefetch to 2 MB (512 bytes * 4096).

The Prefetch Disable Size prevents I/Os of the Disable Size or larger from triggering a prefetch. This causes I/Os of this size and larger to be read directly from the disk uncached.

Prefetch Idle Count limits the execution of prefetch for a very busy LUN. The Prefetch Idle Count is also used in cache flushing; changing this parameter should be done only at the prompting of EMC service personnel.

Watermarks

Watermarks are used to manage write cache flushing. CLARiiON has two watermarks: high and low. These parameters work together to manage the flushing conditions. Watermarks only apply when write cache is activated.

Cache capacity above the high water mark represents a reserve of pages to handle bursts of write I/Os. Watermark flushing starts when the high watermark is exceeded and continues executing until the low watermark is reached. Watermark flushing stops until the high watermark is reached again.

The margin of cache above the high watermark is set to contain bursts of write I/O and to prevent forced flushing. The low watermark sets a minimum amount of write data to maintain in cache. It is also the point at which the storage system stops high water or forced flushing. The amount of cache below the low watermark is the smallest number of cache pages needed to ensure a high number of cache hits under normal conditions. This level can drop when a system is not busy due to idle flushing.

The default Low Watermark is 60 percent. The Default High Watermark is 80 percent.

The difference between the high watermark and the low watermark determines the rate and duration of flushing activity. The larger the difference is, the less often you will see watermark flushing. The flushing activity is highest when the high watermark is passed. Intense flushing increases LBA sorting, coalescing, and concurrency in storage; but it may have an adverse effect on read I/Os. The smaller the difference between the watermarks, the more constant the flushing activity. A lower intensity of flushing permits other I/Os (particularly reads) to execute.

In a bursty workload environment, lowering both watermarks will increase the cache's "reserve" of free pages, allowing the system to absorb bursts of write requests without forced flushing.

The following figure shows how high and low watermarks work together to manage the write cache:

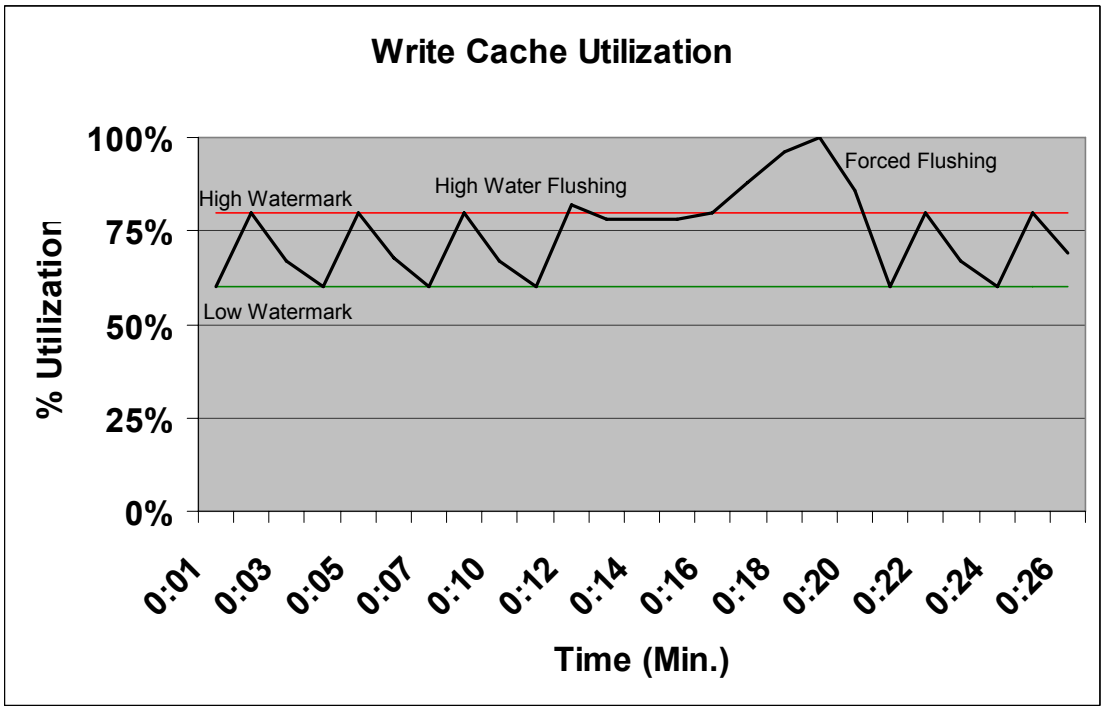

Figure 16 Write Cache Utilization

In Figure 16 the black line is cache utilization, or the percentage of dirty pages. The workload's activity is shown to fill cache to the high watermark (red line). High-water flushing then flushes the cache down to the low watermark (green line), and cache is allowed to fill again. No flushing occurs until the high watermark is again exceeded.

Should the load increase (in the figure, this is the 12-minute mark), steady cache usage can occur as the workload's usage of cache pages and the flushing rate come to equilibrium. If there is a burst of I/O that fills cache and exceeds high water flushing's ability to write pages out to storage--forced flushing occurs. (This is shown at the 17-minute mark.) The combination of forced flushing and high-water flushing make new pages ready for use. At the end of the burst, forced flushing and high water flushing will drive-down cache usage to the low watermark.

The goal of setting watermarks is to avoid forced flushes while maximizing write cache hits. Using the watermarks, cache can be tuned to increase response time while maintaining a reserve of cache pages to match the workload's bursts, and system maintenance.

Write aside size

The Write-Aside size parameter sets the largest I/O that will be cached. Note that I/O needs to *exceed* this parameter to bypass cache. Write aside is set on a per-LUN basis.

Large writes primarily affect write cache bandwidth due to write cache mirroring. The size configured for write-aside should be set large enough to avoid this bottleneck.

The Write-Aside size parameter is numerated in blocks. For example, 4096 blocks results in a write aside of 2 MB (512 bytes * 4096).

The default write aside is 1 MB. That is, 1 MB write requests are cached, and write requests greater than 1 MB are *written through* (written directly to storage).

Setting the write aside higher than the I/O sent by a high bandwidth write application may cause cache to fill quickly, causing forced flushes. Setting the write aside lower will increasing response time for I/Os larger than the Write Aside size. If the target LUN is a parity RAID type, I/O bypassing the cache should be large enough to fill the parity stripe so that parity operations do not result.

Back end

CLARiiON's Fibre Channel back-end buses connect the SPs to storage. The buses and storage are referred to as the *back end.* The bandwidth and throughput of a back-end bus needs to be managed to avoid bottlenecks. Bottlenecks may occur on either the bus itself or at the drives.

Number of back-end buses

CLARiiONs have one or more back-end buses. The number of back-end buses depends on the model. Lower-level models have fewer buses than higher-level models. For example, the entry-level CX4-120 has one back-end bus. The CX4-960 can be configured with up to eight. Refer to your model's product documentation for the number of back-end buses on your storage system.

Back-end buses

Each CLARiiON back-end bus has *redundant* (two) connections between a CLARiiON's SPs and each of its drives.

Each CLARiiON CX4 back-end bus is capable of a 4 Gb/s wire rate. If a RAID group's LUN(s) are accessed by a single SP, the practical usable bandwidth is about 360 MB/s. (This is called the *per-loop* access rate.) Drives on a bus are addressed by LUNs owned by both SPs, have a usable bandwidth of about 720 MB/s per bus (see the RAID section).

The Fibre Channel back-end bus will auto-negotiate to the speed of the slowest device on the bus at the storage system startup. For example, if legacy Fibre Channel hard drives with 2 Gb/s attachments are provisioned in DAEs on a CX4, the bus will reduce speed to 2 Gb/s. Note this will cause all 4 Gb/s drives on that bus to operate at 2 Gb/s.

The utilization of its bandwidth has an effect on bus performance. As the back-end bus approaches its full transfer capacity it is said to *saturate*. The addition of more load will not scale linearly. The result is increased response times for drive access to the bus.

Shared RAID groups

Drives are a shared resource between storage processors. A RAID group may contain one or more LUNs. The LUNs hosted by the RAID group may be owned by either storage processor. Both storage processors can own separate LUNs on a RAID group at the same time.

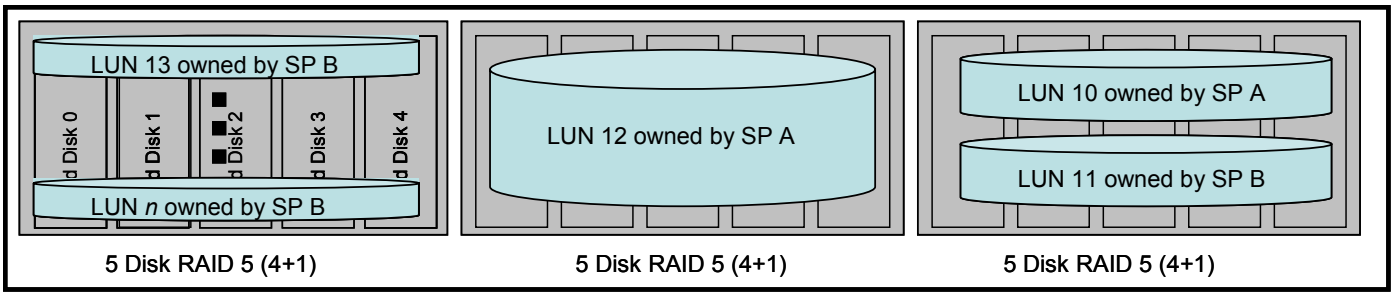

Figure 17 Shared RAID group drives

The simplest organization is to have sole-usage of a RAID group by a single SP. (Shown in the center of Figure 17 Shared RAID group drives) That is, a RAID group's LUNs are owned by a single SP. Sharing RAID groups between storage processors allows for longer queues at the drive. A single storage processor is limited to the number of requests it can send to the drives. Dual-SP access to the drives is one way to get concurrency and achieve the maximum throughput rates from the drives. The tradeoff of longer queues at the disk is longer response times.

Back-end bus performance measurement

The most important performance metric for the back-end bus is the bandwidth from the drives. (Bandwidth is more likely to be exceeded than the limits for IOPS.) The bandwidth being received from the drives implies the bandwidth usage of the back-end bus.

Using Navisphere Analyzer, this information is found in the RAID group **Total Bandwidth** metric**.** A bottle neck of the back end can be diagnosed from the bandwidth being received from the hard drives. Note that it is possible for a RAID group to span two or more buses.

In general the I/O needs of the entire back end should be spread across all of its buses. Navisphere by default distributes RAID groups across back-end buses as evenly as possible. In round-robin order, each RAID group is provisioned on a separate bus, with all member drives on the same bus.

Drives and DAEs

An UltraPoint™ DAE is the hardware enclosure providing the physical protection, organization, and connectivity to CLARiiON's drives. A RAID group's drives may be located in the same Disk Array Enclosure (*DAE*) or in different DAEs. Up to 15 drives may be installed in a DAE (12 for AX4) (see the Disk array enclosure section).

UltraPoint DAEs used in the CLARiiON CX4 series (DAE4P) support 4 Gb/s Fibre Channel back-end bus operation. Legacy 2 Gb/sec UltraPoint enclosures (DAE2P) from previous generation CLARiiONs installed on any CX4 series bus will limit that bus to running at the lower 2 Gb/sec.

Disks can be chosen from any DAE on the storage system to be included in a RAID group. All the disks in the RAID group must be the same type of disk (SATA, SAS, Fibre Channel, or EFD).

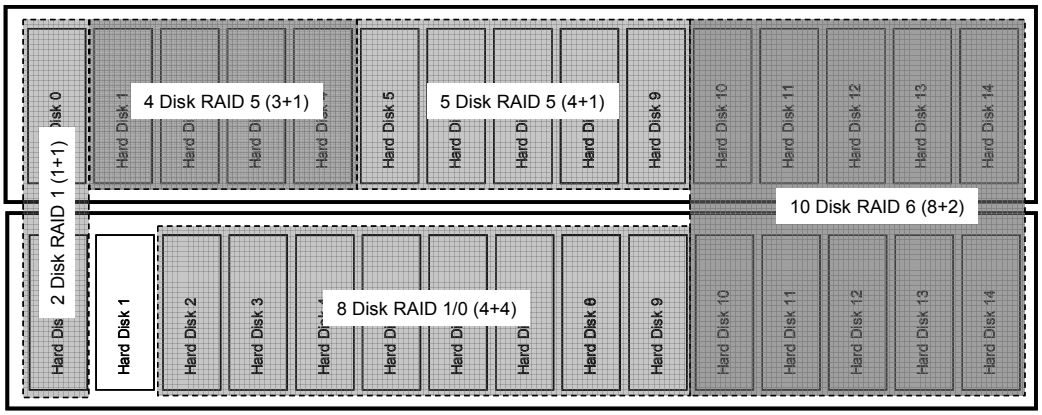

Figure 18 Conceptual RAID group binding

Figure 18 shows a conceptual view of several RAID groups created on two DAEs. Note that RAID groups can span DAEs.

Drives and buses

DAEs and back-end buses

The standard racking of DAEs alternates the buses across adjacent DAEs. On CLARiiONs with more than one back-end bus, drives located in adjacent DAEs will be on different buses. Be sure to verify the number of buses on the CLARiiON, and that standard racking is in use when provisioning RAID groups.

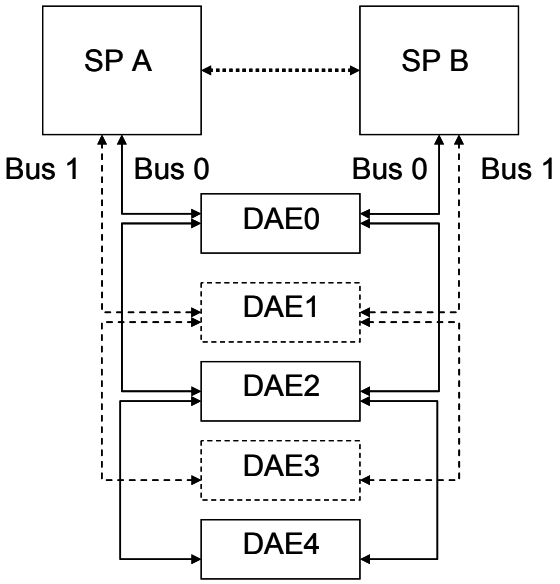

Figure 19 Conceptual view: Back-end bus connection to DAEs

Note that back-end buses have the same identifier despite originating from different SPs. For example, back-end bus 0 originates in both SP A and SP B, and has a connection to DAE0.

Standard racking is not mandatory. DAEs on a back-end bus are not required to be in sequential order. In addition, the bus ordering of the DAEs can have "holes." An empty DAE without drives can be installed in the CLARiiON without causing disruption.

Link controller card (LCC)

An LCC is the interface between the back-end buses and the installed drives in the UltraPoint DAE. An LCC provides Fibre Channel data connectivity, control, and monitoring within a DAE.

There are two LCCs (A and B) per DAE. Each storage processor (A or B) attaches to a DAE through an LCC by a back-end Fibre Channel bus.

Within the DAE, the LCC maintains Fibre Channel loop characteristics while providing a point-to-point connection within the DAE to the drives.

Each drive module in a DAE has two independent Fibre Channel ports. These ports are designed to connect to separate loops. (A loop is an LCC to port connection.) In a DAE each of these ports connects to a separate LCC. Each SP connects through an LCC on its back-end bus to one of the dual ports of a DAE's drive module (sometimes called *drive slots*). Note there is no connection between storage processors or LCCs through the drive installed in a module. The pathway between a storage processor to a drive module is independent of each other.

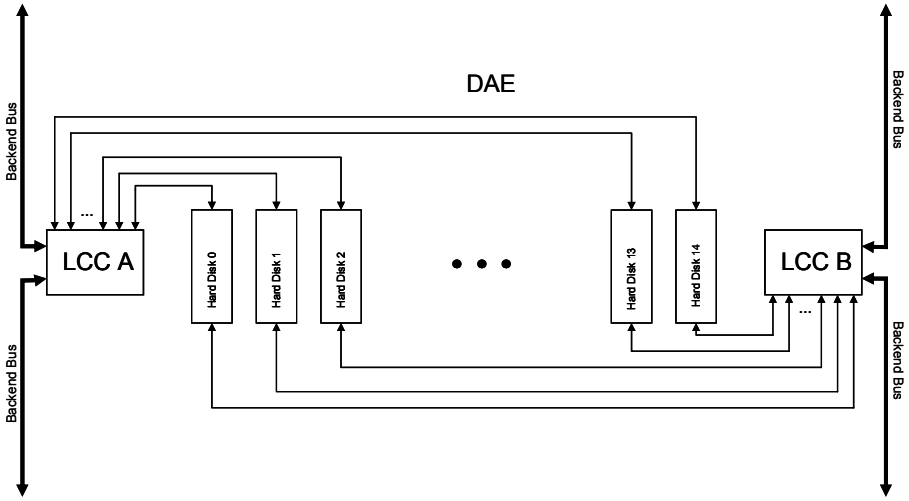

Figure 20 LCC connectivity conceptual diagram

The LCC also monitors the loop for signal connection and protocol errors. Monitoring can detect when bad data is being generated on the loop. Should an error be detected the LCC signals the storage processor. The storage processor can then bypass the erring loop segment allowing the other drives on the loop to continue operation.

An LCC provides control as well as connectivity to each drive installed in the DAE. This includes: power to the drive, maintenance of the DAE's slot configuration status, and the ability to isolate the drive from the storage system. Should a disk be inserted in a DAE module slot, the LCC signals the SP with the installed drive's configuration and status. The LCC can isolate from the storage system: slots without installed drives, powering-up drives, resetting drives, and failed drives.

LCC and availability

Should a DAE's LCC fail, the connected SP loses control and communication with the drives in that DAE and any located behind it in the back-end loop. A failed LCC is a hardware-replaceable item within the DAE. Drive access is provided through failover to the peer storage processor that can access the drives through its surviving LCC.

How the CLARiiON recovers from an LCC failure is affected by the:

- CLARiiON's FLARE firmware version
- RAID group provisioning (multi-bus or single-bus)
- The RAID group's RAID level

Details on LCC failure recovery can be found in *Best Practices*.

Dual-ported drives

All CLARiiON drives are dual ported. They have two independent Fibre Channel ports per drive designed to connect to separate back-end loops. Fibre Channel drives are directly connected. Drives with SATA attachments are adapted for dual port Fibre Channel operation. Each drive port connects to a separate LCC, which in turn is connected through a separate back-end bus to an SP.

Drive and back-end bus bandwidth

The back-end bus has more available bandwidth than a mechanical hard drive's maximum external transfer speed. For example, a 4 Gb/s Fibre Channel back-end bus has about 360 MB/s of bandwidth. The typical mechanical hard drive has a sustained internal transfer rate of about 50 MB/s. (The burst rate is much higher.) The differences in the available bandwidth between the bus and the drive are mediated through the SP's queuing mechanism for Fibre Channel I/O commands and data sent to the drive.

Each drive has two I/O queues, one per storage processor, which is one per drive port. The queues are maintained in software by the SP. SPs send commands and data to a drive and the command is added to the queue. Drives execute the command, and acknowledge it. When the command is acknowledged, it is removed from the queue. Only a limited number of commands are sent per drive per SP. This number is set by the number of possible queue entries. The number of possible queue entries is limited to ensure optimal performance, even though the drive may be capable of handling additional commands, albeit with a response time penalty. Commands are sent until the SP's drive queue fills. Usually, a drive executes its I/O commands without a backlog of commands occurring in its queues. If a drive queue is full, commands are withheld by the storage processor until an enqueued command completes. Serialization rules are enforced by the storage processor issuing the commands to prevent ordering problems with the drive's internal I/O execution.

Performance effect of buses

Horizontal provisioning is the practice of placing all of a RAID group's drives on a *single* bus. Most commonly, this is achieved by creating a RAID group fully contained within a single DAE. Although spanning DAEs on the same bus has the same result. For example, by creating a RAID group with drives in both DAEs 0 and 2 when standard racking is enforced.

The default allocation method in Navisphere is horizontal provisioning. Navisphere by default evenly distributes RAID groups across back-end buses. In round-robin order Navisphere provisions each RAID group to be on a separate bus, with all member drives on the same bus.

Vertical provisioning is a *multi-bus* RAID group provisioning strategy. A RAID group's drives are positioned on two or more buses. This always puts the group's drives within more than one DAE, with each of these DAEs on separate back-end buses. An example is creating a RAID

group with drives in both DAEs 0 and 1 when standard racking is enforced on a multi-bus CLARiiON.

There can be a modest performance advantage with multi-bus provisioning. In particular, large RAID groups, and RAID 1/0 groups can benefit from being distributed over two or more back-end buses.

Multi-bus provisioning is advisable when:

- Bandwidth requirements result in the need for careful load distribution.
- Bursty host I/O is diagnosed to cause back-end bus bottlenecks.
- Write cache destaging of a RAID group's LUNs is diagnosed to interfere with host reads due to bus saturation.
- Large-block I/O requests from a particular RAID group are diagnosed to interfere with I/O response time.

Be aware, that multi-bus provisioning requires more maintenance discipline and planning to set up and maintain than does single-bus provisioning. In addition, not all CLARiiON models have multiple back-end buses

Chapter 10 Availability

This chapter presents these topics:

Availability in the CLARiiON is achieved through a combination of high-reliability components, and redundancy of components and data.

Reliability

Reliability is the ability of the storage system to run for long periods, and under operational stress, without suffering hardware failure or software faults. CLARiiON's architecture contains many features designed to ensure a very high level of reliability. High-quality components, assembly, design verification, and quality testing back up this design. The CX4 series storage systems' systemic reliability is rated at *five 9s* (99.999 percent) uptime based on hundreds of thousands of hours of field experience. The exact reliability measurements depend on the model and components used.

CLARiiON is designed and built to protect data, with full data path protection, throughout the storage system. This data path protection prevents against unreported *data miscompare* errors. The UltraPoint hard drive array enclosures (DAEs: DAE2P, DAE3P, or DAE4P) include a point-to-point design within the drive enclosures. The FLARE operating environment monitors low-level diagnostics at the drive level that include fault detection, isolation, and error correction. RAID protection is supplemented with global and proactive hot sparing; sector-based checksum of all user data; and automatic background verification of the disk media.

Redundancy

Redundancy is the ability of the system to continue operating during failure, without the host losing access to data, while maintaining data integrity. Redundancy is achieved through the duplication of critical components of the storage system and through provisioning choices such as RAID type and front-end connectivity.

In systems built with redundant parts and components, failure events are statistically independent. The occurrence of one failure makes it neither more nor less probable that another failure shall occur. For the user to lose access to data, two or more interdependent components need to fail at the same time. The probability of two statistically independent failures occurring is calculated by multiplying the probability of one failure occurring by the probability of the other failure occurring. In other words, the likelihood of two failures occurring at the same time is very small. This dramatically increases the overall up time.

Active/passive architecture

CLARiiON implements an *active/passive* architecture on a per-LUN basis from the host's. Active/passive means there is a single operational interface active from a storage processor to a LUN. Note this is a logical interface. Both storage processors are typically active for their owned LUNs. Peer storage processors can physically share RAID groups. In a failover (trespass), the passive storage processor assumes temporary ownership of its peer's LUNs. This is often hidden from the host by an internal redirection within the CLARiiON.

However, multiple connections between hosts or networks and CLARiiON front-end ports are encouraged to achieve availability in the case of s path or storage processor failure

Multipathing

Multipathing allows for the *failover* from a failed *I/O path* to an alternate path when more than one path to a LUN on a storage system exists. An I/O path starts at the host bus adapters (HBA) on the host, and includes all the components and connections in-between all the way to a storage system. Hardware components on the I/O path include: cables, switches, front-end ports, storage processors, back-end buses and drives. There are also layers of software on this path, including: hardware drivers, device firmware, and applications running on the host. Any component on the path can affect the availability of storage presented to the host.

The original I/O path is typically the optimal performing path. The failover-availability feature may have an adverse effect on performance. For example, the original path's performance may benefit from special tuning or optimal physical routing through the network or the storage system.

Host-based application support

Path failover requires a host-based application.

EMC's PowerPath® is a host-based application adding multipathing as well as several other performance and availability features. For example, a performance-benefiting feature of PowerPath is that, with multiple active paths to a LUN, I/O load can be balanced across the storage processor's front-end ports and the host's HBAs to increase bandwidth and avoid bottlenecks. An availability feature of PowerPath is auto-failback. *Failback* restores the storage system to its original I/O pathway configuration once it detects a failed I/O path is fixed.

In addition, there are multipathing applications native to the O/S, such as Microsoft Multi-Path I/O (MPIO®), and HP-UX Physical Volume Links (PV Links®) providing similar functionality. The latest versions of MPIO and PV Links support failback.

ALUA

CX4 series CLARiiON storage systems support the industry-standardized Asymmetric Logical Unit Access (*ALUA*) protocol. It provides path management by permitting I/O to stream to either or both storage processors to reach a LUN. ALUA can reduce the effect of some front- and back-end failures to the host. Internally, if a component failure makes a LUN unreachable, ALUA shields the attached hosts from failing over and trespassing LUNs. It does this by routing I/O through the peer storage processor to the storage processor owning the I/O's LUN. All attached hosts receive this benefit whether or not they are set to use ALUA on the front-end ports.

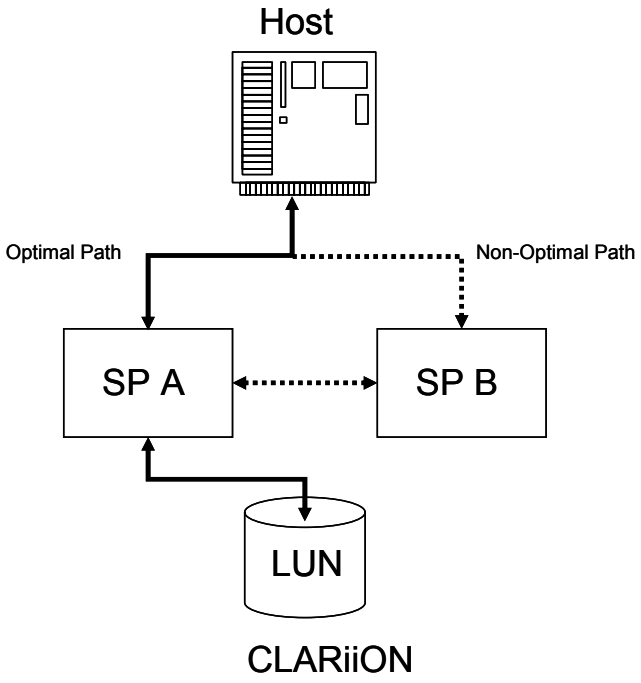

Figure 21 Optimal and non-optimal I/O paths

The original path is the *optimal path*. The path failed over to is the *non-optimal path*. ALUA does not perform automatic failback. A quick return to optimal path usage should always be made. The non-optimal path's I/O routing uses both storage processors for a single I/O. This lengthens host response time and may adversely affect storage system performance.

ALUA must be supported by the host's O/S. ALUA is supported by PowerPath, as well as by MPIO and PV Links.

Measuring reliability and availability

Reliability is the period of time that a device can perform its intended function upon demand. Reliability is calculated based on Mean Time Between Failure (*MTBF*) and Mean Time to Product Replacement.

Expected availability is presented as the percentage of time the system is able to access or transfer any desired block of data along with the number of downtime minutes per year. The reliability metrics for individual components are used to calculate an availability metric Mean Time To Data Loss (MTTDL).

Reliability metrics

MTBF

MTBF is a basic measurement of reliability. It is the period of time that passes before a component or system fails. Manufacturers use this measurement and its value is usually a "best case." Typically, this period is given in hours. For example, a disk manufacture's MTBF for a modern 15k rpm Fibre Channel hard drive is 1.5 million hours.

MTBPR

Expected reliability data is presented as the mean time between part replacement (MTBPR) in hours, which reflects how often, on average, a system will require a part replacement. It is not prudent to wait for devices to fail. A disk's failure is usually predicted (due to a rate of media errors for example) and the product is replaced before it fails.

Availability metrics

MTTR

Related to MTBF is MTTR (*mean time to repair*). This is a *maintainability* metric, and is important to availability in that it defines how long a component or subsystem is in a degraded state. It is measured in hours. It is the amount of time needed to restore a system to functionality once a failure is identified. MTTR typically is repair time made with the assumption the failure is already correctly identified.

In the context of CLARiiON, the most common MTTR is the time it takes for a rebuild to complete. A rebuild occurs in the event of a RAID group drive failure. During the time it takes to complete a rebuild, a RAID group is degraded. The MTTR is an example of and how a rebuild affects availability.

Chapter 11 Storage Object Availability

This chapter presents these topics:

RAID was developed to provide data redundancy. Data redundancy is a technique that provides in the case of a drive failure that data can be reconstructed. On top of RAID, vendors such as EMC have developed other techniques to reduce the effect of media errors and data-transmission errors within the storage system. Lost or corrupted data can be detected and if possible reconstructed. Note that on the CLARiiON the granularity of error detection and reconstruction is at the disk sector level. That is, a sector is the smallest amount of data in which an error can be detected, and reconstructed.

Parity and mirror RAID levels differ in how they handle data redundancy to provide availability.

RAID availability differences

All RAID levels except RAID 0 protect against single drive failures. Data protection is reduced when disk failures occur. Most RAID types are considered degraded when one drive in a RAID group fails. Certain RAID levels provide some level of protection against more than one drive failure. For example, a RAID 1/0 group can survive the failure of one drive from each primary or mirror pair. A RAID 6 group can survive two drive failures and still provide access to data. Data protection is restored upon completion of a rebuild operation, usually to a hot spare.

Mirror RAID groups can survive multiple disk failures – so long as they are in separate primary/mirror pairs, but RAID 6 can survive any two disk failures, making it the most highly available RAID type.

Mirror RAID level availability

CLARiiON's mirror RAID levels are:

- RAID 1
- RAID 1/0

With mirrors, two copies of data are stored on different drives. As data is written to a drive, the storage system automatically writes a second copy to a separate drive. The write is considered complete only after the data is safely written to both drives. Data can be read from either drive. The use of a full data copy for protection means that usable capacity of any mirror group is half the gross available capacity.

For mirrored data protection, should a drive fail or an error be detected, the lost or corrupted data is reconstructed from the copy on the surviving peer. The simple drive-to-drive copying of data during a rebuild affects access to the drive involved in the copy but does not affect other drives on the same bus.

Mirroring is not a substitute for data backup or point-in-time data replication. Erroneous information, user corrupted data, and data deletions are all copied to both drives of the mirrored pair.

RAID 1

RAID 1 is single disk mirroring. RAID 1 has good data availability

RAID 1/0

RAID 1/0 is the mirroring of disk stripes. RAID 1/0 has excellent data availability. A single disk failure results in no data loss. Multiple disk failures may be survived. However, a primary and its mirror cannot fail together or data will be lost. Note, this is *not* double-disk failure protection.

For example, in a four disk RAID 1/0 (2+2), consisting of two primary disks 0 and 1, and their mirrors 0′ and 1′. Both 0 and 1 may fail, preserving the data on the mirrors. Likewise, 0′ and 1′ may fail, preserving the data on the primaries. However, neither 0 and 0′ together, or 1 and 1′ together can fail and have the data survive.

Parity RAID level availability

CLARiiON's parity RAID levels are:

- RAID 3
- RAID 5
- RAID 6

Parity is a data protection function. Parity RAID levels use an algorithmic method of error detection and reconstruction to maintain the integrity of data. The parity technique creates information called parity data and the need to store parity data in the RAID group. This decreases the useable data capacity of the group. However maintaining the parity data uses less capacity than is used for mirroring. Parity data requires at most one drive's capacity per RAID group versus mirroring, which is an entire copy of the user data.

All parity RAID levels are striped. In each stripe, one stripe element contains the parity data. Should a drive fail or an error be detected within the media, the parity-elements and data elements from the working drives in the RAID group are used to reconstruct the contents of the failed drive.

The rebuild of disks from distributed parity information may adversely affect access to other drives on the same bus. The degree depends on the number of drives in the group and the rebuild rate. Rebuilds are a prioritized operation. At the priority of ASAP, rebuilds run at a high rate and a large RAID group can use all a back-end bus's bandwidth.

RAID 3

RAID 3 has good data availability. In RAID 3 all parity data is located on a single dedicated parity drive. If one drive of the group fails, no data is lost.

RAID 5

RAID 5 has good data availability. If one drive of the group fails, no data is lost. RAID 5 parity data is distributed throughout the drives of the group. Parity consumes the equivalent of one disk of the group's capacity.

RAID 6

RAID 6 provides the highest data availability on the CLARiiON. RAID 6 is a dual-distributed parity-RAID level. Two parity elements are computed for each stripe. A single or any two drives may fail in the group and result in no data loss. Note this is true double-disk failure

protection. RAID 6 parity data is distributed throughout the drives of the group. Parity consumes the equivalent of two disks of the group's capacity.

RAID 0

RAID 0 is a special case. There is no data protection built-into RAID 0. It offers only the performance and capacity advantages of a striped RAID. We do not recommend storing any information of business value on RAID 0 groups.

LUN availability

FLARE uses a process called Verify to check the integrity of a storage system's LUNs. The Verify process checks and repairs when possible the entire contents of a LUN location by location by:

- Detecting disk locations with data and parity inconsistencies
- Correcting locations with errors using redundant data before the LUN enters a degraded state and data cannot be reconstructed
- Remapping the location of corrected data
- Reporting uncorrectable disk locations

Location errors

Errors found in the LUN are either correctable by Verify or uncorrectable. Errors in a location are corrected from a RAID group's redundant data.

Any uncorrectable errors found by the Verify process are an indication of location(s) on the LUN that are unreadable by a host system. An example of an uncorrectable error location would be when a LUN's RAID group is operating in degraded mode and during that degraded period, one of the remaining drives encounters a media error. (Note that for RAID-level 6 this would be a correctable error.) The storage system's Verify process cannot determine if an uncorrectable error location on a LUN is in-use by the host or not. However, an uncorrectable error notification should be treated as a serious event that may involve data loss.

LUN verification

There are two types of Verify performed on a LUN. The two LUN verification processes are:

- SNIFF (SNiiFFER): Executes continuously and cyclically on all storage system LUNs at a very low rate as a background process. Simply a media check – it asks the drive to read a series of blocks. No consistency checks of parity or data are done.
- Background Verify (BV): Runs only when triggered by a failure or when requested by the user and executes at a higher rate. BV does a data consistency check.

SNiiFFER

SNiiFFER is not intended to verify the state of LUNs. Instead, it provides a background media checking mechanism that operates without any significant effect on host I/O performance.

When a LUN is bound, SNIFF is enabled at a very low priority. SNIFF does not affect back-end bus bandwidth utilization. It will request that a 512 KB region on each disk in the LUN is read, about every second. Whenever a LUN is idle, SNIFF increases its verification rate, which reduces the overall time required to complete a full pass. The number of drives in the LUN does not matter. The SNIFF I/Os are performed in parallel on every drive in the LUN. Because it is executing at a very low rate, depending on the capacity of the LUN, it may take several days to completely verify a single LUN.

Background Verify (BV)

BV is used to immediately check the consistency of a LUN when there is a concern over the validity or integrity of its data.

A BV is an operation that can be completed in hours. The duration of a BV depends on such factors as priority, host I/O loading, LUN size, RAID group type, and drive type.

BV is a prioritized operation. The priorities are Low, Medium, High, and ASAP. The duration of BVs can be decreased by increasing its priority to either High or ASAP. Note this will consume more storage system I/O resources—primarily disk utilization and back-end bus bandwidth. This may affect host performance. The default setting for BV is FLARE revision dependent. For FLARE revision 28 an higher, the default setting is Medium.

A BV can be scheduled either automatically or manually. A BV will automatically start after a trespass from a faulted SP to its peer SP. BV can be started manually through Navisphere when there are concerns about the validity of a LUN and a more current LUN status than is provided by the SNiiFFER process is needed

Reliability

Reliability is the ability of the storage system to run for long periods, and under operational stress, without suffering hardware failure or software faults. CLARiiON's architecture contains many features designed to ensure a very high level of reliability. High-quality components, assembly, design verification, and quality testing back up this design. The CX4 series storage systems' systemic reliability is rated at *five 9s* (99.999 percent) uptime based on hundreds of thousands of hours of field experience. The exact reliability measurements depend on the model and components used.

CLARiiON is designed and built to protect data, with full data path protection, throughout the storage system. This data path protection prevents against unreported *data miscompare* errors. The UltraPoint hard drive array enclosures (DAEs: DAE2P, DAE3P, or DAE4P) include a point-to-point design within the drive enclosures. The FLARE operating environment monitors low-level diagnostics at the drive level that include fault detection, isolation, and error correction. RAID protection is supplemented with global and proactive hot sparing; sector-based checksum of all user data; and automatic background verification of the disk media.

Chapter 12 CLARiiON Availability

This chapter presents these topics:

Front end

All CX4 series models come standard with multiple iSCSI and Fibre Channel front-end ports on each SP. Connections to iSCSI and Fibre Channel hosts are supported at the same time.

High availability requires multiple connections. In a highly available configuration, I/O connections from a host should be made to more than one port on a storage processor. In addition, an I/O connection between the host and the peer storage processor should be set up.

Configurations allowing the host to connect to more than one of the storage system's front-end ports are a technique referred to as *multipathing*. These additional ports may be used to provide primary access to one LUN and alternate access to another. The existence of alternate paths enables a front-end port to front-end port redirection on the storage system in the event of a path or port failure. *Failover* is the use of an alternate path in the event of a failure. *Failback* is the return to the original path, following the correction of the path's failure.

Typically a failover is the result of a networking error. However, it is possible for failover to occur as a result of a storage system component failure. Failover is initiated on the storage system by the host-based PowerPath, or other host-based multipathing applications (see the Multipathing section for details). In failover, the operational port begins servicing host I/O requests.

Along with redundant network configuration, when provisioning it is prudent to create connections to at least two front-end ports on each storage processor. In addition, at least one or more further connections should be made to a front-end port of the peer SP on the storage system. This configuration allows port-to-port failover on the owning SP, in addition to preparing for a SP failover.

Storage processor

CLARiiON storage systems have two highly-reliable storage processors. When properly configured the storage processors provide a redundant operating environment for storage access.

Active/passive ownership model

Each CLARiiON has two storage processors. For workload balancing the storage system's LUNs are distributed between the storage processors. This results in storage processors efficiently working in parallel to service the workload.

LUNs are assigned to only one storage processor at a time. This is considered *active ownership*. Either storage processor can access a LUN on the storage system's back end. However, I/O is not directed to the non-owning storage processor. The existing but inactive paths to un-owned LUNs are *passively owned* by the peer storage processor. Note this requires the storage processors to mirror each other's write cache (see the Memory section).

If there is a network failure of all I/O to one storage processor, or the unlikely failure of a storage processor, LUNs are *trespassed* to their peer storage processor. Trespass changes the ownership a LUN from passive to active. Trespass is initiated transparently by PowerPath or other host-based multipathing applications (see the Multipathing section). The surviving storage processor assumes ownership and begins servicing host I/O requests as soon as cache page associations with its new LUNs are completed.

Note that for high availability, a single CLARiiON storage processor will need enough CPU resources to handle the workload of the entire storage system.

Back end

CLARiiON's back end is composed of multiple disk enclosures with redundant buses and storage configured to ensure data security.

Vault and write cache availability

The first five hard drives (0 through 4) in DAE 0 in the CLARiiON CX4 series and the first four drives in the AX series have significant capacity reserved for the CLARiiON's operation. These drives are referred to as the system drives. They contain the vault. The vault is capacity reserved for the write cache (so it can be quickly written to this region during certain failures), the storage system's FLARE operating system files, the Persistent Storage Manager (PSM), and the FLARE configuration database.

The CLARiiON boots off of all the system drives. There is actually considerably more private data on those first five drives than just the vault, which is where write cache is de-staged to in case of a power failure or certain component failures. The figure below shows the layout of the five system drives of a CX4. There are several private LUNs on these drives including LUNs for:

- Operating system for both SPs in RAID 1 (SPA - drives 0 and 2, SPB - drives 1 and 3)
- Persistent Storage Manager (PSM)
- FLARE database in RAID 1 triple mirror (drives 0-2)
- The vault in RAID 5 (drives 0-4).
- These drives provide protection from two drive failures of the first four drives.

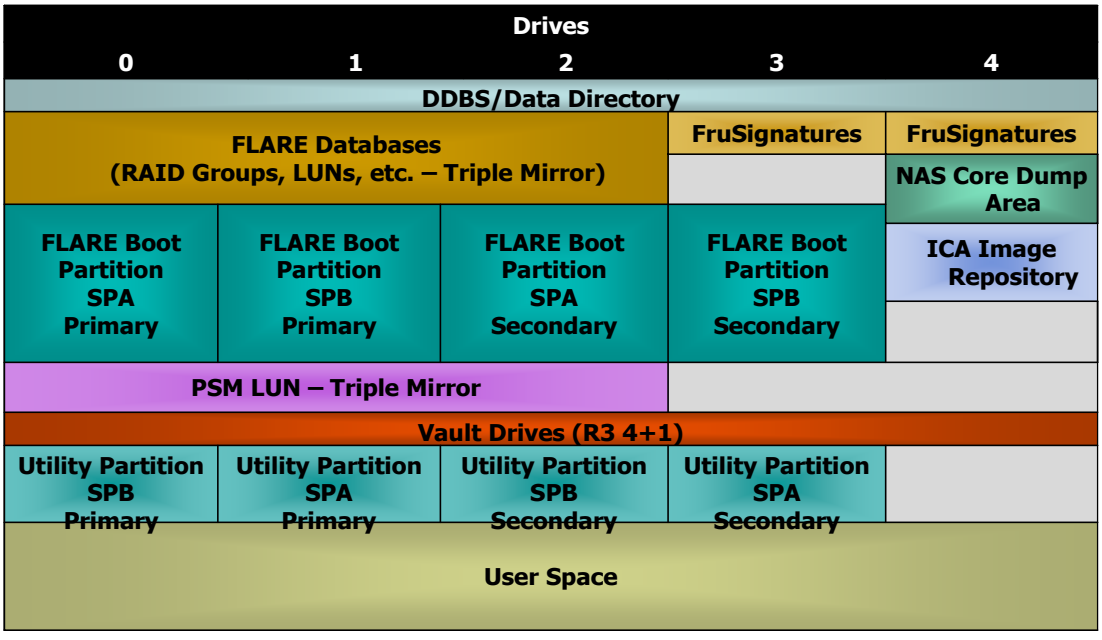

Figure 22 CX4 vault drive layout

System drives can be used just as any other drives on the system. However, system drives have less usable capacity than data drives due to the reserved capacity. In addition, their usage by the CLARiiON can affect the response time for application data stored there. See the *Best Practices* papers for vault drive provisioning performance planning advice.

The OS, PSM, FLARE database, or the vault will not be rebuilt to a hot spare in the case of a single drive failure in any of these private LUNs. Only user LUNs created in the drive's user space are rebuilt to a hot spare; the private LUNs will be rebuilt when the faulted drive is replaced.

For example, if drive 0 on Bus 0 Enclosure 0 fails and there are no user LUNs, no hot spares will be used in a swap. If there is a user LUN that is bound across that drive, only the user LUN data is rebuilt to the hot spare. When the faulted drive is replaced, all data will be copied/rebuilt to the drive 0 (FLARE DB from mirror, vault from parity, PSM from mirror, and user LUN from hot spare).

CX4 back-end bus

As explained in the section Back end, the CX4 CLARiiON's back-end bus redundantly connects both SPs to the storage system's drives housed in the DAEs. The Fibre Channel back end provides two paths to the data on the CLARiiON's drives.

There are two LCCs per DAE. Each SP attaches to a DAE bus through an LCC by its Fibre Channel bus. Should a DAE's LCC fail an SP loses communication with all the drives in the DAE. How the CLARiiON recovers from the failure is affected by the:

- CLARiiON's FLARE firmware version
- RAID group provisioning (multi-bus or single-bus)
- The RAID group's RAID-level

The recovery from an LCC failure may result in the use of either a Background Verify (BV) or Rebuild and possibly a host trespass. Note that recovery from an LCC failure can adversely affect host I/O performance. Details on LCC availability can be found in the *Best Practices* papers.

Global hot sparing

CLARiiON's operating environment constantly monitors the state of its drives. *Hot spares* are drives used to replace failing drives. Proactive sparing is the automatic activation of a hot spare when a drive indicates its pending failure through its SMART reporting feature.

Hot spares replace a failed drive in any redundant RAID group with bound LUNs Hot spares will not replace unbound drives or drives in RAID groups of RAID level 0 or individual disks.

The number of hot spares per storage system is model dependent. Entry-level storage systems may have fewer hot spares allocated than higher models.

Hot sparing occurs when a drive has failed. A spare is selected algorithmically from a pool of previously-designated hot spares. There are some restrictions on which types of drives may be used as hot spares.

In proactive sparing, a suspected failing drive is copied to a spare before it fails (see the SMART section). The suspect drive continues servicing I/Os while this process goes forward. When the

copying completes the suspect drive is failed (powered down) and the hot spare takes its place in the RAID group. If the drive fails before the process completes the remaining missing data is rebuilt to the hot spare by the process described below.

Copying data to the spare proactively is much less resource intensive and usually faster than rebuilding data from parity and the surviving drives. The disk-to-disk copy puts less load on the bus than a parity rebuild. This advantage extends to mirror-type RAID levels, where the copy is performed while the RAID group is not in a degraded mode. This is an availability advantage. For example, if a UER is encountered during the copy, the missing data can be reconstructed. In a rebuild-type data recovery were in-progress, where the failed drive is no longer available, a UER results in loss of data. (The exception being RAID 6 rebuilds, which are immune to double drive failures.)

When a failed drive is manually replaced, a process called *equalization* restores the data from the hot spare to the RAID group's replacement drive. Equalization is a disk-to-disk copy operation and affects the bus and affected drive the same as a RAID 1 rebuild or proactive copy. After equalization the hot spare is returned to the global pool for reuse as a replacement.

Note that a RAID group's operation with a hot spare is not considered a normal operation. Normal operation is restored when the failed drive is replaced, its replacement is equalized, and it begins to handle I/O. No Rebuild occurs if there are no hot spares configured of the appropriate type and size when a drive fails. The RAID group remains in a degraded state until the failed drive is replaced, and then the failed drive's RAID group rebuilds from parity or its mirror drive, depending on the RAID level.

Proactive hot sparing eliminates the risk of a second disk failure during rebuild. A proactively spared RAID group always has its full number of storage devices. Note that it is still considered degraded. In addition, equalization, and not a rebuild, is needed to restore the RAID group's status. Equalization does not use the resources that a rebuild does.

Global hot sparing is not a CLARiiON requirement for normal operation. However, it is a prudent choice if you wish to increase overall availability.

Rebuild logging

The purpose of rebuild logging is to allow a drive that appears to be timing out additional time to recover from whatever it is doing, without delaying host I/O execution for an extended period of time, and to avoid performing a full Rebuild operation to the drive. This is also called *drive probation*.

The situations in which rebuild logging get invoked are specific. If a drive actually faults – for example it no longer accepts commands, reports a fatal condition (such as "Hardware Error"), or is bypassed from the bus – it is immediately retired and a hot spare (if available) is employed. Rebuild logging gets invoked when commands already issued to the drive don't complete in a reasonable amount of time – that is, they appear to be timing out, but the drive is still physically present, and hasn't faulted in any of the conditions listed above.

Software and firmware update

CLARiiON's operating environment can be updated to incorporate new features, and to correct discovered errors.

The disk array operating system (FLARE) can be upgraded online via the Non-Disruptive Update (NDU) process. The DAE LCC Firmware (FRUMON) is upgraded online via the NDU process (contained within FLARE upgrades). Disk firmware is not usually recommended nor required

for upgrade (unless a field change order or ETA is implemented) and requires scheduled downtime for it to be upgraded

Chapter 13 Conclusion

Having read this document, you should now be familiar with CLARiiON's fundamental concepts so that you can understand the details of performance and availability tuning described in *EMC CLARiiON Performance and Availability: Release 29.0 Firmware Update — Applied Best Practices*. (This paper is also referred to as the *Applied Best Practices* paper.)

A lot of information is presented in this paper. However, there are some very simple points that you must take away to get the most out of your CLARiiON and to understand the tuning recommendations in *Applied Best Practices*. They are:

- Performance enables behavior to be evaluated. To do so you need to:
  - Know how to describe your workload.
  - Have anticipated, or past and current performance metrics.
- CLARiiON is a computer-based device with resources to manage.
- Availability is about redundancy and data integrity.

To get the highest performance, you need to understand a storage system's workload. This includes knowledge of the host applications. You should understand how your applications are generating I/O. The workload may be simple, coming from a single application; it might also be complex, meaning it is generated by more than one application on one or more hosts with different I/O profiles. You need to know the throughput, bandwidth, and required response time of your I/O to determine the needed performance. If you can characterize your workload, tuning is a straightforward process. In addition to characterizing the workload, the information gathered is important for creating the historical baselines for future tuning, and investigations. Note that, if the workload's demands exceed the storage system's performance capabilities, applying performance tuning will have little effect.

A CLARiiON is actually two computers with a lot of attached drives. It has CPU, memory (cache), and I/O (buses and storage) resources that need to be managed and provisioned for it to reach optimal performance and highest availability. Be aware of the usable bandwidth of its buses and ports, its usable CPU cycles, and the throughput, bandwidth, and storage capacity of its drives. A part of these resources is consumed by the workload. Another part will be consumed by overall system maintenance, such as backup and storage system management tasks like LUN expansion. In addition, reserves need to be maintained to ensure acceptable performance in degraded modes of operation.

CLARiiON's built-in reliability and redundant hardware make it a highly available system. Redundancy extends to data protection. Careful provisioning is required to ensure the highest data protection *and* performance. Remember that configuring CLARiiON for high availability may adversely affect performance. (Performance and availability are both dependent on the same resources.) The allocation of resources to meet the immediate demands of the workload, while maintaining reserves for maintenance and to ensure high-availability is a continuous process.

This document and others mentioned here are found on Powerlink®, EMC's password-protected extranet for customers and partners. We encourage you to read the best practices information for your workload's applications, and its host hardware and software. The best practices from the vendors of your network elements should also be of interest. An additional networking information resource is the *EMC Networked Storage Topology Guide* found on Powerlink.

Finally, note that a CLARiiON storage system is always an important, but dependent part of a much larger system. Host and networking-related fundamentals are beyond the scope of this document. However, the storage system's performance and availability are greatly affected by the performance and availability of the networks attaching it to hosts, host applications, and the host's hardware and operating system software. In many cases, improving the performance or availability of the network or the host results in better performance and higher availability than tuning the storage system:

Chapter 14 Glossary

`When I use a word, it means just what I choose it to mean -- neither more nor less.'

-- Humpty Dumpty, "Through the Looking Glass (And What Alice Found There)" (1871) by Lewis Carroll

Terms

10 GbE— 10 Gigabit per second Ethernet protocol.

ABQL — Average busy queue length (per drive).

Active-active — Redundant components are active and operational.

Active-passive — Redundant components ready in a standby operational mode.

AFR — Annual Failure Rate.

ALUA — Asymmetric logical unit access protocol.

Allocated Capacity — Actual capacity ready for use in a thin LUN pool. User defined capacity of a Thin LUN.

American National Standards Institute — An internationally recognized standards organization.

ANSI — American National Standards Institute.

Application software — A program or related group of programs performing a business function.

Array — Storage system.

Asymmetric Logical Unit Access — An industry standard multipathing protocol.

Attachment — Drive hardware connector or interface protocol. On a CLARiiON it can be Either Fibre Channel, SAS, or SATA.

Authentication — Verifying the identity of a communication to ensure its stated origin.

Authorization — Determination if a request is allowed access to a resource.

Available capacity — Capacity in a thin LUN pool that is not allocated to thin LUNs.

Availability — Continued operation of a computer-based system after suffering a failure of fault.

Back-end — A logical division of the CLARiiON's architecture from SP to the back-end bus(es) and drives.

Back-end bus — The CLARiiON's Fibre Channel buses connecting the storage processors to the drives

Back-end I/O — I/O between the storage processors and the drives over the back-end buses.

Background Verify — Automated reading of the LUN's parity sectors and verification of their contents by the CLARiiON for fault prevention.

Backup — Copying data to a second, typically lower performance drive as a precaution against the original drive failing.

Bandwidth — A measure of storage-system performance, as measured in megabytes per second (MB/s).

Bind — To associate drives to create a RAID group.

Bit — The smallest unit of data. It has a single binary value, either 0 or 1.

Block — See sector.

Bottleneck — Of all resources needed to perform a task, the bottleneck is the resource that is at maximum performance. It limits overall performance, since other resources are not completely used and could do more work.

Broadband — High-bandwidth networks and interfaces of data and voice communications.

Buffering — Application or file system retention of data for reuse or to optimize data flow.

BURA — Backup, Recovery and Archiving data storage security domain.

Bursty — The characteristic of I/O, which over time is highly variable, or regularly variable, with well-defined peaks.

Bus — An internal channel in a computerized system that carries data between devices or components.

Busy hour — The sliding 60-minute period during which occurs the maximum total traffic load in a 24-hour period.

BV — Background Verify.

Byte — Eight computer bits.

Cache — Memory used by the storage system to buffer read and write data and to insulate the host from disk access times.

Cache Hit — A cache hit occurs when data written from or requested by the host is found in the cache, avoiding a wait for a disk request.

Cache Miss — Data requested by the host is not found in the cache so a disk request is required.

Cache Page Size — The capacity of a single cache page.

CAS — Content Addressable Storage object-based storage as implemented by EMC Centera®.

CHAP — Challenge Handshake Authentication Protocol.

CIFS — Common Internet File System.

CLI — Command Line Interface.

Client — On part of a Client-Server architecture. Typically a user computer or application in communication with a host.

Client-Server — A network architecture between consumers of services (Clients) and providers (Hosts).

Clone — Binary copy of a source LUN.

CMI — Configuration Management Interface.

Coalescing — The grouping of smaller cached I/Os into one or more larger ones.

Command line interface — Mechanism for interacting with a computer operating system or software with text-based commands.

Common Internet File System — Microsoft Windows file sharing protocol.

Concurrent I/O — When more than one I/O request is active at the same time on a shared resource.

Core — A processor unit co-resident on a CPU chip.

Core-switch — Large switch or director with hundreds of ports located in the middle of a SAN's architecture.

CPU — Central Processing Unit.

CRM — Customer Relationship Management.

Customer relationship Management — Applications used to attract and retain customers.

DAE — Disk array enclosure.

DAS — Direct attached storage.

Data — Information processed or stored by a computer.

Data center — A facility used to house computer systems, storage systems, and associated components.

Data link — Digital communications connection of one location to another.

Data mining application — A database application that analyzes the contents of databases for the purpose of finding patterns, trends, and relationships within the data.

Data warehouse — A collection of related databases supporting the DSS function.

DBMS—Data Base Management System.

Decision support system — A database application, used in the "data mining" activity.

Degraded mode — Continued operation after a failure with possible loss of performance.

Departmental system — A storage system supporting the needs of a single department within a business organization.

Destage — Movement of data from cache to drives.

Dirty Page — A cache page not yet written to storage.

Disk array enclosure — The rack-mounted enclosure containing a maximum of 15 CLARiiON drives.

Disk controller — A component of a hard drive. The microprocessor based electronics controlling operation of a hard drive.

Disk crossing — An I/O whose address and size cause it to require access to more than a single stripe element causing two separate back-end I/Os to result.

Disk processor enclosure — Physical rack mounted cabinet containing a CLARiiON storage processor. This enclosure contains drives.

DPE — Disk processor enclosure.

DR — Disaster recovery.

Drive — A hardware storage component from which you can read and write data. Typically a hard drive, but also an EFD.

DSS — Decision support system.

Dump — Copying of the cache image to the vault.

Edge switch —A Fibre Channel switch located on perimeter of a core-edge configured SAN.

EFD — Enterprise Flash Drive.

Enterprise Flash Drive — EMC term for SSD-type drive.

Enterprise resource planning — Applications managing inventory and integrating business processes across the organization.

Enterprise system — A storage system supporting the needs of an entire business organization.

Environment — A computer's state, determined by the hardware platform and which applications and system software are executing.

Equalization — Restoration of the data from a hot spare to the RAID group's replacement drive.

ERP — Enterprise Resource Planning.

ESM — EMC Support Matrix.

Ethernet — A frame-based computer networking architecture for LANs. The IEEE 802.3 standard.

Failure — A malfunction within the hardware components of a system.

Fail back — Multipathing restoration of original path usage after correction of a failure.

Fail over — Multipathing use of an alternate path resulting from a failure

Failure mode — The underlying cause of a failure, or the event that starts a process leading to a failure.

Fan-in — Attachment of many hosts to fewer storage system ports.

Fan-out — Attachment of a few storage system ports to many hosts.

FAQ — Frequently Asked Question.

Fault — An error or failure in the operation of a software program.

Fault tolerance — The ability for a system to continue operation after a hardware or software failure.

FC — Fibre Channel.

FCIP — Fibre Channel traffic over IP networks tunneling technique.

FCP — SCSI Fibre Channel Protocol.

Fibre Channel — A serial data transfer protocol. ANSI X3T11 Fibre Channel Standard.

File — A collection of named data.

File system — The system an O/S uses to organize and manage computer files.

Filer — NAS fileserver accessing shared storage using a file-sharing protocol.

FLARE — Fibre Logic Array Runtime Environment. The CLARiiON's O/S.

FLU — FLARE LUN, pronounced "flu."

Flush — The process of writing data held in the write cache to the drives.

Forced flush — The high priority writing of data to drives to clear a full write cache.

Front-end — A logical division of the CLARiiON's architecture, including the communications ports from hosts to the SP.

GB — Gigabyte.

GB/s — Gigabytes per second.

Gb/s — Gigabits per second.

GbE— Gigabit Ethernet (1 Gb/s Ethernet).

GHz — Gigahertz.

Gigabit — One thousand million bits.

Gigabyte — One billion bytes or one thousand megabytes.

Gigahertz — One billion times per second (1,000,000,000 Hz).

GigE — 1 Gb/s Ethernet.

GMT — Greenwich Mean Time.

Graphical user interface — Mechanism for interacting with a computer operating system or software with screen visual objects.

GUI — Graphical user Interface.

HA — High Availability or Highly Available.

HBA — Host Bus Adapter. A Fibre Channel network interface adapter.

Head-crash — A catastrophic hard drive failure where a read/write head makes physical contact with a platter.

Hertz — Once per second.

Highly available — A system able to provide access to data with a single fault.

Host — A server accessing data or applications from a storage system by means of a network.

Hot spare — An inactive installed hard drive used by the storage system to automatically replace a failed drive.

HP-UX — A proprietary Hewlett Packard Corporation version of the UNIX O/S.

HVAC — Heating, Ventilation, and Air Conditioning.

Hz — Hertz.

IEEE — Institute of Electrical and Electronics Engineers.

IETF — Internet Engineering Task Force

iFCP — Protocol allowing FC devices usage of an IP network as a fabric switching infrastructure.

ICA — Image Copy Application.

IEEE — Institute of Electrical and Electronics Engineers.

iFCP — Protocol allowing FC devices usage of an IP network as a fabric switching infrastructure.

IHAC — I Have A Customer.

Initiators — iSCSI clients.

Institute of Electrical and Electronics Engineers — An international standards organization.

International Organization for Standardization — International organization maintaining standards.

Internet Engineering Task Force — International organization standardizing the TCP/IP suite of protocols.

Internet Protocol — A protocol used with TCP to transmit and receive data on Ethernet networks.

IOPS — Input/Output operations Per Second.

IP — Internet Protocol.

IPSec — Internet Protocol Security.

IPv4 — Internet Protocol version 4.

IPv6 — Internet Protocol version 6.

iSCSI — Internet SCSI protocol. A standard for sending SCSI commands to drives on storage systems.

ISL — Interswitch Link used to connect two or more switches to create a network.

ISO — International Organization for Standardization.

IT — Information Technology. Also, the department that installs and maintains computer-based systems for a business organization

JBOD — Just a Bunch Of Disks.

KB — Kilobyte.

Kb — Kilobit.

Kb/s — Kilobits per sec.

KB/s — Kilobytes per sec.

Kilobits — One thousand bits.

Kilobyte — One thousand bytes.

LAN — Local Area Network.

Large-block — I/O operations with capacities greater than 64 KB.

Layered Apps — Layered Applications.

Layered Applications — CLARiiON application software.

LBA — Logical Block Address.

LCC — Link Control Card

Legacy system — An existing production storage system supporting the current needs of the organization.

Linux — Any of several hardware independent open-systems operating system environments.

Load balancing — The even distribution of the data or processing across the available resources.

Local Area Network — A computer network extending over a small geographical area.

Logical Block Address — A mapping translation of physical drive sector addresses into the logical SCSI block addresses.

Logical unit number— A SCSI protocol entity, to which I/O operations are addressed.

Loop— Both SP A and SP B's shared connection to the same numbered back-end bus.

LUN — Logical Unit Number. Hosts access storage using logical unit numbers, which are exported by a SCSI target.

LVM — Logical Volume Manager. A host-based storage virtualization application. Similar to Microsoft Logical Disk Manager.

MAN — Metropolitan Area Network.

MB — Megabyte.

MB/s — Megabytes per second.

Mb — Megabit.

Mb/s — Megabits per second.

Mean time between failure — The average amount of time that a device or system goes without experiencing a failure.

Mean time to data loss — Probability of when a failure will cause RAID protected data loss.

Mean time to repair— An estimate of the time required to repair a failure.

Media — The magnetic surface of a hard drive's platter used for storing data.

Megabit — One million bits.

Megabyte — One million bytes or one thousand kilobytes.

Megahertz — A million cycles per second (1,000,000 Hz).

Memory Model — Description of how threads interact through memory

Metadata — Any data used to describe or characterize other data.

MetaLUN — A LUN object built by striping or concatenating multiple LUN objects.

MHz — Megahertz.

Mirror — A replica of existing data.

MirrorView — CLARiiON disaster recovery application

MPIO — Microsoft Multi-Path I/O

MR3 Write — The action the CLARiiON RAID engine performs when an entire RAID 5 stripe is collected in the cache and written at one time.

MTBF — Mean Time Between Failures.

MTTDL — Mean Time To Data Loss.

MTTR — Mean Time To Repair.

Multi-path — The provision for more than one host I/O paths between LUNs.

Multi-thread — Concurrent I/O threads.

Name-server — A process translating between symbolic and network addresses, including Fibre Channel and IP.

NAS — Network Attached Storage.

Native Command Queuing — A drive-based I/O execution optimization technique.

Navisphere — CLARiiON's resource management system software.

Navisphere Analyzer — CLARiiON's performance analysis system software.

NCQ — Native Command Queuing

Network — Two or more computers or computer-based systems linked together.

Network element — A device used in implementing a network. Typically refers to a switch or router.

Network file system — A UNIX/Linux file sharing protocol.

Network interface card — A host component that connects it to an Ethernet network.

NDU — Non-Disruptive Update

NFS—Network File System

NIC — Network Interface Card.

Non-disruptive update — Upgrading of system software with minimal effect on performance and no downtime.

Non-optimal path — The ALUA failed over I/O path from host to LUN

O/S — Operating System.

OLTP — OnLine Transaction Processing system.

Online transaction processing — multiuser systems supported by one or more databases handling many small read and write operations.

Operating environment — Operating System.

Operating system — A computer-based system's software responsible for resource management and coordination of tasked functions.

Optimal path — The normal operations I/O path from host to LUN

Ownership — SP management of LUN I/O.

Page — Cache unit of allocation.

Parallel ATA — Disk I/O protocol used on legacy CLARiiONs.

PATA — Parallel ATA disk I/O protocol.

Petabyte — One quadrillion bytes or one thousand terabytes.

PB — Petabyte.

PC — Personal Computer

PCI Express — A bus protocol used by computer-based systems.

PCIe— PCI Express.

PDU — Power Distribution Unit.

Percentage utilization — A measurement of how much of a resource is consumed by operation.

Platform — The hardware, systems software, and applications software supporting a system, for example DSS or OLTP.

Platter — A component of a hard drive; it is the circular disk on which the magnetic data are stored.

Port — An interface device between the storage system and other computers or peripheral devices. Also, an interface between a drive and a bus.

Power distribution unit — A CLARiiON component connecting data center electrical power trunks to the storage system.

Powerlink — EMC's password-protected extranet for customers and partners.

PowerPath — EMC host-based multipathing application.

Prefetch — A caching method by which some number of blocks beyond the current read are read and cached in the expectation future use.

Private LUN — A LUN managed by FLARE and not addressable by a host.

Protocol — A specification for device communication.

PSM — Persistent Storage Manager.

PV Links —HP-UX Physical Volume Links

QA — Quality Assurance.

Quality Assurance — The department, or policies and procedures for verifying promised performance and availability.

QFULL — Queue Full.

QoR — Quality of Result.

QoS — Quality of Service.

Quality of result — A term used in evaluating technological processes or implementations.

Quality of service Agreement — A defined, promised level of performance in a system or network.

Queue full — An iSCSI protocol signal sent to hosts indicating a port or LUN queue cannot accept an entry.

RAID — Redundant Array of Independent Disks.

RAID group — A logical association of between two to 16 drives with the same RAID level.

RAID level — An organization of drives providing fault tolerance along with increases in capacity and performance.

Random I/O — I/O written to locations widely distributed across the file system or partition.

Raw drive — A hard drive without a file system.

RDBMS — Relational Database Management System.

Read Cache — Cache memory dedicated to improving read I/O.

Read/write head — Component of a hard drive that records information onto the platter or read information from it.

Read-ahead — See prefetch

Rebuild — The reconstruction of a failed drives data from through either parity or mirroring.

Recovery time objective — The estimated amount of time to restore a system to full operation after a fault or failure.

Redundancy — The ability of the storage system to continue servicing data access after a failure through the use of a backup component or data protection mechanism.

Relational database management system — A database typically hosted on a storage system, for example, Oracle, DB2, Sybase and SQL Server.

Reliability — The ability of the storage system to run for long periods, and during stress, without suffering either hardware or software faults.

Request size — In a file system, the size of the block actually read from the drive.

Response time — A measure of performance including cumulative time for an I/O completion as measured from the host.

RFC — Request for Comments.

Rich media — A workload that allows for active participation by the recipient. Sometimes called interactive media.

Rotational latency — The time required for a disk drive to rotate the desired sector under the read head.

Rpm — Revolutions per minute.

RPQ — Request for product qualifier.

RTO — Recovery time objective.

RV — Rotational vibration.

SAN — Storage area network.

SAN Copy — CLARiiON storage system to storage system copy application.

SAP — The enterprise resource planning application produced by the software company, SAP AG.

SAS—Serial attached SCSI.

SATA—Serial ATA disk I/O protocol.

Saturation — The condition in which a storage system resource is loaded to the point where adding more I/O dramatically increases the system response time but does not result in additional throughput.

SCSI — Small computer system interface.

Sector — The smallest addressable information unit on a hard drive. A region on a hard drive track on which 512 bytes of data.

Sequential I/O — A set of IO requests whose pattern of address and size result in serial access of a complete region of data in monotonically increasing addresses.

Serial Attached SCSI - A point-to-point serial protocol for moving data to drives.

Service Time — The interval it takes a drive or resource to perform a single I/O.

Shelf — **DAE**.

Short stroking — A LUN performance optimization technique of only using a portion of a RAID group.

SLA — Service Level Agreement. A contract between a service provider and a customer providing a measurable level of service or access to a resource.

SLIC — Small I/O Card. The generic name for the CX4 UltraFlex I/O modules, either Fibre Channel or iSCSI.

Small Computer System Interface — Set of standards for physically connecting and transferring data between hosts and drives.

Small block — I/O operations up to 16 KB.

SnapView — CLARiiON point-in-time copy application.

Snapshot — Backup copy of how a LUN looks at a particular point in time.

Solid State Disk — A drive using non-volatile semiconductor memory for data storage.

SP — Storage processor.

SPE — Storage Processor Enclosure.

Spike — A sudden, sharp, and significant increase in load on the storage system.

Spin down —Setting inactive hard drives into a low-power “sleep” mode.

Spindle — A component of a hard drive; it is the axel platters are mounted on. Also, spindle sometimes refers to a hard drive.

SPS —Standby Power System.

SSD — Solid State Disk.

Stack — Layered protocols.

Storage area network — A network specifically designed and built for sharing drives.

Storage array — A storage system.

Storage object — A logical construct or physical device supporting both read and write data accesses.

Storage pool — A logical construct of drives supporting both read and write data accesses.

Storage processor — A logical division of the CLARiiONs architecture including the CPUs and memory.

Storage processor enclosure — Physical rack mounted cabinet containing CLARiiON Storage Processor. This enclosure contains no drives.

Storage system — A system containing multiple hard drives, cache and intelligence for the secure and economical storage of information and applications.

Stripe crossing — If a back-end I/O is not contained in an entire stripe, a stripe crossing is incurred, which takes more than one stripe

Stripe element — Capacity allocated to a single device of a stripe.

Stripe size — The usable capacity of a RAID group stripe.

Stripe width — The number of hard drives in a RAID group stripe.

Stripe — Distributing sequential chunks of storage across many drives in a RAID group.

Stroke — Movement of a hard drives' read/write head across the platter.

Switch — A Layer 2 device providing dedicated bandwidth between ports and switching functions between storage network devices.

System software — Operating system and applications used for a computer's management.

TB — Terabyte.

TCP — Transmission Control Protocol. A protocol used with IP to transmit and receive data on Ethernet networks.

TCP/IP — The pair of communications protocols used for the Internet and other similar networks

TCP/IP offload engine — A co-processor based host component that connects it to an Ethernet network.

Terabyte — One trillion bytes or one thousand gigabytes.

Thread — An independent I/O request that may execute in parallel with other requests.

Throughput — A measure of performance of I/Os over time; usually measured as I/Os per second (IOPS).

TLU — Thin Provisioning LUN.

TOE —TCP/IP Offload Engine.

Topology — How parts of a system component, subsystem, or system are arranged and internally related.

Track — A ring-like region of a hard drive platter on which data is organized and stored.

Tray — DAE.

Trespass — A multipathing initiated change in SP LUN ownership as a result of a failure or command.

UER — Unrecoverable Error Rate

UNIX — Any of several open system operating system environments.

Unrecoverable error rate — Bit error reliability metric for hard drives.

UPS — Uninterruptable Power Supply.

User — An individual who operates application software.

Vault — Special area on drives of DAE0 for storage of CLARiiON system files and cache dumps.

Virtual machine — A software application emulating a server hardware environment.

Virtual Provisioning — Explicit mapping of logical address spaces to arbitrary physical addresses. Ex: Presenting an application with more capacity than is physically allocated.

VLAN — Virtual Local Area Network.

VLAN Tagging — Mechanism for segregating VLANs.

VM — Virtual Machine.

VMware — EMC's family of virtual machine applications.

Volume — LUN.

WAN —Wide Area Network**.**

Watermark — A cache utilization set point.

WCA —Write Cache Availability.

Wide area network — A computer network extending over a large, possibly global, geographical area.

Windows — Any of several proprietary Microsoft operating system environments.

Wintel — Industry term used to describe computers based on an Intel hardware architecture and a Microsoft Windows operating system.

Wire Rate — The maximum bandwidth for data transmission on the hardware without any protocol or software overhead. Also known as "Wire Speed."

Workload — The characteristics of a pattern of IO requests presented to the storage system to perform a set of application tasks, including IO size, address pattern, read to write ratio, concurrency, and burstiness.

World wide name — A unique identifier in a Fibre Channel or SAS storage network.

WORM — Write Once Read Many.

Write Cache — Cache memory dedicated to improving host write I/O response time by providing quick acknowledgement to the host while destaging data to disks later in the background.

Write-aside — Bypass of the write cache, where the RAID engine dispatches a write immediately to the disks.

Write-aside Size —The largest request size, in blocks, written to cached for a particular LUN.

WWN — World Wide Name