Technical white paper

A Deep Dive into Continuous Access Synchronous Software

HP XP7 Storage

Table of contents

Click here to verify the latest version of this document

Terms

AutoLUN XP—Internal XP software used to migrate data from different classes of storage without any interruption from the host. Can only be used in manual mode (with External Storage XP).

Business Copy XP—The XP's internal mirroring software. Users can make up to 9 full (clone) copies from one primary volume.

XP Continuous Access—The XP's 'array to array' mirroring product. Provides synchronous, asynchronous sidefile or asynchronous journal methods of replication.

CLPR—Cache Logical Partition. A virtual private data cache partition such that a 'cache hungry' application should not affect the performance of other applications (in other CLPRs).

CVS—The XP's software product used to carve up an LDEV into smaller fractions, for use as multiple LUs.

External Device Group—A grouping of External Storage XP volumes under a single name.

Port—An array connection point through a Fibre Channel GBIC initiates connections to the SAN, like an HBA on a server.

LDEV—An XP logical device manifesting in a particular RAID format and emulation type (e.g. CU 1, ldev 2 may be an Open-3 LDEV in 3Data+1Parity RAID5). Once an LDEV is registered within the XP, it can either mapped directly to a port and LU, aggregated with other LDEVs to create a larger LU (LUSE), or carved up via CVS to create a smaller LU. External LUs map in as a VDEV and consequently an LDEV.

LU—Logical unit or disk volume.

LUSE—"Logical Unit Size Expansion". The XP's software product that allows multiple LDEVs to be aggregated into a larger LU.

LUN—Logical Unit Number (often misused in place of LU)

PAIR States—SMPL (Simplex) means NO pair relationship exists. PAIR means data is actively updating from the P-VOL to the S-VOL. SUS (suspend) means data movement between the P-VOL and S-VOL is temporarily suspended (allowing host alteration of both the P-VOL and S-VOL). PSUE – the P-VOL relationship is suspended in an error state.

P-VOL—The primary or "production" side volume of a pair.

RAID Manager XP—A host-based program capable of communicating with XP disk arrays for the purpose of CLI monitoring and manipulation of BC XP or Cnt Ac XP volume pairs. Raid Manager XP is able to manage FlexCopy and External Storage XP resources beginning with version 1.12.06. This is done in part with the newly added "-fe" option to the Raid Manager *raidscan* command.

S-VOL—The backup or "mirror" secondary volume of a pair.

VDEV—An intermediate XP architectural level between LDEV and the PDEV (physical disk devices). Under External Storage XP, the VDEV level keeps track of the external storage array attributes & paths. For internal disks, the VDEV keeps track of the disk speed and type, etc.

RPO—Recovery Point Objective. A storage disaster recovery metric that indicates the tolerable amount of data and/or number of transactions that can be lost in the event of a catastrophe where a site-to-site failover has occurred.

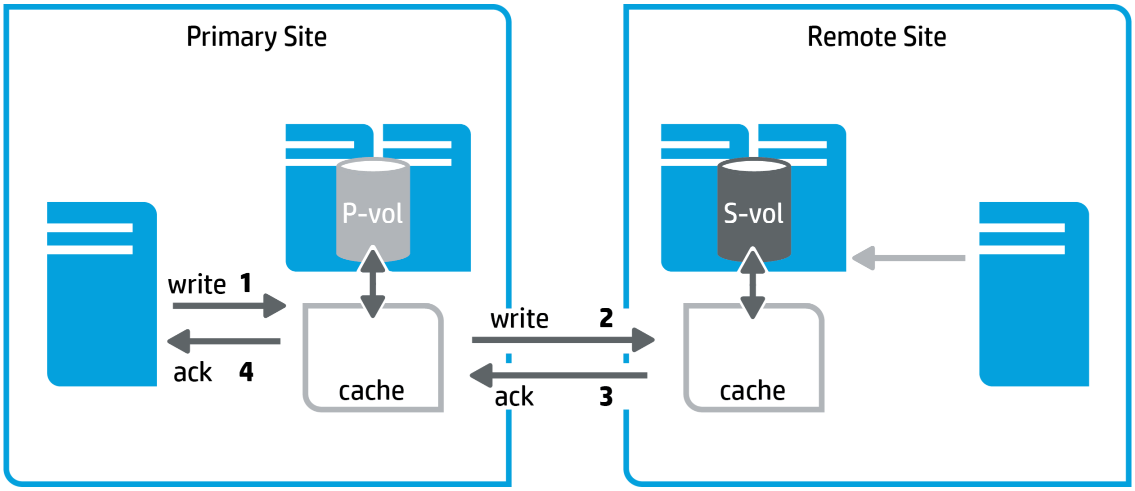
RTO—Recovery Time Objective. A storage disaster recovery metric that indicates the amount of time to successfully execute a site-to-site failover and restore business functionality.

Introduction to XP Continuous Access

This paper is intended to assist you in understanding XP Continuous Access, its setup and configuration, possible topologies and how it may be tuned for performance. It is intended as a supplement to the XP7, XP P9500 and XP24000/XP20000 and XP12000/XP10000 user guides for Continuous Access and assumes you are familiar with the concepts explained therein as well as the concepts explained in the *HP XP Raid Manager User's Guide*.

Brief Overview of XP Continuous Access Synchronous (a.k.a. Cnt Ac-S, or 'SYNC')

The diagram below illustrates the most notable characteristics of Cnt Ac-S. That is, a host write to the primary site (MCU) will not be completed (or ACK'd to the host) until the successful remote replication has been acknowledged by the remote site (RCU). Sync's strongest point is that the host can typically[1] be assured of the deterministic replication of every I/O to the remote site (assured currency). The weakest point of Cnt Ac-S is that the latency experienced by the host application is directly related to the electronic delay and the effective bandwidth between sites; making Sync performance intolerable over long distances or links with insufficient bandwidth. For some applications (e.g. highly reliable campus replication , etc.), Cnt Ac-S remains the best product for the job. For replication distances up to about ~30 km, Cnt Ac-S may also offer better throughput than Continuous Access Asynchronous, due to a shorter firmware code path.



Operation

1. Server writes data to primary array cache.
2. Primary Array writes data to Remote Array cache.
3. Remote Array acknowledges write to Primary Array.
4. I/O is acknowledged by Primary Array to server.

Key Solutions

- Seamless, reliable Campus/Metropolitan Disaster Recovery/Clustering

Weakest points of SYNC:

- Latency intolerant: data transmission latencies are summarily added to the response times experience by replicated Host I/O write operations.

Strong points of SYNC:

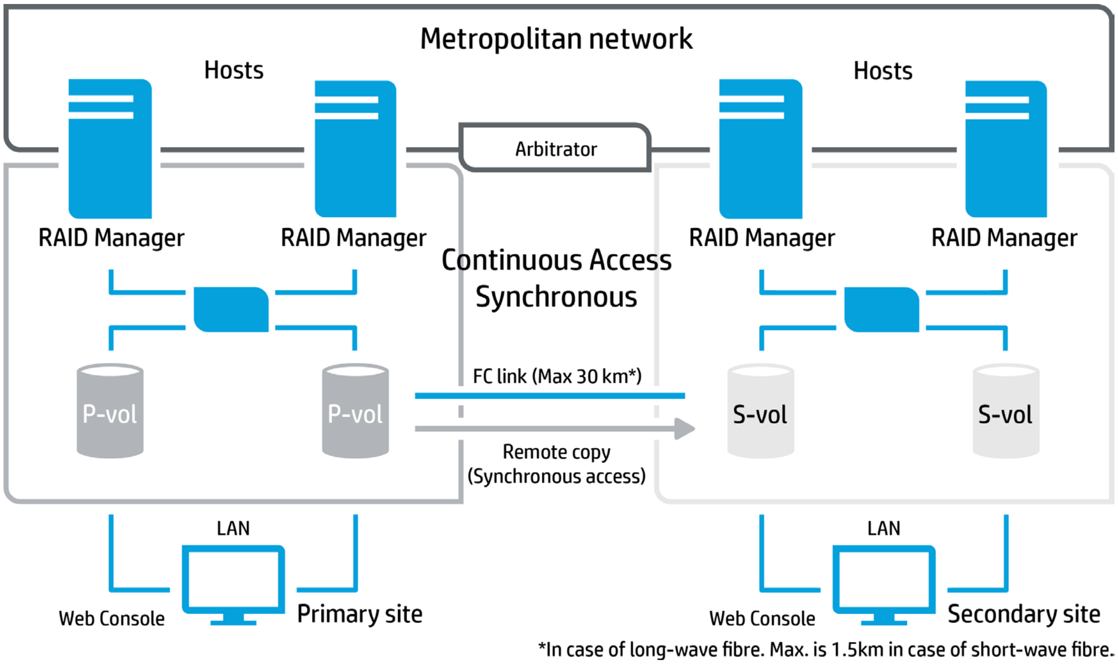
- Highest site-to-site replication concurrency available.

[1] There are low probability corner cases (e.g. SCSI abort), where the resulting data on both sides is not identical.

Use of Continuous Access in Clustered Environments

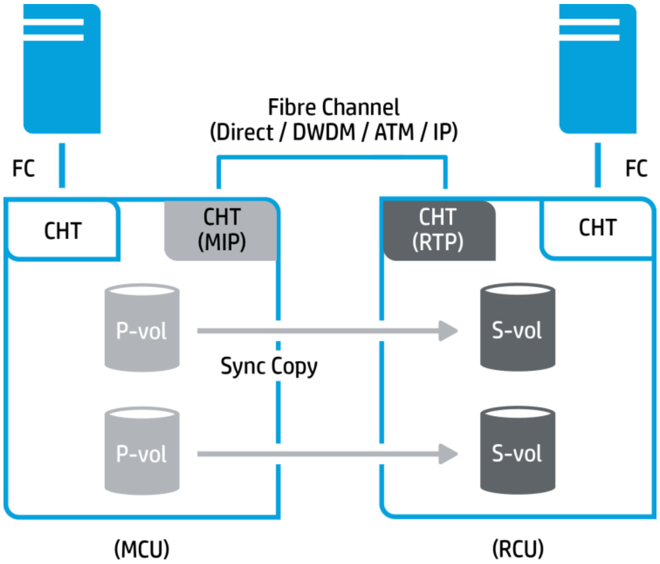
Use of XP Continuous Access Synchronous in clustered environments would allow:

- Disaster Recovery capability using Metrocluster or Cluster Extension (CLX) products.
- Maximum 30km (via long wave fibre) with cascade connection of two FC switches.
- Paired Volume Control via *RAID Manager XP or Replication Manager for HP Command View Advanced Edition*



Basic Interface between the MCU and RCU

The diagram below illustrates the basic connection paradigm used to connected the Main Control Unit (a.k.a. local, or “MCU”) and the Remote Control Unit (a.k.a. remote or “RCU”) arrays for the purpose of deploying XP Continuous Access. The interface for the array itself must be Fibre Channel; however one can use any of the popular communications mediums such as DWDM, ATM, IP, etc., if routers/extenders are deployed in addition to the XPs. This is valid for XP7, XP P9500 and XP24000/XP20000/XP12000/XP10000 arrays.

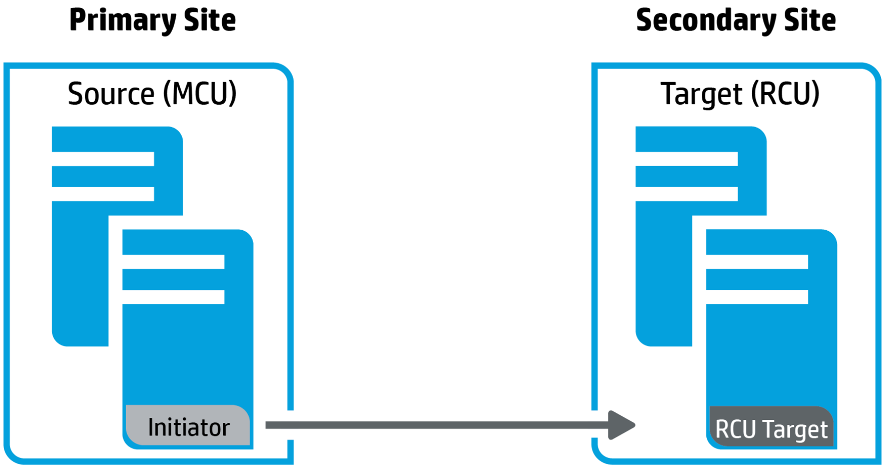


A Detailed Overview of XP Continuous Access Synchronous

XP Continuous Access Synchronous Operational "Push Model"

Sync uses a **PUSH** model to move data from the MCU to the RCU. This method **only** requires a "MCU initiator" to "RCU target" link (1 physical and 1 logical path) to enable replication in one direction. However for failover/failback support, a RCU to MCU path is also required, but will only be used if the application fails back from the secondary site to the primary site.

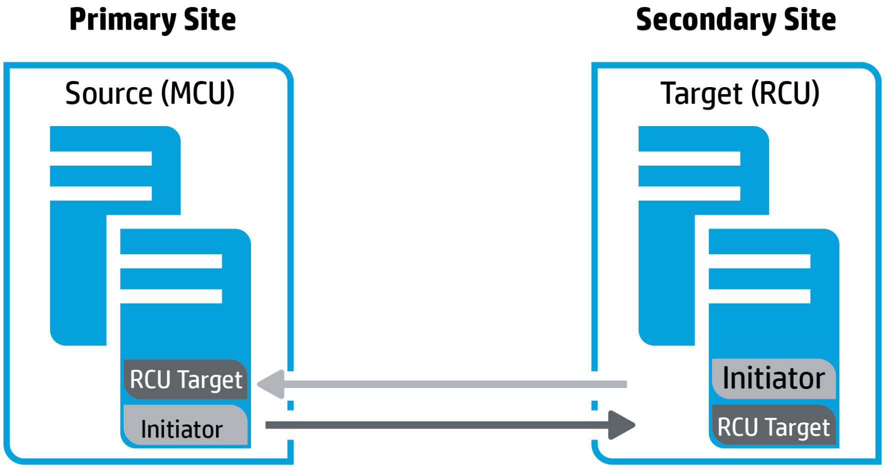
Single direction replication:



This type of physical connection limits replication options to only one source and one target array. A failover with a change of replication direction will not be possible in this configuration. However, this configuration is useful for data migration, or disaster recovery operations, without the option to replicate back to the production site.

For most failover, failback, and cluster integrated solutions, it is required to have another (push model, physical and logical) link in the opposite direction (from the secondary site to the primary site, see below). This will allow both arrays to act as both source (MCU) and target (RCU) Sync/Async arrays. In addition, the direction of replication can be changed by using the failover command.

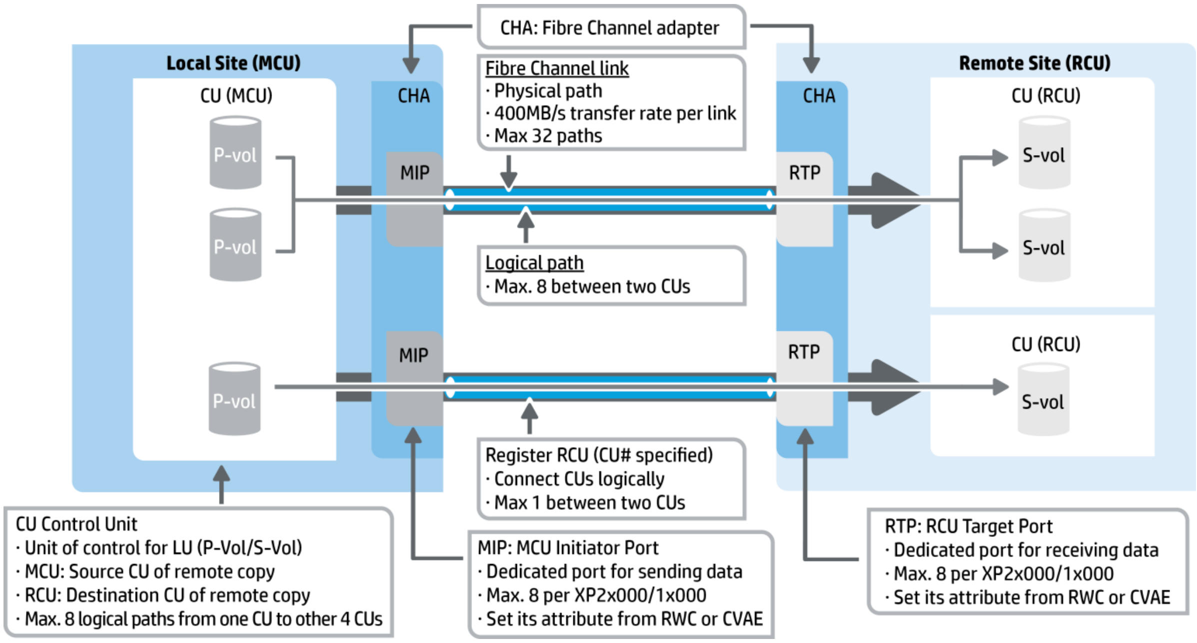
Bi-directional replication:



For most highly available solutions, it is recommended to have at least two links in each direction, each originating from different power clusters in the array. This will ensure that replication functions can continue without interruption during a single component failure or link maintenance. Multiple physical links may utilize the same intermediate link infrastructure. For even higher availability, the physical link should be routed through different link infrastructures, routed via different paths.

Using the "CU" Model to Configure Sync Replication

There are two models for defining XP Continuous Access Synchronous replicated storage. They are called "CU" and "CU Free". The older and more restrictive "CU" model is used when it is desired only to replicate LDEVs between the MCU and RCU that are in a specific control unit, or "CU" in the XP LDEV map. The figure below illustrates this paradigm.


CHA: Fibre Channel adapter
Local Site (MCU)
CU (MCU)
CHA
P-vol
P-vol
P-vol
MIP
MIP
Fibre Channel link
· Physical path
· 400MB/s transfer rate per link
· Max 32 paths
Logical path
· Max. 8 between two CUs
Register RCU (CU# specified)
· Connect CUs logically
· Max 1 between two CUs
Remote Site (RCU)
CHA
RTP
RTP
CU (RCU)
S-vol
S-vol
CU (RCU)
S-vol
CU Control Unit
· Unit of control for LU (P-Vol/S-Vol)
· MCU: Source CU of remote copy
· RCU: Destination CU of remote copy
· Max. 8 logical paths from one CU to other 4 CUs
MIP: MCU Initiator Port
· Dedicated port for sending data
· Max. 8 per XP2x000/1x000
· Set its attribute from RWC or CVAE
RTP: RCU Target Port
· Dedicated port for receiving data
· Max. 8 per XP2x000/1x000
· Set its attribute from RWC or CVAE

In addition, the following points apply to the use of the "CU" paradigm:

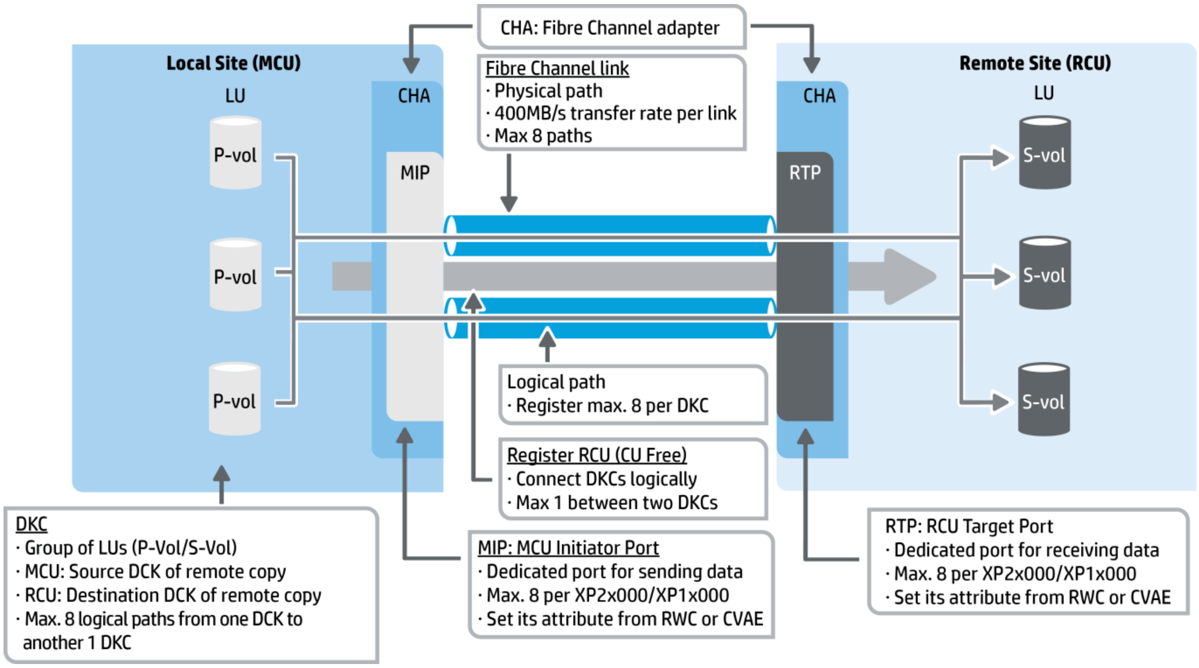
- Control of replications pairs can be done via Remote Web Console (RWC), Raid Manager XP, and Replication Manager.
- One LU can belong to only one Continuous Access pair.
- The ratio of P-VOL : S-VOL must be 1 : 1.
- Maximum 4 remote CUs can be registered for one local CU.
- The user can select S-VOL(s) from maximum 4 remote CUs.
- Up to 65536 Cnt. Ac-S volume pairs are allowed between two XP7s, or XP P9500s.
- Up to 32,768 Cnt. Ac-S volume pairs are allowed between two XP24000/XP20000s.
- Up to 8,192 Cnt. Ac-S volume pairs are allowed between two XP12000/XP10000s.
- Supported emulation types: Open-3/8/9/E/L/V.
- LUSE and Open Volume Management volumes are supported.
- Two LDEVs which compose a pair must be of the same emulation type and same capacity.
- Independent of RAID levels and HDD types (Different levels or types can coexist).

CU number	0	1	---	FD	FE
Max. number of CUs assigned to one logical path	4	4	---	4	4
Max. number of LDEVs	256	256	---	256	256
LDEV ID range	0.00 – 0.FF	1.00 – 1.FF	---	FD.00 – FD.FF	FE.00 – FE.FF
Max. number of SSIDs* (One SSID per 64 LDEVs)	4	4	---	4	4

* "1" is recommended although max. number of SSIDs is "4".

Using the "CU Free" to Configure Sync Replication

The newer and less-restrictive "CU Free" model is used when it is desired to replicate LDEVs between the MCU and RCU that may belong to any existing control unit, or "CU". The following figure illustrates this point. "CU Free" is available for XP7, XP P9500, and XP24000/XP20000/XP12000/XP10000 arrays.


CHA: Fibre Channel adapter
Local Site (MCU)
LU
CHA
P-vol
MIP
P-vol
P-vol
Fibre Channel link
· Physical path
· 400MB/s transfer rate per link
· Max 8 paths
Remote Site (RCU)
CHA
LU
RTP
S-vol
S-vol
S-vol
Logical path
· Register max. 8 per DKC
Register RCU (CU Free)
· Connect DKCs logically
· Max 1 between two DKCs
DKC
· Group of LUs (P-Vol/S-Vol)
· MCU: Source DCK of remote copy
· RCU: Destination DCK of remote copy
· Max. 8 logical paths from one DCK to another 1 DKC
MIP: MCU Initiator Port
· Dedicated port for sending data
· Max. 8 per XP2x000/XP1x000
· Set its attribute from RWC or CVAE
RTP: RCU Target Port
· Dedicated port for receiving data
· Max. 8 per XP2x000/XP1x000
· Set its attribute from RWC or CVAE

In addition, the following points apply to the use of "CU Free" model:

- LUs can be paired regardless of whether they are on the same CU or not
- Volume pair control via RWC, Raid Manager XP, or Replication Manager.
- One LU can belong to only one Sync pair.
- P-VOL : S-VOL = 1 : 1
- Up to 65536 pairs Sync volume pairs are allowed between two XP7s or XP P9500s.
- Up to 32768 Sync volume pairs are allowed between two XP24000s.
- Supported emulation types: Open-3/8/9/E/L/V.
- LUSE and Open Volume Management volumes are supported.
- Two LDEVs which compose a Sync volume pair must be of the same emulation type and same capacity.
- Independent of RAID levels and HDD types. (Different levels or types can coexist.)
- Can configure an LUSE by combining LDEVs on different (CUs).
- Can configure an LUSE by combining LDEVs that have different sizes (capacities).

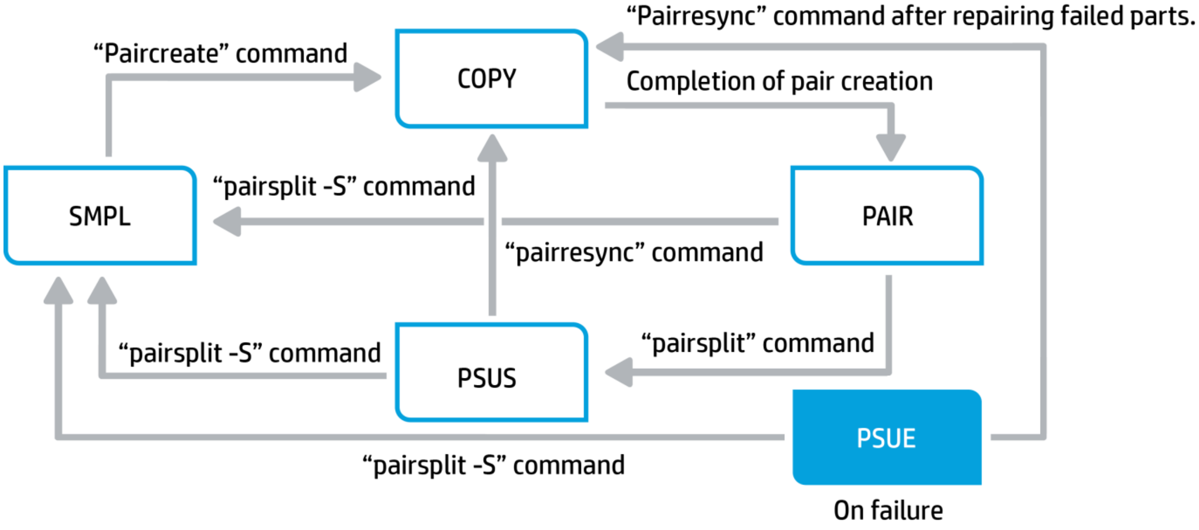
(The P-VOL must have the same capacity as the S-VOL. The P-VOL and the S-VOL must contain the same number of LDEVs.)

XP Continuous Access Synchronous Pair States

The diagram and description below communicates the possible states XP Continuous Access Synchronous pairs can be placed in either by deliberate commands or by XP array events, such as faults or errors that affect the state of the replication.

Possible pair states:

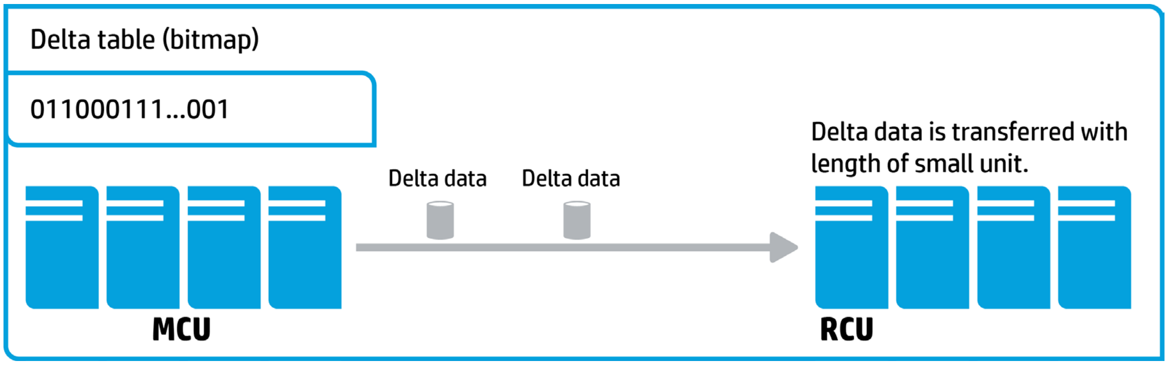
- SMPL: This volume does not belong to any Sync volume pair.
- COPY: This volume is changing to the PAIR state.
- PAIR: This volume belongs to a Sync volume pair.
- PSUS: By issuing “pairsplit” command, remote copy for this Sync volume pair is suspended. In this state, changes to the P-VOL are NOT copied to the S-VOL, but are kept track of at the logical track level in a bitmap defined in the MCU XP’s shared memory (covered later).
- Updated data of P-VOL and S-VOL are managed by delta table (bitmap).
- PSUE: Remote copy for this Sync volume pair was suspended due to a failure. Only updated data of P-VOL is managed by delta table (bitmap).



It should be noted that, unlike HP XP Business Copy pairs, which are intended to remain in a suspended, or “PSUS” state the vast majority of the time, Sync pairs are intended to remain paired, or in “PAIR” state the vast majority of the time and should only be in the other states either during initialization, in error, or recovering from a bitmap state.

Change Tracking in Suspended State via Bitmaps in Shared Memory

When either XP Business Copy, or Continuous Access replication pairs are in a suspended, or “PSUS” state, changes resulting from Host I/O write to the pair P-VOLs or S-VOLs are isolated, but are tracked and maintained in bitmaps located in the shared memory of the MCU XP. These bitmaps are used to track changes at the logical “track” level by default (other choice is by logical “Cylinder”). When the pair is resynchronized, these P-VOL and S-VOL bitmaps are merged and used to update the S-VOL. The diagram below illustrates this process:



Using Consistency Groups (CTGs) with Sync to Ensure Data Consistency

To ensure data consistency, two or more XP Continuous Access Synchronous pairs may be placed in their own consistency group (also referred to as a "CTG").

IMPORTANT NOTE CONCERNING CTGs:

If a CTG is used, the pairs in a CTG may be created, split, and resynchronized as a group, meaning that the array will hold any future pending commands issued for a given CTG until the current command for that CTG has completed for all its pairs. Although it may seem with some configurations that a given command may be in fact be executed atomically (applied at the exact-same time for all the pairs in a CTG), it is important to understand that this is not the case.

For the XP7, XP P9500 AND XP24000/XP20000/XP12000/XP10000 arrays, up to 256 CTGs may be defined. CTGs may be used with the "CU" or "CU Free" models discussed previously. CTGs may be utilized via the RWC, Raid Manager XP, or Replication Manager.

Continuous Access Synchronous for the XP7

Overview

The XP7 is the most recent generation of XPs and has SYNC available at first release. It will be released at V01 with ALL of the features of XP P9500 V05. The XP7 contains several enhancements, such as a new Remote Web Console that has all of the replication controls implemented in flash (as opposed to having it split between flash and Java as is done with the XP P9500).

General Specification

The table below gives the overall specifications for the XP7 by explaining the changes, or 'deltas' between the XP7 and XP P9500. Significant changes are noted in ***Italic***.

Table. XP7 Specifications

No.	Item	XP P9500	XP7 (V01)
1	Required P.P.	OPEN: Continuous Access MF: Continuous Access for Mainframe	←
2	Capacity-based charging	Applied	←
3	DKC combinations	MCU/RCU support connection with XP P9500, XP24000, and XP12000.	***MCU/RCU support connection with XP7 XP P9500, and XP24000.***
4	Sync/Async	Only Sync is supported (Async is not supported)	←
5	Physical drives	No restriction (Comply with product spec)	←
6	Maximum number of pairs	32k pairs	***64k pairs***
7	Supported LDEV#	0 to 65279 (0x0000 to 0xfeff)	←
8	Supported capacity	Up to 1918TB	← ***(With V02, capacity that can be defined by Thin Provisioning will be supported in Continuous Access, and capacity that can be defined in MF will be supported in Continuous Access for Mainframe)***
9	SM area	Need to install SM for CA/CA-MF	←
10	Supported RAID levels	RAID1/RAID5/RAID6	←

No.	Item	XP P9500	XP7 (V01)
11	Supported emulation types	OPEN-3/8/9/K/E/L/V 3380-J/-E/-K/-F 3390-1/-2/-3/-3R/-9/-L/-A/-M	***OPEN-3/8/9/K/E/L/V*** ***3390-1/-2/-3/-9/-L/-A/-M***
12	LDEV size	OPEN: 4TB MF: 223GB	***OPEN: 4TB*** ***MF: 223GB*** ***(With V02, capacity that can be defined by Thin Provisioning will be supported in Continuous Access, and capacity that can be defined in MF will be supported in Continuous Access for Mainframe)***
13	CVS setting	Supported	←
14	LUSE setting	Supported	***Not supported*** ***(Pair creation with LUSE of older models is not supported)***
15	Copy from small P-VOL→ large S-VOL	CA: Not supported CA-MF: Supported	←
16	Connection between DKCs	Fibre direct connection (including SW connection) Fibre+DWDM connection Fibre+Extender connection	←
17	Ports between DKCs	Initiator → RCU Target	←
18	SW/DWDM/Extender connection	Complies w/ XP P9500 product spec.	***Complies w/ XP7 product spec.***
19	Number of physical paths per DKC (Port:Port)	4096	←
20	Type of paths between DKCs	CU path/DKC path	***In OPEN, only DKC path is supported (CU path is not supported). In MF, only CU path is supported (Same as XP P9500)***
21	Number of CU paths registered	1020 (Local CU(255)×4)	←
22	Number of DKC paths registered	64	←
23	Number of connected CU paths/DKC paths	2 to 8 paths	←
24	Number of CTGs	128	***256***
25	CTG type	MF-CTG, OPEN/MF CTG, CTG between multiple DKCs	←
26	Number of pairs in CTG	8192	←
27	Number of DKCs per CTG	4 DKCs (Configuration with more than 4 DKCs is supported when necessary)	←
28	Mixing models in CTG	XP P9500/XP7 can be mixed	←
29	Mixing OPEN/MF in CTG	Supported	←

No.	Item	XP P9500	XP7 (V01)
30	DF connection	Not supported	←
31	PP combinations	Complies w/ XP P9500 product spec.	***Complies w/ XP7 product spec.***
32	Difference management method	SM method	***V01: SM method*** ***V02: SM method/hierarchical memory method***
33	Unit of difference management	Track/Cylinder	***Track only (Cylinder is not supported)***
34	Effective range of pair options (Priority only)	System (per DKC)	←
35	Effective range of system options	System (per DKC) or MP	←

Sync for the XP P9500 Disk Array

Overview

There were many architectural changes that distinguish the XP P9500 from the XP24000, however there were very few changes that one needs to be concerned about for Continuous Access Synchronous and for replication in general.

Array Port Dedications

As with the XP24000 and earlier, the XP P9500 still requires you to assign at a minimum one (1) Initiator and one (1) RCU-Target port on the MCU and RCU arrays respectively. One very-positive change is that since the XP P9500 array microprocessors are not dedicated to any one port, assigning either of the above port attributes does NOT cost the customer two array ports like it does with the XP24000 and earlier arrays.

LDEV Ownership Concerns

Unlike earlier XPs, the XP P9500 maintains LDEV ownership at the Microprocessor Blade (MPB) level. It is possible to span a single Continuous Access Synchronous consistency group (CTG) across multiple MPBs. Take caution in doing this so to avoid overloading any one single MPB.

Notes on Configuring Shared Memory

For the XP P9500, the shared memory element is a mirrored and partitioned segment of cache; whose size is defined by the storage administrator. The amount of shared memory configured is of paramount importance for most of the XP P9500 program products (Continuous Access included) as failure to provide shared memory will restrict its ability to use its program products (PPs). The table below gives the amounts of mirrored shared memory on a per-PP combination basis.

Table 1. XP P9500 Mirrored Shared Memory Amounts per Program Product Combination

	Judgment Factor of SM Capacity (*3)									
NO.	Number of CU (Configuration of LDEV)	Program Product (*2)				SI/VM Extension		CoW Extension	TC/UI Extension	SM Capacity
		SI/VM	FCv2/DP/ CoW/TPF	TC/UR	HDT	1	2			
1	1-64 (16KLDEV)	0	X	X	X	X	X	X	X	8 GB
2	1-64 (16KLDEV)	0	X	0	X	X	X	X	X	16 GB
3	1-64 (16KLDEV)	0	X	0	X	X	X	X	0	24 GB
4	1-255 (64KLDEV)	0	0	X	X	0	X	X	X	16 GB
5	1-255 (64KLDEV)	0	0	X	X	X	0	0	X	24 GB
6	1-255 (64KLDEV)	0	0	0	X	0	X	X	X	24 GB
7	1-255 (64KLDEV)	0	0	X	0	0	X	X	X	24 GB
8	1-255 (64KLDEV)	0	0	X	0	X	0	0	X	32 GB
9	1-255 (64KLDEV)	0	0	0	X	X	0	0	X	32 GB
10	1-255 (64KLDEV)	0	0	0	X	0	X	X	0	32 GB
11	1-255 (64KLDEV)	0	0	0	0	0	X	0	X	32 GB
12	1-255 (64KLDEV)	0	0	0	0	X	0	0	X	40 GB
13	1-255 (64KLDEV)	0	0	0	X	X	0	X	0	40 GB
14	1-255 (64KLDEV)	0	0	0	0	0	X	X	0	40 GB
15	1-255 (64KLDEV)	0	0	0	0	X	0	0	0	40 GB

*1:	TrueCopy:	TC	ShadowImage:	SI
	Flash Copy Version2:	FCV2	Universal Replicator:	UR
	Dynamic Provisioning:	DP	Volume Migration (*2):	VM
	CoW Snapshot	CoW	Hitachi Dynamic Tiering:	HDT

For the above table:

UR = Continuous Access Journal, TC = Continuous Access Synchronous,

SI = Business Copy, DP = Thin Provisioning, CoW = SnapShot, VM = AutoLUN/TSM,

HDT = SMART Tiering.

Please note that these values are for the mirrored amounts of shared memory. The actual amount dedicated per-power cluster is exactly half of what is listed above.

For more information, please contact HP.

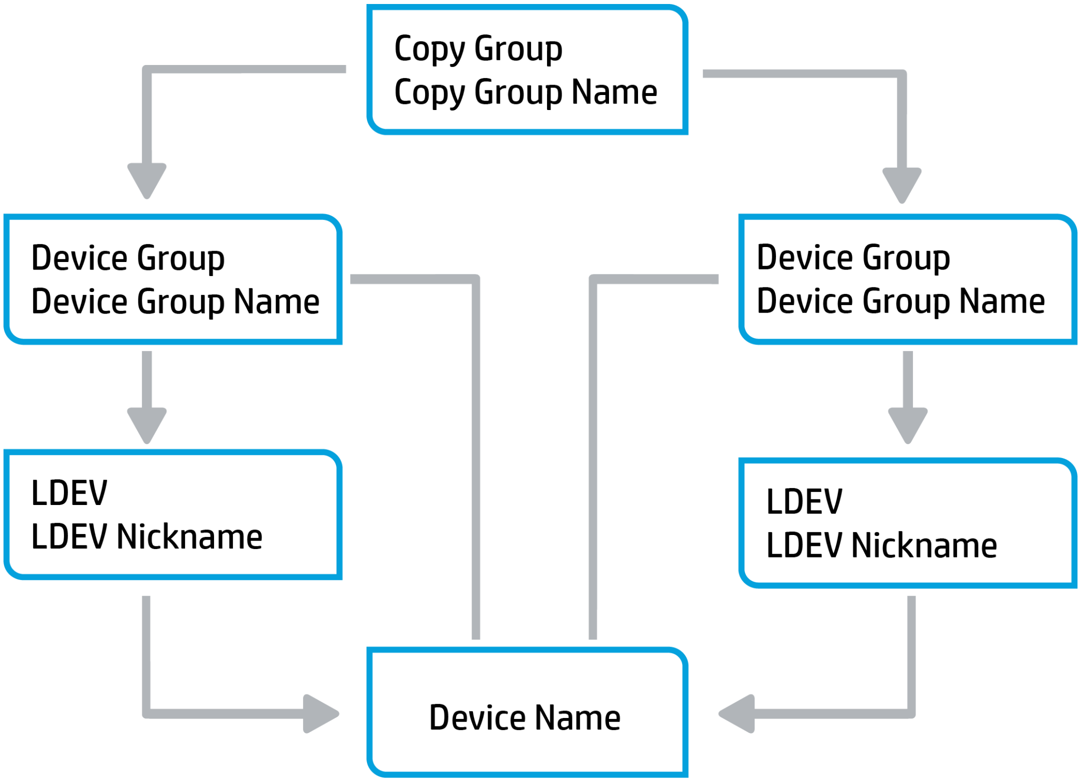
Notes on the Status of the XP P9500 Remote Web Console (RWC)

Currently, most of the array resource provisioning and event monitoring is found in the flash-based portion of the RWC, while items such as replication control and licensing are still in a java-based application launched from flash-based portion of the RWC. The java-based application has exactly the same look and feel of that of the XP24000; so there will be almost no learning curve initially. The ultimate goal, however, is for the RWC to be completely flash-based; improving both usability and responsiveness seen by the user.

Notes on Using Device Groups and Copy Groups via Raid Manager CLI with XP P9500 V02 or Later

With the availability of firmware 60-02-xx-xx/xx or simply "V02" for the XP P9500 comes the ability to take advantage of a new way to organize and manage replication pairs (Continuous Access and Business Copy) via the use device groups and copy groups. The basic paradigm for using device groups and copy groups is that they may be used to establish a hierarchy of replicated device management. The figure below illustrates this hierarchy:

Basic Device Group and Copy Group Hierarchy:



Under this new paradigm, individual LDEVs are grouped into device groups (similar to what is already done in Raid Manager) and these device groups are in turn joined in a pair relationship known as a copy group. Under this paradigm, one Copy Group is equivalent to a consistency group, or "CTG".

The primary value-add of this new paradigm is that it is now longer necessary to created detailed and error-prone horcm files. Instead, all LDEV, device group, and copy group data and status is now stored on the XP P9500 DKC and can be created and manipulated via the Raid Manager CLI 'raidcom' command.

For more information of using device groups and copy groups with XP P9500, see the Raid Manager User's Guide.

Notes on 'Virtual' Command Capability with XP P9500/Raid Manager CLI

With the XP P9500, it is now possible to configure a raid manager instance to connect to the XP P9500 SVP out-of-band (via UDP/IP), thus establishing a 'Virtual Command Device' that functions that same as those which are in-band. The format for specifying this via a horcm file is:

\\.\IPCMD-IPaddr-Port[-unitid] or IPCMD-IPaddr-Port[-unitid]

- Port: port number for HORCM
- unitid: this must be specified for Multiple DKCs

See the Raid Manager User's Guide for more information.

Notes on the "Performance Accelerator" Program Product available with XP P9500 V04a Firmware

The new "Performance Accelerator" Program Product available with V04a has the following important features:

- Improves overall processing of replicated and un-replicated workloads through several firmware changes specifically designed to increase I/O processing efficiency.
- Further optimizes the already-efficient XP P9500 array cache usage for workloads and working sets that are heavily random as well as cache-avoidant and resident on SAS solid-state and spinning media.
- Yields significant performance improvements as much as 2X (not typical) over the current un-replicated and replicated workload processing for workloads and working sets described above.
- Does NOT further optimize for segmented or sequential workloads or working sets that are significantly cache-resident.
- Does NOT optimize for workloads on external storage working sets.
- Does NOT require the dismantling of replication pair relationships prior to installation or de-installation.
- Offered to our customers in firmware "V04a" in both a 30-day trial as well as a permanently assigned frame product license key labeled "Performance Accelerator".

The two major performance improvements made to the firmware that yields most the above are:

1. More efficient cache handling between the two processes that handle I/O; causing them to process data more quickly.
2. A targeted improvement for SAS (not SATA) backend disks hosting working sets for largely cache-avoidant and random workloads. This improvement will be most evident on SSD-resident working sets/workloads, but should yield some improvement for SAS spinning media.

While the changes are not specifically meant to improve replication processing or replicated performance, there will be some benefit that can be derived if the customer is running working sets and workloads like those described above.

Configuring XP Continuous Access Synchronous on the XP7, XP P9500 and XP24000/XP20000/XP12000/XP10000 Disk Arrays

The section is intended to be a supplement to the *HP Continuous Access User's Guide* and is only meant to expose and communicate best practices and nuances that are not covered in depth in any other source.

The basic pre-requisites to successfully configuring Sync on the XP P9500 and any XP are:

- XP Continuous Access Synchronous software licenses installed on the MCU and RCU arrays (see user's guide).
- Correctly sized and "spec'd" Telcom-provided communications link (discussed later) between the MCU and RCU array sites. This has direct bearing on the replication capabilities you can expect to use in the present and future tense.
- Supported Fibre Channel switches and/or routers to facilitate communication between sites. See HP's online *SAN Design Guide*.
- Standard IPv4 of IPv6 network between sites if it is desired to use Raid Manager XP to manage replication operations.
- At least one port on the MCU and RCU dedicated to Sync inter-array communications. Note that dedicating a port on the XP24000/XP20000 arrays also implies the dedication of its associated microprocessor (MP). This is NOT the case with the XP P9500. Note that if the arrays are XP24000/XP20000s only allowing for one MP on each array for inter-array replication will severely limit possible replicated storage performance (also discussed later).

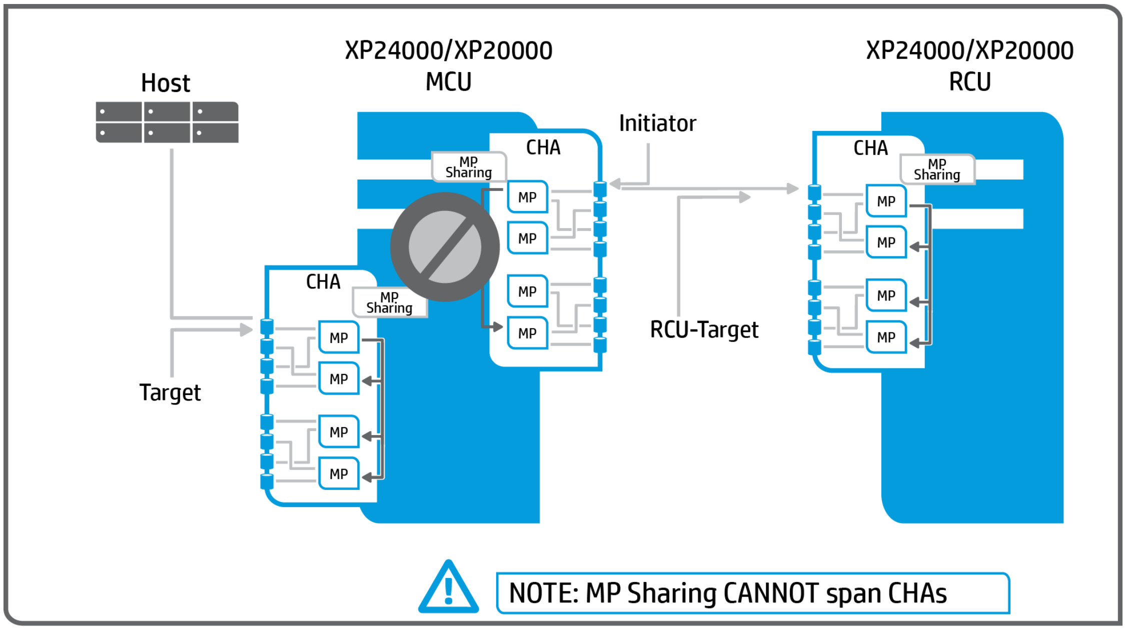
Once these pre-requisites are met, Sync may be installed and pairs may then be created. Optionally, cluster applications such MSCS (with Cluster Extension XP), and MetroCluster may be modified to failover storage and services between the local and remote sites while operating CntAc-S in the process.

XP24000/XP20000 Microprocessor Load Distribution (a.k.a. "MP Sharing")

For the XP24000/XP20000 disk arrays, it is possible to configure the array so that the write component of Host I/O and replication I/O workload may be distributed across the MPs of a single CHA provided the following conditions are met:

- Only applies to write I/O (reads are NOT distributed).
- Other MPs must be < 50% utilized.
- Other MPs must be set to same attribute (Initiator, etc.)
- MP Sharing CANNOT span CHAs.

The following diagram illustrates the possible MP positions that are used for XP24000/XP20000 Continuous Access Synchronous.



Note that the two possible positions for MP sharing are MPs dedicated for Host I/O (Target) on the MCU and the RCU "RCU-Target" MPs/ports.

Sync: Designing for Performance

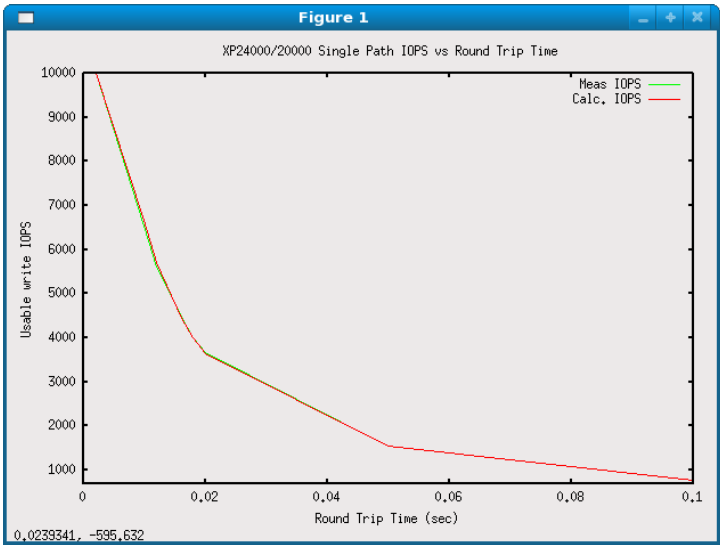
Inter-Array Communications Link Concerns

There are three important aspects of the communications link that can either permit or prohibit Sync replication for a given configuration and set of storage performance requirements. They are:

- Link Bandwidth: This is the number of concurrent bits that can be sent down the link. It is typically measured in Megabits per Second, or “Mbit”
- Link Latency or “Round Trip Time”: This is the time in milliseconds that it takes for a single bit to travel from the MCU to the RCU array PLUS the time for an acknowledgement to be returned from the RCU back to the MCU.
- Error Rate: This is a measure of how many transmissions in a given time period will fail due to aspect of communication such as line noise, old or malfunctioning equipment, etc. This is typically measured as a percent of packets sent (%). This measurement is most relevant when considering the use of IP LANs or WANs for use in replication.

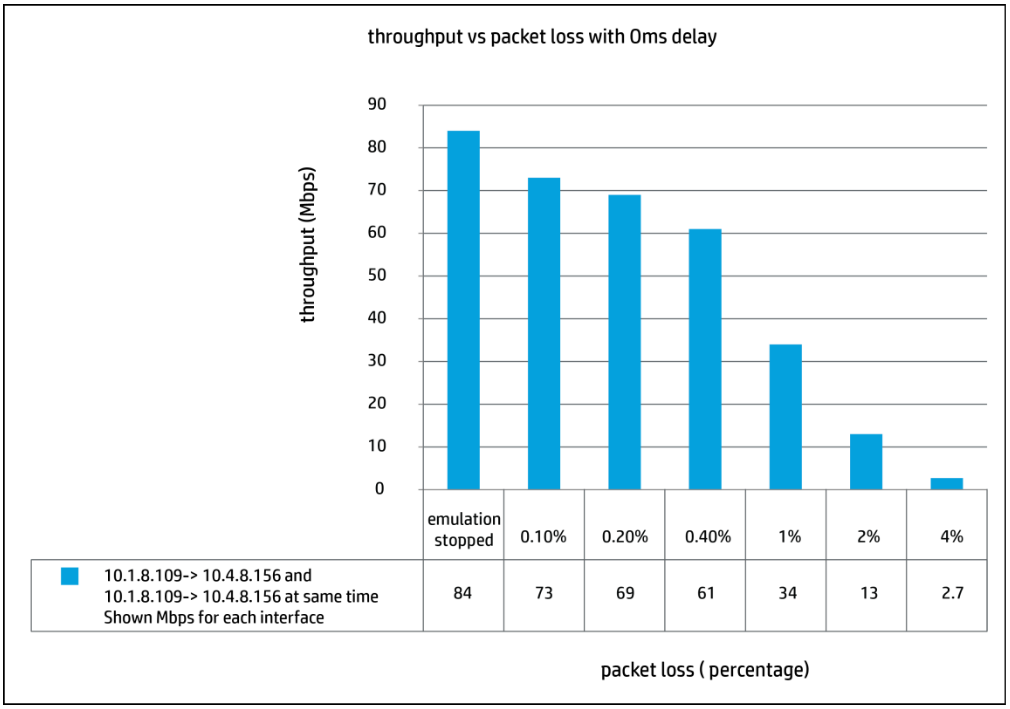
When considering the amount of bandwidth to specify for a given configuration, one must remember that there is an approximate 20-25% overhead typically associated with the transmission of replicated storage I/O from one array to the other. This overhead typically manifests itself in the transmission of checksums and other housekeeping data necessary for FC and any other protocol layers (such as IP) that must be used.

The link latency, which is typically measured as a “Round Trip Time” in milliseconds (msec) is a very important factor when designing a Sync replication configuration because it must be effectively doubled and added to the expected response time of replicated Host I/O. Another side effect of this latency is that, as it increases it begins to affect the total possible IOPs that can be replicated by the dedicated MPs/ports (designated Initiator and RCU-Target) on the MCU and the RCU. The graph in the illustration below illustrates this point.



As the figure clearly shows, the total usable write IOPs that can be replicated by a link drops to less than 10% when the round trip time is more than 80msec.

The link error rate is also a major factor that can negatively affect replication performance. As the diagram below indicates, even a link error rate of 0.4% can reduce the total usable bandwidth to only 60% of its design capability.



	emulation stopped	0.10%	0.20%	0.40%	1%	2%	4%
🟦 10.1.8.109-> 10.4.8.156 and 10.1.8.109-> 10.4.8.156 at same time Shown Mbps for each interface	84	73	69	61	34	13	2.7

This aspect of link performance tends to be of greater concern when an IP link (as opposed to dark fibre or other photon-based communications) is used as it has the greatest potential to be a non-zero value. As a rule, it is best to ensure that the link error rate for any link is as close to zero as possible.

In general, the best practice to remember is to make the best effort possible to understand the nature of the communications links available.

XP7 Sync Performance Design

HP is still in the process of determining configuration best-practices and rules of thumb for the XP7. However, given that the XP7 does not represent a fundamental shift in either architecture or SYNC-related firmware capabilities, it is recommended that you follow the guidelines below for the XP P9500 until this information can be updated. HP will be publishing a new version of the XPCAPET for the XP7 which can be used for sizing exercises in the new future.

XP P9500 Sync Performance Design

HP has published version 7.0.1 of the XPCAPET tool which is capable to sizing both SYNC and JOURNAL configurations based on inputs or both an academic nature as well as directly from an exported PA database as well as from the host-based utility known as XP P9000Watch (csv file).

XP24000/XP20000 Sync Performance Design

One of the primary concerns when designing an XP Continuous Access Synchronous installation is what the performance of the replicated storage will be after replication has been started (pairs are in “PAIR”, or duplex state). To assist in the design, the XP Continuous Access Performance Estimator Tool (XP CAPET) was developed. This tool will assist you in sizing both XP Continuous Access Synchronous and Journal implementation for current and existing customers.

The primary limiting factor for Sync replication performance is generally the number of I/O operations per second, or “IOPS” that a particular Initiator port can transmit. If sufficient MPs are deployed in the two areas described in the previous section (using the XP CAPET as a guide), it is theoretically possible that given sufficient bandwidth, link quality, and XP resources, one would be able to recover any potentially lost performance (within the limits of link latency) due to the instigation of Sync (or Journal) replication on a given set of array LDEVs.

Suggestions for XP12000/XP10000 Performance Design

As the XP12000/XP10000 is not capable of sharing it's workloads across MPs, there is a significant difference in performance between the XP12000/XP10000 generation and the current generation. Thus, it is preferred to have any potential customer who wants to conduct XP Continuous Access Synchronous to upgrade to the XP24000/XP20000 generation. However, if it is not possible, the following guideline should be applied:

- Design the configuration such that each configured outbound MCU Initiator Port and inbound RCU RCU-Target Port is replicating no more than 2,200 write IOPS. If more than 2,200 write IOPs is required, dedicate additional initiator ports and the same number of RCU-Target Ports.

Following the above guidelines should yield stable configurations for SMALL BLOCK RANDOM WORKLOADS given sufficient MP, ECC group, and link resources. If more aggressive performance configuration information is desired, or a different workload is needed, please contact the factory.

Important System and Host Options Modes for SYNC

This section is intended to communicate information on both System Option Modes (SOMs) and Host Option Modes (HMOs) designed to vary the functionality of SYNC to better-address the needs of individual customers.

HMO 0x51: Reduction of Round Trip Time during SYNC Replication, or 'FastWrite' Implementation

For the XP7 and XP P9500, enabling this mode via the RWC causes the MCU change the state machine it uses for sending data via SYNC. Specifically, it causes the MCU stop waiting for the 1st acknowledgement normally expected after the MCU->RCU write command has been sent. Instead, the sequence of communication events will follow this paradigm:

1. MCU Sends Write Command to RCU.
2. MCU Sends SYNC data to RCU.
3. RCU acknowledgement to MCU.

This implementation can be thought of as an implementation of a 'FastWrite' paradigm. This mode should be enabled on the initiator ports for both the MCU and RCU. See HMO 0x65 below if you have the maximum-configured number of MP Blades.

HMO 0x65: Extension of HMO 0x51 for XP P9500 and XP7 Maximum MP Blade Count

Basically, if you enable HMO 0x51 and you have a maximally configured number of MP Blades, you should enable this mode as well.

SOM 732: Prioritization of SYNC Replication I/O De-Stage

For XP7/XP P9500/XP24000/XP20000. Increases the cache de-stage priority of replication I/O (RIO) on arrays with write-centric workloads. If set to ON, the prioritization occurs when the arrays write-pending cache reaches 10%. If left OFF, the prioritization occurs when the write-pending reaches 60%. Should be set on both MCU and RCU.

Use of “Min. Copy” to Shorten Lengthy Initial Copy Operations

Overview

Beginning with the XP24000/XP20000 V05 firmware release, it is now possible for a customer to shorten lengthy initial copy operations that must occur once sync is configured. This is particularly useful if the customer has links that are low-bandwidth and/or have a high latency; potentially affecting the performance of the XP’s replicated storage for longer periods of time depending on amount of data to be replicated.

Basic Paradigm

Min. Copy works by allowing the user to complete the majority of the sync initial copy task by doing the following steps:

1. Make a raw vol. backup of the proposed Sync P-VOL(s) for safety purposes.
2. Set Function Switch 32 via the RWC sync “Optional Operation” window on both the MCU and RCU.
3. Create a Business Copy within the MCU where the BC P-VOL (s) are the future sync P-VOLs.
4. Create the sync pairs with the “nocopy” option, thus skipping the initial copy and transitioning the pairs immediately to PAIR state.
5. Suspend the sync pairs.
6. Suspend the BC pairs and make a backup copy either to tape or a host-resident image.
7. Transport the raw vol. P-VOL backup (via tape or internet) to the remote site.
8. Restore the backup on to the sync S-VOL at the remote site.
9. Re-sync the sync pairs.

Once the “re-sync” command is issued, sync reaches PAIR state by only transmitting the delta of what changed in during the time they were in a suspended (PSUS) state. Note that once you re-synchronize the pairs in question that Function Switch 32 will automatically reset to OFF.

Other considerations

Customers that are considering using Min. Copy to establish their sync pairs are strongly encouraged to checksum their S-VOL images prior to transfer and again following their being restored to new sync S-VOLs just before the Continuous Access “resync” command is issues. Any variance in the checksum calculation during the transport or after the restore likely indicates that the data is corrupted. The MD5 checksum has 128 bits of protection and is recommended for this operation.

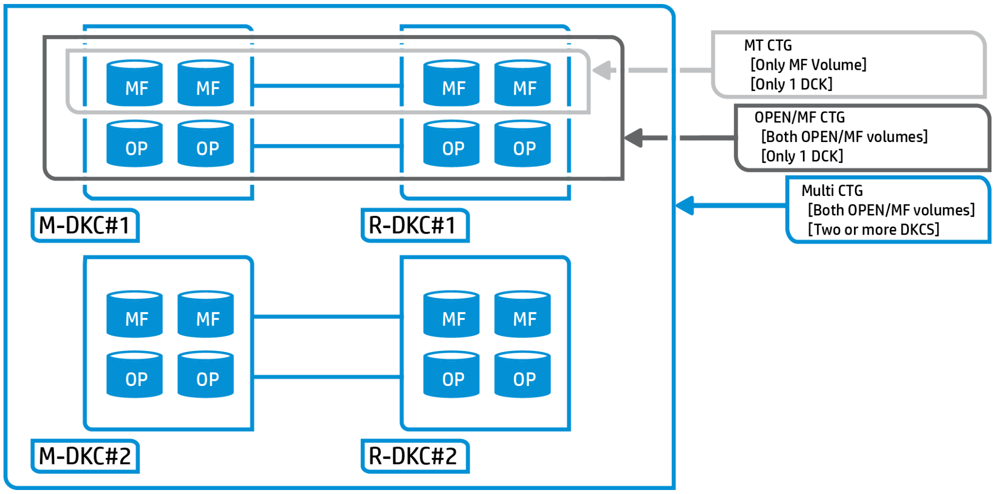
Use of Consistency Groups with SYNC

Unlike JOURNAL (Continuous Access Journal), SYNC does NOT by-default implement a consistency group when the paircreate operation is executed against a set of P-VOLs/S-VOLs. If it is desired to implement a CTG, it may be either selected in the RWC at the time of pair creation, or specified using the ‘-fg [CTG]’ argument within the Raid Manager paircreate command.

Implementing OPEN MxN using SYNC

OPEN MxN is the method by which a single Consistency Group (CTG) can be expanded or ‘stretched’ across multiple MCUs and RCUs (see figure below). Though originally designed using Continuous Access Journal, OPEN MxN is also available using SYNC as its underlying replication transport model. Additionally, CTGs can be shared between OPEN systems and Mainframe systems using this paradigm.

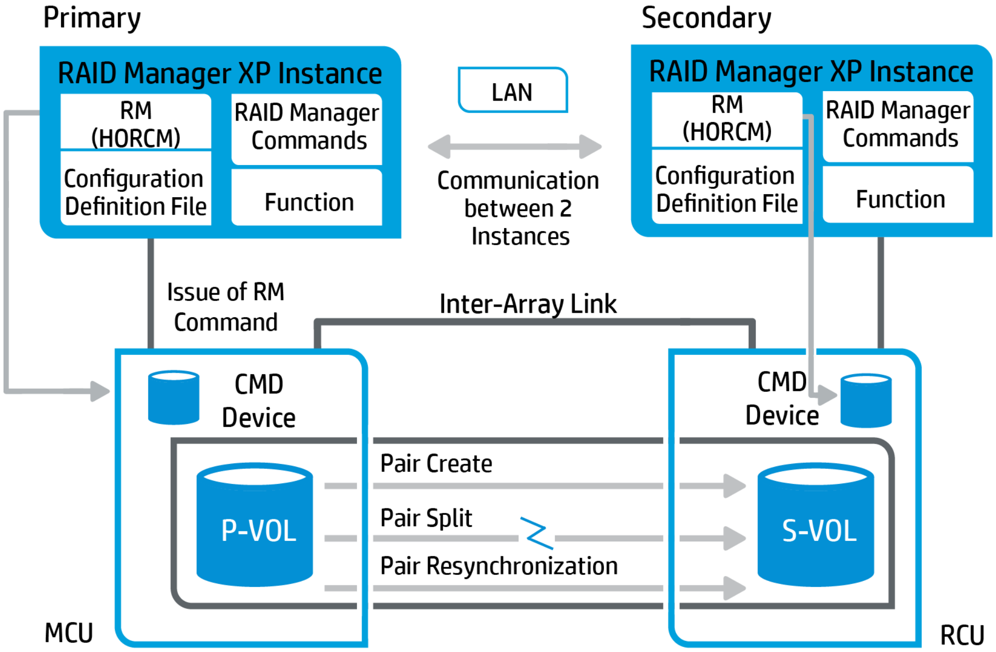
Fig. OPEN MxN using SYNC



For more information on implement OPEN MxN, see the Continuous Access User Guide as well as the Raid Manager User Guide.

Using Raid Manager: A Brief Overview

The diagram below illustrates the basic components and operations of Raid Manager XP.



Raid Manager XP (a.k.a. RMXP) is the typical choice for customers who desire to manage their XP replication from a host-based CLI. For the purpose of managing XP Continuous Access on two sites, LDEVs that have been designated “Command Devices” via the XP’s Remote Web Console, or other capable software tool are made visible to at least one host connected to the MCU and one host connected to the RCU. On each host, at least one Raid Manager XP service, or “instance” is started. These instances each correspond to a host-resident configuration file whose name is in the “horcmXXX.conf” format, where “XXX” corresponds to the instance number being used. Once the instances have been started (typically by script upon host start-up) and the correct shell variables have been set, RMXP may be used to display the status of, create, split, and resynchronize XP Continuous Access pairs. For more information on the use of RMXP to manage XP Continuous Access replication, see the *HP Raid Manager XP User’s Guide* located on the public HP storage website hp.com/go/storage. Raid Manager XP does NOT require a separate license to function with the XP24000/XP20000/XP12000/XP10000 disk arrays, however, it does require the arrays in question have valid replication licenses (Continuous Access, etc.).

Answers to Common Questions (Q&A)

This section is a compilation of XP Continuous Access Synchronous–related questions and their answers. Please contact the factory if you are not able to find your question here.

Q: Why is there no coverage of XP Continuous Access Asynchronous in this whitepaper?

A: Continuous Access Asynchronous has been placed in "maintenance mode" and will no longer be offered as of the XP generation that succeeds the XP24000/XP20000. If you require information concerning XP Continuous Access Asynchronous, please consult the older internal whitepapers or contact the factory.

Q: If I issue "pairsplit–S" (SMPL) or "pairsplit" (PSUS) to CntAc-Sync, CntAc-Async or BC pairs, under what conditions will the volumes be the same at the completion of the command.

A: Either command to CntAc-Sync or CntAc-Async volumes will make sure the volumes are synchronized before command completion. For CntAc-Sync they are always the same. For CntAc-Async, delta data is copied before command completion. For BC, you must suspend the pairs (PSUS) before deleting them (SMPL) in order for them to be synchronized.

Q: How is data copy different between initial copy pairresync updates and normal updates?

A: Initial CntAc copies are full copies in ascending cylinder order (as are differential copies from some extreme failures in PSUE [error] state). Pairresync copies from PSUS (suspend) state are differential (delta, out of order) copies based on a bit map of changed cylinders. Data is copied based on the logical OR of the bit maps in the S-VOL and P-VOL. Normal (pair state) updates are a block (512B) at a time. This can be more efficient that one competitor who has a minimum transfer size of 32KB. **(Competitive Advantage)**

Q: Is CntAc loop-back into the same array allowed?

A: Yes.

Q: Can an array be a MCU for some CntAc pairs and a RCU for others at the same time?

A: Yes. Array fan-in can be up to 8:1 per CU (64 per array). Fan-out can be up to 1:8 per CU (64 per array). No association between the volumes outside of one array is allowed (e.g. CntAc-Async Consistency groups are not allowed to span multiple arrays. CntAc-Sync device groups may, but there is NO data consistency/ordering provide across S-VOLs.)

Q: If a host read fails due to a bad P-VOL, can CntAc provide the data from the S-VOL through the inter-array link?

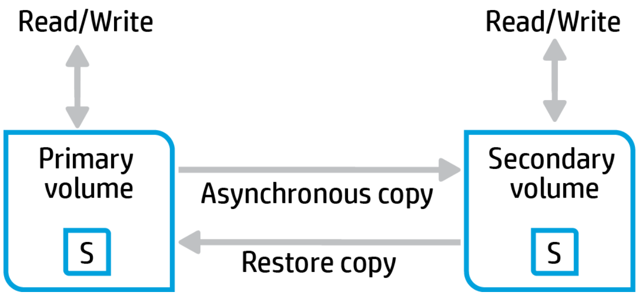
A: No

Q: What could cause a PSUE [error] state.

A: When the CntAc volumes cannot be synchronized (e.g. update copy error). PSUE at the MCU can also be a result of a pair delete (pairsplit–S) at the RCU side or a MCU power failure. In the case of a MCU power failure, the RCU must be operating before power to the MCU is restored. [See more on this subject under CntAc-Async Q&A].

Q: Can a host write to a CntAc S-VOL?

A: Only if it's in a suspend (PSUS) status. The S-VOL is read-only in most states (i.e. pair, copy). This is the official list of internal states and when P-VOLs/S-VOLs can be written to:



Status	Pairing status	Primary	Secondary
SMPL	Unpaired volume	R/W enabled	R/W enabled
PAIR	Paired duplicated volume. Data in the primary and secondary volumes are not assured to be identical.	R/W enabled	**R*** enabled
COPY	In paired state, but copying to the secondary volume is not completed.	R/W enabled	**R*** enabled
RCPY	This state shows copying from secondary to the primary volume via restore option. [BC only]	**R*** enabled	R enabled
PSUS	In paired state, but updating the secondary volume data is suspended. The primary volume controls the differences of the updated data.	R/W enabled	R/W enabled
PSUE (Error)	"PSUS" status owing to an internal failure. The primary volume may not maintain a delta table.	R/W enabled (See Note 1.)	**R*** enabled

Pairing Statuses

Note 1: Reading and writing are enabled, as far as no failure occurs in the primary volume.

R*: Reading are disabled when specified "-m noread" option of the paircreate command. [BC Only]

Q: During an initial CntAc copy operation, can the host write normally to the MCU P-VOL?

A: Yes, but the paircreate initial copy priority can have a performance impact.

Q: Are you allowed to delete a RCU from the array list when volumes are active?

A: No. This operation will only succeed if all CntAc S-VOLs associated with that RCU have been deleted ("pairsplit–S" to set them to SMPL state).

Q: Does it matter whether I do a pair delete (pairsplit–S) from the MCU or RCU?

A: Yes. When executed from the MCU, both the P-VOL and S-VOL go to smpl (simplex, un-paired) state without a problem. When executed from the RCU, only the S-VOL goes to smpl. The P-VOL goes to PSUE when the MCU detects this. To complete the deletion of the P-VOL, you must do (a forced, "pairsplit–S") a delete from the MCU.

Q: Why don't we allow RAID0 with CntAc (or BC)? Can't the combination of RAID0 (if it were allowed) with CntAc (or BC) protect data while saving disk space?

A: RAID0 is data striped across multiple disks without a parity disk (i.e. if you lose a disk, data is lost). RAID1 is total duplication with 100% disk overhead. RAID5 is striped disks with a parity disk (about 25% disk overhead). **Customers have lost data using RAID0**. Even if BC or CntAc mirrors were connected, if you lose one RAID0 P-VOL disk, either the application stops or an expensive failover occurs. With RAID1/5 the application keeps going undisturbed. If BC or CntAc is temporarily disconnected due to a link failure (or backup, etc.) and you lose a disk in RAID0, you have lost all the data since the split. With RAID1/5 in this scenario, the application continues. If a failover takes place with CntAc, the remote S-VOL is still RAID1/5 protected. With EMC RAID0, the data is vulnerable unless you suffer the 100% overhead of a BCV mirror. So the correct answer (normally) is to use RAID5 for solid data protection with only a minimum of overhead. EMC does not make this recommendation because their RAID5 is slower than their RAID1. Our RAID5 is not slower than our RAID1. EMC can read (slowly) through the SRDF link if they lose a RAID0 disk and the link happens to be up. If a disaster hits the remote site under those conditions, the customer is in a very bad data loss position. With HP, that RAID0 situation can't occur, so slow reads through the link are not supported. **(Competitive Advantage)**

Q: Should I mirror my entire Oracle dB via CntAc or just the online redo logs?

A: Oracle Standby Database allows the option to automatically (as of Oracle8i) send archived redo logs to be applied to an Oracle dB in standby mode at a remote site. In this case, CntAc is used only to mirror the online redo logs (for Geo-Mirrored Standby dB), which would be used in a disaster to bring the dB current. The alternative is to use CntAc to mirror all the dB (logs, tables, etc). Each has its merits but Oracle suggests the first approach:

Pros/Cons—CntAc mirror the online logs only:

Uses less CntAc link bandwidth with higher dB performance (if CntAc-Sync)

Logical corruptions will not automatically propagate to the mirror copy (caught at redo application)

Slower failover time (must apply online redo logs)

Failback may be faster (may or may not need to copy the dB across the CntAc link)

More complex approach

Pros/Cons—CntAc mirror the entire dB:

Uses more CntAc link bandwidth with lower dB performance (if CntAc-Sync)

Logical corruptions will automatically propagate to the mirror copy

Faster failover time (just do a dB crash recovery)

Failback is slower (may need to copy the entire dB across the link)

Simpler approach

Are able to use CntAc-Sync for redo log, CntAc-Async for table spaces

(2007 Update)

Q: Is it true that CntAc adds no value when Oracle DataGuard is being used?

A1: Correct, if it is the same data or instance of Oracle. If it is a different instance of Oracle, it may make sense. Some customers will CntAc one of the instances and DataGuard the other depending on the recovery objective. The recovery is much faster if using DataGuard.

A2: The mechanism described above is more akin to log shipping whereas DataGuard (DG) ships individual redo entries. DG can be configured to not allow a transaction to complete until it has been written to two or more destinations. Thus it provides the same level of protection as CntAc sync. Alternatively DG can be configured in async mode. In principal, the option to (1) use DG to copy the archived redo logs, and CntAc to only transmit the online redo log, or (2) to use CntAc to copy everything would be better achieved by DG alone to synchronously transmit redo log entries - as DG needs less bandwidth and has less of a latency impact than CntAc (at least if Oracle are to be believed).

Q: If several pairs are mirrored in Sync (status fence) and the link fails, can the P-VOL host still write?

A: The pairs that are not currently active would initially be P-VOL_pair, S-VOL_pair. The P-VOL pair that’s written to would go to P-VOL_PSUE with a write error because the S-VOL can’t be changed to PSUE over the link. In order to write to that P-VOL, you would need to do either a P-VOL horctakeover (which would change all the P-VOLs in that device group to PSUE) or “pairsplit –S” the P-VOL(s) to simplex.

Q: Can an XP connect to a HDS/Hitachi branded array via CntAc?

A: The GUI & CLI allow for minimal (but not full) functionality.

Q: SRDF can “automatically” re-synchronize the data between the primary and the backup site when there is a link break during an SRDF operation. Why can’t we?

A: So far, nobody has asked for such an enhancement because of the data integrity issues. Once the link is lost and a bit map starts getting used, data ordering information is lost. Since the resync operation will be out of order and (with the S-VOL inconsistent until completion) users often want to take actions first, such as break off a consistent S-VOL BC mirror before starting the resync. Thus, many customers would not want automatic resync. A customer does have the option of sensing the error and sending down a Pairresync from a host script. EMC can get away with a default to auto resync the links because their links are unidirectional. Since XP links can switch directions (e.g. horctakeover) some smart software like MetroCluster or CLX needs to make the call of when it's safe or not safe to resync based on which side has the most current data.

Q: Why do the S-VOL changes disappear so quickly at resync? Situation:

(1) Sync CntAc PAIR/PAIR (2) Write file A, B, C to P-VOL

(3) Suspend CntAc (4) Write large file D to S-VOL

(5) Resync CntAc (6) Suspend CntAc

(7) Notice that file D is gone from S-VOL VERY quickly. How is this done so fast? I know here a bitmap on both sides for CntAc-Sync. If the host looks for file D on the S-VOL while it’s still in COPY mode for the resync (without pairevtwait-pair), what happens?

A: The resync copies P-VOL data to the S-VOL cylinders that were updated by file D. If there is a lot of update to copy file D, the copy time might be long. A S-VOL read in copy state returns whatever data happens to be on the S-VOL side at that instant (if the customer does not wait for the copy to finish first [i.e. pairevtwait–pair]). The data consistency of the host side Read operation cannot be guaranteed during the resync copy state. That read data is not usable. Currently we are required to make the host read wait for the completion of resync. Most customers use the P-VOL side, and it is OK to safely R/W the P-VOL during the resync time.

Q: How should bad block reallocation (BBR) be set on the XP vs. LVM? For EMC it’s turned **off** at the array and **on** for LVM. XP protects data via Raid. No need for any alternative block protection facility from the SW side. We guess the reason of LVM tool may be for the recovery of permanent error data within cache storage in case of EMC. If there were no tools, LVM and DKC side may leave the error status forever, even though the system level recovery is completed in some cases. The XP has a Pinned track recovery tool for DKC maintenance so LVM BBR is not needed.

A: HP/Hitachi suggests turn off BBR @ LVM. It can’t be disabled on the XP. Don’t want them working across purposes.

Q: Does EMC Farpoint (SRDF) have an advantage over CntAc XP?

A: No, like the XP, Farpoint allows multiplexing multiple device I/O command and data operations via a single link at the same time. The benefit is a minimum delay over a long distance using a single link.

Q: How do HW errors effect BC/CntAc?

A: The following is a summary. Raid Manager XP Documentation shows the data consistency scenarios.

- Data consistency is defined by the combination of both sides status, and the fence level.
- Single Point Failure can be avoided.

The exact behavior of BC or CntAc depends on the failure;

(case 1) Failure of either array cluster (e.g. CL1 or 2)

The subsystem can continue operation although a SIM is reported. RM also can work. During pair status of CntAc, blockade of either cluster of cache memory or shared memory causes suspend status of volumes.

(case 2) Failure of both array clusters (e.g. CL1 or 2)

The failure part:

- CHP ports: It is impossible to access RAID from host. But consistency of data can be retained.
- Shared memory: In case of CntAc, S-VOL takeover is executed by the cause of DKC failure, then Sync/Async S-VOL data sequence can be kept after it. In case of BC, P-VOL and S-VOL data consistency cannot be kept. BC XP guarantees data consistency in only PSUS/SSUS state, and BC cannot do so in PAIR. This case is PSUE/PSUE and it is impossible to manage the delta table.
- Cache memory: It is impossible to update S-Vol.
- ACP ports: The subsystem can operate with correction access or LDEV blockade occur. It depends on a part of the failure.

Q: Is a firmware upgrade on a XP possible while its links to the remote site are down and the CntAc pair status suspended (error)?

A: Yes.

Q: What's the story on HDS NanoCopy?

A: HDS NanoCopy is for mainframe only today. Think of it as BC (at the end of a Cont. Acc.) split by appointment. It is of value in MF environment to compensate for no online dB backup ability (such as would be found in Oracle on-line Backup) on mainframes. Since that ability exists with Oracle on open systems, nobody yet sees the need for NanoCopy on open systems.

Q: If a CntAc P-VOL is a BC S-VOL and I execute a BC fast split to PSUS followed by a CntAc split to PSUS, so that the CntAc S-VOL will be consistent, does the CntAc split have to be delayed for the true BC background suspend copy to complete (if so, can pairvolchk track the percent done of the BC suspend copy?)

A: BC and CntAc combination does not allow pair & pair, or copy & copy state at the same time. When you specified BC fast split, CntAc operation can start immediately. However the CntAc resync or paircreate will be started after BC real background split finishes.

Q: Why is a SRDF failback slower that CntAc?

A1: Let's take the non-HW-failure scenario (e.g. when a customer has just failed over for primary site host for maintenance). Both SRDF and CntAc failover to the remote side and start updating the remote copy of the pair. Now its failback time.

With CntAc:

The original site disk (current S-VOL) is kept up to date by the failover disk (current P-VOL). [A swap-takeover took place at failover time, leaving both in PAIR state].

Shut down the application at the remote site

Swap the CntAc personalities via swap-takeover (a few seconds)

Start the application up on the primary site with all the latest data

With SRDF:

The R1->R2 data direction can't be swapped (without a CE visit and re-configure of MetroCluster, which are not likely to happen).

Do several point-in-time re-syncs from R2 back to R1 while the application still runs on and alters R2. This will always miss copying the data that changed after the resync command. This may take significant time and still be incomplete. If this optional/manual step is not done, the final resync will be very long.

Shutdown the application at the R2 (failover) side.

Do R2 to R1 resync (could take significant time, depending on application activity since the last resync)

Startup the application on the R1 side. You could start the application up before the re-sync completes but read-throughs will be slowed to link speed and the customer will have a false sense of security in that R1 (during the re-sync) is not actually current or consistent and R2 is out of date. If the link is lost during the re-sync, R1 is missing data and is inconsistent (unusable). If R2 site is lost during the re-sync, data is lost and R1 is unusable. There is no recovery from this.

A2: Now, let's take the case of a synchronous link failure that appears as a primary side storage failure. New CntAc "fast failback" functionality causes the original S-VOL to go to SSWS state and start taking I/Os from the application on the remote host (even though it can't "see" the original P-VOL. The S-VOL tracks the changes in a bitmap. The link is restored.

With CntAc (**Competitive Advantage**):

Do "pairresync–swaps" [fast failback] which (1) copies the latest delta data to the primary site (P-VOL), (2) swaps the P-VOL/S-VOL designations (3) makes them PAIR state.

Stop the application on the remote side

Do a swap-takeover

Start the application on the primary side. All of this was done via RAIDmgr. No on-site assistance was needed.

With SRDF:

Do several point-in-time re-syncs from R2 back to R1 while the application still runs on and alters R2. This will always miss copying the data that changed after the resync command. This may take significant time and still be incomplete. If this optional/manual step is not done, the final resync will be very long.

Shutdown the application at the R2 (failover) side.

Do R2 to R1 resync (could take significant time, depending on application activity since the last resync)

Startup the application on the R1 side.

A3: If the P-VOL/R1 disk is lost and must be replaced

With CntAc:

While the application continues to run on the remote side, do paircreate which does a full copy from the remote (P-VOL) side to the primary side S-VOL.

Stop the application on the remote side

Do a swap-takeover (very fast)

Restart the application on the primary side.

With SRDF:

A full R2 to R1 resync copy is required. Can the application start on R1 during the copy? With MetroCluster, NO. It is possible manually. EMC allows users to access the R1 volume while it is being updated by the R2 volume. If you change (write) data on R1 while this update is in process you will have some significant issues: 1.) Performance—the link is already fully utilized for the resync from R2->R1 and now you also change data on R1 which has to be transferred to R2. 2.) If either of the two arrays goes down (or the link) both copies are in an inconsistent (unusable) state. Because of that MetroCluster doesn't use this approach and nobody working in HA environments would recommend this unless you have a consistent copy (BCV) split before you start this process. If that BCV exists, its probably on the R2 side. If the whole R2 array is lost, the customer has nothing usable anywhere.

Q: I heard that a write error using sync data fence (due to down links) allows a write of the failing I/O on the P-VOL side (and obviously not the S-VOL side). If I cut the CntAc redo log mirroring links between a production database and a standby database, bring both up and compare, I may find that the primary site database has a few more I/Os that had not been committed up to the application. Please elaborate.

A: The P-VOL will write the I/O to disk that caused the write error (and write fencing) to the host. Thus, the P-VOL may be one write ahead of the S-VOL. Hitachi notes that this is correct SCSI disk behavior in that the write I/O to any regular SCSI disk that is returned with an error, may or may not have actually written data to media. In this design, it does write data to media. The Hitachi basic spec says that mirror consistency is assured (because an error is returned). The application is expected to correctly use that error information at a higher solution level. It does not say that a write will either go to either both sides, or neither, as many of us has interpreted. Our documentation must be reviewed for this misconception (see the section entitled: Geo-Mirrored Oracle to a Standby dB).

Q: What does the '-s' option to horctakeover do?

A: With '-s', the S-VOL RM will not even try to talk to the P-VOL RM to do a swap-takeover. It will instead, do just a S-VOL takeover.

Q: When can CntAc be resync'ed p->S-VOL, and when can it be resync'ed s->P-VOL?

A: If a CntAc link is suspended (SSWS) due to a horctakeover with the latest microcode, the only direction allowed for resync is S-VOL to P-VOL via the "pairresync -swaps[p]" command. If a CntAc pair is suspended (PSUS) due to a pairsplit command, the P-VOL data is still dominant. Can you do a P-VOL to S-VOL resync? – YES. Can you do a S-VOL to P-VOL resync (-swaps[p])? – YES.

Q: Can a older generation of XP replicate to a newer XP with CntAc?

A: With some limitation. See the online STREAMS document for Cont. Access XP

Q: I understand your logic about CntAc's behavior during aborts. I can see that SCSI ABTS are likely to leave writes on the P-VOL that are not reflected on the S-VOL and that this may be un-avoidable due to the nature of SCSI abort and the CntAc 2-phase commit process. I'm wondering if continued aborts, either due to a failing HBA or a factory test would cause the P-VOL to become more and more different from the S-VOL?

A: Your case is rare; Write aborted before the S-VOL gets the data, with no host retry. If this case continues, there may be an increasing difference. However the purpose of Abort is cancel the previous I/O and it must be retried by the host per the SCSI Standard. If you consider only an Abort sequence, without a retry, you should be concerned. However SCSI protocol says your case should never happen.

Q: When a SCSI abort causes a P-VOL write that does not get reflected on the S-VOL, does the P-VOL bitmap bit for that cylinder get set so that the P-VOL change will eventually migrate to the S-VOL?

A: Yes.

Q: If yes, will paircurchk or pairvolchk indicate any consistency problem until they become the same?

A: Yes. Your case is for only Abort. And your case changes SCSI sequence as you like.

Q: For Data fence, can CntAc link errors result in one side being different?

A: For Data Fence, if the reply from the RCU never makes it back to the MCU, the I/O will:

(A) be returned to the host with an error, and

(B) be destaged from cache to disk on the P-VOL and S-VOL side

If the request to the MCU never makes it to the RCU, the I/O will:

(A) be returned to the host with an error, and

(B) be destaged from cache to disk only on the P-VOL side

Q: Does MetroCluster support the HDS arrays?

A: No. It may or may not work, but it is definitely not supported.

Q: My thinking was that if the user doesn't use LUN security and RM security, he shouldn't pay the price of additional scanning, whether it's via ioscan or SCSI inquiry. If the security is off, I think one can avoid the automatic scanning by setting HORCMPERM=HORCMNOINST.

A: As of RM 01.06.03, during startup, it still uses ioscan and raidscan (raidscan will do scsi inquiry to list of rdsk, in this case the list from ioscan but the list can be provided by a text file pointed by HORCMPERM). There are 2 ways to turn on RM Security. By setting env. var. “HORCMPROMOD” from the host or at LUN configuration of command device by turning on security for command device.

cmddev security	HORCMPROMOD	RM Security
ON	set	ON
ON	Not set	ON
OFF	set	ON
OFF	Not set	OFF

Whenever RM Security is ON by either means of above, and if HORCMPERM=MGRNOINST, no volume/pair is permitted. This is to prevent user to get around the security by not allow RM to do Inquiry. Truly, if the user is not using RM security(OFF), they can set HORCMPERM=MGRNOINST to avoid scanning or HORCMPERM=/etc/horcmpermXX.conf to shorten scanning time.

In summary,

If user does not use RM Security,

- set HORCMPERM=MGRNOINST, for fastest RM startup (no ioscan, no raidscan at all)
- set HORCMPERM=/etc/horcmpermXX.conf, for faster RM startup (no ioscan, raidscan only which takes little time)

If user uses RM Security,

- if no HORCMPERM set or /etc/horcmpermXX.conf not provided, RM may take longer to startup (ioscan and raidscan activity)
- if HORCMPERM set and /etc/horcmpermXX.conf is provided, RM will start quicker (no ioscan, raidscan only;)
  However, make sure horcmpermXX.conf is up-to-date.

Also if RM is used in cluster environment, make sure above is done accordingly when you start RM manually depending on your scenario. In the case that your cluster software handles RM startup, make sure the cluster software is supplied with correct RM instance number, horcmXX.conf, HORCMPERM setting and horcmpermXX.conf etc.

Q: In case of Primary-XP has only RCPs (senders) and Remote-XP has only LCPs (receivers), Can Ack data return from Remote-XP to Primary-XP?

A: Yes.

Q: I'd like to reconfirm CC operation in the above CntAc Environment. In case of Primary-XP has only RCPs and Remote-XP has only LCPs, is CC support?

A: This is really a question for the HA lab, but I will say “yes”. CC is supported if Primary XP has all RCPs and Secondary XP has all LCPs.

Q: And, When a Disaster occurs at the Primary Site, Can I perform “cmrecoverycl” (horctakeover) on Remote Site as normal?

A: Yes. This should work.

Q: What are the guidelines for online XP upgrades regarding having different RM or FW versions during the upgrade (with or without MetroCluster).

A: ***Rules for Online Upgrades (different RM versions. Different FW versions)***
Online firmware upgrade XP Firmware upgradescan be done online.

XP Continuous Access data replication can continue during firmware upgrade (i.e. P-VOL and S-VOL can remain paired in status PAIR).

Raid Manager XP/XP Firmware version backward and forward compatibility

The latest firmware *will* fall back to the CntAc functionality of an earlier firmware in case the firmware on both arrays differs.

Raid Manager 'may' or 'may not' tolerate talking to another RM of a different vintage.

The use of earlier Raid Manager XP versions with newer firmware version is not supported.

Note
This is somewhat confusing since it goes against the fact that Raid Manager XP 1.3.3 was supported on XP512 for MetroCluster environment!

Different versions of Raid Manager XP (e.g. OS type, or HP vs. HDS) cannot communicate with each other.

Metrocluster and Raid Manager XP interaction

MetroCluster and Continuous Access interact via Raid Manager only at package startup.

In particular, no action is taken in package shutdown.

Q: If the P-VOL contained a bootable image, would the S-VOL (in PSUS state) be assured to be also bootable?

A: (Hitachi) What CntAc does is to keep the contents of P-VOL and S-VOL identical. CntAc does not assure that the content of volume is bootable. I think that it is possible in few special cases but is generally impossible.

Q: What is the XP1024 Controller ID and to what value can this field be set?

A: The "Controller ID" is used to specify DKC type of RCU side. Expected values are as follows (which should eventually make it into the User's Guide):

DKC type: Controller ID

------------------------------------

XP512/XP48 2

XP1024/XP128 3

XP12000/XP10000 4

XP24000/XP20000 5

Q: I have a question on how CntAc XP resynchronization works when the resync is in progress and the host continues writing new I/O to the local XP during the resync.

A: While re-sync is in progress, if a new I/O from host modifies a track that is already copied to the remote XP, this new I/O will be written to both local XP and remote XP via CntAc. If it modifies a track that has not been copied to the remote XP, it will only be written to the local XP. The new change will be copied as a part of resync process. If the fence level is 'never' or 'data', the data resync and new I/O will be copied to the remote XP synchronously. If the fence level is 'async', the data resync and new I/O will be copied to the remote XP asynchronously with a new data sequence number.

Q: If I want a one-way replication then I can configure a single port as initiator at the primary site and a single target port at the remote site. Any attempt to reverse the role would require that I:

- delete all pairings
- reverse roles (target local and initiator remote)
- create new pairs
- sync all the data back from the remote site

A: Yes, with one difference: If you didn't access the disks after you split them (both P-VOL and S-VOL) you can just do a pair creation without copying the data (paircreate-nocopy) to re-instate the disk pairs.

Q: I also thought that I could have multiple ports on the local site designated as initiators to multiple ports at the remote site designated as targets and the array will load balance across the multiple links. I realize that the performance is not linear as ports are added, but I could have load balancing and auto-link failover.

A: That is our experience, too. However that highly depends on the disk distribution. CUs are assigned certain paths (local FC port to remote FC port and local CU to remote CU). If you are able to balance the paths according to the CUs you can distribute IO load. Another factor is the number of array groups in the CU and such. (Like the performance guys say stripe as much as you can. Use as many different CUs as possible.)

Q: Ok, assuming that I have the above correct, what value is there in having an initiator and target pair at both sites other than to have data replicated both ways (frameA to frameB, frameB to frameA)? Are there benefits regarding failback of the data?

A: That's it. Bi-directional failover, no other reason, no other benefits.

Actually, thinking about it, if you mirror in both directions you use the parallel power of two arrays (and their controllers) to manage the mirror processes. There are 16 processes at a time per CU. If you mirror bi-directional you double the number of processes for mirroring and that decreases wait time for the hosts and you get therefore more IOPS.

Q: One other question, could I do the following:

1) quiesce the local frame (shutdown the apps)

2) create a backup of the local frame

3) pair the volumes to the remote site without initiating a data copy

4) split the pairs to initiate the creation of the delta track table

5) start up the apps

6) restore the backup at the remote site

7) resync the pairs

8) the frame copies over all the dirty tracks since the split

Is that possible? I am basically looking for options as to how to perform a new sync without doing it over the wire.

A: I just tested your procedure and it works. I am just not sure if it does a full copy or only the delta. It picks up on the sync percentage where it was before. But what happens is that it marks both sites with a "W" as in "wrote" :-) and, I guess, merges the track tables. The problem with your procedure is that you need to make sure that less than 50% of the data on each disk changes otherwise the array will do a full copy (faster than update at that point). What you can also do is mirror the arrays locally. Split the pairs. Put one array in a box and ship it to the remote site in less time than the batteries die (2 days, sometimes longer). Power up the array at the remote site and do delta copy (pairresync). This has been done in Europe successfully.

Q: I know CLX can be used with metro-distance (same cluster) solutions, but can it also be used with continental-distance (separate cluster) solutions?

A: Yes. CLX can be used between any kind of cluster (metro or continental reach). IBM's HAGeo is based on HACMP and works with CLX. Veritas's Global Cluster Manager is based on VCS and works with CLX. HP's Serviceguard for Linux doesn't offers ContinentalCluster for Linux but if you have one cluster on one site and the other on the other site it will work with CLX. Same applies to Windows. (Here, it's a bit more difficult but possible. We would need to change the install procedures but the code would work.)

Q: Does CLX work with EMC Symmetrix?

A: No. CLX does NOT support EMC Symmetrix. But if the customer pays for it, we might think about it.

For more information

HP XP7 Storage

HP XP7 Business Copy Software

HP XP7 Continuous Access Suite

Sign up for updates
hp.com/go/getupdated

Share with colleagues

Rate this document

4AA5-1154ENW, April 2014