

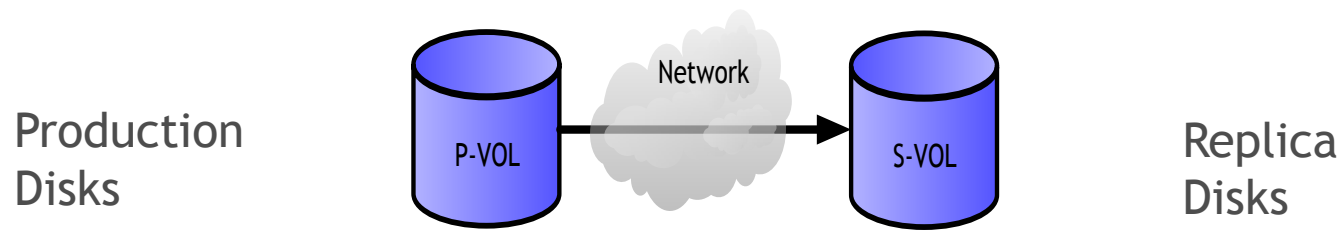
Universal Replicator Knowledge Transfer

June 24, 2014

HDS Global Account Team



Replication Basics



- Universal Replicator is a remote copy replication product
- As write traffic arrives at the production subsystem, it is transmitted continuously (as opposed to a batch transfer) over a network connection to a replica copy on a remote storage device
- In a perfect world, the primary to secondary network connection should be fast and wide; you want writes to transfer to the secondary without delay or congestion

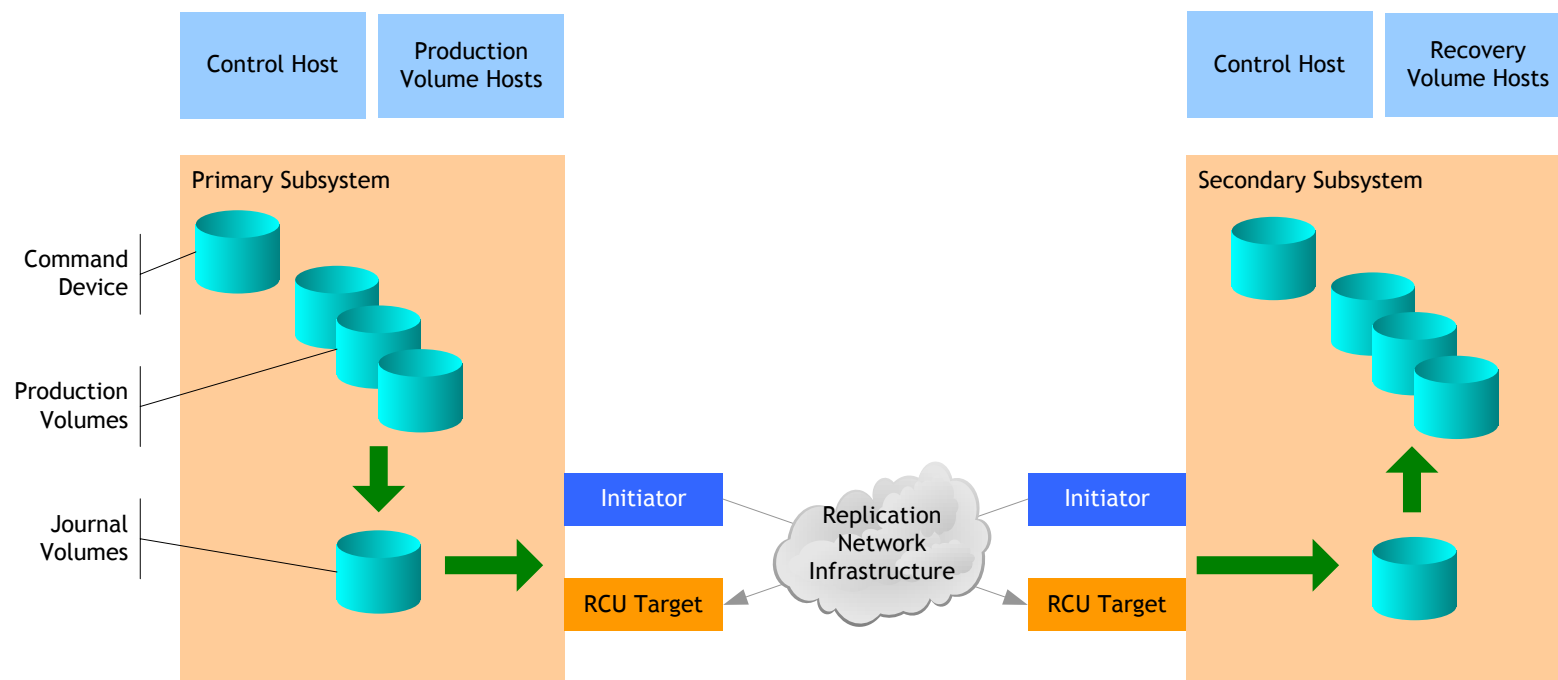
Background

- Universal Replicator is an asynchronous disk replication product from Hitachi Data Systems that has existed since 2005
- Replica copies of production disks are maintained on a secondary, remote subsystem
- Write traffic is transmitted continuously to the remote subsystem (i.e., this is not a batch transfer)
- Network links between the primary and secondary subsystems must provide sufficient bandwidth to sustain replication traffic

Components and Definitions

- Primary Volumes (PVOL)
- Secondary Volumes (SVOL)
- MDKC (Production Subsystem)
- RDKC (Recovery Subsystem)
- MJNL (Journal volumes on the production subsystem)
- RJNL (Journal volumes on the recovery subsystem)
- Initiator Ports
- RCU Target Ports
- Channel Extension
- Bandwidth, Latency
- Command Devices
- Host Control Software
 - RAID Manager/CCI
 - Business Continuity Manager
 - HRpM Hitachi Replication Manager

Universal Replicator Architecture



JNL CACHE

Journal Cache Specifications

- Generally Universal Replicator requires a smaller amount of cache than other asynchronous remote replications
 - The recommended minimum is **25% to 30%** additional cache for replicated data, plus an additional 1GB per Journal Group
- JNL Cache — MCU
 - In normal use, MCU journal cache contains mainly control (metadata) information
 - If the **data change rate** rises above link throughput, **MCU journal cache write pending ratio will climb**
 - Destaging of journal cache to journal disk will begin after Cache Write Pending reaches the 30-40% range
 - If **JNL volume throughput** is insufficient, the **journal cache write pendings may reach priority destage level**

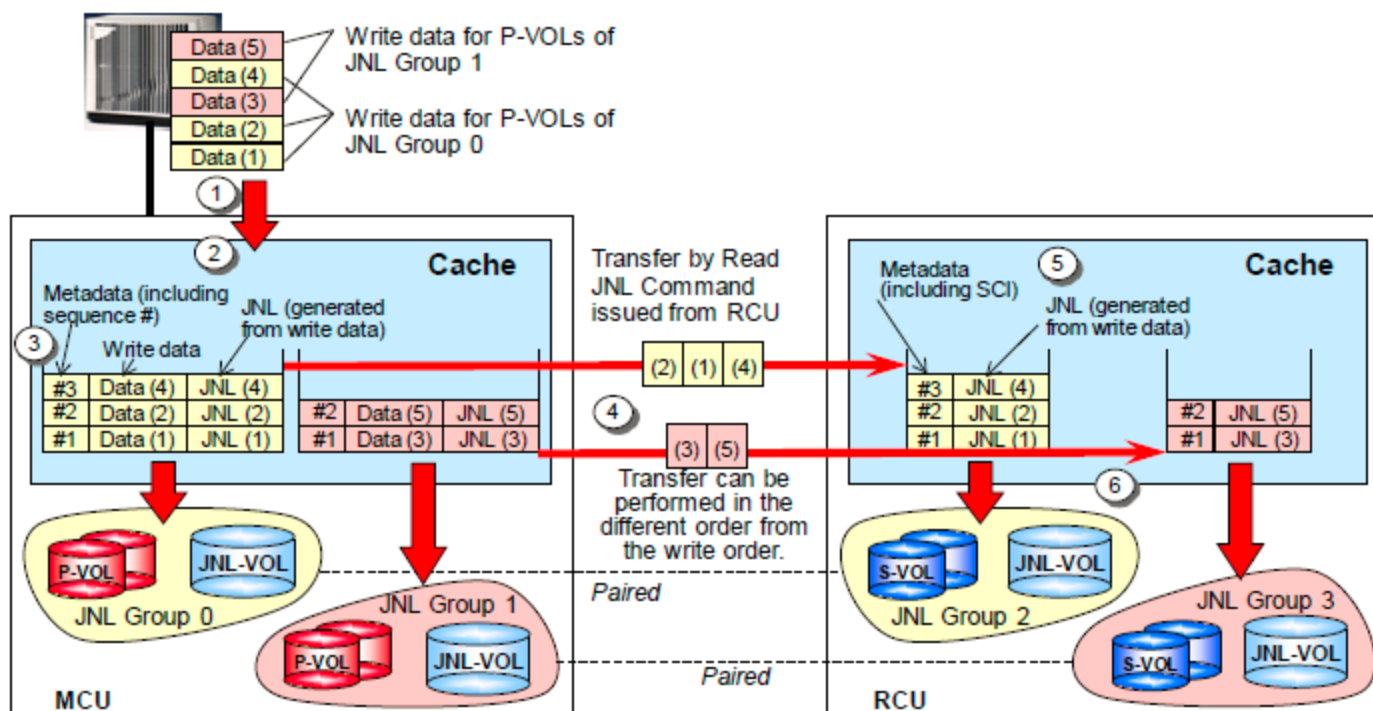
30

JNL CACHE

- **Use of Cache** parameter setting: Determines whether or not the RCU will use journal cache. Default is USE
 - Requires 25% to 30% additional cache in RCU, plus 1GB per journal group
 - Provides significant RCU performance benefit
 - If set to NOT USE, all journal updates go to R-JNL
- For both MCU and RCU, the only way to increase journal cache size is to increase total cache size

End to End HUR Process

- Sequencing guarantees that write order fidelity is maintained for each JNL group



15

Steps Explained

1. P-VOL receives write command from the host.
2. MCU stores the received write data to the cache.
3. For each write data, MCU obtains sequence# in the JNL Group and creates the metadata including the sequence#. MCU also allocates the area of JNL VOL and records the location the area to the metadata. (The write data and write order of them are managed with metadata for each JNL Group.)
MCU also creates JNL that can be stored in JNL VOL from write data.
4. RCU issues Read Journal Commands and MCU sends the metadata and the JNL as a response to the commands. They can be performed in a different order from the write order.
5. RCU stores the metadata and the JNL to the cache.
6. RCU sorts the JNL in the write order by referring metadata (sequence#) for each JNL Group and reflects (writes) them to the paired S-VOL according to the write order.

How Initial Copy Works

- **Base Journal process – Initial Copy**
 - The **Base Journal** process reads P-VOL data, creates metadata, and sends data
 - P-VOL is online, accepting host I/O. Changes in the data are noted in the bitmaps
 - When all data have been copied (including all changes that occur while copying), pair status changes to **PAIR**
 - **Nocopy** option is available if desired (*see note*)
 - **Base Journal Metadata - MCU** creates metadata pointers for P-VOL data locations as it is read. Those pointers provide a restart point if the Base Journal is interrupted

Initial Copy Explained

- Initial copy operations synchronize data in the primary data volume and data in the secondary data volume. They are performed independently from the host I/Os, when you create a data volume pair or when you resynchronize a suspended pair. The initial copy operation copies the Base Journal data that is obtained from the primary data volume at the Primary storage system to the Secondary storage system.
- The Primary storage system reads all data of the primary data volume as the Base Journal data, in sequence. The Base Journal contains a replica of the entire data volume or a replica of updates to the data volume.
- The Base Journal process denotes changes in primary volume by noting changed data locations in the Bitmap Areas. Once all changes are copied, PAIR status is declared.
- When a data volume pair is suspended, the primary volume Bitmap Areas is again used to note changes to the primary volume.

What happens once initial copy is done

▪ **Journal Obtain** process

- Begins after Base Journal is complete (PAIR status is achieved)
- Builds **Journal Updates** as data comes in from host
 - Metadata (control data) is created in **Journal Cache** for each update
 - Metadata includes **Sequence numbers** (and **CTQ-Markers** furnished by **CCI** if MxN Consistency Group is specified) to preserve write order consistency across all volumes in the R-JNL groups
- Metadata and data blocks are **queued for destage** to the primary journal volumes (**M-JNL Group**)
- **Journal Obtain** notification is sent to RCU
- Metadata and data blocks are held in cache as long as possible before destaging to M-JNL volumes

Update Copy - Cont..

▪ **Journal Copy process**

- Upon receipt of Journal Obtain notification, secondary storage system (RCU) responds with **Read Journal** Command to pull data from primary storage system (MCU)
- MCU responds by sending journal data (from cache if possible) as the **Journal Copy**
 - Data is sent with **Sequence Numbers** (and CTQ-Markers if MxN Consistency Group is set)
 - Even if Journal Copy sends journal data from cache, destage to the M-JNL occurs
- Sent data is received in **RCU Cache**, waiting to be settled. Data may be stored on secondary journal volumes (**R-JNL**) if necessary because of RCU constraints. (See note)
- **Read Journal** commands are sent repeatedly to primary storage system until MCU sends highest sequence number

13

Note: Among other causes, journal updates arriving at the RCU may be written to R-JNL due to inadequate RCU Journal Cache, or low data volume write I/O throughput due to fewer or slower disks used on RCU side than MCU side.

Update Copy - Cont..

- **Journal Restore process**

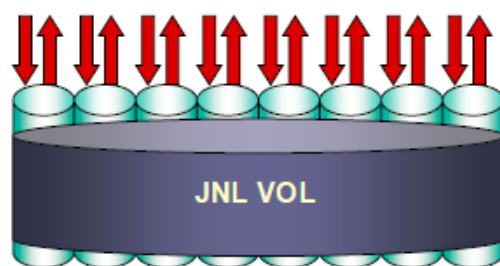
- After data is received in **RCU Journal Cache**, RCU initiates **Journal Restore**. Updates are reassembled by sequence number (see note)
- RCU notifies MCU of highest sequence number, along with the number of updates received
- Sequence numbers are compared to insure that all updates sent by MCU were in fact received by RCU
- After successful comparison of sequence numbers between RCU and MCU:
 - RCU writes data to S-VOLs in sequence
 - RCU **discards** metadata and any retained R-Journal entries
 - MCU **discards** entries in M-Journal

JNL - Best Practices

Journal Volume Physical Characteristics

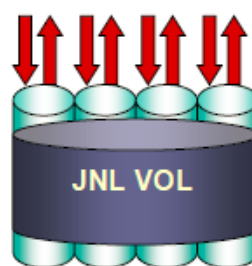
- RAID configuration

- RAID-5 (7D+1P) is recommended
 - 7+1 parity group is more efficient; spreads I/O across multiple parity groups, BEDS (see note)
 - Better than two RAID-5 (3D+1P) groups of same capacity
 - RAID-1 sequential write performance is low in comparison



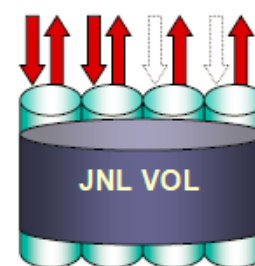
RAID-5 (7D+1P)

BEST!



RAID-5 (3D+1P)

Better



RAID-1 (2D+2P)

Worst!

29

Note: RAID-6 14D+2P is also supported. No guidance on suitability for Journals has been provided.

JNL - Best Practices

Best Practice Sizing for Journal Volumes

■ JNL VOL Throughput — Requirement

- JNL VOL throughput of M-JNL group must exceed the peak inflow data rate for data to be replicated
- If JNL VOL throughput is less than the peak data rate, Universal Replicator will suspend groups (PFUS)
- Components of JNL VOL throughput:
 - Journal Volume performance
 - High performance HDD
 - Number of journal volumes in the group (more is better)
 - Size of journal volumes (roughly equal is better)
 - Overall parity group utilization (low I/O load of other volumes in the parity group)
 - Best Practice: One JNL Volume per parity group per Journal Group
 - Transfer Bandwidth
 - Number of transfer paths (links)
 - Speed of transfer pipe

36

Initial Copy

How Universal Replicator Works

- Replication process

- **Initial Copy** (also called Base Journal Copy)

- During Initial Copy process, metadata pointers to data on the P-VOL and write sequence numbers are stored in the metadata area of the journal volume
 - The base journal data is obtained from the P-VOL and sent to the R-DKC in reply to Read Journal commands issued by the R-DKC
 - The data in the S-VOL synchronizes with the data in the P-VOL using the sequence numbering scheme stored as metadata on the primary journal volume
 - Only one synchronization pass is required for the initial copy and resynchronization
 - This operation is conceptually similar to Initial Copy in TrueCopy

34

Update Copy

How Universal Replicator Works

- **Journal Update process – After Base Journal completes:**
 1. **Journal Obtain** process is invoked. JNL data is queued for destage to MCU JNL volumes
 2. MCU sends **Journal Obtain** notification to RCU
 3. RCU issues **Read Journal** Command to initiate Journal Copy
 4. **Journal Copy** pulls from MCU cache (may have to initiate disk access if JNL data has been destaged to JNL volume) and sends data to RCU
 5. RCU executes **Journal Restore** and sorts JNL data by sequence number
 6. When complete, RCU compares **sequence numbers** with MCU. If agreement is reached, data is written to S-VOLs and JNL data on both M-JNL and R-JNL volumes are discarded

Initial Copy

Best Practice Sizing for Journal Volumes

■ JNL VOL Throughput — Requirement

- JNL VOL throughput of M-JNL group must exceed the peak inflow data rate for data to be replicated
- If JNL VOL throughput is less than the peak data rate, Universal Replicator will suspend groups (PFUS)
- Components of JNL VOL throughput:
 - Journal Volume performance
 - High performance HDD
 - Number of journal volumes in the group (more is better)
 - Size of journal volumes (roughly equal is better)
 - Overall parity group utilization (low I/O load of other volumes in the parity group)
 - Best Practice: One JNL Volume per parity group per Journal Group
 - Transfer Bandwidth
 - Number of transfer paths (links)
 - Speed of transfer pipe

36

Background

- Universal Replicator is an asynchronous disk replication product from Hitachi Data Systems that has existed since 2005
- Replica copies of production disks are maintained on a secondary, remote subsystem
- Write traffic is transmitted continuously to the remote subsystem (i.e., this is not a batch transfer)
- Network links between the primary and secondary subsystems must provide sufficient bandwidth to sustain replication traffic

Questions?

Thank You!