KAMINARIO DATA REPLICATION DEEP DIVE

Author(s): Chris M Evans

A DISCUSSION OF THE FEATURES OF A MATURE DATA REPLICATION SOLUTION

First Published: October 2015
Latest Update: October 2015
Document ID - LB1CD0081
Release 3

This White Paper was commissioned by Kaminario Inc and written and distributed by Langton Blue Ltd.

Table of Contents

Executive Summary

Data replication is an essential feature of modern shared storage arrays, delivering high levels of application availability. It provides the ability to mitigate the risk of data loss or system outage in the case of hardware failure, natural disaster or where data centre or hardware maintenance is required.

The process of remote replication protects data by creating an additional copy of volumes or LUNs on a secondary or remote (and physically distinct) storage array using an IP network. This array is usually deployed at a geographical distance away from the primary system to avoid the risk of both locations suffering from related issues, such as power outages.

Replication comes in three flavours; synchronous, asynchronous streamed and asynchronous snapshot. Synchronous replication provides a 100% guarantee of capturing all application updates, but at the expense of increased application latency. Asynchronous streamed replication removes the latency barrier but can’t guarantee 100% data integrity, with a delay in consistency of seconds or minutes. Asynchronous snapshot replication provides greater control over application consistency while reducing the network overhead of streamed replication protocols.

Kaminario K2 systems implement asynchronous replication based on existing native snapshot technology. Application consistency across multiple volumes or LUNs is achieved through the use of replication groups, which provide policy-based implementation of recovery point objectives (RPOs). Snapshots have the additional benefit of enabling the easy validation of replicated data copies through duplicating and mounting snapshot images at the remote site.

A “failover” is used to bring a secondary copy of data into use, while tracking updates or differences from the last good image at the production site. Once the DR issue has passed, normal operation can be resumed through a “failback” which copies any updates back to the primary site. Failover and failback can be operated at the replication group level.

Replication, failover and failback can be driven from either the K2 web GUI, CLI or REST-based API, making it possible to automate these processes through scripting or via existing replication solutions such as VMware’s vSphere SRM (supported by K2 using the Storage Replication Adaptor).

As with all Kaminario K2 features, replication is provided at no additional cost to the customer, making it a worry-free solution for delivering high-quality enterprise-class data protection.

Introduction

Objective

This report discusses the use of remote replication as a tool for providing data protection in enterprise environments. It explores how replication can be implemented in different ways based on requirements and physical constraints. Finally, the report looks at how replication is implemented on Kaminario's K2 storage array.

Audience

Decision makers in organisations looking to evaluate the implications of using array-based replication with all-flash storage arrays will find this report provides information on both the technology and the K2 platform. The report provides a basis for decision makers to develop a comparison methodology and selection criteria to aid them in their choice of solution.

Contents of This Report

- **Executive Summary** – a summary of the background and conclusions reached in this report.
- **Enterprise Replication Requirements** – a discussion of the need to implement array-based replication in enterprise environments.
- **Replication Explained** – a detailed look at how replication is implemented and what features to look out for.
- **Kaminario K2 Replication: Deep Dive** – an in-depth look at how K2 implements replication.

Enterprise Replication Requirements

Information is the lifeblood of the modern enterprise, forming a critical part of the intellectual property of large organisations. Pretty much without exception in today's world this information is data stored on computer systems that require curation and protection.

Data protection comes in many forms, with solutions such as snapshots used to protect data locally within an external storage system; tape or disk-based backups used to provide an offsite component to recovery and remote replication for providing ongoing business continuity. Remote replication (whether performed in the storage array, at the operating system level or the application) provides a secondary clone copy of the data sitting in the production environment. In the event of a problem in the main site, the remote copy can be used to continue business operations with the minimal amount of outage.

Data protection (and in particular business continuity) doesn't have to be used in a "disaster" scenario. There are many reasons why having a second copy of data can be extremely useful, including:

- Hardware issue with primary storage or systems – this could be related to the hardware itself, or be a result of environmental issues (loss of power/cooling, fire, flood), or problems with the network.
- Pre-emptive disaster avoidance – this includes moving the primary operations to another location to cater for an impending storm or other environmental risk.
- Maintenance work – production workloads can be moved to cater for system upgrades, building upgrades or other planned work. This can include load-balancing systems between sites, data migration and the creation of development images from production applications.
- Application load balancing – for example moving production processing of data to another location, for reporting or data analysis (Bi/data warehouse).

The service-level objectives of an application are typically defined by the following metrics:

- **RPO** – Recovery Point Objective is a measure of the amount of acceptable data loss within a recovery scenario and is measured in units of time (typically seconds, minutes or hours). An RPO of 0 indicates no loss of data. An RPO of 12 hours indicates the recovered data can be up to 12 hours old.
- **RTO** – Recovery Time Objective is a measure of the acceptable time allowed to return the application back to the original state before the failure and so describes the outage time to complete the restore process. An RTO of 0 indicates data recovery should be instantaneous, whereas an RTO of 4 hours indicates a maximum 4-hour recovery period to return to normal operations.

It is impractical to consider remote replication for all applications as many systems don't justify the cost of fast recovery. However, mission-critical applications (such as those in finance or banking) may require an RPO=0/RTO=0 in order to ensure customers can be served 24 hours a day. Reporting systems using copies or extracts of production data may tolerate a longer outage and be recoverable from older backups (e.g. RPO=12 hours, RTO=4 hours). Note that a recovery point objective of zero implies a "crash consistent" copy of data, which may require a longer recovery process – more on this later.

The challenge for enterprises running large numbers of applications with many interdependencies is in understanding their own recovery process. Application-based recovery provides the highest level of consistency but can be complex and difficult to manage at scale. For this reason, many companies prefer to use array-based replication, which is more efficient, easier to implement, but does require some work to implement.

Replication Explained

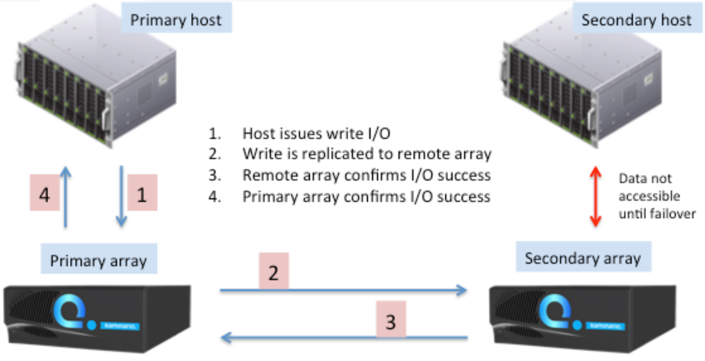
Data replication can be implemented in a number of ways. Local replication in the form of snapshots and clones provides point-in-time copies of data that enable recovery from data loss, corruption or other reasons that require the need to revert back to a previous copy or image of the data.

Remote replication provides the ability to maintain a secondary copy of data that can be used in the event that either the primary data or the primary systems are inoperative. Typically, remote replication achieves the following two goals:

- **Creating an independent copy of data** – providing a recovery copy on a completely separate piece of equipment to avoid problems that occur with hardware failure.
- **Creating a remote copy of data** – ensuring that a recovery copy is located remotely to the primary system and would not be impacted in the event of a disaster.

There are typically two ways to implement replication; either synchronous or asynchronous. Synchronous replication guarantees both primary and secondary copies data will be identical at all times, but is achieved at the cost of increased latency for the application. Asynchronous replication streams updates to a remote array at a point in time after write I/O completion has been confirmed to the primary host. In the event of a total disaster, this means some data may not have made it to the remote site.

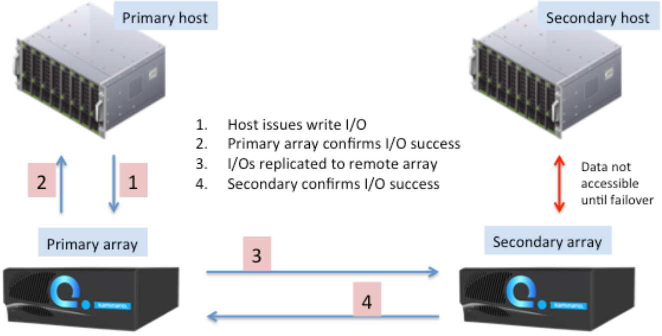
Synchronous Replication


Primary host
Secondary host
1. Host issues write I/O
2. Write is replicated to remote array
3. Remote array confirms I/O success
4. Primary array confirms I/O success
4
1
Data not accessible until failover
Primary array
Secondary array
2
3

SYNCHRONOUS REPLICATION

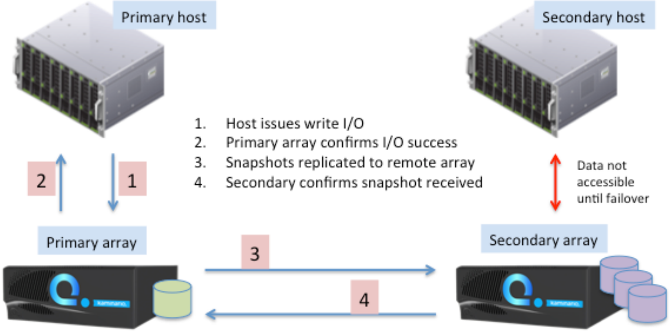
Synchronous replication moves data from a primary to a secondary array with a 100% guarantee that the data is accurate in both locations. This is achieved by delaying confirmation of the I/O completion to the host until the remote array confirms receipt of the data in the secondary location. From a consistency perspective, synchronous replication ensures all updates have been captured and duplicated for resiliency. However, this facility comes at a cost. The I/O completion time or latency from a host perspective includes the round-trip time to move the data to the remote location. The further away the secondary array is located, the higher the increase in latency. In all-flash arrays, I/O responses are of the order of 1 millisecond or less. The network hardware for a single round-trip link can double this figure, even before distance latency is added, making synchronous replication a significant overhead. Where there is a desire to replicate data out of a local or "metro" area to protect against a regional disaster, synchronous replication results in too much overhead to the application to be practical. Synchronous replication produces "crash consistent" copies of an application so may require extended time to bring the DR copy of the application into operation.

Asynchronous Replication


Primary host
Secondary host
1. Host issues write I/O
2. Primary array confirms I/O success
3. I/Os replicated to remote array
4. Secondary confirms I/O success
2
1
Data not accessible until failover
Primary array
Secondary array
3
4

ASYNCHRONOUS REPLICATION (STREAMED)

There are two forms of asynchronous replication; streamed and snapshot. Streamed replication transmits all updates at the primary array to the secondary array in sequence as they occur. Depending on the level of updates and the latency of the link, the amount of outstanding data to be transmitted to the remote site could be different by milliseconds, seconds or even minutes. Unconfirmed data in the primary site has to be tracked through metadata and may be retained in cache until confirmed at the remote site. This can cause problems if the remote array becomes significantly out of date with the primary system (for instance if replication links are unreliable).


Primary host
Secondary host
1. Host issues write I/O
2. Primary array confirms I/O success
3. Snapshots replicated to remote array
4. Secondary confirms snapshot received
2
1
Data not accessible until failover
Primary array
Secondary array
3
4

ASYNCHRONOUS REPLICATION (SNAPSHOT)

Snapshot-based replication uses periodic snapshots to coalesce all updates on the primary system and replicates the changes by shipping the coalesced updates to create the entire snapshot image at the remote array. Snapshot shipping is performed on a customer-defined time metric, typically every minute or greater. The benefit of using snapshot-based replication is two-fold; first, only the last update to any piece of data is sent across the wire. On systems with heavy update activity, this represents a significant bandwidth saving. Second, snapshots can be retained at the remote site and replayed to revert the remote site (or a clone) to a previous point in time. This is especially useful if logical corruption has occurred on the primary site – with synchronous and async-streamed the corruption will be faithfully replicated and applied to the remote copy before the issue has been identified.

The trade-off between the different array replication techniques are be summarised in Table 1 - Replication Characteristics.

Table 1 - Replication Characteristics

Characteristic	Synchronous	Async-Streamed	Async-Snapshot
Efficiency	Good – no additional local disk space required	OK, no additional disk space required – metadata required to track unsync'd updates	OK, additional space required to store and ship snapshots on both arrays
Latency	Bad, primary host latency affected by array<->array traffic	Good, primary host latency not affected	Good, primary host latency not affected
RPO	Good (RPO=zero)	OK – RPO dependent on bandwidth of links	OK, RPO dependent on snapshot interval and bandwidth
Overhead	OK – all updates streamed – throughput related to write-I/O update rate	OK – all updates streamed so throughput related to primary site update rate.	Good – updates batched/consolidated, eliminating duplicate updates.
Granularity	Good – all updates streamed to the remote site	Good, all updates streamed to remote site	Good – multiple updates consolidated into single snapshot.
Flexibility	OK – replication is on or off and will copy logical data corruptions.	OK – replication is on or off; may be possible to prevent replication of data corruptions.	Good – ability to choose specific snapshots to use for recovery/failover, plus consistency.

Using the Remote Copy

Data replication is pretty much a background task, until a need arises to use the remote copy. Typically, in a remote pair (a primary and secondary volume replication relationship), the primary volumes will be enabled for read/write operations while the remote or secondary volume will be read-only. To use the remote copy, a failover operation occurs. This swaps the status of the primary and secondary volumes, allowing the secondary to be updateable and act as the main copy for production workload. Once the incident is resolved at the main site, a failback occurs, moving operations back to the primary site. It's important to ensure that the failover/failback process operates as quickly as possible and this means making sure any updates occurring at the secondary site during the failover are quickly replicated back to the primary. With synchronous and async-streamed modes, the array pair must keep track of updates. With async-snapshot, updates can be replicated back by making and applying a snapshot of the updates since the failover took place.

Additional Features

In order to deliver efficient replication, there are a number of key features that should be looked out for.

- **WAN Optimisation -** Bandwidth can be expensive and so every attempt should be made to minimise the traffic traversing the network. Asynchronous-snapshot can reduce that workload, as can compression and de-duplication of the data before transmission. Consolidation of updates into a snapshot also provides significant bandwidth savings. Other techniques can be used to ensure that latency doesn't cause an issue, by loading or queuing up data in transit to make sure the replication link is fully utilised.

- **Application Integration –** Array-based replication produces a data copy at the secondary site that is best described as "crash consistent". Booting a server from the remote image will appear as if the server or application crashed or the server was powered off. Where possible, the application or hypervisor should be involved in the failover process in order to ensure a cleaner transition to the remote location. Integration with the database layer of the application, for instance, can ensure that application consistency is maintained by quiescing updates during the snapshot process.
- **Diversity –** replication at the array level shouldn't be constrained by the system model or current software release within the same product family, but should provide flexibility in allowing replication to be used as part of the upgrade or migration process.
- **Cost –** replication needs to be cost effective. The implementation of replication is expensive in the first instance, as an entire second copy of both the data and the server hardware have to be put in place. Charging for replication adds another layer of cost into the equation.

K2 Replication Deep Dive

Kaminario K2 systems implement remote replication using asynchronous snapshot technology. This is the same native technology used to deliver local snapshots. The use of snapshot technology provides a number of benefits:

- Replication is optimised to coalesce updates to the frequency of snapshots, with a granularity of less than 60 seconds. This means only the data changed between snapshots is replicated, reducing the WAN traffic load on systems with high write I/O activity.
- Volumes can be grouped together into a single replication or consistency group, ensuring that the replication process is consistent across a number of volumes that comprise an application. Replication groups also allow RPO policies to be applied separately to each group.
- Snapshot-based replication can be used to provide application consistent data copies through synchronising the snapshot process with pausing the application. This leads to lower recovery times (RTO) compared to using crash-consistent images.
- Up to 32 snapshot copies of a replication group can be retained at the remote site. This provides the ability to perform historical restores (to avoid a data corruption) or to use some snapshots for application consistent copies and others for general data protection (with little or no impact to the application). This puts the customer in charge to obtain the most flexible RTO/RPO option for recovery.

The replication process between two Kaminario K2 systems is created in the management GUI. A peer relationship between two systems is established using the management IP interface, with replication data traversing the front-end 10GbE connections. The system administrator simply needs to provide a target IP address and credentials (user & password) for the remote system. The amount of bandwidth used for replication can be controlled both at the physical layer (the WAN IP link) and at the replication group level. K2 systems use compression on the WAN links in order to reduce the amount of network traffic between systems.

Replication relationships can be established between multiple K2 systems, any of which can be a primary and secondary system in a remote pair at the same time. This provides the capability to support bi-directional replication.

Volumes or LUNs are placed into replication groups, which are used to establish the policy that is applied to replicated data. Replication snapshots are applied at the replication group level, which ensures that all related or connected applications in the group have their data consistently copied in one atomic action. The replication group also applies an RPO (recovery point objective) SLA to the replicated data, based on the interval between snapshots. Some applications may have less stringent replication requirements and so require less frequent replication, whereas other systems may need application integration to ensure data integrity.

When defining replication relationships, the target of the replication can be either an existing set of volumes or can be defined and created automatically at the time of defining the replication pair.

The Failover/Failback Process

In normal operations, the primary volumes in a replication group are marked as read/write and handle production traffic. The equivalent volumes at the remote site will be marked as read-only. During a controlled failover, updates to the primary volume(s) are suspended (by marking them read-only) and the secondary volumes(s) are enabled for writes. The administrator has the ability to choose which remote snapshot is used to bring the secondary volumes online. At this point, production work can continue and the systems keep track of any updates to the active secondary volumes. In a true disaster scenario, the primary system may no longer be available and so the secondary volumes are simply made available from the chosen remote snapshot.

If the reason for DR failover isn’t related to the storage systems, then in it may be desirable to reverse the direction of replication during the failover. This provides offsite data protection while in DR mode. K2 systems are capable of reversing the direction of replication for any replication group, while continuing to keep track of the differences between primary and secondary volumes.

In the event of a disaster at the primary site affecting the primary array, the secondary K2 system can keep track of updates and apply those back to the primary system once that system is restored. Failback returns systems to normal operation.

K2 systems integrate with VMware SRM through installation of the Kaminario SRA (Storage Replication Adaptor) on the SRM server at both the production and DR sites. K2 snapshots are application consistent and can be used to provide more reliable recovery than standard “crash consistent” replication copies.

DR Testing and Validation

One essential feature of replication is the ability to test the validity of the DR copies in the remote site. With synchronous or asynchronous streamed replication, the testing process can only be achieved by either suspending replication or taking a snapshot from the replicated LUN/volume (assuming that facility is permitted – some systems require replication to be suspended to take a snapshot). A copy taken in this way will not be consistent, or will be at best “crash consistent” and require the application to go through file system recovery at start-up. On K2 systems, a local snapshot can be taken from the DR snapshot at the remote site and used to test the validity of the DR image. This snapshot image can, of course, be application consistent too, providing both a high level of assurance in the DR process without risking a copy of the DR data. All of this is achieved without stopping or impacting the application or replication process.

K2 replication provides the ability to do more than simply provide a contingency copy of data. The ability to replicate data and snapshots provides the option to build out test and development environments, including either hardware shared or separate from the production system. All of this is achieved with the minimal amount of overhead, as snapshots are simply tracked differences from the production data.

Automation and Reporting

In enterprise environments operating at scale, the ability to automate the replication and failover process is essential. Automation enables the delegation of replication processes to the application layer, providing a greater level of data integrity. It also enables failover and failback to be managed through tools such as VMware vSphere SRM (Site Recovery Manager).

Kaminario K2 systems provide the capability to manage administration through either a web GUI, SSH CLI (command line interface) or REST-based API. This allows administrators to automate all parts of the replication and failover process through scripting. Kaminario provides support for VMware vSphere Site Recovery Manager (SRM) version 5.5 using the K2 SRA (Storage Replication Adaptor). This supports vSphere ESXi hypervisor releases 5.1, 5.5 and 6.0. SRM enables administrators to manage the failover process from within the VMware vCenter management GUI.

K2 systems show the status of replication through the management web GUI. A summary is provided on the “Replication” tab of the main Dashboard, with the ability to highlight more detailed replication information by clicking on the replication tiles within the dashboard window.

Licensing

Remote replication is delivered within K2 systems at no additional cost to the customer. This makes the technology cost effective and simple to deploy, with only the cost of additional storage capacity to consider. The efficiency of replication reduces bandwidth and in many cases can eliminate the need to purchase additional WAN optimisation software or hardware. This all-inclusive nature of features such as replication is especially attractive to service providers looking to add value for their customers.

More Information

Full details of the Kaminario K2 architecture can be found in the white paper “K2 Architecture White Paper” available on the Kaminario website.

For additional technical background or other advice on the use of flash in the enterprise, contact enquiries@langtonblue.com for more information.

Langton Blue Ltd is hardware and software independent, working for the business value to the end customer. Contact us to discuss how we can help you transform your business through effective use of technology.

Website: www.langtonblue.com
Email: enquiries@langtonblue.com
Twitter: @langtonblue
Phone: (0) 330 220 0128

Post:

Langton Blue Ltd
133 Houndsditch
London
EC3A 7BX
United Kingdom

The Author

Chris M Evans has worked in the technology industry since 1987, starting as a systems programmer on the IBM mainframe platform, while retaining an interest in storage. After working abroad, he co-founded an Internet-based music distribution company during the .com era, returning to consultancy in the new millennium. In 2009 he co-founded Langton Blue Ltd (www.langtonblue.com), a boutique consultancy firm focused on delivering business benefit through efficient technology deployments. Chris writes a popular blog at http://blog.architecting.it, attends many conferences and invitation-only events and can be found providing regular industry contributions through Twitter (@chrismevans) and other social media outlets.

No guarantees or warranties are provided regarding the accuracy, reliability or usability of any information contained within this document and readers are recommended to validate any statements or other representations made for validity.