

HP-UX File System Replication

What is it, how does it work, and what are its benefits?

Jim Kramer

John Saylor

Quest Software, Inc.

610 Newport Center Drive

Suite 1400

Newport Beach, California 92660

(800) 306-9329

Introduction

The move towards a more distributed environment together with the proliferation of decentralized heterogeneous databases, has brought a challenge to managers of information systems. As systems and databases continue to grow, managers face the question of how to deliver, replenish and extract data to and from independent data stores. These problems escalate when users require delivery of information in a timely, consistent and meaningful manner.

Some of the solutions can be provided by data replication.

Data replication is a class of techniques for copying data as it is being generated. Typically, the destination for the data is a different host than the source. Data replication can be used to provide both *wide* availability of data and *high* availability. Wide-availability data replication provides copies of data for current use to satisfy load balancing (performance off-loading), warehousing, or other data distribution needs. High-availability data replication provides online backup copies of data for standby use or for archiving to tape.

Quest Software has been providing data replication capabilities to the MPE community for many years, and now, with its SharePlex for HP-UX product, is making them available for HP-UX hosts. SharePlex provides replication for UNIX files and for Oracle databases. This paper discusses data replication in general and SharePlex file replication for UNIX files on HP-UX in particular. A companion paper, *Data Replication for HP-UX and Oracle DBMS*, discusses Oracle replication on HP-UX.

What is Data Replication?

Here is a general definition of data replication:

Data replication is a process of modifying one set of data (the destination set) according to changes being made to another set of data (the source set), in a well-defined way.

Thus the destination set of data is uniquely determined by the source set, allowing for the time needed to capture, propagate and process the changes. Data replication may keep the destination set identical to the source set; this is often what is needed for disaster recovery. Data replication also can be useful for other purposes: load balancing, distributed processing, etc. For such uses, the destination data may

be a subset of the source set (only selected rows or columns, for example), or the data may have been transformed in a well-defined way according to user specifications.

Uses for Wide-Availability Replication

Wide-availability replication provides these possible uses:

- **Load balancing** Replication can be used to provide the data for use on additional hosts if it is undesirable or impossible for a single host to handle the entire application load. For example, transactions can be made on the source host, while reporting and queries are done on destination hosts.
- **Distribution** Sometimes data is more conveniently accessed on a destination host (for example, for communications reasons) than on the source host. Data replication can provide a convenient solution in such cases.
- **Warehousing** A data warehouse typically contains snapshots of operational data for the use of management information systems. Data replication can be used to provide a continuous feed of operational data for warehouses.
- **Inter-Application** One application system may provide data that is of use to another application system. For example the order entry system may generate information about customer orders that is needed by the manufacturing system. This system may run on a different host. Data replication can be used to send the information to the manufacturing system.

Uses for High-Availability Replication

When a host goes down, its data is unavailable for use. The data may also be at risk for permanent loss. The business information stored on the host is relied upon for making informed decisions. Manual functions performed in the past are long forgotten, so the impact can be devastating. Many aspects of the business come to a halt, which usually translates into late deliveries, unhappy customers, loss of revenue, and loss of goodwill. Moreover, many computing environments today are required to run 24 hours a day. A critical issue for IS professionals to address is: if disaster strikes, can the business recover?

Disruption to normal business activities can come in many forms: natural threats, technical threats, and human threats. Every business is vulnerable.

High-availability data replication can provide a solution. It can back up data generated by critical applications to a standby host in real time. If the application host fails, the standby host can substitute for it. If desired, the backup copy of the data can be archived to tape without disrupting activities on the source host.

Elements of Data Replication

Now that we know how data replication is used, it is time to find out more about what it does.

A replication technique must do the following things:

- *Capture* the change data that describes the changes being made to the source set.
- *Propagate* the changes to the destination.

When changes are made to files, a replication technique can propagate either the entire files or just the changed parts of the files. For obvious efficiency reasons, SharePlex and most other replication techniques propagate only the changed parts.

There are various ways to do the capture and propagation, as discussed below.

Techniques for Capturing Data

There are many different techniques of data capture. One way to characterize a technique is by the *level* at which it operates. The highest level is to capture within the application itself. The lowest level is to capture at or near the hardware, perhaps as part of the hardware itself.

Replication at the application level can be very efficient because of the tailored control it allows over what data is to be replicated. The problem is that it requires extra programming and may therefore be quite expensive.

Hardware-level replication avoids programming but is typically not selective about what is replicated. Entire volumes or sets of volumes may have to be replicated. Transmission bandwidths may have to match the bandwidths of the storage devices. For these reasons, hardware and communications costs are high. With many types of hardware replication, files and databases cannot be accessed while replication proceeds. In such cases, this approach is useful only for high-availability uses, not for wide availability.

There are also software-based approaches that operate somewhere between these levels. SharePlex is one of these. It provides data replication selectivity without programming, as discussed below.

Techniques for Propagating Data

The captured changes must be propagated to the destination. Propagation requires, at a minimum, transmission of the change data to the destination host and posting of the changes to files and databases. There are three main approaches to propagation:

- **No storage** The change data is propagated immediately to the destination without being stored at an intermediate location. If there are propagation delays for whatever reason, changes are lost or activity against the source set must be suspended. Two-phase commit is a form of no-storage propagation in which the committing of changes to the source set depends on successful committing of changes to the destination set.
- **One storage** The change data may be stored at the source location prior to transmission to the destination. This allows changes against the source set to proceed even if transmission and posting to the destination are impossible for some reason.
- **Two storage** The change data may be stored at the source location and/or the destination location prior to posting. Storing at the source has the same purpose as for the one-storage case. Storing at the destination permits the full transmission bandwidth to be used even if posting cannot keep up with transmission. This is critical if replication is being used for high-availability reasons. Data transmission can proceed even if posting is stopped on the destination (for example, to permit backup of the destination copy).

SharePlex for HP-UX uses the two-storage propagation method (as does SharePlex on MPE). This method provides optimum operational flexibility and makes the best use of transmission facilities. It makes it possible to select transmission capabilities based on average rather than peak bandwidth, thus reducing costs.

Multiple Updateable Masters

So far, we have referred to replication as an asymmetrical operation in which the source set defines the destination set. However, we may wish to have two or more sets of data, each of which acts as the source to the others. In other words, we may wish to be able to change any set and have the change reflected in the other sets. This situation is often referred to as having “multiple updateable masters.” “Master” here just means data source.

The main issue in such a situation is what should be done if two or more of the sources disagree; that is, there are coincident and contradictory changes to the same areas of the different sources. There must be a resolution policy to determine which, if any, of the changes will be applied to the source and destination sets.

Synchronization

When file replication is proceeding properly, each destination data set is modified as needed to reflect what its source was at an earlier time. When this is true we say that the source and destination are *synchronized* or *in synch*. When this is not true, the files are *unsynchronized* or *out of synch*.

There may be a short delay (a few seconds or less) between the time a change is made to a source and the time the destination reflects the change. However, for various reasons, the time could be much longer. An example of this might be unavailability of the destination host.

Keeping the files in synch is the mission of data replication. If it doesn't do this, it is not replicating. Therefore, a replication technique must pay proper attention to safeguarding the data at all times, even if hosts, transmission lines, or processes should fail. Furthermore, there should be satisfactory methods of checking for synchronization.

SharePlex Replication for HP-UX

Now that we have discussed replication in general, let's take a look at SharePlex replication. SharePlex is a family of products on MPE and HP-UX that provides various approaches to meeting needs for high-availability and wide-availability data. SharePlex on MPE has been meeting the needs of MPE customers for more than ten years. It provides data replication, remote file access, spool file distribution, and other data-availability capabilities.

SharePlex on HP-UX is a new product that provides UNIX file and Oracle database replication on HP-UX hosts. Its design has had the benefit of Quest Software's unmatched experience with replication in the MPE marketplace.

Here are the key characteristics of SharePlex for HP-UX:

- SharePlex replicates UNIX files and Oracle databases in real time between two or more hosts. It does this by capturing changes made to data residing on the source hosts, and propagating those changes to all designated destination hosts.
- The granularity of replication is a table within a database or a directory (and its descendants) within a UNIX file system.

- Because SharePlex uses a two-storage approach to propagation, the user session does not wait while the data is being propagated to all destinations, nor does the user session wait if one or more destinations are unavailable.
- Only incremental file changes are sent to destinations to minimize host overhead and network traffic.
- Oracle replication maintains transaction integrity.
- File synchronization is continually checked during data posting.
- A tiered architecture allows SharePlex to grow with the network. A host can simultaneously be used as both source and destination. Replication configurations include unidirectional, bidirectional, broadcast, and consolidation.
- Replication configuration is done at the source host and only at the source host. Configuration information flows from the source host to destination hosts as needed for automatic configuration of the destination hosts.

SharePlex uses a software-based data capture technique that operates at an intermediate level; that is, below applications and above hardware. SharePlex monitors file and database activities of programs. Unlike replication at the application level, no changes to application programs are required. Since only data changes are replicated, SharePlex yields faster data transfers than batch data extract techniques while requiring far less host and network resources.

SharePlex Administration and Monitoring

Administration and monitoring are centralized. Administrators can easily manage and monitor replication from a single host. Destination hosts transmit status to source hosts, making centralized monitoring possible.

Two administration programs are available. One runs on Windows and provides a convenient multi-window point-click-drag-drop interface. The other runs on HP-UX and provides a TTY-compatible command-line interface.

Host and Network Configuration

A SharePlex solution requires a configuration consisting of one or more HP-UX hosts. Most solutions require at least two hosts, although online backup applications may be achieved by replicating to a separate file system of the source host.

A single host may serve as both a source host and a destination host. As a source host, it captures and propagates change data to other hosts according to its own configurations. As a destination host, it receives and posts data propagated by other hosts according to their configurations.

There is no restriction on source to destination combinations. Consider the following possibilities:

- Unidirectional: one host replicates to another.
- Bidirectional: two hosts replicate to one another.
- Broadcast: one host replicates to many.
- Consolidation: many hosts replicate to one.

SharePlex transport is based on TCP/IP. Source and destination hosts must be networked together with TCP/IP. Either a wide or local area network can be used to provide the connection. Line requirements

depend on the application. One MPE application had a continuous flow of one million transactions per day. They were sent across a T1 line and the transport layer only consumed 5% of the network. Some customers have even replicated transactions through satellite links.

SharePlex: Oracle Database Replication

Each Oracle database that serves as a source of data must have its own configuration. A configuration specifies what tables are to be replicated and the destinations for that table. Any table can be replicated to an arbitrary number of destination tables on as many hosts and as many databases on a host as desired. A table can even be replicated to multiple tables within the same database.

Replication is controlled separately for each database; for example, replication can be started and stopped for each database independently of what is happening with the others, and of what is happening with UNIX file replication.

The approach SharePlex uses for capturing data provides much higher performance than other competing approaches to Oracle database replication.

Oracle replication is covered in greater detail in the companion paper at this conference, *Data Replication for HP-UX and Oracle DBMS*.

SharePlex: UNIX File Replication

A UNIX file replication configuration specifies which directories are to be replicated and to which destination directories on which hosts. By default, when a directory is replicated, all files and directories that descend from it are replicated also, and to the same destinations. However, this default can be overridden to exclude a subdirectory from replication, or to replicate it to different destinations.

SharePlex supports replication of directories and regular files for both HFS and JFS. Creation and deletion of named pipes are also replicated.

UNIX file replication captures data using code that is added to the UNIX kernel. The code is added to the kernel in the same way that a driver would be. Hewlett-Packard source code is neither modified nor recompiled. The SharePlex kernel code monitors all modification activity to disk files on a source host. It captures changes to the files being replicated and stores the changes in a capture log, ready for propagation to destinations. When replication is disabled, the presence of SharePlex code in the kernel has no measurable performance effect.

Synchronization in SharePlex

Before replication starts, the files and tables being replicated must be synchronized between source and destination. This is a systems administration task external to SharePlex. In a future version, SharePlex itself will handle this synchronization.

As replication proceeds, it is critical that as much checking as possible be done to assure continued synchronization of source and destination. As data is being posted on the destination host, SharePlex performs the following types of synchronization checking and reporting:

- **File absence** If an operation (for example, remove, rename, write) is done against a file that doesn't exist, a synchronization error is reported.
- **File existence** If an attempt is made to create a file that already exists, a synchronization error is reported.

- **Preimage** If the block of data that is being overwritten does not match the corresponding data overwritten on the source, a synchronization error is reported.

With the current version of SharePlex, resynchronization of out-of-synch data is a systems administration task external to SharePlex. In a future version, resynchronization will be done automatically by SharePlex.

SharePlex was designed with the utmost concern for data safety in order to maintain synchronization. Synchronization survives SharePlex process crashes, destination host crashes, and network failures. Synchronization usually, but not always, survives crashes of a source host.

Routing

SharePlex transport mechanisms are based on TCP/IP connections; that is, data and other information are transmitted from source host to destination host via a TCP/IP connection between them. A replication configuration may specify that data may propagate to one or more intermediate hosts prior to reaching destinations, thus allowing fanout from the intermediate hosts. This would allow the data to be propagated to hundreds or even thousands of hosts.

Using SharePlex Replication for HP-UX

SharePlex replication provides the tools for high-availability and wide-availability solutions, such as fast disaster recovery, performance off-loading, concurrent backups, and 24-hour access to critical business data. SharePlex does not require networks and destination hosts to be continuously available. Transactions that take place during a network or destination host failure are queued on the source host until they can be propagated.

SharePlex saves time and development costs since there is no need to code custom replication applications. SharePlex is application independent; it does not require code changes to applications. Therefore, it avoids creating dependencies on specific packages.

Because it automatically and instantaneously updates destination copies, SharePlex replication is more powerful and convenient than data-extract or batch-transfer techniques. It also consumes less network bandwidth. SharePlex can replace many current FTP applications. It can also provide many of the benefits of NFS across wide area networks where NFS would not be feasible.

Using SharePlex: High Availability

In the event of a disaster, a destination copy of the applications, files and databases can be accessed immediately. Users simply reconnect to the destination host and resume normal processing. Typically, PC users have a second icon to run the application from the destination host. A switching front end can be used for terminal users.

Using SharePlex: Backups

Backups can be performed on a destination host while the source host is being accessed. Posting to the destination files and databases can be stopped to allow backup. Therefore there is no need for special online backup tools. Meanwhile, because of the two-storage architecture of SharePlex, data can continue to be received and stored on the destination host during backup, thus maintaining protection against disruptions on the source host.

Backups on a destination host remove the overhead on a source host that would be used by an online backup tool. Usually, the overhead of a backup can be tolerated more easily on a destination host.

Using SharePlex: Load Balancing and Data Distribution

SharePlex is a wide-availability solution; it distributes data where needed. Users can have the most current data locally instead of having to access one central source of data. Performance balancing or off-loading can be achieved by directing all reporting and inquiry requests for one group of users to the destination copy of the data, while another group of users can still access the source copy. Websites can be easily replicated to allow users of read-intensive sites to be serviced by multiple and perhaps smaller hosts.

Using SharePlex: Data Warehousing

SharePlex provides an easy path to data warehousing. It is the perfect solution for capturing data from production databases: since data is continuously updated, decision support users benefit from having a current snapshot of the business activity.

SharePlex: Future Directions

SharePlex will support additional hosts: Sun, IBM, Windows NT, and others.

SharePlex will support additional databases: Sybase RDBMS, Informix RDBMS, and others.

Future SharePlex offerings will provide additional features:

- Complete automatic checking that source and destination are fully synchronized.
- Automatic resynchronization of source and destination when they are not synchronized, without need for operator intervention.
- Subsetting and transformation of data.
- Multiple updateable masters.
- Selection by file as well as by directory. Selection by pattern as well as by name.

Summary

Data replication is an important tool for providing high-availability and wide-availability solutions. SharePlex replication for HP-UX draws on Quest Software's extensive experience with data replication to provide flexible, efficient, affordable and easily implemented replication of UNIX files and Oracle databases on HP-UX.

About the authors

Jim Kramer is project leader for SharePlex at Quest Software. He has been in the HP community for twenty years doing development work on HP 3000's and HP 9000's, and support work as an HP employee. He is the author of Quad and Diogenes, two well-known HP 3000 programs.

John Saylor joined Quest Software in October 1993. John brings to Quest a well-rounded background of Information Systems experience. He has assisted major corporations in several industries to plan solutions to their business problems, and has maximized the return on investment in current and future computing resources. He has held a variety of positions in the information technology world, ranging from 11 years at Hewlett-Packard as a Sr. Technical Consultant focused as a Performance Specialist, Capacity Planner and Data Center / Disaster Planning Advisor. His prior position was the MIS Manager for Western Digital Corporation. At Quest, John works with Hewlett-Packard customers to create disaster-tolerant environments and to bridge their HP environments with future technologies.

John has published papers in INTERACT magazine and presented papers at several HP User Groups and at HPWORLD '96.