Synchronizing Large Engineering Source Code Repositories

Scalable, multidirectional synchronization over distance with Aspera Sync

WHITE PAPER

aspera
an IBM® company

Synchronizing Large Engineering Source Code Repositories

aspera
an IBM® company

Scalable, multidirectional synchronization over distance with Aspera Sync

TABLE OF CONTENTS

HIGHLIGHTS

Overview

Aspera Sync is purpose-built by Aspera for high-performance, scalable, multidirectional asynchronous file replication and synchronization. Designed to overcome the performance and scalability shortcomings of conventional synchronization tools like rsync, Aspera Sync can scale up and out for maximum speed replication and synchronization over WANs, for today's largest big data file stores—from millions of individual files to the largest file sizes.

Use Cases

- Synchronization of remote source code repositories
- Sharing of design documents
- Distribution of build images
- Replication for disaster recovery and business continuity
- File and product archiving and storage
- Server and VM migration and replication

Benefits

- Built on Aspera FASP™ technology for maximum transfer speed regardless of file size, transfer distance, and network conditions
- Flexible deployments with support for one-to-one, one-to-many, and full-mesh synchronization
- Highly scalable synchronization architecture supports synchronization of millions of files and multi-terabyte file sets
- Locally detects changes and compares them to file system snapshot without having to check with remote systems
- Built-in deduplication saves transfer time, bandwidth and storage capacity
- Familiar rsync-like interface shrinks learning curve and simplifies deployment

Synchronizing Large Engineering Source Code Repositories

aspera
an IBM® company

Scalable, multidirectional synchronization over distance with Aspera Sync

OVERVIEW

Developing new technology, be it hardware- or software-based, is inherently a team-oriented activity. Writing innovative software or designing and building programmable hardware components requires engineers with a high degree of specialization. More and more, these multi-discipline engineering teams live and work in different remote offices, often times in different countries separated by distance and time zones. They remain connected by a wide area network (WAN) that supports all communication, data transfer, and sharing. To operate as a single high-performance virtual team requires a high degree of integration and collaboration, and the need to regularly distribute and share source code, software builds, detailed design documents, CAD files, test cases and tools, and other forms of IP.

However, sharing and distributing all this IP over long haul WANs with high-latency and packet loss using conventional tools and solutions creates technical and operational challenges, which require significant human and financial resources to create and maintain. The latency and packet loss created over distance disrupts replication over TCP. While TCP can be optimized, it remains an unreliable approach to large-scale delivery or acquisition of file-based assets for businesses and industries that require predictable performance, highly efficient bandwidth utilization, reliability, and security.

Multi-layered systems that rely on rsync for data replication, WAN optimization to improve network throughput, and data de-duplication to optimize storage are fragile and expensive to deploy and maintain. This unmanageable complexity leads to high failure rates and lack of visibility and control in the overall process and state of the underlying systems, resulting in huge inefficiencies and loss of productivity. System failures leave gaps in the workflow that prevent teams from interacting, and create waste when engineers work on outdated versions of the underlying source code. Finally, the money spent on maintaining these complex systems detracts from the main engineering goal of creating new value for the end-user and impacts the effectiveness and predictability of development.

Aspera's patented FASP™ high-speed transfer protocol has become the standard in bulk data transfer in industries such as Media and Entertainment, Government and Life Sciences to name a few. The innovative transfer software eliminates the fundamental shortcomings of conventional, TCP-based file transfer technologies. As a result, FASP maximizes the utilization of available bandwidth to deliver unmatched transfer speeds and to provide a guaranteed delivery time regardless of file size, transfer distance or network conditions.

The high-speed transfer technology is now available in Aspera Sync, purpose-built for highly scalable, multidirectional asynchronous file replication and synchronization. Aspera Sync is designed to overcome the bottlenecks of conventional synchronization tools like rsync and scale up and out for maximum speed replication and synchronization over WANs, for today's largest big data file stores — from millions of individual files to the largest file sizes.

CHALLENGES WITH TYPICAL SOLUTIONS

Today's development environments can be large-scale and complex. Source code repositories may contain from hundreds of thousands of files up to several million, and file sizes can range from tens of bytes to large multi-MB graphics and images to multi-GB tars. Collectively, a repository can grow to reach hundreds of GBs and when combined with build images and with versioning enabled, can easily approach a TB in size. Development teams are typically located in multiple remote locations, requiring collaboration and file sharing at a distance.

The typical solutions used to interconnect these teams with the underlying files and data they need are multi-layered, with many distinct technologies, working together to collectively solve the problem.

TCP ACCELERATION

There has been significant innovation in TCP optimization and acceleration techniques using in-band hardware appliances and storage devices. Mainly, these approaches utilize compression and de-duplication to minimize network traffic (e.g., chatty protocols) and incrementally send block-changes between appliances in support of application-specific workflows. This approach effectively optimizes certain applications from sending unnecessary information and data over the network, including some file synchronization scenarios.

Synchronizing Large Engineering Source Code Repositories

aspera
an IBM® company

Scalable, multidirectional synchronization over distance with Aspera Sync

However, these acceleration appliances do not expedite big data freighting and synchronization scenarios where very large files (such as graphics and build files) and very large collections of small files (such as source code and configuration files) need to be transferred, replicated, or synchronized, in bulk. When improvements in throughput are achieved, they often come at the expense of quality of service as other network traffic is effectively squeezed.

AD HOC SOFTWARE TOOLS

Today, a number of low-cost TCP-based file replication and synchronization tools are available through the open source community and software vendors. These range from open source tools like rsync to an entire ecosystem of do-it-yourself (often unsupported) tools for Mac and Windows. Typically this software is host-based (i.e., runs on each server) and can be used to replicate files point-to-point (unidirectional or one way) over TCP.

For small data sets across relatively low-latency wide-area connections, rsync provides an efficient unidirectional approach; it can also be configured to run back and forth between endpoints. For larger file systems and long-haul WAN connections, replication speed is limited by two factors: the time it takes to scan the local file system, and TCP. As files are created, moved, or changed on the source file system, the scanning process takes progressively longer to execute. As distance between source and target servers increases, latency and packet loss worsen on average, limiting the speed of the comparison of the local and remote systems to find the deltas, and the speed of replication over TCP. Running multiple, parallel rsync jobs improves throughput, but also significantly increases system complexity. For multi-site repositories with millions of files, thousands of concurrent rsync jobs are needed to achieve the desired throughput.

Thus, rsync works fine for small data sets over low-latency networks. But, as data sets grow and network latency increases, rsync quickly becomes impractical.

STORAGE DEDUPLICATION

As source code repositories grow, engineering teams have responded with data deduplication strategies in the storage layer, with everything from compression to file-level single-instance storage and even block-level deduplication for backup and recovery sites.

Both in-line and post-process storage-based data deduplication can effectively reduce the amount of storage needed for a given set of files. However, the effectiveness is dependent on the underlying structure of the data, and they yield little-to-no benefit over the WAN as the files must be pushed to the target system before the deduplication processing can take place. When storage dedupe is paired with the network-based deduplication of the TCP accelerator, complexity rapidly increases and system reliability decreases.

BENEFITS OF ASPERA SYNC

The approach Aspera has taken is founded on the following design principles:

SINGLE APPLICATION LAYER

In order to simplify the system architecture, Aspera Sync was designed to meet the requirements of speed and efficiency as a single solution. Built on FASP high-speed transport, data is transferred at maximum speed so sync jobs complete in a fraction of the time of TCP-based solutions. The highly scalable synchronization architecture supports synchronization of millions of files and multi-terabyte file sets at multi-gigabit transfer speeds on commodity hardware, without requiring thousands of concurrent sessions. Built-in data deduplication is designed to achieve equal if not superior deduplication on the file system as compared to the dedupe offered by proprietary storage systems, and effective compression ratios are equal to or better than rsync.

Synchronizing Large Engineering Source Code Repositories

aspera
an IBM® company

Scalable, multidirectional synchronization over distance with Aspera Sync

MAXIMUM THROUGHPUT OVER DISTANCE

Aspera's FASP transfer technology provides the core transport for Aspera Sync, eliminating network-layer bottlenecks, irrespective of distance. Aspera FASP scales transfer speeds over any IP network, linearly. The approach achieves complete throughput efficiency, independent of the latency of the path and robust to packet losses. For known available bandwidth, file transfer times can be guaranteed, regardless of the distance between endpoints or network conditions — even over satellite and long distance or unreliable international links.

Compared to TCP/Accelerated TCP, Aspera FASP adaptive rate control is:

- Loss tolerant. Reacts only to true congestion, while remaining immune to inherent channel loss.
- Perfectly efficient. Ramps up to fill unclaimed bandwidth, regardless of latency and packet loss.
- TCP fair. Quickly stabilizes at a TCP-friendly rate when links are congested without squeezing small traffic.
- Stable. Zeroes in on the available bandwidth.

Aspera performed a set of tests to baseline performance compared to rsync. The tests provided the following:

- **Reproducing real-world WAN conditions—with 100mls delays and 1% packet loss over a 1Gbps link.**
- **Small files: measuring the performance and results of replicating many small files—into the millions.**
- **Large files: measuring the performance and results of replicating large files—into terabytes.**
- **Time: measuring how long it takes to perform both small and large file replication scenarios.**
- **Throughput: measuring the overall throughput utilized during the replication job.**
- **Change replication: changing a fraction (10%) of the files in the data set and measuring the time and throughput to replicate the changes.**

Table 1 - Performance Comparison Synchronizing Many Small Files (Average size 100 KB) over WAN of RTT 100ms / Packet Loss 1%

Small File Performance	Number of Files	Data set size (GB)	Sync Time	Throughput (Mbps)
Aspera sync	978944	93.3	9,968 sec (2.8 hours)	80.4
Rsync	978944	93.3	814,500 sec (9.4 days)	0.99
			Aspera Speed Increase	81X

Table 2 - Performance Comparison Synchronizing Large Files (Average Size 100MB) over WAN of RTT 100ms / Packet Loss 1%

Large File Performance	Number of Files	Data set size (GB)	Sync Time	Throughput (Mbps)
Aspera sync	5194	500.1	4664 sec (1.3 hours)	921
Rsync	5194	500.1	4,320,000 sec (50 days)	0.98
			Aspera Speed Increase	940X

FAST SYNCHRONIZATION OF NEW FILES AND CHANGES

Aspera Sync quickly captures file changes through snapshots and file change notifications. This enables Aspera Sync to quickly acquire and replicate files based on changes. Aspera Sync both quickly captures and synchronizes file changes within a very large and dense directory tree.

Table 3 - Synchronization time when adding 31,056 files to 1 million small files (100 KB each) over WAN of RTT 100ms / Packet Loss 1%

Change File Performance	No. Existing Files	No. Files Added	Total Size (GB)	Sync Time	Throughput (Mbps)
Aspera sync	978,944	31,056	2.97	947 sec (16 min)	26.9
Rsync	978,944	31,056	2.97	37,076 sec (10.3 hours)	0.68
				Aspera Speed Increase	39X

Synchronizing Large Engineering Source Code Repositories

aspera
an IBM® company

Scalable, multidirectional synchronization over distance with Aspera Sync

Table 4 - Synchronization time when adding new files to set of large files (100 MB each) over WAN of RTT 100ms / Packet Loss 1%

Change File Performance	No. Existing Files	No. Files Added	Total Size (GB)	Sync Time	Throughput (Mbps)
Aspera sync	5194	54	5.49	54 sec	871
Rsync	5194	54	5.49	54,573 sec (15 hours)	0.86
				Aspera Speed Increase	1000X

FILE-LEVEL DEDUPLICATION

Duplicate files on the source system are identified and Aspera Sync instructs the target system to create a hard link rather than resending the file, eliminating unnecessary transfers, reducing sync times and reducing storage use.

OPEN, CROSS-PLATFORM SOFTWARE

Aspera Sync is compatible with industry-standard networks, storage, servers, and operating systems. This enables the utilization of any type of storage supported by the host operating system—and interoperability across multiple vendor solutions.

HONORING DIRECTORY SEMANTICS

A key capability of Aspera Sync is the ability to rename, move, delete, and copy files and directories on any source and target endpoint included in a synchronization job. This means end users can safely interact with files and directories exactly as they would on Linux, Windows, or Mac. If a user accidentally deletes or renames a file or directory, the file or directory can be re-synchronized to restore.

CENTRALIZING MANAGEMENT AND REPORTING

Aspera Console, the optional web-based interface for centralized transfer management, provides real-time notification, logging and reporting capabilities, while maintaining a centralized transfer history database for detailed auditing and customized reporting. The console enables administrators to centrally manage, monitor, and control transfers and replications across endpoints, called nodes, and precisely control and monitor all transfer and bandwidth utilization parameters.

COMPLETE SECURITY

The FASP protocol provides complete built-in security without compromising transfer speed. The security model, based on open standards cryptography, consists of secure authentication of the transfer endpoints using the standard secure shell (SSH), on-the-fly data encryption using strong cryptography (AES-128) for privacy of the transferred data, and integrity verification per data block, to safeguard against man-in-the-middle attacks. The transfer preserves the native file system access control attributes between all supported operating systems, and is highly efficient.

Aspera offers an option to integrate with LDAP servers and Active Directory. Both users and groups can be created in LDAP and used in conjunction with Aspera's client and server products.

Aspera transfers and synchronization jobs can be tunneled within Virtual Private Networks (VPNs) over any standard IP network. Firewalls and other perimeter security measures supporting UDP such as Network Address Translation (NAT) are fully compatible.

RESILIENCE AND AVAILABILITY

File replication and synchronization are resilient to end-to-end failures, from the host system (source and target) and across the network. If a failure occurs on the source, jobs quickly restart where the job left off. Aspera Sync can also be load-balanced across end-points. If a synchronization job is interrupted, Aspera Sync will resume at the point at which the last file was transferred. (Some tools like rsync require the entire job to start over.)

DEPLOYMENT MODES, MODELS AND USE CASES

Aspera Sync provides two modes of operation: one-time synchronization and continuous.

ONE-TIME SYNCHRONIZATION

Ideal for testing and replicating in ad hoc ways, one-time sync can be scheduled or manually initiated. In this mode, all endpoints defined in the job will synchronize once. When the job is complete, no further action will be taken.

Synchronizing Large Engineering Source Code Repositories

aspera
an IBM® company

Scalable, multidirectional synchronization over distance with Aspera Sync

CONTINUOUS SYNCHRONIZATION

Most ongoing replication operations will run in continuous mode. Continuous sync can be scheduled or manually initiated. Once scheduled, synchronization occurs transparently in the background and continues as files or directories are added, updated, or changed.

UNIDIRECTIONAL

Much like the rsync model, Aspera Sync supports replicating files and directories from a source to a target. If updates are being maintained at a single location and the sole goal is to propagate the updates to a target server, this scenario may be adequate. Two-way unidirectional is also supported.

BIDIRECTIONAL

Bidirectional synchronization occurs between two endpoints. While any number of use cases could be supported, some examples include:

- Point-to-Point synchronization: files are kept current and in sync between two servers or between two remote sites.
- Disaster recovery: a primary site replicates to a remote secondary site, used for backup.

MULTIDIRECTIONAL

Multidirectional replication can be used in support of many advanced synchronization scenarios. These include but are not limited to:

- Hub-and-spoke synchronization, where a core source (hub) server or cluster replicates to *n*-number of endpoints (spokes). This configuration could support distributed workflows, remote or branch office replication, and any other scenario requiring replication to multiple endpoints concurrently.
- Source code distribution: in cases where a customer needs to synchronize downstream caching points, regional trunks, or other peering points, Aspera Sync can be used exclusively or in combination with other products.
- Cloud ingest and distribution: Aspera Sync can be used in situations where local files need to be uploaded and subsequently synchronized with multiple repositories within the cloud data center.
- Collaborative file exchange: Aspera Sync can also be used in support of collaboration, where one user may upload a file and need to synchronize the file with multiple users or remotely located sites.

SUMMARY

Development teams that need to replicate and share large repositories of source code, engineering specifications, and design documents across multiple remote locations have had to integrate multiple different technologies to achieve the speed and efficiencies needed for collaboration. The resulting systems are expensive to create and maintain, and remain fragile with frequent failures that lead to waste and loss of productivity.

Aspera Sync is purpose-built by Aspera as a single application-layer solution for highly scalable, multi-directional asynchronous file replication and synchronization of large-scale file repositories typical of today's development environments. Sync is designed to overcome the bottlenecks of conventional synchronization tools like rsync and scale up and out for maximum speed replication and synchronization over WANs, efficiently utilizing bandwidth and storage. The optional Aspera Console provides global visibility into Sync sessions and activity across the entire network. The complete solution meets the performance requirements of today's high-performance development teams, while significantly reducing the complexity and cost of multi-layered systems developed in-house.