AVANCE PERFORMANCE WITH 10G NICs

Contents

EXECUTIVE SUMMARY

In Avance software release 2.1, Stratus has implemented support for 10 Gigabit Ethernet NIC's for the data synchronization link between the two Avance PMs. Avance guarantees consistency of guest O/S (VM) data by replicating it between the local storage attached to each PM. The time required to replicate data between nodes contributes to the total time required to complete each I/O operation performed by a guest O/S. Decreasing the wire-time for replicated data would be expected to yield some performance benefit. If the total disk subsystem throughput exceeds the bandwidth of the replication link, then the replication link can become the performance-limiting factor. This paper describes the conditions under which a faster replication link is likely to translate into improved performance, and offers some metrics which customers can use to decide whether to install 10 Gb cards in their Avance systems.

I/O READ PERFORMANCE

Avance performs read I/O from the local disks, and does not normally send read I/O data across the replication link. As a consequence, installing a 10Gb replication link will have no effect on read performance. If the nature of your performance bottle-neck or the affected application is mainly related to read operations, this alone would not be a compelling reason to upgrade to 10Gb links.

ENHANCE PERFORMANCE - RECOVERY, INSTALLATION, IMPORT OPERATIONS, SYNCHRONIZATION

If the disk subsystem can support write throughput in excess of a 1 Gb private link (~ 100 MBytes/second), certain types of operations will likely benefit from a faster replication link. Operations that result in large, sequential, sustained writes, such as database restores, software installation, and large file copy operations fit this description. Likewise, after a physical Avance node has been powered down and requires resynchronization with the primary node, this operation will proceed up to 10 times faster with a 10Gb replication link. Failover, when Avance needs to migrate running VM's between nodes is another operation which will benefit from the faster link, because the migration traffic (which flows over the primary 1 Gb private link) will not compete with the replication traffic flowing over the 10 Gb replication link. Use of 10 Gb Ethernet cards can also offer increased system hardware redundancy, since Avance will utilize more than one 10 Gb port for replication, and will fail-over between links if either of them fails.

PERFORMANCE ANALYSIS – A DEEPER DIVE

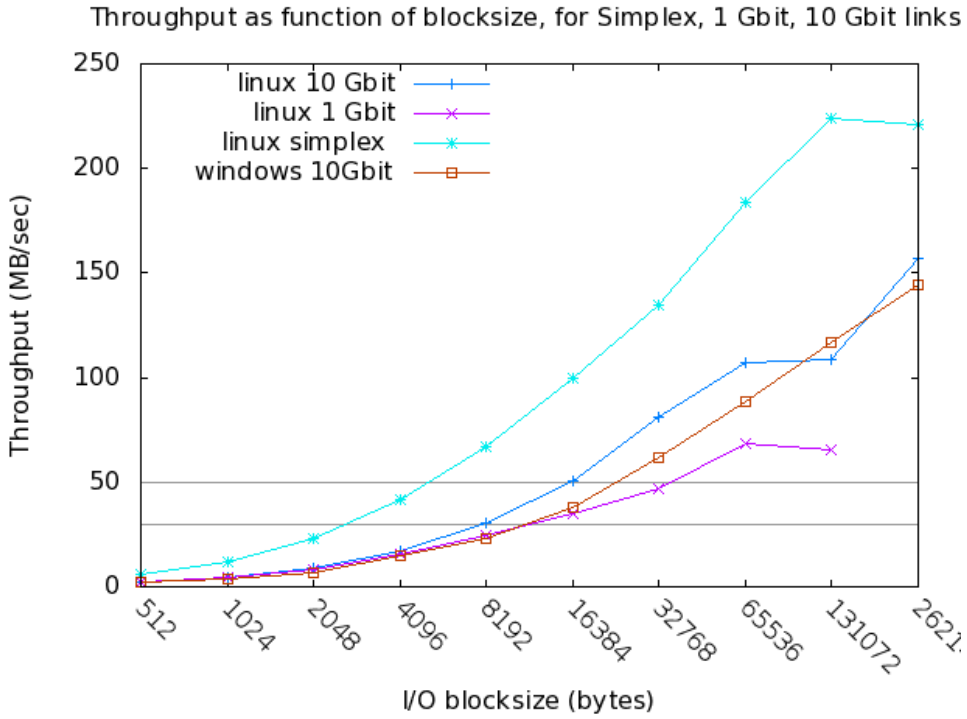
EFFECT OF I/O SIZE ON THROUGHPUT

In addition to the total disk I/O write throughput, the size and quantity of the I/O operations dramatically affects whether a faster replication link will deliver an application performance gain. In general, large, sequential writes will perform better than small, random operations. This is because the latency required to perform network replication is better amortized over large I/O blocks, and the wire-transit time consumes a larger fraction of the total I/O replication time, as compared with smaller I/O sizes. Applications such as databases, which issue small (8 KB) writes, requiring each one to complete before starting the next, will be much more affected by the replication latency than applications which performs pipe-lined operations, for example a file-server. Replication latency consists of the time required to

actually transmit the data over the wire, plus the overhead of moving the data through the guest and host operating systems and protocol stacks. At small I/O sizes the wire-time becomes an insignificant portion of the total latency, so a faster network link offers comparatively little benefit.

In order better to advise customers of the benefit of installing 10Gbit links, Stratus has conducted several types of tests in our labs. The simplest test was a simple write I/O test, performed on a single VM, using the 'dd' utility. The graph below shows the results of this test, for several runs. Each colored line shows the total throughput obtained when writing I/O at a given block sizes. The effect of 1Gb 10Gb, and 'simplex' links are shown. In this case, "simplex" means that one Avance node is powered off, completely removing the data replication from the I/O path. (Of course, with only one node running, the VM's and their data are not HA, so this is not a recommended real-world configuration. But it does show us the effective "upper limit" on the speed improvement, assuming an infinitely fast replication operation. Testing with SQL Server workloads shows that Avance's performance in simplex mode approaches 90 - 95% of bare-metal.) Note that the theoretical upper limit on I/O throughput for a 1 Gb link is approximately 100 MBytes/sec. In fact, the 1 Gb test case shows that a single VM can achieve approximately 75% of this maximum.

We can draw several conclusions from the data shown in the following graph. First, we see that at I/O block sizes below 8K, there is little performance difference between 1Gb and 10Gb links. At a block size of 16 KB, we begin to see a significant improvement due to the 10 Gb link. Another way to read this graph is to look at the total I/O bandwidth at which we see significant performance gains due to the 10 Gb



link. Below 35 MBytes/sec (the lower horizontal line), there is little difference between 1 and 10 Gb (note where the purple (1Gb) line crosses the horizontal). Above 50 MBytes/sec, we see almost a factor of two improvement.

PERFORMANCE SCALING IMPROVEMENTS WITH MULTIPLE VMS

Some customers run more than one performance-sensitive application on a single Avance system. In these cases we are interested in knowing whether upgrading the replication link will have a benefit for aggregate performance. The table below shows the results of two tests: a file copy test on four Windows VM's, and a file creation test on five Windows VM's.

Test	4 VMs file copy	5 VMs file creation
1 Gb	85 MB/sec	95 MB/sec
10 Gb	140 MB/sec	441 MB/sec

Notice that the file copy test over 10 Gb on four VM's is already performing faster than possible with a 1 Gb link. At the larger number of VM's, the file creation test over 10 Gb shows that write performance continues to scale well past the maximum 100 MB/sec possible over a 1 Gb link. This data shows how a system running multiple VM's can benefit from the faster replication link, assuming that the system is capable of saturating a substantial portion of the 1 Gb link in its current configuration. Notice also, that even with five VM's running a write-intensive operation, the guests are not able to consume even half of the 10 Gb link bandwidth. So the replication link is no longer a performance bottle-neck for this system.

COLLECTING DATA FROM DEPLOYED SYSTEMS

In order to predict whether a particular system may benefit from a faster replication link, we must collect some statistics from the running system. For customers already running or benchmarking Avance, the best place to start is at the Avance U/I. By navigating to the "Statistics" tab of the "Storage Groups" page, the total disk write bandwidth can be determined for each shared mirror. Be sure to un-check the box marked "Disk Read", since read I/O does not involve the replication link, as mentioned above. By summing the values obtained for each shared mirror, we can determine whether enough replication traffic is being generated to suggest this system may experience a performance boost with 10Gb links.

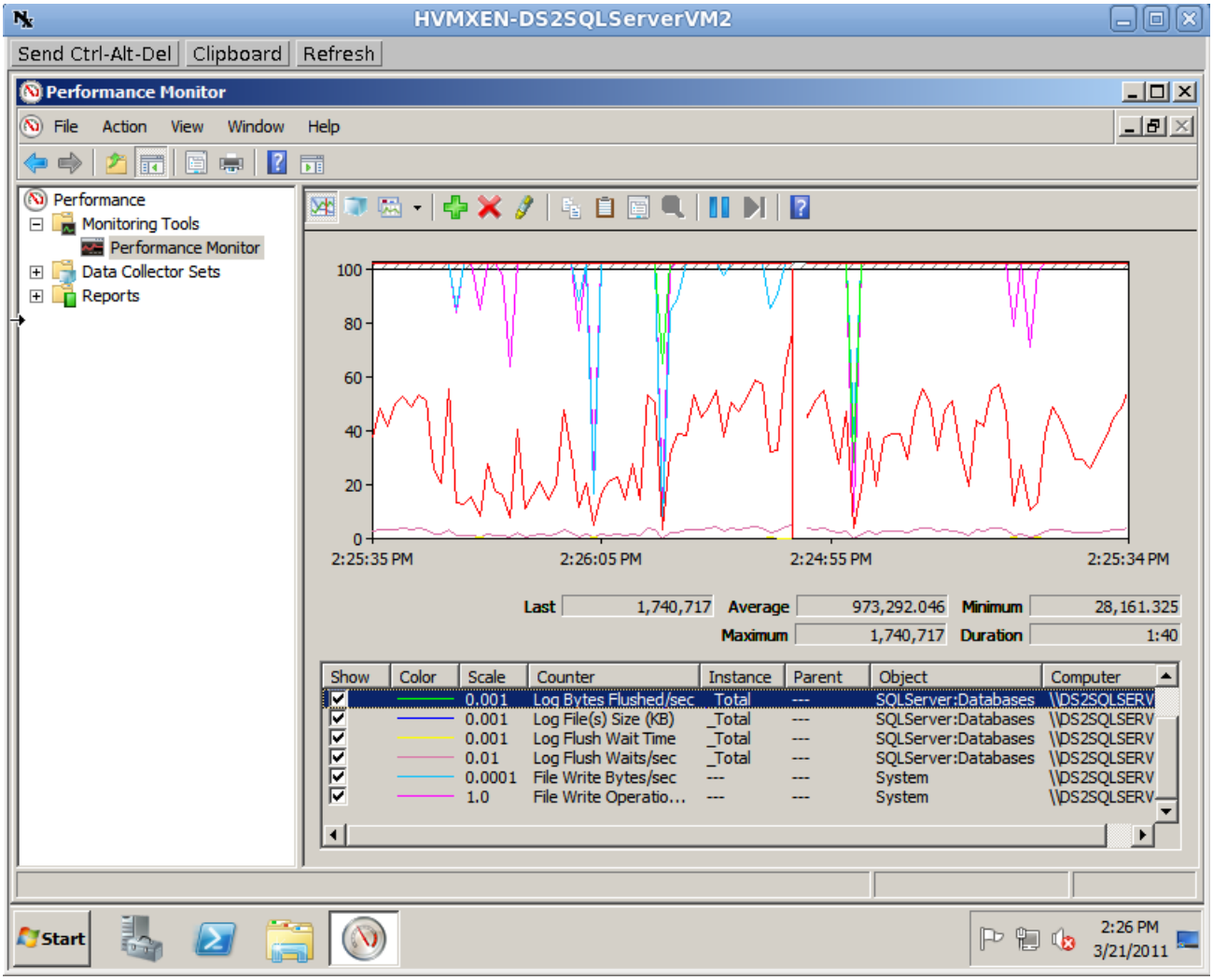
LOOKING AT APPLICATION I/O PROFILES IN THE VM

In some cases, an application is currently running on bare-metal, and a customer is planning to deploy it on Avance. We can determine which specific applications may benefit from 10Gb links by looking into their I/O profiles. In order to gather this data, we use Windows Performance Monitor, as illustrated in the diagram below. We suggest gathering the average and maximum values for the following statistics:

System:

File Write Bytes/sec

File Write Operations/sec



Be sure to collect the performance counters while the system is under its typical peak load. If several VM's share similar application profiles, it may suffice to capture the data for a single VM and multiply the numbers by the number of similar VM's. If VM's have different application profiles, then performance numbers should be captured for each VM. Now that we have the data collected, we are able to make the following calculations and comparisons:

DISK WRITE I/O THROUGHPUT

We compute the total disk write I/O throughput as the sum of the File Write Bytes/sec value over all VM's. We would expect to see a benefit from installing 10G links if this value exceeds 35 MB/second.

DISK WRITE I/O AVERAGE BLOCK SIZE

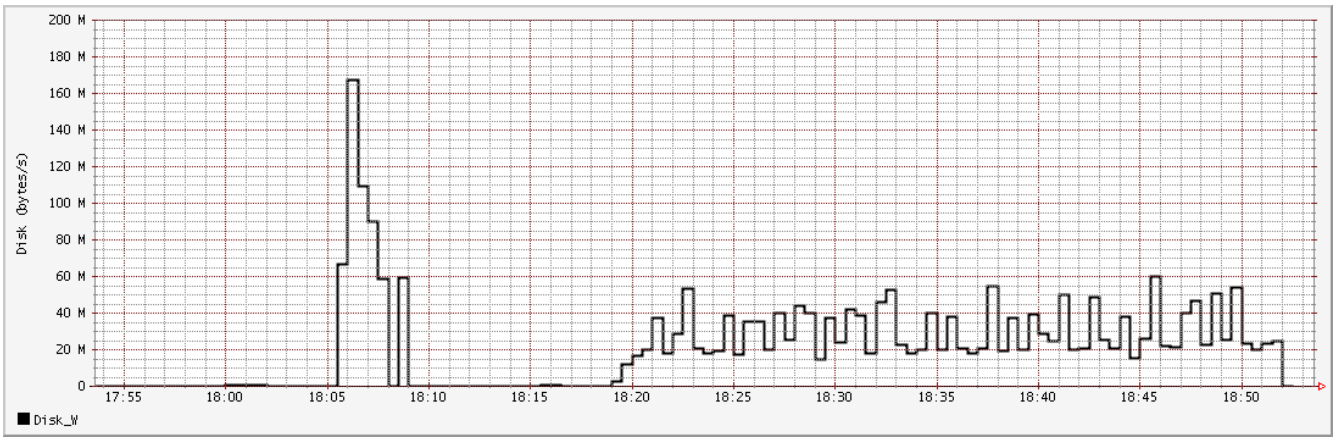
We also compute the average write I/O block size for a particular VM as the File Write Bytes/sec value divided by the File Write Operations/second. We compare this value to 16 KBytes. If the average write block size is greater than 16 KBytes, then a performance boost is more likely. If the average write block size is less than 8 KB, then a 10Gbit link is unlikely to benefit the performance of this VM's application. If the average block size falls between 8 and 16 KB, then a performance boost for this application is possible but not certain. (Although an application may perform writes in larger block sizes, the writes may get fragmented by the Avance software, so we cannot be certain about the performance boost.) As larger numbers of VM's are deployed, the likelihood of performance for applications characterized by small block writes increases.

DATABASE APPLICATIONS

Customers running databases are faced with a more complex situation. Whether a database will be accelerated by a faster replication link depends on the I/O profile of that system. The most straightforward way to predict whether an application will benefit from 10Gb links is to examine the current replication link usage (utilizing a 1Gb link).

EXAMPLE: SQL SERVER, SIX VM'S

The Disk I/O graph shown below shows the behavior of an Avance system with 10Gb links with six VM's running SQL Server. The initial peak of I/O between 18:05 to 18:10 corresponds to restoring a 1 GByte database. The period of I/O from 18:20 onwards shows the databases responding to queries. Clearly, when restoring data, this system is able to take advantage of the faster replication link. With a 1Gb link the I/O bandwidth would have been throttled back to approximately 75 MBytes/second, so the restore operation would have taken perhaps twice as long. However, under steady usage, this application does not generate a heavy I/O traffic pattern, so the steady-state use case does not benefit from the faster links. Additionally it is known that the average write block size during steady operation is 8 KBytes, which also suggests that the application is not likely to be accelerated by the faster link. Testing in this configuration with a 1 Gb link confirms that the steady state operation did not benefit significantly from the faster link.



CONCLUSION

Deploying 10Gb replication links in an Avance system will result in faster guest software installation, re-synchronization, and fail-over. Aside from the straightforward cases of large file copy, database restore, or other sustained write-intensive operations, predicting whether an application will benefit from 10Gb replication links requires studying the I/O characteristics of the system. Customers with questions about performance enhancements with 10Gb links are encouraged to observe the I/O utilization statistics in Avance and Windows, and consult the guidelines presented in this paper. Totaling the aggregate write I/O bandwidth across all shared mirrors, or capturing similar data at the Windows O/S level is likely to be the most reliable metric for predicting whether a particular system may experience a performance boost after installing 10Gb replication links. To simplify the decision criterion, the following table can be used as a guideline. Based upon the value obtained for the sum of the shared mirror I/O throughput, consult the following table to determine whether your current utilization falls in to the "green", "yellow", or "red" zones.

Throughput	**Zone**	**Application performance benefit from 10G links?**
< 35 Mbytes/sec	Green	Unlikely
35 - 50 Mbytes/sec	Yellow	Possible, but uncertain
> 50 Mbytes/sec	Red	Likely

NOTICE

The information contained in this document is subject to change without notice.

UNLESS EXPRESSLY SET FORTH IN A WRITTEN AGREEMENT SIGNED BY AN AUTHORIZED REPRESENTATIVE OF STRATUS TECHNOLOGIES, STRATUS MAKES NO WARRANTY OR REPRESENTATION OF ANY KIND WITH RESPECT TO THE INFORMATION CONTAINED HEREIN, INCLUDING WARRANTY OF MERCHANTABILITY AND FITNESS FOR A PURPOSE. Stratus Technologies assumes no responsibility or obligation of any kind for any errors contained herein or in connection with the furnishing, performance, or use of this document.

Software described in Stratus documents (a) is the property of Stratus Technologies Bermuda, Ltd. or the third party, (b) is furnished only under license, and (c) may be copied or used only as expressly permitted under the terms of the license.

Stratus documentation describes all supported features of the user interfaces and the application programming interfaces (API) developed by Stratus. Any undocumented features of these interfaces are intended solely for use by Stratus personnel and are subject to change without warning.

Stratus and the Stratus logo, Stratus Avance and the Stratus Avance logo are registered trademarks of Stratus Technologies Bermuda, Ltd.

The Stratus Technologies logo, the Stratus 24 x 7 logo, and ActiveService are trademarks of Stratus Technologies Bermuda, Ltd.

The registered trademark Linux is used pursuant to a sublicense from the Linux Mark Institute, the exclusive licensee of Linus Torvalds, owner of the mark on a world-wide basis.

All other trademarks are the property of their respective owners.

Stratus Technologies, Inc.
111 Powdermill Road
Maynard, Massachusetts 01754-3409