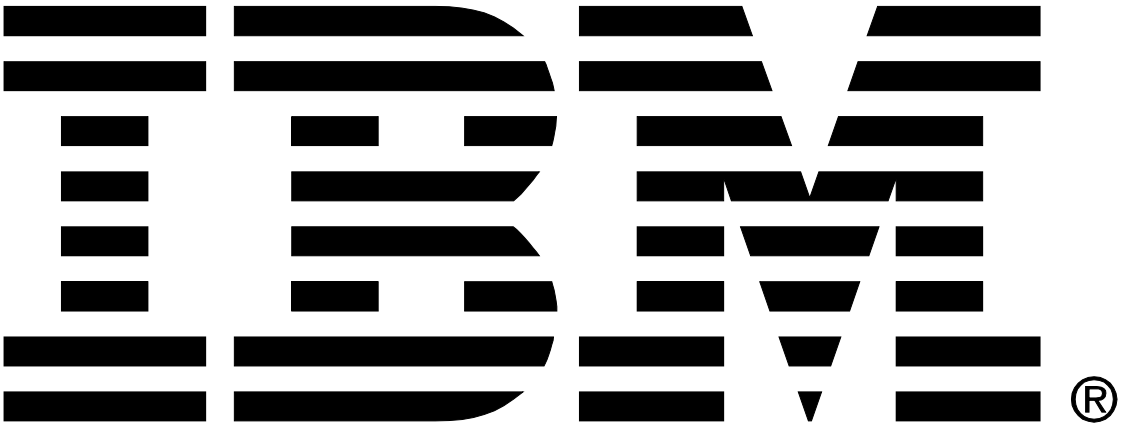


IBM MQ Appliance Performance Report

Model: M2004

Version 1.0 - November 2026

Sam Massey
IBM MQ Performance
IBM UK Laboratories
Hursley Park
Winchester
Hampshire



1 Notices

Please take Note!

Before using this report, please be sure to read the paragraphs on “disclaimers”, “warranty and liability exclusion”, “errors and omissions”, and the other general information paragraphs in the "Notices" section below.

First Edition, November 2026.

This edition applies to *IBM MQ Appliance* (and to all subsequent releases and modifications until otherwise indicated in new editions).

© Copyright International Business Machines Corporation 2026. All rights reserved.

Note to U.S. Government Users

Documentation related to restricted rights.

Use, duplication or disclosure is subject to restrictions set forth in GSA ADP Schedule contract with IBM Corp.

DISCLAIMERS

The performance data contained in this report was measured in a controlled environment. Results obtained in other environments may vary significantly.

You should not assume that the information contained in this report has been submitted to any formal testing by IBM.

Any use of this information and implementation of any of the techniques are the responsibility of the licensed user. Much depends on the ability of the licensed user to evaluate the data and to project the results into their own operational environment.

WARRANTY AND LIABILITY EXCLUSION

The following paragraph does not apply to the United Kingdom or any other country where such provisions are inconsistent with local law:

INTERNATIONAL BUSINESS MACHINES CORPORATION PROVIDES THIS PUBLICATION “AS IS” WITHOUT WARRANTY OF ANY KIND, EITHER EXPRESS OR IMPLIED, INCLUDING, BUT NOT LIMITED TO, THE IMPLIED WARRANTIES OF NON-INFRINGEMENT, MERCHANTABILITY, OR FITNESS FOR A PARTICULAR PURPOSE.

Some states do not allow disclaimer of express or implied warranties in certain transactions, therefore this statement may not apply to you.

In Germany and Austria, notwithstanding the above exclusions, IBM's warranty and liability are governed only by the respective terms applicable for Germany and Austria in the corresponding IBM program license agreement(s).

ERRORS AND OMISSIONS

The information set forth in this report could include technical inaccuracies or typographical errors. Changes are periodically made to the information herein; any such change will be incorporated in new editions of the information. IBM may make improvements and/or changes in the product(s) and/or the program(s) described in this information at any time and without notice.

INTENDED AUDIENCE

This report is intended for architects, systems programmers, analysts and programmers wanting to understand the performance characteristics of *IBM MQ Appliance*. The information is not intended as the specification of any programming interface that is provided by IBM. It is assumed that the reader is familiar with the concepts and operation of IBM MQ Appliance.

LOCAL AVAILABILITY

References in this report to IBM products or programs do not imply that IBM intends to make these available in all countries in which IBM operates. Consult your local IBM representative for information on the products and services currently available in your area.

ALTERNATIVE PRODUCTS AND SERVICES

Any reference to an IBM product, program, or service is not intended to state or imply that only that IBM product, program, or service may be used. Any functionally equivalent product, program, or service that does not infringe any IBM intellectual property right may be used instead. However, it is the user's responsibility to evaluate and verify the operation of any non-IBM product, program, or service.

USE OF INFORMATION PROVIDED BY YOU

IBM may use or distribute any of the information you supply in any way it believes appropriate without incurring any obligation to you.

TRADEMARKS AND SERVICE MARKS

The following terms used in this publication are trademarks of their respective companies in the United States, other countries or both:

- **IBM Corporation :** IBM

Other company, product, and service names may be trademarks or service marks of others.

EXPORT REGULATIONS

You agree to comply with all applicable export and import laws and regulations.

2 Contents

1 Notices 2

2 Contents 4

3 Introduction 5

4 Comparison of M2004A to M2003A 8

5 Comparison of M2004A to M2002A 11

6 Request/Responder Scenario 12

6.1 Test Scenario C1 – 10 Applications per QM, 1 QM, Non-persistent 14

6.2 Test Scenario C2 – 10 applications per QM, 1 QM, Persistent 15

6.3 Test Scenario C3 – 10 applications per QM, 10 QM, Non-persistent 16

6.4 Test Scenario C4 – 10 applications per QM, 10 QM, Persistent 17

7 Connection Scaling 18

7.1 Connection Test 18

8 HA Scenarios 19

8.1 Test Scenario HA1 – 10 Applications per QM, 1 QM, Persistent 20

8.2 Test Scenario HA2 – 10 applications per QM, 10 QM, Persistent 21

8.3 How does HA perform over larger distances? 22

9 DR Scenarios 25

9.1 Test Scenario DR1 – 10 Applications per QM, 1 QM, Persistent 26

9.2 Test Scenario DR2 – 10 Applications per QM, 10 QM, Persistent 27

9.3 How does DR perform over larger distances? 28

10 HA and DR Scenarios 29

11 Additional M2004A vs M2004B scenarios 30

12 Filesystem encryption performance 36

12.1 Single QM 36

12.2 Multiple QM 38

13 TLS 41

14 AMS 44

15 Frequently Asked Questions 45

16 Appendix A – Client machine specification 46

17 Appendix B – QM Configuration 46

3 Introduction

This performance report at version 1.0 contains performance data based on the MQ Appliance models M2004A and M2004B. The M2004 model is based on new hardware and is designed for optimal performance; it features an updated CPU and faster RAM, faster NVMe storage and increased 100Gb connectivity. This report covers standalone, HA and DR messaging performance, TLS performance and includes the following highlights:

- The M2004A offers up to 25% higher throughput than M2003A in Non-persistent messaging scenarios. See section 4
- Up to 20% higher throughput over M2003A in Persistent messaging scenarios. See section 4
- Up to 30% higher throughput over M2003A in HA Persistent messaging scenarios. See section 4
- Over 600,000 round trips/second peak messaging rate achieved in a Non-persistent messaging scenario, equating to over 2.4 million MQ put/get operations per second. See section 6.3
- Over 200,000 round trips/second peak messaging rate achieved in an HA enabled scenario equating to over 800,000 MQ put/get operations per second. See section 8.2
- Up to 4x higher persistent messaging throughput than M2002A. See section 5

The M2004 hardware components and how they compare to the previous model M2003 are shown below:

Model	M2003A	M2003B	M2004A	M2004B
CPU	2x16 Core HT (2.9GHz)	1x8 Core HT (2.9GHz)	2x16 Core HT (2.8GHz)	1x8 Core HT (2.8GHz)
RAM	256GB	256GB	256GB	256GB
Storage	6.4TB NVMe SSD	3.2TB NVMe SSD	6.4TB NVMe SSD	3.2TB NVMe SSD
IO Subsystem	RAID 10	RAID 10	RAID 10	RAID 10
Workload and replication network connectivity	8x1Gb 4x10Gb 4x40Gb 2x100Gb	8x1Gb 4x10Gb 4x40Gb 2x100Gb	8x1Gb 4x10Gb 2x40Gb 4x100Gb	8x1Gb 4x10Gb 2x40Gb 4x100Gb
Management	2x1Gb	2x1Gb	2x1Gb	2x1Gb
Chipset	Icelake	Icelake	Emerald Rapids	Emerald Rapids
RAID	PCIe 4 x8 16GB/s	PCIe 4 x8 16GB/s	PCIe 4 x8 16GB/s	PCIe 4 x8 16GB/s

The MQ appliance combines all the core MQ functionality with the convenience, ease of install and simplified maintenance of an appliance.

There are local disks within the appliance to enable efficient persistent messaging by the local Queue Managers. The four 3.2TB NVMe SSD drives are configured in a RAID10 configuration so that data is protected should one of the drives suffer a failure. High Availability (HA) may be achieved by the pairing of two MQ appliances which results in the Queue Manager (QM) log and queue files being distributed synchronously across the pair of appliances. Disaster Recovery (DR) may be achieved by the addition of a remote appliance to which QM data is distributed synchronously or asynchronously. This report will also illustrate the HA and DR capabilities of the new model.

The MQ appliance can be purchased in two variants: M2004A and M2004B. There are two main differences for the M2004B as highlighted in the table above, reduced CPU capacity and reduced filesystem storage space.

As before, you can purchase an upgrade to convert an M2004B appliance to an M2004B+ appliance, which has the same capacity as an M2004A appliance.

The majority of the tests use the M2004A variant of the MQ Appliance and this is the default hardware unless stated otherwise. A number of tests were also conducted using the M2004B variant and provide comparative data points to the main testing to provide appropriate capacity planning information.

The M2004A and M2004B appliances are supplied with 4x100Gb Ethernet network links, 2x40Gb Ethernet network links, 4x10Gb Ethernet network links and 8x1Gb Ethernet network links. If the appliances are configured for redundant HA, 2x1Gb links would be reserved for use by the appliance in addition to another interface to perform the HA replication (this can be configured to use any of the interfaces available or indeed an aggregated interface), leaving a potential total of 266Gb for customer workloads (If a 40Gb port is used for replication). In NonHA mode, all 288Gb connectivity can be utilised for workload traffic. There are a further two separate 1Gb links that are explicitly reserved for appliance administration.

There are two modules that support 100Gb network connectivity with two ports available in each. There is a capacity limit of 100Gb per module. This report utilises 2 of the 100Gb links for workload traffic; and 1 100Gb port for replication traffic.

There is one module that supports 40Gb network connectivity with two ports available. There is a capacity limit of 40Gb across those two ports.

All the scenarios featured in this report utilise Request Responder messaging scenarios and the published messaging rate is measured in round trips/sec, which involves 2 message puts and 2 message gets. If you are only utilising one-way messaging (using a message sender, queue and message receiver to perform 1 message put and 1 message get), and you can avoid queue-lock contention, then you may achieve up to double the published rates.

The version of the MQ Appliance as tested in this report is M2004A MQ 10.0.0 and where a comparison is made to the restricted appliance configuration, this uses the MQ Appliance M2004B MQ 10.0.0

4 Comparison of M2004A to M2003A

The new model has provided significant performance and capacity benefits. The majority of scenarios should benefit from either the updated CPU, faster RAM and increased IO performance.

The M2003 model itself saw significant performance improvements over the M2002, and the M2004 again provides further improvement. The graphs in this section will illustrate that uplift by comparing the M2004A and M2003A models.

The scenarios from sections 6.3, 6.4 and 8.2 are used to compare the performance of M2004A with M2003A.

The following graph shows the improvement provided in Non-persistent messaging:

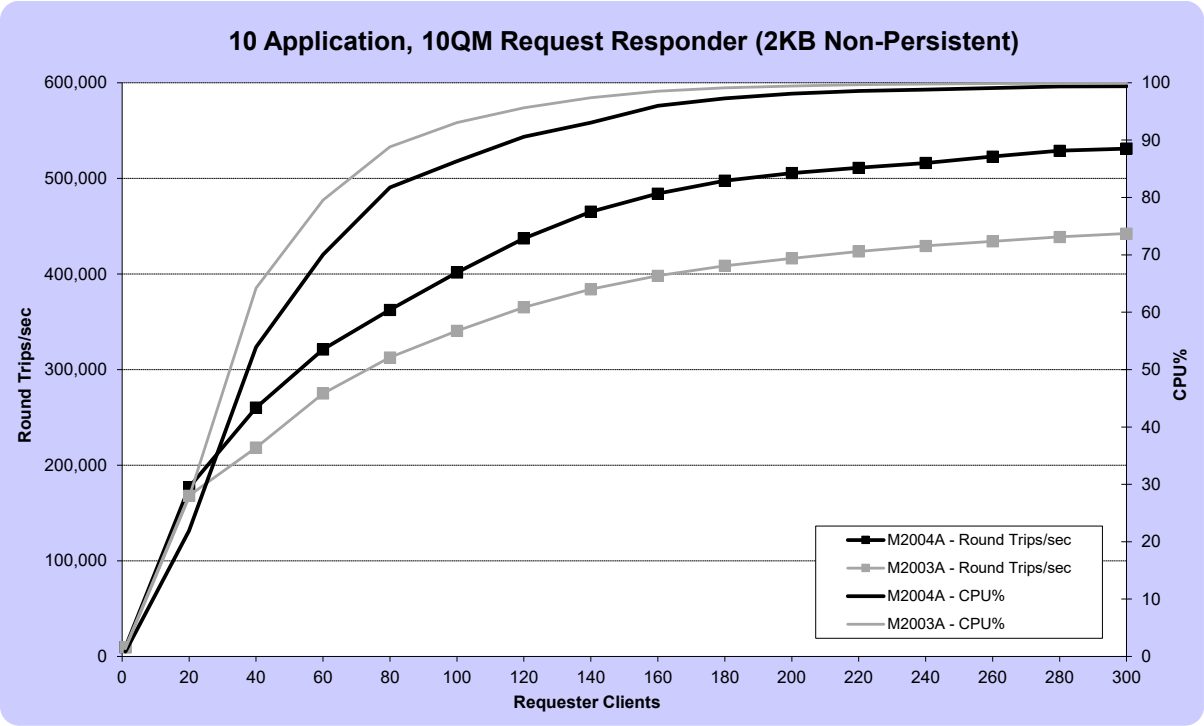


FIGURE 1 - PERFORMANCE RESULTS FOR 2KB, 10QM NON-PERSISTENT MESSAGING, M2004A VS M2003A

Figure 1 (and the tabulated data below) shows that the M2004A appliance is capable of up to 25% increased throughput of Non-persistent messaging when compared with M2003A in CPU limited scenarios.

The additional 100Gb network connectivity provided in the M2004 appliance offers increased connectivity options and larger overall capacity for messaging bandwidth as can be seen in the table below.

Test	M2004A			M2003A		
	Max Rate*	CPU%	Clients	Max Rate*	CPU%	Clients
10Q 10QM Request Responder (256b Non-persistent)	607,134	98.51	240	527,399	99.8	240
10Q 10QM Request Responder (2KB Non-persistent)	531,067	99.37	300	442,280	99.93	300
10Q 10QM Request Responder (20KB Non-persistent)	401,773	99.86	300	217,535	88.77	140
10Q 10QM Request Responder (200KB Non-persistent)	56,789	62.47	70	28,027	34.85	60

*Round trips/sec

TABLE 1 - PEAK RATES FOR 10QM NON-PERSISTENT MESSAGING, M2004A VS M2003A

The following graph shows the improvement provided in Persistent messaging:

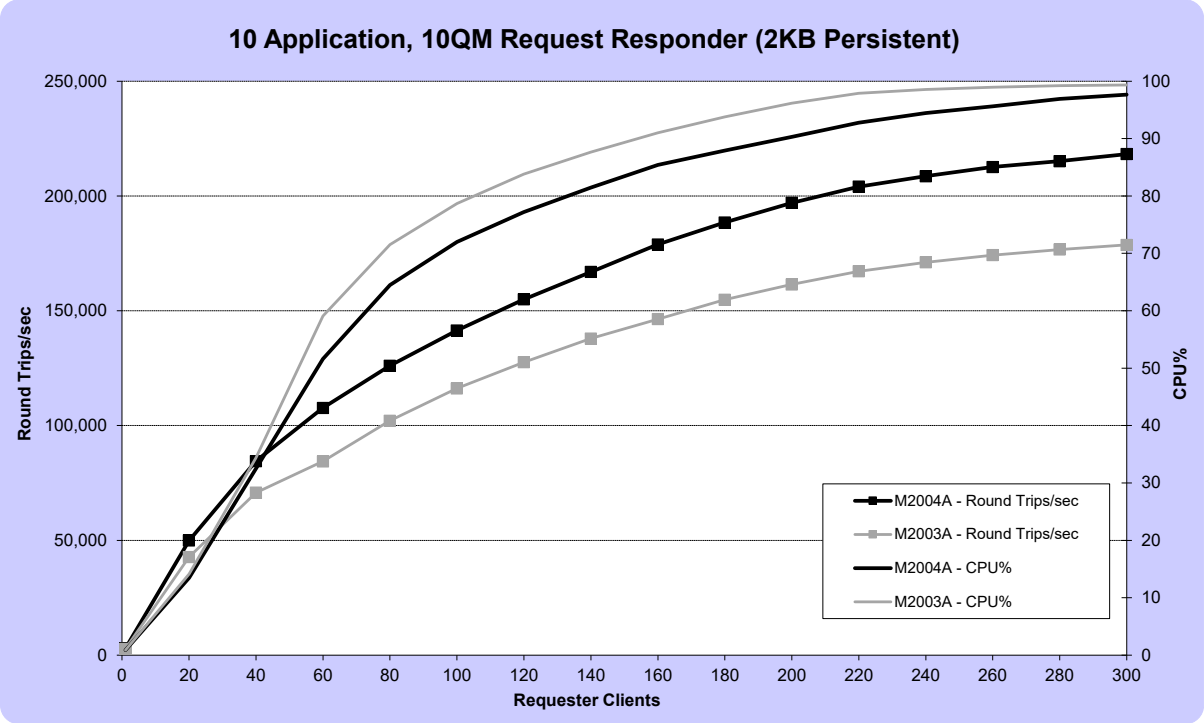


FIGURE 2 - PERFORMANCE RESULTS FOR 2KB, 10QM PERSISTENT MESSAGING, M2004A VS M2003A

Figure 2 shows that there is up to 20% performance improvement when using the 2KB message size. A similar improvement is also seen for 20KB and 200KB message sizes, as shown in the table below.

Test	M2004A			M2003A		
	Max Rate*	CPU%	Clients	Max Rate*	CPU%	Clients
10Q 10QM Request Responder (256b Persistent)	240,354	96.84	300	196,819	99.31	300
10Q 10QM Request Responder (2KB Persistent)	218,257	97.67	300	178,693	99.32	300
10Q 10QM Request Responder (20KB Persistent)	151,930	92.16	300	123,875	97.47	300
10Q 10QM Request Responder (200KB Persistent)	16,156	22.79	110	15,728	32.95	120

*Round trips/sec

TABLE 2 - PEAK RATES FOR 10QM PERSISTENT MESSAGING, M2004A VS M2003A

The following graph shows the improvement provided in Persistent messaging with Queue Managers enabled for HA:

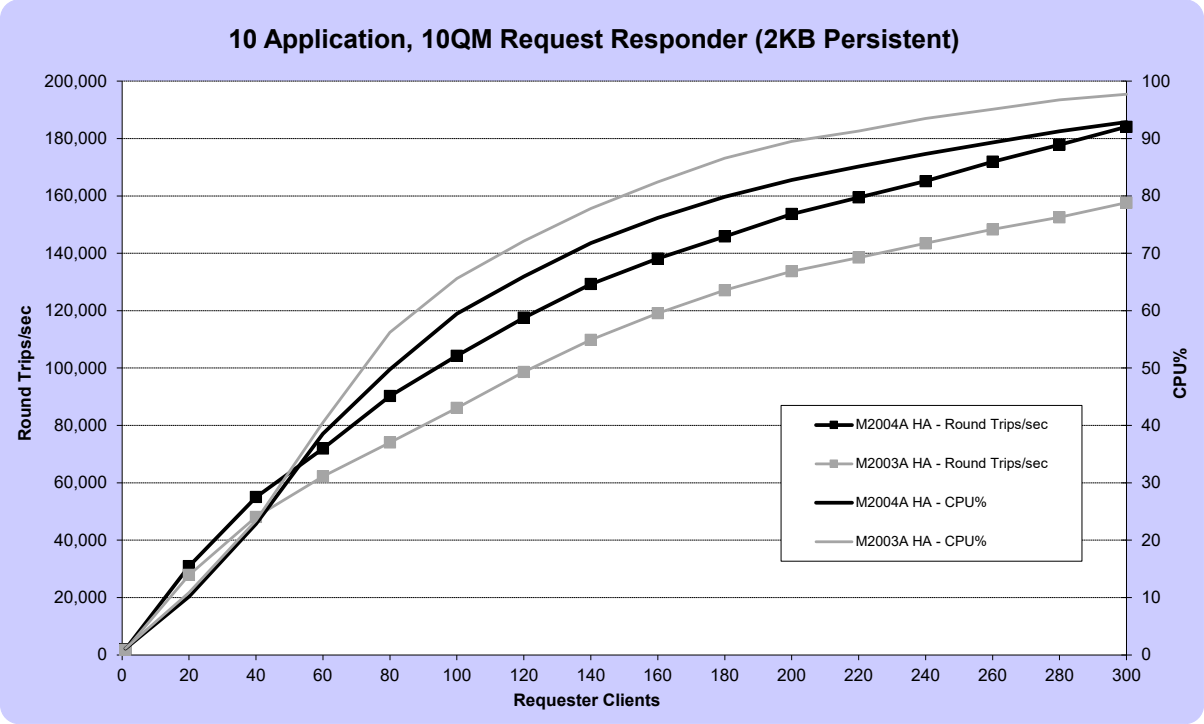


FIGURE 3 - PERFORMANCE RESULTS FOR 2KB, 10QM HA PERSISTENT MESSAGING, M2004A VS M2003A

Figure 3 shows that there is more than 15% performance improvement when using the 2KB message size. For message sizes of 20K and above, there is an improvement of up to 30%.

Test	M2004A HA				M2003A HA			
	Max Rate*	CPU%	Clients	Latency#	Max Rate*	CPU%	Clients	Latency#
10Q 10QM Request Responder (256b Persistent)	203,927	92.2	300	0.5	172,432	97.42	300	0.5
10Q 10QM Request Responder (2KB Persistent)	184,089	92.89	300	0.5	157,645	97.71	300	0.6
10Q 10QM Request Responder (20KB Persistent)	116,647	80.9	300	0.6	95,542	85.49	300	0.7
10Q 10QM Request Responder (200KB Persistent)	14,788	26.15	120	1.2	11,296	23.97	110	1.4

*Round trips/sec
#Single thread round trip latency (ms)

TABLE 3 - PEAK RATES FOR 10QM PERSISTENT MESSAGING, M2004A VS M2003A

5 Comparison of M2004A to M2002A

With the end of support date for the M2002A MQ Appliance set at 30th September 2027, it is anticipated that many M2002A users may seek to migrate to the new M2004 MQ Appliance.

The following points indicate the likely improvement in Performance that the M2004A offers in comparison to the M2002A:

- The M2004A offers up to 2x more throughput than the performance of M2002A in Non-persistent messaging scenarios. (2KB)
- Up to 4x increase in performance over M2002A in Persistent messaging scenarios. (20KB)
- Up to 4x increase in performance over M2002A in HA Persistent messaging scenarios. (20KB)

This is based on data collected at V9.4.5 for the MQ M2002 Appliance and V10.0.0 for the MQ M2004 Appliance.

6 Request/Responder Scenario

The scenario that will be used in this report reflects the most common, anticipated usage patterns for the MQ appliance and provides guidance for those customers performing capacity planning or migration activities.

Each test was initially conducted and graphs produced using a 2KB (2048 byte) message size. Additional tests were also conducted using 256 byte, 20KB and 200KB to provide further data for capacity planning and are found in the accompanying data table.

As customers replace their existing MQ QM infrastructure, they may consolidate their MQ configuration from separate MQ QM servers (possibly running on different hardware and different MQ Versions) onto a single MQ appliance. They may have a mix of applications tightly bound to their existing QM and also a number of applications that connect using the MQ client API. To migrate to the MQ appliance all applications will need to connect via the MQ client API.

The following tests use MQ client connections and present the performance of MQ as deployed on the Appliance.

The test scenario in Figure 4 is a Request Responder scenario that simulates several applications that interact with a single QM. A request queue and a reply queue will be created for each application, so ten pairs of queues are created for this test. One or more requester applications will send messages to one of the application request queues and will wait for a reply on the associated reply queue. Responder applications will listen for messages on the request queues before sending them to the correct reply queue.

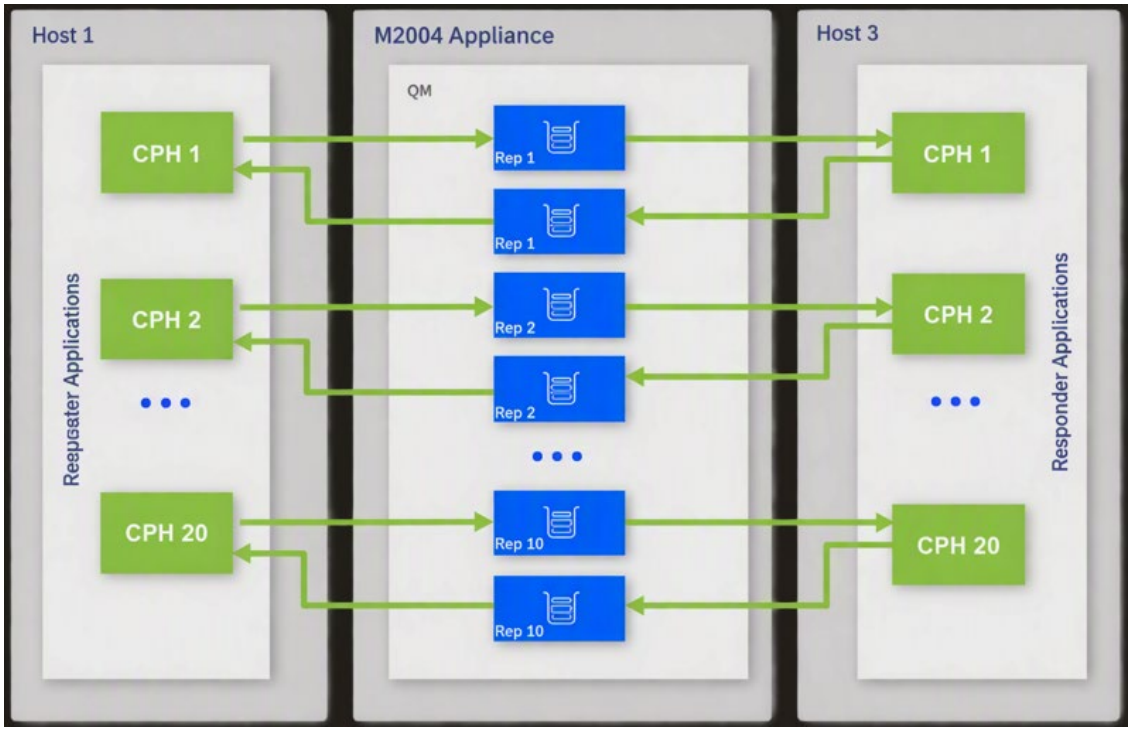


Figure 4 - Requester-responder with remote queue manager on MQ Appliance

Subsequent requester applications will send and receive messages from the set of application queues on a round-robin basis i.e. distributing the messages produced and consumed across the set of application queues.

Results are presented for various numbers of producer threads distributed across the 10 applications (using 10 pairs of queues), 300 fixed responder threads (30 responders per request queue) will send the replies to the appropriate reply queue, and the report will show the message rates achieved (in round trips/second) as the number of producers is increased.

For the 10QM tests, there are 10 QM with 10 applications per QM (again using 10 pairs of queues). There are still 300 overall responder threads, but as we now have 100 pairs of queues, we have 3 responders per request queue.

6.1 Test Scenario C1 – 10 Applications per QM, 1 QM, Non-persistent

The following graph shows how the scenario detailed in section 6 performs with Non-persistent messaging against a single QM.

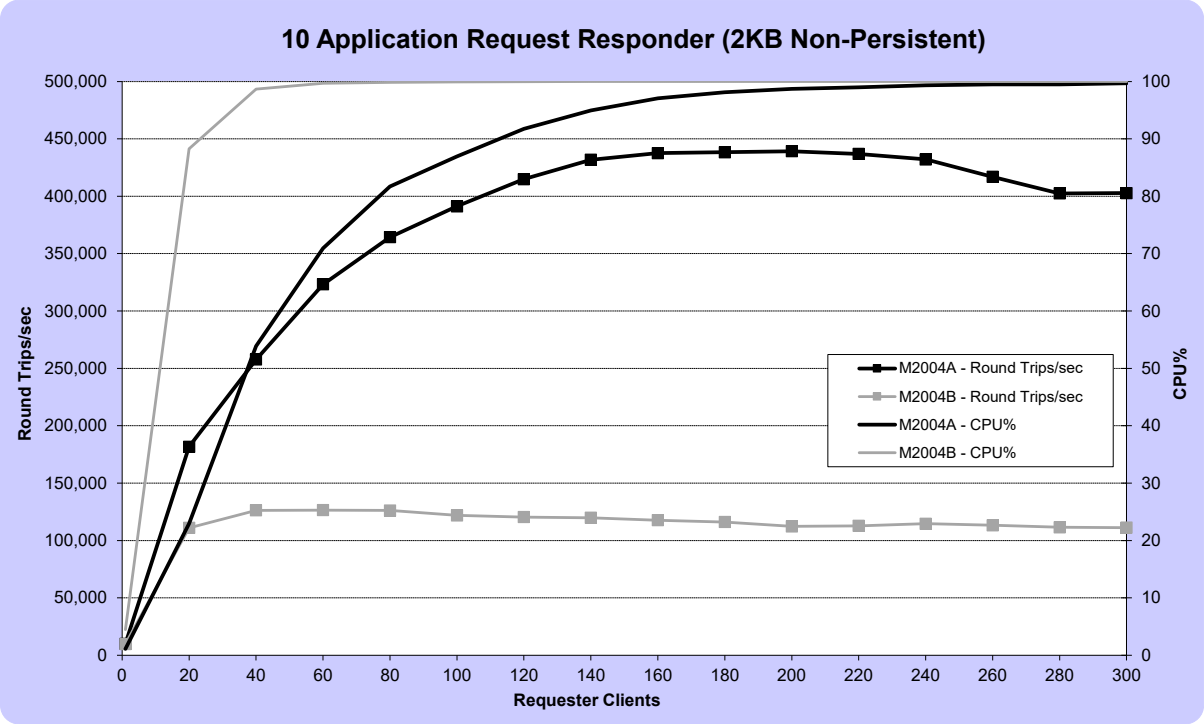


FIGURE 5 – PERFORMANCE RESULTS FOR 2KB NON-PERSISTENT MESSAGING

Figure 5 shows how by increasing the workload on the appliance (by increasing the number of concurrent requester clients), the throughput rate increases until the CPU capacity of the appliance is exhausted.

The M2004A appliance can achieve approximately 3 times the throughput of the M2004B appliance.

Test	M2004A			M2004B		
	Max Rate*	CPU%	Clients	Max Rate*	CPU%	Clients
10Q Request Responder (256b Non-persistent)	508,337	97.55	180	148,164	99.54	60
10Q Request Responder (2KB Non-persistent)	439,223	98.69	200	126,499	99.68	60
10Q Request Responder (20KB Non-persistent)	338,233	98.46	180	97,901	98.74	40
10Q Request Responder (200KB Non-persistent)	56,616	67.98	70	25,149	99.35	40

*Round trips/sec

TABLE 4 - PEAK RATES FOR NON-PERSISTENT MESSAGING

6.2 Test Scenario C2 – 10 applications per QM, 1 QM, Persistent

This test repeats the test C1 featured in section 6.1, but utilises persistent messaging on the appliances local RAID10 disk subsystem.

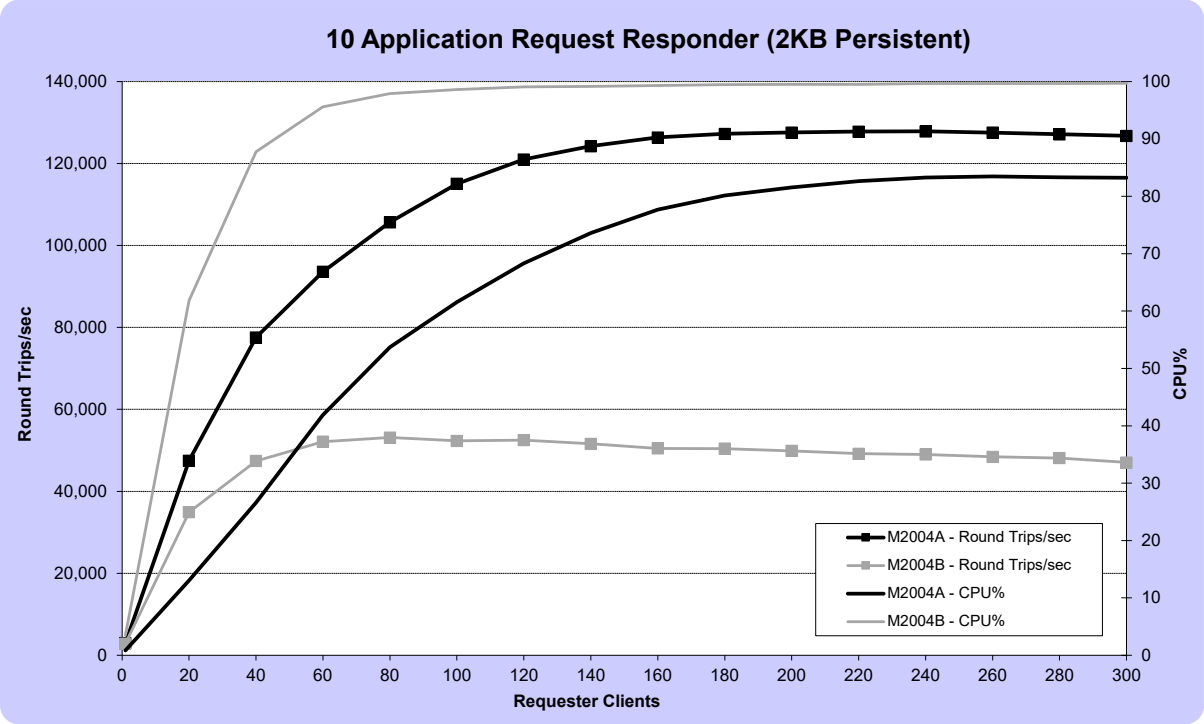


FIGURE 6 - PERFORMANCE RESULTS FOR 2KB PERSISTENT MESSAGING

Figure 6 shows that as the workload increases, a maximum throughput is achieved (~128,000 Round trips/sec for 2KB message size) and the limits of the local disk subsystem have become the limiting factor for a single QM utilising a single writing thread.

If using a message size of 2KB, the M2004A appliance can achieve approximately 2.5 times the persistent throughput of the M2004B appliance.

Test	M2004A			M2004B		
	Max Rate*	CPU%	Clients	Max Rate*	CPU%	Clients
10Q Request Responder (256b Persistent)	134,527	83.09	240	58,673	97.95	80
10Q Request Responder (2KB Persistent)	127,828	83.27	240	53,145	97.88	80
10Q Request Responder (20KB Persistent)	77,127	54.32	160	41,099	96.41	120
10Q Request Responder (200KB Persistent)	9,364	10.74	30	8,231	49.04	25

*Round trips/sec

TABLE 5 – PEAK RATES FOR PERSISTENT MESSAGING

6.3 Test Scenario C3 – 10 applications per QM, 10 QM, Non-persistent

This test is equivalent to test C1 featured in section 6.1 with 10QM instead of 1QM. A total of 100 applications will be distributed across the 10 QM.

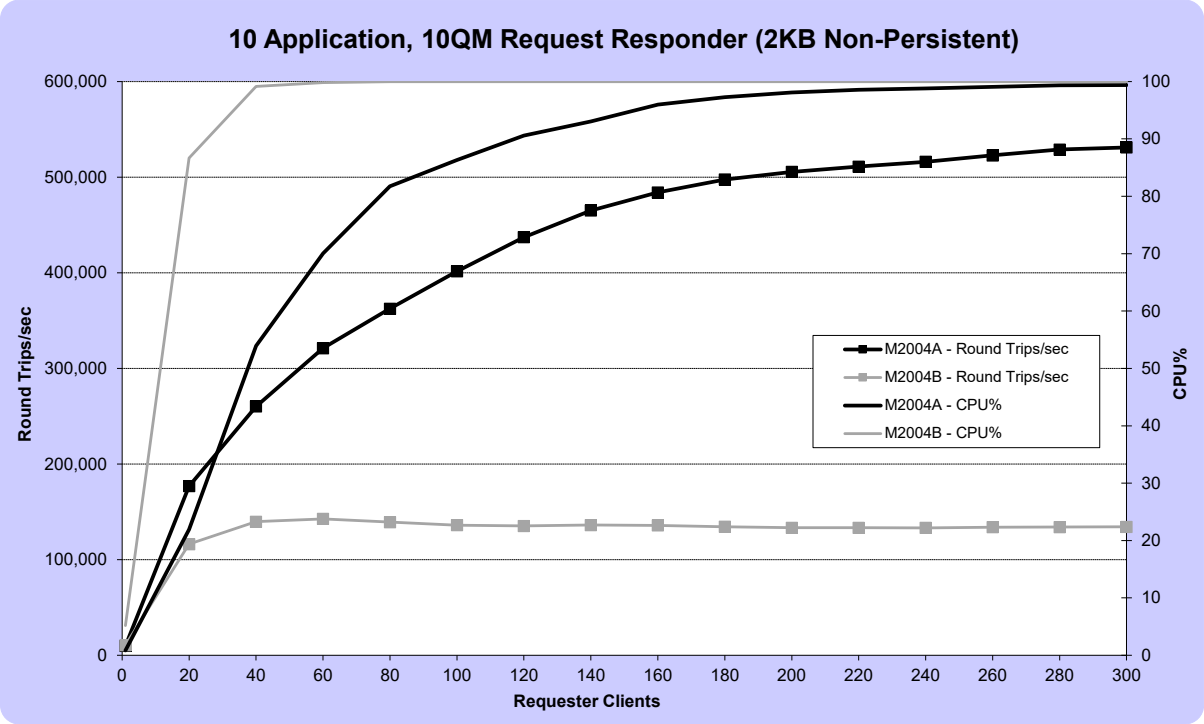


FIGURE 7 - PERFORMANCE RESULTS FOR 2KB, 10QM NON-PERSISTENT MESSAGING

Figure 7 shows that similar or improved performance can be obtained when running Non-persistent messaging through 10QM as compared with the single QM scenario.

The M2004A appliance can achieve approximately 3 times the throughput of the M2004B appliance.

Test	M2004A			M2004B		
	Max Rate*	CPU%	Clients	Max Rate*	CPU%	Clients
10Q 10QM Request Responder (256b Non-persistent)	607,134	98.51	240	169,761	98.88	40
10Q 10QM Request Responder (2KB Non-persistent)	531,067	99.37	300	142,669	99.82	60
10Q 10QM Request Responder (20KB Non-persistent)	401,773	99.86	300	108,073	99.9	60
10Q 10QM Request Responder (200KB Non-persistent)	56,789	62.47	70	26,903	99.73	40

*Round trips/sec

TABLE 6 - PEAK RATES FOR 10QM NON-PERSISTENT MESSAGING

6.4 Test Scenario C4 – 10 applications per QM, 10 QM, Persistent

This test repeats the test C3 featured in section 6.3, but utilises persistent messaging on the appliances local RAID10 disk subsystem. The graph and the accompanying data table illustrate that to utilise all of the available IO capacity on the appliance, multiple QM are required.

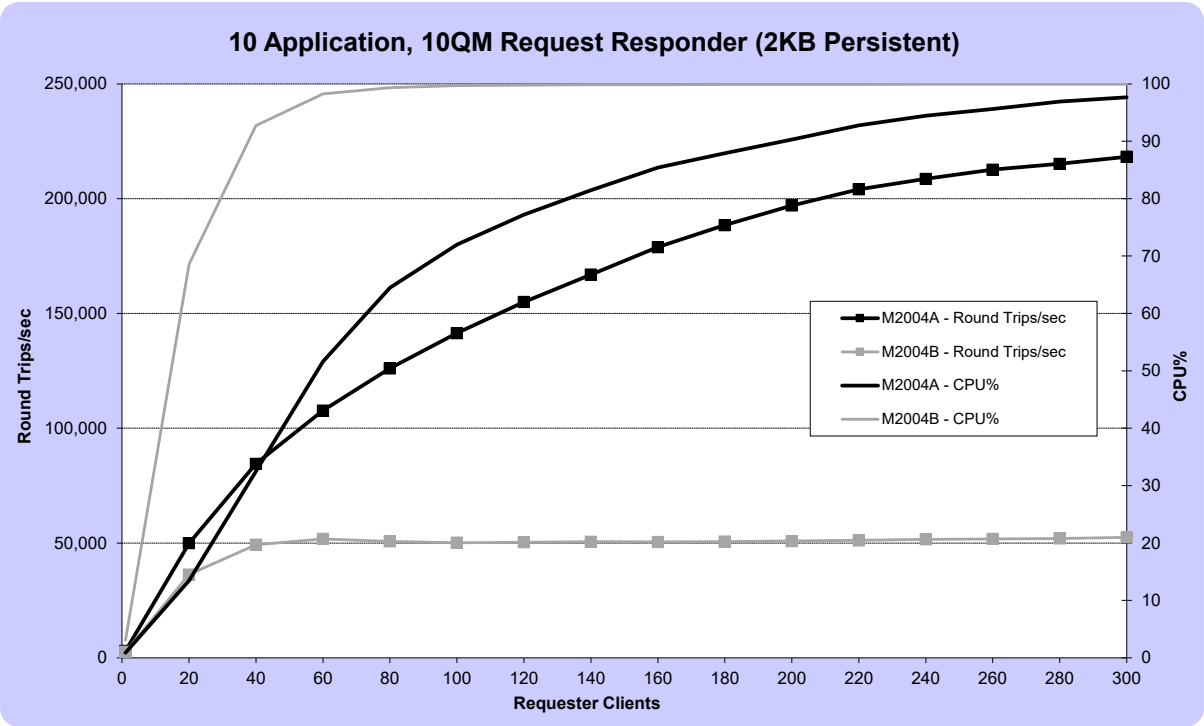


FIGURE 8 - PERFORMANCE RESULTS FOR 2KB, 10QM PERSISTENT MESSAGING

Figure 8 shows that when we have multiple QM performing persistent messaging, the peak messaging rate obtainable on the M2004A is over 210,000 Round trips/sec for 2KB message size.

If using a message size of 2KB, the M2004A appliance can achieve approximately 4 times the persistent throughput of the M2004B appliance.

Test	M2004A			M2004B		
	Max Rate*	CPU%	Clients	Max Rate*	CPU%	Clients
10Q 10QM Request Responder (256b Persistent)	240,354	96.84	300	57,554	99.87	300
10Q 10QM Request Responder (2KB Persistent)	218,257	97.67	300	52,462	99.93	300
10Q 10QM Request Responder (20KB Persistent)	151,930	92.16	300	42,643	98.53	60
10Q 10QM Request Responder (200KB Persistent)	16,156	22.79	110	14,283	98.36	80

*Round trips/sec

TABLE 7- PEAK RATES FOR 10QM PERSISTENT MESSAGING

7 Connection Scaling

The scaling measurement in this section is designed to demonstrate a scenario where there are a larger number of clients attached. Whereas the previous tests are optimised for throughput, this test defines a scaling challenge to test from a performance perspective.

7.1 Connection Test

This test uses the Requester Responder workload as described in section 6. The requester applications are rated at 1 message every 100 seconds and 60,000 client bound requester applications are connected as fast as possible to determine the overall connection time for those clients to the MQ Appliance.

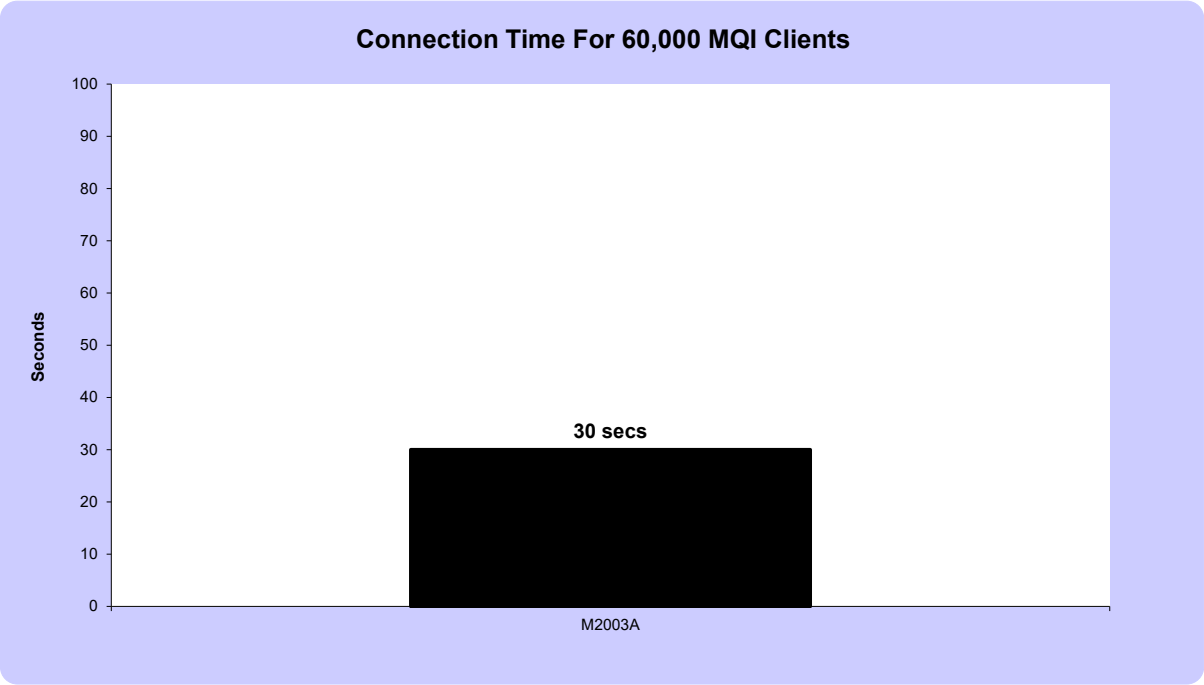


FIGURE 9 – PERFORMANCE RESULTS FOR MQI CLIENT CONNECTION TEST

The M2004 appliance with its range of 10,40 and 100Gb networking supports rapid connections from concurrent client bound applications and the MQ appliance can support 60,000 clients initiating a connection to a single queue manager in approximately 30 seconds.

8 HA Scenarios

High Availability (HA) can be enabled by pairing two MQ Appliances together to provide continuous availability in the event that one of the appliances suffers a failure. The Queue Manager (QM) log and queue files are synchronously replicated across the pair of appliances.

If separate networks (and switches) are used to connect the pair of appliances, then the pair can also continue to operate in the event of a partial network outage.

To ensure clients reconnect to the QM on either of the pair of appliances, the clients should be made aware of the IP addresses assigned to the workload interfaces of both appliances; or a Virtualised IP address in the case that a suitable load balancer component is employed; or a floating IP if it is configured on the appliance for the QM.

To illustrate the performance profile of enabling the HA infrastructure, tests will be performed on two of the scenarios featured earlier in the report.

1) Request Responder 1QM Persistent (Test C2)
2) Request Responder 10QM Persistent (Test C4)

Each test will be conducted with both a standalone QM and a QM incorporated into an appliance HA group, so that the cost of the synchronous replication can be evaluated.

This section utilises the following connections:

Primary Appliance	Secondary Appliance	Notes
eth13	eth13	Connected directly between appliances with 1Gb copper patch cable
eth17	eth17	Connected directly between appliances with 1Gb copper patch cable
eth41	eth41	Connected directly between appliances with 100Gb Fibre cable
eth40		Workload driven via this 100Gb interface
eth42		Workload driven via this 100Gb interface

8.1 Test Scenario HA1 – 10 Applications per QM, 1 QM, Persistent

This test is identical to test C2 in section 6.2 and is presented here with results from running tests against a standalone QM and against a QM that is included in an HA group.

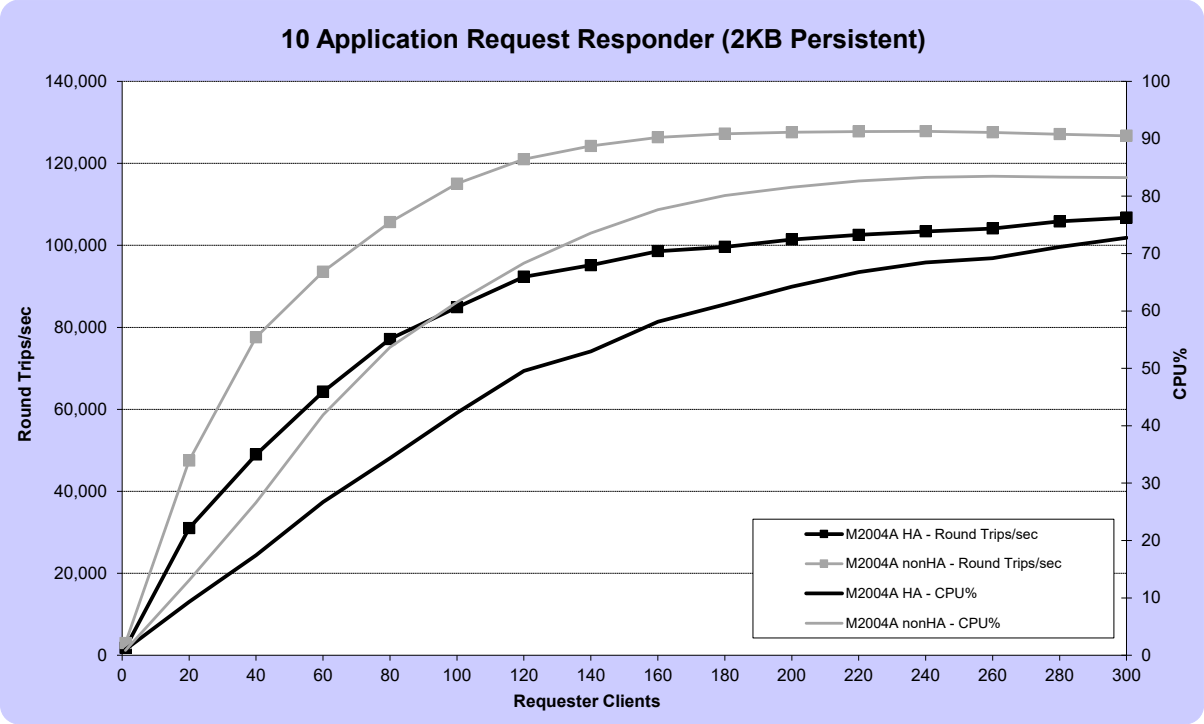


FIGURE 10 – PERFORMANCE RESULTS FOR 2KB PERSISTENT MESSAGING

Figure 10 shows that by enabling HA capability, the maximum throughput achieved in this single queue manager scenario with a 2KB message size is reduced by less than 20%. A small penalty for ensuring your data is synchronously written to both primary and secondary appliances in the HA group.

Test	M2004A HA				M2004A nonHA			
	Max Rate*	CPU%	Clients	Latency#	Max Rate*	CPU%	Clients	Latency#
10Q Request Responder (256b Persistent)	133,940	84.76	280	0.5	134,527	83.09	240	0.3
10Q Request Responder (2KB Persistent)	106,746	72.74	300	0.6	127,828	83.27	240	0.3
10Q Request Responder (20KB Persistent)	29,164	23.75	200	0.6	77,127	54.32	160	0.4
10Q Request Responder (200KB Persistent)	3,658	5.59	30	1.1	9,364	10.74	30	0.8

*Round trips/sec
#Single thread round trip latency (ms)

TABLE 8 - PEAK RATES FOR PERSISTENT MESSAGING

8.2 Test Scenario HA2 – 10 applications per QM, 10 QM, Persistent

This test repeats test C4 and is presented here with results from running tests against a standalone set of Queue Managers and against a set of Queue Managers that are included in an HA group.

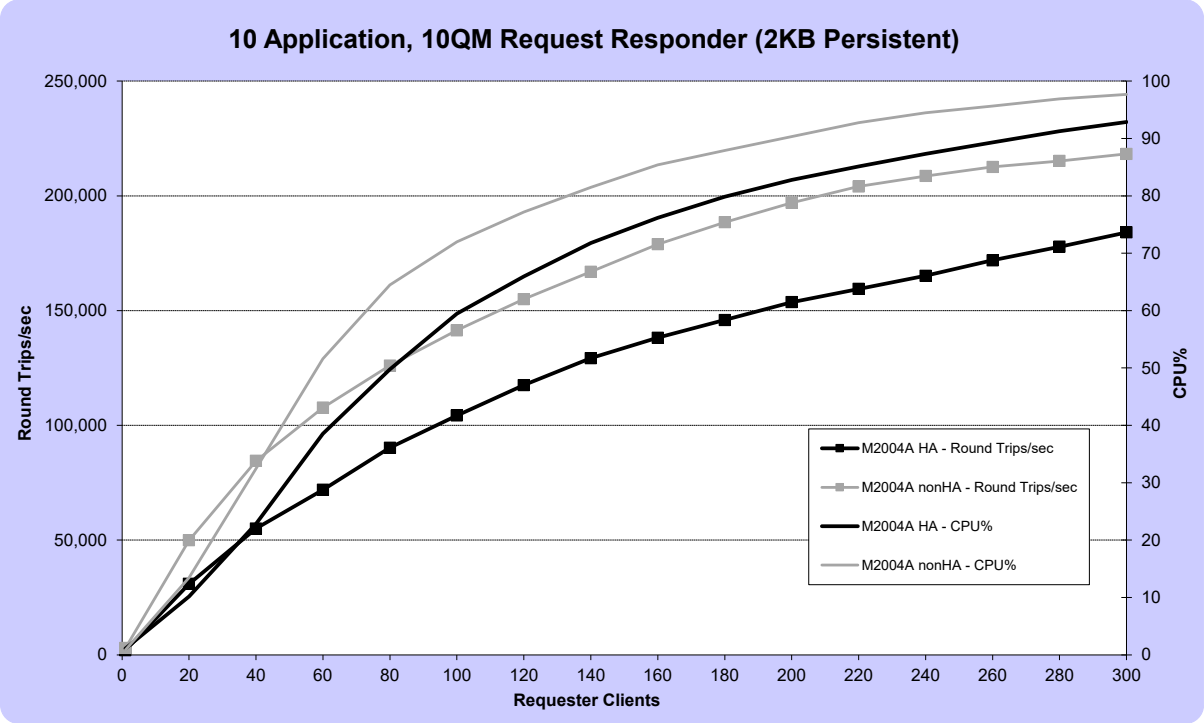


FIGURE 11 – PERFORMANCE RESULTS FOR 2KB, 10QM PERSISTENT MESSAGING

Figure 11 shows that when we have multiple QM performing 2KB persistent messaging across a pair of HA appliances, the messaging rate is only 15% less than when distributed across a set of non HA Queue Managers.

Test	M2004A HA				M2004A nonHA			
	Max Rate*	CPU%	Clients	Latency#	Max Rate*	CPU%	Clients	Latency#
10Q 10QM Request Responder (256b Persistent)	203,927	92.2	300	0.5	240,354	96.84	300	0.3
10Q 10QM Request Responder (2KB Persistent)	184,089	92.89	300	0.5	218,257	97.67	300	0.3
10Q 10QM Request Responder (20KB Persistent)	116,647	80.9	300	0.6	151,930	92.16	300	0.4
10Q 10QM Request Responder (200KB Persistent)	14,788	26.15	120	1.2	16,156	22.79	110	0.8

*Round trips/sec
#Single thread round trip latency (ms)

TABLE 9 - PEAK RATES FOR 10QM PERSISTENT MESSAGING

8.3 How does HA perform over larger distances?

The previous section shows how the MQ appliance HA capability might perform if both appliances were located in the same data centre (i.e. 3m distance between the appliances). How would the HA performance differ if the pair of appliances were located a larger distance apart? Due to testing limitations, we need to simulate the latency that might be experienced as the distances between the appliances grows.

If the appliances are located 100Km apart, you might expect the smallest increase in packet transmission latency for each leg to be calculated as follows:

distance / speed = time

100,000m / 300,000,000m/s [1] = 0.000333s = 333 microseconds

There must also be an allowance for the refraction index of the cable

333 * 1.5 = 500 microseconds

Switching hardware and non-linear cable routing will likely further increase the latency between the pair of HA appliances. It is currently advised to customers to site a pair of HA appliances so that the latency between the two appliances is no greater than 10ms and preferably within the same data centre.

A delay can be inserted into the sending network layer of both appliances to simulate such latency and let us examine how this impacts the HA performance. The following chart repeats test HA2 from section 8.2 and shows the effect of a 1ms round trip latency introduced into the network layer between the two HA appliances.

[1] Assuming speed of light to be $3x10^8$m/s

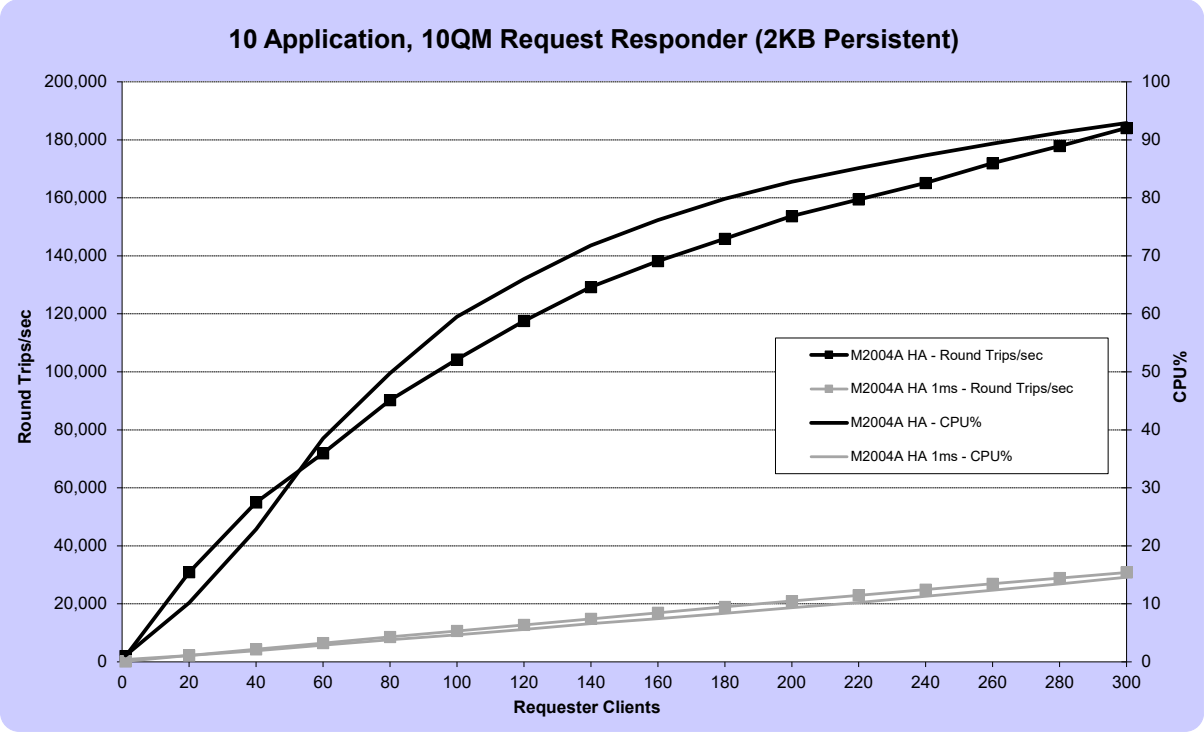


FIGURE 12 - PERFORMANCE RESULTS FOR 2KB, 10QM PERSISTENT MESSAGING WITH/WITHOUT 1MS LATENCY

Figure 12 shows that an additional 1ms latency on the round trip time of the HA replication interface results in a ~80% reduction in performance than compared with the direct connection (no additional latency) between the appliances.

Test	M2004A HA Max Rate*	CPU%	Clients	Latency#	vs Direct
10Q 10QM Request Responder 1ms Latency (256b Persistent)	30,931	13.37	300	4.6	15.2%
10Q 10QM Request Responder 1ms Latency (2KB Persistent)	30,822	14.59	300	5.5	16.7%
10Q 10QM Request Responder 1ms Latency (20KB Persistent)	28,142	17.68	300	5.8	24.1%
10Q 10QM Request Responder 1ms Latency (200KB Persistent)	8,122	15.05	120	6.4	54.9%

*Round trips/sec
#Single thread round trip latency (ms)

TABLE 10 - PEAK RATES FOR 10QM PERSISTENT MESSAGING WITH 1MS SIMULATED LATENCY

The data in the following tables show additional data points with simulated latency delays of 2, 5 and 10ms.

Test	M2004A HA Max Rate*	CPU%	Clients	Latency#	vs Direct
10Q 10QM Request Responder 2ms Latency (256b Persistent)	16,084	6.88	300	8.7	7.9%
10Q 10QM Request Responder 2ms Latency (2KB Persistent)	16,089	7.61	300	10.3	8.7%
10Q 10QM Request Responder 2ms Latency (20KB Persistent)	15,341	9.42	300	10.9	13.2%
10Q 10QM Request Responder 2ms Latency (200KB Persistent)	4,960	9.07	120	11.5	33.5%

*Round trips/sec
#Single thread round trip latency (ms)

Test	M2004A HA Max Rate*	CPU%	Clients	Latency#	vs Direct
10Q 10QM Request Responder 5ms Latency (256b Persistent)	6,570	3.1	300	20.9	3.2%
10Q 10QM Request Responder 5ms Latency (2KB Persistent)	6,600	3.36	300	25.0	3.6%
10Q 10QM Request Responder 5ms Latency (20KB Persistent)	6,431	4.11	300	26.3	5.5%
10Q 10QM Request Responder 5ms Latency (200KB Persistent)	2,155	4.05	120	26.9	14.6%

*Round trips/sec

#Single thread round trip latency (ms)

Test	M2004A HA Max Rate*	CPU%	Clients	Latency#	vs Direct
10Q 10QM Request Responder 10ms Latency (256b Persistent)	3,314	1.8	300	41.3	1.6%
10Q 10QM Request Responder 10ms Latency (2KB Persistent)	3,322	1.91	300	49.4	1.8%
10Q 10QM Request Responder 10ms Latency (20KB Persistent)	3,288	2.4	300	51.9	2.8%
10Q 10QM Request Responder 10ms Latency (200KB Persistent	1,117	2.31	120	52.7	7.6%

*Round trips/sec

#Single thread round trip latency (ms)

TABLE 11 - PEAK RATES FOR 10QM PERSISTENT MESSAGING WITH 2, 5 AND 10MS SIMULATED LATENCY

9 DR Scenarios

Users can configure a Queue Manager for Disaster Recovery (DR) to ensure that the QM data is distributed to a recovery appliance. This configuration allows the Queue Manager on the recovery appliance to resume work should an outage occur that results in the main appliance becoming unavailable.

Users can also configure a QM for both HA and DR. The performance of this configuration will be examined in section 10, whereas in this section we will look at the standalone performance of DR.

The Queue Manager data is replicated asynchronously to the recovery appliance, which can result in messaging data loss (up to a maximum of 4MB per QM is held in the TCP send buffer) should the main appliance become unavailable. The Queue Manager at the recovery appliance must be manually started before it can start accepting connections from clients.

To illustrate the cost of enabling the DR infrastructure, tests will be performed on two of the scenarios featured earlier in this report.

1) Request Responder 1QM Persistent (Test C2)

2) Request Responder 10QM Persistent (Test C4)

Each test will be conducted with both a standalone QM and a QM configured with a remote DR appliance, so that the cost of the asynchronous replication can be evaluated.

This section utilises the following connections:

Appliance A	Appliance B	Appliance C (DR)	Notes
eth13	eth13		Connected directly between appliances with 1Gb copper patch cable. Used in section 10 only
eth17	eth17		Connected directly between appliances with 1Gb copper patch cable. Used in section 10 only
eth43	eth43	eth43	Connected via ethernet switch with 100Gb Fibre cables for DR
eth41	eth41		Connected directly between appliances with 100Gb Fibre cable for HA. Used in section 10 only
eth40			Workload driven via this 100Gb interface
eth42			Workload driven via this 100Gb interface

9.1 Test Scenario DR1 – 10 Applications per QM, 1 QM, Persistent

This test is identical to test C2 in section 6.2 and is presented here with results from running tests against a standalone QM and against a QM that is configured for Disaster Recovery (although the recovery appliance is located 3m from the main appliance).

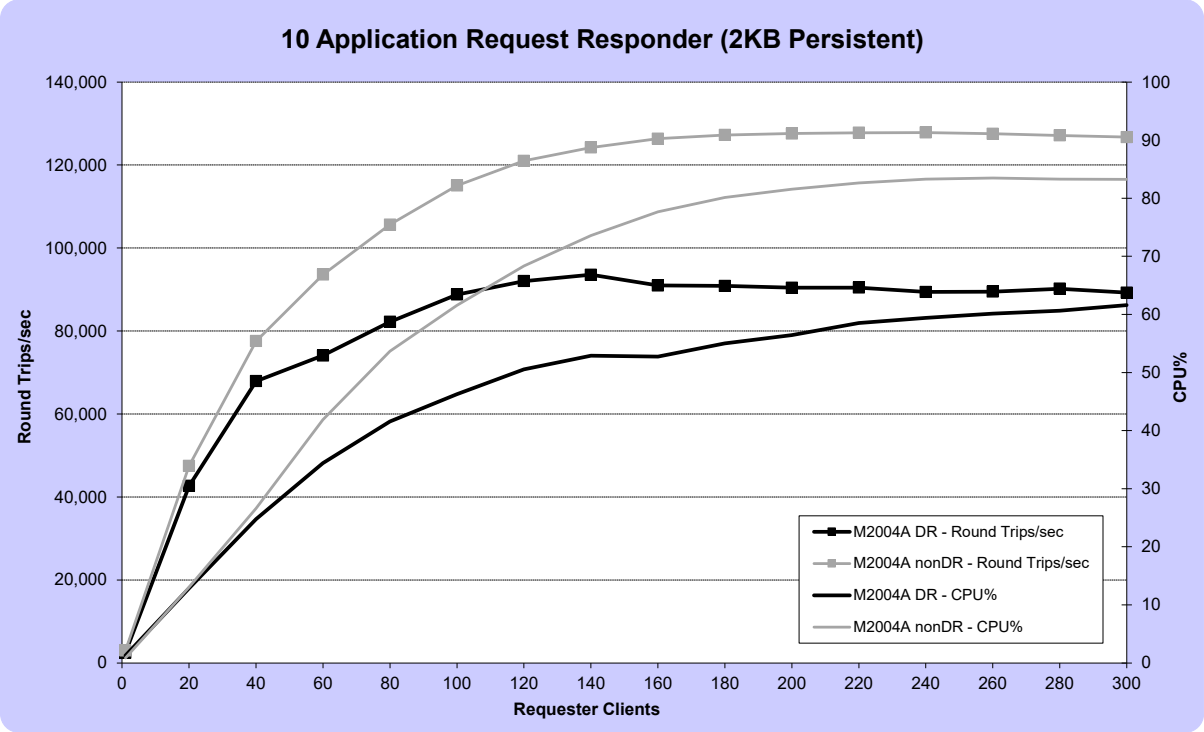


FIGURE 13 – PERFORMANCE RESULTS FOR 2KB PERSISTENT MESSAGING

Figure 13 shows that by enabling DR capability, the maximum throughput achieved with a 2KB message size on a single Queue Manager is reduced by up to 25%. This is similar to the impact of enabling HA replication, but utilises less CPU.

Test	M2004A DR				M2004A nonDR			
	Max Rate*	CPU%	Clients	Latency#	Max Rate*	CPU%	Clients	Latency#
10Q Request Responder (256b Persistent)	128,261	80.57	300	0.3	134,527	83.09	240	0.3
10Q Request Responder (2KB Persistent)	93,556	52.9	140	0.4	127,828	83.27	240	0.3
10Q Request Responder (20KB Persistent)	40,185	33.02	160	0.5	77,127	54.32	160	0.4
10Q Request Responder (200KB Persistent)	4,281	6.48	30	1.1	9,364	10.74	30	0.8

*Round trips/sec
#Single thread round trip latency (ms)

TABLE 12 - PEAK RATES FOR PERSISTENT MESSAGING

9.2 Test Scenario DR2 – 10 Applications per QM, 10 QM, Persistent

This test is identical to the test C4 in section 6.4 and is presented here with results from running tests against ten standalone QM and against ten QM that are configured for Disaster Recovery (although the recovery appliance is located 3m from the main appliance).

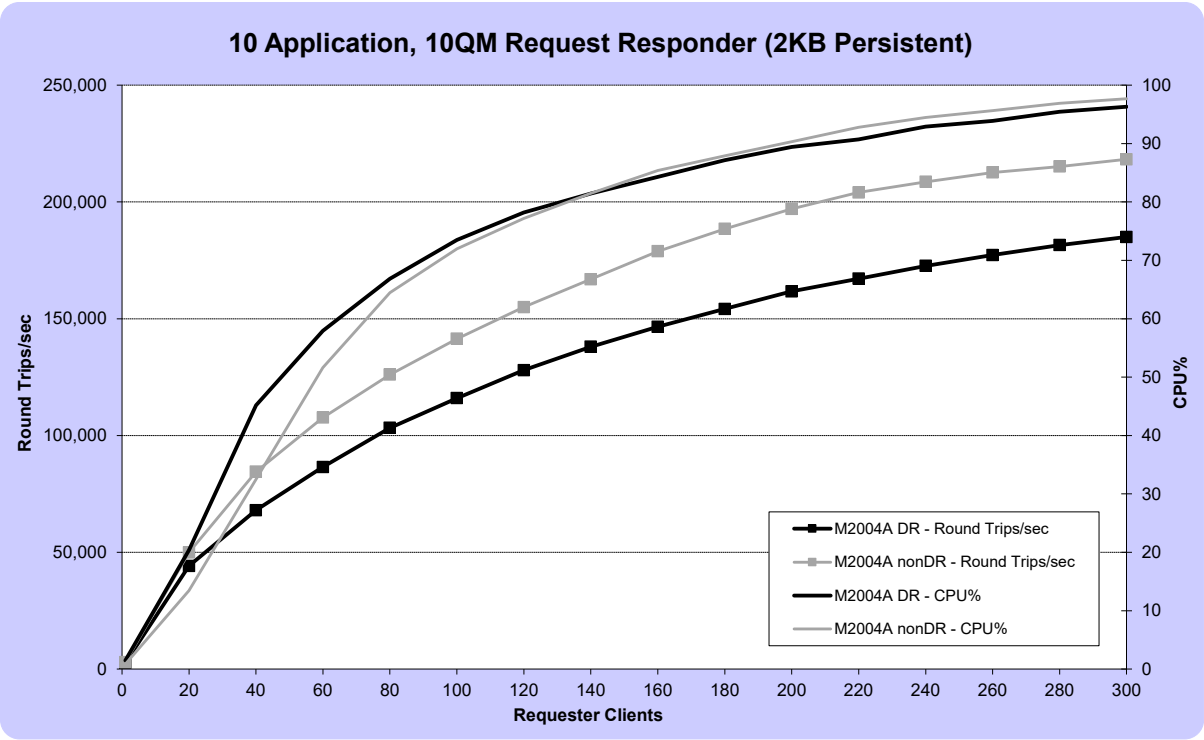


FIGURE 14 - PERFORMANCE RESULTS FOR 2KB, 10QM PERSISTENT MESSAGING

Figure 14 shows that when we have multiple QM configured for Disaster Recovery performing 2KB persistent messaging, the peak messaging rate is within 15% of the rate achieved when distributed across a set of NonDR Queue Managers.

Test	M2004A DR				M2004A nonDR			
	Max Rate*	CPU%	Clients	Latency#	Max Rate*	CPU%	Clients	Latency#
10Q 10QM Request Responder (256b Persistent)	204,272	96.68	300	0.4	240,354	96.84	300	0.3
10Q 10QM Request Responder (2KB Persistent)	185,003	96.32	300	0.4	218,257	97.67	300	0.3
10Q 10QM Request Responder (20KB Persistent)	100,664	78.86	300	0.5	151,930	92.16	300	0.4
10Q 10QM Request Responder (200KB Persistent)	15,869	33.04	110	1.2	16,156	22.79	110	0.8

*Round trips/sec
#Single thread round trip latency (ms)

TABLE 13 - PEAK RATES FOR 10QM PERSISTENT MESSAGING

9.3 How does DR perform over larger distances?

DR configuration usually requires a remote appliance to be situated a large distance from the primary location so that any particular event that might affect the primary location would be hoped not to affect the remote recovery appliance.

The data in the following tables show the results from the test scenario featured in the previous section but with additional data points using simulated latency delays of 10, 20, 50 and 100ms. A comparison against the DR scenario in which the MQ Appliances are directly connected is also included.

Test	M2004A DR Max Rate*	CPU%	Clients	Latency#	vs Direct
10Q 10QM Request Responder 10ms Latency (256b Persistent)	199,633	94.91	300	0.4	97.7%
10Q 10QM Request Responder 10ms Latency (2KB Persistent)	180,899	92.26	300	0.3	97.8%
10Q 10QM Request Responder 10ms Latency (20KB Persistent)	48,657	29.83	240	0.4	48.3%
10Q 10QM Request Responder 10ms Latency (200KB Persistent	5,051	9.38	90	1.7	31.8%

*Round trips/sec
#Single thread round trip latency (ms)

Test	M2004A DR Max Rate*	CPU%	Clients	Latency#	vs Direct
10Q 10QM Request Responder 20ms Latency (256b Persistent)	191,887	91.34	300	0.3	93.9%
10Q 10QM Request Responder 20ms Latency (2KB Persistent)	148,887	79.59	240	0.4	80.5%
10Q 10QM Request Responder 20ms Latency (20KB Persistent)	20,534	10.97	220	0.6	20.4%
10Q 10QM Request Responder 20ms Latency (200KB Persistent	2,461	4.52	70	3.1	15.5%

*Round trips/sec
#Single thread round trip latency (ms)

Test	M2004A DR Max Rate*	CPU%	Clients	Latency#	vs Direct
10Q 10QM Request Responder 50ms Latency (256b Persistent)	107,670	44.77	280	0.4	52.7%
10Q 10QM Request Responder 50ms Latency (2KB Persistent)	41,985	15.53	140	0.6	22.7%
10Q 10QM Request Responder 50ms Latency (20KB Persistent)	7,554	3.65	60	1.4	7.5%
10Q 10QM Request Responder 50ms Latency (200KB Persistent	744	1.53	100	7.0	4.7%

*Round trips/sec
#Single thread round trip latency (ms)

Test	M2004A DR Max Rate*	CPU%	Clients	Latency#	vs Direct
10Q 10QM Request Responder 100ms Latency (256b Persistent)	52,891	18.17	260	0.9	25.9%
10Q 10QM Request Responder 100ms Latency (2KB Persistent)	22,172	7.57	180	1.2	12.0%
10Q 10QM Request Responder 100ms Latency (20KB Persistent	3,832	1.94	100	2.7	3.8%
10Q 10QM Request Responder 100ms Latency (200KB Persisten	387	0.85	60	14.5	2.4%

*Round trips/sec
#Single thread round trip latency (ms)

TABLE 14 - PEAK RATES FOR 10QM PERSISTENT MESSAGING WITH 10, 20, 50 AND 100MS SIMULATED LATENCY

10 HA and DR Scenarios

Users can configure a Queue Manager for both HA and DR. The data that is asynchronously replicated for disaster recovery is sent from the currently active instance of the HA pair.

This configuration allows the Queue Manager on the DR recovery appliance to resume work should an outage occur that results in both the appliances in the HA group becoming unavailable.

The performance of these scenarios is close to that which has been measured in the HA scenarios featured in sections 8.1 and 8.2; the throughput is slightly reduced due to the additional work performing the asynchronous replication in addition to the synchronous HA replication.

Test	M2004A HA and DR				M2004A nonHA			
	Max Rate*	CPU%	Clients	Latency#	Max Rate*	CPU%	Clients	Latency#
10Q Request Responder (256b Persistent)	120,623	78.17	300	0.4	134,527	83.09	240	0.3
10Q Request Responder (2KB Persistent)	85,305	52.28	160	0.5	127,828	83.27	240	0.3
10Q Request Responder (20KB Persistent)	21,136	14.37	80	0.7	77,127	54.32	160	0.4
10Q Request Responder (200KB Persistent)	2,332	5.31	15	1.3	9,364	10.74	30	0.8

*Round trips/sec
#Single thread round trip latency (ms)

TABLE 15 - PEAK RATES FOR PERSISTENT MESSAGING, HA AND DR

Test	M2004A HA and DR				M2004A nonHA			
	Max Rate*	CPU%	Clients	Latency#	Max Rate*	CPU%	Clients	Latency#
10Q 10QM Request Responder (256b Persistent)	181,523	91.63	300	0.5	240,354	96.84	300	0.3
10Q 10QM Request Responder (2KB Persistent)	165,807	92.15	300	0.6	218,257	97.67	300	0.3
10Q 10QM Request Responder (20KB Persistent)	92,448	78.67	300	0.6	151,930	92.16	300	0.4
10Q 10QM Request Responder (200KB Persistent)	14,110	38.28	120	1.3	16,156	22.79	110	0.8

*Round trips/sec
#Single thread round trip latency (ms)

TABLE 16 - PEAK RATES FOR 10QM PERSISTENT MESSAGING, HA AND DR

11 Additional M2004A vs M2004B scenarios

In the earlier sections of this report detailing NonHA and NonDR scenarios, graphical and numerical data was also provided for the M2004B model. In the later sections detailing HA and DR scenarios, the comparison points were the equivalent NonHA and NonDR measurements.

This section has been added to illustrate the performance comparison of running HA and DR scenarios on either the M2004A or M2004B appliances.

We will initially look at the results of running test HA1 from section 8.1:

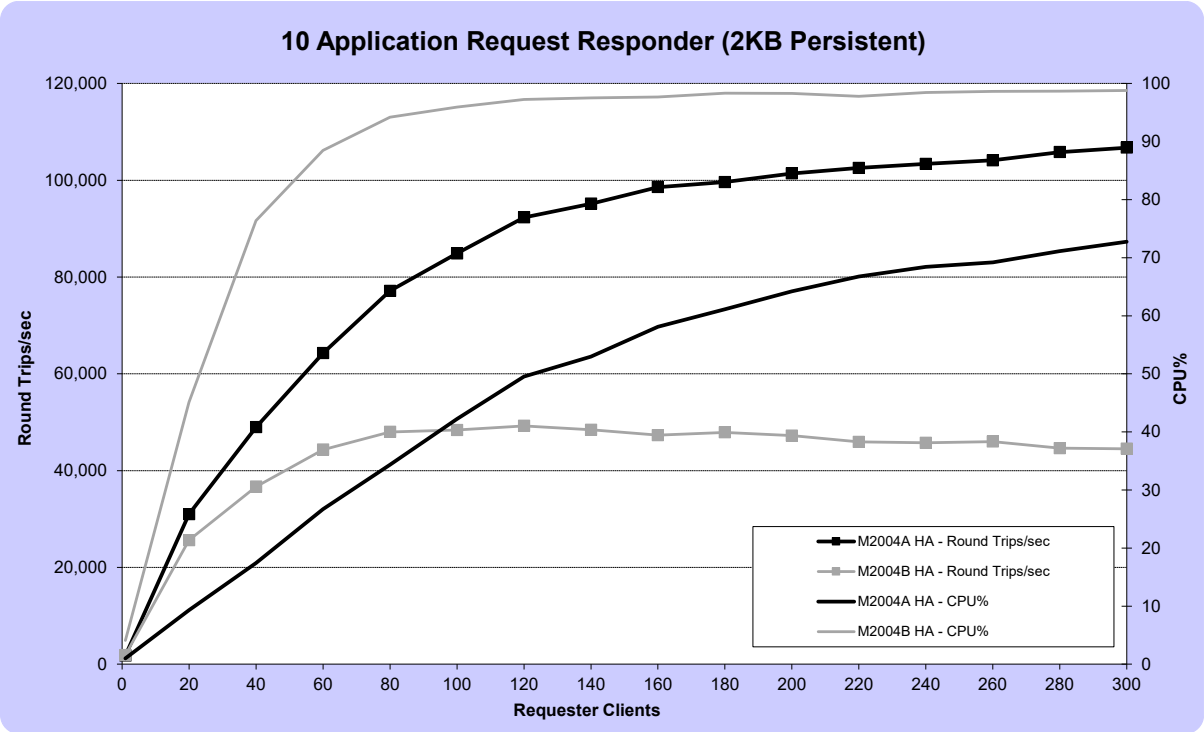


FIGURE 15 - PERFORMANCE RESULTS FOR 2KB PERSISTENT MESSAGING

Figure 15 shows that the M2004A appliance can achieve over twice the throughput of the M2004B appliance in this single HA QM scenario.

Test	M2004A HA				M2004B HA			
	Max Rate*	CPU%	Clients	Latency#	Max Rate*	CPU%	Clients	Latency#
10Q Request Responder (256b Persistent)	133,940	84.76	280	0.5	54,329	97.76	120	0.5
10Q Request Responder (2KB Persistent)	106,746	72.74	300	0.6	49,261	97.24	120	0.5
10Q Request Responder (20KB Persistent)	29,164	23.75	200	0.6	24,682	69.9	160	0.6
10Q Request Responder (200KB Persistent)	3,658	5.59	30	1.1	3,299	25.18	30	1.2

*Round trips/sec
#Single thread round trip latency (ms)

TABLE 17 - PEAK RATES FOR PERSISTENT MESSAGING

The following graph shows the results of running test HA2 from section 8.2:

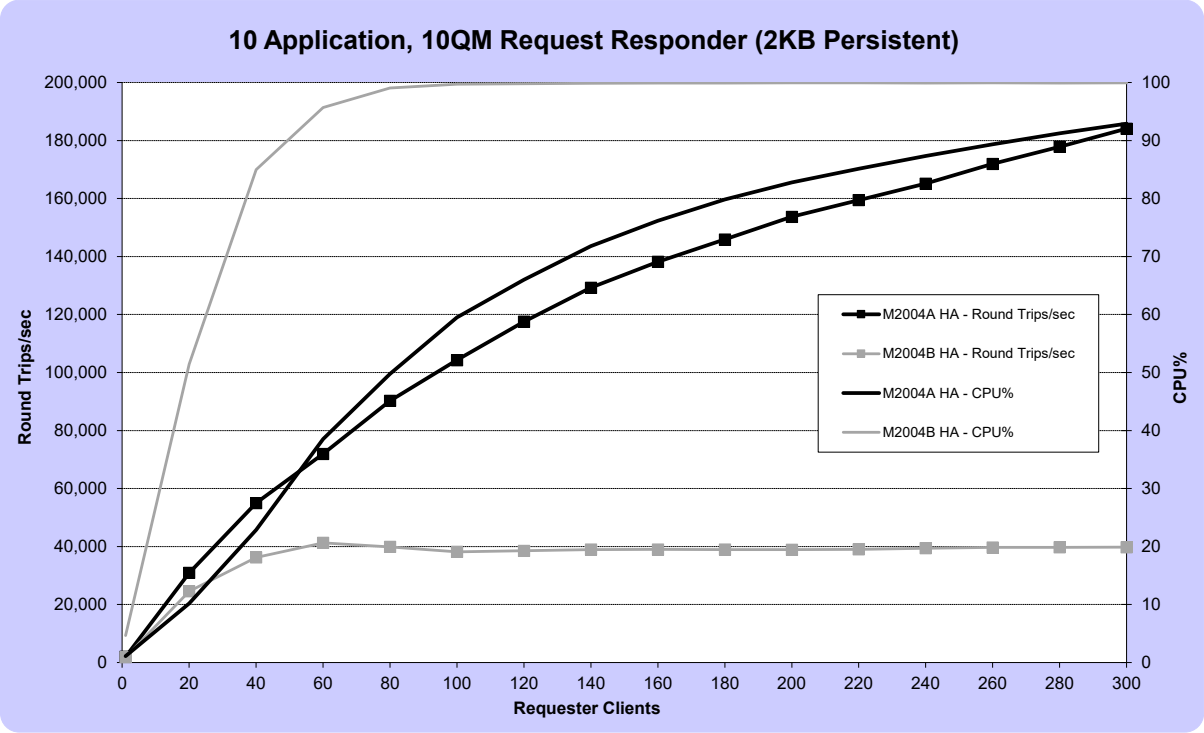


FIGURE 16 - PERFORMANCE RESULTS FOR 2KB, 10QM PERSISTENT MESSAGING

Figure 16 shows that with multiple HA Queue Managers, the M2004A appliance can achieve over 4 times the throughput of the M2004B appliance.

Test	M2004A HA				M2004B HA			
	Max Rate*	CPU%	Clients	Latency#	Max Rate*	CPU%	Clients	Latency#
10Q 10QM Request Responder (256b Persistent)	203,927	92.2	300	0.5	44,485	94.73	60	0.5
10Q 10QM Request Responder (2KB Persistent)	184,089	92.89	300	0.5	41,259	95.69	60	0.6
10Q 10QM Request Responder (20KB Persistent)	116,647	80.9	300	0.6	32,872	98.91	80	0.6
10Q 10QM Request Responder (200KB Persistent)	14,788	26.15	120	1.2	11,691	97.98	120	1.2

*Round trips/sec
#Single thread round trip latency (ms)

TABLE 18 - PEAK RATES FOR 10QM PERSISTENT MESSAGING

The following graph shows the results of running test DR1 from section 9.1:

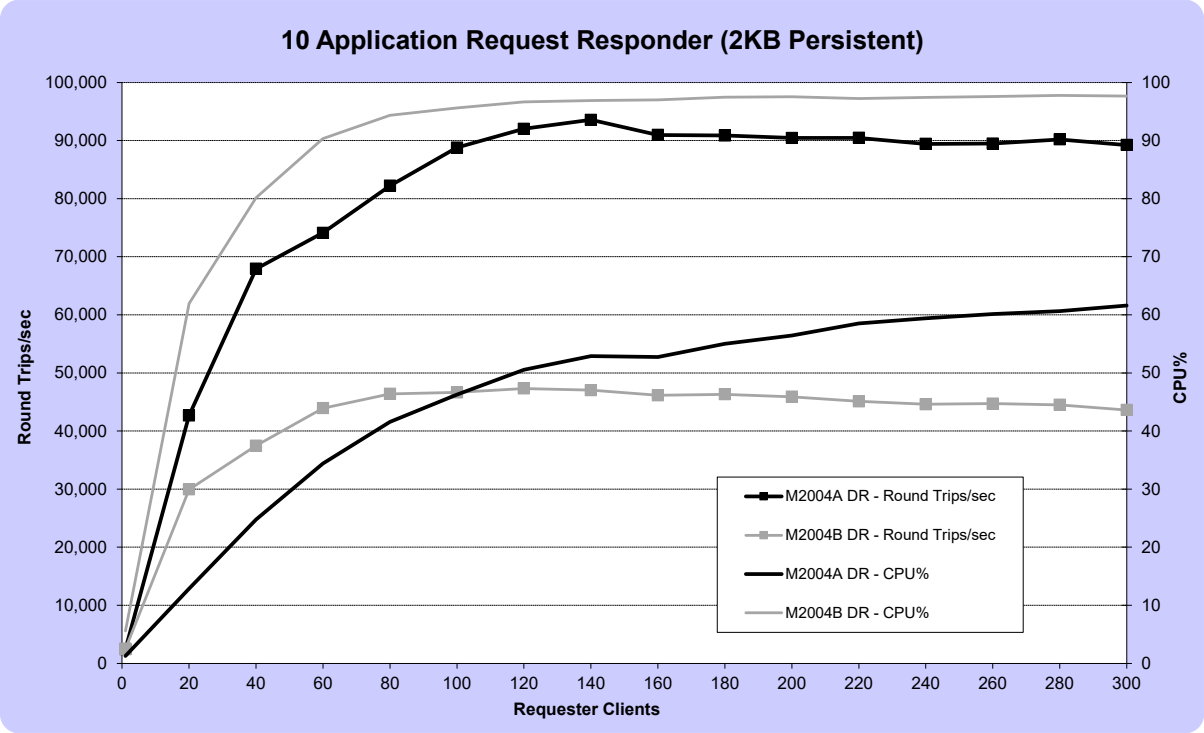


FIGURE 17 - PERFORMANCE RESULTS FOR 2KB PERSISTENT MESSAGING

Figure 17 shows that the M2004A appliance can achieve more than double the throughput of the M2004B appliance in this single DR QM scenario.

Test	M2004A DR				M2004B DR			
	Max Rate*	CPU%	Clients	Latency#	Max Rate*	CPU%	Clients	Latency#
10Q Request Responder (256b Persistent)	128,261	80.57	300	0.3	53,490	96.46	80	0.3
10Q Request Responder (2KB Persistent)	93,556	52.9	140	0.4	47,318	96.65	120	0.4
10Q Request Responder (20KB Persistent)	40,185	33.02	160	0.5	18,388	54.5	100	0.5
10Q Request Responder (200KB Persistent)	4,281	6.48	30	1.1	3,453	29.02	30	1.2

*Round trips/sec
#Single thread round trip latency (ms)

TABLE 19 - PEAK RATES FOR PERSISTENT MESSAGING

The following graph shows the results of running test DR2 from section 9.2:

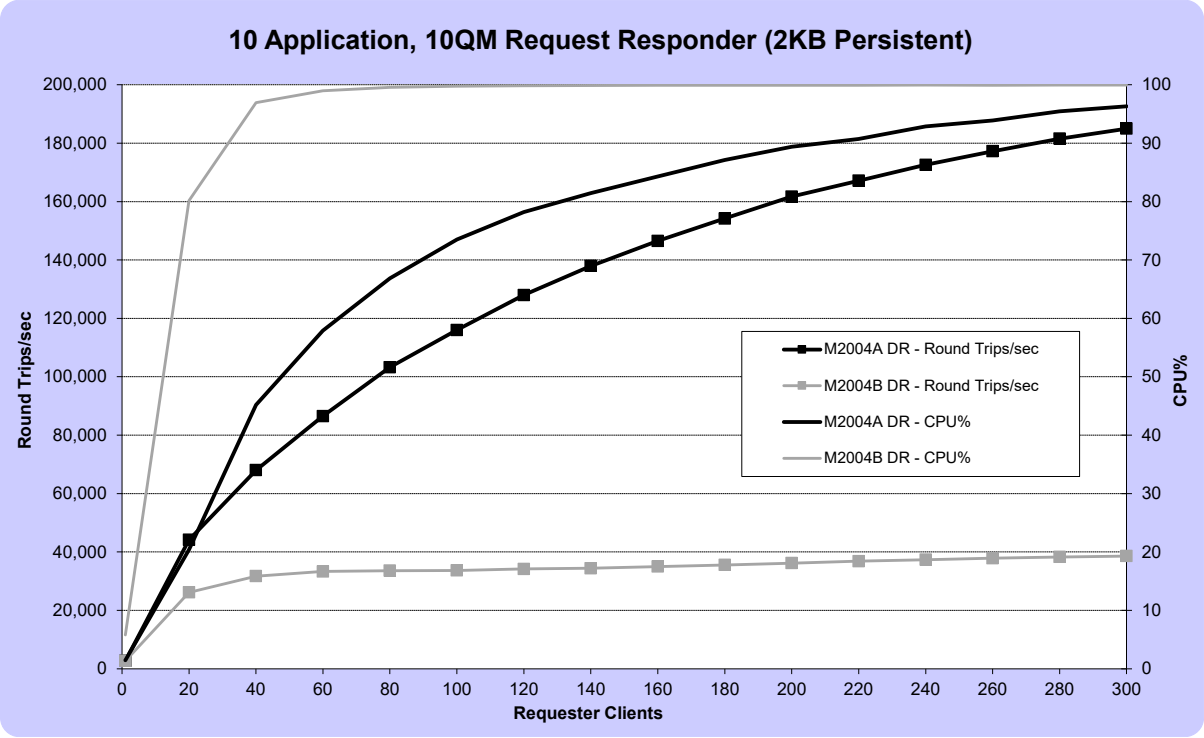


FIGURE 18 - PERFORMANCE RESULTS FOR 2KB, 10QM PERSISTENT MESSAGING

Figure 18 shows that with multiple DR Queue Managers, the M2004A appliance can achieve up to 4 times the throughput of the M2004B appliance.

Test	M2004A DR				M2004B DR			
	Max Rate*	CPU%	Clients	Latency#	Max Rate*	CPU%	Clients	Latency#
10Q 10QM Request Responder (256b Persistent)	204,272	96.68	300	0.4	42,018	99.94	300	0.4
10Q 10QM Request Responder (2KB Persistent)	185,003	96.32	300	0.4	38,592	99.96	300	0.4
10Q 10QM Request Responder (20KB Persistent)	100,664	78.86	300	0.5	28,521	99.95	300	0.4
10Q 10QM Request Responder (200KB Persistent)	15,869	33.04	110	1.2	8,880	99.39	120	1.1

*Round trips/sec
#Single thread round trip latency (ms)

TABLE 20 - PEAK RATES FOR 10QM PERSISTENT MESSAGING

The following graph shows the results of running the combined HA and DR scenario from section 10 for a single Queue Manager:

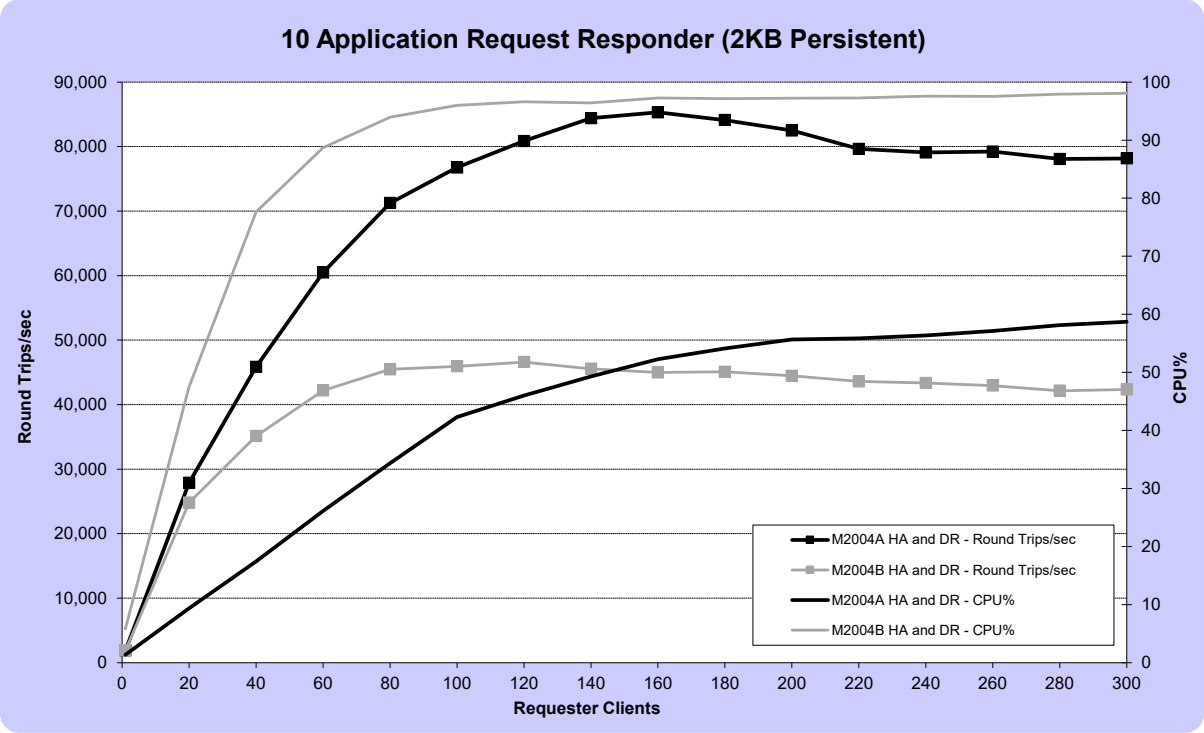


FIGURE 19 - PERFORMANCE RESULTS FOR 2KB PERSISTENT MESSAGING

The table below shows that the M2004A appliance can achieve more than double the throughput of the M2004B appliance in this single HA and DR QM scenario.

Test	M2004A HA and DR				M2004B HA and DR			
	Max Rate*	CPU%	Clients	Latency#	Max Rate*	CPU%	Clients	Latency#
10Q Request Responder (256b Persistent)	120,623	78.17	300	0.4	52,782	97.94	120	0.5
10Q Request Responder (2KB Persistent)	85,305	52.28	160	0.5	46,601	96.61	120	0.5
10Q Request Responder (20KB Persistent)	21,136	14.37	80	0.7	17,709	56.27	120	0.6
10Q Request Responder (200KB Persistent)	2,332	5.31	15	1.3	2,321	24.65	15	1.2

*Round trips/sec
#Single thread round trip latency (ms)

TABLE 21 - PEAK RATES FOR PERSISTENT MESSAGING

The following graph shows the results of running the combined HA and DR scenario from section 10 against ten Queue Managers:

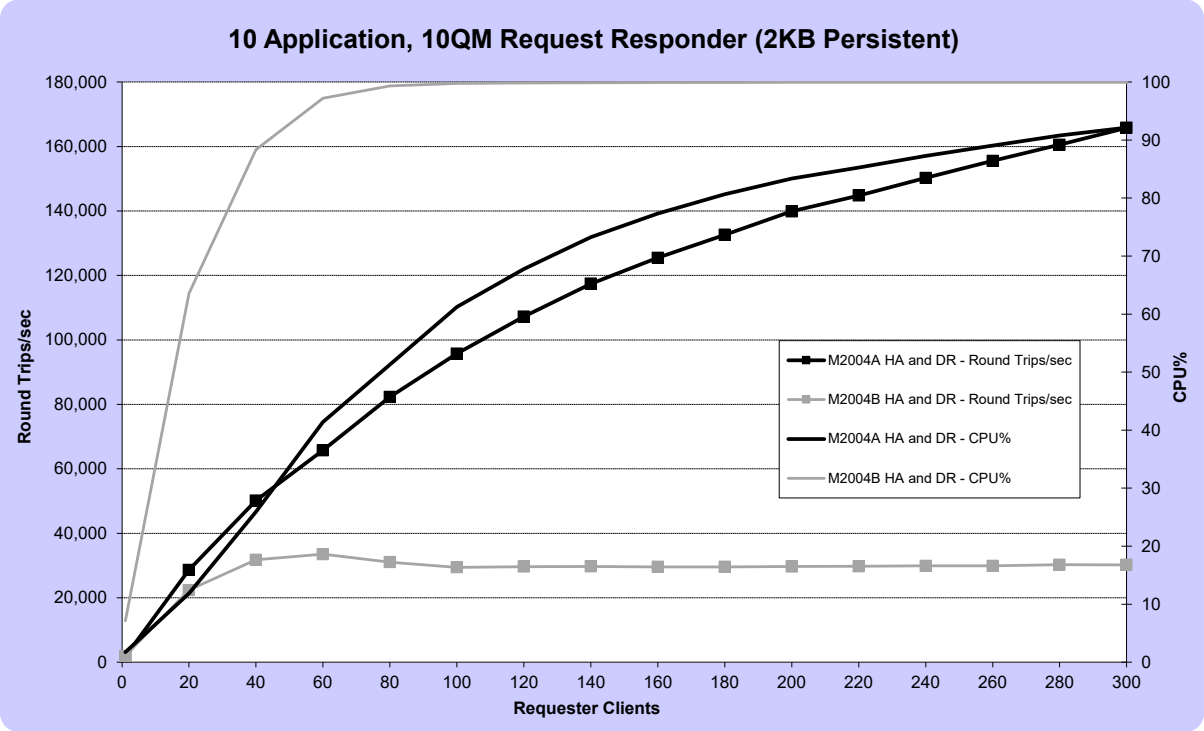


FIGURE 20 - PERFORMANCE RESULTS FOR 2KB, 10QM PERSISTENT MESSAGING

Figure 20 shows that with multiple HA and DR Queue Managers, the M2004A appliance can achieve over 4 times the throughput of the M2004B appliance.

Test	M2004A HA and DR				M2004B HA and DR			
	Max Rate*	CPU%	Clients	Latency#	Max Rate*	CPU%	Clients	Latency#
10Q 10QM Request Responder (256b Persistent)	181,523	91.63	300	0.5	36,028	97.26	60	0.5
10Q 10QM Request Responder (2KB Persistent)	165,807	92.15	300	0.6	33,507	97.19	60	0.5
10Q 10QM Request Responder (20KB Persistent)	92,448	78.67	300	0.6	24,604	96.09	60	0.6
10Q 10QM Request Responder (200KB Persistent)	14,110	38.28	120	1.3	7,595	97.96	120	1.2

*Round trips/sec
#Single thread round trip latency (ms)

TABLE 22 - PEAK RATES FOR 10QM PERSISTENT MESSAGING

12 Filesystem encryption performance

It has been noted that more IBM MQ customers require message payloads to be encrypted while in transit and at rest to comply with various security mandates. Before the release of filesystem encryption, customers who wished to ensure their data was encrypted at rest had to use AMS (Advanced Message Security).

Customers can now use TLS to protect the message data in transit and an encrypted QM filesystem to protect that data when at rest in the appliance. Note that the use of an encrypted filesystem does not require that all messaging is TLS enabled.

It is also possible to enable an encrypted filesystem for an HA and/or DR QM as well as a standalone QM. The passphrase used to encrypt the filesystem will be propagated to the HA partner automatically but will need to be manually added to any DR recovery appliance.

The following sections show the impact of utilising an encrypted filesystem for Persistent messaging across several different scenarios.

12.1 Single QM

The graph below shows the results from the single QM test using a 2KB message size on an M2004A appliance:

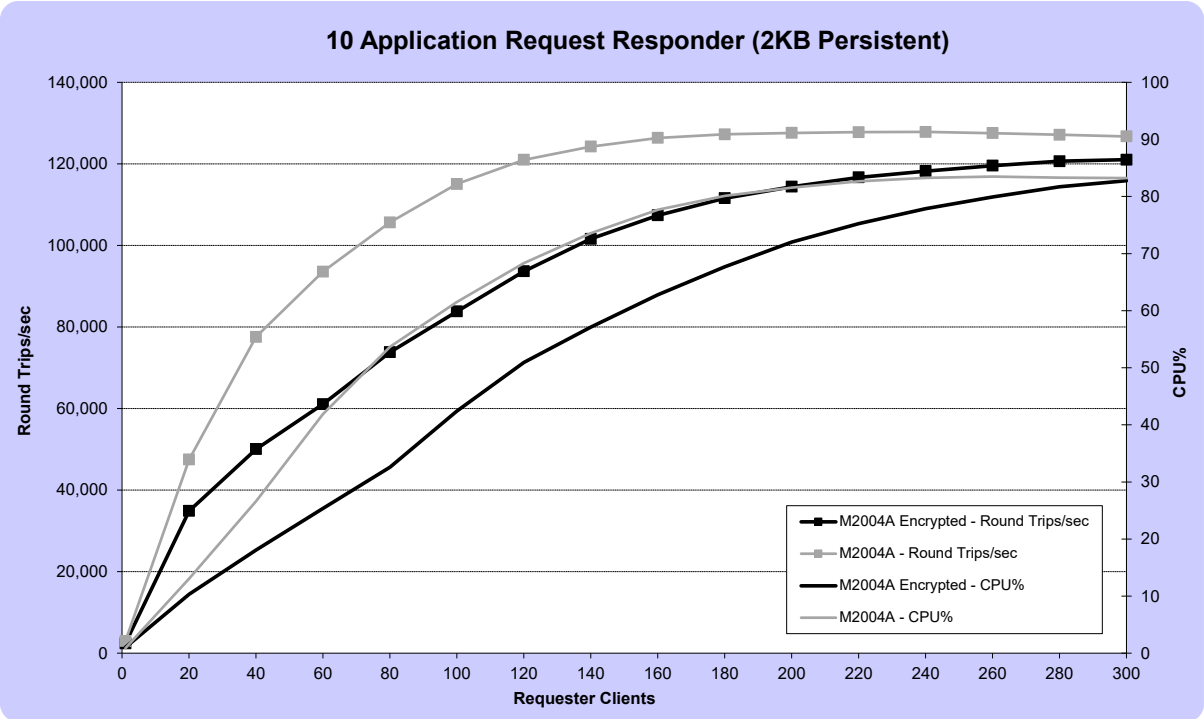


FIGURE 21 - PERFORMANCE RESULTS FOR 2KB, 1QM PERSISTENT MESSAGING, M2004A

Figure 21 shows the impact of enabling encryption on a single QM when sending and receiving 2KB messages. The peak throughput achieved is reduced by a minimum of 5%.

The following table contains the datapoints for the other message sizes in this scenario:

Test	M2004A Encrypted			M2004A		
	Max Rate*	CPU%	Clients	Max Rate*	CPU%	Clients
10Q Request Responder (256b Persistent)	132,880	86.3	300	134,527	83.09	240
10Q Request Responder (2KB Persistent)	121,031	82.76	300	127,828	83.27	240
10Q Request Responder (20KB Persistent)	56,831	47.66	140	77,127	54.32	160
10Q Request Responder (200KB Persistent)	6,340	13.48	25	9,364	10.74	30

*Round trips/sec

TABLE 23 – PEAK RATES FOR ENCRYPTED AND NON ENCRYPTED SINGLE QM ON M2004A

The following graph shows the results from the single QM test using a 2KB message size on an M2004B appliance:

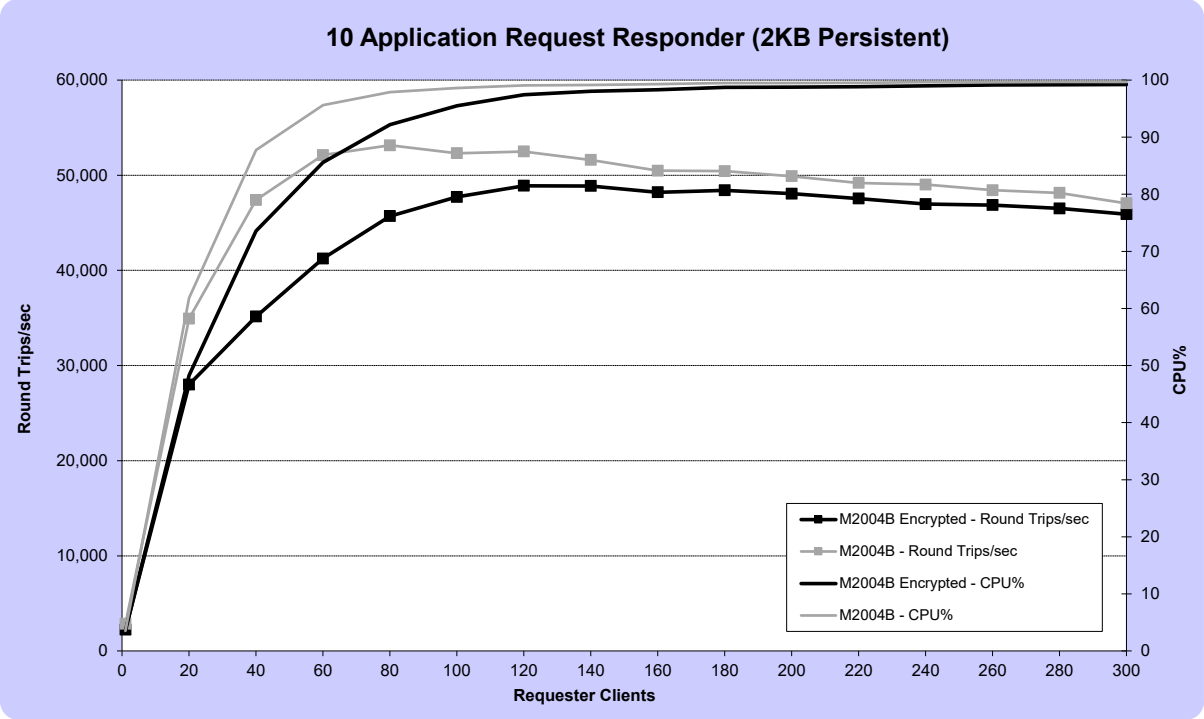


FIGURE 22 - PERFORMANCE RESULTS FOR 2KB, 1QM PERSISTENT MESSAGING, M2004B

Figure 22 shows the impact of enabling encryption on a single QM when sending and receiving 2KB messages on the M2004B appliance. The peak throughput achieved is reduced by a minimum of 8%.

The following table contains the datapoints for the other message sizes in this scenario:

Test	M2004B Encrypted			M2004B		
	Max Rate*	CPU%	Clients	Max Rate*	CPU%	Clients
10Q Request Responder (256b Persistent)	55,411	97.97	120	58,673	97.95	80
10Q Request Responder (2KB Persistent)	48,895	97.44	120	53,145	97.88	80
10Q Request Responder (20KB Persistent)	30,845	88.96	140	41,099	96.41	120
10Q Request Responder (200KB Persistent)	5,197	54.8	25	8,231	49.04	25

*Round trips/sec

TABLE 24 - PEAK RATES FOR ENCRYPTED AND NON ENCRYPTED SINGLE QM ON M2004B

12.2 Multiple QM

The graph below shows the results from the multiple QM test using a 2KB message size on a M2004A appliance:

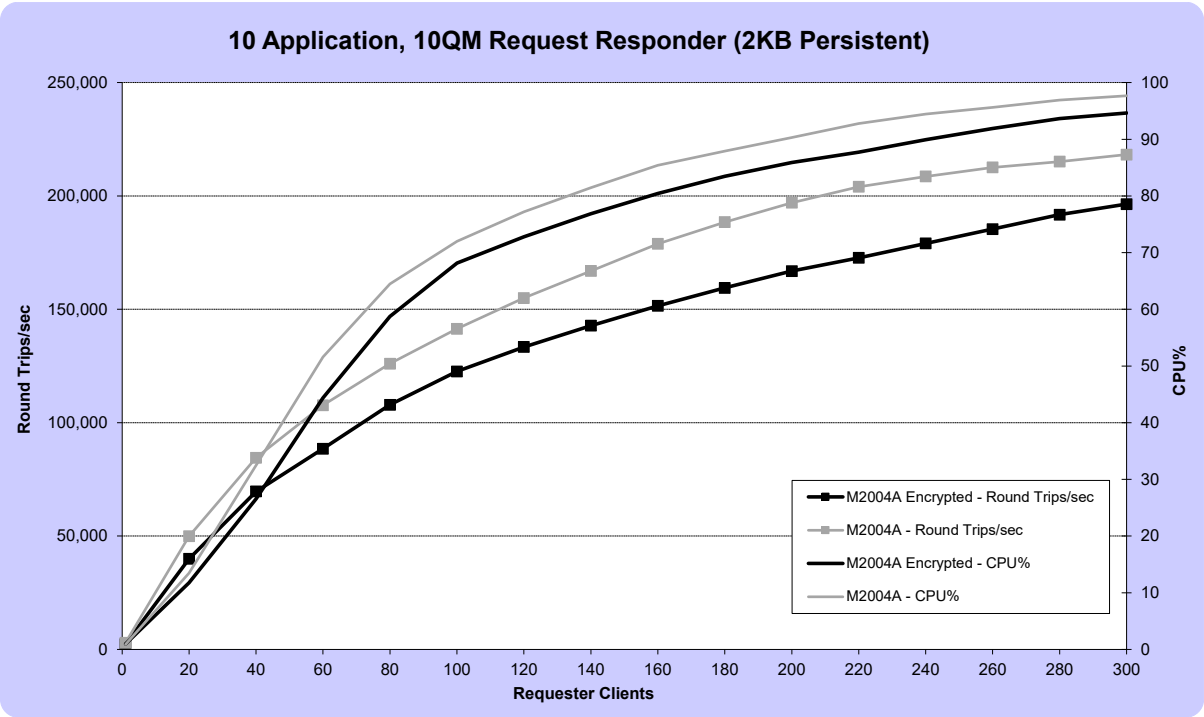


FIGURE 23 - PERFORMANCE RESULTS FOR 2KB, 10QM PERSISTENT MESSAGING, M2004A

The impact of enabling encryption has a reduced effect on multiple QM because there are multiple threads processing the encryption and recovery log writes. The peak throughput achieved across a varying number of requester threads is reduced by a minimum of 10%.

The following table contains the datapoints for the other message sizes in this scenario:

Test	M2004A Encrypted			M2004A		
	Max Rate*	CPU%	Clients	Max Rate*	CPU%	Clients
10Q 10QM Request Responder (256b Persistent)	214,931	95.18	300	240,354	96.84	300
10Q 10QM Request Responder (2KB Persistent)	196,419	94.62	300	218,257	97.67	300
10Q 10QM Request Responder (20KB Persistent)	114,217	84.69	300	151,930	92.16	300
10Q 10QM Request Responder (200KB Persistent)	15,645	44.29	110	16,156	22.79	110

*Round trips/sec

TABLE 25 - PEAK RATES FOR ENCRYPTED AND NON ENCRYPTED MULTIPLE QM ON M2004A

The graph below shows the results from the multiple QM test using a 2KB message size on a M2004B appliance:

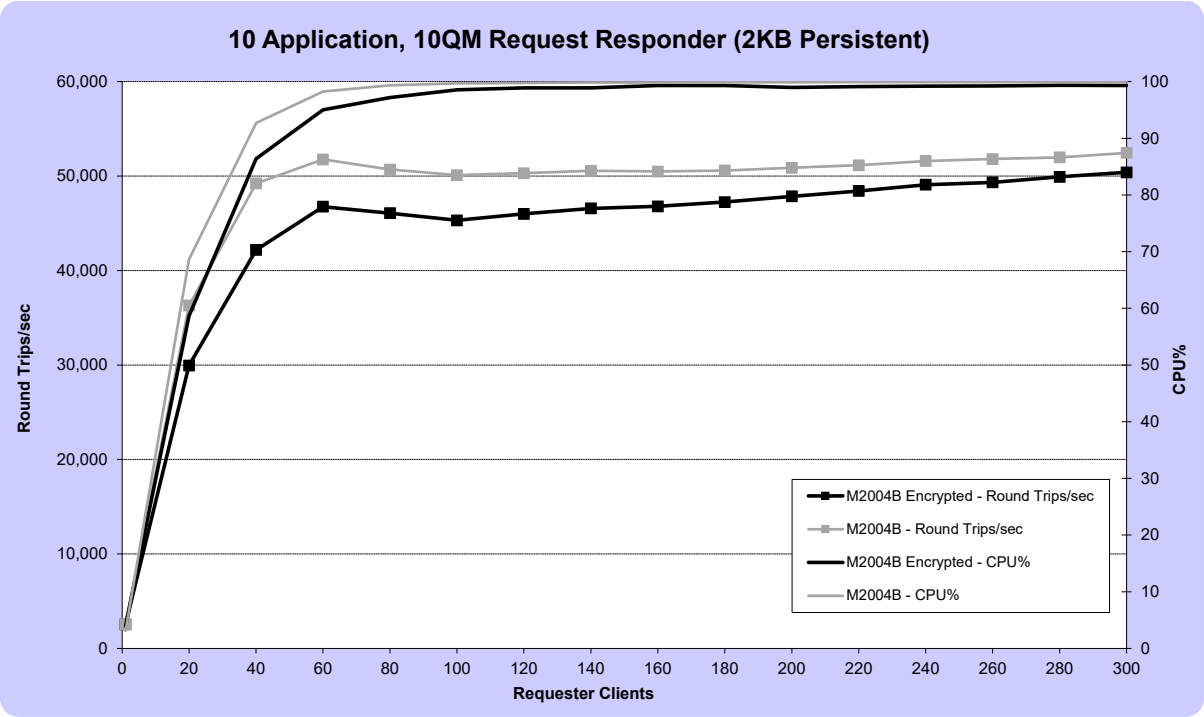


FIGURE 24 – PERFORMANCE RESULTS FOR 2KB, 10QM PERSISTENT MESSAGING, M2004B

The impact of enabling encryption has a reduced effect on multiple QM because there are multiple threads processing the encryption and recovery log writes. The peak throughput achieved across a varying number of requester threads is reduced by a minimum of 12%.

The following table contains the datapoints for the other message sizes in this scenario:

Test	M2004B Encrypted			M2004B		
	Max Rate*	CPU%	Clients	Max Rate*	CPU%	Clients
10Q 10QM Request Responder (256b Persistent)	54,817	99.25	300	57,554	99.87	300
10Q 10QM Request Responder (2KB Persistent)	50,398	99.31	300	52,462	99.93	300
10Q 10QM Request Responder (20KB Persistent)	36,836	99.77	300	42,643	98.53	60
10Q 10QM Request Responder (200KB Persistent)	9,259	99.03	120	14,283	98.36	80

*Round trips/sec

TABLE 26 - PEAK RATES FOR ENCRYPTED AND NON ENCRYPTED MULTIPLE QM ON M2004B

The filesystem encryption functionality offers protection for your data at rest within the MQ appliance. There is a small increase in CPU which reflects the cost of encrypting the message payload before persisting that data to storage. There is a similar cost when decrypting the message data after retrieval from storage. Note that in many scenarios, the QM (and the OS) optimize message retrieval by avoiding reading (and therefore decryption) from the IO subsystem.

13 TLS

This section illustrates the cost of enabling TLS communication between the clients and the QM. We will use scenarios C1 and C2 from Section 6 and apply three TLS 1.2 CipherSpecs and one TLS 1.3 CipherSpec to compare their performance.

The following TLS 1.2 CipherSpecs were tested (all are FIPS compliant):

- ECDHE_ECDSA_AES_256_CBC_SHA384
- ECDHE_ECDSA_AES_256_GCM_SHA384 (Suite B compliant)
- ECDHE_ECDSA_AES128_CCM8

The following TLS 1.3 CipherSpec was tested (FIPS compliant):

- TLS_AES_256_GCM_SHA384

Results for all the above CipherSpecs are plotted below. As will be seen in the tables below, the remaining tested CipherSpecs exhibited a performance profile like one of these.

Queue Manager authentication is used to setup the TLS conversation.

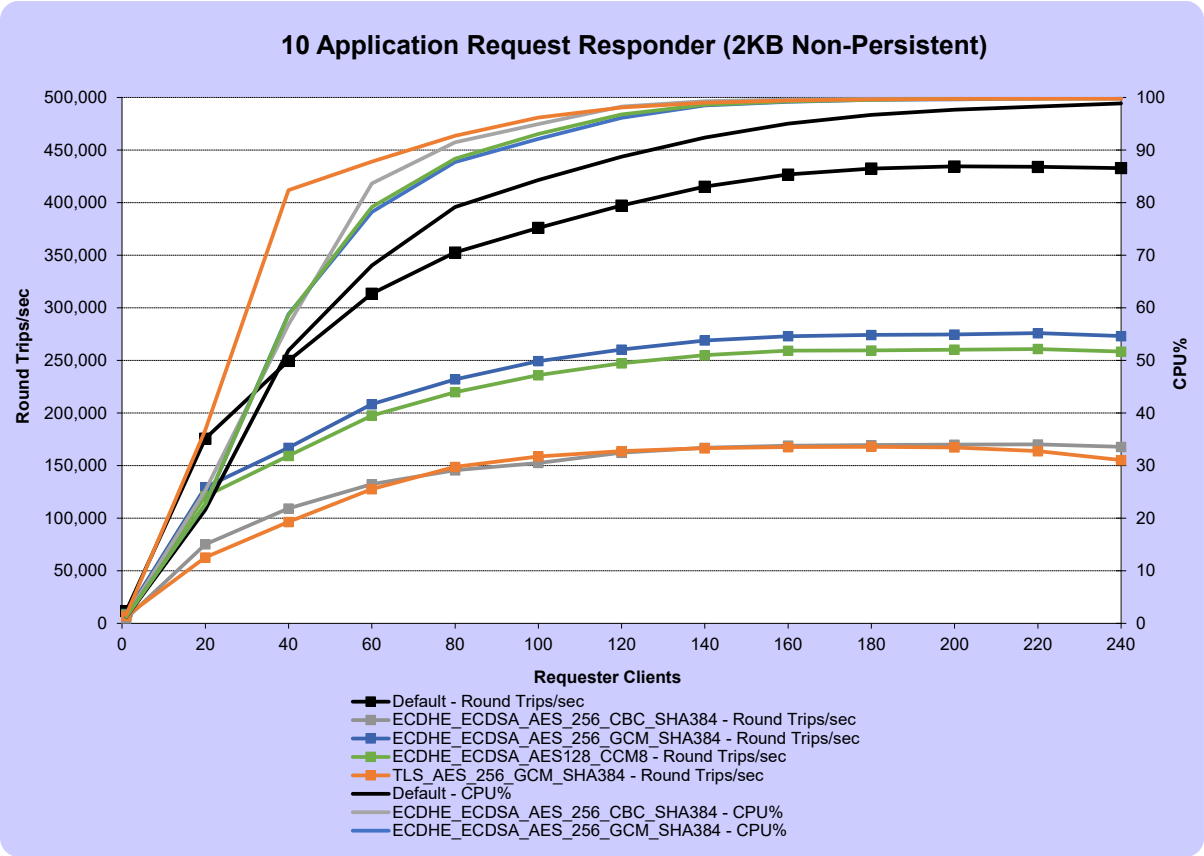


FIGURE 25 - PERFORMANCE RESULTS FOR 2KB NON-PERSISTENT MESSAGING WITH TLS

The graph in FIGURE 25 illustrates that ECDHE_ECDSA_AES_256_GCM_SHA384 was the best performing TLS12 CipherSpec, whilst the TLS12 CBC and TLS13 Ciphers perform similarly.

The table below shows the peak throughput for all the referenced CipherSpecs. It should be noted that all TLS13 CipherSpecs performed similarly.

TLS Cipher	Default Max Rate*	CPU%	Clients
ECDHE_ECDSA_AES_256_CBC_SHA384	170,142	99.87	220
ECDHE_ECDSA_AES_256_GCM_SHA384	275,881	99.8	220
ECDHE_ECDSA_AES128_CCM8	260,824	99.83	220
TLS_AES_256_GCM_SHA384	167,917	99.63	180
No TLS	434,502	92.67	260

*Round trips/sec

TABLE 27 - PEAK RATES FOR 2KB NON-PERSISTENT MESSAGING WITH TLS

The following chart illustrates the impact of TLS encryption on Persistent messaging scenarios:

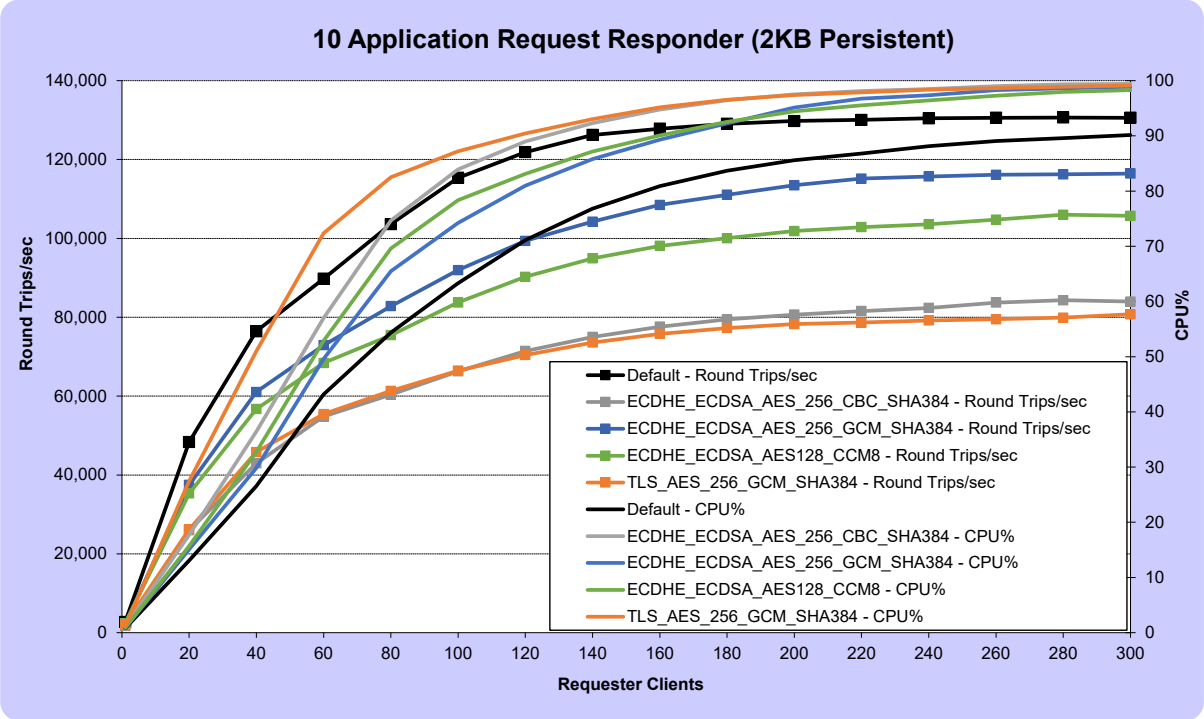


FIGURE 26 - PERFORMANCE RESULTS FOR 2KB PERSISTENT MESSAGING WITH TLS

The graph in FIGURE 26 illustrates that ECDHE_ECDSA_AES_256_GCM_SHA384 again was the best performing TLS12 CipherSpec, whilst the TLS12 CBC and TLS13 Ciphers perform similarly. TLS encryption has less impact on Persistent scenarios than Non-Persistent scenarios.

The table below shows the peak throughput for all the referenced CipherSpecs. It should be noted that all TLS13 CipherSpecs performed similarly.

TLS Cipher	Default Max Rate*	CPU%	Clients
ECDHE_ECDSA_AES_256_CBC_SHA384	84,300	99.33	280
ECDHE_ECDSA_AES_256_GCM_SHA384	116,441	98.88	300
ECDHE_ECDSA_AES128_CCM8	105,977	97.97	280
TLS_AES_256_GCM_SHA384	80,744	99.13	300
No TLS	130,648	89.6	280

*Round trips/sec

TABLE 28 - PEAK RATES FOR 2KB PERSISTENT MESSAGING WITH TLS

14 AMS

This section illustrates the cost of enabling AMS to protect the message contents in transit between the clients and the QM and at rest at the QM. We will use the scenario C2 from Section 6, and compare Integrity, Privacy and Confidentiality mode with the Non-AMS performance.

The certificate key size used is 2048 bytes and the key reuse limit in Confidentiality mode was set to 32. The symmetric key encryption algorithm used was AES256. The cryptographic hash function used was SHA512.

Note that client CPU (rather than server CPU) is featured on the graph below as that shows the increase in computation performed by the clients in encrypting and decrypting the AMS protected messages.

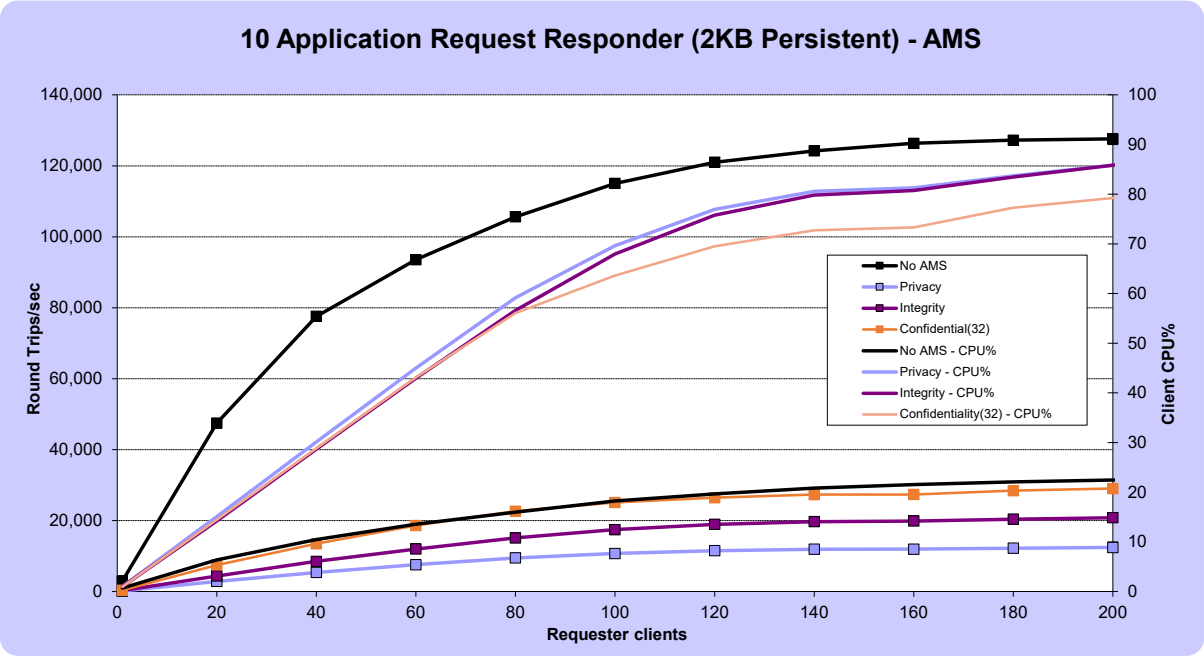


FIGURE 27 - PERFORMANCE RESULTS FOR 2KB PERSISTENT MESSAGING WITH AMS

The graph in **FIGURE 27** shows that the performance of AMS confidentiality is approximately 25% of the Non-AMS performance. There is further impact should Integrity or Privacy mode be selected.

The peak throughput achieved for the AMS Confidentiality(32) measurement at 200 clients was approximately 30,000 round trip/s. The request/responder scenario utilizes a request and a reply queue, so for each round trip, 2 message put and 2 message get operations take place. For a single put/single get scenario, the peak performance that you might obtain in the same environment is over 60,000 msg/sec.

15 Frequently Asked Questions

Will I be able to use FASTPATH channels to send/receive messages into the MQ Appliance?

Yes - this is now the default MQIBindType as specified in the Channels stanza in the qm.ini configuration file.

How do I view and change QM settings on the MQ Appliance?

You can use the *dspmqini* command to view the QM configuration and *setmqini* to alter any configuration options. There are similar *dspmqvar* and *setmqvar* commands to view/alter environment variables.

What type of logging is used on the MQ Appliance?

Only circular logging is supported on the MQ Appliance, and thus facilities to monitor/prune QM recovery logs (useful for managing linear logs), are not required.

Can I run my existing user exits?

No – for appliance integrity, user exits will not be supported on the MQ Appliance. Many historic reasons for using code exits have now been resolved by product features.

What is throttling my messaging scenario?

If customers experience throttled performance when driving high throughput workloads on M2004A, they should check the following:

- Persistent workloads - Customers might encounter the limits of the RAID10 subsystem as illustrated in this document, although this is less likely with the fast NVMe storage available in the M2004 appliance
- Larger message (10KB+) Non-persistent workloads - Customers might encounter network limits depending on which interfaces are selected for workload traffic. Customers can select higher bandwidth connectivity or aggregate multiple interfaces.
- Small message (2KB-) Non-persistent workloads – Customers might encounter CPU saturation (Check MQ Console or CLI)
- All - Ensure that they have a sufficient number of concurrently connected clients to drive workload on the M2004 appliance

I have an M2003A/B, can I upgrade to M2004A/B?

The M2004 appliance has very different hardware to the M2003 appliance, therefore there is no hardware upgrade option.

16 Appendix A – Client machine specification

The two client machines used for the performance tests in this report have the following specification:

Category	Value
Machine	Lenovo ThinkSystem SR630 V4
OS	Red Hat Enterprise Linux Server 9
CPU	2x16 (3.6Ghz)
RAM	512GB RAM
Network	10/100Gb Ethernet
Disks	2x 3TB NVMe SSD in RAID-0 array
RAID	Linux mdraid MQ Logs hosted on RAID-0 partition

17 Appendix B – QM Configuration

The following commands and expect scripts were used to create the standalone Queue Managers for this report:

```
crtmqm -lp 64 -lf 16384 -h 5000 -fs 16 PERF0
setmqini -m PERF0 -s TuningParameters -k DefaultPQBufferSize -v 10485760
setmqini -m PERF0 -s TuningParameters -k DefaultQBufferSize -v 10485760


proc configureQM { QMname QMport QMqueues } {
    send "runmqsc $QMname\n"
    send "define listener(L1) trptype(tcp) port($QMport) control(qmgr)\n"
    send "start listener(L1)\n"
    send "alter channel(SYSTEM.DEF.SVRCONN) chltype(SVRCONN) sharecnv(1) maxmsgl(104857600)\n"
    send "alter qmgr maxmsgl(104857600)\n"
    send "alter qlocal(system.default.local.queue) maxmsgl(104857600)\n"
    send "alter qmodel(system.default.model.queue) maxmsgl(104857600)\n"
    send "alter qmodel(system.jms.model.queue) maxmsgl(104857600)\n"
    send "alter qmodel(system.jms.tempq.model) maxmsgl(104857600)\n"
    send "alter qlocal(system.dead.letter.queue) maxmsgl(104857600)\n"
    send "define channel(SYSTEM.ADMIN.SVRCONN) chltype(SVRCONN)\n"
    send "alter qmgr chlauth(disabled)\n"
    send "alter authinfo(SYSTEM.DEFAULT.AUTHINFO.IDPWOS) authtype(IDPWOS) chckclnt(OPTIONAL)\n"
    send "refresh security type(CONNAUTH)\n"
    send "define qlocal(queue) maxdepth(5000) replace\n"
    send "define qlocal(request) maxdepth(5000) replace\n"
    send "define qlocal(reply) maxdepth(5000) replace\n"
    for {set j 0} {$j <= $QMqueues} {incr j 1} {
        send "define qlocal(request$j) maxdepth(5000) replace\n"
        send "define qlocal(reply$j) maxdepth(5000) replace\n"
    }
    send "end\n"
}
```