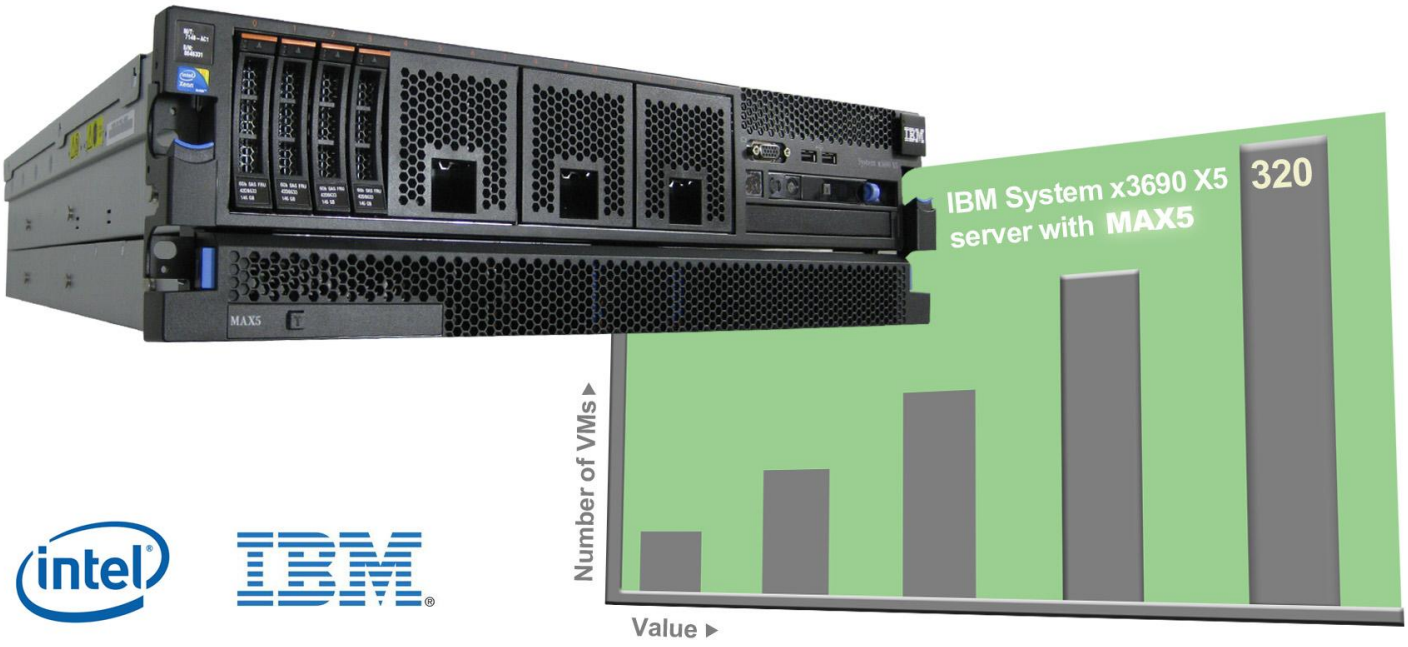
IBM System x server with MAX5: High virtualization density with confidence



Contents

Overview **4**

Performance and safety **4**

Why high density? **6**

Space 6

Power and cooling 6

IT capital expenditures (CAPEX) 6

eX5 technology extends the industry standard with IBM leadership to run your business-critical x86 workloads **7**

Memory capacity for the eX5 family 7

Processor power and features 7

Hardware features 7

Management software 8

Networking and Storage 8

Our testing scenario **9**

Workload 9

Topology 10

Our testing process 10

Catastrophic outage scenario 11

Planned shutdown scenario 11

Planned maintenance scenario 11

Detailed results **12**

Performance results 12

Catastrophic outage scenario 13

Planned shutdown scenario 13

Planned maintenance scenario 14

Timing results 17

Catastrophic outage scenario 17

Planned shutdown scenario 18

Planned maintenance scenario 19

Summary **19**

Appendix A – Detailed timing results **21**

Appendix B – Server and storage configuration information **26**

Appendix C – Detailed setup **28**
Physical server setup and MAX5 cabling 28
IBM XIV configuration and cabling 28
VMware vSphere (ESX) installation 29
Installing Windows Server and vCenter on the management server 29
Creating and configuring the cluster 29
Creating the VMs 30
Setting resource reservations 30
Appendix D – VMware vCenter Server and HA best practices **31**
VMware vCenter Server 31
General recommendations 31
Sizing recommendations 31
Database best practices 32
HA performance tuning 32
HA VM startup concurrency 33
Tuning Power On Frequency 33
Powering on additional VMs 33
HA best practices 33
Appendix E – IBM XIV Best Practices for VMware **36**
Key considerations 36
vStorage APIs for Array Integration (VAAI) 36
Queue depth 36
Multipathing 37
Zoning and cabling 37
LUN size and datastore 37
About Principled Technologies **38**

OVERVIEW

If you're responsible for virtualizing corporate applications, the prospect of putting a great many virtual machines (VMs) on a handful of servers may raise two concerns: performance and reliability.

Worry no more. We assembled a cluster of two IBM® System x3690 X5 servers with high-end Intel® Xeon® processors and MAX5 memory expansion to see how well the servers and IBM supporting technologies could handle the job. The systems successfully ran our Web server workload with increasing VM densities—up to 320 VMs on one server—with CPU and memory capacity to spare. We also simulated both planned and unplanned outages at each VM density level and found that, after failover and moving all VMs to the remaining system, a single IBM System x3690 X5 could support the maximum VMware-supported number of VMs: an astounding 320 VMs. With today's data centers and applications requiring ever-greater amounts of RAM, the IBM System x3690 X5 and MAX5 memory capacity are critical elements in providing necessary resources to high-density VM environments.

Results like these demonstrate that the IBM System eX5 systems with MAX5 memory expansion are ideal virtualization platforms for your enterprise applications. Get more for your server dollar, rest easy about outages, and boost your utilization—the IBM solution lets you go high density with confidence.

PERFORMANCE AND SAFETY

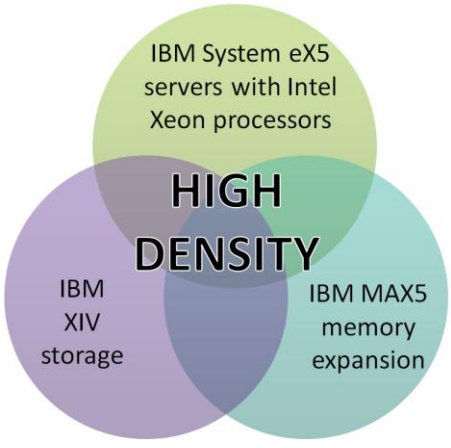


Figure 1: A winning combination. The IBM System eX5 servers, taking full advantage of the IBM MAX5 memory expansion drawer and using the IBM XIV storage, deliver ultra-high virtualization densities with ease.

For years, consolidating older servers and workloads through virtualization has been the order of the day. Studies show how using new servers running the latest virtualization technology to replace your old servers can save time, space, power usage, and of course, money. Maxing out a server with all the VMs that will fit sounds good in theory, and might work for a business that can afford the occasional chunk of downtime.

Your applications are essential to your business, though. If you've stuck a toe in the water of virtualized consolidation, you're likely underutilizing your hardware—typical enterprise data center usage hovers around only 15 to 20 VMs per server—and crossing your fingers that you won't have an outage. That has made sense, but IBM now offers a solution that will let you go to high virtualization densities and relax about the consequences of high availability events due to things such as power loss, network isolation, or storage isolation.

To investigate the possibilities of high-density virtualization in a real-world scenario, Principled Technologies set up a cluster of two IBM System x3690 X5

servers running VMware® vSphere™ 4.1 with High Availability (HA) enabled. This solution, which takes advantage of IBM MAX5's ability to effectively double available RAM, successfully ran from 18 to 160 virtual machines per server prior to our simulated outage, and up to 320 virtual machines—the VMware maximum—on one server after our simulated outage, and still had CPU and memory capacity to spare. While we used a Web server workload, other workloads—including but not limited to Active Directory®, file and print services, and even a virtual desktop infrastructure (VDI) solution—would also be excellent candidates for this highly dense virtualization scenario. The IBM System eX5 server family with MAX5 is also a prime candidate for other VM density scenarios, where, for example, you might need to run fewer VMs with larger VM RAM requirements.

To learn about this cluster's reliability, we simulated three scenarios at increasing VM density levels on our cluster: 36 VMs, 60 VMs, 100 VMs, 200 VMs, and 320 VMs. These three scenarios were a catastrophic outage scenario, which exercised VMware HA; a planned shutdown scenario, which exercised VMware HA; and a planned maintenance window, which exercised VMware vMotion technology. We measured performance of the Web workload and the HA or vMotion timings for all three scenarios.

In response to both the catastrophic outage scenario and the planned shutdown scenario, both of which exercised VMware HA, the cluster flawlessly restarted the VMs on the remaining server, where all VMs ran easily. For these HA events, it made these adjustments automatically. For our specific timing results on the VMware HA event, see our Detailed results section.

To test our planned maintenance scenario by exercising VMware vMotion, we initiated planned host evacuations of 18, 30, 50, 100, and 160 VMs, where we placed half of the total cluster VMs on each server. In each case, the VMs were migrated from one server to the other server, transparent to the user. The remaining server then ran each of the scenarios' VM workloads with no issues. For specific vMotion timing results, see the Detailed Results section.

With performance and reliability capabilities like these, IBM System eX5 servers with MAX5 memory expansion can position your enterprise to reap the benefits of high-density virtualization without the risk of losing data or revenue.

Note: We used two servers for this test to maximize the workload on the servers and show the high densities that the solution is capable of supporting. In a real-world implementation, adding three or more servers would enable

administrators to expand on this concept and achieve even greater balance, and spread the load more evenly while maintaining a strong high availability position.

WHY HIGH DENSITY?

In the late 1990s, the dominance of multi-tier applications and the drop in prices of x86 architecture servers fueled physical server sprawl for many companies. With this server sprawl came problems such as high power consumption, increasing physical space utilization, and cooling issues as companies faced ever-growing business demands. Intra-company organizations "siloed" their applications to avoid co-locating with other business applications. Virtualization began to address some of these issues in the mid-2000s as companies learned to consolidate their workloads and applications in order to increase utilization of their hardware. But until now, with massive amounts of RAM available, such as the 1TB capacity available on the IBM System x3690 X5 with MAX5 (using 16GB DIMMs), companies have not been able to reliably consolidate in a highly dense manner.

By using the IBM System eX5 platform, you can move towards a highly dense virtualization configuration, increasing utilization of your hardware and lowering total overall IT cost by addressing some of the issues we note below.

Space

As you consolidate servers, you can reduce the number of racks or even the number of data centers that house corporate servers.

Power and cooling

Fewer servers consume less power, especially when those newer servers, such as the IBM System x3690 X5, use power more efficiently than older servers.

IT capital expenditures (CAPEX)

CAPEX, the cost of obtaining physical assets for business operations, is directly applicable to the dense virtualization discussion. Buying, powering, and supporting fewer servers brings obvious CAPEX savings. Other potential CAPEX cost savings include a need for less ancillary equipment, such as racks and network switches; as the number of servers decreases, these costs decrease as well. Additionally, over time, the decreased costs for these CAPEX items will bring decreases in operational expenses, such as administrative and maintenance time, power consumption, and software licensing.

eX5 TECHNOLOGY EXTENDS THE INDUSTRY STANDARD WITH IBM LEADERSHIP TO RUN YOUR BUSINESS-CRITICAL X86 WORKLOADS

The IBM ecosystem brings a unique convergence of hardware features, management software, and storage solutions that enable reliable high-density virtualization. Along with the IBM System x3690 X5 server, the IBM eX5 server family includes the IBM System x3850 X5 and the IBM BladeCenter HX5. The IBM System x3690 X5, which we tested for this study, possesses these features that boost virtualization capabilities, which we highlight below.

MAX5 increases memory capacity by

32 DIMM slots

Memory capacity for the eX5 family

The more VMs you have, the more memory you need. Unlike other hardware platforms, where this means greater capital expenditure for additional server capacity, the IBM System eX5 family of servers with MAX5 lets you save your money. In the case of the IBM System x3690 X5 server, the IBM MAX5 memory expansion drawer doubles the already large amount of memory available to the system without adding more processors, providing extreme memory expansion for virtualized environments. This scalable 1U memory expansion drawer delivers an additional 32 DIMM slots to the IBM System x3690 X5 with a memory controller for added performance, a key innovation in the IBM System eX5 architecture. IBM's unique ability to separate the memory infrastructure of the server from the rest of the server components allows you to support more RAM without purchasing additional server infrastructure—processors, network cards, storage controllers, and so on. This makes IBM System eX5 servers flexible and well suited to rapidly changing virtualized environments.

Processor power and features

With more than 20 new reliability, availability, and serviceability (RAS) features; many cores with hyper-threading; Turbo Boost Technology; and advanced virtualization technologies at the processor level, the latest generation Intel Xeon processors powerfully equip your IBM systems to reliably run your applications continuously, day after day, and to support high VM densities. Now, because of the new reliability features in Intel Xeon processors, you can virtualize with confidence and high performance.

Hardware features

- **Light-path-diagnostic LEDs.** These LEDs, located at various spots on the server, provide diagnostic information on components—processor, memory, power supply, fan, and more—without interrupting system operations. This reduces time for hardware repairs.

- **Predictive Failure Analysis (PFA).** This feature detects when components or hard drives are operating outside of set of pre-defined thresholds, and helps increase uptime by letting you receive proactive alerts via IBM Systems Director.
- **Memory reliability features through IBM Active Memory™.** Enhanced error recovery with Chipkill error correction, memory scrubbing, Memory ProteXion, and memory mirroring help maintain data integrity.

Management software

IBM Systems Director VMControl™ Enterprise Edition utilizes a workload-optimized approach to decrease infrastructure costs and improve service levels. VMControl Enterprise Edition provides simplified virtualization management, which enables faster problem solving, a higher return on investment, and rapid responses to changing business goals and strategies. IBM Systems Director and VMControl Enterprise Edition work together to help you more effectively utilize your virtual environment. VMControl Enterprise Edition helps you to create and modify virtual farms, make dynamic virtual workload adjustments, and move workloads within system pools, resulting in an optimized virtual environment with increased resilience to cope with planned or unplanned downtime. By combining VMware technologies with IBM VMControl technology, an organization can lower response times to outages and increase management capabilities.

Networking and Storage

IBM Blade Network Technology (BNT) brings incredible networking power, speed, and efficiency to the IBM ecosystem. In addition, IBM VMready® technology, functional with multiple hypervisors, provides a powerful new set of networking features. With VMready, organizations can enable their networks to be VM aware, and minimize networking administration. To learn more about IBM networking technologies, visit http://www-03.ibm.com/systems/networking/.

Successful high virtualization density requires a large amount of high-quality storage. In our testing, we paired the IBM System x3690 X5 servers with the IBM XIV® Storage System. IBM XIV uses a new architecture that safely allows for massive amounts of capacity with simplified storage management. We provide IBM XIV best practices when implementing this storage solution with VMware in Appendix E. To learn more, visit http://www-03.ibm.com/systems/storage/disk/xiv/.

OUR TESTING SCENARIO

How many VMs can a server support under regular working conditions? And what happens in an HA event if an outage occurs? To answer these vital questions, we configured a cluster of two IBM System x3690 X5 servers and IBM MAX5 memory expansion modules running VMware vSphere 4.1 with HA enabled. We also configured the vCenter Server to manage cluster activity, configured a Web load balancer, and configured our workload machines.

To start, we configured the cluster with 36 VMs—18 per server, typical in enterprises—and measured performance in those VMs using a benchmarking tool that simulates a Web server workload. Next, we tested our three outage scenarios: catastrophic outage, planned shutdown, and planned maintenance. For each scenario, we measured performance before the HA or vMotion event, noted how long it took the VMs on that server to either restart on (HA) or migrate to (vMotion) the remaining server, and then measured the performance of all VMs on that server. We repeated this scenario four times, increasing the number of VMs on each server. We tested VM densities up to the VMware-supported maximum of 320 VMs per host.

Workload

For our Web workload, we used WebBench 5.0 (128-bit US version), an industry-standard benchmark for Web server software and hardware, which uses workstation clients to send Web requests to servers. Our WebBench workload ran on workstation-class machines, and the workload targeted static homogenous Web sites on each VM consisting of HTML and JPG images. Each VM ran Microsoft® Windows Server® 2008 R2 along with the IBM HTTP Server, a Web server based on the Apache HTTP Server.

For our WebBench test runs, we configured the benchmark to run for two 15-minute periods, totaling 30 minutes for the test scenarios. The only exception to this test duration was for the vMotion scenarios at 200 and 320 VMs, which we ran for four 15-minute periods, totaling 60 minutes. This allowed time for the vMotion migrations to complete, as VMware vSphere 4.1 supports eight concurrent vMotions.

In WebBench, we configured each client to run four engines, or instances, of the benchmark. As our VM density increased, we increased the

number of WebBench clients accordingly. Figure 2 shows the number of WebBench clients and engines we used at each VM density.

	36 VMs	60 VMs	100 VMs	200 VMs	320 VMs
WebBench clients	2	3	5	10	16
Engines per WebBench client	4	4	4	4	4
Total engines	8	12	20	40	64

Figure 2: Number of engines and clients per VM density test run.

Topology

Figure 3 shows our test topology, with varying numbers of clients targeting our varying numbers of VMs on the IBM x3690 X5 cluster running VMware vSphere.

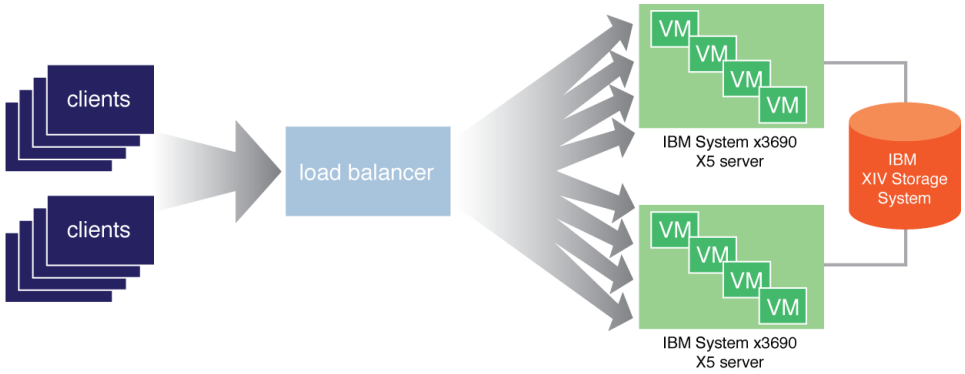


Figure 3: Our test topology.

To balance the Web load evenly amongst all VMs during our benchmarking period, we used a separate physical machine running VMware vSphere and a large-footprint Linux VM running HAProxy. HAProxy is a fast, reliable, and free load balancing and proxy application for HTTP applications. We reconfigured HAProxy between each test run, adjusting the number of backend servers to equal the number of VMs for that test run. This HAProxy VM used four virtual processors and 8 GB of RAM. For more information on HAProxy, see the HAProxy Web site at http://haproxy.1wt.eu/.

Our testing process

We used the following structured process for each of our three test scenarios:

1. We configured the appropriate number of VMs per server: 18, 30, 50, 100, or 160.

2. We powered on the VMs on their respective nodes, and restarted our client workstations.
3. We reconfigured our load balancer running HAProxy to target the appropriate number of backend VMs (36, 60, 100, 200, or 320).
4. We stopped the IBM HTTP Server service on each VM, removed the access.log and error.log files, and restarted the IBM HTTP Server service. We used these log files to analyze our results.
5. We started the WebBench workload.
6. At the 10-minute mark, we induced our event type: either catastrophic, planned, or vMotion migration (see below).
7. We monitored the servers' performance and VM recovery times and gathered output data upon test completion.

Catastrophic outage scenario

No matter your situation, you prepare for the remote chance of an unplanned catastrophic outage, including power loss to a server or storage or network cable damage. To test this scenario, we performed the steps listed above for each VM density: 36, 60, 100, 200, and 320. At the 10-minute mark, we removed power to the server by using the "Power Off Immediately" option in the IBM Integrated Management Module (IMM), the equivalent of pulling the plug. The cluster correctly identified the loss of the node and initiated an HA failover process on the cluster.

Planned shutdown scenario

Some situations, such as receiving an urgent PFA alert, can require administrators to quickly shut down their servers. To test this scenario, we performed the steps listed above for each VM density: 36, 60, 100, 200, and 320. At the 10-minute mark, we cleanly shut down one of the servers containing half the VMs by using the "shutdown" command at the service console in VMware vSphere. The cluster again correctly identified the loss of the node and initiated an HA failover process on the cluster.

Planned maintenance scenario

Most organizations have the need for planned maintenance windows on hardware, but uptime for users is a great concern. VMware's vMotion technology allows administrators to migrate VM workloads from one server to another transparently without end users being affected. During the vMotion process, the source server prepares the VM for migration, the state of the active virtual machine is transmitted over the VMkernel network designated for vMotion, and the VM is registered on the new host, all transparent to the user

with no data loss. To test this scenario, we performed the steps listed above for each VM density: 36, 60, 100, 200, and 320. At the 10-minute mark, in the vCenter console we selected all VMs on one of the servers containing half the VMs and initiated a migration. This started the vMotion evacuation process, and all VMs moved to the remaining node.

DETAILED RESULTS

Performance results

For all testing scenarios, we configured the cluster with increasing numbers of VMs—36, 60, 100, 200, and 320. In all scenarios, we placed half the VMs on one of the IBM System x3690 X5 servers and half the VMs on the remaining IBM System x3690 X5 server. We then started our benchmark to place load on the VMs, induced either a catastrophic outage, planned shutdown, or planned maintenance event depending on the scenario, and measured recovery time and performance during the test run. In this section, we show results for each scenario.

In Figure 4, we show the average requests per second the cluster serviced both before the HA or vMotion event while running on two nodes, then after the HA or vMotion event with the workload running on one node. To calculate the average requests per second after the HA event, we waited until all VMs had restarted on or migrated to the destination server.

Scenario	Number of nodes	36 VMs	60 VMs	100 VMs	200 VMs	320 VMs
Catastrophic outage	Two (before HA event)	2,104	2,820	3,717	4,593	5,417
	One (after HA event)	2,028	2,667	3,607	4,439	5,305
Planned shutdown	Two (before HA event)	2,076	2,824	3,732	4,573	5,376
	One (after HA event)	2,035	2,652	3,639	4,441	5,339
Planned maintenance	Two (before vMotion)	2,061	2,828	3,702	4,552	5,380
	One (after vMotion)	1,965	2,680	3,662	4,352	5,323

Figure 4: Overview of performance, in requests per second, of all HA test scenarios.

Catastrophic outage scenario

In Figure 5, we show the cluster-wide requests per second as reported by the IBM HTTP Server access log file at each VM density for our catastrophic outage scenario. The catastrophic outage, in this case a power loss on our first node, can clearly be seen at the 10-minute mark. From the 10-minute mark, onwards throughout the duration of the test, all Web requests were serviced by one node. VMware HA automatically restarted the failed node's VMs on the remaining server node. Note: While a specific VM on the failed server was restarting on the other node, the specific VM could not process any Web requests until the VM restart time was complete. We present these VM restart times in the Timing results section. Because the load balancer in our scenario immediately distributed the incoming requests to our remaining node's VMs, requests could be serviced throughout the recovery period. After the recovery period, all VMs ran on one node.

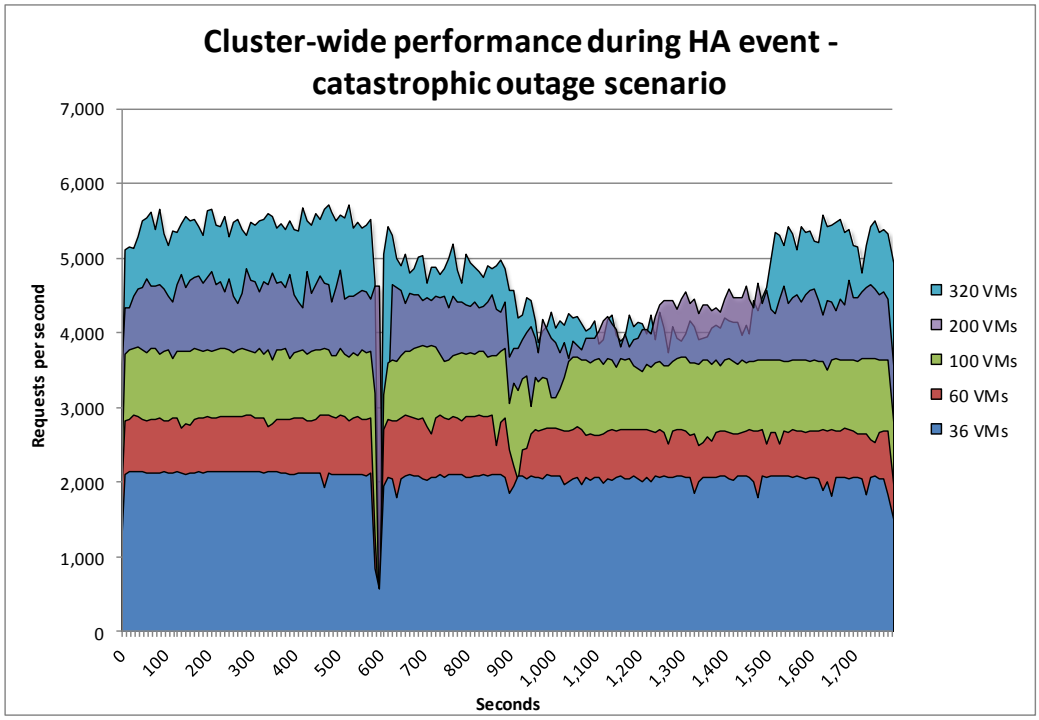


Figure 5: Cluster-wide requests per second at each VM density for our catastrophic outage scenario.

Planned shutdown scenario

In Figure 6, we show the cluster-wide requests per second as reported by the IBM HTTP Server access log file at each VM density for our planned shutdown scenario. The shutdown of our first node, which triggers the HA event, can be clearly seen at the 10-minute mark. From the 10-minute mark,

onwards throughout the duration of the test, one node serviced all Web requests. VMware HA automatically restarted the failed node's VMs on the remaining server node. Note: While a specific VM on the failed server was restarting on the other node, the specific VM again could not process any Web requests until the VM restart time was complete. We present these VM restart times in the Timing results section. Because the load balancer in our scenario immediately distributed the incoming requests to our remaining node's VMs, requests could be serviced throughout the recovery period. After the recovery period, all VMs ran on one node.

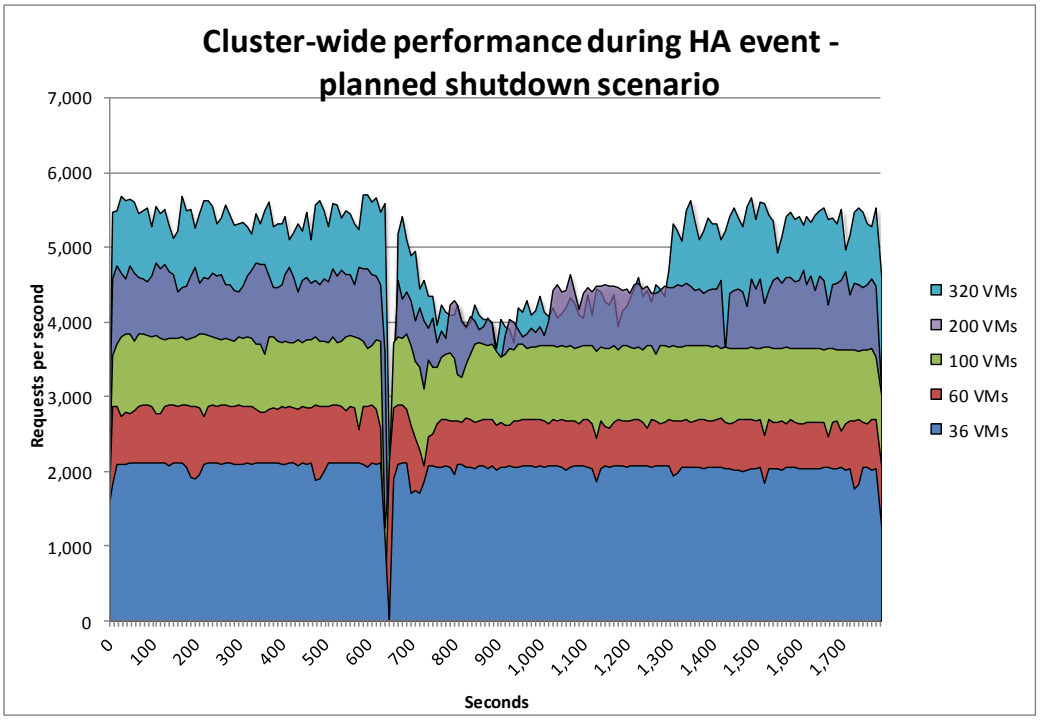


Figure 6: Cluster-wide requests per second at each VM density for our planned shutdown scenario.

Planned maintenance scenario

In Figure 7, we show the cluster-wide requests per second as reported by the IBM HTTP Server access log file at the 36, 60, and 100 VM densities for our planned maintenance scenario. The start of our vMotion event can be clearly seen at the 10-minute mark. From the 10-minute mark, the cluster migrated VMs from one server to the other, transparently transitioning the VMs to the destination node with no VM restart required. For the 200 and 320 VM density tests, we ran our test for a 60-minute period. We present those results in Figure 8.

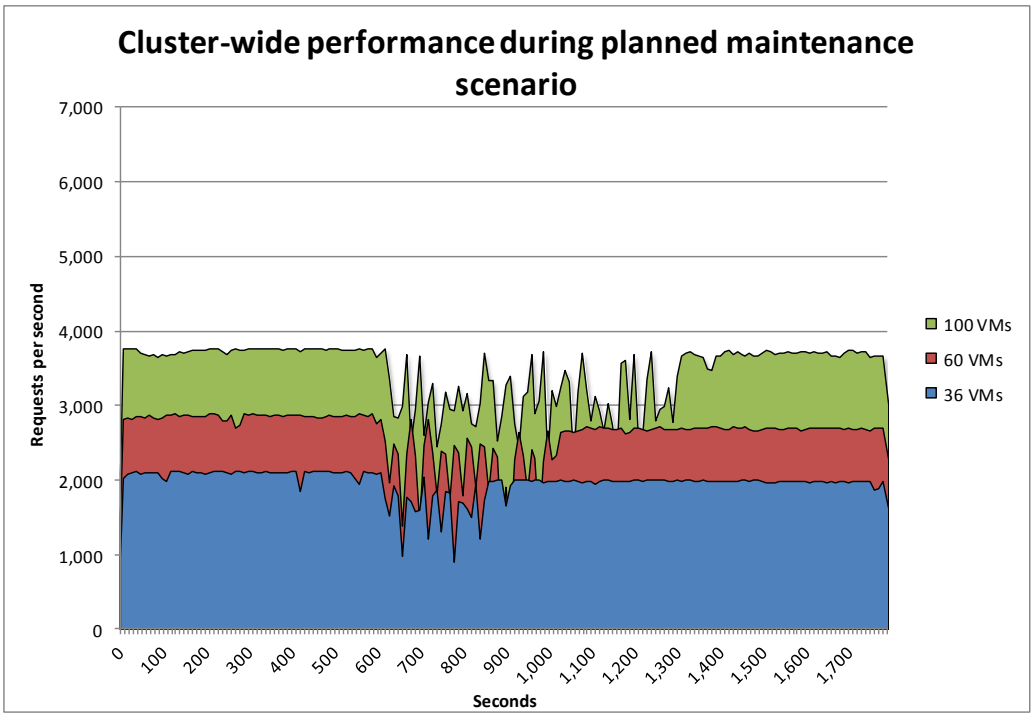


Figure 7: Cluster-wide requests per second for 36 VMs, 60 VMs, and 100 VMs for our planned maintenance scenario.

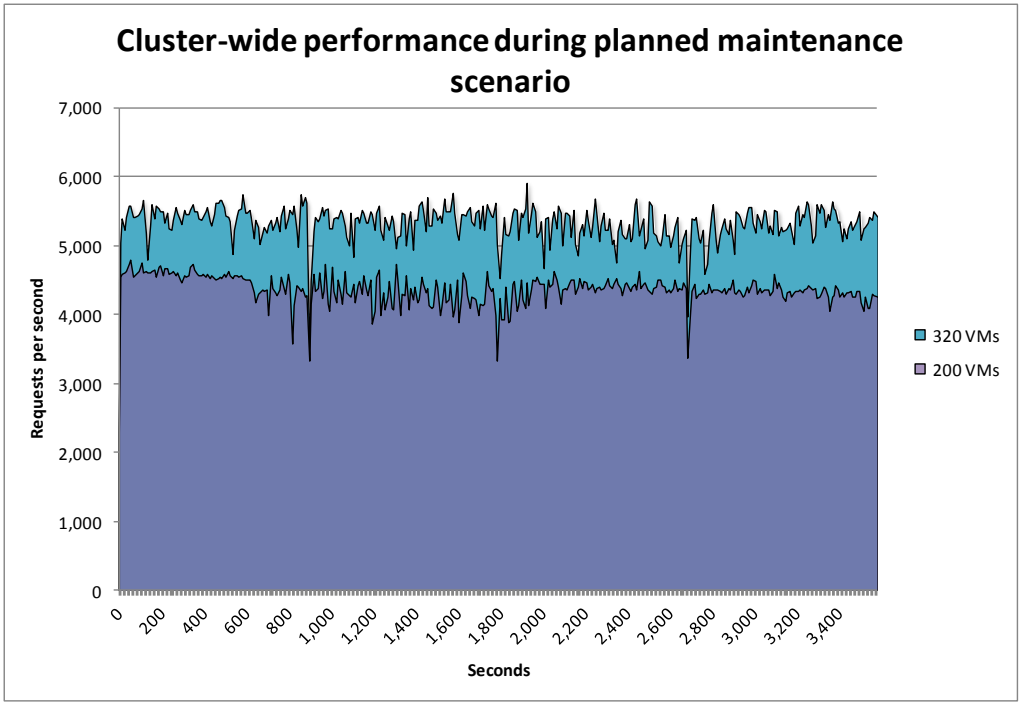


Figure 8: Cluster-wide requests per second for 200 VMs and 320 VMs for our planned maintenance scenario.

When vMotion occurs on a VM, the vSphere cluster uses shared storage to allow multiple hosts access to a particular VM's files. Active memory and the execution state are then immediately transferred over the vMotion network, and control of that VM is passed to the new node. This node transition time is extremely rapid. For most Web-based applications, as well as other applications, this movement would not be noticeable to end users, an enormous benefit for this type of event. In our test scenario, after the migration period, all VMs ran on one node.

Figure 9 further demonstrates the vMotion maintenance host-evacuation process, and its effects on each system's performance. You can see the CPU core utilization percentage on each server during the 160-VM evacuation during our testing. As each VM is migrated off the source server, the CPU utilization drops on the source server while the destination server's CPU utilization increases.

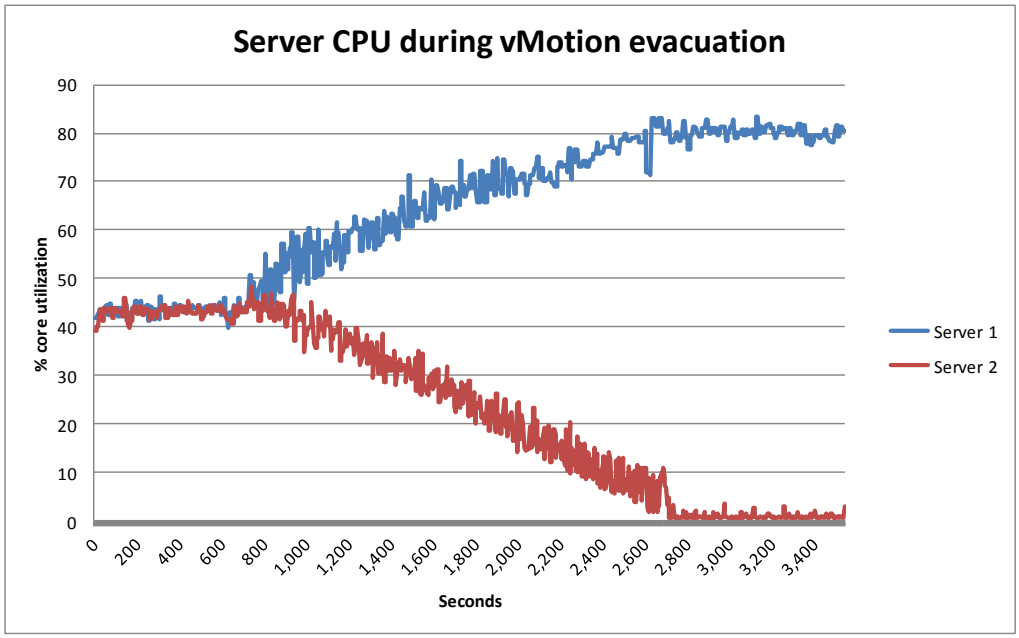


Figure 9: CPU core utilization on each of our two servers during our planned maintenance vMotion maintenance testing.

Timing results

For all testing scenarios, in addition to measuring performance data in requests per second, we measured the time it took for the VMs to either restart on (HA) or migrate to (vMotion) the remaining cluster node.

In Figure 10, we show the minimum, maximum, and average times it took at each VM density for these events to occur. For the HA scenarios where the VM actually restarted, we measure from either the time of power loss or shutdown command until the IBM HTTP Server service was available on the VM after its restart. For the vMotion scenario, we measure the migration time of each VM as reported by the vCenter console, with the start time coinciding with our executing the vMotion command.

Scenario/event		36 VMs	60 VMs	100 VMs	200 VMs	320 VMs
Catastrophic outage—Time to restart VM	Minimum	326	326	345	334	359
	Maximum	465	369	466	670	922
	Average	366	343	402	496	630
Planned shutdown—Time to restart VM	Minimum	115	121	126	133	147
	Maximum	128	160	241	443	717
	Average	120	136	181	278	407
Planned maintenance—Time to migrate VM	Minimum	70	69	60	63	74
	Maximum	410	478	740	1,413	2,313
	Average	187	272	403	750	1,212

Figure 10: Timing results for each test scenario by VM density, in seconds.

Catastrophic outage scenario

Figure 11 shows the failover times of all VMs per density for our catastrophic outage scenario. Failover time stays roughly the same through 100 VMs, well under 10 minutes.

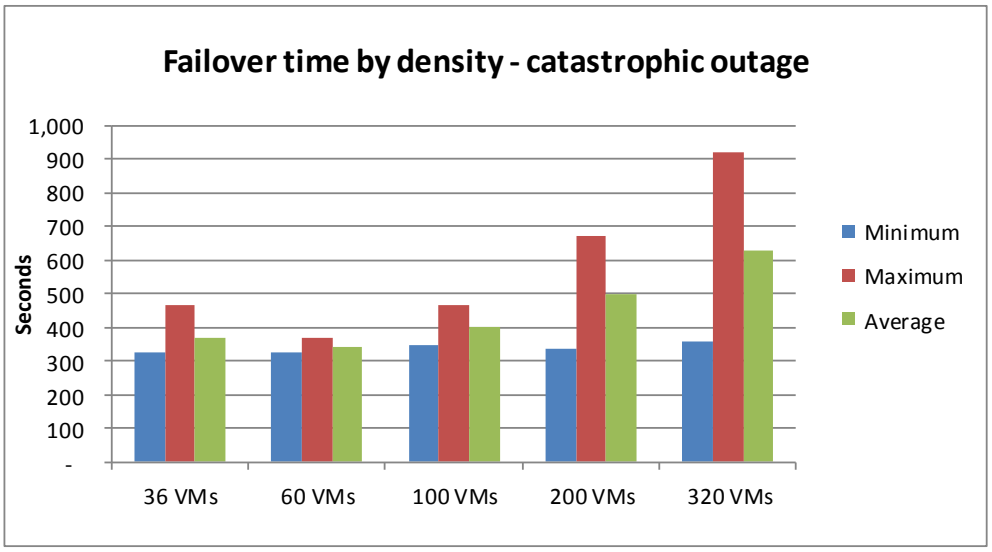


Figure 11: Failover time for the catastrophic outage scenario for each VM density, in seconds.

Planned shutdown scenario

In Figure 12, we show the failover times of all VMs per density for our planned shutdown scenario. Average failover times are all well under 10 minutes.

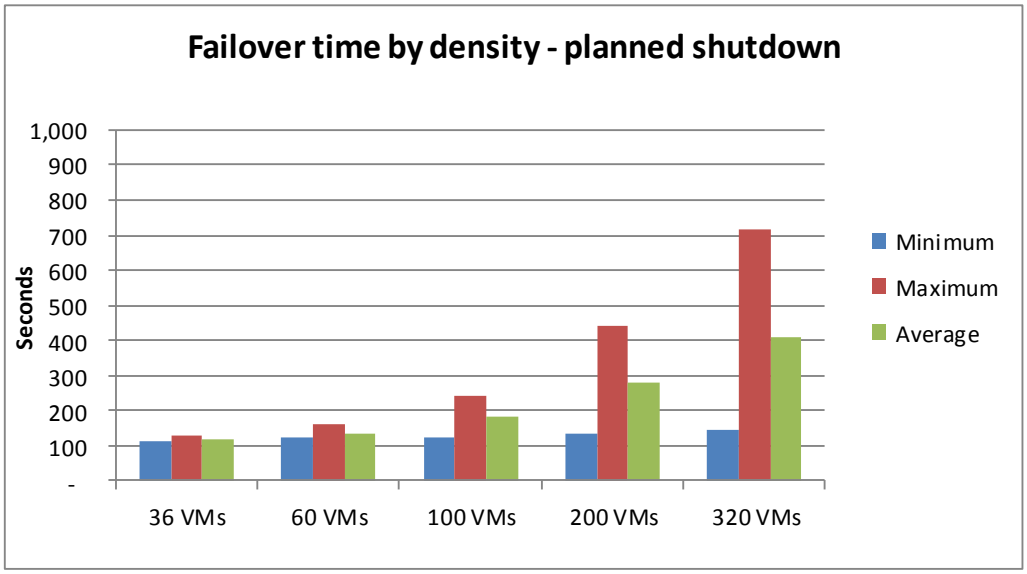


Figure 12: Failover time for the planned shutdown scenario for each VM density, in seconds.

Planned maintenance scenario

In Figure 13, we show the minimum, maximum, and average vMotion migration times of all VMs per density for our planned maintenance scenario. Although some vMotion migrations took longer than some of the HA VM restarts, the advantage of vMotion is that Web users lose no work, as the VMs are not restarted, but migrated.

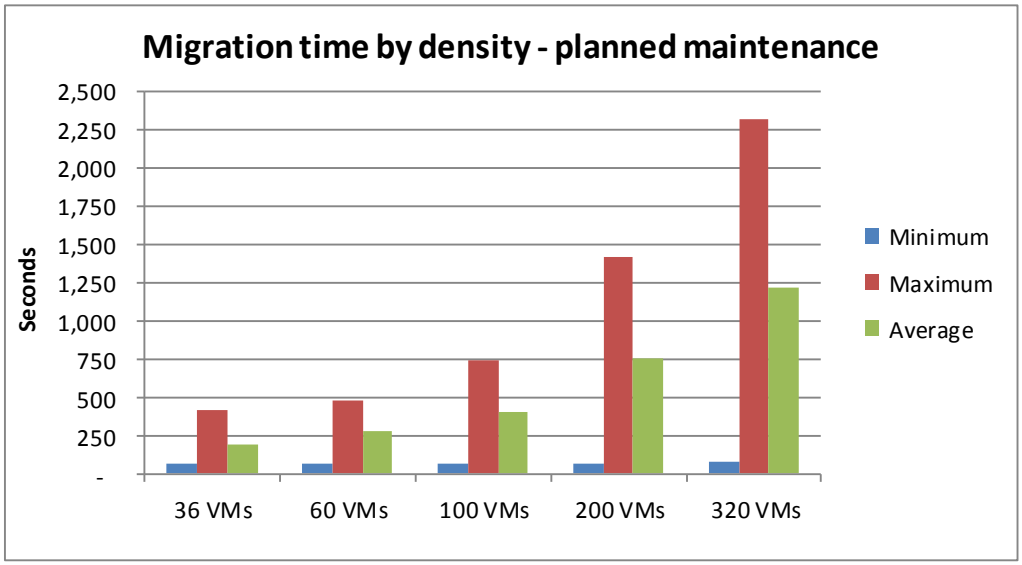


Figure 13: Migration time for the planned maintenance scenario for each VM density, in seconds.

SUMMARY

As our tests demonstrated, the IBM System x3690 X5 server with the latest high-end Intel Xeon processors and MAX5 memory expansion running VMware vSphere 4 with High Availability delivered excellent performance and quick failover and migration rates across all three scenarios we tested.

In the catastrophic outage and planned shutdown scenarios, the affected VMs were back, on average, in 10.5 minutes or less (less than 10 minutes for up to 200 VMs), and were performing at very close to pre-event levels. This means that if a 10-minute outage falls within the bounds of your service level agreement, you could comfortably run up to 100 VMs per server. In the planned maintenance scenario, the solution took slightly longer to achieve pre-event performance levels, which would be acceptable in a real-world environment because users lose no work when using vMotion.

This data proves that the risks of high density virtualization are actually less than typically perceived – higher VM densities will drive virtualization return on investment (ROI) up, enabled by the IBM eX5 server family .

These findings demonstrate that the IBM System x3690 X5 server with the latest high-end Intel Xeon processors and MAX5 memory expansion running VMware vSphere 4 with High Availability is a robust and reliable platform for virtualizing your enterprise applications. It provides excellent scaling to allow you to get the most for your server dollar and handles failover seamlessly, letting you rest easy and know your data and revenue are safe.

Not only does the solution provide unique functionality and confidence, it can save you money. The IBM System x3690 X5 is a two-socket system that delivers solid performance and extreme memory scalability.

Simply put, the IBM System x3690 X5 with MAX5 offers superlative scalability and reliability for your enterprise applications.

APPENDIX A – DETAILED TIMING RESULTS

Figure 14 shows the aggregated average, minimum, and maximum breakdown of the VM timings the IBM System x3690 X5 server cluster handled during each type of scenario event while running 36, 60, 100, 200, and 320 VMs across the two servers. Figure 15 shows the individual results for each VM. For the catastrophic outage scenario, we measure the number of seconds from the time the power loss on one server occurred until the IBM HTTP Server service was available and ready to receive requests again on the relevant VM after their reboot on the remaining cluster node. For the planned shutdown scenario, we measure the number of seconds from the time the shutdown command was issued on one server until the IBM HTTP Server service was available and ready to receive requests again on the relevant VM after their reboot on the remaining cluster node. For the planned maintenance scenario, we measure the number of seconds from the time the vMotion migration was initiated on one server until the migration was complete.

	Seconds for VM to reboot - catastrophic outage					Seconds for VM to reboot - planned shutdown					Seconds for VM to migrate - planned maintenance				
	18 VMs	30 VMs	50 VMs	100 VMs	160 VMs	18 VMs	30 VMs	50 VMs	100 VMs	160 VMs	18 VMs	30 VMs	50 VMs	100 VMs	160 VMs
Avg	366	343	402	496	630	120	136	181	278	407	187	272	403	750	1,212
Min	326	326	345	334	359	115	121	126	133	147	70	69	60	63	74
Max	465	369	466	670	922	128	160	241	443	717	410	478	740	1,413	2,313

Figure 14: Aggregated VM timing results for each scenario.

	Seconds for VM to reboot - catastrophic outage					Seconds for VM to reboot - planned shutdown					Seconds for VM to migrate - planned maintenance				
VM	18 VMs	30 VMs	50 VMs	100 VMs	160 VMs	18 VMs	30 VMs	50 VMs	100 VMs	160 VMs	18 VMs	30 VMs	50 VMs	100 VMs	160 VMs
1	331	368	454	664	908	124	122	236	437	710	140	109	60	113	158
2	463	360	366	392	364	117	150	126	141	175	410	69	150	93	94
3	464	356	394	356	635	117	134	176	163	432	160	109	150	63	199
4	333	338	401	506	840	125	121	204	322	582	260	109	180	163	124
5	465	337	423	394	359	117	123	129	174	228	140	148	120	143	125
6	327	335	438	598	886	128	151	218	389	669	290	278	90	143	74
7	326	330	398	536	480	124	122	180	322	266	160	248	90	143	134
8	327	343	455	614	547	119	152	218	416	323	230	148	120	483	75
9	464	329	374	435	700	116	157	128	233	508	230	148	340	253	245
10	327	327	396	549	505	118	131	219	343	291	190	69	380	283	595
11	327	329	424	532	838	124	122	176	324	617	70	298	180	723	198
12	326	356	432	598	534	124	130	233	400	317	70	188	272	93	225
13	333	338	345	460	726	123	122	149	241	508	110	248	320	193	334
14	463	358	345	334	366	117	132	127	141	150	190	278	180	303	454
15	328	337	377	429	727	116	124	148	226	488	110	448	400	364	788
16	326	356	373	495	798	124	154	170	290	593	110	378	386	843	474
17	327	350	398	343	666	116	122	153	161	463	260	188	740	423	434
18	328	340	377	435	726	115	132	130	234	532	230	218	211	513	384
19		343	346	454	442		131	128	251	237		449	272	664	384
20		329	347	360	692		124	148	198	461		378	430	603	546

	Seconds for VM to reboot - catastrophic outage					Seconds for VM to reboot - planned shutdown					Seconds for VM to migrate - planned maintenance				
VM	18 VMs	30 VMs	50 VMs	100 VMs	160 VMs	18 VMs	30 VMs	50 VMs	100 VMs	160 VMs	18 VMs	30 VMs	50 VMs	100 VMs	160 VMs
21		340	452	588	852		157	218	383	671		218	211	765	1,139
22		329	348	501	768		151	177	275	565		338	430	333	595
23		328	391	354	360		130	151	184	199		298	381	623	405
24		329	345	519	457		131	154	308	268		338	460	754	855
25		360	453	659	595		160	239	436	355		409	490	543	1,022
26		369	449	615	881		156	234	406	694		449	610	334	788
27		362	447	601	542		129	230	389	338		409	271	543	299
28		360	398	657	397		123	148	143	154		338	520	663	494
29		326	434	533	488		129	174	324	277		378	321	223	674
30		328	434	588	875		152	220	361	656		478	611	664	299
31			454	653	576			234	428	366			680	283	948
32			394	340	665			140	182	429			670	813	818
33			455	620	576			239	425	344			580	573	1,020
34			348	340	604			128	183	387			290	843	1,515
35			398	391	420			130	158	178			580	943	1,247
36			376	492	460			129	285	240			520	223	1,787
37			466	665	595			236	437	374			681	943	862
38			433	559	500			220	351	288			731	873	1,054
39			398	502	757			129	290	566			231	1,185	674
40			347	525	786			178	311	592			550	773	754
41			460	654	922			239	443	717			231	943	1459
42			454	619	547			241	422	339			340	983	1,694
43			436	604	889			219	399	676			550	694	1,144
44			376	670	625			176	144	395			490	573	1,579
45			423	531	804			178	325	610			460	453	924
46			375	435	394			155	220	152			610	1,283	724
47			376	526	779			172	289	566			640	883	1,213
48			433	575	843			217	364	660			520	873	974
49			346	504	770			211	265	537			731	1,263	1,097
50			347	340	901			146	143	153			700	1,014	1,184
51				343	638				133	429				513	494
52				502	757				282	567				193	1,514
53				656	906				181	154				1,095	866
54				572	819				359	621				484	1,894
55				515	459				283	261				843	1,274
56				594	544				381	314				1,074	333
57				363	698				176	462				1,013	634
58				390	395				182	180				1,263	298
59				337	602				158	397				1,363	1,246
60				559	510				355	301				1,413	1,418
61				398	699				218	496				1,124	1,614
62				647	914				435	700				394	1,477
63				357	656				182	434				724	697
64				485	756				266	539				1,033	525
65				389	622				188	424				1,153	1,213
66				391	670				225	468				364	697

	Seconds for VM to reboot - catastrophic outage					Seconds for VM to reboot - planned shutdown					Seconds for VM to migrate - planned maintenance				
VM	18 VMs	30 VMs	50 VMs	100 VMs	160 VMs	18 VMs	30 VMs	50 VMs	100 VMs	160 VMs	18 VMs	30 VMs	50 VMs	100 VMs	160 VMs
67				437	419				241	177				1,033	1,328
68				573	504				344	291				253	1458
69				359	668				161	451				1,323	224
70				453	734				260	508				1,073	1,694
71				358	419				159	195				1,223	2,014
72				459	459				268	262				423	1,894
73				620	573				413	349				793	1,054
74				596	869				380	657				1,363	537
75				565	508				357	317				693	1,665
76				649	903				420	701				1,135	1,972
77				500	470				310	272				304	1,827
78				397	658				196	445				1,263	1,665
79				460	415				251	238				1,283	1,097
80				390	630				133	409				1,323	1,877
81				606	541				406	342				1,386	1,053
82				640	597				434	353				1,185	753
83				665	392				151	177				1,185	1,613
84				455	415				259	235				1,203	1,814
85				638	574				424	348				394	652
86				533	452				325	289				1,125	1,387
87				463	728				237	509				1,087	1,246
88				535	840				345	616				793	1,363
89				339	366				188	203				1,323	404
90				494	771				259	536				913	2067
91				454	418				253	243				453	1,417
92				550	818				325	619				754	2,198
93				356	441				157	202				1,386	2,112
94				638	546				425	349				983	2,045
95				391	420				160	148				513	1,534
96				456	418				237	196				623	1,742
97				338	398				163	176				1,223	298
98				353	363				143	149				603	553
99				428	696				200	465				913	2,100
100				342	908				140	177				983	404
101					506					267					974
102					421					175					1,392
103					623					401					1,967
104					738					508					1,768
105					883					700					1,151
106					487					267					1,363
107					820					650					1,760
108					545					326					574
109					733					539					1,705
110					469					246					2,047
111					900					683					1,314
112					370					152					974

	Seconds for VM to reboot - catastrophic outage					Seconds for VM to reboot - planned shutdown					Seconds for VM to migrate - planned maintenance				
VM	18 VMs	30 VMs	50 VMs	100 VMs	160 VMs	18 VMs	30 VMs	50 VMs	100 VMs	160 VMs	18 VMs	30 VMs	50 VMs	100 VMs	160 VMs
113					840					652					1,274
114					880					703					1,845
115					573					378					1,164
116					699					497					2,137
117					888					691					2,078
118					796					565					1,073
119					442					194					404
120					601					407					884
121					798					565					2,066
122					419					197					2,265
123					732					510					724
124					481					239					1,966
125					603					407					1,314
126					898					661					1,634
127					601					385					1,742
128					695					449					1,579
129					578					363					2,051
130					779					589					1,417
131					480					292					824
132					591					384					1,815
133					771					602					884
134					665					455					1,877
135					920					702					2,264
136					394					177					1,579
137					505					287					1,934
138					788					608					1,163
139					486					271					2,014
140					779					605					2,197
141					443					222					2,237
142					538					314					2,274
143					360					155					924
144					394					205					2,168
145					917					713					2,137
146					399					198					2,313
147					623					428					1,894
148					755					564					2,238
149					921					181					2,168
150					394					198					537
151					598					355					1,797
152					665					480					1,477
153					867					645					1,513
154					468					239					2,168
155					811					592					2,213
156					529					315					634
157					697					519					860
158					397					147					1,634

	Seconds for VM to reboot - catastrophic outage					Seconds for VM to reboot - planned shutdown					Seconds for VM to migrate - planned maintenance				
VM	18 VMs	30 VMs	50 VMs	100 VMs	160 VMs	18 VMs	30 VMs	50 VMs	100 VMs	160 VMs	18 VMs	30 VMs	50 VMs	100 VMs	160 VMs
159					671					470					1,934
160					904					708					1,144

Figure 15: Individual VM timing results for each scenario.

APPENDIX B – SERVER AND STORAGE CONFIGURATION INFORMATION

Figure 16 provides detailed configuration information about the test servers, while Figure 17 provides detailed configuration information about the IBM XIV storage system.

System	IBM System x3690 X5	IBM System x3690 X5
CPU		
Vendor and model number	Intel Xeon X7560	Intel Xeon X7560
Number of processor packages	2	2
Core frequency (GHz)	2.26	2.26
Number of cores per processor	8	8
Hyper-threading	Yes	Yes
Platform		
Vendor and model number	IBM System x3690 X5 7148-AC1	IBM System x3690 X5 7148-AC1
BIOS name and version	IBM MLE131AUS 1.31 (02/10/2011)	IBM MLE131AUS 1.31 (02/10/2011)
BIOS settings	Default	Default
Memory module(s)		
Total RAM in system (GB)	512	512
Vendor and model number	Hynix HMT42GR7BMR4C-G7	Samsung M393B2K70CM0-CF8
Type	PC3-8500	PC3-8500
Speed (MHz)	1,066	1,066
Size (GB)	16	16
Number of RAM module(s)	32	32
Rank	Quad	Quad
MAX5 memory modules		
Total RAM in system (GB)	512	512
Vendor and model number	Samsung M393B2K70CM0-CF8	Samsung M393B2K70CM0-CF8
Type	PC3-8500	PC3-8500
Speed (MHz)	1,066	1,066
Size (GB)	16	16
Number of RAM module(s)	32	32
Rank	Quad	Quad
Hard disk		
Vendor and model number	Hitachi Ultrastar HUC103014CSS600	Hitachi Ultrastar HUC103014CSS600
Number of disks in system	4	4
Size (GB)	146	146
Buffer size (MB)	64	64
RPM	10,000	10,000
Type	SAS 6Gb/s	SAS 6Gb/s
Disk controller		
Vendor and model number	IBM ServerRAID M5015	IBM ServerRAID M5015
Controller firmware	2.120.03-1108	2.120.03-1108
RAID configuration	10	10

System	IBM System x3690 X5	IBM System x3690 X5
Fibre Channel HBA		
Vendor and model number	QLogic® QLE2562-CK	QLogic QLE2562-CK
Type	8Gb Dual-port Fibre channel	8Gb Dual-port Fibre channel
Firmware	4.03.001	4.03.001
Additional NIC		
Vendor and model number	Brocade 1020	Brocade 1020
Type	10 Gb/s dual-port CNA	10 Gb/s dual-port CNA
Operating system		
Name	VMware vSphere 4.1 Update 1	VMware vSphere 4.1 Update 1
Build number	348481	348481
File system	vmfs	vmfs
Ethernet		
Vendor and model number	Broadcom® BCM5709C NetXtreme® II GigE	Broadcom BCM5709C NetXtreme II GigE
Type	Integrated	Integrated

Figure 16: Configuration information for the test servers.

System	IBM XIV Storage System
General	
Vendor and model number	IBM XIV A14/2812
System Version	10.2.4.a
Number of modules	15
Number of disks per module	12
Hard disk	
Vendor and model number	Hitachi HUA721010KLA330
Number of disks in system	180
Size (GB)	1,024
Buffer size (MB)	32
RPM	7,200
Type	SATA 3.0Gb/s

Figure 17: Configuration information for the IBM XIV Storage System.

APPENDIX C – DETAILED SETUP

This section details the setup of hardware and software we used in this study.

Physical server setup and MAX5 cabling

We installed an IBM System x3650 M3 and two IBM System x3690 X5 servers each with an IBM MAX5 memory expansion module in a server rack. We cabled each x3690 X5 to a MAX5 unit using an IBM QPI cable kit.

In each IBM System x3690 X5, we installed a Brocade 1020 10Gb Converged Network Adapter (CNA). For networking, then, each IBM System x3690 X5 had two onboard NICs and two ports available on their respective CNA. We cabled the two Brocade CNAs together with an active SFP 10Gb copper cable in a point-to-point configuration, which we used for vMotion. In a real-world configuration, consult your networking administrator and perform this configuration through an approved 10Gb redundant switch. We cabled the two onboard NICs on each ESX server into distribution switches for management and VM Network traffic. For full details on VMware best practices, refer to Appendix D.

IBM XIV configuration and cabling

As Figure 18 shows, we cabled each IBM System x3690 X5 to the IBM XIV via two Fibre channel switches. From each switch, we cabled one cable to each input module (modules 4, 5, 6, 7, 8, and 9) on the IBM XIV, for a total of 6 paths from each fabric. We cabled Switch A to port 1 on each relevant XIV module, and Switch B to port 3 on each relevant module.

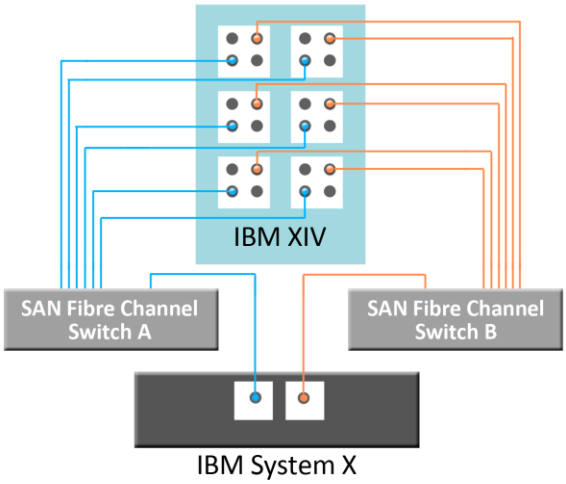


Figure 18: IBM XIV cabling to each system.

Using the XIVGUI application, we created a new cluster in the Hosts and Clusters view. We then added each ESX host to the new cluster and presented both Fibre ports from each host to the XIV. Still in the XIVGUI application, under Storage Pools, we created four 17TB storage pools. Within each storage pool, we then created four 2TB volumes, for a

total of 16 volumes. To present the volumes to the hosts in the cluster, we mapped each individual volume to the cluster created in the XIVGUI application.

VMware vSphere (ESX) installation

We installed VMware ESX 4.1 Update 1 on the two IBM System x3690 X5 servers using default settings. To utilize the MAX5 memory expansion, we set the boot parameter allowInterleavedNUMAnodes to TRUE as specified in VMware and IBM documentation[1]. We set the appropriate network configurations and IP addresses on each ESX server.

Installing Windows Server and vCenter on the management server

We created a three-disk RAID 5 virtual disk on the IBM System x3650 M3 and installed Windows Server 2008 R2 SP1 Enterprise Edition using the default settings. We then installed VMware vCenter Server 4.1.0 Update 1 using the default settings. We also installed VMware vSphere Client 4.0.1 Update 1 using the default settings. Before creating the cluster, we assigned the proper network settings and operating system settings on the server.

Creating and configuring the cluster

We logged into our vCenter server using the administrator credentials and the vSphere Client. Within the vSphere client, we created a new datacenter and cluster using the default settings. We then added our two ESX hosts to the new cluster. With both hosts added to the cluster, we edited the cluster settings and enabled VMware HA and VMware DRS using the default settings. For more information on setting up a VMware HA cluster, see http://www.vmware.com/pdf/vsphere4/r41/vsp_41_availability.pdf. For more information on HA best practices, see http://www.vmware.com/files/pdf/techpaper/VMW-Server-WP-BestPractices.pdf.

After adding our hosts to the cluster, we connected to the IBM XIV storage from the vSphere client and created 16 datastores using the volumes from the XIV.

We configured our vMotion network on each host and connected it to our 10Gb Brocade CNA. We configured our VM networks on each host and connected it to our 1Gb onboard NICs. Additionally, we enabled jumbo frames on the vSwitch designated for vMotion traffic by following the information provided by VMware at this link: http://kb.vmware.com/selfservice/microsites/search.do?language=en_US&cmd=displayKC&externalId=1003712.

Finally, we installed VAAI storage drivers on each ESX host using the information provided by IBM at this link: http://delivery04.dhe.ibm.com/sar/CMA/SDA/02ift/1/IBM_Storage_DD_v1.1.0_for_VMware_VAAI_IG.pdf. We also updated the queue depth of each fibre channel HBA to be 256, as recommended by IBM best practices.

[1] http://kb.vmware.com/selfservice/microsites/search.do?language=en_US&cmd=displayKC&externalId=1021454
http://www-947.ibm.com/support/entry/portal/docdisplay?brand=5000020&lndocid=MIGR-5084842

Creating the VMs

We created one VM to use as our gold image. We allocated the VM 1 vCPU, 3GB RAM, and created a 20GB virtual disk. For the VM operating system, we installed Windows Server 2008 R2 SP1 Enterprise Edition. We then installed IBM HTTP Server 7.0 using the default settings and copied our static Web content into the root directory of the Web server. For more information on IBM HTTP Server, see http://www-01.ibm.com/software/webservers/httpservers/.

Because these VMs would be undergoing multiple hard shutdown processes as part of our HA testing, we adjusted the amount of time that Windows warns the user that an improper shutdown occurred from the default of 30 seconds to 1 second.

Once we finalized the VM configuration, we converted the VM to a template and deployed 320 VMs from the template, assigning 20 VMs per datastore. vCenter customized each VM's OS using a saved customization specification, which renamed the hostname of the VM and set the IP address. We then ensured that each VM had network connectivity and split them between the two hosts as necessary for each test scenario.

Setting resource reservations

To account for VMware's slot calculation in a cluster, we assigned memory and CPU reservations to each VM. We set the memory reservation to 2,560 MB and the CPU reservation to 100 MHz. For more information on slot calculations, see http://kb.vmware.com/selfservice/microsites/search.do?cmd=displayKC&docType=kc&externalId=1010594&sliceId=1&docTypeID=DT_KB_1_1&dialogID=181201672&stateId=0%200%20181203617.

APPENDIX D – VMWARE VCENTER SERVER AND HA BEST PRACTICES

VMware vCenter Server

In this section, we present guidelines and recommendations for achieving the best performance with vCenter Server. This is not intended to replace the installation and management guides from VMware, but only to supplement them and offer useful tips for getting the best performance in a scale-up vSphere environment.

General recommendations

- Use an enterprise database located on a separate server or VM. Use SQL Express for test or lab environments only.
- Avoid running other applications with vCenter Server to avoid resource contention.
- vCenter Server can be run on a dedicated host or in a virtual machine. In either case, be sure to provide enough resources as outlined below based on the size of the environment.
- Minimum requirements are a host with 2 64-bit CPUs or 1 64-bit dual-core CPU, 3GB RAM, and 1Gb Ethernet NIC.
- vCenter Server is now a 64-bit application, requiring a 64-bit version of Windows Server operating system and a 64-bit system DSN connection to a database.

Sizing recommendations

Figure 19 provides three vCenter Server deployment sizes: *Medium*, *Large*, and *Extra-large*, based on the number of hosts and powered-on VMs in the environment:

Deployment size	Hosts	Powered-on VMs
Medium	50	500
Large	300	3,000
Extra-large	1,000	10,000

Figure 19: Sizing guides for vCenter Server.

Based on the three deployment sizes, Figure 20 offers minimum system recommendations for good performance for the Windows operating system running vCenter Server, and the client system running the vSphere Client.

Deployment size	Component	Cores	Memory	Disk
Medium	vCenter Server	2	4 GB	5 GB
	vSphere Client	1	200 MB	1.5 GB
Large	vCenter Server	4	8 GB	10 GB
	vSphere Client	1	500 MB	1.5 GB
Extra-Large	vCenter Server	8	16 GB	10 GB
	vSphere Client	1	500 MB	1.5 GB

Figure 20: Hardware recommendations for vCenter Server and vSphere client.

Database best practices

vCenter Server uses a database to store state information and historical performance statistics about the VMware vSphere environment. Performance statistics are among the largest and the most resource-intensive components of the vCenter database and can take up to 90 percent of the overall database size, and hence are a primary factor in the performance and scalability of the vCenter Server database.

vCenter Server comes bundled with a Microsoft SQL Server Express Edition database. With the SQL Server Express Edition database, vCenter can support up to five VMware ESX hosts. However, we do not recommend using SQL Server Express for long-term production use due to lack maintenance tools and scalability.

Supported enterprise-class database servers for use with vCenter are IBM DB2 9.5, Oracle 10*g* and 11*g*, and Microsoft SQL Server 2005 and 2008.

If the vCenter Server database is running on a different machine than the vCenter Server, the network latency between the machines is something that can adversely affect the performance of your vCenter Server installation. Make sure there is adequate network bandwidth available between the vCenter Server and the database server.

For the vCenter Server database, separate database files for data and for logs onto drives backed by different physical disks, and make sure statistics collection times are set conservatively so that they will not overload the system.

HA performance tuning

HA uses network heartbeats to determine liveliness of the hosts. By default, heartbeats are sent every 1 second by the hosts, and if no heartbeat is received in 15 seconds from a host the host will be declared as. The heartbeat interval and failure detection time can be changed through the vSphere Client via the HA Advanced Settings. Using the default values is recommended in most cases, but if there are a large number of hosts and heavy network traffic has been observed, consider increasing the heartbeat interval to reduce the amount of heartbeat traffic. In doing so, you should also increase the failure detection time to avoid false failover.

The heartbeat interval cannot be larger than the failure detection time, and we do not recommend setting it to a very large value, in which case missing a few heartbeats would possibly lead to undesirable failover. Also, for a high-latency network or network with intermittent problems, you might want to increase the failure detection time to avoid false failover.

Advanced option name	Description	Default value
das.failuredetectiontimeinterval	Heartbeat interval	1 second
das.failuredetectiontime	Waiting time before declaring host as failed	15 seconds

Figure 21: Advanced tuning options for VMware HA.

Note: You will need to disable and re-enable HA in the cluster for the changes to take effect.

HA VM startup concurrency

When multiple VMs are restarted on one host, up to 32 VMs will be powered on concurrently by default. This is to avoid resource contention on the host. This limit can be changed through the HA advanced option: das.perHostConcurrentFailoversLimit. Setting a larger value will allow more VMs to be restarted concurrently in situations where VM density per host is higher. This might reduce the overall VM recovery time, but the average latency to recover individual VMs might increase as well.

Tuning Power On Frequency

HA polls the system periodically to update the cluster state with such information as how many VMs are powered on, and so on. The polling interval is 1 second in vSphere 4.0, and the default setting is 10 seconds in vSphere 4.1. You can tune the advanced option das.sensorPollingFreq, and set it to a value between 1 and 30 seconds.

If you see transient "Communication Timeout" errors in the cluster, this means too many state updates are happening in the cluster and HA is having a hard time catching up. You can tune the advanced option to a larger value to eliminate this error. A smaller value leads to faster VM power on, and a larger value leads to better scalability if a lot of concurrent power operations need to be performed in a large cluster.

Powering on additional VMs

When HA is enabled, spare resources are reserved in order to fulfill the failover requirements. If you cannot power on a VM in an HA cluster, then it is likely because powering on this VM will violate the HA admission control policy and result in insufficient failover capacity in the cluster. HA admission control is done based on the policy you select, the cluster capacity, and the VMs' resource settings. There are advanced options that you can set to adjust the number of VMs that can be powered on in an HA cluster. Please see VMware's vSphere Availability Guide for more details.

HA best practices

- **Network redundancy.** Network isolation is perhaps one of the most common HA-related incidents that occur in a datacenter environment. We highly recommend using a series of network redundancy backups – multiple physical NICs, VMware vSphere Service Console redundancy, and redundant physical switches and cabling.

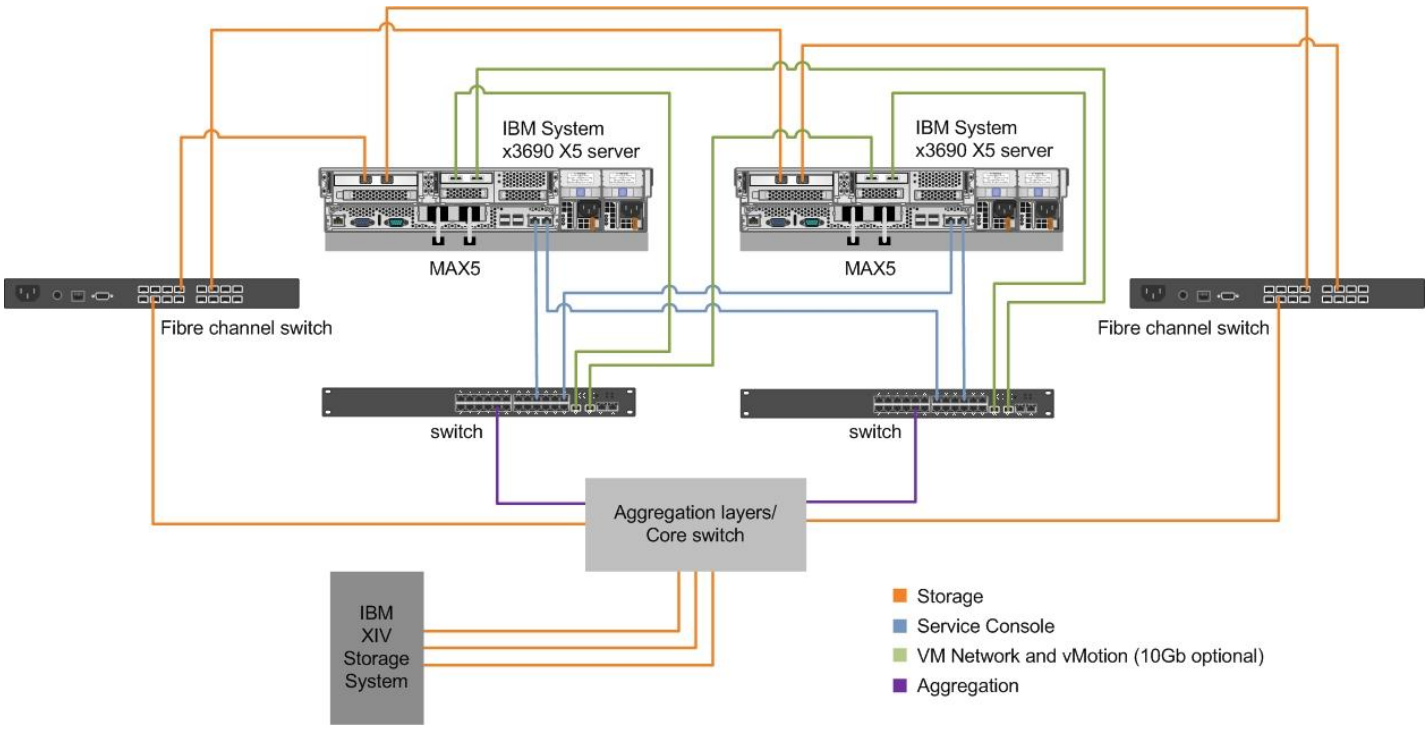


Figure 22: Example diagram of network cabling.

The figure above shows a sample redundant cabling configuration. Keep in mind that each environment is different and you may wish to prioritize VM network traffic over vMotion, or vice versa, so allocate your 10Gb ports as necessary. Also, in our setup, we cabled our Fibre connections directly to the IBM XIV storage, but as the image above shows, you may have core switching that you must connect to. Work with you network administrator to ensure your organization's best practices and policies are followed.

- **Power redundancy.** To minimize the chances of unwanted power loss, split server power supplies among multiple circuits. In the event that one circuit fails, the redundant power supplies running on the separate circuits will keep servers running while the issue is being resolved.
- **Cluster monitoring with alarms.** VMware vCenter Server monitors clusters and hosts using a series of alarm triggers. We strongly recommend fully customizing these alarms to fit your monitoring needs and configuring email notifications.
- **Storage multipathing.** To ensure a consistent and reliable connection to storage, equip each host with multiple connections to storage. Multiple paths to the storage will keep virtual machine storage consistent with compute resources on the host. ESX and ESXi both support multipathing.
- **Host isolation response.** VMware HA setup offers three options for host isolation response: Power Off, Shutdown, Leave Powered On. To protect the virtual machine's operating system and workload from errors, set the host isolation response to Shutdown.
- **Identical server hardware.** VMware suggests using identical server hardware for all nodes in a cluster. With identical hardware, VMware HA is able to better manage resources in the event of a failover.

- **Host monitoring.** Performing network maintenance can disrupt heartbeat exchanges, which will cause isolation responses and initiate unnecessary failovers. Therefore, we recommend disabling host monitoring within the cluster during your maintenance windows.
- **Using DNS.** We recommend using DNS to resolve host names as opposed to editing hosts files on the individual hosts.
- **Port group names.** Use a documented and consistent naming scheme for port groups across hosts. In order for successful failover to occur during an HA event, matching port group names must be present on other hosts in the cluster.

APPENDIX E – IBM XIV BEST PRACTICES FOR VMWARE

In this appendix, we summarize key considerations when configuring a VMware vSphere infrastructure with IBM XIV storage. For detailed information on IBM XIV with VMware storage best practices, see the attachment redbook located at http://www.redbooks.ibm.com/abstracts/sg247904.html.

Key considerations

vStorage APIs for Array Integration (VAAI)

When using the IBM XIV storage system, install the VAAI driver for VMware ESX server 4.1. This will have a strong impact on the performance of common I/O operations such as migrations, clone from template, and other conditions that request a SCSI reservation to the storage system. To use the VAAI driver with XIV, ensure the XIV microcode is at least 10.2.4a. See www.ibm.com/support/fixcentral/ for more details, and select Disk Systems, XIV Storage System, and select VMware.

When using VAAI, you can observe the impact of the driver by toggling from 1 to 0 under Configuration/advanced settings in vCenter:

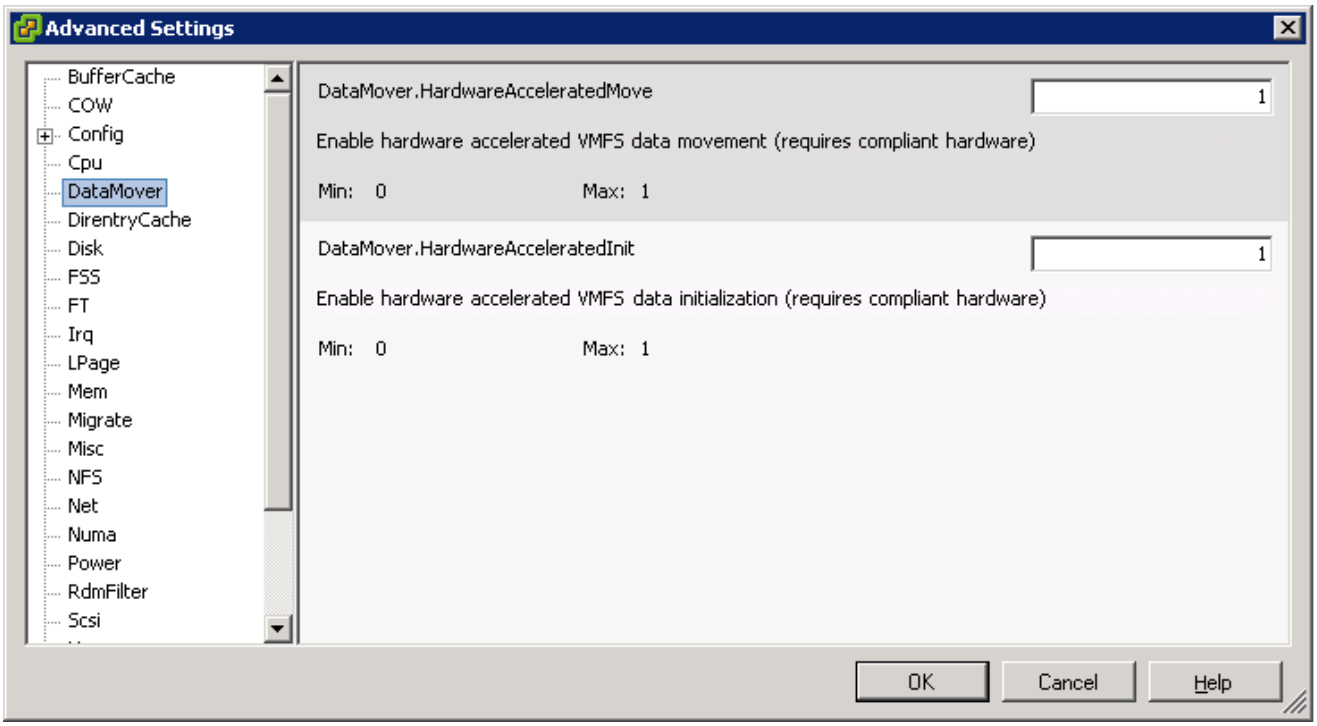


Figure 23: The DataMover.HardwareAcceleratedMove advanced setting in vSphere 4.1 for VAAI.

Queue depth

In general, the queue depth setting on your HBA should be extended as consolidation ratios grow and/or I/O requirements are high. Refer to the above-referenced redbook for examples. Depending on the I/O profile and the consolidation ratios, values will vary from 64, 128, or 256.

Multipathing

You should implement round robin policy for each LUN from the IBM XIV Storage System to achieve the best performance results. The only other supported option is fixed path policy.

Zoning and cabling

Because the IBM XIV Storage System is an active grid, you must ensure that the workload from VMware is balanced across all six I/O controllers. In Figure 18, we show a typical cabling example where a zone is created for each initiator. Depending on SAN vendor recommendations, all six paths would then be added to that zone or separate zones should be created for each initiator/target path. Consult you SAN fabric vendor for best practices. The end result will provide a path from each host initiator to all six targets cabled in the above diagram.

LUN size and datastore

Use large LUNs – 1.5TB to 2TB, the maximum for vSphere servers. This will lower the amount of LUNs to manage in the environment. From a performance perspective on the IBM XIV Storage System, there is no difference between a 500GB LUN and 2TB LUN in a vSphere environment.

Use the default block size when formatting a datastore unless you need a larger than 256GB virtual disk (vmdk). There is no performance impact to changing the default. However, each file on VMFS will take at least the default block size, which could affect storage efficiency noticeably. Other options include:

- 2MB block size – 512GB maximum file size
- 4MB block size – 1024GB maximum file size
- 8MB block size – 2048GB maximum file size (this would be unlikely due to a 2TB limit for LUNs in vSphere)

Lastly, LVM extents are supported, but not recommended. In vSphere with XIV, you can increase the size of a datastore and LUN (up to 2TB) online.

Although these cover important considerations, please consult the following IBM redbook before building out your VMware infrastructure on XIV: http://www.redbooks.ibm.com/abstracts/sg247904.html.

ABOUT PRINCIPLED TECHNOLOGIES

Principled Technologies, Inc.
1007 Slater Road, Suite 300
Durham, NC, 27703
www.principledtechnologies.com

We provide industry-leading technology assessment and fact-based marketing services. We bring to every assignment extensive experience with and expertise in all aspects of technology testing and analysis, from researching new technologies, to developing new methodologies, to testing with existing and new tools.

When the assessment is complete, we know how to present the results to a broad range of target audiences. We provide our clients with the materials they need, from market-focused data to use in their own collateral to custom sales aids, such as test reports, performance assessments, and white papers. Every document reflects the results of our trusted independent analysis.

We provide customized services that focus on our clients' individual requirements. Whether the technology involves hardware, software, Web sites, or services, we offer the experience, expertise, and tools to help our clients assess how it will fare against its competition, its performance, its market readiness, and its quality and reliability.

Our founders, Mark L. Van Name and Bill Catchings, have worked together in technology assessment for over 20 years. As journalists, they published over a thousand articles on a wide array of technology subjects. They created and led the Ziff-Davis Benchmark Operation, which developed such industry-standard benchmarks as Ziff Davis Media's Winstone and WebBench. They founded and led eTesting Labs, and after the acquisition of that company by Lionbridge Technologies were the head and CTO of VeriTest.