Red Hat Reference Architecture Series

Performance and Scalability of the Red Hat Enterprise Healthcare Platform

(based on InterSystems Caché on Quad-Core AMD Opteron™ Processors)

InterSystems Caché 2008.1
Red Hat Enterprise Linux 5
Quad-Core AMD Opteron Processors

Version 1

June 2008

Performance & Scalability of the Red Hat® Enterprise Healthcare Platform (based on InterSystems Caché on Quad-Core AMD Opteron™ Processors)

1801 Varsity Drive
Raleigh NC 27606-2072 USA
Phone: +1 919 754 3700
Phone: 888 733 4281
Fax: +1 919 754 3701
PO Box 13588
Research Triangle Park NC 27709 USA

The information contained herein is subject to change without notice. Red Hat, Inc. shall not be liable for technical or editorial errors or omissions contained herein.

The GPG fingerprint of the security@redhat.com key is:
CA 20 86 86 2B D6 9D FC 65 F6 EC C4 21 91 80 CD DB 42 A6 0E

Table of Contents

1. Introduction & Executive Summary 4
2. Case Study: Switching Caché-based Solution to RHEL/x86-64 6
2.1 Background 6
2.2 Opportunity 6
2.3 Solution 6
2.4 Benefits 7
3. Performance Testing Methodology 9
3.1 Benchmark Description 9
3.2 Performance & Scaling of Well-Behaved Systems 11
4. Caché Performance on Quad-Core AMD Opteron Processors (code named Barcelona) 12
4.1 Hardware / Software Configuration 12
4.1.1 Server 12
4.1.2 OS & DB Software 12
4.2 Performance Results 13
5. Tuning & Configuration Guidelines 17
5.1 Memory Requirement 17
5.2 Hugepages 17
5.3 Caché Shared Memory 17
6. Conclusions & Next Steps 18
7. Appendix: Sysctl Configuration File 19

1. Introduction & Executive Summary

InterSystems is a leading provider of innovative database and integration software in healthcare. This paper addresses the specific integration of Red Hat® Enterprise Linux® (RHEL®) with InterSystems CACHÉ® high-performance database running on Quad-Core AMD Opteron™ Processor based servers.

In a Quad-Core Opteron based server, Caché throughput increased by ~60% when going from 4 active cores to 8 active cores. In both cases, the CPU utilization is >80%. The % of CPU time spent in the system versus user-mode is high and needs further investigation.

The results indicate that despite very limited tuning, Caché was still able to achieve exceptional throughput and scaling up to 8 AMD Opteron cores. Caché running on Red Hat Enterprise Linux & Quad-Core AMD Opteron Processor based servers matches or exceeds the performance of proprietary UNIX and other legacy systems. And because of open source alternatives (running on commodity hardware) replacing proprietary technology it provides this performance at a fraction of the cost.

This represents the first phase of the Caché / RHEL reference architecture for healthcare applications and is the result of a close collaborative effort between Red Hat, Inc., InterSystems Corporation and AMD Corporation.

A key factor that influences customers in choosing a total solution is the size, cost, performance and reliability characteristics of the necessary computing platform. Hence it is essential that customers and ISVs have access to accurate and timely data characterizing their solutions for target applications. Red Hat reference architectures provide enterprise proof points of the capabilities and features of the RHEL-based solutions. They enable Red Hat and its partners to leverage proven RHEL-based architectures in deploying customer-specific solutions.

A Red Hat reference architecture includes a detailed specification of the hardware and software 'solution stack'. After a solution stack is defined, it is assembled and installed and put through a series of rigorous performance and scalability tests. The first version of reference architecture documentation typically focuses on installation / configuration / performance / sizing / tuning guides. Future versions have a more full-function coverage of the "solution ecosystem" including features such as high availability, disaster tolerance, security and system management.

The testing and characterization and the accompanying installation, configuration and tuning guides for a solution stack ensures that all components within a reference architecture function as advertised and that they are capable of withstanding the stresses of a production environment. The testing removes much of the guesswork for the customer, because customers not only know the details of the various components that comprise the reference architecture, but they have the assurance that these components will function as a solution.

This document is intended for field personnel, ISVs (Independent Software Vendors) as well as customers seriously evaluating using InterSystems Caché running on RHEL as a reliable, secure, and affordable alternative to high-priced proprietary OS/server configurations.

The objectives of the follow-on work on the Caché / RHEL reference architecture include further optimizing the existing components and using additional Caché and RHEL technologies to scale beyond the currently tested levels:

1. Scaling-up beyond 8-cores to 12-core and 16-core AMD Opteron based servers
2. Scale-out using Enterprise Caché Protocol (ECP) using Red Hat Enterprise Linux in both the ECP application tier and the ECP database tier
3. Deliver automated failover and migration of Caché virtual machines on Red Hat Enterprise Linux to achieve high availability at an industry leading price-point.

2. Case Study: Switching Caché-based Solution to RHEL/x86-64

A large number of Caché-based healthcare applications still run on proprietary UNIX and legacy platforms like OpenVMS.

This section presents a case study of moving a Caché-based solution from a proprietary platform to RHEL.

2.1 Background

A Harvard Medical School teaching hospital, Beth Israel Deaconess Medical Center (BIDMC) is renowned for excellence in patient care, biomedical research, teaching, and community service. Among independent teaching hospitals, BIDMC is the fourth-largest recipient of biomedical research funding from the National Institutes of Health. With 3,000 doctors and 12,000 employees on staff, the hospital serves nearly one million patients each year and is the official treatment center of the Boston Red Sox. The Information Systems Division at BIDMC maintains a datacenter with 146 mission-critical applications, vital to the functioning of the hospital.

2.2 Opportunity

In 2005, Dr. John Halamka, CIO of Harvard Medical School and BIDMC, wanted to migrate the hospital's IT infrastructure to a more secure, reliable operating system that would reduce operating and capital expenditures. "Our Triple A applications, which are responsible for all of the clinical, financial, administrative, and academic activities in the hospital, ran on HP Unix. But the operating system had memory leaks and required frequent virus patches," said Halamka. "We experienced approximately 20 hours of planned and unplanned downtime last year," added Rob Hurst, Sr. Caché Administrator for BIDMC. The hospital not only wanted to move its applications to a more stable and secure operating system, but also wanted to create a new disaster-recovery system that would increase availability from 99.7 percent to 99.99 percent—improving the hospital's level of patient care even further.

2.3 Solution

Three years prior, BIDMC had begun using Red Hat for the hospital's utility services, including mail exchange, spam filtering, and DNS. "Our internal security team was running Red Hat Enterprise Linux exclusively on its servers, so we knew Red Hat provided rock-solid security," said Hurst. However, executive management questioned whether an open source solution could scale sufficiently while providing the level of reliability needed to support enterprise applications. As the former IT director for another Northeastern hospital, Hurst had gained extensive experience deploying Red Hat for core clinical systems. "Based on my previous experience, I was able to provide BIDMC with benchmark data, demonstrating that Red Hat performance, scalability, and reliability was proven in hospital environments," he said.

After gaining management approval, Hurst spearheaded the migration project, purchasing Red Hat Enterprise Linux from DLT Solutions, one of Red Hat's value-added providers dedicated to healthcare and government environments. Hurst's team deployed Red Hat Enterprise Linux on 11 servers that run Intersystems Caché, as well as the hospital's proprietary Triple A applications. "Red Hat Professional Services helped us review the architecture design, ensuring a smooth transition to our production environment," said Hurst. Within six months, the migration from HP-UX to Red Hat was complete, and Hurst is now leveraging Red Hat solutions, including Red Hat Global File System (GFS) and Cluster Suite, to implement a more robust disaster-recovery strategy. BIDMC currently operates four environments—development, testing, production, and shadow production—and runs 11 servers in a cluster. The hardware used is HP DL385 with AMD dual-core processors.

Using Red Hat Global File System and Cluster Suite, Hurst and his team are creating a multi-tiered architecture that separates the network, applications, and database layers within one stack. "Red Hat GFS creates one file system as if all of the layers are running on one server and redirects files seamlessly as needed. If we have an unplanned outage on one application server, then GFS automatically distributes to another application server or environment, eliminating lengthy wait times," said Hurst. To perform a planned update, such as a security patch or memory upgrade, GFS enables the team to redirect to a different environment easily without having to shut down the system.

2.4 Benefits

Red Hat enabled BIDMC to realize enormous cost savings, while increasing system availability for the hospital's core clinical applications. According to Hurst, "Moving from HP-UX servers to 11 Red Hat Enterprise Linux servers is saving us approximately $200,000 per year, just on system support costs alone." Hurst also explains how system availability not only results in cost savings, but is also critical to providing patient care. "If the hospital experiences more than an hour of downtime, it is considered a state of emergency—because system outages can cost human lives," he said. In the past, BIDMC had to take its system down on Sunday nights for scheduled maintenance. "Whether we perform planned maintenance or experience unplanned outages, Red Hat is reducing our system downtime from 20 hours per year to near zero—enabling BIDMC to provide higher levels of patient care. As we've started implementing our new disaster-recovery system in a lab-controlled environment, we've already seen failover time reduced from 15 minutes to 14 seconds."

Red Hat's open source technology, combined with high-level support, provide BIDMC with the reliability and agility it requires to run a leading-edge hospital. "Red Hat solutions, such as GFS and Cluster Suite, are built into the kernel, providing all of the open source technology we need—affordably and without vendor lock-in," said Hurst. Before moving its Red Hat servers into production, Hurst was impressed when his team was able to communicate directly with the Vice President of Support, 24×7. "Red Hat is an engineering-focused company with executive management and a Global Support Services team that is highly involved and technically capable. This means we can resolve issues quickly and keep our most critical hospital information systems available to ensure leading-edge patient care," he said.

BIDMC's roadmap includes moving other hospital systems from HP-UX to Red Hat. "The hospital is currently considering migrating its Oracle and PeopleSoft database applications to Red Hat. "Moving our core clinical applications to Red Hat was the first step. The reliability and performance gains we've experienced are proof that we're ready to migrate our other applications," said Hurst.

3. Performance Testing Methodology

Hardware / Software Stack Tested

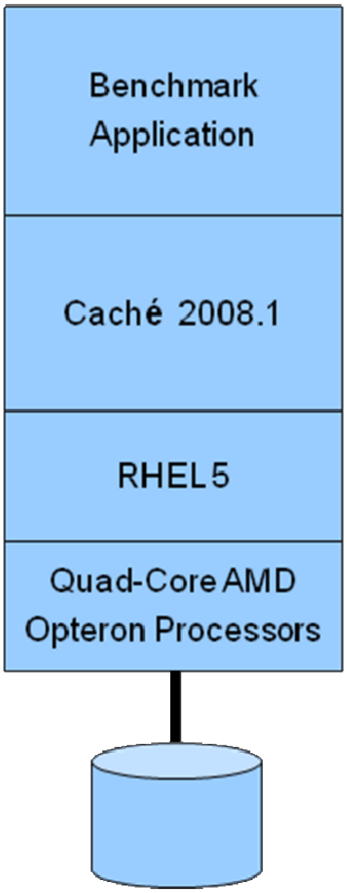

Figure 1

3.1 Benchmark Description

The performance and scalability testing is performed using the InterSystems benchmark application which simulates an interactive transaction processing workload that could represent a hospital information system. Users are emulated using a Remote Terminal Emulator (RTE) which simulates users of the System Under Test (SUT). Each emulated user establishes a Telnet connection to the benchmark application and runs a scripted scenario which includes think time and directives to the benchmark application. This is shown in Figure 2.

When there are very few users, the application response is virtually instantaneous and thus the elapsed time to complete the scenario is determined primarily by the sum of the think times. As the number of users is increased, the application response time increases due to contention for resources and the elapsed time to complete one loop of the script also increases.

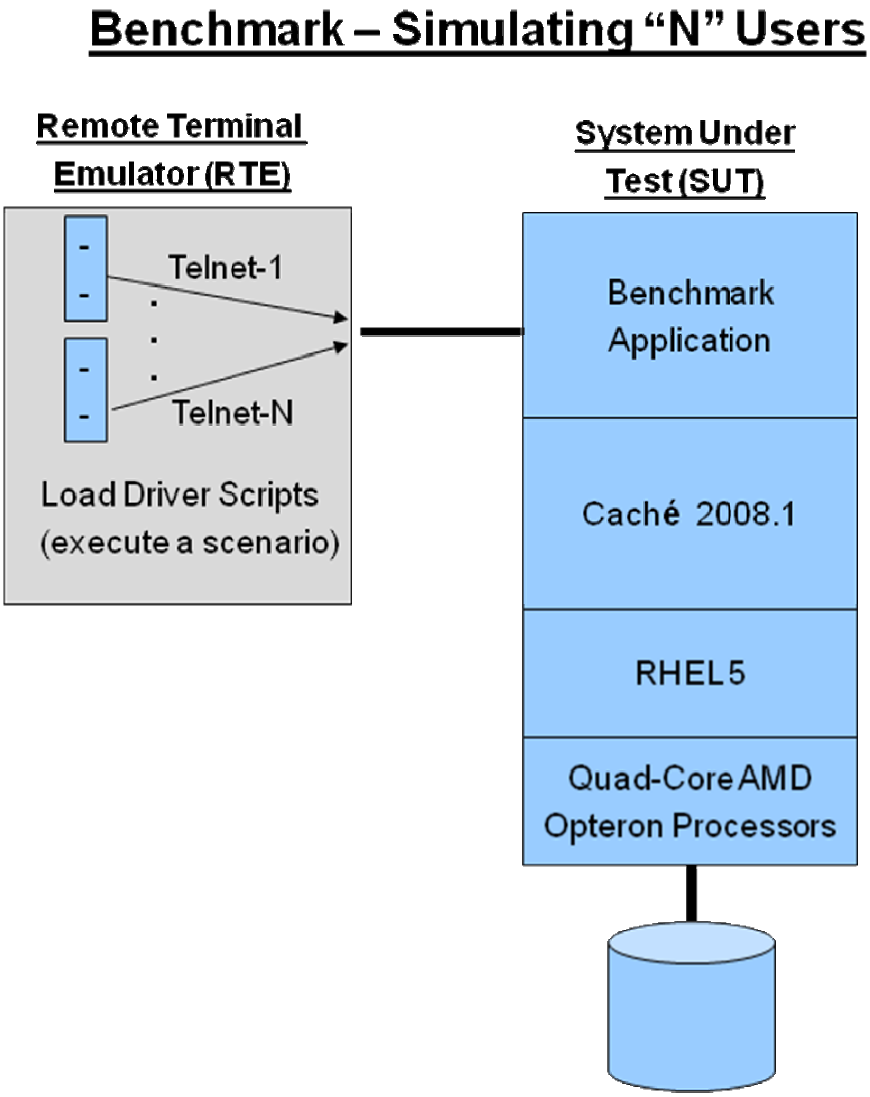
Benchmark – Simulating "N" Users
Remote Terminal Emulator (RTE)
Telnet-1
Telnet-N
Load Driver Scripts (execute a scenario)
System Under Test (SUT)
Benchmark Application
Caché 2008.1
RHEL5
Quad-Core AMD Opteron Processors

Figure 2

The rule used to determine when to stop increasing the number of users any further is when the average elapsed time to complete the scenario reaches a pre-determined threshold (= 250 seconds). At this point the CPU utilization is typically nearing ~100%. The number of users at this point is considered the SUT's rating in terms of number of users. This is shown in Figure 3.

One would never run an actual system under such a heavy load with CPU utilization ~ 80 - 90%. The purpose of this test is to establish well behaved scalability under very heavy loads.

3.2 Performance & Scaling of Well-Behaved Systems

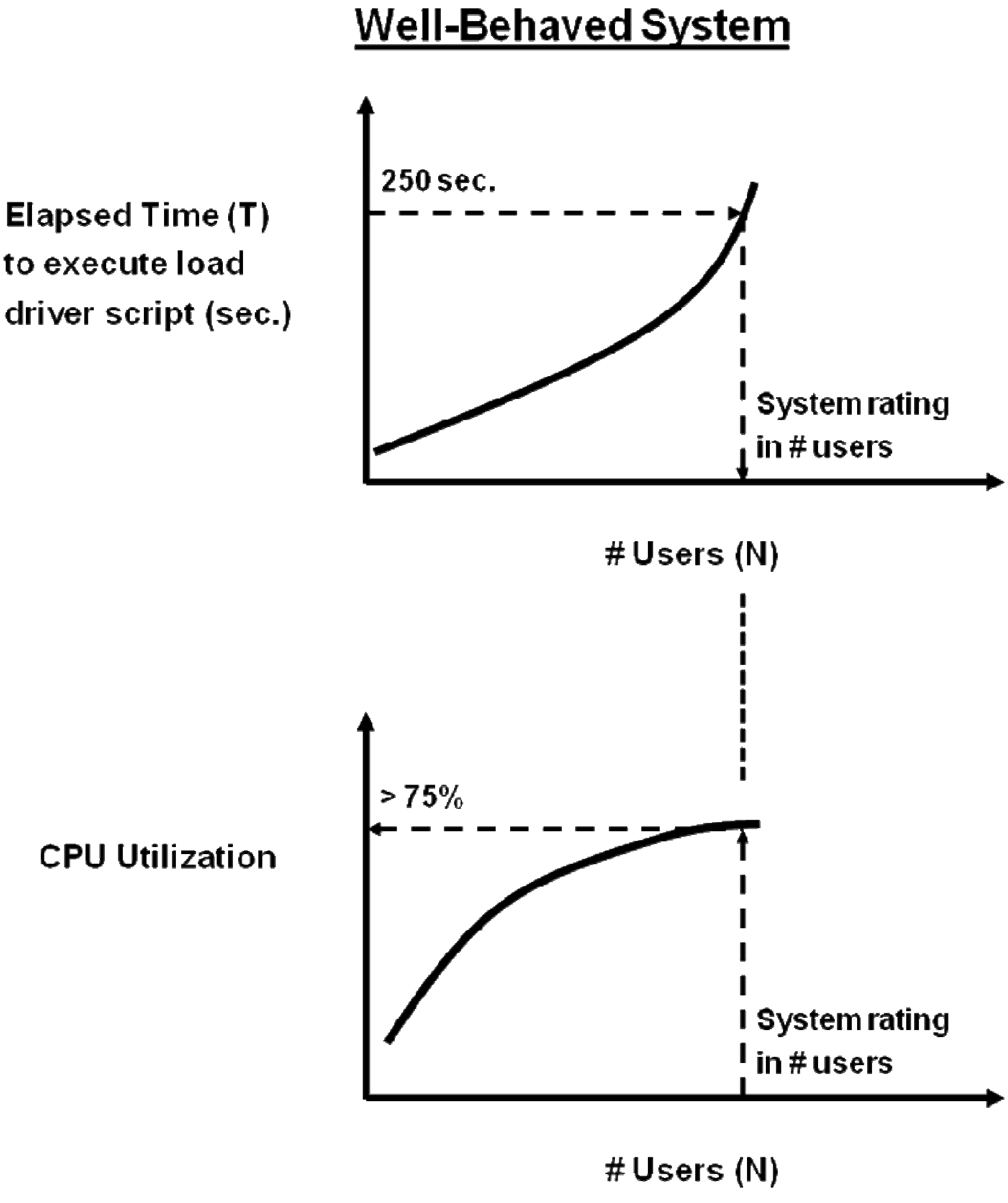

Figure 3

4. Caché Performance on Quad-Core AMD Opteron Processors (code named Barcelona)

4.1 Hardware / Software Configuration

4.1.1 Server

Vendor Model	Quad-Core AMD Opteron™ Processor Model 8356 (code name Barcelona)
Processors	4 sockets, 4 cores/socket = 16 cores
Clock Speed	2.31 GHz
Memory	32 GB
Disks	4 x 160 GB, 7200 RPM SATA Drives (Western Digital WD1600YS)

4.1.2 OS & DB Software

Vendor	Red Hat
Version	Red Hat Enterprise Linux (RHEL) 5.1 - 64 bit
Kernel Version	Kernel v2.6.18-53
File System	EXT3
Other	The RHEL High Availability software solution is not configured

DB Vendor	InterSystems
DB Version	Caché 2008.1 (b386)

4.2 Performance Results

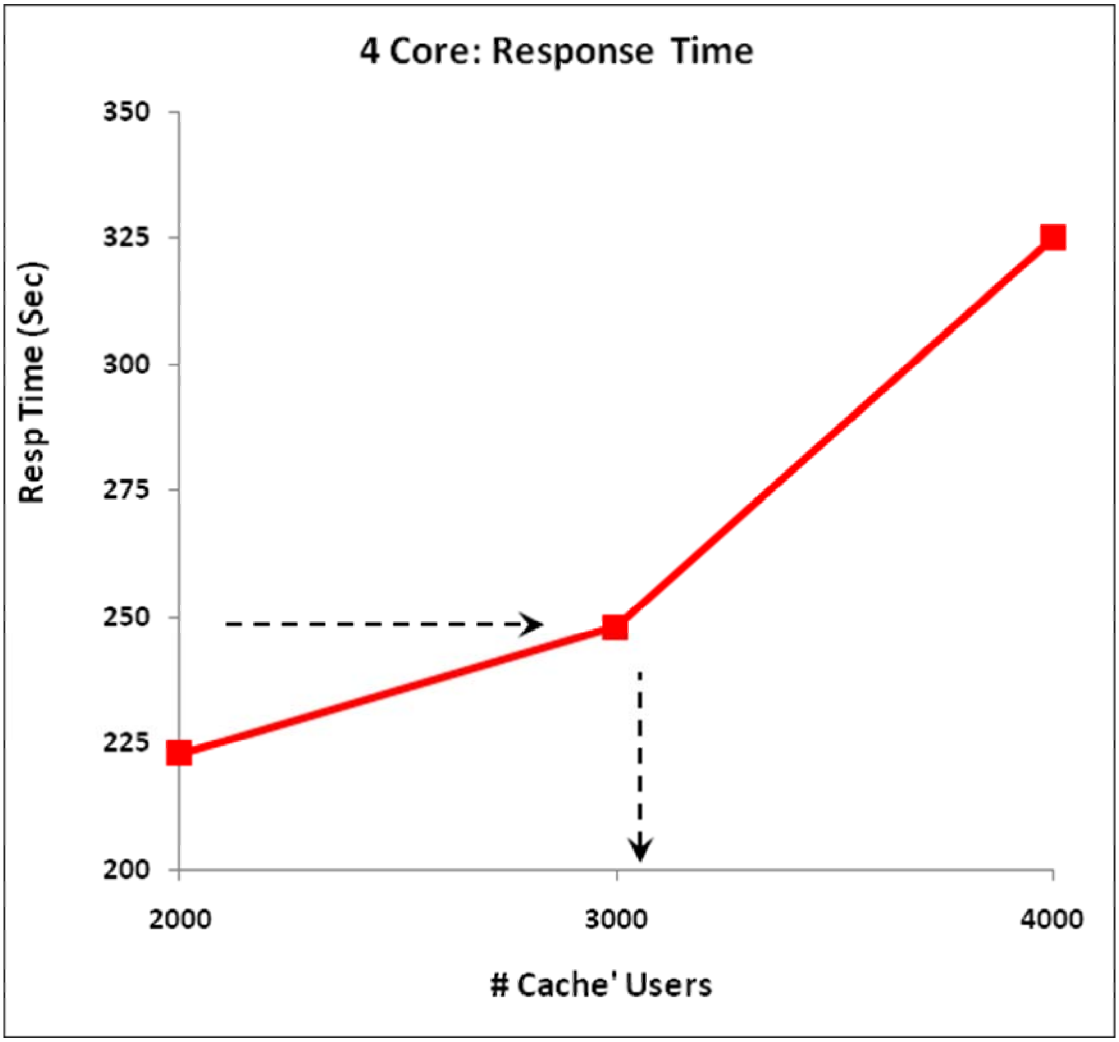

Figure 4

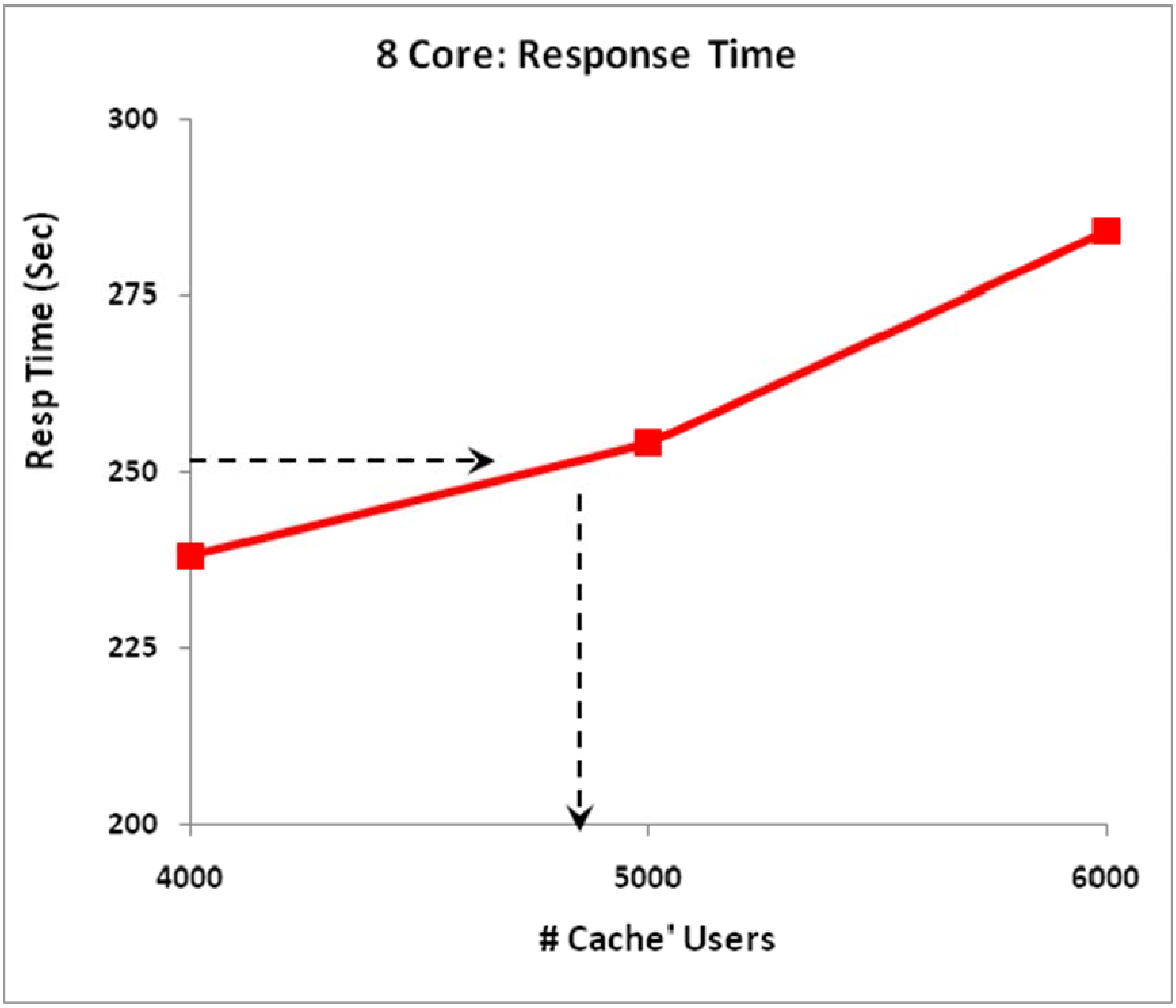
8 Core: Response Time
Resp Time (Sec)
300
275
250
225
200
4000
5000
6000
# Cache' Users

Figure 5

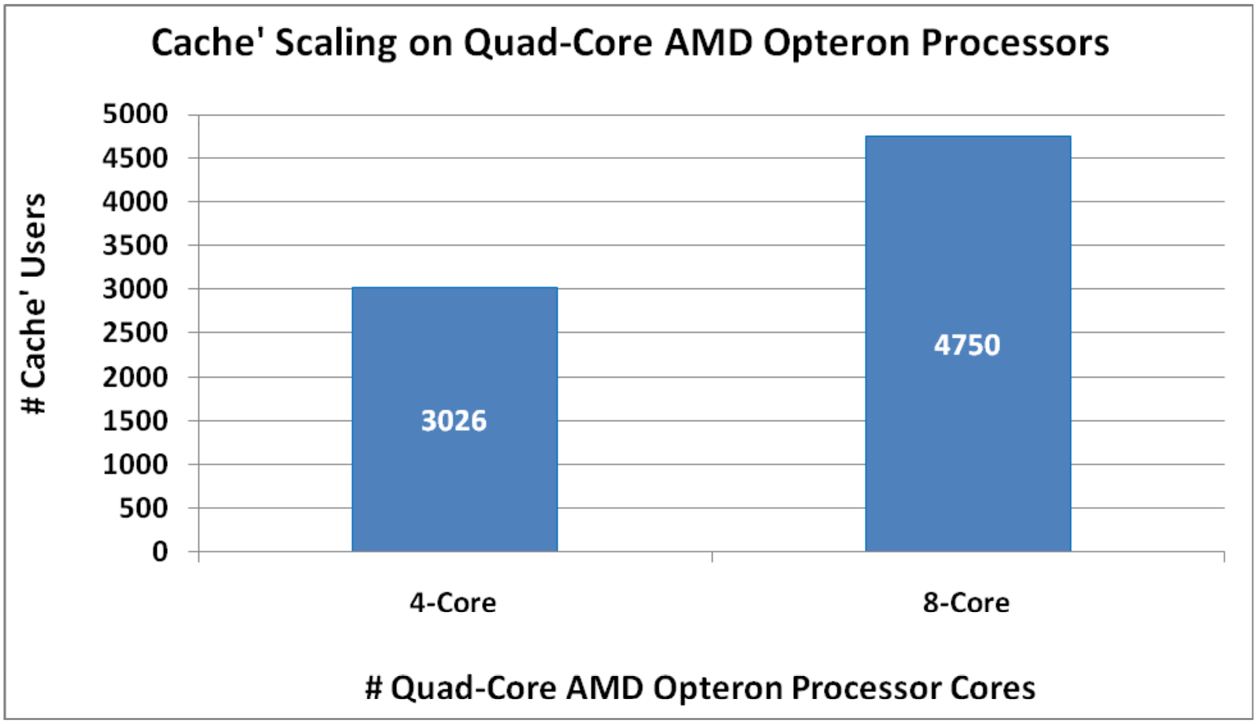

Figure 6

Note: the number of users shown in Figure 6 are for the benchmark used. How this translates into real users supported for a real application will depend on the characteristics of the specific end-user application.

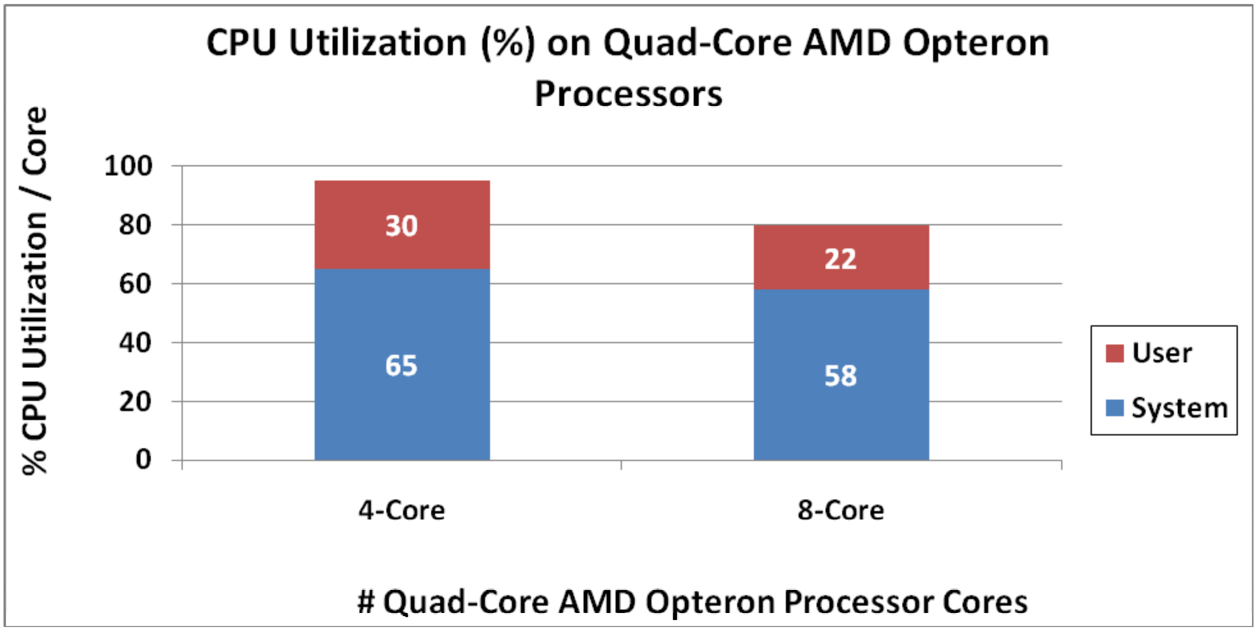
CPU Utilization (%) on Quad-Core AMD Opteron Processors
% CPU Utilization / Core
100
80
60
40
20
0
30
65
22
58
4-Core
8-Core
User
System
# Quad-Core AMD Opteron Processor Cores

Figure 7

5. Tuning & Configuration Guidelines

5.1 Memory Requirement

Systems with up to 8 cpu cores must be configured with 32GB of RAM. Lesser than this leads to memory pressure at higher loads.

5.2 Hugepages

Hugepages benefit systems with 'inadequate RAM'. So, a system with 16GB RAM benefited from hugepages while a system with 32GB RAM showed marginal benefit from hugepages. hugepages are supported in Caché® starting with version 2008.1.

Hugepages have 2 distinct benefits:

1. [2MB] hugepages is more TLB efficient because each TLB entry maps 2MB instead of 4KB or 512 times as much memory. A typical 128 entry TLB can map only 512KB using 4KB pages whereas it can map 256MB using 2MB pages. This obviously reduces the TLB misses or increases the memory footprint before the system incurs TLB misses. This benefit is typically seen as a ~5% performance improvement in user space running a database that used hugepage system V shared memory regions.

2. [2MB] hugepages are not kept on the page lists by the kernel therefore any scalability problems in the VM due to lots of pages on a list protected by one spinlock will not exist for regions of memory that use hugepages. Specifically when the system runs out of RAM (which it always does) the VM system will not process all of the pages in the database cache which are mapped with hugepages simply because those hugepages are not on the page lists. In addition, since the hugepages are not managed memory the kernel does not need to swap any of that memory out when reclaiming memory. This results in a significantly more scalable system when memory exhaustion takes place. If the system does not run out of memory only the TLB benefits to hugepages are realized.

5.3 Caché Shared Memory

shmmax is set to allow sufficient shared memory for Caché®. 4GB was enough for the benchmark application used here, but this value is highly dependent on the specific end-user application. This parameter is set in /proc/sys/kernel/shmmax and/or as kernel.shmmax in /etc/sysctl.conf for a persistent setting.

6. Conclusions & Next Steps

The objective of these benchmark tests was to evaluate the Caché database configured on Red Hat Enterprise Linux (RHEL) running on commodity x86-64 AMD Opteron processors; and to validate this technology stack as a viable, competitive solution.

In a Quad-Core AMD Opteron™ Processor based server, Caché throughput increased by ~60% when going from 4 active cores to 8 active cores. In both cases, the CPU utilization is >80%. The % of CPU time spent in the system versus user-mode is high and needs further investigation.

In objectives of the follow-on work on the Caché / RHEL reference architecture include further optimizing the existing components and using additional Caché and RHEL technologies to scale beyond the currently tested levels:

1. Performance scale-up beyond 8-cores to 12-core and 16-core Quad-Core AMD Opteron™ Processor based servers
2. Performance scale-out (and high availability) using Enterprise Caché Protocol (ECP) on Red Hat Enterprise Linux in both the ECP application tier and the ECP database tier
3. Deliver automated failover and migration of Caché virtual machines on Red Hat Enterprise Linux to achieve high availability at an industry leading price-point.

The results indicate that despite very limited tuning, Caché was still able to achieve exceptional throughput and scaling up to 8 AMD Opteron Processor cores. Caché running on Red Hat Enterprise Linux & AMD Opteron based servers matches or exceeds the performance of proprietary UNIX and other legacy systems. And because of open source alternatives (running on commodity hardware) replacing proprietary technology it provides this performance at a fraction of the cost.

In addition to price/performance leadership, Caché on Red Hat Enterprise Linux provides leadership reliability, availability, interoperability, security and manageability. For more information about Red Hat solutions for Healthcare, visit http:www.redhat.com/healthcare.

7. Appendix: Sysctl Configuration File

```
# Kernel sysctl configuration file for Red Hat Linux
#
# For binary values, 0 is disabled, 1 is enabled.  See sysctl(8) and
# sysctl.conf(5) for more details.

# Controls IP packet forwarding
net.ipv4.ip_forward = 0

# Controls source route verification
net.ipv4.conf.default.rp_filter = 1

# Do not accept source routing
net.ipv4.conf.default.accept_source_route = 0

# Controls the System Request debugging functionality of the kernel
kernel.sysrq = 0

# Controls whether core dumps will append the PID to the core filename
# Useful for debugging multi-threaded applications
kernel.core_uses_pid = 1

# Controls the use of TCP syncookies
net.ipv4.tcp_syncookies = 1

# Controls the maximum size of a message, in bytes
kernel.msgmnb = 65536

# Controls the default maxmimum size of a mesage queue
kernel.msgmax = 65536

# Controls the maximum shared segment size, in bytes
kernel.shmmax = 68719476736

# Controls the maximum number of shared memory segments, in pages
kernel.shmall = 4294967296

# Increase the number of pty.s
kernel.pty.max = 20000

# allocate 8GB huge pages
vm.nr_hugepages = 4112
```