Maximum memcached performance with the Dell PowerEdge R730 and Solarflare technology

John Beckett
Office of the CTO

July 2016 | Version 1.0

Contents

Introduction

Reducing the pressure on your database back end in a high volume website environment has never been more impactful thanks to the combination of Dell® PowerEdge™ servers and Solarflare® Flareon® Ultra server Network Interface Cards (NICs) and their OpenOnload® stack. Thanks to this superior combination of hardware and software, Dell is pleased to announce the following:

- **Dell has demonstrated the capability to service up to 160Gb per second of sustained memcached traffic**
- **Dell has measured up to 40 million memcached transactions per second.**

Compared to other common network adapters, the Solarflare Flareon NICs and their OpenOnload network stack boast enormous increases to both throughput and transactions per second depending on the key/value size. Several configurations were tested during the course of this study, and excellent scaling characteristics were noted as the adapter count was increased.

With proven technology from Dell and Solarflare, the web community has a new model of memcached implementation that will afford the ability for dramatic consolidation, optimal performance characteristics, power savings, and lowered TCO inherent with consolidation activities.

What is memcached?

Introduced by Danga Interactive in 2007, memcached has been a popular open source, distributed memory object caching system primarily used to reduce the load on database back-ends in web application server environments. Memcached is used extensively by companies such as YouTube®, Reddit®, Wikipedia, Facebook® and other high-volume websites. Although nine years old at the time that this white paper was written, memcached remains one of the leading in-memory key/value stores today.

Details of memcached implementation include the following:

- Key size up to 250 bytes / Value size up to 1MB
- Capable of utilizing a very deep but ephemeral cache
- No redundancy, no failover and no authentication
- No coordination with other memcached instances
- Example code in practically any language to create interface with application servers
- Cached data cannot be saved to local storage
- Single listener, multiple worker threads per memcached instance
- Must have another tier behind it — this is a cache, not a database.

Memcached normally resides in a tier behind webservers/application servers, but in front of the database systems. This provides a critical means of easing read pressure on the database tier, as this can be a common bottleneck in high-volume webserver environments. As the web traffic grows, increasing pressure is placed on the databases. Adding memcached as an intermediate tier provides relief to the typically expensive database systems and allows for easy expansion of the caching layer by scaling system count.

Challenges of standard memcached deployments

For the majority of standard memcached deployments around the world, Dell believes that maximum performance is hindered by a variety of factors. Some of the most significant among these factors is the fact that standard interrupt-bound NIC models commonly in use create conditions where smaller key/value sizes tend to overwhelm the CPU cores tasked to service them. This creates a scenario where interrupt handling becomes a bottleneck, and the overall transactions per second performance of memcached is slowed substantially.

In addition, maximum throughput characteristics are typically restricted in these common deployment configurations due to small numbers of relatively slow Ethernet connections, 1Gb Ethernet perhaps still being common in many older deployments. Even when such deployments contain multiple faster NICs, these adapters are often not commonly NUMA aligned, creating performance bottlenecks.

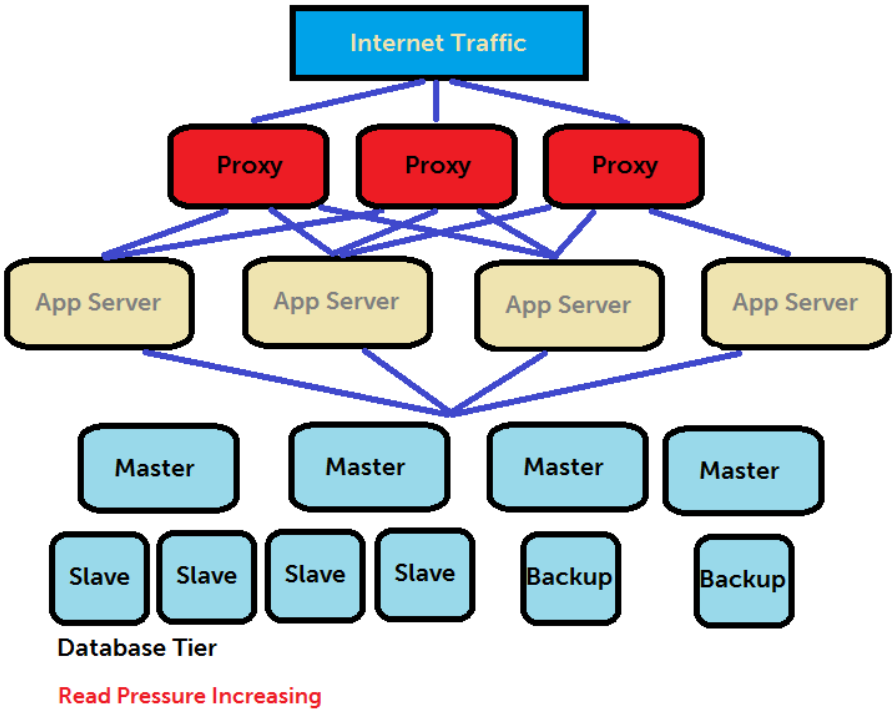
Some other factors include issues with older memcached versions that encounter internal locking inefficiencies that limit worker thread performance scaling. This has been lessened to some extent by more recent versions of memcached. Even current versions suffer a noticeable slowdown in performance when performing write operations, especially with smaller key/value sizes.

Finally, a common implementation scenario for memcached is one where single instances of memcached are applied to "lightly configured" servers, and server numbers tend to be scaled out to service increasing website traffic demands. As these lightly configured systems grow in numbers, pressure is placed on rack space and power utilization, and management becomes more difficult.

Common memcached evolution scenario

In Figure 1 below, we see the beginning of an active website that is starting to experience the first growing pains of increasing readership. We see that due to increasing traffic, the database read pressure continues to grow uncomfortably.

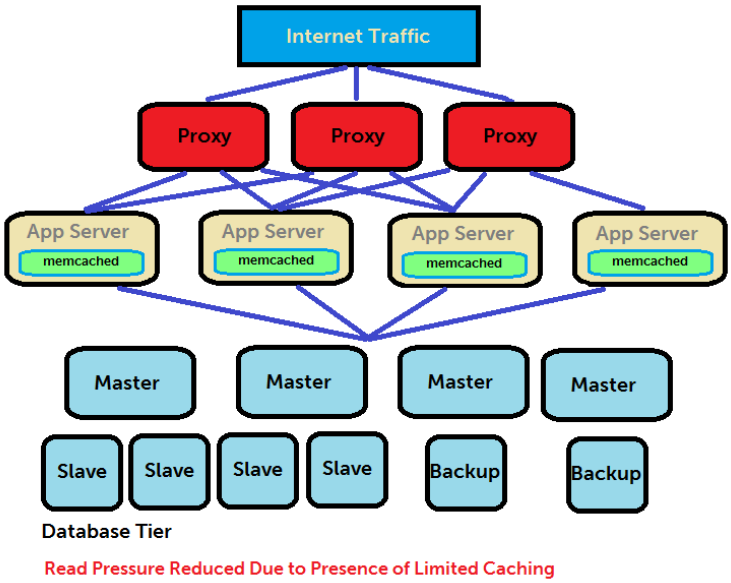
Figure 1. Active website with increased readership



In Figure 2, we see a common stopgap approach to relieving the database pressure being applied to the application servers by adding a memcached instance on the application server hosts themselves. This happens all too often, and the drawbacks of such an approach include reduced memory on the application servers themselves as they are now sharing between two classes of workload on the same hosts. Typically, these types of initial memcached implementations tend to be short-lived due to the inherent inability to scale to the increasing web traffic.

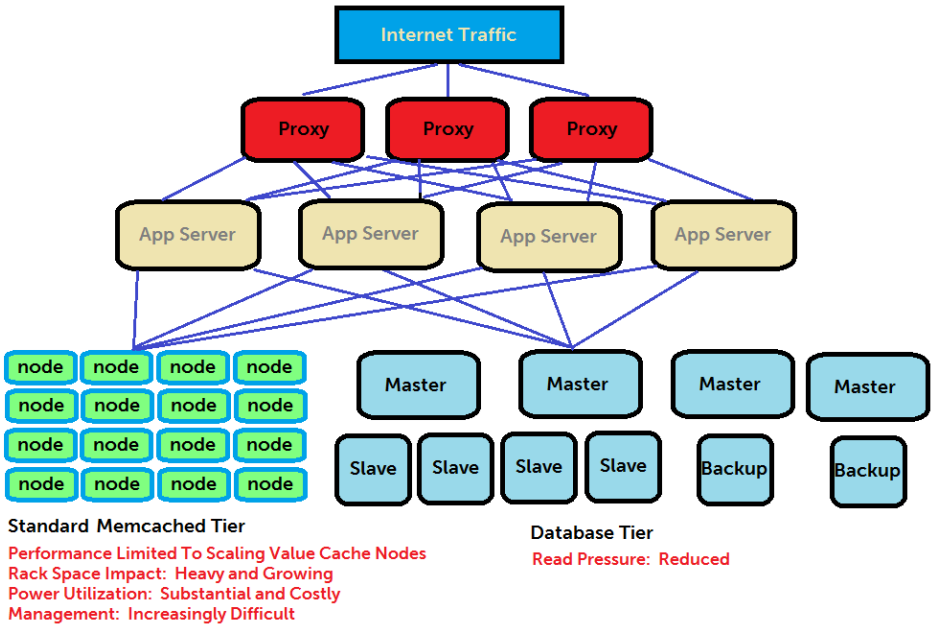
To support this stopgap implementation, the application servers will be modified to support memcached, thanks to example code being prevalent for practically all web-oriented application server software in most popular languages. After memcached integration, the application servers will query the newly installed memcached instance first with a key request. If the key/value pair is not cached, the application servers will go to the database tier to fulfill the request. Finally, the new key/value pair is written to the memcached instance. For requests already in the cache, the databases are not queried at all and the response time for user sessions is improved.

Figure 2. Common stopgap approach to relieving the database pressure



In Figure 3, we see the final common stage of memcached implementation. The memcached instances have been migrated to their own hosts and now comprise a standalone memcached tier. These hosts, which are typically somewhat "value" systems in terms of CPU and memory configurations, are scaled to meet increasing website demands. The databases are protected from read pressure, but as the number of hosts continues to scale it is common to see new issues arise such as rack space utilization growing along with increasing power demand due to the scaling numbers of these relatively weak systems. In addition, management issues tend to arise from the growing host count in this tier.

Figure 3. Final stage of memcached implementation



Enter Solarflare and OpenOnload

Solarflare is a high-performance network interface vendor, providing network solutions that are highly popular for customers who prize high performance and very low latency. Along with advanced Ethernet network interface adapters, Solarflare has open-sourced OpenOnload — a high-performance kernel bypass network stack that dramatically reduces network latency and CPU utilization. OpenOnload runs on Linux® and supports TCP/UDP/IP network protocols with the standard BSD sockets API. In addition, OpenOnload requires no modifications to applications for use. This makes optimizing for performance and low latency dramatically easier than some other competing products.

OpenOnload is able to achieve much of its performance improvements by performing network processing at user-level, bypassing the OS kernel entirely on the network path. For applications that leverage OpenOnload, the spinning model removes traditional interrupts from the equation entirely by configuring Onload to spin on the processor in user mode for a specific time period — something that can dramatically lower CPU utilization. High CPU utilization levels is something that would be common for small to medium key/value size memcached deployments— thus the removal of interrupts means that processor overhead formerly dedicated to interrupt handling can now be leveraged to increase certain types of application performance, including memcached.

Onload is also capable of operating in the traditional interrupt-driven mode. In this mode, the user will still get the benefits of avoiding the user space to kernel context switch experienced with the spinning model. However, this configuration was not tested for this study.

OpenOnload is open source software that is officially supported by Solarflare. In addition, Solarflare offers EnterpriseOnload – a subscription-based version of OpenOnload that includes a support contract, service level agreements, and extended product lifecycles.

Test configuration

The System Configurations used in the Dell memcached testing setup are listed below:

System under test (SUT):

- 1 x Dell PowerEdge R730
- 2 x Intel® Xeon® E5-2699 v3 18-core processors
- 16 x 16GB dual-rank 2133MHz DDR4 RDIMMs (256GB)
- 1/2/4 network adapters (Solarflare Flareon 7142 dual-port 40GbE)
    - OR 1/2/4 “common” network adapters (Intel XL710 dual-port 40GbE)
- Red Hat® Enterprise Linux® 6.5
- OpenOnload 201509-u1
- Memcached 1.4.21 (+ Solarflare move_fd patch)
- 1 memcached instance per populated NIC port, 20GB cache per instance.
- Custom BIOS settings

Client systems:

- 4 x Dell PowerEdge R720
- 2 x Intel Xeon E5-2697 v2 12-core processors
- 8 x 16GB Dual-Rank 1866MHz DDR3 RDIMMs (128GB)
- 2 x Solarflare 6122 dual-port 10GbE network adapters
- Red Hat Enterprise Linux 6.5
- OpenOnload 201509-u1
- Memaslap client tool (from libmemcached-1.0.18)

Network configuration:

- 1 x Dell Networking S4810 10Gb / 40Gb switch
- 4 x VLANs
- 16 x 10Gb ports connected to clients
- 4 x 40Gb ports connected to SUT

One element of the test configuration involved scaling the network adapter count. For both the dual-port 40GbE Solarflare adapters and the dual-port 40GbE "common" adapters we tried 4, 2, and 1 adapter populations. In the case of a 4 adapter population, one NIC port per adapter was populated. In the cases of 2 and 1 adapter population, both NIC ports per adapter were populated.

Each NIC port connected on the SUT represents a separate subnet, and with that a separate instance of memcached. Each subnet has connections to all 4 clients, and these clients are all capable of exercising all 4 networks (and the listening memcached instances) separately or together.

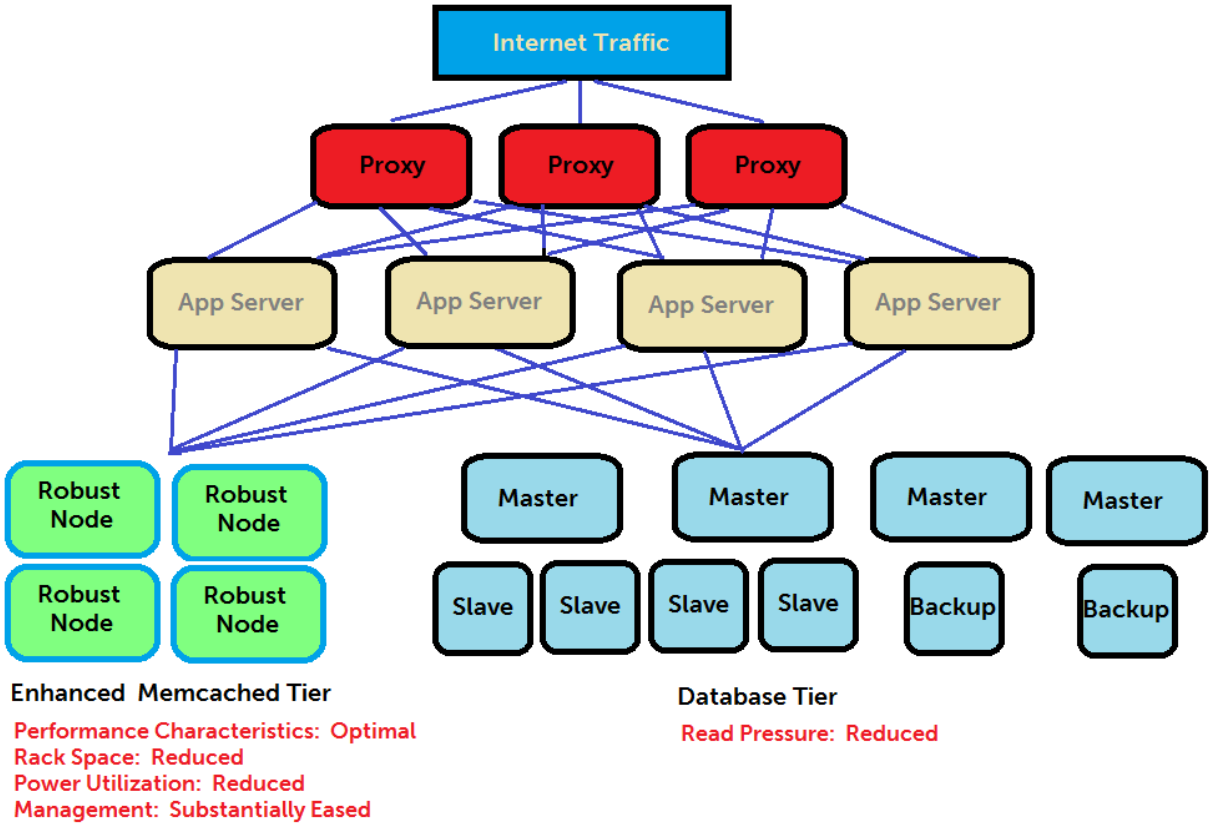
In these memcached tests, Dell did an apples to apples comparison of Solarflare Flareon 7142 adapters to Intel XL710 in the same robustly configured SUT, and tunings were applied optimally to each configuration. The results of these tests illustrate the advantages to the memcached workload that are brought to the table with the addition of Solarflare technology as compared to the Intel 40GbE solution on the same server.

Compared to what Dell believes are common memcached deployment server models, the uplift provided by the Dell/Solarflare robust server model should be much greater than what is displayed in the performance results in this study. This is due to the dramatically lower performance of standard lightly configured memcached server deployments. Compared to this low standard, the Dell/Solarflare combination should provide huge consolidation opportunities due to vastly increased performance.

The new memcached model

In Figure 4 below, we see the new memcached tier model taking shape. The large numbers of relatively weak memcached systems has been replaced by substantially fewer, more robust Dell/Solarflare systems. Each of these systems is populated with a higher core count, larger memory size, and Solarflare NICs. The combination of these new ingredients mean each system is capable of vastly greater performance compared to the common thinly provisioned memcached server model of the past, and thus considerable consolidation is achieved in this tier. The larger memory size means that each memcached instance can have a much deeper cache. Along with reduced rack utilization, implementers would naturally see a concomitant power utilization reduction inherent with consolidation exercises.

Figure 4. New memcached tier model



Memcached access patterns tested

There are four primary memcached access patterns we used for these tests. Broadly, they can be broken up into two groups: 100% Read and 9:1 Get/Set ratio. The 100% Read scenario, as the name implies, measures performance characteristics when executing all read requests from the clients. This scenario would emulate the performance of a well-warmed memcached server, where all incoming requests are already in cache. Separately, the 9:1 Get/Set test will inject write operations into the request mix at a ratio of 9 reads (Get) to 1 write (Set). Due to internal locking mechanisms of memcached, adding write commands into the mix has a significantly negative impact to performance.

The other aspect to the tests performed for this white paper involved single (single get or "sget") operations vs. batched (Multiget or "mget") operations. The batched Multiget operations pack multiple messages into a single request at a 48:1 ratio. As we will see with some of the charts in the next section, best case performance is achieved with 100% Multiget operations. Dell recommends this Multiget caching response model be utilized if possible in your environment due to the dramatically higher performance characteristics in a warmed cache environment. Achieving 100% read scenarios should be easier to achieve and sustain for this robust Dell/Solarflare memcached configuration with deep caching capabilities made possible by the large memory population.

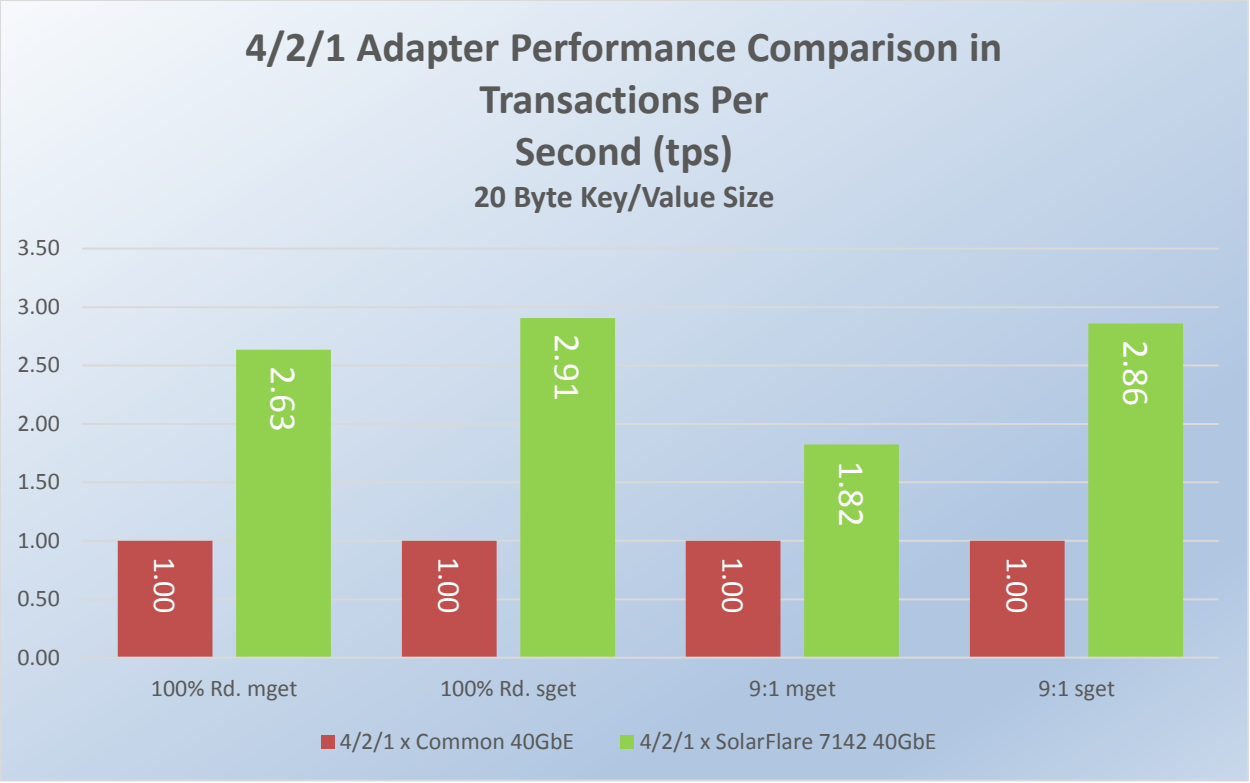
Testing highlights

During the course of testing Dell discovered that the "maximum configuration" SUT with 4 adapters was capable of hitting 160 Gb/sec sustained throughput for the largest key/value sizes. In addition, at the smallest key/value sizes tested we were able to demonstrate performance of over 40 million transactions per second. Evidence of other publicly disclosed memcached systems with similar capabilities has not been found, leading us to conclude that this configuration is currently the most performant memcached configuration.

We analyzed the transactions per second performance characteristics of memcached as the network adapter count was scaled, using both Solarflare 40GbE NICs and "common" 40GbE NICs as a comparison point. As we can see in the following sections of this whitepaper, the performance deltas noted between the Solarflare configurations and the common configurations were quite large. The Solarflare configuration was over 2.5x of the common 40GbE adapter performance across 3 of the 4 request mixes. For the 9:1 Multiget scenario, the Solarflare configuration was capable of over 1.8x the performance of the common adapter configuration. The smallest tested key/value pair size of 20 bytes was used for these comparisons.

Note that the adapter comparisons in Figure 5 below show the same overall performance disparities between the common 40GbE configuration and that of the Solarflare model from 4, 2 and 1 adapter populations. This is due to the fact that at the smallest key/value sizes (20 bytes in this example), adapter count does not impact transactions per second. It is only as the value size increases that we see adapter count play a larger role as throughput constraints become more evident.

Figure 5. 4/2/1 adapter comparison (20 byte value size)

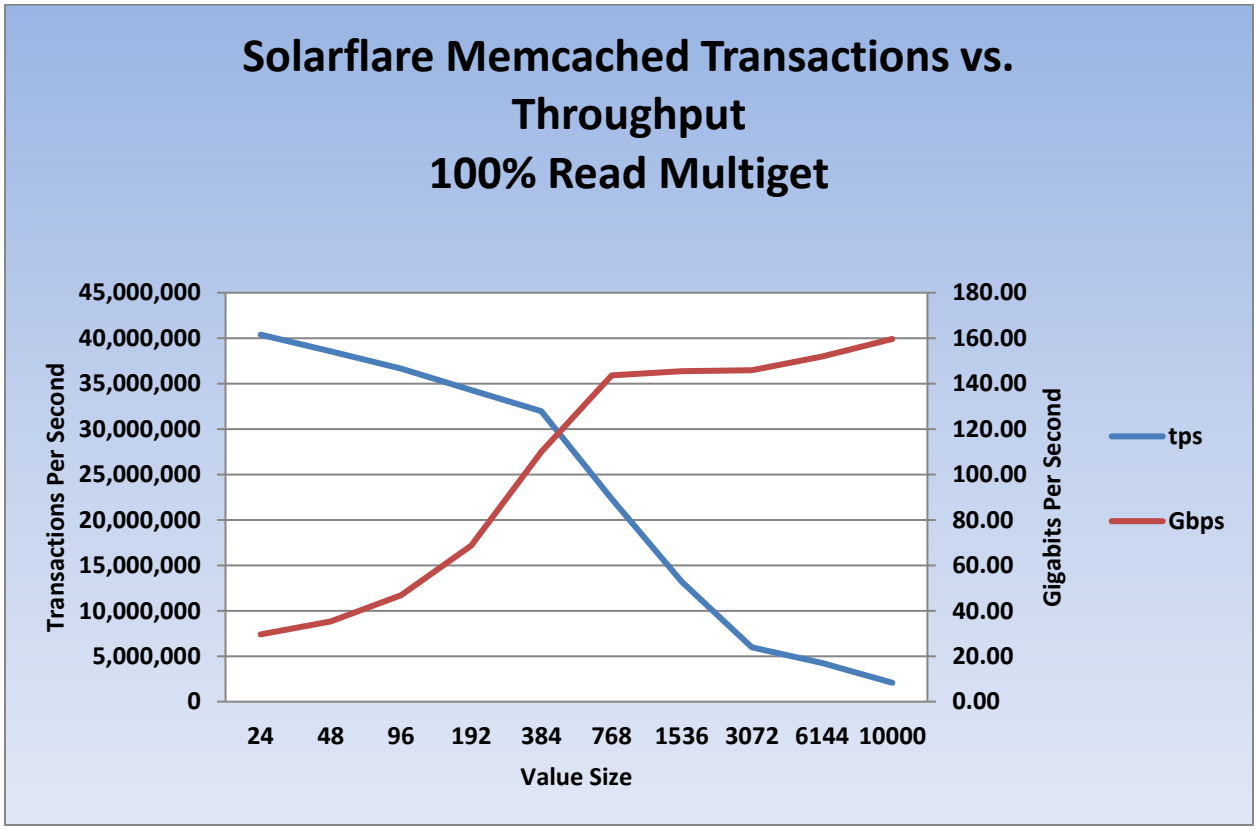


Value size and adapter count

In figure 6 below, the 4 adapter Solarflare configuration shows the relationship between transactions per second and overall throughput as the key/value sizes are scaled from smallest at 20/20 bytes to largest values size that Dell tests which was 64/10,000 bytes. At the smallest key/value sizes, adapter count makes no real difference in our tests. A single adapter is capable of the same transactions per second rate that four adapters are with this value size. However, as the value size scales, so does the effects of adapter count in overall performance as the throughput limitations become increasingly critical. Knowing the maximum value size of a key/value store will allow you to infer the correct adapter count to ensure maximum throughput. Dual-Port 40GbE adapters are limited by the x8 PCI-E Gen 3 slot throughput, so 80 Gb/sec is not realistic to expect from a single dual-port adapter. As all of the testing was done in a NUMA-aligned environment, it is expected that 3 adapter configurations with both ports utilized may start to approach the maximum throughput achieved with a 4 adapter configuration with only a single port per adapter populated across many value sizes. This 3 adapter

configuration, however, is expected to experience increased latency characteristics and represents an unbalanced configuration that will likely be constrained by system architecture. For optimal performance characteristics with value sizes larger than 384 bytes, Dell recommends 4 Solarflare adapters (single port population on each adapter). For value sizes larger than 128 but below 384 bytes, two Solarflare adapters with both ports populated (4 total ports populated) should be capable of optimal performance. For value sizes 128 bytes and below, a single Solarflare adapter should be sufficient with both 40GbE ports populated.

Figure 6. Solarflare Transactions vs. Throughput



Additional adapter count experiments

In a separate set of experiments, the two Network Interface types were compared together while varying the key/value pair sizes. This test, as shown in Figure 7, clearly demonstrates the performance enhancements evident in increased Solarflare adapter count, even at smaller sizes.
At smaller key/value pair sizes, Solarflare adapter count shows a clear scaling trend when compared to the common 40GbE configuration due to the efficiencies of OpenOnload and the high performance Solarflare adapter design. The common 40GbE configuration, seen in Figure 8, shows a much different scaling story for most of the key/value sizes explored as the two and four adapter transactions per second measurements were nearly equivalent. Only at the higher value sizes was there a noticeable benefit for the four adapter count for the common adapter type, unlike the Solarflare configuration which shows strong scaling characteristics across the tested value size range.

Figure 7. Solarflare adapter count tps scaling

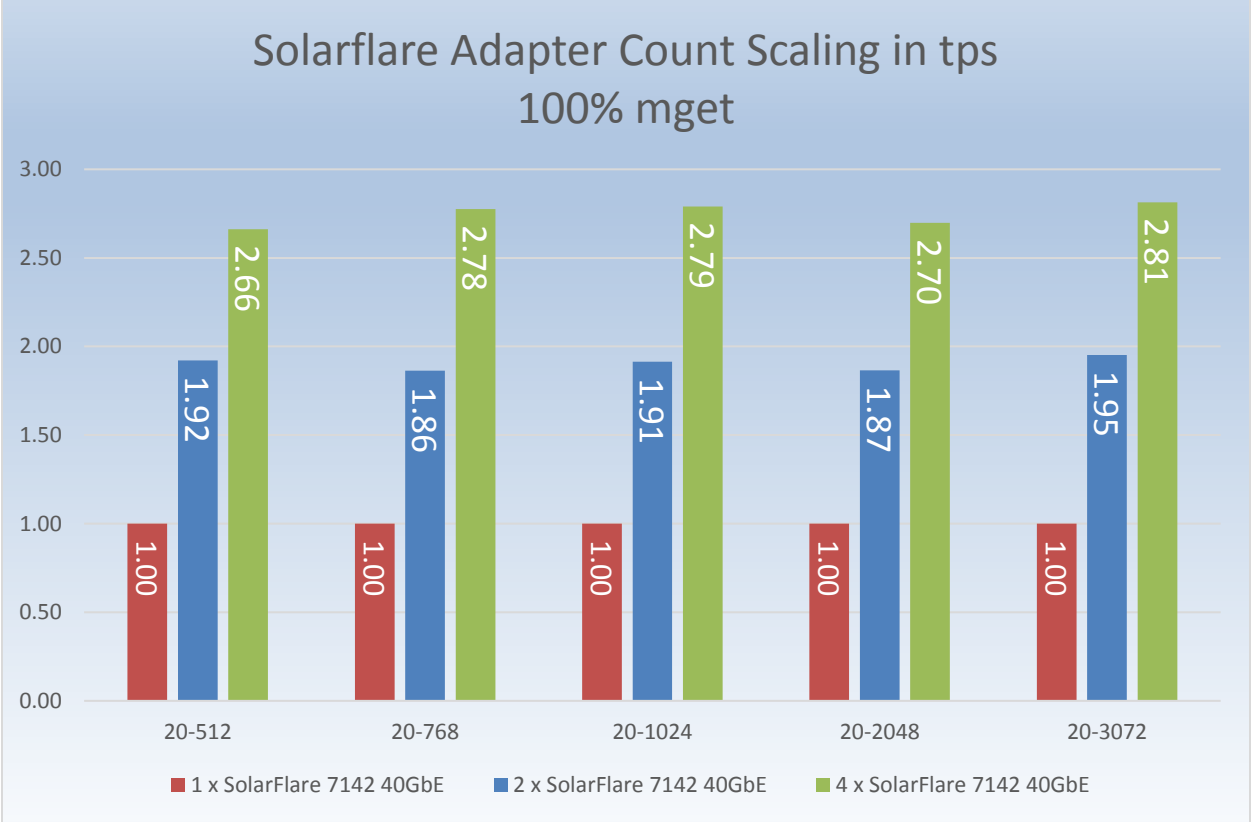
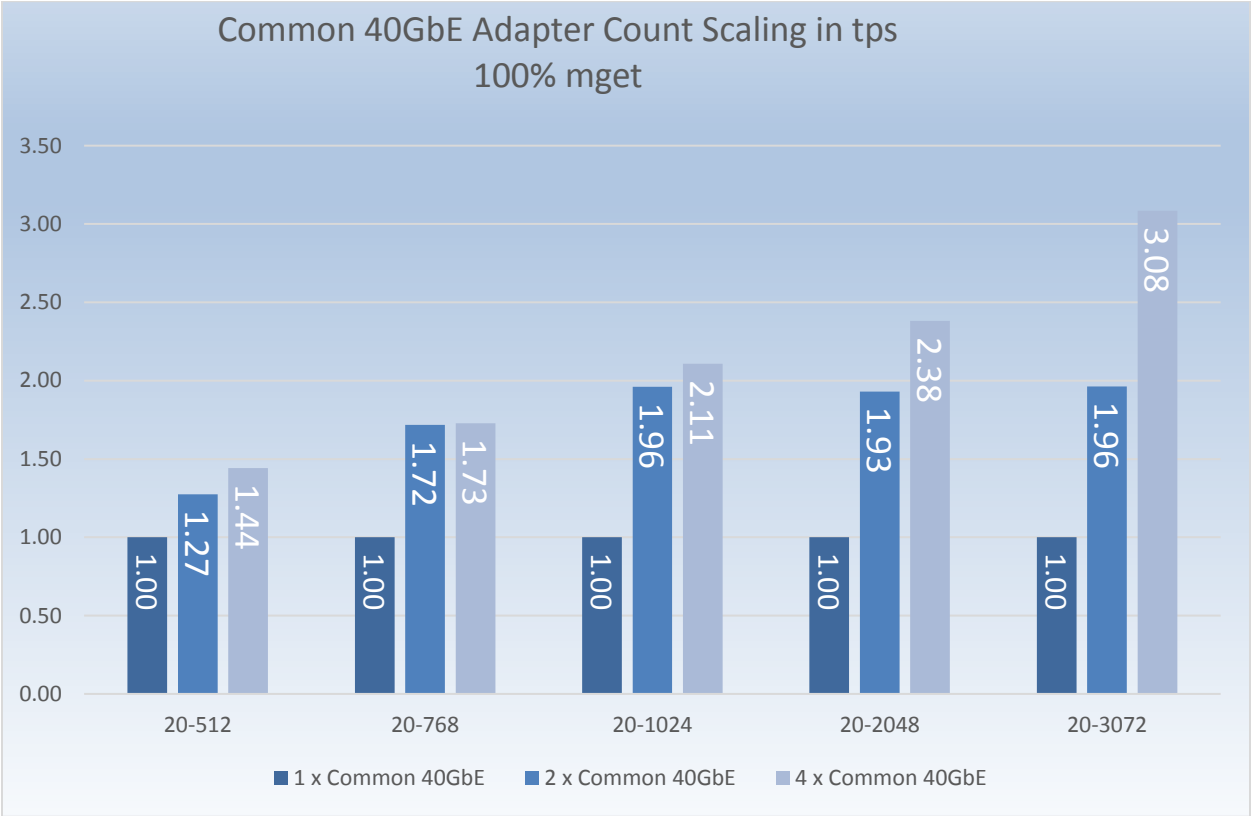


Figure 8. Common 40GbE adapter count scaling in tps



Another aspect of the test results revealed that beyond raw transactions per second performance, the adapter count made a real difference in the throughput performance capabilities of the Solarflare system. This was noted on the common 40GbE configuration to different extent. Both interface types show clear throughput scaling at the higher key/value sizes, indicating that individual adapter throughput limits become more important as the value size increases. This is illustrated in Figures 9 and 10 below. At lower value sizes, the lack of scaling evident for the common adapter type suggests CPU saturation due to driver model inefficiencies, which is something that is not evident in the Solarflare configuration.

Figure 9. Solarflare adapter count

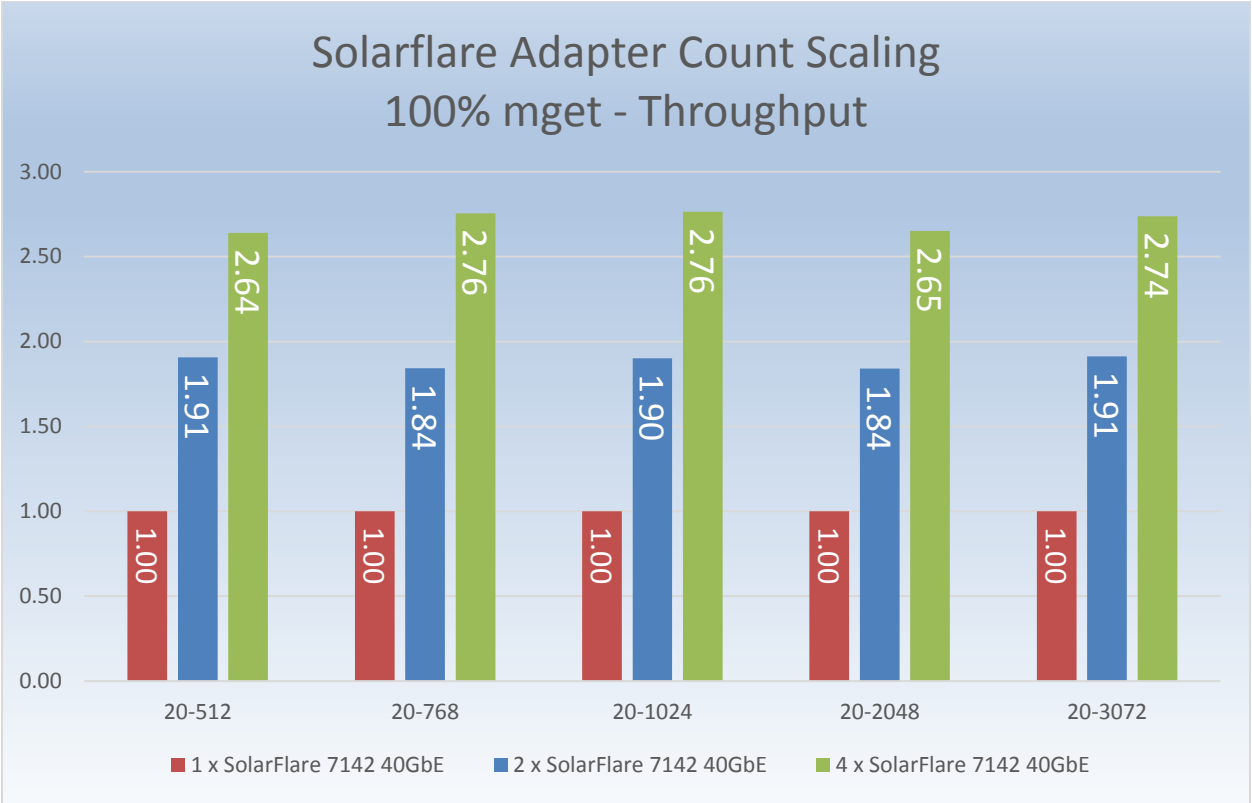
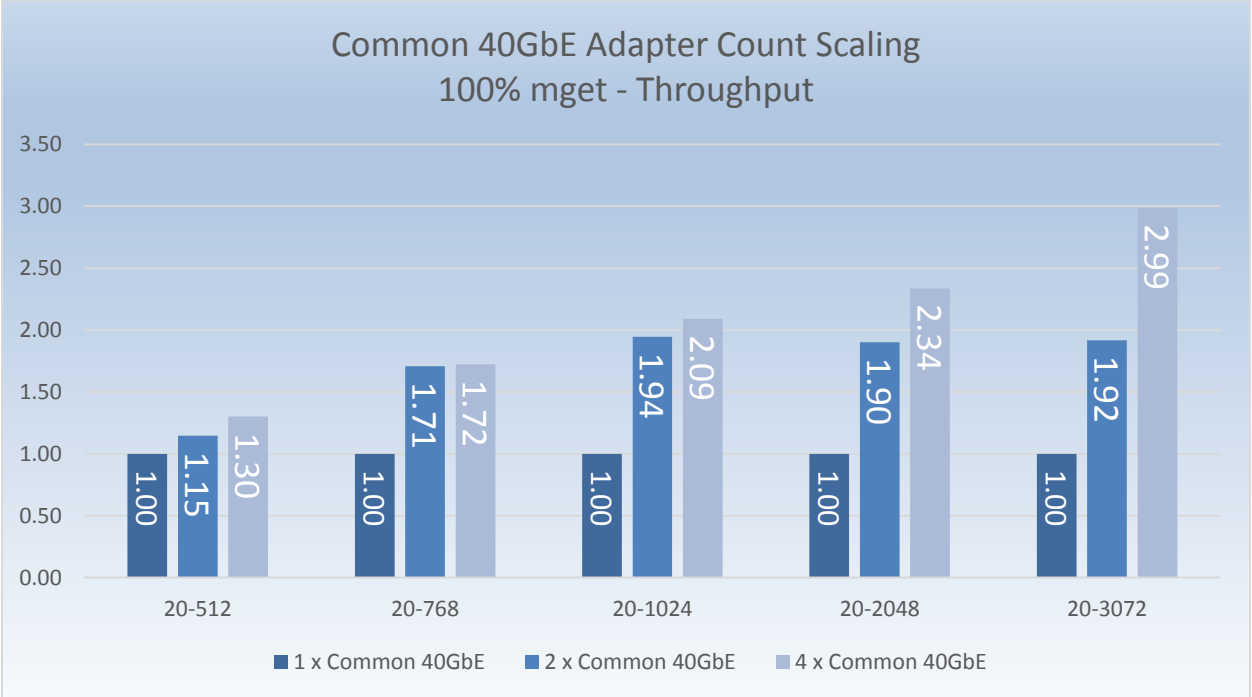


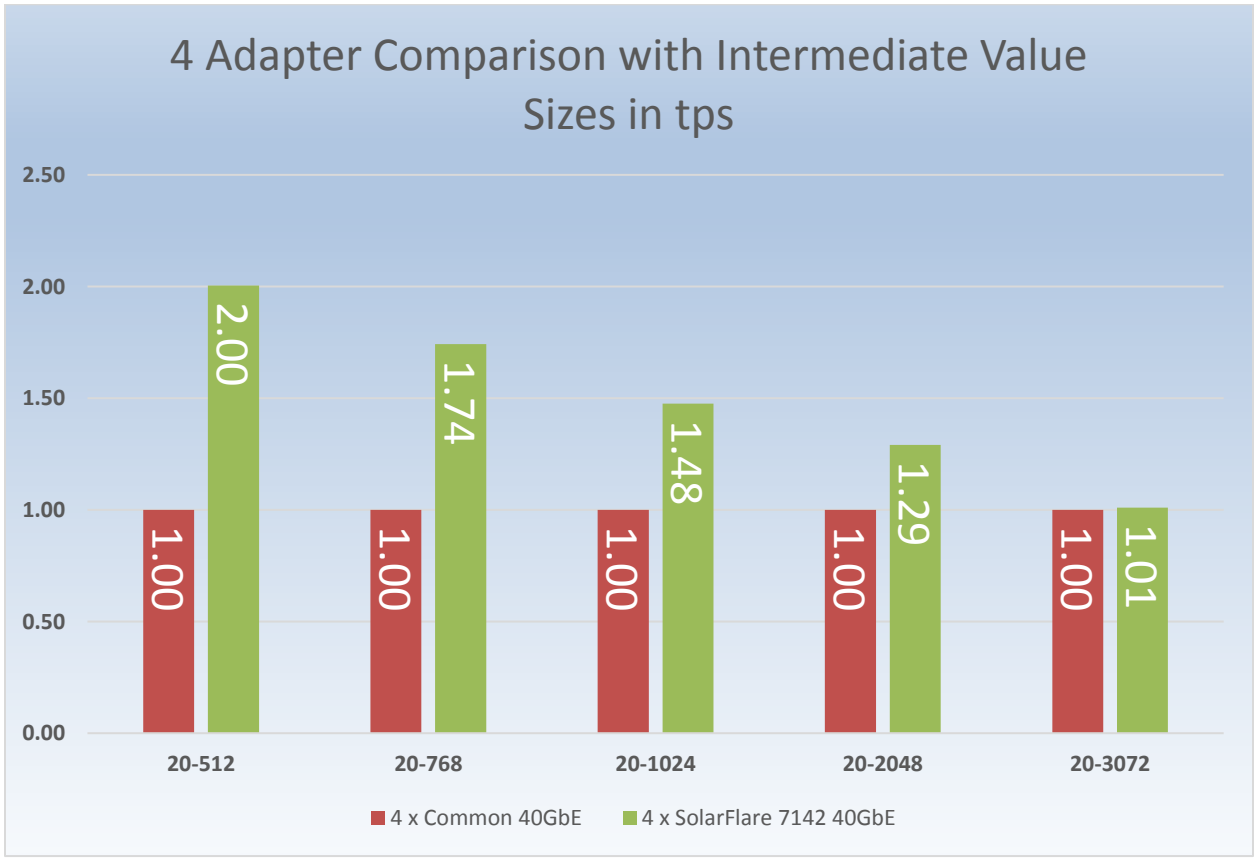
Figure 10. 40GbE adapter count



Intermediate value size performance comparison

Another view of the same experiment will help to understand the performance positioning between the Solarflare and common 40GbE configurations. In the previous section the relative scaling characteristics of both adapter types were contrasted in separate charts. In Figure 11 below, we see the relative transactions per second from a 4 adapter population for both network adapter types. The adapter count chosen for this illustration was intended to represent the best-case capabilities of both adapter types. Examining the results, we see that as the value size increases, the performance disparity starts to decrease. At the very highest value sizes, the relative performance between the two adapter types is diminished due to the reduced system pressure brought about by low transaction rates. It is Dell's understanding that the majority of memcached deployments do not scale to large value sizes, so for the majority of cases the Solarflare configuration will prove dramatically superior to the common 40GbE configuration.

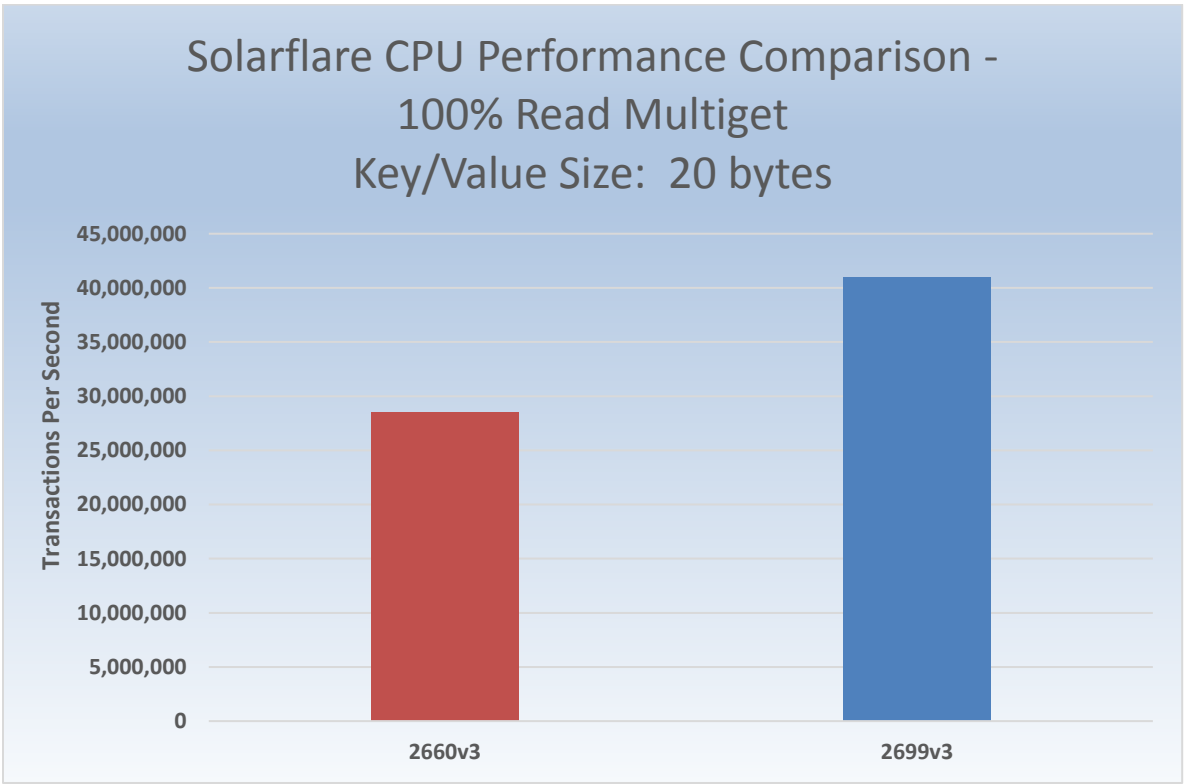
Figure 11. Intermediate Value Size Comparison



CPU choice effects

Dell also performed some experiments using an alternate processor selection — the Intel Xeon E5-2660 v3 10-core processor also configured with Solarflare adapters. We compared this configuration to our Intel Xeon E5-2699 v3 18-core model using the 20 byte key/value size, as this was identified as the area of greatest pressure on the CPU. Although the adapter count varies between the two configurations, Dell determined that for this smallest key/value size adapter count does not factor in the performance capabilities of the system. We see a substantial reduction in performance for 100% Read Multiget, where the Intel Xeon E5-2660 v3 provides approximately 30% less transactions per second than the Intel Xeon E5-2699 v3. With less cores available, memcached instances are forced to use less worker threads. This leads to a large reduction in top-end performance, leading us to recommend a robust configuration choice for best case consolidation activities.

Figure 12. CPU performance comparison



Conclusion

Dell PowerEdge servers combined with Solarflare high-speed Network Interface Cards provide unrivalled memcached performance characteristics. Optimal system configurations are capable of handling almost line-rate throughput at larger key/value sizes, as well as excelling in transactions per second at the lower key/value sizes. Compared to the common 40GbE configurations we tested, Solarflare NICs and the OpenOnload network stack proved enormously capable and were easily leveraged to accelerate application performance.

All testing in this whitepaper compared memcached performance on a single robust SUT with two different network adapter types. Although dramatic performance gains are seen by using Solarflare compared to the common 40GbE configuration, the performance uplifts seen by customers should be propelled even higher if contrasted against an existing memcached deployment. This is due to the typically "value" servers and networks utilized in these older deployments.

The results of the testing indicates that dramatic performance gains can be had by aligning to the Dell/Solarflare memcached model. Consolidating existing memcached deployments are another aspect to this new memcached model, as larger numbers of older and less-capable systems can be consolidated into fewer, more highly capable systems. All of the natural benefits of consolidation such as reducing power utilization, rack footprint, and lowered TCO will be realized by utilizing this new memcached model.

If the Dell/Solarflare memcached model looks promising to you, please contact your Dell Sales Representative for a deeper discussion.