Breaking Through the Storage I/O Barrier For Cloud Computing

How Marvell DragonFly Enables Massively Scalable Virtualization While Reducing Storage Costs By 50 Percent or More

Shawn Kung
Director of Product Marketing
Marvell

August 2012
Overview: The Cloud Computing Storage I/O Bottleneck

It’s no secret that the cloud computing market has been growing rapidly both for public and private deployments. Research from International Data Corporation (IDC) indicates that server sales for public cloud computing will grow to $718 million by 2014 and to $11.8 billion for private clouds during the same time period.¹ IDC also projects public cloud services will grow rapidly at 30 percent per year, which is a 5x higher rate than global IT spending in 2011.²

With this impressive and continuing growth, one of the biggest challenges in cloud computing is how to solve the storage I/O bottleneck that comes with large-scale virtual machine (VM) deployments. As the number of Virtual Machines (VMs.) grows, the corresponding explosion in random read and write I/Os inevitably bring a network attached storage/storage area network (NAS/SAN) array or local direct attached storage (DAS) to its knees as disk I/O or target-side CPU performance become bottlenecked.

For public or hosted clouds, the growth is driven by an explosion of e-commerce, small/medium business (SMB) and distributed enterprises outsourcing their platform infrastructure for cost-efficiency and resource elasticity. The result is a proliferation of virtualized multi-tenant OLTP databases and web servers that create a massive amount of random write and read I/O traffic. For private or enterprise clouds, Fortune 1000 data centers are struggling with spiraling VM scalability pains driven by server consolidation and emerging IT initiatives that accelerate VM density, such as virtual desktop image (VDI) applications. To work around these pains points, storage administrators often resort to buying more storage array hardware, a solution that does not scale effectively in purchase cost, administration overhead and maintenance expenses (power, cooling, floor space).

This paper discusses the rapidly worsening storage I/O bottleneck brought on by the explosive growth of cloud computing and how the Marvell® DragonFly™ is uniquely positioned to solve for this.

Growing Pains: Storage I/O Scalability in Cloud Computing

Enterprise virtualization was initially driven by the desire to improve CPU utilization and other resources for under-utilized servers. At the turn of the 21st century, many data centers saw their server farms yielding only 10 to 20 percent CPU utilization. Not only was this a waste in hardware budget, but the ongoing hidden costs for space, power and cooling to run these servers was a recurring maintenance expense that raised overall IT operations costs. During the past decade, many enterprises deployed virtualization hypervisors such as VMware® ESX and Citrix® Deserver, to successfully consolidate applications onto fewer servers for significant cost savings.

Reinforcing the server consolidation trend is the emergence of public cloud computing service providers. Many SMBs, Internet content providers, e-commerce companies, and even enterprises, have embraced the cost-efficiencies of outsourcing their data center infrastructure needs to a third-party service provider. During this same period, IT infrastructure demand has accelerated due to the proliferation of client devices such as smart phones, tablets, netbooks and other home cloud devices, resulting in unprecedented consumer mobility and Internet accessibility.
Figure 1: The mass consumer market is becoming increasingly “always-on”, driven by mobile and home cloud devices that enable pervasive access to Internet content and e-commerce services.

This has led to the “long tail” effect, whereby the mass consumer market is no longer limited to a few large web businesses, but can now access vast numbers of Internet businesses offering unique products.

Figure 2: Products that have a low sales volume can collectively make up a market share that exceeds the relatively few current bestsellers and blockbusters, if the distribution channel is large enough. The distribution channel opportunities created by the Internet often enable businesses to tap that market successfully. Source: “The Long Tail” by Chris Anderson, October 2004.
As a result, SMBs, Internet companies and enterprises have growing requirements for infrastructure scalability and elasticity to grow and shrink their IT resources on-demand. For public cloud service providers, the combination of rapidly growing multi-tenant resellers and businesses, coupled with unpredictable and growing infrastructure demands, is resulting in an urgent architectural need for cost-effective scalability. Service providers have executed the first step by migrating existing hosting customers from bare-metal DAS islands to virtualized cloud computing and shared storage, resulting in more cost-effective multi-tenant consolidation and server utilization.

But by solving the server utilization challenge, virtualization has inadvertently created a new pain point for both enterprise and public clouds. Enterprise server and storage infrastructures have traditionally experienced three major hardware bottlenecks that limit application performance and scalability:

- Host CPU/Memory
- Network Bandwidth
- Storage I/O

To address Host CPU bottlenecks, the onset of multi-core, multi-threaded x86 processors and DIMM slots has ensured plentiful computational and memory resources to allocate to a large number of virtual machines. And as 10GbE NICs and LAN-on-motherboard (LOMs) with TCP/IP offload (TOE) become increasingly pervasive, network bandwidth is fast becoming a non-factor for VM scalability. The commoditization of CPU and network bandwidth resources is enabling IT to scale to increasingly larger numbers of VMs, placing significantly heavier I/O load on storage.

In the past, a NAS/SAN array mounted by 50 physical servers meant 50 host connections. Today with an average of 10 to 30 virtual machines per host, the same 50 physical hosts yield the equivalent of 500 to 1,500 virtual servers connected to a single storage array. And because each VM is reading/writing data independently from other VMs in the same physical server, the result is a data access pattern that is increasingly random in both writes and reads. This results in two related pain points:

1. Storage Server Bottleneck: NAS/SAN arrays are brought to their knees and overwhelmed by the flood of random write and read I/O requests. Storage CPU and disk I/O rapidly become saturated and bring application performance to a standstill. NAS/SAN arrays cannot serve read/write requests fast enough to service the multitude of application server VMs, resulting in lost end-customer revenue (e.g. unfulfilled e-commerce transactions).
2. Higher Virtual Server Latency: Application server VMs face inconsistent performance as some VMs wait for long periods for write or read I/Os to complete. These VMs become starved for IOPS, jeopardizing performance service level agreements (SLAs) and reducing productivity. In a worst case scenario, slow I/O response can result in application timeouts and unplanned outages.

In order to work around both issues, today’s storage administrators have one of two standard options—either limit the number of VMs connected to a given NAS/SAN target, or spend more IT budget to upgrade their NAS/SAN array. Both alternatives require compromising on scalability or cost, resulting in non ideal solution. Even emerging options, such as a host software cache with SSDs, are very limited. Often host cache software is application-specific, requires tedious porting across multiple operating system versions and burdens the host with significant CPU and memory overhead. Moreover, a dirty little secret is that an SSD-only approach yields worsening random write performance over time. With no good solutions on the horizon, storage admins are increasingly worried about how to break through the storage I/O barrier. As virtualization continues its explosive growth (50 percent of new servers are predicted to be virtualized by 2012), the storage IOPS bottleneck will only get worse with time, as shown in the illustration below.

Figure 4. VM Scalability Bottleneck: As large-scale virtual server clouds become prevalent, VM scalability is bottlenecked by storage I/O.

A Host Cache Solution: Marvell DragonFly

Marvell DragonFly is a revolutionary cloud computing I/O acceleration technology platform that dramatically scales virtual server performance and lowers capital spending costs for networked storage (NAS/SAN) and direct attached storage (DAS). Offered as an adapter level solution powered by Marvell embedded technology, DragonFly is targeted at public and enterprise cloud computing environments. DragonFly leverages intelligent read, write-back caching and the industry’s first write-back caching on tiered non-volatile memory to yield 10x higher virtual machine input/outputs per second (IOPS), while lowering NAS/SAN/DAS disk and CPU overhead by 50 percent or more.
The Marvell DragonFly enables the creation of a next-generation cloud-optimized data center architecture where data is automatically cached in a low-latency, high-bandwidth “host I/O cache” in application servers on its way to/from higher-latency, higher-capacity data storage.

Figure 5: Marvell DragonFly

A unique differentiator for Marvell DragonFly is its use of a sophisticated NVRAM log-structured approach for flash-aware write buffering and re-ordered coalescing. Unlike writing to SSDs that quickly degrade after a certain number of random writes, DragonFly ensures consistently high performance and low-latency with zero write performance degradation over time. An additional unique feature is its ability to create a host I/O cache that is agnostic to all network and local-attached storage protocols. It can be configured to serve as a data cache for DAS, SAN or NAS storage arrays, irrespective of protocols such as iSCSI, SAS or NFS that are used to access the data storage.
The distributed host I/O cache of the Marvell DragonFly employs a two-level cache for very low-latency and high burst throughput. The two levels are comprised of a level-1 (L1) non-volatile DRAM (NVRAM) cache and a level-2 (L2) cache created from off-the-shelf commodity SSD drives. The two-level cache enables consistent high-level performance to the applications running on the server, providing cost-effective incremental IOPS scaling to match your application server growth.

As an industry standard PCIe adapter, DragonFly platform seamlessly fits into all commercially available rack-mounted servers and is operating system independent. DragonFly is incrementally scalable - add an adapter as you add new application servers, and add commercially available 2.5-inch/3.5-inch SSDs to each server as your capacity or I/O needs grow over time. Marvell DragonFly offers a seamless, turnkey solution that accelerates raw performance of hardware cache implementations.

To translate this into practical terms, Cloud Computing customers can achieve the following advantages by deploying DragonFly on application servers:

1. Using your existing NAS/SAN array, **scale the number of VMs by up to 10x** before upgrading to more expensive shared storage systems, or

2. Using your existing application server infrastructure, **downgrade to cheaper modular NAS/SAN arrays** without reducing the number of VMs.

By establishing a high IOPS data cache on the host, the storage array architecture becomes much more affordable. Instead of expensive SAS or Fibre Channel enterprise disk drives or SSDs, the existing array can leverage low-cost, high-capacity SATA drives. Instead of expensive high-end storage platforms, cheaper entry-level or mid-range platforms become more than sufficient.

Alternatively, Cloud storage administrators can get more mileage out of their existing systems before needing to upgrade, thus deferring IT capital expenses. Since DragonFly dramatically reduces I/O to backend capacity-optimized storage, background services, such as snapshot, mirroring, backup, compression and de-duplication, can run seamlessly, without causing resource contention with application I/O requests.
Marvell DragonFly Platform Feature Highlights

Marvell DragonFly embedded cache technology consists of sophisticated software and hardware-assist engines running inside purpose-built System-on-Chips (SoCs) as part of a turn-key hardware caching HBA. Completely independent of host server resources, DragonFly requires only a very thin kernel filter driver. This yields two significant advantages:

1. **Pervasive OS support:** All major operating systems and hypervisors are supported and new major and patch releases are quickly and easily ported.

2. **Near-zero host resources:** DragonFly fully implements the cache technology in the HBA hardware, resulting in near-zero usage of the host server resources.

The following are feature highlights of the Marvell DragonFly:

**Write Cache:** Intelligent write buffering technology prunes overwrites and sequentializes random writes, resulting in improved SSD performance. Dual-level cache employs NVRAM (L1) and SSD (L2) write cache to minimize application latency while maintaining a large SSD write buffer capacity. Data is asynchronously written back to storage to dramatically reduce IOPS to NAS/SAN, resulting in lower storage server disk and CPU utilization. The result is up to 10x improvement in random write (and write-burst) performance. Write cache is powered by two major embedded technologies purpose-built to solve the random write performance limitations of SSDs:

1. **Flash aware write-buffering**— de-stages writes from NVRAM to SSD in a log-structured manner that is aware of SSD wear-leveling and garbage-collection.

2. **Reordered write-coalescing**— de-stages writes from SSD cache to storage in an order that is algorithmically optimized to reduce IOPS to storage.

DragonFly is also VM-aware during guest migrations required for ongoing maintenance and upgrades to the host. To ensure cache consistency and coherence, DragonFly supports virtual machine migration via integrated cache purge and synchronization.

**Write-Through (Read) Cache:** Intelligent population and eviction algorithms enable granular I/O temperature mapping to distinguish hot vs. cold data at a sub-VM level (block and file). Configured as an option to default write-back mode, write-through operates as a read cache so that only random reads benefit. This offers up to 10x improvement in random read performance.

**HA Synchronous Mirroring:** Using a reliable, low-latency protocol, DragonFly employs synchronous peer-to-peer mirroring for data protection in the event of host node failure. Defaulted in write-back caching, the HA synchronous mirror protects against data loss and ensures high availability for mission-critical cloud and data center operations. In the event of HBA or catastrophic node failure, a synchronous mirrored data-cache replica is available on a peer node to guarantee all write updates will successfully be written back to storage, with zero downtime.

**Zero-Warm Up Cache:** Equipped with an onboard supercapacitor, industry-leading 6Gb/s SSD controller and SLC NAND Flash memory, DragonFly includes embedded NVRAM technology to automatically back up cache metadata to Flash in the event of power loss. Upon power restoration, cached metadata can be automatically re-populated in L1 NVRAM to enable zero cache warm up time.

**Multi-Protocol:** DragonFly supports multiple storage protocols for unified I/O caching. For networked storage, both file (NFS) and block (iSCSI, FCP, FCoE) protocols are supported. For direct-attached storage, SCSI protocol is supported for local SAS or SATA drives (direct-attach or SAS expander).
Case Study Example

A sample case study helps to quantify the advantages of the DragonFly solution for public cloud and enterprise cloud computing environments.

Assume a typical IT storage configuration consists of the following:

| Hardware Configuration | 50 rack-mount application servers. Assume $5K/server³  
1 high-end NAS/SAN array. Assume $200K per array⁴ |
|------------------------|--------------------------------------------------|
| Host Configuration     | Linux operating system and hypervisor (Xen or KVM)  
30 guest VMs. 80% Windows, 20% Linux. Mix of OLTP database, Web servers, Collaboration, and VDI applications.  
Average IOPS per Guest VM: 200 IOPS (4K random writes and reads) |
| Current Performance    | Total Host IOPS Demand: 50 hosts x 30 VMs/host x 200 IOPS/VM = 300,000 IOPS  
Total NAS/SAN IOPS Supply: 60,000 IOPS |

The following calculates the acquisition costs required to ensure the NAS/SAN array can supply enough write and read IOPS to meet the demands of the host applications:

- Host applications require 300K IOPS on the 50 hosts
- At 60K IOPS per array, the IT team needs to buy a minimum of 5 arrays to meet performance requirements. Note: if headroom for other storage services such as snapshots, mirroring or backups are required, more arrays may be required.
- 5 arrays x $200K per array = **$1.0M capital cost investment**

By using Marvell DragonFly as a host cache, IT storage admins can significantly reduce the I/O traffic to the storage array. This enables them to use less expensive arrays. In fact, each DragonFly can supply much more than the 6,000 IOPS demanded per host. Assuming 50K write and read IOPS supplied per DragonFly with a baseline 2GB NVRAM and one SSD, storage administrators can now dramatically scale the number of guest VMs per physical host.

The following calculates the acquisition cost when DragonFly adapters are used:

- Host applications require 300K IOPS on the 50 hosts.
- 50 DragonFly Adapters (one on each host) supply 2.5 Million IOPS.
- Assume NAS/SAN array I/O demand is reduced by at least 50 percent, implying that 150K IOPS is required, or 3 arrays.
- 3 arrays x $200K per array = **$600K capital cost investment⁵**

The following exhibit summarizes the VM scalability and storage cost savings of the Marvell DragonFly approach:

<table>
<thead>
<tr>
<th>Scenario</th>
<th>VMs/Host</th>
<th>IOPS Demand</th>
<th>IOPS Supply</th>
<th>Capital Cost (US$)</th>
</tr>
</thead>
</table>
| **Today:**  
50 host servers + Expensive NAS/SAN | 30 Guest VMs | 300K | 5 Arrays: 300K | Hosts: $250K  
Arrays: $1.0M  
**Total: $1.25M** |
| **Future:**  
50 host servers + Marvell DragonFly + Cheaper NAS/SAN | 250 Guest VMs (50K write/read IOPS via DragonFly)  
8x VM scalability | 300K | 50 DragonFlys: 2.5 Million IOPS  
3 Arrays: 180K | Hosts: $250K  
Arrays: $0.6M  
**Total: $0.85M**  
(Savings: 32% Lower CapEx Cost) |
In fact, the cost advantages go well beyond this for storage architects that can evolve their current storage environment over time. Marvell DragonFly gives storage architects and CIOs the rare flexibility to completely re-think their storage architectures and replace high-end NAS/SAN arrays with low-cost capacity storage. Since most application IOPS are handled in the distributed host I/O cache, the back-end array can use cheaper commodity CPUs and SATA disk storage. For example, rather than a $200K high-end SAN, a mid-range NAS/SAN costing $20K to $40K is more than sufficient. Even if the cheaper array can support only 3K to 8K IOPS instead of the 60K IOPS from a higher-end array, this is acceptable since most of the read and write IOPS are serviced from the DragonFly host cache.

An added benefit in shifting to a capacity-optimized NAS/SAN tier is the reduction in ongoing maintenance costs. While capital expenditure (CapEx) cost reduction is an immediate benefit, the DragonFly architecture significantly lowers power, cooling, space and administration costs. Fewer storage arrays with larger-capacity drives yield a much greener data center with a smaller physical and power footprint and less management complexity. With 80 percent of total cost of ownership (TCO) typically occurring after the initial capital purchase, DragonFly offers the potential for CIOs to save millions more over the life cycle of their storage infrastructures. An example of TCO savings is shown in the exhibit below.

### Total Cost of Ownership Comparison

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Capital Cost (existing architecture)</th>
<th>Capital Cost (new architecture)</th>
<th>Ongoing Maintenance Costs (5 yr)</th>
<th>Total Cost Of Ownership (TCO) – 5 yr.</th>
</tr>
</thead>
</table>
| **Today:** 50 host servers + Expensive NAS/SAN | Hosts: $250K  
Arrays: $1.0M | Hosts: $250K  
Arrays: $1.0M | Assume 80% of total costs | Grand Total: $7.5M |
| **Future:** 50 host servers + Marvell DragonFly + Capacity NAS/SAN Storage | Hosts: $250K  
Arrays: $0.6M | Hosts: $250K  
Arrays: $60K  
($20K x 3) | Assume 80% of total costs | Grand Total: $1.86M  
(Savings: 75% Lower TCO over 5-year period) |

**Total Cost of Ownership Comparison**

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Capital Cost (existing architecture)</th>
<th>Capital Cost (new architecture)</th>
<th>Ongoing Maintenance Costs (5 yr)</th>
<th>Total Cost Of Ownership (TCO) – 5 yr.</th>
</tr>
</thead>
</table>
| **Today:** 50 host servers + Expensive NAS/SAN | Hosts: $250K  
Arrays: $1.0M | Hosts: $250K  
Arrays: $1.0M | Assume 80% of total costs | Grand Total: $7.5M |
| **Future:** 50 host servers + Marvell DragonFly + Capacity NAS/SAN Storage | Hosts: $250K  
Arrays: $0.6M | Hosts: $250K  
Arrays: $60K  
($20K x 3) | Assume 80% of total costs | Grand Total: $1.86M  
(Savings: 75% Lower TCO over 5-year period) |

**Conclusion: Rethink Your Cloud Storage Architecture**

Virtualization is fast becoming a staple in enterprise-class computing architectures. Powered by the growth of public-facing managed hosting service providers, and the continuing need for cost-efficient consolidation in private cloud data centers, it is almost certain that virtualized application servers will be the rule and not the exception in coming years.

But the success of virtualization has ushered in a new problem. The rising number of application VMs is causing a storage scalability challenge that will significantly worsen over time. With powerful host CPUs and 10Gb Ethernet ensuring plentiful processor and network resources, the biggest challenge for storage administrators during the next decade is how to incrementally scale storage IOPS in order to support high-density VM architectures, while staying within fixed IT budgets.

Marvell DragonFly offers a cost-effective, turn-key solution that seamlessly fits into existing storage architectures. As the industry’s first unified host cache, DragonFly provides up to 10x write/read IOPS acceleration and application latency reduction and 50 percent or greater reduction in I/O load to NAS/SAN arrays. This immediately solves the virtualization scaling challenge for public and enterprise cloud computing customers. DragonFly uniquely executes sophisticated read and write-back caching technology directly onto scalable tiered memory (NVRAM and SSD) to create an intelligent host write/read cache. As a result, application servers access data from a high-performance, low-latency...
I/O cache directly on the host and shared storage can be cost-optimized as a capacity-optimized storage tier.

For IT storage and virtualization architects looking to develop a highly scalable virtualization architecture, Marvell DragonFly platform offers a quick and easy solution that fits seamlessly into standard rack-mount servers. And since DragonFly allows the company to defer expensive storage capital expenditures by incrementally scaling virtual infrastructure on existing NAS/SAN hardware, it becomes easy to justify budget to the CIO or CFO as immediate payback on investment.

For questions/comments on Marvell DragonFly, please contact your local Marvell representative or visit www.marvell.com for more information. For additional information, please refer to the following useful papers:

- Marvell DragonFly Product Brief
- Marvell DragonFly Platform Overview Paper
- Marvell DragonFly Performance Benchmarks

Footnotes/Sources:

3. Dell PowerEdge R710 configured with Dual Intel® Xeon® E5530, 2.4Ghz, 8M Cache, Turbo, HT, 1066MHz Max Mem, 48GB Memory (6x8GB), 1333MHz Dual Ranked RDIMMs for 2 Processors, 1x 250GB 7.2K RPM SATA 3.5-inch Hot Plug Hard Drive, Citrix XenServer 5.6 Free Version, totaling approximately $4,580 on Dell.com as of February 2011.
5. Estimate each HBA to be approximately equivalent in cost to a higher-end fibre channel adapter. Call Marvell for latest volume pricing.

About the Author:
Shawn Kung
Director of Product Marketing, Marvell

Shawn Kung is director of product marketing at Marvell. He is responsible for enterprise and consumer storage products, including the creation of innovative solutions in solid-state storage technology.

With more than 12 years of Silicon Valley marketing and product management experience, Kung has spearheaded the product vision for several award-winning products in data storage and data management. Prior to Marvell, he was senior director of product management at Aster Data Systems, where his vision for integrating MapReduce with MPP clustered databases led the company from an early-stage 9-engineer startup to a $300M acquisition by Teradata.

Prior to Aster Data Systems, Kung held senior product management positions in core systems at NetApp and enterprise applications at Oracle Corporation.

Kung received his Bachelor of Arts degree from the Woodrow Wilson School at Princeton University and a Master of Science degree in Management Science and Engineering from Stanford University, where he was a Valentine Fellow.