NextIO Enables Multivendor Converged Datacenters

"Best-in-Class" Like the Cloud Giants

Executive Summary

Industry standard x86 server node performance has been improving at an increased pace as core counts and memory capacities have increased. This rapidly increasing compute density is driving a commensurate increase in the costs of implementing in-rack network architectures:

- From 1Gbps Ethernet adapters to much more expensive10Gbps adapters as per-node processing improves and likewise, more Fibre Channel (FC) capacity as storage requirements grow and become more complex
- Increasing cabling costs and cabling complexity as server node count increases and bandwidth requirements per node increase
- Adding additional software layers to known operational software stacks (notably Software Defined Networking – SDN)

Large infrastructure vendors are using this increasing complexity as a means to move customers away from best-in-class rack-level components to vertically integrated racks. Converged and adaptive infrastructures are the new vendor-proprietary lock-in.

The challenge is how to maintain a multivendor, best-in-class datacenter. One good solution is to consolidate first level network and storage resources in a manner that is transparent to existing software stacks. NextIO's vNET I/O virtualization and consolidation appliances are a well-positioned, practical solution to this challenge.

Converged Server Infrastructure

A funny thing happened on the way to cloud infrastructure: server nodes became a commodity. And not only did the compute capacity those nodes represent become a commodity, storage has become a commodity as well.

Today the datacenter infrastructure industry uses the word "converged" in several ways. We use the word "converged" to mean both the ability to share a set of common hardware among multiple applications and the packaging of a complete IT solution that spans multiple applications.

We also define a "server node" as the processors, memory, and I/O complex required to run a specific workload, including the SMP span – the span of processor cores and their associated caches that enable shared memory among the threads in a given task. For at least the next few years, two socket nodes will account for the majority of converged datacenter server node deployments. However, for many cloud, hosting, and

hyperscale workloads, a single processor socket able to run 16 shared-memory hardware threads (plus associated DRAM and I/O) is more than enough SMP span.

The practical reality is that virtualization, consolidation, and Intel's dominating share of multi-vendor converged datacenter server processor sockets have created not just a level playing field for server nodes; it is almost impossible for server node vendors to differentiate based on Intel Xeon processors and Intel chipsets, fiberglass, and bent metal. X86 server nodes have become a non-differentiated commodity.

Cloud and hyperscale technologies are pushing into converged datacenters under private cloud and hybrid cloud models. Using standard Xeon processors, at a rack level the "server nodes per U" count is escalating and with it the number of Ethernet NICs and switches and FC infrastructure. This is partly because the traditional enterprise server blades model of shared I/O resources has run its course, mostly due to challenges with integrating proprietary blade chassis management frameworks into top-level IT management systems and that the ROI promises never really came to fruition. It is also driven by innovative new hyperscale hardware solutions that pack NIC-equipped server nodes into even denser configurations.

While blade chassis pay a manageability price for sharing network resources among blades in a chassis, scale-up approaches seek to return to larger SMP footprints via bigger, hotter, heavily virtualized server nodes. They are running into diminishing returns as the number of hardware threads being shipped in single sockets continues to climb. A CIO needs a lot of justification for an 8P x86 server chassis with 16+ hardware threads per socket, where just ten years ago that 8P x86 box jumped from 8 hardware threads *in total* to 16 with Intel's introduction of Hyper-Threading.

The net effect of these trends is that Ethernet cabling and switching and FC storage bandwidth are becoming a substantial planning, execution, and cost challenge for IT shops.

SDN (software defined networking) is a stopgap measure to address in-rack cable proliferation and rewiring complexity, but it does nothing to alleviate the root cause of in-rack network complexity – the rapidly increasing number of server nodes and storage housed in one rack.

For converged datacenters, the major infrastructure vendors are assembling converged or adaptive infrastructure stories around tuned rack-level performance and interoperability. These are the last bastion of proprietary "better together" sales stories, and like blades they are built around a proprietary core solution. Unlike blades, they are not built around a physical solution – they are built around SDNs.

Pain Points

In-rack network hierarchies capable of supporting increasing compute and storage density, with attendant network cost, power consumption, and cabling complexity, are real problems and they are getting worse.

Modular x86 servers, in-rack switches, and NAS and SAN storage enabled IT to mix and match best-in-class solutions from different vendors for over a decade. Converged infrastructures rely heavily on Ethernet as the "translation layer" for all other protocols. This creates an additional, new management challenge for customers. Instead of managing, for instance, two networks, they now only manage one, but the task of managing the traffic that previously used independent cabling (to ensure priority, quality of service, etc.) over a single converged cable adds complexity.

Vendors are attempting to use SDNs to respond with rack-level proprietary solutions. *We view SDNs as an unnecessary introduction of complexity and inefficiency when used under a top-of-rack (TOR) switch.*

There are two fundamental types of SDN – host-centric (also called "overlay") and network-centric. Network-centric SDNs are implemented at the TOR and above and therefore good matches for promoting innovation in new in-rack switching technologies.

For overlay SDN, a local server thread provides the processing power to run the network processing for local virtual machines. A core assumption for overlay SDN technology is that each server node has "spare" processing power to devote to making the overlay SDN transparent to the rest of the datacenter software stack.

Overlay SDN presents a trade-off in consolidated, heavily virtualized, multi-tenant datacenters, in that the object of consolidation is to boost the utilization of each server node – which by definition reduces the amount of unused compute headroom on each node. In addition, if an overlay is not optimized for the underlying network infrastructure for specific deployments, additional inefficiencies surface. IT managers buy some flexibility at the expense of introducing another layer of software, which siphons off incremental performance while requiring additional certification for converged datacenter software stacks, licensing, and support costs.

A majority of converged datacenters are still maintaining separate end of row networks for risk management. This means that they are only converged in the rack, at the top of the rack the single converged network splits out into separate networks. This is the worst of all worlds; it creates a short-term bottleneck in the rack and then maintains redundant networks between racks and the datacenter backbone network, so there are few opportunities for cost savings.

For datacenter workloads that do not require multi-tenancy, particularly those run by hyperscale SaaS infrastructure owners (web front-ends, analytic and Big Data back ends, no-SQL databases, and many workloads in the middle), overlay SDNs do not provide an operational ROI for their performance overhead and additional complexity.

Giving virtual instances access to real or virtual storage and network resources can be much more straightforward via a combination of scheduler intelligence and smart in-rack network topologies.

A few emerging solutions, epitomized today by companies like SeaMicro and Calxeda, are addressing this challenge by pulling the rack-level network topology into a local chassis fabric. While that approach may have long-term technical merit, in the short-term the root OS and hypervisor must be closely tied to the chassis fabric architecture, effectively creating a proprietary OS distribution for that chassis. The only way to re-provision such a server with a different OS is by asking the chassis vendor to write drivers and qualify another OS distribution for their fabric. This type of solution is optimized for new hyperscale datacenter build-out, but has challenges in typical enterprise converged datacenters.

Enterprise datacenters implementing multi-vendor converged solution stacks and edging into private cloud deployments have a much more stringent set of requirements. For these customers, convergence solutions for managing the costs and complexities of increasing server node density within a rack should have the following attributes:

- Share network and storage I/O resources transparently to application workloads
- Use standard OS and hypervisor distributions
- Use standard management and scheduling frameworks

NextIO Solutions

NextIO's technology strategy is deceptively simple: replace the lowest level Ethernet and Fibre Channel networking switches with a switched PCI-Express (PCIe) fabric. Replacing Ethernet NICs and Fibre Channel NBAs at each server chassis with PCIe adaptors and the in-rack Ethernet switch with a PCIe switch might seem like an even trade, but it enables the above set of critical convergence attributes with a completely different set of optimization trade-offs.

The first is subtle but valuable. PCIe is transparent to the host OS and hypervisor running on every server node. There are no changes to the current software stack. OS, hypervisor, workloads and therefore everything works:

- A datacenter can retain its current policies for re-imaging a server to update the OS or switch to another.
- Each node believes it has access to dedicated network and storage resources. There is no workload code porting, because workloads run with no modification.

NextIO's vNET technology enables all of the server chassis attached to a vNET I/O Maestro PCIe switch to share expensive NIC and HBA resources as needed. It pares a server chassis' NIC and HBA cables down to a single PCIe cable with twice a 10Gbps Ethernet cable's bandwidth. PCIe inherently understands the difference between Ethernet and FC traffic and can route each accordingly without any intermediate protocol conversions. In addition, by routing 10GbE from the vNET switch directly to the

end of row (EOR) switch, top of rack (TOR) switches can be eliminated, further simplifying datacenter network topologies.

vNET switches have transparent, real-time management and configuration control. Instead of imposing more complexity on the network architecture with SDN, for example, virtual I/O resources and quality of service (QoS) settings can be reallocated on-the-fly to accommodate changing workloads without touching the rest of the datacenter's software stack.

We took the photos below during the second annual Dell World event, in December 2012. NextIO's rack funneled 80% of Dell World's event traffic through a single NextIO vNET I/O Maestro switch. The rack implemented both a TOR switch and a vNET switch for some very unusual event requirements – real world datacenter deployments would not implement this configuration. Servers connect to the back of the vNET chassis via PCIe; in front are the Ethernet and Fibre Channel ports. Notice the significant difference in cabling complexity in the back of the NextIO vNET I/O rack – even with the nominal TOR cabling.

Photo 1: NextIO vNET I/O Maestro PCIe switch below top of rack Ethernet switch

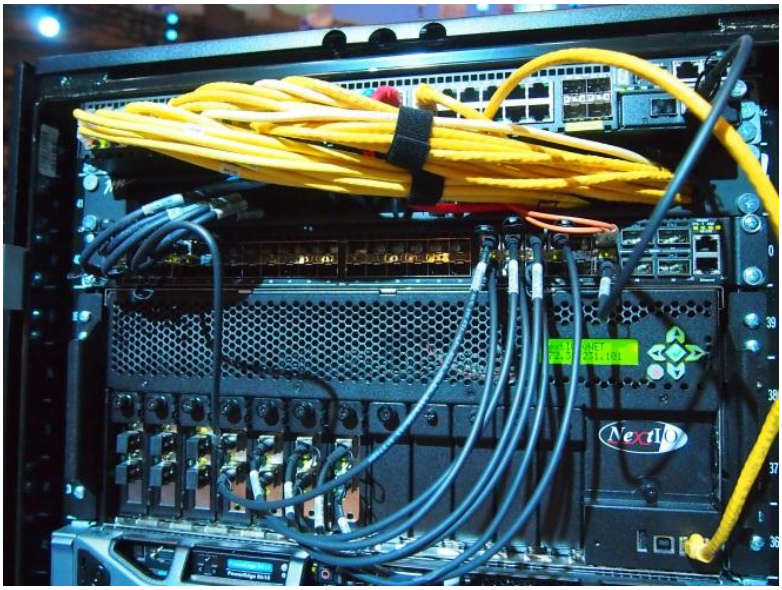

Photo 2 – Left: back of traditional rack, bottom half
Photo 3 – Center: front of Dell full rack with NextIO vNET I/O Maestro PCIe switch
Photo 4 – Right: back of Dell full rack with NextIO vNET I/O Maestro PCIe switch

Is NextIO's PCIe switching strategy a bet against Ethernet?

In a word, "no." Ethernet assumes that a network is a noisy, lossy environment for data traffic between physically distant (many meters) network nodes. Those assumptions are out of date for modern rack-level datacenter architectures. NextIO's PCIe-based adaptors and switching technology prevent systems architects from having to stretch Ethernet's capabilities beyond their intended usage models; hierarchical tree-structured Ethernet topologies are not optimized for addressing massively interconnected local architectures. In fact, from the internal switching inside of the vNET, all server-to server communication over Ethernet is handled within the vNET, creating better overall Ethernet efficiencies.

PCIe already permeates PC and x86 server hardware architectures. It is a stable set of mass-market standards that better match new high-density datacenter requirements.

Conclusion

How can multi-vendor converged datacenters buy into many of the advantages that hyperscale enables and at the same time do so without a high risk "rip and replace" of servers, network, storage, and infrastructure software? Hyperscale datacenter rack-level fabric competition is starting to get interesting. Some products seem viable for a small set of related workloads operating at-scale, others have not yet been tested in production systems for long enough to say which workloads they are optimized for. But hyperscale datacenters are a fundamentally different market than converged infrastructure for enterprise.

NextIO created a best-of-class hybrid solution that uses PCIe for in-rack switching and then hands off to Ethernet and FC between physically distant racks and rows of racks.

For public and hybrid cloud build-outs, NextIO fills a high-value gap. CIOs can buy/retain server hardware and management infrastructure they know and trust, and leverage all of their current software investments, while also reducing CAPEX, simplifying their network topology, and eliminating redundant and underutilized in-rack network capacity. NextIO's vNET technology replaces in-rack Ethernet switches with a flexible rack-level architecture better suited for modern datacenter workloads.

Author
Paul Teich, Analyst at Moor Insights & Strategy

Editor
Patrick Moorhead, President & Principal Analyst at Moor Insights & Strategy

Inquiries
Please contact us at the email address above if you would like to discuss this report and Moor Insights & Strategy will promptly respond.

Licensing

Disclosures
Moor Insights & Strategy has a consulting relationship with NextIO. This paper was commissioned by NextIO. No employees at the firm hold any equity positions with NextIO.

DISCLAIMER
The information presented in this document is for informational purposes only and may contain technical inaccuracies, omissions and typographical errors.