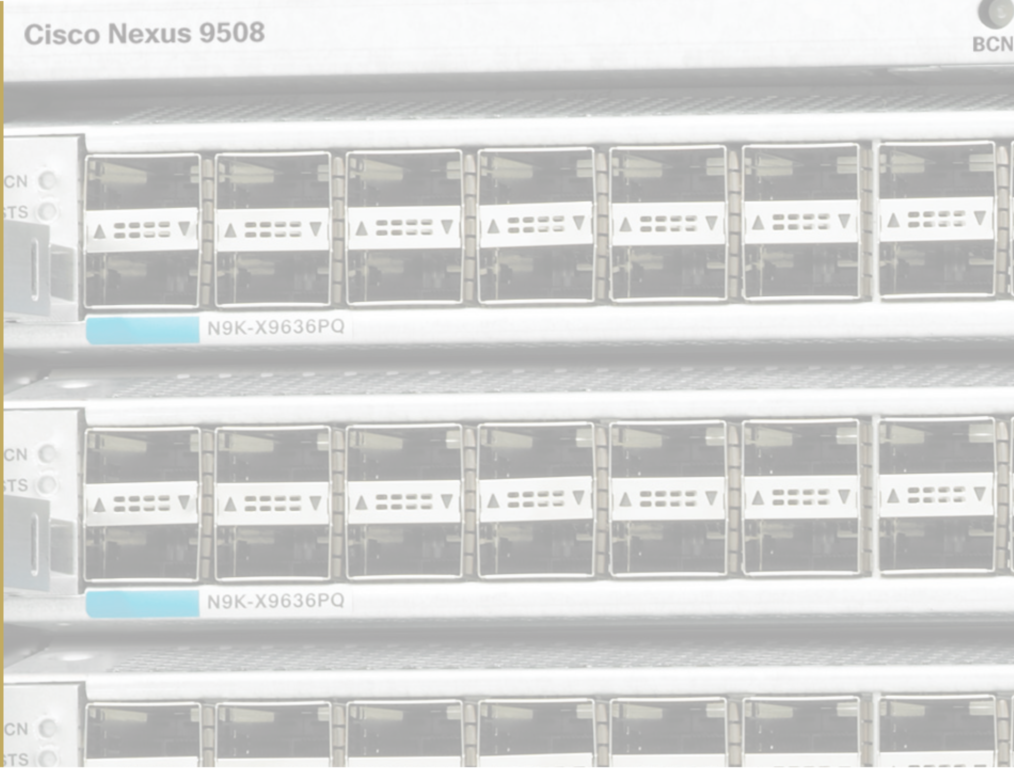
THE LIPPIS REPORT

The Resource for Network/IT Business Decision Makers



Why the Nexus 9000 Switching Series Offers the Highest Availability and Reliability Measured in MTBF

Lippis Report Research Note
November, 2013

At the Open Networking User Group (ONUG) this past October 29th and 30th, hosted by JPMorgan Chase in New York City and attended by over 500 influential IT and industry business leaders, virtualized networking overlays and the need for high performance plus reliable underlays was top of mind. With 2014 and 2015 being the years of open networking pilots and deployments, respectively, much focus was on the upcoming announcements of high-density 40GbE spine switches from Cisco's Nexus 9000 series, Arista's 7500E and HP's 11900/12900. At ONUG, IT leaders expressed multiple drivers for high-density 40GbE spine switches, including connecting thousands, tens of thousands and hundreds of thousands of servers at 10GbE, support for IP storage, supporting high-growth virtual connectivity via network overlays, all over a high-performance and highly reliable underlay.

These requirements mean that spine switches need to offer high resiliency and predictable throughput and latency plus flexible Layer 2 and 3 workload connectivity, independent of physical infrastructure. To address these requirements, this new generation of high-density 40GbE switching needs to offer high Mean Time Between Failure (MTBF), often measured in terms of the number of nines in their availability or time in hours between failures. Availability will be an increasing measurement for spine switches, as they support an ever-increasing workload of virtualized and non-virtualized applications plus IP storage while providing programmable networking features. In fact, this new generation of modular spine switches looks more like servers than traditional switches as they expose their Linux operating system offering automation via programmability.

Merchant or Custom Silicon

Today's networking market has given rise of merchant silicon vendors, such as Broadcom, Intel, Marvell, EZchip, Netronome and others, but Broadcom dominates the market. Every 18 to 24 months, a new generation of merchant silicon enters the market that offers twice as many ports, lowers forwarding latency by nearly 50%, reduces latency jitter to within 10 nanoseconds, lowers power consumption and keeps cost per chip I/O relatively the same. In Broadcom's case, its new Trident II ASIC—scheduled to be widely available in December of 2013—boasts forwarding latency in the sub 5-microsecond range that's also deterministic to within a few nanoseconds. Network switches built with the Trident II's performance at 40GbE speeds will challenge Fiber Channel Storage Area Networking, or SAN, attributes by offering high speed, low and deterministic latency.

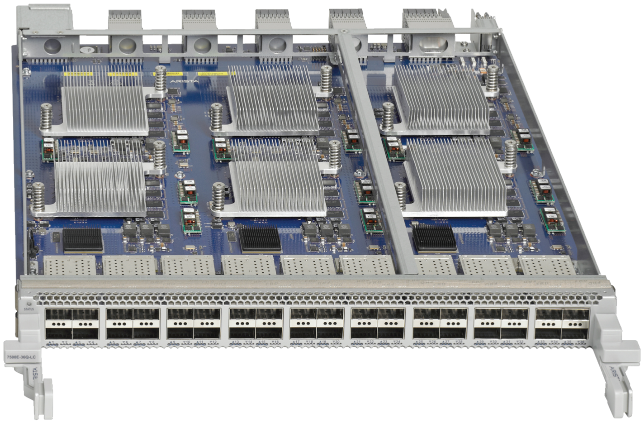
But there is a downside to Broadcom's network ASIC (Application-Specific Integrated Circuit) market share in that an entire industry is held hostage to its innovation cycle. The Trident II ASIC is two to three quarters late, holding back the production of merchant silicon-based leaf and spine switches. In addition, merchant silicon vendors are two steps removed from IT executives and their requirements. As such, merchant silicon vendors have focused primarily on forwarding and power consumption advancements but not on network services. This reality has created an innovation distribution or gap between merchant silicon and network switch vendors, and has given rise to a "hybrid" merchant plus custom ASIC switching architecture.

The new generation of hybrid switching leverages the forwarding and power efficiency gains afforded by merchant silicon and higher value network services of custom ASIC. For example, merchant silicon is based upon 40 nanometer (nm) design while custom ASIC is at 28 nm, allowing custom ASIC to pack much more into the same footprint, which increases network services offered plus lowers board real estate. Most importantly, the hybrid approach offers the best MTBF, as detailed below, which equates to higher availability and reliability just when the market is calling for this in data center underlay network design.

Three Reference Designs to 288 40BgE Spine Switching

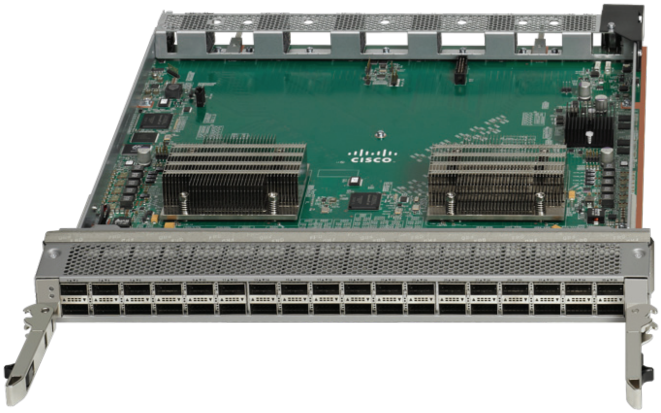
By the end of 2013, there will be three or four spine modular switches that offer eight slots capable of supporting 288 ports of 40GbE, thanks to line card modules equipped with 36 40GbE fitting into each of the eight available chassis slots. Of these four spine switches, there are three basic designs with inherent attributes that stem from the merchant silicon reference design engineers use to build their spine switch. Here are the three basis reference designs:

1) Broadcom Dune (Dune Arad + FE1600 Based Fabric Card): This reference design utilizes the Dune ASICs acquired by Broadcom in late 2009. This design is the most popular of the new 288-40GbE spine switches as many did not want to wait for the Trident II ASIC. This reference design offers line cards with six Dune Arad ASICs plus external buffer architecture based upon 96 DDR3, or Double Data Rate type Three, dynamic RAM memory chips (see figure). The fabric cards are based upon two Dune FE 1600 ASICs. Note that in this design, only 30 out of the 36 40GbE ports per line card is line rate at small packet sizes.



2) Broadcom Trident II: As of this writing, Cisco's Nexus 9000 is the only 288-40GbE spine switch to be based upon the Broadcom Trident II reference design. One of the key benefits of the Trident II is that there are only three ASICs per line card thus offering integrated buffering, reducing the number of chips and real estate used per line card. The fabric cards are engineered with two Broadcom Trident II ASICs too. In addition, all 36 40GbE ports run at line rate at all packet sizes as measured by the Lippis/Ixia test of the Nexus 9000.

3) Hybrid Broadcom Trident II + Custom ASIC: This is the approach used by Cisco/Insieme engineers for the Nexus 9000. This design utilizes two Broadcom Trident II ASICs to engineer the fabric cards plus two Cisco developed custom ASICs within line cards (see figure). From a silicon density point of view, the Nexus 9000 line cards are full featured, line rate forwarding at all packet sizes at 36 40GbE with only two ASIC and integrated buffering.



The above reference designs expose three key differences: 1) performance or line rate forwarding, 2) integrated or separate buffering chips, and 3) the number of ASICs to deliver forwarding plus buffering which impacts MTBF. Based upon a MTBF analysis of the three approaches to 288-40GbE spine switch design, the Broadcom Trident II and Hybrid approaches offer much greater availability than the Dune approach, thanks to the low number of ASICs inherent in the designs. The table shows the number of components required for each spine switch line card reference design. Fewer components for the same or greater functionality is better.

Number of Components for Each Spine Line Card Reference Design
Less Components for the Same or More Operation Is Better

Components	Dune Line Card	Trident II Line Card	Cisco Custom ASIC Line Card
Capacitors	7990	4083	2640
Connectors	18	22	24
Diodes	85	96	97
IC - MPU	3	1	1
IC-Linear	84	62	54
IC-Logic	48	12	18
IC-Mem	103	15	11
Inductors	454	332	157
Misc.	39	29	29
Optics	2	2	2
Resistors	4218	491	524
Transistors	25	19	17
TOTAL:	13067	5164	3574

In addition to the line card component analysis, a detailed analysis of the above three reference designs that counted every capacitor, connector, diode, integrated circuit, inductor, resistor, transistor, fans, etc., was conducted to calculate MTBF. Included in the analysis are all the components of a fully loaded chassis and its eight line cards to deliver 288 40GbE.

For the hybrid model, the chassis, two supervisors, two system controllers in active-standby redundancy, six fabric cards, eight power sharing units in a 7:8 load sharing configuration and three fan trays in a 2:3 load sharing redundancy configuration was modeled.

The Dune model was based upon one chassis, two supervisors, six fabric cards in a 5:6 load-sharing configuration and four power sharing units in a 3:4 load sharing configuration.

Reference Design	System MTBF in Hours
Hybrid	55,996
Trident II	45,162
Dune	19,981

We found that the hybrid reference design's MTBF was 55,996 hours versus the Dune reference design's MTBF of 19,981. That is, the hybrid design's MTBF attribute is 2.8 times more reliable and available than the Dune-based approach. The all-Trident II design is better than Dune but not as reliable as the hybrid approach. The hybrid approach's MTBF will vary from vendor to vendor, as its results are contingent upon its custom ASIC design. Note that the MTBF analysis is based upon the hardware of a single spine switch and does not include software or a network of spine switches.

Cisco Nexus 9500

The New Distribution of Innovation

There is always a distribution of innovation when an industry transitions to the next ASIC generation that partitions scale and/or performance and features. Merchant silicon vendors are best capable to focus on scale/performance as they are two steps removed from the end customer and as such, don't necessarily understand specific operational problems or market requirements. Switch vendors are only one step away from end customers and, thus, can design-in relevant features faster. Secondary consideration is speed of execution on ASICs as custom ASIC is approximately an 18-month process while merchant silicon ASIC tends to be longer in the 18- to 24-month plus time frame. This allows the hybrid model to move faster on relevant features.

For example, in the hybrid design used in the Nexus 9000, VXLAN (Virtual eXtensible LAN) routing capability is afforded by its custom ASIC. Merchant silicon, in this case the Trident II and Dune ASIC, allows for VXLAN bridging but does not support routing in this generation. The Nexus 9000 is able to use VXLAN, not just as an overlay mechanism but also as a label that can be routed throughout the fabric, carrying application and/or tenant relevant context information. This is an important "must have" network design feature to correlate an application to network flow independent of time and location without using IP addresses or VLAN labels. That is, the Cisco engineers designed VXLAN routing into their custom ASIC while Broadcom is used to forward L2 or L3 flows. In essence, Cisco is stitching the overlay and underlay seamlessly together, and provides full visibility throughout the fabric.

As data centers become denser in the number of servers they connect, from tens of thousands to hundreds of thousands, and at the same time, transition server connectivity from 1GbE-to-10GbE, 40GbE demand in the spine will grow significantly. Price points will also drive this transition as 10GbE leaf switches will drop to below $200 to $300 per port while 40GbE in the spine will be $600 to $800 per port during 2014. In addition to 10GbE server connectivity driving 40GbE spine switches, new network features to support the exponential growth of network overlays will become paramount, including VXLAN termination, bridging and routing. A large number of new IP storage firms have emerged over the past 18 months to address key-value, distributed, big data related efforts and new reactive application architectures, which are driving the industry to scale out IP storage. As a result, the underlay network is including features to support IP storage as a converged fabric approach.

These trends, and more, require that the underlay be not only faster and deterministic, but also non-blocking, highly available and reliable. The hybrid model to next generation modular switching at high-density 40GbE offers the best of both worlds, as its leverages the forwarding and power consumption efficiency afforded by merchant silicon vendors and feature acceleration to address new network designs and capabilities via custom ASIC. As this market accelerates through one of the fastest changing periods in the networking industry, the hybrid's inherent distribution of innovation is a must to keep up and be competitive. While Cisco's Nexus 9000 is the only hybrid approach on the market today, we fully expect more to come over the next six to 12 months; they are simply ahead of the market, and leading.

About Nick Lippis

Nicholas J. Lippis III is a world-renowned authority on advanced IP networks, communications and their benefits to business objectives. He is the publisher of the Lippis Report, a resource for network and IT business decision makers to which over 35,000 executive IT business leaders subscribe. Its Lippis Report podcasts have been downloaded over 200,000 times; ITunes reports that listeners also download the ***Wall Street Journal's*** Money Matters, ***Business Week's*** Climbing the Ladder, ***The Economist*** and The ***Harvard Business Review's*** IdeaCast. He is also the co-founder and conference chair of the Open Networking User Group, which sponsors a bi-annual meeting of over 500 IT business leaders of large enterprises. Mr. Lippis is currently working with clients to design their private and public virtualized data center cloud computing network architectures with open networking technologies to reap maximum business value and outcome.

He has advised numerous Global 2000 firms on network architecture, design, implementation, vendor selection and budgeting, with clients including Barclays Bank, Eastman Kodak Company, Federal Deposit Insurance Corporation (FDIC), Hughes Aerospace, Liberty Mutual, Schering-Plough, Camp Dresser McKee, the state of Alaska, Microsoft, Kaiser Permanente, Sprint, Worldcom, Cisco Systems, Hewlett Packet, IBM, Avaya and many others. He works exclusively with CIOs and their direct reports. Mr. Lippis possesses a unique perspective of market forces and trends occurring within the computer networking industry derived from his experience with both supply- and demand-side clients.

Mr. Lippis received the prestigious Boston University College of Engineering Alumni award for advancing the profession. He has been named one of the top 40 most powerful and influential people in the networking industry by ***Network World. TechTarget***, an industry on-line publication, has named him a network design guru while ***Network Computing Magazine*** has called him a star IT guru.

Mr. Lippis founded Strategic Networks Consulting, Inc., a well-respected and influential computer networking industry-consulting concern, which was purchased by Softbank/Ziff-Davis in 1996. He is a frequent keynote speaker at industry events and is widely quoted in the business and industry press. He serves on the Dean of Boston University's College of Engineering Board of Advisors as well as many start-up venture firms' advisory boards. He delivered the commencement speech to Boston University College of Engineering graduates in 2007. Mr. Lippis received his Bachelor of Science in Electrical Engineering and his Master of Science in Systems Engineering from Boston University. His Masters' thesis work included selected technical courses and advisors from Massachusetts Institute of Technology on optical communications and computing.