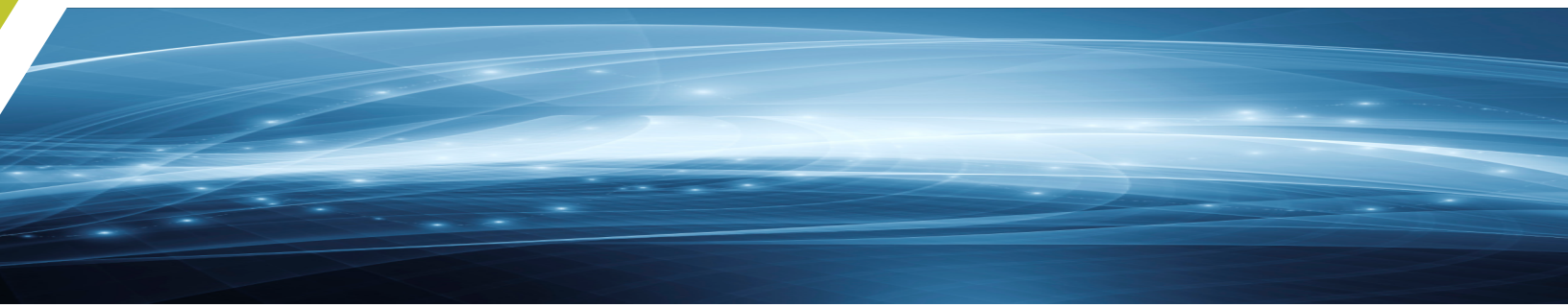

Saving Power in the Modern Data Center

Data consumption is continuing to grow daily, thanks to smartphones, social media, video streaming, big data, and IoT (the 'Internet of Things') applications - all of which require big pipes to cope with the huge quantities of data that are being transferred. In fact, IDC predicts that the sum of the world's data will grow to a 175ZB by 2025[1] .

The growth in high-performance computing and machine-learning cloud, storage and compute requirements has created a demand for larger and stronger data centers, giving rise to more power-consuming network equipment, both in the core and at the edge. To reduce power in the data center, it is important to consider incorporating power-saving mechanisms throughout the fabric.

This paper identifies various power-saving strategies, from deploying less hardware for the same compute needs, through the usage of lower-power interconnect elements, to enabling power-saving mechanisms at the network components level. The Mellanox power-efficient products and solutions described in this paper can significantly reduce power consumption, resulting in higher savings in CAPEX and OPEX.

The 21st Century Data Deluge

The exponential growth of data and the applications that take advantage of real-time massive data processing, data analytics, business intelligence and more, are driving the demand for faster and more efficient interconnect solutions. Higher power consumption together with the maintenance of full connectivity of a network that is typically underutilized, raises both issues of operational costs and carbon footprint, and increases the significance of traditionally overlooked power aspects. In response to these challenges, many of the world's high-performance computing, artificial intelligence, data-intensive and cloud infrastructures leverage InfiniBand's high data throughput, extremely low latency, and smart In-Network Computing acceleration engines to deliver world-leading application performance and scalability.

Power-Savings Using Mellanox® InfiniBand

Saving power in the InfiniBand data-center is very important and can take place on different levels.

Reducing Amount of Hardware—For Given Communication Needs

Whether building or upgrading a data center, significant power savings can be achieved by using less hardware. With Mellanox HDR solutions, one can build a cluster with given compute-communication requirements, but with much fewer switches and cables. Mellanox Quantum®-based HDR (200Gb/s) switches, offer the option that provides the ultimate in scalability and power-savings. It enables data centers to utilize half as many switches and cables as the competition for the same throughput, for example, 1.6X fewer switches and 2X fewer cables to connect a 400-node system. Utilizing two pairs of two lanes per port, Mellanox Quantum switch silicon can support up to 80 ports of 100Gb/s to create the most dense and efficient top-of-rack and modular switches available in the market, and the lowest total cost of ownership for today's data centers and HPC clusters.

1. According to IDC (2018), world-wide data is well on its way to growing from 33 zettabytes in 2018 to 175 zettabytes by 2025.

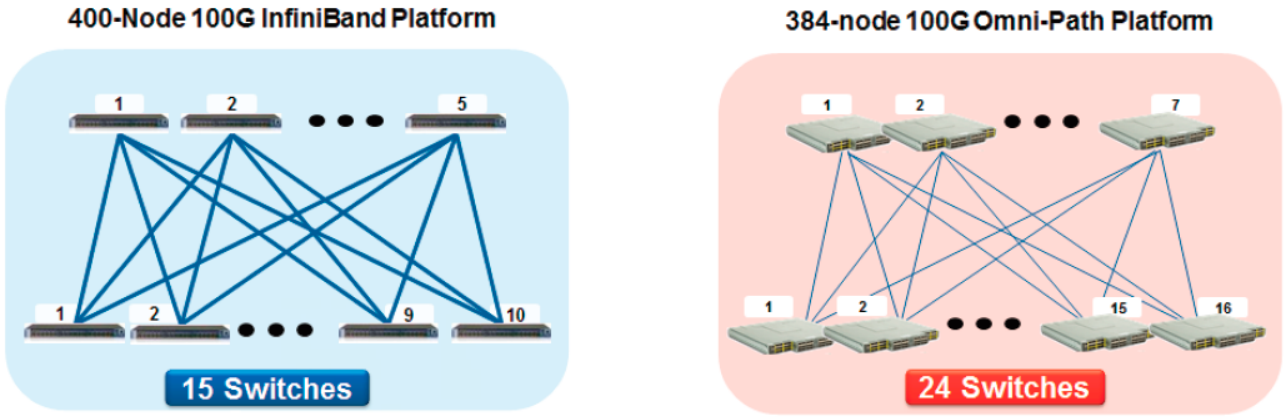

Figure 1. HDR100 Requires 1.6X Fewer Switches and Cables for 400 Nodes

Reducing Amount of Hardware—For Given Compute Needs

It is possible to achieve significant power savings in the modern data center by using less hardware. By deploying Mellanox InfiniBand solutions, customers can buy fewer servers yet achieve the same level of performance as other networks that are comprised of more servers.

Leveraging In-Network Computing

Several factors including higher speed, In-Network Computing based on SHARP™ (Scalable Hierarchical Aggregation and Reduction Protocol), GPUDirect® direct access to GPUs, and many other hardware capabilities, enable customers to offload, and thus accelerate, CPU compute tasks to the network infrastructure. This means that fewer compute nodes are required to run application jobs.

Mellanox Multi-Host® Technology

Mellanox's Multi-Host technology provides high flexibility and major savings when building next generation, scalable high-performance data centers. Reducing the number of switches and cables, Mellanox Multi-host slashes switch port management and power usage such as by reducing the number of cables, HCAs (ConnectX) and switch ports required by several independent servers, from four to one of each.

Connecting multiple compute or storage hosts to a single interconnect adapter, Mellanox Multi-Host separates the adapter PCIe interface into multiple and independent PCIe interfaces with no performance degradation. The technology enables designing and building new scale-out heterogeneous compute and storage racks with direct connectivity between compute elements, storage elements and the network, and better power and performance management, to achieve maximum data processing and data transfers at minimum capital and operational expenses.

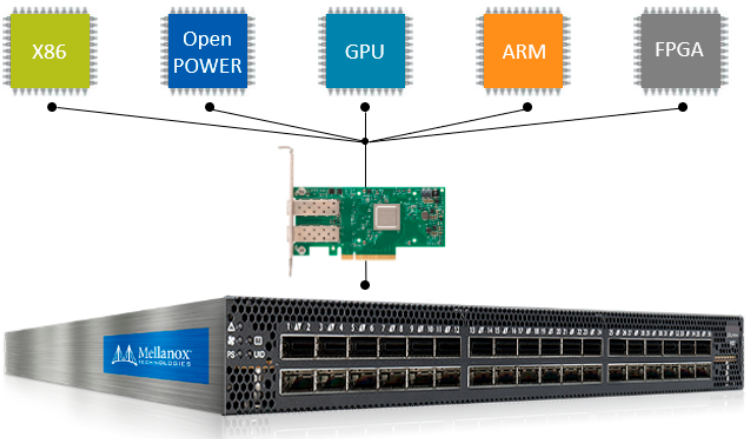

Figure 2. Mellanox Multi-Host Enabling Highest Performance and Scalability for All Compute and Storage Platforms

Mellanox Multi-Host technology also enables IT managers to remotely control the configuration and power state of each host individually, guaranteeing host security and isolation as the management of one host does not affect host traffic performance nor the management of other hosts.

Advanced Topologies

Another method for lowering the total amount of hardware in a cluster is by using advanced topologies, like DragonFly+. With DragonFly+ one can build very large clusters, comprised by interconnections between multiple "islands" of FatTree topology. The interconnections between these islands is achieved within racks of modular switches. By that, the need for long Active Optic Cables (AOCs) to interconnect the islands is eliminated, and power is saved. A good example of DragonFly+ in action is the Niagara system, a partnership between Lenovo, Mellanox and Excelero to deliver Canada's fastest supercomputer based on world's first implementation of DragonFly+ architecture, enabling high levels of compute performance using a reduced number of switches. Click here to learn more.

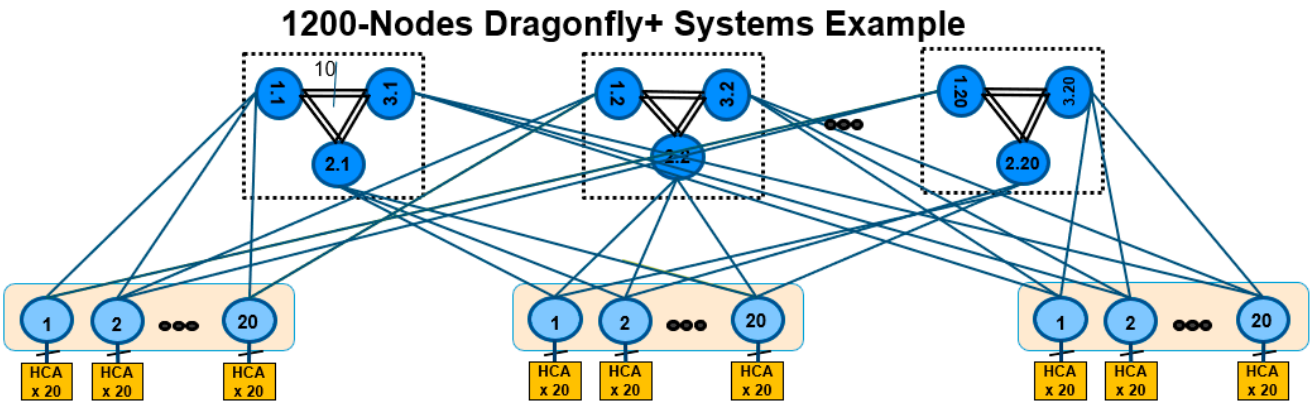

Figure 3. Dragonfly+ topology: reduces total cost of ownership, with fewer long Active Optical Cables

End-to-End InfiniBand

Deploying Mellanox InfiniBand end-to-end solution is another very good method for saving power. Mellanox has optimized the end-to-end signal integrity with lower power. For example, the Mellanox EDR Active Optical Cable (AOC) and SR4 optical module boast a 2.2W power consumption, which is the lowest power rating in the industry.

InfiniBand Interconnect

LinkX AOC

Active Optical Cables (AOC) are widely used in HPCs and have become increasingly popular in hyperscale, enterprise and storage systems as a high-speed, plug & play solution with longer reaches than Direct Attach Copper (DAC) cables. Due to its unique shape, the Mellanox LinkX "H-Cable" HDR AOC enables cross connecting to four InfiniBand switches (or ports) together using a single cable. Reducing the number of layers of InfiniBand switches in large configurations, these H-Cable AOCs enable saving considerable OPEX by enabling a reduction in the number of switches and AOC cables compared to implementing via 100G EDR AOCs, and freeing up four expensive switch ports for other uses.

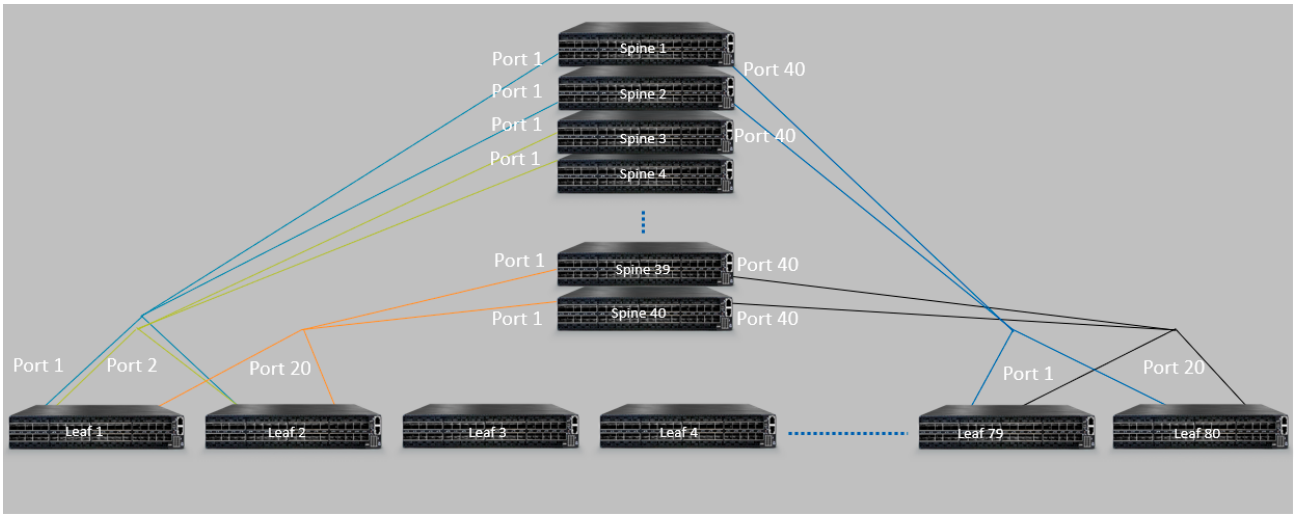

Figure 4. H-cable approach: reduces power and cabling costs as well as the expense of switch ports and transceiver ends.

Additionally, the H-Cable enables reducing the number of actual ports needed to accomplish a cross connect of up to four switches from 8-ports and four AOCs, using individual EDR AOCs to 4-ports and one cable using the new 200G cross connect H-Cable.

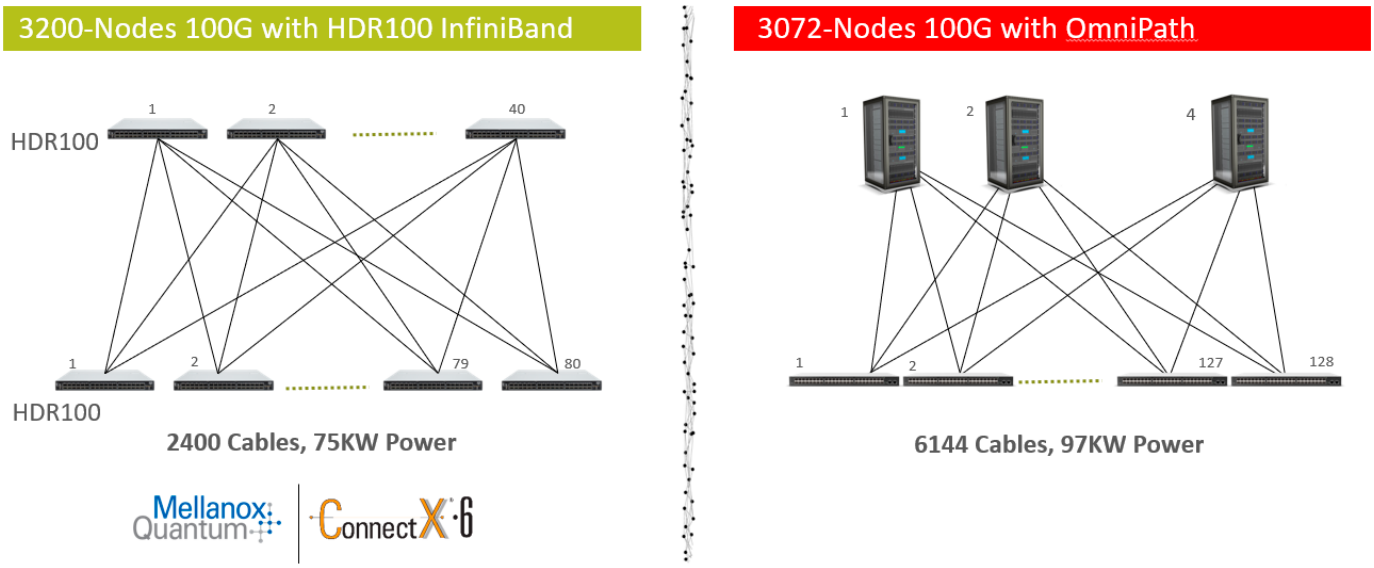

Figure 5. 2X Savings on real estate; 3x savings on cables, and 2x savings on latency

LinkX DAC Cables - Zero Power Consumption

Direct Attach Copper (DAC) cabling is an ideal option for inside data center rack applications to connect InfiniBand top-of-rack switches to network adapters. Apart from their low cost, the other big reason for their enduring popularity is that DAC consumes almost zero power.

With skyrocketing power, every Watt saved at the component level can translate into several Watts of system cooling and related power consumption driving fans, and AC equipment. Now, multiply this savings by tens of thousands of cables and a few dollars saved on each cable on the capital expenditure, or (CapEx) and power consumption operating expenditure (OpEx) and the costs adds up fast! Large data centers spend upwards of $4 million per month on electric bills.

Silicon-Level Power Saving

Switch and HCA silicones are designed to support many combinations of network speeds and protocols. As such, each silicon device contains components that can be shut down, scaled, or optimized for each different protocol currently being deployed in the fabric. For instance, a Phase-Locked Loop (PLL) supporting 40Gb/s Ethernet can be shut down when an HCA is working at InfiniBand protocols. Moreover, datapath units can be turned off and on based on the actual packets that will traverse the device. Mellanox includes dynamic HW smart clock gating, which is aware of cells used for each packet processing, and automatically turns off uny used cells, during mission mode and under traffic activity on the port. Since its HW based mechanism, it is a seamless power optimization, not involving any software layers, which are usually slow to react to actual packet processing at full wire speed."

Creating an Adaptive Fabric

An adaptive fabric, ultimately capable of dynamically shutting down unused network elements and self-optimizing its topology, will increase energy efficiency, decrease CO^2 emissions, and reduce the data center cost of operation. Mellanox dedicates development efforts[1] toward incorporating power-saving features into all fabric levels, starting from the silicon level and through the Host Channel Adapter (HCA), switch port, and switch systems.

System-level Power Savings

Fan Power Savings

Reducing the system fan speed is an effective way to save power. When the switch operates at low capacity, the temperature does not require full fan Revolutions Per Minute (RPM). Mellanox is developing smart algorithms to optimize power consumption of fans. Figure 6 shows the reduction of power consumption when reducing RPM.

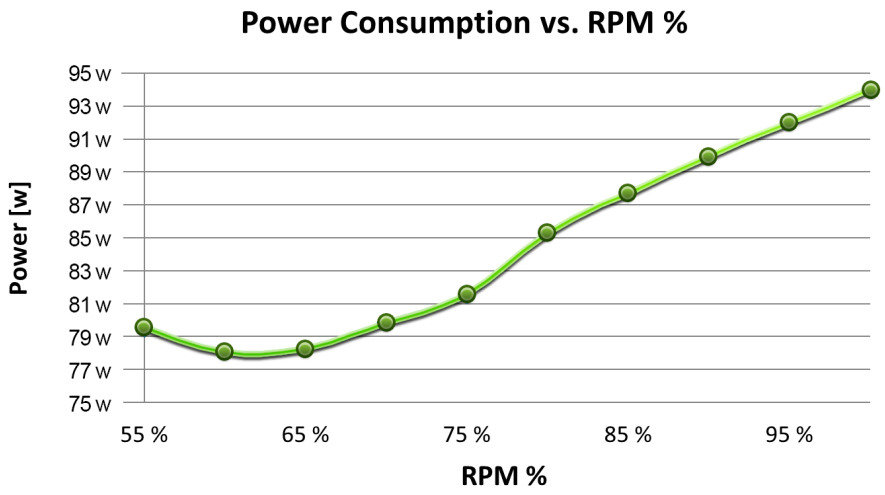

Figure 6. Reducing power consumption when reducing RPM

Conclusion

As the data center's power consumption and carbon footprint become increasingly critical, new power-saving methods are required. Mellanox InfiniBand solutions bring a comprehensive and holistic power management approach throughout all fabric layers. The multiple power-saving methods implemented in the fabric components help reduce TCO and lower the carbon footprint of the data center network.

1. For specific features availability, please refer to the products' Release Notes.

350 Oakmead Parkway, Suite 100, Sunnyvale, CA 94085
Tel: 408-970-3400 • Fax: 408-970-3403
www.mellanox.com