WHITE PAPER

January, 2013

Power Saving Features in Mellanox Products

In collaboration with the European-Commission ECONET Project

Introduction

The growth in Cloud and Web 2.0 storage and compute requirements in recent years has led to an increase in demand for larger, stronger, and more cost efficient data centers.

As data centers are growing in size, the interconnecting fabric is becoming larger and more complex, resulting in more power consuming network equipment, both in the core and on the edges.

This higher power consumption increases both operational costs and carbon footprint, increasing the significance of traditionally overlooked power aspects. Furthermore, rarely all nodes in the data center are active, requiring full network connectivity to be maintained 24/7.

An adaptive fabric, capable of shutting down currently unused network elements and self-optimizing its topology, will increase energy efficiency, decrease CO_2 emissions, and reduce the data center cost of operation.

This paper introduces the "green" fabric concept, presents the Mellanox power-efficient features under development as part of the European-Commission ECONET project, displays a real-world data center scenario, and outlines additional steps to be taken toward "green" fabrics.

The features described in this paper can reduce power consumption by up to 43%. When summed over a real-world data center scenario, a total reduction of 13% of all network components' power consumption is demonstrated. This reduction can amount to millions of dollars in savings over several years.

The Multi-Layered "Green" Fabric

An adaptive fabric, 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.

In order to provide an effective policy, it is important to incorporate power-saving features into all fabric levels, starting from the silicon level, through the Host Channel Adapter (HCA) and switch port, through switch systems, and finally in the Unified Fabric Manager (UFM®)/Software Defined Networking (SDN) engine.

Silicon-Level Power Saving

Switch and HCA silicones are designed to support many combinations of network speeds and protocols. As a result, 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 10GbE or InfiniBand protocols

Mellanox integrated circuits incorporate several power-saving features:

- Unused SerDes (Serializer/Deserializer) PLLs and clock buffers closure
- Disabled port optimizations
- Closing unused PCI Express lanes (x16 -> x8 -> x4) while unsupported by board
- Scaling of core frequency, depending upon protocol used for interconnect
- Scale ASIC voltage supply

These optimizations reduce the power consumption of Mellanox ConnectX®-3 silicon by up to 40%, as seen in Figure 1:

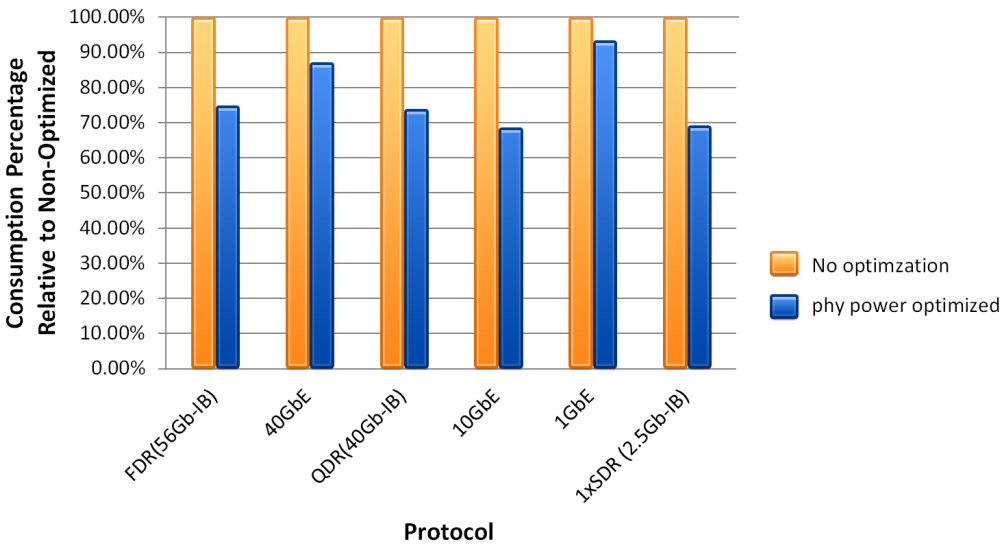

Figure 1. *Reduction of Power in ConnectX-3 IC Due to Silicon Power-Saving Features*

Link-Level Power Saving

When a fabric is underutilized, many of its links are up, yet operate at a fraction of their maximum throughput. Reducing each link's speed (for example, 10GbE -> 1GbE) and width (40GbE -> 10GbE or 4xQDR -> 1xQDR) can considerably reduce power consumption for the entire fabric.

The Mellanox ConnectX-3 HCA and SX6036 Top-of-Rack (ToR) switch can save power by optimizing each port's link width and speed.

These optimizations will be embedded in the HCA and switch hardware and enabled via the firmware. Mellanox is also adding the ability to remotely control and manage these features upon its management software solution – MLNX-OS™ (local device management) – and to the UFM, its fabric-wise central management platform.

One of the main methods in this aspect is the Width Reduction Power Saving (WRPS)[1], enabling a 40Gb/s 4xQDR port running at an effective 10Gb/s rate to shut down three inactive QDR lanes, thereby saving power. Figures 2 and 3 show the percentage of power saved due to WRPS for a SX6036 switch and a ConnectX-3 HCA, respectively.

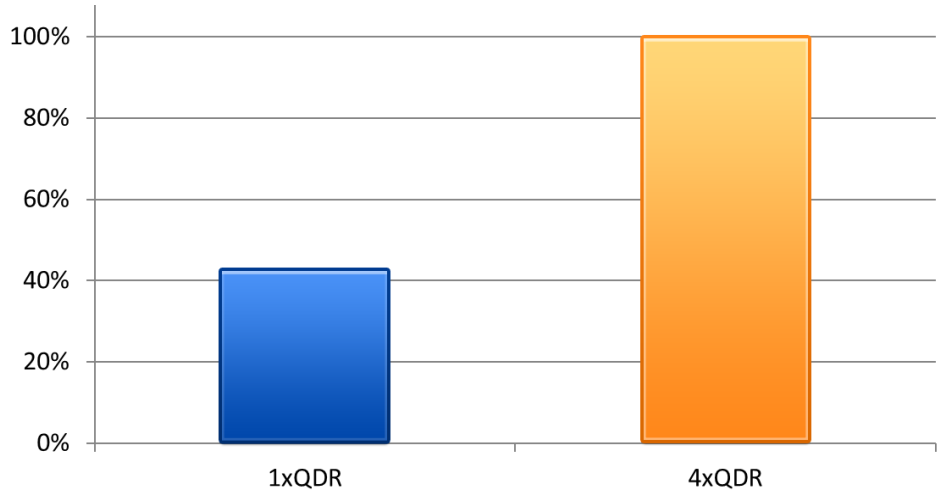

Figure 2. *Width Reduction Power Saving Capability in a Mellanox SX6036 Switch*

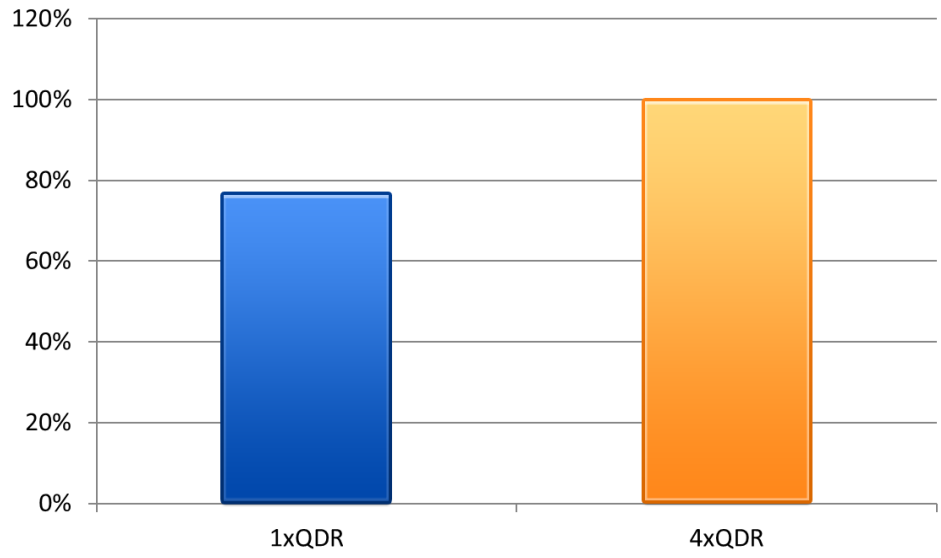

Figure 3. *Width Reduction Power Saving Capability in a Mellanox ConnectX-3 HCA*

System-Level Power Saving

Internal Port and Module Suspend/Shutdown

Large switch systems are comprised of several modules, with many ports each. For instance, the SX6506 InfiniBand director switch allows up to 108 FDR ports, amounting to 12.1Tb/s aggregate switching capacity. In reality, not all ports are active all the time. Some ports are fully utilized, while some are not active at all. This results in a frequently underutilized, yet fully power-consuming switch.

In such a case, some of the internal ports can be shut down without compromising the fabric performance.

The Mellanox system has the ability to set a port suspension policy that enables saving power on ports that are not being used, while quickly bringing up all needed internal ports once the fabric utilization rises. Figure 4 shows a reduction of ~1% from the SX6506 power consumption for every internal switch port that is closed.

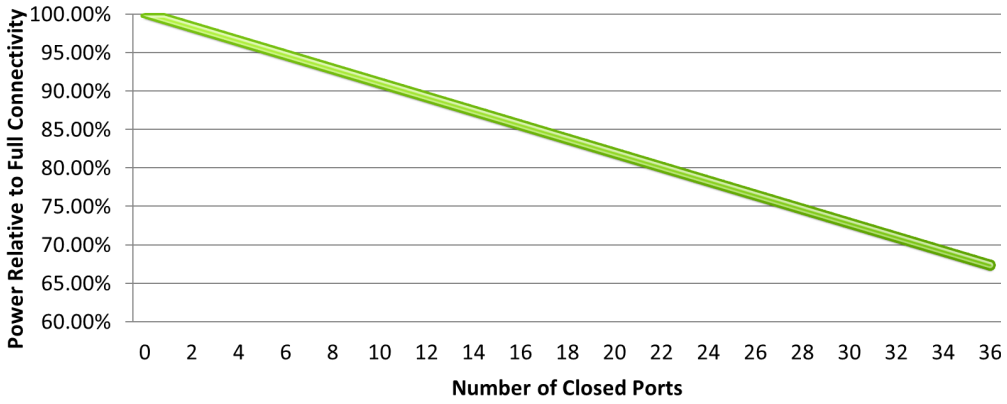

Figure 4. *SX6506 Power Reduction as a Function of Closed Ports*

With enough unutilized internal ports, a large switch can shut down an entire internal switching module, resulting in a 14% power reduction.

Fan Power Savings

Another effective way to save power is to reduce the system fan speed. 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 5 shows the reduction of power consumption when reducing RPM.

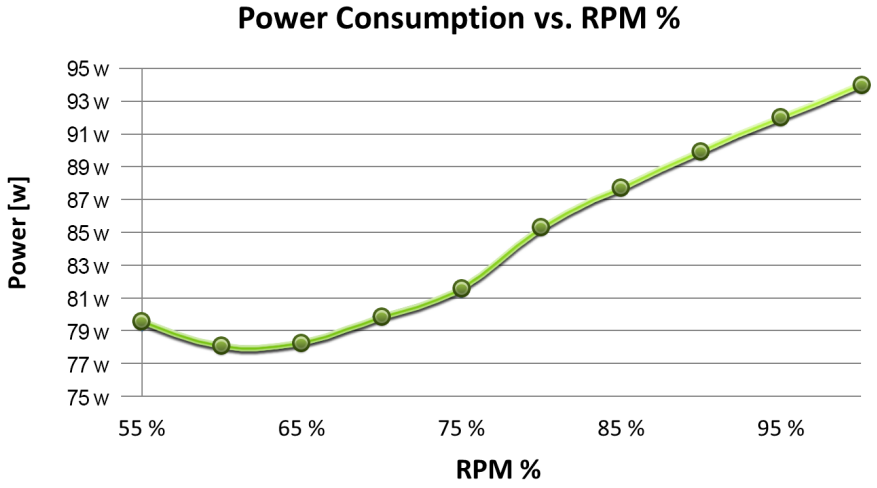

Figure 5. *Power Consumption as a Result of Fan RPM*

Wake-on-LAN

The ConnectX-3 host channel adapter supports the Wake-on-LAN (WOL) standard, allowing it to "sleep" when receiving an appropriate request from the server and quickly awaken the system once a dedicated magic packet arrives, without losing any data. Figure 6 demonstrates that more than 60% of the HCA's power consumption can be saved in comparison to a fully "awake" inactive HCA.

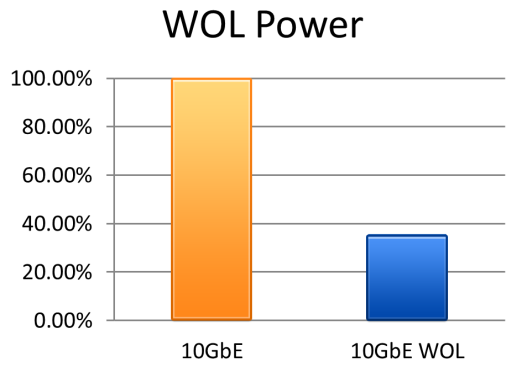

Figure 6. *ConnectX-3 Power Consumption with/without Wake-on-LAN*

Fabric-Wise Power Management

The InfiniBand fabric is centrally managed by the Subnet Manager (SM), with advanced management and monitoring provided by Mellanox's UFM. The UFM is constantly aware of the state of nodes in the fabric, and monitors traffic patterns. Power saving features already implemented in Mellanox systems will be controllable from the UFM in the near future. The UFM will be able to make "green" configurations and optimizations to the fabric: modifying routing tables and allowing shut down of unused switch ports, modules, and even entire switch systems, while keeping fabric integrity and connectivity intact. This can only be achieved due to the centralized topology management of the InfiniBand fabric.

Topology decisions for traditional Ethernet networks are made independently in each node, resulting in a distributed, often very power-inefficient fabric. No element exists in the network that is capable of shutting down ports, nodes, or entire switch systems, while re-engineering the fabric to "route around" the disabled nodes.

SDN calls for unified management of an Ethernet fabric, thereby allowing all of the aforementioned power-related optimizations.

Mellanox switches, silicon, and host channel adapters support both InfiniBand and Ethernet on the same platform. When combined with the future SDN-engine that will be integrated into the UFM, a complete Ethernet/InfiniBand green fabric could be built from Mellanox products, with superior performance, lower TCO (total cost of ownership), and a reduced carbon footprint.

Data Center Example

Figure 7 shows a data center of 1008 servers, interconnected via a 56Gb/s non-blocking FDR InfiniBand fabric. Such a data center is constantly underutilized, running at 10Gb/s most of the time. As a result, the Mellanox HCAs and edge switches can apply the Width Reduction feature, and the core switches can shut down internal ports.

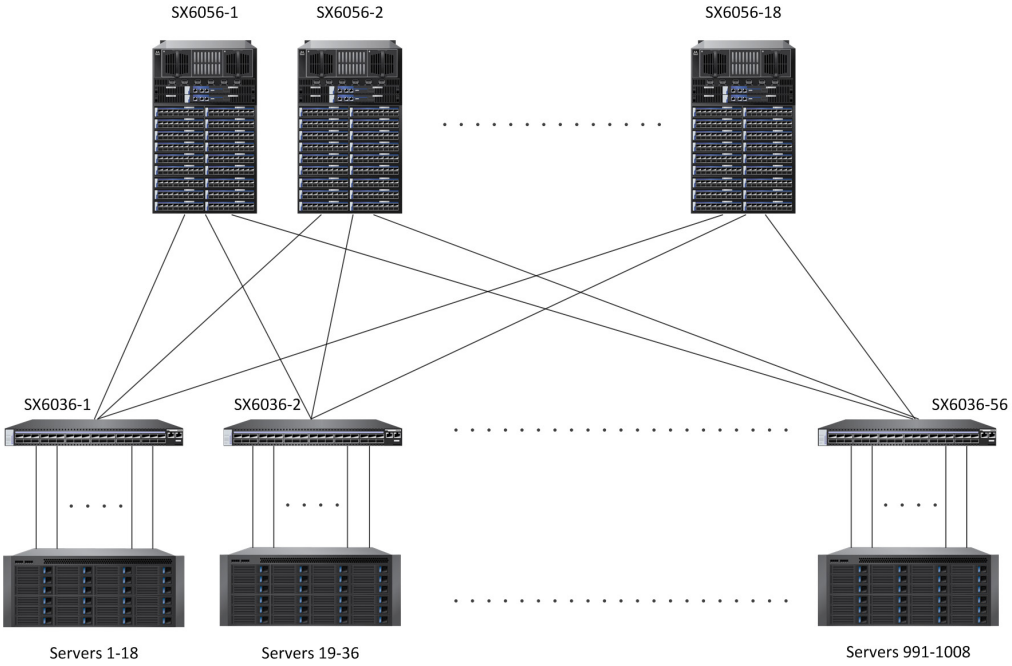

Figure 7. *Data Center with 1008 Servers Connected via a 56Gb/s InfiniBand Fabric. The Fabric Contains 56 Edge Switches and 18 Core Switches.*

Figure 8 depicts savings per component. Power savings features can save up to 57% from a ConnectX-3 silicon's power consumption.

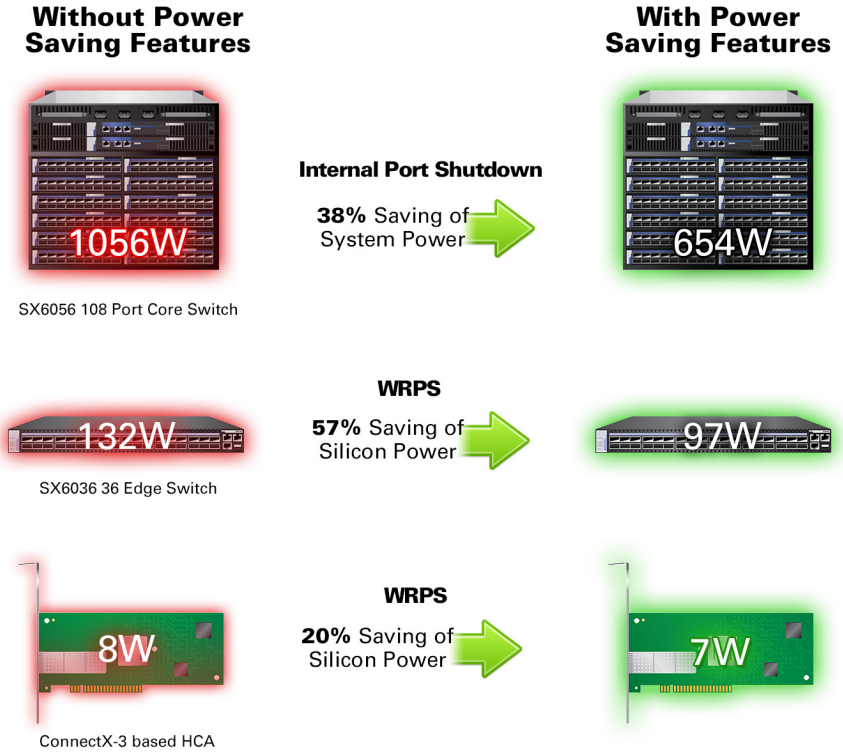

Figure 8. *Power Saving in Each Component of the Data Center*

In our example of 1008 servers (56 edge and 18 core switches), we can achieve $51,000 savings per 3 year cycle, as detailed in the following table:

Regular Data Center Consumption	132.2KW
Green Data Center Consumption	114.5KW
Difference	13%
Electricity Price	$0.11/KWh
TCO Reduction for 3 year cycle	$51,000

Table 1. *Total Power Saving for 1008-Server Data Center over a Three-Year Lifecycle (Calculated Using Average Electricity Prices in the US[2], and a Cooling Factor of 0.7 Added to Total KW)[3]*

Today's large Web 2.0 data centers include hundreds of thousands of servers. For a data center containing 100,000 servers, savings can amount to $5,000,000 over a three-year cycle, which is an ecological savings equivalent to 51 kilotons of CO_2, 1.3 million trees, or 10 thousand cars.

Summary

As the data center's power consumption and carbon footprint become increasingly critical, new power-saving methods are required.

Mellanox "Multi-Tier" Green Fabric brings an innovative 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.

Mellanox, as a member of the European-Commission ECONET project, continues to advance toward eco-friendly networks by developing more innovative power-saving features and implementing the "green" fabric vision.

About ECONET

The ECONET (low Energy COnsumption NETworks) project is a three-year IP project (running from October 2010 to September 2013) co-funded by the European Commission under the Framework Programme 7 (FP7).

The ECONET project aims at studying and exploiting dynamic adaptive technologies (based on standby and performance scaling capabilities) for wired network devices that allow energy to be saved when a device (or part of it) is not in use.

The overall idea is to introduce novel green network-specific paradigms and concepts, enabling the reduction of energy requirements of wired network equipment by 50% in the short- to mid-term (and by 80% in the long run).

More information is available at https://www.econet-project.eu.

Mellanox is an active partner of the ECONET project.

About Mellanox

Mellanox Technologies (NASDAQ: MLNX, TASE: MLNX) is a leading supplier of end-to-end InfiniBand and Ethernet interconnect solutions and services for servers and storage. Mellanox interconnect solutions increase data center efficiency by providing the highest throughput and lowest latency, delivering data faster to applications and unlocking system performance capability. Mellanox offers a choice of fast interconnect products: adapters, switches, software and silicon that accelerate application runtime and maximize business results for a wide range of markets including high performance computing, enterprise data centers, Web 2.0, cloud, storage and financial services.

More information is available at www.mellanox.com.

References

[1] Width Reduction Power Saving technology is protected by U.S. Patent Application #12/686,401

[2] http://www.eia.gov/beta/enerdat/#/topic/7?agg=2,0,1&geo=g&freq=M&start=200101&end=201209&ctype=linechart&maptype=0&rse=0

[3] http://www.apcmedia.com/salestools/VAVR-5TDTEF_R1_EN.pdf

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