SDN switches from Mellanox with the Cumulus operating system

Freely combined

Martin Gerhard Loschwitz

With its Spectrum switches, ConnectX adapter cards and matching cables, Mellanox can provide everything you need to equip computer centers with Ethernet. As the operating system running on the switches is Cumulus Linux, it is easy to integrate the devices into an SDN environment.

Special print from iX released in the issue 12/16

When they hear the name "Mellanox" in the network environment, most administrators first think of InfiniBand: while Mellanox did not invent this process, it is now one of the main suppliers active in the InfiniBand market. The company offers InfiniBand HBAs (Host Bus Adapters) with comprehensive Linux support and the necessary switches.

For some time now, Mellanox has been putting Ethernet products onto a market dominated by Cisco and Juniper and their hardware. This article presents the current Mellanox products in the Ethernet segment: switches with the Spectrum chip set and up to 32 100-GbE ports as well as ConnectX-4 network cards. The latter are interesting for cloud computing in particular because they are capable of VXLAN offloading (Virtual Extensible LAN, an encapsulation of data traffic for virtual machines) and soon probably Open-vSwitch offloading too.

Mellanox offers customers a complete package for the computer center – an "end-to-end" range in fact: switches, network cards and cables with the matching (Q)SFP modules (optical or electric transceivers, also known as miniature GBICs). This is not a unique selling point, but the company includes special technology in its equipment.

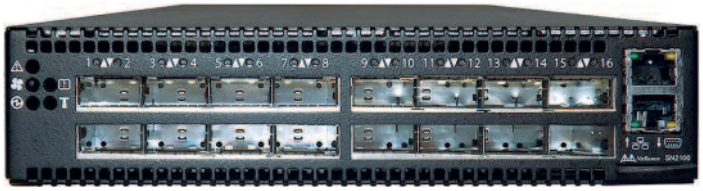
End-to-end range

Mellanox has broken with the standard procedure in the Ethernet environment and does not work as a reseller for chip sets produced by other companies – even though administrators regularly have to do with unlabeled chip sets from large manufacturers such as Broadcom in the "white label" switches in particular. A large number of different devices can be purchased in the market, but they all basically have the same technical restrictions. Instead, Mellanox has developed a chip series of its own for its Ethernet hardware: ASICs (Application Specific Integrated Circuits) under the name of "Spectrum". If necessary, the company also produces customized units upwards of a certain minimum order size.

Technically, the Spectrum chips have a lot of power under their bonnet: they

iX-TRACT

- Switches from Mellanox are equipped with the Spectrum chip set and allow high data rates at a small footprint.
- The Switch Abstraction Interface allows you to install operating systems of your own on the switches. Mellanox relies on Cumulus Linux for this.
- The Ethernet cards of the manufacturer can relieve the system's CPU by taking over specific offloading functions.
- Switch operating system and offloading on the hardware side are interesting in the SDN environment in particular.


Source: Mellanox

The SN2100 is a real space-saver: on half a height unit, it offers a total of 64 25-GbE ports per breakout cable (Fig. 1).


Source: Mellanox

Breakout cables turn one 100-GbE port at the switch into four ports with 25 GbE (Fig. 2).

supply 128 real physical connections (PHY) and allow each port to be connected with up to four lanes each. Each PHY achieves a speed of 25 Gbit/s, i.e. a port has a maximum of 100 Gbit/s. At 128 available 25-Gigabit Ethernet PHYs (GbE), a maximum of 32 100-GbE ports can be used per switch – non-blocking, of course. The manufacturer promises up to 6.4 Tbit/s, so the switch doesn't falter even when under full load. Of course, realistic statements on the capabilities of the Spectrum chip can only be made in the context of the devices in which it is installed.

Three models

Mellanox offers three switch models: the SN2100, which only takes up half a height unit, the top-of-rack model SN2410 and the SN2700. They are all designed to supplement each other in a computer center network.

Space-saver: SN2100

Even at first sight, the SN2100 (Fig. 1) differs considerably from the other switches of the Spectrum series – and from most of the other units on the market. This is because the device only takes up half a 19" height unit (HU). It offers 16 100-GbE ports, with at least one port being required for the uplink to another switch or a router. Using QSFP28 breakout cables (Fig. 2) allows you to double the number of ports if you are happy with 50 Gigabits per second, and to quadruple them at 25 Gbit/s. If you insert two SN2100s into the rack next to each other, you have up to 128 network ports in one height unit – Mellanox markets this product as a "high-density switch".

Beast of burden: SN2410

In contrast to the SN2100, the SN2410 (Fig. 3) is designed as a top-of-rack or "leaf" switch: it offers 48 25-GbE ports and eight additional 100-GbE ports. The concept is simple: the 48 "slow" ports are for cabling within the rack. The 100-GbE ports ensure the connection to the next switch level, the "spine" switches.

Heavyweight: SN2700

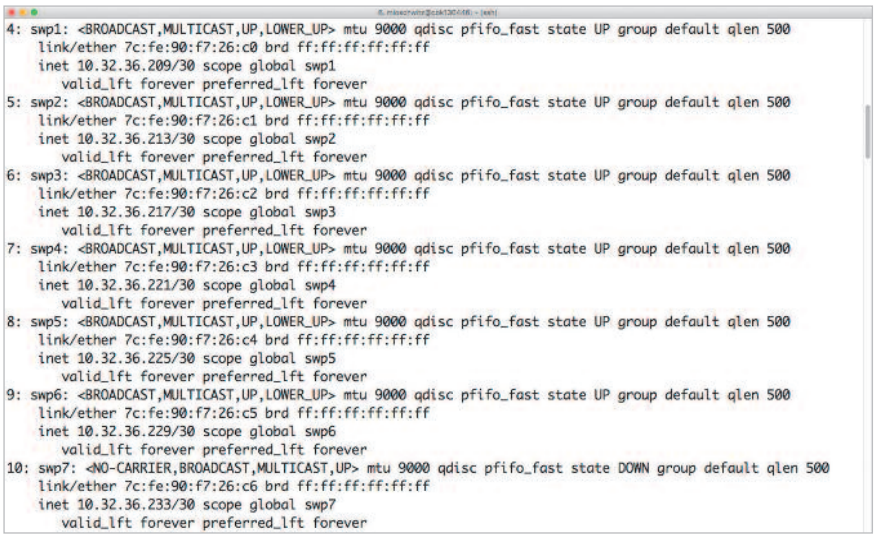
The front panel of the largest representative of the new Mellanox switches (Fig. 4) has 32 real 100-GbE ports based on the Spectrum ASIC. The unit takes


Source: Mellanox

Mellanox markets the SN2410 as a beast of burden: it has 48 25-GbE ports and 8 100-GbE ports (Fig. 3).

Source: Mellanox

The SN2700 offers 32 real 100-GbE ports and is designed as a spine switch for large cloud setups (Fig. 4).



Cumulus Linux turns a switch into a normal Linux server that can be managed and controlled using Debian tools. The image shows the output of the ip a command (Fig. 5).

up one HU. The 6.4 Tbit/s backplane mentioned above works in the background. Incidentally, the SN2100 has to make do with 3.2 Tbit/s, and the SN2410 with 3.6 Tbit/s. This makes the SN2700 suitable as a switch for the spine level of a leaf-spine architecture in which the leaf switches of many racks communicate via an SN2700. Alternatively, breakout cables can be used to coax the SN2700 to provide 64 ports of lower speed – the Spectrum ASIC can't handle any more than this. The practical relevance of a setup of this kind is low, however, as Mellanox also offers the SN2100 with twice as many ports in one height unit.

Low latencies and jumbo frames

Besides the differences described, the switches have a lot in common: one of the core functions stressed by the manufacturer is the fact that the Spectrum ASIC is optimized for low latencies. According to Mellanox, a package needs 300 nanoseconds to get from one port to the other. Typical network functions such as Jumbo Frames, VLANs or LACP are no problem for the units either.

Mellanox not only presents its MLNX-OS for the Spectrum series, but has also given its devices the Switch Abstraction Interface (SAI) that allows other operating systems to be installed in them. This means that a Mellanox switch can be integrated into the computer center like any Linux server you care to name, and you can even edit it with tools such as Ansible – if the correct operating system is running on it, that is.

The Switch Abstraction Interface

The SN series is not the first to allow different operating systems to be installed. This is because Mellanox belongs to the Open Compute Project (OCP, [a]), the declared aim of which is to develop an open software platform for the switches of different manufacturers. In addition, Mellanox made a considerable contribution towards designing the SAI and implementing it in a first version.

Typically, the firmware for switches is proprietary – so vendor lock-in is a potential danger. It is often not even possible to combine the devices of different manufacturers with a platform in such a way that they can function without difficulties. For example, Jumbo frames are not standardized, making them a problem for heterogeneous network architectures.

This is where the SAI comes in: as part of the switch firmware, it communicates with ASIC while offering a standardized interface for external programs allowing you to take information directly from the ASIC or send commands to it. An "external program" can for example be the Linux kernel: the SAI produces normal Netlink events which a Linux kernel can handle with ease. In other words, Linux can be used as an operating system on a switch in the same way as it could on a normal PC – as an alternative to the proprietary manufacturer software. For example, the entry of the ip a command in a switch of this kind provides one network interface per port (Fig. 5).

Devices with SAI can be operated reliably in mixed environments and do not have the described lock-in effect. For example, a Mellanox switch with Linux could easily be replaced by a model from Dell.

No great variety

Admittedly, there is not yet much of a range of different operating systems for the switches with SAI. At present, Mellanox is only working with Cumulus, a supplier already established in the market [b], to get the Linux of the same name onto the switches of the SN series. The cooperation has been successful: if you want to use Cumulus, you can buy the devices from Mellanox ready-made including software and the necessary license.

After installing the switch, administrators use Linux on the basis of Debian Jessie: network interfaces for example can be administrated via /etc/network/interfaces. If necessary, administrators are free to fetch any desired Debian packages into the system using apt-get install – even though Cumulus doesn't offer any support here. If a firmware update is due, it can be performed using apt-get update && apt-get dist-upgrade.

High flexibility

As services for the Broader Gateway Protocol (BGP) are available with Quagga (preinstalled) or Bird (optional), it is possible to operate a Mellanox switch of the SN series in the L3 router mode. This is common practice in scaled environments such as clouds: if you put your network into a leaf-spine architecture, you usually wire one server using a dual-port network card with switches in different racks and perform the announcement of the fastest path via BGP at routing level. Proprietary firmware is capable of this function too – but the suppliers often make you pay a high price for it in the form of an additional license.

In addition, a setup of this kind makes (physical) scaling easier: an SN2700 allows a total of 31 racks to be connected if you assume that one 100-GbE port is required for the uplink to the core router. If it later turns out that the setup is to consist of more racks, the need for a further switch level can be included in the L3 router setup – even when the system is up and running.

The subject of automation plays a role with Cumulus too: precisely because the switch is "only" a normal computer, the usual automators are available. Like Ansible, for example: using a suitable playbook, the administrator can automate the deployment of a switch including Bird or Quagga for a Layer 3 setup – reproducibly and without the need for any additional modules. The configuration can be reproduced reliably at any time. Also, in contrast to the proprietary systems offered by the competition, the administrator does not need to busy himself with the APIs intended for automation there.

And what about the clients?

When you plan a new network infrastructure, it is worthwhile to keep an eye on the clients besides the switch level so as to use the necessary bandwidth to the full. For this purpose, Mellanox offers an Ethernet version of its ConnectX-4 card, which was originally designed for InfiniBand. The successor generation has already been announced, but it can probably not be expected before spring 2017.

The most potent variant of the card possesses two 100-GbE ports. If you can make do with less speed – for operation with an SN2410, for example – you will no doubt reach for the smaller model. ConnectX-4 cards expect an x16 PCIe slot according to PCI Express version 3.0. However, it is recommendable to use a 2-port variant (Fig. 6) so that Layer 3 routing can be set up as shown.

Hardware offloading

In the scope of functions offered by the ConnectX-4 chip set, the offloading capabilities for the operators of large distributed systems are especially striking. VXLAN offloading, for example: as soon as the function has been activated, the chip set of the network card itself takes over the job of evaluating any VXLAN information of individual packages and treating them accordingly. The same work should normally be carried out by the kernel of the active operating system that receives the package. However, this always takes considerably longer than when it receives the completely "predigested" package from the card. Offloading in the hardware has an effect on the data rate, and in particular on the latency, which drops considerably as a result of comprehensive offloading.

VXLAN is used for cloud computing in particular – so this technology is of special interest here. The next generation, ConnectX-5, will require PCIe 4.0. The company says that Open vSwitch, a standard popular in the cloud and SDN environment, is to be realized by the network card via offloading in the case of ConnectX-5 in the same way as with VXLAN.

Functioning drivers, useful firmware

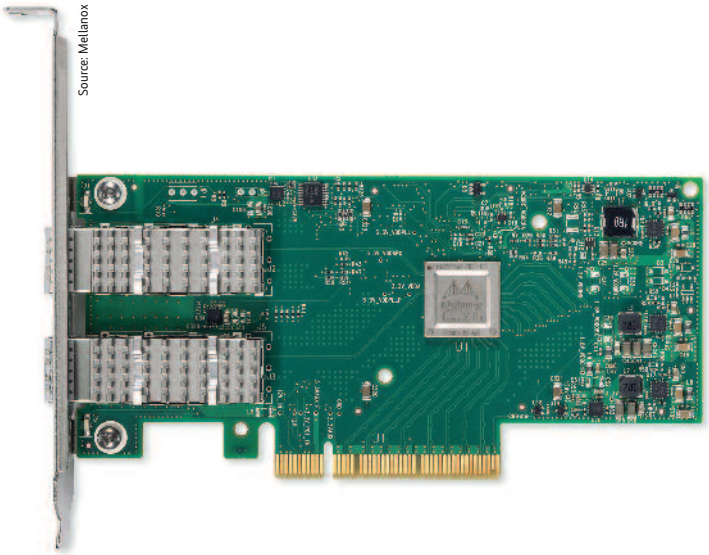
Drivers for the Mellanox NICs are to be found in the Linux kernel, and as an alternative, the manufacturer offers complete packages including more recent versions, which usually offer more functions. In addition, the firmware image allows administrators to define whether a ConnectX-4 card is to use NetBoot via DHCP and PXE or the more recent UEFI standard.

On demand, Mellanox can also supply the cables necessary for operation as well as suitable modules – Active Optical Cable (AOC) and Direct Attached Copper (DAC). Administrators should at least take this into account in their budgets, for depending on the technology and the cable length required, the cables can increase the total cost of the solution considerably.

Made for the cloud

Which target group is the company actually addressing? Who might find it worthwhile to have a closer look? The Spectrum ASIC is obviously suitable for classical HPC environments in which large amounts of data have to be transferred via a network. The SN2700 offers more performance than most of today's servers are able to make use of. Limiting factors here are local storage media such as SSDs on the one hand and the installed buses on the other, in the majority of cases the PCI Express bus according to Standard 3.0. It is no accident that, in connection with the Connect C5 cards, Mellanox is already looking to the new PCIe 4.0 standard, which provides considerably more bandwidth.

However, the SN Series will no doubt not be focusing so much on HPC environments – for years now, Mellanox has been supporting these with InfiniBand, which offers not only high data rates but extremely low latencies too. In spite of all optimizations, the Spectrum chip cannot match these. The product de-


Source: Mellanox

Mellanox offers not only Spectrum switches but also network cards of its own. Shown here is a ConnectX-4 Lx EN with two 25-GbE ports (Fig. 6).

scriptions of the Spectrum switches and the Ethernet versions of the ConnectX-4 cards are more informative in terms of the companies the manufacturer sees as potential customers: words like "cloud", "DevOps" and "automation" are often to be found there.

Many companies are currently dealing with the subject of cloud computing and are giving some serious thought to entering the business themselves, either by building a private cloud for their own use or a public cloud. Large, distributed installations make two principal demands on the hardware used: it has to be powerful enough to stand up to the workload of a large number of virtual machines at the same time, and it must be possible to integrate it into the setup in such a way that it does not cause any special maintenance effort.

Mellanox fulfils criterion 1: ConnectX-4 cards already offer enough power for an individual system if "only" the version with two 25-GbE ports is used (or with one port if the subject of redundancy is not important). On the rack level, the administrator can choose between the SN2100 and the SN2410; the SN2700 with its various outer designs was obviously intended as a core switch to connect the switches installed in the racks.

With regard to the subject of maintenance effort, the Switch Application Interface comes into play, as does Cumulus as the operating system: this means that a switch can be administrated like any other device – because it is a completely normal computer with a large number of network connections. It can even be a central part of the setup: OpenStack clouds for example are usually based on the principle that external traffic flows from VMs via gateway nodes which take care of DNAT and SNAT settings on the one hand while announcing the external cloud network to the outside world via BGP on the other. All this can be done by a switch in the case of Cumulus – and as any desired packages can be installed, it would not even be a problem to install the necessary components of OpenStack directly on the switch.

A setup of this kind no longer needs an expert for each device. Here Mellanox appears to be pushing ahead with a process of change in the DevOps environment: network professionals with solid basic know-how in matters of system administration are increasingly making way for generalists with a solid basic knowledge of subjects such as network connections.

If you're planning a network for a distributed system, you should definitely have a good look at the Mellanox hardware. It's hard to obtain a comparable combination of flexibility on the one hand and power on the other at present – Mellanox is considerably upping the benchmark for the competition here.

Who's to pay for all this?

The price is a critical factor for the purchase of hardware. For the SN2410 in the 25-GbE version with Cumulus for example, Mellanox names a price of about 14,000 euros. Then there are the matching cables necessary for attaching it to a fictitious SN2700: these cost about 840 euros each (cable length: three meters). For the redundant connection to two core switches, that comes to about 1680 euros in total. Also, 100 euros are charged per 3-meter cable fitted with suitable plugs. A fully equipped SN2410 with a redundant uplink therefore adds up to a total price of over 20,000 euros net. Then there are the costs for Cumulus support, which Mellanox can sell along with the unit if necessary: the 2500 euros for three years that they charge for this on the Gold level are moderate, however, and they include comprehensive support for the switch and the Cumulus.

A second example is shown by the pricing of the SN2700: the manufacturer offers the device itself for around 19,500 euros. The full equipment with 32 100-GbE cables (3 m) costs another 27,000 euros. A cable with a length of 50 meters for connecting racks positioned a long distance away costs about 1650 euros. Completely equipped, therefore, an SN2700 reaches a net price of around 46,500 euros. Here, too, the support costs for Cumulus must be added to the total.

The network cards are almost negligible here – especially in comparison with the products of the other manufacturers: Mellanox sells a ConnectX-4 card with two 25-GbE ports for about 400 euros, with a small extra charge if you require a support contract. The breakout cables already mentioned on several occasions should not be forgotten either: a 1-meter cable for the transition from 100 GbE to 4 x 25 GbE costs around 150 euros.

Conclusion

With its Spectrum switches and ConnectX network cards, Mellanox has got what it takes to be a game-changer in the Ethernet market: while it is true that large manufacturers can keep up when it comes to hardware – at least in terms of switches – Mellanox is practically out on its own when it comes to network cards. Something which virtually no other manufacturer can supply are high-performance switches with the flexibility of a free operating system such as Cumulus. The combination of installed technology, useful software and attractive price is successful on the whole – and it is suitable for use in modern DevOps and cloud scenarios in particular. If you are planning a cloud, you should have this Israeli company on your radar. (jab)

Martin Gerhard Loschwitz

is Teamlead OpenStack at SysEleven in Berlin, where he deals mainly with the subjects of distributed storage, software-defined networking and OpenStack.

All links: www.ix.de/ix1612066