We present the new compact hybrid switch with densely integrated on board optical modules. We show that four such switches can be packed into a single 1RU 19" box, and operate with moderate power consumption. The integrated box can be easily thermally managed with forced air cooling thanks to the substitution of the front panel pluggable transceivers by on board optics modules. Further, by introducing a programmable power control and relying on an SDN controller, we suggest that the integrated box can be used in different power modes saving considerable power when scaling to mega size data centers.

Introduction
The optical front panel density per rack unit of data center switches, in terms of number of ports and bandwidth, is limited by the number of pluggable transceivers that fit into the rack unit front panel. This is sometimes referred as to the front panel bottleneck. Despite this constraint, the use of front panel transceivers in data center switches is widely extended mainly due to two reasons: the presence of developed standards allowing modules selection from different manufacturers, and the easy replacement in case of failure, without the need of removing the rack unit from the rack. For instance, with 10G technology, it is common to package a single 128-ports switching ASIC with 32 QSFP optical transceivers (4 x 10G), providing a total of 1.28 Tbps bandwidth per rack unit. More recently, the industry is moving to 25G technologies with the same amount of ports, 32 QSFP28 optical transceivers (4 x 25G), and 3.2 Tbps bandwidth per rack unit.

On-Board Optics (OBO) transceivers were proposed as an alternative to front panel transceivers many years ago, However, the application has been restricted mainly to High Performance Computing (HPC), where performance is preferred over standardization and replacement flexibility. There are few examples of On-Board Optics modules used in data center switches. However, they only scale up to tens of thousands of hosts, and use Infiniband. Since we are interested in data centers above 100k hosts, and Ethernet prevails over InfiniBand, those devices are not especially interesting in our scenario.

This research also explores the use of On-Board Optics transceivers as optical interfaces of data center switches, aiming to find out the highest level of integration that can be achieved with current technology in a single rack unit. The constraint to a single rack unit arises from our scale-out approach, by which it is relevant to have the smallest single-board design, Software Defined Networking (SDN) enabled platform that can be used as building block for the whole data center network.

To the authors knowledge, the proof-of-concept design with a size of only 20cm x 20cm achieves the smallest single-board implementation of a 128-ports switch, and the packaging of four switches in a 19" rack unit achieves the highest level of integration per rack unit with a total of 4 x 128-ports.
Compact switch design with high density optical transceivers

It is well-known that switching ASIC’s with the largest port count are highly desirable because they reduce the amount of switches required to interconnect a certain amount of servers, and they lead to flatter data centers with reduced latency\(^8\). In a similar way, transceivers with a large number of ports are also especially relevant in large data center deployments, since a large port count reduces the number of transceivers needed, and also, it reduces the power and cost per Gbps\(^7\). For instance, three times more transceivers are required when comparing 4-ports and 12-ports modules (32 QSFP/QSFP28 vs 11 OBO-10G/OBO-25G for a 128-ports switching ASIC).

Our research shows that the reduced amount and small size of the OBO transceivers, closely placed to the switching ASIC, allows very compact designs, like our proof-of-concept implementation with 20cm x 20cm shown in Fig. 1. It includes eleven 12x10G OBO transceivers\(^10\) surrounding the 128x10G electronic switching ASIC\(^11\) with a total of 1.28 Tbps, covered by a fan in the image. The control plane processor, attached in the top right corner of the board, manages the data plane switching ASIC thanks to the decoupling between control and data planes provided by SDN. The rest of the board, in the top left corner, includes the required interface connectors, and the voltage converters generating the internal voltages. A similar approach could be followed with the same-size 25G version of those devices\(^12,13\) to reach 3.2 Tbps.

Multiplying optical front panel bandwidth

It is widely common to package a single switching ASIC per rack unit, with the corresponding amount of front panel transceivers\(^2\). This coupling limits the optical front panel density of the rack unit. In contrast, our research shows that decoupling switching ASIC and front panel bandwidth allows front panel densities several times higher than the underlying switching ASIC characteristics. For instance, Fig. 2 shows a possible packaging of four of our switches into a 19” rack case. Being only limited by the case size, up to 6 switches could be placed in the bigger OpenRack unit of OpenCompute\(^2\).

As a result, the optical front panel density scales up to 512 / 768 ports and 5.12 / 7.68 Tbps, accessible through the MPO connectors provided by the 4 / 6 switching ASIC’s and 44 / 66 OBO transceivers packaged into the 19” / OpenRack case. The number of ASIC’s, transceivers and ports would be the same when implementing the 25G version of the hybrid switch, but in that case the bandwidth would scale up to 12.8 Tbps / 19.2 Tbps.
Thermal management of rack units densely populated with OBO

Thermal management of multiple switches packaged into a single rack unit requires special attention because the performance and reliability of passively cooled optical transceivers is highly dependent on temperature. Fig. 3 summarizes the experimental results obtained by reading the temperature sensors included in the transceivers of one of the boards during a set of experiments with 2, 4, and 6 fans configurations. Every extra pair of fans reduces the thermal stress in the transceivers, not only by reducing their maximum operating temperature, but also by reducing the temperature range variation. For instance, transceiver six has a maximum temperature of 50.7°C / 40.5°C / 35.2°C, a minimum temperature of 36°C / 30°C / 27.3°C with 2/4/6 fans, resulting in operating ranges of 14.7°C / 10.5°C / 7.9°C respectively. The measured maximum operating temperature of the transceivers is 51.6°C with 2 fans, 40.5°C with 4 fans, and 36°C with 6 fans. The changes introduced in the packaging experiment previously shown in Fig. 2 explain those results. The main modification consists of removing the power supplies from the rack case, getting rid of all the heat generated by them. Without the power supplies, more fans can be installed in the back panel for extra forced cooling, like in our example with six fans. In addition, OBO modules are distributed all around the rack case, and not concentrated in the front panel, so they receive additional benefit from the improved thermal system of the case. All this, combined with the fact that also the front panel has some free area for extra ventilation being not fully populated with fiber array connectors, explains the sufficient airflow and heat removal of the rack unit.

Fig. 3: (left) Measured temperature of the transceivers in the implemented hybrid switch
Fig. 4: (right) Measured power consumption of the rack unit under different operational conditions

Software controlled power saving schemes

The control plane processor included in the switches can configure the data plane sections of the rack unit in different operational modes. Thus, the optical front panel capacity and power consumption of the rack unit can be adjusted according to the utilization scheme of the data center optical network. Fig. 4 compares the total power consumption of the 512 optical ports 19” rack unit, with two, four and six fans. Every line is composed by eight segments, each one of them corresponding to a different operational mode. The first segment corresponds to the lowest power consumption of the rack unit, and it is achieved when only the four control planes boards are running to maintain communication with the data center central controller, but all the switching ASIC’s and the optical transceivers are powered off. The power consumption for the whole system in this state is as low as 25.8W. This mode is mainly interesting for
software defined data center networks where it is desired to keep the control plane network always on, and under certain conditions, to power down sections of the data plane network to save in power. The second segment of the three lines shows the power increase when enabling two, four or six fans respectively. The third section shows the boot up peak produced when powering up ASIC and transceivers. Finally, the fourth to eighth segments display the power consumption of the system when the OpenFlow controller is running, for configurations with no ports, 25%, 50%, 75% and 100% enabled (and increasingly loaded) ports. For instance, with six fans, the total power consumption increases from 341.3W with no ports, to 367.4W with 25% ports, 400.6W with 50% ports, to 427.3W with 75% ports, and to 457.2W with 100% ports.

**Conclusion**

We have shown how to overcome the front panel bottleneck inherent to front panel transceivers by substituting them by On-Board Optics transceivers. This type of transceivers, especially those with a large number of ports, allows higher levels of integration compared with front panel transceivers. Our proof-of-concept design with OBO scale this number up to 512 ports and 4 switching ASIC’s per rack unit, effectively multiplying optical front panel density. The approach is experimentally validated, both thermally and in terms of power. In terms of temperature, On-Board Optics transceivers, distributed all around the rack case and not concentrated in the front panel, benefit from any additional forced air cooling, which brings down their maximum operating temperature and their temperature variation range. In terms of power, the control plane processor included in the switches allows to configure and operate the data plane hybrid switches included in the rack unit in different operational modes, adjusting the optical front panel capacity and power consumption of the rack unit according to the utilization scheme of the data center optical network.

**References**