Converged photonic data storage and switch platform for exascale disaggregated data centers

R. Pitwon*[a], K. Wang[a], A. Worrall[a]
[a]Seagate, Langstone Road, Havant, Hampshire, PO9 1SA, United Kingdom;

ABSTRACT

We report on a converged optically enabled Ethernet storage, switch and compute platform, which could support future disaggregated data center architectures. The platform includes optically enabled Ethernet switch controllers, an advanced electro-optical midplane and optically interchangeable generic end node devices. We demonstrate system level performance using optically enabled Ethernet disk drives and micro-servers across optical links of varied lengths.

Keywords: Disaggregation in data centers, optical interconnect, disk drives, micro-servers

1. INTRODUCTION

Projections of digital information growth show that by 2020 44 ZB of digital information will have been created, of which at least 15.4 ZB will be useful if stored[1]. However the total amount of data that installed data storage capacity will be able to hold, including the combined total capacity of Hard Disk Drives (HDDs) and Solid State Drives (SSDs), will be 9.9 ZB. This will give rise to an estimated 5.5 ZB discrepancy between data that should be stored and data that will be able to be stored. Furthermore a dramatic shift is taking place in where data is located. As shown in **Figure 1**a, in 2010, 65% of data was stored on the client-side, mostly residing on PCs. However, as shown in **Figure 1**b, the proliferation of more mobile compute and communications platforms, such as smart phones, tablets and ultra-slim notebooks, is changing this distribution dramatically.

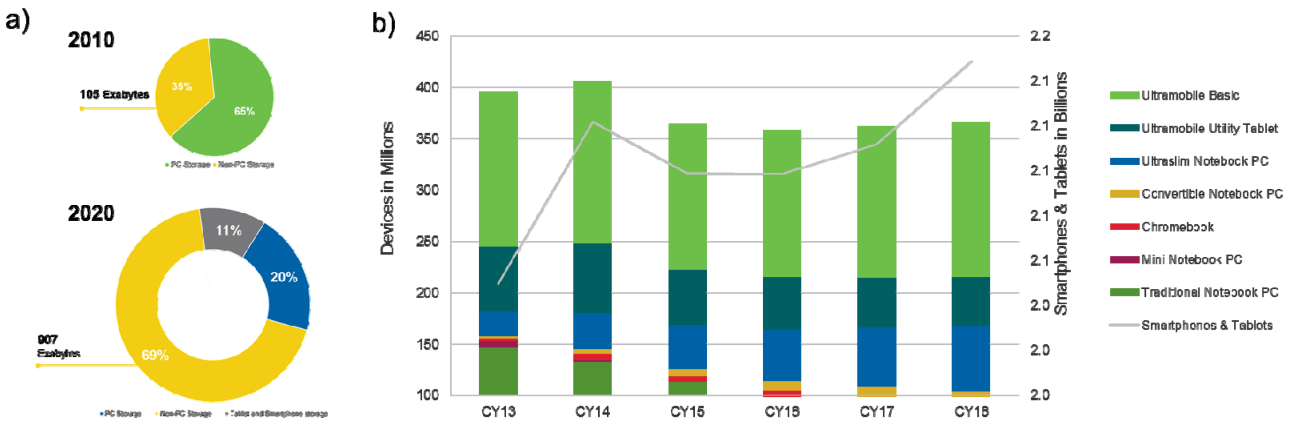

Figure 1. Data storage and compute mobility trends, a) Location of data storage shipped in 2010 versus projected location distribution in 2020[2], b) Compute device categorization 2015[3]

Thus by 2020 it is projected that only 31% of data will be stored on PCs, smartphones and tablets, while 69% of data will be stored in remote data centers (**Figure 1**a). The reason for this is that greater amounts of data are being routinely generated, which more quickly exceed the local storage capacities on the mobile devices, thus necessitating remote storage on Cloud data centers.

Furthermore according to the latest Cisco Global Cloud Index[4] projections, around 79% of digital communication within data centers will comprise so called "East-West" traffic, that is communication between nodes in the same data center. As data centers increase in size to exascale capacities, data center architectures need to evolve accordingly to manage the corresponding increase in complexity of internode communication, while reducing power consumption and total cost of ownership (TCO).

In order to sustain exponential global digital data growth, exascale Cloud data center providers are turning to disaggregated schemes, such as OpenCompute, in which distributed computing and storage resources from across the data center can be dynamically assigned to tasks as required, rather than being restricted to a localized pool of dedicated resources. Such disaggregation architectures have the potential to substantially reduce total cost of ownership, but will require ubiquitous optical connectivity to overcome the distance between resources. While current schemes target higher single mode connectivity tiers in the data center, the continuing depreciation in transceiver cost is fuelling the migration of optical interconnect deeper into the cost sensitive, high volume, low margin equipment under the Top-of-Rack including data storage and switch platforms.

We report on a converged optically enabled Ethernet storage, switch and compute platform, which could support future disaggregated architectures, but goes further than traditional models in that it provides deeply migrated optical links right down to the storage and switch subsystem nodes themselves. We validate system performance across optical links of different length up to 150 meters using optically enabled Ethernet disk drives and micro-servers.

2. EUROPEAN HORIZON 2020 PROJECT: NEPHELE

Nephele ("End to end scalable and dynamically reconfigurable optical architecture for application-aware SDN cloud datacenters") is a collaborative research project on optical data center network technologies, supported by the Horizon2020 Framework Programme for Research and Innovation of the European Commission. The three-year project started on February 2015 and brings together seven leading European universities, research centers and companies with National Technical University of Athens acting as coordinator[5].

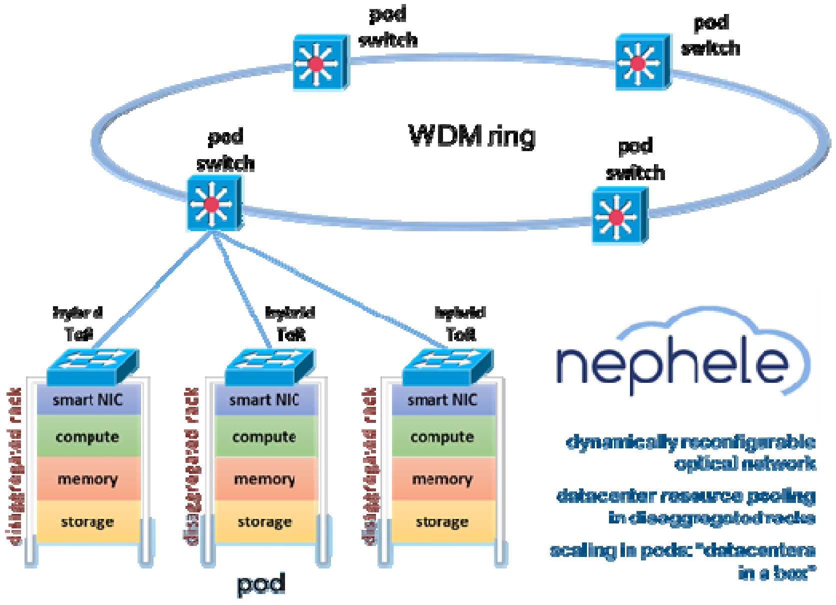

Figure 2. Nephele advanced hybrid electro-optical scale out data center architecture

The aim of the Nephele project is to develop an advanced dynamic optical network architecture for future scale-out, disaggregated data centers as shown in Figure 2. Nephele seeks to leverage the enormous bandwidth capacities of installed optical fiber infrastructures and deploy hybrid optical switching to attain the ideal combination of high bandwidth at reduced cost and power compared to current data center networks.

The project performs multidisciplinary research, extending from the data center architecture to the overlaying control plane and to the interfaces with the application, in order to deliver a fully functional networking solution. Driven by user needs, Nephele's end-to-end development path aims to bridge innovative research with near-market exploitation, achieving transformational impact in data center networks that will pave the way to exascale infrastructures.

3. DESIGN OF CONVERGED OPTICALLY ENABLED PLATFORM

As part of the Nephele project, Seagate developed a converged switch, storage and compute platform, NephDem06.01 (Figure 3), which comprises optically enabled and interchangeable switch controllers (FireBird), an advanced Prizm MT® compliant electro-optical midplane and optically interchangeable storage and compute end nodes (Figure 3). In order to satisfy the need for redundancy in enterprise class data storage array systems, the interconnect topology is based on a dual star configuration, whereby each end node subsystem supports two bidirectional data links on the midplane, one to each of two separate FireBird controller modules.

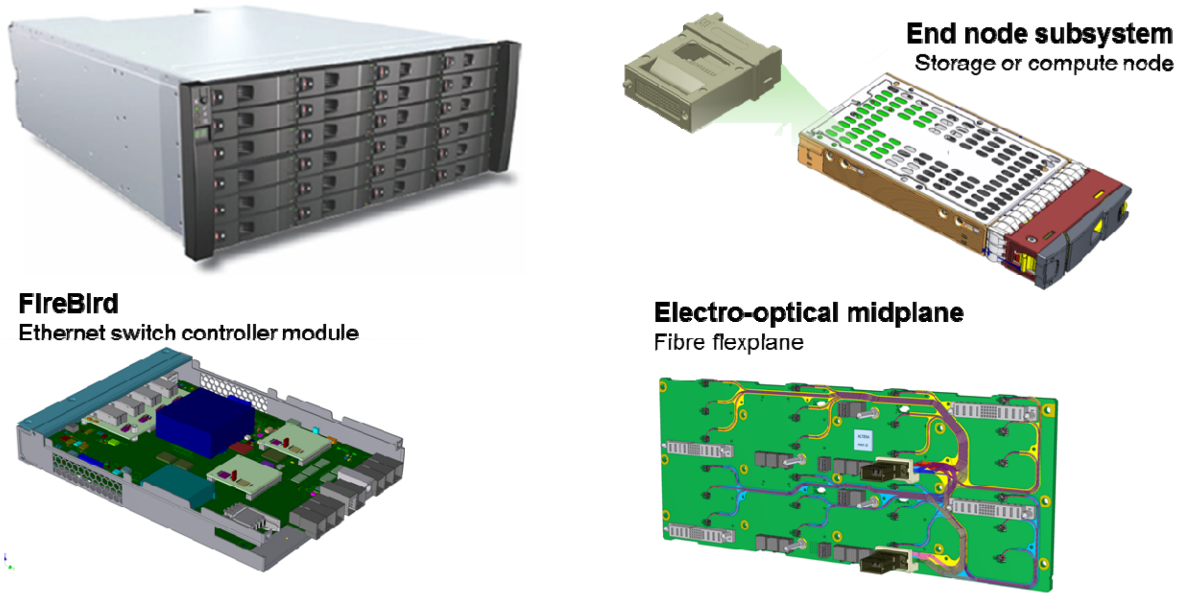

Figure 3. Nephele converged optically enabled data storage and switch platform, NephDem06.01

3.1 Data center rack system form factor

The NephDem06.01 platform was designed to fit into a 19" wide enclosure (Figure 4), which is a common data center rack width, and a height of 7" corresponding to 4 Rack Units (4U).

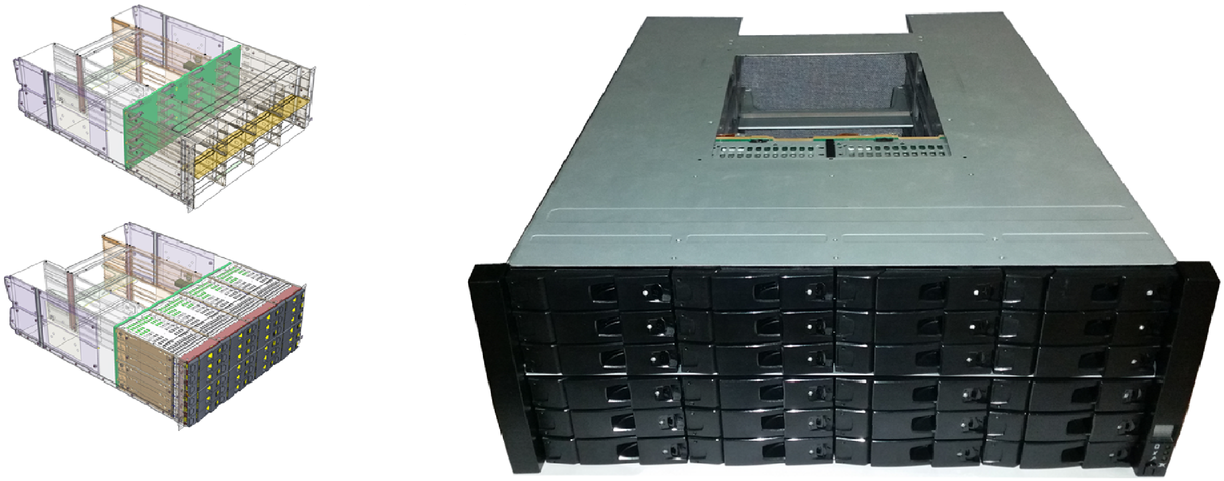

Figure 4. Data storage and switch enclosure for 19" data center racks based on 4U24 form factor

The enclosure accommodates an electro-midplane positioned to receive 24 optical end node slots on the front side shown and 2 optical controller slots in the rear.

3.2 Storage switch controller with generic optical mezzanine slots

The optically enabled storage controllers (NephDem06.01.SC1) are designed to accommodate Ethernet switches, allowing object oriented connectivity between up-stream systems such as Top-of-Rack switches and downstream devices, in this case the 24 optical end node device slots, which on Nephele can be either optically disaggregated storage or compute systems.

The NephDem06.01.SC1 controller boards include midboard mezzanine slots (Figure 5a) for both upstream and downstream optical communication. The slots are designed to hold small pluggable midboard optical transceiver mezzanine boards (Figure 5b). One larger mezzanine board slot (shown on left edge) is provided for upstream connections and uses a default card with 4 SFP+ cages assembled onto it. Two smaller mezzanine board slots are provided for downstream connections and are each designed to accommodate at most 12 bidirectional links, thereby collectively providing 24 bidirectional links as required by the system topology. By default a mezzanine card has been designed, which houses a Finisar BOA® board mounted optical transceiver module with 12 bidirectional channels[6].

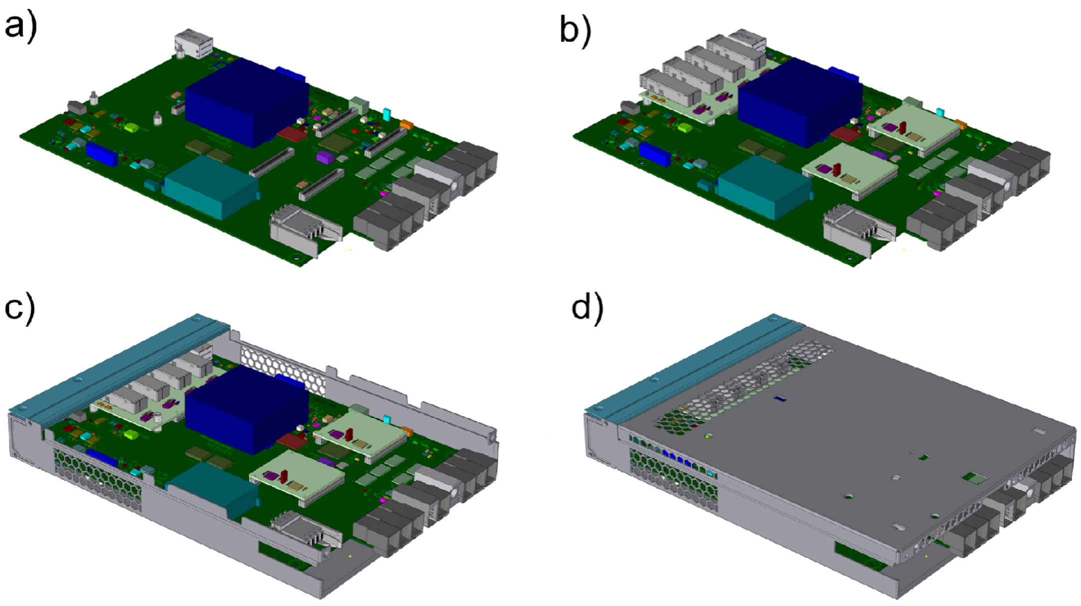

Figure 5. Nephele NephDem06.01.SC1 Ethernet storage controller with mezzanine slots for upstream and downstream optical interconnect slots a) NephDem06.01.SC1 controller card with unpopulated optical mezzanine card slots, b) NephDem06.01.SC1 with upstream and downstream optical mezzanine cards populated, c) NephDem06.01.SC1 partly assembled into Storage Bridge Bay (SBB) canister with no top lid, d) NephDem06.01.SC1 fully assembled into SBB canister

The generic form factor for both types of mezzanine card will allow different variants of mezzanine card to be developed to house different types of transceiver, from either commercial or research sources. This feature is crucial as it would allow lower TRL transceiver technologies, such as those based on silicon photonic chip sets, developed on other international research projects to be immediately deployed and characterized in a fully functional data center system.

As shown in Figure 5c and Figure 5d, the NephDem06.01.SC1 controller is designed to fit into a canister form factor, which complies with the the Storage Bridge Bay Specification[7], defining mechanical, electrical and low-level enclosure management requirements for a data center enclosure controller slot.

3.3 Electro-optical midplane with PrizmMT® compliant terminals

The electro-optical midplane comprised an electrical midplane and a separate interchangeable flexible optical fiber circuit. The fiber circuit (Figure 6a) was designed to be optimally compliant with the electrical midplane form factor and the positions of the optical ports for the controller and end node device slots. Each fiber lead was terminated with a lensed PrizmMT® ferrule. The 192 fiber circuit provided 4 bidirectional links to each end node slot, even though the end node storage and compute devices developed only required two bidirectional links in this architecture. This allows easy accommodation of future end node devices requiring 4 bidirectional links, such as those using PCI express interfaces. Figure 6b and Figure 6c respectively show schematic and actual views of the electrical midplane with the 192 fiber flexplane attached.

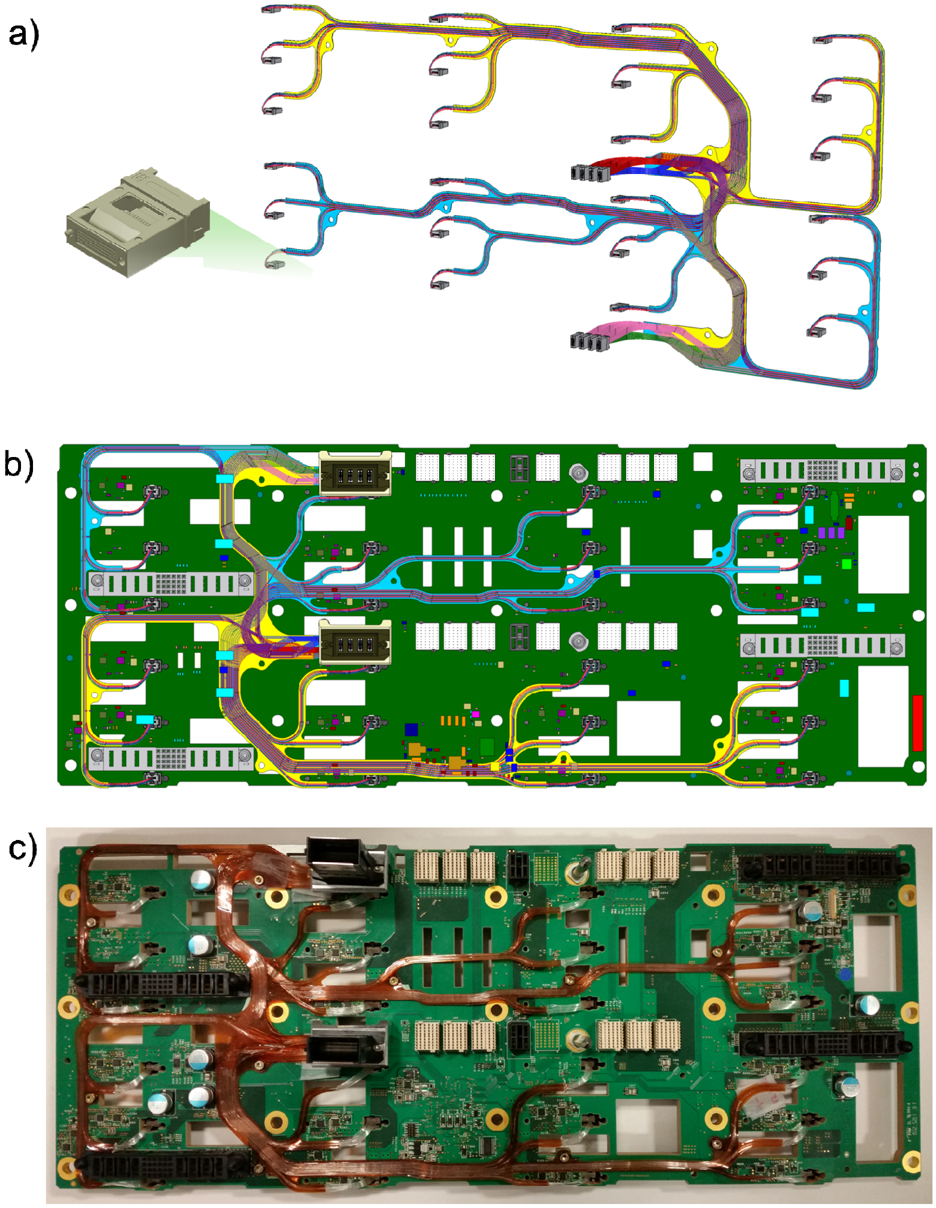

Figure 6. Electro-optical midplane, a) CAD view of 192 fiber flexplane with Prizm MT terminations, b) CAD view of electrical midplane with flexplane attached, c) photo of electrical midplane with flexplane attached

3.4 Storage and compute end node devices

The optically enabled end node devices were designed to fit into a carrier canister for a 3.5” hard disk drive, allowing them to be easily incorporated into standard data storage enclosures, which form the building blocks of modern data centers. Two varieties of optically interchangeable end node device were developed: 1) a storage device comprising a 3.5” hard disk drive with an Ethernet communications interface and 2) a compute device, which comprised a microserver platform with an Ethernet communications interface. Both included a specially designed active optical interface card and PrizmMT® compliant optical connector allowing them to be optically pluggable and interchangeable.

4. LINK PERFORMANCE CHARACTERIZATION

In order to validate the performance of optical links passed directly to the end nodes through the NephDem06.01.SC1 system, three levels of system link testing were carried out over 4 different optical link lengths: <1 meter, 50 meters, 100 meters and 150 meters.

4.1 End node device discovery over optical links

The first level of performance validation was the discovery of the end node devices by an Ethernet server. Both storage and compute end nodes were discoverable over all optical link lengths by a DHCP server running on a PC attached to the platform.

4.2 Ping testing over optical links

A ping test is a troubleshooting method of checking both connectivity and response time between two devices. As part of the second level of performance validation, ping tests were carried out between a PC attached to the platform to all end node devices in all 24 slots, whereby the optical link to each end node device was changed according to the desired link length under test. All ping tests were successful with no noticeable change in reported response time recorded over the optical link lengths measured.

4.3 Ethernet communication link over all optical links

Finally the performance of write and read data communication to end node devices over the four optical link lengths under test was characterized using a bespoke Ethernet traffic test script generator configured such that 50% of the test data was written to the end node (“put” data in object oriented parlance) and 50% of the data was read from the end node (“get” data in object oriented parlance). For each link length under test, the rates of put and get data provided to and from the optical end node device was directly compared with put and get data rates to an end node in the same platform connected through a standard short electrical PCB link. For optical link lengths the put and get data rates were comparable between the optically connected end node device and the electrically connected end node device.

5. PHOXLAB CROSS-PROJECT INITIATIVE

On the FP7 PhoxTroT project[8], Seagate and Fraunhofer IZM are developing a cross-project initiative called “PhoxLab”, which will allow advanced technologies and demonstration platforms from different international projects to be tested with respect to one another. PhoxLab defines sub-system card form factors, which can be disseminated to allow other projects to design test board platforms enabling advanced optical transceivers or switches (such as silicon photonics devices), optical circuit boards and connectors to be mounted into different systems of varying complexity and Technology Readiness Levels (TRLs).

It is envisaged that Nephele and other European, as well as international projects developing technologies in this sphere will be able to vastly expand their performance evaluation by participating in the initiative.

As shown in Figure 7, Seagate have developed the first converged PhoxLab test rack called “Nexus”, which includes both Nephele and PhoxTroT demonstration platforms. The Nexus test platform includes a NephDem06.01 platform and would allow the advanced transceiver technologies (for instance based on silicon photonic chips) and optical circuit board technologies (for instance based on singlemode planar glass waveguides[9]) to be evaluated within a Nephele architecture.

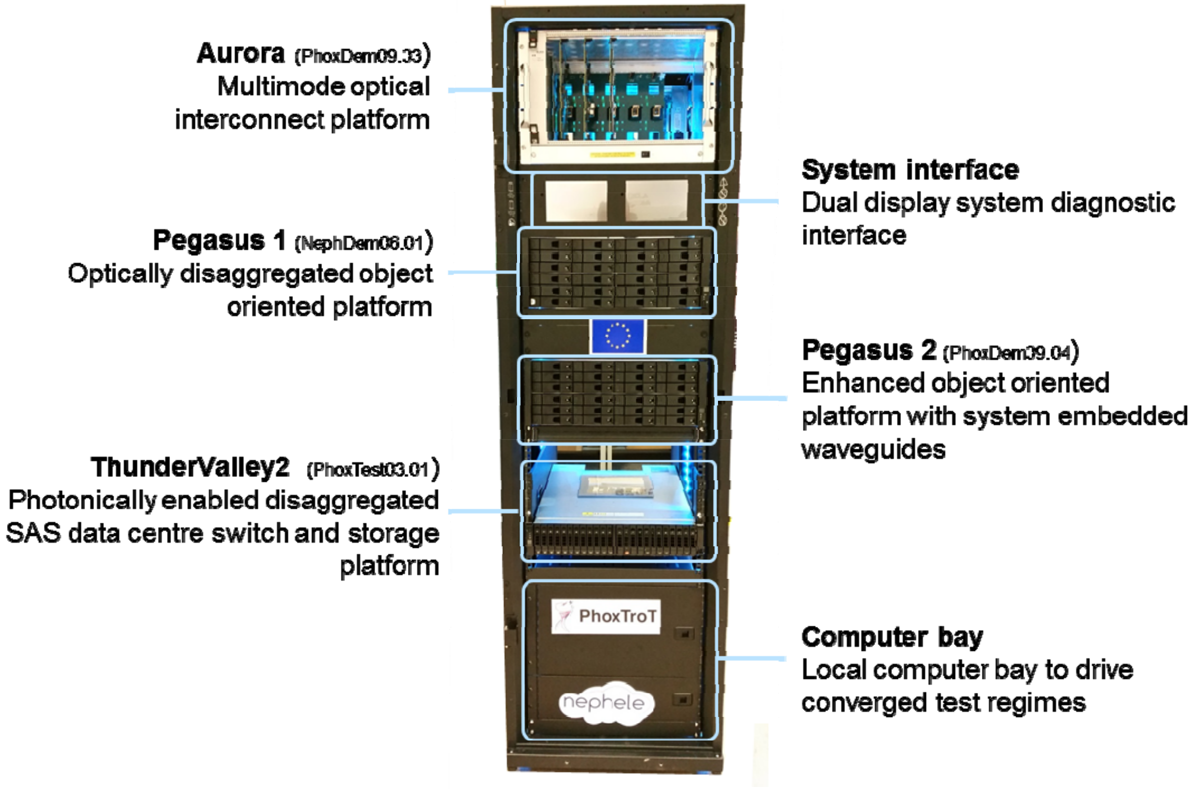
Aurora (PhoxDem09.33)
Multimode optical interconnect platform
Pegasus 1 (NephDem06.01)
Optically disaggregated object oriented platform
ThunderValley2 (PhoxTest03.01)
Photonically enabled disaggregated SAS data centre switch and storage platform
System interface
Dual display system diagnostic interface
Pegasus 2 (PhoxDem39.04)
Enhanced object oriented platform with system embedded waveguides
Computer bay
Local computer bay to drive converged test regimes
PhoxTroT
nephele

Figure 7. Nexus converged data center test platform

6. CONCLUSION

We have reported on the design and development of a converged optically enabled, object-oriented storage switch system, which will form a central pillar of the advanced data center architecture under development in the European Nephele project.

One crucial requirement throughout the technology development was the incorporation of generic mezzanine card slots at different points in the platform, allowing different future transceiver and interconnect technologies to be deployed and characterized within the Nephele data center architecture as well as other test environments.

The performance of the system was characterized and validated with different optical link lengths to the end nodes up to 150 meters.

This work was funded by the European Commission Horizon2020 Nephele project (Grant No. 645212).

REFERENCES

[1] D. Reinsel, “Where in the World Is Storage: A Look at Byte Density Across the Globe,” *Ind. Dev. Model.*, vol. Oct, no. 243338, 2013.

[2] Seagate, “Strategic Marketing & Research End of Quarter Database, Product Portfolio TAM Publication.”

[3] Gartner, “WW PC Forecast September 4Q2015; Forward Insights – NAND Insights, 4Q15 (Tablets & Mobile Phones).”

[4] Cisco, “Global Cloud Index (GCI).” [Online]. Available: http://www.cisco.com/c/en/us/solutions/service-provider/global-cloud-index-gci/index.html.

[5] National Technical University of Athens, “Nephele Project.” [Online]. Available: http://www.nepheleproject.eu. [Accessed: 08-Dec-2015].

[6] K. Schmidtke, F. Flens, A. Worrall, R. Pitwon, F. Betschon, T. Lamprecht, and R. Krraehenbuhl, “960 Gb/s Optical Backplane Ecosystem Using Embedded Polymer Waveguides and Demonstration in a 12G SAS Storage Array,” *Journal of Lightwave Technology*, vol. 31, no. 24. pp. 3970–3975, 2013.

[7] “Storage Bridge Bay (SBB) Specification,” *Storage Bridge Bay Working Group Inc.*, 2008. [Online]. Available: http://www.sbbwg.org.

[8] Tolga Tekin, “PhoxTrot project.” [Online]. Available: http://www.phoxtrot.eu/. [Accessed: 16-Nov-2015].

[9] L. Brusberg, D. Manessis, M. Neitz, B. Schild, H. Schröder, and T. Tekin, “Development of an electro-optical circuit board technology with embedded single-mode glass waveguide layer,” in *Proc. of ESTC,* 2014.