ERICSSON WHITE PAPER
284 23-3281 Uen | May 2016

ERICSSON

HYPERSCALE CLOUD

REIMAGINING DATA CENTERS FROM HARDWARE TO APPLICATIONS

The hyperscale approach can help industries meet the IT capacity demands of the Networked Society. Transforming data centers in this way goes beyond hardware to impact the software layer, including hypervisors, operating systems, cloud platforms and – ultimately – applications.

INTRODUCTION

As technologies such as broadband, mobility and cloud transform businesses and societies around the world, demand for IT capacity continues to grow. The rapid digitalization of industries, combined with the rise of the Internet of Things and the ongoing transition to cloud computing, are just a few of the factors that require vastly increased compute, data and networking resources.

In consequence, global spending on data center systems is growing [1]. However, the current ratio between capacity and cost means that even increased spending will not deliver the capacity required.

The solution is to go hyperscale: to allow infrastructure to scale beyond the cost and capacity limitations of today's data center architecture [2]. And the first step in this journey is to rethink that architecture in order to make it more modular, flexible and smart. These characteristics bring new functional opportunities, but are also a step toward reduced total cost of ownership (TCO).

While demand for this type of infrastructure is becoming increasingly apparent, and initial solutions have been made available, there are certain challenges that must be overcome if the hyperscale vision is to be fully realized.

This paper discusses the path toward hyperscale. It presents the scope of the impact of this data center transformation and reveals how the benefits can go beyond the hardware layer to impact the software layer, including hypervisors, operating systems (OSs), cloud platforms and – ultimately – applications.

TODAY'S HARDWARE-DEFINED INFRASTRUCTURE

A key challenge relating to today's data centers is the low level of resource utilization. This is partly caused by resource stranding [3]: a fragmentation problem created by varying application profiles (such as compute-intensive, memory-intensive and networking-intensive applications). To address this challenge, data center operators employ virtualization technologies such as hypervisors and containers to enable resource sharing and facilitate scaling of resources.

But despite advances in virtualization, data centers still operate with low resource utilization. This is because existing technologies cannot overcome the hardware-defined infrastructure (HDI) boundaries of today's data centers.

This hardware setup is a result of the monolithic nature of component integration, leading to a one-to-one mapping between a physical server chassis and a system executing on that chassis (which is referred to here as the host – see Figure 1). Hosts therefore have fixed configuration properties, and this makes it extremely difficult for them to adapt to varying applications since, in most cases, the entire physical server chassis needs to be planned for each application profile.

Another effect of the current hardware component integration is that it tightly binds data center life cycle management to the life cycle of an entire server. This causes problems for data center providers who wish to upgrade part of their infrastructure, since different resources within a server may have different life cycles.

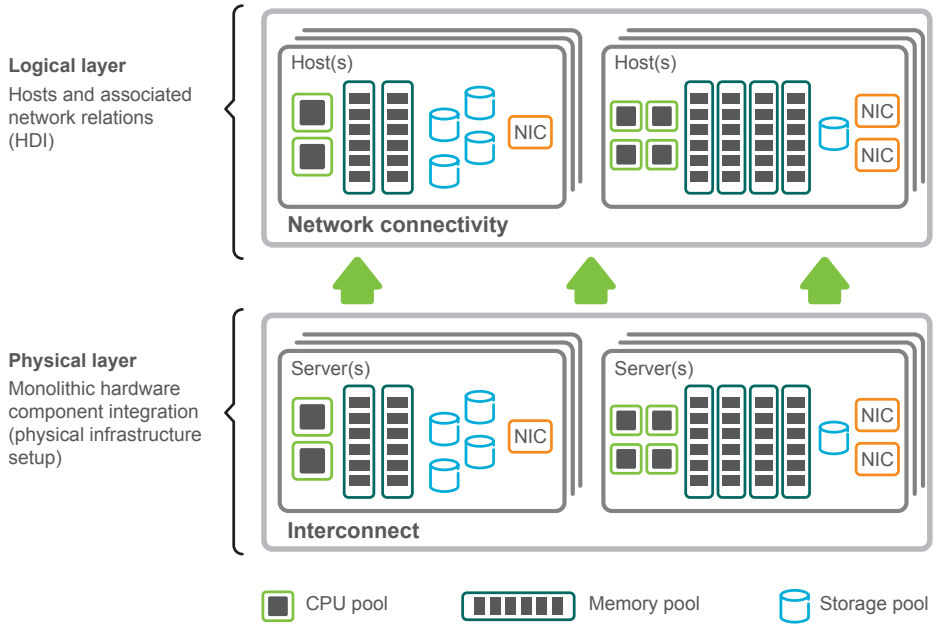

Figure 1: HDI.

SOFTWARE-DEFINED INFRASTRUCTURE: THE FIRST STEP TOWARD HYPERSCALE

ICT industry players and research communities have been working toward realizing hyperscale through new data center architectures that rely on the principles of hardware disaggregation and programmable infrastructure.

With this approach, physical server limitations are removed, and resources (for example, compute, memory and storage) are considered as individual, modular components. Disaggregation has been the term most often used to describe this phenomenon, where resources are organized in pools of common resource kinds. The approach brings greater modularity, flexibility and extensibility to the data center infrastructure, opening the way for software-defined infrastructure (SDI) rather than HDI.

This breaks the existing one-to-one mapping between a physical server chassis and a host executing on that chassis – as depicted in Figure 2. In this context, host composition is done independently from hardware-specific dependencies. With such infrastructure, data center operators will be able to employ resources in more efficient ways (in terms of increased utilization and lower power consumption, for example), ultimately leading to TCO reductions.

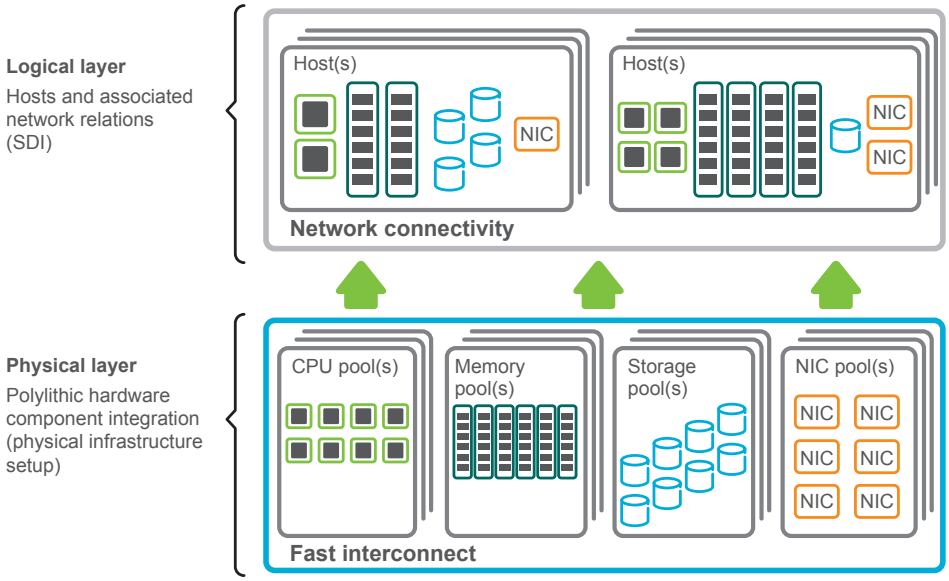

Figure 2: SDI.

HYPERSCALE DATA-CENTER-DRIVEN USE CASES

In a flexible and modular data center environment, numerous use cases can arise. These include the following:

Individual component replacement

One of the first use cases enabled by disaggregation is the ability to replace individual physical components in a more efficient manner. Replacement may be necessary due to malfunction or scheduled upgrades. Today, the data center operator is in most cases required to replace much

more than the individual component (in some cases the entire server), which creates unnecessary costs.

Flexible host setup

By breaking down the boundary of the physical chassis, hosts can be set up and adjusted dynamically according to the number and type of components required. Today, scalability is applied at the virtualization layer and is severely constrained by the capacity of the physical chassis. SDI and hardware disaggregation will overcome this and enable scaling properties at the physical level.

Dynamically optimized data centers

From a data center operator's perspective, the ultimate goal is to apply all possible policies that lead to cost reductions, while still fulfilling Service Level Agreements. The ability to adjust hosts dynamically allows for the application of policies that optimize resource allocation based on aspects such as utilization and power consumption.

To realize these and other future use cases, software-driven management, operation and execution systems need to be developed and deployed, as well as enabling hardware components.

THE FULL STACK IMPACT OF HYPERSCALE

At first sight, the impact and benefits of a new type of physical infrastructure may appear to be confined to the infrastructure layer and its management. However, its impact will be much broader than that.

The following section explains how the evolution toward SDI impacts every layer, including:

- infrastructure
- composition and resource orchestration
- workload execution
- data and applications.

It presents the key considerations for data center operators at every layer, including enablers, benefits and challenges.

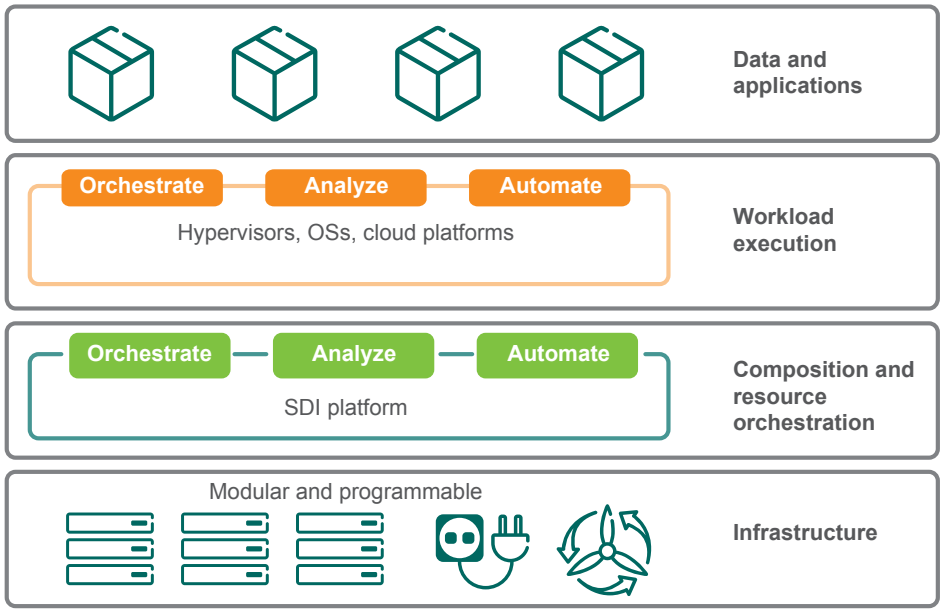

Figure 3: The hyperscale cloud stack.

INFRASTRUCTURE

Redesigning infrastructure to make it modular and programmable can relax today's data center constraints and allow for much more flexible operations. The fundamental elements that enable this transition include hardware disaggregation, networking, resource slicing, sharing and aggregation.

Hardware disaggregation (a breakthrough)

Until very recently, networking technologies were setting the constraint for the different components to be closely integrated (for example, in the case of bandwidth and latency requirements for CPU-CPU and CPU-memory integration among others – see Table 1). Today's networking technologies are relaxing this constraint, and the industry is aiming for more flexible integration through disaggregation.

One of the fundamental challenges lies in understanding the level of disaggregation possible.

Disaggregation can be seen on two levels: on one level, the hard mapping between logical and physical components within a single chassis is broken; and on the other, dependencies between components are removed, allowing for host composition across different chassis (for example, a host CPU unit in one chassis can now connect to a memory unit in another chassis).

Hardware disaggregation is in its early stages, and there is still a long way to go before its full potential can be achieved. Storage disks have been fully disaggregated, but challenges remain in relation to how to realize decoupling between the remaining components such as memory, CPU and the network.

Communication type	Delay (ns)	Bandwidth (Gbps)
CPU-CPU	10	200
Memory-CPU	20	300
CPU-10G NIC	>103	10
CPU-SSD disk	>104	5
CPU-HDD disk	>105	1

Table 1: Approximate communication requirements between resources within a server. The values differ between individual hardware sets [4].

Networking (the key enabler)

Networking is the true enabler for hardware disaggregation. Communication between different resources has highly demanding requirements, such as very fast speeds with low latency and no packet loss. Copper board traces have supported this communication, but have practical limitations.

Advances in silicon photonics have triggered a radical change in the economics of fiber-optic communication. Silicon photonics chips can now be created using the same manufacturing techniques that revolutionized CPU manufacturing. Complementary metal oxide semiconductor (CMOS) technology embedded onto the motherboard like any other chip eliminates the need for expensive optical transceivers. This greatly reduces the cost of fiber-optic communication in data centers and brings economies of scale to the optical communication [5].

These advances allow a shift to optical interconnection of components. In addition to the ability to move larger amounts of data at very high speeds, optics are more power-efficient than electrical signals over copper cables [6].

In a completely disaggregated environment, each resource (or pool of resources) will ultimately have a direct interface to the data center's network and utilize it to communicate with other resources. Communication previously restricted to a physical motherboard will be carried across the data center's network. The network is therefore required to handle a variety of interconnect technologies and protocols, such as the Intel QuickPath Interconnect (QPI), the Serial Attached SCSI (SAS), the Peripheral Component Interconnect Express (PCIe) and Ethernet.

As a result, there are fundamental questions that need to be answered, such as:

- Is there a need to modify or enhance the interconnect protocols?
- How can network resources be shared between different traffic types requested for communication between different resources composing hosts (for example, CPU to memory), and between different composed hosts?
- How can the requirements for component communication be met?
- How can forwarding and congestion control be performed?

Resource slicing, sharing and aggregation (unlocking the potential)

To maximize resource utilization, it is essential to be able to slice, share and even aggregate physical resources across different logical systems. In today's server-based architecture, this is accomplished – to a certain extent – by means of virtualization.

With a modular and programmable infrastructure, the data center operator can improve resource utilization – and eliminate resource stranding and fragmentation – by performing resource sharing at a lower level, such as the hardware layer or close to it. Figure 4 shows four examples of how resource slicing, sharing and aggregation could be realized in such an infrastructure:

- Example A: a dedicated physical unit is provided for a host and no slicing or sharing is actually done.

- Example B: resource units are sliced and each part is provided to different hosts as dedicated logical units.
- Example C: slicing is realized and sharing is also enabled.
- Example D: multiple resource units are aggregated and provided to the host as a single logical resource.

Other examples could be considered in which the different approaches are mixed (for example, providing dedicated memory and storage units, and slicing and sharing a CPU unit). Moreover, traditional virtualization technologies based on hypervisors and containers can still be applied at the host level.

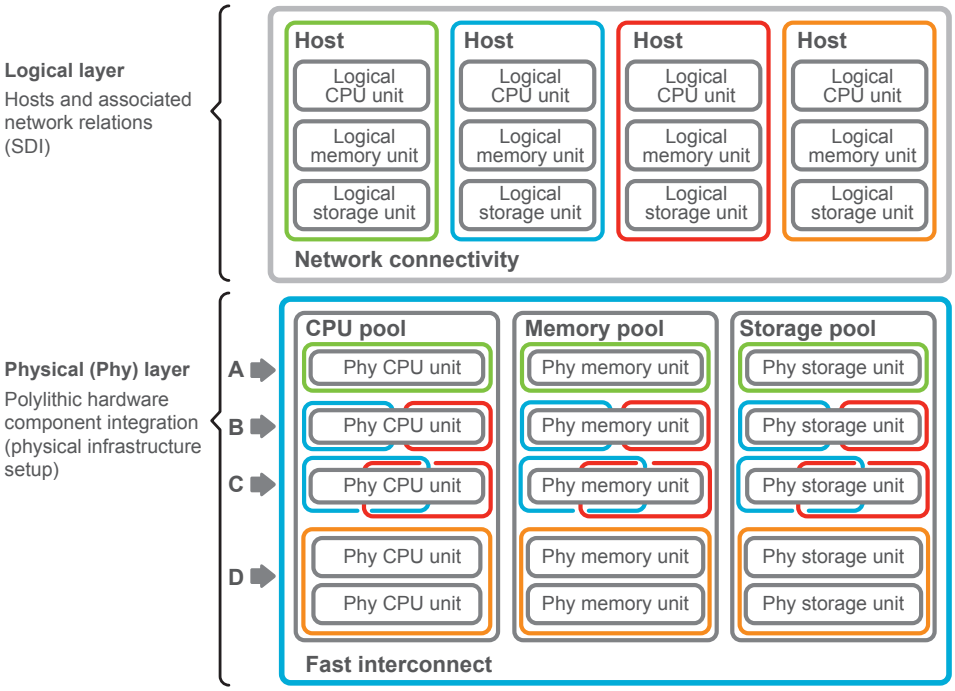

Figure 4: Resource slicing and sharing.

COMPOSITION AND RESOURCE ORCHESTRATION

The effect of making infrastructure flexible may be minimal unless proper decisions are taken at higher levels of the stack.

There are several key composition and orchestration elements that should be taken into account: resource scheduling; analytics; application-aware orchestration; and hardware planning and setup.

Resource scheduling (taking appropriate decisions)

Due to the limitations of current data center architecture, scheduling involves the selection of the most suitable server to host a particular job according to server properties such as its physical and topological information, as well as data center operational policies. This operation might seem simple compared with a future disaggregated environment.

Scheduling mechanisms can become extremely complex – especially when each individual component (compute, memory, storage, networking and so on) has to be considered.

Regardless, scheduling will significantly change in these environments, enabling more fine-grained decisions to be made. Many factors need to be taken into account, including power consumption, resource defragmentation, performance and utilization. Moreover, operations should not remain static; optimization processes should be carried out periodically to maximize overall data center efficiency.

Analytics (increasing understanding)

The cornerstone of any software-defined system designed ultimately to become fully automated is analytics. Prediction through analytics enables infrastructure to be managed in a proactive manner, taking resilience and robustness to a higher level.

However, events (for example, network-driven or hardware-driven events) cannot always be predicted, so reactive mechanisms – triggered by anomaly detection systems – also need to be in place. Live, accurate information – from the physical infrastructure as well as from the logically composed hosts – must be taken into account by these mechanisms.

Application-aware orchestration (becoming smarter)

Infrastructure can behave in a smarter way when it knows what it is running. In other words, applications should be described through application profiling of functional and non-functional requirements toward the infrastructure. The composition and resource orchestration layer should then be responsible for deriving, creating and maintaining the associated infrastructure resources according to the application's needs.

Hardware planning and setup (the physical reality)

When setting up a data center environment, dimensioning and distribution of resources across racks and chassis go hand-in-hand. In an environment that is extremely flexible in terms of physical resource distribution, it is even more important to find the optimal physical distribution ("the sweet spot") of resources.

Distribution brings an associated networking cost, making it necessary to find a balance between benefit and cost. For example, separation of compute and memory at long distances might be possible, but an expensive interconnect technology will compromise the potential utilization benefits.

Initial resource distribution is performed based on a data center forecast of the workload that will run in the infrastructure. However, since workload patterns are highly likely to change over time, the physical redistribution of resources will also be beneficial. This may involve moving a certain memory pool from one rack to another, for example. Having systems in place to carry out projections, periodically assess resource distribution and propose updates will help to maximize the benefits of a flexible infrastructure.

WORKLOAD EXECUTION

Execution environments are deployed according to a specific infrastructure paradigm. As a result, changing infrastructure principles requires support from the execution side. Virtualization, OSs, and cloud platforms all need to be considered.

Virtualization (abstraction and portability)

The ability to realize resource slicing and sharing at the physical level is a key accomplishment, as it allows resources to be used more efficiently. However, running traditional virtualization technologies such as hypervisors and containers on top of logically composed hosts will still be necessary for a number of reasons. For example, virtualization technologies can provide hardware abstraction and easy portability of certain applications; and many platforms (cloud platforms, for example) have a close dependency on hypervisors today.

In the long run, it is nevertheless important to explore how virtualization technologies can best be integrated into this new type of infrastructure architecture.

OSs (executing properly)

From an execution environment perspective, the initial impact of a new hardware architecture is felt at the OS level. It is necessary to understand how to cope with this impact in terms of CPU scheduling, memory drivers, network drivers, storage drivers and scalability.

For example, a fully disaggregated environment (where the distance – or access time – between individual components can be heterogenous) will require the OS to schedule operations (for example, read or write) accordingly. Moreover, transparent resource scaling (including the addition and removal of components) and/or migration must be possible in order to take full advantage of the infrastructure's flexibility.

Cloud platforms (unlocking on-demand flexibility)

Current workload management platforms (such as cloud platforms like OpenStack and big data frameworks like Apache Hadoop) were built on traditional server-based architectures. For their integration to be maintained, logical hosts are required to be composed in advance, and to remain static. However, this does not fully exploit the potential of a flexible infrastructure.

It is therefore important to understand the alternative approaches to integration that are possible, which can make best use of such an infrastructure. One example is the ability for a cloud platform to interface with an SDI orchestration layer and to request logical components when required.

DATA AND APPLICATIONS

When an infrastructure becomes more flexible, scalable, robust and proactively manageable, it becomes important to understand how these characteristics can be leveraged by applications.

For example, if the infrastructure is robust enough and able to scale extremely effectively by creating logical instances with high capacity, the data center operator should consider whether the splitting of an application into several components is still required (other than for geographical reasons, for example).

WHAT ABOUT LEGACY INFRASTRUCTURE?

The transition to a new infrastructure paradigm should be as smooth as possible. There is a well-established infrastructure paradigm that cannot be neglected, and the coexistence of both is therefore mandatory.

From the physical infrastructure perspective, it is up to the SDI composition and resource orchestration layer to ensure the binding with legacy servers is possible. In other words, it should allow management of legacy hardware (with its inherent limitations).

If a cloud platform (OpenStack, for example) is used to manage the infrastructure and it is tightly integrated with the remaining management system in a data center, it must be possible to keep them integrated. Logical servers can be composed, and cloud platforms integrated in the same way as today.

CONCLUSION

The ever-growing need for new IT infrastructure, along with the pressure on companies to reduce TCO, should lead to a fundamental rethink of the nature of this infrastructure. The data center based on SDI and characterized by resource disaggregation, programmability and modularity is poised as the lead solution on the road toward hyperscale.

The ability to tailor infrastructure to a particular workload improves resource utilization, a key challenge in today's data centers. The possibility to replace components rather than whole servers reduces the cost of infrastructure upgrades, while programmable features contribute to reduced operational costs.

The benefits of this new architecture are not restricted to hardware and its management. While being backward-compatible, the architecture also allows for new approaches to building, managing and running services. But this introduces challenges in many areas:

Modular and programmable infrastructure: hardware disaggregation is a breakthrough and networking its key enabler, however the extension of how far disaggregation can and should go needs to be fully understood.

Composition and resource orchestration: making infrastructure flexible might be insignificant if proper decisions are not taken at higher levels in terms of scheduling and analytics, for example. Decisions are required for physical planning and setup as well as real-time adjustments.

Workload execution: changing infrastructure principles requires support from the execution side. Proper integration of OSs, hypervisors and cloud platforms leverages the benefits of infrastructure flexibility.

Data and applications: the advances along the entire stack will lead to an assessment of how data and applications can take advantage of them.

On account of the broad impact of this evolution, it is important to understand it fully. This paper explains a key part of the path toward hyperscale data centers that are more flexible, smarter and able to achieve increased resource utilization levels while reducing TCO.

GLOSSARY

CPU	central processing unit
HDD	hard disk drive
HDI	hardware-defined infrastructure
NIC	network interface card
OS	operating system
SCSI	Small Computer System Interface
SDI	software-defined infrastructure
SSD	solid-state drive
TCO	total cost of ownership

REFERENCES

[1] ZDNet, IT spending 2016: The cloud drives software and data center increases, January 2016, available at: http://www.zdnet.com/article/it-spending-2016-the-cloud-drives-software-and-data-center-increases/

[2] Intel, The Intel® Rack Scale Architecture Vision, accessed May 2016, available at: http://www.intel.com/content/www/us/en/architecture-and-technology/rack-scale-architecture-vision-brochure.html

[3] Chaowei Phil Yang and Qunying Huang, Spatial Cloud Computing: A Practical Approach, CRC Press, 2013.

[4] Sangjin Han, Norbert Egi, Aurojit Panda, Sylvia Ratnasamy, Guangyu Shi and Scott Shenker, Network Support for Resource Disaggregation in Next-Generation Datacenters, Proceedings of the Twelfth ACM Workshop on Hot Topics in Networks (HotNets-XII), ACM, New York, US, 2013.

[5] Business Wire, Luxtera Debuts 1310nm 100G-PSM4 QSFP28 Module and Silicon Photonics Chipset at OFC 2015, March 2015, available at: http://www.businesswire.com/news/home/20150323005646/en/Luxtera-Debuts-1310nm-100G-PSM4-QSFP28-Module-Silicon#.VRG203NmD2p.mailto

[6] Intel® Silicon Photonics Research, accessed May 2016, available at: http://www.intel.com/content/www/us/en/research/intel-labs-silicon-photonics-research.html

FURTHER READING

[1] Mainstay, An Economic Study of the Hyperscale Data Center, January 2016, available at: http://cloudpages.ericsson.com/hubfs/Content-Offers/Economic-Study-Hyperscale-Datacenter.pdf