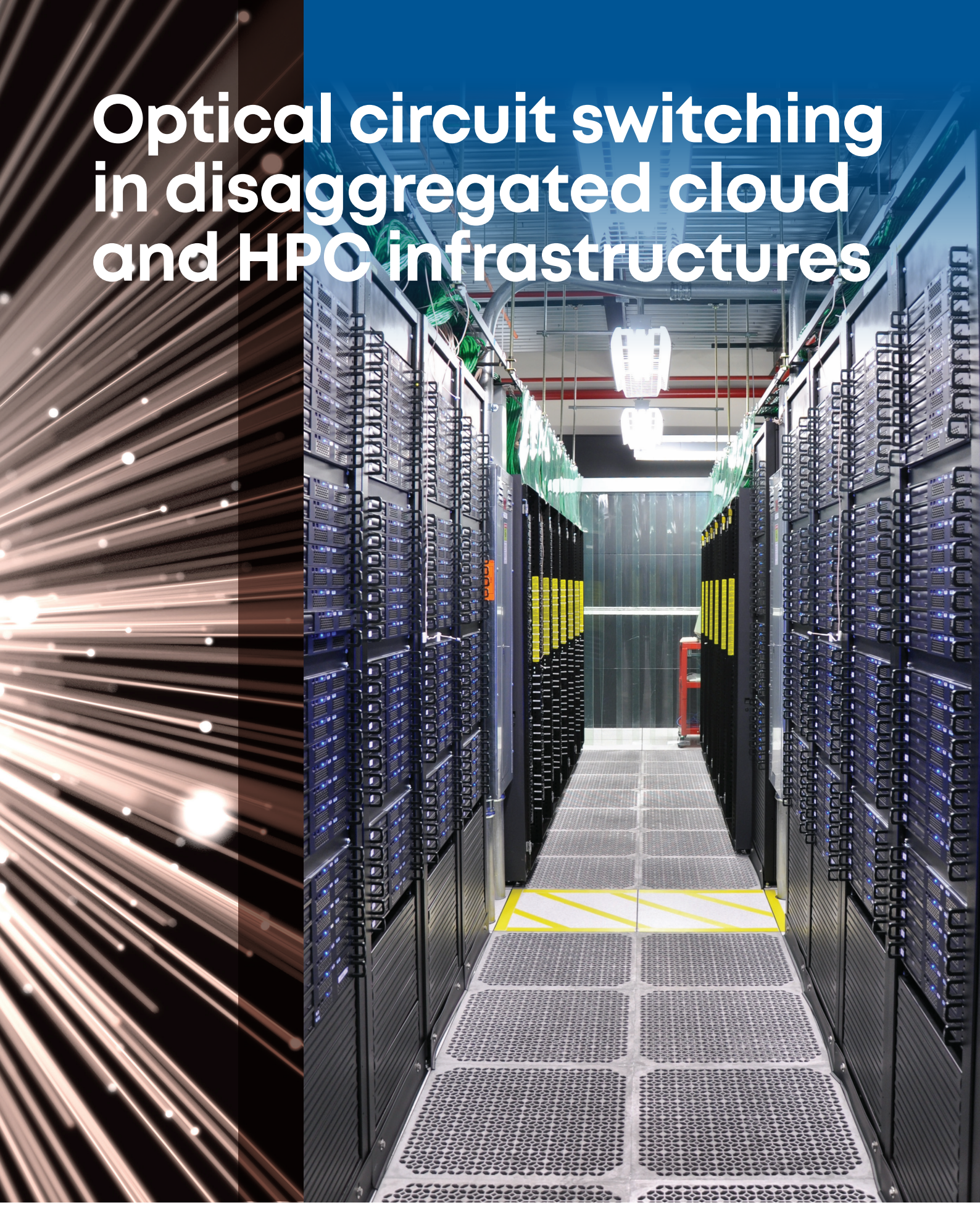
Optical circuit switching in disaggregated cloud and HPC infrastructures



White Paper

HUBER+SUHNER

AI, Machine Learning and Big Data driving performance needs

As applications such as Artificial Intelligence (AI) have moved from the academic and government domains into the commercial mainstream to drive the next generation of consumer and industrial applications, the pressure has grown on the Hyperscalers in cloud computing and other providers of high-performance computing (HPC) services to architect and scale their computing platforms to meet this client demand.

However, to remain competitive and ride this wave of growth, these providers need to find ways to control CAPEX and reduce power requirements in the light of environmental pressures and burgeoning demand on power grids.

The demands of applications such as training on large language models (LLM) or libraries of medical imaging, or sifting through mountains of raw intelligence data, have forced a re-evaluation of the structure of these high performance cloud computing networks, where the processing power required has increased by orders of magnitude. This in turn has seen the rise of accelerators such as GPUs that have to be deployed in large clusters.

Solution

Disaggregation of resources holds the key to lower costs and reducing power

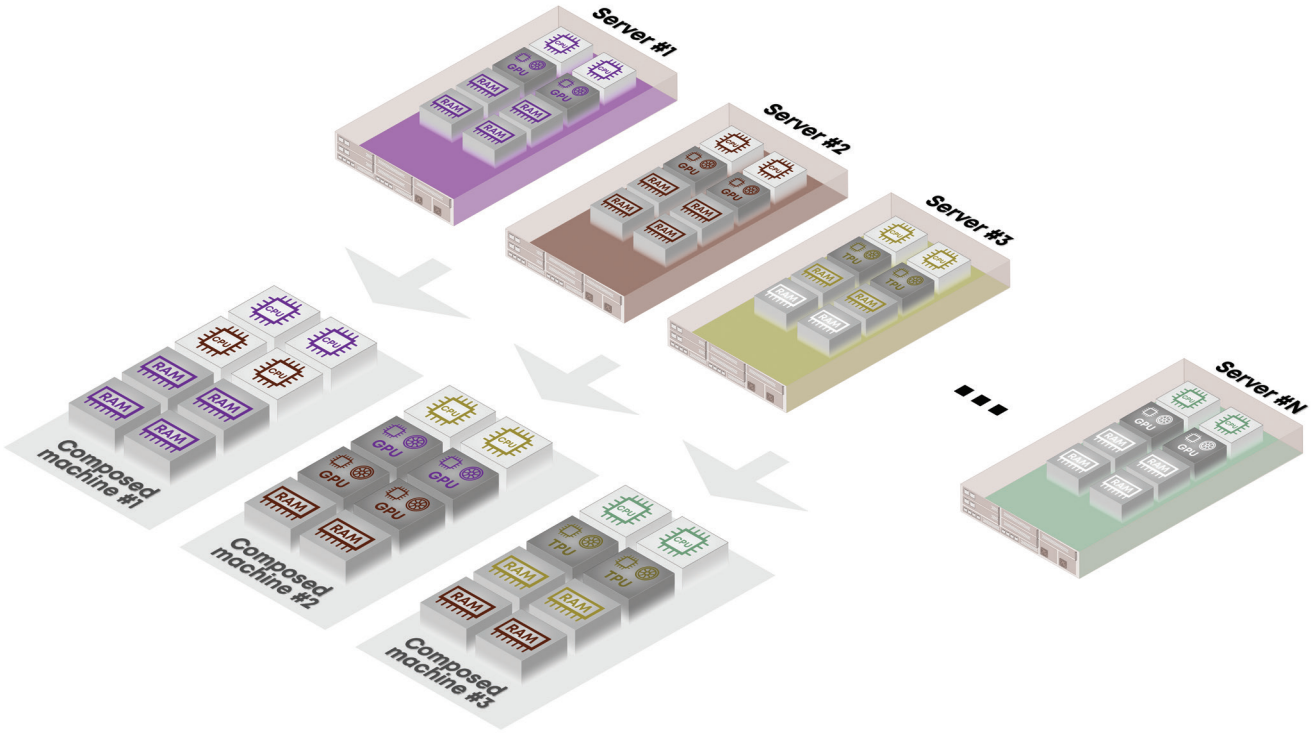
The building blocks of a typical hyperscale cloud or HPC platform are tightly and often inflexibly bundled in a common monolithic platform such as a standard server chassis. Growing the infrastructure to meet new application demands simply by adding more servers leads to increased inefficiency and underutilisation of some of the underlying compute and storage resources, and importantly excessive power consumption.

An innovative way of implementing flexible compute infrastructures that overcome those issues has emerged over the last fifteen years. It is based on the paradigm of **'disaggregating'** the requisite component parts or sub-systems.

A disaggregated architecture is one in which the key building blocks are flexibly combined by interconnecting them using integrated high-speed digital transceivers and a dedicated interconnect fabric based on appropriate transport media and switching technologies. These key building blocks, such as processors, memory, storage and accelerators can then be combined and appropriately scaled, independently of each other, to meet the demands of the expected workloads.

Disaggregated compute platforms at a glance

The principle of disaggregating HPC and cloud computing platforms is shown in the figure below.



Solution

Disaggregated computing platforms

Flexible resource utilisation

In place of conventional server structures, the required resources are bundled together in bespoke ratios to form flexibly proportioned 'bare metal' hardware hosts. This can be seen as a form of 'hardware virtualisation' as multiple physical hosts can be 'composed' on-the-fly, usually under the control of a dedicated orchestration function, using a common pool of underlying finely grained resources. Such platforms can then be used to support single tenants or multiple tenants using 'virtual machines' (VMs) through conventional 'software virtualisation' or cloud technologies.

The key building blocks in this case are resource elements that comprise computer hardware used to support high performance or commodity cloud computing applications, namely CPUs, memory, storage and various kinds of acceleration hardware such as GPUs and FPGAs.

In the figure on page 3, Composed Machine #1 is based on four units of CPU which are sourced from Servers #1 and #2 in the resource pool and four units of RAM which are sourced only from Server #1. The two units of GPU resource in Server #1, which in a non-disaggregated platform would be stranded and unused, are in this case used to supply half of the GPU resource required by Composed Machine #2 with the other half coming from Server #2.

The efficiencies and performance optimisations that can be achieved using this form of computer hardware disaggregation will depend on the granularity at which the resource blocks can be accessed and consumed. In its most granular form, this implies that each resource block (e.g. a bank of DRAM, a CPU, an accelerator) has onboard hardware to facilitate the necessary high-speed, low-latency connection of its resources to the interconnect platform. Clearly this will require new hardware designs and manufacturing of these resource blocks.

However, as explained below, there is also the possibility to deploy less granular forms of resource disaggregation that are more compatible with current hardware implementations and may be seen as a way to facilitate a more gradual transition towards fully disaggregated platforms.

The interconnect fabric

In a disaggregated compute platform the technology selected to build the interconnect fabric will have a significant effect on the overall system performance and power consumption.

The fabric itself could take a number of forms including:

- Packet switching.
- Electrical cross-point switching.
- Transparent optical switching.

The latter two provide deterministic, circuit-switched, fixed bandwidth data paths which are clearly well suited to interconnect hardware resource elements that would otherwise be directly and deterministically interconnected at a low level on a server motherboard or via a specific bus technology such as PCI Express.

Optical interconnect fabric

The benefits of choosing an optical interconnect fabric

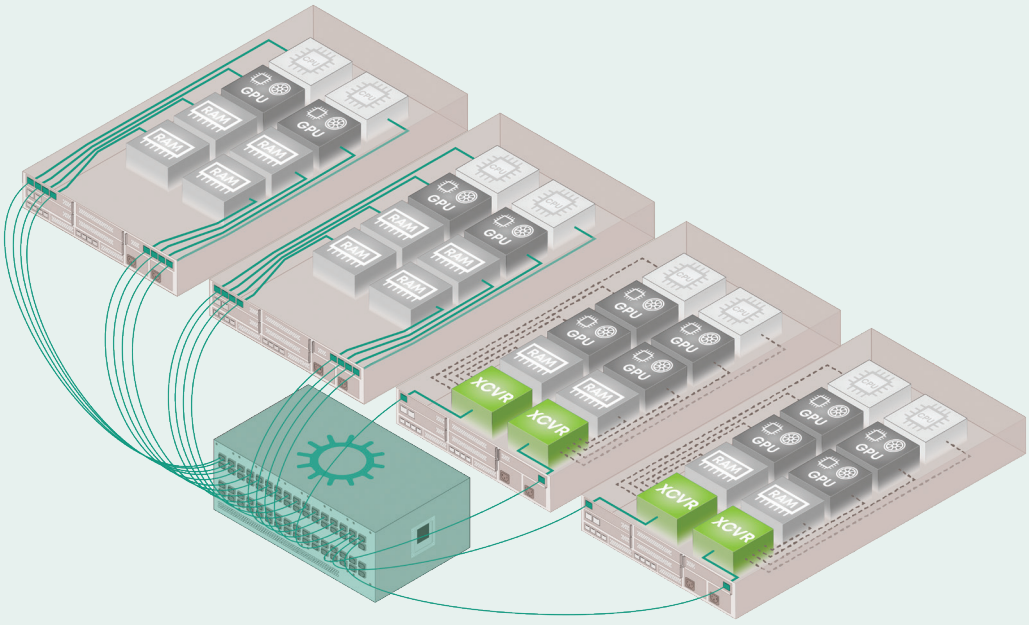
Using an all-optical (or photonic) interconnect fabric built from optical circuit switches, also often referred to as all-optical or photonic cross-connects, meets the above requirement of implementing deterministic, very high capacity end-to-end data paths.

Furthermore, an optical interconnect fabric brings with it many other benefits. These include:

- Significant reductions in power consumption of the fabric itself compared to an electrical fabric.
- Much lower latencies associated with the data paths through it.
- A much better ability to physically scale the fabric up and out.
- Significantly better future-proofing resulting from an inherent transparency of the fabric to the formats and line rates of the serialized data traffic between the disaggregated resource elements.

Minimising or eliminating Forward Error Correction from the optical links through the switching fabric is essential to minimise end-to-end latency through the fabric. This requires that the optical switching technology used exhibits the lowest possible losses to ensure the best bit error rate performance.

The figure below shows two methods by which disaggregated pools of server resources can be interconnected by an optical switching fabric for composition into flexibly proportioned 'hardware virtual machines'. Both are explained in more detail later.



Solution

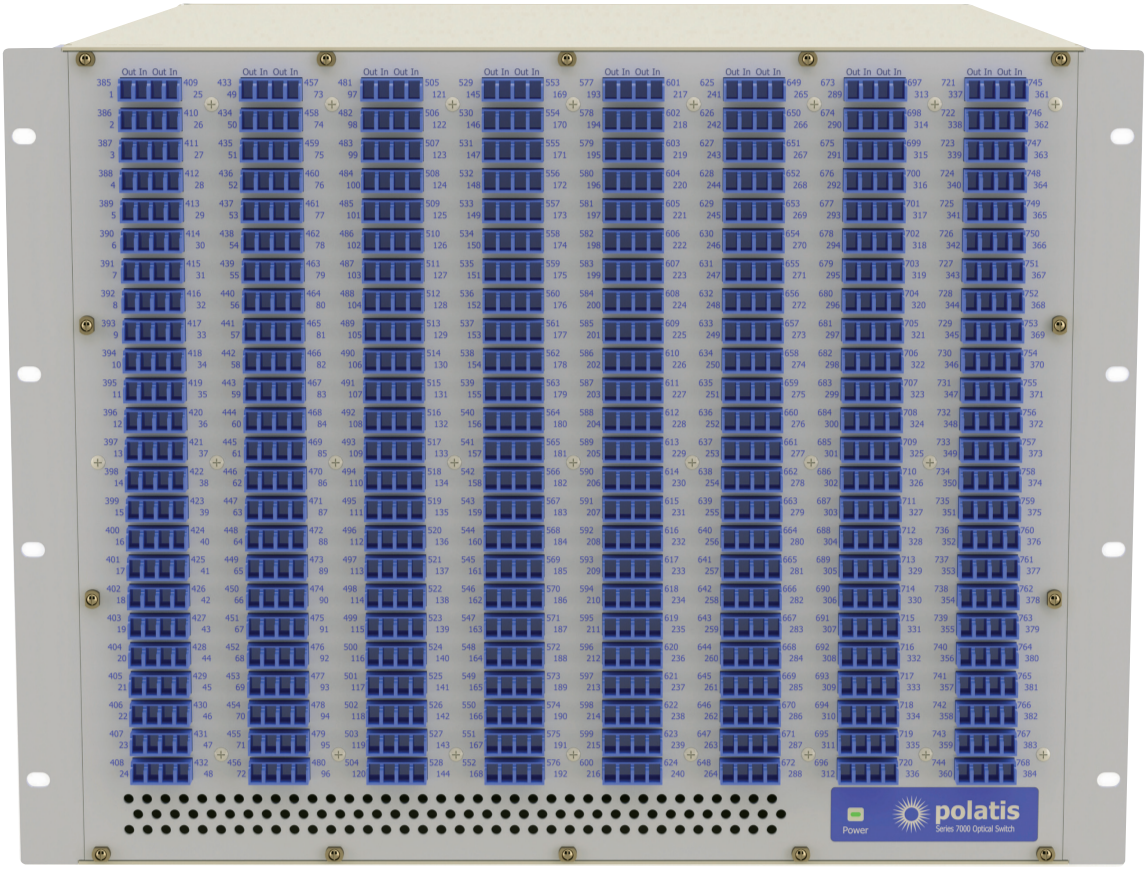
Optical circuit switching

Scaling the interconnect fabric

The scalability mentioned above derives from the fact that low-loss optical circuit switches can be used to build multi-stage switching fabrics that can in turn scale to support large numbers of optical end points.

The lowest loss optical circuit switches, such as POLATIS DirectLight™ switches, allow for fabrics to be constructed with up to four or more stages of switching whilst keeping within the optical loss budgets of typical transceivers used with disaggregated compute resource elements.

This allows optical switch fabrics to be architected and built up from larger numbers of smaller switches on a pay-as-you-grow basis, as an alternative to the more capital intensive approach of deploying a smaller number of high port count switches with the associated upfront CAPEX.



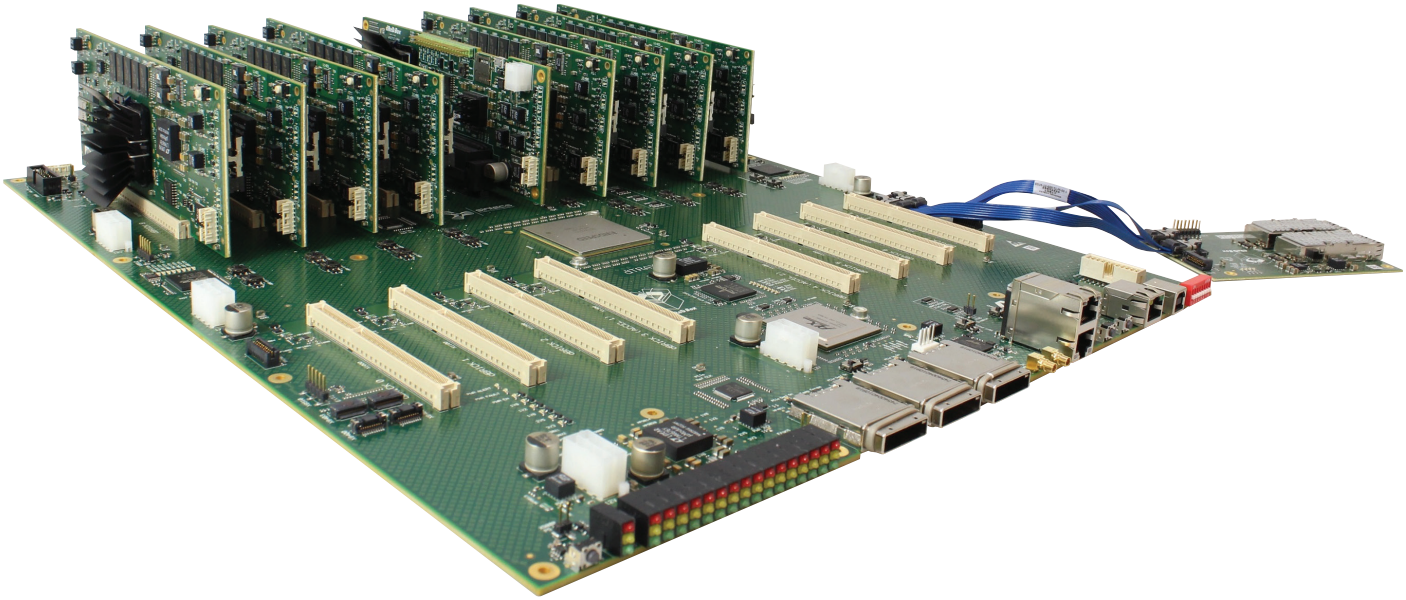
POLATIS 384x384 with LC connectors

Solution

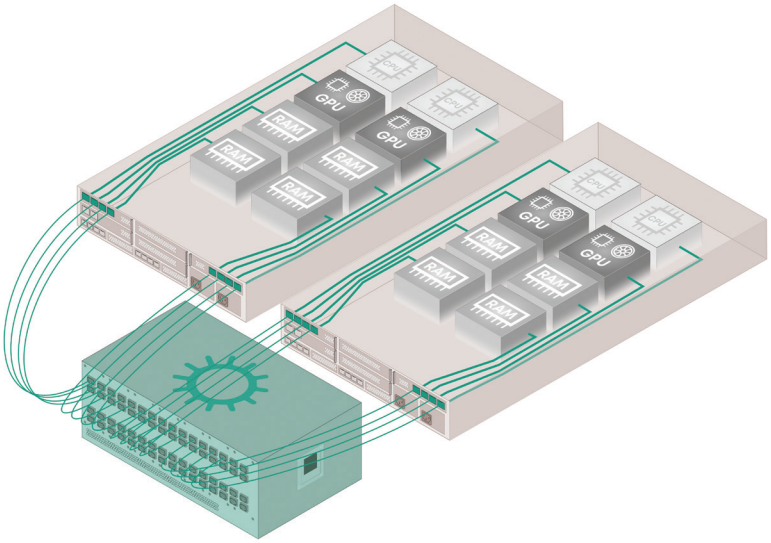
A realistic path towards disaggregated computing

Proof of concept

Between 2016 and 2019 HUBER+SUHNER Polatis explored disaggregation as part of a consortium of industrial and academic partners in a European Commission funded R&D project called dReDBox (disaggregated Recursive Datacentre in a Box) during which prototype hardware, orchestration software and representative user applications were successfully demonstrated.



A tray hosting up to 16 of the granular resource blocks (termed ‘bricks’) is shown above. Each brick was fitted with high capacity, multichannel silicon photonics-based, mid-board optics transceiver arrays that allowed the brick resources to be shared over a scalable low-loss optical circuit switching fabric. This represents the most granular form of resource disaggregation mentioned above and is illustrated in the figure on page 5 by the two leftmost server trays providing the pooled resources, and by the figure below.



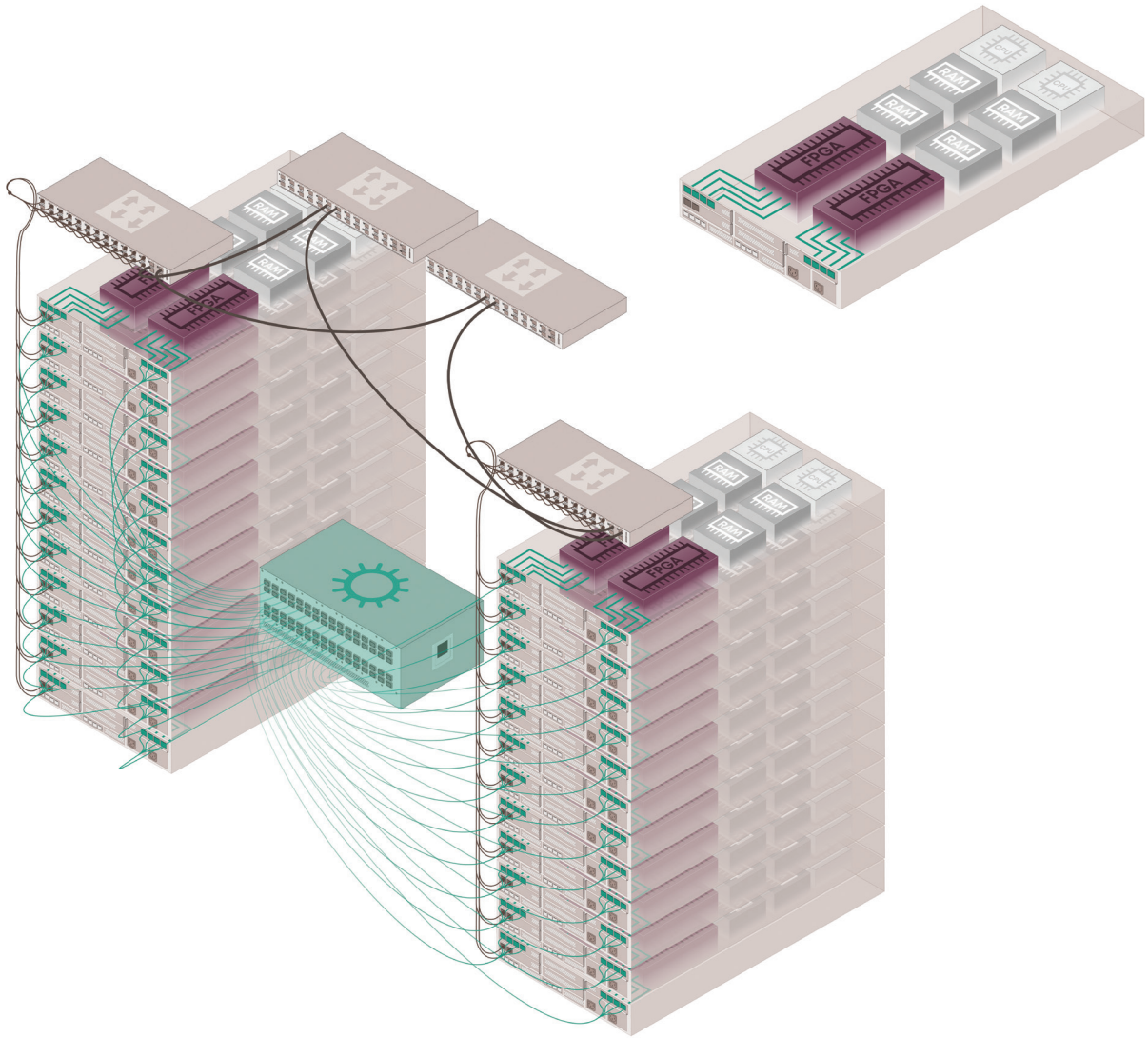
Optical interconnect overlaid on packet-switching fabric

The ultimate form of disaggregated computing exemplified in the dReDBox project is not the only option. There are a number of intermediate forms that can also act as stages of a more gradual migration towards a realisation of full disaggregation.

An application regularly being deployed today is one in which the dynamically interconnected compute resource components are limited to accelerator cards only - usually of the FPGA type. These cards are installed in standard hosts that will collectively comprise a HPC cluster or in dedicated chassis that are optimised to host multiple PCI Express (PCIe) cards.

However, they nearly always have a number of onboard standard pluggable transceiver cages, typically two or four in number. Where these cages are fitted with single mode optical transceivers they can be flexibly and directly interconnected with those in other hosts using a dedicated optical switching fabric that effectively acts as an 'overlay' to a packet switching fabric that will already be used to provide most of the interconnect between the hosts in the cluster.

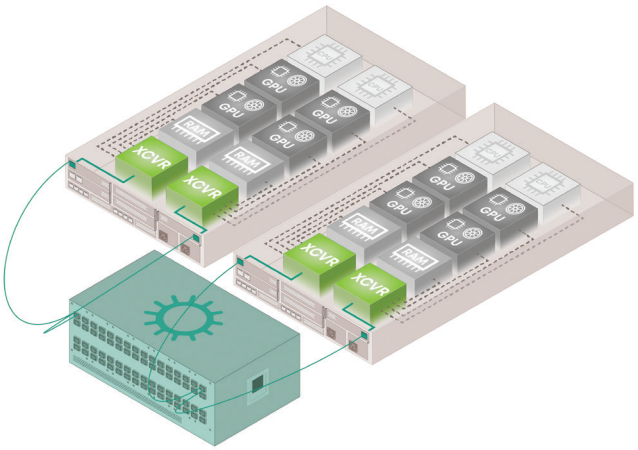
This hybrid architecture is illustrated below with a single server tray and then two racks of such servers. The packet interconnect fabric is shown above the racks and the optical switch in the centre.

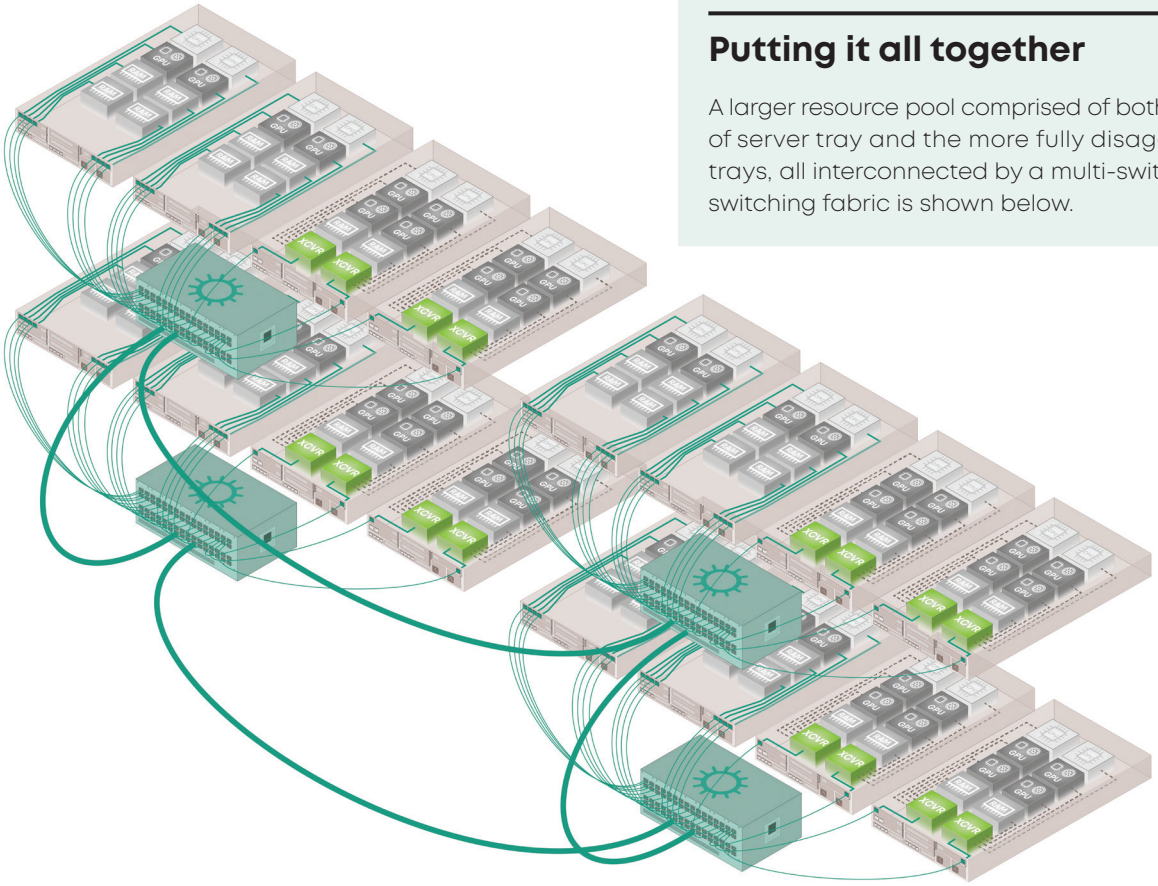


Repurposing conventional servers

Alternatively, or as a subsequent stage to the interconnection of FPGA accelerator cards alone, it is now also possible to realise a disaggregated compute platform by accessing more of the resources already present in fleets of conventional servers using an existing high performance bus technology like PCI Express. This involves installing a dedicated PCIe interconnection card fitted with specialised SerDes processing hardware and firmware and high-density, high-speed optical transceivers, all of which acts as a high performance gateway between the PCIe-connected compute resources in that chassis and the optical interconnect fabric.

This is illustrated by the two rightmost server trays in the figure on page 5 and by the figure to the right, in which the PCIe interconnection cards are represented by the green blocks labelled XCVR. The optical interconnect on these cards can again take the form of multiple cages accommodating standard pluggable modules such as QSFP-28 or, to realise the highest bandwidth densities, at least one vendor now deploys co-packaged optics.





Putting it all together

A larger resource pool comprised of both this type of server tray and the more fully disaggregated trays, all interconnected by a multi-switch optical switching fabric is shown below.

The benefits of disaggregated computing

A number of distinct benefits arise from flexibly interconnecting resources using a dedicated switching fabric.

Foremost amongst these are that:

- Hardware computing platforms can be composed on-the-fly.
- Platforms can be scaled to whatever size and ratio of the available resource types is appropriate to the kinds of workloads that will be run on the hardware.
- Platforms can be resized during the course of running a particular workload as the resource consumption requirements evolve.
- Resources not required can be powered down if they do not share common backplane or motherboard services, such as power supply rails and communications buses, resulting in OPEX savings.

Operators can:

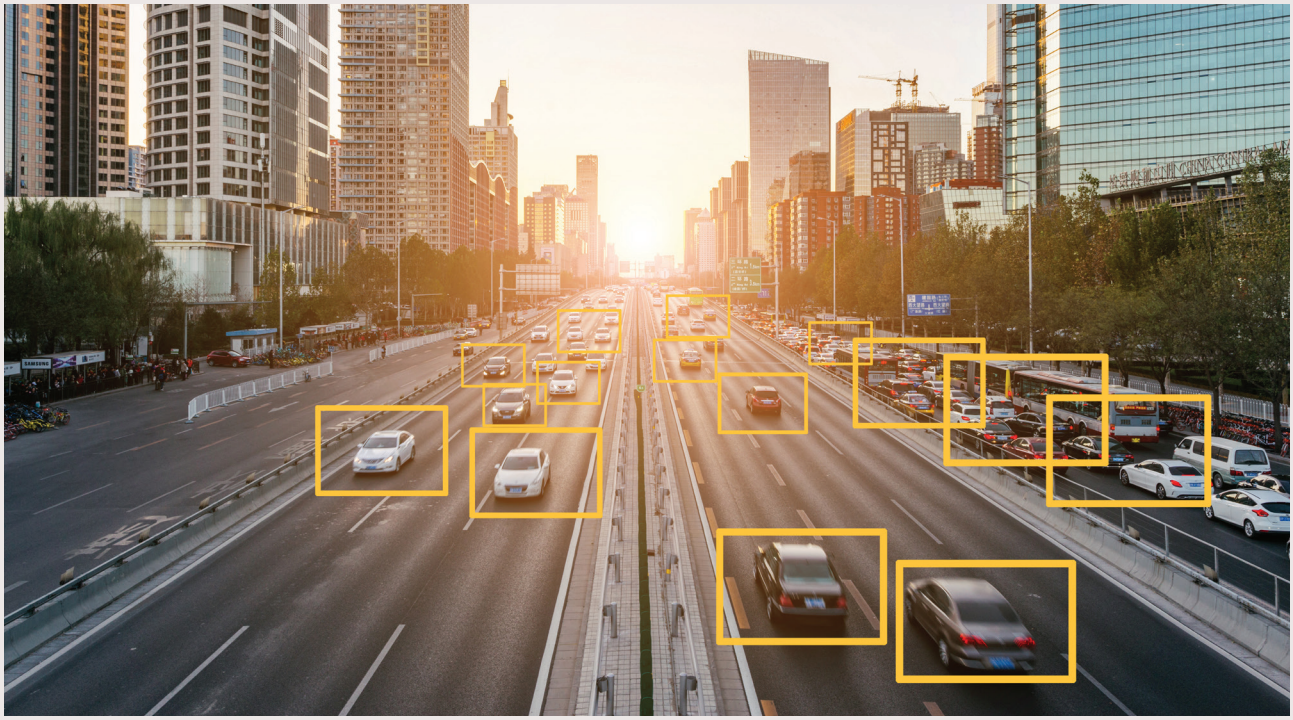
- Select best-of-breed vendors for the various component building blocks.
- Use those resources that support only the specific functions they need.
- Upgrade different types and/or blocks of resource element as and when required.

Together these can yield some substantial CAPEX savings as well as mechanisms to avoid excessive service downtime during hardware replacement programmes.

This is similar to some of the benefits recently described by one of the large Hyperscalers which has published a number of papers on the adoption of optical circuit switching into their intra-datacenter packet-based switching fabrics.

Hence a disaggregated platform is in many cases better suited to efficiently service the widely ranging demand profiles experienced by commodity cloud and HPC service providers than conventionally architected systems. Since the different hardware building blocks can be independently upgraded and the orchestration software is fully under the control of the operator, disaggregation can be seen as a key enabler of the transition from hardware-defined infrastructures (HDI) to much more agile software-defined infrastructures (SDI).

With the development of commercial products, the disaggregation of cloud compute/HPC platforms with optical circuit switching at its core has very much come of age. The benefits of operating such infrastructures are clear from a conceptual standpoint and from the numerous academic studies that have modelled expected system performance characteristics and operational costs.

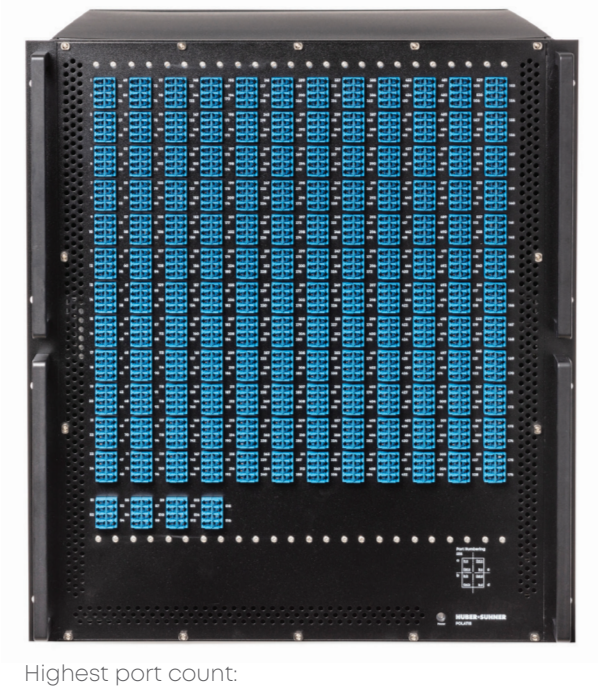


POLATIS® optical circuit switching for network disaggregation

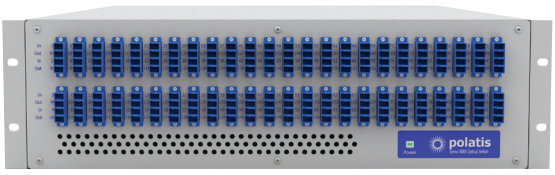
Advanced and proprietary fiber optic switching technology

POLATIS® has significant advantages over other all-optical (OOO) switching solutions for disaggregation, including:

- The industry's lowest optical loss and superior performance in stability.
- The broadest range of symmetric (NxN) switches, in matrix sizes from 16x16 to 576x576 ports, essential to support the evolving needs of network disaggregation, with modular scalability to connect thousands of fiber endpoints.
- High density switch matrices occupying very little rack space.
- Protocol and data rate agnostic so can switch signals of any type.
- Switching time <50ms for a single connection.
- Near-zero signal latency for fastest delivery to resources.
- True dark fiber switching, which requires no light to make and hold connections, enabling preprovisioning of future paths.
- Fully software-controlled for a seamless interface with leading orchestration solutions.
- Support for the broadest range of Software Defined Networking (SDN) interfaces including NETCONF and RESTCONF.
- Robust by design to be highly reliable for mission critical applications, with dual redundant, hot-swappable network interface controllers and power supplies.
- New POLATIS 576x576 switch also has dual redundant controllers and optional field addressable spare ports for increased resilience.
- Eco-friendly, low power consumption counterbalancing the high power density of disaggregated racks.



Highest port count:
POLATIS 576x576 optical circuit switch



Intermediate port count:
POLATIS 96x96 optical circuit switch

The HUBER+SUHNER advantage

The POLATIS team at HUBER+SUHNER has made a significant contribution to academic studies, projects and industry initiatives relating to disaggregation and worked with many of the pioneer vendors of disaggregation tools. Our network experts can provide advice and guidance on the best way to integrate optical circuit switches into disaggregated infrastructures.

HUBER+SUHNER offers a broad range of products for data centers such as fiber cables, patch cords, fiber management, structured cabling solutions, POLATIS® optical circuit switches, WDM components and more.

Worldwide sales and support are available to make sure data center systems continue to operate day in and day out.

HUBER+SUHNER
POLATIS® optical circuit switches
Americas: +1 781 275 5080
EMEA/Rest of World: +44 (0)1223 424200
info.polatis@hubersuhner.com
polatis.com
hubersuhner.com

4014/4115/03.2024