# Spatial Division Multiplexing for High Capacity Optical Interconnects in Modular Data Centers [Invited]

Matteo Fiorani, Massimo Tornatore, Jiajia Chen, Lena Wosinska and Biswanath Mukherjee

***Abstract*—Modular design has recently emerged as an efficient solution to build large data center (DC) facilities. Modular DCs are based on stand-alone prefabricated modules (i.e., PODs) that can be easily installed and interconnected. PODs can generate a large amount of traffic and thus require an ultra-high capacity interconnection network. However, current electronic and optical interconnect architectures applied to modular DCs may experience major scalability problems in terms of high energy consumption and cabling complexity. To address these problems, we investigate five optical interconnect architectures based on spatial division multiplexing (SDM), and for each architecture, we propose a resource allocation strategy. We also present an extensive comparison among the SDM architectures in terms of cost and performance (i.e., blocking probability and throughput), with the objective to find the architecture offering the best trade-off between cost and performance for given DC sizes and traffic load values. Our results demonstrate that, in small modular DCs with low traffic load, an architecture based only on SDM is the best option, while in medium DCs with medium traffic load, an architecture based on coupled SDM and flexgrid wavelength division multiplexing (WDM) with spectral flexibility is the best solution. Finally, for large DCs with high traffic load values, the best trade-off between cost and performance is achieved by an SDM architecture that is based on uncoupled SDM and flexgrid WDM.**

***Index Terms*—Spatial division multiplexing (SDM), optical interconnects, data centers, resource allocation, cost analysis.**

## I. Introduction

The growing adoption of cloud services is driving the demand for large data centers (DCs) hosting hundreds of thousands of servers [1]. These DCs are often located in remote areas in proximity of green energy sources that can guarantee continuous power supply and low energy cost. An example is the Facebook Arctic DC in Northern Sweden located close to the polar circle and entirely powered by locally produced hydro-electric energy [2]. Building and maintaining large DC facilities in remote areas might be challenging due to the high costs for transporting and installing electronic equipment. Recently proposed modular DCs [3] based on prefabricated stand-alone modules, referred to as PODs, which can be easily transported and installed, are considered as a promising solution to this problem. Each POD is composed of a predefined set of compute, storage and network resources, and it is optimized to guarantee high energy efficiency. The modular DC design ensures (*i*) better economy of scale, (*ii*) faster deployment, and (*iii*) higher cost and energy efficiency with respect to conventional DC designs. It is reported in [3] that the modular DC market will grow at a compound annual rate (CAGR) of 31% in the next 5 years, with new modular solutions from major vendors entering the business. Examples are the *HP performance optimized data center* [4] and the *Cisco containerized data center* [5].

PODs can contain several hundreds or even thousands of blade servers, each equipped with network interface cards (NICs) operating at capacity of 10 Gbps or higher. For this reason each POD can generate a large amount of traffic (i.e., in the order of several Tbps). In addition, the traffic inside the modular data centers is expected to increase in the future with a very high compound annual growth rate [6]. Consequently, future modular DCs will require ultra-high capacity networks to interconnect the PODs. With today traffic levels, conventional DC interconnects based on electronic packet switches are very efficient and relativity cheap. However, it has already been shown that, in the future, electronic DC interconnects will suffer from major scalability problems, especially in terms of energy consumption and cabling complexity [7]. Optical switching architectures have been recently proposed to address these limitations in conventional DCs [7], [8]. Some of these approaches can be applied to modular DCs to offer higher scalability and energy efficiency. However, the optical interconnect architectures proposed so far might still not provide the ultra-high capacity required by modular DCs, and they do not solve the problem of cabling complexity.

Optical spatial division multiplexing (SDM) has been recently proposed as a solution to increase fiber transmission capacity. SDM refers to the controllable arrangement of optical signals in the spatial domain, and it is based on the use of fibers equipped with multiple spatial elements, e.g., multi-mode, multi-core or multi-element fibers [9]. SDM has high potential to solve the scalability problems in modular DCs because it ensures ultra-high capacity and low cabling complexity, in terms of reducing the amount of fiber cables required to interconnect the PODs. Moreover, given the relatively short reach of communications inside DCs, impact of physical layer impairments (e.g., crosstalk) in intra-DC interconnects can be negligible, which simplifies the manufacturing of SDM components for DCs and shortens their time to the market.

An optical interconnect architecture for conventional DCs

Manuscript received June, 2016.

M. Fiorani, J. Chen and L. Wosinska are with KTH Royal Institute of Technology, School of ICT, Stockholm, Sweden, email: fiorani@kth.se.

M. Tornatore is with Politecnico di Milano, Department of Electronics, Information and Bioengineering, Milan, Italy.

B. Mukherjee is with University of California Davis, Department of Computer Science, Davis, CA, USA.

based on SDM has been proposed in [10]. Here, a large port count (LPC) spatial switch is employed to interconnect the top-of-rack (ToR) switches inside the DC. In this architecture, SDM is used in place of wavelength division multiplexing (WDM) to reduce the cost of the network infrastructure. However, this solution might incur scalability problems when applied to modular DCs, due to the fact that PODs have much higher capacity requirements (i.e., in the order of several Tbps) with respect to conventional ToR switches. To obtain the required capacity for modular DCs, SDM can be combined with flexgrid WDM. Several schemes for combining SDM and flexgrid WDM have been proposed [11], [12], e.g., depending on whether optical signals on different spatial elements are coupled to each other to form spatial superchannels or not. Each scheme provides a different level of network flexibility and imposes a different level of complexity on the network architecture. Consequently, the choice of the SDM scheme will significantly impact both the performance and the cost of the DC network.

In [12] we analyzed four SDM schemes for modular DCs. For each scheme we proposed a possible network architecture and a resource allocation strategy. We then performed a preliminary assessment of the cost and performance of each architecture. This paper extends the work in [12] by (*i*) analyzing a new SDM architecture, (*ii*) providing more details about the network design and resource allocation strategies, and (*iii*) presenting an extended set of simulation results. The objective is to identify which of the proposed SDM architectures provides the best trade-off between cost and performance for a given traffic load and DC size (i.e., in terms of number of PODs).

## II. Related Works

The concept of using the spatial dimension to increase the fiber transmission capacity is several decades old [9]. Yet only recently, due to the expected capacity crunch and the technological advances, operators are looking into SDM for upgrading their network infrastructure [13], and there has been an increase in research addressing SDM network technologies. Significant progress has been reported on the realization of fibers supporting multiple spatial elements, such as multi-core and multi-mode fibers [9]. For instance the design of multi-core fibers with 19 cores was reported for the first time in [14]. More recently, research studies have targeted the design of optical systems (e.g., transceivers and switches) for SDM networking [15]–[17]. In [15], several optical transceivers supporting different SDM schemes (e.g., coupled and uncoupled SDM) have been proposed and analyzed. Also, a possible design of a spatial and wavelength selective switch has been reported. In [16], several switch architectures for supporting different SDM schemes have been proposed and compared in terms of complexity, flexibility and scalability. Finally, in [17] the authors have reported an extensive discussion of the challenges of deploying SDM transceivers and switches in different network scenarios. The authors have concluded that the first deployments of SDM network technologies are likely to be performed in DCs where, due to the short reach of the communications, physical layer impairments are not a significant problem. Other recent works have instead moved the focus towards the efficient spatial and spectral resource allocation in SDM networks with spatial and spectral flexibility [18]–[20]. Authors in [18] have proposed for the first time an integer linear programming formulation that optimizes the use of spatial and spectral resources in SDM networks. Meanwhile, authors in [19], [20] have compared spectral and spatial superchannel allocation policies for SDM networks, considering different SDM switching schemes and modulation formats. An important conclusion is that resource allocation strategies that prioritize the creation of spectral superchannels are more efficient than those that prioritize the creation of spatial superchannels. However, these studies also show that spatial superchannel allocation using joint switching can offer significant benefits in terms of cost savings.

To the best of our knowledge, the only SDM architecture for DCs available in the literature is reported in [10]. In this architecture, SDM is used in place of WDM to reduce the cost of the network infrastructure. However, this solution might not provide the ultra-high capacity required by large modular DCs. For this reason, in our previous work [12], we proposed and investigated four SDM architectures tailored specifically for modular DCs. In this paper we extend the work in [12] by presenting a new SDM architecture and a more detailed simulation study, aiming at understanding which SDM schemes offer the best trade-off between cost and performance.

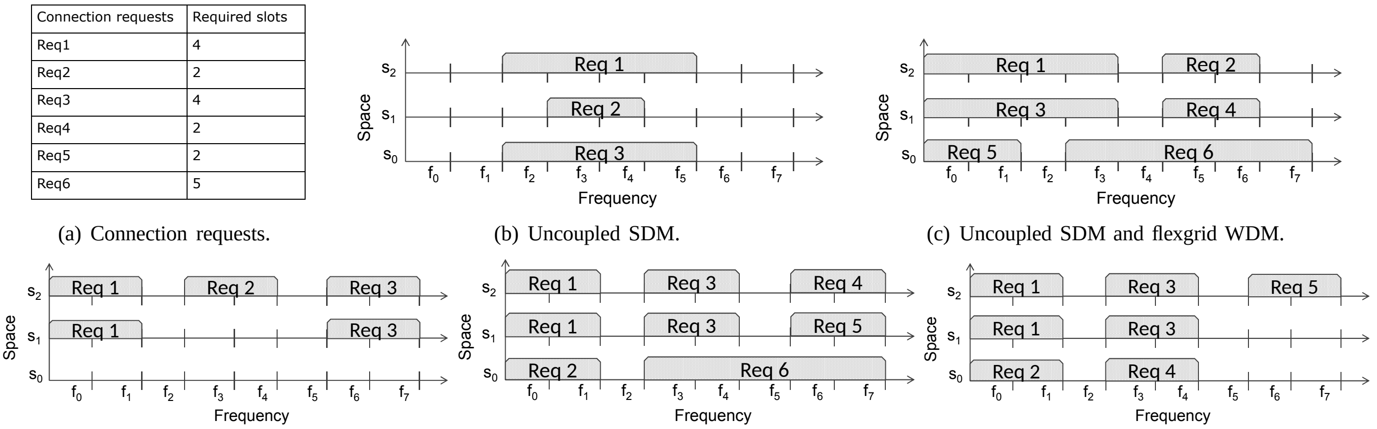
## III. SDM Switching Schemes

Our study analyzes five possible SDM switching schemes to be used in modular DCs. In the following, we describe each of the considered SDM schemes. We indicate with $N$ the number of spatial elements per fiber and with $M$ the number of spectral (frequency) slots per fiber.

The first scheme is referred to as **uncoupled SDM** and corresponds to the SDM solution proposed in [10]. In this SDM scheme, each spatial element carries a single independent optical signal (see Fig. 1(b)). Hence, SDM is used in place of WDM to establish parallel channels on the same fiber. The maximum number of channels that can be established on a single fiber is in this case equal to the number of spatial elements $N$ (in the example illustrated in Fig. 1(b) $N$=3). We assume that, by using flexible (i.e., bandwidth variable) transceivers, the capacity of each channel can be varied according to the traffic demand.

The second scheme is referred to as **uncoupled SDM and flexgrid WDM**. Here, each spatial element operates as an independent flexgrid WDM fiber where it is possible to establish multiple independent spectral superchannels (see Fig. 1(c)). The optical signals on different spatial elements are independent on each other (i.e., uncoupled). In this scheme a single fiber can carry up to $M \cdot N$ independent channels. This scheme represents the natural evolution of current flexgrid WDM transmission systems in the SDM domain and enables the reuse of conventional flexgrid WDM transceivers at the end-points of the communication.

The third scheme is referred to as **coupled SDM with spectral flexibility**. In this SDM scheme, spectral superchannels



| Connection requests | Required slots |
|---|---|
| Req1 | 4 |
| Req2 | 2 |
| Req3 | 4 |
| Req4 | 2 |
| Req5 | 2 |
| Req6 | 5 |

(a) Connection requests.

(b) Uncoupled SDM.

(c) Uncoupled SDM and flexgrid WDM.

(d) Coupled SDM with spectral flexibility.

(e) Coupled SDM with spectral and spatial flexibility.

(f) Coupled SDM with restricted spectral and spatial flexibility.

Fig. 1. Example of a set of connection requests (a) and their possible mapping to a multi-spatial element fiber using different SDM schemes (b-f).

can be expanded in the spatial domain to create spectral-spatial superchannels with increased capacity (see Fig. 1(d)). In a spectral-spatial superchannel, the optical signals on different spatial elements can be coupled to each other if new multi-input multiple-output (MIMO) optical transceivers are used at the end-points of the communication. We assume that each spectral-spatial superchannel is allocated all the spatial elements in the fiber, even if some of the spatial elements are not utilized. As a consequence, the flexibility is restricted only to the spectral domain, and the maximum number of parallel channels that can be established on the same fiber is $M$. This restriction is imposed to limit the complexity of the network.

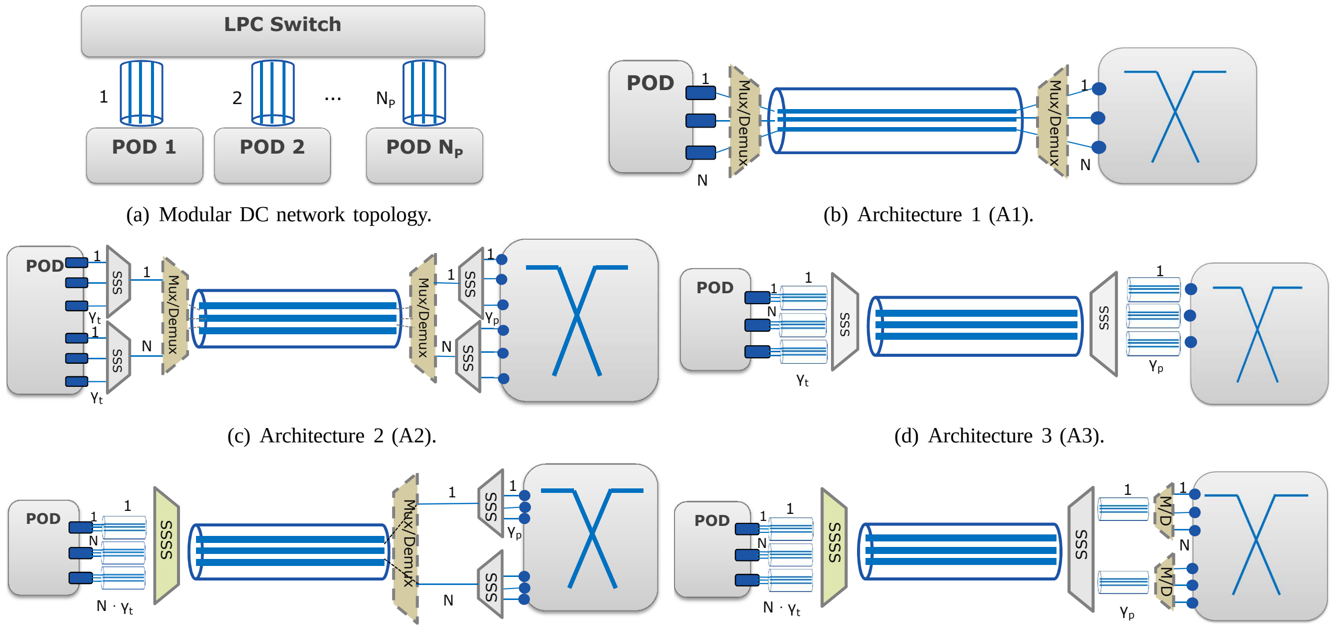
The fourth scheme is referred to as **coupled SDM with spectral and spatial flexibility**. In this case, we exploit unrestricted flexibility in both spectral and spatial domains (see Fig. 1(e)). Flexible spectral-spatial superchannels can be established, leading to the highest possible degree of network flexibility and enabling the implementation of advanced resource allocation schemes. However, this comes on the expense of higher network complexity. MIMO transceivers might be required to transmit and receive the flexible spectral-spatial superchannels, and complex switching components are needed within the network. Using this SDM scheme, it is possible to establish up to $M \cdot N$ parallel channels over the same fiber.

Finally, the fifth SDM scheme considered in this paper is referred to as **coupled SDM with restricted spectral and spatial flexibility**. This SDM scheme allows to establish flexible spectral-spatial superchannels, but with the restriction that the superchannels need to be organized in spectral groups (see Fig. 1(f)). The spectral-spatial superchannels that belong to the same spectral group utilize the same spectral resources (i.e., frequency slots) on one or multiple spatial elements. The spectral group restriction limits the complexity of the network, but at the expenses of lower flexibility. This SDM scheme can be seen as a combination between the other two coupled SDM schemes described before. The maximum number of parallel channels that can be established over the same fiber is $M \cdot N$. The use of this SDM scheme in modular DCs is analyzed for the first time in this paper.

To clarify how the resource allocation is performed in the five SDM schemes, we illustrate in Fig. 1 an example of how a set of connection requests can be accommodated over a multi-spatial element fiber using each scheme. The set of connection requests is shown in Fig. 1(a). Each connection request may require a different amount of spectral slots which can be distributed over one or multiple spatial elements. For simplicity, in this example, we assume $N$=3 and $M$=8. We also assume that each superchannel requires one spectral slot as guard band. The guard bands limit the maximum number of requests that can be served over the same fiber using the different SDM schemes. Fig. 1(b) shows how the connection requests can be served using uncoupled SDM. Due to the fact that WDM is not utilized, only the first three connection requests can be accommodated, one over each spatial element. Fig. 1(c) illustrates how the connection requests can be served using uncoupled SDM and flexgrid WDM. In this case six spectral superchannels are utilized to serve successfully all the requests. Fig. 1(d) shows that, using coupled SDM with spectral flexibility, only the first three connection requests can be accommodated by generating three spectral superchannels expanded in the spatial domain. It is clear from this example that, using this SDM scheme, the spatial dimension is not used efficiently. Fig. 1(e) shows an example of how the six connection requests can be accommodated using coupled SDM with spectral and spatial flexibility. Here, two spectral-spatial and four spectral superchannels are utilized to serve all the requests. Finally, Fig. 1(f) shows how the connection requests can be served using coupled SDM with restricted spectral and spatial flexibility. Here, only five connection requests can be accommodated using two spectral-spatial and three spectral superchannels.

## IV. SDM Architectures

In this section we propose a possible reference network architecture and resource allocation strategy, for each SDM scheme described in Section III. The reference scenario is a modular DC in which the PODs are connected to each



(a) Modular DC network topology.

(b) Architecture 1 (A1).

(c) Architecture 2 (A2).

(d) Architecture 3 (A3).

(e) Architecture 4 (A4).

(f) Architecture 5 (A5).

Fig. 2. Reference modular DC network topology (a) and proposed SDM architectures (b-f). (b) A1 supports uncoupled SDM; (c) A2 supports uncoupled SDM and flexgrid WDM; (d) A3 supports coupled SDM with spectral flexibility; (e) A4 supports coupled SDM with spectral and spatial flexibility; (f) A5 supports coupled SDM with restricted spectral and spatial flexibility. LPC: large port count.

other through a single optical LPC switch, as illustrated in Fig. 2(a). Each POD is connected to the LPC switch with a single bidirectional fiber that supports $N$ spatial elements and $M$ spectral slots.[1] We indicate with $N_p$ the number of PODs in the DC.

Fig. 2(b) shows architecture 1 (A1) which is designed to support uncoupled SDM. Each POD is equipped with $N$ flexible transceivers whose bandwidth can be varied to accommodate connection requests with different capacity demands. Spatial mux/demux are used to multiplex/demultiplex the parallel optical signals on the multi-spatial element fiber. A possible realization of the spatial mux/demux using 3D light waveguide technologies is reported in [10], [21]. This design is very compact and guarantees lower cost with respect to conventional mux/demux devices used in WDM networks, such as array waveguide gratings (AWGs). The LPC switch is equipped with $N$ ports per POD. Concerning the resource allocation strategy, for this SDM scheme, we consider a simple first-fit (FF) spatial element allocation. According to this strategy, for every new connection request between PODs, the first available spatial element is allocated both at source and destination fibers. Note that the spatial element used in the source fiber might be different than in the destination fiber.

Fig. 2(c) shows architecture 2 (A2) which is designed to support uncoupled SDM and flexgrid WDM. In this architecture, the PODs are equipped with flexgrid and tunable transceivers to transmit/receive spectral superchannels. To exploit the maximum degree of flexibility enabled by this SDM scheme, we assume that each POD is equipped with $N \cdot \gamma_t$ transceivers, where $\gamma_t = \min(M; N_p - 1)$. In this way, it is possible to establish up to $\gamma_t$ spectral superchannels on each of the $N$ spatial elements. The architecture requires $2 \cdot N$ spectrum selective switches (SSS) and two spatial mux/demux per each POD to multiplex/demultiplex optical signals on the multi-spatial element fiber and connect toward the LPC switch. The LPC switch is equipped with $N \cdot \gamma_p$ ports per POD, where $\gamma_p = \min(M; 2N_p - 1)$. As a resource allocation strategy, we propose a FF spatial and spectral allocation. This strategy is a straightforward extension of the FF spectral allocation utilized in conventional flexgrid WDM networks [22]. For every new connection request between PODs, the spatial elements and spectral slots are checked iteratively at both source and destination fibers. The iteration is stopped when a sufficient number of available contiguous spectral slots are identified in both fibers. The spectral slots on the source and destination fibers are required to be on the same frequency range, but can be on different spatial elements. This resource allocation strategy ensures: (*i*) spectral continuity, (*ii*) spectral contiguity and (*iii*) spectral non-overlapping.

Fig. 2(d) illustrates architecture 3 (A3) which is designed to support coupled SDM with spectral flexibility. Each POD is equipped with advanced flexgrid, tunable MIMO transceivers to transmit/receive the spectral-spatial superchannels. These transceivers are more complex with respect to the ones used in A2. A possible design of these transceivers is reported in [15], which is based on a single laser at both transmitter and receiver to transmit/recive over multiple spatial elements. To exploit the maximum degree of flexibility enabled by this SDM scheme, we assume that each POD is equipped with $\gamma_t$ transceivers.

[1] An electronic packet switch (not shown in Fig. 2(a) and out of the scope of our study) might be used in addition to the optical LPC switch to connect the PODs. The electronic packet switch can be used for instance to transmit short-lived traffic flows and to provide connectivity among PODs that cannot be directly connected through the LPC switch (e.g., because there are not enough optical resources).

Two large SSSs are employed to connect the PODs to the LPC switch, which comprises $\gamma_p$ ports per POD. Note that in A3 the number of transceivers and the number of LPC switch ports are not dependent of $N$, which makes A3 simpler than A2, A4 and A5. As a resource allocation strategy, we consider a conventional FF spectral allocation [22] used in flexgrid WDM networks. However, differently than in conventional flexgrid WDM systems, in A3 we exploit the spatial dimension to create spectral superchannels expanded over the spatial elements. This reduces the amount of spectral resources required to serve a given traffic demand. For example, with reference to Fig. 1(d), a connection request for six slots can be served using one spectral-spatial superchannel utilizing only two separate frequency slots and expanded over the three spatial elements. As a consequence, spectral resources can be saved with respect to a conventional flexgrid WDM system and more traffic can be carried over a single fiber.

Fig. 2(e) illustrates architecture 4 (A4) which is designed to support coupled SDM with spectral and spatial flexibility. Also in this case, PODs are equipped with advanced flexgrid, tunable, MIMO transceivers to transmit/receive the spectral-spatial superchannels. To guarantee maximum flexibility, the number of transceivers per POD is set to $N \cdot \gamma_t$. In addition, A4 requires a large spectral and spatial selective switch (SSSS) to connect each POD to the multi-spatial element fiber. The SSSS is a complex component that enables to route each spectral slot and each spatial element independently. Possible realization options for the SSSS have been studied in [15], [16]. On the other hand, a spatial mux/demux and $N$ SSSs can be employed to connect the multi-element fibers to the LPC switch, which is equipped with $N \cdot \gamma_p$ ports per POD. Concerning the resource allocation strategy, A4 allows for a large number of different approaches which have been widely investigated in the recent literature [20]. In [20], it is demonstrated that an approach that utilizes spectrum resources first (SpeF) is more efficient than an approach that utilizes spatial resources first (SpaF). For this reason in our work we assume for A4 the same resource allocation strategy as for A2, i.e., we try to accommodate each new connection request using a spectral superchannel. If this approach fails to identify free resources, we try to create spectral-spatial superchannels using an increasing number of spatial elements until a feasible solution is found to successfully serve the connection request. Note that serving a request using a spectral-spatial superchannel introduces an additional constraint with respect to the ones discussed for A2, which is referred to as non-spatial-separation constraint. This constraint is enforced to ensure that the spectral-spatial superchannels are routed correctly through the LPC switch, i.e., all the optical signals over different spatial elements are routed to the same destination fiber. In A3, this constraint is not necessary because all the spatial elements are always routed together and spatial separation is impossible.

Fig. 2(f) illustrates architecture 5 (A5) which is designed to support coupled SDM with restricted spectral and spatial flexibility. Each POD is equipped with $N \cdot \gamma_t$ flexgrid, tunable, MIMO transceivers. An SSSS is required to connect each POD to the multi-spatial element fiber. A single SSS and $\gamma_p$ spatial mux/demux can be utilized to connect the fiber to the LPC switch. This is possible due to the fact that in A5 the spectral-spatial superchannels are organized in spectral groups. Comparing A4 and A5, it can be observed that the main difference is in the way the fibers are connected to the LPC switch: A4 requires a single spatial mux/demux and $N$ SSSs, while A5 requires a single SSS and $\gamma_p$ spatial mux/demux. The spatial mux/demux is a much simpler and less expensive component with respect to SSS and, consequently, A5 is less expensive than A4. The LPC switch in A5 is equipped with $N \cdot \gamma_p$ ports per POD. Regarding resource allocation, A5 enables a large number of possible strategies. However, the spectral group constraint limits the flexibility with respect to A4 and makes the implementation of SpeF strategies that try to accommodate new requests using only spectral superchannels less efficient. As a consequence, in our work, we employ a SpaF strategy that tries to maximize the use of the spatial resources. Accordingly, for each new connection request, we apply the same resource allocation strategy as described for A3. If this strategy fails to identify free resources, we try to create spectral-spatial superchannels using an increasing number of spectral slots until a feasible solution is found. If the request is served using a spectral-spatial superchannel, the non-spatial-separation constraint needs to be enforced to ensure that the optical signals are routed correctly through the LPC switch.

## V. SDM Network Modeling

In this section, we first present an analytical cost model developed to assess the cost of the proposed architectures, and then we present a traffic model used in the simulation experiments that were carried out in order to evaluate the performance of the architectures.

### A. Cost Model

The cost of the SDM architectures is obtained by summation of the cost of the required network components. We assume that the costs of the spatial mux/demux and the switching elements (i.e., SSS, SSSS and LPC switch) depends linearly on their number of ports[2] Based on this assumption, the cost of A1 can be obtained using the following formula:

$$C_{A1} = N_p \cdot N \cdot (C_{tr}^{f} + C_{sp} + 2 \cdot C_{sm}), \tag{1}$$

where $C_{tr}^{f}$ is the cost of a flexible (i.e., bandwidth variable) transceiver, $C_{sp}$ is the cost per port of an LPC switch and $C_{sm}$ is the cost per port of a spatial mux/demux. Similarly, the cost of A2 can be obtained using the following formula:

$$\begin{aligned} C_{A2} = N_p \cdot N \cdot (\gamma_t \cdot (C_{tr}^{f,t} + C_{sss}) + \\ \gamma_p \cdot (C_{sp} + C_{sss}) + 2 \cdot C_{sm}), \end{aligned} \tag{2}$$

where $C_{tr}^{f,t}$ is the cost of a flexgrid and tunable transceiver and $C_{sss}$ is the cost per port of an SSS. The cost of A3 is given by the following equation:

$$C_{A3} = N_p \cdot (\gamma_t \cdot (C_{tr}^{f,t,m} + C_{sss}) + \gamma_p \cdot (C_{sp} + C_{sss})), \tag{3}$$

[2] We make this simplifying assumption because our objective is to evaluate the relative cost difference of the proposed SDM architectures, and not their exact cost values. Our model takes into account the difference among the number and the complexity of the components that are required in the SDM architectures..

where $C_{tr}^{f,t,m}$ indicates the cost of a flexgrid, tunable, MIMO transceiver required to transmit/receive spectral-spatial super-channels. It can be seen from formula (3) that the cost of A3 is independent on the number of spatial elements $N$. In practice the cost of a MIMO transceiver $C_{tr}^{f,t,m}$ might be dependent on the number of spatial elements. However, we assume that using the design in [15] the dependence on $N$ will be marginal; and for this reason, we consider a fixed value for $C_{tr}^{f,t,m}$ (see Tab. I). The cost of A4 can be obtained through the following equation:

$$\begin{aligned} C_{A4} = N_p \cdot N \cdot (\gamma_t \cdot (C_{tr}^{f,t,m} + C_{ssss}) \\ + \gamma_p \cdot (C_{sp} + C_{sss}) + C_{sm}), \end{aligned} \tag{4}$$

where $C_{ssss}$ is the cost of an SSSS port. Finally, the cost of A5 is calculated using the following formula:

$$\begin{aligned} C_{A5} = N_p \cdot (N \cdot (\gamma_t \cdot (C_{tr}^{f,t,m} + C_{ssss}) + \\ \gamma_p \cdot (C_{sp} + C_{sm})) + \gamma_p \cdot C_{sss}). \end{aligned} \tag{5}$$

In our study, we consider the normalized cost values for the network components shown in Tab. I. Since components for SDM are not yet commercially available, we evaluate their cost using a forecast methodology based on the model proposed in [23]. The model in [23] can be used to estimate the cost of novel optical components based on their relative complexity with respect to commercially available ones. The model in [23] is applied to flexgrid optical transport networks, but it can be potentially utilized also to different network scenarios [24]. According to this model, the cost of a device based on a new technology is three times higher than the cost of the same device using the most advanced existing commercial technology. As a consequence, we estimate the cost of a MIMO transceiver ($C_{tr}^{f,t,m}$) to be three times higher than the cost of a commercial flexgrid transceiver ($C_{tr}^{f,t}$). Similarly, we assume that the cost per port of a SSSS ($C_{ssss}$) is three times higher than the cost per port of a SSS ($C_{sss}$). We also perform an extensive sensitivity analysis on the cost values in Tab. I to evaluate the dependency of our results on the input data. The main conclusions from this sensitivity analysis are discussed in Section VI. In the following, we describe the traffic model used for performance evaluation of the SDM architectures.

TABLE I
REFERENCE COST VALUES FOR THE PROPOSED SDM ARCHITECTURES [12], [23]. CU=COST UNIT.

| **Component** | **Cost [CU]** |
|---|---|
| Transponder (flexible) ($C_{tr}^{f}$) | 1 |
| Transponder (flexgrid, tunable) ($C_{tr}^{f,t}$) | 1.2 |
| Transponder (flexgrid, tunable, MIMO) ($C_{tr}^{f,t,m}$) | 3.6 |
| LPC Switch port ($C_{sp}$) | 0.8 |
| Spectral selective switch ($C_{sss}$) | 0.8 |
| Spatial spectral selective switch ($C_{ssss}$) | 2.4 |
| Spatial Mux/Demux ($C_{sm}$) | 0.001 |

TABLE II
MAPPING BETWEEN CAPACITY AND NUMBER OF SPECTRAL SLOTS. WE ASSUME DP-QPSK MODULATION FORMAT [22].

| **Capacity** | **Slots (12.5 GHz)** |
|---|---|
| 1 Gbps | 1 slot |
| 10 Gbps | 2 slots |
| 100 Gbps | 3 slots |
| 200 Gbps | 4 slots |
| 400 Gbps | 6 slots |
| 1000 Gbps | 16 slots |

### *B. Traffic Model for Modular DCs*

To the best of our knowledge, there is no well-accepted traffic model for modular DCs available in the literature. Our study relies on the data provided in [25], [26] that are based on measurements collected from a number of conventional DCs worldwide. According to these data, the traffic pattern in the core tier of conventional DCs[3] varies slowly over time (i.e., the variation is on the order of several seconds or higher) [25], [26]. In addition, the traffic is almost uniformly distributed among the aggregation switches in the core tier. In modular DCs, the network interconnecting the PODs presents similar characteristics as the core tier in the conventional DCs. This is due to the fact that, both the traffic among the aggregation switches (in conventional data centers) and among PODs (in modular data centers) is aggregated by the switches in the top of rack. As a consequence, our work assumes that the traffic in modular DCs (i.e., traffic demands between PODs) varies slowly with time. Hence, the transceivers and switches in the network are reconfigured periodically after fixed time intervals. These time intervals can be for instance on the order of several seconds or tens of seconds. Utilizing the estimated traffic pattern as an input, the spectral and spatial resources allocation is performed offline at the beginning of each time interval and the network elements (i.e., transceivers and switches) are configured accordingly.

To assess performance of the proposed SDM architectures, we developed a Monte Carlo simulator. The Monte Carlo simulator was implemented specifically for the purpose of our study using the C++ programming language. Based on the data in [25], [26] we assume that the traffic is uniformly distributed among the PODs. For this reason, we generate the traffic pattern at the beginning of each time interval assuming that each POD $i$ requires an optical connection through the LPC switch toward $x_i$ other PODs, where $x_i \in [0, N_p - 1]$ is a variable extracted from a random uniform distribution. In our simulation, we changed both the mean of the uniform distribution and the number of PODs in the DC ($N_p$) to evaluate the performance of the SDM architectures under different traffic loads and DC sizes. The capacity of each connection request is randomly distributed in the range of [1, 10, 100, 200, 400, 1000] Gb/s and follows a normal distribution with mean 100 Gb/s. Based on the consideration that inside DCs the transmission reach is short and physical layer impairments are not significant, we assume a fixed modulation format, i.e., dual-polarization quadrature phase shift keying (DP-QPSK). For spectral slots of 12.5 GHz, it is then possible to map each capacity to a number of required spectral slots. The considered mapping between capacity and spectral slots is shown in Tab. II. Observe that, based on [22], we assume

[3] The core tier is the network segment in charge of interconnecting the aggregation switches inside the DC among themselves and to the inter-DC network.

that the dependency between capacity and number of spectral slots is not linear.

## VI. Numerical Results

In this section, we present the numerical results showing the cost and the performance, in terms of blocking probability and throughput, of the proposed SDM architectures, based on the cost and simulation models described in Section V. We define the load as the average number of optical connections requested by each POD, normalized with respect to the total number of possible destinations ($N_p - 1$). The blocking probability is defined as the probability that a connection request between two PODs cannot be served because the required optical resources are not found in the network.[4] We assume that the maximum acceptable blocking probability is $10^{-2}$. This is based on the consideration that, when an SDM architecture offers a blocking probability lower than $10^{-2}$, it guarantees the almost full bisection bandwidth (i.e., the same bisection bandwidth of a non-blocking non-oversubscribed electronic packet switching network) [27]. On the other hand, when an SDM architecture offers a blocking probability higher than $10^{-2}$, it means that it is not able to guarantee anymore the full bisection bandwidth [27]. For example, a blocking probability of $10^{-1}$ indicates that the SDM architecture offers only 90% of the full bisection bandwidth. The throughput is defined as the amount of traffic carried by each SDM network architecture, and is obtained by subtracting the blocked traffic from the offered traffic. The sample size of the Monte Carlo simulations is 5000, and the presented results have a confidence interval not exceeding 5%, with 95% confidence level. Our study assumes $N$=10 and $M$=320. In addition, we assume that one spectral slot is used as guard band per each superchannel (as shown in Fig. 1).

Fig. 3 shows the blocking probability of the proposed SDM architectures as a function of the load and for different sizes of the DC. The boxes inside the graphs show the costs of the architectures (which depend on the size of the DC, but not on the load).

In Fig. 3(a), a modular DC of small/medium size ($N_p = 15$) is considered. This is the typical size of modular DCs owned by large private enterprises [28]. A1 shows some blocking probability for traffic loads higher than 60% while we didn't observe any request blocking in the other architectures. The blocking probability for A1 exceeds $10^{-2}$ only at high loads (i.e., higher than 75%) which means that A1 can perform well in normal DC working conditions. Regarding cost, a difference of some orders of magnitude can be observed between A1, A3 and A2/A4/A5. Specifically, A1 is 5 times cheaper than A3 and at least 40 times cheaper than A2/A4/A5. We conclude that A1 represents the best solution in this case.

In Fig. 3(b), a relatively large modular DC ($N_p = 75$) is analyzed. It could represent the DC owned by a medium-sized cloud provider. In this case A1 exhibits high blocking probability even at relatively low loads; thus it does not represent a feasible solution. A3 shows some request blocking probability at medium/high loads, which exceeds $10^{-2}$ for loads higher than 60%. On the other hand, we didn't observe any request blocking for A2, A4 and A5. From a cost perspective, A3 is more than 6 times cheaper than A2 and more than 10 times cheaper than A4/A5. We can conclude that A3 is the best option if the traffic load of the DC is normally lower than 60%. Otherwise A2 is the best option .

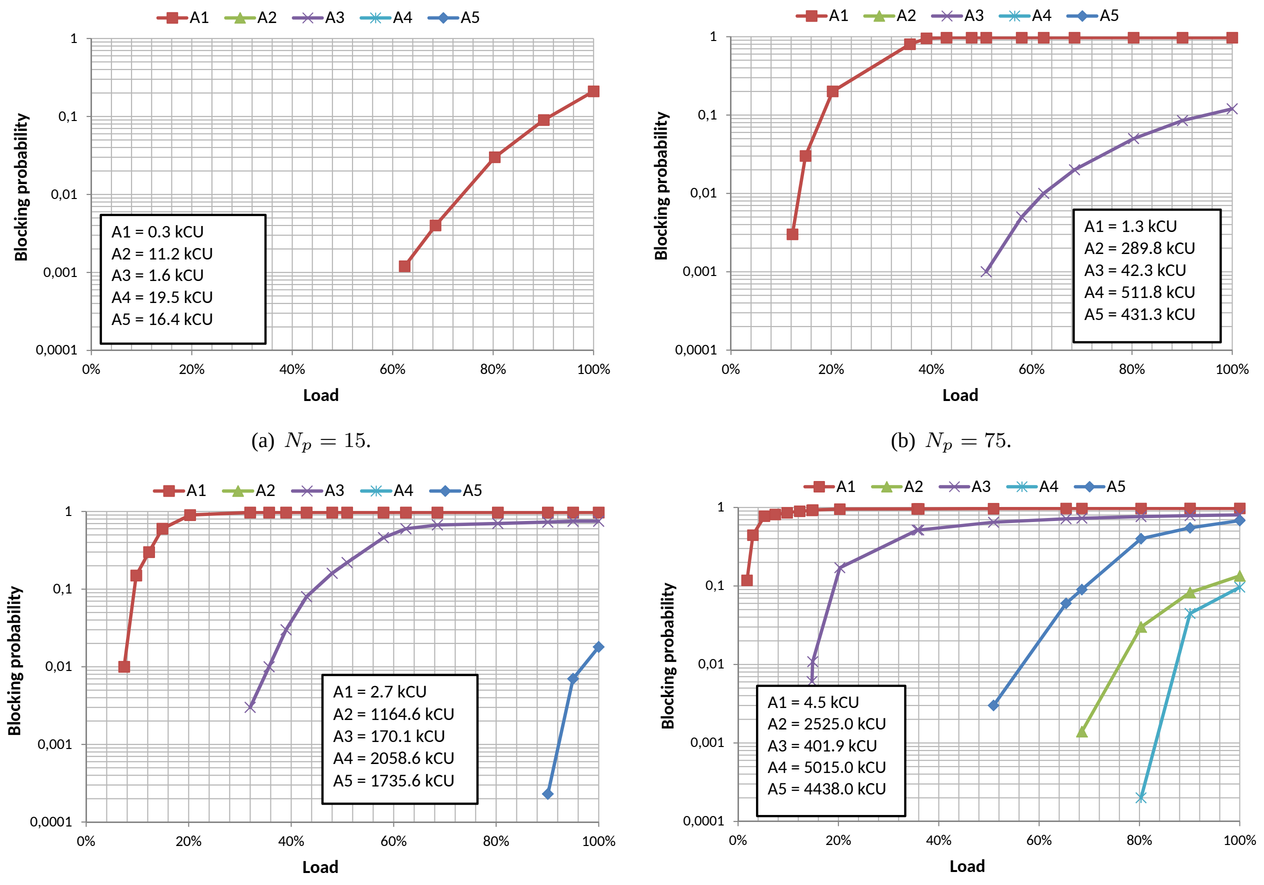
In Fig. 3(c), we show the blocking probability for a large modular DC ($N_p = 150$). This could represent the size of a DC currently owned by a large cloud provider (e.g., Facebook, Google and Microsoft). In this case, both A1 and A3 show high blocking probability even at low/moderate loads and thus they cannot be considered as feasible options. A5 exhibits some request blocking at very high loads (i.e., higher than 90%) while we didn't observe any blocking for A2/A4. In A5 the group restriction limits the flexibility leading to lower performance with respect to A2 and A4. The performance of A5 could be improved by defining a more efficient resource allocation strategy, but will most likely not reach lower blocking probability than A2 and A4. We plan to investigate this aspect and devise more advanced resource allocation schemes for A5 in our future work. Regarding cost, A2 is the cheapest among the three feasible architectures in this scenario (i.e., 32% cheaper than A5 and 43% cheaper than A4) and thus represents the best option. The reason for A2 being cheaper than A4/A5 is mostly due to the fact that A4 and A5 require expensive MIMO transceivers to transmit/receive spectral-spatial superchannels.

Fig. 3(d) shows the results for a very large modular DC ($N_p = 250$). This could represent the size of DCs that will be operated in the future by large cloud providers. In this case, A1 and A3 show very high blocking probability and are not feasible options. Also A5 show unacceptable blocking probability at medium/high loads and could be feasible only if the DC operates at relatively low loads. A2 performs well up to very high loads (i.e., A2 blocking probability is lower than $10^{-2}$ up to around 75% load). Finally, A4 shows the best performance and its blocking probability is lower than $10^{-2}$ up to 90% load. However, A2 is probably the best option because it performs well in realistic working conditions while it is 43% cheaper than A4.

Note that, for $N_p < (M-1)$, the ratio between the cost of A2/A3/A4/A5 and the cost of A1 increases linearly with $N_p$, while for $N_p \geq (M-1)$ it is almost constant. Similarly, for $N_p < (M-1)$ the ratio between any of A2/A4/A5 and A3 increases with $N_p$, while for $N_p \geq (M-1)$ the ratio becomes constant. This is due to the fact that, when $N_p \geq (M-1)$, the value of $\gamma_t$ is limited by the number of spectral elements in the fiber ($M$) and does not increase with increasing the number of PODs in the modular DC ($N_p$).

Fig. 4 shows the blocking probability of the proposed SDM architectures as a function of the size of the modular DC ($N_p$) and for two load values. In Fig. 4(a), the results with load equal to 30% are shown. It can be observed that A1 offers good blocking performance for relatively small DCs (i.e., up to $N_p = 20$) while A3 can support medium/large DCs (i.e.,

[4] As discussed in Section III we consider that if two PODs cannot be connected through the optical LPC switch they can still communicate using a parallel electronic packet switch. However, the electronic switch might not be able to provide the entire capacity required by the connection. The analysis of the performance of the electronic packet switch is out of the scope of this study.



(a) $N_p = 15$.

(b) $N_p = 75$.

(c) $N_p = 150$.

(d) $N_p = 250$.

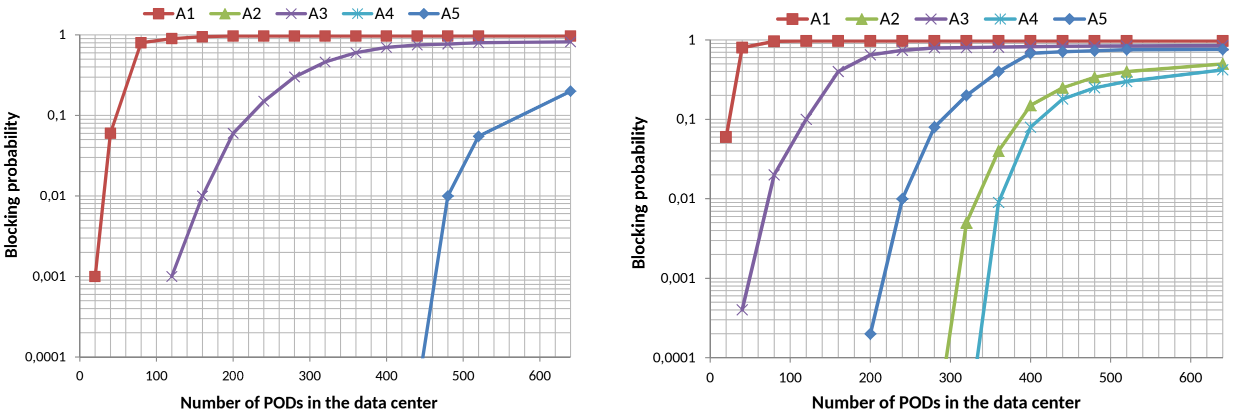
Fig. 3. Blocking probability of the proposed SDM architectures as a function of load and for different sizes of the modular DC. Boxes indicate the cost of the architectures ($kCU = CU \cdot 10^3$).

up to $N_p = 120$). On the other hand, A5 works well even for extremely large DCs (i.e., up to $N_p = 480$) while A2 and A4 can support any practical modular DC size.

In Fig. 4(b), the results with load equal to 60% are shown. It can be observed that A2 provides acceptable blocking performance only for very small DCs (i.e., up to $N_p = 8$) and A3 can support small/medium DCs (i.e., up to around $N_p = 50$). A5 provides good blocking performance even for large DCs (i.e., up to $N_p = 200$), but might not support very large future cloud DC sizes. Finally, A2 can be scaled to support very large future DCs with up to $N_p = 320$ and A4 can be further scaled to support up to $N_p = 340$. With more than 340 PODs in the DC, all the architectures exhibit blocking probability higher than $10^{-2}$.

In Fig. 5, we show the ratio between cost and throughput as a function of the size of the modular DC ($N_p$) and for different load values. For each architecture, the curve is terminated when the respective blocking probability becomes higher than $10^{-2}$. The boxes show the architectures that provide the best cost-throughput trade-off for a given range of DC sizes. These results summarize the main findings of our study. Fig. 5(a) shows the results with load equal to 30%. In this case A1 represents the best option for modular DCs with up to 20 PODs (e.g., private enterprise DCs); A3 is the best option for DCs with a number of PODs between 20 and 120 (e.g., medium-sized cloud provider DCs); and A2 is the best option for DCs with more than 120 PODs (e.g., large cloud provider DCs). Fig. 5(b) shows the results with load equal to 60%. In this case A1 represents the best option only for very small DCs with up to 8 PODs (medium/small private enterprise DCs); A3 is the best option for DCs with a number of PODs between 8 and 50 (e.g., large private enterprise or small cloud provider DCs); A2 is the best option for DCs with a number of PODs between 50 and 320 (e.g., medium and large cloud provider DCs); and A4 is the only option for DCs with a number of PODs between 320 and 340, while larger DCs would require some extensions in the proposed SDM architectures.

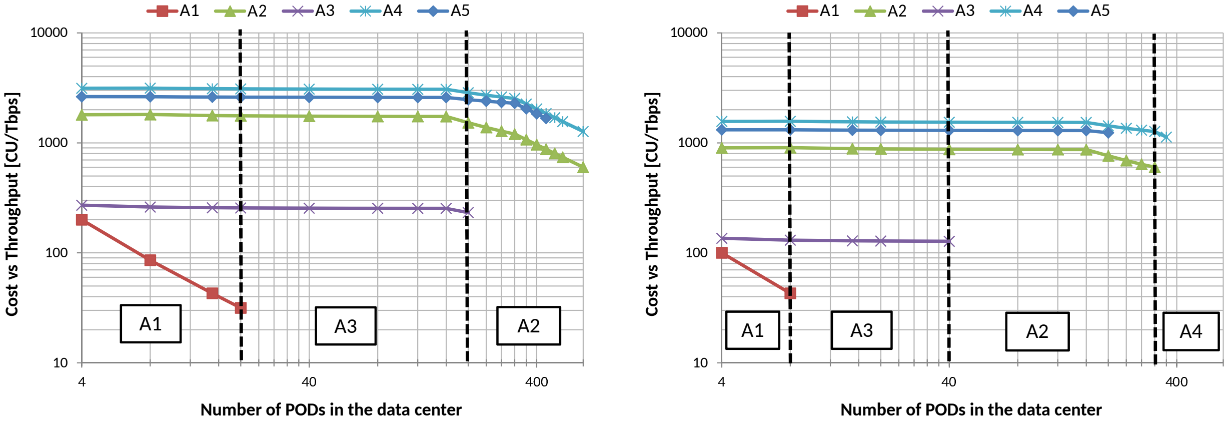
We performed an extensive sensitivity analysis on the cost of the architectures to check how much our results depend on the input values. We started by increasing and decreasing the cost of each individual component reported in Table I by 50%. We observed that, in all cases, the ratios between cost and throughput of the SDM architectures present the same trends shown in Fig. 5, and the conclusions drawn above remain valid. The reason is that there is a relevant difference in the amount of equipment required by each SDM architecture, therefore changing the cost of a single component does not affect the conclusions. We then changed by 50% the cost of



(a) Load = 30%.

(b) Load = 60%.

Fig. 4. Blocking probability of the proposed SDM architectures as a function of the size of the modular DC ($N_p$) and for different traffic load values.



(a) Load = 30%.

(b) Load = 60%.

Fig. 5. Ratio between cost (CU) and throughput (Tbps) of the proposed SDM architectures as a function of the size of the modular DC ($N_p$) and for different traffic load values. Boxes indicate the SDM architecture that returns the best trade-off between cost and throughput for a given range of DC sizes.

two components at the same time. Again in almost all cases the same conclusions discussed above remain valid. However, when reducing the cost of both MIMO transceivers ($C_{tr}^{f,t,m}$) and SSSS ports ($C_{ssss}$) by 50%, A5 becomes slightly less expensive than A2. Consequently, there is a range of DC sizes in which A5 returns the best trade-off between cost and throughput.

We also performed a sensitivity analysis with respect to the number of spatial elements per fiber ($N$). Specifically, we varied $N$ between 2 and 20, and we evaluated the ratio between cost and throughput for the SDM architectures. The results show that $N$ affects the exact range of DC sizes in which each SDM architecture returns the best trade-off. Increasing $N$ leads to increasing almost linearly the range of DC sizes for which A1 represents the best option. On the other hand, increasing $N$ leads only to marginal increase in the performance of A3; thus the range of DC sizes for which A3 is the best option is not significantly affected by the value of $N$. Finally, increasing $N$ increases almost at the same pace the performance of A2, A4 and A5. In our sensitivity study, A2 usually returns the best trade-off for large DCs, A4 is qualified as the best architecture only in a few cases of extremely large DCs, and A5 can never be considered as the best option.

Another factor that might have some impact on the results is the resource allocation strategy utilized in each of the SDM architectures. We plan to investigate this aspect in our future work.

## VII. Conclusions

In this paper, we analyzed the applicability of five different SDM schemes for the interconnection of modular DCs. For each SDM scheme, we proposed a possible network architecture and resource allocation strategy. We devised cost and simulation models to evaluate the cost and the performance (blocking probability, throughput) of the proposed solutions. Our results show that the SDM architecture returning the best cost-performance tradeoff mainly depends on (*i*) network load and (*ii*) DC size ($N_p$).

A1 is the best option for small DCs and relatively low load values. Examples are modular DCs with up to 20 PODs with working load of 30% and up to 8 PODs with working load of 60%. A3 is the best solution for medium DCs and medium load values, such as modular DCs with 20 to 120 PODs with working load of 30% and DCs with 8 to 50 PODs with working load of 60%. A2 is the best architecture for large DCs and large load values. Examples are modular DCs with more than 120 PODs with working load of 30% and DCs with 50 to 320 PODs with working load of 60%. For some very large modular DCs and high load values, A4 is the only architecture that can provide acceptable performance. Finally, A5 never represents the best solution in our considered scenarios. On the other hand, A5 could become interesting if and when it will be possible to realize low-cost SDM devices (i.e., MIMO transceivers and SSSS switches).

## Acknowledgment

This work was partly supported by the Swedish Research Council (VR), the Swedish Foundation for Strategic Research (SSF), and the Göran Gustafsson Foundation.

## References

[1] Jeff Clark, "The rise of mega data centers," *The Data Center Journal*, May 2012. Available at http://www.datacenterjournal.com/the-rise-of-mega-data-centers/.

[2] https://www.facebook.com/LuleaDataCenter/

[3] http://www.datacenterknowledge.com/

[4] https://www.hpe.com/us/en/integrated-systems/pods.html

[5] http://c3378910.r10.cf0.rackcdn.com/aag-c45-657373.pdf

[6] Cisco, "Global cloud index: forecast and methodology, 20142019," *White Paper*, August 2016.

[7] J. Chen, Y. Gong, M. Fiorani, and S. Aleksic, "Optical interconnects at the top of the rack for energy-efficient data centers," *IEEE Communications Magazine*, vol. 53, no. 8, pp. 140-148, August 2015.

[8] M. Fiorani, S. Aleksic, P. Monti, J. Chen, M. Casoni, and L. Wosinska, "Energy efficiency of an integrated intra-data-center and core network with edge caching," *IEEE/OSA Journal of Optical Communications and Networking*, vol. 6, no. 4, pp. 421-432, April 2014.

[9] D. J. Richardson, J. M. Fini, and L. E. Nelson, "Space-division multiplexing in optical fibres," *Nature Photonics*, vol. 7, no. 2, pp. 354-362, May 2013.

[10] S. Yan et al., "Archon: A function programmable optical interconnect architecture for transparent intra and inter data center SDM/TDM/WDM networking," *IEEE/OSA Journal of Lightwave Technology*, vol. 33, no. 8, pp. 1586-1595, April 2015.

[11] D. Klonidis et al., "Spectrally and spatially flexible optical network planning and operations," *IEEE Communications Magazine*, vol. 53, no. 2, pp. 69-78, Feb. 2015.

[12] M. Fiorani, M. Tornatore, J. Chen, L. Wosinska, and B. Mukherjee, "Optical spatial division multiplexing for ultra-high-capacity modular data centers," Proc. of *IEEE/OSA OFC*, March 2016.

[13] R.J. Essiambre, and R. W. Tkach, "Capacity trends and limits of optical communication networks," in Proc. of *IEEE*, vol. 100, no. 5, pp. 1035-1055, May 2012.

[14] J. Sakaguchi, et al., "19-core fiber transmission of 19x100x172-Gb/s SDM-WDM-PDM-QPSK signals at 305 Tb/s," in Proc. of *IEEE/OSA OFC*, March 2012.

[15] R. Ryf, S. Chandrasekhar, S. Randel, D.T. Neilson, N.K. Fontaine, and M. Feuer, "Physical layer transmission and switching solutions in support of spectrally and spatially flexible optical networks," *IEEE Communications Magazine*, vol. 53, no. 2, pp. 52-59, Feb. 2015.

[16] D. Marom, and M. Blau, "Switching solutions for WDM-SDM optical networks," *IEEE Communications Magazine*, vol. 53, no. 2, pp. 60-68, Feb. 2015.

[17] T. J. Xia, H. Fevrier, T. Wang, and T. Morioka, "Introduction of spectrally and spatially flexible optical networks," *IEEE Communications Magazine*, vol. 53, no. 2, pp. 24-33, Feb. 2015.

[18] A. Muhammad, G. Zervas, D. Simeonidou, and R. Forchheimer, "Routing, spectrum and core allocation in flexgrid SDM networks with multi-core fibers," Proc. of *IEEE ONDM*, May 2014.

[19] D. Siracusa, et al., "Spectral vs. spatial superchannel allocation in SDM networks under independent and joint switching paradigms," Proc. of *IEEE ECOC*, Sept. 2016.

[20] P. S. Khodashenas et al., "Comparison of spectral and spatial super-channel allocation schemes for SDM networks," *IEEE/OSA Journal of Lightwave Technology*, vol. 34, no. 11, pp. 2710-2716, June 2016.

[21] P. Mitchell, G. Brown, R. Thomson, N. Psaila, and A. Kar, "57 channel (193) spatial multiplexer fabricated using direct laser inscription," Proc. of *IEEE/OSA OFC*, March 2014.

[22] O. Gerstel, M. Jinno, A. Lord, and S. J. B. Yoo, "Elastic optical networking: a new dawn for the optical layer?," *IEEE Communications Magazine*, vol. 50, no. 2, pp. s12-s20, February 2012.

[23] J.L. Vizcano, Y. Ye, V. Lopez, F. Jimenez, R. Duque, and P. M. Krummrich, "Cost evaluation for flexible-grid optical networks," Proc. of *IEEE Globecom*, December 2012.

[24] M. Fiorani, S. Aleksic, M. Casoni, L. Wosinska, and J. Chen, "Energy-Efficient elastic optical interconnect architecture for data centers," *IEEE Communications Letters*, vol. 18, no. 9, pp. 1531-1534, September 2014.

[25] T. Benson, A. Akella, and M. Zhang, "Understanding data center traffic characteristics," Proc. of *IEEE/ACM SIGCOMM*, August 2009.

[26] T. Benson, A. Akella, A. Akella, and D. A. Maltz, "Network traffic characteristics of data centers in the wild," Proc. of *IEEE IMC*, November 2010.

[27] K. Chen et al., "OSA: An Optical Switching Architecture for Data Center Networks With Unprecedented Flexibility," *IEEE/ACM Transactions on Networking*, vol. 22, no. 2, pp. 498-511, April 2014.

[28] C. Kachris, K. Bergman, and I. Tomkos, "Optical Interconnects for Future Data Center Networks," *Springer Science & Business Media*, 2012.