Resource Partitioning in the NEPHELE Datacentre Interconnect

K. Yiannopoulos[1], K. Kontodimas[2], K. Christodoulopoulos[2,3], E. Varvarigos[2]
1: Department of Informatics and Telecommunications, University of Peloponnese, Greece
2: School of Electrical Engineering and Computer Science, National Technical University of Athens, Greece
3: Department of Computer Engineering and Informatics, University of Patras, Greece
vmanos@central.ntua.gr

ABSTRACT
We present heuristic algorithms for the efficient resource partitioning in the NEPHELE datacentre optical interconnect. The algorithms aim to segment the network into smaller and isolated virtual datacentres (VDCs), where all racks are able to communicate at full capacity irrespective of their placement. Since the NEPHELE architecture relies on shared optical rings, the isolation of VDC traffic is challenging. Observing its close resemblance to finding a bi-clique on a bipartite graph, which is NP-hard, we propose heuristic algorithms which find a solution by limiting either the spatial spread of racks that construct each VDC or their wavelength allocation. If a solution cannot be found, then the algorithms invoke a second de-fragmentation phase, where they re-allocate the racks of existing VDCs to concentrate them spatially and reduce traffic on the shared optical rings. It is demonstrated via simulation that the proposed heuristics can achieve very high utilization and also exhibit low VDC request blocking probability for typically expected VDC sizes.
Keywords: virtual datacentres, resource partitioning, optical rings, bi-cliques.

1. INTRODUCTION

Resource partitioning is a key functionality that enables offering the DC as a service to customers that operate independently and require a diverse spectrum of resource sizes [1]. This calls for the deployment of virtual datacentres (VDCs) that share a common physical infrastructure (servers and switches), but are also operating in full isolation from one another [2], [3]. In the NEPHELE datacentre interconnect an almost hierarchical structure is deployed, where servers are grouped in racks and multiple racks are grouped to form PODs [4]. The racks of each POD are identified by their receiving wavelengths, while racks that are located in different PODs interconnect over multi-wavelength optical rings. It is then straightforward to create VDCs in NEPHELE by assigning the available racks to them, along with the required rings that enable rack communications whenever racks are located in more than one POD.

The valid rack and ring combinations that will form a new VDC are, however, limited by the shared nature of the optical rings. If an existing VDC utilizes a specific wavelength over a ring, then this wavelength/ring assignment should not be used for any other VDC. Moreover, a similar problem arises within the VDC itself: if the same wavelengths are assigned by multiple racks in the VDC then collisions will occur when rack communications take place over a common ring segment (see Fig. 1). Therefore, VDC allocation requires a careful examination of the rack/ring combinations that are available in NEPHELE. Finding an appropriate rack/ring allocation for a VDC request resembles the "balanced complete bipartite subgraph (bi-clique)" problem, which is NP-complete [5]. To make things harder, the bipartite graph that summarizes the available racks (vertices) and rings (edges) constantly changes during the formation of the VDC (bi-clique), since the selection of a rack rules out the inclusion of all colliding racks in the VDC; thus edges are constantly removed from the bipartite graph, and existing bi-clique finding algorithms that rely on fixed graphs are not applicable.

Within this context, we present and evaluate two heuristics that are compatible with rack and ring collision-free allocation in NEPHELE: the first heuristic only allows two PODs per VDC, since all available wavelengths can be used without collisions between any two PODs (refer to Fig. 1(c)). The second heuristic allows multiple PODs per VDC, but each wavelength is only allowed once in the VDC; intra-VDC collisions similar to those of Fig. 1(a), (b) are not possible in this approach. Both heuristics are simulated in Matlab and simulation results show that the heuristics achieve reasonable utilization (~90%) of the NEPHELE rack capacity, while simultaneously VDC requests are accepted and facilitated with high probability (95%) whenever the VDC size amounts to approximately 10 racks. Smaller utilization and higher blocking probabilities are observed for increasing VDC sizes.

2. FORMULATION OF THE RING AND RACK ASSIGNMENT PROBLEM

2.1 Ring Constraints

The racks (or equivalently top-of-rack switches – TORs) inside the NEPHELE PODs are identified by their unique receiving wavelength and W=80 wavelengths are present in each of the P=20 PODs; for wavelength re-use purposes, the same set of W wavelengths is used in all PODs. Intra-POD communications are collision free and a WxW arrayed waveguide grating (AWG) routes wavelengths inside the POD. Inter-POD communications take place over R=20 multi-wavelength optical rings that are accessed by all PODs and are capable of

transporting all W wavelengths simultaneously. The physical connections between the TORs and the rings are implemented via cyclic AWGs; a WxR AWG multiplexer routes traffic from TORs on the optical rings, while a second RxW AWG demultiplexer directs traffic from the rings to the TORs with the appropriate wavelength.

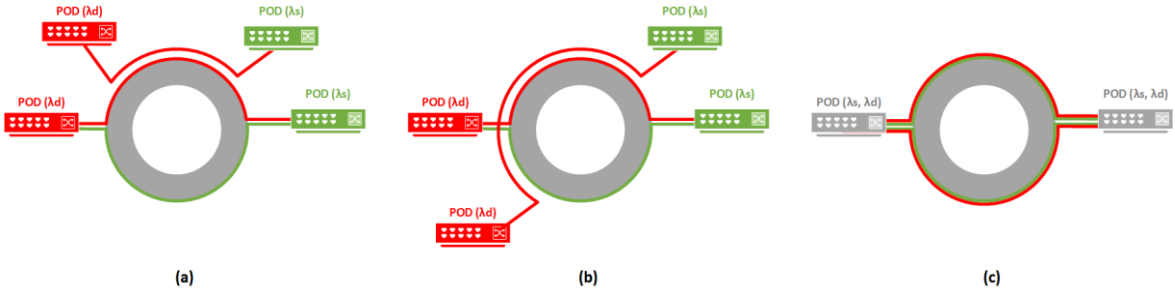

Figure 1. Rack and ring assignments: assignments (a) and (b) create collisions on the shared optical rings, while (c) is collision free.

Let us now assume that a ring is utilized for the communication of two racks $TOR_{s,k}$ and $TOR_{d,l}$ inside the VDC; for ease of notation, we will refer to TORs as $TOR_{w,p}$ meaning that the corresponding TOR is tuned at wavelength λ_w and is located at POD_p. The cyclic AWG will route this communication to the ring with index r that is calculated from

$$r = (\lambda_s + \lambda_d) \bmod R. \tag{1}$$

However, communications from any other source $TOR_{s+n\cdot R,k}$, (with $1 \le s + n \cdot R \le W$) in NEPHELE towards a destination with the same wavelength $TOR_{d,l}$, will also utilize the same ring and a collision will take place, unless one of the following constraints are met when forming the VDC:

- VDC constraint 1: If more than one TORs in the wavelength set $\lambda_{s+n\cdot R}$ are allocated in the VDC, then only one of the TORs with wavelength λ_d may also be allocated.
- VDC constraint 2: If wavelength λ_d is assigned to more than one TORs in the VDC, then only one of the TORs in the wavelength set $\lambda_{s+n\cdot R}$ may also be assigned.

A similar constraint is also true when considering communications that belong to different VDCs and an existing communication between $TOR_{s,k}$ and $TOR_{d,l}$ "blocks" all communications between $TOR_{s+n\cdot R,k}$, and $TOR_{d,l}$, in subsequent VDC allocations.

2.2 Allocation Equivalence with a Bi-clique

The TOR connections that are available given the existing allocations in VDCs can be summarized by a Boolean $W{\cdot}P$x$W{\cdot}P$ connection availability matrix. The matrix entries are either '1' if the TOR connection is available, or '0' otherwise. Establishing a VDC of a specific rack size m then corresponds to finding an equally-sized set of matrix columns and rows with all intersections being equal to '1', since all TORs must be able to communicate with each other in the VDC. Using a bi-partite graph representation of the availability matrix as in Fig. 2, the VDC allocation is equivalent to finding a $K_{m,m}$ bi-clique (i.e. a complete bipartite subgraph with $2m$ vertices) on the connection availability bipartite graph. Additionally, the ring constraints must be also observed in the bi-clique, i.e. two vertices with wavelengths that introduce collisions must not be allowed to co-exist inside the same or different bi-cliques, as dictated by the aforementioned constraints.

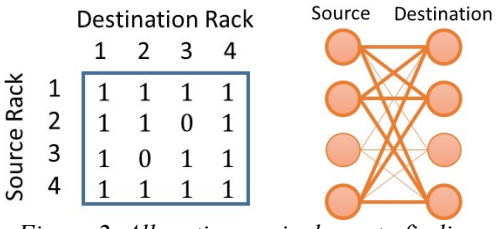

Figure 2. Allocation equivalence to finding a bi-clique on a bipartite graph.

3. HEURISTICS AND EVALUATION

The problem of finding $K_{m,m}$ bi-cliques on the bipartite graph has been shown to be NP-complete [5]. This stems from the fact that all possible combinations must be exhaustively checked before a solution can be found; the exhaustive search of allocation is practically impossible given the targeted size of NEPHELE. Moreover, the constant update of the bipartite graph to avoid collisions further aggravates the execution time of exhaustive searches. As a result, we resort to heuristics that limit the combinations that are evaluated in order to calculate VDC allocations within a limited time.

3.1 Two-POD Heuristic

The heuristic constraints the VDC allocation to one or two PODs, thus reducing the possible TOR combinations that need to be examined. Moreover, intra-VDC collisions are not possible when only two PODs communicate

and there is no need for a constant update of the bi-partite graph (i.e. any bi-clique between two PODs is a valid VDC allocation), leading to execution times that equal a couple of secs per allocation. Still, collisions between VDCs are still possible and the availability matrix is updated after the two-POD bi-clique has been found.

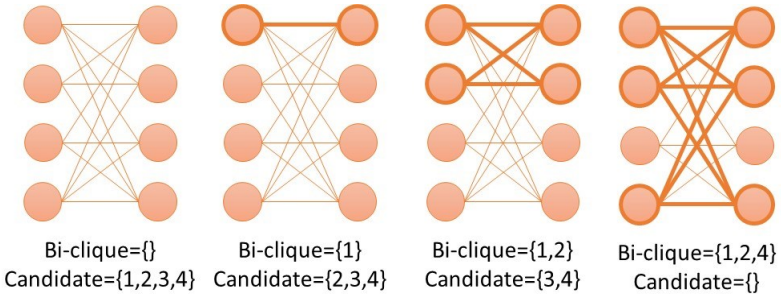

Figure 3. Exhaustive search of bi-cliques.

The heuristic initially seeks a single-POD solution; the corresponding bi-clique is trivial to find since intra-POD communications are collision free and if m or more TORs are available in a POD then a solution has been found. If a single-POD solution is not feasible, then reduced size allocation matrices are formed for all possible two-POD combinations. These new matrices are exhaustively searched for bi-cliques of the required size m by gradually increasing the size of smaller bi-cliques, in a process that is illustrated in Fig. 3. Given the reduced size of the problem, it is well possible to find a solution within a few seconds at the price of rejecting VDC requests when the available resources are located at more than two PODs.

The heuristic was simulated in Matlab for VDCs that requested a uniformly distributed number of TORs ranging from 1 to M=10, 40 and 80. To better evaluate the dynamic response of the heuristic, we considered VDC requests that arrived according to a Poisson process with arrival rate λ, and had an exponentially distributed lifetime $1/\mu$. The arrival and departure rates were set following the desired average occupancy ρ ("load") of PODs in NEPHELE following the relation

$$\rho = \frac{\lambda}{\mu}\frac{M+1}{2WP}. \tag{2}$$

Fig. 4 summarizes the results that were obtain in terms of the rack utilization and the blocking probability. The results shown in the figure show that the heuristic is performing almost without utilization loss for a VDC size of 10 TORs, which is typically anticipated in contemporary DCs, and loads of up to 0.9. A similar behaviour is also observed for bigger VDC sizes, but the heuristic becomes less capable of utilizing the available TORs at lower loads. This behaviour can be explained by the fact that a new VDC request finds TORs and rings partially filled, thus a suitable TOR and ring combination must be found. However, it is not possible to a-priori know the VDC requests and plan ahead; there are possible scenarios where the new VDC is "blocked" by existing VDC placements in NEPHLE because a suitable TOR/ring combination cannot be found, either because such a combination does not exist or because the heuristic limits itself to examining two PODs. In agreement with the results for utilization, a VDCs are blocked with increasing probability as their average size increases, and the blocking probability can get as high as 18% , 12% and 4% for maximum VDC sizes of 80, 40 and 10, respectively.

The performance of the heuristic can be improved if some of the VDC allocations that have been decided in earlier time are changed so as to release some of the utilized rings and/or increase the contiguous occupation of a POD. This *de-fragmentation* process requires the migration of the VDC virtual machines from the TORs the VDC currently occupies to another set of TORs located at a different POD (or PODs). Since VM migration is a costly process that disrupts service, re-allocation should be avoided. Within this rationale, we also

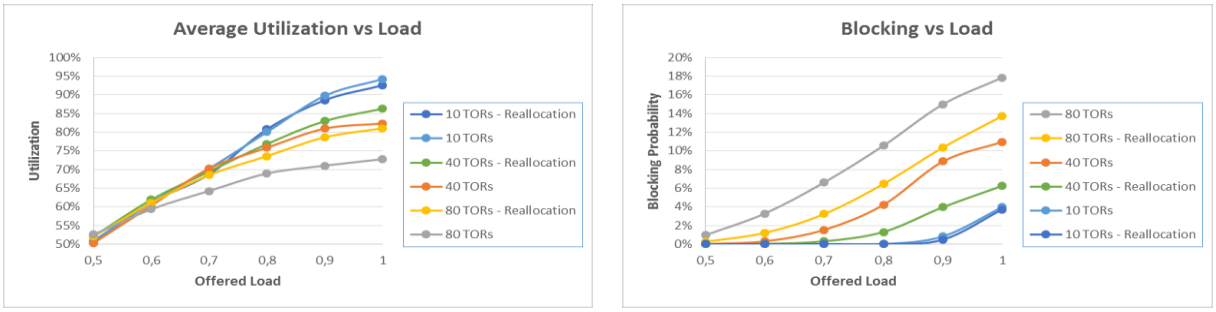

Figure 4. Simulation results for the two-POD heuristic.

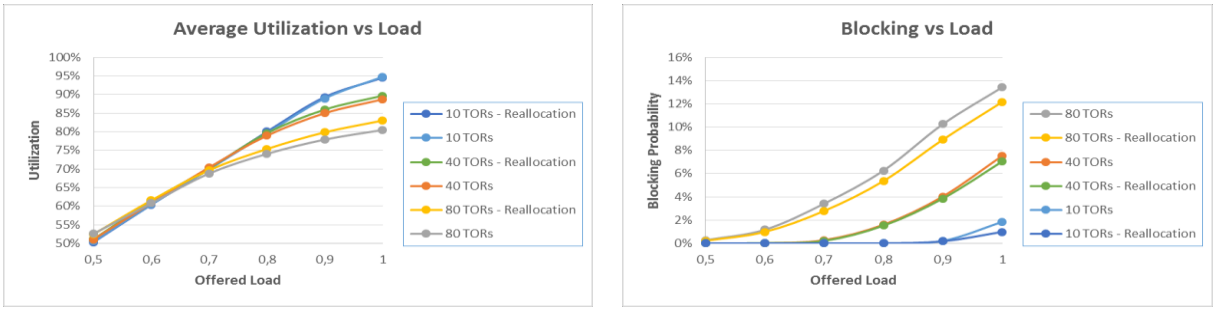

Figure 5. Simulation results for the unique-wavelength heuristic.

examined an extension of the heuristic to allow limited re-allocation of VDCs. The utilization of TORs and blocking that is achieved by the re-allocation algorithm is also presented in Fig. 4. It becomes clear that re-allocation has no meaningful impact for small size VDCs, where blocking is limited. On the other hand, a utilization improvement of almost 5% and 9% is evident as the VDC size increases to 40 and 80 TORs, respectively, at high loads. Blocking is also improved by 4% in both cases. Re-allocation also becomes less efficient at lower loads, where the original heuristic version is capable of finding a solution with high probability.

3.2 Unique-Wavelength Heuristic

The second heuristic does not limit the number of PODs that participate in a VDC, and a VDC can potentially reside in all PODs. They key difference is that a wavelength is never used twice in the same VDC, and this ensures that intra-VDC collisions can be avoided. From an implementation perspective, the heuristic needs to keep track which wavelengths are available on each ring segment for each new VDC request. This introduces a significant execution time overhead and to mitigate this we assume that wavelength pairs occupy the whole circumference of the optical ring, rather than ring segments (see Fig. 1). In this approach, the heuristic is now only required to remember the wavelength pairs (λ_s, λ_d) that have been used by previous VDCs (or blocked in the rings), since they are sufficient to determine the ring that is occupied according to Eq. (1). A reduced WxW wavelength pair availability matrix is now formed to keep track of available wavelength pairs and the bi-clique is found on the this matrix, which further speeds-up the execution time of the heuristic to <1 sec per allocation.

The simulation results for the unique-wavelength heuristic are summarized in Fig. 5. It becomes evident that the heuristic performs similar to the two-POD heuristic (without re-allocation) for small VDC sizes (M=10) but provides better utilization as the VDC size increases. An improvement of 6% and 8% is observed for M=40 and 80, respectively, at ρ=1.0. The blocking probability is also improved by approximately 4% for these parameter values. The improvement is expected by the fact that the second heuristic does not limit itself to one or two PODs; in some cases it is possible to have all twenty PODs participating in the VDC formation. Still, the utilization is not 100% because the heuristic only considers TORs that fully inter-connect on the optical rings and misses out on allocations with a strong intra-POD component that is not also available on rings. As it can be verified from Fig. 5, the re-allocation process does not aid the heuristic in a significant manner towards this direction, and no measurable improvement is observed. Unfortunately, these allocations can be only found by examining bi-cliques on the original WPxWP allocation matrix.

4. CONCLUSIONS

We presented two heuristics for the allocation of racks and optical rings in the NEPHELE datacentre interconnect. The heuristics complete adequately fast and their simulation revealed that high utilization and low request rejection (blocking) can be expected, but their performance becomes worse with increasing VDC sizes. A utilization of ~90% of the available racks and an allocation success probability of ~95% is predicted for contemporary commercial VDC sizes that require ~10 racks per VDC.

REFERENCES

[1] Amazon Elastic Compute Cloud (Amazon EC2). https://aws.amazon.com/ec2/

[2] M. F. Bari *et al.*: Data Center Network Virtualization: A Survey, IEEE Communications Surveys & Tutorials, vol. 15, no. 2, pp. 909-928, Second Quarter 2013.

[3] S. Peng *et al.*: Multi-Tenant Software-Defined Hybrid Optical Switched Data Centre, *J. Lightwave Technol.*, vol. 33, no 15, pp. 3224-3233, 2015.

[4] K. Christodoulopoulos *et al.*: Bandwidth allocation in the NEPHELE hybrid optical interconnect, in *Proc. ICTON 2016*.

[5] M. S. Garey and D. S. Johnson: Computers and Intractibility: A Guide to NP-Completeness, Freeman, New York, 1979.