Available online at www.sciencedirect.com

ScienceDirect

IERI Procedia 7 (2014) 8 – 14

Procedia IERI

www.elsevier.com/locate/procedia

2013 International Conference on Applied Computing, Computer Science, and Computer Engineering

Complex Cloud Datacenters

Sonja Filiposka[a,b]*, Carlos Juiz[b]

[a]*Faculty of Computer Science and Engineering, Ss. Cyril and Methodius University – Skopje, 1000 Skopje, R. Macedonia*
[b]*Computer Science Department, University of Balearic Islands, 07122 Palma de Mallorca, Spain*

Abstract

The network architecture deployed in a data center is of extreme importance and should provide the data center with the high throughput while keeping overprovisioning at the minimum costs. In order to provide these features new complex network based models have been proposed recently. In this paper we give an overview of the proposed models and their characteristics, advantages and problems. We then propose a new datacenter network model that provides easy subgrouping while preserving high bandwidth and low average path length and network diameter.

Selection and peer review under responsibility of Information Engineering Research Institute

Keywords: data center; complex networks; performances.

1. Introduction

The rise of cloud computing and the anything-as-a-service models offered by large-scale, distributed datacenters have become the driving force behind the continuous growth and enormous overprovisioning of datacenters. The topology of the datacenter interconnection network is the cornerstone for ensuring scalability and implementation of optimal resource management policies. Thus, the network architecture deployed in a

* Corresponding author. Tel.: +389-2-3099-153; fax: +389-2-3088-222.
E-mail address: sonja.filiposka@finki.ukim.mk.

Selection and peer review under responsibility of Information Engineering Research Institute
doi:10.1016/j.ieri.2014.08.003

data center is of extreme importance and should provide the data center with the necessary performances especially in terms of high throughput while keeping overprovisioning at the minimum in order to cut costs.

State-of-the-art data center architectures today have precisely designed and rigid symmetric structures made out of homogenous equipment. Typical network architectures examples include fat-tree [1], BCube [2], DCell [3], and other tree like and symmetrical regular architectures. However, the extremely symmetric design and the use of homogenous equipment are clashing with today's dynamic and constant need for expansion in the data centers. Growth with these designs can be done only on a large scale involving a great number of changes, while the datacenter needs (put into economic terms) demand the ability to extend the number of servers or network equipment proportionally to the available resources.

In order to introduce increased flexibility inside the data center topology, there have been some approaches that remodel the datacenter's network topology following a complex network pattern that has been discovered in many large-scale natural networks. The common characteristics shared by all complex networks are the ability for evolutive growth and highly efficient communication due to the small diameter despite the huge number of nodes. Previous examples following this line of research include Jellyfish [4], small-world datacenters (SWDC) [5] and Scafida [6]. All of these complex network topologies offer solution to the problem of small average path length between the network nodes, but are different in the ways they treat the homogenous aspect of the network equipment and the possibility for group formations. In order to work towards a unified solution that will result with a network topology with all of the desired characteristics, we intend to further develop the proposed Scafida data center network topology model proposed in [6], by extending it with another very important feature: clustering. In this paper, we provide a network topology that besides having small diameter and high bandwidth, will also be clustered, i.e. divisible into groups, which can help increase the performances of the datacenter by localizing traffic.

2. Cloud datacenter topologies based on complex networks

Cloud data center network architectures consist of thousands and more servers that provide services to the customers. Most popular data centers networks are based on regular topologies [1] and are arranged in three layers: core, medium, and server layers and can be built using commodity Ethernet switches, where the flows are leveraged using multiple paths. Many of the proposed symmetric architectures (especially fat-tree based) allow expansion at a very coarse rate (in terms of several thousand servers) and require servers with free ports reserved for future expansion. The main goal of all regular topology solutions is to offer high bandwidth.

Because of the high limitations on growth imposed by their regular structure, recently a number of datacenter network topologies based on the fundamental properties of complex networks have been proposed.

2.1. Jellyfish

Jellyfish [4] is a degree-bounded random graph topology that can be built using top-of-rack switches. The design theoretically supports heterogeneity and also allows construction of arbitrary-size networks. It was shown that Jellyfish supports more servers than a fat-tree built using the same network equipment while providing at least as high per-server bandwidth. In addition, Jellyfish has lower average path length, and is resilient to failures and miswirings.

However, it must be noted that the bandwidth and throughput measurements where based on calculations of the theoretical bounds for bisection bandwidth for regular random graphs which stand and can be used for very large-scale networks only. Also, the construction of the network itself is very hard to deploy.

All of the benefits presented with the Jellyfish architecture are actually the properties of the underlying and very popular Erdös-Rény random graph model [7] wherein the edges connecting vertices are created

randomly. The small average path length and the remarkable resilience to failures are inherent properties of these random graphs. Although the Jellyfish model implements the random graph for degree-bounded nodes (because of the limited number of links that a switch can support), the main influence of these characteristics still stands especially for the presented large-scale cases using homogenous switches.

2.2. Small-world datacenters

Regular lattice graphs wherein a fraction of edges are rewired with a probability p have been proposed by Watts and Strogatz and are known as small-worlds networks [8]. The small-world datacenters (SWDC) model [5] proposes a topology for data centers inspired by this small-world complex network topology. They also propose a network model based on a locally-regular foundation that is easy to wire, such as a ring, grid, cube or torus, and amend this regular structure with a number of random network links placed throughout the datacenter. The result is a datacenter topology with a low average path length, uniform node degree distribution and tendency towards structural grouping (clustering). SWDC can provide higher bandwidth than conventional fat-tree based hierarchical datacenters, and are very resilient to random node failures.

However, SWDCs rely on a collection of NICs (the authors propose 6) in each server node in order to provide the rich topology, while switches are not utilized at all. Also, adding new nodes in an SWDC requires care to ensure that the resulting graph has retained its beneficial properties. Another problem are the long distance links, which are difficult to implement and can easily become bottlenecks in the network if very careful routing is not provided. This will mean that the route lengths will be far longer than the average path length of the graph.

2.3. Scafida

The authors in [6] proposed a data center topology generation algorithm called Scafida by modifying the scale-free network creation algorithm given by Barabási and Albert (BA) [9]. The network structure is generated iteratively, i.e. the nodes are added one by one; a new node is attached to an existing node using preferential attachment, i.e. with probability that is proportional to the existing node's degree. The modification of the original model is the artificial constrain put on the number of links that a node can have, so-called degree bounding of the nodes, in order to meet the port number of switches and servers which are defined at the beginning of the algorithm.

Scale-free networks have proven to be inherently highly resistant to random failures, i.e. the diameter of the network does not increase until significant number of the nodes fails. The degree bounded homogenous networks investigated in [6] still have a fairly small diameter and average path length. Also, the bandwidth offered by Scafida is close to the one obtained using the symmetric topologies. However, more investigation needs to be done in order to see how these properties change in a heterogeneous environment. Another problem with the proposed Scafida model is the complete lack of clustering and grouping in the graph.

3. Clustered scale-free data center

A cloud data center is a complex system with a very large number of shared resources subject to user requests. Thus, cloud resource management is extremely challenging. We argue that a clustered scale-free network architecture will provide a core subset of nodes and build groups of service nodes around them. In this way optimal resource management policies like capacity allocation, load balancing, energy optimization, and QoS guarantees can be efficiently implemented. Thus, we propose a data center network model that is based on both the scale-free evolution with small average path length and high bandwidth, together with

clustering. By enabling grouping in the data center architecture we can obtain lots of benefits like, for example, isolating services from each other by group placing and localization of traffic.

3.1. Clustered scale-free model

We propose a clustered scale-free algorithm that generates a data center topology with servers and a number of different types of switches in terms of ports. Starting from a core of m interconnected nodes, the network is growing iteratively, i.e. the nodes are added one by one to the network just as in the scale-free BA model. During the network growth, the degree-bound conditions are met according to the constraints given by the defined pool of switches, which is a newly added node cannot create a link towards a switch that has no more free ports.

According to the Scafida model every new node in the network attaches it self to m other existing nodes via the preferential attachment (PA) method proposed by Barabási and Albert. The preferential attachment method will however result in a network that has an extremely low clustering, thus making it extremely difficult to divide into groups. In order to increase the clustering, we propose a different type of linking, which is based on the clustered scale-free model proposed in [10]. When a node v with m edges is added to the network, we first perform one PA step, and then perform a triad formation (TF) step with the probability p_{tf} or a PA step with the probability 1-p_{tf}. The average number m_t of the TF trials per added node is then given by $m_t=(m - 1)p_t$. It should be noted that our model reduces to the original Scafida model when $m_t = p_{tf}= 0$.

The rule of triad formation is as follows: If an edge between v and w was added in the previous step using preferential attachment, then add one more edge from v to a randomly chosen neighbor of w, and thus effectively create a triad. If there remains no pair to connect, i.e., if all neighbors of w are already connected to v, do a preferential attachment (PA) step instead. The reasoning behind the triad formation is that the actual clustering coefficient of a network is measured as the average number of triads in the network, since the triads have been show to be the foundation of all tightly connected groups. We must note that this type of algorithm results with some self-loops which are afterwards discarded. Thus the clustered network may have somewhat smaller number of links compared to the Scafida model.

3.2. Characteristics

In order to make a thorough analysis of the model properties we defined a number of different example heterogeneous networks with different critical parameters. Our aim was to investigate how the model properties change when changing the number of nodes in the network, the switch types involved (for the presented results we used three types of switches with 8, 16 and 32 ports), and the probability for triad formation. All of the presented values are averaged over a set of 50 generated topologies for each case, where the value of m was set to 2. For all of the sets, the clustering control parameter p_{tf} was changed from 0 to 0.75. While reviewing the properties of the new model, we will focus our performances comparison of the model with the Scafida model. As the servers of a data center communicate with each other, the average length of the paths between the nodes fundamentally impacts the performance of the data center network. While the *average path length* (APL) represents the average of the length of all possible end-to-end pairs of nodes, the *network diameter* is actually represented by the maximum of all possible paths.

Our clustered model shows only slight increasing of the APL and diameter compared to both the degree bounded Scafida (p_{tf} = 0.0) and the original scale-free network. In Fig. 1a we present the impact of limiting node degrees on the average shortest path length and network diameter for different network sizes. The increasing network size is gradually increasing the APL and diameter, which is expected. However, irrespective of the size of the networks, the average lengths of the paths are negligently increased compared to

the not-clustered network, while the network diameter increases somewhat more prominently, but remains small compared to the network overall size (e.g. diameter of 12 hops for a network with over 20000 nodes).

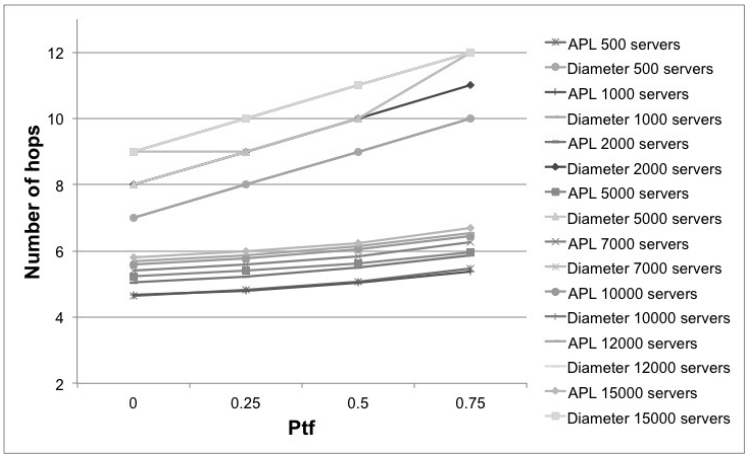

Fig. 1. (a) Average path length (APL) and network diameter;

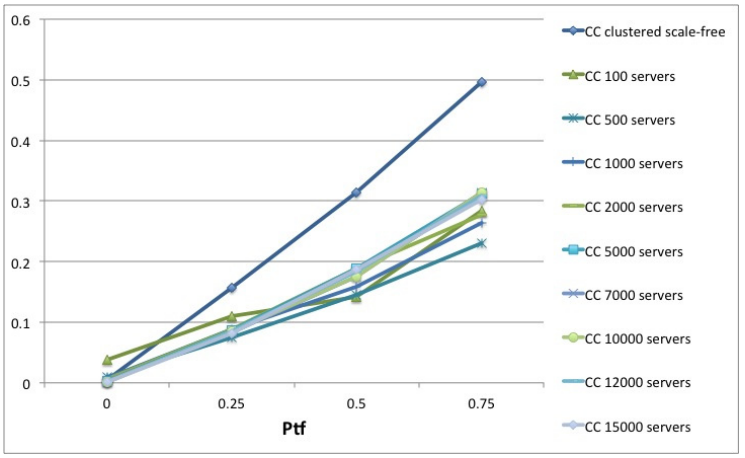

Fig. 1. (b) clustering coefficient (CC) for different datacenter sizes

The main motivation in the presented work was to introduce structured clustering into the network topology and create a network that will be easily partitioned into groups that will be tightly interconnected inside, with smaller number of interconnections. Thus, we utilized the triad formation in order to boost the clustering coefficient of the network and make its structure more group oriented. In Fig. 1b, we present the changes observed in the value for the clustering coefficient (CC) when using our model. The figure also presents the change in the clustering coefficient in case no degree bounding is present (pure clustered scale-free network). The comparison shows that for our degree bounded case the change in the clustering coefficient is less prominent, but however still quite substantial compared to the scale-free network model without clustering ($p_{tf} = 0.0$) for which case (with and without degree bounding) the clustering coefficient is always ~ 0.0 representing no structural tendency for group formations. The figure also presents that the

clustering coefficient increases with p_{tf} being under almost no influence of the network size. Also after introducing clustering in the network, the routes between the nodes are significantly decreased since the intra-group communication takes place on a very smaller number of hops compared to the average path length, while the inter-group communication is more rarely needed.

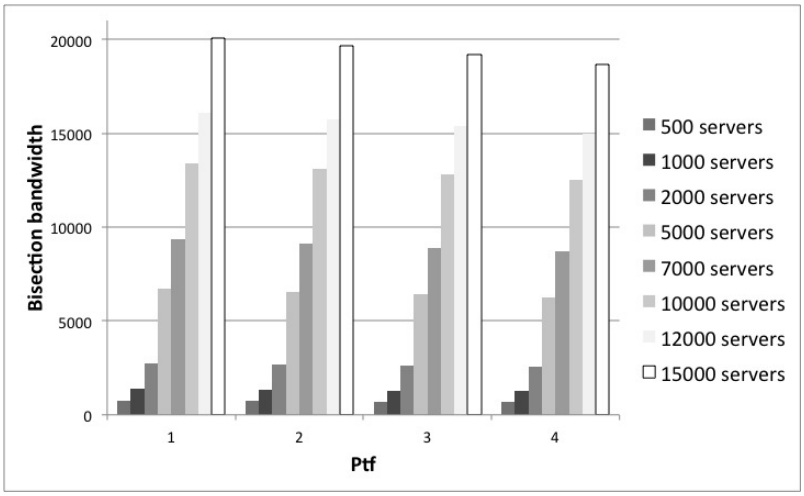

Fig. 2. Bisection bandwidth for different example clustered scale-free datacenters

Another crucial aspect of data center's architecture is its throughput capability. One of the ways to measure the throughput capability of a data center is via the bisection bandwidth. For the purposes of measuring the bisection bandwidth, the network nodes are divided into two random groups (averaged in total of 200 different random divisions) and afterwards the maximal flow between the two parts is computed as the number of links between the two groups while the link capacity is considered to be even. Fig. 2 shows the impact of degree limitation and triad formation on the distribution of bisection bandwidth in case of different example data centers. With the increasing clustering in the network, the bisection bandwidth somewhat decreases linearly, but only for 2%, 4% and 7% when p_{tf}= (0.25; 0.5; 0.75) respectively. The reason for this small decrease in bisection bandwidth is due to the smaller number of links in the network. However, since the clustering allows for traffic localization, the overall throughput of the datacenter should improve.

The proposed topology generation method is extremely scalable and flexible. On one hand, the number of servers (i.e. the size of the data center) can be set on a very fine-grained scale unlike the typically offered solutions with symmetric regular structure. Thus, the algorithm is also extremely flexible in terms of the type of network equipment that is going to be used (e.g. different switches with port numbers). On the other hand, as presented here, the performance is very high compared to other architectures.

4. Conclusion

The significant advantages of grouping, in particular the ability to implement effective admission control and QoS policies, seem important enough to justify the need for a change in cloud datacenters. Indeed, over-provisioning used by exiting clouds is not a sustainable strategy to guarantee QoS. We believe that by designing a datacenter network based on a clustered scale-free model we offer solutions and possibilities for further improvement of the datacenter performances by taking advantage of the underlying network properties. The presented results show promising characteristics that we intend to improve in our future work.

Acknowledgements

This work has been partially supported by the EUROWEB project (http://www.mrtc.mdh.se/euroweb) funded by the Erasmus Mundus Action II programme of the European Commission. This work is also partially supported by the Spanish Ministry of Economy and Competitiveness under grant TIN2011-23889.

References

[1] Al-Fares M, Loukissas A, Vahdat A. A scalable, commodity data center network architecture. SIGCOMM 2008; 63:74.
[2] Guo C, Lu G, Li D, Wu H, Zhang X, Shi Y, Tian C, Zhang Y, Lu. S. Bcube: a high performance, server-centric network architecture for modular data centers. SIGCOMM 2009; 63:74.
[3] Guo C, Wu H, Tan K, Shi L, Zhang Y, Lu S. Dcell: a scalable and fault-tolerant network structure for data centers. SIGCOMM 2008; 75:86.
[4] Singla A, Hong C-Y, Popa L, Godfrey PB. Jellyfish: networking data centers randomly. 9th USENIX conference on Networked Systems Design and Implementation 2012; 17:27
[5] Shin J-Y, Wong B, Sirer EG. Small-world datacenters. 2nd ACM Symposium on Cloud Computing 2011.
[6] Gyarmati L, Trinh TA. Scafida: a scale-free network inspired data center architecture. ACM SIGCOMM Computer Communication Review 2010; 40-5-5:12.
[7] Erdös P, Rényi A. On random graphs. Publicationes Mathematicae 1959; 6-290:297.
[8] Watts DJ, Strogatz SH. Collective-dynamics of small-world networks. Nature 1998; 393-440:442.
[9] Barabási A-L, Albert R, Jeong H. Scale-free theory of random networks; the topology of World Wide Web. Physica A, 2000; 281-69:77.
[10] Holme P, Kim BJ. Growing scale-free networks with tunable clustering. Physical Review E 65 2002.