1-1-2012

HyScaleII: A High Performance Hybrid Optical Network Architecture for Data Centers

Shivashis Saha
University of Nebraska - Lincoln, ssaha@cse.unl.edu

Jitender S. Deogun
University of Nebraska - Lincoln, deogun@cse.unl.edu

Lisong Xu
University of Nebraska - Lincoln, xu@cse.unl.edu

2012 35th IEEE Sarnoff Symposium (SARNOFF),
Digital Object Identifier: 10.1109/SARNOF.2012.6222725

HyScaleII: A High Performance Hybrid Optical Network Architecture for Data Centers

Shivashis Saha, Jitender S. Deogun, Lisong Xu
Department of Computer Science and Engineering,
University of Nebraska-Lincoln, Lincoln, NE 68588-0115, U.S.A.
Email: {ssaha, deogun, xu}@cse.unl.edu

***Abstract*—Tremendous growth in data-intensive cloud applications have resulted in an increased demand for highly scalable data center network (DCN) architectures with high throughput and low network complexity. In this paper, we propose *HyScaleII* to improve the performance of *HyScale* [17]. *HyScaleII* is a *switch-centric* high performance hybrid optical network based DCN architecture that has most of the desirable properties of a data center, e.g. high scalability, low diameter, high bisection width, fault-tolerance, and low network complexity. We also present an efficient and simple routing scheme called *HySII* routing, which exploits the structural properties of *HyScaleII*. In our experiments, *HyScaleII* has lower packet loss ratio and higher average aggregate throughput by an average of 50% and 13.8% respectively as compared to *HyScale* [17].**

I. Introduction

Mega data centers supporting 100,000 or more servers have received significant interest in recent years due to the tremendous growth and popularity of data-intensive cloud applications [1]-[8]. This motivates the investigation of *data center network* (DCN) architectures for efficiently interconnecting large number of servers. The three important design goals of such architectures are: *scalability*, *fault-tolerance*, and *high network bandwidth* [1]-[8].

The amount of data transferred within a data center has also exponentially increased in recent years. It is reported that for about every byte of data communicated over the Internet, at least 1MB of data is communicated within a data center [9]. Thus, DCN architectures are expected to provision reliable high-bandwidth communications. Electrical networks in data centers are increasingly becoming a bottleneck for supporting high-volume high-speed data transfers [10]. Therefore, the use of Optical Circuit Switching (OCS) in DCN architectures have been recently propagated for supporting such high on-demand bandwidth communications [9]-[16]. However, optical burst switching (OBS) has not received much attention in DCN architectures [17].

DCN architectures are broadly classified into *server-centric* [4]-[8] and *switch-centric* [1]-[3] architectures. In *server-centric* architectures, servers perform both computation and routing of data. The design goals of servers are not intended to support high-speed, high-bandwidth routing. Moreover, using servers as both computing and routing nodes may prevent the servers not needed for computation from being turned off or put to a low-energy state. *Server-centric* architectures are also unable to entirely exploit the benefits of optical networks [11]-[13]. Thus, such architectures are not deemed suitable for DCN architectures using optical networks even though they are reported to have better scalability and performance than most *switch-centric* architectures [4]-[5]. *Switch-centric* networks are therefore the preferred choice for DCN architectures using optical networks [9]-[15].

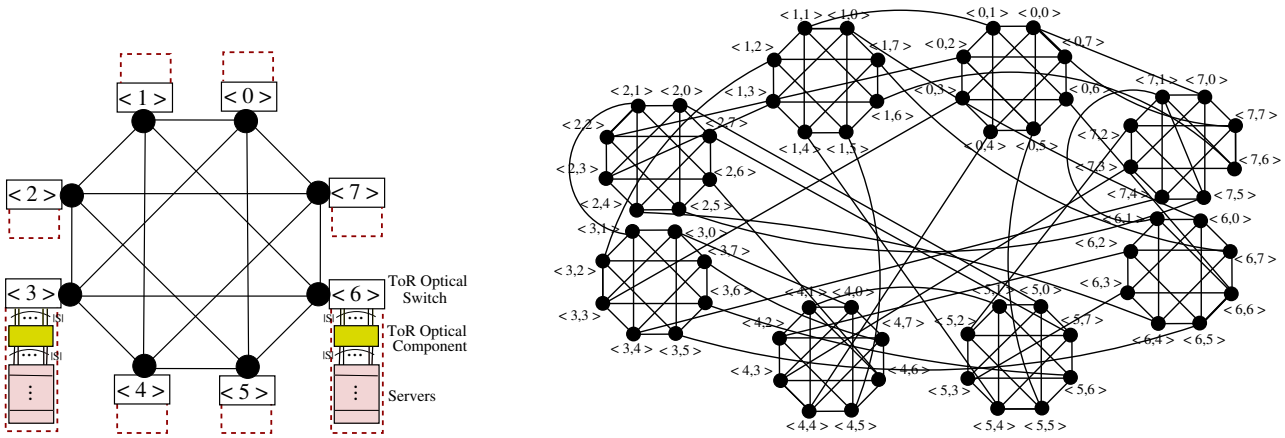
Most of the existing DCN architectures using optical networks are *switch-centric* hierarchical tree based structures. Electrical packet switching (EPS) and OCS are used in these architectures for supporting low and high volumes of data transfer respectively [11]-[13]. It is reported that hierarchical tree based DCN architectures do not have good scalability and performance [4]-[8]. Moreover, the use of electrical networks can also restrict the exploitation of optical networks in these architectures [14]. In grid environments, *hybrid optical networks* integrating OCS and OBS technologies are reported to have good performance in transmitting high and low volumes of data respectively [18]-[20].

In this paper, we present *HyScaleII*, a *switch-centric* high performance DCN architecture using hybrid optical networks. The objective is to improve the performance of *HyScale* that we proposed recently [17]. We accomplish our objective by tweaking the topology of HyScale [17] in order to lower the expected load per link in HyScaleII and thus lowering the probability of packet loss. HyScaleII still retains most of the desirable properties of a data center, e.g. scalability, fault tolerance, high bisection width, low network complexity, and low diameter. Based on the structural properties of HyScaleII, we also present an efficient and simple routing scheme called *HySII* routing. In our experiments, HyScaleII has lower packet loss ratio and higher average aggregate throughput by an average of 50% and 13.8% respectively as compared to HyScale [17]. Therefore, HyScaleII achieves higher traffic balance and has better performance than HyScale [17].

II. Related Work

BCube [4], DCell [5], FiConn [6], DPillar [7], and BCN [8] are examples of *server-centric* DCN architectures. In these architectures, servers are used as both computing and routing nodes. The switches in these architectures are never directly connected. Whereas, the servers are connected using different switches. Thus, servers are used for relaying data in such DCN architectures. Most of these architectures are recursively defined and are highly scalable. The recursive definition of

978-1-4673-1466-4/12/$31.00©2012 Crown



(a) Basic Construction Unit, $\Psi(0, \Omega, 8)$

(b) Design of $\Psi(1, \Omega, 8)$

Fig. 1. HyScaleII Topology Design

these architectures embeds the concept of *locality*, i.e. many servers are in close proximity of each other [21]. This may increase the communication efficiency of these architectures.

In *switch-centric* architectures, the servers are typically placed in a *rack* and are connected with the "Top of Rack" (ToR) switch. One or more levels of switches are used to interconnect the ToR switches. Thus, servers are not needed for routing the data and can be turned on and off as only needed for computation. Typically each rack has around 10 to 40 or more servers [12]. Fat-Tree [1], VL2 [2], and PortLand [3] are examples of such DCN architectures. Most of the *switch-centric* architectures are hierarchical tree based structures. The size of the switches may significantly increase with an increase in the number of servers supported by these architecture. This can restrain the scalability of these architectures [4].

In order to guarantee high bandwidth communication, DCN architectures provide full bisection bandwidth between several pairs of nodes [11]-[14]. It has been observed from data center traffic traces that 80% of the flows are smaller than 10KB in size and 95% of the bytes transferred are in the top 10% of the *elephant flows* (flows with large amount of data) [22]-[24]. Thus, full bisection bandwidth between several pairs of nodes is rarely required [11]-[14]. OCS networks have been propagated as an option to transmit *elephant flows* in data centers [9]-[16]. Helios [11], HyPaC [12], and Proteus [13] are examples of such DCN architectures. These architectures employ OCS and EPS for transmitting *elephant* and *mice* flows (flows with low amount of data) respectively and are *switch-centric* hierarchical tree based structures.

III. Structure of HyScaleII

In this section we give the construction of HyScaleII and briefly review the structure of HyScale [17].

To exploit the high-bandwidth of optical networks, the ToRs are optically interconnected with each other [12]-[13]. Each ToR has an optical crossconnect and an *optical component* [12]-[13] for connecting the servers in the corresponding rack with the crossconnect. The optical crossconnects can employ OCS and OBS technologies for transmitting *elephant* and *mice* flows respectively. Optical Packet Switching (OPS) is not yet deployed in practical optical networks [12], [18], and thus OBS is the preferred choice for transmitting *mice* flows. For brevity, ToR optical crossconnect is denoted as ToR.

HyScaleII is a recursively defined topology denoted by $\Psi(k, \Omega, \Gamma)$, where k is the number of levels of the topology, and Γ (power of 2) is the number of sub-topologies defining Ψ. Ω defines the address space of all nodes (or ToRs) in Ψ, where the address of each node is a $(k+1)-$tuple. $\Psi(k, \Omega, \Gamma)$ is obtained by interconnecting Γ of $\Psi(k-1, \Omega, \Gamma)$ sub-topologies. This recursive definition embeds the concept of *locality* [21] in Ψ. $\Psi(0, \Omega, \Gamma)$ is the basic construction unit of HyScaleII.

A. Basic Construction Unit, $\Psi(0, \Omega, \Gamma)$

$\Psi(0, \Omega, \Gamma)$ is a vertex-transitive and edge-transitive graph which is isomorphic to $K_{\Gamma/2,\Gamma/2}$ (complete bipartite graph). The address space of each node in $\Psi(0, \Omega, \Gamma)$ is defined by a $1-$tuple, i.e. $\Omega = \{\langle i_0 \rangle | i_0 \in [0, 1, \ldots, \Gamma - 1]\}$. The links (or edges) in $\Psi(0, \Omega, \Gamma)$ are called $level_0$ links. Fig. 1(a) shows the basic construction unit of Ψ when $\Gamma = 8$.

B. Construction of $\Psi(k, \Omega, \Gamma)$ $(k \geq 1)$

$\Psi(k, \Omega, \Gamma)$ is a symmetric, vertex-transitive, and regular graph. A level k topology of Ψ, $\Psi(k, \Omega, \Gamma)$ is constructed by interconnecting Γ of $\Psi(k-1, \Omega, \Gamma)$ sub-topologies. The address space of each node in $\Psi(k, \Omega, \Gamma)$ is defined as a $(k+1)-$tuple, i.e. $\Omega = \{\langle i_k, \ldots, i_0 \rangle | i_l \in [0, 1, \ldots, (\Gamma - 1)], \forall l = 0, \ldots, k\}$. Two nodes $I = \langle i_k, \ldots, i_0 \rangle$ and $J = \langle j_k, \ldots, j_0 \rangle$ in $\Psi(k, \Omega, \Gamma)$ are connected by a $level_k$ link if $F(i_k, i_0) = j_k$, and $i_0 = j_0$, where function F is defined as

$$F(x, y) = \begin{cases} x \oplus y & \text{If} \quad y \neq 0 \\ \overline{x} & \text{Otherwise} \end{cases} \tag{1}$$

Fig. 1(b) shows the design of $\Psi(1, \Omega, 8)$.

C. Structure of HyScale [17]

HyScale is a recursively defined topology denoted by $\Phi(k, \Omega, \Gamma)$. It is a symmetric, vertex-transitive, regular, and bi-partite graph. $\Phi(0, \Omega, \Gamma)$ is isomorphic to $K_{\Gamma/2,\Gamma/2}$. $\Phi(k, \Omega, \Gamma)$ $(k \geq 1)$ is constructed by interconnecting Γ of $\Phi(k-1, \Omega, \Gamma)$ sub-topologies, and two nodes $I = \langle i_k, \ldots, i_0 \rangle$ and $J =$

TABLE I
COMPARISON BETWEEN DIFFERENT DATA CENTER NETWORK ARCHITECTURES

	HyScaleII	HyScale [17]	Fat-Tree [1]	BCube [4]	DCell [5]	FiConn [6]
Number of Servers	$S\Gamma^{k+1}$	$S\Gamma^{k+1}$	$\frac{m^3}{4}$	n^{k+1}	$(n+1)^{2^k}-1$	$2^{k+2}\frac{n}{4}2^k$
Number of Switches	Γ^{k+1}	Γ^{k+1}	$\frac{5m^2}{4}$	$n^k(k+1)$	$\frac{(n+1)^{2^k}-1}{n}$	$\frac{2^{k+2}\frac{n}{4}2^k}{n}$
Diameter (or bound ($\star$))	$3k+2$ ($\star$)	$4k+2$ ($\star$)	4	$k+1$	$2^{k+1}-1$ ($\star$)	$2^{k+1}-1$ ($\star$)
Bisection Width	$\frac{\Gamma^{k+1}}{4}$ $(k \geq 1)$; $\left\lfloor \frac{\Gamma^2}{8} \right\rfloor$ $(k=0)$	$\frac{\Gamma^{k+1}}{4}$ $(k \geq 1)$; $\left\lfloor \frac{\Gamma^2}{8} \right\rfloor$ $(k=0)$	$\frac{m^3}{8}$	$\frac{n^{k+1}}{2}$	$\frac{(n+1)^{2^k}-1}{4*\log_n((n+1)^{2^k}-1)}$	$\frac{2^{k+2}\frac{n}{4}2^k}{4*2^k}$

k is the number of levels of the recursive structure; n is the number of servers in the basic construction unit of BCube [4], DCell [5], and FiConn [6]; Γ is the number of ToRs in the basic construction unit of HyScale [17] and HyScaleII; S is the number of servers in a rack; m is the number of pods in Fat-Tree [1]

$\langle j_k, \ldots, j_0 \rangle$ in $\Phi(k, \Omega, \Gamma)$ are connected by a $level_k$ link if $F_1(i_k) = j_k$, $F_2(i_k) = i_0$ or $F_3(i_k) = i_0$, and $i_0 = j_0$, where the functions F_1, F_2, and F_3 are defined as

$$F_1(i) = (2.z + 1 + i) \bmod \Gamma, \text{ where } z \in [0, 1, \ldots, \Gamma/2 - 1]$$
$$F_2(i) = (z + i \bmod \Gamma/2) \bmod \Gamma, \text{where z satisfies } F_1(i_k) = j_k$$
$$F_3(i) = (F_2(i) + \Gamma/2\ ) \bmod \Gamma \quad (2)$$

IV. ROUTING IN HYSCALEII

In this section, we present a routing scheme for HyScaleII, called *HySII* routing. Similar to *HyS routing* for HyScale [17], *HySII* routing exploits the structural properties of HyScaleII topology (or Ψ). Thus, comparing the performance of *HySII* and *HyS* routing schemes is in fact comparing the performance of HyScaleII and HyScale [17] topologies.

If the source and destination servers are in the same rack, then the intra-rack communication does not use any link in Ψ. Otherwise for all other inter-rack communications, a path between two ToRs (or nodes) in Ψ will be computed. Given a pair of source and destination nodes in $\Psi(k, \Omega, \Gamma)$, if they are in same $\Psi(k-1, \Omega, \Gamma)$ sub-topology, then the route between them will be contained in that $\Psi(k-1, \Omega, \Gamma)$ sub-topology. Otherwise, the route between them will include exactly one $level_k$ link for traversing between the two $\Psi(k-1, \Omega, \Gamma)$ sub-topologies. Such $level_k$ links may not be incident on the source node. But, a node with such a $level_k$ link is always at most two $level_0$ links away. A high-level description of the routing algorithm for $\Psi(k, \Omega, \Gamma)$ is given in Algo. 1. Its time complexity is $O(k)$.

Algorithm 1 HySII Routing

// src = $\langle i_k, i_{k-1}, \ldots, i_1, i_0 \rangle$; dst = $\langle j_k, j_{k-1}, \ldots, j_1, j_0 \rangle$
// (a,b) denotes the link between nodes a and b
// path = **Route(src,dst,k)**
Route(src,dst,l)
if src == dst **then**
  return
else
  if (l == 0) **then**
    if src and dst are connected with a $level_0$ link **then**
      return (src,dst)
    else
      temp = $\langle i_k, i_{k-1}, \ldots, i_1, (i_0 + 1) \bmod \Gamma \rangle$
      return (src,temp) + **Route(temp,dst,0)**
    end if
  else if ($i_l \neq j_l$) **then**
    if ($F(i_l, i_0) == j_l$) **then**
      temp = $\langle j_k, j_{k-1}, \ldots, j_{l+1}, j_l, i_{l-1}, \ldots, i_1, i_0 \rangle$
      return (src,temp) + **Route(temp,dst,l-1)**
    else
      $x = F(i_l, j_l)$
      temp1 = $\langle j_k, j_{k-1}, \ldots, j_{l+1}, i_l, i_{l-1}, \ldots, i_1, x \rangle$
      temp2 = $\langle j_k, j_{k-1}, \ldots, j_{l+1}, j_l, i_{l-1}, \ldots, i_1, x \rangle$
      return **Route(src,temp1,0)** + (temp1,temp2)
        + **Route(temp2,dst,l-1)**
    end if
  else
    return **Route(src,dst,l-1)**
  end if
end if

V. PROPERTIES OF HYSCALEII

In this section, we show that HyScaleII is scalable, fault tolerant, and has low network complexity. The network diameter, the switch size, and the number of disjoint routes between any pair of ToRs in HyScaleII all increase linearly with an exponential increase in the number of servers. Thus, HyScaleII is highly scalable and fault-tolerant. The topology of HyScaleII can be realized with small size switches. Therefore, HyScaleII has low overall network complexity [12]. The following theoretical results support the above assertions. The proofs are omitted due to space limitations.

Theorem 5.1: Total number of nodes in $\Psi(k, \Omega, \Gamma)$ is Γ^{k+1}.

Corollary 5.2: Assuming each rack has S servers, the total number of servers in $\Psi(k, \Omega, \Gamma)$ is $S\Gamma^{k+1}$.

Theorem 5.3: $\Psi(k, \Omega, \Gamma)$ is symmetric, vertex-transitive, and $(\frac{\Gamma}{2} + k)-$regular graph.

Corollary 5.4: The size of the ToR optical crossconnects in Ψ is $W(\frac{\Gamma}{2} + k) + S$.

Theorem 5.5: $\Psi(k, \Omega, \Gamma)$ requires $\frac{\Gamma^{k+1}(\frac{\Gamma}{2}+k)}{2}$ links for interconnecting the nodes.

Corollary 5.6: $\Psi(k, \Omega, \Gamma)$ has $\frac{\Gamma^{k+2}}{4}$ $level_0$ links and $\frac{\Gamma^{k+1}}{2}$ $level_i$ links $\forall i = 1, \ldots, k$.

Theorem 5.7: $\Psi(k, \Omega, \Gamma)$ is $(\frac{\Gamma}{2} + k)-$vertex connected.

Theorem 5.8: Diameter of $\Psi(k, \Omega, \Gamma)$ is bounded by $(3k+2)$ links.

Theorem 5.9: Bisection width of $\Psi(k, \Omega, \Gamma)$ is $\frac{\Gamma^{k+1}}{4}$ $(k \geq 1)$.

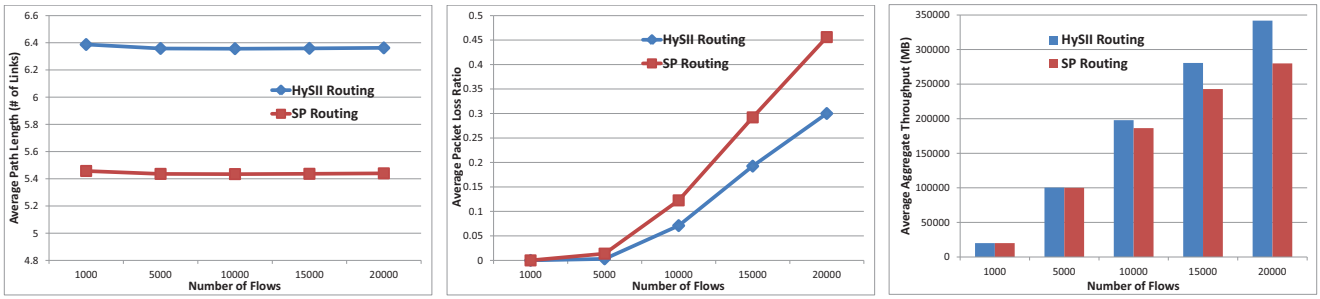
Corollary 5.10: Bisection width of $\Psi(0, \Omega, \Gamma)$ is $\left\lfloor \frac{\Gamma^2}{8} \right\rfloor$.

We believe the following result holds but do not have the complete proof. Thus, it is presented as a conjecture.

Conjecture 5.11: Diameter of $\Psi(k, \Omega, \Gamma)$ $(k \geq 1)$ is $(3+2k)$.

A. Comparison of HyScaleII with other DCN architectures

Table I shows a comparison of few important properties of HyScaleII with other DCN architectures.



(a) Average Path Length (b) Average Packet Loss Ratio (c) Average Aggregate Throughput

Fig. 2. Comparison between HySII and SP routing schemes

Proposition 5.12: HyScaleII has better expected performance as compared to HyScale [17].

Proof: The total number of links and nodes in both HyScaleII and HyScale [17] are $\frac{\Gamma^{k+1}(\frac{\Gamma}{2}+k)}{2}$ and Γ^{k+1} respectively. The maximum number of flows in these topologies is $\Gamma^{k+1}(\Gamma^{k+1}-1)$. The diameter of HyScaleII and HyScale [17] is bounded by $3k$+2 and $4k$+2 links respectively. As all links are identical, the maximum number of links traversed by the maximum number of flows in HyScaleII and HyScale [17] are $\Gamma^{k+1}(\Gamma^{k+1}-1)(3k+2)$ and $\Gamma^{k+1}(\Gamma^{k+1}-1)(4k+2)$ respectively. Therefore, each link in HyScaleII carries at most about $\frac{2(\Gamma^{k+1}-1)(3k+2)}{\frac{\Gamma}{2}+k}$ flows. Whereas, each link in HyScale [17] carries at most about $\frac{2(\Gamma^{k+1}-1)(4k+2)}{\frac{\Gamma}{2}+k}$ flows. Thus, HyScaleII has lower maximum load per link and is expected to have lower packet loss probability as compared to HyScale [17]. ■

VI. Experimental Results and Discussion

In this section, we describe the experiments and analyze the results obtained. The results represent an average of 20 runs and all results have 95% confidence level.

A. Experimental Setup

We simulate a $\Psi(k,\Omega,\Gamma)$ topology with Γ=8, k=3, and S=32. Therefore, HyScaleII (or Ψ) has 131,072 servers and 4,096 ToRs. Each link in Ψ is a bidirectional fiber with uniform transmission capacity of 1Gbps, and has four wavelengths per fiber. Thus, each ToR is a $60-$port optical cross-connect with no fiber delay lines or wavelength converters.

Inter-rack traffic are the most likely candidates for over-subscription [24] and intra-rack traffic will not use the links in Ψ. Therefore, we only consider inter-rack traffic in our simulation. Real data center traffic traces show that there are at most 10,000 active flows at any given interval, 80% of the flows are smaller than 10KB in size, and 95% of the bytes transferred are in the top 10% of the large flows [22], [24]. Accordingly, we consider the average size of a flow is exponentially distributed with mean 20MB. The source and destination of the flows are uniformly distributed among all the ToRs. The flows with more than 25MB of data are classified as *elephant* flows [23]. At least 20% of the flows at any given interval are *elephant* flows in our simulation.

As discussed earlier, OCS and OBS are used for transmitting *elephant* and *mice* flows respectively. The burst header is processed all-optically by the intermediate nodes [25] and is encapsulated in the burst [20]. The burst assemble time is 20ms, burst processing time is 20μs, packet processing time is 1μs, and OCS switching time is 10ms [15], [20]. The maximum length of a burst and the packet size are fixed at 12.5 MB and 12.5 KB respectively [20].

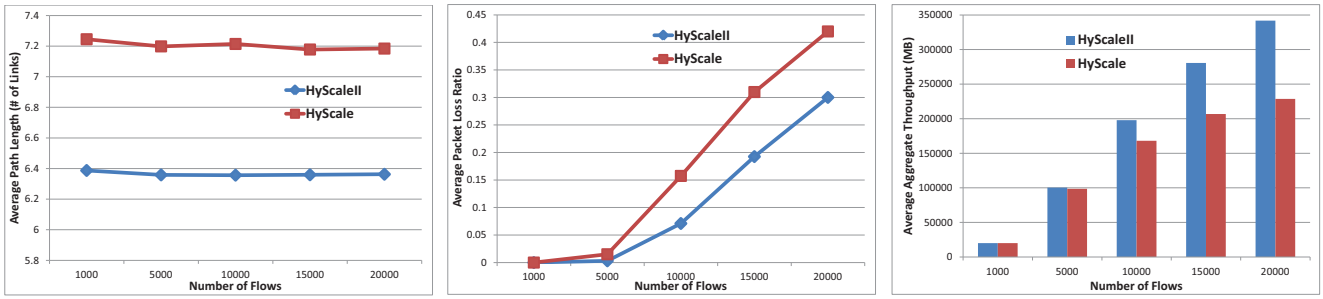
B. Comparison of HySII Routing with Shortest Path Routing

HyS routing was reported to have better performance than *Shortest Path* (SP) routing in HyScale [17]. Thus, it is expected that *HySII* routing will also have better performance than *SP* routing in HyScaleII. Below, we compare the performance of *HySII* and *SP* routing schemes.

The route of a circuit for an *elephant* flow is computed at the source. If the circuit can not be successfully established, then all the packets in the flow are dropped. Otherwise the circuit is reserved for the duration of the flow. The *mice* flows are assembled into one or more bursts at the source. The route of a burst is also computed at the source and is encapsulated in the burst header. Just-Enough-Time (JET) protocol [18] is used to schedule the bursts. If a burst is blocked due to contention, then all the packets in the bursts are dropped. The packet loss ratio (PLR) is defined as the ratio of the total number of packets dropped to the total number of packets in all the flows.

Fig. 2(a) shows a comparison of the average path length traversed by the packets using *HySII* and *SP* routing schemes. The path length is defined as the number of links (or fibers) traversed by the packets between the ToRs. *HySII* routing does not always compute the shortest path between the source and destination nodes. Thus the average path length of *SP* routing is always less than that of *HySII* routing. The time complexity of *HySII* routing is $O(k)$ (Algo. 1). Whereas, the time complexity of *SP* routing is $O(\Gamma^{2(k+1)})$. Moreover, *HySII* routing exploits the structural properties of HyScaleII and computes a route of at most $k-1$ links longer than the shortest path. As compared to the number of nodes in HyScaleII (Γ^{k+1}), $k-1$ is a negligible number. For example, when $\Gamma=8$, the time complexity of *HySII* routing is $O(k)$, while that of *SP* routing is $O(2^{6(k+1)})$. In our experiments with $k=3$, the difference in the average path length of the two routing schemes is always less than a link. Thus there is a negligible increase in the average end-to-end delay of *HySII* routing as compared to the *SP* routing.

The performance of the two routing schemes in terms of PLR is shown in Fig. 2(b). *HySII* routing has a lower PLR in these experiments. Thus, it achieves better traffic balance



(a) Average Path Length (b) Average Packet Loss Ratio (c) Average Aggregate Throughput

Fig. 3. Comparison between HyScaleII and HyScale [17] architectures

with negligible increase in the length of the routes. In our experiments, *HySII* routing has a lower PLR by an average of 52% as compared to the *SP* routing. As expected, the PLR of both the routing schemes increases with an increase in the maximum number of flows.

Fig. 2(c) shows the comparison of the average aggregate throughput of the routing schemes. As expected, the average aggregate throughput increases with an increase in the number of flows. The lower PLR of *HySII* routing also results in higher average aggregate throughput as compared to *SP Routing*. In our experiments, *HySII* routing has a higher average aggregate throughput by an average of 5.5% as compared to the *SP* routing. Therefore, *HySII* routing has better performance and better traffic balance as compared to the *SP* routing.

C. Performance Comparison of HyScaleII with HyScale [17]

In this section, we compare the performance of HyScaleII with HyScale [17]. As discussed earlier, comparison between *HySII* and *HyS* [17] routing schemes in fact is a comparison between HyScaleII and HyScale [17] topologies. Thus, for comparing the two topologies, we simulate HyScale topology ($\Phi(3, \Omega, 8)$) [17] with equal number of servers and ToRs as that in HyScaleII and use *HySII* and *HyS* [17] routing schemes respectively in HyScaleII and HyScale [17] topologies.

As shown in Fig. 3(a), HyScaleII has a lower average path length than HyScale. This is due to the fact that the bound on the diameter of HyScaleII is smaller than that of HyScale [17]. Fig. 3(b) compares the performance of the two architectures in terms of PLR. Thus, the empirical results are in line with the proposition 5.12. HyScale has a higher number of bottlenecks resulting in higher PLR. As expected, PLR increases with an increase in number of flows. In our experiments, HyScaleII has a lower PLR by an average of 50% as compared to HyScale [17]. Lower PLR in HyScaleII also results in its higher average aggregate throughput, shown in Fig. 3(c). In our experiments, HyScaleII has a higher average aggregate throughput by an average of 13.8% as compared to HyScale [17]. Therefore, HyScaleII has better performance than HyScale [17].

VII. Conclusion

We propose HyScaleII to improve the performance of HyScale [17]. HyScaleII is a highly scalable, recursively defined, fault-tolerant DCN topology with low network complexity. We also present an efficient and simple routing scheme called *HySII* routing, which exploits the structural properties of HyScaleII. In our experiments, HyScaleII has lower PLR and higher average aggregate throughput by an average of 50% and 13.8% respectively as compared to HyScale [17]. In future, we want to develop a testbed and analyze HyScaleII's performance with real data center traffic traces.

References

[1] M. Al-Fares, A. Loukissas, A. Vahdat, "A scalable, commodity data center network architecture," *ACM SIGCOMM*, pp.63–74, 2008.

[2] A. Greenberg, *et al.*, "VL2: a scalable and flexible data center network," *ACM SIGCOMM*, pp.51–62, 2009.

[3] R. Mysore, *et al.*, "PortLand: A Scalable Fault-Tolerant Layer 2 Data Center Network Fabric," *ACM SIGCOMM*, pp.39–50, 2009.

[4] C. Guo, *et al.*, "BCube: A High Performance, Server-centric Network Architecture for Modular Data Centers," *SIGCOMM*, pp.63–74, 2009.

[5] C. Guo, *et al.*, "Dcell: A Scalable and Fault-tolerant Network Structure for Data Centers," *ACM SIGCOMM*, pp.75–86, 2008.

[6] D. Li, *et al.*, "Scalable and Cost-Effective Interconnection of Data-Center Servers Using Dual Server Ports," *IEEE/ACM Trans. on Netwk.*, vol.19, no.1, pp.102–114, 2011.

[7] Y. Liao, D. Yin, L. Gao, "DPillar: Scalable Dual-Port Server Interconnection for Data Center Networks," *IEEE ICCCN*, 2010.

[8] D. Guo, *et al.*, "BCN: Expansible network structures for data centers using hierarchical compound graphs," *IEEE INFOCOM*, 2011.

[9] R. Ho, *et al.*, "Optical systems for data centers," *OFC/NFOEC* 2011.

[10] K. Barker, *et al.*, "On the Feasibility of Optical Circuit Switching for High Performance Computing Systems," *IEEE Supercomputing*, 2005.

[11] N. Farrington , *et al.*, "Helios: a hybrid electrical/optical switch architecture for modular data centers," *ACM SIGCOMM*, pp.339–350, 2010.

[12] G. Wang, *et al.*, "c-Through: part-time optics in data centers," *ACM SIGCOMM*, pp.327–338, 2010.

[13] A. Singla, A. Singh, K. Ramachandran, L. Xu, Y. Zhang, "Proteus: a topology malleable data center network," *HotNets*, 2010.

[14] M. Glick, "Optical switching and routing for the data center," *IEEE Photonics Society Winter Topicals Meeting Series (WTM)*, 2010.

[15] A. Vahdat, H. Liu, X. Zhao, C. Johnson, "The emerging optical data center," *OFC/NFOEC*, 2011.

[16] L. Schares, D. Kuchta, A. Benner, "Optics in Future Data Center Networks," *IEEE HOTI*, 2010.

[17] S. Saha, J. Deogun, L. Xu, "HyScale: A Hybrid Optical Network based Scalable, Switch-centric Architecture for Data Centers," *accepted to appear in IEEE ICC*, pp.1–6, 2012.

[18] C. Xin, C. Qiao, Y. Ye, S. Dixit, "A hybrid optical switching approach," *IEEE GLOBECOM*, vol.7, pp.3808–3812, 2003.

[19] C. M. Gauger, *et al.*, "Hybrid optical network architectures: bringing packets and circuits together," *IEEE Comm Mag*, vol.44, pp.36–42, 2006.

[20] Y. Wang, S. Wang, S. Xu, X. Wu, "A new hybrid optical network design consisting of lightpath and burst switching," *Int. Conf. on Advanced Communication Technology (ICACT)*, vol.3, pp.1873–1876, 2009.

[21] Y. Zhang, A. Su, G. Jiang, "Evaluating the impact of data center network architectures on application performance in virtualized environments," *IEEE IWQoS*, 2010.

[22] T. Benson, A. Akella, D. Maltz, "Network traffic characteristics of data centers in the wild," *ACM IMC*, 2010.

[23] A. Curtis, K. Wonho, P. Yalagandula, "Mahout: Low-overhead datacenter traffic management using end-host-based elephant detection," *IEEE INFOCOM*, pp.1629–1637, 2011.

[24] S. Kandula, J. Padhye, P. Bahl, "Flyways To De-Congest Data Center Networks," *HotNets*, 2010.

[25] Y. Liu, *et al.*, "Ultrafast all-optical signal processing: towards optical packet switching," *Proc. SPIE*, vol. 6353, pp. 635312–8, 2006.