2015

Usage Based Topology for DCNs

Qing Yi
Portland State University

Suresh Singh
Portland State University

Usage Based Topology for DCNs

Qing Yi
Department of Computer Science
Portland State University
Portland, Oregon 97207
Email: yiq@cs.pdx.edu

Suresh Singh
Department of Computer Science
Portland State University
Portland, Oregon 97207
Email: singh@cs.pdx.edu

***Abstract*—Fat-tree network topology can support full bisection bandwidth for hundreds of thousands of servers in data centers and makes it possible for high network throughput and server agility. However, the low utilization rate of DCNs (average 10%)**

I. Introduction

One of the primary design goals of many Data Center Networks (DCN) has been to provide very high throughput at maximum (or 100%) loading of the links at a low cost. This led to the deployment of fat-tree network topologies that provide full bisection bandwidth between each layer of the network and can be built using identical sized ncommodity switches. Over the last few years, some studies of DCN loading have appeared in the literature which show that typical DCNs are greatly under-utilized with 10% loading being a common occurrence [**?**]. Relatd studies have also shown that the traffic flows between servers varies greatly depending on the application domain of the data center. Combining these observations together, it is clear that the original design goal of full bisection bandwidth is overly pessimistic and leads to high capital cost as well as electricity cost as a majority of network equipment remains unused or under-utilized. In order to address the issue of electricity cost, researchers have proposed that idle network equipment be in low power mode to save energy [**?**]. Some use left-most routing to consolidate the traffic to the left and keep active a smaller sub-tree [**?**]. While these approaches do save energy, we believe that it is more appropriate to consider redesigning the network so as to use fewer switches while still providing needed performance for *realistic loads*.

In this paper we systematically construct a DCN that supports the expected loading for different application domains but has lower cost. Specifically,

1) We first begin with fat-tree DCNs and examine the sub-graph of these networks that is used for loading as high as 70% for different types of applications (educational, cloud, and private data centers) when using left-most routing as in [?]. The results indicate that we can indeed reduce the number of switches at different levels of the network without incurring any loss or increased latency.
2) Next we consider the possibility of moving flows to fewer servers, particularly for low loads. This approach is interesting since it can inform job schedulers about how and where to place jobs in order to minimize network energy cost. Consolidating flows further reduces the needed switches in the network.
3) Our analysis shows that edge switches (i.e., switches connected to servers) account for a high energy cost as they are always powered on, even at tiny loads. Given that a significant energy cost of a switch is static (in the chassis, power supply, processor, interconnect fabric), by using high cardinality switches (and thus fewer switches) we can save significant amount of energy even if they are always on.

Putting all these studies together we obtain a new DCN in which edge switches have high-port density and where the other switches are connected in a left-skewed topology which is a subgraph of the fat-tree. This type of topology has a lower capital cost and lower operational cost as well.

The remainder of the paper is organized as follows. In Section II, we formulate the mathematical model of the number of switches in each layer of a fat-tree network for given traffic load and pattern. Section II-A shows the sun-trees generated from simulations with traffic from diffent types of data centers. In Section III, we propose an approach of using high-radix edge switches to further consolidate the traffic and the resulting sub-trees show he number of active switches and links are reduced to half. We compare the power consumption using the reality power data for commercial switches in Section 12 and show that using high-radix edge switches achieves better overall energy efficiency of CDCNs. Section VI introduces some related work and background and Section VII summarizes the conclusions and the main contributions.

II. Sub-Trees for Different Traffic Patterns and Loadings

A k-ary fat-tree consists of k pods in which each pod has $k/2$ edge switches and $k/2$ aggregation switches. Each edge switch is connected to $k/2$ servers and to each of the $k/2$ aggregation switches. There are $k^2/4$ core switches and each aggregation switch in each pod is connected to $k/2$ core switches. Thus, a DCN contains $5/4k^2$ switches with k ports each.

A. Traffic Model

Benson *et al.* [**?**] analyzed traffic data characteristics of ten data centers, including three university data centers (EDU), two private enterprise IT data centers (PRV) and five commercial cloud data centers (CLD). The EDU data centers serve students and staff on campus. The main applications in EDU data centers include distributed file systems and Web services. The private enterprise IT data centers mainly serve corporate users

and developers. Besides hosting traditional Web services, these data centers also run customized applications. The cloud data centers are purposely-built to support specific applications and serve external users. For example, two of the CLD data centers are mainly running MapReduce style jobs and the other three are primarily running Internet-facing applications, including Messaging, Webmail, Web portal and searching.

By observation, a significant part of the traffic in the EDU data centers is distributed file system traffic across the entire network. On average, about 30% of the three EDU data center traffic is within the same rack. The applications in private data centers have shown a degree of emerging patterns of consolidation and virtualization and around 45% of the traffic is within the same rack. The MapReduce job in the cloud data centers is scheduled to pack into the same rack to reduce core interconnect and nearly 75% of the traffic is confined in the same rack.

For our study, we need to have traffic traces that not only follow these different patterns of EDU, PRV and CLD but also have different loading. Therefore, we created traffic generators for each of these types of traffic and fed this traffic to our fat-tree simulator. Traffic in a fat-tree can be characterized by two probabilities – p_1 and p_2. p_1 denotes the probability that the destination of a packet is connected to the same edge switch. Of the other packets, there is a probability p_2 that their destination is within the same pod and thus will need to traverse an aggregation switch. The rest of the packets are destinated to servers in other pods and thus need to pass through a core switch. By varying the probabilities p_1 and p_2 we can simulate different types of traffic models. We generate synthetic traffic traces using an On/Off process with the On period and Off period following the lognormal distributions. The packet interarrival time is also a lognormal process (this is based on the results of Benson *et al* [?]). The source and destination nodes are chosen *uniformly* from servers in each pod. The specific parameters we use are:

- EDU ($p_1 = 0.3$, $p_2 = 0.2$)
- EDU1 ($p_1 = 0.3$, $p_2 = 0.5$)
- PRV ($p_1 = 0.45$, $p_2 = 0.2$)
- CLD ($p_1 = 0.75$, $p_2 = 0.2$)

In order to study the benefits of consolidating jobs into fewer servers, we distribute different traffic loads to each pod. For example, say we have 12 pods. We generate 70% pod load for the first five pods and 10% pod load for the sixth pod and leave all other pods with zero traffic coming in and out. The overall load will be 10% for the entire network. This is called the *non-uniform* case. For each of the traffic models we create seven loads from 10% to 70% of network capacity.

For our simulations, we assume a $k = 12$ fat-tree data center network that supports 432 end hosts, which are connected via 180 12-port switches. These switches are grouped in 12 pods and each pod contains six edge switches and six aggregation switches. The core layer consists of 36 core switches.

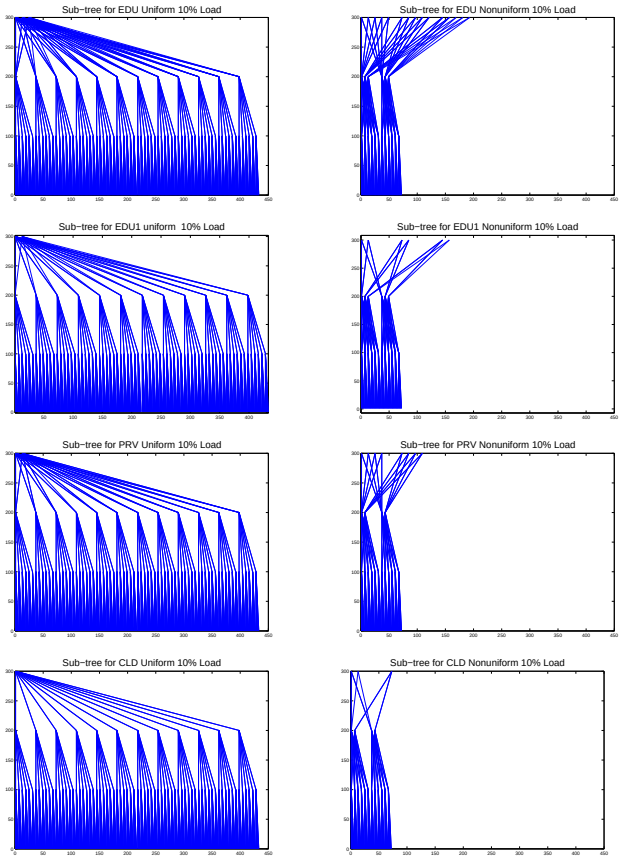

Fig. 1. Minimum fat-tree with uniform and nonuniform traffic of load 10%.

B. Active Sub-Trees

We feed the synthetic packet traces to the simulated fat-tree and use left-most routing. The sub-trees obtained are illustrated in Figure 1 and 2. We demonstrate the sub-trees with two border loads of 10% and 70% for all traffic patterns. The fraction of the total number of switches and links needed are shown in Figure 3. As we can see, when the uniform load is as low as 10%, a minimum spanning tree is sufficient for the cloud data center because only around 25% of the traffic leaves the rack. Even when the load increases to 70%, there are still a significant number of switches and links in idle state. For the nonuniformly distributed traffic, it is obvious that if we can pack the communicating jobs into fewer number of pods, a large number of edge and aggregation switches can be powered off. The number of core switches is determined by the heaviest inter-pod load. Therefore, the job placement should target to evenly sharing the inter-pod traffic among all the involved pods.

The fraction of total switches and links are shown in Figure 3. For evenly distributed traffic, less than 50% of switches and links of the fat-tree are needed. 20% of switches and 25% of links will never be used when the traffic load is less than 70%. The cloud data centers are usually designed with meticulous job placement in order to decrease cross-pod traffic, so the fraction of idle switches and links is even greater. For the nonuniformly distributed traffic, even fewer switches and links are used since jobs are consolidated into fewer pods.

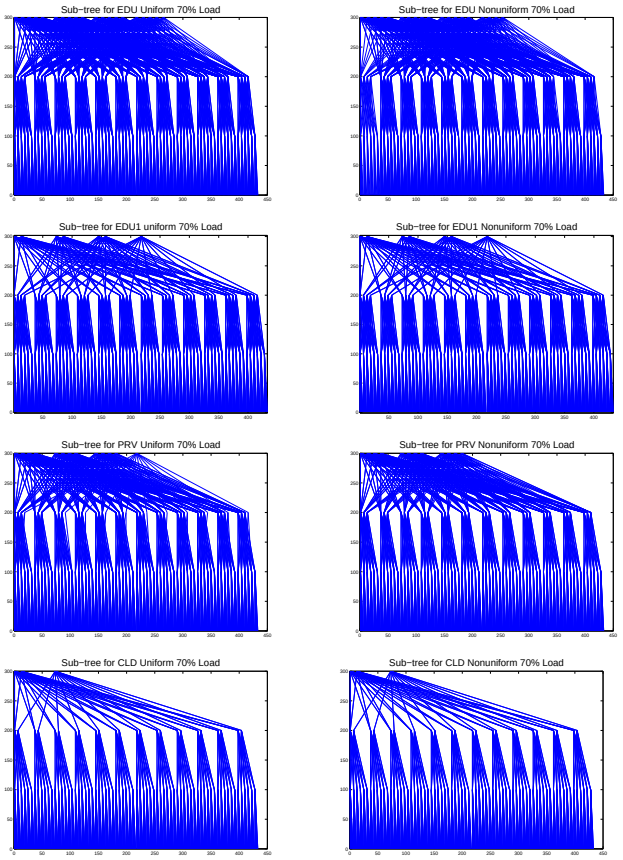

Fig. 2. Minimum fat-tree with uniform and nonuniform traffic of load 70%.

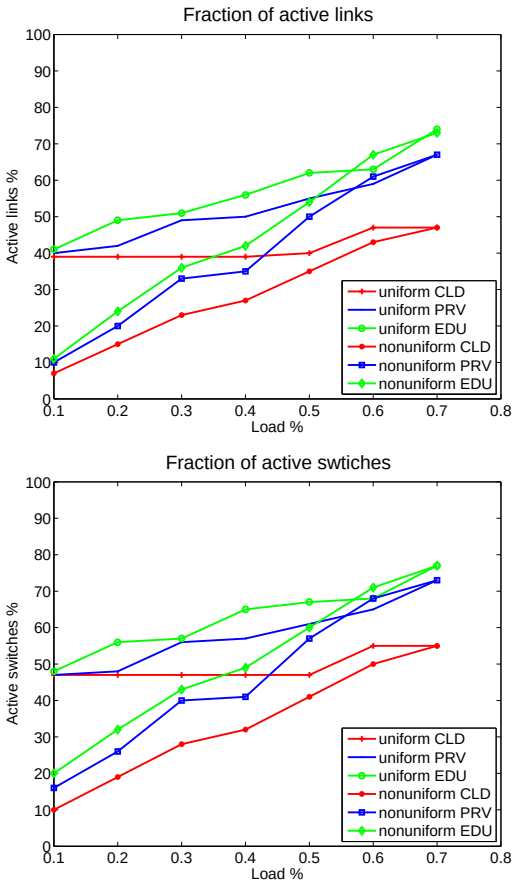

Fig. 3. Number of active links and active switches for uniform and nonuniform traffic load.

C. Analytical Model of Sub-Tree Size

The simulations above clearly show that significant parts of a typical fat-tree can be dispensed with without affecting performance. However, the simulation results were conducted for relatively small DCNs. In this section, we provide a theoretical model that can be generalized to arbitrary sized DCNs. Let us assume that the traffic load generated by the k pods is $\lambda_1, \lambda_2, ..., \lambda_k$ (represented as a fraction of full load). We use two parameters p_1 and p_2 to represent the probability of traffic travelling between servers connected to the same edge switch, and traffic travelling within the same pod, respectively. Therefore, $(1-p_1-p_2)$ of traffic goes to other pods. We assume the p_1 and p_2 for each pod are written as $p_1^1, p_1^2, ..., p_1^k$ and $p_2^1, p_2^2, ..., p_2^k$. Then for pod i, λ_i of traffic load is generated of which $\lambda_i(1-p_1^i)$ goes up to the aggregation layer and $\lambda_i(1-p_1^i-p_2^i)$ arrives the core layer switches.

The edge-layer switches are almost constantly active since they are connected to the end servers. Therefore, we assume that for each pod i,

$$e_i = k/2 \tag{1}$$

where e_i denotes the number of edge switches that are powered on. Since the traffic from the edge switches takes the left-most available aggregation switches first, and the total capacity of each aggregation switch is $\frac{1}{\frac{k}{2}}$ of the pod load, the total number of aggregation switches for pod i to handle traffic load $\lambda_i(1-p_1^i)$ is:

$$a_i = \lceil \frac{\lambda_i(1-p_1^i)k}{2} \rceil \tag{2}$$

For the core layer, we consider two scenarios. The first scenario is when $p_2 = 0$. In this case, all the traffic arrives in the aggregation switches is going up to the core layer. The number of core switches is determined by the maximum load from one of the k pods. Suppose pod j has the maximum load going to the core layer, $\lambda_j(1-p_1^j)$. Since each core switches can handle a fraction of $\frac{1}{\frac{k^2}{4}}$ pod load, the total number of core switches for the entire network is:

$$c = \lceil \frac{\lambda_j(1-p_1^j)k^2}{4} \rceil \tag{3}$$

When $p_2 \neq 0$, the number of core switches varies with the traffic load going to the core layer. However, we can determine the range of the number of core switches needed. If for all the pods, all the load going to the core layer is distributed on the left-most aggregation switches, the minimum number of core switches needed is calculated from the maximum core load from one of the k pods. Similarly, we suppose the maximum load going to the core layer is from pod j with the load $\lambda_j(1-p_1^j-p_2^j)$ and the number of core switches is:

$$c_{min} = \lceil \frac{\lambda_j(1-p_1^j-p_2^j)k^2}{4} \rceil \tag{4}$$

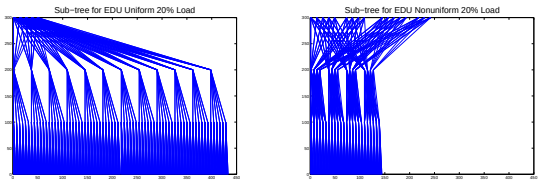

Fig. 4. Minimum fat-tree with uniform and nonuniform traffic of load 10%.

If the load going to the core layer is distributed randomly on active aggregation switches of each pod, the maximum number of active core switches is dependent on the maximum number of active aggregation switches. Suppose pod j has the most active aggregation switches and $a_j = \lceil \frac{\lambda_i(1-p_1^j)k}{2} \rceil$. Each of these aggregation switches can send traffic to $k/2$ core switches. Therefore, we can estimate that the upper bound of the core swathes is:

$$c_{max} = \lceil \frac{\lambda_j(1-p_1^j)k}{2} \rceil \times \frac{k}{2} \quad (5)$$

For example, for a $k = 12$ fat-tree with 10% overall load. If the load is uniformly distributed among all the pods, then each of the pods has 10% load. We take the example of a university data center with $p_1 = 0.3$ and $p_2 = 0.2$, the active sub-tree network after the left-most routing is demonstrated in Figure 4, left. However, if the 10% load is nonuniformly distributed. For instance, the pod 1 and pod 2 have 70% pod load and 50% pod load respectively, and all the other pods have no traffic generated. The overall traffic load is still 10%. However, the resulting sub-tree is much different (Figure 4, right).

Using the above formulation we compute the number of active switches for a $k = 12$ fat tree in each layer using formulas given in Equation (1) to (5) and show the results in Table I. We give the range of the number of core switches from Equation (4) and (5). The significant conclusions we can draw are as follows:

- Even at 70% loading, in both the uniform and non-uniform trafic cases, no more than 50% of aggregation switches are used.
- For CLD data centers, only a *third* of the aggregation switches are used because of the jb placement policies.
- At 70% loading, no more than 50% of core switches are used ahile for CLD this percentage is even smaller at 33%.

Based on the above results and the observation from various studies that show most loadings of DCNs at below 30%, we can reduce the number of aggregation switches and core switches by 50% from current fat-tree networks. One approach may be to keep each pod symmetric (for the uniform traffic model) and discard the rightmost 50% of aggregation switches from each pod. For the non-uniform traffic case, the leftmost pods would not be modified but the rightmost pods would have only a few aggregation switches as computed by eqn (2). Similarly, we can discard rightmost core switches based on eqns (4 and 5).

III. Right sizing the edge switches

A regular fat-tree uses the same size switches over the entire network. While this is a useful feature when purchasing

TABLE I. Number of active switches.

λ	EDU Uniform			EDU Nonuniform		
	edge	aggr	core	edge	aggr	core
10%	72	12	[2, 6]	12	6	[2, 18]
20%	72	12	[4, 6]	24	11	[4, 18]
30%	72	24	[6, 12]	36	16	[6, 18]
40%	72	24	[8, 12]	42	21	[8, 18]
50%	72	36	[9, 18]	54	26	[9, 18]
60%	72	36	[11,18]	66	31	[11,18]
70%	72	36	[13,18]	72	36	[13,18]
λ	EDU1 Uniform			EDU1 Nonuniform		
	edge	aggr	core	edge	aggr	core
10%	72	12	[2, 6]	12	6	[2, 18]
20%	72	12	[4, 6]	24	11	[4, 18]
30%	72	24	[6, 12]	36	16	[6, 18]
40%	72	24	[8, 12]	42	21	[8, 18]
50%	72	36	[9, 18]	54	26	[9, 18]
60%	72	36	[11,18]	66	31	[11,18]
70%	72	36	[13,18]	72	36	[13,18]
λ	PRV Uniform			PRV Nonuniform		
	edge	aggr	core	edge	aggr	core
10%	72	12	[2, 6]	12	5	[2, 18]
20%	72	12	[3, 6]	24	10	[3, 18]
30%	72	12	[4, 6]	36	16	[4, 18]
40%	72	24	[6, 12]	42	20	[6, 18]
50%	72	24	[7, 12]	54	26	[7, 18]
60%	72	24	[8, 12]	66	31	[8, 18]
70%	72	36	[9, 18]	72	36	[9, 18]
λ	CLD Uniform			CLD Nonuniform		
	edge	aggr	core	edge	aggr	core
10%	72	12	[1, 6]	12	3	[1, 12]
20%	72	12	[1, 6]	24	7	[1, 12]
30%	72	12	[1, 6]	36	11	[1, 12]
40%	72	12	[1, 6]	42	13	[1, 12]
50%	72	12	[1, 6]	54	17	[1, 12]
60%	72	12	[2, 6]	66	21	[2, 12]
70%	72	24	[2, 12]	72	24	[2, 12]

switches in bulk from OEMs, we show that it is not the best approach from an energy efficiency standpoint. Consider the benefits of increasing the degree of edge switches. The immediate impact of this is to increase p_1 and decrease p_2 thus we will need fewer aggregation switches. The second benefit comes about in energy cost of the edge switches. The energy cost of a switch can be viewed as the cost of the chassis, switching fabric, line-cards and ports [**?**]. As we show in section IV we can increase the port density of switches by adding new line-cards which has the net effect of scaling the energy cost sub-linearly with the number of ports. Thus using a single switch with twice as many ports is more energy efficient as compared to using two switches with half as many ports each.

As we analyze in Section II, the number of switches and links required in any pod for a given traffic is dependent on the traffic load λ and traffic pattern parameters of p_1 and p_2. If we increase the size of edge switches, more servers will directly connect to it and by definition, p_1 will be greater, and thus more traffic is transferred directly through the edge switches. As a result, the number of required aggregation layer switches will decrease to $a_i' = \lceil \frac{\lambda_i(1-p_1^{i'})k}{2} \rceil$. If we keep the size of the pod unchanged, then less edge switches are required with

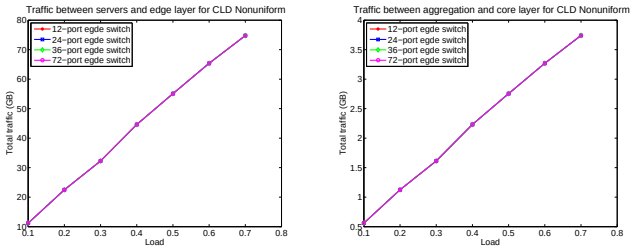

Fig. 5. Traffic between servers and edge layer is unchanged, same as traffic between agregation and core layer.

larger-sized edge switches. At the same time, since the pod size does not change, the amount of inter-pod traffic, $1-p_1-p_2$ will remain the same. Therefore, the lower bound of required core switches is the same as we calculate in Equation 4. However, since a smaller number of aggregation switches are used, the inter-pod traffic is moved to the left side of the aggregation switches, so the maximum number of required core switches will decrease. to $c'_{max} = \lceil \frac{\lambda_j(1-p_1^{j'})k}{2} \rceil \times \frac{k}{2}$. In particular, when p_1 keeps increasing and $p_2 = 0$, then all the traffic reaching aggregation layer is directed to the core switches. The number of required core switches is the minimum as $c_{min} = \lceil \frac{\lambda_j(1-p_1^j)k^2}{4} \rceil$.

We simulate a fat-tree using different sizes edge switches. The original $k = 12$ fat-tree has six 12-port edge switches in each pod. In our experiment, we use three 24-port edge switches, two 36-port edge switches, or one 72-port edge switch in each pod. The edge switches still use half of the ports connected to end servers and the other half connected to aggregation switches. For example, the 24-port edge switch connects to 12 servers and connects to six aggregation switches with *two links* between each edge switch and aggregation switch. We compare the traffic transfered in bwtween two layers of the network and show the result in Figure 5 and Figure 6. It shows that the traffic between the end hosts and the edge switches is the same for different sizes of edge switches. Also, the traffic between the aggregation layer and the core layer is also unaffected. While as shown in Figure 6, there is less traffic transferred between the edge layer and aggregation layer when the size of edge switches increase because more flows are transferred with in the same rack when the edge switch (ToR) becomes larger. The fraction of active switches is illustrated in Figure 7. We can see even the total number of switches is less when we use larger-sized edge switches.

From Figure 7, we conclude that as the size of edge switches increases, the fraction of the total number of required switches decreases. The fraction of total core switches required is shown in Figure 8. The EDU data centers have the most inter-pod traffic and the number of core switches can decrease by $5\% \sim 7\%$ when the overall load is greater than 30%. Also, when we use a 72-port edge switch in the pod, all the servers in the pod are connected to the same edge switch and the intra-pod traffic become traffic within the same subnet, thus $p_2 = 0$. The traffic arriving in the aggregation layer is all inter-pod traffic and is all moving up to the core layer. With the left-most routing, the core switches are filled in a left-to-right order. This feature makes it possible to choose the right sub-tree for a given traffic load. Figure 9 shows examples of

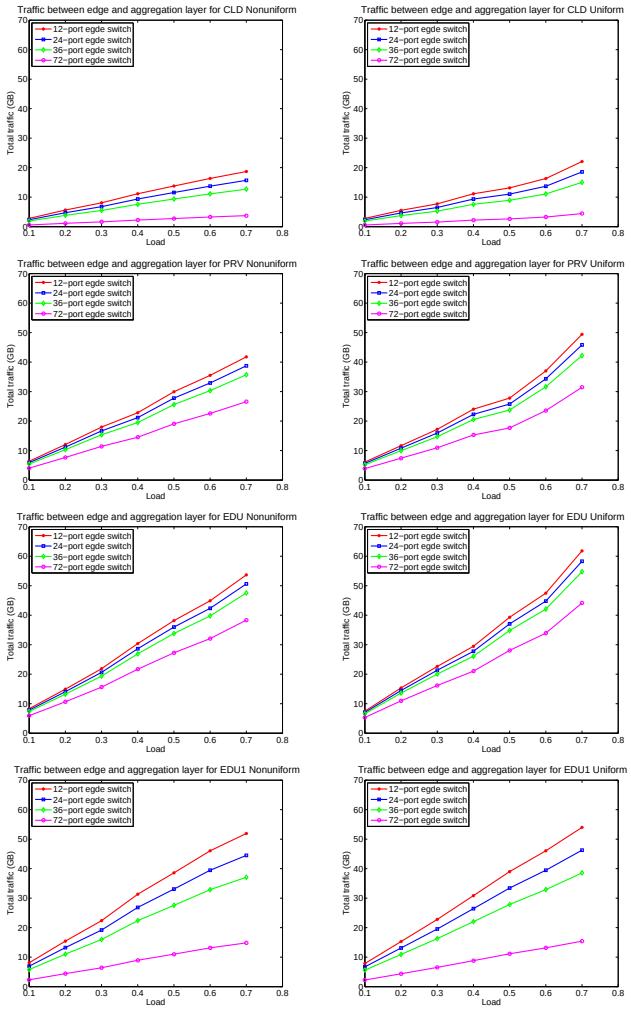

Fig. 6. Traffic between edge layer and aggregation layer is less when the size of edge swithces increases.

the resulting sub-trees when we use either 12-port or 72-port edge switches for the EDU data center in the $k = 12$ fat-tree. We can conclude that *using the highest port-density switches for the edge layer minimizes the overall number of aggregation and core layer switches required.*

IV. Energy Cost of Practical Data Center Switches

Data centers have different types and sizes. The campus data centers usually have hundreds of end hosts. A cloud data center running MapReduce type jobs or Internet-facing Web services may have tens of thousands of servers. A fat-tree data center topology uses switches of the same size to connect all servers. For example, a fat-tree network using all 12-port switches can support 432 servers, which can be deployed as a typical campus data center, while a fat-tree with 48-port switches can support 27,648 servers. To understand the power consumption of data center networks, we investigate the composition of the power cost of some widely used commodity switches for DCNs. Data center switches include fixed switches and modular switches. Modular switches are chassis based and designed and built for performance and reliability. A

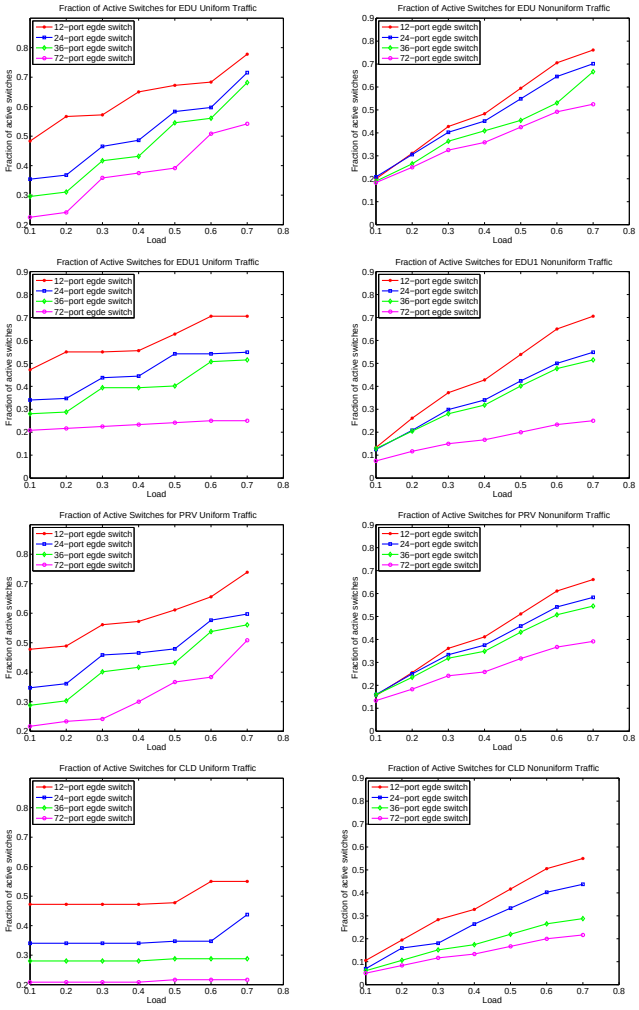

Fig. 7. Fraction of active switches with larger-sized edge switches.

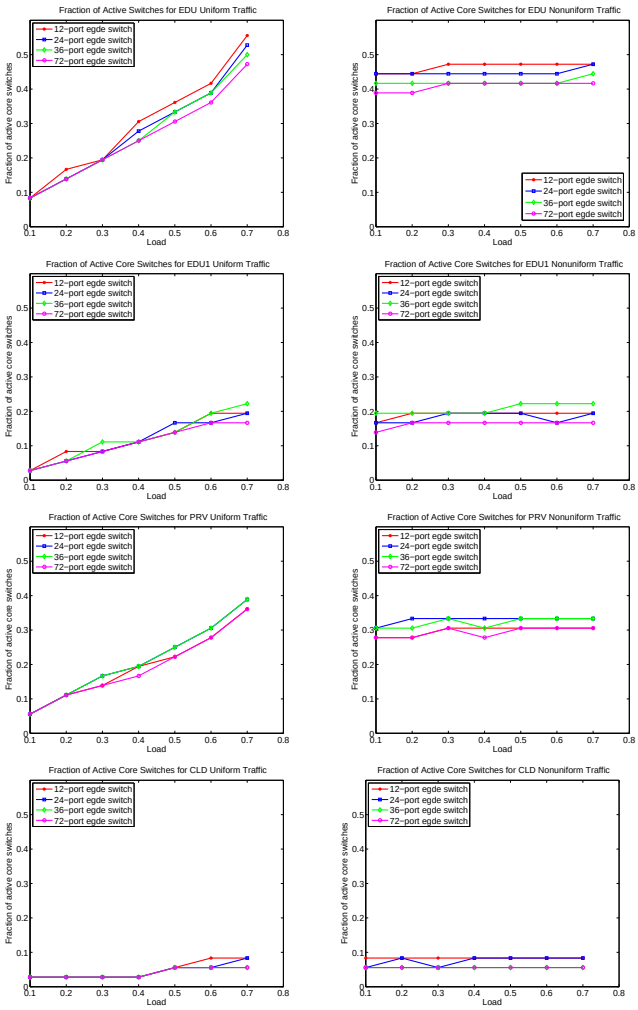

Fig. 8. Fraction of core switches with larger-size edge switches.

modular switch consists of chassis, supervisor engines and line cards. A chassis include fans, system controller, fabric modules and slots for supervisor engines and slots for line cards. The power consumption of the chassis, supervisor engines and the line-cards are considered as the static power consumption of a switch. The power consumption of port is dependent on whether the port is idle or active. We list the power consumption of typical Cisco data center modular switches using the data from Cisco Power Calculator [**?**]. Table II lists the power consumption of Cisco Catalyst 4503-E switches with 1 Gigabit line cards supporting 12, 24, 36 and 72 ports. We calculate the fixed part of power consumption of all the switches in a $k = 12$ data center with 12-port, 24-port, 36-port and 72-port edge switches in Figure 10. The results show that with the same chassis and supervisor engine, around 31% power savings can be achieved on the data center switches by insering different line cards and thus configure different size of edge switches (red line and blue line). If we choose the the switches configurations that cosume the lest power among those 12-port, 4-port, 36-port and 72-port 4503-E switches, we can still achieve around a 27% power saving for the entire network (green line).

Large cloud data centers may have tens of thousands of

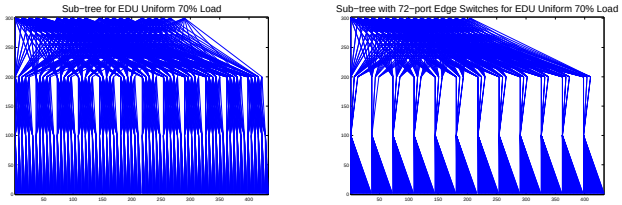

Fig. 9. Sub-tree for EDU with 12-port or 72-port edge switches, 70% load, uniform traffic.

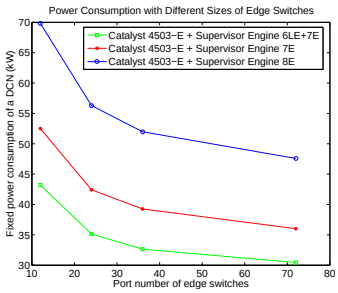

Fig. 10. Power comsumption of a $k = 12$ fat-tree DCN with different sizes of edge switches.

TABLE II. POWER CONSUMPTION OF DATA CENTER MODULAR SWITCH - CISCO CATALYST 4503-E

Chassis		Supervisor Engine		Line Card		Line Card		Port		Total Power Used
model	power	model	power	model	power	model	power	number	power	
Catalyst 4503-E	48W	7E/ 7LE	223.68W	4712-SFP-E	19.97W			12		291.65W
				4724-SFP-E	31.97W			24		303.65W
				4712-SFP-E	19.97W	4724-SFP-E	31.97W	36		323.62W
				4724-SFP-E	31.97W	4748-SFP-E	73.63W	72		377.28W
		8E	319.97W	4712-SFP-E	19.97W			12		387.94W
				4724-SFP-E	31.97W			24		399.94W
				4712-SFP-E	19.97W	4724-SFP-E	31.97W	36		419.9W
				4724-SFP-E	31.97W	4748-SFP-E	73.63W	72		473.57W

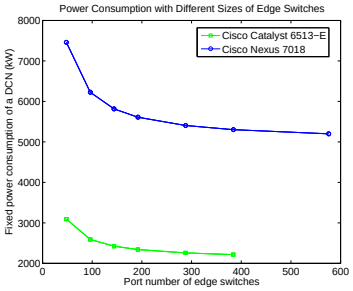

Fig. 11. Power comsumption of a $k = 48$ fat-tree DCN with different sizes of edge switches.

servers and require switches with higher port-count switches. For example, a $k = 48$ fat-tree topology uses 48-port switches and each pod has 24 edge switches. To merge the edge switches, high-port-density switches are required. For example, we need 192-port edge switches if we merge four 48-port edge switches, such as Cisco catalyst 6513-E or Cisco Nexus 7018. A Catalyst 6513-E switch chassis has 11 line-card slots and can support 528 1 GE ports in total. A Cisco Nexus 7018 switch has 16 line-card slots and can provide 768 1 GE ports. We choose the switch configuration for each switch size that consumes the least power among the switch configurations in each series, and we calculate the overall power consumption of the DCNs, which is shown in Figure 11. We find almost 30% power savings when we use 576-port Nexus 7018 and 384-port Catalyst 6513-E switches as edge switches compared with original 48-port edge switches.

V. ENERGY SAVINGS OF LARGER-SIZED EDGE SWITCHES

A network switch is composed of chassis, line-cards and ports. The chassis includes fans and line-card slots and it consumes a fixed amount of power. Low-end switches with 24 or fewer ports have built-in line-cards which can not be changed. In high-end switches, the number of ports can be extended by inserting more line-cards. Line-cards consume different amount of power depending on the number of ports and the size of memory buffer they have. A port usually consumes $1 \sim 3W$ of power when it is active and consumes as low as $0.1W$ when idle. Besides, the port capacity setting, port utilization and switch firmware version also affect the power consumption of a switch [**?**]. For simplicity, we only consider the power consumption of the chassis, line-cards and active ports in this work. We have the power model of the edge switches as:

$$P_{switch} = P_{chassis} + numCard \times P_{linecard} + numPort \times P_{port}$$

The power benchmarking study in [?] proposes a linear power model for switches used in data center racks. Derived from this model, the base chassis of a switch that has line-card slots supporting 12/24/48-port line-cards consumes around $60W$ power. Each line-card consumes about $35 \sim 40W$ power and each active port with 1Gbps capacity costs around $1 \sim 2W$. Since we need edge switches with 12 ports, 24 ports and 36 ports, we can use a switch that has three slots, with each slot supporting a 12-port line-card. Each port at the line-card is required to have 1 Gbps maximum capacity. We assume each active port consumes $1W$ power and each line-card consumes $35W$. For the 72-port edge switch, we can use a commercial switch with three 24-port line-cards. Each line-card in this switch consumes $40W$. The power model we use for estimating the switch power consumption can be formulated as:

$$P_{switch} = \begin{cases} 60 + 35 + x & \text{12-port switch} \\ 60 + 2 \times 35 + x & \text{24-port switch} \\ 60 + 3 \times 35 + x & \text{36-port switch} \\ 60 + 3 \times 40 + x & \text{72-port switch} \end{cases}$$

where x is the number of active ports of the switch. The power consumption of the sub-trees is compared in Figure 12. With the uniform traffic load, the homogeneous fat-tree can achieve more than 50% power savings through left-most routing. By replacing the 12-port edge switches with larger-sized switches, traffic flows can be further consolidated at the edge and core layer, and thus achieving a skinner sub-tree with more energy savings of the DCN.

VI. RELATED WORK

VII. CONCLUSIONS

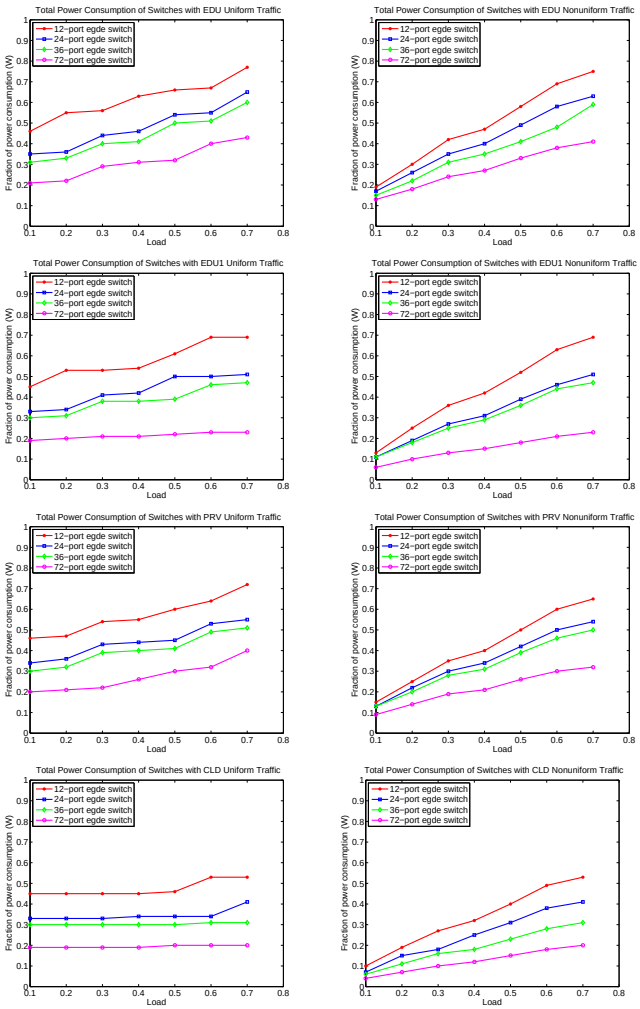

Fig. 12. Fraction of total power consumption of network switches with larger-sized edge switches for different traffic load and pattern.