Networks of Tiny Switches (NoTS): In search of network power efficiency and proportionality

Joseph Chabarek^{+}, Sujata Banerjee*, Puneet Sharma*, Jayaram Mudigonda*, Paul Barford^{+}
* HP Labs $^{+}$University of Wisconsin

ABSTRACT

The era of cloud computing is driving deployments of dense, high bandwidth, high port count network devices whose energy consumption is growing super-linearly. This trend diverges from the objective of power proportional, cloud-scale data center and enterprise networks, and implies a looming network power wall that will require re-thinking the design and configuration of networks and network systems. In this paper, we present a new architecture called *Networks of Tiny Switches* (NoTS). Our approach is based on the idea of building large networks from low port count, low power devices. We provide an overview of the hardware, software, and management required by a NoTS architecture. We also describe a case-study for NoTS in the enterprise context, and conduct a thorough evaluation of using small switches in a data center.

1. INTRODUCTION

Network energy is becoming an increasingly important and significant fraction of the energy consumed by the overall Information and Communication Technology (ICT) infrastructure. As far back as 2008, networks consumed 18 and 51 TWh in the United States and worldwide respectively [15]. There are several reasons for this increasing share of power consumed by networks.

First, the recent past has seen a considerable amount of work on–and hence great improvements in–the energy efficiency of other dominant components such as servers [4, 6]. In contrast, the pursuit of an always-connected world that depends on cloud services only resulted in ever increasing energy consumption in the network because the higher bandwidths and sophisticated packet processing capabilities required increase the power demand. The projected 20 to 30% per year growth in traffic far outpaces the expected improvements of about 10 to 20% in network equipment efficiency [14]. This widening gap can only be addressed by innovations at network architecture level.

Second, it is well established that network traffic exhibits consistent and substantial time-of-day variations [7]. However, unlike modern servers, today's network equipment cannot exploit these variations in offered load because it is largely not power proportional - *i.e.*, cannot lower the energy consumption to match the drop in traffic. This means the network can represent a large fraction of the overall data center power usage [3]. This non-proportionality becomes even more limiting given the modern trend of making data centers largely reliant on local renewable energy [11].

In this paper we present Networks of Tiny Switches (NoTS). The central idea in NoTS is to replace large high port-count switches with many tiny low port-count switches. NoTS reduces network power consumption by simultaneously improving (i) power proportionality and (ii) energy efficiency. Deploying large numbers of tiny switches enables devices to be turned off with finer granularity, thereby allowing the network as a whole to be more power proportional. Tiny switches are often more energy efficient than larger devices. For instance, in a current commercial product family the per-port power almost doubles from about 0.75 Watts in an 8-port switch to about 1.375 Watts in a 48-port switch. The challenge in NoTS is to deploy configurations that enable power consumption to be minimized while maintaining performance objectives

We describe a framework for designing NoTS configurations. The framework is based on matching the network design to a general traffic profile using low-radix devices. To demonstrate the potential impact of this approach, we conducted a simulation study in enterprise and data center settings. We investigated a large design space consisting of a wide range of power models, switch configurations, traffic patterns (synthetic as well as real-world) and topologies. We show that in an enterprise setting based on a live network, NoTS can realize up to 30% power savings without port consolidation (preserves the overall number of ports at the distribution layer) and 55% if ports are consolidated and packed into the smallest number of devices possible. In a data center setting, we show that NoTS can achieve as much as 32% of the bisection bandwidth of a full Fat-tree network while consuming a mere 17% of the Fat-tree's power.

2. EMERGING POWER CONSUMPTION TRENDS

It is evident from multiple recent studies that large network switches and routers consume large amounts of power. There is significant variability in the power consumption of these network devices depending on the device vendor, function (core, distribution, access), feature set, and the generation of the ASIC used. While the backplane fabric consumes a significant amount of power, the feature set and processing logic core switches result in higher power consumption. A feature rich core router carrying multi-terabits of traffic may consume several kilowatts of power (*e.g.*, the Juniper T4000 core router). Typical enterprise switches consume power in the range of a few hundred watts (*e.g.*, HP 5400, Cisco 4500), depending on size, features, etc.

Surveying a large number of switches utilizing the same generation of technology and with similar features, we observe that the power consumption increases faster than linear with respect to the switch forwarding capacity. Using data from [12], we show this super-linear increase in Figure 1(a), for the 80Gbps per port switches as the radix and hence the switch forwarding rate is increased. The 160Gbps per port and 320Gbps per port data is similar but is omitted for space constraints.

The following high level trends emerge that support the intuition behind the NoTS architecture. *(i)* Small switches (with lower forwarding rates) consume very low power, are more power efficient (low Watts/Gbps) and more power proportional (small no-load power, about 20-30% rather than the 60%-80% of maximum rated power for larger switches). *(ii)* Small switches come with many features that are commonly found in enterprise and data center environments. This is partly because the number of ports can be traded off in favor of advanced features on the same silicon real estate.

We provide anecdotal evidence of the above trends using some of the most power efficient, yet feature rich switches available. Figures 1(b) and (c) show the power consumption (Watts) and power efficiency (Watts/Gbps) respectively as the forwarding rate is increased for both managed and unmanaged small switches from TRENDnet [2]. For the unmanaged switch with the 16Gbps forwarding rate, the static no load power was about 17% of the max rated power of 3.5 Watts. Fixing the port rate, the per port power consumption does not increase much with increase in switch size. Thus, the fixed power costs increase much faster than linearly as a switch is scaled up to larger number of ports and results in non-proportional behavior. As another example, the NetGear platform advertises a maximum power consumption of 6W for a managed 8 port 1Gbps switch with a 16 Gbps backplane. From the same family of switches, the 16-port, 24-port and 48-port switches consume a maximum of 16.5 Watts, 21.5 Watts and 66 Watts, an increase of 2.75x, 3.58x and 11x respectively over the 8-port switch.

3. NOTS ARCHITECTURE

The core principle of the NoTS architecture is to build networks with simple, low radix switches referred to as NoTS devices. The physical architecture consists of inherently low power NoTS devices interconnected in a topology that achieves the necessary performance requirements at a fraction of the energy consumption. While there have been previous network architectures that rely on low radix switches *e.g. CLOS*, **NoTS is differentiated by offering topologies that closely match the topology construction to the network demand in order to lower overprovisioning**.We do not claim that NoTS is universally applicable, the purpose of enterprise and data center evaluation is to identify traffic patterns and topologies where small switches make sense.

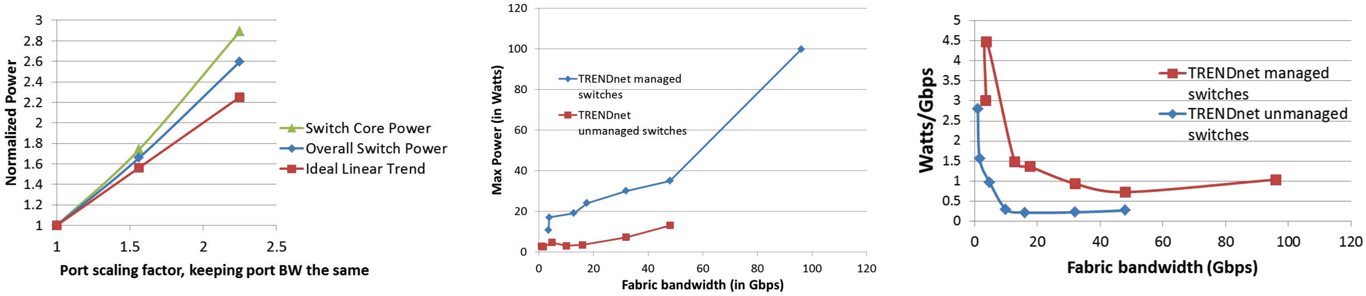
A typical NoTS device is a single ASIC, relatively low radix (of the order of 4-8 ports), and inexpensive in terms of power consumption and capital. This is in contrast to the current trend of building large, monolithic switches, where an implicit assumption is that all switch ports need to be connected via a high speed backplane irrespective of demand. The NoTS switch has a fully provisioned backplane, however the lower radix of the switch results in fewer crossing points and a simple architecture.

The smaller building blocks of the NoTS architecture allow us to match backplane capacity to traffic patterns. We discuss a number of the trade-offs inherent in using NoTS as the base switching platform. **Backplane:** NoTS devices have low powered backplanes. They have fewer communication channels, slower communication channels (as larger switches require speedup over a larger number of ports), and less arbitration overhead. **Processor:** The NoTS backplane requires less throughput because of limited port density, lower power processors can be used in a NoTS device but there will be control plane at each NoTS node. **Features:** NoTS devices must support commonly implemented enterprise and data center switch features. These common features include VLANs, spanning tree, link aggregation, PoE supply, SNMP functionality, and have the logic for some intelligent forwarding capacity, etc. There are a few feature-rich small switches in the market today[1]. **Power Supply/Fans:** Most low radix devices have simple a/c to d/c converters and typically do not have fans, which reduces the overall power consumption. As NoTS devices are aggregated into larger topologies shared power supplies and fans may become necessary.

To evaluate NoTS we need to model the behavior and power consumption of a set of candidate small switches. Currently these are based on off-the-shelf desktop switches, though future iterations of NoTS could use custom hardware. We introduce 4 potential conservative NoTS device power models from a survey that we conducted, along with a traditional high radix switch power model in table 1. The standard model (labeled Model C) was used in a prior study [16] and has the smallest chassis power of the models discussed in that paper. Model C is a chassis based switch that can accommodate up to 6 line cards; a common configuration is to have 5 of the linecards to be 1Gbps 24-port cards

---

[1] For example, the NetGear ProSafe 8-port Gigabit Smart Switch is a managed switch with nearly all of the feature set common in enterprises today. The maximum power consumption for this switch is 6 Watts [18].



(a) Normalized power vs. port scaling

(b) Power vs. Forwarding Rate

(c) Efficiency vs. Forwarding Rate

Figure 1: Power consumption trends in (a) future switches and (b),(c) current switches

Model	Line-cards	Chassis Power(W)	1Gbps port(W)	Host/Uplink ports
ModelC	Y	55	0.9	120/4
NoTS1	N	5	0.65	8/2
NoTS2	N	5	0.9	6/2
NoTS3	N	5	0.9	6/1
NoTS4	N	0.7	0.3	8/8

Table 1: Device power models measured for Model C and inferred for NoTS 1-4

and the sixth one can host 4 10Gbps uplink ports.

We conservatively assume a base cost of 5W for the NoTS1-3 device chassis and assign the port cost for each of the available ports to be equal to one of the measured port costs from 3 models from [16]. This puts the total power consumption of the NoTS models to be about double that of the equivalent NetGear small switches, but lower than the linearly scaled down versions of Model C. We vary the subscription levels (host to uplink bandwidth) of NoTS 1-3 to be 4:1, 3:1, and 6:1 respectively as a result of selecting port constraints and assigning the available 1Gbps ports for hosts or uplinks. We only evaluate NoTS4 in the context of data center topologies, where the uplink to downlink relationship varies with the topology. This model was inspired from measurements of a low radix desktop switch.

4. APPLYING NOTS TO ENTERPRISE NETWORKS

Enterprise networks are primarily designed in a three tiered *core, distribution, and access* topology. The access ports for the various compute devices are connected to the closest switching closet which in turn connects to a distribution layer which aggregates links and passes traffic to the core. The *distribution, and core* switches are interconnected amongst themselves to provide complete reachability as well as ingress/egress to the Internet and data centers hosting various enterprise services.

Most enterprise access ports [16] have very low utilizations with little traffic between end hosts (*e.g.*, desktops of employees). Inactivity is caused by ports in public spaces such as conference rooms as daily diurnal work patterns. Traffic primarily *(greater than 95% in our measurements)* flows *north-south*, from access to core where servers providing various enterprise services and peering points are located and back to the hosts in the access. We focus on the access portion of an enterprise network as a test case for NoTS. Our initial simple enterprise transformation is as follows: **replace large access switches with a larger number of NoTS switches that connect hosts directly to the distribution layer.**

Evaluation: We concentrate on a small slice of a production enterprise network with five switches (that are well represented by Model C) that was provisioned for 600 end-hosts/devices. Our measurements indicate that out of these 600 ports only 355 had been allocated and only 218 were being actively used by users on one floor of a building. These *access* switches aggregate traffic from 1Gbps links connected to host machines to a 10Gbps uplink that connects to a *distribution* switch that provides connectivity to the rest of the network via 10Gbps fiber. The 5 access level switches are directly connected with 1Gbps links.

In all replacement schemes we remove any previous direct connectivity between access layer switches and connect the *access* layer directly to the *distribution* layer. This follows from our assumption that inter-access communication is low. In the worst case, *access*-to-*access* packets would incur an additional hop at the *distrbution* layer. The devices in the NoTS topology use 1Gbps links to transit from the *access* to the *distribution* as opposed to the previous 10Gbps link, but aggregate traffic from fewer links leads to similar ratio of “uplink” to “downlink” capacity. If additional line cards or chassis are required at the distribution layer to accommodate additional 1Gbps links, we add those costs to the power cost of the NoTS topology.

We evaluated 3 different scenarios. In the "-allocated" scenarios, we ensure that there are enough chassis and respective uplink capability to support 600 ports while we only count host port costs for the 218 host ports that are administratively and operationally up in the original deployment. For the "-packing" scenarios, we wire the network such that all of the 355 host ports that are administratively registered as having an end host are wired contiguously to the fewest switches as possible. The access devices that do not have ports listed as allocated are powered off along with the associated uplink capability. Additionally, for comparison we add a deployment-realized scenario where linecards are powered down if they do not have any active ports. Figure 2 summarizes the power required for each of the configurations where the NoTS configurations can provide up to 30% reduction in power.

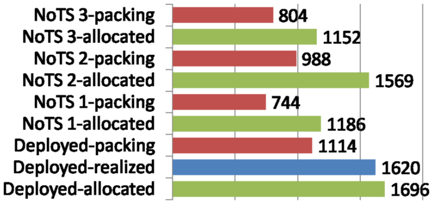


Figure 2: Scenario based power consumption

5. APPLYING NOTS TO DATA CENTER NETWORKS

The NoTS enterprise transformation will not be acceptable in data centers where significant traffic flows "east-west." Hence a different network transformation is required for the NoTS paradigm to be useful in the data center context. In this section we present and evaluate a range of NoTS interconnection topology options for connecting low radix devices.

There are existing data center topologies such as BCube [8], the 3d torus, and random graphs which can be parameterized to support low radix switches. We analyze a variety of topologies describing their static properties such as the number of hosts a topology can scale to for a given switch radix, bisection bandwidth, network interface cards required, etc. Next, we provide an initial simulation of specific topologies. We assume where appropriate that individual network elements (*i.e.*, switches and links) can be powered up or down as needed.

Our simulation tool uses the graph library `networkx` [1] to generate graphs and provides a mechanism to analyze traffic from empirical traffic traces (`pcap` format) or synthetically generated random traffic matrices. For the initial results below, we evaluate the topologies of interest with a synthetically generated traffic matrix so that we can easily vary the size of the network under test. We generate this traffic matrix with a parametrized percentage of hosts from the graph to be powered up and randomly select the source and destination of flows. Each randomized test is conducted 50 times with the average reported. The simulator finds the parameters that construct the most compact network available for each topology type given the number of requested hosts.

5.1 Static Topology Analysis

There are a number of data center network topologies that have been proposed recently, targeting high bisection bandwidths. In our study, we pick a set of important topologies - Fat-tree [19], BCube [8], a three-dimensional torus network, and random regularized graphs (RRG) [20]. We summarize the parameters and properties of the Fat-tree, BCube, torus, and random regularized graphs in Table 2, where we connect N end hosts. A torus has d dimensions and we represent the distance between a source and destination in dimension i with (δ_i). A BCube is parametrized with k, and n where k is the number of levels and n is the size of each BCube cell. A Fat-tree is parametrized with s, p where s is the number of stages in the tree and p is the number of ports for each switch. Random topologies are constructed by randomly connecting top of rack (ToR) switches where k is the radix of each ToR switch, r is the number of unique ToR switches each switch is connected to and T is the number of racks.

The Fat-tree and BCube topologies are constructed such that the bisection bandwidth is high regardless of the number of hosts, while the torus' changes with the number of servers attached to the network. Fat-trees are full bisection bandwidth by construction. BCubes have an aggregate throughput of: (n/(n-1))*(N-1) where n is the switch port radix and N is the number of servers. The bisection bandwidth of the torus is a function of the number of nodes where if the length of each side is $len = N^{(1/d)}$ then an equal partition is bisected by $k^{(d-1)}$ links. The random topologies support at least as many servers per switch on the same hardware as the Fat-tree.

5.2 Traffic Driven Tradeoff Analysis

Using low radix switches in NoTS may increase the average path length and also impact the bisection bandwidth. However, in many realistic scenarios, smaller networks with less than full bisection bandwidth will suffice and the radix can be lowered to match the requirement - our goal is to identify the power-performance trade-offs for a set of NoTS topologies.

For the purposes of this analysis we consider a data center with a maximum size of 1800 hosts and generate the test topologies with the appropriate parameters (listed below). *BCube:* 8-port switches ($n = 8$) arranged in 3 levels ($k = 3$). *Fat-tree:* A 3-tier Fat-tree [19] requiring switches with $p = 20$ ports. *Fat-tree2:* Same as Fat-tree except with a commonly sold switch radix, 24 rather than 20. *RRG9:* In RRG9 3 ports connect to servers and 6 ports connect to other ToR switches. *RRG42:* In RRG42 each ToR switch connected to 12 servers and 30 other ToR switches.

With the above parameters, the torus, BCube and RRG9 topologies satisfy the NoTS criteria of using low radix switches, while the remaining three topologies - Fat-tree, Fat-tree2 and RRG42 are non-NoTS. Next we compute shortest paths between hosts, step through the traffic matrices, and compute the power consumption for the network given the devices powered to support the traffic ma-

Table 2: Data Center Topologies Studied

Topology	Params	Hosts	#Switches	#Links	Radix	disjoint paths	sp's
Torus	d	Inf	N	dN	2d	$max(d)min = 1$	$(\sum_{i=1}^{i=d} \delta_i)!/\prod_{i=1}^{i=d}(\delta_i)!$
BCube	k,n	$n^{(k+1)}$	$(k+1)(n^k)$ n:N hosts	$Nlog_n(N)$	$n$and k+1	k+1	k+1
Fat-tree	s,p	$(p^3)/4$	$(5p^2)/4$	$5p^3/4$	p	1	$p^2/4$
RRG	k,r,T	T(k-r)	T	Tk	k	*	*

Topology	Normalized Bisection Bandwidth
Torus	0.19
BCube	1.14 [9]
Fat-tree	1.0
RRG9	0.32
RRG42	0.87

Table 3: Bisection bandwidths for data center networks generated for 1800 servers normalized by aggregate bandwidth of servers in one set of bisection.

trix for the particular time step.

Power Consumption/Efficiency: We first examine the power consumption of the NoTS and non-NoTS topologies with 1800 hosts. In Figure 3a, we plot the power consumption against the fraction of active hosts for two different power models. The non-NoTS switches are given the advantage of using Model C with the low chassis cost from Table 1 and then scaled down based on the number of ports (35W for the 20 and 24 port switches; 55W for the 42 port switches). The port cost for non-NoTS is matched with the port cost of the NoTS models. In Figures 3a and 3b we use the NoTS1 and NoTS4 power models with per port power of 0.65W and 0.3W respectively. Note that a perfectly power proportional network would have a linear consumption profile starting at the origin in figures 3a and 3b and ending at each topologies maximum value. In both cases, the NoTS configurations of RRG9, torus, and BCube use less power than the non-NoTS configurations. Among the NoTS topologies, RRG42 performs better than the Fat-trees.

Bisection Bandwidth: In Table 3 we report the bisection bandwidth for the same network topologies. Results are normalized by the aggregate bandwidth of servers in one of the two disjoint sets. The bisection bandwidth results for the regularized random graphs are computed from Bollobas' lower bound [5]. RRG9 and torus sacrifice bisection bandwidth for better power efficiency and proportionality. BCube emerges as the ideal NoTS choice for low power, power proportionality and high bisection bandwidth.

Average Path Length: In Figure 3c we show the average path length for the topology classes for the random traffic matrices discussed before. Notice that the low power NoTS topologies incur a penalty in the average path length for the random flows. The average increase of about 1.5 hops between the Fat-tree and the NoTS cases is likely not significant for most applications. Also, note that the RRG42 outperforms the Fat-tree topologies while the RRG9 has a longer average path length than the Fat-trees.

6. RELATED WORK

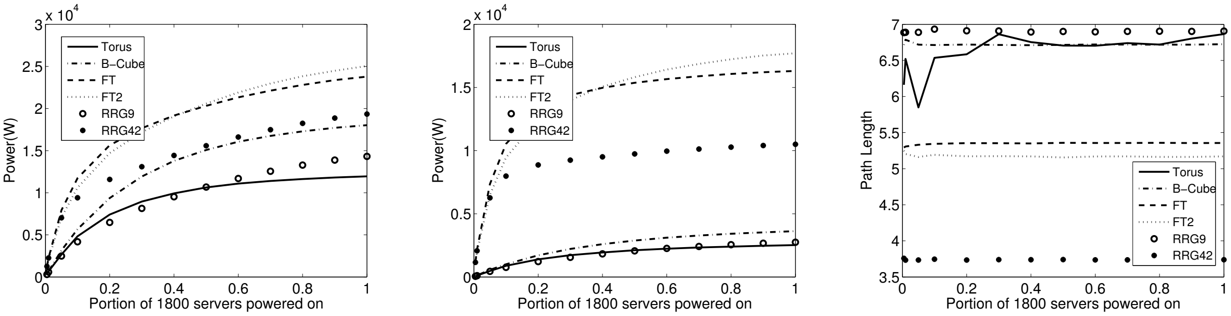
Recent work to build efficient, proportional networks includes the characterization of network devices as well as constructing power-aware topologies. In [17], Mahadevan *et al.* modeled the power profiles of current generation network hardware and highlighted the poor power proportionality of network devices. Hardware enhancements at the component level have been proposed in standards bodies to enable ports to automatically adjust their rate or go into sleep states [13]. In [10] and [7] authors optimize network deployments to minimize power. Finally, [16] discusses enterprise network operational practices that lead to power inefficiencies. Many of these techniques can be leveraged in NoTS.

The BCube data center network [8] used mini-switches connected in the BCube topology and showed that it was possible to build high performance networks with small switches offloading significant network functions to the servers. In [3], authors consider an energy proportional data center network. Abts et. al. propose using a flattened butterfly with link-level rate modulation of Infiniband links. The NoTS approach explores multiple topologies built from low radix commodity hardware.

7. DISCUSSION

The challenges of meeting current and future network performance and feature demands, while simultaneously moving toward energy-proportional ICT infrastructures are significant. The current generation of networking components used in large ICT infrastructures constitute a *growing* percentage of overall energy budgets and continue to lack capabilities for proportional power use that are critical to the objectives of green ICT.

In this paper, we present NoTS, a novel network design methodology which takes an important step towards efficient, power proportional networks with low power low radix switches deployed to match traffic patterns. We seek to demonstrate the feasibility and potential impact of NoTS over a broad design space to highlight tradeoffs and provide insights on the methodology. In an enterprise network we show the potential savings for a 600-port network that has 200 active ports is between 12% (NoTS 2) and 36% (NoTS 1 and NoTS 3). In the data center, we show that the energy improvements from the NoTS principles do not have to come at the cost of decreased bisection bandwidth - for example, using the BCube topology. The average increase in



(a) NoTS1 scaling

(b) NoTS4 scaling

(c) Path Length

Figure 3: Power consumption for network with varying number of hosts powered using using the NoTS1 in Figure 3a with a per port cost of 0.65W, NoTS4 in Figure 3b with a per port cost of 0.3W. Figure 3c shows simulated average path length

path length of about 1.5 hops will not impact many current data center applications.

There are several practical issues for NoTS which are planned as future work including using software defined networking (SDN) principles to manage traffic routes and investigating further hardware platforms that support the NoTS ideal. Additionally, current NoTS devices have form factors that would make them cumbersome to deploy in large numbers. This practical matter could be addressed in a fashion similar to blade servers, which are popular in data centers.

Acknowledgements

This work was initiated at HP Labs and supported in part by an HP Labs internship. This work was supported in part by NSF grants CNS-0831427, CNS-0905186, ARL/ARO grant W911NF1110227 and the DHS PREDICT Project. Any opinions, findings, conclusions or other recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF, ARO or DHS.

8. REFERENCES

[1] NetworkX. `http://networkx.lanl.gov/`.

[2] TRENDnet switch. `http://www.trendnet.com/`.

[3] D. Abts, M. Marty, P. Wells, P. Klausler, and H. Liu. Energy Proportional Datacenter Networks. In *Proc. Intl. Symp. on Computer Architecture (ISCA)*, 2010.

[4] L. A. Barroso and U. Holzle. The case for energy-proportional computing. In *IEEE Computer*, December 2007.

[5] B. Bollobas. The isoperimetric number of random regular graphs. Eur. J. Comb, 1988.

[6] A. K. et. al. Napsac: Design and implementation of a power proportional web cluster. In *Proc. of ACM SIGCOMM Workshop on Green Networking*, 2010.

[7] B. H. et. al. ElasticTree: Saving Energy in data Center Networks. In *NSDI*, April 2010.

[8] C. G. et. al. BCube: A High Performance, Server-centric Network Architecture for Modular Data Centers. In *ACM SIGCOMM*, August 2009.

[9] D. L. et. al. Hyper-bcube: A scalable data center network. In *ICC*, 2012.

[10] J. C. et. al. Power awareness in network design and routing. In *Proc. Of INFOCOM*, April 2008.

[11] M. A. et. al. Towards the Design and Operation of Net-Zero Energy Data Centers. In *Proc. of the 13th Intersociety Conf. on Thermal and Thermomechanical Phenomena in Electronic Systems*, 2012.

[12] N. B. et. al. The role of optics in future high radix switch design. In *Proc. of ISCA*, 2011.

[13] IEEE 802.3 energy efficient ethernet task force (formerly study group). `http://grouper.ieee.org/groups/802/3/az/index.html`.

[14] T. E. Klein. Next-Generation Energy Efficient Networks: Overview of the GreenTouch Consortium. 2011 KAIST Institute International Workshop on Smart Media, September 2011.

[15] S. Lanzisera, B. Nordman, and R. Brown. Data Network Equipment Energy Use and Savings Potential in Buildings. ACEEE Summer Study on Energy Efficiency in Buildings, 2010.

[16] P. Mahadevan, S. Banerjee, and P. Sharma. Energy Proportionality of an Enterprise Network. In *Proc. First ACM SIGCOMM Workshop on Green Networking*, Aug 2010.

[17] P. Mahadevan, P. Sharma, S. Banerjee, and P. Ranganathan. A Power Benchmarking Framework for Network Devices. In *Proc. IFIP Networking*, May 2009.

[18] Netgear. ProSafeÂő 8-port Gigabit Smart Switch Data Sheet. `http://www.netgear.com/images/gs108tv2_ds_10dec0918-5336.pdf`.

[19] R. N. Mysore et al. PortLand: A Scalable Fault-Tolerant Layer 2 Data Center Network Fabric. In *ACM SIGCOMM*, August 2009.

[20] A. Singla, C.-Y. Hong, L. Popa, and P. B. Godfrey. Jellyfish: Networking data centers randomly. In *9th USENIX NSDI*, April 2012.