2013 IEEE 12th International Symposium on Network Computing and Applications

Incorporating Rate Adaptation into Green Networking for Future Data Centers

Lin Wang*§, Fa Zhang*, Chenying Hou*§, Jordi Arjona Aroca†‡ and Zhiyong Liu*§
*Institute of Computing Technology, Chinese Academy of Sciences, China
†Institute IMDEA Networks, Spain
‡University Carlos III of Madrid, Spain
§University of Chinese Academy of Sciences, China
¶State Key Laboratory of Computer Architecture, ICT, CAS, China

***Abstract*—Despite some proposals for energy-efficient topologies, most of the studies for saving energy in data center networks are focused on traffic engineering, i.e., consolidating flows and switching off unnecessary network devices. The major weakness of this approach is network oscillation brought by the frequent change of network topology when traffic fluctuates very fast. In this paper, we propose to incorporate rate adaptation into green data center networks. With rate adaptive network devices, we aim at approaching network-wide energy proportionality by routing optimization. We formalize the problem with an integer program and propose an efficient approximation algorithm – TSRR, solving the problem quickly while guaranteeing a constant performance ratio. Extensive range of simulations confirm that more than 40% of the energy can be saved while introducing very slight stretch on network delay.**

I. Introduction

The popularization of cloud computing drives the extensive deployment of data centers in order to provide fundamental computing infrastructures. As a result, the energy consumed in these massive data centers is tremendous and emerges as a big concern. To achieve reliability and scalability, most data centers are designed with a significant level of resource redundancy and over-provisioning, leading to a mismatch between common workloads and power consumption. For example, the CPU utilization of a typical Google cluster is within the $10\% - 50\%$ range during a large portion of the time [1]. Therefore, efficient energy saving strategies are urgently demanding in data centers.

Besides the enormous number of servers, the energy consumed by the large amount of network devices used for connecting servers also emerges as a major concern. It has been reported in [2] that the total power consumed by network elements in data centers in 2006 in the U.S. was about 3 billion kWh and is still increasing rapidly. In general, up to 45% of the total energy used in a data center is consumed by the servers (CPUs, memories and storage systems etc.) [3]. However, with some advanced energy efficient strategies being involved, the power consumed by the network will also become a first-order issue [4]. Regardless of whether servers are energy proportional, making the network energy proportional can bring considerable savings, implying tremendous economic benefit.

Traditional data center networks are mostly structured with 2N tree topology [5]. In a 2N tree the effective bisection bandwidth can be easily cut down by a small number of failures. Alternatives such as FatTree ([6], [7]), VL2 [8], BCube [9] and DCell [10] provide much richer connectivity and can handle failures more gracefully. However, the static provisioning irrespective to real workloads forces most of the designs into the quandary that high connectivity comes with high power consumption as current network devices are unlikely energy proportional. It has been verified in [11] that 60% of the time, the average traffic stays quite small and the time in which traffic peaks is less than 5%. As a result, the power consumption of the network should be proportional to the workload to conserve energy.

The topic of greening the data center network has been widely explored. Most of the work can be categorized into two groups in general. The most straightforward way is designing energy-efficient topologies which can provide similar connectivity while using less network devices [4], [12]. However, the potential of this approach is limited since we have to guarantee sufficient bandwidth for traffic bursts. Another option consists in reducing the amount of active devices in current networks. This is generally accomplished by consolidating traffic flows and turning off unnecessary devices (e.g. [13], [14], [15], [16]). The key observation behind this line is the connectivity redundancy and the traffic load variation in current data center networks. However, when we switch off devices, the network topology will be changed. Since this topology transformation cannot be completed in a short time, the network may suffer from oscillation. Consequently, maintaining the quality of service in the network will become tricky.

In this paper, we propose to incorporate rate adaptation into data center networks to achieve energy conservation. To the best of our knowledge, this has not been deeply explored before. Rate adaptation was proposed by Nordman and Christensen [17] and has been widely studied. The main idea of rate adaptation is to approach energy proportionality by varying the link rate adaptively to meet its carried load. Since being proposed, rate adaptation starts to be supported by some

This research was partially supported by the National Natural Science Foundation of China grant 61020106002, 61161160566 and 61202059, and the Comunidad de Madrid grant S2009TIC-1692, Spanish MICINN grant TEC2011-29688-C02-01.

978-0-7695-5043-5/13 $26.00 © 2013 IEEE
DOI 10.1109/NCA.2013.24

production devices, such as InfiniBand [18]. Although there are still some limitations in applying rate adaptation directly, we believe that future network devices will provide a wide set of operating rates in order to keep up with the green computing trend. In this sense, this work also reveals the potential of saving energy by using rate adaptation in future data center networks.

We also emphasize the opinion that a good energy saving solution comprises not only single-device energy proportionality, but also needs network-wide optimization. The basic idea is to globally optimize the scheduling and routing of flows and dynamically adjust the rates of network devices according to their loads. Based on rate adaptation, we aim at exploring efficient routing algorithms for improving the energy efficiency in data center networks. Specifically, we model this rate-adaptive energy-efficient routing problem and provide a constant approximation algorithm to solve it. This is the most significant difference compared with the most relevant work [19] which provides an energy-efficient traffic engineering solution for general networks by utilizing rate adaptation heuristically.

Our main contributions are highlighted as follows: 1) We provide a formal model to describe the energy-efficient routing problem using rate adaptation in data center networks. 2) We propose a two-step relaxation and rounding algorithm, solving the energy saving problem approximately within a constant ratio to the optimum. 3) We carry out extensive simulations to verify the performance of our algorithm, where we show that up to 40% energy reduction can be achieved by incorporating rate adaptation into greening data center networks.

The remainder of this paper is organized as follows. Section II models the routing optimization problem. Section III presents our main algorithm. Section IV verifies the performance of the proposed algorithm by simulations. Section V concludes the paper.

II. Problem Description

Consider a data center network $G = (V, E)$, where V is the set of nodes and E is the collection of edges. Here nodes represent the chassises of network devices, while edges represent the network links and the corresponding line cards. For the sake of simplification, we assume identical network devices, which is applicable for FatTree, BCube etc. All the edges in E are assumed to have the capability of adapting their operating rates according to their carried traffic loads. It is quite reasonable that manufactures will provide many different operating rates in future network devices, due to the trend of being green. Assume we have given m discrete rates $R = (r_1, r_2, ..., r_m)$ for all the edges in E, where $r_i < r_{i+1}$ for $i \in [1, m-1]$. Each rate $r_i \in R$ $(1 \leq i \leq m)$ has a cost defined by function $f(r_i)$ which is uniform in the network, representing the power consumed by edge e when working at rate r_i.

We have a set of traffic demands $D = (d_1, d_2, ..., d_k)$ to be routed on G. For the i-th demand, a d_i units of bandwidth are requested to be provisioned from a source node s_i to a destination node t_i. In order to avoid packet reordering, we assume that all the demands will be routed in an unsplittable way, i.e., follow a single path (one TCP flow). The total cost of the network is defined as the summation of the costs of all the edges, which can be formalized as

$$C = \sum_{e \in E} C_e = \sum_{e \in E} f(z_e) \tag{1}$$

where z_e is the operating rate chosen for edge e. This cost represents the total power consumption of the whole network. The rate-adaptive energy-efficient routing problem now can be formulated with the following integer program.

$$\begin{aligned} (P_1) \quad & \min \quad C = \textstyle\sum_{e \in E} f(z_e) \\ \text{subject to} \quad & \\ & x_e = \textstyle\sum_i \varphi_{i,e} \cdot |d_i| && \forall e \\ & x_e \leq z_e && \forall e \\ & z_e \in \{r_1, r_2, ..., r_m\} && \forall e \\ & \varphi_{i,e} \in \{0, 1\} && \forall i, e \\ & \varphi_{i,e} : \text{ flow conservation} \end{aligned}$$

where $\varphi_{i,e}$ is an indicator to show whether the i-th demand will be routed through link e which follows the flow conservation. Flow conservation means that for the i-th demand, the source s_i generates a flow of d_i units and the sink absorbs it. For any other vertices, the ingress and egress flows are the same. x_e and z_e are the total load and the chosen operating rate for edge e respectively. For any rate $z_e \in \{r_1, ..., r_m\}$, we have $f(x_e) = f(z_e)$ if we choose z_e such that $x_e \leq z_e$. In other words, if we have $r_{j-1} < x_e \leq r_j$, then $f(x_e) = f(r_j)$, which inevitably results in the discreteness of the power cost function. In fact, $f(x)$ is a non-decreasing step function of the amount of the total traffic going through an edge.

III. Approximation

Not surprisingly, solving P_1 is NP-hard, which can be easily proved using a reduction from the edge-disjoint path problem. Since the optimum cannot be found in polynomial time, we seek good approximations. It can be observed from P_1 that the complexity of this problem mainly arises from the non-convexity of the objective function C. The observation we can make here is that C will become convex if we transform $f(\cdot)$ into a convex function. This will make the problem easier to solve. Suppose we have such a transformation, obtaining a convex function $g(x)$ instead of $f(x)$. The power cost now becomes $C = \sum_e g(x_e)$, and P_1 turns to be an integer convex program. The usual way to solve it is by randomized rounding, which 1) relaxes the binary variables, 2) solves the relaxed program and 3) rounds the fractional solution to a feasible one.

Following the above method, we devise an approximation algorithm called two-step relaxation and rounding (TSRR). In this algorithm, we first carry out a two-step relaxation in order to make the problem solvable. After solving the relaxed problem, we transport the solution into the feasible region by applying a two-step rounding operation. It can be proved that using this algorithm, a constant ratio of approximation can be

achieved. The details can be found in [20] and here we just provide a sketch of the algorithm.

(RLX1) Convert the step function $f(\cdot)$ into a convex continuous function $g(\cdot)$ by using a special kind of interpolation. From the superadditive property of power consumption, we assume $g(x)$ can be formulated as μx^{β} where μ and β are parameters and $\beta \in (1, 3]$ [21]. We then obtain the values for the parameters that minimize the interpolation error. Denote the new program after this step as P_1'.

(RLX2) Remove the binary constraint $y_{i,e} \in \{0, 1\}$ and solve P_1' by convex programming, obtaining the fractional solution, denoted by $y^*_{i,e}$.

(RD1) Use the Raghavan-Thompson randomized rounding algorithm to obtain a single routing path for each traffic flow. This process basically is to randomly select a path from the candidate paths using a weight (obtained from $y^*_{i,e}$) as the probability of choosing it. Denote the obtained solution by $\hat{y}_{i,e}$.

(RD2) Choose an appropriate rate for each link. Denote $\hat{x}_e = \sum_{i \in [1,k]} \hat{y}_{i,e}$. Then, the rate s_e for link $e \in E$ is chosen following $s_e = \min\{R_i | (i \in [1, m]) \wedge (\hat{x}_e \leq R_i)\}$. This solution is feasible for the original problem.

IV. Evaluation

We carry out comprehensive simulations to evaluate the energy saving performance of our proposed algorithm in this section. Both synthetic and real network traces are used.

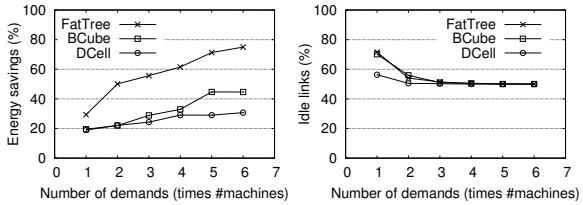
A. Experimental Settings

We use a synthetic power function $f(\cdot)$ which we believe is similar to the fashion power being consumed by future rate-adaptive commodity network devices. The maximum operating rate of network devices is set to be 60 units. We also believe that the precise form of $f(\cdot)$ can be easily obtained from vendors in the future. The alternative convex function $g(\cdot)$ is obtained by applying the interpolation method we proposed in section III.

The efficiency of energy saving of our algorithm is verified with both synthetic traffic conditions and real network traces. In both cases, the numbers of physical machines are chosen as 128 which is determined by the real network traces. We conduct our simulations using three popular kinds of data center network topology: FatTree, BCube and DCell. The traffic condition for the synthetic testing is generated using the same setting as described before, while for the real testing it is extracted from a university data center [22]. The traffic patterns of the real network traces is worth 15 minutes long and the endpoints in the traces are mapped to the three topologies we use.

B. Potential Energy Savings

We evaluate now the energy saving efficiency of the proposed routing algorithm when being applied in real data centers. To be fair, we assume that all the network devices are capable of rate adaptation no matter what routing algorithm is used. We compare our algorithm with the shortest path based routing which is a common practice in most networks. We focus on two aspects of interest - the energy savings and the ratio of idle links (not used for routing any demand). The energy saving ratio is calculated as the ratio between the energy consumption of our algorithm and the shortest path based algorithm, while the ratio of idle links is calculated using the number of idle links divided by the total number of links.



(a) (b)

Fig. 1. Simulation results for testing the energy saving efficiency using synthetic network conditions based on three different topologies: FatTree, BCube and DCell.

1) Synthetic Traffic: The simulation results are shown in Figure 1. It can be observed that TSRR can achieve up to 60% energy savings in FatTree and more than 30% in both BCube and DCell. Moreover, this savings tends to be stable with the increase of the number of demands, confirming that a global energy-efficient routing optimization can help reduce a substantial amount of power consumption.

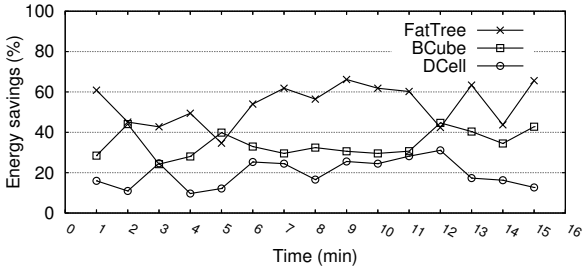
We also observe that the link utilization in a data center stays quite low during most of the time. As shown in Fig. 1(b), with a reasonable amount of demands, only 50% of the links are needed. This reveals that for normal traffic patterns, all the traffic can be carried by only half of the total links, as a consequence of the high link redundancy in the network.

2) Real Traces: We repeat the above experiments using real traces from a university data center. The numerical results are shown in Fig. 2. Regarding the energy savings, we have similar findings as before. The average energy saving is more than 50% in FatTree, while this value is 30% and 20% in BCube and DCell respectively. At the same time, the ratios of idle links presented in Fig. 2(b) also proves that we can use about half of the links to carry all the traffic on a real data center network.

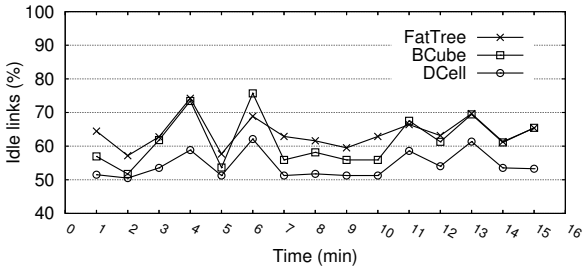
Recall that in our model we do not have any assumption on powering down network devices. Meanwhile, the second rounding phase in TSRR provides some extra capacity for each edge, leading to better stability in the network while compared with power-down based approaches. Nevertheless, we believe that our proposed method can also be integrated with power-down based approaches for the reason that the ratio of idle links in data center networks is usually quite high as indicated above. As a result, further energy can be saved by switching off these unused links.

C. Delay Stretch

It is quite possible that the proposed energy-efficient routing will use more hops to route a demand than the shortest path

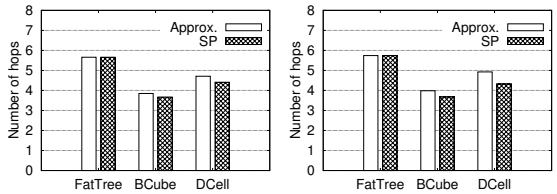


(a)



(b)

Fig. 2. Simulation results for testing the energy saving efficiency using real network traces based on three different topologies: FatTree, BCube and DCell.



(a) (b)

Fig. 3. The average number of hops for routing each demand under a) the synthetic traffic, b) the real network traces.

based routing, bringing about lager network delay, thus the degradation of the network quality of services. In order to justify how bad it can be, we compare the average hops for routing each demand with the two routing algorithms. This comparison is quite sensible since the maximum number of hops is bounded by the topology diameter in most data center networks. The results are illustrated in Fig. 3(b). We can notice that the overhead brought by the energy-efficient routing is always within one hop, being acceptable then. Furthermore, there is no delay overhead in FatTree topology due to the special hierarchical property of FatTree.

V. Conclusions

In this paper, we showed the potential of saving energy by introducing rate adaptation in data center networks. Compared with power-down based approaches, rate adaptation is advantageous because of the better stability when being applied in networks. It has been shown that network-global optimization is necessary in order to achieve energy proportionality for the whole network with energy-proportional network devices. However, it is also known that this network-global optimization problem is usually hard to solve. To this end, we devise an approximation algorithm basing on a two-step relaxation and rounding process. The proposed algorithm performs very well by obtaining nearly optimal solutions in practice, while guaranteeing a constant approximation ratio in theory. Comprehensive simulations show that by incorporating rate adaptation into data center networks, the network-global routing optimization can bring up to 40% energy savings, even without switching off any network devices.

References

[1] L. A. Barroso and U. Hölzle, "The case for energy-proportional computing," *IEEE Computer*, vol. 40, no. 12, pp. 33–37, 2007.

[2] U.s. environmental protection agency's data center report to congress. [Online]. Available: http://tinyurl.com/2jz3ft

[3] A. G. Greenberg, J. R. Hamilton, D. A. Maltz, and P. Patel, "The cost of a cloud: research problems in data center networks," *Computer Communication Review*, vol. 39, no. 1, pp. 68–73, 2009.

[4] D. Abts, M. R. Marty, P. M. Wells, P. Klausler, and H. Liu, "Energy proportional datacenter networks," in *ISCA*, 2010, pp. 338–347.

[5] Cisco data center network topology. [Online]. Available: http://www.cisco.com/en/US/docs/solutions/Enterprise/Data_Center/DC_3_0/DC-3_0_IPInfra.html

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

[7] R. N. Mysore, A. Pamboris, N. Farrington, N. Huang, P. Miri, S. Radhakrishnan, V. Subramanya, and A. Vahdat, "Portland: a scalable fault-tolerant layer 2 data center network fabric," in *SIGCOMM*, 2009, pp. 39–50.

[8] A. G. Greenberg, J. R. Hamilton, N. Jain, S. Kandula, C. Kim, P. Lahiri, D. A. Maltz, P. Patel, and S. Sengupta, "Vl2: a scalable and flexible data center network," in *SIGCOMM*, 2009, pp. 51–62.

[9] C. Guo, G. Lu, D. Li, H. Wu, X. Zhang, Y. Shi, C. Tian, Y. Zhang, and S. Lu, "Bcube: a high performance, server-centric network architecture for modular data centers," in *SIGCOMM*, 2009, pp. 63–74.

[10] C. Guo, H. Wu, K. Tan, L. Shi, Y. Zhang, and S. Lu, "Dcell: a scalable and fault-tolerant network structure for data centers," in *SIGCOMM*, 2008, pp. 75–86.

[11] X. Meng, V. Pappas, and L. Zhang, "Improving the scalability of data center networks with traffic-aware virtual machine placement," in *INFOCOM*, 2010, pp. 1154–1162.

[12] L. Huang, Q. Jia, X. Wang, S. Yang, and B. Li, "Pcube: Improving power efficiency in data center networks," in *IEEE CLOUD*, 2011, pp. 65–72.

[13] B. Heller, S. Seetharaman, P. Mahadevan, Y. Yiakoumis, P. Sharma, S. Banerjee, and N. McKeown, "Elastictree: Saving energy in data center networks," in *NSDI*, 2010, pp. 249–264.

[14] Y. Shang, D. Li, and M. Xu, "Energy-aware routing in data center network," in *Green Networking*, 2010, pp. 1–8.

[15] P. Mahadevan, S. Banerjee, P. Sharma, A. Shah, and P. Ranganathan, "On energy efficiency for enterprise and data center networks," *IEEE Communications Magazine*, vol. 49, no. 8, pp. 94–100, 2011.

[16] X. Wang, Y. Yao, X. Wang, K. Lu, and Q. Cao, "Carpo: Correlation-aware power optimization in data center networks," in *INFOCOM*, 2012, pp. 1125–1133.

[17] B. Nordman and K. Christensen, "Reducing the energy consumption of network devices," *IEEE 802.3 tutorial*, pp. 1–30, 2005.

[18] Infiniband. [Online]. Available: http://www.infinibandta.org/

[19] N. Vasic and D. Kostic, "Energy-aware traffic engineering," in *e-Energy*, 2010, pp. 169–178.

[20] L. Wang, A. Fernández Anta, F. Zhang, C. Hou, and Z. Liu, "Routing for energy minimization with discrete cost functions," *CoRR*, vol. abs/1302.0234, 2013.

[21] M. Andrews, A. Fernández Anta, L. Zhang, and W. Zhao, "Routing for energy minimization in the speed scaling model," in *INFOCOM*, 2010, pp. 2435–2443.

[22] T. Benson, A. Anand, A. Akella, and M. Zhang, "Understanding data center traffic characteristics," *Computer Communication Review*, vol. 40, no. 1, pp. 92–99, 2010.