Green Data Center: a survey of existing techniques CSC547/447 Green Computing

Junxiao Shi University of Arizona
Gavin Simons University of Arizona
Nick Crutchfield University of Arizona
Chris Gniady University of Arizona
Beichuan Zhang University of Arizona

I. Introduction

Servers and data centers in U.S. consumed 61 billion kWh in 2006, which is 1.5% of total U.S. electricity consumption, and similar to the amount of electricity consumed by approximately 5.8 million average U.S. households [1]. Google, the Internet giant, consumed 2 billion kWh in 2010 [2]. As the Internet grows, this power consumption is likely to increase.

Most data center operate servers at 4% average utilization, which is a consequence of the biggest design priority being reliability and performance; aggressive cooling system keeps the room at freezer temperature [3]. There is significant potential for energy-efficiency improvement in data centers through better design and operation.

Energy-consuming equipment in a typical data center includes equipment that performs primary IT functions, as well as equipment that ensures continuous operation. IT equipment, including servers, storage devices, and network switches, consumes 59% of the power. Cooling system, typically computer room air conditioner (CRAC) units, consumes 33% of the power [4]. Power delivery equipment (such as Uninterrupted Power Supply) and office space also account for a small portion of electric power.

Energy efficiency of data centers is an important problem. In terms of money, 25% in a data center's cost of ownership is recurring cost which is mainly electric power. More importantly, the power density in data centers are increasing with the growth of server and network capacity, but heat removal is becoming a limiting factor.

The efficiency of a data center is represented by its power usage effectiveness (PUE):

$$PUE = \frac{\text{total data center power use}}{\text{total IT equipment power use}} \quad (1)$$

During the years many research works target at data center energy efficiency. Individual elements are replaced with more efficient ones. Agility is better supported at the data center design, so that resource utilization can be improved, and/or cooling cost can be reduced.

II. Energy-Efficient Elements

The major power-consuming equipment in data centers includes servers, network switches, and cooling system. Making these individual elements more power-efficient can reduce the overall power consumption of data centers.

Servers are custom designed and built from the ground up in Facebook's Open Compute Project [5]. The servers and storage valets are specifically optimized for energy efficiency. They do not contain anything not contributing to efficiency. As a result, a data center with Open Compute elements is 38% more efficient and 24% less expensive to build and run than other state-of-the-art data centers.

Network power consumption is reduced in IEEE Energy Efficient Ethernet (EEE) standard [6] by automatically adjusting energy use based on actual network traffic. An EEE-compliant product can enter a sleep mode during low activity, and re-engage when data transmission occurs. These are done while retaining full compatibility with existing equipment. HP's EEE-compliant line cards lower power consumption of 10G ports by 56%, and 30% on Gigabit Ethernet ports [7].

Newer network topologies are designed to save network power. Folded Clos network topology (aka fat tree) is widely used in data centers that need full bisection bandwidth to support the all-to-all traffic pattern of MapReduce applications. In a data center of 32K servers, 8235 switch chips is needed to build a Floded Clos topology. Flattened Butterfly (FBFLY) is an alternate network topology that also provides full bisection bandwidth, but uses only 4096 switch chips to build a network of the same size [8]. Therefore, FBFLY can reduce network power by half.

Natural cooling is being used to reduce the usage of mechanical cooling. Facebook's Oregon data center builds an outside air cooling system: cool outside air is filtered and pushed into the server room, and hot air goes out; this results in a PUE of 1.12 [9]. Google's Hamina data center uses sea water air conditioning [10]; their Douglas County facility uses recycled water [11] which conserves both energy and clean water. Such systems have a high up-front cost, but offers huge savings over the long run.

New heat transfer mediums are proposed, instead of the air circulating in a Hot Aisle / Cold Aisle layout. Intel is testing "submerged servers" [12]. Servers are submerged into mineral oil, which can carry more heat than air. An oil based data center can have much lower PUE than an air cooled data center [13]; it can operate at much higher temperature because a component is less likely to overheat; the effect of of "hot spots" is almost completely removed. However, liquid-ready components are much more expensive, and it takes longer to replace a failed component.

III. Energy-Efficiency by Agility

Agility inside a data center means that any server can be dynamically assigned to any service anywhere in the data center, while maintaining proper security and performance isolation between services [4]. Agility is enabled by server virtualization, virtual machine migration, and soft-defined networking.

Job scheduling algorithms can make use of this agility. Traditionally, job scheduling algorithms focus on job completion time. Power-aware job scheduling algorithms will consider energy-efficiency in addition to job completion time.

A. Basic Underlying Techniques

Server virtualization is the most important innovation in data center architecture. Many services require servers dedicated to them, for purposes of performance and security isolation, as well as avoiding software conflicts. The same number of physical machines (PMs) were needed, along with low utilization at most times. Server virtualization enables us to deploy services onto virtual machines (VMs), and pack multiple VMs onto one PM. Therefore, we need far less PMs, and have higher utilization on physical resources.

VM migration further improves the agility of virtualized servers. VMs can be moved among PMs with no service interruption. Therefore, we can overbook a PM by placing many VMs with low resource usage, and migrate a VM away when it demands more resource. During off-peak hours, many VMs are placed on a small number of PMs, and other PMs are put to sleep; in peak hours, more PMs are activated and run VMs. The user apps are unaware about the migrations and do not need any changes, because the number of VMs does not change over time.

Soft-Defined Networking (SDN) decouples network control plane from data forwarding plane. OpenFlow, an enabler of SDN, allows the path of network packets through the network of switches to be determined by software running on routers. This flexibility on network traffic management provides the security and performance isolation on network communication between a set VMs, and allows a job scheduling algorithm to choose appropriate paths so that some network switches can go to sleep.

B. Factors in Job Scheduling

Possible inputs to a job scheduling algorithm include:

- static configuration
  - server: physical resources, power consumption
  - network: topology, link capacity, power consumption of switches / line cards / ports
  - cooling: physical layout of racks / IT equipment / air conditioners
- current state
  - server: current utilization of each type of physical resource, actual power consumption
  - network: link utilization, queue length
  - cooling: thermometer readings
- jobs
  - arrival time: could be known a priori, or arrive dynamically (following a distribution, or bounded with a maximum change rate)
  - execution time: most jobs execute until computation completes, some jobs (such as web server) may execute forever, some jobs (such as video capturing during a football game) always stop at fixed time
  - physical resource demand: each process's (or VM's) demand for each type of physical resource; could be known a priori (in HPC), or change dynamically (most other scenarios)
  - traffic matrix: amount of network traffic between processes (or VM) of the same job; could be known a priori (in HPC), or change dynamically (most other scenarios)
  - constraints: certain process (or VM) must run on a different rack / row for reliability purpose
  - migration cost: whether a running process (or VM) can be migrated to another server, and the physical resource and network traffic required to do so

A job scheduling algorithm should take care of the following for each job:

- scheduling: which physical machines should the processes (or virtual machines) be placed, when should the job start execution
- network: on which path should the network route traffic from one process (or virtual machine) to the other
- migration: when a process (or virtual machine) should be migrated, which physical machine to place it, and which network path to transmit the state

C. Partial Optimization

Several power-aware job scheduling algorithms have been proposed that target at saving some but not all power components.

Spatial subsetting saves server power: loading is concentrated onto as few servers as possible, servers run at almost full utilization, and unused servers go to sleep; its disadvantage is that these running servers are much hotter than sleeping servers, and cooling power increases with the temperature of hottest servers.

Inverse-temperature assignment saves cooling power: loading is distributed onto all servers such that a cooler server does more work than a hotter server, making temperature balanced; this saves cooling power but increases server idle power.

PowerTrade [14] jointly optimizes server and cooling power: The configuration goes towards inverse-temperature assignment on a loading increase, and goes towards spatial subsetting on a loading decrease. In each round of a successive refinement process, a group of 10 servers is activated or deactivated, after 5 minutes the temperature is measured, and the change of the total of server and cooling power is calculated to determine whether the refinement should continue or stop.

ElasticTree [15] saves network power by traffic engineering. It doesn't know about jobs, but selects network paths for traffic

between physical machines so that some line cards can go to sleep.

TVMPP [16] places a cluster of virtual machines with high traffic demands among them onto physical machines close to each other, so most traffic stay in the rack instead of going across the data center. TVMPP aims at avoiding network congestion, but it can help ElasticTree achieve higher power saving, because less bisection bandwidth is needed.

Common issues with existing works are:

- reduces one power component at the cost of increasing another power component: spatial subsetting creates hot spots, increasing cooling power.
- degrades response time: spatial subsetting keeps just enough servers, so any increase needs activating more server and causes latency.
- does not account for the cost of migration: TVMPP shuffles virtual machines periodically.
- does not honor job constraints: a highly available service must distribute its replicas in multiple rows, but TVMPP is likely to place them into the same rack.

D. Overall Optimization

An ideal power-aware job scheduling algorithm should consider all power component, have minimal/bounded response time degradation, account for migration cost, and honor job constraints. Such a job scheduling algorithm does not exist yet.

REFERENCES

[1] Enengy Star, "Report to Congress on Server and Data Center Energy Efficiency," US Environmental Protection Agency, Tech. Rep., 2007. [Online]. Available: {http://www.energystar.gov/index.cfm?c=prod_development.server_efficiency_study}

[2] Katie Fehrenbacher, "Google reveals electricity use, aims for a third clean power by 2012," 2011. [Online]. Available: {http://gigaom.com/cleantech/google-reveals-electricity-use-aims-for-a-third-clean-power-by-2012/}

[3] Greening Greater Toronto, "Nine lessons in Greening IT," Tech. Rep., 2010. [Online]. Available: {http://www.greeninggreatertoronto.ca/pdf/GGT-Green-Exchange-IT-Summary.pdf}

[4] A. Greenberg, J. Hamilton, D. A. Maltz, and P. Patel, "The cost of a cloud: research problems in data center networks," *SIGCOMM Comput. Commun. Rev.*, vol. 39, no. 1, Dec. 2008.

[5] Facebook, "Open Compute Project - Hacking Conventional Computing Infrastructure." [Online]. Available: {http://opencompute.org/}

[6] Institute of Electrical and Electronics Engineers. and IEEE-SA Standards Board., "Ieee standard for information technology–telecommunications and information exchange between systems–local and metropolitan area networks–specific requirements part 3: Carrier sense multiple access with collision detection (csma/cd) access method and physical layer specifications amendment 5: Media access control parameters, physical layers, and management parameters for energy-efficient ethernet," *IEEE Std 802.3az-2010 (Amendment to IEEE Std 802.3-2008)*, 2010.

[7] Jim Duffy, "HP's green switch modules support Energy Efficient Ethernet standard," 2010. [Online]. Available: {http://www.networkworld.com/news/2010/120810-hp-green-eee-switch-modules.html}

[8] D. Abts, M. R. Marty, P. M. Wells, P. Klausler, and H. Liu, "Energy proportional datacenter networks," *SIGARCH Comput. Archit. News*, vol. 38, no. 3, Jun. 2010.

[9] Katie Fehrenbacher, "A rare look inside Facebook's Oregon data centers," 2012. [Online]. Available: {http://gigaom.com/cleantech/a-rare-look-inside-facebooks-oregon-data-center-photos-video/}

[10] Rich Miller, "Google Using Sea Water to Cool Finland Project," 2010. [Online]. Available: {http://www.datacenterknowledge.com/archives/2010/09/15/google-using-sea-water-to-cool-finland-project/}

[11] Jim Brown, "Helping the Hooch with water conservation at our Douglas County data center," 2012. [Online]. Available: {http://googleblog.blogspot.com/2012/03/helping-hooch-with-water-conservation.html}

[12] Rich Miller, "Intel Embraces Submerging Servers in Oil," 2012. [Online]. Available: {http://www.datacenterknowledge.com/archives/2012/09/04/intel-explores-mineral-oil-cooling/}

[13] Green Revolution Cooling, "The CarnotJet System - total fluid submersion cooling for OEM servers." [Online]. Available: {http://www.grcooling.com/docs/Green-Revolution-Cooling-CarnotJet-System-Pamphlet.pdf}

[14] F. Ahmad and T. N. Vijaykumar, "Joint optimization of idle and cooling power in data centers while maintaining response time," *SIGARCH Comput. Archit. News*, vol. 38, no. 1, Mar. 2010.

[15] B. Heller, S. Seetharaman, P. Mahadevan, Y. Yiakoumis, P. Sharma, S. Banerjee, and N. McKeown, "Elastictree: saving energy in data center networks," in *Proceedings of the 7th USENIX conference on Networked systems design and implementation*, ser. NSDI'10, 2010.

[16] X. Meng, V. Pappas, and L. Zhang, "Improving the scalability of data center networks with traffic-aware virtual machine placement," in *Proceedings of the 29th conference on Information communications*, ser. INFOCOM'10, 2010.