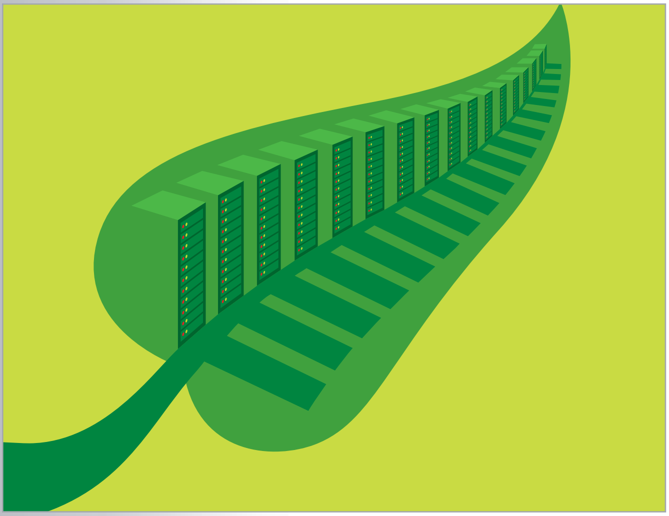

Energy-Efficient Data Centers

Weisong Shi
Wayne State University

Thomas F. Wenisch
University of Michigan

Data centers are the core of modern business environments as computation rapidly moved into the cloud in the last decade. Data centers are among the fastest-growing users of electricity in the US, consuming an estimated 91 billion kilowatt-hours of electricity in 2013, with a projection to increase to roughly 140 billion kilowatt-hours annually by 2020. This is the equivalent annual output of 50 power plants, costing US businesses $13 billion annually in electricity bills and emitting nearly 100 million metric tons of carbon pollution per year. When operating a data center of hundreds of thousands of servers, effective operation is essential to improve energy efficiency and environmental sustainability.

With cloud-based computing's aggressive adoption, the demands on data centers are growing exponentially, and both academia and industry must rethink how data centers are designed, built, and operated to be sustainable. Despite a decade of research and industrial innovation, a recent report from the Natural Resources Defense Council (NRDC) indicates typical small and mid-size data centers hosting private clouds still hold many wasteful practices. Although best practices at mega-scale commercial cloud operators (such as Facebook, Microsoft, Google, and Amazon) have addressed the most egregious wastes (inefficient cooling, for instance), best practices still need to permeate the data center landscape and address the remaining performance and efficiency challenges that afflict even the largest installations.

Around the mid-2000s, the advent of mega-scale Internet services and public cloud offerings led to a redesign of data center architectures, which addressed key inefficiencies, particularly in electrical and mechanical infrastructure. At the same time, the accelerated need for efficient servers spurred a generation of research on CPU, memory, network, and storage power-management techniques, which have led to a marked improvement in server efficiency and energy proportionality. However, this first generation of improvement has plateaued; further opportunity in the large-scale mechanical infrastructure is limited and no single server or network component stands out as the key source of inefficiency. Hence,

1089-7801/17/$33.00 © 2017 IEEE

it's time for a second, holistic, clean-slate redesign of the data center, encompassing new server architectures, heterogeneous computing platforms, radical networking paradigms, new mechanical and electrical designs, intelligent cluster management, and radical rethinking of software architectures while considering changing usage patterns (such as hybrid private/public clouds).

In This Issue

In this special issue, many of these problems are under active investigation. For example, in the context of multitenant data centers, Shaolei Ren's article "Managing Power Capacity as a First-Class Resource in Multitenant Data Centers" explores how to leverage market approaches for maximizing power capacity use. Multitenant data centers are a crucial but underexplored type of data center, where tenants individually manage their own physical servers and workloads while the operator is mainly responsible for facility support. The study supplants the current practice of simply allocating power capacity in a fixed manner with a dynamic, scalable, and coordinated market-based design.

Concerning the operation of a green data center, Prateek Sharma and colleagues' article "Design and Operational Analysis of a Green Data Center" describes the design and analysis of a state-of-the-art green university data center, the Massachusetts Green High-Performance Computing Center (MGHPCC), from the perspective of power, water, and carbon usage. The article offers several insights about MGHPCC's operational and efficiency characteristics.

Energy storage devices (ESDs) enable data centers to set smaller power budgets, because they can provide additional energy when the demand is expected to be higher than the budget. In "Data Center Peak Power Management with Energy Storage Devices," Baris Aksanli analyzes the economic feasibility of this methodology from three different perspectives, including comparing different ESDs based on their ability to manage data center peak power; demonstrating that peak shaving benefits might differ because of data center usage and ownership; and analyzing whether data centers should participate in electric utility programs with their ESDs to obtain additional savings.

In a data center, a significant portion of the initial capital expenditures and recurring operating expenditures are devoted to cooling. The article by Matt Skach and colleagues, "Thermal Time Shifting: Decreasing Data Center Cooling Costs with Phase-Change Materials," proposes to use phase-change materials (PCMs) to shape the thermal load of a data center, absorbing and releasing heat when it's advantageous to do so. The study finds that PCMs can reduce the necessary cooling system capacity by up to 12 percent without impacting peak throughput, or increase the number of servers by up to 14.6 percent without increasing the cooling load.

Regarding data center networks (DCNs), in "BEEP: Balancing Energy, Redundancy, and Performance in Fat-Tree Data Center Networks" Antonio Cleber de S. Araujo and colleagues propose an energy-efficient strategy, called BEEP, which combines multipath routing and the global overview offered by software-defined networks to balance energy efficiency, equipment redundancy level, and DCN performance. Experimental results show that BEEP accomplishes energy savings in DCNs without compromising the required redundancy level or network performance.

We hope that readers will find these articles interesting and informative. We also thank all of the authors for their submissions and all of the reviewers for their timely and high-quality reviews.

Weisong Shi is a professor of computer science at Wayne State University. His research interests mainly focus on big data systems, edge computing, energy-efficient computer systems, and mobile health. Shi has a PhD from the Chinese Academy of Sciences. He has received a US National Science Foundation CAREER award and served as the founding steering committee chair of IEEE/ACM Symposium on Edge Computing (SEC). He is an IEEE Fellow and an ACM Distinguished Scientist. Contact him at weisong@wayne.edu.

Thomas F. Wenisch is an associate professor of computer science and engineering at the University of Michigan, specializing in computer architecture. His ongoing work focuses on server and data center architectures, programming models for persistent memory, and architectural support for 3D medical image reconstruction. Wenisch has a PhD from Carnegie Mellon University. He has received a US National Science Foundation CAREER award. Contact him at twenisch@umich.edu.