Power Minimization Techniques for Networked Data Centers

Data centers can achieve energy-use goals through load balancing among server farms, which strikes an optimal tradeoff between power and delay.

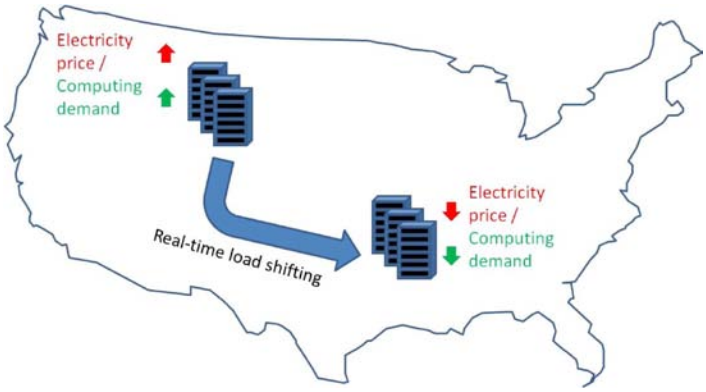
Introduction

Many large companies use a massive array of servers across multiple data centers, without having a global method to manage energy consumption based on customer demand. This study is creating algorithms designed to recognize the demand for computing services and to balance tasks across servers and data centers according to energy-use goals. This project targets "volume servers," the largest consumers of data center energy, where efficiency gains can yield substantial energy savings.

The fundamental problem of power optimization in data centers is the tradeoff between energy consumption and service delay. Therefore, a complete solution must include technologies that optimize power use and reduce delay in the computation and delivery of results. The algorithms developed in this project will reduce power by exploiting idle or low-load periods and select the optimal tradeoff when energy consumption and delay objectives are in conflict. This solution will optimize individual server performance and coordinate the energy consumption of server clusters within a data center and across geographically distributed data centers, minimizing an enterprise's overall energy costs.

Benefits for Our Industry and Our Nation

Although the national demand has daily peaks and valleys for both data center computing and electricity supply, no currently available technology can match the two demand curves and best manage cost and supply. This project addresses the need to match these demand curves for volume servers. The ability to shift computing resource use from high-usage resources to idle resources—and from peak-rate electricity locations to non-peak electricity locations—can reduce the need for power-plant growth and increase the responsiveness of cost-efficient computing assets.



Load balancing based on user demand and electricity cost. *Illustration courtesy of U.S. Department of Energy's Industrial Technologies Program.*

Applications in Our Nation's Industry

This technology will be most useful to organizations that rely upon large, networked data centers. It will provide a means to balance the use of computing equipment and data centers based on energy cost and user demand. Utility companies may be interested in partnering with users of this technology so as to manage their renewable energy generation and optimize demand-side management programs. The target customers for this technology include:

- Internet application and service providers including providers of cloud computing (a shared pool of configurable computing resources) infrastructure and services
- Organizations relying on cloud computing
- Organizations with multiple data centers
- Utility companies with demand-side management programs or renewable energy portfolios

Project Description

Volume servers are typically under-utilized, so there is a great deal of potential to reduce their energy consumption. For instance, virtualization increases server utilization and decreases the size of the required server pool, which reduces energy consumption for the server and its associated infrastructure. This project is creating complementary technologies to optimize energy consumption under any given load pattern. The goal is to minimize energy consumption and manage response delay time. This will support an optimal tradeoff between energy consumption and performance.

Guided by this framework, the project spans three synergistic technologies that will accomplish the following:

- Optimize the scheduling and speed scaling at individual servers
- Minimize delay in the delivery of computation results
- Balance loads across multiple servers in one or more data centers based on electricity costs and client locations.

These technologies can be implemented by servers, acceleration appliances (or software), and load balancers within a data center or across multiple data centers.

Barriers

The main barrier to this ubiquitous computing vision is that, to satisfy stringent delay requirements, protocol inefficiency limits the distance between the server and the client to within a few hundred miles. This constrains the architecture of content delivery infrastructure and reduces energy saving opportunities. In this case, without protocol inefficiency impeding performance, data centers can be located in areas with the lowest power and cooling costs and client requests can be routed to wherever intermittent renewable energy is available at the time. Delay-minimizing technology is thus central to low-cost, decentralized, and energy-efficient content delivery architecture for ubiquitous computing (computing services not tied to a specific computer or system).

Pathways

The expected outcomes include the following:

1. Theoretical foundation

- Mathematical models that capture multiple facets of fundamental tradeoffs between power and delay optimization based on control, optimization, and game theories
- Decomposition and characterization of optimal and suboptimal solutions to the global optimization problem

2. Algorithm development

- Development of decentralized algorithms to solve the formulated global optimization problems
- Performance evaluation of these algorithms via mathematical analysis and/or simulations

Milestones

- Completion of the mathematical framework for the global optimization problem (Completed)
- Formulation of algorithms that are practical for implementation
- Completion of performance analysis of algorithms

Commercialization

This project will result in scientific publications, patent applications, and potential software implementations. Non-confidential research results will be disseminated in technical journals, conferences, workshops, and on the web in a timely fashion.

Commercialization of this research will be pursued through FastSoft which has an established market presence with large data-center operators and web-based companies.

Project Partners

California Institute of Technology
Pasadena, CA
Principal Investigator: Steven Low
E-mail: slow@caltech.edu

Cornell University
Ithaca, NY

For additional information, please contact

Gideon Varga
Technology Manager
U.S. Department of Energy
Industrial Technologies Program
Phone: (202) 586-0082
E-mail: Gideon.Varga@ee.doe.gov

Information & Communications Technology

Equipment and Software

Research & Development

U.S. DEPARTMENT OF ENERGY
Energy Efficiency & Renewable Energy

EERE Information Center
1-877-EERE-INFO (1-877-337-3463)
eere.energy.gov/informationcenter

DOE/EE-0502 • May 2011
Printed with a renewable-source ink on paper containing at least 50% wastepaper, including 10% post consumer waste.