Enterprise data centers consume an alarmingly-high fraction of the world’s energy. The total carbon footprint of the world’s data centers is roughly the same as the CO2 emissions of the entire Czech Republic. In the US, the EPA estimates that data center energy consumption will reach over 100 billion kWh this year (2011), 2.5% of domestic power generation, at an estimated cost of $7.4 billion. Improving enterprise computing energy efficiency is a critical challenge.

Recently, there has been explosive growth in the deployment of “Web 2.0” applications in data centers. User experience has rapidly transitioned from accessing static web pages to providing a dynamic social experience where users tag, comment and share content. This drastic shift has fundamental implications on the performance, scaling, and energy requirements on data center hardware, software, and physical infrastructure.

**Challenge 1: Energy-efficient Platforms for OLDI Applications.** A central driver of the growth in the demand for Warehouse-Scale Computers (WSCs) is the popularity of a class of workloads that I call *On-line Data-Intensive* (OLDI) services. These workloads perform significant computing over massive data sets per user request but, unlike their off-line counterparts (such as MapReduce computations), they require responsiveness in the sub-second time scale at high request rates. Large search products, social networking, on-line advertising, and machine translation are examples of workloads in this class. Although the load in OLDI services can vary widely during the day, their energy consumption sees little variance due to the lack of energy proportionality of the underlying machinery. The scale and latency sensitivity of OLDI workloads also make them a challenging target for current power management techniques. Cluster-grain approaches that scale cluster size in response to load variation are inapplicable to OLDI because the number of servers provisioned in a cluster is determined primarily by the application’s aggregate memory footprint instead of the throughput of incoming requests. For a cluster to process an OLDI data set for even a single query with acceptable latency, the data set must be partitioned over thousands of nodes that act in parallel.

There is a pressing need for systems researchers to investigate energy-efficient platforms for OLDI applications. As an initial step, we at the University of Michigan are investigating energy-efficient server and cluster architectures for **memcached** workloads. Memcached provides a conceptually simple distributed main-memory key-value store to act as a cache for more expensive durable stores (e.g., database management systems). By aggregating the main-memory capacity of hundreds of commodity servers, datasets reaching into the terabytes can be available at extremely low latencies across a data center network. A number of companies already have sizable deployments—Facebook reports memcached clusters with over 20 TB of DRAM across more than 800 servers that service 150 million requests per second at per-request latencies in the hundreds of microseconds.

Despite its apparent simplicity (memcached comprises only a few thousand lines of code), performance and energy management for memcached proves to be surprisingly complex—optimal server and cluster architectures must balance network, memory, and CPU bandwidth while minimizing overall energy consumption under rapidly-varying load. Like other OLDI workloads, obvious power management schemes (e.g., shutting down part of the cluster) are not easily applicable because of the way in which the cache is distributed across nodes. Our preliminary investigations reveal that there is no one-size-fits-all energy-efficient server or cluster architecture for memcached—the optimal balance of CPU, network, and memory bandwidth
depends heavily on the average size of objects stored in the cache. Further work is needed to
determine how to design scalable, efficient servers and clusters for memcached and other OLDI
applications.

Challenge 2: Scalable Data Center Evaluation Methodology. Though the systems community
has great experience in modeling individual servers, we lack quantitative methods that can answer
questions about how applications behave at scale. Conventional architectural simulators do not
directly address data center-level issues and often require hours to simulate only seconds of real
time for a single machine; attempting to simulate tens, let alone thousands, of machines quickly
becomes prohibitive. Whereas existing simulators might be able to discern a memory system
bottleneck by modeling the leaf node of a web-search cluster, present-day infrastructure could not
be used to determine the level of traffic at which data-center-wide power capping mechanisms
throttle performance to the point where 99th-percentile response-time latencies exceed a desired
bound. The latter is an example of a design challenge that emerges only when modeling large
server ensembles.

To enable quantitative evaluation of WSC infrastructure, the systems community must develop
models of system behavior that target a higher level of abstraction than conventional (e.g.,
instruction-level) architectural exploration tools. One promising approach to WSC modeling that
we at the University of Michigan are pursuing is the use of stochastic simulation [MW10].
Stochastic simulation is an approach for system characterization, based on discrete-event
simulation of probabilistic models of workload/system behavior, to enable quantitative
exploration of data center-level challenges, such as performance optimization and power
management. Rather than simulate workloads at the granularity of an instruction, memory, or disk
access as in conventional tools, stochastic simulation builds on the theoretical framework of
queuing theory, where the fundamental unit of work is a task. Tasks can be characterized by a set
of statistical properties—random variables that describe their length, resource requirements,
arrival distribution, or other relevant properties—which are collected through observation of real
systems. Stochastic simulation abstracts the data center as an interrelated network of queues and
power/performance models describing the relevant behaviors of software/hardware components.
Our preliminary work [MSB+11] has demonstrated that stochastic simulation can be used to
study the performance/power characteristics of web search. Nevertheless, numerous challenges
remain, including (1) demonstrating stochastic simulation for workloads with dependencies
between tasks, (2) collecting/validating stochastic models for additional workloads, and (3)
developing power/performance models for servers and physical infrastructure.

Participant Bio. Thomas F. Wenisch is an Assistant Professor of CSE at the University of
Michigan, specializing in computer architecture. Tom's prior research includes memory streaming
for commercial server applications, store-wait-free multiprocessor memory systems, and rigorous
sampling-based performance evaluation methodologies. More recently, his work has focused on
server and data center energy efficiency, including the PowerNap server architecture [DGW11],
PowerRouting [PMZ+10] to reduce data center infrastructure costs, and memory system energy-
efficiency improvements [DMR+11]. Prior to his academic career, Tom was a software developer
at American Power Conversion, where he worked on data center thermal topology estimation.
Tom received his Ph.D. in ECE from Carnegie Mellon University.

2011.
center,” In ASPLOS 2010.