Sensing Data Centers for Energy Efficiency

By
Jie Liu — Microsoft Research
Andreas Terzis — Johns Hopkins University

Abstract

Data centers are large energy consumers today and their consumption is expected to increase further, driven by the growth in cloud services. The large costs and the environmental impact of this consumption have motivated data center operators to optimize data center operations. We argue that one of the underlying reasons for the low energy utilization is the lack of visibility into a data center's highly dynamic operating conditions. Wireless sensor networks promise to remove this veil of uncertainty by delivering large volumes of data collected at high spatial and temporal fidelities. The paper summarizes data center operations in order to describe the parameters that a data center sensing network would need to collect and motivate the challenges that such a network would face. We present technical approaches for the problems of data collection and management and close with an overview of Data Center Genome, an end-to-end data center sensing system.

1. Introduction

Internet services, such as news, e-mail, on-line storage, social networking, entertainment, e-commerce, search and advertising, have become intrinsic parts of people's lives. As a result, the IT infrastructure that powers these services is experiencing a rapid growth in recent years. A central part of this IT infrastructure are data centers, hosting servers and storage devices. According to an EPA survey, in 2006, the total energy consumption for data centers in the US was 61 billion kWh, enough to power 5.8 million US households [65]. Perhaps more importantly, IT power consumption is the fastest growing sector and is expected to double by 2011. Thus, improving data center efficiency not only reduces a company's operating cost but can also be considered as part of its social responsibility.

Data centers can have drastically different sizes and form factors, from a few server cabinets, to shipping containers, to dedicated warehouse-size buildings. The largest of data centers can consume more than 100 MW of electricity and cost a few hundred million dollars to build. Roughly speaking a dedicated data center has three main subsystems, *data systems*, including servers, storage, and network infrastructure, *power systems*, including transformers, energy storage units, backup generators, and the power distribution network, and *cooling systems*, including chill water towers, computer room air conditioning (CRAC) units, and air ducts.

Data centers have been traditionally over-provisioned: the power system is under subscribed; server rooms (called colocations, or *colos* for short) are over cooled; and server utilization is low on average. As a quantitative metric of this condition, the Power Utilization Efficiency (PUE), defined as the total facility power divided by the total power used by the data systems, of some data centers is greater than 2. In an effort to reduce operation overhead the data center industry is exploring ways to bring PUE as close to 1 as possible and to increase computing resource utilization as much as possible. Doing so requires the operators to have a detailed and up to date understanding of where and how power is consumed within a data center and specifically where heat is generated and how equipment are cooled.

Maintaining an up to date image of a data center is complicated by the observation that a data center's operating conditions are not static. Servers have typically three to five years of lifetime to leverage the latest innovations in hardware. Moreover, the services deployed on these servers change over time—some of the largest online services have continuous release cycles [40]. Last but not least, service workload varies driven by user demands [10, 11]. For example, twice as many users are logged in the Windows Live Messenger service during peak time, compared to off-peak periods. In turn, this workload variation causes the power consumption and heat output of the servers to vary. Typically, a heavy loaded server may consume twice as much power as an idle server. Therefore, depending on the physical layout and load distribution in a data center, the power and cooling systems are stressed at different times and locations. The existence of these data center dynamics introduces challenges to some of the most important decisions that operators have to make, including:

- *Capacity Planning*: While online services continue to grow, the more servers an existing data center can host, the fewer data centers need to be built. Given the immense capital costs of building new data centers and the inherent facility overhead in power consumption, improving the utilization of existing facilities is crucial.

- *Change Management*: Hardware and software components in data centers change over time. In this context, data center operators need to make informed decisions about where to place new equipment and replace existing ones to optimally use the space. In cloud computing and other online services, the management software needs to decide where to place virtual machines (VMs), applications, and user workload. In addition to server capacity and availability, power distribution and cooling availability are key factors that influence management decisions.

- *Real-time Control*: Given the reduced safety margins associated with aggressive power, cooling, and workload optimization, the role of real-time control is critical. The massive spatial distribution of physical variables and the interaction between physical and cyber activities make the control problem challenging.

- *Diagnostics*: Unsafe operating conditions, such as high temperatures, fast temperature variations and excessive vibrations, may lead to server performance throttling and degraded equipment reliability. Failures in some cases develop over long periods of time. For this reason, accumulating long term data about operating conditions and performance statistics can power retrospective analyses that help identify the failures' root causes and improve future operations.

While cyber-properties such as server workload, network traffic, and application performance are routinely monitored in software, the visibility into data center physical conditions and the cyber-physical interactions is rather poor. Traditional data centers have a small number of sensors associated with key pieces of equipment and inside the colos. These sensors however do not provide the granularity that is necessary to capture the physical conditions at the rack or server level and answer the previously mentioned questions. We argue that improving data center utilization and efficiency, requires a marked increase in our ability to sense physical conditions in side and across data centers – including space configuration, power consumption, air temperature, humidity, air flow speed, chill water temperature, and AC utilization – together with performance and workload management.

In this paper, we summarize data center operations in Section 2 and discuss the challenges of data center sensing in Section 3. Section 4 presents systems for data center sensing, data stream management as well as decision and control. Section 5 presents a case study, while Section 6 reviews some the other work in this area. We conclude in Section 7.

2. Data Center Background

To motivate the need for fine-grained data center sensing, we first describe a data center's physical infrastructure and discuss the dynamics of physical properties driven by computing and environmental variations.

(*a*) *Power Distribution and Cost*

A Tier-2 data center, which includes $N+1$ redundant infrastructure components providing 99.741% availability [25], is typical for hosting Internet services. In these data centers, power drawn from the grid is transformed and conditioned to charge the UPS system (based on batteries or flywheels). The un-interrupted power is distributed through Power Distribution Units (PDUs) to supply power to the server and networking racks. This portion of power is called the critical power, since it is used to perform "useful work." Power is also used by water chillers, computer room air conditioning (CRAC) systems, and humidifiers to provide appropriate temperature and humidity conditions for IT equipment.

The power capacity of a data center is primarily defined by the capacity of its UPS system, both in terms of the steady load that it can handle as well as the surges that it can withstand. For well-managed data centers, it is the maximum instantaneous power consumption from all servers allocated to each UPS unit that determines how may servers a data center can host.

The availability and cost of electricity may vary over time. For example, local green energy production such as from solar or wind can change daily or seasonally. Electricity price fluctuate dramatically in the real-time market. While data centers sign bulk rate contracts with utility companies, exceeding power cap can introduce huge financial cost. With proper sensing and control, a collaborative group of geo-distributed data centers can take advantage of energy price variation to reduce the total energy expense [48].

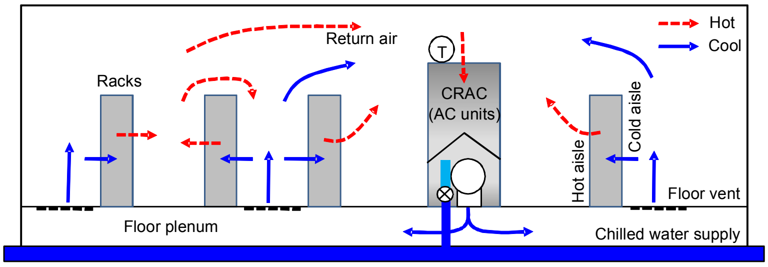

Figure 1. An illustration of an air-cooled data center on raised floors.

(*b*) *Cooling System*

Most data centers are air cooled, i.e. blowing cold air through the servers to maintain the electronic components within their operating ranges for temperature and humidity. Figure 1 shows the cross section of a typical data center with a cold-aisle-hot-aisle arrangement on a raised floor. The CRACs in the room draw warm air into the AC unit, where heat is exchanged with chilled water, and cold air is blown to the sub-floor. Perforated tiles allow the cold air to flow into the room, from where server fans drive it through the servers' chassis to cool key components.

Air cooling systems have slow dynamics. To avoid over-reaction and oscillations, CRAC units usually react every 15 minutes. Furthermore, their actions reach the servers after long propagation delays, depending on air dynamics, the volume of air in the room, and the thermal properties of servers and the building's materials.

Appropriate temperature and humidity ranges are important to maintain servers in reliable working conditions. Servers have protective temperature sensors which will throttle the CPU or even shut down the server if key components overheat. To prevent such overheating events from happening and considering the lack of fine-grain information about their environmental conditions, most data centers operate with conservative temperature and humidity settings. The American Society of Heating, Refrigerating, and Air-Conditioning Engineers (ASHREA) recommends that data centers operate between 20^oC to 25^oC and 30% to 45% relative humidity. However, excessive cooling does not necessarily improve reliability or reduce device failure rates [47]. Morden data centers start to agressively relax operation conditions and use ambient air to cool data centers directly as much as possible.

(*c*) *Workload Dynamics*

Resource utilization in data centers naturally fluctuates depending on service demands. For example, Figure 2 shows the total number of users connected to Windows Live Messenger and the rate of new user logins over a week, normalized to five million users and a login rate of 1,400 users/second. One can see that the number of users in the early afternoon is almost twice as large as the one after midnight and that the total demand in weekdays is higher than during weekends. One can also see flash crowds, during which a large number of users login in a short period of time. Other research studies recently reported similar demand variations [10,15,22].

Armbrust et al. reported another example of demand variation [4]: *"When Animoto made its service available via Facebook, it experienced a demand surge that*

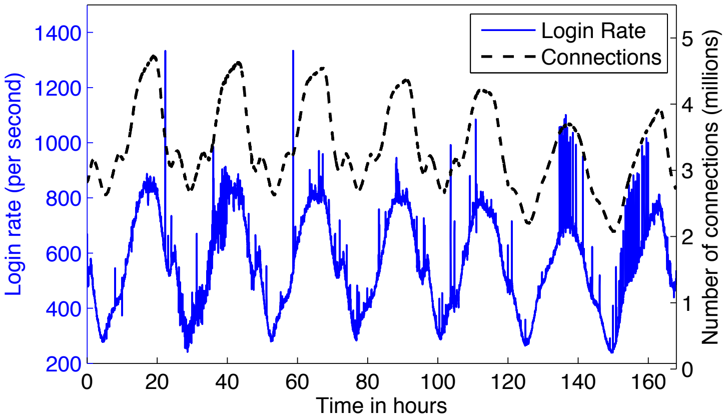

Figure 2. Load variation, in terms of total number of users and rate of new user logins, for the Microsoft Live Messenger service, normalized to 5 million users.

resulted in growing from 50 servers to 3500 servers in three days... After the peak subsided, traffic fell to a level that was well below the peak."

As a rough approximation, current servers consume ∼ 60% of their peak power when idle and power consumption above this point scales linearly with CPU utilization [5]. It is then easy to see that service type and workload variations cause the servers' power consumption to change and consequently impact the amount of heat they generate.

(*d*) *Interaction Between Cyber and Physical Systems*

The heat the servers generate does not dissipate evenly over space. Instead, a typical server's 4 to 10 fans direct heat towards its rear side. These air streams coupled with those generated by CRAC ventilation fans, generate complex air flow patterns within a data center.

Figure 3 presents heat maps generated from 24 sensors placed along the front and back of a row of server racks. In the cold aisle (i.e., front side), the temperature difference between the hottest and coldest spots is as high as 10°C. It is evident that the racks' mid sections, rather than their bottoms, are the coolest areas, even though cool air blows from the floor. This counter-intuitive heat distribution exists in almost all data centers and is driven by Bernoulli's principle. This principle states that an increase in fluid speed decreases its pressure. Thereby, fast cold air near the floor creates low pressure pockets which draw warm air from the back of the rack. The high temperature at the top right corner is due to uneven air flow which prevents cool air from reaching that area. As a consequence, hot air from the back of the rack flows to the front.

Air dynamics can be further complicated when a server is turned off, creating a local tunnel that connects the hot aisle at the server's rear to the cold aisle in the front. Figure 4 illustrates an example of this interaction; shutting down the controlled server causes an increase in the intake air temperature of the server below it. While few servers are affected by the actions of one server, a framework that predicts temperatures should consider these interactions.

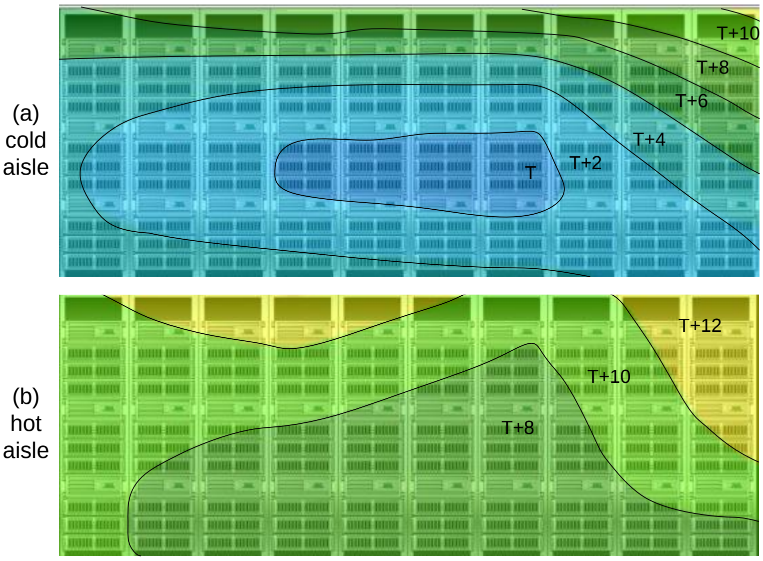

Figure 3. Cold aisle and hot aisle heat maps generated from sensor data.

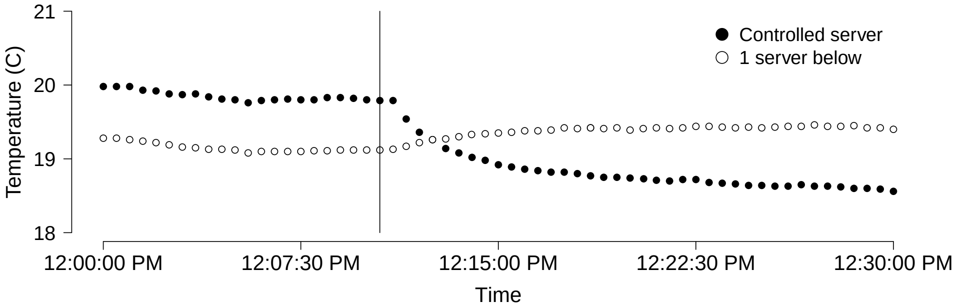

Figure 4. The intake air temperature of the controlled server decreases after the server is shut down, while the temperature of the server below increases. The vertical line indicates the time when the controlled server was shut down.

3. What to Sense

A system that provides a comprehensive view of a data center's conditions and is used to facilitate planning, management, and control needs to continuously monitor numerous physical and cyber variables, such as:

- **Physical Configuration.** The physical configuration information includes the location and type of servers, the topology of the data network, as well as the topology of the power distribution network. Although these specifications appear static, in practice they change continuously as old servers are decommissioned, new servers are installed, and the facility is upgraded in general. It is however possible to infer the topology of the power distribution and data network by generating signals with specific signatures from one side of the network and detecting them at the side [12, 32]. With tens of thousands pieces of equipment and hundreds of millions of dollars investment in a mega data center, asset management should be highly reliable and low cost. Traditional RFID technologies face challenges since data center environments are not RF-friendly due to multiple metal barriers.

- **Environmental conditions.** Temperatures and to some extent, humidities are critical for the safe operation of servers and to maintain high performance. Other factors, such as vibration and corrosive chemical substances may shorten equipment lifetime. The key temperature to control is the servers' internal temperature. However, those internal temperatures fluctuate widely depending on the workload and are therefore hard to use for controlling CRAC. More actionable sensor streams are the servers' intake temperatures. Humidity is important only to reduce the risk of condensation and electrostatic buildup due to temperature variations. Since there are no data center components that actively increase humidity, there is no need to sample humidity as densely as temperature.

- **Workload.** Workload can be measured in an application independent way (via performance counters that are widely supported by server operating systems), or through application-specific logs. In general, the more one knows about application semantics, the easier it is to control power consumption. For example, Chen et al. showed that controlling connection-oriented server that hold state, is quite different from controlling stateless servers (e.g., web farms) [11].

- **Power consumption.** Power is a scarce resource in data centers, especially when circuits are over-subscribed. Typical power distribution systems have built-in meters to measure aggregate power consumption at the circuit breaker level or above. These measurements can be used for safety and diagnostic purposes. Nevertheless, monitoring power consumption at the level of individual servers becomes increasingly important for power capping and control purposes. When a circuit approaches its safety boundary, the operator needs to know which service to migrate or which servers to shut down to maximize benefit and minimize performance degradation. Power consumption can be measured directly by sensors on a server motherboard, or using in-line sensors at the power plug. Perhaps more interestingly, since server power consumption is directly related to the utilization of its key components (e.g., CPU, disk, memory, network, etc.), if one builds a regression model from performance counters to power consumption, then it is possible to derive the power consumption of servers with the same type and configuration without physical sensors [26].

Among these key variables, workload and server power consumption, as inferred by performance counters, can be measured through the host OS. On the other hand, wireless sensor network technologies are more suitable for collecting asset data and environmental conditions. The remainder of the paper elaborates on such technologies.

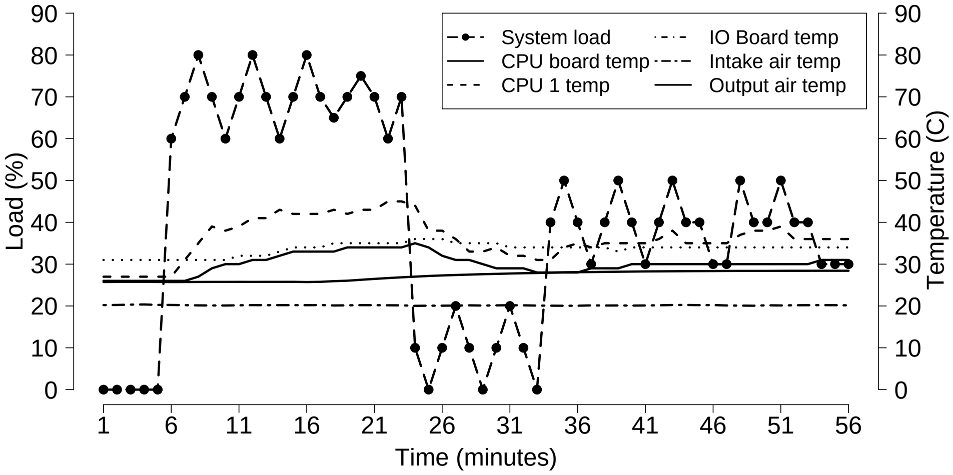

Figure 5. Temperature measured at different locations in and around an HP DL360 server. Also shown is the server's CPU load. Internal sensors reflect the server's workload instead of ambient conditions.

4. Research Challenge Highlights

The scale and density of the equipment in data center and the temporal-spacial variations of operation parameters bring significant challenges to data center sensing.

(*a*) *Data Collection*

(i) *Wired vs Wireless sensors*

There are seemingly several options for measuring the temperature and humidity distributions inside a data center. For one, thermal imagers can visualize temperature variations over the camera's view frame. However, considering the cramped data centers layout and the direct field of view requirement of infrared cameras, continuously capturing thermal images throughout the data center is prohibitively expensive. Alternatively, modern servers have several onboard sensors that monitor the thermal conditions near key server components, such as the CPUs, disks, and I/O controllers. Provisioning all servers with environmental sensors can be an overkill and an unnecessary increase in server cost. On the other hand, it is difficult to accurately estimate the room temperature and humidity from other onboard sensors. Figure 5 plots the temperature measured at various points along with the CPU utilization for an HP DL360 server with two CPUs. Air intake and output temperatures are measured with external sensors near the server's front grill and its back cover. It is evident from this figure that internal sensors are quickly affected by changes in the server's workload, rather than reflecting ambient conditions.

Intake air temperature (IAT) is important also because it can be used for auditing purposes. Server manufacturers and data center facility management contracts usually specify server operation conditions in terms of IAT. For example, the HP ProLiant DL360 (G3) servers require IAT to range from 10 ° to 35 °C. It is therefore necessary to place external sensors at regular intervals across the servers' air intake grills to monitor IAT.

More importantly, the communication mechanism used to retrieve the collected measurements is the other crucial aspect of the system design. Options in this case are divided in two categories: in-band vs. out-of-band. In-band data collection routes measurements through the server's operating system to the data center's (wired) IP network. The advantage of this approach is that the network infrastructure is (in theory) available and the only additional hardware necessary are relatively inexpensive USB-based sensors. However, data center networks are in reality complex and fragile. They can be divided into several independent domains not connected by gateways. Traversing network boundaries can lead to serious security violations. Finally, the in-band approach requires the host OS to be always on to perform continuous monitoring. Doing so however would prevent turning off unused servers to save energy.

Out-of-band solutions use separate devices to perform the measurements and a separate network to collect them. Self contained devices provide higher flexibility in terms of sensor placement, while a separate network does not interfere with data center operations. However, deploying a wired network that would connect each sensing point is undesirable as it would add thousands of network endpoints and miles of cables to an already cramped data center.

For this reason, wireless networks are an attractive method for collecting the desired environmental conditions. Moreover, networks based on IEEE 802.15.4 radios [24] (or 15.4 for short) are more attractive compared to Bluetooth or WiFi radios. The key advantage is that 15.4 networks have simpler network stacks compared to the alternatives. This simplicity has multiple implications. First, sensing devices need only a low-end MCU such as the MSP430 [62] thus reducing the total cost of ownership and implementation complexity. Second, the combination of low-power 15.4 radios and low-power MCUs leads to lower overall power consumption.

At the same time, there are significant challenges when using 15.4 networks for data center sensing, due to low data throughput and high packet loss rate. The maximum transmission rate of a 15.4 link is 250 Kbps, while effective data rates are usually much lower due to MAC overhead and multi-hop forwarding. Furthermore, the lower transmission power† can lead to high bit error rates especially in RF-challenging environments such as data centers. In fact, a quantitative survey of the RF environment in a data center by Liang et al. showed that significant challenges exist [29]. The following paragraphs summarize the results of that study.

(ii) *Data center RF environment*

Data centers present a radio environment different from the ones that previous sensor network deployments faced. This is intuitively true as metals are the dominant materials in a data center. In addition to switches, servers, racks, and cables, other metallic obstacles include cooling ducts, power distribution systems, and cable rails. Given this departure from RF environments studied in the past (e.g., [56,68]), characterizing this environment is crucial to understanding the challenges it poses to reliable data collection protocols.

† The TI CC2420 802.15.4 radio we use, transmits at 0 dBm, or 1 mW [61].

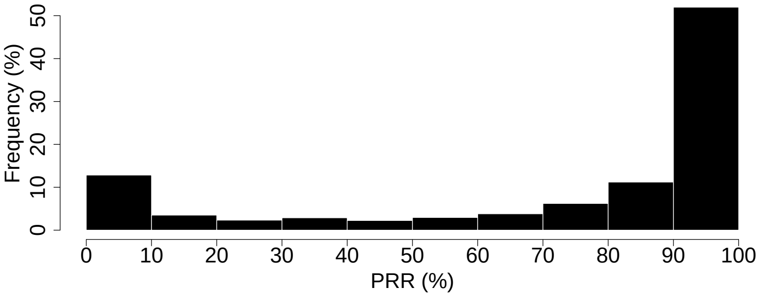

Figure 6. Distribution of packet reception ratios (PRR) across all the links from a 52-node data center site survey described in [29]. A large percentage of the network's links exhibit non-trivial loss rates.

For this reason Liang et al. performed a site survey by uniformly distributing 52 motes in an production data center spanning an area of approximately 1,100 m^2 [29]. The motes were placed at the top of the racks, following a regular grid pattern with adjacent nodes approximately 10m from each other. During the experiment, all nodes took turns broadcasting 1,000 128-byte packets with an inter-packet interval of 50 ms. All nodes used the 802.15.4 frequency channel 26 and transmitted their packets without performing any link-layer backoffs. Upon receiving a packet, each receiver logged the Received Signal Strength Indication (RSSI), the Link Quality Indicator (LQI), the packet sequence number, and whether the packet passed the CRC check.

We summarize the results from this survey below:

Neighborhood Size. The survey found that on average 50% of all the nodes are within a node's communication range and that a node's neighborhood can include as many as 65% of the network's nodes. Moreover, the neighborhood size in production deployments will be significantly higher as they consist of hundreds of nodes deployed over the same space. It is thereby imperative to devise mechanisms that minimize packet losses due to contention and interference.

Packet Loss Rate. Figure 6 illustrates the distribution of packet reception ratios (PRR) over all the network links. While the majority of the links have low loss rate (i.e., $< 10\%$), a significant percentage of links experience high number of losses. This observation suggests that even in dense networks data collection protocols must discover high-quality links to build end-to-end paths with low loss rates.

Link Qualities. Both RSSI and LQI measurements have been used to estimate link qualities [57, 64]. RSSI measures the signal power for received packets, while LQI is related to the chip error rate over the packet's first eight symbols (802.15.4 radios use a Direct Sequence Spread Spectrum encoding scheme). Indeed, the results shown in Figure 7 indicate that there is a strong correlation between RSSI/LQI and packet reception rates. Based on these results, one can use an RSSI threshold of -75 dBm to filter out potential weak links. Selecting this conservative threshold removes a large number of links. Fortunately, the network remains connected because each node has many neighbors with high RSSI links.

The results of the site survey also revealed that approximately 3.43% of losses in

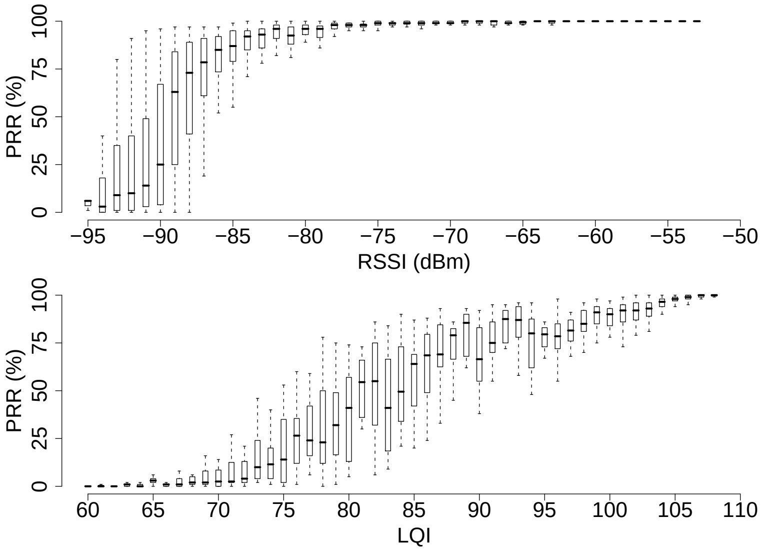

Figure 7. Boxplots of link PRR as a function of RSSI and LQI values from the data center RF survey in [29]. Boxplots show the sample minimum, first quantile, median, third quantile, and sample maximum. Links with RSSI $>$ −75 dBm and LQI $>$ 90 have persistently low PRR.

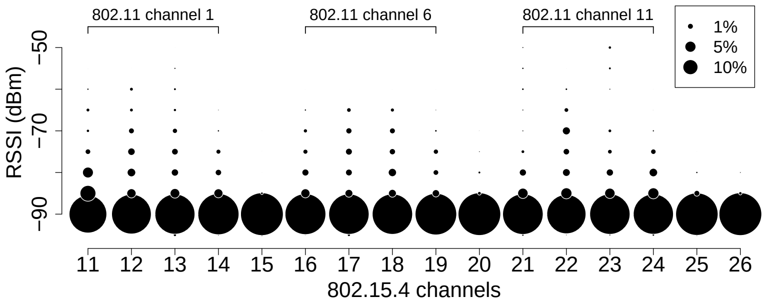

Figure 8. Background noise distribution across all 802.15.4 frequency channels in the 2.4 GHz ISM band. Each of the circumferences is proportional to the occurrence frequency of the corresponding RSSI level. Channels 15, 20, 25, and 26 are relatively quiet compared to other channels.

the network were due to CRC failures. However, since the 16-bit MAC-layer CRC used by the 802.15.4 standard is relatively weak, it might not detect all corrupted packets. To understand the extent of this potential problem an additional 16-bit CRC was included that covered the application-level payload to every packet transmission. As many as 1% of the total number of packets that passed the MAC-layer CRC failed the one at the application level. It is thereby crucial for applications that require data integrity to include an application-level CRC.

Background RF Interference. Figure 8 shows the background noise distribution measured on each of the sixteen 802.15.4 frequency channels available on the

2.4 GHz ISM band. The measurements were collected by a mote that sampled its RSSI register at a frequency of 1 KHz while no other 802.15.4 radios were active. A total of 60,000 samples were collected on each channel. Because the data center in which the measurements were taken has considerable levels of 802.11 traffic, 802.15.4 channels that overlap with 802.11 channels experienced higher noise levels. On the other hand, 15.4 channels 15, 20, 25, and 26 are relatively quiet. This results motivates the use of of all the quiet channels simultaneously.

(*b*) *Data Management*

The scale of data center sensing does not only bring challenges to data collection, but to data management as well. If a hundred performance counters and physical variables of four bytes each were collected at every server every 30 seconds on average, then a large online service provider with a million servers would have to manage data streams totaling more than 1 TB per day. In order to perform long term analysis, planning, and diagnostics, these data need to be archived over years. This scale introduces challenges not only to storing the data, but processing as well. For example, even reading the data to perform a rudimentary histogram query can take hours.

Recent advances in databases and data management in general address these challenges with data parallelism such as NoSQL (e.g., Google Big Table [9], Amazon SimpleDB [2], BerkeleyDB [39], and Tokyo Cabinet [18]), streaming databases (e.g., Borealis [1], Microsoft StreamInsight [34], and IBM InfoSphere [23]), and hybrid approaches like DataGarage [33].

NoSQL approaches are best for archiving large volumes of data with simple structures and performing embarrassingly parallelizable queries. While NOSQL does not preserve query accuracy over massive amounts of data, DataGarage combines relational databases (SQL Embedded) with MapReduce-type of distributed processing [14]. DataGarage is specifically designed to manage performance counters collected from servers in data centers. Realizing that most queries on performance counters only touch a single server and a continuous period of time, DataGarage divides time series into small and distributable chunks. For example, it uses a SQL database to store performance counters from one machine over a day. Upon receiving a query, such as identifying servers with CPU utilization higher than 80%, it spawns multiple SQL queries in parallel and then combines the results.

Streaming databases are useful for real-time monitoring and event detection. While the collected data flow through the database engine, predefined patterns and events can be detected. Streaming databases usually do not offer data archiving intrinsically, arguing that it is the high level events, rather than raw data, that are of interest to users. However, in many data center design and planning applications, it is important to perform hypothesis testing on historical data. Nevertheless, commercial software exists for archiving time series. For example, the PI System from OSI [41] was initially designed for compressing and archiving process control measurements. However, it has been used in recent years to archive power and environmental measurements.

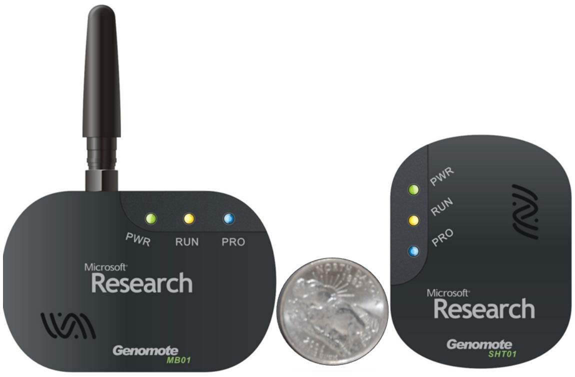

Figure 9. Two types of sensors designed for DC Genome. The wireless node (on the left) controls several wired nodes (on the right) to reduce the number of wireless sensors within the same broadcast domain.

5. Case Study

We use the Data Center (DC) Genome project at Microsoft Research (MSR) as a case study of using wireless sensor networks in a production data center.

The deployment site is a single 1,100 m^2 colo provisioned for 2.8 MW. There are roughly 250 racks in the colo arranged in 32 rows. Some racks are double sided with half-depth servers. They take cold air from both front and back, and exhaust hot air to the top of the rack.

(*a*) *Genomotes*

The DC Genome project uses Genomotes – temperature/humidity sensing nodes designed at MSR and shown in Figure 9. To reduce the density of wireless nodes and thus interference among them, the deployment uses two kinds of sensing nodes. The wireless master node (shown on the left) and several wired sensors (one example shown on the right) form a (wired) daisy chain covering one side of a rack, collecting data at different heights. The wireless nodes self-organize into an ad-hoc wireless network, called RACnet. This design satisfies the sensing density requirement while reducing the number of contending radios, without sacrificing deployment flexibility.

Both the master and the slave nodes use the MSP430 low power microcontroller from Texas Instruments (TI) [62]. Each master nodes also has a TI CC2420 802.15.4 compatible radio [61] and a flash memory chip that caches data locally to prevent data loss during disconnected periods. The whole chain is powered by a USB port connected to a server or a wall charger. Using a USB connection to power the whole mote chain means that unlike many previous sensor networks, power is not a critical concern in this application. The master node has a rechargeable battery to ensure continuous operation during the time that the server needs to be rebooted. The maximum current that one can draw from a USB port by a foreign device is 100 mA. This limitation means that it would impossible to use a server's USB port to power multiple (or even a single) WiFi-based sensing devices. Thus an IEEE 802.15.4 network is used to achieve low power, low cost, and communication flexibility. Finally, we note that using the same USB port to carry measurements is not

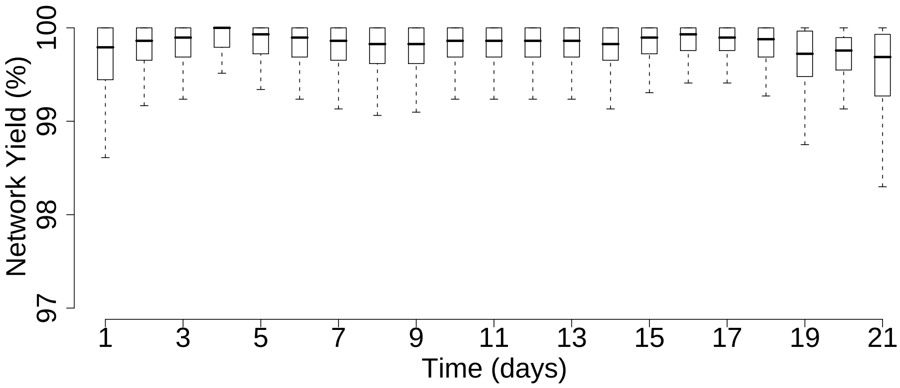

Figure 10. Daily network yield from the production deployment over a period of 21 days.

an option because it requires the installation of additional software on the servers – something that is not administratively possible in the Microsoft environment.

(*b*) *RACNet*

The data reliability challenge in DC Genome comes from the density of the sensor nodes. Even with the chain design, there are easily several hundred wireless master nodes within a data center colocation facility. Without carefully scheduling packet transmissions, the collisions among the sensors can significantly reduce the data yield [30].

To achieve reliable data collection in a dense wireless sensor network, RACNet uses a multi-channel collision avoidance protocol called Wireless Reliable Acquisition Protocol (WRAP). The main design rationale behind WRAP is to coordinate transmissions over multiple frequency channels. IEEE 802.15.4 standard defines 16 independent channels. WRAP uses multiple basestations as the receivers, each on a different channel. The basestations compute average packet delays over the network to estimate the congestion level and the network quality. If the delays among the channels are significantly different, WRAP tries to balance load among the basestations by allowing nodes to opportunistically switch to different channels.

Within each channel, WRAP builds a data collection tree, similar to protocols such as Collection Tree Protocol (CTP) [21]. However, in order to reduce self interference, WRAP uses a token ring schedule on top the collection tree. That is, the basestation will release a token to one of its children to give it permission to transmit. The token is passed among the nodes in a depth-first search way. At any given time, there is only one token per channel. Only the node that holds the token can originate a data transmission. All parent nodes in the tree forward the packet in a collision-free environment. After receiving an end-to-end acknowledgment, the sender can confirm that its data was successfully delivered and pass the token to the next node.

Figure 10 shows the per-node data yield over a period of three weeks in the production data center deployment. There were 696 sensing points in the deployment, with 174 wireless nodes. The sampling rate was one sample per sensor every 30 seconds. The data is considered successfully collected if it reaches the basestation within 30 seconds.

The median yield across all days was above 99.5%, while the lowest yield was always above 98%. This small packet loss is due to the fact that WRAP limits the number of end-to-end retransmission requests to five before it stops the attempt to recover the packet. Note that even though some data are not delivery by the deadline, they are saved locally in the master nodes' flash memory, and are retrieved at a later time.

(*c*) *Data compression and analysis*

A key insight when dealing with massive time-series datasets such as those collected in DC Genome is that not all queries look for precise answers. For example, when computing histograms, or performing trending analyses and classifications, users can often tolerate a certain degree of inaccuracy in exchange for prompt answers.

Cypress [52] is a system designed to compress and archive time series by decomposing them in both the frequency and time domains seeking for sparse representations. The decomposed time series are called *trickles*. An interesting feature of Cypress is that common queries such as histograms and correlations can be answered directly using compressed data. Furthermore, by using trickles, the search space of signals with high pairwise correlation can be drastically reduced, accelerating processing of such queries [38].

(*d*) *Using The Data*

We give one example of how the data collected by RACNet is used for analyzing data center operations.

Thermal runaway is a critical operation parameter, which refers to the temperature changes when a data center loses cool air supply. Predicting thermal runaway temperatures through simulations is very hard because their accuracy depends on the thermal properties of IT equipment that are difficult to obtain. On the other hand, RACNet collected actual thermal runaway data, during an instance when a CRAC was temporarily shut down for maintenance.

Figure 11 plots the temperature evolution at various locations across a row of ten racks during that maintenance interval. The CRAC was turned off for 12 minutes. It is evident that the mid sections –although the coolest spots normally– experience fast temperature increases when the CRAC stops. In contrast, temperature changes moderately at the two ends of the row, especially at the top and bottom of the rack. This is because those racks have better access to room air, which serves as a cooling reserve. This is an important finding because large temperature changes in a short period of time can be fatal to hard drives. For example, according to the specification of the Seagate SAS 300GB 15K RPM hard drive, the maximum safe rate of temperature change is 20^{o}C/hr. Notice that, in the middle of rack 7, the rate of temperature change is almost 40^{o}C/hr. This implies that storage intensive servers need to be placed carefully if the data center has a high risk of losing CRAC power.

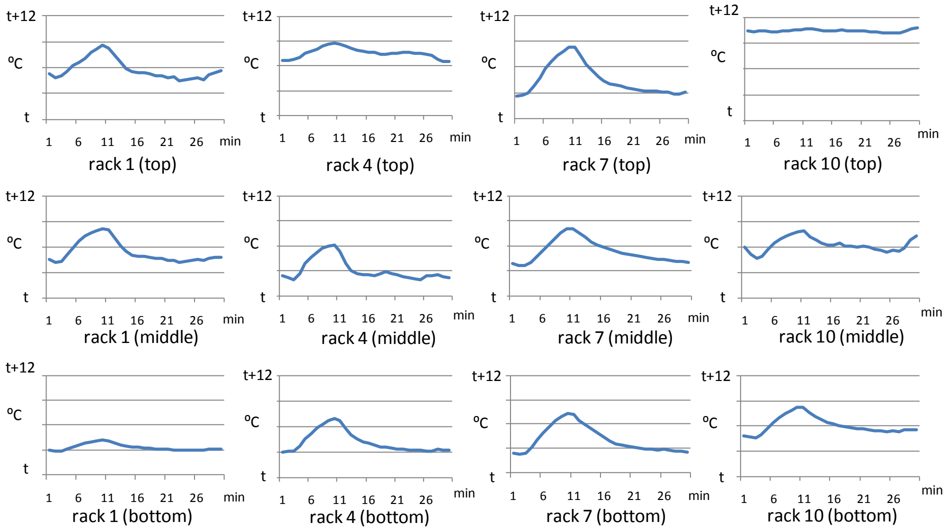

Figure 11. Temperature data from a row of ten racks, labeled from 1 to 10, during a thermal runaway event. Each rack has three sensors at the top, middle, and bottom, respectively. Temperature changes depend on locations.

6. Related Work

Multiple commercial products provide sensing of various environmental conditions and server performance parameters [3,19,54,58]. Federspiel Controls [19] uses OEM sensors from Dust Networks, which incorporate a frequency-hopping protocol called Time Synchronized Mesh Protocol (TSMP) [16]. TSMP network can support up to 255 nodes with a fixed TDMA schedule. SynapSense [58] provides the LiveImaging solution for monitoring data center environment conditions. To the best of our knowledge, LiveImaging supports only five minute sampling intervals. Both solutions use battery powered sensors, which limit their sampling rate and system lifetime. Sensicast [54] and Arch Rock [3] offer environmental sensors such as temperature, humidity, air pressure, air flow, and water chiller flow rate. Little is known about the architecture and performance of these commercial systems other than the fact that the Arch Rock system uses the recent 6LowPAN networking standard. While all the systems described so far use wireless networks to deliver the sensor measurements, Bash et al. used a network of wired sensors in an experimented that tested the feasibility of controlling CRAC units based on real-time sensor readings [6].

In the absence of measurements about the environment inside a data center thermal modeling has been used to predict conditions over space and time. Traditional thermal modeling of interior spaces is based on the zonal approach. It subdivides a region into blocks where temperature is assumed to be uniform and well-mixed, [31,53] and uses energy balances to predict the temperature time evolution. For data center applications, zonal models have been described in Refs. [51,63,66]. Our proposed work *ThermoCast* (see below) also falls into this category. These models are sometimes called "gray-box" models since some of the relevant physics are included (the term "black-box" methods refers to approaches in which temporal data mining is used to generate thermal response models without any reference to the underlying physics [37,45,59]).

Often, it is also important to predict the spatial patterns of cooling airflow, especially when buoyant motions occur, or when highly three-dimensional currents result from aggregate effects of many fans. Thus, modeling of airflow and heat transfer in data centers using Computational Fluid Dynamics and Heat Transfer (CFD/HT) has played an increasingly important role in their design and thermal optimization (e.g., [17,44,49,50,55]). Rambo & Joshi [50] provide a review of modeling approaches of the various layouts and tools employed. Most models rely on standard Reynolds averaged Navier-Stokes solvers with the $k - \epsilon$ turbulence model to account for turbulence. A detailed analysis coupled with multi-variable optimization is presented in Ref. [8]. The objective function of their optimization process was to minimize the rack inlet air temperature. Prior works often motivate the simulations by the need to gain detailed insights into flow and thermal state inside data centers, since little actual data is normally available. Also, they simulate steady-state conditions and do not take into account time-dependencies and dynamical effects.

Responsive thermal monitoring will enable us to understand and respond to machine room dynamics at time and space scales not previously realized. Previous studies of the interaction between workload placement, machine configuration, and the data center heat load have ignored that data centers are complex thermo-fluid systems. Current approaches use the global energy balance—inputs and outputs at nodes/chassis and the CRAC—to characterize the entire machine room. Initial efforts examined the contribution of components by permutation of components [7], or by modeling the data center as a linear combination of zones [36]. Some systems enhance global energy with heat-flow models across individual nodes [60], including a circuit representations of nodes [67]. Several represent the machine room as a set of chassis and model interactions among them as a cross inference problem [42,66]. Some insight into spatial properties can be realized by creating thermal maps using machine learning techniques applied to temperature data collected inside processing nodes [35]. Systems that do use fluid and heat models have explored configurations offline to determine the effect of failed components in a single rack [13] or machine room [7].

The relationship between workload placement, performance, and total energy use is poorly understood. The simplest issues are controversial, e.g. how does one place a fixed set of jobs across racks in order to minimize overall power consumption. Intuition dictates that since each active rack has fixed power costs (including PDUs and networking) jobs should be placed densely on racks to minimize fixed costs. Heat-flow based on global energy balance [66] has this property. However, measurements have shown that low-density job placement across many racks reduces node temperature and, thus, decreases cooling load and improves performance [27]. We believe that both of these guidelines apply, but to different thermal configurations that can only be resolved by sensing and modeling. Moreover, this example shows how the goals of reducing compute power and cooling power conflict at times and motivates the need for unified power management for all components in the data center.

Modeling and sensing will provide insight into the spatial and temporal distribution of energy and how to resolve the tension between cooling and performance. The potential savings are large and many different factors interact to govern power usage. Under certain conditions, active management of data centers at runtime saves up to 20% [43], dynamically sizing the allocation of nodes saves up to 20% [28], shaping/scaling workload on each node can save 20% [20], and an optimized data center can be more than 50% [46] more efficient. Because the heating and cooling system is turbulent, small changes can have large effects on energy draw and system performance [7]. Sensing and modeling will allow us to uncover the small configuration and workload modifications to realize energy savings.

7. Summary

The rapid growth in cloud services coupled with the large energy consumption of existing data centers has motivated data center operators to optimize their operations. In this paper we argue that one of the underlying reasons for the existing inefficient use of energy is the lack of visibility into a data center's operating conditions. However, recent innovations in wireless sensor networks can lift the veil of uncertainty, providing streams of measurements with high temporal and spatial fidelities. We present the requirements for such sensor networks and outline some of the technical challenges data center sensing introduces. Finally, we present DC Genome, an end-to-end data center sensing system deployed at a Microsoft data center.

References

[1] D. J. Abadi, Y. Ahmad, M. Balazinska, U. Cetintemel, M. Cherniack, J.-H. Hwang, W. Lindner, A. S. Maskey, A. Rasin, E. Ryvkina, N. Tatbul, Y. Xing, and S. Zdonik. The design of the borealis stream processing engine. In *2nd Biennial Conference on Innovative Data Systems Research (CIDR)*, 2005.

[2] Amazon Web Services LLC. Available from: `http://aws.amazon.com/simpledb/`, 2011.

[3] Arch Rock Corporation. Arch Rock Energy Optimizer. Available at: `http://www.archrock.com/downloads/datasheet/AREO_DS_web.pdf`, 2010.

[4] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph, R. H. Katz, A. Konwinski, G. Lee, D. A. Patterson, A. Rabkin, I. Stoica, and M. Zaharia. Above the clouds: A berkeley view of cloud computing. Technical Report UCB/EECS-2009-28, EECS Department, University of California, Berkeley, Feb 2009.

[5] L. A. Barroso and U. Hölzle. The case for energy-proportional computing. *IEEE Computer*, 40:33–37, December 2007.

[6] C. Bash, C. Patel, and R. Sharma. Dynamic thermal management of air cooled data centers. In *The Tenth Intersociety Conference on Thermal and Thermomechanical Phenomena in Electronics Systems (ITHERM)*, pages 452–461, June 2006.

[7] A. H. Beitelmal and C. D. Patel. Thermo-fluids provisioning of a high performance high density data center. *Distributed and Parallel Databases*, 21:227–238, 2007.

[8] S. Bhopte, D. Agonafer, R. Schmidt, and B. Sammakia. Optimization of data center room layout to minimize rack inlet air temperature. *Transactions of the ASME*, 128:380–387, 2006.

[9] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. E. Gruber. Bigtable: A distributed storage system for structured data. In *Seventh Symposium on Operating System Design and Implementation (OSDI)*, 2006.

[10] J. Chase, D. Anderson, P. Thakar, A. Vahdat, and R. Doyle. Managing energy and server resources in hosting centers. In *Proceedings of the ACM Symposium on Operating Systems Principles (SOSP)*, 2001.

[11] G. Chen, W. He, J. Liu, S. Nath, L. Rigas, L. Xiao, and F. Zhao. Energy-aware server provisioning and load dispatching for connection-intensive internet services. In *Proceedings of the 5th USENIX Symposium on Networked Systems Design and Implementation (NSDI)*, pages 337–350, Berkeley, CA, USA, 2008. USENIX Association.

[12] K. Chen, C. Guo, H. Wu, J. Yuan, Z. Feng, Y. Chen, S. Lu, and W. Wu. Generic and automatic address configuration for data center networks. In *ACM SIGCOMM*, 2010.

[13] J. Choi, Y. Kim, A. Sivasubramaniam, J. Srebric, Q. Wang, and J. Lee. A CFD-based tool for studying temperature in rack-mounted servers. *IEEE Transactions on Computers*, 57, 2008.

[14] J. Dean and S. Ghemawat. MapReduce: Simplified Data Processing on Large Clusters. In *Sixth Symposium on Operating System Design and Implementation*, 2004.

[15] R. Doyle, J. Chase, O. Asad, W. Jin, and A. Vahdat. Model-Based Resource Provisioning in a Web Service Utility. In *In Proceedings of the 4th USENIX Symposium on Internet Technologies and Systems*, 2003.

[16] I. Dust Networks. Technical overview of time synchronized mesh protocol (tsmp). Available at: `http://www.dustnetworks.com/sites/default/files/TSMP_Whitepaper.pdf`, 2006.

[17] E.Samadiani, Y. Joshi, and F. Mistree. The thermal design of a next generation data center: A conceptual exposition. *Journal of Electronic Packaging*, 130, 2008.

[18] FAL Labs. Available from: `http://fallabs.com/tokyocabinet/`, 2011.

[19] Federspiel Controls. Optimizing data center uptime with dynamic cooling.

[20] V. W. Freeh, N. Kappiah, D. K. Lowenthal, and T. K. Bletsch. Just-in-time dynamic voltage scaling: Exploiting inter-node slack to save energy in MPI programs. *Journal of Parallel and Distributed Computing*, 68(9):1175–1185, Sept. 2008.

[21] O. Gnawali, R. Fonseca, K. Jamieson, D. Moss, and P. Levis. Collection Tree Protocol. In *Proceedings of the 7^{th} ACM Conference on Embedded Networked Sensor Systems (SenSys)*, pages 1–14, Nov 2009.

[22] T. Heath, B. Diniz, E. V. Carrera, W. M. Jr., and R. Bianchini. Energy conservation in heterogeneous server clusters. In *Proceedings of ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP)*, 2005.

[23] IBM Corp. InfoSphere Streams. Avaiblable from: `http://www-01.ibm.com/software/data/infosphere/streams/`, 2011.

[24] IEEE Standard for Information technology – Telecommunications and information exchange between systems – Local and metropolitan area networks. Specific requirements – Part 15.4: Wireless Medium Access Control (MAC) and Physical Layer (PHY) Specifications for Low-Rate Wireless Personal Area Networks (LR-WPANs). Available at `http://www.ieee802.org/15/pub/TG4.html`, May 2003.

[25] W. P. T. IV and K. G. Brill. Tier classifications define site infrastructure performance. Technical report, Uptime Institute, White Paper, 2008.

[26] A. Kansal, F. Zhao, J. Liu, N. Kothari, and A. Bhattacharya. Virtual machine power metering and provisioning. In *ACM Symposium on Cloud Computing (SOCC)*, 2010.

[27] K. Karavanic. Scalable methods for performance and power data collection and analysis. In *Los Alamos Computer Science Symposium*, 2009.

[28] J. Leverich and C. Kozyrakis. On the energy (in)efficiency of hadoop clusters. In *Proceedings of HotPower*, 2009.

[29] C.-J. M. Liang, J. Liu, L. Luo, A. Terzis, and F. Zhao. RACNet: a high-fidelity data center sensing network. In *Proceedings of the 7^{th} ACM Conference on Embedded Networked Sensor Systems (Sensys)*, pages 15–28, 2009.

[30] C.-J. M. Liang, J. Liu, L. Luo, A. Terzis, and F. Zhao. Racnet: A high-fidelity data center sensing network. In *Proceedings of The 7th ACM Conference on Embedded Networked Sensor Systems (SenSys 2009)*, 2009.

[31] P. Linden. The fluid mechanics of natural ventilation. *Annual Review of Fluid Mechanics*, 31:201–238, 1999.

[32] J. Liu. Automatic server to circuit mapping with the red pills. In *Workshop on Power Aware Computing and Systems (HotPower '10)*, 2010.

[33] C. Loboz, S. Smyl, , and S. Nath. DataGarage: Warehousing Massive Performance Data on Commodity Servers. In *36th International Conference on Very Large Data Bases, Very Large Data Bases (VLDB)*, 2010.

[34] Microsoft Corp. Microsoft StreamInsight. Available from: `http://msdn.microsoft.com/en-us/library/ee362541.aspx`, 2011.

[35] J. Moore, J. Chase, and P. Ragananthan. Consil: Low-cost thermal mapping of data centers. In *Workshop on Tackling Computer Systems Problems with Machine Learning Techniques*, 2006.

[36] J. Moore, J. Chase, P. Ranganathan, and R. Sharma. Making scheduling "cool": temperature-aware workload placement in data centers. In *USENIX Annual Technical Conference*, 2005.

[37] J. Moore, J. S. Chase, and P. Ranganathan. Weatherman: Automated, online and predictive thermal mapping and management for data centers. In *Proceedings of the 2006 IEEE International Conference on Autonomic Computing*, 2006.

[38] A. Mueen, S. Nath, and J. Liu. Fast approximate correlation for massive time-series data. In *Proceedings of the 2010 ACM SIGMOD international conference on Management of data (SIGMOD)*, 2010.

[39] Oracle. Oracle Berkeley DB 11g. Available from: `http://www.oracle.com/technetwork/database/berkeleydb/overview/index.html`.

[40] T. O'Reilly. What is Web 2.0: Design Patterns and Business Models for the Next Generation of Software. *Communications & Strategies Magazine*, (1), 2007.

[41] OSISoft. The PI System. Available from: `http://www.osisoft.com/software-support/what-is-pi/What_Is_PI.aspx`, 2011.

[42] E. Pakbaznia, M. Ghasemazar, and M. Pedram. Temperature-aware dynamic resource provisioning in a power-optimized datacenter. In *Proceedings of Design, Automation, and Test in Europe*, 2010.

[43] L. Parolini, N. Tolia, B. Sinopoli, and B. H. Krogh. A cyber-physical systems approach to energy management in data centers. In *Proceedings of the International Conference on Cyber-Physical Systems*, 2010.

[44] C. Patel, C. E. Bash, C. Belady, L. Stahl, and D. Sullivan. Computational fluid dynamics modeling of high compute density data centers to assure system inlet air specifications. In *Proceedings of IPACK*, 2001.

[45] D. Patnaik, M. Marwah, R. Sharma, and N. Ramakrishnan. Sustainable operation and management of data center chillers using temporal data mining. In *Proceedings of KDD*, 2009.

[46] S. Pelley, D. Meisner, T. F. Wenisch, and J. W. VanGilder. Understanding and abstracting total data center power. In *In Proceedings of the Workshop on Energy Efficient Design*, 2009.

[47] E. Pinheiro, W.-D. Weber, and L. A. Barroso. Failure trends in a large disk drive population. In *Proceedings of the 5th USENIX Conference on File and Storage Technologies (FAST 2007)*, February 2007.

[48] A. Qureshi, R. Weber, H. Balakrishnan, J. Guttag, and B. Maggs. Cutting the electric bill for internet-scale systems. In *Proceedings of ACM SIGCOMM 2009*, New York, NY, USA, 2009. ACM.

[49] J. Rambo and Y. Joshi. Reduced order modeling of steady turbulent flows using the pod. In *Proceedings of HT2005, ASME*, 2005.

[50] J. Rambo and Y. Joshi. Modeling of data center airflow and heat transfer: State of the art and future trends. *Distributed and Parallel Databases*, 21:193–225, 2007.

[51] L. Ramos and R. Bianchini. C-Oracle: Predictive thermal management for data centers. In *Proceedings of the 14th IEEE International Symposium on High Performance Computer Architecture (HPCA)*, 2008.

[52] G. Reeves, J. Liu, S. Nath, and F. Zhao. Managing massive time series streams with multi-scale compressed trickles. In *VLDB'2009: Proceedings of 35th Conference on Very Large Data Bases*, August 2009.

[53] P. Riederer, D. Marchio, J. Visie, A. Husaunndee, and R. Lahrech. Room thermal modelling adapted to the test of hvac control system. *Building and Environment*, 37:777–790, 2002.

[54] Sensicast, Inc. Sensinet. Available at: `http://www.sensicast.com/data_center.php`.

[55] A. Shah, V. Carey, C. Bash, and C. Patel. Exergy analysis of data center thermal management systems. *Journal of Heat Transfer*, 130, 2008.

[56] K. Srinivasan, P. Dutta, A. Tavakoli, and P. Levis. Some implications of low power wireless to IP networking. In *HotNets '06*, Nov. 2006.

[57] K. Srinivasan and P. Levis. RSSI is Under Appreciated. In *EmNets '06*, May 2006.

[58] SynapSense Corporation. LiveImaging: Wireless Instrumentation Solutions. Available at: `http://www.synapsense.com/`, 2008.

[59] Q. Tang, S. K. S. Gupta, and G. Varsamopoulos. Energy-efficient thermal-aware task scheduling for homogeneous high-performance computing data centers: A cyber-physical approach. *IEEE Transactions on Parallel and Distributed Systems*, 19:1458–1472, 2008.

[60] Q. Tang, T. Mukherjee, S. K. S. Gupta, and P. Cayton. Sensor-based fast thermal evaluation model for energy efcient high-performance datacenters. In *Intelligent Sensing and Information Processing*, 2006.

[61] Texas Instruments. 2.4 GHz IEEE 802.15.4 / ZigBee-ready RF Transceiver. Available at `http://www.chipcon.com/files/CC2420_Data_Sheet_1_3.pdf`, 2006.

[62] Texas Instruments. MSP430x1xx Family User's Guide (Rev. F). Available at `http://www.ti.com/litv/pdf/slau049f`, 2006.

[63] T.Heath, A.P.Centeno, P.George, L.Ramos, Y.Jaluria, and R. Bianchini. Mercury and freon: Temperature emulation and management for server systems. In *Proceedings of ASPLOS*, 2006.

[64] TinyOS. MultiHopLQI. Available from: `http://www.tinyos.net/tinyos-1.x/tos/lib/MultiHopLQI`, 2004.

[65] U.S. Environmental Protection Agency. EPA Report on Server and Data Center Energy Efficiency: ENERGY STAR Program, 2007.

[66] N. Vasic, T. Scherer, and W. Schott. Thermal-aware workload scheduling for energy efficient data centers. In *Proceedings of International Conference on Autonomic Computing*, 2010.

[67] L. Wang, G. von Laszewski, J. Dayal, X. He, A. J. Younge, and T. R. Furlani. Towards thermal aware workload scheduling in a data center. In *Proceedings of the Symposium on Pervasive Systems, Algorithms and Networks*, 2009.

[68] J. Zhao and R. Govindan. Understanding Packet Delivery Performance In Dense Wireless Sensor Networks. In *Sensys '03*, 2003.