PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A
MATHEMATICAL, PHYSICAL & ENGINEERING SCIENCES

Sensing data centres for energy efficiency

Jie Liu and Andreas Terzis

Phil. Trans. R. Soc. A 2012 **370**, 136-157
doi: 10.1098/rsta.2011.0245

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY A

Phil. Trans. R. Soc. A (2012) **370**, 136–157
doi:10.1098/rsta.2011.0245

Review

Sensing data centres for energy efficiency

By Jie Liu[1,*] and Andreas Terzis[2]

[1] *Microsoft Research, Redmond, WA 98052, USA*
[2] *Department of Computer Science, Johns Hopkins University, Baltimore, MD 21218, USA*

Data centres are large energy consumers today, and their consumption is expected to increase further, driven by the growth in cloud services. The large monetary cost and the environmental impact of this consumption have motivated operators to optimize data centre management. We argue that one of the underlying reasons for the low-energy utilization is the lack of visibility into a data centre'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 centre operations in order to describe the parameters that a data centre sensing network should collect and motivate the challenges that such a network faces. We present technical approaches for the problems of data collection and management and close with an overview of a data centre genome, an end-to-end data centre sensing system.

Keywords: sensor networks; data centres; energy efficiency

1. Introduction

Internet services, such as news, e-mail, online storage, social networking, entertainment, e-commerce, search and advertising, have become intrinsic parts of people's lives. As a result, the information technology (IT) infrastructure that powers these services has been experiencing a rapid growth in recent years. A central part of this IT infrastructure are data centres, hosting servers and storage devices. According to an Environmental Protection Agency survey, in 2006, the total energy consumption for data centres in the US was 61 billion kWh, enough to power 5.8 million US households [1]. Perhaps more importantly, IT power consumption is the fastest growing sector and is expected to double by 2011. Thus, improving data centre efficiency not only reduces a company's operating cost, but can also be considered as part of its social responsibility.

Data centres can have drastically different sizes and form factors, from a few server cabinets, to shipping containers, to dedicated warehouse-size buildings. The largest data centres can consume more than 100 MW of electricity and cost a few hundred million dollars to build. Roughly speaking, a dedicated data centre has three main subsystems: *data systems*, including servers, storage and

*Author for correspondence (jie.liu@microsoft.com).

One contribution of 11 to a Theme Issue 'Sensor network algorithms and applications'.

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 centres 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 centres is greater than two. In an effort to reduce operation overhead, the data centre industry is exploring ways to bring PUE as close to one 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 centre and specifically where heat is generated and how equipment is cooled.

Maintaining an up-to-date image of a data centre is complicated by the observation that a data centre's operating conditions are not static. Servers have typically 3–5 years of lifetime to leverage the latest innovations in hardware, yet facilities are in operation for decades. Moreover, the services deployed on these servers change over time—some of the largest online services have continuous release cycles [2]. Last but not least, service workload varies, driven by user demands [3,4]. For example, twice as many users are logged in the Windows Live Messenger service during peak time, compared with 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 centre, the power and cooling systems are stressed at different times and locations. The existence of these data centre dynamics introduces challenges to some of the most important decisions that operators have to make. These challenges include the following.

— *Capacity planning*. While online services continue to grow, the more servers an existing data centre can host, the fewer data centres need to be built. Given the immense capital costs of building new data centres and the inherent facility overhead in power consumption, improving the utilization of existing facilities is crucial.
— *Change management*. Hardware and software components in data centres change over time. In this context, data centre 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, applications and user workload. In addition to server capacity and availability, power distribution and cooling availability are key factors that influence placement quality.
— *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 root causes of the failures and improve future operations.

While cyber-properties, such as server workload, network traffic and application performance, are routinely monitored in software, visibility into data centre physical conditions and cyber-physical interactions is rather poor. Traditional data centres 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 centre utilization and efficiency requires a marked increase in our ability to sense physical conditions inside and across data centres—including space configuration, power consumption, air temperature, humidity, air flow speed, chill water temperature and air conditioning (AC) utilization—together with performance and workload management.

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

2. Data centre background

To motivate the need for fine-grained data centre sensing, we first describe a data centre'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 centre, which includes $N+1$ redundant infrastructure components providing 99.741 per cent availability [5], is typical for hosting Internet services. In these data centres, power drawn from the grid is transformed and conditioned to charge the uninterruptible power supply (UPS) system (based on batteries or flywheels). The uninterrupted 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, CRAC systems and humidifiers to provide appropriate temperature and humidity conditions for IT equipment.

The power capacity of a data centre 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 centres, it is the maximum instantaneous power consumption from all servers allocated to each UPS unit that determines how may servers a data centre can host.

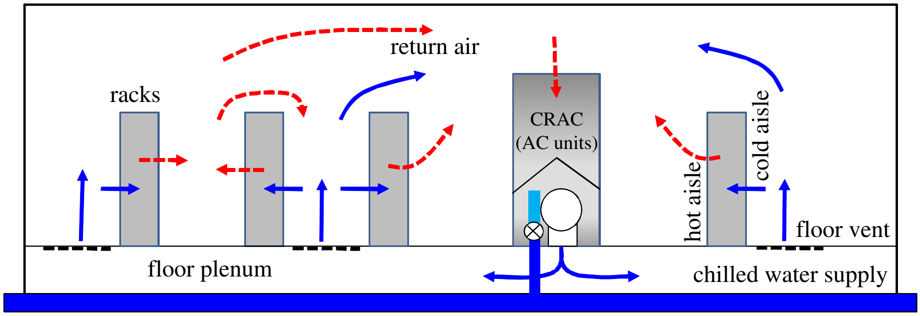

Figure 1. An illustration of an air-cooled data centre on raised floors. Dashed arrows indicate 'hot' and solid arrows indicate 'cool'. (Online version in colour.)

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 prices fluctuate dramatically in the real-time day-ahead market. While data centres sign bulk rate contracts with utility companies, exceeding power caps can introduce huge financial costs. With proper sensing and control, a collaborative group of geographically distributed data centres can take advantage of energy price variations to reduce the total energy expense [6].

(*b*) *Cooling system*

Most data centres 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 a cross section of a typical data centre 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 server chassis to cool key components.

Air cooling systems have slow dynamics. To avoid over-reaction and oscillations, CRAC units usually react every 15 min. Furthermore, their actions reach the servers after long propagation delays, depending on air dynamics, the volume of air in the room as well as 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 that will throttle the central processing units (CPUs) 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 centres operate with conservative temperature and humidity settings. The American Society of Heating, Refrigerating and Air-Conditioning Engineers (ASHRAE) recommends that data centres operate between 20 and 25°C and 30–45% relative humidity. However, excessive cooling does not necessarily improve reliability or reduce device failure rates [7]. More recently, data centres have started to aggressively relax operation conditions and use ambient air to cool equipment directly as much as possible.

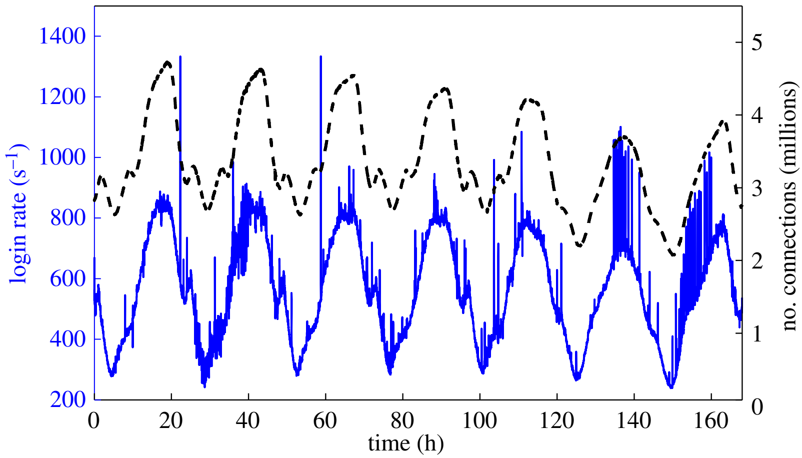

Figure 2. Load variation, in terms of total number of users and rate of new user logins, for the Microsoft LIVE MESSENGER service, normalized to five million users. Dashed lines indicate connections and solid lines indicate login rate. (Online version in colour.)

(c) *Workload dynamics*

Resource utilization in data centres 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 1400 users per 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. Studies have recently reported similar demand variations [4,8,9].

Armbrust *et al.* [10] reported another example of demand variation:

> 'When Animoto made its service available via Facebook, it experienced a demand surge that 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'. [10], p. 11

As a rough approximation, current servers consume approximately 60 per cent of their peak power when idle and power consumption above this point scales linearly with CPU utilization [11]. It is then easy to see that service type and workload variations cause the power consumption of the servers 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 4–10 fans direct the heat towards its rear side. These air streams, coupled with those generated by CRAC ventilation fans, generate complex air flow patterns within a data centre.

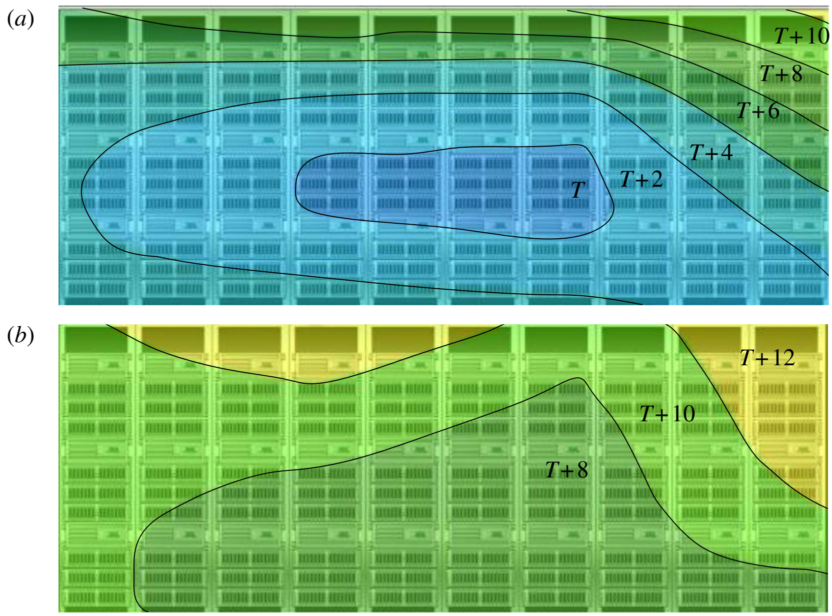

Figure 3. (*a*) Cold aisle and (*b*) hot aisle heat maps generated from sensor data. (Online version in colour.)

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 counterintuitive heat distribution exists in almost all data centres 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 that draw warm air from the back of the rack. The high temperature at the top right corner is owing 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 (IAT) 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.

3. What to sense

A system that provides a comprehensive view of a data centre's conditions and is used to facilitate planning, management and control needs to continuously monitor numerous physical and cyber variables. These include the following.

— *Physical configuration.* The physical configuration information includes the location and type of servers, the topology of the data network, as well as the

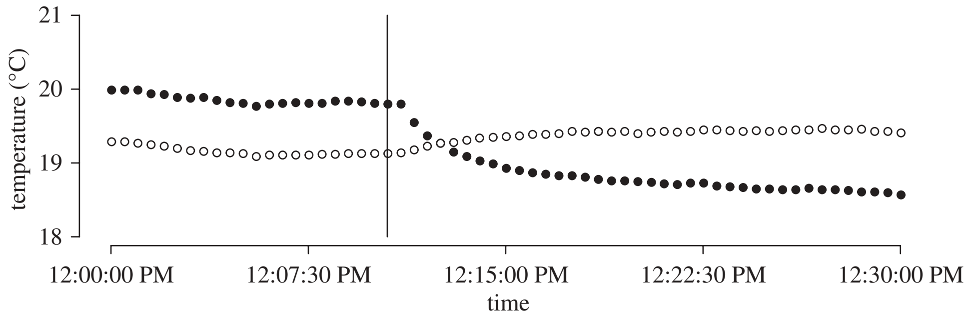

Figure 4. The intake air temperature of the controlled server decreases after the server is shut down, while the temperature of the server below increases. The vertical line indicates the time when the controlled server was shut down. Filled circles, controlled server; open circles, one server below.

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 gradually. With tens of thousands of pieces of equipment and hundreds of millions of dollars investment in a mega data centre, asset management should be highly reliable and low cost. Typical barcode scanning and spreadsheet copying methods for asset management are slow and error prone. Traditional radio-frequency (RF) identification technologies face challenges since data centre environments are not RF-friendly owing to multiple metal barriers. 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 [12,13].

— *Environmental conditions.* Temperature and, to some extent, humidity 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 internal temperature of the servers. However, those internal temperatures fluctuate widely depending on the workload and are therefore hard to use for controlling the CRAC(s). More actionable sensor streams are the intake temperatures of the servers. Humidity is important only to reduce the risk of condensation and electrostatic buildup owing to temperature variations. Since there are no data centre 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.* [3] showed that controlling connection-oriented servers that hold state is quite different from controlling stateless servers (e.g. web farms).
— *Power consumption.* Power is a scarce resource in data centres, 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, it is then possible to derive the power consumption of servers with the same type and configuration without physical sensors [14].

Among these key variables, workload and server power consumption, as inferred by performance counters, can be measured through the host operating system (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 these technologies.

4. Research challenge highlights

The scale and density of the equipment in a data centre and the temporal–spatial variations of operation parameters introduce significant challenges to data centre sensing.

(*a*) *Data collection*

(i) *Wired versus wireless sensors*

There are seemingly several options for measuring the temperature and humidity distributions inside a data centre. For one, thermal imagers can visualize temperature variations over the camera's view frame. However, considering the cramped layout of data centres and the direct field of view requirement of infrared cameras, continuously capturing thermal images throughout the data centre 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 input–output (IO) 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 ProLiant 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.

IAT is important also because it can be used for auditing purposes. Server manufacturers and data centre facility management contracts usually specify server operation conditions in terms of IAT. For example, the HP ProLiant DL360

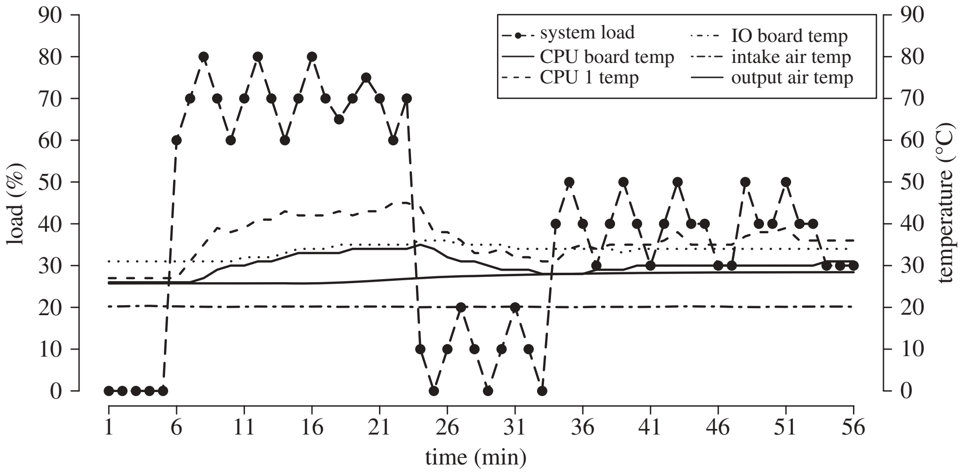

Figure 5. Temperature measured at different locations in and around an HP DL360 server. Also shown is the server's CPU load. Internal sensors reflect the workload of the servers instead of ambient conditions.

(G3) servers require IAT to range from 10 to 35°C. It is therefore necessary to place external sensors at regular intervals across the air intake grills of the servers to monitor IAT.

More importantly, the communication mechanism used to retrieve the collected measurements is the other crucial aspect of system design. Options in this case are divided into two categories: in-band versus out-of-band. In-band data collection routes the measurements through the server's OS to the data centre's (wired) internet protocol network. The advantage of this approach is that the network infrastructure is (in theory) available and the only additional hardware necessary are relatively inexpensive sensors plugging into the universal serial bus (USB) ports of servers. However, data centre 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 production data centre 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 centre.

For this reason, wireless networks are an attractive method for collecting the desired environmental conditions. Moreover, networks based on Institute of Electrical and Electronics Engineers (IEEE) 802.15.4 radios [15] (or 15.4 for short) are more attractive compared with Bluetooth or WiFi radios. The key advantage is that 15.4 networks have simpler network stacks compared with the alternatives. This simplicity has multiple implications. First, sensing devices need only a low-end microcontroller unit (MCU) such as the MSP430 [16], 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 centre sensing, owing to low data throughput and high packet loss rate. The maximum transmission rate of a 15.4 link is 250 kilobits per second, while effective data rates are usually much lower owing to medium access control (MAC) overhead and multi-hop forwarding. Furthermore, the lower transmission power[1] can lead to high bit error rates, especially in RF-challenging environments such as data centres. In fact, a quantitative survey of the RF environment in a data centre by Liang *et al.* [18] showed that significant challenges exist. The following paragraphs summarize the results of that study.

(ii) *Data centre radio-frequency environment*

Data centres present a radio environment different from those that previous sensor network deployments faced. This is intuitively true as metals are the dominant materials in a data centre. 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. [19,20]), characterizing this environment is crucial to understanding the challenges it poses to reliable data collection protocols.

For this reason, Liang *et al.* [18] performed a site survey by uniformly distributing 52 wireless embedded sensor nodes, sometimes called motes, in a production data centre spanning an area of approximately 1100 m^2. The motes were placed at the top of the racks, following a regular grid pattern with adjacent nodes approximately 10 m from each other. During the experiment, all nodes took turns broadcasting 1000 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 cyclic redundancy check (CRC) check.

We summarize the results from this survey below.

— *Neighbourhood size.* The survey found that on average, 50 per cent of all the nodes are within a node's communication range and that a node's neighbourhood can include as many as 65 per cent of the network's nodes. Moreover, the neighbourhood 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 owing 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. less than 10%), a significant percentage of links

[1]The commonly used TI CC2420 802.15.4 radio transmits at a maximum power of 0 milliwatts in decibels (dBm), or 1 mW [17].

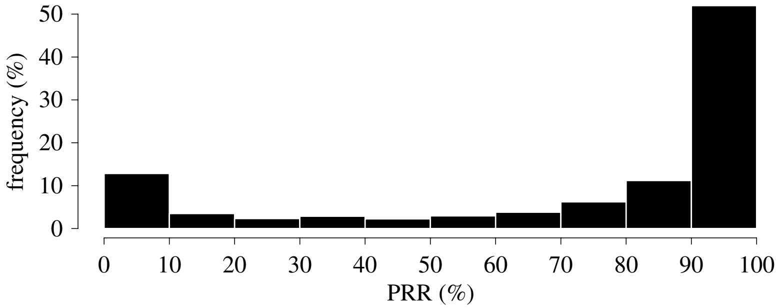

Figure 6. Distribution of packet reception ratios (PRR) across all the links from a 52 node data centre site survey described by Liang *et al.* [18]. A large percentage of the network's links exhibit non-trivial loss rates.

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 [21,22]. 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 neighbours with high RSSI links.

The results of the site survey also revealed that approximately 3.43 per cent of losses in the network were owing to CRC failures. However, as 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 per cent 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 industrial, scientific and medical (ISM) radio 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 centre 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, channel numbers 15, 20, 25 and 26 in the IEEE 802.15.4 band are relatively quiet. This result motivates the use of all the quiet channels simultaneously.

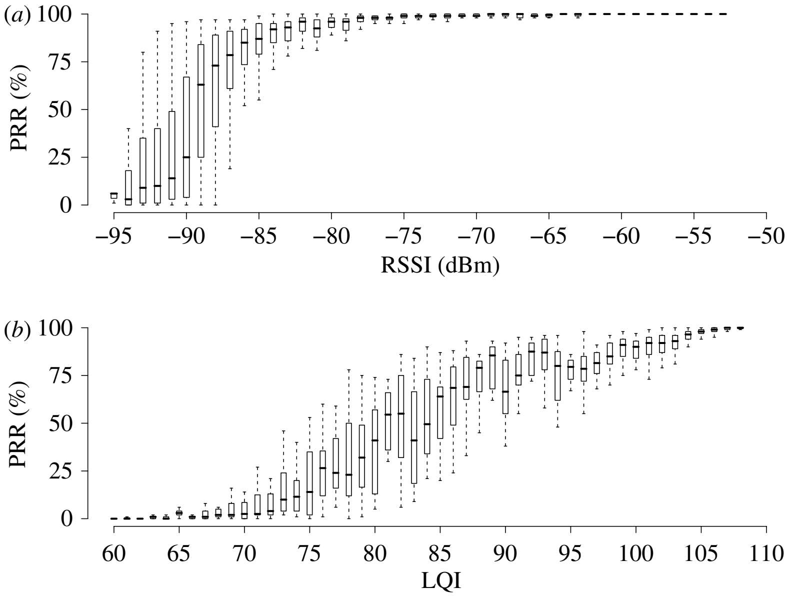

Figure 7. Boxplots of link PRR as a function of (*a*) received signal strength indication (RSSI) and (*b*) LQI values from the data centre RF survey in [18]. Boxplots show the sample minimum, first quantile, median, third quantile and sample maximum. Links with RSSI ≥ 75 dBm and LQI > 90 have persistently low PRR.

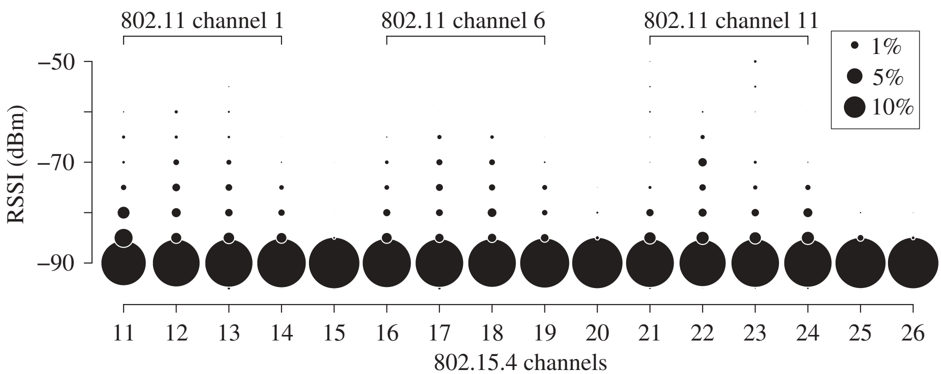

Figure 8. Background noise distribution across all 802.15.4 frequency channels in the 2.4 GHz ISM band. Each of the circumferences is proportional to the occurrence frequency of the corresponding RSSI level. Channels 15, 20, 25 and 26 are relatively quiet compared with other channels.

(*b*) *Data management*

The scale of data centre sensing not only brings challenges to data collection, but to data management as well. If a hundred performance counters and physical variables of 4 bytes each were collected at every server every 30 s on average, then a large online service provider with a million servers would have to manage data streams totalling 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 Not only structured query language (NoSQL) (e.g. Google BIGTABLE [23], Amazon SIMPLEDB [24], BERKELEYDB [25] and Tokyo CABINET [26]), streaming databases (e.g. BOREALIS [27], Microsoft STREAMINSIGHT [28], IBM INFOSPHERE [29]) and hybrid approaches such as DATAGARAGE [30].

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 [31]. DATAGARAGE is specifically designed to manage performance counters collected from servers in data centres. 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 an 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 per cent, 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 centre 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 [32] was initially designed for compressing and archiving process control measurements. However, it has been used in recent years to archive power and environmental measurements.

5. Case study

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

The deployment site is a single 1100 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 (figure 9*a*) and several wired sensors (figure 9*b*) form a (wired) daisy chain covering one side of a rack, collecting data at

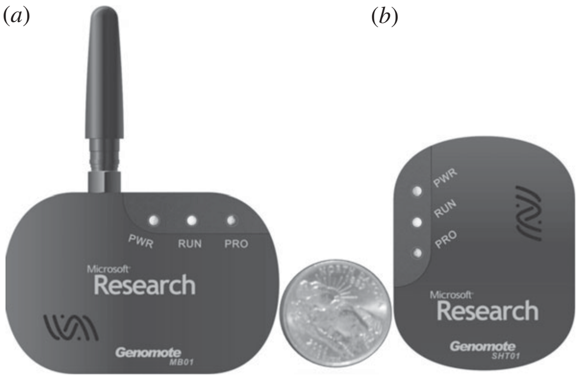

Figure 9. Two types of sensors designed for DC Genome. (*a*) The wireless node controls (*b*) several wired nodes to reduce the number of wireless sensors within the same broadcast domain.

different heights. The wireless nodes self-organize into an ad hoc wireless network, called a reliable acquisition network (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) [16]. Each master node also has a TI CC2420 802.15.4 compatible radio [17] 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 be 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 an option because it requires the installation of additional software on the servers—something that is not administratively possible in the Microsoft environment.

(*b*) *Reliable acquisition network*

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 centre colocation facility. Without carefully scheduling packet transmissions, the collisions among the sensors can significantly reduce the data yield [18].

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. The IEEE 802.15.4 standard defines 16 independent channels. WRAP uses multiple basestations as

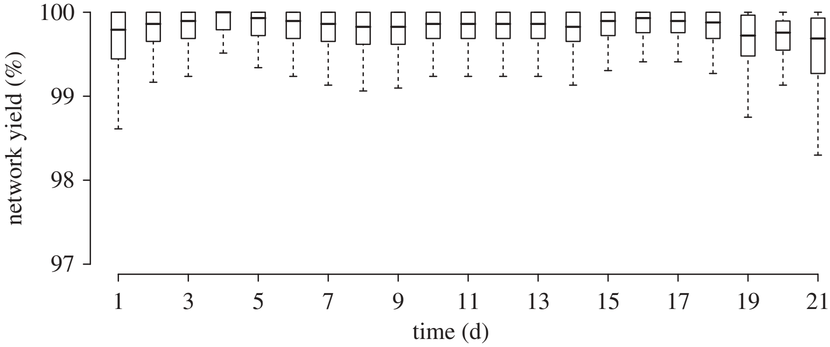

Figure 10. Daily network yield from the production deployment over a period of 21 days.

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, then 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 [33]. However, in order to reduce self-interference, WRAP uses a token ring schedule on top of the collection tree. That is, the basestation will release a token to one of its children to give it the 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 were 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 centre deployment. There were 696 sensing points in the deployment, with 174 wireless nodes. The sampling rate was one sample per sensor every 30 s. The data are considered successfully collected if it reaches the basestation within 30 s.

The median yield across all days was above 99.5 per cent, while the lowest yield was always above 98 per cent. 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 delivered by the deadline, they are saved locally in the flash memory of the master nodes, 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 [34] 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 [35].

(*d*) *Using the data*

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

Thermal runaway is a critical operation parameter, which refers to the temperature changes when a data centre loses its cool air supply. Predicting thermal runaway temperatures through simulations is very hard because their accuracy depends on the thermal properties of IT equipment, which 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 min. 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 300 GB 15 K RPM hard drive, the maximum safe rate of temperature change is $20^{\circ}\mathrm{C\,h^{-1}}$. Notice that, in the middle of rack 7, the rate of temperature change is almost $40^{\circ}\mathrm{C\,h^{-1}}$. This implies that storage-intensive servers need to be placed carefully if the data centre has a high risk of losing CRAC power.

6. Related work

Multiple commercial products provide sensing of various environmental conditions and server performance parameters [36–39]. Federspiel Controls [36] uses original equipment manufacturer sensors from dust networks, which incorporate a frequency-hopping protocol called time-synchronized mesh protocol (TSMP) [40]. The TSMP network can support up to 255 nodes with a fixed TDMA schedule. SynapSense [37] provides the LiveImaging solution for monitoring data centre environment conditions. To the best of our knowledge, LiveImaging supports only 5 min sampling intervals. Both solutions use battery powered sensors, which limit their sampling rate and system lifetime. Sensicast [39] and Arch Rock [38] 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

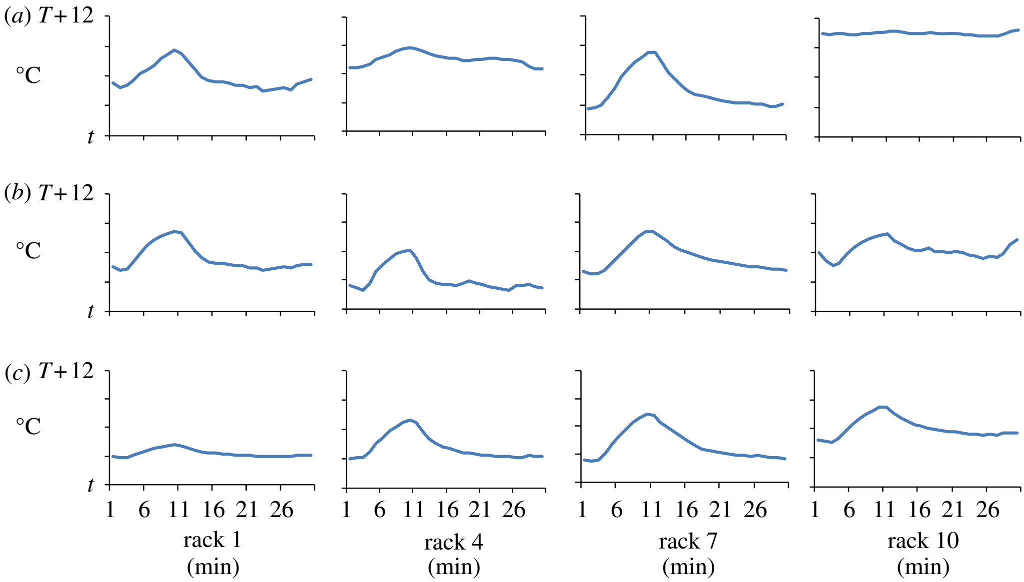

Figure 11. Temperature data from a row of ten racks, labelled from 1 to 10, during a thermal runaway event. Each rack has three sensors at the (*a*) top, (*b*) middle and (*c*) bottom, respectively. Temperature changes depend on locations. (Online version in colour.)

the sensor measurements, Bash *et al.* [41] used a network of wired sensors in an experiment that tested the feasibility of controlling CRAC units based on real-time sensor readings.

In the absence of measurements about the environment inside a data centre, thermal modelling has been used to predict conditions over space and time. Traditional thermal modelling 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, [42,43] and uses energy balances to predict the temperature time evolution. For data centre applications, zonal models have been described in earlier studies [44–46]. These models are sometimes called 'grey-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 [47–49]).

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 the aggregate effects of many fans. Thus, modelling of airflow and heat transfer in data centres using computational fluid dynamics, and heat transfer has played an increasingly important role in their design and thermal optimization (e.g. [50–54]). Rambo & Joshi [52] provide a review of modelling approaches of the various layouts and tools employed. Most models rely on standard Reynolds-averaged Navier–Stokes solvers with the k–ϵ turbulence model to account for turbulence. A detailed analysis coupled with multi-variable optimization is presented in Bhopte *et al.* [55]. 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 centres, as little actual data are 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 centre heat load have ignored that data centres 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 [56], or by modelling the data centre as a linear combination of zones [57]. Some systems enhance global energy with heat flow models across individual nodes [58], including circuit representations of nodes [59]. Several represent the machine room as a set of chassis and model interactions among them as a cross-inference problem [46,60]. Some insight into spatial properties can be realized by creating thermal maps using machine learning techniques applied to temperature data collected inside processing nodes [61]. Systems that do use fluid and heat models have explored configurations offline to determine the effect of failed components in a single rack [62] or machine room [56].

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 [46] 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 [63]. We believe that both of these guidelines apply, but to different thermal configurations that can only be resolved by sensing and modelling. Moreover, this example shows how the goals of reducing compute power and cooling power conflict at times and motivate the need for unified power management for all components in the data centre.

Modelling 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 centres at runtime saves up to 20 per cent [64], dynamically sizing the allocation of nodes saves up to 20 per cent [65], shaping/scaling workload on each node can save 20 per cent [66], so overall an optimized data centre can be more than 50 per cent [67] more efficient. Because the heating and cooling system is turbulent, small changes can have large effects on energy draw and system performance [56]. Sensing and modelling 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 centres have motivated data centre 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 centre'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 centre sensing introduces. Finally, we present DC Genome, an end-to-end data centre sensing system deployed at a Microsoft data centre.

We thank the anonymous reviewers whose comments helped improve the quality of the paper. We also thank Joe Polastre for his insightful comments during the early stages of this manuscript. Last but not least, we thank our collaborators in the DC Genome project, Chieh-Jan Mike Liang, Suman Nath, Bodhi Priyantha and Feng Zhao.

References

1 US Environmental Protection Agency. 2007 Report to congress on server and data center energy efficiency law 109-431. US Environmental Protection Agency ENERGY STAR program. See http://www.energystar.gov/ia/partners/prod_development/downloads/EPA_Datacenter_Report_Congress_Final1.pdf.

2 O'Reilly, T. 2007 What is web 2.0: design patterns and business models for the next generation of software. *Commun. Strateg. Mag.* **1**, 17–38.

3 Chen, G., He, W., Liu, J., Nath, S., Rigas, L., Xiao, L. & Zhao, F. 2008 Energy-aware server provisioning and load dispatching for connection-intensive internet services. In *Proc. 5th USENIX Symp. on Networked Systems Design and Implementation (NSDI) San Francisco, CA, 16–18 April 2008*, pp. 337–350. Berkeley, CA: USENIX Association.

4 Chase, J., Anderson, D., Thakar, P., Vahdat, A. & Doyle, R. 2001 Managing energy and server resources in hosting centers. In *Proc. ACM Symp. on Operating Systems Principles (SOSP), Banff, Canada, 21–24 October 2001.* New York, NY: ACM.

5 Pitt Turner IV, W. & Brill, K. G. 2008 Tier classifications define site infrastructure performance. Technical Report (white paper). Uptime Institute, NY, New York.

6 Qureshi, A., Weber, R., Balakrishnan, H., Guttag, J. & Maggs, B. 2009 Cutting the electric bill for internet-scale systems. In *Proc. ACM SIGCOMM 2009, Barcelona, Spain, 17–21 August 2009.* New York, NY: ACM.

7 Pinheiro, E., Weber, W.-D. & Barroso, L. A. 2007 Failure trends in a large disk drive population. In *Proc. 5th USENIX Conf. on File and Storage Technologies (FAST 2007), San Jose, CA, 13–16 February 2007.* USENIX Association.

8 Heath, T., Diniz, B., Carrera, E. V., Meira Jr, W. & Bianchini, R. 2005 Energy conservation in heterogeneous server clusters. In *Proc. ACM SIGPLAN Symp. on Principles and Practice of Parallel Programming (PPoPP), Chicago, IL, 15–17 June 2005.* New York, NY: ACM.

9 Doyle, R., Chase, J., Asad, O., Jin, W. & Vahdat, A. 2003 Model-based resource provisioning in a web service utility. In *Proc. 4th USENIX Symp. on Internet Technologies and Systems, Seattle, WA, 26–28 March 2003.* USENIX Association.

10 Armbrust, M. *et al.* 2009 Above the clouds: a Berkeley view of cloud computing. Technical report no. UCB/EECS-2009-28, EECS Department, University of California, Berkeley, CA.

11 Barroso, L. A. & Hölzle, U. 2007 The case for energy-proportional computing. *Comput.* **40**, 33–37. (doi:10.1109/MC.2007.443)

12 Liu, J. 2010 Automatic server to circuit mapping with the red pills. In *Workshop on Power Aware Computing and Systems (HotPower '10), Vancouver, BC, 3 October 2010.* USENIX Association.

13 Chen, K., Guo, C., Wu, H., Yuan, J., Feng, Z., Chen, Y., Lu, S. & Wu, W. Generic and automatic address configuration for data center networks. In *Proc. ACM SIGOMM 2010, New Delhi, India, 30 August–3 September.*

14 Kansal, A., Zhao, F., Liu, J., Kothari, N. & Bhattacharya, A. 2010 Virtual machine power metering and provisioning. In *ACM Symp. on Cloud Computing (SOCC), Indianapolis, IN, 10–11 June 2010.* New York, NY: ACM.
15 IEEE standard for information technology—telecommunications and information exchange between systems local and metropolitan area networks. 2003 Specific requirements: part 15.4: wireless medium access control (MAC) and physical layer (PHY) specifications for low-rate wireless personal area networks (LR-WPANs). See http://www.ieee802.org/15/pub/TG4.html.
16 Texas Instruments. 2006 MSP430x1xx family user's guide (rev. F). See http://www.ti.com/litv/pdf/slau049f.
17 Texas Instruments. 2006 2.4 GHz IEEE 802.15.4/ZigBee-ready RF transceiver. See http://www.chipcon.com/files/CC2420_Data_Sheet_1_3.pdf.
18 Liang, C.-J. M., Liu, J., Luo, L., Terzis, A. & Zhao, F. 2009 RACNet: a high-fidelity data center sensing network. In *Proc. 7th ACM Conf. on Embedded Networked Sensor Systems (Sensys), Berkeley, CA, 4–6 November 2009*, pp. 15–28. New York, NY: ACM.
19 Srinivasan, K., Dutta, P., Tavakoli, A. & Levis, P. 2006 Some implications of low power wireless to IP networking. In *Proc. 5th Workshop on Hot Topics in Networks (HotNets-V), Irvine, CA, 29–30 November 2006.* New York, NY: ACM.
20 Zhao, J. & Govindan, R. 2003 Understanding packet delivery performance in dense wireless sensor networks. In *Proc. 2nd ACM Conf. Embedded Networked Sensor Systems (SenSys 2003), Los Angeles, CA, 5–7 November 2003.* New York, NY: ACM.
21 Srinivasan, K. & Levis, P. 2006 RSSI is under appreciated. In *3rd Workshop on Embedded Networked Sensors (EmNets 2006), Cambridge, MA, 30–31 May 2006.*
22 TinyOS. 2004 MultiHopLQI. See http://www.tinyos.net/tinyos-1.x/tos/lib/MultiHopLQI.
23 Chang, F., Dean, J., Ghemawat, S., Hsieh, W. C., Wallach, D. A., Burrows, M., Chandra, T., Fikes, A. & Gruber, R. E. 2006 Bigtable: a distributed storage system for structured data. In *7th Symp. on Operating System Design and Implementation (OSDI), Seattle, WA, 6–8 November 2006.* USENIX Association.
24 Amazon Web Services LLC. 2011 Amazon SimpleDB. See http://aws.amazon.com/simpledb/.
25 Oracle. 2011 Oracle Berkeley DB 11g. See http://www.oracle.com/technetwork/database/berkeleydb/overview/index.html.
26 FAL Labs. 2011 See http://fallabs.com/tokyocabinet/.
27 Abadi, D. J. *et al.* 2005 The design of the borealis stream processing engine. In *2nd Biennial Conf. on Innovative Data Systems Research (CIDR), Asilomar, CA, 4–7 January 2005.* VLDB Foundation.
28 Microsoft Corporation. 2011 Microsoft StreamInsight. See http://msdn.microsoft.com/en-us/library/ee362541.aspx.
29 IBM Corporation. 2011 InfoSphere Streams. See http://www-01.ibm.com/software/data/infosphere/streams/.
30 Loboz, C., Smyl, S. & Nath, S. 2010 DataGarage: warehousing massive performance data on commodity servers. In *36th Int. Conf. on Very Large Data Bases, Very Large Data Bases (VLDB), Singapore, 13–17 September 2010.* VLDB Endowment.
31 Dean, J. & Ghemawat, S. 2004 MapReduce: simplified data processing on large clusters. In *6th Symp. on Operating System Design and Implementation, San Francisco, CA, 6–8 December 2004.* USENIX Association.
32 OSIsoft. 2011 The PI system. See http://www.osisoft.com/software-support/what-is-pi/What_Is_PI.aspx.
33 Gnawali, O., Fonseca, R., Jamieson, K., Moss, D. & Levis, P. 2009 Collection tree protocol. In *Proc. 7th ACM Conf. on Embedded Networked Sensor Systems (SenSys), Berkeley, CA, 4–6 November 2009*, pp. 1–14. New York, NY: ACM.
34 Reeves, G., Liu, J., Nath, S. & Zhao, F. 2009 Managing massive time series streams with multi-scale compressed trickles. In *VLDB'2009: Proc. 35th Conf. on Very Large Data Bases, Lyon, France, 24–28 August 2009.* VLDB Endowment.
35 Mueen, A., Nath, S. & Liu, J. 2010 Fast approximate correlation for massive time-series data. In *Proc. 2010 ACM SIGMOD Int. Conf. on Management of data (SIGMOD) Indinapolis, IN, 6–11 June 2010.* New York, NY: ACM.

36 Federspiel Controls. 2010 Optimizing data center uptime with dynamic cooling. See http://federspielcorporation.com/downloads/white_papers/white_paper_data_center.pdf.
37 SynapSense Corporation. 2008 LIVEIMAGING: wireless instrumentation solutions. See http://www.synapsense.com/.
38 Arch Rock Corporation. 2010 Arch Rock energy optimizer. See http://www.archrock.com/downloads/datasheet/AREO_DS_web.pdf.
39 Sensicast, Inc. 2009 Sensinet. See http://www.sensicast.com/data_center.php.
40 Dust Networks. 2006 Technical overview of time synchronized mesh protocol (TSMP). See http://www.dustnetworks.com/sites/default/files/TSMP_Whitepaper.pdf.
41 Bash, C., Patel, C. & Sharma, R. 2006 Dynamic thermal management of air cooled data centers. In *The 10th Intersociety Conf. on Thermal and Thermomechanical Phenomena in Electronics Systems* (*ITHERM*)*, San Diego, CA, 30 May–2 June 2006*, pp. 452–461. IEEE.
42 Linden, P. 1999 The fluid mechanics of natural ventilation. *Annu. Rev. Fluid Mech.* **31**, 201–238. (doi:10.1146/annurev.fluid.31.1.201)
43 Riederer, P., Marchio, D., Visie, J., Husaunndee, A. & Lahrech, R. 2002 Room thermal modelling adapted to the test of HVAC control system. *Build. Environ.* **37**, 777–790. (doi:10.1016/S0360-1323(02)00052-5)
44 Ramos, L. & Bianchini, R. 2008 C-Oracle: predictive thermal management for data centers. In *Proc. 14th IEEE Int. Symp. on High Performance Computer Architecture* (*HPCA*)*, Salt Lake City, UT, 16–20 February 2008.* IEEE.
45 Heath, T., Centeno, A. P., George, P., Ramos, L., Jaluria, Y. & Bianchini, R. 2006 Mercury and freon: temperature emulation and management for server systems. In *Proc. ASPLOS, San Jose, CA, 21–25 October 2006.* ACM.
46 Vasic, N., Scherer, T. & Schott, W. 2010 Thermal-aware workload scheduling for energy efficient data centers. In *Proc. Int. Conf. on Autonomic Computing, Washington, DC, 7–11 June 2010.* IEEE/ACM.
47 Patnaik, D., Marwah, M., Sharma, R. & Ramakrishnan, N. 2009 Sustainable operation and management of data center chillers using temporal data mining. In *ACM SIGKDD Conf. on Knowledge Discovery and Data Mining, Paris, France, 28 June–1 July 2009.* ACM.
48 Moore, J., Chase, J. S. & Ranganathan, P. 2006 Weatherman: automated, online and predictive thermal mapping and management for data centers. In *Proc. 2006 IEEE Int. Conf. on Autonomic Computing, Karlsruhe, Germany, 14–18 June 2006.* IEEE.
49 Tang, Q., Gupta, S. K. S. & Varsamopoulos, G. 2008 Energy-efficient thermal-aware task scheduling for homogeneous high-performance computing data centers: a cyber-physical approach. *IEEE Trans. Parallel Distrib. Syst.* **19**, 1458–1472. (doi:10.1109/TPDS.2008.111)
50 Patel, C., Bash, C. E., Belady, C., Stahl, L. & Sullivan, D. 2001 Computational fluid dynamics modeling of high compute density data centers to assure system inlet air specifications. In *Proc. IPACK, Kauai, Hawaii, 8–13 July 2001.* ASME.
51 Rambo, J. & Joshi, Y. 2005 Reduced order modeling of steady turbulent flows using the POD. In *Proc. ASME-HT05–ASME Summer Heat Transfer Conf., San Francisco, CA, 17–22 July 2005.* ASME.
52 Rambo, J. & Joshi, Y. 2007 Modeling of data center airflow and heat transfer: state of the art and future trends. *Distrib. Parallel Databases* **21**, 193–225. (doi:10.1007/s10619-006-7007-3)
53 Shah, A., Carey, V., Bash, C. & Patel, C. 2008 Exergy analysis of data center thermal management systems. *J. Heat Transf.* **130**, 1–10. (doi:10.1115/1.2787024)
54 Samadiani, E., Joshi, Y. & Mistree, F. 2008 The thermal design of a next generation data center: a conceptual exposition. *J. Electron. Packag.* **130**, 93–102. (doi:10.1115/1.2787024)
55 Bhopte, S., Agonafer, D., Schmidt, R. & Sammakia, B. 2006 Optimization of data center room layout to minimize rack inlet air temperature. *J. Electron. Packag.* **128**, 380–387. (doi:10.1115/1.2356866)
56 Beitelmal, A. H. & Patel, C. D. 2007 Thermo-fluids provisioning of a high performance high density data center. *Distrib. Parallel Databases* **21**, 227–238. (doi:10.1007/s10619-005-0413-0)
57 Moore, J., Chase, J., Ranganathan, P. & Sharma, R. 2005 Making scheduling 'cool': temperature-aware workload placement in data centers. In *USENIX Annual Technical Conf, Anaheim, CA, 10–15 April 2005.* USENIX Association.

58 Tang, Q., Mukherjee, T., Gupta, S. K. S. & Cayton, P. 2006 Sensor-based fast thermal evaluation model for energy efficient high-performance datacenters. In *Proc. 4th Int. Conf. Intelligent Sensing and Information Processing, Bangalore, India, 15–18 December 2006.* IEEE.

59 Wang, L., von Laszewski, G., Dayal, J., He, X., Younge, A. J. & Furlani, T. R. 2009 Towards thermal aware workload scheduling in a data center. In *Proc. Symp. on Pervasive Systems, Algorithms and Networks, Kaohsiung, Taiwan, 14–16 December 2009.* IEEE.

60 Pakbaznia, E., Ghasemazar, M. & Pedram, M. 2010 Temperature-aware dynamic resource provisioning in a power-optimized datacenter. In *Proc. Design, Automation, and Test in Europe, Dresden, Germany, 8–12 March 2010.* European Design and Automation Association.

61 Moore, J., Chase, J. & Ragananthan, P. 2006 Consil: low-cost thermal mapping of data centers. In *Workshop on Tackling Computer Systems Problems with Machine Learning Techniques, Saint-Malo, France 27 June 2006.*

62 Choi, J., Kim, Y., Sivasubramaniam, A., Srebric, J., Wang, Q. & Lee, J. 2008 A CFD-based tool for studying temperature in rack-mounted servers. *IEEE Trans. Comput.* **57**, 1129–1142. (doi:10.1109/TC.2008.52)

63 Karavanic, K. 2009 Scalable methods for performance and power data collection and analysis. In *Los Alamos Computer Science Symp, Santa Fe, New Mexico, 13–14 October 2009.* Los Alamos National Laboratory.

64 Parolini, L., Tolia, N., Sinopoli, B. & Krogh, B. H. 2010 A cyber-physical systems approach to energy management in data centers. In *Proc. Int. Conf. on Cyber Physical Systems, Stockholm, Sweden, 13–14 April 2010.* ACM/IEEE.

65 Leverich, J. & Kozyrakis, C. 2009 On the energy (in)efficiency of Hadoop clusters. In *Workshop on Power Aware Computing and Systems* (*HotPower '09*), *Big Sky, MT, 10 October 2009.* ACM.

66 Freeh, V. W., Kappiah, N., Lowenthal, D. K. & Bletsch, T. K. 2008 Just-in-time dynamic voltage scaling: exploiting inter-node slack to save energy in MPI programs. *J. Parallel Distrib. Comput.* **68**, 1175–1185. (doi:10.1016/j.jpdc.2008.04.007)

67 Pelley, S., Meisner, D., Wenisch, T. F. & VanGilder, J. W. 2009 Understanding and abstracting total data center power. In *Proc. Workshop on Energy Efficient Design, Austin, TX, 20 June 2009.*