Building and Environment 82 (2014) 151–159

Contents lists available at ScienceDirect

Building and Environment

journal homepage: www.elsevier.com/locate/buildenv

Assessing the environmental impact of data centres part 1: Background, energy use and metrics

Beth Whitehead [a, *, 1], Deborah Andrews [a], Amip Shah [b], Graeme Maidment [a]

[a] Faculty of Engineering Science and The Built Environment, London South Bank University, 103 Borough Road, London, SE1 0AA, UK
[b] HP Labs, 1501 Page Mill Rd, Palo Alto, CA, 94304, USA

ARTICLE INFO

Article history:
Received 15 May 2014
Received in revised form
13 August 2014
Accepted 14 August 2014
Available online 23 August 2014

Keywords:
Data centres
Environmental impact
Carbon footprint
Life cycle assessment
Data centre metrics
Energy consumption

ABSTRACT

Data centres consume high levels of energy to power the IT equipment contained within them, and extract the heat they produce. Because of the industry's heavy reliance on power, data centre metrics have historically used operational efficiency as a proxy for sustainability. More recently the industry has begun to recognise that its focus needs to go beyond energy consumption, with the creation of metrics for issues such as carbon, water and compute efficiency. However, single-issue metrics often consider only the operational phase, omitting impacts from other issues, during other stages in a facility's lifetime. Further approaches exist to assess more holistically the impact of data centres, such as building environmental assessment methods, but none have the capacity to capture fully the interlinked nature of a system, where improvements in one area and to one impact, can adversely affect a totally different area and totally different impacts.

The following review of literature summarises the approach of the data centre industry to environmental impact, and provides direction for future research. Part 1 describes the energy consumption of the ICT industry and in particular data centres; current knowledge on the environmental impact of the industry; and how single-issue metrics have risen to prominence.

1. Introduction

Data centres contain IT equipment used for the processing and storage of data, and communications networking [1]. They are the backbone of IT networks across the globe [2,3] and include extensive supporting infrastructures required to power and cool the IT equipment. A data centre can be as simple as a single rack in a server closet or as complex as a large warehouse, typically having built-in redundancy for the avoidance of downtime.

Data centres are high energy consumers. In 2007 the ICT industry was estimated to account for 10% of total UK electricity consumption [4], and 2% of global anthropogenic CO_2 [5], approximately equal to the direct emissions of the aviation industry operation. The operation of data centres already accounts for around a quarter of these emissions [4], and is believed to have the fastest growing carbon footprint from across the whole ICT sector [5].

This energy consumption has drawn the attention of data centre owners and operators. Firstly because of the cost of energy bills, and more recently because of it's impact on the environment. However, exclusive consideration of energy consumption has meant that other impacts and stages in a data centres life cycle are not well understood.

This two-part literature review seeks to present the current energy consumption and environmental impact of the data centre industry, and how it is monitored, assessed and benchmarked, and concludes the need for a more holistic approach to the management of environmental impact in the future. It does not seek to establish ways in which to reduce the impact. The review aims to focus the industry on why it has approached environmental issues in the current manner, highlight the need for a change in approach, and suggest further research and work required to enable this. Part 1 describes the energy consumption of the ICT industry and in particular data centres; current knowledge on the environmental impact of the industry; how the industry benefits the environment; and how single-issue metrics

* Corresponding author. Tel.: +44 (0) 7739 019960.
E-mail addresses: bethfwhitehead@gmail.com, BethWhitehead@dc-oi.com (B. Whitehead).
[1] Permanent address: Operational Intelligence, 74 Kelvedon Close, Kingston upon Thames, Surrey, KT2 5LF, UK.

http://dx.doi.org/10.1016/j.buildenv.2014.08.021
0360-1323/

have risen to prominence. Part 2 builds on this foundation to describe the use of building environmental assessment methods and tools; and based on both parts of the review, concludes the need to apply life cycle thinking to assess the environmental impact of data centres.

2. Data centres

Data centres house servers, and networking and storage equipment, and are considered the central nervous system of the 21st century. They contain comprehensive mechanical and electrical infrastructures to support the energy intensive computing required to perform one or more of the following functions [6,7]:

- The physical housing of IT equipment such as computers, servers, switches, routers, data storage devices, racks, and related equipment.
- The storage, management, processing and exchange of digital data.
- The provision of application services or management for data processing, such as web hosting, internet, intranet and telecommunication.

Data centres vary in size from a single rack in a server closet to huge server farms with floor areas reaching 150,000 m^2. Some occupy floors within offices and others are steel sheds on dedicated sites like that shown in Fig. 1. Large facilities contain data halls as shown in Fig. 2, which contain racks of IT equipment; the remaining floor space houses power and cooling equipment. Typically, the extensive floor space required for the supporting infrastructures can be as much as two [9] to four (when there are no external services) [10] times greater than the data halls themselves. Tight controls on air quality mean that the data halls do not include windows, and in the UK are often built using a steel frame and concrete floor construction resulting in large, windowless boxes. They are high energy consumers, both for power and the extraction of the heat dissipated from the IT equipment, and although some have huge floor plates, they are incredibly low occupancy facilities.

Data centres are used by businesses, corporations, educational establishments and governments, to provide web hosting and the internet, the storage of company information, and the processing of business transactions. They can be public (accessible to all, such as those for Google searches) or private (for the storage of company information on network drives) and, based on the importance of continued access to the data, display varying levels of [12]:

Fig. 1. Facebook data centre [8].

Fig. 2. Inside a data hall [11].

- Reliability – probability that a component/system/data centre operates without failure over a set time period. Facilities can have the same availability, but a facility that has one outage per year is more reliable than a facility that has many failures lasting the same amount of time.
- Availability – the average time per time period (for example a year) that a component/system/data centre operates as designed, without downtime. For example 0.99999 availability is a facility that has a total yearly downtime of 315 s.
- Redundancy – the topology of supporting infrastructures that ensure a component/system/data centre remains available in the event of a failure.

Facilities are described using Tier classifications [13] – Tier I to Tier IV – which refer to the topology of the facility's supporting infrastructures (power and cooling), and reflect how the building performs under planned and unplanned outages. The ability of a data centre to continue to perform its function in the event of a problem is determined by the amount of redundancy (spare plant) built into the design. For example, two mains power feeds to a site would ensure continued operation if one feed is lost, because operations can be switched to the other. The amount of redundancy incorporated into a data centre is dependent on whether or not the business linked to the facility can continue relatively unharmed in the event of a fault.

The Tiers were established by the Uptime Institute (an industry research body), to provide a common language across which the availability and reliability (redundancy) of different facilities can be compared, and are described in Table 1.

3. ICT and the internet

Worldwide, the number of data centres is growing, in part, due to the increase in access to PCs (personal computers) and the internet. The global PC installed base (including laptops) is well documented, and has grown rapidly from 242 million in 1995 [15,16] to 592 million in 2002 [5] and 1 billion in 2009 [17]. Furthermore, projections up to 2004 [15,16] 2014 [17] and 2020 [5], and shown in Fig. 3, fit with a pattern of exponential growth suggested by these early figures; most of which will require access to the internet and networks supported by data centres [17].

Furthermore, at the end of 2012, 34.3% of the global population were internet users [18], a penetration that grew from less than 1% in 1995 [19], as shown in Fig. 4, and equated to a rise from 0.04% to 6% in less developed countries [15]. Between 2000 and 2012,