The Electricity Journal 29 (2016) 61–69

Contents lists available at ScienceDirect

The Electricity Journal

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

Power usage effectiveness in data centers: Overloaded and underachieving

Nathaniel Horner*, Inês Azevedo

Department of Engineering and Public Policy, Carnegie Mellon University, Baker Hall 129, Pittsburgh, PA 15213, USA

ARTICLE INFO

Article history:
Available online 7 June 2016

Keywords:
Data center
Energy metric
PUE
Information technology

ABSTRACT

The power usage effectiveness (PUE) metric has become an industry standard for reporting energy performance of data centers. However, it is an incomplete metric, failing to address hardware efficiency, energy productivity, and environmental performance. The industry should focus on adopting and systematically reporting more comprehensive metrics, which would allow more insight into data center energy performance.

1. Introduction

The advent of the digital computer brought about the “Information Age,” an era in which information has come into its own as a valuable commodity. The accuracy, relevance, and timeliness of an organization’s information are—as they have always been—keys to its success. However, the higher speeds, greater traffic, and increased access on the “information superhighway” have made firms hungrier for ever-increasing volumes of data.

Data services lie at the heart of operations for many companies and constitute a core product for others. Google indexes the web to provide search services to users while simultaneously collecting information about search activity to deliver advertising. Social networking entities such as Facebook collect personal data in exchange for hosting virtual communities, aiding interaction among groups. Online retailers like Amazon.com house online inventories and bring buyers and sellers together. Data service providers such as Dropbox and Apple allow users to store their documents, files, and digital content “in the cloud,” a distributed storage network. Even traditional retailers like Walmart and Whole Foods use data-intensive processes to manage their inventories in real time.

The hardware and software used to store, transmit, and utilize data to provide e-services are collectively known as information technology (IT) or information and communications technology (ICT). ICT includes computers and software, mobile devices, and communication networks and their components. As digital content has proliferated, so too have the storage mechanisms grown, moving from the lone server to the server closet, the server room, and now the server farm. These storage repositories are collectively known as data centers, which not only provide static storage but also dynamically provide a wide variety of services including hosting web pages and email, streaming multimedia, and running complex business applications like banking management software. Individuals, private firms, universities, and government entities all use data centers of varying scale and complexity to manage the digital information they need to operate.

Expansion of data and services has meant exponential growth in needed storage capacity. IBM reports that global daily data production is 2.5 quintillion bytes and that 90% of the world’s data has been produced in the previous two years (IBM, 2013). In addition to the magnitude of data produced, increasing complexity of software has increased its size. As the prevalence of data centers has grown, so have public concerns about their aggregate energy consumption (Markoff and Hansell, 2006). A series of bottom-up estimates has generally found that data centers use on the order of 1–2% of U.S. and global electricity consumption (Koomey, 2008, 2011). An industry report examining the greenhouse gas (GHG) emissions attributable to ICT estimates that, while data center carbon emissions are a negligible part of the global total, they have grown at 8.6% annually since 2001 and will continue to outpace both the global footprint growth rate and the that of other ICT subcategories (networks and end user devices) through 2020 (BCG, 2012).

In response to these concerns, and also to get a better handle on operational costs, the industry has worked to establish metrics to assess data center performance. The most prevalent of these is

* Corresponding author.
E-mail address: nch@cmu.edu (N. Horner).

http://dx.doi.org/10.1016/j.tej.2016.04.011
1040-6190/© 2016 Elsevier Inc. All rights reserved.

power usage effectiveness (PUE), which is the ratio of facility-wide power consumption to power used by the IT equipment. This article discusses the main problems associated with this metric. We first provide a brief background on data center types, as different metrics may be more or less relevant to different sizes and applications of data centers (Section 2). Section 3 defines PUE and highlights its main drawbacks as an overall measure of data center energy performance. Alternative metrics are discussed in Section 4, while Section 5 briefly summarizes different ways in which data center energy performance—once assessed with an appropriate metric—can be improved. Finally, Section 6 summarizes the discussion.

2. Background: data center taxonomy

What is a data center? Lawrence Berkeley National Laboratory (LBNL) describes a data center as a special-purpose facility with the following characteristics and functions (LBNL, 2013):

- “Houses various equipment, such as computers, servers (e.g., web servers, application servers, database servers), switches routers, data storage devices, load balancers, wire cages or closets, vaults, racks, and related equipment.
- Store[s], manage[s], processe[s], and exchange[s] digital data and information;
- Provide[s] application services or management for various data processing, such as web hosting internet, intranet, telecommunication and information technology.”

While the LBNL definition is focused on facility contents and function, ICT consulting firm Gartner uses a definition more focused on the organizational role the data center plays, defining a data center as “the department of an enterprise that houses and maintains the back-end [IT] systems and data stores—its mainframes, servers, and databases. In the days of large, centralized IT operations, this department and all the systems resided in one physical space.” (Gartner, 2013)

These definitions make no statements about size, and include few restrictions on function. Data centers can range from small server closets to huge server farms and can host a variety of IT services, such as corporate email and filesystems, data archives, and cloud services. The main criterion for a data center seems to be that it houses “back-end” ICT equipment—equipment accessed indirectly by users via a network.

The variety of size and function can make clear classification difficult, though most classification systems rely on some combination of size, criticality of service, and service type. We adopt the following nomenclature for context, loosely based on the taxonomy used by IDC (Patterson, 2010):

1. **Server closets**, or “ad-hoc” data centers, support small businesses or individual projects at larger companies. They may get some support from a corporate-level IT department but may also be configured and operated by non-experts.
2. **Server rooms** are small data centers that support small businesses or special groups or projects of larger entities. They may be administered by central IT staff or “owned” by each project or division.
3. **Localized data centers** provide business-critical applications and have some power and cooling redundancy, though downtime is not catastrophic. Restoration of service on the order of hours is acceptable.
4. **Mid-tier data centers** are large-to-medium-size data centers used to host enterprise-wide applications in support of operations or human resources (e-mail accounts, filesystems, internal data). The data is critical, but incidental to the primary business line. Downtime lasting longer than a few minutes has significant impact on the business. These facilities are operated by the company’s central IT department.
5. **Enterprise data centers** are large facilities used, usually by non-ITC companies, in support of core business operations (e.g., banks, health care companies, etc.). These data centers are often in special-purpose facilities and operated under a separate business unit or division. Downtime is catastrophic, and these facilities have highly redundant infrastructure.
6. **Hyperscale data centers, server farms, or warehouse-scale computers (WSCs)** are the very large data centers, usually constructed in their own physical plants, built by ICT companies with a primary business line focused on data (e.g., Google, Apple, Facebook, Amazon, et al.) and, increasingly, cloud-based services. Barroso and Hölzle (2009) coined the term warehouse-scale computer to emphasize the distinguishing large economies of scale, extreme parallelism, hardware and software homogeneity, and aggressive focus on efficiency of these data centers.

Generally, size, infrastructure redundancy, quality of service, and criticality increase as one moves down the list, though these distinctions are necessarily qualitative and somewhat fuzzy in nature. Note that these data center types can be deployed in very different domains, ranging from corporate entities to university- and research-based enterprises.

The Uptime Institute, a corporate data center consultancy, has published a data center classification focused on infrastructure redundancy, ranging from “basic” to “fault-tolerant” (Turner et al., 2008; C7 Data Centers, 2012). Tier I data centers have no redundant systems, whereas Tier IV facilities have duplicate active power and cooling distribution paths, with redundant components on each, so that the center can withstand any single equipment failure. The tiers roughly follow the functional/size-based classification above, with Tier I & II data centers being appropriate for businesses that have no or low obligated quality-of-service requirements and Tier III and IV data centers being appropriate for businesses that need to deliver round-the-clock services with serious consequences for downtime.

When analyzing the energy performance of a data center, it can be difficult to draw a clear boundary around the system, particularly if it is part of a larger multiuse building, as is often the case. Generally the data center encompasses the computing load and associated equipment in such cases, not the entire building. However, if cooling, lighting, and HVAC systems are shared between the data center and spaces dedicated to other uses (e.g., offices), which is especially common in the smallest data centers, then it can be hard to accurately assess energy performance. Inconsistency in what is included in the measurements can make comparisons among different data centers difficult.

3. Measuring energy efficiency: PUE

Estimates of aggregate energy usage provide context for determining if data center energy consumption is a cause for concern. However, these sorts of studies cannot reveal *how* that energy is being used—that is, how efficiently do data centers use energy to deliver the services they provide? To assess energy performance, we must first define a suitable metric.

There have been many different metrics proposed by various industry and research organizations. One review paper cited no less than 30 metrics proposed by various organizations to measure different aspects of data center energy efficiency (Jamalzadeh and Behravan, 2011). However, the industry has converged on power