CONTRIBUTED
P A P E R

Green Cloud Computing: Balancing Energy in Processing, Storage, and Transport

For processing large amounts of data, management and switching of communications may contribute significantly to energy consumption and cloud computing seems to be an alternative to office-based computing.

By Jayant Baliga, Robert W. A. Ayre, Kerry Hinton, and Rodney S. Tucker, *Fellow IEEE*

ABSTRACT | Network-based cloud computing is rapidly expanding as an alternative to conventional office-based computing. As cloud computing becomes more widespread, the energy consumption of the network and computing resources that underpin the cloud will grow. This is happening at a time when there is increasing attention being paid to the need to manage energy consumption across the entire information and communications technology (ICT) sector. While data center energy use has received much attention recently, there has been less attention paid to the energy consumption of the transmission and switching networks that are key to connecting users to the cloud. In this paper, we present an analysis of energy consumption in cloud computing. The analysis considers both public and private clouds, and includes energy consumption in switching and transmission as well as data processing and data storage. We show that energy consumption in transport and switching can be a significant percentage of total energy consumption in cloud computing. Cloud computing can enable more energy-efficient use of computing power, especially when the computing tasks are of low intensity or infrequent. However, under some circumstances cloud computing can consume more energy than conventional computing where each user performs all computing on their own personal computer (PC).

KEYWORDS | Cloud computing; core networks; data centers; energy consumption

I. INTRODUCTION

The increasing availability of high-speed Internet and corporate IP connections is enabling the delivery of new network-based services [1]. While Internet-based mail services have been operating for many years, service offerings have recently expanded to include network-based storage and network-based computing. These new services are being offered both to corporate and individual end users [2], [3]. Services of this type have been generically called "cloud computing" services [2]–[7].

The cloud computing service model involves the provision, by a service provider, of large pools of high-performance computing resources and high-capacity storage devices that are shared among end users as required [8]–[10]. There are many cloud service models, but generally, end users subscribing to the service have their data hosted by the service, and have computing resources allocated on demand from the pool. The service provider's offering may also extend to the software applications required by the end user. To be successful, the cloud service model also requires a high-speed network to provide connection between the end user and the service provider's infrastructure.

Cloud computing potentially offers an overall financial benefit, in that end users share a large, centrally managed pool of storage and computing resources, rather than owning and managing their own systems [5]. Often using existing data centers as a basis, cloud service providers invest in the necessary infrastructure and management

Manuscript received November 26, 2009; accepted July 3, 2010. Date of publication August 30, 2010; date of current version December 17, 2010. This work was supported by the Australian Research Council and by Cisco Systems.
The authors are with the Department of Electrical and Electronic Engineering, University of Melbourne, Melbourne, Vic. 3010, Australia (e-mail: jbaliga@gmail.com; r.ayre@ee.unimelb.edu.au; k.hinton@unimelb.edu.au; r.tucker@unimelb.edu.au).
Digital Object Identifier: 10.1109/JPROC.2010.2060451

systems, and in return receive a time-based or usage-based fee from end users [6], [8]. Since at any one time, substantial numbers of end users are inactive, the service provider reaps the benefits of the economies of scale and from statistical multiplexing, and receives a regular income stream from the investment by means of service subscriptions [6]. The end user in turn sees convenience benefits from having data and services available from any location, from having data backups centrally managed, from the availability of increased capacity when needed, and from usage-based charging [2], [3]. The last point is important for many users in that it averts the need for a large one-off investment in hardware, sized to suit maximum demand, and requiring upgrading every few years [5].

There are many definitions of cloud computing, and discussion within the IT industry continues over the possible services that will be offered in the future [8], [10]. The broad scope of cloud computing is succinctly summarized in [11]:

> *Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources that can be rapidly provisioned and released with minimal management effort or service provider interaction.*

Cloud computing architectures can be either public or private [8], [9]. A private cloud is hosted within an enterprise, behind its firewall, and intended only to be used by that enterprise [8]. In such cases, the enterprise invests in and manages its own cloud infrastructure, but gains benefits from pooling a smaller number of centrally maintained high-performance computing and storage resources instead of deploying large numbers of lower performance systems. Further benefits flow from the centralized maintenance of software packages, data backups, and balancing the volume of user demands across multiple servers or multiple data center sites. In contrast, a public cloud is hosted on the Internet and designed to be used by any user with an Internet connection to provide a similar range of capabilities and services [8]. A number of organizations are already hosting and/or offering cloud computing services. Examples include Google Docs [12], Amazon's Elastic Compute Cloud and Simple Storage services [13], Microsoft's Windows Azure Platform [14], IBM's Smart Business Services [15], Salesforce.com [16], and Webex [17].

But while its financial benefits have been widely discussed, the shift in energy usage in a cloud computing model has received little attention. Through the use of large shared servers and storage units, cloud computing can offer energy savings in the provision of computing and storage services, particularly if the end user migrates toward the use of a computer or a terminal of lower capability and lower energy consumption. At the same time, cloud computing leads to increases in network traffic and the associated network energy consumption. In this paper, we explore the balance between server energy consumption, network energy consumption, and end-user energy consumption, to present a fuller assessment of the benefits of cloud computing.

The issue of energy consumption in information technology equipment has been receiving increasing attention in recent years and there is growing recognition of the need to manage energy consumption across the entire information and communications technology (ICT) sector [18]–[20]. It is estimated that data centers accounted for approximately 1.2% of total United States electricity consumption in 2005 [20]. The transmission and switching networks in the Internet account for another 0.4% of total electricity consumption in broadband-enabled countries [21]. In addition to the obvious need to reduce the greenhouse impact of the ICT sector [4], [19]–[22], this need to reduce energy consumption is also driven by the engineering challenges and cost of managing the power consumption of large data centers and associated cooling [23], [24]. Against this, cloud computing will involve increasing size and capacity of data centers and of networks, but if properly managed, cloud computing can potentially lead to overall energy savings.

The management of power consumption in data centers has led to a number of substantial improvements in energy efficiency [25], [26]. Cloud computing infrastructure is housed in data centers and has benefited significantly from these advances. Techniques such as, for example, sleep scheduling and virtualization of computing resources in cloud computing data centers improve the energy efficiency of cloud computing [24].

While it is important to understand how to minimize energy consumption in data centers that host cloud computing services, it is also important to consider the energy required to transport data to and from the end user and the energy consumed by the end-user interface. Previous studies of energy consumption in cloud computing [4], [24], [27] have focused only on the energy consumed in the data center. However, to obtain a clear picture of the total energy consumption of a cloud computing service, and understand the potential role of cloud computing to provide energy savings, a more comprehensive analysis is required.

In this paper, we present an overview of energy consumption in cloud computing and compare this to energy consumption in conventional computing. For this comparison, the energy consumption of conventional computing is the energy consumed when the same task is carried out on a standard consumer personal computer (PC) that is connected to the Internet but does not utilize cloud computing. We consider both public and private clouds and include energy consumption in switching and transmission, as well as data processing and data storage. Specifically, we present a network-based model of the switching and transmission network [21], [28], [29], a model of user

computing equipment, and a model of the processing and storage functions in data centers [7], [30], [31]. We examine a variety of cloud computing service scenarios in terms of energy efficiency. In essence, our approach is to view cloud computing as an analog of a classical supply chain logistics problem, which considers the energy consumption or cost of processing, storing, and transporting physical items. The difference is that in our case, the items are bits of data. As with classical logistics modeling, our analysis allows a variety of scenarios to be analyzed and optimized according to specified objectives.

We explore a number of practical examples in which users/customers outsource their computing and storage needs to a public cloud or private cloud [8], [9]. Three cloud computing services are considered, including *storage as a service* [3]–[6], [8], *processing as a service* [2]–[6], [8], and *software as a service* [2]–[4], [6], [8]. As the name implies, storage as a service allows users to store data in the cloud. Processing as a service gives users the ability to outsource selected computationally intensive tasks to the cloud. Software as a service combines these two services and allows users to outsource all their computing to the cloud and use only a very-low-processing-power terminal at home.

We show that energy consumption in transport and switching can be a significant percentage of total energy consumption in cloud computing. Cloud computing can enable more energy-efficient use of computing power, especially when the users' predominant computing tasks are of low intensity or arise infrequently. However, we show that under some circumstances cloud computing can consume more energy than conventional computing on a local PC. Our broad conclusion is that cloud computing can offer significant energy savings through techniques such as virtualization and consolidation of servers [25], [32] and advanced cooling systems [26], [33]. However, cloud computing is not always the greenest computing technology.

II. CLOUD SERVICE MODELS

We focus our attention on three cloud services—storage as a service, processing as a service and software as a service. In the following sections, we outline the functionality of each of the three cloud services. Note that we use the terms "client," "user," and "customer" interchangeably.

A. Software as a Service

Consumer software is traditionally purchased with a fixed upfront payment for a license and a copy of the software on appropriate media. This software license typically only permits the user to install the software on one computer. When a major update is applied to the software and a new version is released, users are required to make a further payment to use the new version of the software. Users can continue to use an older version, but once a new version of software has been released, support for older versions is often significantly reduced and updates are infrequent.

With the ubiquitous availability of broadband Internet, software developers are increasingly moving towards providing software as a service [2]–[4], [6]. In this service, clients are charged a monthly or yearly fee for access to the latest version of software [2], [3]. Additionally, the software is hosted in the cloud and all computation is performed in the cloud. The client's PC is only used to transmit commands and receive results. Typically, users are free to use any computer connected to the Internet. However, at any time, only a fixed number of instances of the software are permitted to be running per user. One example of software as a service is Google Docs [12].

When a user exclusively uses network- or Internet-based software services, the concept is similar to a "thin client" model, where each user's client computer functions primarily as a network terminal, performing input, output, and display tasks, while data are stored and processed on a central server. Thin clients were popular in office environments prior to the widespread use of PCs.

In Section IV, we explore the opportunity for reduced energy consumption in the client's PC when we only use software services. In this scenario, data storage and processing is always performed in the cloud and we are thus able to significantly reduce the functionality, and consequently, the power consumption, of the client's PC.

B. Storage as a Service

Through storage as a service, users can outsource their data storage requirements to the cloud [3]–[6]. All processing is performed on the user's PC, which may have only a solid state drive (e.g., flash-based solid-state storage), and the user's primary data storage is in the cloud. Data files may include documents, photographs, or videos. Files stored in the cloud can be accessed from any computer with an Internet connection at any time [5]. However, to make a modification to a file, it must first be downloaded, edited using the user's PC and then the modified file uploaded back to the cloud. The cloud service provider ensures there is sufficient free space in the cloud and also manages the backup of data [5]. In addition, after a user uploads a file to the cloud, the user can grant read and/or modification privileges to other users. One example of storage as a service is the Amazon Simple Storage service [13].

C. Processing as a Service

Processing as a service provides users with the resources of a powerful server for specific large computational tasks [2]–[6]. The majority of tasks, which are not computationally demanding, are carried out on the user's PC. More demanding computing tasks are uploaded to the cloud, processed in the cloud, and the results are

Table 1 Summary of Cloud Services

	Software as a Service	Storage as a Service	Processing as a Service
Location of Processing	Cloud	Client	Short tasks at client, large tasks in cloud
Location of Storage	Cloud	Cloud	Client
Function of Transport	Transmit commands and receive results	All files/documents	Files for large tasks

returned to the user [6]. Similar to the storage service, the processing service can be accessed from any computer connected to the Internet. One example of processing as a service is the Amazon Elastic Compute Cloud service [13].

When utilizing a processing service, the user's PC still performs many small tasks and is consequently required to be more powerful than the "thin client" considered in the software service (Section II-A). However, the user's computer is not used for large computationally intensive tasks and so there is scope to reduce its cost and energy consumption, relative to a standard consumer PC, by using a less powerful computer.

D. Summary of Models

Table 1 provides a summary of the location of processing, location of storage, and function of transport for each of these cloud services. In a storage service, the majority of processing occurs at the user's PC (the client) and the majority of storage is in the cloud. The transmission and switching network transports the user's files between the data center and the user. With a processing service, the user's computer processes only short tasks and the cloud processes large computationally intensive tasks. Long-term storage of data is on the user's computer and transport is required to transfer the files relevant to each large task. In a software service, processing and storage are performed in the cloud. Transport is required for all tasks to enable transmission of commands to the cloud and to return the results.

III. MODELS OF ENERGY CONSUMPTION

In this section, we describe the functionality and energy consumption of the transport and computing equipment on which current cloud computing services typically operate. We consider energy consumption models of the transport network, the data center, plus a range of customer-owned terminals and computers. The models described are based on power consumption measurements and published specifications of representative equipment [7], [21], [22], [30]. Those models include descriptions of the common energy-saving techniques employed by cloud computing service providers.

The models are used to calculate the energy consumption per bit for transport and processing, and the power consumption per bit for storage. The energy per bit and power per bit are fundamental measures of energy consumption, and the energy efficiency of cloud computing is the energy consumed per bit of data processed through cloud computing. Performing calculations in terms of energy per bit also allows the results to be easily scaled to any usage level.

We consider both public and private clouds. Fig. 1 shows schematics of a public cloud computing network [Fig. 1(a)] and a private cloud computing network [Fig. 1(b)]. For the public cloud, the schematic includes the data center as well as access, metro and edge, and core networks. The private cloud schematic includes the data center as well as a corporate network. These schematics form the basis for the analysis in the following sections of this paper. From a hardware perspective, the key difference between public cloud computing and private cloud computing is the network connecting the users to the respective data center. As described earlier, a data center for a public cloud is hosted on the Internet and designed to be used by anyone with an Internet connection.

Public cloud users are typically residential users and connect to the public Internet through an Internet service provider's (ISP) network. Looking forward, it is expected that the access portion of such networks will increasingly use passive optical network (PON) technologies, which are particularly energy efficient [34]. Within the ISP's network, Ethernet switches aggregate user traffic, broadband network gateways (BNGs) regulate access and usage, and provider edge routers form the gateway to the global Internet, which comprises many large core routers and high-capacity transport networks.

Data centers in turn connect to the core network through their own gateway router. A typical data center comprises a gateway router, a local area network, servers, and storage [7], [30]. As shown in Fig. 1(a), the BNG routers, provider edge routers, and the data center gateway routers typically dual-home to more than one core router, in order to achieve higher service availability through network redundancy. Although only a single data center is shown, a cloud service provider would normally maintain several centers with dedicated transport between these centers for redundancy and efficient load balancing.

Private clouds, as described earlier, are intended only to be used by the enterprise that owns the private cloud. In the center of Fig. 1(b) is a schematic of a corporate network connecting users, who are shown on the left, to a data

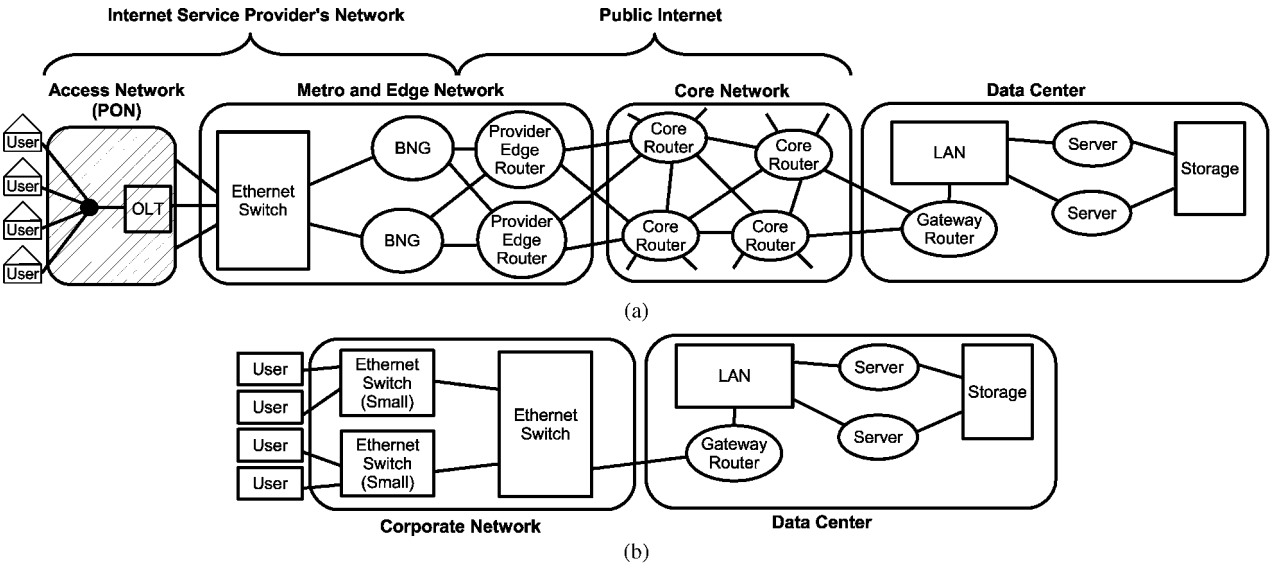

Fig. 1. *Schematic of networks connecting users to a cloud and the data center infrastructure used to host those clouds. (a) Public cloud. (b) Private cloud.*

center hosting a private cloud, which is shown on the right. Each user connects to a small Ethernet switch, which connects to one or more larger Ethernet switches to make up a private core network. The data center, which is similar to the one used for public cloud computing, is typically connected directly to this large Ethernet switch. Similar to a public cloud service, typically multiple data centers would be deployed for redundancy.

A key factor in the calculation of energy consumption of switching and data centers is the energy consumed in cooling and other overheads [35]. The power usage effectiveness (PUE) is the ratio of the total power consumption of a facility (data or switching center) to the total power consumption of IT equipment (servers, storage, routers, etc.) [35]. A 2003 benchmark of 15 data centers found that the average PUE was approximately 1.93 [36]. A second more recent benchmark of nine data centers, performed in 2005, found an improved PUE of 1.63 [36]. Some data centers have since achieved even higher efficiency, with Google reporting that one of their data center's was operating with a PUE of 1.15 [33]. In the present analysis, we use a PUE of 1.5.

In Sections III-A–C, we describe the functionality and energy consumption of the user, network, and data center equipment in greater detail.

A. User Equipment

A user may use a range of devices to access a cloud computing service, including a mobile phone (cell phone), desktop computer, or a laptop computer. In this paper, we focus on desktop computers and laptops. These computers typically comprise a central processing unit (CPU), random access memory (RAM), hard disk drive (HDD), graphical processing unit (GPU), motherboard, and a power supply unit. Peripheral devices including speakers, printers, and visual display devices are often connected to PCs. These peripheral devices do not influence the comparison between conventional computing and cloud computing and so are not included in the model. In our analysis, we assume that when user equipment is not being used it is either switched off or in a deep sleep state (negligible power consumption).

Table 2 lists a range of commonly used classes of computers that users may use for personal computing and/or to

Table 2 Hardware in Model of User Equipment

		Power Consumption	
Equipment	Parts	Idle	Full Load
Modern mid-range computer	Intel E5200 @ 2.5 GHz, 2 GB RAM	70 W	110 W
Old mid-range computer	Intel Pentium 4 @ 2.86 GHz, 1 GB RAM	125 W	210 W
Modern high-end computer	Intel Q6600 @ 2.4 GHz, 4 GB of RAM	110 W	175 W
Low-end laptop	Intel Atom @ 1.6 GHz, 2 GB RAM	18 W	18 W
Terminal	-	8 W	8 W
2.5" HDD	Western Digital Scorpio Blue	0.25 W	2.5 W

access cloud computing services. Table 2 also lists the power consumption of a modern 2.5″ HDD. Power consumption data for this equipment was complied through measurements of the current drawn when each computer was idle and also under full load. The midrange description of the older computer refers to its classification at the time of its release. The computational capacity of the midrange older computer is significantly lower than that of the modern midrange computer.

A terminal typically only accepts simple commands from users and communicates with a server via an IP network. The server returns raw or lightly compressed video data for display. The terminal relies on the server for all processing. Based on the power consumption of a consumer network switch and new low-end laptops, we estimate that the terminal consumes 8 W.

B. Data Centers

A modern state-of-the-art data center has three main components—data storage, servers, and a local area network (LAN) [7], [30], [31]. The data center connects to the rest of the network through a gateway router, as shown on the right-hand side of Fig. 1(a) and (b) [7], [30]. Table 3 lists equipment typical of that used in data centers, as well as the capacity and power consumption of this equipment. Power consumption figures for the LAN switches, routers, and storage equipment are the figures quoted in their respective product data sheets. The power consumption data for each server was obtained by first calculating the maximum power using HP's power calculator [39], then following the convention that average power use for midrange/high-end servers is 66% of maximum power [20]. In the following, we outline the functionality of this equipment as well as some of the efficiency improvements in cloud computing data centers over traditional data centers.

Long-term storage of data in a data center is provided by hard disk arrays, together with associated equipment. Hard disk arrays include supporting functionality such as cache memories, disk array controllers, disk enclosures, and redundant power supplies. In a cloud computing data center, all the storage space in the data center is consolidated and hard disk usage is centrally coordinated [9], [42]. Consolidation and central coordination minimizes the total number of hard disks used, greatly increasing the overall energy efficiency of storage. In addition, files that are not accessed regularly are stored in a different set of capacity optimized hard disks [43]. These hard disks enter a low-power mode when not in use and consume negligible energy. To reflect these gains in energy efficiency, our analysis attributes storage power only to those files that are being regularly accessed.

While infrequently used data files are stored on a disk, the data rate and latency of disk read operations is generally inadequate for services such as file hosting, which entail frequent accesses to the file. Data for these services are cached in RAM on one or more servers. Additional servers perform data center management and, in a high-performance computing facility, provide on-demand computing. The server performance depends on the computational characteristics of the task being performed, including the number of floating-point operations, memory accesses, and suitability for parallel processing.

Through server virtualization/consolidation, a very large number of users can share a single server, which increases utilization and in turn reduces the total number of servers required [25], [32], [44]. Users do not have or need any knowledge of the tasks being performed by other users and utilize the server as though they are the only user on the server. During periods of low demand, some of the servers enter a sleep mode which reduces energy consumption [26]. To reflect the efficiency gains from sleeping, virtualization, and consolidation, in our analysis, the computation servers and file hosting servers are fully utilized.

A LAN inside the data center aggregates the traffic from the servers into higher capacity (typically 10 GE) links and connects to the network core through a gateway router [7], [30], [31]. The LAN today is typically a three-tier/layer aggregation network with Ethernet switches at both layers, however data centers are moving towards two-tier aggregation networks [7], [30]. In our analysis, we consider a two-tier aggregation network.

C. Network

In this section, we describe the corporate and Internet IP networks in greater detail and outline the functionality of the equipment in those networks. Table 4 lists equipment used in our calculations of energy consumption in the corporate network and the Internet IP network as well as the capacity and power consumption of this equipment.

Table 3 Equipment in Model of Data Centers

	Equipment	Capacity	Power Consumption
Storage	HP 8100 EVA	604.8 Tb [37]	4.9 kW [37]
Content Server	HP DL380 G5	800 Mb/s [38]	225 W [39]
Computation Server	HP DL380 G5	-	355 W [39]
LAN	Cisco 6509	320 Gb/s [40]	3.8 kW [40]
Gateway Router	Juniper MX-960	660 Gb/s [41]	5.1 kW [41]

Table 4 Equipment in Model of Network

	Equipment	Capacity	Power Consumption
Ethernet Switch (Small)	Cisco 4503	64 Gb/s [40]	474 W [40]
Ethernet Switch	Cisco 6509	160 Gb/s [40]	3.8 kW [40]
BNG	Juniper E320	60 Gb/s [41]	3.3 kW [41]
Provider Edge	Cisco 12816	160 Gb/s [40]	4.21 kW [40]
Core router	Cisco CRS-1	640 Gb/s [40]	10.9 kW [40]
WDM (800 km)	Fujitsu 7700	40 Gb/s [45]	136 W/channel [21]

1) Corporate Network: The corporate network comprises several Ethernet switches interconnected in a hierarchical configuration, as shown on the left-hand side of Fig. 1(b). A small Ethernet switch at the lower layer might aggregate traffic on one building floor, and several higher layer switches aggregate traffic from buildings or campuses.

The energy E_C required to transport one bit from the data center to a user through a corporate network is

$$E_C = 3 \times 3 \times \left(\frac{P_{les}}{C_{les}} + \frac{3P_{es}}{C_{es}} + \frac{P_g}{C_g}\right) \quad (1)$$

where P_{es}, P_{les}, and P_g are the powers consumed by the Ethernet switches, small Ethernet switches, and data center gateway routers, respectively. C_{es}, C_{les}, and C_g are the capacities of the corresponding equipment in bits per second. The left most factor of three accounts for the power requirements for redundancy (factor of 2) as well as cooling and other overheads (factor of 1.5). The typical average utilization of Ethernet links in LANs is less than 5% [46]. However, a private cloud would significantly increase network traffic and so in our model we assume an average utilization of 33%. The second factor of three in (1) is to account for this underutilization of corporate networks. The factor of three for Ethernet switches is to include the Ethernet switches in the corporate LAN as well as the Ethernet switches in the LAN inside the data center.

Using power consumption figures for representative commercial equipment, given in Table 4, we estimate the per-bit energy consumption of transmission and switching for a private cloud to be around 0.46 μJ/b.

2) Internet: The access network is modeled as a PON [47]. The energy consumption of the access network is largely independent of traffic volume [34]. Thus, the access network does not influence the comparison between conventional computing and cloud computing. Therefore, it is omitted from consideration and is not included in our calculations of energy consumption. Table 4 lists the equipment used in our model of the IP network as well as the capacity and power consumption of this equipment [21]. These values were obtained from manufacturer's data sheets [40], [45]. In the following, we outline the functionality of this equipment.

On the network side, the access network typically connects to an Ethernet aggregation switch, which is the entry point to the metro and edge network, as shown in Fig. 1(a). The Ethernet switches perform traffic aggregation and connect to two or more BNG routers, which perform traffic management and authentication functions. The minimum of two uplinks is for redundancy, and in this model, we include redundancy for all network elements on the network side of the edge Ethernet switch. The BNG routers connect to provider edge routers, which groom and encapsulate the IP packets into a packet over SONET/SDH (PoS/SDH) format for transmission to the network core. The core network typically comprises a small number of large routers. These core routers perform all the necessary routing and also serve as the gateway to neighboring core routers.

High-capacity wavelength division multiplexed (WDM) fiber links interconnect core routers. WDM fiber links also connect edge routers to core router. Edge routers are presumed to be located within 80 km of a core router and so do not require additional WDM transponder systems. We model a core network with major core routers in cities an average of 800 km apart. In this topology, the 800-km link requires seven intermediate line amplifiers and two terminal system for all 44 optical channels. Each optical channel operates at 40 Gb/s.

The energy E_I required to transport one bit from a data center to a user through the Internet is

$$E_I = 6\left(\frac{3P_{es}}{C_{es}} + \frac{P_{bg}}{C_{bg}} + \frac{P_g}{C_g} + \frac{2P_{pe}}{C_{pe}} + \frac{2\times 9P_c}{C_c} + \frac{8P_w}{2C_w}\right) \quad (2)$$

where P_{es}, P_{bg}, P_g, P_{pe}, P_c, and P_w are the powers consumed by the Ethernet switches, broadband gateway routers, data center gateway routers, provider edge routers, core routers, and WDM transport equipment, respectively. C_{es}, C_{bg}, C_g, C_{pe}, C_c, and C_w are the capacities of the corresponding equipment in bits per second. The power consumption and capacities of this equipment is given in Table 4. The factor of six accounts for the power requirements for redundancy (factor of 2), cooling and other overheads (factor of 1.5), and the fact that today's network typically operate at under 50% utilization [48] while still consuming almost 100% of maximum power [49] (factor

of 2). The factor of three for Ethernet switches is to include the Ethernet switches in the metro network as well as the Ethernet switches in the LAN inside the data center. The factor of two for provider edge routers is to include the edge router in the edge network and the gateway router in the data center. The factor of two for core routers allows for the fact that core routers are usually provisioned for future growth of double the current demand [50]. In today's Internet, packets traverse an average of 12–14 hops between source and destination [51]. In the model we consider, customer traffic must traverse three hops to reach the network core, two hops from the network core to the server, and a further average of eight core hops, which leads to an average of 13 hops in total. The factor of nine for routers and factor of eight for WDM transport equipment account for the eight core hops. However, we also halve the number of hops for core WDM transport equipment because many of the core hops are between co-located switches or routers and so WDM transport is not used.

Using power consumption figures for representative commercial equipment, given in Table 4, we estimate the per-bit energy consumption of transmission and switching for a public cloud to be around 2.7 μJ/b [21].

IV. ANALYSIS OF CLOUD SERVICES

In this section, we compare the per-user energy consumption of each cloud service outlined in Section II using the energy model described in Section III. The energy consumption of each cloud service is also compared against the energy consumption of conventional computing.

As described earlier, the key difference between public cloud computing and private cloud computing is the transport network connecting users to the data center. In the following, E_T is the per-bit energy consumption of transport in cloud computing. If we are considering a private cloud model, $E_T = E_C$ (transport through a corporate network), and if we are considering a public cloud model, $E_T = E_I$ (transport through the Internet).

A. Storage as a Service

In this section, we analyze the energy consumption of storage as a service. We consider, as an example, a file storage and backup service, where all processing and computation is performed on the user's computer but user data are stored in the cloud. Files are downloaded from the cloud for viewing and editing and then uploaded back to the cloud for storage. The per-user power consumption of the storage service P_{st}, calculated as a function of downloads per file per hour, is

$$P_{st} = \frac{B_d D}{3600}\left(E_T + \frac{1.5P_{st,SR}}{C_{st,SR}}\right) + 2B_d\frac{1.5P_{SD}}{B_{SD}} \qquad (3)$$

where B_d (bits) is the average size of a file and D is the number of downloads per hour. $P_{st,SR}$ is the power consumption of each content server and $C_{st,SR}$ (bits per second) is the capacity of each content server. The power consumption of hard disk arrays (cloud storage) is P_{SD} and their capacity is B_{SD} (bits). Power consumption and capacity of content servers, and hard disk arrays, is described in Section III-B. The per-bit energy consumption of transmission and switching is E_T. The division by 3600 converts hours to seconds, the multiplication by a factor of 2 in the third term accounts for the power requirements for redundancy in storage, and the multiplication by a factor of 1.5 in the second and third terms accounts for the power requirements for cooling as well as other overheads. As outlined earlier, in our model, we assume that only files that are regularly accessed consume energy when stored. Files that are not accessed regularly are stored in other disk drives that sit at a low-power mode and consume negligible power.

Fig. 2 is a plot of the percentage of total power consumed in transport, storage, and servers/computation, as a function of number of downloads per hour, for a public cloud storage service. Fig. 3 presents equivalent results for a private cloud storage service. Note that the results presented in Figs. 2 and 3 are applicable for all file sizes. The file size is independent of distribution of energy consumption between storage, servers, and transport, which can be seen from (3).

The number of times a file is downloaded per hour would depend on the nature of the file. A word processing document or spreadsheet might be required a few times per hour, but photograph downloads might take place many times per hour. At a low download rate of 10^{-2}/h, for the public cloud storage service, approximately 75% of power is consumed in storage (principally in the hard disk

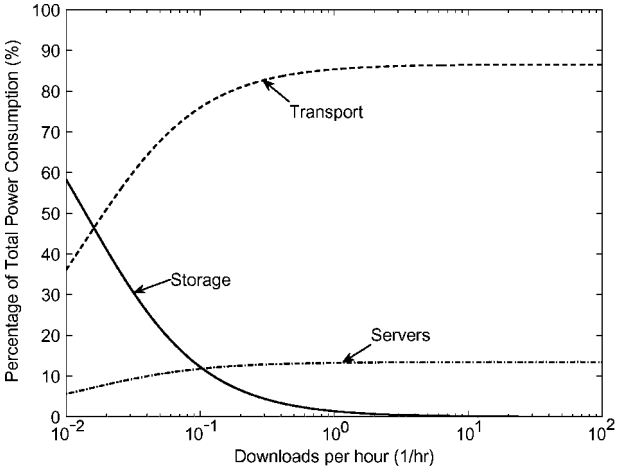

Fig. 2. *Percentage of total power consumption of transport, storage, and servers of a public cloud storage service as a function of download rate.*

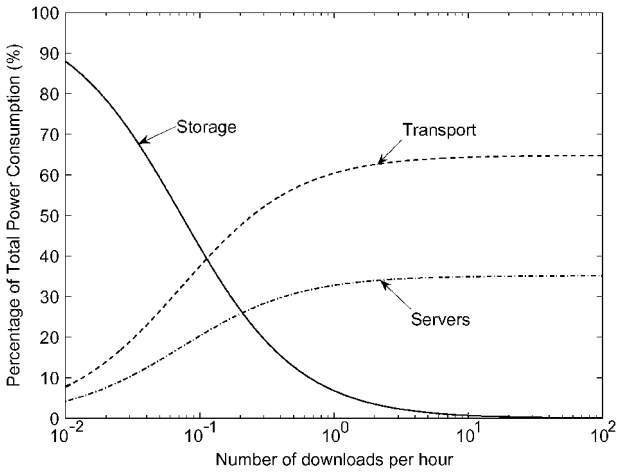

Fig. 3. ***Percentage of total power consumption of transport, storage, and servers of a private cloud storage service as a function of download rate.***

arrays), approximately 25% is consumed in transport, and the remainder is consumed by servers (including data caches). At the same download rate, for a private cloud storage service, approximately 90% of power is consumed in storage, approximately 10% is consumed in transport, and the remainder is consumed by servers. Thus, power consumption in storage dominates total power consumption for both the public and private cloud storage services at low usage levels. Archiving infrequently used data on to capacity optimized HDDs is a useful tool to minimize this energy consumption in storage. In addition, these capacity optimized HDDs could be spun down and put into a sleep/idle state to further reduce energy consumption.

As the average download rate increases, an increased number of servers, routers, and switches are required to support this additional traffic. Storage requirements are independent of the download rate. Thus, as the average download rate increases, the percentage of total power consumed in servers and transport increases, while the percentage of total power consumed in storage decreases.

At more than one download per hour for a public cloud storage service, servers consume approximately 10%, storage consumes less than 1%, and the remaining power is consumed in transport. For a private cloud storage service, at a download rates above one download per hour, servers consume 35%, storage consumes less than 7%, and the remaining 58% of total power is consumed in transport. These results indicate that transport dominates total power consumption at high usage levels for public and private cloud storage services. The energy consumed in transporting data between users and the cloud is therefore an important consideration when designing an energy-efficient cloud storage service. Energy consumption in servers is also an important consideration at high usage levels. The percentage of total power consumed in servers is greater in private cloud computing than that in public cloud computing. In both public and private cloud storage services, the energy consumption of storage hardware is a small percentage of total power consumption at medium and high usage levels.

We now consider the total per-user power consumption of a storage service by scaling the per-file power consumption P_{st} by the average number of files in use by each user. The per-user power consumption of a storage service with F files per user is FP_{st}. Fig. 4 shows the total per-user power consumption of a private cloud storage service and a public cloud storage service as a function of download rate, when the number of active files per user is 20. Here the download rate is the number of downloads per hour, per user, per file. The average file size is 1.25 MB. The power consumption of the storage services is below 1 W at low download rates (< one download per hour per file). As the download rate increases, due primarily to the increased power consumption in transport, the power consumption of the storage services increases towards 10 W. The power consumption of the public cloud storage service is about 2.5 times that of the power consumption of the private cloud storage service, at medium and high download rates, due primarily to the increased energy consumption in transport. Note that the results presented here can easily be linearly scaled by size of the files. For example, a storage service storing on average two active files per user, where each file's size is 12.5 MB, is equivalent to a storage service storing on average ten active files per user, where each file's size is 1.25 MB. This equivalence can be seen from (3).

It is interesting to compare the energy consumed by a cloud storage service to an HDD in a home computer. Included in Fig. 4 is the power consumption of a modern

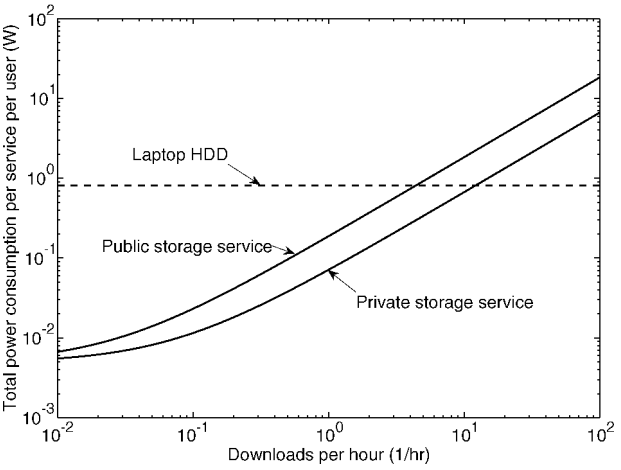

Fig. 4. ***Per-user power consumption of public and private cloud storage services as a function of download rate. Also included is the power consumption of a modern laptop HDD. The average document size is 1.25 MB.***

laptop HDD (2.5″ HDD) that is idle (low-power state) 75% of the time and active 25% of the time. Comparing the power consumption of the laptop HDD and the storage service, it is clear that at low download rates, the storage service is more efficient, but this benefit vanishes if the number of regularly used files is larger, and if downloaded frequently. However, we note that because the per-user savings are less than 1 W, there are bigger opportunities elsewhere for large energy savings through a cloud service.

B. Software as a Service

Users access a software service (sometimes referred to as virtual desktop infrastructure) via a terminal ("dumb client" computer) that communicates with its server via simple commands transmitted through the Internet. The server in turn transmits video data to the terminal that is output on a monitor. As mentioned in Section III-A, the power consumption of the visual display unit, speakers, and peripheral devices is not included in the model as they would be common to all alternative configurations. All data processing is performed at a remote server. The per-user power consumption P_{sf} of the software service, including the terminal, as a function of the bit rate A (bits per second) between each user and server is

$$P_{sf} = P_{sf,PC} + \frac{1.5P_{sf,SR}}{N_{sf,SR}} + 2B_d\frac{1.5P_{SD}}{B_{SD}} + AE_T \qquad (4)$$

where $P_{sf,PC}$ is the power consumption of the user's terminal, $P_{sf,SR}$ is the power consumption of the server, P_{SD} is the power consumption of the hard disk arrays, $N_{sf,SR}$ is the number of users per server, and B_{SD} is the capacity of the hard disk arrays. As with the storage service, the multiplication by a factor of 2 in the third term accounts for the power requirements for redundancy in storage and the multiplication by a factor of 1.5 for data center equipment (second and third terms) accounts for the energy consumption in cooling as well as other overheads.

Each user has a monitor running at a resolution of 1280×1024 with 24-b color, giving a total of $1280 \times 1024 \times 24$ b/frame. If Y is the number of new frames every second (frames/s), the data rate between each user and the server is $A = 1280 \times 1024 \times 24 \times Y$ b/s. At a refresh rate of 1 frame/s, the server must deliver ~31.5 Mb/s to the user. We calculate the power consumption of the cloud service in terms of the frames per second capacity of the network, henceforth referred to as "frames per second" or "frame rate." We consider frames per second capacity because power consumption of transport networks is determined by capacity and not usage. Note that if 100% of the user's screen changes every second, this corresponds to one frame per second. If only a small percentage of the user's screen is changing, then only a portion of a frame is transmitted and the frame rate falls below one frame per second.

The responsiveness of such a system to inputs from end users depends both on job queuing delays and network latency. Queuing delay depends on the computational intensity of the users' tasks, the memory/disk access requirements of the task, the task frequency, the number of users sharing a server, and the performance of the server. Network latency is controlled by ensuring the network is not congested and limiting the distance between the server and the end user. We model a data center with computation servers (described in Section III-B) and consider two scenarios. In the first scenario, each server is able to support 20 users and in the second scenario each server is able to support 200 users. The utilization of the Internet is 50% and the utilization of the corporate network is 33%, which is sufficiently low to minimize latency. Our analysis assumes that servers are sufficiently close (geographically) to ensure that propagation delay is small.

Fig. 5 is a plot of the percentage of total power consumption of each component (storage, transport, servers) of a public cloud software service as a function of the frame rate. The percentage of power consumed in the user terminal is not shown. On average 10 GB of data is stored in the cloud per user. The plot includes curves for 20 users per server and 200 users per server. We have assumed that the power consumption of the server is determined solely by the number of users per server and that increasing the frame rate has a negligible effect on the server. Note that with 200 users per server, the curves stop at 0.11 frames/s because the maximum transmission capacity of each server is 800 Mb/s [38].

If the software service only requires frame rates below 10^{-2}, less than 10% of total power is consumed in

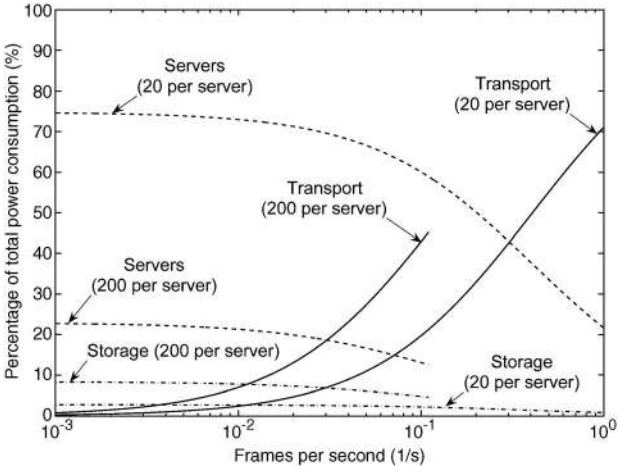

Fig. 5. *Percentage of total power consumption of transport, storage, and servers of a public cloud storage service as a function of download rate with 20 and 200 users per server. The percentage of total power consumed by the user terminal is not shown.*

transport. As the frame rate increases, the percentage of power consumed in transport significantly increases. At 0.1 frames/s, transport consumes 20% of total power with 20 users per server. At the same frame rate, with 200 users per server, transport consumes 42% of total power. Cloud storage consumes less than 15% of total power at all frame rates.

Fig. 6 is a plot of the percentage of total power consumed in each of transport, storage, and servers/processing, as a function of the frame rate, for a private cloud software service. As with the public cloud software service, on average, 10 GB of data is stored in the cloud per user and the percentage of power consumed in the user terminal is not shown in Fig. 6. The plot includes curves for 20 users per server and 200 users per server. With 200 users per server, transport, storage, and servers together consume less than half of total power consumption. The remaining power is consumed in the terminal. With 20 users per server, at frame rates less than 0.1 frames/s, transport consumes less than 5% of total power, increasing to 40% of total power consumption as frame rates increase to 1 frame equivalents per second. With 20 users per server, the majority of power is consumed in the servers.

Fig. 7 is a plot of the total per-user power consumption of the public and private software services, including the terminal, as a function of frames per second. The power consumption of the cloud services with 20 users per server is 35–45 W when the frame rate is small (< 0.1 frames/s). If the transport component of the public cloud service is required to support the equivalent of 1 frame/s, the power consumption of the service rises to 129 W due to the high transport requirements. The power consumption of the private cloud service with 20 users per server does not exceed 60 W even at high frame rates. The lower power consumption of the private software service is due to the lesser transport infrastructure involved. The power consumption of the cloud services with 200 users per server is 12–23 W. The power consumption in transport increases as the frame rate increases, but the transmission rate limit of each server of 800 Mb/s limits the frame rate to 0.11 frames/s.

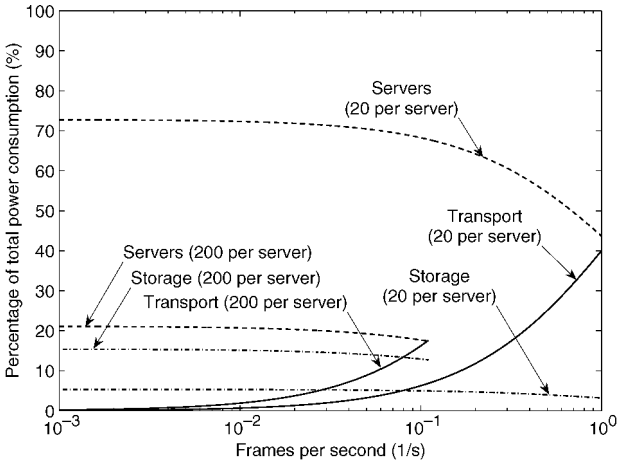

Fig. 6. *Percentage of total power consumption of transport, storage, and servers of a private cloud storage service as a function of download rate with 20 and 200 users per server. The percentage of total power consumed by the user terminal is not shown.*

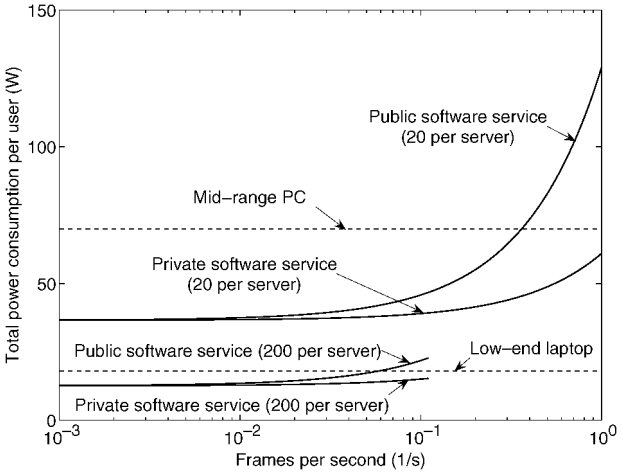

Fig. 7. *Per-user power consumption of public and private cloud software services as a function of download rate. Also included is the power consumption of a low-end laptop and the power consumption at idle of a modern midrange computer.*

Included in Fig. 7 is the power consumption at idle of a low-end laptop (18 W) and the power consumption at idle of a modern midrange computer (70 W). A low-end laptop consumes the least power but also has the least functionality and processing capacity. The cloud service in both scenarios is more efficient than the modern midrange PC at low frame rates. However, as the frame rate increases, the power consumption of the public cloud service with 20 users per server approaches and then exceeds the power consumption of the midrange PC. Users utilizing a software service consume up to approximately 35–55 W less than users with a midrange PC, when the frame rate is low and the number of users per server is high. As the frame rate increases or the number of users per server decreases, the energy savings diminish.

The number of users per server is the most significant determinant of the energy efficiency of a cloud software service. Cloud software services are more efficient than modern midrange PCs for simple office tasks, where the number of users per server can be high. However, if the user's tasks are intensive and high frame rates are required, then public software services are not energy efficient relative to a modern midrange PC. Due to the low transport energy consumption with private software services, even intensive computing tasks with high frame rates are more

energy efficient than midrange PCs. Our results show that corporations should strongly consider private software services instead of standalone PCs to reduce energy consumption.

C. Processing as a Service

We model processing as a service with each user having a low-end laptop that is used for routine tasks and compare the energy consumption with the use of a higher capacity desktop machine. In the cloud, there are computation servers that are used for computationally intensive tasks. Data for computationally intensive tasks are uploaded to a cloud service, and the completed output is returned to the user. As an example of a computationally intensive task, we model the task of converting and compressing a video file. We calculate the per-week energy consumption of the processing service as a function of the number of encodings per week N. The per-user energy consumption (watt hours) E_{ps} of the processing service, including the user's PC, is

$$E_{ps} = 40P_{ps,PC} + 1.5NT_{ps,SR}P_{ps,SR} + 168AE_T \qquad (5)$$

where $P_{ps,PC}$ is the power consumption of the user's laptop, $P_{ps,SR}$ is the power consumption of the server, and $T_{ps,SR}$ is the average number of hours it takes to perform one encoding. The user's PC is used on average 40 h/week for common office tasks (factor of 40 in first term). A factor of 1.5 is included in the second term to account for the energy consumed to cool the computation servers, as well as other overheads. In the third term, A is the per-user data rate (bits per second) between each user and the cloud, E_T is the per-bit energy consumption of transport, and the factor of 168 converts power consumption in transport to energy consumption per week (watt hours).

We take as a model for a demanding task the processing of a 2.5-h DVD-sized video stored in MPEG-2 (8.54 GB) and encoding it into the H.264 (MPEG-4 Part 10) format [52]. Re-encoding this 2.5-h MPEG-2 video file into H.264 takes 1.25 h on the computation server [53]. If on average N encodings are performed each week, the average data rate between each user and the data center is $N \times 8 \times 8.54 \times 10^9/144\,000$ b/s, where the factor of 8 converts bytes to bits and 8.54×10^9 is the size of the video file. We assume that users submit jobs to the processing service sometime during the week while they are using the computer for office tasks (144 000 is the number of seconds in 40 h).

Fig. 8 is a plot of the percentage of total power consumption of transport and servers in a public cloud processing service as a function of the number of times a user performs such encodings each week. Fig. 9 is a plot of the percentage of total power consumption of transport and servers in a private cloud processing service as a function of encodings per week. In both cases, the total power consumption includes the power consumed in the user's laptop and the percentage of total power consumed in the user's laptop is shown in both figures. These figures and the subsequent figures intentionally extend to improbably high numbers of video program encodings per week to show the effect of applications requiring substantial computation and input/output resources on the energy performance of a cloud service.

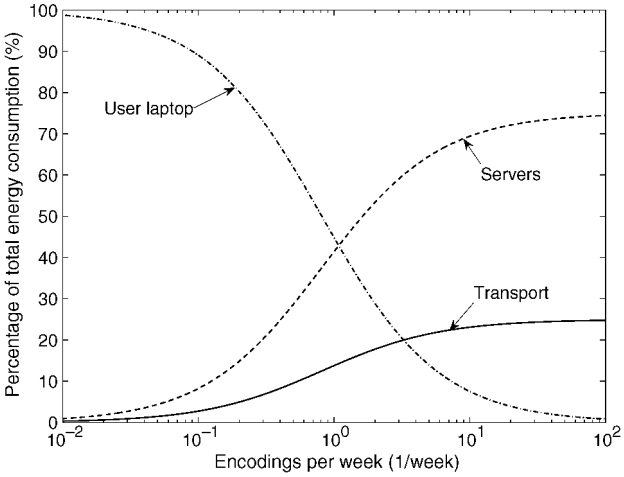

Fig. 8. *Percentage of total power consumption of transport, storage, and servers of a public cloud processing service as a function of encodings per week.*

At a fewer than 10^{-1} encodings per week over 90% of power is consumed in the user's laptop for both the public and private cloud processing services. As the number of such encoding per week increases, the energy

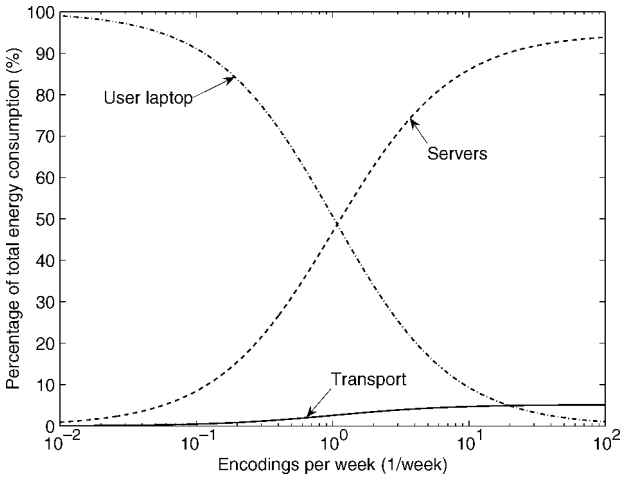

Fig. 9. *Percentage of total power consumption of transport, storage, and servers of a private cloud processing service as a function of encodings per week.*

consumption in transport and processing increases. The user's laptop is modeled as being used for 40 h/week regardless of the number of encodings and so its energy consumption remains constant as the number of encodings increases. At one encoding per week with a public cloud processing service, approximately 40% of total energy is consumed in servers, approximately 15% of total energy is consumed in transport, and the remainder is consumed in the user laptop. For a private cloud processing service with one encoding per week, half of the total energy is consumed in the user laptop, approximately 50% of total energy is consumed in servers, and the remainder is consumed in transport. The trend of increased energy consumption in servers and transport continues as the number of encodings per week increases.

The results indicate that in a public cloud processing service, even for the computationally intensive task of video encoding, transport consumes a significant percentage of total energy consumption at medium and high usage rates. However, the percentage of energy consumed in transport with a private cloud processing service is less than 5% at all usage rates.

Fig. 10 is a plot of the per-user per-week total energy consumed with public and private cloud processing services, as a function of the number of video encoding per week. The total energy consumption of both cloud processing services includes the energy consumed in the user's low-end laptop. The user's low-end laptop consumes 0.72 kWh/week when used for 40 h to perform common office tasks.

At less than 10^{-1} encodings per week, total energy consumption with the public and private cloud processing services is similar. At this usage level, the total energy consumption is dominated by the energy consumption in the user's laptop, as seen in Figs. 8 and 9. If on average one encoding is performed each week, the total energy consumption with the public processing service is 1.6 kWh/week. At the same number of encodings per week, the total energy consumption of the private processing service is 13% lower at 1.4 kWh/week due to the lower energy consumption in transport. The total energy consumption of the public cloud processing service increases to 10 kWh/week at ten encodings per week and approaches 100 kWh at the extreme of 100 encodings per week. The private cloud processing service is approximately 21% lower, again due to the lower energy consumption in transport.

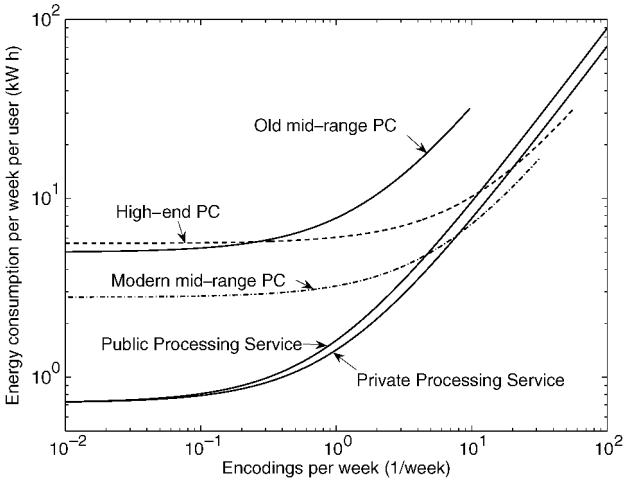

Fig. 10. ***Power consumption of public and private cloud processing services as a function of encodings per week. Also included is the power consumption of a modern midrange PC and the power consumption of a modern idle of a high-end PC.***

We compare these figures for the energy consumption of a low-end laptop supplemented by a processing service with the energy consumed in performing the same set of tasks on a consumer PC. We have measured the power consumption and processing time taken by a range of household and office computers to process such a 2.5-h video file (8.54 GB) and encode it into the H.264 (MPEG-4 Part 10) format. Thus, in Fig. 10, we include the per-week energy consumption when the user has an old midrange PC, a modern midrange PC, or a high-end PC, and processes the video file locally. In these three scenarios, the user's PC is used for 40 h/week for everyday office tasks in addition to the number of hours required to perform the relevant number of video encodings. Our measurements found that to encode a typical DVD video file requires 13.2 h on the old midrange PC, 4 h on the modern midrange PC, and 2.2 h on the high-end PC. Therefore, a maximum of 9.7, 32, and 58 encodings can be performed on the old midrange PC, modern midrange PC, and high-end PC, respectively. Performing 40 h of common office tasks consumes 5, 2.8, and 5.6 kWh/week on the old midrange PC, modern midrange PC, and high-end PC, respectively.

The results in Fig. 10 indicate that a cloud processing service is always more energy efficient than the model older midrange PC. Thus, users with older generation computing equipment could achieve significant energy savings as well as increased computational capability by moving to a combination of a modern low-end laptop and cloud computing. Choosing a modern low-end laptop would also realize upfront cost savings because a low-end laptop is significantly cheaper than a modern midrange PC. Note that the results presented in Fig. 10 are equivalent to a user processing a proportionately larger number of smaller files per week instead of a few large files.

If a user is performing fewer than the equivalent of four such video encodings per week, the most energy efficient option is the combination of a low-end computer and a processing service. At greater than four encodings per week for the public cloud processing service and eight encodings per week for the private cloud processing service, the energy consumption of servers and transport increases to the point that a modern midrange PC is the

most energy efficient option. It is important to note that the computation server is a virtualized server instance running on a very powerful computer system. The virtualized server matches or betters the midrange PC in capability but is less energy efficient. The computation server needs to be computationally powerful to ensure that the total time to encode with a cloud service (transport and encode) is similar to that with a midrange PC (encode only). If a more energy-efficient computer were used in the cloud then the private cloud processing service would have an efficiency similar to that of a modern midrange PC even at high usage rates. However, the transport energy component means that a public cloud service will generally be less efficient than conventional (local) computing for heavy users. Thus, a fundamental consideration when designing an energy-efficient cloud service is the energy consumption of data transport between users and the cloud.

We cannot assume that cloud processing is always more energy efficient than processing through conventional home computing. Cloud computing is only more efficient if the energy consumed in data transport is compensated by savings in the energy consumption of the cloud computation servers (after allowing for management overheads and PUE), and/or by power savings in the home user's computer. If the user performs such computationally intensive tasks only occasionally, a strategy of using a lower capability computer such as a low-end laptop, together with outsourcing the occasional computationally intensive tasks to a cloud service, will deliver savings in energy use as well as cost. However, if the energy savings of a user's computing equipment is negligible or the energy consumed in transport is excessive, cloud processing is less energy efficient than processing through conventional computing.

D. Summary of Results

Table 5 provides a summary of conditions under which energy consumption is significant in transport, storage, and processing for both public and private cloud services. Transport presents a more significant energy cost in public cloud services than in private cloud services. The energy consumption in processing is significant when the nature and frequency of processing tasks dictate that there be fewer users per server. Processing as a service does not involve long-term storage in the cloud.

V. THE FUTURE OF CLOUD COMPUTING

The analysis in previous sections was based on state-of-the-art technology in 2010. In recent years, there have been continuous improvements in the energy efficiency of equipment as new generations of technology come on line. This has led to exponential improvements over time in the energy efficiency of servers [54], storage equipment [55] as well as routers and switches [22], [56], [57]. It is reasonable to expect that future generations of transport and computing equipment will continue to achieve improvements in terms of energy efficiency, largely due to improvements in complementary metal–oxide–semiconductor (CMOS) integrated circuit technology. In this section, we utilize estimates of efficiency gains in technology over time to forecast energy consumption of cloud computing in the future. We also discuss future directions for cloud computing and provide guidelines for how cloud computing can be made as energy efficient as possible.

A. Forecasts of Equipment Energy Consumption

In a commercial environment, especially a data center, many factors dictate the technology in use. Prime objectives are to maximize the delivery of services and hence revenue, at the same time minimizing the costs of support and maintenance, rack space, head load, and power consumption. It is common practice to periodically replace lower performing or high maintenance equipment with state-of-the-art equipment. User equipment in contrast tends to be retained for longer periods and its evolution in the medium-term future is difficult to predict. Our forecasts focus on the energy consumption of the network, servers, and storage and do not consider future generations of user equipment.

Using an exponential model of efficiency improvement [21], [22], [56], if a current piece of state-of-the-art equipment has capacity C_0 and has power consumption P_0, then in t years, a comparable piece of state-of-the-art equipment will have an energy consumption E_Q given by

$$E_Q(t) = \frac{P_Q}{C_Q} = \frac{P_0}{C_0}(1-\alpha)^t \tag{6}$$

Table 5 Conditions Under Which Energy Consumption Is Significant

Energy Component	Service type	Software as a Service	Storage as a Service	Processing as a Service
Transport	Public	High frame rates	Always	Medium to high encodings per week
	Private	Never	High download rates	Never
Storage	Public	Never	Low download rates	-
	Private	Never	Low download rates	-
Processing	Public	Few users per server	Never	Medium to high encodings per week
	Private	Few users per server	High download rates	Medium to high encodings per week

where P_Q is the power consumption in t years, C_Q is the capacity in$t$ years, and α is the annual rate of improvement of state-of-the-art technology. The units for capacity C_0 and C_Q depend on the piece of equipment being considered. The capacity of routers is measured in bits per second, the capacity of storage is measured in bits, the capacity of content servers is measured in terms of transmission capacity (bits per second), and the capacity of computation servers is measured in terms of processing capacity. In the following sections, we provide estimates of α for each class of data center and transport equipment (storage, servers, routers etc.).

1) Networks: In [56], Neilson showed that over the past ten years a 2× increase in throughput of state-of-the-art equipment has been accompanied by a 1.4× increase in power. In addition, router capacity per rack has increased by 1.56×/year. Combining these two trends, Neilson concluded that state-of-the-art router efficiency is improving by 20% per annum. A more recent analysis of router improvement rates [22] presents data suggesting that router energy consumption per bit will decrease by 15% per annum for the next decade. In our analysis, we use the more optimistic router improvement rate of 20% per annum. This corresponds to a technology improvement rate $\alpha = 0.2$ in (6).

The optical components of transport equipment, including optical multiplexers, doped fibres, etc., have more limited scope for improvements in efficiency. Fortunately, the majority of the power consumption in transport equipment is in the optoelectronic components, such as pump lasers, and in the electronic components that perform the multiplexing, control, and management functions for the transport system [21], [22]. These should improve in efficiency in future generations. The results of a recent analysis [22] suggested that the energy consumption per bit of SDH transport systems will decrease by 14% per annum ($\alpha = 0.14$), which we include in our analysis.

2) Data Centers: For the past decade, the energy efficiency of servers (performance per watt) has typically doubled every two to four years [54]. With increasing attention being paid to the need to manage power consumption of data centers, it is reasonable to expect that this trend will continue. In our analysis, we include a doubling of performance per watt in servers every three years, which corresponds to $\alpha = 0.21$.

Increased storage density in HDD platters has achieved exponential reductions in energy consumption [55]. For the past decade, the power consumption per byte (watts per bit) for storage in HDDs has decreased by 30% per annum and this trend is expected to continue [55]. In our analysis, we include a 30% per annum rate of improvement in energy efficiency of storage, which corresponds to $\alpha = 0.3$.

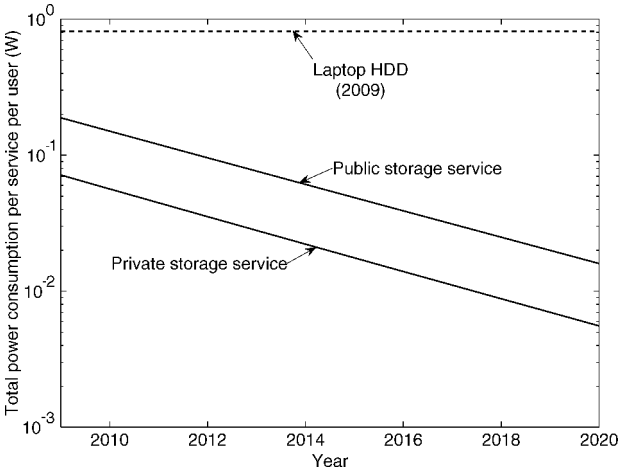

Fig. 11. ***Per-user power consumption of public and private cloud storage services for the years 2009–2020. The cloud storage services store on average 20 active files per user with an average file size of 1.25 MB. The per-user per-file download rate is one download per hour. Also included is the power consumption of a modern laptop HDD.***

B. Storage as a Service

We now forecast the per-user energy consumption of storage as a service. The cloud storage service stores on average 20 active files per user with an unchanging average file size of 1.25 MB. The per-user per-file download rate is one download per hour. Fig. 11 shows the total per-user power consumption trend for such a public or private cloud storage service over the years 2009–2020. For reference, included in Fig. 11 is the power consumption of a modern laptop HDD (2.5″ HDD) in 2009. At one download per hour, we saw in Fig. 2 for the public cloud service and Fig. 3 for the private cloud service, that the energy consumption of transport dominates total power consumption. Improvements in technology should lead to a factor of 10 improvement over time for both types of services. However, as previously noted, the absolute energy savings from the service are small and there are better opportunities for large energy savings elsewhere.

C. Software as a Service

Our power consumption forecast of software as a service considers public and private cloud software services with 20 and 200 users per server. Fig. 12 shows the total per-user power consumption trend for such cloud software services over the years 2009–2020. The power consumption of the software services includes the power consumed by servers, storage, transport, and the user terminal. The user terminal is built using 2009 technology and its estimated power consumption is also included in Fig. 12. Although it is reasonable to expect user terminals to become more energy efficient in the future, in this analysis, we focus on net gains that will be achieved through improvements in server and transport equipment.

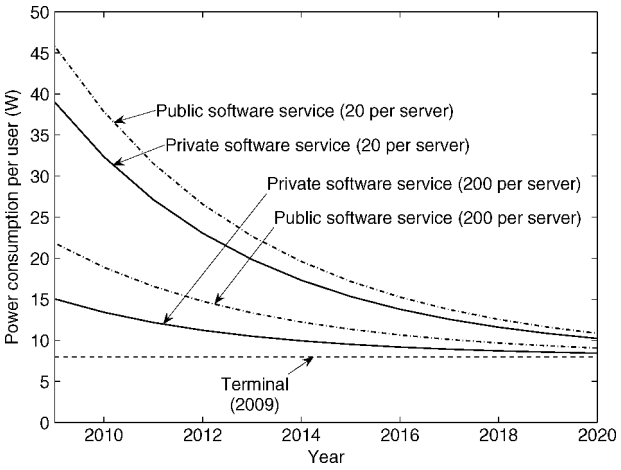

Fig. 12. ***Per-user power consumption of public and private cloud software services with 20 and 200 users per server for the years 2009–2020. Also included is the power consumption of a user terminal from 2009.***

Improvements in the energy efficiency of server, storage, and transport technology should lead to reductions of approximately 76% and 74% in the power consumption of the public and private cloud software services with 20 users, respectively. Technology improvements should also lead to the power consumption of public and private cloud software services with 200 users falling by 59% and 44%, respectively. This plot is based on a 2009 user terminal for all years, and so the total power consumption of each cloud service asymptotes to the power consumption of the user terminal, which is 8 W. The results suggest that to achieve energy consumption reductions in the long-term future, improvements in the user terminal are required. This could be a significant challenge as it requires end users to replace old equipment with new equipment, despite gaining no benefit in equipment functionality.

D. Processing as a Service

To forecast the energy consumption of processing as a service, we again consider a processing service used for computationally intensive tasks; in this case, the encoding of 2.5 h of video material 0.5 times per week. Fig. 13 shows the total per-user per-week energy consumption trends of such public and private cloud processing services for the years 2009–2020. The total energy consumption includes the energy required to perform common office tasks on a low-end laptop dating from 2009. As with software as a service, we keep the power consumption of the user equipment constant because, in this analysis, we focus on net gains that will be achieved through improvements in cloud computing equipment (servers and transport). For reference, Fig. 13 also includes the per-week energy consumption of a modern low-end laptop used 40 h/week and built with 2009 technology. Improvements in server and transport technology should lead to total energy consumption reductions of 35% and 30% for public and private cloud processing services, respectively. The total energy consumption of both cloud services asymptotes toward the energy consumed in the user laptop. The results suggest that to reduce overall energy consumption it will be important to improve the energy efficiency of user computing equipment as well as cloud computing equipment (servers, routers, etc.).

E. Discussion

The level of utilization achieved by a cloud service is a function of the type of services it provides, the number of users it serves, and the usage patterns of those users. Large-scale public clouds that serve a very large number of users are expected to be able to fully benefit from achieving high levels of utilization and high levels of virtualization, leading to low per-user energy consumption. Private clouds that serve a relatively small number of users may not have sufficient scale to fully benefit from the same energy-saving techniques. Our analysis is based on the view that cloud computing fully utilizes servers and storage for both public and private clouds. The results of our analysis indicate that private cloud computing is more energy efficient than public cloud computing due to the energy savings in transport. However, it is not clear whether in general the energy consumption saving in transport with a private cloud offsets the higher energy consumption due to lower utilization of servers and storage.

The logical unification of several geographically diverse data centers assists cloud computing to scale during

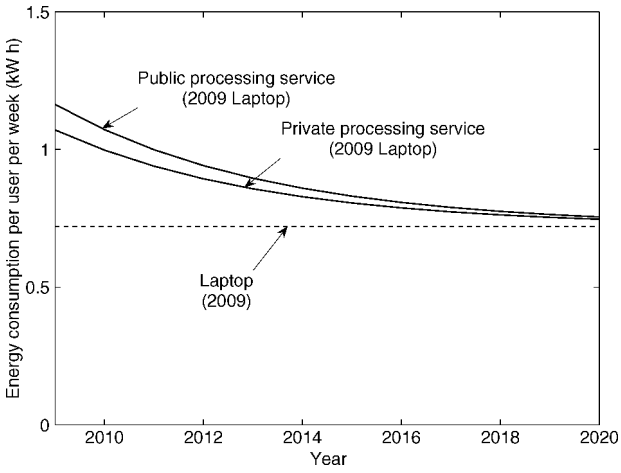

Fig. 13. ***Per-user per-week energy consumption of public and private cloud processing services for the years 2009–2020. The cloud processing service is used to perform an average of 0.5 encodings per week. The total energy consumption includes the energy consumed in the user's laptop. Also included is the power consumption of a 2009 low-end laptop.***

periods of high demand. However, energy-efficient transport between these data centers is necessary to ensure that cloud computing is energy efficient. In our analysis, public clouds consumed more energy than private clouds because users connected to the public cloud through the public Internet. Specifically, the large number of router hops required to traverse the public Internet greatly increases the energy consumption in transport. Optical bypass can be used to reduce the number of router hops through the network [58], [59] and thus the energy consumption in transport [21]. To minimize the energy consumption in transport, cloud computing data centers should be connected through dedicated point-to-point links incorporating optical bypass where possible. Indeed, reducing the number of routings hops and transmission links would yield benefits to all services.

VI. CONCLUSION

In this paper, we presented a comprehensive energy consumption analysis of cloud computing. The analysis considered both public and private clouds and included energy consumption in switching and transmission as well as data processing and data storage. We have evaluated the energy consumption associated with three cloud computing services, namely storage as a service, software as a service, and processing as a service. Any future service is likely to include some combination of each of these service models.

Power consumption in transport represents a significant proportion of total power consumption for cloud storage services at medium and high usage rates. For typical networks used to deliver cloud services today, public cloud storage can consume of the order of three to four times more power than private cloud storage due to the increased energy consumption in transport. Nevertheless, private and public cloud storage services are more energy efficient than storage on local hard disk drives when files are only occasionally accessed. However, as the number of file downloads per hour increases, the energy consumption in transport grows and storage as a service consumes more power than storage on local hard disk drives. The energy savings from cloud storage are minimal.

In cloud software services, power consumption in transport is negligibly small at very low screen refresh rates. As a result, cloud services are more efficient than modern midrange PCs for simple office tasks. At moderate and high screen refresh rates, power consumption in transport becomes significant and energy savings over midrange PCs are reduced. The number of users per server is the most significant determinant of the energy efficiency of a cloud software service. Cloud software as a service is ideal for applications that require average frames rates lower than the equivalent of 0.1 screen refresh frames per second.

Significant energy savings are achieved by using low-end laptops for routine tasks and cloud processing services for computationally intensive tasks, instead of a midrange or high-end PC, provided the number of computationally intensive tasks is small. Energy consumption in transport with a private cloud processing service is negligibly small.

Our broad conclusion is that the energy consumption of cloud computing needs to be considered as an integrated supply chain logistics problem, in which processing, storage, and transport are all considered together. Using this approach, we have shown that cloud computing can enable more energy-efficient use of computing power, especially when the users' predominant computing tasks are of low intensity or infrequent. However, under some circumstances, cloud computing can consume more energy than conventional computing where each user performs all computing on their own PC. Even with energy-saving techniques such as server virtualization and advanced cooling systems, cloud computing is not always the greenest computing technology. ■

Acknowledgment

The authors would like to thank S. Hossain for his helpful comments and suggestions.

REFERENCES

[1] Cisco. (2009). Cisco visual networking index: Forecast and methodology, 2009–2014. White paper. [Online]. Available: http://www.cisco.com.

[2] A. Weiss, "Computing in the clouds," *netWorker*, vol. 11, no. 4, pp. 16–25, 2007.

[3] B. Hayes, "Cloud computing," *Commun. ACM*, vol. 51, no. 7, pp. 9–11, 2008.

[4] T. Singh and P. K. Vara, "Smart metering the clouds," in *Proc. IEEE Int. Workshops Enabling Technol., Infrastructures for Collaborative Enterprises*, Groningen, The Netherlands, Jun.–Jul. 2009, pp. 66–71.

[5] D. Kondo, B. Javadi, P. Malecot, F. Cappello, and D. P. Anderson, "Cost-benefit analysis of cloud computing versus desktop grids," in *Proc. IEEE Int. Symp. Parallel Distrib. Process.*, Rome, Italy, May 2009, DOI: 10.1109/IPDPS.2009.5160911.

[6] R. Buyya, C. S. Yeo, and S. Venugopal, "Market-oriented cloud computing: Vision, hype, and reality for delivering IT services as computing utilities," in *Proc. 10th IEEE Int. Conf. High Performance Comput. Commun.*, Dalian, China, Sep. 2008, pp. 5–13.

[7] A. Greenberg, P. Lahiri, D. A. Maltz, P. Patel, and S. Sengupta, "Towards a next generation data center architecture: Scalability and commoditization," in *Proc. ACM Workshop Programmable Routers for Extensible Services of Tomorrow*, New York, 2008, pp. 57–62.

[8] *Open Cloud Manifesto*. [Online]. Available: http://www.opencloudmanifesto.org/

[9] 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," Electr. Eng. Comput. Sci. Dept., Univ. California, Berkeley, CA, Tech. Rep. UCB/EECS-2009-28, Feb. 2009.

[10] L. M. Vaquero, L. Rodero-Merino, J. Caceres, and M. Lindner, "A break in the clouds: Towards a cloud definition," *SIGCOMM Comput. Commun. Rev.*, vol. 39, no. 1, pp. 50–55, 2009.

[11] P. Mell and T. Grance, *Draft NIST Working Definition of Cloud Computing v14*, Nat. Inst. Standards Technol., 2009. [Online]. Available: http://csrc.nist.gov/groups/SNS/cloud-computing/index.html

[12] *Google Docs*. [Online]. Available: http://docs.google.com

[13] *Amazon Web Services*. [Online]. Available: http://aws.amazon.com

[14] *Azure Services Platform*. [Online]. Available: http://www.microsoft.com/azure

[15] *IBM Smart Business Services*. [Online]. Available: http://www.ibm.com/ibm/cloud

[16] *Salesforce.com*. [Online]. Available: http://www.salesforce.com

[17] *Webex*. [Online]. Available: http://www.webex.com/

[18] GeSI, "Smart 2020: Enabling the low carbon economy in the information age," London, U.K., 2008. [Online]. Available: http://www.smart2020.org/_assets/files/02_Smart2020Report.pdf

[19] M. Gupta and S. Singh, "Greening of the Internet," in *Proc. Conf. Appl. Technol. Architectures Protocols Computer Commun.*, Karlsruhe, Germany, 2003, pp. 19–26.

[20] J. Koomey, *Estimating Total Power Consumption by Servers in the U.S. and the World*. Oakland, CA: Analytics Press, 2007.

[21] J. Baliga, R. Ayre, K. Hinton, W. V. Sorin, and R. S. Tucker, "Energy consumption in optical IP networks," *J. Lightw. Technol.*, vol. 27, no. 13, pp. 2391–2403, Jul. 2009.

[22] O. Tamm, C. Hermsmeyer, and A. M. Rush, "Eco-sustainable system and network architectures for future transport networks," *Bell Labs Tech. J.*, vol. 14, no. 4, pp. 311–327, Feb. 2010.

[23] A. Vukovic, "Data centers: Network power density challenges," *ASHRAE J.*, vol. 47, pp. 55–59, 2005.

[24] J. Liu, F. Zhao, X. Liu, and W. He, "Challenges towards elastic power management in internet data centers," in *Proc. IEEE Int. Conf. Distrib. Comput. Syst. Workshops*, Los Alamitos, CA, 2009, pp. 65–72.

[25] F. Hermenier, N. Loriant, and J.-M. Menaud, "Power management in grid computing with Xen," *Lecture Notes in Computer Science*. Berlin, Germany: Springer-Verlag, 2006, pp. 407–416.

[26] J. S. Chase, D. C. Anderson, P. N. Thakar, A. M. Vahdat, and R. P. Doyle, "Managing energy and server resources in hosting centers," in *Proc. 18th ACM Symp. Oper. Syst. Principles*, New York, 2001, pp. 103–116.

[27] W. Vereecken, L. Deboosere, D. Colle, B. Vermeulen, M. Pickavet, B. Dhoedt, and P. Demeester, "Energy efficiency in telecommunication networks (invited paper)," in *Proc. 13th Eur. Conf. Netw. Opt. Commun.*, Krems, Austria, Jul. 2008, pp. 44–51.

[28] J. Baliga, K. Hinton, and R. S. Tucker, "Energy consumption of the Internet," in *Conf. Opt. Internet/Australian Conf. Opt. Fibre Technol.*, Melbourne, Vic., Australia, Jun. 2007.

[29] R. S. Tucker, R. Pathiban, J. Baliga, K. Hinton, R. W. Ayre, and W. V. Sorin, "Evolution of WDM optical IP networks: A cost and energy perspective," *J. Lightw. Technol.*, vol. 27, no. 3, pp. 243–252, Feb. 2009.

[30] M. Al-Fares, A. Loukissas, and A. Vahdat, "A scalable, commodity data center network architecture," in *Proc. ACM SIGCOMM Conf. Data Commun.*, New York, 2008, pp. 63–74.

[31] J. Baliga, R. Ayre, K. Hinton, and R. S. Tucker, "Architectures for energy-efficient IPTV networks," in *Proc. Opt. Fiber Commun./Nat. Fiber Opt. Eng. Conf.*, San Diego, CA, Mar. 2009, [CD-ROM].

[32] P. Barham, B. Dragovic, K. Fraser, S. Hand, T. Harris, A. Ho, R. Neugebauer, I. Pratt, and A. Warfield, "Xen and the art of virtualization," *SIGOPS Oper. Syst. Rev.*, vol. 37, no. 5, pp. 164–177, 2003.

[33] *Google—Efficient Computing*. [Online]. Available: http://www.google.com/corporate/green/datacenters

[34] J. Baliga, R. Ayre, W. V. Sorin, K. Hinton, and R. S. Tucker, "Energy consumption in access networks," in *Proc. Opt. Fiber Commun./Nat. Fiber Opt. Eng. Conf.*, San Diego, CA, Feb. 2008, [CD-ROM].

[35] The Green Grid. (2007, Feb.). The green grid metrics: Describing data center power efficiency. Tech. Committee White Paper. [Online]. Available: http://www.thegreengrid.org/

[36] U.S. Environmental Protection Agency, "EPA report on server and data center energy efficiency," Aug. 2007.

[37] *Hewlett-Packard Data Sheets*. [Online]. Available: http://www.hp.com

[38] J. S. Domingo and K. Karagiannis, "Servers: More power 2U," *PC Mag.*, Oct. 2004. [Online]. Available: http://www.pcmag.com/print_article2/0,1217,a=135254,00.asp?hidPrint=true

[39] *HP Power Calculator*. [Online]. Available: http://h30099.www3.hp.com/configurator/powercalcs.asp

[40] *Cisco Data Sheets*. [Online]. Available: http://www.cisco.com

[41] *Juniper Data Sheets*. [Online]. Available: http://www.juniper.net

[42] 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 *Proc. 7th Symp. Oper. Syst. Design Implementation*, Berkeley, CA, 2006, pp. 205–218.

[43] D. Colarelli and D. Grunwald, "Massive arrays of idle disks for storage archives," in *Proc. ACM/IEEE Conf. Supercomput.*, Los Alamitos, CA, 2002, DOI: 10.1109/SC.2002.10058.

[44] S. Srikantaiah, A. Kansal, and F. Zhao, "Energy aware consolidation for cloud computing," in *Proc. HotPowerWorkshop Power Aware Comput. Syst.*, Dec. 2008 [CD-ROM].

[45] *Fujitsu Data Sheets*. [Online]. Available: http://www.fujitsu.com

[46] K. J. Christensen, C. Gunaratne, B. Nordman, and A. D. George, "The next frontier for communications networks: Power management," *Comput. Commun.*, vol. 27, pp. 1758–1770, 2004.

[47] P. Chanclou, S. Gosselin, J. F. Palacios, V. L. Alvarez, and E. Zouganeli, "Overview of the optical broadband access evolution: A joint article by operators in the IST network of excellence e-photon/one," *IEEE Commun. Mag.*, vol. 44, no. 8, pp. 29–35, Aug. 2006.

[48] *Spring-Nextel Applied Research GroupIPMon Project*, Jan. 2005. [Online]. Available: http://ipmon.sprint.com

[49] L. Ceuppens, "Planning for energy efficiency—Networking in numbers," *Opt. Fiber Commun./Nat. Fiber Opt. Eng. Conf. Workshop Energy Netws.*, San Diego, CA, Mar. 2009. [Online]. Available: http://www.ict-bone.eu/portal/landing_pages/bone_documents/OFC2009_BONE_Energy_Footprint.zip

[50] E. Desurvire, "Capacity demand and technology challenges for lightwave systems in the next two decades," *J. Lightw. Technol.*, vol. 24, no. 12, pp. 4697–4710, Dec. 2006.

[51] P. V. Mieghem, *Performance Analysis of Communications Networks and Systems*. New York: Cambridge Univ. Press, 2005.

[52] *International Telecommunication Union*. [Online]. Available: http://www.itu.int

[53] *×264 Benchmark*. [Online]. Available: http://www.techarp.com/showarticle.aspx?artno=442

[54] J. G. Koomey, C. Belady, M. Patterson, A. Santos, and K.-D. Lange, "Assessing trends over time in performance, costs and energy user for servers," *IEEE Ann. History Comput.*, 2009.

[55] D. Reinsel, "The real costs to power and cool all the world's external storage," IDC Tech. Rep. IDC #212714, Jun. 2008, vol. 1.

[56] D. T. Neilson, "Photonics for switching and routing," *IEEE J. Sel. Top. Quantum Electron.*, vol. 12, no. 4, pp. 669–678, Jul./Aug. 2006.

[57] G. Epps, "System power challenges," in *Proc. Cisco Routing Res. Symp.*, San Jose, CA, Aug. 2006. [Online]. Available: www.cisco.com/web/about/ac50/ac207/proceedings/POWER_GEPPS_rev3.ppt

[58] A. A. M. Saleh and J. M. Simmons, "Evolution toward the next-generation core optical network," *J. Lightw. Technol.*, vol. 24, no. 9, pp. 3303–3321, Sep. 2006.

[59] R. Parthiban, R. S. Tucker, and C. Leckie, "Waveband grooming and IP aggregation in optical networks," *J. Lightw. Technol.*, vol. 21, no. 11, pp. 2476–2488, Nov. 2003.

ABOUT THE AUTHORS

Jayant Baliga received the B.Sc. degree in computer science and the B.E. degree in electrical and electronic engineering (with first class honors) from the University of Melbourne, Melbourne, Vic., Australia, in 2007, where he is currently working towards the Ph.D. degree in electrical engineering.

His research interests include energy consumption, optical network architectures, and wireless communications.

Robert W. A. Ayre received the B.Sc. degree in electronic engineering from George Washington University, Washington, DC, in 1967 and the B.E. and M. Eng.Sc degrees from Monash University, Melbourne, Melbourne, Vic., Australia, in 1970 and 1972, respectively.

In 1972, he joined the Research Laboratories of Telstra Corporation, working in a number of roles primarily in the areas of optical transmission for core and access networks, and in broadband networking. In 2007, he joined the ARC Special Centre for Ultra-Broadband Networks (CUBIN), University of Melbourne, Melbourne, Vic., Australia, continuing work on networking and high-speed optical technologies.

Kerry Hinton was born in Adelaide, S.A., Australia, in 1955. He received the Honors Bachelor of Engineering degree, the Honors Bachelor of Science degree, and the M.S. degree in mathematical sciences from the University of Adelaide, Adelaide, S.A., Australia, in 1978, 1980, and 1982, respectively, the Ph.D. degree in theoretical physics from the University of Newcastle Upon Tyne, Newcastle Upon Tyne, U.K., and the Diploma in industrial relations from the Newcastle Upon Tyne Polytechnic, Newcastle Upon Tyne, U.K., in 1984.

In 1984, he joined Telstra Research Laboratories (TRL), Vic., Australia, and worked on analytical and numerical modeling of optical systems and components. His work has focused on optical communications devices and architectures, physical layer issues for automatically switched optical networks (ASONs) and monitoring in all-optical networks. He was also a laser safety expert within Telstra. In 2006, he joined the ARC Special Centre for Ultra-Broadband Networks (CUBIN), University of Melbourne, Melbourne, Vic., Australia, where he is undertaking research into the energy efficiency of the Internet and optical communications technologies.

Rodney S. Tucker (Fellow, IEEE) received the B.E. degree in electrical engineering and the Ph.D. degree from the University of Melbourne, Melbourne, Vic., Australia, in 1969 and 1975, respectively.

Currently, he is a Laureate Professor at the University of Melbourne, where he is the Director of the Institute for a Broadband-Enabled Society and the Director of the Centre for Energy-Efficient Telecommunications.

Dr. Tucker is a Fellow of the Australian Academy of Science, a Fellow of the Australian Academy of Technological Sciences and Engineering, and a Fellow of the Optical Society of America. In 1975, he was the recipient of a Harkness Fellowship by the Commonwealth Fund, New York. From 1988 to 1990, he was the Editor-in-Chief of the IEEE Transactions on Microwave Theory and Techniques. From 1991 to 1993, he was with the Management Committee of the Australian Telecommunications and Electronics Research Board, and a member of the Australasian Council on Quantum Electronics. From 1995 to 1999 and from 2009 to the present, he is a member of the Board of Governors of the IEEE Lasers and Electro-optics Society. In 1995, he was the recipient of the Institution of Engineers, Australia, M. A. Sargent Medal for his contributions to Electrical Engineering and was named IEEE Lasers and Electro-optics Society Distinguished Lecturer for the year 1995–1996. In 1997, he was the recipient of the Australia Prize, Australia's premier award for science and technology for his contributions to telecommunications. From 1997 to 2006, he was an Associate Editor of the IEEE Photonics Technology Letters. He is currently Vice-President, Publications of the IEEE Photonics Society. In 2007, he was the recipient of the IEEE Lasers and Electro-optics Society Aron Kressel Award for his pioneering contributions to high-speed semiconductor lasers.