Energy Consumption of Cloud Computing and Fog Computing Applications

by

Fatemeh Jalali

Submitted in total fulfilment of
the requirements for the degree of

Doctor of Philosophy

Department of Electrical and Electronic Engineering
The University of Melbourne
Australia

July 2015

Abstract

Energy Consumption of Cloud Computing and
Fog Computing Applications

by Fatemeh Jalali

Supervisors: Laureate Professor Rod Tucker, Dr Tansu Alpcan, Dr Kerry Hinton

---

A great deal of attention has been paid to the energy consumption of Cloud services and data centers in an endeavor to reduce the energy consumption and carbon footprint of the ICT industry. Since the data in Cloud services is processed and stored in data centers, an obvious focus for studying energy consumption of Cloud services is the data centers. However, the energy consumption of a Cloud service is not just due to data centers, it also includes energy consumption of the transport network that connects end-users to the Cloud and the energy consumption of end-user devices when accessing the Cloud. In most of previous studies on energy consumption of Cloud computing services, the energy consumed in the transport network and end-user devices has not taken into account. To show the importance of energy consumption of these ignored parts, the total energy consumed by three well-known Cloud applications, Facebook, Google Drive and Microsoft OneDrive, is studied using measurements and modeling. The results show that achieving an energy-efficient Cloud service requires improving the energy efficiency of the transport network and the end-user devices along with the related data centers.

The popularity of hosting and distributing content and applications from small servers located in end-user premises (known as nano data centers) is increasing especially with the advent of Internet of Things (IoT) and the Fog Computing paradigm. In this work we study energy consumption of nano data centers since there are different views on the energy consumption of nano data centers. These differences stem from using different energy consumption models and ignoring energy consumed in the transport network.

To fill the knowledge gap in this field, we propose established and measurement based models for network topology and energy consumption to identify parameters that make

nano data centers more/less energy-efficient than centralized data centers. A number of findings emerge from this study, including the factors that enable nano data centers to consume less energy than its centralized counterpart, such as (a) type of access network attached to nano servers, (b) the ratio of nano server's idle time to active time and, (c) type of applications which includes number of downloads, updates and data pre-loading.

This study shows that nano data centers can complement centralized data centers and lead to energy savings for applications that are off-loadable from centralized data centers to nano data centers.

To all smart, talented and brilliant women who have not had the opportunity to pursue their education as they wished

&

to all men who have been supportive of women's endeavors.

This is to certify that

(i) the thesis comprises only my original work,

(ii) due acknowledgment has been made in the text to all other material used,

(iii) the thesis is less than 100,000 words in length, exclusive of table, maps, bibliographies, appendices and footnotes.

Signature_____

Date_____

Acknowledgments

A PhD is a unique journey full of ups and downs that comes with many lessons. Being on the right tracks during the journey would not be possible for me without all those people who helped me along the way. First of all, I would like to express my deepest gratitude to my supervisors Laureate Professor Rod Tucker, Dr Tansu Alpcan and Dr Kerry Hinton for their guidance and support throughout my PhD candidature. Their approaches toward research and life has been a continuous inspiration during my candidature and will be an asset for my life.

I would offer my sincere appreciation to Dr Rob Ayre for his endless support and help during my PhD specially for the practical experiments. I am grateful to Dr Arun Vishwanath, my former co-supervisor, and Dr Leith Campbell for their collaboration and help which greatly influenced this thesis. I also would like to thank the chair of my PhD advisory panel Professor Thas Nirmalathas for his constructive comments and suggestions on improving my work.

I would also like to thank all the past and current members of Centre for Energy-Efficient Telecommunications (CEET) at the University of Melbourne, in particular, Dr Bipin Sankar, Tony Lin, Dinuka Kudavithana, Chrispin Gray, Sascha Suessspeck, Hamid Khodakarami, and Ashrar Matin. A special thanks to my friends from the EEE department at the University of Melbourne Rajitha Senanayake, Dr Ehsan Nekouei and Dr Julien Ridoux. It has been a great pleasure to work with such a talented group of friends and learn from them. I cherish their support and friendship.

I also thank the CEET collaborators Dr Jaime Llorca from Bell Labs (New Jersey) and Grant Underwood from Alcatel Lucent (Sydney) for their excellent reviews and suggestions on improving my research work. I acknowledge the University of Melbourne, The State Government of Victoria and Bell Labs for funding this research.

I am heartily thankful to my parents and siblings for their support and encouragement at all times. Finally, and most importantly, I thank my husband Dr Reza Emdad for his love, inspiration, support, and for making my PhD journey filled with joy and happiness.

List of Publications

Journal Papers

1. F. Jalali, R. Ayre, T. Alpcan, K. Hinton and R. Tucker, "Fog Computing May Help to Save Energy in Cloud Computing", *IEEE Journal on Selected Areas in Communications (J-SAC)*, 2015 (Under second revision).

2. K. Hinton, F. Jalali and A. Matin, "Energy consumption modeling of optical networks", *Photonic Network Communications, Springer*, vol. 30, no. 1, pp. 4 - 16, 2015.

3. A. Vishwanath, F. Jalali, R. Ayre, T. Alpcan, K. Hinton and R. Tucker, "Energy Consumption Comparison of Interactive Cloud-Based and Local Applications", *IEEE Journal on Selected Areas in Communications (J-SAC)*, pp. 616 - 626, vol. 33, Issue 4, 2015.

4. F. Jalali, A. Vishwanath, R. Ayre, T. Alpcan, K. Hinton and R. Tucker, "Energy Consumption of Content Distribution from Nano Data Centers versus Centralized Data Centers", in Proceeding of *ACM SIGMETRICS Performance Evaluation Review (Greenmetrics)*, 2014. **(Best Student Paper Award by IEEE Victorian Section 2014)**

Conference Papers

1. F. Jalali, A. Vishwanath, R. Ayre, T. Alpcan, K. Hinton and R. Tucker, "Energy Consumption of Photo Sharing in Online Social Networks", in Proceeding of *14th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid)*, Chicago, USA, May 2014.

2. A. Vishwanath, F. Jalali, R. Ayre, T. Alpcan, K. Hinton and R. Tucker, "Energy Consumption of Interactive Cloud-Based Document Processing Applications", in Proceeding of *IEEE International Conference on Communications (ICC)*, Budapest, Hungary, June 2013.

3. F. Jalali, “Energy Consumption of Cloud Applications”, in Proceeding of *Asia-Oceania Top University League of Engineering (AOTULE)*, Melbourne, Australia, November, 2014.

Posters

1. F. Jalali, “Hidden Energy Consumption of Photo Sharing in Online Social Networks”, *14th Annual Grace Hopper Celebration of Women in Computing (GHC’14)*, Phoenix, USA, October 2014.

2. F. Jalali, “Home Servers Can Save Energy for IoT Applications”, *15th Annual Grace Hopper Celebration of Women in Computing (GHC’15)*, Houston, USA, October 2015. (Accepted for ACM Student Research Competition)

Contents

List of Figures

List of Tables

List of Acronyms

ADSL	Asymmetric Digital Subscriber Line
BNG	Broadband Network Gateway
CCN	Content Centric Networking
CDN	Content Delivery(/Distribution) Network
CP	Content Provider
CPE	Customer Premises Equipment
DC	Data Center
DSL	Digital Subscriber Line
DHT	Distributed Hash Table
Gbps	Giga bit per second
GB	Giga Byte
GWh	Giga Watt hour
HVAC	Heating, Ventilating, and Air Conditioning
HTTPS	HyperText Transfer Protocol Secure
IaaS	Infrastructure as a Service
ICT	Information and Communication Technology
IoT	Internet of Things
IPTV	Internet Protocol TV
ISP	Internet Service Provider
kbps	Kilo bit per second
kB	Kilo Byte
kWh	Kilo Watt hour
LTE	Long-Term Evolution
Mbps	Mega bit per second
MB	Mega Byte
mW	milli Watt
nJ/b	Nano Joule per bit
OSN	Online Social Network
PON	Passive Optical Network
P2P	Peer-to-Peer
PC	Personal Computer
PaaS	Platform as a Service
PoP	Points of Presence
PACR	Popularity-Aware Content Replication
PUE	Power Use Effectiveness
SLA	Service Level Agreement
TCP	Transmission Control Protocol
TJ	Tera Joule
USB	Universal Serial Bus
VoD	Video on Demand
Wh	Watt hour
WWW	World Wide Web

Chapter 1

Introduction

Cloud computing is an advanced technology that has revolutionized the ICT industry. It has changed the way that services are offered through the World Wide Web (WWW) by providing computing resources such as hardware, application development platforms and computer applications available as services over the Internet. The services made available this way are commonly known, respectively, as Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS). Therefore, Cloud computing offer accessing to the stored data from anywhere anytime and expanding the services independent of end-user hardware. Cloud computing presents numerous benefits compared to the traditional computing in terms of cost, scalability, performance, maintenance, etc [2].

Cloud-based applications and services such as online social networks (OSNs), media sharing and file storage have become increasingly popular among users in recent years. As an example, Facebook users upload more than 350 million photos every day [7] and even more on special occasions such as New Year's Eve and Day uploading up to 1.1 Billion photos [8]. YouTube users upload over 300 hours of video every minute and watch hundreds of millions of hours on YouTube every day [9]. This has led to a significant amount of Cloud traffic and data center traffic which is forecast to increase about threefold between 2013 and 2018 as shown in Figure 1.1 [1]. As the Cloud traffic is increasing, the concern for energy consumption is also rising [10].

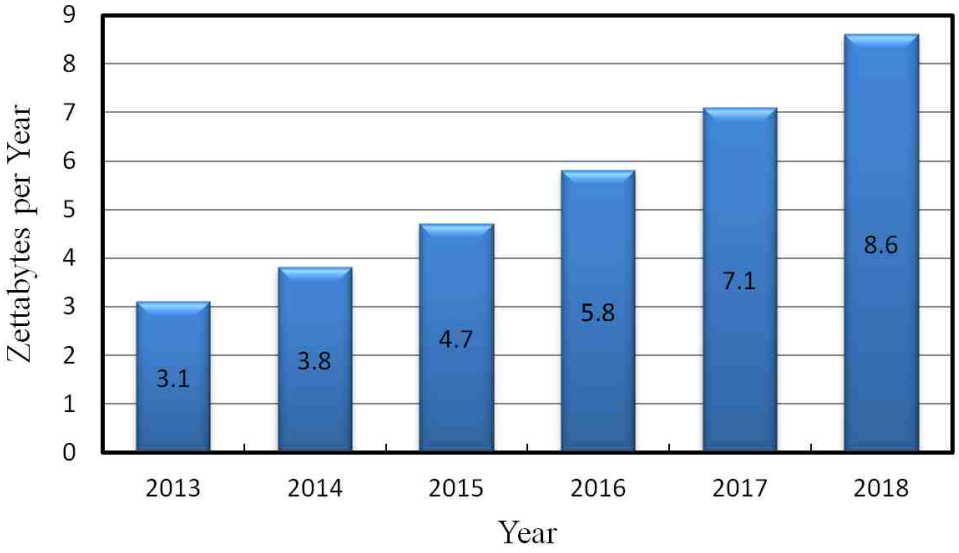


Figure 1.1: Global data center IP traffic growth [1].

1.1 Energy Consumption of Cloud Applications and Services

In order to mitigate energy consumption of Cloud computing applications and services, an obvious focus for reducing energy consumption is data centers since data in Cloud services is processed and stored in the data centers [10]. Therefore, a number of different approaches have been applied to improve the energy efficiency within mega centralized data centers, such as energy proportional computing, dynamic provisioning, cooling method, virtualization of computing resources, etc [10].

In most of studies on energy consumption of the Cloud applications and services, energy consumed within relevant data centers is counted as the total energy consumption of the Cloud applications and services. For example, the energy consumed within Google and Facebook data centers is perceived as the total energy consumption of Google and Facebook applications respectively. However, data centers are not the only component of Cloud computing applications and services. Transport networks and end-user devices are also two important components of Cloud services and applications. Therefore, the total energy consumption of Cloud applications and services includes three components

as shown in Figure 1.2:

a) Energy consumed in end-user devices when accessing the Cloud;

b) Energy consumed in the transport network between end-users and data centers;

c) Energy consumption of Cloud data centers.

In the most of the previous work that studied energy consumption of Cloud applications and services [11, 12, 13, 14, 15, 16], the energy consumption of end-user devices and the transport network has not taken into account. This is one reason Cloud applications and services are often promoted as a green technology. There are some other studies such as [17, 18, 19] that considered energy consumption of end-user devices and the transport network but there is still lack of in-depth energy consumption measurements and modeling for end-user devices and the transport network. In this work, the energy consumed in end-user devices and transport network is calculated for three well known Cloud applications (Facebook, GoogleDocs, Microsoft OneDrive) and compared with their energy consumed within the relevant data centers in order to highlight the importance of energy consumption of these ignored components based on a combination of measurements and modelings. The direct power consumption measurements are conducted for the devices located in end-user premises such as laptops, tablets and Raspberry Pis (very small and low power single board computers) [20]. However, for network equipment such as routers and switches located in the core and edge networks we propose a novel model for energy consumption called "flow-based" model which is based upon proportional allocation of the equipment power consumption over all the flows through the equipment.

The proposed energy consumption model and measurement techniques are applied to Facebook, Google Drive and Microsoft OneDrive in order to obtain the total energy consumption of the applications. Facebook is studied as a representative of online social networks (OSNs) with more than 1 billion users who upload more than 350 million photos every day on Facebook(and even more on the special occasions such as Halloween and New Year's Eve [8]).

In order to consider other types of Cloud applications that differ from OSNs, Google Drive and Microsoft OneDrive are studied as representatives of interactive Cloud-based

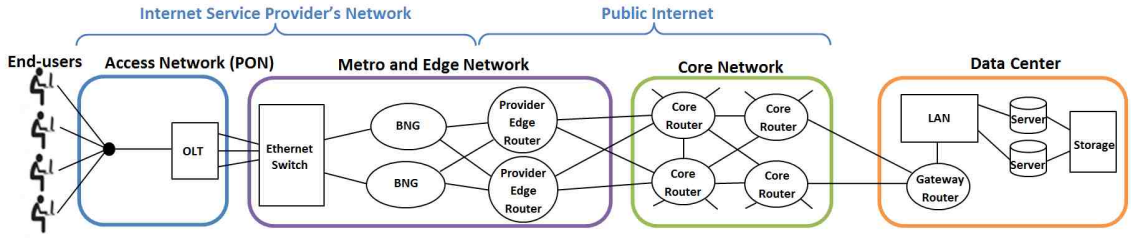


Figure 1.2: Schematic of networks connecting users to a Cloud and the data center infrastructure used to host Cloud services [2].

applications. To compare energy consumption of online interactive Cloud applications to local version of the applications, three scenarios are studied:

a) Creating, editing and saving Word, Presentation and Spreadsheet files in the Cloud;

b) Creating and editing the files locally, and then saving the files in the Cloud;

c) Performing the tasks locally (i.e., the Cloud is absent). All the tasks are performed on the same low-power consuming end-user devices.

Our results reveal that the energy consumed in the end-user devices and the transport network are a significant portion of the total energy consumption of the Cloud applications. A major source of energy-inefficiency in the Cloud applications and services is transport network specifically wireless access network.

Therefore, achieving an energy-efficient Cloud application requires energy efficiency improvement in the transport network and end-user devices along with the related data centers. Additionally, our study shows that performing certain tasks locally and then storing the final results of the tasks to the Cloud is more energy-efficient than doing the same task totally online in the Cloud. It is notable that the results of this work are only based on the consumed energy in the use-phase of applications and equipment not during their entire lifetime. Therefore, the results of this work do not include life-cycle assessment (LCA) [21, 22].

1.2 Energy Consumption of Content and Application Distribution from End-user Premises

The popularity of hosting and distributing content and applications from small servers located in end-user premises, instead of centralized servers, is increasing especially with the introduction of the Fog Computing paradigm [23, 1] and Internet of Things (IoTs). These highly distributed servers that can host and distribute content and applications in a peer-to-peer (P2P) fashion are also known as nano data centers [24, 4].

Despite the growing popularity of nano data centers, their energy consumption is not well-studied and it requires more investigations. There are different points of view on energy consumption of content and application distribution from end-user premises in the literature which stem from using different energy consumption models and ignoring energy consumed in the transport network [4, 25, 26]. In order to have comprehensive models for network topology and energy consumption and to identity cases for which running applications from nano servers is more energy-efficient than running the same applications from centralized data centers, an end-to-end network architecture is constructed which includes all network equipment required for distributing and accessing content from centralized and nano data centers. Then, energy consumption models are proposed and used for shared and un-shared network equipment. In addition to direct measurement and "flow-based" energy consumption models, a new "time-based" energy consumption model is proposed for equipment located in end-user premises that are not shared by many users such as home gateways and home servers.

To apply and validate the proposed models using practical measurements and experiments, the energy consumption of Wordpress application [27] which can host content in servers within centralized data centers or servers in the end-user premises is studied. Nano servers are implemented using Raspberry Pis (very small and low power single board computers) [20] and are characterized by traffic and power consumption measurements. Using the energy models and the measurement techniques, the energy consumption resulting from requesting data from a nano server is compared to that of the same request served from a server within a centralized data center.

The results of this work indicate that while nano servers can save little amount of

energy for some applications by pushing content closer to end-users and decreasing the energy consumption in the transport network, it can consume significant energy when the nano servers are attached to an energy-inefficient access network or when the idle time of dedicated nano servers is much greater than their active time. It is also investigated what type of applications can be run from nano servers to save energy. It is found that parameters such as number of downloads, number of updates and the amount of data pre-loading play an important role on the energy consumption of the applications. The results show that the best energy savings using nano servers come from applications that generate and distribute data continuously in end-user premises which is not frequently accessed such as home video surveillance applications.

Consequently, this study shows that the most energy efficient strategy for hosting and distributing content and application is a combination of centralized data centers and nano servers. By identifying applications (or parts of there-of) best located in nano servers, rather than centralized data centers, the energy efficiency of those applications can be improved.

1.3 Thesis Contributions

The contributions of this thesis can be broadly divided into 4 categories: energy consumption modelings, power consumption measurements and practical experiments in Cloud and Fog computing, energy consumption of three existing and well-known Cloud applications, and calculating energy consumption of Fog computing services. The key contributions are:

- Theoretical analysis of power consumption of network equipment and developing new energy and power consumption models for shared and unshared network equipment. More details on this are presented in Chapter 3.

- Energy consumption of photo sharing on Facebook as a representative of online social networks (OSNs) is computed using the new proposed energy consumption modelings as well as power and traffic measurements. Energy consumed in the end-user devices and the transport network is compared with the energy consumed

within Facebook data centers. More details on this are presented in Chapter 4.

- Energy consumption of interactive online Cloud-based applications such as Google Drive and Microsoft OneDrive is computed and compared with their local version of the applications using the new proposed energy consumption modelings as well as power and traffic measurements. More details on this are explained in Chapter 5.

- Energy consumption of content and application distribution from servers located in end-user premises (Fog computing) is presented using the new proposed energy consumption modelings and practical experiments. The energy consumed in the Fog computing services is compared with their counterparts in Cloud computing in order to identify which applications consume less energy in Fog computing. The details of this contribution is presented in Chapter 6.

1.4 Thesis Organization

The core chapters of this thesis are derived from several papers published during the PhD candidature. The remainder of the thesis is organized as follows:

- Chapter 2 presents a survey on energy consumption of Cloud applications and services as well as a survey on energy consumption of distributed servers located in the core of the network as well as those servers located in end-user premises.

- Chapter 3 presents the theoretical analysis of our proposed energy and power consumption models for various network elements. This chapter is mostly derived from [5, 28, 6]:

  - K. Hinton, F. Jalali and A. Matin, "Energy consumption modeling of optical networks", *Photonic Network Communications, Springer*, 2015.

  - F. Jalali, R. Ayre, T. Alpcan, K. Hinton and R. Tucker, "Fog Computing May Help to Save Energy in Cloud Computing", Submitted to *IEEE Journal on Selected Areas in Communications (J-SAC)*,2015.

  - F. Jalali, A. Vishwanath, R. Ayre, T. Alpcan, K. Hinton and R. Tucker, "Energy Consumption of Content Distribution from Nano Data Centers versus

Centralized Data Centers", in Proceeding of *ACM SIGMETRICS Performance Evaluation Review (Greenmetrics)*, 2014.

- Chapter 4 presents energy consumption of photo sharing on Facebook as a representative of online social networks (OSNs). This chapter is derived from [5, 6, 29]:
  - F. Jalali, A. Vishwanath, R. Ayre, T. Alpcan, K. Hinton and R. Tucker, "Energy Consumption of Photo Sharing in Online Social Networks", in Proceeding of *14th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid)*, Austin, Chicago, USA, May, 2014.
  - F. Jalali, "Hidden Energy Consumption of Photo Sharing in Online Social Networks", *14th Annual Grace Hopper Celebration of Women in Computing (GHC'14)*, Phoenix, USA, October 2014.
  - K. Hinton, F. Jalali and A. Matin, "Energy consumption modeling of optical networks", *Photonic Network Communications, Springer*, 2015.
- Chapter 5 compares energy consumption of interactive online Cloud-based applications such as Google Drive and Microsoft OneDrive with their local version of the applications. This chapter is derived from [5, 30, 31, 32]:
  - A. Vishwanath, F. Jalali, R. Ayre, T. Alpcan, K. Hinton and R. Tucker, "Energy Consumption Comparison of Interactive Cloud-Based and Local Applications", *IEEE Journal on Selected Areas in Communications (J-SAC)*, pp. 616 - 626, vol. 33, Issue 4, 2015.
  - A. Vishwanath, F. Jalali, R. Ayre, T. Alpcan, K. Hinton and R. Tucker, "Energy Consumption of Interactive Cloud-Based Document Processing Applications", in Proceeding of *IEEE International Conference on Communications (ICC)*, Budapest, Hungary, 5-9 June, 2013.
  - F. Jalali, "Energy Consumption of Cloud Applications", in Proceeding of *Asia-Oceania Top University League of Engineering (AOTULE)*, Melbourne, Australia, November, 2014.
  - K. Hinton, F. Jalali and A. Matin, "Energy consumption modeling of optical networks", *Photonic Network Communications, Springer*, 2015.

- Chapter 6 presents energy consumption of content and application distribution from nano servers located in end-user premises. This chapter is derived from [5, 28, 33]:
  - F. Jalali, R. Ayre, T. Alpcan, K. Hinton and R. Tucker, "Fog Computing May Help to Save Energy in Cloud Computing", Submitted to *IEEE Journal on Selected Areas in Communications (J-SAC)*, 2015.
  - F. Jalali, A. Vishwanath, R. Ayre, T. Alpcan, K. Hinton and R. Tucker, "Energy Consumption of Content Distribution from Nano Data Centers versus Centralized Data Centers", in Proceeding of *ACM SIGMETRICS Performance Evaluation Review (Greenmetrics)*, 2014.
  - F. Jalali, "Home Servers Can Save Energy for IoT Applications", *15th Annual Grace Hopper Celebration of Women in Computing (GHC'15)*, ACM Student Research Competition (SRC), Houston, USA October, 2015.
  - K. Hinton, F. Jalali and A. Matin, "Energy consumption modeling of optical networks", *Photonic Network Communications, Springer*, 2015.
- Chapter 7 concludes the thesis with a summary of the main findings, and discussion of future research directions.

Chapter 2

Literature Review

Executive Summary

In recent decades, researchers have been actively investigating energy consumption of Cloud services and data centers in an effort to reduce the energy consumption and carbon footprint of the ICT industry. In this section, we first review the previous research on energy consumption of Cloud based applications and services.

In addition, energy consumption studies of networked distributed servers have recently received significant attention. In this section, we also present a survey on the previous research on energy-saving schemes in content delivery/distribution networks (CDNs) located in the edge of the network as well as small servers located in the end-user premises.

The sections for the literature survey summarized in Figure 2.1.

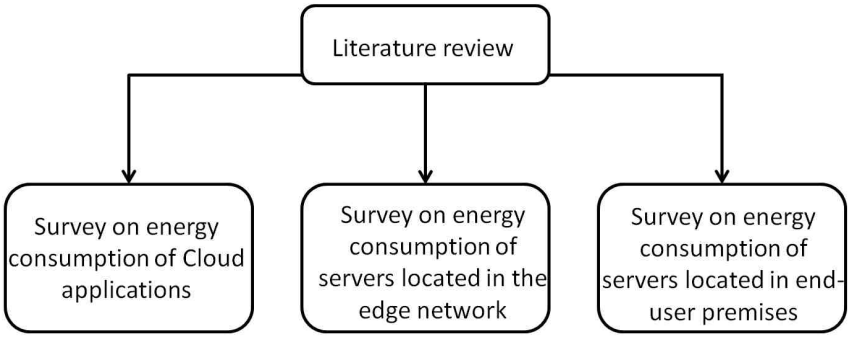


Figure 2.1: Diagram of the literature survey and contributions

2.1 Energy Consumption of Cloud Computing Applications and Services

In order to improve energy consumption of Cloud computing applications and services, the first focus for minimizing energy consumption is data centers since the data in Cloud services is processed, stored and sometimes generated in data centers [10]. A number of different approaches have been applied to improve the energy efficiency within data centers, such as energy proportional computing, improved environmental control, sleep scheduling and virtualization of computing resources, etc [10]. Most of the previous work only considered energy consumed within data centers for the total energy consumption of Cloud applications and services.

In 2009, Liu et al. [11] studied energy consumption of Cloud computing. They proposed a model called GreenCloud to save energy consumption of Cloud computing environment. This proposed model could support consolidate workload and obtain significant energy saving for Cloud computing environment as well as guarantee the real-time performance for various performance-sensitive applications at the same time. The GreenCloud used the state-of-the-art live virtual machine migration technology to achieve these goals. However, the proposed model only focused on energy consumption of data centers for saving energy in the Cloud computing environment and energy consumption of other components such as transport network and end-user terminals was not been included.

In 2009, Ali [12] proposed a Green Cloud scheme for mobile Cloud computing. The proposed scheme is based on "Network as a Service" (NaaS) which is the concept of dynamic bandwidth consumption and quality of server based on the application/service requirement. The proposed Green Cloud scheme modeled the needs for various types of consumers, communities and organizations, with the perspective of being environmentally "Greener" as well as being simple and agile. The proposed scheme had several advantaged for telecommunications operators, enterprise businesses and end users since it provided a new viable revenue generating model that can be sustainable over a long period. The proposed scheme was agile enough to offer the right level of service (SLA) required by the customer, and simple enough to attract a broad range of consumers from different sectors. However, the Green Cloud scheme has not addressed factors such as interaction

between the Clouds, physical network/operation required to enabler dynamic NaaS model, and appropriate pricing model. More importantly, the power consumption of the network between end-users and Clouds during the migration to the Cloud has not been considered.

In 2012, Gu et al. [11] proposed a scheme to save energy in mobile Cloud applications called GMoCA (Green Mobile Cloud Application). This scheme was proposed to prolong the lifetime of the battery in mobile devices which is an important issue for the mobile end-users. The proposal was based on migrating expensive computational tasks to the Cloud and offloading them from thin and mobile devices to powerful and shared devices on the Cloud data centers. However, similar to the previous works, this study did not include energy consumption of transportation between end-user devices and Cloud data centers.

As discussed, most of the previous studies [11, 12, 13] that introduced Cloud computing as a "Green" technology only considered energy consumption of data centers. The rationale for this is that data centers are generally optimized for energy efficiency and migration of applications to the Cloud permits replacing high-power desktop computers by low-power Netbooks and Tablets. Further, the computing resource in data centers is often shared by several users, in contrast to a single user using a desktop computer. Although intuitively reasonable, the above argument ignores two key factors:

a) Energy required to transport data between the end-users and the Cloud;

b) Additional power incurred by the end-user devices when accessing the Cloud.

Although it is crucial to improve energy consumption within data centers that host Cloud computing applications and 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 devices.

In 2011, Baliga et al. [2] studied energy consumption of Cloud computing using a network-based model and revealed that as the data rate between end-users and the Cloud increases, transport energy becomes a dominant fraction of the total energy consumption of Cloud computing.

To obtain a clear picture of the total energy consumption of Cloud computing applications and understand the potential role of Cloud computing to provide energy savings, a more comprehensive analysis is required.

It is worth mentioning that the studies we surveyed here only considered energy consumption in the use-phase of Cloud applications and services however there are other studies such as [21, 22] that considered the environmental footprint of products or services along their entire lifetime not only during its use-phase. LCA (life-cycle assessment) is not within the scope of this study.

2.2 Energy Consumption of Distributed Servers

The Internet services consist of various types of content objects, such as text, image, audio and videos. Traditionally, content objects have been fetched to end-users directly from the origins of their content providers, which are typically in storage servers located in centralized mega data centers. However, hosting content objects in a centralized style is not always the ideal solution because it may lead to undesirable results in user-perceived performance, network reliability and content delivery efficiency, especially under heavy network load [34]. Consequently, Content distribution/delivery network (CDN) was introduced in 2002 to complement centralized data centers by delivering contents to end-users efficiently and reliably [10].

A CDN is based upon content servers located in various places in the telecommunication network. When an end-user requests content, the CDN appoints a server for replying to the end-user request to improve end-user satisfactions in terms of delay and throughput. The current design of CDNs can be differentiated based on the location of content servers as well as the size of servers. In the next subsections, three content distribution strategies which are located in different sites of the Internet network as shown in Figure 2.2 are described. These are:

- Servers located in the core of the Internet network (backbones);
- Servers located in the edge of the Internet network (ISP points of presence (PoPs));
- Servers located in the end-user premises [3].

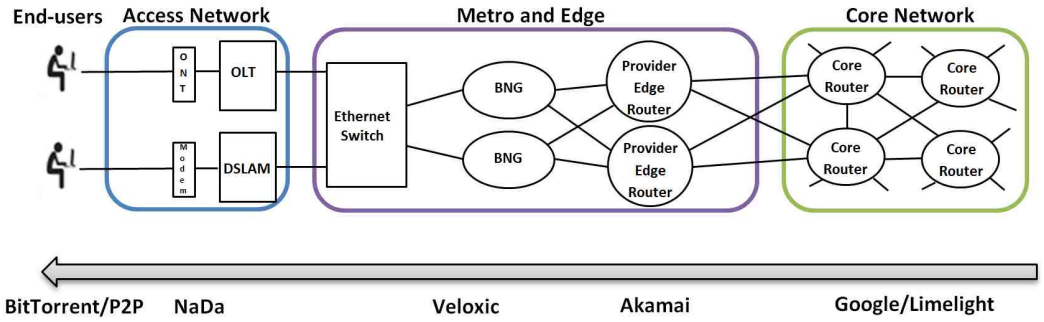


Figure 2.2: Comparison of various content dissemination methods [3].

Content distribution from the core of network

Several content service providers such as Google and Amazon as well as commercial CDN solutions such as Limelight use large server farms located in the core of the network to distribute content and applications. In this strategy servers are located in large data centers at a few strategic locations close to the PoPs of many large ISPs and the data centers are interconnected using private high-speed links (i.e., they tend to bypass tier 1 ISPs and have much smaller tier 1 hop counts) [3, 35, 36].

Content distribution from the edge of network

This proposal distributes a large number of small content servers across the Internet in multiple ISP PoPs located in the edge of the Internet network. This approach is used by hundreds of CDN operators worldwide, and the largest one among them is Akamai Technologies with over 170,000 content servers which are distributed in 102 countries in 2015, indicating the significant growth of the CDN industry [37].

In addition to CDNs by big companies such as Akamai and Limelight, Internet content providers also deploy their own ISP-level CDN solution in their networks (i.e., Velocix Digital Media Delivery Platform) in order to improve video stream quality and provide end-users faster file downloads by putting multimedia content closer to the subscribers. In this case, the content servers are located even closer to the end-users than highly distributed Akamai servers and it is more manageable to support optimized content delivery as ISPs have full knowledge of their networks [3].

Content distribution from the end-user premises

The third approach is based on peer-to-peer (P2P) content distribution from highly distributed and very small content servers located in the end-user premises [24, 25, 4, 26]. The P2P content distribution is used for file sharing applications (i.e, BitTorrent and eMule), and P2P multimedia streaming application (i.e, PPLive, Joost and Zattoo). In addition, the approach of content distribution from end-user premises is generalized to be used for other Cloud applications and services in the concept of nano data center. Nano data center is a distributed P2P content distribution platform based on very small servers located in end-user homes [24, 4, 3]. This concept is also known as in-house Cloud, personal local Cloud [38], and "Fog Computing" [23]. A real example of a simple personal local data center is Samsung Homesync which is currently available in the market [39].

For file sharing in P2P content distribution, the file (or content) is divided into a number of small pieces, and peers cooperatively share their available pieces. A tracker in BitTorrent, or a distributed hash table (DHT) such as Kademlia DHT in eMule can help for file sharing to avoid retrieving the same content from different cooperative peers. Nano data center introduced in [24, 4] uses BitTorrent-like file sharing over nano servers in home gateways. However, the key difference is that the nano servers are managed and coordinated through an ISP. Therefore, the managed P2P system in nano data centers overcome problems such as free-riding, node dynamics, and lack of awareness of underlying network conditions [24, 4, 3].

In the following subsections, we will study the existing work in energy consumption of distributed servers in the edge network and end-user premises. We will not survey energy consumption of servers located in the core network since their energy consumption might be relatively similar to the Cloud data centers discussed in Section 2.1.

2.2.1 Energy consumption of distributed servers located in the edge network

Since a typical CDN located in the edge network often comprises thousands of distributed servers globally [37], the energy cost of operating such an Internet-scale distributed system is substantial. Therefore, energy consumption studies of CDNs have recently gained

significant attention. In this section, we survey the proposed energy-saving schemes in CDNs located in the edge network. The existing strategies in the literature are divided into two categories: dynamic provisioning, request management and caching.

Dynamic provisioning refers to provisioning of content servers and/or network links dynamically in order to map content requests to fewer servers and network paths. Request management refers to mapping content requests to servers. Caching refers to the location of content that can be stored.

Request management

In 2009, Qureshi et al. [40] studied financial costs (not energy) of geographically distributed servers and proposed a new method to reduce the content provider's electricity costs. This method took advantage of two facts: fluctuations in electricity prices in various geographical locations and large distributed systems incorporate request routing and replication.

The proposed method strategically mapped content requests to servers with the minimum electricity cost. The cost was optimized every hour with no knowledge of the future. This rate of cost changing was slow enough to be tuned with existing routing methods but fast enough to respond to electricity cost variations. Content was assumed to be fully replicated on each server. This proposed method is subjected to bandwidth and performance constraints.

A trace-driven simulation was performed with real-world hourly and daily energy prices and real workload trace collected from Akamai. The results of this work showed that electricity bill can reduced by at least 2% under an Akamai-like server distribution with Google-like server energy proportionality which leads to save millions of dollars in electricity costs each year. In order to achieve better performance, this method should be applied in conjunction with other power saving methods.

This method focused on cost, not energy. Therefore, it may cause increase in the overall energy consumption as the access to some cheap servers may require long transmission distance. It would be valuable to use the approach for reducing energy consumption and carbon footprint of the energy used [40, 10].

In 2014, Mathew et al. [41] studied how Internet-scale distributed systems can use

smart grid features such as demand response to propose a new technique for energy cost (dollar, not energy) reduction using smart grid.

There are two possible techniques for reducing energy usage in a distributed system in response to demand-response requests. One technique is to move a portion of the load to other sites and then shutting down a portion of the servers as studied in [40]. The technique in [40] only considered requests for "real-time" applications that need to be serviced immediately. However, there are some requests that do not need immediate response and can be delayed (called elastic load) such as background downloads of software updates by operating systems, distribution of OS-level, security patches and content pre-fetching for local caching. Therefore, another technique is to move load temporal by dimension (rather than spatially or geographically, as has been done in prior work [40]) in order to reduce energy costs for the applications for which real-time interaction is not required. In [41] these kinds of applications are studied.

An optimal offline algorithm for demand response was proposed and evaluated using production workloads from a commercial content delivery network using realistic electricity pricing models. The results showed that energy cost savings for up to 12% for time-of-use electricity pricing with 40% of the elastic load (not real-time load) and 6 hours of service delay. The energy cost savings could increase up to 32% for a peak demand pricing and to 23% considering a combination of time-of-use and demand pricing [41].

Dynamic provisioning and request management

The following researches have used dynamic provisioning strategy as well as request management strategy.

In 2010, Chiaraviglio et al. [42] proposed a new model called "GreenCoop" to reduce power consumption of content providers (CPs) and Internet service providers (ISPs) via dynamically provisioning servers and networking elements. The CP represents a set of servers placed in different locations and the ISP is the owner of a network infrastructure. In this model, CPs and ISPs cooperated to improve the total of power consumption of both CPs' content servers and ISPs' network switches and links.

The results were achieved through optimization modeling in which the objective function was the minimization of the total power consumed by the CP and the ISP and the con-

straint was user delay. The new proposed model was tested based on real ISP topologies and realistic power figures and values. It was observed that up to 71% of power consumption reduction compared to a non-energy aware model with the objective of minimal delay. Although the results of this work indicated a great opportunity to save power through co-operation between CPs and ISPs; to perform dynamic provisioning operations, the reality is that both the CPs and ISPs are not willing to share their sensible data with each other such as the network topology, servers' load or end-to-end traffic demand. Therefore, there is a room to improve the model considering less shared information between CPs and ISPs [42, 10].

In 2011, Chiaraviglio et al. [43] (the same author of [42]) worked on the improvement of "GreenCoop" model to be independent from the shared information between CPs and ISPs. The new model was based on optimization and a distributed algorithm with a dual decomposition. The results of this work showed that the proposed solution is very close to the optimal scenario, with a maximum power efficiency loss less than 17%. Although this model could address the limitation in [42], the computational complexity is very high and it is not feasible for use in practical CDN scenarios [43, 10].

In 2012, Ge et al. [44] proposed a novel approach to minimize power consumption of content servers in large-scale content distribution platforms across multiple ISP domains by putting servers to sleep mode without impacting on content service capability. Decision making on putting geographically distributed servers to the sleep mode is complicated since the decision is based not only on the service capability at the server side, but also availability of network resource to support end-to-end content delivery.

The novelty of this approach is the use of optimized determination of server sleeping mode reconfiguration and smart mapping of user requests to the remaining active servers. The proposed scheme aimed to dynamically optimize power consumption of the servers while strategically putting servers to sleep without violating service constraints on both the server side and the network side (similar to GreenCoop [42]). To achieve this, the scheme is based on the virtual network platform where content providers can lease link bandwidth resources from underlying ISP to provide end-to-end content delivery.

The authors formulated this problem as a nonlinear programming model that can be solved offline. In order to evaluate the proposed scheme, a simulation based on inter-

connected GEANT-Abilene network topologies was used. The results revealed that this scheme is able to run the minimum number of active servers under load capacity constraints. The results indicated that up to 62.2% of overall power consumption reduction when compared to the reference scheme without power awareness. However, the proposed scheme is offline which means it has the knowledge of future whereas in online operation, decisions are made at the current time without any knowledge of the future. Therefore, this scheme is not able to process continuous content requests [44, 10].

In 2012, Mathew et al. [45] proposed a mechanism to reduce energy consumption of CDNs by turning off distributed servers during periods of low load while considering user service level agreements (SLA) and limiting the frequency of on/off server transitions to reduce wear-and-tear on hardware reliability. This work was the first one that proposed an energy-aware load balancing mechanism for CDNs in two levels: global load balancing (across data centers) and local load balancing (within a data center). In this work both an online algorithm (where the load balancing algorithm works with the current time without any knowledge of the future load) and an offline algorithm (where the load balancing algorithm knows the entire load sequence) were developed for the optimization problem. In the online version, a pool of active spare servers was used to absorb increased request volume whereas in the offline version, dynamic programming was used to calculate optimal server provisioning in polynomial time. The mechanism was evaluated using real production workload traces collected over 25 days from 22 geographically distributed clusters across the US from a large commercial CDN.

The results showed that this mechanism can reduce the energy consumption of CDNs by more than 55% while guaranteeing a high level of availability that meets customer SLA requirements with an acceptable number of on/off transitions per server per day. In addition, it was revealed that having about 10% of the servers as hot spares helps absorb load spikes due to global flash crowds with little impact on availability SLAs [45, 10]. Although this work considered both global and local load balancing techniques, the server shutdown technique was applied within local load balancers. Hence, the author of this work proposed another technique in [46] to address this issue.

In 2014, Mathew et al. [46] proposed and evaluated a new technique known as “cluster shutdown” where an entire cluster of servers in a CDN, deployed within a data center, can

be turned off. Cluster shutdown deployed into a global load balancer is capable of moving all loads away from a cluster and shut down all servers within the cluster. It is worth mentioning that the technique is not able to turn off individual servers or a fraction of a cluster as this technique turns off whole clusters or leaves them entirely on. In contrast, the technique introduced in the previous work [45] could shutdown individual servers within the cluster depending on the load. Therefore, these two techniques can complement each other and be implemented together to save energy.

The experimental results using extensive real-world traces from a large commercial CDN showed that the cluster shutdown technique can reduce the system-wide energy usage by 67% in the optimal case. It was also observed that the technique worked very well even for one shutdown per day for each cluster [46].

Dynamic provisioning, request management and caching

In 2010, Xu et al. [47] studied energy consumption of CDN for video distribution. First, the energy efficiency aspect of a video CDN system was investigated by defining and analyzing the energy efficiency of a video CDN system. Then, the authors proposed two theoretical schemes for energy efficiency improvements based on the idea of smart caching algorithms and coordination among video servers located in the edge of the network.

The first scheme improved the ratio of availability of a requested content object in local servers (known as "local hit ratio") without the need to fetch content from non-local servers. In this case, the local hit ratio is improved by pooling and sharing content caches among servers within a cluster. The results in this work revealed that a distributed video CDN scenario consumes more power than a centralized scenario. However, any improvements in the local hit ratio will help to save power in the distributed CDN scenario. The second scheme was to assign requests to less distributed servers and putting the idle servers to sleep mode, which is also a common strategy for power-saving in data centers. All results of this work are purely based on theoretical analyses and the results are yet to be evaluated through real data and practical experiments [47, 10].

In 2013, Llorca et al. [48] proposed an energy-efficient dynamic caching solution that pushes content objects towards interested users considering the minimum energy configuration. The energy efficient dynamic in-network caching solution aims to minimize

overall energy consumption using caching techniques and decreasing the transport network while meeting user requests. The simulation results of this analytical framework showed the potential of the proposed dynamic energy-aware network configuration solutions for significantly reducing energy consumption in content delivery networks. All results of this work are based on an analytical framework and the results are required to be evaluated through real data and practical experiments [48].

2.2.2 Energy consumption of distributed servers located in end-user premises

The extreme case of distributed servers are small servers located in end-user premises. Despite the growing popularity of these highly distributed servers for some applications, their energy consumption has not yet been well-studied. There are two main reasons why the energy consumption of servers for P2P applications has not yet been well-investigated:

- First, the power consumption of the P2P applications in a single location is not high;
- Second, no central entity pays for the power consumption of P2P applications like BitTorrent.

However, there are a few works that have studied energy/power consumption of P2P distributed servers:

In 2008, Nedevschi et al. [25] proposed a model for estimating energy consumption of a P2P architecture (i.e. BitTorrent) and compared the energy consumed by the P2P architecture to its centralized counterpart (i.e. iTunes). This work is one of the leading attempts to model the energy consumption of networked systems running from end-user premises and their initial exploration gave valuable insights for further research in this topic. The P2P system in [25] is formed by a set of PCs (representative of peers) located in end-user premises. In this model only incremental energy consumption (without considering idle energy consumption) of peers when running a P2P application is considered since the authors assumed only powered-on peers participate in the P2P application. Then, the estimated energy of peers was compared with the total energy consumed (both idle and incremental energy consumption) of a server located in a centralized data center.

The results of this work showed that P2P systems consume less energy compared to its centralized counterpart [25]. This work suffers from the following issues:

- Peers are modeled by PCs which are highly energy-inefficient.
- The energy consumption of the access network attached to the home servers (PCs) in the P2P scenario is not taken into account.
- The length of a path between an end-user and a requested content in the P2P scenario is assumed to be longer compared to a path in the centralized scenario whereas there are cases in which a P2P path is shorter.
- The incremental energy consumption of peers is compared to the total energy consumption of centralized servers which is not an appropriate comparison.

In 2010, Valancius et al. [4] proposed an energy consumption model for video-on-demand (VoD) running from nano data center architecture. Nano data center is a highly distributed architecture providing computing and storage services in a P2P fashion. The consumed energy of VoD services in a nano data center was compared to the consumed energy in the traditional centralized data centers. The P2P system in [4] was formed by small storage equipment attached to the access network (home gateway) of end-user premises without processing capability. The key idea of this architecture is to establish a manageable P2P system by storing content locally in end-user premises and enabling services on home gateways. A request for a content is managed and processed using a content server (called tracker) located in the end-users' ISP as shown in Figure 2.3.

The proposed model was evaluated using simulation based experiments and a large set of empirical VoD access data. The results showed that the nano data center approach saves at least 20% to 30% of the energy compared to traditional data centers due to reducing the energy consumption of transport network by bringing the content closer to the end-users, avoiding the cooling cost of a data center and avoiding device over-provisioning [4].

However, [4] is subject to the following limitations:

- The nano servers are small storage equipment (a flash memory or hard disk) attached to the access network (home gateway) of end-user premises without processing capability. However, a nano server is expected do processing in addition to

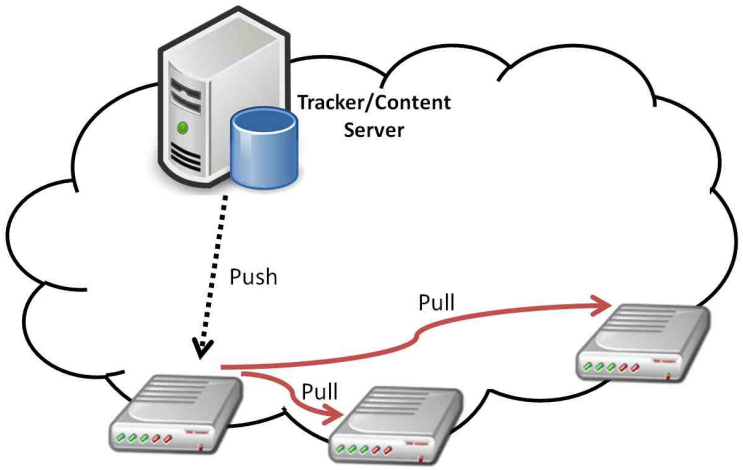


Figure 2.3: High level NaDa architecture. Content is served from home gateways whenever possible [4].

storage. Therefore, the energy consumption of nano servers with processing capability must be taken into account.

- ADSL2+ is the only access technology which was studied in this work although there are several access technologies for end-user premises such as PON, Ethernet, and 3G/4G.

- In the nano data center approach, a content server (Tracker) was used to provide a managed P2P system. However, the consumed energy by the Tracker located in the end-users' ISPs has not been taken into account in estimating the total energy consumption of nano data centers.

- Similar to [25], the incremental energy consumption of nano servers is compared to the total energy consumption of centralized servers which is not an appropriate comparison.

While these two works [25, 4] show the content distribution from end-users' premises can save energy compared to the centralized approach, Baliga et al. in [49] and Feldmann et al. in [26] present different results.

In 2009, Baliga et al. [49] studied energy consumption of video distribution and video delivered by Internet protocol TV (IPTV) from centralized data centers and P2P architec-

tures. The P2P network is formed using set-top boxes attached to the DSL access network located in end-user premises. The results showed that centralized data centers using optical bypass consume less energy (at least three times) compared to P2P for high demand content. However, P2P consumes less energy for movies that are downloaded less than once every few days. It is notable that users are responsible for the cost of energy used for storage in the P2P networks [49].

Similar to [49], Feldmann et al. [26] in 2010 studied the energy consumption of P2P architectures, centralized data center architectures and CDNs by proposing an energy consumption model which includes the transport network and data centers. The results show that a CDN within an ISP consumes less power compared to the two other architectures. It also indicated that although a P2P architecture may reduce the power consumption of the service provider (ISP), it increases the overall energy consumption because data has to cross over the access network twice [26].

These two work subject to the following limitations:

- Various types of access networks are required to be considered.
- Energy consumption of other type of services is needed to be studied as [49] and [26] are only limited to IPTV services and video distribution.
- Set-top-boxes located in end-users premises are considered as home servers. However, comparing the energy consumption of a device without processing capability (i.e. set-top-boxes) with the total energy consumed by a server within a data center is not appropriate. Therefore, more intelligent devices should be studied as home servers.

In 2011, Lee et al. [3] studied energy efficiency of network elements in various part of the Internet network including the core network, the edge network, the access network and end-user premises. The result showed that the ratio "Watt/Gbps" of network elements increases rapidly from core network towards end-users. The results showed power consumption of content distribution from end-user premises is higher than content distributions from the core network since home gateways in nano data center architecture and PCs (representative of home servers) consume 100 and 1000 times more power compared

to core routers respectively under the same traffic load. The authors in [3] proposed a content-centric networking (CCN) architecture help to save power consumption [35, 50]. However, CCN is not in the scope of our work.

In 2013, Mandal et al. [51] presented energy-consumption models and content-placement techniques for reducing energy consumption of content distribution from various locations of the telecommunication network. The authors considered three components to obtain energy consumption of a CDN: transmission energy; storage energy; and energy consumed by heating, ventilation, air conditioning (HVAC). The first two components were the focus of [51] since decreasing transmission and storage energy will generally decrease the HVAC energy. There is a trade-off between energy consumption of data transmission and storage. As an example, by bringing more copies of content closer to end-users the transmission energy can be saved but the storage energy increases since more storage devices are required to host the content copies. Therefore, [51] proposed the appropriate content-placement techniques to balance storage and transmission energy for reducing the overall CDN energy consumption. Two content-placement techniques were introduced: (1) static content-placement which is a simple and fast algorithm based on popularity-aware content replication (PACR) for static traffic and (2) dynamic content-placement which utilizes time-varying traffic irregularities of content-based services. The results of mathematical formulations of the proposed techniques showed that CDNs can save energy by utilizing the difference between peak and off-peak network usage since network elements are generally designed for peak load but their usage varies with the time of the day and with social events. Therefore, shutting down some network elements during off-peak hours and activating them during high traffic reduce total energy consumption [51].

However, the following factors merit further investigation:

- Practical experiments need to evaluate the proposed technique for content-placement.
- Various types of access networks need to be considered.
- Energy consumption of other network elements as home servers rather than set-top-boxes need to be investigated.
- Another important factor is the energy consumption for pre-loading (or updating) and storing multiple copies of the content which is ignored by these studies.

2.3 Conclusions

In this section, we first presented a survey on the existing research on energy consumption of Cloud based services and applications. The knowledge gaps in this field were identified and this survey showed that in order to obtain a clear picture of the total energy consumption of Cloud computing applications and understand the potential role of Cloud computing to provide energy savings, a more comprehensive analysis is required. In this context, the work in this thesis differs from the previous work in five significant ways:

a) Comprehensive and advanced energy/power consumption models for network equipment are presented and applied;

b) Packet traffic and power measurements when using existing interactive Cloud based applications are used and applied to the proposed models to estimate the power/energy consumption involved in accessing the Cloud applications;

c) Power/energy consumed in the transport network and end-user devices are taken into account;

d) Power/energy consumption of various access network technologies (Ethernet, WiFi, 3G/4G) are studied in our calculations;

e) All the models and measurement techniques are applied in current well-known Cloud applications such as Facebook, Google Drive, Microsoft OneDrive;

The energy/power consumption models are described in Chapter 3. The energy consumption of Facebook is discussed in Chapter 4 and the consumed energy by Google Drive and Microsoft OneDrive is explained in Chapter 5.

We then provided a survey on the previous work on energy consumption of distributed servers located in the edge of the network. The previous work studied strategies such as dynamic provisioning, request management and caching to reduce energy consumption of servers in the edge network. The literature showed that while optimizing the energy consumption of servers located in the edge network is well investigated, improving energy

consumption among servers located in end-user premises received comparatively less attention. Therefore, we focused on the servers located in end-user premises to identify the knowledge gaps in this area.

There are different points of view on energy consumption of application distribution from end-user premises (Fog computing) in the literature. For example, [25] and [4] claim that this solution is more energy-efficient than sharing videos from centralized DCs. However, other works [26, 49] show that P2P content distribution from end-user premises consumes more energy than the centralized solution. This difference is largely due to different models for equipment energy consumption in different research work. In addition, some studies have either ignored the transport network or used an overly simple model of the transport network.Therefore, in this work we study energy consumption of applications provided by nano servers located in end-user premises considering following items:

a) Advanced measurement-based energy/power consumption models for shared and unshared network equipment are used in this work;

b) Practical experiments are conducted using devices such as Raspberry Pis, representative of home servers in end-user premises;

c) Total energy consumption of applications provided by nano server are studied which includes energy consumed in end-user terminal, the transport network and nano servers.

d) Energy consumption of different access network technologies (PON, Ethernet, WiFi, 3G/4G) are considered;

e) The energy modelings and measurement techniques in this work is not specified to a particular application such as IPTV or VoD and can be used for various applications.

The energy/power consumption models are explained in Chapter 3 and the energy consumption of Fog computing applications and services provided by servers located in end-user premises is discussed in Chapter 6.

Chapter 3

Energy Consumption Modeling

Executive Summary

Simple, generic and measurement-based energy/power consumption models are described and applied to equipment, networks and services. These models are used to construct power and energy consumption estimates for a diverse range of network scenarios including customer premises equipment and access, edge and core networks and services provided over a network.

The models in this chapter are used in Chapter 4, 5 and 6 for modeling energy/power consumption of Cloud computing and Fog Computing applications.

3.1 Introduction

The rapid growth of the information and communication technology (ICT) industry has increased the concerns of energy consumption and carbon footprint in ICT [52, 53]. The industry is responding to the concern of the ICT energy consumption and carbon emissions by estimating the environmental impact of products and services provided by companies [54, 55, 56]. In order to estimate the energy consumption of ICT and Internet based services, an estimation of the electrical energy consumption of the associated network and service infrastructure is required. It is not feasible to measure power consumption of most telecommunications networks and the Internet directly since they are too large and most of the network equipment are not readily accessible for measurement. Accordingly, a combination of measurement and modeling has been adopted to ascertain energy/power consumption of telecommunication networks [57].

A range of modeling approaches have been applied over the years [57]. Some researchers have taken a "top-down" approach in which equipment inventories across a region (worldwide or a national) are used to estimate network power consumption [58]. Another "top-down" approach uses network power reports from telecommunications providers to extrapolate to total network power [59]. Others have developed "bottom-up" models typically based on simplified network design rules which are used to determine the amount of equipment required to support a given traffic level [60, 61]. Models can also be based on the availability of "coarse-grain" and "fine-grain" network information [62]. Many of the "bottom-up"' models (particularly for core and edge networks) are based on estimates of network energy consumption using a "Joules/bit" approach [60, 63].

This chapter provides a concise overview of concepts and techniques that can be applied to develop a general "bottom-up" power and energy consumption models of telecommunications networks and services. This work is different to previous work because it covers the full "service eco-system" in that previous work has typically focused on a specific component of the service. For example, much work has been done on the the data center. This is the work that develops models for each component of the overall eco-system. In particular a new model for the unshared (user) equipment (i.e. the time based model) and the flow based model for shared equipment. A similar, but not identical,

idea of time-based and flow-based models was considered in [64, 65, 66] independently at the same time.

When analyzing networks a standard approach is to segment the network into the access, metro/edge and core networks as depicted in Figure 3.1. Models of the power consumption of large networks such as the Internet often adopt this approach [60, 6, 67].

There are multiple factors that influence the type of model best suited to assessing the energy consumption of a network or service. For estimating the energy consumption of a network, we would typically adopt models that apply to the total network equipment. For estimating the energy consumption of a service, we need a model that selects out the specific traffic flow for that service.

For the purposes of energy modeling, there are two key factors that distinguish between Customer Premises Equipment (CPE) and access equipment on one hand and edge and core equipment:

a) CPE is typically shared by only one or a few users whereas access, edge and core equipment are shared by increasingly many users.

b) Access, edge and core network equipment are typically specialized machines that undertake a single activity; dealing with packets. Therefore, all their power consumption can be totally allocated to that one activity. In contrast, customer premise equipment may be undertaking multiple tasks apart from communicating with the Internet.

With unshared or lightly shared equipment, the usual approach is to construct a model based on "power per user" [68, 69]. Further, customer equipment such as a laptop may be running several applications, only one or two of which are providing the user access to a network or internet based service. In this case, a further "time-base" resolved model is proposed in this chapter [28].

For highly shared equipment, we have used another approach because it is be very difficult to keep track of the time duration of a highly shared equipment allocates to many users and services. Therefore, with single-function, heavily shared equipment such as

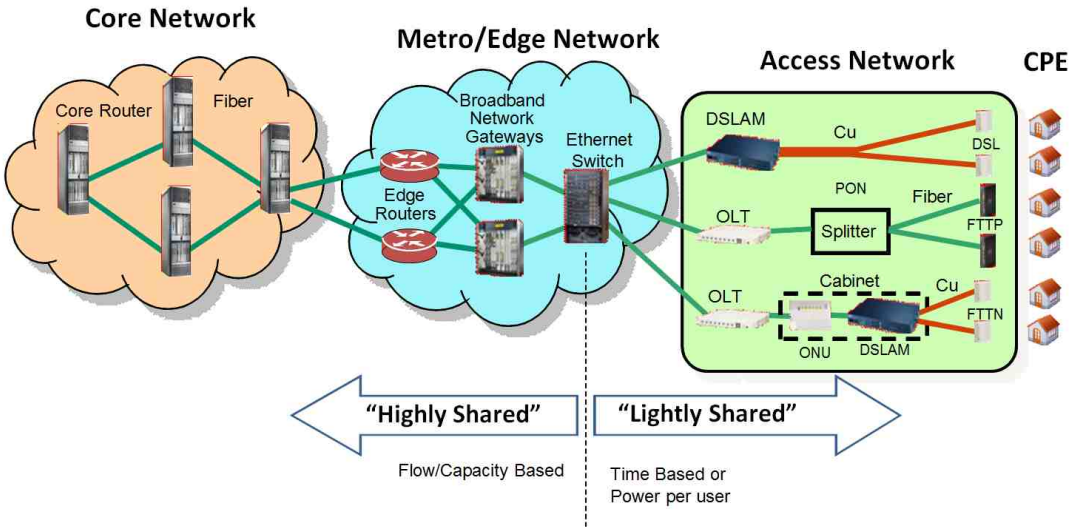


Figure 3.1: Simplified network model. The type of power model depends upon how "shared" the equipment is. For access network equipment that is shared amongst relatively few users, a "time-based" or "power per user" model is typically adopted. For edge and core equipment that is shared over many users, a "flow-based" or "capacity-based" model is typically adopted.

routers and switches, a "flow-based" model is proposed in this thesis which is more practical and uses the machine throughput in bits per second [60, 67].

We need to select a dividing point to delineate where these two models are applied. One option for locating the split between the use of "time-based" model (for unshared equipment) and "flow-based"' model (for shared equipment) is shown in Figure 3.1. In the following sections we will introduce and investigate each of these model types.

A key driver for this work is to give an intuitive introduction to telecommunications network and service power and energy consumption modeling. In addition, since access to detailed network equipment information (such as accurate power consumption and traffic data and detailed network architecture information) for large commercial networks is typically very difficult to acquire, the model provided here is based on typical equipment data (available from vendor equipment data) and simple network architecture parameters (such as number of hops).

Since this field is relatively young (the first publication of a network based energy model of the Internet was Baliga et al. in 2009 [60]), many areas of research are yet to

be developed. In particular the primary focus on ICT energy calculations have focused on networks and equipment. Only a handful of publications have looked at services (e.g. C. Chan et al. in [70]). Further, no estimates have been published for the energy consumption of services such as Facebook, interactive cloud services and the like, despite their growing popularity. This work is the first to undertake this task.

3.2 Highly Shared Network Equipment

For equipment shared by many users and services such as routers and switches in the core and edge of the network, the assessment of the energy consumption is based upon proportional allocation of the equipment's power consumption over all the flows through the equipment.

Ideally an estimate of the energy consumption of a service would identify all the traffic (i.e. bits) that provide the service and then undertake a calculation, based on that identification, to determine the energy of that service. Although this is feasible for unshared equipment (and is the basis of the time based model) it is not feasible in shared equipment. This is because shared equipment will carry the traffic of many customers accessing a diversity of services. Trying to identify which bits are for which service in, say, a core router is not realistic. Therefore we adopt a different approach. We identify the traffic flow (bits/sec) a customer will require to access a service and proportionally allocate the power consumption of the shared equipment to that service based on that flow. This means we use a (joules/bit) x (bits/sec) based model which is the flow based model. The location of the equipment in the network where we apply this model is set by the first piece of equipment in which we decide it is too difficult to track the "bit by bit" use of that equipment by a customer for the service of interest. This is typically the first large scale aggregation point in the local exchange. For example, the aggregation Ethernet switch after the OLT in the exchange. We may be able to allocate power in an OLT based on the traffic flow from individual users, but it is highly unlikely we can do so for the aggregation switch. There will be too many customers being serviced by that switch. In fact, we could adopt a approach in which the output ports of the OLT are too shared for a time based model and we would then apply a time based model to the input ports and a flow based

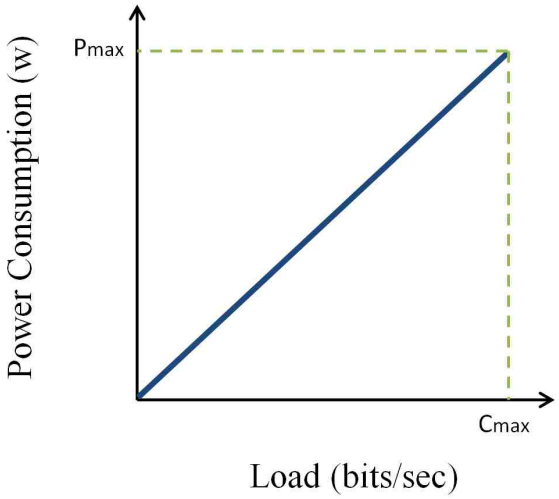


Figure 3.2: Power consumption profile of network equipment such as routers and switches without considering idle power

model to the core of the OLT and its output ports. This is likely to be too complex, so we apply the flow based model to the aggregation switch.

3.2.1 Power and energy consumption

There have been many publications that study the power consumption properties of shared network equipment such as routers, switches and servers [71, 72, 73, 74, 75]. Overall, these studies show that there is an (approximately) affine relationship between power consumption, $P(t)$ (Watts), and load/throughput, $C(t)$ (bit/sec) for most of network equipment such as routers, switches, servers and the equipment in the access network. The energy consumption of these network equipment is studied under a specific amount of load and the results indicate that the power consumption of the network equipment increases when the load increases.

In early studies such as [2, 60], the power consumption profile was modeled as shown in Figure 3.2. This profile is expressed by:

$$P(t) = \frac{P_{\max}}{C_{\max}} C(t) \tag{3.1}$$

where,

$P_{\max}$ is the maximum power consumption of the network equipment; $C_{\max}$ is the maximum load/throughput that the network equipment can handle.

However, more recent studies show that many network equipment such as routers, switches and servers consume some power even when there is no traffic load on them. This idle power consumption, P_{idle}, can be a large fraction of the maximum power consumption, $P_{\max}$, of the device ($P_{idle} \approx 60\% - 95\%$ of $P_{\max}$) [2, 30, 3]. Furthermore, the power consumption of the device increases when the traffic load increases as shown schematically in Figure 3.3. This power profile can be expressed as:

$$P(t) = P_{\text{idle}} + \frac{P_{\max} - P_{\text{idle}}}{C_{\max}} C(t) = P_{\text{idle}} + E_{\text{b-inc}} C(t) \tag{3.2}$$

where,

P_{idle} is the idle power of the equipment which corresponds to the power consumption of the equipment when the throughput is zero $C(t) = 0$;

$P_{\max}$ is the maximum power consumption that occurs when the throughput is at the maximum the equipment is designed to handle, $C_{\max}$.

$E_{\text{b-inc}} = \frac{P_{\max} - P_{\text{idle}}}{C_{\max}}$ is the incremental energy per bit of the network equipment.

This linear profile has been validated by experimental results published in [74]. The linear slope of $E_{\text{b-inc}}$ in Figure 3.3 has dimensions of Joules per bit. We shall refer to this slope as the "incremental energy per bit" which is expressed by:

$$E_{\text{b-inc}} = \frac{P_{\max} - P_{\text{idle}}}{C_{\max}} \tag{3.3}$$

For some network equipment $E_{\text{b-inc}}$ is equal or very close to zero (i.e. $P_{\max} \approx P_{\text{idle}}$). For a fixed configuration, many large routers have $P_{\text{idle}} \geq 0.8 P_{\max}$ (For examples see [73, 74]). For other equipment $E_{\text{b-inc}}$ may be much larger resulting in $P_{\max}$ being noticeably greater than P_{idle}. Examples of $P_{\max}$ is significantly larger than P_{idle} include mobile base-

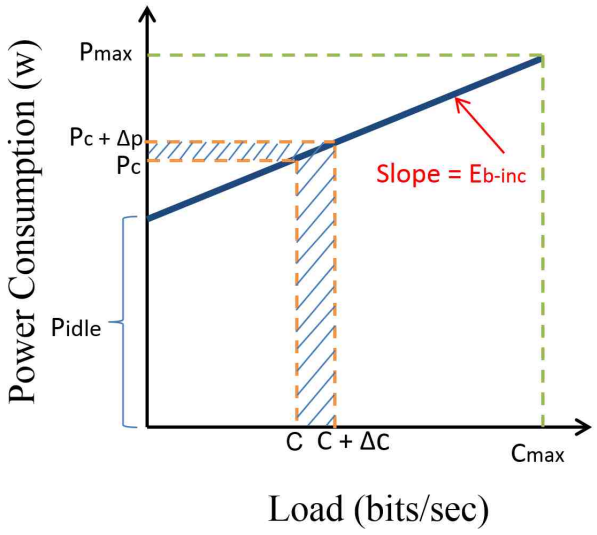


Figure 3.3: Typical power versus load characteristic for network equipment

stations [76, 6].

We refer to the additional power consumption of network equipment above the idle power consumption as the incremental power consumption of the network equipment. The incremental power consumption is given by:

$$\Delta P = E_{\text{b-inc}} \Delta C \tag{3.4}$$

where,
ΔC is the additional load for running a service on the network equipment as depicted in Figure 3.3.

Because the vast majority of network equipment has linear power profile [74], we use the same model for all the equipment in the network which are shared by multiple users and services. The cumulative power consumption of a network can be represented by a staircase curve as shown in Figure 3.4. Each step corresponds to the deployment of additional network equipment once the capacity per network equipment reaches the pre-set maximum operating load utilization, U.

To calculate the joules per bit for the additional traffic generated by a service that

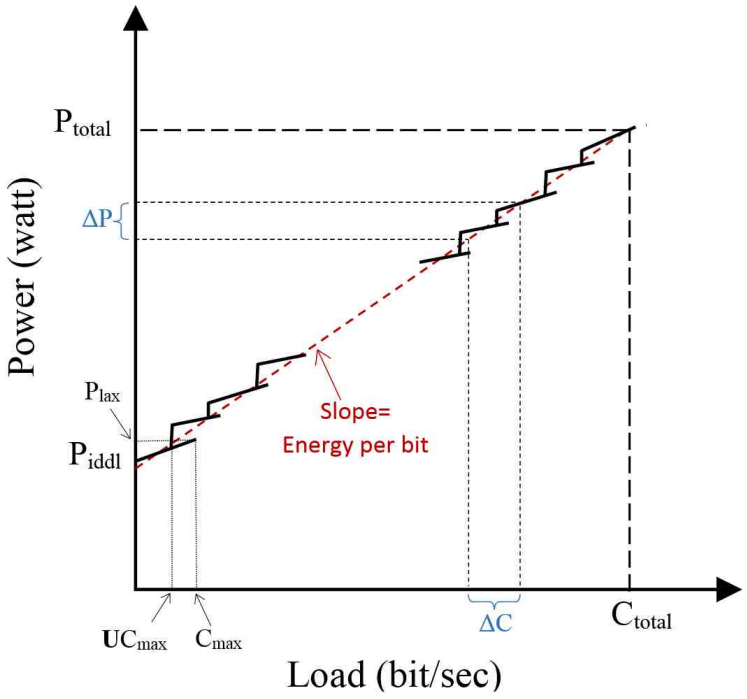


Figure 3.4: Power consumption trend under large-scale equipment deployment [5, 6]

is spread over many machines distributed across a network (such as Facebook photo-sharing), we can adopt the following approach.

We consider a network initially carrying total capacity C as shown in Figure 3.4. To this capacity the service under consideration adds incremental capacity ΔC. This incremental capacity is much greater than the average maximum capacity of the network elements used in the network. Therefore, if $C_{\max}$ is the average maximum capacity of the network elements, then $\Delta C \gg C_{\max}$. Typically, the network elements are not operated at their maximum capacity, they are operated at a fraction, U, of $C_{\max}$. Therefore, the operational capacity per machine is $UC_{\max}$.

We assume the service generates the additional capacity ΔC in the form of many small capacity increases roughly evenly distributed across the metro edge of the entire network. That is, $\Delta C = M\delta C$ where $M \ggg 1$and δC is relatively small (such as the size of a photo file). The parameter M represents the many millions of users who are each simultaneously uploading a file of average size δC.

The additional network capacity ΔC will be supported by the deployment of a number

of machines, N. The total power requirement of deploying this additional equipment to support the additional traffic generated by this service will be:

$$\Delta P = NP_{\text{idle}} + E'_{\text{b-inc}}\Delta C = NP_{\text{idle}} + \frac{P_{\text{total}} - P_{\text{idle}}}{C_{\text{total}}} \tag{3.5}$$

where,

P_{idle}, P_{total} and C_{total} are mean idle power of the additional machines, total power consumption of shared network equipment in a node and total capacity of network equipment in a node, respectively.

$E'_{\text{b-inc}}$ is the mean incremental energy/bit of these machines. This traffic is shared over many machines across the network, so we use a "flow based" approach to model its power consumption.

From this the energy per bit for the additional traffic generated by this service can be approximated by $\frac{\Delta P}{\Delta C}$. To calculate this we need to ascertain the number of machines, N. With the utilization of the machines U, then N is given by:

$$N = \lceil \frac{\Delta C}{UC_{\text{max}}} \rceil \tag{3.6}$$

Therefore, the energy per bit ($E_{\text{b-flow}}$) for traffic generated by this service is given by:

$$E_{\text{b-flow}} = \frac{\Delta P}{\Delta C} = \frac{P_{\text{idle}}}{UC_{\text{max}}} + E \tag{3.7}$$

It is notable that the incremental energy consumption described in this chapter refers to a consequential analysis of energy consumption in environmental sciences and the total energy consumption refers to attributional analysis in environmental sciences [77].

Parameters	**Description**
$P(t)$	Power consumption in time t (w)
$C(t)$	Load/throughput of a network equipment in time t (bps)
P_{max}	Maximum power consumption of a network equipment (w)
P_{idle}	Idle power consumption of a network equipment (w)
C_{max}	Maximum Load/throughput of a network equipment (bps)
$E_{\text{b-inc}}$	Incremental energy per bit (J)
ΔP	Additional power consumption of network equipment (w)
ΔC	Additional load for running a service (bps)
$E'_{\text{b-inc}}$	Mean incremental energy per bit (J/b)
C_{total}	Total capacity of network equipment in a node (bps)
P_{total}	Total power consumption of network equipment a node (w)
N	Average number of network equipment in one node
U	Load threshold
$E_{\text{b-flow}}$	Energy per bit for shared network equipment in a node (J/b)
$N_{\text{bit,k}}$	Number of exchanged bits in service k (byte)

Table 3.1: Notation in energy and power consumption model

3.2.2 Power consumption of a service

A shared network equipment (such as router, switch, optical line equipment located in the network beyond the first aggregation point as shown in the left side of Figure 3.1) typically deals simultaneously with traffic from multiple services. To calculate the power consumption of a specific service, we need to allocate a component of the total network equipment power, $P(t)$, to that service. Allocation of power due to the "incremental energy per bit" is intuitively given by $EC^{(k)}(t)$ where $C^{(k)}(t)$ is the traffic allocated to the $k-th$ service.

The value of $C^{(k)}(t)$ at time t is the traffic load to the $k-th$ service at time t which may or may not equal the actual service traffic at that time. For example, if service k has a "Service Level Agreement" (SLA) that requires a reservation of capacity even if there is no actual traffic, then $C^{(k)}(t)$ corresponds to the reserved capacity rather than the actual "bits/sec" of that service traffic at time t (For example, a protection path may be reserved for a service).

This approach is adopted because the reservation of capacity through a network will incur power consumption of network resources dedicated to providing that reserved capacity even when $C^{(k)}(t)$ is zero. In contrast, a service which has no reservation (e.g. a "best effort" service) requires less dedicated network resources and hence less power. Intuitively we would expect the carbon footprint of services requiring less network resources will be less than that of a service requiring more. We distinguish between "reserved capacity service" in which capacity is allocated to the service. This is sometimes called a "tunnel" because the service is guaranteed a certain amount of capacity. This capacity is "pre-allocated" and reserved, even if the actual traffic of the service is zero. This is a common approach to service provision, and is often referred to as "Committed Information Rate" (CIR). A best effort service typically does not have a CIR (or has a CIR =0) and there is no guarantee of date delivery. This is because the available resources may be otherwise allocated.

We also need to allocate a component of the idle power, P_{idle}, across the multiple services being handled by shared network equipment. The simplest approach is to allocate the idle power in proportion to the total traffic through the machine. We use a notation in

which the superscript in parenthesis identifies the service and a subscript without parenthesis identifies the shared network equipment. Therefore, the power consumption $P_{\mathrm{j}}^{(k)}(t)$ $P_{\mathrm{j}}^{(k)}(t)$ of the $k-th$ service that is being handled by the $j-th$ network equipment is given by:

$$P_{\mathrm{j}}^{(k)}(t) = \frac{P_{\mathrm{idle,j}}}{C_{\mathrm{j}}(t)} C_{\mathrm{j}}^{(k)}(t) + E_{\mathrm{b\text{-}j}} C_{\mathrm{j}}^{(k)}(t) \tag{3.8}$$

where,

$C_{\mathrm{j}}(t)$ is the total throughput of the $j-th$ network equipment (bps);

$P_{\mathrm{idle,j}}$ is the idle power consumption of the $j-th$ network equipment (w);

$E_{\mathrm{b\text{-}j}}$ is the incremental energy per bit of the $j-th$ network equipment (J/b);

Finally, $C_{\mathrm{j}}^{(k)}$ is the traffic of the $k-th$ service that is dealt with by the $j-th$ network equipment (bps). This approach allocates the proportion $\frac{C_{\mathrm{j}}^{(k)}(t)}{C_{\mathrm{j}}(t)}$ of the $j-th$ network equipment' idle power to service k.

3.2.3 Flow-based energy consumption model

For equipment shared by many users and services such as routers and switches in the core of the network which deal with high traffic volume we present a "flow-based" or "capacity-based" energy consumption model. For equipment in this part of the network, the measure of the energy consumption of the Cloud service is based upon proportional allocation of the equipment's power consumption over all the flows through the equipment. The energy consumption of service k, $E_{\mathrm{k\text{-}flow}}$, that uses a network path shared with many other traffic flows, is then approximated by:

$$E_{\mathrm{k\text{-}flow}} \approx m E_{\mathrm{b\text{-}flow}} N_{\mathrm{bit,k}} \tag{3.9}$$

where,

$E_{\mathrm{b\text{-}flow}}$ is the energy per bit of shared network equipment obtained in (3.7) (J/b);

$N_{\mathrm{bit,k}}$ is the number of exchanged bits of service k through the node by the service under

Parameters	Description
$C^{(k)}(t)$	Traffic allocated to the service k at time t (bps)
$C^{(k)}(t)$	Reserved capacity of the service k at time t (bps)
$P_{\mathrm{j}}^{(k)}(t)$	Power consumption of service k handled by network equipment j (w)
$C_{\mathrm{j}}(t)$	Total throughput of the network equipment j (bps)
$P_{\mathrm{idle,j}}$	Idle power consumption of the network equipment j (w)
$E_{\mathrm{b\text{-}j}}$	Incremental energy per bit of the network equipment j (J/bit)
$C_{\mathrm{j}}^{(k)}$	Traffic of service k that is dealt with network equipment j (bps)
$E_{\mathrm{k\text{-}flow}}$	Energy consumption of service k shared with many other traffic flows (J/bit)
$E'_{\mathrm{b\text{-}flow}}$	Energy per bit of shared network equipment (= E'_{b}) (J/bit)
m	Average number of network nodes in the service path

Table 3.2: Notation in service power consumption and flow-based energy consumption model

consideration;

m is the average number of network nodes in the service path.

3.3 Lightly Shared Network Equipment(CPE and Access)

3.3.1 Power per user model

The models described in this section apply to the "lightly shared" network equipment as depicted in Figure 3.1. The "power per user" model developed can represent continuous constant access network power in recognition of the fact that many users tend to leave their CPE permanently powered on and consume power independent of their traffic throughput [68, 69]. If we assume that the power of the equipment involved is constant and if we allocate the power to each user in inverse proportion to the number of users sharing that item of equipment. The resulting power per user for CPE and access network equipment is given by [68, 64]:

$$P_{user,Access} = P_{CPE} + X_{RN}\frac{P_{RN}}{N_{RN}} + X_{TU}\frac{P_{TU}}{N_{TU}} \tag{3.10}$$

In this equation:
P_{CPE} is the customer premise equipment power;
P_{RN} is the remote node power (such as a Digital Subscriber Line Access Multiplexer (DSLAM) or fiber splitter);
P_{TU} is the terminal unit power (before aggregation);
N_{RN} is the number of customers sharing the remote nodes;
N_{TU} is the number of terminal units;
X_{RN} and X_{TU} are factors representing any additional power consumption that may be required for environmental control of the facility housing the equipment, power supplies to the equipment etc. In many circumstances these factors are often referred to as the "Power Use Effectiveness" (PUE). Example values are given in [68, 5].

3.3.2 Time-based energy consumption model

For equipment located at end-user premises, such as home gateways and home servers (nano servers), which perform intermittent network access, we construct a "time-based" energy consumption model based upon the amount of time that equipment spends dealing

with the services of interest. Consider a device such as nano server in a home, a typical nano server's usage for serving services could be represented by the plot in Figure 3.5.

The device is active in serving the services of interest during times $t_l, l = 1, .., n$ (the pink areas) and not serving those services for times $T_l, l = 1, .., n$. The total time under consideration as shown in Figure 3.5 is:

$$T_{tot} = t_{idle} + t_{act} \tag{3.11}$$

where, the total idle time (t_{idle}) for the device is:

$$t_{idle} = \sum_{l=1}^{n} T_l \tag{3.12}$$

and the total active time (t_{act}) for all services is:

$$t_{act} = \sum_{l=1}^{n} t_l \tag{3.13}$$

Therefore, The energy consumption of the customer premises equipment (CPE) including the nano server is given by:

$$E_{\text{cpe}} = P_{\text{idle}} T_{\text{tot}} + \int_{t_{act}} (P(t) - P_{\text{idle}}) \mathrm{d}t \tag{3.14}$$

where, P_{idle} is power consumption of the device in the idle mode.
The incremental energy consumption of CPE is given by:

$$E_{\text{inc-cpe}} = \int_{t_{act}} (P(t) - P_{\text{idle}}) \mathrm{d}t \tag{3.15}$$

In this work we assume that the device can serve one or multiple services. Therefore, the active part of the device (pink areas) can represent one service or multiple services. To determine energy consumption of one specific service running on the device such as

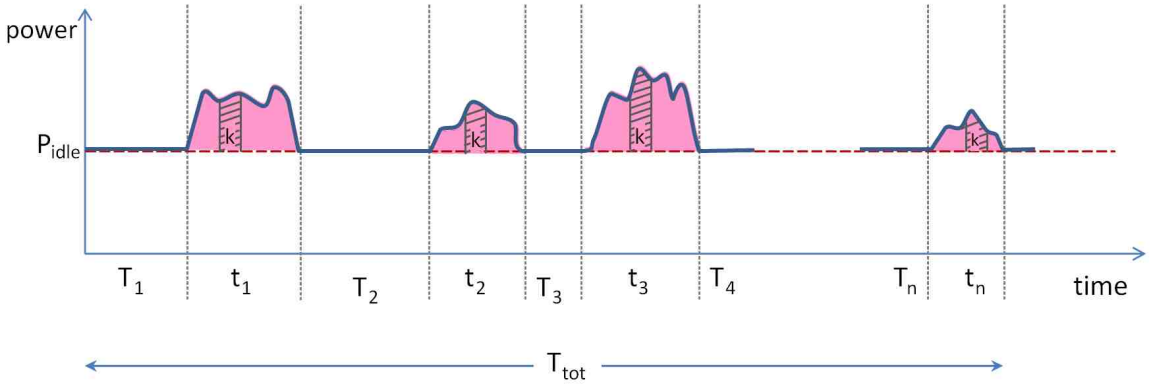


Figure 3.5: Power consumption of a home equipment unit for serving/accessing services

service k (the hatched area in Figure 3.5), two parts are considered:

a) incremental energy consumption due to running this specific service ($E_{\text{inc,k}}$);

b) idle time energy consumed for running the service ($E_{\text{idle,k}}$).

Our approach for allocating the idle power to the service k is proportional. This approach allocates the idle power based on active time of service k ($t_{\text{act,k}}$) to the total active time of the device (t_{act}). Therefore, the total energy consumption of the service over the duration T_{tot} is:

$$\begin{aligned} E_{\text{k-time}} &= E_{\text{idle,k}} + E_{\text{inc,k}} \\ &= E_{\text{idle}} \frac{t_{\text{act,k}}}{t_{\text{act}}} + \int_{t_{act,k}} (P(t) - P_{\text{idle}}) \mathrm{d}t \\ &\approx P_{\text{idle}} T_{\text{tot}} \frac{t_{\text{act,k}}}{t_{\text{act}}} + + \sum_{l} (\bar{P}_{\text{k,l}} - P_{\text{idle}}) t_{\text{act,k,l}} \end{aligned} \tag{3.16}$$

where,

$\bar{P}_{\text{k,l}} = \frac{1}{t_{\text{act,k,l}}} \int_{t_{\text{act,k,l}}} P(t) \mathrm{d}t.$

The data rate of the service during active times is the total exchanged bits ($N_{\text{bit,k}} = \sum_{l} N_{\text{bit,k,l}}$) divided by the total active time ($t_{\text{act,k}} = \sum_{l} t_{\text{act,k,l}}$) of the service which is $C = \frac{N_{\text{bit,k}}}{t_{\text{act,k}}} = \frac{N_{\text{bit,k,l}}}{t_{\text{act,k,l}}}$.

Hence we can re-write the above equation as:

$$E_{\text{k-time}} \approx P_{\text{idle}} T_{\text{tot}} \frac{t_{\text{act,k}}}{t_{\text{act}}} + \sum_{l} (\bar{P}_{\text{k,l}} - P_{\text{idle}}) \frac{N_{\text{bit,k,l}}}{C} \tag{3.17}$$

In the next subsection, we study the effect of idle time and active time of a nano server on the energy consumption of service k.

3.3.3 Ratio of idle time versus active time (α)

In order to show the energy consumption variations in time for a CPE such as a home server, we define the coefficient α which is the ratio of idle time of the equipment to the active time. The coefficient α is given by:

$$\alpha = \frac{t_{\text{idle}}}{t_{\text{act}}} \tag{3.18}$$

The range for α is $0 \leq \alpha \leq \frac{T_{tot} - t_{act,k}}{t_{act,k}}$.

In order to study energy consumption variations of a service depending on the values of α, four different values of α are studied.

(1) $t_{idle} = 0 \Rightarrow \alpha = 0$:

The first scenario is based on the CPE being fully utilized serving multiple services and therefore the idle time is zero ($t_{idle} = 0$). Hence the idle time energy consumption is zero ($t_{idle} = 0$) and $T_{tot} = t_{act}$ in (3.11) as shown in Figure 3.6. According to this assumption, the energy consumption of the service k which is obtained by (3.16) or (3.17) can be expressed by:

$$\begin{aligned} E_{\text{k-time}} &= P_{\text{idle}} t_{\text{act,k}} + \int_{t_{act,k}} (P(t) - P_{\text{idle}}) \mathrm{d}t \\ &= \int_{t_{act,k}} P(t) \mathrm{d}t \end{aligned} \tag{3.19}$$

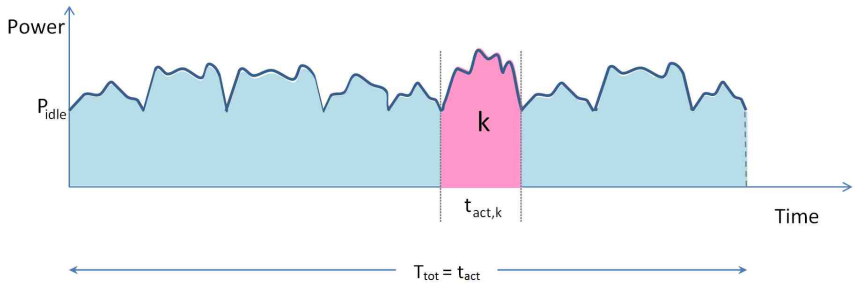


Figure 3.6: Power consumption of a nano server located in end-user premises serving multiple services fully utilized ($t_{idle} = 0$)

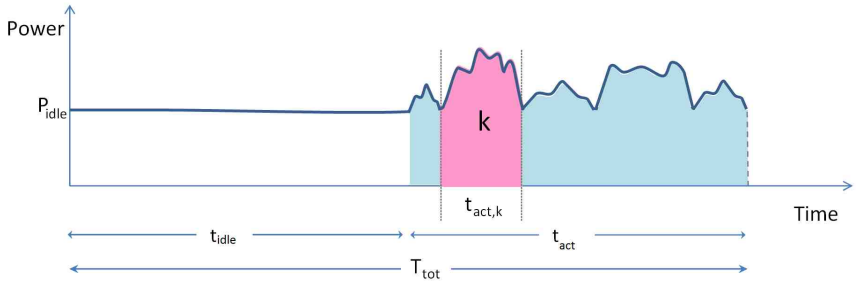


Figure 3.7: Power consumption of a nano server located in end-user premises serving multiple services but not fully utilized ($t_{idle} = t_{act}$)

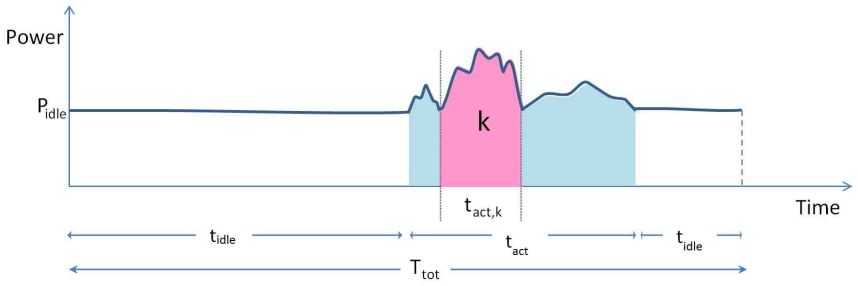


Figure 3.8: Power consumption of a nano server located in end-user premises serving multiple services but not fully utilized ($t_{idle} = 2t_{act}$)

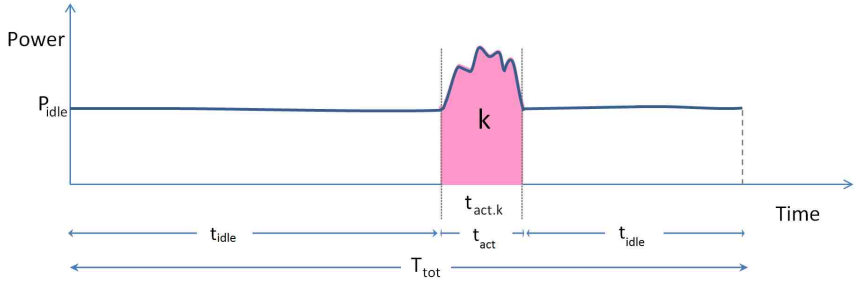


Figure 3.9: Power consumption of a nano server located in end-user premises serving only service K ($t_{idle} = T_{tot} - t_{act,k}$)

However, devices at end-user premises are not usually highly shared. Therefore, to study the effect of active time and idle time of unshared network equipment in end-user premises, consider the CPE with several different idle times ($t_{\text{idle}} = t_{\text{act}}, 2t_{\text{act}}, T_{\text{tot}} - t_{\text{act,k}}$).

(2) $t_{idle} = t_{act} \Rightarrow \alpha = 1$:

The second scenario is based on the CPE serves service k and other services but it is not fully utilized and the idle time is half of the total time ($t_{idle} = t_{act}$) as shown in Figure 3.7. Therefore, based on this assumption, the energy consumption of the service k which is obtained by (3.16) or (3.17) can be obtained by:

$$E_{\text{k-time}} = P_{\text{idle}} 2t_{\text{act,k}} + \int_{t_{act,k}} (P(t) - P_{\text{idle}}) \mathrm{d}t \tag{3.20}$$

(3) $t_{idle} = 2t_{act} \Rightarrow \alpha = 2$:

The next scenario is based on the CPE serves service k and other services but it is not fully utilized and the idle time is one third of the total time ($t_{idle} = 2t_{act}$) as shown in Figure 3.8. Therefore, based on this assumption, the energy consumption of the service k which is obtained by (3.16) or (3.17) can be obtained by:

$$E_{\text{k-time}} = P_{\text{idle}} 3t_{\text{act,k}} + \int_{t_{act,k}} (P(t) - P_{\text{idle}}) \mathrm{d}t \tag{3.21}$$

(4) $t_{idle} = T_{tot} - t_{act,k} \Rightarrow \alpha = \frac{T_{tot} - t_{act,k}}{t_{act,k}}$:

As shown in Figure 3.9, the maximum value of the idle time is when the equipment runs only service k. The idle time is the total time minus the time dedicated to the service k ($t_{idle} = T_{tot} - t_{act,k}$). Therefore, the energy consumption of the service k is given by:

$$E_{\text{k-time}} = P_{\text{idle}} \left(\frac{T_{tot} - t_{act,k}}{t_{act,k}} + 1 \right) t_{\text{act,k}} + \int_{t_{act,k}} (P(t) - P_{\text{idle}}) \mathrm{d}t. \tag{3.22}$$

Consequently, the range for α is $0 \leq \alpha \leq \frac{T_{tot} - t_{act,k}}{t_{act,k}}$ and (3.17) can be rewrite based on α as:

$$E_{\text{k-time}} = P_{\text{idle}} (\alpha + 1) t_{\text{act,k}} + \int_{t_{act,k}} (P(t) - P_{\text{idle}}) \mathrm{d}t. \tag{3.23}$$

Parameters	Description
P_{CPE}	Power consumption of customer premise equipment (CPE) (w)
P_{RN}	Power consumption of s remote node (such as a DSLAM or fiber splitter) (w)
P_{TU}	Power consumption of a terminal unit power (w)
N_{RN}	Number of customers sharing the remote node
N_{TU}	Number of terminal units
X_{RN}	Additional power consumption for environmental control in remote nodes (w)
X_{TU}	Additional power consumption for environmental control in terminal units (w)
t_i	Active time of a CPE or a nano server in interval i (sec)
T_i	Idle time of a CPE or a nano server in interval i (sec)
T_{tot}	Total time for a device located in an end-user premise (sec)
t_{idle}	Idle time for a device located in an end-user premise (sec)
t_{act}	Active time for a device located in an end-user premise (sec)
$t_{\text{act,k}}$	Active time of service k (sec)
E_{cpe}	Total energy consumption of customer premises equipment (CPE) (j)
$E_{\text{inc-cpe}}$	Incremental energy consumption of customer premises equipment (CPE) (j)
$E_{\text{inc,k}}$	Incremental energy consumption allocates to service k (j)
$E_{\text{idle,k}}$	Idle power consumption allocates to service k (w)
$E_{\text{k-time}}$	Total energy consumption of service k over $T_{\text{tot}}(j)$
$C_{\text{act,k}}$	Data rate of the service during active time (bps)
α	Ratio of the idle time of an unshared network equipment to its active time

Table 3.3: Notations in energy/power consumption model for lightly shared/unshared network equipment

3.4 Conclusions

In this chapter energy and power consumption models for various shared/unshared network equipment (as shown in Figure 3.1) and Internet services were examined. First, the total and incremental energy (and power) consumption were determined and distinguished. Then, we categorized network equipment as shared and unshared network equipment and proposed flow-based energy consumption model for shared network equipment and time-based energy consumption model for unshared network equipment. The shared network equipment refers to network equipment such as routers and switches shared among many users and deal with large amounts of traffic. Unshared network equipment refers to network equipment such as home modems and home servers located in end-user premised that are shared with a few users and are based on time rather than traffic.

The measurement-based energy (and power) consumption models were described and shown to apply to equipment, networks and services. These models are used to construct power and energy consumption estimates for a diverse range of network scenarios including customer premises equipment and access, edge and core networks and services provided over a network.

The power/energy models in this chapter are used in Chapter 4, 5 and 6 for modeling energy/power consumption of Cloud computing and Fog Computing applications.

Chapter 4

Energy Consumption of Photo Sharing in Online Social Networks (OSNs)

Executive Summary

Online social networks (OSNs), with their huge number of active users, consume large amounts of energy both in data centers and in the transport network. Existing studies on energy consumption of Cloud-based applications (i.e. OSNs) focus mainly on the energy consumption in the data centers and do not take into account the energy consumption during the transport of data between end-users and data centers. To estimate the amount of the neglected energy, this work provides a framework for network topology and energy consumption model in order to understand the energy consumption of Cloud applications such as photo sharing in social networks. A combination of energy consumption modeling (explained in Chapter3) and measurements are used with the energy models described in Chapter 3 to estimate the energy consumption of sharing photos on Facebook, as an example of a Cloud application.

Our results indicate that the energy consumption involved in the transport network and end-user devices for Facebook photo sharing is approximately 60% of the energy consumption of all Facebook data centers. Therefore, achieving an energy-efficient Cloud service requires energy efficiency improvement in the transport network and end-user devices along with the related data centers.

4.1 Introduction

Cloud computing moves data processing and storage away from end-user devices into data centers [78, 2], and underpins many online social networks (OSNs) such as Facebook, Twitter and LinkedIn. The ubiquity of broadband and wireless networking allows users to instantly connect socially via their PCs or handheld devices.

These Cloud services generate considerable amount of traffic and could readily change the Internet traffic landscape [79]. Associated with this increasing traffic is an increase in energy consumption for transporting, processing, and storing data. Since the data in Cloud services is processed and stored in data centers, an obvious focus for studying energy consumption of Cloud services is the data centers. Cloud provider companies frequently report on their activities to keep their data centers energy-efficient [80]-[82]. However, the energy consumption of a Cloud service includes three components: energy consumption of the data centers, energy consumption of the transport network that connects the users to the Cloud, and the additional energy incurred by end-user devices when accessing the Cloud [2, 30]. Energy consumption of the transport network and end-user devices have been ignored in most studies of energy consumption of Cloud computing [13, 11].

Among cloud based services social networking, and in particular photo sharing services, have become extremely popular and are generating significant network traffic volume. In this work, the energy consumption of a photo sharing service in an OSN is studied. As an example, we chose Facebook which currently hosts more than 240 billion photos, and users upload more than 350 million photos every day [7]. Facebook is becoming the biggest photo sharing platform in the world [83]. We analyze the energy consumption of end-user devices and the transport network when uploading and downloading[1] photos to and from Facebook.

This work is built upon the earlier work in [2], but differs in two significant ways. We improve the previous energy consumption models for Cloud applications, and apply the energy model to a Cloud application based on software as a service (SaaS). In this context, the contributions of this work are:

a) A energy model for shared network elements in the transport network is proposed

[1] We use *download photos* and *view photos* interchangeably in this work.

(explained in Chapter 3);

b) Power consumption and packet-level traffic of end-user terminals are measured to obtain a realistic energy consumption estimate of photo sharing on an OSN.

c) A comprehensive network structure and behavior of photo sharing in OSNs are studied.

The results of this work show that total energy consumption for uploading and downloading photos on Facbook in one year to be about 304 Gigawatt hour(GWh). By comparison, according to Facebook [84], it consumed about 500 GWh of energy in 2012 for the IT facilities in its data centers [84]. Therefore, the energy consumption of the transport network and end-user devices for photo sharing is approximately 60% of the total energy consumption of the Facebook data centers including all services such as photo and video sharing, game, chat and many more.

It is revealed that the energy consumption of Cloud services in the transport network and end-users devices is considerable and should not be ignored when studying the energy consumption of Cloud computing services.

The rest of this chapter is organized as follows. Photo sharing in a social network is described in §4.2. The energy consumption model for photo sharing in OSNs is presented in §4.3. In §4.4 , the relevant traffic measurements are reported. The energy consumption of end-user devices, access network, and edge (and core) network is studied in §4.5, §4.6 and §4.7 , respectively. The energy consumption of Facebook photo sharing over one year is presented in §4.8. Finally, the work is concluded in §4.9.

4.2 Photo Sharing in an Online Social Network

In social networks, new uploaded photos are often more popular than older photos. The term *Hot* is used by Facebook to describe the status of these popular photos. The popularity of the photos typically decreases after a while (the status of the photos changes to *Warm*). After a few days or weeks, there are generally few downloads (the status of the photos changes to *Cold*) [7]. Figure 4.1 shows the percentage of user requests for Facebook photos and the volume of photos stored over time [7]. It can be observed that

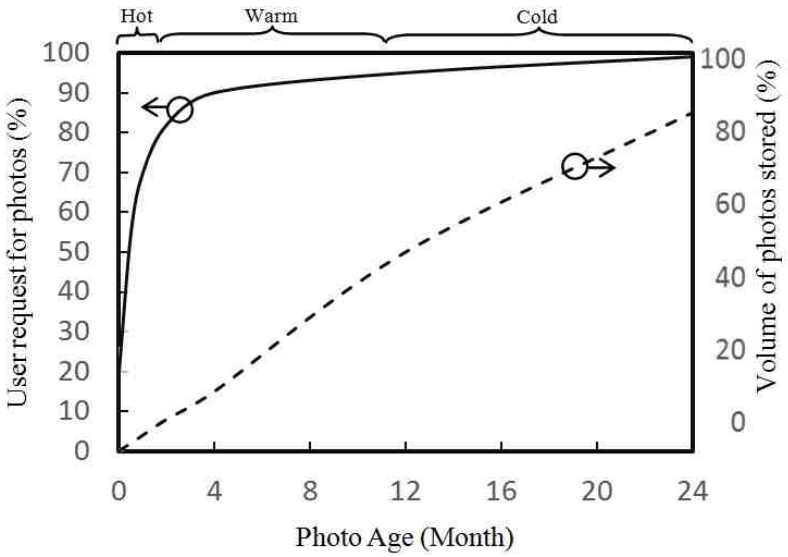


Figure 4.1: Access patterns to photos on Facebook, source: [7]

the majority of user requests are for the *Hot* photos. For example, approximately 82% of requests are for *Hot* photos (photos that are new to the system) which are 8% of the total photos. 13% of photo requests are for *Warm* photos and 5% of requests are for *Cold* photos [7].

Facebook mostly relies on a content delivery network (CDN) for sharing and distributing *Hot* and *Warm* photos (i.e. Akamai) [85, 86]. *Cold* photos are directly served from the Haystack cache (a CDN within Facebook's data center [83]) and are not distributed by the external CDN.

In the following sub-sections, a network model for uploading and downloading photos to and from Facebook is described.

4.2.1 Uploading photos

The uploaded photos are transmitted to the data center closest to the user. Figure 4.2 shows a high-level view of the Facebook network and its connectivity to users. There are a few Facebook data centers which are connected to the core of the Internet.

When a user uploads a photo, the data traverses an access network which might be an ADSL, Ethernet, WiFi, 3G or 4G connection, or a combination of these. Then, the data passes through an edge (metro) network which generally consists of a metro Ethernet

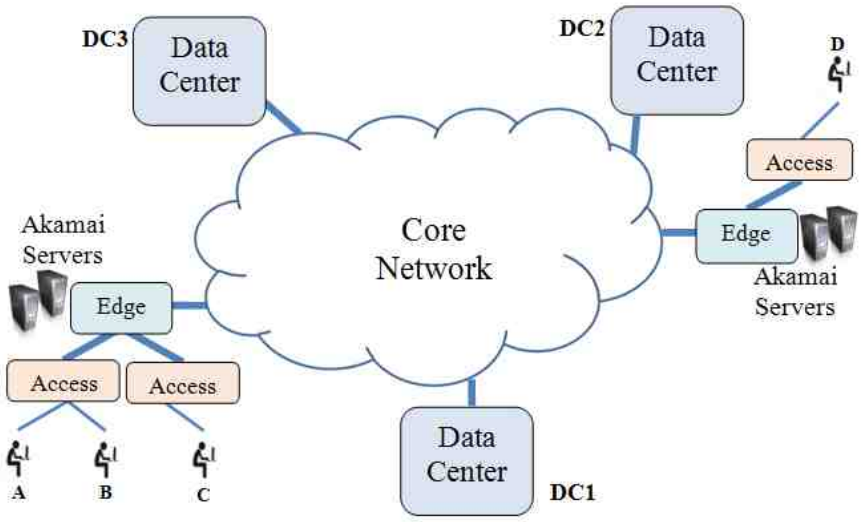


Figure 4.2: Network model of an online social network

switch, broadband network gateways (BNGs) and edge routers [2, 30]. Subsequently, the data traverses the core network comprising large core routers and optical links. The final destination for storing photos is a physical disk drive within a data center. The data center network includes one or a few edge routers, aggregation switches and application servers and storage servers.

4.2.2 Downloading photos

When a user views a photo, the user's browser first sends a request to a web server to find where to download the photo from: a CDN (Akamai) server or a server within the Facebook data center [83]. For *Hot* and *Warm* photos, the browser is directed to Akamai servers. Access to *Cold* photos is directly from the Facebook data center without passing through the Akamai network [83]. Figure 4.2 indicates Akamai servers in the edge of the network collocated with other ISP equipment. Distribution of photos by Facebook is based on the location of friends who are interested in the photos.

When user A in Figure 4.2 wants to share a photo on Facebook, the photo is sent to a Facebook data center (DC1). Then, all friends (user B, C and D) can see the shared photo. When friends request the photo, DC1 sends the photo to Akamai intermediate nodes [34] and then after a few hops it goes to an Akamai server at the edge of the network which is very close to the users. Local friends such as users B and C who are connected to the

same edge network can see the photo from the edge of the network. In contrast, when User D requests the photo, another route is used from Akamai servers in the core of the network to a server at the edge of the network near user D to respond to the request.

4.3 Application of Energy Consumption Model

In order to obtain the energy consumption of a SaaS (Software as a Service) application such as Facebook, a combination of power measurement and energy consumption modeling are required. In this work, incremental energy consumption of network equipment is studied. Incremental energy consumption includes the additional energy consumed by end-user terminal (incurred when accessing the cloud) and the various network elements (incurred when forwarding the application data) along the path between the user and the cloud. All the energy consumption models of network equipment are described in Section 3.

The incremental energy consumption of Software as a Service (SaaS) application ($E_{\text{inc-cloud}}$) in the end-user devices and transport network can be determined as follow:

$$E'_{\text{inc-cloud}} = E'_{\text{terminal}} + N_{\text{bit}}(E'_{\text{b-access}} + E'_{\text{b-edge}}h_{\text{e}} + E'_{\text{b-core}}h_{\text{c}}) \tag{4.1}$$

where,

E'_{terminal} is the incremental energy consumed by the end-user device when interacting with the Cloud service;

$E'_{\text{b-access}}$ is the incremental energy per bit of the equipment in the access network;

$E'_{\text{b-edge}}$ is the incremental energy per bit of the equipment in the edge network;

$E'_{\text{b-core}}$ is the incremental energy per bit of the equipment in the core network;

h_{e} is the number of edge routers traversed;

h_{c} is the number of core routers traversed;

N_{bit} is the number of transmitted and received bits when interacting with the Cloud service.

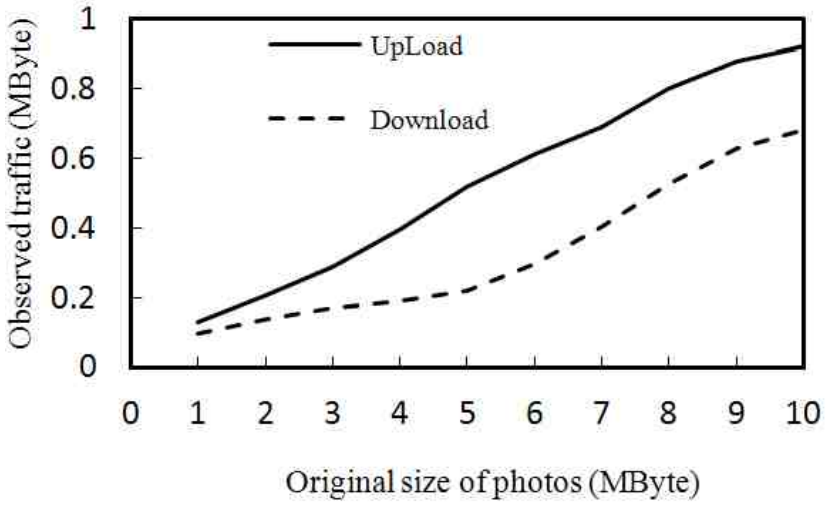


Figure 4.3: Observed traffic during uploading and downloading various sized photos to and from Facebook versus the original sizes of photos

4.4 Traffic Measurement

In order to examine the number of transmitted and received bits (N_{bit}) when sharing a photo on an OSN, we measured the volume of traffic generated for uploading a photo to Facebook and then downloading the same photo from Facebook. To do this, we used a packet analyzer software utility (Wireshark [87], running on the end-user device) to capture all packets exchanged with Facebook.

Photos of different sizes ranging from 1 MB to 10 MB were uploaded to Facebook with normal resolution. Figure 4.3 shows the number of bytes exchanged during uploading and downloading photos versus the size of original photos. The upload curve indicates the traffic volume exchanged during uploading is very much smaller than the original size of the photos. Based on our measurements, we deduced that Facebook compresses photos heavily in the users' browser before sending them to Facebook servers. Photos are compressed to 960×640 pixels for normal quality and 2048×1536 pixels for high quality. However, Facebook does not compress small photos with fewer pixels than the above-mentioned thresholds. In addition, Figure 4.3 shows the upload traffic to this cloud service is greater than the download traffic.

We noted from the Wireshark logs that uploading (or downloading) a photo is sent (or received) as 1314 Byte TCP packets to (or from) the servers followed by ACK packets

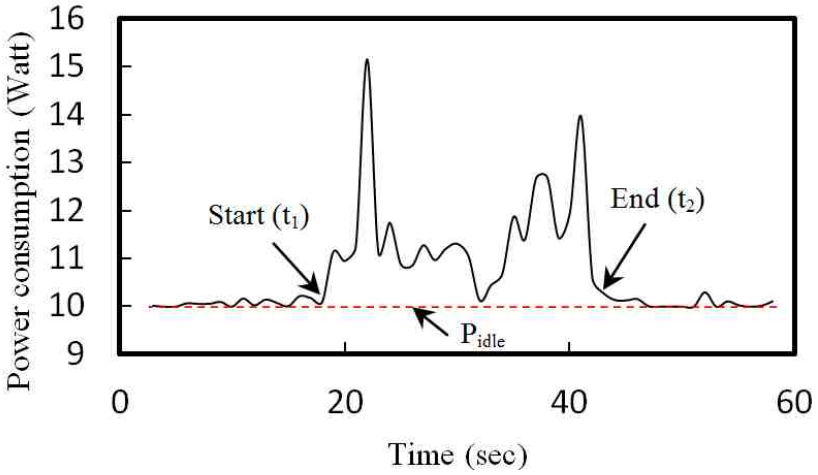


Figure 4.4: Power consumption of a laptop while uploading a photo to Facebook

from the servers (or end-user devices). Both data and ACK packets are included in the traffic count.

The observed traffic for uploading a 5-MB photo in normal quality using a laptop with home WiFi and Ethernet technology is about 500 KB. We also uploaded the same photo using a smart-phone with home WiFi and 4G technologies. The observed traffic was about 1.1 MB.

The download curve also shows that the uploaded photos on Facebook are compressed since the observed traffic during downloading photos is smaller than the original size of photos. The observed traffic for downloading the uploaded photo (5-MB photo) using a laptop with home WiFi and Ethernet technology is 200 KB. The observed traffic when using the Facebook mobile application on a smart-phone (WiFi and 4G) was 120 KB.

Considering the fact that Facebook is not a Storage-as-a-Service [2] service, photo compression is a very effective solution for saving bandwidth, increasing the upload speed and avoiding high traffic in the network.

4.5 Energy Usage of End-user Devices

In a global context, we need to consider all contributions to energy consumption. Therefore, we include the energy consumption of the end-user devices. In order to estimate the incremental energy consumption of end-user devices when interacting with an OSN, we

	Laptop		Mobile phone	
	Ethernet	WiFi (home)	4G	WiFi (home)
Upload	106 J	114 J	40 J	23 J
Download	23 J	33 J	18 J	8 J

Table 4.1: Energy consumption of end-user devices for sharing a photo (with original size of 5MB) in a social network

consider the energy consumption of a low power laptop and a smart-phone.

The laptop used in these experiments is a Sony VAIO Duo 11 running Windows 8 [88], chosen as representative of a modern low energy laptop computer. We used a PowerMate power meter (resolution of 10 mW) [89] and measured the power consumption of the laptop when interacting with the cloud by Ethernet and WiFi connections. Figure 4.4 shows the power consumption of the laptop versus time during uploading a 5-MB photo in normal quality. The power consumption of the laptop when connected to Facebook via wired Ethernet in an idle state is 10 Watt (W).

The incremental energy consumption of this device associated with the upload (or download) is given based on (3.15) in Chapter 3.The incremental energy consumption for uploading a 5 MB photo is 106 J and the energy consumption for a home WiFi connection is 114 J.

The same measurement and calculation methods are used to calculate the energy for downloading photos by the laptop. The results are listed in Table 4.1.

Increasingly, end-users are turning to mobile devices and wireless access networks, rather than PCs/laptop computers and wired connections. Currently, more than half of the users access Facebook via mobile devices [90], the incremental energy for uploading a photo using a smart-phone is obtained by a mobile phone application named *PowerTutor* [91, 92]. The energy consumed by a smartphone with home WiFi and 4G technologies for uploading 1.1 MB are measured to be 23 J and 40 J, respectively.

For viewing the uploaded photo by the smart-phone, the incremental energy by home WiFi and 4G for downloading the photo (file size 120KB) are 8J and 18J, respectively. All results are summarized in Table 4.1.

	Power (Watt)		Capacity (Mbps)		Energy (nJ/bit)	
	Idle	**Max**	**Downlink**	**Uplink**	**Downlink**	**Uplink**
Ethernet Gateway (CPE)	2.8	4.6	100	100	18	18
ADSL2+ Gateway (CPE)	4.1	6.7	24	3.5	108	866
Ethernet Switch (Network edge)	1,589	1,766	256,000	256,000	31.7	31.7
LTE Base Station (Network edge)	333	528	72	12	76,200	19,000

Table 4.2: Energy per bit of equipment in access network

4.6 Energy Consumption of Access Network Equipment

Access network equipment includes customer premises equipment (CPE) , and shared equipment at the network edge. CPE would include an Ethernet gateway, DSL modem, optical fiber network unit, etc while the network edge might include a large Ethernet Switch, an LTE base station, an optical line terminal (OLT), etc.

Table 4.2 lists the energy per bit for access network equipment when receiving data from the users (uplink) and transmitting data to the users (downlink). The data for gateways is from [93] and the energy per bit is calculated based on (3.3) in Chapter 3. The idle power, maximum power and maximum capacity of a typical Ethernet switch is from [30] and the energy-per-bit is obtained according to (3.7) in Chapter 3 assuming a typical utilization of 20% (because they are shared). Finally, to determine the energy-per-bit for LTE base stations, we observe from [76] that the idle and maximum power consumption of a 3-sector 2x2 MIMO 4G/LTE base station deploy in an urban area are 528W and 333W, respectively. In addition, 4G/LTE base stations consume more energy in the downlink direction which is 87% of the total energy consumption according to [76]. The aggregate throughput of this base station is 72 Mbps with 20 MHz spectrum [94]. The average energy-per-bit of this base station is 76.2 μJ/bit in the downlink and 19 μJ/bit in the uplink assuming a typical utilization of 5% over a 24-hour cycle. Should be noted that overall, 4G/LTE as an access technology is much less efficient than the others considered.

The uplink column (the last column in Table 4.2) is used for calculating incremental energy consumption while uploading a photo and the downlink column is used for downloading a photo.

Based on the values in Section 4.4, the traffic for uploading a 5-MB photo by a laptop

	Access via a laptop		Access via a phone	
	Ethernet	**WiFi (home)**	**4G**	**WiFi (home)**
Upload	0.2 J	0.5 J	670 J	1.2 J
Download	0.08 J	1.4 J	18.2 J	0.8 J

Table 4.3: Energy consumption of equipment in access network for sharing a photo in a social network

via Ethernet and home WiFi is 500 KB. Hence, the incremental energy consumption of Ethernet and home WiFi equipment for uploading this photo is 0.2 J and 0.5 J, respectively. For uploading the same photo by a smart-phone via home WiFi and 4G, for which the observed traffic is 1.1 MB, the incremental energy is 1.2 J and 670 J, respectively. Similar calculations have been done for downloading the photo. These results are outlined in Table 4.3.

4.7 Energy Consumption of Edge and Core Network Equipment

The maximum energy consumption, maximum capacity and the energy-per-bit (E_b') of the network equipment in the edge (metro) and core networks are listed in Table 4.4. Although we do not know what equipment is used in ISP networks, those listed in the table are representative of network equipment. The maximum energy consumption and maximum capacity are gathered from Cisco's power consumption calculator [95]. The energy per bit for shared network equipment (E_b') in the edge and core network is obtained based on (3.7) in Chapter 3. We used the value of 60% for U.

By using *traceroute* from end-user device to the Facebook servers, we estimate that on average five core routers and three edge routers are along the path between the users and the servers.

Bringing together the results above for the incremental energy per bit (E_b') and the traffic measurements for uploading the photo, the energy of edge and core equipment for uploading the photo when using a laptop (with home WiFi and Ethernet) is determined to

Type	**Max power** (Watt)	**Max capacity** (Gbps)	$E_{\mathbf{b}}^{'}$ (nJ/bit)
BNG	1890	320	27
Edge router	4550	560	37
Core router	12300	4480	12.6
Server	0.8	225	1037

Table 4.4: Energy per bit of equipment in edge and core networks

be 0.8 J. The incremental energy when using a mobile phone (with home WiFi and 4G) is about 1.8 J. These results are summarized in Table 4.5.

For downloading the photo from a server within a data center, the traffic comes from the data center to core routers, edge router, BNGs, Ethernet switch and access network, in turn. Therefore, the energy consumption of all of this equipment should be considered. The energy for edge and core network is obtained from the numbers in Table 4.4 and the measured traffic from Section 4.4. The energy of equipment in the edge and core networks during downloading the photo (the observed traffic is 200KB) is estimated to be 0.3 J. When the observed traffic is 120KB, the energy is estimated to be 0.2 J. These results are shown in the second row of Table 4.5.

According to [86], the majority of friends using an OSN are relatively closely located geographically so we can assume that half of the friends of a Facebook user are in a local area. For local users in the same geographic region, the photo can be cached to an Akamai server once and then other friends download it from the edge network. Hence, there will be only a few core and edge router hops. The energy consumption in the core and edge networks for downloading one photo for a local friend is summarized in the third row of Table 4.5.

By using Akamai servers in the edge network, the number of hops in core routers and edge routers decreases and energy can be saved. However, the energy consumption of a server in the edge network is added. The maximum power consumption and maximum capacity of a typical content server are gathered from [96] and reported to be 225 W and 800 Mbps, respectively. The idle power consumption of this server is typically 80% of the

	Core & edge via a laptop		Core & edge via a phone	
	Ethernet	**WiFi(home)**	**4G**	**WiFi(home)**
Upload	0.8 J	0.8 J	1.8 J	1.8 J
Download from data center	0.3 J	0.3 J	0.2 J	0.2 J
Download from edge network	0.2 J	0.2 J	0.1 J	0.1 J
Server in edge network	1.7 J	1.7 J	1 J	1 J

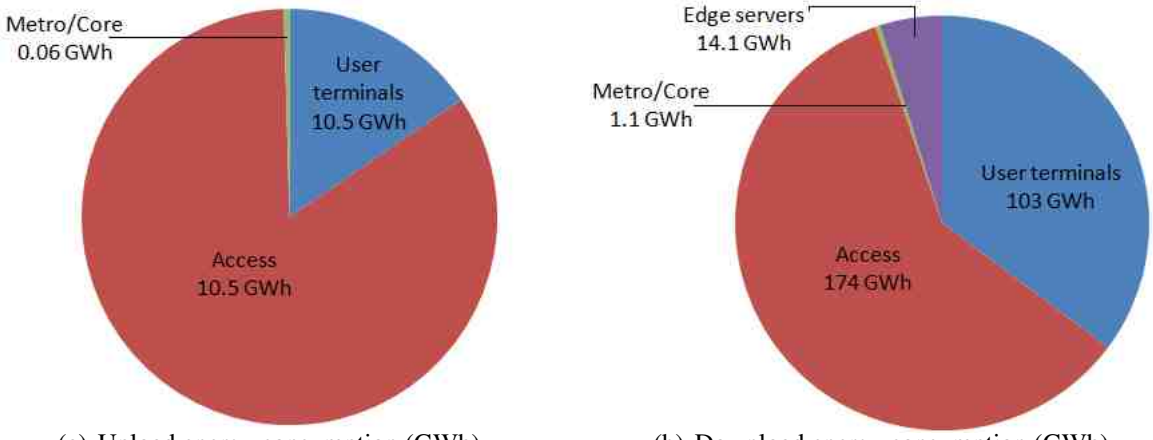
Table 4.5: Energy consumption of equipment in core and edge networks for sharing a photo in a social network

maximum power consumption, therefore the energy per bit, based on (3.7) in Chapter 3, is 1.0 μJ/bit. Then, the power consumption of a server when traffic is 200 KB (the traffic comes from a Laptop) is 1.7 J and when traffic is 120 KB (the traffic comes from a mobile phone) is 1 J. The results are presented in the last row of Table 4.5.

4.8 Photo Sharing Energy Consumption over One Year

We have estimated the total energy consumption for uploading and downloading one average sized photo to and from Facebook including the end-user devices and transport network. The energy consumed for uploading and downloading the photo is 355 J (0.1 Wh) and 100 J (0.03 Wh), respectively. We now use these results to estimate the energy consumption of photo sharing in one year and compare this value to the total energy consumed for IT facilities in entire Facebook data centers in one year which is 500 GWh [84].

Users upload more than 350 million photos to Facebook every day and all the uploaded photos can be downloaded by the users' friends. Each Facebook user has 140 friends on average [97] and we have assumed that 90% of the friends (126 people) view the new uploaded photos. In addition, about 68% of Facebook users are mobile users (751 million of the 1.1 billion) [90]. Since 35% of mobile traffic is WiFi traffic and 65% is cellular traffic [98], we infer 24% (0.68×0.35) of the users are connected to Facebook by home WiFi and 44% (0.68×0.65) of users are connected by 4G. Additionally, we



(a) Upload energy consumption (GWh) (b) Download energy consumption (GWh)

Figure 4.5: Annual energy consumption of photo sharing on Facebook

set the number of users connect to Facebook by Ethernet with a low power device such as laptops/ultrabooks is the same as the number of users by home WiFi with laptops/ ultrabooks [98, 99]. Therefore, 16% is assumed for laptop users connected by Ethernet and 16% for laptop users connected by home WiFi.

Using these data, we estimate the total incremental energy consumption for uploading photos to Facebook in one year to be 12.5 GWh. The energy consumed in end-users devices, access network and edge (and core) network is estimated to be 2 GWh, 10.5 GWh and 0.06 GWh, respectively (as shown in Figure 4.5(a)).

Based on the data presented above, we estimate the total incremental energy consumption for downloading recently uploaded photos (*Hot* photos) from Facebook in one year to be 868 TJ. The request for *Hot* photos is 82% of all requests, 13% of all requests are for *Warm* photos (*Hot* and *Warm* photos are downloaded from the edge network) and 5% of requests are for *Cold* photos (*Cold* photos are downloaded from the data center)[7], the total energy consumption for downloading photos from Facebook is approximately 292 GWh per year. The consumed energy in end-users devices, access network, edge (and core) network and servers in the edge network is estimated to be 103 GWh, 174 GWh, 1 GWh ,and 14 GWh, respectively (as shown in Figure 4.5(b)).

4.9 Conclusion

In this chapter we evaluated the incremental energy consumption of the photo sharing service on Facebook. We studied the incremental energy consumption of user devices, the Akamai servers in the edge network and also the incremental energy consumption of transport network including access, edge and core networks. This additional energy consumption is ignored by most of the works that have evaluated the energy consumption of Cloud computing.

Given the current profile of access technologies used by Facebook users, the estimated annual energy consumption in the transport network and end-user devices for uploading and downloading Facebook photos are about 12.5 GWh and 292 GWh, respectively. Facebook does not explicitly report the energy consumption of their data centers for specific services such as photo sharing. Instead, what they report is the gross data center energy consumption, which is 500 GWh. Comparing our estimate of 304 GWh with 500 GWh, we note that the energy consumption incurred in the transport network and end-user devices is about 60% of the energy consumption of all Facebook data centers. This figure would be higher if we could compare our estimate with just the fraction of data center energy consumption attributed to the photo sharing service alone.

The results in this work show that achieving an energy-efficient cloud service, requires improving the energy efficiency of the transport network and the end-user devices along with that of the data centers. The goal of this study is to gain insights that can inform network designers for future energy-efficient deployment of cloud services and applications. The greatest energy consumption gain would come from improving the energy-efficiency of the access network, especially for wireless 3G/4G/LTE. For example, initiatives for networks to serve wireless users through WiFi hotspots or small cells, in preference to Macro base stations.

The proposed energy model and measurement techniques are not specific to social networks and can be used to estimate the energy consumption of other Cloud applications as well.

Chapter 5

Energy Consumption of Interactive Cloud-Based Applications

Executive Summary

Interactive Cloud computing and Cloud-based applications are a rapidly growing sector of the expanding digital economy because they provide access to advanced computing and storage services via simple, compact personal devices. Recent studies have suggested that processing a task in the Cloud is more energy-efficient than processing the same task locally. However, these studies have generally ignored the power consumption of the network and end-user devices when accessing the Cloud. In this work, we develop a power consumption model for interactive Cloud applications that includes the power consumption of end-user devices and the influence of the applications on the power consumption of the various network elements along the path between the user and the Cloud data center. As examples, we apply our model to Google Drive and Microsoft OneDrive's (previously known as SkyDrive) Word processing, Presentation and Spreadsheet interactive applications. We demonstrate via extensive packet-level traffic measurements that the volume of traffic generated by a session of the application vastly exceeds the amount of data keyed in by the user. This has important implications on the overall power consumption of the service. We show that using the Cloud to perform certain tasks consumes more power (by a Watt to 10 Watts depending on the scenario) than performing the same tasks locally on a low-power consuming computer or tablet.

5.1 Introduction

Cloud computing and web-based Cloud offerings are hailed as the new wave transforming the IT industry. Enterprise customers and home users are increasingly being offered the opportunity to move from running applications on stand-alone computers to using Cloud-based services. As a result, the use of these applications is expected to grow dramatically in the future as more businesses and consumers choose to access applications, documents and content remotely over the Internet [100, 101, 30].

There are three broad flavors to Cloud computing – Infrastructure as a Service (Iaas), Platform as a Service (PaaS) and Software as a Service (SaaS) [102]. This work focuses on SaaS because a large number of Cloud service providers, such as Google, Microsoft and Amazon, promote SaaS products which have the same look-and-feel as desktop applications, to encourage users to make a transition to the Cloud.

Cloud services offer numerous benefits in terms of cost, scalability, performance and maintenance. Several recent studies [11, 12, 13] have suggested that Cloud offerings are "green" in the sense that they save energy relative to traditional desktop computing. The rationale for this is that data centers are generally optimized for energy efficiency, and migration of applications to the Cloud permits replacing high-power desktop computers with low-power consuming computers such as netbooks and tablets. Further, the compute and storage resources in data centers are often shared by many users, in contrast to a single user running a dedicated desktop computer.

While intuitively reasonable, the above argument ignores two key factors: (1) energy required to transport data between the user and the Cloud, and (2) power consumed by the end-user device when accessing the Cloud. Although prior work advocates computation offloading [103, 104, 105], namely techniques to reduce the power consumption of end-user devices (e.g. tablets) when accessing the Cloud, it largely ignores the energy consumed for *transporting* data from the end-user device to the Cloud and back. Using a network-based model it is shown that as the data rate between the user and the Cloud data center increases, the transport energy becomes a dominant fraction of the total energy consumption of Cloud computing, thus reducing the latter's energy efficiency [2].

Numerous interactive Cloud-based applications have become available in recent years.

Moreover, with the widespread deployment of high-bandwidth 3G/4G wireless networks, the number of mobile Cloud users is expected to grow significantly [100, 101]. The large-scale migration to Cloud computing makes it important to quantify the traffic and power consumption implications of using interactive Cloud-based applications.

This work is based on the earlier work in [2] by constructing a measurement based power consumption model for interactive Cloud-based applications. This model includes all components of the interactive Cloud service and the measurements expose the fact that the volume of traffic generated during an online session of the application can be as much as a 1000-times larger than the amount of data keyed in by the user. The model is then used to compare the power consumption of three scenarios:

(i) Creating, editing and saving documents, presentations and spreadsheets in the Cloud;

(ii) Creating and editing the applications locally, and then saving the files in the Cloud;

(iii) Performing the tasks locally (i.e. the Cloud is absent). All the tasks are performed on the same low-power consuming end-user devices.

An important finding of this work is that although migration to the Cloud offers significant benefits, performing tasks in the Cloud may not always be the most energy efficient way to undertake those tasks. The relative merits of using a Cloud service, from the perspective of power consumption, depends on factors such as the power consumption of the end-user device, the access network technology used, the computational complexity of the task to be performed, the volume of traffic exchanged between the user and the Cloud, and factors such as the number of users sharing a compute resource in the Cloud.

The rest of this work is organized as follows. In Section 5.2, we develop a model for quantifying the power consumption per user incurred when using interactive Cloud-based applications. In Section 5.3, we report measurements of traffic, in particular the overhead multiplier. We present estimates of power consumption for various network elements in Section 5.4, and use this to estimate the power consumption per user in Section 5.5. We conclude the work in Section 5.6.

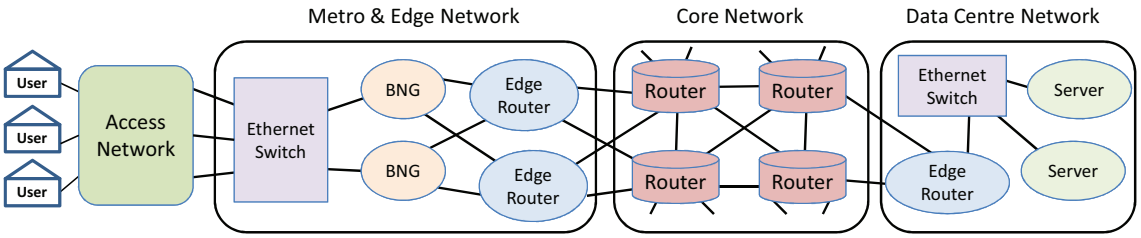


Figure 5.1: Topology of the network between a end-user and the Cloud data center.

5.2 Application of Power Consumption Model

We consider a user accessing the Cloud via the network topology shown in Figure 5.1. The access network includes ADSL Ethernet, WiFi, or in the case of wireless, a 3G/4G (LTE) connection. The metro Ethernet switch aggregates traffic from several users, broadband network gateways (BNGs) regulate access and usage, and edge routers represent the gateway to the global Internet, which consists of many large core routers. Similar architectures have been used in previous studies (e.g. [2, 106]). The data center network comprises an edge router connecting the data center to the Internet, aggregation switches and application servers.

The power consumption per user, P_I, of using an interactive Cloud-based application is a function of the bit-rate of the application and the energy per bit incurred by the various network elements, shown in Figure 5.1, required to deliver the service to the user. This power can be expressed as follows:

$$P_I = P_u + E_a B + (N_c E_c + N_e E_e + E_{bng} + E_{sw})B + E_d B + P_d \tag{5.1}$$

where,

P_u is the power consumed by the end-user device to access the interactive Cloud application;

B is the bit-rate of the application;

N_c (N_e) are the number of core (edge) routers along the path between the user and the application server in the data center;

E_c, E_e, E_{bng}, E_{sw} and E_d denote respectively the energy per bit of the core router, edge

router, BNG, Ethernet and data center switches;

E_a is the energy per bit of the access network;

P_d is the power consumption per user of the server in the data center.

The power consumption of a server is a function of its CPU utilization, which is related to the number of processes running on it. This in turn relates to the number of users assigned to that server. We have thus used power per user to model the server power consumption. For network equipment, power consumption is a function of the load [75], i.e. bits per second flowing through it, and is modeled using energy per bit, as described next.

5.2.1 Energy per bit modeling

The energy per bit of network elements is modeled based on (3.7) in Chapter 3. We assume a realistic utilization ($U = 30\%$) [107], and then apply (5.1) to estimate the power consumption due to the traffic generated when accessing the Cloud application.

5.2.2 Power consumption measurement

The power consumption of end-user devices when interacting with the Cloud (e.g. Google Drive and Microsoft OneDrive[1]) is measured directly using a power meter. In the measurements, we noted that the power consumption of a desktop PC or a high-end laptop was virtually unchanged when interacting with these Cloud applications. In order to accurately isolate the power consumption of a end-user device, we used a MSI Wind U100 netbook computer [108] running Windows XP on a 1.6 GHz Intel Atom processor with 2 GB memory. This netbook computer is representative of Cloud-ready low-power consuming user devices such as Google Chromebook, which consumes 11 W when awake [109] (similar to the netbook). We also performed measurements using a Samsung tablet [110]. A PowerMate power meter [89] (resolution of 10 mW) was used to record the power consumption of the netbook computer with the battery pack removed at intervals of 1 sec during each session. This enabled us to accurately determine the netbook computer's average power consumption. A custom-built power meter was used to record the power

[1]Previously known as SkyDrive

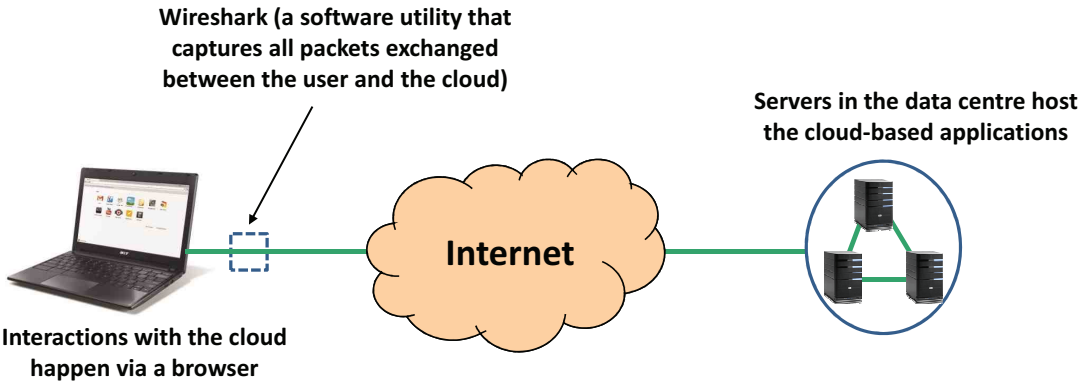


Figure 5.2: Measurement setup to capture the volume of traffic generated when accessing Cloud-based applications.

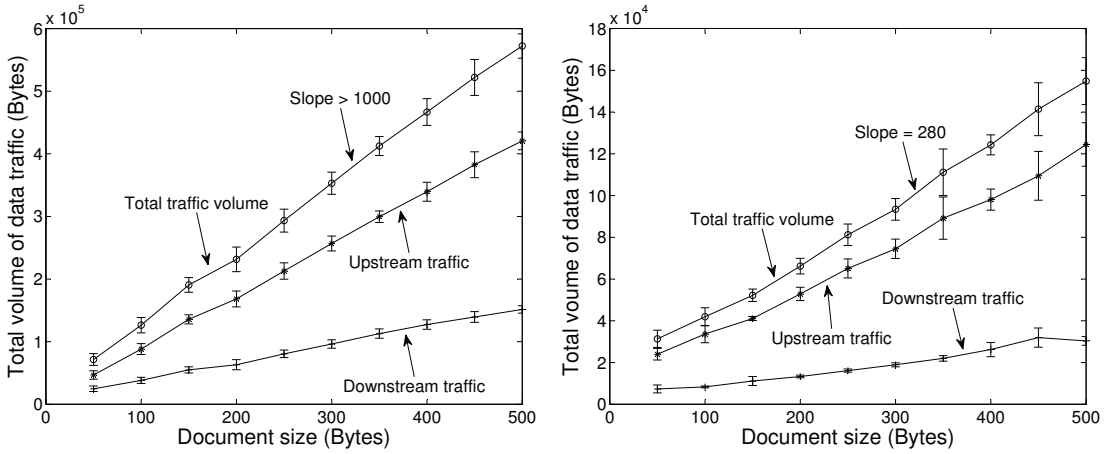
consumption of the tablet.

5.3 Measuring Cloud Application Traffic

We used the setup shown in Figure 5.2 to measure the volume of traffic generated by a session of a Cloud application. A packet sniffer software utility (Wireshark [87]), running on the netbook computer captures statistics of all packets exchanged with the Cloud server during each session. The file size and the number of key strokes when using the Cloud applications were also measured. The applications used for the measurements were office-based applications. The number of characters typed into each application varied from 50 to 500 in steps of 50 characters (equivalently the number of bytes entered varied from 50 to 500 in steps of 50 Bytes). Each session was repeated 10 times to obtain confidence in the results. We automated the typing process using Robosoft record-and-playback software [111]. This enabled us to repeat the experiments consistently across the different applications, ensuring that the typing speed was the same each time; $\approx$ 57 words per minute (speed of a professional typist).

Traffic measurements for two scenarios are considered, corresponding to how the Cloud is used.

(a) Composing and editing Word documents, Presentations and Spreadsheets *online* in Google Drive and Microsoft OneDrive using a web browser (Edit online, Save in the Cloud).



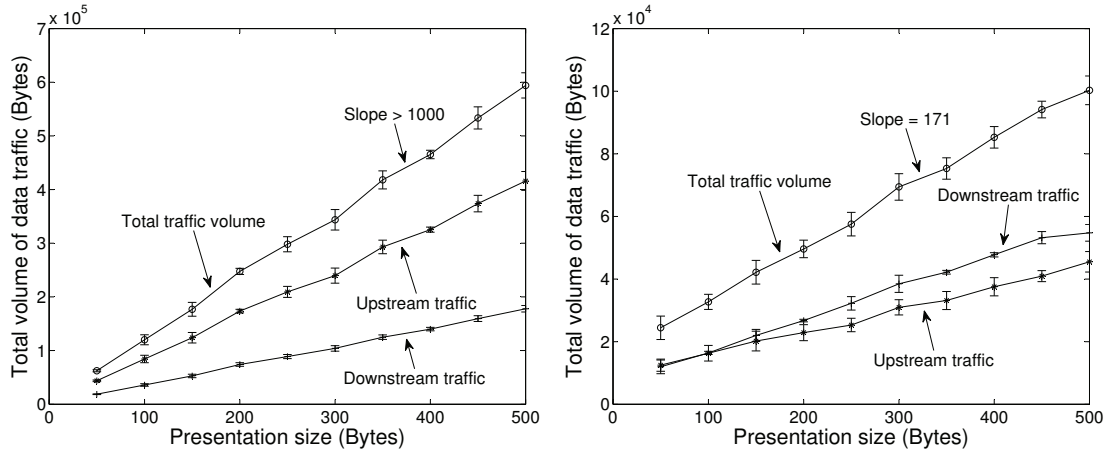
(a) Traffic volume from Google Docs. (b) Traffic volume from Microsoft Docs.

Figure 5.3: Volume of traffic generated vs the size of the document for (a) Google Drive and (b) Microsoft OneDrive word processing applications.

(b) Composing and editing Word documents, Presentations and Spreadsheets *offline* (i.e. locally on the netbook computer), then saving the files in the Google Drive folder on the netbook, and finally synchronizing the folder with the Cloud (Edit offline, Save in the Cloud).

5.3.1 Online interactive Word processing and Presentation applications (edit online, save in the Cloud)

Figures 5.3 and 5.4 show the total volume of data traffic (in Bytes) exchanged between the user and the Cloud for the online interactive Word processing and Presentation applications from Google and Microsoft. The figures also show the traffic volumes in both the upstream and downstream directions. This data was generated after post-processing the Wireshark logs. It can be observed from Figure 5.3 and Figure 5.4 that the total volume of data traffic is substantially larger than the amount of data typed into the application by the user. The overhead multiplier (in terms of the number of bytes) when using Google for both applications is more than a 1000-fold while the overhead multiplier when using Microsoft is 280-fold for Word processing, and 171-fold for Presentation.



(a) Traffic volume from Google Presentation.

(b) Traffic volume from Microsoft Presentation.

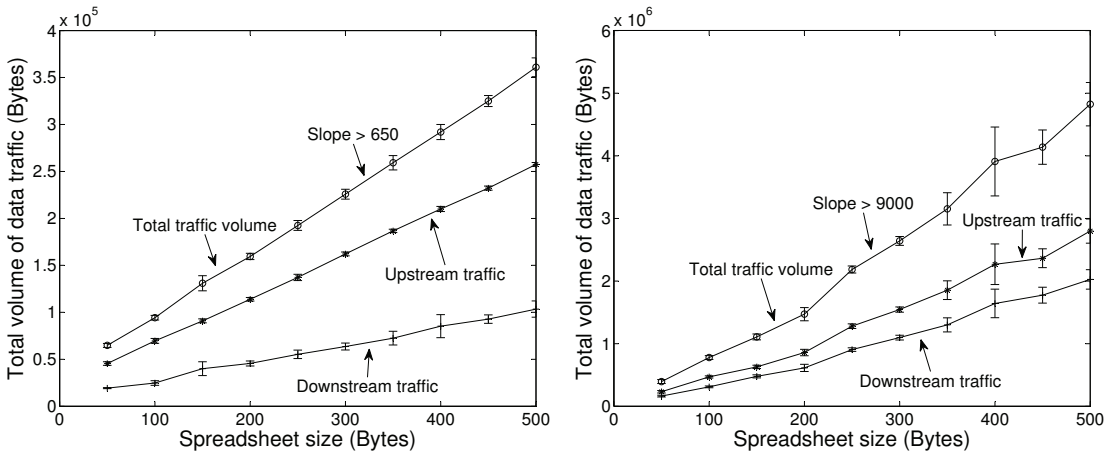
Figure 5.4: Volume of traffic generated vs the size of the presentation for (a) Google Drive and (b) Microsoft OneDrive presentation applications.

5.3.2 Online interactive Spreadsheet applications (edit online, save in the Cloud)

The volume of traffic generated by the Spreadsheet application from Google and Microsoft is shown in Figures 5.5(a) and 5.5(b). The former generates an overhead multiplier of 650, which is smaller than that of the other two applications, while the latter incurs a substantial overhead; in excess of 9000.

5.3.3 Insights into the traffic overhead for online interactive applications (edit online, save in the Cloud)

The Word processing, Presentation and Spreadsheet applications from Google and Microsoft are essentially client-server applications, the browser is the client and the server is accessed via the Cloud. Moreover, their look-and-feel, responsiveness and user experience are very similar to that of local stand-alone applications. To support these features, a considerable amount of communication occurs in the background between the browser and server (a brief overview from Google's applications appears in [112]). We noted from the Wireshark logs and while performing the measurements that changes made to the applications were *automatically* saved in the Cloud server, thereby ensuring no data loss.



(a) Traffic volume from Google Spreadsheet. (b) Traffic volume from Microsoft Spreadsheet

Figure 5.5: Volume of traffic generated vs the size of the spreadsheet for (a) Google Drive and (b) Microsoft OneDrive spreadsheet applications.

Although this provides high service reliability, it incurs a significant traffic overhead.

Word processing and Presentation applications

In the case of Google's Word processing and Presentation applications, logs of the traffic between the user and the data center show that every key stroke triggers an application synchronization event between the user and the server. Figure 5.6 shows a log excerpt from Wireshark for the Word processing application from Google. A single key pressed at the traffic log time 20.63384 sec is sent as a 1314 Byte TCP (Transmission Control Protocol) packet to the server. This is followed by three (relatively small) packets. The packets are transported using HTTPS making it difficult to decipher their content. The traffic logs indicate that the browser could communicate the key that was typed or deleted (for auto-saving), and the position of the cursor in the browser window to the server as part of every synchronization event. This occurs whether the event is an insert or delete operation. The synchronization process ends at time 22.8355 sec at which point the client and server "see" the same document. The next key press event starts at time 25.64 sec and the process repeats.

The behavior of Microsoft's Document and Presentation applications is similar to that of Google's. However, these applications generate less overhead because the latter typically synchronizes with the Cloud following every key stroke (as described above), while

Time	Source	Source port	Destination	Dest port	Protocol	Length	Info
20.633841	101.115.50.114	50359	gg.google.com	https	TCP	1314	[TCP segment of a reassembled PDU]
20.63389	101.115.50.114	50359	gg.google.com	https	TLSv1	448	Application Data
20.634792	101.115.50.114	50359	gg.google.com	https	TLSv1	312	Application Data
21.632119	101.115.50.114	50359	gg.google.com	https	TLSv1	91	Application Data
22.354545	gg.google.com	https	101.115.50.114	50359	TCP	54	https > 50359 [ACK] Seq=4530 Ack=9365 W
22.43447	gg.google.com	https	101.115.50.114	50359	TCP	54	https > 50359 [ACK] Seq=4530 Ack=9660 W
22.434809	gg.google.com	https	101.115.50.114	50359	TLSv1	91	Application Data
22.625126	101.115.50.114	50359	gg.google.com	https	TCP	54	50359 > https [ACK] Seq=9660 Ack=4567 W
22.784899	gg.google.com	https	101.115.50.114	50359	TLSv1	295	Application Data, Application Data
22.824843	gg.google.com	https	101.115.50.114	50359	TLSv1	112	Application Data
22.825003	101.115.50.114	50359	gg.google.com	https	TCP	54	50359 > https [ACK] Seq=9660 Ack=4866 W
22.825195	gg.google.com	https	101.115.50.114	50359	TLSv1	122	Application Data, Application Data
22.825412	gg.google.com	https	101.115.50.114	50359	TLSv1	144	Application Data, Application Data
22.825468	101.115.50.114	50359	gg.google.com	https	TCP	54	50359 > https [ACK] Seq=9660 Ack=5024 W
22.835063	gg.google.com	https	101.115.50.114	50359	TLSv1	229	Application Data, Application Data
22.83542	gg.google.com	https	101.115.50.114	50359	TLSv1	300	Application Data, Application Data
22.835515	101.115.50.114	50359	gg.google.com	https	TCP	54	50359 > https [ACK] Seq=9660 Ack=5445 W
25.647962	101.115.50.114	50359	gg.google.com	https	TCP	1314	[TCP segment of a reassembled PDU]
25.64801	101.115.50.114	50359	gg.google.com	https	TLSv1	448	Application Data
25.648544	101.115.50.114	50359	gg.google.com	https	TLSv1	312	Application Data

Figure 5.6: Wireshark trace following a single key being pressed in Google's interactive Cloud-based Word processing application.

the former synchronizes only when the user pauses or stops typing, as in between words. This results in a smaller volume of traffic exchanged between the user and the Cloud server, reducing the traffic overhead.

Spreadsheet applications

The Google Spreadsheet synchronizes with the Cloud only when the cursor (i.e. focus) shifts from one "cell" in the Spreadsheet to the next. This reduces the frequency of updates, and explains why the overhead (of 650) incurred by the Spreadsheet is smaller than that of the other two applications. In the case of Microsoft OneDrive's Spreadsheet application however, we note that the overhead is significantly larger, as shown in Figure 5.5(b). Post-processing the Wireshark logs revealed that this application generates a large number of TCP sessions and a vast majority of these TCP sessions lasts only a few sec. These sessions handle synchronization of content with the Cloud. For example, it took about 30 sec to enter 50 characters in the Spreadsheet. During this time, there were 20 TCP connections, each lasting on average 4.5 sec. The number of TCP sessions established grew rapidly with the size of the Spreadsheet. Entering 500 characters took 331 sec resulting in 174 TCP sessions, each lasting on average 6.3 sec. We were unable to elicit the content of the sessions because they were encrypted and transported using

HTTPS. The traffic logs indicate that the large traffic overhead is associated with establishing/tearing down TCP sessions very frequently and the volume of data transported to and from the user per session (tens to hundreds of Kilobytes). This behavior was not observed with Google Spreadsheet.

The qualitative explanations above are based on observed traffic measurements. A more precise explanation would require an accurate understanding of the way these applications are designed, which remains proprietary. It is evident that the underlying protocols used by the applications to provide a secure and rich user experience involve frequent and encrypted communication of data between the browser and Cloud server, giving rise to the large traffic overheads.

5.3.4 Word processing, Presentation and Spreadsheet applications (edit offline, save in the Cloud)

The total volume of data traffic exchanged (in Bytes) between the user and the Cloud for editing the Google and Microsoft Word, Presentation and Spreadsheet applications locally and then saving them to the Cloud is only marginally greater than the size of the file stored in the hard disk. The observed extra traffic is only due to the added bytes for secure transmission through the Internet, and the number of key strokes used to compose the file does not impact the traffic generated during the upload, i.e. the overhead multiplier, as described above, is absent in this scenario.

5.4 Power Consumption of Various Components

In this section we determine values of the various parameters in (5.1) needed to estimate the power consumption per user, P_1.

5.4.1 Bit-rate measurements for interactive Cloud-based Word processing applications

We used the setup shown in Figure 5.2 to compose a 2-page document on the Cloud. This experiment is representative of a typical instance where a user accesses the Cloud

to perform a word processing task. The experiment consisted of typing 649 words (4224 characters), inserting a picture, as well as a table comprising 4 rows and 3 columns. Each session on Google and Microsoft lasted on average 12 minutes (± 1 sec), and 11 minutes and 50 second (± 10 sec) respectively, providing us sufficient data to quantify the bit-rate of the applications. We ran a total of 30 sessions for each application.

As explained previously, we used Wireshark to capture all packets generated during each session. We noted from the logs that the bit-rate – i.e. B in (5.1) – for the online interactive Word processing application varied between 45 Kbps and 60 Kbps for Google, and between 10 Kbps and 12 Kbps for Microsoft. The Wireshark post-processing showed that more files with smaller sizes sent to Google data centers. The bit-rates are not a constant because the applications use TCP, and the performance of TCP varies depending on factors such as link congestion, delay and packet loss.

Identical measurements were conducted to determine the bit-rate of Word processing with Google Drive when the files are edited locally (offline) and then saved to the Google Cloud. The bit-rate varied between 1.1 Kbps and 1.5 Kbps. The bit-rate is calculated by dividing the observed exchanged traffic through the total time for editing the file offline plus transferring it to the Cloud.

5.4.2 Bit-rate measurements for interactive Cloud-based Presentation applications

Using the automated setup described above, we composed 5 slides each on the two Presentation applications. The experiment consisted of typing 127 words (735 characters), inserting a picture and a table comprising 4 rows and 4 columns. Each session on Google and Microsoft lasted 4 minutes and 50 second (± 2 sec), and 4 minutes 57 second (± 16 sec), respectively. A total of 30 sessions for each application was performed. From the Wireshark logs we noted that the bit-rate B for the Presentation application varied between 37 Kbps and 40 Kbps for the Google application, and between 25 Kbps and 30 Kbps for the Microsoft application.

Again, identical measurements were conducted to determine the bit-rate of Presentation with Google Drive for the case when the files are edited locally and then saved to the

Google Cloud. The bit-rate varied between 2.5 Kbps and 2.7 Kbps considering the total time for editing the file offline and transferring it to the Cloud. The bit-rate is calculated by dividing the observed exchanged traffic by the total time.

5.4.3 Bit-rate measurements for interactive Cloud-based Spreadsheet applications

We composed a Spreadsheet by entering numbers along 200 rows and 2 columns. The total number of characters (i.e. digits) was 700. We then performed basic numerical operations such as determining the min, max, mean, median and mode of the numbers. Subsequently, we plotted a (x, y) graph, and noted that the graph was updated dynamically as we sorted the numbers in each of the two columns. We repeated this measurement 30 times for each application. Each session on Google lasted 7 minutes and 34 second (± 2 sec), and each session on Microsoft lasted 9 minutes and 8 second (± 5 sec). The bit-rate B, obtained after post-processing the Wireshark logs, of Google Spreadsheet varied between 25 Kbps and 30 Kbps, while for Microsoft it varied between 110 Kbps and 150 Kbps.

These measurements were also repeated to quantify the bit-rate of Spreadsheet when the files are edited locally and then saved to the Google Drive Cloud. The bit-rate varied between 0.3 Kbps and 0.6 Kbps considering the total time for editing the file offline and transferring it to the Cloud. The bit-rate is calculated by dividing the observed exchanged traffic by the total time.

Table 5.1 summarizes the bit-rates of the different applications as obtained from our measurements. The substantial differences in the bit-rate between edit online and edit offline scenarios is due to the cost of incremental updates of file segments that occurs with the edit online scenario.

5.4.4 Average power consumption P_u of the netbook computer

The idle power consumed by the netbook computer with all network interfaces disabled was 10.8 W. We performed experiments at different times during the day (to address the

	Application	Bit-rate
Google Drive Edit online, Save in the cloud	Word processing	45-60 Kbps
	Presentation	37-40 Kbps
	Spreadsheet	25-30 Kbps
Microsoft Skydrive Edit online, Save in the cloud	Word processing	10-12 Kbps
	Presentation	25-30 Kbps
	Spreadsheet	110-150 Kbps
Google Drive Edit offline, Save in the cloud	Word processing	1.1-1.5 Kbps
	Presentation	2.5-2.7 Kbps
	Spreadsheet	0.3-0.6 Kbps

Table 5.1: Summary of bit-rates for Google and Microsoft OneDrive's (previously known as SkyDrive) Word processing, Presentation and Spreadsheet applications.

Application	Access network technology	Average power consumed by the Netbook computer
Google Drive Word Processing Edit online, Save in the cloud	Ethernet	13.6 W
	WiFi	14.0 W
	4G	16.1 W
Microsoft Skydrive Word Processing Edit online, Save in the cloud	Ethernet	14.4 W
	WiFi	14.5 W
	4G	16.7 W
Google Drive Word Processing Edit offline, Save in the cloud	Ethernet	13.2 W
	WiFi	13.7 W
	4G	15.1 W

Table 5.2: Average power consumed by the netbook computer for using Google and Microsoft's Word processing applications.

issue of variability in the situations the user may experience) on the interactive Cloud applications described in the previous section, and noted that the power consumption of the netbook computer was not sensitive to the time-of-day variation. Measurements were performed using three different access technologies available in the netbook, i.e. Ethernet, WiFi and 4G (via a USB dongle), and the power consumed in each of these cases was recorded.

P_u for Word processing applications

Column three in Table 5.2 gives the average power consumed by the netbook, P_u, for composing the 2-pages using Google and Microsoft's Word processing applications. We can see that 13.6 W is consumed when accessing the interactive Word processing application from Google using Ethernet. This increases to 16.1 W when using 4G high-speed wireless technology. A similar trend is observed with the Microsoft application.

P_u for Presentation applications

Table 5.3 shows the netbook's average power consumption to access the Cloud when composing 5-slides in the Presentation applications. We note that the power consumed by the netbook in this scenario is similar to that for the Word processing applications described above.

P_u for Spreadsheet applications

Table 5.4 shows the power consumption when composing the Spreadsheet. We note that P_u of Google Spreadsheet is greater than 16 W regardless of the type of access technology. The results in Table 5.4 also show that the netbook running Google Spreadsheet consumes less power than the netbook running Microsoft Spreadsheet although the bit-rate generated by Microsoft Spreadsheet is higher. It reveals that power consumption of end-user devices running Cloud applications is not only related to the applications bit-rate. Other parameters such as CPU load and utilization due to the design of applications can play determining roles in power consumption.

Energy per bit of routers and switches

Table. 5.5 lists the key network equipment (used in the metro, edge, core and data center networks) corresponding to Figure 5.1. The data was gathered from Cisco's power consumption calculator [95]. Column three represents the maximum capacity (i.e. c_t) of each device, the corresponding maximum power (i.e. p_t) is shown in column four, and the idle power (i.e. p_0), which is typically 90% of the maximum power [113], is denoted in

Application	Access network technology	Average power consumed by the Netbook computer
Google Drive Presentation Edit online, Save in the cloud	Ethernet	14 .0 W
	WiFi	14.2 W
	4G	16.1 W
Microsoft Skydrive Presentation Edit online, Save in the cloud	Ethernet	12.8 W
	WiFi	13.0 W
	4G	15.8 W
Google Drive Presentation Edit offline, Save in the cloud	Ethernet	13.4 W
	WiFi	13.9 W
	4G	15.3 W

Table 5.3: Average power consumed by the netbook computer for using Google and Microsoft's Presentation applications.

column five. The energy per bit (i.e. slope m) is shown in units of nJ/bit in column six. In the network depicted in Figure 5.1, we assume, using the *traceroute* utility, that there are $N_c = 5$ core routers and $N_e = 2$ edge routers on average along the path between the user and the Cloud data center server.

The energy per bit in the case of Ethernet access is approximately 3 nJ/bit; obtained from the data sheet of a Cisco 2960 series switch [114]. The energy per bit for WiFi access is taken to be 128 nJ/bit; obtained from a performance benchmarking study of the Cisco 1250 enterprise WiFi access point [115]. Estimating the energy per bit for a base station is non-trivial since it depends on a variety of different factors such as the number of concurrent users it can support, the deployment area, number of sectors, spectrum allocation, interference, among others. Our energy per bit figures are estimated from [116] by observing that a state-of-the-art 2012-technology 3-sector 2x2 MIMO remote radio head 4G/LTE base station deployed in an urban environment consumes 528 W under full load, and 333 W when idle. The aggregate achievable throughput of this base station is 72 Mbps with 20 MHz spectrum [117]. Further, [116] also reports that base stations consume different amounts of power in each direction (unlike the equipment listed in Table 5.5); roughly 87% of the energy is consumed in the downlink direction and the remaining 13% in the uplink direction. Considering a typical utilization of 5% over a 24-hour cycle, the energy per bit of this base station, on average, can be approximated as 76.2 μJ/bit in the

Application	Access network technology	Average power consumed by the Netbook computer
Google Drive Spreadsheet Edit online, Save in the cloud	Ethernet	16.1 W
	WiFi	16.6 W
	4G	17.8 W
Microsoft Skydrive Spreadsheet Edit online, Save in the cloud	Ethernet	14.3 W
	WiFi	14.7 W
	4G	16.2 W
Google Drive Spreadsheet Edit offline, Save in the cloud	Ethernet	13.4 W
	WiFi	14.3 W
	4G	15.2 W

Table 5.4: Average power consumed by the netbook computer for using Google and Microsoft's Spreadsheet applications.

Type	Model	Max capacity (c_t) (bidirectional)	Max power (p_t)	Idle power (p_0)	Energy per bit (slope m)
Core router	CRS-3	4480 Gbps	12300 W	11070 W	8.5 nJ/bit
Edge router	7609	560 Gbps	4550 W	4095 W	25.2 nJ/bit
BNG	ASR 9010	320 Gbps	1890 W	1701 W	18.3 nJ/bit
Ethernet Switch	Catalyst 6509	256 Gbps	1766 W	1589 W	21.4 nJ/bit
Data Centre Switch	Catalyst 6509	320 Gbps	2020 W	1818 W	19.6 nJ/bit

Table 5.5: Energy per bit of equipment in the metro, edge, core and data center networks of Figure 5.1.

downlink and 19.0 μJ/bit in the uplink.

Power consumption per user P_d of data center server

Obtaining precise information about Google and Microsoft servers is difficult because this information is not publicly available. We instead resort to the following approach to quantify the server power consumption per user. We note that Google's Word processing, Presentation and Spreadsheet applications are a part of the wider Google Apps service suite [109]. The power consumption of a server per user sharing the compute resources, as reported by Google, for the Google Apps services is about 0.25 W [118]. We therefore use this figure of 0.25 W in our calculations. Further, we assume that the per user power

	Word processing locally (i.e. in Microsoft Word)								
Average power consumed by the Netbook to compose document in Microsoft	**11.3 W**								
	Word Processing in Google Drive (Edit online, Save in the cloud)			**Word Processing in Microsoft Skydrive (Edit online, Save in the cloud)**			**Word Processing in Google Drive (Edit offline, Save in the cloud)**		
Power consumption of data centre server (P_d)	0.25 W			0.25 W			0.25 W		
Power consumption of transport network ($N_cE_cB + N_eE_eB + E_{bng}B + E_{sw}B + E_dB$)	8.4×10^{-3} W			1.7×10^{-3} W			0.2×10^{-3} W		
Access network	**4G**	**WiFi**	**Ethernet**	**4G**	**WiFi**	**Ethernet**	**4G**	**WiFi**	**Ethernet**
Power consumption of access network (E_aB)	1.9 W	7×10^{-3} W	0.2×10^{-3} W	0.4 W	1.4×10^{-3} W	35.3×10^{-6} W	0.05 W	0.4×10^{-5} W	4×10^{-6} W
Power consumption of Netbook (P_u)	16.1 W	14 W	13.6 W	16.7 W	14.5 W	14.4 W	15.1 W	13.7W	13.2 W
Average power consumed to use the cloud (i.e. sum of the power consumption of the data centre server, transport network, access network, Netbook)	18.3 W	14.3W	13.9 W	17.4 W	14.8 W	14.7 W	15.4 W	13.9W	13.4 W

Table 5.6: Power consumption per user P_I for using the Word processing application locally and in the Cloud.

consumption of a server in Microsoft's data center is also 0.25 W. This is a reasonable assumption because a typical server from Google or Microsoft that supports the types of applications considered in this study consumes about the same amount of power, i.e. $\approx$ 200 W [109, 119].

Energy per bit of access network

5.5 Power Consumption Per User P_I

We have used the values from the previous section in (5.1) to estimate the power consumption per user, P_I, incurred in using the Cloud applications. The access network power consumption for 4G is calculated as the sum of the power consumption of the 4G base station in the uplink and downlink directions.

5.5.1 P_I for Word processing applications

Table 5.6 summarizes our results for the case when the bit-rate B of the online interactive Word processing application from Google and Microsoft is 55 Kbps and 11 Kbps respectively. The bit-rate B of the Word processing application in Google when editing offline

and saving in the Cloud is 1.3 Kbps. The key points for Word processing from Table 5.6 are:

a) The average power consumption obtained from measurements for composing and saving the document locally on the netbook using Microsoft Word is 11.3 W.

b) When using the Cloud, the power consumption of the transport network is small compared to the contributions made by the other parts of the network. This is because the energy per bit of routers and switches is small (in the order of nJ per bit, see Table 5.5), and so is the bit-rate of the applications (a few tens of Kbps, see Table 5.1).

c) The power consumption of the access network is dominated by 4G (i.e. the 4G base stations), which is three to six orders of magnitude more than a WiFi modem or an Ethernet switch.

d) The power consumption of the netbook computer is a significant fraction of the overall power consumption incurred in using the Cloud applications.

e) We estimate the average power consumption per user – i.e. sum of the power consumption of the data center server, access and transport network, as well as the netbook computer – to use Google Drive and Microsoft OneDrive to vary between 13.9 W and 18.3 W for the former, and between 14.7 W and 17.4 W for the latter (depending upon the access technology used). The power consumption is between 13.4 W to 15.4 W for offline file editing and saving in the Google Drive Cloud.

f) Most importantly, online editing and saving the document in the Cloud consumes more power than offline editing and saving it to the Cloud. Both Cloud scenarios (online and offline editing) consume more power than processing and storing the document locally.

5.5.2 P_I for Presentation applications

Table 5.7 shows data for the Presentation application when the bit-rate B for online interaction with Google and Microsoft is 38 Kbps and 27 Kbps respectively. The important

	Processing presentation locally (i.e. in Microsoft PowerPoint)								
Average power consumed by the Netbook to compose presentation in Microsoft	**11.0 W**								
	Processing presentation in Google Drive (Edit online, Save in the cloud)			**Processing presentation in Microsoft Skydrive (Edit online, Save in the cloud)**			**Processing presentation in Google Drive (Edit offline, Save in the cloud)**		
Power consumption of data centre server (P_d)	0.25 W			0.25 W			0.25 W		
Power consumption of transport network ($N_cE_cB + N_eE_eB + E_{bng}B + E_{sw}B + E_dB$)	5.8×10^{-3} W			4.1×10^{-3} W			0.4×10^{-3} W		
Access network	**4G**	**WiFi**	**Ethernet**	**4G**	**WiFi**	**Ethernet**	**4G**	**WiFi**	**Ethernet**
Power consumption of access network (E_aB)	1.4 W	4.9×10^{-3} W	0.1×10^{-3} W	1.4 W	3.5×10^{-3} W	87×10^{-6} W	0.1 W	0.3×10^{-3} W	7×10^{-6} W
Power consumption of Netbook (P_u)	16.1 W	14.2W	14 W	15.8 W	13 W	12.8 W	15.3W	13.9W	13.4 W
Average power consumed to use the cloud (i.e. sum of the power consumption of the data centre server, transport network, access network, Netbook)	17.8 W	14.6W	14.3 W	17.5 W	13.3 W	13.1 W	15.6 W	14.1W	13.6 W

Table 5.7: Power consumption per user P_I for using the Presentation application locally and in the Cloud.

points for Presentation applications to emerge from Table 5.7 are:

a) The average power consumption for composing 5-slides locally on the netbook computer using Microsoft PowerPoint is 11.0 W.

b) As in the previous example, moving to the Cloud consumes very little power in the transport network, 4G dominates the access network power consumption, and the netbook computer's power consumption is a large fraction of the overall power consumption of the service.

c) The power consumption for using the Presentation application on the Cloud varies between 14.3 W and 17.8 W (for Google) and 13.1 W and 17.5 W (for Microsoft). The power consumption varies between 13.6 W and 15.6 W for offline file editing and saving on Google Drive.

5.5.3 P_I for Spreadsheet applications

Table 5.8 summarizes the results for the online interactive Spreadsheet application when the bit-rate B is 27 Kbps for Google and 130 Kbps for Microsoft. The bit-rate for the Spreadsheet application in Google Drive when editing offline and saving in the Cloud is 0.5 Kbps.

	Processing spreadsheet locally (i.e. in Microsoft Excel)								
Average power consumed by the Netbook to compose spreadsheet in Microsoft	**11.0 W**								
	Processing spreadsheet in Google Drive (Edit online, Save in the cloud)			**Processing spreadsheet in Microsoft Skydrive (Edit online, Save in the cloud)**			**Processing spreadsheet in Google Drive (Edit offline, Save in the cloud)**		
Power consumption of data centre server (P_d)	0.25 W			0.25 W			0.25 W		
Power consumption of transport network ($N_cE_cB + N_eE_eB + E_{bng}B + E_{sw}B + E_dB$)	4.1×10^{-3} W			19.8×10^{-3} W			0.07×10^{-3} W		
Access network	**4G**	**WiFi**	**Ethernet**	**4G**	**WiFi**	**Ethernet**	**4G**	**WiFi**	**Ethernet**
Power consumption of access network (E_aB)	1.0 W	3.5×10^{-3} W	87×10^{-6} W	5.8 W	17×10^{-3} W	0.4×10^{-3} W	0.02 W	0.06×10^{-3} W	0.1×10^{-6} W
Power consumption of Netbook (P_u)	17.8 W	16.6W	16.1 W	16.2 W	14.7 W	14.3 W	15.2W	14.3W	13.4 W
Average power consumed to use the cloud (i.e. sum of the power consumption of the data centre server, transport network, access network, Netbook)	19.1 W	16.9W	16.4 W	22.3 W	15.0 W	14.6 W	15.5 W	14.5W	13.7 W

Table 5.8: Power consumption per user P_I for using the Spreadsheet application locally and in the Cloud.

Composing the spreadsheet locally on the netbook computer using Microsoft Excel incurs 11.3 W, while composing the spreadsheet in the Cloud could incur an additional 11 W if using Microsoft via a 4G wireless access network. Other observations are similar to ones described above.

5.5.4 Key points

These series of measurements using Google Drive and Microsoft OneDrive's Word processing, Presentation and Spreadsheet applications demonstrate that using the Cloud could consume more power than local processing, implying that it is not always more energy-efficient to adopt the Cloud for performing tasks. When making this comparison it is important to note that interactive Cloud applications provide many benefits unrelated to energy efficiency. A prime example being collaborative document drafting and editing by geographically spread team members. Further, the end-user device and the access network, specifically high-speed wireless, can play a major role in determining the overall power consumption involved in using interactive Cloud-based applications.

	Word Processing in Google Drive (Edit online, Save in the cloud)			Word Processing in Microsoft Skydrive (Edit online, Save in the cloud)			Word Processing in Google Drive (Edit offline, Save in the cloud)		
Access network	**4G**	**WiFi**	**Ethernet**	**4G**	**WiFi**	**Ethernet**	**4G**	**WiFi**	**Ethernet**
Average power consumed to use the cloud	3.8 W	2.7 W	2.8 W	2.9 W	3.2 W	3.6 W	0.9 W	2.3 W	2.3 W

	Processing presentation in Google Drive (Edit online, Save in the cloud)			Processing presentation in Microsoft Skydrive (Edit online, Save in the cloud)			Processing presentation in Google Drive (Edit offline, Save in the cloud)		
Access network	**4G**	**WiFi**	**Ethernet**	**4G**	**WiFi**	**Ethernet**	**4G**	**WiFi**	**Ethernet**
Average power consumed to use the cloud	3.3 W	3.0 W	3.2 W	3.0 W	1.7 W	2.0 W	1.1 W	2.5 W	2.5 W

	Processing spreadsheet in Google Drive (Edit online, Save in the cloud)			Processing spreadsheet in Microsoft Skydrive (Edit online, Save in the cloud)			Processing spreadsheet in Google Drive (Edit offline, Save in the cloud)		
Access network	**4G**	**WiFi**	**Ethernet**	**4G**	**WiFi**	**Ethernet**	**4G**	**WiFi**	**Ethernet**
Average power consumed to use the cloud	4.6 W	5.3 W	5.3 W	7.8 W	3.4 W	3.5 W	1.0 W	2.9 W	2.6 W

Table 5.9: Power consumption per user for accessing the Word, Presentation and Spreadsheet applications in the Cloud assuming the user is already online.

5.5.5 Power Consumption when a user is already online

When a user is already online (i.e. connected to the Internet) undertaking other tasks, the network interfaces on the end-user device will already be energized irrespective of use of the interactive Cloud-based applications. Therefore, an equally valid viewpoint would be to calculate the power consumption for using the Cloud applications with ignoring the idle power of the netbook computer as well as the power consumed for enabling the network interfaces. The idle power of the netbook computer is 10.8 W and the power consumed for enabling the Ethernet, WiFi and 4G interfaces are 0.3 W, 0.8 W and 3.7 W respectively. Subtracting these values from the results given in Tables 5.6, 5.7 and 5.8 provides an estimate for the average power consumption involved in using the Cloud applications when a user is already online. These values are shown in Table 5.9.

To make the comparison fair, the power consumed for processing the tasks locally should be the results given in Tables 5.6, 5.7 and 5.8 for local processing less 10.8 W, the idle power consumption of the netbook computer. Thus, to compose a document, presentation and spreadsheet locally on the netbook would require 0.5 W, 0.2 W and 0.2 W. We note from Table 5.9 that the power consumption for Cloud-based processing using any of the three access network technology is still an order of magnitude larger than the

power consumed for local processing.

5.5.6 Power consumption using a Tablet as an end-User device

In addition to using a netbook computer, we carried out measurements using a Samsung Galaxy Tab 3 Lite, 7 inch tablet [110]. We were unable to replicate the scenarios described earlier in the tablet because the tablet-specific offerings of Google Drive and Microsoft OneDrive applications are still under development. For e.g., at the moment, Google does not support inserting pictures or tables in a browser launched from the tablet, and Microsoft does not have the edit online, save in the Cloud feature. We therefore composed a text-only document (same number of words as before) in the Word processing application of Google.

The idle power consumption of the tablet with all network interfaces disabled was 2.3 W. Enabling WiFi and the high-speed wireless interface (3G) increased the power consumption to 2.4 W and 2.5 W respectively; these values denote the baseline power consumption of the tablet. This tablet does not have an Ethernet interface. For the edit online, save in the Cloud scenario, the increase in the power consumption of the tablet, relative to the baseline, was 1.7 W (with WiFi) and 2.2 W (with 3G). For the edit offline, save in the Cloud scenario (performed using the Google Drive app), the increase over the baseline was 1.4 W (with WiFi) and 1.9 W (with 3G). These values give us the P_u in (5.1). Invoking (5.1) and noting that the bit-rate B of the application for each of the two scenarios is 28 Kbps and 5 Kbps on average, gives an estimate of the power consumption incurred in using the Cloud with the tablet. Assuming the user is already connected to the Internet, the power consumed for editing the document online is 2.0 W with WiFi and 3.3 W with 3G. The power consumption for editing the document offline and then saving it in the Cloud is 1.7 W with WiFi and 3.0 W with 3G. The power consumed to compose the document locally in the tablet (using the Polaris Office App) is 1.0 W.

These results show that even when the end-user device is a tablet (an example of a portable mobile device), processing a task in the Cloud could be less energy-efficient than processing the same task locally.

5.6 Discussion and Conclusions

The results of this chapter has shown that for our set of interactive Cloud-based applications, the network power consumption in the core and edge network is only a small fraction (< 1%) of our estimates of overall power consumption. This finding is consistent with the conclusions in [17] and [2] for low-rate traffic flows between the user and the Cloud. As a result, we do not expect our estimates to change significantly if the network topology and/or equipment change.

The model also shows that copying and pasting data from the local editor into the browser does not give rise to the large traffic overheads; the overheads arise from real-time interaction with the Cloud. Therefore, a way of improving service energy efficiency they would edit locally and only store to the Cloud once all the editing is completed. Alternatively there is scope for reducing the traffic overhead multiplier using intelligent client-side caching techniques, and optimizing the frequency with which synchronization of content occurs.

The results in this chapter rely on measurements of a netbook computer and a tablet that is representative of low-end user devices for Cloud access. Repeating the measurements on other devices could alter the estimates. Similarly, the results show that accessing Cloud services via WiFi or Ethernet will generally be less energy consuming than high-speed wireless (3G/4G), however the difference is such that the specific details of the access scenario may change this outcome.

Overall, this work showed that online interactive applications generate high amount of traffic and consume more energy than the same task on a non-interactive environment. Therefore, when online real-time collaboration is not required, it is more energy-efficient to do tasks locally and then save the final version to the Cloud.

In conclusion, we have comprehensively examined interactive Cloud-based applications and developed a model to estimate the average power consumption per user involved in using these applications. We have shown that the volume of traffic exchanged between the user and the Cloud can be considerably larger than that entered by the user, thereby impacting the power consumption of the service. Replacing a 70 W desktop PC (or a 30 W laptop) with a low-power consuming device and adopting the Cloud would indeed be

energy-efficient. However, our measurements demonstrate that simply migrating to the Cloud for processing tasks is not the always energy-wise choice, and it is therefore important to identify the right balance between performing tasks locally and in the Cloud for improving energy efficiency.

Chapter 6

Energy Consumption of Fog Computing Applications

Executive Summary

Tiny computers located in end-user premises are becoming popular as local servers for Internet of Things (IoT) and Fog computing services. These highly distributed servers that can host and distribute content and applications in a peer-to-peer (P2P) fashion are known as nano data centers. Despite the growing popularity of nano servers, their energy consumption is not well-investigated. To study energy consumption of nano data centers, we propose and use flow-based and time-based energy consumption models for shared and unshared network equipment, respectively. To apply and validate these models, a set of measurements and experiments are performed to compare energy consumption of a service provided by nano data centers and centralized data centers.

A number of findings emerge from our study, including the factors in the system design that allow nano data centers to consume less energy than their centralized counterpart. These include the type of access network attached to nano servers and nano server's time utilization (the ratio of the idle time to active time). Additionally, the type of applications running on nano data centers and factors such as number of downloads from users, number of updates, and amount of pre-loaded copies of data influence the energy cost. Our results reveal that number of hops between a user and content has little impact in the total energy consumption compared to the above-mentioned factors.

We show that nano servers in Fog computing can complement centralized data centers to serve certain applications, mostly IoT applications for which the source of data is in end-user premises, and lead to energy saving if the applications (or a part of them) are off-loadable from centralized data centers and run on nano data centers.

6.1 Introduction

Fog computing [23] is a new paradigm that refers to a platform for local computing, distribution and storage in end-user devices rather than centralized data centers [23]. This platform is becoming popular and even critical for wide range of applications, especially Internet of things (IoT), such as geo-distributed, mobile applications, real-time and latency-sensitive applications [23]. In this work we study very small servers known as "nano data centers" located in end-user premises for hosting and distributing content and applications in a peer-to-peer (P2P) fashion [4].

Fog computing is becoming an alternative to cloud computing for some applications [23]. But there has been little analysis, in the literature, of the energy consumption of Fog computing. There are different points of view on energy consumption of data storage and distribution from end-user premises in the literature. For example, in [25] and [4], it is claimed that this solution is more energy-efficient than sharing videos from centralized data centers. However, other works [26, 49] show that P2P content distribution from end-user premises consumes more energy than the centralized solution. This difference is largely due to different models for equipment energy consumption in different research work. In addition, some studies have either ignored the transport network or used an overly simple model of the transport network.

In this work, we aim to identify scenarios for which running applications from nano servers are more energy-efficient than running the same applications from centralized data centers using measurement-based models for network energy consumption that are more accurate than used in previous work. We first consider an end-to-end network architecture that includes all equipment required for distributing and accessing data from centralized data centers and nano data centers. We then derive comprehensive energy consumption models for content distribution. To do this, we propose a flow-based energy consumption model for shared network equipment and a time-based energy consumption model for network equipment located in the end-user premises which is not shared by many users.

To apply and validate our proposed models using experiments, we study the energy consumption of the application Wordpress [27] which can host content in servers within centralized data centers or servers in the end-user premises. Nano servers are implemented

using Raspberry Pi's (very small and low power single board computers) [20] and are characterized by traffic and power consumption measurements. Using the energy models, the energy consumption resulting from requesting data from a nano server is compared to that of the same request served from a server within a centralized data center.

Our results indicate that while nano data centers can save a small amount of energy for some applications by pushing content closer to end-users and decreasing the energy consumption in the transport network, they also can consume more energy when the nano servers are attached to an energy-inefficient access network or when the active time of dedicated nano servers is much greater than the idle time.

We investigate what type of applications can be run from nano servers to save energy. We find that parameters such as the number of downloads, the number of updates and the amount of data pre-loading play a significant role on the energy consumption of the applications. Our results show that the best energy savings using nano servers comes from applications that generate and distribute data continuously in end-user premises with low access data rate such as video surveillance applications.

Consequently, the most energy efficient strategy for content storage and distribution in cloud applications may be a combination of centralized data centers and nano servers. By identifying applications (or parts of there-of) best located in nano servers, rather than centralized data centers, the energy efficiency of those applications can be improved.

The rest of this section is organized as follows. The network topology and energy consumption models are elaborated in §6.2 and §6.3, respectively. §6.4 presents practical experiments and measurements. Energy consumption of centralized data centers and nano data centers is compared in §6.5. Parameters for executing applications efficiently in terms of energy cost on nano servers are explained in §6.6. Finally, this work is concluded in §6.7.

6.2 Network Topology

The end-to-end network topology for both centralized data centers and nano data centers is described in this section.

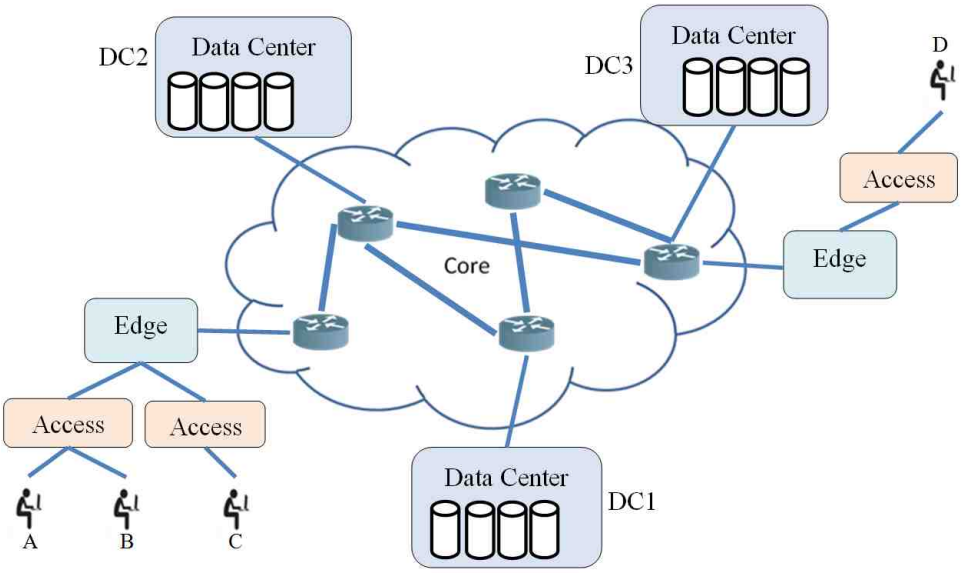


Figure 6.1: Network model of centralized data centers

6.2.1 End-to-end network model for centralized data centers

A cloud service provider has one or a few centralized data centers attached to the core of the network which host content as shown in Figure 6.1. The network within the data centers includes servers, storage, aggregation switches and one or more edge routers. Data center content is transported through large core routers and optical links to the edge network. The edge network generally consists of a metro Ethernet switch, broadband network gateways (BNGs) and edge routers. The content passes through an access network which might be an Ethernet, WiFi, PON, 3G or 4G connection, or a combination of these to reach the end-user terminal [11, 3, 120, 6].

6.2.2 End-to-end network model for nano data centers

In nano data centers architecture, there are no large, centralized data centers attached to the core network. Rather, each end-user is equipped with a device to host and distribute data. We may view the nano data centers approach as data storage and processing distributed amongst users with a piece of data allocated to each user as shown in Figure 6.2. Different network paths will be required for transporting content from the distributed

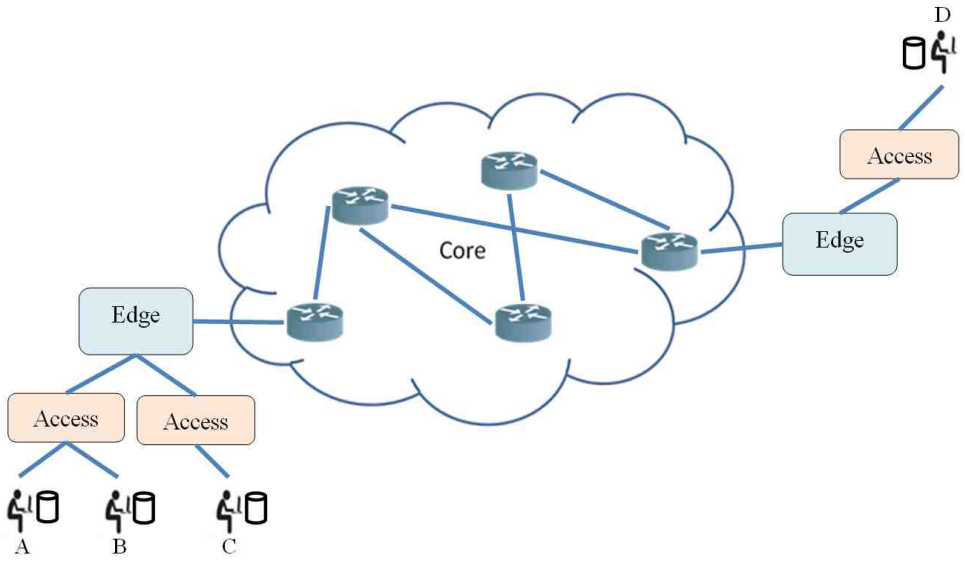


Figure 6.2: Network model of distributed nano servers

servers depending on the user's geographical location. The requests are either sent from (i) "home peers" who are users located in the premises of the nano server (such as user A and user B), (ii) "local peers" who are users located in the same ISP of the nano server (such as user A and user C), or (iii) "non-local peers" who are users located in a different geographical region away from the nano server (such as user A and user D).

As can be seen in Figure 6.2, for local and non-local peers, the content can be accessed by traversing two access networks (one is the access network for the users hosting the content and other is the access network for the users requesting the content). To reach the content from the local peers in the same geographic region, number of hops in the core and edge networks is less than the number required to access the content from a remote centralized data center. However, when accessing the content from a non-local peer, the number of core and edge router hops may be greater than the centralized data center scenario.

6.3 Energy Consumption Models

The network equipment are categorized into two types: 1) Equipment that are shared by many users and 2) customer premises equipment (CPE) dedicated to a single user (or

few users). For the highly shared equipment which deal with high amount of traffic we present a "flow-based" energy model that proportionally allocate the equipment's power consumption over all the flows through the equipment. For the equipment in end-user premises which are not shared by many users and services, we construct a "time-based" energy consumption model based upon the amount of time that equipment spends dealing with a Cloud service. All models are described in details in Chapter 3.

6.3.1 Centralized data centers and nano data centers

The energy consumed when using a service located in a centralized data center can be modeled by splitting it into three components: (a) energy consumption of end-user equipment for accessing the service. This includes the end-user terminals and access technology; (b) energy consumption of the transport network (aggregation, edge and core networks); and (c) energy consumption of the data center including its internal network, storages and servers.

The total energy consumed by service k provided from a centralized data center ($E_{\text{k-dc}}$) can be expressed as:

$$E_{\text{k-dc}} = E_{\text{k-cpe}} + E_{\text{k-access}} + E_{\text{k-edge}} h_{\text{e}} + E_{\text{k-core}} h_{\text{c}} + E_{\text{k-cent}} \tag{6.1}$$

where,

$E_{\text{k-cpe}}$, $E_{\text{k-access}}$, $E_{\text{k-edge}}$, $E_{\text{k-core}}$ and $E_{\text{k-cent}}$ are the energy consumed for service k in devices located in end-user premises, access network, energy per edge network element, energy per core network element and data centers, respectively. Parameters h_{e} and h_{c} are the number of edge and core routers traversed.

We have used the time-based energy consumption model for $E_{\text{k-cpe}}$ and applied the flow-based energy consumption models for $E_{\text{k-access}}$, $E_{\text{k-edge}}$, $E_{\text{k-core}}$ and $E_{\text{k-cent}}$. We also used flow-based energy consumption model for centralized data centers since the centralized data centers are shared by many users and services.

Parameters	**Description**
$E_{\text{k-dc}}$	Total energy consumption of service k provided by data centers
$E_{\text{k-ndc}}$	Total energy consumption of service k provided by nano data centers
$E_{\text{k-cpe}}$	Energy consumption of service k in CPE
$E_{\text{k-access}}$	Energy consumption of service k in the access network
$E_{\text{k-edge}}$	Energy consumption of service k per edge network element
$E_{\text{k-core}}$	Energy consumption of service k per edge network element
$E_{\text{k-cent}}$	Energy consumption of service k in centralized data centers
$E_{\text{k-nano}}$	Energy consumption of service k in nano data centers
$E_{\text{k-access2}}$	Energy consumption of access network attached to nano data centers
h_{e}	Number of hops in the edge network
h_{c}	Number of hops in the core network

Table 6.1: Notation for energy consumption of service k provided by data centers and nano data centers

In the case of nano servers, the energy consumption of the service consists of three components: (a) the energy consumed by end-user devices requested the content; (b) the energy consumption of the transport network between the end-user requesting data and the end-user hosting the data (access network is counted twice for local and non-local peers, once for each user); and (c) the energy consumed by the nano servers located in the end-users premises.

The total energy consumed by service k provided from nano data centers can be expressed as:

$$E_{\text{k-ndc}} = E_{\text{k-cpe}} + E_{\text{k-access}} + E_{\text{k-edge}} h_{\text{e}} + E_{\text{k-core}} h_{\text{c}} + E_{\text{k-access2}} + E_{\text{k-nano}} \tag{6.2}$$

where,

$E_{\text{k-access2}}$ is the energy consumed by access network attached to nano servers for service

k and $E_{\text{k-nano}}$ is energy consumption of service k in nano server devices located in the end-user premises. We have used the time-based energy consumption model for $E_{\text{k-nano}}$ because the nano servers are not shared by many users and services.

Using the expressions for device energy above and comparing (6.1) and (6.2), for a given end-user device and access technology, we note that the differences between energy consumption of a service provided from a centralized data center compared to nano data centers is primarily determined by the following:

- The number of bits exchanged between the user and data center (N_{bit});
- The number of hops for the two cases (h_{e}, h_{c});
- The value of $E_{\text{k-cent}}$ compared to $E_{\text{k-access2}} + E_{\text{k-nano}}$.

To evaluate this difference we require models for each of these contributions.

6.4 Measurements for Energy Models

To quantify the models for $E_{\text{k-dc}}$ and $E_{\text{k-ndc}}$, we use power and traffic measurements undertaken using the Wordpress [27] application which is an open source website and blogging tool. There are two options for Wordpress users: 1) Sign up for an account from the Wordpress website and connect to the Wordpress centralized data centers; 2) Install Wordpress software locally and create a web-server and host the content locally on a nano server.

The nano servers in the end-users premises were implemented using Raspberry Pi's [20]. Each Raspberry Pi has a SD card for storage and , if need be, an external hard drive can be attached to provide additional storage. The Raspberry Pis' low power draw, compact size and silent running make it a good choice for home servers [121].

6.4.1 Traffic measurements (N_{bit})

In order to determine the number of exchanged bits (N_{bit}) between an end-user and a data center or a nano server when uploading files to Wordpress or downloading the same files, we measured the volume of traffic using a packet analyzer (Wireshark) running on the end-user device.

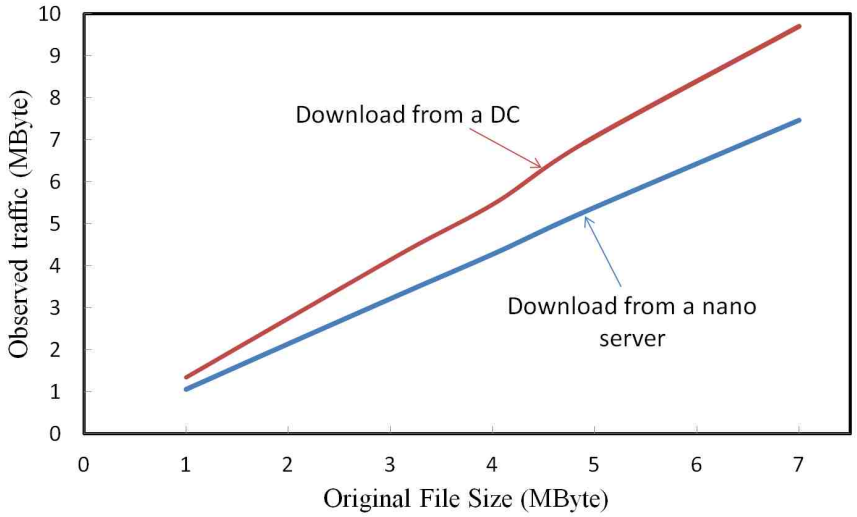


Figure 6.3: Exchanged bytes during downloading files varying in size from Wordpress website versus the original sizes of files

We uploaded files with their original sizes (without compression techniques) to both the data center and the nano server and downloaded the same files. Figure 6.3 shows the number of bytes exchanged during downloading various files, ranging from 1 MB to 7 MB, from the centralized data center and nano server versus their original size. Each session was repeated 10 times and the average traffic is displayed. The download curve for the nano server indicates the traffic exchanged is very similar to the original size of files. However, the traffic for downloading from the data center is higher than the original file size. Post-processing the Wireshark logs reveals that the download traffic from centralized data centers is higher than the original file size due to the existence of third party applications and advertisement traffic.

We also measured the upload traffic and found it was similar to download traffic although there are some cloud applications for which upload and download traffic are not the same; such as Google Drive and Facebook [120, 6].

6.4.2 Power measurements ($P_{\mathbf{cpe}}$)

The power consumption of end-user terminals and nano servers when interacting with the Wordpress website was measured directly using a power meter. We used a PowerMate

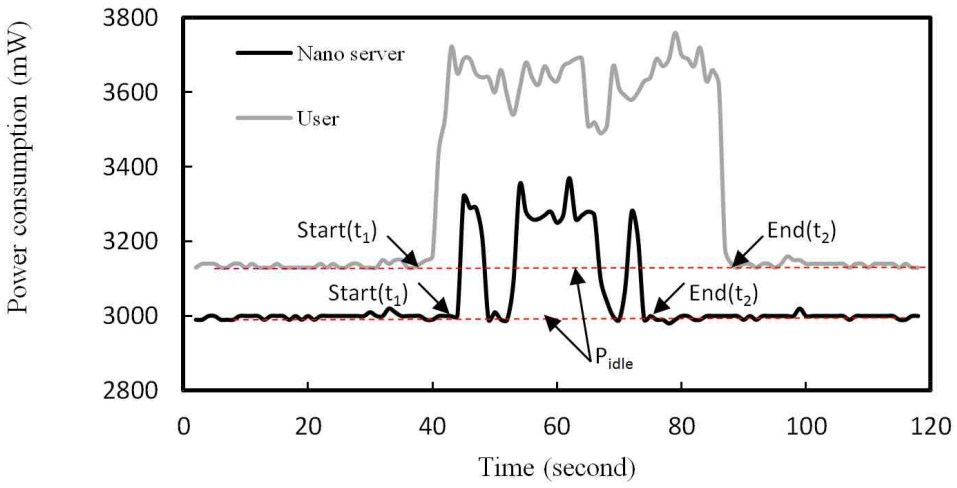


Figure 6.4: Power consumption of an end-user device and a nano server while uploading a file to Wordpress

meter [89] with a resolution of 10 mW during uploading and downloading of data.

We measured the power consumption of end-user devices while uploading and downloading different files to Wordpress data centers and local nano servers. We also measured the power consumption of nano servers. As an example, Figure 6.4 shows the power consumption of two Raspberry Pi's when uploading a 5MB file to the nano server. One Raspberry Pi is set as an end-user device and another is as a nano server. The baseline power consumption of the Raspberry Pi acting as the end-user device is higher than the baseline power consumption of the nano server because of a web browser running on the end-user device. Figure 6.4 displays the power (as a function of time) for uploading a file to the nano server. The sequence of events shown in Figure 6.4 for the upload was: first open the web browser in the end-user device and then upload a file (t_1 in the user curve). After that, the nano server starts to process and store the file (t_1 on the nano server curve). After storing, the local server status switches to idle mode (t_2 in the nano server curve). Then the end-user device completes the final processing after which it also switches to idle mode (t_2 on the user curve).

Similar power measurements have been done for determining the energy consumption for downloading from the nano server to the end-user device.

	Power(Watt)		Traffic(Gbps)		Energy(nJ/bit)	
	Idle	Max	Downlink	Uplink	Downlink	Uplink
Fast Ethernet gateway (CPE)	2.8	4.6	0.1	0.1	N/A	N/A
ADSL2+ gateway (CPE)	4.1	6.7	0.024	0.003	N/A	N/A
4G gateway (CPE)	0.5	1.75	0.024	0.012	N/A	N/A
GPON gateway (CPE)	5.2	8.3	2.4	1.2	N/A	N/A
Ethernet switch	1589	1766	256	256	31.7	31.7
LTE Base-station	333	528	0.072	0.012	76200	19000
OLT	43	48	2.4	2.4	88	179
BNG	1701	1890	320	320	27	27
Edge Router	4095	4550	560	560	37	37
Core Router	11070	12300	4480	4480	12.6	12.6

Table 6.2: Energy per bit of network equipment in access, edge and core networks

6.5 Energy Consumption Comparison

We can now compare the energy consumption of each component in (6.1) for a centralized data center and the corresponding components in (6.2) for a nano data center to ascertain the difference in energy consumption. In both cases we consider a service which is one service (service k) of multiple number of services.

6.5.1 User and access network equipment ($E_{\text{k-cpe}} + E_{\text{k-access}}$)

The access network includes (a) single user customer premises equipment (CPE) such as modems and (b) shared network equipment such as Ethernet switches and LTE base stations. Being customer equipment located in the homes, energy consumption of CPE is modeled using (3.17) in Chapter 3. Energy consumption of shared equipment, such as the OLT, Ethernet switch and base stations, is modeled using (3.9) in Chapter 3 with m set to unity representing a single access node in the data path.

We have studied several technologies by which the CPE may be connected to the access network: Ethernet, WiFi, 4G or PON. As one would expect, the measurement results indicate, for a given connection technology, the energy consumption of end-user device for uploading and downloading data to the centralized Wordpress data center is approximately equal to uploading and downloading data to the nano server.

The first four rows of Table 6.2 list the power consumption and throughput for CPE

when receiving data from the end-users (uplink) and transmitting data to the end-users (downlink). The idle power, maximum power, downlink traffic and uplink traffic of CPE were gathered from [93]. The corresponding values for shared access equipment are also provided in Table 6.2. The idle, maximum power and maximum capacity of Ethernet switch and OLT are gathered from [120] and [122], respectively. The energy per bit values for this equipment are calculated based on (3.7) in Chapter 3 assuming utilization $U = 50\%$.

The energy per bit for an LTE base station depends on factors such as the number of concurrent users, deployment area, spectrum width, interference, etc. The maximum and idle power consumption of a 3-sector 2×2 MIMO 4G/LTE base station deployed in an urban area are reported as 528 W and 333 W by [76]. It is also reported that base stations consume different amounts of power in each direction roughly 87% of the energy is consumed in the downlink direction and the remaining 13% in the uplink direction [76]. The aggregate achievable throughput of this base station is 72 Mbps with 20 MHz spectrum [94]. The energy per bit of this base station, considering a typical utilization of 5% over a 24-hour cycle, would be 76.2 μJ/bit in the downlink and 19.0 μJ/bit in the uplink on average.

6.5.2 Edge and core network equipment ($E_{\mathbf{k\text{-}edge}} h_{\mathbf{e}} + E_{\mathbf{k\text{-}core}} h_{\mathbf{c}}$)

The idle power, maximum power and capacity of equipment in the edge and core networks were gathered from [120] and the energy per bit values calculated using (3.7) in Chapter 3. To determine the values for the key network equipment we set $U = 50\%$. All values for equipment in the edge and core networks are summarized in the last three rows in Table 6.2.

According to (6.1) and (6.2), the energy consumed in the edge and core networks also depend on the number of hops in the edge and core networks (h_e, h_c). Using *traceroute* from end-user devices to the Wordpress servers, we estimate the average number of edge and core routers along the path between the end-users and servers within data centers to be 3 and 5, respectively. However, the number of hops in the case of nano servers depends on the location of end-users requesting the content relative to those hosting the content. The

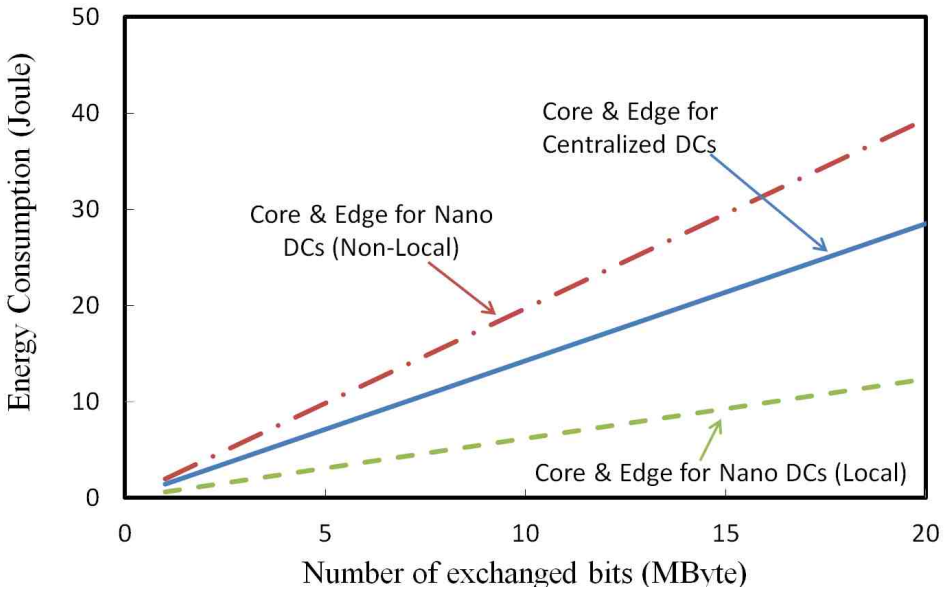


Figure 6.5: Consumed energy in the core and edge equipment for accessing data from different locations

requests are either served from (a) a nano server at the premises of the user placing the request (home peers), (b) a nano server in the same ISP (local peers), or (c) a remote nano server in a different geographical location (non-local peers-longest path). The number of edge and core routers for non-local peers are measured to be 3 edge and 8 core hops and 2 edge hops and 1 core hop for the peers setting in the same ISP (using *traceroute*).

Placing the number of hops and the energy per bit values of BNG, edge and core routers listed in Table 6.2 into (3.9), (6.1) and (6.2), we get Figure 6.5 which shows the energy consumed in the edge and core networks (as a function of N_{bit}) when accessing content from a data center (solid blue line) and a nano server hosted by a local peer (dashed green line) and a nano server hosted by a non-local peer (dot-dash red line). The figure indicates that the energy consumption resulting from requesting data from nano servers can be higher or lower than the energy consumed for accessing the content in centralized data centers depending on distance between the users and the stored content. The transport energy for home peers located in the same premises is zero because they do not pass edge and core routers.

6.5.3 Nano servers ($E_{\text{k-access2}}+E_{\text{k-nano}}$) and centralized servers ($E_{\text{k-cent}}$)

In Section 6.3.1 it was noted that one of the primary factors when comparing the energy consumption of a service provided from a centralized data center with providing it from a nano server was the value of $E_{\text{k-cent}}$ compared to $E_{\text{k-access2}} + E_{\text{k-nano}}$. In this sub-section, we compare the energy consumption of a service provided by a nano server and its attached access network with that of a server within a centralized data center. In this work, we assume there is always at least one service (service k) running from centralized data centers or nano data centers.

Equipment in a centralized data center is highly shared and so is quantified using energy per bit. However, obtaining detailed information about servers within data centers and its associated internal networks to provide a value for energy per bit is very difficult because detailed information on power consumption of the systems within commercial data centers is not publicly available. Two comprehensive articles on data center architecture and dimensioning can be found in [123] and [124], in which a model design, with numbers and types of network equipment and servers, is described. Using the capacity of the data centers described in this model, together with data center traffic characteristics from [125], and several realistic assumptions on server utilization (around 20% [126]) we developed estimates for data center energy consumption in the range 4-7 μJ/bit, excluding factors such as PUE and the need for replication. Including these factors increases the consumption to around 20 μJ/bit [126] for energy-efficient data centers (otherwise, it can be even higher).

In order to estimate the energy consumption of running a service from a Raspberry Pi [20] (as a nano server) it must be recognized that the Raspberry Pi is located in a home and hence connects via the access network. To include this contribution to the energy model, we have used (3.17) adopting the values listed in Table 6.2 for the access network and measurements for the Raspberry Pi.

Figure 6.6 shows energy consumption for serving data from centralized data centers and nano servers versus data traffic. A wide range of energy consumption values for centralized data centers are included in Figure 6.6 ranging from 4 μJ/bit to 20 μJ/bit which is indicated with an orange highlight. Nano servers with different access networks (GPON,

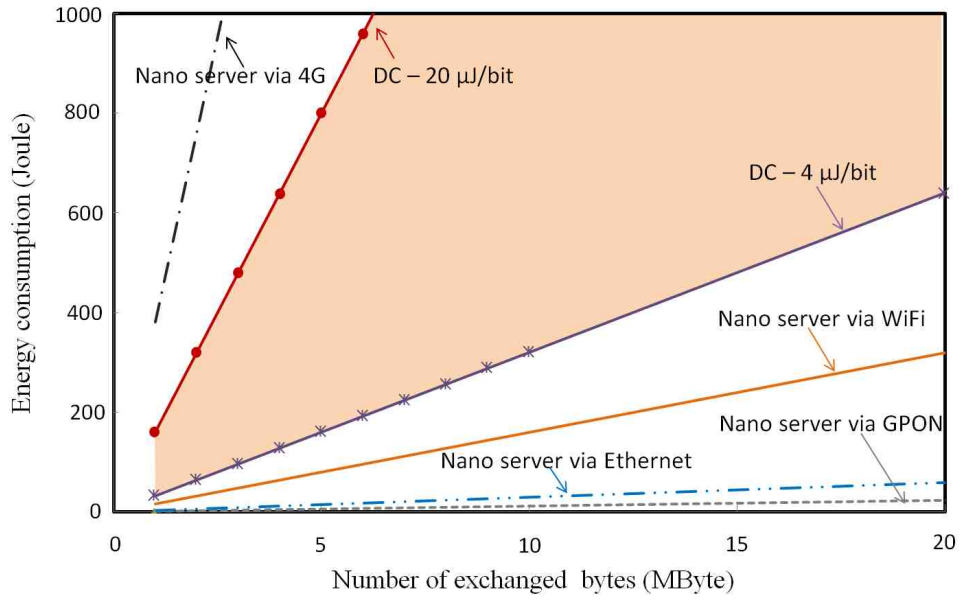


Figure 6.6: Energy consumed by service k in various nano servers and data centers as a function of the volume of data exchanged

Ethernet, WiFi and 4G) are also shown. It can be seen that the nano server attached to a 4G network consumes the greatest energy compared to others options, and a nano server attached to a GPON consumes the least energy. This figure indicates how the energy consumption of the access network can affect the energy consumption of a service provided by nano servers.

The values plotted in Figure 6.6 are based on the nano server being fully utilized serving multiple services. Hence the idle time is zero ($t_{\text{idle}} = 0$) and the ratio of the idle time of the device to the active time is zero ($\alpha = 0$). However, as we discussed in Section 6.3, devices in end-user premises are not highly shared and so may be idle for a significant proportion of time.

Therefore, to study the effect of active and idle time of equipment in end-user premises, we consider a nano server with WiFi access technology (ADSL2+ in end-user homes) and various idle times ($t_{\text{idle}} = 0,\ 5t_{\text{act}},\ t_{\text{act}} \Rightarrow \alpha = 0, 5, 10$). The energy consumption dependence on the data exchange for a service provided by a nano server with different proportions of idle time is compared with centralized data centers in Figure 6.7.

Although Figure 6.6 shows the energy consumption of the nano server connected via

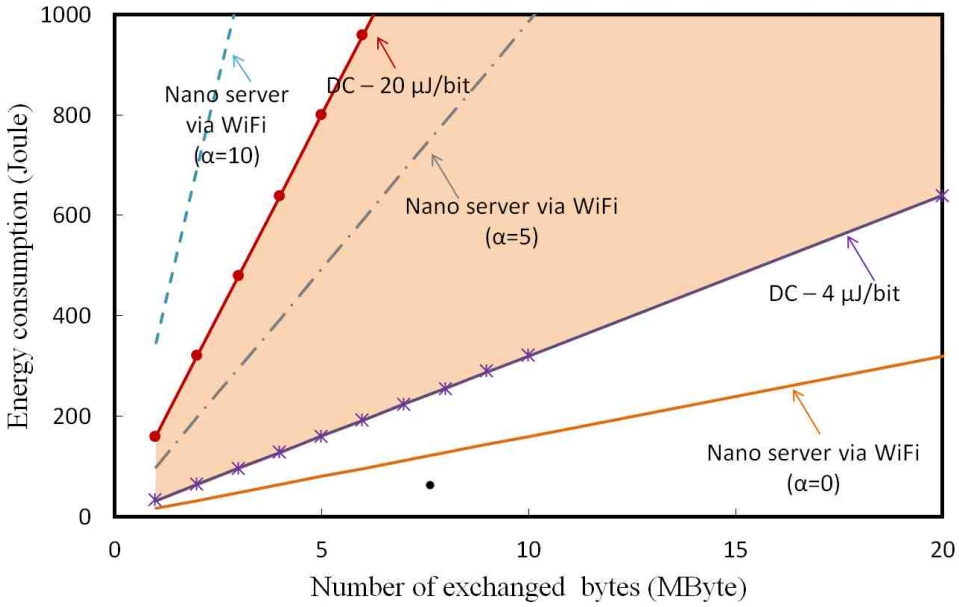


Figure 6.7: Energy consumed by service k provided by WiFi nano servers with different ratios of idle time to active time (α) as a function of the volume of data exchanged

WiFi can be less than that of a relatively energy efficient centralize data center, Figure 6.7 shows that without sharing the idle time of nano server with other services and with assigning more idle time to the service k (increasing α), the energy consumption of the service running on the nano server increases and dominates the energy consumption of running the same service from the data centers.

In Figure 6.7, the total time for the service, T_{tot}, is set to a constant. To calculate the lines for nano server energy consumption with constant α, we have assumed the total active time, t_{act}, is a constant and the amount of active time used by the service k, $t_{\text{act,k}}$, increases in proportion to the number of exchanged bytes for service k. From the results shown in Figure 6.6 and Figure 6.7, we see that the energy efficiency of a service using a nano server compared to a service using a centralized data center is not dependent on the number of byes exchanged. Rather, it is dependent upon factors such as the utilization of the nano server (α), the access technology used by the nano-server and the energy per bit of the centralized data center.

Therefore, managing the idle time of nano servers (i.e. sharing the idle time with multiple services or using sleep mode during the idle time) is a determining factor for

having low energy-consuming service k provided by nano data centers.

6.6 Nano Servers for Improving Energy Efficiency of Applications

We study three different types of applications: (i) applications for which the data source is primarily in end-user premises with static content such as hosting a static website; (ii) applications for which the source of data is primarily in end-user premises with dynamic content such as video surveillance; (iii) applications for which the source of data is not created in end-user premises but must pre-download (pre-load) to nano servers from other source(s) such as Video-on-Demand (VoD) applications.

6.6.1 Applications with static content for which the source of data is primarily in end-user premises

Applications with static content for which the source of data is primarily in end-user premises can be hosted and distributed from either nano servers or a centralized data center. In this case, we consider applications with static content (or with infrequent updates) and users download the content multiple times from a nano or centralize data center. The static content is a data file (such as a video file), which is downloaded N_{dl} times over a set duration. To run the applications with multiple downloads from nano servers and consume less energy than the centralized scenario, the following inequality must be met:

$$\begin{aligned} N_{\text{dl}}(E_{\text{dl-edge}}h_{\text{e}} + E_{\text{dl-core}}h_{\text{c}} + E_{\text{dl-access2}} + E_{\text{dl-nano}}) < \\ N_{\text{dl}}(E^{'}_{\text{dl-edge}}h^{'}_{\text{e}} + E^{'}_{\text{dl-core}}h^{'}_{\text{c}} + E_{\text{dl-cent}}) + \\ + N_{\text{up}}(E^{'}_{\text{up-edge}}h^{'}_{\text{e}} + E^{'}_{\text{up-core}}h^{'}_{\text{c}} + E_{\text{up-cent}}) \end{aligned} \tag{6.3}$$

where,

N_{dl} is number of downloads for the application from end-users and N_{up} is number of up-

dates for the application from its source. $E_{\text{dl-edge}}$, $E_{\text{dl-core}}$, $E_{\text{dl-access2}}$ and $E_{\text{dl-nano}}$ are the energy consumed per download in the edge network per network element, core network per network element, the access network attached to nano servers and nano servers, respectively. h_{e} and h_{c} are the number of hops in the edge and core networks in the nano scenario. $E^{'}_{\text{dl(/up)-edge}}$, $E^{'}_{\text{dl(/up)-core}}$ and $E_{\text{dl(/up)-cent}}$ are the energy consumed per download(/update) for the centralized data center scenario in the edge network, core network and a centralized data center. $h^{'}_{\text{e}}$ and $h^{'}_{\text{c}}$ are the number of hops in the edge and core networks in the centralized data center scenario.

Since we are considering applications with static content (or infrequent updates) in this section, we set $N_{\text{up}} = 1$ (or very low) in (6.3) and $N_{\text{up}}(E^{'}_{\text{up-edge}}h^{'}_{\text{e}} + E^{'}_{\text{up-core}}h^{'}_{\text{c}} + E_{\text{up-cent}})$ has a fixed value.

Figure 6.8 shows plots of the left and right hand sides of (6.3) showing the energy consumption as a function of the number of downloads, for an application running from centralized data centers with 4, 10 and 20 μJ/bit and a nano server attached to home WiFi access network with $\alpha = 5$. The energy consumption in Figure 6.8 includes the energy consumption of transport network and nano and centralized data centers. The size of file to be downloaded and uploaded is 100 MByte. For the nano server scenario, two user distributions are included:

(a) 20% of access events from non-local peers;

(b) 80% of access events from non-local peers.

Figure 6.8 indicates that the ratio of local to non-local requests has little impact in the total energy consumption. This is because the energy consumption of the application is dominated by the access network and data centers (nano or centralized). However, if the initial values of energy consumption of a data center and a nano server for hosting an application are close (such as the data center with 10 μJ/bit and the nano server attached to wireless network with $\alpha = 5$), the energy consumption due to the use of local or non-local peers can be a determining factor for which of centralize data center and nano data centers are more energy consuming. As shown in the figure for a limited number of downloads,

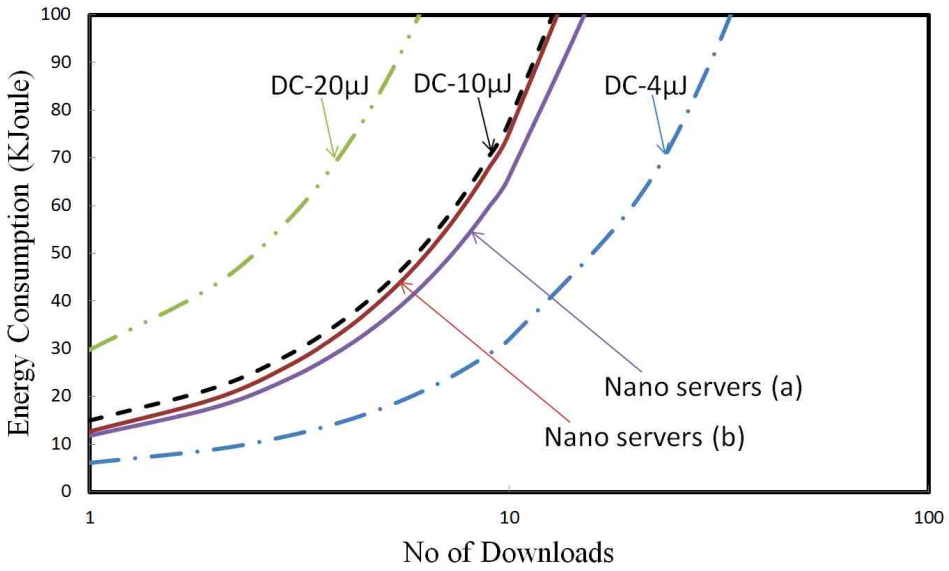


Figure 6.8: Energy consumption of an application running form nano and centralized DCs vs number of downloads to users

energy consumption of the nano server is less than the data center DC-10 μJ/bit and it is more energy-efficient to execute the application from the nano server. However, as the number of downloads from non-local peers rises (the red line in Figure 6.8 with 80% of access from non-local peers), the energy consumption of the transport network in the nano scenario increases quickly and the nano server cannot efficiently serve the application.

Referring to (6.3), if noting that in most cases that the energy consumption in the core and edge networks for each download(/update) is very small compared to energy consumption of a nano server or a data center ($E_{dl-edge}h_e + E_{dl-core}h_c \ll E_{dl-access2} + E_{dl-nano}$ and $E'_{dl(/up)-edge}h'_e + E'_{dl(/up)-core}h'_c \ll E_{dl(/up)-cent}$), then we can approximate (6.3) with $(E_{dl-access2} + E_{dl-nano}) < E_{dl-cent} + (N_{up}/N_{dl})E_{up-cent}$.

Therefore, for applications with low number of updates relative to downloads, $N_{dl} \gg N_{up}$, we get $E_{dl-access2} < E_{dl-cent} - E_{dl-nano}$. It shows that the key factor is the access network energy for the nano server being smaller than the difference between the data center and nano server. Under these circumstances, to first order, the location of the nano servers is not that important. What is important is the utilization of the nano servers (i.e. α) and the technology used to connect them to the network.

6.6.2 Applications with dynamic content for which the source of data is primarily in end-user premises

There are applications whose source of data is in end-user premises and content changes rapidly, such as applications for video monitoring in end-user homes. In this case we have $N_{up}/N_{dl} \geq 1$. We consider the energy consumption as a function of N_{up}/N_{dl} for these applications. To give a perspective on the dependence of energy consumption of a service on the ratio of idle to active time we include the α dependence (replace $E_{dl-nano}$ in (6.3) with (3.23)). We re-write (6.3) in the form:

$$\begin{aligned} E_{\text{dl-edge}}h_{\text{e}} + E_{\text{dl-core}}h_{\text{c}} + E_{\text{dl-access2}} + \\ P_{\text{idle}}(\alpha+1)t_{\text{act,k}} + \int_{t_{act,k}} (P(t) - P_{\text{idle}})\text{d}t < \\ (E^{'}_{\text{dl-edge}}h^{'}_{\text{e}} + E^{'}_{\text{dl-core}}h^{'}_{\text{c}} + E_{\text{dl-cent}}) + \\ (\frac{N_{\text{up}}}{N_{\text{dl}}})(E^{'}_{\text{up-edge}}h^{'}_{\text{e}} + E^{'}_{\text{up-core}}h^{'}_{\text{c}} + E_{\text{up-cent}}) \end{aligned} \tag{6.4}$$

Figure 6.9 shows per download energy consumption of an application running from the data center with 10 μJ/bit and the nano server (attached to home WiFi access network with 80% of downloads from non-local peers) plotted against N_{up}/N_{dl} and α. It can be seen that as the number of updates increases, the nano server is more energy-efficient than the centralized data center for running the application even when the idle time of nano server is relatively high (i.e. $\alpha \gg 1$). Therefore, the ratio of updates to downloads of an application plays an important role in the relative energy consumption of providing an service from a centralized data center compared to a nano data center.

For example applications such as video surveillance for which the image/video is continuously updated, it is not energy-wise to upload every update to the centralized data center. If the data generated by video monitoring is hosted in nano servers even when users access that data remotely (via the network) the energy consumption using a nano data center is still less than uploading the data to a centralized cloud and accessing it from there. Consequently, applications with a higher upload rate and low download rate are more energy-efficient when provided via on the nano servers architecture.

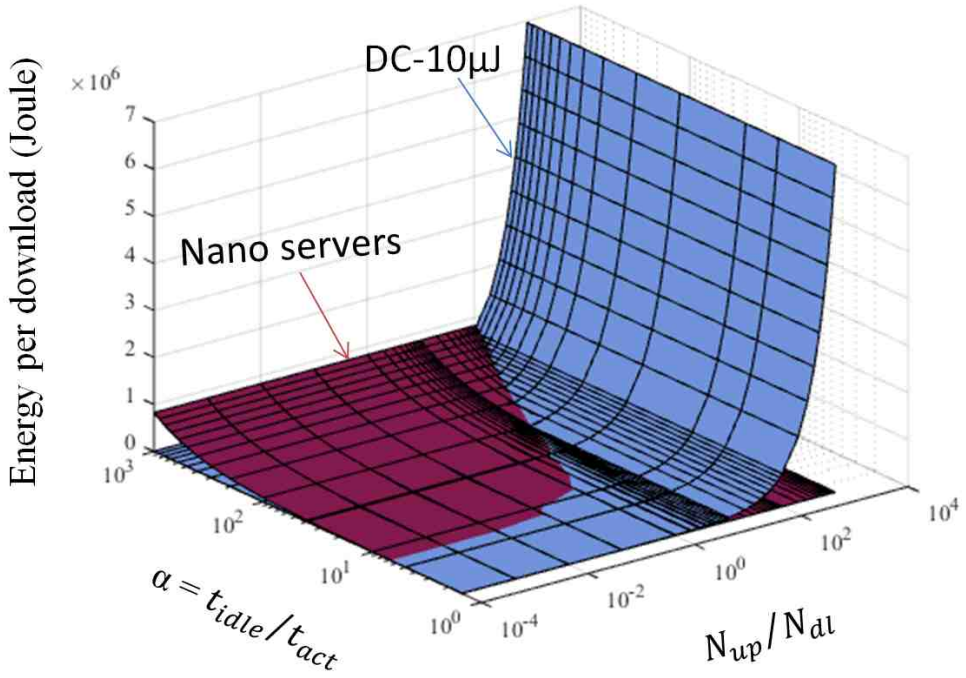


Figure 6.9: Energy consumption of an application running form a nano data center and data center considering number of downloads and updates

6.6.3 Applications requiring data pre-loading

In this section we assume all accesses to the application run on nano servers are 50% from local peers (not home peers) and 50% from non-local peers. Nano servers can also host and distribute data that is sourced outside of end-user premises such as Video on Demand (VoD) data. The general concept of reducing energy consumption by these applications is to push data closer to end-users to reduce the transport network energy consumption.

If we assume we have a nano server attached to an energy-efficient access network and the nano server has enough available time to host VoD efficiently, it is necessary to consider energy consumption of data pre-loading. The source of data to be pre-loaded will be either a server in a centralized data center or a content delivery network (CDN). The pre-loading process consumes energy which needs to be included in the model.

The energy consumption of an application (application k) provided by nano data centers which requires data pre-loading is given by:

$$E_{\text{k-pl}} = N_{\text{pl}}(E_{\text{pl-edge}}h_{\text{e}} + E_{\text{pl-core}}h_{\text{c}} + E_{\text{pl-access2}} + E_{\text{pl-nano}})+ \\ N_{\text{dl}}(E_{\text{dl-edge}}h_{\text{e}} + E_{\text{dl-core}}h_{\text{c}} + E_{\text{dl-access2}} + E_{\text{dl-nano}}) \tag{6.5}$$

where,

N_{pl} is the number of data pre-loadings and N_{dl} is the number of downloads for the application from other end-users. $E_{\text{pl(/dl)-edge}}$, $E_{\text{pl(/dl)-core}}$, $E_{\text{pl(/dl)-access2}}$ and $E_{\text{pl(/dl)-nano}}$ are the energy consumed for each data pre-loading (/per download) in the edge network, core network, access network attached to the nano server and the nano server.

The energy per download for the application with data pre-loading is given by:

$$\frac{E_{\text{k-pl}}}{N_{\text{dl}}} = \frac{N_{\text{pl}}}{N_{\text{dl}}}(E_{\text{pl-edge}}h_{\text{e}} + E_{\text{pl-core}}h_{\text{c}} + E_{\text{pl-access2}} + E_{\text{pl-nano}})+ \\ E_{\text{dl-edge}}h_{\text{e}} + E_{\text{dl-core}}h_{\text{c}} + E_{\text{dl-access2}} + E_{\text{dl-nano}} \tag{6.6}$$

Figure 6.10 shows the energy consumed per download of two nano servers(one requiring data pre-loading and the other not) as a function of the ratio the number of data pre-loading to downloads. As shown in the figure, the energy consumption increases as the number of data pre-loadings to number of downloads increases. As Figure 6.10 indicates, the number of pre-loaded data should be consistent to the number of downloads ($\frac{N_{pl}}{N_{dl}} \leq 1$) to execute an energy-efficient application on nano servers. It means that popular contents with more number of downloads for each data pre-loading are more energy-efficient to be run by nano data centers compared to unpopular contents. Therefore, the number of instances of pre-loaded content to the nano servers without downloads causes energy-efficient nano data centers consume high amount of energy, even when using energy-efficient access networks such as Ethernet or PON.

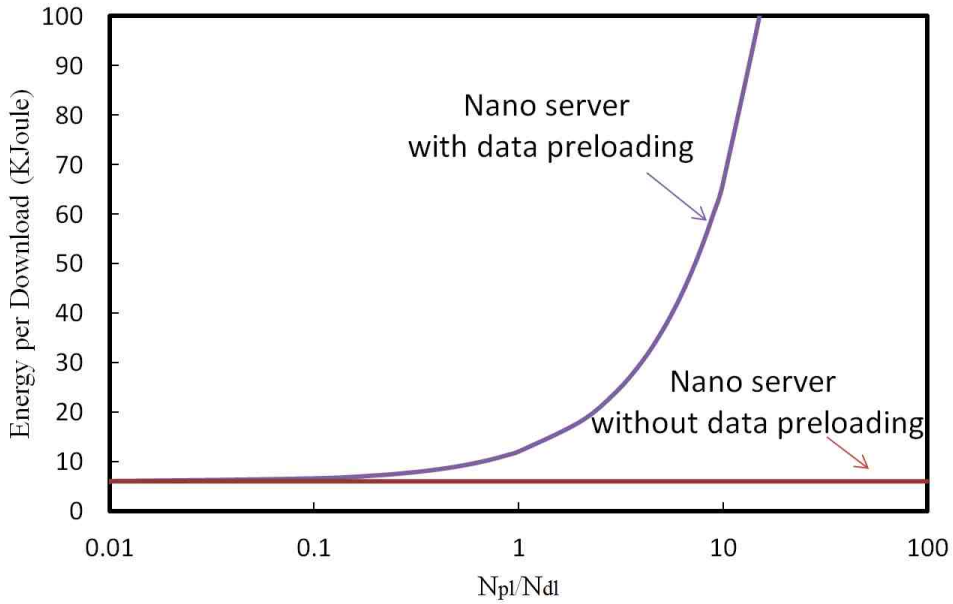


Figure 6.10: Energy consumption versus number of data pre-loading to number of downloads ($\frac{N_{pl}}{N_{dl}}$)

6.7 Conclusion

This work has compared the energy consumption of applications using centralized data centers in cloud computing with applications using nano data centers used in Fog computing. To do this, new energy models for shared and unshared network equipment were introduced and a set of measurements and experiments were used to provide data for the models.

Our results indicate that nano data centers might lead to energy savings depending on system design factors such as (a) type of access network attached to nano servers, (b) the ratio of active time to idle time of nano servers and, (c) type of applications which includes factors like number of downloads from other users, number of updates from the origin(s) and number of data pre-loading. It was also shown that number of hops between users and content has a little impact compared to the above-mentioned factors.

The results of this work show that the best energy savings using nano data centers is for applications that generate and distribute a large amount of data in end-user premises which is not frequently accessed such as video surveillance in end-users homes.

The deployment of nano data centers is occurring with the introduction of Fog com-

puting and implementation of smart devices to end-user homes for Internet of Things (IoT) services. To take advantage of the new architecture and to complement centralized data centers, we should identify applications that are more energy-efficient when provided from nano servers and run them on this platform. In addition to saving energy by running some applications on the nano platform, a portion of energy currently consumed within data centers for serving such applications can be saved.

Chapter 7

Conclusions and Future Directions

This chapter summarizes the research work in this thesis which is about the total energy consumption of Cloud Computing and Fog Computing applications and highlights the main findings. It also discusses open research problems in the area and outlines a number of future research directions.

7.1 Conclusions and Discussion

Cloud computing and content-based applications and services are the new wave transforming the IT industry. Home users and enterprise customers are increasingly being offered applications that run in the Cloud rather than stand-alone computers. Since Cloud computing provides numerous benefits such as accessibility, performance, reduced capital cost, reduced maintenance cost, etc ..., Cloud-based applications and services have become very popular among both home and enterprise customers. This has led to significant increase in traffic in data centers and the network.

With increasing demand and usage of Cloud-based applications and services, their energy consumption has become a major issue. In this thesis, we investigated the energy consumption of Cloud-based applications and made four main contributions to the body of work:

- New measurement-based energy consumption models were proposed (Chapter 3).
- The total energy consumption of online social networks (OSNs) was investigated and analyzed (Chapter 4).
- The total energy consumption of interactive Cloud-based applications was examined and compared with the energy consumption of local counterparts (Chapter 5).

- The energy consumption of content and application distribution from highly distributed servers was examined and the key parameters that could influence the energy consumption of this system were introduced (Chapter 6).

In Chapter 4, the energy consumed in the transport network and end-user devices during photo sharing on Facebook was studied. In order to examine the energy consumed in the transport network, we used an energy consumption model for shared network elements (presented in Chapter 3) and traffic measurements. To obtain the consumed energy in end-user devices, power consumption measurements for various types of end-user devices were performed. Based on the current profile of access technologies used by Facebook users, the estimated annual energy consumption in the transport network and end-user devices for Facebook photo sharing is about 304 GWh. The energy consumption of all Facebook data centers in 2012 was reported 500 Gwh. The energy consumption incurred in the transport network and end-user devices obtained by this work was about 60% of the energy consumption of all Facebook data centers. This figure would be higher if we could compare our estimate with just the fraction of data center energy consumption attributed to the photo sharing service only, however, Facebook did not explicitly report the energy consumption of their data centers for specific services such as photo sharing.

Chapter 4 has shown that the energy consumption of transport network and end-user terminals in Cloud-based applications and services is not trivial and should be taken into account when obtaining the total energy consumption. As a result, achieving an energy-efficient Cloud service requires improving the energy efficiency of the transport network and the end-user devices along with the related data centers.

In Chapter 5, the total energy consumption (including energy consumed in the end-user terminal, transport network and data centers) of interactive Cloud applications such as Google Drive and Microsoft OneDrive was studied. A combination of energy consumption modeling (introduced in Chapter 3) and measurement techniques were performed. It was observed from our traffic measurements that the volume of traffic exchanged between end-users and the Cloud is considerably larger than the size of the document being composed (possibly more than a factor of 1000). In order to compare the energy consumption of online interactive Cloud application to the local version of the applications, three scenarios were considered:

a) Creating, editing and saving Word, Presentation and Spreadsheet files in the Cloud;

b) Creating and editing the applications locally, and then saving the files in the Cloud;

c) Performing the tasks locally (i.e., the Cloud is absent). All the tasks are performed on the same low-power consuming end-user devices without saving to the Cloud.

This section showed that online interactive applications generate a substantial amount of traffic and consume more energy than the same task on a non-interactive environment. As a result, migration to the Cloud is not always more energy-efficient than working locally offline. Performing certain tasks locally can be more energy-efficient than using the Cloud. Overall, the results of this chapter indicated that Cloud application developers, network designers and general users can contribute to saving power consumption using Cloud-based services.

Cloud apps developers can save energy if they can

- Reduce the traffic overhead using intelligent client-side caching techniques;
- Optimize the frequency of synchronization between users and Cloud.

Network designers can gain insights from the results of this work for future energy-efficient deployment of Cloud services and applications by

- Improving the energy-efficiency of the access network especially wireless 3G/4G/LTE.

The results of this chapter can increase awareness of general users for energy-efficient behaviors. For example by

- Editing files locally, then saving to the Cloud;
- Using WiFi rather than 3G/4G if available;
- Using Shared WiFi if possible.

In Chapter 6, energy consumption of Fog computing was investigated. The term Fog computing refers to executing applications from nano servers located at end-user premises rather than centralized data centers. To analyze the energy consumption of this architecture, we developed energy models for shared and unshared network equipment (introduced in Chapter 3) and performed a set of measurements and practical experiments. The

analysis shows that the consumption induced in core and edge network is negligible as compared to the consumption of access networks, and it is thus quite irrelevant whether local or remote peers are used in the distributed case. The relevant parameters for consumption are:

- The access technology used by nano servers (LTE being orders of magnitude worse);
- The idle-to-active ratio for equipment on end-user premises;
- The type of applications, in particular its upload to download ratio.

Therefore, Fog computing can be more efficient than Cloud computing if LTE is avoided as access technology, reducing the idle time of customer premises equipment (by either sharing CPE with other applications or powering down during idle), or for applications with frequent updates and relatively seldom data consumption (i.e. a high upload to download ratio) such as surveillance video, for which nano data centers are more energy efficient even for high idle times.

The chapter has concluded with the outcome that nano servers in Fog computing can complement centralized data centers in Cloud computing to serve certain applications and offload the applications (or a part of them) from centralized data centers and run them on nano data centers in order to save energy.

7.2 Future Research Directions

Despite the contributions of the current thesis in energy consumption of Cloud computing and Fog computing, there are a number of open research challenges that need to be addressed in order to further advance the area.

We investigated the total energy consumption of Google Drive, Microsoft OneDrive and Facebook which are Software-as-a-Service (SaaS). In addition to SaaS, Cloud computing providers offer Platform-as-a-Service (PaaS) and Infrastructure-as-a-Service (IaaS). Therefore, as the energy consumption modeling and measurement techniques proposed in this thesis can be applied to PaaS and IaaS, it is valuable to study energy consumption of PaaS and IaaS in end-user terminals, transport network and data centers. Furthermore,

our results are based on energy consumption of applications during use-phase and we did not study energy consumption of applications and services in their entire life. Research considering a life cycle perspective would be required to examine the total environmental footprint of the applications and services.

Regarding nano data centers and Fog computing, there will be demand for installing security software on local gateways therefore further research is required to justify the additional energy for security. In addition, it is valuable to explore energy consumption of applications that partially run on content delivery network (CDN) architecture and partially run on nano servers. In addition, energy consumption of applications that need to be run on several nano servers in parallel to complete a task merit further investigations. Other factors such as intelligent content placement will require further study optimize the energy efficiency of applications running from nano data centers.

Another area that needs work is how to develop energy models for services that use a combination of real and virtualized network resources. As there is a move toward NFV (Network Function Virtualization), this will become an issue of interest especially as the use of GPPs (General Purpose Processors) becomes widely adopted in place of ASICs (Application Specific Integrated Circuits). It is expected that ASICs consume less energy compared to GPPs, but using virtualization allows multiple services to share GPP resources. Therefore, the trade-off between energy consumption and virtualization merits further study.

Bibliography

[1] Cisco white paper, "Cisco Global Cloud Index: Forecast and Methodology, 2013-2018," 2014.

[2] J. Baliga, R. Ayre, K. Hinton, and R. Tucker, "Green cloud computing: Balancing energy in processing, storage, and transport," *Proceedings of the IEEE*, vol. 99, no. 1, pp. 149–167, 2011.

[3] U. Lee, I. Rimac, D. Kilper, and V. Hilt, "Toward energy-efficient content dissemination," *Network, IEEE*, vol. 25, no. 2, pp. 14–19, March 2011.

[4] V. Valancius, N. Laoutaris, L. Massoulié, C. Diot, and P. Rodriguez, "Greening the internet with nano data centers," in *Proceedings of the 5th international conference on Emerging networking experiments and technologies*, ser. CoNEXT '09, 2009, pp. 37–48.

[5] K. Hinton, F. Jalali, and A. Matin, "Energy consumption modelling of optical networks," *Photonic Network Communications*, vol. 30, no. 1, pp. 4–16, 2015.

[6] F. Jalali, C. Gray, A. Vishwanath, R. Ayre, T. Alpcan, K. Hinton, and R. Tucker, "Energy consumption of photo sharing in online social networks," in *Cluster, Cloud and Grid Computing (CCGrid), 2014 14th IEEE/ACM International Symposium on*, May 2014, pp. 604–611.

[7] Technical talk by Jay Parikh, open compute summit 2013. [Online]. Available: www.opencompute.org/OCP-SUMMIT-IV-VIDEOS/

[8] Social news daily. [Online]. Available: http://socialnewsdaily.com/7064/1-1-billion-facebook-photos-uploaded-new-years-eve-and-new-years-day/

[9] Youtube statistics. [Online]. Available: https://www.youtube.com/yt/press/statistics.html

[10] C. Ge, Z. Sun, and N. Wang, "A survey of power-saving techniques on data centers and content delivery networks," *Communications Surveys Tutorials, IEEE*, vol. 15, no. 3, pp. 1334–1354, Third 2013.

[11] L. Liu et al., "GreenCloud: A New Architecture for Green Data Center," in *Proc. ACM ICAC-INDST*, Spain, Jun 2009.

[12] M. Ali, "Green Cloud on the Horizon," in *Proc. CloudCom*, China, Dec 2009.

[13] Y. Gu, V. March, and B. S. Lee, "GMoCA: Green Mobile Cloud Applications," in *Proc. 1st Int. Wkshp. on Green and Sustainable Software (GREENS)*, Switzerland, Jun 2012.

[14] T. Singh and P. Vara, "Smart metering the clouds," in *Enabling Technologies: Infrastructures for Collaborative Enterprises, 2009. WETICE '09. 18th IEEE International Workshops on*, June 2009, pp. 66–71.

[15] J. Liu, F. Zhao, X. Liu, and W. He, "Challenges towards elastic power management in internet data centers," in *Distributed Computing Systems Workshops, 2009. ICDCS Workshops '09. 29th IEEE International Conference on*, June 2009, pp. 65–72.

[16] W. Vereecken, L. Deboosere, D. Colle, B. Vermeulen, M. Pickavet, B. Dhoedt, and P. Demeester, "Energy efficiency in telecommunication networks (invited paper)," in *Proceedings of NOC2008, the 13th European Conference on Networks and Optical Communications*, 2008, pp. 44–51. [Online]. Available: http://dx.doi.org/1854/12590

[17] D. R. Williams and Y. Tang, "Impact of Office Productivity Cloud Computing on Energy Consumption and Greenhouse Gas Emissions," *Environmental Science and Technology*, vol. 47, pp. 4333–4340, 2013.

[18] D. R. Williams, P. Thomond, and I. Mackenzie, "The greenhouse gas abatement potential of enterprise cloud computing," *Environmental Modelling Software*, vol. 56, pp. 6 – 12, 2014.

[19] E. Masanet, A. Shehabi, L. Ramakrishnan, J. Liang, X. Ma, B. Walker, V. Hendrix, and P. Mantha, "The energy efficiency potential of cloud-based software: A us case study," 2013.

[20] Raspberry pi - a credit-card-sized single-board computer. [Online]. Available: www.raspberrypi.org/

[21] J. Malmodin, s. Moberg, D. Lundn, G. Finnveden, and N. Lvehagen, "Greenhouse gas emissions and operational electricity use in the ict and entertainment media sectors," *Journal of Industrial Ecology*, vol. 14, no. 5, pp. 770–790, 2010.

[22] J. Malmodin, D. Lundn, s. Moberg, G. Andersson, and M. Nilsson, "Life cycle assessment of ict," *Journal of Industrial Ecology*, vol. 18, no. 6, pp. 829–845, 2014.

[23] F. Bonomi, R. Milito, P. Natarajan, and J. Zhu, "Fog computing: A platform for internet of things and analytics," in *Big Data and Internet of Things: A Roadmap for Smart Environments*, ser. Studies in Computational Intelligence. Springer International Publishing, 2014, vol. 546, pp. 169–186.

[24] N. Laoutaris, P. Rodriguez, and L. Massoulie, "Echos: Edge capacity hosting overlays of nano data centers," *SIGCOMM Comput. Commun. Rev.*, vol. 38, no. 1, Jan. 2008.

[25] S. Nedevschi, S. Ratnasamy, and J. Padhye, "Hot data centers vs. cool peers," in *Proceedings of the 2008 conference on Power aware computing and systems*, ser. HotPower'08, 2008, pp. 8–8.

[26] A. Feldmann, A. Gladisch, M. Kind, C. Lange, G. Smaragdakis, and F. Westphal, "Energy trade-offs among content delivery architectures," in *Telecommunications Internet and Media Techno Economics (CTTE), 2010 9th Conference on*, 2010, pp. 1–6.

[27] Wordpress - website and blogging tool. [Online]. Available: http://wordpress.org/

[28] F. Jalali, R. Ayre, A. Vishwanath, K. Hinton, T. Alpcan, and R. Tucker, "Energy consumption of content distribution from nano data centers versus centralized data centers," *SIGMETRICS Perform. Eval. Rev.*, vol. 42, no. 3, pp. 49–54, Dec. 2014.

[29] F. Jalali, "Hidden energy consumption of photo sharing in online social networks," in *14th Annual Grace Hopper Celebration of Women in Computing (GHC'14)*, October 2014.

[30] A. Vishwanath, F. Jalali, R. Ayre, T. Alpcan, K. Hinton, and R. Tucker, "Energy consumption of interactive cloud-based document processing applications," in *Communications (ICC), 2013 IEEE International Conference on*, Hungary, Jun 2013.

[31] A. Vishwanath, F. Jalali, K. Hinton, T. Alpcan, R. Ayre, and R. Tucker, "Energy consumption comparison of interactive cloud-based and local applications," *Selected Areas in Communications, IEEE Journal on*, vol. PP, no. 99, 2015.

[32] F. Jalali, "Energy consumption of cloud applications," in *Asia-Oceania Top University League of Engineering (AOTULE'14)*, November 2014.

[33] F. Jalali, "Home Servers Can Save Energy for IoT Applications," in *15th Annual Grace Hopper Celebration of Women in Computing (GHC'15)*, October 2015.

[34] E. Nygren, R. K. Sitaraman, and J. Sun, "The Akamai network: a platform for high-performance internet applications," *SIGOPS Oper. Syst. Rev.*, vol. 44, no. 3, pp. 2–19, Aug 2010.

[35] U. Lee, I. Rimac, and V. Hilt, "Greening the internet with content-centric networking," in *Proceedings of the 1st International Conference on Energy-Efficient Computing and Networking*, ser. e-Energy '10, 2010, pp. 179–182.

[36] P. Gill, M. Arlitt, Z. Li, and A. Mahanti, "The flattening internet topology: Natural evolution, unsightly barnacles or contrived collapse?" in *Passive and Active Network Measurement*, ser. Lecture Notes in Computer Science. Springer Berlin Heidelberg, 2008, vol. 4979, pp. 1–10.

[37] Akamai, www.akamai.com/html/about/facts_figures.html, 2015.

[38] R. Cascella, C. Morin, J.-P. Banȼtre, and T. Priol, hal.inria.fr/hal-01087558/, 2014.

[39] http://www.samsung.com/us/mobile/cell-phones-accessories/GT-B9150ZKYXAR, 2015.

[40] A. Qureshi, R. Weber, H. Balakrishnan, J. Guttag, and B. Maggs, "Cutting the electric bill for internet-scale systems," *SIGCOMM Comput. Commun. Rev.*, vol. 39, no. 4, pp. 123–134, Aug. 2009.

[41] V. Mathew, R. Sitaraman, and P. Shenoy, "Reducing energy costs in internet-scale distributed systems using load shifting," in *Communication Systems and Networks (COMSNETS), 2014 Sixth International Conference on*, Jan 2014, pp. 1–8.

[42] L. Chiaraviglio and I. Matta, "Greencoop: Cooperative green routing with energy-efficient servers," in *Proceedings of the 1st International Conference on Energy-Efficient Computing and Networking*, ser. e-Energy '10, 2010, pp. 191–194.

[43] L. Chiaravigllio and I. Matta, "An energy-aware distributed approach for content and network management," in *Computer Communications Workshops (INFOCOM WKSHPS), 2011 IEEE Conference on*, April 2011, pp. 337–342.

[44] C. Ge, N. Wang, and Z. Sun, "Optimizing server power consumption in cross-domain content distribution infrastructures," in *Communications (ICC), 2012 IEEE International Conference on*, June 2012, pp. 2628–2633.

[45] V. Mathew, R. Sitaraman, and P. Shenoy, "Energy-aware load balancing in content delivery networks," in *INFOCOM, 2012 Proceedings IEEE*, March 2012, pp. 954–962.

[46] V. Mathew, R. K. Sitaraman, and P. Shenoy, "Energy-efficient content delivery networks using cluster shutdown," *Sustainable Computing: Informatics and Systems*, 2014.

[47] N. Xu, J. Yang, M. Needham, D. Boscovic, and F. Vakil, "Toward the green video cdn," in *Green Computing and Communications (GreenCom), 2010 IEEE/ACM Int'l Conference on Int'l Conference on Cyber, Physical and Social Computing (CPSCom)*, 2010, pp. 430–435.

[48] J. Llorca, A. Tulino, K. Guan, J. Esteban, M. Varvello, N. Choi, and D. Kilper, "Dynamic in-network caching for energy efficient content delivery," in *INFOCOM, 2013 Proceedings IEEE*, April 2013, pp. 245–249.

[49] J. Baliga, R. Ayre, K. Hinton, and R. Tucker, "Architectures for energy-efficient iptv networks," in *Optical Fiber Communication - incudes post deadline papers, 2009. OFC 2009. Conference on*, March 2009, pp. 1–3.

[50] K. Guan, G. Atkinson, D. Kilper, and E. Gulsen, "On the energy efficiency of content delivery architectures," in *Communications Workshops (ICC), 2011 IEEE International Conference on*, June 2011.

[51] U. Mandal, P. Chowdhury, C. Lange, A. Gladisch, and B. Mukherjee, "Energy-efficient networking for content distribution over telecom network infrastructure," *Optical Switching and Networking*, vol. 10, no. 4, pp. 393 – 405, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1573427713000428

[52] C. Ge, Z. Sun, and N. Wang, "A survey of power-saving techniques on data centers and content delivery networks," *Communications Surveys Tutorials, IEEE*, vol. 15, no. 3, pp. 1334–1354, Third 2013.

[53] P. X. Gao, A. R. Curtis, B. Wong, and S. Keshav, "It's not easy being green," *SIGCOMM Comput. Commun. Rev.*, vol. 42, no. 4, pp. 211–222, Aug. 2012.

[54] F. Vanclay, "Impact assessment and the triple bottom line: Competing pathways to sustainability?" *Sustainability and Social Science Round Table Proceedings*, 2004.

[55] The eu emissions trading system (eu ets). [Online]. Available: http://ec.europa.eu/clima/policies/ets/index_en.htm

[56] National carbon offset standard (ncos). [Online]. Available: http://www.environment.gov.au/climate-change/carbon-neutral/ncos

[57] V. C. Coroama and L. M. Hilty, "Assessing internet energy intensity: A review of methods and results," *Environmental Impact Assessment Review*, vol. 45, no. 0, pp. 63 – 68, 2014.

[58] M. F. S. R.-H. B. S. Clemens Cremer, Wolfgang Eichhammer and P. Zoche, "Energy consumption of information and communication technology in germany up to 2010," *Fraunhofer ISI and CEPE Project*, no. 28/01, pp. 63 – 68, 2003.

[59] S. Lambert, W. V. Heddeghem, W. Vereecken, B. Lannoo, D. Colle, and M. Pickavet, "Worldwide electricity consumption of communication networks," *Opt. Express*, vol. 20, no. 26, pp. B513–B524, Dec 2012.

[60] J. Baliga, R. Ayre, K. Hinton, W. V. Sorin, and R. Tucker, "Energy Consumption in Optical IP Networks," *IEEE/OSA Journal of Lightwave Technology*, vol. 27, no. 13, pp. 2391–2403, 2009.

[61] D. Kilper, G. Atkinson, S. Korotky, S. Goyal, P. Vetter, D. Suvakovic, and O. Blume, "Power trends in communication networks," *Selected Topics in Quantum Electronics, IEEE Journal of*, vol. 17, no. 2, pp. 275–284, March 2011.

[62] C. A. Chan, A. F. Gygax, E. Wong, C. A. Leckie, A. Nirmalathas, and D. C. Kilper, "Methodologies for assessing the use-phase power consumption and greenhouse gas emissions of telecommunications network services," *Environmental Science & Technology*, vol. 47, no. 1, pp. 485–492, 2013.

[63] C. Lange, D. Kosiankowski, R. Weidmann, and A. Gladisch, "Energy consumption of telecommunication networks and related improvement options," *Selected Topics in Quantum Electronics, IEEE Journal of*, vol. 17, no. 2, pp. 285–295, March 2011.

[64] D. Schien, P. Shabajee, M. Yearworth, and C. Preist, "Modeling and assessing variability in energy consumption during the use stage of online multimedia services," *Journal of Industrial Ecology*, vol. 17, no. 6, pp. 800–813, 2013.

[65] V. Coroama, D. Schien, C. Preist, and L. Hilty, "The energy intensity of the internet: Home and access networks," in *ICT Innovations for Sustainability*, ser. Advances in Intelligent Systems and Computing, L. M. Hilty and B. Aebischer, Eds. Springer International Publishing, 2015, vol. 310, pp. 137–155.

[66] D. Schien, V. Coroama, L. Hilty, and C. Preist, "The energy intensity of the internet: Edge and core networks," in *ICT Innovations for Sustainability*, ser. Advances in

Intelligent Systems and Computing, L. M. Hilty and B. Aebischer, Eds. Springer International Publishing, 2015, vol. 310, pp. 157–170.

[67] W. Van Heddeghem, F. Idzikowski, W. Vereecken, D. Colle, M. Pickavet, and P. Demeester, "Power consumption modeling in optical multilayer networks," *Photonic Network Communications*, vol. 24, no. 2, pp. 86–102, 2012.

[68] J. Baliga, R. Ayre, K. Hinton, and R. S. Tucker, "Energy Consumption in Wired and Wireless Access Networks," *IEEE Communications Magazine*, vol. 49, no. 6, pp. 70–77, Jun 2011.

[69] GreenTouch White Paper, "GreenTouch GreenMeter Study: Reducing the Net Energy Consumption in Communications Networks by up to 90% by 2020," 2013.

[70] C. A. Chan, A. F. Gygax, E. Wong, C. A. Leckie, A. Nirmalathas, and D. C. Kilper, "Methodologies for assessing the use-phase power consumption and greenhouse gas emissions of telecommunications network services," *Environmental Science & Technology*, vol. 47, no. 1, pp. 485–492, 2013.

[71] L. Niccolini, G. Iannaccone, S. Ratnasamy, J. Chandrashekar, and L. Rizzo, "Building a power-proportional software router," in *Proceedings of the 2012 USENIX Conference on Annual Technical Conference*, ser. USENIX ATC'12, 2012, pp. 8–8.

[72] L. Barroso and U. Holzle, "The case for energy-proportional computing," *Computer*, vol. 40, no. 12, pp. 33–37, Dec 2007.

[73] D. Kharitonov, "Time-domain approach to energy efficiency: High-performance network element design," in *GLOBECOM Workshops, 2009 IEEE*, Nov 2009, pp. 1–5.

[74] A. Vishwanath, J. Zhu, K. Hinton, R. Ayre, and R. Tucker, "Estimating the energy consumption for packet processing, storage and switching in optical-ip routers," 2013, p. OM3A.6.

[75] A. Vishwanath, K. Hinton, R. Ayre, and R. S. Tucker, "Modelling Energy Consumption in High-Capacity Routers and Switches," *IEEE Journal on Selected Areas in Communications*, vol. 32, no. 8, pp. 1524–1532, 2014.

[76] G. Auer et al., "Efficiency Analysis of the Reference Systems, Areas of Improvements and Target Breakdown, Deliverable 2.3, *Energy Aware Radio and neTwork tecHnologies (EARTH)*," https://bscw.ict-earth.eu/pub/bscw.cgi/d71252/EARTH$_$WP2$_$D2.3$_$v2.pdf, 2010.

[77] M. Brander, R. Tipper, C. Hutchison, and G. Davis, "Technical paper: Consequential and attributional approaches to lca: a guide to policy makers with specific reference to greenhouse gas lca of biofuels," 2008.

[78] Y. Gong, Z. Ying, and M. Lin, "A survey of cloud computing," in *Proceedings of the 2nd International Conference on Green Communications and Networks 2012 (GCN 2012)*, 2013, vol. 225, pp. 79–84.

[79] A. Nazir, S. Raza, D. Gupta, C.-N. Chuah, and B. Krishnamurthy, "Network level footprints of facebook applications," in *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement conference*, ser. IMC '09, 2009, pp. 63–75.

[80] Google green. [Online]. Available: www.google.com.au/green/bigpicture/

[81] Open compute project. [Online]. Available: http://www.opencompute.org/

[82] Facebook sustainability. [Online]. Available: http://newsroom.fb.com/sustainability.aspx

[83] D. Beaver, S. Kumar, H. C. Li, J. Sobel, P. Vajgel *et al.*, "Finding a needle in haystack: Facebooks photo storage," *Proc. 9th USENIX OSDI*, 2010.

[84] Facebooks carbon and energy impact 2012. [Online]. Available: www.facebook.com/green/app/$_$439663542812831

[85] A.-J. Su, D. R. Choffnes, A. Kuzmanovic, and F. E. Bustamante, "Drafting behind akamai: inferring network conditions based on cdn redirections," vol. 17, no. 6, pp. 1752–1765, 2009.

[86] F. Zhou, L. Zhang, E. Franco, A. Mislove, R. Revis, and R. Sundaram, "Web-cloud: Recruiting social network users to assist in content distribution," in *Network Computing and Applications (NCA), 2012 11th IEEE International Symposium on*, 2012, pp. 10–19.

[87] Wireshark - packet analyzer. [Online]. Available: www.wireshark.org/

[88] Sony vaio duo 11. [Online]. Available: www.sony.com.au/it-personal-computer/range/VAIO-Duo/561835/

[89] "Power-Mate power meter," www.power-mate.com.au.

[90] Facebook's key facts. [Online]. Available: http://newsroom.fb.com/KEY-FACTS

[91] L. Zhang, B. Tiwana, Z. Qian, Z. Wang, R. P. Dick, Z. M. Mao, and L. Yang, "Accurate online power estimation and automatic battery behavior based power model generation for smartphones," in *Proceedings of the Eighth IEEE/ACM/IFIP International Conference on Hardware/Software Codesign and System Synthesis*, 2010, pp. 105–114.

[92] Powertutor mobile application. [Online]. Available: https://play.google.com/store/apps/details?id=edu.umich.PowerTutor&hl=en

[93] Code of conduct on energy consumption of broadband equipment, version 4.1. [Online]. Available: www.telecom.pt/NR/rdonlyres/75F0D218-04AA-48EA-AA96-8AD6C457E97B/1465560/EnergyConsumptionofBroadbandEquipment.pdf

[94] D. Fritz. The evolving wireless world. alcatel lucent presentation. [Online]. Available: http://ceet.unimelb.edu.au/pdfs/aluteddocument.pdf

[95] Cisco's power consumption calculator. [Online]. Available: http://tools.cisco.com/cpc

[96] J. Baliga, R. Ayre, K. Hinton, and R. Tucker, "Green cloud computing: Balancing energy in processing, storage, and transport," *Proceedings of the IEEE*, vol. 99, no. 1, pp. 149–167, 2011.

[97] One billion fact sheet - facebook. [Online]. Available: http://newsroom.fb.com/PHOTOS-AND-B-ROLL/4227/ONE-BILLION-FACT-SHEET

[98] Cisco white paper- "Cisco Visual Networking Index: Global mobile data traffic forecast update, 2012-2017.

[99] Cisco white paper- "The zettabyte era, 2012-2017". [Online]. Available: www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/VNI\$_$Hyperconnectivity\$_$WP.pdf

[100] Cisco white paper, "Cisco Global Cloud Index: Forecast and Methodology, 2010-2015," 2011.

[101] N. Fernando, S. W. Loke, and W. Rahayu, "Mobile Cloud Computing: A Survey," *Elsevier Future Generation Computer Systems*, vol. 29, no. 1, pp. 84–106, Jan 2013.

[102] R. Buyya, J. Broberg, and A. M. Goscinski, "Introduction to Cloud Computing, Chapter 1," in *Cloud Computing: Principles and Paradigms*, John Wiley and Sons, New York, USA, Feb 2011.

[103] E. Cuervo, A. Balasubramanian, D. Cho, A. Wolman, S. Saroiu, R. Chandra, and P. Bahl, "MAUI: Making Smartphones Last Longer with Code Offload," in *ACM MobiSys*, USA, Jun 2010.

[104] C. Shi, K. Habak, P. Pandurangan, M. Ammar, M. Naik, and E. Zegura, "COSMOS: Computation Offloading as a Service for Mobile Devices," in *ACM MobiHoc*, USA, Aug 2014.

[105] K. Kumar, J. Liu, Y.-H. Lu, and B. Bhargava, "A Survey of Computation Offloading for Mobile Systems," *Springer Mobile Networks and Applications*, vol. 18, no. 1, pp. 129–140, 2013.

[106] F. Jalali, R. Ayre, A. Vishwanath, K. Hinton, T. Alpcan, and R. S. Tucker, "Energy Consumption Comparison of Nano and Centralized Data Centers," in *Proc. ACM Greenmetrics*, USA, Jun 2014.

[107] A. Odlyzko, "Data Networks are Lightly Utilized, and Will Stay That Way," *Review of Network Economics*, vol. 2, no. 3, pp. 210–237, 2003.

[108] "MSI Wind U100 Netbook," http://techreport.com/articles.x/15291.

[109] "Google Apps: Energy Efficiency in the Cloud," http://static.googleusercontent.com/external/content/untrusted/dlcp/www.google.com/en/green/pdf/google-apps.pdf, last visited Nov 2013.

[110] "Samsung Galaxy Tab 3 Lite 7" Tablet," http://www.samsung.com/au/consumer/mobile-phone/tablet/tablet/SM-T111MDWAXSA, last visited Sep 2014.

[111] "Robosoft Automatic Mouse and Keyboard Software," www.robot-soft.com/automatic-mouse-keyboard.html.

[112] "Google Docs Internals," http://googledocs.blogspot.com.au/2010/09/whats-different-about-new-google-docs$_$23.html, last visited Nov 2013.

[113] D. Kharitonov, "Time-Domain Approach to Energy Efficiency in High-Performance Network Element Design," in *Proc. IEEE Globecom Workshop on Green Communications*, USA, Dec 2009.

[114] "Cisco Catalyst 2960 Series Switches," http://www.cisco.com/c/en/us/products/switches/catalyst-2960-series-switches/index.html, last visited Nov 2013.

[115] "Enterprise 802.11n Wireless LAN Access Point Performance Benchmark," http://www.novarum.com/documents/Enterprise802.11nSingleAPBenchmarkTestingv1.3.pdf, last visited Nov 2013.

[116] G. Auer et al., "Efficiency Analysis of the Reference Systems, Areas of Improvements and Target Breakdown, Deliverable 2.3, Energy Aware Radio and neTwork tecHnologies (EARTH)," https://bscw.ict-earth.eu/pub/bscw.cgi/d71252/EARTH$_$WP2$_$D2.3$_$v2.pdf, 2010.

[117] D. Fritz, "The Evolving Wireless World – Alcatel Lucent Presentation."

[118] "Google's Green Computing: Efficiency at Scale," www.google.com/green/pdfs/google-green-computing.pdf, last visited Nov 2013.

[119] "Microsoft Cloud Infrastructure Strategy," https://www.samsung.com/us/business/oem-solutions/events/cio-forum2012/PDFs/Samsung-CIO-Forum-2012$_$DileepBhandarkar$_$Microsoft.pdf, last visited Nov 2013.

[120] "Google Docs Word Processing Service," https://docs.google.com.

[121] E. Upton and G. Halfacree, *Raspberry Pi User Guide.* John Wiley, 2012.

[122] Amn1220 optical line terminal (olt) - data sheet. [Online]. Available: http://hitachi-cta.com/pdf/access/amn1220_gmt_datasheet.pdf

[123] J. Hamilton. Overall data center costs. [Online]. Available: http://perspectives.mvdirona.com/2010/09/18/OverallDataCenterCosts.aspx

[124] J. Hamiltn. Perspectives data center cost and power. [Online]. Available: http://mvdirona.com/jrh/TalksAndPapers/PerspectivesDataCenterCostAndPower.xls

[125] "Cisco Global Cloud Index: Forecast and Methodology, 2012-2017," http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns1175/Cloud_Index_White_Paper.html.

[126] The power of wireless cloud: An analysis of the impact on energy consumption of the growing popularity of accessing cloud services via wireless devices. [Online]. Available: http://ceet.unimelb.edu.au/publications/downloads/ceet-white-paper-wireless-cloud.pdf