Leveraging Workload Relocation and Resource Pruning for Electricity Cost Minimization in Service Provider Networks

PhD Thesis

Muhammad Saqib Ilyas

2005-06-0024

Advisor: Dr. Zartash Afzal Uzmi

Department of Computer Science

School of Science and Engineering

Lahore University of Management Sciences

Dedicated to dedication

Lahore University of Management Sciences

School of Science and Engineering

CERTIFICATE

I hereby recommend that the thesis prepared under my supervision by ***Muhammad Saqib Ilyas*** titled ***Leveraging Workload Relocation and Resource Pruning for Electricity Cost Minimization in Service Provider Networks*** be accepted in partial fulfillment of the requirements for the degree of doctor of philosophy in computer science.

Dr. Zartash Afzal Uzmi (Advisor)

<u>**Recommendation of Examiners' Committee:**</u>

Name **Signature**

Dr. Zartash Afzal Uzmi ______________

Dr. X ______________

Dr. Y ______________

Dr. Z ______________

Acknowledgements

Contents

List of Figures

List of Tables

Abstract

Abstraction

Chapter 1

Introduction

1.1 Networks and systems pervade

Different types of networks play a critical role in our every day lives. We use Public Switched Telephone Networks (PSTN) and cellular networks to communicate with people by making voice calls (and sending text messages in case of cellular networks). PSTN and cellular networks also server as access media for connecting to the Internet, which offers several key services. We communicate and collaborate using email, voice/video calls over Internet Protocol (IP) and social networks. We also use the Internet to access teaching/learning material, course registration systems on campus and even pathological examination reports.

The Internet itself is an interconnection of several different types of networks. First, there are the packet-switched networks operated by Internet Service Providers (ISPs) that provide us access to Internet resources worldwide by carry information between hosts on the Internet. A second type of networks that the Internet is comprised of are the geo-diverse data centers operated by companies like Facebook, Amazon, Microsoft and Google. Servers in these data center networks run applications like Google Search, Gmail, Youtube, Twitter, Bing and Facebook. A third type of network which are also part of the Internet are the

Content Distribution Network (CDN), that place mutlple copies of Internet resources such as web pages across the globe. The role of CDNs is to keep the latency from a user to an Internet resource small (compared to having the resource located at a fixed single location). For instance, if Google's home page were only located at a server in San Jose, CA, the latency (the time it takes for a web browser to send a packet to the server) for users in Pakistan would be hundreds of milliseconds. Placing a replica of the Google home page close to Pakistan lowers the packet latency significantly, thereby allowing the web browser to display the page much faster.

The deployment of these different types of networks involves huge expenses. For instance, Google announced building a data center in Iowa at a cost of $400 Million [1]. Furthermore, according to [2], the capital cost of a typical cellular network site is $550,000[1].

The recurring operational cost of these networks is also quite high. For instance, in 2009, Facebook spent $50 Million on leasing the data center space, alone [3]. In the context of geo-diverse data centers, other contributors to operational expenses include staff salaries, maintenance related costs, the cost of inter-data center network connectivity and electricity bill. Similar trends may also be observed in other types of networks. Optimizing operational costs is critical for network operators in order to offer cost-effective services to consumers and maximize their profit.

1.2 Electricity costs in networks and systems

For many tyeps of networks, electricity costs contribute a significant fraction of operational costs. For instance, electricity costs may be as much as 15% of operational costs in data centers [4]. Similarly, for an operator with 7000 cellular sites in a country as small as Pak-

[1] This does not include spectrum licensing costs. Furthermore, an operator needs to deploy many sites. A site at about every 800 meters is common in urban settings

istan, the annual electricity cost can be roughly estimated at \$9.19 Million[2]. Telecom Italia reported a consumption of 1.793 GWh in 2012 [6], which is significantly higher compared to our estimates for the Pakistani network operator and hence the electricity costs are also expected to be much higher.

1.3 Energy inefficiency characterizes today's networks

For most networks, the power consumption is well-approximated as a linear function of workload [7, 8]. Furthermore, these networks are not energy proportional. In Figure 1.1, the green line shows the ideal energy proportional behaviour where the network consumes no power when there is no workload. No real network exhibits this ideal behaviour for one of the following reasons.

1. The network activity under no workload conditions is not significantly less than that under peak workload. For instance, a cellular network's radio components must continue operating and drawing power to offer uninterrupted connectivity to prospective allers, even when no call is in progress. In packet switching networks, many data link layer technologies continuously transmit frames irrespective of traffic activity.

2. The components of the network may not be energy proportional. For instance, in data centers, server power consumption is a large fraction of the total power consumption and the server idle power consumption is a large fraction of their peak power consumption.

A network that is not energy proportional is energy inefficient (i.e., consumes a lot more energy than it should) in the low-workload regimes. It has been observed for many networks

[2] Using a 1.5 kW draw for a single cellular site [5], Rs. 10 per kWh and Rs. 100 per US\$. Note that the Rs. 10 per kWh is a gross under-estimation, given that it is the approximate current price of grid power, which is note reliable. In the absence of grid power, diesel generators power a cellular site and the resulting cost per kWh is much higher.

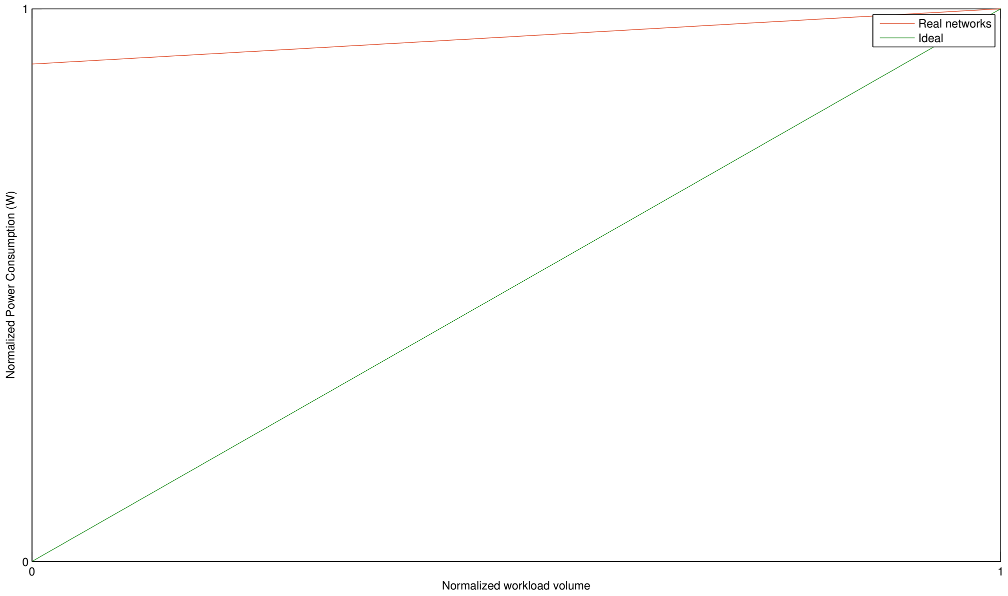


Figure 1.1: Networks lack energy proportionality

that workload is variable and periodic. Figure 1.2 shows the workload for call traffic at a cellular site in a large operational GSM network in Pakistan. It shows that call traffic has diurnal cycles and that traffic peaks for only a short period of time during a day. Furthermore, the workload peak is quite high compared to the trough. ISP [9] and data center [10] traffic also show similar trends. In order to meet peak expected workload amicably, networks are dimensioned according to the peak workload. Since the workload is far from the peak most of the time and networks are not energy proportional, most networks are energy inefficient. Recent years have witnessed significant research effort aimed at improving network energy efficiency in packet networks, cellular networks and geo-diverse data centers. Effectively, such research aims to lower the y-intercept of the red line in Figure 1.1.

We have observed earlier that network electricity costs are quite high. An energy efficient network would only incur high electricity costs if it handled a lot of workload. On the other hand, for energy inefficient networks, such as those prevalent today, high electricity costs

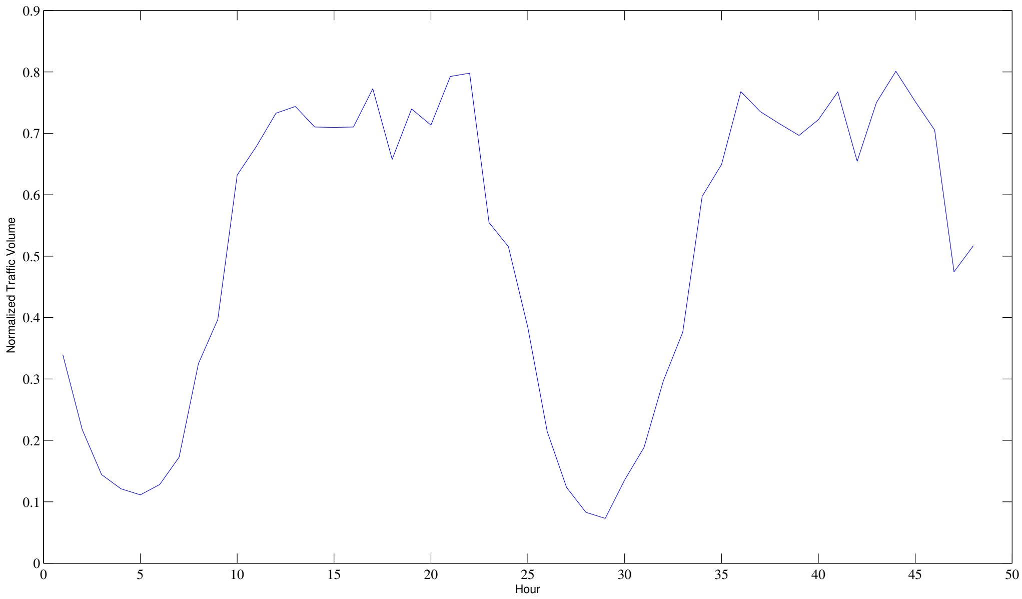


Figure 1.2: Call traffic for an oeprational cellular site over two days

are not justifiable by high workload because the workload is quite variable. In other words, today's networks have very poor performance per Watt characteristics. Therefore, reducing the electricity costs in today's networks is critically important.

1.4 Prevalent electricity cost reduction techniques

The electricity cost for a network during a unit duration of time is given by:

$$\text{Electricity cost} = \text{amount of energy consumed} \times \text{unit price of electricity} \tag{1.1}$$

Conseuqently, the electricity cost for a network may be reduced by minimizing either or both of the terms on the right handside of the above equation. From prior research work and current operational practices in different types of networks, we observe the following techniques to reduce electricity cost in networks by reducing one or both of the two quantities in equation 1.1.

1.4.1 Reducing the amount of energy consumed

1. **Hardware upgrades:** Due to ecological challenges, improved energy efficiency is generally a key requirement when developing new technologies and devices. For a given workload demand, an improvement in device energy efficiency lowers the amount of energy consumed, thereby reducing electricity cost. Therefore, hardware upgrades are a way to reduce electricity costs. An operator would, however, opt for hardware upgrades in their network only after they have obtained the Return on Ivestment (ROI) of the initial deployment. The initial investment not only involves capital cost of equipment but other factors such as spectrum licensing as well. In the cut-throat competition prevalent in most of today's networks, the ROI is slow to achieve. This means that existing energy efficient networks would stay that way for a considerable tiem into the future.

2. **Hardware virtualization:** With the advent of ever faster CPUs, it was observed that servers tend to operate at relatively low CPU utilization most of the time. This was seen as an opportunity to statistically multiplex multiple servers onto a single physical machine by slicing the latter into multiple virtual servers. In this way, virtualization cuts capital costs for procurement of hardware. Since the virtual servers share the same resources (power supply, CPU, network interface, disks), if two servers are multiplexed onto a single physical server, the electricity consumption may be cut by as much as 50%. A more aggressive server consolidation may cut electricity costs by upto 80% [11].

3. **Resource Pruning (RP):**Since network resources must be deployed according to peak demand while the workload peaks only for a short period of time, the excess resource may be deactivated (shutdown or put in power-saving mode depending on

what is supported by the equipment) when workload is low [12, 13, 14, 15, 7]. When evaluating the reduction in electricity costs through resource pruning, it is imperative to consider any costs associated with activation and deactivation of network resources.

1.4.2 Using cheaper electricity - Workload Relocation (WR)

Electricity prices exhibit geographic diversity [16], i.e., the price of electricity varies from one location to another. The variation in electricity price is generally noticeable only at large distances. For instance, the electricity price anywhere within a city is generally the same[3]. Most networks span large enough distances for geographic diversity in electricity prices to be apparent. If the network workload is quite flexible in terms of where it is handled, then the workload originating at a location with high electricity price may be relocated to a different location that has lower electricity price, thereby cutting electricity cost. We call this technique Workload Relocation (WR). We observe that different networks have different levels of geo-flexibility in workload. In geo-diverse data centers, for instance, the workload is highly geo-flexible, i.e., a client's request may be handled close by or even hundreds of miles away. On the other hand, the workload in cellular networks has very low geo-flexibility, i.e., a call mus tbe handled at a cellular base station within a few hundred meters from the caller.

Electricity prices also exhibit temporal diversity [16], i.e., the relative order of electricity prices at different locations keeps changing. If a city in Kansas presently has chepaer electricity than one in Oklahoma, an hour later, the reverse may be true. This means that mapping of workload to locations must be periodically updated. The granularity of these updates depends on how frequently electricity prices change. Electricity markets exhibit price changes at two different time scales (15 minutes for real-time electricity prices and an hour for day-ahead prices).

[3] With the exception of factors such as different tarriffs for domestic, commercial and industrial consumers

1.5 Our thesis

Based on the similarity in workload characteristics and the dependence of power consumption on workload, we opine that a generalized power optimization framework may be formulated that is applicable to many different types of networks. Our generalized electricity cost optimization framework would use workload relocation and resource pruning in tandem to reduce electricity costs[4].

1.6 Contributions

This thesis makes the following contributions:

- We present a generalized model for electricity cost optimization applicable to different types of networks that jointly uses workload relocation and resource pruning. We show that this problem is NP-Hard.
- We present a framework called Relocate Energy Demand to Better Locations (RED-BL), pronounced Red Bull, that solves this problem. We apply RED-BL to geo-diverse data centers as well as cellular networks using real data traces.
- We exactly solve some reasonably-sized instances of this problem using real data. We also propose some heuristics that would be useful for larger instances of the problem.
- We evaluate RED-BL on two different types of networks, namely, geo-diverse data centers and cellular networks.
- Prior efforts in this area had mostly ignored the costs associated with activation and deactivation of network resources. To the best of our knowledge, we are the first to incorporate these in our optimization framework.

[4] Hardware virtualization is complimentary to our framework

- We evaluate the benefits of geographical diversity exhibited by electricity prices and network deployments.

- A network with significant overprovisioning may handle most of the workload at cheaper locations while the more expensive ones may be pruned from the network. In other words, geographic diversity in electricity prices incentivises over-provisioning. We study the benefits of increased over-provisioning and find diminishing returns when increasing over-provisioning.

1.7 Organization

The rest of the document is structured as follows. In Chapter 2, we compare two different types of networks and describe how similar they are in terms of workload handling and power consumption. In chapter 3, we derive a generalized power consumption model, applicable to different types of networks and formulate RED-BL, a generalized electricity cost optimization problem. We present an evaluation of RED-BL on geo-diverse data centers and cellular networks in chapters 4 and 5, respectively. In chapter 6, we draw the conslusions about our thesis and provide some future directions.

Chapter 2

Background - Different Types of Networks and Their Similarities

In this thesis, we claim that many different types of networks are quite similar in terms of power consumption. In this chapter, we take an essentials-only look at two different types of networks with a view to establishing the similarity between them. This similarity motivates the formulation of a generalized electricity cost optimization framework.

2.1 Geo-Diverse Data Centers

Organizations like Microsoft, Facebook, Amazon and Google run a plethora of applications. Some of these applications are accessible by the general public. Google Docs is one such application. Such organizations also run private applications for the consumption of authorized internal users only. These applications run on servers that are hosted at sites called data centers.

A data center is a site that has equipment such as servers, storage and networking equipment, in addition to some allied equipment such as airconditioning and power supplies.

A given data center may host only public applications, only private applications or even both. Furthermore, some public data center operators allow a client to host their own applications, whereas some only offer a fixed set of internally developed applications. For instance, one may run a custom application on a server leased on Amazon's data centers, but on the other hand, Google's search cluster only hosts the Google search application.

Operators typically deploy multiple data centers at different geographic locations. This is done for two reasons. First, having data centers at different locations provides fault tolerance. If one site goes down for some reason, the other site may take over as a backup. Also, multiple remote sites are less likely to affected simultaneously by a natural disaster. A second reason to have multiple data centers is to have low latency to clients at different locations. For instance, Amazon has multiple data centers in different continents, thereby ascertaining that no matter where a client may be, there is an Amazon data center relatively close by compared to the case if Amazon only had one data center in the US. Figure 2.1 shows the locations of Google's data centers across the globe (according to royal.pingdom.com as of April 2008).

2.1.1 Structure

Before delving into the internal structure and composition of a data center, let us consider a data center as a single resource. This view helps provide only the high-level details of an operator's network. At this level, each one of an operator's data centers are inter-connected by means of high-speed inter data center network links. These links serve to carry various types of traffic, some of which are given below:

- **Consistency traffic:** To maintain consistency amongst replicas of an application's servers hosted in different data centers, some overhead in terms of network traffic must be incurred. For instance, a customer's website may be hosted at two different data

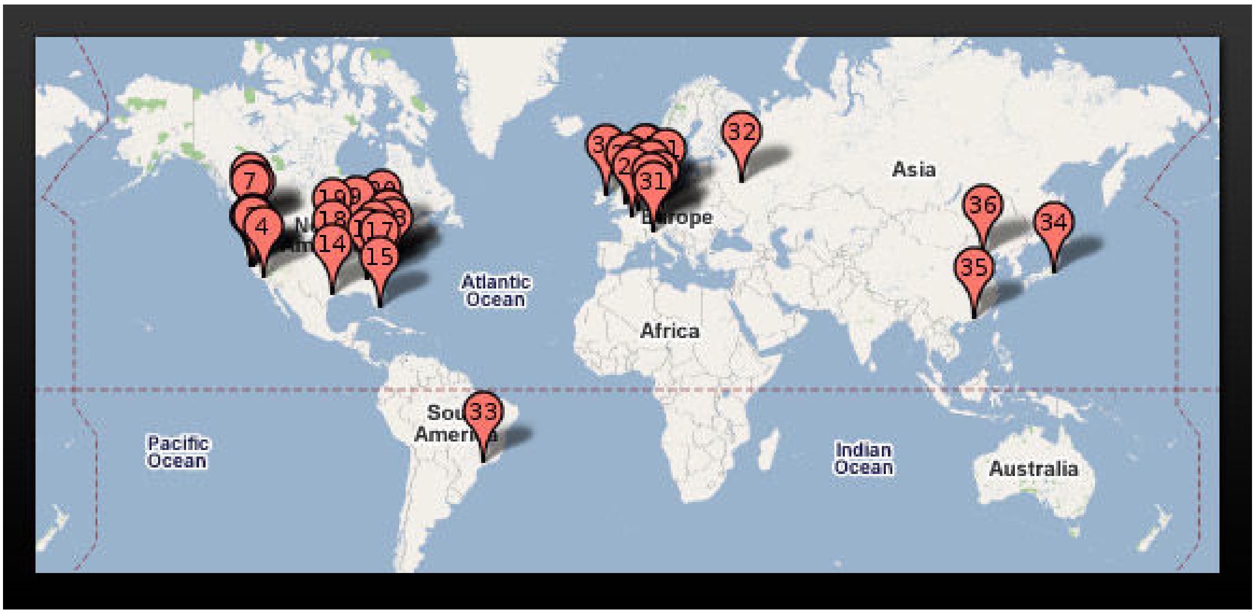


Figure 2.1: Google Data Center Locations - Source: royal.pingdom.com

centers and whenever a change is made to one copy of the website, the same changes must be reflected at the replica as well.

- **Traffic due to load-balancing:** Some traffic on the inter-data center links may be a result of the effort to achieve load-balancing amongst the data centers. For instance, the data centers may be oeprated by a web-based email service provider and the user inboxes may be partitioned over the data centers. In this case, an operator might desire that a roughly balanced amount of storage be used at each of the data centers. To this end, the operator might want to spread the inboxes over the data centers such that the cumulative size of the inboxes at each data center is roughly the same. Over time, due to changes in user behaviour and activity, the operator would need to re-adjust the inboxes assigned to each data center, thus requiring migration of inboxes between data centers.

- **Background traffic:** Yet another source of inter data center traffic is background

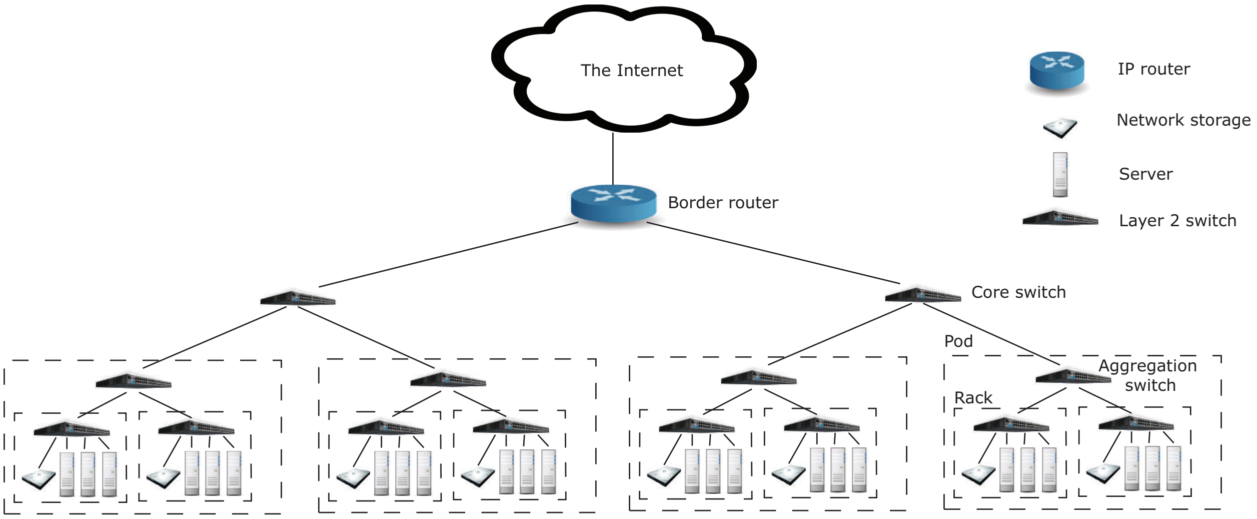


Figure 2.2: The modern data center's architecture

traffic. For isntance, different data centers belonging to an Internet search engine operator may collaboratively compute search results. In this example, the search indices may also be updated periodically in the background.

Having taken a high-level view of a geo-diverse data center operator's network, now let us delve into the internal structure and composition of a data center. Today's data center architecture is heirarchical [17] as shown in Figure 2.2. A typical data center hosts tens of thousands of servers [18]. The servers are installed in vertical racks. Apart from servers, the racks host other equipment as well. In addition to built-in hard drives in the servers, some dedicated storage nodes are also installed in the racks. A high speed Ethernet switch provides interconnection between the devices installed in the rack and connectivity to the rest of the data center and beyond. Power supply and distribution units for the equipment are also installed in the rack.

A group of racks, called a pod (or a cluster), are interconnected by means of aggregattion switches. An aggregation switch allows servers in different racks to communicate with each other. All the pods within a data center are interconnected by core switches. This allows servers in different pods to communicate with each other. The core switches are interconnected through one or more border routers. These border routers are the avenues for traffic coming in and going out of the data center.

All of the equipment is quite tightly packed within a pretty small space in a data center. The equipment generates a lot of heat and to prevent thermal damage to it, cooling must be provided. This is generally done by air-cooling, i.e., heat is transported away from the equipment by circulating cool air around it.

2.1.2 Request routing

As noted in chapter 1, electricity cost depends not only on how much workload is handled, but also where it is handled. In order to develop a model for electricity cost in a geo-diverse data center, we need to first understand how workload from all over the globe is distributed amongst the data centers. In this section, we will use as an example a client request for viewing a web page hosted by a geo-diverse data center operator.

To access a web resource, the user types a uniform resource locator (URL) in the web browser's address bar. The URL typically contains the DNS name corresponding to the web server that hosts the requested resource[1]. Since a single server would hardly be sufficient to handle all traffic for a typical web site, several servers must be mapped to the same DNS name. However, the web browser must connect to exactly one of these servers during a browsing session. Figure 2.3 briefly describes how this web server's IP address is picked. For details on DNS resolution process, see [19, 20].

When the user enters a URL in her web browser, the browser invokes the local Domain Name System (DNS) resolver on the client machine which attempts to determine the IP address corresponding to the DNS name of the remote host specified in the URL. The local DNS server communicates with the DNS server for the client's ISP[2]. The DNS query

[1] It is also possible to specify the IP address of the web server directly in the URL. However, remembering IP addresses for all web sites of interest is not humanly possible

[2] Some people configure other DNS servers, such as Google's Open DNS Servers on their machines. In

eventually reaches the authoritative server for the remote host's domain. In our example, this would be operated by the data center operator. The DNS server for the data center operator resolves the DNS name by returning an IP address corresponding to the DNS name specified by the client. The data center operator's DNS server performs an attempt at load-balancing so that roughly the same amount of workload is sent to each server hosting the requested web site. Notice in Figure 2.3 that caches are available at various DNS resolvers in order to improve the latency of DNS resolution. These caches will keep the IP address corresponding to recently queried DNS names until the timeout specified by the authoritative DNS server expires.

The data center operator has a large pool of IP addresses, also known as IP address space, for their layer 3 devices. This IP address space is segmented over the geo-distributed data centers. The IP address resolved by the operator's DNS server belongs to one of the data centers and the client must now send it's Hyper Text Transfer Prototcol (HTTP) [21] request to the appropriate server at the corresponding data center. The client's web browser now establishes a Transport Control Protocol (TCP) connection with the server. To this end, the client sends packets to the web server's IP address that was just resolved. The packets leave the client's network interface and go to the ISP's gateway. Once in the ISP's network, the routers determine a path to the destination IP address and forward the packets hop by hop until the packets reach the data center where the required web server is hosted.

Having determined the IP address of the web server, the client's web browser establishes an HTTP session with that IP address over a TCP connection. The IP packets belonging to this connection destined to the web server arrive at the border router in the corresponding data center and are forwarded to the server, traversing the core, aggregation and top of rack switches. The response packets are forwarded from the web server to the border router which routes it back to the client machine.

such cases, the local DNS server would communicate with the Google Open DNS Server

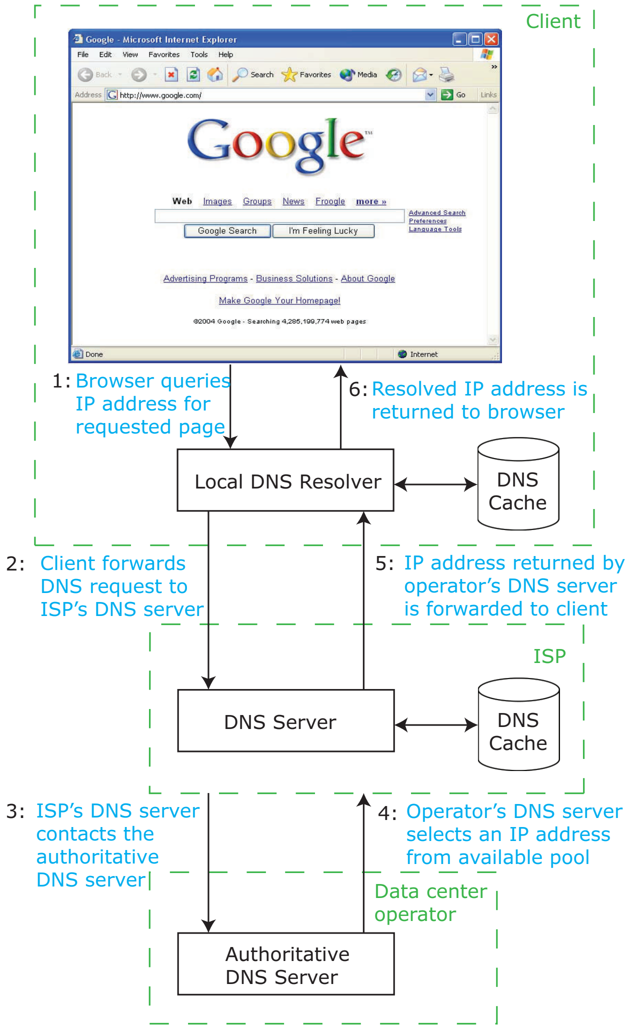

Client
Google - Microsoft Internet Explorer
http://www.google.com/
Google
Web Images Groups News Froogle more »
Google Search
I'm Feeling Lucky
Advanced Search
Preferences
Language Tools
Advertising Programs - Business Solutions - About Google
Make Google Your Homepage!
©2004 Google - Searching 4,285,199,774 web pages
1: Browser queries IP address for requested page
6: Resolved IP address is returned to browser
Local DNS Resolver
DNS Cache
2: Client forwards DNS request to ISP's DNS server
5: IP address returned by operator's DNS server is forwarded to client
ISP
DNS Server
DNS Cache
3: ISP's DNS server contacts the authoritative DNS server
4: Operator's DNS server selects an IP address from available pool
Data center operator
Authoritative DNS Server

Figure 2.3: Resolving the IP address for a server hosted in a data center

2.1.3 Power consumption model

Fan et. al. used the results of a measurement study to show that the power consumption in a data center can be well-modeled as a linear function of the average CPU utilization [8]. As more and more client traffic arrives at servers in a data center, the average CPU utilization increases. If we consider homogenous client requests, the CPU utilization can be modeled as a linear function of workload. In case of heterogenous requests, one can approximate all request types as consisting of an integer number of micro-requests. Using the micro-request as our workload unit, we can still model CPU utilization as a linear function of workload. Since CPU utilization can be modeled as a linear function of workload, server power consumption can be modeled as a linear function of it's workload. The total power consumed by servers in a data center can, therefore, be represented as a linear function of the cumulative workload handled by the servers at the data center.

In our thesis, we wish to minimize the total electricity cost in a data center and servers are not the only power consuming equipment in a data center. Nonetheless, server power consumption is related to total data center power consumption by a measure called Power usage effectiveness (PUE). PUE is a measure of the efficiency with which a data center handles its power. It is defined as the ratio of total facility power to the IT equipment power. IT equipment power consumption includes power consumption by servers, storage and networking equipment. We assume that the power consumption in storage is related to that in servers, i.e., a unit workload consumes a fixed amount of power in storage devices. Networking equipment's power consumption is almost invariable with workload [22, 23, 24]. Therefore, we can consider total IT power as being a constant multiple of server power consumption plus a constant amount (which is not important in our thesis since it plays no role in an electricity cost minimization problem). Therefore, PUE being a constant (depending on how efficiently data center is architected), data center power consumption is proportional to workload handled by the data center.

2.2 Cellular Networks

Being the older sibling of the Internet, telephony services are a more integral part of our every day lives than the Internet. Mobile telephone systems have enabled not only untethered access to traditional telephony services but also new types of services. We make phone calls, send text messages and can even connect to the Internet using our mobile phones. Just as Internet connectivity services are provided by ISPs and Internet applications are powered by data center operators, mobile phone services are provided by mobile network operators (MNOs).

Over the years, mobile networks have been deployed based on different technologies. Literature often categorizes mobile network technologies in terms of *generations*. First generation cellular networks (1G) were based on Advanced Mobile Phone System (AMPS). AMPS networks were deployed starting in 1978. The AMPS system also evolved into Digital-AMPS (D-AMPS) networks. Two technologies were part of the second generation (2G) cellular networks, namely Global System for Mobile communication (GSM) and Code Division Multiple Access (CDMA). Today, 90% of the world's top 20 cellular networks use GSM technology [25]. Anticipating the increased demand for mobile access to data services such as Internet access, vendors introduced General Packet Radio Service (GPRS) as an add-on to GSM networks. GPRS offers data rates between 56 kbps and 114 kbps. 2G networks with GPRS are sometiems referred to as 2.5G. GPRS bit rates are insufficient for many high bandwidth applications such as video calls, video streaming and video conferencing. To enable such services, broadband mobile services were introduced in third generation (3G) networks networks such as High Speed Downlink Packet Access (HSDPA) and Universal Mobile Telephone System (UMTS). The increasing trends in the use of high-bandwidth applications in mobile networks has spawned the fourth generation (4G) cellular networks such as WiMAX and Long Term Evolution (LTE).

2.2.1 Structure

In this thesis, as far as cellular networks are concerned, we focus specifically on GSM networks. Mobile phone networks are also referred to as cellular networks. The term cellular network stems from the fact that the area covered by the operator is logically divided into several small areas called cells. A cell in an urban setting (a macrocell) is typically upto a few hundred meters in radius, whereas in suburban or rural settings, the cell radius may be upto tens of kilometers. A *cell site*, typically situated in the middle of a cell, enables subscribers in that cell to connect to the mobile network. A cell site is also often referred to as a Base Transceiver Station (BTS)[3] or simply a base station. A cell site hosts a number of transceivers (TRXs), radio antennas, power amplifiers and other allied equipment.

Typically a government regulator such as Pakistan Telecommunication Authority (PTA) allocates a frequency band to each of the operators providing cellular service in the host country. The allocation is such that each operator gets a different frequency band. The spectrum allocated to a cellular operator is an integer multiple of the bandwidth of a single GSM channel (200 kHz). A cellular operator distributes their allocated frequencies to cells in their network. The channels allocated to an operator are much fewer than the number of cells in the network. Therefore, a given channel must be reused in an operator's network. Frequency reuse is done in such a way that any two cells that share the same frequency channel are sufficiently far apart so that the radio signal from any one of the cells does not noticably intefere with that in the other. In fact, each cell is typically divided into three sectors (resembling 120 degree pie-slices), therefore, the frequency allocation is done on a per-sector basis. Nonetheless, for a high-level view, the set of frequencies allotted to all sectors in a cell can be considered as allotted to the cell itself. Each TRX at a cell site operates at a distinct frequency.

[3] A single cell site sometimes hosts multiple BTSs, for instance, when multiple network operators share a single site

Given two communicating parties at fixed locations, if the transmitted signal power is kept constant, the received radio signal strength would differ depending on the frequency used. Also, this frequency selective behaviour of the radio communication medium keeps changing with time, i.e., if frequency A receives better propagation compared to frequency B at time t_1, the same will not necessarily be true at time $t_1 + \epsilon$. This means that we can't statically pick the best frequencies to use for a particular cell by considering, for instance, the type of terrain. In order to make decent communication conditions available to all callers, on average, GSM networks also use frequency hopping, whereby the frequency allocation to cells are changed periodically.

To improve GSM's spectrum utilization, each frequency is also time-divided. For each frequency, a 120 ms duration transmission unit is called a GSM multiframe. It is named so because it consists of 26 frames of duration approximately 4.6ms each. Each frame is also divided into 8 bursts of duration approximately 0.577 ms each. The recurrence of a particular burst is what may be called a channel in GSM. In other words, a particular frequency and position within every frame defines a channel.

A MS often receives radio signals from multiple BTSs nearby. The MS picks the BTS from which it receives the strongest signal as it's serving BTS. A MS will do all communication such as call reception and placement through the serving BTS. When a subcriber moves around, the signal from the serving BTS might weaken. In such an event, the MS requests the network to allow it to change it's serving BTS to the one from which it currently receives the strongest signal. This is called a call handoff and is coordinated by a Base Station Controller (BDC).

Whereas a BTS is the access-side of a cellular, having no intelligence and performing only radio transmission and reception, the operations requiring intelligence such as frequency allocation, handoff coordination and frequency-hopping are controlled by a BSC. A BSC is responsible for several BTSs, which are connected to the BSC by means of some backhaul, such as E-1 or microwave links. A cellular network will typically have multiple BSCs, with a BSC being responsible for several BTSs in a vicinity. All BSCs are also interconnected by the cellular network's backbone, so that actions that require global coordination such as frequency assignment, frequency hopping and call handoff can be done smoothly.

Another key component of a GSM network is the Mobile Switching Center (MSC). The MSC is responsible for call routing both within the GSM network and beyond (to a landline phone, for instance). Since the focus of our thesis is power consumption in the network and 50% [26] - 80% [27] of a cellular netowrk's electricity consumption is due to the BTSs, we will not dwell on the MSC and other components of the cellular network any more than necessary.

2.2.2 Call placement

For intelligible wireless communication, only one transmitter may transmit on a given frequency. Therefore, a call to/from a MS requires the allocation of two GSM frequencies, one for uplink (voice traffic from the MS to BTS) and the other for downllink (from BTS to MS). For coordinated acquisition of these frequencies, certain frequencies are reserved in each cell to serve as control channels. In fact, a caller does not get complete access to a particular pair of frequencies. Each of the 8 bursts in a GSM frame for a particular frequency may be used by different callers. Therefore, a particular frequency may be shared between multiple callers at a given time. In GSM terminology, they would all be using different channels, however, because a channel is characterized not only by the frequency but also the position within a GSM frame. Hence, the voice traffic for a call in GSM operates over two channels.

It appears that if n frequencies are assigned to a particular sector, then it can support up to $8n$ simultaneous calls because that is the number of GSM channels available. However, this is not true for two reasons.

- Some channels are reserved for control purposes. The exact number of such channels varies from operator to operator.

- GSM supports two different types of codecs, namely the full-rate codec and the half-rate codec. The full-rate codec corresponds to a caller using a burst in every GSM frame during a call, whereas the half-rate codec corresponds to a caller using a burst in every alternate GSM frame. By default, the full-rate codec is used for every call. However, when traffic congestion rises above an operator-configured threshold, the network attempts to admit every new call using a half-rate codec, if the corresponding MS supports it. If the traffic rises further and crosses a second threshold as configured by the operator, the network also re-assigns current calls to use a half-rate codec depending on the corresponding MS support. This enables a BTS to support more than $8n$ simultaneous calls during times of congestion.

2.2.3 Power consumption model

BTSs account for most of the power consumed in a cellular network. [26] claims that BTSs contribute 50% of overall network power consumption, whereas [27] puts this number at 80%. For this reason, most of the prior work related to power consumption in cellular networks focuses on BTSs.

Lorincz et. al. performed a measurement study of BTS power consumption under real-traffic conditions and concluded that the power consumption may be approximated as a linear function of call traffic [28]. Thus, as traffic varies during a given day, instantaneous power consumption would follow a similar curve as the traffic.

2.3 Similarities between different types of networks

From our discussion of two different types of networks, we can see that they are both essentially a collection of interconnected sites (data centers and BTSs) which are a collection of resources (data centers and TRXs). The workload in both types of networks exhibits diurnal patterns [10, 7]. The network in both cases is provisioned according to peak workload demand. Since the network resources are not energy proportional, this means that in low-workload regimes, the network is heavily over-provisioned. The resulting energy inefficiency can be dealt with by deactivating some resources when the traffic is low.

In terms of power consumption also, the two networks considered in this chapter, namely geo-diverse data centers and cellular network are quite similar. Both have a linear mapping from workload to power consumption. This motivates the possibility of establishing a common framework that smartly schedules the network resources in response to workload variations so as to minimize electricity costs.

2.4 Differences between different types of networks

All characteristics of different network types are not essentially similar. Several attributes are different as well. The following list summarizes some differences between geo-diverse data centers and cellular networks

- **Workload granularity:** The workload capacity of a data center is very large, potentially in millions of client requests per second, whereas the workload capacity of a TRX is less than eight simultaneous calls. For this reason, instead of determining the exact integer number of requests to handle at each data cneter, the fraction of workload to be handled at each data center can be determined as a real-number (which is a much simpler problem to solve), and the resulting number of requests will most likely be an integer or may be rounded off with little change in electricity cost. Meanwhile, in case of cellular networks, the cumulative workload for a cell is a small number of calls and each call must be handled at exactly one of a few candidate cells. Hence, the call to cell mapping must be binary in nature and fractional mapping algorithm will not work.

- **Geo-diversity in electricity prices:** In a geo-diverse data center scenario, the network resources, i.e., data centers are quite far from each other and hence electricity price differential due to geo-diversity in electricity prices is quite likely. However, in case of cellular networks, the cell sites are within a few hundred meters of each other (in an urban setting) and an electricity price differential is highly improbable.

Chapter 3

A generalized framework for electricity cost optimization

In Chapter 1, we have seen that networks, such as geo-diverse data centers and cellular networks, are energy inefficient. This is due to lack of energy proportionality coupled with significant variations in workload. We have also observed that these networks are plagued by high electricity costs. For a given electricity price, if these networks were energy efficient, high electricity costs would only be due to heavy workload. However, energy inefficient networks consume a lot more energy than they ideally should. Hence, high network electricity costs are a source of concern.

In Chapter 2, we developed insights into how these networks operate in order to get a better understanding about their power consumption model. We also saw the similarities and differences between different network types, related to power consumption, that must be kept in mind. in this chapter, we will use this knowledge to formulate an optimization problem that minimizes electricity costs for different types of networks.

3.1 Modeling the electricity cost minimization problem in networks and systems

Let us illustrate network operation from the standpoint of electricity consumption and cost using an example shown in Figure 3.1. The example uses a test tube to represent a network resource and marbles to represent a unit workload. The network resource could be a data center in the context of geo-diverse data center operator, whereas it could be a transceiver in the case of a cellular operator. Similarly, the workload unit could be a client request in the data center context, whereas it could be a call in a cellular network setting. The operator's goal is to assign workload to network resources and, if needed, periodically update this mapping in response to variations in workload.

We consider the largest possible quantum of time for which the workload (and electricity price) remains fixed and term each quantum as an *interval*. We assume that workload for several consecutive intervals is known and term this sequence of intervals as a planning window. The example demonstrates three different ways of mapping this workload to two network resources situated at different locations. For simplicity we assume in this example that the workload is geographically split such that half of it originates near each of the two resource. Figure 3.1 (a) shows the operator workload over a planning window consisting of three consecutive intervals. Meanwhile, Figure 3.1 (b) shows the geo-temporal variation in electricity prices for the two network resources.

One possible operational strategy is to map each workload unit to the nearest available resource as shown in Figure 3.1 (c). In a sense, this is the default strategy in cellular networks, whereby a call is handled by the BTS from which the mobile station (MS) receives the strongest radio signal[1]. In geo-diverse data center settings, this sort of mapping is also

[1] Signal from the physically nearest BTS may be weakened considerably due to natural or man-made obstructions. In such cases, the nearest BTS may not be the one from which the strongest signal is received. Hence, we take ”nearest” to mean the BTS from which the MS receives the strongest signal

often the default strategy because it minimizes the access latency for all clients[2].

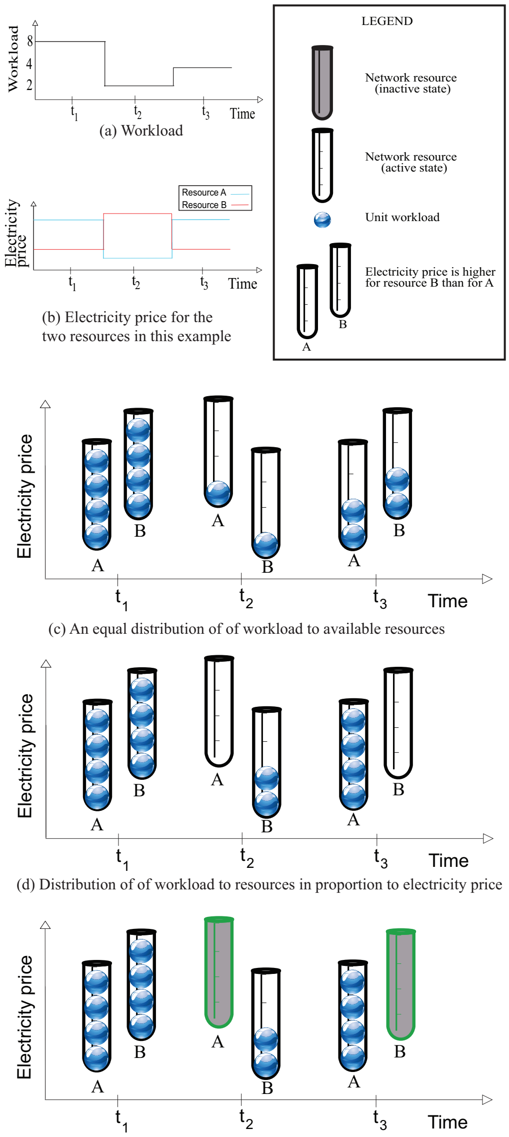
The above workload-resource mapping strategy pays no attention to electricity prices. We can reduce the electricity cost over the planning window by mapping more workload to resources at cheaper locations. We term the change in workload-resource mapping as Workload Relocation (WR). Figure 3.1 (d) shows a mapping strategy which uses WR to map as much workload as possible to resources at cheaper locations. In interval t_1, since the cumulative workload equals the total network capacity, both network resources will be operating at capacity. In interval t_2, on the other hand, we may use WR to move all workload to the network resource at the location with the cheapest electricity price (subject to resource capacity constraints, of course), thereby reducing electricity cost for that interval. In interval t_3, again, the workload may be shifted to the network resource at the location with cheapest electricity price during that interval.

Due to lack of energy proportionality in networks, the power consumption of idle resources is a large fraction of their peak power consumption. Hence, consolidation of workload to cheaper locations has limited benefit in terms of electricity cost reduction. To have considerable savings in electricity cost, one must use RP, i.e., deactivate idle resources. Notice that in the deafult workload-resource mapping strategy of Figure 3.1 (c), there is no opportunity to deactivate idle resources. However, if WR is combined with RP, as shown in Figure 3.1 (e), maximal savings in electricity cost can be achieved, because it not only shifts workload to the cheapest possible resources, but also deactivates as many resources as possible.

3.1.1 The objective function

Provide the mathematical form of the objective function that is designed to solve the optimal state trajectory problem.

[2] Network latency has been shown to have a strong correlation with the physical shortest path distance between two locations on the globe [29]. So, the commonly understood physical measure of "shortest" applies in this case.



(a) Workload

(b) Electricity price for the two resources in this example

(c) An equal distribution of of workload to available resources

(d) Distribution of of workload to resources in proportion to electricity price

(e) Same distribution as in (d) with deactivation of idle resources

Figure 3.1: An example of mapping variable workload to capacity-limited network resources with geo-temporal diversity in electricity prices. Three consecutive intervals t_1, t_2 and t_3 are considered. Workload and electricity prices may only change between two consecutive intervals. (a) Workload considered in this example. (b) Electricity prices for the locations at which the two network resources are situated. (c) A uniform mapping of workload to network resources does not exploit electricity price diversity. (d) Mapping workload to network resources in order of their current electricity price. Due to lack of energy proportionality, only slight savings in electricity cost are possible. (e) Deactivating idle resources alongwith the resource mapping strategy of (d) may result in significant electricity cost savings.

3.1.2 The constraints

Comment on some of the network-specific constraints that the optimization must be subject to.

Chapter 4

Case Study I: Geo-diverse Data Centers

4.1 Instantiating the generalized optimization formulation

Derive the objective function and constraints. Clearly outline the assumptions that we've made about the geo-diverse data centers.

4.2 Experimental setup

4.3 Results

4.3.1 Sensitivity of electricity cost savings to extent of overprovisioning

4.3.2 Sensitivity of electricity cost savings to extent of geo-diversity

4.3.3 Sensitivity of electricity cost savings to magnitude of transition costs

4.3.4 Sensitivity of electricity cost savings to resource pruning granularity

4.3.5 Sensitivity of electricity cost savings to workload estimation errors

4.3.6 Sliding window re-optimization

4.4 Discussion

Chapter 5

Case Study II: Cellular Networks

5.1 Instantiating the generalized optimization formulation

Derive the objective function and constraints. Clearly outline the assumptions that we've made about the geo-diverse data centers.

5.2 Experimental setup

5.3 Results

5.3.1 Sensitivity of electricity cost savings to the duration of an optimization interval

We may optimize at different frequencies, such as once an hour or twice an hour. In this section, we study the sensitivity of electricity cost savings to the frequency of re-optimization

5.3.2 Sensitivity of electricity cost savings to the resource pruning granularity

We may have two states for a BTS: (i) 6+6+6, (ii) 3+3+3. Or, we may have three states: (i) 6+6+6, (ii) 4+4+4, and (iii) 2+2+2. How do the two-state and three-state resource pruning granularity settings comapre in terms of electricity cost savings?

5.3.3 Sensitivity of electricity cost savings to the margin of state-change damping

Suppose that we are using a two-state resource pruning model. If t_{max} is the call capacity of a 6+6+6 site, then the call capacity of the half-pruned site is $t_{max}/2$. If we deactivate TRXs immediately when the instantaneous call volume reaches $t_{max}/2$, we are likely to have many transitions due to short-term variations in call volume. We, therefore, wait until the instantaneous call volume is $t_{max}/2 - \epsilon$ before we switch to a $3 + 3 + 3$ configuration. The value of ϵ is a configurable parameter which can take a value from 0 (very aggressive, lots of transients, perhaps more savings) to $t_{max}/2$ (very conservative, no transients, no savings either). How do the electricity cost savings vary with the value of ϵ.

5.4 Discussion

Chapter 6

Conclusions and Future Work

6.1 Contributions

Describe the contributions made by this thesis

6.2 Limitations

Discuss the limitations of our work

6.3 Future work

Future directions

Bibliography

[1] J. Verge, “Google Pumps $400 Million More into Iowa, Investment Now Tops $1.5 Billion,” April 2013. [Online; accessed 23-May-2013].

[2] “Mobile Broadband: The Benefits of Additional Spectrum,” tech. rep., Federal Communications Commission, October 2010. White Paper.

[3] R. Miller, “Facebook: $50 Million A Year on Data Centers,” September 2010. [Online; accessed 24-May-2013].

[4] A. Greenberg, J. Hamilton, D. A. Maltz, and P. Patel, “The cost of a cloud: Research problems in data center networks,” in *Computer Communications Review*, vol. 39, January 2009.

[5] S. Mbakwe, M. T. Iqbal, and A. Hsaio, “Design of a 1.5kw hybrid wind/photovoltiac power system for a telecoms base station in remote location of benin city nigeria,” in *IEEE NECEC*, November 2011.

[6] T. Italia, “Telecom Italia Annual Report 2012,” January 2013. [Online; accessed 24-May-2013].

[7] C. Peng, S.-B. Lee, S. Lu, H. Luo, and H. Li, “Traffic-driven power saving in operational 3g cellular networks,” in *Proceedings of the 17th annual international conference on*

Mobile computing and networking, MobiCom '11, (New York, NY, USA), pp. 121–132, ACM, 2011.

[8] X. Fan, W.-D. Weber, and L. A. Barroso, "Power provisioning for a warehouse-sized computer," in *ISCA '07: Proceedings of the 34th annual international symposium on Computer architecture*, (New York, NY, USA), pp. 13–23, ACM Press, 2007.

[9] C. Fraleigh, S. Moon, B. Lyles, C. Cotton, M. Khan, D. Moll, R. Rockell, T. Seely, and S. Diot, "Packet-level traffic measurements from the sprint ip backbone," *Network, IEEE*, vol. 17, no. 6, pp. 6–16, 2003.

[10] L. A. Barroso and U. Holzle, "The case for energy-proportional computing," *Computer*, vol. 40, pp. 33–37, 2007.

[11] V. Inc, "Reduce Energy Costs and Go Green With VMWare Green IT Solutions," March 2009. [Online; accessed 26-May-2013].

[12] J. S. Chase, D. C. Anderson, P. N. Thakar, A. M. Vahdat, and R. P. Doyle, "Managing energy and server resources in hosting centers," *SIGOPS Oper. Syst. Rev.*, vol. 35, pp. 103–116, Oct. 2001.

[13] G. Chen, W. He, J. Liu, S. Nath, L. Rigas, L. Xiao, and F. Zhao, "Energy-aware server provisioning and load dispatching for connection-intensive internet services," in *Proceedings of the 5th USENIX Symposium on Networked Systems Design and Implementation*, NSDI'08, (Berkeley, CA, USA), pp. 337–350, USENIX Association, 2008.

[14] D. Meisner, B. T. Gold, and T. F. Wenisch, "Powernap: eliminating server idle power," in *Proceedings of the 14th international conference on Architectural support for programming languages and operating systems*, ASPLOS XIV, (New York, NY, USA), pp. 205–216, ACM, 2009.

[15] M. Lin, A. Wierman, L. L. H. Andrew, and E. Thereska, "Dynamic right-sizing for power-proportional data centers," in *IEEE INFOCOM*, 2011.

[16] A. Qureshi, "Plugging Into Energy Market Diversity," in *7th ACM Workshop on Hot Topics in Networks (HotNets)*, (Calgary, Canada), October 2008.

[17] A. Vahdat, M. Al-Fares, N. Farrington, R. Mysore, G. Porter, and S. Radhakrishnan, "Scale-out networking in the data center," *Micro, IEEE*, vol. 30, no. 4, pp. 29–41, 2010.

[18] D. Abts and B. Felderman, "A guided tour of data-center networking," *Commun. ACM*, vol. 55, pp. 44–51, June 2012.

[19] P. Mockapetris, "Domain names - concepts and facilities." RFC 1034 (INTERNET STANDARD), Nov. 1987. Updated by RFCs 1101, 1183, 1348, 1876, 1982, 2065, 2181, 2308, 2535, 4033, 4034, 4035, 4343, 4035, 4592, 5936.

[20] P. Mockapetris, "Domain names - implementation and specification." RFC 1035 (INTERNET STANDARD), Nov. 1987. Updated by RFCs 1101, 1183, 1348, 1876, 1982, 1995, 1996, 2065, 2136, 2181, 2137, 2308, 2535, 2673, 2845, 3425, 3658, 4033, 4034, 4035, 4343, 5936, 5966, 6604.

[21] T. Berners-Lee, R. Fielding, and H. Frystyk, "Hypertext Transfer Protocol – HTTP/1.0." RFC 1945 (Informational), May 1996.

[22] P. Mahadevan, P. Sharma, S. Banerjee, and P. Ranganathan, "A power benchmarking framework for network devices," in *Proceedings of the 8th International IFIP-TC 6 Networking Conference*, NETWORKING '09, (Berlin, Heidelberg), pp. 795–808, Springer-Verlag, 2009.

[23] 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," in *Opti-*

cal Fiber Communication Conference/National Fiber Optic Engineers Conference 2013, p. OM3A.6, Optical Society of America, 2013.

[24] J. Chabarek, J. Sommers, P. Barford, C. Estan, D. Tsiang, and S. Wright, “Power awareness in network design and routing,” in *In Proc. IEEE INFOCOM*, 2008.

[25] Wikipedia, “List of mobile network operators,” 2013. [Online; accessed 09-July-2013].

[26] J. T. Louhi, “Energy efficiency of modern cellular base stations,” in *IEEE INTELEC ’07*, (New York, NY, USA), pp. 475–476, IEEE, 2007.

[27] E. Oh, B. Krishnamachari, X. Liu, and Z. Niu, “Toward dynamic energy-efficient operation of cellular network infrastructure,” in *IEEE Communications Magazine*, June 2011.

[28] J. Lorincz, T. Garma, and G. Petrovic, “Measurements and modelling of base station power consumption under real traffic loads,” *Sensors*, vol. 12, no. 4, pp. 4281–4310, 2012.

[29] B.-Y. Choi, S. Moon, Z.-L. Zhang, K. Papagiannaki, and C. Diot, “Analysis of point-to-point packet delay in an operational network,” in *INFOCOM 2004. Twenty-third AnnualJoint Conference of the IEEE Computer and Communications Societies*, vol. 3, pp. 1797–1807 vol.3, 2004.