Energy Efficiency of Voice-over-IP Systems*

Salman A. Baset
Department of Computer Science
Columbia University
New York, NY, 10027, USA

Henning Schulzrinne
Department of Computer Science
Columbia University
New York, NY, 10027, USA

November 30, 2010

1 Introduction

Voice-over-IP (VoIP) systems are increasingly prevalent in our lives. These systems come in a wide variety of flavors such as desktop-based software applications (e.g., Skype [26]), systems that replace public switched telephony network (PSTN) as the primary line voice service (e.g., Vonage [29]), and more recently, VoIP over smart phones. In November 2010, the number of concurrent VoIP users on Skype exceeded 25 million [2]. A study estimates that as of February 2010, there are approximately 110 million VoIP hard phone subscribers in the world [11]. Another study estimates that the number of mobile VoIP users will exceed 100 million by 2012 [15]. With such a large existing user base of VoIP and expected user growth, and with constantly increasing costs of energy, we ask ourselves what is the energy efficiency of these systems. To answer that question, we gather information about existing VoIP systems and architectures, build energy models for these systems, and evaluate their power consumption and relative energy efficiency through analysis and a series of experiments.

The core function of a VoIP system is to provide mechanisms for storing and locating the network addresses of user agents and for establishing voice and video

*An earlier version of this chapter titled "How Green is IP-telephony?" appeared in the proceedings of SIGCOMM 2010 Green Networking workshop.

media sessions, often in the presence of restrictive network address translators (NATs) and firewalls. These systems also provide additional functionality such as voicemail, contact lists (address books), conferencing, and calling circuit-switched (PSTN) and mobile phones. From the perspective of energy efficiency, a VoIP system can broadly be classified according to two criteria: whether it is a primary-line phone service replacing PSTN and whether it uses a client-server (c/s) or a peer-to-peer (p2p) architecture. Vonage [29] and Google Talk [10] are examples of c/s architectures, while Skype [26] is an example of a p2p architecture. Of these, only Vonage is a primary-line phone service replacing PSTN.

We begin the chapter by describing the common configurations of deployed c/s and p2p VoIP systems (Section 2). We then devise a simple model for analyzing the energy efficiency of these common configurations (Section 3). This model enables a systematic comparison of c/s and p2p configurations of VoIP systems. We then present measurements for the c/s and p2p VoIP components of these systems (Section 4) which we apply to the model developed for identifying the sources of energy wastage in these systems and the incurred economic costs (Section 5). Based on our analysis, we provide recommendations to improve the energy efficiency of VoIP systems (Section 6). Finally, we present the related work (Section 7).

2 VoIP System Architecture

We briefly explain the main functionalities of VoIP systems and describe how they are typically implemented in c/s and p2p VoIP systems. We then describe in more detail the architecture of a typical Internet telephony service provider (ITSP), an enterprise VoIP system, a softphone based VoIP system, and Skype. The first three are representative of a client-server VoIP architecture, and the latter is representative of a p2p VoIP architecture.

2.1 Functionalities of a VoIP System

The main functionalities of a VoIP system are:

Signaling - storing and locating the reachable address of the user agents, and routing calls between user agents.

NAT keep-alive - sending and processing user agent traffic to maintain state at the NAT devices for receiving incoming requests and calls.

Media relaying - sending VoIP traffic directly between two user agents or through a relay. Relaying is necessary when one or both of the user agents are behind a restrictive NAT/firewall which prevents establishment of a direct VoIP connection.

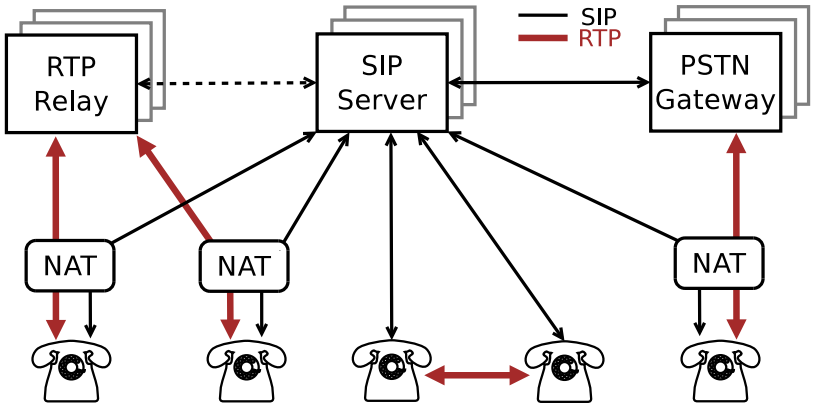

Figure 1: Client-server ITSP architecture.

Authentication, authorization, accounting - verifying that a user agent is permitted to use the system and tracking usage for billing purposes.

PSTN and mobile connectivity - establishing calls between VoIP clients, and PSTN and mobile phones using managed gateways.

Other services - such as voicemail, contact list storage, video calls, and multiparty audio and video conferencing.

Of the services listed above, signaling, NAT keep-alive, and media relaying lend themselves most easily to a p2p implementation. Consequently in the VoIP systems (including Skype) of which we are aware, all but signaling, NAT keep-alive, and media relaying functionality are implemented on centralized servers. As we will see in Section 5, the relative energy consumption of c/s and p2p VoIP systems will be determined by the relative efficiency of c/s and p2p implementations of signaling, NAT keep-alive, and media relaying.

2.2 Client-server VoIP Architecture

We consider three types of client-server VoIP systems. The first type is an Internet telephony service provider (ITSP) that provides telephony service to residential and business customers. The second type is representative of VoIP system deployment in an enterprise. The third type represents softphone-based VoIP systems like Google Talk.

2.2.1 Typical ITSP (T-ITSP)

We surveyed three c/s ITSPs in February 2010 to obtain information about their server systems, subscriber populations and characteristics of the network traffic.

Based on this survey, we present an overview of the largest of these whose architecture is typical for an ITSP. We refer to this ITSP as *T-ITSP* in order to preserve its anonymity.

T-ITSP uses an infrastructure based on open protocols, namely SIP [20] for signaling and RTP [21] for media. It uses a SIP proxy and registrar implementation based on SIP Express Router (SER) [24]. The SIP registrar stores the reachable address of user agents, whereas the proxy server forwards signaling requests between user agents. Users access the system (e.g., place calls) predominantly through hardware SIP phones. Most such phones are audio-capable only, although some also support video. The vast majority of hardphones are connected to the broadband Internet through a broadband modem which in turn is connected to a home or office router. The router is typically configured to act as a NAT/firewall. Over 90% of SIP signaling is carried over UDP. User agents connect to SIP servers, perform SIP digest authentication, and register their reachable address every 50 minutes to receive incoming calls, a process we refer to as a *registration* event.

Because most existing NAT devices maintain UDP bindings for a short period of time [9], hardphones behind NATs need to periodically refresh the binding in order to reliably receive incoming calls. The hardphones achieve this by sending a SIP NOTIFY request [18] every 15 s to the SIP server, which replies with a 200 OK response. While wasteful, this method proved to be the only reliable way of maintaining NAT bindings.

To establish a call, the user agents send the SIP INVITE requests to the SIP proxy servers, which then forward these requests to the destination user agents. The vast majority of hardphones are behind NATs/firewalls and a large proportion of these devices use default settings that prevent user agents from establishing direct VoIP calls. Consequently, T-ITSP needs to operate RTP relay servers to relay these calls, thereby consuming additional energy and network bandwidth. T-ITSP also maintains a number of PSTN servers for calling phones in the traditional telephone network. T-ITSP does not encrypt signaling or media traffic. Figure 1 illustrates the architecture of T-ITSP.

Traffic T-ITSP has a total subscriber base of approximately 100,000 users. The peak call arrival rate is 15 calls per second (CPS) and the systems see no more than 8,000 calls at any instant. Approximately 60% (or 4,800) of the peak calls are to subscribers within the ITSP; the rest are being routed to PSTN/mobile phones. Hardphones register their network address with T-ITSP's SIP registrar every 50 minutes and send a SIP NOTIFY message every 15 s to maintain the NAT binding. For 100 k subscribers, these statistics imply that the SIP registrar needs to process 33 registration events and 6,667 NOTIFY events per second. In Section 4, we extrapolate these peak numbers for a large subscriber base.

Feature	T-ITSP	Enterprise	Google Talk	Skype
User agents (UA)	Hardphone	Hardphone / Softphone	Softphone	Softphone
UAs always on	Yes	Yes	No	No
Signaling	Centralized	Centralized	Centralized	P2P+ Centralized
NAT keep-alive	Centralized	None	Centralized	P2P
Media relaying	Centralized	None	Centralized	P2P
PSTN connectivity	Centralized	Centralized	Centralized	Centralized
Voicemail	Centralized	Centralized	Centralized	Centralized
Contact list	Centralized	Centralized	Centralized	Centralized

Table 1: Comparison of T-ITSP, enterprise VoIP, Google Talk, and Skype features. The value of 'None' in the Enterprise column indicates that the user agents typically do not send NAT keep-alives, nor do they require media relays for establishing calls with user agents within the same enterprise.

2.2.2 Enterprise VoIP

The enterprise VoIP system comprises of SIP proxy and registrar servers, hardphones, and enterprise ethernet switches for connecting hardphones to the proxy server. In addition to the VoIP phones, office computers are also connected to the same ethernet switch. In some installations, the enterprise switches also provide power to the hardphones through Power-over-Ethernet (PoE) [17]. The enterprise VoIP system is connected to the other VoIP, PSTN, or mobile telephony systems through gateways. Typically, the IP address space in an enterprise is flat and the NAT devices are sporadic. Consequently, unlike T-ITSP, the hardphones do not need to periodically send SIP NOTIFY messages to keep the NAT bindings. Further, the enterprise VoIP system does not need to maintain media relay servers. When the IP address space is not flat, the VoIP systems in different departments are typically connected via gateways or call managers [5].

2.2.3 Softphone-based VoIP systems

The softphone-based client-server VoIP systems such as Google Talk are similar in their functionality to T-ITSP, except that the phone runs as a software application on a desktop or a mobile device. Such systems typically do not replace PSTN as the primary phone service.

2.3 P2P VoIP Architecture – Skype

We present an overview of Skype [26] which is representative of a p2p VoIP system. Skype is not advertised as a primary-line phone service. There are two types

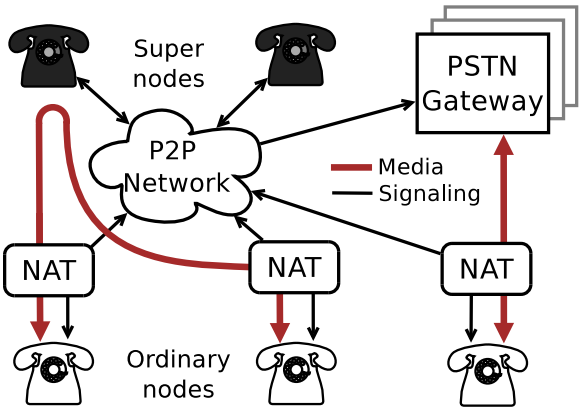

Figure 2: P2P VoIP architecture.

of nodes in a Skype network, super nodes and ordinary nodes. The super nodes form the Skype overlay network, with ordinary nodes connecting to one or more super nodes. Super nodes, which are chosen for their unrestricted connectivity and high-bandwidth, are responsible for signaling, NAT keep-alive, and media relaying. Skype encrypts signaling and media traffic to prevent super nodes from eavesdropping. Skype managed-servers provide functionality for authentication, contact list and voicemail storage, and calling PSTN and mobile phones. Figure 2 shows an illustration of a p2p VoIP system. Table 1 compares the distributed and centralized features of the T-ITSP, enterprise VoIP systems, Google Talk, and Skype.

3 Power Consumption Model

We present a model for understanding the power consumption of c/s and p2p VoIP system architectures. We focus on signaling, NAT traversal, and media relaying as they are accomplished using managed servers in the c/s but through super nodes in p2p VoIP systems. Let N be the total number of online subscribers of a VoIP system and let λ_{INV} be the peak rate of calls per second these subscribers make and d be the average call duration. These calls are either to other subscribers of the VoIP provider or to PSTN or mobile phones. Let p_v be the percentage of VoIP calls. Of these, let p_{relay} be the proportion of calls that need a relay.

3.1 Client-Server

As discussed in Section 2.2.1, a c/s VoIP architecture has dedicated servers for handling the signaling, NAT traversal, and media relaying traffic. Signaling traffic includes registration of user agent network addresses with the SIP registrar and call signaling for establishing media sessions. Let λ_{REG} and λ_{INV} denote the peak number

of SIP registration events and calls per seconds, respectively, that N user agents generate. The NAT traversal traffic (SIP NOTIFY in T-ITSP) is sent by the user agents to refresh NAT bindings and ensuring reliable receipt of incoming calls. Let λ_{NAT} be the rate of these NAT traversal messages per second. λ_{NAT} will be significantly lower for signaling over TCP than over UDP. In most c/s VoIP systems, signaling and NAT traversal are handled on separate servers from those of media-relaying.

Let $S(\lambda_{REG}, \lambda_{INV}, \lambda_{NAT}, PROTO)$ represent the number of signaling servers needed to handle the peak signaling and NAT traversal load under a particular transport protocol $PROTO$. The $PROTO$ may be UDP, TCP, or TLS. An advantage of using permanent TCP connections between user agents and SIP servers is that it reduces the frequency of the traffic to maintain NAT bindings. However, maintaining hundreds of thousands of TCP or TLS connections on a server is costly in terms of the memory needed [23]. Let $M(\lambda_{INV}, d, p_v, p_{relay})$ represent the number of media relay servers needed to relay calls. Let w_s and w_m denote respectively the wattage consumed by signaling and media servers at the peak load. Let c be the system's PUE and r_s and r_m be the redundancy factor used for signaling and media servers. Then the power consumed by the signaling and media-relay servers is given as follows:

$$w_{c/s} = (Sw_s r_s + Mw_m r_m)c \tag{1}$$

3.2 Peer-to-Peer

Recall from Section 2.3 that there are two types of nodes in a p2p communication system, namely, super nodes that forward signaling and routing traffic from other nodes and relay a call between nodes with restrictive network capacity, and secondly, ordinary nodes that do not participate in the overlay routing and connect to one or more super nodes. Let N_S be the number of super nodes in the p2p system with a total population of N subscribers. In contrast to c/s systems, where it is easy to attribute the energy consumption of signaling, NAT traversal and relaying, it is non-trivial to do so for super nodes in p2p systems. We consider two reasonable accounting strategies which apply as well to energy accounting on phones and network devices:

- **delta** - count only the additional power drawn by the signaling and relaying functions of the super node machine above that of the baseline power consumption of the machine.
- **prop** - in addition to delta, attribute to p2p VoIP a fraction of the system baseline power consumption that is proportional to the time the CPU is woken up to handle signaling, NAT traversal and media relaying traffic.

For simplicity, assume that each super node sends and receives λ_{MAINT} messages to maintain the overlay, and receives $\frac{1}{N_S}$ of the total registration, call invites, and NAT traversal. Each super node relays at maximum one call at a time. A node may

use a secure transport protocol such as TLS or DTLS for non-media relaying traffic. Let w_{base} denote the baseline wattage drawn by the super node machine. Let w_{Δ} denote the wattage drawn by the overlay maintenance, registration, signaling, NAT traversal, and media relaying functionality. Let p be the proportion of time the CPU is woken up to serve super node requests if *prop* accounting policy is chosen or zero for the *delta* policy. Then the power consumed by p2p super nodes is

$$w_{p2p} = (w_{\Delta} + w_{base}p)N_S \tag{2}$$

3.3 Comparison Issues in C/S and P2P VoIP Systems

In this section, we highlight the broader issues in comparing c/s and p2p VoIP systems.

3.3.1 PSTN Replacement

The most important consideration for our comparison is whether the systems are used as a replacement for the always-on PSTN system. For an IP-based c/s or p2p system that replaces PSTN as the primary-line phone service, the user agents must always be reachable (or powered on) to receive incoming calls. The total energy consumed by such systems is the sum total of the energy consumed by always-on user agents and servers, if any.

In contrast, systems like Google Talk and Skype run as a software application on a desktop, laptop, or a mobile device. When comparing these architectures, it is important that we examine the power consumed by the machines providing the core functionality (servers in c/s, super nodes in p2p) and not the difference in energy consumed by the user agents.

3.3.2 Network Costs

C/S and p2p communication systems have a different network footprint as in the latter, nodes have to exchange data to maintain the p2p network. Edge and core routers likely incur an energy cost for forwarding traffic for p2p and c/s communication systems. However, these costs are harder to quantify as the edge and core routers are always on. Although an analysis similar to [16] can be used, we focus on quantifying the energy usage of the system itself and not the network. However, we do incorporate the energy costs of broadband modems and network switches to which VoIP user agents are directly connected and that otherwise cannot be powered down without disconnecting the user agent.

4 Measurements and Results

In this section, we describe a set of experiments for measuring the power consumption of signaling and media relay servers, broadband modems and home routers, enterprise ethernet switches, user agents (hardphones and softphones), and Skype super nodes. Our power measurements were taken using a Watts-up .NET power meter [31]. The meter provides 0.1 W precision and claims an accuracy of 1.5% of the measured value.

4.1 Signaling and Media Relay Servers

Based on the architecture and load information of T-ITSP, we set up a test bed consisting of two servers, the first for handling signaling and NAT traversal workload, and the other for handling media relaying. Our goal was to measure the power consumption of these servers under peak load, and extrapolate the number of servers needed and the power consumed based on peak workload, using the model developed in Section 3.1. Although, this extrapolation may be considered an over simplification, it still provides useful insights into the energy consumed by large scale c/s VoIP systems.

4.1.1 Testbed Overview

In our test bed, the SIP server machine was a Dell PowerEdge 1900 server [7] with two quad-core 2.33 GHz Intel Xeon X5345 processors and 4 GB of memory. It was connected to load-generators with two Intel 82545GM Gigabit Ethernet controllers. The machine had six fans. It ran Debian Squeeze (snapshot from 26th February 2010) with Linux kernel 2.6.32. We installed the latest version of SIP-Router, an open source SIP server [24] on the machine and configured it with all the features an ITSP operating in the public Internet would need to use. The SIP server was configured to use 2.5 GB of memory and 16 processes (2 per core). We used MySQL 5.1.41-3 (from a Debian package) configured with 2 GB of query cache. We used SIPp [25] version 3.1.r590-1 to generate SIP traffic according to the model described in Section 2.2.1.

For RTP relay tests, we used an IBM HS22 blade server [12] with 5 blades installed. One of the blades was used as an RTP relay server; remaining four blades and another two desktop-class PCs were used as RTP load generators. Each blade had two Intel Xeon quad-core CPUs running at 2.9 GHz and a 10 GigE Intel NIC with multiple hardware transmission and receive queues and ran a Linux 2.6.31 kernel. We used the latest version of iptrtpproxy [13], a kernel-level RTP relay. The software relays RTP packets using iptables rules. We used a modified version of SEMS [22] to generate a large number of simultaneous RTP sessions.

4.1.2 SIP Server Measurements

We performed a number of measurements to figure out the maximum number of subscribers that our SIP server can support. We wanted to determine the maximum load on this server in three configurations: (1) signaling and NAT keep-alive (SIP NOTIFY) traffic carried over UDP as described in Section 2.2.1; (2) signaling traffic over UDP but without any SIP NOTIFY traffic; (3) signaling traffic over permanent TLS connections. The first configuration allowed us to reason about the maximum ITSP-like workload a server can handle. The second configuration provided insights into peak ITSP-like signaling workload a server can handle, assuming there were no NATs. The third configuration was helpful from the perspective of comparing T-ITSP to Skype, as Skype uses a TLS-like protocol to encrypt signaling and media traffic.

Before running any tests, we provisioned the database of the SIP server with one million unique subscribers. The baseline consumption of the server was 160 W. The machine had 6 fans; each fan consumed 10 W when running at full speed. The power consumption when all fans were removed and the machine was idle was 145 W. To see how CPUs contributed to the overall power consumption of the machine, we ran 8 cpuburn [6] processes (one per core). The machine consumed 332 W when all cores were fully utilized.

For the first configuration, we found out that our server could handle T-ITSP's traffic mix for approximately half a million users. Under this load, the number of calls (λ_{INV}), registrations (λ_{REG}), and NAT keep-alives (λ_{NAT}) events per second were 75 k, 166 k, 33 k, respectively, and the server consumes (w_s) 210 W. For the second configuration, in which there was no NAT traversal traffic, we found that our server could handle load for approximately one million subscribers. w_s was 190 W.

For the third configuration (signaling over TLS) there was no need to exchange frequent keep-alive messages over TCP connections to keep NAT bindings open, so λ_{NAT} was 0. With SIP over TLS, the SIP server used 61 kB of memory per connection and one connection was needed per user agent. Consequently, memory became our bottleneck and a maximum of 43 k simultaneously connected user agents could be supported on a single SIP server. w_s was 209 W.

Based on these measurements, we extrapolate the number of servers needed for these configurations in Table 2. Compared to the first configuration, observe that eliminating the keep-alive traffic reduces the number of servers by half in the second configuration. Although the number of signaling servers needed for the third configuration increases approximately by a factor of 12 as compared to the first configuration, we believe that such limitation can be addressed by (1) tuning the SSL buffer, (2) increasing memory in our server, and (3) using hardware SSL accelerators.

4.1.3 Media Relay Server

We managed to saturate the IBM blade with 15,000 simultaneous calls. Each call had a bit rate of 64 kbit/s for an aggregate bit rate of 960 Mbit/s. At this rate, the resource

Transport	NAT keep-alive	100 k	1 M	10 M	100 M
UDP	YES NOTIFY/s	1	2	20	200
UDP	NO	1	1	10	100
TLS	NO	3	25	250	2500

Table 2: Signaling servers needed by configuration.

% relayed calls	100 k	1 M	10 M	100 M
0%	0	0	0	0
30%	1	2	10	96
100%	1	4	32	320

Table 3: Media servers needed when relayed calls are 0%, 30%, and 100% of ITSP-ITSP calls.

bottleneck appeared to be a single CPU core overloaded by the ksoftirqd kernel thread. It is likely that even greater call volumes could be relayed by optimizing the multi-core scheduling of this machine using techniques such as [8]. At this workload, the media relay server consumed approximately 240 W (w_m). In Table 3, we extrapolate the number of relay servers needed as a function of user population and the number of calls that need relaying.

4.2 Broadband Modems, Middleboxes and Ethernet Switches

A typical residential broadband user is connected to the Internet through a home router (ethernet switch + WiFi) which in turn is connected to the broadband modem (cable, DSL, or fiber). Our measurements indicate that the recent models of WiFi routers with four ethernet switches consume, on average, 3-7 W of power. Similarly, a broadband modem also consumes 3-7 W of power. In our calculations, we use 5 W as an estimate for broadband modem and home router power consumption.

In an enterprise, the VoIP hardphones are connected to an ethernet switch which is typically PoE enabled. A 48 port Cisco switch model C2960S-48LPD-L consumes 70 W of power at five percent throughput [4] and has 370 W of available PoE power or 7.70 W per port.

4.3 User agents

We performed measurements to determine the power consumption for a variety of user agents that included hardware SIP phones and softphones. We also performed power measurements for Skype super nodes.

4.3.1 Hardware SIP Phones

For a variety of SIP-based hardphones, we found that phones consume between 3 W to 6 W of power. We also observed that the phone power consumption does not change when the user is in a voice call.

4.3.2 Softphones

We used Skype and Google Talk as representative of softphones. For several desktop machines running Windows XP and Windows 7, we did not observe any discernible change in the machine baseline power consumption when Skype and Google Talk were idle. The non-discernible change in the power draw when these softphones are idle is partially attributed to the power meter we used which can only measure power up to tenth of a watt with an accuracy of 1.5%. When placing a voice call, we found that on average Skype and Google Talk consumes between 6 W to 8 W on a Windows XP and Windows 7 desktop machine. Similarly, for a video call, Skype and Google Talk consumed between 10 W to 20 W. For laptop machines running Windows XP and Max OS X, we found that Skype and Google Talk, on average, consumed between 1-2 W when placing a voice call. As with the desktop machines, Skype and Google Talk did not cause any discernible power increase when idling. We observed similar power draw behavior for other SIP-based software clients.

4.3.3 Skype's Energy Consumption as a Super Node

Measuring Skype's energy consumption as a super node is not straightforward. First, we need a machine to transition to super node status. Since the Skype client itself decides whether to become a super node, we can only encourage this decision to be made by ensuring that the node has a public IP address, has sufficient bandwidth, and is lightly loaded (which we desired anyway given that we were trying to isolate what we assumed Skype's relatively low power consumption amidst the noise of the machine's hardware and OS). To this end, we ran a Skype client for a few hours on a machine with a public IP address and good network connectivity. To determine if the Skype is relaying a call, we performed measurements using a traffic sniffer running on another machine which is connected to the same hub as the Skype machine. We assumed a call is being relayed if the bit-rate was above a threshold [27]. Although, our meter readings indicated that there was a non-zero power increase, the difference measured was smaller than the measurement error reported by the power meter. Determining when a super node is handling signaling traffic is even harder to detect, and the power draw per event lasts for a shorter interval and is likely smaller in magnitude. We hope to address these challenges in future work. We did find that the machine can go to sleep when Skype is acting as a super node and relaying the call. The calls were either dropped or transferred to another relay; however, it is impossible for us to ascertain the status of those calls due to the closed nature of the

Users	10 k	100 k	1 M	10 M	100 M
Servers (NATs)	0.90	0.90	1.78	13.16	129.68
Servers (no NATs)	0.42	0.42	0.84	4.20	40.20
Broadband modems	50	500	5000	50,000	500,000
Home routers	50	500	5000	50,000	500,000
Hardphones	50	500	5000	50,000	500,000

Table 4: T-ITSP energy consumption as a function of number of users. All numbers are in kilowatts. The wattage for servers includes the PUE factor 'c' of two.

Skype network.

5 Discussion

Our model and measurements allow us to answer the following questions, i.e., (1) what is the total energy consumed by a VoIP system that may or may not replace PSTN as the primary line phone service, (2) where is energy consumed in such a system, (3) are p2p VoIP systems more energy efficient than c/s?

To answer questions (1) and (2), we consider T-ITSP (Section 2.2.1), enterprise (Section 2.2.2), and softphone-based VoIP deployments (Section 2.2.3). Recall that for the T-ITSP workload that include signaling and NAT keep-alive traffic over UDP, our SIP server can handle this workload for 500 thousand subscribers, and consumes 209 W (w_S) under peak load. The RTP relay server under test consumed 240 W (w_M) and can relay 15 thousand calls, with each call having a bit-rate of 64 kb/s. The number of active calls in the system for 500 k users are 24 k (by extrapolating the number of active calls for 100 k T-ITSP users), requiring two relay servers to handle this load (one server can handle 15 k calls). Depending on the actual deployment, not all calls need relaying. Our conversations with various VoIP system providers suggest that using NAT traversal techniques like ICE [19] will likely bring down the relayed sessions under 30%. When relaying 30% of the 24 k calls, only one relay server is needed. We compute $w_{c/s}$ for both 100% and 30% relaying using our c/s model (equation (1)). We plug c (PUE) as 2, and $r_S = 1$ and $r_M = 1$ in our model. For 100% and 30% relaying, the computed w_S is 1.378 kW and 0.89 kW, respectively. Observe that these numbers are approximate for the peak load and will be higher if the servers are under utilized.

Table 4 shows the energy consumed in kilowatts for running the servers, middleboxes, and hardphones. Based on our measurements, we assign 5 W for running the broadband modem and 5 W for the WiFi router with four ethernet ports. These numbers will be higher for a WiFi router with more than four ports. Nevertheless, the energy consumed by these middleboxes cannot be solely attributed to VoIP because the both VoIP and non-VoIP traffic share the same router. A reasonable assumption is that on average, such sharing occurs only for 12 hours in a day. The rest of the

time, these middleboxes must remain powered on so that a VoIP user can receive incoming calls. Using this conservative assumption, we calculate the approximate power required to run a 100 million VoIP system to be 1000.129 MW. The number is calculated by using plugging 500 MW for phones, 500 MW for broadband modems and home routers (discounted by 50% because of our usage assumption) and 129.68 kW for running servers. The monthly cost of running such a system, at 11 cents per kWh [1] is 79.2 million dollars or 80 cents per user per month (rounded up). The energy cost per month of running the servers is $10,270 or less than one thousandth of a cent per user per month.

In enterprise VoIP systems, there are typically minimal or no NATs. Consequently, the hardphones do not need to send SIP NOTIFY packets to the SIP proxy server for keeping the NAT bindings alive nor do they will likely require any media relay servers. However, VoIP hardphones must be connected to the ethernet switches. A 48-port PoE enabled ethernet switch when connected with hardphones that require 5 W per phone consumes 310 W. For an organization with 100,000 hardphones, the total number of such switches needed are at least 2084. If only one half of the ports in each switch are used for VoIP phones and the rest for non-VoIP usages such as Internet, then the number of switches increases to 4168. Assuming that switches solely serve VoIP traffic for one half of the day (ignoring weekends and holidays), the monthly power consumption and economic cost of an enterprise system with 100,000 users is approximately 465,033 kWh and $51,153, respectively. The latter number when rounded up is 52 cents per user per month.

These results indicate always on VoIP phones are a major source of energy waste in T-ITSP and enterprise VoIP systems. Further, the always on broadband modems, home routers, and enterprise switches significantly add to the energy bill. In contrast, the servers only consume a tiny fraction (<0.02%) of the total power consumed by a VoIP system replacing PSTN. Table 4 also illustrates that restrictive NATs and firewalls are wasteful in terms of server power consumption as they increase the total energy consumption of servers by a factor of two and three for number of users below and above one million, respectively.

For softphone-based c/s systems such as Google Talk that do not replace PSTN as the primary line phone service, they incur the same server energy usage for an equivalent load as for the servers in VoIP systems that replace PSTN. However, the softphone energy consumption is harder to quantify in these systems. This is because the softphones typically run on PC's which are powered on any way. If the softphones consume a small fraction of the power consumed by the PC, it is likely that they will still dominate the total power consumption of such a system; however, the relative power fraction of servers will increase. On the other hand, if the users leave their PC's powered on solely for the purpose of receiving calls (such as magicJack [14]), then the power consumption of running these softphones will be much higher than hardphones, making such systems very inefficient. As such, a user study is needed to determine how long the users keep their PC's idle but powered on for receiving

incoming calls.

To answer the third question whether p2p system is more energy efficient than c/s or vice versa, we note this will only hold if the power consumed by all the super nodes assuming a delta accounting policy is less than the total power consumed by the servers in c/s systems, i.e.,

$$w_{\Delta} N_S < w_{c/s} \tag{3}$$

Observe that this equation does not include the power consumed by user agents, broadband modems, home routers, or ethernet switches because we assume that they consume the same amount of power in c/s and p2p VoIP systems. To solve (3) for w_{Δ}, we need to estimate the total number of super nodes in the system that can process signaling, NAT keep-alive and media relaying traffic. We estimate the number of super nodes to be 1% of the total user population, meaning that in a population of 500 k user agents, 5 k are super nodes. This assumption is reasonable since if 30% of the 24 thousand active calls (7,200) need a relay, a super node roughly relays one complete call at any instant. Thus, the power consumption per super node, w_{Δ}, is $\frac{0.89\,k}{5\,k} = 0.178\,W$ in order for c/s and p2p systems to be equivalent in terms of energy efficiency. When the servers are under utilized, say 50%, w_{Δ} is twice its original value (0.356 W). The small value of w_{Δ} suggests that if the super nodes were to consume more power than this value in order to handle the signaling, NAT keep-alives, and media relaying workload, a p2p system using super nodes will become energy inefficient as compared to a c/s VoIP system.

Due to the low precision of our power meter, we are not able to ascertain if Skype super node and relaying power consumption is close to w_{Δ}. However, we speculate that the power consumed by super nodes and relays running on desktop machines may likely be close to the w_{Δ} calculated above. The reason is that the CPU of a relatively unloaded machine running a Skype super node or relay may be woken often to service these requests, thus incurring the small power draw to cause it to go above w_{Δ}. On the contrary, handling an additional job on a loaded server causes almost no additional CPU wakeups.

The analysis reveals that in a VoIP system replacing PSTN, hardphones and switching equipment consume 99.98% of the total energy consumed by the VoIP system. Thus, in order to make VoIP system more energy efficient, we need to take advantage of techniques that allow powering down these devices when idle. In the next section, we discuss the use of these techniques.

6 Recommendations for Reducing Power Consumption of VoIP Systems

In this section, we discuss using a number of existing techniques that can potentially reduce the energy consumption of hardphones, switches and middleboxes, and servers

when these devices are idle. As a result, the devices in a VoIP system will potentially only draw power when making or receiving VoIP calls. Observe that unlike cloud-based systems where services can be aggregated on a smaller number of servers to improve utilization and reduce energy wastage, it is not possible do so in a VoIP system. The reason is simple: the users want to receive and make calls through their telephones and it is simply not possible to aggregate phones similar to aggregating jobs on a server.

Our analysis showed that hardphones, broadband modems, home routers, and enterprise switches comprise the biggest of the total energy consumed in a VoIP system. To reduce the energy consumption of hardphones, the various components of the phone including LCD display, processor, and ethernet jack should be powered down when not in active use. The former two can be accomplished by turning off the LCD display and and by making use of energy efficient processors, whereas the later can be accomplished using energy efficient ethernet [3]. If the phones were only used for eight hours a day and were powered down during the remaining 16 hours, it will bring down the per user per month energy bill from 79 cents to 53 cents in T-ITSP like systems, and from 52 cents to 32 cents in enterprise VoIP systems.

In T-ITSP like systems, the hardphones must send keep-alive messages over UDP every 15 s to keep the NAT bindings alive. Such wasteful traffic prevents the phones and home routers from taking advantage of any sleep modes available on the device. To eliminate such wasteful traffic, the phones can establish a permanent TCP connection with the SIP server. Further, the ISP's can setup a SIP phone on the broadband modem which is typically not behind a NAT device. When the SIP user agent on the broadband modem receives an incoming call, it can wake up the home router and the phone using techniques such as Wake-on-LAN [30] to receive incoming call. This technique can further bring down the per user per month energy bill for running VoIP phones.

Our analysis also indicated the number of servers needed to support a large VoIP user base is fairly small; one SIP server can handle registration events and NAT keep-alive traffic for 500 thousand users, and RTP relay server can relay calls for 15 thousand calls. By setting up the VoIP user agents on cable modems, the NAT keep-alive traffic can potentially be eliminated. By using advanced NAT traversal techniques, such as ICE [19] to allow user agents to detect network conditions, the use of RTP relay server can be further minimized. These techniques will further reduce the power consumption of VoIP servers.

7 Related Work

Nedevschi *et al.* [16] have developed models describing the relative power efficiency of c/s and p2p architectures for generalized network applications (e.g., file-sharing), and conclude that p2p approaches use system energy more efficiently than the c/s ones. Similarly, Valancius *et al.* [28] argue that building p2p nano-data centers on

the Internet gateway devices provides energy savings over traditional centralized data centers. In both papers, the energy savings argument boils down to data center servers (1) needing cooling, network, and other overheads measured by a multiplicative factor called *Power Utilization Efficiency - PUE* and (2) having significant baseline power consumption (i.e., power consumption when idling). Typical data center PUEs range from 1.2–2, while the PUE of a peer is 1 (e.g., home air-conditioning is already running) and peers are on anyway, so processes running on peers escape this baseline cost.

We examined the relative energy efficiency of c/s and p2p VoIP systems, and found, intriguingly, that the energy consumption of a peer does not need to be very large in order for a p2p architecture to be *less* energy efficient than a c/s one.

8 Conclusion

We identified the key components that are implemented on servers in a c/s VoIP system and by super nodes in a p2p VoIP system (Skype). We presented a model for understanding power consumption of c/s and p2p VoIP systems. We performed a number of experiments to determine the power consumption of different components of c/s and p2p VoIP systems. Our model, analysis, and measurements indicate that for VoIP systems used as a replacement for always-on PSTN system, the power consumed by hardphones and connected network devices (broadband modems, home routers, and enterprise switches) overwhelmingly dominate the total power consumed by the VoIP system and the per user per month cost is less than a dollar in such systems. Moreover, when comparing c/s and p2p VoIP systems, our results show that even when super nodes consume relatively small power for system operation, the p2p VoIP system can be less energy efficient than a c/s VoIP system. Further, we demonstrated the presence of NATs as the main obstacle to building energy efficient VoIP systems.

References

[1] Average Retail Price of Electricity to Ultimate Customers by End-Use Sector, by State (URL). `http://www.eia.doe.gov/electricity/epm/table5_6_a.html`, (accessed November 2010).

[2] Celebrating 25 million concurrent users (URL). `http://blogs.skype.com/en/2010/11/25_million.html`, (accessed November 2010).

[3] K. Christensen, P. Reviriego, B. Nordman, M. Bennett, M. Mostowfi, and J. Maestro. Ieee 802.3az: the road to energy efficient ethernet. *Communications Magazine, IEEE*, 48(11):50 –56, November 2010.

[4] Cisco Catalyst 2960-S and 2960 Series Switches with LAN Base Software (URL). http://www.cisco.com/en/US/prod/collateral/switches/ps5718/ps6406/product_data_sheet0900aecd80322c0c.html, (accessed November 2010).

[5] Cisco Unified Communications Manager (CallManager) (URL). http://www.cisco.com/en/US/products/sw/voicesw/ps556/index.html, (accessed November 2010).

[6] CPU burn (URL). http://pages.sbcglobal.net/redelm/, (accessed June 2010).

[7] Dell Power Edge 1900 Server (URL). http://www.dell.com/downloads/emea/products/pedge/en/PE1900_Spec_Sheet_Quad.pdf, (accessed June 2010).

[8] M. Dobrescu, N. Egi, K. Argyraki, B.-G. Chun, K. Fall, G. Iannaccone, A. Knies, M. Manesh, and S. Ratnasamy. RouteBricks: Exploiting Parallelism To Scale Software Routers. In *Proc. of SOSP*, Big Sky, MT, USA, October 2009.

[9] B. Ford, P. Srisuresh, and D. Kegel. Peer-to-Peer Communication Across Network Address Translators. In *Proc. of USENIX Annual Technical Conference*, Anaheim, CA, USA, April 2005.

[10] Google Talk [accessed March 2010]. http://www.google.com/talk/, (accessed June 2010).

[11] Hard VoIP users top 100 million (URL). http://www.theinquirer.net/inquirer/news/1593216/hard-voip-users-100-million, (accessed November 2010).

[12] IBM Blade (URL). http://ibm.com/systems/bladecenter/, (accessed June 2010).

[13] iptrtpproxy (URL). http://www.2p.cz/en/netfilter_rtp_proxy/iptrtpproxy, (accessed June 2010]).

[14] magicJack (URL). http://www.magicjack.com/6/index.asp, (accessed November 2010).

[15] Mobile VoIP users to exceed 100 million by 2012 (URL). http://juniperresearch.com/viewpressrelease.php?pr=187, (accessed November 2010).

[16] S. Nedevschi, J. Padhye, and S. Ratnasamy. Hot Data Centers vs. Cool Peers. In *Proc. of HotPower*, San Diego, CA, USA, December 2008.

[17] Power over Ethernet (PoE) (URL). `http://en.wikipedia.org/wiki/Power_over_Ethernet`, (accessed November 2010).

[18] A. Roach. Session Initiation Protocol (SIP)-Specific Event Notification. RFC 3265, June 2002.

[19] J. Rosenberg. Interactive Connectivity Establishment (ICE). RFC 5245, April 2010.

[20] J. Rosenberg, H. Schulzrinne, G. Camarillo, A. Johnston, J. Peterson, R. Sparks, M. Handley, and E. Schooler. SIP: Session Initiation Protocol. RFC 3261, June 2002.

[21] H. Schulzrinne, S. L. Casner, R. Frederick, and V. Jacobson. RTP: A Transport Protocol for Real-Time Applications. RFC 3550, July 2003.

[22] SEMS (URL). `http://iptel.org/sems`, (accessed June 2010).

[23] C. Shen, E. Nahum, H. Schulzrinne, and C. Wright. The Impact of TLS on SIP Server Performance. In *Proc. of IPTCOMM*, Munich, Germany, August 2010.

[24] SIP Router Project (URL). `http://sip-router.org/`, (accessed June 2010).

[25] SIPp (URL). `http://sipp.sourceforge.net/`, (accessed June 2010).

[26] Skype (URL). `http://www.skype.com/`, (accessed June 2010).

[27] K. Suh, D. R. Figueiredo, J. Kurose, and D. Towsley. Characterizing and Detecting Relayed Traffic: A Case Study using Skype. In *Proc. of IEEE INFOCOM*, Barcelona, Spain, April 2006.

[28] V. Valancius, N. Laoutaris, L. Massoulie, C. Diot, and P. Rodriguez. Greening the Internet with Nano Data Centers. In *Proc. of CoNEXT*, Rome, Italy, December 2009.

[29] Vonage (URL). `http://www.vonage.com/`, (accessed June 2010).

[30] Wake-on-LAN (URL). `http://en.wikipedia.org/wiki/Wake-on-LAN`, (accessed June 2010).

[31] Watts up .NET power meter (URL). `https://www.wattsupmeters.com/`, (accessed June 2010).