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 ***RED-BL: Generalized Workload Relocation and Resource Pruning Procedure 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)

Recommendation of Examiners' Committee:

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

Shed some light on the role networks play in our lives, highlighting different types of networks.

1.2 Electricity costs in networks and systems

The significance of network electricity costs amongst various sources of operational costs in networks of different types [1]. Let the numbers speak for themselves.

1.3 Energy inefficiency characterizes today's networks

Networks are not energy proportional and provisioned for peak workload, therefore, they are not energy efficient. This means that operators are spending way more on electricity costs than they ideally should.

1.4 Prevalent energy efficiency improvement techniques

Electricity cost = amount of energy consumed × unit price of electricity. Therefore, electricity cost may be cut by reducing either or both of the quantities on the right hand side.

1.4.1 Reducing the amount of energy consumed

1. Deploy hardware with better energy efficiency. But that is a capital-intensive and possibly disruptive option. Our focus is on improving electricity cost for operational legacy networks.

2. The role of virtualization in reducing the energy consumption in data centers.

3. Amount of energy consumed may also be reduced using Resource Pruning (RP)

1.4.2 Using cheaper electricity

Using resources at locations with cheaper electricity, i.e., Workload Relocation (WR).

1.5 Our thesis

Our thesis is that RP and WR may be combined into a unified optimization framework that can reduce electricity cost for different types of networks.

1.6 Contributions

This thesis makes the following contributions:

1.7 Organization

The rest of the document is structured as follows

Chapter 2

Background

In this chapter we look at two different network types and observe operational similarities between them from the point-of-view of power consumption.

2.1 Geo-Diverse Data Centers

Data centers host applications that we consume every day. Operators such as Amazon, Google and Microsoft deploy data centers that are geographically dispersed for (i) fault tolerance (ii) low-latency to the clients.

2.1.1 Composition

A really basic introduction covering: composition of a typical data center (racks, pods, networking, cooling etc)

2.1.2 Request routing

front-end server based load balancing and request routing mechanisms such as IP Anycast

2.1.3 Power consumption model

Describe the power consumption model from prior work and derive a more simplified yet equivalent model

2.2 Cellular Networks

Just as data centers enable applications that we rely on every day, cellular networks are an important enabler of another pervasive service: telephony.

2.2.1 Composition

A really basic introduction to cellular networks covering: concept of cells, mobile stations (MSs), Base Transceiver Stations (BTSs) and Base Station Controllers (BSCs)

2.2.2 Call placement

How a call is handled by a BTS (at a very abstract level, i.e., how is the serving BTS chosen). Role of the BSC in cell association and call hand-off

2.2.3 Power consumption model

Describe the power consumption model from prior work

2.3 Similarities

Geo-diverse data centers and cellular networks are similar in the sense that both are built out of network resources to handle workload which results in electricity consumption.

Chapter 3

A generalized framework for electricity cost optimization

3.1 Modeling the electricity cost minimization problem in networks and systems

Discuss how different network types can periodically use RP and WR to minimize electricity costs. Show that this problem is NP-Hard. Describe the concept of network state and motivate a state trajectory optimization problem. Describe transition costs and formulate a mathematical optimization problem.

3.1.1 The objective function

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

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] K. G. Brill, "The invisible crisis in the data center: The economic meltdown of moore's law," tech. rep., The Uptime Institute, 2007.