Spring 5-2016

Joint Resource Provisioning in Optical Cloud Networks

Pan Yi
University of Nebraska-Lincoln, pyi@cse.unl.edu

JOINT RESOURCE PROVISIONING IN OPTICAL CLOUD NETWORKS

by

Pan Yi

A DISSERTATION

Presented to the Faculty of

The Graduate College at the University of Nebraska

In Partial Fulfilment of Requirements

For the Degree of Doctor of Philosophy

Major: Engineering

Specialization: Computer Engineering

Under the Supervision of Professor Byrav Ramamurthy

Lincoln, Nebraska

May, 2016

JOINT RESOURCE PROVISIONING IN OPTICAL CLOUD NETWORKS

Pan Yi, Ph.D.

University of Nebraska, 2016

Adviser: Byrav Ramamurthy

Resource allocation is an evolving part of many Cloud computing and data center management problems. For infrastructure as a service (IaaS) in the Cloud, the Cloud service provider allocates virtual machines (VMs) to the customers with required CPU, memory and disk configurations. In addition to the computing infrastructures, the bandwidth resources would also be allocated to customers for data transmission between reserved VMs. In the near future, users may also want to reserve multiple virtual data centers (VDCs) to construct their own virtual Cloud, which could be called data center as a service (DCaaS). For these two types of services, how to provide guaranteed network bandwidth over an optical network and achieve the joint resource allocation is a challenge to the central resource manager.

In this dissertation, we focus on network-aware resource allocation in Cloud/Grid over optical networks first. We investigate this problem from the provider's perspective and user's perspective. A multi-layer (IP-over-OTN-over-WDM) optical network architecture is utilized for reserving network resources. We develop mixed-integer linear programming (MILP) mathematical models and propose different heuristics for the optimal network-aware resource allocation problem from the Cloud/Grid provider's and the customer's perspectives with different targets.

Furthermore, we investigate the network-efficient virtualized cloud infrastructure provisioning (NE-VCIP) problem in IP-over-EON inter-data center network (DCN) based on the DCaaS model. The elastic optical network (EON) is adopted to provide

spectrum and cost-efficient networking resources for large bandwidth requests. We develop MILP mathematical models for this problem and propose a cost-optimized heuristic to solve this problem. To investigate the cost and blocking rate for the served demands, different modulation formats and optical transponders are compared in the EON layer, and the sliceable bandwidth variable transponders (SBVT) and optical traffic grooming technology are considered.

Finally the network-efficient virtual resource provisioning is investigated for intra-DCN based on different types of optical intra-DCN architectures: a hybrid packet and circuit switched DCN architecture (HyPaC), a novel optical switching DCN architecture (OSA) with reconfigurable optical switching matrix and a pure optical DCN architecture with fully connected non-blocking optical switching matrix. Multi-objective MILP and mixed-integer quadratic programming (MIQP) models are constructed for the optimal resource provisioning problems for the corresponding DCN architectures.

ACKNOWLEDGMENTS

This dissertation would not have been completed without the great support that I have received from so many people over the years during my PhD study. I wish to offer my most heartfelt thanks to the following people.

To my advisor, Dr. Byrav Ramamurthy—I would like to express the deepest appreciation to my advisor. Thank you for the advice and support that allowed me to pursue research on topics that I am passionate about. Thank you for spending a lot of time to discuss the research issues with me and give me suggestions on research directions, approaches and writing skills. Besides the research and work during my PhD studies, I also need to thank you for your support, help and offering guidance for my career after graduation.

To my committee members, Dr. Lisong Xu, Dr. Hongfeng Yu and Dr. Ming Han—I would like to thank you all for spending your time to discuss research problems, to answer my questions regarding my program of studies during school time with your professional expertise and giving me helpful suggestions on life and work besides the study in university. I learned a lot from your detailed review suggestions on my PhD study proposal and dissertation.

To my husband Junjie Qian—I would like to thank you for accompanying me through the happiness and sadness in our life since we met each other. We studied and we worked together in UNL during our PhD studies. Thank you for your support and encouragement when I met difficulties during my 5-year PhD study and my job hunting period.

To my family—I would like to thank my parents and my sister. Thank you for your support and understanding. I would also like to thank my sister, Li Yi, for taking care of our parents while I am abroad.

To my friends and colleagues in University of Nebraska Lincoln—I would like to thank my friends (Ertong Zhang, Hao Luo, Fujuan Guo, Guangdong Liu, Hui Ding, Jianping Zeng, Jun Wu, Lei Tian, Lina Yu, Lin Liu, Ruomeng Zhao, Wei Sun, Xin Liu, Yaoxin Liang, Yaodong Yang, Yu Bai, Zhe Zhang, Zhongyin Zhang, etc), and my colleagues in our UNL Netgroup (Adrian Lara, Bhargav Gorthi, Deepak Nadig Anantha, Mohammad Alhowaidi, Sara El Alaoui and Vishnu Sivadasan) who have graduated or are studying here. We have spent a lot of happy times together. In the work, we discuss and solve problems together. In life, we share our stories. Without you I would not have such an unforgettable 5 years in Lincoln, Nebraska.

Table of Contents

List of Figures **xi**

List of Tables **xv**

1 Introduction **1**

1.1 Grid/Cloud Computing 1

1.2 Resource Allocation Challenges in Grid/Cloud 3

1.3 Performance Isolation for Shared Resources on Cloud 5

1.4 Multi-layer Optical Networks 6

1.4.1 IP/MPLS Layer 6

1.4.2 OTN Layer 7

1.4.3 WDM Layer 7

1.4.4 Elastic Optical Network Layer 9

1.5 Motivation and Contributions 9

1.5.1 Motivation 9

1.5.2 Contributions 12

1.6 Outline 14

2 Related work **15**

2.1 Resource Allocation in Grids 15

2.2 Resource Allocation in Clouds 16

2.2.1 Data Center Management and VM Allocation 16
2.2.2 Resource Allocation with Different Objects 17
2.2.3 Approaches for Resource Allocation 18
2.3 Network Virtualization 19
2.4 Virtual Data Center Embedding in the Cloud 20

3 Provider's Viewpoint: Cost-Optimized Resource Allocation in Grids/Clouds with Multilayer Optical Network 24

3.1 Introduction 24
3.2 Joint Resource Allocation Problem 26
3.2.1 Problem Description 26
3.2.2 Problem Assumptions 27
3.2.3 Network Model 30
3.2.4 Cost Model 32
3.2.4.1 Cost model for IP/MPLS layer 33
3.2.4.2 Cost model for OTN layer 34
3.2.4.3 Cost model for WDM layer 34
3.3 MILP Formulation for Optimal Joint Resource Allocation 36
3.3.1 Inputs of the Model 36
3.3.2 Objective and Constraints 38
3.4 Heuristics for Optimal Joint Resource Allocation 42
3.4.1 Job Scheduling 42
3.4.2 Resource Co-allocation 43
3.4.2.1 Best-Fit Heuristic 43
3.4.2.2 Tabu Search Based Heuristic 44
3.5 Experimental Results and Analysis 46

3.6 Conclusion 53

4 User's Viewpoint: Budget-optimized network-aware joint resource allocation in Grids/Clouds over optical networks 55

4.1 Introduction 55

4.2 Problem Modeling 59

4.2.1 Problem Description 60

4.2.2 Problem Assumptions 60

4.2.3 Optical Network Model 62

4.2.4 Price Model 62

4.3 MILP Formulation for the budget-optimized resource allocation problem 64

4.3.1 Resource Modeling Input 65

4.3.2 Objective and Constraints of the MILP formulations 66

4.3.3 MILP Formulation Complexity Analysis 70

4.4 Heuristic Algorithms 70

4.5 Experimental Results and Analysis 74

4.6 Conclusion 83

5 Provisioning Virtualized Cloud Services in IP/MPLS-over-EON Networks 86

5.1 Introduction 86

5.2 NE-VCIP Problem 89

5.2.1 VDC mapping 90

5.2.2 RSA in EON layer 91

5.2.3 Traffic grooming with sliceable BVT in EON layer 93

5.3 Mathematical Formulation 95

5.3.1 NE-VCIP Problem Setting 95

5.3.2 Network Model 96
5.3.3 Cost Model 98
5.3.4 MILP Model 100
5.3.4.1 Full-Fit Scenario 100
5.3.4.2 Best-Fit Scenario 103
5.4 Heuristic Algorithm 105
5.5 Experimental Results and Analysis 106
5.5.1 Results for BVT-model 107
5.5.2 Results for SBVT-model 109
5.6 Conclusion 115

6 Virtualized Cloud Services Provisioning in Hybrid Optical Data Center Networks 119

6.1 Introduction 119
6.2 Data Center Network Architectures 121
6.3 VM Placement and Routing in Data Center 123
6.4 Problem Settings 124
6.4.1 VDC Demand Submitted by User 124
6.4.2 Physical Resources in Data Center 125
6.4.3 Optical Data Center Network Architecture Adopted 126
6.5 MILP for Fully Connected Non-blocking MEMS DCN Architecture . 129
6.5.1 Parameters for the Fully Connected MEMS DCN Architecture 129
6.5.2 Mixed Integer Linear Program 130
6.6 MIQP for Hybrid Packet and Circuit Switched DCN Architecture . . 132
6.6.1 Parameters for the HyPaC DCN Architecture 132
6.6.2 Mixed Integer Quadratic Program 133

6.7 MILP for OSA DCN Architecture 135
6.7.1 Parameters for the OSA DCN Architecture 135
6.7.2 Flexible bandwidth 135
6.7.3 Mixed Integer Linear Program 136
6.8 Experimental Results and Analysis 139
6.8.1 Approaches for Multiple Objectives MILP/MIQP 139
6.8.2 Experimental Results 140
6.9 Conclusion and future work 143

7 Conclusion and Future Work **145**
7.1 Conclusion 145
7.2 Future Work 147

Bibliography **148**

List of Figures

1.1 Grids and Clouds overview [1]. 3
1.2 Relationship between resource allocation challenges [2]. 4
1.3 An example of basic WDM link. 8
1.4 Flexible grid to support different bit rate demands [3]. 9
1.5 Traffic growth in Cloud. 11

3.1 Resource allocation inputs. 26
3.2 Examples of supported job structures. 29
3.3 Optical transponder mapping. 29
3.4 IP/MPLS-over-OTN-over-WDM layered network architecture. 32
3.5 The 6-node mesh topology. 48
3.6 GCE data center distribution topology constructed from public information on data center locations. 48
3.7 CapEx comparison for 10 input jobs on GCE topology. 49
3.8 Variation of BR of Best-Fit heuristic on GCE topology. 52
3.9 Variation of BR of Tabu search heuristic on GCE topology. 53
3.10 Average blocking rate comparison with the input size of 150 jobs on GCE topology. 54

4.1 The resource allocation simulator. 59
4.2 Job structure – directed multi-stage graph. 61

4.3 10-node Cloud network topology. 75
4.4 Resource utilization and unit cost of each node using MILP method for 10-node topology. 76
4.4 Resource utilization and unit cost of each node using MILP method for 10-node topology. 77
4.5 Expenditure comparison with 5 job inputs on 10-node topology, Best-Fit heuristic. 78
4.6 Expense saving ratio for 5 jobs under distinct job scheduling policies on 10-node topology, Best-Fit method. 79
4.7 The total expense comparison of Best-Fit heuristic and Tabu search heuristic with 15 job inputs on GCE topology. 80
4.8 The total expense saving ratio for different input data set size on GCE topology. 81
4.9 Variation of Blocking Rate (BR) under distinct job scheduling policies on 10-node topology, Best-Fit. 82
4.10 Variation of Blocking Rate (BR) under distinct job scheduling policies on GCE topology, Best-Fit. 83
4.11 The blocking rate of Tabu search heuristic under distinct job scheduling policies on GCE topology. 84
4.12 Blocking rate comparison by Best-Fit and tabu search under SSF job scheduling policy on GCE topology. 85

5.1 The IP/MPLS-over-EON architecture. 90
5.2 VCI demand mapping on the physical Cloud platform. 91
5.3 Multi-layer routing in the Cloud platform. 92
5.4 Optical traffic grooming with SBVTs and BV-WXCs in IP-over-EON. . . 95

5.5 Google data center locations topology (6-node). 107
5.6 NSFNET network topology. 107
5.7 Total cost comparison for BVT-model (10G BVT). 110
5.8 Blocking rate comparison for BVT-model (10G BVT). 111
5.9 Cost comparison in BVT/SBVT models under different modulation formats for demands with bandwidth requirements in: (a) Range (0 Gbps, 40 Gbps], (b) Range (40 Gbps, 100 Gbps], (c) Range (100 Gbps, 400 Gbps]. 113
5.9 Cost comparison in BVT/SBVT models under different modulation formats for demands with bandwidth requirements in: (a) Range (0 Gbps, 40 Gbps], (b) Range (40 Gbps, 100 Gbps], (c) Range (100 Gbps, 400 Gbps]. 114
5.10 Blocking rate comparison in BVT/SBVT models under different modulation formats for demands with bandwidth requirements in : (a) Range (0 Gbps, 40 Gbps], (b) Range (40 Gbps, 100 Gbps], (c) Range (100 Gbps, 400 Gbps]. 116
5.10 Blocking rate comparison in BVT/SBVT models under different modulation formats for demands with bandwidth requirements in : (a) Range (0 Gbps, 40 Gbps], (b) Range (40 Gbps, 100 Gbps], (c) Range (100 Gbps, 400 Gbps]. 118

6.1 A typical three-level tree-based DCN architecture. 121
6.2 Two possible model of a VDC request. 124
6.3 Communication matrix of a demand 125
6.4 Fully connected non-blocking 4×4 MEMS matrix optical switch. 127
6.5 C-through HyPaC DCN architecture. 128
6.6 The OSA architecture [4]. 129
6.7 The OSA overview [4]. 136

6.8 The optimal solution (all demands are accepted with minimal total cost) through two approaches for 30 demands. 141

6.9 Find the suitable value of w for MILP model for fully non-blocking MEMS DCN architecture: (a) 5 demands, (b) 10 demands, (c) 15 demands, (d) 20 demands. 142

6.10 The network traffic flow distribution in data center for 60 demands. . . . 143

List of Tables

3.1 Normalized cost for IP/MPLS layer equipments. 33

3.2 Normalized cost for OTN layer equipments. 34

3.3 Normalized cost for WDM layer equipments. 35

3.4 Parameters for Inputs 37

3.5 Other Constant Parameters 38

3.6 Variables 38

3.7 CapEx ($\times 10^3$) comparisons between OPL and two proposed heuristics with different job scheduling policies on a 6-node topology. 50

3.8 Running time (seconds) comparisons between OPL and two proposed heuristics with different job scheduling policies on a 6-node topology. . . 50

4.1 Price model for processor resource 64

4.2 Price model for storage resource 64

4.3 Price model for network resource 64

4.4 Constant Parameters 66

4.5 Variables 66

4.6 Decision Variables 67

4.7 Total expenditure comparison on 10-node topology 78

5.1 Parameters 97

5.2 Cost Model [5] 99

5.3 Cost and time comparison between CPLEX solver and heuristic for the Full-Fit 108

6.1 MEMS connection configuration between racks 143

Chapter 1

Introduction

1.1 Grid/Cloud Computing

The development of the Grid/Cloud network offers users a powerful platform for large-scale computing and data processing. With Grid/Cloud technologies, users will not execute the tasks on local computers, but on centralized third-party compute and storage facilities. Therefore, how to adopt an effective resource scheduling method to allocate the resources in Grid/Cloud is becoming important.

The Grid enables users to share a large amount of storage, memory and computing resources over a network [6]. A job submitted to the Grid might not execute on a single computer but is separated to execute on several computers. The Grid resource scheduling includes mainly three phases: resource discovery, resource allocation and job execution [7]. Many methods have been developed for Grid resource allocation. In addition, researchers have developed the Grid technology in many practical ways, such as Open Science Grid (OSG) [8] and Global Environment for Network Innovations (GENI) [9], to provide computing power, data and distributed systems research and education. OSG provides distributed computing resources to users to meet their needs of research and academic communities at all scales. To maintain and improve distributed high throughput computing services, managing resources responsibly and efficiently becomes an essential task. HTCondor which is a specialized management

system[10] is used in OSG to take care of scheduling applications and for continually checking the available resources in the Grid. HTCondor acts as a local resource manager which collects the resource information in a certain region of the Grid and maps the submitted jobs in this region to the matched resource pool according to specific requirements by users. However HTCondor does not deal with the network resource allocation. So this is one of the reasons that we investigate the network-aware resource allocation in Grid/Cloud in the work discussed later in this dissertation.

The Cloud is a rapidly developing technology in recent years. The Cloud has some aspects in common with the Grid technology. Both need to manage large facilities and to define methods by which users will request and use resources provided by the facilities in Grid/Cloud. Cloud computing has indeed evolved out of Grid computing, and relies on Grid computing as its backbone and infrastructure support [1]. Hence we refer to the network as *Grid/Cloud network* in this work interchangeably. Clouds provide services at three different levels in general: Infrastructure as a Service (IaaS), Platform as a Service (PaaS) and Software as a Service (SaaS) [11]. IaaS, among the three levels, provisions hardware, software, and equipments to users with usage-based pricing model. The Amazon Elastic Compute Cloud (EC2) [12], Google Compute Engine (GCE) [13] and Windows Azure from Microsoft [14] are successful commercial Cloud technology products. They provide on-demand and reserved infrastructure that scales and adapts to the consumer's needs. A simple example is a request for IaaS where the job submitted by the user requests a number of certain type of virtual machines (VMs), and a certain amount of bandwidth for data transmission between VMs. In this case, the Cloud resource scheduler maintains the status of all the resources in each data center in the Cloud, in order to complete the resource allocation for the requests. The consumers only pay for what they use with the "pay-as-you-go" model in Cloud. Therefore from the user's perspective, what they want

intuitively is to obtain required resources from the Cloud platform for their jobs at a minimum cost.

The Grid computing and Cloud computing technologies overlap with each other, and also with some other technologies such as supercomputers and clusters. The Cloud computing is evolved out of Grid computing and relies on Grid computing as its backbone and infrastructure support [1]. Figure 1.1 shows an overview of the relationship between Cloud, Grid and other distributed technologies. In the dissertation, we refer to Grid/Cloud interchangeably.

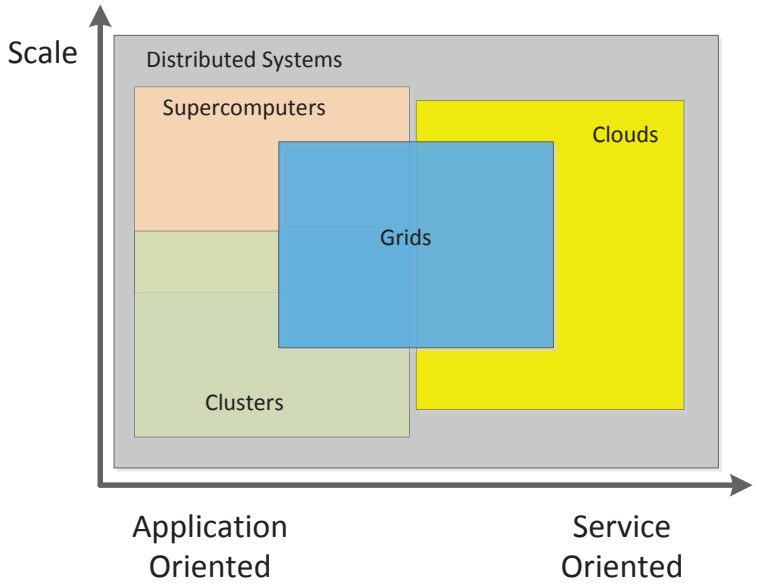


Figure 1.1: Grids and Clouds overview [1].

1.2 Resource Allocation Challenges in Grid/Cloud

In the Grid/Cloud, to complete the resource allocation for users, the resource allocator system should know the status of each type of resources in the distributed Grid/Cloud, and based on the resource status apply efficient algorithms to allocate physical or virtual resources to users while satisfying their requirements. The chal-

lenges for resource allocation in Grid/Cloud mainly lie in the fundamental aspects: resource modeling, resource offering and treatment, resource discovery and monitoring, resource selection and optimization [2]. The first two challenges belong to the conception phase, where the Grid/Cloud provider needs to model resources according to the type of services and resources it will supply. The last two challenges belong to the operational phase. In this phase, the resource allocator needs to monitor the resource status in Grid/Cloud and find available resources that satisfy the current arrival demand. After that it will allocate corresponding resources to serve the demand and update the resource status in Grid/Cloud. Figure 1.2 represents the relationship between resource allocation challenges in the distributed environment [2]. Developing solutions to cope with the resource allocation challenges is still an essential topic in the area of Grid/Cloud computing.

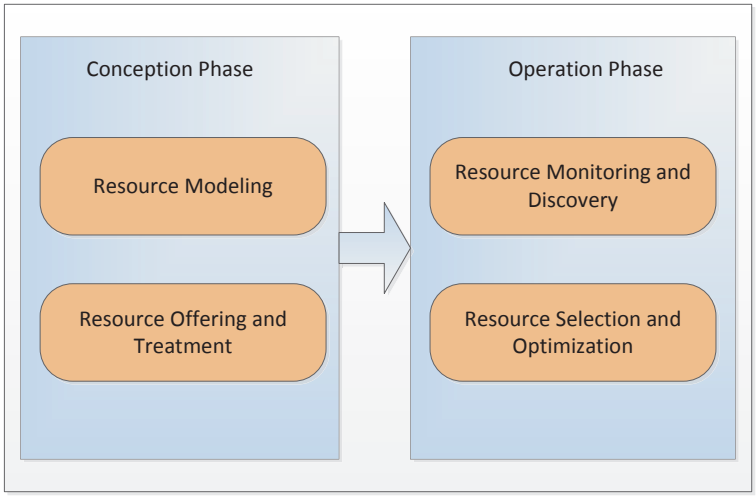


Figure 1.2: Relationship between resource allocation challenges [2].

For the resource selection and optimization, which is one of the four challenges, the provider needs to fulfill the user's requirements and optimize the utilization of the infrastructure when given the information regarding Cloud resource availability at hand. Studies have been carried out on the resource allocation for Grid/Cloud

networks with various distinct objectives. Among them, some studies only targeted the computational resource allocation or merely network resource allocation in a Grid/Cloud. However, in practice, a submitted task might obtain infrastructure resources from several data centers in a Cloud network to complete their execution; in this case they might need to transmit the final and intermediate data between data centers in the Cloud network. This circumstance leads to another challenge of considering network resource in Grid/Cloud networks.

1.3 Performance Isolation for Shared Resources on Cloud

In recent years, the cloud providers have moved from simply supplying computing resources to supplying multiple types of services, including networking, elastic caching, database, analytics [15]. When deal with resources sharing among multiple customers, the performance isolation becomes a challenge for the cloud providers. Significant works have been done to investigate the performance isolation on different aspects. The work in [16] focuses on the performance isolation on shared resources such as processor caches, memory buses and CPU. Another work in [17] focuses on the cloud storage sharing and performance isolation problems. Different from above twos, the work in [18] not only focuses on a single type of resource or multiple resources in a single appliance, this work focuses on the end-to-end performance isolation in multi-tenant data centers at multiple appliances. The abstraction of a dedicated virtual data center (VDC) is proposed in such investigations to deal with virtual resource provisioning and isolation in Cloud. A VDC consists of virtual machines (VMs) and virtual resources lick virtual appliances and virtual network. Two main challenges are presented by providing VDC abstraction. One is that customers can be bottlenecked at different appliances to network links. The other one is that resource consumed by

a demand can vary based on demand characteristics such as type, size [18].

1.4 Multi-layer Optical Networks

An optical (photonics) network is a communications network in which information is transmitted as optical signal through the optical fiber. Compared to the traditional Ethernet network that uses electrical transmission, the optical network has a much higher transmission speed and provides higher bandwidth. In addition, the dynamic provisioning characteristic of the optical network makes optical channels can be split into many high speed wavelengths, allows network managers to increase the capacity of their optical network at very short notice [19]. This is one of the reasons that optical network is widely used in network backbones within buildings and across wide area.

1.4.1 IP/MPLS Layer

The traditional IP routing has several limitations, such as scalability issues to poor support of traffic engineering and poor integration with layer 2 backbones already existing in large service provider networks. Multi-protocol Label Switching (MPLS) is a standard technology for speeding up network traffic flow and making it easier to manage [20]. MPLS allows most packets to be forwarded at layer 2 using switching rather than at layer 3 using routing. An IP/MPLS network is a packet switched network that uses the Internet protocol (TCP/IP) enhanced with the MPLS standard. Compared to traditional IP network and MPLS only network, the IP/MPLS network has several advantages [21]: 1) traffic is no longer delivered by using the destination address. It is labeled at source and based on the label given to the traffic; 2) packets are guaranteed to arrive with a specific cost and time if the specified path is available; 3) an alternative path can be specified in advance in case the first specified path is

not available; 4) there is further enhancement of the quality of service (QoS), etc. The router model of the IP/MPLS layer network consists of two main parts: the basic node and the equipment related to the physical layer interfaces. The basic node in the IP/MPLS router model provides a certain number of bidirectional slots with a fixed bandwidth. The slots must be equipped with a slot card that in turn can connect to different type of port cards [22]. In our work we use a common approach that connects an IP/MPLS router and a WDM system with a short reach interface and a WDM transponder if needed for IP/MPLS and WDM layer connection.

1.4.2 OTN Layer

Optical Transport Network (OTN) is a multiplexing and transmission technology that can provide transport, time division multiplexing and management of optical signals. OTN technology is circuit switched and connection oriented, which means a fixed path is pre-established between an input port and an output port and all frames received on a port follow the fixed path. OTN offers the following advantages relative to synchronous optical networking and synchronous digital hierarchy [23]: 1) stronger forward error correction; 2) more levels of Tandem Connection Monitoring; 3) transparent transport of client signals; and 4) switching capability. In OTN today, switching is provided by electrical cross connects (EXC) in general, which consists of EXC basic node, and line cards. The way that an EXC connects to a WDM system is by using a short reach interfaces at the EXC side and a separate WDM transponder.

1.4.3 WDM Layer

The WDM layer has the functionality of multiplexing and transmitting a number of optical carrier signals with different wavelengths in a single optical fiber, and of switching the signals in transparent optical switches. A traditional 50 GHz WDM

layer link is composed of transponders, muxponders, regenerators, optical amplifiers (OA) and WDM terminals. In addition, in the transparent nodes of WDM layer network, optical switches, such as reconfigurable optical add/drop multiplexer (ROADM) and the optical cross connect (OXC) [24], are needed for supplying optical switching without optical-to-electrical-to-optical (O/E/O) conversion. Figure 1.3 shows a basic WDM link example.

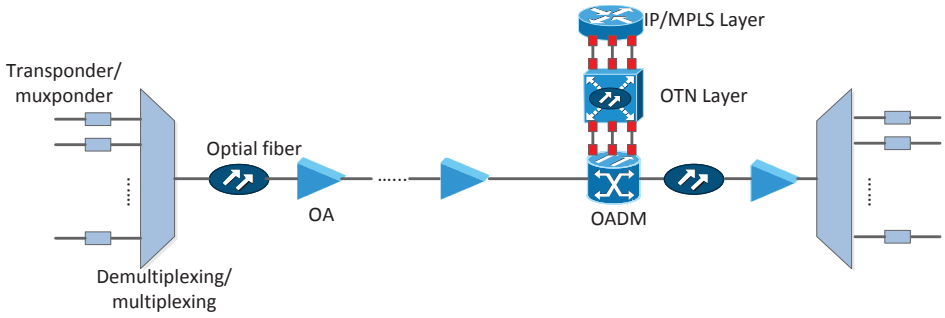


Figure 1.3: An example of basic WDM link.

In the WDM network layer, it is assumed that all wavelength channels are terminated by transponders at the two end nodes of the optical route in the network, and a maximum transparent reach is denoted for each transponder pair. The muxponder is a special device used to realize traffic grooming mechanism. The regenerator is used to provide regeneration on each individual wavelength to restore optical signals that are subject to noise and crosstalk. Regeneration usually occurs at a ROADM location where the wavelength can be dropped and/or demultiplexed [25]. The OA element is capable of amplifying all wavelengths carried by the optical fiber bidirectional. The WDM terminal is responsible for multiplexing/demultiplexing multiple wavelength channels.

1.4.4 Elastic Optical Network Layer

The elastic optical network (EON) has the features of dividing optical spectrum flexibly and generating elastic optical paths, that is paths with variable bit rates, through the new transceivers called bandwidth variable transponders (BVTs). The main motivations for developing EON paradigm are: 1) support for 1 Tbps and other high bit rate demands; 2) satisfy disparate bandwidth needs in the same network (see Figure 1.4); 3) allow for closer spacing of channels, in order to free up spectrum for other demands; 4) trade off the optical signal reach and spectral efficiency well; 5) support dynamic network in the optical layer, that is to response directly to variable bandwidth demands from the client layer [3].

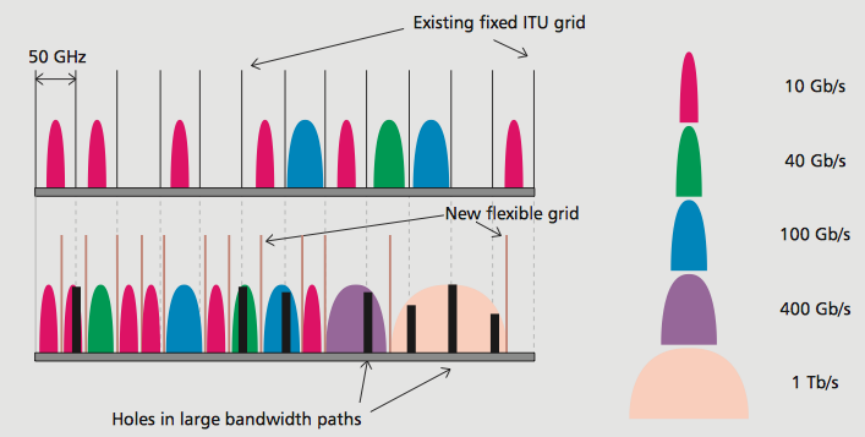


Figure 1.4: Flexible grid to support different bit rate demands [3].

1.5 Motivation and Contributions

1.5.1 Motivation

Resource allocation is an evolving part of many Grid/Cloud computing and data center management problems. In Grid computing, as we discussed above, a specialized management system HTConder who is responsible for resource scheduling and alloca-

tion does not provide network resource allocation. And the OSG Council stated that incorporating the network layer into scheduling is a key distributed high throughput computing (DHTC) research challenge for the next five years [26].

In Cloud computing, based on the infrastructure as a service (IaaS), the Cloud service provider allocates virtual machines (VMs) to the customers according to the CPU, memory and disk requirements of the VMs. In addition to the computing infrastructures the Cloud service provider would allocate bandwidth resources to the customers for data transmission between reserved VMs. The bandwidth resources that are offered by the Cloud providers (e.g. Google, Amazon) today are just the total amount of data you could transmit for a certain time duration (e.g. 100 GB per day). The Cloud service providers do not offer guaranteed bandwidth for the customers for the service period they have reserved. In this case, the smooth data rate for the customer during the service period cannot be guaranteed due to the limited bandwidth when the network communication load is heavy. Moreover, the best-effort data transmission in the Cloud might lead to unpredictable network performance.

The Cisco Global Cloud Index (GCI) [27] is an ongoing effort to forecast the growth of global data center and cloud-based IP traffic. GCI indicates in the forecast and methodology report for 2013-2018 that, the global data center traffic and global Cloud traffic will increase significantly in the future years [27] as shown in Figure 1.5. With the prediction of significantly traffic growth in Cloud, how to manage network resource in Cloud environment becomes important. In fact, the two main telecommunication carriers with their own global IP networks in US, Verizon and AT&T, plan to extend their IaaS service in the Cloud computing with network resources, while considering guaranteed network bandwidth [28] [29].

To construct bandwidth guaranteed circuits for data transmission in the Cloud environment, we will consider involving the optical layer network. Optical networks play

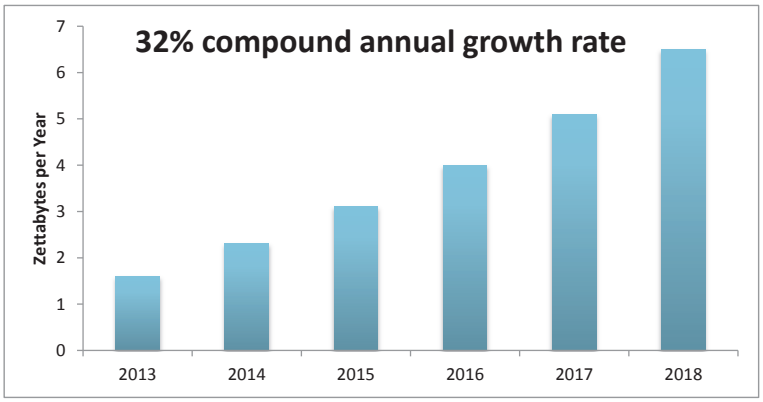


Figure 1.5: Traffic growth in Cloud.

a key role in the realization of Grid/Cloud computing networks. Optical transmission is accepted as the most cost-effective way to realize high-bandwidth connections in the long-haul network [30]. The technology's ability to transfer huge data volumes with low latency has made optical networks the *de facto* standard to connect data centers that provide computing and storage services in Grid/Cloud [31]. Furthermore, the technology of dense wavelength-division multiplexing (DWDM) multiplexes a a number of optical carrier signals onto a single optical fiber and allows optical end-to-end connections over different wavelengths. Commercially available line rates offered by a single wavelength include 10, 40, or 100 Gbps, while channels are typically spaced 50 or 100 GHz [32]. To adapt to the actual traffic needs, more flexible and adaptive networks that equipped with flexible transceivers and network elements are needed for optical layer network. An new approach known as elastic optical network (EON) appears to provide flexible bandwidth for variable bit rate demands from the client layer. An example would be IP-over-EON, in which the bandwidth-variable transponders (BVTs) adjust their bandwidth in line with the IP layer demands [3].

1.5.2 Contributions

In this dissertation, we make the following contributions.

1. Cost-optimized joint resource allocation in Grid/Cloud with multi-layer optical network architecture

We introduce the multi-layer optical network architecture to guarantee the reservation of the network bandwidth resource. We investigate the bandwidth guaranteed joint resource scheduling from the Cloud provider's point of view, which is completing the resource scheduling with minimal capital expenditure (CapEx). The Mixed Integer Linear Programming (MILP) formulations and heuristics (Best-Fit and Tabu search) are developed to solve our problems. The results show that both MILP and heuristics work well to solve the problem, and the heuristics are much more time-efficient. In addition, the Tabu search method achieves the optimal resource allocation, and also reaches a lower blocking rate compared to Best-Fit method [33].

2. Budget-optimized network-aware resource allocation in Grid/Cloud over optical networks

In addition, we focus on network-aware optimal resource allocation in the Cloud from the customer's perspective. We develop a mixed integer linear programming (MILP) optimal mathematical model and heuristics (Best-Fit [34] and Tabu Search [35]) to solve the budget optimized joint-resource allocation problem to minimize the rental cost for each customer. The experimental results show that our heuristics can achieve approximate optimal solution to the MILP solution and can reduce the customer's rental cost by at least 30%. The Best-Fit heuristic with shortest job execution time first (STF) and simplest job structure first (SSF) scheduling policies have a better performance in terms of the traffic blocking rate. The traffic blocking rates under both scheduling policies are 5%$\sim$25% less than other policies. The Tabu

search based heuristic with SSF job scheduling policy has a better performance in terms of the traffic blocking rate than other job scheduling policies. In addition, the Tabu Search based heuristic also reduces the blocking rate by 4%∼30% compared with Best-Fit heuristic under any job scheduling policy [36].

3. Provisioning virtualized Cloud services in IP/MPLS-over-EON networks

In this part, we consider the network-efficient virtualized cloud infrastructure provisioning problem in IP over elastic optical network (IP-over-EON) based on the data center as a service model [37]. The elastic optical network is adopted to provide spectrum and cost-efficient networking resources for large bandwidth requests in our work. We develop mixed integer linear programming (MILP) formulations to construct the mathematic model for this problem and propose a cost-optimized heuristic to solve this problem. To investigate the cost and blocking rate for the served demands, different modulation formats are compared in the EON layer, and the sliceable bandwidth variable transponders and optical traffic grooming technology are considered. The experimental results show that different modulation formats that are adopted in the EON layer will have different impacts on the total cost and demand blocking rate for the same data set size. Also the use of SBVT will reduce the total cost no matter which modulation format is adopted, and the reduction is related to the bandwidth requirement of the demands [38].

4. Virtualized Cloud services provisioning in hybrid optical data center network

In this part, furthermore, we consider the network-aware virtualized cloud services provisioning within data center based on different optical data center network (DCN) architectures. Three types of optical DCN architectures are considered. A simplest pure optical DCN architecture, in which the top of rack switches connect to each other through a fully connected non-blocking MEMS matrix optical switch. A hybrid packet and circuit switched DCN architecture (HyPaC), in which the tradi-

tional packet switching through tree-based architecture is augmented with the high bandwidth, low complexity optical circuit switching through re-configurable MEMS optical switch (one degree connection). A novel optical switching architecture (OSA) through reconfigurable MEMS optical switch ($k \geq 1$ degree connections). We develop MILP and mixed integer quadratic programming (MIQP) models of resource provisioning problem for correlated architectures. Two approaches are adopted to solve the problems with optimal optical switch configuration that could accept maximal number of demands with minimal total cost.

1.6 Outline

The rest of the dissertation is organized as follows. In Chapter 2 we discuss some related work on the resource allocation problems in Grid/Cloud. We investigate the cost-optimized joint resource allocation in Grid/Cloud over multi-layer optical network from the provider's point of view in Chapter 3 and the budget-optimized resource allocation from the customer's point of view in Chapter 4 respectively. In Chapter 5 we investigate the network-efficient virtualized Cloud infrastructure provisioning problem in IP-over-EON network based on the data center as a service model. Furthermore, in Chapter 6, we extend the virtualized Cloud service provisioning for intra-data center network in which an optical architecture data center is considered. Finally, conclusion and future work are presented in Chapter 7.

Chapter 2

Related work

2.1 Resource Allocation in Grids

Many studies have been carried out on resource allocation or task scheduling in the Grid networks [39] with different requirements and objectives in different application fields. The study in [40] focuses on the optimization problem of jointly scheduling computing and network resources, which is called task scheduling and light-path establishment (TSLE), in the Grid to achieve the optimal performance for the data-intensive Grid applications. Two optimization problems are studied with the objectives of minimizing the completion time of a job and minimizing the resource usage/cost to satisfy a job with a deadline respectively. The work in [41] considers the efficient resource allocation problem in ad hoc Grid environment. To reach the goals of both obtaining the optimized quality of service for the agents and maximizing the profit for the Grid resource providers, the ad hoc Grid resource allocation algorithm is proposed which can maximize the global utility of the ad hoc Grid system. The work in [42] focuses on measuring and quantifying the existing resource fragmentation caused by scheduling the jobs in advance, and also on improving the resource usage in the Grid system. Different metrics are proposed to measure the fragmentation presented in a Grid system. These metrics can be applied to trigger the rescheduling of jobs when needed to improve the resource utilization in the Grid.

In addition, the support for advance reservations of resources plays a key role in Grid resource management. The work in [43] investigates the impact of heterogeneity on Grid resource management when advance reservations are supported. And an efficient heterogeneity-aware resource scheduling algorithm which deploys techniques from computational geometry is developed in this work. The work in [44] studies the high performance resource utilization strategies that can be employed in Grid and Cloud networks. It also implements and quantifies strategies including advanced reservation, just-in-time bidding and etc.

2.2 Resource Allocation in Clouds

In the Cloud resource allocation problem, users requires a certain amount of computing resources or VMs, and the resource manager will assign required resources to the users.

2.2.1 Data Center Management and VM Allocation

In the field of Cloud computing, the studies in [45] and [46] propose some solutions for the resource allocation problem which focused on the management of data centers. The study in [45] uses Lyapunov optimization technique to design an on-line admission control, routing, and resource allocation algorithm for a virtualized data center. And the study in [46] proposes an efficient dynamic task scheduling scheme for virtualized data centers. The virtual machine (VM) allocation is a challenging sub-problem as well. The work in [47] investigates the dynamic VM provisioning and allocation problem for the auction-based model. An integer program is formulated and truthful greedy and optimal mechanisms are designed for the problem. The proposed mechanisms achieve promising results in terms of revenue for the Cloud

provider. The work in [48] presents a system that uses virtualization technology to allocate data center resources dynamically based on application demands and support green computing by optimizing the number of servers in use. It introduces the concept of "skewness" to measure the uneven utilization of a server. A work in [49] introduces an efficient network virtualization solution, *CloudMirror*, with three components - a network abstraction, an efficient VM placement strategy and a runtime mechanism that enforces the application bandwidth requirements.

2.2.2 Resource Allocation with Different Objects

In addition, a lot of studies have been also carried out for emphasizing different aspects of the resource allocation problems in Cloud. The work in [50] focuses on the development of dynamic resource allocation that considers the energy between various data center infrastructures to improve energy efficiency and performance. Also the work in [51] proposes an energy efficient virtual network embedding approach to deal with the on-demand allocation of network resources for Cloud. The work in [52] focuses on the load balancing task scheduling in Cloud networks. The author proposes an optimized algorithm based on Fuzzy-GA optimization to achieve better load balancing across all nodes in Cloud networks. The work in [53] develops efficient resource allocation algorithms in distributed Clouds that aim at minimizing the communication costs and latency. To reduce the bandwidth cost, the authors propose an algorithm to choose data centers in Cloud that are close to the user. The objective is to minimize the maximum distance between selected data centers. The work in [54] proposes a new *cloud brokerage service* that reserves a large pool of instances from Cloud providers and serves customers with prices discount. The dynamic strategies are proposed for the broker to make instance reservations with the objective of minimizing its service cost. The work in [55] focuses on the cost-effective resource

allocation in the Cloud. This paper provides answers to three fundamental questions: Given a pub/sub workload, what is the minimum amount of resources needed to satisfy all the subscribers; what is a cost-effective way to allocate resources for the given workload; and what is the cost of hosting it on a public Cloud provider. A problem coined minimum cost subscriber satisfaction (MCSS) is formulated to answer the above three questions. Also related with the cost aspect of resource allocation, the work in [56] investigates how to dynamically allocate resources to optimize resource provisioning cost, while satisfying QoS requirement specified by individual customers simultaneously. The authors propose a decentralized Cloud firewall framework for individual Cloud customers and propose novel queuing models to solve this problem.

2.2.3 Approaches for Resource Allocation

Moreover, to solve the problem of allocating resources to the requests while maintaining high resource utilization, many approaches, such as heuristic algorithms, statistical methods, and soft computing techniques have been investigated by researchers. The work in [57] utilizes a variation of multi dimensional bin packing to model the resource allocation problem and present an efficient resource allocation algorithm using simulated annealing. The work in [58] proposes a Peer to Peer (P2P) resource management approach, which is comprised of a number of agents, to address the problem of resource management for large scale data centers. The work in [59] proposes an important differential evolution algorithm (IDEA) to optimize task scheduling and resource allocation based on the described cost and time models on Cloud computing environment. The work in [60] investigates the problem of joint optimizing the service cost and resource utilization for Clouds. A nonlinear integer programming model is formulated for the optimal reservation problem and a fine-grained heuristic algorithm is proposed to reduce its computational complexity and obtain quasi-optimal

solutions.

2.3 Network Virtualization

Network virtualization is a powerful method to execute multiple experiments simultaneously on a shared infrastructure. However, making efficient use of the underlying resources requires effective techniques for virtual network (VN) embedding–mapping each virtual network to specific nodes and links in the substrate network [61]. In addition, with the growth of data volumes and variety of application demands in the Cloud environment, the data center network virtualization is a promising solution to address the problem of efficiently allocating multiple types of resources (storage, computing, bandwidth) from underlying infrastructures for these demands. The problem of embedding virtual networks in a substrate networks is the main resource allocation challenge in network virtualization [62] and has attracted a lot of attention in both academic and industry.

The work in [63] applies the Markov Random Walk (RW) model to rank a network node based on its resource and topological attributes. And using this node ranking, two VN embedding algorithms are proposed to solve the VN embedding problems with higher long-term average revenue and higher VN request acceptance ratio. Toe solve the VN embedding problem, different heuristics might be proposed. The work in [64] presents ViNEYard–a collection of VN embedding algorithms that leverage better coordination between the two phases. We formulate the VN embedding problem as a mixed integer program through substrate network augmentation. This work devise two on-line VN embedding algorithms D-ViNE and R-ViNE using deterministic and randomized rounding techniques respectively. In addition a generalized window-based VN embedding algorithm (WiNE) is presented to evaluate the

effect of lookahead on VN embedding. Another work in [65] proposes a new scalable embedding strategy named VNE-AC based on the Ant Colony meta-heuristic to solve the VN embedding problem with the target of mapping virtual networks in the substrate network with minimum physical resources while satisfying the required QoS in terms of bandwidth. The simulation results show that the proposed meta-heuristic can enhance the substrate network provider's revenue.

Furthermore, other related works about the VN embedding that focus on different aspects have been investigated as well, such as the work in [66] focuses on the energy efficient VN embedding problem and the work in [67] focuses on the survivable VN embedding problem.

2.4 Virtual Data Center Embedding in the Cloud

Virtualizing data center networks has been considered a feasible strategy to satisfy the requirements of Cloud services. For the VDC allocation in the Cloud, VDC is treated as the unit of resource allocation for multiple users in the Cloud. The mapping of virtual data center (VDC) resources to the physical Cloud resources (facilities), also noted as VDC embedding, can impact the revenue of Cloud providers. Therefore the VDC embedding problem plays an important role in the Cloud resource provisioning area and some studies have been investigated on this area. The work in in [68] proposes a new embedding solution for DCs that considers the relation between switches and links, allows multiple resources to be mapped to a single physical DC, and reduces resource fragmentation in terms of CPU. The work in [69] studies the virtual resource allocation problem for networked cloud environments, incorporating heterogeneous substrate resources, and provides an approximation approach to address the problem. For the node mapping phase, a MIP formulation capable of taking

into accounting QoS requirements is considered. For the link mapping phase, the corresponding flow problem is adopted to solve the problem. The work in [48] presents a system that makes use of virtualization technology to allocate DC resources dynamically and targets optimizing the number of servers in use. A set of heuristics are developed to prevent overload in the system while saving energy used. Moreover, to get the maximum benefit from a distributed cloud system, efficient algorithms are needed for resource allocation which minimize communication costs and latency. The work in [70] develops efficient resource allocation algorithms to address such problems in distributed clouds. The target of this work is to minimize the maximum distance, or latency, between the selected DCs.

In addition, VDC networks have been considered as a feasible alternative to satisfy the requirements of advanced Cloud infrastructure services. Proper mapping of VDC resources to their physical counterparts , also known as VDC embedding, can impact the revenue of cloud providers [68]. In addition to the VM resources, the work in [68] proposes a new embedding solution for DCs that considers the relation between switches and links, and allows multiple resources to be mapped to a single physical DC. The work in [71] focuses on reliable VDC embedding in clouds.The paper presents a technique for computing VDC availability that considers heterogeneous hardware failure rates and dependencies among virtual components. An availability-aware VDC embedding framework, Venice, is proposed for achieving high VDC availability and low operational cost. This work focuses on embedding VDCs onto one physical data center. The work in [72] designs a data center network virtualization architecture called *SecondNet* to enable the VDC abstraction. *SecondNet* is scalable by distributing all virtual-to-physical mapping, routing, and bandwidth reservation state in server hypervisors. *SecondNet* introduces a centralized VDC allocation algorithm for virtual to physical mapping with bandwidth guarantee. The work in [73] about

VDC embedding proposes *Greenhead*, a holistic resource management framework for embedding VDCs across geographically distributed data centers connected through a backbone network. The target of Greenhead is to maximize the cloud provider's revenue while ensuring that the infrastructure is as environmentlly friendly as possible. This work focuses on embedding VDCs onto distributed infrastructures which is different from the work in [74] and [68]. Moreover, the optical network with high throughput and low latency has been used for Cloud environment construction and it could also be used for network resource provisioning in Cloud in the future. The work in [75] presents a cross-functional orchestration platform able to coordinate the provision of cloud-based services with multi-granular data delivery services across flexible optical network. Furthermore, the work in [18] provides a system, *Pulsar*, to give tenants the abstraction of a VDC that affords them the performance stability of an in-house cluster, and the convenience and elasticity of the shared cloud. Pulsar uses a centralized controller to enforce end-to-end throughput guarantees that span multiple appliances and the network.

Different from the work in [73], other works in [76] and [74] focus on mapping all the VDC components within the same data center. The key contribution of [76] is to design virtual clusters as the virtual network abstractions that capture the trade-off between the performance guarantees offered to users, their costs and the provider revenue. The work in [74] proposes VDC Planner, a migration-aware dynamic virtual data center embedding framework that aims at achieving high revenue while minimizing the total energy cost over-time. The proposed framework supports various usage scenarios, including VDC embedding, VDC scaling as well as dynamic VDC consolidation. In our works, we may investigate the VDC mapping problem in distributed data centers and within a single data center while considering the role of optical networks in the Cloud systems.

Furthermore, another work [77] that related with VDC embedding focuses on the virtual infrastructure embedding with reliability guarantee. The reliability is realized through redundant nodes and links. A pooling mechanism opportunistic redundancy pooling (ORP) is introduced to share the redundancies for both independent and cascading types of failures.

No matter for the joint resource allocation problem or for the VDC mapping problem, the network resource virtualization, especially of optical links, plays an essential role in offering elasticity in terms of DC-to-DC data paths and enabling the dynamic allocation of slices of network bandwidth between physical servers in different DCs [78]. A more recent work [79] investigates the problem of joint defragmentation (DF) for the spectrum and IT resources in elastic optical data center interconnections (EO-DCIs). Specifically, in order to reduce the blocking probability in an EO-DCI, the authors re-optimize the allocations of the multidimensional resources jointly with complexity-controlled network reconfigurations. In addition, the work in [78] presents a distributed management platform, namely, the network virtualization management platform (NVMP) for latency aware applications. The anycast-based optimizations are proposed to optimally select the target IT resources and also consider the data transfer performance across the DCs. Different policies are proposed and evaluated that select an inter-DC network path, and accordingly a destination server, so that the VM data transfer can experience the proper delay performance.

Chapter 3

Provider's Viewpoint: Cost-Optimized Resource Allocation in Grids/Clouds with Multilayer Optical Network

3.1 Introduction

In the Cloud network model, resources will be provided according to user's requirements which usually lead to cost reduction. The investigation of solutions to cope with network-aware joint resource allocation is a very important topic in the field of Grids/Clouds in the next five years. We first describe the related work on the Grid/Cloud resource allocation area and then describe the optical network structure which will be adopted in our work. The optimal resource scheduling has also been a great challenge in IaaS Cloud environment and various investigations have been conducted in this area.

The authors of work [53] consider resource allocation algorithms for distributed Cloud systems and develop algorithms for network-aware allocation of virtual machines to achieve good application performance. The objective of this work is to minimize the maximum distance or latency between the selected data centers with the proposed data-center selection algorithms for VM placement. Nowadays, the resource provisioning in the Cloud such that the performance is maximized and the financial cost is minimized is still a challenge in the Cloud environment, and hence many studies have investigated this problem. The work in [80] designs, implements

and evaluates two auto-scaling solutions to minimize the job turnaround time within the budget constraints for Cloud work flows to reach the goal of maximizing the return from the Cloud investment. The work in [81] studies the optimization problem of minimizing resource rental cost for running elastic applications in Cloud while meeting application service requirements. A deterministic resource rental planning (DRRP) model and a stochastic resource rental planning (SRRP) model which considers the price uncertainty, are proposed to generate optimal rental decisions. The study in [82] adopts dynamic capacity provisioning to reduce the energy consumption by dynamically adjusting the number of active machines to match resource demands. A heterogeneity-aware resource management system (HARMONY) is presented for the dynamic provisioning that can strike a balance between energy savings and scheduling delay, while considering the reconfiguration cost.

In this chapter we consider the joint scheduling of processor, storage and network resources in Grid/Cloud networks from the Cloud provider's point of view, while considering guaranteed network bandwidth for inter data center connection. Given the inputs, which are shown in Fig. 3.1, the resource allocator needs to check the real-time resource status in the Cloud and achieve reasonable resource allocation for as many consumers as possible. The objective is to minimize the total capital expenditures (CapEx) for the resource allocation, which include the cost of the network components and initial facility installation costs. We introduce the multi-layer optical network architecture to deal with the guaranteed network bandwidth problem during joint resource allocation. To solve the joint resource allocation, we construct a Mixed Integer Linear Programming (MILP) model with the integer constraints to obtain the optimal solutions and propose two polynomial-time heuristic algorithms.

The rest of this chapter is organized as follows. Section 3.2 describes the problem settings, network model and cost model for the problem. Section 3.3 and 3.4 describe

the MILP formulations and propose heuristics. Section 3.5 describes the experimental results and analysis. Finally, we conclude the work in Section 3.6.

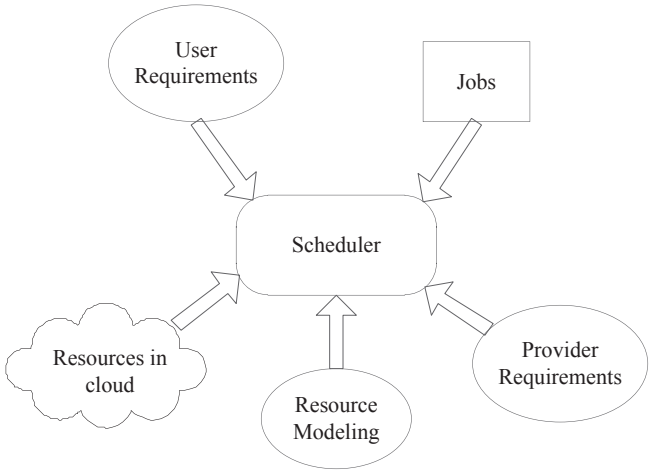


Figure 3.1: Resource allocation inputs.

3.2 Joint Resource Allocation Problem

3.2.1 Problem Description

The Cloud network consists of geographically connected data centers. Based on the user constraints and resource availability, a user request might not obtain all resources from a single data center sometimes. Generally, in a Cloud environment, the resource assignment service for a user request can be divided into several steps. First, we identify the possible candidate data center sites for the request. Second, we select the right data centers that would assign resources for the request. Third, we fix a certain rack in the data center for the request. The last step is to determine the specific VMs in the rack for the request to complete the resource assignment. In this work, we do not consider the details of resource assignment within a data center such as the rack determination and VM placement, but only consider the data center selection and the inter-data center network communications.

A user submits a job request with fixed resource requirements and job execution information to the resource allocator in the Grid/Cloud networks. The job might need certain types of processors and certain capacity storage for execution and certain amount of bandwidth for data transmission as well. To guarantee the bandwidth reserved for data transmission, there's a need to establish a circuit in the optical layer of the network. Several sub-wavelength channels with specific bandwidth are combined together in one single optical fiber circuit. And OTs mark the end points of each WDM sub-wavelength channel in the optical circuit. On the other hand, the centralized resource allocator of the Cloud network has an overall view of the resource status, and maintains the real-time updates on the resources across the whole Grid/Cloud network. The resource allocator will complete the resource scheduling according to the requirements from the submitted jobs and the current resource situation in the Cloud.

3.2.2 Problem Assumptions

We investigate the optimal joint resource allocation in this work from the provider's viewpoint to minimize the total CapEx of resource allocation while considering the optical transport layer as multi-layer architecture. We assume that the transport network adopts the IP/MPLS-over-OTN-over-WDM architecture as shown in Fig. 3.4. To simplify our model and also to ensure its reasonableness for realistic co-scheduling in Grid/Cloud, we make the following assumptions.

- One node in the network topology represents one data center in the Cloud network. Each node has different computational (processor and storage) capacities and different amount of related optical facilities.

- Each link in the topology is bidirectional and has the same bandwidth capacity.

The bandwidth on each link is divided into several sub-wavelengths with equal bandwidth.

- Execution cycle (12 hours), noted as S, is slotted. Each time slot is 1 hour, noted as s. Jobs should be completed within one execution cycle.
- Jobs arrive at the resource allocator one by one and are collected first, then the allocator will schedule them together (batch processing).
- A job consists of a series of dependent or independent tasks. Independent tasks in one job can be executed in parallel, while dependent tasks must be executed sequentially. Fig. 4.2 gives us a visual sense of a job structure. The task t_C and t_D of job $j2$ in Fig. 4.2 are independent from each other, and they both depend on task t_B. Here we note t_B as *parent* task, t_C and t_D as *child* task of t_B.
- The processor and storage resources assigned to one task must be from the same node.
- Task computing (for processor, storage resource) and data transmission (for OT, bandwidth resource) happen synchronously, which means the task will transmit the intermediate data or results right after their generation by task executing.
- Resources reserved by a task will be released once the execution of this task is completed.
- OTs, used as the end points of a sub-wavelength channel along the optical link, must appear in pairs; we call this as transponder mapping.

Here we give an example to illustrate the transponder mapping assumption mentioned above. Parent task t_A and its child task t_B are allocated in nodes $N1$ and $N4$ respectively, and task t_A needs to transmit data to task t_B. Task t_A requires two OTs

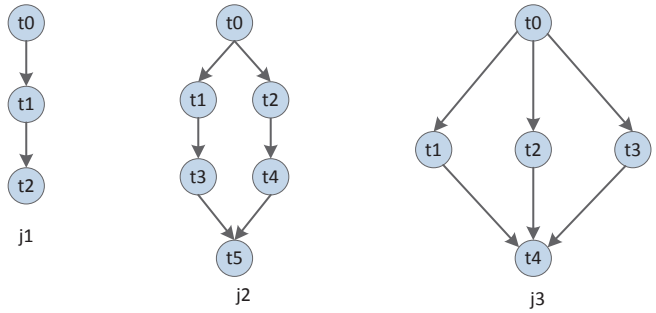


Figure 3.2: Examples of supported job structures.

used for transmission, and they are assigned from node $N1$. Therefore, we need to assign two OTs from node $N4$ for task t_B as the mapping OTs to receive the data from task t_A. As a result, two wavelength channels are established as shown in Fig. 3.3.

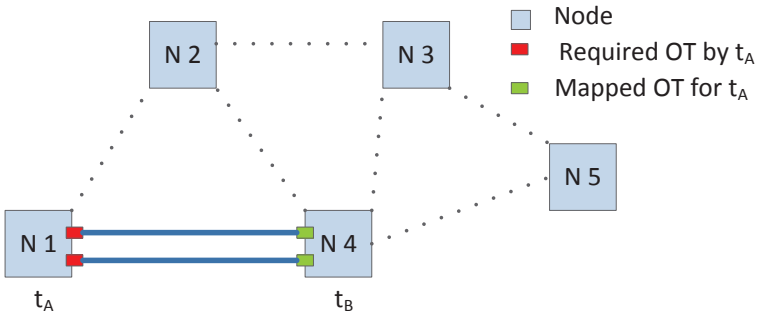


Figure 3.3: Optical transponder mapping.

With the above assumptions, the resource allocator will complete the resource assignment in the Cloud for the user requests. How to make use of the limited resources in the Grid/Cloud to realize an optimal joint resource scheduling for as many consumers as possible is important to both consumers and the Cloud providers.

Intense competition in the Cloud computing service market imposes high pressures on the data center and network operators' revenues. This situation underscores the need to maintain data center and network communication costs under control. It is essential to minimize the total cost of the data centers and the network in the

Cloud environment while supplying services at competitive prices. Therefore, from the provider's point of view, how to complete the joint resource allocation to satisfy the consumers' requirements at a minimal cost is a significant issue, which is the way to increase the profit of Cloud computing service. The cost here is the CapEx of data center and telecommunication network operators. CapEx includes the cost for building the Cloud computing environment infrastructures, such as the construction costs of data centers and network installation costs. In this work, the objective of the resource allocator is to complete the resource scheduling for the request with minimal CapEx.

As mentioned earlier, we only focus on the bandwidth requirement for inter-data center communication when considering the network resource part. In the current Cloud computing service, the network bandwidth resource offered by the providers is not guaranteed. Thus the best-effort data transmission for the tenants is unreliable, which might lead to unpredictable data loss and long delay. We intend to construct an optical route in the Cloud network to transmit data for tenants. Each user can reserve one or several sub-wavelengths with certain amount of bandwidth to transmit data along the optical route. The IP/MPLS-over-OTN-over-WDM multi-layer network architecture is introduced while achieving the routing and bandwidth assignment for guaranteed network resource allocation. We adopt the transparent implementation in the multi-layer network, in which ROADM is used to bypass particular wavelengths at intermediate node in an optical route. In this case, the optical signal can avoid the O-E-O conversion during data transmission between source and destination.

3.2.3 Network Model

In the distributed Cloud environment, it is normal that users will obtain the needed resources from several data centers. In this case, data transmission is necessary

between data centers in order to complete the whole job execution. In the Cloud network, the same set of routers and links are deployed to carry traffic for all customers with no difference. The Cloud providers usually do not supply guaranteed network resources for the tenants. Therefore, the bandwidth for the user to transmit data might vary significantly according to the network load. Thus offering guaranteed network bandwidth for the tenants together with other Cloud resources is critical for Cloud operators. We can utilize the optical network architecture to set up reliable circuit for bandwidth resource reservation. Next we will describe the multi-layer optical network architecture.

Optical multi-layer networks offer a high degree of freedom in network design, adapting to actual network requirements and achieving cost-efficient realizations [22]. The lower layer technologies such as Layer 2 switching and Layer 1 optical networks, with the advantage of high flexibility and agility, are far from their intrinsic physical limits. Therefore considering deploying optical multi-layer network to offer bandwidth guaranteed data transmission in the Cloud at a lower cost would be a viable scheme.

In [22], a detailed CapEx model is given for optical multi-layer networks, which including four layers: Internet protocol/multi-protocol label switching (IP/MPLS), carrier-grade Ethernet, optical transport network (OTN) and wavelength division multiplexing (WDM). All equipment costs discussed in each layer are relative costs that are normalized to the cost of a 10 Gbits/s transponder with a transparent reach of 750 km. Based on the CapEx work presented in [22], we consider the IP/MPLS-over-OTN-over-WDM layered network architecture for our joint resource scheduling problem. Figure 3.4 shows the IP/MPLS-over-OTN-over-WDM multi-layer network structure. We have introduced each layer in Chapter **??** Section 1.3.

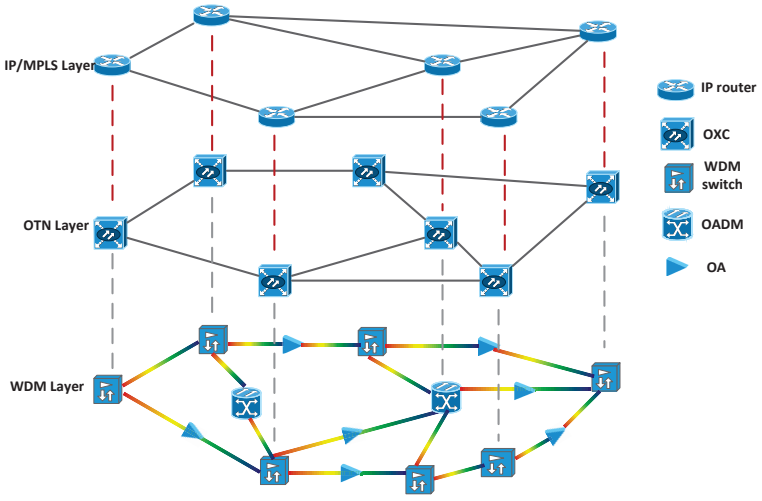


Figure 3.4: IP/MPLS-over-OTN-over-WDM layered network architecture.

3.2.4 Cost Model

In our cost model, the CapEx for joint resource scheduling in the Grid/Cloud environment includes the cost of data center resources (such as processor, storage and bandwidth) and also the cost of equipments in the multi-layer network architecture. According to the analysis of costs in the Cloud [83], the costs in a data center mainly go to the servers, infrastructure, power draw and network. We focus on the costs of the server component since this component takes the greatest part of the total CapEx in a data center. For example, as the per rack lifetime cost for different types of data center infrastructure shown in [84], the per rack lifetime cost is around $70,000 for legacy architecture type. Usually 20 servers are located within one rack and over 90% of the capital cost is typically spent in year 1 (the data center life cycle is 10 years in general) using legacy design approaches [83]. In this case, we probably estimate the capital cost of each server in one operation hour is $0.36, which can be utilized as the data center resources cost part of the total CapEx considered in our problem.

The cost model of the equipment in the multi-layer network architecture in our

problem relies on the multi-layer cost model proposed in [22].

3.2.4.1 Cost model for IP/MPLS layer

For the IP/MPLS layer of the multi-layer network architecture, we involve the IP/MPLS router cost as the CapEx. The IP/MPLS router model is divided into two main blocks: the basic node and the equipment related to the physical layer interfaces. Each basic node supplies a limited number of bidirectional slots with a certain bandwidth for physical layer interfaces. In our problem we assume the slot capacity is 40 Gbits/s in IP router. Each slot is equipped with a slot card that has the capability to connect different types of line cards. Usually the interface cards will be equipped with separate pluggable optical transceiver modules, but for simplicity the cost for optics in our model are aggregated into the line card costs as stated in [22]. The normalized cost values for IP/MPLS network equipment are shown in Table 3.1.

Table 3.1: Normalized cost for IP/MPLS layer equipments.

IP/MPLS router basic nodes		
Capacity	Number of slots (slot capacity = 40 Gbits/s)	Cost
640 Gbits/s	16 slots	16.67
IP/MPLS router slot card		
40 Gbits/s	1 slot/1 slot	9.17
IP/MPLS router port card		
Interface type	Number of slots occupied (slot capacity = 40 Gbits/s)	Cost
4 × PoS STM-16, SR (1310 nm, 2 km reach)	1/4 slot	5.83

3.2.4.2 Cost model for OTN layer

The OTN layer provides the multiplexing and transmission functionalities which grooms TDM signals of distinct granularities within a multiplexing hierarchy. The CapEx of the OTN layer goes to the related switching elements such as OTN electrical cross connects (EXCs) and related interfaces. In our model, we also assume that when an EXC connects to the WDM layer, the method of using short reach interfaces and separate WDM transponders is adopted. The normalized cost values for OTN layer equipment are shown in Table 3.2.

Table 3.2: Normalized cost for OTN layer equipments.

OTN EXC basic nodes		
Capacity	Number of slots (slot capacity = 40 Gbits/s)	Cost
640 Gbits/s	16 slots	13.33
OTN EXC line cards		
Interface type	Number of slots occupied (slot capacity = 40 Gbits/s)	Cost
Gray interface STM-16/ODU1, SR (1310 nm, 2 km reach)	1/16 slot	0.25

3.2.4.3 Cost model for WDM layer

In the WDM layer, transponders, muxponders, WDM multiplexer/demultiplexer terminals, optical amplifiers, regenerators as well as OADM are used in a classical WDM transmission link to achieve transparent optical switching. The capability of each component in a WDM link is declared in Sec. II.B. In our cost model, we assume that the bandwidth of each sub-wavelength in the WDM link is 10 Gbps, and each link has 40 sub-wavelength channels. In addition, we assume that the optical reach

is 750 km. The normalized cost values for related WDM layer components are listed in Table 3.3. The parameter N in the table is the WDM node degree, for which $2 < N \leq 5$.

Table 3.3: Normalized cost for WDM layer equipments.

Component type	Cost
WDM transponders	
10G, LH (750 km reach)	1.00
WDM muxponders	
10G muxponder (2.5G×4), LH (750 km)	1.17
Regenerators (3R)	
10G, LH(750 km reach)	1.40
Optical Line Amplifier (OLA)	
OLA, LH(80 km reach)	1.92
WDM terminals, including booster/receiver amplifier	
40 channel, (LH)	4.17
OADM, including internal amplification stages	
Fixed OADM, 50%, 40 channel system	3.35
OXC, including internal amplification stages	
OXC, N degree, 100%, 40 channel system	$8.33 \times N + 2.5$

As we know that the cost of a component at a specific time needs to be derived using a method that also considers the price variation over time and takes current market into account [24]. A method to model the price variation is described in [85]. The cost values we used in this chapter could be changed in the future.

3.3 MILP Formulation for Optimal Joint Resource Allocation

MILP formulations are developed to complete the optimal joint resource allocation for the requests. Three types of inputs are offered for the MILP formulations. Firstly, the resource modeling contains network topology which indicates the node and link information, and the resource information on each node/link. Secondly, the requests from tenants include submitted jobs which indicate budget, start/finish time and other requirement information. Thirdly, the current traffic in the network contains the information of the current resources consumption on each node and link. In the following, we will discuss the inputs, the constraints and related parameters defined in the MILP formulations.

3.3.1 Inputs of the Model

The resource modeling input involves the multi-layer network architecture, the processor and storage resources on the IP/MPLS layer nodes as well as the bandwidth resource on WDM layer links. In the IP-over-OTN-over-WDM multi-layer network, node/link information of corresponding layer is given. IP/MPLS layer: $Node_i = (n_i, P^{n_i}, D^{n_i}, cp^{n_i}, cd^{n_i}, \alpha^{n_i})$; OTN layer: $Node_o = (n_o, \beta^{n_o})$; WDM layer: $Node_w = (n_w, OT^{n_w}, cot^{n_w}, \gamma^{n_w})$ and $Link_w = (l_w, src^{l_w}, des^{l_w}, Len^{l_w}, cl^{l_w}, B^{l_w}, cb^{l_w})$. We also calculate the shortest path for every pair of nodes in the topology in the pre-processing, and set these shortest paths as one of the inputs. $Path = (psrc, pdes, links\_on\_path, linknum\_on\_path)$, in which $psrc, pdes$ represent the start/end node of the path; $links\_on\_path$ is a set of links that consist the current path; $linknum\_on\_path$ represents how many links are on this path.

The request input involves the jobs submitted by users and the tasks that form a job, noted as $Job = (j, STime_j, FTime_j, Bud_j)$ and $Task = (t, j, tSTime_j^t, tFTime_j^t,$

$RP^s_{jt}, RD^s_{jt}, ROT^s_{jt}, RB^s_{jt}, CID, PID$).

The current traffic inputs consist of the current traffic on nodes and links in the related layer. For IP layer: $currNode_i = (n_i, s, oP^s_{n_i}, oD^s_{n_i})$. For WDM layer: $currNode_w = (n_w, s, oOT^s_{n_w})$ and $currLink_w = (l_w, s, oB^s_{l_w})$.

The detailed parameter information for the inputs described above are listed in the Table 3.4. In addition, other related notations and variables we used in the MILP formulations, are listed in Table 3.5 and 3.6.

Table 3.4: Parameters for Inputs

$P^{n_i}, D^{n_i}, OT^{n_w}, B^{l_w}$	processor, storage, transponder, bandwidth resource capacities on the nodes and links in the corresponding layer
$cp^{n_i}, cd^{n_i}, cot^{n_w}$	unit cost of processor, storage and transponder resource on the nodes in the corresponding layer
cl^{l_w}	link cost per mile which integrates the OA, muxponder.regenerator cost
cb^{l_w}	the unit cost of bandwidth
src^{l_w}, des^{l_w}	source and destination node of link l_w
Len^{l_w}	WDM link length measured by mileage
$\alpha^{n_i}, \beta^{n_o}, \gamma^{n_w}$	unit cost of IP/MPLS, OTN, WDM node terminals
$STime_j, FTime_j$	start/finish time of job j
Bud_j	executing budget of job j
$tSTime^t_j, tFTime^t_j$	start, finish time of task t in job j
$RP^s_{jt}, RD^s_{jt}, ROT^s_{jt}, RB^s_{jt}$	required amount of processor, storage, transponder, bandwidth resources by task t in job j
CID, PID	the child/parent task of task t
$oP^s_{n_i}, oD^s_{n_i}, oOT^s_{n_w}, oB^s_{l_w}$	occupied processor, storage, transponder, bandwidth resources on the nodes and links in the corresponding layer

Table 3.5: Other Constant Parameters

J	set of jobs, $j \in J$
T_j	set of tasks that belongs to job j, $t \in T_j$
N_i	set of IP/MPLS layer nodes in the network topology, $n_i \in N_i$
N_o	set of OTN layer nodes in the network topology, $n_o \in N_o$
N_w	set of WDM layer nodes in the network topology, $n_w \in N_w$
L_w	set of WDM layer links, $l_w \in L_w$
s	one time slot
S	an executing cycle, consisting of certain number of time slots, indicated by scheduler

Table 3.6: Variables

Dep^j_{ik}	binary parameter, equals to 1 if task k is dependent on task i, both are belonged to job j; 0 otherwise
$X^s_{j(i,k)}$	the number of mapped transponders between parent task i and its child task k of job j in time slot s
Cap_j	total CapEx for executing job j
$C^{IP}_j, C^{OTN}_j, C^{WDM}_j$	the CapEx in IP layer, OTN layer and WDM layer for job j
$Drop_j$	binary parameter, equals to 1 if job j cannot be scheduled; zero otherwise
$F^{n_i}_{jt}, F^{n_o}_{jt}, F^{n_w}_{jt}$	binary parameter, equals to 1 if task t of job j are assigned node n_i, n_o and n_w in the corresponding layer; 0 otherwise
$P^{l_w s}_{(t,k)j}$	binary parameter, equals to 1 if link l_w on the path from task t to k of job j in time slot s; 0 otherwise

3.3.2 Objective and Constraints

The objective from the Cloud provider's point of view is to minimize the total CapEx cost for providers to execute all submitted requests that can be scheduled in the

Cloud environment with the transparent IP-over-OTN-over-WDM multi-layer network architecture. This is a way, as a result, for the providers to reach the target of maximizing the total profit.

Objective:

$$Minimize \sum_{j \in J} Cap_j \tag{3.1}$$

$$Cap_j = C_j^{IP} + C_j^{OTN} + C_j^{WDM} \tag{3.2}$$

$$\begin{aligned} C_j^{IP} = & \sum_{t \in T_j, n_i \in N_i} F_{jt}^{n_i} * RP_{jt}^{s} * cp^{n_i} * Dur_j^t \\ & + \sum_{t \in T_j, n_i \in N_i} F_{jt}^{n_i} * RD_{jt}^{s} * cd^{n_i} * Dur_j^t \\ & + \sum_{t \in T_j, n_i \in N_i} \alpha^{n_i} * Dur_j^t * F_{jt}^{n_i} \end{aligned} \tag{3.3}$$

$$C_j^{OTN} = \sum_{t \in T_j, n_o \in N_o} \beta^{n_o} * Dur_j^t * F_{jt}^{n_o} \tag{3.4}$$

$$
\begin{aligned}
C_j^{WDM} = & \sum_{t\in T_j, n_w\in N_w} F_{jt}^{n_w} * ROT_{jt}^{s} * cot^{n_w} * Dur_j^t \\
& + \sum_{i,k\in T_j, n_w\in N_w} F_{jk}^{n_w} * X_{j(i,k)}^{s} * cot^{n_w} * Dur_j^i \\
& + \sum_{k,t\in T_j, l_w\in L_w} Len^{l_w} * cl^{l_w} * P_{(t,k)j}^{l_w s} * Dur_j^t \\
& + \sum_{t\in T_j, t\neq t.CID, l_w\in L_w} P_{(t,k)j}^{l_w s} * RB_{jt}^{s} * cb^{l_w} * Dur_j^t \\
& + \sum_{t\in T_j, t=t.CID, l_w\in L_w} RB_{jt}^{s} * cb_{egress} * Dur_j^t \\
& + \sum_{t\in T_j, n_w\in N_w} \gamma^{n_w} * Dur_j^t * F_{jt}^{n_w}
\end{aligned} \tag{3.5}
$$

where $\forall j \in J, Dur_j^t = tFTime_j^t - tSTime_j^t + 1$.

Time Constraints:

$$tFTime_j^t \leq S, \forall j \in J, t \in T_j. \tag{3.6}$$

$$FTime_j - STime_j + 1 > 0, \forall j \in J. \tag{3.7}$$

Resource Assignment Constraints:

$$\sum_{n_i\in N_i} F_{jt}^{n_i} = 1,\ \forall j \in J, t \in T_j \tag{3.8}$$

$$\sum_{n_o\in N_o} F_{jt}^{n_o} = 1,\ \forall j \in J, t \in T_j \tag{3.9}$$

$$\sum_{n_w\in N_w} F_{jt}^{n_w} = 1,\ \forall j \in J, t \in T_j \tag{3.10}$$

Transponder Mapping Constraints:

$$X_{j(i,k)}^{s} = ROT_{ji}^{s},\ \forall i,k \in T_j, j \in J, m_w \in N_w, s \in [tSTime_j^i, tFTime_j^i] \tag{3.11}$$

Resource Capacity Constraints:

$$\sum_{j\in J,t\in T_j} RP^s_{jt} * F^{n_i}_{jt} \leq P^{n_i} - oP^s_{n_i} \tag{3.12}$$

$$\sum_{j\in J,t\in T_j} RD^s_{jt} * F^{n_i}_{jt} \leq D^{n_i} - oD^s_{n_i} \tag{3.13}$$

$$\sum_{j\in J,t\in T_j} ROT^s_{jt} * F^{n_w}_{jt} + \sum_{j\in J,i,k\in T_j,i\neq k} X^s_{j(i,k)} * F^{n_w}_{jk} \leq OT^{n_w} - oOT^s_{n_w} \tag{3.14}$$

$$\sum_{j\in J,t,k\in T_j} RB^s_{jt} * P^{l_w s}_{(t,k)j} \leq B^{l_w} - oB^s_{l_w} \tag{3.15}$$

where $\forall n_i \in N_i, n_w \in N_w, l_w \in L_w, s \in S$.

Budget Constraints:

$$E_j \leq Bud_j,\ \forall j \in J \tag{3.16}$$

Equation 5.2 states that the total CapEx consists of the CapEx in IP/MPLS layer, OTN layer and WDM layer. Equation 5.3 states that the IP/MPLS layer CapEx includes the costs of processor, storage resources and IP node terminals. Equation 5.4 states that the OTN layer CapEx involves the capital cost of OTN node terminals in the transmission path. Equation 5.5 states that the WDM layer CapEx includes the costs of required optical transponders, mapped optical transponders, bandwidth resource, physical links, and WDM node terminals in the transmission path. Equation 5.13 ensures that each task should complete the execution before the end of the execution cycle. Equation 5.14 guarantees that each job execution time is greater than

zero. Equations 3.8–3.10 guarantee that each task is assigned the required resources by the resource allocator from the same node in each layer. Equation 3.11 guarantees that in the WDM layer the mapped transponder of parent task i will be allocated from the node selected by i's child task k for the whole task i's duration. Equations 3.12–3.15 guarantee that in each time slot s, the cumulative occupied resources by jobs cannot exceed the corresponding available resource capacity on each node/link in the corresponding layer. Equation 3.16 bounds the total payment of each job to the budget given by users.

The MILP formulations can be solved with IBM CPLEX optimization software [86], from which an optimal joint resource scheduling solution is reached for the consumers and Cloud providers. However solving the MILP is a time consuming task, so we develop heuristics to solve the problem as described below.

3.4 Heuristics for Optimal Joint Resource Allocation

Our time-efficient heuristic algorithms solve the joint resource scheduling problem in the Grid/Cloud environment. Given a series of submitted requests, Cloud resource information, and current traffic in the Cloud network, the target of the resource allocator is to complete the optimal resource allocation according to the objectives. To complete such a resource co-allocation for the submitted jobs, we need to consider the job scheduling first, and then realize the resource allocations in the Cloud network for each job.

3.4.1 Job Scheduling

The jobs collected by the resource allocator are scheduled sequentially. Different scheduling orders of the jobs may affect the optimized solution. To investigate the

effect of job scheduling on optimal joint resource allocation, we carry out the experiments with several scheduling policies.

- First come first serve (FCFS). Jobs are scheduled according to their arrival order.
- Shortest job execution time first (STF). The job which occupies the resources for a shorter time is scheduled first.
- Random schedule (Random). Submitted jobs in the queue are scheduled in a random order.
- Early start time job first (ESTF). The job which starts executing earlier is scheduled first.
- Simple job structure first (SSF). The job consisting of fewer sub-tasks is seen as having a simple job structure, and is scheduled first.

3.4.2 Resource Co-allocation

To achieve the target of achieving optimal resource allocation for each job, two heuristics are proposed in this chapter.

3.4.2.1 Best-Fit Heuristic

The Best-Fit heuristic we proposed is a greedy algorithm. The basic idea of Best-Fit is to choose the node with available resources and with lowest resource unit cost for each task in a job. In addition, we would like to allocate resources for tasks of a job from one node or several nodes that are near from each other, to reduce the network cost as much as possible. Based on these ideas the Best-Fit heuristic comprises of two steps: the allocation for computational resources and that of network resources.

For each task in a job, the data center nodes which have available resources to assign processor, storage and transponder resources and have the lowest resource unit cost are selected first. Then, with the chosen nodes for each task, paths between related nodes are set up to allocate bandwidth for data transmission between tasks. Each pair of transponders is used for setting up one sub-wavelength route on the path. The dependency between tasks of a job must be considered as well when allocating resources for them, so that a child task cannot be allocated before the completion of its parent task. The Best-Fit heuristic is shown in **Algorithm 1** and **Algorithm 2**.

In the line 5 of Algorithm 2, we adopt Dijkstra's algorithm to find the pair of shortest paths, which has the time complexity of $O(N_w^2)$. The time complexity of Algorithm 2 is $O(JT^2N_w^2)$. Thus the total time complexity of Best-Fit heuristic is $O(J(T^2N_w^2 + T(N_i + N_w)))$.

3.4.2.2 Tabu Search Based Heuristic

Tabu-search is a "high-level" meta-heuristic procedure for solving optimization problems, designed to guide other methods to escape the trap of local optimality [87], and has been applied to solve resource allocation and other optimization problems. In the proposed Best-Fit algorithm, it is obviously that an optimal solution will be found for a small set of input jobs. For the larger set of input jobs, it will lead to quite high total cost for the latter scheduled jobs in the set since data center nodes with lower resource unit cost have no resource available. In this case, we try to develop Tabu search based method to solve our optimization problem with the hope of obtaining better solutions and improving the traffic blocking rate for the submitted requests.

Based on the study of basic idea of Tabu search, we need to pay attention to several key points in the design of the Tabu search based heuristic, such as initial solution, neighborhoods generation, aspiration satisfaction and termination condition. The

Algorithm 1 Best-Fit Algorithm

Input and Initializations:

Topology information
Current traffics on node/link in each layer
$J = j_1, j_2, ..., j_M$; //set of jobs
$T_j = t_1, ..., t_k$, $j \in J$; //set of tasks belong to job j
$Cap_j = 0$; //initial cost is 0

Output:

Minimize Cap_j, $\forall j \in J$.

update current available resources in the network;
Determine jobs scheduling order according to methods: FCFS, STF, Random, ESTF, SSF;
for $j = 0$; $j < J$, $j++$ **do**
  //computational resources allocation for tasks of job j
  for $t = 0$; $t < T_j$; $t++$ **do**
    if t has parent **then**
      if parent is done **then**
        for $n_i = 0, n_w = 0$; $n_i < N_i, n_w < N_w$; n_i++, n_w++ **do**
          Check resources on IP and WDM layer node for t with minimum cost;
        end for
        if no node available for t **then**
          $Drop_j = 1$, go to next job;
        else
          Update resource on selected node n_i, n_w;
          Update final minimum cost for t, next task;
        end if
      else
        Wait for parent task is done;
      end if
    else
      Use same resource allocation step as above;
    end if
  end for
  if $Drop_j = 0$ **then**
    Bandwidth resource allocation for current job
  end if
  Update total expense Cap_j of current job j;
end for
return Cap_j;

Algorithm 2 Bandwidth resource allocation

```
1: for j = 0; j < J, j++ do
2:   for t = 0; t < T_j;t++ do
3:     for k = 0; k < T_j; k++ do
4:       if t is the parent of k then
5:         Find shortest path from t to k;
6:         if bandwidth available on the path then
7:           Assign bandwidth for t;
8:           Update bandwidth resource of every link on path;
9:           Compute bandwidth expense for task t;
10:        end if
11:      end if
12:    end for
13:  end for
14:  Compute C_j^link;
15: end for
```

pseudo code of the proposed Tabu search based heuristic is shown in **Algorithm 3**. In the Tabu search based heuristic, the procedure of generating solution pool (line 5 in Algorithm 3) is similar with the Best-Fit heuristic. The time complexity of Tabu search heuristic is $O(J(T^2N_w^2+T(N_i+N_w))+K)$, where K is the loop count indicated in the Tabu search termination condition.

3.5 Experimental Results and Analysis

The experiments of our joint resource scheduling problem for both the MILP model and heuristics are carried out with several network topologies. For the MILP models, the IBM OPL CPLEX Optimization Studio is adopted to complete the experiments. The optimized solutions are acquired (when possible) using OPL Optimization first and will be compared with the solutions acquired using heuristic methods. The experiments are carried out respectively on a simple 6-node mesh topology as shown in Fig. 3.5 and a 20-node topology of GCE data center locations as shown in Fig. 3.6.

Algorithm 3 Tabu Search Based Algorithm

Input and Initializations:

Topology information
Current traffic on node/link in each layer
$J = j_1, j_2, ..., j_M$;//Input job requests
$T_j = t_1, ..., t_k$, $j \in J$; //set of tasks belong to job j
$Cap_j = 0$; //initial job cost is 0

Output:

Minimize $\sum_{j \in J} Cap_j$, $\forall j \in J$.

```
Update current available resources in the network;
Select job scheduling policy from: FCFS, STF, Random, ESTF, SSF;
InitialSol := solution by Best-Fit algorithm;
OptSol := InitialSol; //set optimal solution
Generate solutions pool;
Set Tabulist;
push OptSol into Tabulist;
while not-termination conditions do
  Random move to generate neighbor solution: Neighbor;
  if Neighbor ∈ Tabulist then
    Move operation, generate new neighbor;
  else
    CurrSol := Neighbor; //set current solution
    if CurrSol < OptSol then
      OptSol := CurrSol;
      if Tabulist is full then
        Pop out the oldest element in the list;
        Push OptSol into Tabulist;
        Update Tabulist;
        Continue; //move on search
      else
        Push OptSol into Tabulist;
        Update Tabulist;
        Continue;
      end if
    else
      Move operation, generate new neighbor;
    end if
  end if
end while
Return OptSol;
```

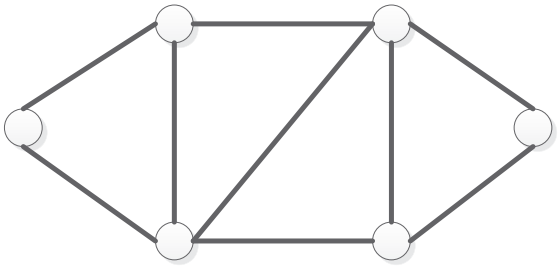


Figure 3.5: The 6-node mesh topology.

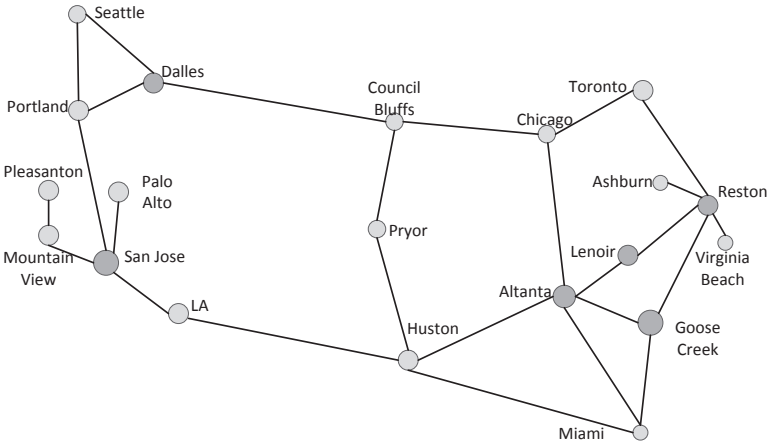


Figure 3.6: GCE data center distribution topology constructed from public information on data center locations.

The optimal solutions for our problem can be generated by solving the MILP formulations with the CPLEX Optimization software. However, the problem solving process using CPLEX is time consuming particularly with the increasing size of the tested network topology and submitted requests. In our experiment, it takes more than one hour to solve the optimal resource allocation for 40 input jobs on a 6-node network topology. Hence we only use the proposed heuristic methods to conduct the experiments for the larger 20-node GCE topology.

We first test the OPL model and proposed Best-Fit, Tabu search heuristics on the 6-node network topology to verify the consistency of MILP and heuristic solutions. Table 3.7 compares the CapEx obtained by OPL, Best-Fit heuristic and Tabu search

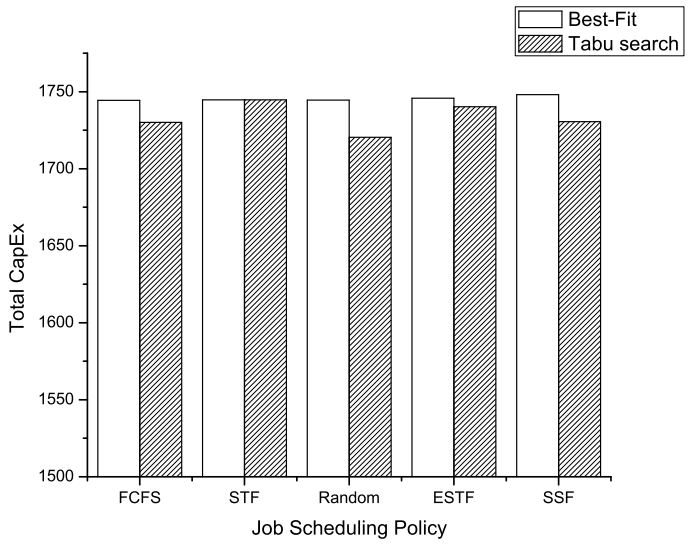


Figure 3.7: CapEx comparison for 10 input jobs on GCE topology.

based heuristic on a 6-node network topology when given different submissions. The data in the table show that the solutions generated by Best-Fit and Tabu search heuristics are the same when given different number of requests under different job scheduling policies on a 6-node network topology. In addition, the comparison shows that the solutions of both heuristics are very close to the optimal solutions obtained by OPL. For the GCE topology, the total CapEx obtained by Tabu search is better than that of the Best-Fit method under different job scheduling policies as shown in Fig. 3.7, but the improvement is not so significant. Table 3.8 compares the running time of OPL, Best-Fit and Tabu search methods with different number of submissions on the same topology. For each heuristic, we use the average running time of all tested job scheduling policies. The comparison indicates that the heuristics are much more time-efficient than the OPL method while generating the optimal solutions as the OPL. Thus our analysis conducted on a larger GCE topology will be carried out using only the heuristics.

We observe the total CapEx obtained by Best-Fit and Tabu search based methods

Table 3.7
CapEx ($\times 10^3$) comparisons between OPL and two proposed heuristics with different job scheduling policies on a 6-node topology.

J	OPL	Both Best-Fit & Tabu search				
		FCFS	STF	Random	ESTF	SSF
5	0.778	0.778	0.778	0.778	0.778	0.778
10	1.679	1.682	1.683	1.681	1.681	1.683
20	2.717	2.723	2.725	2.721	2.722	2.721
30	3.602	3.642	3.611	3.609	3.643	3.605
40	4.543	4.608	4.567	4.546	4.609	4.556

Table 3.8
Running time (seconds) comparisons between OPL and two proposed heuristics with different job scheduling policies on a 6-node topology.

J	OPL	Best-Fit	Tabu search
5	1.502×10^2	0.000	0.000
10	1.192×10^3	0.000	0.000
20	1.941×10^3	0.000	0.050
30	3.001×10^3	0.005	0.660
40	4.631×10^3	0.020	1.080

under different job scheduling policies on the GCE topology. We found that when the size of input sets increasing (larger than 40), the Best-Fit and Tabu search methods under FCFS and Random job scheduling policies can acquire lower CapEx compared to other job scheduling policies.

With the increase in traffic load, the Cloud network topologies with limited resources cannot satisfy all of the jobs' requirements. Thus some jobs will be blocked due to lack of resources. The blocking rate we defined here is $BR = \sum Drop_j/J$, which is the number of blocked jobs divided by the total number of input jobs. In our

problem, we suppose that the dependent tasks automatically block when their parents are blocked. So the whole job that consists of these tasks will be blocked. We investigate the variations of the blocking rate for different input traffic load under different job scheduling policies on the GCE topology. Figures 3.8 and 3.9 correspondingly show the blocking rate variations when employing Best-Fit and Tabu search methods to acquire the optimal solutions. We can observe that when the number of submitted requests is less than 150, Tabu search has a lower blocking rate than Best-Fit under any job scheduling policy. When the number of submitted requests continues to increase, the blocking rates obtained by Tabu search and Best-Fit are same under any job scheduling policy except the SSF job scheduling policy. The blocking rate of Tabu search is 35.3% lower than that of Best-Fit under SSF job scheduling policy when the number of input jobs is 250. In general, the blocking rates for Best-Fit method under FCFS and Random job scheduling policies are relatively less than that under other job scheduling policies; the blocking rates for Tabu search method under FCFS, Random and SSF policies are relatively lower. The heuristics with FCFS and Random job scheduling perform better and more reliably than others with different inputs. In addition, we compare the blocking rates of Best-Fit and Tabu search under the same job scheduling policy, and discover that Tabu search performs better than Best-Fit when the number of input jobs is smaller (e.g., less than 150), but performs the same when the number of input jobs is larger (e.g., more than 150).

Furthermore, for the same input data size (i.e., number of input jobs), different data sets are used to test the blocking rate of the two heuristics in our experiment. Figure 3.10 compares the blocking rate of Best-Fit and Tabu search heuristics for several different input data sets where each set consists of 150 different input jobs. The error bars in the figure indicate the 95% confidence interval for the average with the tested input data sets. We can see from Fig. 3.10 that under different job

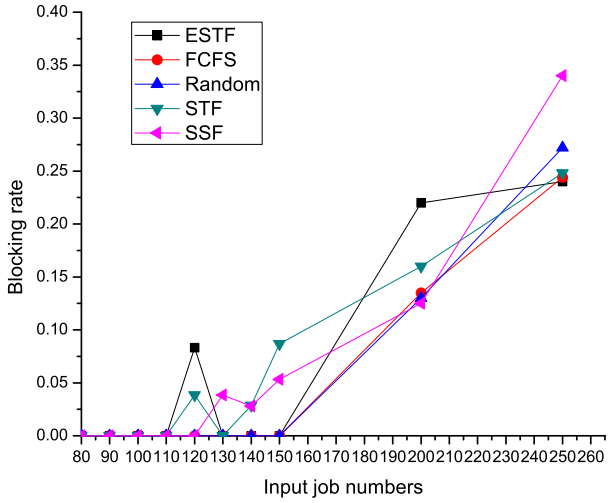


Figure 3.8
Variation of BR of Best-Fit heuristic on GCE topology.

scheduling policies, for different input sets of the same size, the average blocking rates of Tabu search are lower than those of Best-Fit. This result illustrates our conclusion above that Tabu search performs better in terms of blocking rate than the Best-Fit method when the number of input jobs is smaller.

To observe the effects on the total CapEx and blocking rate for two heuristics under different job scheduling policies when the cost model changes, we adopt a new multi-layer network cost model [24] to carry out more experiments on both 6-node and GCE topologies. Similar results are obtained compared with the results we obtained above. With the updated cost model, both heuristics deliver the solutions that are quite close to the optimal solutions by OPL in terms of total CapEx. In addition, both Best-Fit and Tabu search based heuristics under FCFS and Random job scheduling policies can acquire lower CapEx compared to other job scheduling policies when the size of inputs is larger than 40. The blocking rates of the two heuristics under FCFS and Random job scheduling policies are relatively less than those under other job

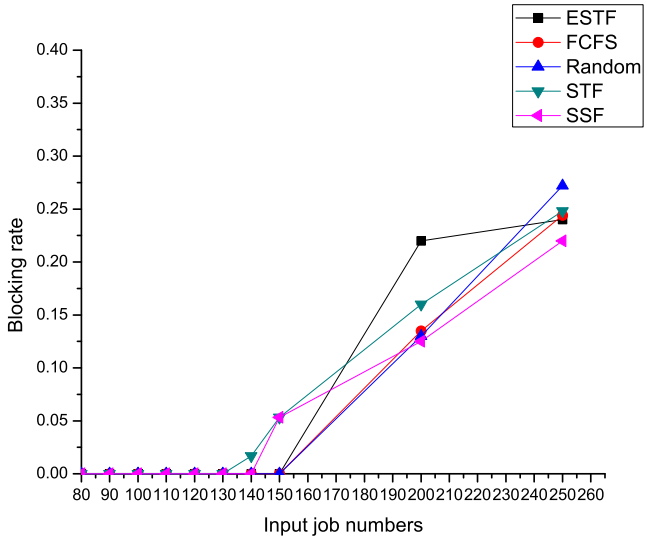


Figure 3.9
Variation of BR of Tabu search heuristic on GCE topology.

scheduling policies with different size of inputs. To sum up, our resource allocation model works well for the updated cost models.

3.6 Conclusion

In this chapter, we develop the MILP models and propose the Best-Fit and Tabu-search-based heuristics with several distinct job scheduling methods to solve the bandwidth-guaranteed optimal joint resource scheduling problem in the Grid/Cloud environment. To offer reliable network bandwidth resource for the Cloud users to transmit data between data centers, the IP-over-OTN-over-WDM multi-layer optical network architecture is introduced to reserve the wavelengths along the constructed optical circuits. We investigate the optimal joint resource allocation problem from the Cloud provider's point of view to minimize the total CapEx for resource allocation. Both MILP and heuristics work well to solve the problem, except that MILP is time consuming. In our study we observe that Tabu search method can obtain the

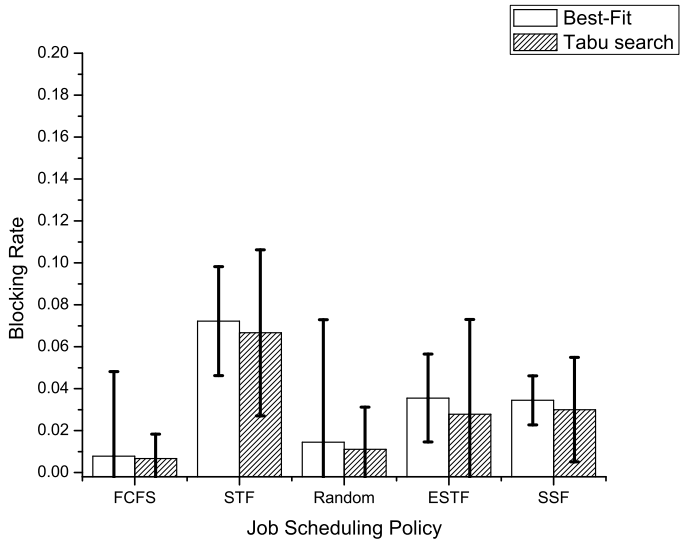


Figure 3.10
Average blocking rate comparison with the input size of 150 jobs on GCE topology.

solutions which are much closer to the optimal solutions by MILP, with less CapEx than that of Best-Fit method. In the blocking rate aspect, we discover that Tabu search has a lower blocking rate than Best-Fit when the number of submitted jobs is less than 150 under any job scheduling policy, while the blocking rates of Tabu search are the same as those of Best-Fit when the number of submitted jobs increases under most of the job scheduling policies.

Chapter 4

User's Viewpoint: Budget-optimized network-aware joint resource allocation in Grids/Clouds over optical networks

4.1 Introduction

The Grid/Cloud computing network model allows resources to be supplied according to users' requirements which could lead to overall cost reduction [88]. Solving resource allocation problems remains a very important topic in the area of Grid/Cloud computing. The challenges for resource allocation in Grids/Clouds mainly include several aspects: resource modeling, resource selection and optimization, resource offering and treatment, resource discovery and monitoring [2]. Thus developing solutions to cope with resource allocation challenges is an important topic in the field of Grid/Cloud computing. For resource selection and optimization, which is one of the four challenges, the provider needs to fulfill all requirements and optimize the usage of the infrastructure when given the information of resource availability in Grids/Clouds. A lot of studies have been carried out on the resource allocation for Grid/Cloud networks with different emphasis. However, some studies only target the computing resource allocation or merely network resource allocation in Grids/Clouds environment. From a practical aspect, users are offered infrastructure services from data centers in a Grid/Cloud network to complete their computing intensive tasks, and they might also have requirements for data transmission between the executing tasks

that distributed in the Grid/Cloud networks. Optical networks are widely used for inter-data center and intra-data center communication. This situation leads to the challenge of considering network-aware joint resource allocation in Grid/Cloud optical networks.

In the field of Grid computing networks, the Open Science Grid (OSG) provides distributed computing resources to users to meet their needs of research and academic communities at all scales [8]. To maintain and improve distributed high throughput computing services, managing resources responsibly and efficiently becomes an essential task. HTCondor [10] is a specialized management system in Grid computing networks that provides a job queuing mechanism, scheduling policy, resource monitoring and resource management. HTCondor system picks a submitted job in the queue, checks the requirements of computing resources by the job and processes the resource allocation for the job (e.g., checks if there are enough resources in Grid to be allocated, updates the resource status in Grid). HTCondor involves many of the emerging Grid and Cloud-based computing methodologies and protocols. However HTCondor does not deal with the network resource allocation. The task of integrating network resource management with current HTCondor system is in progress [26].

In Cloud computing networks, as we know, the IaaS consumers are offered a wide diversity of Cloud resources from multiple, distributed Cloud providers, such as Amazon Elastic Compute Cloud (EC2) [12] and Google Compute Engine (GCE) [13] at distinct hourly cost rates. Customers pay for the resources they need under the "pay-as-you-go" model in current Cloud computing network business model. Therefore from the customer's perspective, what they want intuitively is to obtain resources from the Cloud for their jobs at low rental cost. So how to realize the resource allocation for users under given conditions is what we need to solve. In the related field of Grid [39] computing, investigations such as [40] [43] have been carried out on

resource allocation or task scheduling based on distinct requirements or objectives.

In the field of Cloud computing, the studies in [45] and [46] propose some solutions for the resource allocation problem which focused on the management of data centers. The study in [45] uses Lyapunov optimization technique to design an on-line admission control, routing, and resource allocation algorithm for a virtualized data center. And the study in [46] proposes an efficient dynamic task scheduling scheme for virtualized data centers. The virtual machine (VM) allocation is a challenging sub-problem as well. The work in [47] investigates the dynamic VM provisioning and allocation problem for the auction-based model. An integer program is formulated and truthful greedy and optimal mechanisms are designed for the problem. The proposed mechanisms achieve promising results in terms of revenue for the Cloud provider.

In this chapter, we focus on the network-aware resource selection and optimization problem in the Cloud network from the customer's perspective: minimize the total rental cost (budget) for each user to obtain their required resources. The job collector in our joint resource allocation simulator collects all the submitted jobs from users first. Then the resource allocator who has a whole view of all the resources in the Cloud will allocate required resources for the collected jobs and update the available resources in the Cloud, as shown in Figure 4.1. The resource allocation simulator can be invoked for multiple rounds, and in each round it deals with the resource allocation for a batch of jobs. Given the inputs for the resource allocation problem, the scheduler needs to realize reasonable resource allocation for as many consumers as possible with minimum budget for each user, of course, within the capability of Cloud resources. Due to the emergence of cloud computing and various cloud services which are remote and geographically distributed, data centers interconnected by optical networks have attracted much attention of network operators and service

providers [89]. Optical wavelength division multiplexing (WDM) light paths in the form of "lambda service" offer guaranteed bandwidth connectivity for applications across the Cloud. Scheduling of optical layer resource reservation is an active area of study [90] [91]. Nowadays, the traffic is growing so fast in the Cloud environment and more and more data intensive applications need to transmit a large amount of data. In this case, there is a need for the Cloud provider to offer high bandwidth for data transmission for such applications, with the purpose of reducing data transmission delay or increasing the reliability. For example, On-demand Secure Circuits and Advance Reservation System (OSCARS) has been implemented and deployed on ESnet to provide multi-domain, high-bandwidth virtual circuits that guarantee end-to-end network data transfer performance [92]. Thus we consider utilizing the optical network to provide guaranteed network bandwidth in the Cloud environment according to customers' requirements. In our problem we investigate the wavelength reservation in optical networks to complete the bandwidth resource allocation for the jobs that need high bandwidth to transmit data, such as scientific projects running in the Grid/Cloud environment [93]. Jobs require guaranteed bandwidth service that can be provided, for example, by provisioning a distinct wavelength(s) connection from end-to-end. Thus to deal with the network-aware joint resource allocation, we consider optical circuits and reserve wavelengths to complete the bandwidth resource allocation for the jobs. A Mixed Integer Linear Programming (MILP) model with the linear constraints is constructed and two polynomial time heuristics (Best-fit and Tabu Search based heuristics) are proposed to obtain the optimal solutions for the problem.

The rest of this chapter is organized as follows. Section 4.2 presents the model of the joint resource allocation problem including the optical network model and cost model. Section 4.3 shows the MILP formulations of this problem. Section 4.4

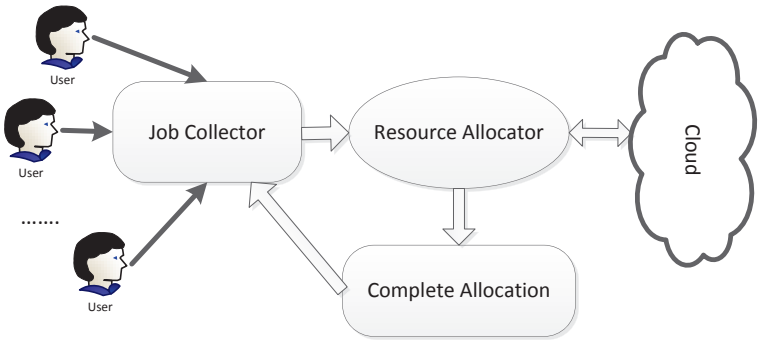


Figure 4.1
The resource allocation simulator.

describes the corresponding optimal heuristics. Section 4.5 evaluates the performance of MILP formulations and heuristics on two topologies. Section 4.6 concludes this chapter.

4.2 Problem Modeling

The work in this chapter, different from other works, first designs a new resource allocation model which combines computation resources and network resources together. Second, submitted jobs that are modeled as directed multi-stage graphs with single source/destination node are considered in this chapter and these bring more constraints for the joint resource allocation problem. Each job consists of a number of sequential tasks or parallel tasks or both. The adopted job structures are reasonable in practical Clouds, since when a user submits a job to the Cloud computing network, the job may contains several tasks which can be executed in parallel or must be executed sequentially. Third, we introduce the temporal parameters for our Cloud resource allocation problem. Fourth, the optical network is adopted to provide guaranteed bandwidth for users by reserving wavelengths along the established optical paths. The joint resource scheduling model proposed by us uses budget minimized resource allocation. The objective of our model is to minimize the budget (total

rental cost) for each user to obtain enough resources for executing their submitted jobs, while allowing the Cloud providers to accept as many job requests from users as possible.

4.2.1 Problem Description

In a Cloud network, a large amount of resources including computing resources are distributed among the various physical hosts or VMs. How to allocate available computing resources (processor, storage) and network resources (optical transponder (OT), wavelength and physical links such as optical fibers) in WDM layer of network to the submitted jobs properly to make sure each user incurs the minimum rental cost is the problem that we investigate.

A user submits a single job which consists of several tasks to the scheduler in the Cloud computing networks. The job needs to obtain a certain amount of processor and storage resources for execution and a certain amount of network bandwidth for data transmission between related tasks with minimum rental cost.

4.2.2 Problem Assumptions

To simplify our model and also to keep it reasonable for realistic joint resource allocation in Cloud computing networks, we make the following assumptions.

- A node in the topology stands for a data center. Each node has potentially, different processor and storage capacities. Each link in the topology is bi-directional.

- Jobs arrive one by one and are collected by the resource allocator first, then the allocator will process them together (batch processing).

- Execution cycle, noted as S_{max}, is slotted into 24 time slots. Each time slot is 1 hour, noted as s. Jobs should be completed within one execution cycle.

- We know that Grid computing tasks are often broken down into multiple sub-tasks and connected using a directed acyclic graph (DAG) to form a grid work flow [94]. So in our work here, we suppose a job consists of one or multiple dependent/independent tasks. Independent tasks in one job can be executed in parallel, while dependent tasks must be executed sequentially. A job structure can be modeled as a directed multi-stage graph with a single source/destination node (a DAG), as shown in Fig. 4.2, similar with the structures we used in our previous work [33].

- The required processor and storage resources by each task, must be allocated from the same data center node.

- The network bandwidth is reserved for the whole task execution duration to guarantee the real time transmission of the generated intermediate data.

- Occupied resources will be released once the execution of a task is completed.

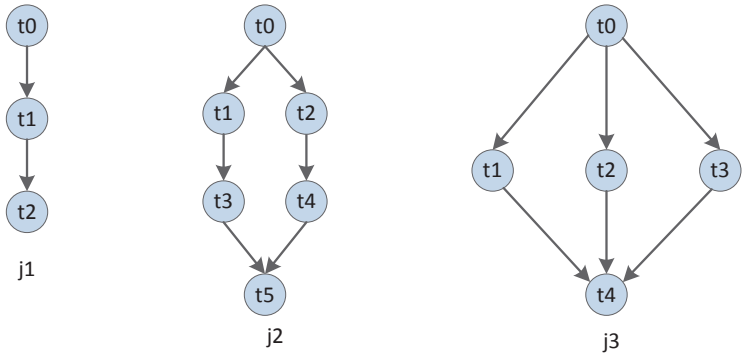


Figure 4.2
Job structure – directed multi-stage graph.

4.2.3 Optical Network Model

To guarantee the bandwidth for the network resource reservation, there is a need to setup an optical circuit in the WDM layer of the network and allocate bandwidth to the user. In our problem, each pair of OTs is used as the two ends of an optical circuit established to transmit optical signals. The optical fiber links in the network topology are bi-directional as described above and consist of several wavelength channels at specific bandwidth. We assume each fiber link in our optical network model has 40 wavelengths and each wavelength has 10 Gbps bandwidth. Each transponder pair is responsible for one optical path that reserves one wavelength along the path. Note that our formulation can be easily adapted to handle higher data rates (e.g. 100 Gbps per wavelength) and various wavelength numbers on each fiber link (e.g. 100 wavelengths on each fiber link). The simplified optical network model in this work does not consider the limitations due to optical signal reach and regeneration of optical signals.

In the Cloud network, each node which owns a large amount of resources can be seen as a data center, with its own intra-data center network. However we consider only bandwidth reservation for traffic across data centers in this work. The optical layer network we considered in this work is the packet optical transport network. The integrated packet optical transport network simplifies the network and increases efficiency. The packet flows can be flexibly delivered directly from DC to DC.

4.2.4 Price Model

All Cloud providers charge users for processor, storage and network resources including API calls and data transfer. Our price model is based on the "pay-as-you-go" method, in which customers pay the resource bills according to how many resources

they use and how long they use the resources. Based on our study of Amazon EC2 and Google GCE price models [95] [96], we propose three price models (for processor, storage and network resources). The values in the price model are parameters and can be changed for specific cloud providers, if needed in the future.

For the processor resource price model, we introduce the concept of compute power of each node, which can be measured by the number of cores of a single processor. We assume that if the compute power is larger (a processor has more cores) at one node, the processor unit price is higher. Two boundaries are proposed to divide processor capacity into three levels as shown in Table 4.1.

For the storage resource price model, the storage resource price depends on the storage amount which is measured in GB, on each node. The more storage resources a node has, the less storage unit price it will have, as is shown in Table 4.2.

For the network resource price model, the network resource price is divided into the price of OT and the price of common cost for using optical fibers. In DWDM networks, wavelength cost is usually modeled by two parts: optical OT cost and the common cost. The common cost includes optical system device cost, fiber cost, optical amplifier cost, installation cost, etc [97]. So in this work we incorporated the cost of optical system device, such as optical amplifier, into the common cost while using optical fibers. The price of common cost is modeled as price per mile of the links. The price of OT resource depends on region and node type. We divide network topology into three regions (US-east, US-west and US-central), in which the east region has lowest transponder unit price followed by central region, and west has the highest transponder unit price. Different node types have price variation within one region. The key junction node which has higher traffic load should maintain more transponder resources with higher unit price. The network resource price model is shown in Table 4.3.

Table 4.1
Price model for processor resource

Number of cores in a processor	Price/processor/time slot
0∼ 5	$ 0.29
5∼20	$ 0.58
20^{+}	$ 1.16

Table 4.2
Price model for storage resource

Storage amount	Price/GB/time slot
$\leq$ 100GB	$ 0.36
100GB $\sim$ 1TB	$ 0.18
1^{+} TB	$ 0.09

Table 4.3
Price model for network resource

Transponder resource cost		
Network region	Node type	Price/transponder/time slot
east region	key junction node	$ 0.08
	general node	$ 0.02
central region	key junction node	$ 0.09
	general node	$ 0.03
west region	key junction node	$ 0.11
	general node	$ 0.05
Common cost: physical links		
price/mileage		$ 0.0001

4.3 MILP Formulation for the budget-optimized resource allocation problem

We develop an MILP mathematical model for our problem to assign resources to the jobs submitted by users. In the MILP formulations, we have three types of inputs: the input of resource model in terms of network topology which indicates the

node/link information, and the resource information on each node/link; the input of submitted jobs from consumers that contains the budget, start/finish time and other information; the input of current traffic in the Grid/Cloud from which the current status of the resources on each node/link could be obtained. In the following we describe the three parts of the input for this network-aware joint resource allocation problem in detail.

4.3.1 Resource Modeling Input

In the following we describe the three parts of the input for this network-aware joint resource allocation problem in detail.

The resource modeling inputs indicate the number of nodes/links in the network topology and also the resources on each node/link. $Node = (n, P_n, D_n, OT_n, cp_n, cd_n, cot_n)$ and $Link = (src_l, des_l, Len_l, cl_l)$. Here src_l is the source node of current link; des_l is the destination node, the meaning of other elements is described in Table 4.5.

The demand inputs indicate the jobs submitted by users. $Job = (j, STime_j, FTime_j, Bud_j)$ and the tasks of a job $Task = (t, j, tSTime^t_j, tFTime^t_j, RP^s_{jt}, RD^s_{jt}, ROT^s_{jt}, C_{ID}, P_{ID})$. t is the task id; j is the job id that current task belongs to; P_{ID} and C_{ID} are the ID of the parent and child tasks of current task. The meaning of other elements is described in Table 4.5.

The current traffic inputs indicate the resources that are being used in the network. Node status $(n, \alpha^s_n, \beta^s_n, \gamma^s_n)$ describes the number of occupied processor/storage/OT resources on node n in time slot s. Link status (l, ω^s_l) describes the occupied wavelength on link l in time slot s.

Some other notations we used in MILP formulation are listed in Tables 4.4, 4.5 and 4.6.

Table 4.4
Constant Parameters

J	set of jobs
T_j	set of tasks that belongs to job j
N	set of nodes in network topology
L	set of links in network topology
s	one time slot
S	an executing cycle, consisting of certain number of time slots, indicated by allocator
cl	fiber link rental price per mileage
cb	network bandwidth price per Gb per time slot

Table 4.5
Variables

P_n, D_n, OT_n	number of processor/storage/transponder resources on node n
cp_n, cd_n, cot_n	price per processor/storage/transponder unit on node n per time slot s
Len_l	length of link l
$STime_j, FTime_j$	start time/finish time of job j
$tSTime_j^t, tFTime_j^t$	start time/finish time of task t in job j
Bud_j	total budget estimated for job j
$RP_{jt}^s, RD_{jt}^s, ROT_{jt}^s$, RB_{jt}^s	number of required processor/storage/transponder and bandwidth resources by task t in job j in time slot s
$\alpha_n^s, \beta_n^s, \gamma_n^s, \delta_l^s$	current occupied processor/storage/transponder and bandwidth resources on node n /link l in time slot s

4.3.2 Objective and Constraints of the MILP formulations

The objective of our problem is to complete the resource allocation for all jobs in the submission set (we call it as full-fit) while minimizing the total rental expenditure for all jobs.

Table 4.6
Decision Variables

$UP_{jt}^{ns}, UD_{jt}^{ns}$, $UOT_{jt}^{ns}, UB_{jt}^{ls}$	finally allocated processor/storage/transponder and bandwidth resources to task t in job j at node n /link l in time slot s.
C_j	total cost for executing job j
Dep_j^{ik}	binary parameter, equals 1 if task k is dependent on task i, both are belonged to job j
X_j	binary parameter, equals 1 if job j is accepted
F_{jt}^n	binary parameter, equals 1 if task t of job j obtains resources on node n

Objective:

$$Minimize \sum_{j \in J} C_j \tag{4.1}$$

$$\begin{aligned} C_j = & \sum_{t \in T_j, n \in N} UP_{jt}^{ns} \cdot cp_n \cdot Dur_j^t \\ & + \sum_{t \in T_j, n \in N} UD_{jt}^{ns} \cdot cd_n \cdot Dur_j^t \\ & + \sum_{t \in T_j, n \in N} UOT_{jt}^{ns} \cdot cot_n \cdot Dur_j^t \\ & + \sum_{l \in L} Len_l \cdot cl + \sum_{l \in L} UB_{jt}^{ls} \cdot cb \end{aligned} \tag{4.2}$$

where $\forall j \in J$, $Dur_j^t = tFTime_j^t - tSTime_j^t + 1$.

Task Dependency Constraint:

$$Dep_j^{ik} = 1, \forall i, k \in T_j, j \in J, tFTime_j^i \leq tSTime_j^k \tag{4.3}$$

Time Constraints:

$$tFTime_j^t \leq S, \forall j \in J, t \in T_j. \tag{4.4}$$

Required Resource Constraints:

$$\sum_{n \in N} UD_{jt}^{ns} = RD_{jt}^s \cdot X_j \tag{4.5}$$

$$\sum_{n \in N} UP_{jt}^{ns} = RP_{jt}^s \cdot X_j \tag{4.6}$$

$$\sum_{n \in N} UOT_{jt}^{ns} = ROT_{jt}^s \cdot X_j \tag{4.7}$$

$$\sum_{l \in L} UB_{jt}^{ls} = RB_{jt}^s \tag{4.8}$$

where $\forall j \in J, t \in T_j, s \in [tSTime_j^t, tFTime_j^t]$.

$$\sum_{n \in N} F_{jt}^n = 1, \forall j \in J, t \in T_j \tag{4.9}$$

Resource Capacity Constraints:

$$0 \leq \sum_{j \in J, t \in T_j} UP_{jt}^{ns} \leq P_n - \alpha_n^s \tag{4.10}$$

$$0 \leq \sum_{j \in J, t \in T_j} UD_{jt}^{ns} \leq D_n - \beta_n^s \tag{4.11}$$

$$0 \leq \sum_{j \in J, t \in T_j} UOT_{jt}^{ns} \leq OT_n - \gamma_n^s \tag{4.12}$$

$$0 \leq \sum_{j \in J, t \in T_j} UB_{jt}^{ls} \leq W - \delta_l^s \tag{4.13}$$

where $\forall n \in N, s \in S$.

Budget Constraints:

$$C_j \leq Bud_j, \forall j \in J : \tag{4.14}$$

Full-fit Constraints:

$$\sum_{j \in J} X_j \quad = \quad |J| \tag{4.15}$$

The total expenditure of a job in the objective consists of processor cost, storage cost and optical transponder cost for tasks, network bandwidth cost and fiber link cost for transporting data which are shown in Equation 5.2, in which Dur_t^j is noted as the duration of task t in job j. The objective function is subjected to the following constraints. Equation 5.3 requires that if task k of job j is dependent on task i in the same job, task k must execute after the execution of task i. Equation 5.4 guarantees that each job should complete its execution in one execution cycles. Equations 4.5–4.8 require that in the indicated time duration, a task obtains the required resources if its job is not dropped. Equation 5.9 guarantees that a task gets resources (processor, storage, transponder) from the same node. Equations 4.10–4.13 show that the allocated resources on each node/link in each time slot cannot exceed the amount of the currently available resources on this node/link. In the formulations, we do not have wavelength continuity constraints. We suppose that wavelength converter is available on each intermediate node along the routing path so that the optical signal can be transmitted via any available wavelength. Equation 5.14 bounds the total expenditure of each job to the budget given by user. Equation 5.15 gives the full-fit constraint which means that all jobs in the submission set need to be satisfied.

4.3.3 MILP Formulation Complexity Analysis

The number of variables can be calculated by $9NS+2L+5J+JT(2+3S+3SN+T+N)$, while the number of constraints can be calculated by $TJ(3S+T+1)+3(J+NS)$. For a 5 jobs inputs and two topologies (10-node and GCE topologies) we investigated in Section VI, the number of variables are 1108974 and 8767244 correspondingly, and the number of constraints are 642377 and 5025037 (the numbers are obtained through the IBM OPL CPLEX Optimization Studio [86] during simulations) correspondingly.

The optimal solution of the MILP model for each user can be obtained by CPLEX optimization software. However as described in Section VI, solving the MILP is a time consuming task.

4.4 Heuristic Algorithms

Two heuristics are developed to solve our joint resource allocation problem while consuming less time. Given a series of submitted jobs, Cloud resource information, and current traffic situation in the Cloud network, our target is to allocate resources to the jobs with minimal budget according to user's distinct requirements. We consider the job scheduling first, and then allocate resources in the Cloud network for each job.

The resource allocator schedules the jobs in the set sequentially. However, different scheduling orders of jobs may impact the final optimized rental cost. To investigate the effect of job scheduling on budget optimization, we investigate experiments with several sorting policies as shown in the following, which are already described in our work [34] and in Chapter 3.

- First come first served (FCFS). Jobs are scheduled according to their arrival order.

- Shortest job execution time first (STF). Jobs which occupies the resources for a shorter time will be scheduled first.
- Random schedule (Random). Submitted jobs in the queue will be scheduled in a random order.
- Early start time job first (ESTF). The job which starts executing earlier will be scheduled first.
- Simple job structure first (SSF). The job consisting of fewer sub-tasks are seen as having a simple job structure, and will be scheduled first.

After the job scheduling order is fixed, the resource allocation procedure needs to be carried out by the resource allocator. We implement the Best-Fit heuristic and Tabu search based heuristic to complete the resource allocation procedure. With the heuristics, we also explore the scenario where jobs are blocked when required resources are not available (Best-Fit). In such a case, the heuristics will attempt to minimize the total budget for the accepted jobs.

A. Best-Fit Heuristic

The Best-Fit heuristic comprises of two steps: computing resource allocation and network resource allocation. For each task in a job, we need to allocate processors, storage and OTs from the distributed data center nodes first. The nodes with lowest rental cost for computing resources and transponder resources will be selected for the tasks of a job. Then, with the selected nodes for each task, we set up paths between related nodes to allocate bandwidth. Each optical transponder is used for setting up one circuit and uses one wavelength on the corresponding link. When allocating resources for tasks in a job, we need to consider the dependency between tasks as well, so that a child task cannot be allocated resources before the completion of its

parent task. If there are not enough resources (on each node) for some task(s) in a job, the whole job will be blocked. The Best-Fit heuristic is shown in *Algorithm* 4 and *Algorithm* 5.

Algorithm 4 Best-Fit Heuristic

Input and Initializations:

$G = (V, E)$;
Current traffic in network;
$J = j_1, j_2, ..., j_M$; //set of jobs
$T_j = t_1, ..., t_k$, $j \in J$; //set of tasks belong to job j

Output:

minimized $\sum_{j \in J} C_j$

```
1: Select a job scheduling policy from: FCFS, STF, Random, ESTF, SSF;
2: for j ∈ J do
3:   //Computational resources allocation for tasks of job j
4:   for t ∈ T_j do
5:     if t has parent then
6:       if parent is done then
7:         for n ∈ V do
8:           Find n for t with minimum resource cost;
9:         end for
10:        if no node available for t then
11:          Block current job j, go to next job;
12:          Release the allocated resources for current job;
13:        else
14:          Update resource on selected node n;
15:          Update final minimum cost for t, next task;
16:        end if
17:      else
18:        Wait for parent task is done;
19:      end if
20:    else
21:      Use same resource allocation step as above;
22:    end if
23:  end for
24:  //bandwidth allocation for current job
25:  Update total cost C_j of current job j;
26: end for
27: return ∑_{j∈J} C_j;
```

Algorithm 5 Bandwidth allocation

```
1: for j ∈ J do
2:   for t ∈ T_j do
3:     Check t's connected adjacent task;
4:     desNode = t.adjacent.selNode;
5:     Compute shortest path for(t, desNode);
6:     Compute path cost;
7:     Update path cost for current job j;
8:   end for
9: end for
```

B. Tabu Search Based Heuristic

The Best-Fit heuristic is a greedy method and we would like to find a better method to solve such optimization problems. So we propose the Tabu search based heuristic to solve our optimization problem with the hope of obtaining solutions with lower budget and reduce the traffic blocking rate for the input demands. The basic concept of tabu search as described by Glover (1986) is "a meta-heuristic superimposed on another heuristic" [87]. The overall approach is to avoid entrenchment in cycles by forbidding or penalizing moves which take the solution, in the next iteration, to points in the solution space previously visited. In our Tabu search based heuristic, we use the solution obtained by Best-Fit heuristic as the initial solution, and adopt random move to find neighbors. The termination condition in the heuristic here is the moving times we required. In general we require the moving times should be twice than the number of the candidates in the solution pool, to increase the probabilities of visiting each candidate solutions through random move [35]. The heuristic is shown in *Algorithm* 6.

Algorithm 6 Tabu Search Based Heuristic

Input and Initializations:

$G = (V, E)$;
Current traffic in network;
$J = j_1, j_2, ..., j_M$; //Input job requests
$T_j = t_1, ..., t_k$, $j \in J$; //set of tasks belong to job j
$C_j = 0$; //initial job cost is 0

Output: Minimize $\sum_{j \in J} C_j$.

```
1: Update current available resources in the network;
2: Select job scheduling policy from: FCFS, STF, Random, ESTF, SSF;
3: Sort the topology nodes according to resource unit cost;
4: InitialSol := solution by Best-Fit heuristic;
5: OptSol := InitialSol; //set optimal solution
6: Generate solutions pool;
7: Set Tabulist;
8: while not-terminate do
9:    Random move to generate neighbor solution: Neighbor;
10:   if Neighbor ∈ Tabulist then
11:      Move operation, generate new neighbor;
12:   else
13:      CurrSol := Neighbor; //set current solution
14:      if CurrSol < OptSol then
15:         OptSol := CurrSol;
16:         Update Tabulist;
17:      else
18:         Move to generate new neighbor;
19:      end if
20:   end if
21: end while
22: Return OptSol;
```

4.5 Experimental Results and Analysis

The simulations are carried out on a Linux server with 16 GB memory for both MILP model and heuristics on two topologies, which are a 10-node mesh topology (Figure 4.3) and a real 20-node GCE data center locations topology in US (available in [34] figure 5). In addition a software tool IBM OPL CPLEX Optimization Studio is used

to simulate the MILP model. In the experiments, each job is generated randomly and consists of a random number of tasks (1-10). The required amount of resources of each task is also generated randomly. Each execution cycle lasts for 24 hours, so the start/end time of the execution of each task in within 24 hours. The size of a data set is defined as the number of jobs in that data set. In the experiments, for every data set size (20 jobs, 100 jobs, etc.), we randomly generate 10 groups of data sets with the same data set size. Therefore, the total expense saving ratio and blocking rate for each number of jobs is represented in terms of the average value of the 10 groups with 95% confidence interval. The joint resource allocation results for MILP

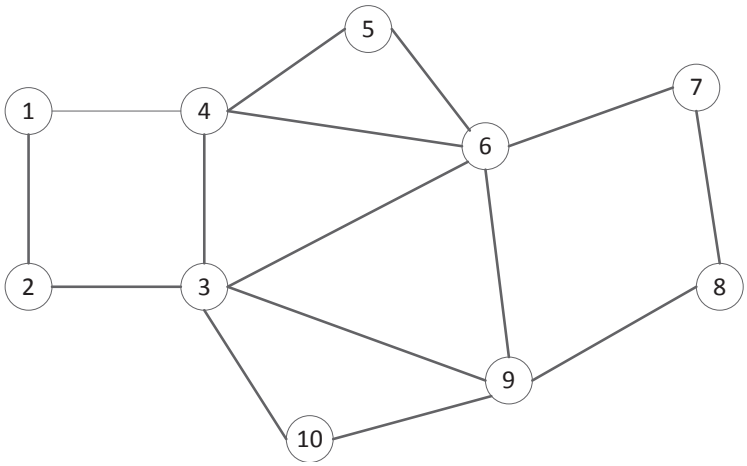
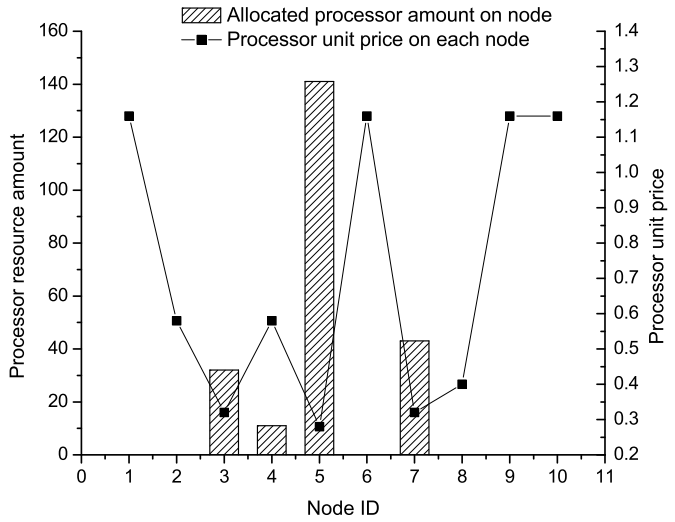


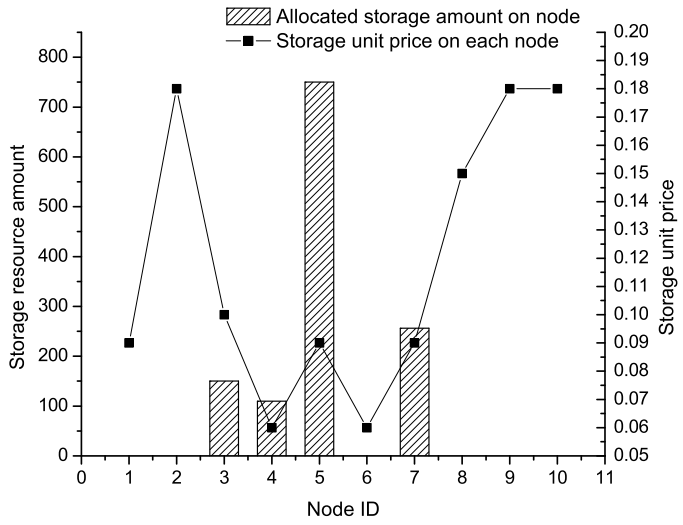
Figure 4.3
10-node Cloud network topology.

model are obtained using OPL Optimization first. This rental cost for each job is minimized and we show the resource allocation situation on each node of 10-node topology with 10 input jobs in Figure 4.4. We know that each node has a different amount of resource and different resource rental cost. From the graph we can see that the nodes with less unit resource cost will be chosen to allocated resources to as many users as possible, and thus the resources on these nodes have higher utilization. The unit rental cost of processor resource is dominant in deciding the node selection

than other two resources.



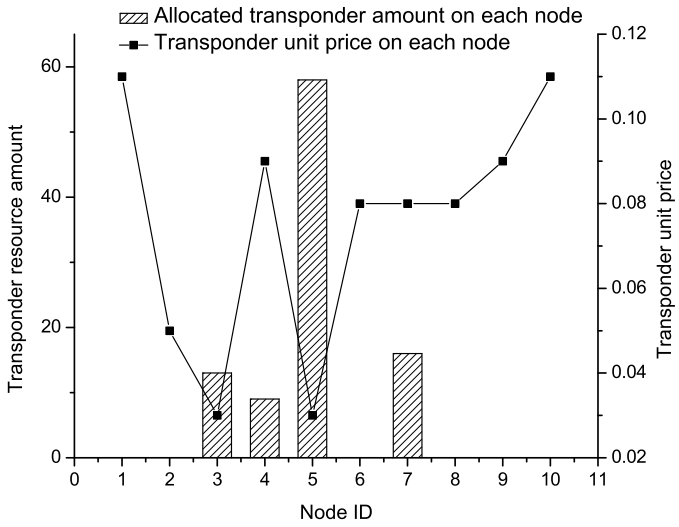
(a) Processor resource



(b) Storage resource

Figure 4.4
Resource utilization and unit cost of each node using MILP method for 10-node topology.

The CPLEX Optimization software can usually return the optimal results for our



(c) OT resource

Figure 4.4
Resource utilization and unit cost of each node using MILP method for 10-node topology.

problem, but it is also very time consuming. In our simulation, more than 1 hour is needed to generate the solution for 10 input jobs on a 10-node network topology. Hence for the larger GCE topology we report results using only our heuristic methods. Figure 4.5 shows the optimal resource allocation results of 5 input jobs obtained using CPLEX and Best-Fit heuristic with different job scheduling policies. The number in the legend is the time used to complete joint resource scheduling for all input jobs with corresponding method. We also compare the total expense obtained by OPL and Best-Fit heuristic on the 10-node network topology for different number of input jobs, see Table 4.7. The comparison in this table and in Figure 4.5 show that the Best-Fit heuristic we implemented with different job sorting policies can complete the resource allocation on Cloud network topology efficiently and fast.

We compare the actual rental expenditure of each job with the original budget for executing this job under different job scheduling policies based on best-fit scheduling

Table 4.7
Total expenditure comparison on 10-node topology

Number of jobs	OPL	Best-Fit				
		FCFS	SFT	Random	ESTF	SSF
10	271.36	287.77	296.16	291.02	285.55	290.41
20	443.88	484.5	493.49	486.58	473.47	480.09
30	620.93	654.02	665.34	657.96	648.08	648.41
40	∗	870.179	878.021	871.387	861.88	852.223
50	∗	1227.19	1242.57	1215.59	1229.4	1202.86

∗ means the CPLEX is running out of memory to generate optimal solutions for MILP formulations.

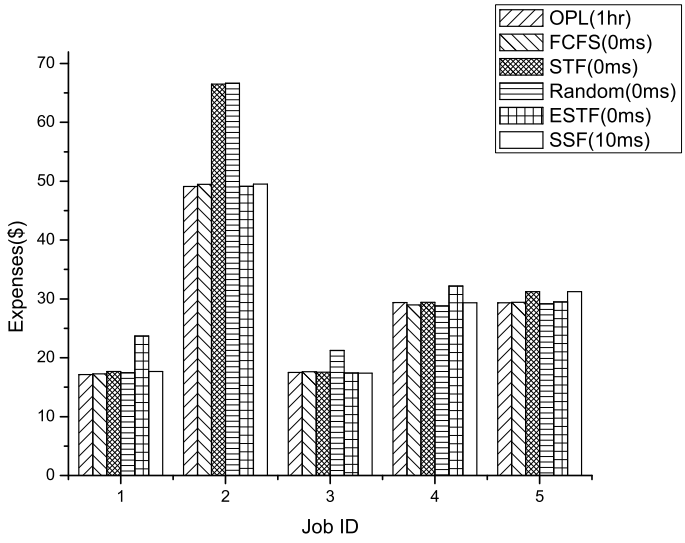


Figure 4.5
Expenditure comparison with 5 job inputs on 10-node topology, Best-Fit heuristic.

algorithm. We define each user's original budget as the total money the user needs to pay if the resources with most expensive unit price are allocated to the job submitted by the user. The expense saving ratio of job j is defined as $ESR_j = \frac{Bud_j - C_j}{Bud_j}$. Figure 4.6 shows each user's expense saving ratio in the 10-node topology with traffic load of 5 input jobs. More tests are carried out under different traffic loads on the 10-node topology, and the results show that for each job submitted by one user, the

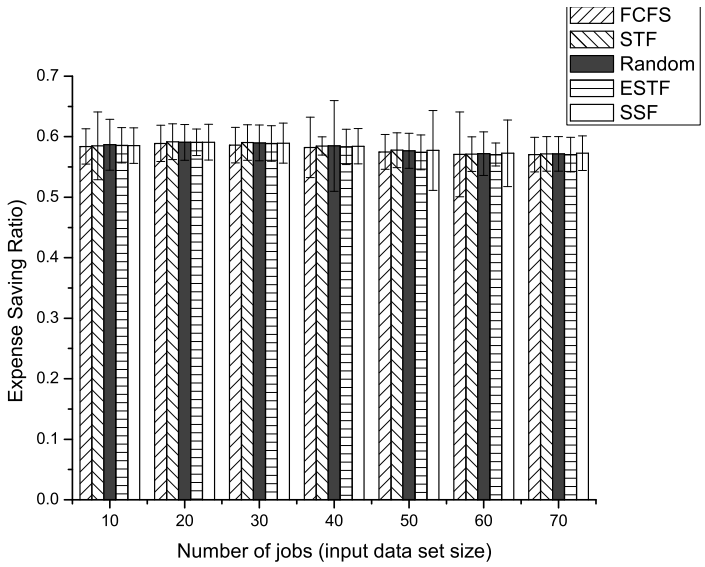


Figure 4.6
Expense saving ratio for 5 jobs under distinct job scheduling policies on 10-node topology, Best-Fit method.

expenditure decreases by least 30%. Especially for smaller jobs, which have simpler job structures and less resource requirements we can achieve a higher expense saving ratio which is nearly 70%. We also test the Best-Fit heuristic with different job sorting methods on GCE topology and obtain similar results. The jobs submitted to the scheduler can reduce their expenditure by 35%∼67.5%.

For the 10-node topology, with distinct input job numbers, the optimal solutions obtained through Tabu search heuristic are as good as those obtained by Best-Fit heuristic, and are approximate with the accurate solutions obtained by CPLEX. Figure.4.7 compares the total expense of Tabu search and Best-Fit heuristic with 15 input jobs, under different job scheduling policies on the GCE topology. Here we did not compare the results with those of CPLEX since it is very slow when solving our problem for a larger network topology. We can see from the figure that when the number of input jobs is 15, the Tabu search results are a little bit better than

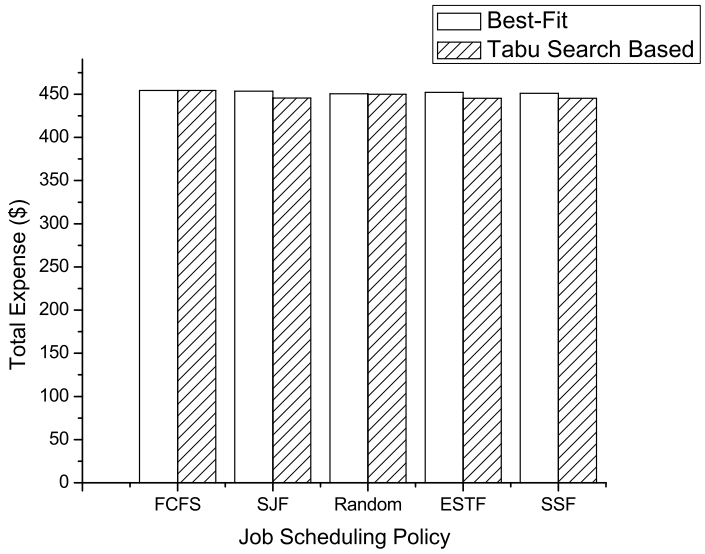


Figure 4.7
The total expense comparison of Best-Fit heuristic and Tabu search heuristic with 15 job inputs on GCE topology.

Best-Fit results under the SJF, ESTF, SSF job scheduling policies.

In the previous figure 4.6, we discuss the expense saving ratio for a single job, and here we will discuss the expense saving ratio for all jobs in the input data set. For the GCE topology, when the size of given input data sets is from 10 to 70 (input sets have 10 to 70 jobs), the Tabu search based heuristic obtains nearly the same results for total cost compared to the Best-Fit heuristic, so we only show the saving ratio obtained by Tabu search results in the following figure 4.8. Figure 4.8 shows the total expense saving ratio when given different number of jobs in the input data set (full-fit with no blocking) under each job scheduling policy. We can see that when given input data set with different size, the total expense saving ratio is around 57%. So our methods can reduce more than half of the original estimated budget for customers.

With the increase in traffic load, Cloud network topologies with limited resources cannot satisfy all of the job's requirements. If one or more tasks in a job cannot obtain

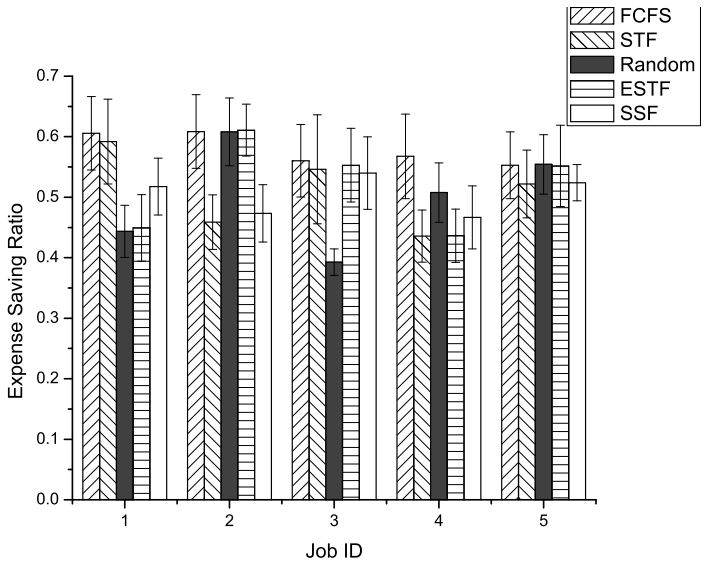


Figure 4.8
The total expense saving ratio for different input data set size on GCE topology.

enough resources during its execution period from any data center, or a child task starts executing before its parent task (see Best-Fit scenario description in Section V. A), the whole job will be blocked. The blocking rate (BR) for an input set is $BR = \frac{J_{block}}{J}$, in which J_{block} is the number of blocked jobs, J is the total number of jobs in a data set. We compare the changes in BR for different input traffic loads under different job sorting policies on the 10-node topology (shown in Figure 4.9). From the graph we can see that when the number of input jobs is less than 70, the blocking rate is 0 for all scheduling methods. After that, along with the increase of the number of input jobs, the blocking rate increases. The blocking rate with ESTF policy increases faster than others. In addition, SSF policy has a better performance in terms of blocking rate compared to other policies. We can see from the graph that when the input consists of 120 jobs, BR with SSF is nearly 66.7% lower than that with ESTF.

The BR comparison on GCE topology is also carried out, which is shown in Figure

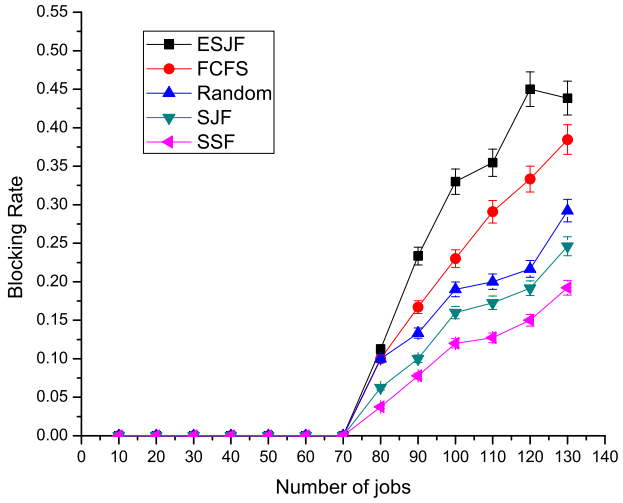


Figure 4.9
Variation of Blocking Rate (BR) under distinct job scheduling policies on 10-node topology, Best-Fit.

4.10. We can see from the figure that ESTF policy still results in higher BR than other job scheduling policies, while SSF always maintains a minimum value of BR. Therefore, compared with other job sorting policies, SSF is a better choice for Best-Fit resource allocation heuristic.

We also examine the BR of our proposed Tabu search heuristic for the optimization problem. The results show that, similar to Best-Fit heuristic, the Tabu search heuristic with SSF job scheduling policy also performs much better in terms of the BR, and has lower BR than other job scheduling policies. Figure 4.11 shows us the BR results for various job scheduling policies for the GCE network topology with Tabu search heuristic (Experimental results on 10-node topology are similar and we do not include the figure here due to space limitations). We can see that the BR under SSF is 50% better than that under ESTF when number of input jobs is 130.

In addition, the Tabu search heuristic reduces the BR significantly compared with

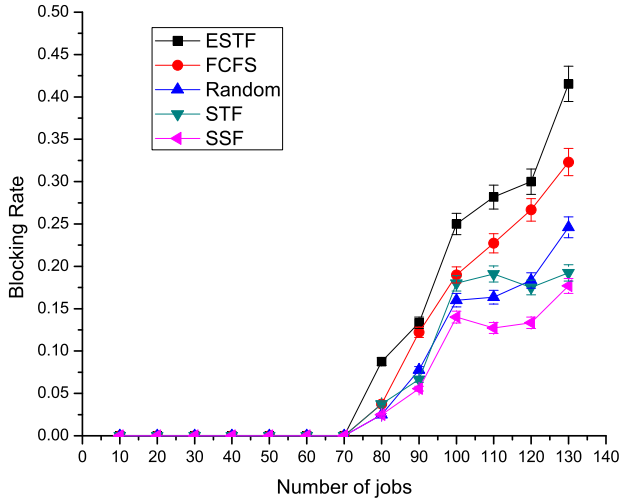


Figure 4.10
Variation of Blocking Rate (BR) under distinct job scheduling policies on GCE topology, Best-Fit.

the Best-Fit heuristic for our problem. In Figure 4.12 we compare the BR of Best-Fit and Tabu search heuristics under SSF job scheduling policy for the GCE topology. The BR is reduced by 4%∼25% than the Best-Fit heuristic. According to the statistics of all the simulation results, the Tabu search heuristic can reduce the BR by 4%∼30% than the Best-Fit heuristic under different job scheduling policies.

4.6 Conclusion

In this chapter, we develop an MILP model, and propose Best-Fit and Tabu search based heuristics based on several distinct job scheduling policies to solve the optimal joint resources scheduling problem in the Grid/Cloud network from the user's point of view. For the input traffic we consider different job structures which consists of parallel or sequential tasks. We also consider the network resource allocation, which is optical transponder allocation and bandwidth reservation for inter-data cen-

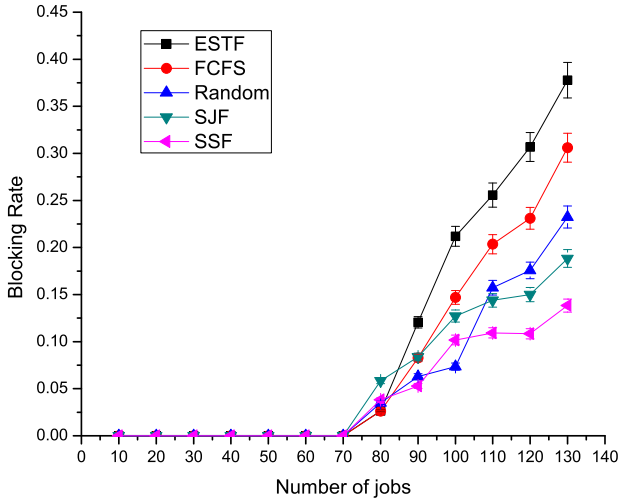


Figure 4.11
The blocking rate of Tabu search heuristic under distinct job scheduling policies on GCE topology.

ter communication in WDM layer of the network. The MILP model can solve our problem with an optimal manner, but it is time consuming when the input size is large. We can obtain an approximate optimal solution through our proposed heuristic algorithms within a very short time. From the experimental results, we observe that two heuristics with different job scheduling policies can reduce the user expense by at least 30% of their original budget. In addition, the Best-Fit algorithm with STF and SSF scheduling policies have a better performance on the traffic blocking rate. The traffic blocking rates under both scheduling methods are 5%∼25% less than other methods. In addition, the Tabu search based heuristic will equal or outperform the Best-Fit heuristic, and both can achieve approximate optimal solutions to the corresponding MILP solver results. The experimental results show that the Tabu search based heuristic with SSF job scheduling policy blocks less traffic, i.e., it has a lower blocking rate than other job scheduling policies. In addition, the Tabu search based

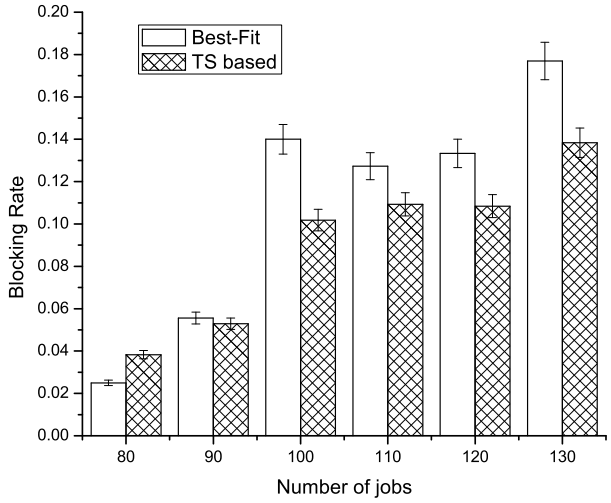


Figure 4.12
Blocking rate comparison by Best-Fit and tabu search under SSF job scheduling policy on GCE topology.

heuristic also reduces the blocking rate by 4%∼30% compared with Best-Fit heuristic under any job scheduling policy.

In this chapter, we only consider joint resource scheduling for submitted jobs during one cycle. We will consider continuous scheduling of the input traffic and the dynamic demands in our future work. We will also consider the use of elastic optical networks [98] [37] in the Cloud. In addition, we will involve the multiplexing technology to make our work support the applications with low bandwidth requirements. Moreover, we will consider the intra-data center network communication for the bandwidth-guaranteed resource allocation problem in the future.

Chapter 5

Provisioning Virtualized Cloud Services in IP/MPLS-over-EON Networks

5.1 Introduction

Cloud computing offers computing resources to a large amount of on-demand service applications. Customers can reserve the required resources through the infrastructure as a service (IaaS) to complete their computing intensive tasks. In the future, customers may not only want to reserve computing resources, such as virtual machines (VMs) and storage, but would also want to reserve their own Cloud environment. A new architecture proposed in [99] to support data center as a service (DCaaS) for the future Cloud computing could satisfy such requirements from customers. DCaaS allows customers to create their own Cloud platforms without constructing the physical DCs. The virtual data center (VDC) service which falls within IaaS enables users to quickly access the Cloud infrastructure from a service provider such as vCloud Suite by VMware [100], VMDC by Cisco [101], etc. A VDC consists of VMs that are connected through virtual switches, and virtual links with certain bandwidth. With the newly proposed DCaaS service model, customers could reserve resources from the physical Cloud computing environment to construct their own virtual Cloud environment. The reserved virtual Cloud environment consists of geographically distributed virtual data centers (VDCs) and backbone networks that connect these VDCs. In this case, we may need to consider the VDC as the unit of resource allocation [72].

The Cisco Global Cloud Index (GCI) [27] is an ongoing effort to forecast the growth of global data center and cloud-based IP traffic. GCI indicates in the forecast and methodology report for 2013-2018 that, the global data center traffic and global Cloud traffic will increase significantly in the future years [27]. For example, the report indicates that the annual global data center IP traffic will reach 8.6 zettabytes by the end of 2018, which will nearly triple over the next 5 years. In addition, the annual global cloud IP traffic will reach 6.5 zettabytes be the end of 2018, which will nearly quadruple over the next five years [27]. To support the large amount of traffic in the Cloud environment (within data center, data center to data center and data center to user) and satisfy the requirement of non-blocking bisection bandwidth among servers, huge bandwidth capacity should be provided by an efficient interconnection architecture. Therefore, the networking, such as optical networking, with scalable bandwidth capacity, low cost and low latency would be desirable [102].

In this chapter, we investigate the bandwidth guaranteed virtualized Cloud infrastructure provisioning (NE-VCIP) in multi-layer network architecture. As we know, the physical Cloud infrastructure comprises the DC infrastructure (i.e., computing, storage, and general IT resources) and the network connectivity interconnecting DCs with each other. In our problem, a virtualized cloud infrastructure (VCI) demand submitted by a user consists of the VDC infrastructures and the virtualized network (VN) connectivity. Each VDC is provided with required amount of computing resources. The VNs are specified with certain amount of bandwidth for data transmission. The bandwidth requirement is an essential addition which provides the significant benefit of performance predictability for distributed computing [72]. The centralized controller needs to map the VDCs and VN to the geographically distributed physical DCs and backbone networks that both have enough related resources. To guarantee the bandwidth requirements by VN, optical circuits are established. In this chap-

ter, we consider the backbone network with IP-over-EON (elastic optical network) architecture. So one important task for the controller is to complete the routing and spectrum assignment (RSA) in the multi-layer network when doing VN mapping. The elastic optical network (EON) has become a promising approach for flexible bandwidth provisioning in optical networks. EON can provide high capacity bandwidth for the demands that cannot be supported well in current WDM networks. In addition, EON allows for adaptive bandwidth provisioning for traffic demands with the use of advanced modulation formats and the bandwidth variable transponder technologies. In this case, the flexible and highly scalable bandwidth provisioning of EON architectures is considered as a significant approach to build effective and cost-efficient cloud-ready transport networks [103]. Furthermore, EON is cost-effective for both single channel and multiple channel modes, and can address the bandwidth waste problem well [104]. Thus we plan to adopt IP/MPLS-over-EON optical network architecture for our cost-optimized network-aware virtual cloud infrastructure provisioning problem in this work.

In this chapter, we made use of the flexible optical network as the backbone network in the Cloud to investigate the virtual cloud resource provisioning problem. We provided the guaranteed bandwidth through layer-1 while dealing with cloud resource provisioning. The objective is to minimize the total cost (CapEx and OpEx) for resource provisioning in the cloud environment. To the best of our knowledge, it is the first work that investigates cost-optimized virtual cloud resource provisioning while utilizing the IP-over-EON network architecture. In this work we further investigate the virtual cloud resource provisioning problem and the contributions are: (1) MILP models for two scenarios are constructed and simulated; (2) to optimize the total cost, sliceable bandwidth variable transponders (SBVT) are utilized and optical traffic grooming is considered in EON.

The rest of this chapter is organized as follows. In Section II, the network-efficient virtual cloud infrastructure provisioning (NE-VCIP) problem we investigated is described in detail. In Section III, two MILP models (Best-Fit and Full-Fit) for the NE-VCIP problem are discussed. In Section IV, a heuristic method for the NE-VCIP problem is discussed as well. In Section V, experiments are carried out for both MILP models and heuristic method, and the simulation results are analyzed. Section VI comes to the conclusion.

5.2 NE-VCIP Problem

The optimal resource provisioning in Cloud has been a challenge in the Cloud computing. Various investigations have been conducted for the resource provisioning problems in Cloud. In addition, VDC networks has been considered as a feasible alternative to satisfy the requirements of advanced Cloud infrastructure services. Proper mapping of VDC resources to their physical counterparts, also known as VDC embedding, can impact the revenue of cloud providers [68].

In the network-efficient virtualized cloud infrastructure provisioning (NE-VCIP) problem, a VCI demand submitted by a customer consists of VDC infrastructures and the virtualized network (VN) interconnecting VDCs. Each VDC requires a certain amount of computing resources (e.g. CPU and storage) and IT resources (e.g. ports for infrastructure connections within a VDC). The VN that connects VDCs requires a certain amount of bandwidth for data transmission. The centralized scheduler needs to map the VDCs and VN to the geographically distributed DCs and backbone networks such that both have enough resources. To guarantee the bandwidth requirements for the VN, optical circuits are established and the spectrums are assigned to the demands. In this work, we consider a backbone network which

uses an IP-over-EON architecture as shown in Fig. 5.1. At the starting point of the data transmission path, the data traffic goes across the IP/MPLS layer node to the connected EON layer node (bandwidth-variable wavelength cross-connects (BV-WXCs)) through bandwidth-variable transponders (BVTs). Then the data traffic travels along the light path in the EON layer, arrives at the EON layer destination node and finally reaches the end point of IP/MPLS layer. Therefore, to perform the VN mapping, an important task is to complete the routing and spectrum assignment (RSA) in the multi-layer network. EON is one of the most exciting future directions for optical networks and also an efficient and cost-effective solution for provisioning of Cloud traffic [105].

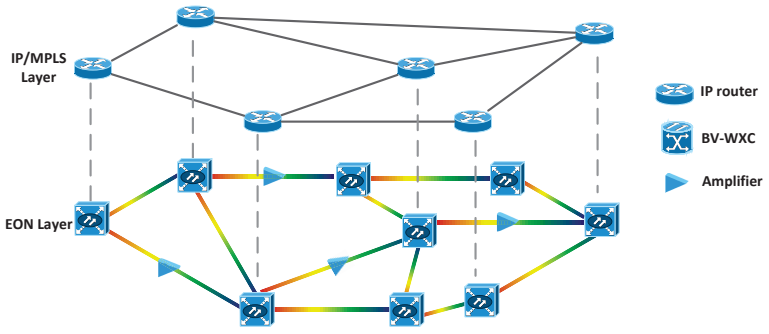


Figure 5.1
The IP/MPLS-over-EON architecture.

5.2.1 VDC mapping

For the VDC mapping, each VDC will be mapped to a physical DC which has enough required computing resources by the VDC. We suppose that no two VDCs in a same VCI demand will be mapped to the same physical DC (as shown in Fig. 5.2) since we would like to avoid the scenario of a disaster at one DC affecting multiple VDCs of a VCI demand. The geographically distributed DCs have different amount of resources with different rental prices. We assume that the DCs in the central region of the

Cloud network have lower rental price compared those in west/east regions, because of the richer resources and lower construction costs.

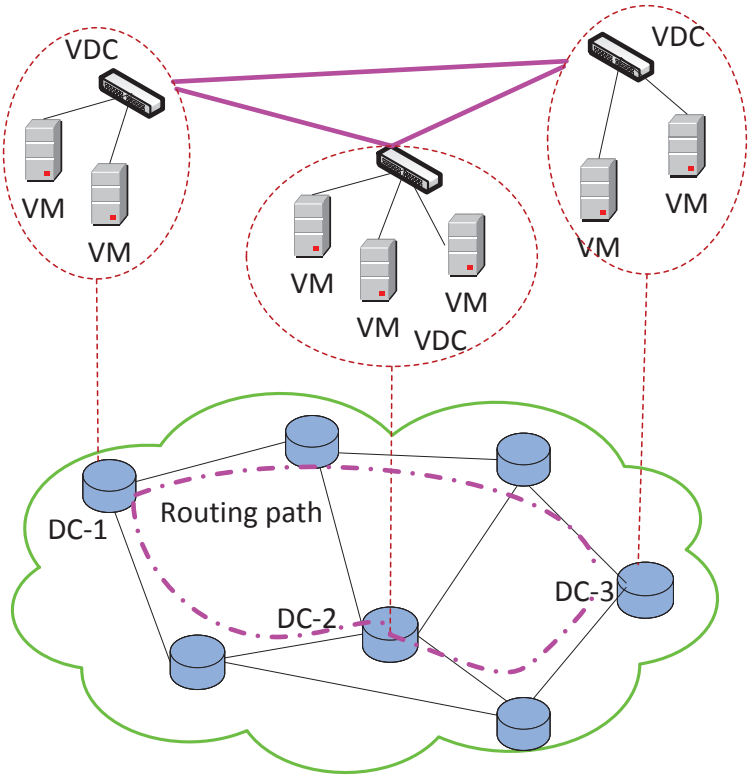


Figure 5.2
VCI demand mapping on the physical Cloud platform.

5.2.2 RSA in EON layer

The RSA problem in flexible grid optical networks consists of both the routing decision for traffic demands and the subcarrier assignment to satisfy the requirements by the corresponding traffic demands [106]. The VN mapping in the EONs is actually a RSA problem, which is *NP-hard* [107]. For the VN requirement of a demand, the central scheduler needs to find the path between two geographically distributed DCs that has the lowest cost and ensure that all the fiber links along this path have enough spectrum resources. Then the scheduler assigns the related frequency slots (FSs)

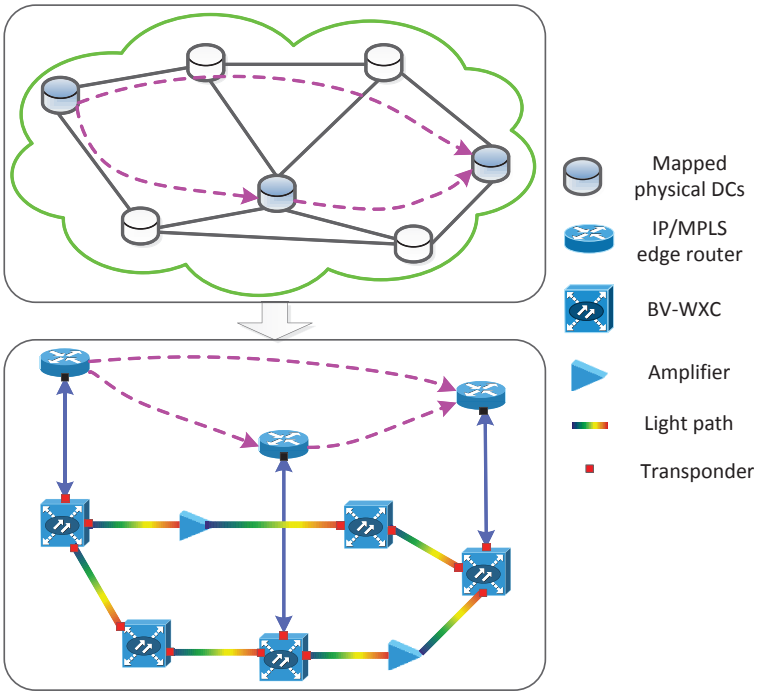


Figure 5.3
Multi-layer routing in the Cloud platform.

from each fiber link along the path for the demands. The required number of FSs must be contiguous in frequency domain and temporal domain for each link on the path. In addition, the links along the routing path must use the same FSs, which is called spectrum continuity. In our model the required bandwidth of a virtual link (VL) in VN is given in bit rate (Gbps). In order to estimate the number of FSs that each VL requires, we convert the required bit rate bandwidth into frequency (GHz) first according to the theoretical bandwidth efficiency limits for the main modulation formats [108]. Formula $F = B/M$ is used in the conversion. Here F is frequency; B is bit rate; M is mod-2 value of modulation formats, $M = 1, 2, 3, 4$ for Binary Phase Shift Keying (BPSK), Quadrature Phase Shift Keying (QPSK), 8 Quadrature Amplitude Modulation (8-QAM) and 16-QAM respectively. The assumed transmission reaches of modulation formats BPSK, QPSK, 8-QAM and 16-QAM are 5000, 3000, 1500 and

700 km respectively [109]. Then according to the frequency grid of the EON, the required number of FSs of a VL can be obtained. In EON, the available optical spectrum is divided into a set of FSs of a fixed finer spectrum width (frequency grid), such as 25 GHz, 12.5 GHz and 6.25 GHz. We use 12.5 GHz as the frequency grid for the operation of the EON in this work with a total of 320 FSs in the C-band on each fiber link.

5.2.3 Traffic grooming with sliceable BVT in EON layer

In our earlier work [37], we use BVTs to provide flexible light paths. An optical channel with any spectral width and central frequency can be established by the BVTs without strictly following the ITU-T fixed grid [110]. However this kind of BVT is non-sliceable, which means that only one optical flow can be transmitted by the BVT. In this case, the transponder utilization is a concern. For example, if 100 Gbps BVTs are adopted in the optical network and the bandwidth requirement of the demands are usually 25 Gbps, so 75 Gbps bandwidth of the transponder will be wasted. To improve the transponder utilization, sliceable BVTs (SBVTs) are adopted. S-BVT is an evolution of the BVT, which is a class of transponders able to dynamically tune the required optical bandwidth and transmission reach by adjusting parameters such as gross bit rate, modulation format, and shaping of optical spectrum [111]. S-BVTs enable the generation of multiple optical flows that can be routed into different media channels (a media channel is a specific portion of the optical spectrum and an optical path through the EON between two end-points) and flexibly directed toward different destinations [112]. A SBVT can be sliced into multiple virtual sub-transponders, and each pair of virtual sub-transponder (the transmitter side and receiver side) is responsible for setting up an independent light path from the source node to the destination node without electronic processing at the intermediate nodes

along the light path. In this case, for the example above, if we use 100 Gbps SBVTs in the optical network, a 25 Gbps virtual transponder can be sliced from the 100 Gbps SBVT, and the remaining 75 Gbps can be used by other demands, which improve the transponder utilization and increase the provisioned traffic [113]. In addition, the use of SBVTs could reduce the total number of transponders needed, thus correspondingly reduce the total network cost. Some previous work has shown that the target cost of 400 Gbps and 1 Tbps SBVTs reduces by 50% the transponder cost in a core network scenario [114] and the Operational expenditure savings related to stock of spare parts can be realized by using SBVTs versus fixed transponder [115].

In addition, the traffic grooming process is often used to reduce the network cost as well. The optical layer traffic grooming can be realized by SBVTs. In the optical traffic grooming, multiple optical flows transmitted from different virtual sub-transponders can be groomed onto one SBVT by an intermediate switching fabric, such as bandwidth variable wavelength cross-connects (BV-WXCs) and then switched as a single unit in the network [116]. Different traffic grooming (electrical traffic grooming and optical traffic grooming) would be conducted according to different types of transponder technologies (BVT, fully sliceable BVT and partially sliceable BVT) that are used [116]. In this work we only consider the fully sliceable BVT (mentioned as SBVT in the following contents for short) and the traffic grooming will be implemented in optical layer (optical traffic grooming). The optical traffic grooming with SBVTs in IP-over-EON networks is shown in Figure 5.4.

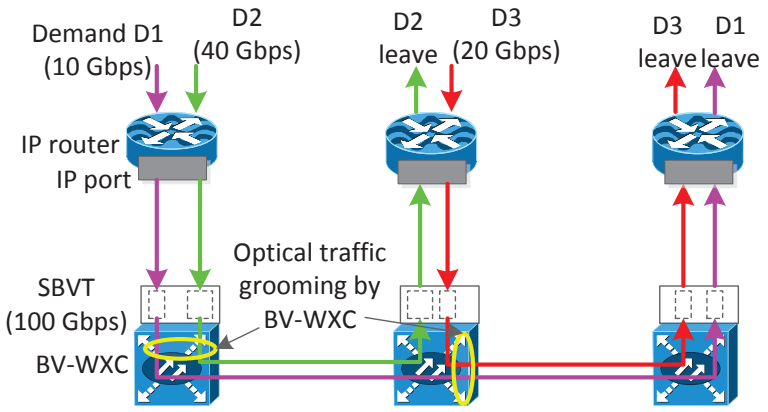


Figure 5.4
Optical traffic grooming with SBVTs and BV-WXCs in IP-over-EON.

5.3 Mathematical Formulation

5.3.1 NE-VCIP Problem Setting

The objective of the NE-VCIP problem is to achieve the minimal cost while satisfying the input demands.

Given:

• A physical Cloud computing infrastructure, modeled as a weighted bi-directional graph $G(V, E)$, V is the set of DCs with a set of computing resources and their unit costs, E is the set of network links. Each DC is described as a tuple data center $= (DC_v, C_v, S_v, P_v, \alpha_v, \beta_v, \gamma_v, i_v, e_v)$ with the capacity and unit cost of types of resources in this data center, the meaning of each item in the tuple is described in Table 5.1. Each edge in E (fiber link) is described as a tuple $e = (u, v, d_{(u,v)})$, which indicates the link between DCs (u, v), and the link distance. Each fiber link has a spectrum capacity at two directions. The network in the modeled Cloud computing infrastructure is in multi-layered;

• A VCI demand d, is modeled as a weighted undirected graph $G^d(V^d, E^d)$, in which V^d is a set of VDCs with specified computing, storage and switch port requirements,

E^d is a set of weighted VLs that indicate the required bandwidth. Each VDC of V^d is described as a tuple VDC $= (d, v', RC_d^{v't}, RS_d^{v't}, RP_d^{v't})$, in which d indicates the demand ID and v' indicates the VDC ID of demand d; the meaning of other items in the tuple is described in Table 5.1. Each VL in E^d is described as a tuple $VL = (d, u', v', RB_d^{(u',v')t})$, in which u', v' indicate the two end VDCs of current virtual link.

• The cost for each optical amplifier (OA) that to be installed in the used fiber links, the cost per km per GHz of using the optical fiber, the cost of IP?MPLS and EON nodes, and the cost for (S)BVT at each IP/MPLS node for connecting optical layer node. All the cost will be described in the cost model in detail (Section 4.3).

• The modulation format for optical signals in EON layer.

Output:

• The mapping for the VDCs in each VCI demand to the physical DCs;

• The routing path for mapped VN with allocated FSs;

• The total cost for satisfying all demands.

Objective:

1. Minimize the total cost for satisfying all the VCI demands.

2. Maximize the total number of accepted VCI demands.

5.3.2 Network Model

In this work, we adopt the IP/MPLS-over-EON architecture for the cloud network as shown in Figure 5.1. In the IP/MPLS-over-EON network, the intermediate node along the routing path could be (1) a multi-layer node with both IP/MPLS and EON capability; (2) only a EON layer node if the transmitted optical signal is not needed to be processed by the IP/MPLS layer; (3) a patch field that only connecting the optical fibers such as optical amplifier if the transmitted optical signal is not needed

Table 5.1
Parameters

C_v, S_v, P_v	CPU, storage and switch port capacities in DC v, $v \in V$
$\alpha_v, \beta_v, \gamma_v$	The unit cost of CPU, storage and switch port in DC v
i_v, e_v	The cost of IP/MPLS, EON layer terminals in DC_v
b	The unit cost (per Gbps) of bandwidth resource
d_e	The distance of link e, $e \in E$
ST_d, ET_d	The start and end time of demand d
V^d	The set of VDCs by demand d
E^d	The set of VN-links by demand d
Bud_d	The budget of demand d
$RC_d^{v't}, RS_d^{v't}, RP_d^{v't}$	The required amount of CPU, storage and switch port for VDC v' by demand d in time slot t, $v' \in V_d$
$RB_d^{(u',v')t}$	The required amount of bandwidth of virtual link between VDC (u', v') by demand d
$T_d^{(u',v')t}$	The required number of frequency slots by demand d between VDC (u', v')
$deg_d^{v'}$	The degree of VDC v' in VCI topology by demand d
c_t	The cost of optical transponder, will be different according to different BVT/SBVT types

to be processed by neither IP/MPLS layer nor EON layer.

In the IP/MPLS layer, an electrical node which can be seen as an IP/MPLS router, consists of main building blocks: the basic node (including switching matrix, power supply and mechanics), line cards (LC), with a different number of ports for transceivers and the transceivers [117]. In the EON layer, a flexible EON node can be seen as a bandwidth variable wavelength cross-connect (BV-WXC), which is used to establish optical cross-connections with various frequency slot width. The BV-WXC which mainly consists of BVT and bandwidth-variable wavelength selective switch (BV-WSS) can provide both sub-wavelength and super-wavelength for the flexible optical network. The EON can provide a granularity of 12.5 GHz instead

of 50 GHz in current WDM systems. Optical transponders can adjust the optical signal transmission rate to the actual traffic demand, by expanding or contracting the bandwidth of an optical path (i.e. varying the number of sub-carriers) and by modifying the modulation format [118].

We investigate our NE-VCIP problem on two optical network models: the optical network model with BVTs (BVT-model) and the optical network model with SBVTs (SBVT-model). In BVT-model, we adopt BVT to set up light path for each data flow and we suppose that the required bandwidth of each data flow does not exceed the maximum data rate of the BVTs used in the EON layer. We will consider using BVTs with capacity of 10 Gbps, 40 Gbps, 100 Gbps and 400 Gbps. In SBVT-model, we adopt SBVT to set up light path for each data flow and we adopt the optical traffic grooming technology to maximize the spectrum utilization. We consider using SBVTs with capacity of 100 Gbps and 400 Gbps. We note that the maximum traffic rate is same in both models, which means that if we use 100 Gbps SBVTs in SBVT-model, the maximum capacity of the BVTs used in BVT-model cannot exceed 100 Gbps. In this case, for example, if a demand requires 20 Gbps bandwidth, in the BVT-model we will use a pair of 40 Gbps BVTs to set up light path for this demand and we know that the remaining 20 Gbps capacity of this pair of BVTs would be wasted; while in the SBVT-model, we will use a pair of 100 Gbps SBVTs and will slice a 20 Gbps logical sub-transponder for this demand, then the remaining 80 Gbps capacity can be used by other demands.

5.3.3 Cost Model

In this work, the cost we considered for the NE-VCIP problem comes from the rental cost for computing resources such as CPU and storage (noted as operating expenditure (OpEx) in this work), and the fixed cost for network equipments and fibers (noted as

capital expenditure (CapEx) in this work). For the OpEx, we refer to the Amazon EC2 cost model to give the unit rental cost (cost per resource unit per slot) of CPU, storage and bandwidth. For the CapEx, we refer to the cost model in [5] for the cost of IP/MPLS nodes, BVTs, optical amplifier, etc (shown in Table 5.2). We assume that the metro node in our topologies consists of a single-chassis router, which consists of a single shelf with 10 line-card slots. All the cost values in our work are normalized.

We assume that for BVT and SBVT with the same capacity, they have the same cost [114]. Therefore, for example, suppose the maximum data rate of transponders we used is 100 Gbps and a demand requires 50 Gbps bandwidth. In the BVT-model, since BVTs with capacities of 10 Gbps and 40 Gbps cannot satisfy the demand, we need to use a pair of 100 Gbps BVTs to set up light path the for this demand. Then the total cost of transponder use for this demand will be $2 \times CostBVT_{100}$. In the SBVT-model, we need to slice out 50 Gbps logical sub-transponders from a pair of 100 Gbps SBVTs to set up light path for this demand. Since the remaining capacity of this pair of SBVTs can be used by other demands, the total cost of transponder use for this demand will be $2 \times \frac{1}{2}CostBVT_{100}$. We can see the transponder cost savings when using SBVT from this example.

Table 5.2
Cost Model [5]

Component	Cost (normalized cost unit)			
IP/MPLS node	9			
BVT	2.5 (10 Gbps)	7.625 (40 Gbps)	20.625 (100 Gbps)	65.625 (400 Gbps)
Optical amplifier	5(reach 80 km)			
Fiber cost	0.02 per km per GHz			

5.3.4 MILP Model

We will describe the MILP formulations of VCI mapping while considering RSA in EON problem. In general the physical frequency filtering requires that various spectrum paths are separated in the spectrum domain by guard frequencies [119] when two spectrum paths share one or more common fiber links. In our problem, to simplify the model, we assume that the size of guard frequencies is zero.

5.3.4.1 Full-Fit Scenario

In the full-fit scenario, when given a set of VCI demands, the resource allocator needs to accept all demands and minimize the total cost of all demands. To construct MILP formulations, we define some variables as shown in the following.

- $x_d^{v'v}$, 1 if required VDC v' by demand d is mapped to DC v; 0 otherwise
- $y_{df(u',v')}^{(u,v)}$, 1 if the FS f is used on physical link (u, v), which is on the mapping path for virtual link (u', v') of demand d; 0 otherwise. $(u, v) \in E, (u', v') \in E^d$
- $COST_d$, The cost for demand d

Objective:

$$\text{Minimize} \sum_d Cost_d \tag{5.1}$$

$$Cost_d = \sum_{t,v',v} (RC_d^{v't} \cdot \alpha_v + RS_d^{v't} \cdot \beta_v + RP_d^{v't} \cdot \gamma_v) \cdot x_d^{(v',v)} \tag{5.2}$$
$$+ \sum_{v',v} (i_v + e_v) \cdot x_d^{(v',v)} \cdot deg_d^{v'} + \sum_{t,e'} RB_d^{e't} \cdot (b + 2c_t)$$
$$+ \sum_{e',e} y_d^{(e',e)} \cdot d_e \cdot comCost \cdot \left\lceil RB_d^{e't}/10 \right\rceil$$

where *comCost* integrates the OA and fiber using cost (Table 5.2) along the fiber links, $t \in [ST_d, ET_d], v' \in V^d, u, v \in V$

Computing Resource Capacity Constraints:

$$\sum_{d,v'} RC_d^{v't} \cdot x_d^{(v',v)} \leq C_v \tag{5.3}$$

$$\sum_{d,v'} RS_d^{v't} \cdot x_d^{(v',v)} \leq S_v \tag{5.4}$$

$$\sum_{d,v'} RP_d^{v't} \cdot x_d^{(v',v)} \leq P_v \tag{5.5}$$

where $t \in [ST_d, ET_d]$.

Resource Allocation Region Constraints:

$$\sum_{v} x_d^{(v',v)} = 1, \forall d \in D, v' \in V^d. \tag{5.6}$$

$$\sum_{v'} x_d^{(v',v)} \leq 1, \forall d \in D, v \in V. \tag{5.7}$$

Spectrum Continuity Constraint:

$$\sum_{f} y^{(u,o)}_{df(u',v')} - x^{u'u}_{d} \times T^{(u',v')t}_{d} = 0, \qquad y^{(i,u)}_{df(u',v')} = 0 \tag{5.8}$$

$$\sum_{f} y^{(i,v)}_{df(u',v')} - x^{v'v}_{d} \times T^{(u',v')t}_{d} = 0, \qquad y^{(v,o)}_{df(u',v')} = 0 \tag{5.9}$$

$$\sum_{f,j\neq v} y^{(i,j)}_{df(u',v')} \quad = \quad \sum_{f,j\neq u} y^{(j,o)}_{df(u',v')} \tag{5.10}$$

where $\forall i, o, j \in V, t \in [ST_d, ET_d]$. We indicate u, v are the source and destination nodes of the mapping route for VL (u', v').

Frequency Slot Consecutiveness Constraint:

$$(y^{(u,v)}_{df(u',v')} - y^{(u,v)}_{d(f+1)(u',v')} - 1) * (-N) \geq \sum_{f'} y^{(u,v)}_{df'(u',v')} \tag{5.11}$$

where $f \in [1, F-1], f' \in [f+2, F], u', v' \in V^d, u, v \in V$.

Frequency Slot Capacity Constraint:

$$\sum_{d,u',v'} y^{(i,o)}_{df(u',v')} \leq 1, \, \forall f, i, o \tag{5.12}$$

$$\sum_{d,f,u',v'} y^{(i,o)}_{df(u',v')} \leq FN, \, \forall i, o \in V \tag{5.13}$$

Equations 5.3–5.5 ensure that the assigned computing resources required to the demand cannot exceed the resource capacity of each node. Equation 5.6 guarantees

that one VDC of a VCI demand can only obtain resources from one physical DC. Equation 5.7 guarantees that a physical DC can only have at most one VDC of a demand to be assigned to itself. Equations in 5.8 guarantee that the number of output frequency slots from the source node equals to the required number of frequency slots, and no input flow to the source node. Equations in 5.9 guarantee that the number of input frequency slots to the destination node equals to the required input frequency slots, and no output flow from the destination node. Equation 5.10 ensures that the spectrum route uses the same spectrum(s) along the routing path. Equation 5.11 ensures that the employed frequency slots are consecutive in frequency domain. The FS consecutiveness constraint requires that, for a spectrum route, the allocated FSs are consecutive in frequency domain. This constraint can be equivalently converted to: if $y^{(i,o)}_{df(u',v')} = 1$ and $y^{(i,o)}_{d(f+1)(u',v')} = 0$, all FSs with index higher than $f+1$ will not be allocated to the VL (u', v') from fiber link (i, o). We introduce a large number N in this constraint. Equation 5.12 ensures that one frequency slot on an fiber link can only be used by one route in a time slot. Equation 5.13 ensures that the used frequency slots cannot exceed the spectrum capacity (noted as FN) of each fiber link.

5.3.4.2 Best-Fit Scenario

When the number of demands are increasing, there might be not enough resources for all demands, so we construct the best-fit MILP model. In the best-fit scenario, when given a set of VCI demands, the resource allocator will accept as many demands as possible to allocate resources for them, and then compute the total cost for resource allocation. (Blocking rate is what we cared about in the best-fit scenario.) Addition variables that are needed to construct the MILP model are listed in the following.

- $w^{v'}_d$, binary variable, 1 if VDC v' in demand d is accepted; 0 otherwise
- z_d, binary variable, 1 if demand d is accepted; 0 otherwise

- $AC_d^{v't}, AC_d^{v't}, AC_d^{v't}$, actual allocated the amount of CPUs/storage/switch ports for VDC V' of demand d in time slot t

Objective:

$$\text{Maximize} \sum_d z_d \tag{5.14}$$

$$AC_d^{v't} = RC_d^{v't} \cdot x_d^{(v',v)} \tag{5.15}$$

$$AS_d^{v't} = RS_d^{v't} \cdot x_d^{(v',v)} \tag{5.16}$$

$$AP_d^{v't} = RP_d^{v't} \cdot x_d^{(v',v)} \tag{5.17}$$

where $d \in D, v' \in V^d, v \in V, t \in [ST_d, ET_d]$.

$$\sum_v x_d^{v'v} = w_d^{v'}, \forall d \in D, v' \in V^d \tag{5.18}$$

$$w_d^{v'} = z_d, \forall d \in D, v' \in V^d \tag{5.19}$$

In the Best-Fit scenario, the objective (equation 5.14) is to maximize the total number of accepted demands. For constraints, except those in Section 5.3.4, we add

additional constraints. Equations 5.15–5.17 ensures that if VDC v' of demand d is mapped to physical DC v, the amount of actual allocated resources will be the same with the amount of required resources, otherwise zero. Equation 5.18 guarantees that if VDC v' of demand d is mapped to a physical DC, it means that this VDC is accepted. Equation 5.19 guarantees that if any VDC of a demand d cannot be mapped, the whole demand d will be droped.

5.4 Heuristic Algorithm

We propose a cost-optimized greedy heuristic for the NE-VCIP problem. Every VCI demand is generated randomly with start time, finish time in [0,24], with required bandwidth and computing resources. In our proposed heuristic, we do not separate the joint resource allocation into two phases: computing resource phase and bandwidth resource phase, but combine them together. In the traditional two-phase allocation process for computing resources and network resources, each VDC in a VCI demand needs to be mapped to a physical DC first according to the cost and availability of computing resources, then we look for the optical circuits with available bandwidth resources between mapped VDCs. However this approach involving considering computing resources first and network resources second, the so called two-phase method, has a deficiency. We found from previous experiments that the network resource is the bottleneck (compared to computing resources in each DC) to complete the joint resource allocation for demands. So the two-phase method may lead to high blocking rates due to lack of network resources along the optical circuit between the mapped VDCs. Compared to the two-phase method, in our new method, we map the first VDC of a VCI demand first, then we consider the network resources availability along the optical circuit between this VDC to its connected ad-

jacent VDC. We also need to consider if the destination physical DC of the optical circuit has enough computing resources for this adjacent VDC (as shown in the following heuristic description in the steps (2), (3) and (4) below). The detailed idea of the heuristic can be found in the following paragraph and in Algorithm 7.

The general ideas of the proposed cost-optimized greedy heuristic are: (1) Map the first VDC (e.g. $v1$) of a demand, map it to the DC (e.g. u) which has enough computing resources and has lowest resource unit cost; (2) check if $v1$ has connections with other VDCs (e.g. $v2$) in the VCI demand graph; (3) if yes, for each connection, the Dijkstra algorithm is adopted to find the shortest path $p(u, des)$ between u and every other DC, and sort the paths in distance ascending; if no, go to (5); (4) map $v2$ to des which is the destination DC of the shortest path if des DC has enough computing resources and all links along path p have required number of FSs; (5) continues until a VCI demand is processed, then go for the next demand. The algorithm details are described in the the following Algorithm 7 which is implemented in [37] of our work . We call the $Dijkstra$ algorithm whose time complexity is $O(|E| + |V|log|V|)$ in our proposed heuristic. The total time complexity of the proposed heuristic is $O((|E| + |V|log|V| + |V|^2)|D||V^d|)$.

5.5 Experimental Results and Analysis

We carry out the computations for ILP model (using IBM ILOG CPLEX Optimization studio) and cost-optimized greedy heuristic on a cluster node which has 2 CPUs/16 cores and 64GB memory with Linux system. Two network topologies are tested for the simulations: a 6-node topology (Google DC locations) shown in Figure 5.5 and NSFNET topology shown in Figure 5.6 with the distance in km. The experiments for BVT-model and SBVT-model are described in 5.5.1 and 5.5.2.

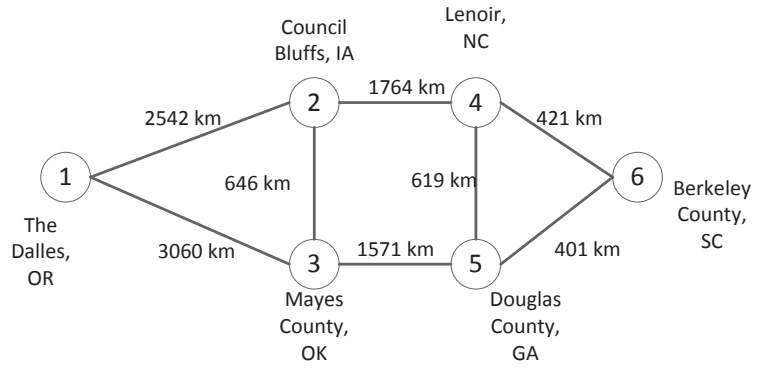


Figure 5.5
Google data center locations topology (6-node).

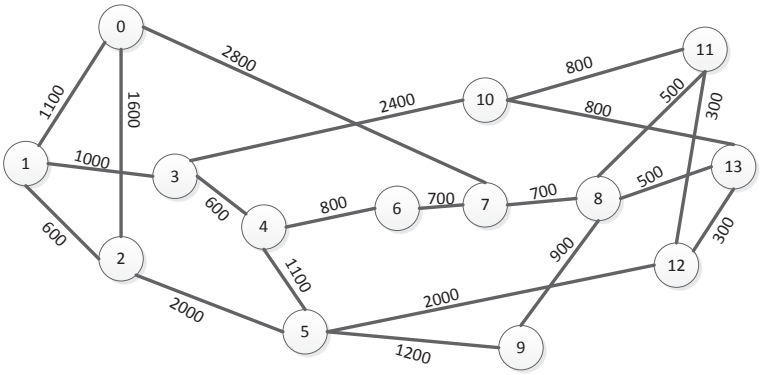


Figure 5.6
NSFNET network topology.

5.5.1 Results for BVT-model

For the experimental results showed in this section, we only consider using 10 Gbps BVT in our tested optical network model, which is part of the short version of our work [37]. In addition, the traffic demands are not categorized by their required bandwidth amount now. The correctness of the proposed greedy heuristic is verified by comparing its results with the Full-Fit ILP results for the small data set for the 6-node topology as shown in Table 5.3. When given one demand, the heuristic can give the near optimal solution compared with that of CPLEX and the computing time is much less than that of CPLEX. When given two or more demands, the CPLEX

method converges much slower to generate optimal solution compared to the heuristic method. In this case, in the later experiments, larger data sets are only tested by the heuristic on two topologies due to the slowness of ILP solution by CPLEX. We

Table 5.3
Cost and time comparison between CPLEX solver and heuristic for the Full-Fit

# of demands	Total Cost (normalized)		Running Time	
	CPLEX	Heuristic	CPLEX	Heuristic
1	1679.8	1692.31	1.2 hours	1.1384 s
2	4196.7461 (gap 18.56%)	4618.15	2 hours	1.1667 s
3	3547.4564 (gap 97.34%)	16788.3	12 hours	1.2685 s
4	*	17184.5	*	1.2841 s
5	*	17646.7	*	1.2943 s

Asterisks indicates that CPLEX was unable to find near-optimal solution within the time allowed.

compare the total cost and demand blocking rate for different data sets with different modulation formats. Due to the space limitation, here we only list the comparison results for the NSFNET topology; similar results are also obtained on the 6-node topology.

In Figure 5.7, all the demands can be accepted and allocated resources from the Cloud by the resource scheduler. By observing Figure 5.7 we can see that for different size of demand set, the total costs decrease with the modulation format order of BPSK, QPSK and 8-QAM, since the required number of FSs of each demand is reduced. But the total cost with 16-QAM increases compared to that with 8-QAM, although each demand has the least number of required FSs with 16-QAM. We note that the required number of FSs for a given bit rate is reduced sequentially with the modulation format orders of BPSK, QPSK, 8-QAM and 16-QAM; and at the same time the optical signal reaches are reduced along the same modulation format order.

In this case, more regenerators are needed along the optical path to regenerate the signals and the total cost will increase instead. It seems that it is a better choice to adopt 8-QAM modulation format to get a lower total cost from Figure 5.7.

With the limited resource capacities, the resource scheduler will drop some demands that cannot be satisfied when the number of demands increases. During the experiment we find that the network resource is a bottleneck compared with other computing resources. Almost every demand that is dropped is due to lack of continuous spectrum resource along its optical path. We observe the blocking rate of different sizes of demand set with four types of modulation formats as shown in Figure 5.8. It is obvious that for the modulation format order of BPSK, QPSK, 8-QAM and 16-QAM, the required number of FSs for each demand reduces significantly, so that the resource scheduler can accept much more demands. When the input number of demands reaches 1800, the blocking rates are nearly 13.6%, 0.7%, 0.022% and 0 with BPSK, QPSK, 8-QAM and 16-QAM respectively in Figure 5.8.

Thus, while considering the total cost and blocking rate together, we find that 8-QAM in our experiment performs best, which has the lowest total cost and has the blocking rate close to 0 for larger data sets.

5.5.2 Results for SBVT-model

To investigate what are the impacts on the total cost and blocking rate when involving SBVT and considering optical traffic grooming, we compare experimental results for BVT-model and SBVT-model. We consider using SBVTs with capacity of 100 Gbps and 400 Gbps as the maximum traffic data rate respectively. In this case, if we test traffics with maximum traffic rate of 100 Gbps, 1) in the BVT-model, BVTs with capacity of 10 Gbps, 40 Gbps and 100 Gbps will be adopted, 2) and in the SBVT-model, SBVTs with the capacity of 100 Gbps will be adopted, as described

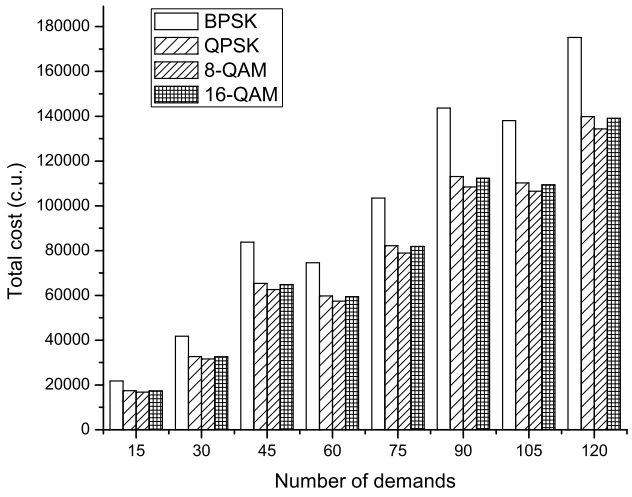


Figure 5.7
Total cost comparison for BVT-model (10G BVT).

in Section 5.3.2. In this part, we divide the randomly generated traffic demands into three categories by their bandwidth requirement as described in the following. Our experimental results show that the results tested on 6-node topology and NSFNET topology have the same trend, so here we only show the results on NSFNET topology.

First we investigate the total cost of solving NE-VCIP problems for submitted demands. Figure 5.9 compares the total cost for VCI demands with different bandwidth requirements (low bandwidth, medium bandwidth, high bandwidth) in both BVT-model and SBVT-model, and four types of formulation formats are considered as well. In Figures 5.9a and 5.9b, the maximum data rate that supported by transponder (BVT and SBVT) is 100 Gbps, while in Figure 5.9c the maximum data rate supported by transponder is 400 Gbps. It means that the required bandwidth of the VCI demands cannot exceeds the supported maximum data rate in both BVT-model and SBVT-model.

Figure 5.9a shows total cost comparison for demands with low bandwidth require-

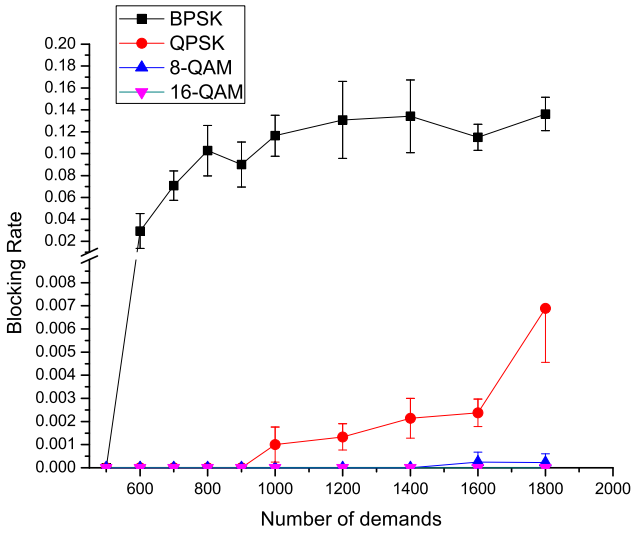


Figure 5.8
Blocking rate comparison for BVT-model (10G BVT).

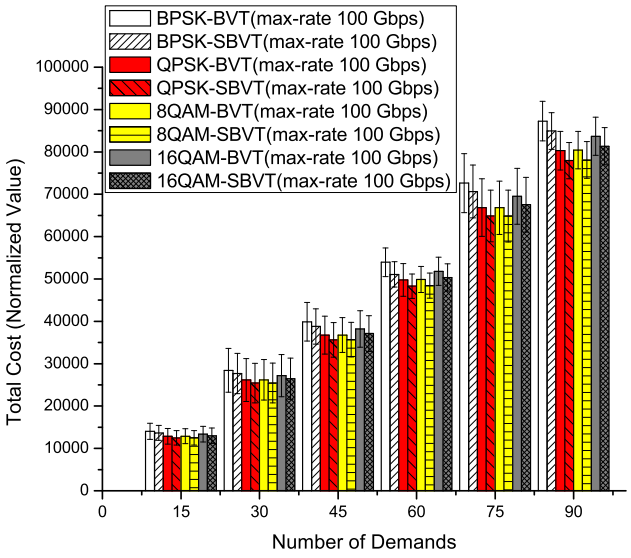
ments. We can observe that for different size of demand set, the total cost with BPSK modulation format is the highest in either BVT-model or SBVT-model. And the total costs with QPSK and 8-QAM modulation formats are nearly the same with each other, and they are the lowest compared to the total cost with other modulation formats. We analyze the data statistically that in BVT-model, the total cost with QPSK and 8-QAM modulation formats can be reduced by 7%∼10% compared to that with BPSK modulation format; and in SBVT-model, the reduction is around 8%. Moreover, we observe from 5.9a that no matter with which type of the modulation format, the total cost in SBVT-model is less that in BVT-model for different size of demand set, and the reduction is around 3%.

Figure 5.9b shows the total cost comparison for demands with medium bandwidth requirements, which has the same trends with that in 5.9a. In addition, the analyzed data shows that, compared to BPSK modulation format, the total cost with QPSK/8-QAM modulation formats can be reduced by 18% and 19% in BVT-model

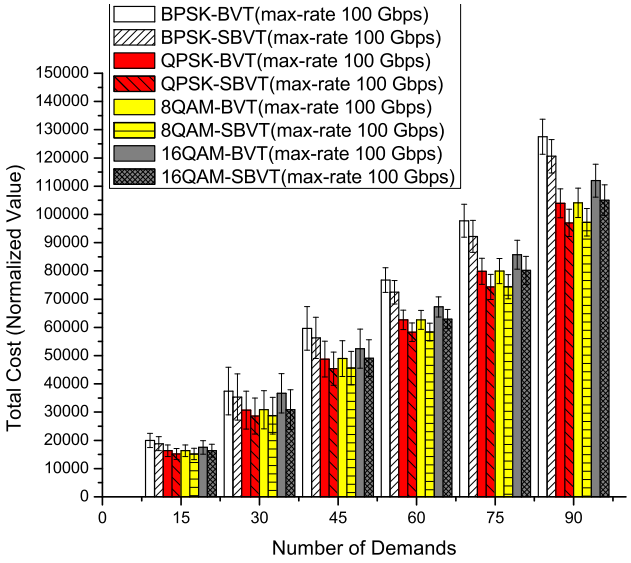
and SBVT-model respectively. Moreover, we observe that the total cost in SBVT-model is reduced by 5% $\sim$ 7.8% compared to that in BVT-model for four types of modulation formats.

Figure 5.9c shows that the total cost comparison for demands with high bandwidth requirements. We can see the the cost decreases with the modulation format order of BPSK, QPSK, 8-QAM and 16-QAM. So for the demands with high bandwidth requirements, they will use the least number of frequency slots under 16-QAM modulation format, and thus will have the lowest total cost (although the reach distance of 16-QAM is the shortest) than that under other modulation formats. The total costs with QPSK, 8-QAM and 16-QAM are reduced by 26%$\sim$28.8%, 35%$\sim$37%, 40%$\sim$42% respectively compared to that with BPSK in BVT/SBVT-models. In addition, we observe that the total cost in SBVT-model is reduced by 6.5%$\sim$10.3% compared to that in BVT-model for four types of modulation formats. To sum up, from Figure 5.9 we can see that the using of SBVTs can reduce the total cost of solving NE-VCIP problem, and such reduction will be more significantly along with the increase of bandwidth requirement.

After the cost analysis, we compare the blocking rate for the demands with different bandwidth requirements (low, medium and high) in Figure 5.10. We test different data sets with different number of demands (from 5 demands in a data set, to 2800 demands in a data set). When the data set has less than 200 demands, no traffic blocking happens. All demands will be processed by allocating required computing and network resources. When the number of demands in a data set goes up to 2800, the blocking rate reaches the relative threshold under each modulation format. The blocking rates in BVT-model and SBVT-model are same since the blocking rate is mainly decided by the computing resource availability in physical data centers and frequency slots availability in optical fiber for the network resource part. Figures



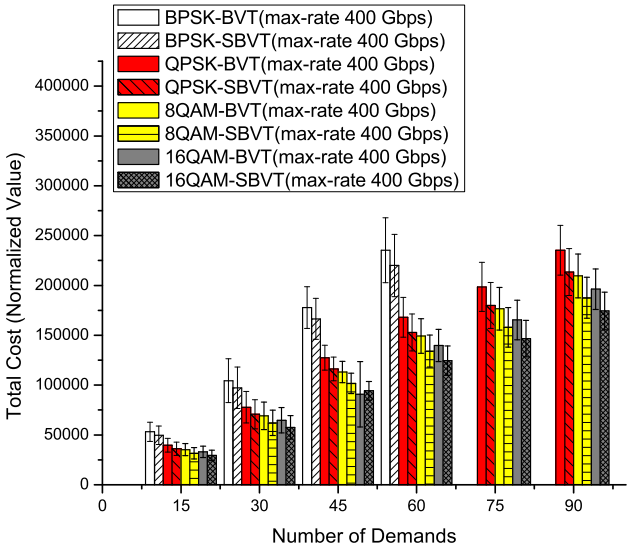
(a)



(b)

Figure 5.9
Cost comparison in BVT/SBVT models under different modulation formats for demands with bandwidth requirements in: (a) Range (0 Gbps, 40 Gbps], (b) Range (40 Gbps, 100 Gbps], (c) Range (100 Gbps, 400 Gbps].

5.10a, 5.10b and 5.10c shows that no matter whether the bandwidth requirement of demands is low or high, BPSK modulation format has the highest blocking rate,



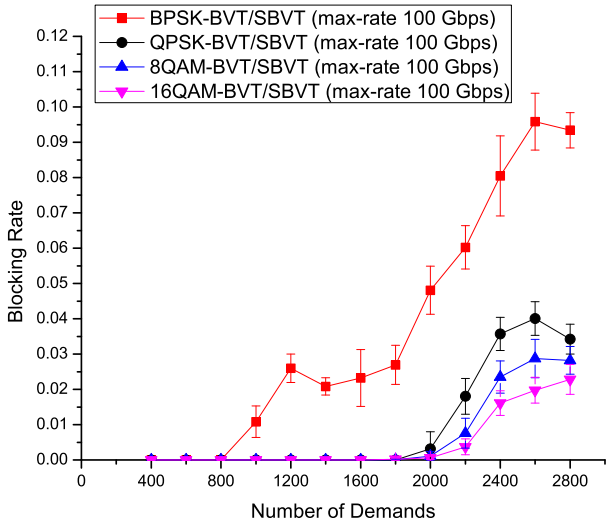
(c)

Figure 5.9
Cost comparison in BVT/SBVT models under different modulation formats for demands with bandwidth requirements in: (a) Range (0 Gbps, 40 Gbps], (b) Range (40 Gbps, 100 Gbps], (c) Range (100 Gbps, 400 Gbps].

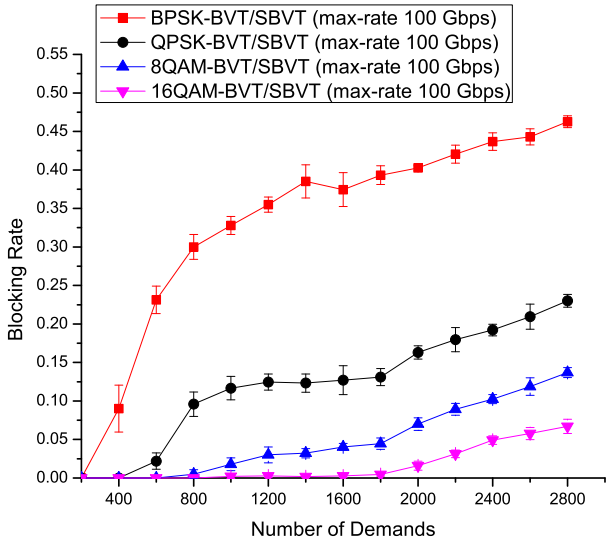
and the blocking rate decreases with the modulation format order of BPSK, QPSK, 8-QAM and 16-QAM. We here note the blocking rate comparison in three stages: QPSK compared to BPSK, 8-QAM compared to QPSK, and 16-QAM compared to 8-QAM. We analyze the data in Figure 5.10 and observe that for demands with low and medium bandwidth requirements, the blocking rates decrease very significantly, and they are decreased by 67%∼84%, 50%∼66.1%, and 35%∼70% for the three stages respectively. For the demands with high bandwidth requirements, the blocking rates are decreased by 26%, 28.4% and 24% for three stages respectively, which are not so significantly as that for demands with lower bandwidth requirements.

5.6 Conclusion

In this work, we propose and investigate the NE-VCIP problem in IP-over-EON network architectures. An ILP mathematical model is constructed and a cost-optimized greedy heuristic is developed to solve the NE-VCIP problem. Different modulation formats that are adopted in the EON layer will have different results for the total cost and the demand blocking rate for the same data set size. So in order to minimize the total cost and also obtain a better system performance (e.g., low blocking rate, high resource utilization), a trade-off needs to be considered between the two. We conclude that for demands with lower bandwidth requirements, adopting 8-QAM in EON layer would be a suitable choice for the resource scheduler to obtain the lowest total cost and also obtain an acceptable lower blocking rate. For demands with high bandwidth requirements, adopting 16-QAM would be a better choice to obtain lower total cost and blocking rate. In addition, in this work we also investigate the effect on total cost and blocking rate while using SBVTs to set up a light path for data transfer and considering traffic grooming technologies. In our experiments, we conclude that the the use of SBVTs (compared to BVTs) and traffic grooming technology will reduce the total cost no matter which one of the four modulation formats are adopted, and this reduction is more significant for the demands with high bandwidth requirements. In future work, we will consider implementing more sophisticated heuristics, such as Tabu search meta-heuristic, to solve the NE-VCIP problem.



(a)



(b)

Figure 5.10
Blocking rate comparison in BVT/SBVT models under different modulation formats for demands with bandwidth requirements in : (a) Range (0 Gbps, 40 Gbps], (b) Range (40 Gbps, 100 Gbps], (c) Range (100 Gbps, 400 Gbps].

Algorithm 7 Cost-optimized Greedy Algorithm

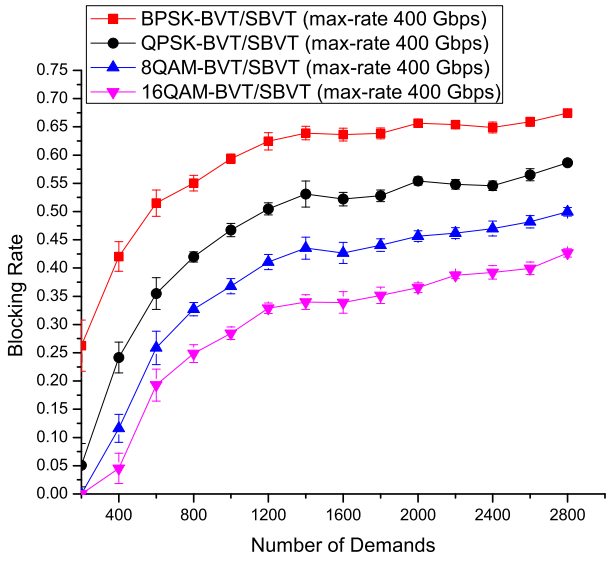
Input and Initializations:

$G(V, E)$ //network topology
D //demand set
$G^d(V^d, E^d)$ //virtual topology of demand d
$Cost_d = 0$; //initial cost for demand d is 0

Output:

Minimize $\sum_d Cost_d$.

```
1: Sort V in ascending order of unit cost for computing resources
2: for all d ∈ D do
3:   for all v^d ∈ V^d do
4:     if v^d is not been processed then
5:       Map v^d on node v with enough resources for v^d;
6:       Allocate computing resources from v for v^d;
7:       Update Cost_d;
8:     end if
9:     Construct set P_v;
10:    for all u ∈ V, u ≠ v do
11:      Find shortest path p(v, u), add p(v, u) into P_v;
12:    end for
13:    Sort paths in P_v in ascending order of distance;
14:    for all u^d ∈ Adjacent(v^d) do
15:      if u^d is not been mapped then
16:        for all p(v, u) ∈ P_v do
17:          if enough resources on u for u^d AND enough spectrums on p(v, u) for
             (v^d, u^d) then
18:            Map u^d on node u;
19:            Allocate computing resources for u^d;
20:            Allocate spectrums and update Cost_d;
21:          end if
22:        end for
23:      else
24:        Check spectrums on route p(v, x);{suppose u^d is already mapped to x ∈
           V}
25:        if p(v, x) has enough spectrums then
26:          Allocate spectrums and Update Cost_d;
27:        else
28:          Drop demand d;
29:          Release the assigned resources for d;
30:        end if
31:      end if
32:    end for
33:  end for
34: end for
35: return ∑_d Cost_d
```



(c)

Figure 5.10
Blocking rate comparison in BVT/SBVT models under different modulation formats for demands with bandwidth requirements in : (a) Range (0 Gbps, 40 Gbps], (b) Range (40 Gbps, 100 Gbps], (c) Range (100 Gbps, 400 Gbps].

Chapter 6

Virtualized Cloud Services Provisioning in Hybrid Optical Data Center Networks

6.1 Introduction

The development of Cloud computing technology has led to the growth in the size of the data centers. Data centers may contain tens of thousands of computers with significant bandwidth requirements. The traditional data center network (DCN) architecture is tree-based hierarchy structure which consists of either three-level or four-lever trees of Ethernet switches and routers. In a typical three-level DCN design (figure 6.1), the core level is at the root of the tree, the aggregation levels are in the middle and the edge level is at the leaves of the tree [76]. The Ethernet layer packet switching solutions are adopted in the traditional tree-based hierarchy DCN architecture to support the data center network communication. However along with the increasing bandwidth requirements by the big data applications running on the Cloud platform, such packet switched tree-based DCN architecture would not provide high performance services in the future. Other packet switching DCN architectures such as Fat-Tree structure [120] and BCube structure [121] also meet such bottleneck.

Involving the optical interconnection networks for the DCN can satisfy the high bandwidth requirements by the big data applications while consuming less power [122]. Optical interconnects support both packet switching and circuit switching.

Circuit switching mainly target that DCN in which long-term bulky data transfers are required between racks. Packet switching optical network can achieve much faster switching times than circuit switching. Thus the packet switching optical network fits better to DCN with burst traffic [123]. Furthermore, introducing optical network to the DCN can help to support such big data applications with high bandwidth requirements and with diverse communication patterns [124]. We can see the optical network is playing an essential role in the current DCN design and will become more important in the future DCNs.

With the increasing requirements of bandwidth resources, the traditional computing resource allocation such as VM allocation in data center needs to involve the network resource allocation to satisfy customers' requirements. The network resource required by the customers are usually used for connecting the customers' private Cloud to the VMs reserved by the customers in DCs, or for connecting the VMs reserved by customers on public Cloud(s). For a Cloud service provider, providing computing resources alone to the customers is not sufficient as a competitive advantage. Other factors have gained more weight, such as offering network solutions to customers. Network performance and resource availability can be the tightest bottleneck for any Cloud [125]. Optical networks with the characteristics of high throughput, low latency and low power consumption, can be adopted to provide the bandwidth guaranteed network service in Cloud.

Going back to the resource provisioning service, the cloud providers have moved from simply supplying computing resources to supplying multiple types of services, including networking, elastic caching, database, analytics [15]. When deal with resources sharing among multiple customers, the performance isolation becomes a challenge for the cloud providers. Significant works have been done to investigate the performance isolation on different aspects, such as Cloud CPU performance isolation

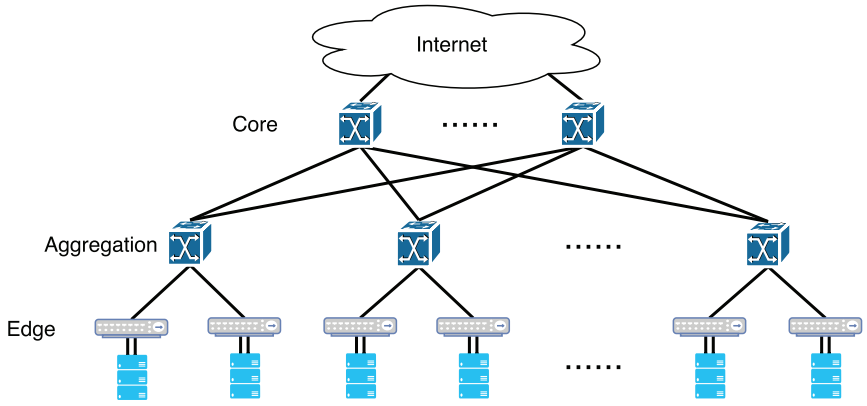


Figure 6.1
A typical three-level tree-based DCN architecture.

[16], end-to-end performance isolation [18] and Cloud storage performance isolation [17]. The abstraction of a dedicated virtual data center (VDC) is proposed in such investigations to deal with virtual resource provisioning and isolation in Cloud.

In this work, we are going to investigate the network-aware resource orchestration in data center with different types of optical data center network architectures. Section 6.2 describes data center network architecture that have been proposed in other works. Section 6.3 discusses the work that have been done on investigating the VM placement and routing problems in data center. Section 6.4 introduces the problem settings and the three DCN architectures we discussed in this work. In the following Sections 6.5–6.7, the mixed integer linear programming (MILP) and mixed integer quadratic programming (MIQP) formulations for the mathematical models based on three DCN architectures correspondingly. Section 6.8 presents the experimental results and analysis. Section 6.9 gives the conclusion for this work.

6.2 Data Center Network Architectures

In this section, we briefly review research works that have been investigated to design and implement the data center network architectures. As introduced in Section

6.1, the commodity data center networks are constructed mainly based on the tree-based hierarchy structure. Along with the increasing big data applications running on Cloud, the supporting for high performance network service of data centers is becoming more and more important. Therefore more and more work are being done on introducing optical network to the data center network architectures to strengthen the capability of providing high bandwidth to correlated applications.

The work in [126] proposed a hybrid packet and circuit switched data center network architecture (HyPaC) which augments the traditional hierarchy of packet switches with a high speed, low complexity, rack-to-rack optical circuit-switched network to supply high bandwidth to applications. The emulation experiments were carried out to show that the HyPaC architecture can provide large benefits to unmodified popular data center applications at a modest scale. Another work in [127] also presented a hybrid electrical/optical switch architecture, called HELIOS, for the DCNs. HELIOS structure can deliver performance comparable to a non-blocking electrical switch with significantly less cost, energy, and complexity. The trade offs and architectural issues were explored in the work in realizing these benefits.

Besides the hybrid architectures for DCNs, another type of DCN architecture in work [4] [128] was proposed. OSA, a novel Optical Switching Architecture was designed, implemented and evaluated in work [4]. The designed OSA can dynamically change it topology and link capacities to achieve unprecedented flexibility to adapt to dynamic traffic patterns. Another optical switching architecture for DCN, named OpenScale, was proposed in work [129]. The idea of "small world" topology is employed to construct a flexible and highly scalable network. Simulations verified that proposed architecture can achieve eminent scalability.

In the following sections, we will introduce three DCN architectures that are adopted in our work in detail.

6.3 VM Placement and Routing in Data Center

We need to consider the VM placement as well as the routing issues when we target the network-aware resource provisioning in data center. A lot of work have been done to investigate the VM placement and routing in data center and Cloud systems. The work in [130] addressed the network-aware VM placement problem by trying to allocate a placement that not only satisfy the predicted communication demand but is also resilient to demand time-variations. The authors introduced several heuristics to solve this new optimization problem called Min Cur Ratio-aware VM placement (MCRVMP). Another work in [131] focused on high performance algorithms to solve the VM placement problem in a network Cloud. A *shadow routing* based approach was proposed for the VM allocation in a large and heterogeneous data centers or server clusters and the good performance, robustness and adaptability of the algorithm was proved analytically and through simulations.

Moreover, a more recent work in [132] focuses on the management of network resources by exploiting joint route selection and VM placement. The paper formalizes the joint route selection and VM placement problem as a static optimization problem and further solve the dynamic version of this problem with the goal of optimizing the long-term-averaged system performance. In our work, we will consider the VM placement in data center as well as routing problems in optical layer network for the network-aware virtual resource provisioning.

6.4 Problem Settings

6.4.1 VDC Demand Submitted by User

A virtual data center (VDC) that a tenant required is an abstraction which afford the tenant convenience of using resources on the shared cloud environment. Each VDC demand consists of VMs which have specified configuration each (number of CPU cores, memory amount) and virtual links that interconnect the VMs. We model a VDC demand as a weighted undirected graph, noted as $G(V, L)$ shown in Figure 6.2.

- D: demand set, $d \in D$

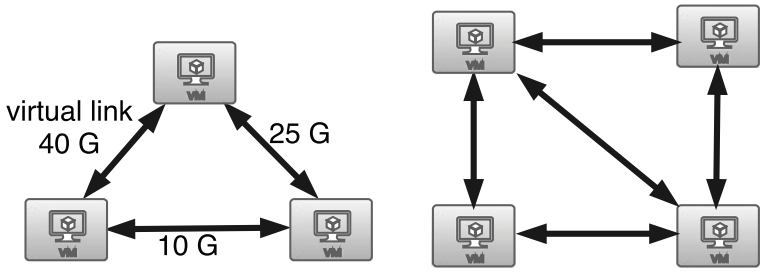


Figure 6.2
Two possible model of a VDC request.

- $G_d(V_d, L_d)$: the modeled weighted undirected graph of demand d, in which V_d is the set of VMs required by demand d, L_d is the set of virtual links that interconnect the VMs

- v_d: a VM of a demand d, $v_d \in V_d$

- l_d: a virtual link of a demand d, $l_d \in L_d$

- RC_{dv}: CPU resource required by VM v of demand d.

- RM_{dv}: memory resource required by VM v of demand d.

- B_d: communication matrix of demand d. $B_d^{l_d}$indicates the bandwidth requirements for the data transmission by virtual link l_d of demand d. A virtual link l_d is represented by a tuple $l_d = (d, VM_i, VM_j, b)$, to indicate that b GB bandwidth is required by this

virtual link between VM_i and VM_j (bi-directional). The following example (Figure 6.3) shows the communication matrix of a demand which has four VMs.

$$\begin{array}{c} \\ VM1 \\ VM2 \\ VM3 \\ VM4 \end{array} \begin{array}{c} \begin{array}{cccc} VM1 & VM2 & VM3 & VM4 \end{array} \\ \begin{pmatrix} 0 & 30 & 0 & 100 \\ 30 & 0 & 50 & 0 \\ 0 & 50 & 0 & 120 \\ 100 & 0 & 120 & 0 \end{pmatrix} \end{array}$$

Figure 6.3
Communication matrix of a demand

6.4.2 Physical Resources in Data Center

In the data center architecture, physical servers are grouped by racks. The servers in a rack are connected to the top-of-rack (ToR) switch of this rack. For the physical resources in data center, we model that each server has certain number of CPU cores and certain capacity of memory. The resource allocator looks for servers that have available resources, and allocate resources to build VMs with required CPU cores and memory capacity on top of the servers for the demands. To reduce the communication overhead between racks, in general, the resource allocator will assign the resources from servers within a same rack for a demand. We define some parameters to described the physical resources in data center in the following.

- N: the number of ports of the MEMS matrix optical switch.
- R: the set of ToR switches in DCN; the identifier of each ToR switch is r, $r \in R$.
- H: the set of physical servers (hosts) in data center; the identifier of each server is h, $h \in H$.
- P: the number of ports of each ToR switch that connect to multiplexer, a port is associated with a wavelength; also indicates the number of servers belonged to a rack.
- Cp_h: the number of CPU cores each physical servers has.

- Cm_h: the amount of memory capacity each physical servers has.

- k: the degree of each ToR switch, indicates that each ToR can communicate with other k ToR switches simultaneously.

- C_{port}: port capacity of ToR switch (ports connect to server side).

- C_λ: wavelength capacity of ToR switch (ports connect to multiplexer side).

- C_{packet}: the bandwidth capacity of the link (packet switching) between ToR switch and aggregation switch.

- W: the number of wavelengths on the connection between ToR switches in both directions ($ToR_i \rightarrow ToR_j$ and $ToR_j \rightarrow ToR_i$)

From the parameters described above for the DCN, more information can be obtained about the DCN. For example, with the given identifier of a server h and the number of servers P in each rack, we can get the rack this server belongs to with equation $r = \lceil h/P \rceil$.

6.4.3 Optical Data Center Network Architecture Adopted

In this work, we solve the resource provisioning problems for three different types of DCN architectures and compare the results of provision resources for the demands on different DCN architectures.

A. DCN with fully connected non-blocking matrix optical switches architecture

Suppose a $N \times N$ MEMS matrix optical switch with fully non-blocking, all optical cross-connect configuration (Figure 6.4 shows an example of a 4×4 MEMS matrix fully connected optical switch example). N ToR switches are connected to the MEMS. In this case, each ToR switch can communicate with other $N-1$ ToR switches directly at the same time.

B. DCN with c-through architecture

The c-through DCN architecture (Figure 6.5) is a hybrid packet switching and

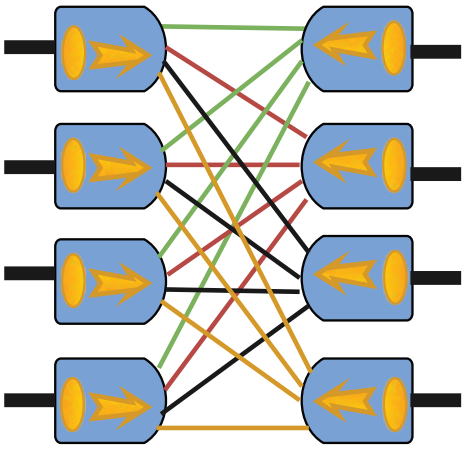


Figure 6.4
Fully connected non-blocking 4×4 MEMS matrix optical switch.

circuit switching (HyPaC) DCN architecture which augments the traditional hierarchy of packet switches with a high speed, low complexity, rack-to-rack optical circuit-switched network to supply high bandwidth to applications [126]. The c-through consists of two parts, the packet-switched tree-based DCN part with Ethernet switches and the high-speed rack-to-rack circuit-switched optical networks with reconfigurable optical paths. The HyPac DCN architecture benefits many kinds of applications, especially those with bulk transfer components, skewed traffic patterns, and loose synchronization [126].

C. DCN with OSA architecture

Based on the optical switch architecture (OSA) for DCNs [4], in this work we will adopt a 160-port optical switching matrix and 40 ToR switches that supported 1280 servers in total. Each ToR electrical switch has 64 ports with fixed supported data rate of 10 GE, in which 32 ports are connected to 32 servers that under this ToR, and the other 32 ports are connected to the optical components (including multiplexer/demultiplexer components and wavelength selective switch (WSS)). Each port that connected to the optical components has a transceiver associated with unique wavelength

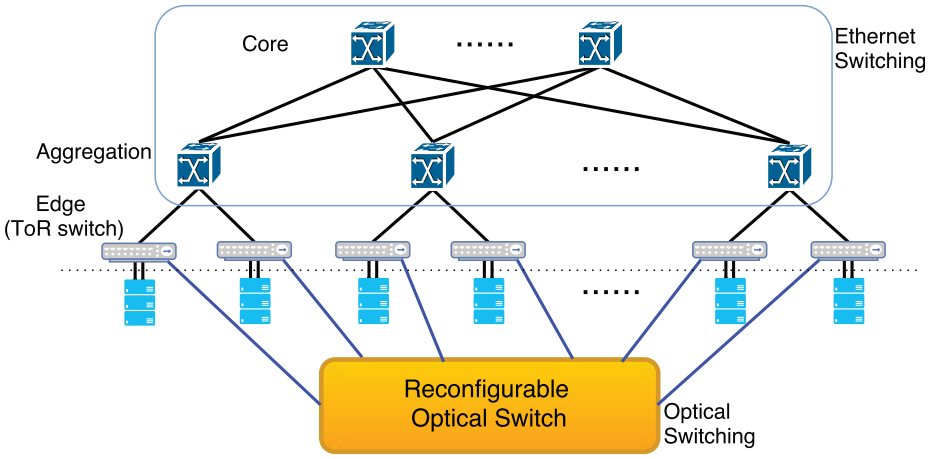


Figure 6.5
C-through HyPaC DCN architecture.

to send/receive data. If we suppose each ToR is 4-degree, then each ToR will be directly switched to another four ToRs through Micro-Electro-Mechanical Systems (MEMS), which means that a ToR can communicate with another four ToRs at the same time. The reconfiguration time for MEMS such devices is a few milliseconds [133].

The MEMES matrix optical switch can be re-configurable means that the optical network topology connecting racks changes. The initialized configuration of the MEMS is related with the current traffic flows in the data center. The optical configuration manager collects the traffic measurements and determines how optical paths should be configured. The rack pair that has high traffic flows are connected directly (one-hop) through the MEMS matrix. The rack pair that has low traffic flows can be connected through multiple hops. The goal of such configuration is to maximize the number of This configuration problem be formulated as a maximum weight perfect matching problem.

Let us look at the connection between one ToR and the MEMS based on previous example. When sending data, the multiplexer groups the data from all 32 ports of

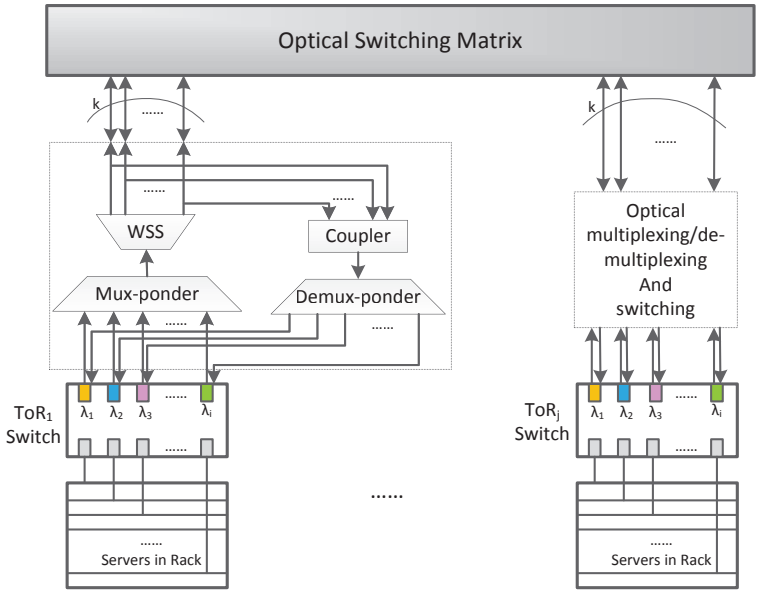


Figure 6.6
The OSA architecture [4].

the ToR with their unique wavelengths into one fiber and send them to the 1×4 WSS. The WSS will split the 32 wavelengths into four groups and each group has its own fiber to transmit the data in that group. The fibers are connected to the corresponding ports of the MEMs switch through optical circulators.

6.5 MILP for Fully Connected Non-blocking MEMS DCN Architecture

6.5.1 Parameters for the Fully Connected MEMS DCN Architecture

In the fully non-blocking MEMS DCN architecture, each ToR switch connects to other ToR switches directly through the MEMS matrix optical switch. So suppose for a $N \times N$ fully non-blocking MEMS, and we have N ToR switches in the data center. Thus for the DCN architecture, the degree k of each ToR switch (degree concept is defined in Section 6.4.2) would be $N-1$, in order to realize non-blocking communication with all other $N-1$ ToR switches simultaneously.

6.5.2 Mixed Integer Linear Program

Input: A set of VDC demands (D) from users as described in Section 6.4.1. The physical resources in the data center as described in Section 6.4.2 and the fully non-blocking MEMS DCN architecture as described in Section 6.4.3.

Output: The allocated computing resources from servers and bandwidth resource from network connections in data center. The wavelength utilization on each fiber link.

Variables:

- f_{dv}^{h}: equals 1 if VM v of demand d is assigned computing resources from physical server h; 0 otherwise.

- F_{dv}^{i}: equals 1 if VM v of demand d is assigned computing resource from rack i, which can be seen under ToR_i as well; 0 otherwise.

- $T_{l_d}^{ij}$: equals 1 if virtual link l_d is mapped to physical fiber connection between two ToRs in the direction of $ToR_i \rightarrow ToR_j$. In the fully non-blocking MEMS DCN architecture, the connection between any two ToRs is one-hop connection.

- $Tw_{l_d}^{ijw}$: equals 1 if wavelength w on fiber connection $ToR_i \rightarrow ToR_j$ is used for virtual link l_d mapping.

- Γ_d equal 1 if demand d can be processed (allocated required resources); 0 otherwise.

Objective: Maximize $a * \sum_{d \in D} \Gamma_d + b * \sum_{d \in D} Cost_d$

Constraints:

1. A single VM should be allocated from a server in one rack.

$$\sum_{h} f_{dv}^{h} \leq 1, \quad \forall d, v. \tag{6.1}$$

$$\sum_{r} F_{dv}^{i} \leq 1, \quad \forall d, v. \tag{6.2}$$

$$\sum_{h} f_{dv}^{h} = \sum_{i} F_{dv}^{i}, \quad \forall d, v \tag{6.3}$$

where $i = \lceil h/P \rceil$ as discussed above in Section 6.4.2.

2. The assigned resources from each server should not exceed its resource capacity.

$$\sum_{d,v} (f_{dv}^{h} \cdot RC_{dv}^{h}) \leq Cp_{h}, \quad \forall h. \tag{6.4}$$

$$\sum_{d,v} (f_{dv}^{h} \cdot RM_{dv}^{h}) \leq Cm_{h}, \quad \forall h. \tag{6.5}$$

3. For a ToR switch, the total in-flow to/out-flow from current ToR switch should equal the total number of VM mappings on this ToR switch.

$$\sum_{j,l_d} T_{l_d}^{ij} + \sum_{i,l_d} T_{l_d}^{ji} = \sum_{d,v} F_{dv}^{i}, \quad \forall i \tag{6.6}$$

4. Only if virtual link l_d is mapped to the fiber connection from ToR_i to ToR_j, the wavelength can be used for the virtual link on this fiber connection.

$$Tw_{l_d}^{ijw} <= T_{l_d}^{ijw}, \quad \forall i, j, w, i \neq j. \tag{6.7}$$

5. The needed wavelength amount is restricted by the required bandwidth and wavelength capacity.

$$\sum_{w} Tw_{l_d}^{ijw} = B_{d}^{l_d}/C_{\lambda}, \quad \forall i, j, d, l_d. \tag{6.8}$$

6. The used number of wavelengths should not exceed the total number of wavelengths W on every fiber link that connects ToRs.

$$\sum_{l_d,w} Tw_{l_d}^{ijw} \leq W, \quad \forall i,j,i \neq j. \tag{6.9}$$

7. A wavelength on a fiber connection can only be used by one virtual link at a time.

$$\sum_{l_d} Tw_{l_d}^{ijw} <= 1, \quad \forall i,j,w,i \neq j. \tag{6.10}$$

8. The demand is accepted only when the resource allocation for all VMs of this demand is successfully.

$$\sum_{h} f_{dv}^{h} \leq \Gamma_d, \quad \sum_{h} f_{dv}^{h} \geq \Gamma_d, \quad \forall d,v \tag{6.11}$$

6.6 MIQP for Hybrid Packet and Circuit Switched DCN Architecture

6.6.1 Parameters for the HyPaC DCN Architecture

In the hybrid packet and circuit switched DCN (HyPaC), in addition to the traditional hierarchy of packet switches, the high speed, rack to rack optical circuit switched network is adopted to offer high bandwidth to applications. The optical network part is implemented through the re-configurable optical switch. The reconfiguration of the optical switch is based on the current traffic flow in DC, in order to carry maximum number of traffics.

The optimal configuration for the optical switch could be modeled as the bipartite graph maximum weight perfect matching problem. We can model the bipartite graph

as a complete weighted bipartite graph with bipartition $(R1, R2)$. The nodes in each bipartition are same and represent the racks in the DC. The weight of each edge represents the required bandwidth between two racks. The weight of the edge will be zero if the edge connects two same racks or there is no bandwidth requirement between two different racks obviously.

6.6.2 Mixed Integer Quadratic Program

In the MIQP formulation, our target is to allocate resources to as many demands as possible. So we need to look for VMs with enough available computing resources for the demands. In addition, we need to configure the MEMS matrix as well to adjust the topology and to find routes between ToR-pairs, in order to carry as many traffic demands as possible while allocating required bandwidth resource for these demands.

Input: A set of VDC demands (D) from users as described in Section 6.4.1. The physical resources in the data center as described in Section 6.4.2 and the HyPaC DCN architecture described in Section 6.4.3.

Output: The MEMS matrix configuration topology. The allocated computing resource from servers and bandwidth resource from network connections in data center.

Variables: Some of the variables used in the following MILP formulations are already defined in Section 6.5.2. In addition, we defined some new variables to be used.

- M_{ij}: bandwidth traffic matrix, indicates the desired bandwidth from ToR_i to ToR_j.
- l_{ij}: equals 1 if ToR_i is connected to ToR_j through MEMS matrix optical switch directly (bi-direction connection, $l_{ij} = l_{ji}, i \neq j$), 0 otherwise.

Objective: Maximize the total number of accepted demands and maximize the

bandwidth traffic flow (optimal MEMS configuration):

$$Maximize \sum_{d} \Gamma_d + \sum_{i,j} M_{ij} \cdot l_{ij}. \tag{6.12}$$

1. A ToR switch can only connect to another ToR switch at a time.

$$\sum_{j} l_{ij} = 1, \quad \forall i, \qquad \sum_{i} l_{ij} = 1, \quad \forall j. \tag{6.13}$$

2. The desired bandwidth between any two ToRs is exactly the total bandwidth requirement by virtual links mapped to the route between two ToRs.

$$\sum_{d,l_d} T_{l_d}^{ij} \cdot B_d^{l_d} = M_{ij}, \quad \forall i, j, i \neq j \tag{6.14}$$

3. The total desired bandwidth of the bandwidth traffic matrix cannot exceed total bandwidth requirements by all demands.

$$\sum_{i,j} M_{ij} \leq \sum_{d,l_d} B_d^{l_d}, \quad i \neq j. \tag{6.15}$$

4. The demand traffic between two ToR racks through packet-switching cannot exceed the capacity of electrical network link that connecting ToRs.

$$\sum_{d,l_d,j} T_{l_d}^{ij} \cdot (1 - l_{ij}) \leq C_{packet}, \qquad \sum_{d,l_d,i} T_{l_d}^{ij} \cdot (1 - l_{ij}) \leq C_{packet}. \tag{6.16}$$

Other constraints on the computing resource allocation are same with those already presented in Section 6.5.2 (see constraints 1, 2, 4–8 in Section 6.5.2).

6.7 MILP for OSA DCN Architecture

6.7.1 Parameters for the OSA DCN Architecture

In the OSA DCN architecture, we adopt a re-configurable N-port MEMS matrix optical switch. The N ports are divided into N/k groups. k is the degree of each ToR switch, which means a ToR switch can communicate with other k ToRs simultaneously. As described in Section 6.4.2, we suppose a ToR switch has $2 \times P$ ports in total, in which P ports connect to P servers in this rack and the other P ports connect to multiplexer, each port is associate with a wavelength.

In order to facilitate the construct the DCN architecture for mixed integer linear programming (MILP) model and heuristics in the following sub-sections, we define some parameters in the following section to described the OSA DCN architecture, in addition to the ones we represent in the above sections.

6.7.2 Flexible bandwidth

In the optical switching architecture (OSA) for the data center network we adopted in this work, each ToR can be connected to other k ToRs directly through the optical switching matrix. Each fiber connecting a ToR to the optical switching matrix can support different bandwidth through carrying different number of wavelengths in a single fiber. For example, ToR_i wants to communicate with ToR_j with bandwidth B, and B is larger than the capacity of a single wavelength (Let us suppose the capacity of a single wavelength is w). In this case, ToR_i will use $\lceil p = B/w \rceil$ ports, each associate with a wavelength to support this request with B bandwidth requirement. These p wavelengths together with other wavelengths that for other ToRs communications are multiplexed into one optical fiber which is connected to the WSS through the

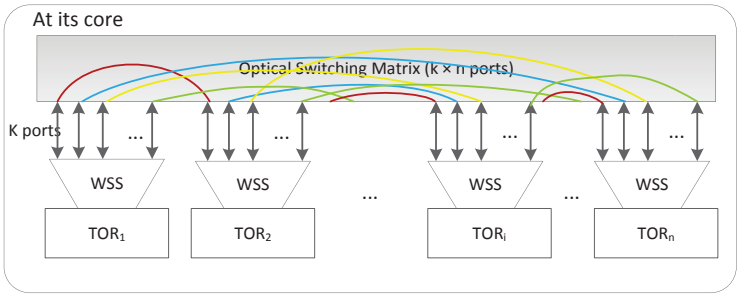


Figure 6.7
The OSA overview [4].

WDM technology. The WSS will split these p wavelengths from other wavelengths carried in the input fiber, and send the p wavelengths to the appropriate port in the optical switching matrix which has a circuit to ToR_j. Thus a circuit with B capacity from ToR_i to ToR_j is set up. Overall, the OSA architecture can support the ToRs communication with distinct bandwidth requirements for the demands. Similarly, if two ToRs are not directly connected through the optical switching matrix, the same wavelength selection and multiplexing are conducted on the hop-by-hop routing along the multi-hop paths.

6.7.3 Mixed Integer Linear Program

In the MILP formulation, our target, the same as mentioned in Section 6.6, is to allocate resources to as many demands as possible. We also need to optimally configure the MEMS matrix optical switch. Different from the MEMS configuration in the HyPaC DCN architecture that each ToR can communicate with only another ToR at the same time, each ToR in the OSA DCN architecture can communicate with other k ToRs at the same time.

Input: A set of VDC demands (D) from users as described in Section 6.4.1. The physical resources in the data center as described in Section 6.4.2 and the OSA DCN

architecture described in Section 6.4.3.

Output: The MEMS matrix configuration topology. The allocated computing resource from servers and bandwidth resource from network connections in data center.

Variables: Some of the variables used in the following MILP formulations are defined in Sections 6.5.2 and 6.6.2. In addition, we defined some new variables to be used.

- l_{ij}: equals 1 if ToR_i is connected to ToR_j through MEMS matrix optical switch directly (bi-direction connection, $l_{ij} = l_{ji}, i \neq j$), 0 otherwise.
- δ_{ij}^{w}: equals 1 if l_{ij} carries wavelength λ_w from ToR_i to ToR_j, 0 otherwise.
- S_{ij}: the bandwidth provisioned from ToR_i to ToR_j (may be over multiple-hops along the routing path from ToR_i to ToR_j).
- v_{ij}^{w}: the volume of traffic flow carried by wavelength λ_w from ToR_i to ToR_j (one-hop connection from ToR_i to ToR_j).
- $T_{l_d}^{ij}$: equals 1 if virtual link l_d with bandwidth requirement is mapped to the route $ToR_i \rightarrow ToR_j$ and $ToR_j \rightarrow ToR_i$ in both directions. In the OSA DCN architecture, the route between any two ToRs could be one-hop route or multi-hop route.

For the above variables, $w \in \{1, 2, ..., W\}$; $i, j \in \{1, 2, ..., R\}, i \neq j$.

Objective: Maximize the number of demands and the bandwidth traffic that can be served:

$$Maximize \sum_{d} \Gamma_d + \sum_{i,j} S_{ij}. \tag{6.17}$$

Constraints:

1. The finally served bandwidth matrix is at most the required bandwidth matrix

by the demands.

$$S_{ij} \leq M_{ij}, \quad \forall i, j \tag{6.18}$$

2. A wavelength between ToR_i and ToR_j can only be used if the two ToRs are connected.

$$\delta_{ij}^{w} \leq l_{ij}, \quad \forall i, j, w. \tag{6.19}$$

3. ToR_i can receive/send wavelength λ_w from/to one ToR at most.

$$\sum_{j} \delta_{ij}^{w} \leq 1, \sum_{i} \delta_{ij}^{w} \leq 1, \quad \forall i, w \tag{6.20}$$

4. We assume in this model, the degree of a ToR switch is k, so a ToR connects to exactly k other ToRs directly through MEMS.

$$\sum_{j} l_{ij} = k, \quad \forall i \tag{6.21}$$

5. The carried bandwidth amount by each wavelength for hop-to-hop connection is limited by the port capacity and the wavelength capacity of a ToR switch port.

$$v_{ij}^{w} \leq \min\{C_{port}, C_{\lambda} \times \delta_{ij}^{w}\}, \quad \forall i, j, w, i \neq j. \tag{6.22}$$

6. The traffic flow balance constraint: the incoming transit flow to a ToR equals

the outgoing transit flow from this ToR.

$$\sum_{j,w} v_{ji}^{w} - \sum_{j} S_{ji} = \sum_{j,w} v_{ij}^{w} - \sum_{j} S_{ij}, \quad \forall i \tag{6.23}$$

Other constraints on the computing resource allocation are same with those already presented in Section 6.5.2 (see constraints 1, 2, 4–8 in Section 6.5.2). The constraints on desired bandwidth traffic matrix are same as with the ones in Section 6.5.2 (see constraints 2 and 3 in Section 6.6.2).

6.8 Experimental Results and Analysis

6.8.1 Approaches for Multiple Objectives MILP/MIQP

In this work, we model our problems as multiple objective MILP problems. The first objective is to maximize the number of accpeted traffic demands. The second objective is to minimize the total cost for all accepted traffic demands. We can see that the two objectives have dependencies. We adopt two approaches to solve the dependent multiple objective MILP model.

Approach 1: Formulate the problem as a weighted sum of two linear objectives

Maximize $obj1 + w * obj2$

S.t. constraints

Then adjust weight w from small negative number to large positive number and resolve the problem for different values of w using CPLEX warming start techniques.

Approach 2: Add constraint for the first primary objective

Minimize $a * obj1 + b * obj2$

Initialize:

$a = -1$; $b = -1$; $objVal = -1$;

S.t. constraints

S.t. if $(objVal >= 0)$ $obj1 = objVal$

Then we will conduct the CPLEX solving process for two rounds using CPLEX script flow control. In the first CPLEX solving round, we maximize the $obj1$. In the second round, without affecting the result of $obj1$ using additional constraint listed above and changing the value for a and b in the CPLEX script flow control, we minimize the $obj2$.

6.8.2 Experimental Results

The experiments are carried out on a Linux Server, the IBM OPL CPLEX tool is used to generate optimal solutions for the MILP and MIQP mathematical models. All the demands which are connected un-directional graphs are generated automatically by a self implemented random generation algorithm in the experiments. We allow each demand can have 1 to 5 VMs with the bandwidth requirement from 1 GB (low bandwidth requirement) to 100 GB (high bandwidth requirement) between related VMs. Two kinds of data center topology scales are used in the experiments: one is in small scale with 4 racks and each rack has 4 servers; the other one is in medium scale with 10 racks and each rack has 10 servers. We will develop dynamic heuristics to solve this virtualized resource provisioning problem in optical DCNs for large scale data centers in the future.

First, we test the MILP model for fully non-blocking MEMS DCN architecture through two approaches discussed in 7.1. Through approach 2, the optimal solution with minimal total cost for the maximal number of accepted demands can be obtained one time. Through approach 1, we ad just the value of weight w from positive value to negative value to obtain the optimal solution. Figure 6.8 shows that when we

adjust the value of w to -0.1, we can get the optimal solution (no blocking happens and with minimal total cost) for 30 demands. From the result, we can see that when w increases its value, the cost will increase, when w decreases its value (less than -0.1 in the figure) continuously, the blocking will happen in which some demands will be dropped. In the figure we used infinity cost to show the blocking situation when w is less than -0.1.

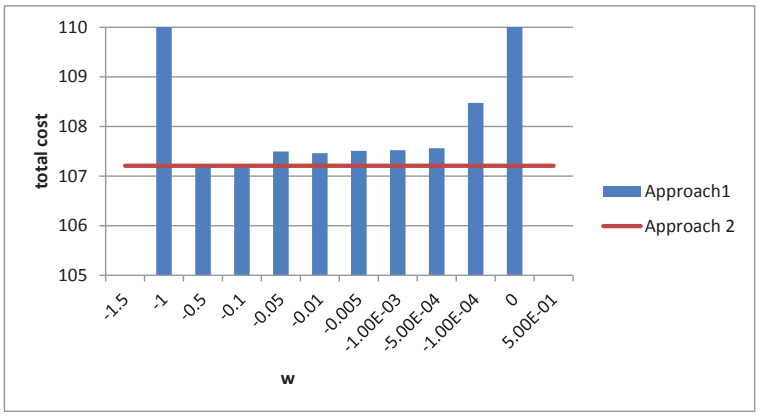


Figure 6.8
The optimal solution (all demands are accepted with minimal total cost) through two approaches for 30 demands.

For the fully non-blocking MEMS DCN architecture, we tested different number of demands to find the suitable w that could be used for different number of demands on different data center topology scales. In figure 6.9 we only show that result obtained from the small scale data center topology since it is the same with what is obtained from the large scale data center topology. To sum up, for approach 1, we found that with weight $w = -0.1$ we can get the optimal solution, which is same with that is obtained by approach 2, for different number of demands for different data center topology scales for the MILP model on fully connect non-blocking MEMS DCN architecture.

In addition, we test the MIQP model for the HyPaC DCN architecture through two

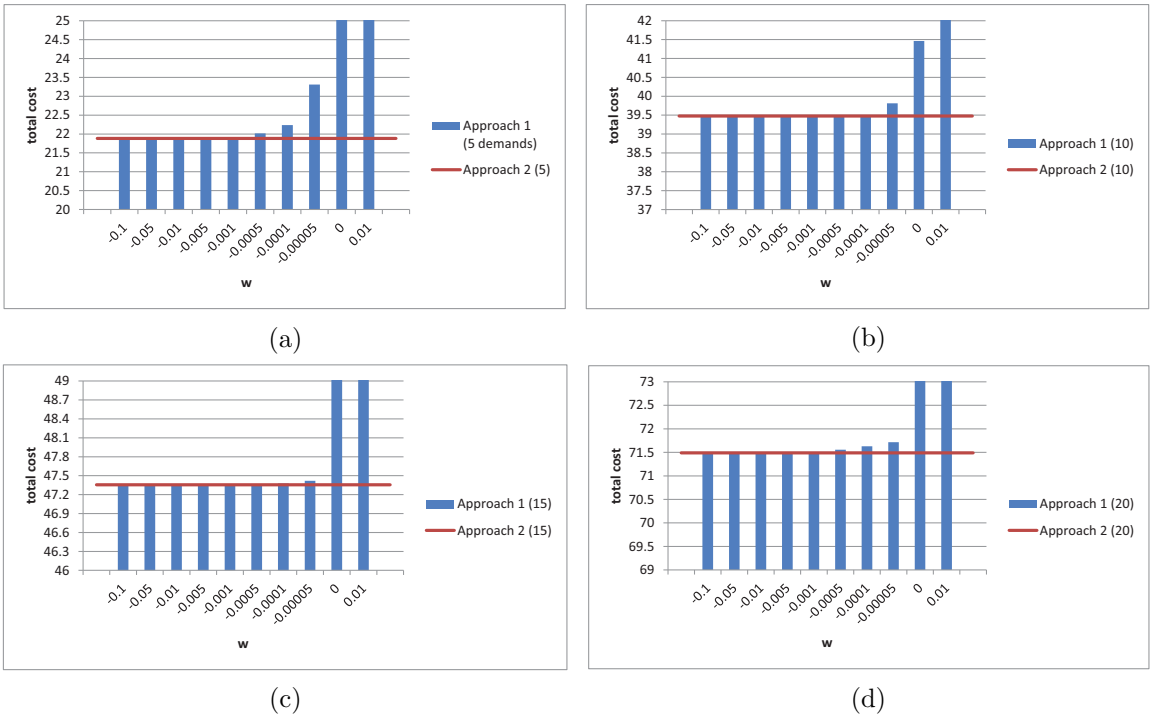


Figure 6.9
Find the suitable value of w for MILP model for fully non-blocking MEMS DCN architecture: (a) 5 demands, (b) 10 demands, (c) 15 demands, (d) 20 demands.

approaches as well. For the HyPac DCN, the MEMS optical switch is not configured initially, so we do not know the connections between the racks in data center through the MEMS. To configure the MEMS, the MIQP model maps the VMs of each demand to the servers and computes the network traffic flow between racks, finally decide the MEMS configuration so that maximum number of demands can be accepted with minimal cost (the cost for network traffic through optical circuit switching is less than that through Ethernet packet switching). Figure 6.10 shows the network traffic distribution in the data center (10 racks with 10 servers under each rack) for 60 demands when the CPLEX solver allocates resources for all demands with minimal total cost. And the MEMS connection is shown in table 6.1. From the network traffic distribution figure and the MEMS configuration table, we can see that to minimize the

total cost, around 81.45% network traffic flows are switched through MEMS optical switch and only 8% traffic flows are switched through Ethernet packet switching. The remaining network traffics are within one rack.

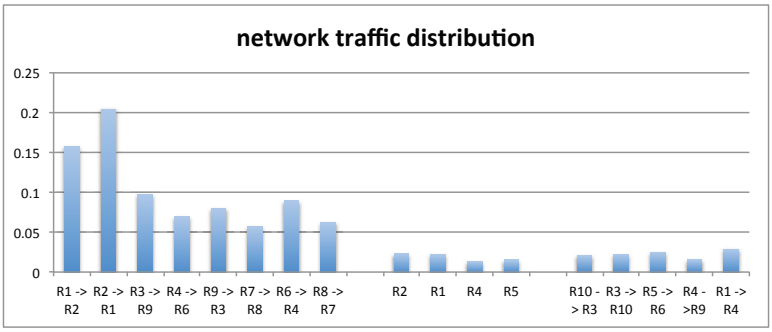


Figure 6.10
The network traffic flow distribution in data center for 60 demands.

Table 6.1
MEMS connection configuration between racks

Rack to Rack	Connected through MEMS
R1 ↔ R2	YES
R3 ↔ R9	YES
R4 ↔ R6	YES
R7 ↔ R8	YES
Other connections	NO

6.9 Conclusion and future work

In this chapter, we investigate the virtualized resource provisioning problems in optical DCNs. Based on different types of optical DCN architectures, different MILP/MIQP model are constructed. The target for the resource provisioning problems is to allocate resources for as many demands as possible and minimize the total cost for providing all these resources. Currently, we have conducted experiments for different MILP/MIQP mathematical models. Two approaches are adopted to find

the optical solutions for the models. However, such models only work well for small scale experiments, such as smaller number of demands and small DCN. In the future, we will design more complete experiments and develop dynamic heuristics for solving the problems for different optical DCN architectures.

Chapter 7

Conclusion and Future Work

7.1 Conclusion

In this dissertation, we investigate the network-aware resource allocation and virtual data center resource provisioning problems in Grid/Cloud. Resource allocation and management is an evolving part of many Grid/Cloud computing and data center management problems. Along with the increasing number of Big Data applications that run in Grid/Cloud, the network resource becomes an essential aspect that needs to be considered and could be the bottleneck for the resource provisioning performance for Grid/Cloud providers.

We focus on the joint resource scheduling for the submitted jobs which consist of number of sequential and parallel sub-tasks in Grid/Cloud networks in the first two parts (Chapters 3 and 4) of this dissertation. Grid network users can access a shared set of resources for scientific computing tasks. Cloud tenants are offered IT infrastructure through Infrastructure as a Service (IaaS). An efficient resource scheduling mechanism across the network, as a result, will improve the resource utilization and also reduce the cost of scheduling in the Grid/Cloud significantly. We investigated the bandwidth guaranteed joint resource scheduling from both the Grid/Cloud provider's point of view and the customer's point of view, in which the multi-layer optical network architecture is introduced to guarantee the reservation of the network bandwidth

resource. From the Grid/Cloud provider's point of view, we completed the joint resource scheduling for as many submitted jobs as possible with the minimal overall capital expenditure for providers. From the customer's point of view, we allocate the resources to each tenant with minimal rental cost. Making use of the advantages of optical networks, the cost optimized joint resource scheduling can be realized with low cost and high throughput. We modeled the joint resource scheduling problems mentioned above as optimization problems and developed both MILP optimal mathematical model and efficient heuristics to solve the problems. We also found that different job scheduling policies would affect the total cost of resource allocation and the total job acceptance rate by the Grid/Cloud providers.

Along with the advent of techniques for virtual cloud and virtualized data centers, the Cloud service is not limited to providing computing resources such as VMs to the customers based on the infrastructure as a service (IaaS). The virtual data center and virtual cloud service enable customers to quickly construct their own cloud platform for running their applications. In this case, in the second two parts (Chapters 5 and 6) of this dissertation we focus on the network-aware virtualized cloud and virtualized data center provisioning problems through optical network technology. The IP over elastic optical network architecture is adopted for the inter-data center network connections in the virtual cloud provisioning. The hybrid optical network architecture and complete optical switching network architecture are introduced for the intra-data center network connection in the virtual data center provisioning. We model the problems as optimization problems, construct MILP mathematical model and propose heuristics to solve them. All of the resource allocations problems we discussed in the dissertation are NP-hard problems. In addition we only deal with static demands from customers which leave us the one expansion for possible future work.

7.2 Future Work

One possible future work is we plan to consider the dynamic demand traffics for the resource provisioning problems. We will involve the queuing theory models to analyze the dynamic demand traffic and study how our resource provisioning simulator processes the dynamic requests from customers. The dynamic demand traffic would mimic the real traffic in the current cloud, such as Google cloud and Amazon cloud.

In addition, another future work is mainly based on the second two parts in the dissertation. In Chapters 5 and 6, we consider the virtual resource provisioning for inter-data center network and intra-data center network connectivities separately. In the future work, we will investigate the end-to-end virtual resource provisioning across multiple cloud network domains, in which the detailed inter-data center and intra-data center communications will be dealt with together. The end-to-end resource provisioning idea would be important for our future goals. In our current work, what we have implemented are resource provisioning simulators. In the future, what we want is to implement resource provisioning emulators that would emulate the real network and hardware and would work over the real operating systems. In addition we want to test our emulators on real testbeds such as GENI (an academic testbed) and Amazon AWS (an industry testbed).

Moreover, one more possible future work is to involve the software defined networking (SDN) technique for the network-aware resource provisioning system. In our dissertation, we focus on the network-aware resource provisioning for different scenarios, in which the on-demand provisioning of network resource plays an important role in the problem. We would like to involve the SDN technique for the network provisioning with the advantages of reducing network provisioning time and reducing service costs through improved network management efficiency.

Bibliography

[1] I. Foster, Y. Zhao, I. Raicu, and S. Lu, "Cloud computing and Grid computing 360-degree compared," in *Grid Computing Environments Workshop (GCE)*, November 2008, pp. 1–10.

[2] P. T. Endo, A. V. de Almeida Palhares, N. N. Pereira *et al.*, "Resource allocation for distributed cloud: concepts and research challenges," *IEEE Network*, vol. 25, pp. 42–46, July-August 2011.

[3] O. Gerstel, M. Jinno, A. Lord, and S. J. B. Yoo, "Elastic optical networking: a new dawn for the optical layer?" *Communications Magazine, IEEE*, vol. 50, no. 2, pp. s12–s20, February 2012.

[4] A. Singla, A. Singh, and Y. Chen, "OSA: An optical switching architecture for data center networks with unprecedented flexibility," in *Presented as part of the 9th USENIX Symposium on Networked Systems Design and Implementation (NSDI 12)*, 2012, pp. 239–252.

[5] O. Pedrola, A. Castro *et al.*, "CAPEX study for a multilayer IP/MPLS-over-flexgrid optical network," *IEEE JOCN*, vol. 4, no. 8, pp. 639–650, Aug 2012.

[6] I. Foster and C. Kesselman, Eds., *The Grid: Blueprint for a New Computing Infrastructure*. Morgan Kaufmann Publishers Inc., 1999.

[7] J. M. Schopf, "Grid resource management," J. Nabrzyski, J. M. Schopf, and J. Weglarz, Eds. Kluwer Academic Publishers, 2004, ch. Ten Actions when Grid Scheduling: The User As a Grid Scheduler, pp. 15–23.

[8] "Open Science Grid." [Online]. Available: https://www.opensciencegrid.org/bin/view

[9] "GENI exploring networks of the future." [Online]. Available: https://www.geni.net/

[10] "HTCondor high throughput computing." [Online]. Available: http://research.cs.wisc.edu/htcondor/

[11] M. Armbrust, A. Fox, R. Griffith, A. D. Joseph *et al.*, "A view of cloud computing," *Communications of the ACM*, vol. 53, pp. 50–58, April 2010.

[12] "Amazon Elastic Compute Cloud." [Online]. Available: http://aws.amazon.com

[13] "Google Compute Engine." [Online]. Available: https://developers.google.com/compute

[14] "Windows Azure." [Online]. Available: http://www.windowsazure.com

[15] "Amazon AWS Cloud Products." [Online]. Available: https://aws.amazon.com/products/

[16] X. Zhang, E. Tune, R. Hagmann, R. Jnagal, V. Gokhale, and J. Wilkes, "CPI-2: CPU performance isolation for shared compute clusters," in *Proceedings of the 8th ACM European Conference on Computer Systems*. ACM, 2013, pp. 379–391.

[17] D. Shue, M. J. Freedman, and A. Shaikh, "Performance isolation and fairness for multi-tenant cloud storage," in *Presented as part of the 10th USENIX Symposium on Operating Systems Design and Implementation (OSDI 12)*. Hollywood, CA: USENIX, 2012, pp. 349–362. [Online]. Available: https://www.usenix.org/conference/osdi12/technical-sessions/presentation/shue

[18] S. Angel, H. Ballani, T. Karagiannis, G. O'Shea, and E. Thereska, "End-to-end performance isolation through virtual datacenters," in *11th USENIX Symposium on Operating Systems Design and Implementation (OSDI 14)*, 2014, pp. 233–248.

[19] "Optical network (photonic network) definition." [Online]. Available: http://searchnetworking.techtarget.com/definition/photonic-network

[20] "IP/MPLS networks." [Online]. Available: http://www.commverge.com/solutions/IPCoreEdgeNetworks/IPMPLSNetwork

[21] "IP/MPLS network, coppernet solutions." [Online]. Available: http://www.mct.gov.zm/index.php?option=com_content&view=article&id=36:multi-protocol-label-switching&catid=3:networking

[22] R. Huelsermann, M. Gunkel, C. Meusburger, and D. A. Schupke, "Cost modeling and evaluation of capital expenditures in optical multilayer networks," *Journal of Optical Networking*, vol. 7, pp. 814–833, Sep. 2008.

[23] T. P. Walker, "Optical transport network tutorial," ITU-T standard.

[24] F. Rambach, B. Konrad, L. Dembeck, U. Gebhard, M. Gunkel, M. Quagliotti, L. Serra, and V. Lopez, "A multilayer cost model for metro/core networks,"

IEEE/OSA Journal of Optical Communications and Networking, vol. 5, pp. 210–225, March 2013.

[25] B. Ramamurthy, R. K. Sinha, and K. K. Ramakrishnan, "Multi-layer design of IP over WDM backbone networks: impact on cost and survivability," in *The Conference on the Design of Reliable Communication Networks (DRCN)*, March 2013, pp. 60–70.

[26] "OSG Council Document 1106-v1." [Online]. Available: http://osg-docdb.opensciencegrid.org/cgi-bin/ShowDocument?docid=1106

[27] Cisco, "Cisco global cloud index: forecast and methodology, 2013-2018," http://www.cisco.com/c/en/us/solutions/collateral/service-provider/global-cloud-index-gci/Cloud_Index_White_Paper.pdf, Cisco, 2014.

[28] "AT&T NetBond." [Online]. Available: https://www.synaptic.att.com/clouduser/html/productdetail/ATT_NetBond.htm

[29] "Verizon brings cloud offering into self-service era." [Online]. Available: http://www.itworld.com/cloud-computing/376880/verizon-brings-cloud-offering-self-service-era

[30] J. Berthold, A. Saleh, L. Blair, and J. M. Simmons, "Optical networking: Past, present, and future," *Lightwave Technology, Journal of*, vol. 26, no. 9, pp. 1104–1118, May 2008.

[31] C. Develder, M. De Leenheer, B. Dhoedt, M. Pickavet, D. Colle, F. De Turck, and P. Demeester, "Optical networks for grid and cloud computing applications," *Proceedings of the IEEE*, vol. 100, no. 5, pp. 1149–1167, May 2012.

[32] ITU, "G.694.1: Spectral grids for WDM applications: DWDM frequency grid," *International Telecommunication Union, ITU*, 2012.

[33] P. Yi, H. Ding, and B. Ramamurthy, "Cost-optimized joint resource allocation in grids/clouds with multilayer optical network architecture," *Optical Communications and Networking, IEEE/OSA Journal of*, vol. 6, no. 10, pp. 911–924, Oct 2014.

[34] ——, "Budget-minimized resource allocation and task scheduling in distributed grid/clouds," in *The Conference on Computer Communications and Networks (ICCCN)*, July/August 2013.

[35] ——, "A tabu search based heuristic for optimized joint resource allocation and task scheduling in grid/clouds," in *IEEE Conference on Advanced Networks and Telecommunications Systems (ANTS)*, Dec. 2013.

[36] ——, "Budget-optimized network-aware joint resource allocation in grids/clouds over optical networks," *Journal of Lightwave Technology*, Jan. 2016, to appear.

[37] P. Yi and B. Ramamurthy, "Provisioning virtualized cloud services in IP/MPLS-over-EON networks," in *Optical Network Design and Modeling (ONDM), 2015 International Conference on*, May 2015, pp. 45–50.

[38] ——, "Provisioning virtualized cloud services in IP/MPLS-over-EON networks," *Photonic Network Communications*, pp. 1–14, December 2015. [Online]. Available: http://dx.doi.org/10.1007/s11107-015-0588-x

[39] A. Ravula and B. Ramamurthy, "Grid networking," in *Next-generation Internet: architectures and protocols*, B. Ramamurthy, G. N. Rouskas, and K. M. Sivalingam, Eds. Cambridge University Press, 2011, ch. 5, pp. 88–103.

[40] X. Liu, W. Wei, C. Qiao, T. Wang, W. Hu, W. Guo, and M. Wu, "Task scheduling and lightpath establishment in optical grids," in *INFOCOM*, April 2008, pp. 1966–1974.

[41] C. Li and L. Li, "An efficient resource allocation for maximizing benefit of users and resource providers in ad hoc grid environment," *Information Systems Frontiers*, vol. 14, pp. 987–998, 2012.

[42] L. Tomas, B. Caminero, and C. Carrion, "Improving grid resource usage: metrics for measuring fragmentation," in *The IEEE/ACM Symposium on Cluster, Cloud and Grid Computing (CCGrid)*, May 2012, pp. 352–359.

[43] C. Castillo, G. Rouskas, and K. Harfoush, "Efficient resource management using advance reservations for heterogeneous grids," in *Parallel and Distributed Processing, 2008. IPDPS 2008. IEEE International Symposium on*, April 2008, pp. 1–12.

[44] K. Chard, K. Bubendorfer, and P. Komisarczuk, "High occupancy resource allocation for grid and cloud systems, a study with drive," in *Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing (HDPC)*, 2010, pp. 73–84.

[45] R. Urgaonkar, U. C. Kozat, K. Igarashi, and M. J. Neely, "Dynamic resource allocation and power management in virtualized data centers," in *IEEE Network Operations and Management Symposium (NOMS)*, April 2010, pp. 479–486.

[46] X. Kong, C. Lin, Y. Jiang, W. Yan, and X. Chu, "Efficient dynamic task scheduling in virtualized data centers with fuzzy prediction," *Journal of Network and Computer Applications*, vol. 34, pp. 1068–1077, July 2011.

[47] M. Nejad, L. Mashayekhy, and D. Grosu, "Truthful greedy mechanisms for dynamic virtual machine provisioning and allocation in clouds," *IEEE Transactions on Parallel and Distributed Systems*, vol. 26, no. 2, pp. 594–603, Feb 2015.

[48] Z. Xiao, W. Song, and Q. Chen, "Dynamic resource allocation using virtual machines for cloud computing environment," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 24, no. 6, pp. 1107–1117, June 2013.

[49] J. Lee, Y. Turner, M. Lee, L. Popa, S. Banerjee, J.-M. Kang, and P. Sharma, "Application-driven bandwidth guarantees in datacenters," in *Proceedings of the ACM Conference on SIGCOMM (SIGCOMM)*, 2014, pp. 467–478.

[50] R. Buyya, A. Beloglazov, and J. Abawajy, "Energy-efficiency management of data center resources for cloud computing: A vision, architectural elements, and open challenges," in *Proceedings of the International Conference on Parallel and Distributed Processing Techniques and Applications (PDPTA)*, Las Vegas, USA, July 2010.

[51] L. Nonde, T. El-Gorashi, and J. Elmirghani, "Energy efficient virtual network embedding for cloud networks," *Lightwave Technology, Journal of*, vol. 33, no. 9, pp. 1828–1849, May 2015.

[52] S. Tayal, "Tasks scheduling optimization for the cloud computing systems," *International Journal of Advanced Engineering Science and Technologies*, vol. 5, pp. 111–115, Feb. 2011.

[53] M. Alicherry and T. V. Lakshman, "Network aware resource allocation in distributed clouds," in *INFOCOM, 2012 Proceedings IEEE*, March 2012, pp. 963–971.

[54] W. Wang, D. Niu, B. Li, and B. Liang, "Dynamic cloud resource reservation via cloud brokerage," in *IEEE Conference on Distributed Computing Systems (ICDCS)*, July 2013, pp. 400–409.

[55] V. Setty, R. Vitenberg, G. Kreitz, G. Urdaneta, and M. van Steen, "Cost-effective resource allocation for deploying pub/sub on cloud," in *IEEE International Conference on Distributed Computing Systems (ICDCS)*, June 2014, pp. 555–566.

[56] M. Liu, W. Dou, S. Yu, and Z. Zhang, "A decentralized cloud firewall framework with resources provisioning cost optimization," *Parallel and Distributed Systems, IEEE Transactions on*, vol. 26, no. 3, pp. 621–631, March 2015.

[57] D. Pandit, S. Chattopadhyay, M. Chattopadhyay, and N. Chaki, "Resource allocation in cloud using simulated annealing," in *Applications and Innovations in Mobile Computing (AIMoC), 2014*, Feb 2014, pp. 21–27.

[58] M. Sedaghat, F. Hernandez-Rodriguez, and E. Elmroth, "Autonomic resource allocation for cloud data centers : a peer to peer approach," in *International Conference on Cloud and Autonomic Computing (ICCAC)*, 2014, pp. 131–140.

[59] J.-T. Tsai, J.-C. Fang, and J.-H. Chou, "Optimized task scheduling and resource allocation on cloud computing environment using improved differential evolution algorithm," *Computers & Operations Research*, vol. 40, no. 12, pp. 3045 – 3055, 2013.

[60] K. Liu, J. Peng, W. Liu, P. Yao, and Z. Huang, "Dynamic resource reservation via broker federation in cloud service: A fine-grained heuristic-based approach," in *IEEE Global Communications Conference (GLOBECOM)*, Dec 2014, pp. 2338–2343.

[61] M. Yu, Y. Yi, J. Rexford, and M. Chiang, "Rethinking virtual network embedding: Substrate support for path splitting and migration," *SIGCOMM Comput. Commun. Rev.*, vol. 38, no. 2, pp. 17–29, Mar. 2008.

[62] A. Haider, R. Potter, and A. Nakao, "Challenges in resource allocation in network virtualization," in *20th ITC Specialist Seminar*, vol. 18, 2009, p. 20.

[63] X. Cheng, S. Su, Z. Zhang, H. Wang, F. Yang, Y. Luo, and J. Wang, "Virtual network embedding through topology-aware node ranking," *SIGCOMM Comput. Commun. Rev.*, vol. 41, no. 2, pp. 38–47, Apr. 2011.

[64] M. Chowdhury, M. R. Rahman, and R. Boutaba, "Vineyard: Virtual network embedding algorithms with coordinated node and link mapping," *IEEE/ACM Transactions on Networking (TON)*, vol. 20, no. 1, pp. 206–219, 2012.

[65] I. Fajjari, N. Aitsaadi, G. Pujolle, and H. Zimmermann, "VNE-AC: Virtual network embedding algorithm based on ant colony metaheuristic," in *Communications (ICC), 2011 IEEE International Conference on*, June 2011, pp. 1–6.

[66] J. F. Botero, X. Hesselbach, M. Duelli, D. Schlosser, A. Fischer, and H. De Meer, "Energy efficient virtual network embedding," *Communications Letters, IEEE*, vol. 16, no. 5, pp. 756–759, 2012.

[67] M. R. Rahman and R. Boutaba, "SVNE: Survivable virtual network embedding algorithms for network virtualization." *IEEE Transactions on Network and Service Management*, vol. 10, no. 2, pp. 105–118, 2013.

[68] M. G. Rabbani, R. P. Esteves *et al.*, "On tackling virtual data center embedding problem," in *IFIP/IEEE International Symposium on IM*, May 2013, pp. 177–184.

[69] C. Papagianni, A. Leivadeas *et al.*, "On the optimal allocation of virtual resources in cloud computing networks," *IEEE Transactions on Computers*, vol. 62, no. 6, pp. 1060–1071, June 2013.

[70] M. Alicherry and T. V. Lakshman, "Network aware resource allocation in distributed clouds," in *INFOCOM*, March 2012, pp. 963–971.

[71] Q. Zhang, M. Zhani, M. Jabri, and R. Boutaba, "Venice: Reliable virtual data center embedding in clouds," in *INFOCOM, 2014 Proceedings IEEE*, April 2014, pp. 289–297.

[72] C. Guo, G. Lu *et al.*, "SecondNet: A data center network virtualization architecture with bandwidth guarantees," in *Co-NEXT '10*, New York, NY, USA, 2010, pp. 15:1–15:12.

[73] A. Amokrane, M. F. Zhani *et al.*, "Greenhead: Virtual data center embedding across distributed infrastructures," *IEEE Transactions on Cloud Computing*, vol. 1, no. 1, pp. 36–49, Jan 2013.

[74] M. F. Zhani, Q. Zhang *et al.*, "VDC planner: Dynamic migration-aware virtual data center embedding for clouds," in *IFIP/IEEE International Symposium on IM*, May 2013, pp. 18–25.

[75] B. Martini, M. Gharbaoui, and P. Castoldi, "Cross-functional resource orchestration in optical telco clouds," in *International Conference on Transparent Optical Networks (ICTON)*, 2015, p. to appear.

[76] H. Ballani, P. Costa, T. Karagiannis, and A. Rowstron, "Towards predictable datacenter networks," in *Proceedings of the ACM Conference on SIGCOMM (SIGCOMM)*, 2011, pp. 242–253.

[77] W.-L. Yeow, C. Westphal, and U. Kozat, "Designing and embedding reliable virtual infrastructures," in *Proceedings of the Second ACM SIGCOMM Workshop on Virtualized Infrastructure Systems and Architectures*, ser. VISA '10, 2010, pp. 33–40.

[78] M. Gharbaoui, B. Martini, and P. Castoldi, "Anycast-based optimizations for inter-data-center interconnections [invited]," *Optical Communications and Networking, IEEE/OSA Journal of*, vol. 4, no. 11, pp. B168–B178, Nov 2012.

[79] W. Fang, M. Lu, X. Liu, L. Gong, and Z. Zhu, "Joint defragmentation of optical spectrum and it resources in elastic optical datacenter interconnections," *Optical Communications and Networking, IEEE/OSA Journal of*, vol. 7, no. 4, pp. 314–324, April 2015.

[80] M. Mao and M. Humphrey, "Scaling and scheduling to maximize application performance within budget constraints in cloud workflows," in *IEEE Symposium on Parallel Distributed Processing (IPDPS)*, May 2013, pp. 67–78.

[81] H. Zhao, M. Pan, X. Liu, X. Li, and Y. Fang, "Optimal resource rental planning for elastic applications in cloud market," in *IEEE Symposium on Parallel Distributed Processing (IPDPS)*, May 2012, pp. 808–819.

[82] Q. Zhang, M. F. Zhani, R. Boutaba, and J. L. Hellerstein, "Harmony: dynamic heterogeneity-aware resource provisioning in the cloud," in *IEEE Conference on Distributed Computing Systems (ICDCS)*, July 2013, pp. 510–519.

[83] A. Greenberg, J. Hamilton, D. A. Maltz, and P. Patel, "The cost of a cloud: research problems in data center networks," *SIGCOMM Computer Communication Review*, vol. 39, pp. 68–73, December 2008.

[84] "Determining total cost of ownership for data center and network room infrastructure," American Power Conversion, Tech. Rep., 2005.

[85] B. T. Olsen and K. Stordahl, "Models for forecasting cost evolution of components and technologies." [Online]. Available: http://www.business-planning-for-managers.com/Worpress/wp-content/uploads/2014/03/T04_4.pdf

[86] "IBM ILOG CPLEX Optimization Studio." [Online]. Available: http://www.ibm.com/software/products/en/ibmilogcpleoptistud/

[87] F. Glover and M. Laguna, *Tabu Search*. Kluwer Academic Publishers, 1997.

[88] G. Goncalves, M. Santos, G. Charamba, P. Endo *et al.*, "D-CRAS: Distributed cloud resource allocation system," in *Network Operations and Management Symposium, IEEE Network*, April 2012, pp. 16–20.

[89] H. Chen, J. Zhang, Y. Zhao, J. Deng, W. Wang, R. He, X. Yu, Y. Ji, H. Zheng, Y. Lin, and H. Yang, "Experimental demonstration of datacenter resources integrated provisioning over multi-domain software defined optical networks," *Lightwave Technology, Journal of*, vol. 33, no. 8, pp. 1515–1521, April 2015.

[90] P. Angu and B. Ramamurthy, "Continuous and parallel optimization of dynamic bandwidth scheduling in wdm networks," in *Global Telecommunications Conference (GLOBECOM 2010)*. IEEE, 2010, pp. 1–6.

[91] N. Charbonneau and V. M. Vokkarane, "A survey of advance reservation routing and wavelength assignment in wavelength-routed WDM networks," *Communications Surveys Tutorials, IEEE*, vol. 14, no. 4, pp. 1037–1064, Fourth 2012.

[92] "OSCARS." [Online]. Available: https://www.es.net/engineering-services/oscars/

[93] U. B. G. Bauer, B Beccati and K. Biery, "The CMS online cluster: It for a large data acquisition and control cluster."

[94] J. Atlas, M. Swany, and K. S. Decker, "Flexible grid workflows using TAEMS," in *Workshop on Exploring Planning and Scheduling for Web Services, Grid and Autonomic Comp*, 2005.

[95] "Amazon EC2 pricing." [Online]. Available: http://aws.amazon.com/ec2/pricing/

[96] "Coogle compute engine pricing." [Online]. Available: https://cloud.google.com/compute/pricing

[97] D. Shen, G. Li, A. Chiu, D.-M. Hwang, D. Xu, D. Wang, C.-K. Chan, and R. Doverspike, "On multiplexing optimization in DWDM networks," in *Optical Fiber Communication Conference and Exposition (OFC/NFOEC)*, March 2011, pp. 1–3.

[98] K. Wen, X. Cai, Y. Yin, D. J. Geisler, R. Proietti, R. P. Scott, N. K. Fontaine, and S. J. B. Yoo, "Adaptive spectrum control and management in elastic optical

networks," *Selected Areas in Communications, IEEE Journal on*, vol. 31, no. 1, pp. 39–48, 2013.

[99] S. Peng, R. Nejabati *et al.*, "Role of optical network virtualization in cloud computing," *IEEE/OSA JOCN*, vol. 5, no. 10, pp. A162–A170, Oct 2013.

[100] VMware, "vCloud Suit," 2015. [Online]. Available: http://www.vmware.com/ap/products/vcloud-suite

[101] Cisco, "Cisco Virtualized Multiservice Data Center," 2013. [Online]. Available: http://www.cisco.com/c/en/us/solutions/enterprise/data-center-designs-cloud-computing/landing_vmdc.html

[102] C. Kachris and I. Tomkos, "Optical interconnection networks for data centers," in *ONDM*, April 2013, pp. 19–22.

[103] L. M. Contreras, V. Lopez *et al.*, "Toward cloud-ready transport networks," *IEEE Communications Magazine*, vol. 50, no. 9, pp. 48–55, September 2012.

[104] M. Xia and S. Dahlfort, "Cost analysis for elastic optical networking: Single channel vs. multi channels," in *IEEE Globecom Workshops*, Dec 2012, pp. 364–368.

[105] M. Klinkowski and K. Walkowiak, "On the advantages of elastic optical networks for provisioning of cloud computing traffic," *Network, IEEE*, vol. 27, no. 6, pp. 44–51, November 2013.

[106] C. Politi, V. Anagnostopoulos, C. Matrakidis, and A. Stavdas, "Routing in dynamic future flexi-grid optical networks," in *ONDM*, April 2012, pp. 1–4.

[107] Y. Wang, X. Cao *et al.*, "A study of the routing and spectrum allocation in spectrum-sliced elastic optical path networks," in *INFOCOM*, April 2011, pp. 1503–1511.

[108] F. Xiong, *Digital Modulation Techniques*. Norwood, MA, USA: Artech House, Inc., 2006.

[109] Infinera, "Super-channels: DWDM transmission at 100Gb/s and beyond," http://www.infinera.com/pdfs/whitepapers/superchannel_whitepaper.pdf, Infinera, 2012.

[110] J. Zhang, Y. Zhao, X. Yu, J. Zhang, M. Song, Y. Ji, and B. Mukherjee, "Energy-efficient traffic grooming in sliceable-transponder-equipped IP-over-elastic optical networks [invited]," *IEEE/OSA JOCN*, vol. 7, no. 1, pp. A142–A152, Jan 2015.

[111] N. Sambo, A. D'Errico, C. Porzi, V. Vercesi, M. Imran, F. Cugini, A. Bogoni, L. Potì, and P. Castoldi, "Sliceable transponder architecture including multi-wavelength source," *J. Opt. Commun. Netw.*, vol. 6, no. 7, pp. 590–600, Jul 2014.

[112] N. Sambo, P. Castoldi, A. D'Errico, E. Riccardi, A. Pagano, M. Moreolo, J. Fabrega, D. Rafique, A. Napoli, S. Frigerio, E. Salas, G. Zervas, M. Nolle, J. Fischer, A. Lord, and J.-P. Gimenez, "Next generation sliceable bandwidth variable transponders," *Communications Magazine, IEEE*, vol. 53, no. 2, pp. 163–171, Feb 2015.

[113] M. Dallaglio, A. Giorgetti, N. Sambo, and P. Castoldi, "Impact of SBVTs based on multi-wavelength source during provisioning and restoration in elastic optical

networks," in *Optical Communication (ECOC), 2014 European Conference on*, Sept 2014, pp. 1–3.

[114] L. Velasco, O. Gonzalez de Dios, V. Lopez, J. Fernandez-Palacios, and G. Junyent, "Finding an objective cost for sliceable flexgrid transponders," in *Optical Fiber Communications Conference and Exhibition (OFC), 2014*, March 2014, pp. 1–3.

[115] B. de la Cruz, O. Gonzalez de Dios, V. Lopez, and J. Fernandez-Palacios, "Opex savings by reduction of stock of spare parts with sliceable bandwidth variable transponders," in *Optical Fiber Communications Conference and Exhibition (OFC), 2014*, March 2014, pp. 1–3.

[116] J. Zhang, Y. Ji, M. Song, Y. Zhao, X. Yu, J. Zhang, and B. Mukherjee, "Dynamic traffic grooming in sliceable bandwidth-variable transponder-enabled elastic optical networks," *Lightwave Technology, Journal of*, vol. 33, no. 1, pp. 183–191, Jan 2015.

[117] F. Rambach, B. Konrad *et al.*, "A multilayer cost model for metro/core networks," *IEEE/OSA JOCN*, vol. 5, no. 3, pp. 210–225, March 2013.

[118] V. Lopez, O. Gonzalez de Dios *et al.*, "Target cost for sliceable bandwidth variable transponders in a real core network," in *Future Network and Mobile Summit*, July 2013, pp. 1–9.

[119] M. Svaluto Moreolo, J. Fabrega, L. Nadal, F. Vilchez, and G. Junyent, "Bandwidth variable transponders based on OFDM technology for elastic optical networks," in *Transparent Optical Networks (ICTON), 2013 15th International Conference on*, June 2013, pp. 1–4.

[120] M. Al-Fares, A. Loukissas, and A. Vahdat, "A scalable, commodity data center network architecture," in *Proceedings of the ACM SIGCOMM 2008 Conference on Data Communication*, ser. SIGCOMM '08, 2008, pp. 63–74.

[121] C. Guo, G. Lu, D. Li, H. Wu, X. Zhang, Y. Shi, C. Tian, Y. Zhang, and S. Lu, "BCube: A high performance, server-centric network architecture for modular data centers," in *Proceedings of the ACM SIGCOMM Conference on Data Communication*, 2009, pp. 63–74.

[122] S. J. B. Yoo, Y. Yin, and K. Wen, "Intra and inter datacenter networking: The role of optical packet switching and flexible bandwidth optical networking," in *Optical Network Design and Modeling (ONDM), 2012 16th International Conference on*. IEEE, 2012, pp. 1–6.

[123] C. Kachris, K. Kanonakis, and I. Tomkos, "Optical interconnection networks in data centers: recent trends and future challenges," *Communications Magazine, IEEE*, vol. 51, no. 9, pp. 39–45, 2013.

[124] H. Wang, Y. Xia, K. Bergman, T. Ng, S. Sahu, and K. Sripanidkulchai, "Rethinking the physical layer of data center networks of the next decade: Using optics to enable efficient *-cast connectivity," *ACM SIGCOMM Computer Communication Review*, vol. 43, no. 3, pp. 52–58, 2013.

[125] M. Abu Sharkh, M. Jammal, A. Shami, and A. Ouda, "Resource allocation in a network-based cloud computing environment: design challenges," *Communications Magazine, IEEE*, vol. 51, no. 11, pp. 46–52, 2013.

[126] G. Wang, D. G. Andersen, M. Kaminsky, K. Papagiannaki, T. E. Ng, M. Kozuch, and M. Ryan, "c-Through: Part-time optics in data centers,"

in *Proceedings of the ACM SIGCOMM 2010 Conference*, ser. SIGCOMM '10, 2010, pp. 327–338.

[127] N. Farrington, G. Porter, S. Radhakrishnan, H. H. Bazzaz, V. Subramanya, Y. Fainman, G. Papen, and A. Vahdat, "Helios: A hybrid electrical/optical switch architecture for modular data centers," in *Proceedings of the ACM SIGCOMM 2010 Conference*, 2010, pp. 339–350.

[128] K. Chen, A. Singla, A. Singh, K. Ramachandran, L. Xu, Y. Zhang, X. Wen, and Y. Chen, "OSA: An optical switching architecture for data center networks with unprecedented flexibility," *IEEE/ACM Transactions on Networking*, vol. 22, no. 2, pp. 498–511, April 2014.

[129] D. Zhang, J. Wu, H. Guo, and R. Hui, "An optical switching architecture for intra data center interconnections with ultra-high scalability," in *Optical Interconnects Conference, 2014 IEEE*, May 2014, pp. 45–46.

[130] O. Biran, A. Corradi, M. Fanelli, L. Foschini, A. Nus, D. Raz, and E. Silvera, "A stable network-aware VM placement for cloud systems," in *Cluster, Cloud and Grid Computing (CCGrid), 2012 12th IEEE/ACM International Symposium on*, May 2012, pp. 498–506.

[131] Y. Guo, A. L. Stolyar, and A. Walid, "Shadow-routing based dynamic algorithms for virtual machine placement in a network cloud," in *INFOCOM, 2013 Proceedings IEEE*. IEEE, 2013, pp. 620–628.

[132] Y. Zhao, Y. Huang, K. Chen, M. Yu, S. Wang, and D. Li, "Joint VM placement and topology optimization for traffic scalability in dynamic datacenter networks," *Computer Networks*, vol. 80, pp. 109 – 123, 2015.

[133] G. Porter, R. Strong, N. Farrington, A. Forencich, P. Chen-Sun, T. Rosing, Y. Fainman, G. Papen, and A. Vahdat, "Integrating microsecond circuit switching into the data center," *SIGCOMM Comput. Commun. Rev.*, vol. 43, no. 4, pp. 447–458, Aug. 2013.