ABSTRACT

MA, XUEZHOU. Towards an Optimized Internet Backbone Network. (Under the direction of Khaled Harfoush.)

The explosive growth of traffic demand, fueled by the expansion of the Internet in reach and capacity, imposes severe stress on backbone networks. A tier-1 Internet Service Provider (ISP) like AT&T delivers more than 7 petabytes of data per day through its backbone fabric where as little as 1% higher traffic loss will incur several terabytes of retransmissions, almost equivalent to the daily load of a medium-sized regional network, and billions of dollars are spent annually on infrastructure construction and equipment upgrade to satisfy the increasing bandwidth requirements. An optimized design for Internet backbone networks thus benefits both service providers and Internet users worldwide.

Despite the large body of research targeted at optimizing backbone networks, it remains challenging to identify the actual major factors driving the design and create a realistic and tractable model with appropriate design metrics. The development of networking technologies further complicates the problem by continuously shifting the bottleneck factors from some elements to others. Moreover, with the advent of the concept of *green* Internet, legacy models aimed at maximizing the throughput are tuned to a new energy-smart perspective aimed at minimizing the energy footprint. Such rapid evolution of backbone networks leaves many critical issues unsolved, inspiring more investigations and discussions in the research community.

In this dissertation, we study the WDM backbone networks from the bottom up, starting with physical topology and then extending to virtual topology and traffic routing. We make the following contributions. First, we present a physical topology model to determine the number and the choice of constituent fiber links. Our model captures the physical design principles including cost, performance, resilience and geographical constraints, and considers the problem as a tradeoff among all feasible *meshes* that yield best performance for a given budget. The results achieve a similarity of more than 90% with the published ISP structures. Second, existing virtual topology models are highly dependent on the network context. A model that fits one network can perform poor for others. We thus abstract the individual design objectives by iteratively identifying the bottleneck elements and setting up suitable virtual channels accordingly. Our heuristic approach is plug-and-play and averages 28% higher throughput than existing models in all tested networks with different traffic demands and technologies. Third, emerging network design models aimed at maximizing energy savings by aggregating traffic at a small set of resources is offset by legacy models aimed at maximizing the throughput by spreading the load across network resources. We solve this dilemma with a new design perspective which targets optimal power usage for common traffic demand, while accommodating traffic

fluctuations. The proposed heuristic matches the optimality of both factors within 10%, while enjoying polynomial time complexity. Fourth, we make a case that existing backbone traffic routing schemes that minimize the aggregate router power usage are mislead by ignoring the impact of cooling consumption. The actual router power spectrum is polynomial in traffic demand and increases rapidly when the router is loaded. We compare the efficacy of two distinct routing philosophies, traffic *aggregation vs.* traffic *balancing*, and find that mitigating network bottlenecks, rather than creating ones, can save at least 25% energy. Our conclusions challenge the common wisdom about the merit of load concentration strategies.

Towards an Optimized Internet Backbone Network

by
Xuezhou Ma

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Doctor of Philosophy

Computer Science

Raleigh, North Carolina

2012

APPROVED BY:

Rudra Dutta	Harilaos Perros
Douglas Reeves	Khaled Harfoush Chair of Advisory Committee

DEDICATION

This dissertation is dedicated to my wife and my parents who have supported me all the way since the beginning of my studies. This dissertation is also dedicated to my grandmother who passed away in 2007 at age 82 after ten years fighting Alzheimer's disease.

BIOGRAPHY

Xuezhou Ma was born in Xi'an, China in 1981. He attended Xian JiaoTong University from 2000 to 2004, and received his Bachelor of Science in Electrical Engineering. He was then invited into the graduate program of City College of New York in New York City, and received his Master of Engineering with the honors in June 2007. In August 2007, Xuezhou joined the North Carolina State University at Raleigh pursuing his doctoral degree in Computer Science. His major research interests are computer networking, WDM backbone networks, and system optimization.

ACKNOWLEDGEMENTS

First and foremost, I would like to thank my advisor Dr. Khaled Harfoush for his direction and guidance on my research. Without him I would not have achieved my goals for this dissertation. Secondly, I would like to thank my advisory committee members, Dr. Rudra Dutta, Dr. Harry Perros, and Dr. Douglas Reeves, and two other faculty members Dr. Lunardi Leda and Dr. George Rouskas. I sincerely appreciate their constructive suggestions on my work.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

The Internet is a global system of interconnected networks in which a large number of private, public, business, and government networks of local to global scope are linked. An Internet backbone service provider maintains Points of Presence (PoPs) in different locations, typically in densely populated areas, and interconnects these PoPs by means of fiber links, creating a backbone network. Backbone networks interconnect to construct a large network with global reach. In order to gain access to the Internet, edge networks or customers connect to the Internet backbone either directly or through regional networks, which are themselves connected to the backbone.

The explosion of the Internet in reach and capacity is fueled by a significant boost of networking technology at the backbone. Conventionally, each fiber link carries a single signal. Huge optical bandwidth, which is in the order of Tbps, was wasted due to the limited speed of electronic devices at fiber ends. The use of Wavelength Division Multiplexing (WDM) compensates this mismatch by dividing fiber bandwidth into tens of *wavelengths* each able of carrying traffic in the order of Gbps. A virtual channel, *lightpath*, can be established between PoPs by using one wavelength on each link along the path. Once established, it delivers information transparently such that the signal cuts through intermediate PoPs without electronic switching. The set of all lightpaths then forms a *virtual graph* on top of the backbone infrastructure, *physical topology*. Routing and Traffic Engineering (TE) mechanisms are handled over these virtual channels.

A WDM backbone network is a complex engineering structure containing devices in different layers. Without a clear understanding of backbone behavior and the design model at each layer, the resulting Internet backbone networks may lead to 1) increased construction and operation costs and hence a pricier Internet, 2) lower resilience to failures and distributed attacks, 3) higher congestion level and more vulnerability to traffic fluctuations, 4) increased network latency and degraded Quality of Service (QoS), and 5) worse power efficiency and larger Internet energy footprint.

Despite the large body of research targeted at optimizing backbone networks, it remains challenging to identify the actual major factors driving the design and create a realistic and tractable model with appropriate design metrics. The development of networking technologies further complicates the problem by continuously shifting the bottleneck factors from some elements to others. For example, previous research has repeatedly minimized the link/channel utilization to increase the network throughput, while newly deployed infrastructures can support as many as 64 wavelengths on a single fiber with 40 Gbps bandwidth for each. That is several order of magnitude larger than the switching capacity of core routers in the market, implying that the routers are more likely to be the bottleneck in modern networks. Moreover, given that a sizable fraction of total electricity supply in the U.S. is devoured by backbone infrastructure, the concept of green Internet has been highlighted recently. Legacy models aimed at maximizing the throughput are tuned to a new energy-smart perspective aimed at minimizing the energy footprint. Such rapid evolution of backbone networks leaves many critical issues unsolved, such as how to balance the throughput and energy requirements, inspiring more investigations and discussions in the research community.

In this dissertation, we study the WDM backbone networks from the bottom up, starting with physical topology and then extending to virtual topology and traffic routing. We make the following contributions.

1. We present a physical topology model, HINT, to determine the number and the choice of constituent fiber links. Previous models based on the optimization of deployment cost lead to the results not matching the real graphs, while HINT captures the physical design principles including cost, performance, resilience and geographical constraints, and it considers the problem as a tradeoff among all feasible meshes that yield best performance for a given budget. HINT achieves a similarity of more than 90% with the published ISP structures.

2. Existing virtual topology models optimize predetermined objective functions of interest. They are highly dependent on the network context, and a model that fits one network can perform poor for others. Our heuristic approach, BOA, abstracts the individual design objectives by iteratively identifying the bottleneck elements and setting up suitable virtual channels accordingly. BOA is plug-and-play and averages 28% higher throughput than existing models in all tested networks with different traffic demands and technologies.

3. In virtual topology design, a model aimed at maximizing energy savings by aggregating traffic at a small set of resources, to put under-utilized resources to sleep, is offset by legacy models aimed at maximizing the throughput by spreading the load across network resources. We solve this dilemma with a new design perspective which targets optimal

power usage for common traffic demand, while accommodating traffic fluctuations. The proposed heuristic matches the optimality of both factors within 10%, while reducing the computation time to less than 40 minutes compared to 22 hours for optimal solutions.

4. Existing backbone traffic routing schemes that minimize the aggregate router power usage are mislead by ignoring the impact of cooling consumption. We show that the actual router power spectrum is polynomial in traffic demand and increases rapidly when the router is loaded. We compare the efficacy of two distinct routing philosophies, traffic aggregation *vs.* traffic balancing, and make a case that mitigating network bottlenecks, rather than creating ones, can save at least 25% energy. Our conclusions challenge the common wisdom about the merit of load concentration strategies.

The rest of this dissertation is organized as follows. In Chapter 2, we survey research related to the Internet backbone network design. In Chapter 3, we provide the details of the HINT heuristic and use it to emulate ISP physical structures. In Chapter 4, we introduce the BOA model and show how it adaptively fits networks with different features. In Chapter 5, we propose our design formulation, along with the heuristic, to balance the optimality of network throughput and energy savings. In Chapter 6, we describe the details of a novel FBAUR routing to minimize the energy footprint. We finally conclude in Chapter 7.

Chapter 2

Related Works

In general, there are two main approaches to network topology design, namely *statistic modeling* and *optimization*. Statistic modeling depends on an emerging picture of the large-scale statistical properties of networks which are acquired through careful collection and interpretation of topology-related measurements [50, 63]. Measured properties are used in designing and testing new topologies. Optimization techniques [11,24,32,33,59,60] attempt to find the optimal design result for a pre-determined objective function with subject to network resource constraints. The objective function is formalized to reflect metrics of interest. Since statistic modeling represents an average feature map, design models aiming at individual networks mostly use optimization approach. The optimization problem is typically casted as a Linear Programming (LP) problem [24, 60] and solved using software packages such as *Cplex* for small sized cases. To make it tractable for real networks, heuristic algorithms are needed to obtain the approximate or near-optimal solutions.

2.1 Network Physical Topology Design

Previous physical topology designs rely on minimization of the costs. The work in [24, 60] considers the case where the cost is proportional to the number of fiber links regardless of their lengths. The topology is restricted to a regular graph in [24]. In [59], the authors take fiber length into consideration because the installation cost, believed to be the dominant part of the network cost, is generally proportional to link length [12,32]. The cost function in [33] combines the construction and equipment costs by using the metrics, aggregate fiber length and number of wavelengths, respectively. Meanwhile, [13, 39] explore the physical structure based on network survivability. In [39], *Modiano et al.* study the necessary condition for survivable topologies using the concept of *cutset* in graph theory. In [13], the authors introduce a *two-connected mesh* graph for establishing physical links. Sparse meshes then become popular in [26, 33, 45].

Unfortunately, existing models can not justify their results by comparing to real ISP networks. The cost and survivability may not be the only factors in determining network structure.

The problem of designing a physical topology to optimize the number of wavelengths is known to be NP-hard [33]. Minimizing the total link length without ensuring a 2-connected graph is equivalent to searching for the *minimum spanning tree*. The complexity is $O(V^2)$ (V is the number of nodes) by using Prim's algorithm. Computational complexity for examining two-connectivity grows exponentially with the size of the network [26]. Heuristic algorithms are proposed for obtaining a survivable network with minimum cost. Some studies [26, 33] use a *constructing* algorithm starting with a minimum spanning tree and then adding the links which render maximum improvement on the objective. Others [13, 60] rely on *greedy* approaches to exchange the configurations of neighboring nodes during each iteration.

2.2 Network Virtual Topology Design

For years, WDM virtual topology design has been formulated to maximize the network *throughput* subject to constraints on network resources. Many popular design metrics have been considered and each of them was shown optimal for only certain networks or scenarios. In [13], the *packet hop distance* is minimized since the average hop distance is inversely proportional to the network throughput, when traffic demand is balanced across the network. In [31], the *maximum link utilization* is minimized because an overall throughput *scale-up* was often limited by over-congested links. In [53], a virtual layout with the *minimum usage of network resources* (e.g., wavelengths) is selected to fight against traffic fluctuations. In [41], reducing the *demand delay product* (i.e., the product of traffic rate and end-to-end queuing delay), equivalent to reducing the amount of data "on the line", is used to maximize the network throughput. In the case router ports are scarce resources, maximizing the *channel load* by grooming low-rate flows into one channel is shown more appropriate in [66]. Despite the rich literature in this area, relying on one objective function in all cases is expected *not* to lead to the maximum throughput in all networks independent of bottleneck elements in these networks.

The heuristics introduced for virtual topology design are as follows. MMT [41] primarily connects distant nodes with large demand. MLDA/MST [19] establishes lightpaths between node pairs in descending order of their traffic intensity. MRU [66] first places lightpaths between neighboring nodes, then between nodes two-hop away and so on, upon the availability of wavelengths. Two other greedy approaches, simulated annealing [61] (SA in short) and LPLDA [49], apply an iterative algorithm by starting with an initial random virtual graph and adjusting the configuration of two adjacent nodes during each iteration. SA and LPLDA are computationally expensive. The lack of knowledge about the bottleneck location hinders the efficacy of the adjustments, creating a large number of less-relevant intermediate steps.

2.3 Network Robustness and Energy Efficiency

Traffic dynamics are extensively studied and mitigated by by optimizing traffic routing through Traffic Engineering (TE) techniques. Early efforts [22, 23] have investigated the routing optimization problem in the case of a single traffic matrix. In [23], *Gallager* proposes a distributed gradient-based algorithm to solve the problem by changing the routing variables. The work in [22] approximates the link utilization as piece-wise linear functions, and solves the problem using LP. To emulate the different possible traffic spikes, the literature is formalized to minimize the *maximum link utilization (MLU)* in the presence of multiple traffic matrices (Surprisingly, very little work has been found on minimizing the *maximum router utilization*). In [65], *Zhang et al.* consider a set of representative traffic patterns and search for an optimal routing to minimize the MLU over all representative patterns. In [57], the authors distinguish normal operation and link congestion and propose a heuristic to cope both traffic spikes and common demand. While there is a large body of research in the TE field, the proposed solutions do not apply directly to the WDM virtual topology design due to the lack of consideration to the optical layer.

Energy concerns are highlighted recently. Research efforts in WDM network design have shifted to network power models and LP formulations in order to minimize energy footprint [15, 28, 48, 55]. In [48], an objective function that includes the power consumption of routers, transponders, and amplifiers is minimized for given traffic demand. Two heuristics, one with all routers bypassed by the virtual channels and the other with no router bypassed, are shown in [48] to find near-optimal virtual designs. In [28], the authors consider the network power model which consists of electronic routers and optical switches, and propose a flow-based and a interface-based formulations to measure the overall power usage. However, as far as we know, none of existing models attempt to handle both energy savings and network performance.

For traffic routing at IP layer, since routers are primary contributors to energy consumption, the literature has investigated an accurate router power model. In [56], the authors measure the power of a router from a scale of the components by estimating the consumption of the processor, the switching fabric, buffers, *etc.* This method is appropriate for previous-generation routers (e.g., *Alpha 21364*) with simple structure. For modern routers, the authors in [15, 38] use a digital multimeter attached to the power cable of the router to measure system-wide consumption. In [15], 20 *Pentium4* workstations are connected together to generate up to 1.8 Gbps TCP traffic over a 40 Gbps *Cisco 12008 GSR*. The observed router power usage increases linearly with the number of active line cards, or approximately, with the traffic rate crossing the router. Inspired by that, [43, 62] aim at reducing network energy footprint by aggregating most traffic along few routers only. The work in [16] further abstracts the network design problem as an optimization to minimize the set of active routers given traffic demand. While

a large body of research follows the traffic concentration strategy, a recent survey on major ISPs indicates that router racks contribute only a fraction of the utility bill [51]. The power requirements for the cooling system are already a significant part of the overall PoP power requirements [21]. Unfortunately, the impact of cooling consumption is rarely considered in network design problems but more addressed in Mechanical Engineering. For example, Chu *et al.* [17] showed a prototype of a hybrid router cooling system confirmed by Cisco technician lead. Maddren *et al.* [36, 37] implemented this design on Cisco 7609s router and performed a benchmark testing. The detailed fluid mechanic model of a pipe flow can be found in [42].

Chapter 3

HINT: A Realistic Physical Topology Model

The physical topology of a network is expressed as a graph $G = (V, E)$, which identifies the physical layout of devices in the network, V, together with the way devices are interconnected through actual cables, E. In this chapter, we investigate the physical topology of Internet backbone networks, in which V is the set of PoPs and E is the set of fiber links interconnecting them. Specifically, our aim is to answer the following question: *Given the POP locations of a service provider, enabling technologies and traffic demands, how should the service provider lay out fiber links and what are the driving forces behind this layout?*

Identifying the actual major factors driving the design of the physical topologies of Internet backbone networks, and understanding how they interact, is essential since a practical design 1) can reduce the capital investment for ISPs, 2) directly impacts higher layer protocols and applications [13], and 3) provides critical support for Internet researchers in need of practical network models. Most research on physical topology design has focused on deployment cost [24, 33,59,60]. Later studies pointed out that due to technological, cost and performance constraints, Internet backbone networks have a *sparse mesh* structure [11, 32] – Refer to Figure 3.1 for an example. The study in [11, 32], however, does not explain how a sparse mesh is constructed, i.e., which links are established.

Towards this end, in this chapter, we provide a physical topology model, which captures major physical design factors. While it is clear that the set of design factors is far from being unique, our study is focused on the *minimal* set, which we believe is fundamental to any physical topology design. We consider the following factors: 1) the cost of the infrastructure, 2) the expected performance, 3) the geographical constraints, and 4) the resilience of the network to link/node failures (*survivability*). The model results from the solution of a constrained optimization problem. Obtaining an optimal solution to the optimization problem is shown to

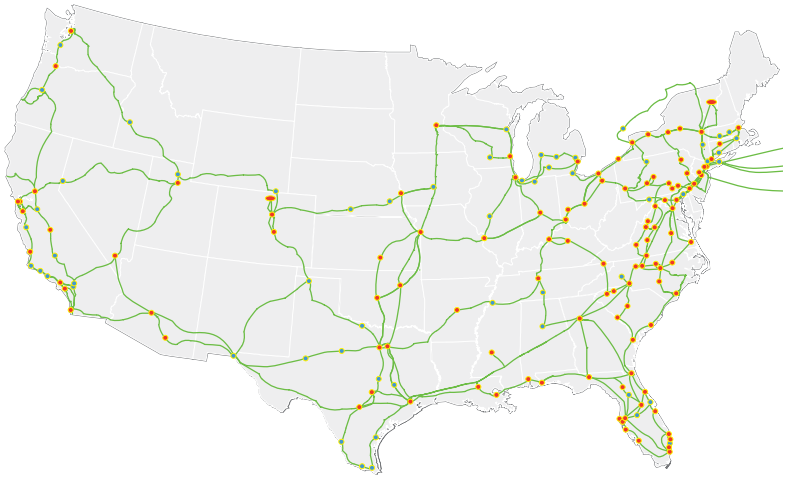


Figure 3.1: Published network map: Level3 network [5].

be NP-hard. We thus introduce a polynomial time heuristic algorithm, HINT, to determine the number and the choice of the constituent links. The efficacy of HINT is established in comparison with the published maps of three major scientific and commercial backbone networks: Internet2 Abilene, AT&T domestic express backbone, and Level3 network. Our results reveal that taking performance and resilience into consideration is necessary to emulate real backbones. The HINT heuristic yields a similarity of more than 90% with the published structures.

The rest of this chapter is organized as follows. In Section 3.1, we discuss the physical topology design factors. In Section 3.2 and 3.3, we formulate the optimization problem and the HINT heuristic. In Section 3.4, we compare HINT results with the published ISP networks. Summary is in Section 3.5.

3.1 Network Design Factors

We consider 1) the cost, 2) the expected performance, 3) the geographical constraints, and 4) the resilience of the network to link/node failures as the basic design factors for Internet backbone networks. Let $G = (V, E)$ be the graph representing the physical topology with V as the set of nodes and E as the set of links. Let the traffic matrix $\Lambda = \{\lambda_{ij}\}$ denote the aggregated traffic (in arbitrary traffic units) between nodes v_i and v_j, with internal traffic $\lambda_{ii} = 0$ for any $v_i \in V$. Let d_{ij} be the fiber length of $e_{ij} \in E$. d_{ij} can also be expressed as propagation delay (in time units) on e_{ij}. Note that $d_{ij} = d_{ji}$, and $d_{ij} = \infty$ if $e_{ij} \notin E$. Let D be the shortest path delay matrix, where D_{ij} denotes the aggregated fiber length over the shortest path between nodes v_i and v_j.

3.1.1 Traffic Model

The demand for Internet service is a major driving force in network design. [18] indicates that the average aggregate traffic between two nodes (PoPs) can be approximated by their population. Since backbone nodes are usually located at major cities, two cities with higher population are likely to exchange more traffic. We thus rely on the traffic model introduced next.

Let δ_i be the population of node v_i. Suppose v_i exports δ_i (in traffic units) to G and imports/downloads δ_i from G. The imported δ_i units are downloaded from nodes of G in proportion to their δ values. The exported δ_i units are uploaded to all other nodes also in proportion to their δ values. Let $\sigma = \sum_{i=1}^{|V|} \delta_i$. Then a node v_i uploads a fraction $\frac{\delta_j}{\sigma}$ of its δ_i traffic units to node v_j and downloads a fraction $\frac{\delta_i}{\sigma}$ of δ_j from v_j. Note that this model generates the same amount of traffic in both directions (originating at v_i and destined for v_j and vice versa). The end-to-end traffic between v_i and v_j can be expressed as:

$$\lambda_{ij} = \delta_i \frac{\delta_j}{\sigma} + \delta_j \frac{\delta_i}{\sigma} = \frac{2\delta_i\delta_j}{\sigma} \tag{3.1}$$

Refer to Figure 3.2 for an example. In Figure 3.2 (left), a directed link from a node v_i to node v_j is annotated with the traffic from v_i to v_j. Figure 3.2 (middle) provides a more compact representation in which we summed up the traffic units in both direction instead of distinguishing them. Figure 3.2 (right) then shows the equivalent traffic carried on fiber links. Self cycles are not considered in our model as they do not affect performance in Section 3.1.3. Note that the sum of the traffic units over all links in this representation equals σ as expected. Also, λ_{ij} in our traffic model is proportional to the product of δ_i and δ_j. At the same time, the aggregate traffic at one node (e.g., $\sum_j \lambda_{ij}$ for node v_i) keeps linear form of δ. That differs from the gravity model in [47].

3.1.2 Economics and Geographical Constraints

Cost plays an important role in determining the physical topology. The cost for constructing a backbone network has many facets. Some costs are *one-time* infrastructure investments such as purchasing fibers and optical switches, digging and installing fibers underground. Other costs are *recurring* such as hiring personnel to run the equipment and the associated overhead.

In general, the digging cost is proportional to the distance between nodes; the fiber cost is decided by both length and quantity. The cost of equipment and the overhead are proportional to the traffic load on nodes. For backbone networks, installing fiber links requires substantial capital investment and the installation process is extremely slow. It is thus widely believed that digging and laying out fibers dominates the remaining costs [12, 32]. Therefore, there is

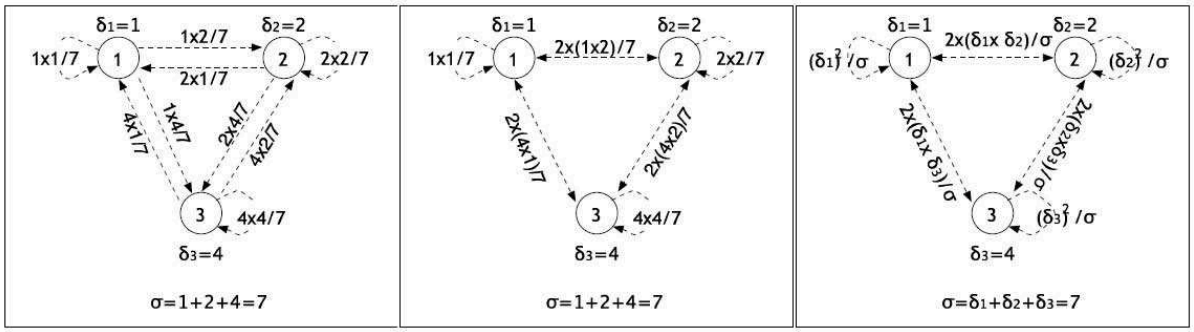


Figure 3.2: An illustration of traffic model in a 3-node graph with $\delta_1 = 1$, $\delta_2 = 2$, and $\delta_3 = 4$ (left). A more compact representation of the traffic flows between nodes (middle); A symbolic representation of flows(right).

tremendous practical incentive in designing the network with minimum total link length, C_l:

$$C_l = \sum_{e_{i,j} \in E} d_{i,j} \tag{3.2}$$

Geographic constraints have been neglected in literatures since they do not lend themselves naturally to direct inspection. In practice, however, geographic limitations (mountains, bridges, etc.) have made it difficult to install fibers following the *Cartesian* distance [29]. Moreover, such limitations may prevent a direct connection between two cities (refer to the connection between Raleigh and Asheville in NCREN [8]) and further restrict their node degree. To tackle this problem, we use the trip distance obtained from online mapping tool (*Google map*) to account for the realistic fiber length of a link. That is possible because the fiber conduit is often installed as part of highway construction project [20]. The trip distance is thus more accurate in reflecting the construction cost.

3.1.3 Quality of Service

Popular metrics such as link utilization and queuing delay are ignored in our model because 1) switching in the optical domain is much faster than in the electronic domain; 2) the technical configurations vary at different PoPs; 3) backbone links are typically over provisioned. Instead, in this chapter, we consider the *latency* as the main performance metric. The end-to-end latency usually consists of two components: propagation delay and queuing delay. The propagation delay for each source-destination pair can be expressed (in time units) as the aggregated length of fibers the signal passes through. The queuing delay defines the processing time at intermediate switching nodes. The average delay for all node pairs, C_d, is then defined as follows:

$$C_d = \frac{1}{\sum_{i \in V} \sum_{j \in V} \lambda_{i,j}} \sum_{i \in V} \sum_{j \in V} \lambda_{i,j} D_{i,j} \tag{3.3}$$

Our latency estimation is based on a model in which end-to-end traffic follows the shortest path route unless some links along the route carry excessive transit traffic. In that case, it will switch to the second shortest path subject to the route length bound α ($1 \leq \alpha < \infty$) which bounds the actual route (and hence the end-to-end delay) between two nodes with respect to the shortest distance between them. The tradeoff between two scenarios is clear: the former configuration provides less latency while the later setup helps to balance the traffic.

3.1.4 Survivability

Survivability is important for Internet backbone networks carrying huge amounts of traffic. With popular WDM technology, it becomes even more critical because multiple wavelength channels traverse the same fiber links and would fail simultaneously in the event of a link failure. It is thus necessary to ensure a survivable physical topology design. Without this, any protection at upper layer protocols will not be effective [26].

A physical topology is considered to be survivable if it can cope with any single failure of network components by rerouting the affected traffic to alternative paths. In graph theory, a *separating* set of a graph G is a set of nodes whose removal renders G disconnected. The connectivity of G, $\kappa(G)$, is the minimum size of the separating set, which means the graph is guaranteed to be still connected even if any $\kappa(G) - 1$ nodes fail [58]. Clearly, a survivable physical topology must be a two-connected graph [26].

Given $v_i, v_j \in V$, a set $S \subset V - \{v_i, v_j\}$ is a v_i, v_j-*cut* if $G - S$ has no v_i, v_j-path. Let $\kappa(v_i, v_j)$ be the minimum size of an v_i, v_j-cut. *Menger*'s theorem [58], given below, determines the connectivity of a network by examining $\kappa(v_i, v_j)$ for every node pairs.

Theorem 1. *A graph $G = (V, E)$ is 2-connected ($\kappa(G) = 2$) if and only if for all $v_i, v_j \in V$, there is no v_i, v_j-cut of size less than 2.*

$$\kappa(v_i, v_j) \geq 2, \forall v_i, v_j \in V \tag{3.4}$$

3.2 Problem Formulation

Our design involves a manipulation of cost, performance, and survivability. Refer to Figure 3.3. Given sufficient budget, one can build a network with redundant links (i.e., complete graph) to provide superior performance and survivability. An excessively large budget is not always available though. On the other hand, a limited budget may require nodes to be connected in

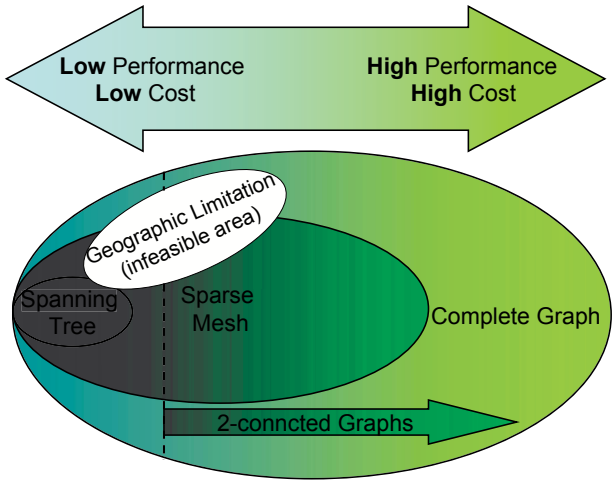


Figure 3.3: The basic idea of our physical topology design.

an economical fashion (i.e., spanning tree) at the cost of the two other factors because tree structure is not 2-connected. And all configurations are subject to geographical constraints (i.e., infeasible area).

The problem is thus aimed at exploring the tradeoff among mesh networks. While there are many possible combinations for a mesh topology, it is clear that each link, once exists, should make substantial contributions towards part or all design factors. Otherwise, if it requires large investment with little impact, then it is likely that service providers will not include it in their plan. In essence, we are asking *for a given budget, what is the survivable physical topology that yielding best performance?*

Given:

- Number of nodes in the network $|V|$
- Traffic matrix $\Lambda = (\lambda_{ij})$
- Trip distance d_{ij} representing the fiber length of $e_{ij} \in E$, $d_{ij} = \infty$ if some e_{ij} does not exist in reality.

Variables:

- Number of links in the network $|E|$
- Physical topology $E = (e_{ij})$: $e_{ij} = 1$ if two v_i and v_j are adjacent and zero, otherwise. $e_{ij} = e_{ji}$ holds for a bidirectional graph, where $e_{ij} \in \{0, 1\}$.

- Shortest path route S: $S_{ij}^{sd} = 1$ if the shortest available path between s and d is routed on physical link e_{ij} and zero, otherwise.
- Second shortest path route SS. Similarly, $SS_{ij}^{sd} = 1$ if the second shortest available path between s and d is routed on physical link e_{ij} and zero, otherwise. If there are more than one path having the same aggregated lengths, select one of them randomly. If no second shortest path exists between s and d, $SS_{ij}^{sd} = \infty$.
- Physical topology route. Let $PM_{ij}^{sd} = 1$ if the actual path between s and d is routed on physical link e_{ij} and zero, otherwise. In a bidirectional graph, a path from s to d is also the path from d to s.
- Link traffic f_{ij} denotes the total traffic being routed through the link e_{ij}. Note that the link traffic is computed as $f_{ij} = \sum_{s,d} PM_{ij}^{sd} * \lambda_{sd}$.

Objective:

$$\text{Minimize: } C = \gamma C_l + |E| C_d \tag{3.5}$$

Subject to:

- Survivability constraint:
  $\kappa(v_i, v_j) \geq 2, \quad \forall v_i, v_j \in V$
- Route constraint:
  $PM_{ij}^{sd} = \begin{cases} SS_{ij}^{sd}, & \sum_{i,j} SS_{ij}^{sd} * f_{ij} < \sum_{i,j} S_{ij}^{sd} * f_{ij} \\ S_{ij}^{sd}, & \text{otherwise} \end{cases}$
- Route length bound:
  $\sum_{i,j} PM_{ij}^{sd} * d_{ij} \leq \alpha D_{sd}, \quad \forall s, d \in V$

Notice that $|E|$, the number of links in G, is multiplied by C_d in Equation (3.5). This parameter is used to normalize the objective function as C_l is the total length of the links, while C_d shows the average performance. It is possible to combine the cost and the performance into a single number because a higher quality of service translates into a higher profit, e.g., ISPs usually increase the monthly charge for residential/commercial users to improve the guarantee of bandwidth. Note that such formalization is not unique. Our objective is designed to be generic such that one can easily weigh more importance on either component by changing the value of γ ($0 < \gamma \leq 1$).

Computational complexity of the optimal solution for physical topology design problem can be easily proved to be NP-hard. In [27], *Hu et al.* argue that getting the minimum value for C_d satisfying certain traffic pattern is an NP-hard problem. Also, the complexity for searching cutsets grows exponentially with the size of the network, since a network with N nodes would yield $2^N - 2$ cutsets [26].

3.3 Our Heuristic Algorithm

Heuristic become important as the size of the network gets larger. In general, there are three different patterns of heuristics for topology design. A *constructing* model (e.g., *HLDA* [46]) has the initial link set empty, $|E| = 0$, and the network grows by adding new links. In contrast, a *de-constructing* model sets initially a full mesh graph by assuming there is a fiber link between all node pairs, then removes the links which are less relevant. *MLDA* [46] and Simulated Annealing [40] represent yet another paradigm which starts with the minimum spanning tree or a random layout, respectively.

In this section, we introduce a heuristic algorithm, HINT, to optimize the objective function in a de-constructing fashion while ensuring that all constraints are held throughout. The HINT algorithm works as follows:

- **Step 1.** Start with a complete graph. Set $D_{i,j} = d_{i,j}, \forall i, j \in V$. Route the traffic following the shortest path routes. Compute the link traffic f_{ij}.
- **Step 2.** Identify the link, $e_{u,v}$, which if removed would reduce the value of C (Equation (3.5)) the most while the survivability constraint and route length bound still hold. If there is no solution, go to Step 5.
- **Step 3.** Remove link $e_{u,v}$ from the network. Recompute the actual path matrix PM upon the removal of $e_{u,v}$.
- **Step 4.** The traffic which was using $e_{u,v}$ is redirected following the new path. Update link traffic f_{ij} accordingly. Goto step 2.
- **Step 5.** Output the link selection and value for C, C_d, and C_l.

In Step 1 of the HINT algorithm, the total link length C_l takes its maximum value while the average delay C_d is minimal. For each iteration (Steps 2 and 3), we attempt to remove the link with large length and small *transit* traffic. Before removing a link, we ensure that G will not lose its 2-connectivity. The traffic carried by the dropped link is then switched to an alternative path. If both the shortest path (S_{ij}) and second shortest path (SS_{ij}) are available, the new alternative path follows the one with less aggregate traffic on the route. Note that during the first rounds, the value for C continuously decreases since the component of total length C_l drops faster than the increase in the average delay C_d. This process continues until the value of C begins to increase.

The survivability constraint in HINT requires to test the cut sizes for all possible node pairs at every iteration, which can easily overwhelm our computational resources. To tackle

this problem, we only verify the $\kappa(v_i, v_j)$, the minimum size of an v_j, v_j-cut, for 2-connectivity of the resulting network in our de-constructing heuristic.

Lemma 1. *In the HINT algorithm, after the removal of any link e_{ij} from graph G, $\kappa(v_i, v_j) = \kappa(G)$.*

Proof. Suppose G is a 2-connected graph. Remove an link e_{ij} from G. To calculate $\kappa(v_i, v_j)$, we select and remove any one node x (other than i and j) from G:

- case 1: i and j become disconnected (i.e., $\kappa(v_i, v_j)$=1). Then $\kappa(G) = \kappa(v_i, v_j) = 1$.
- case 2: i and j are still connected (i.e., $\kappa(v_i, v_j)$=2) and all other node pairs in G are also connected. Then $\kappa(G) = \kappa(v_i, v_j) = 2$.
- case 3: i and j are still connected (i.e., $\kappa(v_i, v_j)$=2) but some other node pair in G, for example, u and v (other than x) becomes disconnected. $\kappa(G) = 1$ in this case. G was originally 2-connected. If we keep the link $e_{i,j}$ for case 3, u and v would have at least one u, v-path after x's leave. That implies that link $e_{i,j}$ is part of the unique path connecting i and j, which conflicts with assumption of case 3 that $\kappa(v_i, v_j) = 2$. Therefore, case 3 does not exist.

□

Lemma 1 helps us reduce the computational complexity in verifying the survivability constraint to $O(P)$ where P is the complexity for searching paths between two nodes, typically $O(VlgV)$. We have initially $|V|(|V|-1)$ possible links. The complexity of the HINT algorithm is thus $O(|V|^5 lg(|V|))$.

3.4 Performance Evaluation

To study the efficacy of proposed problem formulation and heuristic algorithm, we consider three published backbone optical networks in the U.S: 9-node Abilene (2002) [3], 41-node AT&T domestic express backbone (2005) [64] (AT&T for short) and 158-node Level3 network (2008) [5]. For population count, we obtained the data from U.S. census 2004 [4].

Our first experiment targets on the Abilene network. As shown in Figure 3.4, by optimizing only one factor (cost C_l or performance C_d), the resulting physical topologies are not accurate. Optimizing only cost is likely to produce a graph with the *shortest* links, while optimizing only performance will place links between big cities regardless of their distance, e.g., long pipe between NYC and Los Angeles in Figure 3.4(b). Also, without the survivability constraint, one link failure (Chicago-Kansas City in Figure 3.4(c)) will disintegrate the network into two

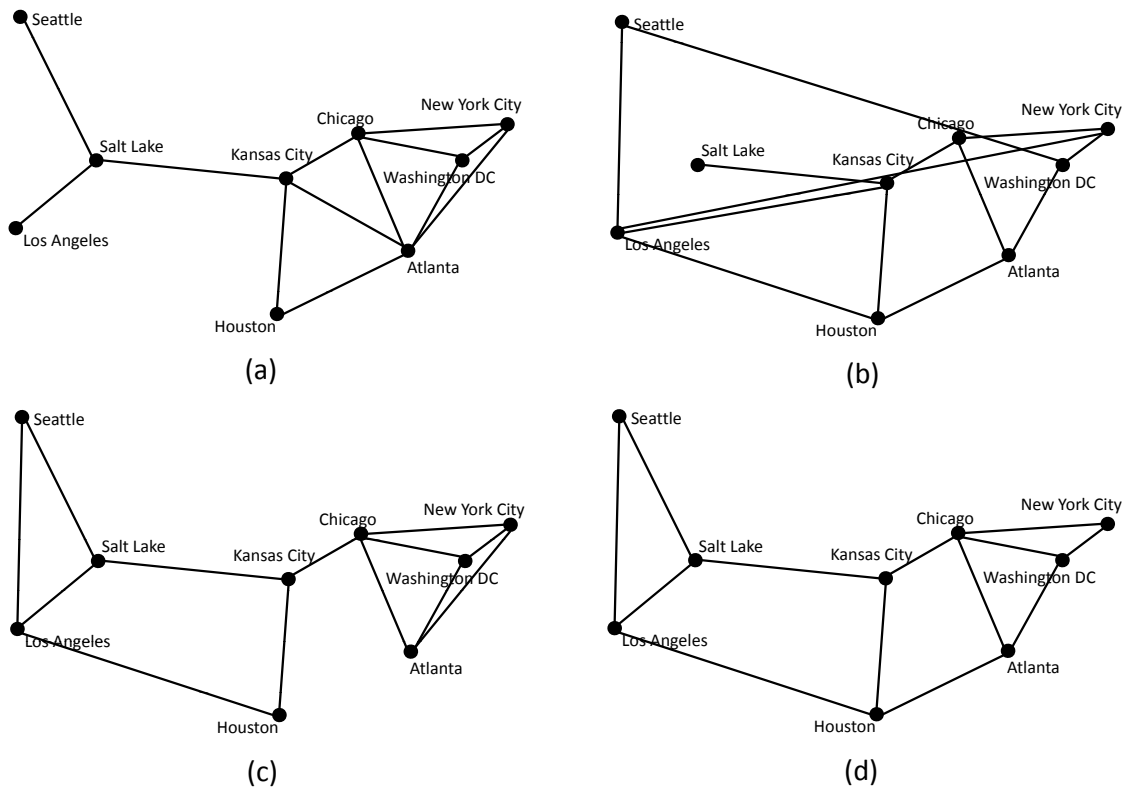


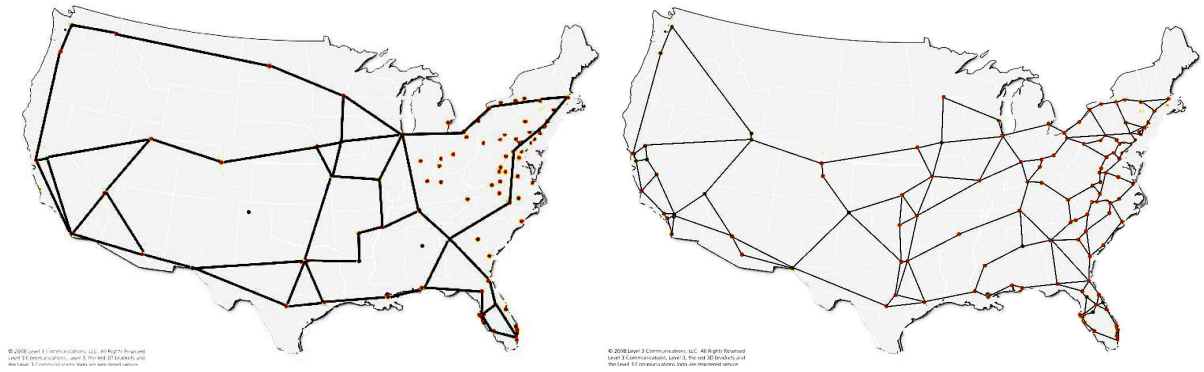
Figure 3.4: Physical topology designs for 13-link Abilene network. (a) optimizing only cost C_l ($\gamma = 10$), (b) optimizing only performance C_d ($\gamma = 0.1$), (c)optimizing both cost and performance ($\gamma = 1$) without survivability constraint, (d) optimizing both cost and performance ($\gamma = 1$) with survivability constraint

components in almost equal size (i.e., Minimum Equally Disconnecting Set (MEDS) = 1) which is not acceptable for most backbone optical networks. It is clear that the combination of cost, performance and survivability is the key to accurate formulation for Abilene. HINT network shown in Figure 3.4(d) has the same physical topology as the published Abilene graph.

Then we run HINT on AT&T and Level3 networks. The graphical results (shown in Figure 3.5) are obtained with the parameters $\gamma = 1$, and $\alpha = 2$. In general, the HINT approach is capable of accurately modeling the topology (link layout) in both cases compared to the published maps [5, 64]. In order to quantify the efficacy of our heuristic in large scale networks, we introduce a measure of *similarity* between the HINT graph H and the published map G. For each node pair (i, j), we refer to the link $e_{i,j}$ a *matching link* if $e_{i,j}$ exists in both H and G; *false positive link* if $e_{i,j}$ exits in H but not G and *false negative link*, otherwise. Let the total number of matching links be l_m, the number of false positive links be l_p. We define $S \equiv \frac{l_m}{l_m+l_p}$ as the *similarity* between H and G. If H and G both have the same number of links, then l_p

Table 3.1: Comparison of HINT networks with published maps

Network	Heuristic	Performance Metrics			
		C_d	C_l	κ	S
Abilene	published	1894	8917	2	-
	HINT	1894	8917	2	100%
AT&T	published	1751	18516	2	-
	HINT	1674	17624	2	91%
Level3	published	1595	26780	2	-
	HINT	1550	25889	2	93%

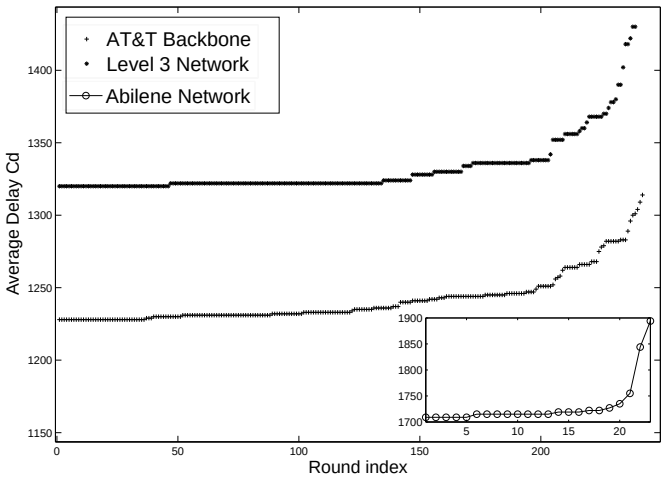


(a) (b)

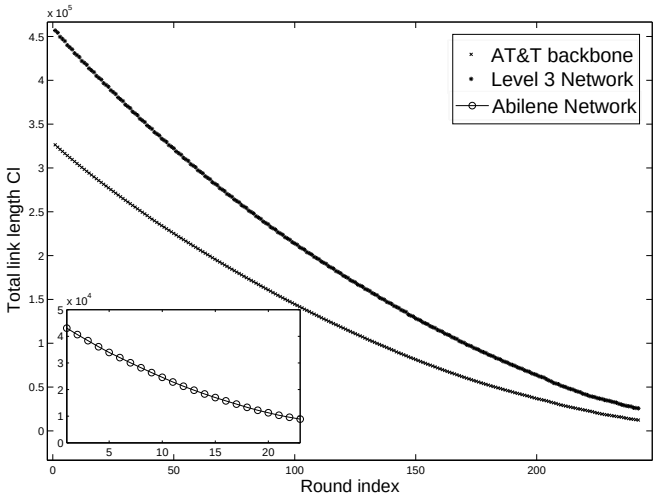
Figure 3.5: HINT results for (a) AT&T and (b) Level3. Shaded background image is from [5] with permission to use.

will always equal to l_n. Thus similarity S quantifies the fraction of matching links among all links.

As shown in Table 3.1 HINT yields 91% and 93% similarity on AT&T and Level3, respectively. The average propagation delays in HINT are 4.4% and 2.8% less than real networks. Cost-wise, HINT lays out 17624 and 25889 miles of fiber link to guarantee 2-connectivity which is 4.8% and 3.3% less than the real networks for the two cases. Interestingly, the resulting HINT networks enjoy better overall performance than the published maps. That is possible considering the simplification we made on the cost model. To exhibit the evolution of metrics in our heuristic, we plot the values of average delay C_d (Figure 3.6 (a)) and total link length C_l (Figure 3.6 (b)) during each iteration of the HINT algorithm. It is clear that C_d does not significantly increase until 95% of links are removed. The curve for C_l, on the other hand, drops



(a)



(b)

Figure 3.6: The change of (a) total link length C_l and (b) average delay C_d through HINT optimization.

sharply at the beginning and gradually slows down as link removal process continues. This suggests that starting from a full mesh, the HINT algorithm successfully optimizes the order of link removal while avoiding sharp performance degradation.

3.5 Summary

In this chapter, we study the physical topology design problem for Internet backbone networks. We introduce a new problem formulation combining cost, performance, and survivability to provide a framework for a realistic design. We use fiber length, average propagation delay, and 2-connectivity to represent important design factors. Furthermore, we introduce a polynomial time heuristic algorithm, HINT, to solve the problem. Preliminary results indicate that each of the three considered factors is necessary. By optimizing them together, HINT accurately models the backbone sparse mesh structures and provides backbone network structures that are almost matching the published ones.

Chapter 4

The Efficacy of WDM Virtual Topology Design Strategies

Existing WDM virtual topology models mainly aim at maximizing the network throughput. Maximizing network throughput is reasonable given that ISPs are profit-driven and a higher throughput translates into a higher profit. Still, distinct models optimize distinct objectives, which result in different virtual graphs. For example, heuristics in [13, 41] primarily connect distant nodes with large demand; while approaches in [54, 66] tend to place lightpaths between neighboring nodes. Refer to Figure 4.1 for two published ISP graphs: National LambdaRail (NLR) and Sprint network. In Sprint, many *end-to-end* lightpaths such as (SJ - RLY) are set up, resulting in a fairly convoluted virtual topology. NLR, on the other hand, exhibits purely *point-to-point* pattern with no lightpath bypassing any intermediate node. This observation, however, creates much confusion as optimizations aimed at the same goal lead to virtual graphs of totally different topological features.

In this chapter we aim at explaining this paradox from a new perspective rather than focusing on individual design strategies. As network throughput is limited by bottleneck elements, whether at routers or links, it is essential for an effective model to identify these elements. We show that none of the existing models identify bottleneck elements in all cases. They abstract the problem with one fixed objective assuming that the throughput hindrance is uniform across the network, and do not consider node structure nor router utilization. As a result, a model suitable for one network (or one branch) may not be suitable for others. Our results reveal that the distinctness of ISP network bottlenecks is the main cause for the distinctness in their corresponding virtual graphs. Furthermore, we introduce a novel algorithm to determine a network bottleneck based on the 1) physical topology, 2) traffic demand and 3) technology constraints, and a topology model leading to optimized network throughput.

The rest of this chapter is organized as follows. In Section 4.1, we discuss the virtual topology

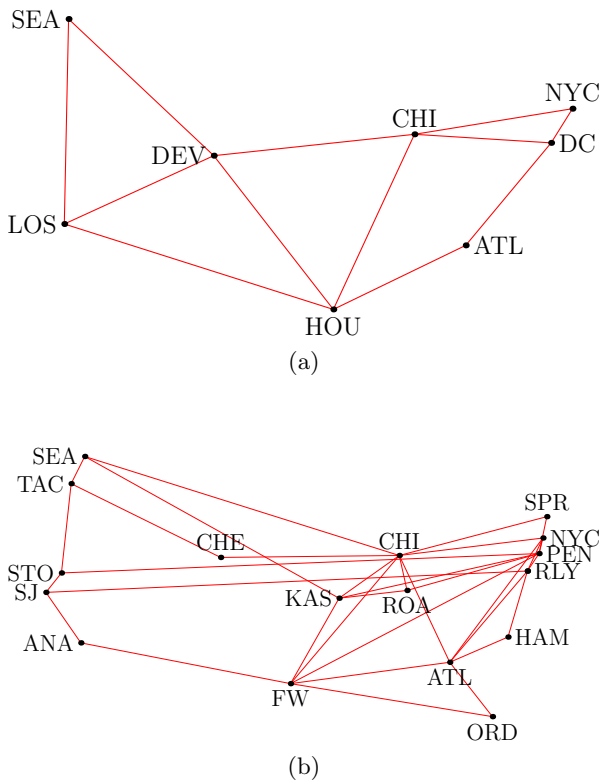


Figure 4.1: Virtual topology maps. (a) NLR [7] and (b) Sprint [44]. Note that each edge in the maps may represent multiple parallel lightpaths.

design factors. In Section 4.2, we propose our approach. In Section 4.3, we present the simulation results on published and synthetic ISP networks. Summary is in Section 4.4.

4.1 WDM Virtual Topology Design Factors

In this section, we make a case that in order to obtain realistic and optimized virtual topologies the following factors need to be considered: 1) node structure, 2) link/router utilization, 3) traffic demand, and 4) technology constraints.

In general, backbone networks consist of two types of nodes: *PoP nodes* which are equipped with both Optical Cross-Connects (OXCs) and routers, and *OXC nodes* which contain only optical switches. As shown in Figure 4.2, the OXCs switch lightpaths in optical domain from input links to output links, and *core routers* switch data packets (when converted into electronic domain) over the lightpaths. Traffic from local area is multiplexed at *access routers*. A pair of *transponders (TX)*, i.e., one transmitter and one receiver, are employed at the endpoints of each lightpath for data transmission and E-O/O-E conversions. Note that OXC nodes can not

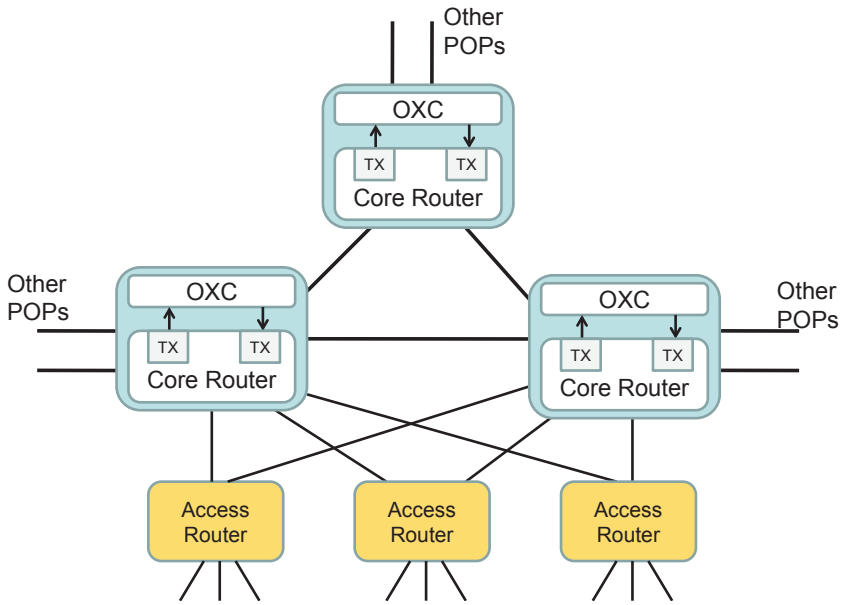


Figure 4.2: A typical structure of the backbone PoP node. The OXC node structure can be obtained by removing the routers from this figure.

originate/terminate lightpaths, and while they should show in a network physical graph, they are absent from the virtual graph. For example, by comparing Sprint physical topology, Figure 4.3, with the corresponding virtual map, Figure 4.1 (b), we find that the outgoing traffic at Atmore is not forwarded directly. Instead, it is first aggregated at Orlando PoP through regional fabric (not shown), then transmitted to the destination by taking one of the two output lightpaths (ORD-FW) and (ORD-ATL). Existing WDM virtual topology models assume that all nodes were created *equal*, as PoP nodes, which leads to unrealistic node connectivity and unoptimized topologies.

Furthermore, in early networking systems, both the number and the capacity of wavelengths were limited. For example, the RACE project [34] in 1998 had only four wavelengths on one link with each running at 2.5 Gbps. That is why researchers have repeatedly minimized the lightpath utilization to increase the throughput. Recently, however, optics technologies have improved tremendously. Newly deployed infrastructures support as many as 64 wavelengths at 40 Gbps each [34], which is far more than the capacity of core routers in the market. As an illustration, Figure 4.4 compares the switching capacity of optical switches and IP routers. We use the Cisco 12000 GSR series to represent typical router specifications in 2006, compared to OXCs from Lucent, Tellium, Nortel, *et al.* in the same time period. It is clear from the figure the optical capacity is several order of magnitude larger than the electronic processing rate, implying that the routers are more likely to be the bottleneck in modern networks. Unfortunately, existing

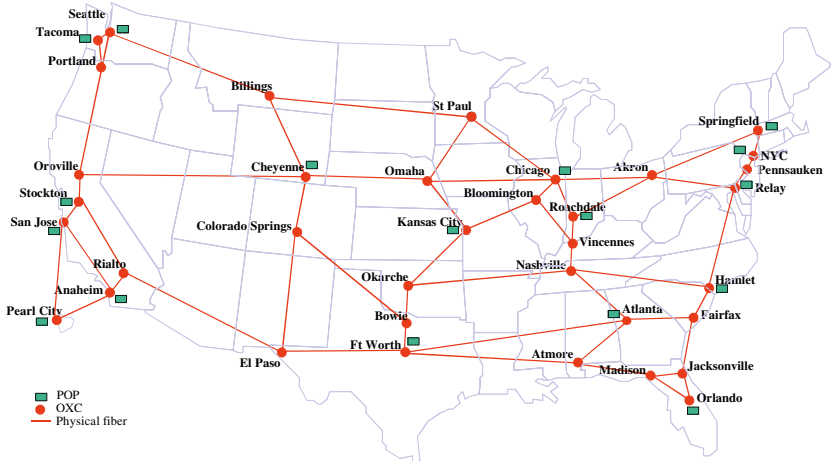


Figure 4.3: Physical topology for Sprint network [44]. Each edge in the graph represents optical fibers. The PoP nodes are marked with green box.

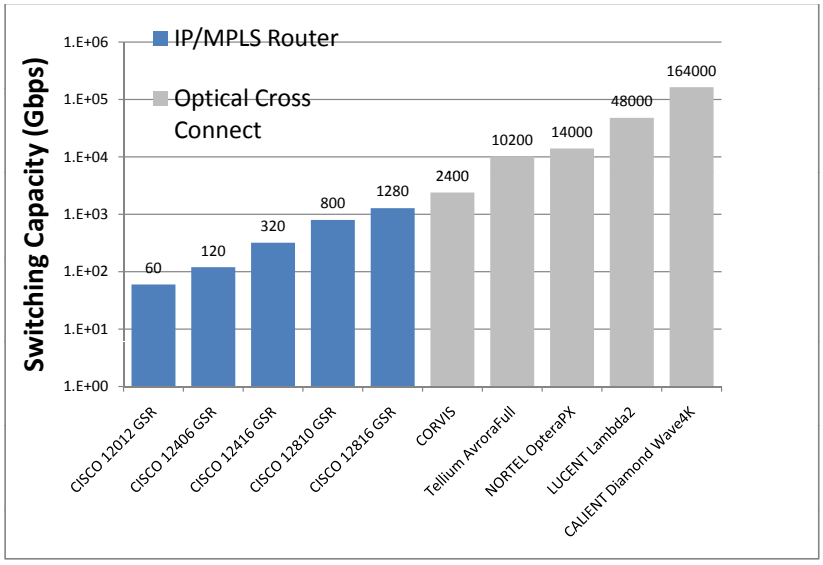


Figure 4.4: Comparison of OXCs and routers in term of switching capacity. Information in this figure is collected from [41] and manufacturer web sites based on data from 2006.

Table 4.1: The bottlenecks of NLR under different technologies

Technologies			Congested Components	
w	c (Gbps)	c_r (Gbps)	Links	Nodes
8	2.5	40	DEV-CHI, LOS-HOU HOU-ATL, CHI-NYC	n/a
8	10	80	DEV-CHI, LOS-HOU HOU-ATL	HOU, CHI
64	10	320	n/a	DEV, HOU, CHI

topology models rarely take router utilization into consideration.

Also, traffic demands are not uniform across the network. PoPs located at major cities with large population can exchange traffic at a much higher rate than others [35]. There are always some nodes and/or links that carry excessive load, creating bottlenecks. The traffic demands, together with the layout of the physical topology, affect the bottleneck location and intensity. For instance, a "hot spot" NYC in Figure 4.3 is expected to employ mostly direct connections to other PoPs in order to avoid congestion at intermediate nodes, while establishing too many end-to-end lightpaths may exhaust the scarce wavelengths and result in new bottlenecks at the output links. Another big city, Chicago, with five output links will not have such problem though. A realistic virtual topology model thus needs to take into consideration the traffic distribution over physical infrastructure.

Moreover, the design of a WDM virtual topology is affected by technology constraints. The choice of the supporting technology is an investment decision balancing the cost and the expected revenue based on traffic volume. While major ISPs can deploy costly switches with large port count to accommodate millions of users, a small ISP with limited number of subscribers will rely on less expensive, hence less capable equipment. We next use an example on NLR to show how easily the network bottleneck will shift as different technologies are deployed. As listed in Table 4.1, we target three specifications: the number of wavelengths per link w; wavelength capacity c and router capacity c_r. The resulting most congested components for each technology are computed and shown in the table (For simplicity, the traffic demand is assumed to be equal for all node pairs and a shortest-path routing is used). In Table 4.1, old and advanced technologies create different network bottlenecks. Previous literature that evaluated their models only on few Tier-1 ISP networks can hardly be effective for others.

Table 4.2: Notations

Notation	Definition
$G_p = (V_p, E_p)$	Physical topology with the set of nodes V_p and the set of fiber links E_p
$G_v = (V_v, E_v)$	Virtual topology with the set of PoP nodes V_v and the set of lightpaths E_v
$\Lambda = \{\lambda^{sd}\}$	Long-term traffic demand between nodes s and d
ρ^L/ρ^N	Maximum lightpath/node utilization
c_r	Electronic switching capacity at a node
c	Capacity of each wavelength
w	Number of wavelengths per fiber
D	Average channel length
H	Average packet hop distance
T	Network throughput

4.2 A Bottleneck-Oriented Design

4.2.1 Problem Definition and Assumptions

Refer to Table 4.2 for the terminology used. Network throughput, T, is the sum of the accommodated traffic between all node pairs. With an over-provisioned network, T is also known as the maximum scale-up factor of given traffic demands. The WDM virtual topology design problem aims at identifying a virtual topology E_v that maximizes the throughput T, given physical topology G_p, PoP node placement V_v, traffic matrix Λ, and technology constraints w, c, and c_r.

We make the following assumptions: 1) Each physical link is bi-directional and composed of a single fiber supporting w wavelengths. 2) The wavelength-continuity constraint is not considered since a design with no or limited converting capability has been addressed in [52]. 3) The number of transponders at a node matches the number of available wavelengths. Otherwise, it will be a network misconfiguration. 4) Traffic flows go through the smallest possible number of lightpaths, i.e., shortest-path-first (SPF) routing protocol. SPF algorithm may not be optimal for all cases, but it holds well for *static* network design problems [29].

4.2.2 Key Idea

In order to maximize the throughput, WDM virtual topology design needs to minimize the utilization of bottleneck elements, whether routers or links. The utilization of a lightpath is defined as the percentage of the wavelength capacity that is used by traffic crossing the lightpath. The utilization of a node is defined as the percentage of the router capacity that is used by

traffic crossing the router. When routers are the bottlenecks, it is desirable to *bypass* electronic processing at routers by creating end-to-end lightpaths. On the other hand, when wavelength capacity is the bottleneck, it is desirable to deploy traffic grooming techniques, thus not bypassing electronic processing, in order to: 1) distribute traffic evenly among parallel lightpaths, and/or 2) spread the load over alternative link paths.

Let ρ^L denote the *maximum lightpath utilization* in the network, and ρ^N denote the *maximum node utilization*. A virtual topology which mostly relies on end-to-end lightpaths (referred to as a *transparent* topology) is thus mostly trying to avoid bottlenecks at routers, and a topology which mostly exploits electronic processing at routers through hop-by-hop lightpaths (referred to as an *opaque* topology) is trying to avoid bottlenecks at links. An optimal topology design should balance the use of these two schemes to minimize $\rho = \max\{\rho^L, \rho^N\}$.

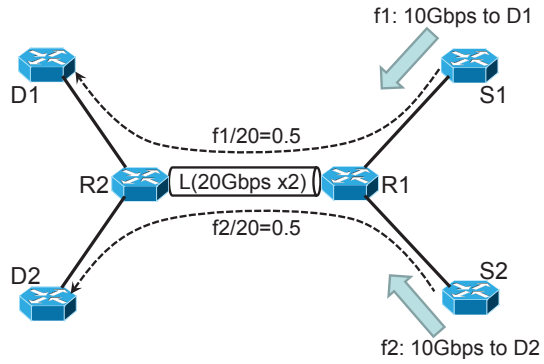
Refer to Figure 4.5 for an example. In this example, two sources $S1$ and $S2$ send two flows with rates $f1$ and $f2$ towards destinations $D1$ and $D2$. The flows cross routers $R1$ and $R2$ and the link L connecting them. In Figure 4.5(a), link L has two 20 Gbps wavelengths and the routers, $R1$ and $R2$, each has 20 Gbps capacity, i.e., w=2, c=20 and c_r=20. If flows $f1$ and $f2$ (10 Gbps each in this case) are processed at $R1$, they will exhaust the router capacity and create node congestion as $\rho^N = (f1 + f2)/20 = 100\%$. Thus, it is desirable to bypass the routers in Figure 4.5(a) with a utilization $\rho = \rho^L = \max\{f1/20, f2/20\} = 50\%$. In Figure 4.5 (b), we reduce the wavelength capacity to 10 Gbps (i.e., w=2, c=10 and c_r=20) and let $f1$ and $f2$ be 8 Gbps and 2 Gbps, respectively. Bypassing routers in this case leads to $\rho = \rho^L = \max\{f1/10, f2/10\} = 80\%$, while applying traffic grooming at R_1, as shown in Figure 4.5(b), provides a lower $\rho = \max\{(f1 + f2)/2/10, (f1 + f2)/20\} = 50\%$. Finally, in Figure 4.5(c), end-to-end lightpaths lead to a $\rho = \rho^L = 80\%$, and hop-by-hop lightpaths lead to $\rho = \rho^N = 100\%$. Neither design is optimal. A layout combining both, as shown in the figure, leads to the best $\rho = 50\%$.

In Table 4.2, the *average channel length*, D, represents the average number of physical links traversed by a lightpath. The *average packet hop distance*, H, represents the average number of lightpaths traversed by a traffic flow. The *average link hop distance*, which is the product of H and D, represents the average number of links traversed by a flow. In following, we derive a lower bound on throughput T in terms of H and D.

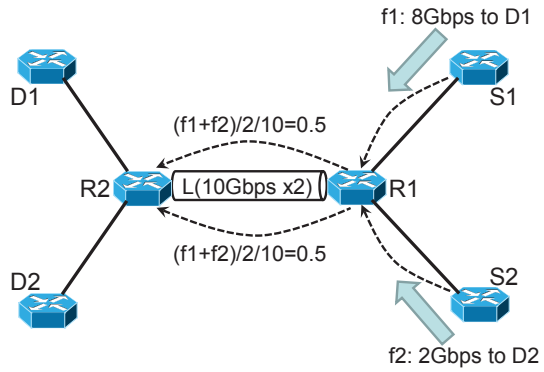
Given the virtual graph G_v and a specific routing scheme, D and H can be calculated as:

$$D = \frac{w \cdot |E_p|}{|E_v|} \tag{4.1}$$

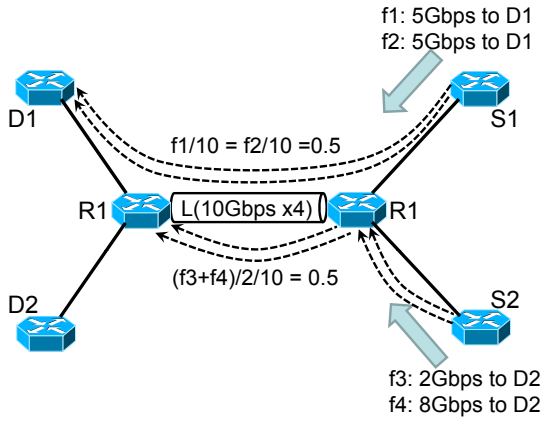
$$H = \frac{1}{\sum_{s,d} \lambda^{sd}} \sum_{s,d} \lambda^{sd} \sum_{(ij)} p_{ij}^{sd} \tag{4.2}$$



(a)



(b)



(c)

Figure 4.5: An illustration of preferred lightpath layout. (a) Router capacity is the bottleneck; (b) Wavelength capacity is the bottleneck; (c) Both router and wavelength capacity are bottlenecks.

where $p_{ij}^{sd} = 1$ if λ^{sd} employs lightpath (i, j) as an intermediate channel, and $p_{ij}^{sd} = 0$ otherwise. $|E_p|$ (or $|E_v|$) denotes the size of the set E_p (or E_v). Let λ_{ij} be the aggregate traffic demand on lightpath (i, j). λ_{ij} can be expressed as

$$\lambda_{ij} = \sum_{s,d} \lambda^{sd} \cdot p_{ij}^{sd} \tag{4.3}$$

Considering the fact that the maximum value is always greater or equal to the average, the lower bounds of ρ^L and ρ^N are given by:

$$\rho^L \geq \frac{1}{|E_v| \cdot c} \sum_{(ij)} \lambda_{ij} = \frac{H \cdot D}{w \cdot c \cdot |E_p|} \sum_{s,d} \lambda^{sd} \tag{4.4}$$

$$\rho^N \geq \frac{1}{|V_v| \cdot c_r} (\sum_i \sum_j \lambda_{ij} - \sum_{s,d} \lambda^{sd}) = \frac{H-1}{|V_v| \cdot c_r} \sum_{s,d} \lambda^{sd} \tag{4.5}$$

Since $1/\rho^L$ and $1/\rho^N$ indicate the maximum *scale-up* factors on links and nodes, respectively, the throughput T is decided by the one with less value:

$$T = min\{\frac{1}{\rho^L}, \frac{1}{\rho^N}\} \cdot \sum_{s,d} \lambda^{sd} \tag{4.6}$$

Combining (4.4), (4.5), and (4.6), we end up with

$$T \leq min\{\frac{|E_v| \cdot c}{\sum_{(ij)} \lambda_{ij}}, \frac{|V_v| \cdot c_r}{\sum_{(ij)} \lambda_{ij} - \sum_{s,d} \lambda^{sd}}\} \tag{4.7}$$

or

$$T \leq min\{\frac{w \cdot c \cdot |E_p|}{H \cdot D}, \frac{|V_v| \cdot c_r}{H-1}\} \tag{4.8}$$

In Equation (4.8), when $\rho^L << \rho^N$ (i.e., routers are the bottlenecks), the upper bound of T is inversely proportional to H, and when $\rho^L >> \rho^N$ (i.e., links are the bottlenecks), the upper bound of T is inversely proportional to $H \cdot D$. This observation implies that none of the existing design objectives, minimizing the packet hop distance, minimizing the channel length, or minimizing the link distance, can work well by itself in all scenarios as it is not always the case that nodes are the bottlenecks or that links are bottlenecks. A general solution requires a careful balance among different bottlenecks. During our survey on several ISPs, tier-1 providers such as Level3 [5] and Sprint are pushing their next-generation DWDM networks towards a transparent design paradigm to avoid router bottlenecks, while regional backbone networks such as NCREN [8] and NEREN [9] mostly implement an opaque topology to reduce link congestion.

Algorithm 1 The BOA Algorithm

1: **INPUT**: Virtual topology G_v (a 2-connected graph)
2: **repeat**
3: Implement SP traffic routing on G_v and compute:
4: $T \leftarrow$ the throughput
5: $k \leftarrow$ the node with maximum utilization ρ^N
6: $(i, j) \leftarrow$ the lightpath with maximum utilization ρ^L
7: **if** $\rho^N \geq \rho^L$ **then**
8: $S \leftarrow$ the list of traffic flows being routed at k
9: **while** $S \neq \emptyset$ **do**
10: $\lambda^{sd} \leftarrow$ pop the largest flow of S
11: **try** to create a new lightpath (s, d). If fails,**do**
12: $c \leftarrow$ connectivity of $G_v - \{\forall(u, v)|p_{uv}^{sd} = 1\}$♠
13: **if** $c > 1$ **then**
14: Replace $\{\forall(u, v)|p_{uv}^{sd} = 1\}$ by a channel (s, d)
15: **end if**
16: **end while**
17: **else**
18: **try** to add one more parallel (i, j). If fails, **do**
19: $(p, q) \leftarrow$ search for the qualified channel♣
20: **if** (p, q) exists **then**
21: Divide (p, q) into segments containing (i, j)
22: **else**
23: Divide (i, j) into pt-to-pt channels
24: **end if**
25: **end if**
26: Recompute the new scale-up factor T'
27: **until** $T' < T$
28: **OUTPUT**: new virtual topology $G_v^{'}$ (a 2-connected graph)

4.2.3 Algorithmic Details

We provide an iterative design algorithm, called bottleneck-oriented approach (BOA), that does not make any assumption about the location of network bottlenecks. In each iteration, ρ^L and ρ^N are computed and maintained. We identify the most congested element and set up the lightpaths accordingly. Refer to Algorithm 1. If the bottleneck is a router, lines 8-11 establish a new end-to-end lightpath to carry the largest flow routing through this router. In case of insufficient wavelengths, line 14 removes the series of intermediate channels currently used by λ_{sd} and replaces them by a direct connection (s, d), with subject to network connectivity constraints shown in lines 12-13. Otherwise, if the bottleneck is a channel, say (i, j), a parallel channel will first be attempted so that the load of (i, j) can be split with the new one (see line

18). In case the creation of the new parallel channel fails due to the shortage of wavelengths, we search for one existing "longer" channel that can be fragmented into (i, j) and other segments (see lines 19-21). As marked by ♣, a qualified channel must traverse all links traversed by (i, j) and we select the one which carries the least traffic load. The initial input G_v can be either a real virtual graph in use, or a baseline configuration such as MMT [41] and MST [19].

The convergence of Algorithm 1 to a lower T is guaranteed since $T' < T$ is checked after each iteration. Furthermore, the algorithm cannot lead to a disconnected graph since we always start with a *2-connected* graph (line 1) and enforce *2-connectivity* in each iteration (as marked by ♠ in Algorithm 1). 2-connectivity is a common property of core networks to ensure resilience [35]. Each iteration of BOA takes $O(|V_v|^3 lg|V_v|)$ of time and $O(|V_v|^2)$ of space in a worst case. The total number of iterations performed depends on input graph G_v and is bounded by $O(|V_v|^2)$.

4.3 Performance Evaluation

Our experiments are based on two real networks: NLR (Figure 4.1) and Sprint (Figure 4.3), and two fictitious networks: *NLR-inspired* and *Sprint-inspired*. According to the data from 2006, NLR carries few hundred terabytes of daily traffic on an eight-wavelength OC-96 WDM infrastructure utilizing Cisco CRS-1 routers [7]. We create a NLR-inspired network by adding three *extra* links into NLR, (SEA-CHI), (LOS-DC), and (HOU-NYC), reflecting NLR's next-generation backbone infrastructure proposal in 2009. Sprint is a petabyte scale Tier-1 ISP. It uses eight-slot Cisco CRS-1 routers and sixteen-wavelength DWDM fibers, with each wavelength 10 Gbps in capacity [10]. To represent emerging DWDM technologies, we create a Sprint-inspired network which has the number of wavelengths on each link upgraded to 64. The detailed network configurations are summarized in Table 4.3.

Table 4.3: ISP Network Configurations

Network	Λ (Tbps)	w	c (Gbps)	c_r (Gbps)
NLR	2	8	5	160
NLR-inspired	2	8	5	160
Sprint	10	16	10	320
Sprint-inspired	50	64	10	320

Since ISP traffic demand is rarely published, we consider two traffic models:

- **Uniform Traffic**: All nodes in the physical graph are considered as PoP nodes, i.e.,

Table 4.4: Network Bottleneck and Design Strategy

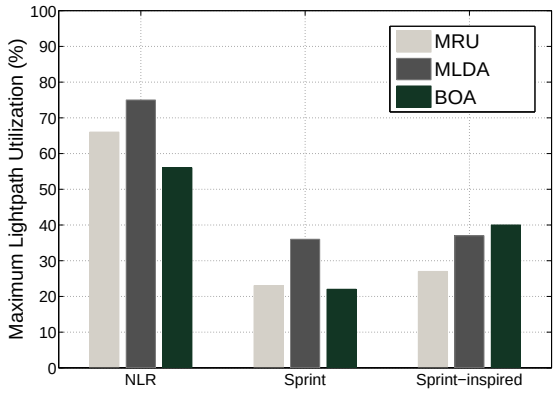
Network	Traffic Pattern	Desired Design	Change of Metric ρ^L (%)	ρ^N (%)
NLR	uniform	opaque	$82 \rightarrow 68$	$46 \rightarrow 56$
	real	opaque	$68 \rightarrow 55$	$42 \rightarrow 50$
NLR-inspired	uniform	opaque	$70 \rightarrow 62$	$44 \rightarrow 52$
	real	combination	$50 \rightarrow 44$	$40 \rightarrow 45$
Sprint	uniform	combination	$32 \rightarrow 27$	$24 \rightarrow 26$
	real	combination	$26 \rightarrow 22$	$20 \rightarrow 23$
Sprint-inspired	uniform	transparent	$28 \rightarrow 40$	$70 \rightarrow 45$
	real	transparent	$27 \rightarrow 40$	$68 \rightarrow 44$

$V_p = V_v$, and all node pairs require the same bandwidth. This scenario ignores intentionally the distinctness of node structure and exhibits a significant degree of traffic uniformity since no hot-spot is present.

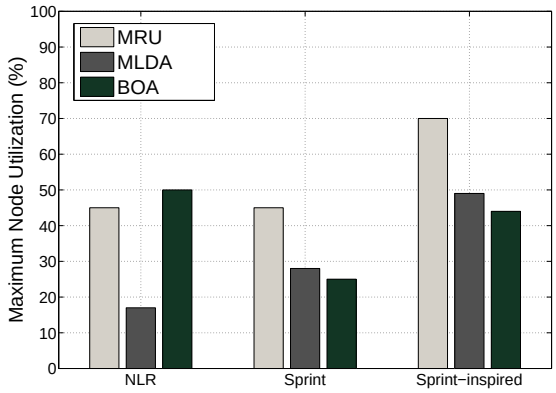
- **Real-World Traffic**: The required bandwidth between two nodes is proportional to the product of their populations (refer to traffic model in Chapter 3), and traffic has to be aggregated at a PoP in the vicinity before being transmitted. This scenario reflects the demand irregularity and hierarchical structure of Internet backbone networks.

Each simulation starts with a simple MMT configuration (see Chapter 2) and then BOA is applied till all traffic demands are accommodated. Depending on the setup, the resulting virtual topology could be transparent, opaque or a combination thereof. The results are listed in Table 4.4. In all cases, BOA balances ρ^L and ρ^N thus reduces the bottleneck utilization. Since NLR has initial link congestion much larger than its node congestion (i.e., $\rho^L = 82\%$ and $\rho^N = 46\%$ for uniform traffic, and $\rho^L = 68\%$ and $\rho^N = 42\%$ for real-world traffic), the bottlencks are on the links. The resulted NLR virtual graph is laid out with all point-to-point connections. Surprisingly, even with an opaque graph, NLR still suffers from link congestion more than from the nodes (i.e., $68\% > 56\%$ for uniform traffic, and $55\% > 50\%$ for real-world traffic). This observation suggests that as traffic volume increases, NLR is likely to set up more parallel channels along the congested links to treat the bottleneck. Also, one observes that by adding three extra links, the bottleneck utilization in the NLR-inspired network decreases significantly. Particularly, $\rho = \rho^N = 45\%$ in NLR-inspired network compared to $\rho = \rho^L = 55\%$ in NLR network. The optimized NLR-inspired network has both transparent and opaque layouts.

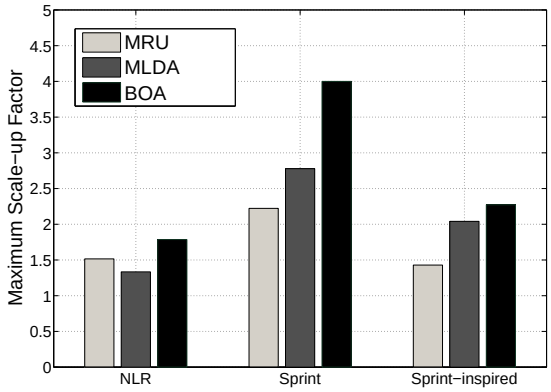
The Sprint network yields comparable link and node congestions in the initial configuration (i.e., $\rho^L = 32\%$ and $\rho^N = 24\%$ for uniform traffic, and $\rho^L = 26\%$ and $\rho^N = 20\%$ for real-world



(a)



(b)



(c)

Figure 4.6: Comparison of BOA with two other design approaches in terms of (a) the maximum lightpath utilization, (b) the maximum node utilization, (c) the network throughput.

traffic). The resulting virtual graph has both point-to-point and end-to-end connections and matches well with the published Sprint map. For the Sprint-inspired network, it is obvious that a significant shortage of switching capability at nodes (i.e., $\rho^N = 70\%$ for uniform traffic, and $\rho^N = 68\%$ for real-world traffic) pushes the resulting graph towards a transparent layout. Our results imply that the Sprint network is likely to evolve towards a transparent virtual graph in order to maximize the throughput.

We next compare the performance of BOA to two other popular heuristics MLDA and MRU (the discussions of MLDA and MRU can be found in Chapter 2) in terms of the maximum lightpath utilization, maximum node utilizations, and network throughput. As shown in Figure 4.6, BOA outperforms the other heuristics, leading to better network throughput. It achieves that by balancing the maximum link utilization and the maximum node utilization. These results reinforce our arguments that none of the existing heuristics performs best in all setups. BOA on the other hand can pinpoint the bottleneck elements, which is crucial to maximize the network throughput.

4.4 Summary

In this chapter we investigate the existing WDM virtual topology models and make a case that none of them fits all ISP networks. This is mainly due to the fact that different networks have different bottleneck elements as they vary in traffic demands and the technology deployed. We further propose an adaptive design model which iteratively identifies the network bottlenecks and sets up suitable lightpaths accordingly. The efficacy of our model is tested on published and synthetic ISP networks. Simulation results highlight the superiority over existing designs. Our approach targets a static design problem and may not be applicable to dynamic switching technologies such as optical burst switching (OBS) and optical packet switching (OPS).

Chapter 5

Towards a Robust and Green Internet Backbone Network

Previous virtual topology design models mainly aim at maximizing the network throughput by *spreading* out the load evenly among network fabrics. That is reasonable because the throughput is limited by the bottleneck elements, whether at routers or links. Recently, however, *energy* concerns are highlighted as a sizable fraction of total electricity supply in U.S. was used by network equipment, and a significant part was devoured by Internet backbone infrastructure. A green Internet with as little as 1% lower power can save more than ten billion dollars per year. Most recent design models thus aim at minimizing energy consumption by *aggregating* traffic along fewer routes while allowing devices on other routes to sleep [25].

Hence, there is an obvious trade-off. On one side, backbone networks are typically over provisioned with bandwidth redundancy. A large part of electricity bill and heat dissipation costs are wasted by under-utilized resources with balanced traffic load. On the other side, Internet traffic is highly fluctuant, containing spikes that ramp up quickly on any links and/or nodes [57]. Concentrating the load on few active routes to save energy may cause the network to become vulnerable to sudden spikes, resulting in severe congestion. A virtual topology design that can handle both network *robustness* and energy conservation is desirable for backbone networks. Unfortunately, existing design models explore only one factor and not the other.

In this chapter, we investigate network power and congestion models, and introduce a Linear Programming (LP) formulation to minimize energy consumption subject to traffic congestion constraints. The optimal solution is shown to be NP-hard. To make the solution feasible for real-sized networks, we propose a heuristic by decomposing the problem into two more tractable subproblems: 1) bounded congestion level for traffic fluctuations, and 2) optimal power usage for common traffic demand. The efficacy of our design is tested on two published backbone networks: NLR and NSFNET. The simulation results reveal that the proposed heuristic leads

to energy savings and resource utilization closely matching the optimal solution but only with polynomial time complexity.

The rest of this chapter is organized as follows. In Sections 5.1 and 5.2, we discuss the network power and congestion models. In Sections 5.3 and 5.4, we show the key idea and propose our design model. In Section 5.5, we present the results on two published ISP networks. We finally summarize in Section 5.6.

5.1 Network Architecture and Power Consumption Model

We target a backbone network running IP-over-WDM. As shown in Figure 5.1, a typical backbone structure consists of nodes interconnected by WDM fiber links. Each node is equipped with both *optical cross-connect (OXC)* and *core router*. At WDM-layer, the OXCs switch lightpaths transparently from input links to output links, and at IP-layer, core routers route data packets (when converted into electronic domain) over the lightpaths. Associated with a core router are several *access routers* that aggregate low-rate flows from local areas. Other devices essential to a WDM system are as follows: 1) A pair of *transponders* (labeled as T/R), i.e., one transmitter and one receiver, are needed at the endpoints of each lightpath for data transmission and E-O/O-E conversions. 2) A pair of optical *multiplexer* and *demultiplexer* (labeled as *De/Mux*) are deployed at fiber ends to multiplex/demultiplex different wavelengths. 3) *Erbium-doped fiber amplifiers (EDFAs)* are placed at fiber ends performing pre and post amplification, and along a fiber at certain distance intervals. In IP-over-WDM networks, virtual channels can be configured in two different ways: *bypass* and *non-bypass*. In the bypass scheme, a lightpath directly connects the source and the destination (e.g., the solid line connecting nodes 1 and 3 in Figure 5.1). Information is delivered end-to-end without in-transit processing. In the non-bypass scheme, traffic undergoes OEO conversions and IP routing at every intermediate nodes (e.g., the dashed line connecting nodes 1 and 3 in Figure 5.1).

An IP-over-WDM network is a complex engineering structure containing different network devices. We next identify the devices that consume power most.[1]

- Core routers are major contributors to total power usage. Energy is consumed for packet level processing such as memory access, scheduling, and table lookups. Modern technology clusters several components together to form one multi-chassis router whose power level increases in discrete steps depending on the number of ports (i.e., line cards) activated. For example, each working *OC-48* port in *Cisco 12008 GSR* consumes 70 watts [15]. We thus approximate router power usage, e^R, as two terms: a fixed term caused by the

[1]The energy is an observed quantity in unit of *joules*, or *kilowatt-hours* for large-scale systems. The power (in unit of *watts*) is the rate at which energy is used. Given the traffic intensity, we measure the energy consumption by aggregated power usage of all network devices.

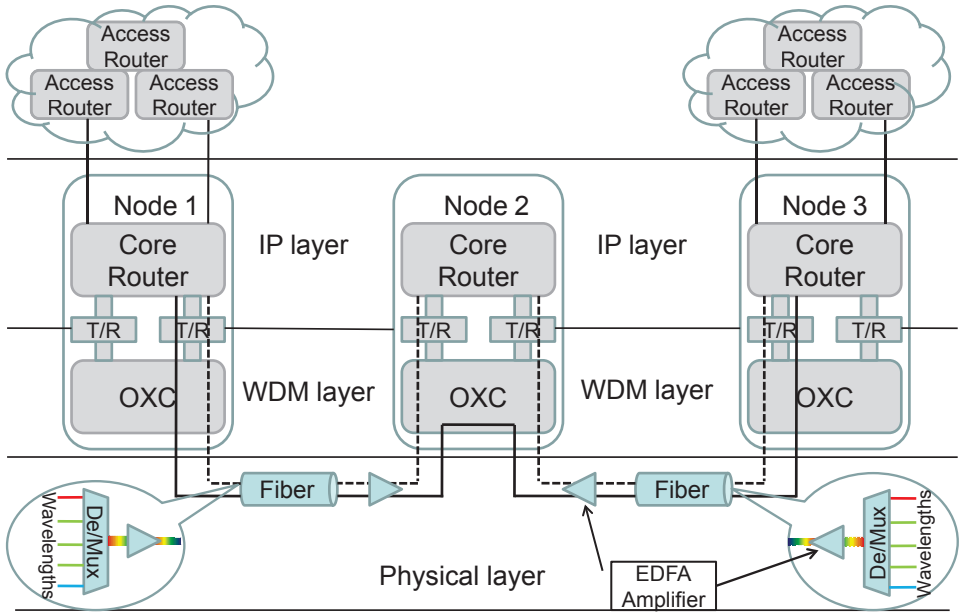


Figure 5.1: Architecture of an IP-over-WDM backbone network.

base system (i.e., chassis plus processor plus switching fabric), and a traffic-dependent term proportional to the amount of traffic passing through. It should be noted that, a router or a line card can be selectively put in standby to conserve energy if it works at lower rate [14]. The standby mode reduces the operation speed and working power to a minimized level to maintain the basic "context information" such as routing tasks. Its power consumption is assumed to be 0.

- Another primary power contributor is the transponder. According to the product data of Alcatel-Lucent WaveStar OLS [6], a pair of transponders of 10 Gbps lightpath consume $e^L = 146$ watts when working and $e^L = 0$ if the lightpath carries no traffic (i.e., standby mode).

Other devices consume minor power. For example, [48] estimates that one 8-watt EDFA is needed for every 80 km of fiber reach. EDFAs have no standby mode, hence a fixed usage for a given network. Each MEMS-based OXC consumes power in the order of 0.45 watt per connection [55], which is negligible compared to core routers. Access routers are also not considered because they are out of our scope as we focus on backbone infrastructure.

In summary, the overall network power usage, E, is expressed as:

$$E = \sum_{i \text{ is a node}} e_i^R + \sum_{(i,j) \text{ is a channel}} e_{(i,j)}^L \tag{5.1}$$

where e_i^R is the router power usage at node i and $e_{(i,j)}^L$ is the power consumed by the channel between nodes i and j.

5.2 Network Robustness to Traffic Spikes

In general, the Internet intra-domain traffic is predictable. It is not difficult for ISPs to estimate the average traffic volume with reasonable accuracy by considering customer subscription, daily peak hours, *etc.* However, estimating traffic fluctuations is difficult. To this end, representative traffic patterns are extracted based on history data and observed trends. These patterns serve as possible traffic spikes within next time window with granularity as fine as hour-to-hour [65]. If a design model does not incorporate this information, it may cause congestion at network devices, creating bottlenecks that limit a network throughput. We refer to the ability to handle traffic spikes as network *robustness*.

To investigate the robustness, we use the following model. Given traffic matrix T and a virtual topology design f, the utilization of a lightpath is the percentage of wavelength capacity that is used by traffic crossing the lightpath; the utilization of a router is the percentage of router capacity that is used by traffic crossing the router. The *maximum lightpath utilization (MLU)*, denoted by $u^L(f,T)$, is the maximum utilization of all lightpaths. The *maximum router utilization (MRU)*, denoted by $u^R(f,T)$, is the maximum utilization of all routers. MLU and MRU were widely used in previous network designs [41, 57, 66].

An optimal design f, which is most robust to traffic T, is the one that minimizes the maximum (lightpath and router) utilization $u(f,T)$. The resulting optimal utilization, $u^*(T)$, is given by

$$u(f,T) = \max\{u^L(f,T),\ u^R(f,T)\} \tag{5.2}$$

$$u^*(T) = \text{Minimize}_f\ u(f,T) \tag{5.3}$$

To compare different designs, the *performance ratio* of an arbitrary f is defined as:

$$p(f,T) = \frac{u(f,T)}{u^*(T)} \geq 1 \tag{5.4}$$

where $p(f,T)$ measures how far f is from being optimal.

Now, to account for different traffic patterns, we extend the defined metric to a set of traffic matrices $\mathbf{X}$. In particular, a design is based on an optimization minimizing $u(f,X)$ for all $X \in \mathbf{X}$, formally:

$$u(f,\mathbf{X}) = \max\{u(f,X) \mid \forall X \in \mathbf{X}\} \tag{5.5}$$

$$u^*(\mathbf{X}) = \text{Minimize}_f\ \ u(f,\mathbf{X}) \tag{5.6}$$

Table 5.1: Notations

Notation	Definition
$G = (V, E)$	physical topology with nodes V and links E
w	number of wavelengths per fiber
c^L	capacity of a wavelength
c^R	electronic switching capacity of a node
$u =$	maximum (link and node) utilization, $\max\{u^R, u^L\}$
E	total network power consumption
$T = \{t^{sd}\}$	traffic matrix for common demand
$\mathbf{X} = \{X\}$	traffic matrices for possible traffic spikes
$\Omega = \{\omega_{ij}\}$	number of parallel lightpaths between nodes i, j
$\Pi = \{\pi_{mn}^{ij}\}$	$\pi_{mn}^{ij} = 1$ if lightpath (i, j) employs link (m, n), $\pi_{mn}^{ij} = 0$ otherwise.
$\Lambda = \{\lambda_{ij}^{sd}\}$	fraction of traffic between s and d that traverses lightpath (i, j)
$\widehat{\Lambda} = \{\widehat{\lambda}_{ij}^{sd}\}$	fraction of traffic between s and d that traverses lightpath (i, j) as part of a non-shortest path
$\widehat{\Omega} = \{\widehat{\omega}_{ij}\}$	number of working lightpaths of ω_{ij}

Note that in (5.5), the aggregation of $u(f, X)$ is based on the worst-case performance. It can also be done, for example, by taking some type of weighted average. The performance ratio of an arbitrary f on $\mathbf{X}$ is:

$$p(f, \mathbf{X}) = \frac{u(f, \mathbf{X})}{u^*(\mathbf{X})} \tag{5.7}$$

Lower $p(f, \mathbf{X})$ translates to a more robust design with regard to the whole set $\mathbf{X}$.

5.3 Problem Statement

5.3.1 Terminology

Refer to Table 5.1, network physical topology is represented by graph, G, with the set of nodes, N, and the set of links, L. It is assumed that each link is bi-directed composing of a single fiber supporting w wavelengths. c^L is the wavelength capacity, and c^R is the capacity of a routing node. Traffic matrix $T = \{t^{sd}\}$ denotes the predicted common demand between node pairs. $\mathbf{X}$ is a set of matrices with each element, $X = \{x^{sd}\}$, representing one possible traffic spike scenario. Ω describes which lightpaths (in terms of end nodes) to be established and Π describes how they are routed over physical links. The routing of traffic over lightpaths is then described by Λ. A virtual topology design f depends on Ω, Π, and Λ.

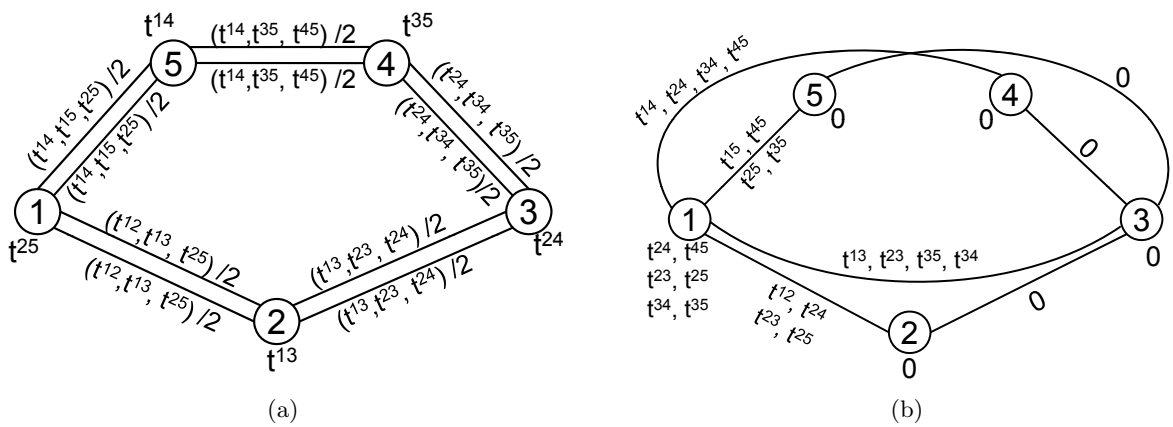


Figure 5.2: Two virtual topology designs on a five-node two-wavelength network (a) minimizing the maximum utilization; (b) minimizing the power consumption. Traffic flows that traverse each lightpath/node are exhibited in the figures.

5.3.2 Problem Formulation

While both network robustness and energy conservation are desirable, the optimization aiming at individual objectives can lead to distinctly different designs. Refer to Figure 5.2 for an example. We assume all node pairs require identical 5 Gbps bandwidth, i.e., $\forall(s,d), t^{sd} = 5$. Let the channel capacity be 20 Gbps and the capacity of each router port be 15 Gbps. Figure 5.2(a) uses purely *non-bypass* scheme with no lightpath bypassing any intermediate node. Flows are balanced across the network. The resulting lightpath utilization $u^L = \frac{3 \cdot t^{sd}}{2 \cdot 20} = 37\%$, and router utilization $u^R = \frac{t^{sd}}{15} = 33\%$. The maximum utilization is thus $u = \max\{u^R, u^L\} = 37\%$. There are totally ten lightpaths and five router ports in use in Figure 5.2(a). On the other hand, Figure 5.2(b) combines *non-bypass* and *bypass* schemes. Flows are aggregated on fewer channels with $u^L = \frac{4 \cdot t^{sd}}{20} = 100\%$. Node 1 processes 30 Gbps in-transit tafffic while all other nodes are in idle. There are totally four lightpaths and two router ports activated in this case. The maximum utilization of Figure 5.2(b) is much worse than in Figure 5.2(a), while its power consumption is 60% less.

A simple way to handle two objectives is to combine them into a single objective function, that is to optimize some function of both objectives. However, there is a significant trade-off between the size of the considered traffic set and energy optimality. In one extreme case, as the set expands to a complete space containing all possible demands, the resulting design is robust to arbitrary traffic spike but likely to produce poor energy savings on common demand. In our survey on several ASs, there are two primary policies towards a realistic design: 1) Profit-driven ISPs are not likely to compromise the guarantee of service for power reduction. Energy savings are pursued on top of the congestion. 2) Common demand lasts for most of the operation time

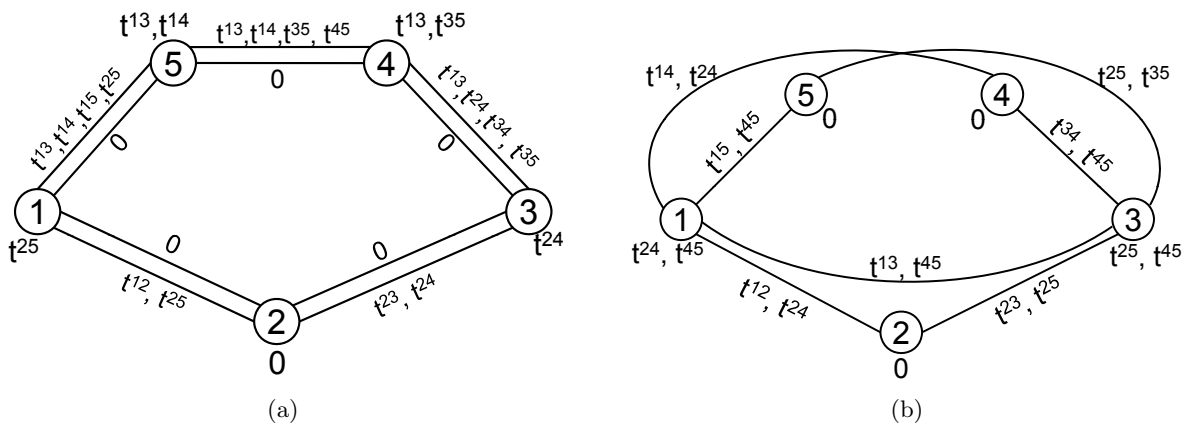


Figure 5.3: Variants of traffic routings in Figure 5.2 with the same virtual graph. (a) a variant of Figure 5.2 (a) minimizing the power consumption, (b) a variant of Figure 5.2 (b) minimizing the maximum utilization.

while traffic spikes have much shorter duration. It is best to focus energy saving on normal operation.

We thus formulate the problem by separating the energy optimization for common traffic demand T and the bounded congestion level for traffic spikes $\mathbf{X}$:

$$\boxed{\begin{array}{c} \textbf{Minimize}_f \text{ on } T: \quad E = \sum_i e_i^R + \sum_{(i,j)} e_{(i,j)}^L \\ \textbf{Subject to: } (1)\ f \text{ is a virtual topology design} \\ (2)\ u(f, \mathbf{X}) \leq 100\% \end{array}} \tag{5.8}$$

Formulation (5.8) is *reducible* to another NP-hard problem – minimizing the maximum lightpath utilization [66] – because testing $u^L \leq 100\%$ for all fs has the same rank as finding the minimum u^L. Solving this problem is numerically intractable for real-sized networks. This inherent complexity leads us to decompose the complete design into two more tractable subproblems: virtual graph layout (VGL) and traffic routing (TR) [41].

5.4 A Two-Phase Heuristic

5.4.1 Key Idea

It should be noted that energy savings and robustness are mostly conflicting objectives when solving the TR subproblem while they are mostly in agreement when solving the VGL subproblem. In particular, VGL relies on *end-to-end* lightpaths to reduce (power and bandwidth) usage at routers; VGL relies on *hop-by-hop* lightpaths to improve (power and bandwidth) efficiency at

link channels through electronic traffic grooming. So the optimality of the two factors is uniform when it comes to VGL. Meanwhile, by comparing Figure 5.3 with Figure 5.2, we observe that a distinctly different TR is deployed as we target one factor instead of the other. For example, Figure 5.3(a) concentrates the flows at one of the two parallel lightpaths on each link to save energy. Figure 5.3(b) balances the load among links/nodes to reduce the maximum utilization. The whole idea is summarized in Figure 5.4.

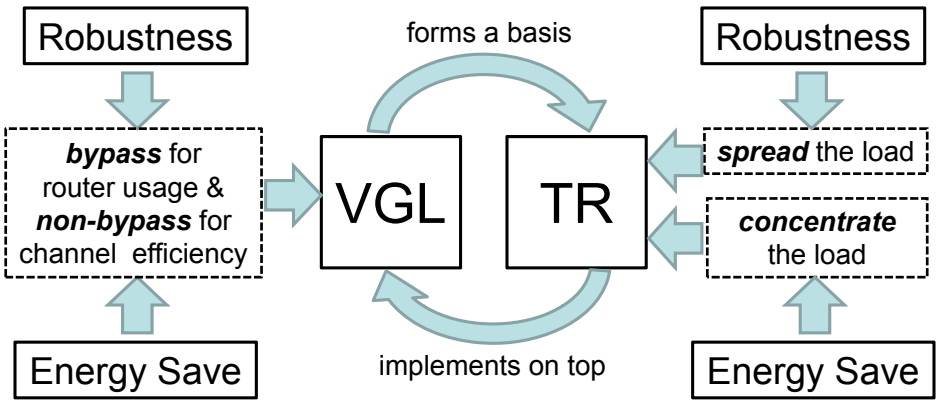


Figure 5.4: Design strategies of two factors on solving each subproblem.

5.4.2 Phase I – Virtual Graphs with Bounded Congestion

In phase I, we find out the virtual graphs with bounded worst-case MLU/MRU against traffic spikes given physical network (i.e., G, c^L, c^R, w) and traffic matrices, $\mathbf{X}$. The MLU/MRU values are computed by assuming the shortest-path-first (SPF) traffic routing. The SPF algorithm is implemented in major ISP networks [29].

Recalling Equation (5.7), a virtual topology design f is said to have *penalty envelope* ϕ if the performance ratio of f on $\mathbf{X}$ is no more than ϕ ($\phi >= 1$) [57], namely:

$$\forall X \in \mathbf{X}, \quad \frac{u(f, X)}{u^*(\mathbf{X})} \leq \phi \tag{5.9}$$

where ϕ restricts the resulted fs to those with the maximum utilization at most ϕ times of the optimal. By choosing ϕ slightly higher than 1, the virtual topologies satisfying constraint (5.9) achieve near-to-optimal congestion. We next develop formula (5.9) in terms of MLU and MRU,

respectively:

$$\forall X \in \mathbf{X}, \forall \text{ channel } (i,j), \ \sum_{s,d} \frac{x^{sd}\lambda_{ij}^{sd}}{\omega_{ij} \cdot c^L} \leq u^*(\mathbf{X}) \cdot \phi \tag{5.10}$$

$$\forall X \in \mathbf{X}, \forall \text{ node } i, \sum_{s,d:s\neq i} \sum_{j:\omega_{ij}>0} \frac{x^{sd}\lambda_{ij}^{sd}}{c^R} \leq u^*(\mathbf{X}) \cdot \phi \tag{5.11}$$

Phase I is then formulated as the following ILP problem and solved by testing if the objective is less than $u^*(\mathbf{X}) \cdot \phi$ where $u^*(\mathbf{X}) \cdot \phi \leq 1$:

Objective:

$$\max \left\{ \max\{\sum_{s,d} \frac{x^{sd}\lambda_{ij}^{sd}}{\omega_{ij} \cdot c^L}, \sum_{s,d:s\neq i} \sum_{j:\omega_{ij}>0} \frac{x^{sd}\lambda_{ij}^{sd}}{c^R}\} \mid \forall X \in \mathbf{X} \right\} \tag{5.12}$$

Variable: $\Omega = \{\omega_{ij}\}$

Subject to: $\forall X \in \mathbf{X}$,

- Traffic routing:

$$\lambda_{ij}^{sd} \in \{0,1\} \text{ applying the SPF algorithm} \tag{5.13}$$

- Total flow on a lightpath:

$$\forall \text{ channel } (i,j), \quad \sum_{s,d} x^{sd}\lambda_{ij}^{sd} \leq \omega_{ij} \cdot c^L \tag{5.14}$$

- Lightpath routing:

$$\pi_{nm}^{ij} \text{ applies } k\text{-shortest-path algorithm} \tag{5.15}$$

- Number of channels on a link:

$$\forall \text{ link } (m,n), \quad \sum_{i,j} \pi_{mn}^{ij} \leq w \tag{5.16}$$

Note that the value of $u^*(\mathbf{X})$ is not known a priori. By default, $u^*(\mathbf{X})$ is computed separately by minimizing the Formulation (5.12), but this is a poor choice considering the repeated computation on the same instances. To avoid the redundancy, we maintain a variable called *best_result* during the search, such that $u^*(\mathbf{X})$ can be obtained on the run. Constraint (5.15) finds k shortest (link) paths between each node pair, and selects one for routing the lightpath. For simplicity, we set $k = 3$. Regarding the complexity, phase I has a total of $O(|V|^2)$ variables and $O(|V|^2)$ constraints.

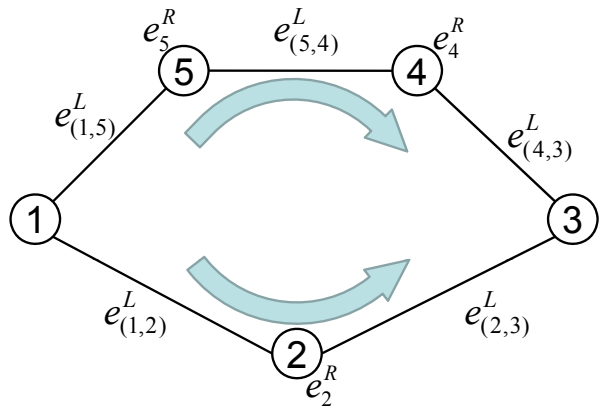


Figure 5.5: Power consumption incurred by SP and non-SP routings.

5.4.3 Phase II – Traffic Routing with Optimized Power Usage

With the penalty envelope as a safeguard, phase II searches for an energy-minimized traffic routing during the normal operation, given common demand T and candidate virtual graphs from phase I.

While a SPF routing was assumed previously, traffic flows here are routed based on the minimization of power usage. As an example shown in Figure 5.5, there are two paths between nodes 1 and 3. The shortest path 1-2-3 consumes E_1, $e_{(1,2)}^L + e_2^R + e_{(2,3)}^L$. The alternative path 1-5-4-3 consumes E_2, $e_{(1,5)}^L + e_5^R + e_{(5,4)}^L + e_4^R + e_{(4,3)}^L$. The alternative path is preferred if $E_1 \geq E_2$. This is possible because the bandwidth of a channel or router port is usually larger than a single flow. t^{13} may cost no power if it can be groomed into the already working components along 1-5-4-3, e.g., $e_{(1,5)}^L = e_4^R = 0$. We use $\widehat{\Omega}$ to indicate the number of working lightpaths in Ω (see Table 5.1).

One remaining issue is the convergence of the MLU/MRU bound. To account for the routing changes, we use $\widehat{\lambda}_{ij}^{sd}$ (see Table 5.1) to measure the load increment produced by a non-SPF routing in phase II. From Constraints (5.10) and (5.11), one observes that $\widehat{\lambda}_{ij}^{sd}$ has to be smaller than $1 - \phi \cdot u^*(\mathbf{X})$ to ensure MLU$<$ 100% and MRU$<$ 100%. Based on this observation, phase II is formulated as the following MILP problem:

Objective:

$$\sum_{s,d:s\neq i} \sum_{j:\omega_{ij}>0} \beta \cdot t^{sd} \lambda_{ij}^{sd} + \sum_{i,j} e_{(i,j)}^L \cdot \widehat{\omega_{ij}} \tag{5.17}$$

Variables: $\Lambda = \{\lambda_{ij}^{sd}\}$

Subject to:

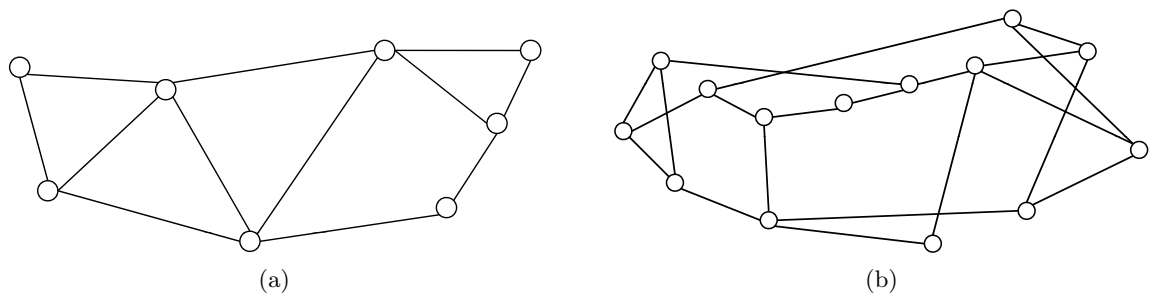


Figure 5.6: Physical topology maps. (a) NLR and (b) NSFNET.

- Flow conservation at each node:

$$\forall s, d, i \in V, \quad \sum_j \lambda_{ij}^{sd} - \sum_j \lambda_{ji}^{sd} = \begin{cases} 1 & i = s \\ -1 & i = d \\ 0 & \text{otherwise} \end{cases} \tag{5.18}$$

- Total flow on a lightpath:

$$\forall \text{ channel } (i, j), \quad \sum_{s,d} x^{sd} \lambda_{ij}^{sd} \leq \omega_{ij} \cdot c^L \tag{5.19}$$

- MLU bound conservation:

$$\forall X \in \mathbf{X}, \ \forall \text{ channel } (i, j), \ \sum_{s,d} \frac{x^{sd} \widehat{\lambda}_{ij}^{sd}}{\omega_{ij} \cdot c^L} \leq 1 - u^*(\mathbf{X}) \cdot \phi \tag{5.20}$$

- MRU bound conservation:

$$\forall X \in \mathbf{X}, \ \forall \text{ node } i, \ \sum_{s,d:s \neq i} \sum_j \frac{x^{sd} \widehat{\lambda}_{ij}^{sd}}{c^R} \leq 1 - u^*(\mathbf{X}) \cdot \phi \tag{5.21}$$

In (5.17), β is effectively the marginal router power usage per unit traffic. Phase II has a total of $O(|V|^3)$ variables and $O(|V|^3)$ constraints for a sparse mesh virtual graph.

5.5 Performance Evaluation

We compare four design models: 1) an energy-minimized design, *Energy-Min*, optimizing the objective function (5.1) only; 2) a congestion-minimized design, *MRU&MLU-Min*, optimizing the objective function (5.5) only; 3) an optimal design, *Optimal*, optimizing the formulation

(5.8); and 4) our heuristic algorithm, *Heuristic*. Model 3) functions as a reference for the best solution it can achieve through model 4). For each model, we use the CPLEX software package to solve the corresponding LP optimization on a desktop with 3.0 GHz CPU and 2G memory. The performance is tested on the two backbone networks shown in Figure 5.6: 8-node 12-link National Lambda Rail (NLR) [7] and 14-node 21-link NSFNET [2]. The physical specifications of the two networks are summarized in Table 5.2. We set $\phi = 1.2$ throughout the simulation.

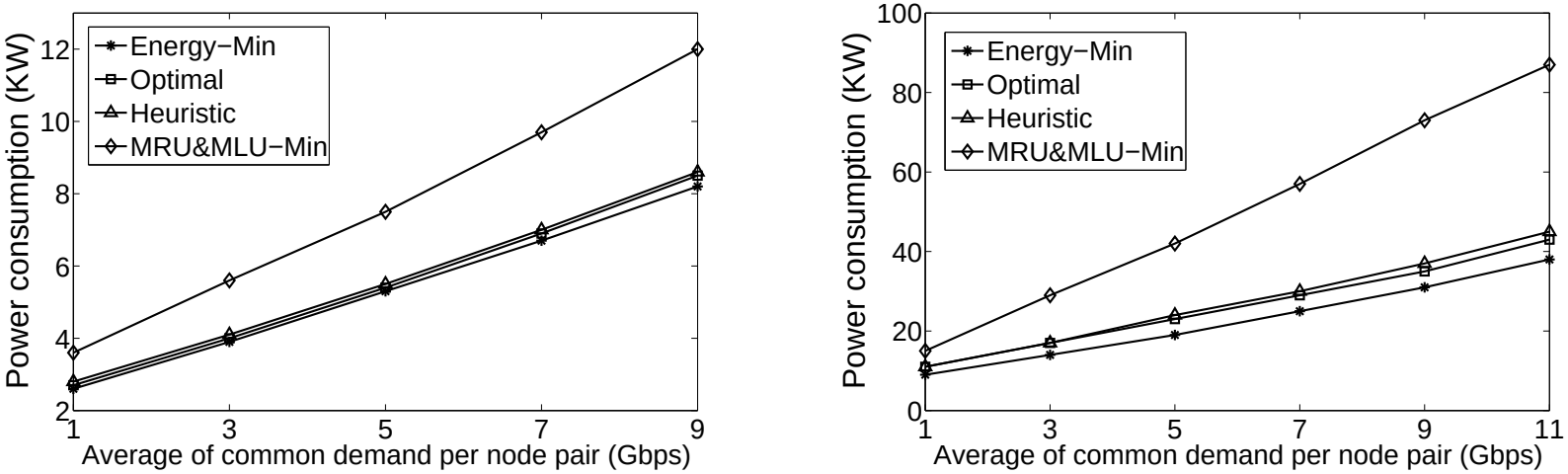
Table 5.2: Network Specifications

Network	Specifications				
	w	c^L (Gbps)	c^R (Gbps)	e^L (Watts)	e^R (Watts)
NLR	8	5	120	75	200/15Gbps port
NSFNET	8	10	160	150	250/20Gbps port

We use the traffic model described in Chapter 3 to generate the predicted common demand. In particular, the required bandwidth between two nodes is proportional to the product of their populations given that PoP nodes are mostly located at major cities. To emulate traffic spikes, we create six representative traffic matrices. In each representative matrix, 10% of all node pairs are *randomly* selected to carry three times of their normal traffic rate. An average of the simulation results from 10 runs are shown in Figure 5.7 and Figure 5.8.

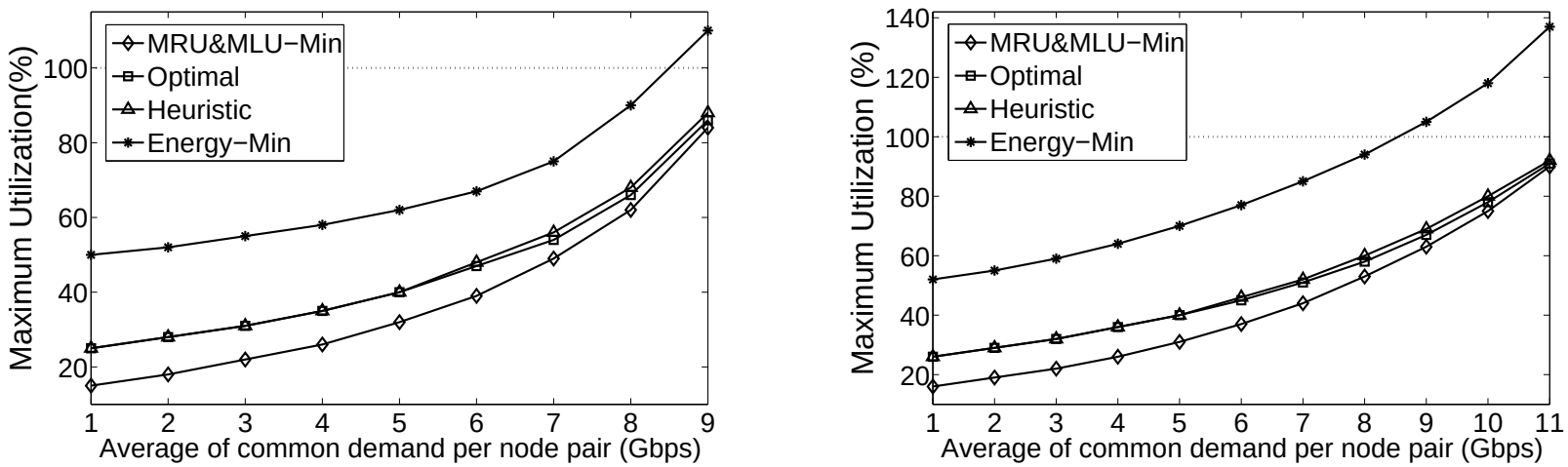
Figure 5.7 compares the resulting power consumption, E, of the design models. It is clear that *Energy-Min* consumes the least power and *MRU&MLU-Min* consumes the most. Four models keep the same power ranks in all tested networks. *Optimal* well tracks the *Energy-Min* curve with no more than 3% higher in NLR and no more than 11% higher in NSFNET. In both cases, *Optimal* is much superior to *MRU&MLU-Min* results. Also, the *Heuristic* is found to perform very closely to *Optimal*, thereby verifying the effectiveness of our heuristic algorithm on energy minimization.

Figure 5.8 compares the models in term of the maximum utilization under the occurrence of traffic spikes. As expected, *MRU&MLU-Min* yields the minimal value. *Energy-Min*, however, has much higher utilization than all other models. In general, the maximum utilization increases as the traffic rate between node pairs, showing nearly exponential growth. *Optimal* follows the trend of *MRU&MLU-Min*, and the two curves begin to merge at large traffic demands. *Heuristic* performs very closely to *Optimal*. *Energy-Min* model without considering congestion goes over the capacity limit as the traffic demand increases (e.g, $t^{sd} = 9$ Gbps in NSFNET), while *Optimal* and *Heuristic* always remain within bound.



(a) (b)

Figure 5.7: Total power consumption. (a) NLR and (b) NSFNET.



(a) (b)

Figure 5.8: The maximum utilization. (a) NLR and (b) NSFNET.

Table 5.3: Comparison of the computation time on NSFNET

Design Model	# of Variables	# of Constraints	Computation Time
Optimal	$O(\|V\|^7)$	$O(\|V\|^6)$	22 hours
Heuristic	$O(\|V\|^3)$	$O(\|V\|^3)$	37 minutes

We finally examine Figure 5.7 and Figure 5.8 together to see the overall performance of the design models. It is clear that an energy-minimized design suffers from severe link/node congestion when traffic fluctuates, and a congestion-minimized design consumes excessive energy. The proposed heuristic well matches the optimal solution. It balances the two metrics, achieving near optimal power consumption with the utilization slightly higher than the minimum. As shown in Table 5.3, NSFNET network ($|V|$=14) requires 10^8 variables and 10^7 constraints, making the exact solution very expensive. Modern technology often has a time window for network reconfiguration no more than an hour, our heuristic which reduces the computation time to less than 40 minutes compared to 22 hours for the optimal algorithm is thus well suited for real-sized networks.

5.6 Summary

In this chapter we introduce a new network design model by considering both energy savings and robustness to traffic spikes, which are *fundamental* challenges to ISP backbone networks. The proposed two-phase heuristic leads to close-to-optimal results on both two factors, while reducing the computation time to less than 40 minutes compared to 22 hours for the optimal algorithm for the simulated networks. Our model considers the power consumption of routers and links but not associated network cooling devices. It is known that heat dissipation has become a primary issue. Supplying sufficient cooling may cost several times more power than that delivered to routers. How to adapt our design to include cooling consumption is interesting for future work.

Chapter 6

Traffic Concentration for a Green Internet

Networking infrastructure accounts for an increasing fraction of global energy footprint. In 2002, about 74 Terawatt-hour (TWh) of energy, which is nearly 2% of total electricity supply in the US, was used by network equipment [30]. This percentage quickly grew to 10% in 2009 [28] and a significant part was devoured by the Internet backbone infrastructure. A green Internet with the total consumption reduced by even a small amount can save billions of dollars and cut the CO_2 emission millions of tons each year. Hence, there is a strong need for an energy-smart design of Internet backbone networks.

Research on greening the Internet requires a precise model on what one backbone network consumes, which is the aggregate consumption of all devices, mainly the *PoP routers*, across the network. One most popular design strategy aims at "aggregating traffic along fewer routes while allowing routers on the other routes to sleep" [25]. This strategy is justified by the linear router power models [15,38] with energy consumed proportional to traffic demand. In this case, traffic concentration is reasonable because the aggregate power consumption is independent of the distribution of traffic among routers.

However, existing router power models depend on the readings of an *ammeter* attached to the power cable of a router. They are mislead by ignoring the impact of external cooling equipment. It is known that heat dissipation at PoPs has fast become the primary contributor to energy consumption [51]. Manufacturers now often package up to sixty-four line cards in a single 2-foot by 2.5-foot router rack, dissipating heat at a rate of 10 KWatts or higher. To supply sufficient cooling, it costs several times more power than that delivered to routers [21]. Without considering the associated cooling consumption, existing router power models are not realistic, leading to sub-optimized designs or even network misconfiguration.

To explore the cooling power usage, there are two main hurdles: 1) the physical layout and

power specifications of PoPs are confidential and hidden by commercial ISPs; 2) lab experiments are typically limited by the maximum traffic rate of few Gbps, which is far too small to observe the whole power spectrum of a PoP router with hundreds of Gbps capacity. To overcome these difficulties, this chapter aims at a hybrid liquid-air cooling system. We analyze the heat transfer map, derive a theoretical bound for router cooling consumption, and verify our model on a real implementation on Cisco routers. Unlike previous power models, the overall energy consumed by a router is shown to be polynomial to traffic demand and increases rapidly when the router is loaded. An energy-smart network design is thus formulated to mostly spread out traffic evenly among routers across the network. We further propose a heuristic and compare our design with popular traffic concentration model on two published backbone networks: NSFNET and AT&T. The simulation results reveal that 1) our design saves at least 25% of the consumed energy, and the proposed heuristic closely matches the optimal results; 2) mitigating network bottleneck routers, rather than creating ones, leads to a greener Internet backbone network.

The rest of this chapter is organized as follows. In Section 6.1, we discuss the router cooling system and power consumption model. In Section 6.2, we show the key idea and propose our network design model. In Section 6.3, we present the results on two published ISP networks. We finally summarize in Section 6.4.

6.1 Router Cooling System and Power Consumption Model

We target a hybrid liquid-air cooling system. Although air fans have been the preferred cooling method for years due to its low cost and high reliability, emerging technology with densely packed router rack requires a more effective solution to a quickly growing heat dissipation demand. A hybrid cooling, due to its high thermal conductivity and low noise level compared to enhanced air, is the choice to meet the expectations of Cisco backbone routers [36].

As shown in Figure 6.1, a hybrid cooling system consists of external equipment including water pump, liquid reservoir, heat exchanger, and air fans, and cold plates attached to electronic computing modules (i.e., line cards) of the router. The liquid, which is driven by the pump and circulated in a sealed pipe connecting all cooling devices, transfers heat from router rack to the outside environment. In particular, a cold plate is usually made of the material of high thermal conductivity (e.g., aluminum) and has a number of evenly spaced channels across it. Low-temperature liquid flowing through the cold plates carries away the heat generated by the line cards to maintain the electronic components operating at a desirable temperature. The heated liquid is then delivered to a liquid-to-air heat exchanger where the heat is extracted and released into ambient air by means of powerful fans. The cooled liquid is stored in the reservoir ready for the next circulation.

One most critical phase in this process is the heat dissipation from the line cards to the

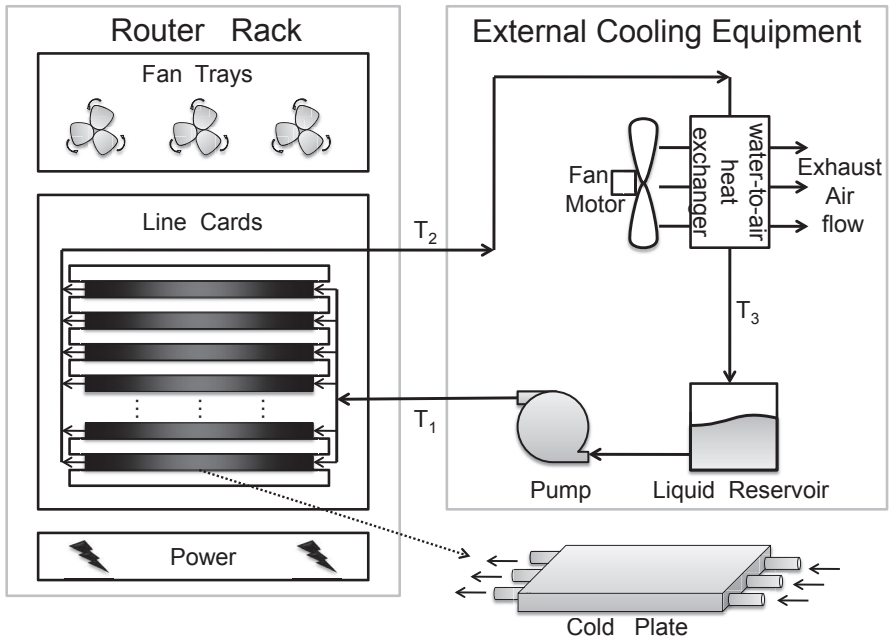


Figure 6.1: Architecture of a hybrid router cooling system.

liquid. The dissipation rate, q, follows the form:

$$q = \mathring{m} \cdot c_p \cdot (T_2 - T_1) \tag{6.1}$$

where $\mathring{m}$ is the mass flow rate of the liquid, c_p is the specific heat capacity (a constant), and T_1 and T_2 ($T_2 > T_1$) are mean inlet and outlet temperatures respectively.[1] Equation (6.1) is also expressed as:

$$q = \rho \cdot v \cdot A \cdot c_p \cdot (T_2 - T_1) \tag{6.2}$$

where ρ and v are the density and the velocity of the liquid, and A is the area of the cross-section of the main pipe. Given q, a required speed v is determined through Equation (6.2) by:

$$v = \frac{1}{\rho \cdot A \cdot c_p \cdot (T_2 - T_1)} \cdot q \tag{6.3}$$

In fluid mechanics, the kinetic energy of a pipe flow is dissipated due to friction, namely, *head loss*. The head loss, h_L, is proportional to the square of the fluid speed and calculated by

[1]Depending on the number of active line cards, the outlet temperatures of cold plates are not uniform. By default, q is computed by applying Equation (6.1) to each cold plate separately. For simplicity, we consider the array of line cards as a whole and use the mean outlet temperature to calculate the overall heat dissipation rate.

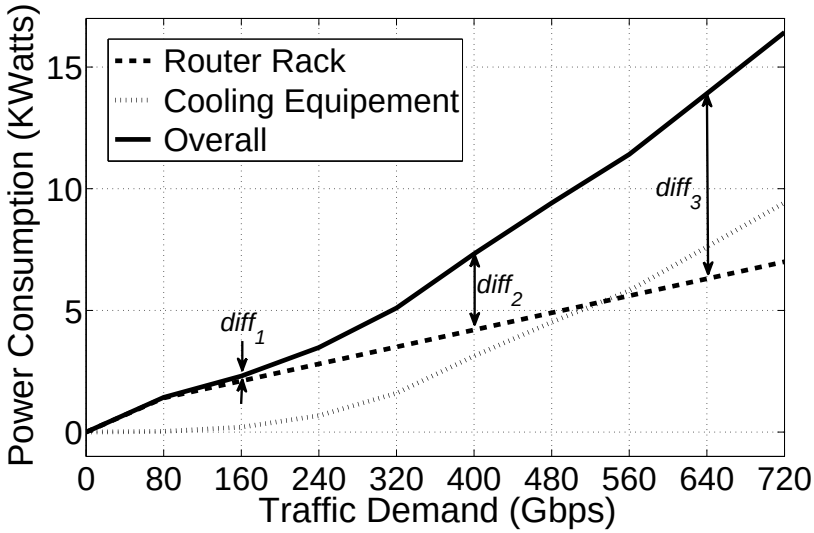


Figure 6.2: Realistic router power consumption.

Darcy-Weisbach equation [42]:

$$h_L = f \cdot \frac{L}{D} \cdot \frac{v^2}{2g} \tag{6.4}$$

where f is a coefficient called Darcy factor, g is gravity acceleration, and L and D are the length and the diameter of the pipe. To drive the liquid flowing at the speed v, a power from water pump, w_p, is needed to compensate the head loss based on energy conservation property:

$$w_p \geq \mathring{m} \cdot g \cdot h_L = (\frac{f \cdot \mathring{m} \cdot L}{2D}) \cdot v^2 \tag{6.5}$$

or, if replacing $\mathring{m}$ by $\rho \cdot v \cdot \pi D$, we have:

$$w_p \geq (\frac{\pi}{2} \cdot f \cdot \rho \cdot L) \cdot v^3 \tag{6.6}$$

It is clear from Equation (6.6) that the minimum power from water pump is proportional to the cube of the liquid speed. We insert (6.3) into (6.6), and obtain:

$$w_p \geq (\frac{\pi}{2} \cdot \frac{f \cdot L}{\rho^2 \cdot A^3 \cdot {c_p}^3 \cdot (T_2 - T_1)^3}) \cdot q^3 \tag{6.7}$$

In Equation (6.7), given the system specifications (i.e., ρ, A, c_p, f, L) and the fixed inlet and outlet temperatures (i.e., T_1 and T_2), w_p is proportional to the cube of q. Recalling that a higher q translates into a larger number of line cards in use, Equation (6.7) makes a case that the power consumption of cooling equipment is approximately proportional to the cube of the

traffic load crossing the router.

Our model presents a theoretical bound. In practice, the temperatures T_1 and T_2 may not be fixed, and the friction f also changes as the flow speed changes. Both these two factors can affect the actual power usage. To this end, we examine a real hybrid cooling project implemented on Cisco 7609s. Cisco 7609s is a 9-slot, half-rack router dissipating approximately 730W per card. The desired operating temperature is no more than 65°. The layout and specifications of all cooling devices installed are outlined in [36, 37]. We model and solve the heat transfer map in Engineering Equation Solver (EES) [1]. The results are shown in Figure 6.2.

The observations from Figure 6.2 are: 1) When a router is lightly loaded, the consumption of cooling devices is only a small fraction compared to router rack (e.g., 0.05 KW compared to 0.73 KW for traffic rate less than 80 Gbps). This is a typical case for most lab experiments. 2) The cooling consumption increases as the traffic demand with a power around 2.8. It exceeds the consumption of router rack when the traffic load approaches 70% of the capacity. In the figure, $diff_3 + diff_1 > 2 \cdot diff_2$ implies that two routers with 400 Gbps traffic on each will cost less cooling power than the case where one is carrying 160 Gbps and the other is carrying 640 Gbps. 3) The overall router power consumption, w, is approximated as three terms: a fixed term w^o caused by the base system (i.e., chassis plus processor plus switching fabric), a traffic-dependent term $w^r(\cdot)$ due to the computing modules activated, and another traffic-dependent term $w^c(\cdot)$ consumed by cooling devices, namely:

$$w = w^o + w^r(t) + w^c(t) \tag{6.8}$$

In Equation (6.8), t is traffic load; w^o equals 0 if $t = 0$ (i.e., the router is in idle); $w^r(t)$ is proportional to t; and $w^c(t)$ is a power of t.

6.2 Problem Statement and Heuristic Algorithm

6.2.1 MILP Formulation

Network topology is represented by graph, G, with the set of routers, N, and the set of links, E. The router capacity is C. Traffic matrix $\Lambda = \{\lambda^{sd}\}$ denotes the traffic demand between routers s and d, $s, d \in N$. Traffic routing $\Phi = \{\phi_{ij}^{sd}\}$ denotes the fraction of λ^{sd} that traverses the intermediate link $(i, j) \in E$. Let t_i be the aggregate in-transit traffic crossing router $i \in N$. The objective is to find an optimal Φ minimizing the aggregate power consumption of all routers while accommodating traffic demands. The problem is formulated in *mixed-integer linear programming (MILP)* as follows:

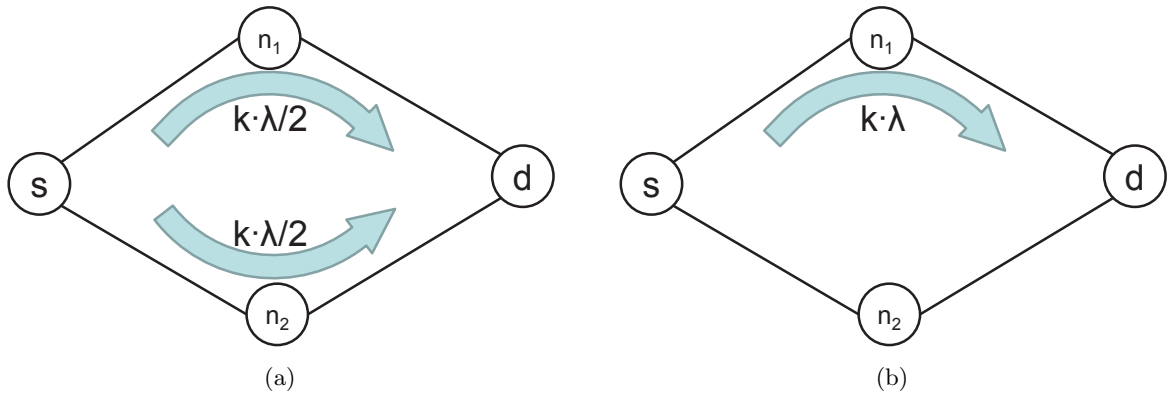


Figure 6.3: Comparison of two traffic routing scenarios: (a) traffic balancing, (b) traffic concentration.

Objective:

$$\textbf{Minimize}: \sum_{i \in N} w^0 + w^r(t_i) + w^c(t_i) \tag{6.9}$$

Variable: $\Phi = \{\phi_{ij}^{sd}\}$

Subject to:

- Flow conservation at each router:

$$\forall s, d, i \in N, \quad \sum_j \phi_{ij}^{sd} - \sum_j \phi_{ji}^{sd} = \begin{cases} 1 & i = s \\ -1 & i = d \\ 0 & \text{otherwise} \end{cases} \tag{6.10}$$

- Capacity of each router:

$$\forall\ i \in N, \quad t_i = \sum_{s,d:s \neq i} \sum_j \lambda^{sd} \phi_{ij}^{sd} \leq C \tag{6.11}$$

This formulation is *reducible* to another NP-hard problem – minimum cost multi-commodity network [67]. An optimal solution can be solved in polynomial time if fractional flows are allowed, but it is still numerically expensive for real-sized networks.

6.2.2 Key Idea

While a *shortest-path-first (SPF)* algorithm delivers traffic through a minimum number of routers to save energy, when there are alternative paths available, deciding on the traffic distribution among paths is critical for energy savings. Refer to Figure 6.3 for an example. In this

example, the router s sends $k \cdot \lambda$ of traffic towards the router d where $0 < k < \infty$. There are two paths connecting s and d. One path traverses an intermediate router n_1 with load t_1 and the other path traverses an intermediate router n_2 with load t_2.

- **Case 1**: n_1 and n_2 have equal traffic load λ, i.e., $t_1 = t_2 = \lambda$. Figure 6.3 (a) shows a scenario that $k \cdot \lambda$ is spread evenly between the two paths. Based on Equation (6.8), the aggregate power level of n_1 and n_2 is given by: (assume $w^c(t)$ in Equation (6.8) is proportional to the cube of t)

  $$2 \cdot w^o + (k+2) \cdot w^r(\lambda) + 2 \cdot (\frac{k}{2} + 1)^3 \cdot w^c(\lambda)$$

  Figure 6.3 (b), on the other hand, concentrates traffic $k \cdot \lambda$ over n_1 only. The resulting power consumption is:

  $$2 \cdot w^o + (k+2) \cdot w^r(\lambda) + ((k+1)^3 + 1) \cdot w^c(\lambda)$$

  By comparing the results of the two scenarios, traffic balancing saves more energy than traffic concentration regardless of the value of k.

- **Case 2**: n_1 has load λ and n_2 has no traffic, i.e., $t_1 = \lambda$ and $t_2 = 0$. Scenario (a) balances $k \cdot \lambda$ over the two paths such that n_1 and n_2 each carries an equal $\frac{(k+1)\cdot\lambda}{2}$. Two routers consume totally:

  $$2 \cdot w^o + (k+1) \cdot w^r(\lambda) + 2 \cdot (\frac{k+1}{2})^3 \cdot w^c(\lambda)$$

  Scenario (b) forwards all traffic through n_1, consuming:

  $$w^o + (k+1) \cdot w^r(\lambda) + (k+1)^3 \cdot w^c(\lambda)$$

  In this case, traffic balancing is superior if:

  $$k \geq \sqrt[3]{\frac{4 \cdot w^o}{3 \cdot w^c(\lambda)}} - 1 \tag{6.12}$$

  This requirement is very likely to be satisfied. For example, using the results from Figure 6.2, formula (6.12) is equivalent to $k \geq 1.65$ given $w^o = 700$ Watts and $\lambda \geq 80$ Gbps.

- **Case 3:** both n_1 and n_2 have no traffic, i.e., $t_1 = t_2 = 0$. In similar way, we find that scenario (a) is preferred if:

  $$w^c(k \cdot \lambda) \geq \frac{4}{3} \cdot w^o \tag{6.13}$$

Algorithm 2 FBAUR Algorithm

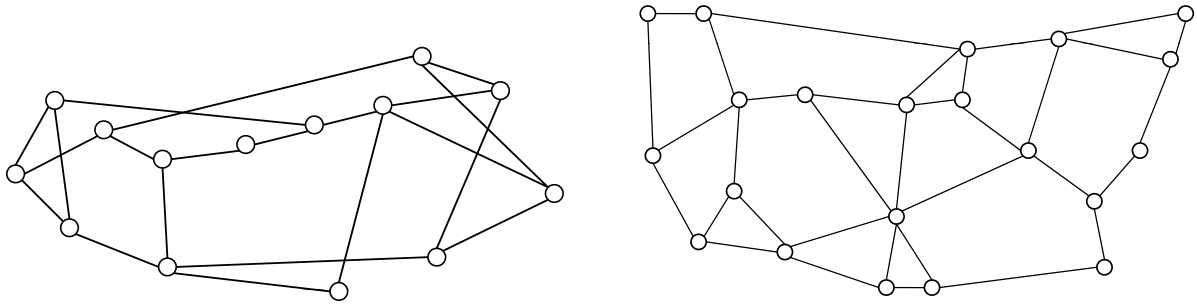
1: **INPUT**: network topology G and traffic matrix Λ
2: Compute two shortest paths (sp_1, sp_2) for each node pair.
3: Sort $\Lambda = \{\lambda^{sd}\}$ in a descending order.
4: **repeat**
5: $\quad\lambda^{sd} \leftarrow$ pop the largest flow in Λ
6: $\quad u_1 \leftarrow$ maximum router load along sp_1 of (s, d)
7: $\quad u_2 \leftarrow$ maximum router load along sp_2 of (s, d)
8: $\quad$**if** $(u_1 = 0, u_2 = 0, w^c(\lambda^{sd}) < \frac{4}{3} \cdot w^o)$ **then**
9: $\quad\quad$Forward λ^{sd} through sp_1.
10: $\quad$**else if** $(u_1 \neq 0, u_2 = 0, \frac{\lambda^{sd}}{u_1} \geq \sqrt[3]{\frac{4 \cdot w^o}{3 \cdot w^c(u_1)}} - 1)$ **then**
11: $\quad\quad$Forward λ^{sd} through sp_1.
12: $\quad$**else if** $(u_1 = 0, u_2 \neq 0, \frac{\lambda^{sd}}{u_2} \geq \sqrt[3]{\frac{4 \cdot w^o}{3 \cdot w^c(u_2)}} - 1)$ **then**
13: $\quad\quad$Forward λ^{sd} through sp_2.
14: $\quad$**else**
15: $\quad\quad$Spread λ^{sd} among sp_1 and sp_2♣.
16: $\quad$**end if**
17: $\quad$Update the traffic load of routers.
18: **until** Λ is empty
19: **OUTPUT**: traffic routing Φ minimizing energy footprint

and scenario (b) is preferred otherwise.

The example in Figure 6.3 tells us: Aggregating traffic along fewer routes is *only* suitable for low-rate flows at under-utilized routers. For all other cases, spreading out the load evenly among routers can better save energy.

6.2.3 Heuristic Algorithm

A heuristic algorithm becomes important as the size of the network gets larger. We introduce an iterative heuristic algorithm, called *flow balancing among utilized routers (FBAUR)*, adapted from [67]. FBAUR (see Algorithm 2) computes two shortest paths, sp_1 and sp_2, for each router pair using the SPF algorithm. During each iteration, we first identify the largest flow λ_{sd} that has not been routed. Line 6 measures the maximum router load u_1 along sp_1 of the pair (s, d), and line 7 measures the maximum router load u_2 along sp_2. Then, lines 8-16 decide on the distribution of λ_{sd} among two alternative paths in four cases based on the discussions in Section 6.2.2. In line 15 (marked by ♠), λ^{sd} is balanced between sp_1 and sp_2 such that the resulting maximum router loads of the two paths are as equal as possible. The complexity of Algorithm 2 is $O(|N|^2)$. While in general, FBAUR can be easily adapted to handle the case



(a) (b)

Figure 6.4: Physical topology maps. (a) NSFNET and (b) AT&T.

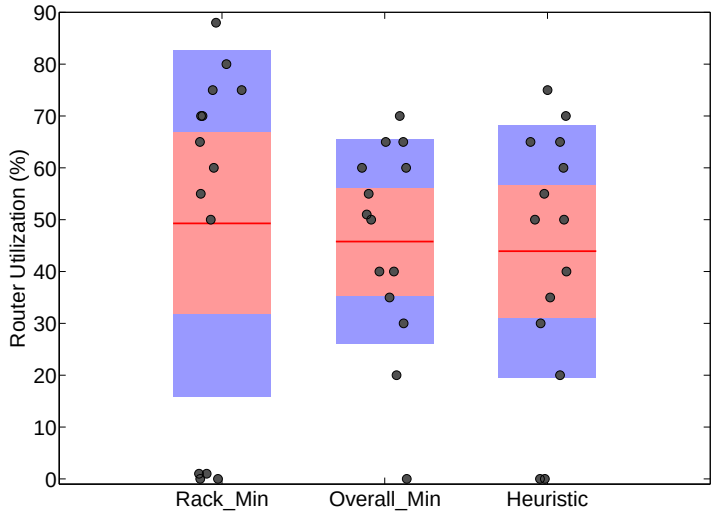
with k-shortest paths, it will increase the complexity significantly without much improvement on performance because most backbone networks are sparse meshes [35].

6.3 Performance Evaluation

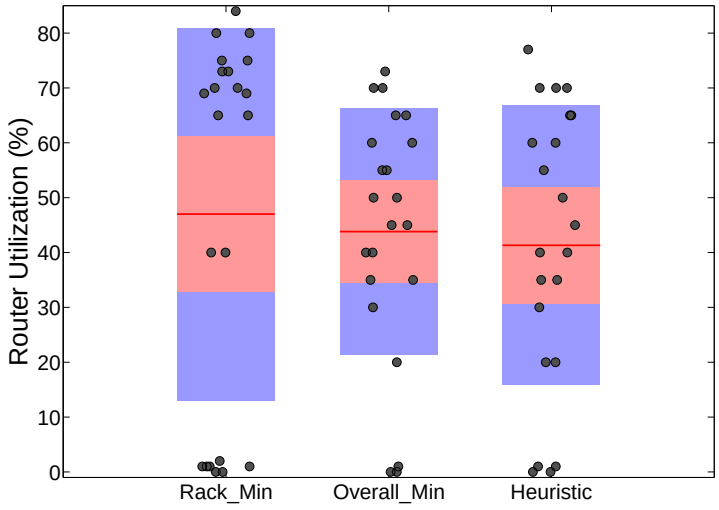
We compare three design models: 1) a traffic-concentration design [15], referred as *Rack_Min*, without considering cooling equipment; 2) an optimal design, *Overall_Min*, minimizing the overall power consumption as formulated in (6.9)-(6.11); 3) the proposed FBAUR heuristic, *Heuristic*, described in Algorithm 2. Model 1 assumes that router chassis and line cards dominate the power profile of a network. Model 2 functions as a reference for the best solution that can be achieved. For the first two models, we use the CPLEX software package to solve the corresponding MILP optimization on a desktop with 3 GHz CPU and 2G memory. The performance is tested on the two ISP networks shown in Figure 6.4: 14-node 21-link NSFNET and 21-node 34-link AT&T express backbone (US). It is assumed that each node is a PoP router with the power consumption model shown in Figure 6.2.

We use a gravity traffic model described in Chapter 3 to generate the traffic matrix. In particular, the required bandwidth between two routers is proportional to the product of their populations given that PoP routers are mostly located at major cities, and cities with larger population are expected to exchange traffic at a much higher rate. The metric, *router utilization*, denotes the percentage of router capacity that has been utilized. To reveal the distinct routings of the design models, we draw the utilization of all routers in a *boxplot* where the average utilization is shown, along with interquartile and lower/upper quartile of the average.

Figure 6.5 (a) and (b) show the simulation results on NSFNET and AT&T respectively. *Rack_Min* aggregates most traffic along fewer routers. The router utilization is thus mainly distributed at the two poles of the band. For example, in AT&T, *Rack_Min* has 13 (out of 21) "hot" routers with more than 65% utilization while there are 6 routers in idle. On the other



(a)



(b)

Figure 6.5: Traffic distribution among routers. (a) NSFNET and (b) AT&T.

hand, *Overall_Min* has the router utilization mostly distributed within a narrow band around 50% utilization, hence a balanced traffic routing among routers. *Heuristic* acheives similar pattern as in *Overall_Min* with few more routers away from the center of the band.

The resulting aggregate router power consumption is shown in Table 6.1. *Overall_Min* consumes 25% less energy than *Rack_Min* in NSFNET and 27% less energy than *Rack_Min* in

Table 6.1: Aggregate Power Consumption (KWatts)

Networks	Design Models		
	Rack_Min	Overall_Min	Heuristic
NSFNET	185	140	152
AT&T	356	260	280

AT&T. Therefore, traffic balancing outperforms traffic concentration in the tested networks. Also, *Heuristic* well tracks the *Overall_Min* results with no more than 9% higher in NSFNET and no more than 8% higher in AT&T, thereby verifying the effectiveness of our heuristic algorithm in energy minimization.

6.4 Summary

Cooling equipment accounts for a sizable fraction of overall router power consumption. Without incorporating cooling power consumption, existing traffic concentration models lead to unoptimized network designs. This chapter proposes a new MILP design formulation, along with a heuristic, reflecting a more realistic router power model. The simulation results suggest to mostly spread out the load among routers across the network. This is also a common objective in Traffic Engineering aiming at maximizing the network throughput. The two factors (energy savings and network throughput), surprisingly, are in agreement with each other, highlighting the need for thorough investigations.

Chapter 7

Conclusion

This dissertation is dedicated to investigating the WDM Internet backbone networks. A complete design consists of network planning at three layers: physical topology, virtual topology and traffic routing. At physical layer, by taking performance, resilience and geographical constraints into consideration, the HINT model emulates the sparse mesh structures of the published AT&T and Lelvel3 networks with more then 90% similarity. For virtual topology design, while the literature has been aimed at maximizing the throughput, none of them fits all ISP networks. Our BOA model abstracts the individual design strategies from a point view of network bottleneck elements, leading to optimized throughput independent of network contexts. The superiority of BOA is justified on NLR and Sprint and two other synthetic networks. Then, we consider the energy factor and show a new virtual design formulation balancing the network robustness and the power consumption. Our formulation is realistic because profit-driven ISPs usually pursue energy savings on top of the guarantee of service. The proposed two-phase heuristic reduces the computation time to less than 40 minutes compared to 22 hours for the optimal algorithm of the simulated networks. For traffic routing, we challenge the most popular load concentration strategy – *aggregating traffic along fewer routes while allowing routers on the other routes to sleep* – because the actual router power spectrum is non-linear in traffic demand. To reflect the real router power model, FBAUR mostly spreads out the traffic among routers across the network. The simulation results make a case that, by mitigating network bottleneck routers, FBAUR consumes more than 25% less energy than existing energy-saving protocols.

By comparing Chapter 4 and Chapter 6, one observes that a balanced traffic load is pursued by both designs maximizing the network throughput and minimizing the network power consumption. The optimalities of the two factors become, surprisingly, in agreement with each other. This observation challenges the common wisdom about the tradeoff between the performance and the energy, and highlights the need for more thorough investigations. The other future works are as follows. First, we will explore the cooperation of the design in different

layers and understand how they interact together towards an optimized result. Second, the efficacy of our design model depends on the estimation (with reasonable accuracy) of input traffic matrices. Given the fact that these matrices are not always available, we will consider the feasibility of adapting our model to a dynamic network context and verify the stability of the solutions.

REFERENCES

[1] Engineering Equation Solver (EES). http://www.fchart.com/ees.

[2] NSFNET network, 1998. http://www.nsfnet-legacy.org/about.php.

[3] Internet2 network, 2002. http://www.internet2.edu.

[4] County and city data book, 2007. http://www.census.gov.

[5] Level3 network, 2008. http://www.level3.com.

[6] Alcatel-Lucent WaveStar OLS 1.6T specifications, 2009. http://www.alcatel-lucent.com.

[7] National LambdaRail network, 2009. http://noc.nlr.net/nlr/network-status.html.

[8] NCREN regional network, 2009. https://www.mcnc.org.

[9] NEREN regional network, 2009. http://www.neren.net.

[10] Sprint network, 2009. http://www2.sprint.com/mr/mrhome.do.

[11] D. Alderson, L. Li, W. Willinger, and J. C. Doyle. Understanding Internet topology: principles, models, and validation. *IEEE/ACM Transactions on Networking*, pages 1205–1218, 2005.

[12] D. Bailey and E. Wright. *Practical Fiber Optics*. Butterworth-Heinemann, 2003.

[13] D. Banerjee and B. Mukherjee. Wavelength-routed optical networks: linear formulation, resource budgeting tradeoffs, and a reconfiguration study. *IEEE/ACM Transactions on Networking*, 8:598–607, 2000.

[14] R. Bolla, R. Bruschi, A. Cianfrani, and M. Listanti. Enabling backbone networks to sleep. *IEEE Network*, 25(2):26–31, 2011.

[15] J. Chabarek, J. Sommers, P. Barford, C. Estan, D. Tsiang, and S. Wright. Power awareness in network design and routing. In *Proceedings of IEEE 27th Conference on Computer Communications (INFOCOM)*, pages 457–465, 2008.

[16] L. Chiaraviglio, D. Ciullo, E. Leonardi, and M. Mellia. How much can the Internet be greened? In *Proceedings of IEEE GLOBECOM Workshops*, pages 1–6, 2009.

[17] R.C. Chu, R.E. Simons, M.J. Ellsworth, R.R. Schmidt, and V. Cozzolino. Review of cooling technologies for computer products. *IEEE Transactions on Device and Materials Reliability*, 4(4):568–585, 2004.

[18] A. Dwivedi and R. E. Wagner. Traffic model for USA long-distance optical network. In *Proceedings of Optical Fiber Communication Conference*, volume 1, pages 156–158, 2000.

[19] F.E. El-Khamy, M. Nasr, H.M.H. Shalaby, and H.T. Mouftah. The performance for heuristic algorithms for virtual topology design in all-optical WDM networks. In *Proceedings of 11th International Conference on Transparent Optical Networks (ICTON)*, pages 1–4, 2009.

[20] Eshoo, Waxman, Boucher, and Markey. Broadband conduit deployment act of 2009. The library of congress. http://thomas.loc.gov/cgi-bin/query/z?c111:H.R.2428:.

[21] Mark Fontecchio. Power Usage Effectiveness (PUE). http://searchdatacenter.techtarget.com/definition/power-usage-effectiveness-PUE.

[22] B. Fortz and M. Thorup. Internet traffic engineering by optimizing OSPF weights. In *Proceedings of IEEE 19th Annual Joint Conference of Computer and Communications Societies (INFOCOM)*, volume 2, pages 519–528, 2000.

[23] Robert G. Gallager. A minimum delay routing algorithm using distributed computation. *IEEE Transactions on Communications*, pages 73–85, 1977.

[24] C. Guan and V. Chan. Efficient physical topologies for regular WDM networks. In *Proceedings of Optical Fiber Communication Conference*, volume 1, 2004.

[25] Maruti Gupta and Suresh Singh. Greening of the Internet. In *Proceedings of ACM conference on Applications, technologies, architectures, and protocols for computer communications (SIGCOMM)*, pages 19–26, 2003.

[26] D. Habibi, H. N. Nguyen, Q. V. Phung, and Kung meng Lo. Establishing physical survivability of large networks using properties of two-connected graphs. In *Proceedings of IEEE TENCON*, pages 1–5, 2005.

[27] T. C. Hu. Optimum communication spanning trees. *SIAM Journal on Computing*, 3:188–195, 1974.

[28] Shu Huang, D. Seshadri, and R. Dutta. Traffic grooming: A changing role in green optical networks. In *Proceedings of the 28th IEEE conference on Global telecommunications (GLOBECOM)*, pages 1–6, 2009.

[29] G. Iannaccone, Chuah Chen-Nee, S. Bhattacharyya, and C. Diot. Feasibility of IP restoration in a tier 1 backbone. *IEEE Journal on Networking*, 18(2):13–19, 2004.

[30] K. Kawamoto, J. Koomey, R. Nordman B.and Brown, M. Piette, M. Ting, and A. Meier. Electricity used by office equipment and network equipment in the U.S. In *ACEEE Summer Study Conference on Energy Efficiency*, 2000.

[31] R. M. Krishnaswamy and K. N. Sivarajan. Design of logical topologies: A linear formulation for wavelength-routed optical networks with no wavelength changers. *IEEE/ACM Transactions on Networking*, 9(2), 2001.

[32] L. Li, D. Alderson, W. Willinger, and J. Doyle. A first-principles approach to understanding the Internet's router-level topology. In *Proceedings of ACM conference on Applications, technologies, architectures, and protocols for computer communications (SIGCOMM)*, pages 3–14, 2004.

[33] H. Liu and F. A. Tobagi. Physical topology design for all-optical networks. In *Proceedings of 3rd International Conference on Broadband Communications, Networks and Systems (BROADNETS)*, pages 1–10, 2006.

[34] K.H. Liu. *IP over WDM*. John Wiley & Sons, 2002.

[35] X. Ma, Kim S., and K. Harfoush. Towards realistic physical topology models for Internet backbone networks. In *Proceedings of IEEE 6th International Symposium on High-Capacity Optical Networks and Enabling Technologies (HONET)*, pages 36–42, 2009.

[36] J. Maddren and C. Pascual. Next generation router system cooling design, analysis and testing: Phase 1 study report. Technical report, California Polytechnic State University, 2005.

[37] J. Maddren and C. Pascual. Next generation router system cooling design, analysis and testing: Phase 2 study report. Technical report, California Polytechnic State University, 2005.

[38] P. Mahadevan, P Sharma, and S. Banerjee. A power benchmarking framework for network devices. In *Proceedings of IFIP Networking*, 2009.

[39] E. Modiano and A. Narula-Tam. Survivable lightpath routing: a new approach to the design of WDM-based networks. *IEEE Journal on Selected Areas in Communications*, 20:800–809, 2002.

[40] B. Mukherjee, D. Banerjee, S. Ramamurthy, and A. Mukherjee. Some principles for designing a wide-area WDM optical network. *IEEE/ACM Transactions on Networking (TON)*, 4(5):684–696, 1996.

[41] Biswanath Mukherjee. *Optical WDM networks*. Springer Science, 2006.

[42] B. R. Munson, D. F. Young, T. H. Okiishi, and W. W. Huebsch. *Fundamentals of Fluid Mechanics*. John Wiley & Sons, 6th edition, 2009.

[43] S. Nedevschi, L. Popa, G. Iannaccone, S. Ratnasamy, and D. Wetherall. Reducing network energy consumption via sleeping and rate-adaptation. In *Proceedings of the 5th USENIX Symposium on Networked Systems Design and Implementation*, pages 323–336, 2008.

[44] Antonio Nucci, Nina Taft, Patrick Thiran, Hui Zang, and Christophe Diot. Increasing the link utilization in IP over WDM networks using availability as QoS. *Photonic Network Communications*, 9:55–75, 2005.

[45] A. Proestaki and M. C. Sinclair. Design and dimensioning of dual-homing hierarchical multi-ring networks. *Proceedings of IEE Communications*, 147:96–104, 2000.

[46] R. Ramaswami and K. N. Sivarajan. Design of logical topologies for wavelength-routed optical networks. *IEEE Journal on Selected Areas in Communications*, 14:840–851, 1996.

[47] M. Roughan, A. Greenberg, C. Kalmanek, M. Rumsewicz, J. Yates, and Y. Zhang. Experience in measuring backbone traffic variability: models, metrics, measurements and meaning. In *ACM SIGCOMM Internet Measurement Workshop*, pages 91–92, 2002.

[48] Gangxiang Shen and R. S. Tucker. Energy-minimized design for IP over WDM networks. *IEEE/OSA Journal of Optical Communications and Networking*, 1(1):176–186, 2009.

[49] Nina Skorin-Kapov. *Greedy Algorithms*, volume 30. InTech, 2008.

[50] N. Spring, R. Mahajan, D. Wetherall, and T. Anderson. Measuring ISP topologies with Rocketfuel. *IEEE/ACM Transactions on Networking*, 12(1):2–16, 2004.

[51] Technical Staff. Heat density trends in data processing, computer systems and telecommunications equipment. White paper, The Uptime Institute, 2000. http://www. itcrisis.com /pdf/library/Heat_Density1. pdf.

[52] J. Teng and G. N. Rouskas. Wavelength selection in OBS networks using traffic engineering and priority-based concepts. *IEEE Journal on Selected Areas in Communications*, 23(8):1658–1669, 2005.

[53] M. Tornatore, G. Maier, and A. Pattavina. WDM network design by ILP models based on flow aggregation. *IEEE/ACM Transactions on Networking*, 15(3):709–720, 2007.

[54] P. N. Tran and U. Killati. Design of logical topology for IP over WDM networks: network performance vs. resource utilization. *INOC*, 2007.

[55] F. Vismara, V. Grkovic, F. Musumeci, M. Tornatore, and S. Bregni. On the energy efficiency of IP-over-WDM networks. In *Proceedings of the IEEE Latin-American Conference on Communications*, pages 1–6, 2010.

[56] H. S. Wang, L. S. Peh, and S. Malik. A power model for routers: modeling Alpha 21364 and InfiniBand routers. In *Proceedings of IEEE 10th Symposium on High Performance Interconnects*, pages 21–27, 2002.

[57] Hao Wang, Haiyong Xie, Lili Qiu, Yang Richard Yang, Yin Zhang, and Albert Greenberg. COPE: traffic engineering in dynamic networks. In *Proceedings of ACM onference on Applications, technologies, architectures, and protocols for computer communications (SIGCOMM)*, pages 99–110, 2006.

[58] Douglas B. West. *Introduction to Graph Theory*. Prentice Hall, 2000.

[59] G. Xiao, Y. Leung, and K. Hung. Two-stage cut saturation algorithm for designing all-optical networks. *IEEE Transactions on Communications*, 49:1102–1115, 2001.

[60] Y. Xin, G. N. Rouskas, and H. G. Perros. On the physical and logical topology design of large-scale optical networks. *Journal of Lightwave Technology*, 21:904–915, 2003.

[61] Sugang Xu, K. Sezaki, and Y. Tanaka. A two-stage simulated annealing logical topology reconfiguration in IP over WDM networks. In *Proceedings of IEEE 11th International conference on Telecommunications Network Strategy and Planning Symposium (NETWORKS)*, pages 327–332, 2004.

[62] Emre Yetginer and George N. Rouskas. Power efficient traffic grooming in optical WDM networks. In *Proceedings of the 28th IEEE conference on Global telecommunications (GLOBECOM)*, pages 1838–1843, 2009.

[63] S. Yook, H. Jeong, and A. Barabasi. Modeling the Internet's large-scale topology. *PNAS*, 99:21, 2002.

[64] G. Young. Objectives for service provider shared transport of 802.3 higher speed Ethernet. Technical report, AT&T, 2005. http://www.ieee802.org/3/hssg/public/nov06//young_01_1106.pdf.

[65] C. Zhang, Y. Liu, W. Gong, J. Kurose, R. Moll, and D. Towsley. On optimal routing with multiple traffic matrices. In *Proceedings of the 24th Annual Joint Conference of the IEEE Computer and Communications Societies (INFOCOM)*, pages 607–618, 2005.

[66] K. Zhu and B. Mukherjee. Traffic grooming in an optical WDM mesh network. *IEEE Journal on Selected Areas in Communications*, 20(1):122–133, 2002.

[67] Keyao Zhu, Hongyue Zhu, and Biswanath Mukherjee. *Traffic Grooming in Optical WDM Mesh Networks*. Springer Science, 2005.