UNIVERSITY OF CALIFORNIA, SAN DIEGO

Physical Layer Driven Optical Switching for Data Center Networks

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Electrical Engineering (Photonics)

by

William Maxwell Mellette

Committee in charge:

Professor Joseph E. Ford, Chair
Professor George Papen
Professor George Porter
Professor Stojan Radic
Professor Alex Snoeren

2016

The dissertation of William Maxwell Mellette is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California, San Diego

2016

DEDICATION

To my parents.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I consider myself fortunate to have had the opportunity to work with many talented people over the last four years. I would first like to thank my advisor, Professor Joseph Ford, for his guidance, insight, and steadfast support over the course of my research. His perspective has been invaluable in shaping the direction of my research. I would like to thank the other members of my committee: George Papen, George Porter, Stojan Radic, and Alex Snoeren for investing their time and effort in me. Although Professor Ford was my sole advisor, I have worked closely with George Porter, George Papen, and Alex Snoeren, who have all helped me broaden the scope of my research with their expertise in the systems aspects of communication networks.

Thanks also go to my friends, colleagues, and to the members of the Photonic Systems Integration Lab. Special thanks to Rachel whose support and positivity had a daily impact on my endeavors. Finally, I would like to thank my parents, Carol and Gary, who have devoted their lives to me and provided me with their love and support.

The interdisciplinary nature of my research involved a number of collaborators, who are listed next.

Chapter 2, in part, reprints material from the paper titled: "Scaling Limits of MEMS Beam-Steering Switches for Data Center Networks," published in the *Journal of Lightwave Technology*, 33(15), pp 3308-3318, 2015, by W. M. Mellette and J. E. Ford.

Chapters 2 and 3, in part, reprint material submitted for publication in a paper titled: "A Scalable, Partially Configurable Optical Switch for Data Center Networks," submitted to the *Journal of Lightwave Technology*, by W. M. Mellette, G. M. Schuster, G. Porter, G. Papen, and J. E. Ford.

Chapter 4, in part, is being prepared for submission in a paper tentatively titled: "SelecToR: A Scalable, Partially Configurable Data Center Network Architecture," by W. M. Mellette, J. R. McGuinness, A. Forencich, G. Papen, A. Snoeren, J. E. Ford, and G. Porter.

Appendix A reprints material from the paper titled: "Planar waveguide LED illuminator with controlled directionality and divergence," published in *Optics Express*, 22(S3), pp A742-A758, 2014, by W. M. Mellette, G. M. Schuster, and J. E. Ford.

VITA

2012	Bachelor of Science in Engineering Physics *magna cum laude* with department honors, University of California, San Diego
2014	Master of Science in Electrical Engineering (Photonics), University of California, San Diego
2015	Candidate in Philosophy in Electrical Engineering (Photonics), University of California, San Diego
2012–2016	Research Assistant, Photonic Systems Integration Lab, Department of Electrical and Computer Engineering, University of California, San Diego
2016	Doctor of Philosophy in Electrical Engineering (Photonics), University of California, San Diego

JOURNAL PUBLICATIONS

William M. Mellette, Glenn M. Schuster, George Porter, George Papen, and Joseph E. Ford, "A Scalable, Partially Configurable Optical Switch for Data Center Networks," submitted to *Journal of Lightwave Technology*.

William M. Mellette and Joseph E. Ford, "Scaling Limits of MEMS Beam-Steering Switches for Data Center Networks," *Journal of Lightwave Technology*, 33(15), pp 3308-3318, 2015.

Salman Karbasi, Ashkan Arianpour, Nojan Motamedi, William M. Mellette, and Joseph E. Ford, "Quantitative analysis and temperature-induced variations of moiré pattern in fiber-coupled image sensors," *Applied Optics*, 54(17), pp 5444-5452, 2015.

William M. Mellette, Glenn M. Schuster, and Joseph E. Ford, "Planar waveguide LED illuminator with controlled directionality and divergence," *Optics Express*, 22(S3), pp A742-A758, 2014.

CONFERENCE PROCEEDINGS

Joseph E. Ford, Salman Karbasi, Ilya Agurok, Igor Stamenov, Glenn Schuster, Nojan Motamedi, William M. Mellette, Adam R. Johnson, Ryan Tennill, and Ron A. Stack, "Panoramic imaging with monocentric lenses and curved fiber bundles," to be published in *SPIE Defense and Commercial Sensing*, 2016.

William M. Mellette, Glenn M. Schuster, George Porter, and Joseph E. Ford, "61 Port 1×6 Selector Switch for Data Center Networks," in *Optical Fiber Communication Conference*, pp M3I 1-3, 2016.

William M. Mellette and Joseph E. Ford, "Scaling Limits of Free-Space Tilt Mirror MEMS Switches for Data Center Networks," in *Optical Fiber Communication Conference*, pp M2B 1-3, 2015.

Stephen J. Olivas, Michal Sorel, Ashkan Arianpour, Igor Stamenov, Nima Nikzad, Glenn Schuster, Nojan Motamedi, William M. Mellette, Ronald A. Stack, Adam R. Johnson, Rick Morrison, Ilya Agurok, and Joseph E. Ford, "Digital image processing for wide-angle highly spatially variant imagers," in *SPIE Optical Engineering and Applications*, pp 91930B 1-3, 2014.

Stephen J. Olivas, Nima Nikzad, Igor Stamenov, Ashkan Arianpour, Glenn Schuster, Nojan Motamedi, William M. Mellette, Ronald A. Stack, Adam R. Johnson, Rick Morrison, Ilya Agurok, and Joseph E. Ford, "Fiber Bundle Image Relay for Monocentric Lenses," in *Computational Optical Sensing and Imaging*, pp CTh1C.5 1-3, 2014.

William M. Mellette, Glenn M. Schuster, Ilya P. Agurok, and Joseph E. Ford, "Planar waveguide Illuminator with Variable Directionality and Divergence," in *Solid-State and Organic Lighting*, pp DT3E-3, 2013.

ABSTRACT OF THE DISSERTATION

Physical Layer Driven Optical Switching for Data Center Networks

by

William Maxwell Mellette

Doctor of Philosophy in Electrical Engineering (Photonics)

University of California, San Diego, 2016

Professor Joseph E. Ford, Chair

Today's data center networks operate at the cutting edge of fiber optic link and electronic packet switching capabilities. The immense bandwidth requirements of next-generation data centers will stress the limits of electronic switching, providing an opportunity for transparent optical switching to deliver an overall cost-bandwidth advantage. However, current optical switching approaches are not optimal for data center networks because they either do not scale to large port count, reconfigure too slowly, or introduce insertion loss or crosstalk levels incompatible with cost-effective optical transceivers. This dissertation presents the design and demonstration of a novel optical switch architecture more well-suited to data centers, along with the design of overall network architectures that employ this new switch architecture.

The dissertation begins at the physical layer with a scalability assessment of conventional microelectromechanical systems (MEMS) based beam-steering optical switching. MEMS beam-steering cross-connects are the only optical switching technology which has demonstrated the large port count and broadband, polarization-

insensitive transmission necessary to approach the scale and link power budgets of modern data center networks. The shortcoming of conventional cross-connects is their slow reconfiguration time, which prevents them from effectively provisioning bandwidth on the timescales necessary for a potentially large fraction of data center traffic. First-principles analysis at the device level indicates that, rather than a straightforward redesign of existing crossbar switches, entirely new switch architectures are necessary to meet the optical switching performance required for data centers.

Motivated by physical layer analysis, a novel *selector switch* architecture is presented which, through an unconventional approach of relaxing the degree of switch configurability, allows MEMS beam-steering switching elements to scale to microsecond-class response speeds while supporting large port count and low loss switching. The switch is *partially configurable* in that it selects port mapping patterns from a small hardware library of preconfigured mappings, rather than implementing arbitrary mappings like a crossbar. The physical architecture of the switch uses pupil-division and relay imaging, permitting designs compatible with single-mode or multi-mode fiber optics. The design, fabrication, and experimental characterization is presented for a proof-of-principle prototype using a single MEMS comb-driven micromirror to achieve 150 μs switching of 61 single-mode ports between 4 preconfigured port mappings. The scalability of this switch architecture is demonstrated with the detailed optical design of a low-loss 2,048-port selector switch with 20 μs switching time.

Because conventional network architectures are typically based on crossbar switches, new overall network architectures are required to utilize the partial configurability of selector switches. The dissertation concludes with an investigation of network architectures based on selector switches, showing, perhaps unexpectedly, that partially configurable networks can deliver aggregate bandwidth approaching that of a fully-provisioned electronically-switched network for common network traffic patterns, but for reduced cost, cabling complexity, and power consumption.

The approach taken in this dissertation of developing switch and network architectures which balance scalability at the physical layer and performance at the network layer will hopefully aid in the design of future optical data center networks.

Chapter 1

Introduction

High performance data center networks interconnecting tens- to hundreds-of-thousands of servers enable the modern web applications, storage capabilities, and cloud computing platforms provided by companies like Google, Facebook, Microsoft, Amazon, and many others. The sustained growth in demand for these services continues to stress the scalability of the underlying network infrastructure, which today relies on commodity electronic packet switching. Bandwidth demand within data centers is now growing at a faster rate than in the wide area Internet, spurred by the need to process ever-larger datasets. Paralleling the wide area telecom networks of the 1980's, today's data center operators have already replaced copper cables with point-to-point fiber optic links throughout most of the data center to support the growing interconnection bandwidth requirements. Links carrying 100 Gb/s over hundreds to thousands of meters are currently being installed in warehouse- and campus-sized data centers. Communication at these data rates and distances can only be supported by fiber optic transceivers.

The evolution of data center networks will likely once again parallel that of telecom networks in a paradigm shift from using optics for point-to-point communication to incorporating switching functionality at the optical layer. Similar to telecom operators who found it more cost-effective to replace electronic switches with transparent optical switches as network data rates increased, the growing aggregate bandwidth demands and impending capacity limitations of electronic packet switches in data centers will motivate operators to adopt optical switching in their networks.

While similar underlying technology trends in optical links and electronic switches may correlate the move to optical switching in telecom and data center networks, data centers are subject to entirely different cost and operating models than telecom networks. For example, laying new fiber in a wide area telecom network may be prohibitively expensive, necessitating complex and expensive but spectrally-efficient transceivers along with optical amplifiers to efficiently use the existing fiber plant. Conversely, fiber is relatively abundant within warehouse-sized data centers, with short (100 m) spans permitting inexpensive multimode signaling. High transceiver density in data centers requires low power operation, meaning uncooled, non-retimed, course-wavelength-division-multiplexed (CWMD) transceivers with minimal optical link power budgets are preferred. Wide area networks provision high bandwidth lightpaths between geographically separated endpoints on long timescales, permitting slow optical switch reconfiguration on the order of milliseconds. Microsecond-scale protection switches require only a small number of fail-over ports in telecom networks. In the data center, traffic patterns between many thousands of endpoints change on short timescales and exhibit multicast, requiring microsecond-scale optical switching to provision bandwidth between hundreds to thousands of ports. Telecom components are designed to stringent specifications for long operating lifetimes without service, while data center components are upgraded every few years. For these reasons among others, the network architectures and optical switching hardware used in wide area telecom networks are not well-suited to data center networks. Other contemporary research-level approaches to optical switching also face significant barriers to entry in data centers because they either do not scale to large port count, reconfigure too slowly, and/or introduce insertion loss or crosstalk levels incompatible with cost-effective optical transceivers.

This dissertation, through modeling and experiment, addresses the design of optical switches and optically-switched networks subject to the practical constraints of data centers. To briefly summarize the findings, novel optical switch and network architectures are identified which, based on the principle of partial configurability, can realize next-generation networks with better cost, complexity, and bandwidth scaling properties than existing approaches. The dissertation is organized as follows.

Chapter 2 studies the tradeoffs between the port count, switching speed, and optical transmission of conventional MEMS cross-connects using a first-principles physical-layer model. The theoretical results show that switching speed is inversely proportional to port count, and indicate that optical signal transmission is a weak mediator of that proportionality. The model also suggests that the switching speed of commercial cross-connects cannot be substantially improved without tighter alignment tolerances, multilayer electrical routing, higher drive voltage, small lithographic feature size, and more complex actuator structures, all of which increase manufacturing cost. Multistage switch architectures are analyzed which can overcome some of the scaling limitations of the conventional cross-connect architecture, but do so at the expense of increased insertion loss.

Motivated by the analysis in Chapter 2, Chapter 3 presents a novel *selector switch* architecture which, through an unconventional approach of relaxing the requirement of arbitrary switch configurability, allows MEMS beam-steering micromirrors to scale to microsecond-class response speeds while supporting large port count and low loss switching. This *partially configurable* optical switch does not retain the non-blocking properties of a crossbar, and instead selects between a set of preconfigured interconnection patterns. The design, fabrication, and experimental characterization is presented for a proof-of-principle prototype using a single MEMS comb-driven micromirror to achieve 150 μs switching of 61 single-mode fiber ports between 4 preconfigured interconnection port mappings. The scalability of this switch architecture is demonstrated with the detailed optical design of a low-loss 2,048-port selector switch with 20 μs switching time.

Because conventional network architectures are typically based on crossbar switches, new overall network architectures are required to leverage the partial configurability of selector switches. Chapter 4 investigates network architectures based on selector switches, showing, perhaps unexpectedly, that partially configurable networks can deliver bandwidth approaching that of a fully-provisioned electronically-switched network for common network traffic patterns, but for reduced cost, cabling complexity, and power consumption.

The approach taken in this dissertation of exploring the joint design space of the physical and network architecture layers will hopefully find applications in the design of future high speed data center networks.

1.1 Scope

This dissertation primarily focuses on the physical layer aspects of optical switches and the architectural aspects of optical data center networks. There are a number of closely-related topics which are outside the scope of this dissertation.

One important topic related to optically-switched networks is the design of robust control planes. Data can enter an electronic packet switch at any time because it can be buffered as the switch prepares to forward the data to its destination. In this case, each packet's destination is read by the switch and data is stored physically as electrons in transistors. Unfortunately, there is no practical way to buffer photons in an optical switch, and the transmission of data through the switch must be synchronized with the state of the switch in order for data to reach its proper destination. Further, data must not be sent through an optical switch during the reconfiguration of light paths. Fortunately, all hardware is typically under common ownership in a data center, permitting the required degree of synchronization. A number of recent proposals address this topic [1], [2], [3]. An open question in this area is whether centralized or distributed control planes will yield better performance in large scale networks.

Another topic, related to control planes, is the choice of communication protocol. Protocols define how information is addressed, routed, and checked for errors or drops. For example, TCP (Transmission Control Protocol) is a popular protocol, designed for the wide area internet to allow efficient communication between endpoints with different network interface hardware. In a data center where the network interfaces of all endpoints are under common ownership and control, modified protocols may be more effective. The interplay between communication protocols, control planes, and network hardware can perhaps be addressed through the larger topic known as software-defined networking (SDN). Many researchers are working in this field, and this effort may

provide insights into how to most effectively control the transmission of data in optically-switched data center networks.

This dissertation focuses on optical *circuit* switching, meaning the optical switch does not decode data for routing purposes. Optical *packet* switching is another approach being investigated, in which the header of each packet is read (typically electronically) by the switch and the packet's payload (data) is switched optically to the appropriate port. While it promises fine switching granularity, optical packet switching faces a number of implementation challenges, such as realizing practical optical packet buffering and packet-level parsing. These challenges may be overcome in the future, but in this dissertation we focus instead on the design of practical and cost-effective optical circuit switches aimed for relatively near-term adoption.

1.2 Related Work

Perhaps owing to its large potential impact, there are many research groups working on optical switching and optically-switched systems. Some prominent work at the system and device levels is reviewed below.

1.2.1 Related System-level Work

One prominent class of hybrid optical-electronic networks proposed recently is based on conventional optical cross-connect switches [4], [5], [6]. Because these optical switches take tens of milliseconds to reconfigure, the optical portion of these networks can only support the most stable and sparse communication patterns. Traffic stability is required because the large reconfiguration time would impose a significant duty-cycle penalty if the switch state were altered too frequently. Sparse traffic is required because optical switches cannot establish multicast connections. This type of network is well-suited to large scale data migration between racks or clusters of servers, or for altering the network topology on long timescales. However, a significant amount of data center

traffic is short-lived and exhibits multicast connection patterns, both of which are not efficiently served by this first class of optical networks.

More recent work has demonstrated a hybrid network with faster optical switching using wavelength-selective optical switches in a ring topology [7], [8]. The switches had a 10 μs reconfiguration time, allowing the optical network to efficiently serve a more diverse set of traffic patterns, including all-to-all type workloads. However, the increased system level performance came at the cost of a more expensive and complex physical network with limited scalability. The high insertion loss of the cascaded wavelength selective switches and the splitting losses inherent to the ring topology required (expensive) optical amplification. Further, each server required an optical transceiver with a unique wavelength, limiting the number of servers to the number of unique wavelength in the erbium-based optical amplification window (only 88 wavelengths assuming 50-GHz spaced channels). Beyond the 10 Gb/s links employed in [8], modern 40 and 100 Gb/s transceivers use four wavelengths modulated at 10 and 25 Gb/s each, which further reduces the effective pool of unique wavelengths (and the number of network end hosts) in order to support these higher data rates.

Other approaches propose the use of freespace optics in data centers [9], [10]. These networks may reduce cabling complexity and have the potential for large fan-out. However, freespace optical links are typically low bandwidth due to the relatively large area (and high-capacitance) photodetectors required. Lenses can be used to focus the centimeter-diameter optical beams required to traverse a large-scale data center onto the micrometer-scale high speed photodetectors required for high speed data transmission, but this leads to extremely tight alignment tolerances across the data center. The major practical limitation of these approaches is that the entire data center becomes an optical switch which requires a carefully-controlled (warehouse-sized) environment. Active alignment would be necessary to compensate for thermal expansion and highly skilled technicians would be needed to set up and troubleshoot connections.

The works mentioned above make a first cut at several points in the system level design space of optically-switched data center networks, but significant work remains to further explore this design space in search of practical and scalable solutions.

Table 1.1: Key performance metrics of demonstrated optical switches

Technology	Ports	Speed (μs)	Crosstalk (dB)	On chip loss (dB)	Fiber to fiber loss (dB)	Ref.
Semiconductor optical amplifier	16	< 0.01	-10	30	40	[11]
Electro-optic Mach Zehnder	8	~ 0.01	-15	-	20	[12]
Thermo-optic Mach Zehnder	8	30	-20	6.5	13.7	[13]
MEMS actuated waveguide	64	1	-60	4	10 (est.)	[14]
3-D beam-steering MEMS	1,100	1×10^5	-60	N/A	4	[15]

1.2.2 Related Device-level Work

There are two primary classes of optical switches: integrated planar waveguide switches and fiber-coupled freespace beam-steering switches.

Freespace beam-steering switches operate by coupling lightwave signals form fiber optics into free space, and perform switching on the freespace signals, typically leveraging all three spatial dimensions. Switches based on MEMS beam-steering are commercially available today with hundreds of ports, low insertion loss, and reconfiguration times of tens of milliseconds. This is a result of the large effort into developing MEMS switches for telecommunication networks [16]. MEMS optical cross-connects with over 1,000 ports and 4 dB worst-case insertion loss have been demonstrated in research [15].

Contemporary research in optical switching has shifted away from freespace beam-steering to integrated planar waveguide switches fabricated from silicon or III-V materials. These switches keep light closely confined to optical waveguides and switch optical signals between waveguides, typically confining the switch to a two dimensional geometry. Nanosecond to microsecond switching speeds are possible, limited by the response time of the switching material. While inherently fast, each switching element is typically a 1 × 2 or 2 × 2 device, requiring multistage architectures to scale to larger port count. Signal loss and crosstalk accumulates from cascaded waveguide crossings and the switching elements themselves, leading to an undesirable tradeoff between port count

and insertion loss and crosstalk. Crosspoint architectures are more scalable, but have device count and chip area which scales as the square of the number of ports. The transmission of any type of fiber-coupled waveguide switch is limited by the efficiency of fiber/chip coupling. Practical polarization-splitting grating couplers can contribute an insertion loss of 6 dB per pair [17], already consuming the entire link budget of a standard long reach optical transceiver.

Table 1.1 shows the key metrics for a number of research-level optical switches. Missing from the list is a microsecond-class switch which is scalable to hundreds (or thousands) of ports with a low insertion loss and crosstalk compatible with cost-effective optical transceivers. The goal of this dissertation is to provide such switches, along with the overall network architectures to effectively employ them.

Chapter 2

Scaling Limits of MEMS Beam-steering Cross-connects

Commercial MEMS beam-steering cross-connects were designed to provision bandwidth in wide area telecommunications networks, requiring millisecond-scale reconfiguration speeds. Microsecond-scale protection switches were also developed for telecom networks, but required only a small number of fail-over ports. Due to their scale and traffic patterns, data center networks require fast switching between a large number of ports. The cost-effective optical transceivers used in data centers also necessitate low switch insertion loss and crosstalk. This chapter explores the design space of MEMS beam-steering switches, with a particular focus on the tradeoffs between the switch port count, switching speed, and optical transmission and crosstalk. First-principles analysis at the device layer indicates that, rather than a straightforward redesign of conventional telecom switches, entirely new switch architectures will be necessary to meet the optical switching performance required for data center networks.

A number of physical-layer characteristics parameterize the design space of MEMS beam-steering switches. The nomenclature used in this chapter to describe these parameters is summarized below in Table 2.1.

Table 2.1: Nomenclature used in Chapter 2

Symbol	Explanation	Main occurrence
α	Mirror array fill factor	Sec. 2.2.4
A	Area of electrode overlap	Fig. 2.4
β	Half divergence angle of optical beam	Fig. 2.5
η_{ac}	Angular confinement efficiency	Sec. 2.2.4
η_f	Fiber coupling efficiency	Sec. 2.2.4
η_{sc}	Spatial confinement efficiency	Eq. (2.15)
η_{switch}	Overall optical efficiency of switch	Eq. (2.16)
ε_0	Permittivity of free space	Sec. 2.2.2
E	Young's modulus	Eqs. (2.12) & (2.13)
f_0	Natural resonant frequency	Eq. (2.9)
g	Finger to finger air gap (comb actuator)	Fig. 2.4
G	Shear modulus	Eq. (2.11)
h_m	Electrode-mirror air gap (plate actuator)	Fig. 2.4
i	Tilt axis, spanning x and y	
I	Moment of inertia	Sec. 2.2.3
k	Rotational spring constant	Eqs. (2.11) & (2.12)
λ	Wavelength	
l_f	Comb finger length	Fig. 2.4
l_s	Spring length	Sec. 2.2.3
M	Number differentiable optical mirror states	Eq. (2.14)
N	Number of ports in switch	
N_f	Number of comb fingers	Eq. (2.4)
ψ	Electrode ramp angle (plate actuator)	Fig. 2.4
r_m	Mirror radius	
R	Mirror reflectivity	Sec. 2.2.4
S	Number of active switching stages	Eqs. (2.1) & (2.2)
θ, θ_{max}	Mechanical tilt angle, maximum tilt angle	
t_f	Comb finger thickness	Fig. 2.4
t_m	Mirror thickness	Fig. 2.4
t_s	Spring thickness	Eq. (2.11)
τ	Torque	Sec. 2.2.2
V	Applied electrode voltage	Sec. 2.2.2
w_0	Waist of Gaussian beam	Sec. 2.2.4
w_s	Spring width	Eq. (2.11)
z_R	Rayleigh range of Gaussian beam	

2.1 Introduction

Optical circuit switching may augment or replace electronic switching and meet the size and bandwidth demands of future data center networks [4], [7], [18], [19]. For optical switching to be adopted in the data center, however, it must provide energy-efficient switching that reduces the net capital and operational cost of the overall network

without degrading overall performance. At the physical layer, three aspects of an optical switch impact its practical feasibility: insertion loss, port count, and switching speed.

Optical transceivers account for a large fraction of total data center network cost [20]. Replacing electronic switches with transparent optical switches reduces the number of transceivers required and can reduce total network cost. However, the optical switch insertion loss requires a larger link power budget, which will become increasingly expensive at higher data rates. In fact, datacom manufacturers are already developing 100 Gb/s transceivers [21] with lower power budget and cost than standard long reach (10 km) transceivers with link margins of approximately 6 dB. This technology trend correlates the insertion loss of an optical switch with the network cost, and makes minimizing signal attenuation, crosstalk, distortion, and polarization sensitivity key aspects of optical switch design.

Today's data center networks use a multi-stage folded Clos topology and electronic packet routers to interconnect servers [20], [22]. Each additional stage in the network requires a set of switches and optical interconnection links to the preceding and subsequent stages, increasing cost and cabling complexity. Optical circuit switches have the potential to scale to higher port count and higher per-port bandwidth than electronic switches, reducing cost and cabling complexity by flattening the network. Given that data center networks in production today connect 100,000 servers [20], providing direct connectivity between servers with a monolithic switch is impractical. Instead, transparent optical switches may be used to connect electronically-aggregated groups of servers (e.g. racks, pods, or clusters). Smaller aggregation groups require fewer stages of electronic switching, leading to flatter and less expensive networks, but require optical switches with more ports to interconnect the groups. For example, 2,000 port switches would be required to connect the racks of a 100,000 server network assuming 50 servers per rack.

Planar waveguide optical switches fabricated with Silicon or III-V materials are being investigated by a number of research groups. Nanosecond to microsecond reconfiguration speeds are possible with optical switches based on electro-optic modulation, semiconductor optical amplification, or thermo-optic modulation. However, the accumulated loss and crosstalk induced by their multistage architectures have limited

these switches to small port counts (≤8) [23], [24], [13] or high loss (>15 dB) [25]. Alternatively, MEMS-actuated silicon waveguide switching structures with microsecond response times have recently been reported and integrated into a 64-port cross-point matrix with 4 dB on chip loss [14]. However, because the size and complexity of planar cross-point architectures scale as the square of the port count, scaling these switches to hundreds of ports presents chip-area, loss, and yield challenges. Further, significant fiber-chip coupling losses preclude multi-chip topologies. Today, practical packaging approaches use polarization-splitting coupling structures to interface with standard transceivers and fiber, introducing a total insertion loss of over 6 dB [17].

Fiber-coupled free-space optical switches based on microelectromechanical systems (MEMS) beam-steering elements have an extensive publication record [16], and have proven successful in telecommunications networks for bandwidth provisioning and fault protection, which requires large port counts and low loss, but relatively slow switching speeds. MEMS beam-steering switches have been fabricated with over 1,000 ports and less than 4 dB worst-case insertion loss [15], approaching the port count and transmission requirements for deployment in data center networks. However, beam-steering cross-connects have response times on the order of 10 to 100 milliseconds, limiting their role to provisioning point-to-point bandwidth on second-long timescales [4]. While useful for latency-insensitive data migration, many data center applications exhibit short-lived communication patterns between many end-points [26], and cannot effectively utilize slow switching.

MEMS devices are not intrinsically slow; electrostatically actuated MEMS structures can have GHz resonant frequencies [27], but the optical requirements on beam-steering MEMS devices limits their response speed. Digital MEMS tilt mirrors are the fastest optical beam-steering devices, switching in 20 microseconds or less [28]. However, bistable operation has limited their use to small port-count switching. In data centers, both switching speed and port count are critical figures of merit, and sub-millisecond response times are essential to meet the network demands [7], [19], [18].

Here we re-examine canonical MEMS tilt mirror devices to quantify the tradeoffs between switching speed, port count, and optical transmission. From a network-level

perspective, our results can also be interpreted in terms of the number of reconfigurable ports achievable per second, which can be approximated by taking the product of the device resonant frequency and switch port count. A basic $1 \times N$ MEMS switch directs light from a single input fiber through free space to an electrostatically actuated mirror, which redirects the light to couple to one of N output fibers. $N \times N$ switches, with N inputs and N outputs, can be thought of as a collection of $1 \times M$ switches which use free space and relay optics to refocus light between a series of mirrors. While the specific switch layouts can differ, we can still compare MEMS device performance based on the fundamental requirement that each micromirror discriminates between optical switch states.

Using fundamental physical mechanics, electrostatics, and free-space optics, we investigate how the response speeds of canonical 1- and 2-axis tilt mirror devices scale as a function of switch port count, crosstalk, and insertion loss. The electrostatic, mechanical, and optical properties of the MEMS devices as well as switch topologies are discussed in Section 2.2. In Section 2.3, we describe the numerical approach used to quantify device performance, analyze the results by considering specific design cases in more detail, and compare the modeling results to a commercial switch. The findings motivate us to explore new overall optical switching configurations, which are discussed in Section 2.4.

2.2 Generalized MEMS Beam-steering Switch Model

Beam-steering switches fall into two major categories, those which incorporate wavelength selectivity using spectral demultiplexing and $1 \times N$ port topologies, or those that use wavelength-independent $N \times N$ port topologies (see Figure 2.1). A typical tilting micromirror device consists of a flat region to reflect a beam of light, a supporting structure to suspend the mirror and provide angular restoring force, and a set of nearby electrodes which apply electrostatic force to tilt the mirror. To explore a wide variety of switch configurations, we consider a switch as consisting of four modular parts: 1) an

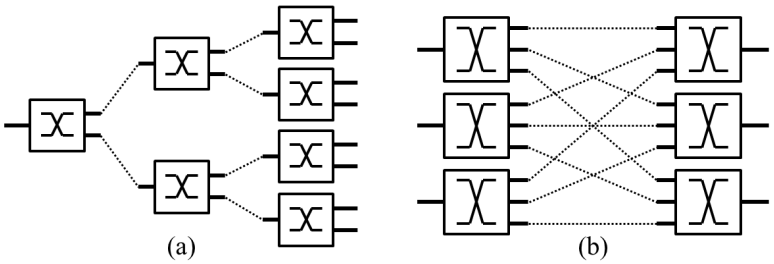

Figure 2.1: Schematic of common optical switch architectures. (a) 1×8 switch with 3 stages of 1×2 switching elements. (b) 3×3 switch using 2 stages of 1×3 elements. The latter exhibits the topology of a conventional $N \times N$ 3D-MEMS OXC.

overall switch architecture, 2) a MEMS actuation structure, 3) a mirror structure, and 4) a set of resolvable optical beam paths which meet transmission and crosstalk requirements.

Our analysis framework needed to be general enough to cover the scope of MEMS tilt mirror actuators, but include enough detail to accurately capture the behavior of each actuator type considered. The theoretical basis of our numerical Matlab model used a straight-field approximation, Euler-Bernoulli beam theory, Hooke's law, and Gaussian beam optics. This model is less accurate than a device-specific finite element simulation, but provides orders of magnitude faster execution. This allowed a search for optimal MEMS device designs over a large parameter space and the observation of scaling behavior over a large range of switch structures.

2.2.1 Overall Switch Architecture

The port count of a cross-connect is determined by the switch architecture. $1 \times N$ switches typically use a tree topology with one or more switching stages, where one input node branches sequentially into N output nodes (or vice versa), with each micromirror acting as a branching node in the tree (see Figure 2.1(a)). The number of stages in the tree, S, is related to the number of optical mirror states of each mirror, M, and the number of output ports, N, by

$$S = \log_M (N). \tag{2.1}$$

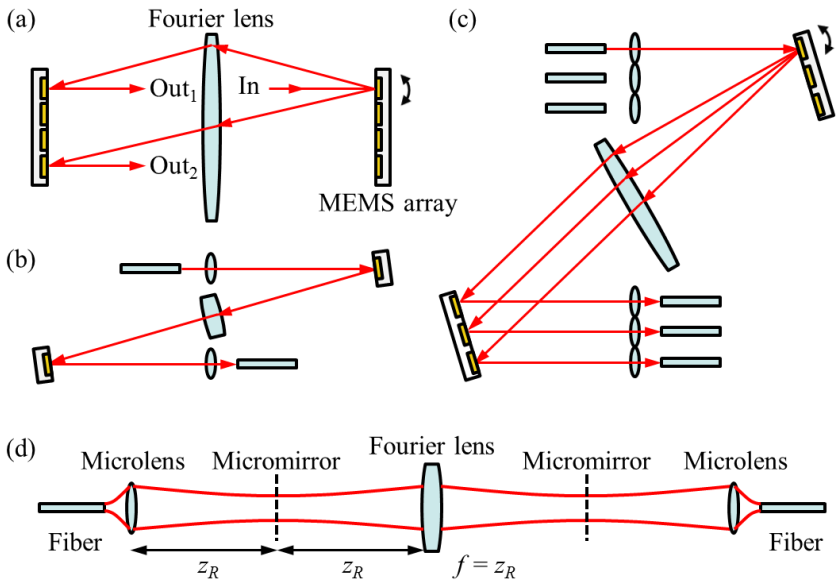

Figure 2.2: Operation of an $N \times N$ cross-connect. (a) Top view using a Fourier lens to make full use of the tilt range of all micromirrors in the MEMS array. Side views show separation of beam paths in the switch using (b) 1-axis and (c) 2-axis micromirrors. (d) Gaussian beam profile through the unfolded system, showing relaying of the beam waist between micromirror planes when the focal length of the Fourier lens and distance between MEMS array and Fourier lens equal the Rayleigh range of the beam.

$N \times N$ switches typically use a folded multi-rooted tree topology with at least two switching stages (see Figure 2.1(b)). In this topology,

$$S = 2\log_M(N). \tag{2.2}$$

Conventional free-space OXCs use $S = 2$ stages of N-state mirror elements ($M = N$), allowing N input and N output ports. Figure 2.2 illustrates $N \times N$ OXC geometries using 1- and 2-axis micromirrors. 1-axis switches use a linear array of mirrors and 2-axis switches use a two-dimensional array of mirrors. Introducing passive optics to aim the beam paths toward the center of the second array makes full use of the micromirrors' scan range, increasing the port count of a 2-axis switch by 4× compared to designs which do not incorporate passive beam aiming. There are a number of nearly equivalent techniques to aim the beams with passive optics, including field lenses at the collimator arrays, field lenses at the micromirror arrays, or a Fourier lens between micromirror arrays. Here we focus on the Fourier lens switch geometry for subsequent modeling and

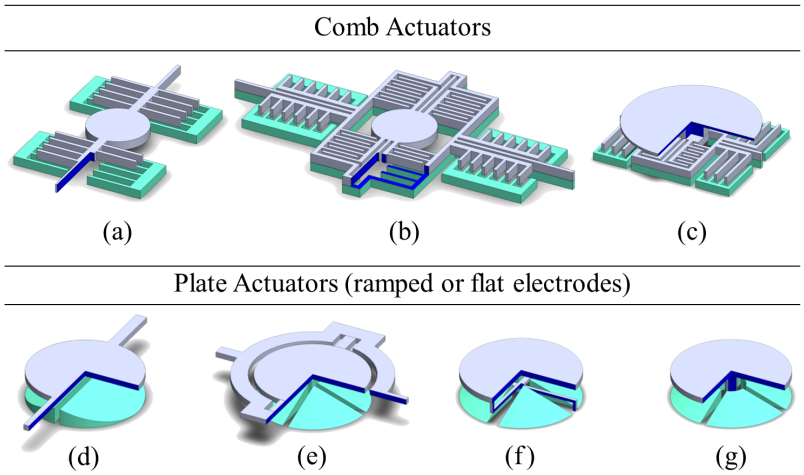

Figure 2.3: Illustrations of canonical MEMS beam-steering actuators. These actuators were considered in the design study. (a) 1-axis comb, (b) 2-axis in-plane comb, (c) 2-axis hidden comb, (d) 1-axis plate, (e) 2-axis plate with gimbal, (f) 2-axis plate with hidden "crossbar" springs, and (g) 2-axis plate with hidden "post" spring. Plate actuators are shown with ramped electrodes, but were also analyzed with flat electrodes. Partial cross sections have been taken to reveal the structure.

analysis [15]. Choosing the focal length of the Fourier lens to equal the Rayleigh range, z_R, of the optical beam and placing the lens one focal length from each MEMS array relays the beam waist between micromirrors. This reduces the aperture requirement on the mirrors, which lowers inertia and increases switching speed. Prior work has assessed the scaling of beam-steering cross-connects which do not employ a Fourier lens or other means of passive beam aiming [29].

2.2.2 Tilt Mirror Electrostatic Actuation

Choosing from the large number of actuators which have been proposed in the literature, we analyzed a set of commonly employed 1- and 2-axis torsional actuators using gap-closing plates and vertically offset combs [30] (Figure 2.3). Gap closing actuators are typically fabricated with parallel (flat) plate electrodes [31], but we also considered a ramped electrode design, which has been shown to have improved voltage response [32]. In addition to in-plane gimbaled 2-axis tilt mirrors, we considered

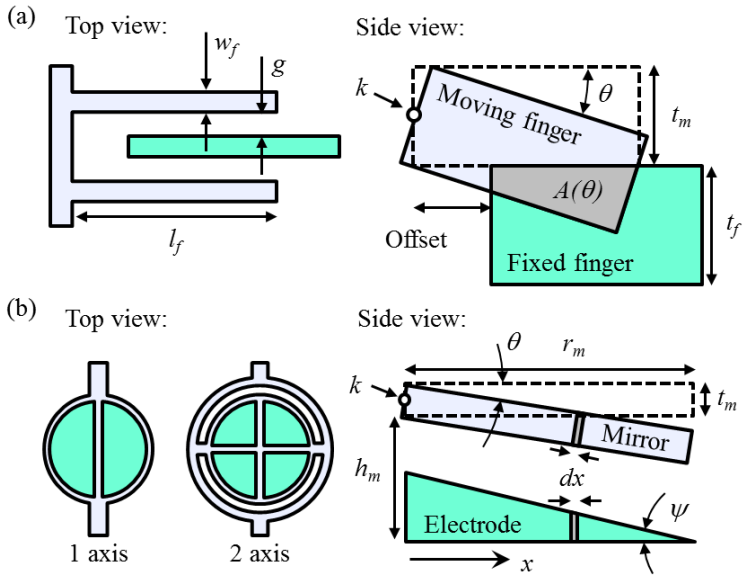

Figure 2.4: Geometric parameters of MEMS beam-steering actuators. Cross sectional illustrations of (a) comb and (b) plate actuation mechanisms.

variations with the support structures hidden under the mirror [33], [34]. Although more difficult to fabricate, designs with hidden springs reduce the rotational inertial and increase the density of mirrors in the array. There are alternative hidden actuator designs with different design constraints (e.g. [35]) which are not considered here.

The maximum optical beam-steering angle is determined by the mechanical tilt range of the mirror, which we found by balancing the restoring torque of the supporting springs with the electrostatic torque applied by the electrodes. The mechanical restoring torque is approximated by Hooke's law,

$$\tau_i = k_i \theta_i , \tag{2.3}$$

where k_i is the rotational spring constant and θ_i is the mechanical tilt angle about the i^{th} axis, where i spans x and y.

The driving torque of the comb actuator, calculated by differentiating the stored energy in the effective capacitor, is

$$\tau_{comb,i} = N_{f,i} \frac{\varepsilon_0 V_i^2}{2g} \frac{dA_i}{d\theta_i} , \tag{2.4}$$

where N_f is the number of comb fingers, ε_0 is the vacuum permittivity, V is the applied voltage, g is the air gap between comb finger electrodes, A is the area of electrode

overlap, and θ is the tilt angle, all for the i^{th} axis [36]. As shown in Figure 2.4(a), the electrode overlap area A depends on θ, the finger thicknesses, finger length, and fixed finger offset. For the in-plane comb actuator, we assume the mirror, comb arm, comb fingers, and torsion spring are fabricated from the same device layer in a single etch step, and must therefore have the same thickness, t_m. The fixed finger thickness, t_f, is defined by a separate device layer, and can have a different thickness. We found that thicker fixed fingers ($t_f > t_m$) increased the performance of the 1-axis comb actuator by allowing larger tilt angles, because large tilt angles are necessary to achieve large port counts in 1-axis actuators. 2-axis actuators have an increased dimensionality of tilt, and do not require such large tilt angles along each axis. We found that when optimizing the 2-axis comb actuators for speed, the optimal devices always tilted slightly less than the thickness of the moving comb finger. This can be explained because the rate of change in capacitance with angle begins to diminish when the top of the moving finger tilts below the top of the fixed finger, reducing the applied torque past this point. Operation in this regime allows larger tilt, but requires softer torsion springs (and lower resonant frequency) for the same electrode voltage. The tradeoff between spring stiffness and tilt angle favored stiffer springs for the 2-axis devices, and did not require the fixed fingers to be thicker than the moving fingers, at least for the port counts considered here. Note that optimized comb actuators still had larger scan angles than plate actuators, fulfilling the expected design advantages of comb drives.

Small structural asymmetries can excite lateral failure modes of the comb drive, imposing additional limits on the maximum tilt angle [37]. We consider these effects as inherent to comb drives and include them in our model. We set the gap between comb fingers, g, and the comb finger width, w_f, to be 2 micrometers to maximize comb density [37].

The torque generated by the plate actuator was calculated by integrating the forces exerted on the mirror by the electrode, neglecting fringing fields. For 1-axis tilt, we integrated over the electrode in a radial direction, giving an applied torque of

$$\tau_{plate,x} = \int_0^{r_m} x dF \,, \tag{2.5}$$

where r_m is the mirror radius, and x is the direction normal to the rotation axis. The incremental applied force dF is

$$dF = \frac{\varepsilon_0 V^2 \sqrt{r_m^{\,2} - x^2}\, dx}{\left(h_m - r_m\psi - x\left(\theta - \psi\right)\right)^2}, \tag{2.6}$$

where h_m is the nominal air gap between the mirror and a flat electrode and ψ is the electrode ramp angle, defined in Figure 2.4(b) [38]. Note that $\psi = 0$ for a flat electrode. In the 2-axis plate actuator, the ramped electrode is conical in shape and we integrate in two dimensions, giving an applied torque of

$$\tau_{plate,i} = \int_{\rho_1}^{r_m} \int_{\varphi_1}^{\varphi_2} \rho \cos\varphi \, dF, \tag{2.7}$$

where ρ is the radial coordinate and φ is the azimuthal coordinate of a cylindrical coordinate system and

$$dF = \frac{\varepsilon_0 V^2}{\left(h_m - r_m\psi - \rho\left(\theta\cos\varphi - \psi\right)\right)^2} \rho\, d\varphi\, d\rho. \tag{2.8}$$

The mechanical tilt range of plate actuators considered here was less than ±13°, so (2.6) and (2.8) remain reasonably valid.

We assumed the use of four quadrant (90°) electrodes as in Figure 2.4(b), such that when tilting the mirror in a direction centered on a quadrant (45° from a quadrant boundary line), higher torque is applied by activating three electrodes rather than a single electrode. The gimbal design allows $\rho_1 = 0$ in (2.7), but for the post design ρ_1 must be greater than the post radius. In the hidden crossbar design, the width of the springs cuts into the area of the electrodes, and we must modify the integration limits in (2.7) accordingly. We found that maximal electrode ramp angles produced the highest performing plate actuator devices, except in the case of the hidden post spring, where the center cut-out in the electrodes to allow for the post negated the benefit of the ramped electrodes. Complex electrode designs can improve device performance [39], [40], but to maintain the large scope of our study, we focused on the most common designs, illustrated in Figure 2.4.

We limited the applied electrode voltage to 275 V to avoid electrostatic breakdown [41]. Care was taken in calculating the maximum tilt angle for the plate

actuator; past some tilt angle (typically 44% of the maximal angle), the nonlinearity in the torque exerted by the electrode overcomes the linear restoring torque of the spring and the mirror is snapped down to the substrate. This is the well-known "pull-in" phenomenon [31].

The plate actuator can be purposely operated in the pull-in regime, allowing the mirror to be snapped to a discrete number of mechanical states in a "digital" fashion [28], where the mirror structure accelerates until it reaches contact with a mechanical stop. This mode of operation allows switching on microsecond time scales, but the small number of mechanical and optical states limits the port count of the switch. Alternatively, the mirror can be operated with continuous "analog" positioning over a smaller angular range, allowing more optical states but with a slower reconfiguration rate. Digital vs. analog actuation is a critical switch design choice.

2.2.3 Tilt Mirror Dynamics

The maximum device switching speed is primarily limited by the resonant frequency at which the mirror structure oscillates. While driving the mirror faster than its natural resonant frequency is possible, this requires sophisticated high-voltage closed-loop control which is likely to be impractical to implement at high switching speeds due to the necessarily high device driver currents. The natural resonant frequency, f_0, of the device is given as

$$f_{0,i} = \frac{1}{2\pi}\sqrt{\frac{k_i}{I_i}}, \tag{2.9}$$

where I_i is the moment of inertia about the i^{th} axis. Mass located farther from the axis of rotation has a larger contribution to the rotational inertia. There can be multiple resonances of the structure [42], and the actuator design must ensure the desired torsional mode has the lowest resonant frequency in order to suppress unwanted motion in parasitic modes. The operational resonant frequency is proportional to the natural

resonant frequency, but depends on both damping, Γ, and driving torque, τ_{drive}, and can be found by solving the full equation of motion given by

$$I\frac{d^2\theta}{dt^2}+\Gamma\frac{d\theta}{dt}+k\theta=\tau_{drive}(\theta). \tag{2.10}$$

The transient solution can be found by detailed calculation of the damping term [43]. The damping can be fine-tuned by changing the ambient gas pressure or shape of the cavity beneath the mirror, or by etching small holes in the mirror [44]. For our analysis, the driven resonant frequency in the absence of damping is a sufficient metric for comparing the response speeds of different devices because damping effects establish a proportionality between driven resonant frequency and response time, and because that proportionality factor is tunable, it can be made similar in all devices considered.

Comb actuators have a nearly linear response because the driving torque is nearly constant as a function of θ (up to the angle at which the comb fingers are fully interdigitated). In the absence of damping, then, the driven resonant frequency of the comb actuator is its natural frequency. The driving torque of the plate actuator, on the other hand, is highly nonlinear in θ. The driven resonant frequency drops as a function of tilt angle, approaching zero at the pull-in angle. Closed-loop control can extend the analog tilt range of the mirror [45], but providing sufficient voltage and current for closed-loop control becomes extremely challenging with fast switching devices, so we assumed open-loop control. This simplifies drive electronics and allows a direct comparison to the comb actuator, which does not exhibit the same vertical pull-in effect. This means some angular margin must be maintained between the maximum operational tilt angle and the pull-in angle in order to drive the plate actuator in an analog fashion at high speeds.

The physical geometry of the suspension structures determines their stiffness. The rotational spring constant of the torsional elements is given by

$$k_{torsion}=\frac{2G}{l_s}ab^3\left(\frac{1}{3}-0.21\frac{b}{a}\left(1-\frac{b^4}{12a^4}\right)\right), \tag{2.11}$$

where G is the shear modulus of the material (polysilicon), l_s is the length of the spring, and a and b are the longer and shorter dimensions of the beam cross section, respectively.

Depending on the design, either a or b can assume the spring width, w_s, or the spring thickness, t_s. For in-plane devices, the spring thickness was set equal to the mirror thickness so both structures could be fabricated from the same device layer in a single etch step. The hidden actuator devices decouple mirror thickness from spring thickness. We approximated the flexure structure in the hidden post design as having a rotational spring constant given by

$$k_{flexure} = \frac{\pi E r_{post}^{\ 4}}{l_s}, \tag{2.12}$$

where E is Young's modulus of the material (polysilicon), r_{post} is the post radius, and l_s is the spring length. Nonlinear springs are commonly used in MEMS structures and have been shown to extend the tilt range of micromirrors [46]. However, to maintain the scope of the study, we used linear springs in our model because they do not require case-by-case optimization.

In practice, the finite translational stiffness of the springs means that an applied electrostatic force will contribute to moving the mirror vertically (in a piston mode), and will slightly reduce the torsional deflection. We found the torsional spring constant was at least an order of magnitude weaker than the flexure spring constant for the high-aspect ratio springs considered here, so we approximated that all applied force contributed to the torsional mode. Case-by-case spring optimization could further suppress the piston mode.

Because of its finite stiffness, the micromirror bends under static and dynamic actuation. We constrained the mirror to maintain a flatness of $1/8^{th}$ the wavelength to satisfy the Rayleigh criterion. Because the plate actuator applies force directly onto the mirror, we required the mirror thickness to increase with mirror radius and spring constant to maintain flatness under static deflection:

$$t_m = \left(\frac{16 r_m k}{E} \frac{\tan\theta}{\lambda} \right)^{1/3}. \tag{2.13}$$

The comb actuated mirror does not experience direct electrostatic force, and can typically be thinner for the same radius. In this case, the limiting thickness is determined by the dynamic deformation of the mirror [47].

2.2.4 Tilt Mirror Optical Response

The micromirror must be able to discriminate between a discrete number of optical switch states without excessive optical loss or crosstalk. For the tilt angles used in MEMS beam-steering switches, the number of optical states resolved along each rotational axis can be approximated as

$$M_i = \frac{2\theta_{\max,i}}{\beta} + 1, \tag{2.14}$$

where $\pm\theta_{max,i}$ is the maximum mechanical tilt angle along the i^{th} axis, and β is the half divergence angle of the optical beam.

In $N \times N$ switches, as shown in Figure 2.2, the beam propagates to a second array of mirrors. The physical size of the spring structures, gimbal, and comb fingers surrounding the mirror all contribute to the footprint of a single device, and limit how close adjacent mirrors can be positioned in the array. We assumed plate actuators to be separated by at least twice the mirror height, h_m, to prevent electrical crosstalk. We define the linear mirror fill factor, α_i, as the ratio of the mirror diameter to the mirror pitch along the i^{th} dimension. This value changes with the physical structure of the actuator, and was calculated on a case-by-case basis for each actuator design. Mathematically, α_i scales the first term on the right-hand side of (2.14), so that lower fill factors reduce the number of addressable optical states. We did not include the potential reduction in fill factor from electrical routing because it is highly design dependent. Multilayer electrical routing can increase device density significantly compared to planar routing [48]. We did not include the skew angle of the MEMS array or path length variability in our optimization model. The impacts on tilt angle and insertion loss are a second-order correction to the model, and these impacts are quantified for example design cases in Section 2.3.2 using physical optics modeling in Zemax. We also note that skew can be completely removed by choice of switch geometry, while still using a Fourier lens configuration [49].

The number of optical states for devices with two rotation axes can be approximated by determining the number of states along each principle axis of rotation.

The 2-axis comb actuator has two independent axes, each with its own torsion springs and comb fingers. The maximum tilt angle of the comb actuator traces a rectangle in angular space (for small angles) and the device can resolve $M_{comb} = M_x M_y$ beams, where M_x and M_y are the number of one-dimensional resolvable beams along each axis, given by (2.14). The two axes of the plate actuator have independent torsion springs, but are coupled by a common electrode. The maximum tilt angle of the plate actuator traces an ellipse in angular space, resolving $M_{plate} = \pi M_x M_y / 4$ beams. The distributions of mass and spring constants differ between the two axes, so in general both the resonant frequencies and maximum tilt angles differ for each axis (i.e. $f_{0,x} \neq f_{0,y}$ and $M_x \neq M_y$). Because the maximum response rate of the device is limited by its slowest axis, devices optimized for speed tend to have comparable resonant frequencies (and different angular ranges) along both axes. This leads to an inequality in the number of resolvable optical states between axes, which skews the shape of the mirror array to rectangular instead of square.

We modeled the light emitted by the input single mode fiber as a Gaussian beam parameterized by a nominal wavelength, λ, of 1550nm. The Gaussian approximation of a fiber mode is less accurate far from the optical axis, and more detailed analysis may be necessary for systems where very high extinction ratios are required [50]. Using the Fourier lens OXC geometry (Figure 2.2), we place the waist of the beam at the micromirror. The Gaussian beam is infinite in spatial extent and is clipped at each mirror, resulting in a spatial confinement efficiency, η_{sc}, at the mirror given by

$$\eta_{sc} = 1 - \exp\left(-\frac{2r_m^{\,2}}{w_0^{\,2}}\right), \tag{2.15}$$

where w_0 is the beam waist. The angular distribution of light reflected from the mirror is altered as a result of diffraction from the edges of the mirror, and is no longer a pure Gaussian beam. We calculated the far-field intensity distribution of light reflected from a mirror by convolving the Fourier transforms of the Gaussian field and the mirror aperture. Using the far-field diffraction pattern (Figure 2.5(a)), we defined a nominal angular subtense, β, to distinguish the signal portion of the beam from the surrounding potential crosstalk (Figures 2.5(b) and 2.5(c)). The fraction of power encircled within the

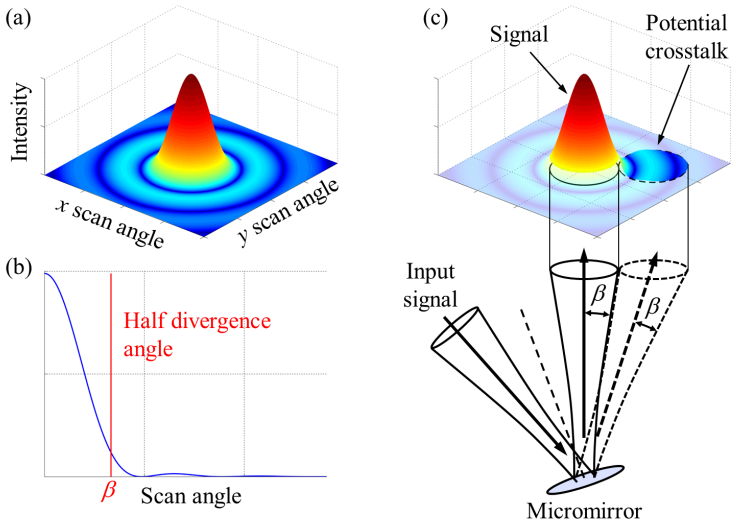

Figure 2.5: Diffraction of light from a MEMS micromirror. (a) Far field angular intensity diffracted from a micromirror. (b) Cross section of diffraction pattern and choice of half divergence angle, β. (c) An input beam diffracts light into signal and crosstalk beams, who's distinguishability is defined in angular space by β.

signal portion defines the angular confinement efficiency, η_{ac}, while the nearest neighbor crosstalk is found by integrating the appropriate region of the surrounding power. Thus, the nominal divergence angle of the beam is related to the mirror radius, beam waist, angular confinement, and crosstalk.

Our approximation of crosstalk using encircled energy was necessary to limit computation time during optimization. This method gives an upper bound on the crosstalk, and becomes more accurate in the limit of high insertion loss, which is where crosstalk becomes a significant concern. A more accurate assessment of crosstalk requires a mode overlap calculation between the fiber and the potential crosstalk signal after it is focused by the corresponding microlens. We perform this detailed analysis for design examples in Section 2.3.2.

For a given confinement efficiency, the overall insertion loss of the switch is driven by the number of stages. We modeled the overall switch throughput efficiency, η_{switch}, as a series of lumped element efficiencies at each micromirror:

$$\eta_{switch} = \eta_f \left(R\eta_{sc}\eta_{ac} \right)^S , \tag{2.16}$$

where η_f is the fiber coupling efficiency at the output (assumed to be 90% based on experimental demonstrations in large port count OXCs [15]), R is the mirror reflectivity (assumed to be 97% for gold at 1550nm), and S is the number of micromirror stages in the switch (see (2.1) and (2.2)).

2.3 Switch Scaling Study

The governing equations outlined in the previous section lay the framework for our scaling analysis of MEMS cross-connects. To assess the accuracy of our first-principles model and the practicality of the results, we perform detailed optical design work on a number of representative solutions predicted by the model.

2.3.1 Optimization and Theoretical Scaling Limits

We implemented the model discussed in Section 2.2 numerically in Matlab. The initial goal was to determine how the resonant frequency of each MEMS device scales with switch port count and optical transmission and crosstalk levels in a conventional OXC. The geometrical form of the device, including the mirror radius and thickness, spring width and length, electrode shape and air gap, as well as the optical beam parameters all constitute a design space which determines the resonant frequency and optical properties of a device.

For a given device, switch port count, insertion loss, and crosstalk, the problem of determining the optimal values of all free design variables is underdetermined. Consequently, we implemented a global search over the design space, with the geometrical form and optical beam parameters as inputs to the algorithm. Although computationally slower than other optimization methods, such a brute-force search is immune to local maxima and does not require assumptions about the optimization space other than its value limits. We bounded the search algorithm on the bottom end by assuming a minimum feature size of 1 micrometer and a minimum beam waist of 3

micrometers (twice the wavelength at 1550 nm). The upper end was bounded by the size of the mirror, and corresponding mechanical structures, necessary to achieve the maximum port count we considered (N = 4,096 ports). We calculated that mirrors larger than 4 mm in diameter had excessive optical performance to meet the maximum port count, and would be unnecessarily slow due to increased inertia. We checked the solutions to ensure that the imposed boundaries did not arbitrarily constrain the design space. We discretized each design variable linearly or logarithmically with sufficiently fine sampling that we saw convergence in the solution.

The electrostatics, mechanics, and optics coupled many of the design variables. To save computation time, we separated the algorithm into an electromechanics component and an optics component which were coupled through the mirror radius. In the electromechanics code, we used the geometrical form parameters of the mirror, springs, and electrodes to compute the resonant frequency using (2.9), (2.11), and (2.12) and the maximum mechanical tilt angle using (2.2)-(2.8) for every realization of each device within the design space. Dual axis devices took into account that the resonant frequency of the outer axis depends on the parameters of the inner axis, and that different spring constants are needed to achieve the same resonant frequency along both axes. The number of unique electromechanical realizations of a single device ranged from 10^5 to 10^7, depending on the number of design variables. The optical portion of the code used the mirror radius, beam waist, and beam divergence angle as inputs to compute the spatial confinement, angular confinement, and crosstalk. We considered roughly 10^5 unique optical configurations.

For a specified switch topology and number of ports, we used (2.16) to calculate the insertion loss of the switch for each optical realization. We then eliminated any optical realizations which did not satisfy specified levels of insertion loss and crosstalk. Next, we used (2.14) to eliminate a portion of the electromechanical realizations based on the number of resolvable optical states required. Finally, we sorted the remaining device realizations by driven resonant frequency to determine the fastest device capable of meeting the specified switch parameters. This process was repeated for different devices, port counts, and optical transmission and crosstalk parameters.

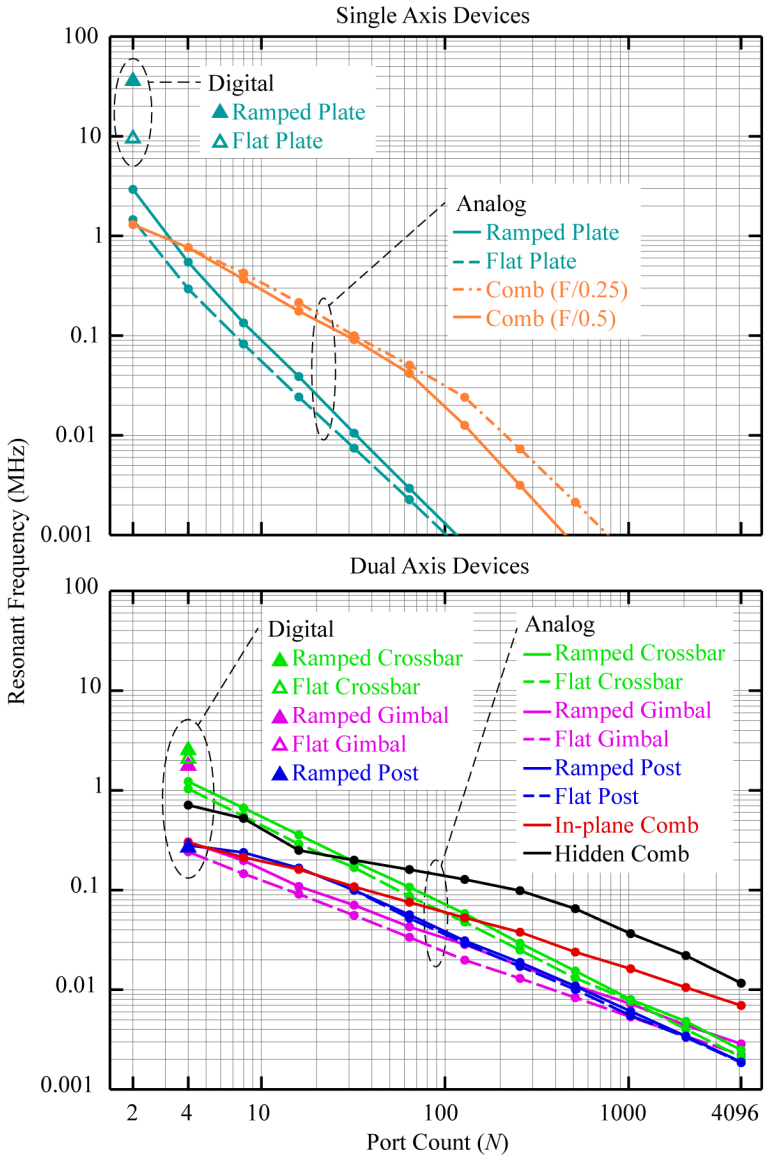

Figure 2.6: Device resonant frequency vs. switch port count. (upper) Single-axis and (lower) dual-axis devices (see Figure 2.3) arranged in a conventional $N \times N$ free-space OXC (see Figure 2.2), constrained for insertion loss better than 3dB and crosstalk better than -20dB. The tilt angle of the 1-axis comb drive must be limited for compatibility with a reasonable F/# Fourier lens.

Figure 2.6 shows how each device's resonant frequency scales with switch port count for a conventional $N \times N$ 3D-MEMS OXC (Figure 2.2). The optical performance was constrained to have 3 dB insertion loss using (2.16) and less than -20 dB crosstalk. We found that insertion loss imposed the stronger constraint, and that all optimal designs had approximately 3 dB loss and much less than -20 dB crosstalk. There is a clear

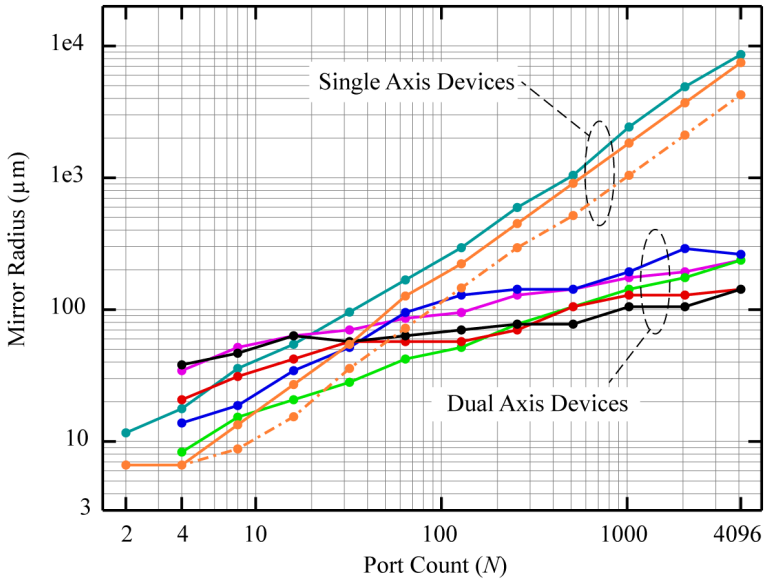

Figure 2.7: Micromirror radius vs. switch port count. Shown for the devices in Figure 2.6.

tradeoff between switching speed and port count, which can be understood through two functional relationships. First, the tilt range of a mirror is inversely proportional to resonant frequency through the spring constant in (2.9) and directly proportional to port count in (2.14). Second, the mirror radius is inversely proportional to resonant frequency through rotational inertia and directly proportional to port count through diffraction and beam divergence angle.

From Figure 2.6, ramped plate actuators always outperform parallel (flat) plate actuators, and digitally operated plate actuators always outperform their analog counterparts. The latter can be understood by considering the additional optical steering range gained by allowing the mirror to snap down to the substrate, and the independence of resonant frequency and tilt angle when driving past pull-in. The drawback of digital devices is that they do not scale beyond a few ports in a conventional OXC. 1-axis devices are faster than two axis devices in the small port count regime, where lower inertia makes up for the reduced dimensionality of tilt.

One interesting result seen in Figure 2.6 is that within the single- and dual-axis subgroups, the fastest actuator changes as a function of port count. Focusing on single-axis devices, the plate actuator operated digitally has 10× the resonant frequency of the

next fastest device. The comb actuator is faster than the plate actuator for larger port counts, when, as a consequence of diffraction, the mirror radius has become sufficiently large to overcome the inertial impact of the comb fingers. The 1-axis comb drive naturally optimizes to large tilt angles (>20°), so we imposed restrictions on the tilt angle to maintain compatibility with the f-number (F/# = focal length divided by full aperture) of the Fourier lens. F/0.25 may be impractical to achieve due to lens aberrations, but shows the theoretically allowed scaling limit. F/0.5 may be achievable with an aspheric curved mirror. Examining the dual-axis devices, we see that more complex designs (hidden comb and crossbar) have better performance than simpler designs. The scaling trends of resonant frequency with port count are largely explained by those of the mirror aperture. Figure 2.7 shows the corresponding mirror apertures for each device necessary

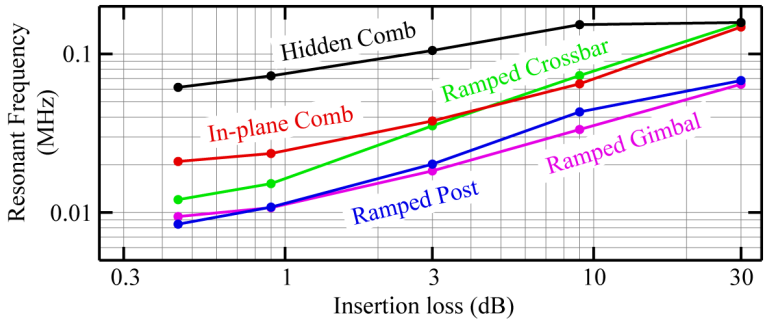

Figure 2.8: Device resonant frequency vs. switch insertion loss. Port count was set to 256. Crosstalk was constrained to be better than -10dB. Similar trends were seen for all port counts considered.

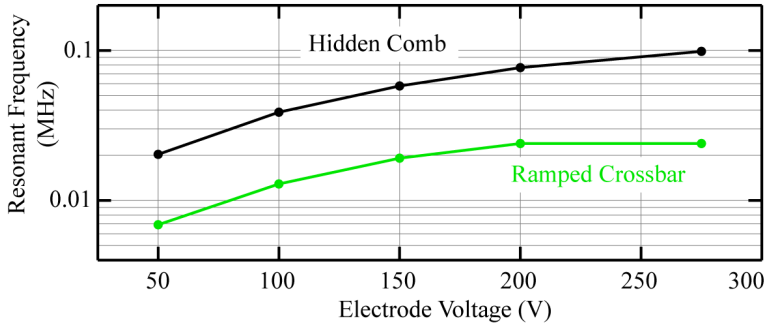

Figure 2.9: Device resonant frequency vs. electrode drive voltage. Port count was set to 256. Insertion loss was constrained to be ≤ 3dB and crosstalk better than -20dB. Similar trends were seen for all port counts considered.

to achieve the performance shown in Figure 2.6. We see that single axis devices require larer mirrors than dual axis devcies to reach high port count, accouting for the different scaling trends in resonant frequency.

We used our model to investigate the tradeoff between optical performance parameters (insertion loss and crosstalk) and switching speed in the $N \times N$ OXC topology to determine the speed increase that can be gained by surrendering optical performance. Physically, the rotational inertia can be reduced, and the resonant frequency increased, by shrinking the size of the mirror. The smaller mirror, however, spatially clips more of the optical beam and diffracts more light into adjacent ports. Our model showed that a significant increase in switching speed cannot be achieved by a reasonable sacrifice in optical performance. Figure 2.8 shows that in the best case, while maintaining a crosstalk of better than -10 dB in a 256 port switch, the resonant frequency (switching speed) of a device can only be improved by about 3×, and requires 27 dB excess insertion loss. Similar scaling trends hold for all port counts considered.

Finally, because the finite slew rate of the MEMS driver can limit the speed of the device, we investigated how the resonant frequency varies as a function of electrode voltage. Figure 2.9 shows the scaling trends for two of the highest performing devices in a 256 port OXC.

2.3.2 Detailed Analysis & Comparison to Commercial Switches

To assess the accuracy of our optimization model and the practicality of the optimal systems, we extracted the actuator and switch parameters from our model for the designs shown in Figure 2.6. We used Zemax to construct 3D switch models to account for the skew of the MEMS arrays, the variability of optical path length, and the aberrations associated with the microlenses and Fourier lens. Figure 2.10 shows the micromirror and switch system for a 132 port OXC based on the in-plane 2-axis comb drive. We tiled the mirrors into an array, accounting for the asymmetric fill factors and tilt angles along each dimension, then used physical optics propagation to model the

single mode fiber coupling and crosstalk for both an ideal (paraxial) and biconvex silicon Fourier lens. The results for this and a few other selected designs are summarized in Table 2.2. These devices all operate at 275 V. Based on the Zemax results, we found that our model accurately predicted optical performance for moderate port counts, and still maintained reasonable accuracy at extreme port counts. Our approximation of crosstalk in Section 2.2 was conservative, and did not impose an unintended constraint during optimization. We used a reflective Fourier mirror to achieve the F/0.5 requirement for the 256 port switch using the 1-axis comb drive. Corrections to tilt angle to account for array skew and the non-paraxial Fourier lens were less than 10% of the model output value in all cases. Using an optimized triplet Fourier lens instead of a simple biconvex singlet could further improve performance, but is beyond the scope of this analysis.

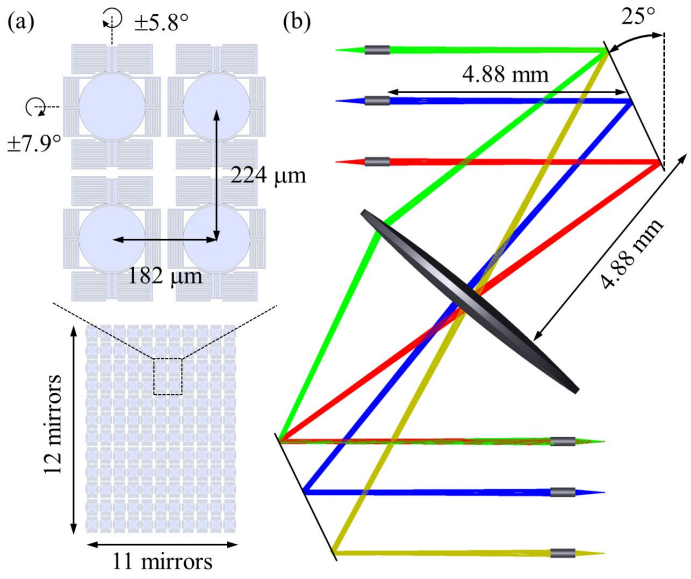

Figure 2.10: Detailed model of a 132-port cross-connect. (a) Optimal 2-axis in-plane comb actuator for a 132 port switch, arranged in an array accounting for the asymmetric fill factors and tilt angles. (b) Zemax model of the corresponding system including skew angle, microlenses, and Fourier lens.

Table 2.2: Selected optimal modeled actuator and switch parameters

	Actuator Parameters													Switch Parameters						
Type	f (kHz)	θ_{mech} (± °)	r_m (μm)	t_m (μm)	h_m (μm)	Fill factor	k (Nm/rad)	l_s (μm)	w_s (μm)	t_s (μm)	N_f	l_f (μm)	t_f (μm)	Ports [layout]	Skew (°)	z_R (mm)	Insertion loss (dB) Parax.	Insertion loss (dB) Biconv.	Crosstalk (dB) Parax.	Crosstalk (dB) Biconv.
2-axis In-plane comb	52.8	7.9 (5.8)	57	6.4	-	0.69 (0.56)	1.5×10^{-8} (6.6×10^{-8})	24 (41)	1.1 (1.7)	6.4 (13)	6 (10)	46 (63)	6.4	132 [11×12]	25	4.9	2.4	3.3	-39	-39
2-axis Hidden comb	11.6	10.9 (11.6)	143	1	55	0.95 (0.95)	6.7×10^{-9} (1.2×10^{-8})	123 (123)	1.6 (1.5)	5.2 (10)	30 (30)	18 (22)	5.2	4154 [62×67]	35	23.4	3.0	4.2	-29	-27
2-axis Hidden crossbar	2.5	10.2 (10.2)	237	2.6	105	0.70 (0.70)	6.7×10^{-9} (6.7×10^{-9})	237 (237)	1.1 (1.1)	26 (26)	-	-	-	4096 [64×64]	23	85.6	2.9	3.5	-34	-33
1-axis Comb (F/0.5)	3.2	12.8	449	22	-	1	6.7×10^{-7}	12	2	22	236	126	22	256 [1×256]	0.5	240	3.6	5.2	-31	-26

Table showing design parameters extracted from the optimization model, at a few interesting points shown in Figure 2.6. The insertion loss and crosstalk are calculated using physical optics propagation in Zemax for a paraxial Fourier lens and a biconvex singlet Fourier lens. For 2-axis actuators, the value for the inner axis is indicated without parentheses and the value for the outer axis is indicated with parentheses. The layout of the micromirror array is indicated by brackets.

The idealized switch structures in this model lead to a theoretically achievable performance which is not necessarily compatible with a practical switch constrained by the many design factors required for a manufacturable and cost-effective product. However, it is useful to discuss these factors to understand how a theoretical design can be translated into reality. Consider the optimized 132 port switch shown in Figure 2.10. The predicted 3 dB insertion loss requires perfect alignment of the collimator and fiber arrays. To account for misalignments between the fiber array, microlens array, and MEMS array while keeping insertion loss low, we must redesign for low theoretical insertion loss and let the mirror aperture grow. Redesigning for 0.2 dB nominal loss requires the mirror aperture to increase by 2×, and the resonant frequency is reduced by 2×. To reduce cost, the Fourier lens may be omitted, which requires the tilt range to increase by 2× in each dimension. Because the tilt range is already large, we can instead double the distance between mirror arrays and double the mirror aperture. Also, because the beam is no longer focused onto the mirror, the mirror aperture must increase by 1.5× and the mirror pitch must increase by 2×, again requiring larger tilt, or a larger mirror. The omission of the Fourier lens cumulatively increases mirror aperture by 4× and reduces resonant frequency by 4×. Next, to maximize the reliability of drive electronics, the electrode voltage might be reduced to 150 V. This requires weaker springs to maintain the same tilt range, and reduces the resonant frequency by 2×. To account for imperfect fiber and MEMS array yield, we can add redundant elements to the arrays. Assuming an 80% yield for fiber and MEMS arrays, we must increase the number of elements in the array by 50%. Finally, to avoid complicated multilayer electrical routing, we might use planar routing and increase the pitch between mirror elements. To account for yield and planar routing, we may let the mirror radius increase by 1.5× and reduce the spring constant further, reducing the resonant frequency by 2×. Accouting for all these factors, the resonant frequency is reduced to 1.6 kHz and the mirror radius is increased to 680 μm. Commercial switches are often operated at some fraction ($1/10^{th}$) of the mirror's resonant frequency to allow mirror ringing to subside. This gives a response time of 6.2 ms, which is more comparable to that of commercial switches with ~100 ports. Besides

the changes in switching speed and physical scale, the switch still looks very similar to the one shown in Figure 2.10 (ignoring the Fourier lens), but the cost of manufacture is significantly reduced. Supposing that similar relationships hold for all port counts considered in our study, the relative changes in the theoretical limits calculated here (Figure 2.6 and Figure 2.8) should be reflected in real switches.

2.4 Multistage Switch Architectures

Our scaling study indicates that new overall optical switch architectures will be needed to achieve microsecond-scale switching with the large port count necessary for data center networks, as opposed to simply modifying device structures within existing telecommunications switches. Here we describe two switch architectures to illustrate how multistage topologies can allow better scaling properties. We used the skew ray representation of Gaussian beams [51] to design the switches and physical optics propagation in Zemax to model single mode fiber coupling.

2.4.1 Multiport Wavelength Selective Switch

Telecom multiport wavelength switches typically use digital beam steering and aperture division [52]. Introducing an array of relay lenses located near a digital micromirror array can extend the port count of $1 \times N$ wavelength selective switches while retaining the microsecond-scale reconfiguration rate of two-state micromirrors.

Figure 2.11 illustrates this switch structure. The input signal is spectrally demultiplexed by a reflective dispersive element in the Fourier plane, with the fiber and MEMS arrays located at the focal plane of the Fourier lens. Each wavelength channel is spatially separated in y at the MEMS array and is independently relayed laterally in x by the tilting mirrors. By purposely adding spherical power to the micromirrors (possible by greyscale lithography [53] or stress induced bending [54]) and using an array of reflective micro optics to form a 4-f relay, the beam parameters can be relayed between

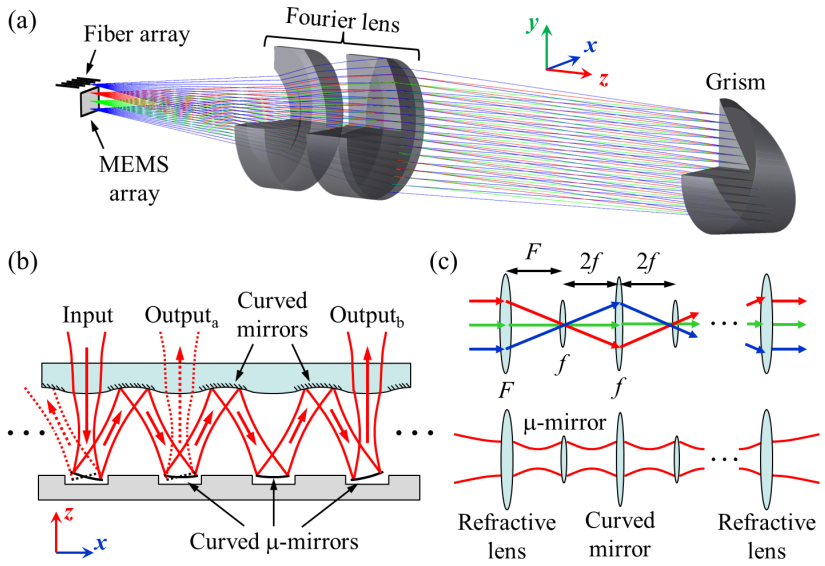

Figure 2.11: Multistage wavelength selective switch. (a) Zemax model showing dispersion along y at the MEMS array. Ray color indicates different wavelength channels. (b) Schematic x-z cross section at the MEMS array showing a single wavelength channel being relayed to the right or left in x using an array of curved mirrors. (c) Unfolded schematic showing the simultaneous relaying of three spectral components of a single channel and the Gaussian beam parameters when $z_R = f$ and $F = 2f$.

mirrors with a minimal number of reflections (Figure 2.11(b)). A refractive microlens at the input of the relay focuses the spectral components of a single wavelength channel onto the micromirror to maintain a wide passband (Figure 2.11(c)). Because the 4-f relay spatially inverts the spectral components of the passband with each pass through the relay, an output port can only be located at every second micromirror.

The size of the relay lenses imposes a spectral separation in y, and to keep the overall system length to a minimum we used a grism (grating-prism combination) with stronger dispersion than a standard grating. We modeled a 1×12 port switch supporting 128 50-GHz-spaced C-band channels in Zemax (Figure 2.11(a)). We assumed all surfaces had a high reflectivity coating, such as Newport DM.8, which is 99.5% reflective up to 45° at 1550nm. Performing Gaussian beam propagation in Zemax, and assuming a Grism efficiency of 80% for a custom blaze angle, we found the worst case insertion loss was 3.1 dB, with a 25 GHz passband with 0.5 dB excess loss.

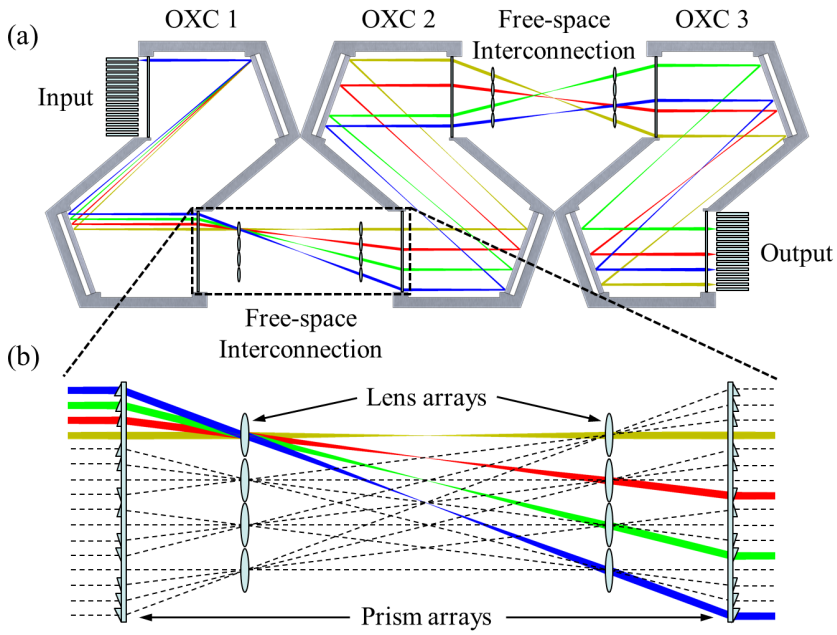

Figure 2.12: Multistage cross-connect. (a) Cross section of three 256 port OXCs, each with 4× reduced scan angle, interconnected with free-space optics. Ray color indicates different optical paths through the system, depending on mirror states. (b) Detail of the passive free-space interconnection with possible connections.

2.4.2 Multistage Optical Cross-connect

In Section 2.3 and Figure 2.6, we observed that the switching speed of an $N \times N$ OXC can be increased by reducing the port count. The second illustrative switch geometry uses free-space optics to interconnect many small port count OXC "sub-switches" in a multistage network to form an $N \times N$ switch which retains the faster reconfiguration rate of the small sub-switches.

Figure 2.12 shows an illustration of the switch structure. The drawing seems to show a cascade of three fully interconnected $N \times N$ switches. In fact, the tilt range of every micromirror in each OXC switch has been reduced, sacrificing full connectivity within a single OXC, but allowing a faster reconfiguration rate through the inverse relationship between tilt range and resonant frequency. Full, non-blocking connectivity between all ports is regained by interconnecting three active switching stages in a Clos network [55], provided the stages are interconnected with a suitable port-mapping

structure. This could be done with fiber cabling, but would triple the insertion loss. Instead, the necessary port mapping between stages can be accomplished with an optical transpose interconnection implemented by relay imaging [56], reducing optical loss and complexity compared to fiber optic connections. Figure 2.12(b) shows the passive optics required, using two prism arrays and two lens arrays to redirect and relay light output from the first switch to the input of the next switch. For the 2-axis hidden crossbar device, our model predicts a 4× reduction in tilt angle in a 256 port switch increases the switching speed by 3×. The reduction in tilt allows a higher resonant frequency due to stiffer springs, but the gain in speed is limited because stiffer springs require a thicker mirror to prevent bending of the mirror under actuation force. We performed Gaussian beam analysis in Zemax and found the worst-case insertion loss after 3 stages was 7.7 dB accounting for the accrued path length differences.

A more substantial increase in speed can be achieved by reducing the mirror aperture, as this increases resonant frequency by decreasing inertia and does not require an increase in mirror thickness. Our model indicates a 4× reduction in mirror radius can increase switching speed by 10× for a 256 port switch. However, the smaller micromirror aperture requires a larger beam divergence to maintain high spatial confinement efficiency at the mirrors. The increased beam divergence can be accommodated by adding relay optics between collimators and MEMS arrays and between MEMS arrays in the system shown in Figure 2.12, providing a more substantial increase in switching speed at the cost of increased optical complexity.

2.5 Discussion

This chapter quantified the theoretical relationships between speed, port count, and optical transmission for MEMS beam-steering cross-connects. Based on our prediction that conventional telecom switches may not scale far beyond their current port counts and switching speeds without significant increase in optical loss or cost of manufacture, we suggested two multistage switch architectures that help extend the performance of MEMS tilt mirror technology. Most important, however, is the

understanding of physical layer parameters governing MEMS beam-steering devices developed in this chapter, as this serves as the theoretical motivation for a novel switch (and ultimately network) architecture studied in Chapters 3 and 4 of this dissertation.

Chapter 2, in part, reprints material as it appears in the paper titled: "Scaling Limits of MEMS Beam-Steering Switches for Data Center Networks," published in the *Journal of Lightwave Technology*, 33(15), pp 3308-3318, 2015, by William M. Mellette and Joseph E. Ford.

Chapter 2, in part, reprints material that has been submitted for publication in a paper titled: "A Scalable, Partially Configurable Optical Switch for Data Center Networks," submitted to the *Journal of Lightwave Technology*, by W. M. Mellette, G. M. Schuster, G. Porter, G. Papen, and J. E. Ford.

Chapter 3

A Scalable, Partially Configurable Optical Selector Switch

This chapter builds upon the results of Chapter 2, which indicated that the conventional optical cross-connect architecture will not gracefully scale to meet the requirements of data center networks. Analysis in Chapter 2 also suggests that while multistage crossbar switches have better scaling properties than single-stage switches, accumulated loss will limit switch performance.

This chapter presents an optical *selector switch* architecture which, through an unconventional approach of relaxing the requirement of arbitrary switch reconfigurability, allows MEMS beam-steering switching elements to scale to microsecond-class response speeds while supporting large port count and low loss switching. The physical architecture of the switch uses pupil-division switching, permitting designs for single-mode or multi-mode fiber optics. The design, fabrication, and experimental characterization is presented for a proof-of-principle prototype using a single MEMS comb-driven micromirror to achieve 150 μs switching of 61 single-mode ports between 4 preconfigured interconnection *matchings*. Here, as in graph theory, the word matching is taken to mean a bipartite mapping between input and output ports, so that every input port is connected to one output port. The scalability of this switch architecture is demonstrated with a detailed optical design of a low-loss 2,048-port selector switch with 20 μs switching time. This chapter focuses on the physical layer

aspects of selector switches, and Chapter 4 follows up with a discussion of network architectures that demonstrate their utility.

3.1 Introduction

Previous work has exposed opportunities for optical switching with microsecond-scale reconfiguration times in data center networks [7], [18], [19]. Unfortunately, as discussed in Chapter 2, there is a fundamental tradeoff between switching speed, insertion loss, and port count in beam-steering cross-connects: to scale to large port count while maintaining low loss requires micromirrors with larger apertures (and inertia) and/or larger tilting ranges (requiring softer torsion springs), both of which reduce the response speed of the switch [57]. Multistage beam-steering switches can achieve faster response times, but inevitably accumulate loss and crosstalk from cascaded switching stages in order to realize large port counts.

Here, we investigate a novel optical *selector switch* architecture which forgoes non-blocking crossbar configurability, instead enabling rapid selection between a relatively small set of preconfigured interconnection matchings. This concept can be implemented in multiple switching technologies, and previous work demonstrated a similar concept using wavelength switching for fast selection of interconnection patterns recorded as volume holograms [58]. However, the change to a selector switch architecture allows MEMS beam-steering micromirrors to scale to microsecond response speeds while supporting a large number of ports and low-loss switching between the broadband single- or even multi-mode transceivers used in data center networks. Chapter 4 examines the network architecture aspects of selector switches.

The chapter is organized as follows. We discuss the basic switch architecture, applications, and pupil-division switching in Section 3.2. In Section 3.3 we show the design of a 61-port proof-of-principle prototype switch based on commercial-off-the-shelf (COTS) optical components, which we fabricate and characterize in Section 3.4. In Section 3.5 we present the design of a low-loss 2,048-port switch with a 20 μs response time which uses a custom MEMS device and micro-optic port matching structures.

3.2 Selector Switch Architecture

The proposed *selector switch* differs from a conventional optical cross-connect in both its architecture and its basic optical switching principle.

3.2.1 Partial Configurability

Instead of implementing all $N!$ possible port matchings of an $N \times N$-port crossbar, the selector switch selects between a small subset $k << N!$ of these matchings. In this sense, the selector switch can be thought of as *partially configurable*. While many interconnection networks have been designed to leverage the arbitrary configurability of crossbars, partially configurable switches can be used to implement a number of useful network topologies. For instance, a network with full connectivity can be constructed from partially configurable circuit switches. The set of shuffle-equivalent network topologies (e.g. Banyan, Perfect Shuffle, Crossover), are typically implemented in space as multistage interconnection networks with $\log_2 N$ stages [59]. These networks provide full connectivity between N ports using a minimum number of connections. The network diameter (number of *hops* data makes as it traverses the network) is $\log_2 N$ in these networks. A partially configurable switch can realize these network topologies by multiplexing in time (rather than space), cycling through a set of $k = \log_2 N$ port matchings. Other network topologies with $k = O(N)$ port matchings [60] can also be constructed from partially configurable switches. Any partially configurable network design will have tradeoffs between the number of switches, number of port matchings, and the network diameter. This network-level design space is covered in more detail in Chapter 4.

To ground our discussion and establish a starting point in the design space of partially configurable switches, we focus on selector switches with $k = \log_2 N$ port matchings. This configuration provides a balance between the hardware complexity ($\log_2 N$ physical port matchings) and network diameter (at most $\log_2 N$ hops). There are a

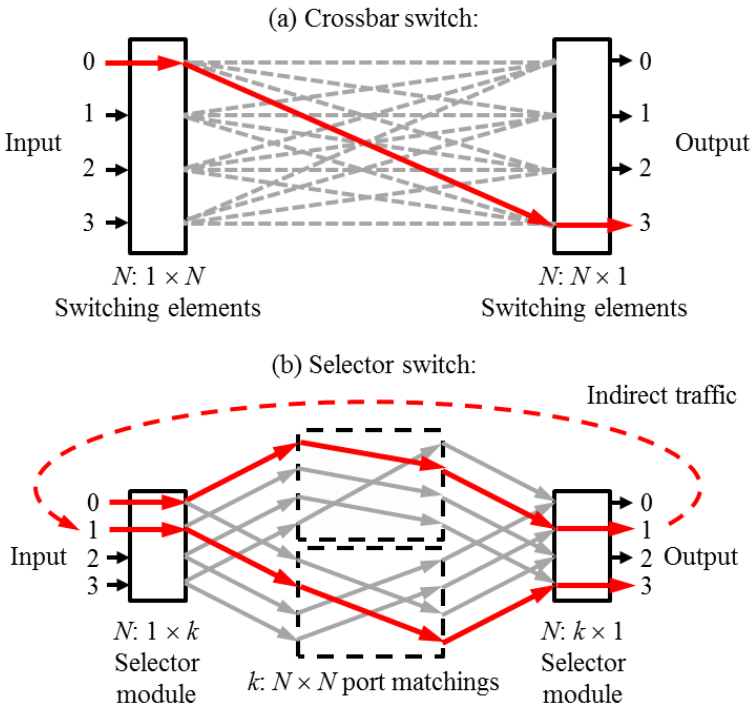

Figure 3.1: Crossbar and selector switch architectures. A crossbar connects any two ports in a single hop through the switch using $1 \times N$ switching elements. A selector switch uses $1 \times k$ switching elements (here $k = \log_2 N$) to select amongst k port matchings. With $k = \log_2 N$ port matchings, data passes through the switch up to $\log_2 N$ times. For example, two hops are required to send from node 0 to 3, with data passing through intermediate node 1.

number of equivalent sets of port matchings with logarithmic network diameter [59]; a simple example is to form sets of matchings from port p to ports $p + 2^0, 2^1, \ldots, 2^{\log_2 N - 1}$ modulo N, where p is indexed from 0 (shown for $N = 4$ in Figure 3.1(b)) [61]. With only $\log_2 N$ matchings, data will generally traverse the switch multiple times, but will be electronically forwarded at intermediate terminal nodes between each optical hop.

Figure 3.1 shows the optical crossbar and selector switch architectures, each requiring two stages of $1 \times N$ and $1 \times \log_2 N$ switching elements, respectively. When implemented with MEMS micromirror switching elements, each micromirror in the selector switch needs to resolve $k = \log_2 N$ optical states, as opposed to $k = N$ optical states for the crossbar. This reduction in the number of optical states significantly reduces the aperture and tilt requirements of the micromirror, allowing it to be redesigned for

higher speed operation. Previous work quantified the theoretical limits of switching speed for single- and dual-axis MEMS micromirror actuators based on the number of resolvable optical states [57], indicating a significant reduction in switching speed is possible by reducing the number of resolvable optical states by a logarithmic factor. Sections 3.4 and 3.5 quantify the achievable switching speeds in selector switch designs based on previously demonstrated beam-steering micromirror devices, showing between two and three orders of magnitude improvement over conventional optical cross-connects depending on the specific micromirror design.

In this work, we consider selector switches with fixed port matchings which are implemented with either low-loss micro-optic or fiber optic interconnections. Alternatively, the switch could be designed to quickly select between k slowly-reconfigurable interconnection patterns by replacing the hard-wired port matchings in Figure 3.1(b) with crossbar switches. This would enable arbitrary configurability while reducing the loss-of-light time during switching, but would increase hardware cost and insertion loss.

The primary component of a selector switch is the *selector module* (see Figure 3.1(b)). The module can be designed with N $1\times k$ *individually-switched* elements, allowing signals from each port to be routed through an independently selected port matching, or as a monolithic $1\times k$ *gang-switched* element which simultaneously selects one matching for all ports. An individually-switched selector module provides greater flexibility because it can select from and combine different port matchings to expand the effective set of selectable matchings, but requires a micromirror array with individually-controllable mirrors. A gang-switched selector module, on the other hand, is less flexible in that it can only select from the $\log_2 N$ hard-wired matchings, but is simpler to control and less expensive to implement because it requires only a single micromirror and control signal. A prototype gang-switched selector module is designed in Section 3.3 and characterized in Section 3.4. Section 3.5 discusses a design compatible with individual switching.

3.2.2 Pupil-division Switching

Conventional cross-connects typically employ a non-imaging layout using two micromirror arrays to steer essentially collimated beams through a freespace volume where beams may intersect (Figure 3.2(a)). Arbitrary bijective port matchings are possible because the path of each beam is defined by a pair of dedicated micromirrors. An alternate design uses a Fourier lens between micromirror arrays to achieve $\sqrt{2}$ smaller beam diameter at the micromirrors, but operates under the same principle as above [8].

We designed the selector module based on a radically different optical configuration, incorporating relay imaging and pupil division instead of the collimated beam steering used in current MEMS cross-connects. Figure 3.2(b) shows a schematic cross section of a fiber-coupled selector module, with a 4-f imaging relay and prism array located near the intermediate pupil of the relay. Light from a two dimensional (2-D)

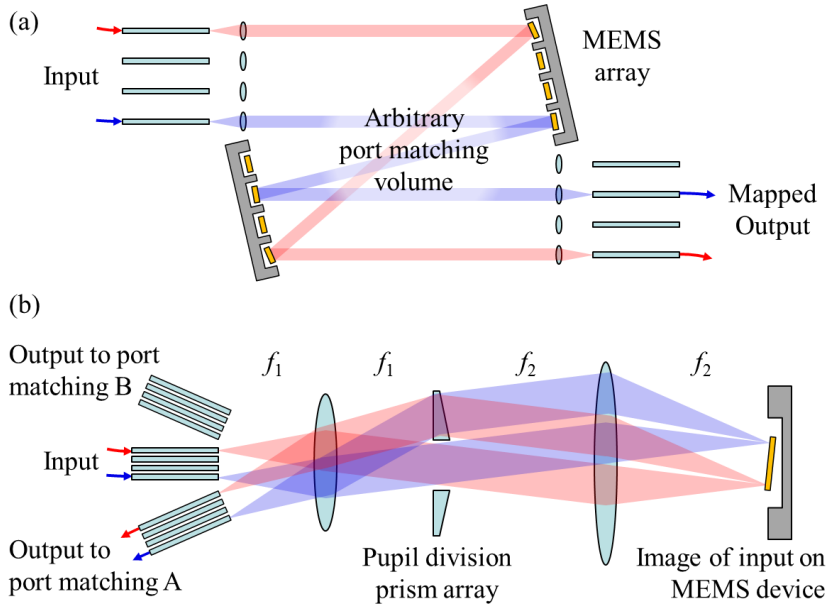

Figure 3.2: Optical crossbar and selector switch physical architectures. (a) Schematic of conventional optical cross-connect. (b) Schematic cross section of fiber-coupled selector module. The MEMS device tilts to select prism apertures which refract the arrayed image to couple into different fiber arrays. Port matchings are implemented externally in (b).

array of input fibers is imaged onto a MEMS tilt-mirror device, which tilts in 2-D to direct reflected light through discrete prism apertures, each of which refracts the output image position to couple into a different output fiber array. Each output fiber array interfaces to an external fiber optic port matching. The telecentricity of the 4-f relay minimizes fiber coupling loss by ensuring that each optical beam couples at normal incidence into its corresponding output fiber core. Because switching occurs by aperture selection in the pupil plane, as opposed to spatially scanning across the fiber array in a conventional cross-connect, the system is more tolerant to angular misalignments of the micromirror. This relaxes the required drive electronics precision and sensitivity to underdamped mirror ringing. The 4-f image relay makes the design compatible with both multimode fiber and space-division multiplexed signals, but we use single mode fiber in the switch designs considered here.

3.3 Fiber-interconnected Selector Module Design

The most straightforward implementation of a selector switch uses fiber optics to realize the desired port matchings between two selector modules (see Figures 3.1(b) and 3.2(b)). In this section, we describe *selector module* designs based on current commercially available fiber arrays and MEMS beam-steering micromirror.

3.3.1 Prototype Design Using Commercial Optics

Maximizing the spatial density of the arrayed input signals maximizes the port count of the selector module. We based our design on a 61-core pitch reducing fiber array [62] commercially available from Chiral Photonics. The array maintains the mode-field diameter and numerical aperture of single mode fiber, but is tapered to position the fiber cores in a 2-D hexagonal array with a 37 μm core pitch. The distance "r_0" from the center to corner channel in the array was 148 μm. We designed for the C band, with a nominal center-band wavelength of λ = 1550nm and a mode waist of w_0 = 5.2 μm at the

fiber. We used a MEMS device with a single micromirror instead of a micromirror array. This allows the micromirror to be surrounded by a large-area (and correspondingly fast) actuator structure, but also means the relayed image of all 61 signals had to be encircled by the micromirror radius "r_m". We allowed the magnification $|M| = f_2 / f_1$ (see Figure 3.2(b)) of the 4-f relay to vary to accommodate different micromirror radii. To first order, the mirror radius must be at least

$$r_m = |M| \cdot r_0 . \tag{3.1}$$

The prism apertures were chosen to be a factor of $\xi = 1.3$ larger than the Gaussian beam mode width at the Fourier plane, yielding 97% power transmission through each prism aperture. The micromirror needs to tilt over a mechanical angular range $\pm \theta_m$ sufficient to select between prism apertures. With 61 ports, we need $\log_2 61 = 5.931 \rightarrow$ at most 6 port matchings to implement the logarithmic matchings discussed in Section 3.2. Hexagonally tiling the 6 prism apertures minimized the required mechanical tilt range of the micromirror:

$$\theta_m \approx \frac{1}{|M|} \cdot \frac{\lambda \xi}{\pi w_0} . \tag{3.2}$$

For a given drive voltage and actuator structure, the switching speed of the micromirror is primarily a function of its radius and tilt range, which respectively determine its mass and torsional stiffness [57]. While equations (3.1) and (3.2) could be used to optimize a micromirror for our system, instead we used them to guide our search for a commercially available micromirror with the fastest response which would meet the system requirements. We chose a USB-powered 2-axis electrostatic comb driven micromirror [63] (Mirrorcle Technologies part #A7M8.1) with r_m = 400 μm, $\theta_m = \pm 4°$, and sub-millisecond response speed (exact settling time depends on the drive signal, and is measured in Section 3.4).

The final design steps were to choose relay lenses and the refraction angle of the prisms. The micromirror parameters allowed relay magnifications between 1.8 and 2.7. We targeted a magnification closer to 2.7 because although it increased system length, it maximized the F/# (focal length divided by clear aperture) and minimized the aberrations

of Lens 2. We chose commercially available doublet lenses with f_1 = 25 mm and f_2 = 60 mm, for a relay magnification of M = 2.4. To avoid mechanical interference, the fiber arrays needed to be separated laterally by at least the end-face diameter of 600 μm. Minimizing the fiber array separation would also minimize the field-of-view requirement of Lens 1 as well as the refraction angle and chromatic dispersion induced by the prisms, which ultimately limits the spectral bandwidth of the switch. We chose commercially available fused silica prisms with a 5° wedge, which required a fiber array spacing of 1 mm and a field of view of ± 2.3° at Lens 1. Finally, in order to minimize the off-axis aberrations in Lens 1, we positioned the prism array near Lens 2 instead of at the Fourier plane so the returning beams would enter the lens closer to the optical axis.

Figure 3.3 shows the prototype selector module with N = 61 input ports and six 61-port output arrays modeled in Zemax optical design software. We modeled the spectral transmission of the selector module, including single-mode fiber coupling loss, with the Zemax physical optics propagation tool. The modeled transmission for the center and edge channels in the array is shown in Figure 3.4, along with the modeled transmission for the custom optical design described in the following section. The

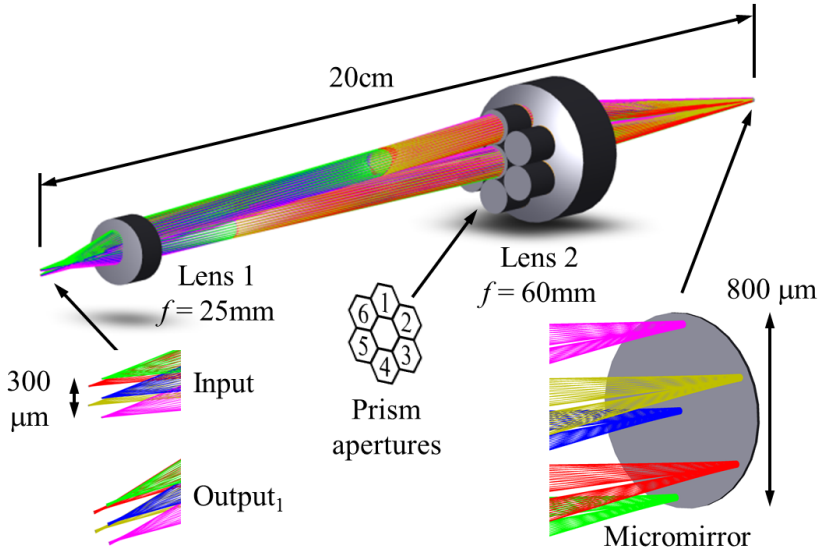

Figure 3.3: Zemax model of prototype selector module. Design is configured for 61-ports and 1×6 selection using commercial doublet lenses and fused silica prisms. Ray colors correspond to the center and edge channels of the 61-core fiber array.

bandwidth is primarily limited by chromatic dispersion from the fused silica prisms. The fiber arrays were aligned to maximize transmission at 1560 nm (corresponding to gain peak of the erbium light source used to characterize the prototype in Section 3.4), but the transmission peak can be tuned to any waveband by refocusing the lenses and translating the fiber arrays. Peak transmission is limited by lens aberrations and reflection losses from the uncoated fiber arrays.

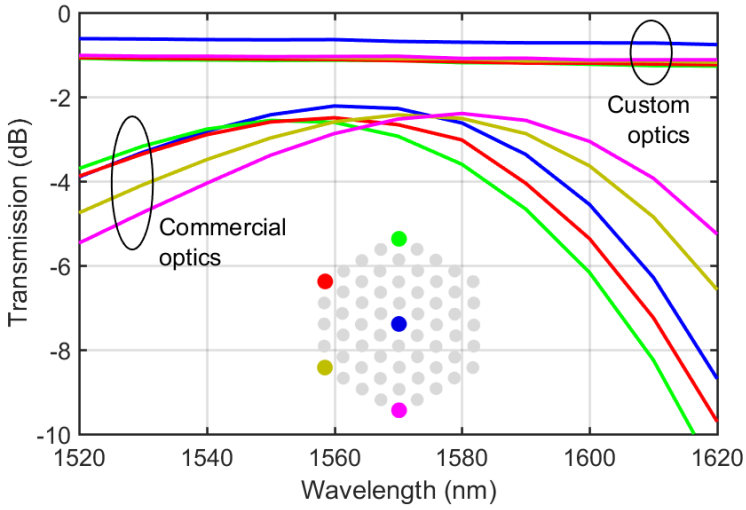

Figure 3.4: Modeled transmission of COTS- and custom-optics prototype. Color inset shows location of channels in 61-core fiber array.

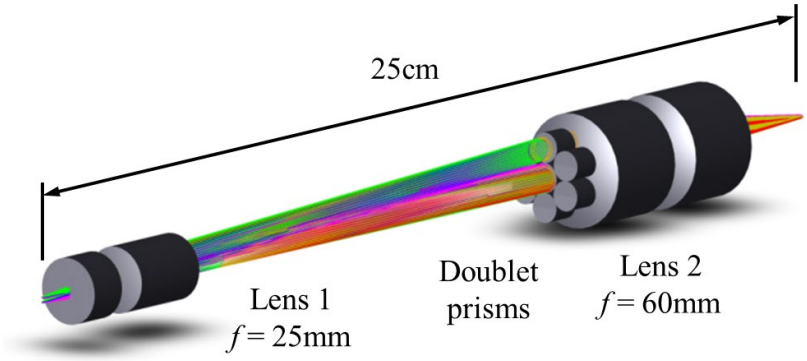

Figure 3.5: Zemax model of custom-optics prototype selector module. 61-port 1×6 selector module has been achromatized with custom triplet lenses and doublet prisms.

3.3.2 Achromatized Prototype Design with Custom Optics

The modeled insertion loss and bandwidth limits from the COTS lens and prims designs (above) would be unacceptable in a practical data center, especially since the transmission of the full selector switch is half that of the selector module, as light must pass through two selector modules in the switch architecture shown in Figure 3.1(b). These losses can be significantly reduced by a minor redesign and customized lenses and coatings.

The single-glass prisms were the dominant source of chromatic dispersion in the commercial optics design. We designed a doublet prism using standard crown and flint glasses (Calcium Fluoride and N-BASF64) to provide the same refraction angle but with negligible dispersion over the C-band. The COTS doublet lenses also contributed insertion loss due to aberrations. We designed a pair of triplet lenses in Zemax using standard glasses with diffraction-limited performance over the C-band and the field of view required by the system. We assumed all refractive surfaces (including the fiber arrays) were coated with commercially-available antireflection coatings and the MEMS mirror was coated with gold (97% reflective). The Zemax model of the achromatized selector module is shown in Figure 3.5 and the transmission is shown in Figure 3.4. The modeled transmission is above -1 dB and substantially flat over >100 nm, which would yield an excellent overall switch transmission of greater than -2 dB.

3.4 Prototype Fabrication and Characterization

3.4.1 Optomechanical Assembly

We fabricated the 61-port selector module designed in Section 3.3.1 based on COTS optics, and the assembled prototype selector module is shown in Figure 3.6(a). The main fabrication challenge was ensuring precise mechanical alignment of the fiber arrays to minimize insertion loss. Positional alignment is the primary driver of fiber

coupling efficiency, requiring micrometer accuracy, while angular misalignment (tip/tilt) of up to a few degrees has little impact on coupling [64]. The end faces of the fiber arrays needed to be brought to within one millimeter of contact while avoiding mechanical interference (Figures 3.6(b) and 3.6(c)). To accomplish this, we machined custom aluminum mounts. The central "input" fiber array was attached to a mount fixed to the optical breadboard. The "output" fiber arrays were mounted to goniometers attached to 3-axis roller bearing translation stages. This provided the one rotational and three linear degrees of freedom required to align the output arrays to the fixed input array. Adjustment for tip and tilt of the arrays was not needed.

We fabricated a custom prism array for the switch (Figure 3.6(d)) by dicing sections from commercially available antireflection coated fused silica wedges via a diamond saw to form the prism facets. These facets were arranged and bonded to a fused silica flat with optical epoxy to create the hexagonally-tiled array. As described in Section 3.3.1, we used commercially available fiber arrays, lenses, and MEMS micromirror (Figure 3.6(e)).

In addition to the input array, we populated four of the six available outputs with fiber arrays. The loss in the interconnection matchings can be minimized by fusion-splicing the output fibers to the required patterns, but for convenience our prototype used bulkhead connectors for both input and output paths. The left wall of the enclosure provides the 61-fiber patch panel interface to the input array, and the right wall holds the four 61-fiber interfaces to the output arrays. Only four arrays were populated because this provided enough port matchings for integration into a 16 server network testbed. The four output arrays were sufficient to characterize all important aspects of the prototype.

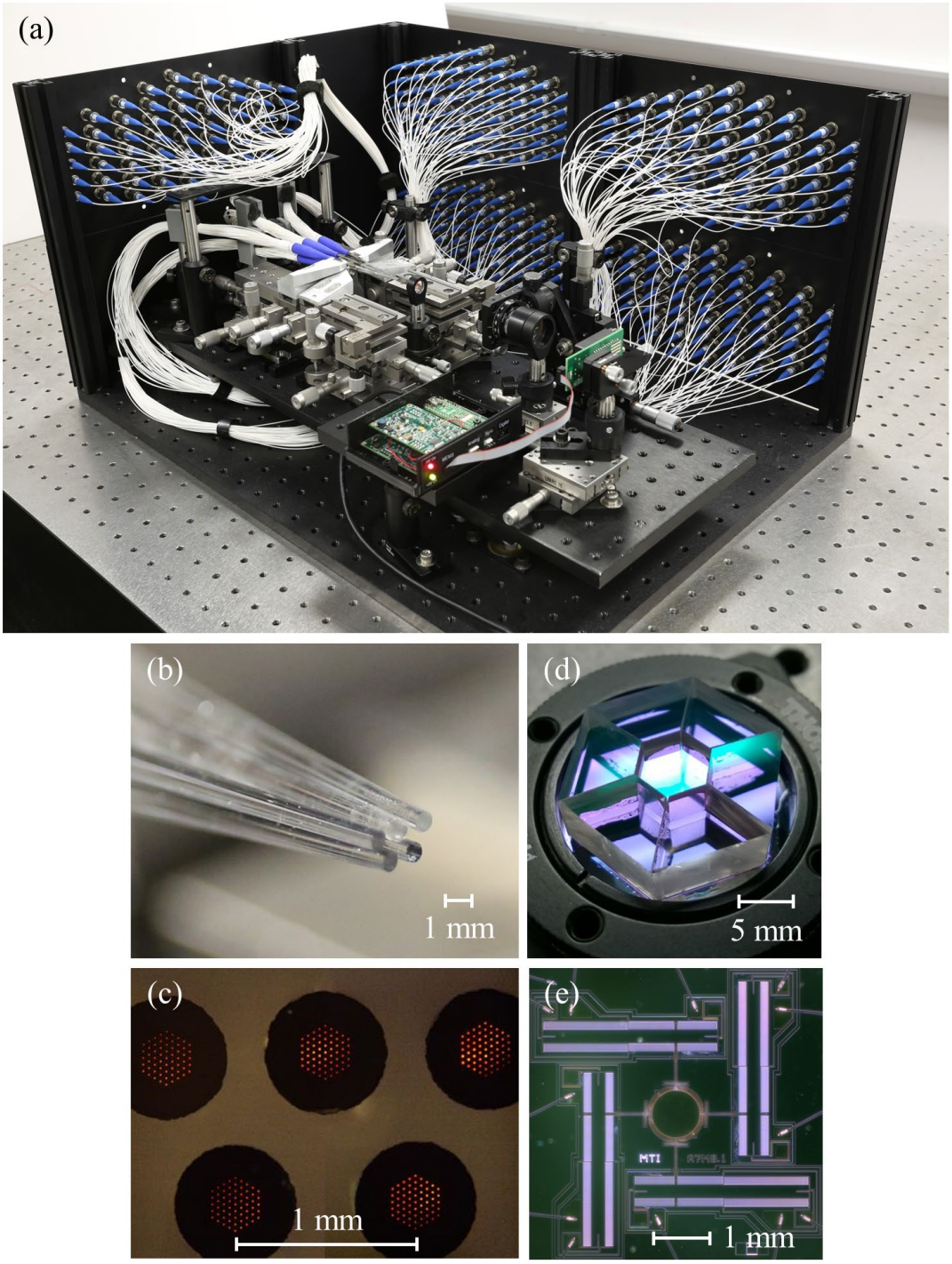

Figure 3.6: Fabricated prototype selector module. (a) Overall selector module. (b) Close-up of fiber arrays. (c) Microscope image of fiber array end faces. (d) Custom antireflection-coated fused silica prism array. (e) Micromirror with large-area comb drive actuators.

3.4.2 Characterization

An erbium amplified spontaneous emission (ASE) light source was used to measure the transmission spectrum of the selector module. The fiber arrays were aligned to maximize transmission at the 1560 nm gain peak of the source. Figure 3.7 shows the measured transmission spectra for all 61 channels for each of the four output fiber arrays. The passband can be shifted arbitrarily by a simple realignment of the system. The measurements are generally in good agreement with the modeled transmission shown in Figure 3.4 (also aligned for maximum transmission at 1560 nm). Array 2 has the most uniform transmission across all channels, while there is the most variance between the channels of array 4. The transmission variance between arrays and between channels within each array was attributed to irregularities in the as-fabricated fiber core positions in the arrays. Modeling showed that core pitch irregularities of ± 2 μm can reduce coupling by up to 2 dB and/or shift the transmission spectrum by 35 nm, accounting for the observed variability in transmission. Using a laser diode source with a mechanical polarization rotator we confirmed that the selector module is polarization insensitive, with a loss variation of less than 0.01 dB.

Next, using the same ASE light source, the intra- and inter-array crosstalk of the selector module was measured. The best and worst case nearest neighbor crosstalk was -40 dB at the center of the array, and -30 dB for channels at the edge of the array. The fiber array manufacturer quotes -35 dB crosstalk between adjacent fiber cores, and given that light propagates through two arrays in the selector module, at least -32 dB crosstalk is expected due to the fiber arrays themselves. We measured a worst case inter-array crosstalk of -50 dB.

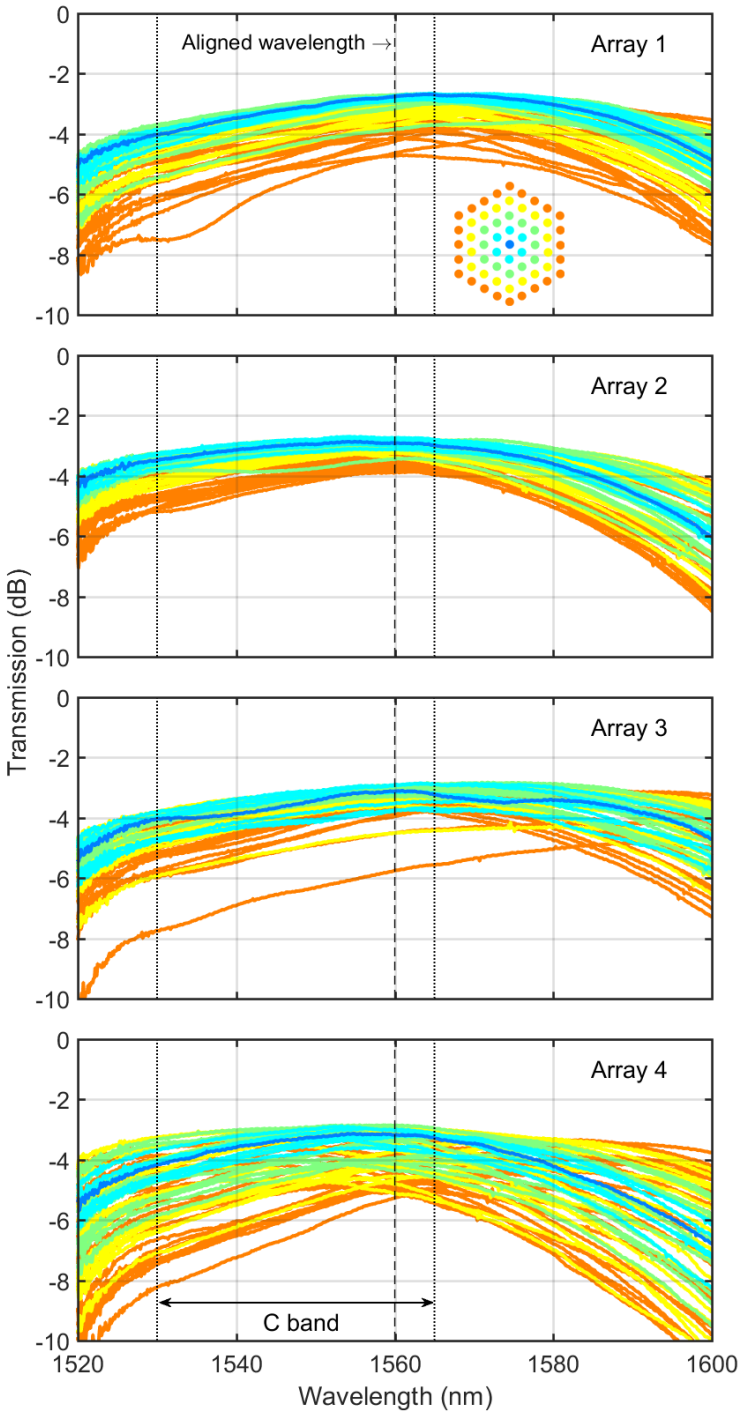

Figure 3.7: Measured transmission spectrum of prototype selector module. Transmission of all ports is shown. Arrays were aligned for maximum transmission at 1560 nm. The C band is shown for bandwidth reference. Color inset shows location of channels in 61-core fiber array.

The response times of the switch measured for each switch transition are shown in Figure 3.8. The voltage waveforms were digitally filtered with the inverse transfer function of the micromirror to provide fast switching while suppressing mechanical

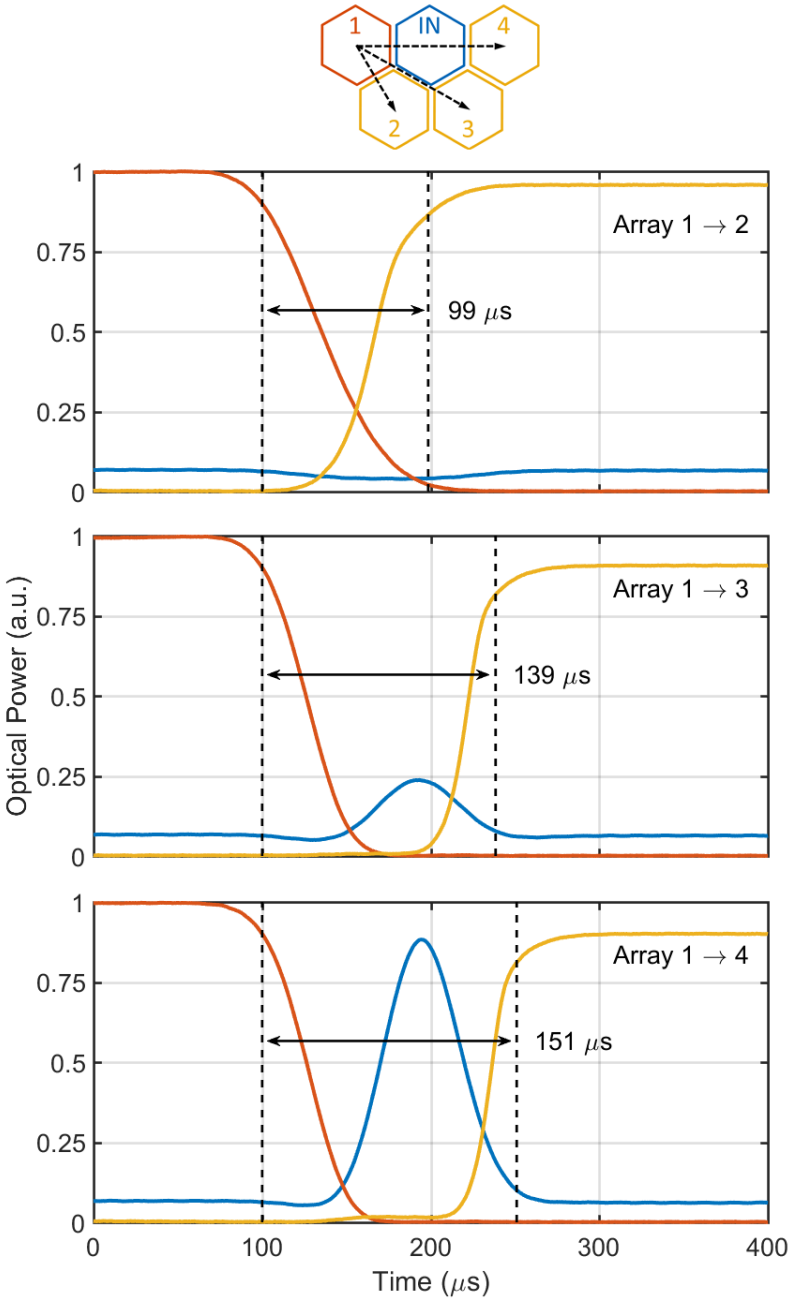

Figure 3.8: Measured switch time of prototype selector module. (Top) illustration of prism apertures with beam trajectories for three representative switching transitions, and (lower) temporal response of switch, with times measured from 90% to 90% power.

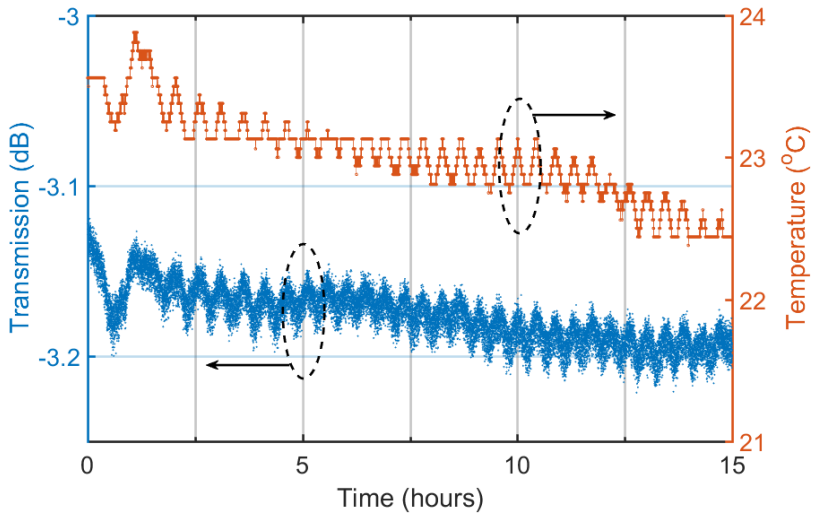

Figure 3.9: Stability of prototype selector module. Measured under laboratory conditions. Loss varies by 0.05 dB / °C, or 0.1 dB / °C for the full switch.

overshoot. The longest switching time of 151 μs (90% – 90% optical power) occurred while moving the micromirror over its longest angular travel between arrays 1 and 4. This is two orders of magnitude faster than commercial MEMS optical cross-connects with comparable port count. The uncoated fiber/air interface accounts for the return loss during steady state, and can be suppressed with antireflection coatings on the fiber array surface. While the mirror is in motion, it can return a significant fraction of the signal power to the transmitter for certain switch transitions (e.g. 1 → 4). In our use of the switch in a testbed with commercial datacom transceivers, we found this had no effect on data transmission because no data was communicated during switching, and also because any spurious oscillations in the laser cavity subsided in less than a single 100 ps bit interval.

Finally, we measured the optomechanical stability of the prototype. Figure 3.9 shows the switch transmission and ambient laboratory temperature over a 15 hour period with no active adjustment. Based on the observed correlation, the selector module transmission varies by 0.05 dB / °C, or 0.1 dB / °C for the full switch. Even lower temperature dependence could be achieved by considering the coefficients of thermal expansion in design of the optomechanical package.

3.5 Freespace-interconnected Selector Switch Design

The proof-of-principle prototype described in the previous section established feasibility of the fast pupil-switched optics, but the fiber-based interconnects would present a significant cabling cost for larger switches because the number of interconnection fibers scales as the product of the port count and number of port matchings in the switch. For example, using logarithmic port matchings in a 2,048 port switch would require 2,048 fibers for each of the 11 port matchings, totaling 22,528 interconnection fibers. To circumvent this issue, we describe a design using freespace micro-optics to integrate the port matchings into a single compact optical assembly which combines the entire selector switch diagrammed in Figure 3.1(b).

3.5.1 Monolithic Switch Assembly

Figure 3.10 shows a cross sectional illustration of the switch layout with two different light paths through the system. Like the fiber-interconnected switch in Figure 3.2(b), this design is also based on 4-f imaging relays and pupil-division switching. However, a number of differences allow this design to incorporate both selector module stages and port matchings into a single assembly. First, this design shares a single fiber array with half the fibers as inputs and half as outputs. Instead of a reduced pitch array, it uses a 2-D fiber array with corresponding microlenses, similar to those employed by cross-connect switches [15]. 2-D arrays with core pitches as low as 170μm and up to 4,096 elements are available from commercial suppliers [65]. The microlenses are designed so they: 1) form a larger beam waist at their output (plane A) to increase signal density, and 2) have a focal point at the fiber face to map any positional misalignment of output beams into angular misalignment at the fiber cores to increase tolerances elsewhere in the system. The beam waists are relayed with demagnification from A onto a micromirror array at B. Each micromirror tilts to direct its beam into one of several 4-f relays, which are defined by lens apertures at C. The beam waists are relayed from plane

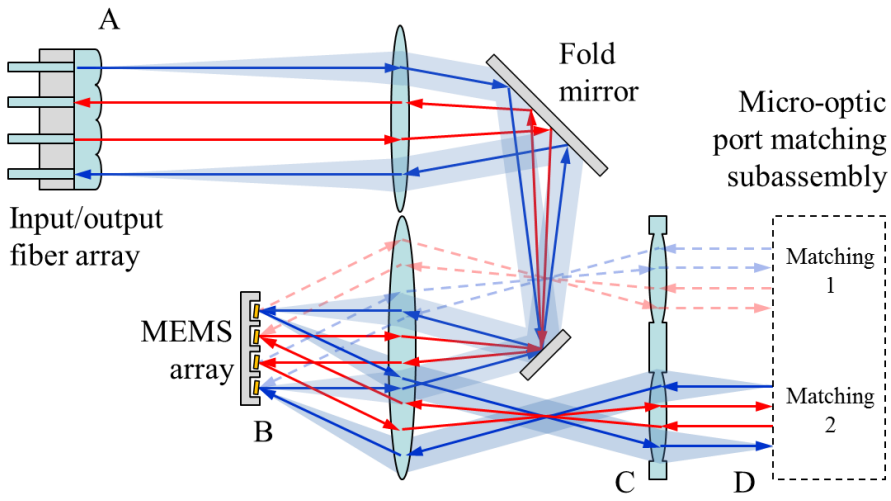

Figure 3.10: Schematic of freespace-interconnected selector switch. Signals are relayed onto a micromirror array, steered, and relayed again onto a micro-optic port matching assembly. After rearrangement, signals traverse the switch again to couple to output fibers in a shared input/output array. Solid and dashed chief rays indicate light paths through different matchings.

B to D, where the signals enter the micro-optic interconnection assembly. The interconnection assembly is a set of stacked micro-patterned substrates which spatially rearranges signals through refraction and reflection and sends them back through the switch to couple into output fibers. The different regions of the assembly are uniquely patterned to implement different port matchings.

3.5.2 Arbitrary Port Matching Subassembly

Figure 3.11(a) illustrates a port matching subassembly capable of arbitrary matchings and which can be integrated into the switch layout shown in Figure 3.10. It uses a single substrate with patterned prism facets on the front side to route signals and curved mirrors on the back side to reflect and refocus the optical beams. Each set of prism facets that constitute a port matching shares a common curved mirror. To maintain the beam shape across all ports, the curvature of the mirror must match the wavefront

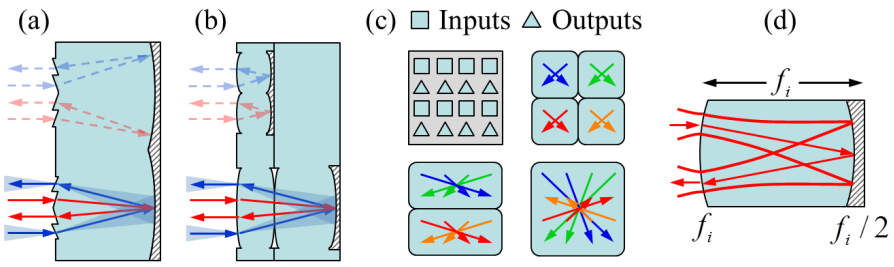

Figure 3.11: Schematics of micro-optic port matching subassemblies. (a) Arbitrary port matching subassembly using prism arrays. Solid and dashed lines show light paths through different port matchings. (b) Logarithmic port matching subassembly using microlens arrays. (c) Head-on view of input/output port arrangement and microlens patterns implementing Crossover port matchings. Ray color indicates port groupings. (d) Side view of section of the logarithmic port matching imaging relay.

curvature of the beam and the maximum differential path length of the beams must be small. A fabrication challenge of this design is the large sag of the curved mirrors, possibly necessitating assembly from bulk optics.

3.5.3 Logarithmic Port Matching Subassembly

While arbitrary port matchings may be required in some situations, logarithmic port matchings support a number of useful interconnection architectures and can be realized in a microlens-based geometry which is more amenable to lithographic fabrication. Here, we examine an implementation of the Crossover network topology, which is isomorphic to the set of logarithmic shuffle-equivalent topologies [66].

Figure 3.11(b) illustrates the port matching subassembly, which uses refractive and reflective microlenses in a stack of substrates to define a set of port matching imaging relays. Figure 3.11(c) shows a head-on view of how signals are mapped in the Crossover topology in an 8-port example using $\log_2 8 = 3$ port matchings. Figure 3.11(d) shows a side view of a section of the port matching imaging relay with a Gaussian beam passing through. The refractive microlens acts as a field lens to route signals through the center of the reflective microlens. The curvature of the reflector is chosen so the beam

waists are relayed to the surface of the refractive microlens as the beams exit the relay. The numerical aperture (NA) of the largest microlens increases as more ports and port matchings are added to a single port matching substrate, contributing aberrations, path length difference, and lens sag. To lower the NA requirements, multiple substrates can be stacked to accommodate a larger range of microlens focal lengths. Similar stacked assemblies of micro optics have been previously demonstrated [67]. As shown in Figure 3.11(b), microlenses can be placed at the intermediate refractive surfaces in the stack to split the optical power required by each surface (and thus reduce the maximum surface sag).

3.5.4 Logarithmically-interconnected 2,048-port Switch

Based on components discussed above, we used Zemax to design a 2,048-port switch and model its transmission. Figure 3.12(a) shows the switch layout. We based the design on a previously demonstrated 4,096-element micromirror array with 20 μs response, ± 4.3° mechanical tilt, and 120 μm diameter micromirrors [35]. We designed for a 64 × 64 single mode fiber array with 250 μm core spacing, with 2,048 cores acting as inputs and 2,048 as outputs. A BK7 microlens array attached to the fibers was designed to create an 83.3 μm beam waist at the microlens face for a pitch-to-waist ratio of 3 (for 99% power confinement in the aperture). A 200 mm focal length 2-glass telephoto lens, 96 mm 3-glass lens, and pair of fold mirrors are used to relay the beam waists onto the micromirror array. The lenses were designed with standard glasses and optimized for diffraction-limited performance over the C-band. Beam waists of 40 μm are formed on the 120 μm diameter high-fill-factor micromirrors (again a pitch-to-waist ratio of 3). The focal length ratio of the relay can be modified to accommodate different fiber or micromirror pitches.

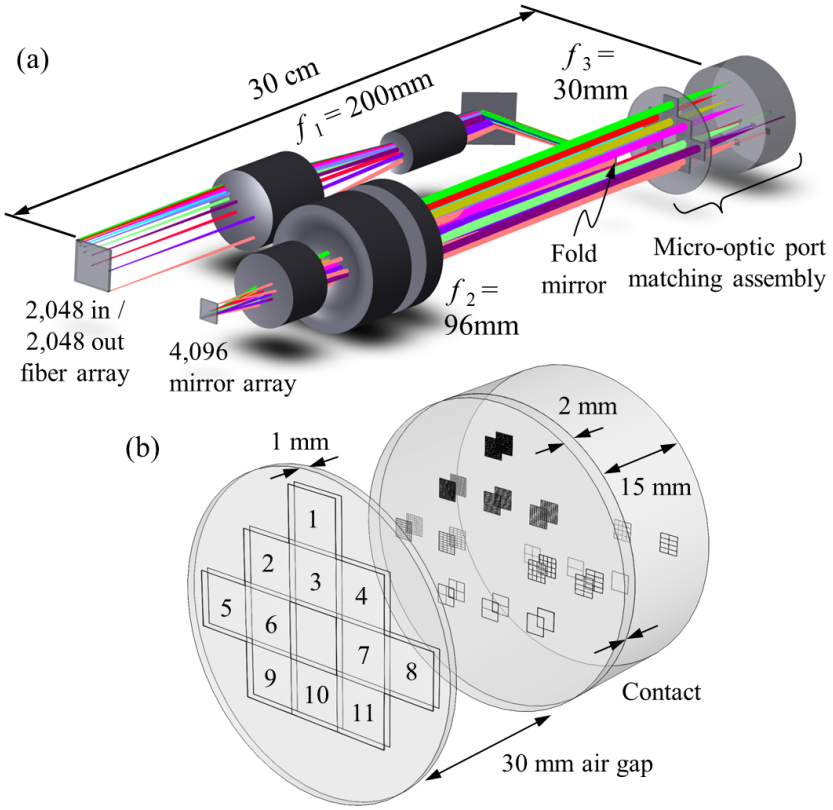

Figure 3.12: Zemax model of 2,048-port freespace-interconnected selector switch. (a) Full switch. (b) Close up of port matching assembly, indicating the 11 port matching apertures.

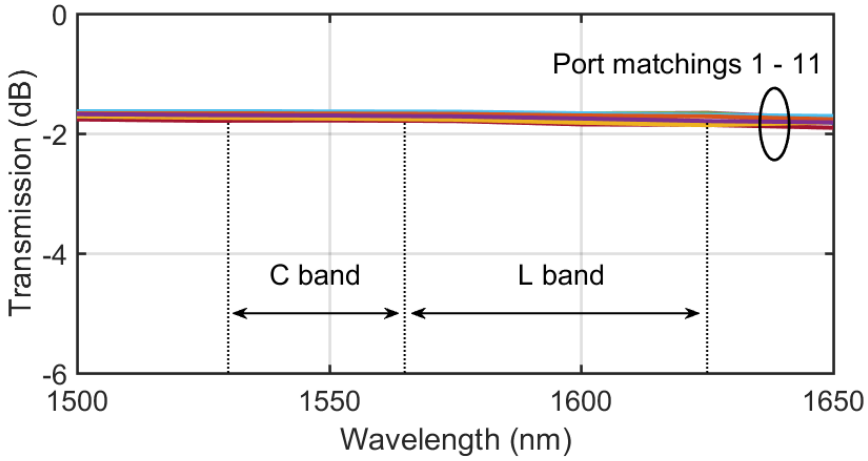

Figure 3.13: Modeled transmission of 2,048-port selector switch. Curves show worst-case transmission through all 11 port matchings. C and L bands are shown for bandwidth reference.

Figure 3.12(b) shows an enlarged view of the port matching assembly. An array of eleven 30 mm focal length silicon lenses defines the locations of the $\log_2 2{,}048 = 11$ port matchings. Two micro-patterned silicon substrates, 2 and 15 mm thick, are required for the port matching relay subassembly in order to keep the maximum microlens sag under 50 μm (sag previously demonstrated in silicon [67]).

Figure 3.13 shows the modeled transmission of the switch, including fiber coupling, which is > -2 dB over both the C and L bands. All refractive surfaces were assumed to be antireflection coated for 99.75% transmission. The fold mirrors were modeled with enhanced refection coatings (99.5% reflective) and the micromirrors and reflective microlenses were assumed to be gold coated (97% reflective). Tolerancing analysis indicated that a fiber-collimator misalignment of + 2 μm at the input and – 2 μm at the output introduced 3 dB excess loss. A misalignment of 5 μm between any of the substrates in the port matching assembly introduced 3 dB excess loss.

3.6 Discussion

This chapter presented a novel partially configurable optical switch architecture which is highly scalable in speed and port count without compromising transmission performance, potentially meeting the needs of data center networks. The design and experimental characterization were presented for a 61-port prototype selector module with 150μs switching time using commercial off the shelf components with a center-band overall switch loss of less than 10 dB. Detailed optical designs indicate the loss of the prototype could be reduced to 2 dB with custom optics. The switch can scale to 2,048 ports and a 20 μs response with 2 dB loss using micro-optic port matchings and a previously demonstrated micromirror array.

This chapter focused primarily on the physical-layer aspects of the selector switch, in order to establish its practical feasibility and scalability through modeling and prototyping. Details on how the switch could be used in data center networks and the performance of such networks are presented in Chapter 4.

Chapter 3, in part, reprints material that has been submitted for publication in a paper titled: “A Scalable, Partially Configurable Optical Switch for Data Center Networks,” submitted to the *Journal of Lightwave Technology*, by W. M. Mellette, G. M. Schuster, G. Porter, G. Papen, and J. E. Ford.

Chapter 4

SelecToR: A Partially Configurable Optical Data Center Network

This chapter investigates novel network architectures based on *partially configurable* selector switches. As discussed in Chapter 3, a selector switch can select port matchings from a small hardware library of pre-configured matchings, which make up a (tiny) subset of all possible bijective matchings. Limiting the configurability of the switch affords significant increases in port count and switching speed at the physical layer, but can partially configurable switches be used to construct performant networks? Despite the apparent connectivity limitations of selector switches, we show that throughput performance approaching that of a fully-provisioned packet switched network is possible through careful selection of the pre-configured matchings in novel network topologies based on selector switches. Two topology classes are presented: one based on logarithmically-spaced matchings which relies on indirect routing to restore complete connectivity, and another which uses parallelism to provide full connectivity without requiring indirection. Indirection in the second topology class can be used to load balance traffic, providing improved performance for sparse or skewed traffic patterns. A full scale network design and the integration of a prototype selector switch into a small scale network testbed are presented. An approach to distributed flow control is also discussed, with performance approach that of an optimal offline linear program solver.

4.1 Introduction

Today's data center networks are based on a crossbar switching model, where each switch provides arbitrary connectivity between ports. Building on Chapter 3, this chapter investigates how partially configurable selector switches can be used to best effect in network architectures. Because a selector switch is only partially configurable, a network built from selector switches will be partially configurable, meaning the arbitrary interconnectivity between endpoints enabled by a conventional packet switched network based on a non-blocking folded-Clos topology is not possible. Our analysis focuses on determining how many port matchings, and which matching patterns, are necessary to restore full connectivity in a network built from selector switches.

One approach uses a single switch pre-configured with a logarithmic number of interconnection matchings to connect all network endpoints. Because the switch cannot provide direct connectivity between all ports, we rely on indirection, allowing data to make multiple *hops* through the switch, to regain full connectivity. We analyze the reduction in network throughput due to multi-hop forwarding as the network scales. We also present results from a small-scale network testbed using a prototype selector switch and scheduling algorithm to communicate data between 16 servers.

Another approach uses a group of parallel selector switches to provide an expanded set of matchings without increasing the number of matchings in each switch. This approach can provide enough matchings for direct connectivity between all endpoints, eliminating the network capacity reduction due to indirection. However, we show that indirection is still useful under sparse (or skewed) traffic conditions. Specifically, a distributed control scheme based on the principle of load balancing can approach the performance of an offline solver for sparse traffic. More logical connections are required to implement this parallel-switch topology, but strategic network packaging can provide the modest degree of parallelism required for a large-scale deployment.

This chapter focuses on the optical portion of an optical-electronic parallel hybrid network. Throughput, or the achievable aggregate bandwidth, is the primary metric used to assess the topologies considered here. Latency, the time it takes data to traverse the

network, is another important network metric. At the time of writing, it has not been shown that the optical portion of the network can provide the latency performance necessary to entirely discard the electronically-switched portion. This topic is the subject of ongoing work. Here, as in other work, we assume that an under-provisioned packet switched network exists to handle any extremely latency-sensitive traffic, though it may be possible to route latency-sensitive traffic over multi-hop optical paths using more sophisticated routing and flow control methods.

We conclude the chapter with a large-scale network design example, showing that a selector switch based network, "SelecToR," can provide larger aggregate bandwidth capacity than a conventional packet switched folded Clos network for comparable cost and cabling complexity.

4.2 Network Throughput Model

In any communication network, there is a theoretical maximum throughput of information for a given communication pattern. In real networks, there are also practical limits to throughput due to the specific implementations of routing, flow control, network protocol, buffering, and a host of other factors. Before considering some of these implementation details, we first develop a framework to determine the idealized limits to throughput in a number of novel partially configurable network topologies. We use the throughput of a crossbar topology representing a strictly-non-blocking packet switched network as a baseline for comparison in subsequent sections of this chapter.

We used a commercially available linear program (LP) solver, Gurobi [68], as the basis for our theoretical throughput model. We construct a multicommodity network flow optimization problem by feeding variables, constraints, and an optimization criterion into the solver. The solver returns the optimal flow routing and throughput for the modeled network. This approach affords the flexibility to model arbitrary network topologies using the same framework by simply modifying the variables and constraints of the model. Because we wanted to model the throughput under different network traffic patterns, we enforced fairness constraints between flows so the resulting placement of

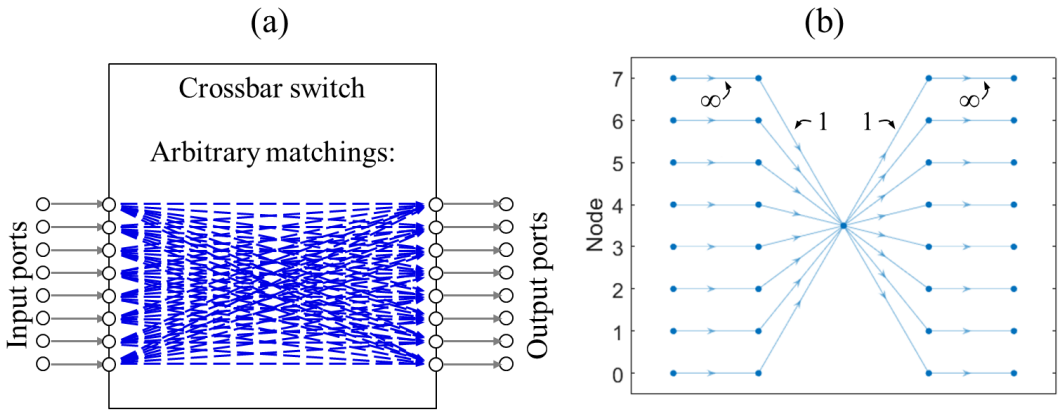

Figure 4.1: A crossbar and its graph representation. (a) 8-port crossbar switch showing arbitrary internal connectivity between input and output ports. (b) The graph representation of (a).

flows was representative of the traffic demands. We chose to enforce constraints based on the well-known principle of max-min fairness [69], so that contentious flows are not starved but flows with less contention are not unduly restricted. The details of the model are described below. We begin by outlining the method used to construct the network graphs, and then describe the flow constraints and optimization criterion placed on the graphs.

4.2.1 Graph Construction

In the framework of graph theory, a network can be represented as a collection of *vertices* and *edges*, which together form a *graph*. Vertices, or *nodes*, represent the source and destination endpoints which generate and sink data as well as the switches which redirect data. Edges represent the communication links which carry data through the network, and typically have a finite *capacity* to carry data. In this section, we describe how we form graphs to represent both crossbar as well as time-multiplexed partially configurable topologies. The next section discusses graph constraints and optimization.

Figure 4.1(a) shows a network with 8 end-points connected to a non-blocking crossbar switch, and Figure 4.1(b) shows the corresponding graph representation. In our

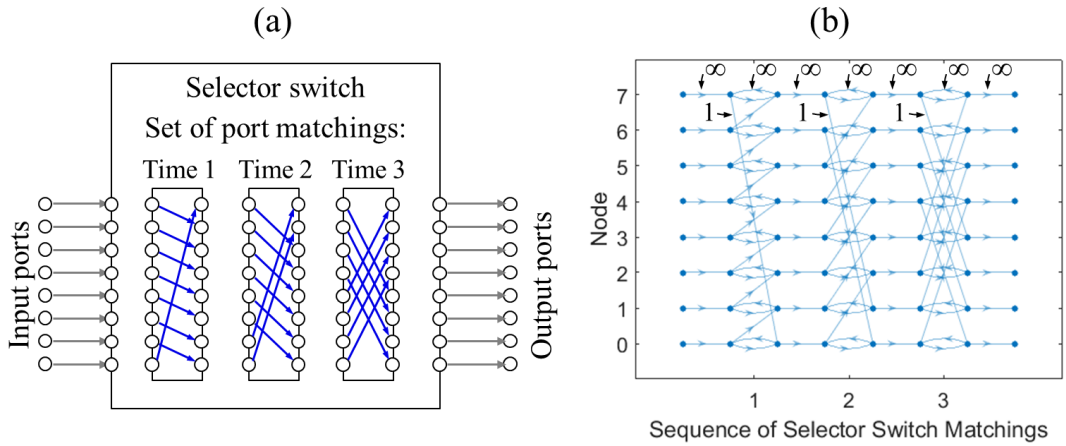

Figure 4.2: A selector switch and its graph representation. (a) Example logarithmically-interconnected 8-port selector switch showing example pre-configured matchings which are multiplexed in time by the switch. (b) The graph representation of (a), where the time multiplexing of matchings is represented by cascading matchings in the directed graph.

model we use *directed* graphs, meaning that data can only flow in one direction along each edge. The vertices on the left (numbered 0-7) represent the data sources, and those to the right are the destinations. The vertex in the center represents the crossbar switch, which is able to redistribute data entering from edges to the left across any of the edges leaving to the right. To represent the finite flow of data through each port on the crossbar, we assign finite capacities to the edges touching the crossbar vertex. For analysis we normalize these capacities to unity, but in a real network they would be the link data rate.

Figure 4.2(a) shows an 8 end-point network connected to a selector switch, and Figure 4.2(b) shows the graph representation. Unlike the crossbar, which could be represented by a single vertex able to forward data arbitrarily between input and output ports (edges), we must explicitly define the limited set of matching patterns internal to the selector switch. Further, because the matching patterns will be multiplexed in time by the switch, we represent them as a series of stages in the graph. The directed edges in the graph give the concept of time, allowing flows to interact with one matching at a time. Infinite capacity edges are placed to allow flows to be stored at a node before being sent at a later time. Depending on the edge constraints, data may be stored by intermediate nodes and later forwarded indirectly to its destination. We also provision the graph with

"loop-back" edges, so that flows can be forwarded through intermediate nodes while a matching is in place, allowing so-called "cut-through" indirection. The amount of cut-through allowed can be tuned through the edge capacities on the loop-back edges.

One artifact of our method of constructing graphs for selector switch based networks is that the order of port matchings is fixed by the topology of the graph. A chosen ordering may not in general be optimal for all traffic patterns. However, the predictability ensured by fixing the order may yield a simpler overall network control plane. In any case, the average throughput may be regarded as a lower bound on that achievable with variable ordering.

4.2.2 Solver Constraints and Optimization Criterion

In our multicommodity flow optimization problem, each source-destination pair communicates a unique "commodity" through the graph. In a network with N endpoints, there can be up to $N^2 - N$ commodities (assuming no endpoints send information to themselves over the network), and we must track the flow of these commodities to ensure they originate from the correct source and arrive at the correct destination. To do so, we define flow variables for each commodity along each edge in the graph. Each flow variable is indexed by its source, destination, edge source, and edge sink, where the edge source and edge sink are the start and end vertices of a given edge in the graph.

Next, we enforce constraints on the flow variables. The first constraint is that the sum of flows along each edge cannot exceed the capacity of the edge:

$$\forall i, j \in \text{edges} : \sum_{g,h \in \text{commodities}} \text{flow}[g,h,i,j] \leq \text{capacity}[i,j] \tag{4.1}$$

Next, the flow of commodities through each node must be conserved. This only applies to "internal" nodes in the graph which cannot source or sink data, and not to the designated source and destination nodes:

$$\forall g, h \in \text{commodities}, \forall j \in \text{internalNodes} : \sum_{i,j \in \text{edges}[*,j]} \text{flow}[g,h,i,j] = \sum_{i,j \in \text{edges}[j,*]} \text{flow}[g,h,i,j] \tag{4.2}$$

Next, we place constraints on the source nodes to ensure that flow commodities with source *s* can only be generated by source *s*:

$$\forall s \in \text{sources}, \forall g, h \in \text{commodities} \mid g \neq s : \text{flow}\left[g, h, s, *\right] = 0 \qquad (4.3)$$

Finally, we constrain the destination nodes to ensure that commodities bound for destination *d* cannot flow to any other destination:

$$\forall d \in \text{destinations}, \forall g, h \in \text{commodities} \mid h \neq d : \text{flow}\left[g, h, *, d\right] = 0 \qquad (4.4)$$

With the flow variables and constraints in place, we next set up the optimization criterion. Our overall goal is to enforce fairness amongst flows according to bandwidth demands from a given network traffic pattern. With this in mind, we adopt an optimization criterion from the Maximum Concurrent Flow Problem [70] as follows. We define a variable *z* which is the fraction of demanded bandwidth *D* assigned to a flow. All flows are subject to the constraint:

$$\text{flow}\left[\text{source}_i, \text{destination}_i, *, \text{destination}_i\right] = z \cdot D_i \qquad (4.5)$$

The optimization function is to maximize *z*. In this way, all flows are fairly allocated an equal fraction of their demanded bandwidth. While this means no flows are starved, it may also unduly restrict the bandwidth of flows which could have been allotted more bandwidth without negatively impacting any other flows. We avoid this case with an iterative approach, outlined next.

4.2.3 Iterative Max-Min Fairness

In each iteration, the LP solver assigns bandwidth to flows so the most heavily contended bandwidth is distributed fairly amongst its contending flows. However, multiple iterations of the solver are necessary to ensure the bandwidth of flows under less contention is not unduly restricted. The overall iterative algorithm is structured as follows.

First, the original bandwidth demand is input. A copy of the demand is made and all non-zero demands are scaled to unity. This scaled demand is input to the LP solver, which returns the path and allocated bandwidth for each flow through the network. The

allocated bandwidth capacity along each edge is subtracted from the edge's current capacity. Next, the allocated bandwidth is compared flow-wise to the current bandwidth demand to determine if any flows were allocated too much bandwidth. If so, the over-allocated capacity is added back to the corresponding edges in the graph. The allocated bandwidth is then subtracted from the current bandwidth demand to update the current bandwidth demand. Next, a single-commodity flow is run on each flow with remaining demand to determine if the flow can still traverse the graph after the graph capacities were updated. The set of flows that can still traverse the graph are input again to the LP solver, and the entire process is repeated until all demand is served or no more flows can traverse the graph. The end result is an allocation of bandwidth which is fair to flows under heavy contention, but which does not arbitrarily restrict the bandwidth of flows under light contention.

4.3 Logarithmically-interconnected Topologies

This section investigates a class of partially configurable network topologies with logarithmic diameter. In this case, an N-port selector switch is pre-configured with $\log_2 N$ port matchings. With only $\log_2 N$ matchings, each input port on the switch can only send data to at most $\log_2 N$ output ports, a small subset of all N output ports. We describe a number of equivalent topologies which use indirect routing of data to recover full connectivity between all input and output ports, requiring data make at most $\log_2 N$ hops through the switch. We also present a scheduling algorithm to determine a sequence of port matchings and indirectly route flows through those matchings. Measurements from a small-scale testbed using a prototype selector switch are also presented.

4.3.1 Chord and Shuffle-equivalent Matchings

Logarithmic network topologies have been studied extensively due to their ability to provide connectivity between all endpoints of a network with a minimum number of

connections. A prominent class of logarithmic topologies is known as *shuffle-equivalent*, and includes the Perfect Shuffle, Banyan, Crossover, and others. It has been shown that these topologies are isomorphic to one another [66]. Shuffle-equivalent topologies have been traditionally implemented as a binary-logarithmic number of interconnection stages cascaded in space with switching stages between each interconnection stage. Data makes a logarithmic number of hops, one through each stage, to reach its destination.

A similar class of topologies can be implemented using selector switches, but now with the interconnection patterns multiplexed in time rather than cascaded in space. Instead of establishing multi-hop paths in a spatial fabric, data is routed through space

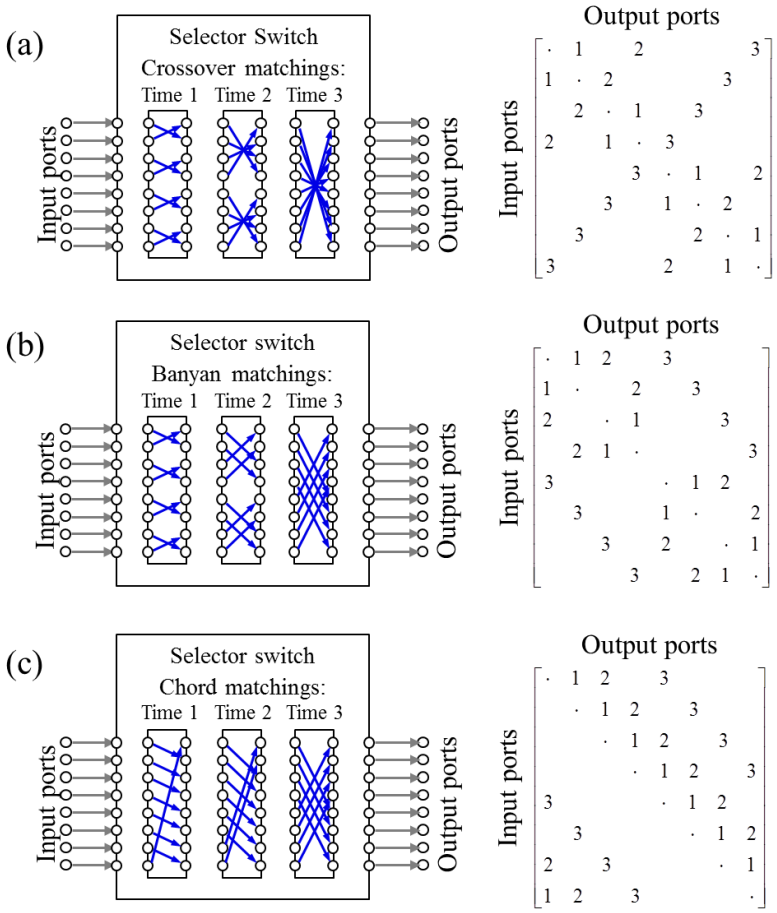

Figure 4.3: Logarithmically-interconnected selector switch topologies. Examples of 8-port selector switches with matchings inspired by (a) Crossover, (b) Banyan, and (c) Chord. The adjacency matrices are shown to the right, indicating the connectivity provided by each matching in time.

and time, making at most $\log_2 N$ hops through time-multiplexed matchings in an N-port selector switch. Such multi-hop routing requires memory at intermediate nodes along the path to store data until the correct matching is configured to establish the next hop along the path. This topology class is a good match for the selector switch hardware because a logarithmic number of matchings permits the design of fast, large port count, and low loss switches at the physical layer, as demonstrated in Chapter 3.

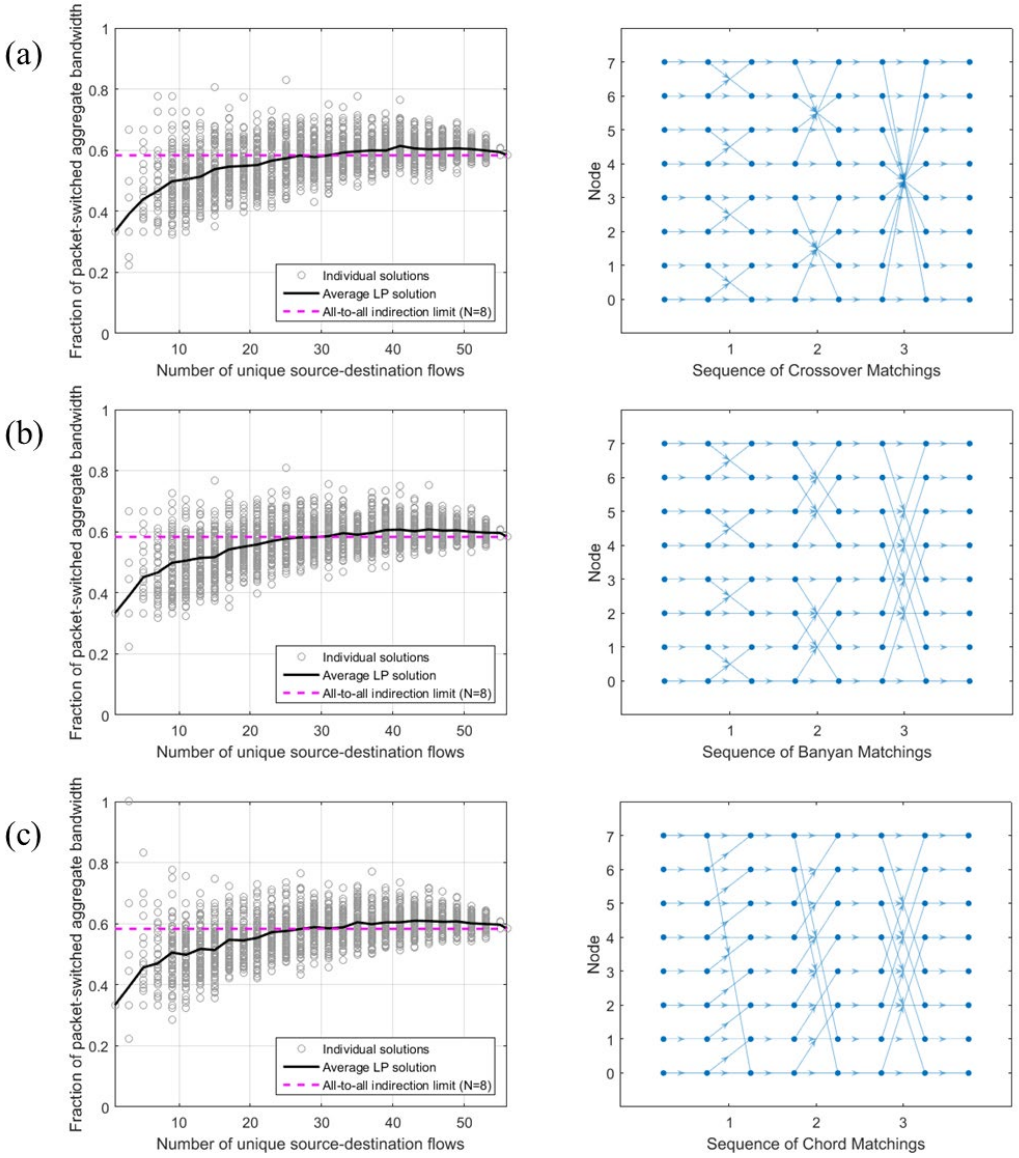

Figure 4.4: Throughput of logarithmically-interconnected selector switches. Modeled for the 8-port selector switch topologies in Figure 4.3, relative to that of a packet switched network. The number of flows ranges from one (sparse demand) to 56 (all-to-all demand). The graph for each network is shown to the right.

Figure 4.3 shows three example choices of logarithmic port matchings for 8-port selector switches, inspired by the Crossover, Banyan, and Chord interconnection patterns. The left illustrations show the set of $\log_2 8 = 3$ matchings in the selector switch, and the adjacency matrices are shown to the right, representing the connectivity established by each matching.

Figure 4.4 shows the modeled throughput for 8-port selector switched networks with Crossover, Banyan, and Chord inspired matchings using the iterative multicommodity flow solver outlined in Section 4.2. The throughputs are normalized to the modeled throughput of a crossbar-based packet-switched network. The x-axis sweeps the number of flows with unique source-destination pairs in the network, from 1 to 56, with 56 flows being an all-to-all communication pattern. For each point along the x-axis, 96 randomly generated bandwidth demand matrices with the corresponding number of non-zero flows are generated and run through the solver. The values in the bandwidth demand matrices are binary, representing the presence or absence of a "heavy-hitter" flow. Skewed demand matrices can be formed by the scaled superposition of binary demand matrices with different numbers of flows. For example, a demand matrix with low-bandwidth all-to-all traffic and a small number of high-bandwidth heavy-hitter flows can be decomposed into a weighted sum of all-to-all and sparse demand matrices.

The model assumes that the switch stays in each matching configuration much longer than the reconfiguration time of the switch, so that the duty cycle of the switch is 100%. In a real system, the duty cycle will be less than 100%, and will simply scale the throughput results reported in Figure 4.4.

The results indicate, not surprisingly, that the three logarithmic topologies have the same throughput performance. The average throughput is about 60% that of a fully-provisioned packet network for dense communication patterns and converges to 33% throughput for a single flow.

Unfortunately, the throughput performance diminishes as the number of switch ports increases. This is discussed in more detail in the following subsection, but is fundamentally due to an *indirection penalty* caused by the multi-hop routing necessary to recover connectivity between all ports with only $\log_2 N$ matchings. Flows making

multiple hops use up more bandwidth than those with direct connections to their destinations. As derived below, the maximum throughput for an all-to-all traffic pattern in a logarithmically-interconnected N-port selector switch is approximately $2 / \log_2 N$, relative to a fully-provisioned packet network. The logarithmic onset of the indirection penalty means these topologies may find utility in networks with a modest numbers of endpoints, or for connecting a small number of heavily aggregated endpoints.

4.3.2 An Approach to Scheduling Matchings and Routing Flows

The previous section used the multicommodity flow solver to model the throughput of logarithmically-interconnected selector switched networks. In this section, we investigate the properties of this selector switch topology class in more detail. In the process, we derive throughput scaling properties and develop a heuristic approach for scheduling flows given a traffic demand matrix.

The Chord [61] matchings are perhaps the easiest logarithmic matchings to reason about, and we focus on them in the following analysis without loss of generality. Chord matchings are constructed by matching port p to ports $p + 2^0, 2^1, \ldots, 2^{\log_2 N - 1}$ modulo N, where p is indexed from 0. Figure 4.3(c) shows an 8-port selector switch pre-configured with Chord matchings and the corresponding adjacency matrix. In Figure 4.3(c), matching 1 connects each port to its nearest neighbor, matching 2 to its second nearest neighbor, and matching 3 to its fourth nearest neighbor. In other words, the matchings are spaced by powers of two. The lack of symmetry about the main diagonal in matchings 1 and 2 means those connections are not bidirectional.

The switch multiplexes the matchings in time, allowing data to reach nodes without a direct connection by making multiple hops through the switch. For example, if a node wants to send to its sixth nearest neighbor, it can first send the data to its second nearest neighbor through matching 2, and then that node can forward the data to the destination through matching 3 for a total of 2 hops. We refer to the path through the matchings as a *routing*. In general, the routing problem is equivalent to an integer composition problem: to send data to a node X ports away, we need to find the

composition of X using powers-of-two. There are a potentially large number of compositions if we allow the composing matchings to be repeated. Returning to our example, the sixth nearest neighbor can be reached in 6 hops through matching 1 or 3 hops through matching 2. However, these paths require data take more hops than the minimum-hop-path, or *minimum-routing*. Routings with more hops than the minimum required are generally not preferred because they create more contention in the network, as each additional hop consumes bandwidth resources. The minimum-routing can be found by expressing the port-distance to be traveled as a binary number, and the minimum number of hops required is simply the Hamming weight of that binary number. In our example of sending to the sixth nearest neighbor, 6 in binary is 110 and the Hamming weight is $H(110) = 2$. The matchings required for this routing can be determined by reading off the 1's positions in the binary number with the least significant bit corresponding to matching 1 and the most significant bit corresponding to matching $\log_2 N$. For example, the routing 110 requires data traverse matchings 2 and 3, but not matching 1. However, either order of matchings 2 and 3 will result in a minimum-routing requiring 2 hops.

To determine the best ordering of matchings, we look to the routings which require the largest number of hops. In an N port switch, the minimum-routing requiring the most hops will occur when sending data a port-distance of N-1, as this requires all $\log_2 N$ matchings be traversed one time each. There are $(\log_2 N)!$ ways to permute the order the matchings. Consider one of those orderings; returning to our 8-port example, consider the ordering {3, 2, 1}. Physically, the switch will sequentially step through this matching order in time. All flows traveling a port-distance of N-1 (8-1 = 7 in this case) will be routed through matchings in this order. Now consider the routes of the remaining flows with lower port-distances, which will all traverse only a subset of the matchings. By defining an ordering of the full set of matchings, we have implicitly defined the ordering of all subsets as well. Returning to our previous example, all flows with a port-distance of 6 will traverse matchings in the order {3, 2} if the longest routing is ordered {3, 2, 1}. We refer to the set of ordered routings for all port-distances as *compatible-routings*. Using compatible-routings ensures all flows can be served in at most one

complete cycle through the matchings, which helps to minimize latency and maximize bandwidth. The optimal ordering of matchings in a compatible-routing will depend on the traffic demand and the degree of control over the switch, as discussed below.

For all-to-all traffic, all compatible-routings yield equivalent bandwidth performance. In this case, the selector switch can be set to repetitively cycle through all matchings in any order with an equal amount of time spent in each matching. Each flow is assigned a sub-division of the time spent in each matching, so that bandwidth is partitioned fairly (and consistently) amongst all flows. The result will be that during each matching's time slot, each node will send 1 unit of direct (1-hop) data and (N/2-1) units of indirect (multi-hop) data, for a total of N/2 units of data. After one cycle through all matchings, each node will have communicated one "original" data unit to each of the other N-1 nodes, but will have send a total of ($\log_2 N$)(N/2) data units to support indirect flows. The throughput is the ratio of the time spent sending original data to the time spent sending all data:

$$Throughput_{Chord,All-to-All} = \frac{2(N-1)}{N \log_2 N}. \tag{4.6}$$

For large N, the all-to-all throughput scales as 2 / $\log_2 N$. While derived here for the Chord matchings, this property applies to all logarithmically-interconnected selector switch topologies.

Next, we consider sparse demands (those with fewer than N^2-N flows). There are two primary options to serve sparse demand: 1) simply use all-to-all routing with equally-weighted round robin access to the matchings, or 2) actively reconfigure the switch based on current demand, allowing unequal time in each matching (potentially skipping some matchings). Approach (1) yields a linear reduction in throughput as the demand becomes sparser:

$$Throughput_{Chord,\ Round-Robin}\left(N_{flows}\right) = \frac{2\left(N_{flows}\right)}{N^2 \log_2 N}. \tag{4.7}$$

Approach (2) can yield higher bandwidth performance for sparse demand, at the expense of a more complex control loop, requiring real-time demand estimation and scheduling.

We take one such approach to constructing a scheduler, building off our discussion of compatible-routings as follows. Given a demand matrix and the set of $(\log_2 N)!$ compatible-routings, we algorithmically compute the time required to drain the demand during each matching of each compatible-routing. The calculation is straight-forward, in that we can compute the transmission time for each flow over each routing, taking into account the indirection penalty resulting from other flows transiting that same routing. The scheduler returns the time-ordered sequence of matchings and their durations which minimize the total time to serve all demand. This requires a search over a $(\log_2 N)!$-sized space, but because each computation is independent the process can be parallelized, potentially admitting GPU-based approaches. Similar scheduling techniques can be applied to the other logarithmically-interconnected selector switch topologies.

The modeled throughput of the centralized scheduler running on an 8-node network is plotted in Figure 4.5 along with that of the simple round robin approach and

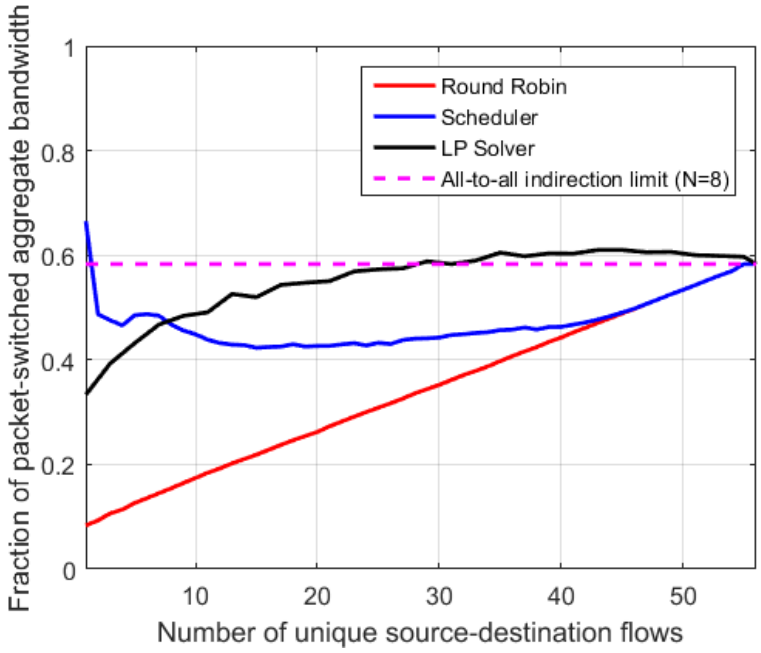

Figure 4.5: Chord-based selector switch throughput under various flow controls. Modeled average throughput for an 8-port selector switch network, relative to a packet switched network. The number of flows ranges from one (sparse demand) to 56 (all-to-all demand). Different flow control schemes are considered, including round robin, a centralized scheduler, and a flow optimization algorithm implemented with an LP solver. See text for details.

that of the LP solver outlined in Section 4.2. All throughputs have been normalized to that of a packet-switched network modeled using the LP solver. Round robin scheduling has equivalent throughput to the other approaches for dense communication patterns, but gives poor performance for sparse demand because it does not adapt to the demand structure. The scheduler yields better performance for sparse demand, even outperforming the LP solver for fewer than about 8 flows. This is because the LP solver assumes all matchings are used for an equal amount of time (yielding a throughput of $1/3^{rd}$ for a single flow), while the scheduler can assign unequal time durations in each matching and even skip matchings. The LP solution outperforms both round robin scheduling and the centralized scheduler for most demand conditions because it allows flows to be subdivided into groups of smaller flows to take advantage of capacity left unused in the other two approaches.

Under all three routing approaches, throughput converges to the all-to-all indirection limit defined in (4.6) at 56 flows (all-to-all demand). For 8 nodes, the limit is approximately 0.583. The throughput for the traffic patterns considered never significantly exceeds the all-to-all indirection limit. As the logarithmically-interconnected topologies discussed here scale to support more endpoints, the indirection penalty grows logarithmically, limiting the throughput of the network. One application of these topologies may be to interconnect a relatively small number of heavily-aggregated server clusters to minimize the effect of the indirection penalty.

4.3.3 Prototype Network Testbed

We used the prototype selector switch designed and fabricated in Chapter 3 in a small-scale network testbed to demonstrate the feasibility of implementing a selector-switched network. The testbed consisted of 8 servers, each with two dual-port 10 Gb/s network interface controllers (NICs), which were virtualized to emulate a total of 16 independent servers. The testbed layout is shown in Figure 4.6. Each server had a 10 Gb/s electrical connection to a control network which was connected to a controller

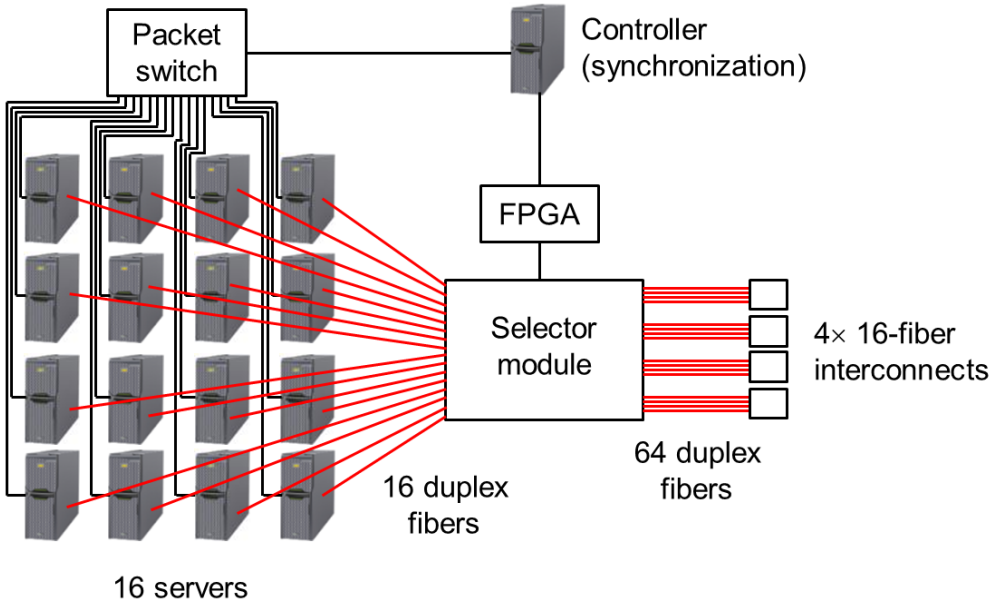

Figure 4.6: Network testbed layout. 16 servers were connected via a 10 Gb/s packet switch to a control server. The 16 servers were also connected via 10 Gb/s optical transceivers to the prototype selector switch. An FPGA was used to interface between the control server and the selector module, with a 1 Gb/s connection from the controller to the FPGA.

server. The controller also had a 1 Gb/s connection to a field-programmable gate array (FPGA) which controlled the selector module. Each virtual server had a commercial 10 Gb/s optical transceiver module which was connected to the selector module with a duplex fiber cable. Critically, no optical amplification was required to send data through the switch and meet the transceiver link budget. We chose to implement a logarithmically-interconnected topology as discussed above. The 16 servers required $\log_2 16 = 4$ matching patterns with 16 connections per matching. We could have manually routed fiber on a fixed patch panel to configure the matchings, but instead chose to use a 64-port MEMS cross-connect as a reconfigurable patch panel to easily change the preconfigured matchings. The state of the cross-connect was unchanged during experiments, and would not be used in a real deployment.

A photograph of the testbed is shown in Figure 4.7(a). Figure 4.7(b) shows how the ports on the prototype selector switch were allocated to support 16 end hosts: 16 of the 61 ports were used as inputs and 16 as outputs by folding the port matchings patterns

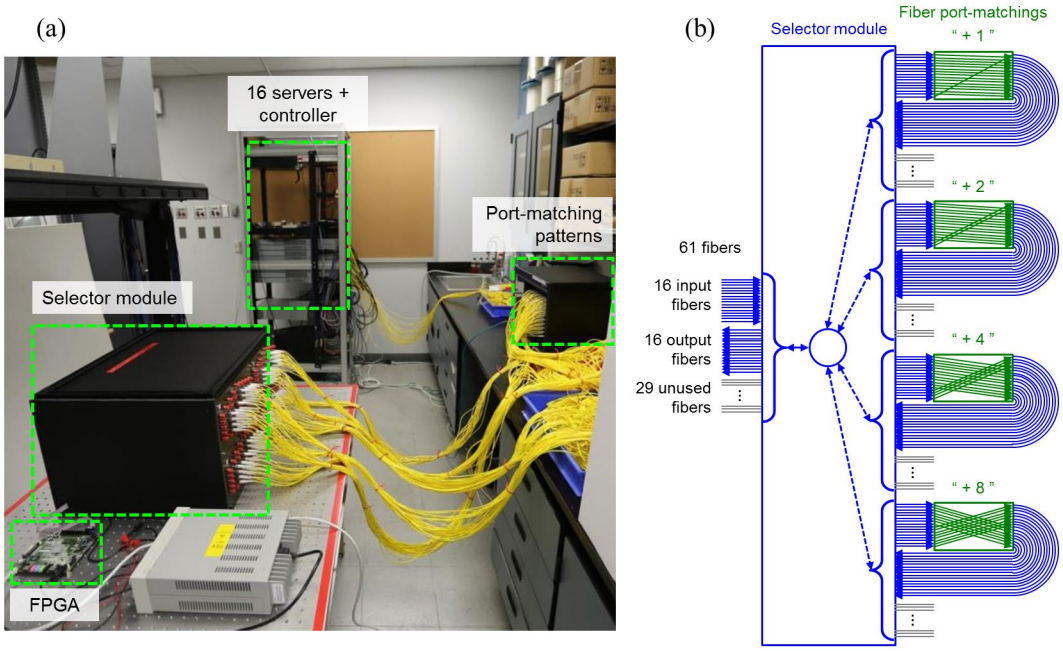

Figure 4.7: Photograph of network testbed and selector switch layout. (a) 16-server network testbed. (b) Schematic of the prototype selector module used as a 16-port 1×4 selector switch.

back through the same selector module. This allowed us to use a single MEMS micromirror as both the input and output selector switching device. The MEMS micromirror routs all 16 optical signals through the same port matching at a given time. We used the Chord port matching patterns in our experiments, shown in Figure 4.7(b).

We used the testbed to experimentally determine the performance of the scheduling algorithm outlined in the previous subsection. This was done by running the scheduler offline on a set of precomputed demand matrices to determine the timing of data flows and switching events, and then replaying the data flows and switching events in real time on the testbed using the controller server for synchronization. We used UDP (user datagram protocol) senders and receivers on each server to communicate data. A custom time division multiple access (TDMA) queuing discipline and precision time protocol (PTP, IEE 1588) were used to synchronize the transmission of data with the reconfiguration of the selector switch.

We ran three traffic patterns on the testbed, varying the transmission timeslot of the optical switch in each case. The results are shown in Figure 4.8. The first traffic

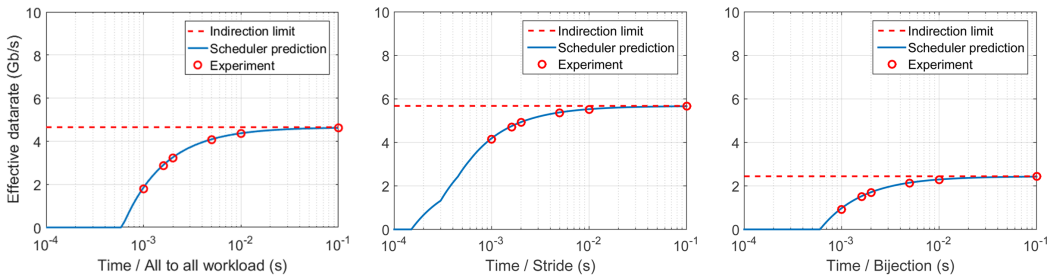

Figure 4.8: Comparison between modeled and measured testbed throughput. Three workloads were considered: all-to-all, stride, and random bijective.

pattern was uniform all-to-all, meaning that every server needed to transmit data to every other server. We see in Figure 4.8(a) that the experimentally measured throughput closely matches the scheduler's predicted throughput. Further, for timeslot lengths which are long relative to the 150 μs reconfiguration time of the switch, the measured throughput approaches the indirection limit of the 16-port switch. The second traffic pattern was a "rolling-stride," where each server sends data to its nearest neighbor for a specified time, then to its second nearest neighbor for the same time and so on, repeating cyclically. The third traffic pattern was a set of 1,000 randomly generated bijective demand matrices implemented sequentially in time. In this scenario, each server only communicates with one other server at a time.

The good agreement between the modeled throughput and experimental results validate our analysis model, and also indicate that (at least small-scale) selector switched networks can be constructed. Of course a remaining undertaking would be to integrate the demand estimation, scheduler, and controller to operate in real time on real application traffic. We leave this work to be explored elsewhere and turn our attention back to the exploration of partially configurable network topologies in the next section.

4.4 Completely-interconnected Topologies

In this section, we investigate another class of partially configurable topologies which overcomes the indirection limit of the logarithmically-interconnected topologies

discussed in the previous section. We call this new topology class "completely-interconnected" because it provisions (time-multiplexed) single-hop connections between *all* network endpoints. This fundamentally requires an expanded set of selectable port matchings. For example, with $N - 1$ matchings, each implementing a unique matching between N ports, all $N^2 - N$ possible input-output connections are provided. Each endpoint must simply wait until the matching with the desired connection is selected by the switch, and then can send at full link rate directly to the destination. However, preconfiguring a single selector switch with $N - 1$ matchings decreases many of the physical layer hardware advantages of a selector switch over a conventional cross-connect. As discussed in Chapter 3, one of the primary reasons the selector switch can scale to fast reconfiguration speeds is that the MEMS switching element discriminates between fewer than the N optical states of a cross-connect. Preconfiguring a single switch with $N - 1 \approx N$ matchings would substantially negate the physical layer speedup.

To sidestep this problem, and keep the number of port matchings per switch approximately equal to $\log_2 N$, we distribute the $N - 1$ matchings amongst a set of parallel selector switches. This requires that each endpoint have a set of parallel communication channels with at least one connected to each selector switch. There is a balance between the number of matchings per switch and the degree of parallelism in this approach. Taken to the extreme, each endpoint could have $N - 1$ hardwired connections, one to every other endpoint in the network, and no switching would be required at all. However, the cost and cabling complexity involved make this solution infeasible. Instead, our approach takes a middle road – few enough matchings per selector switch to make the switch scalable, but enough to keep cabling complexity manageable. In Section 4.5, we describe a network architecture which provides the necessary parallelism to implement our approach at scale. In this section, we discuss two example topologies that use parallel selector switches to provide complete connectivity between network endpoints. A key feature of these networks is that the selector switches can be set to cyclically repeat the same set of matching configurations with equal time spent in each matching, significantly simplifying the synchronization between switches and data transmission.

4.4.1 Rotor Matchings

Perhaps the most straightforward set of port matchings which collectively provide complete connectivity with the fewest number of matchings are those that form the off-diagonals of the network's adjacency matrix. This set is an expansion of the Chord matchings discussed in the previous section, where instead of only including the logarithmically-spaced off-diagonals, we include the full set of $N - 1$ matchings providing +1, +2, +3, …, +($N - 1$) modulo N connectivity. This basic interconnection structure has been explored previously for providing connectivity in an electronic switch [60]. In that work, a single switch repetitively cycled through all $N - 1$ matchings in a

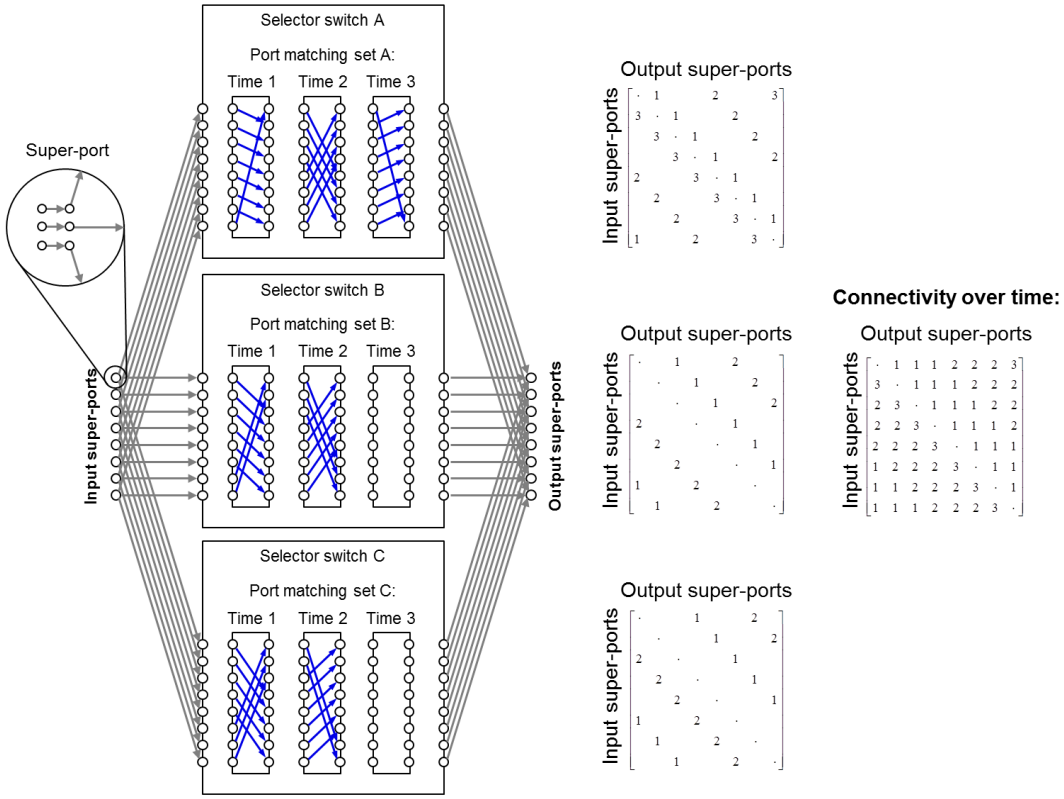

Figure 4.9: Completely-interconnected selector switch with Rotor matchings. Shown for an 8-endpoint network. Each super-port (network endpoint) has three logical connections, one to each of the three selector switches. The Rotor matchings are divided between the three switches, and the time-multiplexed adjacency matrix for each switch is shown to the right. The time-multiplexed adjacency matrix for the overall topology is shown to the far right, showing complete single-hop connectivity between all super-ports.

rotary fashion. Here, we refer to these interconnection patterns as "Rotor" matchings, but instead of preconfiguring a single switch with all $N - 1$ matchings, we install approximately $\log_2 N$ matchings into each of ceiling[$(N - 1)/\log_2 N$] selector switches.

This topology is shown in Figure 4.9 for an 8-port network. We refer to the ports connected to network endpoints as "super-ports." Each super-port has a set of logical sub-channels, with one sub-channel connected to a port on each selector switch. Each selector switch provides only partial connectivity, but when the switches are taken together, full connectivity between all super-ports is realized over time. This can be seen through the superposition of the adjacency matrices of the selector switches, as shown in Figure 4.9. Because the $N - 1$ Rotor matchings may not be evenly divisible into a number of switches, there may be a number of matching "slots" which are not well-defined. For example, in Figure 4.9, the third matching slots in switches B and C are empty because the 7 Rotor matchings could not be equally divided between 3 switches. We could, of course, fill these empty slots with random matchings, but the number of empty matching slots varies with the number of network endpoints, so there is no systematic way to define the matchings for these slots. In order to present the most straightforward analysis, we simply leave these undefined slots empty, and interpret our results as the lower bound on network performance. With some matching slots left empty, the maximum achievable network throughput will be scaled by the ratio of filled matching slots to total matching slots. We refer to this scaling factor as the matching *packing factor*. Fortunately, as we scale the network to more endpoints, the packing factor approaches 1 because the effect of a small number of empty matching slots is outweighed by the growing number of total matchings. This effect is illustrated in Figure 4.10, which shows the maximum all-to-all throughput as a function of the number of endpoints for a network based on Rotor matchings and one based on Chord matchings. The all-to-all throughput of the logarithmically-interconnected network is limited by the indirection penalty, while that of the completely-interconnected network is limited by the packing factor and approaches full throughput at large scale.

Full bisection bandwidth for all-to-all traffic is an attractive feature of completely-interconnected networks, but we'd also like to determine the network

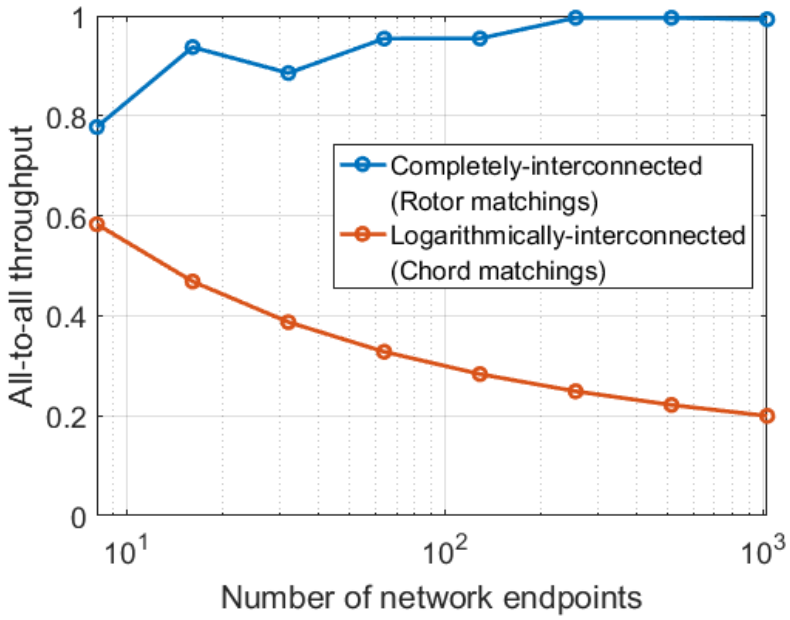

Figure 4.10: All-to-all throughput: logarithmic vs. complete interconnection. Throughputs evaluated for 8- to 1,024-endpoint networks. The throughput of the logarithmically-interconnected network is subject to the indirection limit, whereas that of the completely-interconnected network approaches unity along with the packing factor.

throughput for sparser traffic patterns. In order to simplify the synchronization between the switches and endpoints, we require that the selector switches repetitively cycle through their set of port matchings, spending an equal time in each matching. Achieving high throughput for sparse demand is then simply a matter of how we implement flow control in the network, specifically if we allow store and forward indirection and/or cut-through indirection.

In the simplest case, we do not permit any indirection in the network and each endpoint simply waits to send data to a destination until a direct connection is established to that destination. We refer to this as round-robin flow control. The throughput as a function of the number of flows in the network is given by

$$Throughput_{Rotor,\ Round-Robin}\left(N_{flows}\right)=\frac{N_{flows}}{N_{sw}N_{match}}, \tag{4.8}$$

where N_{sw} is the number of parallel selector switches and N_{match} is the number of preconfigured port matchings in each switch. The throughput is linearly reduced as the traffic pattern becomes more sparse, yielding a throughput of approximately $1/(N-1)$ for a single flow.

Incorporating some degree of indirection into the flow control gives better performance for sparse demand by effectively load-balancing the traffic. We used the LP solver to model the network throughput with only store and forward indirection (and no cut-through). This means that an endpoint can forward data to an intermediate endpoint during each matching period, but that intermediate endpoint cannot forward that data until the next matching period. There is a subtle tradeoff between the bandwidth and latency for sparse traffic in this flow control scheme. The number of available indirect routes over which to send traffic increases with the number of matching cycles. In other words, the bandwidth of sparse traffic increases if that traffic can be delivered at a later point in time.

Finally, we consider a flow control scheme which allows both cut-through and store and forward indirection. Cut-through means that an endpoint can forward data to an intermediate endpoint during a matching period, and that intermediate endpoint can forward that data again during that same matching period. The number of cut-through hops within a single matching period is limited to less than N, although most data takes much fewer than N hops. Still, too much cut-through may be impractical in a real system due to the overhead of setting up the multi-hop path as well as the latency introduced at each hop.

Figure 4.11 shows the modeled throughput for the 8-endpoint Rotor-based network shown in Figure 4.9, under the three flow control schemes discussed above. The throughputs are normalized to that of a fully-provisioned packet switched network. We see that all three schemes converge to full bisection bandwidth (limited only by the matching packing factor) for all-to-all traffic. The LP solution using store and forward indirection outperforms round robin for sparse traffic patterns. The store and forward solution approaches the LP solution using cut-through indirection as the number of matching cycles allotted for storing and forwarding data increases, illustrating the

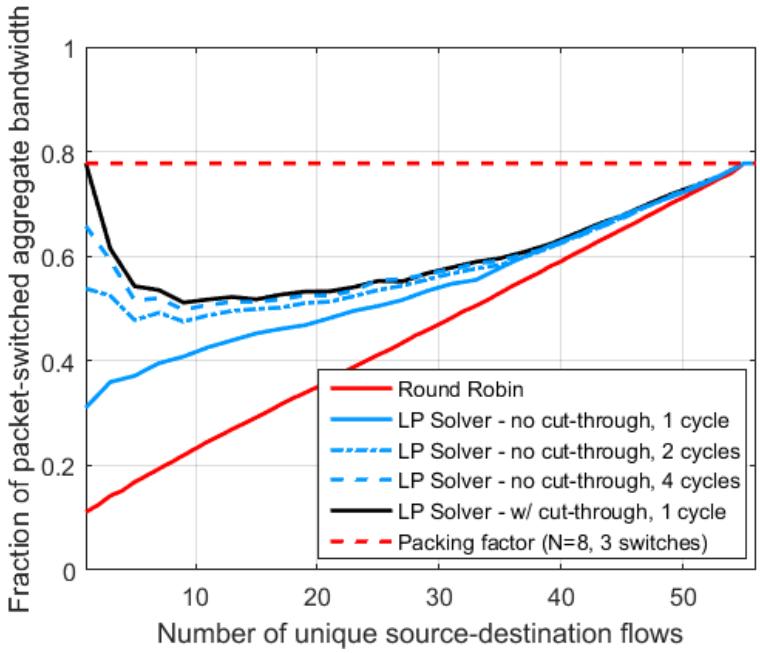

Figure 4.11: Rotor-based selector switch throughput under various flow controls. Modeled average throughput for an 8-port selector switch network pre-configured with Rotor matchings, relative to a packet switched network. The network uses 3 selector switches each configured with 3 matching slots (see Figure 4.9). The number of flows ranges from one (sparse demand) to 56 (all-to-all demand). Different flow control scenarios are considered including round robin (with no indirection), store and forward indirection, and both store and forward and cut-through indirection. See text for details.

tradeoff between space and time. Full link bandwidth (limited only by the matching packing factor) for a single flow is possible using either a large number of cycles with store and forward flow control or in a single cycle using cut-through indirection.

Finally, we considered the throughput of permutation traffic for different network sizes and different approaches to flow control. Permutation traffic is ideal for a crossbar (i.e. packet) switch, because there is no contention for switch resources. In this respect, permutation traffic is adversarial to the Rotor-based selector switch network, which has the best performance for all-to-all traffic. This can be seen in Figure 4.11, noting that the lowest throughput relative to the packet switch occurs around $N_{flows} = 8$ for the 8-endpoint network. The throughput of permutation traffic for 8, 16, 32, and 64-endpoint networks is shown in Figure 4.12. Because the matching packing factor varies with

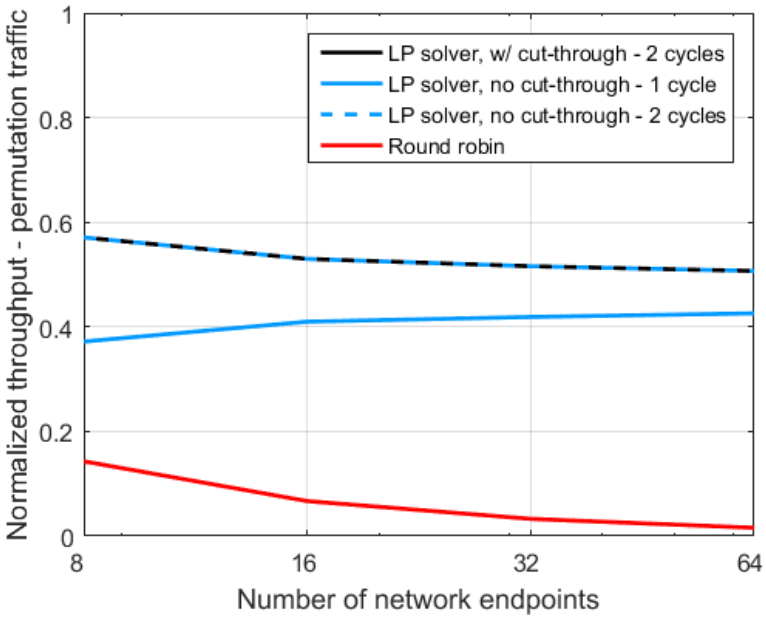

Figure 4.12: Permutation traffic throughput in Rotor-based selector switch. Throughput is normalized to the matching packing factor. The number of network endpoints ranges from 8 to 64. The throughput is shown under round robin (with no indirection), store and forward indirection, and both store and forward and cut-through indirection flow control schemes. See text for details.

network size, we normalized the throughput to the packing factor. We see that for simple round robin routing, the throughput approaches zero as the network scales (scaling as $1/(N-1)$). The LP results show that using cut-through (or only store and forward) indirection significantly increases network throughput. This is due to the load-balancing effect of indirection, which creates the appearance of a more all-to-all type traffic pattern which is well suited to the Rotor-based topology. Further, as the network scales, all flow control approaches using indirection approach approximately 50% throughput relative to a packet switched network.

4.4.2 Permuted Crossover Matchings

The Rotor port matchings discussed above provided nearly full bisection bandwidth for all-to-all traffic patterns (limited by the matching packing factor), and varying degrees of throughput for sparse traffic patterns depending on the flow control mechanism. However, each selector switch must be preconfigured with a unique set of matchings, which may complicate manufacture. Further, the Rotor matching patterns are not compatible with the micro-optic structures used in the design of the highly-scalable selector switch discussed in Chapter 3. The structure of the microlens arrays in that switch were particularly well-suited implement the logarithmic Crossover matching patterns. From a physical layer perspective, it would be desirable to make use of such a

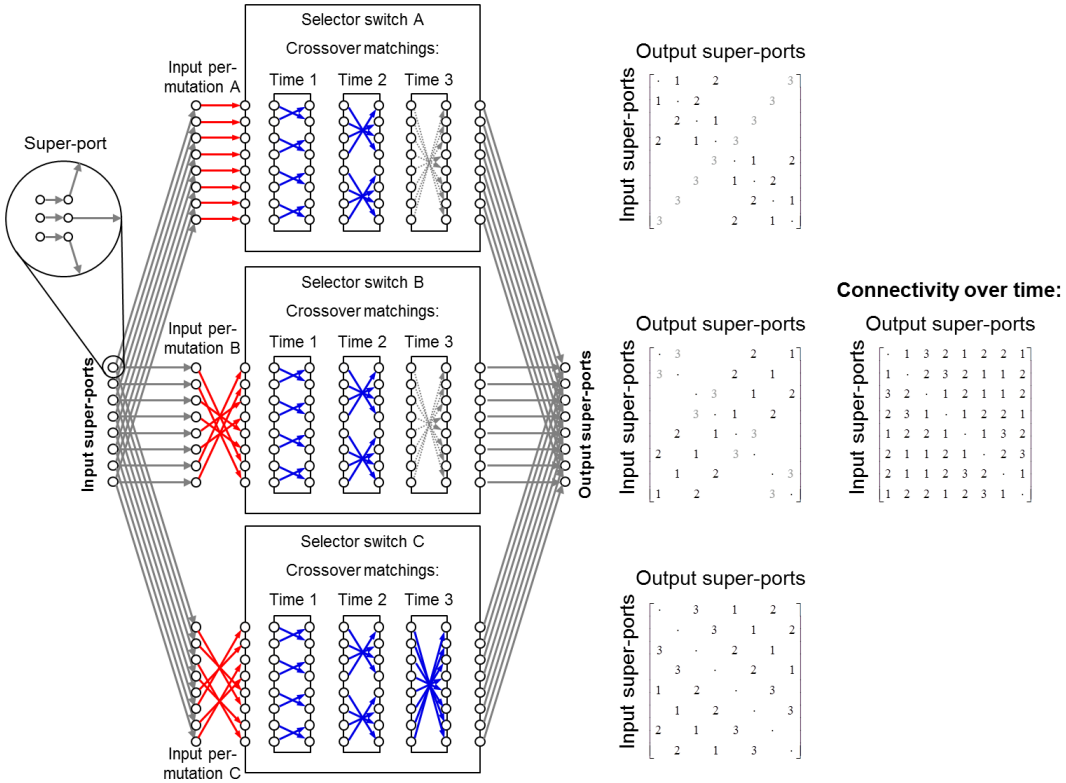

Figure 4.13: Permuted Crossover matching-based selector switch topology. Each selector switch is internally preconfigured with the same logarithmic Crossover matching patterns. The input fibers connected to each switch have been permuted so each switch externally appears to implement a different set of port matchings, which collectively provide complete connectivity between all network endpoints.

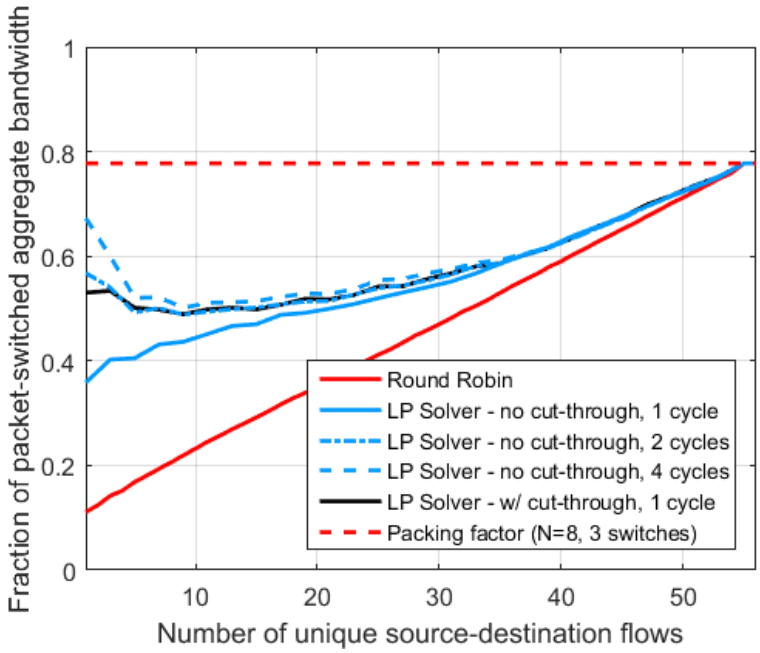

Figure 4.14: Permuted Crossover topology throughput under various flow controls. Modeled average throughput for an 8-port selector switch network using permuted Crossover matchings, relative to a packet switched network. The network uses 3 selector switches each configured with Crossover matchings (see Figure 4.13). The number of flows ranges from one (sparse demand) to 56 (all-to-all demand). Different flow control scenarios are shown, including round robin (with no indirection), store and forward indirection, and both store and forward and cut-through indirection. See text for details.

switch in our network design because of its potential to provide large port count (designed to 2,048 ports), lows loss (2 dB modeled insertion loss), and fast switching (~20 μs).

Fortunately, we can use such switches to construct completely-interconnected network topologies by deliberately permuting the input (or output) fibers connected to each switch. We define the pattern of the permutations so that from the point of view of the network endpoints each switch implements a different set of matching patterns even though their internal matching patterns are identical. Figure 4.13 shows an example 8-port network constructed from 3 selector switches, each preconfigured with Crossover matchings. The permuted adjacency matrices are shown to the right of each switch, and the overall time-multiplexed connectivity is shown to the far right, displaying complete

connectivity between all network endpoints. Note that every connection in each matching is bidirectional (unlike the Rotor matchings). This may be useful in setting up distributed flow control mechanisms because endpoints can perform a "handshake" before communicating data. Many communication protocols, such as Infiniband, require bidirectional channels.

We repeated the throughput modeling analysis discussed in the previous section on the Permuted Crossover network. Figure 4.14 shows the modeled results, again normalized to a fully-provisioned packet switched network. We see that the Permuted Crossover network has very similar properties to the Rotor-based network, with the exception that very sparse traffic patterns have slightly lower throughput when employing cut-through indirection. This is an artifact of the bi-directionality of the matchings. In the Rotor matching set, a single matching which is coprime to the number of endpoints allows any endpoint to be reached in at most $N - 1$ cut-through hops. Because permuted Crossover matchings are bidirectional, a single matching does not have this property. However, we see that store and forward indirection applied over a number of matching cycles recovers bandwidth similar to that observed in the Rotor-based network.

4.4.3 A Distributed Approach to Routing & Flow Control

The previous subsections showed that completely-interconnected selector switch topologies can provide throughput within a factor of two of a crossbar topology for many types of traffic. However, achieving this performance relied on an LP solver-based flow control algorithm with global knowledge of the traffic demand. Such a centralized approach may not scale to support large networks due to the overhead of collecting demand and distributing control signals. Below, we describe a distributed approach to routing and flow control based on the principle of Valiant load balancing (VLB) [69], where each endpoint makes routing decisions based on its local traffic demands. A simple protocol may be needed to provide backpressure for certain traffic patterns.

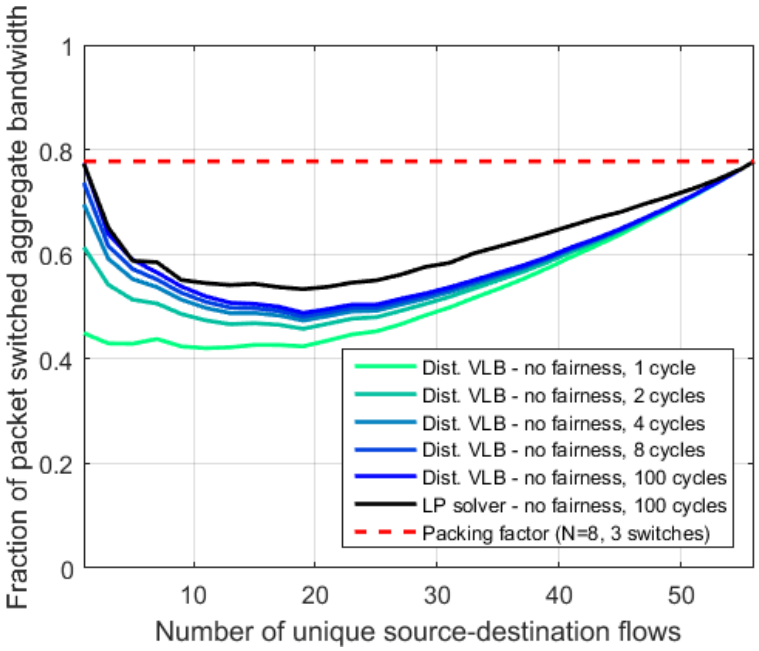

Figure 4.15: Rotor-based selector switch throughput under distributed flow control. Throughput under the centralized LP-based flow control scheme is shown for reference. The throughputs are normalized to that of a packet switched network. The network uses 3 selector switches each configured with Rotor matchings (see Figure 4.9). The number of flows ranges from one (sparse demand) to 56 (all-to-all demand). See text for details.

Our distributed flow control algorithm requires each endpoint maintain two sets of queues: one for *original* traffic generated by that endpoint to each other endpoint, and one for *indirect* traffic being forwarded through that endpoint to each other endpoint. Traffic may only be indirected once (i.e. it may be sent to an intermediate endpoint, but that endpoint must deliver the data to its destination). Indirect traffic is prioritized over original traffic as follows. When a connection is established between endpoints through a selector switch, the sender examines its original and indirect queues corresponding to the currently-connected destination. Any indirected data waiting to be sent to the destination is sent first. Next, if there is a large amount of original data waiting to be sent to the destination, that data is sent directly. Finally, if there is no data waiting to be sent to the current destination, the sender forwards original data destined for other destinations into the indirect queues of the current receiving endpoint in a time-multiplexed fashion. This

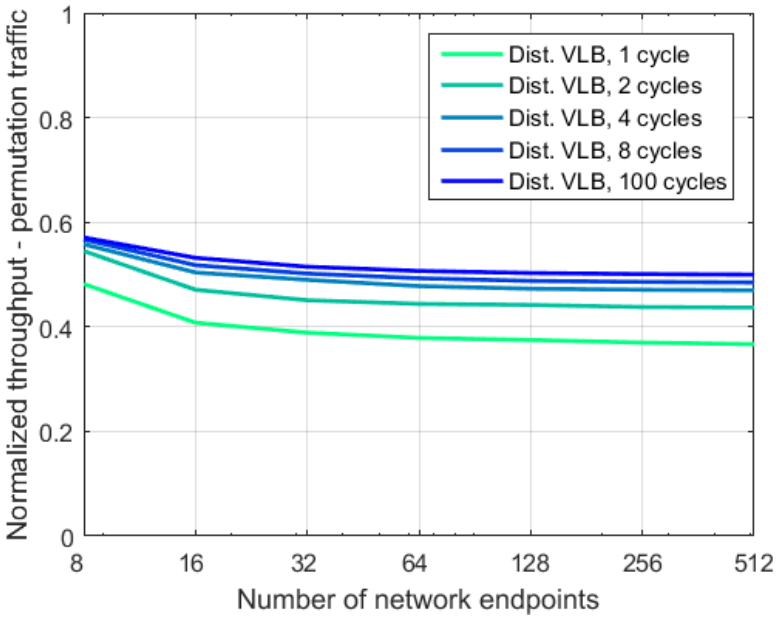

Figure 4.16: Permutation traffic throughput in Rotor-based selector switch under distributed flow control. Throughputs are normalized to the matching packing factor. The number of network endpoints ranges from 8 to 512.

has a load-balancing effect, in that it makes any traffic pattern appear more uniform, allowing it to be more effectively served by completely-interconnected selector switched network.

There is no implicit enforcement of fairness in our distributed algorithm, so we removed the max-min fairness constraints from the LP solver in order to compare the throughput of our distributed approach with that of the packet switch and selector switch with LP-based flow control. The results are shown in Figure 4.15. The throughput of the distributed algorithm approaches that of the theoretical LP solution as the number of matching cycles increases, indicating that our simple and distributed control approach is nearly optimal in terms of throughput.

Unlike the LP solver, the distributed flow control model had a fast run time, allowing us to model the throughputs of networks with more than 64 endpoints. Figure 4.16 shows the throughput of permutation traffic under distributed flow control as the network scales. We chose to study permutation traffic because it simplifies the

throughput calculation of the packet switch (no LP solver required), and because we expect permutation traffic to be adversarial to the selector switched network. The results indicate that the completely-interconnected selector switched network has approximately 50% the throughput of a packet switched network under permutation traffic.

One artifact of truly distributed flow control is that some endpoints may become overloaded with indirect traffic under (statistically unlikely) adversarial traffic conditions. A simple network protocol could prevent this by applying backpressure to senders to prevent them from overfilling the indirect buffers at other endpoints.

4.5 SelecToR Network Architecture

The previous sections established that selector switch based network topologies can provide substantial bandwidth for a number of traffic patterns despite their partial configurability. In this section, we show how a selector switched network, "SelecToR," can be deployed at scale. We show that SelecToR yields a larger bisection bandwidth for the same cost as an electronically-switched network. Equivalently, for the same bisection bandwidth SelecToR reduces cost, cabling complexity, and power consumption.

Given that data center networks in production today contain hundreds of thousands of servers, providing direct connectivity between all servers with a monolithic (optical *or* electronic) switch is impractical. Our goal with SelecToR is to provide high bandwidth connectivity between aggregated groups of servers in a single optically-switched network tier. As long as the aggregation groups are small enough (i.e. don't require multiple electronic switching tiers themselves), the SelecToR architecture can substantially flatten the network relative to an electronically-switched network. In our designs, selector switches connect racks of servers, eliminating multiple tiers of electronic switching and thereby reducing cost and cabling complexity. Drawing a connection to the previous sections, each rack of servers is an *endpoint* with respect to the selector switched network. Before describing the design of SelecToR, we briefly review how conventional electronically-switched data centers are constructed.

Data centers use *racks* to house servers, with one rack typically holding between 30-50 servers. Racks are arranged in rows and columns on the data center floor with cables running in between (or overhead) to connect the racks. Each rack has a so-called "top-of-rack" (ToR) packet switch to which all servers in the rack are connected. The ToR switch serves as a data aggregation point for traffic leaving or entering the rack, as well as a means by which servers within the rack can communicate with each other. The racks are interconnected by a multistage network fabric composed of packet switches and cables connecting those switches. Today's data centers typically use a folded-Clos or "FatTree" topology with multiple switching *tiers* making up the network fabric. By choosing the ratio of upward facing (inter-rack) to downward facing (intra-rack) ports on the ToR switch, an *oversubscription* ratio is defined between the servers within the rack and the rest of the network. With an equal number of upward and downward facing ports, an equal amount of traffic can enter and leave the rack as can be exchanged by servers within the rack. Next, if the entire network fabric were designed to support the required inter-rack bandwidth, all servers in the entire data center could theoretically communicate at full link bandwidth. However, fully-subscribing each rack reduces the number of servers that can be supported by a ToR switch with a given number of ports, and also increases cost and cabling complexity elsewhere in the network. Consequently, data center operators typically oversubscribe ToR switches, leading to reduced cost along with reduced bisection bandwidth.

As an example and baseline for comparison, consider a 24,576-server packet-switched data center with a folded-Clos topology and a 3:1 oversubscription ratio. We assume the packet switches have 32 ports, which is typical for commodity switches. For a 3:1 oversubscription ratio, we allocate 24 downward facing ports and 8 upward facing ports on each ToR switch chip. The ToR on each rack of 48 servers has two switch chips, which together have 16 upward facing ports to the rest of the network and 48 ports facing the servers in the rack. With 48 servers per rack, there are a total of 512 racks. Two additional tiers of 32-port switches are needed above the ToR layer in order to provide the required inter-rack bandwidth. Figure 4.17(a) shows a schematic view of the folded-Clos network.

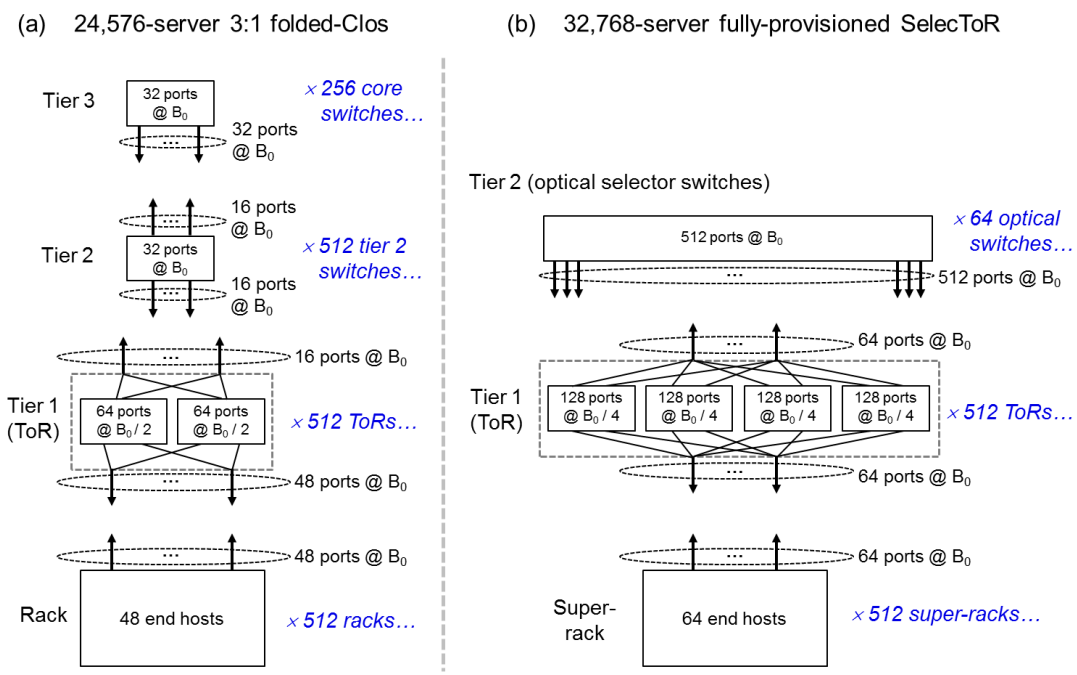

Figure 4.17: Conventional folded-Clos network and proposed SelecToR network. (a) A conventional electronic packet-switched data center network based on a folded-Clos or "FatTree" topology. The network uses 32-port packet switches at bandwidth B_0 per port and supports 24,576 servers with 3 switching tiers. The ToR tier has an oversubscription ratio of 3:1, providing about $1/3^{rd}$ the inter-rack bandwidth of a fully-provisioned network. (b) A 32,768-server SelecToR network, providing fully-provisioned bandwidth between all racks. 64 512-port selector switches are configured with 8 port matchings each. The optical switches provide enough bandwidth to remove all electronic switching tiers above the ToR level.

Figure 4.17(b) shows a 32,768-server SelecToR network. We make the racks slightly larger (64 servers per rack) by using 4 switch chips per ToR, and refer to these larger racks as "super-racks". Depending on the rack space and server size, not all 64 servers may fit into one physical rack. The larger number of switch chips in the ToR switch can be integrated using an electronic backplane [71] or board-integrated interconnect technology such as Intel's Embedded Multi-die Interconnect Bridge (EMIB) [72]. Each ToR in the example SelecToR network is configured with 64 upward and 64 downward facing ports. Each of the upward facing ports on each ToR connects to one of 64 optical selector switches. With 512 super-racks, each selector switch needs 512 ports to connect all the super-racks. Further, to implement a completely-interconnected

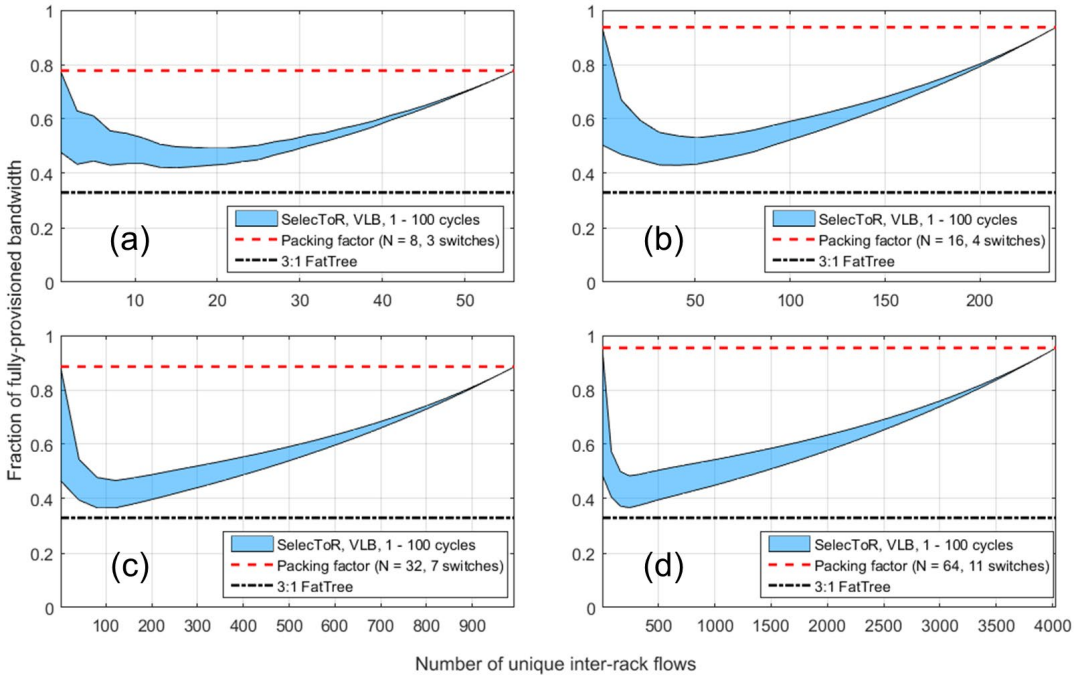

Figure 4.18: Throughput with SelecToR vs. 3:1 FatTree. Throughputs are normalized to a fully-provisioned FatTree. The range of throughputs achievable between 1 and 100 matching cycles are shown for SelecToR. The network scales are (a) 8 racks, (b) 16 racks, (c) 32 racks, and (d) 64 racks. The number of unique inter-rack flows ranges from 1 to N^2-N (all-to-all) in each case.

topology, each selector switch needs to be preconfigured with 8 port matchings. Selector switches at this scale are readily achievable, given that the selector switch designed in Chapter 3 had 2,048 ports and 12 available slots for port matchings.

We could not directly compare the throughputs of SelecToR and the 3:1 FatTree at scale because the run time of the LP solver used to determine the FatTree's throughput was prohibitive past 64 ports (racks). Figure 4.18 shows the inter-rack throughputs of SelecToR and the 3:1 Fat Tree, both normalized to the throughput of a fully-provisioned FatTree. The throughput of networks with 8, 16, 32, and 64 racks are shown. In each case, the number of unique inter-rack flows is swept from 1 to $N_{rack}^2 - N_{rack}$ (all-to-all). The SelecToR throughput is plotted as a range, with the throughput over 1 matching cycle as the lower bound of the range and the throughput over 100 matching cycles as the

Table 4.1: Network components per 1,000 servers and network throughput

Network	Optical transceivers	Switch chips	Optical switches	All-to-all throughput	Permutation throughput
3:1 Folded-Clos (24 k servers)	1,333	73	0	0.333	0.333
SelecToR (32 k servers)	1,000	63	2	1	0.5

upper bound of the range. We expect the performance at scale to look much like that calculated for 64 racks, shown in Figure 4.18(d).

Table 4.1 summarizes the number of required components per 1,000 servers and the throughputs for all-to-all and permutation inter-rack traffic for the folded-Clos and SelecToR networks. The network sizes in our analysis were chosen to provide the fairest comparison possible between the two network architectures, and the relative differences are maintained as the networks scale in size. SelecToR provides higher throughput between more servers than the oversubscribed folded-Clos using a comparable number of components (i.e. for comparable hardware cost). Furthermore, as the per-link data rate increases, the cost of SelecToR scales linearly because the transparent optical selector switches are data rate independent. The cost of the electronic folded-Clos, on the other hand, will begin to scale quadratically as the capacity of electronic packet switches saturates because more switching tiers will be required.

The comparison above highlights just one example of a SelecToR network, and there are other designs that may be considered depending on the deployment scenario. In any SelecToR network, there are subtle design tradeoffs between the number of servers in a super-rack, the number of selector switches, the number of ports per selector switch, and the number of preconfigured port matchings per selector switch. In practice, the design will be driven by the degree of underlying parallelism present in high speed link technologies. 10 Gb/s links are logically addressable as a single channel, but higher speed links today are composed out of multiple underlying channels. For example, a 100 Gb/s link is composed of four 25 Gb/s channels. Future link standards are expected to incorporate even more parallelism. Electronic switches are subject to a similar trend. For example, Broadcom's Tomahawk chip with 3.2 Tb/s switching capacity can be

configured with 32 ports each at 100 Gb/s or 128 ports each at 25 Gb/s. We can exploit this underlying parallelism to electronically aggregate groups of servers using fewer switch chips than would otherwise be required. For example, the super-rack ToR in Figure 4.17 internally splits the link from each server into 4 logically-addressable channels, each at a quarter of the overall link rate. For a switching capacity of $32B_0$ (where B_0 is the nominal server data rate), this allows us to configure each switch chip with 128 ports each running at $B_0 / 4$. Each switch chip has 64 downward and 64 upward facing ports, allowing us to fully provision a rack of 64 servers with four parallel switch chips. We could continue dividing each link and adding switch chips to the ToR to support a larger number of servers in each super-rack. This would reduce the number of super-racks in the network, also reducing the required port count of the selector switches. However, there are practical limits to how many sub-channels a link can be split into as well as how many switch chips can be integrated into a ToR. We leave these details to be determined by future link standards, but expect that those standards will continue to provide higher levels of parallelism. This should allow more servers to be aggregated into super racks, facilitating the design of SelecToR networks.

4.6 Discussion

This chapter investigated how best to use selector switches in overall network architectures. We considered two network topologies: one which used a single selector switch preconfigured with a logarithmic number of port matchings and used indirect routing to recover complete connectivity, and another which used a parallel set of selector switches which collectively contained enough matchings to provide complete, single-hop connectivity. We integrated a prototype 150 μs 16-port selector switch into a 16-server network testbed, and experimentally measured the throughput performance under a centralized scheduling algorithm. We showed that completely-interconnected selector switched networks can achieve throughputs within a factor of two of a packet switched network for a wide variety of traffic patterns. We also discussed a method for

distributed flow control in completely-interconnected selector switched networks. The distributed controller uses load balancing to achieve throughput for sparse or skewed traffic patterns approaching that of an optimal offline LP solver. Finally, we showed how selector switches can be used to interconnect racks of servers at scale in the SelecToR network architecture. SelecToR can provide 2-3× higher throughput than a packet-switched network for similar cost, and scales linearly in cost and complexity to support faster link speeds.

Chapter 4, in part, is being prepared for submission in a paper tentatively titled: "SelecToR: A Scalable, Partially Configurable Data Center Network Architecture," by W. M. Mellette, J. R. McGuinness, A. Forencich, G. Papen, A. Snoeren, J. E. Ford, and G. Porter.

Chapter 5

Conclusion and Future Research Directions

This dissertation presented an investigation into practical optical switches and optically switched networks for data centers. Our approach to realizing networks with better bandwidth scaling properties than current approaches was to design novel network architectures from the ground up, conforming to the properties of a fundamentally scalable optical switch architecture.

Chapter 2 investigated the scaling properties of conventional optical MEMS cross-connect switches using a first-principles physical-layer model. This analysis uncovered the scaling limitations imposed on the physical switching elements by the crossbar switch architecture. While factors such as alignment tolerance, drive voltage, lithographic feature size, and actuator structure had performance impacts, the requirement that each micromirror resolve a unique optical state for each output port was the driver in the tradeoff between port count and response speed.

The physical layer insights gained in Chapter 2 laid the foundation for Chapter 3, which presented a novel switch architecture that improved scalability over the crossbar architecture by limiting the configurability of the switch. This *selector switch* could select port matchings from a small hardware library of preconfigured matchings, fundamentally changing the tradeoff between switching speed and port count. We designed and built a proof-of-principle prototype switch and designed a 2,048-port 20 μs selector switch, demonstrating orders-of-magnitude scaling in response speed and port count over conventional cross-connects.

Chapter 4 investigated how best to use selector switches in overall network architectures. Due to their partial configurability, selector switches cannot be used in conventional networks built around crossbar switches. Two classes of selector switch based topologies were studied: one using a single selector switch preconfigured with a logarithmic number of port matchings coupled with indirect routing to recover complete connectivity, and another using a parallel set of selector switches which collectively contained enough port matchings to provide complete, single-hop connectivity. Analysis showed that a completely-interconnected selector switched network can provide throughput within a factor of 2 of a fully-provisioned packet switch network, but with fewer components, lower cabling complexity, and lower power consumption. We showed an approach to distributed flow control in such networks, and that when deployed at scale, selector switched networks can provide larger aggregate throughput for common communication patterns than conventional approaches for similar cost.

The work presented in this dissertation establishes the framework for a more scalable approach to data center networking based on a novel optical switch architecture. One obvious next step would be to investigate how the switch and network architectures investigated here may benefit electronic switching technologies and other optical switching technologies, such as planar waveguide switches. At first glance, it would appear that the hardware complexity (and cost) of these technologies could be significantly reduced with a shift to a partially configurable switch architecture. Of course, additional work will be required to optimize our approach for deployment into an actual data center. Much of this work will be closely tied to the characteristics of the specific deployment, including the types of applications run on the network, the quality of service requirements, and the other networking hardware present.

Implementation details aside, we expect the approach taken in this dissertation of developing switch and network architectures which balance scalability at the physical layer and performance at the network layer to aid in the design of future optical data center networks.

Appendix A

Unrelated Research Conducted: Planar Waveguide LED Illuminator with Controllable Directionality and Divergence

This appendix is a departure from the topic of data center networking, presenting the author's work on a versatile illumination system with applications in energy-efficient lighting and display. In this system, white light emitting diodes are coupled through a planar waveguide to periodically patterned extraction features at the focal plane of a two dimensional lenslet array. Adjusting the position of the lenslet array allows control over both the directionality and divergence of the emitted beam. We describe an analytic design process, and show optimal designs can achieve high luminous emittance (1.3×10^4 lux) over a 2×2 foot aperture with over 75% optical efficiency while simultaneously allowing beam steering over ± 60° and divergence control from ± 5° to fully hemispherical output. Finally, we present experimental results of a prototype system which validate the design model.

A.1 Introduction

Conventional illumination systems are typically designed to provide either directional or diffuse illumination, spot or flood lighting, using a fixed optical path through collimating or diffusing optics. In settings where the required type of

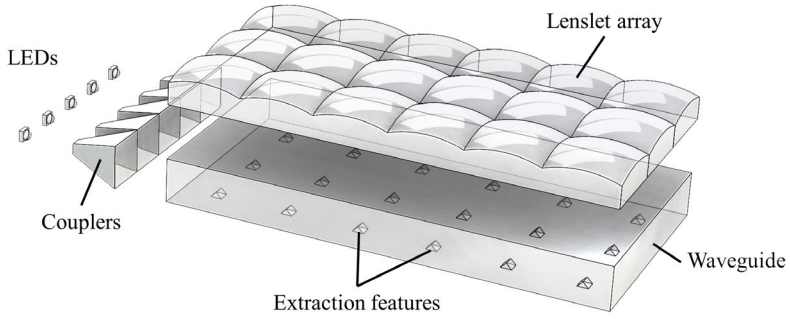

Figure A.1: Conceptual illustration of the planar illumination system. The components have been exploded for clarity.

illumination varies, light energy could be used more efficiently if the source could adapt to provide illumination consistent with the user's immediate need. For example, in home or office lighting the user may want to switch between directional task lighting to illuminate a workspace and diffuse lighting to illuminate an entire room.

Backlights for liquid crystal displays use waveguide illumination, varying the size and shape of features patterned on the light guide plate to control light extraction uniformity [73], and using optical sheets above the light guide to control the directionality of emitted light [74], [75]. Control over directionality allows the display to preferentially direct light into a viewing cone. This viewing cone is fixed, however, because the optical components are designed to provide a single luminance distribution regardless of their relative positioning. Light cannot be actively directed toward an observer moving relative to the device.

Previous work on planar solar concentrators has demonstrated efficient, high-concentration designs that use a two dimensional lens array positioned above a micro-patterned waveguide [76]. The addition of a moveable lens array above the waveguide allows the concentrator to adapt to changing sun angle [77]. The same physical structure can be adapted for a versatile illuminator by reversing the direction of light propagation, and re-optimizing the design for the light source and output constraints.

Figure A.1 shows an illustration of the system in which light emitting diodes (LEDs) are coupled to a planar multimode waveguide such that light is confined by total internal reflection (TIR) defined by Snell's law. As light propagates, it is scattered out of

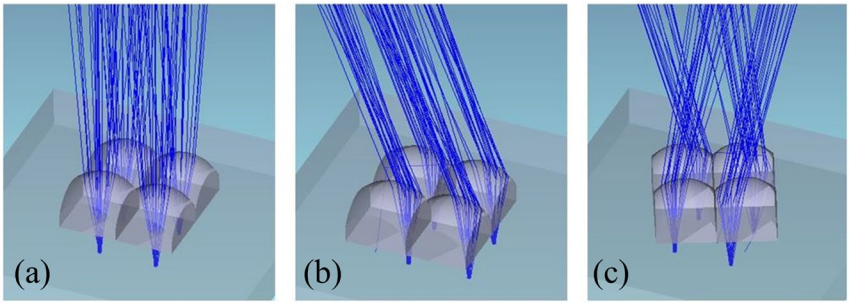

Figure A.2: Section of the array showing a collimated beam when the arrays are aligned (a), a redirected beam when the arrays are translated (b), and a diverging beam when the arrays are rotated (c).

confined modes by periodic extraction features and subsequently interacts with the corresponding lens array, which directs the extracted light toward the target.

Aligning the lenslet and extraction arrays with the extraction features located at or near the focal plane of the lenses produces a collimated output beam (Figure A.2(a)). Laterally translating the lens array relative to the extraction array steers the overall beam by steering all individual beams in the same direction, as shown in Figure A.2(b). Relative rotations between the two arrays alter the overall divergence of the beam by steering the individual beams in a 'spiral' of different directions, as shown in Figure A.2(c). In Figure A.2 the divergence angle of the light extracted from the waveguide has been restricted, because lateral offsets between the arrays would otherwise induce unwanted crosstalk as light spills into adjacent lenses. This crosstalk leads to side lobes in the emitted pattern, which are undesireable for most applications.

The same functionality can be achieved using an array of point-like LED sources directly behind the lens array, which would eliminate the complexity of edge coupling and waveguiding. However, a waveguide-based design has the advantages that it 1) allows a thinner form factor and simplifies electrical routing and heat sinking by moving the LED sources to the edges of the waveguide; 2) clears the aperture opposite to the lens array from LEDs, wiring, and heat sinks, allowing the use of higher performing reflective lenses, discussed in Section A.2.1; and 3) allows the coupling, waveguiding, and extraction structures to perform the necessary angular and spatial mapping of the real sources into an effective array of point-like sources. While the efficacy (electrical to

luminous conversion efficiency) and emittance (spatial power density) of LED dies typically scale inversely with die size within one class of LEDs, so-called ‘high power’ LEDs with apertures larger than 2mm currently have higher performance in terms of emittance than do small package LEDs with apertures less than 1mm. From conservation of radiance, edge coupling a smaller number of high power LEDs will produce a brighter beam than a large number of small LEDs located directly behind the lens array. This edge coupling approach will be adaptable as LED technology improves, up to the point when the emittance of small aperture LEDs matches that of large aperture LEDs, which would warrant the direct array approach.

The thin form factor of the planar illuminator allows conformal mounting to flat surfaces with little or no recessing, making it ideal for retrofitting ceiling fixtures. Further, control over light from a relatively large aperture can be achieved with relatively short range mechanical motion compared to traditional designs. Control over a similar amount of light energy would require an array of traditional luminaires, with each element having its own actuation mechanism. Conventional actuation mechanisms require motion in 3 dimensions, either by moving a lens with radial and axial freedom with respect to the source or by gross actuation of the entire luminaire including the source and heat sink. The planar illuminator uses precise short-range 2D motion of one optical component to achieve the same degree of control.

In the following section we present an analytic model of each element of the system, then in Section A.3 combine the elements to obtain an overall system model to determine the potential performance of optimal designs. In Section A.4 we describe an experimental full-scale ‘proof of principal’ prototype, and compare its performance to the model. We conclude in Section A.5 and discuss future directions of this technology.

A.2 System Design

Typical performance metrics for illumination systems include optical efficiency, efficacy, luminous emittance, and pattern uniformity. In our system, we are also

concerned with the beam steering and divergence ranges conditional on the degree of crosstalk between adjacent lenses. We would also like the system to scale efficiently to large aperture sizes for high flux applications. Here we describe a simple analytic model for each element of the system, beginning at the output where we discuss lens performance, then moving to waveguiding and extraction, and finishing with the source and coupling methods.

A.2.1 Beam Steering and Diverging

The maximum steering angle, minimum divergence angle, and degree of crosstalk of emitted light are driven by two parameters: the lenslet F/# (focal length over aperture diameter) and the divergence of light exiting the waveguide. From geometrical optics, using the paraxial lens approximation, the maximum steering angle with zero geometrical crosstalk is given by:

$$\psi_{\max} = \sin^{-1}\left(n \sin\left(\tan^{-1}\left(\frac{1}{2(F/\#)} - \tan(\theta_2) \right) \right) \right), \tag{A.1}$$

where θ_2 is the half divergence angle of the effective source immersed in refractive index n. Maximizing the steering angle corresponds to minimizing the lens F/# and the divergence angle of the effective source. Also from geometrical optics, we can write the minimum divergence angle due to the spatial extent of the effective source as:

$$\varphi = \sin^{-1}\left(n \sin\left(\tan^{-1}\left(\frac{w_{facet}}{2f} \right) \right) \right), \tag{A.2}$$

where w_{facet} is the full width of the effective source and f is the focal length of the lens. For a small minimum divergence angle, corresponding to a tightly collimated output beam, the lateral extent of the source needs to be small with respect to the focal length of the lens.

In the waveguide solar concentrator, light illuminates the entire face of the lenslets and lenslet aberrations are a critical factor in design. However, for an illuminator it is not necessary to emit from the entire surface area, and illuminating only a fraction of

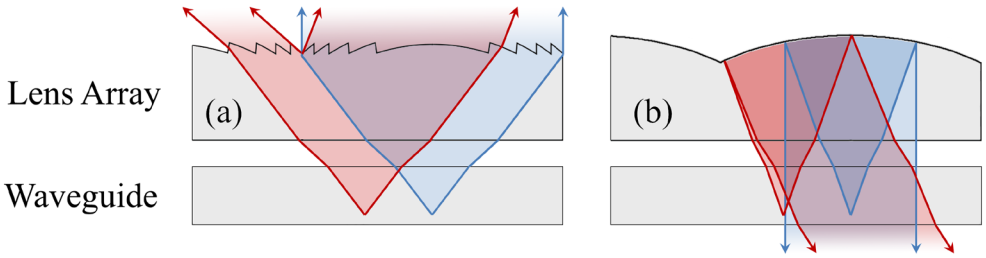

Figure A.3: Lens geometry examples: (a) fully filled refractive Fresnel lens showing crosstalk with lateral translation and (b) partially filled reflective spherical lens showing zero crosstalk with equivalent translation and F/#.

the lens area can be useful to minimize lateral crosstalk (Figure A.3). Lens aberrations affect the performance of the system to the extent that they increase beam divergence. Under-filled lenses contribute fewer aberrations because light only interacts with a localized section of the lens surface. Reflective plano-convex singlets produce lower F/#s than do refractive designs for the same radius of curvature and, consequently, can be driven to lower overall F/#s [77]. Fresnel lenses are a viable option to reduce the F/# of refractive lenses while simultaneously reducing weight, but low F/# Fresnel lenses typically have poor off-axis performance due to increased scatter from zone transitions. Shorter focal length lenses are desirable because aberrations scale with lens dimensions [78] and because they make the illumination pattern more uniform by the nature of having more lenses per unit area. In some designs, it may be beneficial to induce a small fixed defocus by tuning the axial height of the lens in order to blur or 'smooth out' any sharp features present in an otherwise perfectly imaged intensity distribution.

A.2.2 Light Guiding and Extraction

The extraction features act as the effective sources for the lenses by intercepting and redirecting light propagating in the waveguide toward the lens array. Light may be extracted from the waveguide using reflection, refraction, diffraction, or diffuse scattering. Flat faceted features are desirable because they have broadband performance (unlike dispersive gratings) and conserve angular divergence (unlike diffusers or curved

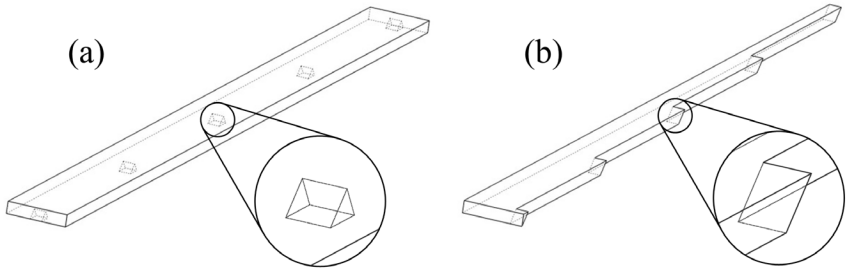

Figure A.4: Constant (a) and stepped (b) mode volume waveguide illustrations for $N = 5$ extraction sites. Each section as drawn supplies light to one row of lenses above the waveguide (not shown).

facets). The conservation of angular divergence is crucial for minimizing crosstalk and generally keeps the system more étendue-limited, leading to more efficient designs.

The waveguide confines light by TIR for a sufficient angular spectrum, allowing light to be efficiently distributed to the extraction sites. The type of waveguide determines the relationship between the waveguide thickness and the dimensions of extraction features. We considered two waveguide designs. One is a constant cross section and *constant mode volume* (CMV) waveguide (Figure A.4(a)) where light is shared between extraction sites, and the other is a laterally tapered *stepped mode volume* (SMV) waveguide (Figure A.4(b)) where each extraction site adiabatically truncates the modal volume [79].

In the SMV design, light makes a single pass through the structure and is extracted uniformly up to a factor determined by the material's absorption coefficient. There is a fixed relationship between the facet and waveguide dimensions given by:

$$w_{facet} = \frac{t_{wg}}{\tan\gamma} = t_{wg}\Big|_{\gamma=45^\circ}, \tag{A.3}$$

where t_{wg} is the waveguide thickness and γ is the angle the facet makes with respect to the waveguide plane. Without loss of generality we set $\gamma = 45°$, corresponding to the case where the average direction of guided propagation is in the plane of the waveguide. Altering this γ will necessitate a split in the angular spectrum (e.g. ± 30° out-of-plane propagation), which does not increase the total radiance in the guide, makes confinement more difficult, and tends to require more complicated coupling structures. We should

also note here that the stepped waveguide has a geometrical relationship limiting its length given the size and number of facets, as will be discussed in Section A.3.2.

In the CMV geometry, light makes multiple passes through the waveguide and extraction is fundamentally non-uniform. We model the percentage of light energy extracted at a facet as the ratio between the facet cross section and the waveguide cross section. This model ignores shadowing effects, which is valid when the divergence is relatively large and the facets are relatively small with respect to their period. First, we determine the facet cross section 'σ_f,' which is the cross sectional area of the facet seen by the average waveguide mode. By the reasoning presented above for the SMV waveguide, we set the facet angle γ to 45°. Constraining the base dimensions of the facet to be square ($w_{facet} \times w_{facet}$) to produce a symmetric beam using a rotationally symmetric lens, the facet cross section is just the product of the facet width and height, where the height is half the width: $\sigma_f|_{\gamma=45°} = w_{facet}^2 / 2$. We then write the distributed absorption and extraction per lens aperture as:

$$\chi = \left(1 - \frac{\sigma_f}{t_{wg} D}\right) \exp(-\alpha D), \tag{A.4}$$

where D is the full lens aperture and α is the absorption coefficient of the waveguide material. Modifying the Beer-Lambert law, where j runs from 1 to N facets, the output power at the j^{th} facet is given by:

$$P_{ext,j} = P_0 \frac{\sigma_f}{t_{wg} D} \cdot \frac{\left(\chi^{j-1} + \eta_2 \chi^{2N-j}\right)}{\left(1 - \eta_1 \eta_2 \chi^{2N}\right)}, \tag{A.5}$$

where P_0 is the power coupled into the waveguide, η_2 and η_1 are the reflection efficiencies from the end of the waveguide and the source, respectively, and N is the total number of extraction sites in the section of waveguide. By symmetry, we consider a section of waveguide that is one lens aperture wide and half the total system aperture long, taking $\eta_2 = 1$ and $\eta_1 = \eta_{coupler}^2 R_{LED}$, where $\eta_{coupler}$ is the coupler efficiency (discussed in Section A.2.3) and is modeled as being equivalent in both forward and reverse directions and R_{LED} is the percentage of light recycled by the LED. The incident light recycled by a typical die is about 50% [80] and the phosphor efficiency can be as

high as 70% per pass [81]. The total recycling efficiency can be approximated by two passes through the phosphor and one reflection from the die, which gives 25% total recycling efficiency. The total extracted power can be determined by evaluating the sum:

$$P_{ext,total} = \sum_{j=1}^{N} P_{ext,j} = P_0 \frac{\sigma_f}{t_{wg} D(\chi - 1)} \cdot \frac{\left(\chi^N - 1\right)\left(1 + \eta_2 \chi^N\right)}{\left(1 - \eta_1 \eta_2 \chi^{2N}\right)}, \tag{A.6}$$

where we consider the term on the right hand side which scales the input power P_0 to be the average extraction efficiency 'η_{ext},' referred to later in Section A.3. In the CMV geometry the relationship between waveguide and facet dimensions is:

$$w_{facet} < 2t_{wg}, \tag{A.7}$$

for $\gamma = 45°$ in order for the facet to fit within the waveguide. Here, unlike for the SMV waveguide, there is no fixed geometrical relationship between facet geometry, number of facets, and waveguide length.

Recalling from (A.2) that minimizing the divergence of emitted light corresponds to minimizing w_{facet}, we find that by the geometry of the SMV waveguide in (A.3) and by the desire for high extraction efficiency in (A.6), we would like to minimize the waveguide thickness 't_{wg}' in both cases.

A.2.3 Light Sources and Couplers

White LEDs currently have superior luminance and efficacy compared to other broadband sources. From conservation of radiance, the brightness at the output of any passive optical system is limited by the brightness of the source. Consequently, LEDs with the highest luminance are desirable because they provide more optical power with the same étendue. These 'high power' LEDs have die sizes exceeding 2mm in width and typically obey Lambert's cosine law, leading us to calculate the fraction of Lambertian power in a beam of half angle θ_1 to be:

$$\eta_{beam} = \sin^2\left(\theta_1\right). \tag{A.8}$$

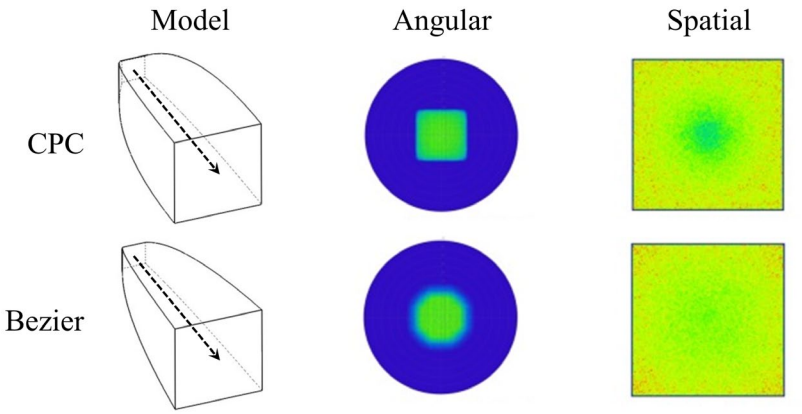

Figure A.5: Angular and spatial output distributions for a conventional CPC and a Bezier collimator both with a uniform Lambertian input.

For example, a Lambertian emitter output clipped at $\theta_1 = \pm 71.65°$ still contains 90% of the total power. Having such a clearly defined beam divergence simplifies étendue calculations.

From the above and per Sections A.2.1 and A.2.2, a high system performance requires coupling large sources with a high divergence angle to a relatively thin waveguide, while minimizing the divergence and maximizing the spatial power density of coupled light. For high optical efficiency the design must conserve étendue. Approaches to solving similar problems have recently been proposed [82], [83]. Our approach was to first collimate the source, allowing a tradeoff between divergence and spatial power density, and then perform a space-variant aperture transformation to interface with the thin waveguide.

The compound parabolic concentrator (CPC) is a standard nonimaging optical component that provides nearly étendue limited concentration and (path-reversed) collimation (Figure A.5, top row) [84]. However, any spatial non-uniformity in the collimated output intensity distribution reduces the uniformity of the waveguide illuminator output. Following previous work [85], we defined a CPC-like collimator with enhanced spatial uniformity at the output using quadratic Bezier curves (Figure A.5, bottom row).

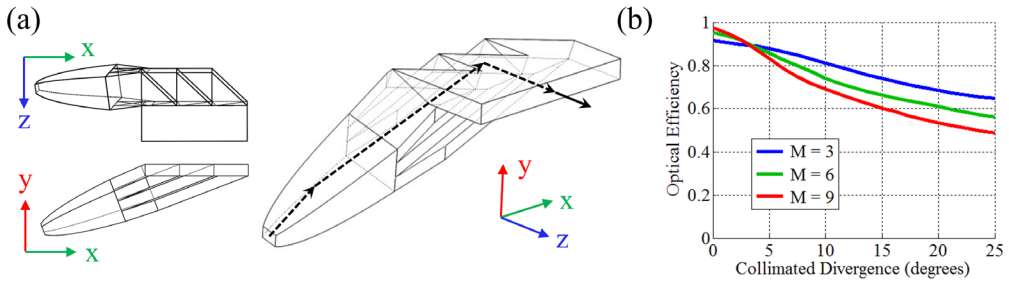

Figure A.6: Wireframe models of faceted coupler with M = 3 segments (a) and corresponding optical efficiency for M = 3, 6, and 9 segments (b).

We can approximate both collimator designs as conserving étendue, so for two square apertures:

$$h_1 \sin(\theta_1) = h_2 \sin(\theta_2), \tag{A.9}$$

where h_1 and h_2 are the full widths of the source and exit apertures and θ_1 and θ_1 are the half divergence angles of light entering and exiting the collimator, respectively.

Next, we consider two designs to transform the exit aperture of the collimator to interface with the waveguide: 'faceted' and 'curled'. Both designs are variants of a stepped mode volume structure where the change in aspect ratio '*M*' from collimator to waveguide is equal to the number of segments:

$$M = \frac{h_2}{t_{wg}}, \tag{A.10}$$

where, as in (A.9), h_2 is the full width of the output aperture of the collimator. The first design uses a series of flat reflective rectangular facets acting like fold mirrors to sequentially redirect segments of light exiting the collimator into the waveguide (Figure A.6(a)). The structure was designed assuming perfectly collimated light and then analyzed in nonsequential Zemax to determine performance as a function of divergence (Figure A.6(b)). A perfect aperture mapping can be achieved using two reflective facets per segment. Our final faceted design used a single facet per segment to reduce complexity and reflective surface loss, because this imperfect mapping approaches the ideal mapping as the aspect ratio *M* increases.

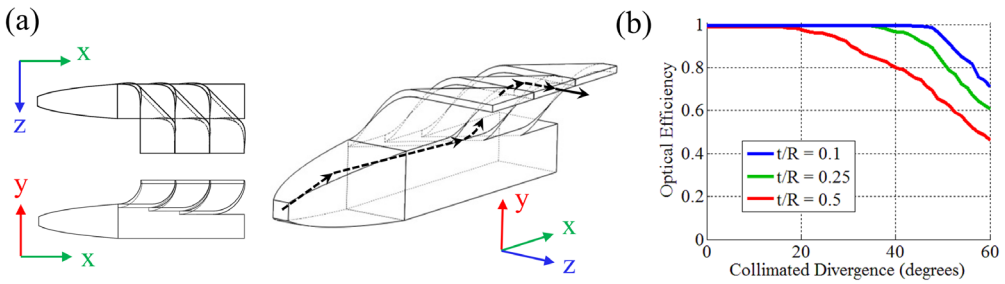

Figure A.7: Wireframe models of curled coupler showing 3 segments (a) and corresponding optical efficiency for a few ratios of t / R (b). The efficiency is independent of aspect ratio.

The 'curled' coupler design we considered uses adiabatic light propagation through curved waveguide sections to 'strip' light energy and transform the aperture (Figure A.7(a)). Following previous work on the confinement properties of curved multimode waveguides by conformal mapping [86], it can be shown that the half divergence angle 'θ_0' incurred from interaction with the curved structure is related to the thickness of the waveguide 't' and the outer bend radius 'R' by:

$$\theta_0 = \cos^{-1}\left(1 - \frac{t}{2R}\right). \qquad \text{(A.11)}$$

For small ratios of t / R, the structure preserves étendue and has nearly equivalent confinement properties to a flat waveguide of the same refractive index. The blue curve in Figure A.7(b) for t / R = 0.1 has nearly 100% optical efficiency up to a half divergence angle of about 46°, compared to the 47.8° TIR angle corresponding to a flat guide with an equal index of 1.49. Unlike the faceted coupler, the optical efficiency of the curled structure is independent of the aspect ratio M. While the curled coupler outperforms the faceted design in terms of optical efficiency, it is less readily manufacturable. It is possible that advances in optical 3D printing technologies will enable inexpensive fabrication of such structures in the future. At present, flexible Corning Willowglass [87] presents a possible fabrication option.

As the aspect ratio M increases, the 'staircase' shaped intermediate aperture in the faceted design (Figure A.6(a), shown with M = 3) approaches a square, as in the curled design (Figure A.7(a)), considerably simplifying the geometry. The efficiency and ease

of manufacture of these couplers will increase as sources with higher luminance and smaller apertures become available through advances in LED technology or other alternatives [88].

A.3 System-level Analytic Model and Optimization

System-level optimization of the planar illuminator is difficult in standard optical design software because of the complex geometries and merit functions. We developed an analytic model based on equations from imaging and nonimaging optics to give an intuitive optimization approach that provided more confidence than a 'black box' method. The designs resulting from the analytic optimization were modeled in Solidworks and ray traced with non-sequential Monte Carlo analysis using Zemax to insure the accuracy of the analytic model. A truly 'optimal' solution is predicated on a detailed list of application-specific constraints and performance metrics. Without the information needed for a quantitative merit function, we optimized according to qualitative ideas of well-balanced performance.

We constrained certain aspects of the design space using parameters from commercially available LEDs and from a comparison lighting fixture. For a comparison fixture, we considered a 2×4 foot 3-tube fluorescent modular ceiling 'troffer' fixture with a luminous flux of 9000 lm, an efficacy of 92.19 lm/W, and an emittance of 1.475×10^4 lux at the aperture. This gave us a target emittance value independent of system aperture size. We chose to set the system aperture to 2×2 feet with the intent of retrofit compatibility with modular ceiling grids. For the waveguide LED source we chose to use the Cree XLamp XM-L2, one of the highest luminous emittance and efficacy single-die LED sources available, delivering 728 lumens at 2A, 3V (about 2/3 max current) in a 2.5×2.5 mm die size. Low-loss BK7 glass was used for the waveguide for its low absorption coefficient of 3×10^{-4} m^{-1} [89].

From conservation of energy, we can relate the luminous emittance 'I_{out}' to the luminous flux of the LED 'P_{LED}' by:

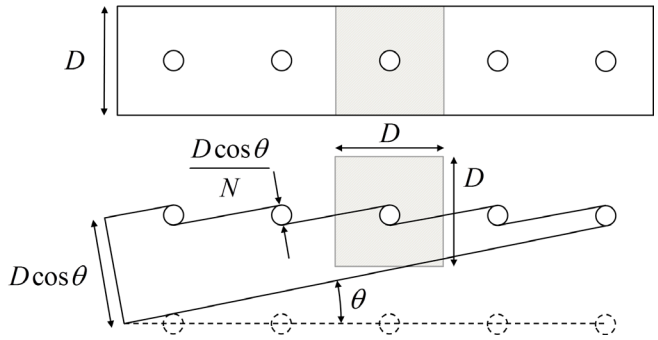

Figure A.8: Top down views of the CMV (top) and SMV (bottom) waveguides with $N = 5$ extraction sites. The grey squares indicate the position and size of a single lens.

$$I_{out} = \eta_{beam}(\theta_1)\eta_{coupler}(M,\theta_2)\eta_{ext}(\sigma_f, D, t_{wg}, N, \chi, \eta_1, \eta_2)\frac{t_{wg}\cos\theta}{NDh_2^{\,2}}P_{LED}, \quad \text{(A.12)}$$

where η_{ext} is the term that scales P_0 in the right hand side of (A.6) and θ is the step angle of the waveguide, as shown in Figure A.8. The second to last term in (A.12) encompasses the ratio between the output area of the coupler and the input area of the waveguide while scaling the output power by the output aperture to convert to emittance.

In the subsequent sections, we consider designs that allow us to solve (A.12) and determine overall system performance. The first, using a constant mode volume waveguide and *faceted* light coupler (CMV-F), is chosen to provide the simplest path to manufacture. The second, using a stepped mode volume waveguide and *curled* coupler (SMV-C), is intended to enable the highest optical performance. We also briefly summarize a third design using a constant mode volume waveguide and curled coupler (CMV-C).

A.3.1 Design 1: Constant Mode Volume with Faceted Coupler

The first design aims for manufacturability at the cost of performance by using the faceted coupling structure and a constant mode volume waveguide. The coupler is

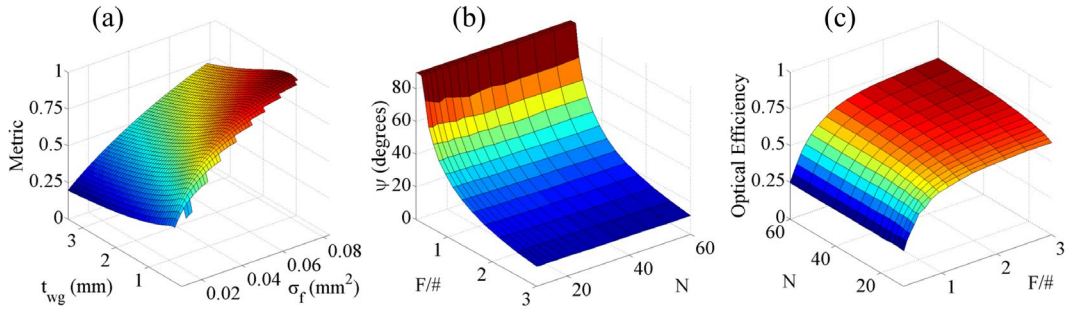

Figure A.9: CMV-F design space for 25% of target emittance. (a) Optimization metric for $N = 60$, F/# = 0.75. (b) Maximum beam steering angle in {F/#, N} space. (c) Optical efficiency in {F/#, N} space. Note that the axes are rotated 90° counterclockwise from (b) to (c) to clearly illustrate the data.

compatible with injection molding and the waveguide with roll processing of glass or plastic sheets.

First, we fit a parameterized 2 dimensional function to the simulated faceted coupler efficiency curves shown in Figure A.6(b). The mathematical form of the function was approximated from knowledge of the shape and boundary conditions of the simulated curves to be:

$$\eta_{coupler}\left(M,\theta_2\right)=\frac{f_1\left(M,\vec{A}_1\right)}{f_1\left(M,\vec{A}_2\right)\theta_2+f_1\left(M,\vec{A}_3\right)}+\frac{f_2\left(M,\vec{A}_4\right)}{\left(f_2\left(M,\vec{A}_5\right)\theta_2\right)^2+f_2\left(M,\vec{A}_6\right)}, \quad \text{(A.13)}$$

where:

$$f_1\left(M,\vec{A}_i\right)=A_{i,1}M^2+A_{i,2}M+A_{i,3}, \quad \text{(A.14)}$$

$$f_2\left(M,\vec{A}_i\right)=\frac{A_{i,1}}{A_{i,2}M+A_{i,3}}, \quad \text{(A.15)}$$

where the 3-element fit vectors $\vec{A}_1$ through $\vec{A}_6$ are determined by least squares minimization. The resulting parametric function is used in the optimization algorithm to give a predicted optical efficiency of the coupler in regimes that were not explicitly simulated beforehand.

From (A.9) and (A.12), setting $\theta = 0$ for the CMV waveguide geometry, we arrive at an implicit transcendental equation for θ_2:

$$\frac{h_1^2 \sin^2\theta_1}{\sin^2\theta_2} = \eta_{beam}(\theta_1)\eta_{coupler}(M,\theta_2)\eta_{ext}(\sigma_f, D, t_{wg}, N, \chi, \eta_1, \eta_2)\frac{t_{wg}P_{LED}}{I_{out}ND}, \quad \text{(A.16)}$$

where we recast $M = (h_1 \sin\theta_1) / (t_{wg} \sin\theta_2)$ using (A.9) and (A.10) so that the optimization problem is constrained to 4 dimensions: $\{t_{wg}, \sigma_f, \text{F/\#}, N\}$, with the remaining variables fixed by design constraints. The optimization algorithm maps the design space by iterating through these 4 dimensions and numerically solving (A.16) over a grid of points in the space. For each point in {F/#, N} space, an optimal point in $\{t_{wg}, \sigma_f\}$ space is found by maximizing a weighted sum of normalized maximum steering angle and normalized system efficiency (Figure A.9(a)). The maximum steering angle is given by (A.1) and the overall optical system efficiency is the product of all efficiency terms in (A.16). We discarded solutions for which the minimum half divergence angle in (A.2) is greater than a design limit of 5° and for which extraction deviation is greater than 1%, where the deviation is given by $\max_j \{|P_{ext,total} - NP_{ext,j}| / P_{ext,total}\}$ using (A.5) and (A.6).

Figures A.9(b) and A.9(c) show the corresponding optimums mapped from $\{t_{wg}, \sigma_f\}$ to {F/#, N} space. There is a clear tradeoff between efficiency and maximum steering angle, which also depends on the target emittance. Higher emittance values drive both the maximum steering angle and efficiency down. High emittance requires a low aspect ratio M to maintain a high spatial power density, which either requires a thick waveguide or a small intermediate aperture from (A.10). To maintain the same minimum divergence angle for the same lens F/# when the waveguide is made thicker, the facet dimension must be held constant (see (A.2)), meaning the extraction efficiency decreases from (A.6). The other alternative, shrinking the intermediate aperture h_2, means that for the same beam efficiency given in (A.8), the divergence angle of coupled light increases in (A.9), which both lowers the maximum steering angle in (A.1) and lowers the coupler efficiency (Figure A.6). Similar balancing forces are present when trying to push the maximum steering angle or the optical system efficiency as well.

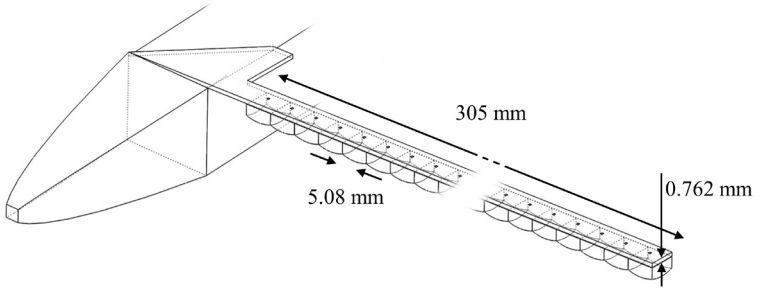

Figure A.10: Single section wireframe model of optimal CMV-F design.

Sweeping emittance values from 1 to 1/10 that of the target value (1.475×10^4 lux), we found that the performance metrics were balanced at about 1/4 of the reference emittance (3.69×10^3 lux). Using this value, we choose an 'optimal' faceted design with N = 60, F/# = 0.75, t_{wg} = 0.762 mm, and σ_f = 0.0762 mm^2 (Figure A.10). This design provided a good tradeoff between efficiency, steering angle, and emittance. Achieving such a low F/# required the use of a reflective lens array.

The physical structure was modeled in Solidworks and imported into Zemax for ray trace analysis. The full system has a 2×2 foot aperture consisting of 120×120 lenslets and 4 source LEDs. The model consisted of a full 3 dimensional structure where rays were stored after being traced through the coupler and re-launched into the waveguide to save repetitive tracing through the coupler. A sufficient number of rays were traced to achieve ergodicity. The far field directionality was simulated as a function of lateral offset (Figure A.11(a)) and the divergence as a function of rotation about the center of the array (Figure A.11(b)). The collimated beam can be steered ± 45° maintaining over 35% optical efficiency, and can be diverged from ± 5° to ± 60° maintaining about 43% optical efficiency. Most of the loss comes from the faceted coupler, which has a relatively large aspect ratio of M = 22. We see good agreement between the analytic model, which assumes a top-hat beam intensity profile characterized by ψ and φ, and the Zemax simulation in Figure A.11(a).

Higher efficiencies can be reached if the minimum divergence requirement is relaxed, as this enables a reduction in the aspect ratio of the coupler, an increase in waveguide thickness, and a corresponding increase in facet size. This allows coupler

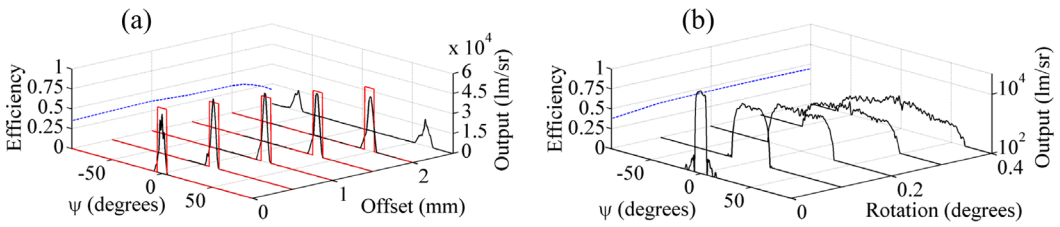

Figure A.11: Far field directivity (a) and divergence (b) simulations of the optimal CMV-F design, with total optical efficiency plotted on the left-hand plane (dashed blue). Part (a) shows good agreement between the Zemax (black) and analytic (red) models. Part (b) shows the Zemax model (black) on a log scale.

efficiency to be increased without reducing extraction efficiency. Similarly, relaxing the uniformity requirement increases the extraction efficiency, which also increases overall system efficiency.

A.3.2 Design 2: Stepped Mode Volume with Curled Coupler

The second design considered uses the light coupling and waveguide structures that may be challenging to fabricate, but offer the maximum efficiency and uniformity. Based on the results of Section A.2.3, we can assume nearly 100% coupling between the LED and waveguide using the curled coupler. This can be achieved for a small enough ratio of t / R independent of aspect ratio and divergence. The fixed relationship between waveguide thickness and facet geometry in (A.3) allows us to write a determined set of relationships describing the geometry of the stepped structure:

$$\theta = \cos^{-1}\left(2N\left(F/\#\right)\tan\varphi\right), \tag{A.17}$$

$$\tan\theta = \frac{N-\cos^2\theta}{N\left(N-1\right)+\cos\theta\sin\theta}, \tag{A.18}$$

$$t_{wg} = \frac{D\cos\theta}{N}, \tag{A.19}$$

where θ is the step angle of the SMV structure (Figure A.8), which decreases with increasing N.

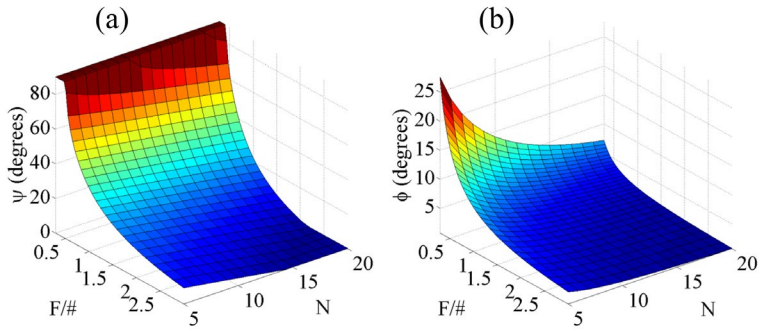

Figure A.12: SMV-C design space for 100% of the target emittance. (a) Maximum steering angle and (b) minimum beam divergence angle, constrained to {F/#, N} space.

Using (A.1), (A.9), (A.12), and (A.19), we can express the maximum steering angle as:

$$\psi_{\max} = \sin^{-1}\left(n \sin\left(\tan^{-1}\left(\frac{1}{2(F/\#)} - \tan\left(\sin^{-1}\left(\frac{h_1 N}{\cos\theta} \sqrt{\frac{I_{out}}{\eta_{coupler} P_{LED}}} \right) \right) \right) \right) \right). \quad \text{(A.20)}$$

During optimization, we iterate through {F/#, N} space, first solving the transcendental equation defined by (A.17) and (A.18) for φ and then for θ, then we solve (A.20) to determine the performance metric. Due to the fixed relationships between the waveguide and extraction feature geometries, the space is constrained to 2 dimensions (Figure A.12). The efficiency is independent of F/# and N and is only determined by (A.8) and parasitic Fresnel losses which were not considered in the anaylic model.

This design benefits greatly from a nearly ideal coupling structure and extraction mechanism. The 1.475×10^4 lux target emittance could be met while retaining a useful portion of the design space. We chose an optimal design with $N = 20$, F/# = 0.5, and t_{wg} = 0.761 mm (Figure A.13). Like the CMV-F design, this design also used a reflective lens array to achieve the necessary F/#. This yielded a predicted maximum steering angle of ±60° and a minimum divergence angle of about ±5°.

The full system has a 2×2 foot aperture consisting of 40×40 lenslets and 6 source LEDs. We used the same modeling technique discussed in Section A.3.1 to simulate the system performance. The result of the Zemax simulations, shown in Figure A.14, confirm that the system can steer the beam ±60° while maintaining over 75% optical

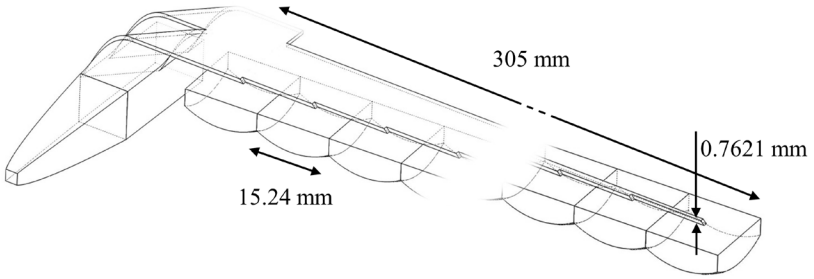

Figure A.13: Single section wireframe model of optimal SMV-C design.

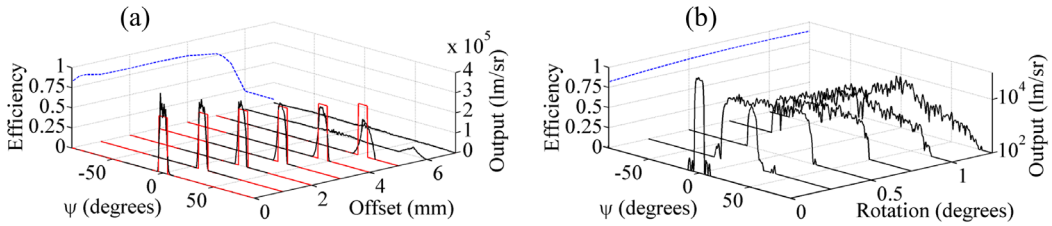

Figure A.14: Far field directivity (a) and divergence (b) simulations of optimal SMV-C design, with total optical efficiency plotted on the left-hand plane (dashed blue). Part (a) shows good agreement between the Zemax (black) and analytic (red) models. Part (b) shows the Zemax model (black) on a log scale.

efficiency and diverge the beam from ± 5° to essentially hemispherical illumination maintaining about 80% optical efficiency. The main source of loss in this design came from Fresnel reflections. To reach higher efficiencies the optics could be anti-reflection coated, at an increased manufacturing cost.

A third design using a constant mode volume waveguide with a curled coupler (CMV-C) was optimized and simulated and occupied a middle-ground between the previously discussed CMV-F (35% optical system efficiency) and SMV-C (75% optical system efficiency) designs in both manufacturability and performance. The optimal CMV-C design emitted 1.22×10^4 lux and could steer the beam ±60°, operating above 62% optical system efficiency, and could diverge the beam from ±5° to hemispherical illumination.

The final step in the design was to compare the overall light emission for the optimized SMV-C design to a benchmark LED troffer fixture. The far field polar

intensity information for the waveguide system was exported from Zemax into Dialux [90] to simulate the illumination pattern in a realistic environment. The result is shown in Figure A.15. The conventional LED fixture (Figure A.15(a)) has a 2×2 foot aperture, consumes 53W, and produces 4000 lm with a nearly Lambertian pattern. The optimized SMV-C design (Figures A.15(b)–A.15(d)) also has a 2×2 foot aperture, consumes 52.84 W, but produces 4800 lm output. The waveguide design can create a similar diffuse illumination distribution (Figure A.15(b)) when configured with a 1° rotation between the lens and extraction arrays. The unique capability of the waveguide system is shown in Figures A.15(c) and A.15(d), in which a collimated spot is steered to each desk in the room, producing a spot more than 10× brighter than any point in the previous two illuminance distributions. Since the LED output level can be controlled, the waveguide system can provide localized task lighting with lower energy consumption.

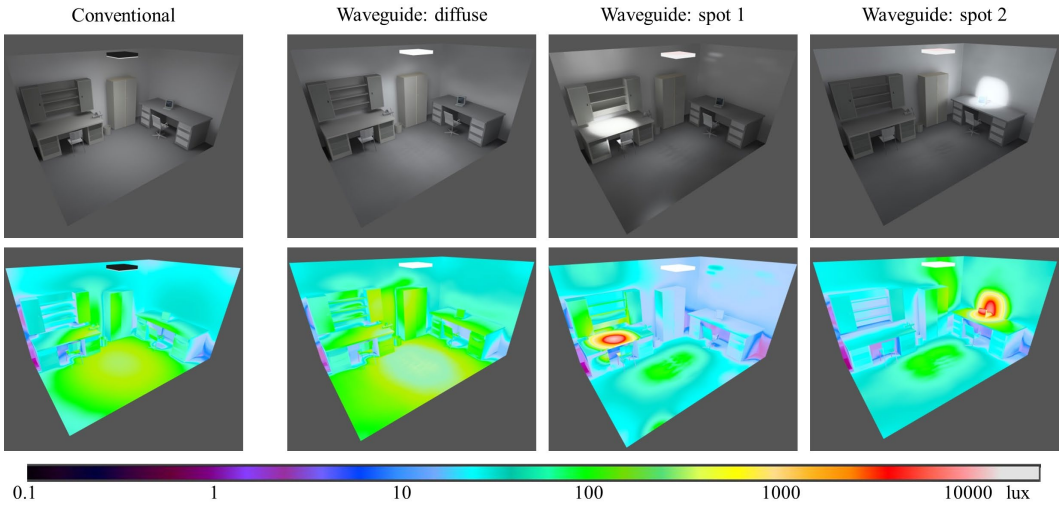

Figure A.15: Dialux simulations of conventional 2×2 foot LED fixture and optimized SMV-C design. The waveguide system was simulated in three configurations: [diffuse] 1° rotation, [spot 1] (Δx, Δy) = (-3, 3) mm, and [spot 2] (Δx, Δy) = (5, 0) mm.

A.4 Prototype Fabrication and Characterization

The modeled systems in Section A.3 used optimized components to achieve high system performance. To demonstrate the concept and compare model with measurement, we constructed a prototype system using commercially available or easily fabricated components. Because alignment tolerances scale with component size, the physical scale of parts was the driving factor in determining our choice of components.

We used F/1.04 refractive Fresnel lenses molded from poly methyl methacrylate (PMMA) available in 4×4 arrays measuring 3×3 inches. To reduce F/# and increase steering range we increased the lens power by stacking two lens layers for a final F/0.7 lens, measured in the PMMA waveguide. The Fresnel lenses were oriented so that the grooved sides were both facing away from the source. For the extraction features, we used 1mm diameter steel ball bearings epoxied into hemispherical recesses machined into the waveguide. The spherical symmetry of the bearings translates into relaxed alignment tolerance and a higher degree of repeatability compared to flat facets, which would require precise 3 dimensional alignment. The spatial extent of the 1mm diameter hemispheres gives a 3.2° half divergence angle of emitted light. For the waveguide, we used a 2.54 mm thick planar sheet of PMMA, where the thickness was chosen to produce uniform and efficient extraction. A 10.6 mm thick PMMA substrate was glued to the bottom of the lens array to minimize the air gap between the waveguide and lens structure while keeping the total optical distance between lens and extraction feature equal to the focal length. We found that an air gap of 100-300 μm between the lenses and waveguide was sufficient to minimize undesirable divergence, and could be achieved using a small number of thin Teflon spacers distributed across the system aperture.

The curled and faceted couplers discussed previously provide a relatively collimated and axially symmetric angular spectrum, which is ideal for use with flat facets. However, when using spherical extraction features, there is no need for the illumination to be collimated or axially symmetric due to the scattering properties of a sphere. From an étendue perspective, the spheres are more efficiently illuminated by light with a larger divergence angle and a higher spatial power density. Additionally, the

extraction efficiency of spherical facets was found to increase when light propagates with a large average angle with respect to the waveguide plane, so long as the TIR condition is obeyed. Based on these observations, we used a linear array of closely-spaced 0.43 mm thick LEDs attached to a 1-D CPC to reduce the divergence in the plane normal to the waveguide while allowing full divergence in the plane of the waveguide. The CPC bar was attached to the waveguide at a 36° angle with respect to the waveguide plane. The CPC couplers were machined out of polycarbonate and vapor polished to produce a specular surface finish, and later sputtered with 1 micron thick silver reflector (measured to be >85% efficient) to increase reflectivity in regions of the CPC that were not TIR limited. The LEDs were chosen for their thin form factor, allowing adequate collimation defined by the 1-D étendue relation, and for their high flux of 4.38 lm from a 2.3×0.3 mm aperture. The LEDs were reflow-soldered onto a printed circuit board (PCB) while using an alignment fixture machined from FR-4 to register the LEDs to about 200 μm positional tolerance. This tight alignment tolerance allowed efficient interface with the CPC coupler.

A.4.1 Unit Cell Device

Prior to fabrication of a full 2×2 foot aperture system, we constructed a 'unit cell' consisting of a waveguide with a single 1 mm hemispherical extraction feature, a small section of the lens array, and 3 LEDs (Figure A.16(a)). The lens array was mounted onto

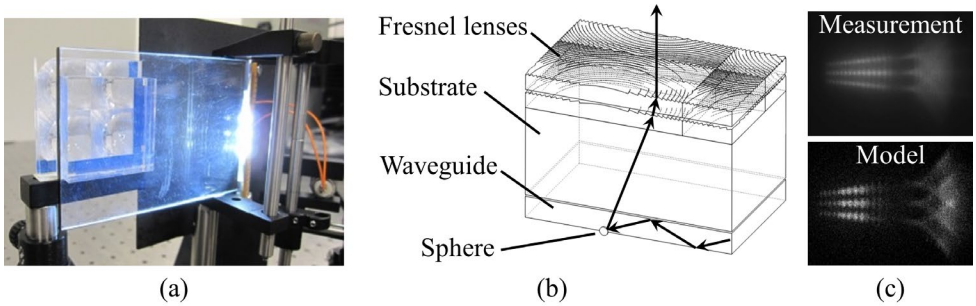

Figure A.16: (a) Unit cell system. (b) Cut-away schematic drawn to scale and illustrative ray path. (c) Measured (top) and simulated (bottom) far field intensity patterns.

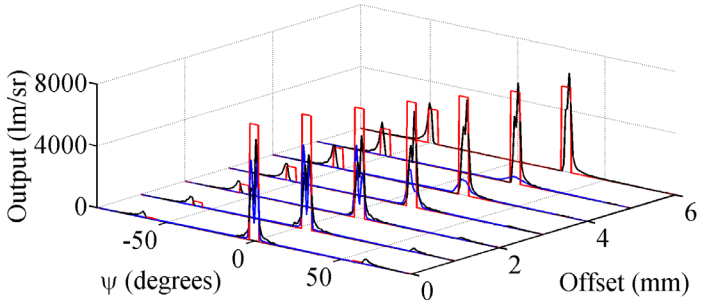

Figure A.17: (a) Far field directivity of the unit cell system: analytic model (red), Zemax simulation (black), and lab measurement (blue). Measured drop in off-axis intensity is due to poor off-axis lens performance.

a 3-axis translation stage for accurate positioning relative to the waveguide. The far field intensity pattern was measured 1 meter from the lens aperture. The intensity pattern is a superposition of 3 patterns from the 3 LEDs, with some fine structure because the coupled waveguide modes had not fully homogenized before striking the facet. An equivalent system was modeled in Zemax and its corresponding far field pattern shows excellent agreement with measurement (Figure A.16(c)).

The unit cell system was also used to characterize the directional capabilities of the system by taking intensity line scans 1 meter from the aperture for different lateral offsets between the lens array and extraction feature (Figure A.17). The data is plotted against curves from a corresponding polar far field Zemax simulation of a full 2×2 foot aperture system (black) and a modified semi-analytic version of the CMV model discussed in Section A.3.1 (red). The measured data (blue) is scaled to arbitrary units because the output power of the full aperture system cannot be directly inferred from the unit cell device. We also cannot determine the divergence capabilities because only one lens/extraction feature pair is present. We see relatively good agreement between both models and measurement, with the exception that the measured off-axis intensity falls dramatically compared to either model. The attenuation is significant at high field angles and completely eliminates the crosstalk lobe seen in both the analytic and Zemax models. This inconsistency can be explained by the poor off-axis Fresnel lens performance compared to the ideal paraxial lens used in both models.

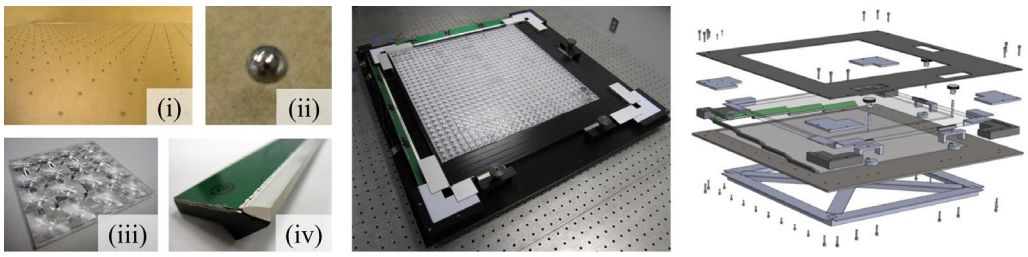

Figure A.18: (a) System components: (i) waveguide, (ii) ball-bearing extraction feature, (iii) lenses, and (iv) PCB, LEDs, and CPC coupler; (b) assembled system (shown without cover); and (c) exploded CAD model.

A.4.2 Full Aperture System

Next we fabricated a full 2×2 foot aperture prototype composed of a 26×26 element extraction array and 28×28 lens array, both with a 19mm pitch, and 304 source LEDs. The lens array was larger than the extraction array to prevent clipping at the corners during rotation. Light was coupled into the waveguide from two edges, allowing room for mechanical control from the opposite edges. We attached high strength neodymium magnets to the lens array at 3 points on the edges opposite to the sources and used ferromagnetic eccentric cams seated on the magnets to translate and rotate the lens array relative to the extraction array. Rotation of the cam through a 180° angle produced the 20 mm travel required for operation. Our prototype used manual control, but could easily be fitted with motorized controllers to enable remote electrical operation. The computer-aided-design (CAD) model as well as the physical realization of the system components and fully assembled system is shown in Figure A.18.

Qualitative (Figure A.19) and quantitative (Figure A.20) measurements were taken 3 meters from the system aperture using a camera and calibrated photodiode, respectively, demonstrating good agreement with both the semi-analytic and Zemax models. The top-hat profile beam calculated with the semi-analytic model was mapped from polar far field space to physical space using simple radiometric calculations. The scattering of light from Fresnel zone transitions accounts for the main discrepancy

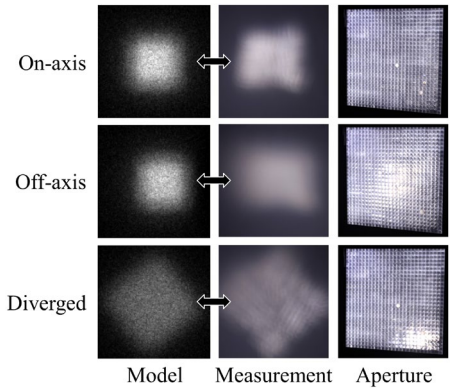

Figure A.19: Simulation (left column) and measurement (center column) of on-axis, off-axis, and diverged spots 3 meters from the aperture. The right column shows the corresponding view of the aperture from an angle.

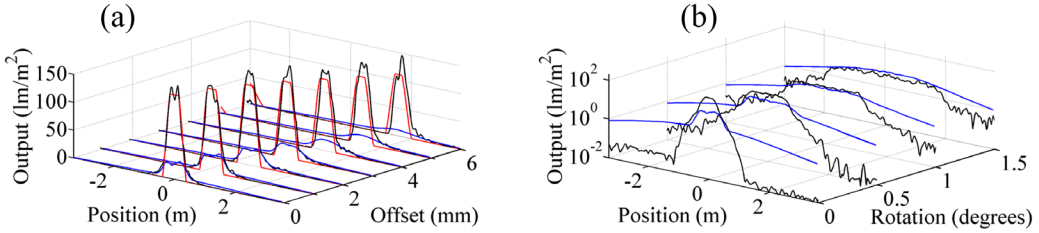

Figure A.20: Near field directionality (a) and divergence (b) of the prototype system 3 meters from the aperture. Part (a) shows the analytic model (red), Zemax model (black), and measurements (blue). Part (b) shows the Zemax model (black) and measurement (blue) on a log scale.

between model and measurement. From lens cross section measurements the zone transitions were estimated to obscure about 30% of the clear lens aperture, accounting for the reduction in central beam power and resultant increase in the noise pedestal surrounding the beam. This effect becomes more pronounced as the beam is steered to more extreme angles. This also explains the behavior observed for extreme rotations, where we find the system acts more like a diffuse emitter instead of preferentially 'spreading out' the light according to the Zemax model.

Polar integration of the illuminance line scan measurements yields a total output of 98 lm, corresponding to an optical system efficiency of 7.6%, which agrees well with the simulated optical efficiency of 7.56%. The major source of loss in the prototype came

Table A.1: System Efficiencies and Loss Mechanisms

Design	System efficiency	Dominant sources of loss
SMV-Curled	75%	Fresnel reflections from uncoated interfaces
CMV-Curled	62%	+ Imperfect extraction in the CMV waveguide
CMV-Faceted	35%	+ Suboptimal coupler efficiency
Lab prototype	7.6%	+ Large material absorption of PMMA waveguide

from the high absorption coefficient of the PMMA waveguide, measured and simulated to be 0.5 m^{-1}.

Zemax simulations showed that using a BK7 waveguide with an absorption coefficient of 3×10^{-4} m^{-1} (used in the optimized theoretical designs) would increase the overall optical system efficiency of the prototype to 31%. Secondary sources of loss in the prototype were coupling mirror loss, waveguide surface scattering, and small misalignments in the couplers and lens array. While the prototype system is highly inefficient compared to optimal designs, the consistency between measurement, model, and simulation indicates that the predicted high efficiencies for optimized designs (Table 1) are credible. This agreement also supports the accuracy of the analytic model in representing the system during design and optimization.

A.5 Summary

We showed how a planar waveguide illuminator with periodically patterned extraction features and lens array can be used to control both the directionality and divergence of light output using short-range mechanical motion.

The system performance depends on a large number of variables, which led us to develop an analytic model compatible with the two coupling and two waveguiding designs considered in order to perform system-level optimization. The analytically optimized designs were ray traced in Zemax and the resulting performance was in good agreement with the analytic model. We found that the optimal design used a stepped mode volume glass waveguide and curled coupler. This design could steer a collimated beam over ± 60° and diverge the beam from ± 5° to fully hemispherical illumination,

while maintaining over 75% optical efficiency, for a total output of 4800 lumens from a 2×2 foot aperture.

We constructed a proof-of-principle prototype from commercially available components which successfully demonstrated both the beam steering and diverging principle in a 2×2 foot aperture embodiment. Although the optical efficiency of the device was only 7%, good agreement between the measurement, Zemax simulation, and analytic model was established, supporting the predictions of high efficiency and high output power in optimal designs which used fully custom optical components. The next step would be to fabricate an efficient system using the optimized optical structures, and using electrical controllers to allow remote actuation.

In future research, the same basic concept could be extended to provide a thin energy efficient flat panel display where light energy is actively directed toward one or more users, whose position may be tracked using a video camera and face-tracking software. Given accuracy sufficient to selectively illuminate each of the user's eyes, this approach may be used for multi-user glasses-free 3D display. This research was made possible with support from CogniTek. The authors would also like to thank Dr. Ilya Agurok for helpful discussions.

Appendix A, in full, reprints material as it appears in the paper titled: "Planar waveguide LED illuminator with controlled directionality and divergence," published in *Optics Express*, 22(S3), pp A742-A758, 2014, by William M. Mellette, Glenn M. Schuster, and Joseph E. Ford.

Bibliography

[1] H. Liu, F. Lu, A. Forenchich, R. Kapoor, M. Tewari, G. Voelker, G. Papen, A. Snoeren and G. Porter, "Circuit Switching Under the Radar with REACToR," in *11th USENIX Symposium on Networked Systems Design and Implementation*, 2014.

[2] H. Liu, M. Mukerjee, C. Li, N. Feltman, G. Papen, S. Savage, S. Seshan, G. Voelker, D. Anderson, M. Kaminsky, G. Porter and A. Snoeren, "Scheduling Techniques for Hybrid Circuit/Packet Networks," in *ACM CoNEXT*, 2015.

[3] S. Bojja, M. Alizadeh and P. Viswanath, "Costly Circuits, Submodular Schedules and Approximate Caratheodory Theorems," in *SIGMETRICS*, 2016.

[4] N. Farrington, G. Porter, S. Radhakrishnan, H. Bazzaz, S. V., Y. Fainman, G. Papen and A. Vahdat, "Helios: A Hybrid Electrial/Optical Switch Architecture for Modern Data Centers," in *ACM SIGCOMM*, 2010.

[5] G. Wang, D. Anderson, M. Kaminsky, K. Papagiannaki, T. Ng, M. Kozuch and M. Ryan, "c-Through: Part-time Optics in Data Centers," in *ACM SIGCOMM*, 2010.

[6] K. Chen, A. Singla, A. Singh, K. Ramachandran, L. Xu, Y. Zhang, X. Wen and Y. Chen, "OSA: An Optical Switching Architecture for Data Center Networks With Unprecedented Flexibility," *IEEE/ACM Transactions on Networking,* vol. 22, no. 2, pp. 498-511, 2014.

[7] G. Porter, R. Strong, N. Farrington, A. Forencich, P. Chen-sun, T. Rosing, Y. Fainman, G. Papen and A. Vahdat, "Integrating microsecond circuit switching into the data center," in *ACM SIGCOMM*, 2013.

[8] N. Farrington, A. Forencich, G. Porter, P.-C. Sun, J. E. Ford, Y. Fainman, G. Papen and A. Vahdat, "A Multiport Microsecond Optical Circuit Switch for Data Center Networking," *IEEE Photonics Technology Letters,* vol. 25, no. 16, pp. 1589-1592, 2013.

[9] N. Hamedazimi, Z. Qazi, H. Gupta, V. Sekar, S. Das, J. Longtin, H. Shah and A. Tanwer, "Firefly: A reconfigurable wireless data center fabric using free-space optics," in *ACM SIGCOMM*, 2014.

[10] M. Ghobadi, R. Mahajan, A. Phanishayee, N. Devanur, J. Kulkarni, G. Ranade, P.-A. Blanche, H. Rastegarfar, M. Glick and D. Kilper, "ProjecToR: Agile Reconfigurable Data Center Interconnect," in *ACM SIGCOMM*, 2016.

[11] R. Stabile, A. Albores-Mejia and K. Williams, "Monolithic active-passive 16 x 16 optoelectronic switch," *Optics Letters,* vol. 37, no. 22, pp. 4666-4668, 2012.

[12] B. Lee, A. Rylyakov, W. Green, S. Assefa, C. Baks, R. Rimolo-Donadio, D. Kuchta, M. Khater, T. Barwicz, C. Reinholm, E. Kiewra, S. Shank, C. Schow and Y. Vlasov, "Monolithic Silicon Integration of Scaled Photonic Switch Fabrics, CMOS Logic, and Device Driver Circuits," *Journal of Lightwave Technology,* vol. 32, no. 4, pp. 743-751, 2014.

[13] K. Suzuki, K. Tanizawa, T. Matsukawa, G. Cong, S. Kim, S. Suda, M. Ohno, T. Chiba, H. Tadokoro, M. Yanagihara, Y. Igarashi, M. Masahara, N. S. and H. Kawashima, "Ultra-compact 8 x 8 strictly-non-blocking Si-wire PILOSS switch," *Optics Express,* vol. 22, no. 4, pp. 3887-3894, 2014.

[14] T. Seok, N. Quack, S. Han, R. Muller and M. Wu, "Large-scale broadband digital silicon photonic switches with vertical adiabatic couplers," *Optica,* vol. 3, no. 1, pp. 64-70, 2016.

[15] J. Kim, C. Nuzman, B. Kumar, D. Lieuwen, J. Kraus, A. Weiss, C. Lichtenwalner, A. Papazian, R. Frahm, N. Basavanhally, D. Ramsey, V. Aksyuk, F. Prado, M. Simon, V. Lifton, H. Chan, M. Haueis, A. Gasparyan, H. Shea, S. Arney, C. Bolle, P. Kolodner and R. Ryf, "1100 x 1100 Port MEMS-Based Optical Crossconnect With 4-dB Maximum Loss," *Photonics Technology Letters,* vol. 15, no. 11, pp. 1537-1539, 2003.

[16] M. Wu, O. Solgaard and J. Ford, "Optical MEMS for Lightwave Communication," *Journal of Lightwave Technology,* vol. 24, no. 12, pp. 4433-4454, 2006.

[17] A. Mekis, S. Gloeckner, G. Masini, A. Narasimha, T. Pinguet, S. Sahni and P. D. Dobbelaere, "A Grating-Coupler-Enabled CMOS Photonics Platform," *IEEE Journal of Selected Topics in Quantum Electronics,* vol. 17, no. 3, pp. 597-608, 2011.

[18] A. Vahdat, H. Liu, X. Zhao and C. Johnson, "The Emerging Optical Data Center," in *OFC Conference*, 2011.

[19] N. Farrington, Y. Fainman, H. Liu, G. Papen and A. Vahdat, "Hardware Requirements for Optical Circuit Switched Data Center Networks," in *OFC Conference*, 2011.

[20] A. Singh, J. Ong, A. Agarwal, G. Anderson, A. Armistead, R. Bannon, S. Boving, G. Desai, B. Felderman, P. Germano, A. Kanagala, J. Provost, J. Simmons, E. Tanda, J. Wanderer, U. Holzle, S. Stuart and A. Vahdat, "Jupiter Rising: A Decade of Clos Topologies and Centralized Control in Google's Datacenter Network," in *ACM SIGCOMM*, 2015.

[21] "100G CWDM4 MSA Technical Specifications," [Online]. Available: http://www.cwdm4-msa.org/wp-content/uploads/2015/12/CWDM4-MSA-Technical-Spec-1p1-1.pdf.

[22] C. Clos, "A Study of Non-blocking Switching Networks," *Bell System Technical Journal,* vol. 32, no. 2, pp. 406-424, 1953.

[23] L. Lu, L. Zhou, Z. Li, X. Li and J. Chen, "Broadband 4x4 Nonblocking Silicon Electrooptic Switches Based on Mach-Zehnder Interferometers," *IEEE Photonics Journal,* vol. 7, no. 1, pp. 1-8, 2015.

[24] Q. Cheng, A. Wonfor, J. Wei, R. Penty and I. White, "Low-Energy, High-Performance Lossless 8x8 SOA Switch," in *OFC Conference*, 2015.

[25] K. Tanizawa, K. Suzuki, M. Toyama, M. Ohtsuka, N. Yokoyama, K. Matsumaro, M. Seki, K. Koshino, T. Sugaya, S. Suda, G. Cong, T. Kimura, K. Ikeda, S. Namiki and H. Kawashima, "Ultra-compact 32 x 32 strictly-non-blocking Si-wire optical switch with fan-out LGA interposer," *Optics Express,* vol. 23, no. 13, pp. 17599-17606, 2015.

[26] A. Roy, H. Zeng, J. Bagga, G. Porter and A. Snoeren, "Inside the social network's (datacenter) network," in *ACM SIGCOMM*, 2015.

[27] Y. Xie, S.-S. Li, Y.-W. Lin, Z. Ren and C. Nguyen, "1.52-GHz micromechanical extensional wine-glass mode ring resonators," *IEEE transactions on ultrasonics, ferroelectrics, and frequency control,* vol. 55, no. 4, pp. 890-907, 2008.

[28] R. Knipe, "Challenges of a digital micromirror device: Modeling and design," in *SPIE Lasers, Optics, and Vision for Productivity in Manufacturing I*, 1996.

[29] R. Syms, "Scaling laws for MEMS mirror-rotation optical cross connect switches," *Journal of Lightwave Technology,* vol. 20, no. 7, pp. 1084-1094, 2002.

[30] U. Krishnamoorthy, D. Lee and O. Solgaard, "Self-aligned vertical electrostatic combdrives for micromirror actuation," *Journal of Microelectromechanical Systems,* vol. 12, no. 4, pp. 458-464, 2003.

[31] G. Nielson and G. Barbastathis, "Dynamic pull-in of parallel-plate and torsional electrostatic MEMS actuators," *Journal of Microelectromechanical Systems,* vol. 15, no. 4, pp. 811-821, 2006.

[32] H. Ishii, M. Urano, Y. Tanabe, T. Shimamura, J. Yamaguchi, T. Kamei, K. Kudou, M. Yano, Y. Uenishi and K. Machida, "Fabrication of optical microelectromechanical-system switches having multilevel mirror-drive electrodes," *Japanese Journal of Applied Physics,* vol. 43, no. 9A, pp. 6468-6472, 2004.

[33] J. Tsai and M. Wu, "Gimbal-less MEMS two-axis optical scanner array with high fill-factor," *Journal of Microelectromechanical Systems,* vol. 14, no. 6, pp. 1323-1328, 2005.

[34] I. Jung, U. Krishnamoorthy and O. Solgaard, "High fill-factor two-axis gimbaled tip-tilt-piston micromirror array actuated by self-aligned vertical electrostatic combdrives," *Jounal of Microelectromechanical Systems,* vol. 15, no. 3, pp. 563-571, 2006.

[35] F. Pardo, R. Cirelli, E. Ferry, W. Lai, F. Klemens, J. Miner, C. Pai, J. Bower, W. Mansfield, A. Kornblit, T. Sorsch, J. Taylor, M. Baker, R. Fullowan, M. Simon, V. Aksyuk, R. Ryf, H. Dyson and S. Arney, "Flexible fabrication of large pixel count piston-tip-tilt mirror arrays for fast spatial light modulators," *Microelectronic Engineering,* vol. 84, pp. 1157-1161, 2007.

[36] D. Hah, H. Toshiyoshi and M. Wu, "Design of electrostatic actuators for MOEMS applications," in *Symposium on Design, Test, Integration, and Packaging of MEMS/MOEMS*, 2002.

[37] D. Lee and O. Solgaard, "Pull-in analysis of torsional scanners actuated by electrostatic vertical combdrives," *Journal of Microelectromechanical Systems,* vol.

17, no. 5, pp. 1228-1238, 2008.

[38] R. Sattler, F. Plotz, G. Fattinger and G. Wachutka, "Modeling of an electrostatic torsional actuator: Demonstrated with an RF MEMS switch," *Sensors and Actuators A,* vol. 97/98, pp. 337-346, 2002.

[39] C. Pu, S. Park, P. Chu, S. Lee, M. Tsai, D. Peale, N. Bonadeo and I. Brener, "Electrostatic actuation of three-dimensional MEMS mirrors using sidewall electrodes," *Journal of Selected Topics in Quantum Electronics,* vol. 10, no. 3, pp. 472-477, 2004.

[40] U. Krishnamoorthy, K. Li, K. Yu, D. Lee, J. Heritage and O. Solgaard, "Dual-mode micromirrors for optical phased array applications," *Sensors and Actuators A,* vol. 97/98, pp. 21-26, 2002.

[41] H. Shea, A. Gasparyan, H. Chan, S. Arney, R. Frahm, D. Lopez, S. Jin and R. McConnell, "Effects of electrical leakage currents on MEMS reliability and performance," *IEEE Transactions on Device and Materials Reliability,* vol. 4, no. 2, pp. 198-207, 2004.

[42] H. Urey, C. Kan and W. Davis, "Vibration mode frequency formulae for micromechanical scanners," *Journal of Micromechanics and Microengineering,* vol. 15, pp. 1713-1721, 2005.

[43] M. Bao and H. Yang, "Squeeze film air damping in MEMS," *Sensors and Actuators A,* vol. 136, no. 1, pp. 3-27, 2007.

[44] D. Greywall, P. Busch and J. Walker, "Phenomenological model for gas-damping of micromechanical structures," *Sensors and Actuators A,* vol. 72, no. 1, pp. 49-70, 1999.

[45] P. Chu, I. Brener, C. Pu, S.-S. Lee, J. Dadap, S. Park, K. Bergman, N. Bonadeo, T. Chau, M. Chou, R. Doran, R. Gibson, R. Harel, J. Johnson, C. Lee, D. Peale, B. Tang, D. Tong, M. Tsai, Q. Wu, W. Zhong, E. Goldstein, L. Lin and J. Walker, "Design and nonlinear servo control of MEMS mirrors and their performance in a large port-count optical switch," *Journal of Microelectromechanical Systems,* vol. 14, no. 2, pp. 261-273, 2005.

[46] D. Burns and V. Bright, "Nonlinear flexures for stable deflection of an electrostatically actuated micromirror," in *Microelectronic Structures and MEMS*

for Optical Processing III, 1997.

[47] P. Brosens, "Dynamic mirror distortions in optical scanning," *Applied Optics,* vol. 11, no. 12, pp. 2987-2989, 1972.

[48] Y. Low, R. Scotti, D. Ramsey, C. Bolle, S. O'Neill and K. Nguyen, "Packaging of Optical MEMS Devices," *Journal of Electronic Packaging,* vol. 125, pp. 325-328, 2003.

[49] M. Kozhevnikov, R. Ryf, D. Neilson, P. Kolodner, D. Bolle, A. Papazian, J. Kim and J. Gates, "Micromechanical optical crossconnect with 4-F relay imaging optics," *Photonics Technology Letters,* vol. 16, no. 1, pp. 275-277, 2004.

[50] A. Ankiewicz and G. Peng, "Generalized Gaussian approximation for single-mode fibers," *Journal of Lightwave Technology,* vol. 10, no. 1, pp. 22-27, 1992.

[51] P. Colbourne, "Generally astigmatic Gaussian beam representation and optimization using skew rays," in *SPIE IODC*, 2015.

[52] J. Ford, V. Aksyuk, D. Bishop and J. Walker, "Wavelength add/drop switching using tilting micromirrors," *Journal of Lightwave Technology,* vol. 17, no. 5, pp. 904-911, 1999.

[53] C. Chen, S. Tzeng and S. Gwo, "Silicon microlens structures fabricated by scanning-probe gray-scale oxidation," *Optics Letters,* vol. 30, no. 6, pp. 652-654, 2005.

[54] P. Townsend and D. Barnett, "Elastic relationships in layered composite media with approximation for the case of thin films on a thick substrate," *Journal of Applied Physics,* vol. 62, no. 11, pp. 4438-4444, 1987.

[55] W. Dally and B. Towels, "Non-blocking networks," in *Principles and Practices of Interconnection Networks*, San Francisco, Elsevier, 2004.

[56] G. Marsden, P. Marchand, P. Harvey and S. Esener, "Optical transpose interconnection system architectures," *Optics Letters,* vol. 18, no. 13, pp. 1083-1085, 1993.

[57] W. M. Mellette and J. E. Ford, "Scaling limits of MEMS beam-steering switches for data center networks," *Journal of Lightwave Technology,* vol. 33, no. 15, pp. 3308-

3318, 2015.

[58] J. Ford, Y. Fainman and S. Lee, "Reconfigurable array interconnection by photorefractive correlation," *Applied Optics,* vol. 33, no. 23, pp. 5363-5377, 1994.

[59] C. Wu and T. Feng, "On a class of multistage interconnection networks," *Transactions on Computers,* vol. 29, no. 8, pp. 694-702, 1980.

[60] T. Beth and V. Hatz, "A restricted crossbar implementation and its applications," *ACM SIGARCH Computer Architecture News 19.6,* pp. 12-16, 1991.

[61] I. Stoica, R. Morris, D. Karger, M. Kaashoek and H. Balakrishnan, "Chord: a scalable peer-to-peer lookup service for internet applications," in *ACM SIGCOMM*, 2001.

[62] V. Kopp, J. Park, M. Wlodawski, J. Singer, D. Neugroschl and A. Genack, "Chiral fibers: microformed optical waveguides for polarization control, sensing, coupling, amplification, and switching," *Journal of Lightwave Technology,* vol. 32, no. 4, pp. 605-613, 2014.

[63] "Mirrorcle Technologies," [Online]. Available: http://www.mirrorcletech.com/.

[64] R. Wagner and W. Tomlinson, "Coupling efficiency of optics in single mode fiber components," *Applied Optics,* vol. 21, no. 15, pp. 2671-2688, 1982.

[65] "FiberGuide Specification Sheet," [Online]. Available: http://www.fiberguide.com/wp-content/uploads/2012/08/V-Grooves_Arrays_FINAL.pdf.

[66] J. Janhs and M. Murdocca, "Crossover networks and their optical implementation," *Applied Optics,* vol. 27, no. 15, pp. 3155-3160, 1998.

[67] T. Suleski and R. T. Kolste, "Fabrication trends for free-space microoptics," *Journal of Lightwave Technology,* vol. 23, no. 2, pp. 633-646, 2005.

[68] "Gurobi Optimization," [Online]. Available: www.gurobi.com.

[69] W. Dally and B. Towles, Principles and Practices of Interconnection Networks, Elsevier, 2004.

[70] Y. Dong, E. Olinick, T. Kratz and D. Matula, "A compact linear programming

formulation of the maximum concurrent flow problem," *Networks,* vol. 65, no. 1, pp. 68-87, 2015.

[71] N. Farrington and A. V. E. Rubow, "Data center switch architecture in the age of merchant silicon," in *17th IEEE Symposium on High Performance Interconnects*, 2009.

[72] "Intel Custom Foundry EMIB," [Online]. Available: http://www.intel.com/content/www/us/en/foundry/emib.html.

[73] J.-G. C. a. Y.-B. Fang, "Dot-pattern design of a light guide in an edge-lit backlight using a regional partition approach," *Optical Engineering,* vol. 46, no. 4, pp. 10984-10995, 2012.

[74] D. Feng, Y. Yan, X. Yang, G. Jin and S. Fan, "Novel integrated light-guide plates for liquid crystal display backlight," *Journal of Optics A Pure and Applied Optics,* vol. 7, no. 3, pp. 111-117, 2005.

[75] T. Teng and J. Ke, "A novel optical film to provide a highly collimated planar light source," *Optics Express,* vol. 21, no. 18, pp. 21444-21455, 2013.

[76] E. T. J. F. J. Karp, "Planar micro-optic solar concentrator," *Optics Express,* vol. 18, no. 2, pp. 1122-1133, 2010.

[77] J. Hallas, K. Baker, J. Karp, E. Tremblay and J. Ford, "Two-axis solar tracking accomplished through small lateral translations," *Applied Optics,* vol. 51, no. 25, pp. 6117-6124, 2012.

[78] A. Lohmann, "Scaling laws for lens systems," *Applied Optics,* vol. 28, no. 23, pp. 4996-4998, 1989.

[79] D. Moore, G. Schmidt and B. Unger, "Concentrated photovoltaic stepped planar light guide," in *International Optical Design Conference*, 2010.

[80] J. K. Kim, T. Gessmann, H. Luo and E. F. Schubert, "GaInN light emitting diodes with RuO2/SiO2/Ag omnidirectional reflector," *Applied Physics Letters,* vol. 84, no. 22, pp. 4508-4510, 2004.

[81] H. Luo, J. K. Kim, E. F. Schubert, J. Cho, C. Sone and Y. Park, "Analysis of high-power packages for phosphor-based white-light-emitting diodes," *Applied Physics*

Letters, vol. 86, no. 24, 2005.

[82] H. J. Cornelissen, H. Ma, C. Ho, M. Li and C. Mu, "Compact collimators for high brightness blue LEDs using dielectric multilayers," in *SPIE Optical Engineering and Applications*, 2011.

[83] T.-C. Teng, W.-S. Sun, L.-W. Tseng and W.-C. Chang, "A slim apparatus of transferring discrete LEDs' light into an ultra-collimated planar light source," *Optics Express,* vol. 21, no. 22, pp. 26972-26982, 2013.

[84] R. Winston, J. C. Minano and P. Benitez, Nonimaging Optics, Academic Press, 2005.

[85] F. Fournier, W. J. Cassarly and J. P. Rolland, "Method to improve spatial uniformity with lightpipes," *Optics Letters,* vol. 33, no. 11, pp. 1165-1167, 2008.

[86] M. Heiblum and J. H. Harris, "Analysis of curved optical waveguides by confromal transformation," *IEEE Journal of Quantum Electronics,* vol. 11, no. 2, pp. 75-83, 1975.

[87] S. Garner, H. Fong, M. He, P. Cimo, X. Li, Y. Cai, S. Ouyang, Y. Xie, Q. Shi and S. Cai, "Flexible glass substrates for display and lighting applications," in *IEEE Photonics Conference (IPC)*, 2013.

[88] K. A. Denault, M. Cantore, S. Nakamura, S. P. DenBaars and R. Seshadri, "Efficient and stable laser-driven white lighting," *AIP Advances,* vol. 3, no. 7, 2013.

[89] A. O. Marcano, C. Loper and N. Melikechi, "High-sensitivity absorption measurement in water and glass samples uisng a mode-mismatched pump-probe thermal lens method," *Applied Physics Letters,* vol. 78, no. 22, pp. 3415-3417, 2001.

[90] P. S. Chechurov and G. E. Romanova, "Using the ZEMAX software complex to form photometric models of LED illuminator devices," *Journal of Optical Technology,* vol. 79, no. 5, pp. 302-304, 2012.