HIGH PERFORMANCE SWITCHES AND ROUTERS

H. JONATHAN CHAO and BIN LIU

WILEY-INTERSCIENCE
A JOHN WILEY & SONS, INC., PUBLICATION

HIGH PERFORMANCE SWITCHES AND ROUTERS

The Wiley Bicentennial–Knowledge for Generations

Each generation has its unique needs and aspirations. When Charles Wiley first opened his small printing shop in lower Manhattan in 1807, it was a generation of boundless potential searching for an identity. And we were there, helping to define a new American literary tradition. Over half a century later, in the midst of the Second Industrial Revolution, it was a generation focused on building the future. Once again, we were there, supplying the critical scientific, technical, and engineering knowledge that helped frame the world. Throughout the 20th Century, and into the new millennium, nations began to reach out beyond their own borders and a new international community was born. Wiley was there, expanding its operations around the world to enable a global exchange of ideas, opinions, and know-how.

For 200 years, Wiley has been an integral part of each generation's journey, enabling the flow of information and understanding necessary to meet their needs and fulfill their aspirations. Today, bold new technologies are changing the way we live and learn. Wiley will be there, providing you the must-have knowledge you need to imagine new worlds, new possibilities, and new opportunities.

Generations come and go, but you can always count on Wiley to provide you the knowledge you need, when and where you need it!

William J. Pesce
President and Chief Executive Officer

Peter Booth Wiley
Chairman of the Board

HIGH PERFORMANCE SWITCHES AND ROUTERS

H. JONATHAN CHAO and BIN LIU

WILEY-INTERSCIENCE
A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey
Published simultaneously in Canada

Library of Congress Cataloging-in-Publication Data.

Chao, H. Jonathan, 1955-
High performance switches and routers / by H. Jonathan Chao, Bin Liu.
p. cm.
ISBN-13: 978-0-470-05367-6
ISBN-10: 0-470-05367-4
1. Asynchronous transfer mode. 2. Routers (Computer networks)
3. Computer network protocols. 4. Packet switching (Data transmission)
I. Liu, Bin. II. Title.
TK5105.35.C454 2007
621.382′16- -dc22 2006026971

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

CONTENTS

PREFACE

As increasing voice, audio, video, TV, and gaming traffic is carried over IP, Internet traffic continues to grow rapidly. Many network-related applications are emerging for portable devices. As smart cellular phone technology advances, the price decreases, and the infrastructure to support wireless applications (voice, data, video) is being deployed ubiquitously to meet unprecedented demands from users. All of these fast-growing services translate into the high volume of Internet traffic, stringent quality of service (QoS) requirements, large number of hosts/devices to be supported, large forwarding tables to support, high speed packet processing, and large storage capability. When designing/operating next generation switches and routers, these factors create new specifications and new challenges for equipment vendors and network providers.

Jonathan has co-authored two books: *Broadband Packet Switching Technologies—A Practical Guide to ATM Switches and IP Routers* and *Quality of Service Control in High-Speed Networks*, published by John Wiley in 2001. Because the technologies in both electronics and optics have significantly advanced and because the design specifications for routers have become more demanding and challenging, it is time to write another book. This book includes new architectures, algorithms, and implementations developed since 2001. Thus, it is more updated and more complete than the two previous books.

In addition to the need for high-speed and high-capacity transmission/switching equipment, the control function of the equipment and network has also become more sophisticated in order to support new features and requirements of the Internet, including fast re-routing due to link failure (one or more failures), network security, network measurement for dynamic routing, and easy management. This book focuses on the subsystems and devices on the data plane. There is a brief introduction to IP network management to familiarize readers with how the network is managed, as many routers are interconnected together.

The book starts with an introduction to today's and tomorrow's networks, the router architectures and their building blocks, examples of commercial high-end routers, and the challenging issues of designing high-performance high-speed routers. The book first covers the main functions in the line cards of a core router, including route lookup, packet classification, and traffic management for QoS control described in Chapters 2, 3, and

4, respectively. It then follows with 11 chapters in packet switching designs, covering various architectures, algorithms, and technologies (including electrical and optical packet switching). The last chapter of the book presents the state-of-the-art commercial chipsets used to build the routers. This is one of the important features in this book—showing readers the architecture and functions of practical chipsets to reinforce the theories and conceptual designs covered in previous chapters.

A distinction of this book is that we provide as many figures as possible to explain the concepts. Readers are encouraged to first scan through the figures and try to understand them before reading the text. If fully understood, readers can skip to the text to save time. However, the text is written in such a way as to talk the readers through the figures.

Jonathan and Bin each have about 20 years of experience researching high-performance switches and routers, implementing them in various systems with VLSI (very-large-scale integration) and FPGA (field-programmable gate array) chips, transferring technology to the industry, and teaching such subjects in the college and to the industry companies. They have accumulated their practical experience in writing this book. The book includes theoretical concepts and algorithms, design architectures, and actual implementations. It will benefit the readers in different aspects of building a high-performance switch/router. The draft of the book has been used as a text for the past two years when teaching senior undergraduate and first-year graduate students at the author's universities. If any errors are found, please send an email to chao@poly.edu. The authors will then make the corresponding corrections in future editions.

Audience

This book is an appropriate text for senior and graduate students in Electrical Engineering, Computer Engineering, and Computer Science. They can embrace the technology of the Internet so as to better position themselves when they graduate and look for jobs in the high-speed networking field. This book can also be used as a reference for people working in the Internet-related area. Engineers from network equipment vendors and service providers can also benefit from the book by understanding the key concepts of packet switching systems and the key techniques of building high-speed and high-performance routers.

ACKNOWLEDGMENTS

This book would not have been published without the help of many people. We would like to thank them for their efforts in improving the quality of the book.

Several chapters of the book are based on research work that was done at Polytechnic University and Tsinghua University. We would like to thank several individuals who contributed material to some sections. They are Professor Ming Yu (Florida State University) on Section 1.5, Professor Derek C. W. Pao (City University of Hong Kong) on Section 2.4.2, and Professor Aleksandra Smiljanic (Belgrade University) on a scheduling scheme she proposed in Chapter 7. We would like to express our gratitude to Dr. Yihan Li (Auburn University) for her contribution to part of Chapter 7, and the students in Bin's research group in Tsinghua University for their contribution to some chapters. They are Chenchen Hu, Kai Zheng, Zhen Liu, Lei Shi, Xuefei Chen, Xin Zhang, Yang Xu, Wenjie Li, and Wei Li. The manuscript has been managed from the beginning to the end by Mr Jian Li (Polytechnic University), who has put in tremendous effort to carefully edit the manuscript and serve as a coordinator with the publisher.

The manuscript draft was reviewed by the following people and we would like to thank them for their valuable feedback: Professor Cristina López Bravo (University of Vigo, Spain), Dr Hiroaki Harai (Institute of Information and Communications Technology, Japan), Dr Simin He (Chinese Academy of Sciences), Professor Hao Che (University of Texas at Arlington), Professor Xiaohong Jiang (Tohoku University, Japan), Dr Yihan Li (Auburn University), Professor Dr Soung Yue Liew (Universiti Tunku Abdul Rahman, Malaysia), Dr Jan van Lunteren (IBM, Zurich), Professor Jinsoo Park (Essex County College, New Jersey), Professor Roberto Rojas-cessa (New Jersey Institute of Technology), Professor Aleksandra Smiljanic (Belgrade University, Serbia and Montenegro), Professor Dapeng Wu (University of Florida), and Professor Naoaki Yamanaka (Keio University, Japan).

Jonathan would like to thank his wife, Ammie, and his children, Jessica, Roger, and Joshua, for their love, support, encouragement, patience, and perseverance. He also thanks his parents for their encouragement.

Bin would like to thank his wife, Yingjun Ma, and his daughter, Jenny for their understanding and support. He also thanks his father-in-law for looking after Jenny to spare his time to prepare the book.

CHAPTER 1

INTRODUCTION

The Internet, with its robust and reliable Internet Protocol (IP), is widely considered the most reachable platform for the current and next generation information infrastructure. The virtually unlimited bandwidth of optical fiber has tremendously increased the data transmission speed over the past decade. Availability of unlimited bandwidth has stimulated high-demand multimedia services such as distance learning, music and video download, and videoconferencing. Current broadband access technologies, such as digital subscriber lines (DSLs) and cable television (CATV), are providing affordable broadband connection solutions to the Internet from home. Furthermore, with Gigabit Ethernet access over dark fiber to the enterprise on its way, access speeds are expected to largely increase. It is clear that the deployment of these broadband access technologies will result in a high demand for large Internet bandwidth. To keep pace with the Internet traffic growth, researchers are continually exploring faster transmission and switching technologies. The advent of optical transmission technologies, such as dense wave division multiplexing (DWDM), optical add-drop multiplexers, and ultra-long-haul lasers have had a large influence on lowering the costs of digital transmission. For instance, 300 channels of 11.6 Gbps can be wavelength-division multiplexed on a single fiber and transmitted over 7000 km [1]. In addition, a 1296×1296 optical cross-connect (OXC) switching system using micro-electro-mechanical systems (MEMS) with a total switching capacity of 2.07 petabits/s has been demonstrated [2]. In the rest of this chapter, we explore state-of-the-art network infrastructure, future design trends, and their impact on next generation routers. We also describe router architectures and the challenges involved in designing high-performance large-scale routers.

High Performance Switches and Routers, by H. Jonathan Chao and Bin Liu

1.1 ARCHITECTURE OF THE INTERNET: PRESENT AND FUTURE

1.1.1 The Present

Today's Internet is an amalgamation of thousands of commercial and service provider networks. It is not feasible for a single service provider to connect two distant nodes on the Internet. Therefore, service providers often rely on each other to connect the dots. Depending on the size of network they operate, Internet Service Providers (ISPs) can be broken down into three major categories. Tier-1 ISPs are about a dozen major telecommunication companies, such as UUNet, Sprint, Qwest, XO Network, and AT&T, whose high-speed global networks form the Internet backbone. Tier-1 ISPs do not buy network capacity from other providers; instead, they sell or lease access to their backbone resource to smaller Tier-2 ISPs, such as America Online and Broadwing. Tier-3 ISPs are typically regional service providers such as Verizon and RCN through whom most enterprises connect to the Internet. Figure 1.1 illustrates the architecture of a typical Tier-1 ISP network.

Each Tier-1 ISP operates multiple IP/MPLS (multi-protocol label switching), and sometimes ATM (asynchronous transfer mode), backbones with speeds varying anywhere from T3 to OC-192 (optical carrier level 192, ∼10 Gbps). These backbones are interconnected through peering agreements between ISPs to form the Internet backbone. The backbone is designed to transfer large volumes of traffic as quickly as possible between networks. Enterprise networks are often linked to the rest of the Internet via a variety of links, anywhere from a T1 to multiple OC-3 lines, using a variety of Layer 2 protocols, such as Gigabit Ethernet, frame relay, and so on. These enterprise networks are then overhauled into service provider networks through edge routers. An edge router can aggregate links from multiple enterprises. Edge routers are interconnected in a pool, usually at a Point of Presence (POP)

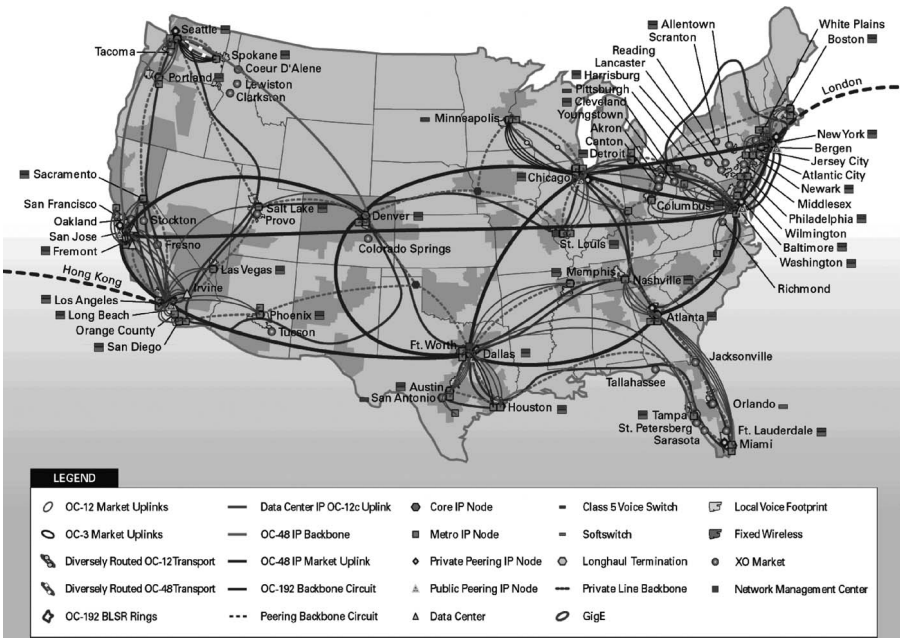


Figure 1.1 Network map of a Tier-1 ISP, XO Network.

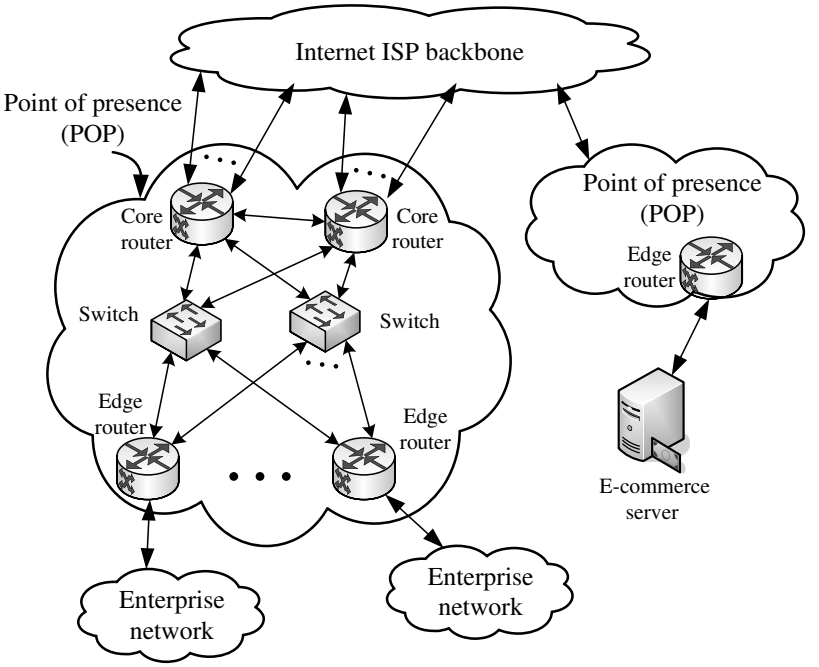
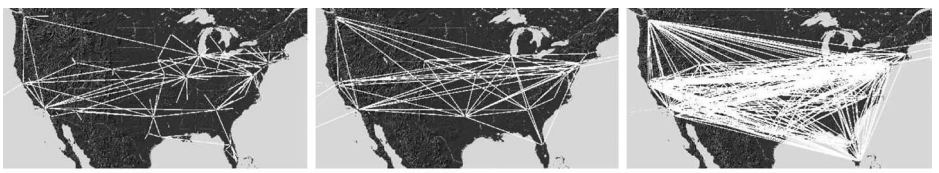


Figure 1.2 Point of presence (POP).

of a service provider, as shown in Figure 1.2. Each POP may link to other POPs of the same ISP through optical transmission/switching equipment, may link to POPs of other ISPs to form a peering, or link to one or more backbone routers. Typically, a POP may have a few backbone routers in a densely connected mesh. In most POPs, each edge router connects to at least two backbone routers for redundancy. These backbone routers may also connect to backbone routers at other POPs according to ISP peering agreements. Peering occurs when ISPs exchange traffic bound for each other's network over a direct link without any fees. Therefore, peering works best when peers exchange roughly the same amount of traffic. Since smaller ISPs do not have high quantities of traffic, they often have to buy *transit* from a Tier-1 provider to connect to the Internet. A recent study of the topologies of 10 service providers across the world shows that POPs share this generic structure [3].

Unlike POPs, the design of backbone varies from service provider to service provider. For example, Figure 1.3 illustrates backbone design paradigms of three major service providers



(*a*) (*b*) (*c*)

Figure 1.3 Three distinct backbone design paradigms of Tier-1 ISPs. (*a*) AT&T; (*b*) Sprint; (*c*) Level 3 national network infrastructure [3].

in the US. AT&T's backbone design includes large POPs at major cities, which in turn fan out into smaller per-city POPs. In contrast, Sprint's backbone has only 20 well connected POPs in major cities and suburban links are back-hauled into the POPs via smaller ISPs. Most major service providers still have the AT&T backbone model and are in various stages of moving to Sprint's design. Sprint's backbone design provides a good solution to service providers grappling with a need to reduce capital expenditure and operational costs associated with maintaining and upgrading network infrastructure. Interestingly, Level 3 presents another design paradigm in which the backbone is highly connected via circuit technology such as, MPLS, ATM or frame relays. As will be seen later, this is the next generation of network design where the line between backbone and network edge begins to blur.

Now, let us see how network design impacts on the next generation routers. Router design is often guided by the economic requirements of service providers. Service providers would like to reduce the infrastructure and maintenance costs while, at the same time, increasing available bandwidth and reliability. To this end, network backbone has a set of well-defined, narrow requirements. Routers in the backbone should simply move traffic as fast as possible. Network edge, however, has broad and evolving requirements due simply to the diversity of services and Layer 2 protocols supported at the edge. Today most POPs have multiple edge routers optimized for point solutions. In addition to increasing infrastructure and maintenance costs, this design also increases the complexity of POPs resulting in an unreliable network infrastructure. Therefore, newer edge routers have been designed to support diversity and are easily adaptable to the evolving requirements of service providers. This design trend is shown in Table 1.1, which lists some properties of enterprise, edge, and core routers currently on the market. As we will see in the following sections, future network designs call for the removal of edge routers altogether and their replacement with fewer core routers to increase reliability, throughput, and to reduce costs. This means next generation routers would have to amalgamate the diverse service requirements of edge routers and the strict performance requirements of core routers, seamlessly into one body. Therefore, the real question is not whether we should build highly-flexible, scalable, high-performance routers, but how?

1.1.2 The Future

As prices of optical transport and optical switching sharply decrease, some network designers believe that the future network will consist of many mid-size IP routers or MPLS

TABLE 1.1 Popular Enterprise, Edge, and Core Routers in the Market

Model	Capacity[a]	Memory	Power	Features
Cisco 7200	–	256 MB	370 W	QoS, MPLS, Aggregation
Cisco 7600	720 Gbps	1 GB	–	QoS, MPLS, Shaping
Cisco 10000	51.2 Gbps	–	1200 W	QoS, MPLS
Cisco 12000	1.28 Tbps	4 GB	4706 W	MPLS, Peering
Juniper M-320	320 Gbps	2 GB	3150 W	MPLS, QoS, VPN
Cisco CRS	92 Tbps	4 GB	16,560 W	MPLS, Qos, Peering
Juniper TX/T-640	2.5 Tbps/640 Gbps	2 GB	4550 W/6500 W	MPLS, QoS, Peering

[a]Note that the listed capacity is the combination of ingress and egress capacities.

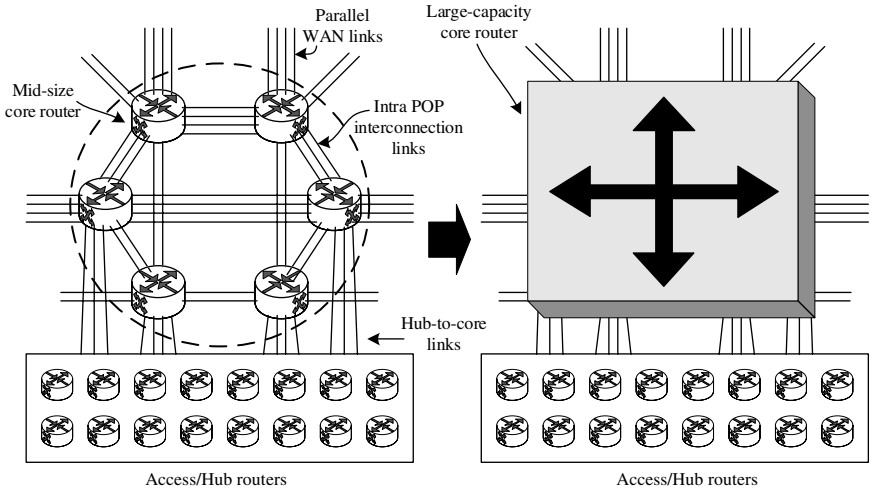


Figure 1.4 Replacing a cluster of mid-size routers with a large-capacity scalable router.

switches at the network edge that are connected to optical crossconnects (OXCs), which are then interconnected by DWDM transmission equipment. The problem for this approach is that connections to the OXC are usually high bit rates, for example, 10 Gbps for now and 40 Gbps in the near future. When the edge routers want to communicate with all other routers, they either need to have direct connections to those routers or connect through multiple logical hops (i.e., routed by other routers). The former case results in low link utilization while the latter results in higher latency. Therefore, some network designers believe it is better to build very large IP routers or MPLS switches at POPs. They aggregate traffic from edge routers onto high-speed links that are then directly connected to other large routers at different POPs through DWDM transmission equipment. This approach achieves higher link utilization and fewer hops (thus lower latency). As a result, the need for an OXC is mainly for provisioning and restoring purposes but not for dynamic switching to achieve higher link utilization.

Current router technologies available in the market cannot provide large switching capacities to satisfy current and future bandwidth demands. As a result, a number of mid-size core routers are interconnected with numerous links and use many expensive line cards that are used to carry intra-cluster traffic rather than revenue-generating users' or wide-area-network (WAN) traffic. Figure 1.4 shows how a router cluster is replaced by a large-capacity scalable router, saving the cost of numerous line cards and links, and real estate. It provides a cost-effective solution that can satisfy Internet traffic growth without having to replace routers every two to three years. Furthermore, there are fewer individual routers that need to be configured and managed, resulting in a more efficient and reliable system.

1.2 ROUTER ARCHITECTURES

IP routers' functions can be classified into two categories: datapath functions and control plane functions [4].

The datapath functions such as forwarding decision, forwarding through the backplane, and output link scheduling are performed on every datagram that passes through the router. When a packet arrives at the forwarding engine, its destination IP address is first masked by the subnet mask (logical AND operation) and the resulting address is used to lookup the forwarding table. A so-called longest prefix matching method is used to find the output port. In some applications, packets are classified based on 104 bits that include the IP source/destination addresses, transport layer port numbers (source and destination), and type of protocol, which is generally called 5-tuple. Based on the result of classification, packets may be either discarded (firewall application) or handled at different priority levels. Then, time-to-live (TTL) value is decremented and a new header checksum is recalculated.

The control plane functions include the system configuration, management, and exchange of routing table information. These are performed relatively infrequently. The route controller exchanges the topology information with other routers and constructs a routing table based on a routing protocol, for example, RIP (Routing Information Protocol), OSPF (Open Shortest Path Forwarding), or BGP (Border Gateway Protocol). It can also create a forwarding table for the forwarding engine. Since the control functions are not performed on each arriving individual packet, they do not have a strict speed constraint and are implemented in software in general.

Router architectures generally fall into two categories: centralized (Fig. 1.5*a*) and distributed (Fig. 1.5*b*).

Figure 1.5*a* shows a number of network interfaces, forwarding engines, a route controller (RC), and a management controller (MC) interconnected by a switch fabric. Input interfaces send packet headers to the forwarding engines through the switch fabric. The forwarding engines, in turn, determine which output interface the packet should be sent to. This information is sent back to the corresponding input interface, which forwards the packet to the right output interface. The only task of a forwarding engine is to process packet headers and is shared by all the interfaces. All other tasks such as participating in routing protocols, reserving resource, handling packets that need extra attention, and other administrative and maintenance tasks, are handled by the RC and the MC. The BBN multi-gigabit router [5] is an example of this design.

The difference between Figure 1.5*a* and 1.5*b* is that the functions of the forwarding engines are integrated into the interface cards themselves. Most high-performance routers use this architecture. The RC maintains a routing table and updates it based on routing protocols used. The routing table is used to generate a forwarding table that is then downloaded

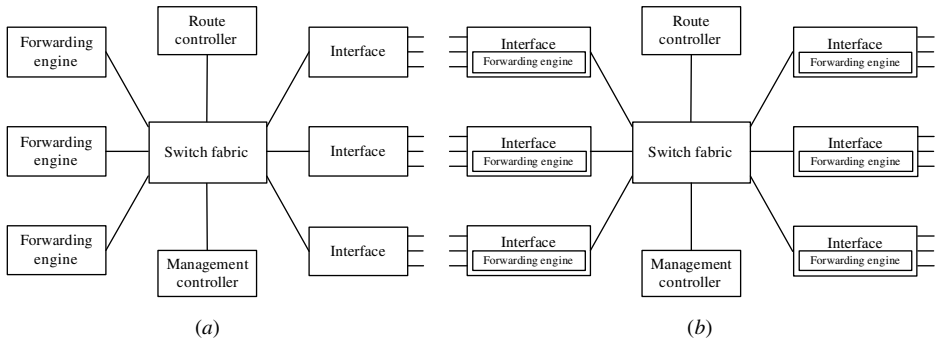


Figure 1.5 (*a*) Centralized versus (*b*) distributed models for a router.

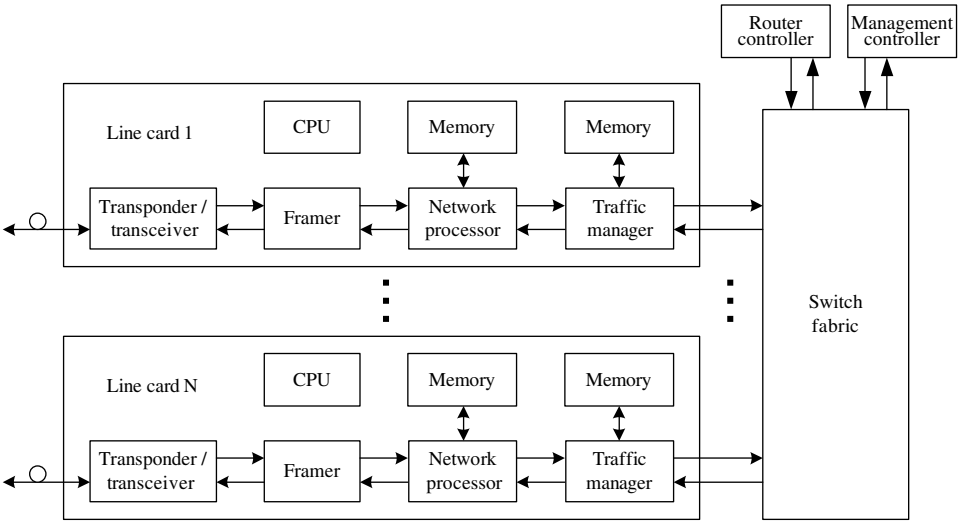


Figure 1.6 Typical router architecture.

from the RC to the forwarding engines in the interface cards. It is not necessary to download a new forwarding table for every route update. Route updates can be frequent, but routing protocols need time, in the order of minutes, to converge. The RC needs a dynamic routing table designed for fast updates and fast generation of forwarding tables. Forwarding tables, on the other hand, can be optimized for lookup speed and need not be dynamic.

Figure 1.6 shows a typical router architecture, where multiple line cards, an RC, and an MC are interconnected through a switch fabric. The communication between the RC/MC and the line cards can be either through the switch fabric or through a separate interconnection network, such as a Ethernet switch. The line cards are the entry and exit points of data to and from a router. They provide the interface from physical and higher layers to the switch fabric. The tasks provided by line cards are becoming more complex as new applications develop and protocols evolve. Each line card supports at least one full-duplex fiber connection on the network side, and at least one ingress and one egress connection to the switch fabric backplane. Generally speaking, for high-bandwidth applications, such as OC-48 and above, the network connections support channelization for aggregation of lower-speed lines into a large pipe, and the switch fabric connections provide flow-control mechanisms for several thousand input and output queues to regulate the ingress and egress traffic to and from the switch fabric.

A line card usually includes components such as a transponder, framer, network processor (NP), traffic manager (TM), and central processing unit (CPU).

Transponder/Transceiver. This component performs optical-to-electrical and electrical-to-optical signal conversions, and serial-to-parallel and parallel-to-serial conversions [6, 7].

Framer. A framer performs synchronization, frame overhead processing, and cell or packet delineation. On the transmit side, a SONET (synchronous optical network)/SDH (synchronous digital hierarchy) framer generates section, line, and path overhead. It performs framing pattern insertion (A1, A2) and scrambling. It

generates section, line, and path bit interleaved parity (B1/B2/B3) for far-end performance monitoring. On the receive side, it processes section, line, and path overhead. It performs frame delineation, descrambling, alarm detection, pointer interpretation, bit interleaved parity monitoring (B1/B2/B3), and error count accumulation for performance monitoring [8]. An alternative for the framer is Ethernet framer.

Network Processor. The NP mainly performs table lookup, packet classification, and packet modification. Various algorithms to implement the first two functions are presented in Chapters 2 and 3, respectively. The NP can perform those two functions at the line rate using external memory, such as static random access memory (SRAM) or dynamic random access memory (DRAM), but it may also require external content addressable memory (CAM) or specialized co-processors to perform deep packet classification at higher levels. In Chapter 16, we present some commercially available NP and ternary content addressable memory (TCAM) chips.

Traffic Manager. To meet the requirements of each connection and service class, the TM performs various control functions to cell/packet streams, including traffic access control, buffer management, and cell/packet scheduling. Traffic access control consists of a collection of specification techniques and mechanisms that (1) specify the expected traffic characteristics and service requirements (e.g., peak rate, required delay bound, loss tolerance) of a data stream; (2) shape (i.e., delay) data streams (e.g., reducing their rates and/or burstiness); and (3) police data streams and take corrective actions (e.g., discard, delay, or mark packets) when traffic deviates from its specification. The usage parameter control (UPC) in ATM and differentiated service (DiffServ) in IP performs similar access control functions at the network edge. Buffer management performs cell/packet discarding, according to loss requirements and priority levels, when the buffer exceeds a certain threshold. Proposed schemes include early packet discard (EPD) [9], random early packet discard (REPD) [10], weighted REPD [11], and partial packet discard (PPD) [12]. Packet scheduling ensures that packets are transmitted to meet each connection's allocated bandwidth/delay requirements. Proposed schemes include deficit round-robin, weighted fair queuing (WFQ) and its variants, such as shaped virtual clock [13] and worst-case fairness WFQ (WF^2Q+) [14]. The last two algorithms achieve the worst-case fairness properties. Details are discussed in Chapter 4. Many quality of service (QoS) control techniques, algorithms, and implementation architectures can be found in Ref. [15]. The TM may also manage many queues to resolve contention among the inputs of a switch fabric, for example, hundreds or thousands of virtual output queues (VOQs). Some of the representative TM chips on the market are introduced in Chapter 16, whose purpose it is to match the theories in Chapter 4 with practice.

Central Processing Unit. The CPU performs control plane functions including connection set-up/tear-down, table updates, register/buffer management, and exception handling. The CPU is usually not in-line with the fast-path on which maximum-bandwidth network traffic moves between the interfaces and the switch fabric.

The architecture in Figure 1.6 can be realized in a multi-rack (also known as multi-chassis or multi-shelf) system as shown in Figure 1.7. In this example, a half rack, equipped with a switch fabric, a duplicated RC, a duplicated MC, a duplicated system clock (CLK), and a duplicated fabric shelf controller (FSC), is connected to all other line card (LC) shelves, each of which has a duplicated line card shelf controller (LSC). Both the FSC and the

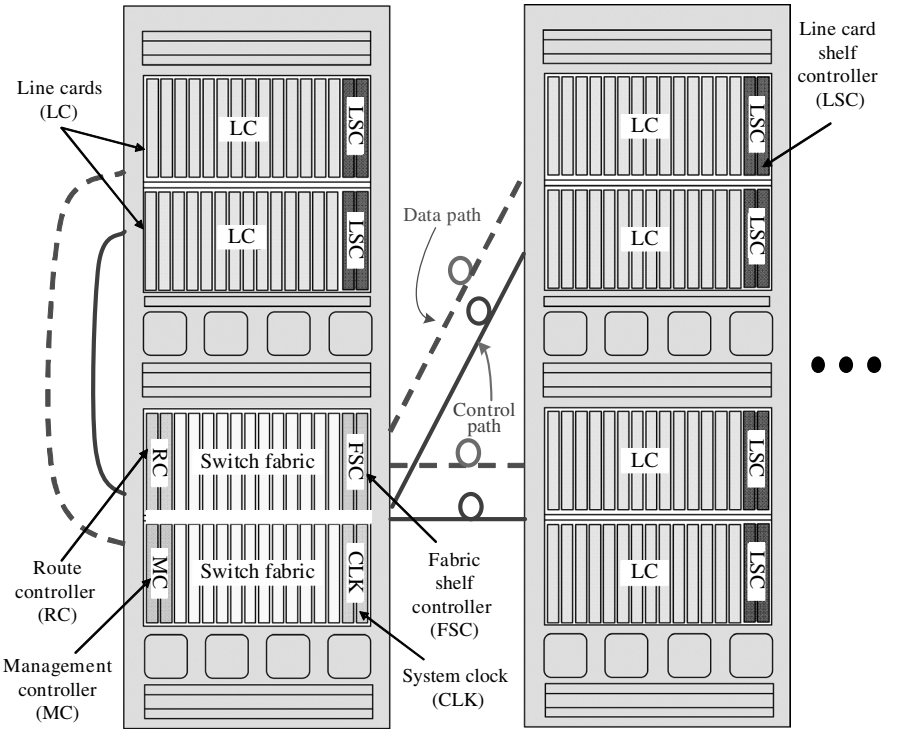


Figure 1.7 Multi-rack router system.

LSC provide local operation and maintenance for the switch fabric and line card shelves, respectively. They also provide the communication channels between the switch/line cards with the RC and the MC. The duplicated cards are for reliability concerns. The figure also shows how the system can grow by adding more LC shelves. Interconnections between the racks are sets of cables or fibers, carrying information for the data and the control planes. The cabling usually is a combination of unshielded twisted path (UTP) Category 5 Ethernet cables for control path, and fiber-optic arrays for data path.

1.3 COMMERCIAL CORE ROUTER EXAMPLES

We now briefly discuss the two most popular core routers on the market: Juniper Network's T640 TX-Matrix [16] and Cisco System's Carrier Routing System (CRS-1) [17].

1.3.1 T640 TX-Matrix

A T640 TX-Matrix is composed of up to four routing nodes and a TX Routing Matrix interconnecting the nodes. A TX Routing Matrix connects up to four T640 routing nodes via a three-stage Clos network switch fabric to form a unified router with the capacity of 2.56 Terabits. The blueprint of a TX Routing Matrix is shown in Figure 1.8. The unified router is controlled by the Routing Engine of the matrix which is responsible for running routing protocols and for maintaining overall system state. Routing engines in each routing

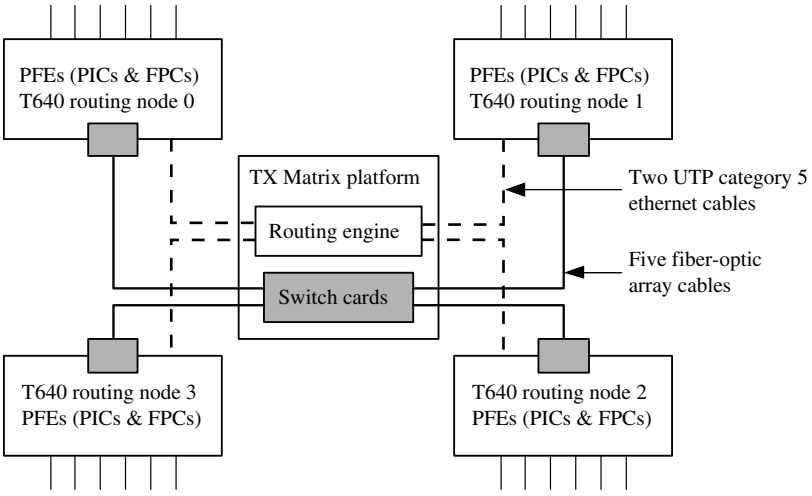


Figure 1.8 TX Routing Matrix with four T640 routing nodes.

node manage their individual components in coordination with the routing engine of the matrix. Data and control plane of each routing node is interconnected via an array of optical and Ethernet cables. Data planes are interconnected using VCSEL (vertical cavity surface emitting laser) optical lines whereas control planes are interconnected using UTP Category 5 Ethernet cables.

As shown in Figure 1.9, each routing node has two fundamental architectural components, namely the control plane and the data plane. The T640 routing node's control plane is implemented by the JUNOS software that runs on the node's routing engine. JUNOS is a micro-kernel-based modular software that assures reliability, fault isolation, and high availability. It implements the routing protocols, generates routing tables and forwarding tables, and supports the user interface to the router. Data plane, on the other hand, is responsible for processing packets in hardware before forwarding them across the switch fabric from the ingress interface to the appropriate egress interface. The T640 routing node's data plane is implemented in custom ASICs in a distributed architecture.

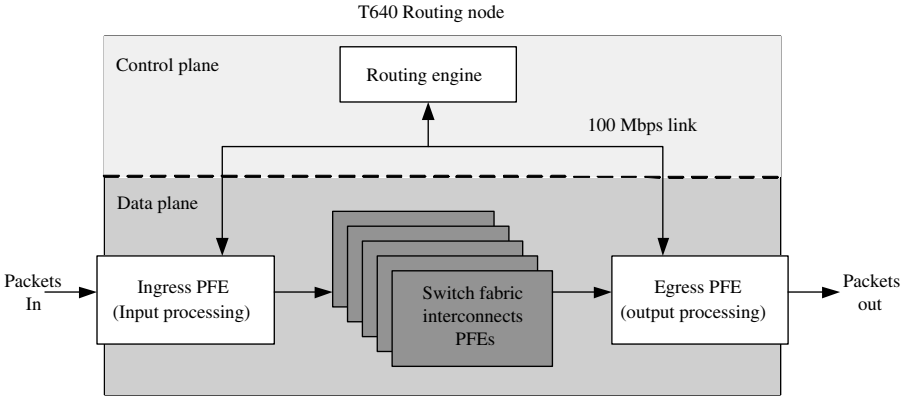


Figure 1.9 T640 routing node architecture.