The Network Core

Queuing Delays and Packet Loss

- For each attached link, the packet switch has an output buffer (also called an output queue), which stores packets that the router is about to send into that link.
- In addition to the store-and-forward delays, packets suffer output buffer queuing delays.
- These delays are variable and depend on the level of congestion in the network.
- Since the amount of buffer space is finite, an arriving packet may find that the buffer is completely full with other packets waiting for transmission. In this case, packet loss will occur—either the arriving packet or one of the already-queued packets will be dropped

The Network Core

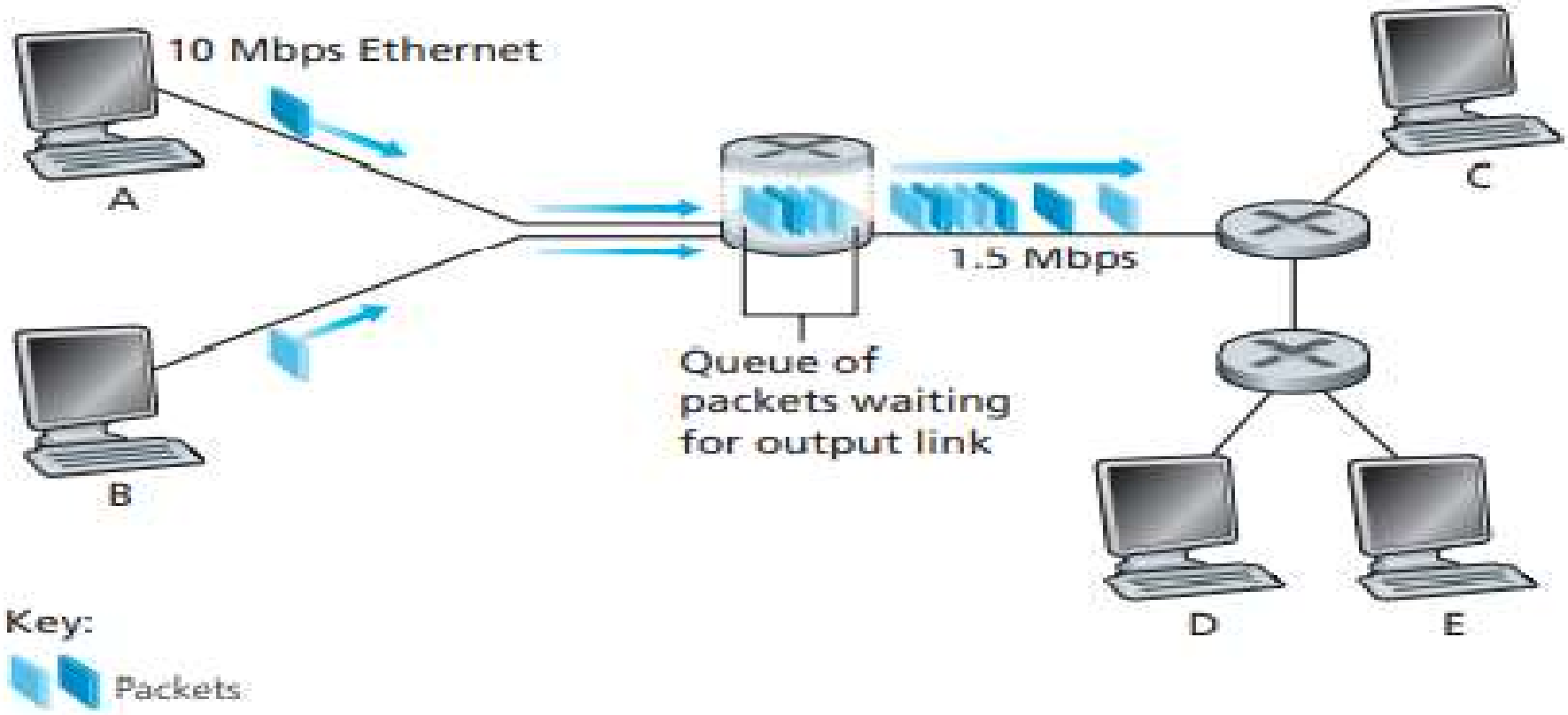

Figure 1.24 Packet switching

- Figure 1.24 illustrates a simple packet-switched network.

The Network Core

Circuit Switching and Packet Switching

- There are two fundamental approaches to moving data through a network of links and switches: circuit switching and packet switching.
- In circuit-switched networks, the resources needed along a path (buffers, link transmission rate) to provide for communication between the end systems are reserved for the duration of the communication session between the end systems.
- In packet-switched networks, these resources are not reserved; a session's messages use the resources on demand, and as a consequence, may have to wait (that is, queue) for access to a communication link.

The Network Core

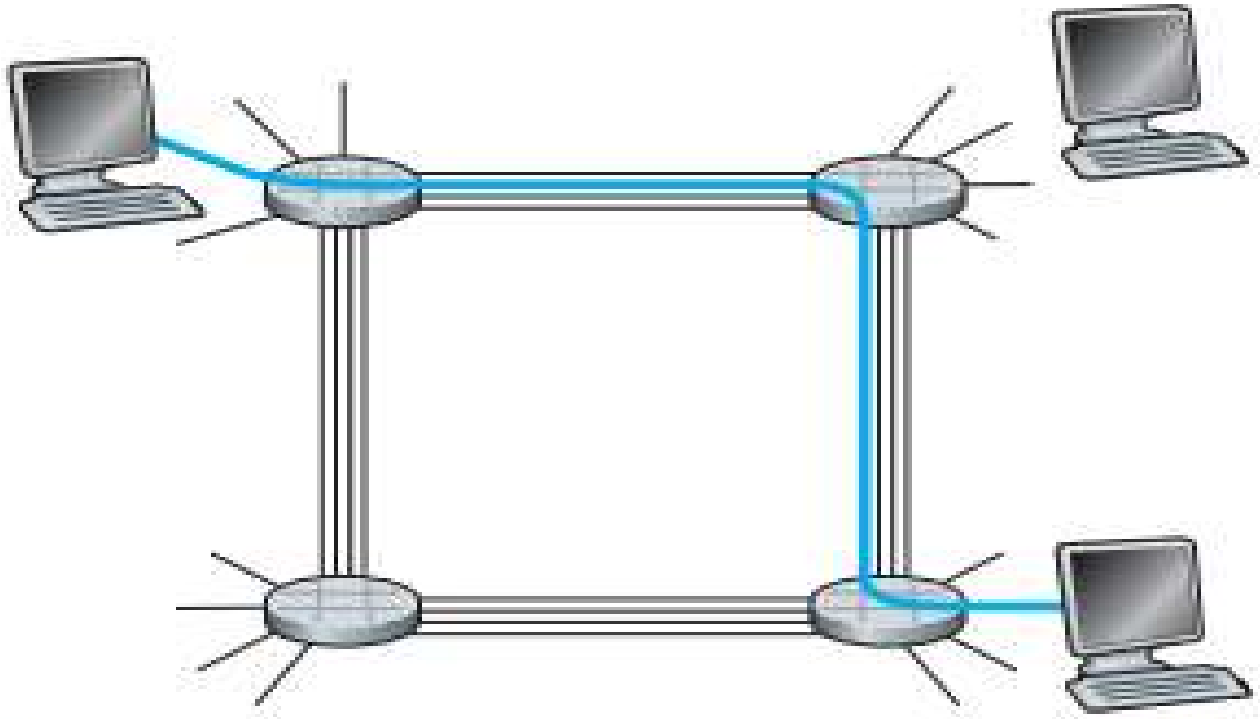

Figure 1.25 A simple circuit-switched network consisting of four switches and four links

The Network Core

Multiplexing in Circuit-Switched Networks

- A circuit in a link is implemented with either frequency-division multiplexing (FDM) or time-division multiplexing (TDM).
- With FDM, the frequency spectrum of a link is divided up among the connections established across the link.

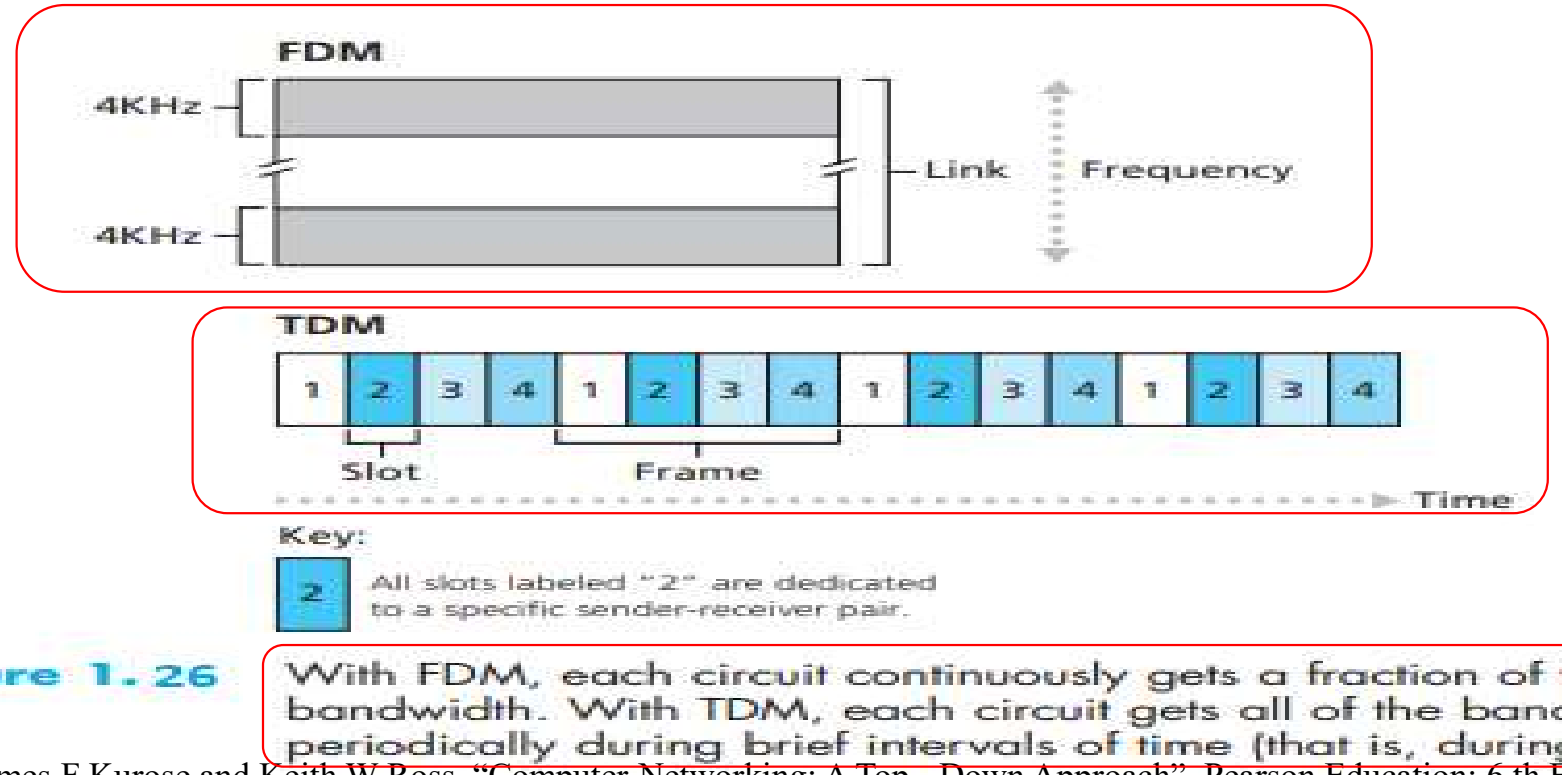

Figure 1.26 With FDM, each circuit continuously gets a fraction of the bandwidth. With TDM, each circuit gets all of the bandwidth periodically during brief intervals of time (that is, during slots)

Delay, Loss, and Throughput in Packet-Switched Networks

Overview of Delay in Packet-Switched Networks

- Packet suffers from several types of delays at each node along the path.
- The most important of these delays are the nodal processing delay, queuing delay, transmission delay, and propagation delay; together, these delays accumulate to give a total nodal delay.

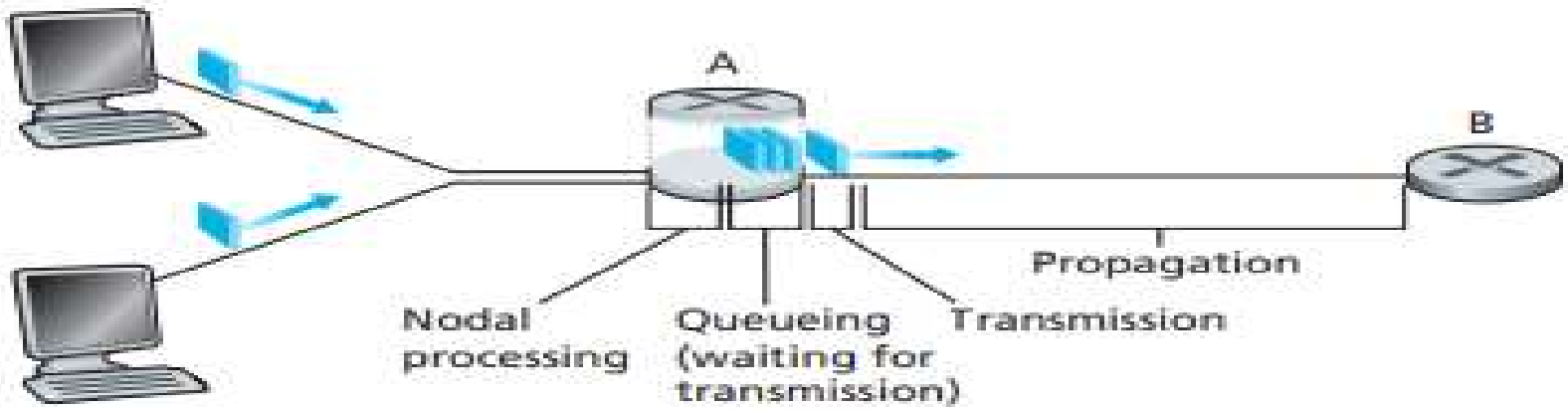

Figure 1.27 The nodal delay at router A

Types of Delay

Processing Delay

- The time required to examine the packet's header and determine where to direct the packet is part of the processing delay.
- The processing delay can also include other factors, such as the time needed to check for bit-level errors in the packet that occurred in transmitting the packet's bits from the upstream node to router A.
- Processing delays in high-speed routers are typically on the order of microseconds or less.
- After this nodal processing, the router directs the packet to the queue that precedes the link to router B.

Types of Delay

Queuing Delay

- At the queue, the packet experiences a queuing delay as it waits to be transmitted onto the link.
- The length of the queuing delay of a specific packet will depend on the number of earlier-arriving packets that are queued and waiting for transmission onto the link.
- If the queue is empty and no other packet is currently being transmitted, then our packet's queuing delay will be zero.
- On the other hand, if the traffic is heavy and many other packets are also waiting to be transmitted, the queuing delay will be long.
- Queuing delays can be on the order of microseconds to milliseconds in practice.

Types of Delay

Transmission Delay

- Assuming that packets are transmitted in a first-come-first-served manner, the packet can be transmitted only after all the packets that have arrived before it have been transmitted.
- Denote the length of the packet by L bits, and denote the transmission rate of the link from router A to router B by R bits/sec.
- For example, for a 10 Mbps Ethernet link, the rate is R = 10 Mbps; for a 100 Mbps Ethernet link, the rate is R = 100 Mbps.
- The transmission delay is L/R. This is the amount of time required to push (that is, transmit) all of the packet's bits into the link.
- Transmission delays are typically on the order of microseconds to milliseconds in practice.

Types of Delay

Propagation Delay

- Once a bit is pushed into the link, it needs to propagate to router B. The time required to propagate from the beginning of the link to router B is the propagation delay.
- The bit propagates at the propagation speed of the link.
- The propagation speed depends on the physical medium of the link (that is, fiber optics, twisted-pair copper wire, coaxial cable and so on) and is in the range of $2x10^8$ meters/sec to $3x10^8$ meters/sec which is equal to, or a little less than, the speed of light.

Types of Delay

Propagation Delay

- The propagation delay is the distance between two routers divided by the propagation speed.
- That is, the propagation delay is d/s, where d is the distance between router A and router B and s is the propagation speed of the link.
- Once the last bit of the packet propagates to node B, it and all the preceding bits of the packet are stored in router B.
- The whole process then continues with router B now performing the forwarding.
- In wide-area networks, propagation delays are on the order of milliseconds.

Types of Delay

Comparing Transmission and Propagation Delay

- The transmission delay is the amount of time required for the router to push out the packet; it is a function of the packet's length and the transmission rate of the link, but has nothing to do with the distance between the two routers.

- The propagation delay is the time it takes a bit to propagate from one router to the next; it is a function of the distance between the two routers, but has nothing to do with the packet's length or the transmission rate of the link.

Types of Delay

- Calculate the propagation time and transmission time for a 5Mbytes message if the bandwidth of the network is 1Mbps. Assume that distance between the sender and receiver is 12000km and light travels at 2.4×10^8 m/s.

- Propagation time=d/s=$(12000 \times 1000)/ 2.4 \times 10^8 = 50$ms
- Transmission time=L/R=$(5 \times 10^6 \times 8)/ 10^6 = 40$s

Types of Delay

- If d_{proc}, d_{queue}, d_{trans}, and d_{prop} denote the processing, queuing, transmission, and propagation delays, then the total nodal delay is given by $d_{nodal} = d_{proc} + d_{queue} + d_{trans} + d_{prop}$
- The contribution of these delay components can vary significantly.
- For example, d_{prop} can be negligible (for example, a couple of microseconds) for a link connecting two routers on the same university campus; however, d_{prop} is hundreds of milliseconds for two routers interconnected by a geostationary satellite link, and can be the dominant term in d_{nodal}.

Types of Delay

- Similarly, d_{trans} can range from negligible to significant. Its contribution is typically negligible for transmission rates of 10 Mbps and higher (for example, for LANs); however, it can be hundreds of milliseconds for large Internet packets sent over low-speed dial-up modem links.
- The processing delay, d_{proc}, is often negligible; however, it strongly influences a router's maximum throughput, which is the maximum rate at which a router can forward packets.

Queuing Delay

- The most complicated and interesting component of nodal delay is the queuing delay, d_{queue}.
- Unlike the other three delays (namely, d_{proc}, d_{trans}, and d_{prop}), the queuing delay can vary from packet to packet.
- For example, if 10 packets arrive at an empty queue at the same time, the first packet transmitted will suffer no queuing delay, while the last packet transmitted will suffer a relatively large queuing delay (while it waits for the other nine packets to be transmitted).
- Therefore, when characterizing queuing delay, one typically uses statistical measures, such as average queuing delay, variance of queuing delay, and the probability that the queuing delay exceeds some specified value.

Queuing Delay

When is the queuing delay large and when is it insignificant?

Queuing Delay

When is the queuing delay large and when is it insignificant?

- The answer to this question depends on the rate at which traffic arrives at the queue, the transmission rate of the link, and the nature of the arriving traffic, that is, whether the traffic arrives periodically or arrives in bursts.
- Let a denote the average rate at which packets arrive at the queue (a is in units of packets/sec).
- R is the transmission rate; that is, it is the rate (in bits/sec) at which bits are pushed out of the queue.
- Suppose all packets consist of L bits. Then the average rate at which bits arrive at the queue is La bits/sec.

Queuing Delay

When is the queuing delay large and when is it insignificant?

- Assume that the queue is very big, so that queuing delay will approach infinity and can hold essentially an infinite number of bits.
- The ratio La/R, called the traffic intensity, often plays an important role in estimating the extent of the queuing delay.
- If La/R > 1, then the average rate at which bits arrive at the queue exceeds the rate at which the bits can be transmitted from the queue. In this situation, the queue will tend to increase without bound and the queuing delay will approach infinity!
- Therefore, one of the golden rules in traffic engineering is: Design your system so that the traffic intensity is no greater than 1.($La/R \leq 1$)

Queuing Delay

When is the queuing delay large and when is it insignificant?

- Now consider the case $La/R \leq 1$.
- The nature of the arriving traffic impacts the queuing delay.
- For example, if packets arrive periodically—that is, one packet arrives every L/R seconds—then every packet will arrive at an empty queue and there will be no queuing delay.
- If packets arrive in bursts but periodically, there can be a significant average queuing delay.
- For example, suppose N packets arrive simultaneously every (L/R)N seconds. Then the first packet transmitted has no queuing delay; the second packet transmitted has a queuing delay of L/R seconds; and more generally, the nth packet transmitted has a queuing delay of (n-1)L/R seconds.

Queuing Delay

- As the traffic intensity approaches 1, the average queuing delay increases rapidly.
- A small percentage increase in the intensity will result in a much larger percentage-wise increase in delay.
- If some event causes an even slightly larger-than-usual amount of traffic, the delays you experience can be huge.

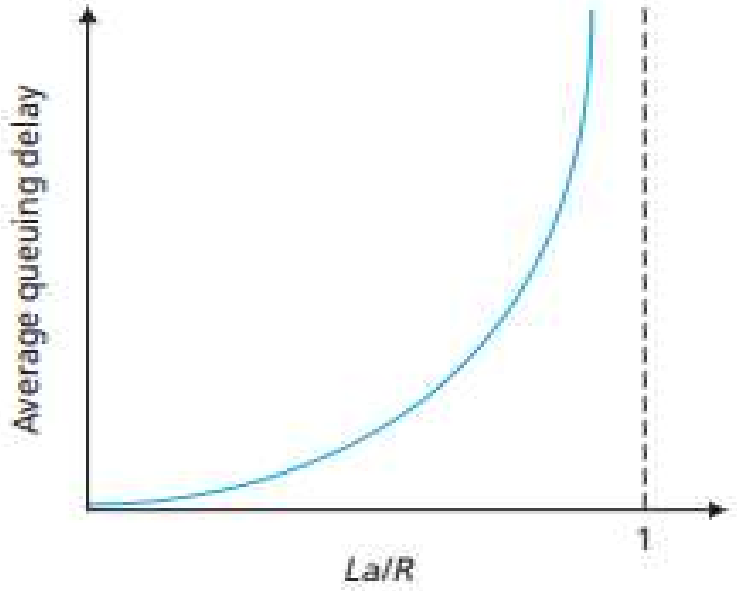

Figure 1.28 Dependence of average queuing delay on traffic intensity