To see the importance of looking at performance from top to bottom, including both hardware and software, consider the following example.

Performance of Two Networks

Example

Consider the following measurements made on a pair of SPARCstation 10s running Solaris 2.3, connected to two different types of networks, and using TCP/IP for communication:

Characteristic	Ethernet	ATM
Bandwidth from node to network	1.125 MB/sec	10 MB/sec
Interconnect latency	15 µs	50 µs
HW latency to/from network	6 µs	6 µs
SW overhead sending to network	200 µs	207 µs
SW overhead receiving from network	241 µs	360 µs

Find the host-to-host latency for a 250-byte message using each network.

Answer

We can estimate the time required as the sum of the fixed latencies plus the time to transmit the message. The time to transmit the message is simply the message length divided by the bandwidth of the network.

The transmission times are

$$\text{Transmission time}_{\text{Ethernet}} = \frac{250 \text{ bytes}}{1.125 \times 10^6 \text{ bytes/sec}} = 222\ \mu\text{s}$$

$$\text{Transmission time}_{\text{ATM}} = \frac{250 \text{ bytes}}{10 \times 10^6 \text{ bytes/sec}} = 25\ \mu\text{s}$$

So the transmission time for the ATM network is about a factor of nine lower.

The total latency to send and receive the packet is the sum of the transmission time and the hardware and software overheads:

$$\text{Total time}_{\text{Ethernet}} = 15 + 6 + 200 + 241 + 222 = 684\ \mu\text{s}$$

$$\text{Total time}_{\text{ATM}} = 50 + 6 + 207 + 360 + 25 = 648\ \mu\text{s}$$

The end-to-end latency of the Ethrnet is only about 1.06 times higher, even though the transmission time is almost 9 times higher!