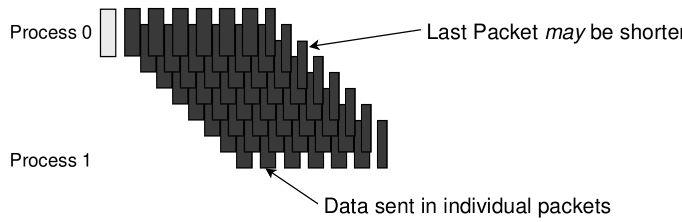
Packetization

- Some networks send data in discrete chunks called *packets*

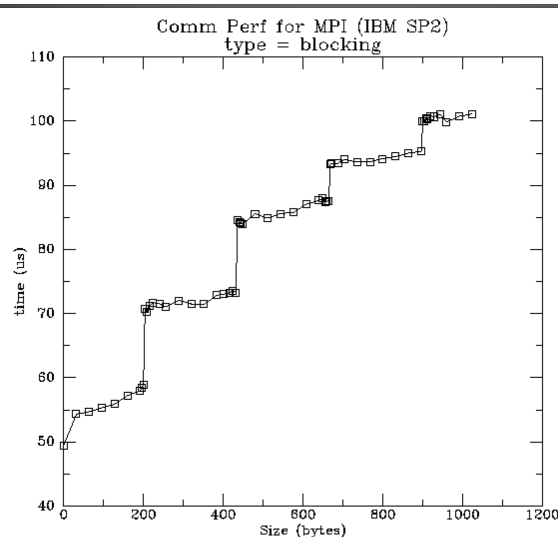


Introduces a ceil(n/packet_size) term

Staircase appearance of performance graph

1

Example of Packetization



Packets contain 232 bytes of data. (first is 200 bytes, so MPI header is probably 32 bytes).

Data from mpptest, available at ftp://ftp.mcs.anl.gov/pub/mpi/misc/perftest.tar.gz

2

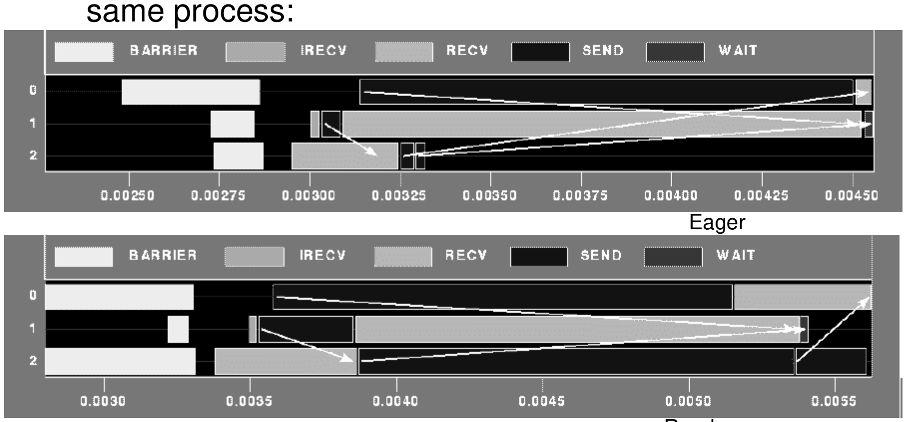
Synchronization Delays

- Message passing is a cooperative method — if the partner doesn't react quickly, a delay results
- There is a performance tradeoff caused by reacting quickly — it requires devoting resources to checking for things to do

3

Observing Synchronization Delays

- 3 processors sending data, with one sending a short message and another sending a long message to the same process:



Eager

Rendezvous

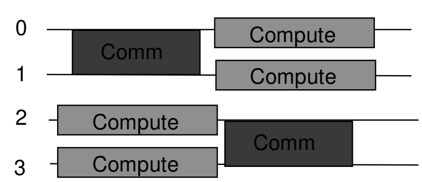
4

Contention

- Point-to-point analysis ignores fact that communications links (usually) shared
- Easiest model is to equally share bandwidth (if K can shared at one time, give each 1/K of the bandwidth).
- "Topology doesn't matter anymore" is *not* true, but there is less you can do about it (just like cache memory)
- MPI has processor topology routines ...

Scheduling for Contention

- Many programs alternate between communication and computation phases
- Contention can reduce effective bandwidth
- Consider restructuring program so that some nodes communicate while others compute:



Effect of contention

- IBM SP2 has a multistage switch. This test shows the point-to-point bandwidth with half the nodes sending and half receiving

Processors	Bandwidth (MB/sec)
2	34
4	34
8	34
16	31
32	25
64	22

IBM SP with TB3 also demonstrates contention effects (as will any multistage system)

7