Timing MPI Programs

- The elapsed (wall-clock) time between two points in an MPI program can be computed using **`MPI_Wtime`**:

```
double t1, t2;
t1 = MPI_Wtime();
...
t2 = MPI_Wtime();
printf( "time is %d\n", t2 - t1 );
```

- The value returned by a single call to **`MPI_Wtime`** has little value.
- Times in general are local, but an implementation might offer synchronized times. See attribute **`MPI_WTIME_IS_GLOBAL`**.

1

Measuring Performance

- Using MPI_Wtime
  - » timers are *not* continuous — MPI_Wtick
- MPI_Wtime is local unless MPI_WTIME_IS_GLOBAL attribute is true
- MPI Profiling interface
- Performance measurement tools

2

Sample Timing Harness

- Average times, make several trials

```
for (k<nloop) {
    t1 = MPI_Wtime();
    for (I<maxloop) {
        <operation to be timed>
        }
    time = MPI_Wtime() - t1;
    if (time < tfinal) tfinal = time;
    }
```

- Use MPI_Wtick to discover clock resolution
- Use getrusage to get other effects (e.g., context switches, paging)

3

Pitfalls in timing

- Time too short:

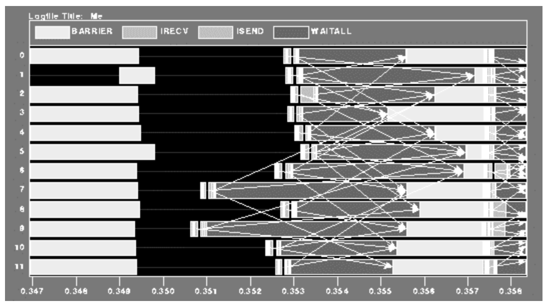
```
t = MPI_Wtime();
MPI_Send(…);
time = MPI_Wtime() - t;
```

- Underestimates by MPI_Wtick, over by cost of calling MPI_Wtime
- "Correcting" MPI_Wtime by subtracting average of MPI_Wtime calls overestimates MPI_Wtime
- Code not paged in (always run at least twice)
- Minimums not what users see
- Tests with 2 processors may not be representative
  - » T3D has processors in pairs, pingpong give 130 MB/sec for 2 but 75 MB/sec for 4 (for MPI_Ssend)

4

Example of Paging Problem

- Black area is *identical* setup computation



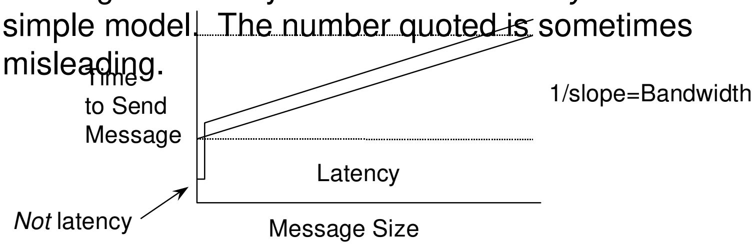
Latency and Bandwidth

- Simplest model $s + r\, n$
- s includes both hardware (gate delays) and software (context switch, setup)
- r includes both hardware (raw bandwidth of interconnection and memory system) and software (packetization, copies between user and system)
- Head-to-head and pingpong values may differ

3

Interpreting Latency and Bandwidth

- Bandwidth is the inverse of the slope of the line
  time = latency + (1/rate) size_of_message
- For performance estimation purposes, latency is the limit(n➜0) of the time to send n bytes
- Latency is sometimes described as “time to send a message of zero bytes”. This is true *only* for the simple model. The number quoted is sometimes misleading.



Exercise: Timing MPI Operations

- Estimate the latency and bandwidth for some MPI operation (e.g., Send/Recv, Bcast, Ssend/Irecv-Wait)
  - » Make sure all processes are ready before starting the test
  - » How repeatable are your measurements?
  - » How does the performance compare to the performance of other operations (e.g., memcpy, floating multiply)?