03-Apr-2012 JEC

MPI Bcast test

MPI_bcast_test is a code for timing mpi_bcast operations. A buffer of 4096*80 characters is broadcast to all processes, firstly as a single operation, then split into 4, 8, 16....4096 segments. The elapsed time for each set of transfers is measured using MPI_WTIME, & the results are gathered & printed by process 0.

A shell script **mpi_bcast_test.csh** runs the executable on a set of nodes, for a specified number of times separated by a set interval, and concatenates the output files from process 0. (The output is formatted as .csv, for processing by a spreadsheet)

```
     Wtime,      nt,       nbt,      Tmax,      Tmin,  jac-64    , jac-65    , jac-66
  143638.984  ,    1,    327680,    108.49,     57.34,      1.34,     57.34,    108.49,
  143639.018  ,    4,     81920,     33.27,     33.19,      1.76,     33.27,     33.19,
  143639.050  ,   16,     20480,     30.85,     30.56,      1.56,     30.85,     30.56,
  143639.085  ,   64,      5120,     34.42,     34.27,      8.68,     34.42,     34.27,
  143639.143  ,  256,      1280,     57.05,     57.03,     33.06,     57.03,     57.05,
  143639.214  , 1024,       320,     70.96,     70.95,     64.62,     70.96,     70.95,
  143639.378  , 4096,        80,    135.31,    135.31,    132.91,    135.31,    135.31,
```

Sample Output

Results

The code was tested on the JET analysis cluster, running Open MPI 1.4. No special tuning was attempted, and all parameters were as set by configure. The mechanisms by which MPI effects the transfers (buffering, block sizes etc) have not been considered, but clearly the transfer size must be less than or equal to the request size.

Figure 1(below) shows a histogram of transfer times for the entire 327.68Kb of data, for 12 transfers between 17 processes (each running on a different machine). The 4096*80 and 1024*320byte modes are significantly slower, while the other modes of operation are all largely below 50mSec, but with significant numbers of outliers with longer times.

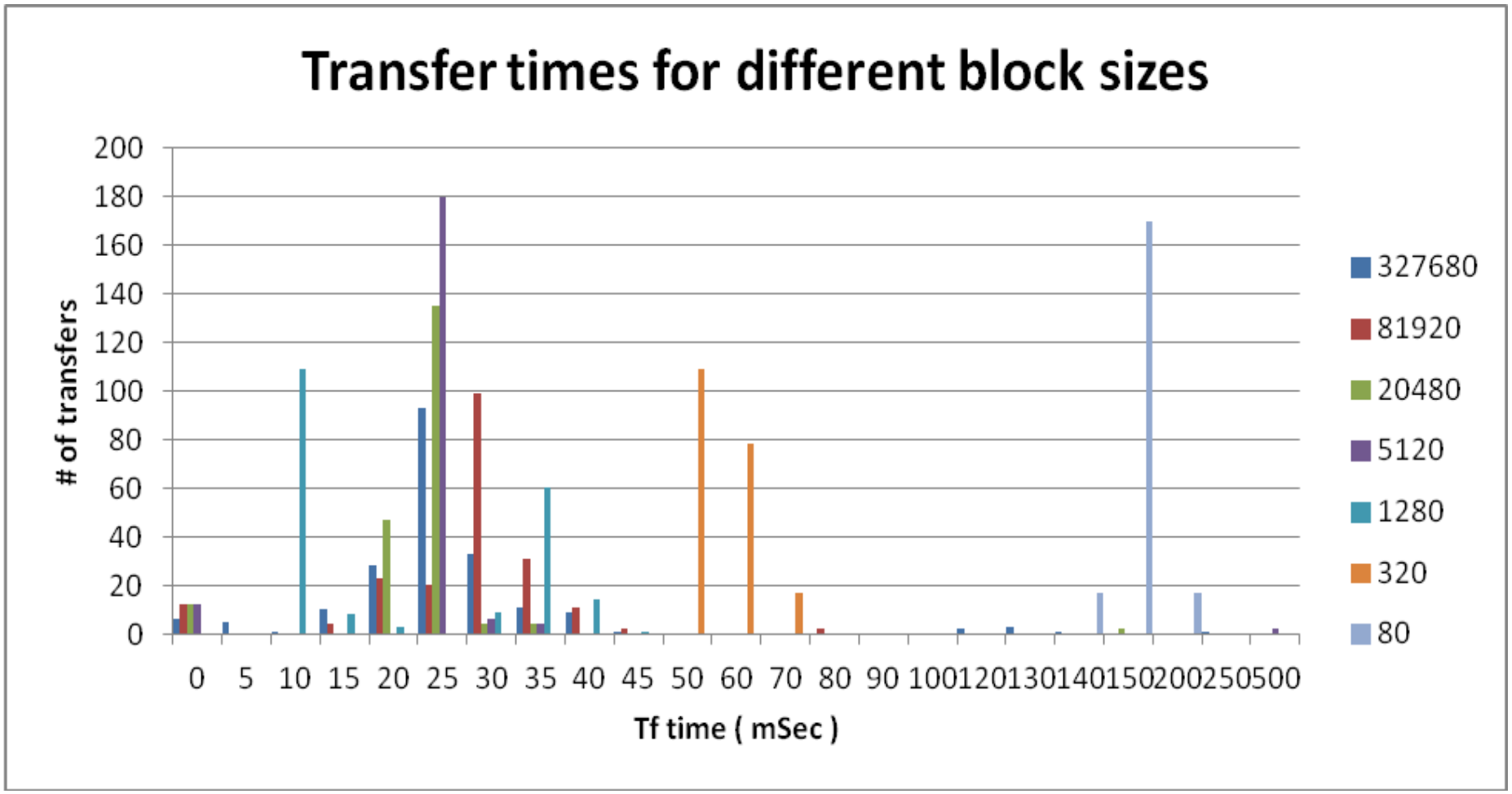

Figure 1 Transfer times between 17 “batch” nodes, 12 iterations at 30minute intervals

The parameter of practical interest is the maximum transfer time, since at some point the results from all processes must be combined, and all will have to wait for the slowest. The sample maximum is a poor estimator of the population maximum, so the time taken for 90% of transfers to complete has been used as a measure for comparison. This is shown in figure 2.

Four different runs are compared, two on the "batch" nodes and two more on "jac" (interactive and batch) nodes. Jac12x24 (red) and batch17x12a (green) were performed during normal working hours, while the other two were run late on a Sunday. This may account for difference between the two jac runs.

Transfer sizes of 5.12 Kb or higher appear to give the best performance, but with some degradation for single 327 Kb transfer on the jac hosts. The optimum settings offer an improvement of nearly 4 over the 80 byte transfers

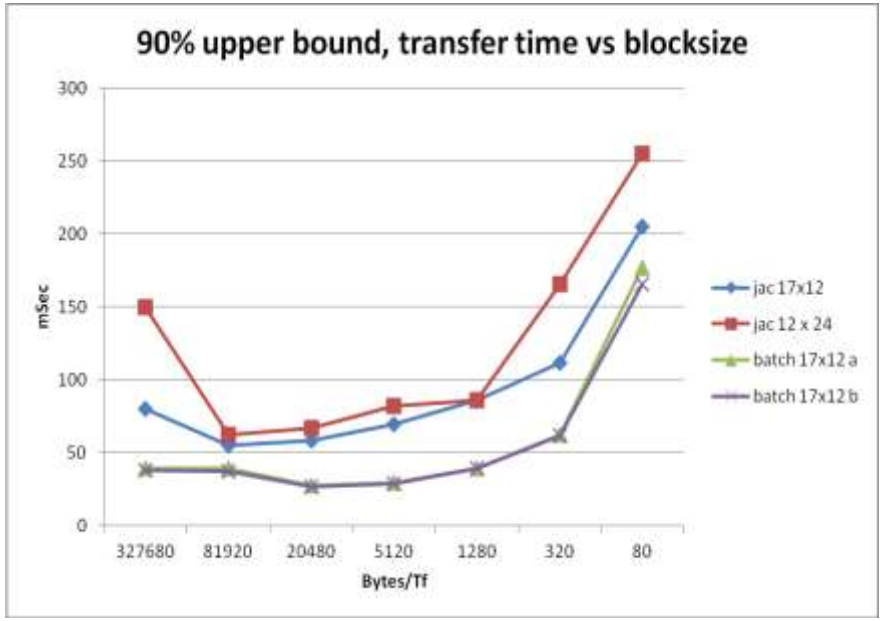

Figure 2 Transfer times