

Mpipview: An MPI Performance Profile Viewer

John Gyllenhaal

Livermore Computing

John May

Center for Applied Scientific Computing

Lawrence Livermore National Laboratory



MpiP Overview

- Scalable, light-weight MPI profiling library
 - Generates detailed text summary of MPI behavior
 - Time spent at each MPI function callsite
 - Bytes sent by each MPI function callsite (where applicable)
 - MPI I/O statistics (in version 2.7, released Aug 3rd, 2004)
 - Configurable traceback depth for function callsites
 - Controllable from program using MPI_Pcontrol
 - Allows you to profile just one code module or cycle
 - Allows mpiP profile dumps mid-run (in version 2.7)
 - Requires only a relink with mpiP libraries
 - IBM's MPI_trace libraries provide similar functionality



MpiP Overview (cont.)

- Freely available and well documented
 - Written by Jeff Vetter and Chris Chembreau
 - Download from <http://www.llnl.gov/CASC/mpip/>
- Supports multiple platforms
 - Power/AIX variations, Pentium4/Linux, IA64/Linux, Alpha Tru64 (see web page for details)
 - Other mpiP ports in the works
 - Function traceback mechanism platform specific
- Polished Text Output
 - Well organized, tuned for ease of finding important info
 - Output text file can become very large
 - E.g., 512 task IRS run yields 12406 lines of text output



MpiP Text Output Examples

@--- Callsite statistics (all, milliseconds): 8 -----

Name	Site	Rank	Count	Max	Mean	Min	App%	MPI%
Barrier	1	0	1	0.107	0.107	0.107	0.00	44.03
Barrier	1	*	4	0.174	0.137	0.107	0.00	0.00
Barrier	2	0	1	0.136	0.136	0.136	0.00	55.97
Barrier	2	1	1	1e+04	1e+04	1e+04	99.92	100.00
Barrier	2	2	1	1e+04	1e+04	1e+04	99.92	100.00
Barrier	2	3	1	1e+04	1e+04	1e+04	99.92	100.00
Barrier	2	*	4	1e+04	7.5e+03	0.136	74.94	100.00

@--- Callsite statistics (all, sent bytes) -----

Name	Site	Rank	Count	Max	Mean	Min	Sum
Send	5	0	80	6000	6000	6000	4.8e+05
Send	5	1	80	6000	6000	6000	4.8e+05
Send	5	2	80	6000	6000	6000	4.8e+05
Send	5	3	80	6000	6000	6000	4.8e+05
Send	5	*	320	6000	6000	6000	1.92e+06

Mpipview: An MpiP Output Viewer

- Organizes and condenses mpiP output
 - Allow users to find key mpiP data quickly
 - Hides complexity of large scale runs until needed
 - Shows source code for the MPI callsites reported on
 - Design based on our experience using mpiP on ASC apps
- Open source, portable, part of Tool Gear
 - Download from http://www.llnl.gov/CASC/tool_gear
 - Requires Qt (download from <http://www.trolltech.com/>)
 - Tested on AIX, Linux, Tru64, and Mac OS X
- Easy to use - parses mpiP text output file
 - `mpipview irs.8.default.mpiP`



Initial View: MPI Timing Summaries

```
1: /g/g0/gyllen/irs_mpiP/magenta/irs.8.default.mpiP on berg01.llnl.gov
File Edit Help
Data read complete
Message List Displayed:
MpiP Callsite Timing Statistics (all, milliseconds) [14 items]
Allreduce[12] 48.70% of MPI 1.23% of App 8/8 Tasks .MPI_Allreduce_Wrapper:145 (FunctionTimer_mpi_wrappers.c)
Isend[6] 22.49% of MPI 0.57% of App 8/8 Tasks .MPI_Isend_Wrapper:389 (FunctionTimer_mpi_wrappers.c)
Waitany[11] 21.76% of MPI 0.55% of App 8/8 Tasks .MPI_Waitany_Wrapper:502 (FunctionTimer_mpi_wrappers.c)
Irecv[4] 4.52% of MPI 0.11% of App 8/8 Tasks .MPI_Irecv_Wrapper:363 (FunctionTimer_mpi_wrappers.c)
Waitall[3] 1.67% of MPI 0.04% of App 8/8 Tasks .MPI_Waitall_Wrapper:480 (FunctionTimer_mpi_wrappers.c)
Beast[5] 0.56% of MPI 0.01% of App 8/8 Tasks .MPI_Beast_Wrapper:202 (FunctionTimer_mpi_wrappers.c)
Wait[10] 0.18% of MPI 0.00% of App 8/8 Tasks .MPI_Wait_Wrapper:524 (FunctionTimer_mpi_wrappers.c)
Allgather[2] 0.05% of MPI 0.00% of App 8/8 Tasks .MPI_Allgather_Wrapper:103 (FunctionTimer_mpi_wrappers.c)
Isend[6] Source
FunctionTimer_mpi_wrappers.c:389 (MPI_Isend_Wrapper)
Unable to locate the source file:
'FunctionTimer_mpi_wrappers.c'
The search path may be modified via 'Edit->Set Search Path...' or 'Ctrl-P'
```

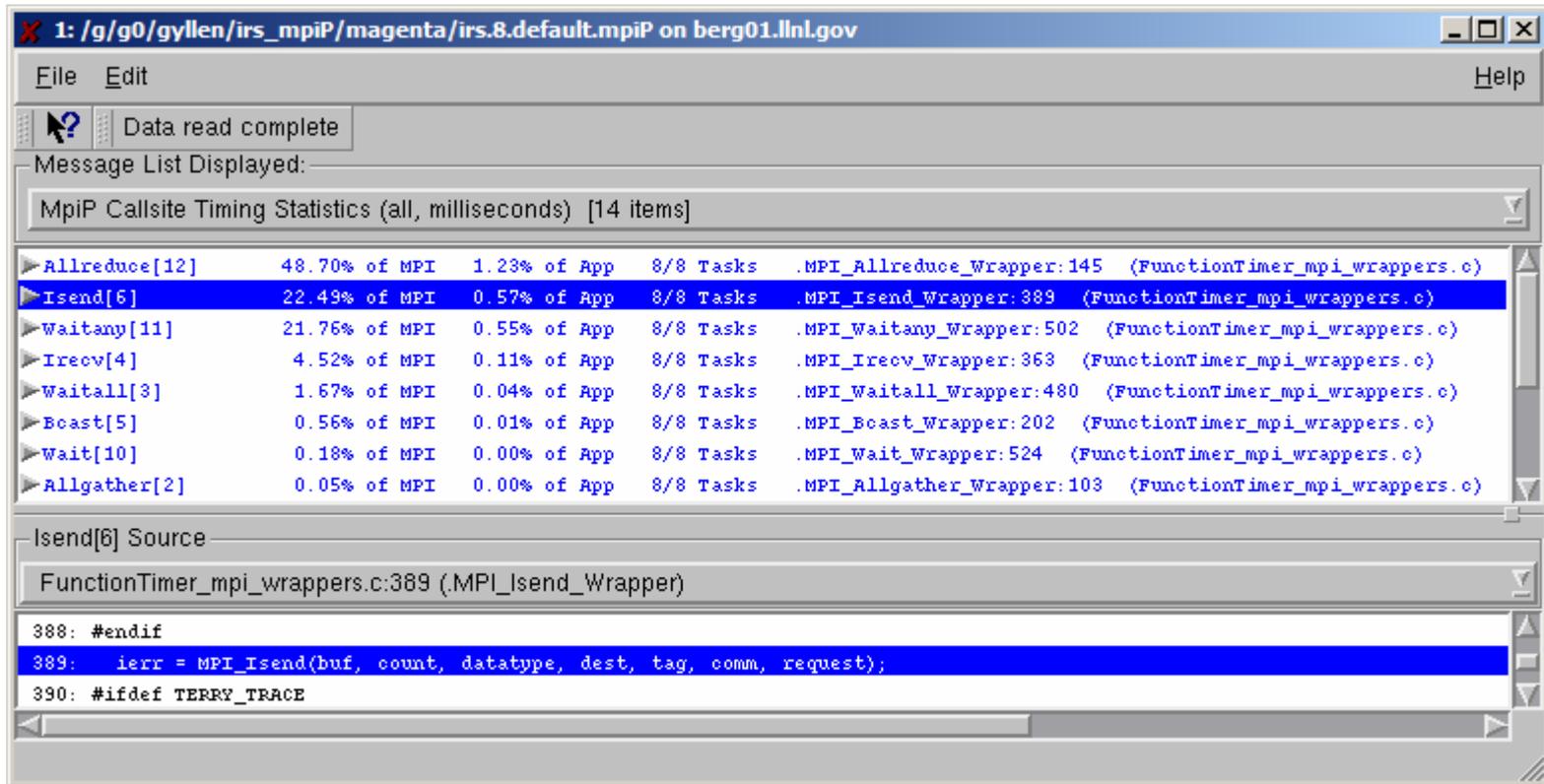
- Shows timing stats summaries, sorted by % of MPI
- May need to set search path to find source code
 - Setting ‘MPIP’ env variable option ‘-n’ will put full path in mpiP output file, if executable contains full path info (-qfullpath for IBM)

Setting Source Code Search Path

The screenshot shows a terminal window on the left and a 'Set search path' dialog box on the right. The terminal window displays MPI timing statistics for various operations, with 'Isend[6]' highlighted. Below the statistics, it shows the source file for 'Isend[6]' as 'FunctionTimer_mpi_wrappers.c:38'. An error message follows: 'Unable to locate the source file: 'FunctionTimer_mpi_wrappers.c''. The dialog box, titled 'Set search path', has a tree view of directories. The selected path is '/g/g0/gyllen/irs-1.4/sources'. Below the tree, there are buttons for 'Copy to path', 'Remove from path', 'Show directory', and 'Paste from clipboard'. A text field contains the path '/g/g0/gyllen/irs-1.4/sources'. To the right, there is a 'Recursive search?' section with a 'Search directories' list containing three entries: '/g/g0/gyllen/irs-1.4/sources', '/g/g0/gyllen/benchmarks', and '/g/g0/gyllen/demo'. The first two are checked. At the bottom, there is an 'Apply...' section with three radio button options: '...to current session only', '...and save in current directory', and '...and save in home directory'. The 'OK' and 'Cancel' buttons are also present.

- Modified via 'Edit->Set Search Path...' or 'Ctrl-P'
- Supports recursive search of complex directory tree
 - Can specify search directory order and which to recursively search
- Can save settings in current or home directory

MPI Callsite Timing Summaries



The screenshot shows a terminal window titled "1: /g/g0/gyllen/irs_mpiP/magenta/irs.8.default.mpiP on berg01.llnl.gov". The window displays a message list with the following data:

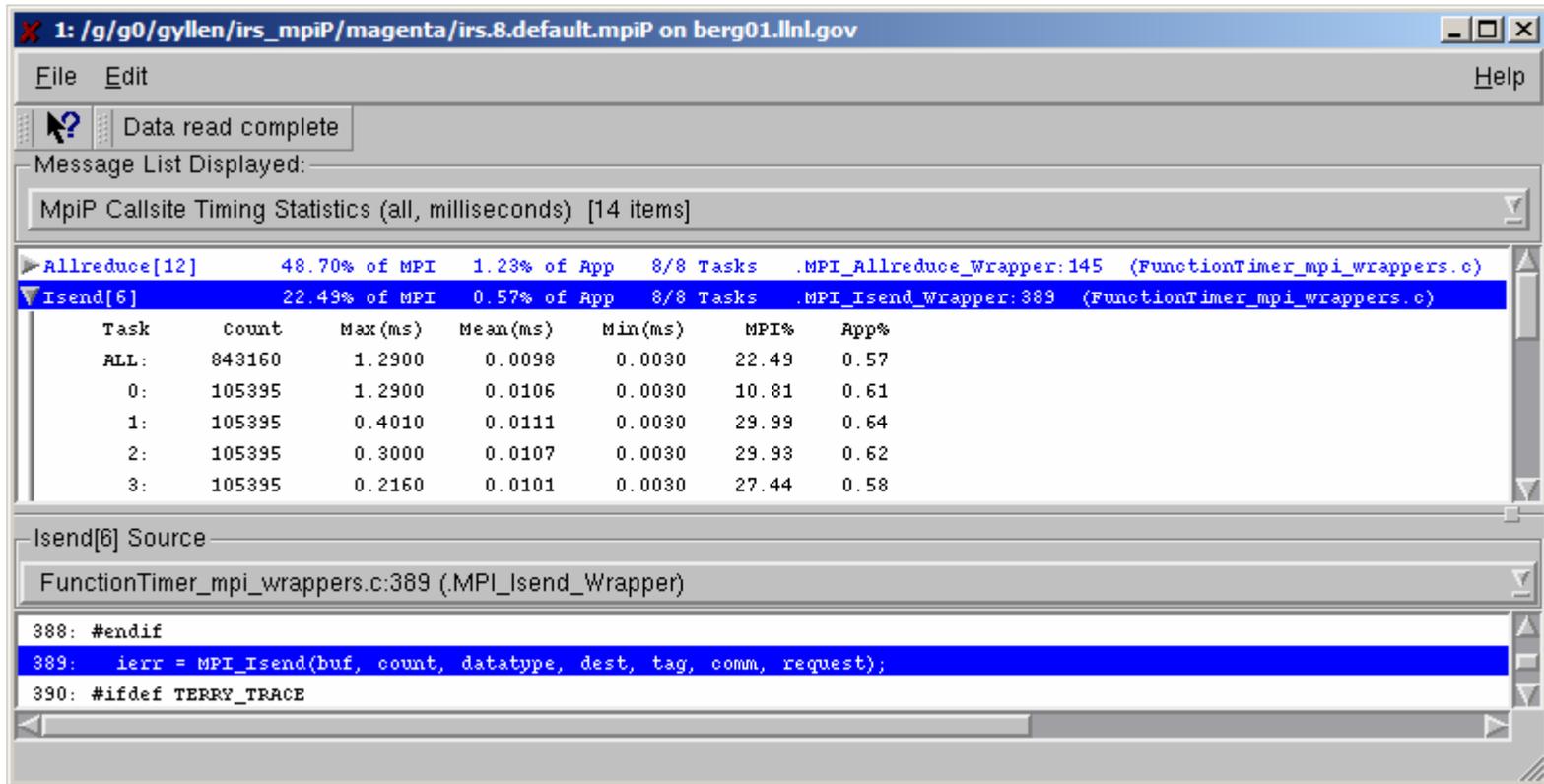
Operation	% of MPI	% of App	Tasks	Wrapper	Source
Allreduce[12]	48.70%	1.23%	8/8	.MPI_Allreduce_Wrapper:145	(FunctionTimer_mpi_wrappers.c)
Isend[6]	22.49%	0.57%	8/8	.MPI_Isend_Wrapper:389	(FunctionTimer_mpi_wrappers.c)
Waitany[11]	21.76%	0.55%	8/8	.MPI_Waitany_Wrapper:502	(FunctionTimer_mpi_wrappers.c)
Irecv[4]	4.52%	0.11%	8/8	.MPI_Irecv_Wrapper:363	(FunctionTimer_mpi_wrappers.c)
Waitall[3]	1.67%	0.04%	8/8	.MPI_Waitall_Wrapper:480	(FunctionTimer_mpi_wrappers.c)
Beast[5]	0.56%	0.01%	8/8	.MPI_Beast_Wrapper:202	(FunctionTimer_mpi_wrappers.c)
Wait[10]	0.18%	0.00%	8/8	.MPI_Wait_Wrapper:524	(FunctionTimer_mpi_wrappers.c)
Allgather[2]	0.05%	0.00%	8/8	.MPI_Allgather_Wrapper:103	(FunctionTimer_mpi_wrappers.c)

Below the table, the source code for the selected Isend[6] callsite is shown:

```
FunctionTimer_mpi_wrappers.c:389 (.MPI_Isend_Wrapper)
388: #endif
389: ierr = MPI_Isend(buf, count, datatype, dest, tag, comm, request);
390: #ifdef TERRY_TRACE
```

- Clicking on summary displays callsite's source code
 - Callsites indicate where an MPI call was called from
 - Each callsite is tracked separately in mpiP
 - Isend[6] indicates the 6th MPI callsite reached was an MPI_Isend

MPI Callsite Timing Details



The screenshot shows a window titled "1: /g/g0/gyllen/irs_mpiP/magenta/irs.8.default.mpiP on berg01.llnl.gov". The window displays "MpiP Callsite Timing Statistics (all, milliseconds) [14 items]". The statistics are as follows:

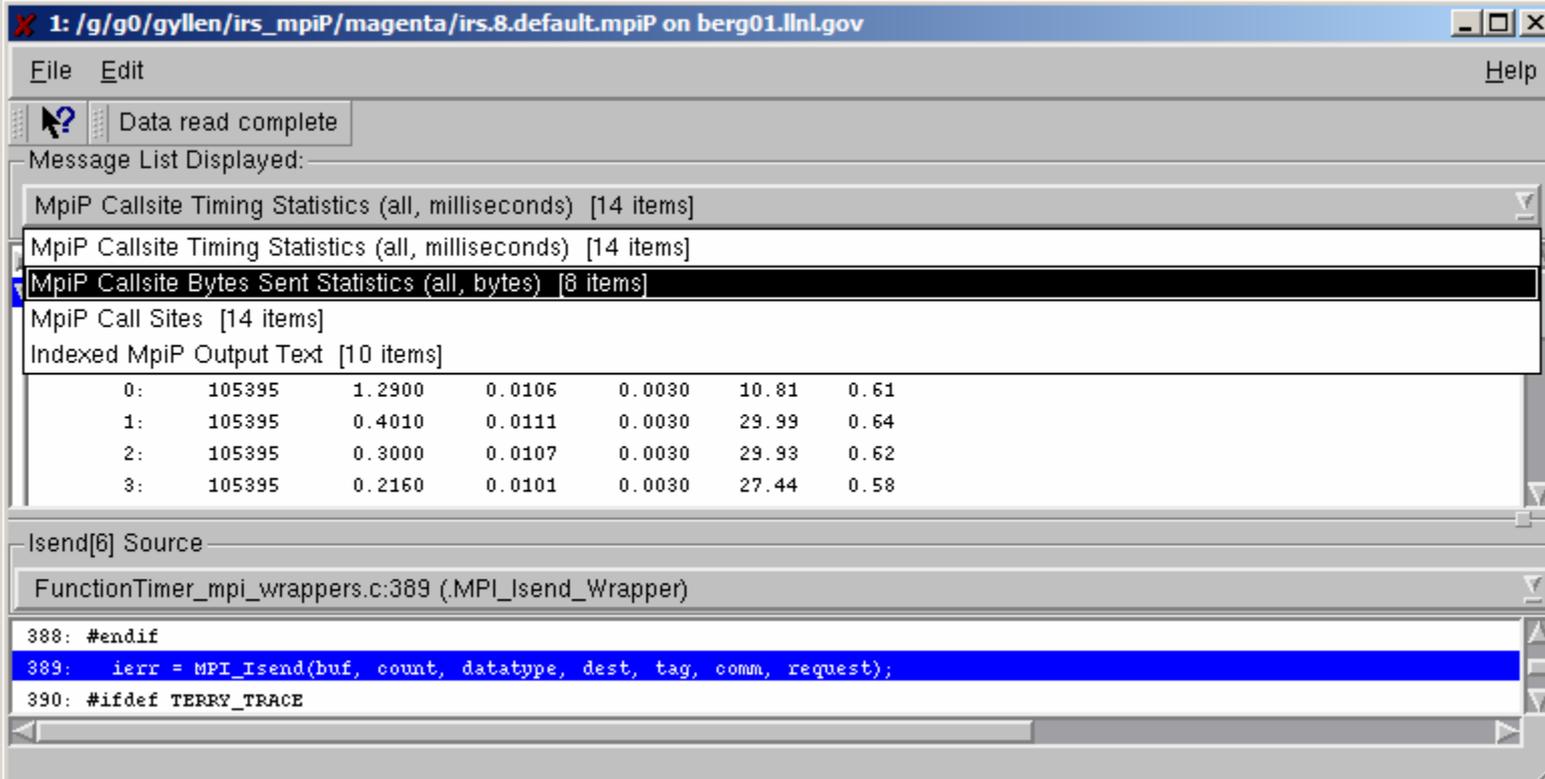
Task	Count	Max(ms)	Mean(ms)	Min(ms)	MPI%	App%
ALL:	843160	1.2900	0.0098	0.0030	22.49	0.57
0:	105395	1.2900	0.0106	0.0030	10.81	0.61
1:	105395	0.4010	0.0111	0.0030	29.99	0.64
2:	105395	0.3000	0.0107	0.0030	29.93	0.62
3:	105395	0.2160	0.0101	0.0030	27.44	0.58

The window also shows the source code for the Isend[6] operation, highlighting the MPI_Isend function call on line 389:

```
388: #endif  
389: ierr = MPI_Isend(buf, count, datatype, dest, tag, comm, request);  
390: #ifdef TERRY_TRACE
```

- Double click on summary (or click arrow) to show/hide
 - ALL: Aggregate over all tasks that reached callsite (8/8 Tasks)
 - 2: Displays task 2's details (only shows tasks that reached callsite)

Selecting MpiP Output Sections



1: /g/g0/gyllen/irs_mpiP/magenta/irs.8.default.mpiP on berg01.llnl.gov

File Edit Help

Data read complete

Message List Displayed:

- MpiP Callsite Timing Statistics (all, milliseconds) [14 items]
- MpiP Callsite Bytes Sent Statistics (all, bytes) [8 items]
- MpiP Call Sites [14 items]
- Indexed MpiP Output Text [10 items]

0:	105395	1.2900	0.0106	0.0030	10.81	0.61
1:	105395	0.4010	0.0111	0.0030	29.99	0.64
2:	105395	0.3000	0.0107	0.0030	29.93	0.62
3:	105395	0.2160	0.0101	0.0030	27.44	0.58

Isend[6] Source

FunctionTimer_mpi_wrappers.c:389 (MPI_Isend_Wrapper)

```
388: #endif
389: ierr = MPI_Isend(buf, count, datatype, dest, tag, comm, request);
390: #ifdef TERRY_TRACE
```

- Several other mpiP “message lists” are displayable
 - [8 items] indicates there are eight “Byte Sent” statistics available
 - Select list (clicking, arrow keys, mouse scroll wheel) to view
 - Lists available depend on MPI calls exercised (e.g., no MPI I/O)

MPI Callsite Data Sent Summaries

1: /g/g0/gyllen/irs_mpiP/magenta/irs.8.default.mpiP on berg01.llnl.gov

File Edit Help

Data read complete

Message List Displayed:

MpiP Callsite Bytes Sent Statistics (all, bytes) [8 items]

MPI Call	% of MPI	Total Bytes	Mean Bytes	Tasks	Wrapper
Allreduce[12]	48.70%	4607000	40.96	8/8	.MPI_Allreduce_Wrapper:145 (FunctionI
Isend[6]	22.49%	3.169e+09	3759	8/8	.MPI_Isend_Wrapper:389 (FunctionI
Beast[5]	0.56%	516400	140.6	8/8	.MPI_Beast_Wrapper:202 (FunctionI
Allgather[2]	0.05%	2528	4	8/8	.MPI_Allgather_Wrapper:103 (Funct
Gatherv[14]	0.04%	13660	7.146	8/8	.MPI_Gatherv_Wrapper:312 (Functio
Gather[13]	0.03%	9728	6.909	8/8	.MPI_Gather_Wrapper:297 (Function
Reduce[7]	0.00%	4416	69	8/8	.MPI_Reduce_Wrapper:435 (Function

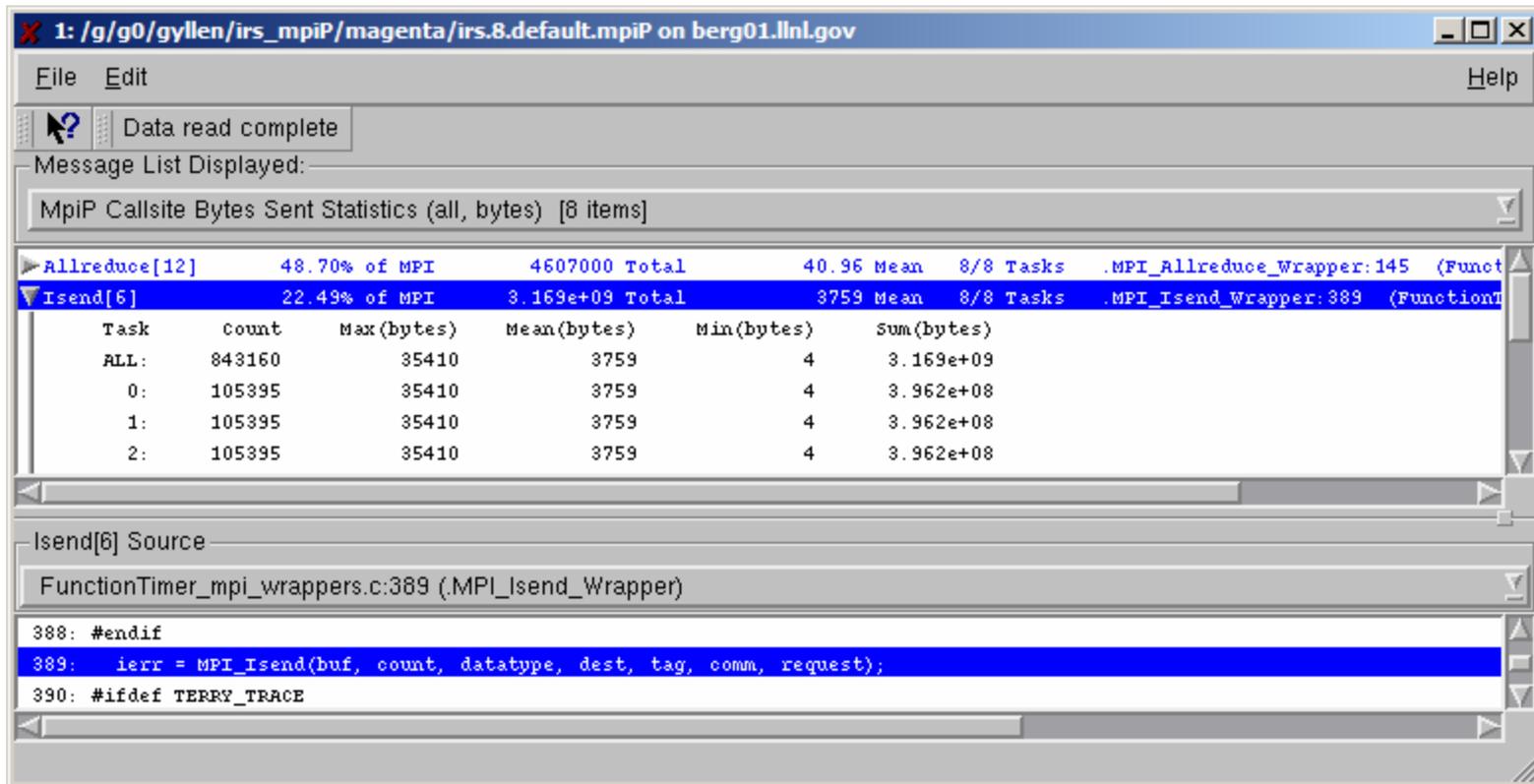
Isend[6] Source

FunctionTimer_mpi_wrappers.c:389 (.MPI_Isend_Wrapper)

```
388: #endif
389:  ierr = MPI_Isend(buf, count, datatype, dest, tag, comm, request);
390: #ifdef TERRY_TRACE
```

- Summary shows aggregate and mean bytes sent
 - Summaries ordered by % MPI, not bytes sent
 - Only data sending MPI calls shown (i.e., no barriers, receives, etc.)

MPI Callsite Byte Sent Details



The screenshot shows a window titled "1: /g/g0/gyllen/irs_mpiP/magenta/irs.8.default.mpiP on berg01.llnl.gov". The window displays MPI Callsite Bytes Sent Statistics. The summary table is as follows:

Operation	Percentage of MPI	Total Bytes	Mean (bytes)	Tasks	Function
Allreduce[12]	48.70%	4607000	40.96	8/8	.MPI_Allreduce_Wrapper:145 (FunctionT
Isend[6]	22.49%	3.169e+09	3759	8/8	.MPI_Isend_Wrapper:389 (FunctionT

Task	Count	Max(bytes)	Mean(bytes)	Min(bytes)	Sum(bytes)
ALL:	843160	35410	3759	4	3.169e+09
0:	105395	35410	3759	4	3.962e+08
1:	105395	35410	3759	4	3.962e+08
2:	105395	35410	3759	4	3.962e+08

The source code for Isend[6] is shown below:

```
FunctionTimer_mpi_wrappers.c:389 (.MPI_Isend_Wrapper)  
388: #endif  
389: ierr = MPI_Isend(buf, count, datatype, dest, tag, comm, request);  
390: #ifdef TERRY_TRACE
```

- Double click on summary (or click arrow) to show/hide
 - ALL: Aggregate over all tasks that reached callsite (8/8 Tasks)
 - 2: Displays task 2's details (only shows tasks that reached callsite)

Dealing with MPI Wrappers

The screenshot shows a terminal window titled "1: /g/g0/gyllen/irs_mpiP/magenta/irs.8.4deep.mpiP on berg01.llnl.gov". The window displays MPIIP Callsite Timing Statistics for 114 items. The statistics are as follows:

Call	% of MPI	% of App	Tasks	Wrapper	Source
Allreduce[111]	17.90%	0.50%	8/8	.MPI_Allreduce_Wrapper:145	(FunctionTimer_mpi_wrappers.c)
Isend[27]	9.22%	0.26%	8/8	.MPI_Isend_Wrapper:389	(FunctionTimer_mpi_wrappers.c)
Waitany[110]	8.92%	0.25%	8/8	.MPI_Waitany_Wrapper:502	(FunctionTimer_mpi_wrappers.c)

Below the statistics, the source of Isend[27] is shown with a traceback of 4 levels:

```
[1] FunctionTimer_mpi_wrappers.c:389 (.MPI_Isend_Wrapper)
===== Isend[27] Source =====
[1] FunctionTimer_mpi_wrappers.c:389 (.MPI_Isend_Wrapper)
[2] combuf.c:316 (.postbuf)
[3] rbndcom.c:737 (.rbndcom)
[4] MatrixSolve.c:206 (.MatrixSolveCG)
===== Raw MpiP Data =====
/g/g0/gyllen/irs_mpiP/magenta/irs.8.4deep.mpiP:1084
```

- Adding '-k 4' to MPIP env var selects 4 levels traceback
 - May have significantly more callsites, one for each distinct traceback
- Useful when MPI calls are buried in user MPI wrappers
 - Some codes require very long tracebacks to get useful information

GUI Handles Large MpiP Files Well

1: /g/g0/gyllen/irs_mpiP/magenta/irs.512.default.mpiP on berg01.llnl.gov

File Edit Help

Data read complete

Message List Displayed:

MpiP Callsite Timing Statistics (all, milliseconds) [14 items]

Task	Count	Max(ms)	Mean(ms)	Min(ms)	MPI%	App%
511:	80866	224.0000	6.1300	0.1410	70.12	32.32
Waitany[11]		41.83% of MPI	17.88% of App	512/512 Tasks	.MPI_Waitany_Wrapper:502	(FunctionTimer_mpi_wrappers.c)
ALL:	883975664	231.0000	0.1590	0.0060	41.83	17.88
0:	664144	154.0000	0.2400	0.0070	23.86	10.40
1:	996216	157.0000	0.1810	0.0060	27.51	11.78
2:	996216	157.0000	0.1910	0.0070	29.15	12.44

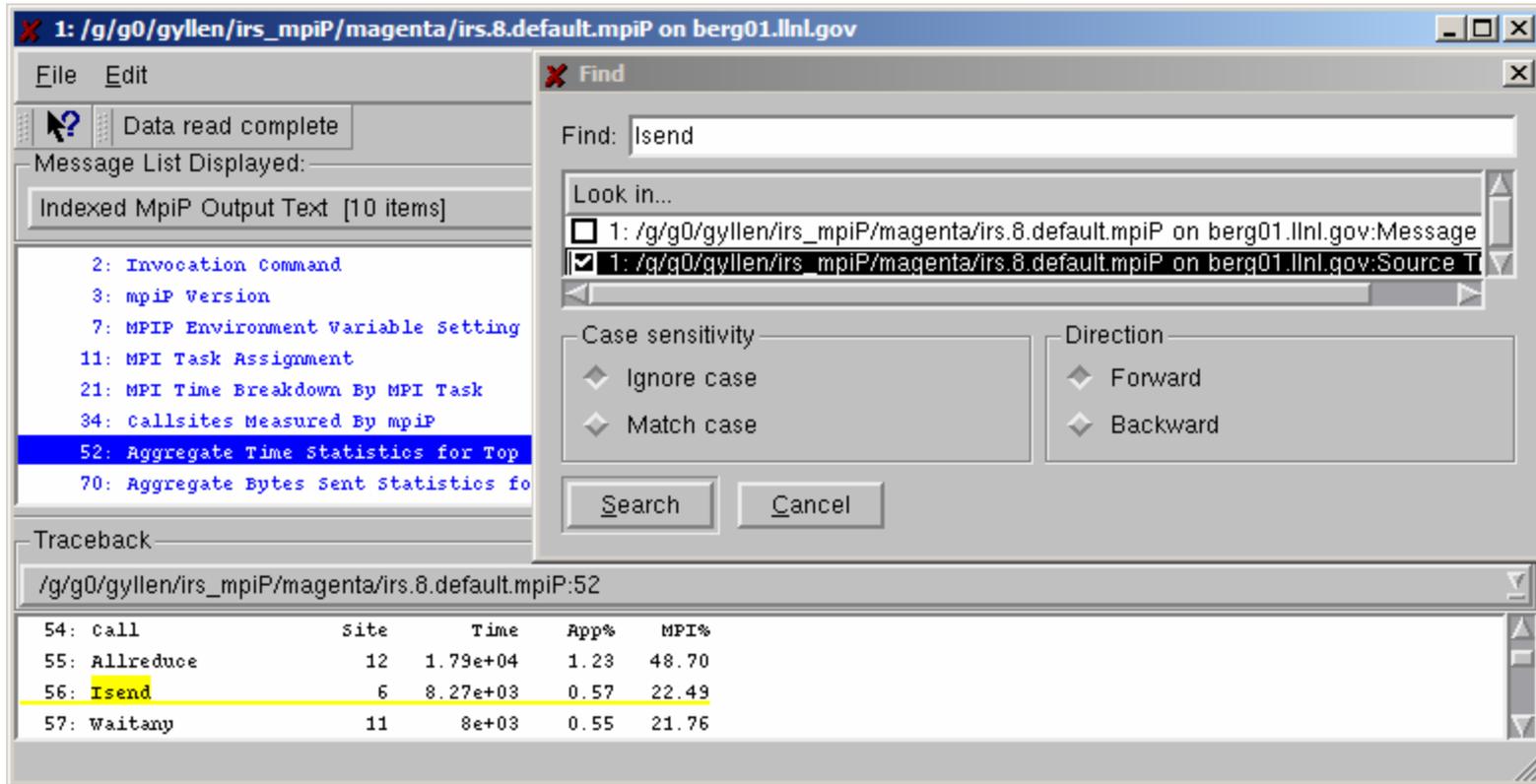
Waitany[11] Source

FunctionTimer_mpi_wrappers.c:502 (.MPI_Waitany_Wrapper)

```
502: ierr = MPI_Waitany(count, array_of_requests, index, status);
```

- 512 task IRS mpiP output is 12406 lines of text
 - Navigating large outputs is where mpipview becomes essential
- Mpipview's GUI is tuned for performance
 - Displayed > 1 million lines of messages in ~6 seconds on Power4

Finding Specific Text



The screenshot shows a software window titled "1: /g/g0/gyllen/irs_mpiP/magenta/irs.8.default.mpiP on berg01.llnl.gov". The window has a menu bar with "File" and "Edit". Below the menu bar, there is a status bar that says "Data read complete" and a message list that says "Message List Displayed: Indexed mpiP Output Text [10 items]". The main area of the window displays a list of MPI statistics, with the following items:

- 2: Invocation Command
- 3: mpiP Version
- 7: MPIP Environment Variable Setting
- 11: MPI Task Assignment
- 21: MPI Time Breakdown By MPI Task
- 34: Callsites Measured By mpiP
- 52: Aggregate Time Statistics for Top
- 70: Aggregate Bytes Sent Statistics fo

The "Find" dialog box is open, showing the search term "Isend". The "Look in..." section has two options: "1: /g/g0/gyllen/irs_mpiP/magenta/irs.8.default.mpiP on berg01.llnl.gov:Message" (unchecked) and "1: /g/g0/gyllen/irs_mpiP/magenta/irs.8.default.mpiP on berg01.llnl.gov:Source T" (checked). The "Case sensitivity" section has two options: "Ignore case" (checked) and "Match case" (unchecked). The "Direction" section has two options: "Forward" (checked) and "Backward" (unchecked). The "Search" and "Cancel" buttons are visible at the bottom of the dialog box.

Below the "Find" dialog box, the "Traceback" section shows the path "/g/g0/gyllen/irs_mpiP/magenta/irs.8.default.mpiP:52". The main area of the window displays a table of MPI statistics, with the following data:

	Site	Time	App%	MPI%
54: Call				
55: Allreduce	12	1.79e+04	1.23	48.70
56: Isend	6	8.27e+03	0.57	22.49
57: Waitany	11	8e+03	0.55	21.76

- Find dialog via “Edit->Find...” or Ctrl-F
 - Useful for finding specific MPI calls, locations, or tasks ids
 - Only searches “open” messages and source currently displayed

Future work

- Support for mpiP 2.7 (released Aug 3rd, 2004)
 - Support tweaked mpiP 2.7 output format (done)
 - MPI I/O profile support (started)
 - Release of Mpipview 1.2 targeted for end of August, 2004
- Developing message viewers for other tools
 - Tool Gear's streamlined interface makes it easy
 - Just three API calls to pass all the data, Tool Gear does rest
 - Our next target: Umpire
 - MPI correctness tool written by Bronis de Supinski
 - Generates voluminous text output in multiple files
 - Requires support for multiple source tracebacks (started)
 - Support for other tool developers that are using Tool Gear
 - We are willing to reprioritize Tool Gear's implementation plan



UCRL-PRES-205849

Work performed under the auspices of the U. S. Department of Energy by the University of California, Lawrence Livermore National Laboratory under Contract W-7405-Eng-48

MpiP Libraries and documentation available at:
<http://www.llnl.gov/CASC/mpip/>

Mpipview tool and documentation available at:
http://www.llnl.gov/CASC/tool_gear

