The Message Passing Interface: MPI 3.1 and Plans for MPI 4.0

Martin Schulz
LLNL / CASC
Chair of the MPI Forum

MPI Forum BOF @ SC14

http://www.mpi-forum.org/

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344.

Overview

- **Current State of MPI**
  - Features in MPI 3.0
  - Implementation status
- **Timeline for MPI 3.1**
- **Initiatives for MPI 4.0**
  - Fault tolerance
  - Support for hybrid programming
  - Stream/channel communication
- **How to contribute to the MPI Forum**

Let's keep this interactive – Please feel free to ask questions!

Where We Are

- **MPI 3.0 ratified in September 2012**
  - Available at http://www.mpi-forum.org/
  - 852 pages, 430 functions
  - Adaption in most MPIs progressing fast

MPI 3.0 has 430 Functions

MPI_ABORT
MPI_ACCUMULATE
MPI_ADD_ERROR_CLASS
MPI_ADD_ERROR_CODE
MPI_ADD_ERROR_STRING
MPI_ADDRESS
MPI_ALLGATHER
MPI_ALLGATHERV
MPI_ALLOC_MEM
MPI_ALLOC_MEM_CPTR
MPI_ALLREDUCE
MPI_ALLTOALL
MPI_ALLTOALLV
MPI_ALLTOALLW
MPI_ATTR_DELETE
MPI_ATTR_GET
MPI_ATTR_PUT
MPI_BARRIER
MPI_BCAST
MPI_BSEND
MPI_BSEND_INIT
MPI_BUFFER_ATTACH
MPI_BUFFER_DETACH
MPI_CANCEL
MPI_CART_COORDS
MPI_CART_CREATE
MPI_CART_GET
MPI_CART_MAP
MPI_CART_RANK
MPI_CART_SHIFT
MPI_CART_SUB
MPI_CARTDIM_GET
MPI_CLOSE_PORT
MPI_COMM_ACCEPT
MPI_COMM_C2F
MPI_COMM_CALL_ERRHANDLER
MPI_COMM_COMPARE
MPI_COMM_CONNECT
MPI_COMM_CREATE
MPI_COMM_CREATE_ERRHANDLER
MPI_COMM_CREATE_GROUP
MPI_COMM_CREATE_KEYVAL
MPI_COMM_DELETE_ATTR
MPI_COMM_DISCONNECT
MPI_COMM_DUP
MPI_COMM_DUP_FN
MPI_COMM_DUP_WITH_INFO
MPI_COMM_F2C
MPI_COMM_FREE
MPI_COMM_FREE_KEYVAL
MPI_COMM_GET_ATTR
MPI_COMM_GET_ERRHANDLER
MPI_COMM_GET_INFO
MPI_COMM_GET_NAME
MPI_COMM_GET_PARENT
MPI_COMM_GROUP
MPI_COMM_IDUP
MPI_COMM_JOIN
MPI_COMM_KEYVAL_CREATE
MPI_COMM_NULL_COPY_FN
MPI_COMM_NULL_DELETE_FN
MPI_COMM_RANK
MPI_COMM_REMOTE_GROUP
MPI_COMM_REMOTE_SIZE
MPI_COMM_SET_ATTR
MPI_COMM_SET_ERRHANDLER
MPI_COMM_SET_INFO
MPI_COMM_SET_NAME
MPI_COMM_SIZE
MPI_COMM_SPAWN
MPI_COMM_SPAWN_MULTIPLE
MPI_COMM_SPLIT
MPI_COMM_SPLIT_TYPE
MPI_COMM_TEST_INTER
MPI_COMM_WORLD
MPI_COMPARE_AND_SWAP
MPI_CONVERSION_FN_NULL
MPI_DIMS_CREATE
MPI_DIST_GRAPH_CREATE
MPI_DIST_GRAPH_CREATE_ADJACENT
MPI_DIST_GRAPH_NEIGHBOR_COUNT
MPI_DIST_GRAPH_NEIGHBORS
MPI_DIST_GRAPH_NEIGHBORS_COUNT
MPI_DUP_FN
MPI_ERRHANDLER_C2F
MPI_ERRHANDLER_CREATE
MPI_ERRHANDLER_F2C
MPI_ERRHANDLER_FREE
MPI_ERRHANDLER_GET
MPI_ERRHANDLER_SET
MPI_ERROR_CLASS
MPI_ERROR_STRING
MPI_EXSCAN
MPI_F_SYNC_REG
MPI_FETCH_AND_OP
MPI_FILE_C2F
MPI_FILE_CALL_ERRHANDLER
MPI_FILE_CLOSE
MPI_FILE_CREATE_ERRHANDLER
MPI_FILE_DELETE
MPI_FILE_F2C
MPI_FILE_GET_AMODE
MPI_FILE_GET_ATOMICITY
MPI_FILE_GET_BYTE_OFFSET
MPI_FILE_GET_ERRHANDLER
MPI_FILE_GET_GROUP
MPI_FILE_GET_INFO
MPI_FILE_GET_POSITION
MPI_FILE_GET_POSITION_SHARED
MPI_FILE_GET_SIZE
MPI_FILE_GET_TYPE_EXTENT
MPI_FILE_GET_VIEW
MPI_FILE_IREAD
MPI_FILE_IREAD_AT
MPI_FILE_IREAD_SHARED
MPI_FILE_IWRITE
MPI_FILE_IWRITE_AT
MPI_FILE_IWRITE_SHARED
MPI_FILE_OPEN
MPI_FILE_PREALLOCATE
MPI_FILE_READ
MPI_FILE_READ_ALL
MPI_FILE_READ_ALL_BEGIN
MPI_FILE_READ_ALL_END
MPI_FILE_READ_AT
MPI_FILE_READ_AT_ALL
MPI_FILE_READ_AT_ALL_BEGIN
MPI_FILE_READ_AT_ALL_END
MPI_FILE_READ_ORDERED
MPI_FILE_READ_ORDERED_BEGIN
MPI_FILE_READ_ORDERED_END
MPI_FILE_READ_SHARED
MPI_FILE_SEEK
MPI_FILE_SEEK_SHARED
MPI_FILE_SET_ATOMICITY
MPI_FILE_SET_ERRHANDLER
MPI_FILE_SET_INFO
MPI_FILE_SET_SIZE
MPI_FILE_SET_VIEW
MPI_FILE_SYNC
MPI_FILE_WRITE
MPI_FILE_WRITE_ALL
MPI_FILE_WRITE_ALL_BEGIN
MPI_FILE_WRITE_ALL_END
MPI_FILE_WRITE_AT
MPI_FILE_WRITE_AT_ALL
MPI_FILE_WRITE_AT_ALL_BEGIN
MPI_FILE_WRITE_AT_ALL_END
MPI_FILE_WRITE_ORDERED
MPI_FILE_WRITE_ORDERED_BEGIN
MPI_FILE_WRITE_ORDERED_END
MPI_FILE_WRITE_SHARED
MPI_FINALIZE
MPI_FINALIZED
MPI_FREE_MEM
MPI_GATHER
MPI_GATHERV
MPI_GET
MPI_GET_ACCUMULATE
MPI_GET_ADDRESS
MPI_GET_COUNT
MPI_GET_ELEMENTS
MPI_GET_ELEMENTS_X
MPI_GET_LIBRARY_VERSION
MPI_GET_PROCESSOR_NAME
MPI_GET_VERSION
MPI_GRAPH_CREATE
MPI_GRAPH_GET
MPI_GRAPH_MAP
MPI_GRAPH_NEIGHBORS
MPI_GRAPH_NEIGHBORS_COUNT
MPI_GRAPHDIMS_GET
MPI_GREQUEST_COMPLETE
MPI_GREQUEST_START
MPI_GROUP_C2F
MPI_GROUP_COMPARE
MPI_GROUP_DIFFERENCE
MPI_GROUP_EXCL
MPI_GROUP_F2C
MPI_GROUP_FREE
MPI_GROUP_INCL
MPI_GROUP_INTERSECTION
MPI_GROUP_RANGE_EXCL
MPI_GROUP_RANGE_INCL
MPI_GROUP_RANK
MPI_GROUP_SIZE
MPI_GROUP_TRANSLATE_RANKS
MPI_GROUP_UNION
MPI_IALLGATHER
MPI_IALLGATHERV
MPI_IALLREDUCE
MPI_IALLTOALL
MPI_IALLTOALLV
MPI_IALLTOALLW
MPI_IBARRIER
MPI_IBCAST
MPI_IBSEND
MPI_IEXSCAN
MPI_IGATHER
MPI_IGATHERV
MPI_IMPROBE
MPI_IMRECV
MPI_INEIGHBOR_ALLGATHER
MPI_INEIGHBOR_ALLGATHERV
MPI_INEIGHBOR_ALLTOALL
MPI_INEIGHBOR_ALLTOALLV
MPI_INEIGHBOR_ALLTOALLW
MPI_INFO_C2F
MPI_INFO_CREATE
MPI_INFO_DELETE
MPI_INFO_DUP
MPI_INFO_F2C
MPI_INFO_FREE
MPI_INFO_GET
MPI_INFO_GET_NKEYS
MPI_INFO_GET_NTHKEY
MPI_INFO_GET_VALUELEN
MPI_INFO_SET
MPI_INIT
MPI_INIT_THREAD
MPI_INITIALIZED
MPI_INTERCOMM_CREATE
MPI_INTERCOMM_MERGE
MPI_IPROBE
MPI_IRECV
MPI_IREDUCE
MPI_IREDUCE_SCATTER
MPI_IREDUCE_SCATTER_BLOCK
MPI_IRSEND
MPI_IS_THREAD_MAIN
MPI_ISCAN
MPI_ISCATTER
MPI_ISCATTERV
MPI_ISEND
MPI_ISSEND
MPI_KEYVAL_CREATE
MPI_KEYVAL_FREE
MPI_LOCK_ALL
MPI_LOOKUP_NAME
MPI_MESSAGE_C2F
MPI_MESSAGE_F2C
MPI_MPROBE
MPI_MRECV
MPI_NEIGHBOR_ALLGATHER
MPI_NEIGHBOR_ALLGATHERV
MPI_NEIGHBOR_ALLTOALL
MPI_NEIGHBOR_ALLTOALLV
MPI_NEIGHBOR_ALLTOALLW
MPI_NULL_COPY_FN
MPI_NULL_DELETE_FN
MPI_OP_C2F
MPI_OP_COMMUTATIVE
MPI_OP_CREATE
MPI_OP_F2C
MPI_OP_FREE
MPI_OPEN_PORT
MPI_PACK
MPI_PACK_EXTERNAL
MPI_PACK_EXTERNAL_SIZE
MPI_PACK_SIZE
MPI_PCONTROL
MPI_PROBE
MPI_PUBLISH_NAME
MPI_PUT
MPI_QUERY_THREAD
MPI_RACCUMULATE
MPI_RECV
MPI_RECV_INIT
MPI_REDUCE
MPI_REDUCE_LOCAL
MPI_REDUCE_SCATTER
MPI_REDUCE_SCATTER_BLOCK
MPI_REGISTER_DATAREP
MPI_REQUEST_C2F
MPI_REQUEST_F2C
MPI_REQUEST_FREE
MPI_REQUEST_GET_STATUS
MPI_RGET
MPI_RGET_ACCUMULATE
MPI_RPUT
MPI_RSEND
MPI_RSEND_INIT
MPI_SCAN
MPI_SCATTER
MPI_SCATTERV
MPI_SEND
MPI_SEND_INIT
MPI_SENDRECV
MPI_SENDRECV_REPLACE
MPI_SIZEOF
MPI_SSEND
MPI_SSEND_INIT
MPI_START
MPI_STARTALL
MPI_STATUS_C2F
MPI_STATUS_C2F08
MPI_STATUS_F082C
MPI_STATUS_F082F
MPI_STATUS_F2C
MPI_STATUS_F2F08
MPI_STATUS_SET_CANCELLED
MPI_STATUS_SET_ELEMENTS
MPI_STATUS_SET_ELEMENTS_X
MPI_T_CATEGORY_CHANGED
MPI_T_CATEGORY_GET_CATEGORIES
MPI_T_CATEGORY_GET_CVARS
MPI_T_CATEGORY_GET_INFO
MPI_T_CATEGORY_GET_NUM
MPI_T_CATEGORY_GET_PVARS
MPI_T_CVAR_GET_INFO
MPI_T_CVAR_GET_NUM
MPI_T_CVAR_HANDLE_ALLOC
MPI_T_CVAR_HANDLE_FREE
MPI_T_CVAR_READ
MPI_T_CVAR_WRITE
MPI_T_ENUM_GET_INFO
MPI_T_ENUM_GET_ITEM
MPI_T_FINALIZE
MPI_T_INIT_THREAD
MPI_T_PVAR_GET_INFO
MPI_T_PVAR_GET_NUM
MPI_T_PVAR_HANDLE_ALLOC
MPI_T_PVAR_HANDLE_FREE
MPI_T_PVAR_READ
MPI_T_PVAR_READRESET
MPI_T_PVAR_RESET
MPI_T_PVAR_SESSION_CREATE
MPI_T_PVAR_SESSION_FREE
MPI_T_PVAR_START
MPI_T_PVAR_STOP
MPI_T_PVAR_WRITE
MPI_TEST
MPI_TEST_CANCELLED
MPI_TESTALL
MPI_TESTANY
MPI_TESTSOME
MPI_TOPO_TEST
MPI_TYPE_C2F
MPI_TYPE_COMMIT
MPI_TYPE_CONTIGUOUS
MPI_TYPE_CREATE_DARRAY
MPI_TYPE_CREATE_F90_COMPLEX
MPI_TYPE_CREATE_F90_INTEGER
MPI_TYPE_CREATE_F90_REAL
MPI_TYPE_CREATE_HINDEXED
MPI_TYPE_CREATE_HINDEXED_BLOCK
MPI_TYPE_CREATE_HVECTOR
MPI_TYPE_CREATE_INDEXED_BLOCK
MPI_TYPE_CREATE_KEYVAL
MPI_TYPE_CREATE_RESIZED
MPI_TYPE_CREATE_STRUCT
MPI_TYPE_CREATE_SUBARRAY
MPI_TYPE_DELETE_ATTR
MPI_TYPE_DUP
MPI_TYPE_DUP_FN
MPI_TYPE_EXTENT
MPI_TYPE_F2C
MPI_TYPE_FREE
MPI_TYPE_FREE_KEYVAL
MPI_TYPE_GET_ATTR
MPI_TYPE_GET_CONTENTS
MPI_TYPE_GET_ENVELOPE
MPI_TYPE_GET_EXTENT
MPI_TYPE_GET_EXTENT_X
MPI_TYPE_GET_NAME
MPI_TYPE_GET_TRUE_EXTENT
MPI_TYPE_GET_TRUE_EXTENT_X
MPI_TYPE_HINDEXED
MPI_TYPE_HVECTOR
MPI_TYPE_INDEXED
MPI_TYPE_LB
MPI_TYPE_MATCH_SIZE
MPI_TYPE_NULL_COPY_FN
MPI_TYPE_NULL_DELETE_FN
MPI_TYPE_SET_ATTR
MPI_TYPE_SET_NAME
MPI_TYPE_SIZE
MPI_TYPE_SIZE_X
MPI_TYPE_STRUCT
MPI_TYPE_UB
MPI_TYPE_VECTOR
MPI_UNPACK
MPI_UNPACK_EXTERNAL
MPI_UNPUBLISH_NAME
MPI_WAIT
MPI_WAITALL
MPI_WAITANY
MPI_WAITSOME
MPI_WIN_ALLOC
MPI_WIN_ALLOCATE
MPI_WIN_ALLOCATE_CPTR
MPI_WIN_ALLOCATE_SHARED
MPI_WIN_ALLOCATE_SHARED_CPTR
MPI_WIN_ATTACH
MPI_WIN_C2F
MPI_WIN_CALL_ERRHANDLER
MPI_WIN_COMPLETE
MPI_WIN_CREATE
MPI_WIN_CREATE_DYNAMIC
MPI_WIN_CREATE_ERRHANDLER
MPI_WIN_CREATE_KEYVAL
MPI_WIN_DELETE_ATTR
MPI_WIN_DETACH
MPI_WIN_DUP_FN
MPI_WIN_F2C
MPI_WIN_FENCE
MPI_WIN_FLUSH
MPI_WIN_FLUSH_ALL
MPI_WIN_FLUSH_LOCAL
MPI_WIN_FLUSH_LOCAL_ALL
MPI_WIN_FREE
MPI_WIN_FREE_KEYVAL
MPI_WIN_GET_ATTR
MPI_WIN_GET_ERRHANDLER
MPI_WIN_GET_GROUP
MPI_WIN_GET_INFO
MPI_WIN_GET_NAME
MPI_WIN_LOCK
MPI_WIN_LOCK_ALL
MPI_WIN_NULL_COPY_FN
MPI_WIN_NULL_DELETE_FN
MPI_WIN_POST
MPI_WIN_SET_ATTR
MPI_WIN_SET_ERRHANDLER
MPI_WIN_SET_INFO
MPI_WIN_SET_NAME
MPI_WIN_SHARED_ALLOCATE
MPI_WIN_SHARED_QUERY
MPI_WIN_SHARED_QUERY_CPTR
MPI_WIN_START
MPI_WIN_SYNC
MPI_WIN_TEST
MPI_WIN_UNLOCK
MPI_WIN_UNLOCK_ALL
MPI_WIN_WAIT
MPI_WTICK
MPI_WTIME

Notable Additions to MPI 3.0

- Non-blocking collectives
- Neighborhood collectives
- RMA enhancements
- Shared memory support
- MPI Tool Information Interface
- Non-collective communicator creation
- Fortran 2008 Bindings
- New Datatypes
- Large data counts
- Matched probe

MPI-3 Impl. as of Novemer 2014 (thanks to Pavan Balaji)

	MPICH	MVAPICH	Open MPI	Cray MPI	Tianhe MPI	Intel MPI	IBM BG/Q MPI [1]	IBM PE MPICH [2]	IBM Platform	SGI MPI	Fujits u MPI	MS MPI
NB collectives	✔	✔	✔	✔	✔	✔	✔	Q4 '14	✔	✔	✔	
Neighborhood collectives	✔	✔	✔	✔	✔	✔	✔	Q4 '14	Q3 '15	✔	Q2 '15	
RMA	✔	✔	✔	✔	✔	✔	✔	Q4 '14	Q3 '15	✔	Q2 '15	
Shared memory	✔	✔	✔	✔	✔	✔	✔	Q4 '14	Q3 '15	✔	Q2 '15	✔
Tools Interface	✔	✔	✔	✔	✔	✔	✔ [3]	Q4 '14	Q3 '15	✔	Q2 '15	*
Non-collective comm. create	✔	✔	✔	✔	✔	✔	✔	Q4 '14	Q3 '15	✔	Q2 '15	
F08 Bindings	✔	✔	✔	Q4 '14	✔	Q4 '14	✔	Q4 '14	Q3 '15	✔	Q2 '15	
New Datatypes	✔	✔	✔	✔	✔	✔	✔	Q4 '14	Q3 '15	✔	Q2 '15	*
Large Counts	✔	✔	✔	✔	✔	✔	✔	Q4 '14	Q3 '15	✔	Q2 '15	*
Matched Probe	✔	✔	✔	✔	✔	✔	✔	Q4 '14	Q3 '15	✔	✔	*

Release dates are estimates and are subject to change at any time.

Empty cells indicate no *publicly announced* plan to implement/support that feature.

[1] Open source, but unsupported [2] Beta release [3] No MPI_T variables exposed * Under development

(*) Platform-specific restrictions might apply for all supported features

Where We Are

- **MPI 3.0 ratified in September 2012**
  - Available at http://www.mpi-forum.org/
  - 852 pages, 430 functions
  - Adaption in most MPIs progressing fast
- **Working towards MPI 3.1**
  - Inclusion for errata (mainly RMA, Fortran, MPI_T)
  - Minor updates and additions (address arithmetic and non-block. I/O)
  - Currently planned for March 2015

Where We Are

- **MPI 3.0 ratified in September 2012**
  - Available at http://www.mpi-forum.org/
  - 852 pages, 430 functions
  - Adaption in most MPIs progressing fast
- **Working towards MPI 3.1**
  - Inclusion for errata (mainly RMA, Fortran, MPI_T)
  - Minor updates and additions (address arithmetic and non-block. I/O)
  - Currently planned for March 2015
- **Concurrently discussions for MPI 4.0**
  - Major additions as gating items
  - Schedule tbd. (depends on features)

Larger Topics under Discussion

- **Fault Tolerance support in MPI**
  - Avoid job failure when one node fails
  - Suitable fault detection mechanism
  - Ability to reason about state of MPI and continue execution
- **Support for Hybrid Programming models**
  - Improved thread support
  - Easier integration with MPI+X approaches
  - Extensions to be able to treat threads as MPI endpoints/processes
- **New point to point mechanisms like Streams/Channels**
  - Dedicated point to point connections
  - Continued streams of data without the need for individually matching send/recv pairs

Larger Topics under Discussion

- **Fault Tolerance support in MPI**
  - Part of the Fault Tolerance WG
  - Presentation by Wesley Bland, ANL
- **Support for Hybrid Programming models**
  - Part of the Hybrid WG
  - Presentation by Pavan Balaji, ANL
- **New point to point mechanisms like Streams/Channels**
  - Part of the Point to Point WG
  - Presentation by Daniel Holmes, EPCC