

Software Standards at the Example of the Message Passing Interface (MPI)

Martin Schulz

Lawrence Livermore National Laboratory
Chair of the MPI Forum

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

4th ENES HPC Workshop, April 7th, 2016



LLNL-PRES-688667

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

 Lawrence Livermore
National Laboratory

The Message Passing Interface (MPI)

- MPI is ~~an~~ **the** API specification for message passing
 - Targeted at distributed process environments
 - Available on basically any HPC platform
 - Vendor and platform agnostic – just an API specification
 - Bindings for C, C++ (for a while), a various Fortran variants
- Basic Features (MPI-1)
 - Operations for sending and receiving messages
 - Support for asynchronous messages
 - Collective operations (for optimization)
 - Communication based on communicators (collection of processes)
- Later additions (MPI-2/3)
 - Support for I/O
 - One-sided communication (RMA)
 - Dynamic processes (although, without much uptake)
 - Non blocking and neighborhood collectives

What it is Not / General Misconceptions

- MPI \neq SPMD
 - Most programs are written in SPMD or BSP style
 - BUT: MPI supports any style that requires message exchanges
- MPI \neq ABI
 - MPI only specifies an API, but, e.g., not the types used
 - WHY? Enables widely varying implementations
- MPI \neq Fixed environment
 - The MPI standard explicitly omits standardization of the environment
 - Example: “mpirun” is just a suggestion
 - WHY? Enables portability
- MPI \neq Fixed forever (and hence dead for exascale)
 - MPI is a living standard that can be modified
 - BUT: additions must be based on “Best Practices”

Participate and help us evolve MPI

MPI: How It All Began

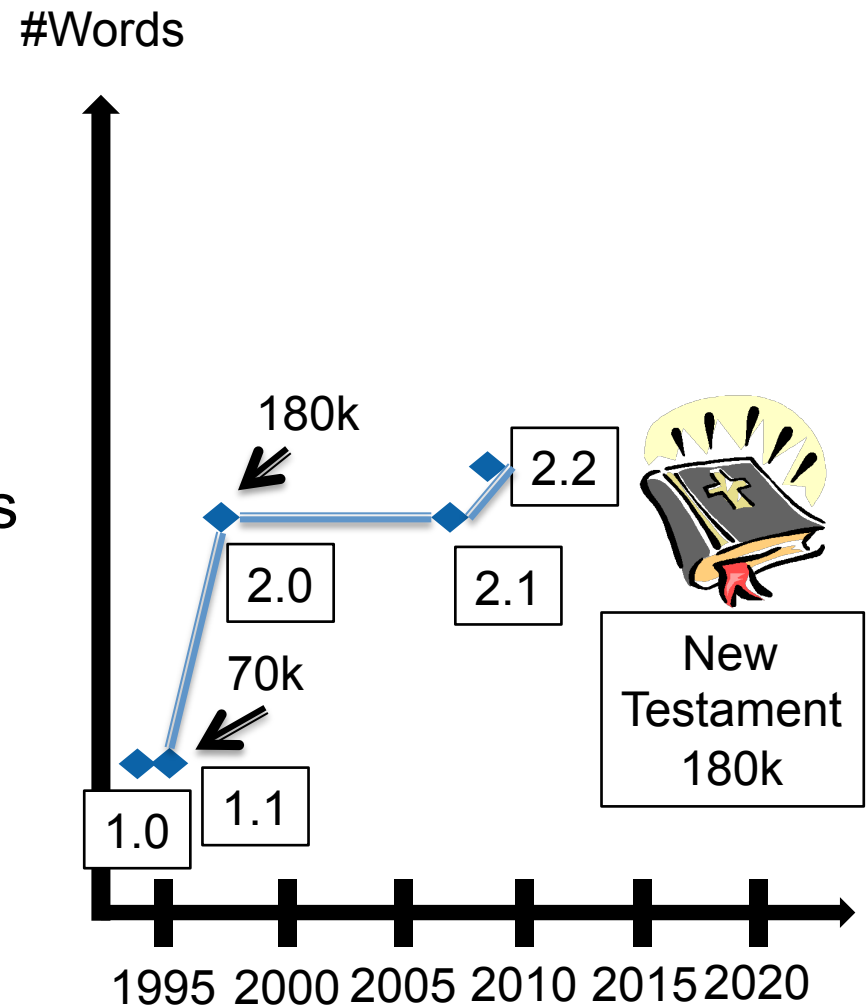
- Multiple message passing APIs in the late 80's
 - Increasing importance of distributed memory machines
 - Parallel Virtual Machine (PVM), LAM, P4
 - Wide range of vendor APIs
- Workshop on Standards for Message Passing in a Distributed Memory Environment
 - Held on April 29–30, 1992 in Williamsburg, Virginia
 - Led to very early strawman draft
 - Follow-on organization at Supercomputing 1992
- Group of interested people got together after that
 - Meeting in person every 6 weeks for 9 months
 - Comment draft by Supercomputing 1993
 - First version ratified by May 1994

The Growth of the MPI Standard

- MPI 1.0 May 1994 – 228 pages
- MPI 1.1 Nov 1995 – 238 pages (128 functions)
- MPI 2.0 Nov 1997 – 608 pages

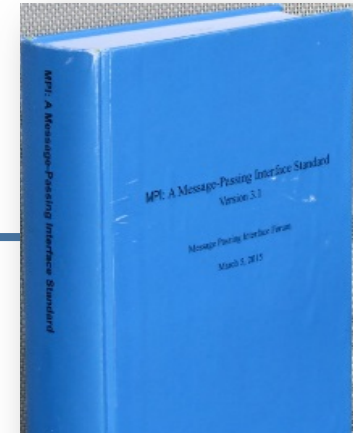
10 year break

- MPI 2.1 June 2008 – 608 pages
 - Merged document
- MPI 2.2 Sep 2009 – 647 pages
 - Small, conforming additions
- In parallel, start on MPI 3.0
 - Major new concepts
 - Mainly driven by working groups



The Latest Versions

- MPI 3.0 ratified in September 2012
 - Available at <http://www.mpi-forum.org/>
 - Several major additions compared to MPI 2.2
- MPI 3.1 ratified in June 2015
 - Inclusion for errata (mainly RMA, Fortran, MPI_T)
 - Minor updates and additions (address arithmetic and non-block. I/O)
 - Adaption in most MPIs progressing fast



Available through HLRS
-> MPI Forum Website



Major Additions to MPI 3

- 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
- Non-blocking I/O

MPI-3.1 Impl. as of November 2015 (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	Fujitsu MPI	MS MPI	MPC
NBC	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	(*)	Q4'15
Nbrhood collectives	✓	✓	✓	✓	✓	✓	✓	✓		✓			Q4'15
RMA	✓	✓	✓	✓	✓	✓	✓	✓		✓			*
Shared memory	✓	✓	✓	✓	✓	✓	✓	✓		✓		✓	*
Tools Interface	✓	✓	✓	✓	✓	✓	✓	✓		✓		*	Q4'16
Comm-creat group	✓	✓	✓	✓	✓	✓	✓	✓		✓			*
F08 Bindings	✓	✓	✓	✓	✓		✓			✓			Q2'16
New Datatypes	✓	✓	✓	✓	✓	✓	✓	✓		✓		✓	Q4'15
Large Counts	✓	✓	✓	✓	✓	✓	✓	✓		✓		✓	Q2'16
Matched Probe	✓	✓	✓	✓	✓	✓	✓	✓		✓	✓	✓	Q2'16
NBC I/O	✓	Q1'16	✓	Q4'15						Q2'16			

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

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

Platform-specific restrictions might apply for all supported features

¹ Open Source but unsupported

² No MPI_T variables exposed

* Under development

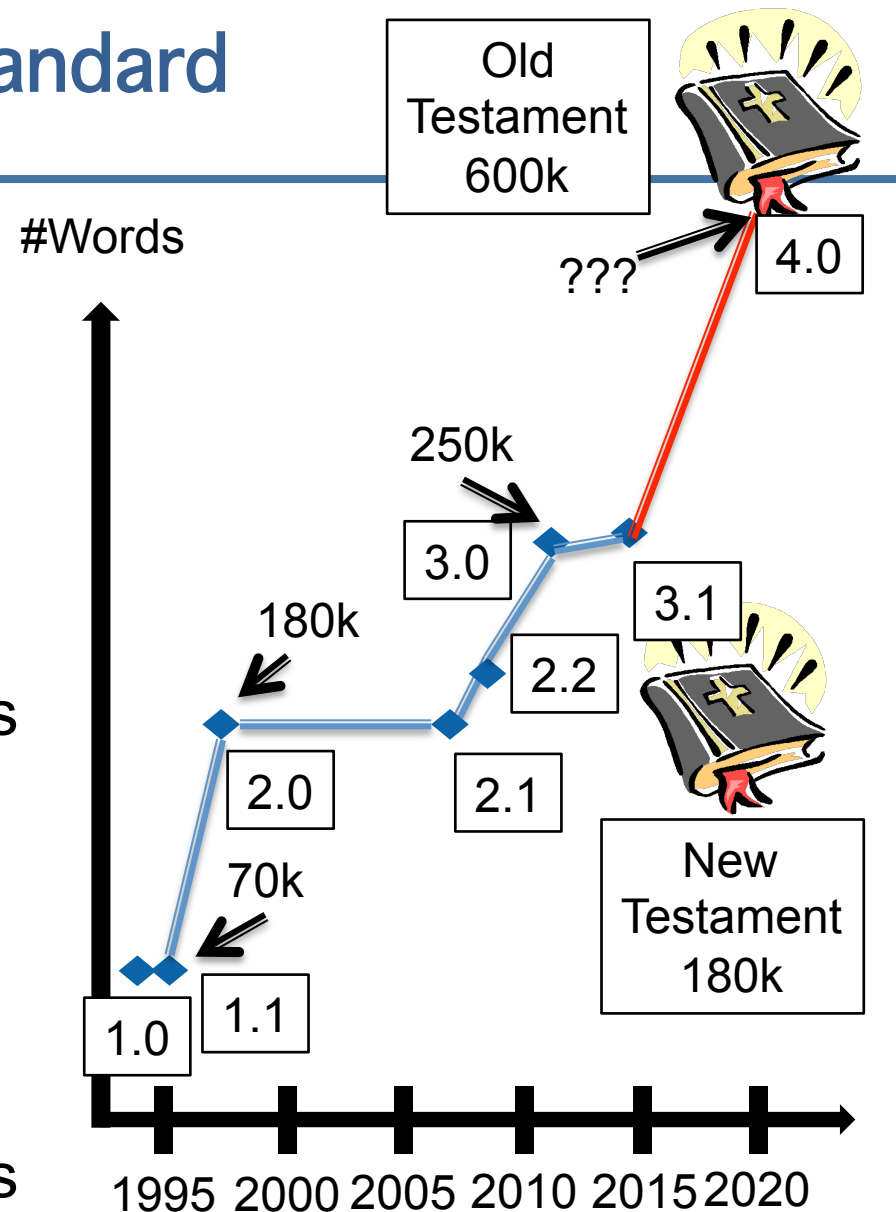
(*) Partly done

The Growth of the MPI Standard

- MPI 1.0 May 1994 – 228 pages
- MPI 1.1 Nov 1995 – 238 pages (128 functions)
- MPI 2.0 Nov 1997 – 608 pages

10 year break

- MPI 2.1 June 2008 – 608 pages
 - Merged document
- MPI 2.2 Sep 2009 – 647 pages
 - Small, conforming additions
- MPI 3.0 Sep 2012 – 852 pages
- MPI 3.1 June 2014 – 868 pages



Towards MPI 4.0

- Deliberate decision to continue after MPI 3.0
 - Major points not included in MPI 3
 - Continued interest in participation
 - Adjustment of meeting schedule
- Several major initiatives
 - Fault tolerance
 - Support for hybrid models
 - MPI Sessions to enable isolation
 - Hints and asserts
 - Persistent collectives
 - Advances in tool interfaces
- Tradeoffs
 - Useful extensions vs. unlimited growth
 - General standard vs. additions for particular use cases

The MPI Forum Drives MPI

- Standardization body for MPI
 - Group of volunteers paid by their home organizations
 - Intended to be a good mix
 - Implementors, users and user communities, HPC centers (procurements)

MPI Forum Participants (list for MPI 3.0)

William Gropp
Richard Graham
Torsten Hoefler
George Bosilca
David Solt
Bronis R. De Supinski
Rajeev Thakur
Darius Buntinas
Jeffrey M. Squyres
Rolf Rabenseifner
Tatsuya Abe
Tomoya Adachi
Sadaf Alam
Reinhold Bader
Pavan Balaji
Purushotham V.
Bangalore
Brian Barrett
Richard Barrett
Robert Blackmore
Aurelien Bouteiller
Ron Brightwell
Greg Bronevetsky
Jed Brown
Darius Buntinas
Devendar Bureddy
Arno Candel
George Carr
Mohamad Charawi

Raghunath Raja
Chandrasekar
James Dinan
Terry Dontje
Edgar Gabriel
Balazs Gero
Brice Goglin
David Goodell
Manjunath Gorentla
Erez Haba
Je Hammond
Thomas Herault
Marc-Andre Hermanns
Jennifer Herrett-
Skjellum
Nathan Hjelm
Atsushi Hori
Joshua Hursey
Marty Itzkowitz
Yutaka Ishikawa
Nysal Jan
Bin Jia
Hideyuki Jitsumoto
Yann Kalemkarian
Krishna Kandalla
Takahiro Kawashima
Chulho Kim
Dries Kimpe
Christof Klausecker

Alice Koniges
Quincey Koziol
Dieter Kranzmueller
Manojkumar Krishnan
Sameer Kumar
Eric Lantz
Jay Lofstead
Bill Long
Andrew Lumsdaine
Miao Luo
Ewing Lusk
Adam Moody
Nick M. Maclaren
Amith Mamidala
Guillaume Mercier
Scott McMillan
Douglas Miller
Kathryn Mohror
Tim Murray
Tomotake Nakamura
Takeshi Nanri
Steve Oyanagi
Mark Pagel
Swann Perarnau
Sreeram Potluri
Howard Pritchard
Rolf Riesen
Hubert Ritzdorf
Kuninobu Sasaki

Timo Schneider
Martin Schulz
Gilad Shainer
Christian Siebert
Anthony Skjellum
Brian Smith
Marc Snir
Raele Giuseppe
Solca Shinji
Sumimoto Alexander
Supalov
Sayantan Sur
Masamichi Takagi
Fabian Tillier
Vinod Tipparaju
Jesper Larsson Traff
Richard Treumann
Keith Underwood
Rolf Vandevaart
Anh Vo
Abhinav Vishnu
Min Xie
Enqiang Zhou

The MPI Forum Drives MPI

- Standardization body for MPI
 - Group of volunteers paid by their home organizations
 - Intended to be a good mix
 - Implementors, users and user communities, HPC centers (procurements)
- Organization consists of chair, secretary, convener, steering committee, and member organizations
 - Plus working groups and chapter committees
- Tasks
 - Discusses additions and new directions
 - Oversees the correctness and quality of the standard
 - Enforces the “Best Practice” rules
 - Ensures “minimality”

 - Represents MPI to the community
 - Holds events like BoFs at SC or ISC

Rules of the MPI Forum

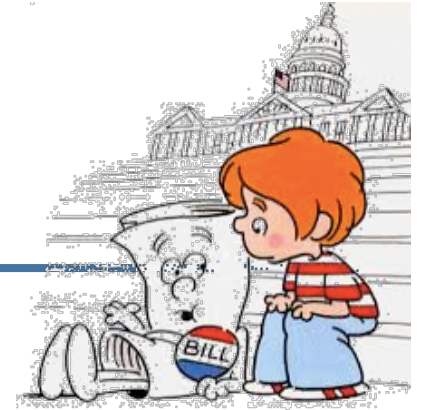
- Evolved over the years
 - Free form in the beginning
 - Hear-say rules with some notes in a side document
 - Since MPI 3: evolving formal rule document
- MPI rules are public and available on the website
 - Mainly deals with memberships and voting procedures
 - Published under open source license and intended to be reused
- Open membership
 - Any organization is welcome to participate
 - Consists of working groups and the actual MPI forum
 - Physical meetings 4 times each year (3 US, one at EuroMPI/Asia)
 - Working groups meet between forum meetings (via phone)
 - Plenary/full forum work is done mostly at the physical meetings
 - Voting rights depend on attendance
 - An organization has to be present two out of the last three meetings (incl. the current one) to be eligible to vote

Differences to Other Standards

- MPI is not backed by a formal legal entity
 - No organization like ISO or IEEE
 - Provides more flexibility and allows quicker turn-around
 - Market too small
 - Copyrights held in good faith by universities
 - Makes financial transaction hard
 - No “official” membership or membership dues
- Closest comparison: OpenMP
 - Closed (for pay) membership
 - Legal entity governed by the member organizations
 - Pro:
 - Easier rules for defining members, votes, ...
 - Allows for procedures like phone votes
 - Cons:
 - More heavy weight, legal costs, ...
 - Only possible with much stronger industry participation

Typical Way to Add Items to MPI

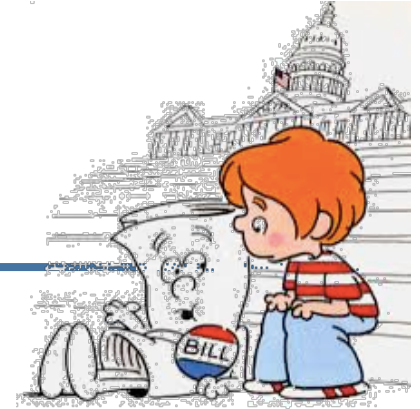
1. New items should be brought to a matching working group for discussion
 - Creation of preliminary proposal
 - Simple (grammar) changes are handled by chapter committees



Current Working Groups

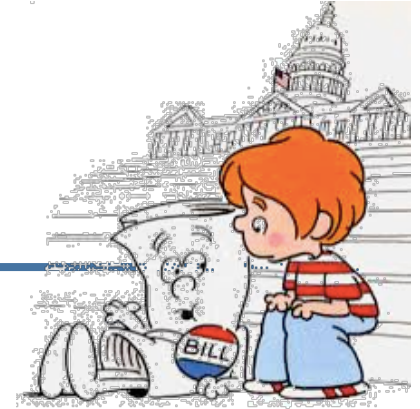
- **Collectives & Topologies**
 - Torsten Hoefler, ETH
 - Andrew Lumsdaine, Indiana
- **Fault Tolerance**
 - Wesley Bland, ANL
 - Aurelien Bouteiller, UTK
 - Rich Graham, Mellanox
- **Fortran**
 - Craig Rasmussen, U. of Oregon
- **Generalized Requests**
 - Fab Tillier, Microsoft
- **Hybrid Models**
 - Pavan Balaji, ANL
- **I/O**
 - Quincey Koziol, HDF Group
 - Mohamad Charawi, HDF Group
- **Large count**
 - Jeff Hammond, Intel
- **Persistence**
 - Anthony Skjellum, Auburn Uni.
- **Point to Point Comm.**
 - Dan Holmes, EPCC
 - Rich Graham, Mellanox
- **Remote Memory Access**
 - Bill Gropp, UIUC
 - Rajeev Thakur, ANL
- **Sessions**
 - Jeff Squyres
- **Tools**
 - Kathryn Mohror, LLNL
 - Marc-Andre Hermans, RWTH Aachen

Typical Way to Add Items to MPI



1. New items should be brought to a matching working group for discussion
 - Creation of preliminary proposal
 - Simple (grammar) changes are handled by chapter committees
2. Socializing of idea driven by the WG
 - Could include plenary presentation to gather feedback
 - Focused on concepts not details like names or formal text
 - Make proposal easily available through WG wiki
 - Important to keep overall standard in mind
3. Development of full proposal
 - Latex version that fits into the standard
 - Creation of a github issue to track voting and a matching pull request
4. MPI forum reading/voting process

The MPI Voting Process



- Quorum
 - 2/3 of eligible organizations have to be present
 - 3/4 of present organization have to vote yes
 - Goal: standardize only if there is consensus

- Steps
 1. Reading: “Word by word” presentation to the forum
 2. First vote
 3. Second vote

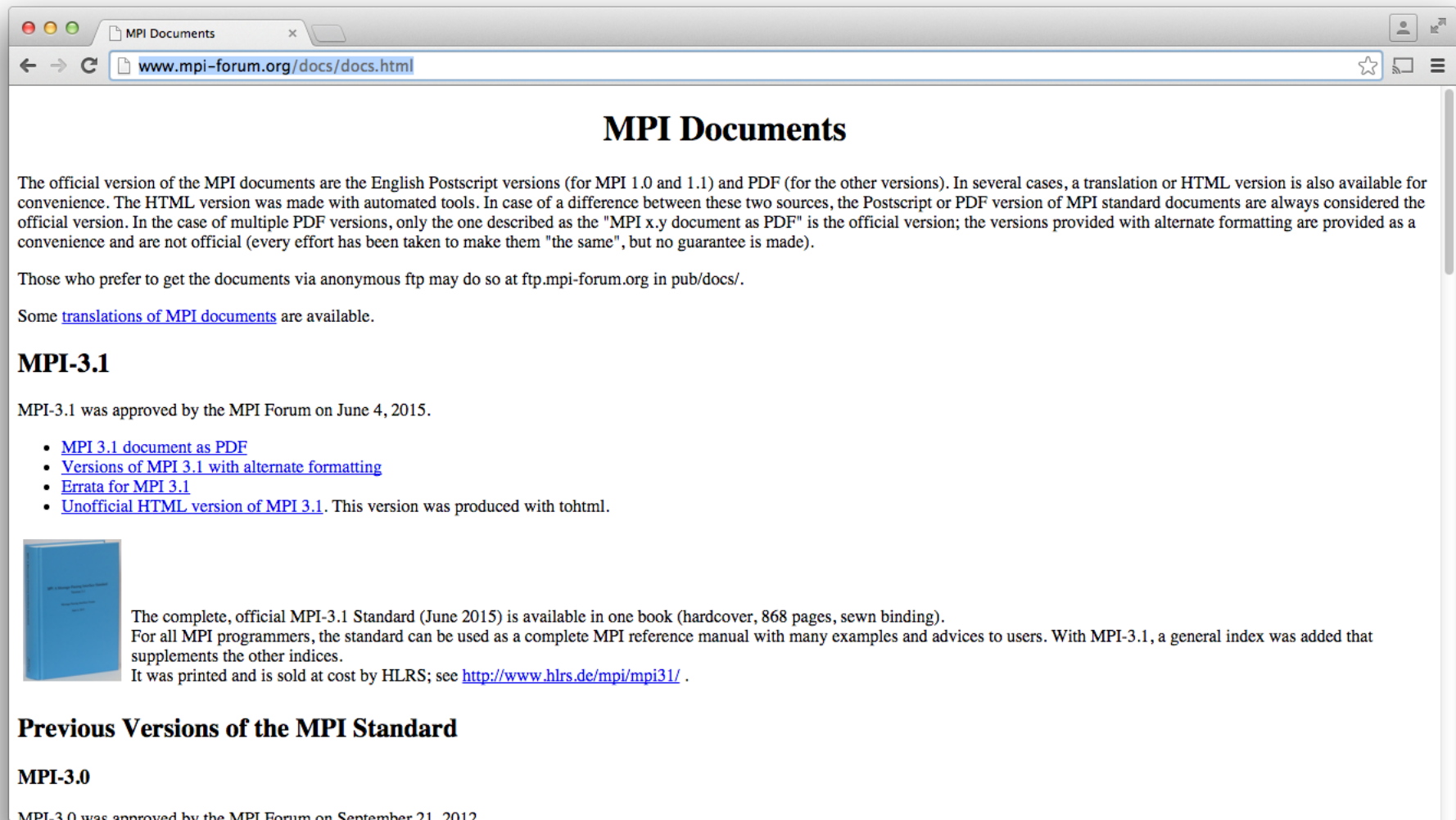
- Each step has to be at a separate physical meeting
 - Ensure people have time to think about additions
 - Avoid hasty mistakes, which are hard to fix
 - Prototypes are encouraged and helpful to convince people

- Separate (extended) rules for document votes

How To Get Involved

- Submit comments to the MPI forum
 - mpi-comments@mpi-forum.org
 - Feedback on prototypes / proposals as well as the existing standard
- Subscribe to email lists to see what's going on
 - Each working group has its own mailing list
- Join a working group
 - Check out the respective Wiki pages
 - Participate in WG meetings (typically phone conference)
 - Contact the WG chairs to introduce yourself
- Participate in physical MPI forum meetings
 - June 2016, Bellevue, WA, USA
 - September 2016, Edinburgh, UK
 - Logistics and agendas available through the MPI forum website
 - Drop me an email if you have questions or are interested
- More information at: <http://www.mpi-forum.org/>

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



The screenshot shows a web browser window with the address bar containing www.mpi-forum.org/docs/docs.html. The page title is "MPI Documents". The main content includes a paragraph explaining the official versions of MPI documents, a link to translations, and a section for MPI-3.1. The MPI-3.1 section mentions its approval date and provides links to PDF, alternate formatting, errata, and an unofficial HTML version. There is also a section for the MPI-3.1 standard book, including an image of the book cover and text about its availability and purchase. The page concludes with a section for previous versions of the MPI standard, specifically MPI-3.0.

MPI Documents

The official version of the MPI documents are the English Postscript versions (for MPI 1.0 and 1.1) and PDF (for the other versions). In several cases, a translation or HTML version is also available for convenience. The HTML version was made with automated tools. In case of a difference between these two sources, the Postscript or PDF version of MPI standard documents are always considered the official version. In the case of multiple PDF versions, only the one described as the "MPI x.y document as PDF" is the official version; the versions provided with alternate formatting are provided as a convenience and are not official (every effort has been taken to make them "the same", but no guarantee is made).


Those who prefer to get the documents via anonymous ftp may do so at [ftp.mpi-forum.org](ftp://ftp.mpi-forum.org/pub/docs/) in `pub/docs/`.

Some [translations of MPI documents](#) are available.

MPI-3.1

MPI-3.1 was approved by the MPI Forum on June 4, 2015.

- [MPI 3.1 document as PDF](#)
- [Versions of MPI 3.1 with alternate formatting](#)
- [Errata for MPI 3.1](#)
- [Unofficial HTML version of MPI 3.1](#). This version was produced with tohtml.



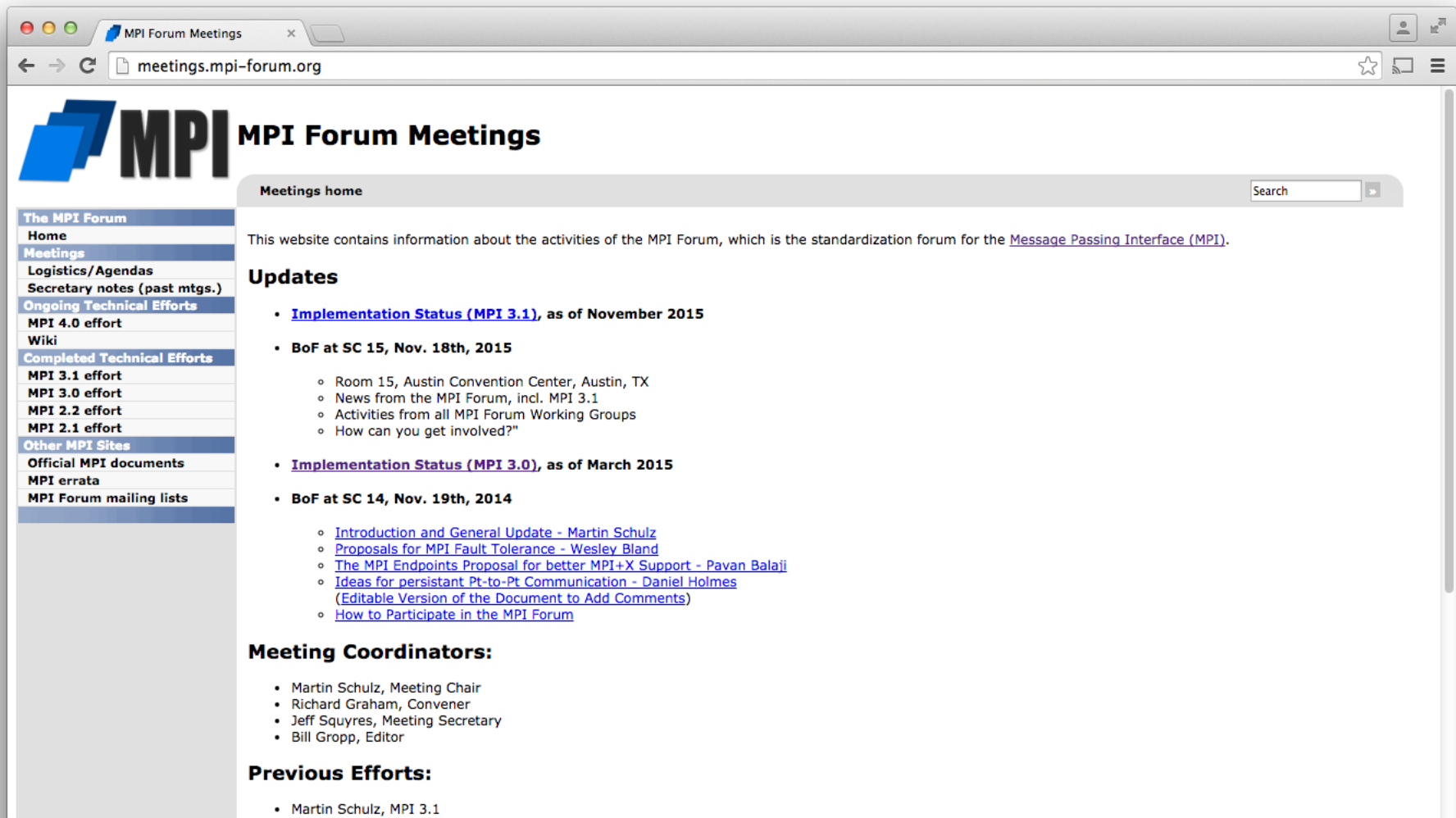
The complete, official MPI-3.1 Standard (June 2015) is available in one book (hardcover, 868 pages, sewn binding). For all MPI programmers, the standard can be used as a complete MPI reference manual with many examples and advices to users. With MPI-3.1, a general index was added that supplements the other indices. It was printed and is sold at cost by HLRS; see <http://www.hlrs.de/mpi/mpi31/>.

Previous Versions of the MPI Standard

MPI-3.0

MPI-3.0 was approved by the MPI Forum on September 21, 2012.

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



The screenshot shows a web browser window with the URL `meetings.mpi-forum.org`. The page features the MPI Forum Meetings logo and a navigation menu on the left. The main content area includes a search bar, a description of the website, and several sections: Updates, Meeting Coordinators, and Previous Efforts.

Meetings home

This website contains information about the activities of the MPI Forum, which is the standardization forum for the [Message Passing Interface \(MPI\)](#).

Updates

- [Implementation Status \(MPI 3.1\), as of November 2015](#)
- **BoF at SC 15, Nov. 18th, 2015**
 - Room 15, Austin Convention Center, Austin, TX
 - News from the MPI Forum, incl. MPI 3.1
 - Activities from all MPI Forum Working Groups
 - How can you get involved?"
- [Implementation Status \(MPI 3.0\), as of March 2015](#)
- **BoF at SC 14, Nov. 19th, 2014**
 - [Introduction and General Update - Martin Schulz](#)
 - [Proposals for MPI Fault Tolerance - Wesley Bland](#)
 - [The MPI Endpoints Proposal for better MPI+X Support - Pavan Balaji](#)
 - [Ideas for persistent Pt-to-Pt Communication - Daniel Holmes](#)
([Editable Version of the Document to Add Comments](#))
 - [How to Participate in the MPI Forum](#)

Meeting Coordinators:

- Martin Schulz, Meeting Chair
- Richard Graham, Convener
- Jeff Squyres, Meeting Secretary
- Bill Gropp, Editor

Previous Efforts:

- Martin Schulz, MPI 3.1

GitHub Organization

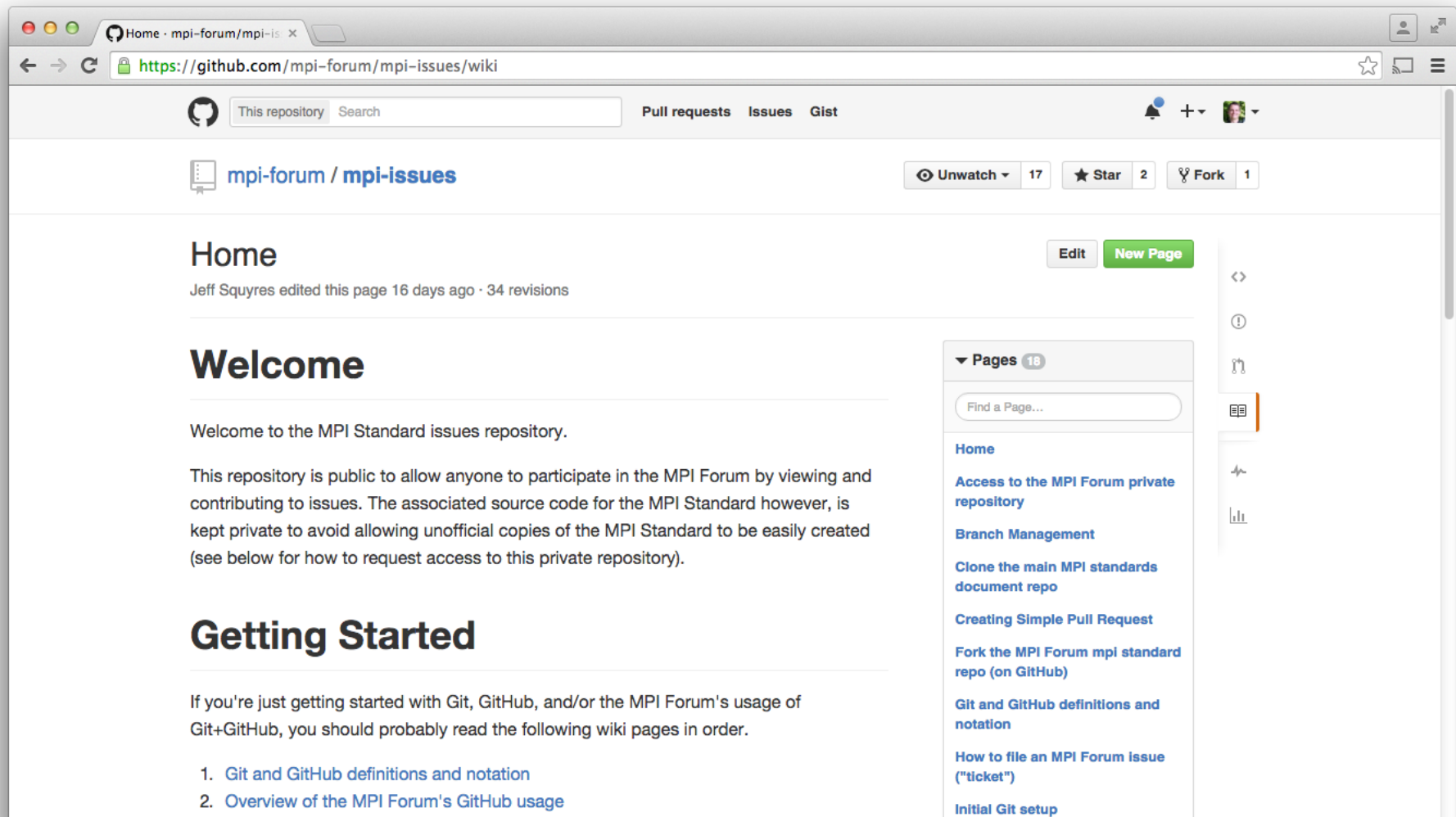
<https://github.com/mpi-forum>

The screenshot shows the GitHub organization page for MPI Forum. The browser address bar displays <https://github.com/mpi-forum>. The page header includes the GitHub logo, a search bar, and navigation links for Pull requests, Issues, and Gist. The main content area features the MPI Forum logo and the text: "Home of the MPI Forum's Repositories. See mpi-issues for general issues and wiki pages." Below this, there are links for the website (<http://www.mpi-forum.org>) and email (mpi-forum@lists.mpi-forum.org). The page is divided into sections: "Repositories" (with a search bar and "Filters" dropdown), "People" (16 members), and "Teams" (4 teams). The "Repositories" section lists three repositories:

- meetings.mpi-forum.org** (PHP, 0 stars, 0 forks) - The meetings.mpi-forum.org web site, updated 4 days ago.
- mpi-standard** (PRIVATE, TeX, 0 stars, 13 forks) - Official repository for the MPI Standard source, updated 11 days ago.
- mpir** (PRIVATE, TeX, 0 stars, 4 forks) - Repository for the MPIR Interface.

The "People" section displays a grid of 16 member avatars.

https://github.com/mpi-forum/mpi-issues/wiki



Home · mpi-forum/mpi-iss x

https://github.com/mpi-forum/mpi-issues/wiki

This repository Search Pull requests Issues Gist

mpi-forum / mpi-issues Unwatch 17 Star 2 Fork 1

Home

Jeff Squyres edited this page 16 days ago · 34 revisions

Edit New Page

Welcome

Welcome to the MPI Standard issues repository.

This repository is public to allow anyone to participate in the MPI Forum by viewing and contributing to issues. The associated source code for the MPI Standard however, is kept private to avoid allowing unofficial copies of the MPI Standard to be easily created (see below for how to request access to this private repository).

Getting Started

If you're just getting started with Git, GitHub, and/or the MPI Forum's usage of Git+GitHub, you should probably read the following wiki pages in order.

1. [Git and GitHub definitions and notation](#)
2. [Overview of the MPI Forum's GitHub usage](#)

Pages 18

Find a Page...

- Home
- [Access to the MPI Forum private repository](#)
- [Branch Management](#)
- [Clone the main MPI standards document repo](#)
- [Creating Simple Pull Request](#)
- [Fork the MPI Forum mpi standard repo \(on GitHub\)](#)
- [Git and GitHub definitions and notation](#)
- [How to file an MPI Forum issue \("ticket"\)](#)
- [Initial Git setup](#)

Open Tickets/Issues

Search/filter for: “is:open is:issue user:mpi-forum”

The screenshot shows a web browser window displaying the GitHub Issues page for the repository 'mpi-forum/mpi-issues'. The search bar at the top contains the query 'is:open is:issue user:mpi-forum'. Below the search bar, there are filters for 'Created', 'Assigned', and 'Mentioned'. The main content area shows a list of 17 open issues, with 5 closed issues. The issues are sorted by date, with the most recent at the top. Each issue entry includes the issue title, the repository name, the issue number, the date it was opened, the user who opened it, and the location. The issues are as follows:

Issue Title	Repository	Issue Number	Opened	User	Location	Labels	Comments
User-Level Failure Mitigation: Files	mpi-forum/mpi-issues	#22	7 days ago	wesbland	2016-02 Chicago, USA	mpi-4.0, not ready, wg-ft	0
User-Level Failure Mitigation: RMA	mpi-forum/mpi-issues	#21	7 days ago	wesbland	2016-02 Chicago, USA	mpi-4.0, not ready, wg-ft	0
User-Level Failure Mitigation	mpi-forum/mpi-issues	#20	7 days ago	wesbland	2016-02 Chicago, USA	mpi-4.0, not ready, wg-ft	1
make MPI_THREAD_MULTIPLE a requirement for all implementations	mpi-forum/mpi-issues	#19	9 days ago	jeffhammond			5
Longer types for use with MPI_MINLOC and MPI_MAXLOC	mpi-forum/mpi-issues	#18	9 days ago	jeffhammond		not ready	2
clarify what functions can be called on uncommitted datatypes	mpi-forum/mpi-issues	#17	9 days ago	jeffhammond		not ready	6
deprecate MPI_GRAPH_CREATE	mpi-forum/mpi-issues	#16	9 days ago	jeffhammond		not ready	0
MPI_Reduce_scatter advice to users is wrong	mpi-forum/mpi-issues	#15	9 days ago	jeffhammond		errata	3
deprecate MPI_COMM_JOIN	mpi-forum/mpi-issues	#14				mpi <next>, not ready	0



25-28 SEPTEMBER
EURO  **MPI**
EDINBURGH 2016



www.EuroMPI2016.ed.ac.uk

- Call for papers open by end of November 2015
- Full paper submission deadline: 1st May 2016
- Associated events: tutorials, workshops, training
- Focusing on: benchmarks, tools, applications, parallel I/O, fault tolerance, hybrid MPI+X, and alternatives to MPI and reasons for not using MPI

Summary: The MPI Forum in a Nutshell

- MPI Forum is an open forum
 - Everyone / every organization can join
 - Want/Need/Encourage community feedback
- MPI Forum rules are public
 - Available on the MPI forum web site
 - Under open source license (!)
 - If reused, please let us know and help us evolve them!
- Major work in the next few years on MPI 4
 - Several major initiatives, incl.
 - Fault Tolerance
 - Better support for hybrid programming
 - Performance Hints and Assertions
 - Many other proposals as well
- Get involved in MPI
 - Let us know what is needed for climate and weather codes
 - Provide a user perspective and a balance to MPI implementors
 - Help close these gaps!



www.mpi-forum.org