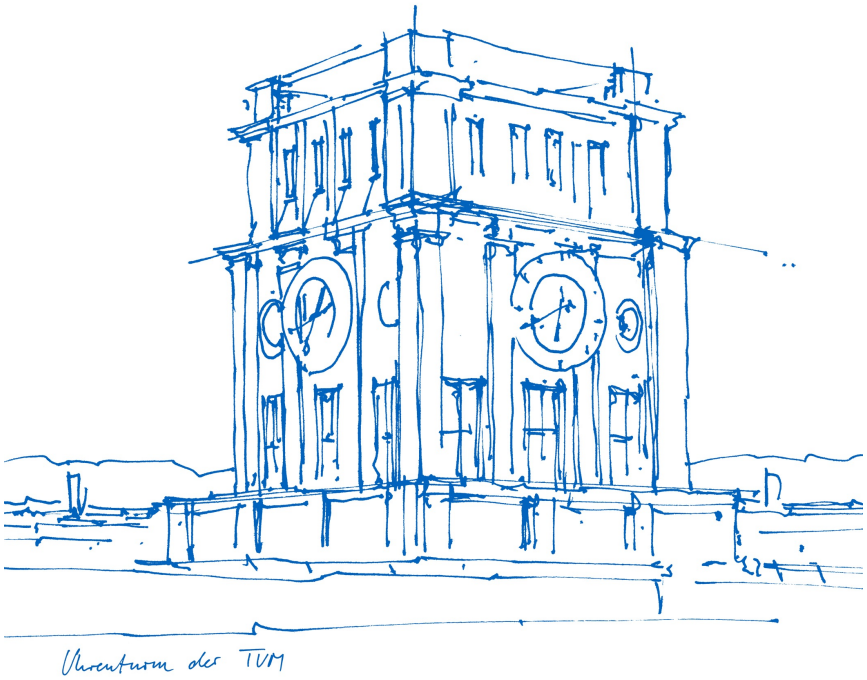
The State of MPI
Current Standard and Future Plans



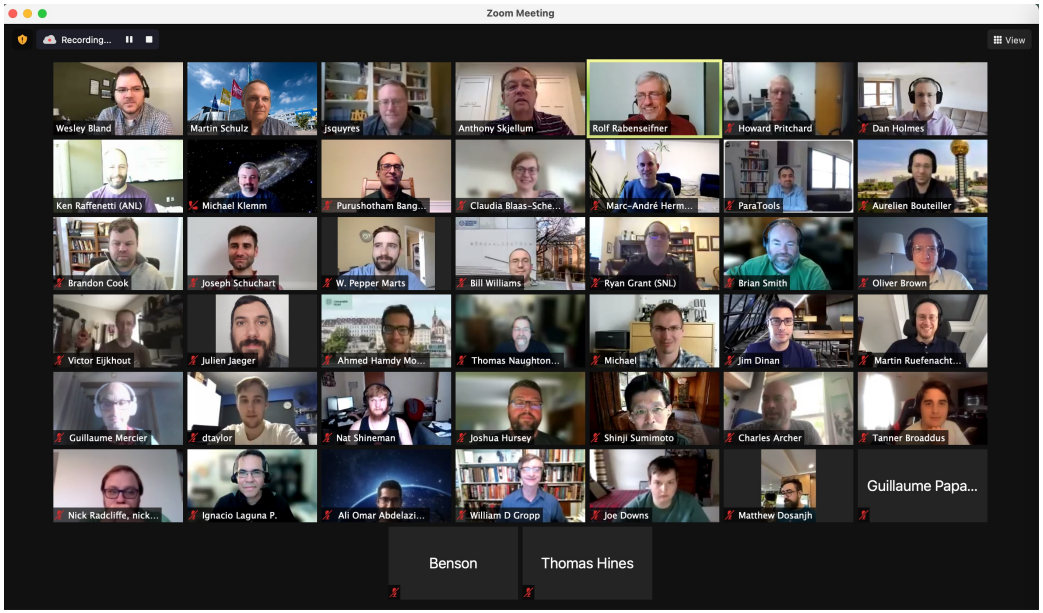
Martin Schulz
Chair for Computer Architecture and Parallel Systems
Technical University of Munich

Plasma-PEPSC Seminar

Where Are We?

MPI 4.0 was published 2.5 years ago (in the middle of Corona)

- Solution for “Big Count” operations
- Partitioned Communication
- Persistent Collectives
- Improved Error Handling
- Topology Solutions
- New init options via MPI Sessions
- And much more …



The “Embiggenment“

Big Count aka. Embiggenment

Problem: in previous interface “count” arguments are “int”

- Limits communication volumes to 32bit x Datatype
- Significant number of applications need more
- Initial datatype “trick” no longer sufficient

Solutions discussed included:

- Just changing “int” arguments to “MPI_Count” arguments → ☹ ☹ ☹
- Polymorphic bindings → ☹ ☹
- Duplication of interfaces: with int and with MPI_Count (“_c” suffix) → ☹

Last option was selected

- Update of the general type rules for bindings
- Verification of all bindings, which led to errata tickets
- Addition of many new routines with “_c”

Example: MPI_Recv

Language Independent Binding

MPI_RECV(buf, count, datatype, source, tag, comm, status)

OUT	buf	initial address of receive buffer (choice)
IN	count	number of elements in receive buffer (non-negative integer)
IN	datatype	datatype of each receive buffer element (handle)
IN	source	rank of source or MPI_ANY_SOURCE (integer)
IN	tag	message tag or MPI_ANY_TAG (integer)
IN	comm	communicator (handle)
OUT	status	status object (status)

C binding

```
int MPI_Recv(void *buf, int count, MPI_Datatype datatype, int source,
             int tag, MPI_Comm comm, MPI_Status *status)

int MPI_Recv_c(void *buf, MPI_Count count, MPI_Datatype datatype,
             int source, int tag, MPI_Comm comm, MPI_Status *status)
```

MPI 3.1 Version

BigCount Version

Fortran 2008 binding

```
MPI_Recv(buf, count, datatype, source, tag, comm, status, ierror)
    TYPE(*), DIMENSION(..) :: buf
    INTEGER, INTENT(IN) :: count, source, tag
    TYPE(MPI_Datatype), INTENT(IN) :: datatype
    TYPE(MPI_Comm), INTENT(IN) :: comm
    TYPE(MPI_Status) :: status
    INTEGER, OPTIONAL, INTENT(OUT) :: ierror

MPI_Recv(buf, count, datatype, source, tag, comm, status, ierror) !(_c)
    TYPE(*), DIMENSION(..) :: buf
    INTEGER(KIND=MPI_COUNT_KIND), INTENT(IN) :: count
    TYPE(MPI_Datatype), INTENT(IN) :: datatype
    INTEGER, INTENT(IN) :: source, tag
    TYPE(MPI_Comm), INTENT(IN) :: comm
    TYPE(MPI_Status) :: status
    INTEGER, OPTIONAL, INTENT(OUT) :: ierror
```

Fortran binding

```
MPI_RECV(BUF, COUNT, DATATYPE, SOURCE, TAG, COMM, STATUS, IERROR)
    <type> BUF(*)
    INTEGER COUNT, DATATYPE, SOURCE, TAG, COMM, STATUS(MPI_STATUS_SIZE),
              IERROR
```

Big Count = Big Task

List of functions affected (133)

MPI_Accumulate
MPI_Allgather
MPI_Allgatherv
MPI_Allreduce
MPI_Alltoall
MPI_Alltoallv
MPI_Alltoallw
MPI_Bcast
MPI_Bsend
MPI_Bsend_init
MPI_CONVERSION_FN_NULL
MPI_Comm_spawn_multiple
MPI_Dist_graph_neighbors_count
MPI_Exscan
MPI_File_iread
MPI_File_iread_all
MPI_File_iread_at
MPI_File_iread_at_all
MPI_File_iread_shared
MPI_File_iwrite
MPI_File_iwrite_all
MPI_File_iwrite_at
MPI_File_iwrite_at_all
MPI_File_iwrite_shared
MPI_File_read
MPI_File_read_all
MPI_File_read_all_begin
MPI_File_read_at
MPI_File_read_at_all
MPI_File_read_at_all_begin
MPI_File_read_ordered
MPI_File_read_ordered_begin
MPI_File_read_shared
MPI_File_write
MPI_File_write_all
MPI_File_write_all_begin
MPI_File_write_at
MPI_File_write_at_all
MPI_File_write_at_all_begin
MPI_File_write_ordered
MPI_File_write_ordered_begin
MPI_File_write_shared
MPI_Gather
MPI_Gatherv
MPI_Get
MPI_Get_accumulate
MPI_Get_count
MPI_Get_elements
MPI_Get_elements_x
MPI_Graph_neighbors_count
MPI_Iallgather
MPI_Iallgatherv
MPI_Iallreduce
MPI_Ialltoall
MPI_Ialltoallv
MPI_Ialltoallw
MPI_Ibcast
MPI_Ibsend
MPI_Iexscan
MPI_Igather
MPI_Igatherv
MPI_Imrecv
MPI_Ineighbor_allgather
MPI_Ireduce
MPI_Ireduce_scatter
MPI_Ireduce_scatter_block
MPI_Irsend
MPI_Iscan
MPI_Iscatter
MPI_Iscatterv
MPI_Isend
MPI_Issend
MPI_Mrecv
MPI_Neighbor_allgather
MPI_Neighbor_allgatherv
MPI_Neighbor_alltoall
MPI_Neighbor_alltoallv
MPI_Neighbor_alltoallw
MPI_Pack
MPI_Pack_external
MPI_Pack_external_size
MPI_Pack_size
MPI_Put
MPI_Raccumulate
MPI_Recv
MPI_Recv_init
MPI_Reduce_local
MPI_Reduce_scatter
MPI_Reduce_scatter_block
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_Ssend
MPI_Ssend_init
MPI_Startall
MPI_Status_set_elements
MPI_Status_set_elements_x
MPI_T_cvar_handle_alloc
MPI_T_pvar_handle_alloc
MPI_Testall
MPI_Testany
MPI_Testsome
MPI_Type_contiguous
MPI_Type_create_hindexed
MPI_Type_create_hindexed_block
MPI_Type_create_hvector
MPI_Type_create_indexed_block
MPI_Type_create_struct
MPI_Type_get_extent_x
MPI_Type_get_true_extent_x
MPI_Type_indexed
MPI_Type_size_x
MPI_Type_vector
MPI_Unpack
MPI_Unpack_external
MPI_Waitall
MPI_Waitany
MPI_Waitsome

We already have a few "_X" functions

„Pythonization“

All bindings are generated via embedded Python

- Initially introduced as in a neutral way
- Vetted by all WGs
- Then used for BigCount „flipped a switch“

```
\begin{mpi-binding}
    function_name("MPI_Send")

    parameter(
        "buf", "BUFFER", desc="initial address of
         send buffer", constant=True)
    parameter(
        "count",
        "XFER_NUM_ELEM_NNI_SMALL",
        desc="number of elements in send buffer",)
    parameter(
        "datatype", "DATATYPE", desc="datatype of
         each send buffer element)
    parameter("dest", "RANK", desc="rank of
              destination")
    parameter("tag", "TAG", desc="message tag")
    parameter("comm", "COMMUNICATOR")
\end{mpi-binding}
```

Consequences

- Slightly different rendering
- More consistency
- Uncovered errors

New opportunities

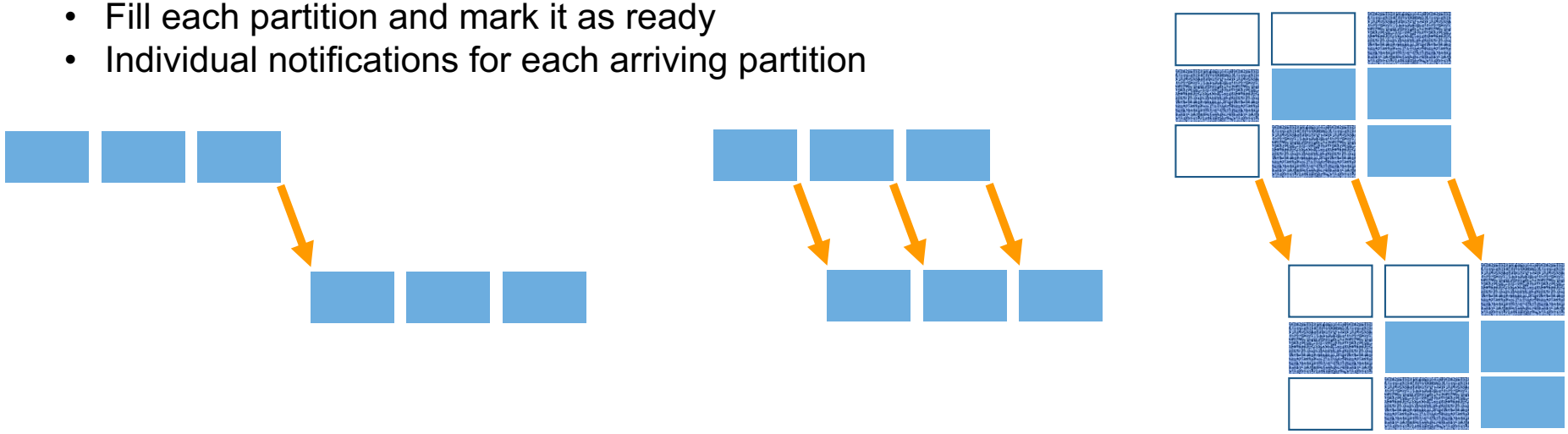
- API / Query mechansism is being developed
- Machine readable description of all MPI routines
- Automatic extraction of the interface
- Eases future standard-wide changes
- Enables better tool support (e.g., generation of PMPI tools)

Partitioned Communication

Partitioned Communication

Core idea – efficient highly concurrent communication

- Built on the concept of persistent P2P communication
- Send and receive buffers are split into (possibly different) partitions
  - Fill each partition and mark it as ready
  - Individual notifications for each arriving partition



Notifications - on send and receive side – are light-weight

- May be driven from light weight environments, without entire MPI stack
- May need additional synchronization to trigger message transfer safely

Partitioned Communication for Thread Support

```
MPI_Psend_init(..., &request);
for (...) {
  MPI_Start(&request);
  #pragma omp parallel
  {
          kernel(..., request);
  }
  MPI_Wait(&request);
}
MPI_Request_free(&request);
```

```
Thread:

kernel(..., MPI_Request request)
{
  int i = my_partition[my_id];
  /* Compute and fill partition i then mark ready: */
  MPI_Pready(i, request);
}
```

Heavy weight MPI communication outside of parallel region

- Inside only light-weight triggering (easier thread safety)
- Partition for each thread
- Each thread signals when it is ready
- MPI can optimize for latency or bandwidth (or shift)

Persistent Collectives

Persistent Collective Operations

Use-case: a collective operation is done many times in an application

- The specific sends and receives represented never change (size, type, lengths, transfers)
- Opportunities
  - Fixed cost for making optimizations can be amortized
  - Static resource allocation can be done
  - Special limited hardware can be allocated if available

Basics

- Mirror regular nonblocking collective operations
- For each nonblocking MPI collective, MPI now has a persistent variant
- For every MPI_I<coll>, add MPI_<coll>_init
- Parameters are identical to the corresponding nonblocking variant
  - Plus additional MPI_INFO parameter
- All arguments “fixed” for subsequent uses
- Persistent collective operations cannot be matched with other collective calls

Example

```
for (i = 0; i < MAXITER; i++) {
    compute(bufA);
    MPI_Ibcast(bufA, …, rowcomm, &req[0]);
    compute(bufB);
    MPI_Ireduce(bufB, …, colcomm, &req[1]);
    MPI_Waitall(2, req, …);
}
```

Nonblocking collectives API

```
MPI_Bcast_init(bufA, …, rowcomm, &req[0]);
MPI_Reduce_init(bufB, …, colcomm, &req[1]);
for (i = 0; i < MAXITER; i++) {
    compute(bufA);
    MPI_Start(req[0]);
    compute(bufB);
    MPI_Start(req[1]);
    MPI_Waitall(2, req, …);
}
```

Persistent collectives API

Better Error Handling

Improved Error Handling

Goal: allow applications to limit impact of failures to avoid terminations

- Specify that MPI_SUCCESS indicates only the result of the operation, not the state of the MPI library.
- Localize error impact of some MPI operations. MPI_ALLOC_MEM will now raise an error on COMM_SELF, not COMM_WORLD
- Specify that MPI should avoid fatal errors when the user doesn't use MPI_ERRORS_ARE_FATAL
- New MPI Error Handler - MPI_ERRORS_ABORT
- Allow the user to specify the default error handler at mpiexec time.

What can you do with this?

- Point to Point communication with sockets-like error handling
- Enables master/worker and other non-traditional types of applications
- Enterprise applications that want to move from sockets to MPI can do so.

BUT: Not full fault tolerance for MPI!

Topology Solutions

New Ways to Adapt to Topologies

New systems are very hierarchical

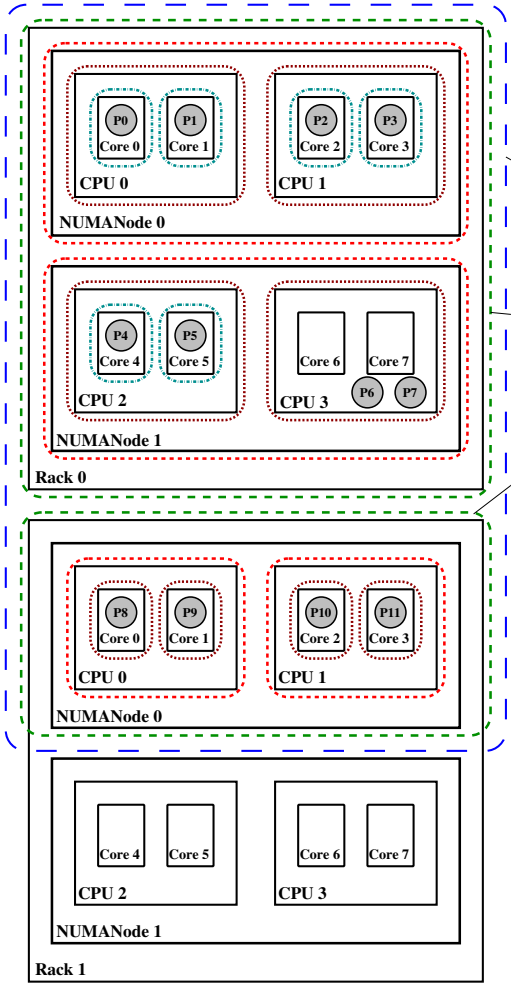
- On node and whole system
- Application mapping is critcal
- Need topology-aware communicators

Guided Mode

- MPI_COMM_SPLIT_TYPE
- Special split type with info key to specify level

Unguided Mode

- Start at MPI_COMM_WORLD
- Step-wise go to lower levels until leaf is reached

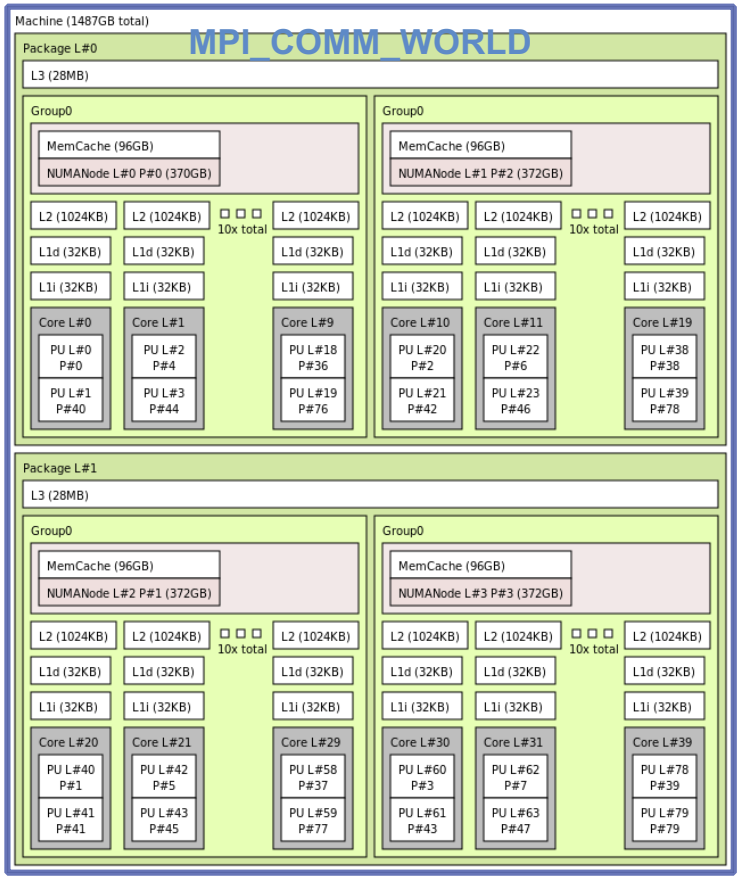


Graphics/Example from Guillaume Mercier, INRIA

Example

MPI_Init(…);

MPI_Comm_rank(MPI_COMM_WOLRD, &rank);



Graphics/Example from Guillaume Mercier, INRIA

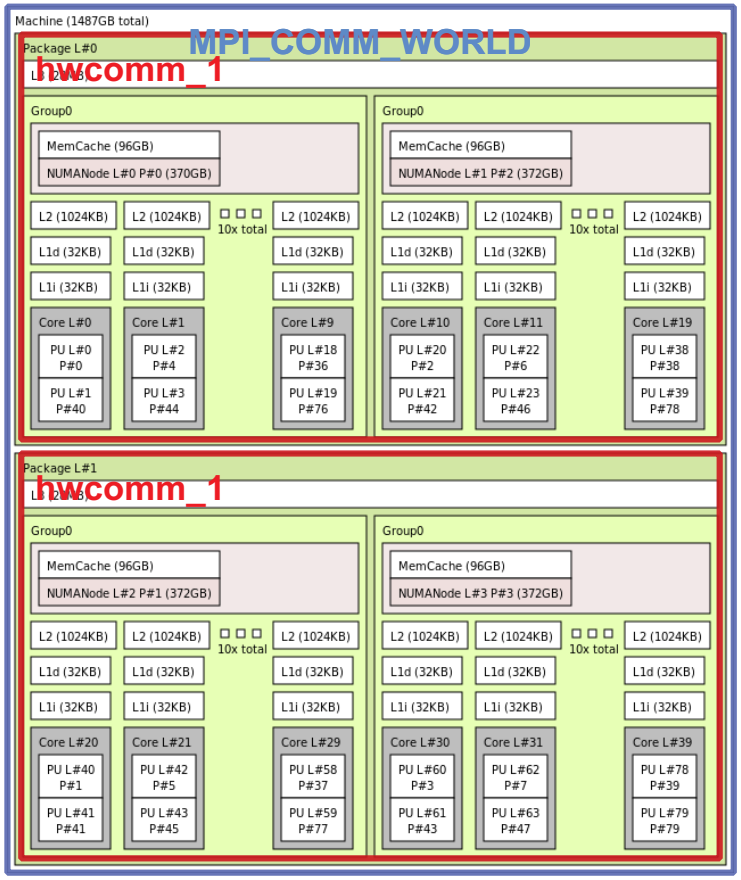
Example

MPI_Init(…);

MPI_Comm_rank(MPI_COMM_WOLRD, &rank);

MPI_Comm_split_type(MPI_COMM_WORLD,

MPI_COMM_TYPE_HW_UNGUIDED,
rank,info,&hwcomm_1);



Graphics/Example from Guillaume Mercier, INRIA

Example

MPI_Init(…);

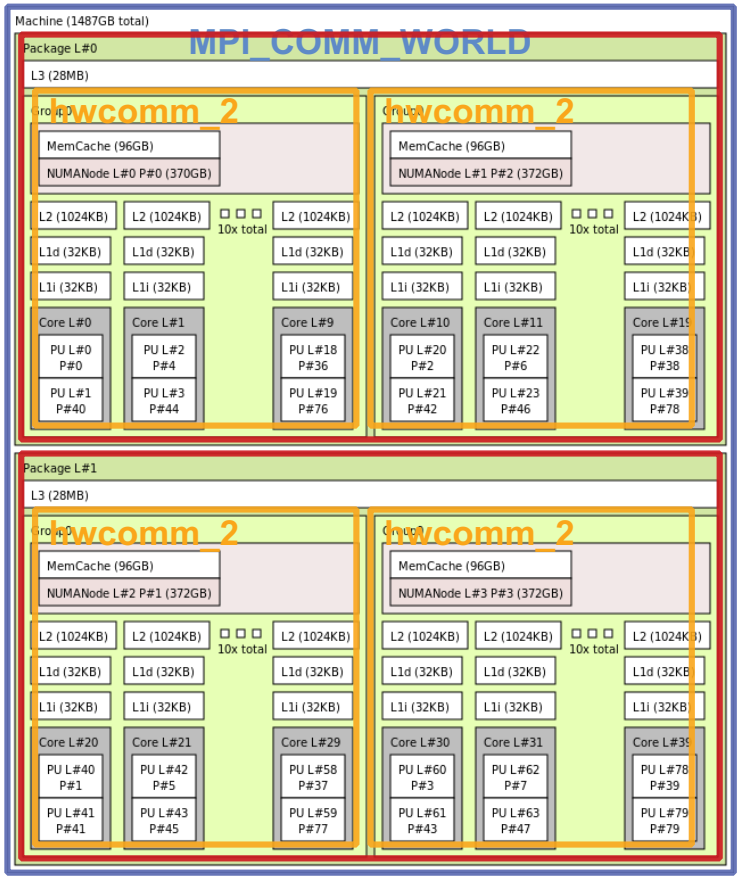
MPI_Comm_rank(MPI_COMM_WOLRD, &rank);

MPI_Comm_split_type(MPI_COMM_WORLD,

MPI_COMM_TYPE_HW_UNGUIDED,
rank,info,&hwcomm_1);

MPI_Comm_split_type(hwcomm_1,

MPI_COMM_TYPE_HW_UNGUIDED,
rank,info,&hwcomm_2);



Graphics/Example from Guillaume Mercier, INRIA

MPI Sessions

MPI Sessions

Attacking some fundamental problems in MPI

- MPI_COMM_WORLD is a very static resource
  - Minimized the complexity
- MPI_COMM_WORLD is immutable
  - Avoids many issues and reduces overhead (and made it not PVM ☺)
- MPI has no ability to isolate resources
  - This was not necessary in the past, very little to isolate
- MPI has no ability to “talk” to the runtime system
  - Explicit choice to improve portability, separation of concerns and matched HPC setups
- MPI is seen as not supporting new communities and industrial applications
  - HPC was the initial target and is still the main area

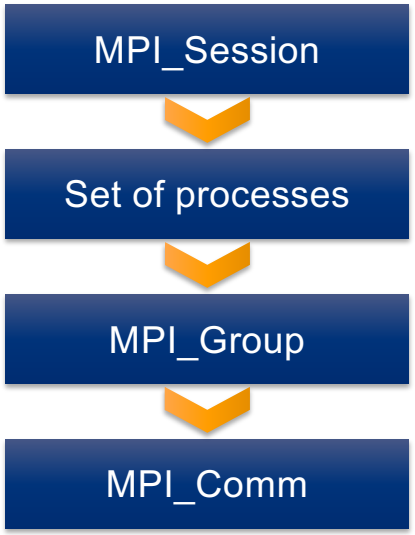
Question: How can we overcome this, without compromising MPI?

- Backwards compatible, but that is only part of it
- Same look and feel of MPI, maintain the learning curve
- Enabe code reuse for existing codes

MPI Sessions

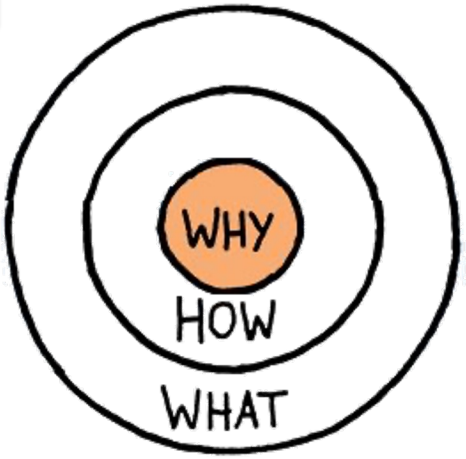
Instead of MPI_Init / MPI_COMM_WORLD:

1. **Get local access to the MPI library**
   Get a Session Handle
2. **Query the underlying run-time system**
   Get a "set" of processes
3. **Determine the processes you want**
   Create an MPI_Group
4. **Create a communicator with just those processes**
   Create an MPI_Comm



What does this do?

- Deliver runtime information of (changing) information to the MPI library
- Enables the ability to provide resource isolation between sessions
- Eliminate the need for a static resource MPI_COMM_WORLD



It's a starting point!

Where Are We?

MPI 4.0 was published 2.5 years ago (in the middle of Corona)

- Solution for “Big Count” operations
- Partitioned Communication
- Persistent Collectives
- Improved Error Handling
- Topology Solutions
- New init options via MPI Sessions
- And much more …

MPI 4.1 was published November 2023

- Clean-Up & terminology adjustments
- Automatic buffer for MPI_Bsend
- Support for different memory kinds, including GPU memory

What's New in MPI 4.1? Updates

MPI 4.1 is a minor update to MPI, mostly bug fixes and clarifications

Many Clarifications

- Terms and behavior of routines in some corner cases.
- See the change log for specifics.

Errata items

- Revert change to MPI_CART_COORDS made by mistake in MPI 4.0
- MPI_STATUS_SET_ELEMENTS for large count added

Deprecated (MPI features that might be removed in a later version)

- Some obscure functions deprecated (MPI_TYPE_SIZE_X and such)
- MPI_HOST (attribute key) deprecated
- mpif.h deprecated (Fortran programmers should use the mpi or mpi_f08 module)

What's New in MPI 4.1? New Features

Added functions to query/set status fields

- E.g., MPI_STATUS_GET_SOURCE. Can still use direct access.

Improved Bsend support

- Automatic, "unlimited" buffering for BSEND added
- New buffer attach/detach for communicators and sessions.
- New routines to complete communication (e.g., MPI_COMM_FLUSH_BUFFER)

Added functions to query status on a request

- E.g., MPI_REQUEST_GET_STATUS_ANY
- Without freeing the request

Additions to HW Topologies

- Added MPI_COMM_TYPE_RESOURCE_GUIDED for MPI_COMM_SPLIT_TYPE
- Added MPI_GET_HW_RESOURCE_INFO

Added routines to remove error codes, classes, and strings

- For those added by the user

New in MPI 4.1: Allocator Kind Info

Use MPI info to provide users with a portable solution to:

1. Detect whether accelerator memory is supported by the MPI library
2. Request support for accelerator memory from the MPI library (when using Sessions)
3. Constrain usage of accelerator memory to specific communicators, windows, etc.

mpi_request_memory_alloc_kind

- Request support for memory allocator kind from the MPI library

mpi_assert_memory_alloc_kind

- Assert memory kinds used by the application on the given MPI object

mpi_memory_alloc_kind

- Memory kinds supported by the MPI library

Slide by Jim Dinan, NVIDIA

New in MPI 4.1: Allocator Kind Info

Use MPI info to provide users with a portable solution to:

1. Detect whether accelerator memory is supported by the MPI library
2. Request support for accelerator memory from the MPI library (when using **Sessions**)
3. Constrain usage of accelerator memory to specific communicators, windows, etc.

mpi_request_memory_alloc_kind

- Request support for memory allocator kind from the MPI library

mpi_assert_memory_alloc_kind

- Assert memory kinds used by the application on the given MPI object

mpi_memory_alloc_kind

- Memory kinds supported by the MPI library

Slide by Jim Dinan, NVIDIA

Request Support for CUDA Allocated Memory

```
bool cuda_aware = false;
int len = MPI_MAX_INFO_VAL, flag = 0;
char *val = malloc(MPI_MAX_INFO_VAL);
MPI_Info info;

MPI_Info_create(&info);
MPI_Info_set(info, “mpi_memory_alloc_kinds”, “cuda:device”);
MPI_Session_init(info, MPI_ERRORS_ARE_FATAL, &session);
MPI_Info_free(&info);

MPI_Session_get_info(session, &info);
MPI_Info_get_string(info, “mpi_memory_alloc_kinds”, &len, val, &flag);

// Check mpi_memory_alloc_kind for “cuda:device”
while (flag && (kind = strsep(&val, ",")) != NULL) {
  if (strcasecmp(kind, "cuda:device") == 0) {
    cuda_aware = true;
    break;
  }
}
```

Slides by Jim Dinan, NVIDIA

Where Are We?

MPI 4.0 was published 2 years ago (in the middle of Corona)

- Solution for “Big Count” operations
- Partitioned Communication
- Persistent Collectives
- Improved Error Handling
- Topology Solutions
- New init options via MPI Sessions
- And much more …

MPI 4.1 planned for November 2023

- Clean-Up & terminology adjustments
- Automatic buffer for MPI_Bsend
- Support for different memory kinds, including GPU memory

Possible MPI 4.2 planned soon

- Efforts to define an MPI ABI are fairly advanced

Moving Towards MPI 5.0

Building on the new concepts of MPI 4.0/4.1

- Partitioned Communication, MPI Sessions, Memory kinds
- Increased presence of persistency

New features that are under discussion

- Better support for accelerated architectures

Partitioned Communication for Thread Support

```
MPI_Psend_init(..., &request);
for (...) {
  MPI_Start(&request);
  #pragma omp parallel
  {
          kernel(..., request);
  }
  MPI_Wait(&request);
}
MPI_Request_free(&request);
```

```
Thread:

kernel(..., MPI_Request request)
{
  int i = my_partition[my_id];
  /* Compute and fill partition i then mark ready: */
  MPI_Pready(i, request);
}
```

Extensions to basic concept

- Guarantees for readiness at receiver
- Different send type options
- Collective versions
- Do non persistent version make sense?

Partitioned Communication for Thread Support

```
MPI_Psend_init(..., &request);
for (...) {
  MPI_Start(&request);
  #pragma omp parallel
  {
          kernel(..., request);
  }
  MPI_Wait(&request);
}
MPI_Request_free(&request);
```

```
Thread:

kernel(..., MPI_Request request)
{
  int i = my_partition[my_id];
  /* Compute and fill partition i then mark ready: */
  MPI_Pready(i, request);
}
```

Extensions to basic concept

- Guarantees for readiness at receiver
- Different send type options
- Collective versions
- Do non persistent version make sense?

Extensions to accelerators

- Heavy weight MPI on host process
- Light-weight triggers on accelerators (without OS)
- Requires translation of requests

Requires GPU bindings for MPI

Partitioned Communication for Thread Support

```
MPI_Psend_init(..., &request);
for (...) {
  MPI_Start(&request);
  #pragma omp target
  {
          GPU_kernel(..., request);
  }
  MPI_Wait(&request);
}
MPI_Request_free(&request);
```

```
Accelerator:

kernel(..., MPI_GPU_Request request)
{
  int i = my_partition[my_id];
  /* Compute and fill partition i then mark ready: */
  MPI_Pready(i, request);
}
```

Extensions to basic concept

- Guarantees for readiness at receiver
- Different send type options
- Collective versions
- Do non persistent version make sense?

Extensions to accelerators

- Heavy weight MPI on host process
- Light-weight triggers on accelerators (without OS)
- Requires translation of requests

Requires GPU bindings for MPI

Accelerator Bindings for MPI Partitioned APIs

CUDA and SYCL Language Bindings Under Exploration

- Consequence of moving partitioned triggers to GPU
- Opens many cans of worms (What is a process? Which bindings? …)

```
int MPI_Psend_init(const void *buf, int partitions, MPI_Count count,
                   MPI_Datatype datatype, int dest, int tag, MPI_Comm comm,
                   MPI_Info info,MPI_Request *request)

int MPI_Precv_init(void *buf, int partitions, MPI_Count count,
                   MPI_Datatype datatype, int source, int tag, MPI_Comm comm,
                   MPI_Info info, MPI_Request *request)

int MPI_[start,wait][_all](...)
```

Keep host only

Add device bindings

```
__device__ int MPI_Pready(int partition, MPI_Request request)

__device__ int MPI_Pready_range(int partition_low, int partition_high, MPI_Request request)

__device__ int MPI_Pready_list(int length, const int array_of_partitions[], MPI_Request
request)

__device__ int MPI_Parrived(MPI_Request request, int partition, int *flag)
```

Separation of Bindings from MPI Concepts

New user communities use new/other languages

- C++, Python, Java, …
- MPI should be available to them, in their „native usage pattern“, not as C wrappers

One idea

- Split the MPI standard into semantics and bindings
- One central semantics document
- One document per language to describe the mapping

Side effects

- Would make us think more about the details of the semantics of the standard
- Question: what does this mean for scripted languages?

Early discussions / New working group

Moving Towards MPI 5.0

Building on the new concepts of MPI 4.0/4.1

- Partitioned Communication, MPI Sessions, Memory kinds
- Increased presence of persistency

New features that are (or IMHO should be) under discussion

- Better support for accelerated architectures
- New language support to enable new communities
- Better integration with task-based runtimes

Proposal for Thread Continuations

Idea: **Treat the completion of an MPI operation as continuation of some activity**
Ability to couple with OpenMP events and dependencies

```
3
void release_event(MPI_Status status, void *data)
{
  omp_event_handle_t event = (omp_event_handle_t)(uintptr_t)data;
  omp_fulfill_event(event);
}
```

```
1
MPI_Request cont_req;
MPIX_Continue_init(&cont_req);

omp_event_handle_t event;
int value;
#pragma omp task depend(out:value) detach(event)
{
  2
  MPI_Request req;
  MPI_Irecv(&value, ..., &req);
  MPIX_Continue(&req, &release_event, event, MPI_STATUS_NULL, cont_req);
}

#pragma omp task depend(in: value)
{
  4
  // process value
}
```

"Callback-based completion notification using MPI Continuations,"
Joseph Schuchart, Christoph Niethammer, José Gracia, George Bosilca, Parallel Computing, 2021.

"MPI Detach - Asynchronous Local Completion,"
Joachim Protze, Marc-André Hermanns, Ali Demiralp, Matthias S. Müller, Torsten Kuhlen. EuroMPI '20.

Moving Towards MPI 5.0

Building on the new concepts of MPI 4.0/4.1

- Partitioned Communication, MPI Sessions, Memory kinds
- Increased presence of persistency

New features that are (or IMHO should be) under discussion

- Better support for accelerated architectures
- New language support to enable new communities
- Better integration with task-based runtimes
- Improved tool support

Improving the Profiling Interface: QMPI

MPI supports the tools via the MPI profiling interface

- Ability to intercept any MPI call, including parameters
- Redirection of MPI calls for new functionality
- Used by many tools

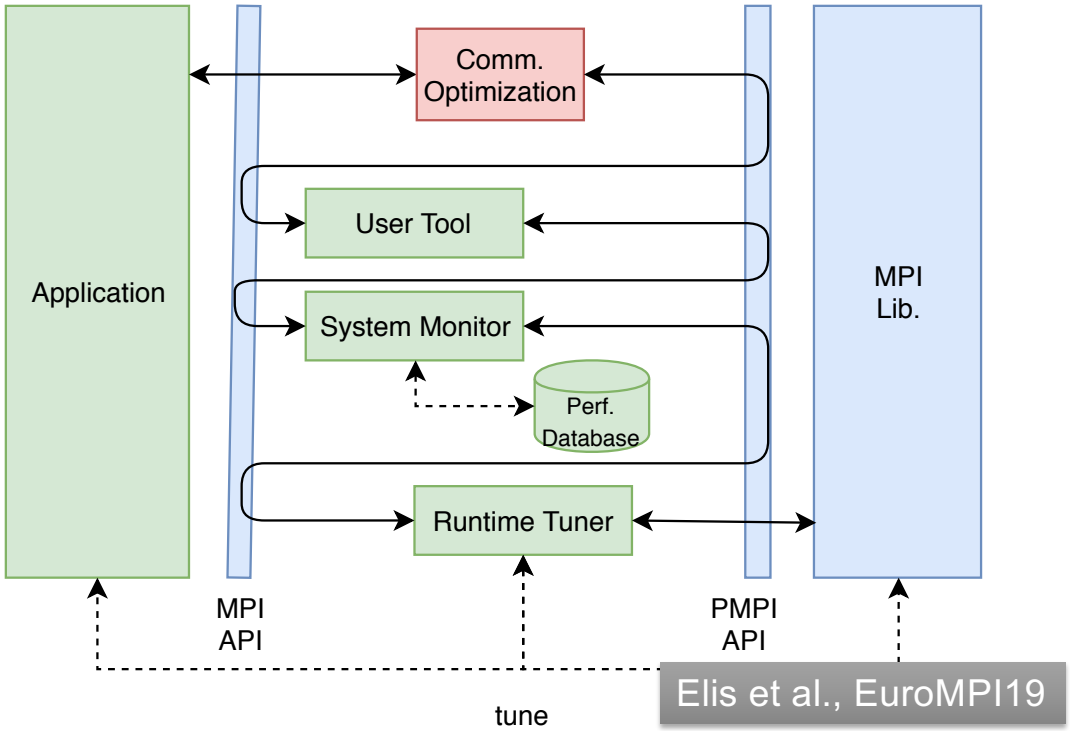
Limitations

- Single tool and user drive
- Usage in Fortran/non-C problematic

Proposal to revamp the interface

- Tools as dynamic wrappers
- Language independent
- Chains of multiple tools

Challenge: Sessions (!)



Moving Towards MPI 5.0

Building on the new concepts of MPI 4.0/4.1

- Partitioned Communication, MPI Sessions, Memory kinds
- Increased presence of persistency

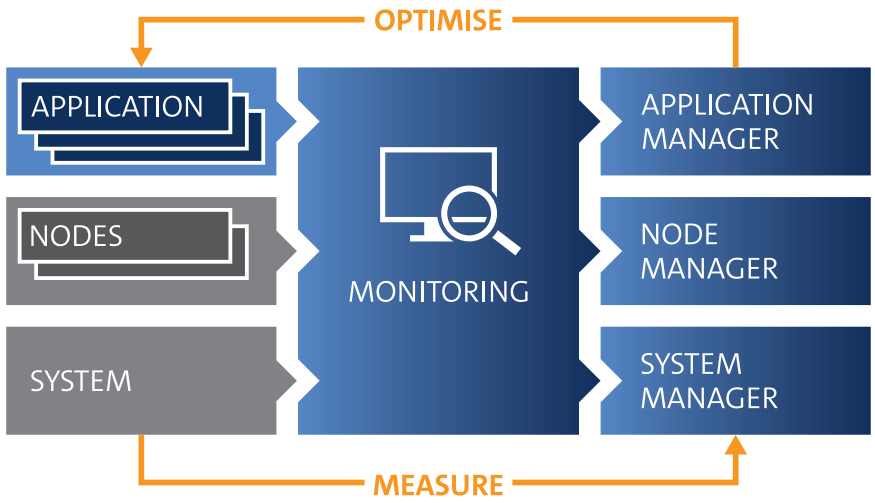
New features that are (or IMHO should be) under discussion

- Better support for accelerated architectures
- New language support to enable new communities
- Better integration with task-based runtimes
- Improved tool support
- Malleability

Malleability / Dynamic Execution

Support for more flexible HPC environments

- Applications to adjust to their sweetspot
- Systems to adjust based on resources
- Accelerated systems with complex nodes
- Communication with disaggregated accelerators
- Complex workflows



New user communities

- Going beyond typical application-driven messaging
- Communicatipon in system software
- Commercial applications beyond HPC (“faster sockets”)

How to change MPI to support malleability?

- Fundamental change of abstraction away from MPI_COMM_WORLD
- Need interactions with the runtime system
- Maintain basic “look & feel” of MPI

SPONSORED BY THE
Federal Ministry of Education and Research

EU Grant #955606
BMBF #16HPC014
DEEP-SEA

EU Grant #956560
BMBF #16HPC039K
REGALE

Time-X #955701
BMBF #16HPC050
TIME-X

Building on top of MPI Sessions

Sessions can provide the needed abstractions

- Transitive closure around all resources in a set of sessions → “MPI bubble”
- Within a bubble normal MPI rules apply
- BUT: bubbles can come and go, interaction with the runtime

Option 1: MPI Session form ”MPI Bubbles”

- “All resources that are derived from a set of resources across a set of MPI processes”
- Implicitly derived from MPI application using sessions
- Within an MPI bubble, normal MPI
- Can invalidate and recreate new bubble,
  while maintaining state

Option 2: Process sets can change

- Ability for the runtime to “tell” something to the application
- Enable process sets ot grow or shrink
- Names are local to MPI Sessions
- Agreement protocol/versioning to agree on new set

Some of the Open Issues

Resource change detection APIs

- Do we want/need notification vs. Polling?
- How to version process sets?
- How to avoid full comparisons of all process sets?

Resource description APIs

- Which resources should be requested?
- Which resources are available?
- Should this be part of MPI or external?

Handshake negotiation APIs

- How to request a change?
- How to inform applications of a change?
- How to capture agreement?
- Central manager or collective?

Connection to fault tolerance proposals ?!?!

Moving Towards MPI 5.0

Building on the new concepts of MPI 4.0/4.1

- Partitioned Communication, MPI Sessions, Memory kinds
- Increased presence of persistency

New features that are (or IMHO should be) under discussion

- Better support for accelerated architectures
- New language support to enable new communities
- Better integration with task-based runtimes
- Improved tool support
- Malleability
- Fault Tolerance

Fault Tolerance for MPI

Initial approaches failed

- Too heavy weight
- Too much oriented to one model
- …

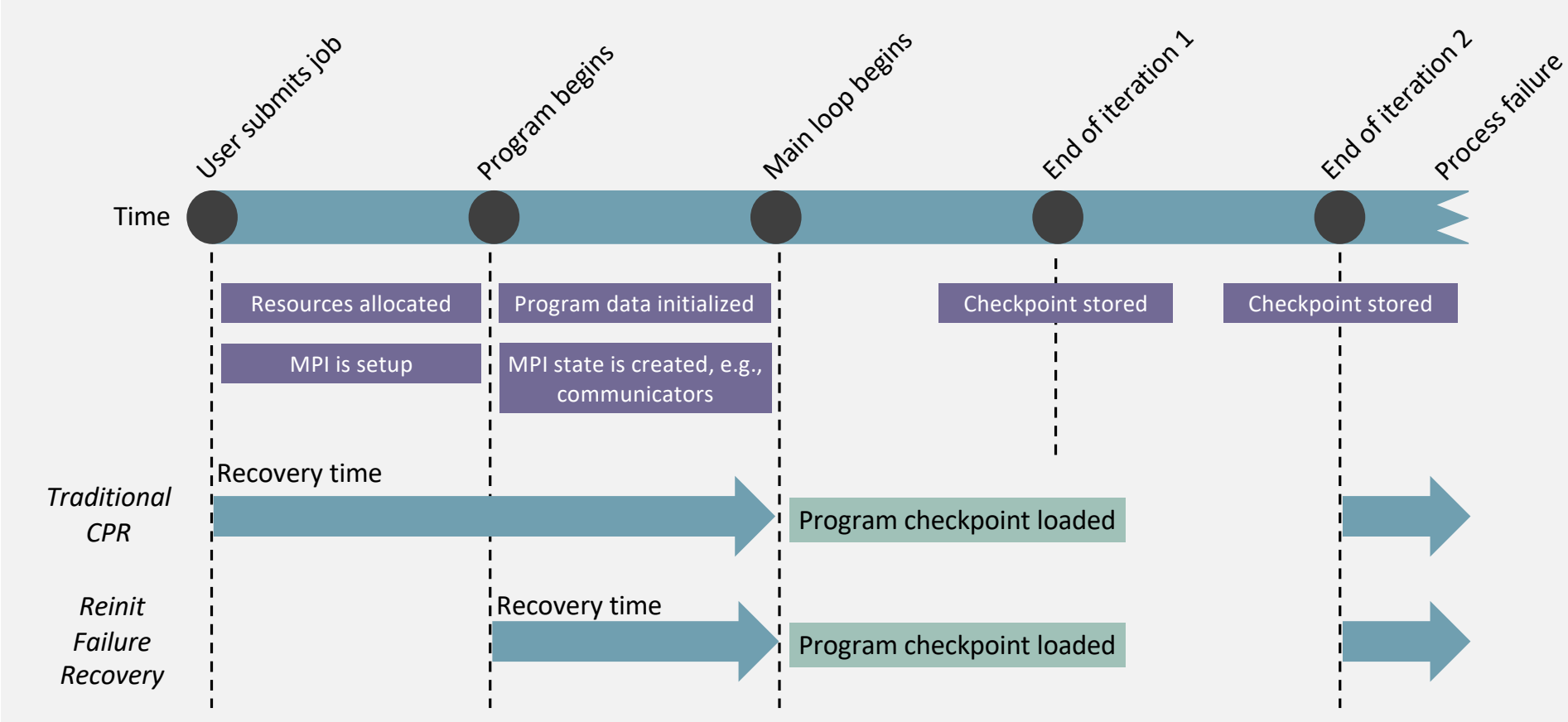
Nevertheless, interest is large

- Continued work in the working group
- Goal: minimal building blocks (many concepts stay, though)
- First step: better error handling in MPI 4.0

Support for several FT models

- Fine grained → ULFM (for new apps)
- Coarse grained → ReInit (to support existing C/R-based apps)
- Session-based

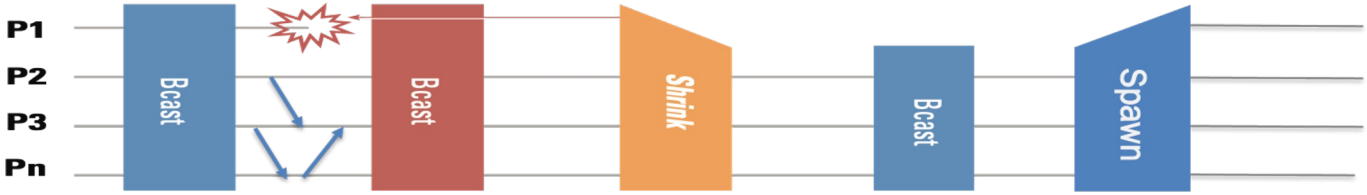
Coarse-grained Recovery (Reinit)



Slide: Ignacio Laguna, LLNL

Fine-grained Fault Tolerance

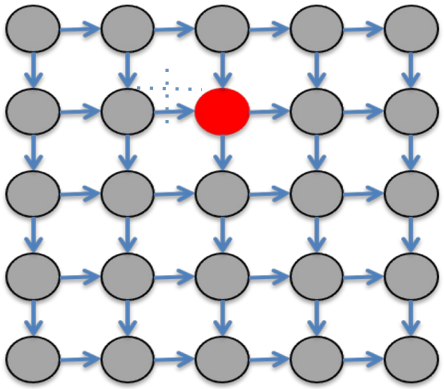
ULFM MPI Crash Recovery (Background)



- Some applications can **continue** w/o recovery
- Some applications are **malleable**
  - Shrink creates a new, smaller communicator on which collectives work
- Some applications are *not* **malleable**
  - Spawn can recreate a "same size" communicator
  - It is easy to reorder the ranks according to the original ordering
  - Pre-made code snippets available

- Failure Notification
- Error Propagation
- Error Recovery
- Respawn of nodes
- Dataset restoration

Not all recovery strategies require all of these features, that's why the interface should split notification, propagation and recovery.

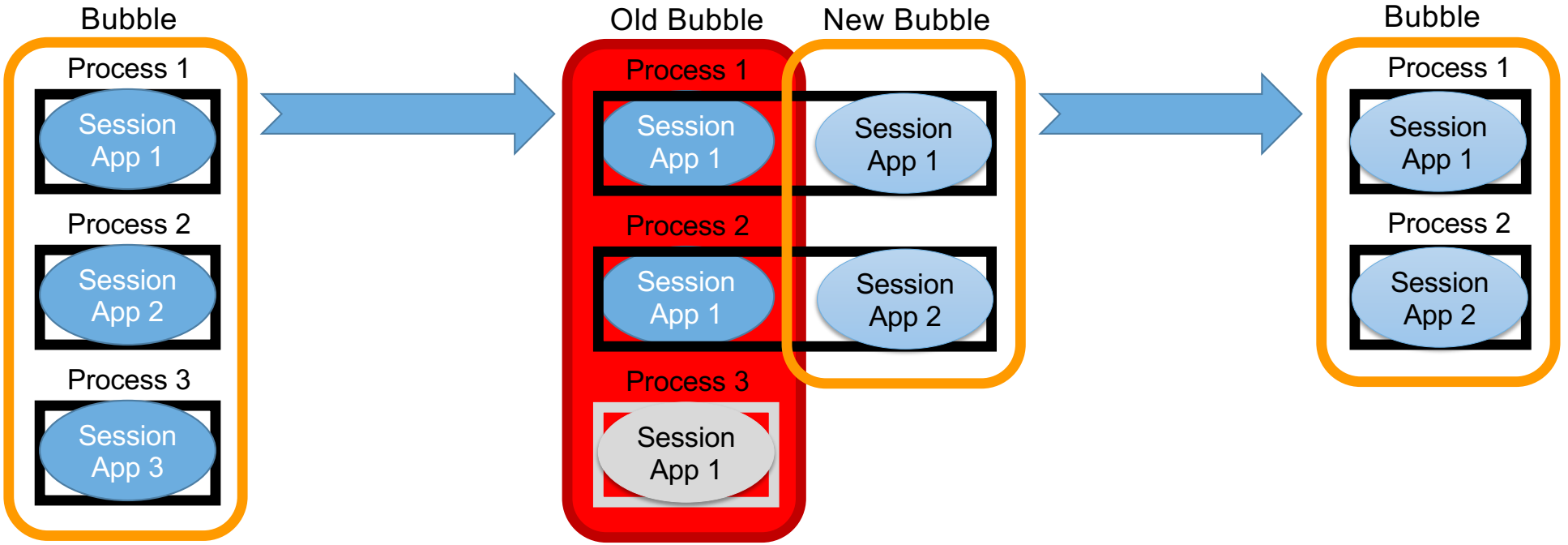


Who should be notified of a failure?
What is the scope of a failure?
What actions should be taken?

- Adds **3 error codes** and **5 functions to manage process crash**
  - **Error codes:** interrupt operations that may block due to process crash
  - **MPI_COMM_FAILURE_ACK / GET_ACKED:** continued operation with ANY-SOURCE RECV and observation known failures
  - **MPI_COMM_REVOKE** lets applications interrupt operations on a communicator
  - **MPI_COMM_AGREE:** synchronize failure knowledge in the application
  - **MPI_COMM_SHRINK:** create a communicator excluding failed processes
  - More info on the MPI Forum ticket #20: https://github.com/mpi-forum/mpi-issues/issues/20

Slide: Aurelien Bouteiller, UTK

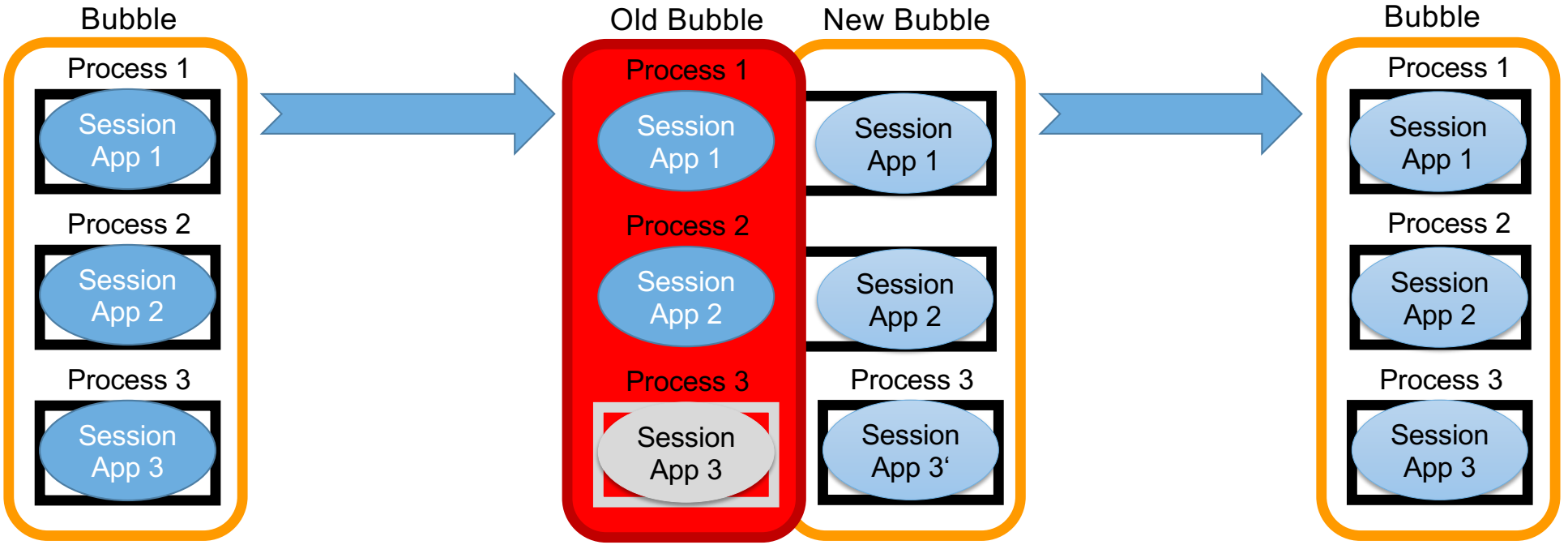
Sessions Can Support Isolation of Failed Processes



Simple support for "Shrinking Recovery"

- Needs functionality to "pop" a bubble
- Supports cleanup, as resources are properly isolated

Sessions Can Support Restart of Resources



Simpe support for “Non-Shrinking Recovery”

- Preferred/required by most applications → default case
- Needs negotiation interface with runtime

The MPI Standard and the MPI Forum

Where we are?

- MPI 4.0 provides powerful new features
- MPI 4.1 offers cleanup and adds minor features
- Strong push and support for an ABI standardization

What is next?

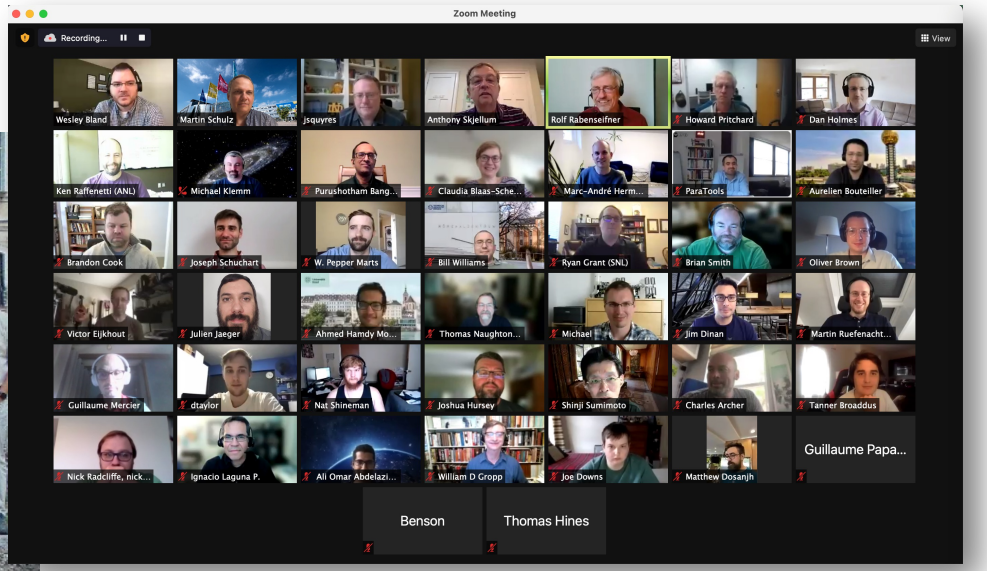
- Many research topics on optimizing new features
- New concepts in MPI 4.1 drive new innovations
  - New support for accelerated systems
  - New language and tooling support
  - Drive towards malleability
- Many more open issues

How can you participate?

- The MPI forum is an open organization
- Join a WG, participate in meetings
- Introduce new concepts so all can benefit

It takes a team, or rather teams …

The TUM CAPS Team



The MPI Forum

Bayerisches Staatsministerium für Wissenschaft und Kunst

Bayerische Forschungsstiftung

SPONSORED BY THE

Federal Ministry of Education and Research

Federal Ministry for Economic Affairs and Climate Action

The MPI Standard and the MPI Forum

Where we are?

- MPI 4.0 provides powerful new features
- MPI 4.1 offers cleanup and adds minor features
- Strong push and support for an ABI standardization

What is next?

- Many research topics on optimizing new features
- New concepts in MPI 4.1 drive new innovations
  - New support for accelerated systems
  - New language and tooling support
  - Drive towards malleability
- Many more open issues

How can you participate?

- The MPI forum is an open organization
- Join a WG, participate in meetings
- Introduce new concepts so all can benefit

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