MPIBlib: Benchmarking MPI Communications for Parallel Computing on Homogeneous and Heterogeneous Clusters

Alexey Lastovetsky Vladimir Rychkov Maureen O'Flynn
{Alexey.Lastovetsky, Vladimir.Rychkov, Maureen.OFlynn}@ucd.ie

Heterogeneous Computing Laboratory
School of Computer Science and Informatics, University College Dublin,
Belfield, Dublin 4, Ireland
http://hcl.ucd.ie

The 15th European PVM/MPI Users Group conference
September 9, 2008, Dublin, Ireland

- Accurate estimation of the execution time of MPI communication operations plays an important role in optimization of parallel applications:
  - Design of parallel applications
  - Tuning collective communication operations
  - **Heterogeneous platforms**

- Accurate estimation of the execution time of MPI communication operations plays an important role in optimization of parallel applications:
  - Design of parallel applications
  - Tuning collective communication operations
  - **Heterogeneous platforms**
- MPI benchmarking suites
  mpptest, NetPIPE, IMB(PMB), SKaMPI, MPIBench
  - Measurement of the execution time of MPI functions - **fixed set of communication operations to be measured (except SKaMPI)**
  - A benchmark methodology - **a single timing method**
  - Not much interpretation of results - **executables and plotting**

- Communication performance modeling - **interpretation of results**
  The procedure of the estimation of parameters determines **what amount of experimental results** *and* **what communication experiments** *are required*

- Communication performance modeling - **interpretation of results**
  The procedure of the estimation of parameters determines **what amount of experimental results** *and* **what communication experiments** *are required*

  - Results of experiments should be available dynamically - **MPI benchmarking library**
  - The communication operations measured by benchmarking suite should be customized - **user-defined communication experiments**
  - The efficiency of measurements is crucial for the modeling at runtime (less accurate can be acceptable) - **selection of timing methods**

- **Benchmark methodology**

  Gropp, W., Lusk E.: **Reproducible Measurements of MPI Performance Characteristics**. *In: Dongarra, J., Luque, E., Margalef, T. (eds.) EuroPVM/MPI 1999. LNCS, vol. 1697, pp. 1118, Springer (1999)*

  - Repeating the communication operation multiple times to obtain the reliable estimation of its execution time
  - Selecting message sizes adaptively to eliminate artifacts in a graph of the output
  - Testing the communication operation in different conditions: cache effects, communication and computation overlap, communication patterns, non-blocking communication etc.

- **Benchmark methodology**

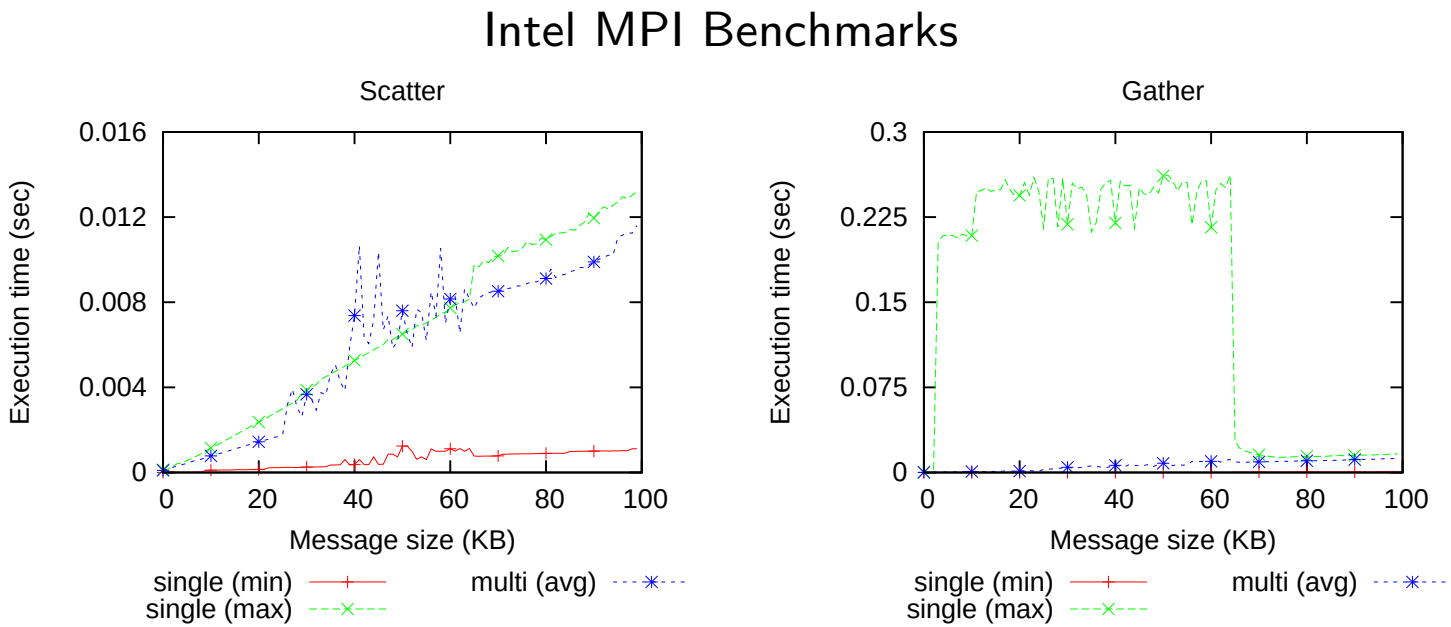
  Gropp, W., Lusk E.: **Reproducible Measurements of MPI Performance Characteristics**. *In: Dongarra, J., Luque, E., Margalef, T. (eds.) EuroPVM/MPI 1999. LNCS, vol. 1697, pp. 1118, Springer (1999)*

  - Repeating the communication operation multiple times to obtain the reliable estimation of its execution time
  - Selecting message sizes adaptively to eliminate artifacts in a graph of the output
  - Testing the communication operation in different conditions: cache effects, communication and computation overlap, communication patterns, non-blocking communication etc.

- Common features on MPI benchmarking suites
  - computing an average, minimum, maximum execution time of a series of the same communication experiments to get accurate results;
  - measuring the communication time for different message sizes - the number of measurements can be fixed or adaptively increased for messages when time is fluctuating rapidly;
  - performing simple statistical analysis by finding averages, variations, and errors.

Scheduling the communication experiment

- Series of communications - **overlapping**



- **Isolation of communication operations from each other - barrier, reduce, short acknowledgments**
  overlapping with these communications

Timing methods - based on MPI_Wtime

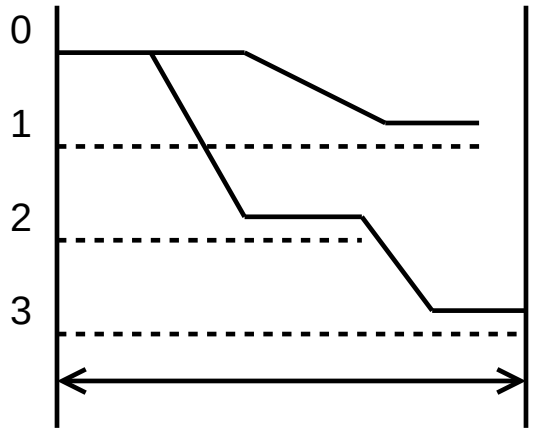
- General - the time between two events:
  - on a single designated processor (*root*)
  - on all participating processors (*max*)
  - on different processors (*global*)

Global timing is the most accurate but the costliest if MPI global timer is not supported by a platform (regular clock synchronization required)

Timing methods - based on MPI_Wtime

- General - the time between two events:
  - on a single designated processor (*root*)
  - on all participating processors (*max*)
  - on different processors (*global*)

  Global timing is the most accurate but the costliest if MPI global timer is not supported by a platform (regular clock synchronization required)
- Operation-specific
  Supinski, B. de, Karonis, N.: **Accurately measuring MPI broadcasts in a computational grid**. *In: The 8th International Symposium on High Performance Distributed Computing, pp. 2937 (1999)*



MPIBlib benchmarking suite

- Implemented as a library - can be integrated into applications
- Provides general and operation-specific timing methods
- Supports extension of the communication operations to be measured

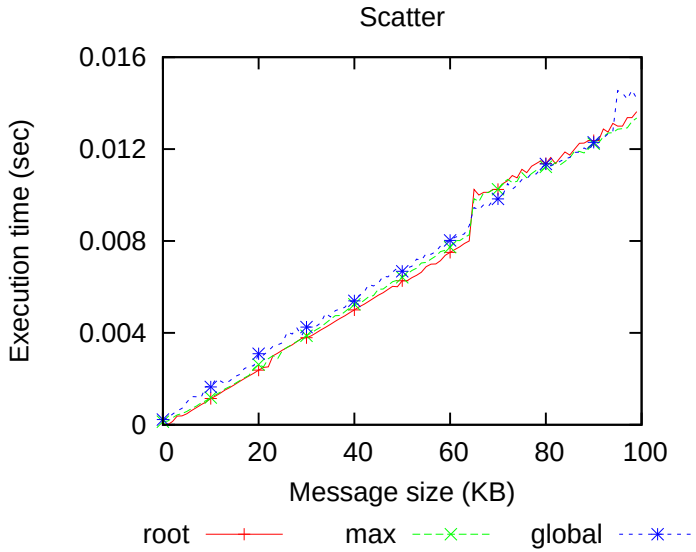
Input accuracy parameters

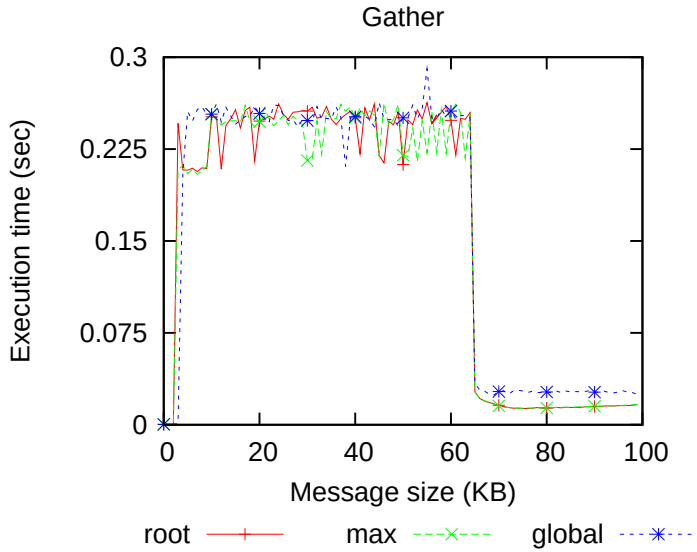
- minimum/maximum numbers of repetitions
  if min_reps == max_reps, the fixed number of measurement
- confidence level and error of estimation
  if min_reps < max_reps, the number of measurement depends on statistics

Output accuracy parameters

- number of repetitions
- confidence interval

Different timing methods on 16 node heterogeneous cluster





Timing method	**Scatter** 0..100KB, 1KB stride, 1 rep (sec)	**Gather** 0..100KB, 1KB stride, 1 rep (sec)
Global	28.7	44.7
Maximum	0.8	15.6
Root	0.8	15.7

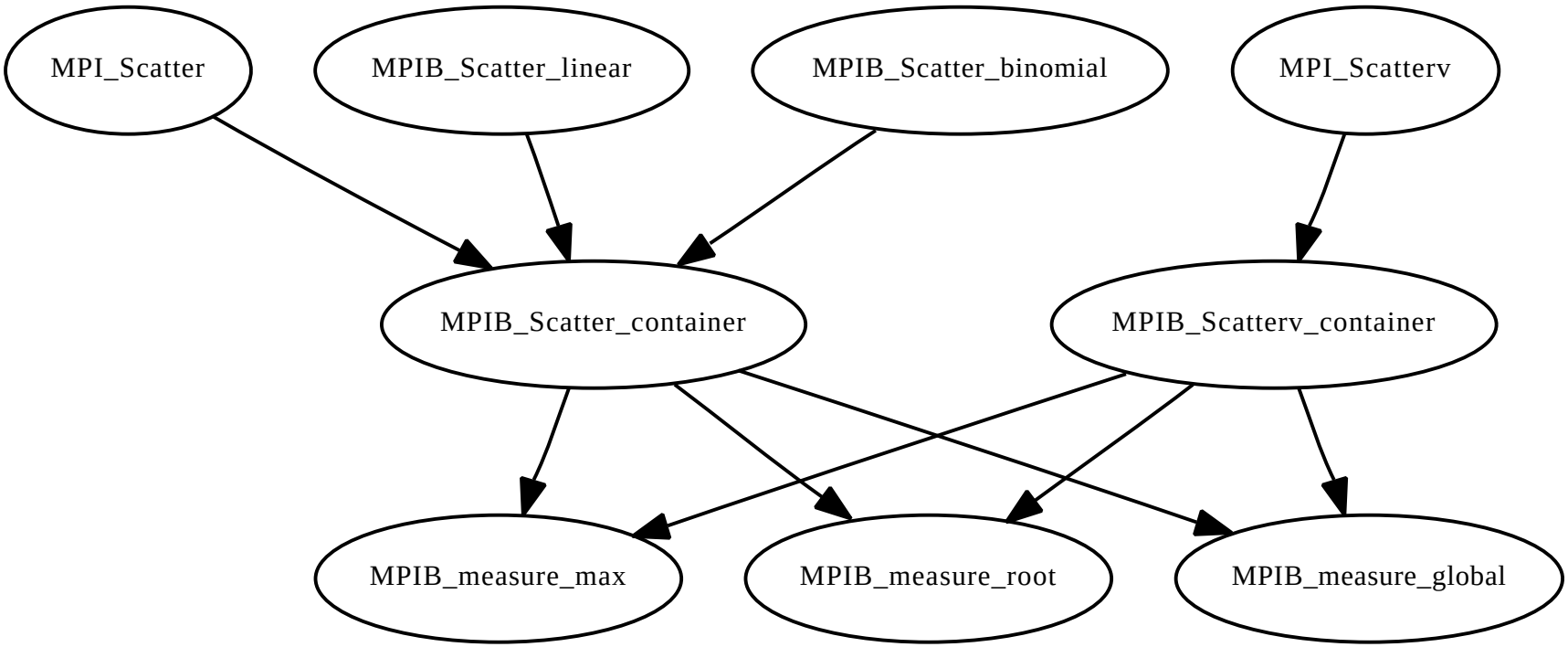
Encapsulation - Special data structure

```
struct MPIB_coll_container {|
 void (*initialize)(void* this, MPI_Comm comm, int root, int M);|
 void (*execute)(void* this, MPI_Comm comm, int root, int M);|
 void (*finalize)(void* this, MPI_Comm comm, int root);|
 void (*free)(void* this);|
}|
```

- Allocation and deallocation of buffers required for the communication operation
- Communication operation
- Release of data structure

```
struct MPIB_Scatter_container {|
 struct MPIB_coll_container base;|
 char* buffer;|
 int (*scatter)(void* sendbuf, int sendcount, MPI_Datatype sendtype,...);|
}|
```

Customization of communication operations



MPI Benchmarking library was used for communication performance modeling on heterogeneous clusters

- Measurement of roundtrips with empty and non-empty messages - *sequential, parallel (clusters with a single switch)*
- Measurement of linear scatter/gather - *root timing*
- User-defined communication operations - *one-to-two* - *sequential, parallel (clusters with a single switch)*

University College Dublin

Science Foundation Ireland

IBM Dublin CAS