



MPI

Framework care facilitează

- Pornirea programelor distribuite (procese pe același sistem sau pe sisteme diferite, dar strâns conectate – ideal aceeași rețea)
- Conectarea proceselor unui program distribuit (accept, bind, connect)
- Simplificarea identificării (identificatori în loc de IP, port)
- Simplificarea comunicării (oferă funcții gen Send/Recv, Broadcast)
- Asigură comunicarea corectă pe sisteme cu arhitecturi de calcul diferite (little/big endian problems)



MPI memoria

- Nu avem memorie partajată în MPI. Arhitectură NUMA
- Toate variabilele sunt locale proceselor.
- Pentru a muta informație de la un proces la altul vor trebuie folosită comunicație, prin apelul funcțiilor oferite de MPI:
 - Send/Recv
 - Broadcast
 - Scatter
 - Gather



Instalare OpenMPI

apt-get install libopenmpi-dev openmpi-bin openmpi-doc openmpi-common



Compiling and running MPI programs

mpicc test.c

mpirun –np 4 a.out mpirun –np 3 date Pornește 4 procese.

Dacă este setat, va porni procesele pe mașini diferite.

Procesele sunt identice dar au id-uri diferite. Funcționează parțial și cu programe care nu sunt implementate pentru MPI.

./a.out ←

Funcționează dar pornește un singur proces.



```
#include<mpi.h>
#include<stdio.h>
int main(int argc, char * argv[])
      int rank;
      int nProcesses;
      MPI Init(&argc, &argv);
      MPI Comm rank(MPI_COMM_WORLD, &rank);
      MPI Comm size(MPI COMM WORLD, &nProcesses);
      printf("Hello from %i/%i\n", rank, nProcesses);
      MPI Finalize();
      return 0;
```



```
#include<mpi.h>
#include<stdio.h>
int main(int argc, char * argv[])
                                   Pornește procesele MPI
      int rank;
      int nProcesses;
      MPI Init(&argc, &argv);
      MPI Comm rank(MPI COMM WORLD, &rank);
      MPI Comm size(MPI COMM WORLD, &nProcesses);
      printf("Hello from %i/%i\n", rank, nProcesses);
      MPI Finalize();
      return 0;
```



```
#include<mpi.h>
#include<stdio.h>
int main(int argc, char * argv[])
                                       Întoarce ID-ul
                                       procesului (rank-ul)
      int rank;
      int nProcesses;
      MPI Init(&argc, &argv);
      MPI Comm rank(MPI COMM WORLD, &rank);
      MPI Comm size(MPI COMM WORLD, &nProcesses);
      printf("Hello from %i/%i\n", rank, nProcesses);
      MPI Finalize();
      return 0;
```



```
#include<mpi.h>
#include<stdio.h>
int main(int argc, char * argv[])
                                     Întoarce numărul total
      int rank;
                                     de procese
      int nProcesses;
      MPI Init(&argc, &argv);
      MPI Comm rank(MPI_COMM_WORLD, &rank); `
      MPI Comm size(MPI COMM WORLD, &nProcesses);
      printf("Hello from %i/%i\n", rank, nProcesses);
      MPI Finalize();
      return 0;
```



```
#include<mpi.h>
#include<stdio.h>
int main(int argc, char * argv[])
      int rank;
      int nProcesses;
      MPI Init(&argc, &argv);
      MPI Comm rank(MPI COMM WORLD, &rank);
      MPI Comm size(MPI COMM WORLD, &nProcesses);
      printf("Hello from %i/%i\n", rank, nProcesses);
      MPI Finalize();
                                   Afișează hello (pentru
      return 0;
                                   fiecare proces pornit).
```



```
#include<mpi.h>
#include<stdio.h>
int main(int argc, char * argv[])
      int rank;
      int nProcesses;
      MPI Init(&argc, &argv);
      MPI Comm rank(MPI COMM WORLD, &rank);
      MPI Comm size(MPI COMM WORLD, &nProcesses);
      printf("Hello from %i/%i\n", rank, nProcesses);
      MPI Finalize();
                                   Oprește programul
      return 0;
                                   MPI.
```



MPI example executed

```
#include<mpi.h>
#include<stdio.h>

int main(int argc, char * argv[])
{
    int rank;
    int nProcesses;
    MPI_Init(&argc, &argv);
    MPI_Comm_rank(MPI_COMM_WORLD, &rank);
    MPI_Comm_size(MPI_COMM_WORLD, &nProcesses);
    printf("Hello from %i/%i\n", rank, nProcesses);
    MPI_Finalize();
    return 0;
}
```

Hello from 0/4

```
#include<mpi.h>
#include<stdio.h>

int main(int argc, char * argv[])
{
    int rank;
    int nProcesses;
    MPI_Init(&argc, &argv);
    MPI_Comm_rank(MPI_COMM_WORLD, &rank);
    MPI_Comm_size(MPI_COMM_WORLD, &nProcesses);
    printf("Hello from %i/%i\n", rank, nProcesses);
    MPI_Finalize();
    return 0;
}
```

Hello from 2/4

```
#include<mpi.h>
#include<stdio.h>

int main(int argc, char * argv[])
{
    int rank;
    int nProcesses;
    MPI_Init(&argc, &argv);
    MPI_Comm_rank(MPI_COMM_WORLD, &rank);
    MPI_Comm_size(MPI_COMM_WORLD, &nProcesses);
    printf("Hello from %i/%i\n", rank, nProcesses);
    MPI_Finalize();
    return 0;
}
```

Hello from 3/4

```
#include<mpi.h>
#include<stdio.h>

int main(int argc, char * argv[])
{
    int rank;
    int nProcesses;
    MPI_Init(&argc, &argv);
    MPI_Comm_rank(MPI_COMM_WORLD, &rank);
    MPI_Comm_size(MPI_COMM_WORLD, &nProcesses);
    printf("Hello from %i/%i\n", rank, nProcesses);
    MPI_Finalize();
    return 0;
}
```

Hello from 1/4





MPI_Send/MPI_Recv

```
int MPI_Send( ↓ void *b, ↓ int c, ↓ MPI_Datatype d, ↓ int reciever, ↓ int t, ↓ MPI_Comm)
```



MPI_Send/MPI_Recv

```
int MPI_Recv( ↑ void *b, ↓ int c, ↓ MPI_Datatype d, ↓ int sender, ↓ int t, ↓ MPI_Comm, ↑ MPI_Status * )
                                             [0, ..)
          &v[3]
                                         MPI ANY TAG
           &a num el(v)
          v+5 [0,...)
                                  [ 0, num_tasks )
                                 MPI ANY SOURCE
                         MPI INT
                                              MPI COMM WORLD
                        MPI CHAR
                        MPI FLOAT
                                                           &Stat
                        MPI LONG
                                                   MPI STATUS IGNORE
                                             Stat.MPI SOURCE, Stat.MPI TAG
```



MPI_Bcast

```
int MPI_Bcast ( ↑ void *b, ↓ int c, ↓ MPI_Datatype d, ↓ int root, ↓ MPI_Comm )
```

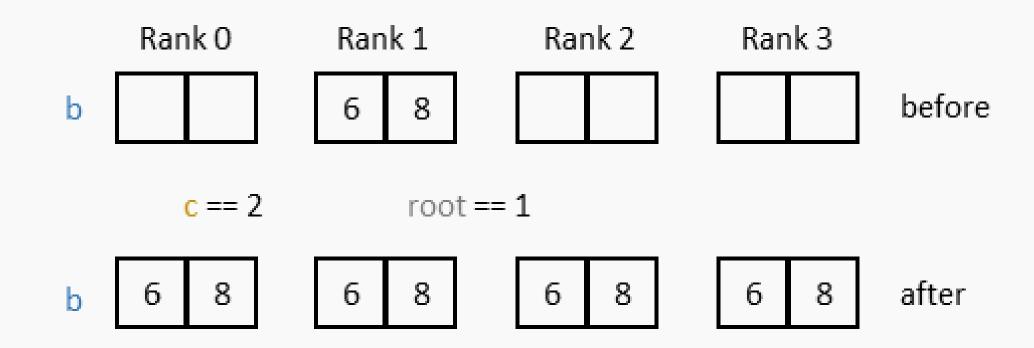
```
num_el(v)
v [0,...)

&v[3]
&a
v+5

MPI_INT
MPI_COMM_WORLD
MPI_CHAR
MPI_FLOAT
MPI_LONG
```



MPI_Bcast





MPI_Scatter

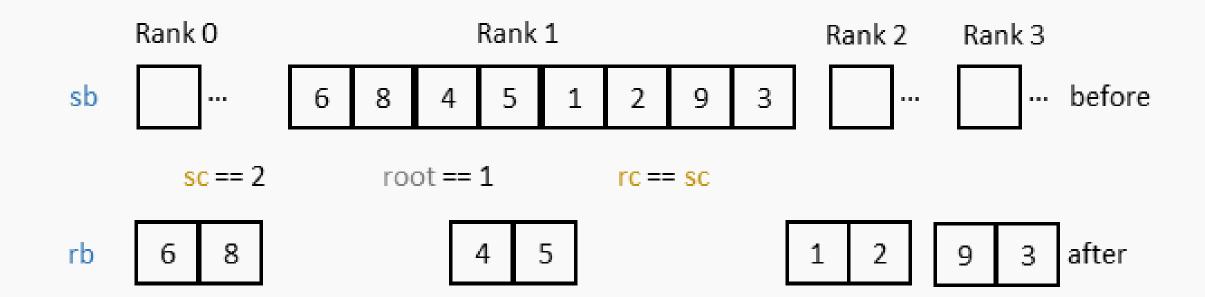
```
int MPI_Scatter ( ↓ void *sb, ↓ int sc, ↓ MPI_Datatype sd, ↑ void *rb, ↓ int rc, ↓ MPI_Datatype rd, ↓ int root, ↓ MPI_Comm )
```

```
[ 0, num_tasks )
num_el(v)/num_tasks
                        num_el(v)/num_tasks
       [0,...)
                                [0,...)
               MPI INT
                                       MPI INT
                          &v[3]
 &v[3]
              MPI_CHAR
                                      MPI CHAR
                           &a
  &a
             MPI FLOAT
                                      MPI FLOAT
                          v+5
  v+5
              MPI LONG
                                      MPI LONG
```

MPI_COMM_WORLD



MPI_Scatter





MPI_Gather

```
int MPI_Gather ( ↓ void *sb, ↓ int sc, ↓ MPI_Datatype sd, ↑ void *rb, ↓ int rc, ↓ MPI_Datatype rd, ↓ int root, ↓ MPI_Comm )
```

```
      num_el(v)/num_tasks
      num_el(v)/num_tasks
      [ 0, num_tasks )

      [0,...)
      v

      &v[3]
      &v[3]

      &a
      &a

      v+5
      v+5
```

MPI_INT MPI_CHAR MPI_FLOAT MPI_LONG MPI_INT MPI_CHAR MPI_FLOAT MPI_LONG

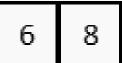
MPI_COMM_WORLD



MPI_Gather

Rank 0

sb



Rank 1



Rank 2

Rank 3

9 3 before

$$sc == 2$$

rb



6

8

4

1

9

.

··· after





MPI blocking/non-blocking send/recv

Proces 1

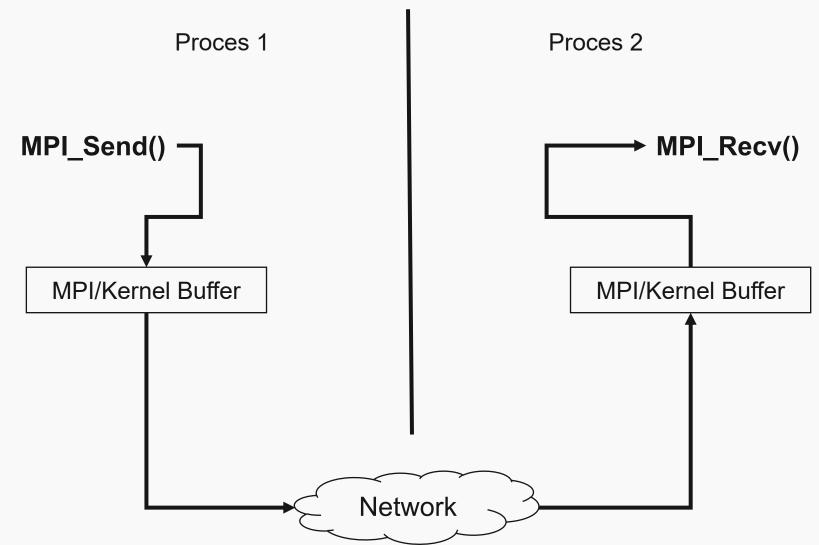
MPI_Send()

Proces 2

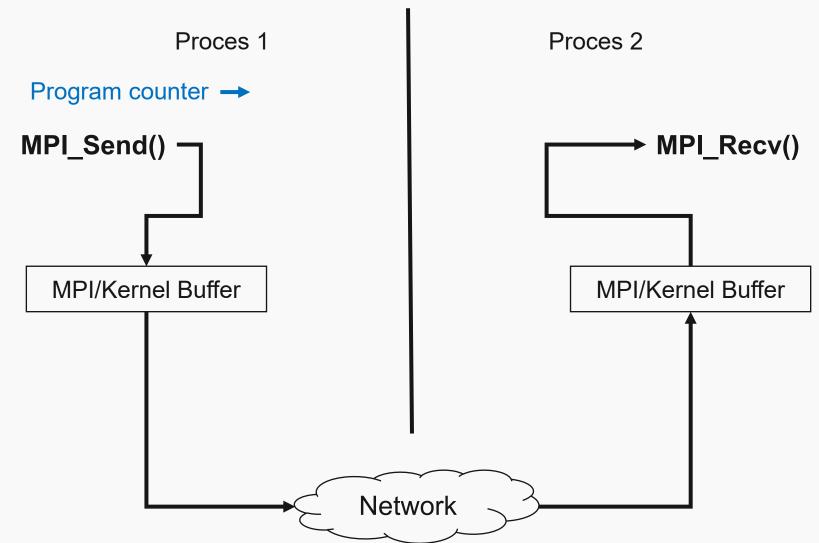
MPI_Recv()



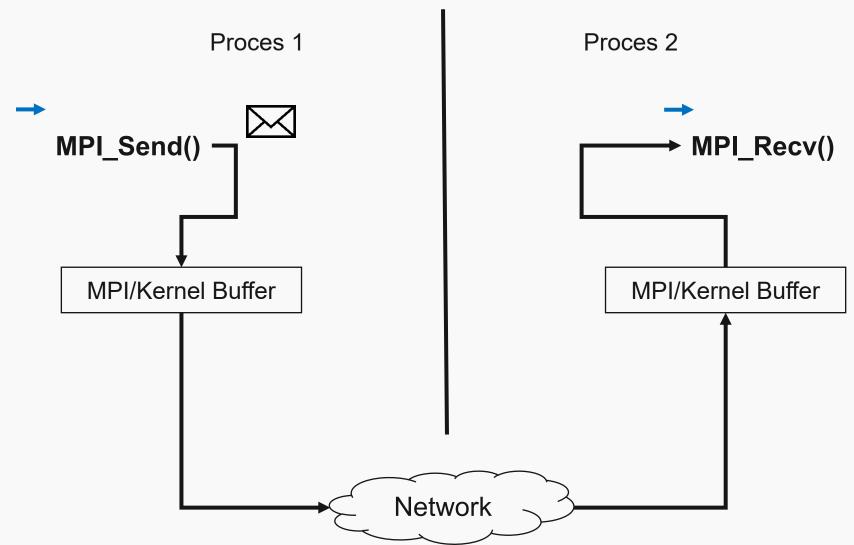
MPI blocking/non-blocking send/recv



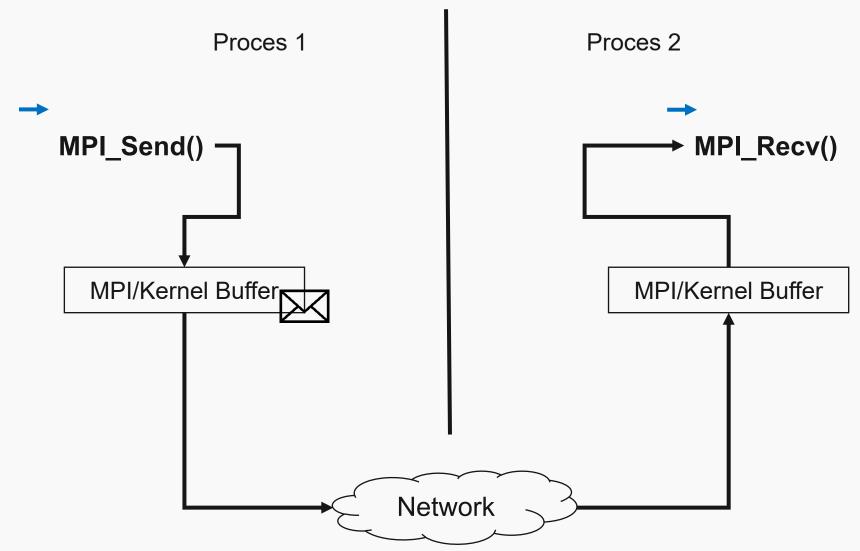




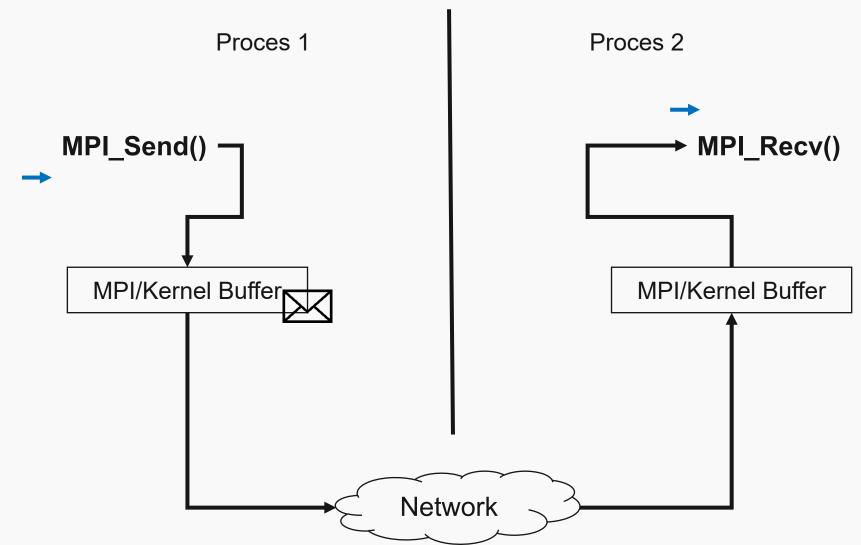




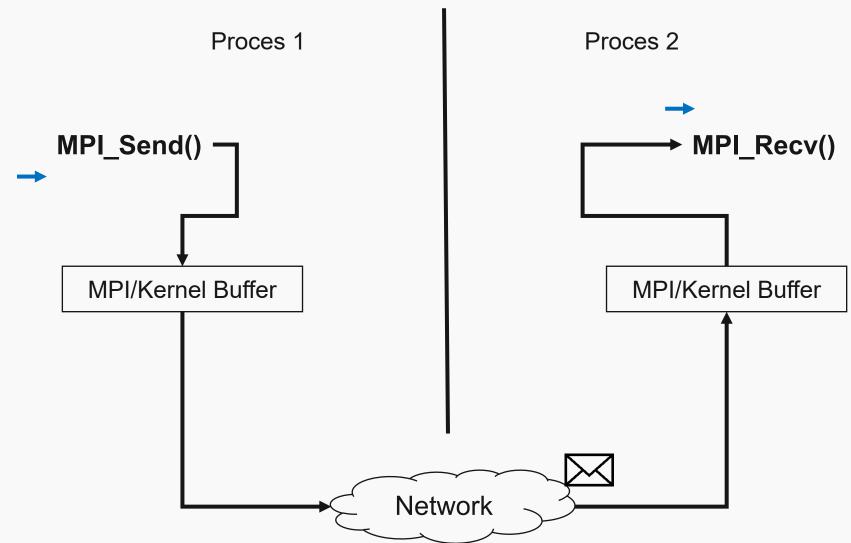




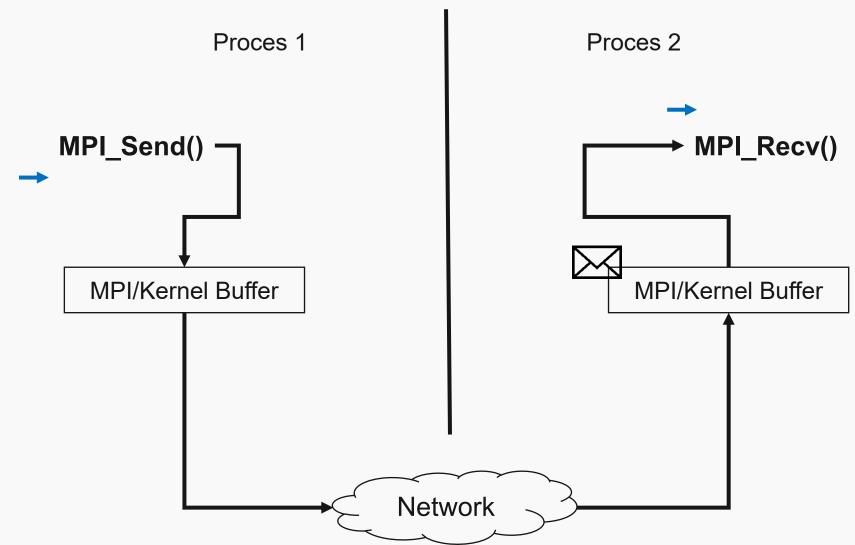




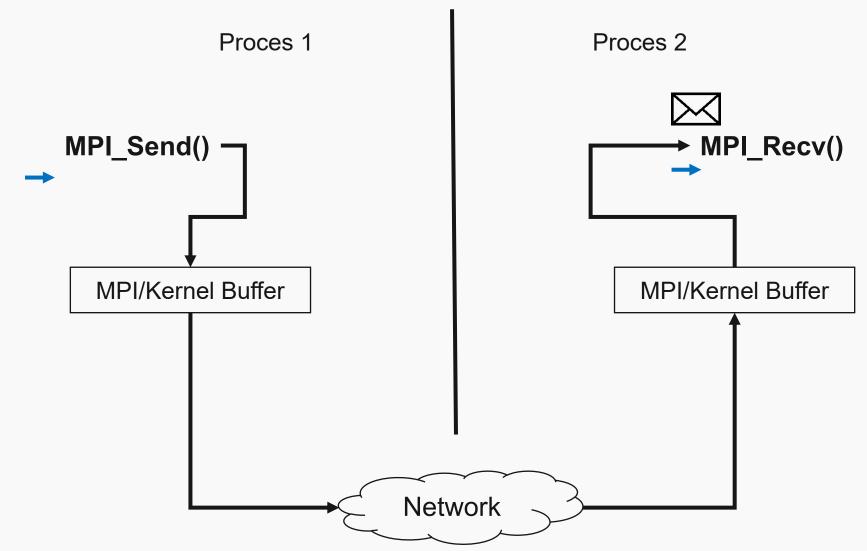




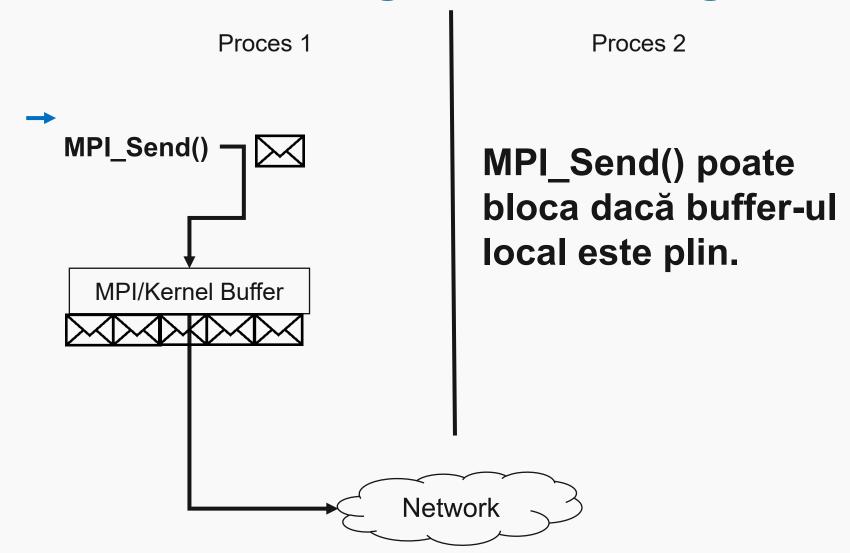




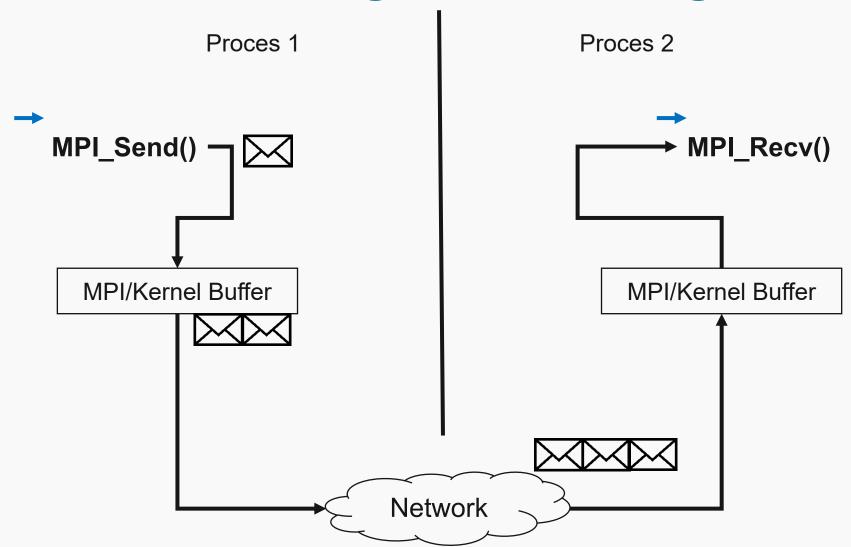




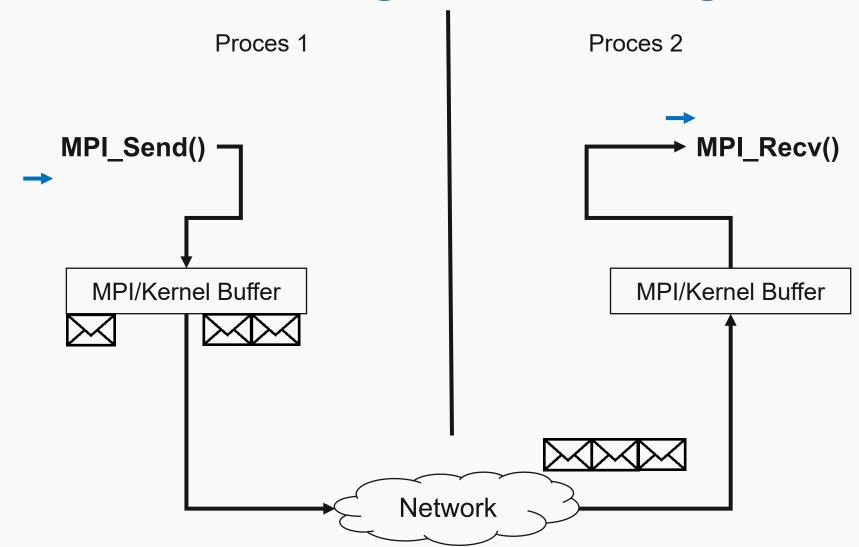










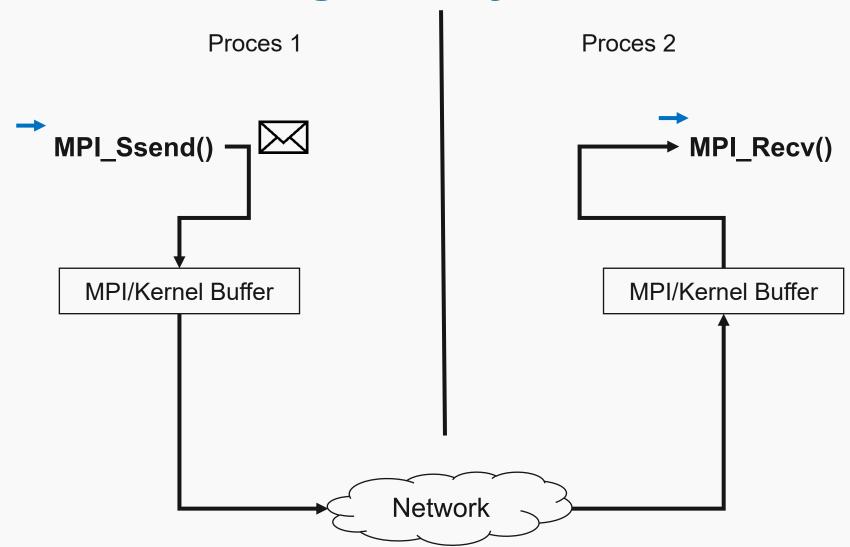




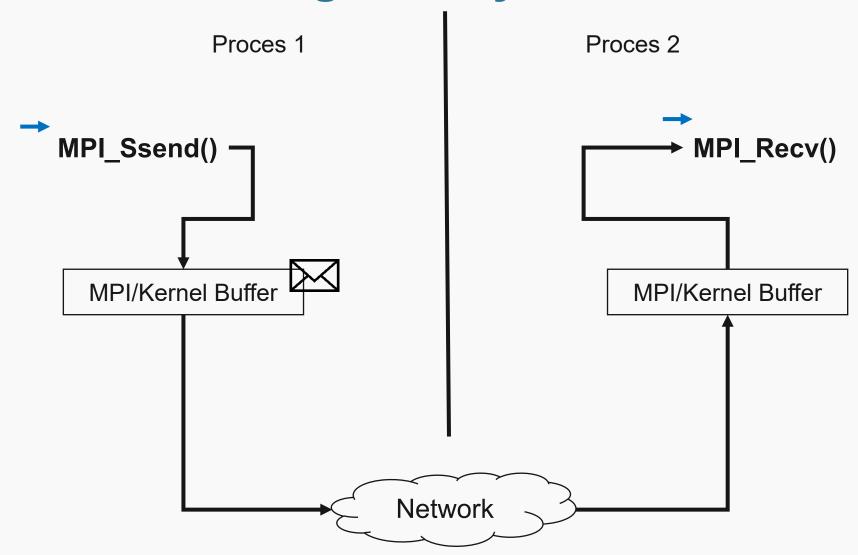
MPI blocking recv/synchronized send



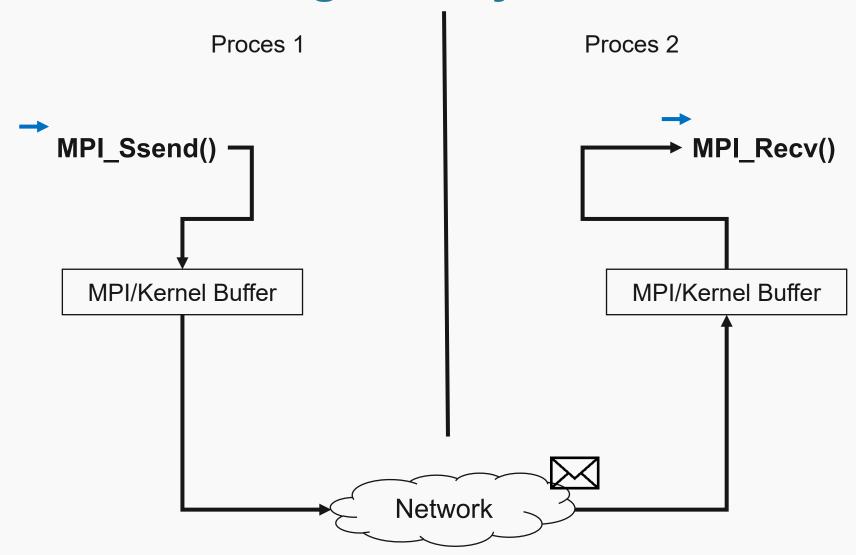
MPI blocking recv/synchronized send



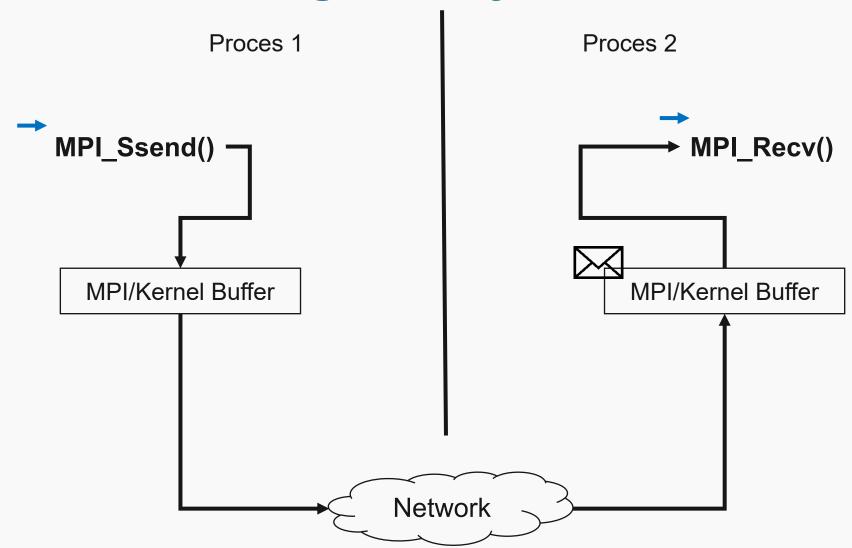




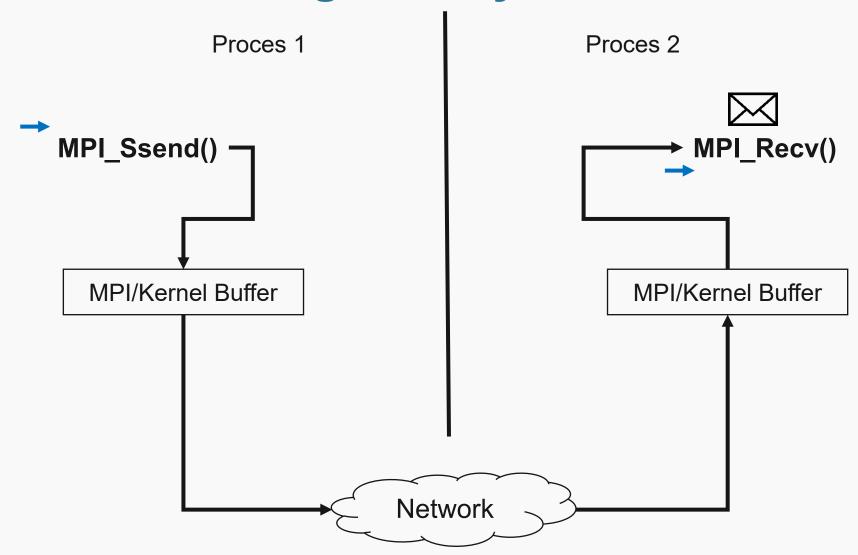




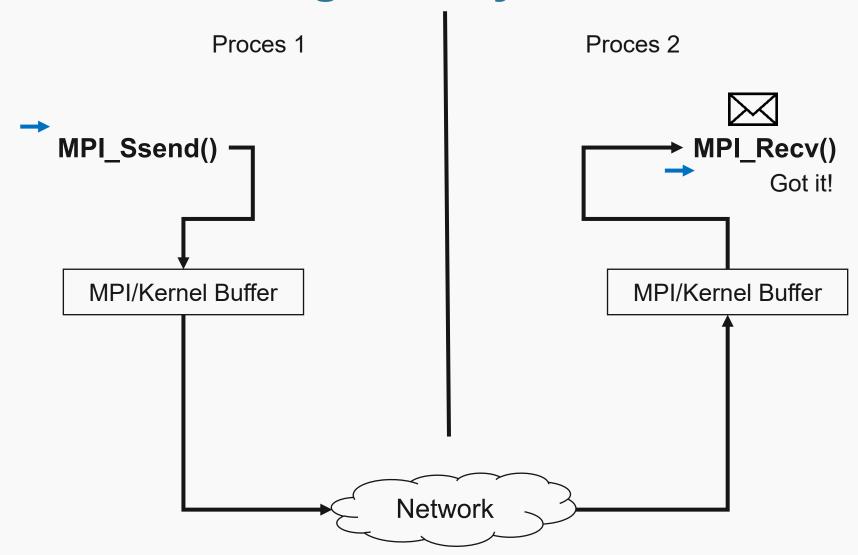




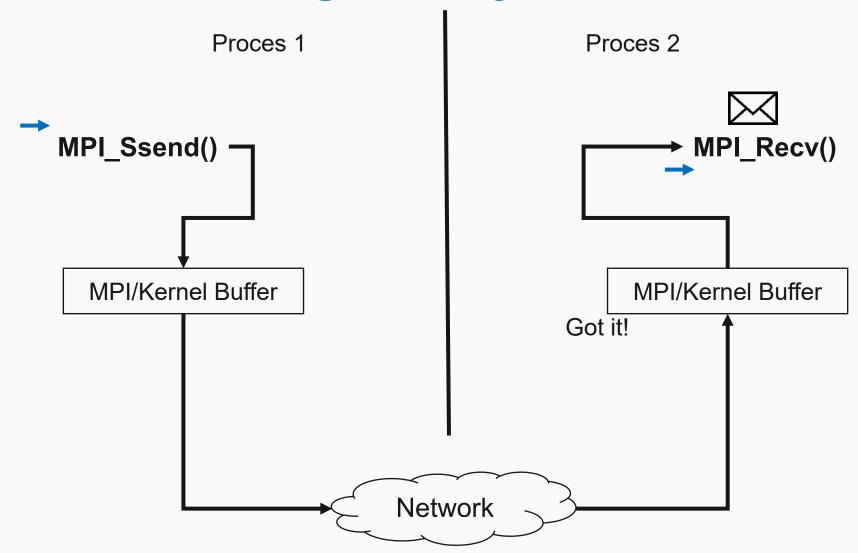




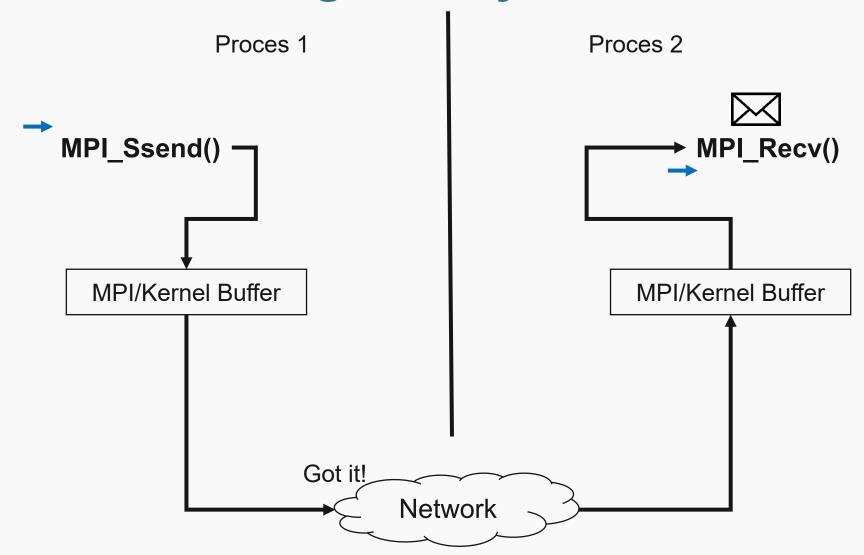




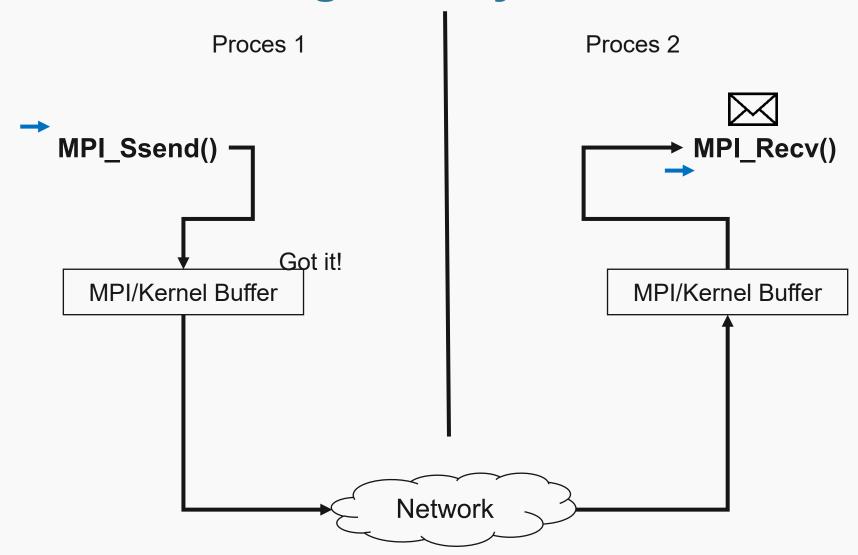




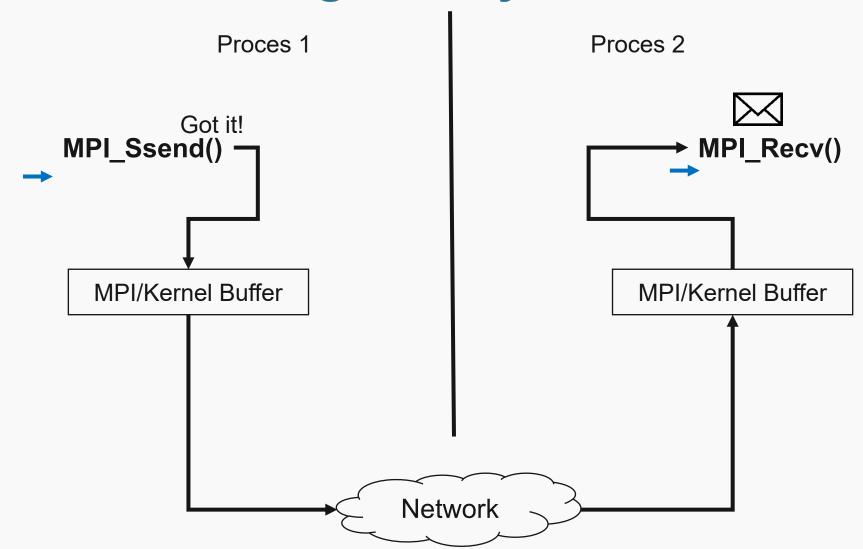






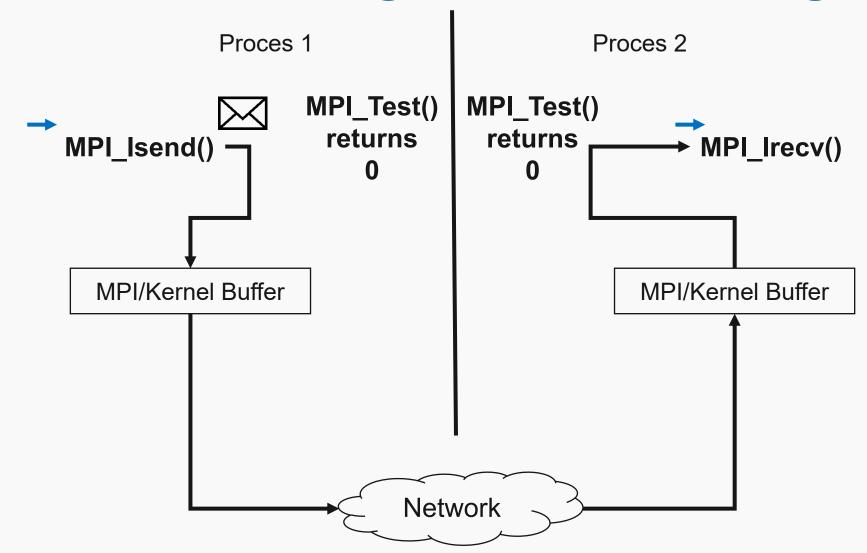




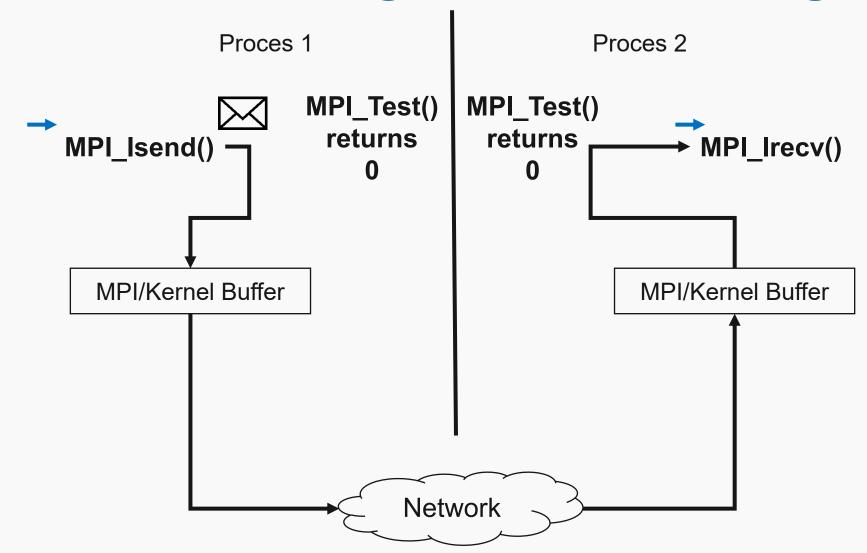




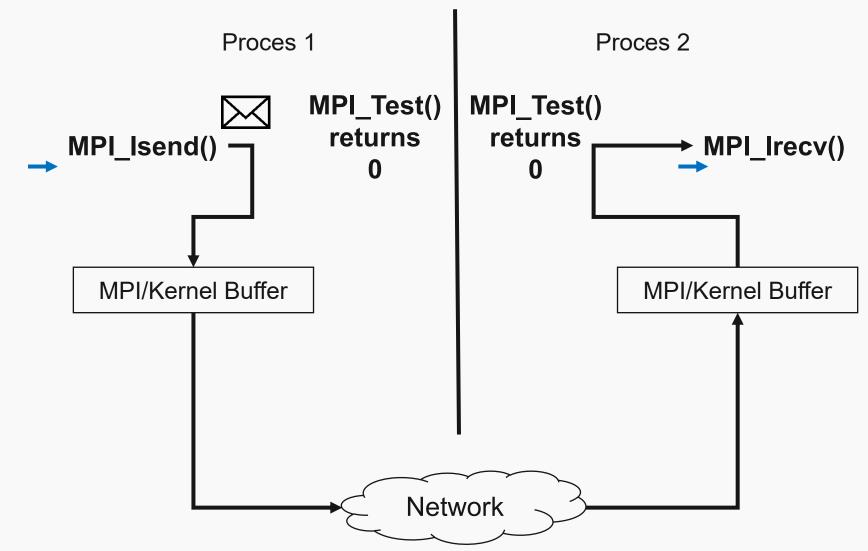




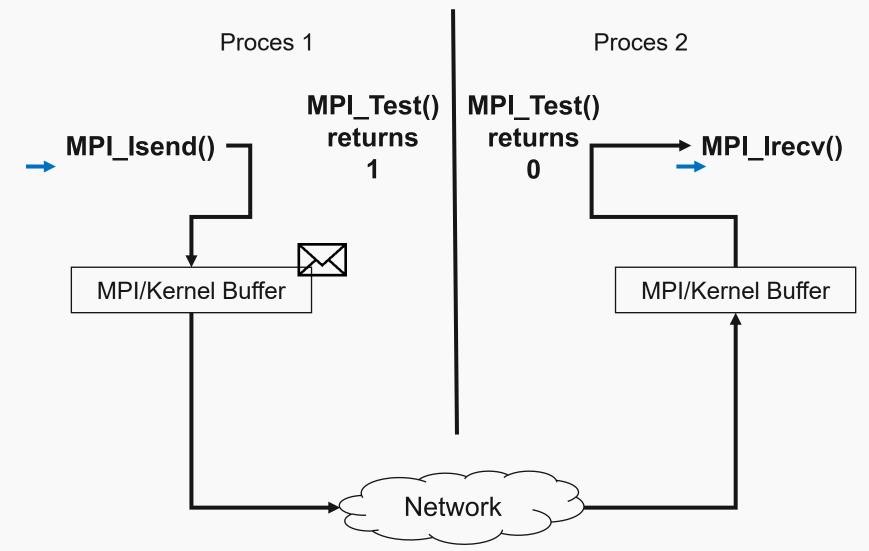




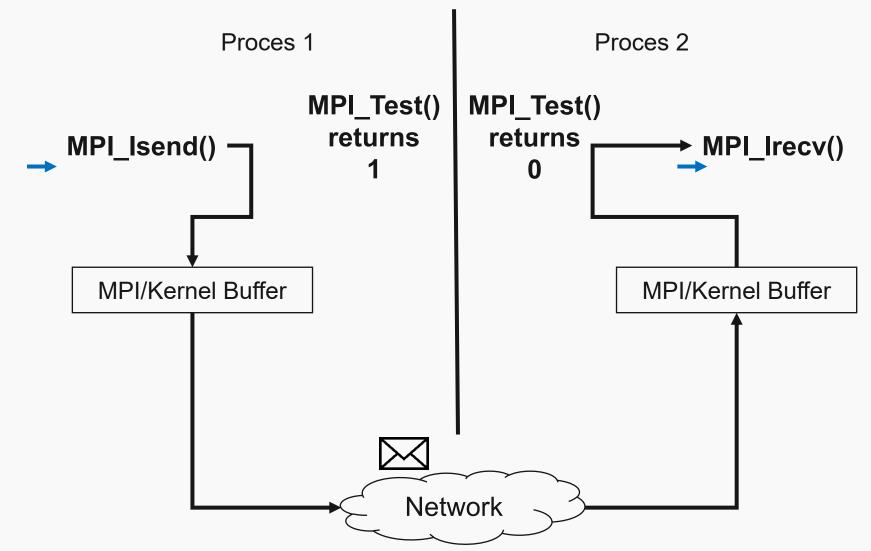




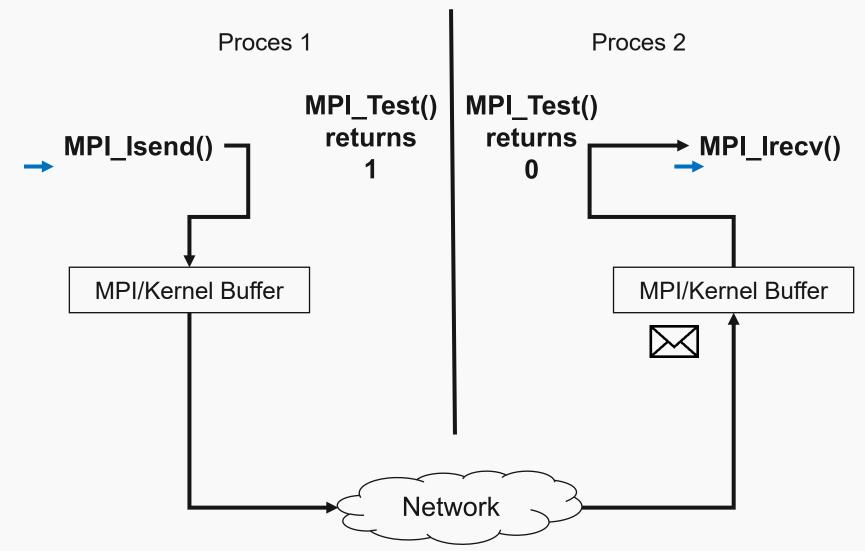




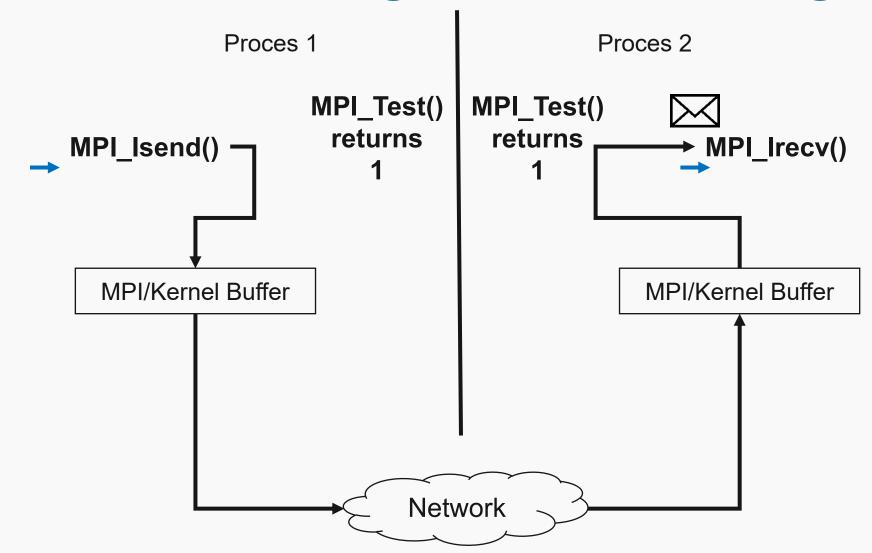














Comunicația non-blocantă

```
int MPI Isend( ↓ void *b, ↓ int c, ↓ MPI Datatype d, ↓ int recv, ↓ int t, ↓ MPI Comm, ↑ MPI Request *)
int MPI_Recv( ↑ void *b, ↓ <mark>int c</mark>, ↓ MPI_Datatype d, ↓ int sender, ↓ <mark>int t</mark>, ↓ MPI_Comm, ↑ MPI_Request * )
        MPI Test(↓ MPI Request *, ↑ int * flag, ↑ MPI_Status *)
        MPI Testall()
        MPI Testany()
        MPI Testsome()
                                      MPI_Wait(↓ MPI_Request *, ↑ MPI_Status *)
                                      MPI Waitall()
                                      MPI Waitany()
                                      MPI Waitsome()
```





Calcul Complexitate



Modelul Foster

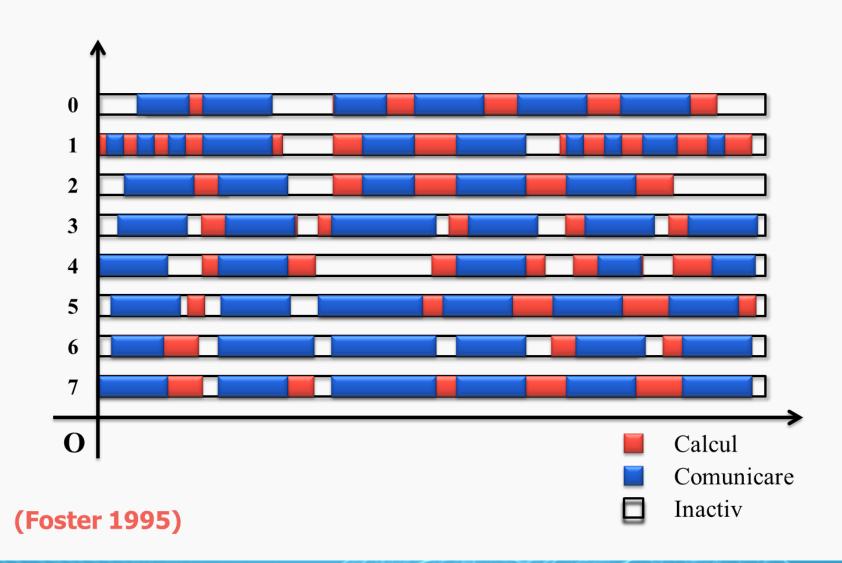
Designing and Building Parallel Programs

Ian Foster





Modelul Foster





Modelul Foster

Definiţie

 Timpul scurs de la începerea execuţiei primului proces până la terminarea execuţiei ultimului proces.

$$T = f(N, P, U, ...)$$
$$= T^{j}_{comp} + Tjcommu_{n} + Tjidle$$

Unde j un proces arbitrar. SAU

$$T = \left(\frac{1}{P}\right) * \left(\sum_{i=0}^{P-1} T_{comp}^{i} + \sum_{i=0}^{P-1} T_{commun}^{i} + \sum_{i=0}^{P-1} T_{idle}^{i}\right)$$

$$= \left(\frac{1}{P}\right) * (T_{comp} + T_{commun} + Tcommun)$$



LogP model

LogP: Towards a Realistic Model of Parallel Computation*

David Culler, Richard Karp, David Patterson, Abhijit Sahay, Klaus Erik Schauser, Eunice Santos, Ramesh Subramonian, and Thorsten von Eicken

> Computer Science Division, University of California, Berkeley

Abstract

A vast body of theoretical research has focused either on overly simplistic models of parallel computation, notably the PRAM, or overly specific models that have few representatives in the real world. Both kinds of models encourage exploitation of formal loopholes, rather than rewarding development of techniques that yield performance across a range of current and future parallel machines. This paper offers a new parallel machine model, called LogP, that reflects the critical technology trends underlying parallel computers. It is intended to serve as a basis for developing fast, portable parallel algorithms and to offer guidelines to machine designers. Such a model must strike a balance between detail and simplicity in order to reveal important bottlenecks without making analysis of interesting problems intractable. The model is based on four parameters that specify abstractly the computing bandwidth, the communication bandwidth, the communication delay, and the efficiency of coupling communication and computation. Portable parallel algorithms typically adapt to the machine configuration, in terms of these parameters. The utility of the model is demonstrated through examples that are implemented on the CM-5.



David Culler



David Patterson



LogP model

- L Limita superioară a **latenței (latency)** sau întârzierea de transmitere a unui mesaj de la sursă la destinație
- o **overhead**, durata de timp în care procesorul execută transmiterea sau recepţia fiecărui mesaj; În acest timp procesorul nu poate efectua alte operaţii
- g gap, intervalul minim de timp între două transmiteri succesive sau două recepţii succesive la acelaşi procesor. Reciproca lui g este echibalentă cu lungimea de bandă (bandwidth)
- P numărul de module **procesor / memorie**. Presupunem că funcționează la aceeași unitate de timp, numită ciclu.





Unele probleme nu pot fi paralelizate/distribuite

Calculating the hash of a hash of a hash ...of a string.

Deep First Search

Huffman decoding

Outer loops of most simulations

P complete problems



Paralelizare prin împărțirea problemei

Sunt o serie de probleme care sunt extrem de uşor de paralelizat/distribuit.

Embarrassingly parallel



Multiplicare unui vector cu un scalar

9 6 9 4 2 7 6 5 6 1

* 3

27 18 27 12 6 21 18 15 18 3



Toate calculele pot fi efectuate în același timp

* 3

27 18 27 12 6 21 18 15 18 3



Câte elemente sunt?





Câte elemente sunt?



. 1



Câte elemente sunt?

969427656 ...



Dar câte elemente de procesare?



. 1



Dar câte elemente de procesare?



. 1



Dar câte procese?



1



Dar câte procese?



. |



Cum este P față de N?



1



P << N

9 6 9 4 2 7 6 5 6

. .



Caz concret: P = 2 Cum împărțim?

```
9 6 9 4 2 7 6 5 6
```

. 1



Caz concret: P = 2 Cum împărțim?

9 6 9 4 2 7 6 5 6

1

Proces 1



Caz concret: P = 2 Cum împărțim?

9 6 9 4 2 7 6 5 6

. 1

Proces 1



Caz concret: P = 2 Cum împărțim?

9 6 9 4 2 7 6 5 6

. .

Proces 1



Caz concret: P = 2 Cum împărțim? Putem și random

9 6 9 4 2 7 6 5 6

1

Proces 1



Caz concret: P = 2 Cum împărțim?

9 6 9 4 2 7 6 5 6

.. 1

Proces 1



Caz concret: P = 2 Cum împărțim?

9 6 9 4 2 7 6 5 6

.. 1

Proces 1

Proces 2

Este utilă?



Caz concret: P = 2 Cum împărțim?

9 6 9 4 2 7 6 5 6

.. 1

Proces 1

Proces 2

Ce ne dorim?



Caz concret: P = 2 Cum împărțim?

.. 1

Proces 1

Proces 2

Ce ne dorim? Aproximativ același număr elemente



Aproximativ N/P elemente pe fiecare proces

9 6 9 4 2 7 6 5 6 ... 1

Proces 1



Aproximativ N/P elemente pe fiecare proces

9 6 9 4 2 7 6 5 6 ... 1

Dacă N nu se divide perfect la P?

Proces 1



Aproximativ N/P elemente Dacă N nu se divide perfect la P?

1

6

4 2 7

9 6 9

Proces 1

8

4 9 2

5 6 3



floor(N/P) elemente floor(15/2) = 7

1

6

4 2 7

9 6 9

Proces 1

8

4 9 2

5 6 3



ceil(N/P) elemente ceil(15/2) = 8

6 5

4 | 2 | 7

9 6 9

Proces 1

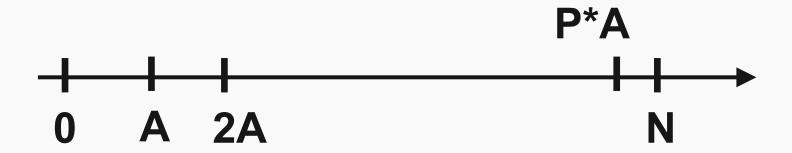
8 1

4 9 2

6 3

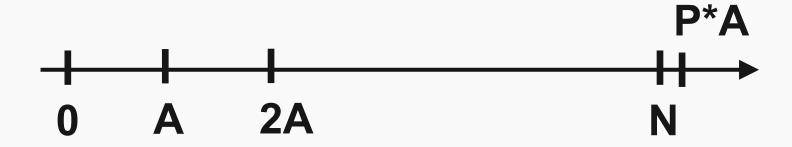


$$A = floor(N/P)$$





$$A = ceil(N/P)$$





Formule elegante:

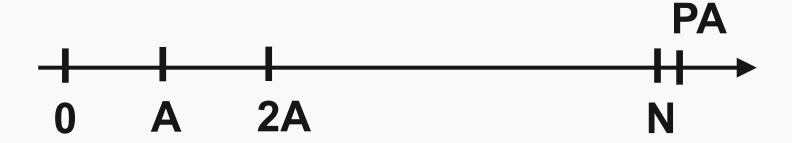
rank este identificator de proces, are valori de la 0 la P





Formule elegante:

rank este identificator de proces, are valori de la 0 la P



```
Funcţionează şi:

start = round(rank * N / P)

end = round((rank+1) * N / P) De ce?
```





Măsurarea timpului de execuție

time mpirun -n NUM_PROC ./executabil p a r a m e t r i



time sleep 5

real 0m5.001s user 0m0.000s sys 0m0.001s

time sleep 5

sleep 5 0.00s user 0.00s system 0% cpu 5.002 total

/usr/bin/time sleep 5

0.00user 0.00system 0:05.00elapsed 0%CPU (0avgtext+0avgdata 2076maxresident)k 0inputs+0outputs (0major+73minor)pagefaults 0swaps



time sleep 5

real 0m5.001s

user 0m0.000s

sys 0m0.001s

Wall clock time – Timpul trecut de la pornirea programului – Pe acesta îl folosim

time sleep 5

sleep 5 0.00s user 0.00s system 0% cpu 5.002 total

/usr/bin/time sleep 5

0.00user 0.00system 0:05.00elapsed 0%CPU (0avgtext+0avgdata 2076maxresident)k 0inputs+0outputs (0major+73minor)pagefaults 0swaps



time sleep 2

real 0m2.0**21**s user 0m0.000s sys 0m0.000s time sleep 2

real 0m2.0**18**s user 0m0.000s sys 0m0.016s

Timpii măsurați nu sunt exacți. Pentru a măsura corect trebuie să facem medie a timpilor după mai multe rulări sau să considerăm doar timpi mari – peste o secundă.

time sleep 2

real 0m2.0**16**s user 0m0.000s sys 0m0.000s

time sleep 2

real 0m2.0**15**s user 0m0.000s sys 0m0.000s



time sleep 5

real 0m5.001s

user 0m0.000s

sys 0m0.001s

Suma timpului petrecut în user space pe fiecare core.

time sleep 5

sleep 5 0.00s user 0.00s system 0% cpu 5.002 total

/usr/bin/time sleep 5

0.00user 0.00system 0:05.00elapsed 0%CPU (0avgtext+0avgdata 2076maxresident)k 0inputs+0outputs (0major+73minor)pagefaults 0swaps



time sleep 5

real 0m5.001s

user 0m0.000s

sys 0m0.001s

Suma timpului petrecut în kernel pe fiecare core.

time sleep 5

sleep 5 0.00s user 0.00s system 0% cpu 5.002 total

/usr/bin/time sleep 5

0.00user **0.00system** 0:05.00elapsed 0%CPU (0avgtext+0avgdata 2076maxresident)k 0inputs+0outputs (0major+73minor)pagefaults 0swaps



Operațiile de I/O sunt executate de Kernel

```
time dd if=/dev/zero of=file.txt count=1024 bs=1 048576
1024+0 records in
1024+0 records out
```

1073741824 bytes (1.1 GB) copied, 9.4847 s, 113 MB/s

real 0m9.490s

user 0m0.000s

sys 0m0.992s



Măsurare timp cu sau fără I/O?



Performanța

- Timp de execuție
- Memorie ocupată
- Număr de procese (thread-uri)
- Scalabilitate
- Toleranță la defecte
- Cost



Măsuri

- T Timpul total necesar execuției programului paralel
- P Numărul de procesoare utilizate
- S Speedup

$$\Box S = \frac{G}{T}$$

G – Timp execuție cel mai rapid algoritm secvențial



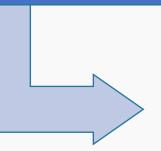
Workflow - Testarea programelor



 Test mici, rapide pentru a salva timp dacă sunt probleme majore

Stress test consistency

 Singura metodă "acceptabilă" de a confirma că programul nu are buguri ce apar rar



Measure time

 Pentru a determina că programul e scalabil şi întradevăr implementat în paralel