

Programação Paralela OPRP001

Programação com MPI

Prof. Guilherme Koslovski Prof. Maurício Pillon





Referências

- Cursos da ERAD
 - https://www2.sbc.org.br/erad/doku.php?id=start
- MPI Fórum
 - http://www.mpi-forum.org
- Open MPI Documentation
 - https://www.open-mpi.org/doc/current/



Agenda

Programação Paralela com Troca de Mensagens

MPI

Exemplos

Considerações Finais



Programação Paralela com Troca de Mensagens

- Opções de Programação
 - Linguagem de programação paralela (específica)
- Occam (Transputer)
 - Extensão de linguagens de programação existentes
- CC++ (Extensão de C++)
- Fortran M



Programação Paralela com Troca de Mensagens

- Geração automática usando anotações em Código e compilação (FORTRAN)
 - Linguagem padrão com biblioteca para troca de mensagens
- MPI (Message Passing Interface)
- PVM (Parallel Virtual Machine)



Programação Paralela com Troca de Mensagens

- Geração automática usando anotações em Código e compilação (FORTRAN)
 - Linguagem padrão com biblioteca para troca de mensagens
- MPI (Message Passing Interface)
- PVM (Parallel Virtual Machine)



Linguagem padrão com troca de mensagens

- Descrição explícita do paralelismo e troca de mensagens entre os processos
- Métodos Principais
- Criação de processos para execução em diferentes computadores
- Troca de mensagens (send/recv) entre processos
- Sincronização entre os processos



Modelos

- SPMD Single Program Multiple Data
 - Existe somente um programa
 - Um único programa executa em diversos hosts sobre um conjunto de dados distinto
- MPMD Multiple Program Multiple Data
 - Existem diversos programas
 - Programas diferentes executam em hosts distintos
 - Cada máquina possui um programa e um conjunto de dados distintos



Criação de processos

Criação Estática

- Os processos são especificados antes da execução
- Número fixo de processos
- Modelo SPMD é o mais comum

Criação Dinâmica

- Os processos são criados durante a execução (spawn)
- Encerramento dos processos é dinâmico
- Número variável de processos
- Modelo MPMD é o mais comum



Troca de mensagens

- Realizada através de primitivas send e receive
- Comunicação síncrona/bloqueante
 - Send bloqueia o emissor até o receptor executar receive
- Comunicação assíncrona/não bloqueante
 - Send n\u00e3o bloqueia o emissor
 - Receive pode ser utilizado durante a execução



Troca de mensagens

- Seleção de mensagens
- Filtro para receber uma mensagem de um determinado tipo (message tag) ou de um emissor específico
- Comunicação em Grupo
- Broadcast
- Gather/Scatter
 - Envio de partes de uma mensagem de um processo para diferentes processos de um grupo (distribuir), e recebimento de partes de mensagens de diversos processos de um grupo por um processo (coletar)

DO ESTADO DE

Sincronização

Barreiras

- Permitem especificar pontos de sincronia entre os processos
- Ao chegar na barreira, o processo fica esperando todos os demais processos do seu grupo chegarem na barreira.
- O ultimo processo libera todos os demais processos que estão bloqueados.



Agenda

Programação Paralela com Troca de Mensagens

MPI

Exemplos

Considerações Finais



Message Passing Interface (MPI)

- Padrão definido em 1994 pelo MPI Fórum
- Utiliza troca de mensagens entre os processos
- Versão Atual 3.1.2
- Implementações mais utilizadas
 - MPICH
 - IAMMPI
 - JAVA-MPI



Message Passing Interface (MPI)

- Ambiente
 - O processo principal inicia a execução
 - A execução ocorre em um conjunto de computadores pré-definidos
- Possui mais de 125 funções
- Compiladores
 - Mpicc (linguagem C)
 - Mpicc++ (linguagem C++)
 - Mpi77 (linguagem Fortran)



MPI – Diretivas Básicas

 MPI_INIT inicia um processo MPI, estabelecendo o ambiente e sincronizando os processos para iniciar a aplicação paralela.

int MPI_Init(int *argc, char* argv[])

Array de argumentos

Ponteiro com o número de argumentos

- O MPI deve ser inicializado uma única vez (não realizer chamdas subsequentes de MPI_Init ou MPI_Init_Thread)
- MPI_FINALIZE encerra o um processo MPI. Utilizado para sincronizar os processos para o término da aplicação paralela.

Int MPI_Finalize()



MPI – Exemplo

```
#include <mpi.h>
#include <stdio.h>
int main(int argc, char **argv){
        int rank, size;
        MPI_Init(&argc, &argv);
        MPI_Comm_rank(MPI_COMM_WORLD, &rank);
        MPI_Comm_size(MPI_COMM_WORLD, &size);
        printf("Hello World! I'm %d of %d\n",rank,size);
        MPI_Finalize();
        return 0;
```



MPI – Diretivas Básicas

 MPI_COMM_SIZE retorna o número de processos dentro de determinado grupo.

int MPI_Comm_size(MPI_Comm comm, int *size)

Tamanho a ser atualizado

Comunicador a ser analisado

 Caso o comunicador seja o MPI_COMM_WORLD o MPI_COMM_SIZE retorna quantidade total de processos.



MPI – Diretivas Básicas

 MPI_COMM_RANK retorna o rank do processo em determinado comunicador.

int MPI_Comm_rank(MPI_Comm comm, int *rank)

Rank a ser atualizado

Comunicador a ser analisado

- Utilizado amplamente em programas estilo mestre escravo
 - O processo com rank 0 é o mestre e os demais processos escravos



MPI – Exemplo

```
#include <mpi.h>
#include <stdio.h>
int main(int argc, char **argv){
   int rank, size;
   MPI_Init(&argc, &argv);
   MPI_Comm_rank(MPI_COMM_WORLD, &rank);
   MPI_Comm_size(MPI_COMM_WORLD, &size);
   printf("Processo %d de %d\n",rank,size);
   if(rank == 0)
      printf("(%d) -> Primeiro a escrever!\n",rank);
     MPI_Barrier(MPI_COMM_WORLD);
  }else{
     MPI_Barrier(MPI_COMM_WORLD);
      printf("(%d) -> Agora posso escrever!\n",rank);
   MPI_Finalize();
   return 0:
```



MPI – Diretivas Básicas

 MPI_COMM_SIZE retorna o número de processos dentro de determinado grupo.

int MPI_Comm_size(MPI_Comm comm, int *size)

Tamanho a ser atualizado

Comunicador a ser analisado

 Caso o comunicador seja o MPI_COMM_WORLD o MPI_COMM_SIZE retorna quantidade total de processos.



MPI – Exemplo

```
#include <mpi.h>
#include <stdio.h>
int main(int argc, char **argv){
   int rank, size;
   int tag=0;
   MPI_Status status;
   char msg[20];
   MPI_Init(&argc, &argv);
   MPI_Comm_rank(MPI_COMM_WORLD, &rank);
   MPI_Comm_size(MPI_COMM_WORLD, &size);
   if(rank == 0) 
     strcpy(msg,"Hello World!\n");
     for(i=1;i<size;i++) {
         printf("0 enviando 20 para %d\n", i);
         MPI_Send(msg,20,MPI_CHAR,i,tag,
         MPI_COMM_WORLD);
```

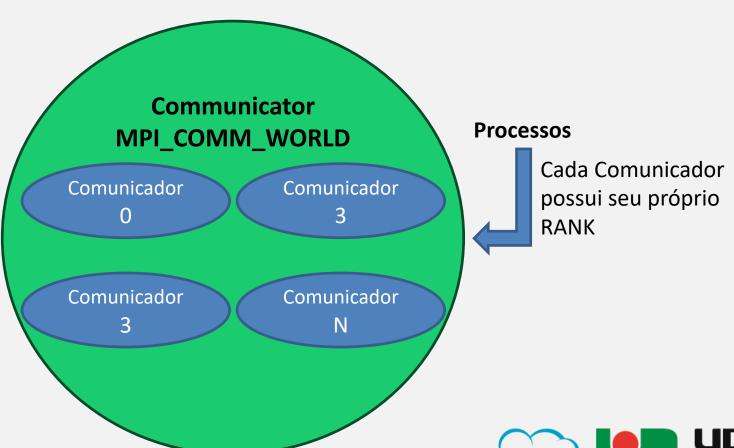
```
else{
      printf("%d esta
            esperando\n", rank);
MPI_Recv(msg, 20, MPI_CHAR, 0,
          tag,
          MPI COMM WORLD,
          &status);
      printf("Message received:
            %s\n", msg);
   MPI_Finalize();
   return 0;
```







MPI - Componentes





MPI – Tipos de Dados

- Dados do tipo MPI_Datatype
 - MPI_CHAR
 - MPI_DOUBLE
 - MPI_FLOAT
 - MPI_INT
 - MPI_LONG
 - MPI_LONG_DOUBLE

- MPI_SHORT
- MPI_UNSIGNED_CHAR
- MPI_UNSIGNED
- MPI_UNSIGNED_LONG
- MPI_UNSIGNED_SHORT



MPI – Comunicação

Basicamente as mensagens possuem a seguinte configuração



MPI – Funções Bloqueantes

Quantidade de elementos enviados

Endereço inicial do buffer

• MPI_Send(void* buf, int count, MPI_Datatype datatype, int dest, int tag, MPPI_Comm comm)

Rank do processo de destino



MPI – Funções Bloqueantes

Quantidade de
elementos
enviados

Tipo de dados dos
elementos no buffer

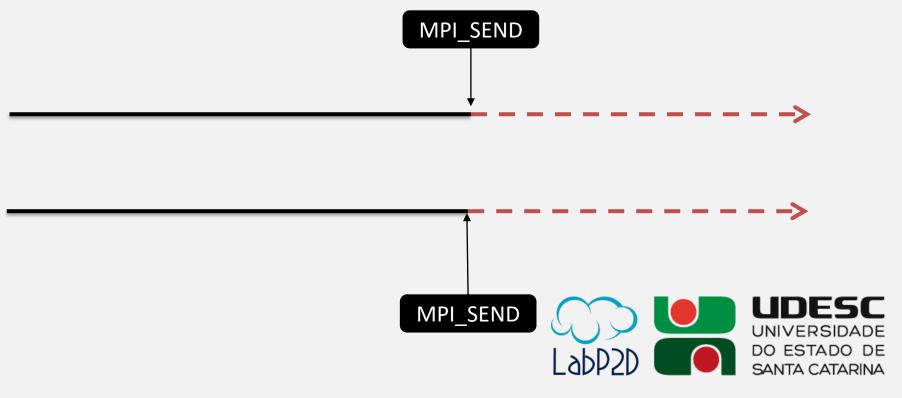
MPI_Recv(void* buf, int count, MPI_Datatype datatype, int source,
int tag, MPPI_Comm comm)

Rank do processo
de origem da
mensagem



MPI – Funções Bloqueantes DEADLOCK

 Este fenômeno ocorre quando todos os processos estão aguardando eventos que não foram iniciados



MPI – Funções Não Bloqueantes

Quantidade de elementos enviados

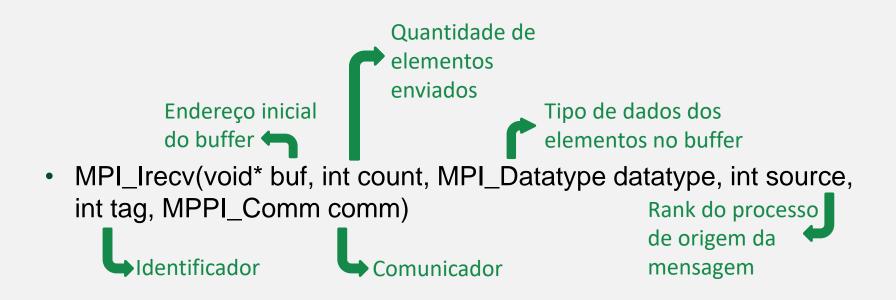
Endereço inicial do buffer

MPI_Isend(const void* buf, int count, MPI_Datatype datatype, int dest, int tag, MPPI_Comm comm)

Rank do processo de destino

SANTA CATARINA

MPI – Funções Não Bloqueantes





MPI – Exemplo

```
#include <stdio.h>
#include <mpi.h>
#include <string.h>
int main(int argc, char **argv){
   int rank, size, i;
   int tag=0;
   MPI Status status;
   char msg[20];
   MPI_Init(&argc, &argv);
   MPI_Comm_rank(MPI_COMM_WORLD, &rank);
   MPI Comm size(MPI COMM WORLD, &size);
   if(rank == 0) {
      strcpy(msg,"Hello World!\n");
      for(i=1;i<size;i++)
```

```
MPI_Send(msg, 13,
MPI_CHAR,i,tag,
MPI_COMM_WORLD);
  } else {
    MPI_Recv(msg,20,
MPI_CHAR,0,tag,
MPI_COMM_WORLD, &status);
  printf("Message received:
  %s\n", msg);
MPI Finalize();
return 0;
```







MPI – Exercício 01

- Construa um programa com comunicação em anel utilizando MPI
 - Cada processo deve receber uma mensagem do processo anterior (o rank)
 - Some seu rank com o valor da mensagem recebida
 - Envie o valor atualizado para o processo seguinte do anel
- O processo com rank 0 inicia com token 0
- Utilize as funçõe bloqueante MPI_Send() e MPI_Recv()
- Realize testes com 4,8,12,16 hosts



MPI - Exercício 02

- Crie um programa em MPI que realize um somatório do produtório das linhas de uma matriz M[10][30]
 - O master envia uma linha da matriz para os processos filhos
 - Os slaves realizam a multiplicação da linha da matriz
 - Ao terminar, o slave envia seu resultado ao master
 - O master realiza a soma dos resultados obtidos pelos slaves
 - OBS: Tomar cuidado para não iniciar um slave que não irá receber uma linha da matriz (rank_slave < qnt_linhas)



MPI - Exercício 03

- Implementar em MPMD o exercício 02
- A partir do Código do exercício 02, separe em dois arquivos as funções do master e slave.
 - O primeiro código executa o envio dos dados, coordenação entre os processos e o somatório final.
 - O Segundo código recebe os dados, processa a tarefa de multiplicação de matriz e devolve os resultados ao processo pai.
 - Será necessário incluir as funções de finilização do MPI.
- Para executar
 - Mpirun –n 1 ./master_code : -n qnt_filhos ./slave_code



MPI – Funções Não Bloqueantes

 Espera a requisição de envio ou recebimento de mensagem seja completada

Requisição de send ou recv



MPI_Wait(MPI_Request *request, MPI_Status *status)



MPI – Funções Não Bloqueantes

Testa se o envio ou recebimento de mensagem foi completada

Requisição de send ou recv

Status da requisição

True se a requisição

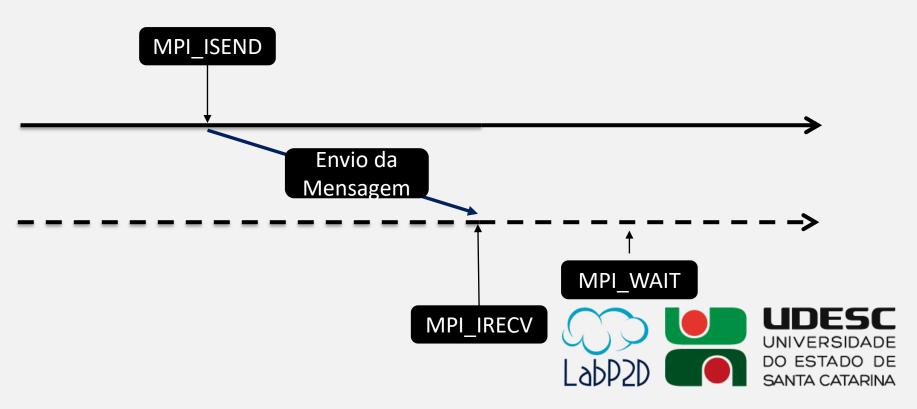
foi completada, falso
caso contrario

MPI_Test(MPI_Request *request, int flag, MPI_Status ** status)



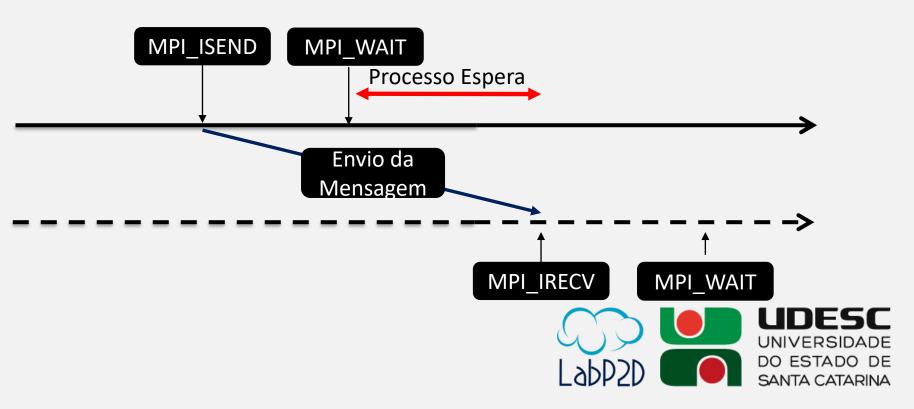
MPI – Funções Não Bloqueantes DEADLOCK (?)

 O processo n\u00e3o espera que o recebimento da mensagem esteja concluído.



MPI – Funções Não Bloqueantes DEADLOCK (?)

 O processo espera que o recebimento da mensagem esteja concluído.



MPI – Barreiras

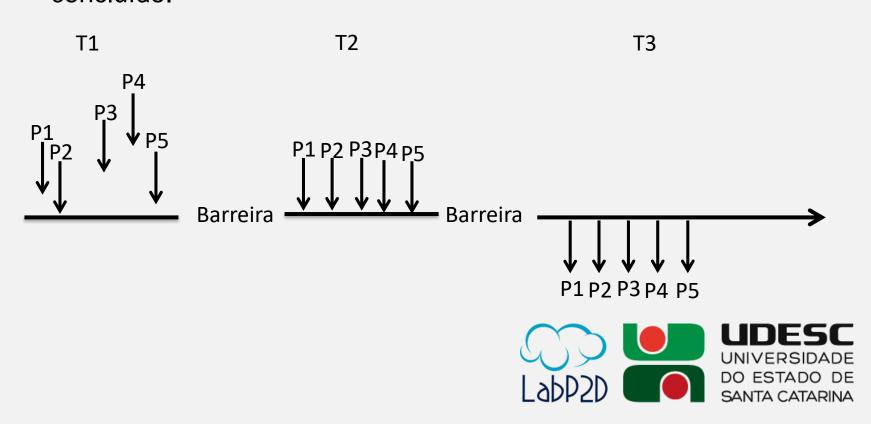
- A função MPI_BARRIER realiza a sincronização explícita de todos os processos de um determinado comunicador/grupo
- O processos que utilize MPI_BARRIER para de executar até que todos os membros de seu grupo também executem o MPI_BARRIER





MPI – Barreiras

 O processo espera que o recebimento da mensagem esteja concluído.



MPI – Exemplo

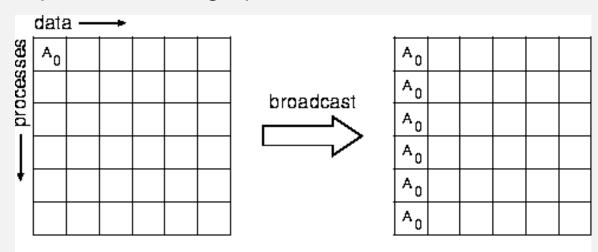
```
#include <stdio.h>
#include <mpi.h>
#include <string.h>
int main(int argc, char **argv){
         int rank, size, i;
         MPI_Init(&argc, &argv);
         MPI_Comm_rank(MPI_COMM_WORLD, &rank);
         MPI_Comm_size(MPI_COMM_WORLD, &size);
         printf("I'm %d of %d\n",rank,size);
         if(rank == 0) {
                   printf("(%d) -> Primeiro a escrever!\n",rank);
                   MPI_Barrier(MPI_COMM_WORLD);
         }else{
                   MPI_Barrier(MPI_COMM_WORLD);
                   printf("(%d) -> Agora posso escrever!\n",rank);
         MPI_Finalize();
         return 0;
                                                                           UNIVERSIDADE
                                                                           DO ESTADO DE
                                                                           SANTA CATARINA
```

MPI – Comunicação em grupo Broadcast

 A função MPI_Bcast permite o processo enviar dados para todos os memberos do grupo

Todos os processos do grupo devem utilizer o mesmo comm e

root









MPI – Comunicação em grupo Broadcast

Endereço inicial do buffer Comunicador/grupo

Quantidade de elementos elementos no buffer

Tipo de dados dos elementos no buffer

Tipo de dados dos elementos no buffer

Tipo de dados dos elementos no buffer

Rank do emissor do broadcast



MPI – Comunicação em grupo Scatter e Gather

Scatter

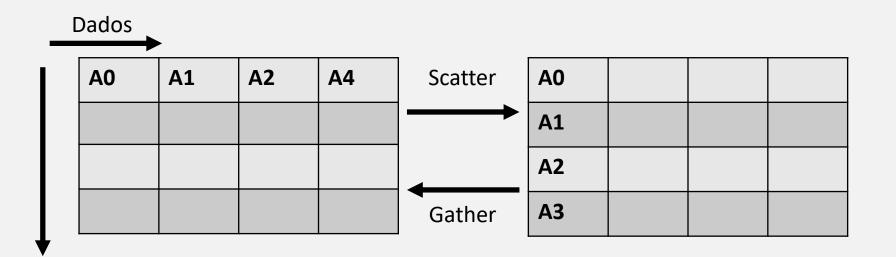
- Um processo necessita distribuir dados em n segmentos iguais.
- O segmento N é enviado para o processo N

Gather

Um processo necessita coletar dados de n processos do grupo



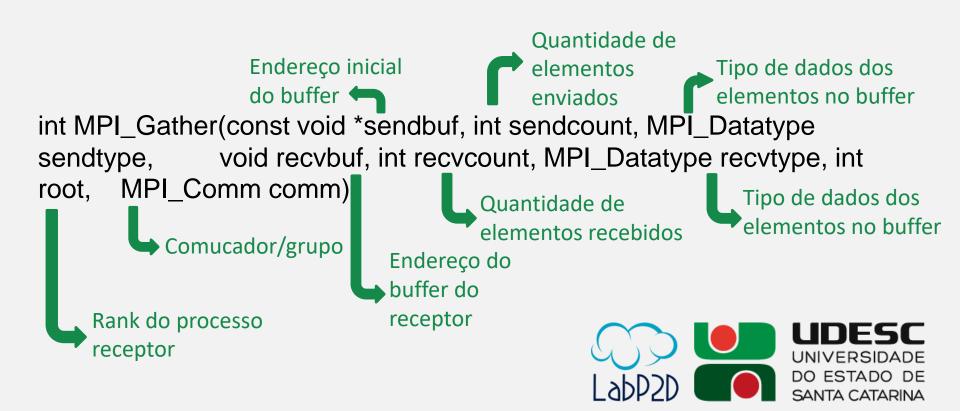
MPI – Comunicação em grupo Scatter e Gather



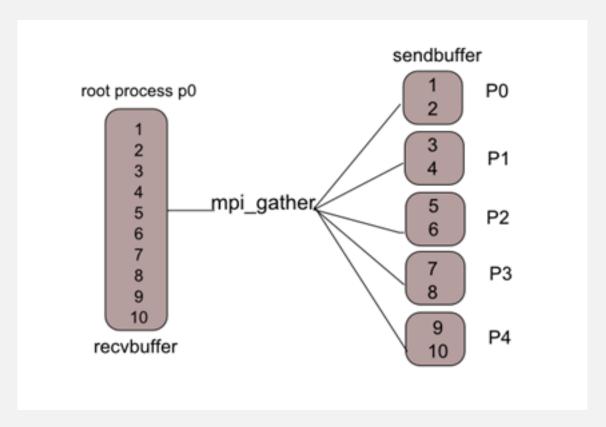
Processos



MPI – Comunicação em grupo Gather



MPI – Comunicação em grupo Gather





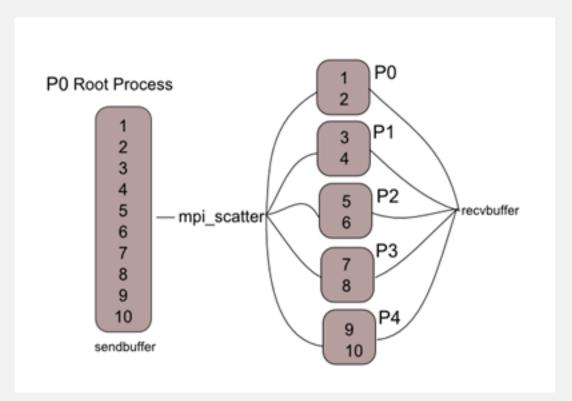
MPI – Exemplo

```
#include <stdio.h>
#include <mpi.h>
#include <string.h>
int main(int argc, char **argv){
          int sndbuffer, *recvbuffer, rank, size, i;
          MPI_Init(&argc, &argv);
          MPI_Comm_rank(MPI_COMM_WORLD, &rank);
          MPI Comm size(MPI COMM WORLD, &size);
          recvbuffer = (int *)malloc(size*sizeof(int));
          sndbuffer = rand*rank;
          MPI_Gather(&sndbuffer, 1, MPI_INT, recvbuffer, 1, MPI_INT, 0,
                      MPI COMM WORLD):
          if(rank == 0)
                    printf("(%d) - Recebi vetor: ", rank);
                    for(i=0;i< size;i++)
                              printf("%d",recvbuffer[i]);
          MPI Finalize();
                                                                              UNIVERSIDADE
                                                                              DO ESTADO DE
          return 0;
                                                                              SANTA CATARINA
```

MPI – Comunicação em grupo Scatter



MPI – Comunicação em grupo Scatter







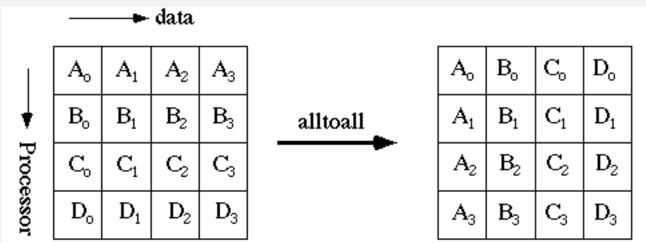


MPI – Exemplo

```
#include <stdio.h>
#include <mpi.h>
#include <string.h>
int main(int argc, char **argv){
          int *sndbuffer, recvbuffer, rank, size, i;
          MPI_Init(&argc, &argv);
          MPI_Comm_rank(MPI_COMM_WORLD, &rank);
          MPI Comm size(MPI COMM WORLD, &size);
          sendbuffer = (int *)malloc(size*sizeof(int));
          if(rank == 0) {
                   for(i=0): i<size: i++) sndbuffer[i] = i*i:
          MPI_Scatter(sndbuffer, 1, MPI_INT, &recvbuffer, 1, MPI_INT, 0,
                      MPI COMM WORLD);
          if(rank != 0)
                    printf("(%d) - Received %d\n", rank, recvbuffer);
          MPI_Finalize();
          return 0:
                                                                              UNIVERSIDADE
                                                                              DO ESTADO DE
                                                                              SANTA CATARINA
```

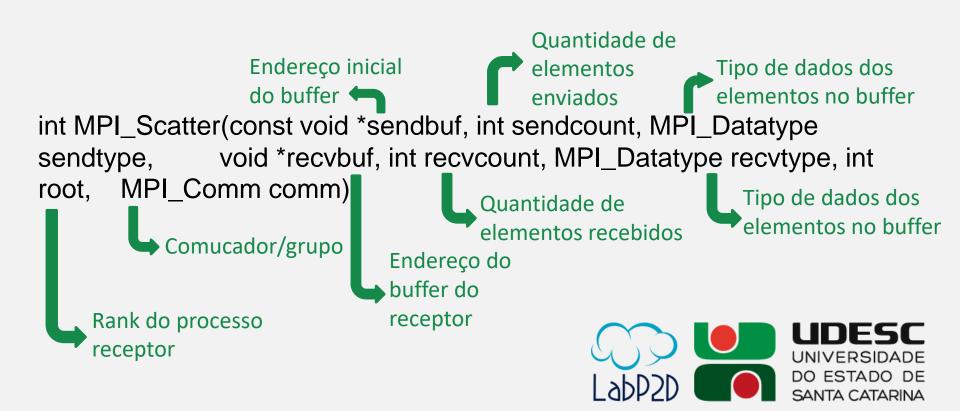
MPI – Comunicação em grupo MPI_Reduce

- Esta função faz com que todos os processos enviem e coletem dados de cada processo da aplicação.
- Cada processo efetua um Broadcast.





MPI – Comunicação em grupo MPI_AIIToAII



MPI – Exemplo

```
void printvector(int rank,
                                                                              int *buffer, int size){
#include <stdio.h>
#include <mpi.h>
                                                               int i;
#include <string.h>
                                                               printf("Vetor:\n");
                                                               for (i=0; i<size - 1; i++)
int main(int argc, char **argv){
                                                                  printf("%d\n", buffer[i]);
          int *sndbuffer, *recvbuffer, rank, size, i;
          MPI_Init(&argc, &argv);
          MPI_Comm_rank(MPI_COMM_WORLD, &rank);
          MPI Comm size(MPI COMM WORLD, &size);
          sendbuffer = (int *)malloc(size*sizeof(int));
          recvbuffer = (int *)malloc(size*sizeof(int));
          if(rank == 0)
                    for(i=0; i < size; i++) sndbuffer[i] = i*i;
          for(i=0; i < size; i++) sndbuffer[i] = i*i+rank;
          printvector(rank, sndbuffer, size);
          MPI_Alltoall(sndbuffer, 1, MPI_INT, recvbuffer, 1, MPI_INT,
                       MPI_COMM_WORLD);
          printvector(rank, sndbuffer, size);
                                                                                 UNIVERSIDADE
                                                                                 DO ESTADO DE
          MPI_Finalize();
                                                                                 SANTA CATARINA
          return 0;
```

MPI – Comunicação em grupo MPI_Reduce

- O resultado parcial de um processo em determinado grupo é combinado e retornado para um processo específico
- Utiliza um tipo de função de operação

Função	Resultado
MPI_MAX	Valor máximo
MPI_MIN	Valor mínimo
MPI_SUM	Somatório
MPI_PROD	produto



MPI – Comunicação em grupo MPI_Reduce

Endereço inicial do buffer do receptor int MPI_Alltoall(const void *sendbuf, void *recvbuf , int count, int sendcount, MPI_Datatype datatype, MPI_Op op, int root, MPI_Comm comm)

Tipo de dados dos elementos no buffer

Comucador/grupo

Endereço do buffer do receptor de elementos

Operação de reduce



Agenda

Programação Paralela com Troca de Mensagens

MPI

Exemplos

Considerações Finais



Exemplos

• Os exemplos estão disponíveis no moodle



Considerações Finais

pThreads, OpenMP e MPI são amplamente utilizados

 MPI define uma interface padrão para troca de mensagens entre processos distribuídos

Diversas bibliotecas implementam esta API

