

高级MPI编程技术

Lecture 09: MPI编程——集合通信

肖俊敏

中国科学院计算技术研究所

MPI内容目录

- 基本概念
- 点到点通信
- 自定义数据类型 
- 集合通信
- 虚拟拓扑
- 文件IO

MPI数据类型分类

- 预定义数据类型
- 自定义数据类型

MPI预定义数据类型

MPI预定义数据类型	相应的C数据类型
MPI_CHAR	signed char
MPI_SHORT	signed short int
MPI_INT	signed int
MPI_LONG	signed long int
MPI_UNSIGNED_CHAR	unsigned char
MPI_UNSIGNED_SHORT	unsigned short int
MPI_UNSIGNED	unsigned int
MPI_UNSIGNED_LONG	unsigned long int
MPI_FLOAT	float
MPI_DOUBLE	double
MPI_LONG_DOUBLE	long double
MPI_BYTE	无对应类型
MPI_PACKED	无对应类型

为什么自定义数据类型

- MPI 的消息收发函数**只能处理连续存储的同一类型的数据**.
- 不同系统有不同的数据表示格式。MPI预先定义一些基本数据类型，在实现过程中在这些基本数据类型为桥梁进行转换。
- 派生数据类型:允许消息来自不连续或类型不一致的存储区域，如数组散元与结构类型等的传送。
- 它的使用可有效地减少消息传递的次数, 增大通信粒度, 并且在收/发消息时避免或减少数据在内存中的拷贝、复制。

数据类型的定义

■ MPI 数据类型由两个 n 元序列构成, n 为正整数.

- 第一个序列包含一组数据类型, 称为类型序列 (type signature):

Typesig = {**type**₀, **type**₁, . . . , **type** _{$n-1$} }.

- 第二个序列包含一组整数位移, 称为位移序列 (type displacements):

Typedisp = {**disp**₀, **disp**₁, . . . , **disp** _{$n-1$} }.

位移序列中位移总是以字节为单位计算的

数据类型的定义(续)

- 构成类型序列的数据类型称为基本数据类型, 它们可以是**原始数据类型**, 也可以是任何**已定义的数据类型**.
- 因此MPI的数据类型是嵌套定义的. 为了以后叙述方便, 我们称非原始数据类型为复合数据类型.

类型图(type map)

- 类型序列刻划了数据的类型特征，位移序列则刻划了数据的位置特征。类型序列和位移序列元素的一一配对构成序列的类型图。

Typemap = {(type₀, disp₀), (type₁, disp₁), ... , (type_{*n*-1}, disp_{*n*-1})}.

- 假设数据缓冲区的起始地址为buff₀, 则由上述类型图所定义的数据类型包含 *n* 块数据, 第 *i* 块数据的地址为
buff₀ + disp_{*i*}, 类型为type_{*i*}; *i* = 0, 1, ... , *n*-1.

类型图(type map)(续)

- MPI 的原始数据类型的类型图可以写成{(类型, 0)}. 如 MPI_INTEGER 的类型图为{(INTEGER, 0)}.
- 位移序列中的位移不必是单调上升的, 表明数据类型中的数据块不要求按顺序排放. 位移也可以是负的, 即数据类型中的数据可以位于缓冲区起始地址之前.

类型图的表示

类型图 = {

< 基类型, 偏移 >

< 基类型, 偏移 >

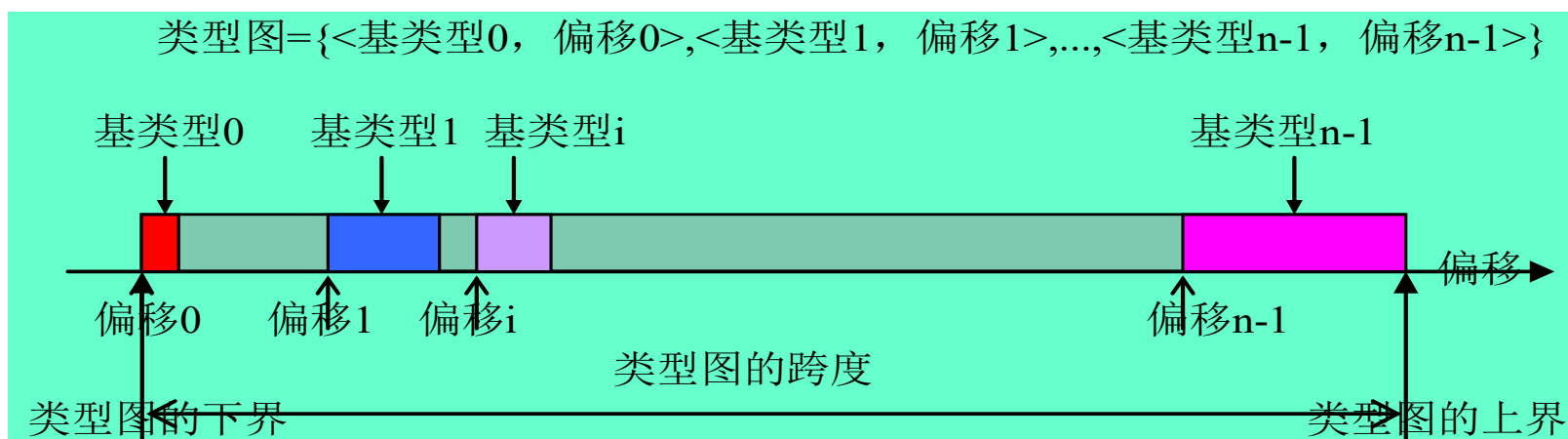
< 基类型, 偏移 >

...

< 基类型, 偏移 >

}

类型图的图示



例1

- 假设数据类型TYPE 类型图为：
{(integer, 4), (integer, 12), (integer, 0)}

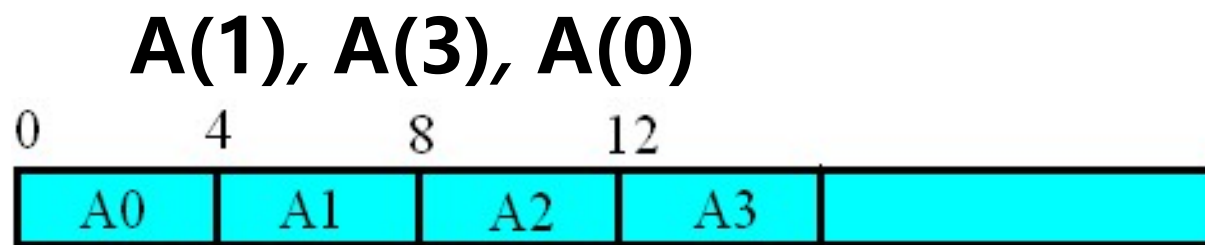
则语句：

```
int A(100)
```

```
... ..
```

```
MPI_Send(A, 1, TYPE, ...)
```

将发送？



数据类型的大小

- 指该数据类型中包含的数据长度(字节数), 它等于类型序列中所有基本数据类型的大小之和。 数据类型的大小就是消息传递时需要发送或接收的数据长度。
- 假设数据类型type的类型图为:
 $\{(\text{type0}, \text{disp0}), (\text{type1}, \text{disp1}), \dots, (\text{typen}-1, \text{dispn}-1)\}$
则该数据类型的大小为:

$$\text{Sizeof}(\text{type}) = \sum_{i=0}^{n-1} \text{sizeof}(\text{type}_i)$$

下界、上界与域

- **下界(lower bound)**: 数据的最小位移
- **上界(upper bound)**: 数据的最大位移加1, 再加上一个使得数据类型满足操作系统地址对界要求(alignment)的修正量 ε .
- **域(extent)**: 上界与下界之差

数据类型的对界量

- 原始数据类型的对界量由编译系统决定
- 复合数据类型的对界量则定义为它的所有基本数据类型对界量的最大值
- 地址对界要求一个数据类型在内存中的(字节) 地址必须是它的对界量的整数倍.

C语言中的对界 (例2)

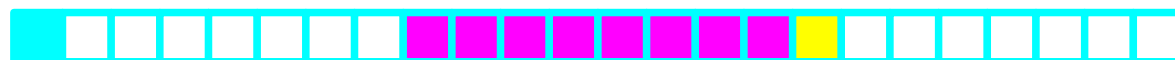
```
typedef struct
{
    char a;
    double b;
    int c;
} T;
```

```
main(){
    T m;
    printf( "sizeof(T)=%d,sizeof(m)=%d\n",
            sizeof(T), sizeof(m) );
    printf( "m.a=%d, m.b=%d, m.c=%d\n",
            (char *)&m.a - (char *)&m,
            (char *)&m.b - (char *)&m,
            (char *)&m.c - (char *)&m
    );
}
```


C语言中的对界

```
zf@loginNode:~  
zf@gnode1:~> ./a.out  
sizeof(T)=24, sizeof(m)=24  
m.a=0, m.b=8, m.c=16  
zf@gnode1:~> 
```

struct(char, double, int)



char

double

int

MPI LB和MPI UB

- MPI提供了两个特殊数据类型MPI_LB 和MPI_UB, 称为伪数据类型(pseudo datatype). 它们的大小是0, 作用是让用户人工指定一个数据类型的上下界.
- MPI 规定: 如果一个数据类型的基本类型中含有MPI_LB, 则它的下界定义为:

$$lb(type) = \min_i \{disp_i \mid type_i = MPI_LB\}$$

- 如果一个数据类型的基本类型中含有MPI_UB, 则它的上界定义为:

$$ub(type) = \max_i \{disp_i \mid type_i = MPI_UB\}$$

例3

- 类型图 $\{(\text{MPI_LB}, -4), (\text{MPI_UB}, 20), (\text{MPI_DOUBLE}, 0), (\text{MPI_INTEGER}, 8), (\text{MPI_BYTE}, 12)\}$

问题：下界为 -4，上界为 20，域为 24。

数据类型查询函数

■ 查询指定数据类型的大小:

```
int MPI_Type_size (  
    MPI_Datatype Datatype      /* in */,  
    int*          size         /*out*/)

```

■ 查询指定数据类型的域:

```
int MPI_Type_extent (  
    MPI_Datatype Datatype      /* in */,  
    MPI_Aint*    extent        /*out*/)

```

数据类型查询函数

■ 查询指定数据类型的上界:

```
int MPI_Type_ub (  
    MPI_Datatype      Datatype      /* in */,  
    MPI_Aint*         displacement  /*out*/)

```

■ 查询指定数据类型的下界:

```
int MPI_Type_lb (  
    MPI_Datatype      Datatype      /* in */,  
    MPI_Aint*         displacement  /*out*/)

```

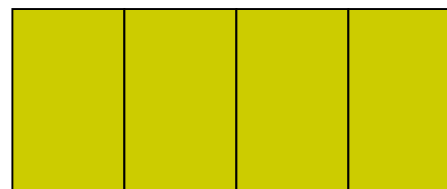
MPI自定义数据类型

- 连续数据类型
- 向量数据类型
- 索引数据类型
- 结构数据类型

连续数据类型的创建

```
int MPI_Type_contiguous(  
    int                count        /* in */,  
    MPI_Datatype        oldtype     /* in */,  
    MPI_Datatype*       newtype     /*out*/)
```

同一类型的多次重复



作用:将连续的基类型重复作为一个整体看待

新类型的递交和释放

■ 提交:

int **MPI_Type_commit**(MPI_Datatype* datatype)

- 将数据类型映射进行转换或“编译”
- 一种数据类型变量可反复定义，连续提交

■ 释放:

int **MPI_Type_free**(MPI_Datatype* datatype)

- 将数据类型设为MPI_DATATYPE_NULL

定义矩阵的一行(例4)

■ 在C中定义矩阵的一行:

```
float a[4][4];  MPI_Datatype C_R;
```

```
MPI_Comm_dup (MPI_COMM_WORLD, &comm);
```

```
MPI_Type_contiguous (4,MPI_FLOAT,&C_R);
```

```
MPI_Type_commit (&C_R);
```

```
MPI_Send(&(a[2][0]),1,C_R, right, tag, comm);
```

```
MPI_Recv(&b[i][j],4,MPI_FLOAT,.....);
```

定义矩阵的一行图示

```
count = 4;  
MPI_Type_contiguous(count, MPI_FLOAT, &rowtype);
```

1.0	2.0	3.0	4.0
5.0	6.0	7.0	8.0
9.0	10.0	11.0	12.0
13.0	14.0	15.0	16.0

`a[4][4]`

?

如何发送一行?

```
MPI_Send(&a[2][0], 1, rowtype, dest, tag, comm);
```

9.0	10.0	11.0	12.0
-----	------	------	------

1 element of
rowtype

MPI自定义数据类型

- 连续数据类型
- **向量数据类型**
- 索引数据类型
- 结构数据类型

向量数据类型的生成

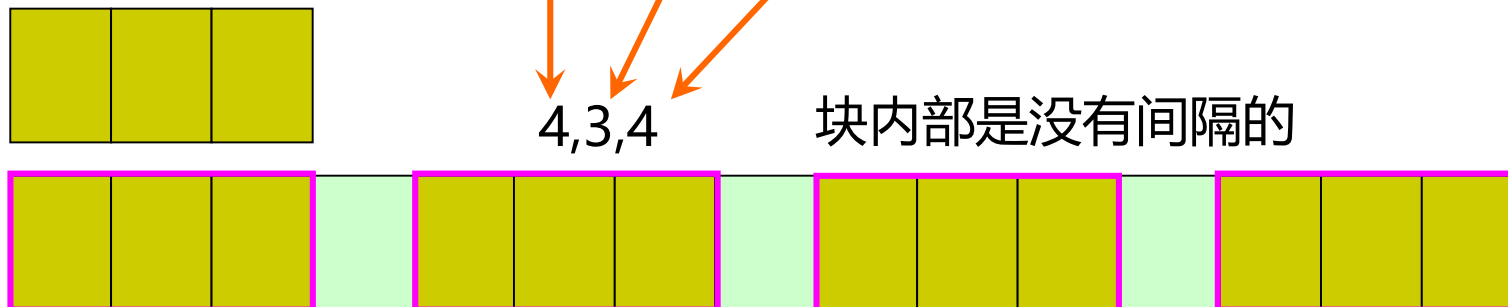
```
int MPI_Type_vector(  
    int      count          /* in */,  
    int      blocklen       /* in */,  
    int      stride         /* in */,  
    MPI_Datatype oldtype     /* in */,  
    MPI_Datatype* newtype    /*out*/)
```

- count 块数（非负整数）
- blocklength 每块中的元素个数（非负整数）
- stride每块开始间隔的**元素个数** (integer)

向量数据类型的生成

MPI_Type_vector(count, blocklength, stride, oldtype, newtype)

- 类型重复形成块，多个块按一定的**间隔**排列间隔可以是以本类型为单位，也可以是以字节为单位



向量数据类型的生成

```
int MPI_Type_hvector(  
    int                count        /* in */,  
    int                blocklen     /* in */,  
    int                stride       /* in */,  
    MPI_Datatype        oldtype     /* in */,  
    MPI_Datatype*      newtype      /*out*/)
```

- **MPI_Type_vector** 中 stride以oldtype 的域为单位
- **MPI_Type_hvector** 中stride以字节为单位

定义矩阵的一列(例5)

■ 在C中定义矩阵的一列

```
float a[4][4]; MPI_Datatype C_C;
```

```
MPI_Type_vector (4,1,4,MPI_FLOAT,&C_C);
```

```
MPI_Type_commit(&C_C);
```

```
MPI_SEND(&(a[0][1]),1,C_C, right, tag, comm)
```

定义矩阵的一系列图示

```
count = 4;  blocklength = 1;  stride = 4;  
MPI_Type_vector(count, blocklength, stride, MPI_FLOAT,  
                &columntype);
```

1.0	2.0	3.0	4.0
5.0	6.0	7.0	8.0
9.0	10.0	11.0	12.0
13.0	14.0	15.0	16.0

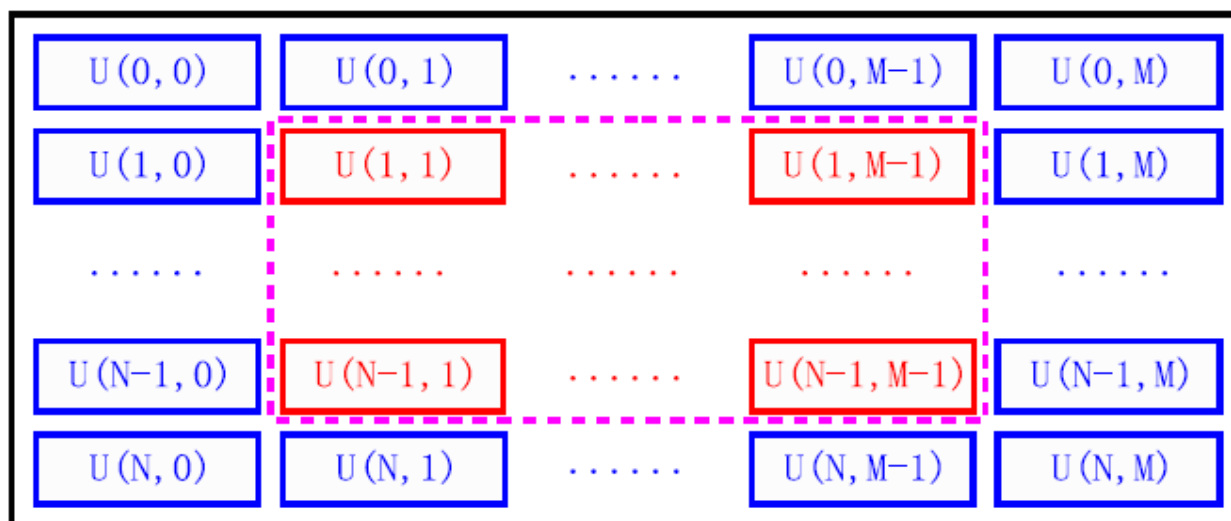
$a[4][4]$

```
MPI_Send(&a[0][1], 1, columntype, dest, tag, comm);
```

2.0	6.0	10.0	14.0
-----	-----	------	------

1 element of
columntype

思考



答案

```
int U[n+1][m+1]; MPI_Datatype N_T;
```

```
MPI_Type_vector (n-1,m-1,m+1,MPI_INT,&N_T);
```

```
MPI_Type_commit(&N_T);
```

```
MPI_SEND(&(U[1][1]),1,N_T, right, tag, comm);
```

MPI自定义数据类型

- 连续数据类型
- 向量数据类型
- **索引数据类型**
- 结构数据类型

索引数据类型

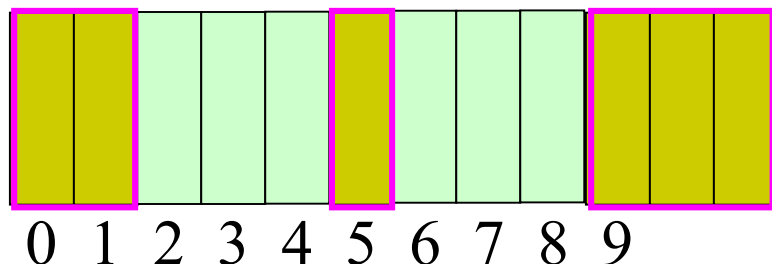
```
int MPI_Type_indexed(  
    int      count          /* in */,  
    int      blocklens[]    /* in */,  
    int      indices[]      /* in */,  
    MPI_Datatype oldtype    /* in */,  
    MPI_Datatype* newtype    /*out*/)
```

- **count** number of blocks -- also number of entries in indices and blocklens
- **blocklens** number of elements in each block
- **indices** displacement of each block in multiples of old_type

索引数据类型

- **MPI_TYPE_INDEXED**(count,array_of_blocklengths,array_of_displacements,oldtype,newtype)
- 重复形成块，不同的块放到不同的位置，位置的指定可以是以旧数据类型为单位

3,{2,1,3},{0,5,9}



索引数据类型

int **MPI_Type_hindexed**(

int count /* in */,

int blocklens[] /* in */,

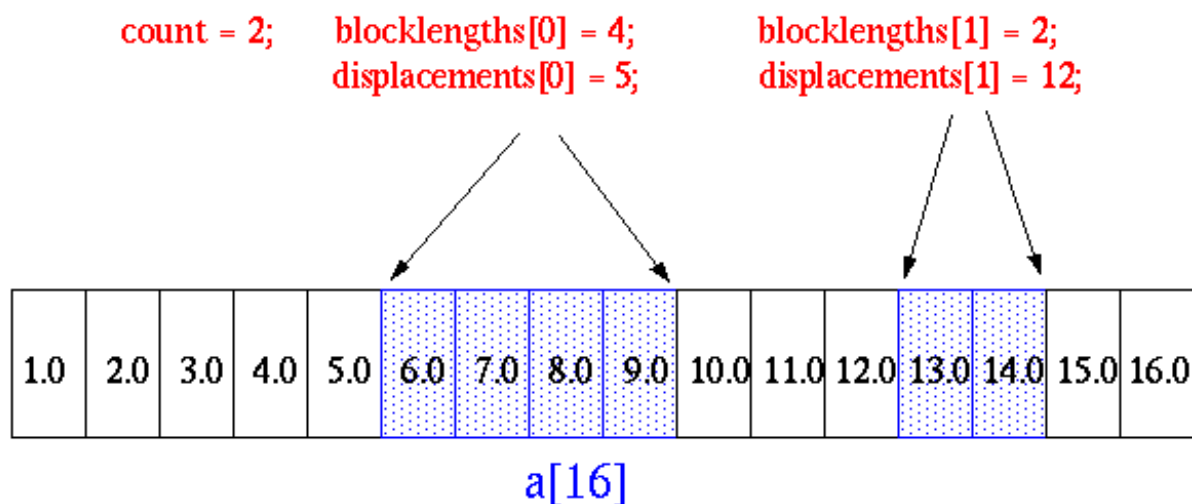
int indices[] /* in */,

MPI_Datatype oldtype /* in */,

MPI_Datatype* newtype /*out*/)

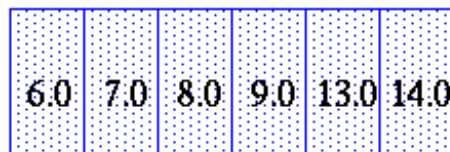
- **count** number of blocks -- also number of entries in indices and blocklens
- **blocklens** number of elements in each block
- **indices** displacement of each block in *bytes*

例6



```
MPI_Type_indexed(count, blocklengths, displacements, MPI_FLOAT, &indextype);
```

```
MPI_Send(&a, 1, indextype, dest, tag, comm);
```



1 element of
indextype

问题

- **MPI_type_vector()**与**MPI_type_indexed()**的区别?
- **MPI_Type_indexed** 与**MPI_Type_vector** 的区别在于每个数据块的长度可以不同, 数据块间也可以不等距.

MPI自定义数据类型

- 连续数据类型
- 向量数据类型
- 索引数据类型
- **结构数据类型**

结构数据类型

```
int MPI_Type_struct (
    int                count        /* in */,
    int                blocklens[]   /* in */,
    int                indices[]     /* in */,
    MPI_Datatype        types[]      /* in */,
    MPI_Datatype*       newtype      /*out*/ )
```

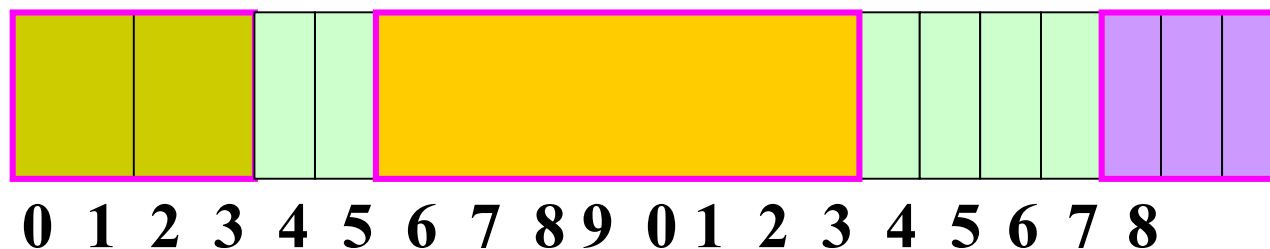
- **count** number of blocks (integer)
- **blocklens** number of elements in each block
- **indices** byte displacement of each block (array)
- **types** of elements in each block

结构数据类型

`MPI_Type_struct(count,array_of_blocklens,array_of_displacements, array_of_types, newtype)`

将多个不同的旧数据类型进行组合，而前面的数据类型生成方法都是对一个旧数据类型进行重复。

3,{2,1,3},{0,6,18}

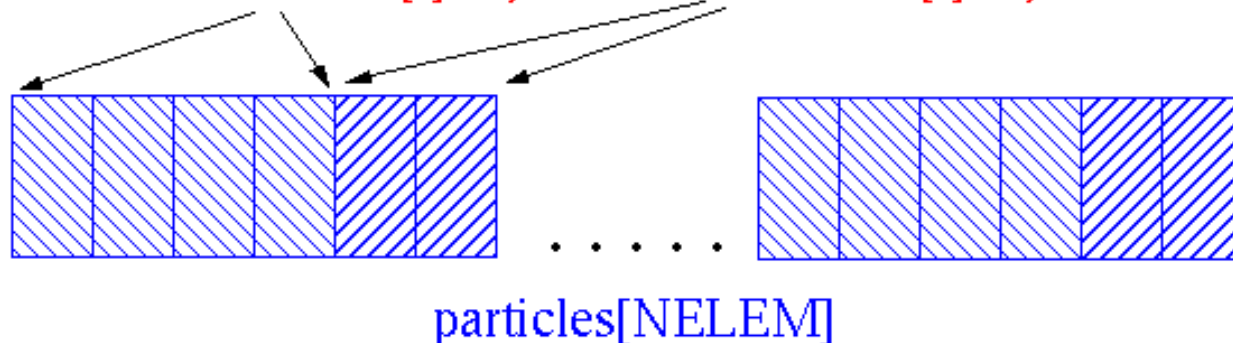


例7

```
typedef struct { float x,y,z,velocity; int n,type; } Particle;  
Particle particles[NELEM];
```

```
MPI_Type_extent(MPI_FLOAT, &extent);
```

```
count = 2; oldtypes[0] = MPI_FLOAT; oldtypes[1] = MPI_INT  
offsets[0] = 0; offsets[1] = 4 * extent;  
blockcounts[0] = 4; blockcounts[1] = 2;
```



```
MPI_Type_struct(count, blockcounts, offsets, oldtypes, &particletype);
```

```
MPI_Send(particles, NELEM, particletype, dest, tag, comm);
```

Sends entire (NELEM) array of particles, each particle being comprised four floats and two integers.

数据的打包与拆包

- 在MPI 中, 通过使用特殊数据类型MPI_PACKED, 用户可以将不同的数据进行打包后再一次发送出去, 接收方在收到消息后再进行拆包.
- 为了与早期其它并行库兼容
- **MPI不建议用户进行显式的数据打包**

数据的打包

```
int MPI_Pack (  
    void*                inbuf        /* in */,  
    int                  incount      /* in */,  
    MPI_Datatype          datatype    /* in */,  
    void*                outbuf       /* in */,  
    int                  outsize      /* in */,  
    int*                 position     /*in/out*/,  
    MPI_Comm              comm        /* in */)
```

- 该函数将缓冲区inbuf 中的incount 个类型为datatype 的数据进行打包. 打包后的数据放在缓冲区outbuf 中. outsize 给出的是outbuf 的总长度(字节数, 供函数检查打包缓冲区是否越界用).

数据的打包

- position 是打包缓冲区中的位移, 第一次调用 MPI_Pack 前用户程序将position 设为0
- 随后MPI_Pack 将自动修改它, 使得它总是指向打包缓冲区中尚未使用部分的起始位置
- 每次调用MPI_Pack 后的position 实际上就是已打包的数据的总长度

例8

```
MPI_Comm_dup(MPI_COMM_WORLD, &com);  
if(my_rank == 0){  
    position = 0;  
    MPI_Pack(n,1,MPI_FLOAT,buff,64,position,com);  
    MPI_Pack(m,1,MPI_INT,buff,64,position,com);  
    MPI_Pack(A,5,MPI_FLOAT,buff,64,position,com);  
    MPI_Send(buff,64,MPI_PACK,1,111,com);  
}
```


数据的拆包

```
int MPI_Unpack (  
    void*                packbuf                /* in */,  
    int                  insize                  /* in */,  
    int*                 position                /*in/out*/,  
    void*                outbuf                  /*out*/,  
    int                  outcount                /* in */,  
    MPI_Datatype          datatype                /* in */,  
    MPI_Comm              comm                   /* in */) 
```

- 从packbuf 中拆包outcount 个类型为datatype 的数据到outbuf 中. 函数中各参数的含义与MPI_Pack 类似, 只不过这里的packbuf 和insize 对应于MPI_Pack中的outbuf 和outsize, 而outbuf 和 outcount 则对应于MPI_Pack 中的inbuf 和incount.

例9

```
MPI_Comm_dup(MPI_COMM_WORLD, &com);  
if(my_rank == 1){  
    MPI_Recv(buff1,64,MPI_PACK,0,111,com,&status);  
    position = 0;  
    MPI_Unpack(buff1,64,position,n1,1,MPI_FLOAT,com);  
    MPI_Unpack(buff1,64,position,m1,1,1,MPI_INT,com);  
    MPI_Unpack(buff1,64,position,A1,5,MPI_FLOAT,com);  
}
```

数据类型函数汇总

函数类型	函数表达
连续数据	MPI_Type_contiguous
向量数据	MPI_Type_vector
	MPI_Type_hvector
索引数据	MPI_Type_indexed
	MPI_Type_hindexed
结构数据	MPI_Type_struct
类型查询	MPI_Type_size
	MPI_Type_extent
	MPI_Type_lb
	MPI_Type_ub
类型递交	MPI_Type_commit
类型释放	MPI_Type_free

上机实验

■ 将矩阵A的转置拷贝到矩阵B中

Type1

$a(0,0)$	$a(0,1) \dots\dots$	$a(0,n)$
$a(1,0)$	$a(1,1) \dots\dots$	$a(1,n)$
$\dots\dots$		
$a(m,0)$	$a(m,1) \dots\dots$	$a(m,n)$

矩阵转置

多种实现方法

1	2	3	
5	6	7	
9	10	11	
13	14	15	



1	5	9	13
2	6	10	14
3	7	11	15

矩阵转置

```
#include <stdio.h>
#include "mpi.h"
#define N 10
int main(int argc, char* argv[]) {
    int my_rank, tag = 0, a[N][N], b[N][N], i, j;
    MPI_Status status;
    MPI_Comm comm;
    MPI_Datatype column_type;
    MPI_Datatype row_type;
    MPI_Init(&argc, &argv); /*初始化MPI*/
    MPI_Comm_dup (MPI_COMM_WORLD, &comm);
    MPI_Comm_rank (comm, &my_rank); /*获得进程的ID*/
    for (i=0; i<N; i++){ /*Initial the matrix a&b*/
        for (j = 0; j < N; j++) {
            a[i][j] = tag++;    b[i][j] = 0;
        }
    }
}
```

矩阵转置

```
MPI_Type_vector(N,1,N,MPI_INT,&column_type); /*定义新数据类型
*/
MPI_Type_commit(&column_type); /*递交新数据类型 */
if(my_rank == 0){ /* P0 send to P1*/
    for (i = 0; i<N; i++)
        MPI_Send(&(a[0][i]),1,new_type,1,tag+i,comm);
    打印矩阵A;
}

if(my_rank == 1){ /* P1 recv from P0 */
    for(i=0; i<N; i++)
        MPI_Recv(&b[i][0],N,MPI_INT,0,tag+i,comm,&status);
    打印矩阵B;
}

MPI_Finalize();
}
```

矩阵转置（接收方的改进策略）

```
MPI_Type_vector(N,1,N,MPI_INT,&column_type); /*定义新数据类型 */
MPI_Type_contiguous(N,MPI_INT,&raw_type);
MPI_Type_commit(&column_type); /*递交新数据类型 */
MPI_Type_commit(&raw_type); /*递交新数据类型 */
if(my_rank == 0){ /* P0 send to P1*/
    for (i = 0; i<N; i++)
        MPI_Send(&(a[0][i]),1,column_type,1,tag+i,comm);
    打印矩阵A;
}
if(my_rank == 1){ /* P1 recv from P0 */
    for(i = 0; i < N; i++)
        MPI_Recv(&b[i][0],1,raw_type,0,tag+i,comm,&status);
    /* MPI_Recv(&b[i][0],N,MPI_INT,0,tag+i,comm,&status); */
    打印矩阵B;
}

MPI_Finalize();
}
```


MPI内容目录

- 基本概念
- 点到点通信
- 自定义数据类型
- 集合通信
- 虚拟拓扑
- 文件IO

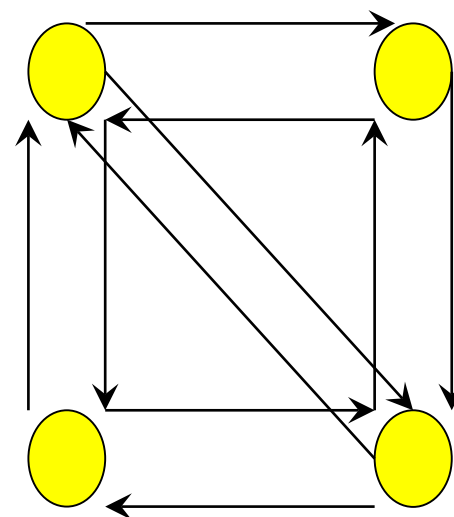
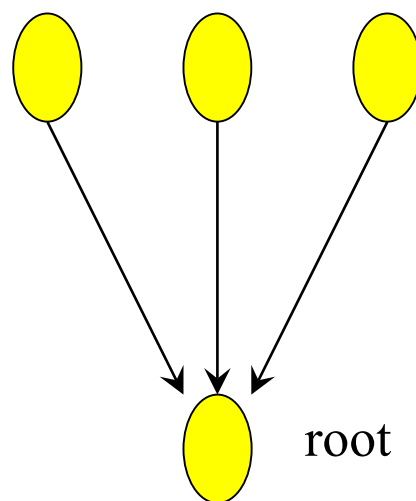
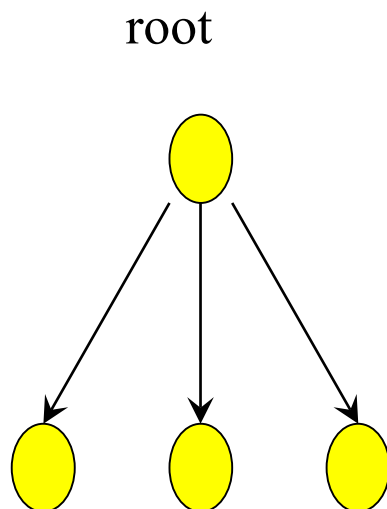


组通信概述

- 通信域限定哪些进程参加以及组通信的上下文
- 组通信调用可以和点对点通信共用一个通信域
 - MPI保证由组通信调用产生的消息不会和点对点调用产生的消息相混淆
- **在组通信中不需要通信消息标志参数**
- 组通信一般实现三个功能**通信、同步和计算**
 - 通信功能主要完成组内数据的传输
 - 同步功能实现组内所有进程在特定地点在执行进度上取得一致
 - 计算功能要对给定的数据完成一定的操作

三种通信方式

■ 一对多、多对一、多对多 (按通信方向)



组通信中的同步

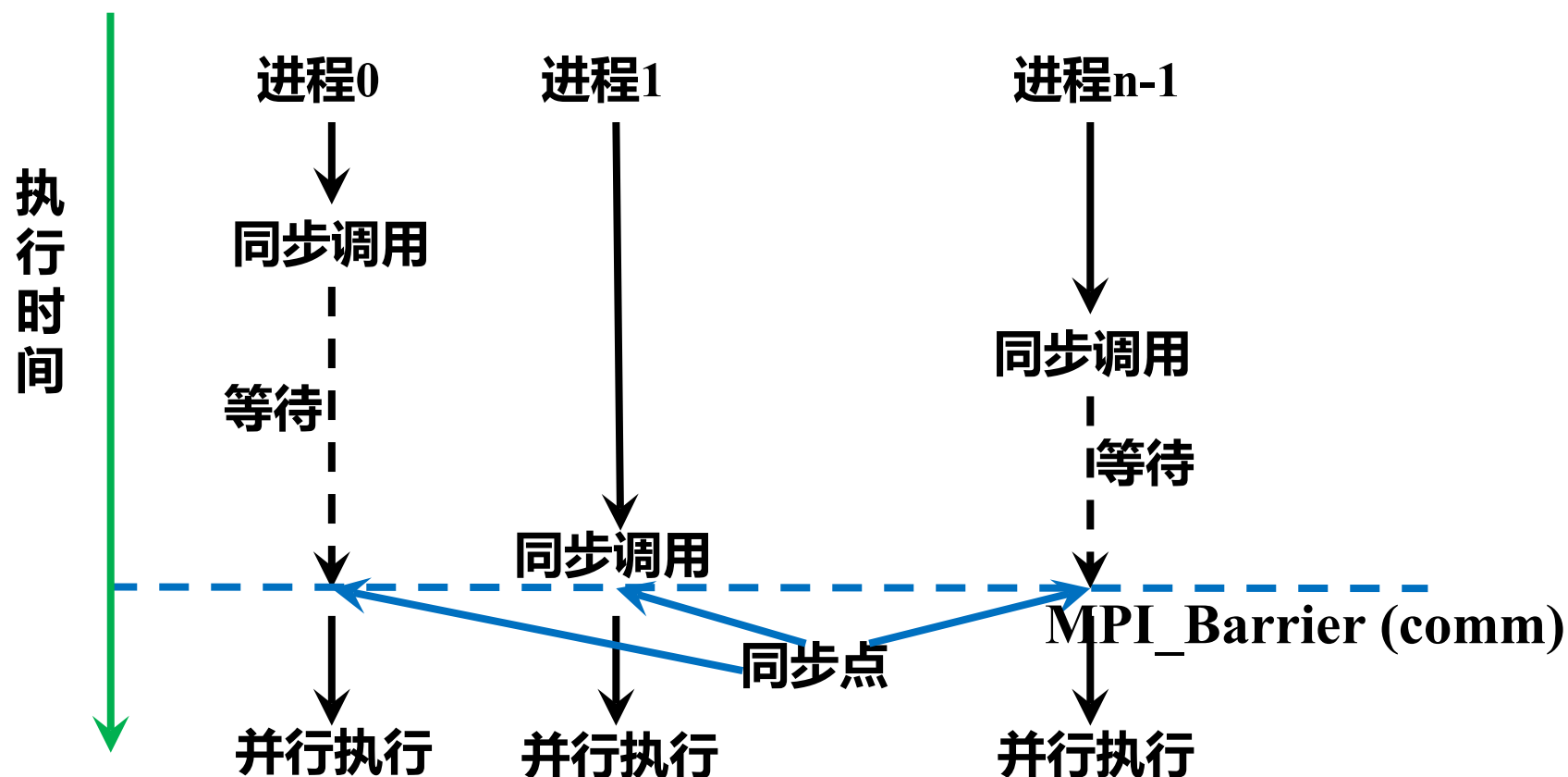
- 点到点通信的完成，重新使用缓冲区
- 一个进程组通信的完成，并不表示其他所有进程的组通信都已经完成。
- 同步操作，完成各个进程之间的同步，协调各个进程的进度和步伐。

同步函数 MPI_Barrier

```
int MPI_Barrier (MPI_Comm comm /* in */)
```

- *MPI* 唯一的一个同步函数, 当comm中的所有进程都执行这个函数后才返回。
- 如果有一个进程没有执行此函数, 其余进程将处于等待状态。在执行完这个函数之后, 所有进程将同时执行其后的任务。

组通信中的同步



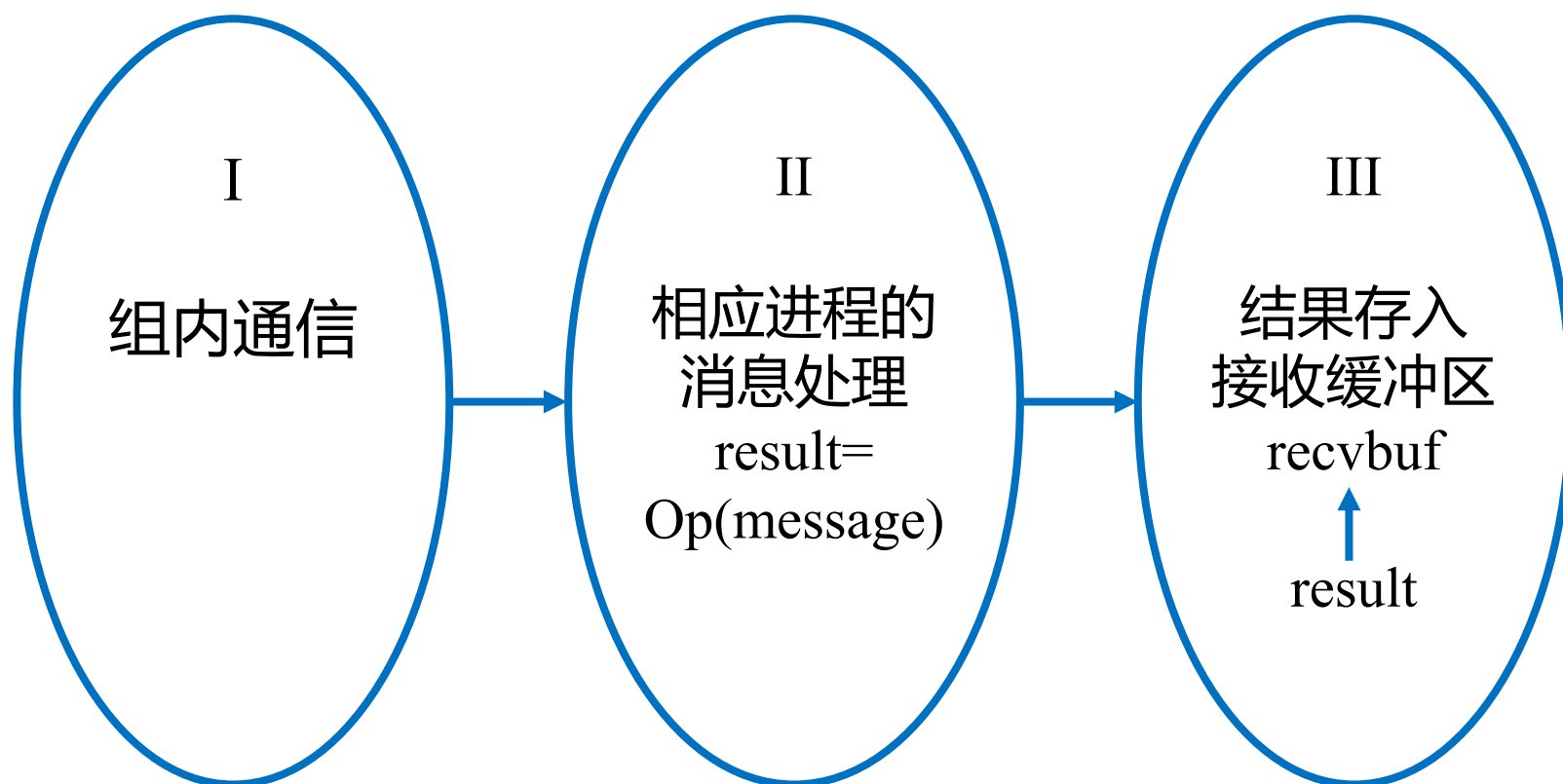
实例: MPI_Barrier

```
MPI_Init( &argc, &argv );
MPI_Comm_rank( MPI_COMM_WORLD, &rank );
MPI_Comm_dup ( MPI_COMM_WORLD, &comm);
if (rank == 0)      value = 100;
MPI_Bcast( &value, 1, MPI_INT, 0, comm);
                /*将该数据广播出去*/
printf( "Process %d got %d\n", rank, value );
MPI_Barrier (MPI_COMM_WORLD);
                /* 同步 */
MPI_Finalize( );
```

组通信中的计算

- **组通信除了通信和同步之外，还可进行计算。MPI组通信的计算功能是分三步实现的**
 - 首先是通信的功能，即消息根据要求发送到目的进程，目的进程也已经接收到了各自所需要的消息
 - 然后是对消息的处理即计算部分。MPI组通信有计算功能的调用都指定了计算操作，用给定的计算操作对接收到的数据进行处理
 - 最后一步是将处理结果放入指定的接收缓冲区

组通信中的计算图示



全局归约MPI Reduce

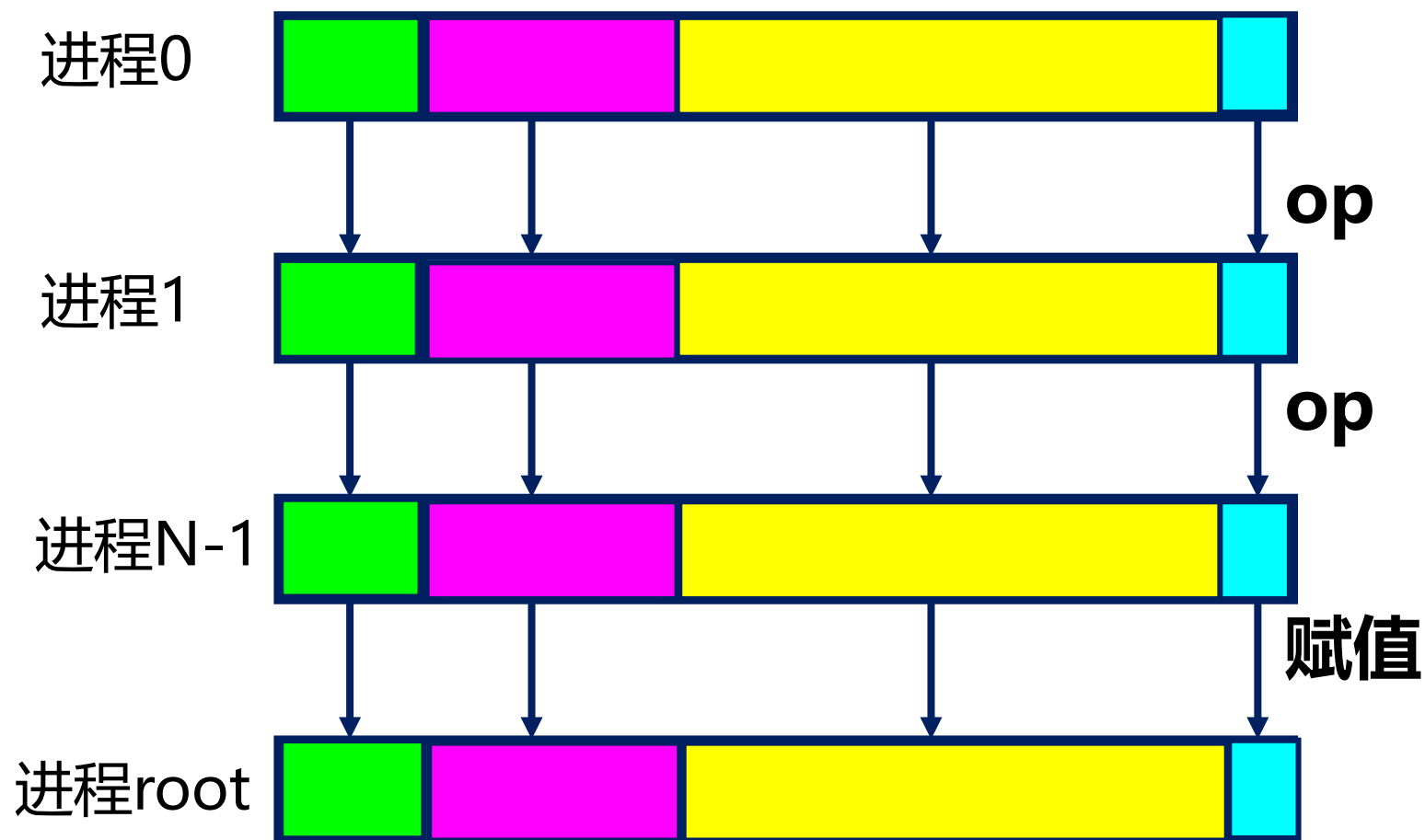
```
int MPI_Reduce (  
    void*          Sendbuf      /* in */,  
    void*          Recvbuf      /* in */,  
    int            count        /* in */,  
    MPI_Datatype    datatype    /* in */,  
    MPI_Op          operator    /*out*/,  
    int            root         /* in */,  
    MPI_Comm        comm        /* in */)
```

Sendbuf	发送缓冲区起始地址
Recvbuf	接收缓冲区(结果)的地址
count	发送缓冲区数据个数
operator	归约操作符

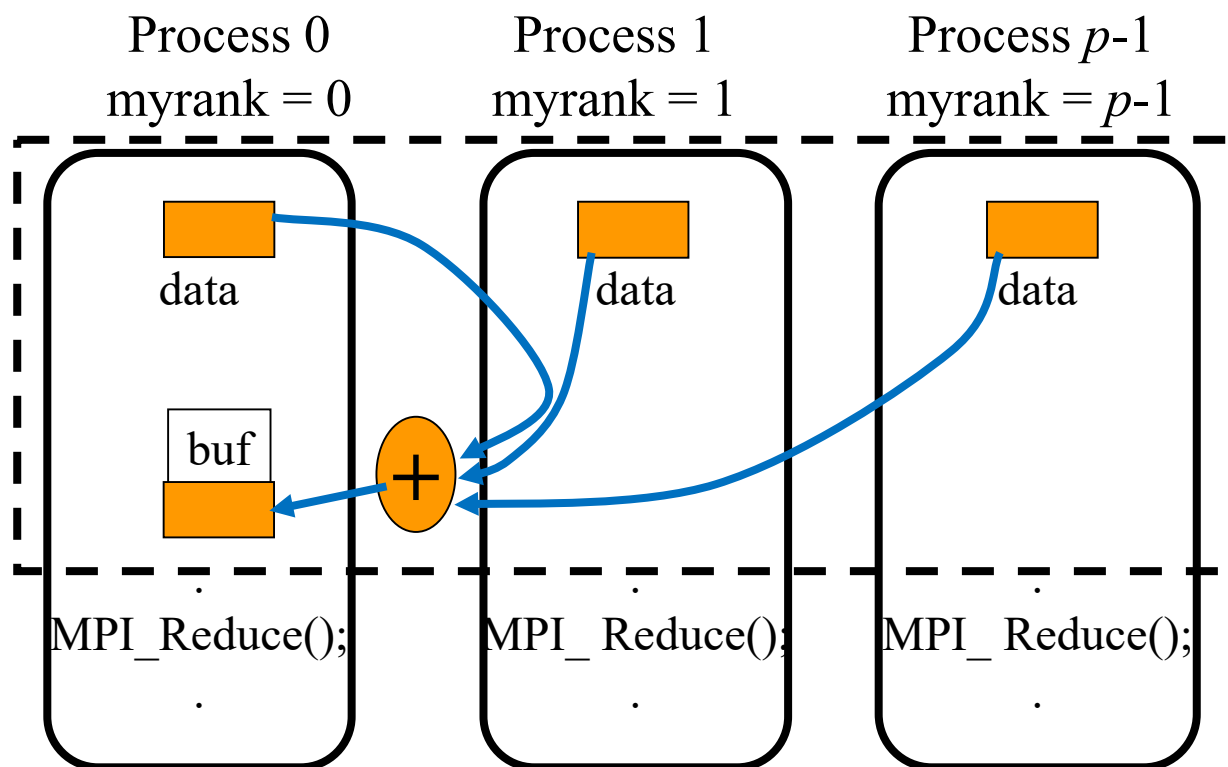
全局数据运算Reduce

- MPI Reduce将组内每个进程输入缓冲区中的数据按给定的操作op进行运算，并将结果返回到根进程的输出缓冲区中。
- 输入缓冲区由参数sendbuf、count和datatype定义。输出缓冲区由参数recvbuf、count和datatype定义
- 要求两者的元素数目和类型都必须相同。所有组成员都用同样的count、datatype、op、root和comm来调用此例程，故所有进程都提供长度相同、元素类型相同的输入和输出缓冲区
- 每个进程可能提供一个元素或一系列元素，组合操作依次针对每个元素进行

MPI Reduce图示

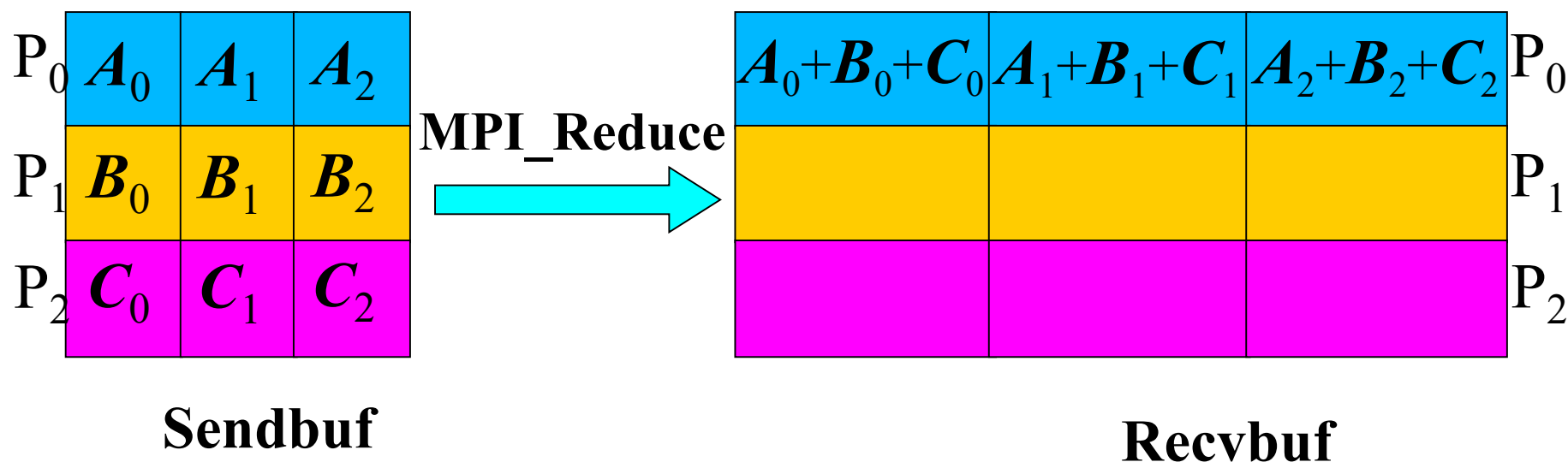


归约求和图示



归约求和图示

MPI_Reduce: np = 3; count = 3; Op = MPI_SUM



MPI预定义操作

名字	含义	名字	含义
MPI_MAX	最大值	MPI_LOR	逻辑或
MPI_MIN	最小值	MPI_BOR	按位或
MPI_SUM	求和	MPI_LXOR	逻辑异或
MPI_PROD	求积	MPI_BXOR	按位异或
MPI LAND	逻辑与	MPI_MAXLOC	求最大值位置
MPI_BAND	按位与	MPI_MINLOC	求最小值位置

组归约函数MPI_Allreduce

```
int MPI_AllReduce (  
    void*          Sendbuf      /* in */,  
    void*          RecvBuf      /* in */,  
    int            count        /* in */,  
    MPI_Datatype    datatype     /* in */,  
    MPI_Op          operator     /*out*/,  
    MPI_Comm        comm        /* in */) 
```

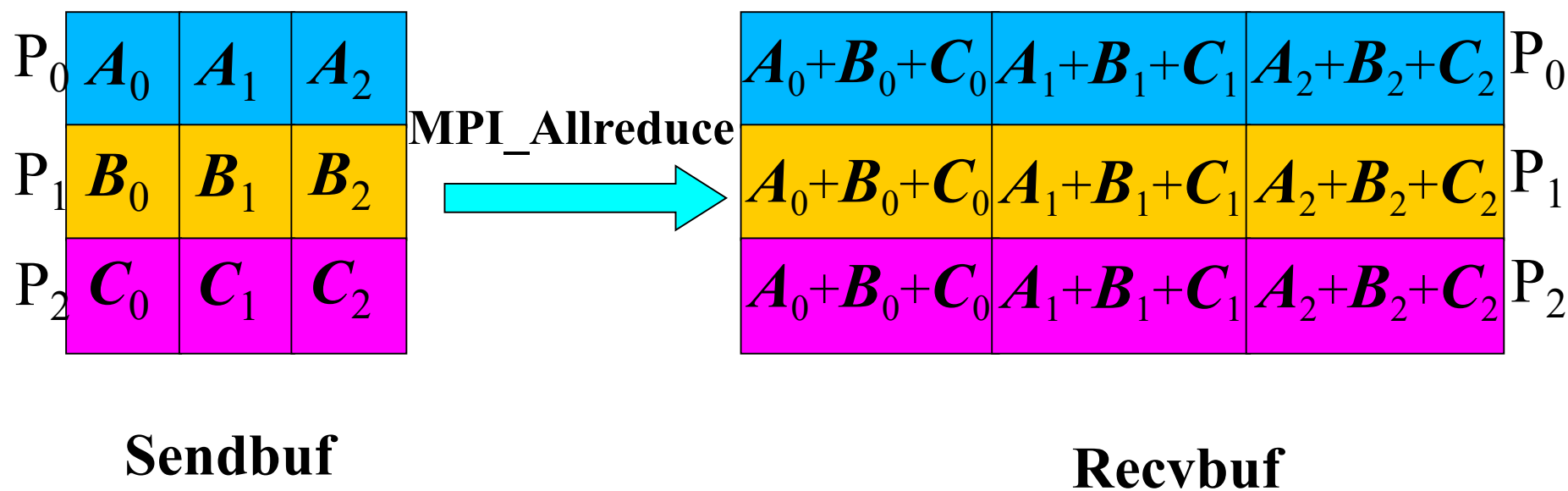
- **MPI_Allreduce**比**MPI_Reduce**少一个root 参数, 其它参数及含义与后者一样

MPI Allreduce函数详解

- **MPI Allreduce**相当于组中每一个进程都作为**root**分别进行了一次归约操作。
- **MPI Allreduce** 相当于在**MPI_Reduce** 后马上再将结果进行一次广播**MPI_Bcast**
- 归约的结果不只是某一个进程所有而是所有的进程都所有。
- **MPI Allreduce** 与 **MPI Reduce**的规约在某种程度上和组收集与收集的关系很相似

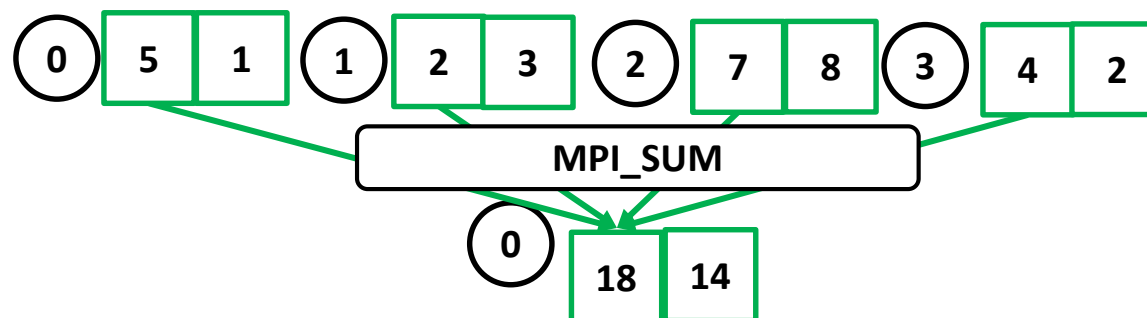
MPI_Allreduce函数图示

MPI_Allreduce: np = 3; count = 3; Op = MPI_SUM

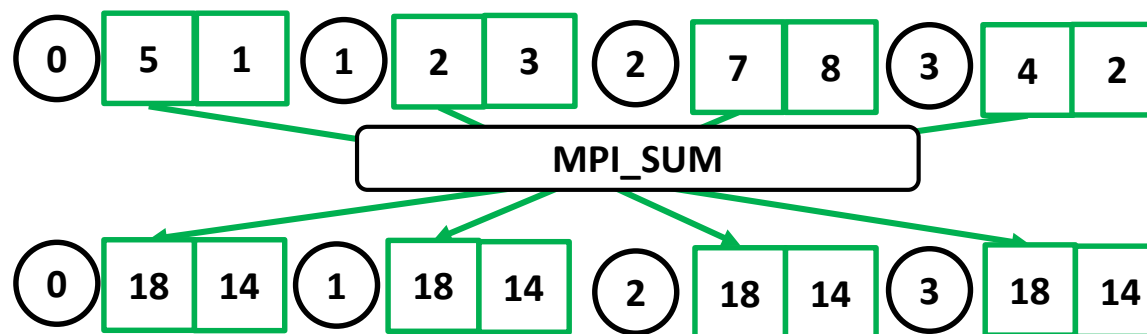


Exercise: Reduce

MPI_Reduce



MPI_Allreduce



数据广播MPI Bcast

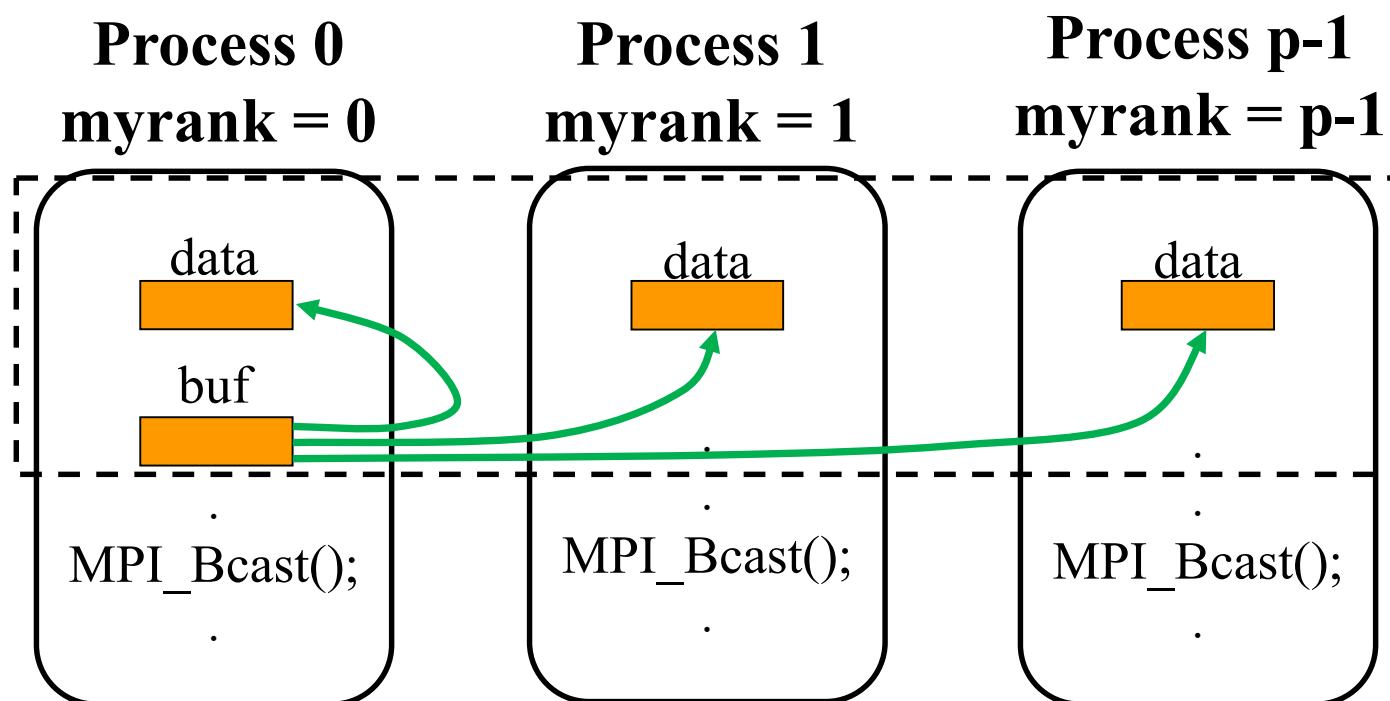
- **MPI Bcast完成从root进程将一条消息广播发送到组内的所有进程，包括它本身在内**
 - 其执行结果是将根进程通信消息缓冲区中的消息拷贝到其他所有进程中去
 - 组内所有进程不管是root进程本身还是其他的进程都使用**同一个通信域comm和根标识root**
 - 数据类型datatype可以是预定义或派生数据类型
 - 其它进程指定的**通信元素个数count、数据类型datatype**必须和根进程指定的count和datatype**保持一致**

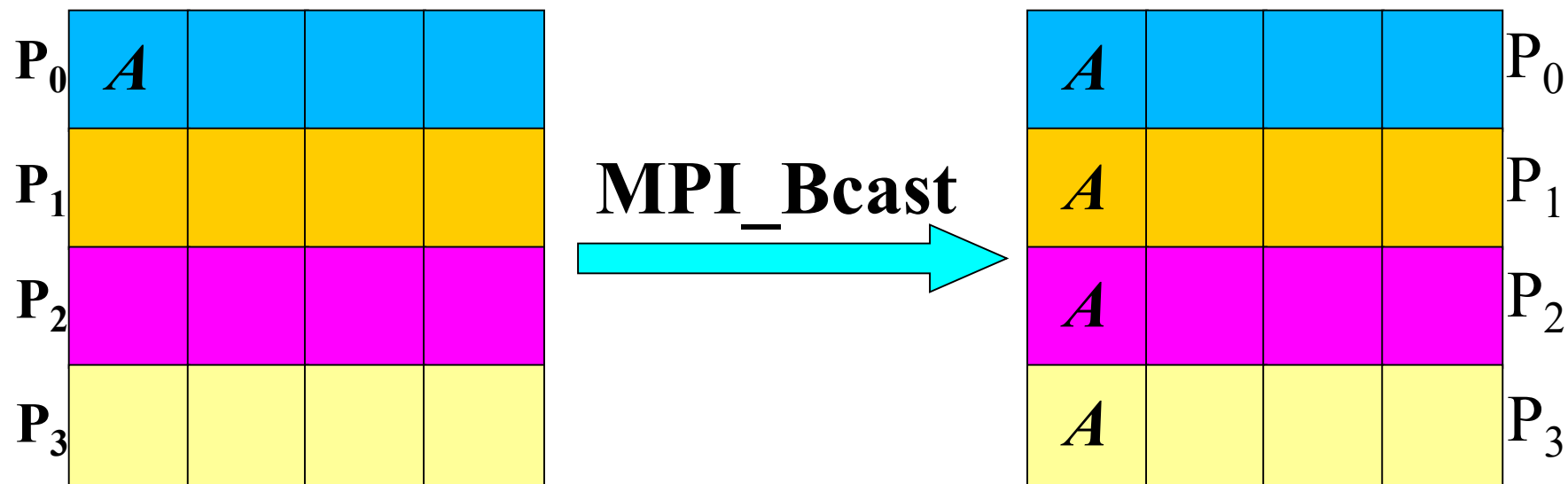
MPI Bcast

```
int MPI_Bcast (  
    void*          buffer      /*in/out*/  
    int            count       /* in */  
    MPI_Datatype   datatype    /* in */  
    int            root        /* in */  
    MPI_Comm       comm        /* in */)
```

buffer	通信消息缓冲区的起始地址
count	将广播出去/或接收的数据个数
datatype	广播/接收数据的数据类型
root	广播数据的根进程的标识号
comm	通信域

MPI Bcast图示



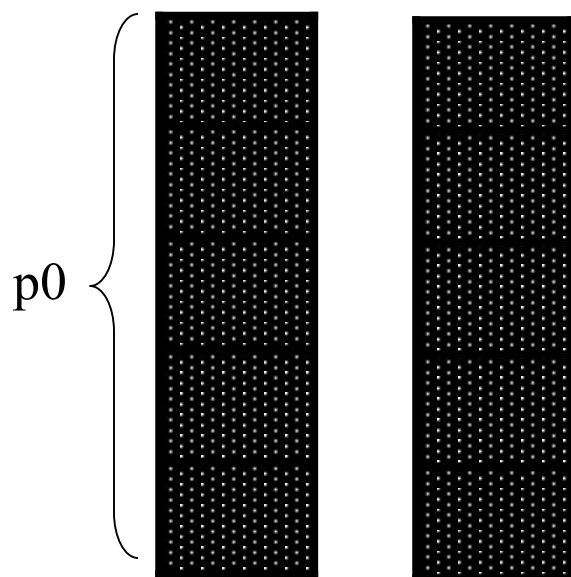


实例: MPI_Bcast

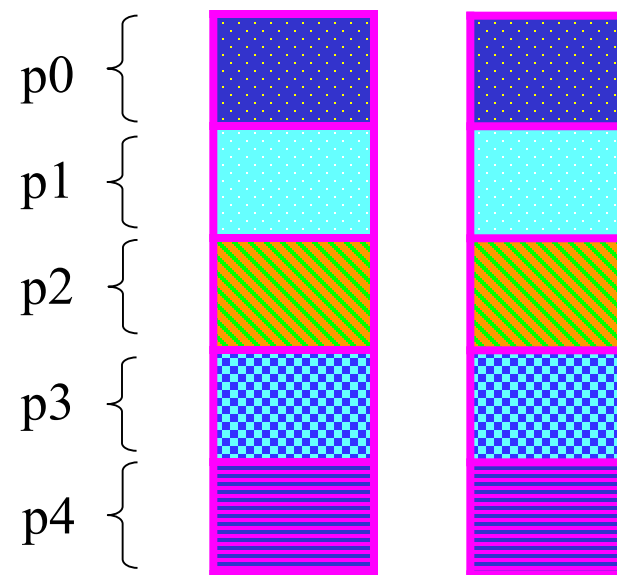
```
MPI_Init( &argc, &argv );
MPI_Comm_rank( MPI_COMM_WORLD, &rank );
MPI_Comm_dup ( MPI_COMM_WORLD, &comm);
if (rank == 0)                /*进程0读入需要广播的数据*/
    scanf( "%d", &value );
MPI_Bcast( &value, 1, MPI_INT, 0, comm);
                                /*将该数据广播出去*/
printf( "Process %d got %d\n", rank, value );
                                /*各进程打印收到的数据*/
MPI_Finalize( );
```


实例:求向量点积

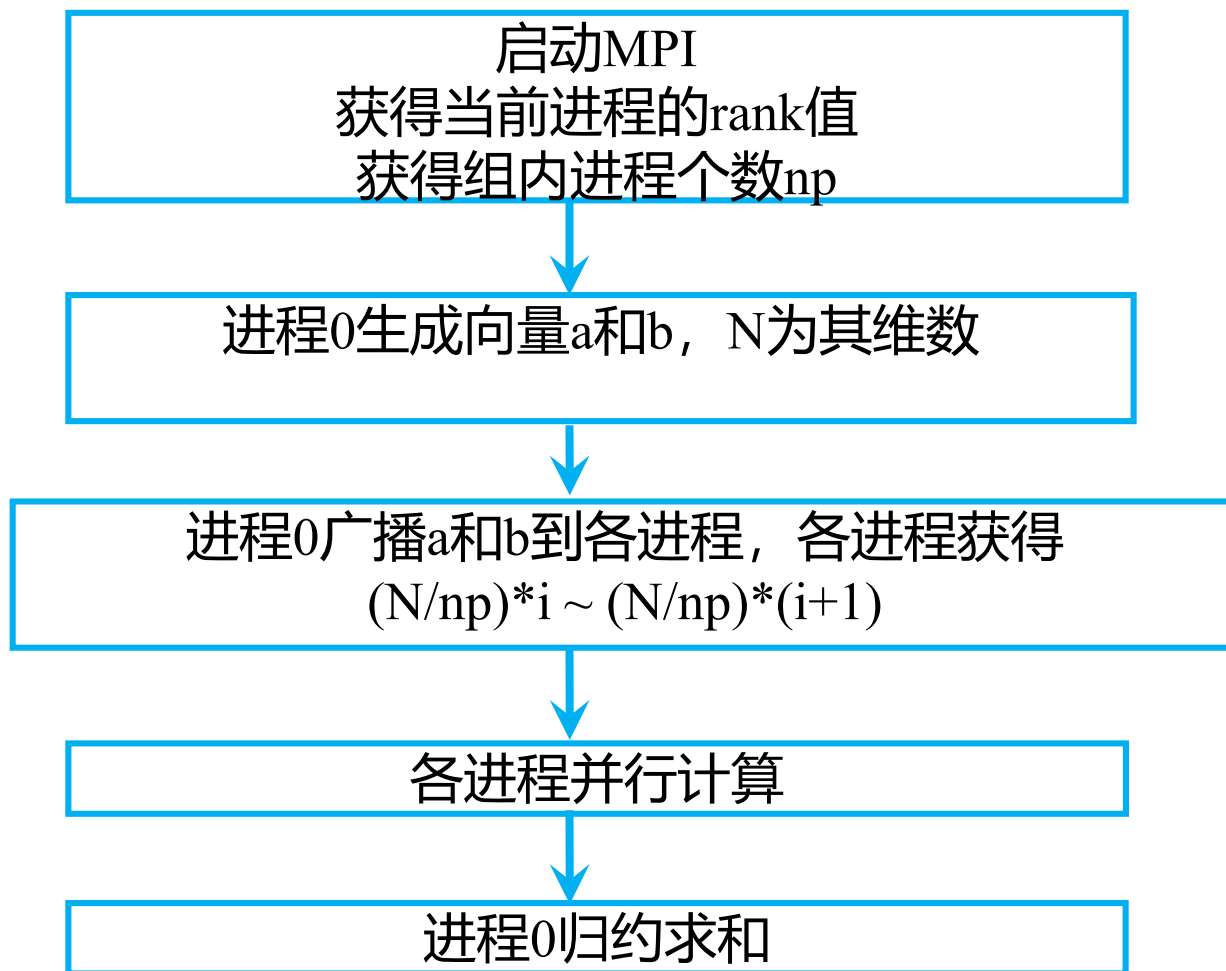
$$c = \sum_{i=0}^{n-1} a_i \cdot b_i$$



$$c = \sum_{j=0}^{n/p} \sum_{i=0}^{n_j-1} a_i \cdot b_i$$



求向量点积计算流程



向量点积代码1

```
#define N 20000
```

```
main(int argc, char** argv) {  
    int *x, *y, gsize, size, myrank, i;  
    float local_sum=0.0, sum;  
    MPI_Status status;    MPI_Comm comm;  
  
    MPI_Init(&argc, &argv);  
    MPI_Comm_dup(MPI_COMM_WORLD, &comm);  
    MPI_Comm_rank(comm, &myrank);  
    MPI_Comm_size(comm, &gsize);
```

```
    x=(int*)malloc(N * sizeof(int));  
    y=(int*)malloc(N * sizeof(int));
```

向量点积代码2

```
if (myrank == 0)          /*给两个向量x, y 赋值*/
    for (i=0; i<N; i++) { x[i] = i +1;      y[i] = i +1; }
    /* 进程0广播向量x和y到各个进程*/
    MPI_Bcast(x, N, MPI_INT, 0, comm);
    MPI_Bcast(y, N, MPI_INT, 0, comm);
    size = N / gsize;
    for (i=0; i<size; i++) { /*各进程并行计算局部向量点积*/
        local_sum =
        local_sum+x[myrank*size+i]*y[myrank*size+i];}

    MPI_Reduce(&local_sum, &sum, 1, /*进程0归约求和*/
               MPI_INT, MPI_SUM, 0, comm);

if (myrank==0) printf("the sum of dot produce
is:%d\n", sum);
free(x);
free(y);
MPI_Finalize();
}
```

数据收集MPI Gather

- 把所有进程(包括root) 的数据聚集到root 进程中, 并且按顺序存放在接收缓冲区中.
- 其结果就象一个进程组中的N个进程(包括root)都执行了一个发送调用, 同时根进程执行了N次接收调用.

MPI_Gather函数

```
int MPI_Gather (  
    void*          sendbuf          /* in */  
    int            sendcount        /* in */  
    MPI_Datatype    sendtype        /* in */  
    void*          recvbuf         /*out*/  
    int            recvcount        /* in */  
    MPI_Datatype    recvtype        /* in */  
    int            root             /* in */  
    MPI_Comm        comm            /* in */) 
```

sendbuf	发送缓冲区起始地址
sendcount	发送缓冲区数据个数
sendtype	发送缓冲区数据类型

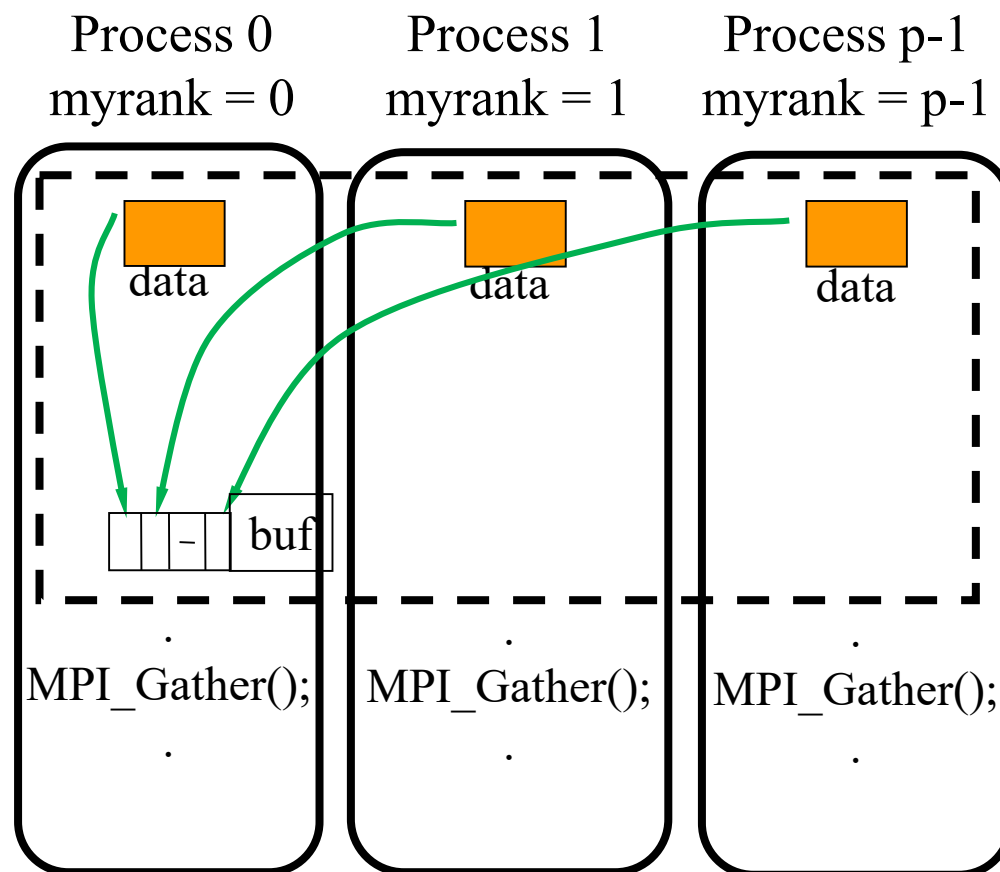
recvbuf	接收缓冲区起始地址
recvcount	接收缓冲区数据个数
recvtype	接收缓冲区数据类型

Root ONLY

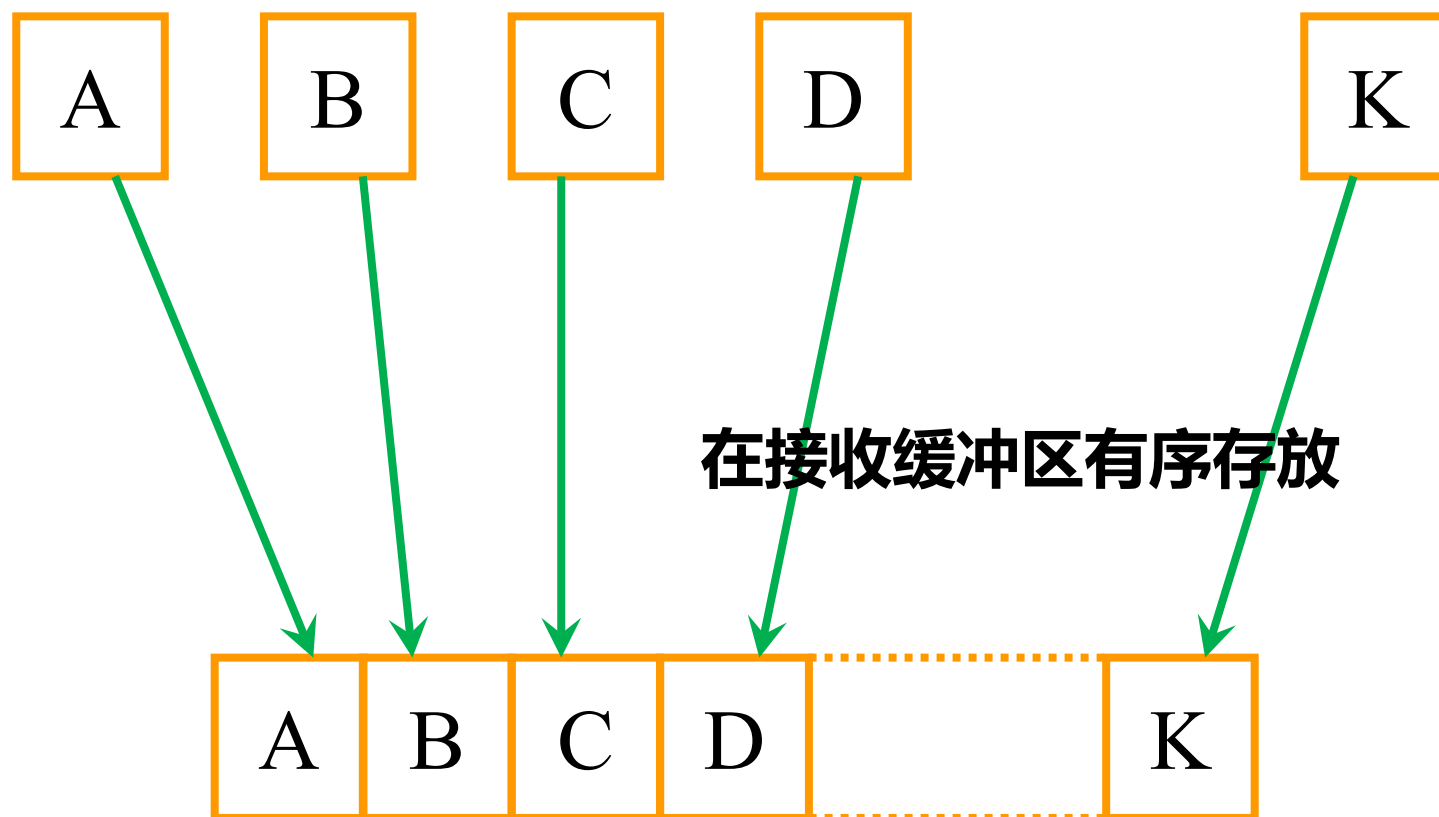
MPI_Gather函数详解

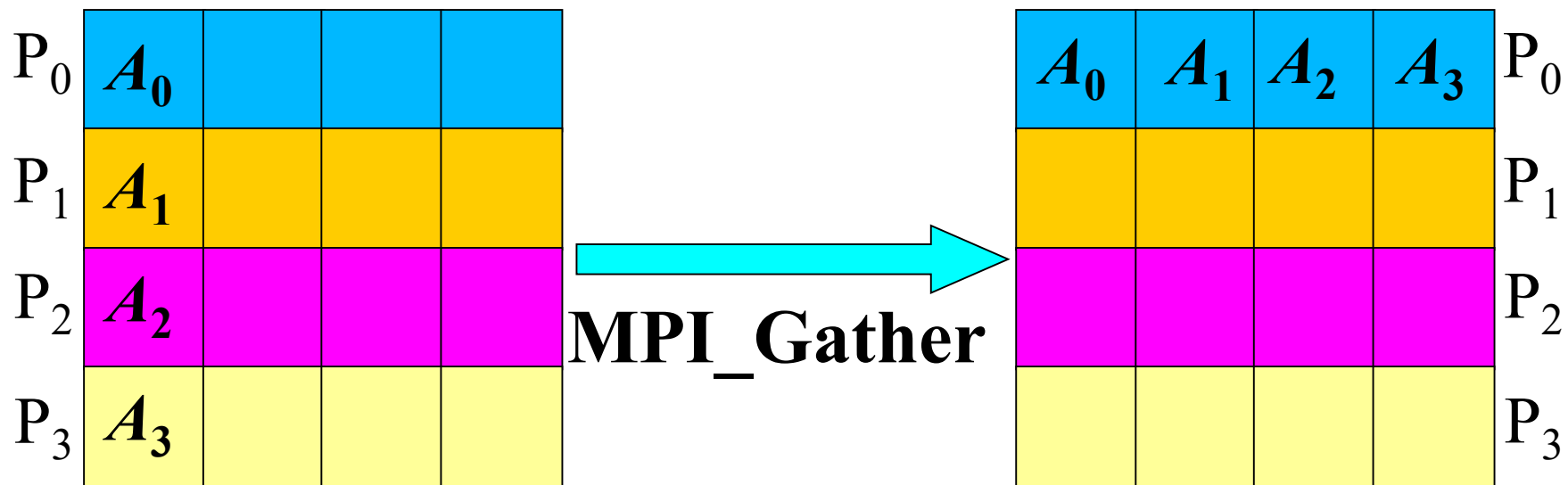
- 从各个进程收集到的数据一般是互不相同的
- 收集调用每个进程的发送数据个数sendcount和发送数据类型sendtype都是相同的，都和根进程中接收数据个数recvcount和接收数据类型recvtype相同。
- root和comm在所有进程中都必须是一致的
- 根进程中指定的接收数据个数是指从每一个进程接收到的数据的个数而不是总的接收个数
- 对于所有非根进程接收消息缓冲区被忽略但是各个进程必须提供这一参数
- 所有参数对根进程都是有意义的，而对于其它进程只有sendbuf、sendcount、sendtype、root和comm有意义，其它的参数虽没有意义但却**不能省略**

MPI Gather图示



MPI Gather图示





实例: MPI_Gather

```
MPI_Comm      comm;
int           size, root, s_data[100], *rbuf;
...
MPI_Comm_size( comm, &size );
rbuf = (int *) malloc( size * 100*sizeof(int) );
           /* 申请接收缓冲区 */
MPI_Gather( s_data, 100, MPI_INT, rbuf, 100,
           MPI_INT, root, comm );
...
MPI_Finalize( );
```

数据散发MPI_Scatter

- MPI_Scatter是一对多的组通信调用
- 但是和广播不同，Root向各个进程发送的数据可以是不同的
- MPI_Scatter和MPI_Gather的效果正好相反两者互为逆操作

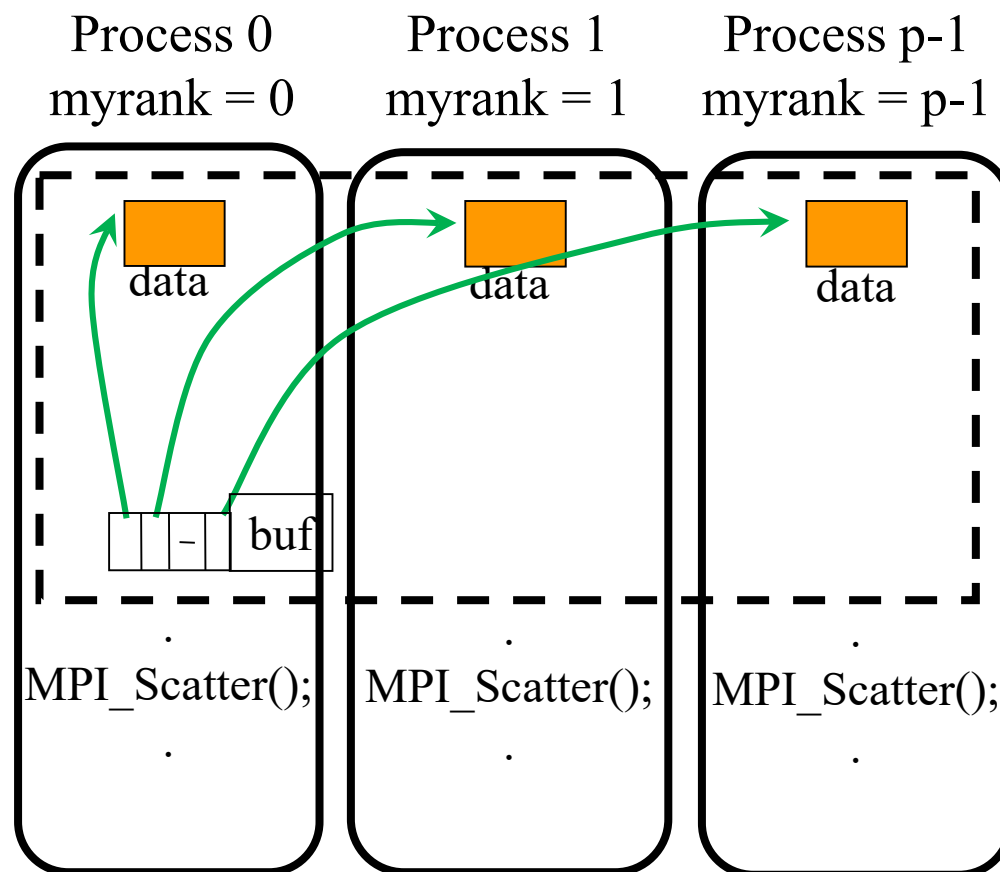
Scatter函数

```
int MPI_Scatter (  
    void*          sendbuf          /* in */  
    int            sendcount        /* in */  
    MPI_Datatype   sendtype         /* in */  
    void*          recvbuf          /* out */  
    int            recvcnt          /* in */  
    MPI_Datatype   recvttype        /* in */  
    int            root             /* in */  
    MPI_Comm       comm             /* in */)   
    sendbuf  发送缓冲区起始地址  recvbuf  接收缓冲区起始地址  
    sendcount 发送缓冲区数据个数  recvcnt  接收缓冲区数据个数  
    sendtype  发送缓冲区数据类型  recvttype 接收缓冲区数据类型
```

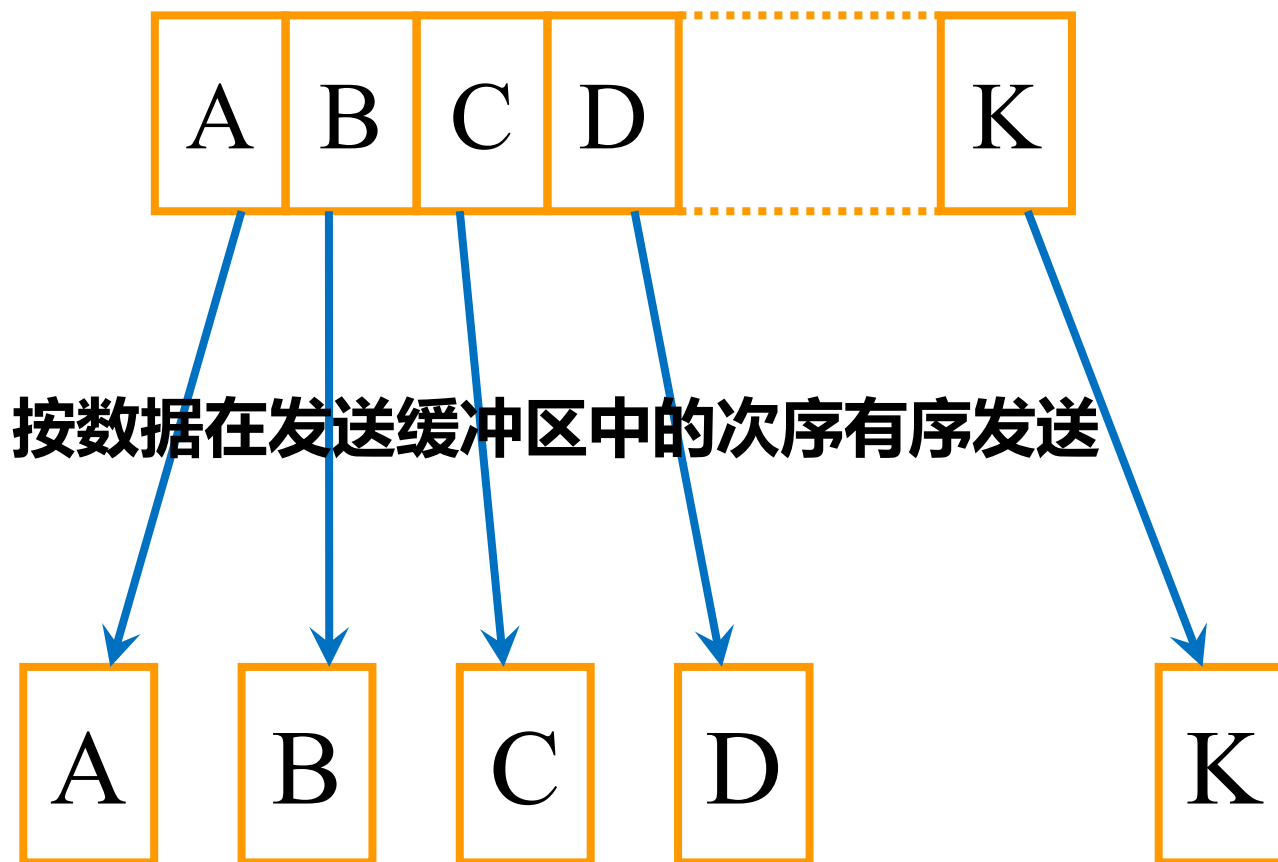
MPI Scatter函数详解

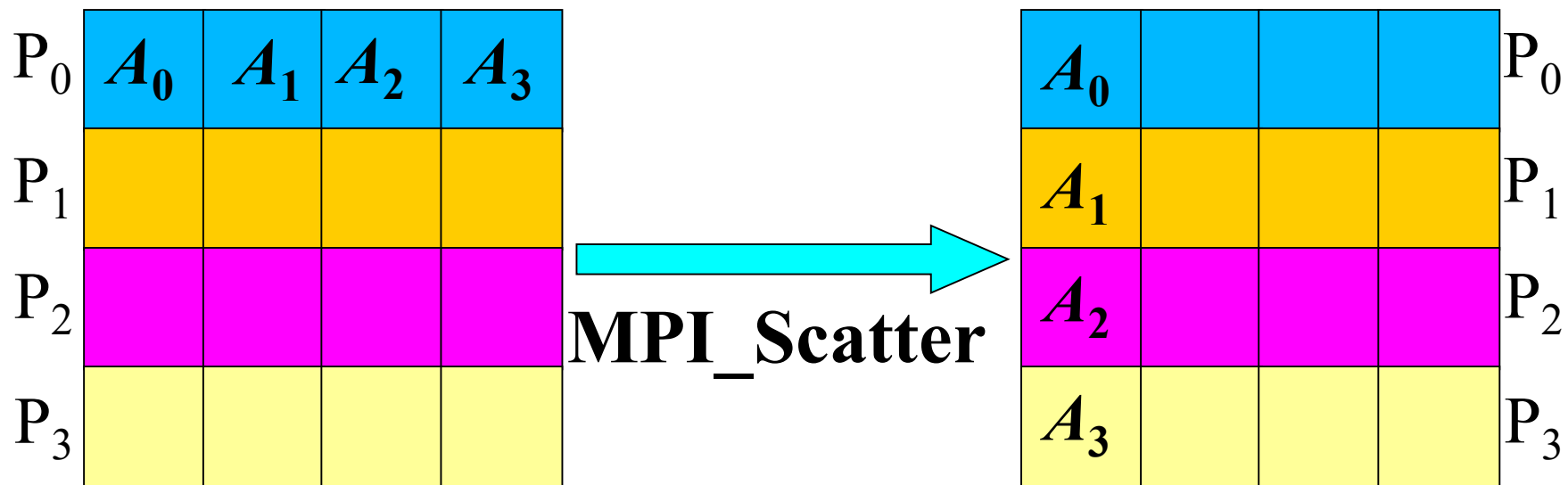
- 根进程中的发送数据元素个数sendcount和发送数据类型sendtype必须和所有进程的接收数据元素个数recvcount和接收数据类型recvtype相同
- 对于所有非根进程，发送消息缓冲区被忽略
- 根进程发送元素个数指的是发送给每一个进程的数据元素的个数而不是总的元素个数
- 此调用中的所有参数对根进程来说都是有意义的而对于其他进程来说只有recvbuf、recvcount、recvtype、root和comm是有意义的参数
- root和comm在所有进程中都必须是一致的

MPI Scatter图示



MPI Scatter图示






实例: MPI_Scatter

```
MPI_Comm comm;
int size, *sendbuf;
int root, rbuf[100];
.....
MPI_Comm_size(comm, &size);
sendbuf = (int *)malloc(size * 100 * sizeof(int));
.....
MPI_Scatter(sendbuf, 100, MPI_INT, rbuf,
            100, MPI_INT, root, comm);
```

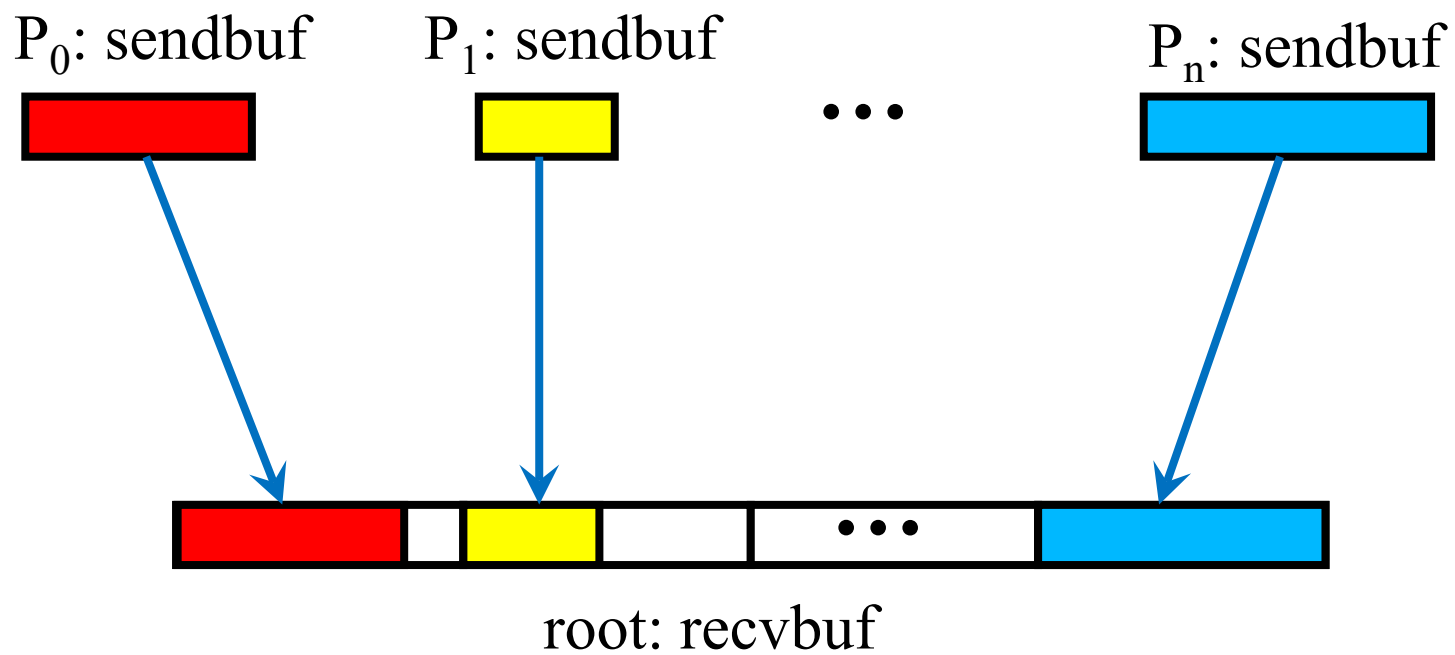
MPI_Gatherv函数

```
int MPI_Gatherv (  
    void*          sendbuf          /* in */  
    int            sendcount        /* in */  
    MPI_Datatype   sendtype        /* in */  
    void*          recvbuf         /*out*/  
    int            recvcnts[]    /* in */  
    int            displs[]     /* in */  
    MPI_Datatype   recvttype       /* in */  
    int            root            /* in */  
    MPI_Comm       comm           /* in */)
```



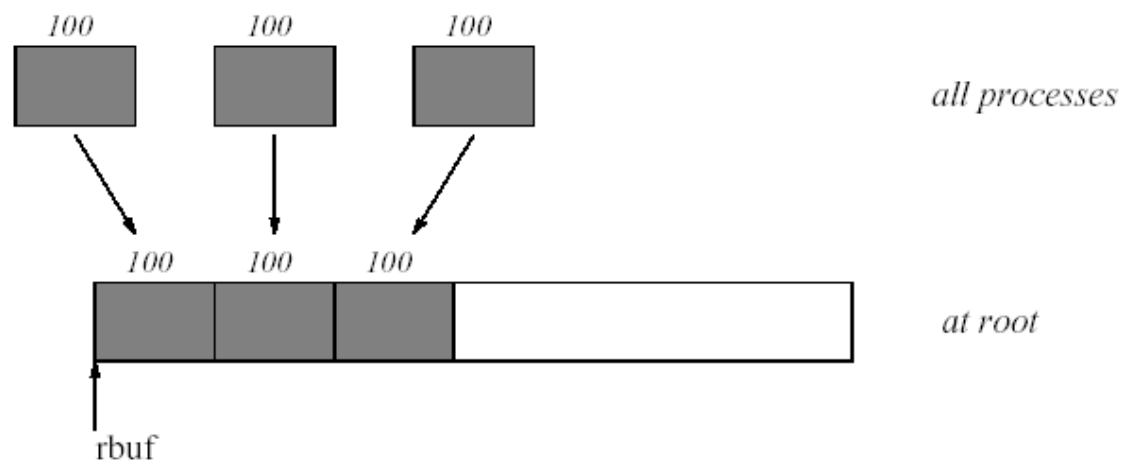
MPI_Gatherv函数详解

- 收集不同长度数据块。从不同进程接收不同数量的数据，为此接收数据元素的个数`recvcounts`是一个数组，用于指明从不同的进程接收的数据元素的个数
- 根从每一个进程接收的数据元素的个数可以不同，但是发送和接收的个数必须一致
- 除此之外它还为每一个接收消息在接收缓冲区的位置提供了一个位置偏移`displs`数组用户可以将接收的数据存放到根进程消息缓冲区的任意位置
- MPI_Gatherv明确指出了从不同的进程接收数据元素的个数以及这些数据在Root接收缓冲区存放的起始位置

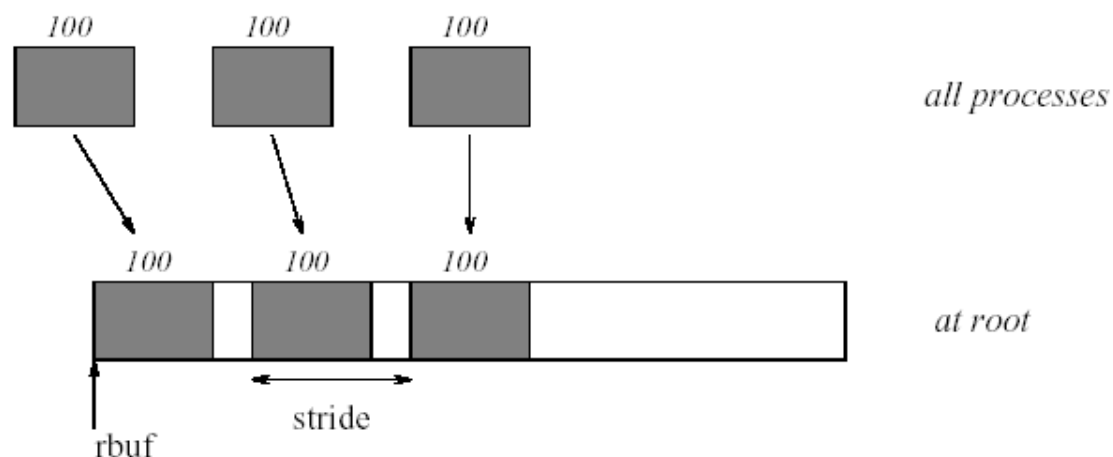


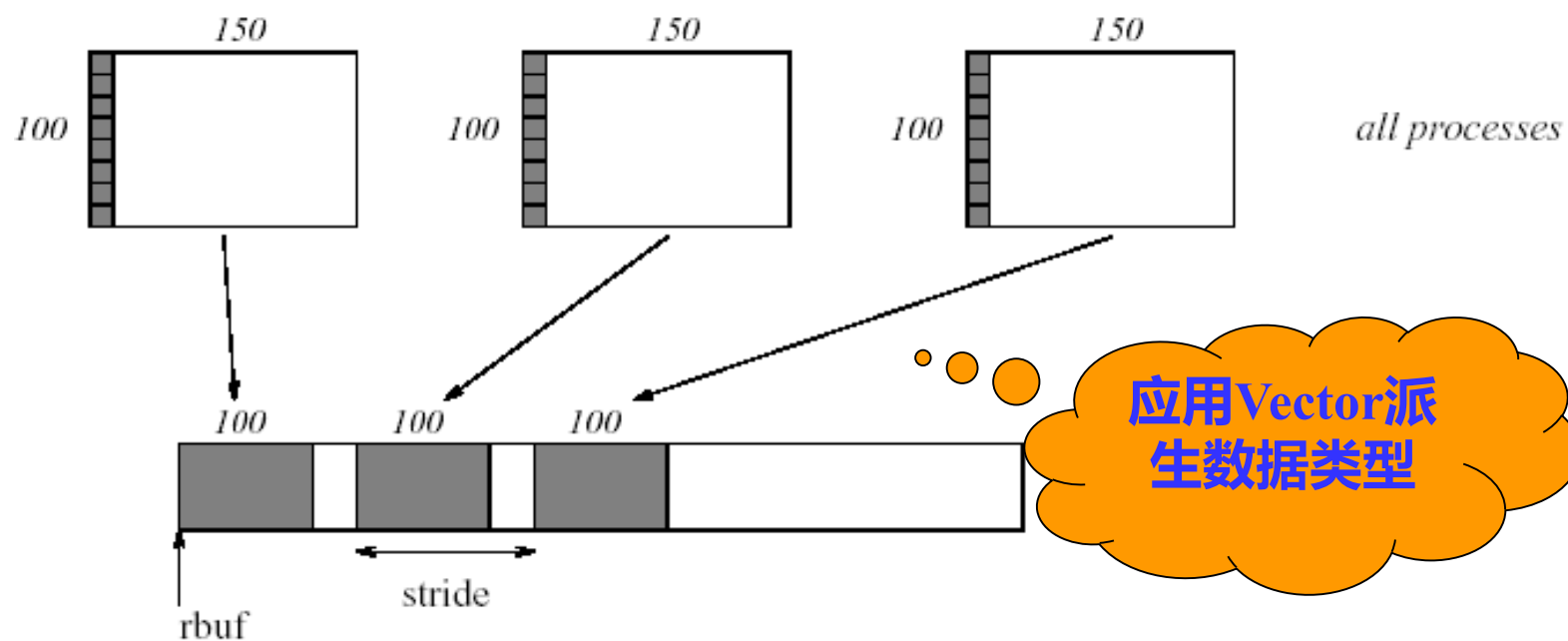
MPI Gather vs MPI Gatherv

Gather



Gatherv





实例:MPI_Gatherv


```
int    size, root, s_data[100], *rbuf, *displs,
      *rcount;
int          stride = 120;
...
MPI_Comm_size( comm, &size );
rbuf = (int *) malloc( size * stride *
    sizeof(int));
displs = (int *) malloc( size * sizeof(int)); /*申请缓冲区*/
rcounts = (int *) malloc( size * sizeof(int));
for (i = 0; i < size; i ++){
    displs[i] = i * stride;  rcounts[i] = 100;}

MPI_Gatherv( s_data, 100, MPI_INT, rbuf, rcounts,
    displs, MPI_INT, root, comm);
...
```


MPI_Scatterv函数

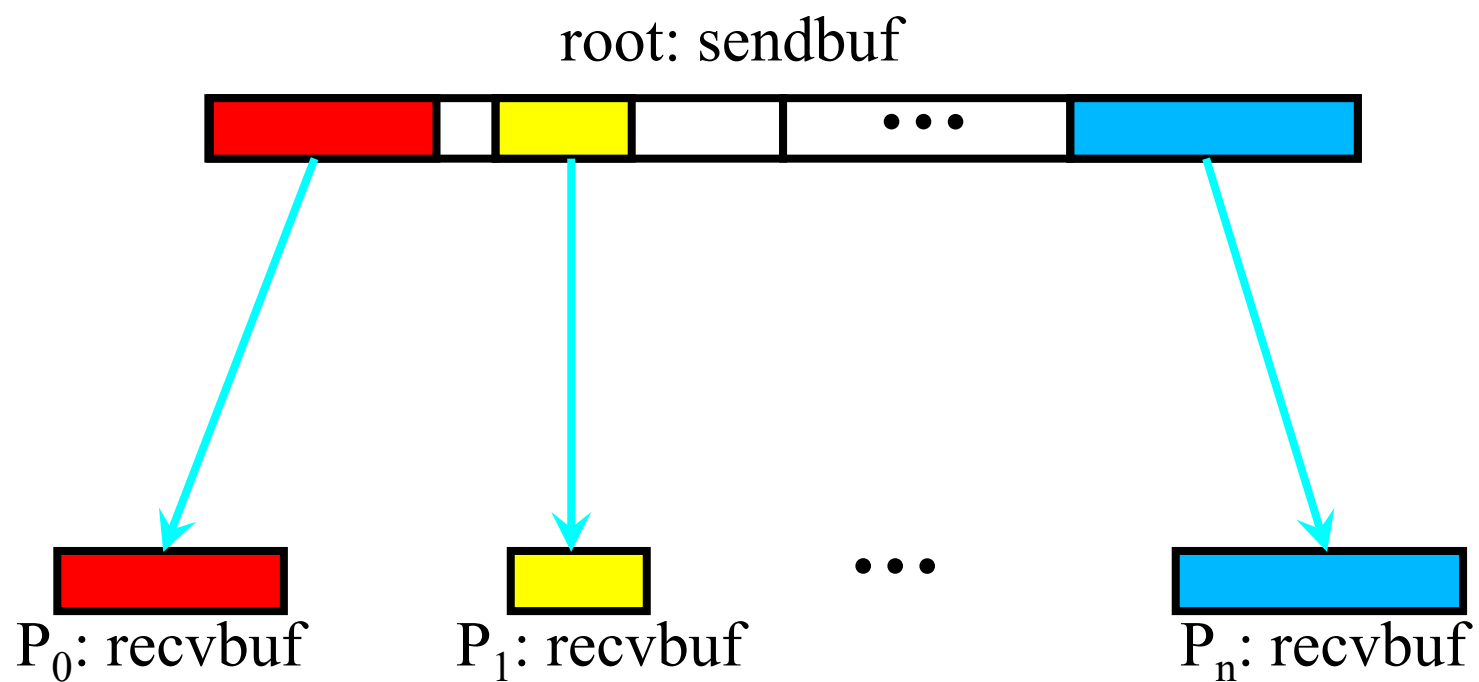
```
int MPI_Scatterv (  
    void*                sendbuf                /* in */  
    int                  sendcounts[]           /* in */  
    int                  displs[]               /* in */  
    MPI_Datatype          sendtype              /* in */  
    void*                recvbuf/*out*/  
    int                  recvcnt               /* in */  
    MPI_Datatype          recvtype             /* in */  
    int                  root                 /* in */  
    MPI_Comm              comm                /* in */)

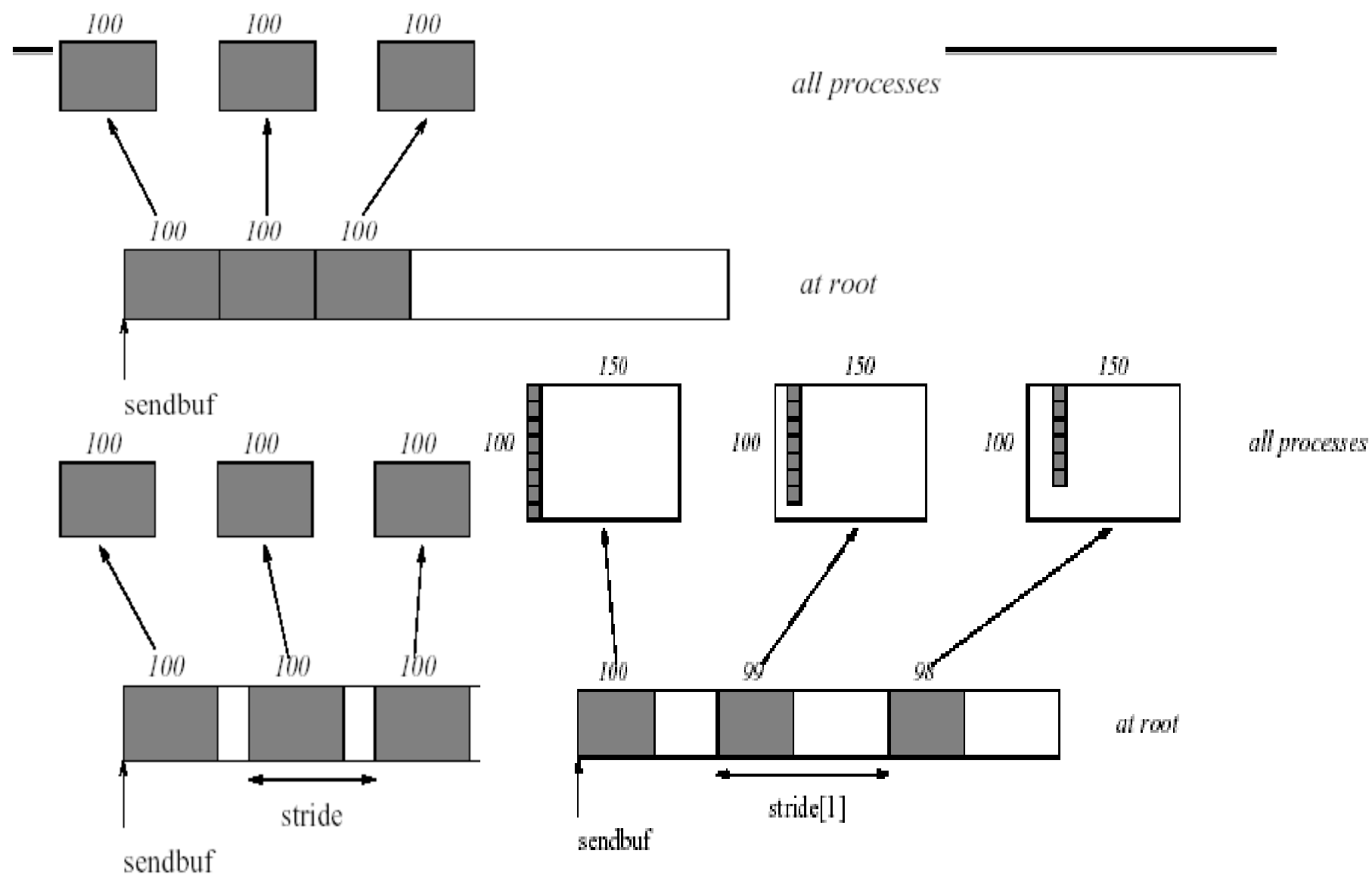
```



MPI_Scatterv函数详解

- MPI_Scatterv允许Root向各个进程发送个数不等的数。因此要求**sendcounts是一个数组**同时还提供一个新的参数**displs**指明根进程发往其它不同进程数据在根发送缓冲区中的**偏移位置**
- 根进程中sendcount[i]和sendtype的类型必须和进程i的recvcount和recvtype的类型相同这就意味着在每个进程和根进程之间发送的数据量必须和接收的数据量相等
- 对于所有非根进程，发送消息缓冲区被忽略。
- 此调用中的所有参数对根进程来说都是很重要的而对于其他进程来说只有recvbuf、recvcount、recvtype、root和comm是有意义的
- 参数root和comm在所有进程中都必须是一致的





实例: MPI Scatterv

```
int i, size, root, *sendbuf,
    rbuf[100], *displs, *counts;
int stride = 120;
...
MPI_Comm_size(comm, &size);

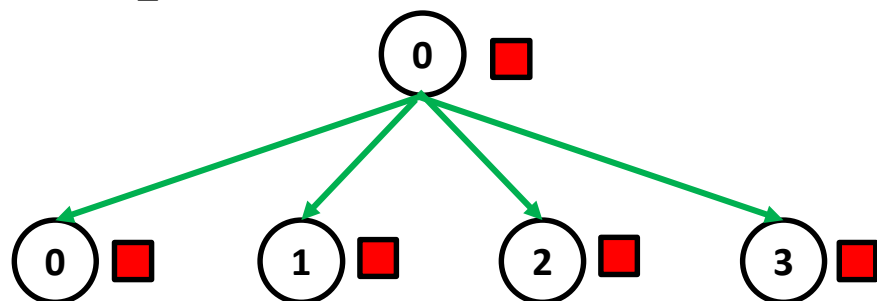
sendbuf = (int *)malloc(size * stride *
    sizeof(int));
displs = (int *)malloc(size * sizeof(int)); /* 申请缓冲区 */
counts = (int *)malloc(size * sizeof(int));

for (i=0; i<size; ++i) {
    displs[i] = i*stride;    counts[i] = 100;}
```

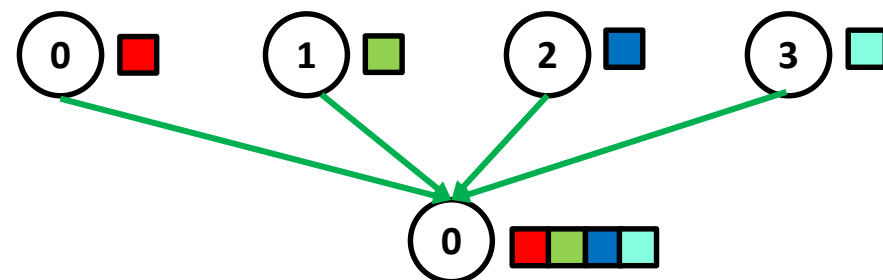
```
MPI_Scatterv(sendbuf, counts, displs, MPI_INT,
    rbuf, 100, MPI_INT, root, comm);
```

小结

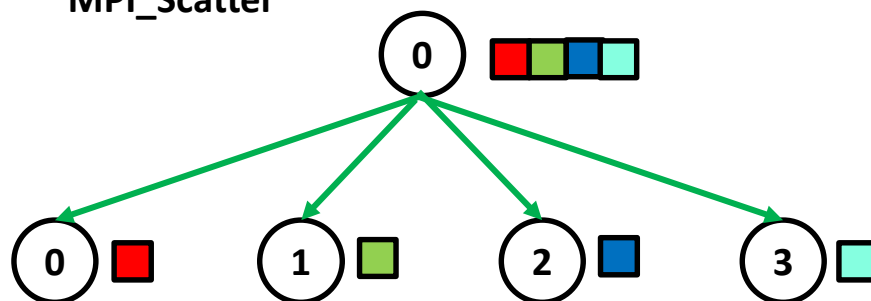
MPI_Bcast



MPI_Gather



MPI_Scatter



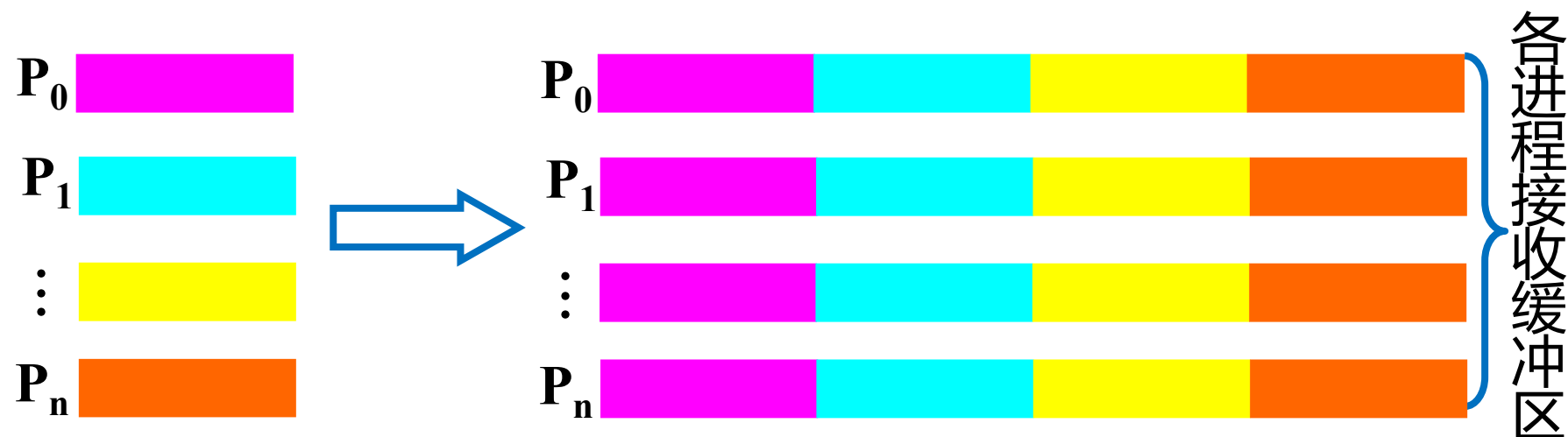
MPI Allgather函数

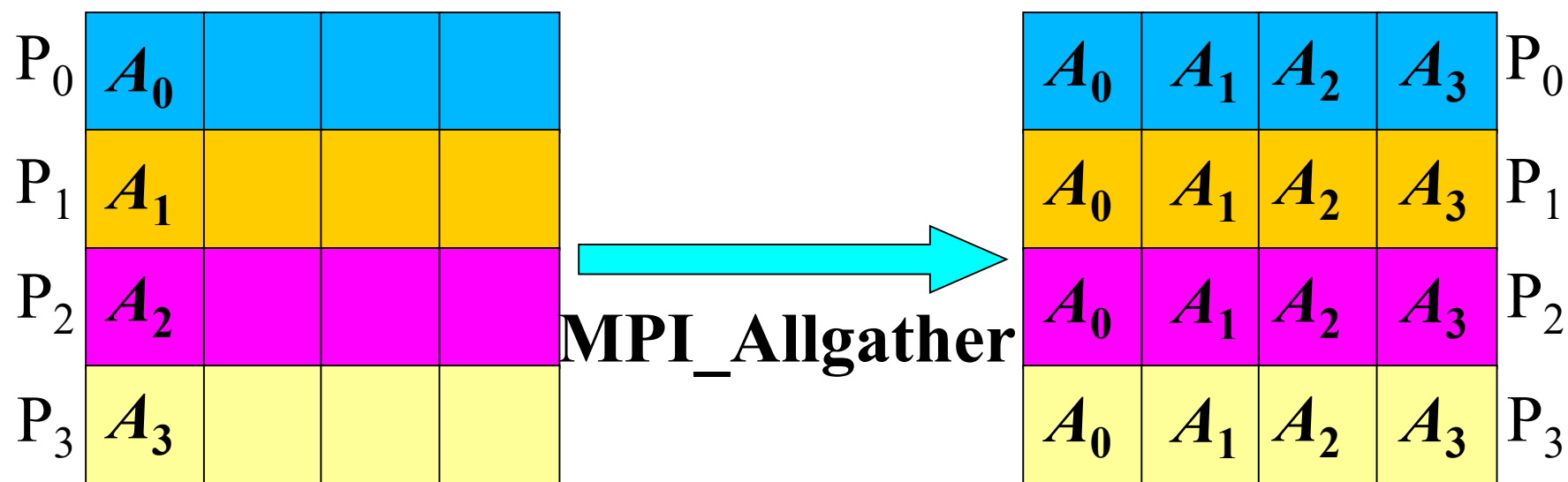
```
int MPI_Allgather (  
    void*          sendbuf          /* in */  
    int            sendcount        /* in */  
    MPI_Datatype   sendtype         /* in */  
    void*          recvbuf          /*out*/  
    int            recvcount        /* in */  
    MPI_Datatype   recvtype         /* in */  
    MPI_Comm       comm             /* in */) 
```

MPI_Allgather函数详解

- **MPI_Gather**将数据收集到根进程, **MPI_Allgather**相当于每一个进程都作为Root执行了一次**MPI_Gather**调用, 即每一个进程都收集到了其它所有进程的数据
- 从参数上看**MPI_Allgather**和**MPI_Gather**完全相同, 只不过在执行效果上对于**MPI_Gather**执行结束后, 只有Root进程的接收缓冲区有意义;
- **MPI_Allgather**调用结束后所有进程的接收缓冲区都有意义。
- 它们接收缓冲区的内容是相同的。


MPI_Allgather函数图示





MPI_Allgather函数

```
int MPI_Allgather(  
    void*      sendbuf          /* in */,  
    int        sendcount        /* in */,  
    MPI_Datatype sendtype        /* in */,  
    void*      recvbuf          /*out*/,  
    int         recvcnts[]       /* in */,  
    int         displs[]        /* in */,  
    MPI_Datatype recvtype        /* in */,  
    MPI_Comm    comm            /* in */)
```



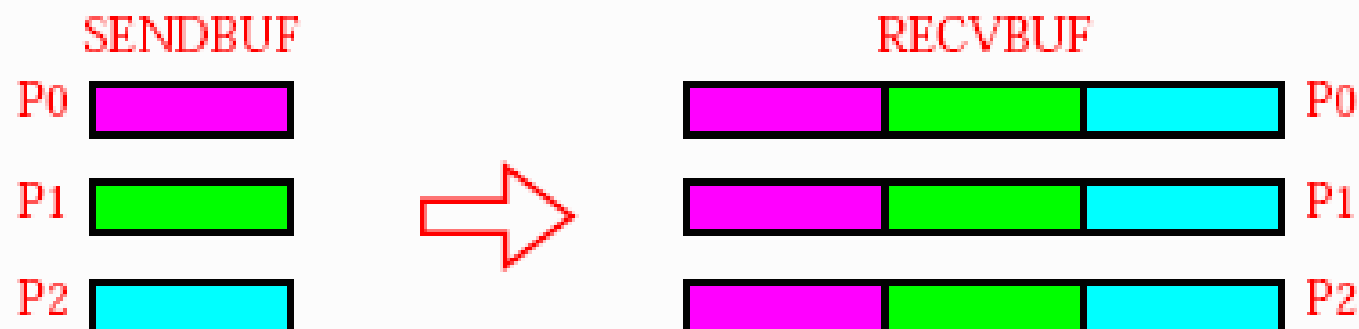
Root ONLY

MPI_Allgatherv函数详解

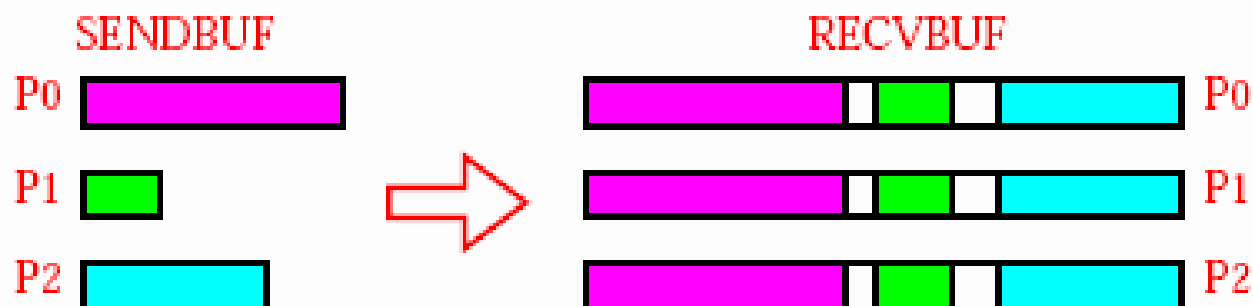
- **MPI_Allgatherv**也是所有的进程都将接收结果，而不是只有根进程接收结果；
- 从每个进程发送的第**j**块数据将被每个进程接收，然后存放在各个进程接收消息缓冲区**recvbuf**的第**j**块，进程**j**的**sendcount**和**sendtype**的类型必须和其他所有进程的**recvcounts[j]**和**recvtype**相同

MPI_Allgather vs MPI_Allgatherv

Allgather, NPROCS=3



Allgatherv, NPROCS=3



实例:MPI_Allgather

```
MPI_Comm comm;  
int size, sendarray[100], *rbuf;  
.....  
MPI_Comm_size(comm, &size);  
rbuf = (int *)malloc(size * 100 * sizeof(int));
```

```
MPI_Allgather(sendarray, 100, MPI_INT,  
              rbuf, 100, MPI_INT, comm);
```

实例:MPI_Allgatherv

```
int size, sendarray[100], *rbuf, *displs, i, *rcounts;
int stride = 120;
MPI_Comm_size(comm, &size);

rbuf = (int *) malloc(size * stride * sizeof(int));
displs = (int *) malloc(size * sizeof(int)); /*申请缓冲区*/
rcounts = (int *) malloc(size * sizeof(int));
for (i=0; i<size; ++i) {
    displs[i] = i * stride;   rcounts[i] = 100;}

MPI_Allgatherv(sendarray, 100, MPI_INT,
               rbuf, rcounts, displs, MPI_INT, comm);
```

MPI全散发收集函数

- 每个进程散发自己的一个数据块, 并且收集拼装所有进程散发过来的数据块。称该操作为数据的“**全散发收集**”。
- 它既可以被认为是数据全收集的扩展, 也可以被认为是数据散发的扩展

MPI_Alltoall函数

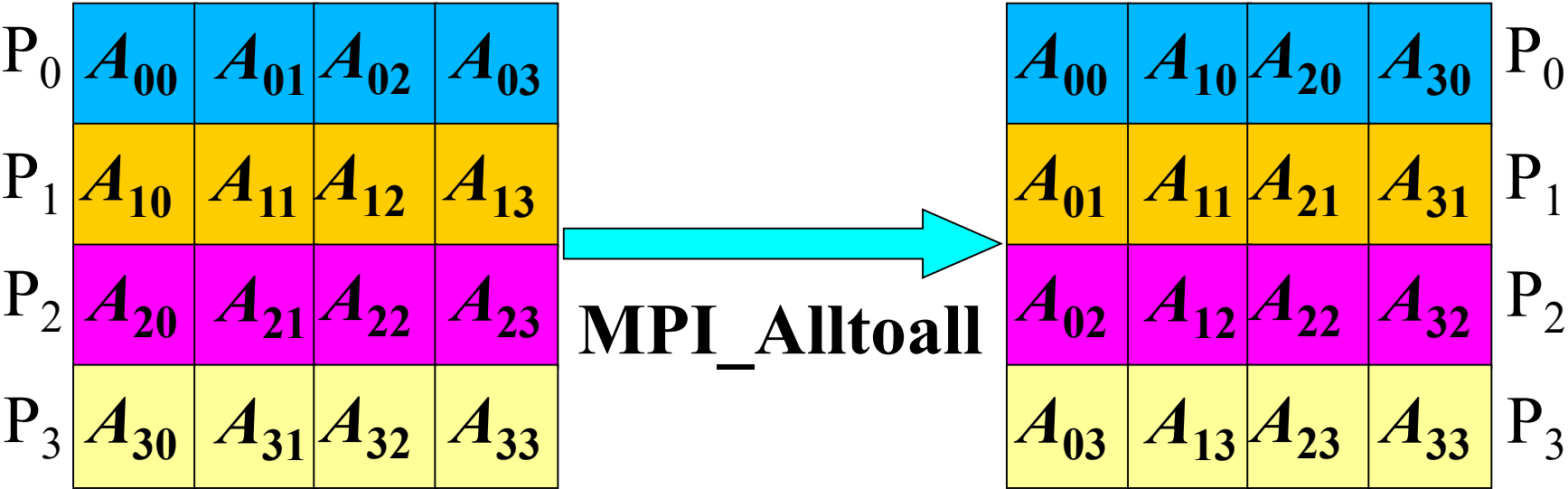
```
int MPI_Alltoall(  
    void*,  
    int,  
    MPI_Datatype,  
    void*,  
    int,  
    MPI_Datatype,  
    MPI_Comm,  
    sendbuf,  
    sendcount,  
    sendtype,  
    recvbuf,  
    recvcount,  
    recvtype,  
    comm,  
    /* in */,  
    /* in */,  
    /* in */,  
    /*out*/,  
    /* in */,  
    /* in */,  
    /* in */)
```

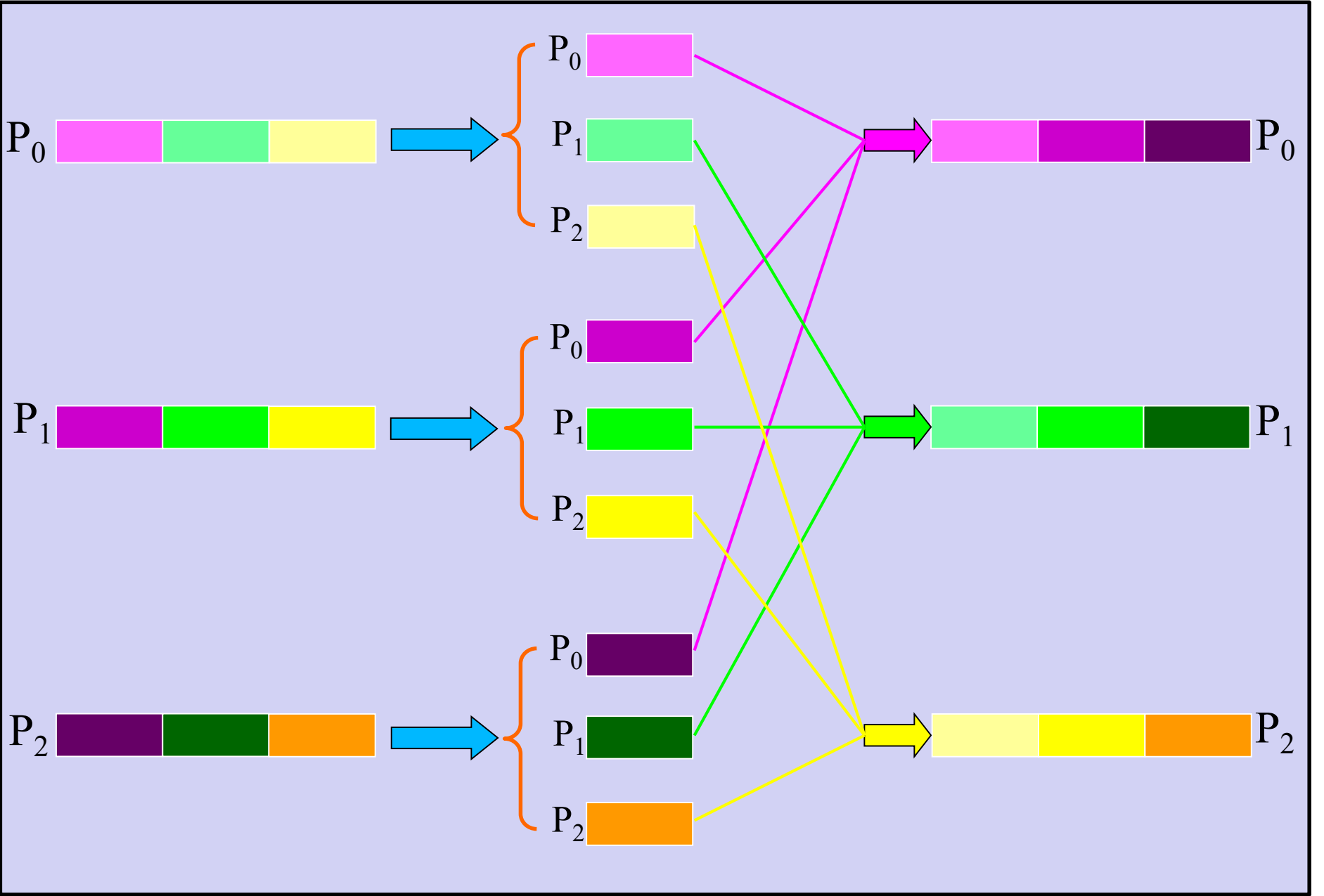
MPI Alltoall函数详解

- MPI Alltoall是组内进程之间完全的消息交换，每一个进程都向其它所有的进程发送消息，同时每一个进程都从其它所有的进程接收消息。
 - MPI_Allgather每个进程散发一个相同的消息给所有的进程
 - MPI_Alltoall散发给不同进程的消息是不同的。因此它的发送缓冲区也是一个数组
- 调用MPI Alltoall相当于每个进程依次将它的发送缓冲区的第i块数据发送给第i个进程，同时每个进程又都依次从第j个进程接收数据放到各自接收缓冲区的第j块数据区的位置。

MPI_Alltoall函数详解

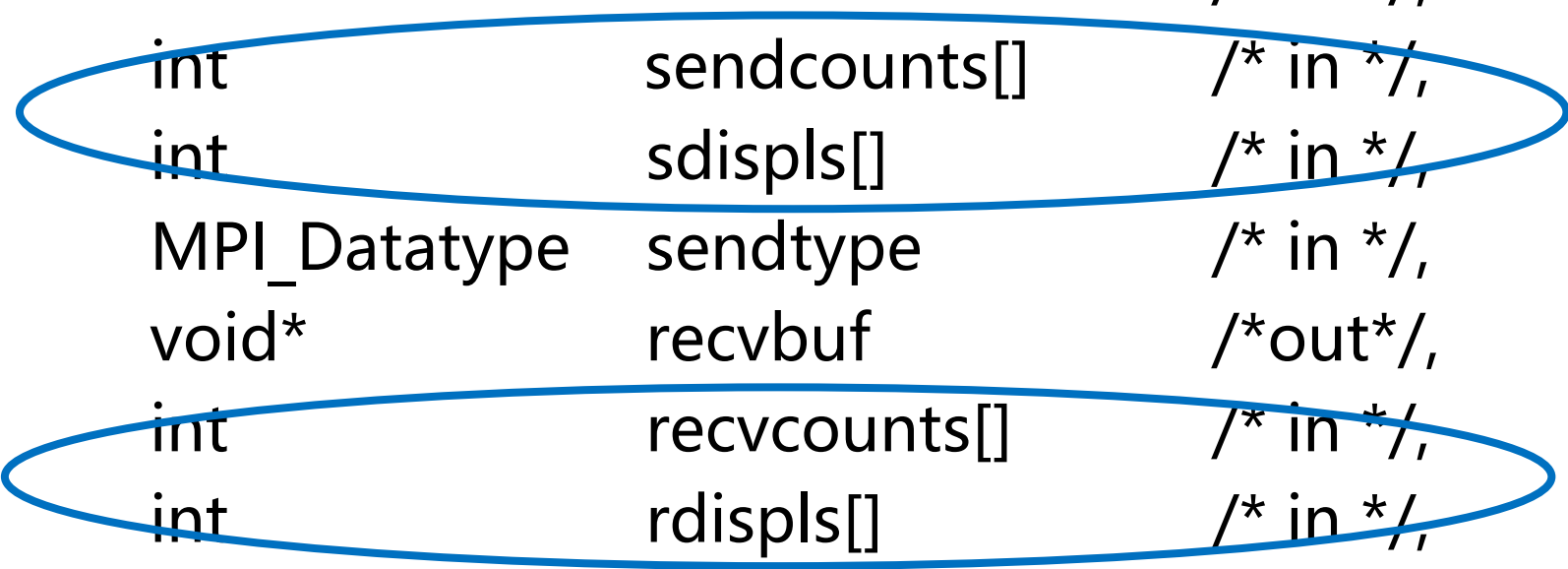
- MPI_Alltoall的每个进程可以向每个接收者发送数目不同的数据，第i个进程发送的第j块数据将被第j个进程接收，并存放在其接收消息缓冲区recvbuf的第i块。
- 每个进程的sendcount和sendtype的类型必须和所有其他进程的recvcount和recvtype相同这就意味着在每个进程和根进程之间发送的数据量必须和接收的数据量相等。





MPI_Alltoallv函数

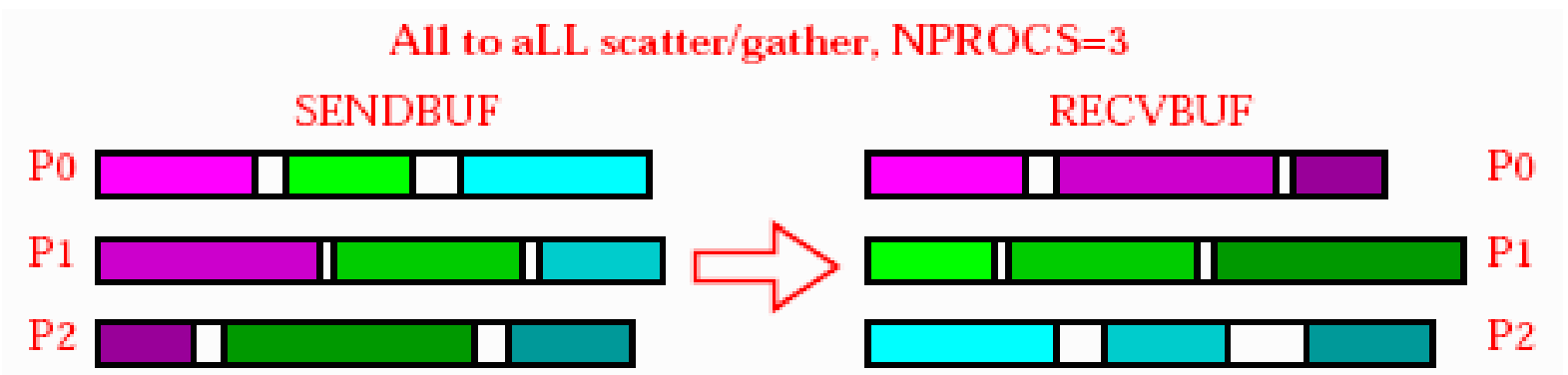
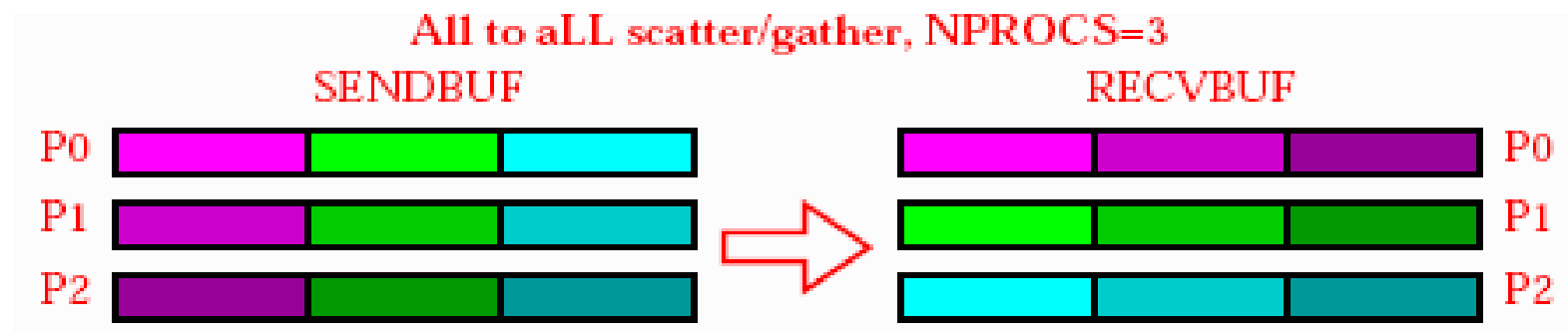
```
int MPI_Alltoallv(  
    void*          sendbuf          /* in */,  
    int            sendcounts[]      /* in */,  
    int            sdispls[]         /* in */,  
    MPI_Datatype    sendtype         /* in */,  
    void*          recvbuf          /* out */,  
    int            recvcnts[]        /* in */,  
    int            rdispls[]         /* in */,  
    MPI_Datatype    recvtype         /* in */,  
    MPI_Comm        comm             /* in */)
```

The diagram shows the function signature of MPI_Alltoallv. Two blue ovals are drawn around the parameters. The first oval encircles the send parameters: int sendcounts[], int sdispls[], MPI_Datatype sendtype, and void* recvbuf. The second oval encircles the receive parameters: int recvcnts[], int rdispls[], MPI_Datatype recvtype, and MPI_Comm comm. The first oval is slightly larger and more prominent than the second.

MPI_Alltoallv函数详解

- 正如MPI_Allgatherv 和MPI_Allgather 的关系一样 MPI_Alltoallv在MPI_Alltoall的基础上进一步增加了灵活性。它可以由sdispls指定待发送数据的位置，在接收方则由rdispls指定接收的数据存放在缓冲区的偏移量
- 所有参数对每个进程都是有意义的，并且所有进程中的comm值必须一致
- MPI_Alltoall和MPI_Alltoallv可以实现n次独立的点对点通信但也有限制：1)所有数据必须是同一类型；2)所有的消息必须按顺序进行散发和收集

MPI Alltoall vs MPI Alltoallv

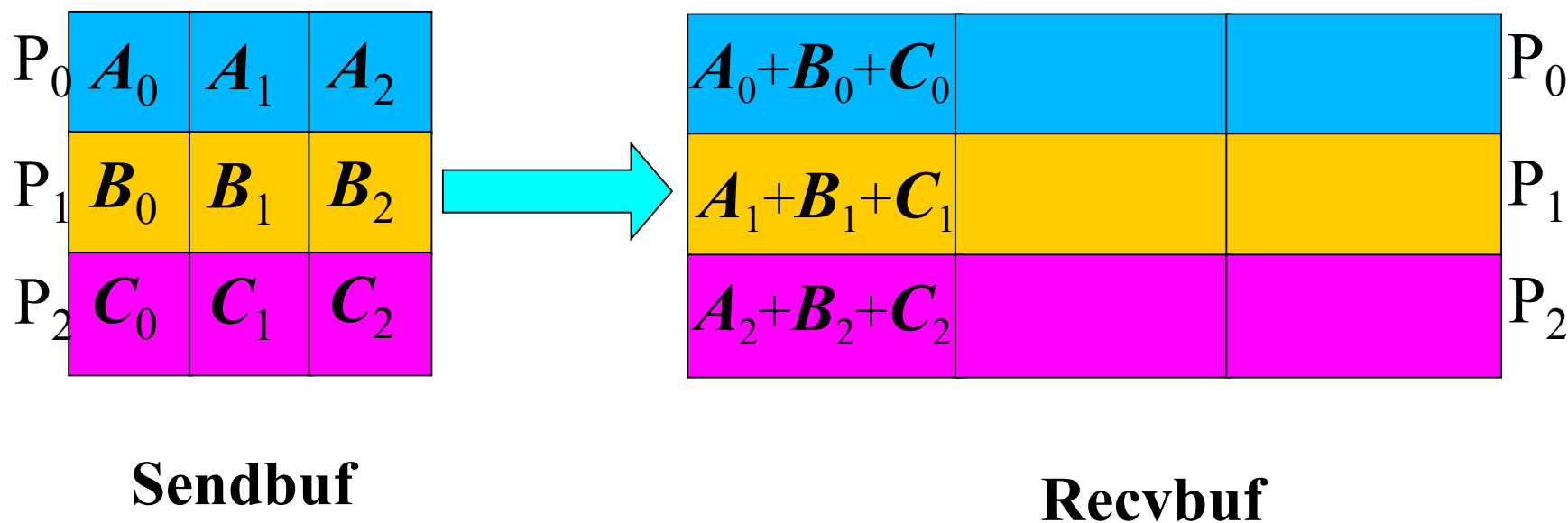


归约散发函数MPI Reduce scatter

```
int MPI_Reduce_scatter (  
    void*          sendbuf          /* in */,  
    void*          recvbuf          /* in */,  
    int*           recvcounts       /* in */,  
    MPI_Datatype    datatype        /* in */,  
    MPI_Op          operator        /* out */,  
    MPI_Comm        comm            /* in */)
```

MPI_Reduce_scatter函数图示

MPI_Reduce_scatter: np = 3; count = 3;
Op = MPI_SUM



MPI_Reduce_scatter函数详解

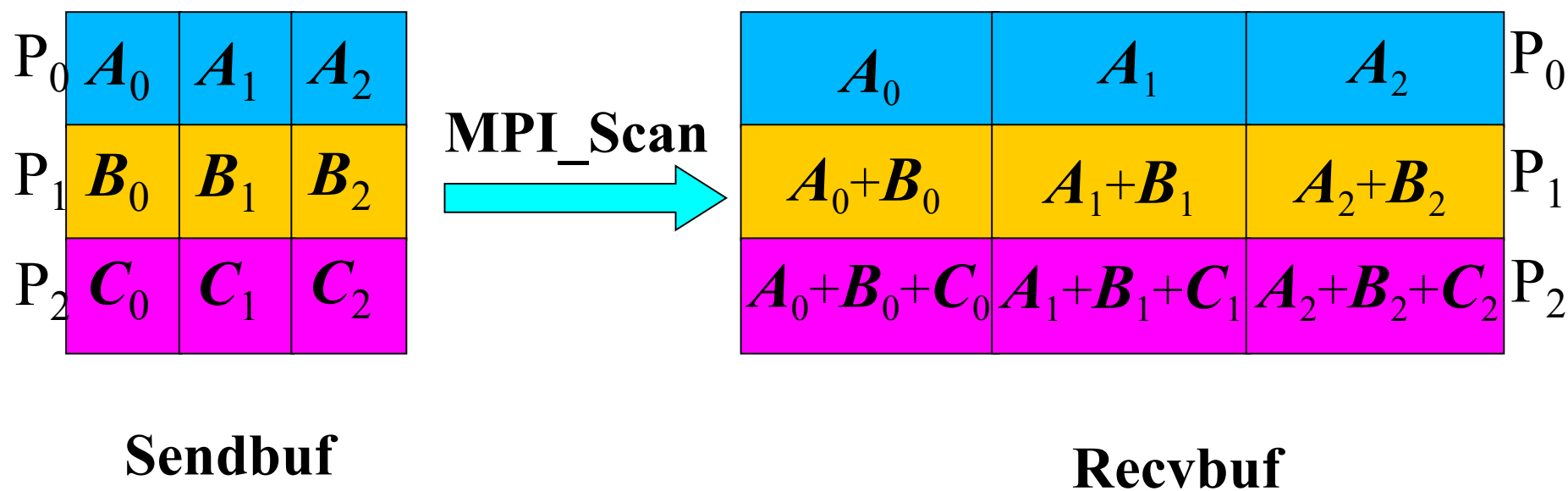
- MPI_Reduce_scatter可以认为是MPI对每个归约操作的变形，它将归约结果分散到组内的所有进程中去而不是仅仅归约到root进程
- MPI_Reduce_scatter对由sendbuf、count和datatype定义的发送缓冲区数组的元素逐个进行归约操作，发送缓冲区数组的长度count= recvcnt[i]
- 然后再对归约结果进行散发操作，散发给第*i*个进程的数据块长度为recvcnts(i). 其余参数的含义与MPI_Reduce一样.

前缀扫描函数MPI Scan

```
int MPI_Scan (  
    void*          Sendbuf      /* in */,  
    void*          Recvbuf      /* in */,  
    int            count        /* in */,  
    MPI_Datatype    datatype    /* in */,  
    MPI_Op          operator     /*out*/,  
    MPI_Comm        comm        /* in */)
```

MPI_Scan函数图示

MPI_Scan: np = 3; count = 3; Op = MPI_SUM



MPI_Scan函数详解

- **MPI_Scan**前缀扫描, 或前缀归约, 与归约**MPI_Reduce**操作类似, 但各处理器依次得到部分归约结果。
- 确切地说, 操作结束后第 i 个处理器的recvbuf 中将包含前 i 个处理器的归约运算结果。
- 各参数的含义与**MPI_Allreduce** 基本相同

MPI集合通信函数

类型	函数	功能
数据移动	MPI_Bcast	一到多，数据广播
	MPI_Gather	多到一，数据汇合
	MPI_Gatherv	MPI_Gather的一般形式
	MPI_Allgather	MPI_Gather的一般形式
	MPI_Allgatherv	MPI_Allgather的一般形式
	MPI_Scatter	一到多，数据分散
	MPI_Scatterv	MPI_Scatter的一般形式
	MPI_Alltoall	多到多，置换数据
	MPI_Alltoallv	MPI_Alltoall的一般形式
数据收集	MPI_Reduce	多到一，数据归约
	MPI_Allreduce	上者的一般形式，结果在所有进程
	MPI_Reduce_scatter	结果scatter到各个进程
	MPI_Scan	前缀操作
同步	MPI_Barrier	同步操作

-
- **All**表示最后的结果存放到所有的进程中,
MPI_Allgather、**MPI_Alltoall**、**MPI_Allreduce**
 - **V:Vector**,操作以及被操作的数据对象更加灵活,
MPI_Gather(v)、**MPI_Scatter(v)**、
MPI_Allgather(v)、**MPI_Alltoall(v)**

Exercises: Collectives

■ Directory Collectives

- **Matrix vector multiplication**
- optional: computePi: compute the value of PI in parallel
- optional: average: same as reduce exercise, but now have to write it yourself.

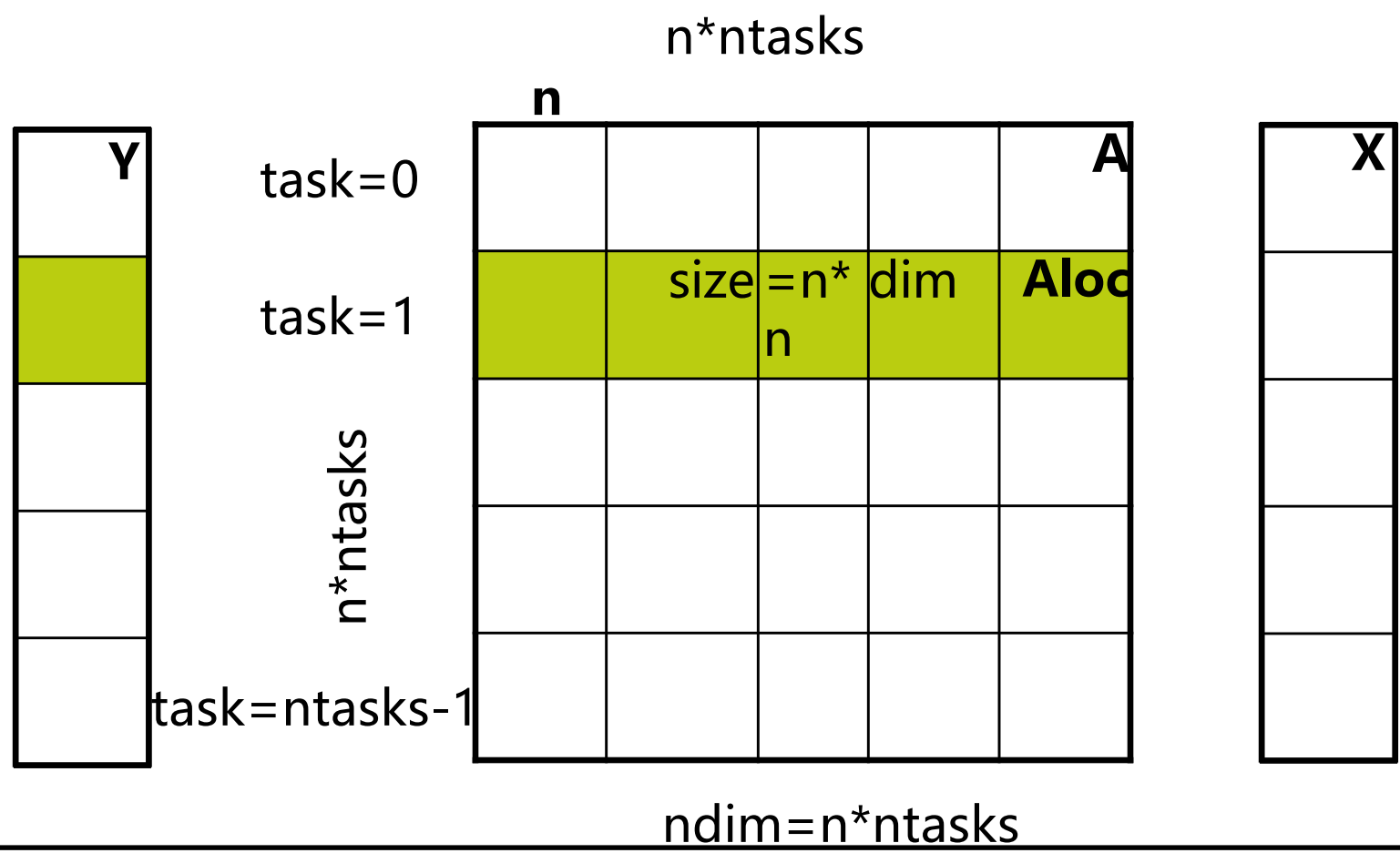
矩阵-向量乘法

□ $A = (a_{ij})_{n \times n}$ X, Y 为 n 维向量, $Y = A \times X$

$$Ax = \begin{pmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & & a_{2,n} \\ \vdots & \vdots & & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,n} \end{pmatrix} \cdot \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{pmatrix} a_{1,1}x_1 + a_{1,2}x_2 + \cdots + a_{1,n}x_n \\ a_{2,1}x_1 + a_{2,2}x_2 + \cdots + a_{2,n}x_n \\ \cdots \\ a_{n,1}x_1 + a_{n,2}x_2 + \cdots + a_{n,n}x_n \end{pmatrix}$$

$$y_i = \sum_{j=0}^{n-1} a_{ij} x_j$$

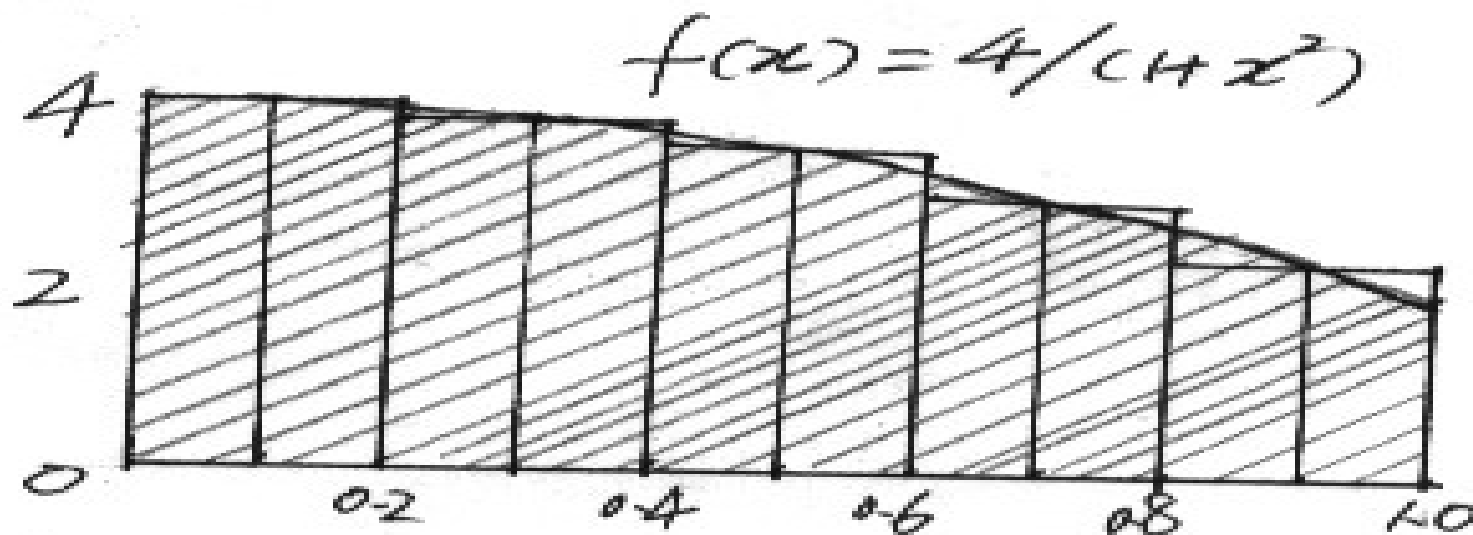
Collectives: Matrix-Vector



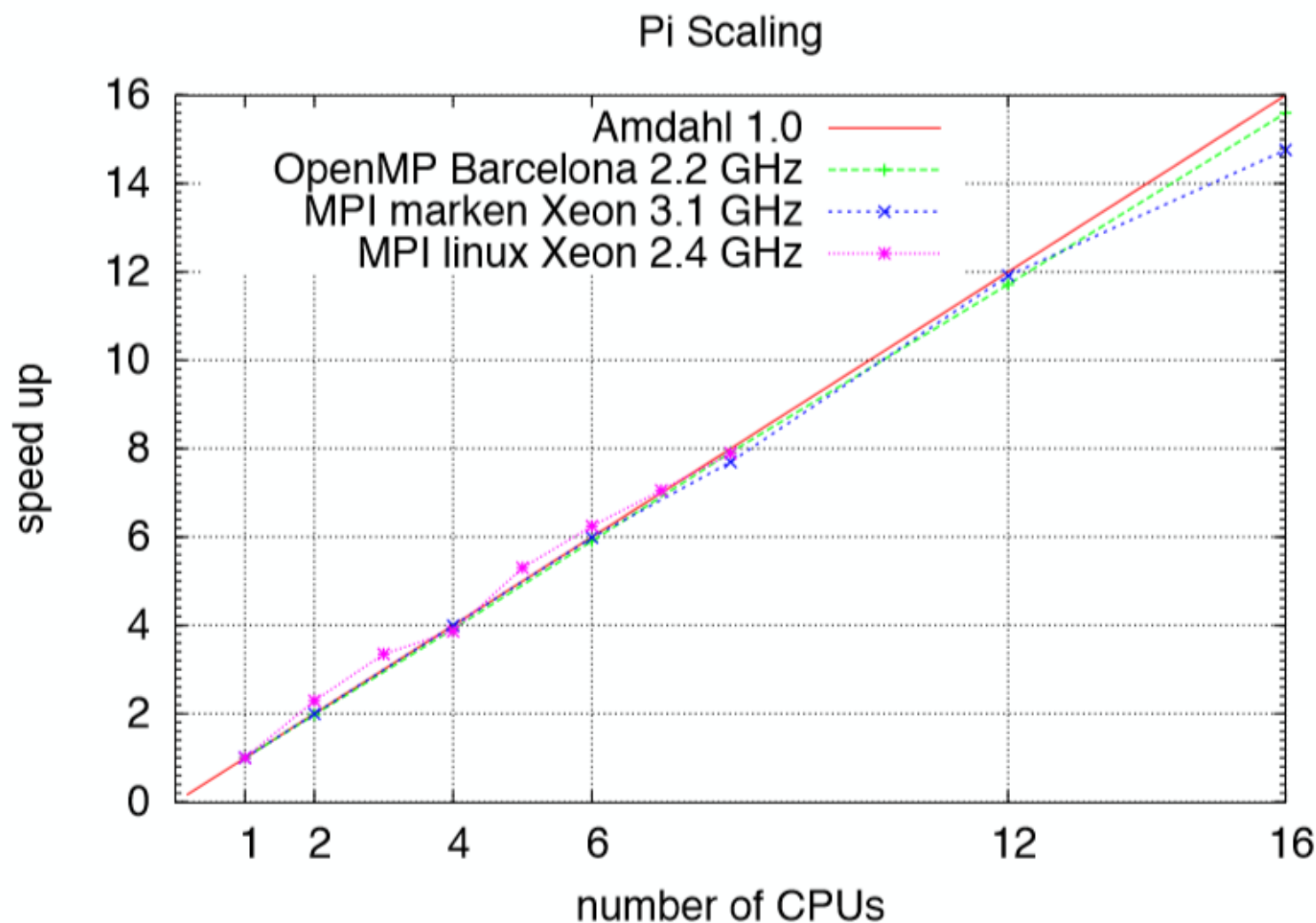
计算圆周率 π

$$\int_0^1 \frac{1}{1+x^2} dx = \arctan(x) \Big|_0^1 = \arctan(1) - \arctan(0) = \arctan(1) = \frac{\pi}{4}$$

令函数 $f(x) = 4/(1+x^2)$, 则: $\int_0^1 f(x) dx = \pi$



Exercise: PI with MPI and OpenMP



THANKS