



\Rightarrow D АЛГОРИТМЫ

Занятие 1. Технология MPI



Составитель: Герасимов А.С.

Учебный кластер МФТИ

head.vdi.mipt.ru

remote.vdi.mipt.ru:52960

ssh login@head.vdi.mipt.ru

- Узлы: 1 головной (head) и 7 вычислительных
- Узлы идентичны: 4 ядра, 15 ГБ ОЗУ
- Система очередей – Torque/PBS

Пример PBS-задачи

job.sh

```
#!/bin/bash
```

```
#PBS -l walltime=00:10:00,nodes=7:ppn=1
```

```
#PBS -N job_name
```

```
#PBS -q batch
```

```
uname -n
```

Запуск задачи

```
qsub job.sh
```

Выход задачи:

- `<job_name>.o<ID>` – выход stdout
- `<job_name>.e<ID>` – выход stderr

Ограничения:

- 5 заданий / пользователя
- 10 минут выполнения
- 1 ГБ памяти

Просмотр текущих задач в очереди

qstat

```
[kolya@head mpi]$ qstat
```

Job id	Name	User	Time Use	S	Queue
-----	-----	-----	-----	-	-----
25.localhost	my_job	kolya	0	R	batch
26.localhost	my_job	kolya	0	R	batch
27.localhost	my_job	kolya	0	R	batch
28.localhost	my_job	kolya	0	R	batch
29.localhost	my_job	kolya	0	R	batch

Удаление задачи

qdel <ID>

MPI (Message Passing Interface)

- Библиотека функций, предназначенная для поддержки работы параллельных процессов.
- Базовый механизм связи между процессами – передача и приём сообщений.
- Ориентирован на системы с распределенной памятью
- Состав сообщений:
 - *отправитель — ранг (номер в группе) отправителя;*
 - *получатель — ранг получателя;*
 - *признак(и);*
 - *коммуникатор — код группы процессов.*
- Блокирующие / неблокирующие передачи

Общие процедуры MPI. Инициализация

```
int MPI_Init (int* argc, char*** argv)
```



- MPI_SUCCESS
- код ошибки



Аргументы функции main()

```
int MPI_Finalize (void)
```

Общие процедуры MPI. Инициализация

Основа программы

```
#include "mpi.h"
```

```
int main(int argc, char** argv) {
```

```
    MPI_Init(&argc, &argv);
```

```
    ...
```

```
    MPI_Finalize();
```

```
}
```


Общие процедуры MPI. Размер группы

```
int MPI_Comm_size  
    (MPI_Comm comm, int* size)
```




- Коммуникатор группы
- MPI_COMM_WORLD



(OUT) Размер группы

Общие процедуры MPI. Ранг процесса

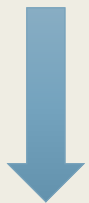
```
int MPI_Comm_rank  
    (MPI_Comm comm, int* rank)
```

- 
- Коммуникатор группы
 - MPI_COMM_WORLD

(OUT) Номер процесса в группе [0; size-1]

Общие процедуры MPI. Подсчет времени

```
double MPI_Wtime (void)
```



Некоторое время в секундах

Общие процедуры MPI. Оценка ускорения

Закон Амдала

$$a \leq \frac{1}{(1 - p) + \frac{p}{n}}$$

a – оценка ускорения

p – распараллеливаемая часть программы (доля общего времени выполнения)

n – количество процессов

Компиляция программы

```
mpicc superhot.c -o hot
```

Запуск программы

```
mpirun -np <thread_num> ./hot
```

job.sh

#!/bin/bash

#PBS -l walltime=00:01:00,nodes=1:ppn=3

#PBS -N my_job

#PBS -q batch

cd \$PBS_O_WORKDIR

mpirun --hostfile \$PBS_NODEFILE -np 3 ./hot

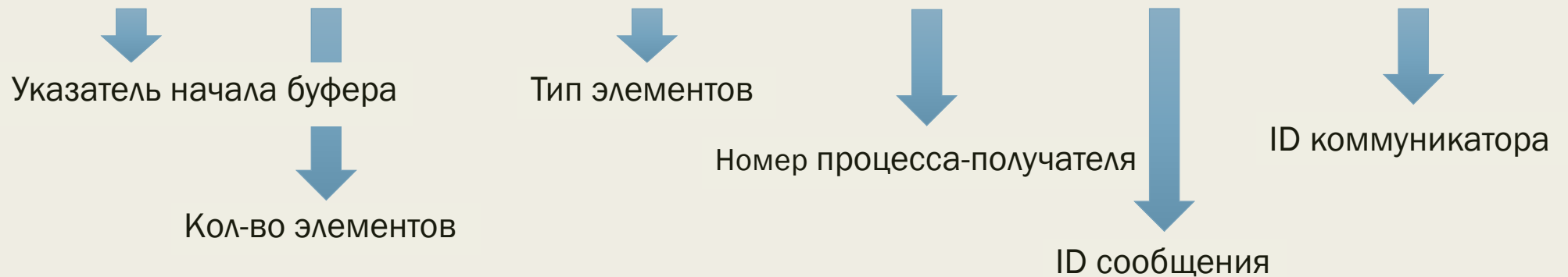
Общие процедуры MPI. Задача 1

- Составить и запустить программу «Hello, world!»
- Вывести размер своего коммуникатора и своего процесса

Общие процедуры MPI. Блокирующие передачи

int MPI_Send

(void* buf, int count, MPI_Datatype datatype, int dest, int msgtag, MPI_Comm comm)

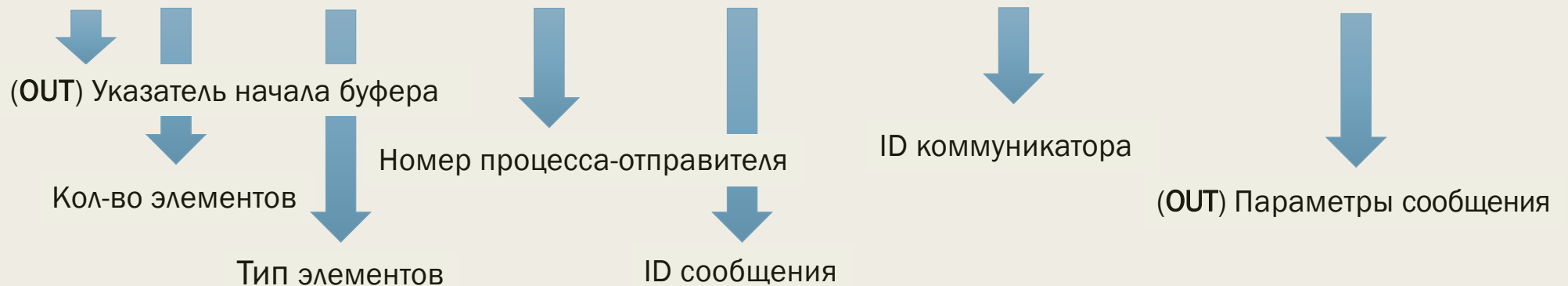


- Все элементы сообщения расположены подряд в буфере
- Значение count может быть нулем
- Тип элементов должен указываться с помощью констант типа
- Разрешается передавать сообщение самому себе
- Никаких гарантий передачи/приема

Общие процедуры MPI. Блокирующие передачи

int MPI_Recv

(void*, int, MPI_Datatype, int source, int msgtag, MPI_Comm comm, MPI_Status *status)



- Число элементов в сообщении не должно превосходить count
- Если нужно узнать точное число элементов в сообщении → MPI_Probe
- Гарантия, что после возврата все элементы сообщения приняты и расположены в *buf*
- В качестве *source* можно указать predetermined константу MPI_ANY_SOURCE
- В качестве *msgtag* можно указать константу MPI_ANY_TAG
- Из двух приходящих сообщений выбирается то, что отправлено раньше

Общие процедуры MPI. Предопределенные константы

Константа MPI	Тип в C
<i>MPI_CHAR</i>	signed char
<i>MPI_SHORT</i>	signed int
<i>MPI_INT</i>	signed int
<i>MPI_LONG</i>	signed long int
<i>MPI_UNSIGNED_CHAR</i>	unsigned char
<i>MPI_UNSIGNED_SHORT</i>	unsigned int
<i>MPI_UNSIGNED</i>	unsigned int
<i>MPI_UNSIGNED_LONG</i>	unsigned long int
<i>MPI_FLOAT</i>	float
<i>MPI_DOUBLE</i>	double
<i>MPI_LONG_DOUBLE</i>	long doubl

Общие процедуры MPI. Предопределенные константы

MPI_Status - атрибуты сообщений

MPI_Source (номер процесса-отправителя)

MPI_Tag (ID сообщения)

MPI_Error (код ошибки)

Константы-пустышки

MPI_COMM_NULL

MPI_DATATYPE_NULL

MPI_REQUEST_NULL

Код успешного завершения процедуры

MPI_SUCCESS

Общие процедуры MPI. Асинхронные передачи

```
int MPI_Isend
```

```
(void*, int, MPI_Datatype, int, int, MPI_Comm, MPI_Request *request)
```



(OUT) ID операции

- Возврат происходит сразу после инициализации
- Нельзя повторно использовать буфер для других целей без получения информации о завершении посылки
- Окончание процесса передачи можно определить с помощью *request* и процедур MPI_Wait и MPI_Test
- Сообщение, отправленное любой из процедур MPI_Send и MPI_Isend, может быть принято любой из процедур MPI_Recv и MPI_Irecv

Общие процедуры MPI. Асинхронные передачи

int MPI_Irecv

(void*, int, MPI_Datatype, int, int, MPI_Comm, MPI_Request *request)



(OUT) Указатель начала буфера



(OUT) ID операции

- Возврат происходит сразу после инициализации
- Окончание процесса можно определить с помощью *request* и процедур MPI_Wait и MPI_Test
- Сообщение, отправленное любой из процедур MPI_Send и MPI_Isend, может быть принято любой из процедур MPI_Recv и MPI_Irecv

Общие процедуры MPI. Задача 2

- Составить параллельную программу, суммирующую все натуральные числа от 1 до N
- Каждый процесс получает свой диапазон чисел для суммирования
- Главный процесс выводит результат на экран.
- N задается аргументом запуска
- Вести подсчет времени выполнения всей программы + времени подсчета на каждом процессе

Общие процедуры MPI. Задача 3

- Составить параллельную программу, в которой массив данных (размером NP) передается по кольцевой топологии.
- Каждый процесс выводит на экран элемент массива, соответствующий своему номеру процесса.
- Значения массива задаются аргументом запуска
- Реализовать передачу по топологии «звезда» с обратной связью

Общие процедуры MPI. Коллективные взаимодействия

```
int MPI_Bcast  
(void *, int count, MPI_Datatype, int source, MPI_Comm)
```



- Рассылка сообщения происходит от *source* всем процессам, включая рассылающий процесс
- При возврате из процедуры содержимое *buf* процесса *source* будет скопировано в локальный буфер процесса
- Значения параметров *count*, *datatype* и *source* должны быть одинаковыми у всех процессов.

Общие процедуры MPI. Коллективные взаимодействия

MPI_Scatter

(void* send_buf, int send_count, MPI_Datatype,



Буфер отправки



Кол-во отправляемых элементов
в пакете



Тип отправляемых элементов

void* recv_buf, int recv_count, MPI_Datatype, int root, MPI_Comm)



(OUT) Буфер приема



Кол-во принимаемых элементов
в пакете



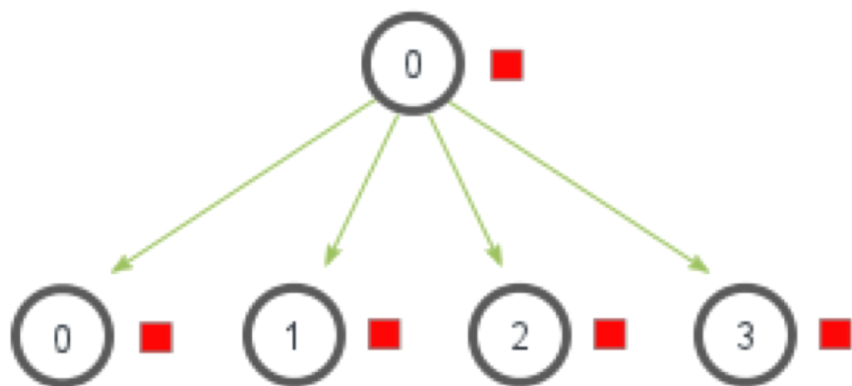
Тип принимаемых элементов



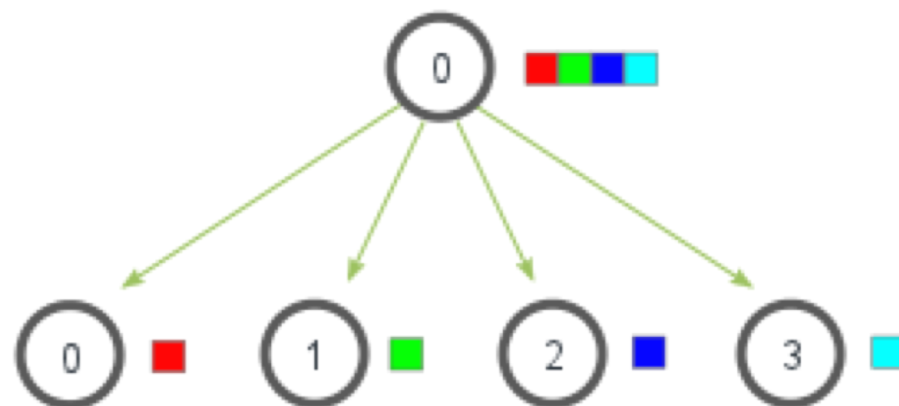
Номер рассылающего процесса

Общие процедуры MPI. Коллективные взаимодействия

MPI_Bcast



MPI_Scatter



Общие процедуры MPI. Коллективные взаимодействия

MPI_Gather

(void* send_buf, int send_count, MPI_Datatype,



Буфер отправки



Кол-во отправляемых элементов
в пакете



Тип отправляемых элементов

void* recv_buf, int recv_count, MPI_Datatype, int dest, MPI_Comm)



(OUT) Буфер приема



Кол-во принимаемых элементов
в пакете



Тип принимаемых элементов



Номер принимающего процесса

Общие процедуры MPI. Коллективные взаимодействия

MPI_Gather

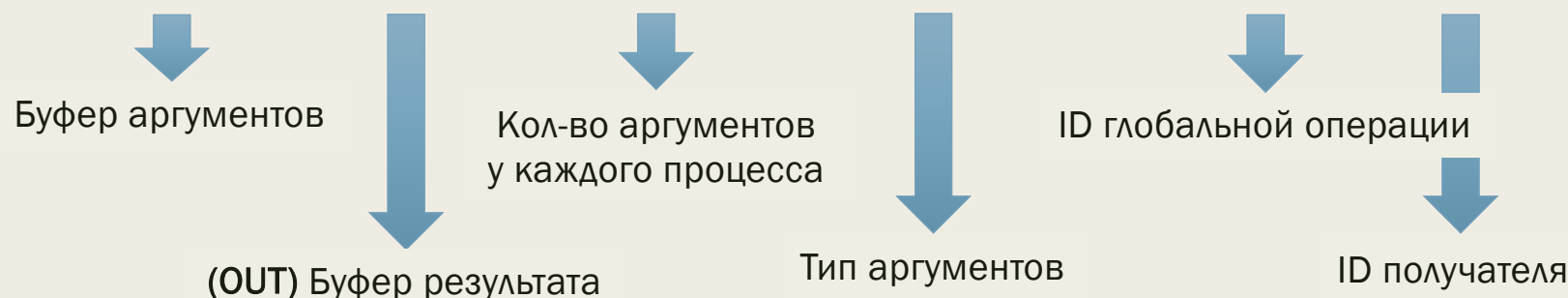
```
(void* send_buf, int send_count, MPI_Datatype, void*  
recv_buf, int recv_count, MPI_Datatype, int dest, MPI_Comm)
```

- Собирающий процесс сохраняет данные в *recv_buf*, располагая их в порядке возрастания номеров процессов
- Параметр *recv_buf* имеет значение только на собирающем процессе и на остальных игнорируется
- Значения параметров *count*, *datatype* и *dest* должны быть одинаковыми у всех процессов

Общие процедуры MPI. Коллективные взаимодействия

int MPI_Reduce

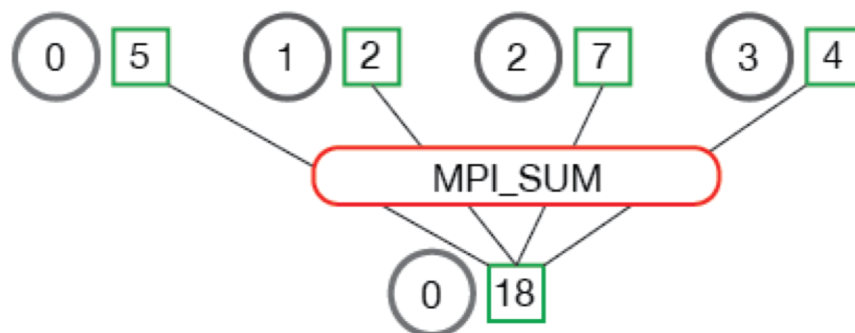
(void *sbuf, void *rbuf, int count, MPI_Datatype, MPI_Op op, int root, MPI_Comm)



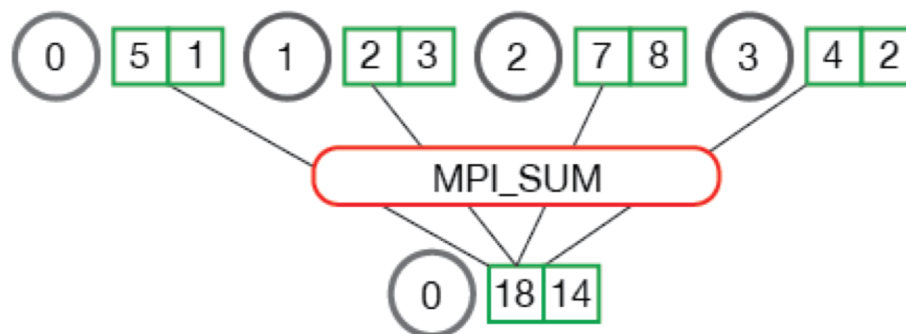
- Выполнение *count* глобальных операций *op* с возвратом результата в буфер *rbuf* процесса *root*
- Операция выполняется независимо над соответствующими аргументами всех процессов.
- Значения *count* и *datatype* у всех процессов должны быть одинаковыми
- Из соображений эффективности реализации предполагается, что операция *op* обладает свойствами ассоциативности и коммутативности

Общие процедуры MPI. Коллективные взаимодействия

MPI_Reduce



MPI_Reduce



Общие процедуры MPI. Задача 4

- Составить параллельную программу, вычисляющую сумму ряда (вычисление экспоненты) с использованием операций коллективного взаимодействия
- Число слагаемых задается аргументом запуска
- Вести подсчет времени выполнения всей программы + времени подсчета на каждом процессе
- Вычисление числа Π
- Вычисление более сложного ряда