Лекция 8 Параллельные сеточные вычисления

Курносов Михаил Георгиевич

E-mail: mkurnosov@gmail.com WWW: www.mkurnosov.net

Курс «Параллельные вычислительные технологии» Сибирский государственный университет телекоммуникаций и информатики (г. Новосибирск) Осенний семестр, 2019



Расчет стационарного распределения тепла

- С течением времени в теле устанавливается некоторое не зависящее от времени распределение температуры (тепловое состояние выходит на стационарный режим)
- Распределение температуры в таком случае описывается уравнением теплопроводности
- Стационарное двумерное уравнение Лапласа (Laplace equation)

$$\Delta U = \frac{d^2 U}{dx^2} + \frac{d^2 U}{dy^2} = \mathbf{0} \tag{1}$$

- Функция U(x, y) неизвестный потенциал (теплота)
- Задано уравнение (1), значения функции U(x, y) на границах расчетной 2D-области
- Требуется найти значение функции U(x, y) во внутренних точках расчетной 2Dобласти

Расчет стационарного распределения тепла

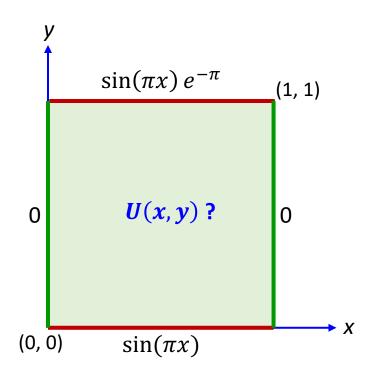
■ Найти решение стационарного двумерного уравнения Лапласа (Laplace equation)

$$\frac{d^2U}{dx^2} + \frac{d^2U}{dy^2} = \mathbf{0}$$

- Расчётная область (domain) квадрат [0, 1] x [0, 1]
- Граничные условия (boundary conditions):
 - $U(x,0) = \sin(\pi x)$

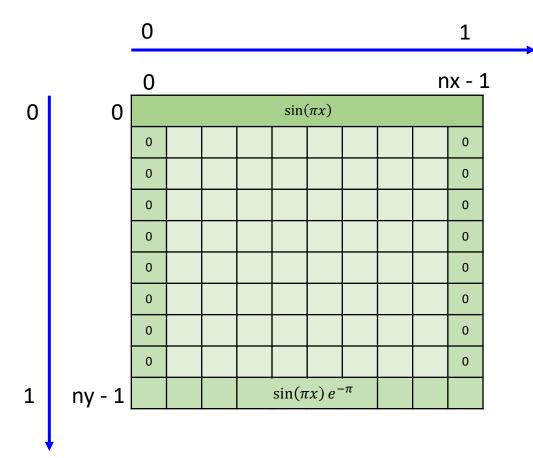
 - $\Box U(0,y) = U(1,y) = 0$
- Для данной задачи известно аналитическое решение

$$U(x,y) = \sin(\pi x) e^{-\pi y}$$



Дискретизация расчетной области

■ Расчетная область [0, 1] x [0, 1] покрывается прямоугольной сеткой с постоянным шагом: пх точек по оси ОХ и пу точек по оси ОҮ



- Расчетная сетка массив [пу, пх] чисел (температура)
- Переход от индекса ячейки [i, j] к координатам в области [0, 1] х [0, 1]:

$$x = j * 1.0 / (nx - 1.0)$$

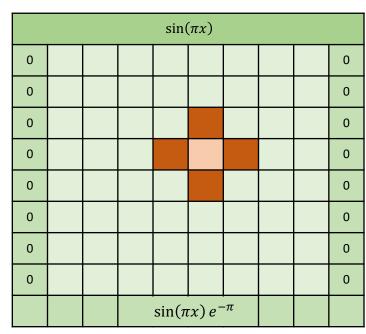
 $y = i * 1.0 / (ny - 1.0)$

Разностная аппроксимация оператора Лапласа

 Вторые производные аппроксимируются на расчетной сетке разностным уравнением с применением четырехточечного шаблона

$$\Delta U = \frac{d^2 U}{dx^2} + \frac{d^2 U}{dy^2} = 0$$

 Новое значение в каждой точке сетки равно среднему из предыдущих значений четырех ее соседних точек (схема «крест»)



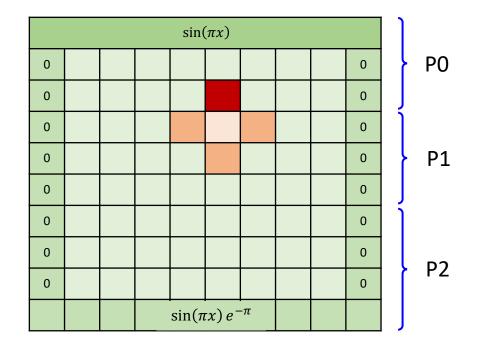
- Вычисляем новое значение в каждой точке [i, j] сетки среднее из предыдущих значений четырех ее соседних точек (схема «крест»), результат записываем в новую сетку (массив)
- На следующей итерации текущей делаем новую сетку предыдущей итерации
- Заканчиваем итерационный процесс, если разность между каждым текущим и предыдущим значениями по модулю не больше EPSILON

```
#define EPS 0.001
#define PI 3.14159265358979323846
#define IND(i, j) ((i) * nx + (j))
int main(int argc, char *argv[])
    int rows = 100;
    int cols = 100;
    int ny = rows;
    int nx = cols;
    double *local grid = xcalloc(ny * nx, sizeof(*local grid));
    double *local newgrid = xcalloc(ny * nx, sizeof(*local newgrid));
    double dx = 1.0 / (nx - 1.0);
    // Initialize top border: u(x, 0) = \sin(pi * x)
    for (int j = 0; j < nx; j++) {
        int ind = IND(0, j);
        local newgrid[ind] = local grid[ind] = sin(PI * dx * j);
    // Initialize bottom border: u(x, 1) = \sin(pi * x) * \exp(-pi)
    for (int j = 0; j < nx; j++) {
        int ind = IND(ny - 1, j);
        local newgrid[ind] = local grid[ind] = sin(PI * dx * j) * exp(-PI);
```

```
int niters = 0;
for (;;) {
   niters++;
   for (int i = 1; i < ny - 1; i++) { // Update interior points
       for (int j = 1; j < nx - 1; j++) {
            local newgrid[IND(i, j)] =
                (local grid[IND(i - 1, j)] + local grid[IND(i + 1, j)] +
                 local grid[IND(i, j - 1)] + local grid[IND(i, j + 1)]) * 0.25;
    // Check termination condition
    double maxdiff = 0;
   for (int i = 1; i < ny - 1; i++) {
       for (int j = 1; j < nx - 1; j++) {
            int ind = IND(i, j);
            maxdiff = fmax(maxdiff, fabs(local_grid[ind] - local_newgrid[ind]));
   // Swap grids (after termination local grid will contain result)
   double *p = local grid;
    local_grid = local_newgrid;
   local newgrid = p;
    if (maxdiff < EPS)</pre>
       break;
ttotal += wtime();
```

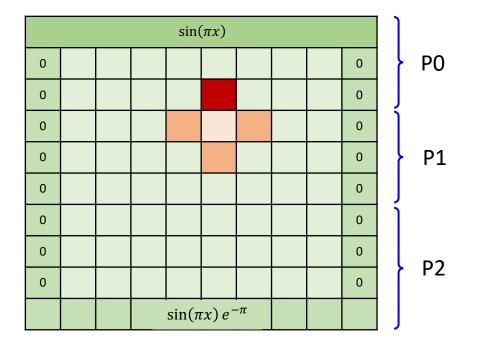
```
printf("# Heat 2D (serial): grid: rows %d, cols %d\n", rows, cols);
printf("# niters %d, total time %.6f\n", niters, ttotal);
// Save grid
if (filename) {
    FILE *fout = fopen(filename, "w");
    if (!fout) {
        perror("Can't open file");
        exit(EXIT FAILURE);
    for (int i = 0; i < ny; i++) {
        for (int j = 0; j < nx; j++)
            fprintf(fout, "%.4f", local grid[IND(i, j)]);
        fprintf(fout, "\n");
    fclose(fout);
return 0;
```

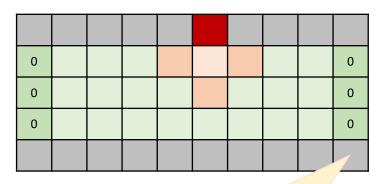
- Одномерная (1D) декомпозиция расчётной области на горизонтальные полосы
- Каждому процессу назначается ny / р строк расчетной сетки
- Сетка хранится в памяти в распределенном виде
- Проблема для расчета значений некоторых точек требуются данные соседних полос, которые находятся в памяти других процессов



- На каждой итерации перед вычислением значений в точках обмениваемся пограничными строками с соседними процессами
- Как хранить эти строки?
- Как выполнять обмены?

- Одномерная (1D) декомпозиция расчётной области на горизонтальные полосы
- Каждому процессу назначается ny / р строк расчетной сетки
- Сетка хранится в памяти в распределенном виде
- Проблема для расчета значений некоторых точек требуются данные соседних полос, которые находятся в памяти других процессов





Теневые ячейки (halo, ghost cells) для хранения строк соседних процессов

```
#define NELEMS(x) (sizeof((x)) / sizeof((x)[0]))
#define IND(i, j) ((i) * cols + (j))
int get block size(int n, int rank, int nprocs)
    int s = n / nprocs;
    if (n % nprocs > rank) s++;
    return s;
int main(int argc, char *argv[])
    int commsize, rank;
    MPI_Init(&argc, &argv);
    double ttotal = -MPI Wtime();
    MPI_Comm_size(MPI_COMM_WORLD, &commsize);
    MPI_Comm_rank(MPI_COMM_WORLD, &rank);
    int rows, cols; // Broadcast command line arguments
    if (rank == 0) {
        rows = (argc > 1) ? atoi(argv[1]) : commsize * 100;
        cols = (argc > 2)? atoi(argv[2]): 100;
        if (rows < commsize) {</pre>
            fprintf(stderr, "Number of rows %d less then number of processes %d\n", rows, commsize);
            MPI Abort(MPI_COMM_WORLD, EXIT_FAILURE);
        int args[2] = {rows, cols};
        MPI_Bcast(&args, NELEMS(args), MPI_INT, 0, MPI_COMM_WORLD);
    } else {
        int args[2];
        MPI Bcast(&args, NELEMS(args), MPI INT, 0, MPI COMM WORLD);
        rows = args[0];
        cols = args[1];
```

```
// Allocate memory for local 1D subgrids with 2 halo rows [0..ny + 1][0..cols - 1]
int ny = get block size(rows, rank, commsize);
double *local grid = xcalloc((ny + 2) * cols, sizeof(*local grid));
double *local newgrid = xcalloc((ny + 2) * cols, sizeof(*local newgrid));
// Fill boundary points:
// - left and right borders are zero filled
// - top border: u(x, 0) = \sin(pi * x)
// - bottom border: u(x, 1) = \sin(pi * x) * \exp(-pi)
double dx = 1.0 / (cols - 1.0);
if (rank == 0) {
    // Initialize top border: u(x, 0) = \sin(pi * x)
    for (int j = 0; j < cols; j++) {</pre>
        int ind = IND(0, j);
        local newgrid[ind] = local grid[ind] = sin(PI * dx * j);
if (rank == commsize - 1) {
    // Initialize bottom border: u(x, 1) = \sin(pi * x) * \exp(-pi)
    for (int j = 0; j < cols; j++) {
        int ind = IND(ny + 1, j);
        local newgrid[ind] = local grid[ind] = sin(PI * dx * j) * exp(-PI);
```

```
// Neighbours
int top = (rank > 0) ? rank - 1 : MPI PROC NULL;
int bottom = (rank < commsize - 1) ? rank + 1 : MPI PROC NULL;</pre>
// Top and bottom borders type
MPI Datatype row;
MPI Type contiguous(cols, MPI_DOUBLE, &row);
MPI Type commit(&row);
MPI Request reqs[4];
double thalo = 0;
double treduce = 0;
int niters = 0;
for (;;) {
    niters++;
    // Update interior points
    for (int i = 1; i <= ny; i++) {
        for (int j = 1; j < cols - 1; j++) {
            local newgrid[IND(i, j)] =
                (local grid[IND(i - 1, j)] + local_grid[IND(i + 1, j)] +
                 local grid[IND(i, j - 1)] + local grid[IND(i, j + 1)]) * 0.25;
```

```
// Check termination condition
double maxdiff = 0;
for (int i = 1; i <= ny; i++) {
    for (int j = 1; j < cols - 1; j++) {
        int ind = IND(i, j);
        maxdiff = fmax(maxdiff, fabs(local grid[ind] - local newgrid[ind]));
// Swap grids (after termination local grid will contain result)
double *p = local grid;
local grid = local newgrid;
local newgrid = p;
treduce -= MPI Wtime();
MPI Allreduce(MPI IN PLACE, &maxdiff, 1, MPI DOUBLE, MPI MAX, MPI COMM WORLD);
treduce += MPI Wtime();
if (maxdiff < EPS)</pre>
    break:
// Halo exchange: T = 4 * (a + b * cols)
thalo -= MPI Wtime();
MPI_Irecv(&local_grid[IND(0, 0)], 1, row, top, 0, MPI_COMM_WORLD, &reqs[0]);
                                                                                      // top
MPI Irecv(&local grid[IND(ny + 1, 0)], 1, row, bottom, 0, MPI COMM WORLD, &reqs[1]); // bottom
MPI Isend(&local grid[IND(1, 0)], 1, row, top, 0, MPI COMM WORLD, &reqs[2]);
                                                                                      // top
MPI Isend(&local grid[IND(ny, 0)], 1, row, bottom, 0, MPI COMM WORLD, &reqs[3]);
                                                                                      // bottom
MPI Waitall(4, regs, MPI STATUS IGNORE);
thalo += MPI Wtime();
```

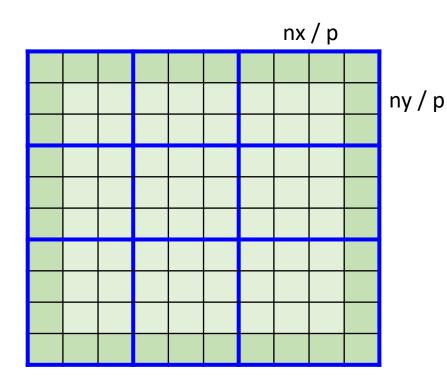
```
MPI Type free(&row);
free(local newgrid);
free(local grid);
ttotal += MPI Wtime();
if (rank == 0)
    printf("# Heat 1D (mpi): grid: rows %d, cols %d, procs %d\n", rows, cols, commsize);
int namelen;
char procname[MPI_MAX_PROCESSOR_NAME];
MPI Get processor name(procname, &namelen);
printf("# P %4d on %s: grid ny %d nx %d, total %.6f, mpi %.6f (%.2f) = allred %.6f (%.2f) + halo %.6f (%.2f)\n",
       rank, procname, ny, cols, ttotal, treduce + thalo, (treduce + thalo) / ttotal,
       treduce, treduce / (treduce + thalo), thalo, thalo / (treduce + thalo));
double prof[3] = {ttotal, treduce, thalo};
if (rank == 0) {
   MPI Reduce(MPI IN PLACE, prof, NELEMS(prof), MPI DOUBLE, MPI MAX, 0, MPI COMM WORLD);
    printf("# procs %d : grid %d %d : niters %d : total time %.6f : mpi time %.6f : allred %.6f : halo %.6f\n",
           commsize, rows, cols, niters, prof[0], prof[1] + prof[2], prof[1], prof[2]);
} else {
    MPI Reduce(prof, NULL, NELEMS(prof), MPI DOUBLE, MPI MAX, 0, MPI COMM WORLD);
MPI Finalize();
return 0;
```

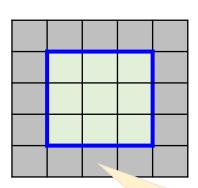
- Анализ времени выполнения информационных обменов при одномерной декомпозиции
- Время передачи сообщения размером *m* байт: $t(m) = \alpha + \beta m$
- Пусть сетка квадратная и содержит *п* строк и столбцов
- На каждой итерации выполняется два send и два recv:

$$t_{iter} = 4\alpha + 4n\beta$$

Минус: время обменов не зависит от числа р процессов – при любом числе процессов время будет одним и тем же (не будет убывать)

- Двумерная (2D) декомпозиция расчётной области на прямоугольные области
- Каждому процессу назначается подмассив [ny / p, nx / p] строк расчетной сетки
- Сетка хранится в памяти в распределенном виде
- Проблема для расчета значений некоторых точек требуются данные соседних полос, которые находятся в памяти других процессов





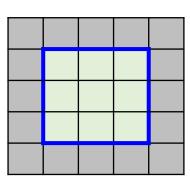
Теневые ячейки (halo, ghost cells) для хранения значений из соседних процессов

```
#define EPS 0.001
#define PI 3.14159265358979323846
#define NELEMS(x) (sizeof((x)) / sizeof((x)[0]))
#define IND(i, j) ((i) * (nx + 2) + (j))
int main(int argc, char *argv[])
   int commsize, rank;
   MPI Init(&argc, &argv);
    double ttotal = -MPI Wtime();
   MPI Comm size(MPI_COMM_WORLD, &commsize);
   MPI Comm rank(MPI COMM WORLD, &rank);
   // Create 2D grid of processes: commsize = px * py
   MPI Comm cartcomm;
   int dims[2] = \{0, 0\}, periodic[2] = \{0, 0\};
   MPI Dims create(commsize, 2, dims);
   int px = dims[0];
   int py = dims[1];
   if (px < 2 || py < 2) {
        fprintf(stderr, "Invalid number of processes %d: px %d and py %d must be greater than 1\n",
                commsize, px, py);
        MPI Abort(MPI COMM WORLD, EXIT_FAILURE);
   MPI Cart create(MPI COMM WORLD, 2, dims, periodic, 0, &cartcomm);
   int coords[2];
   MPI Cart coords(cartcomm, rank, 2, coords);
   int rankx = coords[0];
    int ranky = coords[1];
```

```
int rows, cols;
// Broadcast command line arguments
if (rank == 0) {
    rows = (argc > 1) ? atoi(argv[1]) : py * 100;
    cols = (argc > 2) ? atoi(argv[2]) : px * 100;
    if (rows < py) {</pre>
        fprintf(stderr, "Number of rows %d less then number of py processes %d\n", rows, py);
        MPI Abort(MPI COMM WORLD, EXIT FAILURE);
    if (cols < px) {</pre>
        fprintf(stderr, "Number of cols %d less then number of px processes %d\n", cols, px);
        MPI Abort(MPI COMM WORLD, EXIT FAILURE);
    int args[2] = {rows, cols};
    MPI Bcast(&args, NELEMS(args), MPI INT, 0, MPI COMM WORLD);
} else {
    int args[2];
    MPI_Bcast(&args, NELEMS(args), MPI_INT, 0, MPI_COMM_WORLD);
    rows = args[0];
    cols = args[1];
```

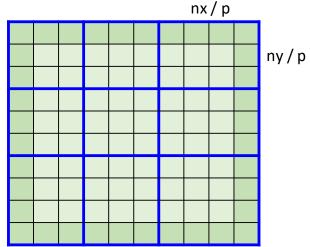
```
// Allocate memory for local 2D subgrids with halo cells [0..ny + 1][0..nx + 1]
int ny = get block size(rows, ranky, py);
int nx = get block size(cols, rankx, px);
double *local grid = xcalloc((ny + 2) * (nx + 2), sizeof(*local_grid));
double *local newgrid = xcalloc((ny + 2) * (nx + 2), sizeof(*local newgrid));
// Fill boundary points:
    - left and right borders are zero filled
     - top border: u(x, 0) = \sin(pi * x)
     - bottom border: u(x, 1) = \sin(pi * x) * \exp(-pi)
double dx = 1.0 / (cols - 1.0);
int sj = get sum of prev blocks(cols, rankx, px);
if (ranky == 0) {
    // Initialize top border: u(x, 0) = \sin(pi * x)
    for (int j = 1; j <= nx; j++) {
        // Translate col index to x coord in [0, 1]
        double x = dx * (sj + j - 1);
        int ind = IND(0, j);
        local newgrid[ind] = local grid[ind] = sin(PI * x);
if (ranky == py - 1) {
    // Initialize bottom border: u(x, 1) = \sin(pi * x) * \exp(-pi)
    for (int j = 1; j <= nx; j++) {
        // Translate col index to x coord in [0, 1]
        double x = dx * (sj + j - 1);
        int ind = IND(ny + 1, j);
        local newgrid[ind] = local grid[ind] = sin(PI * x) * exp(-PI);
```

```
// Neighbours
int left, right, top, bottom;
MPI_Cart_shift(cartcomm, 0, 1, &left, &right);
MPI Cart_shift(cartcomm, 1, 1, &top, &bottom);
// Left and right borders type
MPI Datatype col;
MPI_Type_vector(ny, 1, nx + 2, MPI_DOUBLE, &col);
MPI Type commit(&col);
// Top and bottom borders type
MPI Datatype row;
MPI Type contiguous(nx, MPI DOUBLE, &row);
MPI_Type_commit(&row);
MPI Request reqs[8];
double thalo = 0;
double treduce = 0;
```



```
int niters = 0;
for (;;) {
    niters++;
    // Update interior points
    for (int i = 1; i <= ny; i++) {
        for (int j = 1; j <= nx; j++) {
            local newgrid[IND(i, j)] =
                (local_grid[IND(i - 1, j)] + local_grid[IND(i + 1, j)] +
                 local grid[IND(i, j - 1)] + local grid[IND(i, j + 1)]) * 0.25;
    // Check termination condition
    double maxdiff = 0;
    for (int i = 1; i <= ny; i++) {
        for (int j = 1; j <= nx; j++) {
            int ind = IND(i, j);
            maxdiff = fmax(maxdiff, fabs(local grid[ind] - local newgrid[ind]));
    // Swap grids (after termination local grid will contain result)
    double *p = local grid;
    local grid = local newgrid;
    local newgrid = p;
    treduce -= MPI Wtime();
    MPI Allreduce(MPI IN_PLACE, &maxdiff, 1, MPI_DOUBLE, MPI_MAX, MPI_COMM_WORLD);
    treduce += MPI Wtime();
    if (maxdiff < EPS)</pre>
        break;
```

```
// Halo exchange: T = 4 * (a + b * (rows / py)) + 4 * (a + b * (cols / px))
 thalo -= MPI Wtime();
 MPI Irecv(&local grid[IND(0, 1)], 1, row, top, 0, cartcomm, &reqs[0]);
                                                                           // top
 MPI Irecv(&local grid[IND(ny + 1, 1)], 1, row, bottom, 0, cartcomm, &reqs[1]); // bottom
 MPI_Irecv(&local_grid[IND(1, 0)], 1, col, left, 0, cartcomm, &reqs[2]);
                                                                           // left
 MPI_Irecv(&local_grid[IND(1, nx + 1)], 1, col, right, 0, cartcomm, &reqs[3]); // right
 MPI Isend(&local grid[IND(1, 1)], 1, row, top, 0, cartcomm, &reqs[4]);
                                                                            // top
 MPI_Isend(&local_grid[IND(ny, 1)], 1, row, bottom, 0, cartcomm, &reqs[5]);
                                                                            // bottom
 MPI Isend(&local grid[IND(1, 1)], 1, col, left, 0, cartcomm, &reqs[6]); // left
 MPI Isend(&local grid[IND(1, nx)], 1, col, right, 0, cartcomm, &reqs[7]); // right
 MPI Waitall(8, regs, MPI STATUS IGNORE);
 thalo += MPI Wtime();
// iterations
```



```
MPI Type free(&row);
MPI Type free(&col);
free(local newgrid);
free(local grid);
ttotal += MPI Wtime();
if (rank == 0)
    printf("# Heat 2D (mpi): grid: rows %d, cols %d, procs %d (px %d, py %d)\n", rows, cols, commsize, px, py);
int namelen;
char procname[MPI MAX PROCESSOR NAME];
MPI Get processor name(procname, &namelen);
printf("# P %4d (%2d, %2d) on %s: grid ny %d nx %d, total %.6f, mpi %.6f (%.2f) = allred %.6f (%.2f) + halo %.6f (%.2f)\n",
       rank, rankx, ranky, procname, ny, nx, ttotal, treduce + thalo, (treduce + thalo) / ttotal,
       treduce, treduce / (treduce + thalo), thalo, thalo / (treduce + thalo));
double prof[3] = {ttotal, treduce, thalo};
if (rank == 0) {
    MPI_Reduce(MPI_IN_PLACE, prof, NELEMS(prof), MPI_DOUBLE, MPI_MAX, 0, MPI_COMM_WORLD);
    printf("# procs %d : grid %d %d : niters %d : total time %.6f : mpi time %.6f : allred %.6f : halo %.6f\n",
           commsize, rows, cols, niters, prof[0], prof[1] + prof[2], prof[1], prof[2]);
} else {
    MPI Reduce(prof, NULL, NELEMS(prof), MPI DOUBLE, MPI MAX, 0, MPI COMM WORLD);
MPI Finalize();
return 0;
```

- Анализ времени выполнения информационных обменов при двумерной декомпозиции
- Время передачи сообщения размером m байт: $t(m) = \alpha + \beta m$
- Пусть сетка квадратная и содержит п строк и столбцов, а число р процессов степень числа 2
- На каждой итерации выполняется четыре send и четыре recv:

$$t_{iter} = 8\alpha + \frac{8n}{\sqrt{p}}\beta$$

■ Плюс: время выполнения обменов уменьшается с ростом числа *р* процессов!

Задание

 ■ Сравнить в модели Хокни время дифференцированных обменов при 1Dдекомпозиции и 2D-декомпозиции (k – число итераций алгоритма)

$$t_{\mathrm{1D}} = k(4\alpha + 4n\beta)$$
 vs. $t_{\mathrm{2D}} = k(8\alpha + \frac{8n}{\sqrt{p}}\beta)$

■ При каком *р* 2D-декомпозиция будет характеризоваться меньшим временем обменов?

$$4\alpha + 4n\beta > 8\alpha + \frac{8n}{\sqrt{p}}\beta$$

Задание

 ■ Сравнить в модели Хокни время дифференцированных обменов при 1Dдекомпозиции и 2D-декомпозиции (k – число итераций алгоритма)

$$t_{\mathrm{1D}} = k(4\alpha + 4n\beta)$$
 vs. $t_{\mathrm{2D}} = k(8\alpha + \frac{8n}{\sqrt{p}}\beta)$

■ При каком *р* 2D-декомпозиция будет характеризоваться меньшим временем обменов?

$$4\alpha + 4n\beta > 8\alpha + \frac{8n}{\sqrt{p}}\beta$$

$$p < \left(\frac{2n\beta}{n\beta - \alpha}\right)^2$$