Parallelization of Jacobi algorithm in Fortran-DVMH

It is best of all to start first acquaintance with DVM system with a small example.

Let's consider small program performing some conversions of a matrix. In principle the sense of conversions doesn't matter for us now, but we want they will be executed quickly, and the result of computations will remain the same. How to achieve it? — to parallelize the program. And, of course, for this purpose we will use the DVMH technology.

Consider the source program. Full source codes are available at the end of this article. All computations are done over array B and there is an auxiliary array of A except it:

```fortran
      program jac
      parameter (L=16000, M=6250, ITMAX=200)
      real A(L,M), eps, MAXEPS, B(L,M)

       ...
      MAXEPS = 0.5e-7
```

Before start of computations the array B is initialized:

```fortran
      do j = 1, M
        do i = 1, L
          A(i,j) = 0.
          if(i.eq.1 .or. j.eq.1 .or. i.eq.L .or. j.eq.M) then
            B(i,j) = 0.
          else
            B(i,j) = (1. + i + j)
          endif
        enddo
      enddo
```

Then the iterative Jacobi algorithm is applied:

```fortran
      do it = 1, ITMAX
        eps = 0.

!       variable eps is used for calculation of maximum value
        do j =  2, M-1
          do i = 2, L-1
            eps = max(eps, abs(B(i,j) - A(i,j)))
            A(i,j) = B(i,j)
          enddo
        enddo

        do j = 2, M-1
          do i = 2, L-1
            B(i,j) = (A(i-1,j) + A(i,j-1) + A(i+1,j) + A(i,j+1)) / 4
          enddo
        enddo

        print 200, it, eps
200     format(' it = ', i4, '   eps = ', e14.7)
        if (eps .lt. MAXEPS) goto 3
      enddo
    3 continue
```

To obtain ready-to-run DVM program, it is necessary to do two things:

- to distribute data

- to distribute computations over these data.

As the DVM program can be executed on a cluster, it is possible to imagine that our program will consist of the separate MPI processes launched on different nodes of the cluster. And our task distributes to each MPI process its own portion of data and its own part of job.

To realize the first (to distribute data) it is enough to add directives:

```
      real A(L,M), eps, MAXEPS, B(L,M)

       ...
!dvm$ distribute (block, block) :: A
!dvm$ align B(i,j) with A(i,j)
!     arrays A and B  with block distribution
```

It means that the array A will be uniformly partitioned into parts along two dimensions and obtained submatrixes will be distributed over MPI processes. That is each MPI process will receive submatrix A(i1:i2, j1:j2). Moreover, the array B will be distributed in the same way and each submatrix B(i1:i2, j1:j2) will be received by the same MPI process, as the appropriate submatrix A(i1:i2, j1:j2).

Now, to distribute computations, it is necessary to find code fragments where the distributed data are used. In our case there are three two-dimensional loops which operate with arrays A and B.

Let's begin with an initialization loop. This loop is executed only once at the beginning of the program and doesn't take a lot of time but when we distributed arrays, each MPI process must check there will be no access to array elements, belonging to other MPI processes. Therefore this loop will be executed more slowly and it is quite good to parallelize it. For this purpose it is enough to add **parallel** directive:

```
!dvm$ parallel (j,i) on A(i,j)
!     nest of two parallel loops, iteration (i,j) will be executed on
!     processor, which is owner of element A(i,j)
      do j = 1, M
        do i = 1, L
          A(i,j) = 0.
          if(i.eq.1 .or. j.eq.1 .or. i.eq.L .or. j.eq.M) then
            B(i,j) = 0.
          else
            B(i,j) = 1. + i + j
          endif
        enddo
      enddo
```

It means that the iteration (j,i) should be executed only by that MPI process the element A(i,j) belongs to. Thus, MPI processes, working in parallel, will execute only intended to them iterations of the loop.

Further, it is necessary to parallelize loops inside Jacobi algorithm similarly:

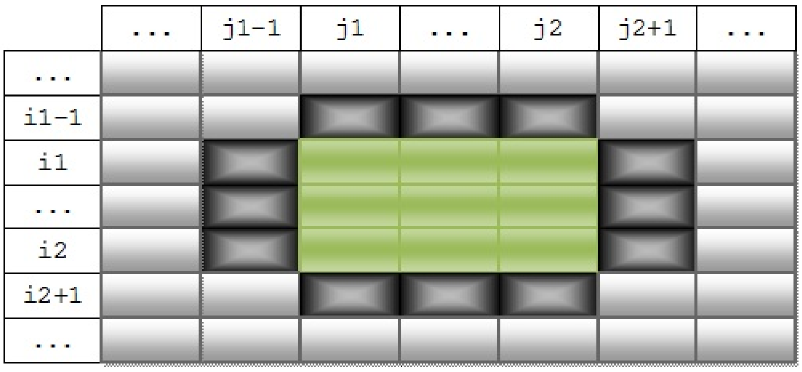
```
!dvm$   parallel (j,i) on A(i,j), reduction(max(eps))
!       variable eps is used for calculation of maximum value
        do j =  2, M-1
          do i = 2, L-1
            eps = max(eps, abs(B(i,j) - A(i,j)))
            A(i,j) = B(i,j)
          enddo
        enddo

```

```
!dvm$   parallel (j,i) on B(i,j), shadow_renew(A)
!       copying shadow elements of array A from
!       neighbouring processors before loop execution
        do j = 2, M-1
          do i = 2, L-1
            B(i,j) = (A(i-1,j) + A(i,j-1) + A(i+1,j) + A(i,j+1)) / 4
          enddo
        enddo
```

There were necessary to use additional clauses **reduction** and **shadow_renew** in these two loops.

In the first loop variable *eps* meets which should contain maximum of the calculated values on each iteration. We know that MPI process executes only its own iterations of parallel loop. Therefore, after the first loop execution its *eps* variable will have the maximum value calculated only among its iterations. Thus to compute a maximum of *eps* among all iterations, it is necessary after the loop finishing to find the maximum among *eps* of all MPI processes. The clause **reduction(max(***eps***))** does it after the loop finishing and perform all necessary actions and all MPI processes will have identical values of own *eps* variables, equal maximum among calculated on all iterations of the loop.

In the second loop there is access from iteration (j,i) to the array elements A(i-1,j), A(i,j-1) and so on. So, MPI process which owns an element A(i,j) and executes iteration (j,i), tries to access to the element A(i-1,j) which can belong to adjacent MPI process. As MPI process can't access directly to memory of other MPI process, it should preliminarily copy data to itself. Such copying can be provided by specifying **shadow_renew**. If MPI process owns submatrixes A(i1:i2,j1:j2), **shadow_renew(***A***)** will copy to it submatrixes A(i1-1,j1:j2), A(i2+1,j1:j2), A(i1:i2,j1-1), A(i1:i2,j2+1) before loop beginning.



The area formed by these submatrixes is called shadow edges. It is possible to change the size of shadow edges by **shadow** directive, but here it isn't required as the current size suits us.

At this step we have DVM program which we can launched, for example, on a cluster and we can see that the program is speed up. However, the program uses only MPI for parallel execution, but if this cluster has graphic accelerators (GPU), they will remain unused. To use them, it is required very little, namely, to specify, what fragments of the program should be executed on GPU and, if necessary, to specify, what data should be copied from a random access memory in GPU memory or back.

Regions (code fragments which can be executed on GPU) should contain parallel loops. If to look at our program, it is visible that the parallel loops are simple and can be included into regions. We do it:

```
!dvm$ region out(A, B)
!dvm$ parallel (j,i) on A(i,j)
       ...
```

```
      enddo
!dvm$ end region
       ...
!dvm$   region
!dvm$   parallel (j,i) on A(i,j), reduction(max(eps))
        ...
        enddo

!dvm$   parallel (j,i) on B(i,j), shadow_renew(A)
        ...
        enddo
!dvm$   end region
```

After specifying regions in the program, it is necessary to supply data copying between random access memory and memory of GPU to guarantee that updated data will be move there where they are necessary.

For initialization loop it's nothing to do: arrays A and B will be created and initiated on the graphic accelerator. Next region is the region in Jacobi algorithm which operates with these arrays already located in GPU memory, therefore they don't need to be copied too. However, *eps* value is calculated in the region and after the region *eps* is printed on the screen. So it is necessary to copy *eps* from the region to random access memory. Directive **get_actual** is used for this purpose:

```
        enddo
!dvm$   end region

!dvm$   get_actual(eps)
        print 200, it, eps
```

Now *eps* value calculated on GPU will be printed, but not 0 written prior to the region beginning. On the following iteration *it*, the same region is executed again. The *eps* variable already met earlier and kept its value on GPU so this value will be used when reading, but not value, assigned before the region. To correct a situation it is necessary to report that new value of variable *eps* should be taken from a random access memory. Directive **actual** can be used for it:

```
        eps = 0.
!dvm$   actual(eps)

!dvm$   region
```

We obtain working implementation of Jacobi algorithm which can be executed on a cluster with GPU, but the calculated matrix B still locates in GPU memory. If we want to print it, for example, it needs to be returned to a random access memory by **get_actual** directive:

```
      enddo
    3 continue

!dvm$ get_actual(B)
```

Now all. We obtained the final version of the DVMH program.

Startup of DVMH program

We will try to compile and launch the obtained DVMH program. To do that we need the computer on which the DVM system is installed. We have a PC with processor Intel Core i7-3770 3.40GHz and graphics card GeForce GTX Titan.

In a directory where our program jac2d.fdv is located, it is necessary to copy dvm script from the directory where DVM system is installed.

```
$ ls
dvm   jac2d.fdv
```

Now we can to compile our program.

```
$ ./dvm f jac2d.fdv
$ ls
dvm   jac2d   jac2d.fdv
```

The executable file jac2d appears in the current directory. It can be launched using dvm script, but previously it is necessary to set start parameters, having edited dvm script.

```
#--------------- DVMH options:
export DVMH_PPN='1' # Number of processes per node
export DVMH_NUM_THREADS='0' # Number of CPU threads per process
export DVMH_NUM_CUDAS='1' # Number of GPUs per process
```

Setting such parameters, we report that we want to execute regions only on one graphic accelerator. Now we can launch the program:

```
$ ./dvm run jac2d
...
 IT =  197    EPS =   0.4086035E+02
 IT =  198    EPS =   0.4062500E+02
 IT =  199    EPS =   0.4044824E+02
 IT =  200    EPS =   0.4021777E+02
 time =     3.176400
```

The program successfully performed in 3.17 seconds. For comparing, we can to compile the same source text by usual compiler, to obtain the sequential program and to launch it:

```
$ cp jac2d.fdv jac2d.f
$ ifort -O3 jac2d.f
$ ./a.out
...
 IT =  197    EPS =   0.4086035E+02
 IT =  198    EPS =   0.4062500E+02
 IT =  199    EPS =   0.4044824E+02
 IT =  200    EPS =   0.4021777E+02
 time =     27.48160
```

The results of execution are the same, but execution time was much more long.

Source code of programs

Serial version of the program

```
      program jac
      parameter (L=16000, M=6250, ITMAX=200)
      real A(L,M), eps, MAXEPS, B(L,M)

      integer clock_rate, clock_start, clock_end

      print *, '**********  test_jacobi   **********'
      MAXEPS = 0.5e-7

      do j = 1, M
        do i = 1, L
          A(i,j) = 0.
          if(i.eq.1 .or. j.eq.1 .or. i.eq.L .or. j.eq.M) then
            B(i,j) = 0.
          else
            B(i,j) = (1. + i + j)
          endif
        enddo
      enddo

      call system_clock(count_rate = clock_rate)
      call system_clock(count = clock_start)
      do it = 1, ITMAX
        eps = 0.

!       variable eps is used for calculation of maximum value
        do j =  2, M-1
          do i = 2, L-1
            eps = max(eps, abs(B(i,j) - A(i,j)))
            A(i,j) = B(i,j)
          enddo
        enddo

        do j = 2, M-1
          do i = 2, L-1
            B(i,j) = (A(i-1,j) + A(i,j-1) + A(i+1,j) + A(i,j+1)) / 4
          enddo
        enddo

        print 200, it, eps
200     format(' it = ', i4, '   eps = ', e14.7)
        if (eps .lt. MAXEPS) goto 3
      enddo
    3 continue

      call system_clock(count=clock_end)
      print *, 'time = ', real(clock_end - clock_start) / clock_rate

      end
```

DVMH version of the program

```
      program jac
      parameter (L=16000, M=6250, ITMAX=200)
      real A(L,M), eps, MAXEPS, B(L,M)

      integer clock_rate, clock_start, clock_end

!dvm$ distribute (block, block) :: A
!dvm$ align B(i,j) with A(i,j)
!     arrays A and B  with block distribution

      print *, '**********  test_jacobi   **********'
      MAXEPS = 0.5e-7

!dvm$ region out(A, B)
!dvm$ parallel (j,i) on A(i,j)
!     nest of two parallel loops, iteration (i,j) will be executed on
!     processor, which is owner of element A(i,j)
      do j = 1, M
        do i = 1, L
          A(i,j) = 0.
          if(i.eq.1 .or. j.eq.1 .or. i.eq.L .or. j.eq.M) then
            B(i,j) = 0.
          else
            B(i,j) = 1. + i + j
          endif
        enddo
      enddo
!dvm$ end region

      call system_clock(count_rate = clock_rate)
      call system_clock(count = clock_start)
      do it = 1, ITMAX
        eps = 0.
!dvm$   actual(eps)

!dvm$   region
!dvm$   parallel (j,i) on A(i,j), reduction(max(eps))
!       variable eps is used for calculation of maximum value
        do j =  2, M-1
          do i = 2, L-1
            eps = max(eps, abs(B(i,j) - A(i,j)))
            A(i,j) = B(i,j)
          enddo
        enddo

!dvm$   parallel (j,i) on B(i,j), shadow_renew(A)
!       copying shadow elements of array A from
!       neighbouring processors before loop execution
        do j = 2, M-1
          do i = 2, L-1
            B(i,j) = (A(i-1,j) + A(i,j-1) + A(i+1,j) + A(i,j+1)) / 4
          enddo
        enddo
!dvm$   end region

```

```
!dvm$   get_actual(eps)
          print 200, it, eps
200       format(' it = ', i4, '   eps = ', e14.7)
          if (eps .lt. MAXEPS) goto 3
        enddo
     3 continue

!dvm$ get_actual(B)
        call system_clock(count=clock_end)
        print *, 'time = ', real(clock_end - clock_start) / clock_rate

        end
```