P3DFFT - Highly scalable parallel 3D Fast Fourier Transforms library

USER GUIDE

Version 2.3

Copyright (C) 2006-2008 Dmitry Pekurovsky
San Diego Supercomputer Center/UC SanDiego

Copyright (C) 2008 Jens Henrik Goebbert (ghost cell support)
RWTH-Aachen, Germany

Acknowledgements

Prof. P.K.Yeung
Dr. Diego Donzis
Dr. Giri Chukkapalli
Dr. Geert Brethouwer

Feedback

Please send your feedback, including bugs and suggestions, to `dmitry@sdsc.edu`

1. `p3dfft` module

The `p3dfft` module declares important variables. It should be included
in any code that calls P3DFFT routines (via `use p3dfft` statement in Fortran).

The `p3dfft` module also specifies `mytype`, which is the type of real and
complex numbers. Its value of 4 corresponds to single precision and 8 for double precision.
You can choose precision at compile time through a preprocessor flag (see Installation Guide).

2. Initialization

Before using the library it is necessary to call an initialization routine `p3dfft_setup`.

Usage: *p3dfft_setup(proc_dims,nx,ny,nz,overwrite)*

Table 1: Arguments of p3dfft_setup

Argument	Intent	Description
proc_dims	Input	An array of two integers, specifying how the processor grid should be decomposed. Either 1D or 2D decomposition can be specified. For example, when running on 12 processors, (4,3) or (2,6) can be specified as proc_dims to indicate a 2D decomposition, or (1,12) can be specified for 1D decomposition. proc_dims values are used to initialize P_1 and P_2.
nx,ny,nz	Input	(Integer) Dimensions of the 3D transform (also the global grid dimensions)
overwrite	Input	(Logical) When set to .true. (or 1 in C) this argument indicates that it is safe to overwrite the input of the btran (backward transform) routine. This may speed up performance of FFTW routines in some cases when non-stride-1 transforms are made.

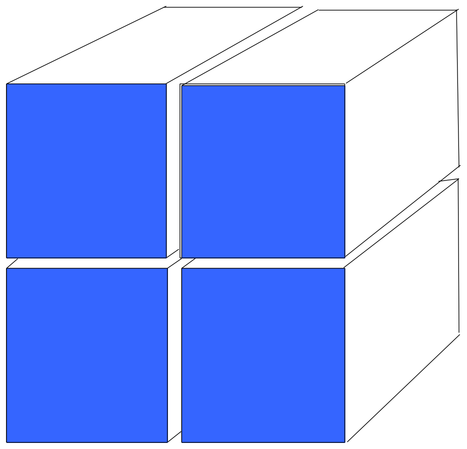
To initialize ghost cells operations, a call to p3dfft_init_ghosts should be made. See example *driver_ghost.f* in *sample/FORTRAN*.

3. Array Decomposition

The `p3dfft_setup` routine sets up the two-dimensional (2D) array decomposition. P3DFFT employs 2D block decomposition whereby processors are arranged into a 2D grid P_1 x P_2, based on their MPI rank. Two of the dimensions of the 3D grid are block-distributed across the processor grid, by assigning the blocks to tasks in the rank order. The third dimension of the grid remains undivided, i.e. contained entirely within local memory (see Fig. 1). This scheme is sometimes called *pencils decomposition*.

A block decomposition is defined by dimensions of the local portion of the array contained within each task, as well as the beginning and ending indices for each dimension defining the array's location within the global array. This information is returned by `get_dims` routine which should be called before setting up the data structures of your program (see *sample/* subdirectory for example programs).

Figure 1: An example of 2D block(a.k.a. *pencils*) decomposition of a 3D grid.



In P3DFFT, the decompositions of the output and input arrays, while both being two-dimensional, differ from each other. The reason for this is as follows. In 3D Fourier Transform it is necessary to transpose the data a few times (two times for two-dimensional decomposition) in order to rearrange the data so as to always perform one-dimensional FFT on data local in memory of each processing element. It would be possible to transpose the data back to the original form after the 3D transform is done, however it often makes sense to save significant time by forgoing this final transpose. All the user has to do is to operate on the output array while keeping in mind that the data are in a transposed form. The backward (complex-to-real) transform takes the array in a transposed form and produces a real array in the original form. The rest of this section clarifies exactly the original and transposed form of the arrays.

Usage: *get_dims(start,end,size,ip)*

Table 2: Arguments of `get_dims()`

Argument	Intent	Description
start	Output	An array containing 3 integers, defining the beginning indices of the local array for the given task within the global grid
end	Output	An array containing 3 integers, defining the ending indices of the local array within the global grid (these can be computed from start and size but are provided for convenience)
size	Output	An array containing 3 integers, defining the local array's dimensions
ip	Input	An integer argument specifying one of the two choices for array types: ip=1: "Original": a "physical space" array of real numbers, local in X, distributed among P_1 tasks in Y dimension and P_2 tasks in Z dimension, where P_1 and P_2 are processor grid dimensions defined in the call to `p3dfft_setup`. Usually this type of array is an input to real-to-complex (forward) transform and an output of complex-to-real (backward) transform. ip=2: "Transposed": a "wavenumber space" array of complex numbers, local in Z, distributed among P_1 tasks in X dimension, P_2 tasks in Y dimension. Usually this type of array is an output of real-to-complex (forward) transform and an input to complex-to-real, backward transform.

Note: the layout of the 2D processor grid on the physical network is dependent on the architecture and software of the particular system, and can have some impact on efficiency of communication. By default, rows have processors with adjacent task IDs (this corresponds to "FORTRAN" type ordering). This can be changed to "C" ordering (columns have adjacent task IDs) by building the library with -DDIMS_C preprocessor flag. The former way is recommended on most systems.

P3DFFT uses 2D block decomposition to assign local arrays for each task. In many cases decomposition will not be entirely even: some tasks will get more array elements than others. P3DFFT attempts to minimize load imbalance. For example is the grid dimensions are 128 x 256 x 256 and the processor grid is defined as 3x4, the original (ip=1) decomposition calls for splitting 256 elements in Y dimension into three processor row. P3DFFT in this case will break it up into pieces of 86, 85 and 85 elements. The transposed (ip=2) decomposition will have local arrays with X dimensions 22, 22 and 21 respectively for processor rows 1 through 3 (the sum of these numbers is 65=(Nx+2)/2 since these are now complex numbers instead of reals, and an extra mode for Nyquist frequency is needed – see Section 5 for an explanation).

It should be clear that the user's choice of P1 and P2 can make a difference on how balanced is the decomposition. Obviously the greater load imbalance, the less performance can be expected.

Note: the two array types are distributed among processors in a different way from each other, but this does not automatically imply anything about the ordering of the elements in memory. Memory layout of the original (ip=1) array uses the "Fortran" ordering. For example, for an array *A(lx,ly,lz)* the index corresponding to *lx* runs fastest. Memory layout for the transposed (ip=2) array type depends on how the P3DFFT library was built. By default, it preserves the ordering of the real array, i.e. (X,Y,Z). However, in many cases it is advisable to have Z dimension contiguous, i.e. a memory layout (Z,X,Y). This can speed up some computations in the wavenumber space by improving cache utilization through spatial locality in Z, and also often results in better performance of P3DFFT transforms themselves. The (Z,X,Y) layout can be triggered by building the library with `-DSTRIDE1` preprocessor flag in the makefile. For more information, see performance section below.

4. Forward (real-to-complex) and backward (complex-to-real) 3D Fourier transforms

Forward transform is implemented by the p3dfft_ftran_r2c subroutine using the following format:

p3dfft_ftran_r2c(IN,OUT)

The input IN is an array of real numbers with dimensions defined by array type with ip=1 (see Table 2 above), with X dimension contained entirely within each task, and Y and Z dimensions distributed among P_1 and P_2 tasks correspondingly. The output OUT is an array of complex numbers with dimensions defined by array type with ip=2, i.e. Z dimension contained entirely, and X and Y dimensions distributed among P_1 and P_2 tasks respectively.

Backward transform is implemented by the p3dfft_btran_c2r subroutine using the following format:

p3dfft_btran_c2r(IN,OUT)

The input IN is an array of complex numbers with dimensions defined by array type with ip=2 (see Table 2 above), i.e. Z dimension is contained entirely, and X and Y dimensions are distributed among P_1 and P_2 tasks correspondingly. The output OUT is an array of real numbers with dimensions defined by array type with ip=1, i.e. X dimension is contained entirely within each task, and Y and Z are dimensions distributed among P_1 and P_2 tasks respectively.

5. Complex array storage definition

Since Fourier transform of a real function has the property of conjugate symmetry, only about half of the complex Fourier coefficients need to be kept. To be precise, if the input array has n real elements, Fourier coefficients F(k) for k=n/2+1,..,n can be dropped as they can be easily restored from the rest of the coefficients. This saves both memory and time. In this version we do not attempt to further pack the complex data. Therefore the output array for the forward transform (and the input array of the backward transform) contains (Nx/2+1) * Ny * Ny complex numbers, with the understanding that Nx/2-1 elements in X direction are missing and can be restored from the remaining elements. As mentioned above, the Nx/2+1 elements in the X direction are distributed among P_1 tasks in the transposed layout.

6. In-place transforms

In and Out arrays can occupy the same space in memory (in-place transform). In this case, it is necessary to make sure that they start in the same location, otherwise the results are unpredictable. Also it is important to remember that the sizes of input and output arrays in general are not equal. The complex array is usually bigger since it contains the Nyquist frequency mode in X direction, in addition to the Nx/2 modes that equal in space to Nx real numbers. However when decomposition is not even, sometimes the real array can be bigger than the complex one, depending on the task ID. Therefore to be safe one must make sure the common-space array is large enough for both input and output. This can be done by finding out dimensions of both original and transposed arrays y calling `get_dims` two times with ip=1 and 2.
In Fortran using in-place transforms is a bit tricky due to language restrictions on subroutine argument types (i.e., one of the arrays is expected to be real and the other complex). In order to overcome this problem wrapper routines are provided, named `ftran_r2c` and `btran_c2r` respectively for forward and backward transform (without `p3dfft_` prefix). There are examples for such in-place transform in the *sample/* subdirectory. These wrappers can be also used for out-of-place transforms just as well.

7. Memory requirements

Besides the input and output arrays (which can occupy the same space, as mentioned above) P3DFFT allocates temporary buffers roughly 3 times the size of the input or output array.

8. Performance considerations

P3DFFT was created to compute 3D FFT in parallel with high efficiency. In particular it is aimed for applications where the data volume is large. It is especially useful when running applications on ultra-scale parallel platforms where one-dimensional decomposition is not adequate. Since P3DFFT was designed to be portable, no effort is made to do architecture-specific optimization. However, the user is given some choices in setting up the library, mentioned below, that may affect performance on a given system. Current version of P3DFFT uses ESSL or FFTW library for it 1D FFT routines. ESSL [1] provides FFT routines highly optimized for IBM platforms it is built on. The FFTW [2], while being generic, also makes an effort to maximize performance on many kinds of architectures. Some performance data will be uploaded at the P3DFFT Web site. For more questions and comments please contact Dmitry@sdsc.edu.

Optimal performance on many parallel platforms for a given number of cores and problem size will likely depend on the choice of processor decomposition. For example, given a processor grid P_1 x P_2 (specified in the first argument to `p3dfft_setup`) performance will generally be better with smaller P_1 (with the product P_1 x P_2 kept constant). Ideally P_1 will be equal or less than the number of cores on an SMP node or a multi-core chip. In addition, the closer a decomposition is to being even, the better load balancing.

Performance is likely to be better when P3DFFT is built using `-DSTRIDE1` flag. This implies stride-1 data ordering for FFTs. Note that using this flag changes the memory layout of the transposed array (see section 3 for explanation). To help tune performance further, two more flags can be used: `-DNBL_X=…` and `-DNBL_Y=…,` which define block sizes in X and Y when doing local array reordering. Choosing suitable block sizes allows the program to optimize cache performance. Recommended value for NBL_Y is 1, while a reasonable value for NBL_X depends on problem size and the cache sizes of the hardware architecture.

Finally, performance will be better if *overwrite* parameter is set to .true. (or 1 in C) when initializing P3DFFT. This allows the library to overwrite the input array, which results in significantly faster execution when not using `-DSTRIDE1` flag.

9. Ghost cell support

Elementary ghost cell operations (nearest-neighbor/halo send-receive) are enabled in this version. See *driver_ghost.f* in sample*/FORTRAN/* for an example of the use.

```
 how to use ghost cell support:
==============
   1) init ghost-support:
     ==>> call p3dfft_init_ghosts(goverlap)
     goverlap = overlap of ghost-cell-slab
             (limitations: only one goverlap supported)
     (of cause we are still assuming periodic boundary conditions)

   2) allocation of memory
     each array which needs ghost-cell-support has to have
     the min. size:      fsize(1)+2*goverlap *
                    fsize(2)+2*goverlap *
                    fsize(3)+2*goverlap

   3) exchange ghost-cells among processors
     to update the ghost-cells among the proceesors
     ==>> call update_rghosts(array).
     This will assume data represents physical space!

   4) use/read ghost-cells
     if you want to access/read ghost-cells
     ==>> call ijk2i(i,j,k)
```

```
      which returns the position in a 1D-array as an integer*8

  array-memory:
 ==============
    1) size if complex-type: fsize(1)+2*goverlap *
                             fsize(2)+2*goverlap *
                             fsize(3)+2*goverlap
    2a) structure in memory:
     |00000000000000000000000000|11111|22222|33333|44444|55555|66666|
            ^^^^                                 ^     ^     ^     ^     ^     ^
       input/output of FFT                       A     B     C     D     E     F

     ghost-cell-slab at each pencil-side:
     A:  -X,Y*Z  A(istart(1)-i, j,k)          => gmem_start(1)
     B:  +X,Y*Z  B(i,j,k)                     => gmem_start(2)
     C:  -Y,X*Z  C(i,j,k)                     => gmem_start(3)
     D:  +Y,X*Z  D(i,j,k)                     => gmem_start(4)
     E:  -Z,X*Y  E(i,j,k)                     => gmem_start(5)
     F:  +Z,X*Y  F(i,j,k)                     => gmem_start(6)

    2b) a more closer look at "A":
     ....|aaaaaabbbbbbccccccdddddd|....
         ^         ^     ^     ^
         a1        b1    c1    d1

     each ghost-cell-slab is build up by no. goverlap planes:
     a1:  Y*Z-plane with lowest X-coord
      |
     d1:  Y*Z-plane with highest X-coord

  important variables:
 ====================
    1) goverlap
      the data overlap between two processors - equal to the thickness
      of a ghost-slab
      This version supports only one fixed overlap value, because it simplifies
      the coding. If different overlaps for different arrays are needed the
      following varables must be dependent on the overlap-value and therefor
      should have one additional dimension.
      A good idea would be to make p3dfft_init_ghosts(..) return an id,which
      has to get set when calling update_ghosts(data,id).

    2) gmem_start(1..6) (-X,-Y,-Z,+X,+Y,+Z)
      start of data in memory for ghost-cells of certain side

    3) gmess_size(1..3) (X,Y,Z)
      size of mpi message needed to send/recieve ghost-slab each dimension(1..3)

    4a) gslab_start(1..3, 1..6)
      relative start coord. of ghost-slab in pencil ijk(1..3), [neg./pos side(1..2) for each
dimension(1..3)](1..6)
         -X   -Y   -Z   +X   +Y   +Z
        i
      j
      k

   4b) gslab_end(1..3, 1..6)
     relative end coord. of ghost-slab in pencil ijk(1..3), [neg./pos side(1..2) for each
dimension(1..3)](1..6)
        -X   -Y   -Z   +X   +Y   +Z
      i
      j
      k

   4c) gslab_size(1..3, 1..3)
     size of ghost-slab for each dimension(1..3), ijk(1..3)

   6) gneighb_(numtasks*(3*2))
     extended processor grid with all processor neighbours in -/+XYZ(1..6)
     including periodic boundary neighbours
      example for 2 processors with gneighb_r(:) = 0,0,0,0,1,1,1,1,1,1,0,0
       pencil-side: -X  +X  -Y  +Y  -Z  +Z
       taskid 0:     0   0   0   0   1   1
       taskid 1:     1   1   1   1   0   0
============================================================
```

10. References

1. ESSL library, IBM, http://publib.boulder.ibm.com/infocenter/clresctr/vxrx/index.jsp?topic=/com.ibm.cluster.essl.doc/esslbooks.html

2. Matteo Frigo and Steven G. Johnson, "The Design and Implementation of FFTW3," *Proceedings of the IEEE* **93** (2), 216–231 (2005). Invited paper, Special Issue on Program Generation, Optimization, and Platform Adaptation.