

Nov 21, 00 16:05

DRI_LIS_07282000.txt

Page 1/22

\$Id: DRI_LIS.txt,v 1.6 2000/11/21 05:01:14 kcairn Exp \$

This is a language-independent specification of a meta-API for data reorganization. This version of the specification is based on the consensus of the Data Reorganization Forum following its June 2000 meeting.

DATE OF THIS DRAFT: 07/28/2000

There are 4 sections to this document:

- 1) Administrative notes
- 2) Change summary between this version and the prior version
- 3) Current version of "critical path" interfaces in the API
- 4) Current version of other interfaces in the API

----- SECTION 1: Administrative Notes -----

These notes are mostly re-iterated from the last version of this draft document (dated 03/07/2000).

NOTE #1: A change has been made in the organization of this file so that the functions that are in the "critical path" to creating and initiating a data reorganization function are presented first. Other functions such as object query functions that are not in this "critical path" are presented later.

As of this date (07/28/2000), this draft contains only the critical path calls (i.e., Section 4 is empty). Since there have been many changes to these functions, we should agree on their specification before handling the low-level ("power user") and non-critical functions.

NOTE #2: A new feature is being added for the committee members - a list of changes that have been introduced in this newer version of the API and the corresponding reasons. The goal is to prevent revisiting issues that have been resolved in past working meetings.

----- SECTION 2: Change Summary -----

***** Changes from March 2000 draft to July 2000 (POST JUNE DR MEETING) draft:

1. Changed capitalization conventions as decided in prior meetings, but has not yet been implemented. The new convention is:

DRI_ prefix for everything (objects/types, function names, etc.)
 DRI_Data_Type (capitalized first letter of object/type names)
 DRI_Data_Type_method (lowercase method/function names)
 DRI_Method (for standalone functions without associated objects)

2. Changed DRI_Distspec object to DRI_Partition, per group's decision at the June 2000 meeting

An associated change is that the pre-defined object DRI_DISTSPEC_INDIVISIBLE has been changed to DRI_PARTITION_WHOLE

(2 changes - DISTSPEC to PARTITION, and INDIVISIBLE to WHOLE)

Nov 21, 00 16:05

DRI_LIS_07282000.txt

Page 2/22

3. Changed DRI_Dist object to DRI_Distribution, per group's decision at the June 2000 meeting
4. Changed DRI_Bufferset_create to DRI_Bufferset_system_create
5. Modified DRI_Bufferset_system_create function to take a DRI_Distribution object parameter (as decided in June 2000 meeting)
6. Created DRI_Bufferset_user_create function to import user-allocated memory into a DRI_Bufferset object (instead of calling DRI_Bufferset_system_create to have the library/run-time perform the memory allocation).
7. Added DRI_Init and DRI_Finalize functions to the API specification
8. Specified a number of pre-defined "NULL" objects (one for each major data type in the API specification). DRI_Init creates these NULL objects at runtime.
9. Added a new function DRI_Partition_whole_create that allows the user to get a DRI_Partition object that specifies no partitioning at all (per June meeting).
10. Changed DRI_Group_create parameters to require an "original_group" from which a subset process group is to be created
11. REMOVED the following language from parts of this document, based on decisions made in June 2000 meeting.

<UNRESOLVED>

What if we are not using standard middleware, but a process set construct exists? Do we want to leverage those (non-portable) approaches to alleviate the need to do process set management in DRI? If we choose to do this, then this object and its methods become unnecessary?

</UNRESOLVED>

***** Changes from December 1999 draft to March 2000 (PRE DR MEETING) draft:

1. Marked appropriate functions as "<META>" to indicate areas where Data Reorg will likely use infrastructure from other middleware

Each meta function now has a "META NOTES" section to try to organize the discussion on the meta-nature of the DRI API.
2. Inserted <UNRESOLVED> notation in this document where some remaining decisions are needed. Before final API is settled, we need to go back and remove these notes and replace with the final decisions
3. Inserted _destroy functions for the following objects:
DRI_Global_Data
DRI_Group
DRI_Overlap
DRI_Distspec
DRI_Bufferset
DRI_Channel.

***** Changes from September 1999 to December 1999 drafts:

Nov 21, 00 16:05

DRI_LIS_07282000.txt

Page 3/22

1. DRI_Group_myrank name changed to DRI_Group_get_rank()
2. DRI_Dist_create - added a note in RESTRICTIONS/POLICY section reiterating that this call may not involve collective communication, at the implementation's discretion. This of course means that erroneous programs could cause DRI_Dist_create to calculate an incorrect data partitioning.
3. DRI_Dist_create - Added useful default layout parameters so the user doesn't have to use DRI_Layout_create to create commonly-needed objects (e.g., DRI_LAYOUT_PACKED_012). All possible packed layouts for 1, 2, and 3 dimensions are provided. This effectively makes DRI_Layout_create a non "critical-path" function in the API.
4. DRI_Dist_create - A NULL pointer argument for the group_dims parameter specifies that the user wants to have the implementation determine an appropriate logical process set topology to use in dividing up the data during the execution of this call.
5. DRI_Dist_create - Added a note in the DESCRIPTION section that says that a valid entry in the distspecs array parameter is DRI_DISTSPEC_INDIVISIBLE (this capability was accidentally removed from the API in prior edits).
6. DRI_Dist_create - Noted that the group_dims array parameter corresponds directly to the dimsizes array parameter of the DRI_Global_Data_create function.
7. Added a DRI_Dist_get_numblocks function
8. Added a DRI_Dist_get_blockinfo function to return a single structure that gives all needed information about a locally owned block of data following the partitioning process. Returns a new DRI_Blockinfo structure type. internally, the DRI_Blockinfo structure contains an array of DRI_Blockdim structures. This pair of structures replaces the old DRI_Part object and bounds_t structure. DRI_Part was not adding anything beyond DRI_Dist, so we have opted directly query DRI_Dist for the low level partitioning information. The bounds_t structure was poorly named, and needed to be more descriptive (we now call it DRI_Blockdim). Additional information was needed beyond what the old bounds_t provided, so we just
9. Changed DRI_Part_calc_local_size to DRI_Dist_calc_local_size, since the DRI_Part object has been removed and we now just query DRI_Dist objects.
10. Added the DRI_Bufferset object and its DRI_Bufferset_create function
11. DRI_Transfer object has been renamed to DRI_Channel

----- SECTION 3: Current API for "critical path" functions -----

```

/***** <META> DRI_Init *****/
DRI_Init - Initialize the data reorganization run-time environment

```

SYNOPSIS

```

DRI_Init(argvp, argcp) - C language binding
DRI_Init(???)         - other language bindings

```

PARAMETERS

Nov 21, 00 16:05

DRI_LIS_07282000.txt

Page 4/22

INOUT: argvp (pointer to array of strings) - application command line arguments

INOUT: argcp (pointer to integer) - address of integer variable that stores the number of command line arguments contained in argvp

META-NOTES

This part of DRI is META in the sense that co-layer implementations with MPI or MPI/RT may be able to accomplish the necessary Data Reorg setup actions within their respective Init functions.

DESCRIPTION

Parses the application command line for implementation-specific data reorganization library options.

Synchronizes with all other data reorganization processes in the environment, and produces the DRI_GROUP_WORLD object that is used to represent all processes in a data reorganization based application.

If they are not already provided at compile-time, this function creates a pre-defined "null" objects:

- DRI_GROUP_NULL
- DRI_GLOBAL_DATA_NULL
- DRI_DATASPEC_NULL
- DRI_OVERLAP_NULL
- DRI_PARTITION_NULL
- DRI_DISTRIBUTION_NULL
- DRI_LAYOUT_NULL
- DRI_CHANNEL_NULL
- DRI_BUFFERSET_NULL
- DRI_BUFFER_ID_NULL

Also, if necessary, creates the following pre-defined objects:

- DRI_PARTITION_WHOLE

COMMUNICATION BEHAVIOR

Collective. Synchronizes with all other data reorganization processes in the environment, and produces the DRI_GROUP_WORLD object that is used to represent all processes in a data reorganization based application.

RESTRICTIONS / POLICY

DRI_Init must be the first data reorganization library function called.

```

/***** DRI_Global_Data_create *****/
DRI_Global_Data_create - Create a global data object

```

SYNOPSIS

```
DRI_Global_Data_create(ndims, dimsizes[ndims], dataspec, gdo)
```

PARAMETERS

IN: ndims (integer) - number of dimensions in the global data

IN: dimsizes (integer array) - size of each dimension of the global data

IN: dataspec (DRI_Dataspec) - data type of each element of the global data

Nov 21, 00 16:05

DRI_LIS_07282000.txt

Page 5/22

OUT: gdo (DRI_Global_Data) - object that describes the global data

DESCRIPTION

Creates a global data object to describe application data. The size information supplied by the user refers to the size of the application data `_without_` considering how the data will eventually be partitioned across a group of processes in the parallel environment

COMMUNICATION BEHAVIOR

Local. All processes that will participate in a future data reorganization involving this data must create this object independently.

RESTRICTIONS / POLICY

All processes that will participate in a data reorganization on the described data must call this function with identical `ndims`, `dimsizes`, and `dataspec` parameters. Implementations may place an upper limit on the `ndims` parameter. However, all implementations must minimally support
`1 <= ndims <= 3`

/***** <META> DRI_Group_create *****/
 DRI_Group_create - Create an object to represent a group of processes

SYNOPSIS

DRI_Group_create(original_group, num_ranks, rank_list, new_group)

PARAMETERS

IN: `original_group` (DRI_Group) - group from which a subset will be taken to produce the `new_group` of processes
 IN: `num_ranks` (integer) - total number of processes in the group to be created
 IN: `rank_list` (array of integer) - list of logical process ranks from the `original_group` that will form the `new_group` of processes
 OUT: `new_group` (DRI_Group) - process group object

META-NOTES

This is meta at the "object level". That is, some implementations may choose to completely leverage constructs from other middleware APIs (e.g., MPI Communicators) as part of a co-layered implementation with Data Reorg. Implementations that take this approach may elect not to implement DRI_Group objects and its associated methods.

DESCRIPTION

Creates an object to represent a group of unique processes in the parallel processing environment. Groups are one-dimensional logical orderings of processes. Each process is assigned an integer rank, numbered between zero and the total number of processes - 1. The `original_group` parameter must be a valid data reorganization group. The pre-defined DRI_Group object `DRI_GROUP_WORLD` must be used to create the first subset group of processes.

COMMUNICATION BEHAVIOR

Local.

Nov 21, 00 16:05

DRI_LIS_07282000.txt

Page 6/22

RESTRICTIONS / POLICY

/***** <META> DRI_Group_get_rank *****/
 DRI_Group_get_rank - Return the rank of the calling process in specified group

SYNOPSIS

DRI_Group_get_rank(group, rank)

PARAMETERS

IN: group (DRI_Group) - group object
 OUT: rank (integer) - rank of the calling process in the group

META-NOTES

This is meta at the "object level". See notes for DRI_Group_create.

DESCRIPTION

Returns the rank (logical process id) in the given group to the caller.

COMMUNICATION BEHAVIOR

Local

RESTRICTIONS / POLICY

Only members of the specified group may call this function successfully

/***** <META> DRI_Group_get_size *****/
 DRI_Group_get_size - Return the size of the specified group

SYNOPSIS

DRI_Group_get_size(group, size)

PARAMETERS

IN: group (DRI_Group) - group object
 OUT: size (integer) - size of the specified group

META-NOTES

This is meta at the "object level". See notes for DRI_Group_create.

DESCRIPTION

Returns the number of participating processes in the given group

COMMUNICATION BEHAVIOR

Local

RESTRICTIONS / POLICY

Nov 21, 00 16:05

DRI_LIS_07282000.txt

Page 7/22

Only members of the specified group may call this function successfully

```

/***** DRI_Overlap_create *****/
DRI_Overlap_create - Create an overlap data partitioning object

```

SYNOPSIS

```
DRI_Overlap_create(ovr_type, num_pos, overlaph)
```

PARAMETERS

IN: ovr_type (DRI_Overlap_type) - overlap policy to implement at the edges of a global data object. Can be one of:

```

DRI_OVERLAP_TRUNCATE
DRI_OVERLAP_TOROIDAL
DRI_OVERLAP_PAD_ZEROS
DRI_OVERLAP_PAD_REPLICATED

```

IN: num_pos (integer) - number of positions to overlap

OUT: overlaph (DRI_Overlap) - overlap object

DESCRIPTION

Creates the overlap attribute used in the data distribution high-level specification. The resulting DRI_Overlap object is to be passed into either DRI_Partition_block_create or DRI_Partition_blockcyclic_create as a left or right overlap argument.

NOTE: Just like the DRI_Partition object, the user is expected to create a DRI_Overlap object specification for each dimension of global data (where a nonzero overlap is desired). In the event that no overlap is requested by the user, DRI_NO_OVERLAP can be passed as the left and right overlap arguments to one of the the DRI_Partition_<block|blockcyclic>create functions.

In general, overlap is the storage of extra data in a processor's local data buffer to hold data that is adjacent in the global data context, but that is assigned to another processor, based on the data partitioning. Overlap therefore refers to data that is stored on processor boundaries in the partitioning of the global data.

There are different overlap policies supported:

1) ovr_type == DRI_OVERLAP_TRUNCATE

The local buffer should contain enough space to store copies of num_pos adjacent, non-local elements. At the ends of the global data object, extra storage is not required in the local data buffer, and is truncated accordingly.

2) ovr_type == DRI_OVERLAP_TOROIDAL

The local buffer should contain enough space to store copies of num_pos adjacent, non-local elements. At the ends of the global data object, extra storage is required in the local data buffer, and will be filled with data from the num_pos elements that start at the opposite end of the global data dimension.

3) ovr_type == DRI_OVERLAP_PAD_ZEROS

The local buffer should contain enough space to store copies of num_pos

Nov 21, 00 16:05

DRI_LIS_07282000.txt

Page 8/22

adjacent, non-local elements. At the ends of the global data object, extra storage is required in the local data buffer, and will be filled with zeros.

4) `ovr_type == DRI_OVERLAP_PAD_REPLICATED`

The local buffer should contain enough space to store copies of `num_pos` adjacent, non-local elements. At the ends of the global data object, extra storage is required in the local data buffer, and will be filled with a copy of the last `num_pos` `_locally_held` elements.

COMMUNICATION BEHAVIOR

Local.

RESTRICTIONS / POLICY

```

/***** DRI_Partition_block_create *****/
/***** DRI_Partition_blockcyclic_create *****/
/***** DRI_Partition_whole_create *****/

```

`DRI_Partition_block_create` - Create a block distribution specification
`DRI_Partition_blockcyclic_create` - Create a block cyclic distribution
`DRI_Partition_whole_create` - Create an indivisible (whole) distribution

SYNOPSIS

```

DRI_Partition_block_create(minsz, mod, lov, rov, part)
DRI_Partition_blockcyclic_create(lov, rov, blksz, part)
DRI_Partition_whole_create(part)

```

PARAMETERS

IN: `minsz` (integer) - minimum number of local elements required
 (user specifies 0 to indicate no preference)

IN: `mod` (integer) - modulo requirement
 (user specifies 1 to indicate no preference)

IN: `lov` (`DRI_Overlap`) - left overlap (`DRI_NO_OVERLAP` specifies no overlap)

IN: `rov` (`DRI_Overlap`) - right overlap (`DRI_NO_OVERLAP` specifies no overlap)

IN: `blksz` (integer) - block-cyclic partitioning block size
 (user specifies 1 for pure cyclic partition)

OUT: `part` (`DRI_Partition`) - high-level data distribution object

DESCRIPTION

These functions create a `DRI_Partition` object that stores information about either a block, blockcyclic, or indivisible (whole) partitioning of global application data. Users must associate a separate `DRI_Partition` object with each dimension of partitioned global data. The output object, `part`, is only a high-level specification of the requested data partitioning. It does not store exact partitioning details such as specific global data indices assigned to a particular process. Because a `DRI_Partition` object is not associated with any single global data array, it can be reused for many different

Nov 21, 00 16:05

DRI_LIS_07282000.txt

Page 9/22

data partitionings. The more exact partitioning information for a particular global data array is stored in the DRI_Distribution object that can be queried for detailed partitioning information following the DRI_Distribution_create operation.

Calling DRI_Partition_whole_create will produce an object equivalent to the pre-defined object DRI_PARTITION_WHOLE. Implementations may in fact return a reference to this pre-defined object as the output of DRI_Partition_whole_create.

Parameter mod specifies that the number of local elements ultimately assigned to the calling process must be some multiple of mod.

Parameters lov and rov specify element overlaps (left and right, respectively). These parameters do not change the mapping of global data indices to processors in the data partitioning. They allow copies of adjacent global data elements at the (left or right) boundaries of the data partitioning to be stored locally. A right overlap refers to overlap in the direction of higher global indices. Consult the section on the DRI_Overlap object for additional details about the overlap specification.

Parameter blksize is used in block-cyclic partitionings to define the size (in number of elements) of the blocks that get assigned to processors in the global data partitioning.

COMMUNICATION BEHAVIOR

Local

RESTRICTIONS / POLICY

This object may NOT be queried until the completion of a subsequent DRI_Distribution_create call.

COMMUNICATION BEHAVIOR

Local

```

/***** <META> DRI_Distribution_create *****/
DRI_Distribution_create - Create a distribution object for a specific
                        global data object over a specific process group

```

SYNOPSIS

```
DRI_Distribution_create(gdo, group, group_dims, parts, layout, disth)
```

PARAMETERS

```

IN:  gdo (DRI_Global_Data) - global data object
IN:  group (DRI_Group) - process group
IN:  group_dims (array of integer) - logical dimensions of process group
IN:  parts (array of DRI_Partition) - high-level data distribution specs
     (one array entry per gdo dimension)

IN:  layout (DRI_Layout) - memory layout of local data buffers
OUT: disth (DRI_Distribution) - data distribution object

```

META-NOTES

This function is meta because it takes parameter of type DRI_Group, which is meta at the "object level". The DRI_Distribution object itself is not meta - it is unique to Data Reorganization. This function, however, is meta at an "interface level".

DESCRIPTION

This function aggregates all of the input objects into a single container, a DRI_Distribution object. It also calculates explicitly the data block(s) of the global data that will be assigned to processes (and stores that detailed information in the resulting DRI_Distribution object). The user will be able to query this low-level information following the execution of this call. Note that the data partitioning performed here guarantees that each global data element is assigned to a process. It is unlikely, but possible that some processes could be assigned NO global data elements as a result of this call.

The layout parameter may have been created by a prior call to the DRI_Layout_create function, or it can be specified with one of the following pre-defined layouts (if the global data described by the gdo parameter is either two or three-dimensional):

- DRI_LAYOUT_PACKED_012
- DRI_LAYOUT_PACKED_021
- DRI_LAYOUT_PACKED_102
- DRI_LAYOUT_PACKED_120
- DRI_LAYOUT_PACKED_201
- DRI_LAYOUT_PACKED_202
- DRI_LAYOUT_PACKED_01
- DRI_LAYOUT_PACKED_10

These pre-defined layout objects specify the order in which a multidimensional local buffer is organized in linear memory space. The numeric codes at the end of each predefined symbol serve this purpose. The most contiguously arranged dimension is indicated by the LAST digit of the predefined object name. The least contiguously arranged dimension is the FIRST digit shown in the object name. The term PACKED refers to data that is compactly stored in memory, meaning that the user is requesting no strides between consecutive elements of the local data.

The group_dims array specifies a logical process set dimensionality for the process group identified by the "group" parameter. The number of elements in group_dims must be equal to the number of dimensions specified for the gdo parameter in a prior call to DRI_Global_Data_create (i.e., group_dims corresponds directly to the dimsizes parameter of DRI_Global_Data_create). The product of all values in group_dims must equal the total number of processes in the process group "group". This parameter gives the caller more explicit control over the global data partitioning process performed by DRI_Distribution_create.

Just like the layout parameter, the group_dims parameter also allows for a Default setting. A NULL pointer can be passed in place of an integer array to specify that the user has no preference for how the process group is viewed logically.

The logical view of processes specified in group_dims is only effective during the execution of the DRI_Distribution_create() function. Other calls to DRI_Distribution_create involving the same DRI_Group object parameter, but a different DRI_Global_Data may use alternative "views" of the process group to yield a different type of partitioning to be applied to the (also different) data set.

Nov 21, 00 16:05

DRI_LIS_07282000.txt

Page 11/22

The parts parameter is an array of DRI_Partition objects, one entry per dimension of data being partitioned. The entries in the array are created prior to this call by using one of the DRI_Partition_<block|blockcyclic>_create functions. An exception is when a data dimension will not be partitioned at all - in that case, the corresponding array entry in the parts parameter here can contain the pre-defined object DRI_PARTITION_WHOLE.

COMMUNICATION BEHAVIOR

At the implementation's discretion, this can be performed either as a collective operation, or as a local operation.

RESTRICTIONS / POLICY

It is possible that an implementation will choose to not communicate collectively among the members of the process group during the execution of this function. It is therefore possible that the constituent processes that make up the group could (erroneously) supply different specifications for the following important parameters: gdo, group_dims, parts. In that case the resulting data distribution that is computed and stored in the DRI_Distribution output parameter may be incorrect.

The user must be able to query the low-level partitioning details that result from this call immediately following completion of this call. This is true even if the implementation does not perform any communication between processes in the specified group during the execution of this function.

```

/***** DRI_Distribution_get_numblocks *****/
DRI_Distribution_get_numblocks - Return the number of locally stored blocks
                                from a data partitioning

```

SYNOPSIS

```
DRI_Distribution_get_numblocks (dsth, nblocks)
```

PARAMETERS

```

IN:  dsth (DRI_Distribution) - data distribution object
OUT: nblocks (integer) - number of blocks associated with the low-level
                        partitioning referred to by the dsth parameter

```

DESCRIPTION

This function returns the number of blocks assigned as part of a low-level data partitioning (described by the dsth parameter that was created in an earlier DRI_Distribution_create call). For block data partitionings, this function will return a value of 1 in the nblocks output parameter. For block-cyclic partitionings, a value greater than 1 may be returned in the nblocks parameter.

COMMUNICATION BEHAVIOR

Local.

RESTRICTIONS/POLICY

Nov 21, 00 16:05

DRI_LIS_07282000.txt

Page 12/22

```

/***** DRI_Distribution_get_blockinfo *****/
DRI_Distribution_get_blockinfo - Get detailed information about a
                                local block of data

```

SYNOPSIS

```
DRI_Distribution_get_blockinfo (dsth, block_num, blockinfo)
```

PARAMETERS

```

IN:  dsth (DRI_Distribution) - data distribution object
IN:  block_num (integer) - local block number for which information is needed
OUT: blockinfo (DRI_Blockinfo) - returned structure containing detailed
                                information about the block

```

DESCRIPTION

For a specified low-level data partitioning object (dsth) and local block index (block_num), allocates and returns a structure to the user containing the following:

```

{
  ndims (integer) - number of dimensions in the local data block described
  first_offset (integer) - offset (in elements) from the beginning of the local
                          application's memory buffer to the first "owned"
                          element of this data block. It therefore in some
                          cases does not identify the first data element in
                          the block, since the first element in storage could
                          be the result of an overlapped data partitioning.
  elem_size (integer) - number of bytes per data element in the local block.
                      This can be obtained by querying other objects
                      (DRI_Global_Data and DRI_Partition), but is provided in
                      this structure for user convenience.
  dims[ndims] (array of DRI_Blockdim structures) - detailed information
            (on a per-dimension basis) about the range of global indices covered
            by the local block of data referred to by this DRI_Blockinfo structure
}

```

The DRI_Blockdim structure referred to above is defined as follows:

```

{
  lov (DRI_Overlap) - left overlap in this dimension
  rov (DRI_Overlap) - right overlap in this dimension
  global_begin_ix (integer) - global index of the first "owned" data element in
                              the block in this dimension
  length (integer) - number of "owned" data elements in this dimension
  stride (integer) - number of elements between consecutive data elements
                    in the local data buffer in this dimension.  If this value is 1, then
                    the data is densely packed, with no spacing between consecutive elements.
}

```

COMMUNICATION BEHAVIOR

Local.

RESTRICTIONS/POLICY

Nov 21, 00 16:05

DRI_LIS_07282000.txt

Page 13/22

```

/***** DRI_Distribution_calc_local_size *****/
DRI_Distribution_calc_local_size - Calculate size of local buffers associated
                                with one side of a data reorganization

```

SYNOPSIS

```
DRI_Distribution_calc_local_size(disth, local_size)
```

PARAMETERS

```

IN:  disth (DRI_Distribution) - low-level data partitioning object
OUT: local_size (integer) - number of bytes specifying the size of data
    buffers that will be used in future data reorganizations

```

DESCRIPTION

This function tells the caller what size (in bytes) is required of application data buffers that participate in data reorganizations associated with the data partitioning object `disth`. The returned `local_size` parameter is calculated based on a combination of user-specified partitioning parameters (at `DRI_Distribution_create`-time) and implementation-imposed decisions regarding local memory layouts.

The number of bytes returned in the `local_size` parameter specifies the size of a memory buffer that is large enough to hold all local blocks from a data partitioning. This is particularly relevant for block-cyclic partitionings, in which it is possible and likely that multiple blocks of data are assigned to a single process.

COMMUNICATION BEHAVIOR

Local.

RESTRICTIONS / POLICY

```

/***** <META> DRI_Bufferset_system_create *****/
DRI_Bufferset_system_create - create shared application/library buffers
                                for processing and data reorganization

```

SYNOPSIS

```
DRI_Bufferset_system_create (nbufs, dist, bufset)
```

PARAMETERS

```

IN:  nbufs (integer) - number of buffers of size bufsize that will make up
    the buffer set to be created by this function
IN:  dist (DRI_Distribution) - distribution object that helps to determine
    buffer sizes that will be created in the set
OUT: bufset (DRI_Bufferset) - buffer set object created

```

META-NOTES

The `DRI_Bufferset` object itself is meta, because other middleware such as MPI/RT has similar constructs that presumably can be leveraged on co-layer implementations instead of using the Data Reorg specific object. This area needs to be studied further by the committees (MPI/RT and Data Reorganization).

Nov 21, 00 16:05

DRI_LIS_07282000.txt

Page 14/22

DESCRIPTION

Creates a buffer set object that will be associated with a later data reorganization (represented by a DRI_Channel object). After this call, the user will never directly query or manipulate the DRI_Bufferset object created. Once the association of the buffer set is made with a channel object (in a later call do DRI_Channel_create), all access to the buffer set's constituent buffers will be made through that associated channel object. In that interaction, the user will work with individual DRI_Buffer_Id objects that are obtained with a call to DRI_Channel_get and returned to the channel with a call to DRI_Channel_put. See the documentation for those functions for additional details on buffer set management.

COMMUNICATION BEHAVIOR

Local.

RESTRICTIONS / POLICY

```

/***** <META> DRI_Bufferset_user_create *****/
DRI_Bufferset_user_create - create shared application/library buffers
                           from user-allocated memory
                           for processing and data reorganization

```

SYNOPSIS

```
DRI_Bufferset_user_create (nbufs, buffer_ptrs[], dist, bufset)
```

PARAMETERS

```

IN:  nbufs (integer) - number of buffers that will make up
      the buffer set to be created by this function
IN:  buffer_ptrs (array of pointer) - addresses of user-allocated buffers
IN:  dist (DRI_Distribution) - distribution object containing buffer
      sizes that should be met or exceeded in the user-supplied buffers
OUT: bufset (DRI_Bufferset) - buffer set object created

```

META-NOTES

This is meta at the "object level". See notes for DRI_Bufferset_system_create.

DESCRIPTION

Creates a buffer set object (to be used in conjunction with an associated DRI_Channel object) from user-allocated memory. Although the application programmer will have access to the addresses of each buffer using this approach, "safe" use of these memory areas must be negotiated by calling DRI_Channel_get and DRI_Channel_put for the associated channel object.

COMMUNICATION BEHAVIOR

Local.

RESTRICTIONS / POLICY

The buffers that are supplied as input parameters here must provide

Nov 21, 00 16:05

DRI_LIS_07282000.txt

Page 15/22

a sufficient amount of memory to insure correct function of the associated channel operation (data reorg) to take place. Users can determine the amount of memory needed in each of the buffers by calling `DRI_Distribution_calc_local_size` (using the `DRI_Distribution` object associated with the data reorg channel).

```

/***** <META> DRI_Channel_create_send *****/
/***** <META> DRI_Channel_create_rcv *****/
DRI_Channel_create - create data reorganization communication channel

```

SYNOPSIS

There are two forms of this call:

```

DRI_Channel_create_send(name, srcDist, srcBufs, channel);
DRI_Channel_create_rcv(name, destDist, destBufs, channel);

```

PARAMETERS

```

IN name: (string/integer?) Identifier for the channel
IN srcDist: (DRI_Distribution) distribution object on the send side
IN destDist: (DRI_Distribution) distribution object on the receive side
IN srcBufs: (DRI_Bufferset) send side data buffers
IN destBufs: (DRI_Bufferset) receive side data buffers
OUT channel: (DRI_Channel) Data reorganization (channel) object created

```

META-NOTES

The `DRI_Channel` object may become meta in some co-layers (e.g., MPI/RT). However, the extent to which this will happen is yet to be determined. A likely scenario for MPI/RT co-layer implementations is that the features associated with MPI/RT channels (QoS, buffer iterators, admission) will be merged with those found in Data Reorg channels (distributions, buffersets, connection). This needs further thought by the two committees.

DESCRIPTION

The send channel object allows the calling process to participate in a data reorganization as a sender. The receive channel object has a similar (obvious) function. To properly set up data reorganizations in which the caller is both a sender and receiver of data, both forms must be called, resulting in two `DRI_Channel` objects.

COMMUNICATION BEHAVIOR

Local. Processes create channel objects independently and in any order.

RESTRICTIONS / POLICY

Buffers supplied here are assumed to be big enough to contain all the data transferred. To find the size of the buffer, use the function `DRI_Distribution_calc_local_size`. Alternatively, the user can have the middleware allocate the associated bufferset using the `DRI_Bufferset_system_create` call. The correct storage is determined by using `DRI_Distribution` object parameter to that call.

Currently, we assume that data reorganizations are either bi-partite (pipeline) or clique-based (Single Program Multiple Data). Intermediate cases, that is, partially overlapping process groups, are disallowed. If any process is both a sender and a receiver, all processes must be both senders and receivers, or

Nov 21, 00 16:05

DRI_LIS_07282000.txt

Page 16/22

an error will result at the time of the subsequent DRI_Channel_connect call.

On a given "side" of a channel (send or receive), all of the participating processes must provide buffersets that contain the same number of local buffers as every other process. The number of buffers on the send side of a channel `_can_` be different than the number of buffers in the bufferset associated with the receive side of the same channel. The reason for the restriction is to enable high performance implementations. The middleware will be able to compute in advance:

- the explicit pairings of send/rcv buffers in the data reorganizations to be performed with this channel
- the precise order in which the pairings will occur (if there are multiple buffers on the send and receive sides of the channel)

```

/***** <META> DRI_Channel_connect *****/
DRI_Channel_connect(chan) - Pipeline channel connect

```

SYNOPSIS

```
DRI_Channel_connect(chan)
```

PARAMETERS

INOUT chan: (DRI_Channel) channel object to be connected

META-NOTES

See notes for DRI_Channel_create.

DESCRIPTION

Enables a given pipeline data reorganization: calculates which processors are Sending to and receiving from which other processors.

COMMUNICATION BEHAVIOR

The connect call is a synchronization point between all processors in the send and receive sides of the data reorganization identified by the chan parameter: it is collective and blocking.

RESTRICTIONS / POLICY

Multiple channel objects must be connected in the correct order by the involved parties or deadlock may (will probably) result.

```

/***** <META> DRI_Channel_connect_sendrecv *****/

```

SYNOPSIS

```
DRI_Channel_connect_sendrecv(c_send, c_rcv) - Clique channel connect
```

PARAMETERS

INOUT c_send: (DRI_Channel) object managing the "send side" of a data reorg
 INOUT c_rcv: (DRI_Channel) object managing the "receive side" of a data reorg

Nov 21, 00 16:05

DRI_LIS_07282000.txt

Page 17/22

META-NOTES

See notes for DRI_Channel_create.

DESCRIPTION

Enables a given clique data reorganization: calculates which processors are Sending to and receiving from which other processors.

COMMUNICATION BEHAVIOR

The connect call is a synchronization point between all processors in the send and receive sides of the given data reorganization: it is collective and blocking.

RESTRICTIONS / POLICY

Multiple channel objects must be connected in the correct order by the involved parties or deadlock may (will probably) result.

```

/***** <META> DRI_Channel_get *****/
/***** <META> DRI_Channel_put *****/

```

SYNOPSIS

```

DRI_Channel_get (chan, buf) - Receive data reorg buffer / Get free buffer
DRI_Channel_put (chan, buf) - Send data reorg buffer / Return used buffer

```

PARAMETERS

```

INOUT chan: (DRI_Channel) channel object managing a data reorganization
OUT buf: (DRI_Buffer_Id) handle to memory buffer

```

META-NOTES

See notes for DRI_Channel_create.

DESCRIPTION

Discussion of DRI_Channel_get:

If the channel object argument refers to the "receive side" of a data reorganization, this function returns a buffer that represents the received data from a set of sending processes. If the channel object argument refers to the "send side" of a data reorganization, this function returns an available buffer to the application so that it can produce the data that will be sent in a subsequent data reorganization operation.

Discussion of DRI_Channel_put:

If the channel object argument refers to the "send side" of a data reorganization, this function initiates the communication using the data provided in the input buffer argument. If the channel object refers to the "receive side" of a data reorganization, then this call simply returns the buffer to the DRI library so that it can be filled up with received data in a subsequent data reorganization operation.

Nov 21, 00 16:05

DRI_LIS_07282000.txt

Page 18/22

General discussion of put and get in context:

In pipeline data reorganizations, incoming buffers are obtained by calling `DRI_Channel_get` with a "receive side" channel object input argument. If, after processing the received buffer, the program needs to send the data "downstream"

in the pipeline, the same buffer can be used as input to a `DRI_Channel_put` call, but with a separate channel object (representing the "send side" of a different data reorganization). In cases where the calling program is at the beginning or end of an application pipeline, the buffer may be returned to the buffer set by calling `DRI_Channel_put` with the same channel object parameter that was used in the earlier `DRI_Channel_get`.

For clique data parallel applications, there are two channel objects associated with the same data reorganization (one for the send side, one for the receive side). To execute clique data reorganizations, the program calls `DRI_Channel_get` with the send-side channel object as input. The returned buffer is filled and a data reorganization is initiated with a call to `DRI_Channel_put` (passing again as input the send-side channel object and the buffer id). The program then calls `DRI_Channel_get`, using the second channel object (associated with the receive side of the data reorganization).

COMMUNICATION BEHAVIOR

`DRI_Channel_get` is a blocking call and does not return until a full buffer of received data is available

`DRI_Channel_put` is a non-blocking call and returns immediately to the calling application, regardless of whether the associated communication has completed. The channel object will manage the availability of the buffers associated with the data reorganization, protecting the buffer from future application use (via `DRI_Channel_get`) until the communication has completed and it is safe to reuse the buffer.

RESTRICTIONS / POLICY

It is possible to use the same `DRI_Channel` object for two different data reorganizations when using a clique data-parallel design. The receive-side channel object from the first data reorganization executed can also act as the send-side channel object for a second, distinct data reorganization. This is permissible when the data buffer sizes do not change as a result of application processing between the two data reorganizations.

----- SECTION 4: Current API for remaining functions -----

/***** DRI_Global_Data_destroy *****/
`DRI_Global_Data_destroy` - destroy a global data object

SYNOPSIS

```
DRI_Global_Data_destroy(global_data)
```

PARAMETERS

INOUT: `global_data` (`DRI_Global_Data`) - object that describes the global data

Nov 21, 00 16:05

DRI_LIS_07282000.txt

Page 19/22

DESCRIPTION

Destroys the global data object referred to by the global_data input parameter.

COMMUNICATION BEHAVIOR

Local.

RESTRICTIONS / POLICY

This function should only free resources associated with the global_data object when necessary. That is, all references to the global data object must be "destroyed" via this call before the actual internal resources used by the global data object are freed and returned to the system.

```

/***** <META> DRI_Group_destroy *****/
DRI_Group_destroy - Destroy an object representing a group of processes

```

SYNOPSIS

```
DRI_Group_destroy(grp)
```

PARAMETERS

INOUT: grp (DRI_Group) - process group object

META-NOTES:

This is meta at the "object level". See notes for DRI_Group_create.

DESCRIPTION

Destroys the process set group object referred to by the grp input parameter.

COMMUNICATION BEHAVIOR

Local.

RESTRICTIONS / POLICY

This function should only free resources associated with the group object when necessary. That is, all references to the group object must be "destroyed" via this call before the actual internal resources used by the group object are freed and returned to the system.

```

/***** DRI_Overlap_destroy *****/
DRI_Overlap_destroy - Destroy an overlap data partitioning object

```

SYNOPSIS

```
DRI_Overlap_destroy(ov)
```

PARAMETERS

INOUT: ov (DRI_Overlap) - overlap object

Nov 21, 00 16:05

DRI_LIS_07282000.txt

Page 20/22

DESCRIPTION

Destroys the object referred to by the ov parameter

COMMUNICATION BEHAVIOR

Local.

RESTRICTIONS / POLICY

This function should only free resources associated with the overlap object when necessary. That is, all references to the overlap object must be "destroyed" via this call before the actual internal resources used by the overlap object are freed and returned to the system.

```

/***** DRI_Partition_destroy *****/
DRI_Partition_destroy - Destroy a data distribution specification object

```

SYNOPSIS

```
DRI_Partition_destroy(part)
```

PARAMETERS

INOUT: part (DRI_Partition) - high-level data distribution object

DESCRIPTION

Destroys the object referred to by the part parameter. This parameter can refer to either a block or block-cyclic distribution object (created by DRI_Partition_block_create or DRI_Partition_blockcyclic_create, respectively).

COMMUNICATION BEHAVIOR

Local

RESTRICTIONS / POLICY

This function should only free resources associated with the part object when necessary. That is, all references to the part object must be "destroyed" via this call before the actual internal resources used by the part object are freed and returned to the system.

```

/***** <META> DRI_Bufferset_destroy *****/
DRI_Bufferset_destroy - destroy shared application/library buffers
                        for processing and data reorganization

```

SYNOPSIS

```
DRI_Bufferset_destroy (nbufs, bufsize bufset)
```

PARAMETERS

INOUT: bufset (DRI_Bufferset) - buffer set object destroyed

META-NOTES

See DRI_Bufferset_create notes.

Nov 21, 00 16:05

DRI_LIS_07282000.txt

Page 21/22

DESCRIPTION

Destroys the object referred to by the bufset parameter.

COMMUNICATION BEHAVIOR

Local.

RESTRICTIONS / POLICY

This function should only free resources associated with the bufferset object when necessary. That is, all references to the bufferset object must be "destroyed" via this call before the actual internal resources used by the bufferset object are freed and returned to the system.

```

/***** <META> DRI_Channel_destroy *****/
DRI_Channel_destroy - destroy data reorganization communication channel

```

SYNOPSIS

```
DRI_Channel_destroy(chan);
```

PARAMETERS

INOUT chan: (DRI_Channel) Data reorganization (channel) object destroyed

META-NOTES

See DRI_Channel_create notes.

DESCRIPTION

Destroys the channel referred to by the chan parameter. Frees all internal resources used by the channel, including temporary buffers that may have been created during the earlier DRI_Channel_connect() call.

COMMUNICATION BEHAVIOR

<UNRESOLVED>

It would be nice to be able to "shut down" a channel gracefully (i.e., in a "collective" fashion). This could be difficult with respect to process synchronization in pipeline application architectures, where many processes participate in two data reorganization channels. This scenario forces a specific order in which channels must be destroyed (or else deadlock could occur). Since this will apparently be pushed to the application level, a proposal would be to make DRI_Channel_destroy have local communication behavior, and to have applications use other middlewares for the necessary "graceful synchronization".

</UNRESOLVED>

RESTRICTIONS / POLICY

This destroy call is different than most others because all internal resources associated with the channel can be freed without checking for other "references". Channels in the Data Reorganization API cannot be referenced more than once.

Nov 21, 00 16:05

DRI_LIS_07282000.txt

Page 22/22

```
/***** <META> DRI_Finalize *****/  
DRI_Finalize - Free resources used by the data reorganization  
run-time environment
```

SYNOPSIS

```
DRI_Finalize()
```

PARAMETERS

META-NOTES

This part of DRI is META in the sense that co-layer implementations with MPI or MPI/RT may be able to accomplish the necessary Data Reorg finalize actions within their respective Finalize functions.

DESCRIPTION

Frees any internal resources used by the data reorganization implementation.

COMMUNICATION BEHAVIOR

Local

RESTRICTIONS / POLICY