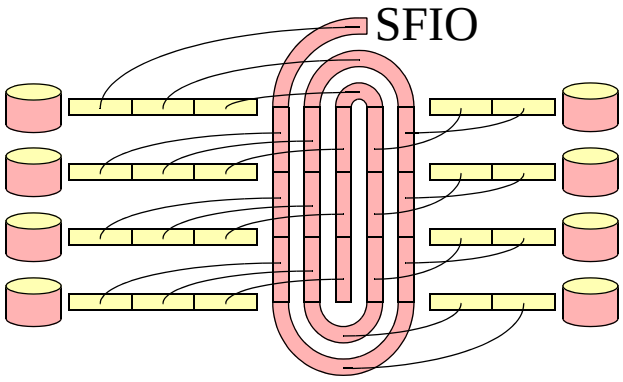
5th Workshop on Distributed Supercomputing: Scalable Cluster Software
May 23-24, 2001, Sheraton Hyannis, Cape Cod, Hyannis Massachusetts, USA

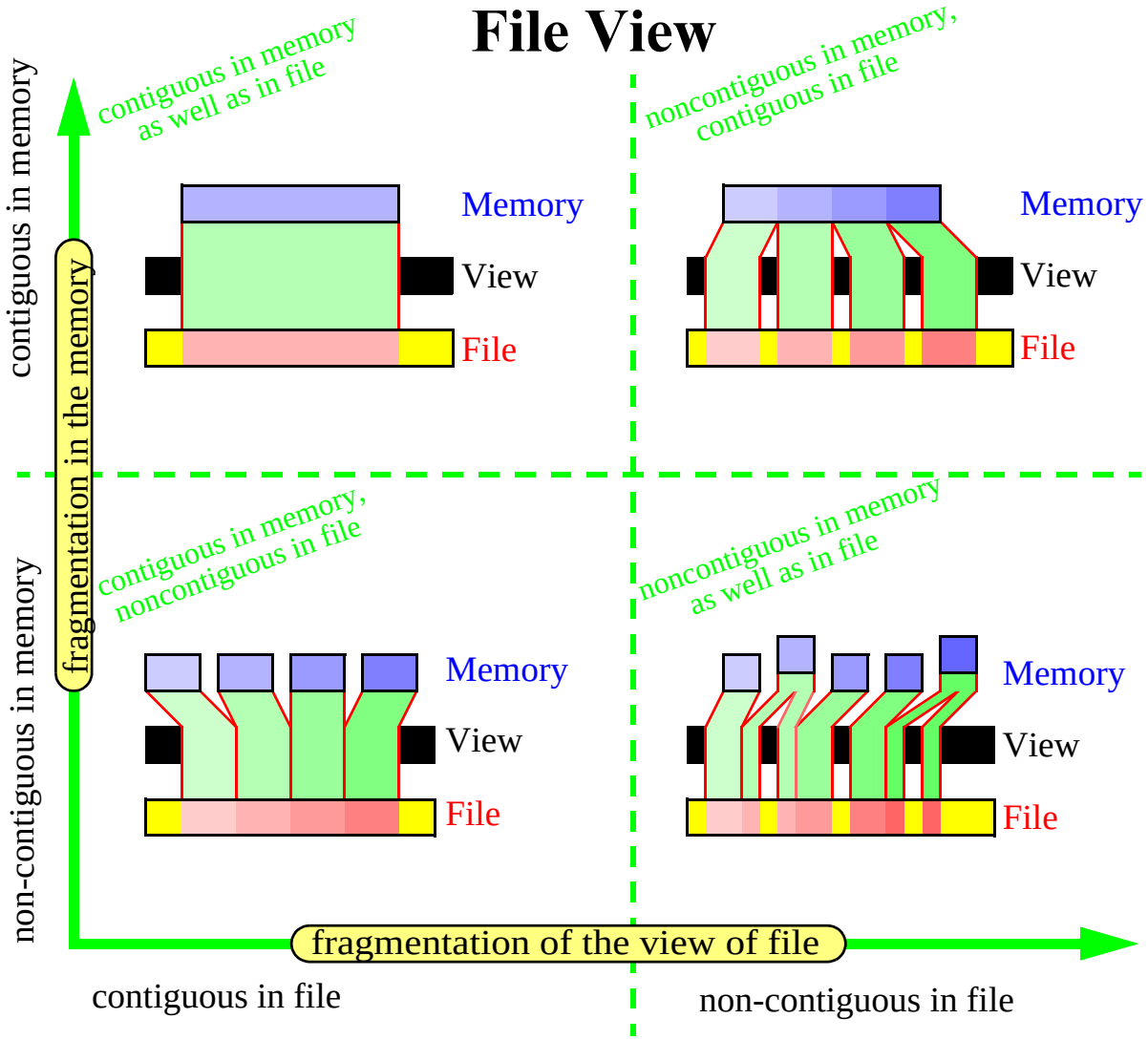
ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Isolated MPI-I/O for any MPI-1

Emin Gabrielyan, Roger D. Hersch
École Polytechnique Fédérale de Lausanne, Switzerland
{Emin.Gabrielyan,RD.Hersch}@epfl.ch



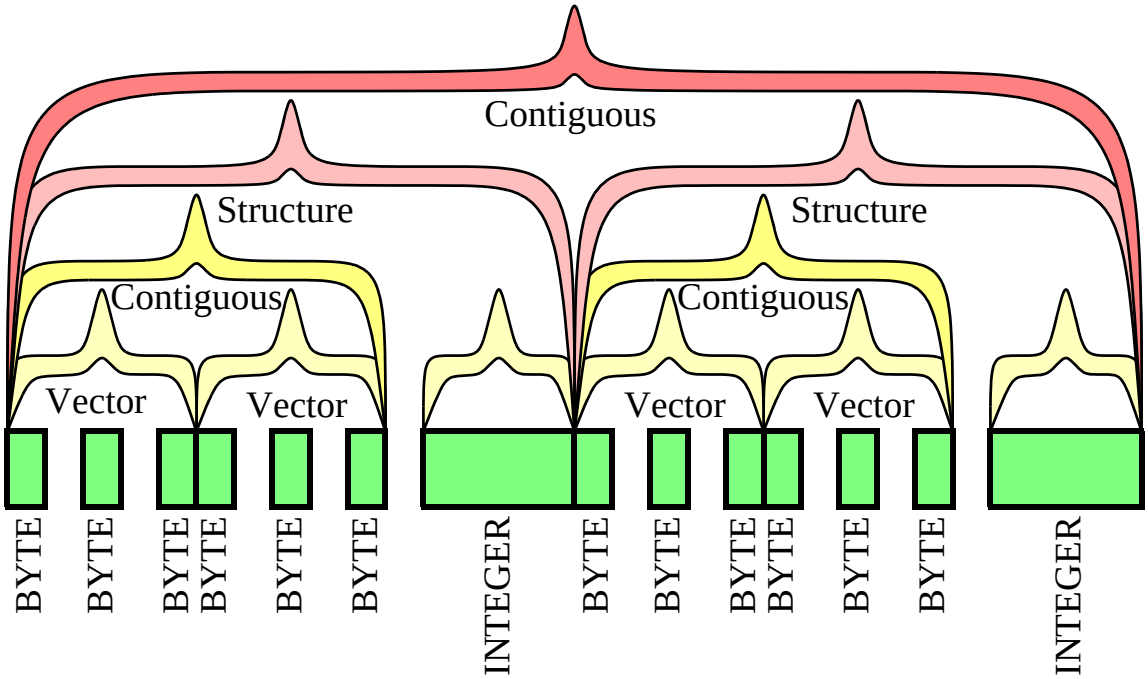
.



The file view is a global concept, which influence all data access operations. Each process obtains its own view of the shared data file. In order to specify the file view the user creates a derived datatype. Since each access operation can use another derived datatype that specifies the fragmentation in memory, there are two orthogonal aspects to data access: the fragmentation in memory and the fragmentation of the file view.

.

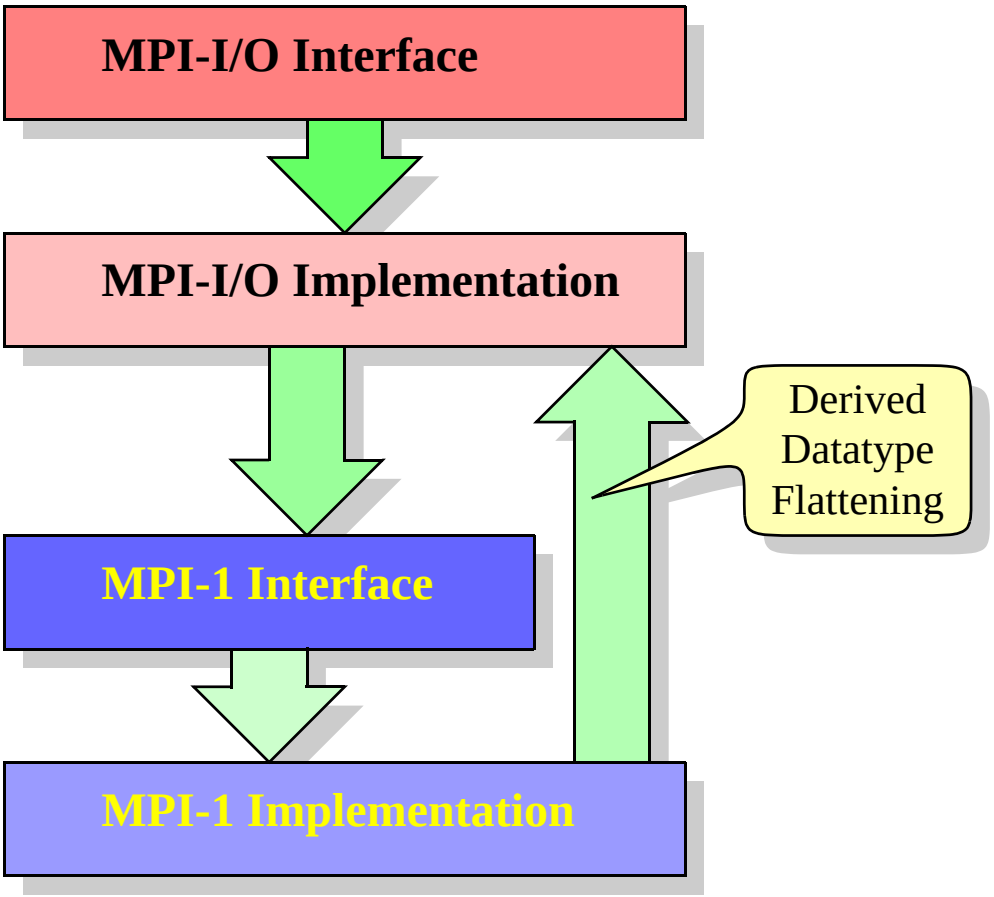
Derived Datatypes



- **MPI-1 provides techniques for creating datatype objects having an arbitrary data layout in memory.**
- **A derived opaque datatype object can be used in various MPI operations.**
- **But the layout information, once encapsulated in a derived datatype, can not be extracted with standard MPI-1 functions.**

.

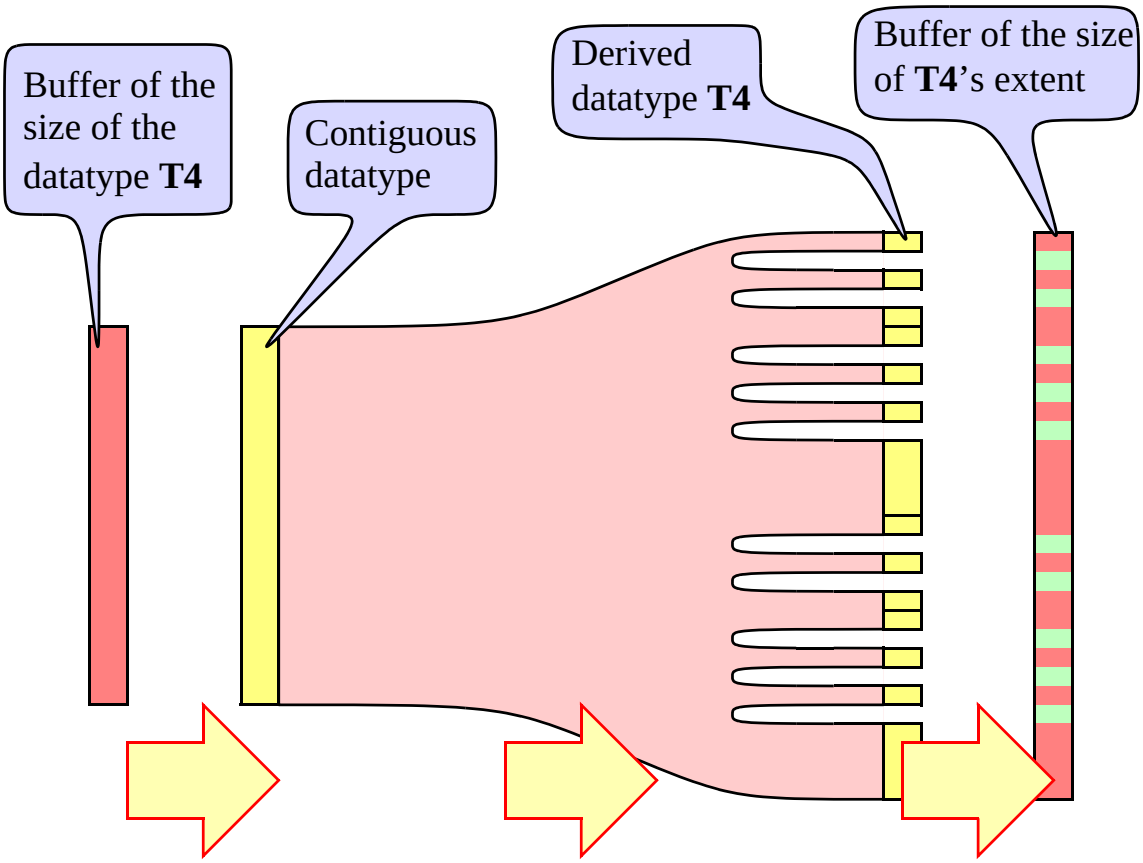
MPI-I/O Implementation



MPI-2 operations and the MPI-I/O subset in particular form an extension to MPI-1. However a developer of MPI-I/O needs access to the source code of the MPI-1 implementation, on top of which he intends to implement MPI-I/O. For each MPI-1 implementation, a specific integration of MPI-I/O is required.

.

Reverse Engineering



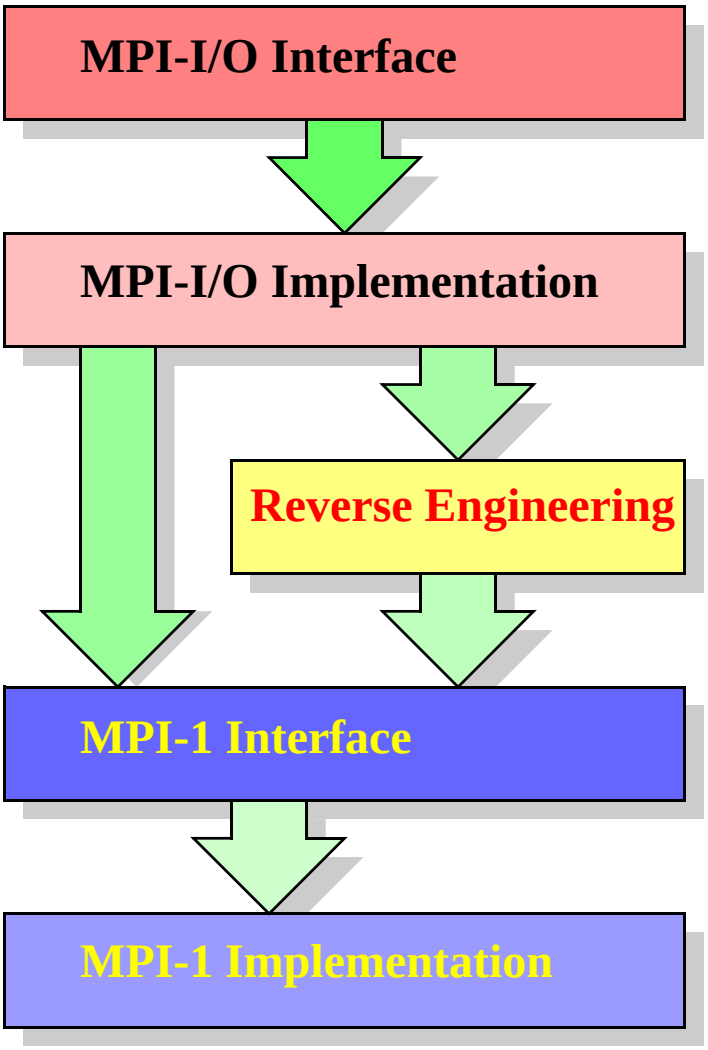
MPI_Send(source,size,MPI_BYTE,...)

MPI_Recv(destination-LB,1,**T4**,...)

MPI_Recv operation receives a contiguous network stream and distributes it in memory according to the data layout of the datatype. If the memory is previously initialized with a "green colour", and the network stream has a "red colour", then analysis of the memory after data reception will give us the necessary information on the data layout. Instead of sending and receiving, it is possible to use the MPI_Unpack standard MPI-1 operation.

.

Portable MPI-I/O Solution



Once we have a tool for derived datatype decoding, it becomes possible to create an isolated MPI-I/O solution on top of any standard MPI-1. The Argonne National Laboratory's MPICH implementation of MPI-I/O is used with our datatype reverse engineering technique and a limited subset of MPI-I/O operations has been implemented.

.

Conclusion

The presented isolated MPI-I/O package automatically gives to every MPI-1 owner an MPI-I/O, without any requiring to change or modify his current MPI-1 implementation.

Future work

- **Implementation of blocking collective file access operations.**
- **Implementation of non-blocking file access operations.**
- **Integration to EPFL's parallel file striping library SFIO.**

Thank You !