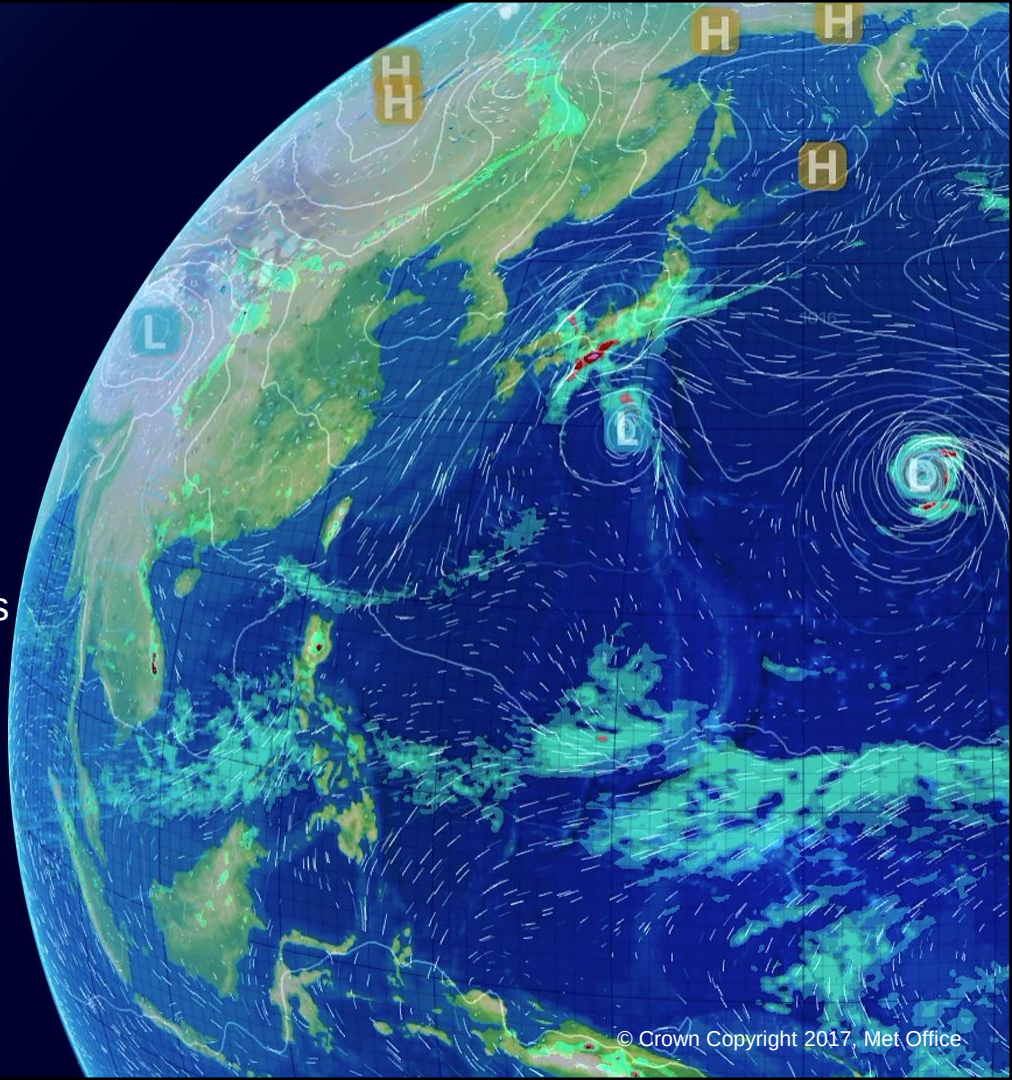
Met Office

Parallel I/O in the LFRic Infrastructure

Samantha V. Adams

Workshop on Exascale I/O for Unstructured Grids

25-26th September 2017, DKRZ, Hamburg.



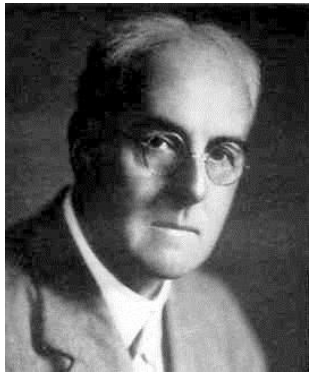
Met Office

Talk Overview

- Background and Motivation for the LFRic project
- Addressing two key issues:
  - Scalability
  - Flexible deployment for future HPC architectures
- Why parallel I/O is important in the early stages of LFRic
- Integration of the XIOS parallel I/O framework into LFRic
- Progress and headline results
- Next steps...and some challenges

Met Office

LFRic Background



- **L**ewis **F**ry **Ric**hardson
- GungHo Project Recommendations (*Met Office, NERC, STFC*)
- Aim to develop the science for a new dynamical core
  - Keep the best of current MO dynamical core (EndGame)
  - Improve where possible (e.g. Conservation)
- Scalability (Moore's Law => more cores). Current UM will not scale indefinitely
- What will future HPC architectures be like and how to adapt quickly?

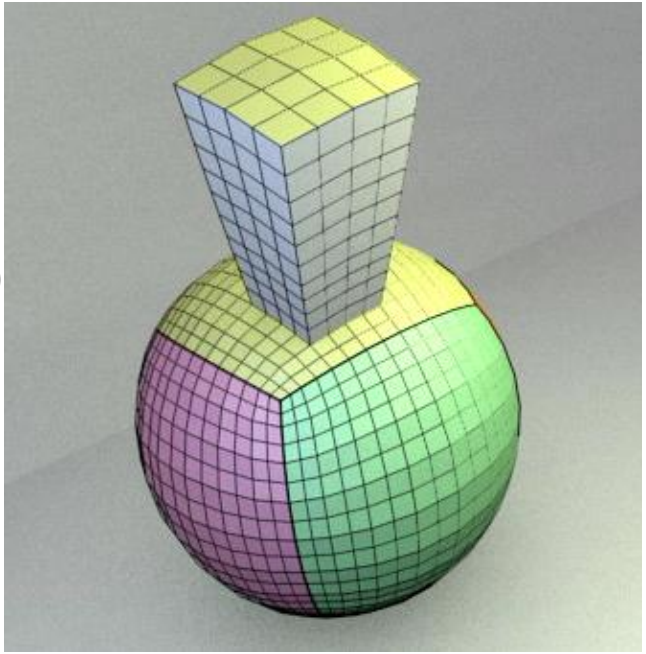
Scalability - Meshes

- Regular lat/lon meshes are a problem (pole singularities)
- From GungHo recommendations, use ***semi-structured*** meshes to get rid of poles…

Which sport do you prefer!?

Scalability - Meshes

- LFRic infrastructure potentially supports **unstructured** meshes
- Currently mainly use Cubed Sphere (and also planar biperiodic)
- Cubed Sphere is **semi-structured** in horizontal
- Created externally to model as 2D, UGRID format
- Partitioned inside LFRic model and extruded to 3D



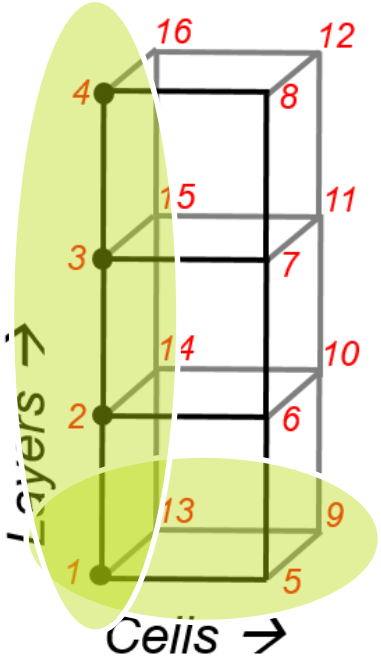
Scalability – FEM formulation

- Regular lat/lon meshes give nice properties, so need to use different computational methods
- Mixed Finite Element methods can retain desirable properties for a dynamical core (Cotter and Shipton, 2012; Staniforth and Thuburn, 2012)
- Data held on different elements of mesh: nodes, edges, faces
- Currently lowest-order FEM (but LFRic infrastructure supports higher order)

Cotter, C.J. and Shipton, J. 'Mixed finite elements for numerical weather prediction'. J. Comput. Phys.,2012.
Staniforth, A. and Thuburn, J. 'Horizontal grids for global weather prediction and climate models: a review'. Q. J. R. Meteorol. Soc., 2012.

Scalability – Data layout

- As the horizontal mesh is now (potentially) unstructured, we have indirect addressing in the horizontal
- To maintain reasonable cache use and sensible vector lengths, the dynamical core uses a ‘k-contiguous’ data layout - Data points (‘dofs’) are column ordered
- Work currently done on whole columns
- Some impact on I/O as the data needs restructuring into a ‘layer ordered’ format



Met Office

Flexible Deployment

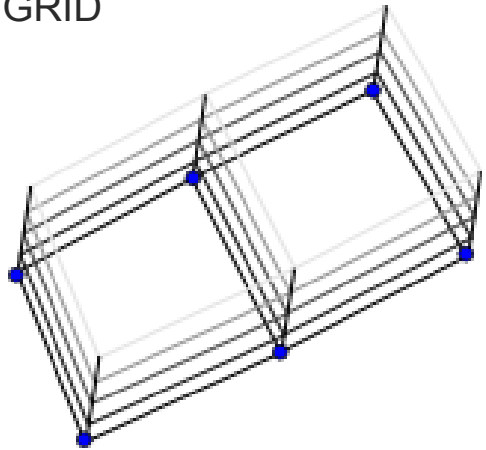
- Future HPC architecture ? (we know many more cores, but CPU, GPU, hybrid??)
- 'Separation of concerns'. Allows scientists to write science without worrying about machine architecture and specific optimisations
- LFRic solution is **PSyclone** – developed by our collaborators **STFC Daresbury, UK**.
- PSyclone is a python based Fortran parser and code generator – takes 'science' code, examines and processes it to apply appropriate optimisations
- Currently applies MPI by default, OpenMP optionally. Currently investigating OpenACC for GPU

Parallel I/O

- Not considered in GungHo or early stages of LFRic, but becoming more important as we need to:
  - Assess likely impact on compute performance
  - Facilitate Science tests on larger jobs
  - Provide information to other Met Office teams and UM partners about our future output formats
- Prime requirement is to handle parallel read/write efficiently and be scalable (i.e. not destroy the compute performance we are working hard to achieve!)
- Decided to adopt an existing framework rather than write from scratch

File formats

- As mentioned previously LFRic mesh is in 2D UGRID format
- Output diagnostic format is currently ‘3D layered’ UGRID



```
dimensions:
nMesh2_node = 6 ; // nNodes
nMesh2_edge = 7 ; // nEdges
nMesh2_face = 2 ; // nFaces
nMaxMesh2_face_nodes = 4 ; // MaxNumNodesPerFace
Mesh2_layers = 10 ;

Two = 2 ;

variables:
// Mesh topology
integer Mesh2 ;
Mesh2:cf_role = "mesh_topology" ;
Mesh2:long_name = "Topology data of 2D unstructured mesh" ;
Mesh2:topology_dimension = 2 ;
Mesh2:node_coordinates = "Mesh2_node_x Mesh2_node_y" ;
Mesh2:face_node_connectivity = "Mesh2_face_nodes" ;
Mesh2:face_dimension = "nMesh2_face" ;
Mesh2:edge_node_connectivity = "Mesh2_edge_nodes" ;
Mesh2:edge_dimension = "nMesh2_edge" ;
Mesh2:edge_coordinates = "Mesh2_edge_x Mesh2_edge_y" ;
Mesh2:face_coordinates = "Mesh2_face_x Mesh2_face_y" ;
Mesh2:face_edge_connectivity = "Mesh2_face_edges" ;
Mesh2:face_face_connectivity = "Mesh2_face_links" ;
Mesh2:edge_face_connectivity = "Mesh2_edge_face_links"
integer Mesh2_face_nodes(nMesh2_face, nMaxMesh2_face_nodes) ;
Mesh2_face_nodes:cf_role = "face_node_connectivity" ;
Mesh2_face_nodes:long_name = "Maps every face to its corner nodes." ;
Mesh2_face_nodes:_FillValue = 999999 ;
Mesh2_face_nodes:start_index = 1 ;
integer Mesh2_edge_nodes(nMesh2_edge, Two) ;
Mesh2_edge_nodes:cf_role = "edge_node_connectivity" ;
Mesh2_edge_nodes:long_name = "Maps every edge to the two nodes that it connects." ;
Mesh2_edge_nodes:start_index = 1 ;
```

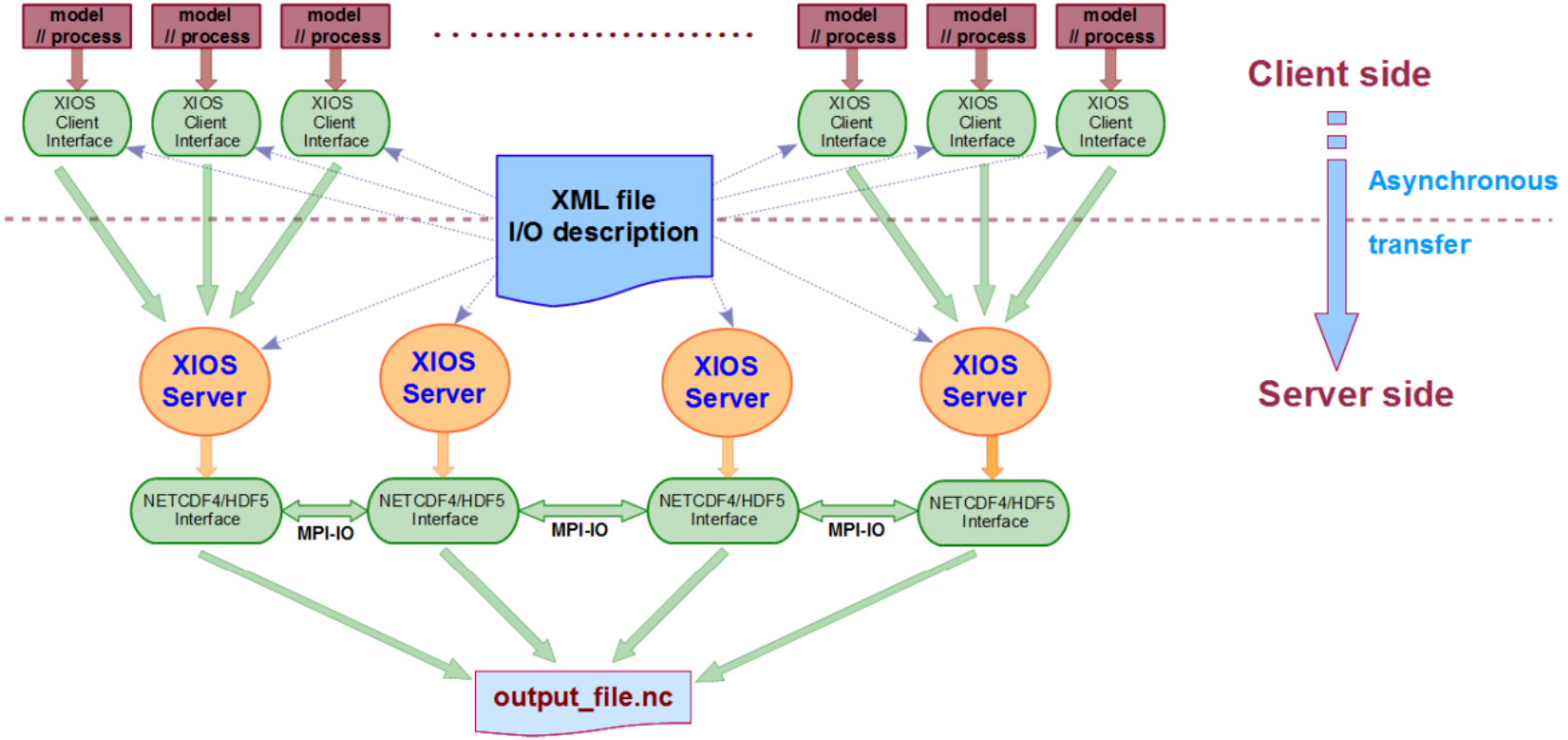
http://ugrid-conventions.github.io/ugrid-conventions/#3d-layered-mesh-topology

Why XIOS?

- There are a few candidate parallel I/O frameworks
- XIOS already supports parallel read and write
- Proven on jobs ~10K cores in weather and climate domain
- Already in use in Met Office (NEMO ocean model)
- Works with OASIS coupler now and coupling functionality is being added
- Prior to last year, no frameworks supported UGRID output
- We have collaborated with IPSL to add UGRID support for LFRic

The XIOS parallel IO framework

- Authored by Institut Pierre Simon Laplace
- Freely available from http://forge.ipsl.jussieu.fr/ioserver
- Client-Server IO architecture that can run with or without server (but requires server for fully asynchronous parallel IO)
- Supports on-the-fly processing operations such as regridding, daily and monthly meaning
- Supports unstructured grids
- Output formats CF netCDF and (since our work) UGRID extension to netCDF



Met Office

Progress over the last year

- **XIOS development to output UGRID in serial and parallel (IPSL)**
- **Integration of XIOS into LFRic infrastructure. Not trivial!**
  - FEM computational space → real world coordinate space with data on appropriate mesh elements
  - Ensuring that parallel == serial. Is all the data ***and*** topology in the correct place?
- **Preliminary performance evaluation with new UGRID output**
  - Does XIOS show reasonable scaling?
  - What happens when 100s of fields / 100s of Gigabytes of data are output to one file?

Headline Results

Scaling

- Run out to 14k cores with little/no I/O penalty. This is more cores than previous benchmark jobs with XIOS (NEMO ocean model, ~8k cores) ***BUT*** much smaller mesh / lower resolution and lower output frequency

How many I/O servers to use?

- Generally more == better. For a fixed size job (3,456 cores), increasing XIOS servers reduces client wait time and no increase in overall run time…..BUT…

Impacts of a Lustre file system?

- With appropriate striping, can achieve low client wait time with fewer I/O servers

Diagnostic output loading

- With each 100 field (~112Gb) increase, approx +5% I/O penalty

Next Steps

- We are optimistic, but not complacent as there is still much to do!
- Yearly LFRic compute performance report will now include I/O as part of the benchmark
  - LFRic science will change
  - Optimisations will change
- Tuning, tuning, tuning…for larger jobs with more cores, longer runs, more frequent output, etc.
- We have done the ‘O’ part but not the ‘I’…reading files in parallel is also important

Our Challenges

- We use ESMF for halo exchange operations on partitioned mesh
  - ESMF ascribes a globally-unique 32-bit integer index to each data point. When LFRic runs with high resolution, higher order and/or many vertical levels, it quickly has more data points than there are 32-bit integers!
- UGRID diagnostic output format
  - Currently convert to a 'finite volume' equivalent to hold data on cells/faces and with face/node topology.
  - Essentially throwing away information!
  - We are exploring ‘mimetic post-processing’ with UM partner organisation NIWA
- Visualisation of unstructured netCDF / UGRID
  - What options are available?
  - I have been in contact with Felicia Brisc and Niklas Röber about Paraview plugins already!

Acknowledgements

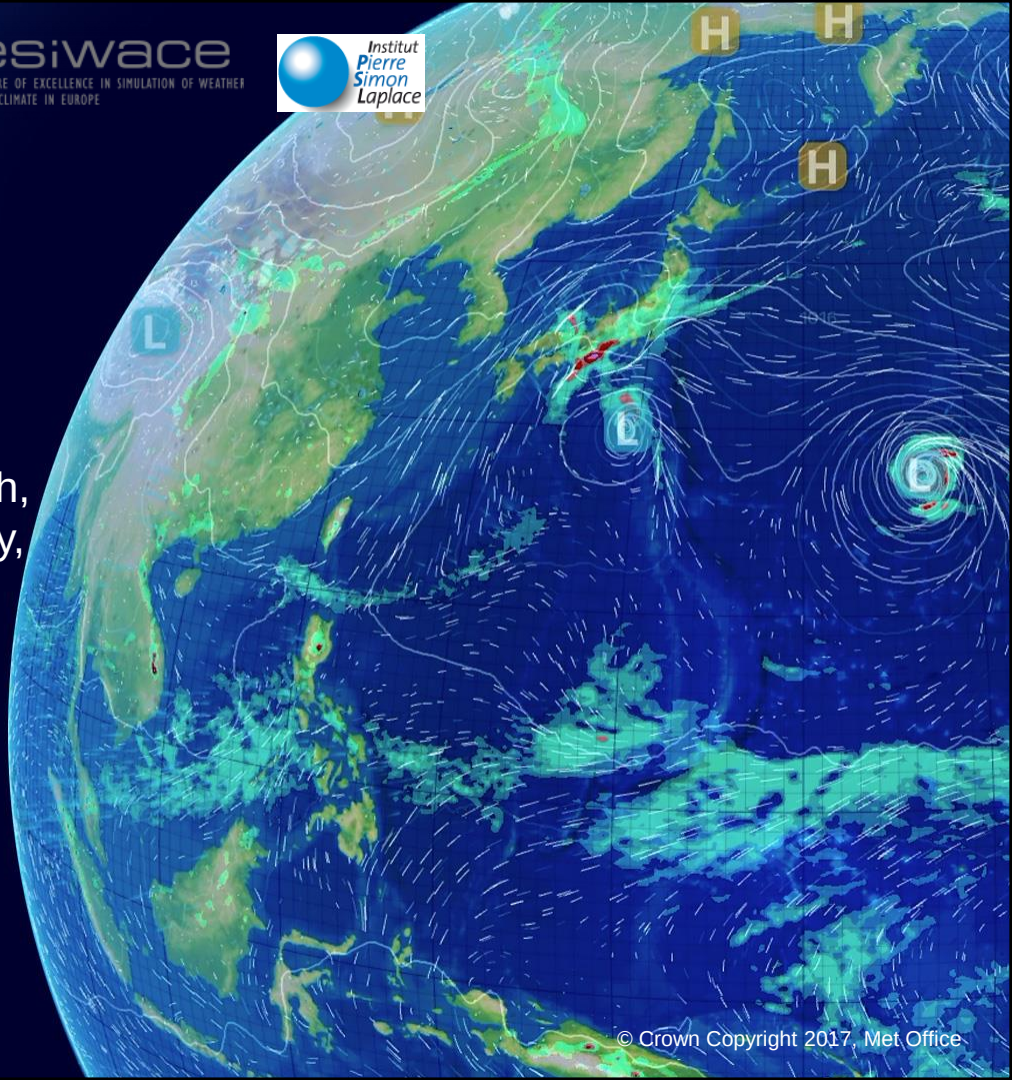
Met Office UK LFRic team:
Sam Adams, Tommaso Benacchio, Matthew Hambley, Mike Hobson, Iva Kavcic, Chris Maynard, Tom Melvin, Steve Mullerworth, Stephen Pring, Steve Sandbach, Ben Shipway, Ricky Wong

STFC (Hartree Centre), UK:
Rupert Ford, Andy Porter

University of Bath, UK: Eike Mueller

Monash University, Australia: Mike Rezny

IPSL (LSCE/CEA), France: Olga Abramkina, Yann Meurdesoif



Met Office

Thank You!

Questions?