Investigation of parallel IO module in WRF

MuQun Yang, Mike Folk

NCSA HDF Group
May 8, 2003

1. Introduction

The Weather Research and Forecasting (WRF) Model is a limited-area weather model for research and prediction. It uses a multi-layer domain decomposition method to run on distributed computing system. MPI and OpenMP are used in WRF for parallel computing. Currently there is no real parallel IO module in the model to boost I/O performance. Furthermore, it is not even known whether parallel IO can even work in the model. The purpose of this report is to provide some insights on the parallel IO module in WRF.

The contents of the paper include

- to investigate the current WRF IO module,
- to describe the requirement of adding a parallel IO module in the model,
- to provide possible approaches for parallel HDF5-WRF IO module
- to discuss the current approach

2. Investigation of current WRF IO module

There are several IO modules, here we only discuss two of them: NetCDF and IO quilt server.

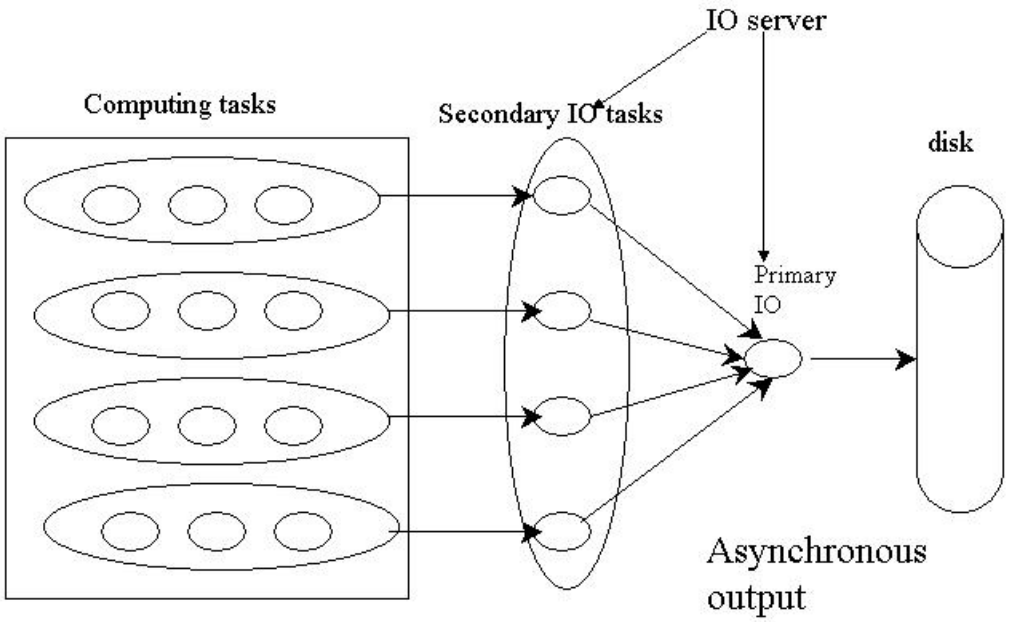
1) NetCDF

Only sequential NetCDF IO is provided due to limitation of the NetCDF library. The WRF I/O module assigns a node to gather and scatter data from other nodes for writing and reading.

2) IO quilt server

IO quilt server does not use MPI-IO. It uses separate I/O nodes to generate output asynchronously while computing continues in computing nodes. The working procedure is illustrated in the schematic below:

IO Quilt server schematic



IO Quilt server can improve performance. Sometimes it can also waste some computing resources. If IO is too slow, computing nodes have to wait until the finishing of IO in the previous timestamp.

3. Requirement of parallel HDF5-WRF IO module

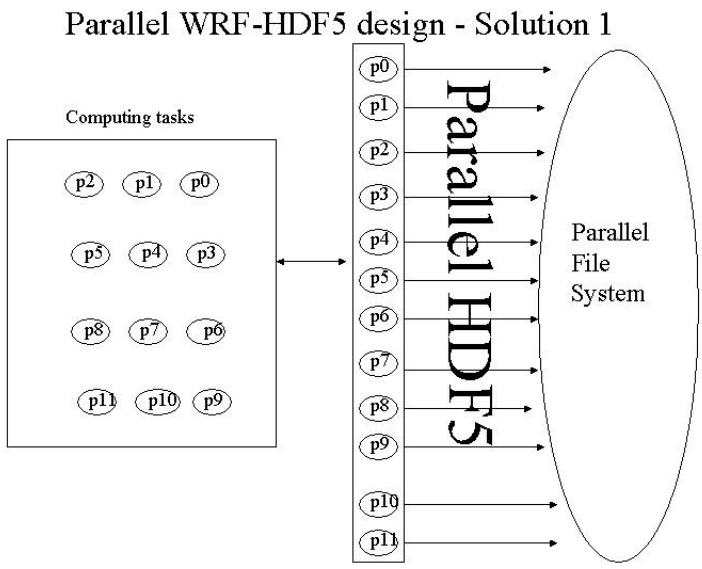
Since HDF5 is the only data format that supports MPI-IO, we would like to investigate the requirements of parallel HDF5-WRF IO module. After reading the WRF software design document and the model source codes, we find WRF I/O to be sufficient for supporting parallel I/O; there are no additional requirements for adding parallel HDF5 IO module to WRF from the WRF model side.

4. Three possible approaches for parallel HDF5-WRF IO module

Enlightened by IO quilt server, we proposed three possible approaches for parallel HDF5-WRF IO module.

Approach 1

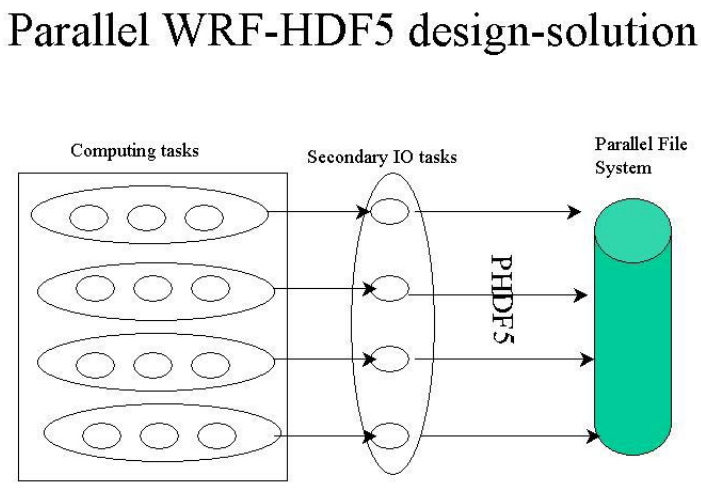
Every node will be used for IO through HDF5. The schematic is shown below:



This approach may be easy to implement, but with many processors, overhead may be a problem.

Approach 2

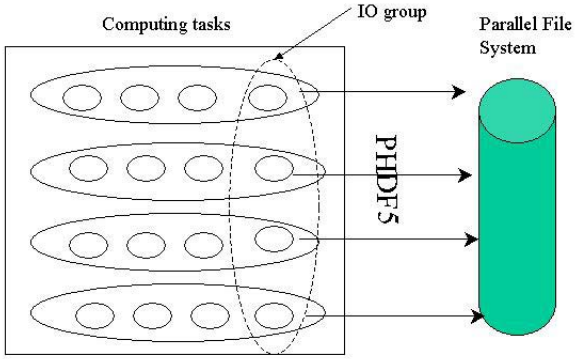
Like I/O quilt server, we divide nodes into computing nodes and I/O nodes. WRF data fields are written to the disk through I/O nodes via parallel HDF5, as illustrated below. The I/O overhead small in this case since there are fewer nodes doing I/O, and thus the approach is more scalable. However, extra work is needed to transfer data from computing nodes to IO nodes. For compute bound simulations, the use nodes exclusively for I/O can waste computing resources.



Approach 3

This approach is similar to approach 2, but all nodes will be used to compute; some nodes will also be used for IO.

Parallel WRF-HDF5 design-solution 3



5. The current approach

HDF5 will implement two-phase IO in its parallel module in the future and the two-phase IO idea is very similar to that in approach 2 and approach 3. Therefore, we have decided not to implement approach 2 or 3 at this time. Instead, since there is a short-term need for parallel I/O for WRF, we decided to implement approach, which is easy and may actually give good performance. We will report our performance result and challenges we met in a later report.