Bringing Compute to the Data

Bryan Lawrence
National Centre for Atmospheric Science &
Department of Meteorology, University of Reading &
Centre for Environmental Data Archival, STFC Rutherford Appleton Laboratory
(bryan.lawrence@ncas.ac.uk)

January 13, 2015

Bringing analysis computation to data will be a necessary part of next generation earth system simulation workflows. Massive data growth is expected, in the presence of slower growth in storage technology. Systems based on hybrid batch-cloud environments provide a suitable environment into which such computation can be brought, but there are technical challenges ahead in order to provide satisfactory storage capacity and performance.

Introduction

In 2012, the US National Academy produced a report advancing a national strategy for climate modelling (Dunlea & Elfring, 2012), which amongst other things noted that

> "Without substantial research effort into new methods of storage, data dissemination, data semantics, and visualization, all aimed at bringing analysis and computation to the data, rather than trying to download the data and perform analysis locally, it is likely that the data might become frustratingly inaccessible to users"

The context of course, is the massive increases of data associated with international model intercomparison projects (MIPs), such as CMIP5 (Taylor et al., 2011) and planned successors, such as CMIP6 (Meehl et al., 2014).

It is difficult to accurately predict the volumes of data which will be produced for CMIP6, but initial estimates based on numbers of years to be simulated suggest a global archive of around 10PB. However, interest in extended diagnostics and higher temporal resolution output could lead to higher volumes of output, without requiring more simulated years. These estimates suggest that *for global model inter comparison projects*, output will increase by about a factor of 10 in the 6 or 7 years between the two MIPs (by contrast, from CMIP3 to CMIP5, volumes increased by a factor of 50 or so over a similar period).

Table 1: *Storage Trends in terms of doubling (halving) times (from a paper in preparation, and not expected to be very precise numbers, caveat lector).*

Requirements & Infrastructure	
CMIP requested output	15-22 months
NCAR archive	29 months
DKRZ archive	23 months
DKRZ disk	10-15 months
JASMIN disk	12 months
Storage	
Capital Cost of Disk	(29 months)
(Historic) Bandwidth to disk	28-35 months
(Future) Bandwidth to disk	up to 90 months
Bandwidth to Tape	32 months

With all the appropriate caveats associated with extrapolating from two imperfectly known points, one might conclude that the data volumes associated with global climate modelling have a doubling time of between 15 and 22 months. These are not inconsistent with the historic doubling times in archive storage capacity and storage bandwidth at a selection of major institutions — Table 1 — but looking ahead, most places report expectations of needing faster storage growth to copes with their entire predicted workload (beyond global MIPs).

In the remainder of this short position paper, prepared for the 2015 Barcelona Big Data and Extreme Computing (BDEC) workshop, we present a discussion of storage infrastructure, before a discussion of the required architecture for a computing environment for exploiting earth system simulation data.

Storage Infrastructure

Like compute, storage components have benefited from many years of increasing performance, but like clock speed, rapid increases in disk bandwidth have probably stalled. The best way to get improved performance from disk I/O is from parallelisation. It looks like there is still mileage in improvements in tape performance, but it is clear that these improvements are likely to lag behind the data production rate, so although tape can (non-intuitively) have higher bandwidth than disk, the relative bandwidth from tape is falling.

Disk

I/O bandwidth has always been a problem for earth simulation, even in the presence of parallel file systems. The scale of the problem can easily be seen by considering the performance of a state-of-the-art file system in an I/O benchmark - Figure 1. We can see that the aggregate bandwidth to the storage scales linearly, but the bandwidth can easily be swamped by relatively few compute nodes. This much is not news, but the consequences need to be thought through.

In a typical simulation workload, I/O begins and ends a block of work, and in between there is a long period of compute - which means that I/O can be hidden by using I/O server technology (e.g `http://forge.ipsl.jussieu.fr/ioserver/`) in parallel with compute loads. However, we assume that such hiding isn't nearly as possible in analysis workloads. While we don't yet have the numbers to back up this assumption, it seems reasonable to assume that many analysis tasks are I/O bound, with relatively little computation behind which one could hide I/O.

The filesystem itself is probably a significant constraint too, removing POSIX is a desirable goal. However, even without going that far, and recognising that some techniques to accelerate I/O such as keeping metadata on SSD are less efficient for the large files that we would like to use, we are considering developing new techniques to accelerate access to key formats such as NetCDF.

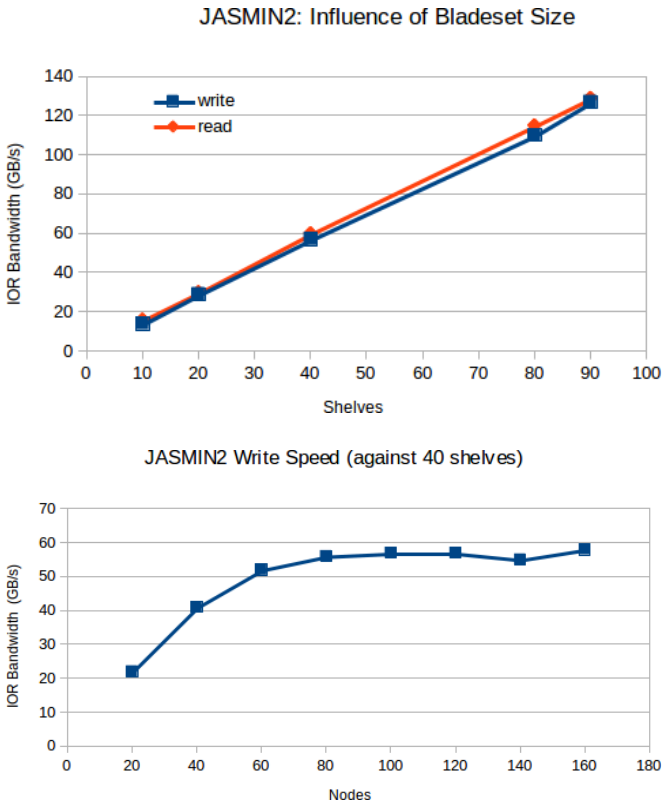

Figure 1: *IOR benchmark results from the JASMIN acceptance testing. The top panel shows the influence of the size of the storage pool (bladeset size, measured in shelves) on bandwidth - effectively performance scaled linearly. The bottom panel shows that for a given storage pool, it's very easy to swamp the filesystem with relatively few compute nodes.*

Tape

It is important to note that traditional disk to tape backup is not possible or desirable at and above petascale with analysis workloads — there is too much volatile data on the disk which occurs as intermediate data products. Such data is and should be transient, and which if backed up would simply reside in backup un-needed and un-read.

Tape is mostly used in the simulation community as long-term archive, providing file-level storage (e.g. the CERN CASTOR tape system which we use, `http://castorwww.web.cern.ch`), although both the UK Met Office and the European Centre for Medium Range Weather Forecasting (ECMWF) have sophisticated bespoke systems which allow tape operations to access subset data (MASS and MARS respectively). In both cases, the use of sophisticated systems mean that they have smaller analysis disk systems than they would otherwise need - but this comes at the cost of maintaining sophisticated bespoke systems and constraints on file content and layout.

Looking forward, our projections of storage requirements suggest that the physical size of disk subsystems, coupled with energy demands will further limit the proportion of data we can store on parallel disk, which means we are likely to make more use of tape within active workflows as is done with MARS and MASS. There is scope for an open-source alternative.

While disk-to-tape backup isn't desirable, reliability within the tape system is important. As we reach exabytes of data, the desirability of providing such reliability by mirroring data becomes less attractive, and alternative methods built on Redundant Array of Independent Tape to use erasure codes become desirable. However, they have their drawbacks too, not least they can influence performance via fragmentation, and diminishing the potential concurrent service load (one process will monopolise multiple drives) - although some solutions are becoming apparent Cappello et al., 2011). Solutions which provide both reliability and increased throughput would be desirable!

A Computing Infrastructure for Exploiting Simulation Data

The solution to avoiding download, is clearly to bring the compute to the data, which has been a mantra for years. However, there are important subsidiary questions. What if the data isn't in one place? How does one provide a general purpose platform in that context - what characteristics, in terms of hardware and software should it have?

Figure 2 describes a generic environment that maps onto the specific environment we have deployed with our JASMIN platform (Lawrence et al., 2013), serving (a heterogeneous) UK environmental science community. The key assumptions we have made in this design are:

1. We should bring all the data to one place, and given we have multiple HPC platforms, there isn't a special advantage in co-locating with one of them (so we haven't).

2. Different user communities will need different types of hardware for analysis, so we provide a range of compute nodes with memory (currently) up to 2 TB per node in the fat nodes.

3. We expect that users need to control what data they have on tape, so we have provided an "elastic tape" facility, where users decide what is on tape, and how much.

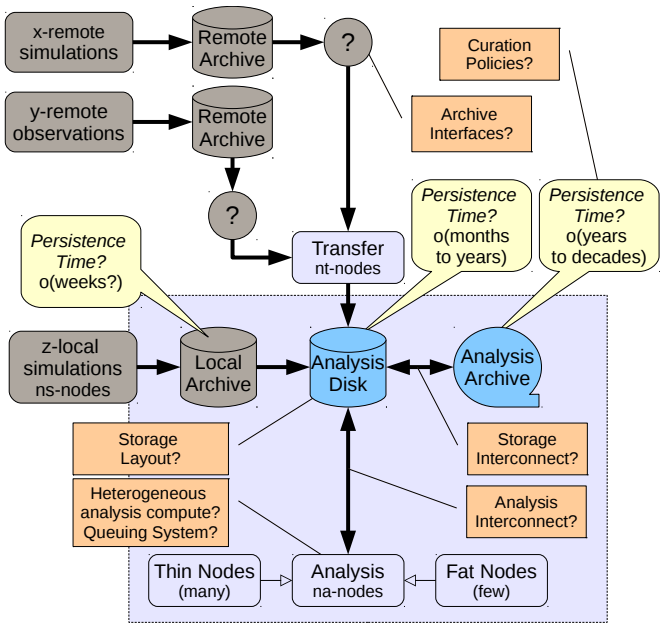

Figure 2: *A generic environment for supporting analysis workflow for earth system analysis. In this diagram, data flow is shown using solid black lines. We start from the assumption that it is no longer possible to assume that all the data is locally generated - we will be dealing with x remote simulation datasets and y observational datasets. It is assumed that the analysis compute will be heterogenous, and that the the persistence time of data in the archive will be long enough to need active curation to avoid massive duplication.*

4. We need dedicated network links to key remote sites (in our case, HPC platforms in Edinburgh and Exeter), key partner sites (at universities), and possibly soon, ESA. Such network links need to exploit fast transfer protocols, which of course, depend as much on the remote sites, as on us.

5. We expect residence times for large datasets to be of the order of months to years on the analysis disk, and years to decades in the archive. Accordingly, storage policies are important, as are formal digital curation concerns.

We think these generic principles will remain for any analysis platform. However, there are more detailed configuration choices that we have made, and that we expect to differ for other big data analysis environments, or indeed, in our own environment in the future. These include:

1. Compute and Queuing: We have made the choice to split our compute into a traditional batch environment (with a specific software installed) and a more flexible cloud environment.

Apart from meaning we are using two very different schedulers, this means we do not have to be responsible for all the possible system environments needed by our users. We divide our cloud into two notional types of systems: the managed environment (where we control the systems) an the unmanaged environment (where we offer Infrastructure as a Service, and remote individuals control the systems).

2. We have chosen to exploit parallel file systems for the bulk of our disk resource.

We believe a parallel file system gives the most flexibility for users who can exploit our managed cloud (being highly performant, but flexible). We hope this will help us avoid some of the seven deadly sins of cloud computing (Schwarzkopf et al., 2012), insofar as we have optimised our environment for a very varied environment, not cherry picked it for one or two applications. The unmanaged systems currently use bulk disk deployed on servers, since there are security issues with parallel access from unmanaged systems that we have yet to address - but they can still access data from the fast disk via applications running in the managed cloud.

3. The internal analysis interconnect consists of a fully unblocked 10Gbit ethernet network.

We have a notional internal bandwidth of several TBit/s, but while this is all one filesystem, it is broken into storage pools (e.g. see Figure 1), which together aggregate to that performance. These storage pools allow us to guarantee users of one portion of the filesystem un-hindered I/O performance (at least on the server side — there are issues to be understood on the client side from the cloud, e.g massive I/O jitter as noted in Armbrust et al., 2010).

4. The internal interconnect between disk and tape currently uses a lightweight home-grown "elastic-tape" protocol layered over CASTOR.

The current JASMIN system is supporting a very heterogeneous and growing community with good results, but it is clear that there are problems ahead. In particular, we do not have a sound theoretical basis to choose and balance our load scheduling/virtualisation methods. We do not understand the details of I/O performance in our multi-user environment, so we do not know if we are getting good value from our filesystem choices. We know we cannot continue to add disk at the rate our projections require, if for no other reason than we will eventually need a new building - at some point the right answer is not new buildings (whether ours, or those in a public cloud). The right answer must include better use of tape in our workflow, which needs better tape software exploiting sophisticated reliability and performance strategies. Finally, we know we have to better understand and improve the commonly used analysis algorithms, particularly as bandwidth becomes more problematic into the future.

Some have suggested that when dealing with petascale systems, academia should learn from industry, where they are already handling exabytes. While this is undoubtedly true, many industrial applications are of the sift/query-return-result nature, and few industrial environments provide hundreds of users access to petabytes of data for analysis in a common environment, where such analysis may consist of handling and generating petabytes, within the workflow. To that end, there is much learning yet to be done.

References

Armbrust, M., Fox, A., Griffith, R., Joseph, A. D., Katz, R., Konwinski, A., ... Zaharia, M. (2010, April). A View of Cloud Computing. *Commun. ACM*, *53*(4), 50–58. doi: 10.1145/1721654.1721672

Cappello, F., Jacquelin, M., Marchal, L., Robert, Y., & Snir, M. (2011). Comparing archival policies for Blue Waters. In *High Performance Computing (HiPC), 2011 18th International Conference on* (pp. 1–10). IEEE. doi: 10.1109/HiPC.2011.6152428

Dunlea, E., & Elfring, C. (Eds.). (2012). *A National Strategy for Advancing Climate Modelling*. The National Academies. Retrieved from `http://www.nap.edu/catalog.php?record_id=13430`

Lawrence, B., Bennett, V., Churchill, J., Juckes, M., Kershaw, P., Pascoe, S., ... Stephens, A. (2013, October). Storing and manipulating environmental big data with JASMIN. In *2013 IEEE International Conference on Big Data* (pp. 68–75). doi: 10.1109/BigData.2013.6691556

Meehl, G. A., Moss, R., Taylor, K. E., Eyring, V., Stouffer, R. J., Bony, S., & Stevens, B. (2014). Climate Model Intercomparisons: Preparing for the Next Phase. *Eos, Transactions American Geophysical Union*, *95*(9), 77–78.

Schwarzkopf, M., Murray, D. G., & Hand, S. (2012). The seven deadly sins of cloud computing research. In *Proceedings of the 4th USENIX conference on hot topics in cloud computing*. USENIX Association. Retrieved from `http://dl.acm.org/citation.cfm?id=2342763.2342764`

Taylor, K. E., Stouffer, R. J., & Meehl, G. A. (2011, October). An Overview of CMIP5 and the Experiment Design. *Bulletin of the American Meteorological Society*, *93*, 485–498. doi: 10.1175/BAMS-D-11-00094.1

Acknowledgements: Much of the thinking discussed here has arisen from conversations with Jonathan Churchill, Phil Kershaw, and Matt Pritchard, of STFC, and colleagues in the EC FW7 IS-ENES consortium.