ECMWF Feature article

from Newsletter Number 143 – Spring 2015

COMPUTING

Supercomputing at ECMWF

tounka25/iStock/Thinkstock

www.ecmwf.int/en/about/news-centre/media-resources

doi:10.21957/szfnyqb5

This article appeared in the Computing section of ECMWF Newsletter No. 143 – Spring 2015, pp. 32–38.

Supercomputing at ECMWF

Mike Hawkins, Isabella Weger

ECMWF's latest High-Performance Computing Facility (HPCF), two Cray XC-30s with over 160,000 processor cores in a resilient configuration, is at the heart of ECMWF's production of weather forecasts and cutting-edge research in numerical weather prediction (NWP). The two clusters, which are among the most powerful supercomputers in Europe, have been producing ECMWF's operational forecasts since September 2014. In addition to ECMWF's operational and research activities, ECMWF's Member States also have access to the HPCF and scientists may apply to use it for Special Projects. Figure 1 shows 'Ventus', one of the two Cray clusters installed in ECMWF's data centre.

Numerical weather prediction has always relied on state-of-the-art supercomputers to run a complex numerical model of the atmosphere in the shortest possible period of time. ECMWF's current Integrated Forecasting System (IFS) is based on a numerical model with 293 million grid points, ingests 40 million observations per day and takes 2 hours and 10 minutes to produce a 10-day high-resolution global forecast.

ECMWF's first operational forecast in 1979 was produced on a single-processor Cray-1A. The fastest supercomputer at its time, it had a peak performance of 160 million floating-point operations per second, around a tenth of the computing power of a modern smartphone. The peak performance of the Cray XC-30 system is 21 million times greater, equivalent to a stack of smartphones more than 15 kilometres tall.

The demand for more accurate and reliable forecasts and for better early warnings of severe weather events, such as windstorms, tropical cyclones, floods and heat waves, requires continual improvements of ECMWF's numerical models and data assimilation systems. Finer model grid resolutions, a more realistic representation of physical processes in the atmosphere, and the assimilation of more observations are the main drivers for better computational performance. ECMWF's forecasting system has also developed towards a more comprehensive Earth-system model: the atmospheric model is coupled to ocean, wave, sea ice and land-surface models and now includes the composition of the atmosphere (e.g. aerosols and greenhouse gases).

The growing computational requirements resulting from ECMWF's scientific and operational strategy require ECMWF to replace and upgrade its HPCF on a regular basis. Competitive procurements are run every four to five years and contracts have built-in upgrade cycles of about two years to take advantage of improvements in technology and to better match the system to operational and research needs.

Figure 2 shows the evolution of sustained HPC performance at ECMWF, with the Cray XC-30 as the most recent system. Sustained performance is measured by ECMWF benchmark codes that represent the current operational version of the IFS.

Figure 1 'Ventus' is one of the two Cray XC-30 clusters installed in ECMWF's data centre.

doi:10.21957/szfnyqb5

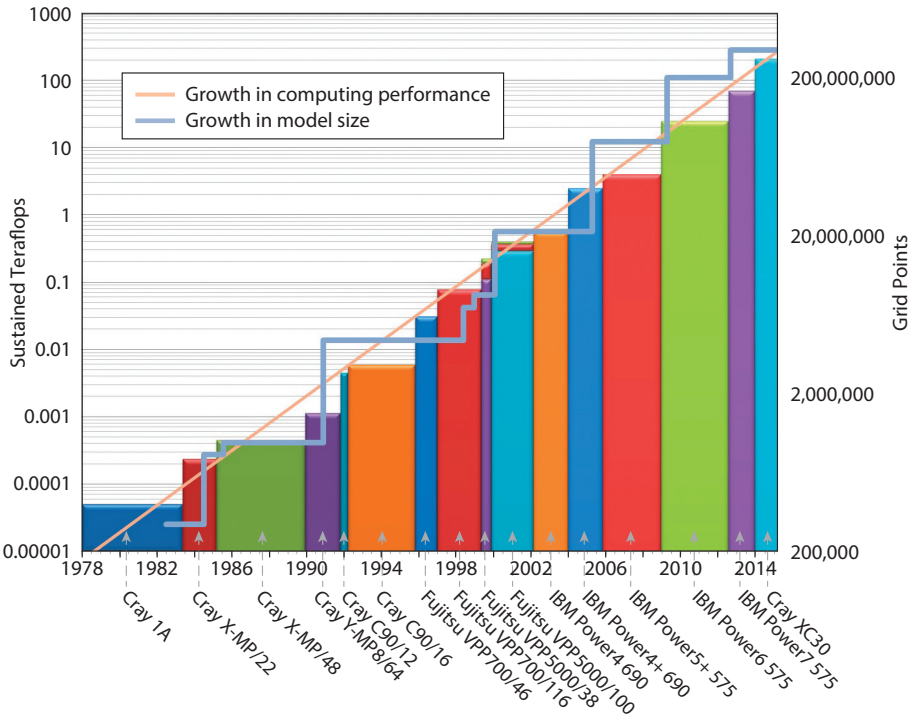

Figure 2 HPC growth versus size of ECMWF's high-resolution forecast model.

Looking back to the beginnings of supercomputing

From 1976 to 1978, ECMWF leased access to a Control Data Corporation (CDC) 6600 computer, hosted by CDC at John Scott House in Bracknell. The CDC6600 was one of the most powerful systems available at the time and considered the first successful supercomputer. It allowed the development of the first version of ECMWF's weather forecasting model, but it still needed 12 days to produce a 10-day forecast.

The CDC6600 experience showed that, provided a suitably powerful computer could be acquired, useful forecasts could be produced. Figure 3 shows an example architecture envisaged at the time. The first supercomputer owned by ECMWF was installed on 24 October 1978 at the new Shinfield Park site. This system was a CRAY-1A, serial number 9, manufactured by Cray Research. Before then, the Centre's scientists also had access to an IBM 360/195 at the British Meteorological Office and later to 'Serial 1', hosted at the Rutherford Laboratory, the first production model of the CRAY 1 series, in order to test out all the programs required to produce a real operational forecast.

From the 70s to the 90s – the first Cray era

The Cray-1A was a single-processor computer with a memory of eight megabytes and a 2.4 gigabyte storage system. The processor could produce two results per cycle, with a cycle time of 12.5 nanoseconds, giving a theoretical peak performance of 160 megaflops (160 million arithmetic calculations per second), about one tenth the performance of a modern smartphone. Running the operational weather model, the machine was capable of a sustained performance of 50 megaflops, allowing an operational 10-day forecast to be produced in five hours.

The era of Cray systems at ECMWF lasted 18 years, until 1996. In that time, a succession of systems advanced the total sustained performance the model could achieve from 50 megaflops to 6,000 megaflops. Despite the relatively long time period, most of the systems were quite similar in design. They all had a small number of fast processors, 16 in the last system, and each of these processors had access to all of the memory on the machine. This 'shared memory' configuration is the basic building block of the large systems we use today.

The other important feature of the Cray systems was the use of vector instructions, single instructions that could work on single-dimensional arrays of data. With a vector instruction, to add 10 numbers together, the set of 10 numbers is loaded and then added up in one go. This parallelism gives much better performance than doing the additions one after the other in a loop, as would be done on a 'scalar' system. Vector instructions are also used in modern processors. They were the key building block of the next era of supercomputing at ECMWF, the Fujitsu vector systems.

doi:10.21957/szfnyqb5

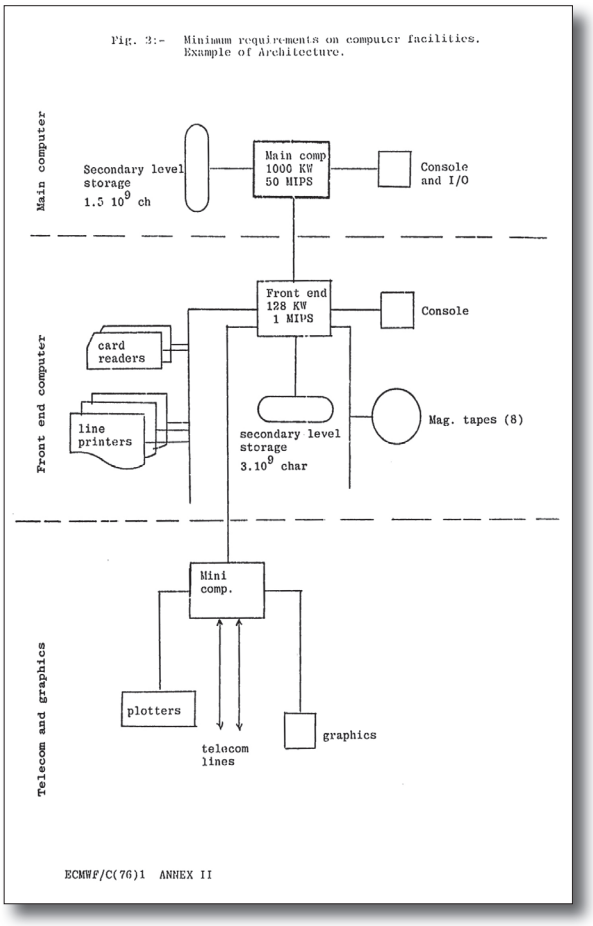

Figure 3 Example architecture to meet the minimum requirements of the computer facility – from a paper to the ECMWF Council, April 1976.

The last of the Cray machines was quite different from the others. Delivered in 1994, the Cray T3D was a massively parallel system. Rather than the 16 specially designed and built processors of its predecessor, the system had 128 Alpha processors, produced by Digital Equipment Corporation (DEC). Each processor had 128 megabytes of memory. The name of the system came from the network, which was arranged in the shape of a torus. With the memory distributed between the processors rather than shared into one big pool, substantial changes to the weather forecasting system were required to operate efficiently on this type of architecture.

From the 90s to 2002 – the Fujitsu era

In 1996, a small Fujitsu test system, a VPP300/16, was installed. This was followed by a larger VPP700/46 system, which started producing operational forecasts from 18 September 1996. The VPP700 had 39 processing elements for computing, six for input/output (I/O), and one running the batch system and interactive work. Unlike the Cray systems, the VPP systems had a distributed-memory, parallel architecture. Each of the vector processing elements only had direct access to their own two gigabytes of memory. A built-in, fully non-blocking crossbar switch acted as a high-speed network that allowed the processing elements to communicate with each other. The model ran on 18 processing elements and achieved a sustained performance of around 30 gigaflops, a fivefold increase over the last Cray. The system as a whole had a peak performance equivalent to around 60 modern smartphones.

The distributed memory of the Fujitsu necessitated a rewrite of the forecast model. Cray-specific code features had to be removed and the standard Message Passing Interface (MPI) adopted so that the code would work efficiently on the new system. The rewrite made the code fully portable to different architectures, an important feature retained to this day. Figure 4 shows the structure of ECMWF's computer systems towards the end of the Fujitsu era. The Fujitsu systems continued successfully for six years, increasing the sustained performance 13-fold to 400 gigaflops.

doi:10.21957/szfnyqb5

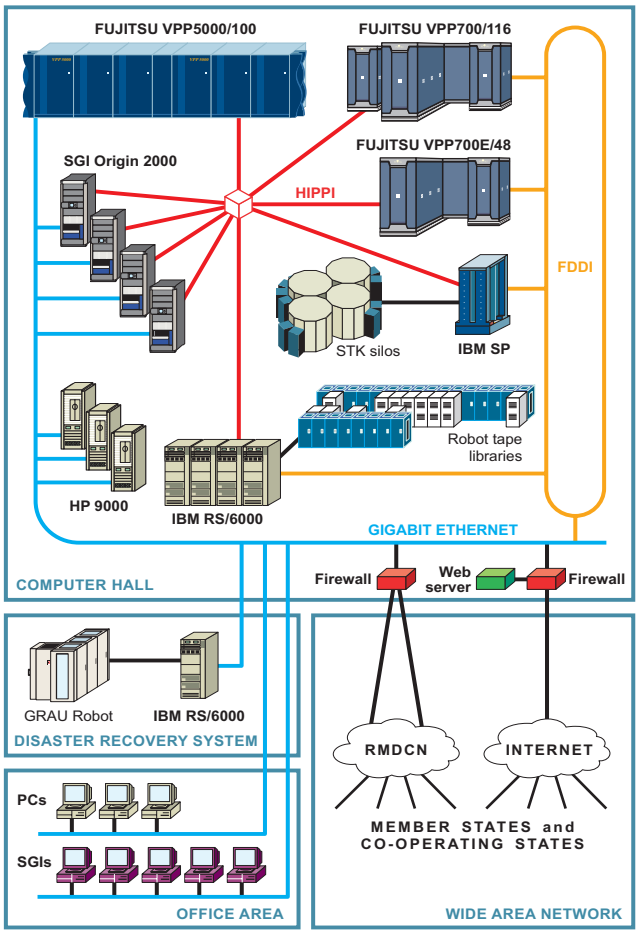

Figure 4 ECMWF computer systems in 2000.

From 2002 to 2014 – the IBM era

In 2002, following a competitive tender exercise started in 2001, new machines from IBM replaced the Fujitsu systems. Two IBM Cluster 1600 systems, consisting of 30 p690 servers connected by an SP2 switch, produced their first operational forecasts on 4 March 2003. These machines differed from the Fujitsu systems in two important ways. First, they had no vector-processing capability, and second, they were high-volume production, standard computers linked by a special high-performance interconnect.

IBM systems provided ECMWF's computing service until the current Cray system replaced them in 2014. They took the sustained performance from the gigaflop range into the terascale, achieving 70 teraflops of sustained performance on the POWER7 system in 2012.

The current Cray HPCF

The current Cray system is the result of a competitive procurement carried out in 2012 and 2013. This resulted in ECMWF awarding a two-phase service contract to Cray UK Ltd to supply and support this HPCF until mid-2018. The contract was signed on 24 June 2013.

The first-phase system started producing operational forecasts on 17 September 2014. The layout comprising two identical Cray XC30 systems continues ECMWF's successful design of having two self-sufficient clusters with their own storage, but with equal access to the high-performance working storage of the other cluster. This cross-connection of storage provides most of the benefits of having one very large system, but dual clusters add significantly to the resilience of the system. They enable flexibility in performing maintenance and upgrades and, when combined with separate resilient power and cooling systems, they provide protection against a wide range of possible failures. Figure 5 shows a high-level diagram of the system, the parts of which are described more fully below.

Each compute cluster weighs almost 45 metric tonnes and provides three times the sustained performance on ECMWF codes of the previous system. Table 1 compares the current system's specification with that of the previous HPCF.

doi:10.21957/szfnyqb5

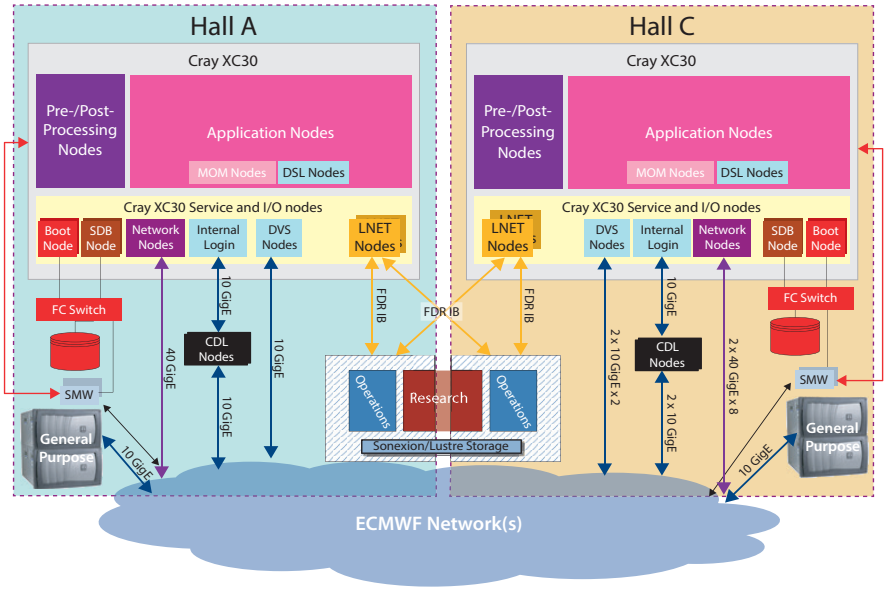

Figure 5 High-level diagram of the Cray HPCF showing major components.

	Previous system	New system
Compute clusters	2	2
Peak performance (teraflops)	1,508	3,593
Sustained performance on ECMWF codes (teraflops)	70	200
EACH COMPUTE CLUSTER		
Compute nodes	768	3,505
Compute cores	23,424	84,120
Operating system	AIX 7.1	Cray CLE 5.2
High-performance interconnect	IBM HFI	Cray Aries
High Performance Parallel Storage (petabytes)	3.14	6.0
General-purpose storage (terabytes)	Not applicable	38
EACH COMPUTE NODE		
Memory in compute node (gibibytes)	64 (20 nodes with 256)	64 (60 x 128, 4 x 256)
Processor type	IBM POWER7	Intel E5-2697 v2 'Ivy Bridge'
CPU chips per node	4	2
Cores per CPU chip	8	12
Clock frequency (gigahertz)	3.8	2.7

Table 1 Comparison of the current system's specification with that of the previous HPCF.

doi:10.21957/szfnyqb5

System description

The bulk of the system consists of compute nodes, which each have two 12-core Intel processors. As shown in Figure 6, up to four compute nodes sit on a blade. Sixteen blades sit in a chassis, and there are three chassis in a cabinet. This gives a maximum of 192 nodes or 4,608 processor cores per cabinet. The number of compute nodes in a cabinet will sometimes be less than the maximum since, as well as compute nodes, each cluster has a number of 'service nodes'. These have space for a PCI-Express card and are twice the size of a compute node so that only two fit on a blade. There are 19 cabinets in each of ECMWF's two clusters.

The Intel Xeon EP E5-2697 v2 'Ivy Bridge' used in the system was released in September 2013. It is an update of the original Xeon E5 'Sandy Bridge' processor following Intel's 'tick-tock' development strategy. In Intel terms, the Sandy Bridge processor was a 'tock', an introduction of a new microarchitecture and new features. Ivy Bridge is a 'tick' as it takes the architecture from the 'tock' and builds it using a new manufacturing technology, in this case a shrink to a 22-nanometre process that gives a greater transistor density on the chip. This allows more to be packed onto a chip while retaining the same overall energy consumption. For comparison, a 22-nanometer transistor is so small that about 4,000 can fit across the average width of a human hair.

The peak performance of one processor core is around 22 gigaflops per second. This is more than the peak performance of the Cray C90/16 system ECMWF had in 1996. There are more than 84,000 such cores in each of the XC30 clusters.

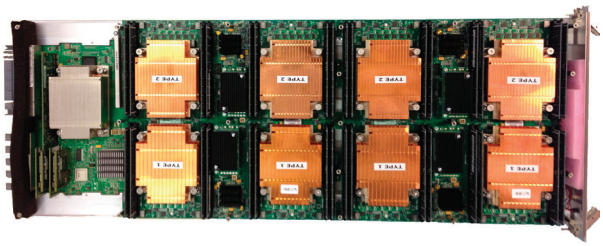

Figure 6 An XC30 compute blade. On the main part of the blade you can see the heat sinks for the eight CPU chips of the four nodes. At the back of the blade is the Aries router.

High-performance interconnect

Connecting all of the processing power together is the Aries™ interconnect developed by Cray. This interconnect uses a 'dragonfly' topology, shown in Figure 7. The name stems from the shape of the dragonfly's body and wings, which represent local electrical connections on the one hand and longer-distance optical connections on the other.

Each blade in the system has a single Aries chip, and all the nodes on the blade connect to it via PCI-Express Gen3 links capable of a transfer rate of 16 gigabytes per second in each direction. Each Aries chip then connects via the chassis backplane to every other blade in the chassis. A chip has five other electrical connections, one to each chassis in a group of two cabinets. Cray describe this as an 'electrical group'. As shown in Figure 8, a further network level uses optical links to connect every electrical group to every other electrical group in the system. Electrical connections are cheaper than optical ones but are limited in length to a few metres. The Aries chip design also removes the need for external routers.

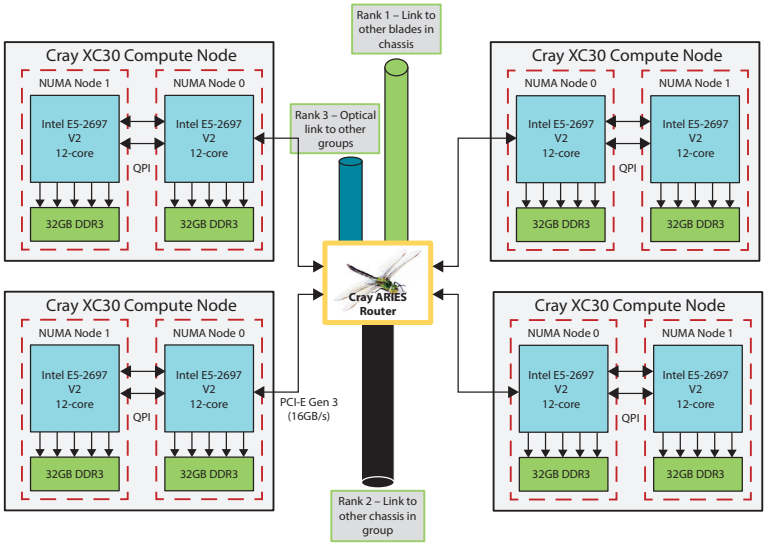

Figure 7 A diagram of an XC30 compute blade. Each blade has four dual-socket nodes and an Aries router chip in a 'dragonfly' arrangement.

doi:10.21957/szfnyqb5

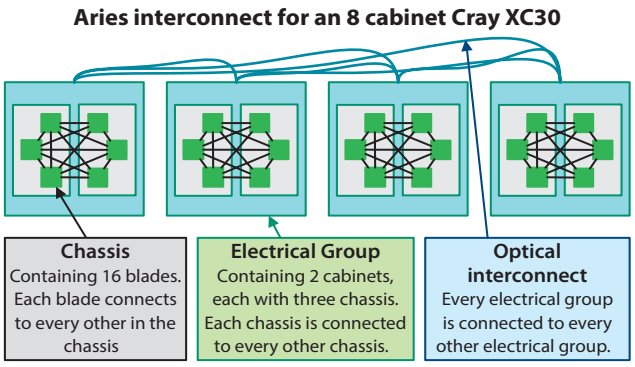

Figure 8 The Cray Aries interconnect has a large number of local electrical connections and a small number of longer-distance optical connections.

Operating system

The nodes of the Cray system are optimised for their particular function. The bulk of the nodes run in 'Extreme Scalability Mode'. In this mode, each node runs a stripped-down version of the Linux operating system. Reducing the number of operating system tasks running on a node to the minimum is a key element of providing a highly scalable environment for applications to run in. Any time spent not running the user's application is a waste. If the application is a tightly coupled parallel one, where results need to be exchanged with processes running on other nodes in order to progress, then delays caused by operating system interruptions on one node can cause other nodes to go idle waiting for input, increasing the runtime of the application.

The other two types of nodes in a Cray system are 'Service' nodes and 'Multiple Applications Multiple User (MAMU)' nodes.

MAMU or Pre-/Post-processing nodes (PPN) for ECMWF run full versions of the Linux operating system and allow more than one batch application to be run on a node. This mode is important: approximately four-fifths of the jobs run on the ECMWF systems require less than one full node to run on. These jobs are the preparation and clean-up for the main parallel jobs. While there are a huge number of these jobs, they account for less than 1% of the processing time offered by the system.

Service nodes are generally not visible to users. They perform a number of functions, such as connecting the compute system to the storage and the ECMWF networks, running the batch scheduler and monitoring and controlling the system as a whole.

Figure 9 A Cray Sonexion storage appliance. The rack contains a metadata server and six storage building blocks.

doi:10.21957/szfnyqb5

High-performance storage

High-performance working storage for the compute clusters is provided by Lustre file systems from integrated Cray Sonexion appliances, shown in Figure 9. Each cluster has two main pools of storage, one for time-critical operational work, the other for research work. Segregating time-critical from research storage helps avoid conflicts between workloads and thus limits the variability of run times for time-critical work. While each cluster has its own high-performance working storage and is self-sufficient, it also has access, at equal performance, to the storage resources of the other cluster. This cross-mounting allows work to be flexibly run on either cluster, in effect making it in many regards a single system. There is a risk that an issue on one storage system can affect both compute clusters but, if necessary, the cross-mounts can be dropped to limit the impact of the instability to just one compute cluster. Each of our XC30 systems has about 6 petabytes of storage and offers up to 250 gigabytes per second of I/O bandwidth.

Lustre file system

The Lustre architecture has been developed in response to the requirement for a scalable file system for large supercomputers.

A Lustre file system, shown schematically in Figure 10, has several components. A metadata server (MDS) supports the directory hierarchy and information about individual files, such as who owns them and who can access them. The MDS stores its data on a metadata target (MDT), a small RAID array connected to the primary and backup server for resilience. Object Storage Servers (OSS) handle the actual data. Each OSS has a number of Object Storage Targets (OSTs) where the data is actually stored. When a Lustre client wants to do something like write a file, it contacts the MDS. This checks the authorisation of the user and that they have permission to access the file location. If successful, it sends back a list of OSTs from which the file can be read or to which it can be written. The client can then deal with the OSSs that host the OSTs. Since the OSTs are independent of each other, if the client has been given more than one, then it is possible for it to use them in parallel for higher performance. How many OSTs are given out for each file is a configurable parameter called ‘stripe count’, which can be set on a file, a directory or the entire file system. When more than one OST is used, data is striped across all of them in chunks controlled by the stripe size parameter.

General-purpose storage

The second type of storage in the HPCF is the general-purpose storage provided by a NetApp Network File System. This storage provides space for home file systems and for storing applications. At 38 terabytes, its capacity is relatively small compared to the Lustre file systems, but the general-purpose storage is very reliable and offers a number of advanced features, such as file system snapshots and replication, which Lustre currently does not implement.

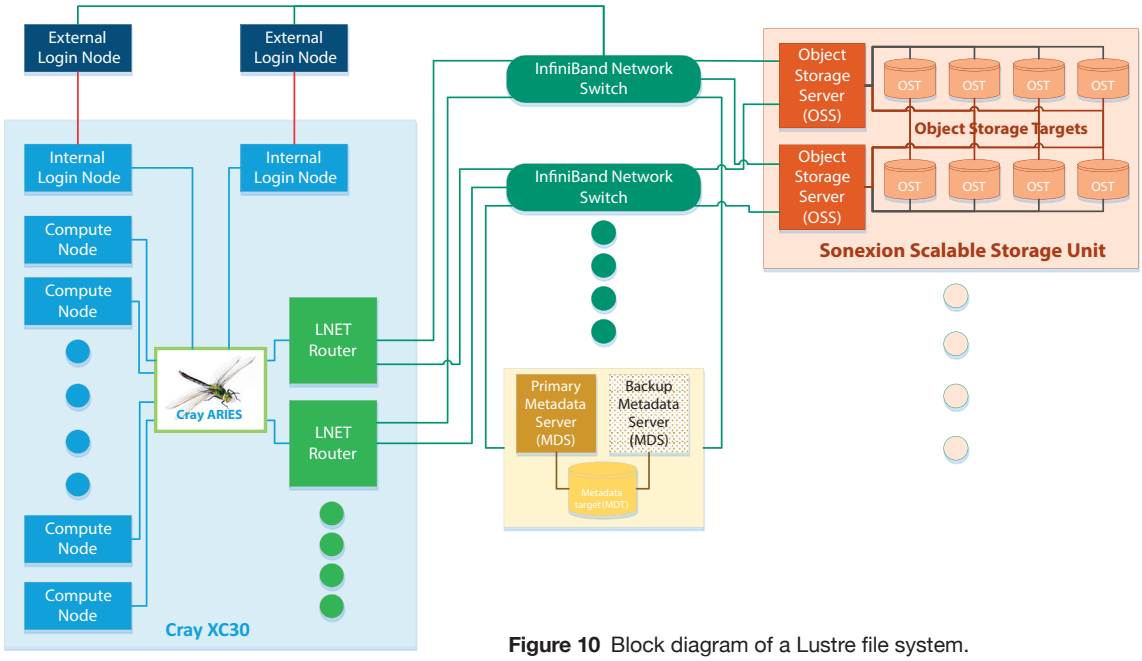

Figure 10 Block diagram of a Lustre file system.

Outlook

Improving ECMWF's forecasts will require further advances in modelling the Earth's physical processes, the use of more observational data and finer model grid resolutions. This means that ECMWF will have to continue to develop its computing capability.

A key requirement for the future will be the scalability of the forecasting system to prepare it for the next generation of HPCFs: exascale facilities performing a billion billion calculations per second. Space is also an issue. ECMWF is looking for new premises to accommodate the kind of supercomputer centre it will need to maintain its leading position in global numerical weather prediction.

 doi:10.21957/szfnyqb5