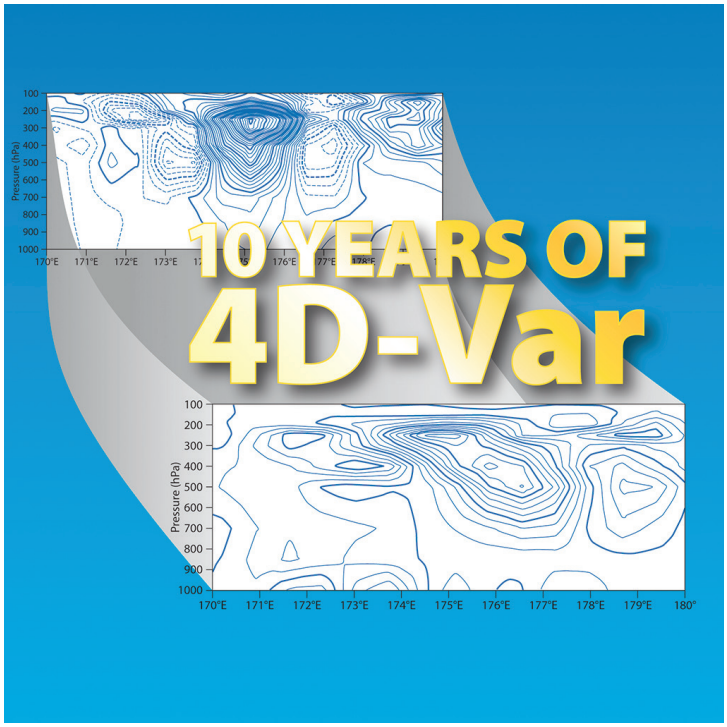
ECMWF Feature article

from Newsletter Number 115 – Spring 2008

COMPUTING

ECMWF's Replacement High Performance Computing Facility 2009–2013



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

doi:10.21957/enu9zmqwfc

This article appeared in the Computing section of ECMWF Newsletter No. 115 – Spring 2008, pp. 44-49.

ECMWF's Replacement High Performance Computing Facility 2009–2013

Neil Storer, Isabella Weger

In 2007, ECMWF carried out a procurement to replace its High Performance Computing Facility (HPCF) from 2009 onwards. Within the framework of a service contract with IBM United Kingdom Ltd., a POWER6 system will be provided for the period 2009–2011, delivering five times the performance over the current system, which will be replaced with a POWER7 system for the period 2011 to mid-2013.

Summary of the procurement process

The ECMWF Strategy 2006 to 2015 defines the principal goal for the coming years as maintaining the current rapid rate of improvement of its global medium-range weather forecasting products, with particular emphasis on improving early warnings of severe weather. The Strategy includes specific targets for improvements in forecast skill and highlights the fact that the rate of increase in the performance of the HPCF is an essential factor for achieving these targets. It calls for a sustained performance of 20 teraflops from early 2009, with a gradual increase to between 150 and 200 teraflops sustained by 2015.

The Strategy requirements equate to an improvement over the existing HPCF system by a factor of 5 from 2009 (Phase 1 of the new system) and a factor of 10 to 12.5 from 2011 (Phase 2 of the new system).

Since the existing service contract for the current HPCF expires in early 2009, in December 2006 Council approved the proposal to procure a replacement system. Recognizing that the performance of the HPCF is crucial to the successful implementation of the strategy, Council accordingly approved an increased money stream, with the proviso that all ancillary costs (infrastructure enhancements, data archive system upgrades and increased electricity consumption) arising from this increase in HPC funding are to be covered out of this money stream. The Invitation to Tender (ITT) therefore provided vendors with a way of calculating the ancillary costs associated with their particular offer and required them to make a corresponding deduction from the funds available for the new contract.

In March 2007 ECMWF published an ITT (reference: ECMWF/2007/192) for the procurement of a High Performance Computing Facility, with a closing date of 1 June. A number of working groups were established, each charged with performing an in-depth analysis and comparison of specific aspects of the tenders received. The process was carried out over the summer months and included visits to reference sites and a dialogue with tenderers to refine the solutions bid.

In December 2007, Council authorized ECMWF's Director to sign a contract with the successful tenderer, IBM UK Ltd. The new HPCF will attain a performance improvement factor of 5 for Phase 1 and about 10.5 for Phase 2, based on a benchmark comprising the following set of applications that are representative of the most computationally demanding components of ECMWF's planned operational activities:

- 4D-Var at resolution T799L91 and T95/159/255.
- Deterministic forecast at resolution T1279L126.
- EPS at resolution T639L91.

doi:10.21957/enu9zmqwfc

The new HPCF

The new HPCF will be supplied in two phases; Phase 1 will cover the period 2008 to 2011, while Phase 2 will cover the period 2011 to 2013. Each Phase will be delivered and installed in two stages (Stage 1A and Stage 1B for Phase 1, Stage 2A and Stage 2B for Phase 2). The system delivered for each phase will comprise two identical independent subsystems. Phase 1 for example will comprise two identical POWER6 compute clusters along with two identical I/O storage clusters, so subsystem-1A will comprise one of the compute clusters and one of the I/O storage clusters, while subsystem-1B will comprise the other compute cluster and I/O storage cluster.

Table 1 outlines the main characteristics of one of the two identical subsystems of Phase 1.

Stage 1A will start its formal acceptance tests towards the end of 2008, then, once Stage 1B has been installed, the whole of Phase 1 will start its formal acceptance tests a few months later. Phase 2 will start its formal acceptance tests in 2011.

The staged delivery approach has two main advantages.

- It puts less strain on ECMWF's power and cooling infrastructure, as it will be possible to ensure that during periods of overlap, when systems are being commissioned alongside production systems, only a maximum of three subsystems (two old and one new, or two new and one old), rather than four, need be powered and cooled.
- It will enable one subsystem to reside for a while at IBM's manufacturing and testing plant in Poughkeepsie, USA at the same time as the other subsystem resides at ECMWF. This will enable the two subsystems to be tested concurrently. ECMWF will run a workload that tries to emulate what normally runs on the existing HPCF. At the same time engineers in Poughkeepsie can concentrate more on benchmark and diagnostic testing, as well as having an identical system on which to solve any problems experienced in testing at ECMWF.

Figure 1 is a schematic of the Phase 1 system, showing the two subsystems and the various clusters and components. The only physical difference between 'application' nodes and 'pre/post-processing' nodes is the amount of memory they have. Although the I/O storage clusters are shown as belonging to one subsystem or the other, the multi-cluster GPFS filesystem enables both compute clusters to access data on either I/O cluster as if the data was resident on that compute cluster. Potentially all of the data is accessible to any node in either subsystem and the data will be transferred at a rate that is independent of which of the two I/O clusters serves the data. Also, the I/O storage clusters are independent of the compute clusters, so it will be possible for example to shutdown a compute cluster without affecting either of the I/O clusters or the other compute cluster.

Figure 2 illustrates the independence of the I/O storage clusters, as well as their constituent parts.

	Subsystem attribute	Description
COMPUTE	Compute nodes	240 32-core 4.7GHz POWER6 'application' nodes
		8 32-core 4.7GHz POWER6 'pre/post-processing' nodes
		2 32-core 4.7GHz POWER6 'spare' nodes
	Memory	2 GB/core on the 'application' and 'spare' nodes
		8 GB/core on the 'pre/post-processing' nodes
	Network nodes	9 32-core 4.7GHz POWER6 (mainly for I/O routing)
	Interconnect	8-plane 4×DDR Infiniband
I/O	I/O nodes	28 (14 pairs of) 4-core 4.2 GHz POWER6
	Disk space	0.9 Petabytes (900,000 GB, or 900 TB)

Table 1 The main characteristics of one of the two identical subsystems of Phase 1.

doi:10.21957/enu9zmqwfc

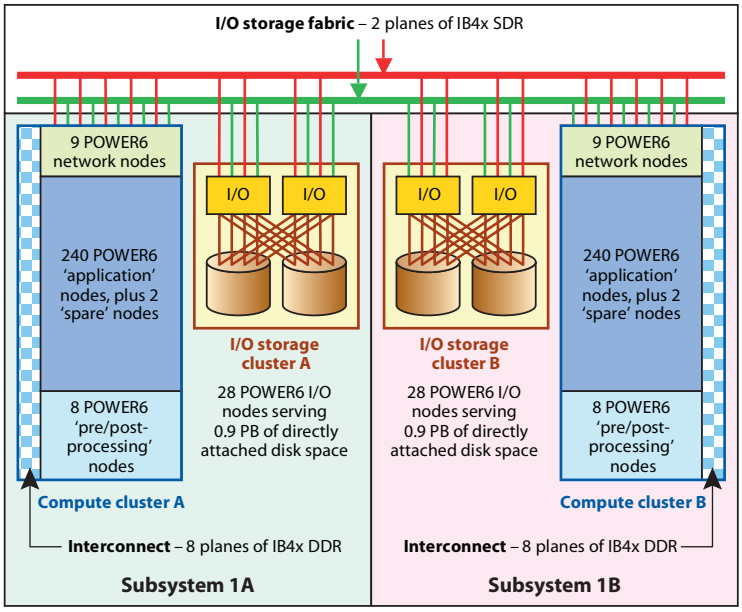


Figure 1 An overview of the Phase 1 system.

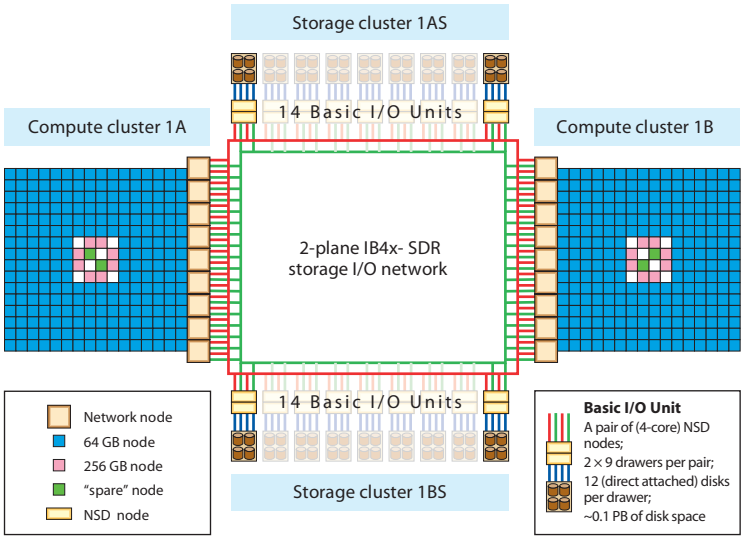


Figure 2 The Phase 1 system in more detail.

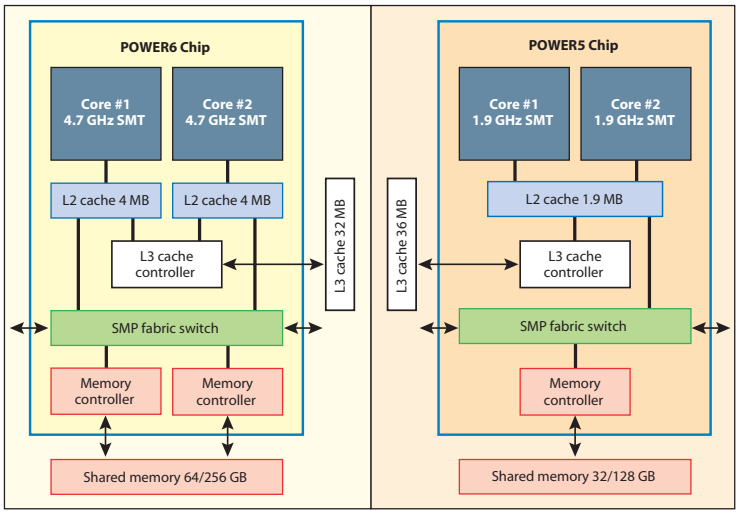


Figure 3 The architecture of the IBM POWER6 and POWER5 at ECMWF.

doi:10.21957/enu9zmqwfc

Comparing the current and new HPCFs

The main component of the system is the POWER6 node, which is an evolution of the p5-575 POWER5 node that is used in the existing HPCF. In the compute clusters the POWER6 node has 32 cores, running at 4.7 GHz, while the nodes in the I/O clusters each have 4 cores with a frequency of 4.2 GHz. Figure 3 is a simplified schematic which shows the differences in architecture of the POWER6 and POWER5 chips.

While the POWER6 processor is binary compatible with the POWER5 processor that is used in ECMWF's current HPCF, it has several additional and enhanced architectural features that are outlined in Table 2, which compares the two architectures.

There are other differences between ECMWF's new HPCF and the existing one besides the generation of the processor. The POWER5 nodes in the clusters that comprise ECMWF's current HPCF are interconnected using the pSeries High Performance Switch (also known as the Federation Switch). This is a proprietary switch designed and manufactured by IBM solely for clusters made out of the pSeries POWER range of shared memory processor (SMP) systems. The compute clusters in the new HPCF will use a non-proprietary switch for their interconnect, based on eight planes of 4x Infiniband running at double data rate (i.e. IB4x DDR). "4x" means that four lanes are aggregated into a single link, providing a theoretical peak transfer rate of 2 GB/s bi-directional per IB4x DDR adapter. So an 8-plane IB4x DDR has a theoretical peak transfer rate of 16 GB/s per connection in each direction. This compares to the 2-plane, 4 GB/s Federation switch on the existing HPCF.

The current HPCF has a single GPFS "controlling" cluster, which along with VSD (virtual shared disk) I/O nodes in the compute clusters is connected via a fibrechannel storage area network to RAID storage subsystems having multiple disks and controllers. However, this "controlling" cluster does not serve data to the compute clusters; it solely controls multi-cluster GPFS metadata and tokens. The new HPCF has two separate I/O clusters and only they are directly connected to disks, via RAID adapters. When a compute node in the new HPCF needs to read/write data, the actual disk I/O is performed by the NSD (network shared disk) nodes in the I/O cluster. These nodes route the data over a dedicated dual-plane IB4x SDR network to/from network nodes in the compute cluster, which in turn route the data to/from the compute node requesting the I/O over the cluster's internal Infiniband interconnect.

The use of storage clusters, rather than a fibrechannel storage area network, enables greater flexibility in deployment of the most appropriate technologies; also the simplicity of this scheme should improve reliability. It also means that since neither compute cluster "owns" data it should be relatively simple to perform maintenance on the compute clusters without affecting the I/O capabilities of the other one. Certain datasets will be replicated on the two I/O storage clusters to ensure that the operational suite of jobs can run even if, for whatever reason, one of these clusters becomes unavailable.

	POWER6	POWER5
SMP packaging	Dual chip module (DCM)	Dual chip module (DCM)
Transistors per chip	790 million	276 million
Execution style	Mostly in-order, with special case out-of-order execution	General out-of-order execution
Symmetric multi-threading	2 threads/core, priority-based dispatch, simultaneous dispatch from 2 threads (up to 7 instructions)	2 threads/core, priority-based dispatch, alternate dispatch from 2 threads (up to 5 instructions)
Frequency	4.7 GHz (18.8 GFLOPS/core peak)	1.9 GHz (7.6 GFLOPS/core peak)
Functional Units	2 floating-point multiply-add 2 fixed-point (integer) 2 load/store 1 branch 1 decimal-point 1 SIMD (VMX)	2 floating-point multiply-add 2 fixed-point (integer) 2 load/store 1 branch
L1 cache	On-core: 64 KB 4-way instruction 64 KB 8-way data	On-core: 64 KB 2-way instruction 32 KB 4-way data
L2 cache	On-chip, private to a core: 2 × 4 MB 8-way LRU, 150 GB/s	On-chip, shared by the cores: 1.9 MB 10-way LRU, 30 GB/s
L3 cache	Off-chip, but on-DCM, shared by the 2 cores (unused cache can be "borrowed" by other chips): 32 MB 16-way LRU, 32 GB/s	Off-chip, but on-DCM, shared by the 2 cores: 36 MB 12-way LRU, 12 GB/s

Table 2 Comparison of the POWER6 and POWER5 architectures.

Another difference between the new and the existing HPCF is the operating system. The current HPCF runs the AIX 5.3 operating system, but the new one is likely to run AIX 6.1. While in general, users of the system should see very little difference regarding the software, AIX 6.1 has features that improve the security and availability of the system by enabling concurrent kernel updates to be done as well as tracing, debugging and recovery. It also includes improved support for memory management.

The technology of the new p6-575 servers is denser than the current technology. Although the new HPCF has five times the sustained performance compared to the current one, it will still only take up about the same amount of floor space. However, power and cooling requirements will increase significantly. ECMWF will upgrade its infrastructure in 2008 to accommodate an additional uninterruptible power supply machine and two additional chillers. These are needed to power and cool the new system.

New cooling technologies are used within the servers to remove the heat from the denser systems. The processor chips in the new p6-575 servers will be water-cooled, which is a departure for IBM in that up until now their pSeries systems have been air-cooled. The memory DIMMs will continue to be air-cooled, but the heated air will pass through a water jacket attached to the rear door of the servers. These two water cooling features together will enable more than 70% of the heat to be extracted directly by water, with very little remaining to be dissipated into the machine room. The flow of the water is regulated through copper pipes to the heat sink on the chip to match the heat output of the processors. Figure 4 shows how dense the packaging within a frame is, and how the water is fed into and out of each of the individual p6-575 servers, of which there will be twelve per frame.

Table 3 compares the main features of the two systems.

	New Phase 1 HPCF	Current HPCF
No. of clusters	2 compute clusters 2 I/O storage clusters 1 small test cluster	2 compute clusters 1 MC-GPFS "controlling" cluster 1 small test cluster
Performance	~20 TFLOPS (sustained)	~4 TFLOPS (sustained)
Energy requirement	~2.5 MW	~1.0 MW
Each compute cluster		
Operating System	AIX 6.1 (probably)	AIX 5.3
Compute nodes	248 × 32-core POWER6 (SMT)	155 × 16-core POWER5 (SMT)
Compute processors	~8000	~2500
Network nodes	9 × 32-core POWER6 (connected to the LAN and the I/O storage fabric)	2 × 16-core POWER5 (connected to the LAN)
I/O nodes	None	8 × 16-core POWER5 (VSD nodes connected to the fibrechannel SAN)
Interconnect	8-plane IB4x-DDR (16 GB/s)	dual-plane pSeries HPS (4 GB/s)
I/O		
Paradigm	Independent I/O storage clusters, each with 28 × 4-core POWER6 NSD nodes transferring data over a dual-plane IB4x SDR network	Fibrechannel SAN
Disk types	Directly attached RAID6 storage	DS4500 RAID5 storage systems
Disk space	1.8 Petabytes (total HPCF) ~6000 disks	100 Terabytes (total HPCF) ~3500 disks
Each compute server (node)		
Memory	64 Gigabytes (8 with 256 GB)	32 Gigabytes (4 with 128 GB)
Dual-core chips	16	8
Processors (cores)	32	16
Each processor (core)		
Lithography	65 nm	90 nm
No. of transistors	790 million	276 million
Clock frequency	4.7 GHz	1.9 GHz
Peak performance	18.8 GFLOPS (~290 TF total HPCF)	7.6 GFLOPS (~37 TF total HPCF)

Table 3 Comparison of the new Phase 1 HPCF and the current facility.

doi:10.21957/enu9zmqwfc

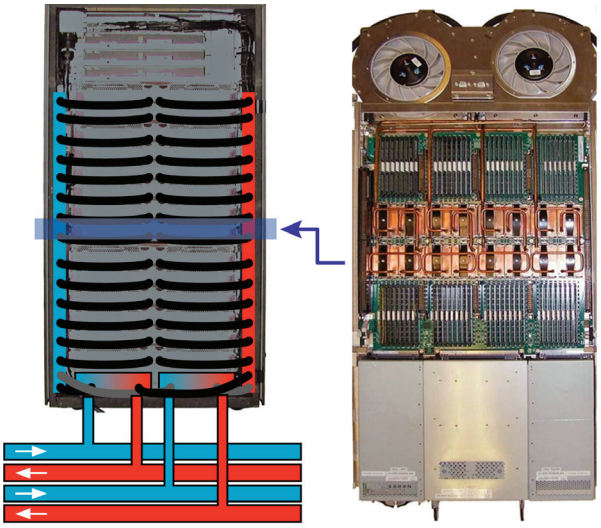


Figure 4 The cooling system based on new technologies.

Phase 2

In mid-2011, the POWER6 based Phase 1 system will be replaced by one based on IBM's follow-on processor generation, POWER7. This will be similar to a "petaflop" system being delivered under a contract awarded to IBM by the US National Science Foundation. The Phase 2 system macro-architecture will be identical to that of Phase 1, again comprising of two identical compute clusters. An article on the Phase 2 system will be published in the ECMWF Newsletter closer to the time of its installation.

The Phase 1 system is crucial for ECMWF to meet the strategic target of increasing the resolution of all its forecasting systems in 2010. The Phase 2 system will support further improvement in the operational forecasting system in 2011 and beyond. It will provide the compute resources to support the research and development needed to prepare for the next resolution enhancement in 2015. It will also enable continued efforts to make optimal use of the wealth of information supplied by the observing system, especially for the provision of early warnings of severe weather events.

Further Reading

IBM POWER6 Microprocessor Technology.
IBM Journal of Research Development
http://researchweb.watson.ibm.com/journal/rd51-6.html.

doi:10.21957/enu9zmqwfc