Porting the UK Met Office's Unified Model to the Cray X1

Paul Burton, *Centre for Global Atmospheric Modelling,*
University of Reading (UK) and **Bob Carruthers**, *Cray UK.*

ABSTRACT: *The UK Met Office's Unified Model (UM) is one of the world's leading weather forecasting and climate prediction models. It is used by the Met Office's Hadley Centre and a large group of academic researchers both within and outside the UK for prediction and research into climate change. In this paper we will present a port of the latest generation of the UM to the Cray X1 and compare and contrast the performance with a number of other platforms available to the UK academic community.*

1. Introduction

The Centre for Global Atmospheric Modelling, based at the University of Reading, UK is a key element of the UK's Natural Environment Research Council's (NERC) research on climate change issues. It acts as a centre of expertise for climate science within the UK academic community, and takes the lead in many major research strands, encompassing a broad range of areas, investigating time scales from seasonal to multi-century. CGAM has close links with the Met Office's Hadley Centre – sharing the same numerical modelling system, the Unified Model (UM), and collaborating on many research projects.

As well as providing a lead in scientific research, CGAM is also responsible for the computational science issues involved in a major programme of climate change research using one of the world's leading climate models. CGAM provides support and training for the many users of the UM within the UK academic community, and ports and supports the UM on the various HPC platforms that are available to the community.

A limited amount of optimisation is carried out to enable scientists to take best advantage of the platforms available to them, but the major consideration to most of the scientists we support is the "correctness" of port (the ability to reproduce a climate on different platforms) rather than obtaining the last drop of performance. CGAM do not own the UM code, and rely on the Met Office to supply them with regular releases of new versions of the UM, which tend to be optimised for whichever platform they are currently running the model on.

2. UK HiGEM

CGAM, together with a number of the other NERC academic climate research groups have recently entered into a national, high-profile collaborative "Grand Challenge" programme with the Met Office's Hadley Centre, entitled "UK-HiGEM" (**Hi**gh Resolution **G**lobal **E**nvironmental **M**odel). The aim of the project is to develop the UM to a new high resolution climate configuration, with a 1º atmosphere model and a ⅓º ocean model.

The expectation is that this new configuration will allow a better understanding and better predictability of extreme events, predictability, some of the less well understood climate feedback processes and climate "surprises" (such as rapid changes in crucial phenomena such as the gulf stream). The higher resolution will also allow a much better understanding of the regional impacts of climate change.

3. Computer Systems

CGAM has access to a number of HPC systems which are used for running its climate research workload. Most of its CPU time comes from the systems provided for the UK scientific research community by EPSRC (Engineering and Physical Sciences Research Council), but it also has access to some other HPC systems for specific project work.

CSAR (University of Manchester, UK)

CSAR host a number of SGI machines, which CGAM mostly uses for lower resolution climate production work, which are typically run in a throughput mode – with many different jobs being run by a number of users. CSAR's Cray T3E service was retired last year, and the service is now provided by two machines:

- SGI Origin 3800 with 512 CPUs, and a total of 512 Gb memory (0.4 Tf peak)
- SGI Altix Itanium2 with 256 CPUs, and a total of 384 Gb memory (1.3 Tf peak)

HPCx (Daresbury, UK)

HPCx host a single IBM p690+ based machine, which provides a capability service to the academic community. CGAM use HPCx for much of the HiGEM development and production work, as well as its investigations into ensemble climate model systems. The machine is currently being upgraded to the latest generation of IBM technology, and in this paper we present performance results from both the old and new configurations:

- Phase 1 : IBM p690 / POWER4 with 1280 CPUs, organised as 8 way LPARs, and a total of 1.28Tb memory (6.6 Tf peak)
- Phase 2 : IBM p690+ / POWER4 / Federation switch with 1600 CPUs, organised as 32 way LPARs, and a total of 1.6Tb memory (10.8 Tf peak)

Earth Simulator (JAMSTEC, Japan)

CGAM and the Met Office have negotiated time on the Earth Simulator to be used for the HiGEM project production runs. Architecturally, the machine is similar to the NEC SX6, but has less memory per node and a modified version of the interconnect (to incorporate the large number of nodes). The machine can be rather complicated to use, due to the configuration of the disk systems, and the fact that it is not connected to the outside world – necessitating visits to Japan, and CGAM/Met Office staff permanently on site.

- 5120 SX6 CPUs, organised as 8 way nodes, and a total of 10Tb memory (40 Tf peak)

Met Office (Exeter, UK)

The Met Office have recently taken delivery of an NEC SX6 system, which replaces their Cray T3E systems. The Met Office will be using the SX6 for their forthcoming IPCC production runs, and for their contributions to the HiGEM project.

- 240 CPUs, organised as 8 way nodes, and a total of 1Tb memory (1.9Tf peak)

Cray

Cray have offered CGAM time on X1 system, to test a port of the UM and investigate its performance.

- sn702 : Cray X1 with 3x(16 SSP / 4 MSP) 800MHz production nodes

4. The Unified Model

The UM was initially developed by the UK Met Office in the early 1990's and was designed to run efficiently on the Office's Cray YMP. Since then it has undergone continuous development and expansion, including conversion to message passing parallelism (for the Cray T3E), limited scalar optimisation and most recently a completely new dynamical core. Currently is consists of around a million lines of portable Fortran77/90 and C and many thousands of lines of supporting scripts.

Throughout all the developments to the UM, the core aim of having a unified code base and infrastructure for both operational NWP (global and limited area modelling) and climate prediction (including a coupled ocean model) has been maintained. More recently, functionality has been added to allow external models to coupled using the OASIS coupler, and over the next few years the Met Office plan to implement the PRISM coupler within the UM system.

Having a single unified model for a wide range of applications results in a very complex and highly configurable model system – for example there are typically many different versions of each of the physical parameterisation schemes, and each one of these has a number of tuneable parameters which may need to be tweaked for specific applications. This complexity is largely hidden from the user by a hierarchical graphical user interface, which presents a simplified representation of the model configuration to the user, and then generates the appropriate Fortran namelists and configuration scripts.

Parallelisation

The UM was parallelised using message passing in the mid-90's, with this parallelisation further optimised for the Cray T3E that was procured. A portable interface library is used which can be configured to use MPI or Cray SHMEM as the underlying communications library. This allowed full portability to all distributed memory systems, whilst still achieving high performance from the Cray T3E's fast interconnect.

Both the atmosphere and ocean models are parallelised using a regular domain decomposition, with the atmosphere employing a two dimensional decomposition (in the two horizontal dimensions) and the ocean model a one dimensional decomposition (in the North-South direction).

The communication pattern of the UM is characterised by many short messages and regular barrier synchronisations. This is partly a reflection of the algorithms used by the UM, but is also a recognition that such a communications pattern was handled very efficiently by the Cray T3E. Today's machines struggle to maintain the balance between latency, bandwidth and processor speed that the T3E achieved, and the UM is highly sensitive to the communications performance on the platforms it is run on.

Atmosphere Model

The UM's atmosphere model employs a regular latitude/longitude grid (with the side effect the grid points converge towards the poles), with user-definable vertical level structure. The main prognostic model variables are winds, temperatures, moisture and pressure.

As was stated earlier, the dynamical core of the UM was recently upgraded. One of the consequences of this is that the dynamics of the UM now employs a 2D array structure in the horizontal dimension, compared to the previous 1D array structure, which meant that the inner loop is now relatively short, and sensitive to decomposition in the East-West direction.

Fields are advected using a semi-lagrangian advection scheme. Much of the computational effort in this scheme is taken in calculating the value of fields at the "departure point" (a precise location where the field at the current grid point originated from at the previous timestep) using a 3rd or 5th order interpolation scheme. This is quite a difficult algorithm to vectorise well, as each departure point must be calculated separately, and the number of points required to calculate the value is relatively small. A small amount of communication is required where departure points lie outside of the halos on a processor – this is a more significant effect on processors near the pole (due to the convergence of grid points), and when the model has been decomposed in the East-West direction.

The dynamical forces are balanced using a semi implicit 3D Helmholtz solver. The algorithm used can be vectorised relatively well and is moderately sensitive to the communications network, particularly in the climate configuration, where each iteration of the solver requires global summations to be carried out over all processors. The NWP configuration of the model removes some of the global communications for greater efficiency, at the cost of a loss of reproducibility if the processor configuration is changed.

The physical parameterisations in the UM generally work with a 1D array structure in the horizontal (as their code structure reflects their heritage in Cray YMP code). The biggest problem with many of the parameterisations is of load balance – as they represent processes which are by nature different at different grid boxes. Three of the most computationally expensive parameterisations are discussed later in the paper:

- *Radiation* – short wave radiation deals with the incoming solar radiation, which at any one time only covers half of the earth's surface
- *Boundary Layer* – dealing with the interaction of the atmosphere with the earth's surface. The computations required for land surfaces and ice are much more complex than the sea surface, so there are load balance issues where some processors have more land points than others.
- *Convection* – parameterises the sub grid scale convective processes which occur as warm moist air rises rapidly into the atmosphere – these processes are much more active in equatorial regions, and around active weather systems at mid-latitudes.

Ocean Model

The UM's ocean model is contained within the same executable as the atmosphere model, and uses the same processors as the atmosphere model uses. The atmosphere model runs for 24 hours and passes coupling fields to the ocean model which then runs for the same 24 hour period before passing coupling fields back to the atmosphere model. At present there is no facility to run the two models as separate executables, and they both must use all the processors allocated to the model. If the models are not equally scalable, this means there can be an efficiency trade off as one cannot necessarily run both models with the optimum number/configuration of processors.

The ocean model uses a 4th order advection scheme, and an iterative conjugate gradient solver. This solver turns out to be very important in the model's performance, as it carries out a number of latency sensitive operations at each timestep. Filtering is carried out around each pole to remove numerical noise generated due to the convergence of grid points.

5. Climate Configuration Results

Configuration

The configuration used here is the latest version of the Hadley Centre's HadGEM configuration, featuring an N48 (270km resolution) atmosphere model and a 1º ocean model. This configuration has been developed on the Cray T3E, and so does not contain any specific SX6 porting optimisations. The configuration is only just reaching finalisation, and was only received from the Met Office shortly before this paper was written, so reducing the amount of detailed investigation that has been possible.

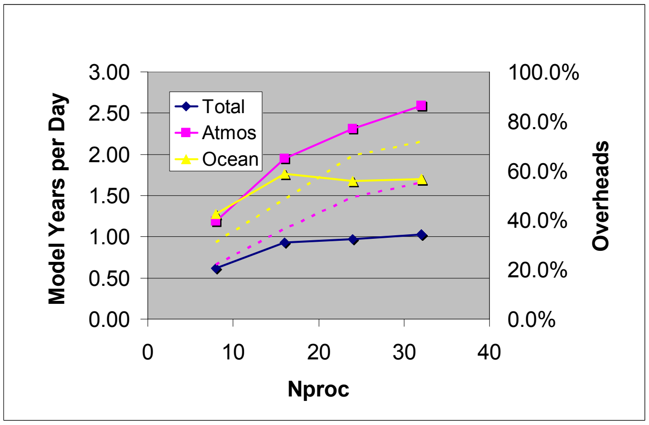
We have successfully ported and run this configuration of the UM on the IBM p690 "HPCx" systems, the Met Office SX6, and the Cray X1. Unfortunately, we were unable to run this configuration on the SGI Newton system. The Fortran compiler (developed by Intel) on this platform was very fussy, and refused to compile many routines which compiled successfully on all the other platforms we tested. Even after the code was modified to allow compilation, the runs failed – there was not sufficient time to investigate whether the problem was with the code (quite possible since there are a number of components which were still under active development in the version tested) or the compiler.

The ports to the IBM HPCx systems, Met Office SX6 and Cray X1 were all relatively painless. Due to time constraints, very little platform specific optimisation has been carried out on any of these ports. The following levels of compiler optimisation were applied:

- IBM : `-O3 -qstrict` (a fairly conservative but safe level of optimisations which gives bit identical results compared to –O3)
- SX6 : `-Cvopt` (a safe level of vector and scalar optimisations)
- X1 : `-Oaggress,scalar3,vector2,stream0, nopattern,ssp ,task0`

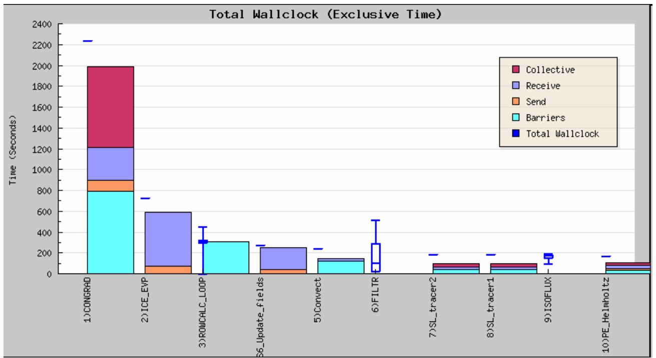
It should be noted that the X1 was used in SSP mode rather than MSP mode. Complexity in many areas of the UM code mean that the Cray compiler was unable to successfully automatically generate MSP code. Introducing the necessary compiler directives to allow a successful MSP compilation would require similar effort to an OpenMP port. At the time of writing, there has not been time to make the necessary additions of compiler directives to allow a successful MSP compilation

HPCx IBM p690 (Phase 1)



The solid lines show the absolute performance (expressed as simulated model years per wallclock day), while the dotted lines show the "overheads" expressed as a percentage of the total time spent in the area of code concerned. The "overheads" are the time spent in communications routines (including barriers). It can be seen that the overall model does not scale much beyond 1 model day/year at around 16 CPUs, which is a rather disappointing result. Looking at the individual contributions from the atmosphere and ocean models it can be seen that the lack of scalability can be attributed to the ocean model, which slows down after more than 16 CPUs.

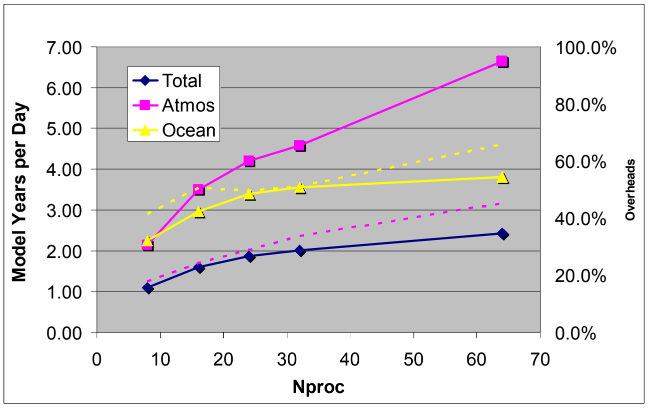
In order to understand what is causing the ocean model to scale so badly we looked at a more detailed timing analysis of the UM run. This bar chart shows the the top 10 routines in the 32 CPU run on the IBM p690 (Phase 1), with the bars indicating the contribution from the various message passing overheads.



The "number 1" routine, which obviously dominates the time of the model run is the ocean model's conjugate gradient solver, and it can be seen that most of the time is being spent in barriers and collective (global) communications. The conjugate gradient solver appears to be requiring O(1000) iterations on every timestep of the ocean model before necessary convergence is achieved. Every timestep requires a global reduction operation. The barrier cost is incurred because many more iterations are required for rows near the poles, and processors away from the poles are just waiting for the polar processors to complete (hence the barrier).

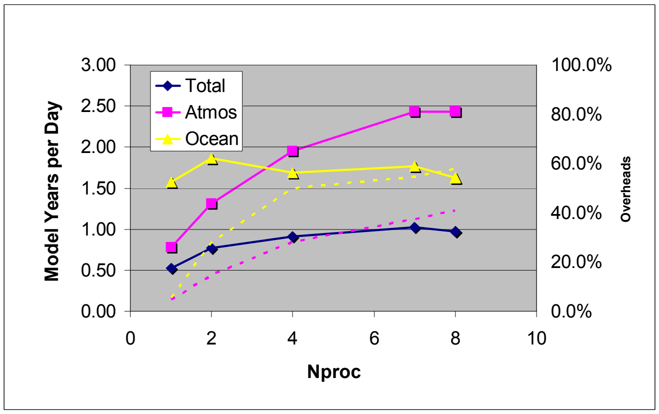
The ocean conjugate gradient solver is obviously a problem, and the Met Office are planning to address this problem – as the current algorithm will be completely unusable at higher resolutions. There are a number of potential solutions, such as using a suitable preconditioner to reduce the number of iterations required for convergence, or using a completely different kind of algorithm to solve the equations.

HPCx IBM p690+ (Phase 2)



These initial results from the upgraded machine are very encouraging, and show almost a two times speedup, much of which is obtained from the reduction in communication overheads. Up to 32 CPUs everything is in-node, and even up to 64 CPUs where the switch is used, the atmosphere model scales reasonably well. Even the ocean model, which was slowing down on the Phase 1 machine shows a small speedup on 64 CPUs, but it is still badly hindered by the conjugate gradient solver.

Met Office SX6



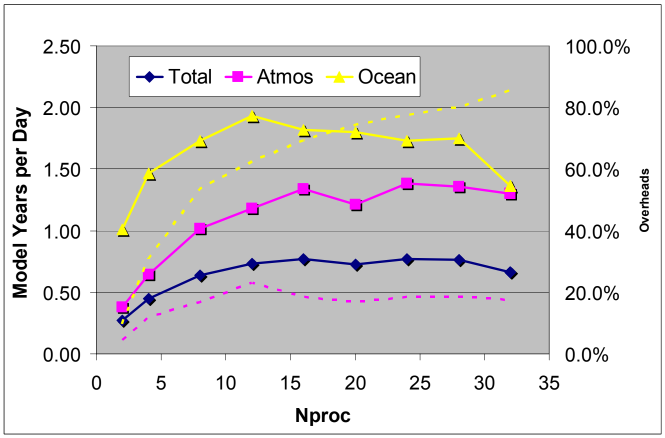
These results demonstrate the difficulty in running relatively low resolution (ie. Small data size and short vector length) models on powerful vector machines such as the SX6. Even though this model is only run on a single node, so no use is made of the switch, it can be seen that the overheads still account for a sizable proportion of the total run time (over 50% for the ocean model on 8 CPUs). This is a reflection of the relatively poor scalar performance of the SX6.

It can be seen from the graph above that the results on 8 CPUs are worse or at the least no better than the results on 7 CPUs. This is a persistent problem on the SX6, and the Met Office generally use no more than 7 CPUs per node for production work, to allow one CPU to be kept free for system related tasks which would otherwise interfere with the run.

Using these relatively small numbers of CPUs it was possible to use a "1xn" decomposition for the atmosphere model, ensuring that the inner loop has the maximum possible vector length (typically 96). The graph shows that the atmosphere model scales up to 7 CPUs, but once again, the ocean model has real problems. Closer investigation of timing results shows that as well as poor performance from the conjugate gradient solver, there is a serious load balance problem – with the polar processors being much more expensive. There are a number of reasons for this, the most dominant being effects due to grid point convergence around the poles (which causes greater number of iterations of solvers, and the requirement for filtering to remove spurious noise) and the use of ice models around the poles. There is also a larger scale load imbalance created by the uneven distribution of ocean over the planet's surface – with the southern hemisphere having more ocean points to calculate than the northern hemisphere.

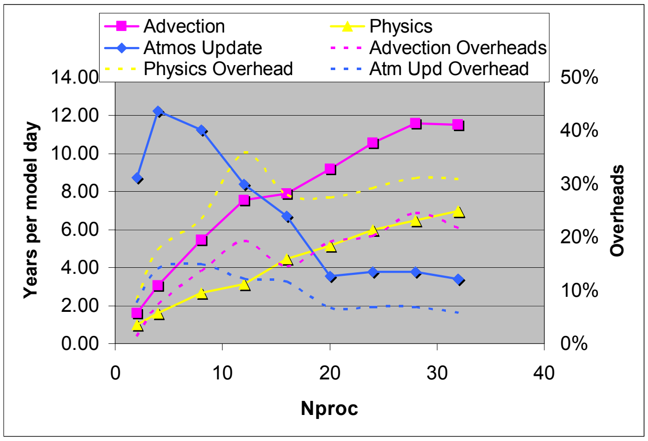
It should be remembered that these results were obtained with a very basic port from the Cray T3E code. The Met Office have also been working on an SX6 optimised version of this configuration – improving the basic level of compiler optimisation where it was safe to do so, rearranging some code to improve vectorisation and inlining certain routines to reduce the scalar overhead of calling subroutines. The Met Office informs us that on 7 CPUs they can now achieve almost twice the performance that we demonstrate here with the "straight" port.

Cray X1



As we have used SSPs rather than MSPs for these runs, a relatively higher number of CPUs (compared to the SX6) is required to achieve the desired level of performance. This is particularly bad for the ocean model which we have already shown to have serious scalability problems – the load balance issues discussed on the SX6 results are even worse here, as the polar processors are a tiny proportion of the total CPUs on a 32 SSP run. Most of the "overhead" shown in the graph is due to the load imbalance than any actual message passing!

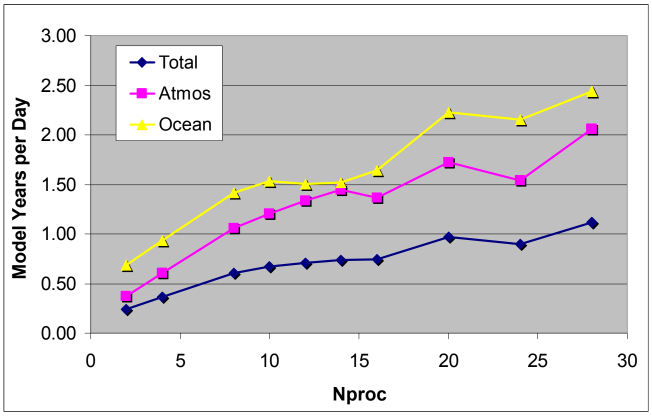
The atmosphere scalability results look rather disappointing, with no scalability beyond 16 SSPs (1 node). It is instructive to break the atmosphere model into a number of its individual dominant components to better understand where the loss of scalability is coming from.



Here we can see that both the main dynamical routine (the advection), and the physical parameterisations both scale reasonable well up to 32 SSPs. However, something called the "Atmos Update" is performing extremely badly, and dominating costs by 32 SSPs. This piece of code is not performing any real science – it just does a whole lot of "tidying up" at the end of an atmosphere time step, such as adding on the field increments calculated by all the preceding parameterisation schemes, and updating the halos of fields. From the "overheads" shown on the graph, this cost does not seem to be due to message passing overheads, so we must conclude the poor performance is due to single processor effects. Initial indications are that this is due to a number of unvectorised loops which are performing very poorly.

A drawback of using SSPs rather than MSPs is that one cannot use a 1xn decomposition for all CPU counts. The maximum value of "n" for the climate configuration is 14, which means that we must decompose in the East-West direction as we increase the CPU count. This can be seen in the performance of the advection, where there is a plateau at 14 and 28 CPUs – each time the decomposition in the East-West direction is increased, and the vector length of the inner loop is halved. Although the effect can be observed, it is comforting that it does not seem to be having a very dramatic effect on performance – even at 32 SSPs where the inner loop length is as low as 24.

We have also made a run of this configuration after carrying out a little optimisation, including replacing MPI with SHMEM and adding vectorisation directives to enable vectorisation in some critical areas (including the "Atmos Update" region discussed previously).



This is a much more encouraging result than the initial scalability graph. The benefit of the low latency SHMEM communications can be seen most clearly in the improvement in scalability of the ocean model.

However, there is clearly still much work to be done to enable this configuration to efficiently utilise the X1's full capabilities. The SX6 optimisation from the Met Office should help this effort, and there is definitely scope to make better use of the low latency communications offered by co-array Fortran, especially within the ocean model.

6. NWP Global Configuration Results

Configuration

The configuration used here is based on the configuration used for operational NWP forecasting at the UK Met Office, featuring an N216 (60km) global atmosphere model. Although CGAM are not concerned with running operational NWP models, this configuration is of interest as it represents the higher resolution of atmospheric models that we are aiming towards with the HiGEM project.

This configuration has been ported to the SX6 by the Met Office where some effort has been made to optimise the configuration to obtain the performance necessary to meet operational deadlines.

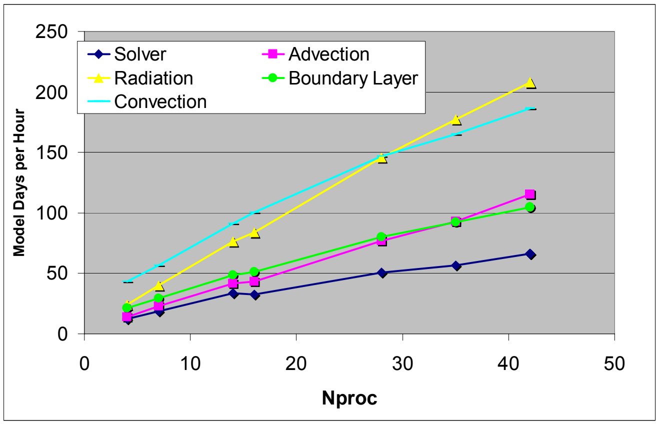
The SX6 optimisations carried out by the Met Office can be classified as follows:

- Addition of compiler directives to improve vectorisation
- Rearrangement of small number of key loops to allow vectorisation
- Automatic and manual inlining of regularly called subroutines/functions
- Optimisation of halo update routine (directly calls asynchronous MPI communications rather than using the interface library)
- Use of the highest level of vector and scalar optimisation `(-Chopt)` where it was safe to do so.

Unfortunately the SX6 optimised version is not yet available to users external to the Met Office, so the version that we have ported and run is based around the T3E configuration, which does not contain the SX6 optimisations. At the time of writing, we have only had the opportunity to port and run this configuration on the Cray X1.

The port to the Cray X1 was again relatively straightforward. A small amount of platform specific optimisation was carried out for the X1, in order to improve the vectorisation in two critical areas; the dynamics solver and the radiation parameterisation scheme. As with the climate configuration, the X1 port uses the machine in SSP mode.

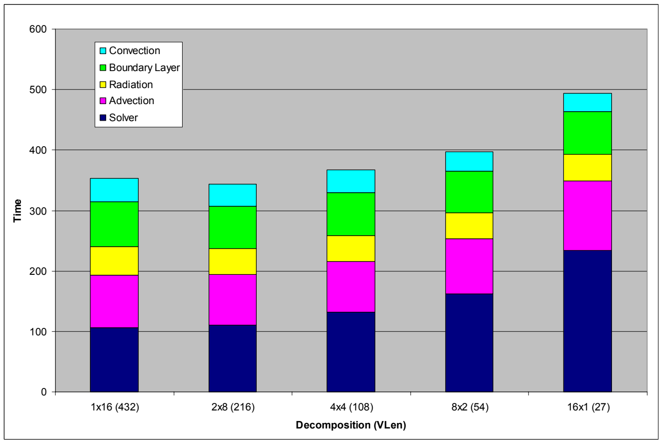
Met Office SX6



These results show good scalability for all the major model components. Vector length was maintained with a 1xn decomposition throughout the range of CPU counts tested. For the reasons described earlier in this paper, a maximum of 7 CPUs per node was used to achieve maximum performance.

It can be seen that the convection and radiation physical parameterisations show better scalability than the dynamical core routines; advection and solver. This is largely due to the fact that these parameterisations contain very little message passing communications, and also because their 1D array structure facilitates longer vectors than are available to the dynamical routines which use a 2D array structure.

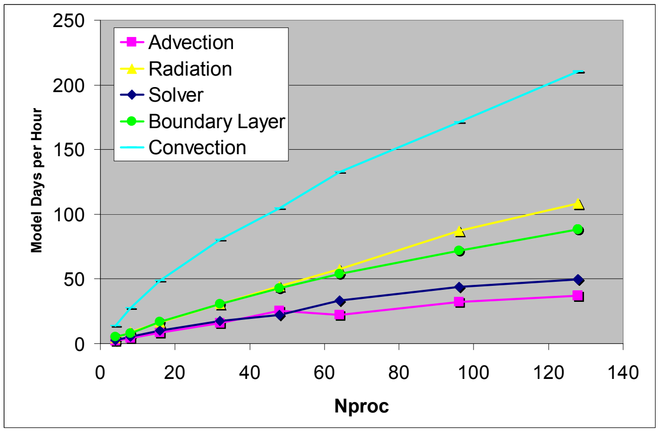
The solver is the worst scaling routine here, which is perhaps unsurprising as it contains many short communications and global reduction operations, both of which test the limitations of the SX6 interconnect.



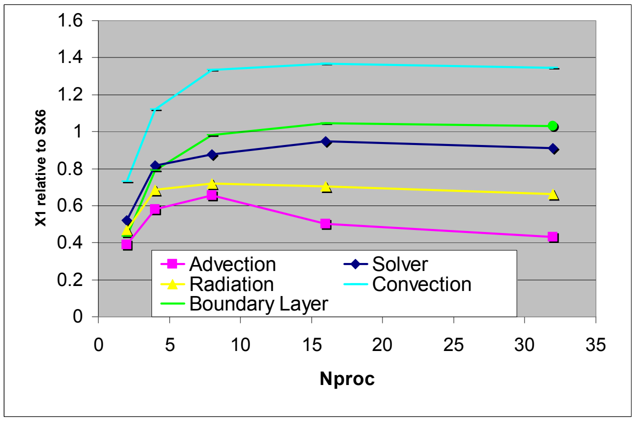
It is interesting to examine the effect of decomposition in the East-West direction, which effectively reduces the vector length in the dynamics. It can be seen that the general trend is an overall increase in CPU time taken by the model as the number of processors in the East-West dimension is increased. It is also obvious from the graph that this is due mostly to the solver routine - this routine is very sensitive to vector length as all the major loops are over “i” – the number of points in the East-West. The physical parameterisations are largely insensitive to the East-West decomposition as they use a 1D array structure.

Close inspection of the graph shows that the best performance is actually achieved on a 2x8 decomposition. Here it appears that the vector length is still long enough not to harm the performance of the solver, but the decomposition actually helps reduce the load imbalance in the boundary layer parameterisation (the land points are better spread over CPUs with this decomposition), and the convection scheme (more processors now cover the expensive equatorial region).

Cray X1

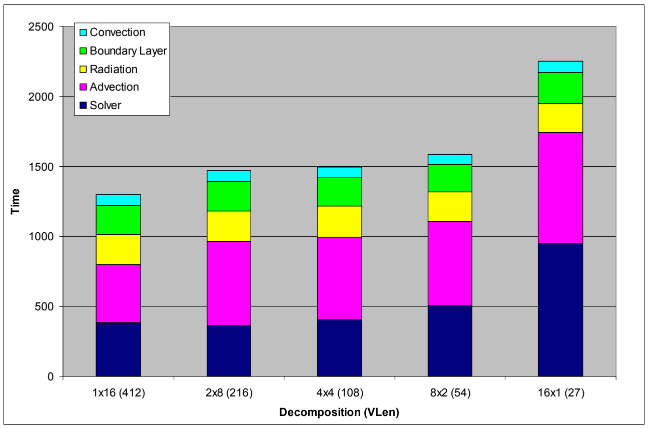


Promising scalability is shown for all the major model components, and the results compare quite favourably with the SX6 results, especially considering the relatively small amount of optimisation carried out for the X1 port. The following graph compares the performance of each of the code sections with the SX6, showing the ratio of X1 times relative to SX6 times (a factor > 1 indicates that the X1 is faster than the SX6). Note that the X1 performance is expressed in terms of MSPs rather than SSPs for ease of comparison even though the code was run in SSP mode.



It can be seen that the relative cost of some components is different between the SX6 and X1. This is perhaps most evident in the performance of the radiation and the advection, which are both performing relatively poorly on the X1, and will need some further attention. The graph indicates that the X1 is benefiting from better performance on large number of processors, indicating better performance of the interconnect. On small number of processors, the relative performance is poorer, indicating that there is much work to be done to improve vector performance of the UM on the X1.

Once again we are using SSPs which limits the use of the 1xn decomposition, so it is useful to understand the affect of increasing the level of decomposition in the East-West direction.



The results are similar to the SX6 results, in that the solver appears to be most sensitive to the decomposition. Unlike the SX6 results, the advection also seems to show some sensitivity to decomposition – most of the extra cost appears as soon as we decompose by 2 CPUs in the East-West. The message passing overheads of the advection do not rise significantly as the number of East-West processors is increased, so this is probably a vectorisation issue, which may well be addressed when the SX6 optimisations are available.

7. Summary

We have demonstrated that it is possible to run the Met Office Unified Model successfully on the Cray X1 system, and that early results show a promising potential for the performance. The lack of an MSP version of the code is, however, a hindrance to performance, as it forces the use of a decomposition in the East-West direction, which decreases vector length in many routines, and increases the amount of communications required. The addition of compiler directives to allow MSP usage, or inclusion of an OpenMP parallelisation would therefore benefit the performance on the X1.

Unsurprisingly, we found better scalability in the higher resolution NWP configuration, where vector lengths are longer, and the ratio between communications and computations is more favourable. However, most parts of the code do scale well even in the low resolution climate configurations, and the overall performance is only brought down by certain areas of codes, whose performance can, no doubt, be improved by small changes to improve the vectorisation potential and make best use of the low latency communications offered by the X1.

We found the communications overhead on the Cray X1 to be smaller than either the IBM p690+ system or the NEC SX6 system, demonstrating Cray's ability to deliver a low latency, high bandwidth communications network, which is of crucial importance to the Met Office's Unified Model. SHMEM shows better performance than MPI, and a co-array Fortran solution would be expected to improve scalability even further.

The return to vector is, unsurprisingly, not painless. However, with the Met Office also moving back to a vector platform, we can be sure that the Unified Model will become increasingly vector friendly, meaning that the X1 and its successors will become very attractive platforms to run this application on.

8. About the Authors

Paul Burton is a computational scientist at CGAM, and previously to this was Manager of the Unified Model System at the UK Met Office. He can be reached at Room 2L46, Dept. of Meteorology, University of Reading, PO Box 243, Reading, Berks, UK. RG6 6BB. Email: Paul@met.rdg.ac.uk

Bob Carruthers is an Applications Consultant with Cray (UK), and has been involved in supercomputing for over 25 years, particularly concentrating on the environmental sector over the last 10 years. He can be reached at Cray UK Ltd, 2 Brewery Court, High Street, Theale, Reading, Berks, UK. RG7 5AH. Email: crjrc@cray.com