Performance Analysis and Optimisation of

METOFFICE UNIFIED MODEL ON A CRAY XC30

EASC 2015 Edinburgh

Karthi Sivalingam & NCAS - CMS, University of Reading
The National centre for Atmospheric Science (NCAS) - Computational Modelling services (CMS) 
NCAS is one of the Natural Environment Research Council's (NERC) research centres.

AIMS
- The science of climate change, including modelling and predictions
- Atmospheric composition, including air quality
- Weather, including hazardous weather
- Technologies for observing and modelling the atmosphere
## HPC & CLIMATE

**Top 500 list Nov 2014**

<table>
<thead>
<tr>
<th>RANK</th>
<th>SITE</th>
<th>SYSTEM</th>
<th>CORES</th>
<th>RMAX (TFLOPS)</th>
<th>RPEAK (TFLOPS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>25</td>
<td>EPSRC/University of Edinburgh United Kingdom</td>
<td>ARCHER - Cray XC30, Intel Xeon E5 v2 12C 2.700GHz, Aries interconnect Cray Inc.</td>
<td>118,080</td>
<td>1,642.5</td>
<td>2,550.5</td>
</tr>
<tr>
<td>28</td>
<td>ECMWF United Kingdom</td>
<td>Cray XC30, Intel Xeon E5-2697v2 12C 2.7GHz, Aries interconnect Cray Inc.</td>
<td>83,160</td>
<td>1,552.0</td>
<td>1,796.3</td>
</tr>
<tr>
<td>29</td>
<td>ECMWF United Kingdom</td>
<td>Cray XC30, Intel Xeon E5-2697v2 12C 2.7GHz, Aries interconnect Cray Inc.</td>
<td>83,160</td>
<td>1,552.0</td>
<td>1,796.3</td>
</tr>
<tr>
<td>30</td>
<td>Science and Technology Facilities Council - Daresbury Laboratory United Kingdom</td>
<td>Blue Joule - BlueGene/Q, Power BQC 16C 1.60GHz, Custom IBM</td>
<td>131,072</td>
<td>1,431.1</td>
<td>1,677.7</td>
</tr>
<tr>
<td>43</td>
<td>University of Edinburgh United Kingdom</td>
<td>DiRAC - BlueGene/Q, Power BQC 16C 1.60GHz, Custom IBM</td>
<td>98,304</td>
<td>1,073.3</td>
<td>1,258.3</td>
</tr>
<tr>
<td>—</td>
<td>Met Office United Kingdom (BBC)</td>
<td>Cray XC40, Intel Haswell</td>
<td>480,000</td>
<td>16,000</td>
<td>16,000</td>
</tr>
</tbody>
</table>
OBJECTIVE

- Message passing, Threading and Vectorisation
- Message passing has scaling limitations
- Single node performance - Threading and Vectorisation
- Intel x86, IBM Power, Coprocessors, Accelerators, ARM 64 bit
- Need for a HPC library that abstracts the architecture
The UM is a numerical modelling system, developed by the UK Met Office, and used for operational weather forecasting and climate prediction.

It is licensed to the UK academic community for research.

Joint Weather and Climate Research Programme (JWCRP), a strategic partnership between NERC and the Met Office for model development.

It is used by forecast centres and climate agencies around the world.
CLIMATE MODELLING

courtesy: Trenberth et al, 2007, 2009

have been made to assemble, analyze, derive, and assess global datasets of water vapor (Trenberth et al. 2005), cloud (Dai et al. 1999b, 2006), precipitation (amount, frequency, intensity, type) (Trenberth 1998; Dai et al. 1999a; Dai 2001a; Trenberth et al. 2003), evapotranspiration (evaporation plus transpiration from plants) (Qian et al. 2006), soil moisture, runoff, streamflow and river discharge into the oceans (Dai and Trenberth 2002, 2003), atmospheric moisture flows and divergence (Trenberth and Guillemot 1998; Dai and Trenberth 2002; Trenberth and Stepaniak 2003a), atmospheric moisture storage (Trenberth and Smith 2005), and freshwater flows in the ocean (Dai and Trenberth 2003). Related issues are the effects of temperature and water-holding capacity, relative versus specific humidity (Dai 2006), covariability of temperature and precipitation (Trenberth and Shea 2005), recycling of moisture (which is taken to mean the fraction of precipitation in a given region, such as a river basin, that comes from moisture evaporated within that basin as opposed to advected in from outside the region) (Trenberth 1999), combinations of temperature and precipitation such as in the Palmer drought severity index (PDSI) (Dai et al. 2004), the diurnal cycle (Dai et al. 1999a,b; Dai 2001b; Trenberth et al. 2003), and forcings of the hydrological cycle, such as solar radiation (Qian et al. 2006). It is well established that latent heating in the atmosphere dominates the structural patterns of total diabatic heating (Trenberth and Stepaniak 2003a,b) and thus there is a close relationship between the water and energy cycles in the atmosphere.

Water vapor is the dominant greenhouse gas (Kiehl and Trenberth 1997) and is responsible for the dominant feedback in the climate system (Karl and Trenberth 2003). However, it also provides the main resource for clouds and storms to produce precipitation, and most precipitation comes from moisture already in the atmosphere at the time a storm forms (Trenberth 1999).
PARALLEL IMPLEMENTATION

- Regular, Static, Lat-Long Decomposition
- Mixed mode MPI/OpenMP
- Asynchronous I/O servers
- Communications on demand for advection
- Multiple halo sizes (up to 8)
# Global Models

<table>
<thead>
<tr>
<th></th>
<th>N96</th>
<th>N144</th>
<th>N216</th>
<th>N320</th>
<th>N512</th>
<th>N768</th>
<th>N1024</th>
<th>N2048</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(192 x 145)</td>
<td>(288 x 217)</td>
<td>(432 x 325)</td>
<td>(640 x 481)</td>
<td>(1024 x 769)</td>
<td>(1536 x 1152)</td>
<td>(2048 x 1536)</td>
<td>(4096 x 3073)</td>
</tr>
<tr>
<td></td>
<td>~135 km</td>
<td>~90 km</td>
<td>~60 km</td>
<td>~40 km</td>
<td>~25 km</td>
<td>~17 km</td>
<td>~12 km</td>
<td>~6 km</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Run length</th>
<th>NWP</th>
<th>Climate</th>
</tr>
</thead>
<tbody>
<tr>
<td>10 day operational forecast, 15 day ensemble forecast</td>
<td>Months (seasonal) Years, decades, centuries+</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Global resolution</th>
<th>NWP</th>
<th>Climate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Testing: N320 (40 km) with 15 min ts</td>
<td>Low resolution: N96 (135 km) with 20 min ts</td>
<td></td>
</tr>
<tr>
<td>Operational: N768 (17 km) with 7.5 min ts</td>
<td>High resolution: N512 (25 km) with 15 min ts</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Dynamics</th>
<th>NWP</th>
<th>Climate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Non-bit reproducible</td>
<td>Bit-reproducible</td>
<td></td>
</tr>
</tbody>
</table>
NCAS SUPPORTED MACHINES

- **Polaris**
  - SGI
- **HPC Wales**
  - Fujitsu
- **UM submission**
  - Code repositories
  - Web server (Trac)
  - User /home space
- **Clustervision**
- **Lander**
- **MONSooN**
- **IBM Power 775**
- **MASS**
- **Jasmin**
- **NOC**
- **Fujitsu**
- **Cray XC30**
- **RDF**
- **Local**

User ssh to all
Climate modelling on Cray XC30

ARChER - Cray XC30, Intel Xeon E5 v2 12C 2.700GHz, Aries interconnect Cray Inc.
Cray Performance Analysis Tools

AIMS
- Performance scaling of the UM at different resolutions
- Performance analysis using Cray PAT tools
- Optimisations for MPI
- Cray Reveal for OPENMP
### UM JOBS

<table>
<thead>
<tr>
<th>JOB</th>
<th>COLUMN</th>
<th>ROWS</th>
<th>LAND</th>
<th>VERTICAL</th>
<th>TIME</th>
<th>RESOLUTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>N96</td>
<td>192</td>
<td>144</td>
<td>11271</td>
<td>85</td>
<td>20 min</td>
<td>135 km</td>
</tr>
<tr>
<td>N216</td>
<td>432</td>
<td>324</td>
<td>52614</td>
<td>85</td>
<td>15 min</td>
<td>60 km</td>
</tr>
<tr>
<td>N512</td>
<td>1024</td>
<td>768</td>
<td>280592</td>
<td>85</td>
<td>10 min</td>
<td>25 km</td>
</tr>
</tbody>
</table>

Number of columns and rows describes the grid of the global model in North-South and East-West (horizontal) direction respectively.
Land points refers to the number of simulated land points.
Vertical levels describes the vertical grid of atmosphere.
Time steps refers to the number of physics timesteps per simulated day.
Resolution refers to resolution of the global grid.
In measuring the performance of UM models, we will use a baseline for further performance analysis. Table 2 lists three standard UM jobs at different resolution. Number of columns and rows describes the grid of the global model in North-South and East-West (horizontal) direction respectively. Land points refers to the number of simulated land points. Vertical levels describes the vertical grid of the model. For global climate modelling, **Timesteps** refers to the number of physics timesteps per simulated day. Resolution refers to resolution of the global grid. For MPI parallelism, we can set the number of processes and rows of a job (as listed in table 2). Processor decomposition that can be used. This is very expensive and on the Cray XC30 we find that the peak wallclock is the time taken for modelling 5 model days.

**$M_{\text{year}} = \frac{1200}{T_{\text{model}}}$**  \[1\]

**$T_{\text{model}} = T_{\text{wallclock}} - T_{\text{initial}}$**  \[2\]

**$C = \frac{1}{M_{\text{year}}} \times n_{\text{core}} \times 24$**  \[3\]

$M_{\text{year}}$ - the number of model years simulated per day.

$T_{\text{model}}$ - time to simulate 5 model days

$T_{\text{wallclock}}$ - total wallclock time

$T_{\text{initial}}$ - time to initialise

$C$ - Cost of simulating a model year in core-hours

$n_{\text{core}}$ - number of physical cores

In this study, we use the IBM Power 775 machine as a baseline for comparison. For MONSooN , we assume that the UM standard jobs will be scaled by 1/1000 for ease of representation (as listed in table 2).

**TABLE 2**

In this paper, a model year is assumed to be 360 days by default unless stated otherwise. These jobs are used as a baseline for further performance analysis. For global climate modelling, this project is considered to be a quantum leap for climate science. In this paper, we study how the optimizations discussed in the UPSCALE project can be improved further for a Cray XC30 machine.
MODEL SETUP

Hyper threading or Symmetric Multithreading

Hyper threading slows UM; SMT achieved ~30% speedup

Bit reproducibility

-e m -s real64 -s integer64 -h O2 -hflex_mp=intolerant -h omp
Perfomance scaling of UM job on ARCHER (ARC) and MON-SooN (MON). Cores refers to the actual number of physical cores used and performance is measured as number of model years simulated in a day (Myear). MON PS and ARC PS refers to perfect scaling that can be expected on MONSooN and ARCHER respectively.
Cost of simulating a model year (C) of the UM job on ARCHER (ARC) and MONSoN (MON) compared to the number of physical cores - ncore (left) and Model years in a day Myear (right).
PERFORMANCE ANALYSIS

The inner pie chart shows the overall profile in which UM includes profile of all UM user code and ETC includes all other library calls.
MPI RANK REORDER

Nearest neighbour communications

MPI ranks - PE configuration 24 x 36

SMP Rank Order (Nodes 0,4,8,12)

MPI rank 37 (and 75) along with the ranks involved in nearest neighbour communications are highlighted. Rank order is based on using 12 MPI ranks per node on ARCHER.

GRID Rank Order (Nodes 0,1,2,3)

I2 PEs per node
Comparison of performance scaling of the N512 job with different MPI rank orders. %Speedup refers to the relative speedup achieved by using GRID rank order instead of SMP rank order.
Imbalance percentage of UM jobs: N96 job running on a single node; N512 job running on 73 nodes; N512 job running on 241 nodes; Thread imbalance percentage are relative to the set of threads used. Imbalance percentage of UM functions are relative to set of PEs.
Cray Reveal - integrated performance analysis and code optimisation tool.

- provides loop analysis and scoping of serial loops.
- suggests OPENMP directive that can be inserted to a loop.
- can attach the performance data collected during execution to identify profile of loops.
- requires knowledge of OPENMP to resolve conflicts and issues.
- works only with Cray compiling environment.
- does not support tasks, barrier, critical or atomic regions.
- For more details - refer Cray documentation (not much)
CRAY REVEAL

Source - ...

Function View

Loop @ 849
Loop @ 851
Loop @ 853
Loop @ 1263
Loop @ 1303
Loop @ 1305
Loop @ 1320
Loop @ 1330
Loop @ 1418
Loop @ 1420
Loop @ 1422

Scope Loops

Name | Type | Scope | Info
--- | --- | --- | ---
qdims | Scalar | Private |
qdims | Scalar | Private |
qdims | Scalar | Private |
qdims | Scalar | Shared
qdims | Scalar | Shared
T_DIMS | Scalar | Shared
T_DIMS | Scalar | Shared

OpenMP Directive

Directive inserted by Cray Reveal. May be incomplete.

```
$OMP parallel do default(none)
$OMP private(i,j,k)
$OMP shared(t_work, QDIMS, T_DIMS)
```

---

First/Last Private -

Enable FirstPrivate

Enable LastPrivate

Find Name:

Reduction:

None
Performance scaling of the UM jobs using increasing number of OpenMP threads on ARCHER/MONSooN. Wallclock time refers to the time taken to complete 2 model days. UM 8.6 refers to the original UM code and UM Reveal to the code with new OpenMP directives. % Speedup is measured as a relative performance improvement achieved by adding new OpenMP directives.
%Speedup of UM functions on MONSooN and ARCHER. %Speedup is measured as the relative improvement achieved by adding OpenMP regions. Performance of UM is measured using 6 MPI and 4 OpenMP threads per node. In X-axis labels, the calltree of the function is specified using ‘:’ delimiter. For example in AS:AP1, ‘Atm Step’ calls function ‘Atmos Physics 1’.
VECTORISATION

- ARCHER compute nodes contain two 2.7 GHz, 12-core E5-2697 v2 (Ivy Bridge) series processors
- supports AVX Instruction set extensions - 256-bit vector SIMD extension
- AVX floating point arithmetic 8x faster compared to scalar
- AVX2 and AVX512 also available
## Vectorisation in UM

<table>
<thead>
<tr>
<th>Category</th>
<th>LOOPS</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total Loops</td>
<td>70000</td>
<td>100.00</td>
</tr>
<tr>
<td>Vectorized</td>
<td>9769</td>
<td>13.96</td>
</tr>
<tr>
<td>Fused</td>
<td>6418</td>
<td>9.17</td>
</tr>
<tr>
<td>Replaced with Library Calls</td>
<td>2207</td>
<td>3.15</td>
</tr>
<tr>
<td>Partially Vectorized</td>
<td>2089</td>
<td>2.98</td>
</tr>
<tr>
<td>Not Vectorized</td>
<td>49522</td>
<td>70.75</td>
</tr>
<tr>
<td>Reason</td>
<td>Count</td>
<td>Vectorization Score</td>
</tr>
<tr>
<td>-----------------------------------------------------------------------</td>
<td>-------</td>
<td>---------------------</td>
</tr>
<tr>
<td>Not vectorized because a better candidate was found</td>
<td>22382</td>
<td>31.97</td>
</tr>
<tr>
<td>Not vectorized because of a potential reassociation issue</td>
<td>10418</td>
<td>14.88</td>
</tr>
<tr>
<td>Not vectorized because a recurrence</td>
<td>5762</td>
<td>8.23</td>
</tr>
<tr>
<td>Not vectorized because it contains a call to subroutine/function/irregular expression</td>
<td>4876</td>
<td>6.97</td>
</tr>
<tr>
<td>Not vectorized because it does not map well onto the target architecture</td>
<td>3482</td>
<td>4.97</td>
</tr>
<tr>
<td>Was not vectorized because the target array (x1) would require rank expansion</td>
<td>371</td>
<td>0.53</td>
</tr>
<tr>
<td>Not vectorized because it contains a reference to a non-vector intrinsic</td>
<td>365</td>
<td>0.52</td>
</tr>
<tr>
<td>Not vectorized because the iteration space is too irregular</td>
<td>408</td>
<td>0.58</td>
</tr>
<tr>
<td>Not vectorized because of unknown reason</td>
<td>1458</td>
<td>2.08</td>
</tr>
</tbody>
</table>
not vectorized because of a potential reassociation issue

If -h flex_mp=intolerant is specified on the command line, then loops are rejected as vector candidates if they contain floating point or complex operations which can potentially cause subtle result differences due to optimization variances between the main vector loop body and any left-over remainder iterations.

```fortran
DO k=1,qdims%k_end
   DO j=1,rows
      DO i=1,row_length
         qrain_inc_step(i,j,k) = qrain_star(i,j,k) - qrain_n(i,j,k)
         qrain_conv(i,j,k) = qrain_n(i,j,k) + fraction_step*qrain_inc_step(i,j,k)
      END DO ! i
   END DO ! j
END DO ! k
```
**VECTORISATION IN UM**

*not vectorized because of recurrence*

Scalar code was generated for the loop because it contains a linear recurrence. The following loop would cause this message to be issued:

```fortran
DO I = 1, 100
   A(I) = A(I-1)
ENDDO

DO j = first_row, last_row
   DO I = row_start_pt, row_end_pt
      iloc = LBC_address(i,j)
      rho_lbc(iloc,k) = p_zero/(R*temp) &
      *exner_lbc(iloc,k)**((1.0-kappa)/kappa)
      exner_lbc(iloc,k+1) = exner_lbc(iloc,k)
   END DO
ENDDO
```

<table>
<thead>
<tr>
<th>Code Count</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>22382</td>
<td>31.97</td>
</tr>
<tr>
<td>10418</td>
<td>14.88</td>
</tr>
<tr>
<td>5762</td>
<td>8.23</td>
</tr>
<tr>
<td>4876</td>
<td>6.97</td>
</tr>
<tr>
<td>3482</td>
<td>4.97</td>
</tr>
<tr>
<td>371</td>
<td>0.53</td>
</tr>
<tr>
<td>365</td>
<td>0.52</td>
</tr>
<tr>
<td>408</td>
<td>0.58</td>
</tr>
<tr>
<td>1458</td>
<td>2.08</td>
</tr>
</tbody>
</table>
not vectorized because it does not map well to the target architecture

The loop contains too many operations that have no clean vector equivalent in hardware for the targeted architecture, and has been left to run as a purely scalar loop. Although the loop is vectorizable from a dependency and idiom standpoint, it would be unprofitable to emulate vector execution using scalar operations.

On other architectures with different hardware capabilities, this loop may be cleanly vectorized.

```fortran
DO k = 1, dim_k_out
  DO j = 1, dim_j_out
    DO i = 1, dim_i_out

      i_out(i,j,k) = i_out(i,j,k) - datastart(1) + 1

    END DO  ! i = 1, dim_i_out * dim_j_out * dim_k_out
  END DO
END DO
```

```
END DO  ! i = 1, dim_i_out * dim_j_out * dim_k_out
END DO
END DO
```
SUMMARY

On Cray XC30, UM performance is characterised by

- 2 OPENMP threads has a load imbalance of 46% that increases as the UM is scaled to higher resolution

- Newly added OPENMP directives result in 5 to 19% speedup

- ~57% of the loops cannot be vectorised by the compiler. Can be improved by not enforcing bit reproducibility

- Message passing does not scale well above 2880 MPI ranks as it consumes more than 50% of the total wallclock time.

- Using a GRID rank reorder results in 5 to 12% speedup.