SEVENTH FRAMEWORK PROGRAMME
Research Infrastructure

FP7-INFRASTRUCTURES-2010-2 – INFRA-2010-1.2.3:

Virtual Research Communities

Combination of Collaborative Project and
Coordination and Support Actions (CP- CSA)

LinkSCEEM-2
Linking Scientific Computing in Europe and the Eastern Mediterranean – Phase 2

Grant Agreement Number: RI-261600

D10.2
Report on Improvements in Scalability of the Climate Modelling Software Tools

Final

Version: 1.0
Author(s): Hendrik Merx, MPG
Date: 31/08/2012

Project and Deliverable Information Sheet

LinkSCEEM Project	**Project Ref. No.:**	**RI-261600**
	Project Title:	**LinkSCEEM-2**
	Project Web Site:	http://www.linksceem.eu/
	Deliverable ID:	**D10.2**
	Deliverable Nature:	**Report**
	Deliverable Level: PU*	**Contractual Date of Delivery:** 31 / 08 / 2012
		Actual Date of Delivery: 12 / 09 / 2012
	EC Project Officer:	**Sonia Spasova**

* - The dissemination level are indicated as follows: **PU** – Public, **PP** – Restricted to other participants (including the Commission Services), **RE** – Restricted to a group specified by the consortium (including the Commission Services). **CO** – Confidential, only for members of the consortium (including the Commission Services).

Document Control Sheet

Document	**Title:**	**Report on Improvements in Scalability of the Climate Modelling Research Tools**
	ID:	**D10.2**
	Version: 0.1	**Status:** Final
	Available at:	http://www.eniac.cyi.ac/
	Software Tool:	Microsoft Word 2011
	File(s):	LinkSCEEM Deliverable D10.2.docx
Authorship	**Written by:**	Hendrik Merx, MPG
	Contributors:	Klaus Klingmüller, CyI-CaSToRC
	Reviewed by:	
	Approved by:	Jens Wiegand, CyI-CaSToRC

Document Status Sheet

Version	**Date**	**Status**	**Comments**
0.1	31/08/2012	Draft	
1.0	12/09/2012	Final version	approved

Document Keywords

Keywords:	LinkSCEEM-2, Computational Science, HPC, e-Infrastructure, Eastern Mediterranean

Table of Contents

PROJECT AND DELIVERABLE INFORMATION SHEET II

DOCUMENT CONTROL SHEET II

DOCUMENT STATUS SHEET II

DOCUMENT KEYWORDS III

TABLE OF CONTENTS IV

LIST OF FIGURES IV

REFERENCES AND APPLICABLE DOCUMENTS V

LIST OF ACRONYMS AND ABBREVIATIONS V

EXECUTIVE SUMMARY 6

1 INTRODUCTION 7
2 THE EMAC MODEL 7
2.1 PARALLEL DECOMPOSITION OF ECHAM 7
2.2 PARALLEL DECOMPOSITION OF ECHAM/MESSY 8
3 LOAD IMBALANCE 8
3.1 EMAC LOCAL IMPLICIT LOAD BALANCING 8
3.2 NON-LOCAL IMPLICIT LOAD BALANCING 9

List of Figures

Figure 1: Parallel decomposition in the ECHAM model using 256 processors. Grid points are coloured according to their assigned processor. The base grid measures 128 points in longitudinal direction and 64 points in latitudinal direction. The processor grid measures 16 processors in each direction. 10
Figure 2: Distribution of the vertically summed computational load for one time step of the homogeneous chemistry submodule MECCA. The time in seconds required for each grid point to complete was measured with the MPI internal timer and added along each column. 10
Figure 3: Distribution of the longitudinally summed computational load for one time step of the MECCA submodule. The time required to complete each grid point was measured and added along the latitudes of the Eastern hemisphere showing the load imbalance of sunset. The heavy load near the Earth's surface is caused by air pollution emissions on the Northern hemisphere. 11
Figure 4: Distribution of the latitudinally summed computational load for one time step of the MECCA submodule. The time required to complete each grid point was measured and added along the longitudes between 30 degrees Southern and 30 degrees Northern latitude. 11
Figure 5: Distribution of the run time in seconds of 256 processors running the MECCA submodel for 96 time steps. Most processors finish in under one second, processors assigned to complex chemical processes need more than two seconds. The run time of the whole simulation is determined by the right flank of the distribution showing the slowest processors at 2.2 seconds. The total number of measurements is 24576. 12
Figure 6: Non-local parallel decomposition for 256 processors. Grid points are coloured according to their assigned processor. 12
Figure 7: Distribution of the run time in seconds of 256 processors running the MECCA submodel following the non-local implicit load balancing policy for 96 time steps. The dispersion has been reduced considerable compared to Figure 5. The right flank of the distribution has been shifted to 1.2 seconds. 13
Figure 8: EMAC run time for one simulated day. Both local and non-local load balancing policies are implemented using both ECHAM and UNITRANS communication. 13

References and Applicable Documents

[1] E. Roeckner, G. Bäuml, L. Bonaventura, R. Brokopf, M. Esch, M. Giorgetta, S. Hagemann, I. Kirchner, L. Kornblueh, E. Manzini, A. Rhodin, U. Schlese, U. Schulzweida, and A. Tompkins, *The Atmospheric General Circulation Model ECHAM5,* Report No. **349** of the Max Planck Institute for Meteorology, Hamburg, Germany, 2003.

[2] P. Jöckel, R. Sander, A. Kerkweg, H. Tost, and J. Lelieveld, *Technical Note: The Modular Earth Submodel System (MESSy) — A new approach towards Earth System Modelling,* Atmos. Chem. Phys. **5**, 433 – 444, 2005.

[3] M. Geimer, F. Wolf, B. J. N. Wylie, E. Ábrahám, D. Becker, and B. Mohr, *The Scalasca performance toolset architecture,* Concurrency Computat. **22**, 702 – 719, 2010.

[4] V. Sandu, A. Sandu, M. Damian, F. Potra, and G. R. Carmichael, *The kinetic preprocessor KPP — A software environment for solving chemical kinetics,* Comput. Chem. Eng. **26**, 1567 – 1579, 2002.

[5] *Ferret* is a product of the National Oceanic and Atmospheric Administration's (NOAA) Pacific Marine Environmental Laboratory (PMEL), Seattle, U.S.A.

[6] *ScalES* has been funded by the German Federal Ministry of Education and Research (BMBF) with reference 01IH08004.

List of Acronyms and Abbreviations

BMBF	German Federal Ministry of Education and Research
CaSToRC	Computation-based Science and Technology Research Centre, part of CyI
CyI	The Cyprus Institute, Lefkosia, Cyprus
DLR	German Aerospace Centre, Oberpfaffenhofen, Germany
EC	European Commission
ECHAM	European Centre/HAMburg, a climate model
ECMWF	European Centre for Medium-Range Weather Forecast, Reading, UK
EMAC	ECHAM/MESSy Atmospheric Chemistry, a climate and atmospheric chemistry model
HPC	High Performance Computing
KPP	Kinetic Pre-Processor, a software tool for the simulation of chemical kinetics
LinkSCEEM	Linking Scientific Computing in Europe and the Eastern Mediterranean
LinkSCEEM-2	Linking Scientific Computing in Europe and the Eastern Mediterranean – Phase 2
MESSy	Modular Earth Submodel System, an atmospheric chemistry model
MPG	Max Planck Society for the Advancement of Science, Germany
MPI	Message Passing Interface, a distributed-memory communication library
MPI-C	Max Planck Institute for Chemistry, Mainz, Germany, part of MPG
MPI-M	Max Planck Institute for Meteorology, Hamburg, Germany, part of MPG
NOAA	National Oceanic and Atmospheric Administration, U.S.A.
PMEL	Pacific Marine Environmental Laboratory, Seattle, U.S.A., part of NOAA
ScalES	Scalable Earth System Models, project funded by BMBF, ref. 01IH08004
UNITRANS	UNIversal TRANSposition library, developed in ScalES

Executive Summary

The ECHAM/MESSy Atmospheric Chemistry (EMAC) model is used by the European climate and atmospheric chemistry modelling community to simulate the Earth’s atmosphere and its interactions with land, ocean, and space. It is based on the general circulation climate model European Centre/HAMburg (ECHAM) that has been extended by the Modular Earth Submodel System (MESSy) to include a variety of atmospheric processes such as homogeneous and heterogeneous chemistry, photochemistry, and aerosols. Some of these processes require computationally complex calculations that are unequally distributed over the simulated atmosphere.

Running the EMAC model requires large amounts of computing resources and is usually performed on parallel supercomputers. While the climate model ECHAM provides the simulations with a parallelisation that has been tailored to the computational load distribution of a standard climate model, the distribution of computational complexity of an atmospheric chemistry model differs markedly creating a load imbalance that wastes valuable computing resources. Furthermore, the load imbalance problem worsens with increasing numbers of processors of future computer architectures that due to limitations in semiconductor physics rely on higher parallelisation rather than faster single processors. Thus, the simulations do not scale well to higher numbers of processors.

The load imbalance and it’s consequent reduction in effective computer performance required a re-evaluation of the load balancing strategies in EMAC and lead to a different parallelisation strategy of the model space for physically local processes. Moreover, the implementation also provided for higher allowed numbers of processors for low-resolution simulations that had been limited by neighbourhood relations for the exchange of data between simulation cells. The new, highly-scalable load distribution policy was implemented in the EMAC model and has been made available to the climate and atmospheric modelling community of the Eastern Mediterranean on BA and CaSToRC systems.

1 Introduction

Aiming for the transfer of state-of-the-art techniques in three thematic areas to regional user communities in the Eastern Mediterranean the LinkSCEEM-2 project contains work packages on research and development in the areas of cultural heritage, climate-related research, and synchrotron radiation. A recent but major problem in climate-related research has been the low scalability of climate and atmospheric chemistry models running on current and future high-performance computing (HPC) architectures. These are characterised by an increase in the number of computational cores rather than increasing the speed of a single processor core.

While different solution strategies leading to new models are presently pursued elsewhere, the ECHAM/MESSy Atmospheric Chemistry (EMAC) model has been used extensively in the European climate and atmospheric chemistry research community. It seemed therefore beneficial to improve the existing model by re-evaluating and re-designing the existing parallelisation strategy. This report describes the recent progress in analysis, design and implementation of the parallelisation policy in the EMAC model.

2 The EMAC Model

The European Centre/HAMburg (ECHAM) model was derived from the operational weather forecast model used at the European Centre for Medium-Range Weather Forecast (ECMWF) and has been adapted for climate experiments at the Max Planck Institute for Meteorology (MPI-M) in Hamburg, Germany. It uses spherical harmonics for the solution of the equations governing the dynamical behaviour of the atmosphere [1] and calculates radiation, clouds, and convection processes as mean values within approximately rectangular volumes of the atmosphere.

The Modular Earth Submodel System (MESSy) framework connects the ECHAM base model to a number of submodels that treat localised atmospheric processes such as homogeneous and heterogeneous chemistry, photochemistry, and aerosols [2]. It has been developed at the Max Planck Institute for Chemistry (MPI-C) in Mainz, Germany and the German Aerospace Centre (DLR).

2.1 Parallel Decomposition of ECHAM

In the ECHAM model the dynamical state of the atmosphere of the Earth is described by variables on points of a grid covering the surface of the planet. In order to integrate the state of the atmosphere in the spectral model ECHAM the model variables have to be transformed between the grid-point description in a polar horizontal and hybrid vertical coordinate system and the space of spherical harmonics in a truncated coefficient representation. These transformations have been accelerated by dividing the data into segments that are worked on in parallel. Data segments resulting from each transformation are redistributed between the parallel processors and fed into the subsequent transformation.

Computational load imbalances resulting from different states of the atmosphere at different locations are being distributed onto the parallel processors according to a north-south load balancing policy assigning to each processor grid points from both hemispheres and hence reducing the load imbalance due to the Earth’s seasons (Figure 1). The parallel decomposition is implemented using the Message Passing Interface (MPI) standard on a distributed-memory architecture.

2.2 Parallel Decomposition of ECHAM/MESSy

The MESSy framework covers localised physical-chemical processes of the atmosphere and inherits its parallel decomposition from ECHAM. Although separate decompositions for single submodules have been shown to be possible, each new data distribution requires a pair or transformations with the accompanying communication demands. Given that new decompositions have to differ significantly from preceding ones in order to have different load balancing properties, their data re-distribution communication pattern approaches an all-to-all structure with maximum data transfer requirements. Furthermore, data re-distribution adds non-scalability to an otherwise scalable approach to the simulation of physical systems due to the locality of physical and chemical processes.

It is therefore advantageous to implement any improved load balancing policy using the inevitable transformations connected to the ECHAM base model.

3 Load Imbalance

In order to solve the load balancing problem it was necessary to measure its extent and try to find its cause. This was achieved by first using the Scalasca parallel performance toolset [3] to determine the amount distribution of computational effort over the base model and the different submodels. For a typical simulation in the resolution T42L90MA the computational load is shared between ECHAM and MESSy in 3 against 7 parts with the homogeneous chemistry submodel MECCA consuming nearly 9 out of 10 parts of the MESSy load.

The MECCA submodel uses the Kinetic Pre-Processor (KPP) [4] to solve the differential equations governing the chemical kinetics. It employs a Rosenbrock-3-type adaptive time step integrator that adjusts its time step size and hence the number of time steps needed to the stiffness of the differential equations manifesting the complexity of the simulated chemical regime. This method is very effective for serial computations as the computational effort scales with the complexity speeding up the simulation for low-complex grid points. In parallel environments it creates a load balancing problem as the computation has to wait for the slowest process leaving the other processes to idle.

Subsequently, the submodel MECCA was fitted with timing analysis routines in order to specify the distribution of load with grid-point accuracy. Plotted using the Ferret visualisation software [5] the distribution shows a distinct load maximum at sunrise and sunset where due to the changing level of light intensity the kinetic differential equations become very stiff and require more and shorter time steps to be solved at a given accuracy (Figure 2). This load maximum is located in the lower stratosphere where the difference between the daytime regime of the chemical kinetics and the night-time regime is greatest (Figure 3 and Figure 4). The grid points situated at the dawn or dusk line experience an up to 100-fold increase in computational complexity against grid points in permanently dark or bright volumes of the stratosphere.

In the figures the tropopause separating the troposphere from the stratosphere can be seen as discontinuity close to level 70 at high latitudes and 65 at low latitudes. The stratopause marking the boundary between stratosphere and mesosphere is not resolved.

3.1 EMAC Local Implicit Load Balancing

The ECHAM model employs an effective implicit load balancing scheme that assigns neighbouring grid points to the same processor. The parallel decomposition splits the total work load along both horizontal directions into segments to be distributed to the processors.

The imbalanced load distribution is offset by averaging the computational load in three dimensions over submodules on each grid point, along vertical columns and horizontal neighbours, the first dimension being non-spatial but rather temporal.

This parallelisation policy is frequently applied to physical problems where a short-range order exists as the data transfer between neighbouring grid points residing on different processors is minimised. It is, however, a localised function that maps a local load imbalance caused by local physical processes onto the local processor. Therefore, its load balancing capacity decreases with growing number of processors as different load levels are concentrated on particular processors. Yet, for moderate parallelisation the ECHAM decomposition is able to reduce the load imbalance from 1:100 to 1:3 as is shown in a run time histogram for each processor (Figure 5).

3.2 Non-Local Implicit Load Balancing

Although EMAC's local implicit load balancing implementation achieves acceptable results for low to moderate parallelisation, its disadvantages are the localisation of physical load imbalance, its reliance on simple MPI point-to-point library functions for data exchange, and an artificial restriction in the ghosting data exchange with neighbours in latitudinal direction. This latter limitation was solved independently from the load balancing problem by allowing lists of neighbours to be used for communication.

While locality does not seem to be advantageous for moderate to high parallelisation, implicitness certainly is, as it can be implemented statically without the need for sorting and re-assigning the work load explicitly to particular processors. Therefore, a non-local implicit load balancing policy was implemented in a simple prototype using the existing ECHAM communication routines. Non-locality was achieved by assigning the physically imbalanced load in a non-local decomposition to the whole set of processors (Figure 6). Each local imbalance is assigned to a multitude of processors and each processor is tasked with work loads from a multitude of locations. Accordingly, the run time histogram shows a rather narrow distribution of each processor's implicitly balanced work load (Figure 7).

While the prototype implementation shows some improvement for low numbers of processors (Figure 8), it doesn't scale well due to the increased communication demands for higher parallelisation. Caused by using simple MPI point-to-point routines with explicit buffer construction data receiver designation, this implementation doesn't use modern MPI features such as collective communication and data types.

Having addressed the first and last disadvantages mentioned above, we have used the UNITRANS library developed in the Scalable Earth-System Models (ScalES) project [6] to improve the data exchange. It provides a simple but efficient abstract interface to the standard MPI routines and enables the non-local load balancing scheme to be implemented with favourable speed-up and scalability. This is achieved by using both collective communication and MPI data types, and by performing the expensive data exchange pattern calculation once at the start of the program and using this communication template for all later data exchanges.

In summary, load balancing in the EMAC model has been improved by combining a non-local implicit load balancing policy with an efficient communication implementation. Neither of these alone achieves both significant speed-up and scalability.

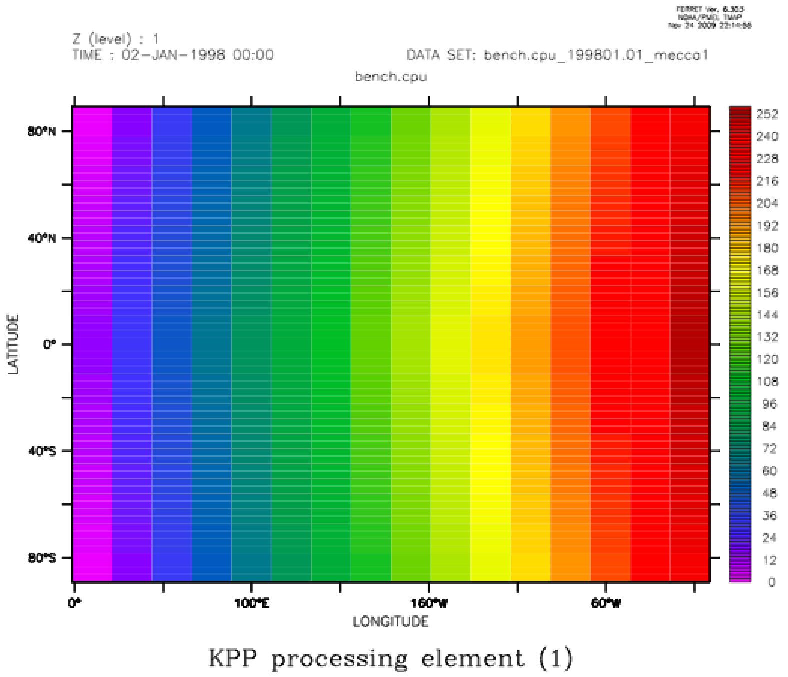

Figure 1: Parallel decomposition in the ECHAM model using 256 processors. Grid points are coloured according to their assigned processor. The base grid measures 128 points in longitudinal direction and 64 points in latitudinal direction. The processor grid measures 16 processors in each direction.

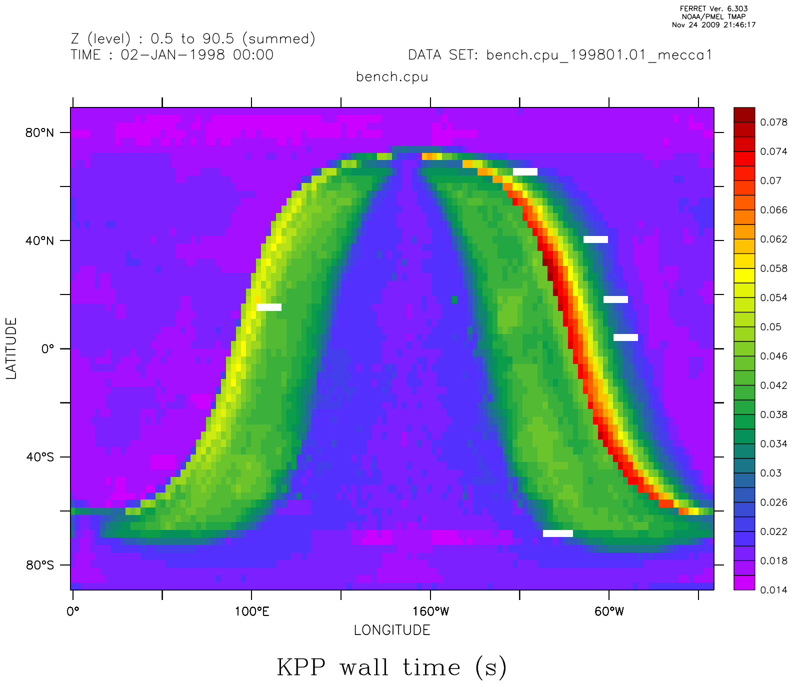

Figure 2: Distribution of the vertically summed computational load for one time step of the homogeneous chemistry submodule MECCA. The time in seconds required for each grid point to complete was measured with the MPI internal timer and added along each column.

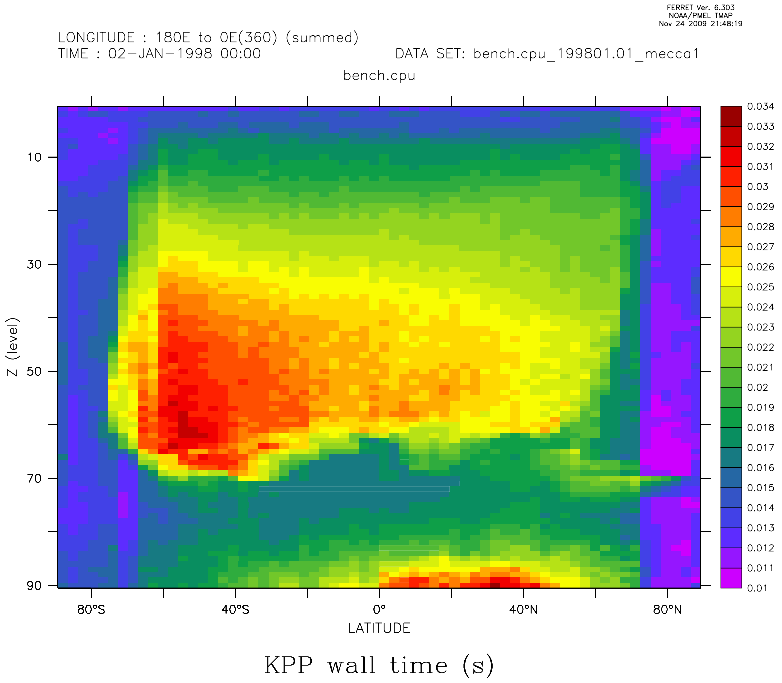

Figure 3: Distribution of the longitudinally summed computational load for one time step of the MECCA submodule. The time required to complete each grid point was measured and added along the latitudes of the Eastern hemisphere showing the load imbalance of sunset. The heavy load near the Earth's surface is caused by air pollution emissions on the Northern hemisphere.

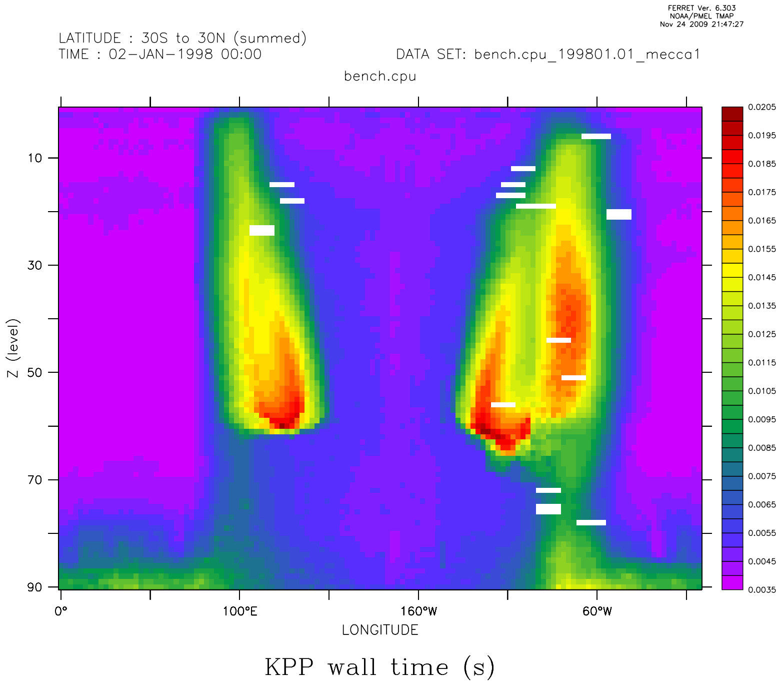

Figure 4: Distribution of the latitudinally summed computational load for one time step of the MECCA submodule. The time required to complete each grid point was measured and added along the longitudes between 30 degrees Southern and 30 degrees Northern latitude.

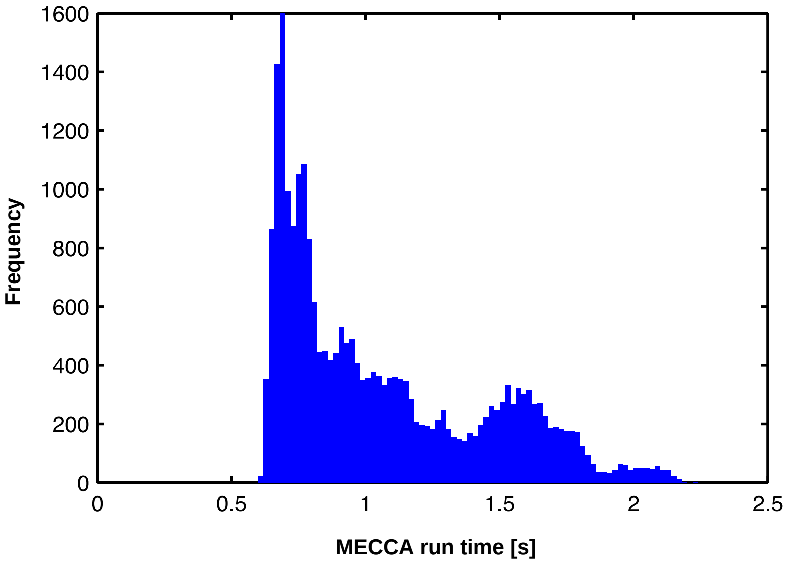

Figure 5: Distribution of the run time in seconds of 256 processors running the MECCA submodel for 96 time steps. Most processors finish in under one second, processors assigned to complex chemical processes need more than two seconds. The run time of the whole simulation is determined by the right flank of the distribution showing the slowest processors at 2.2 seconds. The total number of measurements is 24576.

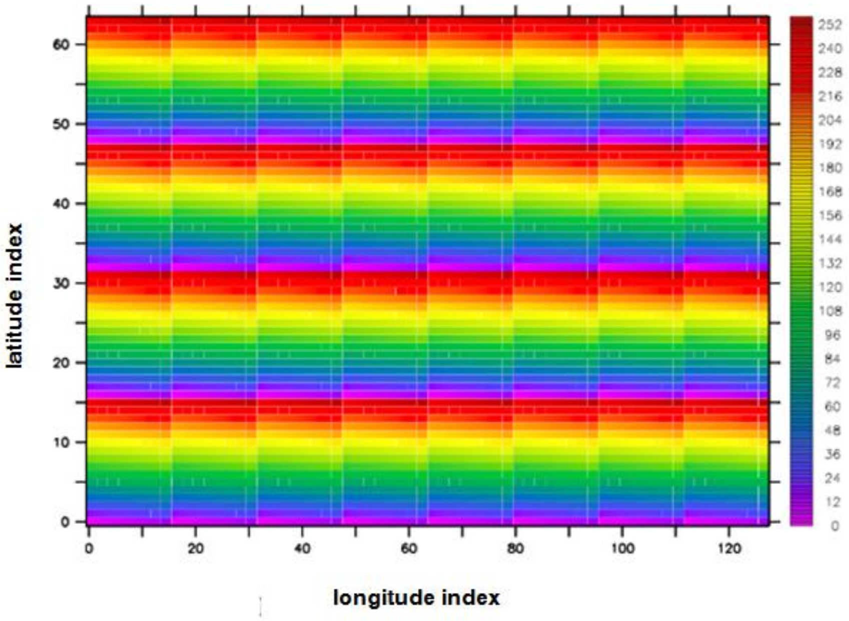

Figure 6: Non-local parallel decomposition for 256 processors. Grid points are coloured according to their assigned processor. Coordinates are internal matrix indexes covering all latitudes from 90 degrees South to 90 degrees North and all longitudes from 0 degrees East to 360 degrees East.

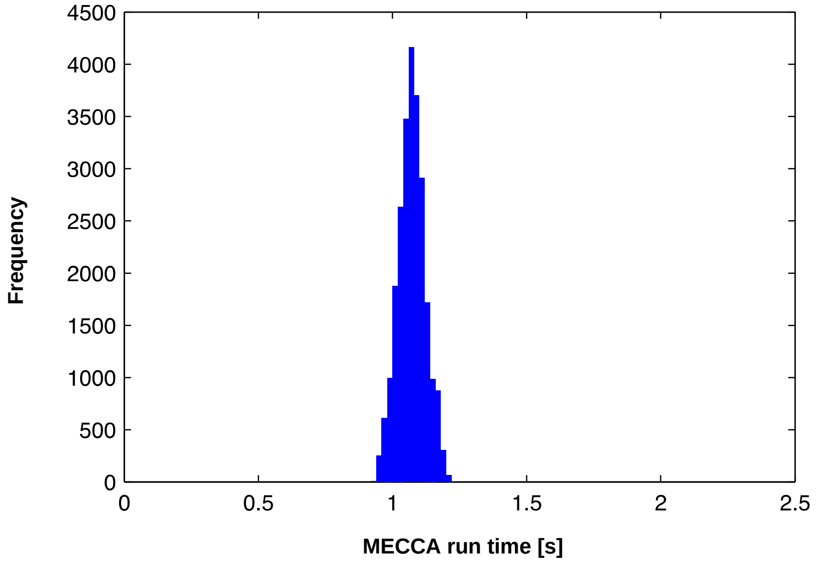

Figure 7: Distribution of the run time in seconds of 256 processors running the MECCA submodel following the non-local implicit load balancing policy for 96 time steps. The dispersion has been reduced considerable compared to Figure 5. The right flank of the distribution has been shifted to 1.2 seconds.

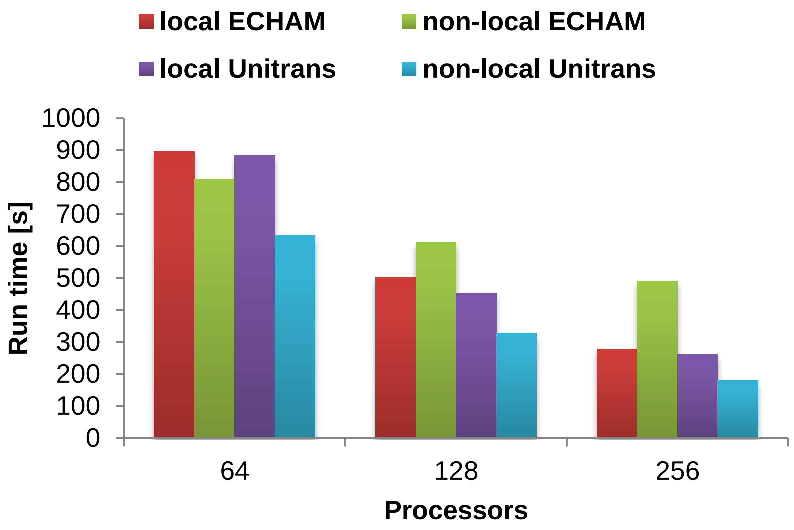

Figure 8: EMAC run time for one simulated day. Both local and non-local load balancing policies are implemented using both ECHAM and UNITRANS communication.