Sensitivity Analysis Study of a Large-scale Air Pollution Model – Computational Problems and High Performance Solutions

Tzvetan Ostromsky*, Ivan Dimov*,
Pencho Marinov*, Rayna Georgieva*, Zahari Zlatev†

* Institute of Information and Communication Technologies (IICT) – Bulgarian Academy of Sciences, Acad. G. Bonchev st., bl. 25-A, 1113 Sofia, Bulgaria
e-mail: *ceco@parallel.bas.bg*, *ivdimov@bas.bg*, *pencho@bas.bg*, *rayna@parallel.bas.bg*

† National Environmental Research Institute, Århus University, Frederiksborgvej 399, Roskilde, Denmark
e-mail: *zz@dmu.dk*

Outline of the talk

- Introduction to sensitivity analysis
- Total Sensitivity Index
- Variance-based methods, Sobol approach
- Monte Carlo calculations for SA
- Application in air pollution modelling, computational difficulties
- The Danish Eulerian Model
- Numerical treatment of the model
- Sensitivity analysis with respect to the chemical rate constants
- UNI-DEM, SA-DEM and their up-to-date parallel implementations
- Scalability of the new SA-DEM implementation on the IBM Blue Gene/P
- Implementation and results on the EGEE Grid
- Conclusions and plans for future work

Introduction to sensitivity analysis (SA)

- Definition of sensitivity analysis (in A. Saltelli et al. (2008)):
  "The study of how uncertainty in the output of a model can be apportioned to different sources of uncertainty in the model input."
- Purpose of our particular SA study:
  To evaluate the changes in the output results, caused by perturbation of certain internal parameters in the chemical scheme of a large-scale air polution model – the Danish Eulerian Model, which determine the speed of the chemical reactions.

Mathematical representation of the SA study

Let $u = f(\mathbf{x}),$ where $\mathbf{x} = (x_1, x_2, \ldots, x_d) \in U^d \equiv [0,1]^d$ is a vector of input parameters with a joint probability distribution function $p(\mathbf{x}) = p(x_1, \ldots, x_d)$.

Total sensitivity index

Total Sensitivity Index (TSI) of an input parameter $x_i, \quad i \in \{1, \ldots, d\}$:

$$S_{x_i}^{tot} = S_i + \sum_{l_1 \neq i} S_{il_1} + \sum_{l_1, l_2 \neq i, l_1 < l_2} S_{il_1l_2} + \ldots + S_{il_1 \ldots l_{d-1}},$$

where $S_{il_1 \ldots l_{j-1}}$ – j^{th} order sensitivity index for the parameter x_i $(1 \leq j \leq d)$,

$j = 1$: S_i – the *main effect* of x_i.

General approaches (overview)

- Local sensitivity analysis (one-at-a-time experiments)
- Screening methods
- Variance-based methods - Sobol approach
  (subject to our study)
- Derivative-based global sensitivity measures

ANOVA - HDMR (Sobol approach)

Analysis of Variances (ANOVA) - based on HDMR of a square-integrable function $f(\mathbf{x})$:

High Dimensional Model Representation (HDMR):

$$f(\mathbf{x}) = f_0 + \sum_{s=1}^{d} \sum_{l_1<\ldots<l_s} f_{l_1\ldots l_s}(x_{l_1}, x_{l_2}, \ldots, x_{l_s}),$$

where

- f_0 - constant,
- $\int_0^1 f_{l_1\ldots l_s}(x_{l_1}, x_{l_2}, \ldots, x_{l_s})\, dx_{l_k} = 0, \quad 1 \le k \le s, \quad s = 1, \ldots, d.$

Sobol (1969), Sobol (1993)

Sobol approach

Therefore

- $\int_{U^d} f_{i_1,\ldots,i_s} f_{j_1,\ldots,j_l}\, d\mathbf{x} = 0, \quad (i_1,\ldots,i_s) \neq (j_1,\ldots,j_l), \quad s,l \in \{1,\ldots,d\}$

and the functions in the right-hand side are defined in a unique way:

- $f_0 = \int_{U^d} f(\mathbf{x}) d\mathbf{x}$
- $f_{l_1}(x_{l_1}) = \int_{U^{d-1}} f(\mathbf{x}) \prod_{k \neq l_1} dx_k - f_0, \; l_1 \in \{1,2,\ldots,d\}$
- $f_{l_1 l_2}(x_{l_1}, x_{l_2}) = \int_{U^{d-2}} f(\mathbf{x}) \prod_{k \neq l_1, l_2} dx_k - f_0 - f_{l_1}(x_{l_1}) - f_{l_2}(x_{l_2}),$
  $l_1, l_2 \in \{1,2,\ldots,d\}$

Sobol global sensitivity indices

Definition (Sobol):

$$S_{l_1 \ \ldots \ l_s} = \frac{\mathbf{D}_{l_1 \ \ldots \ l_s}}{\mathbf{D}}, \qquad s \in \{1, \ldots, d\},$$

where

- partial variances $\mathbf{D}_{l_1 \ \ldots \ l_s} = \int f^2_{l_1 \ \ldots \ l_s} dx_{l_1} \ldots dx_{l_s}$,
- total variance $\mathbf{D} = \int_{U^d} f^2(\mathbf{x}) d\mathbf{x} - f_0^2$,

and the following properties hold:

- $S_{l_1 \ \ldots \ l_s} \geq 0$, $\qquad \sum_{s=1}^{d} \sum_{l_1 < \ldots < l_s}^{d} S_{l_1 \ \ldots \ l_s} = 1$

Various methods for evaluating global sensitivity indices exist (Sobol (1993, 2001), Saltelli (2002), etc.)

Sobol Monte Carlo algorithm for evaluating global sensitivity indices

Let $\mathbf{x} = (\mathbf{y}, \mathbf{z}) \in \mathbb{R}^d, \mathbf{y} = (x_{k_1}, \ldots, x_{k_m}) \in \mathbb{R}^m, K = (k_1, \ldots, k_m)$.
Variance of the subset $\mathbf{y}$: $\mathbf{D_y} = \sum_{n=1}^{m} \sum_{(i_1 < \ldots < i_n) \in K} \mathbf{D}_{i_1,\ldots,i_n}$.

Theorem (Sobol):

$$\mathbf{D_y} = \int f(\mathbf{x})\ f(\mathbf{y}, \mathbf{z}')\mathrm{d}\mathbf{x}\mathrm{d}\mathbf{z}' - f_0^2$$

Combined approach

- Choose a constant $c \sim f_0$ and set the function $\varphi(\mathbf{x}) = f(\mathbf{x}) - c$.
- Use $\varphi(\mathbf{x})$ rather than $f(\mathbf{x})$:

$$\mathbf{D_y} = \int \varphi(\mathbf{x})\ [\varphi(\mathbf{y}, \mathbf{z}')\mathrm{d}\mathbf{z}' - \varphi(\mathbf{x}')\mathrm{d}\mathbf{x}']\ \mathrm{d}\mathbf{x}$$

$$\mathbf{D} = \int \varphi(\mathbf{x})[\varphi(\mathbf{x}) - \varphi(\mathbf{x}')]\ \mathrm{d}\mathbf{x}\mathrm{d}\mathbf{x}'$$

(For more detail, see Saltelli (2002), Kucherenko (2007)).

Computational difficulties in SA study of an air polution model

- The air pollution model is a complicated PDE system - needs splitting with respect to the various physical and chemical processes;
- The numerical methods to solve it require spatial and time discretization with sufficient resolution (to ensure stable calculations);
- Discretization result is a large computational domain (spatial grid, covering the whole Europe, long time period (one year), small time-step, esp. on the chemistry stage);
- Huge amount of I/O data – the main data streams require special treatment in dependence with the target hardware in order to get the best performance of it and to avoid slow-down of the computstional process;
- Need of many similar experiments with changing perturbaton coefficients for each of the studied parameters.

Technology and tools applied in order to solve efficiently the heavy computational problems

The following state-of-the-art technology and tools are used in order to deal with the problem and to overcome the above computational difficulties:

- Parallel computing – portable algorithm with three nested levels of parallelism; high-performance supercomputers.
- Grid computing – powerful heterogeneous computing system, based on low-cost computing nodes (mainly, ordinary PC).

The Danish Eulerian Model

$$\frac{\partial c_s}{\partial t} = -\frac{\partial (u c_s)}{\partial x} - \frac{\partial (v c_s)}{\partial y} - \frac{\partial (w c_s)}{\partial z}$$

$$+ \frac{\partial}{\partial x}\left(K_x \frac{\partial c_s}{\partial x}\right) + \frac{\partial}{\partial y}\left(K_y \frac{\partial c_s}{\partial y}\right) + \frac{\partial}{\partial z}\left(K_z \frac{\partial c_s}{\partial z}\right)$$

$$+ E_s - (k_{1s} + k_{2s}) c_s + Q_s(c_1, c_2, \ldots c_q), \quad s = 1, 2, \ldots q \ .$$

q	– number of chemical species
c_s	– concentrations of the chemical species,
u, v, w	– components of the wind along the coordinate axes,
K_x, K_y, K_z	– diffusion coefficients,
E_s	– emissions in the space domain,
k_{1s}, k_{2s}	– coefficients of dry and wet deposition,
$Q_s(c_1, c_2, \ldots, c_q)$	– non-linear functions that describe the chemical reactions between the species involved.

DEM is described in detail in the books of Zlatev (1995), Zlatev & Dimov (2006).

Splitting into submodels

According to the major physical / chemical processes:

$$\frac{\partial c_s^{(1)}}{\partial t} = -\frac{\partial (u c_s^{(1)})}{\partial x} - \frac{\partial (v c_s^{(1)})}{\partial y} + \frac{\partial}{\partial x}\left(K_x \frac{\partial c_s^{(1)}}{\partial x}\right) + \frac{\partial}{\partial y}\left(K_y \frac{\partial c_s^{(1)}}{\partial y}\right)$$

$$= A_1 c_s^{(1)}(t)$$

horizontal advection & diffusion

$$\frac{\partial c_s^{(2)}}{\partial t} = E_s + Q_s(c_1^{(2)}, c_2^{(2)}, \dots c_q^{(2)}) - (k_{1s} + k_{2s}) c_s^{(4)} = A_2 c_s^{(2)}(t)$$

chemistry, emissions & deposition

$$\frac{\partial c_s^{(3)}}{\partial t} = -\frac{\partial (w c_s^{(3)})}{\partial z} + \frac{\partial}{\partial z}\left(K_z \frac{\partial c_s^{(3)}}{\partial z}\right) = A_3 c_s^{(3)}(t)$$

vertical transport

Related work: Strang (1968); Marchuk (1982);
McRae, Goodin & Seinfeld (1984);
Lancer & Verwer (1998); Dimov, Farago & Zlatev (1999);
Dimov, Farago, Havasi & Zlatev (2001).

The chemical scheme: Condensed CBM IV

- Chemistry is of primary importance for the model and the tuffest computational task.
- It is nonlinear and stiff – often requires reduced time-step, compared to the advection time-step, in order to ensure stability.
- Without chemistry the model brakes out in a number of independent advection-diffusion subproblems.
- Condensed Carbon Bound Mechanism (CBM IV) - Gery et al. (1989)

  35 species (pollutants and other),

  116 chemical reactions:

  - 47 time-independent,
  - 69 time-dependent, including
    - 19 photo-chemical.
- An improved version of the **QSSA** (Quazi Steady-State Approximation) is used for numerical solution of this chemical scheme - Hesstevedt et al. (1978).

Sensitivity analysis with respect to the chemical rate constants

$$\text{Let} \quad \alpha = (\alpha_i)_{i=1}^{d}, \quad \alpha_i \in \{0.1, 0.2, \ldots, 2.0\}$$

be a set of perturbation coefficients, applied to a chosen set of chemical rate constants in the chemical scheme of DEM.

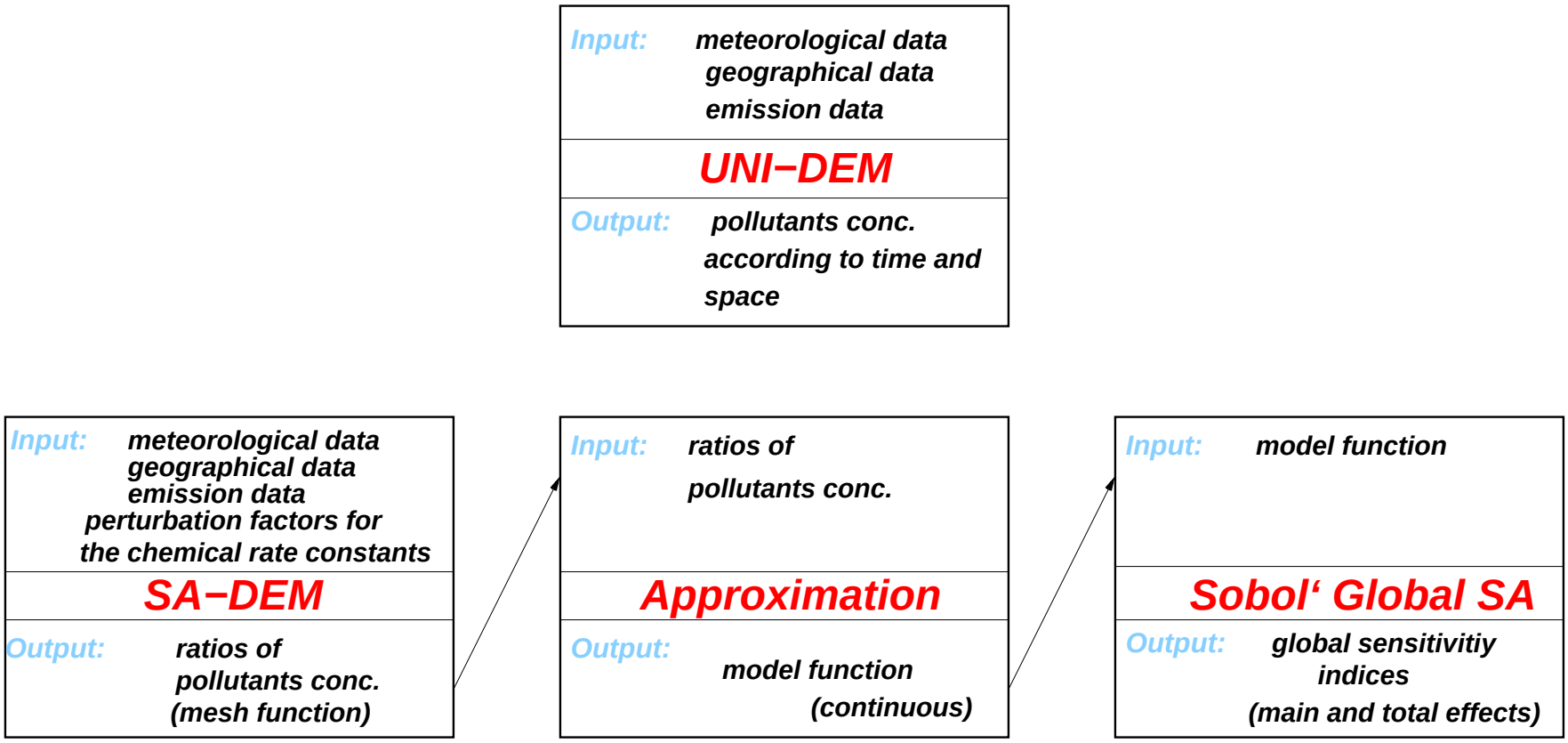
$$r_s(\alpha) = \frac{c_s^{\alpha}(a_s, b_s)}{c_s(a_0, b_0)},$$

Ratios $r_s(\alpha)$ of the above type are calculated to form the initial SA data (model mesh functions). These are a kind of "peak point" mean monthly concentrations, normalized with respect to the central mesh-point $\alpha = (1, ..., 1)$. The "peak point" $(a_0, b_0) \in G$ is located at the position of the c_k true maximum (without any perturbation) in the spatial domain G.

Notation and parameter values

d	– dimension of the SA study, could be any (subset) of the followind sets:
$D^t = \{d_i^t\}_{i=1}^{69}$	– the time-dependent chemical reactions,
$D^c = \{d_i^c\}_{i=1}^{47}$	– the constant chemical reactions,
$D^f = D^t \bigcup D^c$	– all the chemical reactions (full analysis).
$s = 1, \ldots, 35$	– the chemical species (pollutants)
$c_k^{max} = c_k(a_0, b_0) =$ $= \max_{(a,b)\in G}\{c_k(a,b)\}$	– the maximal mean monthly concentration of a fixed pollutant k over the domain G

Phases in the SA study of DEM



UNI-DEM and its parallel implementation features

The development and improvements of DEM throughout the years resulted in a variety of different versions with respect to the grid-size/resolution, the number of layers (2D / 3D) and the number of species in the chemical scheme. The most useful of them has been united under a common driver routine, called UNI-DEM. It provides an uniform and user-friendly access to the available up-to-date versions of the model with an easy way of selecting the appropriate parameters.

The versions, incorporated in UNI-DEM, and the parameters to select one of them are shown below.

Choosable parameters for selecting an optional UNI-DEM version				
Parameter	Description	Optional values		
NX = NY	Grid size (Grid step)	96×96 (50 km)	288×288 (16.7 km)	480×480 (10 km)
NZ	# layers (2D/3D)	1 or 10		
NEQUAT	# chem. species	35, 56 or 168		

SA-DEM components and the top-level parallelization

- A modification of UNI-DEM with direct user access to the chemical rate constants: ability to be modified either separately or in groups in dependence with the dimension of the particular sensitivity analysis study. A small number of input parameters is reserved for this purpose.
- A driver routine that automatically generates a set of tasks to produce the necessary results for a particular SA study. Now it allows to perform in parallel a large number of runs with almost identical input data (reusing it), producing at once a whole set of discrete function values on a regular mesh (used later for calculating the sensitivity indices). This is the coarsest level of parallelism of our algorithm.
- Almost no communications are necessary on this level, except (occationally) synchronization 'by data', as certain temporary data files are used by all the processes. This is implemented (in MPI) mainly by global MPI barriers.

- A new subroutine that splits the global communicator (MPI_COMMON_WORLD) and defines separate communicators for each of the top-level parallel tasks. The communicators are very useful on the lower level of parallelizm (where intensive communications are performed on each time-step).

- An additional program for extracting the necessary mean monthly concentrations and computing the variance mesh functions - not computationally intensive task, but can be done in parallel on this highest level. Currently it is beyond the parallel supercomputer implementation.

  For more efficient use of the processors and from data management viewpoint it might be better to keep it in a separate program (also parallel).

Other important features in the new parallelization scheme of SA-DEM

- Second level of parallelizm (MPI) - based on domain decomposition of the horizontal grid (the traditional MPI parallelization for UNI-DEM)
  - Requires communication on each time step (communication stage) – in **separate communicators**
  - Domain overlapping of the advection-diffusion subproblems – computational overhead, grows up with increasing the number of MPI tasks
  - Improving the data locality for more efficient cache utilization by using **chunks** to group properly the small tasks in the chemistry-deposition stage
  - Additional pre-processing and post-processing stages are needed for **scattering** the input data and **gathering** the results
- Third (finest) level of parallelism (shared memory) - by **OpenMP directives**

Numerical experiments on the IBM Blue Gene/P supercomputer

- Technical specifications of the IBM Blue Gene/P machine:
  - 2048 nodes: quard core PowerPC 450 (850 MHz / 2 GB RAM)
  - 8192 cores in total, theoretical peak performance $>$ 23 TFLOPS
  - 3 modes of node usage (SMP, DUAL, VN)
  - 3-level cache (8 MB L3 cache per node, 32 KB L1 cache per core)
- All experiments are for a time period of one year.
- Each experiment calculates simultaneously 20 different values of certain mesh function of ozone concentration ratios (in 20 different points).
- CHUNKSIZE=48 is used in the experiments. This is a machine-dependent parameter, which needs tunning in dependence with the size of the cache memory of the target machine. It is critical for the performance and has a strong influence on the parallel efficiency. Its optimal value is determined by experiments.

Performance and scalability of SA-DEM on the Bulgarian IBM Blue Gene/P

Time (T) and speed-up (Sp) of SA-DEM on the IBM Blue Gene/P ($96 \times 96 \times 1$) grid, 35 species, CHUNKSIZE=48									
#	Advection		Chemistry		Comm.	I/O	TOTAL		
cores	T [s]	(Sp)	T [s]	(Sp)	T [s]	T [s]	T [s]	(Sp)	**E** [%]
40	3410	(40)	15925	(40)	94	1116	20733	(40)	**100**
80	1715	(79)	7948	(80)	99	1151	11000	(75)	**94**
120	1154	(118)	5291	(120)	138	1051	7664	(108)	**90**
160	870	(157)	3983	(160)	137	1076	6204	(134)	**84**
240	586	(233)	2643	(241)	140	1107	4562	(182)	**76**
320	464	(294)	1974	(323)	153	1131	3810	(218)	**68**
480	344	(396)	1321	(482)	221	1651	3659	(227)	**47**
640	283	(482)	985	(647)	176	1973	3473	(239)	**37**
960	206	(662)	656	(971)	172	1972	3114	(266)	**28**

Time (T) in seconds and speed-up (Sp) of SA-DEM with the 2-level MPI parallelism on the Bulgarian IBM Blue Gene/P (in VN mode). The times for communications and I/O operations are also given.

Time and speed-up of SA-DEM (MPI+OpenMP) on the IBM Blue Gene/P ($96 \times 96 \times 1$) grid, 35 species, CHUNKSIZE=48									
# cores	MPI pr.× OMP thr.	MODE	Advection		Chemistry		TOTAL		
			T [s]	(Sp)	T [s]	(Sp)	T [s]	(Sp)	E [%]
40	40×1	VN	3410	(40)	15925	(40)	20733	(40)	**100**
80	40×2	DUAL	1778	(77)	7972	(80)	11295	(73)	**92**
160	80×2	DUAL	889	(153)	3960	(161)	6153	(135)	**84**
240	120×2	DUAL	647	(211)	2655	(240)	4712	(176)	**73**
320	160×2	DUAL	502	(271)	1978	(322)	4006	(207)	**65**
480	240×2	DUAL	358	(381)	1329	(479)	3418	(243)	**51**
640	160×4	SMP	223	(612)	997	(639)	2768	(300)	**47**
960	480×2	DUAL	218	(626)	659	(967)	2684	(309)	**32**
960	240×4	SMP	158	(863)	667	(955)	2292	(362)	**38**
1280	320×4	SMP	122	(1118)	499	(1277)	2109	(393)	**31**
1920	480×4	SMP	99	(1378)	338	(1885)	2568	(323)	**17**
2560	640×4	SMP	83	(1643)	332	(1919)	2182	(380)	**15**
3840	960×4	SMP	58	(2352)	168	(3792)	1653	(502)	**13**

Time (T) in seconds and speed-up (Sp) of SA-DEM with both MPI and OpenMP parallelizm on the Bulgarian IBM Blue Gene/P

Analysis of the results from the experiments on the IBM Blue Gene/P

- The new 3-level parallel implementation of SA-DEM is a high performance tool for producing sensitivity analysis data, capable to exploit efficiently the computational power of the large IBM Blue Gene/P supercomputer up to its full capacity.
- Chemistry, the most computationally expensive stage of the model, scales almost perfectly in the whole range of experiments.
- Advection stage scales pretty well in most of the experiments, with an expected modest slow-down in the efficiency. It is due to a significant boundary overlapping of the domain partitioning when approaching the inherent partitioning limitations.
- With increasing the number of processors the time for I/O operatios becomes strongly dominant. The problem comes from the insufficient number of I/O devices compared to the CPU number and other resources of the machine. This is the reason for the total efficiensy dropdown in the experiments with extremely high parallelizm.

Numerical experiments on the EGEE Grid /site BG01-IPP at IICT–BAS/

- Resources and access to the Grid site **BG01-IPP**:
  - **Node:** 4 × Intel Xeon X5560 Quard Core Proc. (2.8GHz / 8MB L2 cache)
  - **Size:** 36 nodes (576 cores in total), 23,5 TB storage
  - **CE cluster:** `ce002.ipp.acad.bg`
  - **SE address:** `se001.ipp.acad.bg`
- All experiments are for a time period of one year.
- Each experiment calculates one of 20 different values of certain mesh function of ozone concentration ratios (in one of 20 points).
- CHUNKSIZE=48 is used in the experiments.

Execution times and scalability of SA-DEM on the Bulgarian Grid site BG01-IPP

Time (T) and speed-up (Sp) of SA-DEM on BG01-IPP (96 × 96 × 1) grid, 35 species, CHUNKSIZE=48									
# cores	Advection		Chemistry		Comm.	I/O	TOTAL		
	T [s]	(Sp)	T [s]	(Sp)	T [s]	T [s]	T [s]	(Sp)	E [%]
1	1496	(–)	5214	(–)	0	330	7120	(–)	**100**
2	758	(2.0)	2680	(1.9)	186	237	3909	(1.8)	**91**
4	389	(3.8)	1332	(3.9)	160	184	2094	(3.4)	**85**
8	187	(8.0)	596	(8.7)	191	162	1156	(6.2)	**77**
16	127	(11.8)	367	(14.2)	23	184	725	(9.8)	**61**
32	74	(20.2)	183	(28.5)	22	172	466	(15.3)	**48**

Time (T) in seconds and speed-up (Sp) of the Grid implementation of SA-DEM on the Bulgarian Grid site BG01-IPP at IICT. The times for communications and I/O operations are also given.

Conclusions and plans for future work

- A 3-phase variance-based sensitivity analysis method is proposed and applied to the Danish Eulerian Model. It includes:
  - Generation of input data (mesh functions);
  - Approximation of the mesh functions;
  - Adaptive Monte Carlo algorithm for numerical integration of the approximated functions and computing the global sensitivity indices.
- A special modification of the model (SA-DEM) was developed for the first phase of this method. It was implemented efficiently both as a 3-level parallel algorithm and as a Grid implementation.
- Unllike on the IBM Blue Gene/P, the I/O operations on the Grid does not cause any significant efficiency dropdown in SA-DEM parallel implementation (for not too many processors, at least).
- It is not easy to get on the Grid such a massive parallelism, like on the IBM Blue Gene/P. The performance of a single processor of the BG01-IPP is, however, much higher, compared with a single Blue Gene/P processor.

Plans for future work

- Extending the abilities of SA-DEM (including experiments with more chemical species and on finer resolution grids (storage-permitting).
- Extending the scope of the sensitivity analysis study with respect to a larger set of chemical rate coefficients, as well as with respect to the the emission levels and the boundary conditions.

Thank you for your attention!

Acknowledgments

This research was supported in parts by the Bulgarian NSF via the following Grants:

- DTK 02/44/2009
- DCVP02/1/2010 (SuperCA++)
- DO 02-161/2008