Architectures for Distributed Information Systems Supporting Environmental Simulation

Steffen Unger
GMD FIRST
Kekulèstr. 7, 12489 Berlin, Germany
unger@first.gmd.de

Torsten Asselmeyer
GMD FIRST
Kekulèstr. 7, 12489 Berlin, Germany
torsten@first.gmd.de

Abstract

Different distributed environmental simulation systems are presented. They take advantage of the possibilities of modern network technologies to gather input data, perform simulation and spread the simulation results. The term 'distributed' applies in this context both to the simulation, this is a parallel (distributed) simulation on computer networks, and to the fact, that all components (pre- and postprocessing and the simulation models) may reside in different locations and will be accessed via the Web. The systems described use this architecture to a different extent. Based on this it is shown, that it should be possible to build a system which links various data sources, simulation models, visualization and information tools, all of them running possibly remotely on the corresponding hardware. Such a system would form a distributed virtual laboratory.

1. Introduction

Large simulation models are used for environmental modeling, in particular for air pollution analysis and forecast. Many of them already run in parallel on parallel computers or clusters of workstations or PCs. The term 'distributed' is often used in this context to name a parallel simulation.

On the other hand, pre- and postprocessing, i.e. preparation of different kinds of input data and visualization and interpretation of the results are often an even more challenging task. The support of different institutions like local authorities, weather service, public information sources is required. Most of the sophisticated models nowadays require very special input data, run locally with strongly defined interfaces only, and a large amount of work is necessary to apply such a model to a new domain or link it with different input generators or models. Even if all input data are available, the run of such a model may fail, because often a very specific 'fine tuning' of the control parameters of the model has to be done. For many situations tuned models already exist, which are ready for application, but not open to public access and not prepared to work with a given input data set.

Many users have a strong interest to perform air pollution simulations, even if they don't have their own simulation system. Consider, for instance, the following situations:

- Estimation of the consequences of the release of hazardous substances: A system is required, which is able to run different simulation models at different geographical locations in the air, water and/or soil domain. These models require large sets of input data (local soil structure, orography etc.), which will be available in different databases. The models themselves also will not be available locally at the same place and may have different input requirements and formats. If a common data format and wrappers to the models are defined, they can be linked together to perform a simulation in the given domain.
- Large efforts were made to study consequences of emission reduction measures on the European scale. Based on this transboundary air pollution research, cost/benefit analysis was performed to direct investments into the domains giving most benefit. A number of protocols was signed by the countries to reduce their emissions. It is of broad interest to study the local consequences of such global measures. Therefore, one has to run nested models on the local scale for different locations. The best way to do this is to run the models already available and tuned to these locations. To get comparable results there should be comparable input data (i.e. emission inventories etc.) and the access to the models, which is mostly restricted, must be opened. Such a system can help to standardize the input data sets, and it allows to run the models without violation of their security requirements.

In this paper a concept or architecture of a system is out-

0-7695-0810-3/00 $10.00 © 2000 IEEE

lined, which may allow in future to run the most appropriate model for a given task. A distributed simulation - with respect to this - means, that input data, models, postprocessing and visualization tools are generally distributed over different institutions, universities etc. The necessary workflow, data transfers and wrapping are maintained by the system, the exchange of data will be done via the internet by means of Web servers, which also will provide access to computing power in computing centers. To show, that it is possible to built such a system in the near future, we will discuss 3 projects, where parts of such a system were already realized:

- the automatic forecasting system for ozone in the region of Berlin/Brandenburg, developed and running at GMD FIRST,
- HITERM - HIgh performance computing for Technological and Environmental Risk Management,
- DECAIR - Development of an Earth observation data Converter with application to AIR quality forecast.

The last two were/are funded by the European Community. We will focus on the conceptual and architectural parts oi these projects and show, how they take advantage of the new possibilities offered by the Web.

2. GMD FIRST's Forecast System for Ozone in the Region of Berlin/Brandenburg

This is a very rudimentary system compared to the goal we aim for, but it has already worked automatically, stable and satisfactory for several years (cf. http://www.first.gmd.de/ozon). It consists of a quite simple hydrostatic mesoscale 3 layer model REGOZON [5], simulating local weather conditions and the formation and distribution of air pollutants, in particular ozone.

Input data are the static orography and landuse map, a currently also static emission inventory, and measured and prognostic data (pressure and temperature of a vertical sounding, measured background concentrations for initialization of the concentration arrays, prognostic values for cloud cover and geostrophic wind).

All these current data are gathered by automatic ftp. The main work is to extract the necessary information from these sources, since they are in very special format: meteorological 'temp' for the vertical sounding, listings of the measuring stations with their measurements, textual files giving a forecast of wind speeds, direction and cloud cover, etc. Since this data interpretation has to run automatically, failed measurements have to be sorted out, plausibility test must be performed and in case of failure default values should be provided.

After the successful run of the model, different visualized and textual information is generated, which is placed on a Web site and sent via e-mail to different institutions and a radio station.

Consequently, the Web is used here only for getting/distributing data and information. Special wrappers for the given variable input were developed. The static input data and all the models and tools are locally available at the place where the system runs.

3. HITERM

HITERM was developed to give support to a person/institution in charge of managing accidental release of hazardous substances into environment caused by road or railway traffic accidents or accidents occurring in chemical plants (cf. http://www.ess.co.at/HITERM,[7, 3]). The architecture of the system is shown in Figure 1. The system consists of:

- a main server provided with Graphical User Interface (GUI) and Geographical Information System (GIS) for the specific domain, which runs the prescribed managing scenario for the given type of accident and can connect to different information sources,
- mobile clients at the place of the accident, accessible via the mobile phone network, providing information

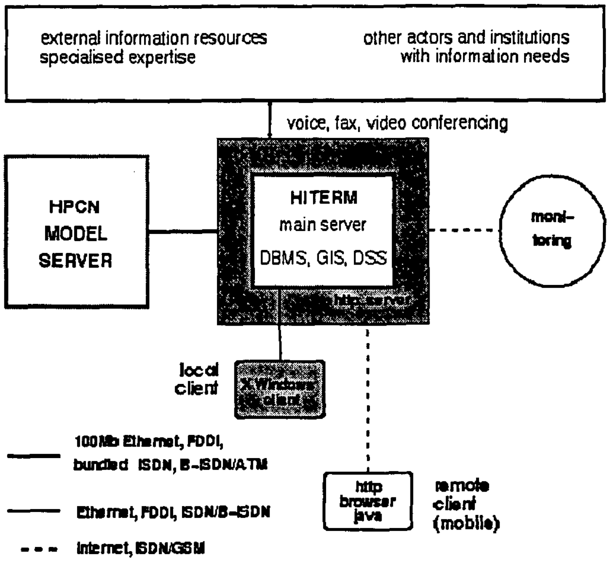

Figure 1. HITERM architecture

about the accident and getting support as result of the information gathered from other sources, in particular simulation results,

- external databases of railway companies, weather service etc.,
- local databases, in particular a chemical database,
- a simulation server possibly on a remote site, connected via the Web, which is able to receive the current input data set (meteorological conditions, spill conditions, parameters of the substance emitted...), to start a simulation and send back the results to the main server. It contains the simulation models, performing on request the simulation of the distribution of the pollutant for the given accident.

Since the simulation is time critical - it has to be much faster than real time - the simulation models were parallelized to run on a workstation cluster.

Consequently, this system already could use clearly defined external data sources, provide and get data to/from a remotely running simulation system and distribute results of its work to different users in different form - text, pictures, diagrams etc.

The corresponding demonstrator is specific for each place it has to be applied (the GIS is built in). It is able to generate a data stream and to send it to other Web servers, which can interpret it and wrap it to input for the simulation models and vice versa.

4. DECAIR

The major objectives of the DECAIR project are to provide data, extracted from satellite data, for air quality simulation models, in order to:

- enhance the quality of simulation results, by enhancing their input data set, and
- ease the implementation of air quality models to new application sites by designing automated input data estimation procedures.

This project is still ongoing (cf. http://www-air.inria.fr/decair). It brings together specialists from air pollution modeling [5, 6], remote sensing research and image processing [2, 1] and information technology [4]. In the first phase it was decided to focus on the generation of detailed, time varying landuse maps for forecast runs and 2D cloud cover arrays for scenario analysis.

One of the major challenges of the project is to define an architecture which is able to support all the functionalities, that are needed to fulfil these objectives. This mainly concerns the remote execution of programs on distributed data and the automatic program control. The DECAIR architecture is designed to be flexible enough in order to easily manage new data sets and new models, as long as a common metadata formulation is adopted. Its potential applications are therefore not limited to the data specified in the project, nor to air quality simulation.

The first step of the project was the analysis of the users (i.e. end-users - institutions or companies responsible of delivering air quality forecasts to the authorities - and model users - scientists developing and applying air quality models) requirements: what are their data needs, what satellite images are able to provide these data and how to process satellite data, what functionalities the architecture must have to support these requirements. This lead to the design of the following system architecture (cf. Figure 2), consisting of:

- a database,
- an access module to the database,
- a database monitor, supervising the freshness of data and controlling the generation of new EO data if new satellite images become available,
- a mediator, providing an interface to the satellite image provider,
- the Web interface, allowing to access the monitor and consequently the database,
- the air quality models.

Since model development and improvement is an ongoing process as well as image processing research and sensor development, the components of the system have to be open and flexible enough for easy adaptation of changing models as well as the incorporation of new data, data types, image processing tools and data converters. Therefore, the DECAIR prototype has to be designed to process and store the data in a way to preserve as much information as possible, irrespective of the current need of the models. All types of input data have to be provided by the DECAIR demonstrator - although only few are determinable at the moment by means of earth observation or in the frame of the project because of sensor availability, lack of proper processing tools, methods or even principle impossibilities. The same applies to the workflow of the preparation of the input data of the air pollution models themselves, even it is not planned to allow to launch remotely the air pollution models in this prototype.

The air quality models and the EO data converter run with huge input data sets; a large part of them will not vary from run to run. Consequently, the core data have to reside nearby the models

- to allow an efficient run of the models and

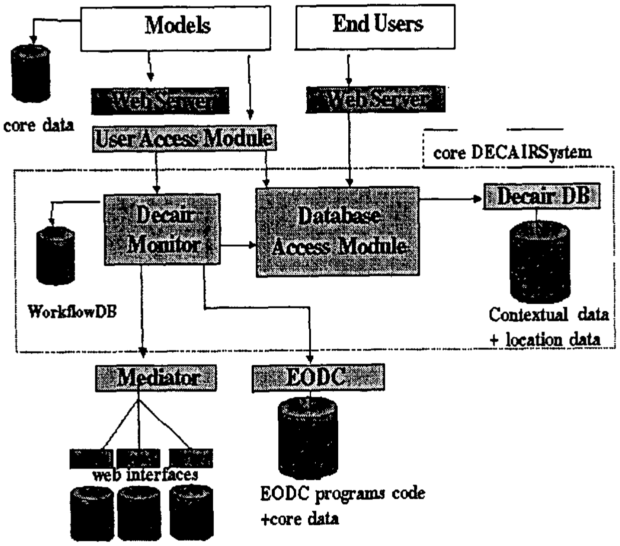

Figure 2. DECAIR architecture

- to avoid frequent data transfers over the Web.

The database will contain only contextual information allowing to query certain data and information on the physical location of the core data. If a user asks for certain data the database is queried and if these data are available in the system the user can decide to actually receive these core data from their physical location. The transfer is also managed by the system.

The EO conversion can be done in automatic mode or by request. In automatic mode (refreshment) as soon as new satellite images are available they are processed, new (fresh) core data are computed and the database stores the corresponding metadata. If a user asks for data for a specific site and date the monitor will first query via the database access module whether these data are already available. If not it will force the mediator to get the necessary satellite images, launch the EO converter to process them and after successful completion inform the user about their availability.

Also, a user can connect to the DECAIR database interactively via the Web interface or automatically query for the availability of new data in forecast mode. He will get information about all the contents of the database and the physical location of the core data. The system manages the requested transfers. If he is allowed, he can also insert data into the database, i.e.:

- current measured data (vertical profiles etc. as in GMD's forecast system; the UPM model [6] works similar) in their original format and/or preprocessed,
- specific data for its application domain and workflow information of his model (assuming there exist wrappers to convert the input data given in the format defined by DECAIR to the format used by the model) for initialization of a new site or model,
- results of simulation runs transferred to the DECAIR format, which then in turn can be queried by authorities or the public.

5. Conclusions

We have shown, that systems already exist where a large part of the features of a distributed virtual laboratory is implemented. With the DECAIR prototype it is even possible to launch programs remotely (EO converters) supplying them with data from a database. To do the same for air pollution models one has to provide the workflow information for the input generation to the database, too. The corresponding system may have the same general structure as the DECAIR system. It will contain a number of different models, linked to the data provided by the database access module. The database also contains information about the necessary data wrappers. The monitor will be able to launch these models. An additional functionality has to be added in case several models are available, which are able to perform the same task (i.e. several air pollution models). Then the monitor has to choose the 'most appropriate' one.

The outlined system can help:

- to standardize input data by supplying a common format for these data and providing wrappers for already included models. This simplifies the adaption of new models to the system.
- to apply models to new domains by providing satellite data in a more or less automatic way and again by help of the standardization of the formats of the input data.
- to perform simulations in a consistent way for different domains.
- to compare the performance of different models for the same task.
- to incorporate as many information as is available and to choose the most appropriate model for a given task. This could improve the results of such a simulation considerably. Many studies become only possible, if these models can be linked together and to the data.

The system is open to include new models, new data, when these become available and the corresponding wrappers/workflows are defined. Thus the problem of linking variety of models in different locations to the data available in distributed data sources can be solved.

6. Acknowledgements

The development of the mentioned systems was a joined work of many groups in Europe and was funded by the European Community. It is not possible to name all the people engaged in the work leading to this paper. We want to acknowledge in particular the groups at ESS Gumpoldskirchen/Austria, the ASIT team from Berne/Switzerland, the people from SYRECO Italy and the LNEC/FCCN/PETROGAL colleagues from Portugal for the close cooperation in the HITERM project. The cooperation in the DECAIR project is still ongoing, the implementation of the database and monitor will be done in the near future. This paper was influenced to a large extend by the discussions during the project meetings about the system architecture and features. We heavily acknowledge the work of Isabelle Herlin, Jean-Paul Berroir, Francois Llirbat and Erik Simon from INRIA (France), Roberto San Jose (UPM, Spain), Jorge Martinez (SICE, Spain), Brian Read (CLRC, Great Britain) and Emmanuel Vavalis (FORTH, Greece).

References

[1] D. Brziat, J.-P. Berroir, S. Bouzidi, I. Herlin, and H. Yahia. Landuse and wind estimation as inputs for air pollution modeling. In A. Sydow, editor, *ERCIM NEWS, Special Issue on Environmental Modeling*. 1998.

[2] I. Cohen and I. Herlin. Non uniform multiresolution method for optical flow and phase portrait models: environmental applications. *International Journal of Computer Vision*, **33**:29–49, 1999.

[3] K. Fedra. Integrated environmental information systems: from data to information. In N. B. Harmancioglu, M. N. Alpaslan, S. D. Ozkul, and V. P. Singh, editors, *Integrated Approach to Environmental Data Management Systems*, pages 367–378. Kluwer, Dordrecht, 1997.

[4] F. Llirbat, R. Hull, B. Kumar, J. Su, G. Zhou, and G. Dong. Optimization techniques for data-intensive decision flows. In *Proc. of Int. Conf. on Data Engineering (ICDE)*, San Diego, February 2000. ICDE.

[5] P. Mieth, S. Unger, and M. Jugel. An environmental simulation and monitoring system for urban areas. *Transactions of the Society for Computer Simulation International*, **15**(3):115–121, 1998.

[6] R. San José, editor. *Measuring and modeling integration of environmental processes*. WIT Press, Southampton, 1999.

[7] S. Unger, I. Gerharz, P. Mieth, and S. Wottrich. HITERM - High-Performance Computing for Technological Risk Management. *Transactions of the Society for Computer Simulation International*, **15**(3):109–114, 1998.