CREATING THE MESSAGE INFRASTRUCTURE

Draft paper for submission to TEC, December 2009

Version 6

[Jeremy Cohen, Robin North, Steven Wilkins, John Darlington, Yike Guo, Neil Hoose, Yajie Ma, John Polak]¹

[Visalakshmi Suresh, Paul Watson, Margaret Bell, Phil Blythe, Jeff Neasham]²

[Mark Calleja, Mark Hayes, Alistair Beresford, Rod Jones, Iq Mead]³

Imperial College London¹, Newcastle University², University of Cambridge³

ABSTRACT

The ever-increasing power of computing infrastructure such as processing, storage and network bandwidth continues to enable new models of data management and analysis. The Mobile Environmental Sensor System Across Grid Environments (MESSAGE) set out to make use of such advances, combined with e-Science techniques, to support the capture, processing, archiving, analysis and visualisation of pollution data. This paper describes the e-Science infrastructure that was developed within MESSAGE to support these tasks.

The Mobile Environmental Sensor System Across GRID Environments (MESSAGE) project is a research project co-funded by EPSRC and the Department for Transport. The project was funded as part of the e-Science initiative with the aim of demonstrating that key outcomes from that programme could be deployed to address real world issues, in this case the detailed monitoring and management of air pollution from road traffic. It is a large project involving five universities; Imperial College London (Centre for Transport Studies, Dept. of Computing, Dept. of Electrical and Electronic Engineering, Dept. of Physics), University of Newcastle (Transport Operations Research Group, Dept. of Electrical and Computer Engineering, School of Computer Science), University of Cambridge (Dept. of Applied Mathematical and Theoretical Physics, Dept. of Chemistry, Computer Laboratory), University of Leeds (Institute for Transport Studies) and University of Southampton (Transportation Research Group). The project ran for 3 years from Oct 2006 with a budget of around £4M and involved more than 30 academic staff, postdoctoral researchers and PhD students. Additionally the project was supported in a variety of ways by 19 non-research organisations from industry and local government and overall co-ordination was provided by Imperial College London.

INTRODUCTION

The ever-increasing power of computing infrastructure such as processing, storage and network bandwidth continues to enable new models of data management and analysis. MESSAGE set out to make use of such advances, combined with e-Science techniques, to support the capture, processing, archiving, analysis and visualisation of pollution data. The project builds on e-Science techniques that have been developed to simplify large-scale, distributed processing within collaborative scientific environments. The techniques used within MESSAGE specifically assist in the management of large-scale data and processing tasks to enable large quantities of data to be captured from a number of geographically distributed, heterogeneous, static and mobile sensor devices. The devices then make use of communications infrastructure to transmit captured data into a distributed processing environment where that data can be processed, analysed and stored in order to support a variety of potential user groups.
Traditional pollution monitoring infrastructure is generally based on an area being instrumented with a sparse set of monitoring stations, however, it has been shown from isolated studies such as [1] that pollution on a street can vary by an order of magnitude over distances of only a few metres and time periods as short as a few seconds. Pollutants such as Nitrogen Oxides (NO/NO$_2$), Carbon Monoxide (CO), Particulate Matter (PM) and Ozone (O$_3$) can be harmful to health, as shown through laboratory exposure research and epidemiological studies [2]. Moreover by combining measurements, multi-species bivariate and multivariate analysis allow for fingerprinting and conditional analysis [3].

MESSAGE takes advantage of the use of smaller, lower cost, sensor devices as part of an integrated network that can potentially be deployed in larger quantities and, where the devices are mobile, they can take advantage of the ability to capture data over a wide spatial as well as temporal range to increase the amount of data available for modelling pollution emissions and dispersion. To further enhance the amount of available data, sensors developed as part of MESSAGE can produce readings at a much higher frequency (potentially as fast as 0.5Hz) than the multi-minute-based averaging of pollution values produced by traditional pollution monitoring equipment.

MESSAGE E-SCIENCE ARCHITECTURE

Figure 1: MESSAGE high-level e-Science Architecture
The MESSAGE e-Science architecture was developed to address the technical challenges facing the full project pipeline, from the sensor devices themselves, through the communications layers, to the real-time storage and long-term data warehousing components. The architecture also supports a potentially unlimited range of higher-level applications and algorithms and makes provision for application specific “data marts” to support them. Integration of third-party data is also possible to support various algorithms and data analysis that may be undertaken. Communication back to the sensors allows for sensor mode operation changes in the event of localised clustering.

A high-level view of the e-Science architecture is shown in Figure 1 and the various elements of the architecture are discussed below.

**Sensors**

A variety of mobile and static sensor devices were developed as part of the MESSAGE project (see Figure 2). These devices were developed by three different project partners. Duvas Technologies Ltd, a spin-out from Imperial College London, developed a high-performance multi-species mobile sensor, based on ultraviolet differential spectroscopy, designed for vehicle mounting or being carried by hand [4]. University of Cambridge developed four broad groups of sensors – types A-D. Type A devices sense CO, NO and NO₂, have onboard GPS and connect to a mobile phone via Bluetooth to transmit data over the mobile phone network. Type B devices sense CO, VOCs (total) and CO₂ and have onboard GPS and GPRS capabilities. Type C devices are similar to type As but with onboard GPRS for direct data transmission without requiring connection to a mobile phone. Type D devices sense CO, NO₂ and temperature and connect to a mobile phone. A microcontroller on the type D devices records pollution and temperature data making use of the analogue microphone input on the phone to pass sensor data to it for transmission over the mobile phone network. University of Newcastle developed a mote device using electrochemical sensors and designed for mounting on street furniture [5, 6]. These devices sense CO, NO₂ and noise and have onboard GPS for positioning and ZigBee for communication.

![Figure 2: A Duvas Technologies/Imperial sensor mounted on a test vehicle (left), A Newcastle mote being mounted on a lamp post (centre), a Cambridge Type A sensor with a 20p piece for scale (right)](image)

The different methods of communication supported by the different device types allowed evaluation of a variety of different data transmission methods when developing and testing the infrastructure. The Duvas Technologies/Imperial devices have the ability to send data using 3G/GPRS via the
mobile phone network enabling greater autonomy and reliability in data transmission but also have support for multi-hop WiFi data transmission with a store and forward configuration enabling data to be buffered locally on the device when no connectivity is available to transmit data. Server-side proxies allowed for the transmission of data to several databases. The University of Cambridge devices use GPRS for data transmission, either directly from the sensor device or via a wireless (Bluetooth) or wired connection to a mobile phone and the University of Newcastle motes use short-range ZigBee peer-to-peer communication to transmit data (potentially routed via other local sensor nodes) to a local gateway device that uses 3G/GPRS or some other available network connectivity to pass data back into the main infrastructure.

Data Capture

Sensors generate packets containing data that they have collected and send this data to an endpoint exposed by the MESSAGE data management infrastructure. The location that sensors send data to may be hard-coded within the sensors’ onboard software or it may be resolved dynamically when a sensor device comes online. Due to the differences in sensor devices, their onboard computing power and technical capabilities, MESSAGE took the approach of developing data capture stacks closely coupled to sensor types in order to ensure maximum efficiency in development and operation. A common database schema was developed for the underlying databases used for data storage and a common interface was developed for data extraction. Co-development between Imperial and Duvas Technologies, enabled a high degree of interoperability between the database and sensor operation. Sensor status could be monitored for diagnostic purposes, operation mode altered, and firmware upgraded remotely.

Imperial College London developed a dynamically scalable data capture infrastructure based on a set of Web Services – a Root Gateway, Sensor Gateways and a Data Storage Service. This infrastructure was developed to handle XML formatted data packets being transmitted by Duvas Technologies’ sensors. Table segmentation was allowed to minimise database operation times on large datasets. The infrastructure was designed with a view to being able to use new, dynamically scalable, pay-per-use Cloud computing infrastructure to allow the number of computing resources to grow or shrink in proportion to the number of sensors active at any point in time. This approach enabled operational costs to be managed much more carefully and removed the need for purchasing and managing large quantities of High Performance Computing (HPC) resources in-house. Deployment of services on Cloud resources was prototyped using the Amazon Web Services Elastic Compute Cloud (EC2) platform [7]. A Root Gateway manages a set of Sensor Gateways and acts a point of contact for sensors connecting to the system. The Root Gateway responds to a sensor’s connection request with details of the Sensor Gateway to which the sensor should connect and begin streaming its data.

University of Cambridge sensors use onboard GPRS to send packets of data to a known endpoint. The server receiving the data is able to carry out pre-processing on the data, where required, and then handles storage of the data into a database.

University of Newcastle’s sensors generate efficiently packed binary packets of data and pass this data, via a gateway, to an Oracle Sensor Edge Server instance. Edge Server is designed to handle feeds from multiple sensors and manages the pre-processing of data and storage into an underlying database.

Data Storage and Database Infrastructure

University of Newcastle led the development of a common database schema for MESSAGE. This schema is based on a modified version of the Urban Traffic Management and Control (UTMC) [8] guidelines. Modifications were made in order to support additional data being generated by certain types of sensor that was not directly supported in the UTMC specification. As a result of legacy requirements or experience at different sites, the project ended up with different groups operating databases on different DBMSs. Given the nature of MESSAGE as an infrastructure that was designed to support easy addition of new sensor types and data stores, this was considered a useful
environment in which to develop the infrastructure providing the challenge of integrating heterogeneous database systems and an appropriate test-bed for the resulting system.

In order to access multiple heterogeneous databases through a common interface, the OGSA-DAI (Data Access and Integration) [9] and DQP (Distributed Query Processing) [10] software, initially developed under the UK e-Science Core Programme and now a product of the UK Open Middleware Infrastructure Institute (OMII-UK), was selected. Within MESSAGE, OGSA-DAI is used as a Web Application that runs within an instance of the Apache Tomcat application server. An OGSA-DAI instance is paired with each database platform that is to be exposed under the common data access interface and a single instance of DQP is then configured to link to these OGSA-DAI nodes (see Figure 3). OGSA-DQP provides a single point of contact that can be used to specify queries referencing data across multiple sites. DQP processes these queries and distributes the relevant elements to the required sites, integrating responses and returning a single result to the end user. A simple web-based interface has been developed in collaboration with OMII-UK to enable SQL queries to be entered into a browser and addressed directly at different MESSAGE databases or at all databases via DQP. The results can be made available in a variety of different formats. Additional plug-ins may be developed for OGSA-DAI to export the data in other formats. Currently CSV, XML and KML data formats can be generated.

Figure 3: Example multi-site OGSA-DAI/DQP deployment with heterogeneous databases

Data Marts and Applications

The top level of the MESSAGE architecture provides an environment for applications and associated data marts. A data mart is a pre-processed data store that is formed from processed sensor data that may have been combined with other third-party data. It is designed to efficiently support higher-level applications by preparing data in the most appropriate format to lower the amount of necessary processing when an application is used. Data marts may be updated at set intervals, or on-demand, refreshing or rebuilding their content based on new data within the raw sensor data store or external third-party data stores. Data marts may be the result of running data through one or more complex algorithms and the process of building/maintaining the content may be time consuming and computationally intensive. Applications may take data directly from a data mart, or from the raw MESSAGE data store, and may have separate user interfaces or simply make their output available as a feed that can be accessed and programmatically used in third-party applications.

Real-time Data Management

The real-time data management element of MESSAGE uses a special instance of a data-mart paired with a real-time data manager that takes a feed of raw sensor data from the data management layer and parses and pre processes this data before pushing it directly to registered clients for display. It
was recognised that real-time data processing could not be linked directly to the main data store due to the latency and inefficiency of storing data and then trying to read it back from the database immediately so an independent Oracle-based data mart services the time window-based analysis/statistical displays shown on the user interface. The real time data manager application is based on the Spring Framework [11] and parses incoming sensor data, converting it to JSON [12] format where necessary. The JSON data is then transmitted to the web-based front end using the BlazeDS [13] open source messaging system. BlazeDS supports both push and polling modes so browsers may poll the data manager to access new data or have it pushed directly to them when it becomes available. The data manager supports a plugin model for parsing code allowing support to be easily added for new sensor types and data formats.

**Distributed Data Mining**

The process of sending data from pollution sensing devices to a processing environment in order to identify patterns highlighting pollution events is reasonable when there are large quantities of data to analyse and when the analysis is not time critical. When analysis is time critical, there are more efficient approaches. MESSAGE took advantage of the potential for localised groups of sensor devices to make ad-hoc communications links by developing methods for in-network data mining where groups of sensor devices could collaborate on the analysis of data they have captured in order to identify problems such as the build up of a pollution hotspot. When such an event is identified, one or more sensors from the collaborating group send a trigger to the master platform highlighting the problem so that it can be logged and remedial action taken. The in-network distributed data mining process is based on the use of a k-means clustering algorithm that is designed to identify small clusters of readings that are outside of acceptable limits and are likely to represent a pollution ‘event’.

**VISUALISATION**

![MESSAGE web-based user interface](Google Maps display Powered by Google, Map data ©2009 Tele Atlas)

Visualisation of data provides many opportunities for the observation of key properties, patterns or
events that may not be immediately obvious when looking at raw data values. In MESSAGE, various different options for visualisations were tested. A web browser-based visualisation (see Figure 4) showing near real-time sensor data and basic analysis of historic data is one of the means of visualisation resulting from the project. The interface is built using Adobe Flex and uses a Google Maps window on which to overlay sensor information. The OGSA-DAI/DQP web query interface described earlier allows data to be output in KML format and this can be imported directly into Google Earth/Maps and visualised. Other visualisation techniques were investigated during the MESSAGE project including using multiple map layers to overlay different pollution data layers on a map, real-time pollution trails, area-averaged grid tiling, and height-based wall maps.

CONCLUSIONS

The MESSAGE project aimed to address many challenging problems and questions in the management of large quantities of data coming from distributed infrastructure using a variety of different types of connectivity. The project tackled the full stack from capture to use covering the difficult process of taking data from sensor nodes, pre-processing it where necessary and then storing it in an infrastructure capable of making it available for use in a wide range of different applications and processes. The display of data from sensors in near real-time was also covered and a web-based interface was developed to show this data and demonstrate a set of general analysis tasks. As well as demonstrating an infrastructure and prototypes of environments to manage and display data, MESSAGE also served to identify important areas where further work is required in developing and maintaining data management platforms and in making these systems efficient.

REFERENCES


**ACKNOWLEDGEMENTS**

The authors would like to thank the sponsors of the MESSAGE project. MESSAGE is a three-year research project which started in October 2006 and is funded jointly by the UK Engineering and Physical Sciences Research Council and the UK Department for Transport. The project also has the support of nineteen non-academic organisations from public sector transport operations, commercial equipment providers, systems integrators and technology suppliers. More information is available from the web site [www.message-project.org](http://www.message-project.org). The views expressed in this paper are those of the authors and do not represent the view of the Department for Transport or any of the non-academic partners of the MESSAGE project. We would also like to thank colleagues working in the MESSAGE project without whose support this work would not have been possible.