

NG-TEPHRA: A Massively Parallel, Nimrod/G-enabled Volcanic Simulation in the Grid and the Cloud

Santiago Núñez*, Blair Bethwaite†, José Brenes§, Gustavo Barrantes‡,
José Castro*, Eduardo Malavassi‡ and David Abramson†

*School of Computer Science and Engineering, Costa Rica Institute of Technology
Cartago, Costa Rica – {snunez,jcastro}@itcr.ac.cr

†MeSsAGE Laboratory, Monash University
Melbourne, Australia – {blair.bethwaite,david.abramson}@infotech.monash.edu.au

‡School of Geographical Sciences, National University of Costa Rica
Heredia, Costa Rica – {gbarran,emalava}@una.ac.cr

§Citizen's Network of Meteorological Stations
San José, Costa Rica – jbrenes54@gmail.com

Abstract—Volcanoes are a principal factor of hazard across the Pacific Rim, with their focus of interest mostly divided into pyroclastic flows and ash deposition. The latter has significantly more impact due to its widespread geographical reach and prolonged effects in human activities and health. TEPHRA is a volcanic ash dispersion model based on a simple version of the advection-diffusion Suzuki model, which has been revisited and modified for the Irazú volcano in Costa Rica. A full parameter exploration is necessary in this particular case (albeit not sufficient) due to scarce observational data. We present in this paper the model, its assumptions and limitations as well as application lifecycle with resulting ash distribution graphics. The computational experimental settings are described, in particular the use of Nimrod/G with respect to non-homogeneous parameter sweeps and its impact on execution time. We also analyze the implementation of a new parameter discard mechanism common to e-Science experiments where sequential generation of new parameter sets has to be complemented with an early verification in order to avoid allocation of CPU time to non-valid scenarios. Finally four sample 100K-scenario runs are analyzed for both traditional HPC clustering and Cloud computing resources in the Amazon EC2 Cloud.

Index Terms—ash dispersion, cloud computing, computational volcanology, grid computing, nimrod, parametric sweep, suzuki advection-diffusion model, tephra

I. INTRODUCTION

Volcanoes play a significant role in human activities and there has been an interest in understanding their dynamics since the beginning of recorded history [12]. In modern times, the rigorous study of volcanoes has led to a broad range of related areas. One of those being ash fall and deposition from volcanic eruptions [24]. The effects on health and aviation are also well known, these being two of the most important motivators in the analysis of volcanic ash dispersion [19], [7]. Also, the regional impact in the economies where volcanic events happen is usually significantly negative.

The geophysical processes and interactions, which lead to such abrupt events as volcanic eruptions and earthquakes, take

place over long periods of geological time. The coupling of volcanism and plate tectonics makes long term prediction of volcanic eruptions infeasible, but there are certain behaviours that signal within more immediate periods a moment of high probability for a volcanic eruption [16], [20]. Application of *ab-initio* methods would require the assembly of resources and information out of reach in terms of computational time, input data and mathematical tools.

A more reasonable approach is to perform simulations that, given the known return periods for volcanic events and their intensity, use probability estimations to generate a *spatial* forecast giving the approximate distribution of materials and damage after a volcanic event. Ash dispersion is a suitable phenomenon for this approach and much benefit is obtained from the application of simulation techniques.

The Pacific Rim contains a varied collection of volcanoes and is the most important region of geological activity on Earth. Costa Rica as part of it has a fair share of volcanic and geological activity: 5 potentially destructive volcanoes with known major events and more than a 100 volcanic structures in total, a subduction zone with boundaries with three major tectonic plates and a microplate. Emergency planning, urban distribution and resource allocation for humanitarian assistance is therefore a large component of the economics in risk mitigation. The last fifteen years have also shown a large increase in economic activities, manufacturing and tourism, all of which are in the path of risk of volcanic eruptions.

Irazú, one of the two volcanoes that lie in the path of winds from North-East to South-West has already shown how such a catastrophe can develop. From 1963 to 1965, it covered the entire metropolitan area with ash. Scarce information was available at the time, with some records from the 18th and 19th centuries collected but little usable data. Even at the time of writing, the authors are required to apply simulation methods due to the uncertainty and scarcity of good data.

This paper describes the ongoing effort to create a Grid-based, Nimrod/G-driven volcanic ash dispersion simulation and spatial forecasting system that can be used in Irazú and where required across the Pacific Rim. Section II summarizes previous efforts and the underlying theoretical background. Section III identifies and describes the methods applied in the current research. Section IV delves into the experimental settings that were used for the analysis, one determined by both the scientific application and the computational test bed in different scenarios, namely the East Cluster at Monash University and the Amazon EC2 Cloud. Section V shows the results from the execution of the simulation model across different platforms. Section VI discusses the findings of the previous sections and their significance with respect to the research problems. Section VII concludes with some final remarks and future work towards the desired system.

II. BACKGROUND

Simulation and modeling of volcanic eruptions is a multi-disciplinary field that encompasses volcanology, geophysics, geology, geography, computational physics and in this case, large-scale, parametric distributed execution of applications in the Grid and the Cloud. This section describes in detail the physical model behind the application, as well as the target simulation area with a historical perspective.

A. Modeling of Volcanic Eruptions: The Suzuki Model

Being this our first approximation to computationally simulate tephra fall we opted to use the Suzuki advection-diffusion semi-analytical model in its simplest form [26], which does not include topography nor particle aggregation, which will be added once the software has been optimized. Hence, it has been used as an introductory exercise to gain insight about the ash fall process and how it is affected by initial data. A more realistic model will require a more refined initial data not available at this point. Furthermore being a semi-analytic model, given fixed parameters and initial conditions its implementation is much more simple than other models with a lower computational cost in comparison.

Based on the experimental finding that vertical diffusion in the volcanic plume is almost null, the model is based on a 2D partial differential equation that describes eddy diffusivity

$$\frac{\partial \chi}{\partial t} = -u \frac{\partial \chi}{\partial x} + \frac{\partial}{\partial x} \left(K \frac{\partial \chi}{\partial x} \right) + \frac{\partial}{\partial y} \left(K \frac{\partial \chi}{\partial y} \right) \quad (1)$$

where u is the diffusive velocity, (x, y) is the position in the 2D plane, positioned in such a way that the origin is on the vent, x points towards the direction of the wind (assumed to be uniform), and z is perpendicularly upward to the ground. χ is the concentration of substance predicted by the model. The function K , which accounts for turbulence, is assumed to be dependent only upon a power of t such that the variance

$$\sigma_r^2 = \frac{\int_0^\infty r^2 \cdot \chi(t, r) \cdot 2\pi r dr}{\int_0^\infty \chi(t, r) \cdot 2\pi r dr} \quad (2)$$

and the apparent diffusivity index

$$A_L \equiv \frac{\sigma_r^2}{4t} \quad (3)$$

approaches empirical relations already known. This is a required standard procedures due to the fact that Eq. 1 is an approximation.

In order to simplify an otherwise mathematically intractably complex phenomenon, Suzuki introduces two adjustable parameters β and λ . The first one, usually within the range $[0.02, 0.5]$ describes the (subjectively observed) distribution of the mass in the vertical z plane. Similarly, λ , to which experimenters have assigned values within the range $[0.25, 2]$ and originally set to 1 by Suzuki, seeks to account the deviation from a ballistic model that ejected particles suffer during their chaotic flight in the column. No unique way of assigning values for β and λ exists since for a known scenario there exist multiple ways of choosing values for both parameters consistently.

Furthermore, equation 2 assumes that the species (in this case, a set of predetermined sizes of ash particles) are fixed throughout the eruption process, readily to be found as a gross approximation if chemistry and particle collisions are taken into account. Indeed, empirical results such as the two peak distribution of fallen ash in the ground, as in the case of Mount Saint Hellens will never be modeled by Suzuki in its current state, as it is the result of particle aggregation while in flight.

Many simulation models have been targeted at modeling volcanic ash dispersion with a different theoretical approach. PUFF [29] and ATHAM [27], are both based on solving variations of Newtonian mechanics, convection, heat transport and fluid mechanics equations. Those normally take full advantage of computational infrastructure based on tightly coupled parallel architectures due to the use of domain decomposition strategies that partitions data for a single scenario.

NG-TEPHRA has three important advantages with respect to PUFF and ATHAM for the current research goals. Albeit less accurate, NG-TEPHRA is simpler and provides initial estimations that can be used to run more detailed models –such as those mentioned above– with refined parameter sets. Second, it provides a simpler theoretical base for understanding ash fall phenomena in different cases and obtain a general picture in a given arbitrary location. Third, it is self-contained and trivially parallelizable, which eases the task of setting up massively distributed simulations in Grid and Cloud environments in contrast with PUFF and ATHAM, designed specifically to run in computing clusters.

B. The Irazú Volcano

Costa Rica lived one of the most important events of its modern volcanic history between 1963-1965 [3], when the Irazú volcano, located NE of the Great Metropolitan Area of San José, erupted violently and produced large amounts of ash. The Great Metropolitan Area of San José received abundant ash falls, up to the point of stressing the economy and health systems. Variable winds at different heights and eruption energies dispersed the ash in very contrasting directions. Ash

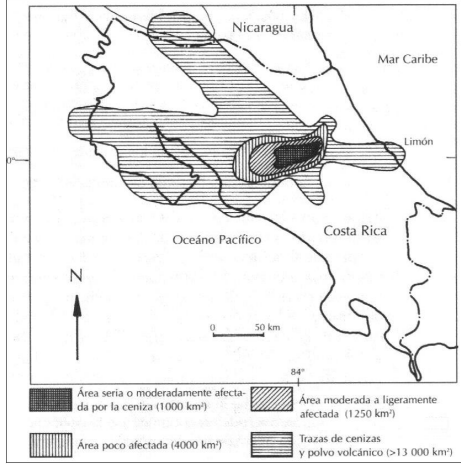


Fig. 1. Irazu volcano location and ash distribution during 1963-1965 [3]

falls were found to reach north of the border with Nicaragua, in the Guanacaste province; to the north-west Pacific coast; and down to Puerto Limón at the central Caribbean coast.

Livestock and agriculture production areas were heavily affected by tephra deposits due to the strong presence of sulfates in their composition. Abruptly, those were forced to relocate to other regions, destroying local economies and delaying potential growth of affected rural areas. Acute medical conditions such as asthma and pneumonia aggravated across the country. Buildings near the vent (usually designed to withstand the rigors of wind and rain only, so having low sloped roofs) succumbed to excessive weight, leading to severe damage to major infrastructure. Limited emergency attention was planned at the time due to scarce information regarding ash distribution and magnitude across the affected areas. At that time little of the risk potential of these events had been taken into account in urban planning, therefore incrementing the potential hazard.

One of the striking features of the eruptions during 1963-1965 was the geographical distribution of the ash (Fig. 1). Most of it followed the direction of NE-to-SW winds, towards the Great Metropolitan Area (GAM). Nonetheless there is additional evidence of ash moving towards the W, NW and E directions when eruptions reached more than 8Km in altitude from the volcanic cone [3]. Wind profiles show Costa Rica has several wind strata relevant to particle dispersion processes [4].

The latter scenario occurred at a time when Costa Rica had less than a million inhabitants –less than half the current population of the GAM alone. Presently, with nearly 3 million inhabitants just in the GAM, it is essential to have a system that allows analysis, characterization and prediction of the impact of volcanic events. It must be integrated with Geographical Information Systems and be parametrizable to other regions of interest where volcanic activity is considered potential and hazardous [13], [25], [28].

III. RESEARCH METHODS AND TOOLS

In order to obtain pertinent and timely results, the appropriate software tools must be utilized. In this section, we describe NG-TEPHRA (Next Generation TEPHRA), the main application which has been fitted to Grid environments. We also describe the use of Nimrod/G, a toolkit that provides the ability to execute large quantities of jobs, in parallel across cluster, Grid and Cloud computational resources.

A. NG-TEPHRA

NG-TEPHRA is a numerical computer simulation that implements the Suzuki model. The original software package was implemented by Connor et al. [11] and it has been extended and extensively modified by the current research group. There currently exist two versions of the application, one targeted to MPI-based cluster computing and another one targeted at large-scale Grid infrastructures using the Nimrod Toolkit.

The implemented Suzuki model used in NG-TEPHRA requires the following inputs: volcano North and East coordinates in CRMT05 (UTM-based projection in meters, WGS84 datum), minimum (ϕ_{min}), maximum (ϕ_{max}), average ($\bar{\phi}$) and standard deviation ($\sigma(\phi)$) of particle sizes in phi units (inverse cm), dimensionless aspect ratio or *skewness* (s) of tephra particles and their mean density (ρ) in gm/cm³, column height (H) in meters measured from the volcanic vent, total erupted mass from one event in kgs, height of the volcanic vent (H_{vent}) in meters above sea level and initial plume velocity ($\bar{v}$) in m/s, the dimensionless beta parameter, average wind speed in m/s and direction in degrees D , V and the probability of the current scenario P .

Basic experiments have been performed approximating the conditions during the period 1963-1965 [5], [6], [21], [22]. The current implementation assumes that (1) particle ascent velocity decreases linearly, (2) diffusion of particles in the volcanic plume has a normal distribution dependent on column height, (3) wind fields are summarized using average direction D and (4) the ash dispersion behaves under the principles of turbulent fluids and gases. It is important to note that for each scenario (run) the wind is thought of as constant at a given height. In order to compensate for the absence of a wind field, multiple wind speed and direction data points are taken, one for each height of interest, therefore approximating layers with significant known differences for these values.

Albeit topography is ignored under the assumption that tephra exhibits a free fall behaviour adjusted by probabilities (an issue addressed by [10]), the simulation permits to obtain results in a simple way and with acceptable accuracy [11]. In terms of numerical simulation, the code is trivially parallelizable – a desirable feature for integration with Nimrod/G. The output of the simulation is the amount of deposited ash $X(x, y)$ for each location (x, y) in the simulation grid, given by the formula

$$X(x, y) = \int_{\phi_{min}}^{\phi_{max}} \int_0^H \frac{5f_z(z)f_\phi(\phi)Q}{8\pi C(t+t_s)^{\frac{5}{2}}} e^{-\frac{[(x-ut)^2+y^2]}{8C(t+t_s)^{\frac{5}{2}}}} dz d\phi \quad (4)$$

where Q is the amount of erupted mass during the eruption, $f_z(z)$ is the diffusion probability of particles outside the plume, a function of height from the volcanic cone ($H = 0$) to the maximum observed height of the plume ($H = H_{max}$). The probability of grain size variation is given by $f_\phi(\phi)$ in ϕ (logarithmic) units. Ash diffusion and fall time are given by t and t_s respectively and are used for estimating the effect of turbulent diffusion in the atmosphere. Wind direction is specified by u . The exponent $\frac{5}{2}$ in the formulas corresponds to solving the eddy diffusivity PDE in 2D.

B. The Nimrod Toolkit

The Nimrod Toolkit enables high-throughput e-Science by allowing researchers to explore complex design spaces using computational models. The various tools support parameter sweeps, numerical search by non-linear optimisation, and even workflows. Experiments can easily be scaled up to span an ad-hoc test bed of many aggregated computational resources, including Grid exposed clusters and more recently, computational Clouds [9]. The Nimrod tools have been successfully applied to much research involving high-throughput e-Science, for example, there has been recent work in fields such as molecular biology [23], cardiology [8] and climatology [15].

In this work we use the distributed execution engine and parameter sweep tool Nimrod/G [1]. Users typically provide Nimrod/G with a plan file that defines the parameter space, and a description on how to execute the application/s. Plan files are declarative and deliberately similar to the job scripts used by batch queue systems, however they also expose file transfer and parameter substitution functionality. Nimrod/G also supports partial parameter sweeps, allowing parameter-sets to be added one-at-a-time via an API, however this requires generating parameters externally.

IV. EXPERIMENTAL SETTINGS

For the validation and organization of input data into a coherent parameter sweep, several sources of data were analyzed, selected and reviewed, and finally a set of parameter limits was developed. In some occasions, when the combined frequency of events was less than 0.5% of observed cases, those were combined with the closest one by means of a weighted average. This provided a reduction from 120 recorded wind directions to only 10 for instance.

Parameters and probabilities are divided according to parameter types, the Volcanic Eruptivity Index to which it belongs, H (column height), D (average wind direction) and V (wind velocity). Parameters such as particle minimum, maximum, average size, clast density and the column shape factor beta are equidistant fixed points with uniform probability in their appropriate ranges. In the case of mass, it is approximated by the formula

$$M = \rho \left(\frac{H}{1670} \right)^4 \Gamma \quad (5)$$

where ρ is the mean clast density, H is the plume height and Γ is the sustained phase time of the eruption [10]. For Irazú, a

Parameter	Total
VEI	4
H	14
D	10
V	7
ρ	5
ϕ_{min}	5
ϕ_{max}	5
β	3
$\vec{v}$	3
Total	4410000

TABLE I
NUMBER OF SCENARIOS AFTER FLATTENING THE PARAMETER TREE

15-minute phase is assumed as the average. The mass is only dependent on height. By following all valid combinations, a total of 217125 scenarios correspond to a full analysis.

NG-TEPHRA does not fit well the existing parameter models in Nimrod/G. The parameters can easily be defined within the existing plan file syntax, however many of the combinations produced by a full cross-product are not geologically sensible but in the execution require new node instances to be spawn only to discard them later (i.e. discard time). Table I shows the cardinality of the set of options for each parameter after the latter mentioned procedure, as well as the total amount of cases to be evaluated by Nimrod/G. These parameter sets represent zero probability scenarios – a result of flattening the parameter tree – and need not be executed. Validating a parameter set is a simple lookup operation and the tables for this already existed from earlier non-Grid enabled versions.

In order to avoid writing new code to externally generate valid parameter sets for NG-TEPHRA and then add them to a Nimrod/G experiment, we added a discard directive to the Nimrod plan file syntax. If specified, the discard line appears as part of the parameters section and specifies the path to a Python module that is loaded dynamically by the Nimrod run file generator. This module must define the method `check_parameters(params_typed, params_strings)`, which should expect two argument dictionaries: one mapping parameters to typed values, and another mapping parameters to their string representations (as in the job environment). `check_parameters` must evaluate to Boolean True for a parameter set to be included in the experiment.

This new feature allows us to easily exploit NG-TEPHRA's existing parameter validation code and can make a significant reduction to overall experiment makespan compared with checking parameter sets as part of each job. In order to perform the current analysis, the parameter sweep has been restricted only to VEI , H , D , V , ϕ_{min} and ϕ_{max} . The latter produces 98000 parameter combinations, where approximately 4.96% are valid scenarios.

Varied resource and experimental configurations were explored. The East Cluster at Monash University was utilized, composed by 3GHz and 1.6GHz machines – a heterogeneous cluster. Also, an Amazon Web Services (AWS) research grant allocation was used to provide EC2 compute time for analysis.

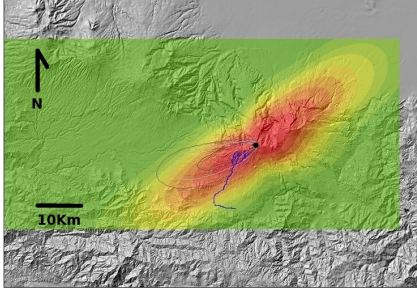


Fig. 2. Sample spatial distribution of accumulated mass for all computed scenarios and superimposed isopacks map from known events during 1963-1965 event. The location of the vent is marked with a black dot. The Reventado river is marked with a blue polyline.

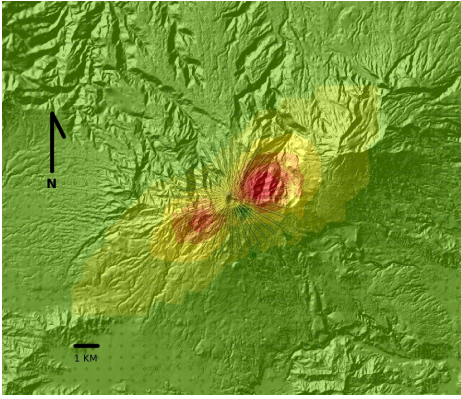


Fig. 3. Close-up at the vent. Mass is scaled to probability, giving affectation areas.

Each run utilized a maximum of 80 concurrent execution slots (dependent on availability), with a single execution slot per CPU-core. On EC2 we used c1.xlarge instances. The following experiments were performed:

- 1) East, no discard, input files copied (**A**)
- 2) EC2, no discard, input files copied (**B**)
- 3) East, discard, input files copied (**C**)
- 4) EC2, discard, input files copied (**D**)
- 5) East, no discard, input files staged using nodestart (**E**)
- 6) EC2, no discard, input files staged using nodestart (**F**)

V. RESULTS

The execution produced a set of 4825 valid files out of 98000 potential scenarios. Those were collected into a single additive sum that represents spatial (not temporal) accumulation assuming no remotion of ash occurs. The plots show amount of expected ash mass deposited per location, expected media grain size and expected grain size standard deviation (Fig. 2-4).

Results are presented for each of the experiments above described. For each case, a plan file was constructed and executed using Nimrod/G. Result sets are divided in two main categories: execution time totals and execution time by node

Metric	A	C	E
T_{CPU}	2,804,394.57	3,256,685.06	3,132,962.00
T_{DSC}	3,432,903.44	0.00	505,144.15
T_{EXP}	6,237,298.01	3,256,685.06	3,638,106.15
F_{CPU}	0.45	1.00	0.86

TABLE II
TOTALS FOR EXPERIMENTS A, C, E

Metric	B	D	F
T_{CPU}	2,235,140.62	2,269,446.66	1,956,799.11
T_{DSC}	6,362,492.60	0.00	707,314.50
T_{EXP}	8,597,633.22	2,269,446.66	2,664,113.61
F_{CPU}	0.26	1.00	0.73

TABLE III
TOTALS FOR EXPERIMENTS B, D, F

type and system. The following metrics were recorded for the aggregate totals:

- Experiment total CPU time [T_{CPU}]
- Experiment total discard time [T_{DSC}]
- Experiment total time (CPU + discard) [T_{EXP}]
- Fraction of CPU time vs total time [F_{CPU}]

For individual processes, the following data were recorded:

- Average process allocation time [$\bar{t}_{ALC}$]
- Standard deviation of allocation time [σ_{ALC}]
- Average process discard time [$\bar{t}_{DSC}$]
- Standard deviation of discard time [σ_{DSC}]
- Average process CPU (ALC - discard) time [$\bar{t}_{CPU}$]

All results in Tables II-VI are given in seconds as the time unit.

VI. DISCUSSION

Results in Fig. 2-3 reveal some important facts about the distribution of ash with respect to probability. Fig. 2 shows that most of the risk of ash deposits remains close to the vent area and in the direction NE to SW winds. Most effect concentrates on eruptions that follow that trend due to having higher probability across the year, therefore contributing more

Metric	A	C	E
$\bar{t}_{ALC}$	374.98	0.00	854.25
σ_{ALC}	7.63	0.00	16.01
$\bar{t}_{DSC}$	36.94	NA	5.42
σ_{DSC}	4.17	NA	1.95
$\bar{t}_{CPU}$	338.04	0	848.83

TABLE IV
PER-PROCESS DATA FOR A, C, E, RUNNING ON 3GHZ NODES (EAST)

Metric	A	C	E
$\bar{t}_{ALC}$	679.39	674.96	649.23
σ_{ALC}	47.30	29.12	13.46
$\bar{t}_{DSC}$	36.80	NA	5.42
σ_{DSC}	3.57	NA	1.95
$\bar{t}_{CPU}$	642.59	674.96	643.81

TABLE V
PER-PROCESS DATA FOR A, C, E, RUNNING ON 1.6GHZ NODES (EAST)

Metric	B	D	F
$\bar{t}_{ALC}$	463.24	470.35	405.55
σ_{ALC}	10.32	9.30	7.45
$\bar{t}_{DSC}$	68.29	0.00	7.59
σ_{DSC}	2.98	0.00	1.4
$\bar{t}_{CPU}$	394.96	470.35	397.96

TABLE VI

PER-PROCESS DATA FOR B, D, F, RUNNING ON AMAZON EC2 C1.XLARGE INSTANCES

to ash deposition than large eruptions in the long term (VEI 3). Fig. 3 presents mass distribution scaled with its accumulated probability for every point. The two spots of higher concentration with respect to probability are close to the vent with displacement dependent upon β . It is worth noting that the predominant directions are in agreement with the expected behaviour. Winds at higher altitudes imply ash traverses longer distances, but violent eruptions –those where ash reaches more than 8Km– are infrequent. The two predominant directions agree with the two most frequent wind patterns in the dry and rainy season at low and medium altitudes.

Table II shows the results obtained from experiments **A**, **C**, and **E**. Experiment **A** has the longest duration and also the highest execution time of the three, and the lowest utilization of nodes across time. Experiment **C** uses Nimrod-discard, therefore having a complete allocation of nodes with useful work (a 55% difference with respect to **A**); the latter reflects sensibly in the total execution time. A remarkable result is that **A** copied execution files for each parameter-set, rather than by using the `nodestart` directive in Nimrod/G to stage them to the execution slot for the first and subsequent sets; experiment **E** shows that, even without the discard mechanism there is a reduction of nearly 680% in discard time if `nodestart` is used. That is, a good file handling scheme is essential to efficient execution in massively distributed, grid experiments. Results in Table III are even more striking. Experiments **B**, **D**, and **F** took longer to complete than the previous ones due to higher latency in the Amazon EC2 Cloud, but for **F** the reduction in discard time was even higher, a 900% with respect to **B**. Figs. 5-8 show job execution over time for scenarios **A**, **B**, **E**, **F** and illustrate the overhead of discarding in situ. The overall trend is consistent with the value of F_{CPU} in each case.

Tables IV, V and VI show disaggregate data per node type. Nodes with 3GHz processors have shorter execution time per process than EC2 instances (except for experiment **E**) and both of them are lower than those for 1.6GHz nodes, as expected. Two observations captured our attention. Results for EC2 instances have very small differences in their standard deviation, this contrasts with findings in our prior work [9] and could indicate the presence of intermediate optimizations geared towards consistent quality of service. The second important result is that by using `nodestart` the standard deviation for each process and the discard time decreased significantly, as expected from the results of Tables II and III.

One of the striking aspects of these results is that it is not obvious at first sight that problems with dependencies amongst

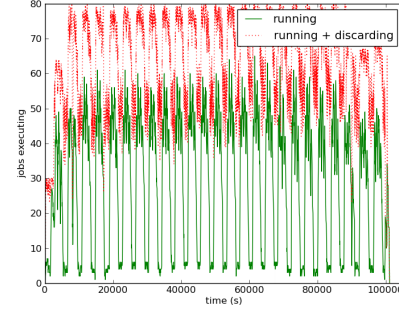


Fig. 4. Task distribution, experiment **A**

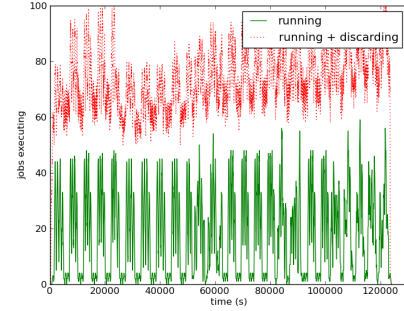


Fig. 5. Task distribution, experiment **B**

parameters behave inefficiently when their parameter sets are flattened and set straight into a plan file. For the case of NG-TEPHRA, after careful analysis it was fairly clear that most scenarios would be discarded since dependencies were still simple enough to keep track of. There might be nonetheless problems where (a) flattening is the best way to generate parameter sets and execute parallel jobs and (b) the complexity of dependencies in parameters are not trivial to determine, therefore increasing the uncertainty on the fraction of useless parameter sets ($1 - \alpha$) and their final impact upon execution

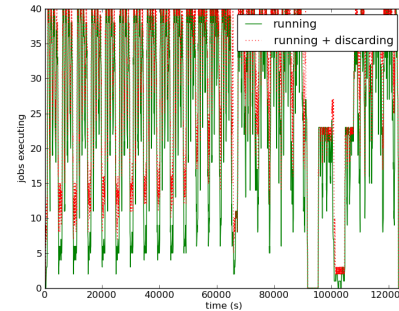


Fig. 6. Task distribution, experiment **E**

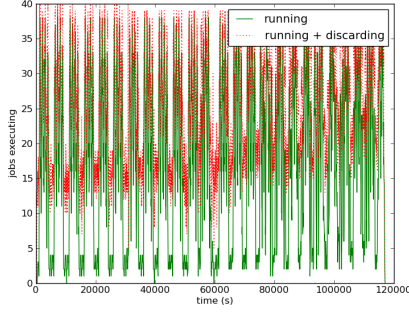


Fig. 7. Task distribution, experiment F

time. Notice that $F_{CPU} = \alpha$ in the above shown results.

In a non-discard execution with N nodes and P processing units, the total execution time T_{NORM} of the simulation in a dedicated environment has a lower bound of

$$T_{NORM} > \frac{N}{P} [\alpha T_{Run} + (1 - \alpha) T_{Disc}] \quad (6)$$

where T_{Run} is the time of a valid task and T_{Disc} is the total time of discarding an invalid task. In contrast, using the discard mechanism, the optimised execution time amounts to

$$T_{OPT} > \frac{N}{P} \alpha T_{Run} \quad (7)$$

Note that in the last equation the $>$ indicates that there is still cost from parallel execution, therefore being an ideal lower bound that does not take into account scheduling and the remote dispatch overhead. The performance ratio

$$\frac{R_{OPT}}{R_{NORM}} = \frac{T_{NORM}}{T_{OPT}} \quad (8)$$

becomes significant under the assumption of pertinence of results, in an emergency scenario.

An important annotation is that the original version of TEPHRA using MPI had much larger figures for execution data (16 minutes per process). Even if the scenarios were generated by a script that guaranteed no zero-probability scenario was produced, the cost of using MPI is prohibitive and grows as more processes are allocated. The latter is due to the fact the cost of using MPI implementations for trivially parallel tasks with respect to the operating system and Grid schedulers is higher than running a single process.

VII. CONCLUSIONS AND NEXT STEPS

This work showed that by using a large-scaled, distributed, Grid-enabled infrastructure it is possible to overcome lack of substantial data and approximate the conditions in a volcanic period of intense eruptivity. Moreover, Nimrod/G provides a robust and efficient platform for doing so. In that regard, two observations comprehend the main lines of work in our current research: (a) efficient scientific workflows are possible both in the Cloud and standard Cluster implementations and (b) setting

up an experiment requires careful design of the information and process flow to be executed. With respect to clusters, the Cloud has a good advantage. If enough resources can be procured and the problem is decomposed in a trivially parallel set of tasks, the total execution time will depend only on the amount of available instances and is not subject to concurrent system load from users. If on the contrary, the problem can not be easily put in a one-instance, one-parameter set, Cloud usage is not the best alternative.

A word on the parameter structure is worthwhile. NG-TEPHRA reveals that parameter sets with a non-homogeneous structure (one with dependencies amongst parameters) are a potential source of inefficiencies in the execution time of a complete experiment. The addition of a discard mechanism to Nimrod/G provided a significant performance gain. Also, a poor design of the data flow can also be responsible for huge loss in performance.

A. Analysis and Visualization

Ongoing research is being performed in the area of visualization and analysis for this particular research problem. Visualization and analysis can be performed by regarding the database as a lattice that can be segmented according to some metric of similarity [2], [14], [18]. The interest in finding similar cases is at the heart of urban planning and emergency mitigation, and common relations among cases can be used to visualize general scenarios in a GIS more succinctly. Needless to say, the implementation of this algorithm will be also a subject for high performance computing and large-scale distributed processing [22].

B. Map Generation

One of the goals is the generation of maps for urban planning and disaster management that can be utilized by regional authorities, city planners and non-technical users. The methodology for producing maps for NG-TEPHRA will be based on that described in [17]. Determination of VEI eruption levels from historical and tephrochronological records will provide input data for the simulation of each eruption type with suitable average environmental parameters. Then, the consolidation of a database of points and ash accumulations allows to generate a GIS surface map containing isolines describing mass and probability accumulation in the region of interest. Once mass and probability has been finally given a location, it is possible to construct maps that are easy to interpret and use for decision-makers with the goal of diminishing material losses facing potential disasters.

Re-classification of data according to ash height (extreme, high, moderate, low) and hazard metrics (extreme, high, moderate, low) will be overlaid on top of political and administrative maps for urban planning purposes and on top of human activity distribution maps for emergency mitigation plans. The symbology in the maps has to be simple using colours from red to green, varying as a function of the level of hazard. The same methodology can be applied to the case of the largest possible event, being reconstructed by studying

paleo-eruptions in order to construct a scenario that anticipate the worst possible outcome. Thus, decision-makers are able to implement any necessary prevention and mitigation actions.

VIII. ACKNOWLEDGMENTS

The authors wish to acknowledge the valuable contribution of several organizations, provided in multiple ways. The Citizen's Network of Meteorological Stations (spanish: *Red Ciudadana de Estaciones Meteorológicas*) provided direct financial support for development and research. UNA provided research funds, access to facilities and cluster infrastructure for debugging purposes. ITCR has provided continuous support for the NG-TEPHRA and other e-Science initiatives. This work has been supported by the Australian Research Council (ARC) under the Discovery grant scheme and by the use of the ARC funded Enterprisegrid computing resource. Thanks go to our colleagues in the MeSSAGE Lab, in particular for providing expertise and access to cluster and Cloud resources, as well as an excellent collaboration disposition. The Pacific Rim Applications and Grid Middleware Assembly (PRAGMA) has provided the international collaboration environment under which this relationship has been established, fostered and maintained. In particular Peter Arzberger, Cindy Zheng and Phil Papadopolous from the PRAGMA community and the California Institute of Telecommunications and Information Technologies have been a source of constant encouragement.

REFERENCES

- [1] ABRAMSON, D., GIDDY, J., AND KOTLER, L. High performance parametric modeling with nimrod/g: Killer application for the global grid? *International Parallel and Distributed Processing Symposium (IPDPS)*, Cancun, Mexico (May 2000), 520–528.
- [2] AHARONOV, D., AND REGEV, O. Lattice problems in $np \cap co-np$. *J. ACM* 52, 5 (2005), 749–765.
- [3] ALVARADO, G. Volcanology and petrology of irazu volcano, costa rica. Doctoral dissertation. University of Kiel, Germany, 261p, 1993.
- [4] ALVARADO L., F. W. Climatología de la atmósfera libre sobre costa rica. *Topics in Meteorology and Oceanography* 8, 2 (2001), 89–115s.
- [5] BARRANTES, G. Development of a geographical information system for hazard evaluation of tephra fall in the irazú volcano. MSc thesis dissertation. ITCR, Costa Rica, 2006.
- [6] BARRANTES, G., GARITA, C., NUNEZ, S., AND CASTRO, J. Aplicación de sistemas de información geográfica en la creación de escenarios de riesgo por caída de tefra. *Conferencia Latinoamericana de Computación e Informática* (2007).
- [7] BAXTER, P. *Encyclopedia of Volcanoes*. Academic Press, Expected: August 2010, ch. Impact of Eruptions on Human Health.
- [8] BERNABEU, M. O., CORRIAS, A., PITT-FRANCIS, J., RODRIGUEZ, B., BETHWAITE, B., ENTICOTT, C., GARIC, S., PEACHEY, T., TAN, J., ABRAMSON, D., AND GAVAGHAN, D. B. Grid computing simulations of ion channel block effects on the ecg using 3d anatomically-based models. *Computers and Cardiology, Park City, Utah* (September 13-16, 2009).
- [9] BETHWAITE, B., ABRAMSON, D., BOHNERT, F., GARIC, S., ENTICOTT, C., AND PEACHEY, T. *Cloud Computing: Principles, Systems and Applications*. Springer, Expected: August 2010, ch. Mixing Grids and Clouds: High-throughput Science using the Nimrod Tool Family.
- [10] BONADONNA, C., CONNOR, C., HOUGHTON, B., CONNOR, L., BYRNE, M. AND LAING, A., AND HINCKS, T. Estimation of volcanic hazards from tephra fallout. *Journal of Geophysical Research* 110, B03203 (2005).
- [11] CONNOR CH., E. A. Estimation of volcanic hazards from tephra fallout. *Natural Hazards Review* 2, 1 (2001), 33–42.
- [12] HARRIS, S. *Encyclopedia of Volcanoes*. Academic Press, Expected: August 2010, ch. Archaeology and Volcanism.
- [13] HEIKEN, G., MURPHU, M., HACKETT, W., AND SCOTT, W. *Hazards and energy infrastructures*. United States. U.S Department of Energy Code EH-33. Office of Risk Analysis and Technology., UR 95-1087, 1995.
- [14] KHOT, S. Hardness of approximating the shortest vector problem in lattices. *J. ACM* 52, 5 (2005), 789–808.
- [15] LYNCH, A., ABRAMSON, D., GÖRGEN, K., BERINGER, J., AND UOTILA, P. Influence of savanna fire on australian monsoon season precipitation and circulation as simulated using a distributed computing environment. *Geophys. Res. Lett.* 34, L20801 (Oct. 2007).
- [16] MCNUTT, S., RYMER, H., AND STIX, J. *Encyclopedia of Volcanoes*. Academic Press, Expected: August 2010, ch. Synthesis of Volcano Monitoring.
- [17] MET-ALARN, P. Erupciones volcánicas: Mapas de amenazas, recomendaciones técnicas para su elaboración. INETER-COSUDE, Managua, Nicaragua, 2005.
- [18] MICCIANCIO, D. The shortest vector problem is np-hard to approximate to within some constant. *SIAM Journal on Computing* 30, 6 (2001), 2008–2035.
- [19] MILLER, T., AND CASADEVAL, T. *Encyclopedia of Volcanoes*. Academic Press, Expected: August 2010, ch. Earth's Volcanoes and Eruptions: An Overview.
- [20] NEWHALL, C. *Encyclopedia of Volcanoes*. Academic Press, Expected: August 2010, ch. Volcano Warnings.
- [21] NÚÑEZ, S., BARRANTES, G., MALAVASSI, E., AND BRENES, J. Ng-tephra: Enabling large-scale volcanic hazard simulations in the pragma grid environment. *18th PRAGMA Workshop. San Diego, CA, USA* (March 3-4 2010).
- [22] NÚÑEZ, S., BARRANTES, G., MALAVASSI, E., BRENES, J., AND CASTRO, J. On the structure of parametric sweeps in large-scale, grid-enabled ash deposition simulations for the irazu volcano in the pragma grid. *Workshop on Enabling Technology for e-Science, FCST 2010 (accepted)* (August 18-22, 2010).
- [23] SCHMIDBERGER, J., BETHWAITE, B., ENTICOTT, C., BATE, M., ANDROULAKIS, S., FAUX, N., REBOUL, C., PHAN, J., WHISSTOCK, J., GARIC, S., GOSCINSKI, W., ABRAMSON, D., AND BUCKLE, A. High-throughput protein structure determination using grid computing. *IEEE Workshop on High Performance Computational Biology, Rome, Italy, co-located with IPDPS 2009* (May 25, 2009).
- [24] SIMKIN, T., AND SIEBERT, L. *Encyclopedia of Volcanoes*. Academic Press, Expected: August 2010, ch. Earth's Volcanoes and Eruptions: An Overview.
- [25] SMITH, K. *Environmental hazards: assessing risk and reducing disaster*. Routledge, London, UK, 1992.
- [26] SUZUKI, T. *Arc Volcanism: Physics and Tectonics*. Terra Scientific Publishing, Tokyo, Japan, 1995, ch. A theoretical model for dispersion of tephra, p. 190.
- [27] TEXTOR, C., GRAF, H., LONGO, A., PAPALE, P., TIMMRECK, C., AND ERNST, G. Numerical simulations of explosive volcanic eruptions from the conduit flow to global atmospheric scales. *Annals of Geophysics* 48, 4/5 (2005).
- [28] TOMLINSON, R. *Thinking about GIS*. ESRI Press, California, USA, 2003.
- [29] WEBLEY, P., DEAN, K., BAILEY, J., DEHN, J., AND PETERSON, R. Automated forecasting of volcanic ash dispersion utilizing virtual globes. *Natural Hazards, Springer Netherlands* 51, 2 (2008).