Self-healing of workflow activity incidents on distributed computing infrastructures

Rafael Ferreira da Silva[a], Tristan Glatard[a], Frédéric Desprez[b]

[a] *University of Lyon, CNRS, INSERM, CREATIS, Villeurbanne, France*
[b] *INRIA, University of Lyon, LIP, ENS Lyon, Lyon, France*

Abstract

Distributed computing infrastructures are commonly used through scientific gateways, but operating these gateways requires important human intervention to handle operational incidents. This paper presents a self-healing process that quantifies incident degrees of workflow activities from metrics measuring long-tail effect, application efficiency, data transfer issues, and site-specific problems. These metrics are simple enough to be computed online and they make little assumptions on the application or resource characteristics. From their degree, incidents are classified in levels and associated to sets of healing actions that are selected based on association rules modeling correlations between incident levels. We specifically study the long-tail effect issue, and propose a new algorithm to control task replication. The healing process is parametrized on real application traces acquired in production on the European Grid Infrastructure. Experimental results obtained in the Virtual Imaging Platform show that the proposed method speeds up execution up to a factor of 4, consumes up to 26% less resource time than a control execution and properly detects unrecoverable errors.

Keywords: error detection and handling, workflow execution, production distributed systems

1. Introduction

Distributed computing infrastructures (DCI) are becoming daily instruments of scientific research, in particular through scientific gateways [1] developed to allow scientists to transparently run their analyses on large sets of computing resources. While these platforms provide important amounts of resources in an almost seamless way, their large scale and the number of middleware systems involved lead to many errors and faults. Easy-to-use interfaces provided by these gateways exacerbate the need for properly solving operational incidents encountered on DCIs since end users expect high reliability and performance with no extra monitoring or parametrization from their side. In practice, such services are often backed by substantial support staff who monitors running experiments by performing simple yet crucial actions such as rescheduling tasks, restarting

Email addresses: rafael.silva@creatis.insa-lyon.fr (Rafael Ferreira da Silva), glatard@creatis.insa-lyon.fr (Tristan Glatard), Frederic.Desprez@inria.fr (Frédéric Desprez)

services, killing misbehaving runs or replicating data files to reliable storage facilities. Fair QoS can then be delivered, yet with important human intervention.

For instance, the long-tail effect [2] is a common frustration for users who have to wait for a long time to retrieve the last few pieces of their computations. Operators may be able to address it by rescheduling tasks that are considered late (e.g. due to execution on a slow machine, low network throughput or just loss of contact) but detection is very time consuming and still approximate.

Automating such operations is challenging for two reasons. First, the problem is online by nature because no reliable user activity prediction can be assumed, and new workloads may arrive at any time. Therefore, the considered metrics, decisions and actions have to remain simple and to yield results while the application is still executing. Second, it is non-clairvoyant due to the lack of information about applications and resources in production conditions. Computing resources are usually dynamically provisioned from heterogeneous clusters, clouds or desktop grids without any reliable estimate of their availability and characteristics. Models of application execution times are hardly available either, in particular on heterogeneous computing resources.

A scientific gateway is considered here as a platform where users can process their own data with predefined applications workflows. Workflows are compositions of *activities* defined independently from the processed data and that only consist of a program description. At runtime, activities receive data and spawn *invocations* from their input parameter sets. Invocations are assumed independent from each other (bag of tasks) and executed on the DCI as single-core *tasks* which can be resubmitted in case of failures. This model fits several existing gateways such as e-bioinfra [3], P-Grade [4], and the Virtual Imaging Platform [5]. We also consider that files involved in workflow executions are accessed through a single file catalog but that storage is distributed. Files may be replicated to improve availability and reduce load on servers.

The gateway may take decisions on file replication, resource provisioning, and task scheduling on behalf of the user. Performance optimization is a target but the main point is to ensure that correctly-defined executions complete, that performance is acceptable, and that misbehaving runs (e.g. failures coming from user errors or unrecoverable infrastructure downtimes) are quickly detected and stopped before they consume too many resources.

Our ultimate goal is to reach a general model of such a scientific gateway that could autonomously detect and handle operational incidents. In this work, we propose a healing process for workflow activities only. Activities are modeled as Fuzzy Finite State Machines (FuSM) [6] where state degrees of membership are determined by an external healing process. Degrees of membership are computed from metrics assuming that incidents have outlier performance, e.g. a site or a particular invocation behaves differently than the others. Based on incident degrees, the healing process identifies incident levels using thresholds determined from platform history. A specific set of actions is then selected from association rules among incident levels. We specifically study the long-tail effect issue, and propose a new algorithm to control task replication.

Section 2 presents related work. Our approach is described in section 3 (general healing process), section 4 (metrics used to quantify incident degrees), section 5 (identification of incident levels), and section 6 (actions). Experimental results are presented in section 7 in production conditions.

2. Related Work

Managing systems with limited intervention of system administrators is the goal of autonomic computing [7], which has been used to address various problems related to self-healing, self-configuration, self-optimization, and self-protection of distributed systems. For instance, provisioning of virtual machines is studied by Nguyen et al. [8] and an approach to tackle service overload, queue starvation, "black hole" effect and job failures is sketched by Collet et al. [9].

An autonomic manager can be described as a so-called MAPE-K loop which consists of monitoring (M), analysis (A), planning (P), execution (E) and knowledge (K). Generic software frameworks were built to wrap legacy applications in such loops with limited intrusiveness. For instance, Broto et al. [10] demonstrate the wrapping of DIET grid services for autonomic deployment and configuration. We consider here that the target gateway can be instrumented to report appropriate events and to perform predefined actions.

Monitoring is broadly studied in distributed systems, both at coarse (traces, archives) and fine time scales (active monitoring, probing). Many workload archives are available. In particular, the grid observatory [11] has been collecting traces for a few years on several grids. However, as noted by Iosup and Epema [12], most existing traces remain at the task level and lack information about workflows and activities. Application patterns can be retrieved from logs (e.g. bag of tasks) but precise information about workflow activities is bound to be missing. Studies on task errors and their distributions are also available [13, 14], but they do not consider operational issues encountered by the gateways submitting these tasks. Besides, active monitoring using tools such as Nagios [15] cannot be the only monitoring source when substantial workloads are involved. Therefore, we rely on traces of the target gateway, as detailed in section 5. One issue in this case is to determine the timespan where system behavior can be considered steady-state. Although this issue was recently investigated [16], it remains difficult to identify non-stationarities in an online process and we adopt a stationary model here.

Analysis consists in computing metrics (a.k.a. utility functions) from monitoring data to characterize the state of the system. System state usually distinguishes two regimes: properly functioning and misfunctioning. Zhang et al. [17] assume that incidents lead to non-stationarity of the workload statistics and use the Page-Hinkely test to detect them. Stehle et al. [18] present a method where the convex hull is used instead of hyper-rectangles to classify system states. As described in section 5, we use multiple threshold values for a given metric to use more than two levels to characterize incidents.

Planning and actions considered in this work deal with task scheduling and file replication. Most scheduling approaches are clairvoyant, meaning that resource, task, error rate and workload characteristics are precisely known [19, 20]. The heuristics designed by Casanova et al. [21] for the case where only data transfer costs are known are an exception, on an offline problem though. Quintin and Wagner [22] also propose an online task scheduling algorithm where only some characteristics of the application DAG are known. Camarasu-Pop et al. [23] propose a non-clairvoyant load-balancing strategy to remove the long-tail effect in production heterogeneous systems, but it is limited to Monte-Carlo simulations.

The general task scheduling problem is out of our scope. We assume that a scheduler is already in place, and we only aim at performing actions when it does not deliver

expected performance. In particular, we focus on site blacklisting and on dynamic task replication [24] to avoid the long-tail effect.

Task replication, a.k.a. redundant requests is commonly used to address non-clairvoyant problems [2], but it should be used sparingly, to avoid overloading the middleware and degrading fairness among users [25]. In this work, task replication is considered only when activities are detected blocked according to the metric presented in section 4. An important aspect to be evaluated is the resource waste, a.k.a. the cost of task replication. Cirne et al. [2] evaluate the waste of resources by measuring the percentage of wasted cycles among all the cycles required to execute the application.

File replication strategies also often assume clairvoyance on the size of produced data, file access pattern and infrastructure parameters [26, 27]. In practice, production systems mostly remain limited to manual replication strategies [28].

3. General Healing Process

An activity is modeled as an FuSM with 13 states shown on Figure 1. The activity is initialized in `Submitting Invocations` where all the tasks are generated and submitted. Tasks consist of 4 successive phases: initialization, inputs download, application execution and output upload. They are all assumed independent, but with similar execution times (bag of tasks). `Running` is a state where no particular issue is detected; no action is taken and the activity is assumed to behave normally. `Completed` (resp. `Failed`) is a terminal state used when all the invocations are successfully completed (resp. at least one invocation failed). These 4 states are crisp (not fuzzy) and exclusive. Their degree can only be 0 or 1 and if 1 then all the other states have a degree of 0. The 9 other states are fuzzy states corresponding to detected incidents.

The healing process sets the degree of FuSM states from incident detection metrics and invocation statuses. Then, it determines actions to address the incidents. If no action is required then the process waits until an event occurs (task status change) or a timeout is reached.

Let $I = \{x_i, i = 1, \ldots, n\}$ be the set of possible incidents (9 in this work) and $\eta = (\eta_1, \ldots, \eta_n) \in [0, 1]^n$ their degrees in the FuSM. Incident x_i can occur at m_i different levels $\{x_{i,j}, j = 1, \ldots, m_i\}$ delimited by threshold values $\tau_i = \{\tau_{i,j}, j = 1, \ldots, m_i\}$. The level of incident i is determined by j such that $\tau_{i,j} \leq \eta_i < \tau_{i,j+1}$. A set of actions $a_i(j)$ is available to address $x_{i,j}$:

$$\begin{aligned} a_i : [1, m_i] &\to \wp(A) \\ j &\mapsto a_i(j) \end{aligned} \tag{1}$$

where A is the set of possible actions taken by the healing process and $\wp(A)$ is the power set of A.

In addition to the incidents themselves, incident causes are taken into account. Association rules [29] are used to identify relations between levels of different incidents. Association rules to $x_{i,j}$ are defined as $R_{i,j} = \{r_{i,j}^{u,v} = (x_{u,v}, x_{i,j}, \rho_{i,j}^{u,v})\}$. Rule $r_{i,j}^{u,v}$ means that when $x_{u,v}$ happens then $x_{i,j}$ also happens with confidence $\rho_{i,j}^{u,v} \in [0, 1]$. The confidence of a rule is an estimate of probability $P(x_{i,j}|x_{u,v})$. Note that $r_{i,j}^{i,j} \in R_{i,j}$ and $\rho_{i,j}^{i,j} = 1$. We also define $R = \bigcup_{i \in [\![1,n]\!], j \in [\![1,m_i]\!]} R_{i,j}$.

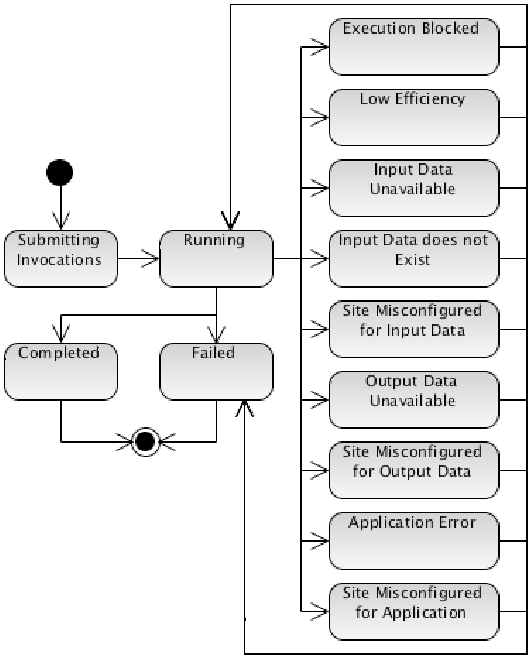

Figure 1: Fuzzy Finite State Machine (FuSM) representing an activity.

Figure 2 presents the algorithm used at each iteration of the healing process. Incident degrees are determined based on metrics presented in section 4 and incident levels j are obtained from historical data as explained in section 5. A roulette wheel selection [30] based on η is performed to select $x_{i,j}$ the incident level of interest for the iteration. In a roulette wheel selection, incident x_i is selected with a probability p_i proportional to its degree: $p(x_i) = \eta_i / \sum_{j=1}^{n} \eta_j$. A potential cause $x_{u,v}$ for incident $x_{i,j}$ is then selected from another roulette wheel selection on the association rules $r_{i,j}^{u,v}$, where x_u is at level v. Rule $r_{i,j}^{u,v}$ is weighted $\eta_u \times \rho_{i,j}^{u,v}$ in the roulette selection. Only first-order causes are considered here but the approach could be extended to include more recursion levels. Note that $r_{i,j}^{i,j}$ participates in this selection so that a first-order cause is not systematically chosen. Finally, actions in $a_u(v)$ are performed.

Input: invocation statuses and history of η
Output: set of actions a

01. wait for event or timeout
02. determine incident degrees η based on metrics
03. determine incident levels j such that $\tau_{i,j} \leq \eta_i < \tau_{i,j+1}$
04. select incident x_i by roulette wheel selection based on η
05. select rule $r_{u,v} = (x_{u,v}, x_{i,j}, \rho_{i,j}^{u,v}) \in R_{i,j}$ by roulette wheel selection based on $\eta_u \times \rho_{i,j}^{u,v}$, where x_u is at level v
06. $a = a_u(v)$
07. perform actions in a

Figure 2: One iteration of the healing process.

Table 1 illustrates this mechanism on an example case where only 3 incidents are considered, and Figure 3 shows it as a MAPE-K loop.

Step 02 and 03: incident degrees and levels are determined:

x_i: incident name	Degree η_i	Level j
x_1: activity blocked	0.8	2
x_2: low efficiency	0.1	1
x_3: input data does not exist	0.4	1

Step 04: $x_{1,2}$ is selected with probability $\frac{0.8}{0.8+0.4+0.1}$.

Step 05: association rules $r_{1,2}^{2,1}$, $r_{1,2}^{3,1}$ and $r_{1,2}^{1,2}$ are considered:

Rule	Confidence
$r_{1,2}^{2,1}$: $x_{2,1} \rightarrow x_{1,2}$	0.8
$r_{1,2}^{3,1}$: $x_{3,1} \rightarrow x_{1,2}$	0.2
$r_{1,2}^{1,2}$: $x_{1,2} \rightarrow x_{1,2}$	1

$r_{1,2}^{2,1}$ is chosen with probability $\frac{0.8\times0.4}{0.8\times0.4+0.2\times0.1+0.8\times1}$.

Step 06: actions in $a_2(1)$ are performed.

Table 1: Example case.

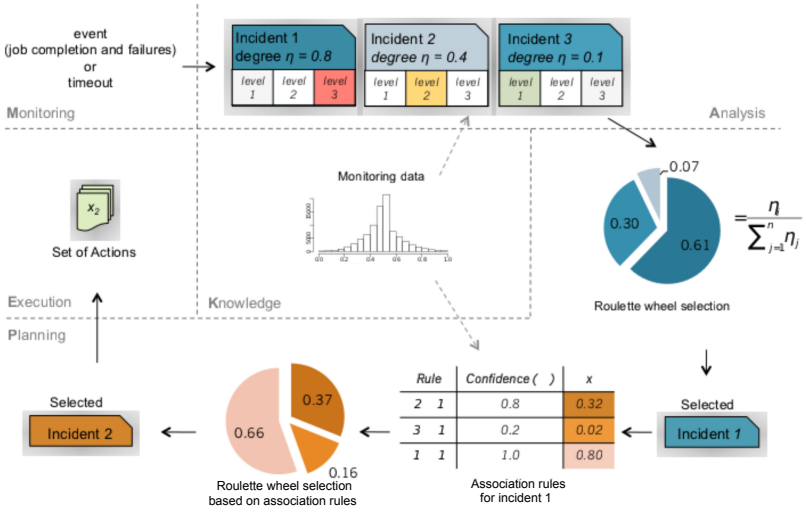

Figure 3: Example case showed as a MAPE-K loop.

4. Incident Degrees

This section describes the metrics used to determine the degree of the 9 considered incidents identified by human operators (step 02 on Figure 2).

4.1. Activity Blocked

This incident happens when an invocation is considered late compared to the others. It is responsible for many operational issues, leading to substantial speed-up reductions.

For instance, it occurs when one invocation of the activity requires more CPU cycles or when the invocation faces longer waiting times, lost tasks or executes on resources with poorer performance. We define the incident degree η_b of an activity from the max of the performance coefficients p_i of its n tasks, which relate the task phase durations (`setup`, `inputs download`, `application execution` and `outputs upload`) to their medians:

$$\eta_{\mathrm{b}} = 2.\max\left\{p_i = p(t_i, \tilde{t}) = \frac{t_i}{\tilde{t} + t_i}, i \in [1, n]\right\} - 1 \tag{2}$$

where $t_i = t_{i\_setup} + t_{i\_input} + t_{i\_exec} + t_{i\_output}$ is the estimated duration of task i and $\tilde{t} = \tilde{t}_{setup} + \tilde{t}_{input} + \tilde{t}_{exec} + \tilde{t}_{output}$ is the sum of the median durations of tasks 1 to n. Note that $\max\{p_i, i \in [1, n]\} \in [0.5, 1]$ so that $\eta_b \in [0, 1]$. Moreover, $\lim_{t_i \to +\infty} p_i = 1$ and $\max\{p_i, i \in [1, n]\} = 0.5$ when all the tasks behave like the median.

The estimated duration t_i of a task is computed phase by phase, as follows: (i) for completed task phases, the actual consumed resource time is used; (ii) for ongoing task phases, the maximum value between the current consumed resource time and the median consumed time is taken; and (iii) for unstarted task phases, the time slot is filled by the median value. Figure 4 illustrates the estimation process of a task where the actual durations are used for the two first completed phases (42s for `setup` and 300s for `inputs download`), the `application execution` phase uses the maximum value between the current value of 20s and the median value of 400s, and the last phase (`outputs upload`) is filled by the median value of 15s, as it is not started yet.

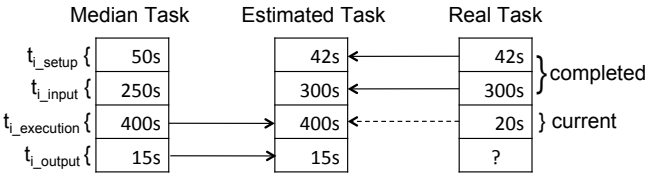

Figure 4: Task estimation based on median values.

4.2. Low Efficiency

This happens when the time spent by all the activity invocations in data transfers dominates CPU time. It may be due to sites with poor network connectivity or be intrinsic to the application. The incident degree is defined from the ratio between the cumulative CPU time C_i consumed by all completed invocations and the cumulative execution time of all completed invocations:

$$\eta_{\mathrm{e}} = 1 - \frac{\sum_{i=1}^{n(t)} C_i}{\sum_{i=1}^{n(t)} (C_i + D_i)}$$

where D_i is the time spent by invocation i in data transfers.

4.3. Input Data Unavailable

This happens when a file is registered in the file catalog but the storage resource(s) is(are) unavailable or unreachable. The incident degree η_{iu} in this state is determined from the input transfer failure rate due to data unavailability. Transfers of completed, failed, and running invocations are considered.

4.4. Input Data does not Exist

This happens when an incorrect data path was specified, the file was removed by mistake or the file catalog is unavailable or unreachable. Again, the incident degree η_{ie} is directly determined by the input transfer failure rate due to non-existent data. Non-existent file is distinguished from file unavailability using ad-hoc parsing of standard error files. Transfers of completed, failed, and running invocations are considered.

4.5. Site Misconfigured for Input Data

This incident happens when sites have utmost input data transfer failure rate. The incident degree η_{is} is measured as follows:

$$\eta_{\text{is}} = \max(\phi_1, \phi_2, \ldots, \phi_k) - \text{median}(\phi_1, \phi_2, \ldots, \phi_k)$$

where ϕ_i denotes the input transfer failure ratio (including both input data unavailable and input data does not exist) on site i and k is the number of white-listed sites used by the activity. The difference between the maximum rate and the median ensures that the incident degree has high values only when some sites are misconfigured. This metric is correlated but not redundant with the two previous ones. If some input data file is not available due to site-independent issues with the storage system, then η_{iu} will grow but η_{is} will remain low because all sites fail identically. On the contrary, η_{is} may grow while η_{iu} and η_{ie} remain low.

4.6. Output Data Unavailable

Output data can also be unavailable. Unavailability happens due to three main reasons: the user did not specify the output path correctly, the application did not produce the expected data, or the file catalog or storage resource are unavailable or unreachable. The incident degree η_{ou} is determined by the output transfer failure rate. Transfers of completed, failed and running invocations are considered.

4.7. Site Misconfigured for Output Data

The incident degree η_{os} in this incident is determined as follows:

$$\eta_{\text{os}} = \max(\psi_1, \psi_2, \ldots, \psi_k) - \text{median}(\psi_1, \psi_2, \ldots, \psi_k)$$

where ψ_i denotes the output transfer failure ratio on site i and k is the number of white-listed sites used by the activity.

4.8. Application Error

Applications can fail due to a variety of reasons among which: the application executable is corrupted, dependencies are missing, or the executable is not compatible with the execution host. The incident degree η_{a} in this state is measured by the task failure rate due to application errors. Completed, failed, and running tasks are considered.

4.9. Site Misconfigured for Application

The incident degree η_{as} in this state is measured as follows:

$$\eta_{\text{as}} = \max(\alpha_1, \alpha_2, \ldots, \alpha_k) - \text{median}(\alpha_1, \alpha_2, \ldots, \alpha_k)$$

where α_i denotes the task failure rate due to application errors on site i and k is the number of white-listed sites used by the activity.

5. Incident Levels and Association Rules

Incident degrees η_i are quantified in discrete incident levels so that different sets of actions can be used to address different levels of the incident. The number and values of the thresholds are determined from observed distributions of η_i. The number m_i of incident levels associated to incident i is set as the number of modes in the observed distribution of η_i. Thresholds $\tau_{i,j}$ are determined from mode clustering. Incidents levels and thresholds are determined offline; thus they do not create any overhead on the workflow execution.

5.1. Training Dataset

Distributions of incident degrees were determined from the science-gateway workload archive [31] available in the grid observatory[1]. These traces were collected from the Virtual Imaging Platform [5] between April and August 2011. Applications deployed on this platform are described as workflows executed using the MOTEUR workflow engine [32]. Resource provisioning and task scheduling is provided by DIRAC [33] using so-called "pilot jobs". Resources are provisioned online with no advance reservations. Tasks are executed on the biomed virtual organization (VO) of the European Grid Infrastructure (EGI)[2] which, as of January 2013, has access to some 90 computing sites of 22 countries, offering 190 batch queues and approximately 4 PB of disk space. Table 2 shows the distribution of sites per country supporting the biomed VO.

This dataset contains $1,082$ executions of 36 different workflows executed by 26 users. Workflow executions contain $1,838$ activity instances, corresponding to $92,309$ invocations and $123,025$ tasks (including resubmissions).

Figure 5 shows the cumulative amount of running activities along this period. It shows that the workload is quite uniformly distributed although a slight increase is observed in June.

5.2. Incident Levels

We replayed the events found in this dataset to compute incident degree values after each event (total of $641,297$ events). Figure 6 displays histograms of computed incident degrees. For readability purposes, only $\eta_i \neq 0$ values are represented. Most of the histograms appear multi-modal, which confirms that incident degrees are quantified. Level numbers and threshold values τ are set from visual mode detection in these histograms and reported on Table 3 with associated actions.

[1] `http://www.grid-observatory.org`

[2] `http://www.egi.eu`

Country	Number of sites	Number of batch queues
UK	13	50
Italy	12	30
France	12	31
Greece	9	11
Spain	5	7
Germany	5	14
Portugal	4	7
Turkey	3	3
Poland	3	4
Netherlands	3	12
Croatia	3	6
Bulgaria	3	3
FYROM	2	2
Brazil	2	3
Vietnam	1	1
Slovakia	1	1
Russia	1	2
Other (.org)	1	1
Moldova	1	1
Mexico	1	1
Cyprus	1	1
China	1	1

Table 2: Distribution of sites and batch queues per country in the biomed VO (January 2013).

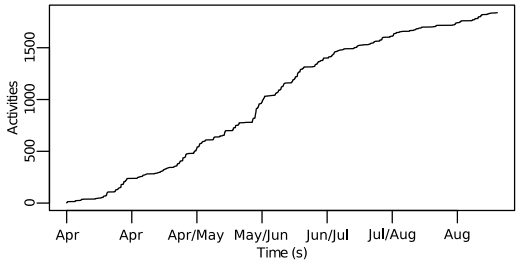

Figure 5: Cumulative amount of running activities from April to August 2011.

Incidents at level 1 are considered painless for the execution and they do not trigger any action. Other levels can lead to radical (completely stop the activity or blacklist a site) or intermediate actions (task or file replication).

The use of historical information to determine the threshold value put on η_b reduces the impact of the assumption that all tasks of a given workflow activity will have the same duration. Indeed, the threshold value quantifies what is an acceptable deviation of the task duration from its median value.

5.3. Association Rules

Association rules are computed based on the frequency of occurrences of two incident levels in the training dataset. The confidence $\rho_{i,j}^{u,v}$ of a rule $x_{u,v} \Rightarrow x_{i,j}$ measures the probability that an incident level $x_{i,j}$ happens when $x_{u,v}$ occurs. Table 4 shows rule samples extracted from the training dataset and ordered by decreasing confidence. The set of rules leading to activity blocked ($x_{1,2}$) and low efficiency ($x_{2,2}$) incidents shows

that they are partially dependent on other "cause" incidents, which is considered by the self-healing process.

At the bottom of the table we find rules with null confidence. These are consistent with common-sense interpretation of the incident dependencies (e.g. no site-specific issue when input data is unavailable).

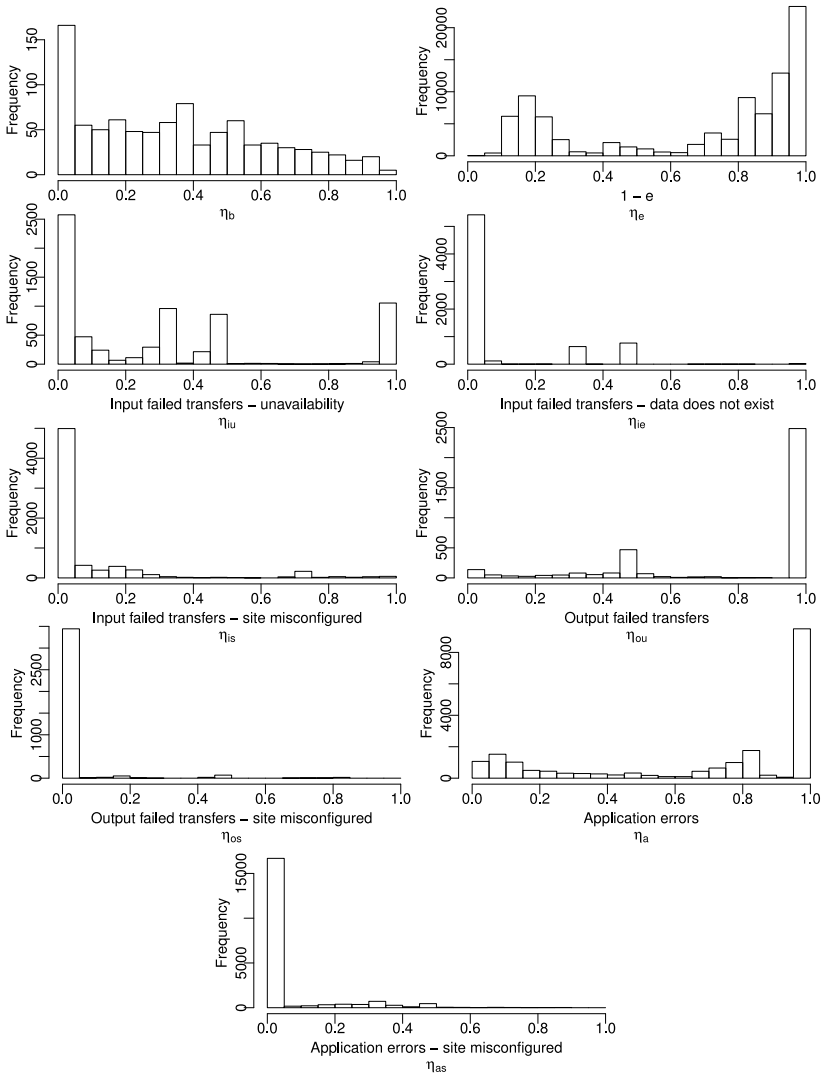

Figure 6: Histograms of incident degrees sampled in bins of 5%.

Incident (x_i)	Number of incident levels (m_i)	Level 1		Level 2		Level 3	
		$\tau_{i,1}$	actions	$\tau_{i,2}$	actions	$\tau_{i,3}$	actions
x_1: activity blocked	2	0	$\emptyset$	0.7	replicate tasks with $p_i > \tau$		
x_2: low efficiency	2	0	$\emptyset$	0.6	replicate input files replicate tasks with $p_i > \tau$		
x_3: input data unavailable	3	0	$\emptyset$	0.2	replicate input files	0.8	stop activity
x_4: input data does not exist	2	0	$\emptyset$	0.8	stop activity		
x_5: site misconfigured for input data	3	0	$\emptyset$	0.3	replicate files on sites reachable from problematic site	0.65	blacklist site
x_6: output data unavailable	2	0	$\emptyset$	0.8	stop activity		
x_7: site misconfigured for output data	2	0	$\emptyset$	0.1	blacklist site		
x_8: application error	2	0	$\emptyset$	0.5	stop activity		
x_9: site misconfigured for application	2	0	$\emptyset$	0.1	blacklist site		

Table 3: Incident levels and actions.

Association rule	$\rho_{i,j}^{u,v}$
$x_{5,2} \Rightarrow x_{2,2}$	0.3809
$x_{7,2} \Rightarrow x_{1,2}$	0.3529
$x_{5,3} \Rightarrow x_{1,2}$	0.3333
$x_{1,2} \Rightarrow x_{2,2}$	0.3059
$x_{3,2} \Rightarrow x_{1,2}$	0.2975
$x_{7,2} \Rightarrow x_{2,2}$	0.2941
$x_{5,2} \Rightarrow x_{1,2}$	0.2608
$x_{9,2} \Rightarrow x_{1,2}$	0.2435
$x_{2,2} \Rightarrow x_{1,2}$	0.2383
...	...
$x_{3,2} \Rightarrow x_{2,2}$	0.1276
$x_{7,2} \Rightarrow x_{3,3}$	0.1250
$x_{3,3} \Rightarrow x_{9,2}$	0.1228
$x_{7,2} \Rightarrow x_{3,2}$	0.0625
...	...
$x_{3,3} \Rightarrow x_{5,2}$	0.0000
$x_{3,3} \Rightarrow x_{5,3}$	0.0000
$x_{4,2} \Rightarrow x_{5,2}$	0.0000
$x_{4,2} \Rightarrow x_{5,3}$	0.0000
$x_{5,2} \Rightarrow x_{3,3}$	0.0000
$x_{5,2} \Rightarrow x_{4,2}$	0.0000
$x_{5,3} \Rightarrow x_{3,3}$	0.0000
$x_{5,3} \Rightarrow x_{4,2}$	0.0000

Table 4: Confidence of rules between incident levels.

6. Actions

Four actions are performed by the self-healing process: task replication, file replication, site blacklisting and activity stop. The first three are described below.

6.1. Task replication

Blocked activities and activities of low efficiency are addressed by task replication. To limit resource waste, the replication process for a particular task is controlled by two mechanisms. First, a task is not replicated if a replica is already queued. Second, if replica j has better performance than replica r (i.e. $p(t_r, t_j) > \tau$, see equation 2) and j is in a more advanced phase than r, then replica r is aborted. Figure 7 presents the algorithm of the replication process. It is applied to all tasks with $p_i > \tau$, as defined on equation 2.

6.2. File Replication

File replication is implemented differently depending on the incident. In case of input data unavailability, a file is replicated to a storage resource selected randomly. The maximal allowed number of file replicas is set to 5. In case a site is misconfigured, replication to the site local storage resource is first attempted. This aims at circumventing inter-domain connectivity issues. If there is no local storage available or the replication process fails, then a second attempt is performed to a storage resource successfully accessed by other tasks executed on the same site. Otherwise, a storage resource is randomly selected. Fig. 8 depicts this process.

Input: Set of replicas R of a task i

01. rep = true
02. **for** $r \in R$ **do**
03. **for** $j \in R$, $j \neq r$ **do**
04. **if** $p(t_r, t_j) > \tau$ **and** j is a step further than r **then**
05. abort r
06. **done**
07. **if** r is started **and** $p(t_r, \tilde{t}) \leq \tau$ **then**
08. rep = false
09. **else if** r is queued **then**
10. rep = false
11. **done**
12. **if** rep == true **then**
13. replicate r

Figure 7: Replication process for one task.

Inputs: File f, set of storage resources S, set of completed tasks on the same site T

01. replicate f to local storage resource
02. **if** replication *not* successful **then**
03. select storage $s_i \in S$ where $t \in T$ could access s
04. replicate f to s_i
05. **if** replication *not* successful **then**
06. select randomly $s_r \in S$
07. replicate f to s_r
08. **done**
09. **done**

Figure 8: Site misconfigured: replication process for one file.

6.3. Site Blacklisting

Problematic sites are only temporarily blacklisted during a time interval set from exponential back-off. The site is first blacklisted for 1 minute only and then put back on the white list. In case it is detected misconfigured again, then the blacklist duration is increased to 2 minutes, then to 4 minutes, 16 minutes, etc.

7. Experiments

The healing process is implemented in the Virtual Imaging Platform (see description in section 5.1) and deployed in production. The experiments presented hereafter, conducted for two real workflow activities, evaluate the ability of the healing process to (i) improve workflow makespan by replicating tasks of blocked activities (*Experiment 1*) and (ii) quickly identify and report critical issues (*Experiment 2*). Another experiment, evaluating the handling of low efficiency, site misconfiguration, and input data unavailability, was reported in [34].

7.1. Experiment conditions and metrics

Experiment 1 aims at testing that blocked activities are properly detected and handled; the other incidents are ignored. This experiment uses a correct execution where all the input files exist and the application is supposed to run properly and produce the expected results. Five repetitions are performed for each workflow activity.

Experiment 2 aims at testing that unrecoverable errors are quickly identified and the execution is stopped. Unrecoverable errors are intentionally injected in 3 different runs: in run `non-existent inputs`, non-existent file paths are used for all the invocations; in `application-error`, all the file paths exist but input files are corrupted; and in `non-existent output`, input files are correct but the application does not produce the expected results.

Two workflow activities are considered for each experiment. `FIELD-II/pasa` consists of 122 invocations of an ultrasonic simulator on an echocardiography 2D dataset. It is a data-intensive activity where invocations use from a few seconds to some 15 minutes of CPU time; it transfers 208 MB of input data and outputs about 40 KB of data. `Mean-Shift/hs3` has 250 CPU-intensive invocations of an image filtering application. Invocation CPU time ranges from a few minutes up to one hour; input data size is 182 MB and output is less than 1 KB. Files are replicated on two storage sites for both activities.

For each experiment, a workflow execution using our method (`Self-Healing`) is compared to a control execution (`No-Healing`). Executions are launched on the biomed VO of the EGI, in production conditions, i.e., without any control of the number of available resources and reliability. `Self-Healing` and `No-Healing` are both launched simultaneously to ensure similar grid conditions. The DIRAC scheduler is configured to equally distribute resources among executions.

The FuSM and healing process are implemented in the MOTEUR workflow engine. The timeout value in the healing process is computed dynamically as the median of the task inter-completion delays in the current execution. Task replication is performed by resubmitting running tasks to DIRAC. To avoid concurrency issues in the writing of output files, a simple mechanism based on file renaming is implemented. To limit infrastructure overload, running tasks are replicated up to 5 times only. MOTEUR is configured to resubmit failed tasks up to 5 times in all runs of both experiments. We use DIRAC v5r12p9 and MOTEUR 0.9.19.

The waste metric used by Cirne et al. [2] does not fit our context because it cannot provide an effective estimation of the amount of resource wasted by self-healing simulations when compared to the control ones. Here, resource waste is assessed by the amount of resource time consumed by the simulations performing the healing process related to the amount of resource time consumed by control simulations. We use the `waste coefficient` (w), defined as follows:

$$w = \frac{\sum_{i=1}^{n} h_i + \sum_{j=1}^{m} r_j}{\sum_{i=1}^{n} c_i} - 1$$

where h_i and c_i are the resource time consumed (CPU time + data transfers time) by n completed tasks for `Self-Healing` and `No-Healing` simulations respectively, and r_i is the resource time consumed by m unused replicas. Note that task replication usually leads to $h_i \leq c_i$. If $w > 0$, the healing approach wastes resources compared to the control. If $w < 0$, then the healing approach consumes less resources than the control, which can happen when faster resources are selected.

7.2. Results and Discussion

Experiment 1. Figure 9 shows the makespan of `FIELD-II/pasa` and `Mean-Shift/hs3` for the 5 repetitions. The makespan is considerably reduced in all repetitions of both activ-

ities. Speed-up values yielded by `Self-Healing` range from 2.6 to 4 for `FIELD-II/pasa` and from 1.3 to 2.6 for `Mean-Shift/hs3`.

Figures 10 and 11 present a cumulative density function (CDF) of the number of completed tasks for `FIELD-II/pasa` and `Mean-Shift/hs3`, respectively. In most cases completion curves of both `Self-Healing` and `No-Healing` executions are similar up to 95%. This confirms that both executions are executed in similar grid conditions. In some cases (e.g. `Repetition 2` in Figure 11) `Self-Healing` execution even presents lower performance than `No-Healing` execution but it is compensated by the long-tail effect produced by the latter.

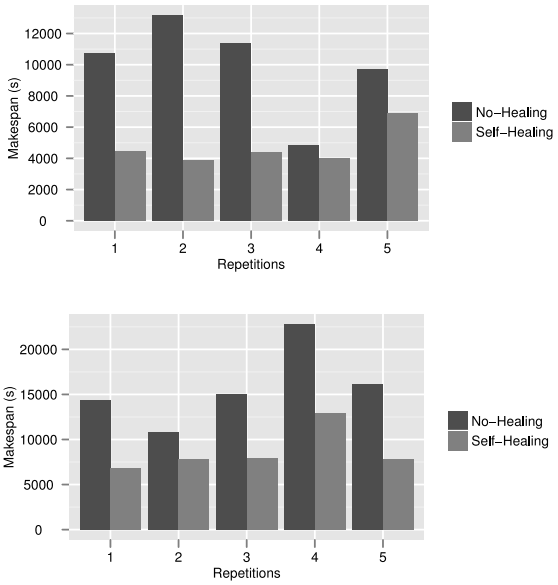

Figure 9: Execution makespan for `FIELD-II/pasa` (top) and `Mean-Shift/hs3` (bottom).

Tables 5 and 6 show the waste coefficient values for the 5 repetitions for `FIELD-II/pasa` and `Mean-Shift/hs3` respectively. The `Self-Healing` process reduces resource consumption up to 26% when compared to the control execution. This happens because replication increases the probability to select a faster resource. The total number of replicated tasks for all repetitions is 172 for `FIELD-II/pasa` (i.e. 0.28 task replication per invocation in average) and 308 for `Mean-Shift/hs3` (i.e. 0.24 task replication per invocation in average).

Repetition	h	r	c	w
1	56,159s	2,203s	64,163s	−0.10
2	60,991s	6,383s	79,031s	−0.15
3	60,473s	10,818s	77,851s	−0.09
4	42,475s	1,420s	41,528s	0.05
5	56,726s	4,527s	82,555s	−0.26

Table 5: Waste coefficient values for `FIELD-II/pasa`.

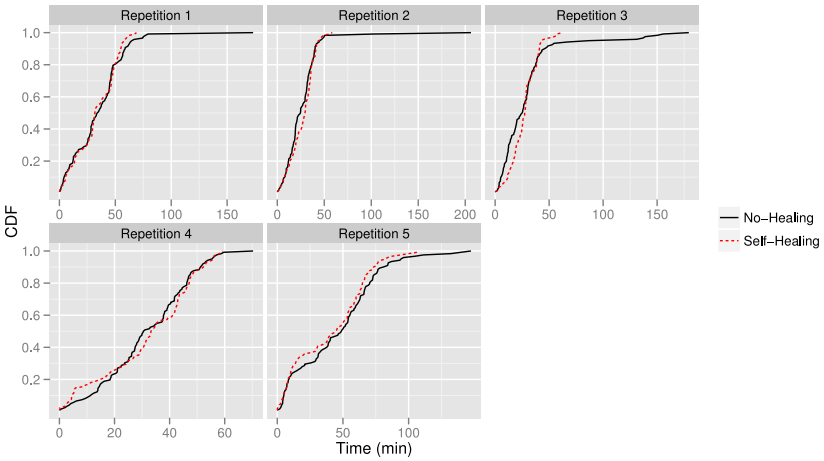

Figure 10: Experiment 1: CDF of the number of completed tasks for `FIELD-II/pasa` repetitions.

Repetition	h	r	c	w
1	119, 597s	5, 778s	126, 714s	−0.02
2	125, 959s	4, 792s	161, 493s	−0.20
3	133, 935s	14, 352s	151, 091s	−0.02
4	147, 077s	2, 898s	152, 282s	−0.02
5	141, 494s	17, 514s	159, 152s	−0.01

Table 6: Waste coefficient values for `Mean-Shift/hs3`.

Experiment 2. Figure 12 shows the makespan of `FIELD-II/pasa` and `Mean-Shift/hs3` for the 3 runs where unrecoverable errors are introduced. `No-Healing` was manually stopped after 7 hours to avoid flooding the infrastructure with faulty tasks. In all cases, `Self-Healing` is able to detect the issue and stop the execution far before `No-Healing`. It confirms that the healing process is able to identify unrecoverable errors and stop the execution accordingly. As shown on Table 7, the number of submitted fault tasks is significantly reduced, which has benefits both to the infrastructure and to the gateway itself.

Run		Number of tasks	
		Self-Healing	No-Healing
application-error	FIELD-II/pasa	196	732
	Mean-Shift/hs3	249	1500
non-existent input	FIELD-II/pasa	293	732
	Mean-Shift/hs3	417	1500
non-existent output	FIELD-II/pasa	287	732
	Mean-Shift/hs3	364	1500

Table 7: Number of submitted faulty tasks.

8. Conclusion

We presented a simple, yet practical method for autonomous detection and handling of operational incidents in workflow activities. No strong assumption is made on the task

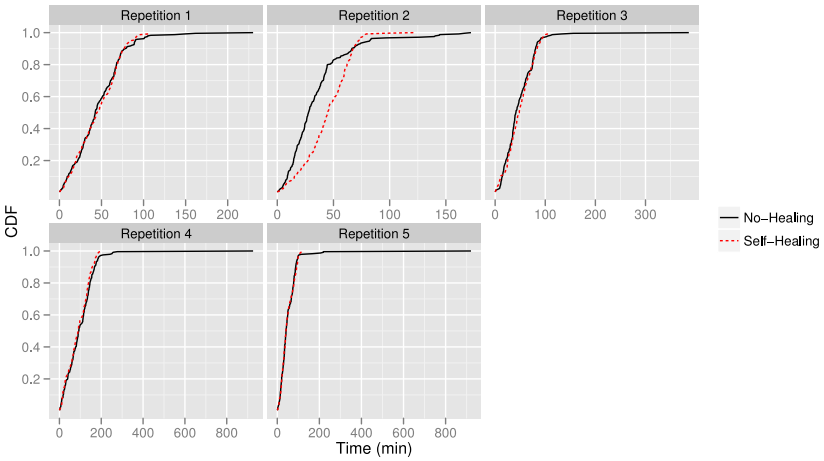

Figure 11: Experiment 1: CDF of the number of completed tasks for `Mean-Shift/hs3` repetitions.

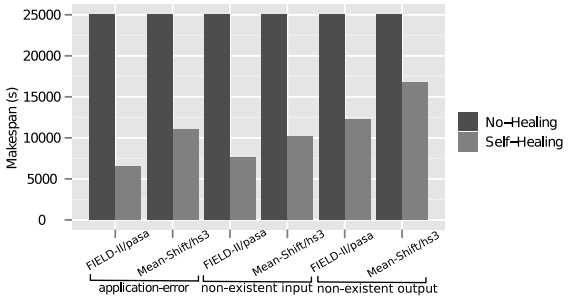

Figure 12: Experiment 2: makespan of `FIELD-II/pasa` and `Mean-Shift/hs3` for 3 different runs.

duration or resource characteristics and incident degrees are measured with metrics that can be computed online. We made the hypothesis that incident degrees were quantified into distinct levels, which we verified using real traces collected during 5 months in our Virtual Imaging Platform. Incident levels are associated offline to action sets ranging from light execution tuning (file/task replication) to radical site blacklisting or activity interruption. Action sets are selected based on the degree of their associated incident level and on confidence of association rules determined from execution history.

This strategy was implemented in the MOTEUR workflow engine and deployed on the European Grid Infrastructure with the DIRAC resource manager. Results show that our handling of blocked activities speeds up execution up to a factor of 4 and consumes up to 26% less resource time than a control execution. A second experiment shows that our self-healing loop properly detects unrecoverable errors.

As a limitation, the mechanism can only handle incidents that have been observed in the historical information. However, the approach can be extended in several ways. First, other incidents could be added, provided that they can be quantified online by a metric ranging from 0 to 1. Possible candidates are infrastructure service downtimes (e.g. file catalog, storage servers, computing sites) and unfairness among workflow exe-

cutions. Action sets could also be extended, for instance with actions related to resource provisioning.

Besides, mode detection used for incident quantification could be improved by (i) automated detection (e.g. with Mean-Shift [35]) and (ii) periodical update from execution history. Using the history of actions performed to adjust incident degree could also be envisaged. For instance, incidents for which several actions already have been taken could be considered more critical.

Finally, other components of science-gateways could be targeted with the same approach. Our future work addresses complete workflow executions, taking actions such as pausing workflow executions, detected blocked workflows beyond activities, or allocating resources to users and executions.

9. Acknowledgment

This work is funded by the French National Agency for Research under grant ANR-09-COSI-03 "VIP'. We thank the European Grid Initiative and National Grid Initiatives, in particular France-Grilles, for providing the infrastructure and technical support. We also thank Ting Li and Olivier Bernard for providing optimization use-cases to the Virtual Imaging Platform.

References

[1] S. Gesing, J. van Hemert (Eds.), Concurrency and Computation: Practice and Experience, Special Issue on International Workshop on Portals for Life-Sciences 2009, Vol. 23, 2011.

[2] W. Cirne, F. Brasileiro, D. Paranhos, L. Goes, W. Voorsluys, On the Efficacy, Efficiency and Emergent Behavior of Task Replication in Large Distributed Systems, Parallel Computing 33 (2007) 213–234.

[3] S. Shahand, M. Santcroos, Y. Mohammed, V. Korkhov, A. C. Luyf, A. van Kampen, S. D. Olabarriaga, Front-ends to Biomedical Data Analysis on Grids, in: Proceedings of HealthGrid 2011, Bristol, UK, 2011.

[4] P. Kacsuk, P-GRADE Portal Family for Grid Infrastructures, Concurrency and Computation: Practice and Experience 23 (3) (2011) 235–245.

[5] R. Ferreira da Silva, S. Camarasu-Pop, B. Grenier, V. Hamar, D. Manset, J. Montagnat, J. Revillard, J. R. Balderrama, A. Tsaregorodtsev, T. Glatard, Multi-Infrastructure Workflow Execution for Medical Simulation in the Virtual Imaging Platform, in: HealthGrid 2011, Bristol, UK, 2011.

[6] D. Malik, J. N. Mordeson, M. Sen, On Subsystems of a Fuzzy Finite State Machine, Fuzzy Sets and Systems 68 (1) (1994) 83 – 92.

[7] J. Kephart, D. Chess, The vision of autonomic computing, Computer 36 (1) (2003) 41 – 50.

[8] H. Nguyen Van, F. Dang Tran, J.-M. Menau, Autonomic virtual resource management for service hosting platforms., in: Workshop on Software Engineering Challenges in Cloud Computing, 2009.

[9] P. Collet, F. Křikava, J. Montagnat, M. Blay-Fornarino, D. Manset, Issues and Scenarios for Self-Managing Grid Middleware, in: Workshop on Grids Meet Autonomic Computing, in association with ICAC'2010, ACM, Washington, DC, USA, 2010.

[10] L. Broto, D. Hagimont, P. Stolf, N. Depalma, S. Temate, Autonomic management policy specification in Tune, in: Proceedings of the 2008 ACM symposium on Applied computing, New York, NY, USA, 2008, pp. 1658–1663.

[11] C. Germain-Renaud, A. Cady, P. Gauron, M. Jouvin, C. Loomis, J. Martyniak, J. Nauroy, G. Philippon, M. Sebag, The grid observatory, IEEE International Symposium on Cluster Computing and the Grid (2011) 114–123.

[12] A. Iosup, D. Epema, Grid computing workloads, Internet Computing, IEEE 15 (2) (2011) 19 –26.

[13] D. Lingrand, J. Montagnat, J. Martyniak, D. Colling, Analyzing the EGEE production grid workload: application to jobs submission optimization, in: 14th Workshop on Job Scheduling Strategies for Parallel Processing (JSSPP'09), Roma, Italy, 2009, pp. 37–58.

[14] D. Kondo, B. Javadi, A. Iosup, D. Epema, The failure trace archive: Enabling comparative analysis of failures in diverse distributed systems, in: 10th IEEE/ACM International Conference on Cluster, Cloud and Grid Computing (CCGrid), 2010, pp. 398 –407.
[15] E. Imamagic, D. Dobrenic, Grid Infrastructure Monitoring System Based on Nagios, in: Proceedings of the 2007 workshop on Grid monitoring, New York, NY, USA, 2007, pp. 23–28.
[16] T. Elteto, C. Germain-Renaud, P. Bondon, M. Sebag, Towards Non-Stationary Grid Models, Journal of Grid Computing.
[17] X. Zhang, C. Germain-Renaud, M. Sebag, Adaptively Detecting Changes in Autonomic Grid Computing, in: Procs of ACS 2010, Belgique, 2010.
[18] E. Stehle, K. Lynch, M. Shevertalov, C. Rorres, S. Mancoridis, On the use of computational geometry to detect software faults at runtime, in: Proceeding of the 7th international conference on Autonomic computing, New York, NY, USA, 2010, pp. 109–118.
[19] A. Benoit, L. Marchal, J.-F. Pineau, Y. Robert, F. Vivien, Scheduling concurrent bag-of-tasks applications on heterogeneous platforms, IEEE Transactions on Computers 59 (2010) 202–217.
[20] C.-C. Hsu, K.-C. Huang, F.-J. Wang, Online scheduling of workflow applications in grid environments, Future Generation Computer Systems 27 (6) (2011) 860 – 870.
[21] H. Casanova, M. Gallet, F. Vivien, Non-clairvoyant scheduling of multiple bag-of-tasks applications, in: Euro-Par 2010 - Parallel Processing, Vol. 6271, 2010, pp. 168–179.
[22] J.-N. Quintin, F. Wagner, Wscom: Online task scheduling with data transfers, Cluster Computing and the Grid, IEEE International Symposium on 0 (2012) 344–351.
[23] S. Camarasu-Pop, T. Glatard, J. T. Moscicki, H. Benoit-Cattin, D. Sarrut, Dynamic Partitioning of GATE Monte-Carlo Simulations on EGEE, Journal of Grid Computing 8 (2) (2010) 241–259.
[24] R. Garg, A. K. Singh, Fault tolerance in grid computing: state of the art and open issues, International Journal of Computer Science & Engineering Survey (IJCSES) 2 (1).
[25] H. Casanova, On the harmfulness of redundant batch requests, International Symposium on High-Performance Distributed Computing 0 (2006) 255–266.
[26] W. H. Bell, D. G. Cameron, R. Carvajal-Schiaffino, A. P. Millar, K. Stockinger, F. Zini, Evaluation of an economy-based file replication strategy for a data grid, in: 3rd IEEE/ACM International Symposium on Cluster Computing and the Grid, 2003, p. 661.
[27] A. H. Elghirani, R. Subrata, A. Y. Zomaya, A proactive non-cooperative game-theoretic framework for data replication in data grids, in: 8th IEEE International Symposium on Cluster Computing and the Grid (CCGRID), 2008, p. 433.
[28] J. Rehn, T. Barrass, D. Bonacorsi, J. Hernandez, I. Semeniouk, L. Tuura, Y. Wu, Phedex high-throughput data transfer management system., in: Computing in High Energy Physics, CHEP'2006, 2006.
[29] R. Agrawal, T. Imielinski, A. Swami, Mining Association Rules between Sets of Items in Large Databases, 1993, pp. 207–216.
[30] K. A. De Jong, An Analysis of the Behavior of a Class of Genetic Adaptive Systems., Ph.D. thesis, University of Michigan, Ann Arbor, MI, USA, aAI7609381 (1975).
[31] R. Ferreira da Silva, T. Glatard, A science-gateway workload archive to study pilot jobs, user activity, bag of tasks, task sub-steps, and workflow executions, in: CoreGRID/ERCIM Workshop on Grids, Clouds and P2P Computing, Rhodes, GR, 2012.
[32] T. Glatard, J. Montagnat, D. Lingrand, X. Pennec, Flexible and Efficient Workflow Deployment of Data-Intensive Applications on Grids with MOTEUR, International Journal of High Performance Computing Applications (IJHPCA) 22 (3) (2008) 347–360.
[33] A. Tsaregorodtsev, N. Brook, A. C. Ramo, P. Charpentier, J. Closier, G. Cowan, R. G. Diaz, E. Lanciotti, Z. Mathe, R. Nandakumar, S. Paterson, V. Romanovsky, R. Santinelli, M. Sapunov, A. C. Smith, M. S. Miguelez, A. Zhelezov, DIRAC3. The New Generation of the LHCb Grid Software, Journal of Physics: Conference Series 219 (6) (2009) 062029.
[34] R. Ferreira da Silva, T. Glatard, F. Desprez, Self-healing of operational workflow incidents on distributed computing infrastructures, in: IEEE/ACM CCGrid 2012, Ottawa, Canada, 2012, pp. 318–325.
[35] D. Comaniciu, P. Meer, Mean Shift: A Robust Approach Toward Feature Space Analysis, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (5) (2002) 603–619.