A systematic review of fault tolerance in mobile agents

Bassey Echeng Isong¹, Eyaye Bekele²

¹Deptmemt of Computer Science, University of Venda, Thohoyandou, South Africa
²School of Computing, Blekinge Institute of Technology, Karlskrona, Sweden

Email address:
bassey.isong@univen.ac.za(B. Isong), eyayeb@gmail.com (E. Bekele)

To cite this article:

Abstract: Mobile agents have engrossed substantial attention in recent years, especially in fault tolerance researches and several approaches have emerged. Fault tolerance design tends to put a stop to incomplete or complete loss of the agent in the face of failures. Despite these developments, reliability issues still remain a critical challenge. Moreover, there is no comprehensive detail bringing together, summaries of the existing efforts of researches in order to focus attention where it is needed most. Therefore, our objective in this systematic literature review (SR) is to explore and analyze the existing fault tolerance implementations in order to bring about the state-of-the-art and the challenges in mobile agent’s fault tolerance approaches. We used studies from a number of relevant article sources, and our results showed the existence of twenty six articles. Our analysis indicates that the existing approaches are not generic and each focuses on a specific aspect of the problem, usually in one or two specific fault models which impacts on agent’s reliability. The implication of the study is to give a clear direction to future researchers in this area for a better reliable and transparent fault tolerance in mobile agents.

Keywords: Mobile Agents, Fault Tolerance, Replication, Check-Pointing, Systematic Review, Platforms

1. Introduction

In the last few decades the field of mobile agents in distributed computing has witnessed substantial attention in both academia and industrial fields. This is known to be stimulated by the exponential growth of the Internet and system dependability. However, in spite of all, mobile agent’s reliability is still a critical issue. Due to their nature, mobile agent’s reliability and execution is not failure-free in the environments they operate. The growth of distributed heterogeneous environments such as the Internet naturally exposes them to abnormal conditions originating from migration request failure, communication exceptions or security violation [3],[4],[6]. Hence, providing reliability is of the essence to integrate agent-driven systems into today’s industrial applications. Mobile agents have to be made reliable through fault tolerance to withstand adverse environmental situations in today’s industrial applications. Fault tolerance is designed to provide reliable execution of mobile agents even in the presence of system failure.

Achieving mobile agent’s fault tolerance requires the adherence to the non-blocking and exactly-once execution [7]. Presently in literature, several mobile agents’ fault tolerance approaches exist in variety of mobile agent platforms. These approaches used different mechanisms to provide the reliability mobile agents’ execution needs especially in the failure detection and recovery aspects. Moreover, majority of the recent approaches are based on optimizations, hybrid-based, while others are based on exception handling. In general, the existing fault tolerance schemes are categorized as either curative in nature (e.g. exception handling) or preventive [8]. [9]. [10]. The preventive schemes are further categorized into check-pointing and replication-based schemes but in some cases a hybrid of both schemes [8].

Despite several efforts and interest in this field, there is no comprehensive detail bringing together, summaries of these efforts. The main gap in this research area lies in the fact that the existing approaches are not generic and each focuses on a specific aspect of the problem, usually in one or two specific fault modes which is known to have huge impacts on agent’s reliability. To improve the reliability of a system, faults originating from different forms need to be addressed in the fault tolerance measure. Unfortunately, there is no fault tolerance framework in the existing literature that serves as a guideline for realizing the state-of-the-art in mobile agent’s fault tolerance. Therefore, bringing together fault tolerance approaches in mobile agent will assist researchers in closing the gaps identified in this study. The objective of this systematic literature review (SR) is to
explore the existing fault tolerance approaches in mobile agents system to identify the current state of research, the techniques and approaches used, and the factors that influence the execution reported in recognized fault tolerance implementations research.

The rest of this article is organized as follows. An overview of mobile agents is given in Section 2. The research methodology of this SR is described in Section 3. The analysis of the results in accordance with the research questions is presented in Section 4. Section 5 provides the study discussions while the conclusions and recommendations are in Section 6.

2. Overview of Mobile Agents

Mobile agents are encapsulated pieces of executing program that have the ability to travel from one host to another and perform certain task autonomously [3]. It is a technology that aimed at shifting computation towards the data other than the other way round [1]. Mobile agents have characteristics that are distinct, thus making them flexible in deployment and desirable for use in distributed applications than other technological paradigms such as client-server, peer-to-peer, and others [2]. (see Figure 1.) These features include ability to naturally operate in heterogeneous environments [9], act autonomously [8], move independently from one host and can effectively make execution decision [14]. Mobile agents heavily rely on the underlying protocol for communications by way of interactions and message exchanges in order to successfully carry out and execute certain task in the agent system [9].

Based on their characteristics, mobile agents provide several benefits such as bandwidth conservation [15], asynchronous and autonomous interactions [8], extended flexibility in disconnected data operations [16] and can improve network latency with better response time [16], robust and fault tolerant [17]. Mobile agents are generally independent of the computer-layer and transport-layer but dependent only on their execution environment [18], and have better scalability [16].

Today, the concept of mobile agents is receiving considerable attention in both research and industrial fields. Several platforms exist that provides operating environments for mobile agents such as Aglets, Agent Tcl, Knowbots, Telescript, Voyager, Mole, Tacoma, Grasshopper, James, Swarm and others [1], [2], [13]. Moreover, they have applications in several areas such as e-commerce and m-commerce, network monitoring and management, distributed information retrieval, telecommunications, remote device control and configuration, Internet computing, etc [2],[5],[8],[19],[20],[22]. To this end, despite the flexibility offered by mobile agents, agents are not isolated from several challenges such as malicious or errant hosts, erratic Internet behaviors or resource scarcity [6]. These therefore, calls for reliability and security mechanisms to be in place [9],[12]. The reliability issue is being addressed by fault tolerance mechanisms, which is the focus of this study.

3. Research Method

Systematic review is a methodology aimed at minimizing the inconsistencies associated with less scientifically rigorous review methodologies through strict qualitative research methods resulting to objective and unbiased results. In this study, we have applied SR to explore the state-of-the-art of fault tolerance in mobile agents by following the guidelines in [23]. The steps involves are discussed in subsequent sections.

3.1. Research Questions

This SR aimed to summarize the existing mobile agent’s fault tolerance approaches in recent years. It would provide a list of reported and recognized techniques and approaches, influencing factors, platform supports and challenges in mobile agent’s fault tolerance. Therefore, the research questions are:

SRQ1. What is the state-of-the-art in research of the recognized mobile agent’s fault tolerance?
SRQ2. What available fault models are considered in designing fault tolerance protocols?
SRQ3. What are the available approaches and their design elements in the available recognized mobile agent’s fault tolerance in current state of research?
SRQ4. What factors influences mobile agent’s fault tolerance execution?
SRQ5. How much supports are offered by the mobile agent’s platforms used in implementing the fault tolerance features?
SRQ6. What challenges exist and how do they affects the implementation of the fault tolerance in mobile agents?

3.2. Search Strategy

A search strategy is designed to ensure that all relevant studies other than irrelevant ones appear in the search result. In this SR, our literature search is limited over the scopes of: publication time period and publications that discusses fault tolerance in mobile agents. We considers the review of 10-years’ efforts in mobile agents fault tolerance, that spanned from January 1998 to December 2008. We selected these periods in order to obtain relevant and sufficient information that are of the essence to this study and provide evidences of the trends in mobile agent fault tolerance then.
Therefore, any paper published after 31 December, 2008 is not included in our search result. We limited our searches to the electronic databases: Compendex/Inspec, IEEE Xplore, Google Scholar, ACM digital library, Springer link and Scirus since they contain peer-reviewed works published in journals, digital libraries, conferences, proceedings and workshops which are of recognized quality within the software engineering research community. In this study, the quality of each selected research article was evaluated against a number of checklist questions. Each of questions is answered based on three options along assigned weights: Yes=1, Partial=0.5 and No=0. The maximum score a particular publication can get is 8. (see Table 1)

**Table 1. List of selected publications by publisher and methodology**

<table>
<thead>
<tr>
<th>Ref.</th>
<th>Authors</th>
<th>Year</th>
<th>Publisher</th>
<th>Published in</th>
<th>Methodology</th>
<th>Quality Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>[27]</td>
<td>Lyu, M.R.</td>
<td>2003</td>
<td>IIIS</td>
<td>Conference</td>
<td>All</td>
<td>5.5</td>
</tr>
<tr>
<td>[31]</td>
<td>Meng, Xuejun</td>
<td>2006</td>
<td>IEEE</td>
<td>Conference</td>
<td>Model and Experiment</td>
<td>5</td>
</tr>
<tr>
<td>[33]</td>
<td>Marin, Olivier</td>
<td>2005</td>
<td>Springer</td>
<td>Workshop</td>
<td>Model and Experiment</td>
<td>8</td>
</tr>
<tr>
<td>[34]</td>
<td>Lyu, Michael R.</td>
<td>2004</td>
<td>IEEE</td>
<td>Journal</td>
<td>Model and Simulation</td>
<td>4.5</td>
</tr>
<tr>
<td>[36]</td>
<td>Silva, Luis Moura</td>
<td>2000</td>
<td>IEEE</td>
<td>Conference</td>
<td>Model and Experiment</td>
<td>6</td>
</tr>
<tr>
<td>[38]</td>
<td>Rothermel, Kurt</td>
<td>1998</td>
<td>IEEE</td>
<td>Conference</td>
<td>Model</td>
<td>3.5</td>
</tr>
<tr>
<td>[40]</td>
<td>Taesoon Park</td>
<td>2004</td>
<td>Springer</td>
<td>Conference</td>
<td>Model and Experiment</td>
<td>5.5</td>
</tr>
<tr>
<td>[41]</td>
<td>Taesoon Park</td>
<td>2004</td>
<td>Springer</td>
<td>Conference</td>
<td>Model and Experiment</td>
<td>6</td>
</tr>
<tr>
<td>[46]</td>
<td>Park, Taesoon</td>
<td>2004</td>
<td>Springer</td>
<td>Conference</td>
<td>Model and Experiment</td>
<td>5</td>
</tr>
<tr>
<td>[48]</td>
<td>Park, Taesoon</td>
<td>2006</td>
<td>Springer</td>
<td>Conference</td>
<td>Model and Simulation</td>
<td>3</td>
</tr>
</tbody>
</table>

The study selection process was individually carried out by the authors involved and any differences were settled by consensus. The search strategies we adopted were iterative in nature and the inclusion/exclusion decisions were checked at least twice and discussed at each stage of execution. We adopted a multi-stage process in selecting the studies in accordance with the guidelines in [23], using different selection criteria. All search terms we created were applied on the selected databases and a total of 6,901 results were found. In the first stage, 6788 articles were excluded based on the relevance of their title or abstract. Furthermore, the titles and abstract of the left over 113 were read and the basic inclusion and exclusion criteria applied, leaving a total of 86 studies. In the second stage, 55 out of the 86 articles
were selected based on the application of the detailed inclusion and exclusion criteria - abstract, the conclusion and in some cases the introduction was reviewed to apply the exclusion criteria. The exclusion criteria were based on inaccessibility, formal or mathematical description, and exception handling. Lastly, in the third stage, we based the selection process on the detailed research questions while the exclusion criteria was based on issues of duplication, application of mobile agents in a different study area, and mobile agent’s platforms articles. As a result, 26 unique studies were selected as primary studies for this SR and 29 studies were discarded.

3.3. Data Extraction and Synthesis

The data extraction strategy was developed in accordance with the research questions, quality assessment checklist, general information associated with the study identification and certain common characteristics in the studies. During the extraction, the authors also checked and re-checked the extracted data to get rid of uncertainties. To assist us find and validate the extracted information and resolve inconsistencies quickly, we tinted all important lines and paragraphs in the selected studies. Accordingly, difficulties encountered were resolved via discussion among the authors. For multiple articles of the same information, articles with the most complete and latest information were used to avoid unbiased findings. To extract relevant information, we created and used data extraction form with the following fields: Title, Authors names, Journal/Conference/Workshop, Year, Research Methodology, Mobile agent fault tolerance Scheme, Protocol, Fault model, Assumptions Detection, Recovery, Fault Tolerance execution, Agent types, Communication, Factors affecting the performance of the proposed model and experiment variables, Platform type, Platform support and Challenges.

With the extracted data from the extraction forms coupled with the nature of this study, we performed descriptive synthesis of the data since it is the only suitable method in such heterogeneous data format.

4. Analysis

In this section, we present analysis of the results of this SR by answering the research questions as follows:

4.1. Mobile Agents Fault Tolerance Research

RQ1: What is the state-of-the-art in research of recognized mobile agent’s fault tolerance?

To answer this research question, analysis will be based on publication years, the qualities of the articles and the methodology used. Table 1 presents list of selected publications by publisher and Methodology as well as the quality score for each study.

4.1.1. Year of Publications

We found 26 studies related to mobile agent fault tolerance where 21 studies were published in conference proceedings, 2 studies in Journals and 3 studies in workshops. Analysis shows that the field of mobile agent fault tolerance was active in research in those periods. Further analysis indicates that years between 2004 and 2006 have showed a remarkable increase in number of publications, though the trend seems to be going down in the last three years (2006 - 2008). Figure 2 and Table 1 show studies by year of publication.

4.1.2. Publication Quality Scores

Analysis shows that more than 75% of the selected publications scored 4.5 or more. Articles with more than 4 as score in the quality assessment generally are selected on the basis of having the most vital information such as detailed description of a model and some form of proof such as results from a real experiment or simulation to support their findings. (see Figure 3)
experiments and simulations has 8%. Only 12% of the studies discussed fault tolerant models with no experiments or simulations. (see Figure 4).

4.2. Fault Models

SRQ2. What available fault models are considered in designing fault tolerance protocols?

In the 26 studies, we found that the ability to observe the agents and detect failure during execution is one of the cardinal features of any fault tolerance in mobile agent approach. This is because mobile agents or its environments are not failure-free. We noticed ‘fault model’ is used to define which set of observations are categorized as failure and which are acceptable operation modes. Analysis shows that all the existing approaches have fault models and there are three classes of agent’s failures: communication, crash and agent software failure. Moreover, we found that most of the studies are designed to either cater for one of the stated failures or multiple of them. Table 2 presents the existing implementations of each study and the failure types designed for them.

Table 2. Failure Types

<table>
<thead>
<tr>
<th>Failure Types</th>
<th>Communication</th>
<th>Crash</th>
<th>Agent/Agent Software</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>[25], [26], [27], [29],[30],[31],[32], [34],[42],[35],[36],[38],[40],[41],[42],[44],[46],[47],[49]</td>
<td>[5],[7],[25],[26],[27],[29],[30],[31],[32],[34],[42],[35],[36],[38],[40],[41],[42],[44],[46],[47],[49]</td>
</tr>
<tr>
<td>[25],[30],[31],[32],[42],[36],[38],[39],[43],[44],[45]</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Based on Table 2, the distribution of fault models in the studies presented in Figure 5 denotes that fault models for communication failures were least addressed (43%). This could depend on the assumption that the network is reliable or agents can eventually resume service even if the network fails. Crash and agent software fault models seem to be supported in most of the existing implementations (84%). Only about 35% of the studies supported all three fault models [44], [27], [38], [30], [31], [32], [42], [36], [45], while 12% of these studies considered the issue of network partitioning and suggested a solution to it [44], [38],[30].

4.3. Mobile Agents Fault Tolerance Approaches and Design

SRQ3. What are the available approaches and their design elements in the available recognized mobile agent’s fault tolerance in current state of research?

This research questions will be answered based on the different approaches of mobile agent’s fault tolerance, the available protocols and the design of the available implementations.

4.3.1. Fault Tolerance Schemes

One of the key features of fault tolerance approach is the ability to recover from failure. In this study, we found that the existing fault tolerance schemes that deal with sources of system failures and recovery were categorized as either the replication-based, checkpoint-based or a hybrid schemes [16], [9], [49]. In this study, we only considered the two most widely used schemes: check-pointing and replication for analysis. The different studies that used these approaches in the perspective of transactional and non-transactional approach are shown in Table 3.

Table 3. Fault tolerance schemes and execution modes

<table>
<thead>
<tr>
<th>Agent’s Execution Modes</th>
<th>Transactional</th>
<th>Non-transactional</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scheme</td>
<td>[5], [38], [36], [27],[34],[45],[48],[46],[32]</td>
<td>[29],[39],[7],[47],[43],[40],[41],[33],[44],[25],[35]</td>
</tr>
<tr>
<td>Hybrid</td>
<td>[14],[31]</td>
<td>[28]</td>
</tr>
</tbody>
</table>

Further analysis shows that both check-pointing and replication-based schemes are used almost equally over the years of consideration (see Figure 6).
4.3.2. Replication Scheme Types

For replication-based schemes: active and passive, analysis in this study shows that the trend of distancing away from active replication-based schemes is quite evident. (See Figure 7) In addition, about 72% of the replication-based approaches used either the passive or semi-active replication, but the semi-active to passive ratio within the passive replication-based category is 3:7. This indicates that semi-active techniques are not used often either. It could be as a consequence of high computation and communication cost they incur. The computation overhead in active replication is higher even among the implementations that support both replication types.

![Figure 7. Replication scheme types distribution](image)

4.3.3. Execution Modes

We found that execution mode in existing agents fault tolerance are either transactional or non-transactional. From all indication, we noticed that the distribution of the transactional and non-transactional executions is someway balanced (see Figure 8), though the transaction-based executions are slightly higher with about 54% than non-transaction-based modes with 46%. The analysis indicates there is no noticeable shift in trend over the years. However, transactional executions are more reliable in maintaining the exactly-once property of agent execution, while its counterpart maintain lower computation overhead. But we recommend transactional execution for application domains that require higher level of consistency.

![Figure 8. Agent's execution modes](image)

4.3.4. Communication Modes

In this study, we found that there are only two communication modes in fault tolerant mobile agent dominated by the asynchronous mode. Analysis shows that about 92% of the studies implemented asynchronous mode execution while only 8% of the studies are both synchronously and asynchronously. (see Figure 9) There was no implementation that solely works on synchronous mode. We believe this could be as a result of the characteristics of the agent’s environment where they are autonomous and migrate usually in open networks with latency. In addition, there is high performance overhead with synchronous when compared to asynchronous.

![Figure 9. Agent’s communication mode distribution](image)

4.3.5. Fault tolerance Protocols

In this SR, we found that protocols in mobile agents fault tolerance implementation model are mostly achieved through effective message passing to coordinate and ensure reliable agent failure detection and recovery. In addition, categorizing these protocols into classes is not easy since some of them exist in association with other protocols. However, they can only be organized with respect to the two execution properties of mobile agents: the exactly-once or non-blocking. Figure 10 shows the class of existing protocols in mobile agents fault tolerance implementation model.

![Figure 10. Agent’s protocols](image)
4.3.5.1. Exactly-One Protocols

This protocol guarantees that an agent must execute the desired action not more and not less than once in a host. This protocol includes the consensus and the transaction-based approaches.

A. Consensus Protocol: This protocol provides means of achieving agreement between the primary agent process and the replicas in the case of failure. This is usually achieved by voting on a certain result by all the participating agents where a consensus is reached when all the participants agree on one outcome. In this case, a replica can become the new primary through its own consensus for either being the first to detect or having the highest priority when the primary fails.

B. Transaction-Based Protocol: Similar to consensus, but uses the transaction commits/abort to reduce the effect of failures on the availability of operational sites during distributed transaction. Commit and abort are both irreversible and unlike consensus, transaction can only be committed if all participating sites vote or agree to commit, otherwise, the transaction is aborted. In this study, analysis shows that most of the studies used the basic transaction-commit protocol such as centralized transaction commit by [26] while [7],[27],[30],[35],[36],[47] used the transaction commit protocol. A hybrid approach of both consensus and transaction-based protocol were implemented by [29],[47],[7],[38].

4.3.5.2. Non-Blocking Protocols

These protocols are used basically for overcoming blocking problems pose by the exactly-once protocols. It is achieved by either restarting a failed process or use duplicate agents that are able to take over in time of primary agent failure without voting. Protocols under this category include the rollback recovery protocols and the replication-based protocols.

A. Rollback Recovery Protocol: This approach tries to reduce the amount of loss in computation by avoiding the restarting of computation from the beginning. It provides a technique that requires a process to periodically record its consistent state, known as check-pointing, into a stable storage. Most of the existing rollback recovery approaches are based on message passing: checkpoint-based and message logging-based.

1. Checkpoint-based: This approach depends strictly on regularly saving agent’s process states and code to a stable storage for future restoration or recovery in the event of failure. Protocols in this category include coordinated, uncoordinated, communication-induced, lazy and the timer-based protocol. In this study we found that most of the rollback-based protocols used this protocol to provide fault tolerance behavior such as communication-induced check-pointing [42], independent check-pointing with receiver-based logging [46], checkpoint-based scheme to restore agents processes back to its consistent state during failure, and checkpoint-based with reliable publisher/subscriber messaging layer[45].

2. Message Log-base: This is the logging of non-deterministic actions preset as determinants in combination with check-pointing to achieve fault tolerance recovery behavior. Protocols under this category include the pessimistic and optimistic logging protocol [7]. Pessimistic execution ensures that changes are applied only if no agent crashes and there is no erroneous result, while in optimistic execution the place modifications can be immediate and transparent but undoing modifications is a complex task [7].

B. Replication-based Protocols: In this approach, there is a live backup agent, either as a duplicate of the primary or as a standby, even before failure is detected. One drawback is that it requires synchronization of replicas with the primary, which is very expensive both in computation and communication. However, several optimizations have been suggested in the studies we considered such as formed a group of replicas and a proxy that communicates with the primary on behalf of the multiple replicas by [39], dynamic adaptive replication scheme by [33] and the sliding window protocol by [25], a technique that controls the number of backup or replicated agents in order to minimize bandwidth consumption.

3.3.6. Implementation Models

In the studies we found in this SR, we noticed that there were several approaches used in designing mobile agent fault tolerance models. These approaches are either integrated into the underlying agent’s codes or platforms. Further analysis shows that most of the approaches are a hybrid of a variety of the protocols but some also include preventive techniques. Figure 11 presents the design approaches of existing implementations.

![Figure 11. Fault tolerance design approach](image)

4.3.6.1. Primary Back-up Model

The primary back-up model (PBM) design requires the replication of the agents or server into primary component (worker component) and one or more observer components (backups/replicas component). The primary component takes charge of execution while the backup components monitor the primary’s computation for any possible failure.

A. Traditional Primary-Backup/Task replication/Data redundancy: In this design, the traditional primary-backup model relies on the replication of system components, task, data, etc. in order to achieve fault tolerance in mobile agents.

B. Rear Guard Agents as Backup: It is chiefly based on the primary-backup principle but instead the backup agent resides on the previously visited host. The backup monitors the primary agent and perform recovery actions to resume
the computation when it detects failure of the primary agent especially in a stage-based partition network. We found several forms of these approaches such as multiple rear-guard in [5] and parallel processing of replicas by [44]. Others are optimizations to the approaches by [43] based on rear-guard protocol and reliable broadcast protocol with election protocol, while [7] discusses the pipe-line mode.

4.3.6.2. Primary Witness Model

The primary witness model (PWM) approach is very similar to the PBM, but here the backup agent is a different type of agent usually for monitoring and creating a backup agent when the primary agent fails. The backup agent that replaces the failed primary agent is only created after the detection of a failure. These approaches include rear guard agents as witness and the collaborative agents.

A. Rear Guard Agents as Witness: The witness agent involves a different type of agent that cannot on its own take over as a primary agent during recovery. The primary agent takes charge of the natural execution while the witness only monitors and recovers. In this case, it creates a new primary and restores normal execution. We found this approach in studies such as Monitor Agent (MA) which detects failure of Execution Agent (EA), it creates Repair Agent (RA) for fixing the error in the EA host [31], and the actual and witness agents that creates a probe agent for recovering log during recovery[27], [34].

B. Collaborative Agents: Here, three or more types of agents with designated responsibilities in the detection and recovery processes work together for achieving fault tolerance action with a clear division of labor. The approach involves the primary agent in charge of the execution, while others participate for specifically task such as monitoring, tracking, checking path, recovering, creating another type of agent, etc. Other agents cannot assume the primary agent position during recovery. In this study, studies that used this approach are [25] using three agent types: observer agents, ping agents and transaction agent used for monitoring, path checking and executing transactions respectively. Also, [26] uses three types of agents which are worker, monitor and tracker.

4.4. Mobile Agents Fault Tolerance Performance Factors

SRQ4. What factors influences mobile agent’s fault tolerance execution?

We noticed in this study that mobile agent’s fault tolerance execution is affected by several factors leading to a decrease in their performance as reported in most studies’ experimental data. However, we only focused on the factors rather than the actual figures from the experiments due to unique scenarios and measurements that cannot be quantitatively compared across the studies. From these studies, the influencing factors are as follows:

4.4.1. Agent Size

Agents are affected by the size of their code and the payload data it takes with itself during mobility. In this study, more than 50% of the experimental studies reported agent size as a performance factor and that the execution time of the agents linearly increases with the increase in size of the agent [14], [40], [41], [42], [36], [46], [7]. In addition, increase in size also lead to replication process having overheads [7],[40].

4.4.2. Numbers of Replica

Also more than 50% of the replication based schemes reported that the number of replica affects their performance especially in the synchronous replication schemes [40]. For instance, a design with consensus requires an increase in agent’s replicas leading to agents spending longer time than expected. In this SR, the studies that reported number of replica/witness factor in their findings are [40], [41], [36], [46], [39], [28], [34].

4.4.3. Message Size

The number of uncontrolled messages can overwhelm fault detection capability and reducing number of messages can produce performance gain [7]. Some studies that reported the number of messages factor in their findings are [14], [38], [45], [34].

4.4.4. Number of Messages

The number of uncontrolled messages can overwhelm fault detection capability and reducing number of messages can produce performance gain [7]. Some studies that reported the number of messages factor in their findings are [14], [38], [45], [34].

4.4.5. Number of Hops/Host

With agent’s characteristics, their survivability decreases with increase in the number of servers the agent’s have to visit [34]. This is believed to have impact on reaching a timeout when the execution time takes a long time especially for schemes that relies on timeout and can initiate an unnecessary recovery process that would affect the performance of the system negatively. In this SR, studies that reported this problem in their findings are [7], [26], [27], [44].

4.4.6. Frequency of Check-pointing

Lastly, another important influencing factor is the degree at which check-pointing are taken. An increase in check-pointing frequency increases overhead while infrequent check-pointing brings about much re-computation in the event of recovery [6],[5].

4.5. Platforms Supports

SRQ5. How much supports are offered by the mobile agent’s platforms used in implementing the fault tolerance features?

A platform in this context is an executing environment for mobile agents. In this SR, we found varieties of agent’s
platforms which are mostly built from Java. In addition, most of the existing platforms provide partial or incomplete facilities for fault tolerance mechanisms. We noticed that several academic and commercial systems are available which differs in their features, architecture and implementations, but more or less offer common facilities for the support of agent migration, inter-agent communication, various forms of security and programming or interpreted languages etc.

Analysis here is based on a qualitative comparison of the current agent platforms. Table 4 presents the various platforms used and the nature of fault tolerance support they have in their implementation.

<table>
<thead>
<tr>
<th>Platform</th>
<th>Prog. Lang.</th>
<th>Mobility</th>
<th>Communication</th>
<th>Fault Tolerance Feature</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aglet [58]</td>
<td>Java</td>
<td>Weak</td>
<td>Asynchronous, Synchronous, Proxy</td>
<td>NA</td>
</tr>
<tr>
<td>Concordia [59]</td>
<td>Java</td>
<td>Weak</td>
<td>Asynchronous</td>
<td>Yes (Checkpoint, Transactional message queue, Proxy)</td>
</tr>
<tr>
<td>FIPA-OS [60]</td>
<td>Java</td>
<td>Weak</td>
<td>Asynchronous</td>
<td>Yes (Replication(clone), Transactional message queue, Proxy)</td>
</tr>
<tr>
<td>Grasshopper [61]**</td>
<td>Java</td>
<td>Weak</td>
<td>Synchronous, Asynchronous, Dynamic method invocation, Multicast,</td>
<td>NA</td>
</tr>
<tr>
<td>JADE [62]</td>
<td>Java</td>
<td>Weak</td>
<td>Asynchronous</td>
<td>NA</td>
</tr>
<tr>
<td>JAMES [21]</td>
<td>Java</td>
<td>Weak</td>
<td>JavaSpace</td>
<td>Yes (Checkpoint)</td>
</tr>
<tr>
<td>MadKit [63], [64]</td>
<td>Java</td>
<td>Weak</td>
<td>asynchronous message passing</td>
<td>Yes (congestion management, agent monitoring mechanisms)</td>
</tr>
<tr>
<td>MOLE [65]</td>
<td>Java</td>
<td>Weak</td>
<td>Asynchronous, Synchronous, Sessions</td>
<td>Yes (Transactional message queue)</td>
</tr>
<tr>
<td>Naptel [66]</td>
<td>Java</td>
<td>Weak</td>
<td>asynchronous message passing</td>
<td>Yes (agent monitoring mechanisms, cloning)</td>
</tr>
<tr>
<td>Tacoma [68]</td>
<td>C/C++, ML, Perl, python</td>
<td>Weak</td>
<td>Asynchronous, Synchronous</td>
<td>Yes (Rear guards)</td>
</tr>
<tr>
<td>Voyager [69]</td>
<td>Java, C#, C++</td>
<td>Weak</td>
<td>Asynchronous, Synchronous, Multicast, Proxy</td>
<td>NA</td>
</tr>
</tbody>
</table>

Note: NA (Not Applicable) implies that either the information or the feature is unavailable. ** Authors have stopped updating the platform.

4.6. Challenges

SRQ6. What challenges exist and how do they affects the implementation of the fault tolerance in mobile agents?

There are lots of challenges that faced in existing fault tolerance implementations that has limited their efficiency or the direct application of fault tolerance strategies. Some of the challenges found in this SR are discussed as follows:

4.6.1. Reliable Fault Detection

The key challenge in fault detection is when a fault tolerance process wrongly detects fault and acts upon it. In such cases, an agent is wrongly assumed failed and a replacement agent is created in place of the failed agent leading to the violation of the "exactly once" property of agent execution [25]. In this study, we see that existing fault tolerance schemes, rely on techniques such as timeout, periodical exchange of heartbeat message, call backs, etc. for reliable detection of faults but none is absolute in detecting the occurrence of failures. Only 20% of the studies specifically mention this problem and considered it as a very critical attribute.

4.6.2. Network Partition

This is due to communication failure where agent’s implementations execute in stages and internal network partitioned into stages or domains. In the event of communication failure, the various stages will be unable to communicate either to advance to the next stage or complete the assigned task, which results in blocking. In this study, only about 12% of the studies specified network partition as a challenge while most studies regard it as a temporary problem.

4.6.3. Lack of Full Process State Capture/Restore

This study found that majority of the existing agent platforms are Java-based systems using the Java Virtual Machine (JVM). With these platforms, analysis indicates that about 90% of them do not allow the capturing and restoring of the full execution state of a process which in turn affects strong mobility support.

4.6.4. Lack of Interoperability

This study noticed fault tolerant agents developed for one agent platform cannot be ported easily to a different agent
platform, which impacts the adoption and full realization of fault tolerance. We found few standards such as Foundation for Intelligent Physical Agents (FIPA) and Mobile Agent System Interoperability Facility (MASIF) and steps have been taken to realize this issue in the future.

4.6.5. Lack of Full Transactional Support

This study found that there is no full transactional support in existing fault tolerant framework especially for various types of failure situations. This could be due to the active nature of mobile agent systems which is somewhat incompatible with the concept of transactions that forces several agents to remain in the same transaction as long as it does not commit or abort [6],[29],[69].

4.6.6. Scalability

This study found that all the studies are from the academia and the implemented mobile agent’s fault tolerance were not large scale. Analysis shows that most experiments and simulations performed are on a small scale setup, though some promise scalability [26],[33],[41],[44]. In all the studies, we did not find large scale experiments.

4.6.7. Lack of Transparency

In majority of the studies we noticed that developers viewed the fault tolerance platforms as not being transparent since it requires the modification of the underlying platform to accommodate the protocols.

5. Discussion

Despite the importance of mobile agents in distributed environments coupled with the increase in research activities in both academia and industry, we found several issues that still affect the complete realization of reliable fault tolerance in mobile agents. In this study, we noticed that the distribution of the publications by year shows sharp increase between 2003 and 2004 and more than 65% of the studies were conducted between 2004 and 2006. This could be due to the increased interest in the area of mobile agents from the research community. However, the trend seems to be going down resulting in only 23% of the studies being conducted in the last three years, which we think is related to the problem of trust in mobile agents as a result of increasing threat from viruses and network worms. Accordingly, the number of studies that were supported by simulation or experiment improved over the years. They provide detailed description of a model used and reporting some results from a real experiment or simulation to support their findings.

Also the existing implementations have classified agent’s failures into communication, crash and agent software failure. Each study provides mechanism to detect these failures based on their fault model strategy, but no known fault model can detect and recover from all types of failures in the existing implementations.

It is evident that different approaches exist that deals with the recovering from failure of mobile agents. These approaches include the check-pointing, replication schemes or the hybrid approach and both were used almost equally over the years of consideration. Among the replication-based schemes, most of the studies tend to favor inactive replication due to it lower computation overhead. In the same vein, the mode of agent’s communication in existing implementations is mostly asynchronous, as well as there appears to be a balance between transactional or non-transactional execution of mobile agents. The reasons could be from the fact that transactional executions are more reliable in maintaining the exactly-once property, while the non-transactional maintain lower computation overhead. In all the studies we considered, there were other elements such as fault tolerance protocols and their design models. Protocols were based on the exactly-once or non-blocking properties, while the implementation design models were based on hybrid approaches dominated by preventive techniques of agent’s replications: primary backup and witness approaches. Each approach has its own strategy to implement fault tolerance.

Other issues we observed were challenges emanating from platform supports and the general challenges in terms of performance affecting mobile agent’s fault tolerance that has not been addressed. Variety of mobile agent platforms exist and are dominated by Java which does not support strong mobility of agents which affects full capturing of the state of collaborating processes. Also, the numbers of replicas, messages, size, etc. were among the factors reported as having impacts on the performance of mobile agent’s execution.

The consequence of this study is that investigating intensely in the state-of-the-art fault tolerance approaches and the challenges in mobile agent will serve as a starting point and give a clear direction to future researchers who will work in this area to improve the existing implementations.

5.1. Strengths and Weaknesses

In this SR, we have covered several numbers of articles, whose authors are listed in Table 1. We are pretty sure that this study truly covers fault tolerance in mobile agents that have been published to date we considered. We have strictly followed Kitchenham et al. [23] guidelines starting from the planning to the reporting of the review results. However, possible threats or weaknesses in this SR could be related to bias in publication, selection of the included studies and the inaccuracy in data extraction.

For publications, we used sources that are credible and trusted by the research community and also conducted trial searches. We also believe publications from 2009 till date that were not considered will not affect the validity of our study since another study will be performed to compare the trends. In addition, though is possible that some relevant papers may have been missed and if they do, we are sure they are not many and their absence has no significant effect to the information reported in this study. Another issue is the search terms we used. If the search terms/strings formulated were not sufficient and effectively utilized, we believe it has
no counter effect on this study. However the two authors worked in collaboration with a librarian and all the selected studies analyzed. In addition, all decisions and results were checked and re-checked.

6. Conclusions

With the increase in system dependability and the exponential growth of the Internet, there is an increasing need to develop reliable mobile agent’s fault tolerance techniques capable of withstanding unfavorable and unpredictable behavior of today’s systems. Several well known approaches and models for failure detection and recovery schemes have been designed, but none is generic. When deciding on how to develop a reliable fault tolerance in mobile agents, it is important to have the knowledge of the existing techniques, and to be able to make a good decision, taking into account the strong and weak points of diverse approaches against each other. To this effect, this SR takes into account the strong and weak points of diverse existing techniques, and to be able to make a good decision, establishing facts in the directions of fault tolerance in mobile agents, recognized techniques/approaches, influencing factors, platform supports and challenges. These were chosen because they together helped in giving a good understanding of the existing findings, identify gaps in existing research and provide recommendations for new research activities.

We have analyzed the existing studies implementations in order to realize the state-of-the-art in mobile agent fault tolerance and trends. In this study, we present the found fault tolerance in mobile agents existing implementations in Table 1. In addition, we have developed taxonomy of the existing protocols and the design model used (Figure 10 and 11). Based on the analysis and results obtained, the study takes a closer look at the available mobile agent’s fault tolerance strategies and the challenges that affects its realization. The cardinal findings are:

- Agent’s fault tolerance fault models generally fall into the three failure types, namely communication, crash and agent software failure though, stated differently in different studies. For instance, place failure, node crash, hardware failure, server failure, host failure; etc were all described as crash failure in the studies.
- Fault tolerance strategies can not address all single-point-of-failures in a system with respect to faults resulting from communication, agent software and crash failures. The complexity and cost of addressing all single-point-of-failures makes the fault tolerance process virtually incomplete. Thus, single-point-of-failures are inevitable in mobile agent and no approach is considered the best in all failure situations.
- Fault tolerance in mobile agent systems can be achieved in a number of ways: replication schemes, check-pointing schemes or a combination of both replication and check-pointing schemes called hybrid approach. However, there is no fault tolerance scheme that is best for all situations since the suitability of an approach heavily depends on the application domain.
- Performance overhead of a fault tolerance strategy is inversely related to recovery time. That is the shorter the recovery time the higher performance overhead. Fault tolerance strategy should try to make a balance between recovery time and performance overhead.
- Mobile agent’s fault tolerance faces enormous challenges such as lack of adequate support from agent’s platforms as well as lack of resource control capabilities which impacts greatly the realization of a reliable agent execution and need to be drastically addressed.
- Fault tolerant mobile agent is gathering momentum without a clear general framework in its agent system. The existing fault tolerance designs are designed to handle a particular set of fault models and not faults in all situations.

Future work includes proposing a general framework for fault tolerance in mobile agents. This will contribute positively to a high level of system dependability and in addressing the challenges and influencing factors affecting the existing fault tolerance models. In addition, we will validate and implement the framework, perform a more in-depth analysis to investigate challenges outside the framework and investigate fault tolerance in other areas.

Based on the above finding, our recommendations are as follows:

- A generic fault tolerance in mobile agents should be designed and developed since it is difficult to measure the completeness of a fault tolerance approach. The fact is that the existing implementations found in this study focus on partial list of failure types, indicating agents cannot tolerate failure of all types. Thus, issues of all single-point-of-failures should be addressed if high reliability or availability is to be achieved.
- Larger scale tests in real world applications other than simulations should be applied to the available fault tolerance schemes so as to better demonstrate their reliability, capabilities and effectiveness. This is because the maturity of fault tolerance approaches depends highly on the level of acceptance of the approaches. The more the approaches are used, the more the approaches evolve and develop and proved that they work.
- The effects of very long itinerary, many collaborative agents, many replicas, many uncommitted transactions, etc during scalability would need to be investigated so as to increase support for the dependability of fault tolerance techniques.
- Existing platforms should be improved by standardizing and including some of the vital fault tolerance protocols such as cloning and resource monitoring in order to reduce interoperability. Moreover, it will be more flexible to detect runaway agents from within the platform than building a
References


[22] O. Kachirski and R. Guha, Intrusion Detection Using Mobile

separate architecture on top of the platforms.
• A better flexibility is if the platforms have a mechanism to selectively apply a fault tolerance protocol to a specific agent or agent place as needed since not all fault tolerance protocols are needed at all times. Developers of agent systems should be able to pick their suitable protocol within their application domain.
• Researchers should focus more on carefully analyzing the unique characteristics of existing agent systems and the applications that are built on them in order to avoid the selection of unsuitable methodology for the recovery of the applications running the agent from faults that affects agents’ execution, migration, and interaction.
• Additional fault tolerance features should be introduced for a better reliability.
• Also, alternative agent’s design approaches should consider minimizing resource consumption such as using existing code in previously visited hosts instead of re-transporting code improve the reusability of code.
• Lastly, developers should have in mind that the application of fault tolerance approach in the real world applications could also introduce a different set of challenges that have never been thought of.


[60] Agents Technology in Europe, ACTS project InfoWin, 1999


