PESARO 2011

The First International Conference on Performance, Safety and Robustness in Complex Systems and Applications

April 17-22, 2011

Budapest, Hungary

PESARO 2011 Editors

Eugen Borcoci, Politehnica University of Bucharest, Romania

Petre Dini, Concordia University, Canada / China Space Agency Center, China

PESARO 2011

Foreword

The First International Conference on Performance, Safety and Robustness in Complex Systems and Applications [PESARO 2011], held between April 17 and 22 in Budapest, Hungary, inaugurated a series of events dedicated to fundamentals, techniques and experiments to specify, design, and deploy systems and applications under given constraints on performance, safety and robustness.

There is a relation between organizational, design and operational complexity of organization and systems and the degree of robustness and safety under given performance metrics. More complex systems and applications might not necessarily be more profitable, but are less robust. There are trade-offs involved in designing and deploying distributed systems. Some designing technologies have a positive influence on safety and robustness, even operational performance is not optimized. Under constantly changing system infrastructure and user behaviors and needs, there is a challenge in designing complex systems and applications with a required level of performance, safety and robustness.

We welcomed technical papers presenting research and practical results, position papers addressing the pros and cons of specific proposals, such as those being discussed in the standard fora or in industry consortia, survey papers addressing the key problems and solutions on any of the above topics short papers on work in progress, and panel proposals.

We take here the opportunity to warmly thank all the members of PESARO 2011 Technical Program Committee, as well as the numerous reviewers. The creation of such a broad and high quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and efforts to contribute to ICDT 2011. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

Also, this event could not have been a reality without the support of many individuals, organizations, and sponsors. We are grateful to the members of the PESARO 2011 organizing committee for their help in handling the logistics and for their work to make this professional meeting a success.

We hope that PESARO 2011 was a successful international forum for the exchange of ideas and results between academia and industry and for the promotion of progress in the areas of performance, safety and robustness in complex systems and applications.

We are convinced that the participants found the event useful and communications very open. We also hope the attendees enjoyed the historic charm of Budapest, Hungary.

PESARO 2011 Chairs:

Miguel Garcia Pineda, Polytechnic University of Valencia, Spain
Juong-Sik Lee, Nokia Research Center - Palo Alto, USA
Chi Harold Liu, IBM Research, China

Liam Murphy, University College Dublin, Ireland
Naoki Wakamiya, University of Osaka, Japan
Piotr Zwierzykowski, Poznan University of Technology, Poland

PESARO 2011

Committee

PESARO Advisory Chairs

Liam Murphy, University College Dublin, Ireland
Naoki Wakamiya, University of Osaka, Japan
Piotr Zwierzykowski, Poznan University of Technology, Poland

PESARO Industry/Research Chair

Juong-Sik Lee, Nokia Research Center - Palo Alto, USA
Chi Harold Liu, IBM Research, China

PESARO Publicity Chair

Miguel Garcia Pineda, Polytechnic University of Valencia, Spain

PESARO 2011 Technical Program Committee

Juan C. Aguero, The University of Newcastle, Australia
Kevin Bauer, University of Colorado, USA
Mehmet Ufuk Çaglayan, Bogazici University- Istanbul, Turkey
Ben-Jye Chang, National Yunlin University of Science and Technology, Taiwan
Harpreet S. Dhillon, The University of Texas at Austin, USA
John-Austen Francisco, Rutgers University - Piscataway, USA
Mina Guirguis, Texas State University - San Marcos, USA
Song Guo, University of Aizu, Japan
Jyri Hämäläinen, Aalto University, Finland
Chi Harold Liu, IBM Research, China
Bilal Khan, City University of New York / John Jay College, USA
Behrouz Khoshnevis, University of Toronto, Canada
Juong-Sik Lee, Nokia Research Center - Palo Alto, USA
Kevin Mills, NIST, USA
Liam Murphy, University College Dublin, Ireland
Asoke K. Nandi, The University of Liverpool, U.K.
Harald Øverby, NTNU, Norway
M. Zubair Rafique, King Saud University - Islamabad, Pakistan
Young-Joo Suh, Pohang University of Science and Technology (POSTECH), South Korea
Zhi Sun, Georgia Institute of Technology, USA
Katarzyna Wac, Carnegie Mellon University, USA / University of Geneva, Switzerland
Naoki Wakamiya, University of Osaka, Japan
Yang Wang, Georgia State University, USA
Yulei Wu, University of Bradford, UK
Gaoxi Xiao, Nanyang Technological University, Singapore
Bashir Yahya, University of Versailles, France
Piotr Zwierzykowski, Poznan University of Technology, Poland

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

Verifying Distributed Algorithms with Executable *Creol* Models

Wolfgang Leister
Norsk Regnesentral
Oslo, Norway
wolfgang.leister@nr.no

Joakim Bjørk and Rudolf Schlatte
Institute of Informatics, University of Oslo
Oslo, Norway
{joakimbj,rschlatte}@ifi.uio.no

Andreas Griesmayer
VERIMAG/UJF
Grenoble, France
andreas.griesmayer@imag.fr

***Abstract*—We show a way to evaluate functional properties of distributed algorithms by the example of the AODV algorithm in sensor networks, *Creol* models and component testing. We present a new method to structure the evaluation work into the categories of techniques, perspectives, arrangements, and properties using executable models. We demonstrate how to use this structure for network simulations and component testing in order to evaluate a large list of properties. We also show which properties are most suited to be evaluated by which technique, perspective, and arrangement.**

***Keywords*—formal analysis; modelling; model checking; testing; routing algorithms**

I. INTRODUCTION

With increasing miniaturisation, computational devices are becoming virtually omnipresent and pose new challenges in software development. We study arising questions on the example of wireless sensor networks (WSN) [1] consisting of spatially distributed autonomous sensor nodes that communicate using radio connections. Each node can sense, process, send and receive data. We concentrate on the verification of a *Distributed Algorithm* for ad-hoc networks between the sensor nodes to route data packets of the participating nodes. There are many requirements for WSN: routing must fulfil properties for quality of service (QoS), timing, delay, and network throughput; furthermore, we are interested in properties like mobility and resource consumption. Autonomous behaviour of the nodes leads to state space explosion during model checking, making evaluation a complex task that requires a combination of techniques from different verification approaches.

In this paper, which is based on a previously published report [2], we present a structured methodology to verification that introduces the categories of *techniques*, *perspectives*, *arrangements*, and *properties*. We combine this novel structure with techniques from simulation, testing, and model checking to create a new, unified method for verification of distributed systems. To this end, we use models in *Creol*, an object oriented modelling language that allows executable models. The novel idea behind our work is to employ one single executable model that is suitable for simulation, testing, and model checking, without the need to develop separate models for each task. We demonstrate the approach by evaluating a large set of properties on a network using the *Ad hoc On Demand Distance Vector* (AODV) routing algorithm [3].

The remainder of this paper is organised as follows: After briefly presenting the *Creol* language and related work, we discuss the AODV model developed previously (Section II), our categories for the validation process (Section III), present results from network simulation and component testing (Section IV), and conclude in Section V.

A. Executable Creol Models

Creol [4], [5] is an *object-oriented* modelling language, which provides an abstract, executable model of the implementation of components. The *Creol* tools are part of the *Credo* tool suite [6] that unifies several simulation and model checking tools. The *Credo* tools support integrated modelling of different aspects of highly re-configurable distributed systems, both structural changes of a network and changes in the components. The *Credo* tools offer formalisms, languages, and tools to describe properties of the model in different levels of detail; these formalisms include various types of automata, procedural, and object-oriented approaches.

To model components, *Creol* provides behavioural interfaces to specify inter-component communication. We use intra-component interfaces together with the behavioural interfaces to derive test specifications to check for conformance between the behavioural model and the *Creol* implementation. Types are separated from classes, and (behavioural) interfaces are used to type objects. *Creol* objects can have active behaviour and are concurrent, i.e., conceptually, each object encapsulates its own processor. During object creation a designated run method is automatically invoked. Asynchronous method calls, explicit synchronisation points, processor release, and under-specified, i.e., nondeterministic, local scheduling of the processes within an object provide flexible object interaction.

Creol includes a compiler, a type-checker, and a simulation platform based on Maude [7], which allow simulation, guided simulation, model testing, and model checking.

B. Related Work

Showing functional correctness and non-functional properties for algorithms employed for WSN helps the developers in their technical choices. Developers use a variety of tools, including measurements on real implementations, simulation, and model-checking. When developing algorithms for packet forwarding in a WSN, simulation results must be compared

 ISBN: 978-1-61208-132-8

with the behaviour of known algorithms to get a result approved [8]. Approaches using simulation, testing, and model checking during the development process use one or more of the following: modelling, traces, runtime monitoring by integrating checking software into the code (instrumentation) [9], or generating software from models automatically [10]. Simulation systems are used to analyse performance parameters of communication networks, such as latency, packet loss rate, network throughput, and other metrics. Most of these systems use discrete event simulation. Examples for such simulation systems include Opnet, OMNeT++, and ns-2 [11], or mathematical frameworks like MathWorks. Many of these tools have specialised libraries for certain properties, hardware, and network types.

The CMC model checker [9] has been applied on existing implementations of AODV by checking an invariant expressing the loop-freeness property. In that work, both specification and implementation errors were found and later corrected in more recent versions of both specification and implementations. CMC interfaces C-programs directly by replacing procedure calls with model-checker code, thus avoiding the need to model AODV. The model checking tools SPIN and *UPPAAL* have been used to verify properties for the correct operation of ad hoc routing protocols [12], such as the LUNAR and DSR algorithms [13]. Both LUNAR and DSR are related to AODV, but use different mechanisms. A timing analysis in *UPPAAL* uncovered that many AODV connections unnecessarily timed out before a route could be established in large networks [14]. Timed automata implemented in *UPPAAL* have been used for validating and tuning of temporal configuration parameters and QoS requirements [15] in network models that allow dynamic re-configurations of the network topology. The model checker *Vereofy* [16], part of the *Credo* tools, has recently been used to analyse aspects of sensor networks and AODV. We also used *Vereofy* as a reference for evaluating the properties and as source for the traces employed for the component testing.

The OGDC-algorithm used in certain sensor networks has been simulated and model-checked in Real-Time Maude [17]. The comparison of these simulation results in Real-Time Maude with simulation results in ns-2 have uncovered weaknesses in a concrete ns-2 simulation.

II. Modelling the Components and the Routing Algorithm

Distributed applications can be described in terms of components interacting in an open environment, based on the mechanisms of *Creol* [18]. This framework models components and the communication between these components, and executes the models in rewriting logic. Different communication patterns, communication properties, and a notion of time are supported. The lower communication layers use tight, loose, and wireless links.

Based on this work, we defined a model of AODV in a WSN using *Creol* [19] that expresses each node and the network as objects with an inner behaviour. The interfaces of the objects describe the communication between the nodes and

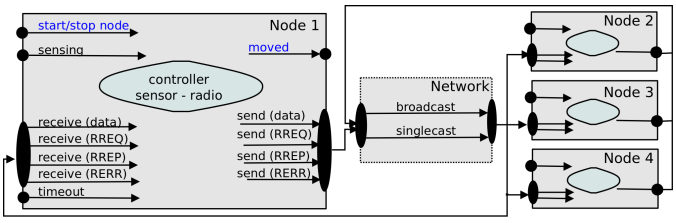

Figure 1. Objects of a WSN model and their communication interfaces.

the network object. In Figure 1, we show the object structure of the model, including the most important interfaces of a node. Within the nodes, its behaviour is implemented in *Creol* as routines that are not unlike real-world implementations.

The purpose of a routing algorithm is to establish a path between a source node and a sink node, so that data can flow from the source node to the sink node. AODV is a reactive routing protocol that builds up the entries in the dynamic routing tables of nodes only if needed. AODV can handle network dynamics, e.g., varying wireless link qualities, packet losses, and changing network topologies.

When a node wants to send a message to a sink node and the next hop cannot be retrieved from the local routing table, it initiates a route discovery procedure by broadcasting RREQ (route request) messages. Nodes that receive a RREQ message will either send a RREP (route reply) message to the node which originated the RREQ message if the route is known; otherwise the node will re-broadcast the RREQ message. This procedure continues until the RREQ message reaches a node that has a valid route to the destination node. The RREP message is unicast to the source node through multi-hop communications; as the RREP message propagates, all the intermediate nodes also establish routes to the destination. After the source node has received the RREP message, a route to the destination has been established, and data packages can be sent along this route.

The essential entries of the routing table include the next hop, a sequence number, and the hop count to the sink node. The latter is the most common metric for routing to choose between routes when multiple routes exist. The sequence number is a measure of the freshness of a route.

When communication failures imply a broken route, the node that is unable to forward a message will inform other nodes, so that the routing tables can be updated. To do this it sends a RERR (route error) message along the reverse route that is also stored in the nodes. Thus the source node will become aware of the broken route, and initiate a new route discovery procedure.

III. Methodology for Simulation, Component Testing, Model Checking

In the following section we show how to evaluate and validate the functional behaviour of the AODV model using *Creol* and the *Credo* methodology [6]. We present the *techniques*, *perspectives*, *arrangements*, and *properties* necessary

 ISBN: 978-1-61208-132-8

for the validation. We also show how to evaluate selected non-functional properties.

A. Techniques for Simulation, Testing, and Model Checking

In order to evaluate the properties of a model, several *techniques* are used to provide the necessary technical measures and procedures to make a model amenable to verification. In general, the following modifications can be applied to the model in preparation for simulation, testing, and model-checking:

Auxiliary variables are added to the model to improve the visibility of a model's behaviour. They must not alter the behaviour and are updated when certain relevant events happen, e.g., a counter is incremented when a new instance is created. When running a simulation these values can be extracted from the state information and visualised in an step-by-step execution, or after the *Creol* model terminates.

Assertions might be necessary depending on the functional requirements to check. While a number of properties can be checked at the final state using auxiliary variables, properties on the transient behaviour of the model require a check during runtime. For such cases, *Creol* provides *assertions* that stop the execution of a model when the condition is violated. The state that caused the violation of the property is then shown for further analysis.

Monitors are pieces of software that run in parallel to the actual model and are used for properties that go beyond simple assertions. A monitor constitutes an automaton that follows the behaviour of the model to decide the validity of a path.

Guarded execution replaces nondeterministic decisions by calls to a guarding object, the *DeuxExMachina*. This allows to check the behaviour of the model under different conditions, while still maintaining reproducibility of the runs. This technique also specifies certain parameters of the environment, like failure rates of the network.

Fault injection adds a misbehaving node (possibly after a certain time) to check error recovery properties. E.g., switch off a node when energy is used up, or inject other errors. This can be implemented by sub-classing nodes and implementing certain misbehaving routines in the subclass.

Property search employs model checking techniques to check whether certain conditions hold for all or a given subset of states. Currently, property search must be specified in *Maude* for a *Creol* model.

B. Perspectives

A *perspective* describes the scope of an evaluation. For the AODV model we developed two perspectives: (*a*) observing the behaviour of the entire network configuration including all nodes and the network; (*b*) observing the behaviour of one node. Testing, simulation, and model checking can be performed from different perspectives and levels of detail for a given model. For AODV, a holistic perspective focuses on the networking aspect of the nodes, implementing all the involved nodes and the network in one model. However, for model checking, such a model leads to a high number of states, and long execution time. Therefore, for realistic models the networking perspective is not feasible.

For the perspective of testing a single node we use the same model code for the nodes in the holistic perspective, but instantiate only one node explicitly. The network is replaced by a *test harness* that impersonates the network and the remaining nodes. The behaviour and responses of the test harness are determined by a rule set that is derived from traced messages between the nodes, as outlined in Section IV-B.

C. Arrangements

An *arrangement* denotes a set of configuration settings that influences how the model operates. Examples for arrangement entities that can be selected in the *Creol* models, together with implementation details for the AODV model are given in the following:

The *communication behaviour* in our model can be set to be either reliable, non-deterministic, or one of several packet loss patterns. Note that non-deterministic behaviour in a simulation currently is not useful due to restrictions in the implementation of the underlying run-time system. Using the differences in communication behaviour, we can study how the algorithm behaves when communication packets can get lost.

Topology changes are used to check the robustness of the protocol. They can be triggered by certain events, e.g., after a certain amount of messages, or after a certain amount of time for timed models. A topology change affects the connection matrix in the network and triggers the AODV algorithm to find new routes in the model.

The *timed model* is realized using discrete time steps and introducing a global clock in the network object and internal clocks in the nodes, which are synchronised when a task is performed in one or more nodes. This allows, e.g., to reason about messages being sent simultaneously, which eventually will lead to packet loss. Also the effect of collisions can be shown without using non-deterministic packet loss. The use of a timed model is most viable together with topology changes since the topology needs to be re-installed for a state when another branch is searched in model checking.

Energy consumption is modelled using an auxiliary variable in each sensor node with an initial amount of energy. For each operation a certain amount of energy is subtracted until the capacity is too low too perform operations on the radio, indicating a malfunction of the sensor node. Such a node does not perform any actions and represents a topology change of the network, since given path are no longer valid. This allows us to identify in which cases an energy-restricted network can perform communication and whether AODV can find routes around an energy-empty node.

Note that arrangements for memory and buffer sizes can be implemented similarly. When maximum memory size is reached, a node will stop to perform certain actions.

Timeouts are modelled nondeterministically by use of a global guarding object and can occur between a message is sent and the corresponding answer is received. AODV employs time-outs in order to work in environments where communication

 ISBN: 978-1-61208-132-8

errors can occur and sends messages repeatedly in case an expected reply has not been received from the network.

D. Properties

A *functional property* is a concrete condition that can be checked for given arrangements, while non-functional properties are values given by metrics. For AODV, we chose the following functional properties: correct-operation, loop-freeness, single-sensor challenge-response properties, shortest-path, deadlock-freeness (both, for node, and for protocol), miscellaneous composed system properties, and non-functional properties.

Correct-operation: For a routing algorithm to be correct, it must find a path if a path exists, i.e., it is *valid* for some duration longer than what is required to set up a route from sender to receiver [12], [13]. Checking this property requires the algorithm-independent predicate whether a route exists. In the absence of topology changes, this predicate can be calculated beforehand. When topology changes are possible, however, we need to check the existence of a path between sender and receiver at any step in the algorithm. Since checking this property in *Creol* involves explicitly visiting all nodes, this increases the reachable state-space of the model. To evaluate this predicate effectively, a suitable implementation would be to interface a *Maude* function. Unfortunately, this is currently not supported by *Creol*.

A related property to evaluate is whether a route is re-established after a transmission error, given a path still exists. We also evaluate how long the path is interrupted after a transmission error occurs.

Loop-freeness: A routing loop is a situation where the entries in the routing tables form a circular path, thus preventing packets from reaching the destination. The invariant for loop-freeness [9] of AODV must be valid for all nodes. It uses *sequence numbers* of adjacent nodes, and the number of hops in the routing tables as input. The loop-freeness property is checked every time a message is transmitted between nodes. To do this, the network-object calls a routine that checks the loop-freeness invariant in an assertion. Since this assertion is complex, and contains nested loops, it should be implemented as a call to a *Maude* function instead of *Creol* code.

Single-sensor challenge-response: The reaction of one node under test is checked using component testing (Section IV-B). Messages are sent to the node under test, and the responses from this node are matched against all correct responses. The correct responses are extracted from specifications or from running simulations using different implementations. The single-sensor properties that can be checked express a certain behaviour of the absence of a certain behaviour after a challenge, e.g., whether an incoming message leads to a specified state change in the node, or whether the node sends an expected response messages.

Shortest-path: Here, we investigate whether the AODV algorithm finds the shortest path for the paths between source and sink node; also other metrics for paths could be checked. While AODV finds the shortest path in the case of no packet loss, it does not always fulfil the shortest-path property in the case of packet loss. To check this property we count the number of hops that each payload-message takes from the source to the sink and compare it with the shortest existing path between source and sink.

Deadlock freedom: Deadlocks in a node, in the protocol or in the model are a threat to robustness, and can reveal errors in the specification, implementation, or model. Deadlocks will make the model execution stop in an error state.

Miscellaneous composed-system properties: Examples are properties that state whether valid routes stay valid, avoidance of useless RREQ messages, number of messages received, timing properties and network connectivity. Most of these are implemented by adding counter variables, and predicates.

Non-Functional Properties: Non-functional properties from the application domain, such as timing, throughput, delivery ratio, network connectivity, energy consumption, memory and buffer sizes, properties of the wireless channel, interferences, mobility, or other QoS properties, can be evaluated by using counter variables and additional *Creol* code.

IV. Holistic and Component Testing

An instrumented *Creol* model can be used for the different verification and testing techniques in the *Credo* methodology: symbolic simulation, guarded test case execution, and model checking. Currently, the auxiliary variables for assertions and the state of the monitors need to be added as *Creol* code that is executed together with the model code. This increases the size of the states and therefore poses a handicap for model checking.

A. Holistic Testing

For our evaluation of the network properties we used simulation using mainly techniques such as auxiliary variables, and assertions. Most of our experiments used a network with symmetrical communication via four sensor nodes and one sink node. We simulated the AODV model using various arrangements using reliable networks, lossy networks, timeouts, energy consumption, and timed modelling. We also checked selected properties from Section III-D.

Reliable communication: As long as the network is connected, the evaluations showed that the modelled AODV algorithm fulfils the properties from Section III-D. We emphasised on the evaluation of packet loss, and loop-freeness assertion. Other predicates for loop-freeness were also used (which failed as expected), and small, faulty changes in the model were introduced (which led to expected failures of the loop-freeness property). The shortest path property was fulfilled in all simulated occasions.

Lossy communication: When simulating lossy communication, both for singlecast, and for broadcast messages the packet loss rate increases as expected. We also observed an increased number of RREQ and RREP messages in the system, using auxiliary variables.

Timeouts: The model allows re-sending of lost RREQ messages up to a certain number of times, using a timeout

 ISBN: 978-1-61208-132-8

TABLE I
NUMBER OF REWRITES AND RUN-TIME FOR SAMPLE ARRANGEMENTS AND PROPERTIES.

#	t steps	energy	loss	timeout	#rewrites	time
	500	–	none	never	$9.4 \cdot 10^6$	17.1s
	5000	–	none	never	$62.8 \cdot 10^6$	114.8s
	500	–	10%	never	$10.7 \cdot 10^6$	19.5s
5	500	–	10%	1/10	$12.1 \cdot 10^6$	22.3s
	500	50	10%	1/10	$8.3 \cdot 10^6$	15.5s
	untimed	50	10%	1/10	$11.6 \cdot 10^6$	17.9s
	untimed	–	10%	never	$32.5 \cdot 10^6$	14.8s
6	untimed	–	10%	never	$90.5 \cdot 10^6$	40.9s
15	5000	–	none	never	$2.7 \cdot 10^9$	31m
30	5000	–	none	never	$24.8 \cdot 10^9$	8h

mechanism. We could observe that this mechanism decreased the packet loss rate, but at the same time does not prevent all packet loss for payload packets.

Energy consumption: Using the energy consumption arrangement we can force a communication failure of certain nodes after some actions. Using this arrangement we can study the re-routing behaviour in detail, including the packet loss rate.

Timed model: Using the timed model we can study the number of time steps needed for sending messages, as well as controlling the number of actions being performed simultaneously. We observed that the packet loss rate is different to the untimed case, which is expected.

Using the timed model, we could observe a model deadlock, which is caused by the way the model is implemented, and certain properties of the current implementation of the *Creol* runtime system. This observation made changes in the model implementation necessary using asynchronous method calls.

The developed *Creol* model was evaluated by using simulation for sample arrangements and properties. The entire model contains about 1600 lines of *Creol* code, excluding comments. After compilation, the resulting code size was about 1050 lines of Maude code, depending on the arrangement. We varied the timing behaviour, the energy consumption, the message loss behaviour, and the timeout behaviour of the model, as well as the number of nodes. The results for the tested cases considering the number of rewrites, and execution time on an AMD Athlon 64 Dual core processor with 1.8 GHz is shown in TABLE I. The timing behaviour, and the number of nodes are the most significant parameters.

While these values may sound high for a simulation system, we emphasise that the purpose of the *Creol* model is to offer one model that is suitable for several perspectives. While the transition from simulation to model checking consists in changing some few Maude statements, the search space during model checking gets combinatorially too high to be viable, already for a low number of nodes.

B. Component Testing of One Node

For component testing, we use one node under test with the same code as for holistic testing. However, we replace the network and all the other nodes using a test harness, as shown in Figure 2. The test is then performed by studying the output messages of a node when given input messages are applied.

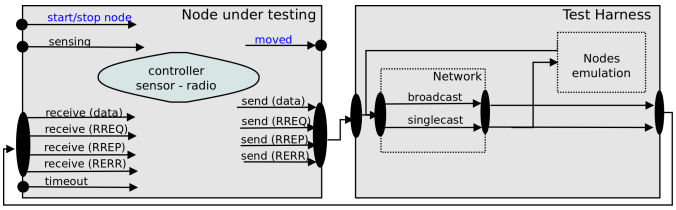

Figure 2. Testing of one node using the network object as a harness.

1) Test harness: The task of the test harness is to send messages to the interfaces of the node under test, and to observe its answers. Both input messages and expected answers can be generated from the specification or from traces of real systems or other simulations.

Although incoming broadcast, singlecast and outgoing packets involve invoking different methods, the *Creol* language, with its object-level parallelism, makes it easy to encode a test case as a single sequential list of statements. Incoming messages are stored in a one-element buffer; the test case simply performs a blocking read on that buffer when waiting for a message from the object under test, before sending out the next message to the object. In that way, both creating a test case by hand and generating test cases from recorded traces become feasible.

A test verdict is reached by running the test harness in parallel with the object under test. If the test harness *deadlocks*, it expects a message from the object under test that is not arriving – in that case, a test verdict of *Fail* is reached. The other reason for test failure is an incoming message that does not conform to the expectations of the test harness; e.g. by being of the wrong type or having the wrong content.

A test verdict of *Success* is reached if the test harness completes the test case and the object under test conforms to the tester's expectations in all cases.

2) Traces: In addition to domain-specific single-object properties test cases can be generated from model implemented with *Vereofy* [16]. To receive the traces from *Vereofy* the content of all variables within the nodes and buffers in the network, before and after each step, and the exchanged data are collected. When the state information is removed we receive a sequence of messages that are exchanged simultaneously.

Traces received from the node under test are tested against *message patterns*, i.e., we remove details that could lead to spurious test failures not expressing a malfunctioning system. For example, the message sequence number can be chosen by the node, the only requirement is that it be monotonically increasing. This property is checked using an invariant in the tester, but a different concrete message number than that used by the *Vereofy* model cannot lead to test failure.

 ISBN: 978-1-61208-132-8

V. Conclusion

We presented the evaluation of a *Creol* model of AODV. We introduced the dimensions of *techniques*, *perspectives*, *arrangements*, and *properties* for this evaluation. We divided the properties used for this evaluation into six property classes, and performed network simulations of the composed system, and component testing of a single node.

Using the network simulation, we evaluated several arrangements. While most most of the properties were fulfilled as expected, some properties did not validate in the simulation, either due to bugs in the model, properties of the modelled AODV algorithm, artificially introduced bugs in the model, or property variants that are not supposed to validate successfully. In one occasion, we could detect deadlocks in the model in a timed-model arrangement, which could be recognised and fixed afterwards. Evaluating other protocols, such as the proactive dynamic routing protocol, is possible, but requires a new model.

Using component testing, we validated the correct behaviour of a single node against properties extracted from the specification of the AODV algorithm. No deviations from specified component behaviour were identified in this process, which is unsurprising since components had already been extensively used for simulation and animation during initial model development at that point in time. However, the test suite served as an excellent help in regression testing during subsequent changes and extensions of the model.

Evaluating the properties of the AODV algorithm, we encountered several challenges, such as modelling suitable abstractions, using language constructs of *Creol*, and observing the properties from a suitable perspective. The major challenge when evaluating the AODV algorithm from a network perspective is to avoid a high number of states in the underlying interpreter. The use of *Creol functions* that can be excluded from the state in model checking would be desirable. This feature is, however, not yet available in the current *Creol* tools.

We found the *Creol* language and the tools useful in the evaluation of the AODV algorithm, and to gain insight in how complex algorithms like AODV work. We observed how small changes in the algorithm, and in the chosen arrangement, imply changes in its behaviour. We also detected the breach of certain properties, which will lead to further investigation of the reasons, removal of this misbehaviour, and, eventually, to a better understanding of AODV, and the algorithms used for sensor networks.

Acknowledgements

This research is part of the EU project IST-33826 CREDO: *Modeling and analysis of evolutionary structures for distributed services*. The authors want to express their thanks to their colleagues involved in the *Credo* project for their support during this work, especially Sascha Klüppelholz, Joachim Klein, Immo Grabe, Bjarte M. Østvold, Xuedong Liang, Marcel Kyas, Martin Steffen, and Trenton Schulz.

References

[1] I. F. Akyildiz, W. Su, Y. Sankarasubramaniam, and E. Cayirci, "Wireless sensor networks: a survey," *Computer Networks*, vol. 38, no. 4, pp. 393–422, 2002.

[2] W. Leister, J. Bjørk, R. Schlatte, and A. Griesmayer, "Validation of creol models for routing algorithms in wireless sensor networks," Norsk Regnesentral, Oslo, Norway, Report 1024, 2010.

[3] C. Perkins, E. Belding-Royer, and S. Das, "Ad hoc On-Demand Distance Vector (AODV) Routing," RFC 3561 (Experimental), Jul. 2003. [Online]. Available: http://www.ietf.org/rfc/rfc3561.txt

[4] E. B. Johnsen and O. Owe, "An asynchronous communication model for distributed concurrent objects," *Software and Systems Modeling*, vol. 6, no. 1, pp. 35–58, 2007.

[5] M. Kyas, *Creol Tools User Guide*, 0.0n ed., Institutt for Informatikk, Universitetet i Oslo, Postboks 1080 Blindern, 0316 Oslo, Norway, May 2009.

[6] I. Grabe, M. M. Jaghoori, B. Aichernig, T. Blechmann, F. de Boer, A. Griesmayer, E. B. Johnsen, J. Klein, S. Klüppelholz, M. Kyas, W. Leister, R. Schlatte, A. Stam, M. Steffen, S. Tschirner, X. Liang, and W. Yi, "Credo methodology. Modeling and analyzing a peer-to-peer system in Credo," in *3rd International Workshop on Harnessing Theories for Tool Support in Software*, 2009.

[7] M. Clavel, F. Durán, S. Eker, P. Lincoln, N. Martí-Oliet, J. Meseguer, and J. F. Quesada, "Maude: Specification and programming in rewriting logic," *Theoretical Computer Science*, 2001.

[8] I. Stojmenovic, "Simulations in wireless sensor and ad hoc networks: matching and advancing models, metrics, and solutions," *IEEE Communications Magazine*, vol. 46, no. 12, pp. 102–107, 2008.

[9] M. Musuvathi, D. Y. W. Park, A. Chou, D. R. Engler, and D. L. Dill, "CMC: a pragmatic approach to model checking real code," in *OSDI, Usenix*, 2002.

[10] M. M. R. Mozumdar, F. Gregoretti, L. Lavagno, L. Vanzago, and S. Olivieri, "A framework for modeling, simulation and automatic code generation of sensor network application," in *Proc. Fifth Annual IEEE Communications Society Conference on Sensor, Mesh and Ad Hoc Communications and Networks, SECON 2008, June 16-20, 2008, San Francisco, USA*. IEEE, 2008, pp. 515–522.

[11] A. Varga and R. Hornig, "An overview of the OMNeT++ simulation environment," in *Proceedings of the 1st international conference on Simulation tools and techniques for communications, networks and systems & workshops*. ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2008, pp. 1–10.

[12] O. Wibling, J. Parrow, and A. N. Pears, "Automatized verification of ad hoc routing protocols," in *FORTE*, ser. Lecture Notes in Computer Science, vol. 3235. Springer, 2004, pp. 343–358.

[13] ——, "Ad hoc routing protocol verification through broadcast abstraction," in *FORTE*, ser. Lecture Notes in Computer Science, vol. 3731. Springer, 2005, pp. 128–142.

[14] S. Chiyangwa and M. Z. Kwiatkowska, "A timing analysis of AODV," in *FMOODS*, ser. Lecture Notes in Computer Science, vol. 3535. Springer, 2005, pp. 306–321.

[15] S. Tschirner, X. Liang, and W. Yi, "Model-based validation of QoS properties of biomedical sensor networks," in *EMSOFT '08: Proceedings of the 7th ACM international conference on Embedded software*. New York, NY, USA: ACM, 2008, pp. 69–78.

[16] C. Baier, T. Blechmann, J. Klein, S. Klüppelholz, and W. Leister, "Design and verification of systems with exogeneous coordination using Vereofy," in *Proc. 4th Intl. Symp. on Leveraging Applications of Formal Methods, Verification and Validation (ISoLA 2010), Part II*, ser. LNCS, vol. 6416. Springer-Verlag, Oct. 2010, pp. 97–111.

[17] P. C. Ölveczky and S. Thorvaldsen, "Formal modeling and analysis of the OGDC wireless sensor network algorithm in Real-Time Maude," in *FMOODS*, ser. Lecture Notes in Computer Science, vol. 4468. Springer, 2007, pp. 122–140.

[18] E. B. Johnsen, O. Owe, J. Bjørk, and M. Kyas, "An object-oriented component model for heterogeneous nets," in *FMCO*, ser. Lecture Notes in Computer Science, vol. 5382. Springer, 2007, pp. 257–279.

[19] W. Leister, X. Liang, S. Klüppelholz, J. Klein, O. Owe, F. Kazemeyni, J. Bjørk, and B. M. Østvold, "Modelling of biomedical sensor networks using the Creol tools," Norsk Regnesentral, Oslo, Norway, Report 1022, 2009.

 ISBN: 978-1-61208-132-8

Identifying Software Hazards with a Modified CHAZOP

Bernhard Hulin
Deutsche Bahn AG
DB Systemtechnik
Munich, Germany
E-mail: bernhard.hulin@deutschebahn.com

Rolf Tschachtli
Deutsche Bahn AG
DB Fahrzeuginstandhaltung
Munich, Germany
E-mail: rolf.tschachtli@deutschebahn.com

***Abstract*—CHAZOP is one of the most popular methods for identifying hazards of software. However, the classical HAZOP methodology as well as the CHAZOP methodology has four technical insufficiencies when applied to software: Ambiguity, incompleteness, nonsensicality and redundancy of HAZOP expressions. This present paper shows a modification of CHAZOP to overcome these insufficiencies. The reasons for these insufficiencies are a non-specified HAZOP language and missing guide words. We therefore, define a HAZOP language and identify missing guide words. The definition of the language is based on the items: Actions, objects, and their attributes. In contrast to the classical HAZOP, the modification defines rules for combining these items with guide words. One of the key ideas of the language is to use HAZOP parameters twice whenever possible: As objects and as attributes. In practice, this means that an attribute is additionally analyzed as if it were a software variable. We call this concept manifestation since in our new method attributes are also manifested in variables. For evaluation, the modified method is compared with the traditional one with the example of a safety-relevant software-controlled system using the windows registry. By means of this example, it is shown that more hazards can be found with the modified CHAZOP than with traditional method.**

Keywords - HAZOP; deviation; parameter-manifestation; hazards.

I. INTRODUCTION

HAZOP is one of the most widely used techniques for the identification of hazards in the production and operation of technical systems. Coming from the chemical industry [1] [2] originally, HAZOP has been adapted to other industrial areas. Later it was adapted for different business areas such as Computers [3][4][5], where it is known as CHAZOP.

The idea behind HAZOP is to combine parameters with guide words to gain indicators of possible failures of a system (we call these combinations HAZOP expressions), then to formulate interpretations of these HAZOP expressions within a team, to extract deviations of the set of interpretations (since a few interpretations are not deviations but are desired), and finally to extract hazards out of the set of deviations [15][16]. HAZOP parameters can be system components, their attributes, as well as actions and their attributes.

We use HAZOP and CHAZOP, respectively, for the identification of hazards induced by the software of railway vehicles. We use these hazards as inputs for the risk assessment [11][12], resulting in a SwSIL classification [17]. These SwSIL classifications are demanded by law for new or modified software used in railway vehicles. They are to be performed by assessors accredited by the German Federal Railway Authority.

Although HAZOP is the most widely used technique for the identification of hazards it has three drawbacks as mentioned by several authors: Amount of time, high costs and safety-gaps. For reducing costs and saving time several authors suggest an electronic system for the management of deviations [7][13]. However, even if the management does not cost any time at all, the expenditure of time for HAZOP meetings remains almost equally high. Our experience in software assessment for railway vehicles is that HAZOP meetings for SwSIL classification for one railway component last about 1.5 days with an average of 4 participants thus an average effort of 6 man-days is spent for this step. Depending on the application, we integrate persons covering the following roles: Operator (or user), maintenance manager, project manager, rollout manager and software developer. Sometimes it is necessary to integrate other roles such as experts for other relevant aspects like fire resistance or EMC or experts that have knowledge about interacting components. Based on our experience, the minimum duration for HAZOP of system modifications is about 4 days.

The efficiency of the meeting can be increased on the one hand by building more meaningful HAZOP expressions, which do not need to be interpreted, and on the other hand by not generating useless HAZOP expressions. These two problems can also be found in literature (see [5] p. 55, [6] pp. 73, 74 and [7] p. 68), but a solution for them has not been shown in literature. Moreover, we found in our meetings that ambiguities of HAZOP expressions can lead to missing

 ISBN: 978-1-61208-132-8

hazards, since there is no proof that all possible interpretations of an ambiguity have been observed.

To overcome these insufficiencies, we modified the generation of expressions. With this modification it is possible to generate meaningful and unambiguous HAZOP expressions manually or alternatively automatically by software. Each HAZOP expression then corresponds to exactly one deviation.

With this modification it is possible to find more deviations and thus more hazards than with the classical generation of deviations, and the workload can be transferred from meetings to the office, which saves manpower. In contrast to earlier publications [10], this paper enhances the modification with concepts necessary for hazard identification in software. The core of this novel concept is developed within Tschachtli's master thesis [14] and considers each HAZOP parameter twice: First as a traditional HAZOP parameter and second as a model expressed in software. For attributes such as pressure or velocity it is a manifestation into an object.

The decision to modify HAZOP instead of developing a new technique or modifying another technique was based on the fact that we haven't found any other method essentially different to HAZOP that performs the identification of hazards in such a structured way.

The paper first describes the traditional method as well as its insufficiencies. Then these insufficiencies are examined in detail. From this examination activities are conducted. They result in new guide words and a new procedure of HAZOP analyzes. We conclude with an application of this modified method.

II. RECENT METHOD

As the initial input for HAZOP we use our knowledge, descriptions of the component, and checklists of former assessments. The descriptions should describe functions of the component, interfaces to other components, its input and output, the internal structure (or architecture), operational conditions, and maintenance modes. Operational conditions should answer at least the questions where, when, how, by whom, and how often the component is used. Information about the output and the interfaces is important to evaluate the effect on other components. For example, a component which does not have any safety-related functions can be cabled to a vehicle bus, which is used for the transfer of safety-related data. For chemical plants, an overview of aspects that have to be contained in a description can be found in [7].

With the use of checklists of former assessments, some HAZOP expressions do not need to be discussed in detail or at all in a meeting. HAZOP expressions are classically generated by combining HAZOP guide words pair-wise with HAZOP parameters. As HAZOP parameters we use objects, attributes of objects and actions (e.g., system functions or user operations). The HAZOP parameters are extracted from documents (descriptions and former assessments), knowledge and discussion. Each HAZOP expression is then analyzed under environmental conditions and operational states.

The set of objects consists of system components, such as trains; subcomponents; subjects, such as humans; and environmental objects, such as tunnels and bridges. Subjects can be members of special groups of persons such as train-drivers, conductors, passengers, disabled persons, or children.

Attributes are attributes of these objects. For example, they contain electrical current, velocity of the train, contrast of the display and so on.

Actions are taken of actions of objects and actions of humans. The main actions, which are always taken into consideration in our assessments, are safety-relevant actions that can be performed by a train or the train driver.

Then the HAZOP expressions are generated and are considered within different scenarios. The set of scenarios contains each operational mode, such as passenger transportation, cargo transportation, cleaning mode, maintenance mode, test mode and so on, locations where this train can be, such as Germany, Austria, France, tunnels, bridges, elevation of track, radius of curves and so on, and hazard-modes of the train, such as onboard-fire, onboard smoke, loudness of noise, and failure of different components.

Weather conditions, such as temperature, fog, rain, snow, and so on are only taken into account as far as the software has to react on it. This is normally the case in displaying software [8] where the contrast may not be enough or on air conditioning software.

A very important analyzes is the change of scenarios where the train transits from one to another scenario. This is for example often the case on European country borders where nearly each country has its own electrical current and train control system.

III. INSUFFICIENCIES OF HAZOP

There are four methodological and technical insufficiencies of HAZOP with respect to its HAZOP expressions. These are ambiguity, incompleteness, nonsensicality, and redundancy of the HAZOP expressions. They result from the assignment of HAZOP expressions to interpretations. Therefore, exactly the four mentioned insufficiencies exist (see Figure 1).

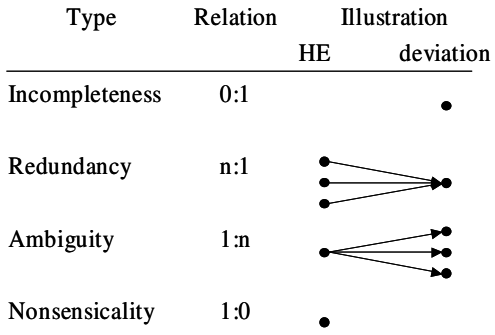

Figure 1. Insufficiencies of HAZOP. How many HAZOP expressions (HE) can be associated with how many deviations.

The danger in ambiguity and incompleteness of HAZOP expressions are missing hazards and thus reduced safety. Nonsensicality and redundancy of HAZOP expressions just results in meetings being disrupted and lasting longer. Therefore, obviation of redundancies and nonsensical HAZOP expressions would improve the method, but this is not safety-critical. Of course, ambiguity also indirectly results in meeting lasting longer, since interpretations have to be found and discussed.

Although three of the four insufficiencies – ambiguity, nonsensicality and redundancy – are known in literature [5] (page 55), [6] (pages 73, 74), [7] (page 68), the problems have not yet been analyzed in depth and no corrective was suggested.

A. *Ambiguity*

Ambiguities arise because the context of the HAZOP expressions is interpretable. There is no one-to-one relation between a HAZOP expression and a deviation. For example the HAZOP expression “no pressure” with the HAZOP parameter “pressure” and the guide word “no” can be interpreted as “no pressure measurable”, “no pressure displayed”, “pressure = 0 Pa” or “there is no variable pressure within an ini-file”.

A second example is the HAZOP expression “more pressure”. Traditionally, this is interpreted as “more pressure than expected”, “more pressure than specified,” or “more pressure than outside the cylinder”. Even the interpretation “more pressure than expected” is ambiguous. It can either mean that the real pressure is higher than expected or that the software variable has a higher value – or both.

B. *Incompleteness*

With traditional HAZOP certain failures and thus hazards cannot or can barely be associated with any HAZOP expression. The conclusion is that the set of HAZOP expressions is not exhaustive and not detailed enough.

For example, each railway train has an identifier such as “de-484-22a-1”. The HAZOP expression “identifier other” can result in a hazard such as train not being reachable. I would be reasonable sure, however, that you have not identified the special deviation “identifier of train 1 is equal to identifier of train 2”. This could be critical if two trains are coupled or have to be dispatched at a central local display within the same district. Certainly, this is a very special case of the HAZOP deviation mentioned but it is very hard to build on this base.

However, the HAZOP expression “identifier other” includes the special case “identifier contains a blank”, too.

C. *Nonsensicality*

Nonsensical HAZOP expressions are also results of the arbitrary combination of each HAZOP parameter with each HAZOP guide word, without considering the context or reasonability of this connection. Examples for this kind of HAZOP expression are “tree early” or “name higher”. Nonsensical HAZOP expressions cost time and even nerves.

D. *Redundancy*

For the redundancies, the reason is similar to the nonsensical expressions. In some HAZOP expressions, the same statements can show up multiple times, because the meaning of the statement is identical. For example, the HAZOP expression “pressure other” includes both “pressure larger” and “pressure smaller”. If you have more than one surname, please make sure that the Volume Editor knows how you are to be listed in the author index.

IV. REASONS OF INSUFFICIENCIES

The main reason for the insufficiencies mentioned is that the HAZOP methodology should induce and support human thinking and interpreting. HAZOP is made for suggesting directions for human thinking with respect to deviations of a system. Therefore, overcoming the insufficiencies mentioned means limiting the degree of interpretation. The possibility to interpret HAZOP expressions is based on at least four aspects.

- Missing ambiguity differentiation for HAZOP parameters
- Unrestricted combination of guide words and HAZOP parameters
- Missing interrelations between HAZOP parameters
- Missing guide word

HAZOP parameters can be ambiguous on their own. One instance of this ambiguity is based on natural language. A prominent example is the data bus within a passenger bus, where “bus” has two meanings. The second kind of ambiguity is based in the dualism between real things and their model. For example the concentration of a gas is a real world attribute while the variable “concentration” within a software program is an object.

Nonsensicality and redundancy of HAZOP expressions are reasoned in unrestrictedly combining guide words and HAZOP parameters. A restriction in combinations, for example, can be that for certain HAZOP parameters the guide word “other” is used, whereas for others the guidewords “less” and “more” are used.

As we found in our HAZOP meeting, missing comparisons between two or more HAZOP parameters result in incompleteness of the set of hazards. Usually, HAZOP expressions are interpreted as a comparison with expectation, e.g. “larger than” “expected”. An analysis has to be done about which HAZOP parameters can be compared, and with which operators they can be compared. In contrast to [10], we widened this topic from comparison to relations.

One reason for undetected deviations is missing guide words. The challenge is to find missing guide words without introducing lots of new guide words.

V. CONCEPT FOR IMPROVING HAZOP

Our strategy for eliminating these reasons is first to differentiate HAZOP parameters and concretize their meanings. After that, rules for combining guide words and HAZOP parameters are generated. This step also focuses on interrelations between HAZOP parameters. Finally missing guide words are added. The idea here is to compare guide

 ISBN: 978-1-61208-132-8

words of HAZOP with the set of mathematical operators, and to add those operators, which are missing.

A. *Differentiation of HAZOP Parameters*

HAZOP parameters can be separated into three different kinds: Objects, actions, and attributes (see Figure 2). The set of objects consists of material and immaterial objects. Examples of material objects can be found in Section 4. Immaterial objects are for example source code (among other object with variables), files, and processes. Actions are discussed elsewhere [18].

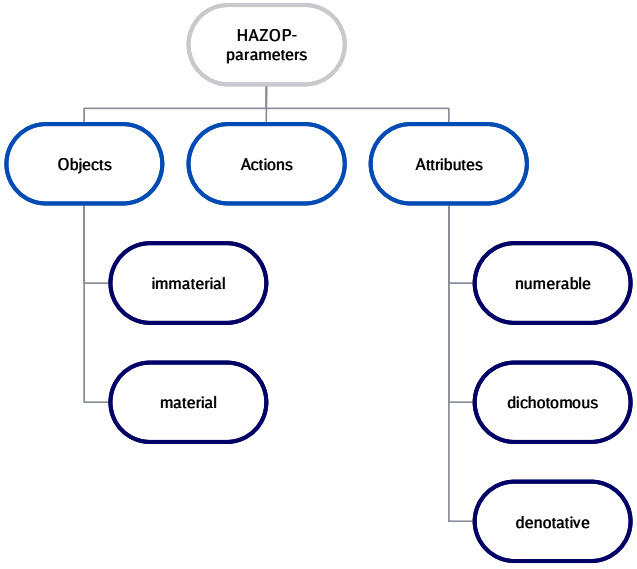

Figure 2. Types of HAZOP parameters.

Attributes belong to objects and can be of type numerable, dichotomous, and denotative. Examples for denoting attributes are names, telephone numbers, characteristic curves of sensitivity of a chip, or charts of shares. Numerable attributes with units for instance are the size in kilograms or liters, while numerable attributes like the amount or degree of capacity might only have a reference object or reference attributes. Dichotomous attributes are like the attribute of a CD – this can be only writable or not writable. We give each dichotomous attribute a name such that it can only have the values true or false.

Up to now, it is still unclear if a HAZOP parameter refers to an attribute or to an immaterial object – for instance, a software variable. For differentiation, we preface each HAZOP parameter an identifier that refers to the type – for example, "variable pressure" and "attribute pressure". A few immaterial objects of software refer to real attributes of objects. In the case of variables, a real attribute is modeled into a variable. This modeling of an attribute into an immaterial object – for instance, a variable – is termed manifestation by us.

Since we do not know a priori if a certain HAZOP parameter is just a real attribute or is implemented as a variable too, we assume in the beginning that each attribute is also implemented as a variable, and thus put them as objects into our list of objects. On the one hand, this induces a duplication of the amount of analyzable items, but on the other hand, this procedure reduces the probability of omitting deviations and thus hazards.

Some points about manifestation are worth noting. One of the most important things is that manifested objects are regular objects. They have attributes and consist of partition objects. In the case of variables, partition objects can be a data-type identifier, the name of variable, and the content of the variable. For example, in Figure 3 where the object "compression control system" is divided into the partition-object "pump" and "control software", the attribute pressure is manifested in a variable pressure.

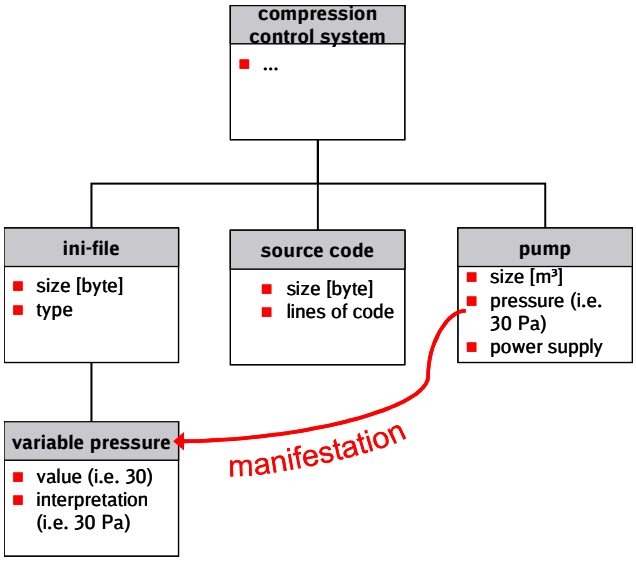

Figure 3. Example of manifestation: The attribute pressure is manifested in a variable pressure that is an object

Attributes of variables are for example, value and interpretation of value. For example, the value "30" of a variable pressure can have the interpretation "30 bar" or "30 Pa". The desired case is that the interpretation of a variable is equal to the real attribute. Note, that attributes of variables are sometimes manifested in other variables.

The second important thing is that attribute type and variable type are independent of each other. For example a numerable attribute is not necessarily a numerable variable but can be modeled in a string, too.

B. *Rules for Generating HAZOP Expressions*

In [10], we analyzed which types of interpretation of HAZOP expressions are in traditional HAZOP. For that, we set up a matrix with a column for each analyzed HAZOP parameter and a row for each HAZOP guide word. Clouds of guide words such as {no, none, not, never} and {reverse, inverse, opposite} were split. HAZOP guide words were collected from publications [1][2][3][4][5][6][9]. HAZOP parameters were taken from recent assessments as well as

 ISBN: 978-1-61208-132-8

from fictive examples. Each of the HAZOP parameters was differentiated with respect to its type. Resulting expressions were characterized by logical, structural, and mathematical aspects.

From that analysis we identified the following types of interpretations of HAZOP expressions in the traditional HAZOP.

1) statement of existence: This kind of expression states that an object does not exist or partially exists. The statement consists of an object and a quantifier or qualifier.

2) comparison of attribute to a special value: This type of statement is received by HAZOP expressions like "no pressure", "not nice" and "no country", which is interpreted as "pressure = 0", "niceness = false" and "country = {}".Type (2) and (3) can be identical but often are not. Special values are software specific values that can cause problems in each kind of software.

3) comparison of attribute to expectation: The difference to a comparison of to a special value is that the expectation is project-specific.

4) junction of two statements of existence: This type occurs if objects are used with the guide words "as well as" and "instead of". Examples are "bus A as well as something else", which can be interpreted as "bus A exists and something else exists".

5) junction of two comparisons of attributes: Here, the guide words "as well as" and "instead of" are used in combination with attributes. For example "beauty as well as something else" can be interpreted as "beauty = true AND another attribute is true ".

C. Completing Guide Words

As mentioned above, interrelations between HAZOP expressions are missing. There can theoretically be the following types.

- Relations between two attributes
- Relations between two objects
- Relations between an attribute and an object

For statements of existence we extend the set of guide words by "completely" and "multiple times". The extensions have the meaning "against expectation one object (already / still) (completely) exists" and "against expectation multiple objects of the same type exist". Sometimes it also makes sense to add "multiple times partially", which is to be interpreted as "against expectation multiple objects partially exist".

Guide words for comparisons of attributes with a special value, with expectation and with another attribute are "=" and "≠". For comparing numerable attributes "<" and ">" can be used, too. For dichotomous and denotative attributes "<" and ">" do not make sense. Moreover, in a few applications for numerable attributes the comparators "l" and "ł", which mean "divides" and "does not divide" can be useful as well. With these comparators one can verify if an attribute is multiple times another attribute or not.

Guide words for the interrelation of objects are "is (not) partition object of" or "is (not) contained in" or "contacts". A specialization of this case is the relation of an object to a predefined special object. This special case is very useful for the analysis of software variables. With the help of this specialization we are able to check a string for blanks, tabulators, quotation marks or other special characters which are often troublesome.

A good example for a relation between an attribute and an object is "manifested in".

The junctions mentioned of rule 4 and rule 5 are very narrow since they join a statement with "something else". They have to be generalized in such a way that two statements can be combined without limitation. However, this problem will be a topic of future papers.

VI. RESULTS

To put the new method into competition with the traditional HAZOP we chose one application for hazard identification where we applied the traditional method and afterwards the new modified method. This order of applying the methods was chosen since we believe that the new method is more complete than the traditional one.

As the application we used a system for displaying the electronic schedule with speed information (including speed reduction intervals) to the train driver. Information displayed to the train driver is in most cases safety-critical.

Some configurations of this system are configured in configuration files like the Windows Registry, ini-files, and so on. Consequently, false configurations within these files can affect information displayed. For an overview of possible failures, see [8].

Entries of the registry are modified remotely by wireless data exchange via the mobile network. Normally this happens if a new version of the system is remotely installed or a new functionality is to be enabled.

Our task is to rate the entries of the registry file according to their safety impacts. Thus, for each entry we have to identify the hazards, the probability of occurrence, the severity of the consequences, the probability of detection in case of occurrence and the chance to escape the critical situation. The result of this analysis is a safety integrity level [19].

With the new method in relation to the conventional HAZOP we have identified the following additional hazards:

- Registry value can be of greater size than the available hard disk space
- Values do not exist
- The data type of a value can change via a software update
- Wrong upper and lower cases inside the value data or the value name

The first two hazards are coped by checking for the completeness of the transferred files and entries and default values set in the program. The third hazard was fought by the cross check and constraining the SIL classification. The constraint is that each change of type has to be reassessed.

 ISBN: 978-1-61208-132-8

Upper and lower cases are not a problem since the program is case-insensitive.

Therefore the SIL classification by the traditional method does not have to be changed. The consecution of SIL can be understood as a good quality of our former work. On the other hand the additional hazard identified is an argument for the quality of the new method.

VII. CONCLUSION

In this paper, we showed an approach to overcome the insufficiencies of hazard identification with HAZOP in the area of software. We introduced the concept of manifestation and added missing types of deviations by adding interrelations and completing the set of operators. The new concept, presented in this paper can eliminate the insufficiencies of HAZOP. We showed the improvement with a railway example.

Although we just described one application of the modified HAZOP our method has been prooven as good and practicable in other projects, too – such as the SwSIL classification of a power-transformation unit and a diagnostic unit for detecting wheel defects. Of course our method also has some limitations. The limitations are that in the preparation phase it is nearly impossible for one person to figure out each important attribute of an object. This is explained by the fact that each object has infinitely many attributes. This limitation is not a special feature of our method but adopted by HAZOP. Thus this issue is not worse in our method than in HAZOP.

Further steps of our work are adding functions, actions, and events to our concept. The guide word application is probably different for them. Furthermore, combinations of expressions have to be analyzed. Moreover, the new method also has to be evaluated with more applications with respect to time consumption.

ACKNOWLEDGMENT

We want to thank Jenny Schulze for her excellent input during her time as master student, and Dr. Dirk Leinhos for enabling and supporting the progress of improving our work as assessors.

REFERENCES

[1] Chemical Industries Association, “A Guide to Hazard and Operability Studies”, 1977.

[2] T. A. Kletz, “Hazop and Hazan: Identifying and Assessing Process Industry Hazards”, 4th edition, Institution of Chemical Engineers, Rugby, UK, 1999.

[3] J. Love, “Process Automation Handbook – A Guide to theory and practice”, Springer Verlag, Berlin, 2007.

[4] S. Mannan, “Lees loss prevention in process industries – hazard identification, assessment and control”, Vol. 1, Elsevier, 2004.

[5] T. Kletz, P. Chung, E. Broomfield, and C. Shen-Orr, “Computer Control and Human Error”, Instn.of Chem.Enginrs, 1995.

[6] F. Redmill, M. Chudleigh, and J. Catmur, “System Safety – HAZOP and Software HAZOP”, Wiley and Sons Ltd., Chichester, U.K, 1999.

[7] I. Faisal, F. Khan, and S. A. Abbasi, “Towards automation of HAZOP with a new tool EXPERTOP”, Environmental Modelling & Software, no 15, Elsevier, pp. 67-77, 2000.

[8] B. Hulin and T. Schulze, “Failure analysis of software for displaying safety-relevant information”, in Reliability, Risk and Safety: Theory and Applications, Taylor and Francis Group, pp. 1327-1331, 2010.

[9] M. Rausand, “HAZOP - Hazard and Operability Study”, Norwegian University of Science and Technology, 2005, www.caia.co.za/files/Hazop_Technique_MarvinRausand.pdf, last access 27th January 2011.

[10] B. Hulin and R. Tschachtli, “Generating unambiguous and more complete HAZOP expressions”, in Reliability, Risk and Safety, Taylor & Francis Group, London, pp. 1-7, 2010.

[11] M. Geisler, „Betriebliche und technische Risiken managen“, Deine Bahn, 10, 2010, pp. 9-14.

[12] B. Milius, “A new classification for risk assessment methods”, in Proceedings of 6th Symposium FORMS/FORMAT 2007, Jan. 2007, pp. 258 – 267.

[13] S. A. McCoy, S. J. Wakeman, F. D. Larkin, M. L. Jefferson, P. W. H. Chung1, A. G. Rushton, F. P. Lees, and P. M. Heino, “HAZID, A Computer Aid for Hazard Identification: 1. The Stophaz Package and the Hazid Code: An Overview, the Issues and the Structure”, Process Safety and Environmental Protection, Volume 77, Issue 6, Nov. 1999, pp. 317-327.

[14] R. Tschachtli, „Entwurf einer Methode zur Identifikation von Fehlermöglichkeiten bei der Entwicklung von Softwarekomponenten“, Master Thesis, Fachhochschule Bingen, Oct. 2009.

[15] Ministry of Defence, “HAZOP Studies on Systems Containing Programmable Electronics”, Defence Standard 00-58, Issue 2, Part 1 and 2, May 2000.

[16] International Electrotechnical Commission, “Hazard and operability studies”, BS IEC 61882, Aug. 2001

[17] CENELEC European Committee for Electrotechnical Standardization, “Railway applications – Communications, signalling and procession systems – Software for railway control and protection systems”, EN50128, Nov. 2000.

[18] J. Schulze, “Improvement of Hazard Identification in Railway Software”, Master Thesis, Chalmers University of Technology, Sep. 2010.

[19] International Electrotechnical Commission, “Functional safety of electrical/electronic/programmable electronic safety-related systems”, IEC 61508, Edition 2.0, Apr. 2010.

 ISBN: 978-1-61208-132-8

Performance-oriented Adaptive Design for Complex Military Organizations

Liu Zhong[+], Huang Jincai[+], Tan Yuejin[+], Wang Chaoyang[+], Ma Jianguang[++], Yang Guoli[+]
[+]School of Information System and Management, National University of Defense Technology
[++]School of Training, National University of Defense Technology
Changsha Hunan, China
e-mail: phillipliu@263.net, {huangjincai,yjtan,cywang,jgma}@nudt.edu.cn, greenyoung@126.com

***Abstract*—Traditional military organizations are designed mainly based on the missions, categorized as mission-oriented design. The paper proposes a new framework that revises the design process, aiming at the organization performance, which comes from the new semantic model eFINC and performance metrics. First, a complete model of military organization is proposed, i.e., eFINC, which extends semantic contents of functional units in complex military organizations, provides the formulization method for the nodes, edges and visual representation. Secondly, the performance metrics of military organizations are defined for eFINC model normatively. The metrics are classified as Response Speed, Coordination Capacity, Execution Capacity and Information Support. And then, the adaptive design model is proposed based on the eFINC model and metrics. Two design strategies are introduced which will lead to a high-performance design of military organizations. Then, Performance Rate is defined as the main reference for the adaptive organization design. The adaptive design procedure for military organizations is illustrated in detail. Finally, the practical case study is conducted to demonstrate the effectiveness of our model.**

Keywords- Complex Military Organization; eFINC model; Performance-Oriented;Adaptive design

I. INTRODUCTION

A. Motivation

The term C4ISRK is used by the US army to refer to the complex systems to carry out missions by military forces [1, 6], can also be viewed as a 'super-system' comprised of varied functional units that are themselves complex, interacting with each other to achieve the common shared goals of military systems [3]. The structure of traditional C4ISRK is typically hierarchy. With the development of networking and computer technology, more and more varied modes of C4ISRK system structures come into being. Meanwhile, the ideas of modern military operations heavily rely on more flexible and more robust structures, such as Network Centric Warfare (NCW) [2]. Network Centric Warfare is becoming the major type of wars, which focus on overall performance of military systems instead of that of individual component for single task.

Military organizations can be viewed as a subset or 'overlay network' of C4ISRK system, which are designed to execute some mission over them. Given a fixed C4ISRK system, how to design an efficient military organization to meet the need of some mission is a challenging problem. Military organizations are assigned to accomplish varied tasks. The challenges that military organization designers are facing upon are how to describe their structures, and how to analyze their performance under uncertain and changing environment. Levchuk had proposed the normative design of task-based organization in the way of three-phase process [7,8,9]; however, the effort is insufficient for organization design, especially in the networking environment. All the subsequent improvements for Levchuk's work are based on the similar ideas [8, 9]. These works can be categorized as mission-oriented design. Military organizations with mission-oriented design strategies are difficult to adapt to the complex and volatile external changing. These organizations will reduce the adaptability of the structure while only aiming at the pursuit of mission efficiency.

It has become an essential need to make design for adaptive military organizations to achieve the high performance for military missions. It have coming into being that the military organizations with adaptive features will play an important role in the war. Traditional organizations are designed mainly based on the organization efficiency. The paper proposes a new framework that revises the design process, aiming at the performance, not just the task efficiency of the corresponding organizations.

B. Related Work

Many researchers have conducted much research work in this field. Anthony put forward four principles for the evaluation of NMO architecture based on Social Network Analysis (SNA) and FINC methods [5, 16, 17]. SNA performs network analysis of relations between individuals within the organization, and is originally inspired from graph theory, and is often applied in military organizations, sociology and anthropology. FINC (Force, Intelligence, Networking and C2) method is used to evaluate effectiveness of different organization architecture, the evaluation metrics include information delay, collaboration delay, intelligence factor and other indices. FINC method proposes some ideas of modeling and analysis of military organizations, but has some difficulties to describe the formal characteristics of military organizations. Anthony also researched the relationship

 ISBN: 978-1-61208-132-8

between the robustness and organization structure based on FINC model [5, 16, 17]. Jeff has also researched the problems of distributed networked operations, building the networked models for military organization [13]. But these models do not have the abilities of performance analysis over the organization structures.

Kathleen have put forward a PCAN method to model C2 organization by using the network form [4, 12]. PCAN model is consisted of multiple networks, but each network is isomorphism (All the nodes are of same types.), and each network is deterministic (all the nodes are connected or disconnected). The work of Kathleen focuses on analysis for networked features of military organizations. The analysis is independant to the design of military organizations.

Levchuk had proposed the normative mission-based design of organization in the way of three-phase process He presents a design methodology for synthesizing organizations to execute complex missions efficiently. It focuses on devising mission planning strategies to optimally achieve mission goals while optimally utilizing organization's resources. Effective planning is often the key to successful completion of the mission, and conversely, mission failure can often be traced back to poor planning. First, corresponding to this framework, military organizations are designed based on highly abstracted system models, without considering constrains of basic characteristics of existing C4ISRK systems. Secondly, Levchuk has provided only simple metrics to evaluate the performance of target organizations, i.e., task accomplishment time and so on. Jincai and Baoxin has researched other metrics to measure the performance, but the improvement can be viewed as the extension of time-based metrics [10, 11, 15]. The effort of these works is insufficient for performance-oriented design.

C. Our Contributions

By extending the FINC model, this paper provides a new approach for organization performance evaluation and builds a new performance-oriented design methodology. Our contributions of this paper are following:

(1) A complete model of military organization is proposed, i.e., eFINC, which comes from the FINC model. Contrary to the traditional FINC model, eFINC extends semantic contents of functional units in organizations, provides the formulization method for nodes, edges and visual representation.

(2) The performance metrics of military organizations are defined for eFINC model normatively. The metrics are classified as *Response Speed*, *Coordination Capacity*, *Execution Capacity* and *Information Support*. Contrary to these of FINC model, the metrics are systematic and meaningful.

(3) The adaptive design model is proposed based on the eFINC model and its metrics. First, two design strategies are introduced which will lead to a high-performance design of military organizations. Then, *Adjusting Value (AV)* is defined, which implies the performance rate, as the main reference for the adaptive organization design. At last, the adaptive design procedure for military organizations is illustrated in detail.

(4) Finally, the practical case study is conducted in Section V to demonstrate the effectiveness of our model.

II. THE EXTENDED FINC MODEL FOR COMPLEX MILITARY ORGANIZATION

The traditional FINC model is leveraged to describe the military organizations consisting of force units, intelligence units, networking units and C2 units. As we mentioned before, FINC model is constrained with its semantic representation for nodes and edges. Based on the graph theory, the paper extends the semantic expression of varied functional units in the military organizations and provides a method of visual representation for the organization topology. The new model here is named as eFINC.

A. Node Model of eFINC

On the basis of FINC model, there are four types of organization functional units: C2 unit (*C2*), intelligence unit node (*I*), force unit (*F*) and communication unit (*Comm*). Communication unit is a special type of units, which builds a relation between different other units. Here, nodes are modeled as *C2*, *I* or *F*. Communication unit will be discussed, together with EDGE model in next section.

NODE ::= < C2| I | F >

(1) *C2*. C2 units receive information transferred from ***I*** or *F*, makes decisions, and takes charge of ***I*** and ***F***. The representation form is as follows:

C2 :: =< Delay, InEdges, OutEdges>

where *Delay* is the time delay for information handling, *InEdges* and *OutEdges* are respectively the input and output edges of the C2 unit.

(2) *I*. Intelligence units includes the detection and surveillance systems that provide space information about entities in the battle fields, receives and transfers these information to C2 unit or force units. Scouts, radars, early-warning aircrafts and satellites are typical examples. The representation form is as follows:

I :: =<Quality, Radius, InEdges, OutEdges>

where *Quality* represents the intelligence quality offered by intelligence units; *Radius* represents the detecting radius accordingly.

(3) *F*. Force units are any entities that can be able to receive orders from C2 units and take actions to the targets and feedback the action effects to C2 units, such as tank bands, armored vehicles, fighters. The representation form is as follows:

F :: =<Radius, InEdges, OutEdges>

where *Radius* is the combat radius of the force units.

B. Edge Model of eFINC

Edge indicates a relation between two different nodes, which is constrained with a communication unit in military organizations. The number of types of arbitrary directed relation between different units with unit types is 9, named as $E_{C2\text{-}C2}$, $E_{C2\text{-}I}$, $E_{C2\text{-}F}$, $E_{I\text{-}C2}$, $E_{I\text{-}I}$, $E_{I\text{-}F}$, $E_{F\text{-}C2}$, $E_{F\text{-}I}$, $E_{F\text{-}F}$. Each edge is dependant with some communication unit. All the

relations are sharing same basic parameter structure, defined as below:

EDGE :: =<EdgeType, Delay, Accuracy, InPort, OutPort>

where *Delay* indicates the delay time from start node *InPort* to end node *OutPort*, *Accuracy* is the information transferring accuracy, and *EdgeType :: = < $E_{C2\text{-}C2}$| $E_{C2\text{-}I}$| $E_{C2\text{-}F}$| $E_{I\text{-}C2}$| $E_{I\text{-}I}$| $E_{I\text{-}F}$| $E_{F\text{-}C2}$| $E_{F\text{-}I}$| $E_{F\text{-}F}$>*.

C. Edge Model of eFINC

The complete model of eFINC can be described as a three-tuple:

eFINC ::= <NODE,EDGE,VP>

where *NODE*, *EDGE* are defined in Section II.A and II.B respectively, which describe the components and structure of military organizations. Accordingly, *VP* defines the visual primitives for nodes and edges in eFINC, as shown in Figure 1. The square nodes represent fire units (*F)* , rounded boxes nodes represent intelligence units (*I)* , circle nodes represent *C2* units *(C2)*, lines with arrows are one-way information flows, while lines without arrows are two way information flows, and the weights on the lines are delay time when the information is transmitted through them.

In order to more clearly illustrate our model, here the paper gives an example of eFINC, shown in Figure 1, with 10 elements [1].

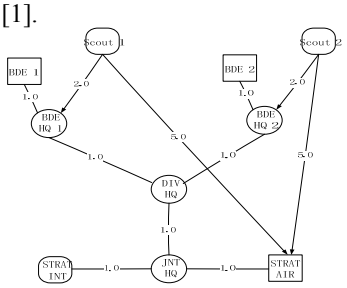

Figure 1. Military Organization Structure based on FINC

According to the organization shown in the figure, the parameters are assumed as below (the delay of each edge is depicted in Figure 1):

Radius (Scout1) = Radius (Scout2) = 100;
Radius (STRAT INT) = 400;
Radius (BDE1) = Radius (BDE2) = 100;
Radius (STRAT AIR) = 400;
Quality (Scout1) = Quality (Scout2) = 0.5;
Quality (STRAT INT) = 0.3;

III. ORGANIZATION PERFORMANCE MODELING

The role of eFINC model is to evaluate organization performance according to organization networking topology and node capacities. These performance metrics directly determinate the adaptability of military organizations while they execute missions. Performance metrics are modeled from four aspects, i.e., response speed, cooperation degree, execution capability and information support. These performance metrics are based on the process of OODA process, which is the C2 model for military organizations [2].

A. Response speed analysis

Response speed is concerned with the mean speed of information flows from intelligence units, via C2 units, to force units, indicating the speed of the whole progress from intelligence-obtaining to intelligence-employing. The main consideration is the delay time on the links between intelligence units and force units.

On the link *<I(p)-F(q)>*, the information delay time from intelligence unit *I(p)* to force unit *I(q)* is the sum of total edge delay time and total C2 delay time, namely, $delay(I(p)\text{-}F(q))=\sum delay(Edge_i) + \sum delay(C2_i)$. The *Information Flow Coefficient* (*IFC*) defined in this paper is used as the metric to measure the organization response speed. When the organization consists of *n* *<I-F>* links, the *IFC* can be represented as below:

$$IFC = \frac{1}{\sum delay(I(p), F(q)) / n} \quad (1)$$

IFC implies the control ability and response speed of military organizations over emergency situations. The larger *IFC* is, the faster the response speed is, and the stronger the control ability is.

In terms of the military organization from Figure 1, the delay time of each *<I-F>* link is:

delay (Scout-BDE1) = 4.0;
delay (Scout2-BDE1) = 8.0;
delay (STRAT INT-BDE1) = 7.0;
delay (Scout1-BDE2) = 8.0;
delay (Scout2-BDE2) = 4.0;
delay (STRAT INT-BDE2) = 7.0;
……

B. Coordination capacity analysis

Coordination capacity is very important especially when the C4ISRK system is highly networked. Coordination implies an organized group of units working together aiming at bringing about a purposeful task such as attacking a plane. Here, only coordination between two units with the same type, such as two force units, C2 units or intelligence units, are considered.

(1) Coordination analysis of force units

The cooperation capacity of force units indicates the cooperation degree while they executing a mission. The main consideration is the delay time in information transmission between force units. The less the delay time is, the faster they exchange information with each other and the higher the degree of coordination is. The shortest link between force unit *F(p)* and *F(q)* is marked as *<F(p)-F(q)>*, and the transmission delay as *delay(F(p),F(q))*. The metrics defined to measure coordination extent between force units is denoted as *Force Coordination Coefficient (FCC)*.

$$FCC = \frac{1}{\sum delay(F(p), F(q)) / n} \quad (2)$$

In terms of the military organization in Figure 1, the delay time of each link is 7, and ***FCC***=0.1433.

(2) Coordination analysis of C2 units

The Coordination analysis of C2 units mainly shows the transmission efficiency of C2 network and the

 ISBN: 978-1-61208-132-8

connectivity of C2 network. *C2 Coordination Coefficient (C2CC)* is defined to weigh the coordination degree between C2 units.

$$C2CC = \frac{1}{\sum delay(C2(p), C2(q))/n} \quad (3)$$

(3) Coordination analysis of intelligence units

Coordination analysis of intelligence units demonstrates the performance of intelligence network, and then reflects the organization capabilities of information obtaining and sharing. Similarly, *Intelligence Coordination Coefficient (ICC)* is defined to weigh the coordination degree between intelligence units.

$$ICC = \frac{1}{\sum delay(I(p), I(q))/n} \quad (4)$$

C. Execution capability analysis

Execution capability is to indicate the capabilitiues of taking orders and executing missions with the use of obtained intelligence information. In this paper, *Execution Capability Coefficient (ECC)* is defined to value the execution capability of force units. The larger ***ECC*** is, the more obvious the advantage of intelligence is, and also the better the execution capability of the force unit is.

$$ECC = \sum_{q} R(F(q))^2 \times EIQ(F(q)) \quad (5)$$

where，*(F(q))* is the combat radius of force unit *F(q)*, and *EIQ(F(q))* is the effective intelligence quality that force unit *F(q)* receives. Effective intelligence quality is the value which *Q(I(p))* is divided by *delay(I(p), F(q))*.

D. Information support analysis

As we all know, intelligence is also an important basis of C2 to carry out situation evaluation and decision making. The analysis of information support capacities is to measure timeliness, accuracy and sufficiency of intelligence. *Information Support Coefficient (ISC)* is defined in this paper to measure the information support capacities.

$$ISC = \sum_{i} IG(C2(i)) \quad (6)$$

IG(C2(i)) is the total information quantity that C2 unit ***i*** obtains. Assume that $IG=R\times R\times Q$ is the initial information quantity that the intelligence unit supplies for. Because of the changing combat environment and the transmission error, the total information quantity *IG* becomes *IG×Accuracy/delay* after transmission.

IV. THE ADAPTIVE DESIGN BASED ON PERFORMANCE RATE

In practice, the structure and performance is hardly optimally matched in the progress of mission executing. Generally speaking, mission enforcing organization could hardly run with the best performance, which calls for an exploration of an adaptive organization design method to adjust the organization structure so as to achieve better performance.

Even the performance metrics of military organizations can be calculated and evaluated; there are still many choices for adjusting the structure. It concerns with the matter of adjusting strategies. The following two strategies are to be complied with to adjust the structures of organizations to achieve better performance.

(1) *Completeness Strategy:* ensuring the completeness of basic command and control relationship, precluding the isolated units. Organizations heavily rely on the proper working state of every basic functional unit, and the task couldn't be executed successfully with the deficiency of any basic command and control relationship. As a result, the completeness of command and control relationship is essential to the performance of organization structure. Meanwhile, as we see in Section III, performance will decrease dramatically while a link is broken (It means that the delay time is infinite.)

(2) *Tightness Strategy:* strengthening the information exchange between different task modules. Different task modules maybe exist simultaneously corresponding to the multiple tasks which are executed at the same time. They have very strong internal connections but weak external connections, which cause negative effect in term of coordination of the whole mission. So it is vital to strengthen the information exchange between different task modules.

The above two strategies will lead the adaptive organization design to the right direction, where the performance is optimal. These strategies emphasize the importance of organization structures, not just the abilities of single unit as in previous models.

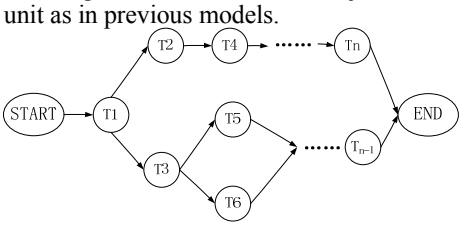

Figure 2. The Tasks Sequence of Mission ***M***

A certain mission *M* is assigned to the C4KISR system, which is composed of multiple tasks. These tasks are organized as shown in Figure 2.

Before the mission to be executed, there exists a certain initial organization O_0 whose structure will turn to O_1, O_2... with the changing of tasks. Here, we only consider the situation where adjustments are triggered by the tasks in turn. When task T_1 over O_0 is in its turn, *IFC*, *CC*, *ECC* and *ISC* can be calculated. After a certain adjustment, O_0 turning to O_0', these four metrics will changed to *IFC´*, *CC´*, *ECC ´*and *ISC´*. Now *Adjusting Value (AV)* is defined, which implies the performance rate, as the main reference for the adaptive organization design.

$$AV = \frac{IFC'-IFC}{IFC} + \frac{CC'-CC}{CC} + \frac{ECC'-ECC}{ECC} + \frac{ISC'-ISC}{ISC} \quad (7)$$

In term of a certain adjustment, the positive *AV* value is indicative of that this adjustment makes the overall performance increase, the greater the value is obtained, the more significant that, after this adjustment, the overall performance of the organization has been enhanced, then

 ISBN: 978-1-61208-132-8

the necessity of this adjustment will become increasingly obvious. Otherwise, the negative value of *AV* shows the downside of the adjustment. The adjustment will be accepted, which accords with the maximum performance value to generate a new organization structure O_1 while it can meet the needs of the task T_1. The process can be formalized as below:

$$O_i = argmax\ \{AV(\ O_{i-1}', O_{i-1})|\ O_{i-1}'\ supports\ task\ T_i\ \} \quad (8)$$

Subsequently, a similar process is used to obtain the structures of T_2, T_3 ... T_n, and their optimal organizational structures O_2, O_3 ...O_n until the mission is finished. The flow of the adaptive organization design based on performance rate is shown as the following Figure 3.

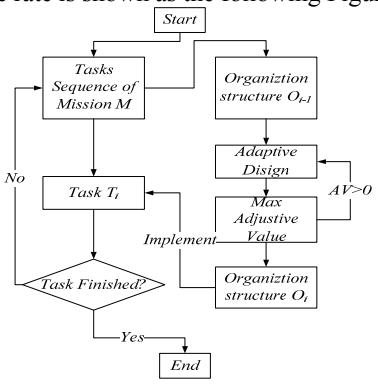

Figure 3. The adaptive organization design flowchart based on performance rate

To save the space of this paper, how to search the satisfied organization for a task is not elaborated here (please refer to [7, 8]). Here, based on the example in Figure 1, the adjustment process is illustrated as below.

Assuming that the organization will execute the task T_i over O_{i-1}, then *IFC*=0.175, *CC*=0.6433, *ECC*=47659.2, and *ISC*=84630.0. Under the precondition that certain link needs to be added and removed in the terms of task T_i, the *Adjusting Value (AV)* corresponding to the adjustment is obtained. Such as, *Adding a link BDE HQ1- BDE HQ2 makes AV*=0.2792; *Removing a link BDE HQ1-DIV HQ, AV*=-2.1719.

The facts show that better performance value can be gained due to some link adjustment. However, negative value can also be acquired. That indicates these adjustments could cause negative effects on the organization performance. In the process of adaptive organization design, the searching of the maximum *AV* could lead the organization to better performance. The optimal organization O_i of Task T_i is acquired when the *Adjusting Value* is not positive any longer.

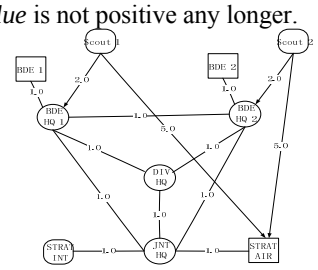

Figure 4. Optimal Structure of the Military Organization for Task T_i

After three adjusting periods, O_i is gained and shown in Figure 4, and it shows that compared with original tree structure, the networking structure has more performance advantage.

V. CASE STUDY

To illustrate our model, here a practical detailed case is given. The proposed model in Section III, IV will be demonstrated.

Figure 5 shows a C4ISRK system architecture for some military operation scenario. In order to block the attack of RED Army, the BLUE army is ordered to defend the towns Alpha and Beta. The main attack mission consisting of force defense and cannon defense provided by *Force A*, *CAN A*, *Force B*, and *CAN B* (square boxes). The secondary mission is asked to support the main attack provided by *AID A* and *AID B* (square boxes). The C2 units are coordinated by small headquarters elements, *A HQ* and *B HQ*, in towns of Alpha and Beta, and by an intermediate headquarters in *GAMMA*, but are ultimately organized form a military headquarters back *MIL HQ*. The secondary attack, on the other hand, is coordinated form the mining town of *DELTA*, and also ultimately organized from aid headquarters *AID HQ*. Form the original organization structure, we cannot find any coordination of main attack and secondary attack, namely there is no military and aid efforts whatsoever. What's more, the intelligence units mainly include *Weather*, *Satellite*, *Political*, *CIS* , etc. [1, 16].

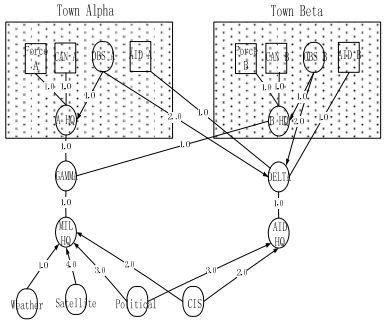

Figure 5. The original organization structure

The detailed parameters of this organization are showed in the following.

Radius (OBS A) = Radius (OBS B) = 100;
Radius (Force A) = Radius (CAN A)
= Radius (Force B) = Radius (CAN B) = 100;
Radius (AID A) = Radius (AID A) = 100;
Accuracy = 0.9; Quality = 1.

Based on the performance model, we now obtain the four main metrics values as shown in Table I (*Value for Figure 5*).

Then the adaptive organization design can be carried out based on the performance rate, in which the *AV* of all the probable organizations can be obtained. According to the model mentioned in Section IV, we will get the improved organization structure shown as Figure 6. The main coordination site now is GAMMA, and then a coordination planning could be done there by main attack

staff and secondary attack staff together. This is a good choice, since *GAMMA* is in fact the most central node in the original structure. All information (other than reports from the field) is now provided directly to the shared headquarters at *GAMMA*.

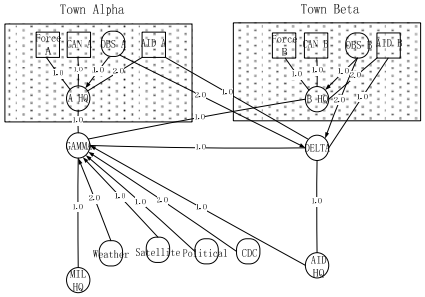

Figure 6. The Improved Organization Structure

The performance metrics improved of organization structure are shown in Table I (*Value for Figure 6*) . The obvious improvements have achieved, which is expressed by the *Advance Extent* in the table.

TABLE I. THE IMPROVED ORGANIZATION EFFICIENCY METRICS

Indices	*Metrics*	*Value for Figure 5*	*Value for Figure 6*	*Advance Extent*
Response Speed	*IFC*	0.145	0.2	37.9%
Coordination Capacity	*CC*	0	0.53	∞
Execution Capability	*ECC*	33285.7	41481.5	24.6%
Information Support	*ISO*	845442	1313100	55.3%

VI. CONCLUSION

We have presented a methodology for representing organization structures, which is called as eFINC. By this methodology, we perform a new way to evaluate the performance of the organization on the basis of functional unit's abilities and network topology. In terms of response speed, coordination capacity, performing capability and information support effectiveness, this paper proposes a quantitative explanation, and proposes four types of metrics. Based on the metrics, the process model of adaptive organization design is put forward. The case study can show that the model is effective.

To the best of our knowledge, there are still not any models which take account into performance while designing the organization. Our method is performance-oriented, not just other methods that are task-oriented. Performance-oriented design will be more suitable to the networking and dynamic combat fields where military organizations are performing the tasks.

VII. ACKNOWLEDGEMENT

The paper is supported by NSFC projects (70771109, 60504036) . And thank for their grateful support.

[1] Dekker A.H. (2002). Applying Social Network Analysis Concepts to Military C4ISR Architectures, Connections, the official journal of the International Network for Social Network Analysis, 24(3) , pp. 93-103

[2] Sean D., Shannon R. B., Ghaith A. R., and Andreas T. (2009). Applying the Information Age Combat Model: Quantitative Analysis of Network Centric Operations. The International C2 Journal, http:// www.dodccrp.org/ files /IC2J_v3n1_06_Deller. Pdf on March 2, 2011

[3] Lyons J. B., Swindler S. D., White J. A. (2008). Dynamic Complexity in System of Systems[R]. International Symposium on Collaborative Technologies and Systems, pp.367-374

[4] Kathleen M. C. (2003). Dynamic Network Analysis, Dynamic Social Network Modeling and Analysis: workshop summary and papers, National Academies Press

[5] Anthony H. D. (2002). C4ISR Architectures, Social Network Analysis and the FINC Methodology: An Experiment in Military Organization Structure, Australia Department of Defense, Defense Science and Technology Organization (DSTO) report: DSTO-GD-0313, Jan 2002

[6] Dekker A.H. (2005). Network Topology and Military Performance. International Congress on Modeling and Simulation, 2005, pp. 2174-2180

[7] Levchuk G. M., Levchuk Y, N., Luo J., Pattipati K. R., and Kleinman D. L. (2002). Normative design of organizations- Part I: Mission planning. IEEE Transactions on Systems, Man, and Cybernetics, Part A, 32(3), pp. 346-359

[8] Levchuk G. M., Levchuk Y, N., Luo J., Pattipati K. R., and Kleinman D. L. (2002). Normative design of organizations-Part II: Organizational Structure. IEEE Transactions on Systems, Man, and Cybernetics, Part A, 32(3), pp. 360-375

[9] Levchuk G. M., Levchuk Y, N., Meirina C., Pattipati K. R., and Kleinman D. L. (2004). Normative design of project-based organizations-Part III: modeling congruent, robust, and adaptive organizations. IEEE Transactions on Systems, Man, and Cybernetics, Part A, 34(3) , pp. 337-350

[10] Jincai H., Shidong Q., and Zhong L. (2010). The Modeling and Evolvement Analysis Method of Operation System of Systems based on Extension Space, IEEE International Conference on ICIS 2010, pp. 113-117

[11] Jincai H., Weiming Z., Guoli Y., and Shidong Q. (2010). The modeling and efficiency analysis method of C2 System Of Systems based on FINC model, IEEE International Conference on MLC 2010, pp. 2026-2030

[12] Kathleen M. C., Jana D., Jeffrey R., and Maksim T. (2007). Toward an interoperable dynamic network analysis toolkit. Decision Support Systems, 43(4), pp. 1324-1347

[13] Jeff C. (2005). Distributed Networked Operations: The Foundations of Network Centric Warfare. Alidate Press

[14] Kathleen M. C. (2006). Modeling Community Containment for Pandemic Influenza: A Letter Report [R], Washington, DC: National Academy Press.

[15] Baoxin X. and ML L. (2010). Robustness analysis of military organization, IEEE International Conference on MLC2010, pp. 1971-1975

[16] Anthony H. D. and Bernard D. C. (2004). Network Robustness and Graph Topology. ACSC '04 Proceedings of the 27th Australasian conference on Computer science, pp. 359-368

[17] Anthony H. D. (2005). Simulating Network Robustness for Critical Infrastructure Networks. ACSC '05 Proceedings of the 28th Australasian conference on Computer Science, pp. 59-67

 ISBN: 978-1-61208-132-8

Study of Polynomial Methods of Finite Differences for the Wavelength Division Multiplex Mesh Networks With Dedicated Optical Path Protection

Stefanos Mylonakis
1-3 Macedonia str, 10433
ATHENS-GREECE
smylo@otenet.gr

***Abstract*-The dedicated protection method uses the spare capacity as a "dedicated" resource and the backup capacity is allocated for the sole use of the connection or link. This method will be used by polynomial methods of the finite differences. The latter provides two methods, the first is the accurate polynomial method and the second is the approximate polynomial one. These methods are arithmetical, solving successfully the survivability problems and used for the verification of their results. The advantage of the polynomial methods is no use of the large matrix which shows the network active links that is used in the linear method. In this paper, the regression analysis is used to study the absolute error of polynomial finite differences methods (statistical method).**

Keywords: finite differences; accurate polynomial; approximate polynomial; dedicated protection; regression.

I. INTRODUCTION

The WDM (Wavelength Division Multiplex) optical mesh networks are high capacity telecommunication networks based on optical technologies that provide routing, grooming and restoration at the wavelength level as well as wavelength based services.

A fiber cut can have massive implications because network planners use more network elements to increase fiber capacity. For a WDM system with many channels on a single fiber, a fiber cut would result in multiple failures, causing many independent systems to fail. The optical path with dedicated 1+1 protection on optical layer of optical WDM mesh networks can perform protection switching faster and more economically. In the present paper the objective is study the absolute error of the polynomial function methods by statistical method. So, we solve the optical path with dedicated 1+1 protection problem using the polynomial methods of the finite differences calculating the final available capacity by two polynomial finite difference methods, the accurate one and the approximate one. Then the absolute error of two methods is calculating. The procedure is executed for a large number of experiments and the regression analysis is used to study the error. Research has been done in relation to the methods and the problems associated with planning, protection and restoration of optical networks. In [1] the authors present OTN (Optical Transport Network) evolution from an operator's point of view, including the history of the transport network, the role of the OTN, and the motivations and requirements for OTN evolution. For a WDM system with many channels on a single fiber, a fiber cut would results multiple failures, causing many independent systems to fail [2][3][5][6][8][9][10]. There are also several approaches to ensure fiber network survivability [2][3][5][6][9]. Network survivability is defined as the capability of a communication network to resist any link or node interruption or disturbance of service, particularly by warfare, fire, earthquake, harmful radiation or other physical or natural catastrophes.The existing methods in solving these problems use special algorithms. We suggest a proposal that it is an approach based on the finite differences polynomial methods and represents the detail algorithm description and its program. The advantage of this approach is that the polynomial methods of finite differences solve the same telecommunication problem (of this kind) by two different methods simultaneously and they can verify each other in accepted tolerances. These methods also can compare with the linear one [11] and it is another verification way.

The following analysis presents the solution of the problems associated with the survival optical networks on the basis of the finite differences and the following problem is solved. The role of the Difference Calculus is in the study of the Numerical Methods. Computer solves these Numerical Methods. The subject of the Difference Equations [6] is in the treatment of discontinuous processes. The network final available capacity is revealed as a difference equation because the final available capacity of the individual working optical fibers is also a difference equation. The reduction of the available capacity of each working optical fiber is a discontinuous process when connection groups of several sizes pass through it. In [10], the authors begin with an overview of the existing strategies for providing transport network survivability and continue with an analysis of how the architectures for network survivability may evolve to satisfy the requirements of emerging networks. In [11], the author presents the finite differences, their methods and their problems when they are used to solve problems of this kind. In [12], two link disjoint paths, a dedicated working path and a shared protection path are computed , for an incoming light path request based on the current network state but the protection approaches to optimize the resource utilization for a given traffic matrix, do not apply because lightpath requests come and go dynamically.

This paper is broken down in the following sections: Section II shows how the finite differences are used for each optical fiber and illustrates the optical fiber final available capacity; Section III describes the problem, its formulation, its algorithm, an example and the proposals with discussion; Section IV draws conclusion and finally ends with the references.

 ISBN: 978-1-61208-132-8

II. THE OPTICAL FIBER AND THE FINITE DIFFERENCES

Before studying finite differences and their use in optical WDM mesh networks survivability, it is necessary to provide a short comprehensive presentation of the finite differences computation. Let's assume that $y_1, y_2, \ldots, y_n$ is a sequence of numbers in which the order is determined by the index n. The number n is an integer and the y_n can be regarded as a function of n, an independent variable with function domain the natural numbers and it is discontinuous. Such a sequence shows the available capacity reduction of a telecommunication fibre network link between two nodes when the telecommunication traffic of 1,2, …, n source-destination node pairs pass through. It is assumed that the telecommunication traffic unit is the optical channel that is one wavelength (1λ). The telecommunications traffic includes optical connections with their protections. The total connections of a node pair form its connection group. The first order finite differences represent symbolically the connection group of each node pair that passes through a fiber. This connection group occupies the corresponding number of optical channels and it is the bandwidth that is consumed by connections of a node pair through this fiber. The first order finite differences are used to represent the connection groups in optical channels of the node pairs that pass through an optical fiber .An equation of the first order finite differences gives the available capacity of an optical fiber network link when a connection group passes through it. This available capacity is provided for the connection groups of the other node pairs that their connections will pass through this optical fiber. When the first connection group of Δy_1 connections passes through an optical fiber network link with installed capacity of y_1 optical channels the first order finite difference equation gives the available capacity $y_{2\,(\,y_{1\,+1)}}$ which is written as following

$$y_{1+1} = y_1 - \Delta y_1$$

The sequence $\Delta y_1, \Delta y_2, \Delta y_3, \ldots, \Delta y_n$ represents the connection groups that pass through this optical fiber network link .When Δy_1 subtracted from y_1, creates y_2, when Δy_2 subtracted from y_2, creates $y_{3,\ldots,}$ when Δy_n subtracted from y_n creates y_{n+1} which is the total unused available capacity of this optical fiber. Thus the total unused available capacity of each network optical fiber is calculated after n connections groups pass through it. The total unused available capacity of each network optical fiber is also written as a polynomial function, and there are two polynomial function methods. The assessment of the polynomial function coefficients is done with the values that the polynomial function represents for 1, 2,…, n, n +1. The values of the function y_{n+1} for each n must be integral because each value represents optical channels. There are more details in the [11] but for helping the reader we write them again.

The general form of a polynomial function that gives the available capacity of the optical fiber network link after the serving n connection groups, is as follows

$$y_{n+1} = \sum_{r=0}^{n} \alpha_r * (n+1)^r \qquad (1)$$

The assessment of the polynomial function coefficients is done with the values that the polynomial function represents for 1,2,…, *n*, *n* +1 .The values of the function y_{n+1} for each *n* must be integral because each value represents optical channels.

-If the equation (1) is written analytically as follows

$y_{0+1} = \alpha_0 * (0+1)^0$

$y_{1+1} = \alpha_0 * (1+1)^0 + \alpha_1 * (1+1)^1$

$y_{2+1} = \alpha_0 * (2+1)^0 + \alpha_1 * (2+1)^1 + \alpha_2 * (2+1)^2$

$y_{3+1} = \alpha_0 * (3+1)^0 + \alpha_1 * (3+1)^1 + \alpha_2 * (3+1)^2 + \alpha_3 * (3+1)^3$

……

$y_{n+1} = \alpha_0 * (n+1)^0 + \alpha_1 * (n+1)^1 + \alpha_2 * (n+1)^2 + \ldots + \alpha_n * (n+1)^n$

The value of the function has high accuracy of 15 decimal digits. This method is an *accurate* one.

-If the equation (1) is written analytically as follows

$y_{0+1} = \alpha_0 * (0+1)^0 + \alpha_1 * (0+1)^1 + \alpha_2 * (0+1)^2 + \ldots + \alpha_n * (0+1)^n$

$y_{1+1} = \alpha_0 * (1+1)^0 + \alpha_1 * (1+1)^1 + \alpha_2 * (1+1)^2 + \ldots + \alpha_n * (1+1)^n$

$y_{2+1} = \alpha_0 * (2+1)^0 + \alpha_1 * (2+1)^1 + \alpha_2 * (2+1)^2 + \ldots + \alpha_n * (2+1)^n$

$y_{3+1} = \alpha_0 * (3+1)^0 + \alpha_1 * (3+1)^1 + \alpha_3 * (3+1)^2 + \ldots + \alpha_n * (3+1)^n$

……

$Y_{n+1} = \alpha_0 * (n+1)^0 + \alpha_1 * (n+1)^1 + \alpha_2 * (n+1)^2 + \ldots + \alpha_n * (n+1)^n$

The value of the function has reduced accuracy and depends of the value of *n*. This method is *an approximated* one.

They are systems of n+1equations with n+1unknown coefficients. The values of the coefficients depend of the number of the connection groups and the connections of each connection group. The 1st method is more accurate than 2nd one because only one factor is added to new equation when the polynomial degree increases versus one factor is added for all equations respectively.

TABLE 1.THE SYMBOLS OF THIS PAPER

S/N	Symbol	Comments
1	q	The node number
2	p	The edge number
3	G(V,E)	The network graph
4	V(G)	The network node set
5	E(G)	The network edge set
6	2p	The number of working and backup fiber for 1+1 line protection
7	n	The number of source – destination nodes pairs of the network
8	n(i)	The total number of the connection groups that passes through the fiber (*i*) and means that each fiber has different number of connection groups pass through it
9	$n(i)_w$	The number of the working connection groups that passes through the fiber (*i*) and means that each fiber has different number of connection groups pass through it
10	$n(i)_p$	The number of the protection connection groups that passes through the fiber (*i*) and means that each fiber has different number of connection groups pass through it
11	K	The number of the wavelengths channels on each fiber that is the WDM system capacity
12	Cinst	The total installed capacity
13	Cav	The total available capacity
14	Cw	The total working capacity
15	Cpr	The total protection capacity
16	C_b	The total busy capacity
17	$\Delta y_{w,i,j}$	The first order of finite difference that corresponds to a group of working optical connections that pass through the optical fiber ***i*** with serial number ***j*** respectively.
18	$\Delta y_{pr,i,j}$	The first order of finite difference that corresponds to a group of protection optical connections that pass through the optical fiber ***i*** with serial number ***j*** respectively.

III. THE PROBLEM AND ITS SOLUTION

A. The problem

The network topology and other parameters are known as WDM and optical fiber capacity, one optical fiber per link with an extension to a 1+1 fiber protection system. So this network is characterized by one working fiber per link, edges of two links, links of two optical fibers, one for working and one for protection. The connections are lightpaths originating in the source nodes and terminating at the destination nodes proceeding from preplanned optical working paths. Additionally, the same number of optical paths is preselected for the preplanned fully disjoint backup paths, (1+1 dedicated protection connection). So the connections are protected. The connections of the same node pair form a group along the network. The preplanned protection paths do the dedicated protection of the connection groups. So a suitable number of wavelengths per link along the network uses. The solution is the calculation of the final available capacity of the network for a given table. This table contains the number of the node pairs, the node pairs and the number of the connections of each node pair when their working and protection paths are preplanned. This procedure is executed for polynomial accurate finite difference method and for the corresponded approximate on and the absolute error is calculated. In [11] I have proved that the linear finite difference method (which is an ILP method with the disadvantage the large matrix A (2pxn) which occupies large memory and has difficult treatment)) and the polynomial accurate method are not introducing errors in the solution of the problem. The polynomial approximate one does it. The absolute error is calculated statistically when the procedure is executed for a large number of experiments and the number of the connections of each node pair takes its value by a random number generation with adjusted maximum value. The maximum value must be small in comparison with the WDM and fiber system capacity so that network is a strictly non blocking network, in which it is always possible to connect any node pair, regardless of the state of the network. So all requests for connection are satisfied and form connections. The regression analysis is used to study this absolute error. The regression analysis focuses more on how much each curve of table 2 is better fitting to the error data that introduced by the approximate method. These fitted curves can aid for data visualization, to infer values of a function where no data are available, and to summarize the relationships among the variables. The curves of the regression analysis are in the table 2.

B. The formulation

The network is assumed to be an optical mesh network with the circuit switched (or packet switched but the packets are adjusted to follow preplanned paths) as a graph. Each vertex represents the central telecommunications office (CO) with the OXC while each edge represents two links. Each edge link has a couple of optical fibers. All optical fibers have the same capacity as the WDM system. All nodes are identical. The numbers of working and protection connections that pass through each optical fiber are different. Two finite difference polynomial equations are calculated for each optical fiber, one for polynomial accurate method and one for the polynomial approximate one. For all network links, the general equation of the polynomial function has two column matrices, the left one that is equals with the right one. When all connections have been set up then each element of the column matrix must be greater or equal to zero. In other cases some connections are not possible. The total final available capacity of the network for the polynomial methods is given by the equation (2). This is the formulation of the *polynomial function* method problem.

$$\sum_{i=1}^{2p} y_{i,n(i)+1} = \sum_{i=1}^{2p} \sum_{r=0}^{n(i)} \alpha_{i,r} * (n(i) + 1)^r \qquad (2)$$

$$y_{i,n(i)+1} >= 0 \quad ,n(i)>0$$

The curves of the regression analysis are in the table 2.

TABLE 2.THE CURVES OF THE REGRESSION ANALYSIS

CURVE NAME	DISTRIBUTION
NORMAL	$Y=A*\exp(((X-B)^2)/C)$
LOG NORMAL	$Y=A*\exp(((\ln(X)-B)^2)/C)$
PARABOLA	$Y=A+B*X+C*X^2$

C. The algorithm

Our algorithm describes the operation of the WDM optical fiber mesh network and we investigate the polynomial finite difference methods. TURBO PASCAL is used to program the model [4]. The algorithm has the following steps and phases.

First step *Network parameters*

Initially the following data are known: network topology, node number, edge number, link number per edge, working optical fiber number per link, protection optical fiber number per link, wavelength number per optical fiber, optical fiber numbering. This information allows the computer to draw a graph and an OXC is on the vertex of the graph [10]. Each edge corresponds to two links with opposite direction to each other. All fibers have the same wavelength number and all links the same fiber number. The computer reads the adjacency matrix and is informed about the network topology.

Second step *Connection selections*

In this step, the number of the connection node pair, the connection node pair selection for connections and the desired connection group size are done. The preplanned working and the protection optical paths for connections of every node pair are also provided. When many experiments are executed a random number generator with adjusted maximum value (max) is activated to give the connection group size.

Failure-free Network Phase

Third step *Wavelength allocation*

In this step, wavelength allocation is initiated. A working connection starts from the source node and progresses through the network occupying a wavelength on each optical fiber and switch to another fiber on the same or other wavelength by OXC, according to its preplanned working optical path up to arrive at the destination node. Simultaneously, the protection connection starts from the source node and progresses through the network occupying a wavelength on each optical fiber and switch to another fiber on the same or other wavelength by OXC, according to its preplanned protection optical path up to arrive at the destination node. So there is full and dedicated protection for this connection. The number of connections of each node pair is equal to its connection group size. After a connection (working as well as protection) has been established, the available capacity is also calculated under both methods of the finite differences. Thus the available capacity of the one method is compared to available capacity of the other method for one connection.

Forth step. *Presentation of the finite differences*

The total available capacity of each optical fiber is calculated and represented under both polynomial methods of finite difference. Thus the total fiber capacity available under the first method is compared to the total fiber capacity available of the other method for all connections.

Fifth step. *Results and comparisons thereof*

Having the desired connection group size the total results are computed under each finite difference method that are the total sum of the individual connection group size, the total installed capacity, the total protection capacity, the total busy capacity and the available residual capacity. These results of the accurate method are compared to the results of the approximated method. The results of these methods must be equal or to have the desired tolerances. In other cases the worst results are rejected.

The *max* value must be small in comparison with the WDM and fiber system capacity so that network is a strictly non blocking network, in which it is always possible to connect any node pair, regardless of the state of the network. So all requests for connection are satisfied and form connections.

Network with failure Phase

When a failure occurs and a link is cut, the optical fibers of this link are also cut and the optical fiber protection 1+1 and the network topology change. The connection groups that passed through the cut link are also cut and the restoration is carried out passing through the preplanned protection paths of other links. The computer is informed of the cut link and modifies suitably the network parameters. The cut optical fiber sets its wavelengths to zero. The connection groups that passing through the cut link set their using wavelengths to zero and through the others to free.

D. Example

The network and the results are presented shortly because the paper must be short. The topology of the network is presented by the graph G (V, E). The vertex set has q=6 elements and the edge set has p=9 elements. Each edge is an optical link of two directions with one working fiber for each direction. Thus there are 2*p=2*9=18 optical fibers. Connection groups transverse the mesh network and correspond to *n* source-destination node pairs. WDM system capacity has 30 wavelengths. The accurate polynomial function method and the approximated polynomial function are presented.

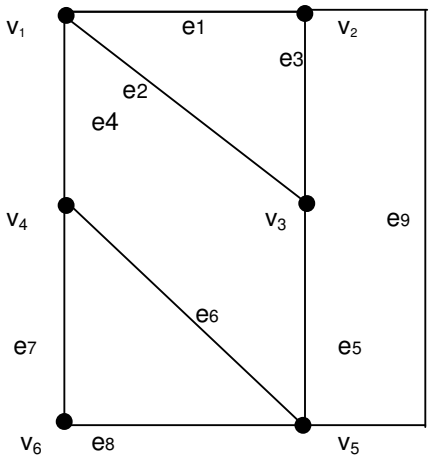

Figure 1.The mesh topology of the network.

TABLE 3.ORDER, SIZE, WORKING PATH, PROTECTION PATH OF EACH NODE PAIR

Node Pair [Si,Di]	Node pair [v_i, v_j]	Working Path	Protection Path	Group size
[S_1, D_1]	[v_1, v_2]	v_1, v_2	v_1, v_3 ,v_2	1
[S_2, D_2]	[v_1, v_3]	v_1 ,v_3	v_1, v_2 ,v_3	2
[S_3, D_3]	[v_1, v_5]	v_1, v_3 ,v_5	v_1, v_2,v_5	5
[S_4, D_4]	[v_2, v_3]	v_2, v_3	v_2, v_1 ,v_3	2
[S_5, D_5]	[v_2, v_4]	v_2, v_1 ,v_4	v_2, v_5 ,v_4	2
[S_6, D_6]	[v_2, v_5]	v_2 , v_5	v_2, v_1, v_4 ,v_5	3
[S_7, D_7]	[v_3, v_4]	v_3, v_1, v_4	v_3, v_5 ,v_4	1
[S_8, D_8]	[v_3, v_6]	v_3,v_5 ,v_6	v_3,v_1 ,v_4 ,v_6	4
[S_9, D_9]	[v_4, v_1]	v_4 ,v_1	v_4,v_5,v_3 ,v_1	1
[S_{10}, D_{10}]	[v_4, v_5]	v_4, v_5	v_4, v_6,v_5	2
[S_{11}, D_{11}]	[v_5, v_4]	v_5, v_4	v_5, v_6 ,v_4	5
[S_{12}, D_{12}]	[v_6, v_1]	v_6, ,v_4,v_1	v_6, v_5 ,v_3,v_1	2

The problem is solved for *n=12 of 30 possible* connection groups. These have their order and sizes for each source-destination node pair, their working paths and their protection paths as shown in *table 3*. The results with finite differences are showed. It is obvious that the dedicated path protection mechanisms use more than 100% redundant capacity because their lengths are longer than their working paths. The total length of working paths is seventeen, (17) and the total length of protection paths is twenty-five, (25).Similarly for the same connections requested group size the capacity that is used by the protection paths is larger than the corresponding working paths.

The synoptic presentation is used for the finite difference tables. So the higher order finite differences and the number of connection groups that pass through each optical fiber are showed in the table 4. (Fiber, i) shows the optical fibers. The n(i) shows the number of the connection groups that pass through each optical fiber. The $\Delta^{m(i)}y_i$ the order finite differences with m(i)=1, 2, 3, 4, 5 of the fiber i. The intermediate order finite differences are not showed. The 16teen has only 1st order differences.

TABLE 4.THE HIGHER ORDER FINITE DIFFERENCES AND THE NUMBER OF CONNECTION GROUPS THAT PASS THROUGH EACH OPTICAL FIBER

Fiber,i	n(i)	m(i)	$\Delta^{m(i)}y_i$
1	3	3	2
2	3	3	1
3	4	4	8
4	4	4	-10
5	2	2	3
6	1	1	1
7	4	4	-8
8	2	2	-1
9	3	3	7
10	2	2	-1
11	3	3	3
12	3	3	5
13	2	2	2
14	2	2	3
15	2	2	-4
16	2	2	0
17	3	3	6
18	0	0	0

 ISBN: 978-1-61208-132-8

When none group goes through the optical fiber, then the degree of the polynomial function is 0, when one group goes through, then the degree of the polynomial function is 1, when two groups go through, then the degree of the polynomial function is 2, etc.

The polynomials that calculate the available capacity of each optical link for the accurate (first) method for all possible values up to four and for the following cases are represented. The number, the order and the size of $\Delta y_{i,j}$ are critical.

-If no one-connection group passes through an optical link the polynomial function is constant, etc.

$y_{j,0+1}= \alpha_0*(0+1)^0$

$y_{j,0+1}=30.$

-If only one-connection group passes through an optical link the polynomial function is of the first degree.

$\Delta y_{i,1}$,	$y_{i,1+1}= \alpha_0*(1+1)^0 +\alpha_1*(1+1)^1$
1	$y_{i,1+1}= 30*(1+1)^0 -(1/2)*(1+1)^1 =29$
2	$y_{i,1+1}= 30*(1+1)^0 -(2/2)*(1+1)^1=28$
3	$y_{i,1+1}= 30*(1+1)^0 -(3/2)*(1+1)^1=27$
4	$y_{i,1+1}= 30*(1+1)^0 -(4/2)*(1+1)^1=26$

-If only two-connection groups pass through an optical link the polynomial function is of the second degree.

$\Delta y_{i,1}$, $\Delta y_{i,2}$, $y_{i,1+1}= \alpha_0*(2+1)^0 +\alpha_1*(2+1)^1+\alpha_2*(2+1)^2$

1 , 1, $y_{i,2+1}= 30*(2+1)^0 -0.5*(2+1)^1- 5.55555555555429E-2 *(2+1)^2 =28$
2 , 1, $y_{i,2+1}= 30*(2+1)^0 -1.0*(2+1)^1-0.000000000000000E+0*(2+1)^2 =27$
1 , 2, $y_{i,2+1}= 30*(2+1)^0 -0.5*(2+1)^1-1.666666666667420E-1*(2+1)^2 =27$
3 , 1, $y_{i,2+1}= 30*(2+1)^0 -1.5*(2+1)^1+5.55555555555429E-2 *(2+1)^2=26$
2 , 2, $y_{i,2+1}= 30*(2+1)^0 -1.0*(2+1)^1-1.11111111111086E-1*(2+1)^2 =26$
1 , 3, $y_{i,2+1}= 30*(2+1)^0 -0.5*(2+1)^1-2.77777777777828E-1*(2+1)^2 =26$
4 , 1, $y_{i,2+1}= 30*(2+1)^0 -2.0*(2+1)^1+1.11111111111086E-1*(2+1)^2=25$
3 , 2, $y_{i,2+1}= 30*(2+1)^0 -1.5*(2+1)^1-5.55555555555429E-2 *(2+1)^2 =25$
2 , 3, $y_{i,2+1}= 30*(2+1)^0 -1.0*(2+1)^1-2.222222222221720E-1*(2+1)^2=25$
1 , 4, $y_{i,2+1}= 30*(2+1)^0 -0.5*(2+1)^1-3.888888888886870E-1*(2+1)^2=25$
4 , 2, $y_{i,2+1}= 30*(2+1)^0 -2.0*(2+1)^1+0.000000000000000E+0*(2+1)^2=24$
3 , 3, $y_{i,2+1}= 30*(2+1)^0 -1.5*(2+1)^1-1.666666666667420E-1*(2+1)^2 =24$
2 , 4, $y_{i,2+1}= 30*(2+1)^0 -1.0*(2+1)^1-3.333333333333485E-1*(2+1)^2=24$

e.t.c

The following matrix provides the residual capacity of all optical fibers, its dimension is (18x1) and the first method or the accurate one is written for all network links as follows.

$$\begin{vmatrix} y_{1,3+1} \\ y_{2,3+1} \\ y_{3,4+1} \\ y_{4,4+1} \\ y_{5,2+1} \\ y_{6,1+1} \\ y_{7,4+1} \\ y_{8,2+1} \\ y_{9,3+1} \\ y_{10,2+} \\ y_{11,3+} \\ y_{12,3+} \\ y_{13,2+} \\ y_{14,2+} \\ y_{15,2+} \\ y_{16,2+} \\ y_{17,3+} \\ y_{18,0+} \end{vmatrix} = \begin{vmatrix} 30 – 0.5*4 -0.167*16-0.05208*64 \\ 30 -1.00 *4 - 0.111*16-0.0191*64 \\ 30 - 0.5 *5 - 0.167*25-0.05208*125 +0.00508*625 \\ 30 – 0.5*5-0.3888*25+0.03472*125-0.0001888*625 \\ 30 - 1 *3 - 0.11111*9 \\ 30 - 0.5*2 \\ 30 - 1*5 - 0.22222*25+0.0243*125-0.003972*625 \\ 30 - 0.5*3 -0.167*9 \\ 30 - 2.5 *4 + 0.1667*16-0.041667*64 \\ 30 - 0.5 *3 - 0.167*9 \\ 30 - 1.5 *4 + 0.055*16-0.013889*64 \\ 30 - 1*4+0*16-0.0625*64 \\ 30 – 2*3 -0*9 \\ 30 - 2.5*3 + 0.0555*9 \\ 30 - 2*3 - 0.333*9 \\ 30 - 1*3 - 0.1111*9 \\ 30 – 2.5*4 +0.05555*16-0.013889*64 \\ 30 \end{vmatrix} = \begin{vmatrix} 22 \\ 23 \\ 20 \\ 22 \\ 26 \\ 29 \\ 20 \\ 27 \\ 20 \\ 27 \\ 24 \\ 22 \\ 24 \\ 23 \\ 21 \\ 26 \\ 20 \\ 30 \end{vmatrix}$$

This method is an accurate one but if all coefficients are not written completely with 15 digits there are errors in the results. The above ones are rounded to the closest integer to agree with the real ones.

The polynomials that calculate the available capacity of each optical link for the approximated (second) method for all possible values up to four and for the following cases are represented. The number, the order and the size of $\Delta y_{i,j}$ are critical.

-If no one-connection group passes through an optical link the polynomial function is constant.

$y_{j,0+1}= \alpha_0*(0+1)^0$

$y_{j,0+1}=30.$

-If only one-connection group passes through an optical link the polynomial function is of the first degree.

$\Delta y_{i,1}$,	$y_{i,1+1}= \alpha_0*(1+1)^0 +\alpha_1*(1+1)^1$
1	$y_{i,1+1}= (30+1)*(1+1)^0 -1*(1+1)^1 =29$
2	$y_{i,1+1}= (30+2)*(1+1)^0 -2*(1+1)^1=28$
3	$y_{i,1+1}= (30+3)*(1+1)^0 -3*(1+1)^1=27$
4	$y_{i,1+1}= (30+4)*(1+1)^0 -4*(1+1)^1=26$

-If only two-connection groups passes through an optical link the polynomial function is of the second degree.

$\Delta y_{i,1}$, $\Delta y_{i,2}$, $y_{i,1+1}= \alpha_0*(2+1)^0 +\alpha_1*(2+1)^1+\alpha_2*(2+1)^2$

1 , 1, $y_{i,2+1}= 31*(2+1)^0 -1.0*(2+1)^1+0.000*(2+1)^2 =28$
2 , 1, $y_{i,2+1}= 33*(2+1)^0 -3.5*(2+1)^1+0.500*(2+1)^2=27$
1 , 2, $y_{i,2+1}= 30*(2+1)^0 +0.5*(2+1)^1-0.500*(2+1)^2 =27$
3 , 1, $y_{i,2+1}= 35*(2+1)^0 -6.0*(2+1)^1+1.000*(2+1)^2 =26$
2 , 2, $y_{i,2+1}= 32*(2+1)^0 -2.0*(2+1)^1-0.000*(2+1)^2=26$
1 , 3, $y_{i,2+1}= 29*(2+1)^0 +2.0*(2+1)^1-1.000*(2+1)^2=26$
4 , 1, $y_{i,2+1}= 37*(2+1)^0 -8.5*(2+1)^1+1.500*(2+1)^2=25$
3 , 2, $y_{i,2+1}= 34*(2+1)^0 -4.5*(2+1)^1+0.500*(2+1)^2 =25$
2 , 3, $y_{i,2+1}= 31*(2+1)^0 -0.5*(2+1)^1-0.500*(2+1)^2=25$
1 , 4, $y_{i,2+1}= 28*(2+1)^0 +3.5*(2+1)^1-1.500*(2+1)^2=25$
4 , 2, $y_{i,2+1}= 36*(2+1)^0 -7.0*(2+1)^1+1.000*(2+1)^2=24$
3 , 3, $y_{i,2+1}= 33*(2+1)^0 -3.0*(2+1)^1+0.000*(2+1)^2 =24$
2 , 4, $y_{i,2+1}= 30*(2+1)^0 +1.0*(2+1)^1-1.000*(2+1)^2=24$

e.t.c

The following matrix provides the residual capacity of all optical fibers, its dimension is (18x1) and the second method or the approximated one is written for all network links as follows.

$$\begin{vmatrix} y_{1,3+1} \\ y_{2,3+1} \\ y_{3,4+1} \\ y_{4,4+1} \\ y_{5,2+1} \\ y_{6,1+1} \\ y_{7,4+1} \\ y_{8,2+1} \\ y_{9,3+1} \\ y_{10,2+1} \\ y_{11,3+1} \\ y_{12,3+1} \\ y_{13,2+1} \\ y_{14,2+1} \\ y_{15,2+1} \\ y_{16,2+1} \\ y_{17,3+1} \\ y_{18,0+1} \end{vmatrix} = \begin{vmatrix} 32 – 3.1574*4 +1.5*16-0.3336*64 \\ 33 - 3.8241*4 + 1*16-0.1669*64 \\ 40 - 19.8339 *5 + 13.1667*25-3.6661*125 +0.3333*625 \\ 12 + 35.33*5-22.083*25+5.1671*125-0.4167*625 \\ 32 - 2*3 +0*9 \\ 31 - 1*2 \\ 20 + 21.662 *5 -15.1666*25+3.8338*125-0.3334*625 \\ 30 + 0.5*3 -0.5*9 \\ 46 – 23.824*4 +9*16-1.167*64 \\ 30 + 0.5*3 - 0.5* 9 \\ 38 - 11.494*4 + 4*16-0.5*64 \\ 38 - 12.6574*4+5.5*16-0.8336*64 \\ 36 – 7*3 +1*9 \\ 38 - 9.5*3+1.5*9 \\ 33 - 2.5*3 - 0.5*9 \\ 32 - 2*3 + 0*9 \\ 42 – 16.824*4+5.5*16-0.667*64 \\ 30 \end{vmatrix} = \begin{vmatrix} 22.02 \\ 23.022 \\ 20.048 \\ 22.032 \\ 26 \\ 29 \\ 20.016 \\ 27 \\ 20.016 \\ 27 \\ 24.024 \\ 22.02 \\ 24 \\ 23 \\ 21 \\ 26 \\ 20.016 \\ 30 \end{vmatrix}$$

The absolute error for all fibers and for both methods is the following.

$$\begin{vmatrix} \text{Error abs } y_{1,4} \\ \text{Error abs } y_{2,4} \\ \text{Error abs } y_{3,5} \\ \text{Error abs } y_{4,5} \\ \text{Error abs } y_{5,3} \\ \text{Error abs } y_{6,2} \\ \text{Error abs } y_{7,5} \\ \text{Error abs } y_{8,3} \\ \text{Error abs } y_{9,4} \\ \text{Error abs } y_{10,3} \\ \text{Error abs } y_{11,4} \\ \text{Error abs } y_{12,4} \\ \text{Error abs } y_{13,3} \\ \text{Error abs } y_{14,3} \\ \text{Error abs } y_{15,3} \\ \text{Error abs } y_{16,3} \\ \text{Error abs } y_{17,4} \\ \text{Error abs } y_{18,1} \end{vmatrix} = \begin{vmatrix} 0.02 \\ 0.022 \\ 0.048 \\ 0.032 \\ 0 \\ 0 \\ 0.016 \\ 0 \\ 0.016 \\ 0 \\ 0.024 \\ 0.02 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0.016 \\ 0 \end{vmatrix}$$

If the degree of polynomials increases then the above writing of the polynomial numerical coefficients has error because they are difficult to be represented.

The total available capacity for the accurate (first) method is $Cavp_1$=426 wavelengths and for the approximated (second) method is $Cavp_2$=426.214 wavelengths. Thus, the total busy capacity for the accurate (first) method is given C_{b1} =114 wavelengths and for the approximated (second) method is given C_{b2} =113.786 wavelengths. The network installed capacity is Cinst=18*30=540 wavelengths. So the following sum is valid Cb1+Cav=Cinst or 114+426=540. The absolute error between the accurate and approximated methods is $\Delta C=|(Cav_2 - Cav_1)|=0.214$.

For the best study of the precision, 100000 experiments are executed. The average is 0.324037, the variance is 0.005072. The width of the variance is 0.47999999515712 with the minimum error equal to 0.088000000454485 and maximum error 0.56799999997019. So the range of the errors is separated in nine intervals (classes) with width 0.06. The classes are showed in the table 5 second column (class). The median of each class is showed in the table 5, third column (x_i). For the better representation of the data that are measured, the graphical method is used figure 2, curve absolute (the measured distribution) and figure 3, curve athristikh (the measured accumulated distribution). The parameters of the measured frequencies are showed in the table 6, second column (absolute).The parameters ASYM and CURV mean the coefficient of asymmetry and the coefficient of curve respectively. After it is tried to understand, model and analyze the relationship between a dependent variable (error frequency) and one independent variable (error class that is showed by its median). It is a regression analysis. The curves of figure 2 are the normal distribution and parabola one. The curves of figure 3 are the log normal distribution and the parabola one. The parameters of the curves of figure 2 are showed in the table 6. The coefficient of correlation between the random variable error class and error frequency is a quantitative index of association between these two variables. In its squared form, as a coefficient of determination R^2 indicates the amount of variance in the criterion variable error frequency that is accounted by the variation in the predictor variable error class. These coefficients for the curves of the Figures 2 and 3 showed in the next Table 7.

TABLE 5.THE CLASSES, THE MEDIAN OF EACH CLASS, THE RELATIVE AND ACCUMULATED RELATIVE FREQUENCIES OF THE ERRORS

S/N	CLASS	x_i	fi'	Fi'
1	(0.054,0.114]	0.084	0.026	0.026
2	(0.114,0.174]	0.144	1.28	1.306
3	(0.174,0.234]	0.204	8.568	9.874
4	(0.234,0.294]	0.264	24.003	33.877
5	(0.294,0.354]	0.324	32.386	66.263
6	(0.354,0.414]	0.384	23.652	89.915
7	(0.414,0.474]	0.444	8.795	98.71
8	(0.474,0.534]	0.504	1.257	99.967
9	(0.534,0.594])	0.564	0.033	100

TABLE 6.THE PARAMETERS OF THE ABSOLUTE, NORMAL AND PARABOLA DISTRIBUTIONS

PARAMETER	ABSOLUT	NORMAL	PARABOL
AVERAGE	0.324037	0.3251	0.324037
DEVIATION	0.005072	0.0041	0.005072
ST.DEVIATION	0.071217	0.064032	0.071217
ASYM	0.0061	0	-0.20141
CURV	2.75117	3	-17.9911

TABLE 7.THE COEFFICIENTS R^2 FOR THE CURVES OF THE FIGURES 2 AND 3

Figure	NORMAL	PARABOLA
2	0.9884	0.7313
Figure	**LOG NORMAL**	**PARABOLA**
3	0.9986	0.9323

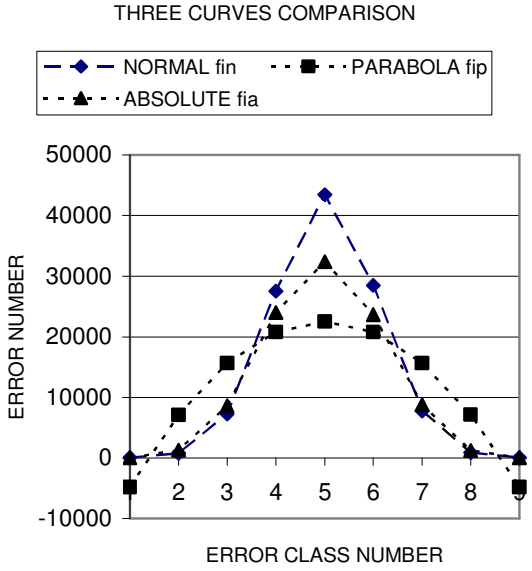

Figure 2.The three distributions.

The curve-fitting techniques will (in most practical cases) produce functions that will not precisely fit all the data points (sometimes none of the points exactly lie on the curve such that the residuals Rin=fin-fia and Rip=fip-fia are existed). These are showed for the curves of figure 2 in the figure 4. There are two fittings with their error patterns. The different two colors are indicating two different fits. It is obvious that the residuals are scattered between negative and positive. This bar chart gives limited visual impressions about the two fitting residuals. We can recognize a random pattern showing us the two fits of the residual distribution for the error class number from 1 to 9. Every abscissa should have two different color bars. The first fit is black and the second white with shadow. We can see that some bars are missing which indicates a zero (or a very small) residual. For example, we may say that at the error class number 1 and 9, the first fit is very accurate (we cannot see the bright black rectangle), while in error class number 5 the black rectangle is larger which means that the first fit function, represented by black color, has a very bad fit there. At the figure 5, the absolute residuals are showed and

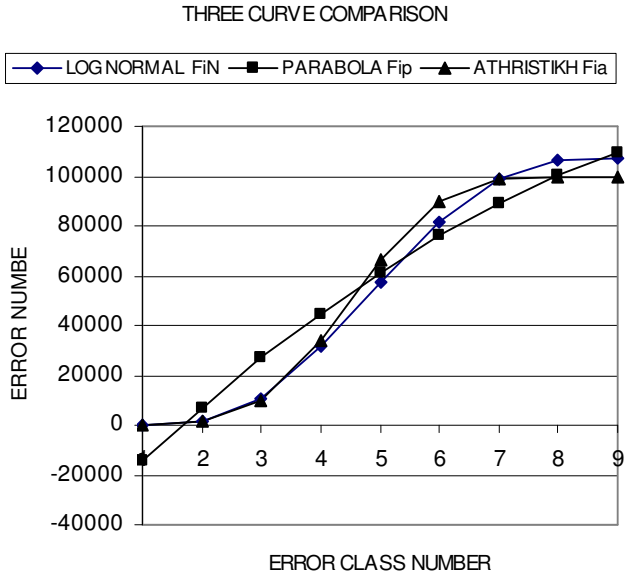

Figure 3.The three distributions (ATHRISTIKH means accumulated in Greek Language).

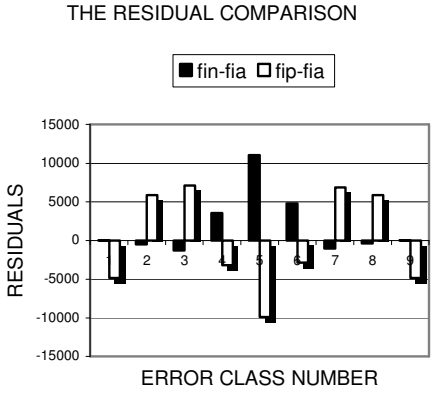

Figure 4.Residuals for curves of figure 2.

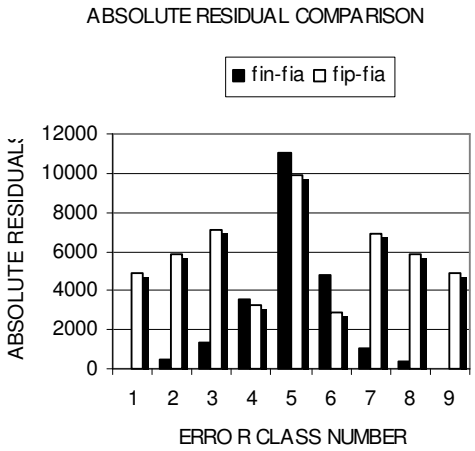

Figure 5.The absolute residuals for the curves of figure 2.

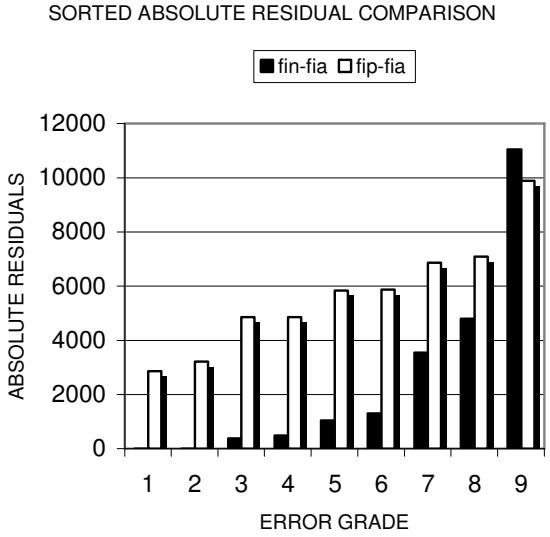

Figure 6.The sorted absolute residuals for the curves of figure 2.

the comparison is presented better. At the figure 6, for the better presentation of the fitting, the error class number is replaced by the error grade and the absolute residuals are sorted. The first error grade (1) represents the minimum absolute residuals (or deviations) for everything of the two fits and last one (9) represents the maximum absolute residuals. The error grade number two (2) also represents residuals that come in the second grade for everyone of the two fits, and so on. Therefore, instead of comparing only maximum errors (mostly occurring at different two points) as indications of how good is the fit, we may compare all the 9 grades of errors and find out which fit would sum up the least errors and that would be our required fit. The same analysis could be represented for the curves of the figure 3.

The complexity of this algorithm for the node number *q* depends on the square of the node number and the total number of the requests for connection (s) so it is written as

O ($s*q^2$). Time complexity of that algorithm is 'order q^2, O($s*q^2$). Consuming time for 100000 experiments for the accurate method is 16' 43'' 10 and for the approximate one 16' 46'' 46. On a 133MHz computer the approximate method consumes more time than the accurate one to solve the same problem in the same network. The worst consuming time depends of network size for the same computer. The matrix of coefficients of the accurate method is low triangular and of the approximate one is square. So the approximate method needs more memory for solving.

E. Discussion and Proposals

In this protection scheme, when a single failure of a cut link occurs and the main connection also cuts then the connection is routed by backup path.

 ISBN: 978-1-61208-132-8

The use of the polynomial methods of finite differences is possible for the study of the problems related to the protection and restoration of connections and has the advantage versus the linear method that is not using the large matrix with the active links of the network. The active links are the links that pass through the optical paths. For better presentation of this research a short example is used that depicts the results in these methods. The algorithm provides for each source-destination node pair and a desired connection group size, a value of the total available capacity of the network. The connection length depends on the number of hops. The network has a complete protection for optical path connection. It is a switch circuit network (or a packet switch network but the packets are adjusted to pass through preplanned paths) so that one lightpath corresponds to one optical connection. Different wavelengths may be used for each connection in each hop, so that wavelength conversion is used at each node.

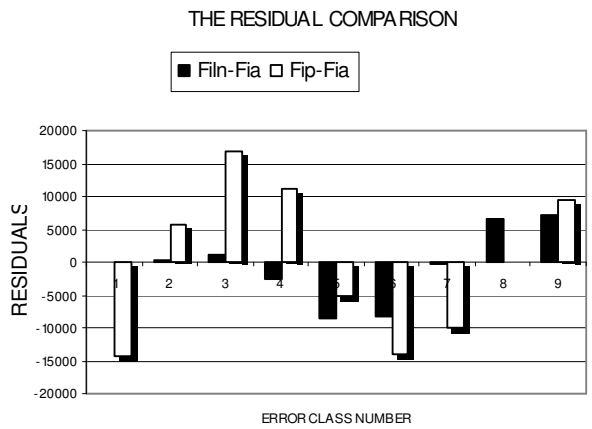

Figure 7.The residuals for the curves of figure 3.

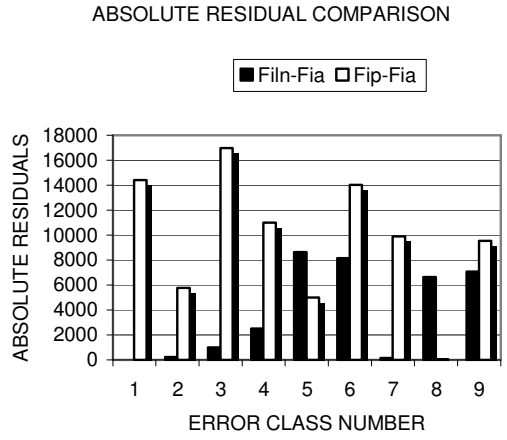

Figure 8.The absolute residuals for the curves of the figure 3.

These methods solve problems with small networks because when the connections groups that pass through a link increases, then the number of the (n+1) equations with n+1 unknown of the system also increases and it is difficult to be solved to calculate the coefficients. The accurate method is easier and more accurate than the approximate one because only one factor is added to each new equation when the polynomial degree increases versus one factor is added for all equations. In Turbo Pascal for the PC, the best precision is 15 decimal digits and some differences that appeared are ought to the difficulty to write all decimal digits for each polynomial coefficient of the polynomial methods. For the precision, the *double* type is used that is a floating point format and provides good dynamic range in addition to high precision.

The coefficient of determination R^2 value is an indicator of how well your data fits a line or curve. This coefficient value is in the range (0, 1). When a distribution (curve) has larger R^2 than another, it has also best dependent for its variables than another one. In other words, if the R^2 value is closer to 1, means more likely your data points are solutions to the equation that defines your curve. This indicates how well your data fits the model you are testing. The normal distribution has best fitting than the parabolic one to the measured distribution and the log normal distribution has also best fitting than the parabolic one to the measured accumulated distribution. These are showed in the figures, 2 and 3. For better presentation of the curve fitting, the residuals are showed in the figures, 4 up to 9.

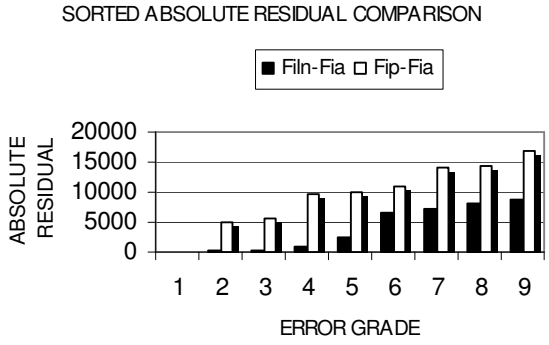

Figure 9.The sorted absolute residuals for the curves of figure 3.

The solution of the optimization problem that means the solution of a problem that more approaches to our demands is more complex. The general solution of this problem is difficult so it is transformed as a NP-complete problem. It is also noted that ILP formulations are practical only for small networks because the number of the equations and the variables should be as less as possible. The statistical method solves the problem but in a long time. For larger networks only heuristic methods are used because they are faster and give a very good, if not optimal, result. A heuristic method tries to solve the optimization problem in one or two steps ignoring whether the solution can be proven to be correct, a good or better solution is produced. So the computation performance is improved obtaining in a short time the solution of this problem at the cost of the accuracy. The simplex method is impractical for larger networks because there are a lot of scenarios, it is an ILP problem and the number of the variables and the number of equations also increase rapidly.

IV. CONCLUSION

 ISBN: 978-1-61208-132-8

In this paper, the dedicated protection optical path method has been researched on the basis of the polynomial methods of finite differences statistically. The methods of finite differences make it possible to research and study the dedicated protection problems of the optical paths. They provide suitable, accurate arithmetical methods to solve telecommunication problems such as planning of a completely protected network, complete protection for any failure occurs on optical path, node or link e. t. c.

REFERENCES

[1] M. Caroll, V. J. Roese and T. Ohara. "The operator's View of OTN Evolution," *IEEE Comms Magazine* September 2010, Vol 48, No 9, pp. 46-51

[2] A. Bononi. Optical Networking. Part 2, SPRINGER, 1992

[3] T. Wu. Fiber Network Service Survivability. ARTECH HOUSE, 1992

[4] J. Burbank "Modeling and Simulation: A practical guide for network designers and developers", *IEEE Comms Magazine*, March 2009, Vol 47, No3, pp. 118

[5] O. Gerstel and R. Ramaswami, Xros. "Optical Layer Survivability-A services perspective," *IEEE Comms Magazine* March 2000, Vol 38, No 3, pp. 104-113

[6] O. Gerstel and R. Ramaswami, Xros. "Optical Layer Survivability-An implementation perspective*," IEEE JSA of Communication*, October 2000, Vol 18, No 10, pp. 1885-1889

[7] H. Levy and F. Lessman. Finite Difference Equations. DOVER, 1961

[8] Canhui (S.), J. Zhang, H. Zang, L. H. Sahasrabuddhe and B. Mukherjee. "New and Improved Approaches for Shared – Path Protection in WDM Mesh Networks," *IEEE Journal of LightWave Technology*, May 2004, Vol 22, No 5, pp. 1223-1232

[9] S. Ramamurthy, L. Sahasrabuddhe and B. Mukherjee. "Survivable WDM Mesh Networks," *IEEE Journal of LightWave Technology*, April 2003, Vol 21, No 4, pp 870-889

[10] G. Carrozzo, St. Giordano, M. Menchise and M. Pagano. "A Preplanned Local Repair Restoration Strategy for Failure Handling in Optical Transport Networks*," Kluwer Academic Publishers Photonic Network Communications* 4:3/4, pp. 345-355, 2002

[11] St. Mylonakis. "WDM Mesh Networks with dedicated optical path protection with finite differences," *Fifth International Conference on Networking and Services, ICNS2009*, Valencia, Spain, April 2009

[12] Canhui (S.), J. Zhang, H. Zang, L. H. Sahasrabuddhe and B. Mukherjee. "New and Improved Approaches for Shared – Path Protection in WDM Mesh Networks," *IEEE Journal of LightWave Technology*, May 2004, Vol 22, No 5, pp. 1223-1232

 ISBN: 978-1-61208-132-8

Performance of Relay-Aided Distributed Beamforming Techniques in Presence of Limited Feedback Information

Alexis A. Dowhuszko, Turo Halinen, Jyri Hämäläinen, and Olav Tirkkonen
Department of Communications and Networking (Comnet), Aalto University
P.O. Box 13000, FI-00076 Aalto, Finland
E-mail: {alexis.dowhuszko, turo.halinen, jyri.hamalainen, olav.tirkkonen }@aalto.fi

***Abstract*—This paper studies the impact of channel signaling resolution on the performance of a (coherent) *distributed beamforming* (DBF) algorithm. This analysis is done in the context of a wireless access network, whose ultimate goal is to give adequate broadband coverage for users inside buildings. In this situation, instead of trying to reach the serving *base station* (BS) directly, we assume that each indoor subscriber receives assistance from a cooperative network that is deployed in its premises. This surrounding cooperative network is formed by a (relative) large number of low-cost *relay nodes* (RNs) with only one antenna. To simplify the analysis, communication in the first link (i.e., from the subscriber's terminal to RNs) is assumed costless, making the bottleneck lay in the second link (i.e., from RNs to serving BS). To carry out the analysis, a suitable closed-form approximation for the outage probability that correspond to a given received *signal-to-noise* power *ratio* (SNR) threshold is derived. Our analysis reveals that the power gain sacrificed when using a small amount of phase feedback information is not considerable in the light of the performance loss that is observed.**

***Keywords*-Cooperative Communications; Distributed Beamforming; Limited Feedback; Outage Probability; Relay Nodes.**

I. INTRODUCTION

Standardization activities on wireless communication networks advance rapidly, fueling the creation of new research topics to cope with the ever-increasing demand of higher data rates for mobile radio access. Given that most part of current voice calls and data usage takes place inside buildings [1], the provision of adequate broadband coverage in indoor environments is a crucial issue [2], [3]. Since indoor users are often behind walls with high attenuation, the penetration losses that radio signals experience put the mobile terminals in a very disadvantageous position, increasing the energy consumption required for signal transmission and reducing the amount of information that can be transfered effectively. Trying to give a solution to this problem, recent developments showed that such difficulties can be overcome by deploying femtocells (i.e., small and low-power wireless access points that connect mobile devices to the cellular network via the wired connection of the subscriber) [4]. Nevertheless, a rather different approach is analyzed in this work, where we propose to boost the communication performance of indoor users deploying a cooperative network in the premises of the subscriber. This network will be formed by a large number of low-cost *relay nodes* (RNs), that will assist the communication between the mobile terminal (i.e., main transmitter) and the network base station (i.e., main receiver) implementing a *distributed beamforming* (DBF) strategy.

The main idea behind DBF is simple: distribute a common message within many low-power single-antenna RNs, and then coordinate the re-transmission of this information in the direction of the intended destination (configuring a virtual transmit antenna array) [5], [6]. When the main transmitter and the RNs share the same environment (e.g., the same room), the communication in the first link can be carried out with almost no cost (in terms of time and power). In this situation, the bottleneck of this system lies in the second link, and can be mitigated by adjusting the channel response (i.e., amplitudes and phases) that the main receiver sees from each individual RN. This enables the coherent combination of the multiple replicas of the original message at the intended destination. The potential benefit of deploying a DBF scheme is well known in the literature: full diversity benefit and M-fold power gain for M active RNs in the network[1] [7]. However, in absence of *channel state information* (CSI) at the RNs, the use of distributed space-time coding was suggested to obtain full diversity gains in the second link (no power gain is possible in this situation) [8].

The main challenges in a DBF scheme are in the synchronization of the RF carriers of all RNs, and in the estimation of each individual channel gain that the main receiver observes in the second link. An adaptive 1-bit feedback DBF algorithm that tries to solve these problems was developed by Mudumbai *et al.* in [9]. The basic idea behind Mudumbai's DBF algorithm is interesting: make independent random phase adjustment at the RNs in each iteration, and retain only those phases that increase the received *signal-to-noise* power *ratio* (SNR) at the main receiver. Even though Mudumbai's DBF algorithm has shown many interesting convergence properties, in this paper we focus in a rather different approach. We assume that the main receiver has the capability to estimate the individual channels from each RN in the second link (using a N-bits uniform quantizer). Since the locations of RNs remain fixed during the whole data communication, only the phase portions of the channel gains are assumed to take random (unknown) values at the beginning of the phase configuration process. A closed-form approximation for the outage probability of this *deterministic* DBF algorithm is derived. Based on this analysis

[1]The power gain increases to a factor of M^2 if each RN transmits always at full power, independently of the number of active RNs in the network.

 ISBN: 978-1-61208-132-8

it is possible to conclude that a relatively small amount of phase signaling information (i.e., $N = 3$ phase feedback bits per channel) is sufficient to obtain a performance close to the one observed in presence of perfect channel phase information at the RNs.

The rest of the paper is organized as follows. Section II presents the system model, the assumptions on the limited feedback DBF algorithm, and the details of the performance criterion that will be used to carry out the analysis. Section III provides expression for the distribution of the received SNR, while Section IV presents the numerical results and studies the impact of the number of feedback bits per channel (i.e., N) and the number of active RNs (i.e., M) to the outage probability of the system. Finally, Section V presents the conclusions of the work.

II. System Model

The general layout of our cooperative relaying system is illustrated in Fig. 1. The system consists of a main transmitter, a main receiver, and M active RNs that share the same physical space with the main transmitter (e.g., the same room or office). All devices are equipped with a single transmit/receive antenna (in accordance with the low-cost requirement for RNs). In our system model, main transmitter and RNs operate in a half-duplex mode in a *decode-and-forward* (DF) fashion. Thus, during the first hop of duration T_1, message intended for the main receiver is sent from main transmitter to the nearby RNs. During the second hop of duration T_2, the message is sent from the RNs to main receiver. Attenuation on this first link is assumed to be small, and channel is either static or slowly varying (e.g., line of sight channel model). This makes possible to assume that communication on the first link can be accomplished with (almost) no cost in terms of power and/or time. The long distance between clustered RNs and main receiver implies a large attenuation on the second link, when compared to the attenuation on first link. This situation makes the second hop the bottleneck of the system, and its analysis the main goal of this paper.

As depicted in Fig. 1, a low-rate, reliable, and delay-free feedback channel exists between the main receiver and the active RNs. Main receiver uses this channel to convey a quantized version of the phase adjustment that each RN should apply in transmission (to maximize SNR in reception). In other words, the limited feedback information that main receiver reports is used to establish a *virtual antenna array* (VAA) in the second link. Note that since all RNs share the same physical location with the main transmitter, no multi-hop strategy is able to provide a better performance than the one obtained with a direct connection between main transmitter and main receiver. Therefore, the only valid option to reach the main receiver is that multiple active RNs transmit cooperatively at the same time, focusing the resulting VAA beam toward the direction of the intended destination over the second link.

Based on the above model, the received signal at transmission time interval i is of the form

$$r[i] = H[i]s[i] + n[i], \tag{1}$$

where $H[i]$ is the resulting sum channel, $s[i]$ is the complex modulation symbol, and $n[i]$ refers to an *additive white Gaussian noise* (AWGN) sample. Power control is not applied in the RNs and thus, the total transmit power P_t in the second hop remains fixed during the whole communication.

In case of unitary noise power, the received SNR in the second hop is given by

$$\Gamma[i] = \Big| \sum_{m=1}^{M} \sqrt{\gamma_m[i]}\, w_m[i] e^{j\psi_m[i]} \Big|^2, \tag{2}$$

where $\gamma_m[i]$ represents the received SNR from the m-th RN, $\psi_m[i]$ is the corresponding channel phase response, and $w_m[i]$ is the transmit weight that the m-th RN applies.

In our system model it is also assumed that:

- All devices admit fixed location. Thus, channel is not changing in time and we have $\gamma_m[i] = \gamma_m$, $\psi_m[i] = \psi_m$.
- Weights $w[i]$ are used to make phase adjustments in RNs. Required feedback message for adjustments may be spread over a time span, denoted by $i = 1, 2, \ldots, \mathcal{I}$.
- Performance analysis considers the resulting sum channel when all phase adjustments are done (i.e. after $\mathcal{I}$ time intervals). The corresponding weights in this situation are denoted by w_m (i.e., time index can be dropped since phase adjustments have been done and channel is static).
- Phases ψ_m are not calibrated in RNs, but they behave as independent, random samples. In the analysis we study the performance over any initial phase configuration. Therefore, we assume that phases ψ_m are *independent and identically distributed* (i.i.d.) uniform *random variables* (RVs) that take values on interval $(-\pi, \pi)$.
- To fulfill the accurate timing requirement, RNs monitor the standard synchronization signals either from destination (i.e., main receiver) or source (i.e., main transmitter).

The signal phase shifts that RNs apply easily create frequency selectivity, which is seen as an increased multipath effect in the resulting wireless channel. Yet, in our system model we assume that the synchronization error of RNs is small when compared to symbol length. Then, it is possible to assume that the duration of the effective channel impulse response is not (considerably) increased in this situation.

A. Assumptions on Limited Feedback Scheme

As shown in Fig. 1, in the second stage of the communication there are M active RNs transmitting a common symbol s to the main receiver. In order to maximize the SNR in the main receiver, each RN adjusts its transmission signal using a complex, individual beamforming weight

$$w_m = \sqrt{\frac{P_t}{M}}\, e^{-j\phi_m}, \quad \phi_m \in \mathcal{Q}, \tag{3}$$

$$\mathcal{Q} = \left\{ \frac{2\pi(n-1)}{2^N} : \; n = 1, \ldots, 2^N \right\}. \tag{4}$$

We note that the individual power (i.e., the module of w_m) is selected based on the number of active RNs in the cooperative system, so that total transmission power is always normalized

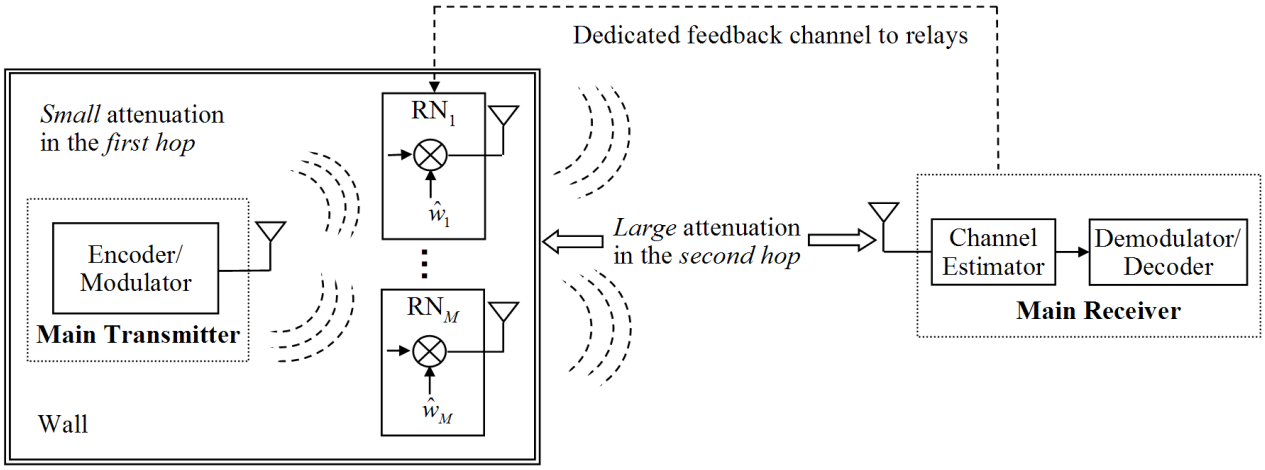

Figure 1. Cooperative relaying system model.

(and fair comparisons are possible). The error free feedback indicating the best index n is provided through the dedicated feedback channel. Number of feedback bits per RN is N.

Phases ϕ_m are selected in the main receiver as follows: Receiver first estimates the phases ψ_m from RN specific reference signals. After that, it selects the phases ϕ_m from quantization set $\mathcal{Q}$ such that $|\theta_m| = |\phi_m - \psi_m|$ is minimized. As a result, adjusted phases θ_m will be uniformly i.i.d. on the interval $(-\frac{\pi}{2^N}, \frac{\pi}{2^N})$ [10]. Thus, phase adjustments are done independently, using a common phase reference at the receiver side.

B. Performance Criterion: Outage Probability

There are two main performance measures that have been defined in the literature to carry out (theoretical) performance analyses: the ergodic capacity and the outage capacity [11]. The ergordic capacity is the long-term average transmission rate, and can be achieved implementing coding schemes that span code words over several coherence time intervals of the fading channel. This measure is feasible for applications with no strict delay constraints. However, in case of constant-rate delay-limited transmissions with coding over a single channel realization, the outage capacity becomes a more appropriate performance indicator. The outage capacity defines the maximum constant rate that can be maintained for a given outage probability.

In practical system implementations, however, a slightly different criterion is widely used. When mobile system performance is evaluated, it is assumed that reception is successful if SNR at the receiver (for the given transmission time interval) is large enough. In other words, a user is said to be supported if its instantaneous received SNR satisfies

$$\Gamma[i] \geq \gamma_0, \tag{5}$$

where the threshold γ_0 is defined to guarantee a certain level of service (for the given transmission rate). In this situation, the statistical performance requirement

$$\Pr\{\Gamma[i] \leq \gamma_0\} = \Pr{}_{\text{out}}(\gamma_0) \tag{6}$$

is defined as the outage probability for a given target SNR threshold. This is the performance measure that will be used throughout this work.

III. Performance of Distributed Beamforming with Limited Feedback

According to the system model presented in Section II, the relevant expression for SNR is of the form

$$\Gamma[i] = |H[i]|^2 = \frac{P_t}{M}\left|\sum_{m=1}^{M}\sqrt{\gamma_m}e^{j\theta_m[i]}\right|^2, \tag{7}$$

where individual received SNRs $\{\gamma_m\}_{m=1}^{M}$ are known beforehand, and remain constant during the whole communication process. Since we want to analyze the effect of the amount of feedback signaling (i.e., N) based on the performance measure presented in (6), a suitable expression for the *cumulative distribution function* (CDF) $F(\Gamma[i]|\gamma_1,\ldots,\gamma_M)$ should be obtained. Unfortunately, a tractable closed-form expression for this distribution can only be obtained for very specific situations (i.e., not for all M and N). However, since in this paper we are interested in studying the outage probability when the number of active RNs is high (i.e., when $M \geq 10$), we will use the central limit theorem to show that RV (7) can be successfully approximated as the sum of two independent *chi-squared* (χ^2) distributed RVs (one central and one non-central) with 1 degree of freedom each.

A. Central and Non-Central Chi-Square Distributions

Let $\{X_k\}_{k=1}^{n}$ be independent Gaussian RVs with common variance σ^2 and non-negative mean μ_k. Then, sum

$$Y = \sum_{k=1}^{n} {X_k}^2 \tag{8}$$

 ISBN: 978-1-61208-132-8

follows the *non-central* χ^2 distribution with n degrees of freedom [12]. The corresponding *probability distribution function* (PDF) expression in this situation is given by

$$f_Y(y)=\frac{1}{2\sigma^2}\left(\frac{y}{s^2}\right)^{\frac{n-2}{4}}\exp\left(-\frac{s^2+y}{2\sigma^2}\right)I_{\frac{n}{2}-1}\left(\frac{s}{\sigma_2}\sqrt{y}\right) \quad y\geq 0, \tag{9}$$

where

$$s^2=\sum_{k=1}^{n}\mu_k^2 \tag{10}$$

is the non-centrality parameter of the distribution, and

$$I_\alpha(x)=\frac{1}{2\pi}\int_{-\pi}^{\pi}\cos(\alpha\theta)\exp\left(x\cos\theta\right)d\theta \tag{11}$$

is the α-th order modified Bessel function of the first kind [13]. The characteristic function is also defined in closed-form, and it is given by

$$\Psi_Y(\omega)=\left(\frac{1}{1-2j\omega\sigma^2}\right)^{\frac{n}{2}}\exp\left(\frac{j\omega s^2}{1-2j\omega\sigma^2}\right). \tag{12}$$

We note that in the particular case when all means are zero (i.e., when $\mu_k=0$ for $k=1,\ldots,n$), the distribution of RV (8) reduces to the so-called *central* χ^2 distribution, whose PDF expression for n degrees of freedom is given by

$$f_Y(y)=\frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)\sigma^n}y^{\frac{n}{2}-1}\exp\left(-\frac{y}{2\sigma^2}\right) \qquad y\geq 0, \tag{13}$$

where

$$\Gamma(y)=\int_0^\infty t^{y-1}\exp\left(-t\right)dt \tag{14}$$

represents the Gamma function [12]. The characteristic function in this situation is given by

$$\Psi_Y(\omega)=\left(\frac{1}{1-2j\omega\sigma^2}\right)^{\frac{n}{2}}. \tag{15}$$

Let us now assume that

$$Z=Y_1+Y_2 \tag{16}$$

is the combination two independent χ^2 RVs: a non-central χ^2 RV with non-centrality parameter ${s_1}^2$ and variance ${\sigma_1}^2$, and a central χ^2 RV with variance ${\sigma_2}^2$, respectively. Let us also consider that the degrees of freedom are also equal in both cases (i.e., $n_1=n_2=n$). A typical way to obtain the distribution of RV Z is to calculate its characteristic function [14], i.e.,

$$\Psi_Z(\omega)=\left[\frac{1}{\left(1-2j\omega{\sigma_1}^2\right)\left(1-2j\omega{\sigma_2}^2\right)}\right]^{\frac{n}{2}}\exp\left(\frac{j\omega{s_1}^2}{1-2j\omega{\sigma_1}^2}\right). \tag{17}$$

This characteristic function can be inverse-Fourier transformed to yield the PDF

$$\begin{aligned} f_Z(z) &= \frac{1}{2{\sigma_1}^2}\left(\frac{\sigma_1}{\sigma_2}\right)^n\left(\frac{z}{{s_1}^2}\right)^{\frac{n-1}{2}}\exp\left(-\frac{z+{s_1}^2}{2{\sigma_1}^2}\right) \\ &\times \left[\sum_{k=0}^{\infty}\frac{\Gamma\left(\frac{n}{2}+k\right)}{k!\Gamma\left(\frac{n}{2}\right)}\left(\frac{\sqrt{z}\left({\sigma_2}^2-{\sigma_1}^2\right)}{s_1{\sigma_2}^2}\right)^k\right. \\ &\times \left.I_{n+k-1}\left(\frac{\sqrt{z}s_1}{{\sigma_1}^2}\right)\right] \qquad z\geq 0. \end{aligned} \tag{18}$$

It is also possible to show that the CDF in this situation admit the form

$$\begin{aligned} F_Z(z) &= \left(\frac{\sigma_1}{\sigma_2}\right)^n\sum_{k=0}^{\infty}\frac{\Gamma\left(\frac{n}{2}+k\right)}{k!\Gamma\left(\frac{n}{2}\right)}\left(\frac{{\sigma_2}^2-{\sigma_1}^2}{{\sigma_2}^2}\right)^k \\ &\times \left[1-Q_{n+k}\left(\frac{s_1}{\sigma_1},\frac{\sqrt{z}}{\sigma_1}\right)\right] \qquad z\geq 0, \end{aligned} \tag{19}$$

where

$$Q_m(a,b)=\int_b^\infty x\left(\frac{x}{a}\right)^{m-1}\exp\left(-\frac{x^2+a^2}{2}\right)I_{m-1}\left(ax\right)dx \tag{20}$$

is the generalized m-th order Marcum Q function [12].

B. Probability Distribution Approximation for Received SNR

Due to the Euler's formula, the RV

$$H[i]=\widetilde{X}_R[i]+j\widetilde{X}_I[i] \tag{21}$$

can be written in terms of its real and imaginary parts:

$$\widetilde{X}_R[i] = \sqrt{\frac{P_t}{M}}\sum_{m=1}^{M}\sqrt{\gamma_m}\cos\theta_m[i], \tag{22}$$

$$\widetilde{X}_I[i] = \sqrt{\frac{P_t}{M}}\sum_{m=1}^{M}\sqrt{\gamma_m}\sin\theta_m[i]. \tag{23}$$

Based on the fact that M is *large*, we use the central limit theorem to claim that both, real and imaginary parts of $H[i]$ are Gaussian with means μ_R and μ_I, respectively [14]. Since the imaginary part of $H[i]$ is a sum of sine functions with symmetrically distributed phases, its mean equals zero[2]. Based on the discussion presented in Section III-A, we observe that it is possible to approximate the stochastic behavior of RV $|\widetilde{X}_I[i]|^2$ as a central χ^2 distribution with 1 degree of freedom. Similarly, it is possible to see that the expected value of the real part of $H[i]$ is non-negative (actually, $\mu_R=0$ only when $N=0$). So, we claim that the stochastic behavior of RV $|\widetilde{X}_R[i]|^2$ can be approximated as a non-central χ^2 distribution with 1 degree of freedom and non-centrality parameter s_1 (unknown for the moment).

One final detail needs to be checked, to use approximation (19) for modeling the probabilistic behavior of main receiver's SNR (i.e., $\Gamma[i]$): we need to show that the real and imaginary parts of $H[i]$ (i.e., $\widetilde{X}_R[i]$ and $\widetilde{X}_I[i]$) are independent RVs. In this particular case, since $\widetilde{X}_R[i]$ and $\widetilde{X}_I[i]$ are modeled as Gaussian distributed RVs (central limit theorem), independence requirement is guaranteed if correlation coefficient between both RVs equals zero [14]. Fortunately, it is possible to show that this condition is satisfied, since the covariance of $\widetilde{X}_R[i]$ and $\widetilde{X}_I[i]$

$$C_{RI}=\mathrm{E}\{\widetilde{X}_R[i]\widetilde{X}_I[i]\}-\mathrm{E}\{\widetilde{X}_R[i]\}\,\mathrm{E}\{\widetilde{X}_I[i]\}=0 \tag{24}$$

in our system setting (detailed proof omitted).

Finally, the parameters that are required to use approximation (19) (i.e., s_1, σ_1, and σ_2) can be obtained from the first

[2]Individual phases $\theta_m[i]$ are uniformly i.i.d. on interval $(-\frac{\pi}{2^N},\frac{\pi}{2^N})$ for all m, and the sine function is an odd function.

 ISBN: 978-1-61208-132-8

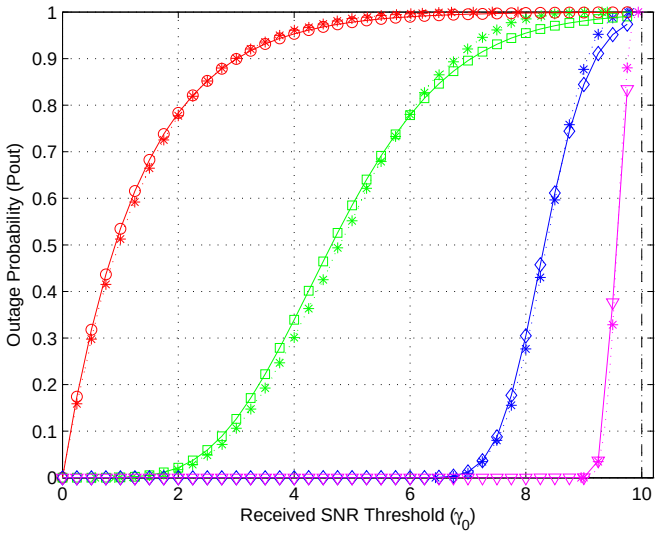

Figure 2. Outage probability as a function of SNR threshold γ_0 for DBF with 10 relays. Solid curves: No CSI ('$\circ$'), $N = 1$ ('$\square$'), $N = 2$ ('$\diamond$'), $N = 3$ ('∇'). Dashed-dotted line: Full CSI at RNs. Simulated values denoted by ('*'). Channel amplitudes are random but fixed samples from Rayleigh statistics.

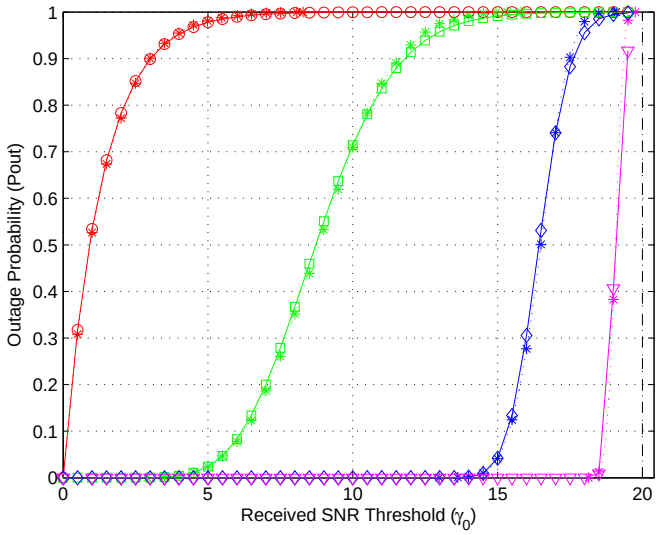

Figure 3. Outage probability as a function of SNR threshold γ_0 for DBF with 20 relays. Solid curves: No CSI ('$\circ$'), $N = 1$ ('$\square$'), $N = 2$ ('$\diamond$'), $N = 3$ ('∇'). Dashed-dotted line: Full CSI at RNs. Simulated values denoted by ('*'). Channel amplitudes are random but fixed samples from Rayleigh statistics.

two raw moments of RVs $\widetilde{X}_R[i]$ and $\widetilde{X}_I[i]$, whose closed-form expressions are obtained through simple but tedious computations:

$$\mu_R = \sqrt{\frac{P_t}{M}}\, C_N \sum_{m=1}^{M} \sqrt{\gamma_k}, \qquad \mu_I = 0, \tag{25}$$

$$\begin{aligned} \mathrm{E}\left\{\widetilde{X}_R^2\right\} &= \frac{P_t}{M}\left[\sum_{m=1}^{M} \gamma_m \left(\frac{1}{2} + \frac{1}{2} C_{N-1}\right)\right. \\ &+ \left. 2 \sum_{l=1}^{M-1} \sum_{m=l+1}^{M} \sqrt{\gamma_l}\sqrt{\gamma_m}\, {C_N}^2\right], \end{aligned} \tag{26}$$

and

$$\mathrm{E}\left\{\widetilde{X}_I^2\right\} = \frac{P_t}{M} \sum_{m=1}^{M} \gamma_m \left(\frac{1}{2} - \frac{1}{2} C_{N-1}\right), \tag{27}$$

with

$$C_N = \frac{2^N}{\pi} \sin\left(\frac{\pi}{2^N}\right). \tag{28}$$

IV. Numerical Results

In this section we analyze the performance of the proposed (coherent) DBF algorithm based on the previously presented approximation. To do so we study the outage probability for different amounts of channel phase signaling (i.e., diverse N), distinct channel amplitude models (dependent on the physical location of the cooperative RNs in the system), and for various numbers of active RNs (i.e., diverse M).

Regarding to the channel amplitude models we note that in all cases, total transmission power over all relays is 0 dB and signal gains $\sqrt{\gamma_m}$ are assumed to be fixed over the whole transmission period. In addition, in those cases where RNs are grouped in two different clusters (with exactly half the number of active RNs in each one), we use notation

$$\delta = \frac{\gamma_{(1)}}{\gamma_{(2)}} \tag{29}$$

to represents the power imbalance situation between both groups. Here, $\gamma_{(1)}$ and $\gamma_{(2)}$ represent the individual SNRs of the active RNs in the first cluster (stronger channel gains) and the second cluster (weaker channel gains), respectively. We will use the following models for the channel amplitudes:

- Amplitudes are random but fixed samples from i.i.d. Rayleigh statistics.
- Amplitudes admit perfect power balance (i.e., $\delta = 0$ dB),
- Medium channel power imbalance (i.e., $\delta = 6$ dB), or
- High channel power imbalance (i.e., $\delta = 10$ dB).

When channel amplitudes are (constant) Rayleigh distributed, it is assumed that individual channel SNRs are i.i.d. exponential distributed with unitary mean value.

Figure 2 and Fig. 3 show the outage probability for a SNR threshold when using the proposed DBF scheme for different amounts of channel phase signaling in case of $M = 10$ and $M = 20$ active RNs, respectively. In this scenario (constant) Rayleigh distributed channel amplitudes were used to model the amplitudes. The solid curves are plotted based on approximation (19) with appropriate fitting parameters, along with asymptotic upper bounds in case of full CSI at RNs (dashed line) [3]. In all cases, simulated point values ('$*$') are also included to verify the validation of the analytical results. Based on the results it is observed that our approximation follows simulated values well. As expected, the accuracy of

[3]Full CSI is actually a synonym of perfect channel phase information because no channel amplitude information is considered in this work.

 ISBN: 978-1-61208-132-8

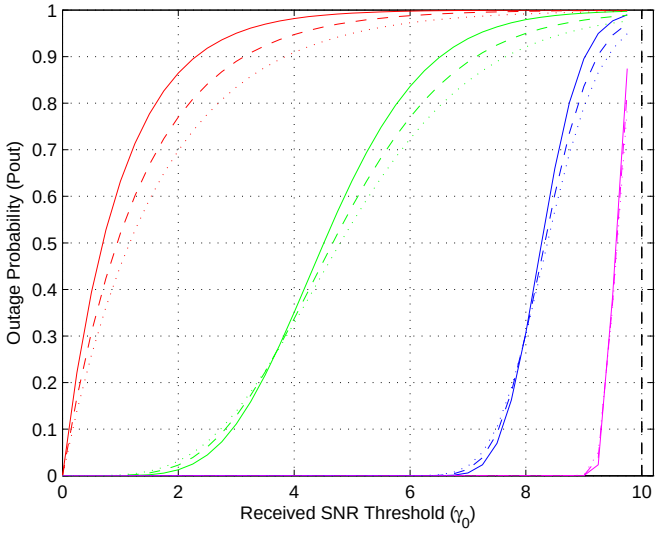

Figure 4. Outage probability as a function of SNR threshold γ_0 for DBF with 10 relays. Solid curves: Perfect channel balance (i.e., $\delta = 0$ dB). Dashed lines: Medium channel imbalance (i.e., $\delta = 6$ dB). Dotted curves: Large channel imbalance (i.e., $\delta = 10$ dB). Channel feedback: No CSI (red), $N = 1$ (green), $N = 2$ (blue), $N = 3$ (magenta). Dashed-dotted lines: Full CSI at RNs.

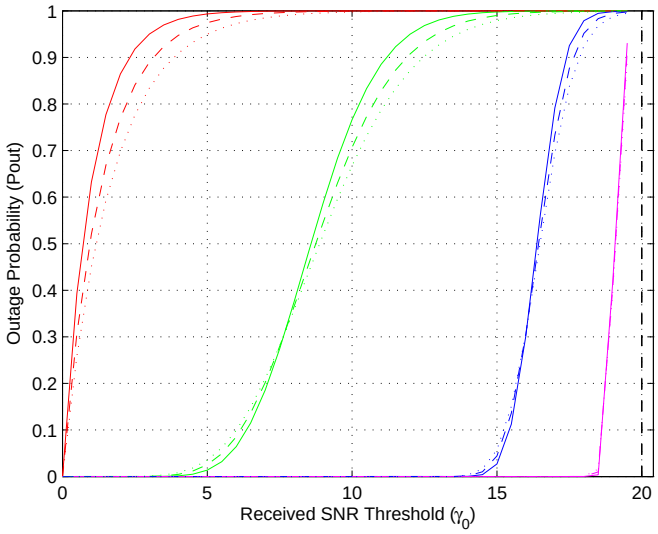

Figure 5. Outage probability as a function of SNR threshold γ_0 for DBF with 20 relays. Solid curves: Perfect channel balance (i.e., $\delta = 0$ dB). Dashed lines: Medium channel imbalance (i.e., $\delta = 6$ dB). Dotted curves: Large channel imbalance (i.e., $\delta = 10$ dB). Channel feedback: No CSI (red), $N = 1$ (green), $N = 2$ (blue), $N = 3$ (magenta). Dashed-dotted lines: Full CSI at RNs.

the approximation is better when the number of active RNs in the system is higher. The outage probability in absence of channel phase signaling is used as a baseline. It is found that performance in terms of outage probability clearly increases with additional phase bits in the feedback link. We also note that if $N = 3$, then the performance of DBF scheme is very close to the one observed with full CSI at RNs.

Figure 4 and Fig. 5 show the outage probability for given SNR threshold when implementing DBF algorithm in different channel power imbalance situations. In this case RNs are

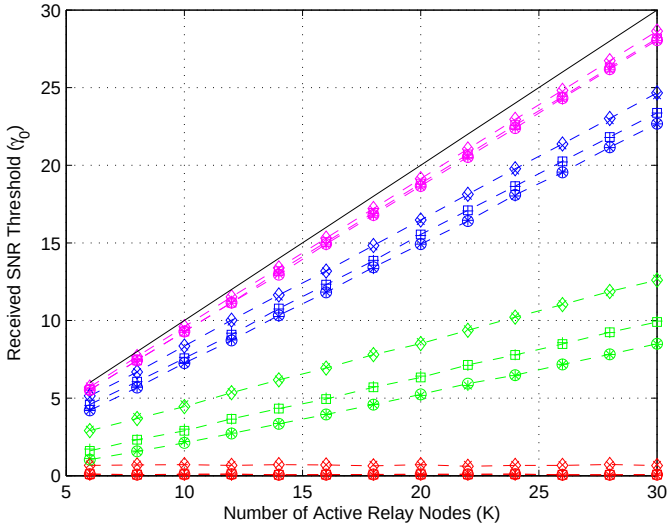

Figure 6. Required SNR threshold γ_0 for DBF to guarantee a given outage probability as a function of number active relays. Outage probability: $\Pr_{\text{out}} = 0.02$ ('$\circ$'), $\Pr_{\text{out}} = 0.1$ ('$\square$'), $\Pr_{\text{out}} = 0.5$ ('$\diamond$'). Channel feedback: No CSI (red), $N = 1$ (green), $N = 2$ (blue), $N = 3$ (magenta). Solid line: Full CSI at RNs. Simulated values denoted by ('*'). Channel amplitudes are fixed with perfect power balance.

grouped in two clusters (of the same size), that are located at different distances from the main receiver. Solid curves, dashed curves, and dotted curves represent perfect channel power balance (i.e., $\delta = 0$ dB), medium channel power imbalance (i.e., $\delta = 6$ dB), and high channel power imbalance (i.e., $\delta = 10$ dB) situations, respectively. Based on the results we observe that, the power imbalance level in the channel amplitude model increases the outage probability of DBF algorithm for low SNR thresholds. The larger is the number of phase bits N, the smaller is this impairment. The same behavior is visible when the number of active RN increases. This is because of the increasing variability that main receiver faces in its SNR in both situations, causing a less abrupt improvement on CDF curve as the value of γ_0 grows.

Finally, Fig. 6 presents the the maximum SNR threshold that can be guaranteed for a given outage probability when implementing our DBF algorithm in a perfect channel power balance case (i.e., $\delta = 0$ dB). These curves admit almost linear behavior with respect to the number of active RNs. Based on these curves we observe that, as N grows, the gap between the different outage probability curves decreases. This is in accordance with the behavior of the expected value of the real part of the sum channel (i.e., μ_R), given in equation (25) and presented in Fig. 7.

In the light of all results we see that there is no reason to use more than $N = 3$ bits for phase feedback per RN. Yet, the performance that is obtained with $N = 1$ bit is not good enough. However, the performance obtained with $N = 2$ bits represents a reasonable tradeoff between *the cost* of signaling overhead, and *the benefit* of the outage probability performance improvement that is observed.

 ISBN: 978-1-61208-132-8

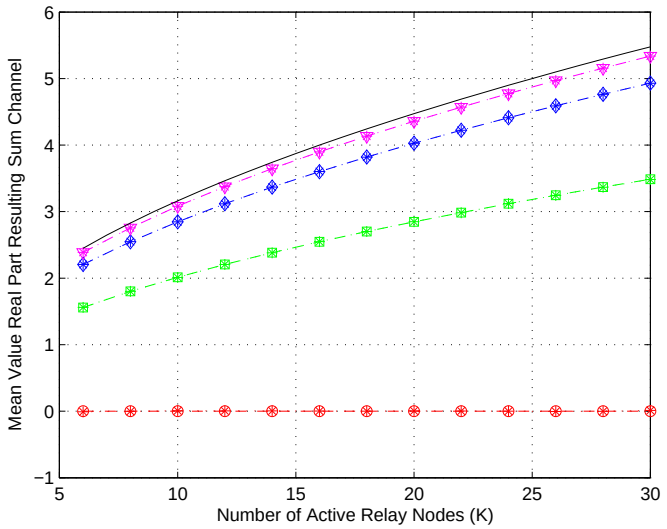

Figure 7. Expected value for the real part of the sum channel μ_R as a function of number active relays. Dashed curves: No CSI ('$\circ$'), $N = 1$ ('$\square$'), $N = 2$ ('$\diamondsuit$'), $N = 3$ ('∇'). Solid line: Full CSI at RNs. Simulated values denoted by ('*'). Channel amplitudes are fixed with perfect power balance.

V. Conclusion

In this paper we studied the performance of a *distributed beamforming* (DBF) algorithm in presence of different amounts of channel phase feedback information. This analysis is done in the context of wireless system, where the subscriber (main transmitter) receives assistance from a cooperative network to boost its communication to base station (main receiver). This cooperative network is formed by a large number of low-cost *relay nodes* (RNs), deployed in the premises of the subscriber. Location of the RNs are assumed to be fixed during the whole duration of the data transmission. Due to short distances, the communication over the first hop (i.e., from main transmitter to RNs) is assumed to be cheap in terms of transmission power and radio resource usage. Therefore, the bottleneck lies in the second hop (i.e., from RNs to the main receiver).

The outage probability for a given target *signal-to-noise* power *ratio* (SNR) is used as performance measure. To carry out the analysis, a suitable closed-form approximation for *cumulative distribution function* (CDF) of received SNR is derived. The parameters for this CDF approximation are obtained from the first two raw moments of the resulting sum channel that main receiver observes. The derived CDF expression is validated using simulations. Our analysis reveals that the use of DBF with a small amount of phase feedback information allows to reap a large fraction of the power gain that is available in the second hop.

VI. Acknowledgement

This work was prepared in *Spectrum Management for Future Wireless Systems* (SMAS) and *Interference Management for Wireless Networks Beyond Present Horizon* (IMANET) project frameworks and supported in part by Academy of Finland and Finnish Funding Agency for Technology and Innovation.

References

[1] G. Mansfield, "Femtocells in the US market–business drivers and consumer propositions," in *FemtoCells Europe*. AT&T, London, U.K., Jun. 2008.

[2] R. Pabst, B. Walke, D. Schultz, P. Herhold, H. Yanikomeroglu, S. Mukherjee, *et al.*, "Relay-based deployment concepts for wireless and mobile broadband radio," *IEEE Commun. Mag.*, vol. 42, no. 9, pp. 80–89, Sep. 2004.

[3] M. Husso, J. Hämäläinen, R. Jänti, J. Li, E. Mutafungwa, R. Wichman, *et al.*, "Interference mitigation by practical transmit beamforming methods in closed femtocells," *EURASIP Journal on Wireless Communications and Networking*, vol. 2010, pp. 1–12, Apr. 2010.

[4] V. Chandrasekhar, J. Andrews, and A. Gatherer, "Femtocell networks: a survey," *IEEE Commun. Mag.*, vol. 46, no. 9, pp. 59–67, Sep. 2008.

[5] Z. Ding, W. H. Chin, and K. Leung, "Distributed beamforming and power allocation for cooperative networks," *IEEE Trans. Wireless Commun.*, vol. 7, no. 5, pp. 1817 –1822, May 2008.

[6] X. Chen, S. Song, and K. Letaief, "Transmit and cooperative beamforming in multi-relay systems," in *Proc. IEEE Int. Conf. on Commun.*, May 2010, pp. 1 –5.

[7] R. Mudumbai, D. Brown, U. Madhow, and H. Poor, "Distributed transmit beamforming: challenges and recent progress," *IEEE Commun. Mag.*, vol. 47, no. 2, pp. 102–110, Feb. 2009.

[8] Y. Jing and B. Hassibi, "Distributed space-time coding in wireless relay networks," *IEEE Trans. Wireless Commun.*, vol. 5, no. 12, pp. 3524 –3536, Dec. 2006.

[9] R. Mudumbai, G. Barriac, and U. Madhow, "On the feasibility of distributed beamforming in wireless networks," *IEEE Trans. Wireless Commun.*, vol. 6, no. 5, pp. 1754 –1763, May 2007.

[10] J. Hämäläinen, R. Wichman, A. A. Dowhuszko, and G. Corral-Briones, "Capacity of generalized UTRA FDD closed-loop transmit diversity modes," *Wireless Personal Communications*, pp. 1–18, May 2009.

[11] E. Biglieri, J. Proakis, and S. Shamai, "Fading channels: information-theoretic and communications aspects," *IEEE Trans. Inf. Theory*, vol. 44, no. 6, pp. 2619 –2692, Oct. 1998.

[12] J. G. Proakis, *Digital Communications*, 4th ed. McGraw-Hill International Editions, 2001.

[13] M. Abramowitz and I. A. Stegun, *Handbook of mathematical functions: with formulas, graphs, and mathematical tables*. Dover Publications, 1970.

[14] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, 3rd ed. McGraw-Hill International Editions, 1991.

 ISBN: 978-1-61208-132-8

Tolerant Control Scheme Applied to an Aerospace Launcher

Nabila Zbiri
Laboratoire d'Informatique, de Biologie Intégrative
et des Systèmes Complèxes (IBISC)
Evry, France
Email: zbiri@iup.univ-evry.fr

Zohra Manseur
Mathematics Department
State University of New York
Oswego, USA
Email: zohra.manseur@oswego.edu

Iskander Boulaabi
Ecole Supérieure Des Sciences et Techniques de Tunis
Tunis, Tunisia
Email: iboulaabi@esstt.rnu.tn

***Abstract*—Autonomous aerospace launchers must carry out their missions safely because any accident can lead to important or dramatic consequences. It is essential to develop robust control solutions that guarantee optimal performances even when failures occur during these missions. The main objective of this paper is to present the development of a fault-tolerant control design in the case of an aerospace launcher. The method consists of defining and carrying out effective procedures for early detection of some critical situations and providing an adequate control that maintains the safe behaviour of the launcher. The improved control performance is obtained by using a sliding mode observer for a robust reconstruction of an actuator fault. This reconstruction is used then to generate an added signal in the initial control law that compensates for the effect of the faults. Simulation results will show the efficiency of the proposed method.**

Keywords-aerospace launcher; observer; stability; fault detection; control.

I. INTRODUCTION

The conquest of space is a technological battle that began decades ago, spurring great interest in researchers. Automatic applications in this field play an important role, particularly in modeling, control, and diagnosis aspects.

For the future needs of the CNES (The French Space Agency), it is useful to develop appropriate methodologies for piloting space vehicle launchers.

This work is within the framework of the PERSEUS Project which is a technology development program, undertaken as part of the research and innovation policy of the CNES Launcher Directorate [1]. The PERSEUS project has three objectives: the search for innovation and the development of promising technology applicable to Space transport systems; the undertaking of this work by young people within a university or association context, in order to encourage them to choose a career in space; and finally, the development of a set of ground-based and flight demonstrators in order to draw up a detailed pre-project file of a system for launching nano satellites.

Recently, some research in the Fault Detection and Isolation (FDI) area has led to systems based on the sliding mode idea [2], [3].

Although uncertainties could reduce the effects of faults in the control system and may cause false alarms, undetected faults could cause catastrophic consequences. In this context, Tan and Edwards [4], in 2003, extended their results obtained in 2000 [5] to design a sliding mode observer (SMO) that minimizes the L_2 gain between the uncertainty and the fault reconstruction signal to implement a robust faults reconstruction system.

The objective of the tolerant control system is to keep a safe behaviour for the system even in the presence of faults. Almost all the existing methods in the literature are divided into two classes: passive and active [6]. Passive techniques deal with an expected set of failures on the actuator and lead to a controller design that makes the closed-loop system insensitive to certain faults. Theses methods may lead to a very complex controller especially when the number of possible failures increases. Moreover, when unexpected failures occur, the controller is not capable of stabilizing the system. Active techniques use an FDI system and a control reconfiguration procedure that takes into account the effect of the fault. Different approaches, as model matching and track trajectory have been developed to improve system performances when a fault occurs.

This work improves the results described in an earlier article [7], where only a control scheme was developed on the launcher but where faults were not taken into account. The tolerant control developed here is based on the active technique. The FDI is built on results from [4] and [5] for a robust fault reconstruction. Unlike many previous active schemes found in the literature, the proposed method can be handled directly without completely reconfiguring the controller. A robust actuator fault reconstruction technique is applied to the process, allowing the compensation of the effect of the faults. The synthesis procedure is expressed in Linear Matrix Inequality terms.

Simulation results demonstrate the ability of the proposed fault tolerant scheme to detect actuator failures in real time,

 ISBN: 978-1-61208-132-8

identify them accurately with low computational overhead, and compensate for those actuator failures to achieve stability of the launcher around zero incidence.

This paper is organised as follows: Section 2 introduces the launcher model and its linear state representation. Section 3 explains the strategy of the detection and control system. Section 4 gives the proposed sliding mode observer, and a robust reconstruction technique for actuator faults is developed in Section 6. Finally, the conclusion is given in Section 7.

II. SPACE LAUNCHER DESCRIPTION

The launcher is assumed to be a rigid structure. Consequently, flexible modes are not considered in the launcher modeling but they may be taken into account as disturbances added to measures.

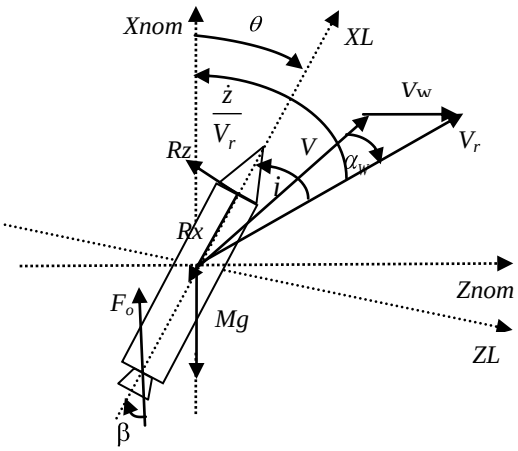

Figure 1. Exterior forces applied on the launcher

V, V_r : Absolute and relative velocity
V_W : Wind velocity
$\dot{z}$: Drift velocity along the pitch axis
i : Incidence of the vehicle
β : Thrust deflection angle
θ : Pitch angle (attitude)
α_w : Angle between V and V_r
C_x : Aerodynamic drag coefficient
C_z : Aerodynamic lift coefficient
R_x : Drag force
R_z : Lift force
F_o : Engine thrust
M : Instantaneous mass of the launcher
ρ : Air density
P_{dyn} : Dynamic pressure
S_{ref} : Reference surface of the launcher
X_F, X_A : Distance of the aerodynamic force and the propulsion control to the gravity center

Figure 1 illustrates the exterior forces acting on the launcher system. These forces are as follows:

- The gravity is given by $F_{grav} = M\, g$. (1)

- The dynamic pressure is expressed as $P_{dyn} = (1/2)\rho V_r^2$. (2)

- The aerodynamic force is given by $F_\alpha = (1/2)\,\rho V_r^2\, S_{ref}$. (3)

The aerodynamic force can be decomposed into two perpendicular forces:

- R_Z, the lift, is the component perpendicular to the trajectory; it is the most important force that carries the launcher:

$$R_z = (1/2)\,\rho V_r^2\, S_{ref} C_z \tag{4}$$

- R_x, the trail, is the weakest component and follows an axis parallel to the trajectory; it pulls up the launcher:

$$R_x = (1/2)\,\rho V_r^2\, S_{ref} C_x . \tag{5}$$

R_Z , the lift component, and R_X , the trail component are given by:

$$R_z = F_a C_z i \tag{6}$$

and

$$R_x = F_a C_x . \tag{7}$$

Considering small angles and applying the dynamic laws leads to the two principal equations modeling the launcher:

$$\begin{cases} \ddot{\theta} = A_6 i + K_1 \beta \\ \ddot{z} = -a_2\theta - a_1 i - (F_o / M)\beta \end{cases} \tag{8}$$

where

$$A_6 = \frac{F_0 C_Z\, X_F}{I_{YY'}},\ K_1 = \frac{F_0 X_A}{I_{YY'}},\ a_1 = \frac{F_0 C_z}{M}, a_2 = \frac{F_0 - R_x}{M} . \tag{9}$$

Finally, the launcher model can be represented by the vector equation [7]:

$$\begin{bmatrix} \dot{\theta} \\ \ddot{\theta} \\ \ddot{z} \end{bmatrix} = \begin{bmatrix} 0 & 1 & 0 \\ A_6 & 0 & \frac{A_6}{V_r} \\ -(a_1 + a_2) & 0 & \frac{-a_1}{V_r} \end{bmatrix} \begin{bmatrix} \theta \\ \dot{\theta} \\ \dot{z} \end{bmatrix} + \begin{bmatrix} 0 \\ K_1 \\ \frac{-a_1}{C_z} \end{bmatrix} \beta + \begin{bmatrix} 0 \\ \frac{-A_6}{V_r} \\ \frac{-1}{V_r} \end{bmatrix} V_w \tag{10}$$

$$i = \theta + \frac{\dot{z}}{V_r} - \frac{V_w}{V_r} \tag{11}$$

where the state vector is $x(t) = \begin{bmatrix} \theta & \dot{\theta} & \dot{z} \end{bmatrix}^T$. The input vector is β(t), the bounded external disturbance is V_w, and the output vector is $y(t) = \theta(t)$.

The coefficients A_6, K_1, a_1 and a_2 are the system variables that make the launcher's model non stationary and, therefore, difficult to control. Typical curves describing the variation of these parameters are shown in Figure 2 [8]. Operating points are chosen.

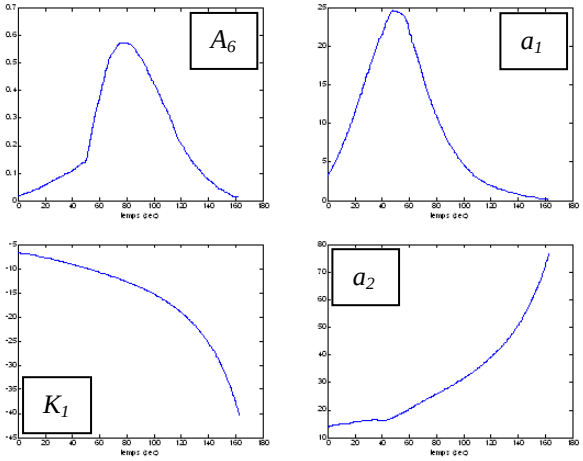

Figure 2. Evolution of the coefficients A_6, K_1, a_1 and a_2

III. Strategy Of Control And Fault Detection

The uncertain system (10) affected by actuator fault $f_a(t)$, can have the following form

$$\begin{cases} \dot{x}(t) = Ax(t) + Bu(t) + Ff_\alpha(t) + M\zeta(t,u,y) \\ y(t) = Cx(t) \end{cases} \quad (12)$$

where $x(t) \in \Re^n$ is the state vector, $u(t) \in \Re^m$ is the input vector, $y(t) \in \Re^p$ is the output vector, $\zeta(t,u,y) \in \Re^k$ includes uncertainties or perturbations affecting the system like the wind effect and $f_a(t) \in \Re^q$ is the actuator faults vector. The system matrices A, B and M are defined in the previous Section . F is the repartition matrix of faults. We assume that $\|f_a(t)\| \le \alpha(t)$ and $\|\zeta(t,u,y)\| \le \beta$, where $\alpha : \Re_+ x \Re^m \to \Re_+$ a known function is and β is a known positive scalar.

In order to eliminate the effect of the actuator fault, a new control law is added to the nominal one. Therefore, the control applied to the system is given by

$$u(t) = \bar{u}(t) + u_0(t) .$$

Then (12) can be rewritten as

$$\dot{x}(t) = Ax(t) + B\bar{u}(t) + M\zeta(t,y,u) + Ff_\alpha(t) + Bu_0(t) \quad (13)$$

where $\bar{u}(t) = -KX(t) = -K_x x(t) - K_z z(t)$ is the control component that minimizes a quadratic functional:

$$J = \int_0^\infty x^T Qx + \bar{u}^T R\bar{u}\ dt \quad (14)$$

where Q and R are diagonal matrices, weighting each state and control variables respectively in the common performance index (14), and the gain matrix K is determined from the expression:

$$K = -R^{-1}B^T P \quad (15)$$

where P is a positive matrix, solution of the well known Riccatti equation.

The additional control law u_0, compensating the effect of faults, can be implemented such that the faulty system (13) is as close as possible to the nominal system, therefore:

$$Ff_a + Bu_0 = 0 \quad (16)$$

and, if the matrix B is of full row rank, then:

$$u_0 = -B^+ Ff_a \quad (17)$$

where $B^+ = (B^T B)^{-1} B^T$ is the pseudo inverse of matrix B. In cases where the matrix B is not of full rank, the SVD theorem can be applied [9].

The control scheme is described in Figure 3.

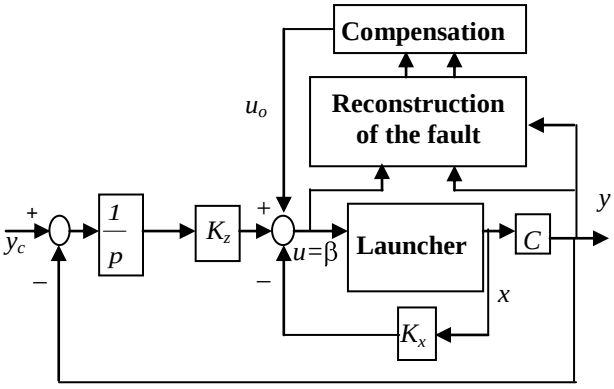

Figure 3. Tolerant control scheme

Using the reconstruction of the actuator fault $f_a(t)$ from the block "Reconstruction" determined in the next section, the component u_o is computed in the block "Compensation" as follows:

$$u_0 = -B^+ F\hat{f}_a$$

and we obtain the control law

$$u(t) = -KX(t) - B^{+}F\ \hat{f}_a(t)\,. \tag{18}$$

IV. SLIDING MODE OBSERVER

In the following work, a design method for a sliding mode observer for uncertain linear systems based methodology inspired from the work of Edwards and Spurgeon [10] is presented. The problem of a robust reconstruction of actuators faults can be implemented as shown on Figure 4.

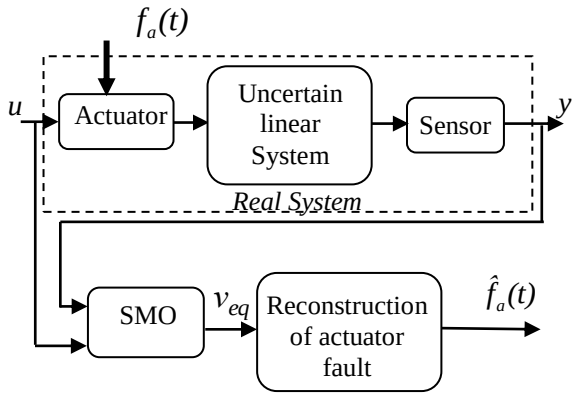

Figure 4. Reconstruction of actuator faults by a Sliding Mode Observer.

For the uncertain system (12), the structure of the SMO is defined by:

$$\dot{\hat{x}}(t) = A\hat{x}(t) + Bu(t) + G_l e_y(t) + G_n \upsilon \tag{19}$$

where $G_l \in \Re^{n\times p}$ is the linear gain and $G_n \in \Re^{n\times p}$ is the non-linear gain. The discontinuous vector υ is defined by

$$\upsilon = \begin{cases} -\rho(t,y,u)\dfrac{\bar{P}_0 e_y(t)}{\left\|\bar{P}_0 e_y(t)\right\|} & \text{if } e_y(t) \neq 0 \\ 0 \quad \text{otherwise} \end{cases} \tag{20}$$

where $e_y = \hat{y} - y$ is the output estimation error, $\bar{P}_0 \in \Re^{pxp}$ is a symmetric positive definite (spd) matrix that will be determined later and the value of the function $\rho : \Re_+ x \Re^p x \Re^m \to \Re_+$ is a known positive scalar that acts as an upper bound on the uncertainties and the faults.

Edwards, Spurgeon, and Patton [11] have shown that a sliding motion exists if:

- rank$(CF) = q$ (21)
- invariant zeros of the system (A, F, C) are stable.

If these conditions are satisfied, then there exists a change of coordinates such that the triplet (A, F, C) will be as follows:

$$\bar{A} = \begin{bmatrix} \bar{A}_{11} & \bar{A}_{12} \\ \bar{A}_{21} & \bar{A}_{22} \end{bmatrix} \qquad \bar{F} = \begin{bmatrix} 0 \\ \bar{F}_1 \end{bmatrix} \qquad \bar{C} = \begin{bmatrix} \end{bmatrix} \quad T^{-} \tag{22}$$

with $\bar{A}_{11} \in \Re^{(n-p)\times(n-p)}$, $\bar{A}_{12} \in \Re^{(n-p)\times p}$, $\bar{A}_{21} \in \Re^{p\times(n-p)}$, $\bar{A}_{22} \in \Re^{p\times p}$, $\bar{F}_1 \in \Re^{q\times q}$ is non singular and $T \in \Re^{p\times p}$ is orthogonal. Define $\bar{A}_{211}$ as the matrix obtained from the upper $(p-q)$ rows of $\bar{A}_{21}$. Tan and Edwards [5] proved that the pair $\bar{A}_{11}, \bar{A}_{211}$ is detectable since the unobservable modes of this pair are the invariant zeros of the system and they are stable. Define also $\bar{F}_2 \in \Re^{p\times q}$ to be the lower p rows of $\bar{F}$ such that $\bar{F}_1 \subset \bar{F}_2$. Then equations (12) are given by:

$$\begin{cases} \dot{\bar{x}}(t) = \bar{A}\bar{x}(t) + \bar{B}u(t) + \bar{F}\ f_\alpha(t) + \bar{M}\ \zeta(t,u,y) \\ \bar{y}(t) = \bar{C}\,\bar{x}(t) \end{cases} \tag{23}$$

Firstly, assume that G_n, in the new coordinates, is given by:

$$\bar{G}_n = \begin{bmatrix} -LT^T \\ T^T \end{bmatrix} \tag{24}$$

where $L = \begin{bmatrix} L_o & 0 \end{bmatrix} \in \Re^{(n-p)\times p}$ with $L_o \in \Re^{(n-p)\times(p-q)}$ and T is defined in (22). For the case, when

$\xi(t,y,u) = 0$ and $\rho = \|CF\|\alpha(t) + \eta_o$ with η_o is a positive scalar, the following results are proven in [5]:

Proposition 1. There exists a Lyapunov symmetric positive definite matrix $\bar{P}$ satisfying:

$$\bar{P}(\bar{A} - \bar{G}_l\bar{C}) + (\bar{A} - \bar{G}_l\bar{C})^T\bar{P} < 0 \tag{25}$$

with

$$\bar{P} = \begin{bmatrix} \bar{P}_1 & \bar{P}_1 L \\ L^T\bar{P}_1 & \bar{P}_2 + L^T\bar{P}_1 L \end{bmatrix} = \begin{bmatrix} \bar{P}_{11} & \bar{P}_{12} \\ \bar{P}_{11}{}^T & \bar{P}_{22} \end{bmatrix} > 0 \tag{26}$$

where $\bar{G}_l = TG_l$, $\bar{P}_1 \in \Re^{(n-p)\times(n-p)}$, $\bar{P}_2 \in \Re^{p\times p}$ and the matrix $\bar{P}_0$ in (20) is given by $\bar{P}_0 = T\bar{P}_2T^T$.

The state estimation error $\bar{e}(t) = T(x(t) - \hat{x}(t))$ is then quadratically stable. Furthermore, a sliding motion occurs in

 ISBN: 978-1-61208-132-8

finite time on $S_g = \{\bar{e} : C\bar{e} = 0\}$, governed by the stable matrix $(\bar{A}_{11} + L_o\bar{A}_{211})$. Then $\bar{e}(t) \to 0$ as $t \to \infty$.

In the following, these results to the case where $\xi(t, y, u) \neq 0$ are generalized. In this context, the state estimation error dynamical system is given by:

$$\dot{\bar{e}}(t) = (\bar{A} - \bar{G}_l\bar{C})\bar{e}(t) + \bar{G}_n v - \bar{F}f_a(t) - \bar{M}\zeta(t, y, u). \quad (27)$$

Suppose there exists a symmetric definite positive matrix $\bar{P}$ witch satisfies proposition 1. Define the positive scalars

$$\mu_o = -\lambda_{\max}(\bar{P}(\bar{A} - \bar{G}_l\bar{C}) + (\bar{A} - \bar{G}_l\bar{C})^T\bar{P}) \quad (28)$$

$$\mu_1 = \sqrt{\lambda_{\max}(\bar{M}^T\bar{P}^2\bar{M})}$$

where $\lambda_{\max}$ is the maximum eigenvalue. Suppose that

$$\rho(t, y, u) \geq \left\|\bar{C}\bar{F}\right\|\alpha(t) + \eta_o \quad (29)$$

where η_o is a positive scalar.

In terms of (27), (28) and (29) we have the following result in lemma 1 [4]:

Lemma 1. The norm of the state estimation error $\bar{e}(t)$ belongs to the set:

$$\Omega_\varepsilon = \left\{\bar{e} : \|\bar{e}\| < \frac{2}{\mu_o}\ \mu_1\beta\ + \varepsilon\right\} \quad (30)$$

where ε is an arbitrary small positive scalar.

Lemma 1 implies that the choice of $\rho(t, y, u)$ guarantees the sliding mode on S_g and provides an explication for the structures of the matrices defined by (22) after the coordinates change.

The application of a second change of coordinates defined in [5] by

$$\tilde{T} : \bar{e} \mapsto \tilde{e} : \qquad \tilde{T} = \begin{bmatrix} I_{n-p} & L \\ 0 & T \end{bmatrix} \quad (31)$$

where L is given by (24), transforms $(\bar{A}, \bar{F}, \bar{C})$ into the following matrices**:**

$$\tilde{A} = \tilde{T}\bar{A}\tilde{T}^{-1} = \begin{bmatrix} \tilde{A}_{11} & \tilde{A}_{21} \\ \tilde{A}_{12} & \tilde{A}_{22} \end{bmatrix}$$

$$\tilde{M} = \tilde{T}\bar{M} = \begin{bmatrix} \tilde{M}_1 \\ \tilde{M}_2 \end{bmatrix} = \begin{bmatrix} \bar{M}_1 + L\bar{M}_2 \\ T\bar{M}_2 \end{bmatrix} \quad (32)$$

$$\tilde{C} = \bar{C}\tilde{T}^{-1} = \begin{bmatrix} 0 & I_p \end{bmatrix}$$

$$\tilde{F} = \tilde{T}\bar{F} = \begin{bmatrix} 0 \\ \tilde{F}_2 \end{bmatrix}$$

where $\tilde{A}_{11} = \bar{A}_{11} + L_o\bar{A}_{211}$ and $\tilde{F}_2 = T\bar{F}_2$.

Thus, the nonlinear gain and the Lyapunov matrix become:

$$\tilde{G}_n = \tilde{T}\bar{G}_n = \begin{bmatrix} 0 \\ I_p \end{bmatrix}. \quad (33)$$

The new Lyapunov matrix is given by

$$\tilde{P} = (\tilde{T}^{-1})^T\bar{P}(\tilde{T}^{-1}) = \begin{bmatrix} \bar{P}_1 & 0 \\ 0 & \bar{P}_o \end{bmatrix}. \quad (34)$$

The new estimation error system is:

$$\dot{\tilde{e}}(t) = (\tilde{A} - \tilde{G}_l\tilde{C})\tilde{e}(t) + \tilde{G}_n - \tilde{F}f_a(t) - \tilde{T}\bar{M}\zeta(t, y, u). \quad (35)$$

Partitioning this error according to the dimensions of (35), we get

$$\dot{\tilde{e}}_1(t) = \tilde{A}_{11}\tilde{e}_1(t) + (\tilde{A}_{12} - \tilde{G}_{l_1})\tilde{e}_y - (\bar{M}_1 + L\bar{M}_2)\zeta(t, y, u) \quad (36)$$

$$\begin{aligned}\dot{\tilde{e}}_y(t) = \tilde{A}_{21}\tilde{e}_1(t) + \ & \tilde{A}_{22} - \tilde{G}_{l_2}\ \ \tilde{e}_y(t) \\ & + v - \tilde{F}_2 f_a(t) - T\bar{M}_2\zeta(t, y, u)\end{aligned} \quad (37)$$

where $\tilde{G}_{l_1}$ and $\tilde{G}_{l_2}$ are appropriate partitions of the matrix $\tilde{G}_l = \tilde{T}G_l$.

Proposition2: If the gain function $\rho(t, y, u)$ from (20) satisfies the inequality:

$$\rho(t, y, u) \geq 2\left\|\tilde{A}_{21}\right\|\mu_1\beta/\mu_o + \left\|\bar{M}_2\right\|\beta + \left\|\tilde{F}_2\right\|\alpha(t) + \eta_o \quad (38)$$

where η_o is a positive scalar, then a sliding mode occurs on S_g in finite time, with the presence of faults and matched uncertainties.

V. ROBUST RECONSTRUCTION OF ACTUATOR FAULT

In this part, assume that the SMO (19) is designed and can give a robust reconstruction of the faults $f_a(t)$ with minimization of the effect of $\zeta(t, y, u)$.

During the sliding motion, $e_y = \dot{e}_y = 0$, equations (36) and (37) become

$$\dot{\tilde{e}}_1(t) = \tilde{A}_{11}\tilde{e}_1(t) - (\bar{M}_1 + L\bar{M}_2)\zeta(t,y,u) \quad (39)$$

$$0 = \tilde{A}_{21}\tilde{e}_1(t) + \nu_{eq} - \tilde{F}_2 f_a(t) - T\bar{M}_2\zeta(t,y,u) \quad (40)$$

where ν_{eq} is the equivalent output error injection. ν_{eq} can be approximated to any degree of accuracy by replacing υ in (20) with:

$$\nu_{eq} = \begin{cases} -\rho(t,y,u)\dfrac{\bar{P}_0 e_y(t)}{\left\|\bar{P}_0 e_y(t)\right\| + \delta} & \text{if } e_y(t) \neq 0 \\ 0 \quad \text{otherwise} \end{cases} \quad (41)$$

where δ is a small positive constant representing the smoothing term. Since ν_{eq} is required for maintaining the sliding motion in presence of faults and uncertainties, the analysis of this term allows us to find the estimated actuator faults $\hat{f}_a(t)$. Now, define an estimate as

$$\hat{f}_a(t) = WT^T\nu_{eq} = G(s)\zeta(t,y,u) + f_a(t). \quad (42)$$

The transfer matrix *G(s)* is defined by

$$G(s) = W\bar{A}_{21}(sI - \tilde{A}_{11})^{-1}(\bar{M}_1 + L\bar{M}_2) + W\bar{M}_2 \quad (43)$$

where $\bar{A}_{21} = T^T\tilde{A}_{21}$ and $WT^TT\bar{M}_2 = W\bar{M}_2$. However, in this case, the transfer matrix *G(s)* links the exogenous input signal $\zeta(t,y,u)$ and the reconstructed faults signal $\hat{f}_a(t)$; thus, obtaining $\hat{f}_a(t) \approx f_a(t)$ (i.e., zero uncertainty case) is equivalent to minimizing the H_∞ norm of *G(s)*, with an appropriately chosen W. To formulate and solve this problem with LMI techniques, the *Bounded Real Lemma* [12] and a numerical development in [4] are used. Then, an optimization problem is address, in which $\left\|G_{\xi\hat{f}}\right\|_\infty < \gamma$, where γ is a positive scalar to be minimized with respect to the variable matrices $\bar{P}$, L, and W subject to the following matrix inequalities:

$$\begin{bmatrix} \bar{P}_{11}\tilde{A}_{11} + \tilde{A}_{11}^T\bar{P}_{11} & -\bar{P}_{11}\tilde{M}_1 & -(W\bar{A}_{21})^T \\ -\tilde{M}_1^T\bar{P}_{11} & -\gamma I & (W\bar{M}_2)^T \\ -W\bar{A}_{21} & W\bar{M}_2 & -\gamma I \end{bmatrix} < 0 \quad (44)$$

and

$$\begin{bmatrix} \bar{P}\bar{A} + \bar{A}^T\bar{P} - \gamma_o\bar{C}^T(D_dD_d^T)^{-1}\bar{C} & -\bar{P}B_d & E^T \\ -B_d^T\bar{P} & -\gamma_0 I & H^T \\ E & H & -\gamma_0 I \end{bmatrix} < 0 \quad (45)$$

where $D_d = [D_1 \;\; 0]$, $H = [0 \;\; H_2]$, $E = [E_1 \;\; E_2]$, $\bar{P}_{11}\tilde{A}_{11} = \bar{P}_{11}\bar{A}_{11} + \bar{P}_{12}\bar{A}_{21}$ and $\bar{P}_{11}\tilde{M}_{11} = \bar{P}_{11}\bar{M}_1 + \bar{P}_{12}\bar{M}_2$.

Note that inequalities (44) and (45) are affine with respect of the variables $\bar{P}_{11}$, $\bar{P}_{12}$, W and γ. Thus, the resulting observer is robust enough for the reconstruction of the faults, which affect the linear uncertain system, assuming that the linear gain $\bar{G}_l$ satisfies

$$\bar{G}_l = \gamma_o\bar{P}^{-1}\bar{C}^T(D_dD_d^T)^{-1}. \quad (46)$$

Inequality (44) is a necessary condition for the feasibility of inequality (45) and imposes the following equations $E_1 = -W\bar{A}_{21}$ and $H_2 = W\bar{M}_2$.

Consequently, this method consists of minimizing γ, with respect to the variables $\bar{P}$ and W subject to (44) and (45) where $\gamma_o \in \Re_+$ and $D_1 \in \Re^{p\times p}$ are arbitrary parameters which adjust the observer's gain. It's clear that when γ_o increases, the value of γ decreases, which results in $\bar{G}_l$ having a larger gain. Decreasing the gain of D_1 has the same effect. Let $\gamma_{\min}$ be the minimum value of γ satisfying (44). Then, equation (44) is a sub-block of (45), so, it is logical to always have $\gamma_{\min} \leq \gamma_o$. Moreover, to solve this convex optimization problem, a software like *MATLAB's LMI Control Toolbox* [13] is available to find γ, $\bar{P}$ and W.

The gain matrices can be obtained from [3] as

$$L = \bar{P}_{11}^{-1}\bar{P}_{12}, \quad \bar{G}_l = \gamma_o\bar{P}^{-1}C^T(D_dD_d^T)^{-1}, \quad \bar{G}_n = \begin{bmatrix} -LT^T \\ T^T \end{bmatrix},$$

$\bar{P}_o = T(\bar{P}_{22} - \bar{P}_{12}^T\bar{P}_{11}^{-1}\bar{P}_{12})T^T$.

The SMO is then completely determined.

VI. SIMULATION RESULTS

The simulation is carried out with Matlab software. The system parameters of the unstable launcher are given as

$$A_6 = 0.57,\ K_1 = -12.2,\ a_1 = 11.3,\ a_2 = 26.5$$

$$V_r = 544.53,\ F_o = 93.81,\ M = 3169.18$$

In Figure 5, the first two curves show the shape of the fault acting on the actuator at $t = 7$ seconds and its reconstruction. The third curve shows the fault reconstruction error. It can be seen that the fault $f_a(t)$ is faithfully reconstructed.

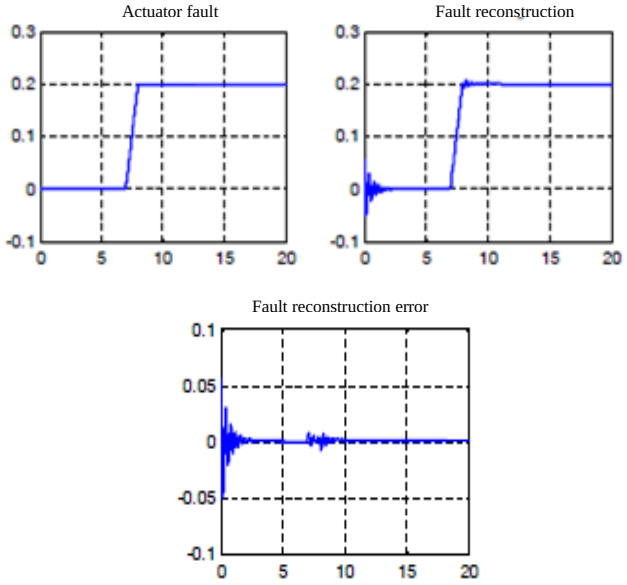

Figure 5. Actuator fault, fault reconstruction and fault reconstruction error

The fault reconstruction is then used to determine the additional term $u_o(t)$ in the control law $u(t)$ according to equation (18).

Figure 6 shows the evolution of the launcher attitude in the normal case (blue curve) and the estimation of its attitude obtained by the SMO observer when a fault appears on the actuator (green curve). To avoid bending forces that can destabilize the launcher, it is important to keep its attitude around zero. It is clear that the control law rejects the fault effect and stabilizes the launcher attitude.

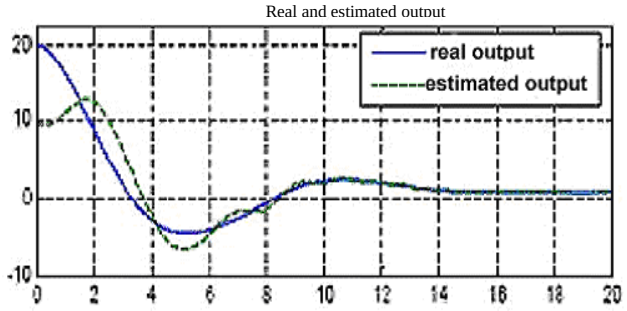

Figure 6. Evolution of the launcher attitude and its estimate

VII. CONCLUSIONS

In this paper, an approach for a robust control system based on fault estimates obtained by reconstruction techniques is proposed for an aerospace launcher system. An SMO was used to reconstruct actuator faults. This approach is based on the minimization of the effect of uncertainty on the faults reconstructed signal by the minimization of the H_∞ norm of the transfer matrix between the unknown inputs and the estimated actuator faults. A signal, built from the fault reconstruction, is then added to the control law and permitted the compensation of the fault effect. A numerical simulation example was provided to verify and validate the developed theoretical results.

Further work will extend the approach to non linear models and will specifically consider the launcher parameters' variations.

REFERENCES

[1] J. Kohlenberg, A. Voiculescu, E. Renault, C. Bac, Q. Vu Dang, "Collaborative Platform for Universities, Foster Clubs and Scientists in Aerospace Research", The Second International Conference on Systems and Networks Communications, ICSNC 2007, August 25-31, France, pp. 29-32, 2007

[2] C. Edwards and S. K. Spurgeon "Sliding Mode Control: theory and Applications," Taylor & Francis, 1998.

[3] V. I. Utkin, "Sliding mode in Control Optimization," Springer-Verlag, Berlin, 1992.

[4] C. P. Tan and C. Edwards, "Sliding mode observers for robust detection and reconstruction of actuator and sensor faults," International Journal of Robust and Nonlinear Control, vol. 13(5), pp. 433-463, 2003.

[5] C. P. Tan and C. Edwards, "An LMI Approach for Designing Sliding Mode Observers," Proceeding of the 39th IEEE Conference on Decision and Control, Sydney, Australia, pp. 2587-2592, 2000.

[6] Y.M. Zhang and J. Jiang, "Bibliographical Review on Reconfigurable Fault-Tolerant Control Systems" IFAC Annual Reviews in Control, pp. 229-252, 2008.

[7] N. Zbiri and Z. Manseur, "Control of an Aerospace Launcher" , Proceeding of the 2th Mediterranean Conference on Intelligent Systems and Automation Mars, pp. 377-380, 2009.

[8] D. Arzelier and D. Peaucelle, "Robust Impulse-to-peak synthesis: Application to the control of an aerospace launcher," IEEE International Symposium on Computer Aided Control Systems Design, pp. 214-219, 2004.

[9] G.H. Golub and C.F. Loan, "Matrix Computations", second edition. The Johns Hopkins University Press, 1989.

[10] C. Edwards and S. K. Spurgeon, "On the development of discontinuous observers," International Journal of control, vol. 59, pp. 1211-1229, 1994.

[11] C. Edwards, S. K. Spurgeon, and R. Patton, "Sliding mode observer for fault detection and isolation," Automatica, 36(4), pp. 541-553, 2000.

[12] M. Chilali and P. Gahinet,"H_∞ design with pole placement constraints : An LMI approach," IEEE Transactions on Automatic Control, vol. 41(3), pp. 358-367, 1996.

[13] P. Gahinet, A. Nemirovski, A. Laub, and M. Chilali,"LMI Control toolbox, User Guide''. Math Works, Inc., 1995.

 ISBN: 978-1-61208-132-8

An Intelligent System to Enhance Traffic Safety Analysis

Andreas Gregoriades
Dept of Computer Science and Engineering
European University Cyprus, Cyprus
a.gregoriades@euc.ac.cy

Neville Parker
Civil Engineering, CCNY-CUNY
Director CUNY ITS
New York, NY, USA
parker@utrc2.org

Kyriacos C. Mouskos
CTL Cyprus Transport and Logistics Ltd
Nicosia, Cyprus
mouskoskc@gmail.com

Ismini Hadjilambrou
CTL Cyprus Transport and Logistics Ltd
Nicosia, Cyprus
topoismini@gmail.com

Natalia Ruiz-Juri
The University of Texas at Austin
Austin, Texas, USA
natiruizjuri@gmail.com

Aneesh Krishna
Dept of Computing
Daculty of Science & Engineering
Curtin University of Technology, Australia
a.krishna@curtin.edu.au

Abstract. **Traffic phenomena are characterized by complexity and uncertainty, hence require sophisticated information management to identify patterns relevant to safety. Traffic information systems have emerged with the aim to ease traffic congestion and improve road safety. However, assessment of traffic safety and congestion requires significant amount of data which in most cases is not available. This work illustrates an approach that aims to alleviate this problem through the integration of two mature technologies namely, simulation-based Dynamic Traffic Assignment (DTA) and Bayesian Belief Networks (BBN). The former generates traffic information that is utilised by a Bayesian engine to quantify accident risk. Dynamic compilation of accident risks is used to gives rise to overall traffic safety. Preliminary results from this research have been validated.**

Keywords - Traffic Safety; Dynamic Traffic Assignment; Bayesian Belief Networks.

I. INTRODUCTION

It is well recognized that traffic accidents contribute substantially to urban congestion and traffic safety. Even a minor accident can cause significant traffic congestion in directly impacted areas which can cause safety issues to the overall road network due to secondary crashes as a result of rerouting [3]. Therefore, predicting traffic dynamics has always been an important issue in road safety. However, predicting the behavior of a road network under extreme conditions using historical records is a complex task. This work addresses this issue through the development of a novel Intelligent Traffic Information System that leverages the capabilities of two mature methodologies namely simulation-based Dynamic Traffic Assignment (DTA) [9] and Bayesian Belief Networks (BBN) [1]. The former is widely used in transportation planning and operations to predict drivers' decisions (where and when to travel on the road network), while the latter is powerful uncertainty modelling technique. Both technologies gained significant acceptance with the invention of powerful computational algorithms that enabled their exploitation.

Simulation-based DTA models depart from the traditional static analysis of traffic phenomena that employ analytic approaches to represent traffic conditions. DTA use traffic simulation to replicate the complex traffic flow dynamics especially for signalized systems where the vehicle and signal interactions are difficult to model analytically. This enables dynamic control and management systems to anticipate problems before they occur rather than simply reacting to existing conditions. The main output of DTA is the dynamic user equilibrium paths of each Origin-Destination (OD) pair [3]. These paths define the optimum route on the network that each vehicle will follow given its origin and destination.

BBN are ideal for modelling problems that are characterised by complexity, uncertainty and incomplete information. They have been used extensively in reliability engineering, risk management and decision support. In our case BBN are used to model and assess accident risk using dynamic and static transportation information.

The motivation of this work resides to the fact that current traffic information systems use multi-state safety databases that contain crash, roadway inventory, and traffic volume data. These are analysed to identify safety issues and evaluate the effectiveness of accident countermeasures. The main limitation of these systems is their retrospective approach. Effective safety management requires a prospective viewpoint. The proposed method combines dynamic modelling of traffic conditions with knowledge-based accident prediction to leverage the benefits of

 ISBN: 978-1-61208-132-8

computational intelligence in road safety and in this way provide forecasts of traffic safety. The paper illustrates the integration of BBN accident risk technique with the VISTA (Visual Interactive Systems for Transport Algorithms) simulated DTA framework. The method is applied to study the future behaviour of a road network in Cyprus.

The paper is organised as follows. Next section gives an overview of the literature. This is followed by the methodology. Subsequent sections concentrate on data pre-processing and BBN development. The integration of VISTA with the BBN along with the results that emerge from the amalgamation of the two technologies is described next. The paper finishes with the conclusions section.

II. RELATED WORK

Traffic information systems mainly provide retrospective analysis of traffic safety using historical data [2] due to, coverage, cost and real-time issues of traditional sensor-based schemes to traffic data collection. To escape from this problem simulation based traffic estimation systems emerged. The DTA simulation method employed herein constitute the state of the art in traffic forecasting. The two DTA approaches commonly used to emulate the path choice behavior of drivers are dynamic assignment en-route and dynamic equilibrium assignment. The former models behavioral rules that determine how drivers react to information received en-route while the latter only pre-trip path choices are considered and the goal is to minimize each driver's travel time by finding optimal or sub optimal paths [5]. This work is based on latter approach. Alternative simulation methodologies such as CORSIM, VISSIM, PARAMICS, WATSIM, AIMSUN do not have a true traveler behavior routing component [10]. They instead move traffic by splitting it probabilistically at every intersection. Thus the above-mentioned simulation models cannot be used to accurately predict traffic flow of each road section.

Related accident prediction approaches such as the one employed by Simoncic [6] utilise probabilistic modelling through BBNs. Their work illustrates the application of BBN to model road accidents and accordingly make inferences for accident analysis. The main limitation of this effort is that it concentrates solidly on the development of the BBN model without providing any substantial evidence of its performance. Work by Hu et al. [7] also uses a probabilistic approach to predicting road accidents through intelligent surveillance of vehicle kinematics; however, their method does not address the causal aspects that lead to observed behaviours and hence cannot be easily generalised. State-of-the-art tools in accident prediction, such as SafeNET 2 (Software for Accident Frequency Estimation for Networks), use traffic flows and geometric information to assess accident risk [8]. However, unlike our method, SafeNET 2 does not address the dynamic aspects of road networks using simulation. Hence, their traffic flow estimates are generic which in effect could lead in inaccurate conclusions.

III. METHODOLOGY AND RATIONALE

The traffic safety assessment system proposed herein is the amalgamation of probabilistic risk assessment [1] with mesoscopic traffic simulation [10]. The need for this integration boils down to the limitations of traditional traffic information systems that mainly concentrate of data warehousing. The methodology proposed utilises data marts to generate projections of future system behaviour. To that end, intelligent information management techniques have been employed to distil knowledge necessary for the development of models that enable the prospective analysis of system behaviour. The two models that emerged from this process are the accident risk assessment model and the traffic simulation model. The accident risk assessment employed is causality-based and uses a technique popular in the AI domain, namely, BBN. BBNs gained widespread acceptance with the introduction of computational algorithms that enabled their exploitation by Pearl [1]. BBNs are causal networks based on the concept of Bayesian probability, and provide a language and calculus for reasoning under uncertainty. A BBN in essence is a directed graph. It consists of vertices or nodes and directed edges (arrows). Each edge points from parent node to child node. In a belief network each node is used to represent a random variable, and each directed edge represents an immediate dependence or direct influence. Inference is achieved by belief propagation through the models topology.BBN technology is used to model how traffic and infrastructural factors influence accident risk. The second component of the approach is a road traffic simulator based on DTA. DTA evolved rabidly over the past two decades. This advancement has been fueled by the needs of application domains ranging from real-time operations to long term planning. DTA models constitute a natural progression in transportation [3] that evolved from static assignment approaches that assume that traffic flow is static and independent of time. One of the main features of DTA models is the dynamic analysis of road networks using time-varying traffic demands. DTA models effectively the complex interactions between supply and demand in a transport network. As a result, they capture the spatio-temporal trajectories (from origin to destination) of every vehicle and in return mimic in real-time the basic driver behaviors of road users. This constitutes a great advantage over traditional models that do not track the movement of individual vehicles but instead split traffic at intersections [3]. The DTA model is used in VISTA through the Dynamic User Equilibrium (DUE) model [4]. DUE assumes that no user can improve his/her travel time by changing their travel paths or by altering their departure or arrival times. DTA methods are divided into two groups the analytical and the simulation-based models. The former uses mathematical techniques to solve traffic problems while the latter

 ISBN: 978-1-61208-132-8

represents problems as a set of interrelated components that dynamically change. The use of DTA model enhances the limitations of existing practices by providing a consistent way of producing estimates of traffic flow conditions of road networks using limited information from traffic flow detectors. Moreover, it produces timely and complete traffic volume estimates for all sections of a road network and hence, can be used to assess accident risk using time varying conditions. The integration of BBN with VISTA in the proposed traffic information system enables the dynamic assessment of accident risk using simulated traffic conditions and prior knowledge embedded in the BBN. A pilot study conducted with the system aimed to assess the safety performance of the Nicosia road network in Cyprus and to investigate how it will behave under different scenarios.

Initially the road traffic model of Nicosia was specified, implemented, verified and validated in VISTA. Models in VISTA are represented by nodes connected by unidirectional links that represent flow of traffic in one direction. It is possible to have more than one link between two nodes to indicate separate lanes and lane direction. The completed VISTA simulation model was integrated with an accident risk assessor implemented in Java. The simulator provided the risk assessor with the traffic volumes of all road sections of the network for every 15 min interval. Traffic volumes in combination with infrastructural properties of the road network were used by the BBN to assess accident risk. Dynamic input to the BBN is provided by the DTA simulation on a step by step basis. For the development of the BBN topology and the parameterization of its prior knowledge, historical road accident data were complied. The BBNs accuracy was evaluated cross validation technique.

IV. SYSTEM ARCHITECTURE.

The proposed system was developed using the component-based software engineering methodology. With the initial specification of the system requirements we proceeded in the identification of suitable software components that matched the initial system requirements. These components were subsequently integrated to implement parts of the system's functionality. In particular the Bayesian inference engine and the charting components were selected after thorough investigation. The glue code that enabled components interaction was implemented in Java. The architecture of the system is comprised of five main components: the BBN accident assessor, the VISTA simulator, the data preprocessor, the results analyser and the results visualiser, as depicted in Figure 1. The risk assessor quantifies accident risk using a Bayesian inference engine that utilises a probabilistic model. Input to the BBN assessor is categorized into static and dynamic. The former is obtained from the VISTA database and the latter is the output of the VISTA simulation. Input evidence is pre-processed before running the BBN model that propagates the evidence down the network to produce the posterior probability. The integration of the VISTA with the BBN model was realised through asynchronous data interchange. Inherently, results from the VISTA simulator were accumulated in a database along with infrastructural information for each road section. The Risk Assessor component accesses the traffic flow and infrastructure records from the database and accordingly propagates the input evidence down the BBN network to produce the posterior probability for accident risk.

To establish communication between VISTA and the risk assessor it was imperative to pre-process VISTA's output data prior to being utilised by the BBN of the risk assessor. Specifically, VISTA variables are continuous by nature, while the BBN model uses categorical/discrete variables. Hence, it was necessary to discretise the output from VISTA prior to instantiating the BBN model. For the discretisation process it was necessary to refer to domain experts that specified the cut-off values for each variable. Specifically, for traffic volume three states were defined, namely, low, average and high. The first corresponding to less than 100 vehicles per 15 time interval, the second to less than 350 and the last to greater than 350.

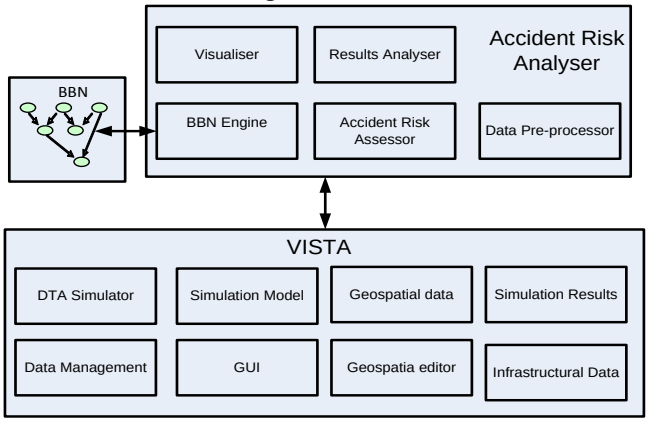

Figure 1. The System architecture

V. DATA COLLECTION AND PRE-PROCESSING

The development of the BBN required the specification of the topology and the parameterisation of its prior knowledge in conditional probability tables. To that end historical accident records were obtained from the Cyprus Police department specialising on traffic safety. Preliminary compilation of the data was performed with the SPSS statistical package. The accident dataset covered all accidents occurred in the Nicosia area from 2002 until 2008 and comprised over 9000 records. Each record consisted of 43 (six continuous and 37 categorical) input parameters covering global, local, temporal, accident, driver and car characteristics collected at the site of the accident by the police officers, eye witnesses and the involved parties. Each record was associated with a single categorical output parameter pertaining to accident severity, namely light,

 ISBN: 978-1-61208-132-8

severe and fatal, as evaluated by the police officer at the site of the accident.

However, for the development of the BBN's topology it was imperative to specify also the influence of infrastructural properties to the accident risk. However, due to limited information regarding infrastructural properties in the accident reports, it was necessary to map each accident on a geospatial GIS platform and subsequently import these on VISTA to obtain more information regarding the infrastructure at the accident location. Once this mapping was achieved additional information regarding the dynamic aspects of the road network at the accident scene was obtained from VISTA. This helped to define the causal relationships of the BBN variables that described the infrastructure and the traffic dynamics. Figure 2 shows the accidents geospatial layer of the dataset on ARCGIS using the baseline map. Each lollypop on the map corresponds to one or more accidents that occurred at the specified location.

Figure 2. Accident data on ARCGIS

The same accident dataset was used to identify locations on the network with high accident frequency the so called black spots. These points were used to validate the system after it was implemented. Specifically, a subset of the original dataset was used to validate the system. Black-spots (Figure 3) that were associated with that subset were used to test its performance on unknown conditions.

VI. DEVELOPMENT OF THE BBN

To develop the BBN model it was imperative to firstly identify the variables that adequately describe the problem, subsequently define the possible states that each variable could take and finally define the dependencies among them. A preliminary analysis of the accident data provided a generic indication of the influence of each variable to road accident. Database pre-processing involved two steps (a) replacement of missing and erroneous (e.g. falling outside the acceptable range) parameter values by the mean value of the parameter values of the other (assumed correct) records, and (b) grouping neighboring or related values of multi-valued (i.e. containing more than 12 values) categorical parameters so as to have a manageable number of intelligible as well as regular categories per parameter.

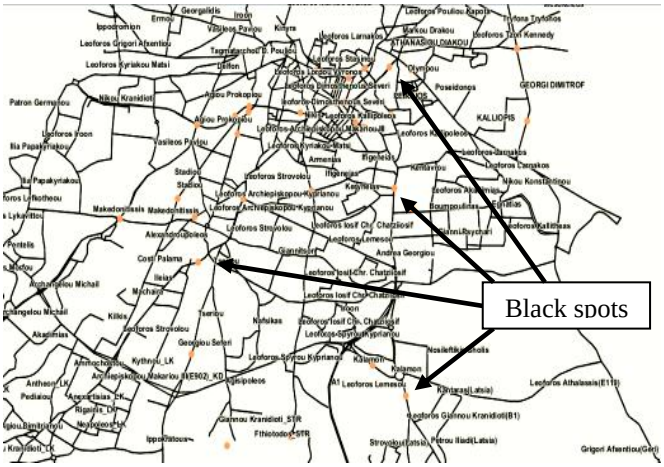

Figure 3. Network's Black Spots as overlaid dots

Statistical analysis relating the 43 input parameters (independent variables) to accident type (dependent variable) reveals that the Spearman correlation coefficient values between the inputs and the output are low (Figure 4), while the Spearman p-values are relatively high. Owing to the sufficient size of the database however, it is still possible for some of the correlations to be statistically significant. In support of that, accident type prediction was found far from satisfactory when only the statistically significant/correlated parameters were employed, thus demonstrating that statistically derived feature selection cannot be performed on a statistical basis for extracting the input parameters that affect the output and discarding those that do not provide accident type-related information.

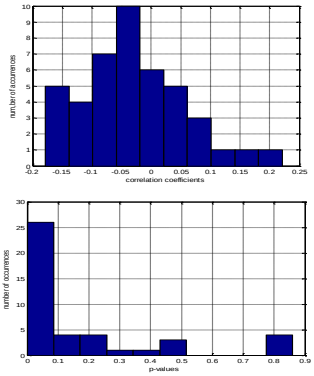

Figure 4. Statistically derived correlation coefficients (left) and p-values (right) between the 43 input parameters and accident type.

To that end, the processed accident data was subsequently used to identify the core variables of the BBN model along with their dependencies. In order to reduce the

complexity of the process and the model itself, the dimensionality of the initial data set was reduced using Principal Component Analysis (PCA). This helped to identify the core variables of the model. In principle, PCA projects the original data into a new set of orthogonal axes in such a way that the original multidimensional dataset with possibly correlated parameters is linearly transformed into a novel dataset of identical dimensions but with totally uncorrelated parameters. Owing to the fact that each new axis is selected so as to maximally expose the (remaining) variability of the dataset, it is not unusual for the first few axes of the PCA mapping to account for most of its variability. Hence, small PCA axes are generally sufficient in representing the original data with minimal loss of information. Results from this process yielded 12 artificial variables.

In addition, a subset of the accident data acquired by the police was mapped into ARCGIS geospatial application and subsequently on VISTA as a GIS layer. Afterward, VISTA was utilized to associate the accident parameters with the infrastructural properties of the road network at each accident point. Finally the traffic volumes along with traffic speed of vehicles were used to calculate traffic density of each road section of the network. These were subsequently associated with each accident record. Results from the dimensionality reduction using PCA together with the data merge that associated accidents locations with traffic density metric and infrastructure, yielded 19 candidate variables for the development of the BBN topology. The topology depicted in Figure 5, was initially learned from these processed data using the Expectation Maximisation algorithm. A refinement of the initial topology was performed using domain knowledge from the literature and subject matter experts.

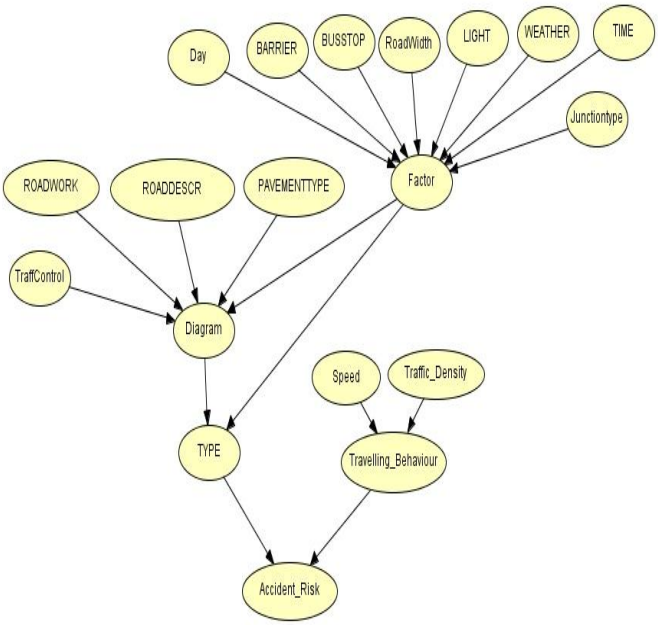

Figure 5. Learned BBN topology

To validate the effect of each variable to the target variable, namely, accident risk, a preliminary validation study was conducted using attribute relevance analysis (Fig. 6). Envisioner, data mining tool was used to compute the relevance between each identified causal factor to accident risk. The relevance of each variable to accident risk was compared against the learned conditional probability that emerged from EM algorithm and the posterior probabilities computed by the BBN with the independent instantiation of each of the leaf node variables.

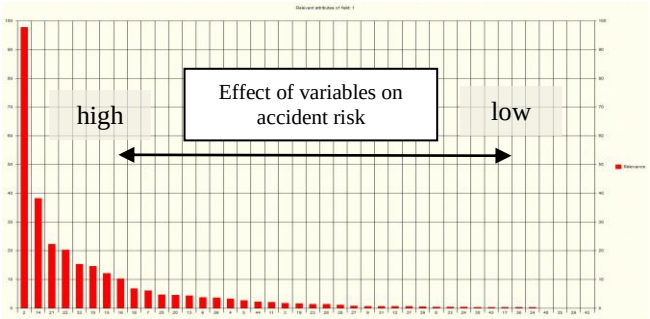

Figure 6. Relevance Analysis results.

VII. MODEL VALIDATION

To estimate the accuracy of the developed BBN model, five-fold cross validation has been employed. To that end the accident database that was enhanced with traffic data was randomly partitioned into five folds of equal number of records. Subsequently, and for each fold, four of the sets were employed for training the model while the remaining set was reserved for testing. Prediction accuracy was calculated by the weighted average of the test results of the five folds. Overall, the results of the validation process demonstrate that the model can accurately predict accident risks. However, the fact that the traffic volume used for validation is based on simulated results biases the outcome. To that end an additional validation study needs to be performed to verify that the model performs well in realistic settings.

VIII. RESULTS

Results from the accident risk assessor were used to calculate the accident risk index of each road section on the network. A road segment was labeled as accident prone if the predicted BBN accident risk probability was above a pre-specified threshold value. BBN estimates that fall below the cutoff value were ignored. This enables the safety engineer to alter the granularity of the analysis by altering the threshold value. To produce the accident risk index is was imperative to normalize the number of accidents that were predicted by the BBN with the traffic volume per time interval, for each road section. This aimed to escape the Simpsons paradox that defines phenomena that falsely prove the reverse of the truth. Inherently the Sympson's Paradox implies false causation, a logical fallacy by which two events that occur together are claimed to be cause and

 ISBN: 978-1-61208-132-8

effect. For example: statistically more accidents occur while the weather is good. Therefore, one erroneously could infer that good weather causes road accidents. The above argument commits to this fallacy, because in fact the explanation is that in good weather more cars are in the road and this causes more accidents. To find the actual effect of weather on accidents it is hence important to normalize the accidents that occurred in relation to the cars that are in the road. To that end, the proposed system uses a systematic approach that utilizes the traffic volume estimates from the VISTA simulation and the accidents predicted using the BBN risk assessor. Traffic volume acts as a normalizing factor for the number of accidents predicted using the BBN risk assessor. This gives rise to the accident risk index for each road section of the network that inherently highlight network's black spots. An illustration of the results produced by the traffic safety system is depicted in Figure 7. This figure illustrates a subset of the results and indicate that sections with IDs, 3, 21 and 47 have the highest accident risk index. The system enables the safety engineer to provide appropriate countermeasures to alleviate the problem. These are reported in the system and subsequently implemented in the simulation model. Each countermeasure then undergoes an evaluation procedure in the system to verify that the problem is eliminated prior to being implemented.

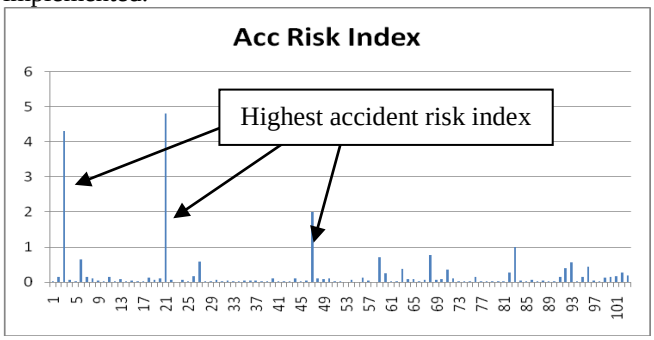

Figure 7. Links on the road network with highest accident risk index

IX. CONCLUSIONS AND FUTURE WORK

The traffic information system described herein illustrates a novel approach to quantifying road safety using probabilistic inference with DTA simulation. Integration of VISTA with BBN, as presented, enables the combination of known and uncertain evidence for accident risk quantification. The system combines state of the art technologies in traffic simulation and accident risk assessment. Integration of these provides the safety engineer with the necessary mean to perform a holistic and intelligent analysis of road safety. The method escapes from the problem of traffic data shortage that most traditional approach are suffering, through the use of DTA simulation. VISTA provides complete traffic volume data estimates for all road sections of the network on a 24 hour basis. This constitutes advancement over existing methods that base their analysis on limited data obtained from a scarce number of traffic sensors on the network. Therefore, the proposed method and the supporting system enable the identification of safety hazards in road networks using dynamic data and thus improved safety analysis that escapes the Sympson's paradox.

It should be noted that, once a black spot is identified, it requires a microscopic safety analysis to examine the behavioral aspects of the vehicle kinematics that leaded to an accident. This requires the development of the corresponding microscopic simulation. This requires detailed specification of roadway geometry, comprehensive traffic control data, lighting, environmental and traffic conditions in a microscopic simulation model. A future enhancement of our method will be developed to incorporate the above attributes into a model that will inherit properties from the mesoscopic analysis described in this study.

ACKNOWLEDGMENT

This research was co-funded by the Cyprus Research Promotion Foundation under the program: Targeted International Collaboration (ΔΙΕΘΝΗ/ΣΤΟΧΟΣ/0308).

REFERENCES

[1] J. Pearl, Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Information. San Francisco: Morgan Kaufmann, 1988.

[2] C. Kapp, "Who acts on road safety to reverse accident trends", The Lancet, vol. 362, 2003, pp. 1125-1125.

[3] V. Sisiopiku and X. Li, "Overview of Dynamic Traffic Assignment Options" Proceedings, 2006 Spring Simulation Multiconference, Society for Modeling and Simulation International, 2006, Huntsville, AL, pp. 110-116.

[4] A. Ziliaskopoulos and S. Lee, "A Cell Transmission Based Assignment-simulation Model for Integrated Freeway/Surface Street Systems", Proc., 75th Transportation Research Board Annual Meeting, 1996, Washington, DC, pp. 427-444.

[5] M. Florian et al., "Application of a simulation-based dynamic traffic assignment model", European Journal of Operational Research, vol. 189, 2008, pp. 1381–1392

[6] M. Simoncic, "A Bayesian Network Model of Two-Car Accidents" Journal of Transportation and Statistics, vol. 7, no. 2,3, 2004, pp13-27.

[7] W. Hu, X. Xiao, D. Xie et al., "Traffic accident prediction using 3-D model-based vehicle tracking", IEEE Transactions on Vehicular Technology, vol. 53, no. 3, 2004, pp. 677-694.

[8] TLR, "SafeNET2", http://www.trlsoftware.co.uk/news/detail.asp?pid=51&iid=29, 2007.

[9] H. Lo, "A dynamic traffic assignment formulation that encapsulates the cell transmission model", In A. Ceder (ed.) Transportation and Traffic Theory, Pergamon, Oxford,1999, pp. 327-350.

[10] FHWA, "Surrogate Safety Measures From Traffic Simulation Models", FHWA-rd-03-050 U.S. Department of Transportation Federal Highway Administration Research, Development, and Technology, 2010

 ISBN: 978-1-61208-132-8