Validation of Performance Data using Experimental Verification Process in Wireless Sensor Network

Tiong Hoo Lim
Department of Computer Science
University of York, York
North Yorkshire, YO10 5DG
United Kingdom
Email: thlim@cs.york.ac.uk

Iain Bate
Department of Computer Science
University of York, York
North Yorkshire, YO10 5DG
United Kingdom
Email: iain.bate@cs.york.ac.uk

Jon Timmis
Department of Computer Science
Department of Electronics
University of York, York
North Yorkshire, YO10 5DG
United Kingdom
Email: jtimmis@cs.york.ac.uk

***Abstract*—Testing a new network protocol experimentally in WSNs is an important step prior to deployment because theoretical models and assumptions made often differ between real environmental properties and performance. It is imperative to ensure that the results obtained from the test are reliable and the performance observed in simulation is a valid representation of the real world. Thus there is a need to perform extensive experimental analysis and evaluation to produce results with an acceptable level of confidence. In this paper, we outline experimental statistical and analysis techniques that allow us to have some confidence in the results obtained are at least relevant to physical deployment. Using the results from hardware and software experiments, we apply our proposed Experimental Verification Process (EVP) to evaluate the performance of the Multimodal Routing Protocol (MRP) against Adhoc On-demand Distance Vector (AODV) and Not So Tiny-AODV (NST-AODV). With the EVP, we have improved the credibility of MRP.**

I. Introduction

Real world deployments of WSNs are usually hard to control and difficult to deploy. It is not always practical to generate all the possible test cases and results to validate the performance of network protocol in a live network or real world deployment. WSNs research community has relied on simulation tools or test-bed to test and evaluate their new algorithm or protocol as it allows significant levels of testing to be performed at reasonable cost. Unfortunately, many current simulators are developed with simplifying assumption about the underlying simulation models, that do not generate the same result as in the real deployment. This can affect the dependability and safety of WSNs [1]. Hence, it is not sufficient to use simulation alone as a validation tool. An alternative approach is to cross validate and evaluate the WSNs using small scale experiment on real hardware in a controlled environment. However, it is difficult and time consuming to test the real experiment sufficiently to have confidence that it will work in practice and pilot studies in the laboratory are not the same as the *real* environment. This has lead to many reports of failures when WSNs are deployed for real [2], [3].

In recent years, the use of hardware and software to test and validate network protocol in WSNs has increased dramatically to ensure that the observed performance improvement is reliable and valid. However, many of the published works are lack of proper evaluation of the experimental results that can lead to false conclusion and affect the credibility of the works. The evaluation is usually not performed thoroughly enough to validate the claim of improvement made by a new protocol. For the experimental finding to be reliable and credible, it is essential that any significant results observed are at a specific *statistical confidence level* sufficient for the reliability targets of the system, *consistent*, *unbiased* and *repeatable* over time [4]. If the results can be reproduced under similar experimental condition, then the protocol under evaluation can be considered to be dependable.

In this paper, an Experimental Verification Process (EVP) is proposed that can be used evaluate the results obtained from both hardware and software in order to validate the performance of a network protocol. The EVP is based on Conceptual Statistical Test Framework (CSTF) that uses extensive testing in simulation to confirm that the simulation is correctly conducted and the obtained results are a valid representation of the real hardware. Using state of the art statistical techniques, the large number of testing helps to reduce statistical variation errors in the simulation. The real world testing allows us to confirm the trends of simulation and understand the degree of similarity between the two. The main objective of this paper is to provide a systematic approach to improve the credibility of an experimental study. To the best of our knowledge, this is the first time that a comprehensive analysis approach has been proposed to evaluate the credibility of the study of WSNs.

The main contributions of this paper are:

- Formulation of a systematic experimental approach to improve the reliability and validity of a WSNs experiment.
- Quantitative evaluation of an experiment to verify the performance of the Multimodal Routing Protocol (MRP) [5] is significantly better than Adhoc On-demand Distance Vector (AODV) [6] and Not So Tiny-AODV (NST-AODV) [7] using the proposed EVP.

The rest of this paper is organised as follows: Section II discusses related works. We present the EVP in Section III and outline the experimental setup in Section IV before the results are evaluated and validated using statistical analysis tools in Section V. We analyse the benefit of the EVP in Section VI and present the conclusion in Section VII.

II. Related Work

Limited work has been done in WSNs to validate the credibility of the results obtained from hardware or software experiments using statistical hypothesis techniques. Work by Ivanov et al. [8] has validated the performance of a link-state ad-hoc routing protocol using the results of 16 real wireless ad-hoc nodes experiment with the results of the ns-2 wireless simulator. The results have shown that the simulated packet delivery ratio is very close (error rate of 1%) to the real emulated results, but the latency results show much deviation from the real experiment due to delay introduced by the hardware and operating systems. However, it is difficult to see the significance of the results due to the lack of statistical tests applied on the results.

In Pham et al. [9], the Castalia WSN simulator is used to evaluate the performance of tunable MAC protocol and validate with the results collected from 9 TelosB motes deployed in a 3 by 3 grid. The validation process is performed by calculating the average of packet reception rate (PRR) for all the links using 50 random runs obtained from simulator and compared against the PRR obtained from real hardware run. The study has found that, even with the complex radio model available in the Castalia simulator, the results obtained are still significantly different. However, the degree of difference is not validated using statistical techniques and the number of run in hardware experiment is small (one run).

In [5][1], the MRP has been proposed to operate in different routing protocols during network void. Individual node in the network can make its own routing decision to switch between routing modes autonomously with minimal network disruption. Significant performance improvements in terms of reliability and efficiency of message delivery over AODV and NST-AODV. The weaknesses of this work are that the evaluation is only performed in simulation making it unclear how well the approach would work on real hardware, and limited statistical analysis of the results was found.

III. Experimental Verification Process

In this section, we present the experimental objectives and framework that are used to test the confidence level of the results obtained from both hardware and software simulation using statistical technique. We apply the framework to affirm the performance of MRP is significantly better than AODV and NST-AODV, and validate the credibility of the experiments.

A. Experimental Objectives and Hypotheses

The work in this paper has three objectives:

Objective 1: to demonstrate that MRP has a better network reliability with lower latency and greater energy efficiency than AODV and NST-AODV.

Objective 2: to demonstrate that the simulation results obtained are a valid representation of the WSN's performance.

Objective 3: to give statistical confidence in the results.

[1]downloadable at http://rtslab.wikispaces.com/file/view/mrp.tar

In order to perform the test to achieve these objectives, we formalise a set of hypotheses:

Hypothesis 1 (H_1): There is no significant improvement in the packet delivery between MRP, AODV and NST-AODV.

Hypothesis 2 (H_2): There is no significant difference in the latency between MRP and AODV, and MRP and NST-AODV.

Hypothesis 3 (H_3): The total routing packet generated by MRP during route failure is no different from AODV and NST-AODV.

Hypothesis 4 (H_4): There is no difference in the results obtained from the simulation and the real hardware implementation.

B. Distribution of Data

It is common in many works to report the mean of a number of runs, and in some cases standard deviation. However, the use of such statistics relies on an underlying assumption that the results follow a normal distribution. This is more than likely not the case, and is very hard to prove. Therefore, a safer alternative is to assume results do not have a normal distribution and employ non-parametric statistics. These non-parametric techniques are suitable for use on data that is normal and non-normally distributed, and therefore in many cases more appropriate.

For results in this paper we present the median. The median is computed by arranging the data in the order of magnitude and is represented by the midpoint of the data set. We also determine the 1st quartile (Q1) and 3rd quartile (Q3) of the data sets by identifying the first quarter of the data set as Q1 and third quarter of the data set as Q3. The quartiles show how the data are distributed on either side of the median and the difference between Q3 and Q1 is known as the inter-quartile range (IQR). As a general rule of thumb, any results outside 1.5 times the IQR can be identified as outliers.

C. Conceptual Statistical Testing Framework

Figure 1 illustrates our statistical testing framework to test the hypotheses and achieve the experimental objectives. The components of this framework are discussed in the following subsection where subsections 1, 2, 3 correspond to the labels (1, 2, 3) in the Figure 1.

1) Statistical Significance: A statistical significance test can be used to determine whether the difference in performance observed in the results is likely to have occurred due to random chance with the samples available, i.e. whether protocol X is really better than Y or whether the results are so close than differences are purely random. In order to determine and compare the relationship between the two samples collected from the experiments, Mann-Whitney-Wilcoxon, also known as rank-sum test, [10] is applied.

This non-parametric test is used, as previously discussed, such test do not make use of any underlying assumption about the distribution of data, avoiding the need to verify the data conform to the test assumption. Another benefit of using the

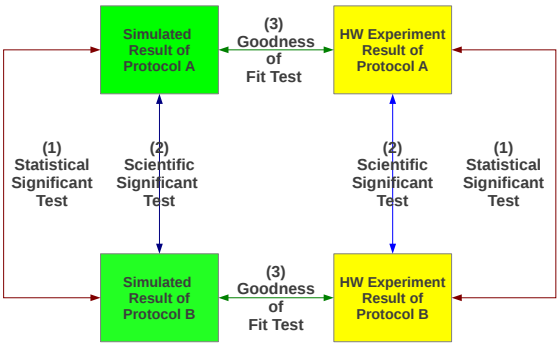


Fig. 1. Conceptual Statistical Test Framework

rank-sum test is the statistics generated from this test can be used to perform scientific significant test.

To perform a rank-sum test, a null hypothesis H_0 is defined. The H_0 states that the results have identical distributions; the alternative hypothesis H_a states that the distributions are different. Using a pre-determined confidence level of X%, H_0 can be rejected if the p-value $<= \alpha$, indicating that any observed difference in the results is unlikely to occur by chance, and our results are statistically significant. An α value of 0.05 is typically used, corresponding to 95% confidence levels [10].

2) Scientific Significance: It is possible for the observed performance improvement between the protocols to be statistically significant but underlying differences are small, i.e. unimportant, and only noticeable due to the large amounts of data obtained. It is also important to examine the scientific significance of results, to measure the difference or the effect size between the protocols. Another non-parametric test known as the Vargha-Delaney A-statistic is used to measure the effect size [11]. A-statistic in the range $[0, 1]$ is obtained using the parameters collected from the previous rank-sum test. Using the guidelines given by Vargha and Delaney in [11], A-statistic value of 0.5 shows no significant difference in effect size for the protocol performance. A-statistic < 0.29 and > 0.71 is required as it indicates a large effect size.

3) Goodness of Fit: The goodness of fit allows us to compare the relationship between the hardware and simulated results. It measures the discrepancy in the results and allows us to deduce the similarity of the simulation and hardware experiments. We apply the Kolmogorov-Smirnov (K-S test) test to determine whether two samples are drawn from identical distributions to measure the goodness of fit. The K-S test is a non-parametric test used to determine whether two samples are drawn from the same distribution, by quantifying the distance between the empirical distribution functions of two results. The null hypothesis H_0 for the K-S test states that the samples are drawn from the same distribution; the alternative hypothesis H_a states that the distributions are different. We reject H_0 if the p-value $<= 0.05$ at 95% confidence levels. Hence, H_a is true.

IV. Experimental Setup

In order to apply EVP to our experiments, we have implemented and evaluated the MRP, AODV and NST algorithms in both TelosB motes programmed using TinyOS and NS-2 simulator. TelosB with TinyOS 2.1.1 is selected for our experiments since NST-AODV was tested and evaluated using the same platform by Gomez et al. [7]. With the source available for download, it can easily be modified to support AODV and MRP. NS-2.34 is selected in our experiments since it has included the IEEE 802.15.4 module, developed by Zheng et al. [12]. Extension to support NST-AODV and MRP in NS-2 is also implemented and available to download[2]. In order to increase the confidence in experimental result, it is necessary to *repeat* the experiments.

A. TelosB Experimental Setup

We have set up a testbed network as shown in Figure 2 in a grid topology using six TelosB motes with the setting given in Table I. A small number of nodes are chosen so that greater control can be provided for the experiments. That is, the experiments are performed in the centre of a large room relatively free from uncontrolled radio sources, however a larger physical network would then be closer to uncontrolled noise sources. Node 1 in the network is configured to collect temperature reading from the sensor and transmit the packet to node 6, at regular intervals of 250ms via the intermediate nodes using multihop routing protocol. Each node is placed in a position of 0.5 metres from the other node so that it can only communicate with its immediate neighbours. The transmission power in each node is also set to a range of about 0.75 metres to avoid any interference to other non-neighbouring nodes. Radio channel 26 is used to avoid any interference with other Wi-Fi operating in the same band. An acknowledgement for each packet transmitted is also enabled for LLN operation.

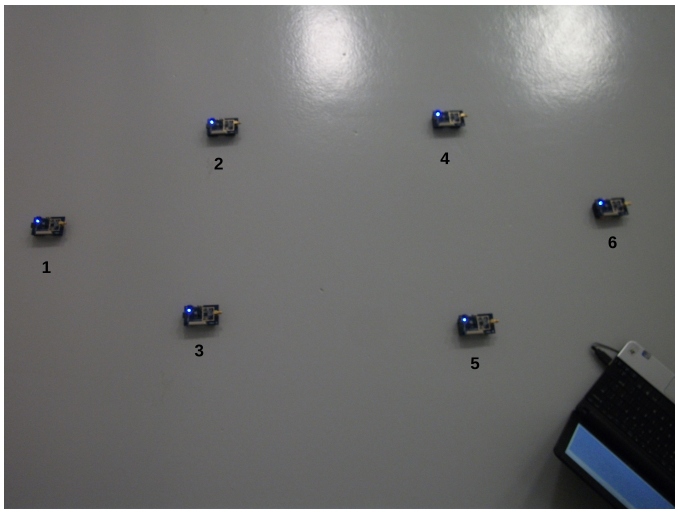


Fig. 2. TelosB network setup

[2]http://rtslab.wikispaces.com/file/view/mrp.tar

Each node also logs its network activities on the on-board flash memory, that are later retrieved for analysis. An extra mote, assigned as the controller, is used to synchronise the clock and collect the network statistics from the flash memory of the nodes. A simple time synchronisation algorithm based on flooding mechanism [13] is implemented in the controller. The controller will communicate with all the nodes on the network without using any multihop routing. A 5s initialisation period is allowed for synchronisation before the actual data communication begins. 10 repeated runs are conducted. Each run takes about 15 minutes to complete.

TABLE I
TINYOS CONFIGURATIONS

Parameters	**Values**
Tx interval:	250ms
Tx Channel:	26
MAC:	802.15.4 (CSMA/CA)
Route Protocol:	AODV, NST-AODV, MRP
Data Queue:	1 (AODV), 1 (NST-AODV), 6 (MRP)
Control Queue:	0 (AODV), 5 (NST-AODV), 0 (MRP)
RREQ Attempts:	1 (AODV), 3 (NST-AODV), 1-3 (MRP)

B. NS-2.34 Simulation Setup

In order to compare the testbed against simulation, extensive simulations were performed using Network Simulator (NS2), based on the same network, using the parameters shown in Table II. Previously, we have evaluated MRP and have achieved better performance on a larger network. For this simulation, we have designed a controlled experiment based on 6 static nodes placed in a 2 x 2 grid that mirrors the real world deployment in Figure 2. Node 0 is configured to transmit to node 5 at every 250ms. 35 repeated run was conducted and each run only takes 10 seconds to complete.

TABLE II
NS SIMULATION PARAMETERS

Parameters	**Values**
Tx interval:	250ms
MAC:	802.15.4 (CSMA/CA)
Routing Protocol:	AODV, NST-AODV, MRP
IFQ Size:	10 packets

C. Simulating Different Transient Failures

WSNs are susceptible to network disruption due to external interference. Depending on the nature of the interference, it can either cause a permanent or periodic failure that can be different in durations. To investigate how these failures can affect the behaviour of the nodes performance, failures are injected to the network by switching off the radio of an active node along the route at different intervals and durations. To ensure our results are *unbiased* toward a specific failure scenario, five different failure durations mainly: 0.25s, 1s, 2s, and 5s, with different intervals were used to represent different types of network activities that can interfere with the motes radio communication.

D. Evaluation Metrics

Choosing the right metrics is crucial to assess the performance accurately. The same metrics proposed in [5] are used as the input for the statistical tests to evaluate the significance of the protocols. These metrics were chosen, based on the work of the extensive models of Tate [14], [15], as they represent the reliability, performance and efficiency of the protocol.

- **Packet delivery ratio:** Packet delivery ratio (PDR) is used to measure the network reliability and is represented as the percentage of the number successful packet received to the total number of packet transmitted.
- **Average Delay:** Average delay measures the network performance and is calculated as the sum of the time required to send each packet over the total number of packets received. For better performance, a low average delay is required.
- **Normalised Routing Overhead:** Routing overhead is calculated as the normalised ratio of the sum of the total number of routing packets send to the total data packet received. It is used as an indicator to measure the amount of energy used to a data packet. A low value is desirable as it represents only a small amount of energy is wasted for communication during route discovery.

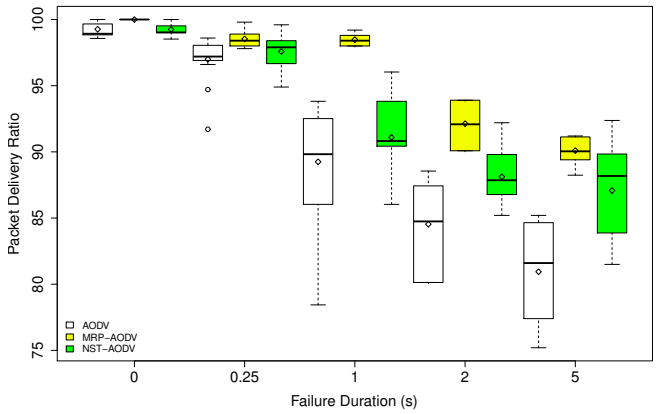
V. VERIFICATION OF EXPERIMENTAL RESULTS

In this section, the results obtained from both simulation and hardware experiments are presented and analysed using EVP. We make use of boxplots, as they present the median and IQR, thus are a presenting results in a non-parametric manner. The centre line in the boxplot represents the median while the bottom and top edges of the box show the first and third quartiles respectively. Outliers are presented by dots outside the whiskers. The mean value is also plotted in the boxplot, represented by a diamond point. In addition, we run statistical tests discussed in III-C to analyse and verify that the results obtained are significant and have scientific values.

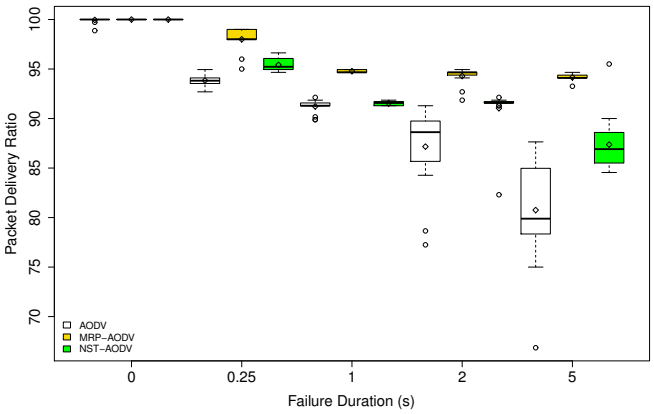
A. Packet Delivery Rate

In order to compare the reliability of each routing protocol, the median and mean value of the PDR is presented using boxplots in Figure 3a and 3b.

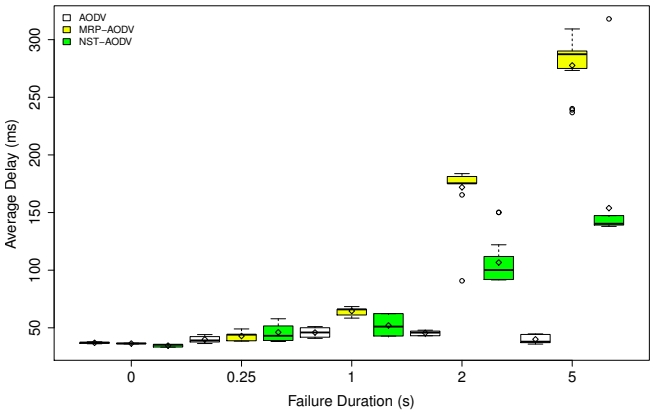
1) TelosB Experiment: In all failure scenarios, above 90% success rate has been achieved by MRP. This performance improvement is significantly different and higher than AODV and NST-AODV as the A-value given in column 4 of Table IV is greater than 0.93 and the p-value < 0.05. Before failures were injected into the network, about 3% of the packets sent were dropped in AODV and 1% in NST-AODV. These packets were dropped during route discovery, where a significant number of RREQ packets were observed. As we have conducted this experiment under a controlled environment in a small network, we believe these packet losses are due to random errors between the motes e.g. through collisions. When the



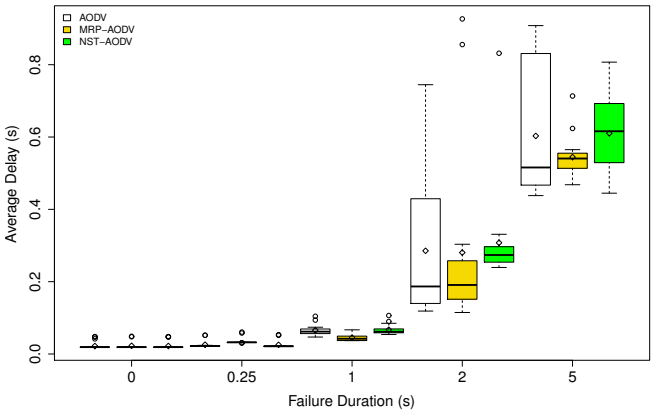
(a) TelosB PDR



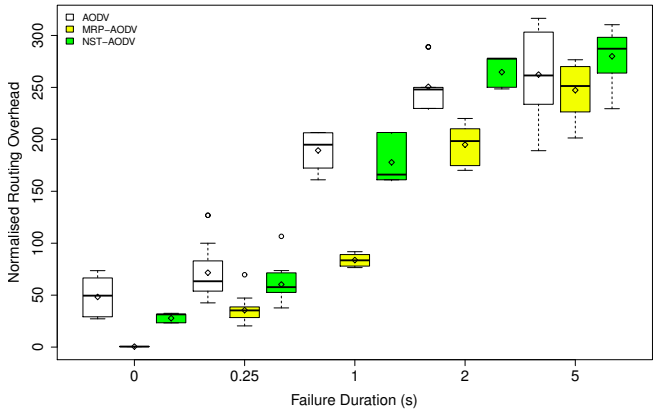
(b) NS-2 PDR



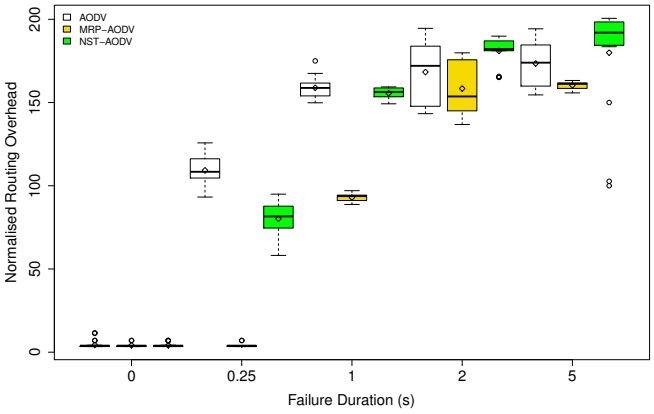
(c) TelosB Average Packet Delay



(d) NS-2 Average Packet Delay



(e) TelosB Normalised Routing Overhead



(f) NS-2 Normalised Routing Overhead

Fig. 3. Box-Whisker plot with Mean, Median and Inter-quartile range

radios of node 4 and 5 are turned off at regular interval for 0.25s, MRP managed to deliver around 98% packet compared to 96% for AODV. More routing packets were observed in AODV causing other node to drop their data packets due to collision. When we gradually increased the failure duration from 0.25s to 5s, MRP has managed to prevent less than 10%

packet loss compared to 15% in AODV, and 11% in NST-AODV as shown in Figure 3a.

2) NS-2 Simulation: Figure 3b shows the boxplot of PDR obtained NS-2 simulations. During normal operation, when no fault is injected to the networks, all the three protocols have achieved 100% PDR. When faults are injected into the node, by turning off an active node along the two possible paths, AODV starts to drop packets by 5% due to the unavailability of next hop neighbour. These numbers keep on increasing as we gradually increase the duration of failure in all the three routing protocols. With our RSM in MRP, we have maintained over 90% packet received in MRP compare to 80% in AODV and 87% in NST-AODV on average. In terms of performance improvement, this is over 10% in AODV and 5% in NST-AODV at 5s. *p*-values in Table III column 3 has demonstrated statistical significance differences between the routing algorithms. The performance of MRP is significantly more reliable than AODV and NST-AODV. A large effect size is also achieved for all failures.

3) Comparing TelosB against NS-2: We also tried to validate NS-2 simulator with the TelosB motes using Kolmogorov-Smirnov goodness of fit test, on the samples collected from hardware and simulation. As highlighted in Table V, a small test case has shown similarity in PDR with p-value > 0.05 supporting H_4. We believe this low number of similarity could be caused by random radio noise in hardware that is not modelled in NS-2. During error-free condition, a few packets have failed to reach the destination in our hardware experiments in AODV and NST-AODV that was not observed in NS-2. The real sensor motes are sensitive to communication failures that can be tolerated by MRP but not NST-AODV and AODV.

4) Discussion: By using EVP, we have achieved above 90% reliability at 95% confidence level with minimal number of test and unit time. The statistical analysis between MRP and NST, and MRP and AODV has shown that the performance improvement between the two algorithms is both scientifically and statistically significant with a large effect size and a small p-value. Hence, H_1 can be refuted. The result obtained from hardware is also better than NS-2 as MRP has maintained a higher packet delivery ratio during short errors. In terms of similarity between hardware and software, PDR can only provide partial evidence to accept H_4.

B. Routing Overhead

In WSNs, additional routing packets are required to establish a route to a destination when a link is unavailable. These additional packets are known to create additional overhead in the network and node. We have analysed the routing overhead between the three routing algorithms using boxplots in Figure 3e and 3f.

1) TelosB Experiment: When we increase the radio failure duration in the active node from 0.25 to 5s in the testbed experiments, the routing overhead increases in AODV. This overhead is less in MRP during small failure durations as shown in Figure 3e. The switching mechanism in MRP can abort RT and switch immediately to RT when the next hop neighbour is available for communication making it capable to operate more efficiently during transient random error. The MRP routing overhead gradually increases with failure duration until it is similar to AODV at 2s. When the cost of RT is higher than RD, MRP switches to RD after successive failures in RT. As for NST-AODV, significantly different routing overheads were observed after 2s with a p-value<0.01 due to additional RREQ packets generated during failure. These differences are scientifically significant as represented by the bold A-value in Table IV.

2) NS-2.34 Simulation: The simulated routing overhead is shown in Figure 3f. During normal condition, each successful packet received requires only 7 routing packets to be sent on average for all the routing protocols. This number increased linearly for NST-AODV and AODV with failure duration where they peaked at 2s. However, the overhead in MRP is less than AODV and NST-AODV as shown by the mean and median in Figure 3f. It is also statistically significant indicated by a low p-value in III. Hence, H_3 can be rejected.

3) Comparing TelosB against NS-2: By analysing the mean and median values in Figure 3f and 3e, each failure scenario has shown differences in routing overhead between the hardware and simulator. Based Kolmogorov-Smirnov Test in Table V, all the failure scenarios have shown a small p-values validating the differences between experimental and simulation results and rejecting H_4.

4) Discussion: From Figure 3e, the median and mean of routing overhead in MRP is smaller than AODV and NST-AODV. This difference is both statistically and scientifically significant as shown by the low p-value and high effect value of A-value. Hence, the results support H_3. However, there is no evidence from the KS-Test available to support H_4 as the overheads from the simulation are less than hardware.

C. Average Packet Delay

Figure 3c and 3d show that the average time delay increases with the failure duration for RD in all the three protocols.

1) TelosB Experiment: The delay observed in the testbed experiment in Figure 3c is smaller in AODV than MRP and is significantly different as shown by the small p-value in Table IV. There are two factors contributing to these observations. First, RD is the only delay in AODV, and this delay is minimal in this small network compared to waiting time incurred by MRP and NST during RT. Secondly, the packets that cannot be transmitted in AODV, were dropped and not accounted for in the delay calculation. The average delay observed in MRP is due to switching and buffering that occur during the fault period. If we normalise the average delay, by assigning a delay value to the dropped packets in all the protocols, the delay in MRP-AODV is less than AODV and NST-AODV in the testbed experiment as seen in Figure 4.

2) NS-2.34 Simulation: In simulation, the results show a different delay pattern was observed when the failure duration is increased in AODV and NST-AODV in Figure 3d. This increased delay pattern is due to the additional queueing of

Failure Duration	Routing Protocols	PDR		AVG DELAY		Routing Overhead	
		p-value	*A*-value	*p*-value	*A*-value	*p*-value	*A*-value
Normal	MRP/AODV	0.3564	0.5278	0.9210	0.5095	0.9210	0.5095
	MRP/NST	1	0.5000	0.9210	0.5095	0.9210	0.5095
	NST/AODV	0.1602	0.5278	0.9865	0.5015	0.9865	0.5015
0.25 sec	MRP/AODV	**1.0306e-06**	**1**	**8.1995e-05**	**0.9102**	**1.7680e-04**	**0.8906**
	MRP/NST	**2.2483e-05**	**0.9336**	**1.7680e-04**	**0.8906**	**1.7680e-04**	**0.8906**
	NST/AODV	**3.8289e-06**	**0.9766**	0.0734	0.3125	0.9249	0.5117
1 sec	MRP/AODV	**8.9372e-07**	**1**	**8.1995e-05**	**0.9102**	**3.6860e-04**	**0.1289**
	MRP/NST	**8.4342e-07**	**1**	**5.0878e-05**	**0.9219**	**1.1195e-04**	**0.9023**
	NST/AODV	0.1097	0.6602	0.9249	0.5117	0.8358	0.5234
2 sec	MRP/AODV	**1.3217e-06**	**1**	**1.5449e-06**	**1**	0.7203	0.5391
	MRP/NST	**1.5946e-06**	**0.9902**	**1.5449e-06**	**1**	**0.0136**	**0.7578**
	NST/AODV	**2.0841e-05**	**0.9375**	0.3365	0.6016	0.1809	*0.6406*
5 sec	MRP/AODV	**1.0306e-06**	**1**	0.5847	0.5586	**0.0014**	**0.8320**
	MRP/NST	**1.9077e-05**	**0.9375**	**0.0400**	**0.2852**	**5.0878e-05**	**0.0781**
	NST/AODV	**1.2815e-04**	**0.8984**	0.0935	*0.7148*	**3.2464e-06**	**0.9844**

TABLE III
p AND *A* VALUES FOR SIMULATION (BOLD HIGHLIGHTS SIGNIFICANCE VALUE)

Failure Duration	Routing Protocols	PDR		AVG DELAY		Routing Overhead	
		p-value	*A*-value	*p*-value	*A*-value	*p*-value	*A*-value
Normal	MRP/AODV	**2.6433e-05**	**0.9000**	**0.0121**	**0.2311**	**6.2648e-07**	**0**
	MRP/NST	**7.4129e-06**	**0.9333**	**2.6191e-06**	**1**	**4.2844e-07**	**0**
	NST/AODV	0.9332	0.5111	**2.6744e-06**	**0**	**0.0179**	**0.2489**
0.25 sec	MRP/AODV	**0.0038**	**0.8111**	0.0457	0.7156	0.2980	0.3867
	MRP/NST	0.0860	0.6844	0.3182	0.3911	0.1832	0.3556
	NST/AODV	0.4288	0.5867	**0.0375**	**0.7244**	0.8029	0.5289
1	MRP/AODV	**2.7362e-06**	**1**	**3.0894e-06**	**1**	**3.0035e-06**	**0**
	MRP/NST	**3.8641e-06**	**1**	**1.9659e-04**	**0.9048**	**4.0074e-06**	**0**
	NST/AODV1	0.3663	0.5958	**0.0155**	**0.7500**	0.4275	0.4167
2 sec	MRP/AODV	**2.5246e-06**	**1**	**2.7592e-06**	**1**	0.7670	0.5333
	MRP/NST	**3.3530e-05**	**0.9422**	**5.2013e-05**	**0.9333**	0.7671	0.5333
	NST/AODV	**0.0052**	**0.8000**	**3.0194e-06**	**1**	**0.0330**	**0.7289**
5 sec	MRP/AODV	**3.1529e-06**	**1**	**3.2829e-06**	**1**	0.5057	0.4267
	MRP/NST	**0.0303**	**0.7333**	**5.1811e-05**	**0.9333**	0.1217	0.3333
	NST/AODV	**4.6293e-04**	**0.8756**	**2.9130e-06**	**1**	0.9334	0.4889

TABLE IV
p AND *A* VALUES FOR HARDWARE EXPERIMENT (BOLD HIGHLIGHTS SIGNICANCE VALUE)

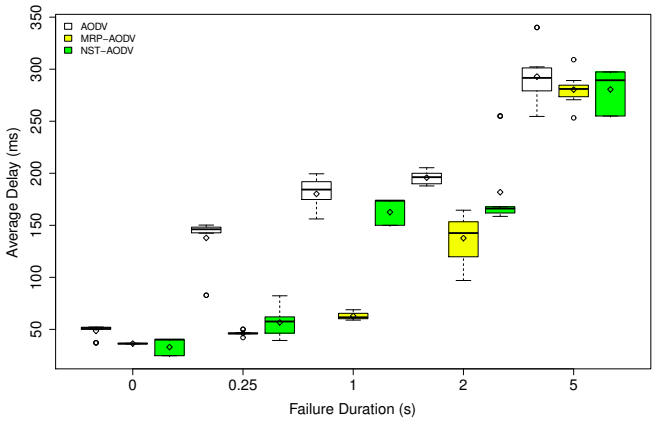


Fig. 4. Testbed Delay with Normalised Delay

the outgoing packet between link and mac layer in NS-2 simulation. During failure, the packets to be transmitted are placed in the queue. This queue size is set to 10 in simulation, while in TelosB, it is set to 1. When a 5s failure duration is injected, over 10 packets with delays between 2 - 9.55s, were found in the simulation during each failure duration in AODV. This packet delay is 2s when we change the failure duration to 2s. Hence, the increased delay in the average packet delay, that was not observed in the hardware experiment, is due to this queue. If we set the IFQ value to 1, no packet was received at the sink.

3) Comparing TelosB against NS-2: There is significant difference between the hardware and simulated results as shown by the small *p*-values (6×10^{-8}) in Table V due to the design and configuration of the NS-2 implementation. In NS-2, failed packets are usually buffered and transmitted when the network is up. In the simulation, if we set the data buffer similar to the hardware, and no packet can be delivered. Hence, the average packet delay metric was not able to accept H_4.

4) Discussion: From the analysis, the large *p*-value allow us to support H_2 when the failure duration is longer than 2s. However, we cannot accept H_4 due to a very small *p*-value. In order to bring the result closer, the NS-2 needs to be reprogrammed and finetuned to incorporate the interface module in TinyOS and radio characteristics observed.

Protocols	F_{rate}	PDR	DLY	RT
MRP	0	**1**	6.0708e-08	6.0708e-08
	0.25	**0.7925**	6.0708e-08	6.0708e-08
	1	1.0778e-07	1.0778e-07	1.0778e-07
	2	3.5093e-06	6.0708e-08	3.5093e-06
	5	6.0708e-08	6.0708e-08	3.5093e-06
NST	0	5.3056e-08	1.6377e-10	1.6377e-10
	0.25	1.8119e-04	6.0708e-08	6.0708e-08
	1	0.0039	6.0708e-08	6.0708e-08
	2	3.9803e-06	6.0708e-08	6.0708e-08
	5	**0.2395**	6.0708e-08	0.0023
AODV	0	5.1399e-06	1.6377e-10	1.6377e-10
	0.25	3.9803e-06	6.0708e-08	6.0708e-08
	1	0.0071	3.5455e-08	3.5455e-08
	2	0.0264	6.0708e-08	6.0708e-08
	5	**0.8467**	6.0708e-08	4.9386e-07

TABLE V
KS TEST p-VALUES (BOLD INDICATES TWO SAMPLES HAVE THE SAME DISTRIBUTION) BETWEEN NS2 AND TINYOS USING MATLAB

VI. ANALYSIS AND BENEFIT

Using a statistical approach to evaluate the results obtained from our experiments, we have shown that the performance improvement between the MRP, and AODV and NST, is both scientifically and statistically significant. The graphs in Figure 3 and small p-value (< 0.05) with a large A-value (> 0.73 *or* < 0.27) in Table IV and III for the PDR, average delay and routing overhead have refuted the hypotheses H_1 and H_3, where MRP has managed to deliver more packets at lower overhead. However, similar to observation made by [9], the KS test shows that the distributions of the results from the hardware and simulation are not the same. As a result, our experiments have refuted hypothesis H_4.

In terms of the benefit of our approach, we have analysed the time taken to perform the whole experiments in order to achieve the level of confidence required. To calculate the total time taken for our experiments, let us assumes that the time taken to run an experiment is equal to *t*, each experiment needs *r* random tests, and *s* scenarios need to be tested. Hence, total time to test one routing protocol $P = t \times r \times s$. If we need to compare *N* protocols, then the total time, (T_e), to test a set of experiment

$$T_e = N \times t \times r \times s \quad (1)$$

Therefore, in our simulation, the total time taken T_{sim}=5250 seconds where, N = 3, s = 5, t = 10 seconds, and r = 35, where in hardware, the actual time taken T_{actual}=202500 seconds, where N = 3, s = 5, t = 900 seconds, r = 15. This implies that our experiments only took approximately 11 days to complete based on 5 hours work per day. However, if the same number of runs performed in simulation is conducted in hardware, the estimated time will take $T_{predict}$=472500 seconds which is equivalent to 26.25 days. Hence, by using EVP, we have reduced the time taken by approximately half.

Our approaches can also be scaled outward as well as in parallel. It is more amenable to automation and the data gathering process in simulation is unobtrusive. We can also scale the network size and perform more tests in simulation without increasing the experimental time in hardware.

VII. CONCLUSION

In this paper, we have proposed the EVP to improve the credibility of the experiments to validate the routing protocol in WSNs. Using a structured statistical and experimental approach to analyse the data from repeated experiments with different test scenarios, it is viable to achieve a level of confidence in the performance results of routing protocol. By applying statistical hypothesis test on both experimental and simulated results, we have demonstrated that the performance improvement of MRP is both statistically and scientifically significant. We have shown, with sufficient confidence level, that the simulation is not a valid representation to the real hardware. This is due the simplistic statistical radio model used in simulation and the small sample size in hardware. Future works will investigate the effect of random sample size on the experiment and the application of real radio characteristics to the simulator to improve the goodness of fit.

REFERENCES

[1] I. Bate, Y. Wu, and J. Stankovic, "Developing safe and dependable sensornets," *Software Engineering and Advanced Applications, Euromicro Conference*, pp. 279–282, 2011.

[2] F. Zhao, "Challenge problems in sensornet research," Keynote at NSF NOSS PI meeting and Distinguished Lectures, Johns Hopkins and Princeton, 2005.

[3] Y. Wu, K. Kapitanova, J. Li, J. Stankovic, S. Son, and K. Whitehouse, "Run time assurance of application-level requirements in wireless sensor networks," in *Proceedings of the 9th ACM/IEEE International Conference on Information Processing in Sensor Networks*, ser. IPSN '10, 2010, pp. 197–208.

[4] S. Kurkowski, T. Camp, and M. Colagrosso, "MANET simulation studies: the incredibles," *SACM SIGMOBILE Mobile Computing and Communications Review*, vol. 9, no. 4, pp. 50–61, 2005.

[5] T. Lim, I. Bate, and J. Timmis, "Multi-modal routing to tolerate failures," in *Proc. of the 7th International Conference on Intelligent Sensors, Sensor Networks and Information Processing*, 2011, pp. 211–216.

[6] C. E. Perkins and E. M. Royer, "Ad-hoc on-demand distance vector routing," in *Proc. IEEE Workshop on Mobile Computing Systems and Applications*, 1999, pp. 90 –100.

[7] C. Gomez, P. Salvatella, O. Alonso, and J. Paradells, "Adapting AODV for IEEE 802.15.4 mesh sensor networks: Theoretical discussion and performance evaluation in a real environment," in *Proc. IEEE International Symposium on World of Wireless, Mobile and Multimedia Networks*, 2006, pp. 159–170.

[8] S. Ivanovn, A. Herms, and G. Lukas, "Experimental validation of the ns-2 wireless model using simulation, emulation, and real network," *ITG-GI Conference on Communication in Distributed Systems*, pp. 1 –12, 2007.

[9] N. Pham, D. Pediaditakis, and A. Boulis, "From simulation to real deployments in wsn and back," in *IEEE International Symposium on a World of Wireless, Mobile and Multimedia Networks*, 2007, pp. 1 –6.

[10] F. Wilcoxon, "Individual comparisons by ranking methods," *Biometrics Bulletin*, no. 6, pp. 80–83, 1945.

[11] A. Vargha and D. Delaney, "A critique and improvement of the CL common language effect size statistics of McGraw and Wong," *Journal of Educational and Behavioral Statistics*, vol. 25, no. 2, pp. 101–132, 2000.

[12] J. Zheng and M. Lee, *A comprehensive performance study of IEEE 802.15.4*, 2006.

[13] M. Miklós, K. Branislav, S. Gyula, and L. Ákos, "The flooding time synchronization protocol," in *Proc. of the 2nd international conference on Embedded networked sensor systems*, 2004, pp. 39–49.

[14] J. Tate, B. Woolford-Lim, I. Bate, and X. Yao, "Comparing design of experiments and evolutionary approaches to multi-objective optimisation of sensornet protocols," in *Proceedings of the 10th IEEE Congress on Evolutionary Computation*, May 2009, pp. 1137–1144.

[15] J. Tate and I. Bate, "Sensornet protocol tuning using principled engineering methods," *The Computer Journal*, vol. 53, pp. 991–1019, 2010.