

國立臺灣大學電機資訊學院資訊工程學系  
碩士論文

Department of Computer Science and Information Engineering  
College of Electrical Engineering and Computer Science  
National Taiwan University  
Master Thesis

智能M2M中介軟體的容錯系統研究  
Research on Fault Tolerant M2M Systems

蘇適

Penn H. Su

指導教授：許永真 博士

Advisor: Jane Yung-Jen Hsu, Ph.D.

中華民國 102 年 3 月  
March, 2013

國立臺灣大學  
資訊工程學系

碩士論文

智能NN中介軟體的容錯系統研究

蘇適撰



國立臺灣大學碩士學位論文  
口試委員會審定書

智能M2M中介軟體的容錯系統研究  
Research on Fault Tolerant M2M Systems

本論文係蘇適君 (R99922157) 在國立臺灣大學資訊工程學研究所完成之碩士學位論文，於民國 102 年 3 月 14 日承下列考試委員審查通過及口試及格，特此證明

口試委員：

---

---

---

---

---

---

---

所主任



# Acknowledgments

I would like to express my greatest gratitude to the people who have helped and supported me throughout my graduate studies. I am grateful to my advisor Prof. Jane Yung-Jen Hsu for her continuous support, from initial advice, contacts in the early stages of conceptual inception, through ongoing advice and encouragement to this day. I am also grateful to my advisor Prof. Kwei-Jay Lin for his undivided attention to details, ongoing advice and encouragement.

To Dr. Yu-Chung Wang and Prof. Chi-Sheng Shih, for their patience and their kindness for answering my many requests.

To Niels Reijers, my colleague who helped me in completing the project with his vast knowledge, black VM magic tricks, ideas, and thoughts that made this journey a lot smoother.

To my colleagues and my friends, thanks for the support and fun. I want to thank iAgent group, George, Jya-Cheng, Joey, Lecheng, Farmer, Bo-Lung, and Sio, who have directly or indirectly help out this effort.

I wish to thank my parents for their undivided support and interest who inspired me and encouraged me to go my own way to this day, without whom I would be unable to complete my graduate studies.





## 中文摘要

對於具有 ”部署一次，永遠運行” 概念的物聯網而言，容錯移轉是這類分散式服務導向網路的必備條件。當設備更換時或系統出狀況時，必須利用資源再規畫去達成容錯移轉的機制。系統在運作時，異質性的或多工性的設備之間若不僅是端對端的訊息傳輸時，不管是設備或是訊息的複製都是昂貴且累贅。特別是當設備某種服務故障時，可由另外一個有能力提供相同服務的同級設備接替其服務，而不一定要由相同設備取代。利用長帶來記錄一連串複製的服務訊息，每一個同級設備都保存一致的長帶記錄。結合常用於失敗偵測的心跳協定，系統由異常回復的機制可以藉由操控分析分散狀態的長帶來達成，而不須利用到中央資料庫。使用Arduino mega 2560相容設備所做的實驗結果顯示，我們已經能夠使小型網路系統故障復原，較大的網路實驗則正在進行中。未來研究方向包括確認網路的可擴展性，網路磁碟分割處理以及解決同步故障的問題。



# Abstract

Failover for service-oriented distributed networks is a prerequisite to enabling Internet-of-Things (IoT) in the sense of “deploy-once, run forever.” Resource reconfiguration is required to achieve failover mechanisms upon replacement of devices or failure of services. It can be particularly challenging when services in applications have more than end-to-end transmissions between devices that are heterogeneous or versatile, for which duplications can be costly and redundant. Specifically, a device with a failed service shall be taken over by another service peer, instead of a device counterpart to recover application as a whole. Strip is introduced to store a list of duplicated services, and, each service peer maintains a consistent view of strips. In combination with the heartbeat protocol which was widely applied for failure detection, recovery from failure can be achieved by manipulating strips in a distributed manner, without a centralized repository. Experiments using Arduino Mega 2560 compatible devices show that our approach is capable of failover in small networks, whereas experiments in larger networks are underway. Future research directions include addressing the scalability issue, network partitions and tackling simultaneous failures.



# Contents

口試委員會審定書	i
Acknowledgments	iii
中文摘要	v
Abstract	vii
<b>1 INTRODUCTION</b>	<b>1</b>
1.1 Goals . . . . .	2
1.2 Challenges . . . . .	3
1.3 Approaches . . . . .	3
1.4 Related Work . . . . .	4
1.5 Thesis Organization . . . . .	5
<b>2 BACKGROUND</b>	<b>7</b>
2.1 Internet-of-Things . . . . .	7
2.2 Machine-to-Machine . . . . .	7
2.2.1 Wireless Sensor Networks . . . . .	8
2.3 Redundancy architecture . . . . .	8
2.4 WuKong: The intelligent middleware for M2M applications . . . . .	9
2.4.1 Goal . . . . .	9
2.4.2 Flow Based Programming . . . . .	9
2.4.3 Sensor Profile Framework . . . . .	11
2.4.4 Compilation and Mapping . . . . .	13
2.4.5 System Progression Framework . . . . .	14
<b>3 SYSTEM DESIGN</b>	<b>15</b>
3.1 User Preference for Fault Tolerance . . . . .	15
3.2 Deploying Application with Fault Tolerance . . . . .	17
3.3 Strip . . . . .	18

3.4	Reconfigurable Redundancy Architecture . . . . .	20
3.4.1	Decentralized Failure Detection . . . . .	20
3.4.2	Failure Recovery . . . . .	21
<b>4</b>	<b>EVALUATION &amp; RESULTS</b>	<b>29</b>
4.1	Application . . . . .	30
4.2	Policy . . . . .	30
4.3	Heartbeat Protocol Arrangement . . . . .	31
4.4	Hardware Platform . . . . .	33
4.5	Experimental Setup . . . . .	33
4.6	Mapping results . . . . .	34
4.7	Results . . . . .	34
<b>5</b>	<b>CONCLUSION</b>	<b>41</b>
5.1	Discussion . . . . .	41
5.2	Future Work . . . . .	42
	<b>Bibliography</b>	<b>45</b>

# List of Figures

2.1	A FBP application . . . . .	10
2.2	WuKong application build flow . . . . .	13
3.1	An example of categories a user policy could impose on a component . .	15
3.2	An example of fault tolerance policy . . . . .	16
3.3	An example network with several strips . . . . .	19
3.4	A failure occurred at node 2 in the network . . . . .	22
3.5	Reconfigure application links . . . . .	25
3.6	Reconfigure heartbeat protocols . . . . .	26
3.7	Update monitored information . . . . .	27
4.1	Heartbeat Protocol Arrangement . . . . .	32
4.2	An WuDevice . . . . .	38
4.3	Average recovery time and message overhead over 5 deployments for each node failure as the first failure . . . . .	39





# List of Tables

4.1	Node setup . . . . .	34
4.2	Strips . . . . .	34
4.3	Strip memory overhead in bytes . . . . .	35



# Chapter 1

## INTRODUCTION

Failover mechanism is essential in any distributed networks. It is especially crucial for enabling Internet-of-Things (IoT) in the sense that applications could "deploy-once, run forever." Where applications degrade gracefully under failures.

Internet-of-Things is being realized in Machine-to-Machine (M2M) systems such as wireless sensor networks (WSN). WSNs are areas filled with network of tiny, resource limited sensors communicating wirelessly. Each sensor is capable of sensing the environment in its proximity. Wireless sensor networks are typically employed in a variety of applications ranging from home automation to military.

Sensor networks offer the ability to monitor real-world phenomena in detail and at large scale by embedding devices into the environment. Deployment is about setting up a sensor network in a real-world environment. Deployment is a labor-intensive and cumbersome task since environmental influences or loose program logic in code might trigger bugs or sensor failures that degrade performance in any way that has not been observed during pre-deployment testing in the lab.

The real world has strong influences on the function of a sensor network that could change the quality of wireless communication links, and by putting extreme physical strains on sensor nodes. Laboratory testbed or simulator can only model to a very limited

extent of those influences.

There have been several reports on sensor network installations where they encountered problems during their deployment[3, 11, 1, 16, 9, 14, 17, 18].

Testbed in laboratory environment can still not model the full extents of the influences a real world environment could do. Deployment still a big problem in wireless sensor network applications.

However, Wireless Sensor Network (WSN) programming can only be performed by experts, because of resource constraints and their high failure probability. Hence, transforming the paradigm of service-oriented architecture (SOA) to sensor networks becomes an important research in the last few years. [7, 6]

WuKong, a middleware for service-oriented M2M systems, has set out to allow developers develop M2M applications with ease by introducing users to a new programming paradigm called flow based programming (FBP) with automatic sensor identifications, node configuration, and system re-configuration. [13]

In this thesis, we present our fault tolerance system for WuKong.

## **1.1 Goals**

Users of WuKong deploy applications written in the form of flow based programming. Applications consist of a set of components linked together. Each component represents a service in network. When application is deployed, each component is mapped to a service in the network. However, partial failures could bring the whole system down, since each service depends on each other to function. However, as users have drastically different requirements even for the same applications, we don't force the system to do in our way, we allow user preference to influence our system via user policy.

The goal of this system is to provide failover for application components such that it meets user's requirements through user policy. Thus the system also needs to design a generally accessible user policy for fault tolerance such that it is easy and intuitive to specify how much fault tolerance a component should have. It is also important that the deployed system would be able to detect, recover from failures, and reconfigure the system autonomously.

## **1.2 Challenges**

The way WuKong presents and models resources and applications, along with user policy and requirements, makes it a unique challenge: User requirements might vary drastically from one to another. System has to embrace a variety of top-level and low-level configurations so components would play along with each other and eventually produce a working failover architecture for the network. Encompassing a variety of devices with different form factors, memory storage, radio bandwidth is also a distinct characteristics in IoT, therefore the system and algorithmic design has to work among nodes with different specifications. In addition, duplications can be costly and redundant when devices are heterogeneous or versatile, such that instead of replaced by a device counterpart, it would be taken over by another local service peers.

## **1.3 Approaches**

Previous work on the problems has not considered failover architecture for multiple services where relationship among services are critical and which would also be influenced by user policies pre-deployment. Furthermore, after every failure recovery, none explained

how existing services could find replaced services in such environment. This thesis proposed a novel algorithm to achieve recovery from failures by combining heartbeat protocol, for failure detection, with Strips, which are used to maintain and track service redundancy.

## **1.4 Related Work**

In WSNs, failover techniques for data replications, and service redundancy are typically adapt for applications only with predefined services, fixed user requirements and services with only end-to-end links: read/write to local data storage. One example of such work with local storage for replications is described in [12]. In this work, some distinguished storage nodes are specified by Hash functions to collect data of certain types. Redundancy is achieved by storing replicas directly on neighbors nodes. But applications are still restricted to end-to-end links, since there is only one type of links in the application.

In [10], a dynamic replication approach for local data storage where replicas are randomly distributed within a predefined replication range influenced by the specific replica number and/or its density is presented. Another dynamic replication approach for local data storage where replicas are selected based on the scoring system defined by several physical resource traits from self-inspection that could reactivate services based on already collected and generated data [8]. However, this approach does not consider applications with dynamic user requirements, services with more complex links.

## **1.5 Thesis Organization**

Our work overlaps many diverse but interconnected domains, each topic being itself a subject of advanced research and abundant literature. Chapter 1 gives an introduction to the system goals and challenges and outline the approach used to solve the problem. Chapter 2 gives a brief background overview of the topic that this work based on. We start by describing IoT, M2M networks and WSNs. The chapter ends with an overview of WuKong. Chapter 3 describes our system design for the WuKong fault tolerance system. In this chapter we give detail description of our method and algorithms, including user policy for fault tolerance, strips, and reconfigurable redundancy architecture. Chapter 4 presents the metrics we would use to measure system performance, which is followed by evaluation of the results. Finally, chapter 5 presents some conclusions of the work, list of contributions and future work.





# **Chapter 2**

## **BACKGROUND**

### **2.1 Internet-of-Things**

Internet-of-Things is a powerful paradigm in the scenario of modern wireless telecommunications. It refers to the pervasive presence of uniquely identifiable objects (things) around us - such as Radio Frequency IDentification (RFID) tags, sensors, actuators, mobile phones, etc, which would be able to interact with each other forming a network like structure. [2]

### **2.2 Machine-to-Machine**

Machine-to-Machine refers to technology that enables both wireless and wired systems to communicate with other devices with the same ability. Thus a device (equipped with sensors) could capture an event (such as photodetector, temperature, etc), which is relayed through a network (wired, wireless or hybrid) to an application which will translate that into a meaningful information to perform action upon (e.g. turn on actuators).

### **2.2.1 Wireless Sensor Networks**

Wireless sensor networks are a network of sensor nodes. Sensor nodes are equipped with low-power, low-cost, and failure-prone sensors or actuators. Sensor networks are networks of sensor nodes that connect to the physical space that are instrumented to produce data that could be meaningful for further research. They collaborate to collect, process and disseminate environmental information[4].

Sensor network could be homogeneous, meaning all nodes are identical with same sensors, actuators and hardware setup. Sensor networks could also be heterogeneous where nodes have different sensors, actuators and hardware setup. Heterogeneous networks require higher level management and organization resources. Wireless sensor networks are nodes that communicate through air by sending electronic signals. Wireless communications aren't stable, as it is highly influenced by environmental factors.

## **2.3 Redundancy architecture**

Sensor networks are usually deployed in large scale and unattended in long period of time. Sensor networks communicate with low-power wireless radios to aid scientists in collecting spatial data that could lead to more understanding of the environment. However, several challenges such as node failures, message loss, and sensor calibration leaves the effectiveness of sensor networks in question. With the assumption of spare homogeneous resources, redundancy is used in sensor networks to increase fault tolerance against node failures. The system is designed with backup nodes that could automatically recover and replace should one node fail.

## **2.4 WuKong: The intelligent middleware for M2M applications**

### **2.4.1 Goal**

Deployment and development for M2M applications are in its infancy today. As many applications are still single purpose in homogeneous networks with specific network protocols. The hardware has a fixed range of sensors, and the applications cannot be easily ported to other platforms.

The existing middleware support that decouple high-level application design abstractions and low-level hardware has not been successful.

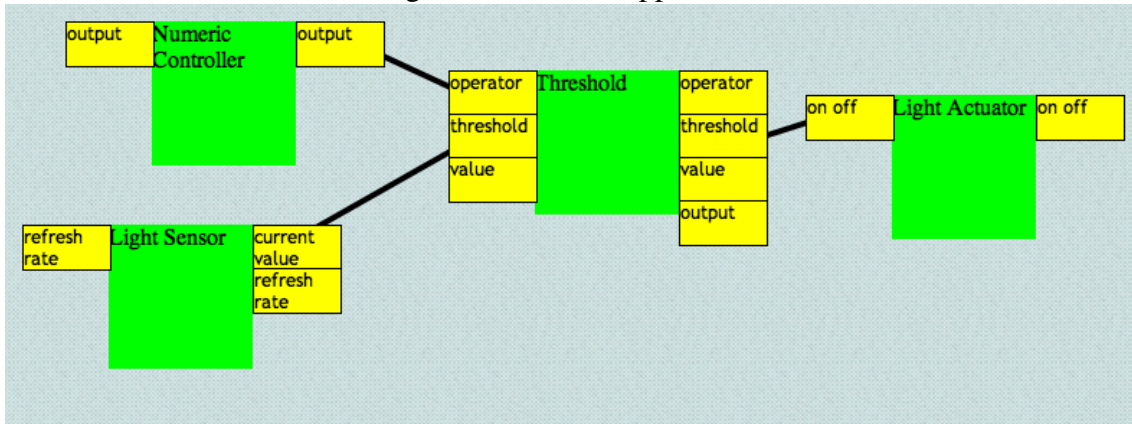
In Intel-NTU Center Special Interests Group for Context Analysis and Management (SIGCAM), we have been collaborating on a project, called WuKong, aiming to develop an intelligent middleware for developing, and deploying machine-to-machine (M2M) applications with ease. The main contribution of this project is to support intelligent mapping from a high-level flow based program (FBP) to self-identified, context-specific sensors in a target environment[13].

### **2.4.2 Flow Based Programming**

M2M applications are by definition distributed where the application requirements involve a network of nodes collaborating for some common goals. M2M applications are typically defined by its flow of information between components, as opposed to more traditional applications that focus more on local information processing.

Flow Base Programming is best suited for describing M2M applications as it allows the developers for the applications to focus more on the abstraction meaning of the components instead letting the unimportant details such as the hardware to stick right in the face. The

Figure 2.1: A FBP application



result application will contain all necessary information for the framework to construct low-level details to implement the flow.

Applications are designed and constructed on FBP canvas by dragging a set of abstract components from the library as illustrated in Figure 2.1 Each component is illustrated by a green block, each block has a set of properties, each with different access modes, such as readonly, writeonly, readwrite. Properties on the left of the greenblocks are properties that could be written, and properties on the right are readable. Components are connected by links, which is drawn by linking two properties in different components.

Some components represent physical hardware such as a sensor, or an actuator while some other components could represent virtual processes such as mathematical computations, comparisons, etc. However, the final physical implementations of the components are only made during application deployment by the Master but not during FBP construction.

Components expose their interface through properties. A link is only made with properties with matching data type. The FBP application in Figure 2.1 illustrates a simple scenario where the light actuator will turn on the light if light level drops below some value. The Numeric Controller component will be assigned to a user input device

used by users to set its desire light threshold, which its output is sent to Threshold component. The light value is sensed from Light Sensor component and sent to Threshold. If the light value sensed is below the threshold value, Threshold will output a boolean to set the on off property of Light Actuator to turn the device, which will be determined during deployment, that it is represented by on or off.

### **2.4.3 Sensor Profile Framework**

While FBP defines the logical view of an application, WuKong profile framework allows tracking, identification of physical resources within the Sensor Network. There are a range of sensors which provide similar functionality with different level of quality, it could model the sensor capability to enable handling heterogeneous sensors and provide a common abstraction for the logical view.

There are two main concepts in Sensor Profile Framework, WuClasses and WuObjects. WuClasss model components by exposing a number of properties describing, and allow access to, a specific resource represented by the class. Drawing from the example in Figure 2.1, the on off property of Light Actuator component is boolean writeonly. WuClass also implements an update function to describe a component's behavior. For example, Threshold has four properties: operator, threshold, value, output. The output value is determined from the previous 3 properties that it returns true when the value is lower or higher than the threshold which depends on the value of the operator, and it returns false otherwise.

WuObjects are the main unit of processing that are hosted on the nodes. Each WuObject is an instance of WuClasses. It allows the framework to achieve 4 responsibilities:

1. Allow the Master to discover the current status of a node with the list of WuClasses

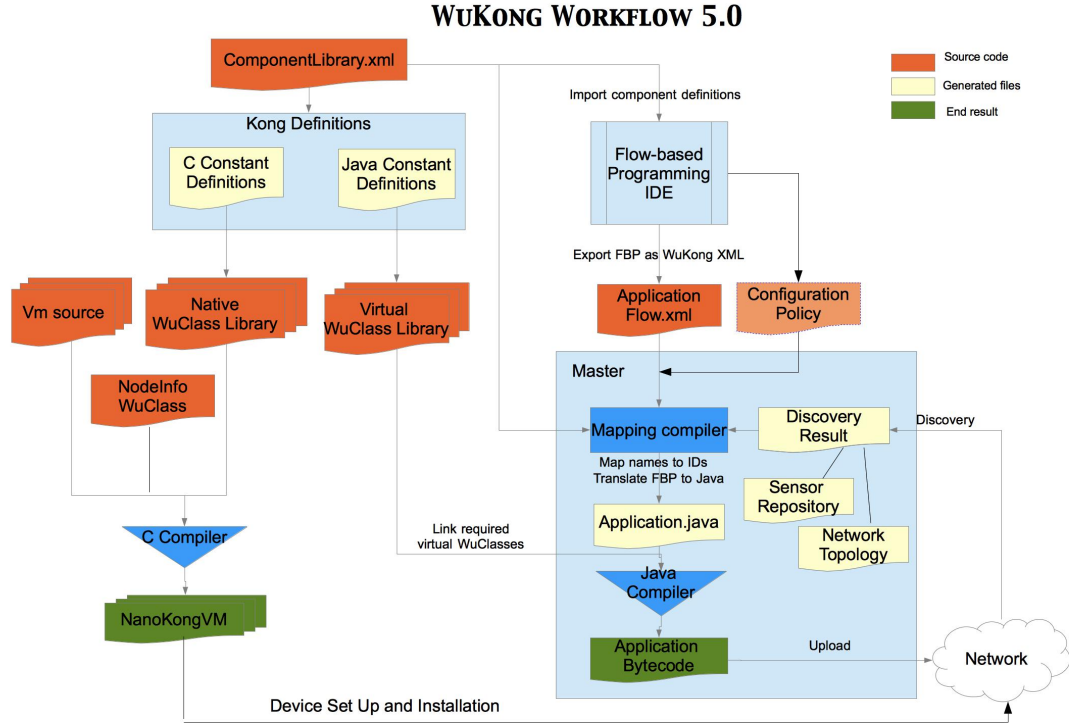
and WuObjects it has.

2. Create new WuObject instances on a node to start receiving data and doing local data processing.
3. Trigger executions in WuObjects, either periodically or as a result of changing inputs.
4. Propagate changes of properties between linked properties in different components, which may be hosted locally or remotely.

### **Property Propagation**

The profile framework is in charge of communication between WuObjects as well, which are not necessarily on the same nodes. Profile Framework monitors the changes in properties and propagate the changes to the connected WuObjects. For example, if a Temperature WuObject is connected to a Threshold WuObject, the changes in Temperature current value property will trigger propagation from the Profile Framework to propagate the new value in current value to the Threshold WuObject connected property, and since Threshold WuObject could be on a different node, the framework will take care of this by initiating a wireless connection between the nodes to send the data over. Once a new value has been set, Threshold WuObject will also trigger its update() function to recompute its output properties which in turn would cause another chain of propagation to the linked WuObjects.

Figure 2.2: WuKong application build flow



#### 2.4.4 Compilation and Mapping

Figure 2.2 shows the overview of WuKong's build flow. The left part represents the build process for NanoKong VM which will be installed on the sensor nodes as part of the WuKong framework. The top part represents the build process for generating component libraries and Virtual WuClass library which will be used in other parts of the process. The right part illustrates the build process for FBP applications from being drawn in the IDE to Java bytecode that will be transmitted to the nodes.

The FBP program from the IDE will be exported as XML to the Master, the Master will then take this XML and passed to Mapper to generate a Java program that will be

executed on the nodes. Finally, the compiled code is then wirelessly uploaded to the nodes in the network.

The Java code consists of many parts from different phases of the mapping process. First, the Java code contains information about links between components that were taken from the FBP XML passed in earlier from the IDE. A link contains the source component id, destination component id. The library code for components corresponding to the component ids are stored in the node if it is written in native language, or uploaded as part of the Java bytecode if it is written in Java language. Second, it contains information about the mapping from application component ids to actual node identifications and positions. The purpose of a mapping which separates from the actual link makes it easier to substitute the actual host of the WuObject later during reconfiguration from the Master. This mapping is created by the Master during discovery phase that probe the network for node's capabilities in terms of available WuClasses, then mapper will decide the final candidates that will be hosting for a component. If no native version of a component is found on the nodes, mapper will substitute with a Java version of it.

### **2.4.5 System Progression Framework**

There are a few popular wireless communication protocols in M2M applications: ZigBee, ZWave. It is expected that in the future more diverse wireless nodes equipped with radios that support protocols such as low-power blueooth and WiFi that all have one or more powerful gateway to connect to the outside world. In WuKong system, one of the gateways will take on the role of higher management decision maker called *Master* to making the decisions for deployment and producing the configuration of wireless sensor networks.



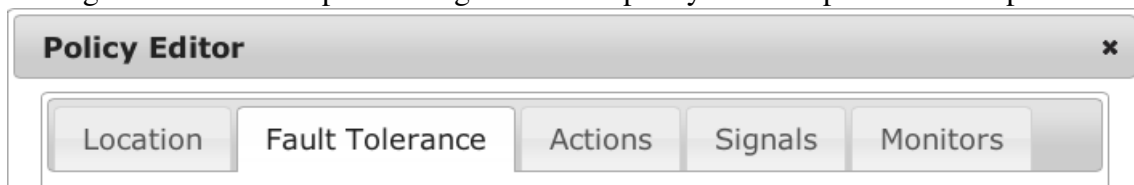
# Chapter 3

## SYSTEM DESIGN

The design for WuKong fault tolerance system is guided by the goals and challenges outlined in chapter 1: We want fault tolerance policy to be expressive, component specific, yet decoupled from physical hardware specifications so it will work with a range of network configurations. Furthermore, we want a reconfigurable redundancy architecture for service-oriented heterogeneous WSNs so components would be resilient to partial failures and degrade gracefully as indicated in the policy while being efficient and simple in terms of engineering design.

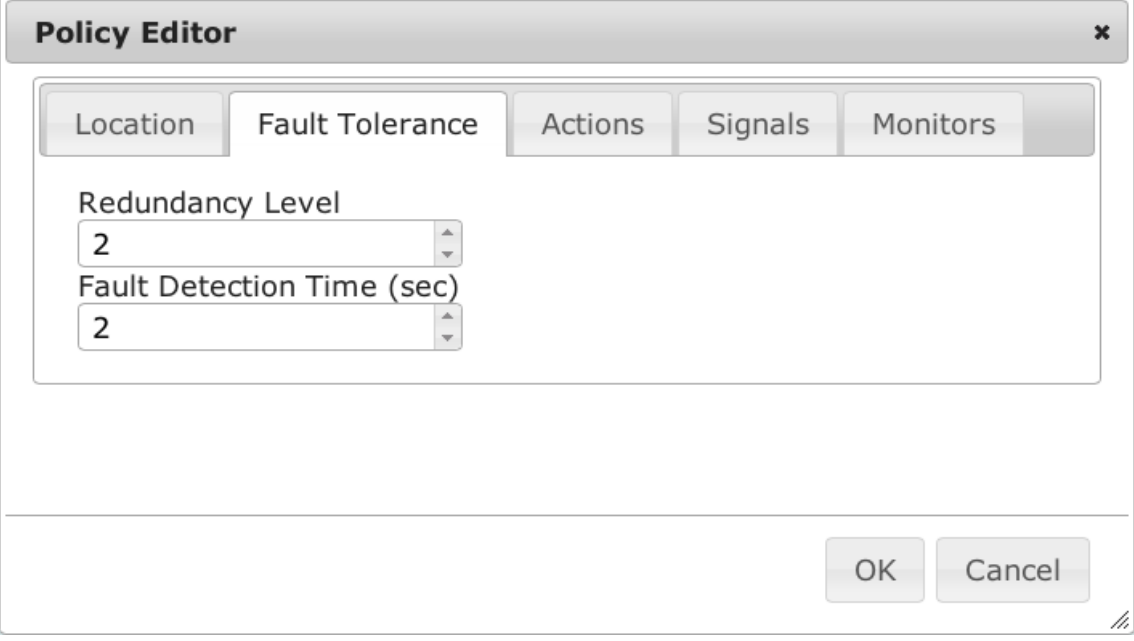
### 3.1 User Preference for Fault Tolerance

Figure 3.1: An example of categories a user policy could impose on a component



User policy guides the decisions and achieves certain outcomes for the system. Specifically, user policy controls how the system could behave in high level concept. When users

Figure 3.2: An example of fault tolerance policy



The image shows a 'Policy Editor' dialog box with a close button (X) in the top right corner. It contains five tabs: 'Location', 'Fault Tolerance', 'Actions', 'Signals', and 'Monitors'. The 'Fault Tolerance' tab is selected. Inside this tab, there are two settings: 'Redundancy Level' with a value of '2' and 'Fault Detection Time (sec)' with a value of '2'. Both values are in input fields with up and down arrows. At the bottom right of the dialog are 'OK' and 'Cancel' buttons, and a small icon in the bottom right corner.

starts up WuKong user interface, they will be confronted with the application where users could specify policy for each application component on, for instance, specific location the application component should be mapped to, or number of redundant devices the application component would have as illustrated in figure 3.1.

Users are also capable of specifying fault tolerance requirements via the user policy for fault tolerance as shown in figure 3.2 Redundancy level indicates the number of devices that will be able to take over when the service failed. Fault Detection Time represents the time the system should take at most to detect an failure in seconds.

There are timeouts set accordingly to prevent deviation in wireless communication quality, internal crystal clock to reduce the chances of getting false positives. Right now there is no way to find the correct timeout for the network, so we set it to right around 2x times of the heartbeat period we set fo the network.

## 3.2 Deploying Application with Fault Tolerance

In WuKong, deployment consists of discovering available resources and network topology, converting application from high-level abstractions to low-level machine instructions, then determine the parameters based on user policy, finally combine all together and deploy to the network. The process was briefly described in section 2, however, the thesis has made some significant changes to the process to support strip, an redundancy abstraction used in failure recovery and reconfiguration algorithms. First, we add a new process in mapper to take the routing topology results in discovered network info, so we could know the neighbors for each device. Next we take the existing mapping process, and convert it to outputting strips. So instead of one-to-one mapping from components to devices, now each component maps to a strip, which could be seen as a list of devices supporting such service ordered by the ordering function used. The default ordering function uses first-fit in a first come first process fashion, thus the ordering of the strip is randomized, as it depends on the order of processing.

An first-fit algorithm for mapping is shown in algorithm 3.2.1

**Algorithm 3.2.1:** FIRSTFITCOMPONENTTOSTRIPMAPPING( $S, F$ )

$C \leftarrow S$

$X \leftarrow F$

**for each**  $c \in C, x \in X$

**do**  $\left\{ \begin{array}{l} \text{if } c \in x \\ \text{then } \left\{ \begin{array}{l} c << x \\ \text{comment: Append } x \text{ to } c \end{array} \right. \end{array} \right.$

**return** ( $C$ )

S is the set of components, F is the set of sets each represents a network node with service capabilities. This algorithm will produce a mapping of C, a set of lists in first-come-first-serve order.

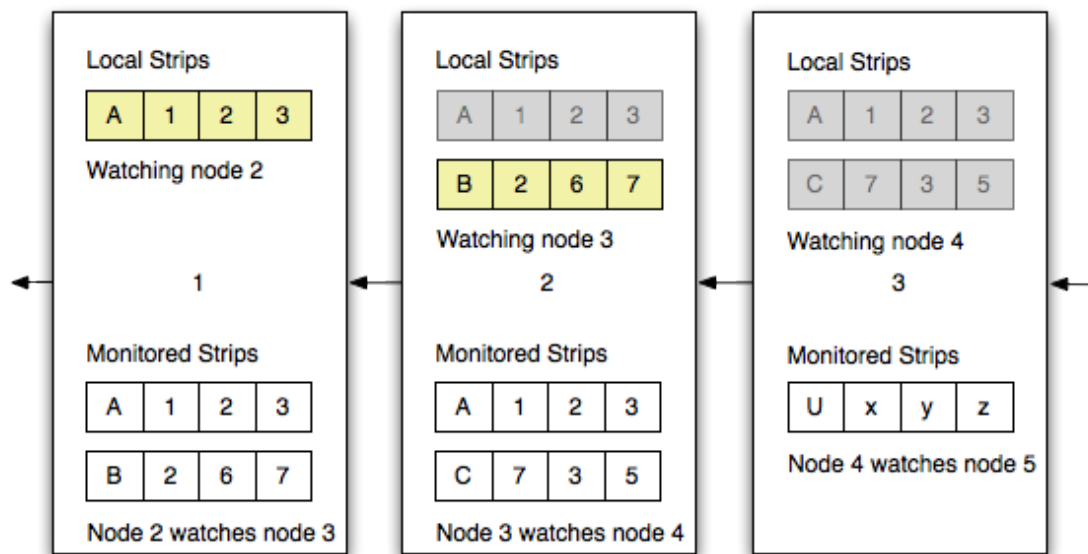
Once the mapping is produced, each device would be reprogrammed to create corresponding WuObjects to host the services.

### 3.3 Strip

Representing a component in WuKong application, each Strip contains a list of node on which the WuObjects representing the component are hosted. As seen from the figure 3.3, each node represented by a big white block contains a copy of strips for components that are specified to have redundancy in fault tolerance policy. Nodes holding a duplicated WuObject are members of the strip. The membership of a strip is called the view. Only the WuObject hosting at the head of the strip will be active, while the rest are backups. As seen from figure 3.3, node 1 has a active WuObject for component A as shown highlighted in yellow, but duplicated WuObjects A in node 2 and 3 are inactive in grey. In Strips, when one member failed, the next one will take over. For instance, if a strip is constructed like this  $\rightarrow 1 - 2 - 3 - 4 - 5$ , when 3 failed 4 will take over the place of 3, and the new chain will look like this:  $\rightarrow 1 - 2 - 4 - 5$ . Now if 1 failed, 2 will take over which would result in  $\rightarrow 2 - 4 - 5$ , since node 2 is now at the head of the strip, its WuObject will be active.

Typically, there would be multiple components deployed to the network and each node could carry multiple WuObjects, therefore many of these strips in the network would crisscross with one and others. It is not unlikely to see a node carrying active and inactive WuObjects at once.

Figure 3.3: An example network with several strips



## 3.4 Reconfigurable Redundancy Architecture

Once the application is deployed, each target device would host one or many WuObjects where each represents a service for an application component. However, the network has to be resilient to failures, it has to detect and recovery autonomously. In the following paragraphs, we will be describing the subsystem which is used to support failure detection and recovery.

### 3.4.1 Decentralized Failure Detection

We want the heartbeat protocol to be able to reach the entire network so any failures will be detected. In order to avoid single point of failures, the protocol will have to be decentralized, thus the failure of one component would not bring the whole failure detection system down.

Distributed failure detection enables high-availability in distributed systems where partial failures are rather common. We utilize heartbeats to detect failures, a common technique widely used to detect failures in high-availability distributed systems. Heartbeats are messages sent periodically until it's unable to send messages anymore. Each node is therefore suspected dead when others stopped receiving messages from it after an extended period of time. Nodes were assumed to fail by stopping and will never come back.

There are abundant literatures on designing heartbeat protocols to ensure high-availability for distributed systems. Our work employed a heartbeat protocol arranged in such a way where each node sends a heartbeat to the previous node and the last node sends back to the first node forming a daisy chain as represented by the black arrows in figure 3.3.

To prevent tight coupling and redundancy in heartbeats, strips are separated from

heartbeats so the order of the strip does not affect the ordering of heartbetas and vice versa. The heartbeat protocol is a support layer below strips, the layers above will take advantage of the given information from the layer below to recover the system.

### **3.4.2 Failure Recovery**

When a failure is detected, there are two tasks that the system would have to do to recover from failures. First it has to make sure all members that carries the strips in the failure nodes will have consistent view of the strips. Second, it would need to propagate the changes to reconfigure other parts of the system that depend on the locations of the heads of the affected strips in order to function. The details of each task is described in the following sections.

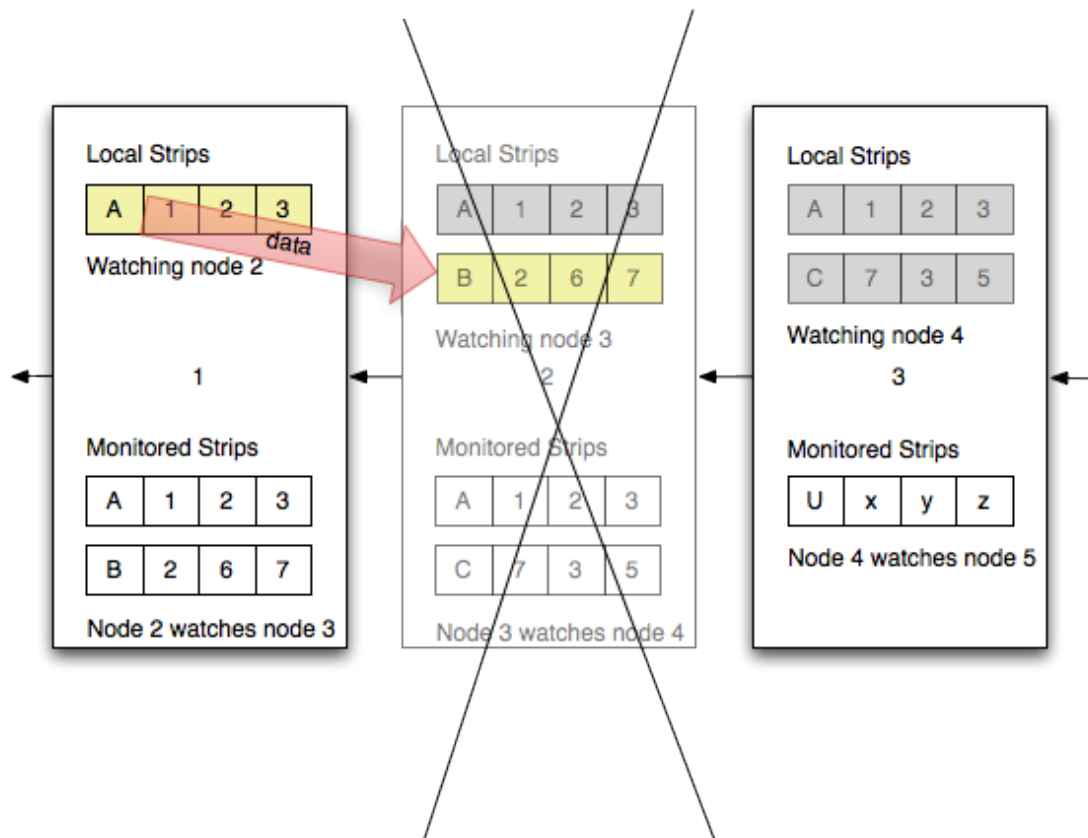
#### **Consistent view of strips**

Consistent view of strips is required to pick a replacement node in the event of failure. Without consistent view, nodes will not be able to know if the component has been replaced. For example, in figure 3.4, when node 2 failed, node 1 will detect this, but without algorithms to maintain consistency, node 3 would not be certain if node 1 or 2 are still alive, thus it might take actions that could compromise the network.

But detector might not know which strips the monitoring node is a member of, so every node will have a copy of its monitoring node's strips view as shown in figure 3.3 below the "Monitored Strips" section where node 1 has knowledge of the strips views of its monitoring node 2.

Our work proposed an algorithm attempting to recover by letting the detector of the failure to initiate the recovery algorithms. Since the detector will be responsible for

Figure 3.4: A failure occurred at node 2 in the network





recovering for the failed node, every node needs to have membership knowledge of the strips from the nodes it is monitoring. For example, if node A is monitoring node B, A would know the members of all strips in node B in addition to its local strips. Strips only specifies the order of recovery, it is not correlated with the network structure for the fault detection, in other words, a strip with A and B doesn't mean B is monitoring A, as B could be monitored by C which depends on the structure of heartbeat protocol layer. In the initial algorithm, the detector node will prepare a update message to inform all members of the strips with which the failed node is associated with. Assuming that every node that monitors other node will have knowledge of the strips that it contains and the members that the strips pertain. The node would send out a marker message first to confirm the nodes which are still functioning, and once all acknowledges have been received, it will proceed to send the update message to update their local knowledge of the strips to reach a consensus. The ordering of the messages wouldn't matter since the end state of any failure sequence for any strip would be the same. For example, given a strip of three members  $\rightarrow 1 - 2 - 3$ , if the updated failure sequence is given in any permutation by  $[1, 2]$  or  $[2, 1]$ , the end results would be the same  $\rightarrow 3$  since the remaining members from those two failure sequence is the same and the relative order of the members would stay the same. Therefore there is no need for extra communication overhead to maintain ordering to guarantee level of consistency between members since they will all come to the same conclusion given each receiver receive the same messages. The overhead are messages required to update each member's internal membership information.

## **Reconfiguration**

Even though consensus of the new view for each strip in the failure nodes has been reached, other devices in the network with connected components would also need to be

updated.

The detector after finishing synchronizing views, will initiate the reconfiguration algorithm. First, the initiator will identify application components that are connected with the components carried by the failed device (linked in the FBP). Then initiator would issue the update message to each head of the strips of the connected components to reconfigure the new heads of the affected strips.

As shown in figure 3.5, node 1 with its updated view of the strips after the failure of node 2 would need to update the new head of component A, which is still node 1 in this case, with the new head of component B since component B is connected to component A; node 1 also needs to update the new head of component B with the new head of component A as well. Thus node 1 will send a reconfiguration message to node 1 and node 6 about the changes in both heads of component A and component B.

To keep the nodes updated after reconfiguration, as most services in WuKong applications does not store any past data, the devices with connected components whose WuObjects originally are sending data to the WuObjects on the failed devices would force a data push to bring the new heads up to date, and vice versa, so instead the devices with connected components whose WuObjects originally are receiving data from the failed devices would force a data pull.

The detector needs to update its knowledge of the "Monitored" section after reconfiguration. The detector knows which node the monitoring node is monitoring, it will send a update heartbeat message request and instruct it to send heartbeats to itself, then it will send a request for its knowledge of local section and update its "Monitored" section, both the strips and the node it is monitoring. As illustrated in figure 3.6, node 1 will get updated with the heartbeat information from node 3, which is monitored by node 2. And also in figure 3.7, node 1 will also get updated with the strip information from node 3.

Figure 3.5: Reconfigure application links

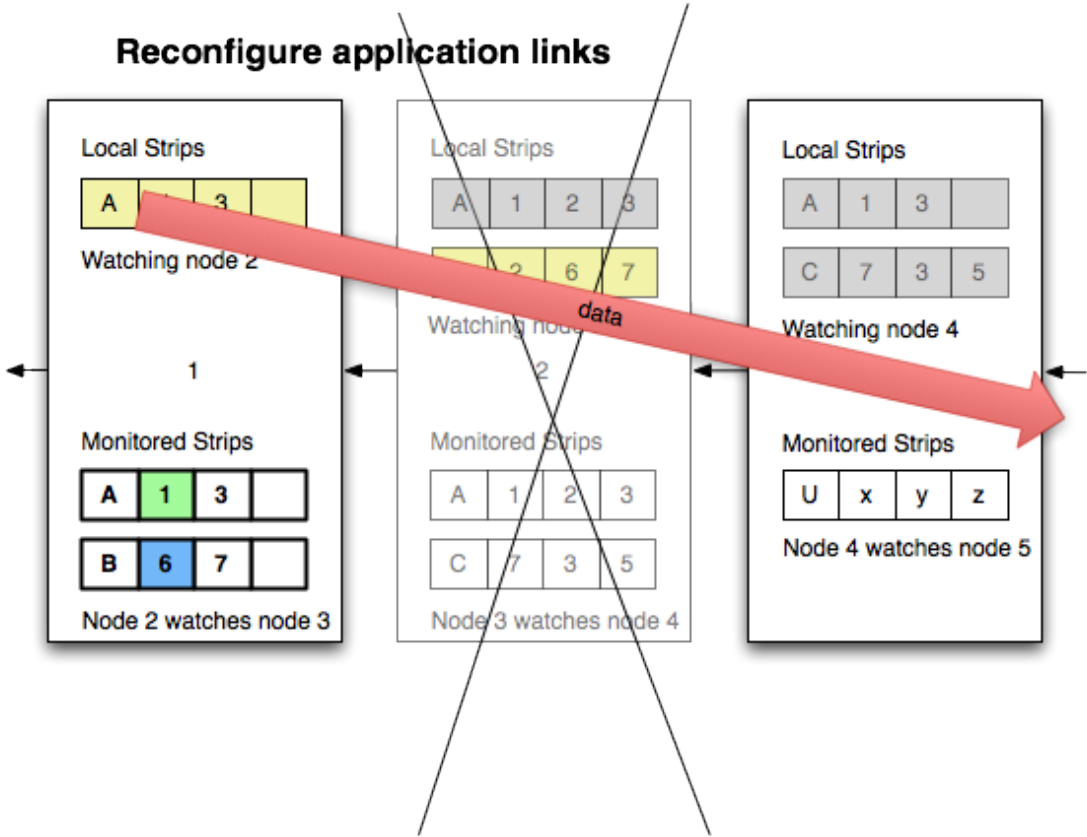


Figure 3.6: Reconfigure heartbeat protocols

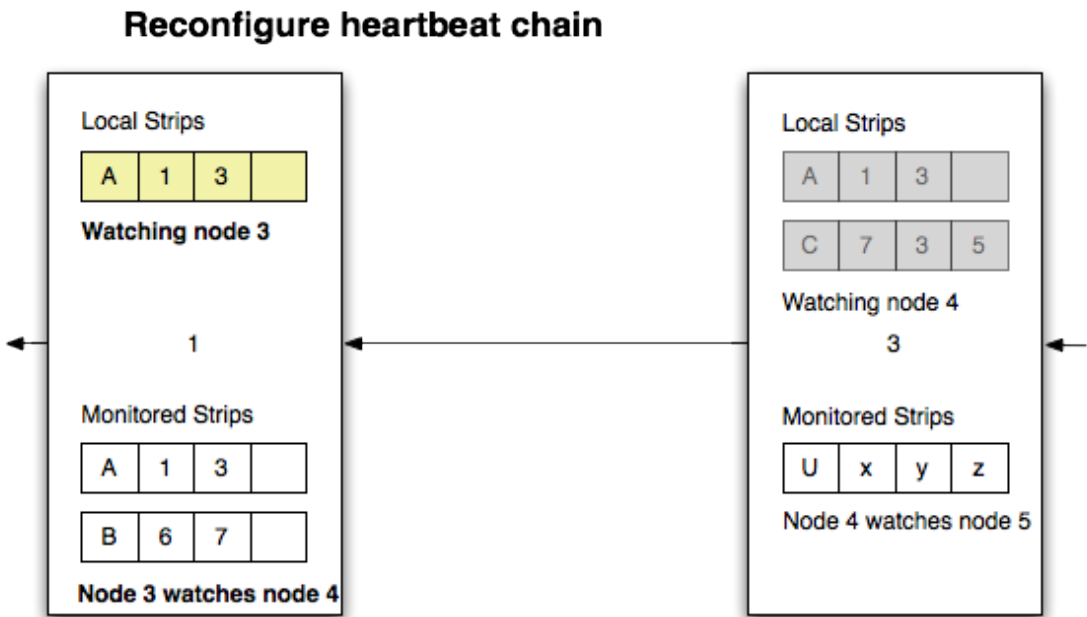
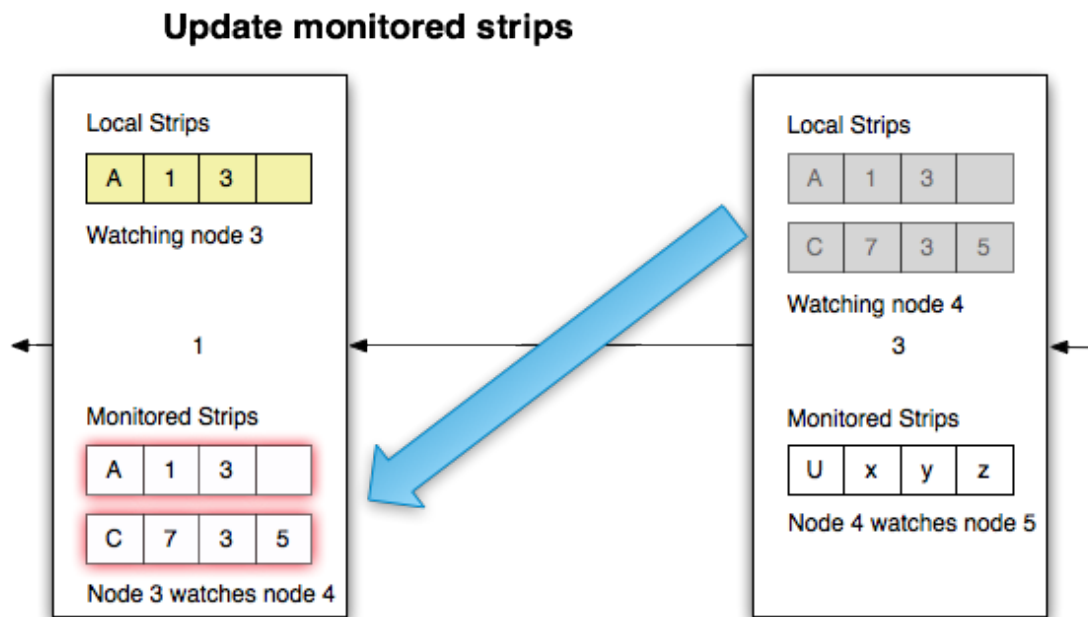


Figure 3.7: Update monitored information





## Chapter 4

# EVALUATION & RESULTS

In order to evaluate the performance of our fault tolerance system, we have introduced some metrics to test how well it perform, and whether after node failures requirements could still be met under small network.

1. Whether the next node in strip take over after failure
2. Memory overhead for Strips
3. Message overhead for failure recovery, including reconfiguration
4. Time to recover from failure

Whenever a node failed, the the next node in the strips it is carrying would take over for the respective services each represents. Time to recover measures the time it takes to recover from the time of failure detection.

We measured system performance live by collecting data from sensor nodes while running. Sensor nodes are programmed to send out their tracking data to a central data sink at appropriate times such as after node initialization or when the failure is resolved.

The application, fault tolerance policy, network topology are described in the following sections.

## 4.1 Application

Application shown in Figure 2.1 will be deployed. There will be four components:

1. Numeric Controller is a user input device which outputs a number from 0 to 255.  
Light Sensor is a photodetector sensor which detects the level of light intensity.
2. Threshold is a conditional function which takes two inputs, Threshold and Value, and, depending on the Operator attribute, return true if the Operator is set to GT (Greater Than) and the Value is higher than the Threshold.
3. Light Actuator is a relay intercepting the power source for a light bulb, it has a property OnOff which turns on the light if it is set to true, otherwise the light will be turned off.
4. Light Sensor is a sensor sensing the light intensity in the surrounding area.

## 4.2 Policy

The component fault tolerance policy for the application is set with the following parameters:

### **Numeric Controller**

Redundancy Level: 1

Fault Detection Time: 2 sec

### **Light Sensor**

Redundancy Level: 2



Fault Detection Time: 2 sec

### **Threshold**

Redundancy Level: 1

Fault Detection Time: 2 sec

### **Light Actuator**

Redundancy Level: 9

Fault Detection Time: 2 sec

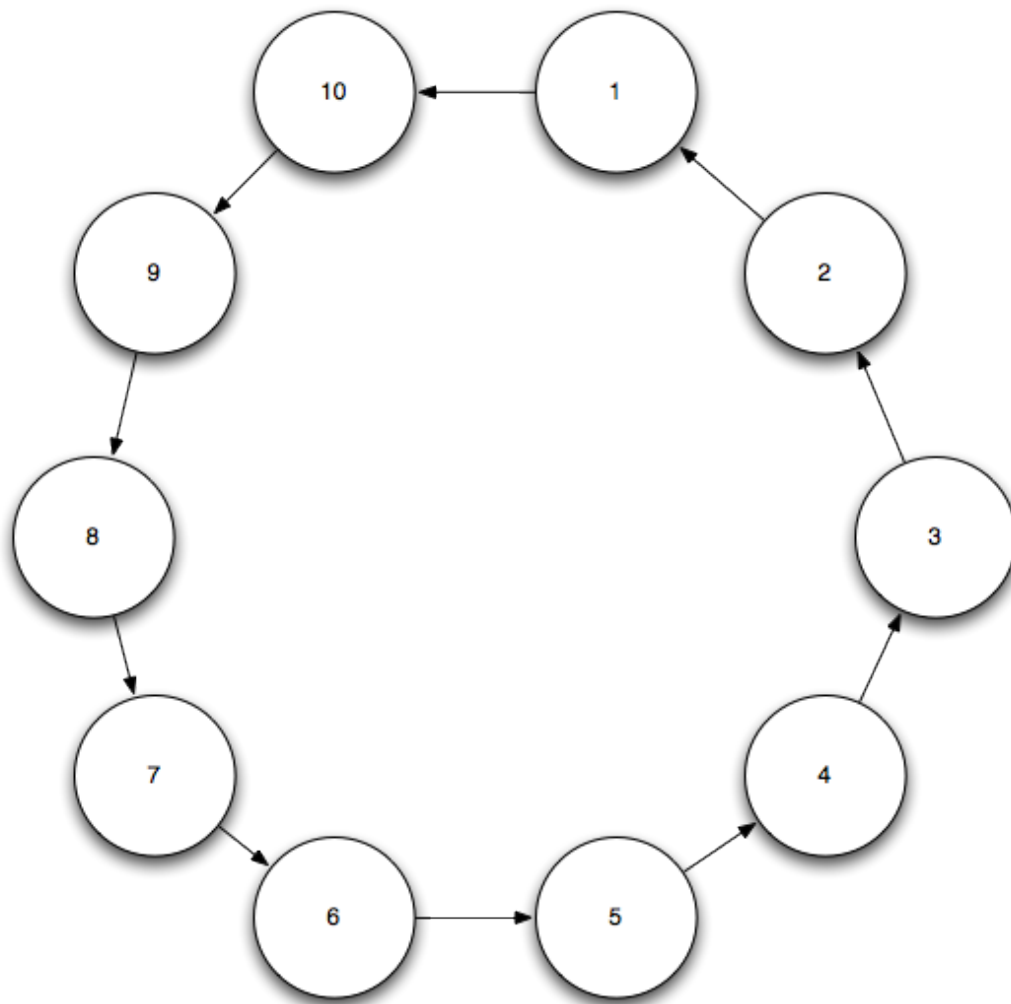
Since we set timeout at 2 times of heartbeat period, assuming the worst time to detect failure takes the full length of fault detection time, the heartbeat period is therefore 1 sec, which is one half of the fault detection time.

## **4.3 Heartbeat Protocol Arrangement**

We deployed 10 nodes in a room in our test lab, which results in a fully connected network. Therefore only one heartbeat chain loop is formed.

The heartbeat protocol arrangement is simple. Every node is sending heartbeat to previous node except the first node, which sends to the last. For example, node 1 receives heartbeat message from node 2, node 2 receives heartbeat message from node 3, etc. Figure 4.1 illustrates the arrangement for this experiment.

Figure 4.1: Heartbeat Protocol Arrangement



## 4.4 Hardware Platform

All boards are equipped with an Atmel ATmega1280-16AU 8-bit microcontroller with 4K of EEPROM and 64k of flash. The boards hardware design is based upon Arduino hardware referenced design, in addition, every board has wires for mounting multiple wireless protocol adapters such as ZWave, ZigBee. In the following experiments, every board is only equipped with a ZWave adapter, and only communicating through ZWave. Every board is also pre-installed with a modified version of NanoVM [5] called “NanoKong” [15] that supports all the basic WuKong framework protocols including the new additions from the work in the previous chapter. A PC with wireless access is dedicated for hosting the WuKong Master software which is responsible for managing WuKong applications for the whole system and serves as a mean to present an interface to the users. Three boards will be used in the experiments below. One of them is equipped with a light sensor that returns a byte indicating the light level around the sensor. The rest are equipped with a relay which each controls the power supply of a lamp. An additional board with the same hardware specification is used as a gateway between the Master and the sensor network.

## 4.5 Experimental Setup

Ten WuDevices are installed throughout our testbed. Every WuDevice will be able to talk to each other directly forming a fully connected network. Eight of them are equipped with light actuators. Two of them have light sensors. Only one of them has user input device (Numeric Controller), and Threshold. We simulate a node failure by removing power supply of a WuDevice. Every device communicates wirelessly through ZWave

Table 4.1: Node setup

<b>Node Id</b>	<b>Equipped resources</b>
1(2)	Light Actuator
2(4)	Numeric Controller, Threshold, Light Sensor
3(5)	Light Actuator
4(6)	Light Actuator
5(7)	Light Actuator, Light Sensor
6(10)	Light Actuator
7(12)	Light Actuator
8(13)	Light Actuator
9(14)	Light Actuator
10(15)	Light Actuator

Table 4.2: Strips

<b>Application Component</b>	<b>Mapped nodes (strip)</b>
Numeric Controller	2
Light Sensor	2, 5
Light Actuator	1, 3, 4, 5, 6, 7, 8, 9, 10
Threshold	2

adapter. The setup is shown in table 4.1

## 4.6 Mapping results

The result of the mapping and the strips are shown in table 4.2. Each row represents each component in the application, where strips are ordered from the left.

## 4.7 Results

The results for memory overhead by strips before failures are consistent, as each node address takes only one byte, the other byte is used in our unique identification system to recognize wuobject on a node. All failovers in deployments have been swift and correct.

Table 4.3: Strip memory overhead in bytes

<b>Application Component Strip</b>	<b>Memory size (bytes)</b>
Numeric Controller	2
Light Sensor	2
Light Actuator	18
Threshold	2

The figure 4.3 illustrates the average recovery time and message overhead over 5 deployments for each node failure in Strip for Light Actuator as first failure in the system. The first failure should on average takes the longest time and higher message overhead compared to consequent failures. Therefore measuring the performance for each node failure as first failure would give us how the system would perform the worst overall. The results carried over even in different strip orders for Light Actuator Strip since the ordering is just a matter of permuting the results as shown in the results.

The recovery time for most nodes were averaging around 2500 milli-seconds, node 3 and node 6 are found out that their radios were a little defective (without antenna) after the experiments therefore it took longer to complete the recovery. It is clear that the results have shown were pretty consistent as there were only a constant number of nodes that needed to contact to recovery regardless of how many strips the node contained. The time it took is reasonable given the small network.

The lower bound of message overhead for heartbeats arranged in a daisy chain loop in a network of size  $n$  is  $\Omega(n)$ , the upper bound is also  $O(n)$ , since there is only one message sent by a node at any given moment.

For an application with  $m$  components deployed to a network with  $n$  nodes where  $n \geq 1$ , the lower bound of memory overhead of the strips for a random member in the network is  $\Omega(1)$  since a node must carry at least one strip and lowest member requirement of a strip is 2. The upper bound is  $O(m * n)$  since the strips could span the whole network.

But in reality, since the nodes typically have fixed memory and fixed capability for the duration of the lifetime in the network, the upper bounds cannot grow indefinitely. The memory overhead a strip could impose on a node is determined by node's capability and memory size.

The lower bound of message overhead by the reconfiguration protocol with one node failure in the same network is  $\Omega(1)$ , since if there is only one component with 2 members on the failed node, the detector would only need to send 1 messages to the remaining strip member of the failing node, and none if the component is not connected to any other components, and finally 1 more message to get the information from the nodes monitored by the failed node. The upper bound is  $O(n + m)$  since if the failed node carried  $m$  components where each strip contains  $n$  members, the detector would have to send  $(n-2)$  messages to all functioning members of the  $m$  strips carried excluding itself plus the messages to all other strip heads connected to the failing node.

The time to recover highly depends on the messages sent for the reconfiguration and the detection time. The lower bound is  $\Omega(1)$  with the same assumption with constant component and members, but the upper bound is  $O(b + n + m)$  where  $b$  is the heartbeat period.

The cost for time for recover is the time itself. Time is the overhead. Thus the longer time it take to recover the higher the overhead. By setting a shorter heartbeat period, it would take a shorter time to recover, thus a lower overhead. I assume every message takes at least a fixed amount of time, and there need at least amount of messages to recover, so the more messages it takes to recover, the longer the time it will take to recover. Here it is assumed that heartbeat messages are never lost/dropped, and in-node computation take negligible amount of time thus it is ignore. If it is about the message overhead within a certain time period, then of course the higher the period the lower the message overhead.

A small network with around 10 nodes with a 5-component application can operate if each node could dedicate at most 50 bytes of memory to strips (assuming one byte node addressing), and equipped with radios and battery capable of handling extra 15 messages for reconfiguration messages per failure. The requirements are reasonable since most embedded devices have at least 4K of EEPROM to store strips, and have radios with throughputs of 40kbps, thus the ideal recovery time without in-node processing and interference and retry would be around 3 milli-seconds.

Depending on the wireless protocol used, a ZWave network can include up to 232 nodes, which is pretty big for most real world deployments. Strips can still be run in network of size like this, since a 4K EEPROM could hold up to 400 members per strip for a 5 component application.

Figure 4.2: An WuDevice

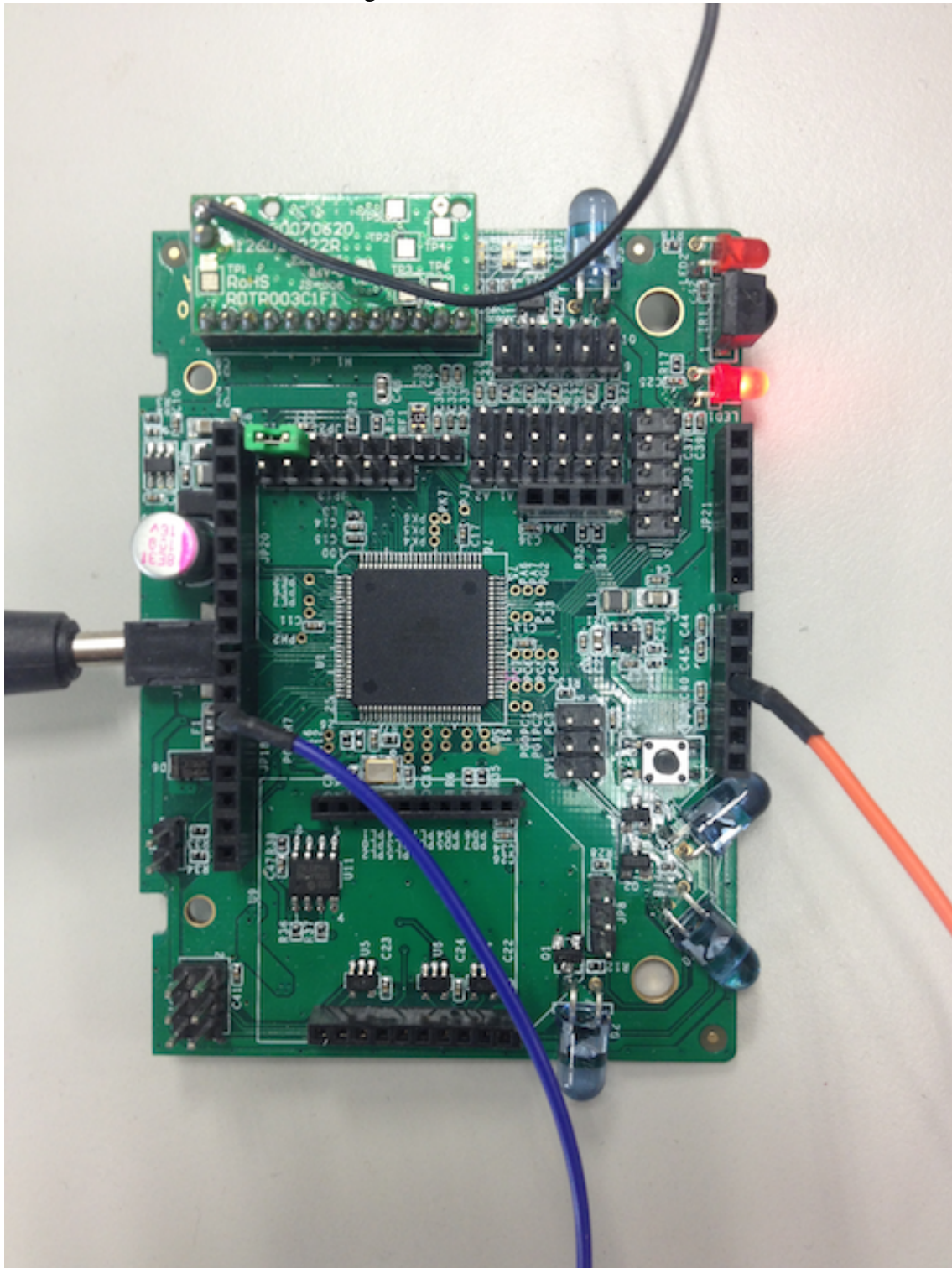
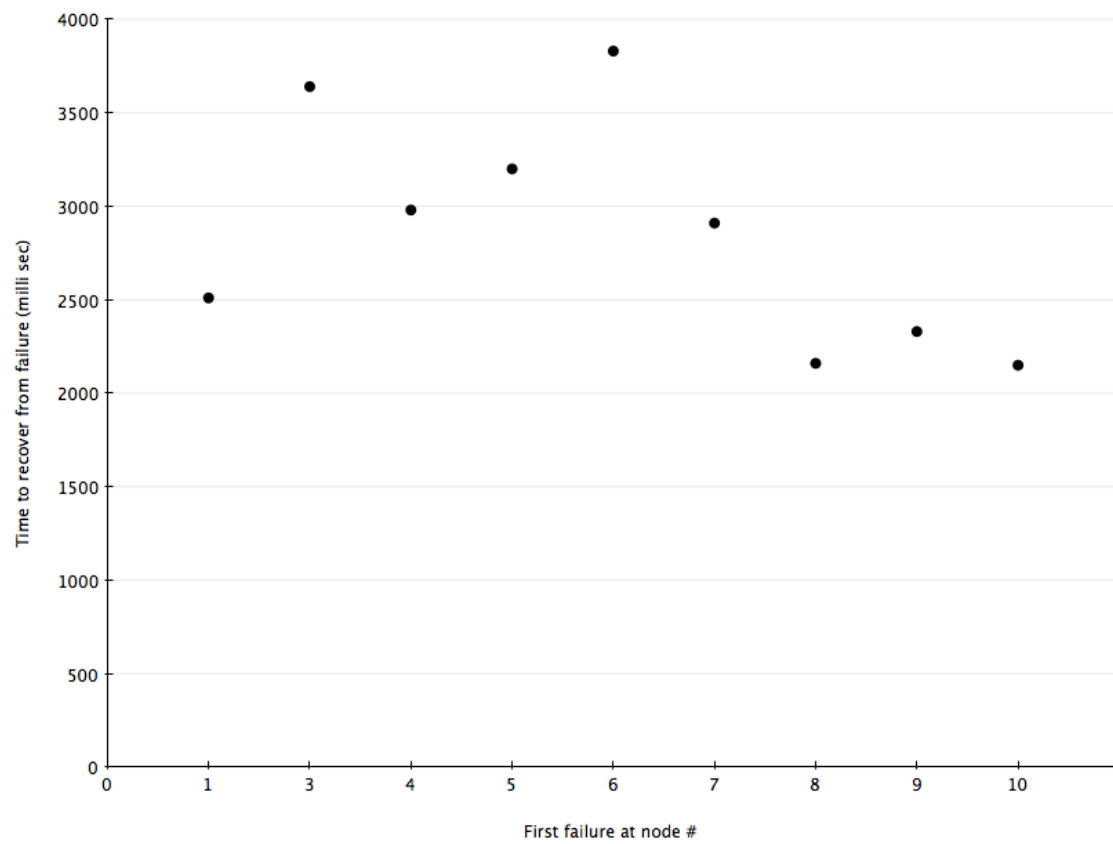




Figure 4.3: Average recovery time and message overhead over 5 deployments for each node failure as the first failure





# Chapter 5

## CONCLUSION

### 5.1 Discussion

This thesis proposed a novel algorithm to achieve recovery from failures by combining heartbeat protocol, for failure detection, with Strips, which are used to maintain and track service redundancy.

We have presented a fault tolerance system able to provide failover for failed services in service-oriented WSNs that comply with user policy requirements. We have also described strip, a redundancy abstraction for service peers along with distributed algorithms to synchronize strip views among members and reconfigure the network for the new structure to recover from node failure. The system allows user intervention through means of user policy, which could directly influence underlying system configurations and structure.

The developed methods adds new and useful solutions to build a fault tolerant system that could be reasoned easily and with a performance as expected on average. This method might not be superior to other methods in terms of completeness and complexity, but it serves as a quick and easy solution to provide practical fault tolerance for WuKong applications.

The experimental results have shown to be consistent and stable among first failures in different rank of members in strip around 2.5 secs. The failover have been successful in all deployments. It is also shown that the performance degraded quickly when the hardware or the wireless communication quality degraded. Therefore it is important that the network setup is as optimized as possible.

## 5.2 Future Work

We have shown a design for a reconfigurable fault tolerant system for WuKong. Strips makes it really easy to describe a component system with redundancy for heterogeneous services and devices. Nevertheless, there is still room for improvements. This section will address some directions future research can take.

WuKong Fault Tolerance System didn't considered for multi-hop networks where it introduces higher latency and network partitions. Network partition is when a network of nodes got partitioned into two subnetworks where none can detect each other for a period of time. One of the possible direction is to create a more complex failure model that could handle network partition. In this thesis we assume failstop model where nodes, once dead, will not come back. Therefore when network partition occurs, each subpart of the network would not be able to recognize each other and would cause conflicts and confusions.

Our current heartbeat protocol is distributed and easy to construct, but it is not shown to work under networks where messages are sent in multiple hops, since the algorithm used to produce where each node should be sending heartbeat messages does not consider the topology of the network. Heartbeat is sensitive on latency, so if a heartbeat message was not received within tolerance period, a failure event could occur and the node is

suspected of failure and will never come back. If the node is still alive, it would be treated as if it is dead. And that will create an artificial network isolation where a few nodes are excluded from the network before of latency.

Although we didn't mention in system design, it is possible to handle multiple consecutive errors by storing in advance more node's strip and heartbeat protocol data, so when consecutive nodes failed simultaneously, the surviving detector would be able to recover all dead services and reconfigure the network correctly. But there is a limitation of the memory a node could store, a possible future direction for research could be to improve the design so it could handle multiple simultaneous failures in large network.

Another important element is the optimization problem for deployment of applications. This thesis didn't consider finding an optimal deployment for the level of redundancy specified in the user policy. The problem of deploying a specific distributed system onto a network structure typically consists of mapping the components of the system onto the hosts of the network. The mapping is subject to constraints. The constraints could be whether a node supports certain service to host certain components, and how much communication overhead would induce from the assignment to maintain consistency for the strips, and from the perspective of WuKong, some components need to separate from other components to achieve fault tolerance, and some need to be placed together to function properly. Determining such an optimal deployment is a combinatorial optimization problem, and combinatorial optimization problems are generally extremely challenging computationally. It is difficult to predict what will and what will not work. It is unlikely that a single approach will be effective on all problems or instances of the same problems. As we also want the system to come up with a solution within a time limit. So finding a good balance between the quality of a solution and the time it takes to come up with a good enough solution is critical.



# Bibliography

- [1] A. Arora, P. Dutta, S. Bapat, V. Kulathumani, H. Zhang, V. Naik, V. Mittal, H. Cao, M. Demirbas, M. Gouda, Y. Choi, T. Herman, S. Kulkarni, U. Arumugam, M. Nesterenko, A. Vora, and M. Miyashita. A line in the sand: A wireless sensor network for target detection, classification, and tracking. *Computer Networks*, 46:605–634, Jan. 2004.
- [2] L. Atzori, A. Iera, and G. Morabito. The Internet of Things: A survey. *Computer Networks*, 54(15):2787–2805, 2010.
- [3] G. Barrenetxea, F. Ingelrest, G. Schaefer, and M. Vetterli. The hitchhiker’s guide to successful wireless sensor network deployments. In *Proceedings of the 6th ACM conference on Embedded network sensor systems (SenSys ’08)*, volume D, pages 43–56. ACM Press, 2008.
- [4] A. Bharathidasan, V. An, and S. Ponduru. Sensor Networks: An Overview. Technical report, Department of Computer Science, University of California, Davis, 2002.
- [5] T. Harbaum. The NanoVM - Java for the AVR, 2006.

- [6] D. Hughes, K. Thoelen, J. Maerien, N. Matthys, W. Horre, J. Del Cid, C. Huygens, S. Michiels, and W. Joosen. LooCI: The Loosely-coupled Component Infrastructure. In *11th IEEE International Symposium on Network Computing and Applications (NCA 2012)*, pages 236–243. IEEE, Aug. 2012.
- [7] X. Koutsoukos, M. Kushwaha, I. Amundson, S. Neema, and J. Sztipanovits. OASiS: A Service-Oriented Architecture for Ambient-Aware Sensor Networks. In *Composition of Embedded Systems. Scientific and Industrial Issues*, pages 125–149. Springer Berlin Heidelberg, 2007.
- [8] J. Neumann, N. Hoeller, C. Reinke, and V. Linnemann. Redundancy Infrastructure for Service-Oriented Wireless Sensor Networks. In *9th IEEE International Symposium on Network Computing and Applications (NCA 2010)*, pages 269–274. IEEE Computer Society, July 2010.
- [9] P. Padhy, K. Martinez, A. Riddoch, H. L. . R. Ong, and J. K. Hart. Glacial Environment Monitoring using Sensor Networks. In *Real-World Wireless Sensor Networks*, pages 10–14, 2005.
- [10] K. Piotrowski, P. Langendoerfer, and S. Peter. tinyDSM: A highly reliable cooperative data storage for Wireless Sensor Networks, 2009.
- [11] J. Polastre, R. Szewczyk, A. Mainwaring, D. Culler, and J. Anderson. Analysis of wireless sensor networks for habitat monitoring. In C. S. Raghavendra, K. M. Sivalingam, and T. Znati, editors, *Wireless Sensor Networks*, pages 399–423. Kluwer Academic Publishers, 2004.
- [12] S. Ratnasamy, B. Karp, L. Yin, F. Yu, D. Estrin, R. Govindan, and S. Shenker. GHT: A Geographic Hash Table for Data-Centric Storage. In *Proceedings of the 1st ACM*



*international workshop on Wireless sensor networks and applications (WSNA '02)*, volume 5, pages 78–87. ACM, 2002.

- [13] N. Reijers, K.-j. Lin, Y.-c. Wang, C.-s. Shih, and J. Y. Hsu. Design of an Intelligent Middleware for Flexible Sensor Configuration in M2M Systems. In *Presented at 2nd International conference on sensor networks (SENSORNETS 2013)*, 2013.
- [14] I. Stoianov, L. Nachman, S. Madden, and T. Tokmouline. PIPENET: A Wireless Sensor Network for Pipeline Monitoring. In *Proceedings of the 6th international conference on Information processing in sensor networks (IPSN '07)*, pages 264–273. ACM, 2007.
- [15] P. Su, N. Reigers, and S. Zhou. NanoKong, 2012.
- [16] J. Tateson, C. Roadknight, A. Gonzalez, T. Khan, S. Fitz, I. Henning, N. Boyd, C. Vincent, and I. Marshall. Real World Issues in Deploying a Wireless Sensor Network for Oceanography. In *Unpublished paper presented at Workshop on Real-World Wireless Sensor Networks (REALWSN '05)*, 2005.
- [17] G. Tolle, J. Polastre, R. Szewczyk, D. Culler, N. Turner, K. Tu, S. Burgess, T. Dawson, P. Buonadonna, D. Gay, and W. Hong. A macroscope in the redwoods. In *Proceedings of the 3rd international conference on Embedded networked sensor systems (Sensys '05)*, pages 51–63. ACM, Jan. 2005.
- [18] G. Werner-Allen, K. Lorincz, J. Johnson, J. Lees, and M. Welsh. Fidelity and yield in a volcano monitoring sensor network. In *Proceedings of the 7th symposium on Operating systems design and implementation (OSDI '06)*, pages 381–396. USENIX Association, Jan. 2006.