

BACHELOR THESIS

Extension of a SCADA Framework to support High Availability and Authenticated Encryption

#Ruby #ØMQ #NaCl

Patrik Wenger, Manuel Schuler

client: mindclue GmbH

supervisor: Prof. Dr. Farhad Mehta

expert: Sören Bleikertz

September – December, 2016

Abstract

Declaration of Originality

We hereby confirm that we are the sole authors of this document, the described changes to the Roadster framework, and libraries developed as a byproduct. Unless stated differently, all illustrations in this document are our creations.

Acknowledgements

Special thanks to Pieter Hintjens † (3 December 1962 – 4 October 2016) for his amazing work and contagious passion within the ØMQ and distributed computing communities. We send our deepest condolences to his family. Rest in peace.

Contents

I	Management Summary	1
1	Context	2
1.1	Initial Situation	2
1.2	Goals	2
1.3	Software development process	2
1.4	Project management infrastructure	3
2	Project Phases	4
2.1	Plan	4
2.2	Inception	4
2.3	Elaboration	4
2.4	Construction	4
2.5	Transition	4
3	Results	5
II	Technical Report	6
4	Scope	7
4.1	Motivation	7
4.1.1	Personal backgrounds	7
4.1.2	Opportunities	8
4.1.3	Open-Source engagement	8
4.2	Initial Situation	8
4.2.1	mindclue GmbH	8
4.2.2	Roadster	9
4.2.2.1	System integration	9
4.2.2.2	Typical hardware	10
4.2.3	ØMQ	11
4.2.4	Software architecture	11
4.2.4.1	Communication Layers	12
4.2.4.2	RMP	13
4.2.4.3	DIM	14
4.2.4.4	Existing CSP in a nutshell	14
4.2.4.5	Persisted data	16
4.3	Goals	16
4.3.1	Security concerns of SCADA applications	16
5	Requirements	18
5.1	Federation	18
5.1.1	DIM extension	19
5.1.1.1	Synchronization	19
5.1.1.2	Access control	20

5.1.2	Autonomy	20
5.1.3	Message routing	21
5.2	High availability	21
5.3	Persistence synchronization	24
5.4	Non-functional requirements	25
5.4.1	Simplicity	25
5.4.2	Testing	25
5.4.3	High availability for OPC UA	26
5.4.4	Encryption	26
5.4.5	Coding Guidelines	27
6	Approach	28
6.1	Getting familiar with Roadster	28
6.2	Testing	28
6.2.1	Setup	28
6.2.2	Unit tests	29
6.2.3	Integration tests	29
6.2.4	Continuous integration	29
6.2.5	System test	29
6.2.5.1	Test scenarios	29
6.3	Port to new ØMQ library	30
6.3.1	Actual port	30
6.4	Federation	30
6.4.0.1	Fallacies of distributed computing	31
6.4.1	DIM synchronization	31
6.4.1.1	CAP theorem	32
6.4.2	Node topology definition	32
6.4.3	Message routing	33
6.4.3.1	Example	33
6.5	High availability	33
6.5.1	Defining reliability	34
6.5.2	Binary Star in a nutshell	34
6.5.3	Failover	35
6.5.3.1	Alarm generation	36
6.5.3.2	Failover from backup to primary node	36
6.5.4	Side benefit: Rolling upgrades	36
6.5.5	A note on dedicated links	36
6.5.5.1	Extending Binary Star Pattern	36
6.5.5.2	Dangerous corner case	37
6.5.6	Single level	37
6.5.6.1	Caveats	37
6.5.6.2	Link failure between Roadster node and field device	38
6.5.6.3	Supporting different field devices	38
6.5.7	Multi Level	38
6.6	Persistence synchronization	38
6.6.1	Aspects	38
6.6.2	Variants	39
6.6.2.1	Polling only	39
6.6.2.2	PUSH-PULL	39
6.6.2.3	PUB/SUB	40
6.6.3	Chosen Variant	40
6.7	Encryption	41
6.7.1	Key generation and distribution procedure	41
6.7.1.1	Client authentication	41
6.7.2	In code	41
6.8	OPC UA Interface: High availability	42

7	Results	43
7.1	Port	43
7.2	Federation	43
7.3	High Availability	43
7.4	Persistence synchronization	43
7.5	Encryption	43
7.6	OPC UA Interface: High availability	43
8	Discussion	44
8.1	Value Added	44
8.2	Limitations	44
8.3	Business Benefits	44
8.4	Ideas for Improvement	44
9	Conclusion	45
III	Appendix	47
A	Self Reflection	48
B	Task Description	49
C	License	55
D	Project Plan	56
D.1	Timetable	56
D.1.1	Estimated time	56
D.1.2	Time Tracking	56
D.1.3	Infrastructure	56
D.1.4	Tools	57
D.1.5	Quality Measure	57
D.1.5.1	Documentation Review	57
D.1.5.2	Meeting Minutes	57
D.1.5.3	Git Policy	57
D.1.6	Meeting	57
D.2	Risks	57
D.2.1	Handling Risks	57
D.3	Listed Risks	58
D.4	Iterations & Phases	61
E	ØMQ	63
E.1	Transport security	63
E.2	Data serialization	64
E.3	Language availability	64
E.4	CZMQ	64
F	Infrastructural Problems	65
F.1	Project Management Software	65

List of Figures

4.1	Roadster's place within the overall system	10
4.2	Roadster's software architecture	12
4.3	Roadster's communication layers	13
4.4	Class diagram for Roadster's meta model used in the DIM	15
5.1	Physical legacy example: a single node and a field device each	19
5.2	Physical federation topology example: supernode, two subnodes, a field device each	20
5.3	Federation example: a HA cluster and redundantly connected field devices	23
5.4	Federation example: HA cluster at root, two subnodes, each with field devices	23
6.1	Federation between a supernode and two subnodes	31
6.2	Single level HA setup between a HA pair and a field device (PLC)	37
6.3	Multi level HA setup between a HA pair and a number of client nodes	38

List of Tables

D.1	Timetable	56
D.2	Timetable	56
D.3	Timetable	57
D.4	Initial Risks	58
D.5	Initial Risk Matrix	59
D.6	Risk-Timeline change protocol	61
D.7	Phase and Iterations	61

List of Listings

5.1	Formal federation feature	19
5.2	Formal DIM synchronization feature	21
5.3	Formal autonomy feature	22
5.4	Formal message routing feature	24
5.5	Formal high availability feature	25
5.6	Formal persistence synchronization feature	26
6.1	Federation DSL example without HA	33
6.2	Fedreation DSL example with HA	34
6.3	Federation DSL example with HA and roles	35
6.4	Starting an authentication handler that allows any clients	42

Part I

Management Summary

Chapter 1

Context

1.1 Initial Situation

Roadster is mindclue GmbH's in-house framework to build modern monitoring and controlling applications in different fields such as traffic systems, energy, and water supply. It is written in Ruby, a modern and expressive scripting language, and is built on a shared-nothing architecture to avoid a whole class of concurrency and scalability issues found in traditional application architectures.

Although considered to be the next generation of its kind, it still lacks important features such as the ability to be run on multiple nodes in a federation, high availability, and secure network communications.

1.2 Goals

Adding the aforementioned, missing features to form the next version of the framework would mean a distinct advantage for mindclue GmbH and thus increase its competitiveness in its sector.

Planning the exact architectural changes and additions, as well as performing the implementation is the students' goal for this bachelor thesis. Using engineering methodology practiced at HSR, solutions for particular problems will be worked out and the best fitting one will be chosen.

Although not exactly part of the requirements, spreading knowledge about Roadster's architecture and code basis is also in the interest of the client, as Andy Rohr is currently the framework's only developer and thus a single point of failure in an increasingly important piece of software.

1.3 Software development process

The [Rational Unified Process \(RUP\)](#) is used to plan and manage this term project. It's an iterative, structured, yet flexible development process which suits this kind of project. At HSR, it's taught as part of the Software Engineering courses and is thus a primary candidate.

Another candidate was Scrum, which we decided against as it's only feasible with teams of three to nine developers.

1.4 Project management infrastructure

The source code of this document and all of our code contributions are hosted on GitHub. The students will organize and perform their work directly on the site as far as possible. This means creating a Project board for each of the development phases, creating, assigning, and closing issues, as well as using the Wiki feature to plan and document meetings with the professor and the client.

Time tracking, as required by the process for bachelor theses [**hsr:thesis-rules**], are done externally on Everhour.

Chapter 2

Project Phases

2.1 Plan

2.2 Inception

2.3 Elaboration

2.4 Construction

2.5 Transition

Chapter 3

Results

Part II

Technical Report

Chapter 4

Scope

The technical goals of this bachelor thesis include extending mindclue GmbH's Roadster framework by adding features such as clustering, high availability and transport security. This chapter outlines the general scope of this project.

4.1 Motivation

4.1.1 Personal backgrounds

To better understand our motivation, it might help to understand our personal backgrounds first.

Patrik Wenger did his apprenticeship in computer science at Swisscom Schweiz AG, and stayed work as a full-time employee for five more years afterwards. In programming he's most fluent in [Ruby](#) and [C](#). During the winter of 2015/2016, he created [CZTop](#) during leisure time because there was no good Ruby binding for [ØMQ/CZMQ](#) available and a side project of his demanded it. Fascinated with event-driven programming and software design patterns such as the [Actor Model](#) (e.g. the [Celluloid](#)¹ library on Ruby, or [Pony](#)², distributed computing and high availability have long been part of his core interests, especially in conjunction with the brilliant [ØMQ](#) library. Having a passion for information security and modern cryptography³, especially in this post-Snowden era, this bachelor thesis couldn't be a better match.

Manuel Schuler did his apprenticeship in computer science at Alcatel-Lucent AG. The most projects during the apprenticeship or other companies involved network monitoring or configuration automation. In programming he's most fluent in Node.js, Java and .NET. He made several projects to keep his life simple. After a while he decided to start his own business. Always keen on learning new things and the fact he often made similar things like Patrik Wenger during his work career motivated him to learn how Patrik did the things he did, so he did not hesitate to join this bachelor thesis at the first opportunity.

In essence, both students are thrilled to gain more experience in the following fields and technologies:

- Distributed Computing

¹a concurrency framework for Ruby based on the actor model, <https://github.com/celluloid/celluloid>

²a young programming language completely based on actors, <http://www.ponylang.org>

³such as [NaCl](#) or [libsodium](#) as used by [ØMQ](#)

- High Availability
- Information Security
- [Actor Model](#)
- [ØMQ](#)
- [Ruby](#)

4.1.2 Opportunities

Coming from different backgrounds and having different levels of experience in each of the above technologies, we can't wait to learn more about them and put them to actual use. The fact that the product of this bachelor thesis is most likely going to be used in the real world only adds to the excitement.

This bachelor thesis involves working with Ruby, the Actor Model, ØMQ, distributed computing with high availability, and state-of-the-art cryptography. Furthermore, in case of successful completion of this thesis, the results will be used in real-world settings like the Ceneri Base Tunnel. It is a huge opportunity for a solution completely based on free and open-source software interacting with other industrial systems over open standards. The students, as well as the client, strongly believe in customized solutions built on reusable, free open-source software.

In addition to that, we look at this bachelor thesis as an opportunity to become more fluent in English, both written and spoken, as well as to improve our skills in crafting scientific documents using \LaTeX .

Depending on how we perform together as a team, further collaboration might result in the future, either between the students themselves, or between the students and the client. Even if our paths will part, this project will serve as a valuable reference for future job hunting.

Last but not least, we feel like Prof. Dr. Mehta is a respected and competent teacher whose opinions we highly value. Due to his polite parlance, discussing project matters, both of the management and the technical kind, has always been an enrichment.

4.1.3 Open-Source engagement

Getting the chance to use [CZTop](#) and watch it perform definitely adds to the motivation as well. Its software design has yet to be proven in more serious settings.

Another personal goal is to create a reusable open-source library as a byproduct. The intention is that the library makes certain ØMQ-based communication protocols readily available for other developers facing the same problems.

4.2 Initial Situation

4.2.1 mindclue GmbH

The company mindclue GmbH, located in Ziegelbrücke GL, provides its partner REMTEC AG with complete [Supervisory Control and Data Acquisition \(SCADA\)](#)⁴ applications. These are then

⁴SCADA software resides in level 2 of the enterprise levels (0–4) modeled by the [ISA-95](#) standard, https://en.wikipedia.org/wiki/Enterprise_control

used to supervise and control operation and safety equipment found in:

- national freeways, e.g. emergency phones
- tunnels, e.g. lights and ventilation
- water supply systems
- energy facilities
- many other specialized fields

To build these customized applications, their in-house creation Roadster, a next-generation SCADA framework, is used.

4.2.2 Roadster

Roadster is a SCADA framework written in Ruby. It was, and still is, developed to produce next-generation SCADA applications to replace legacy solutions based on its predecessor found in numerous tunnel facilities in Switzerland.

A Roadster installation combines the following responsibilities:

- interaction with subordinate field devices (monitoring & controlling)
- persisting data (e.g. certain sensor data, and events)
- sophisticated alarm (*case*) management
- providing a machine-to-machine interface to higher level systems
- providing a modern, customized web UI for interaction with operational and executive personnel

Among others, the field devices include various kinds of [Programmable Logic Controllers \(PLCs\)](#)⁵ as well as emergency call systems⁶. These are interacted with over numerous proprietary and standardized protocols.

4.2.2.1 System integration

This section briefly describes the big picture of Roadster's place within typical production environments and its relationship with other systems.

In some deployments, Roadster has no higher-level client systems. The only clients would be its human users.

In other deployments, there are higher level systems which act as clients of a Roadster instance, communicating over protocols including [Service Oriented Application Protocol \(SOAP\)](#) and [OPC UA](#). Their purpose is to collect and aggregate supervisory data from larger regions. At the top of the hierarchy are the [Federal Roads Office \(FEDRO\)](#) (German: [Bundesamt für Strassen \(ASTRA\)](#)) which combine the information of all subsystems to provide a nationwide overview.

[Figure 4.1](#) illustrates the typical, overall system. The line style meanings also apply to the remaining physical topology illustrations. The acronyms used, which are part of the [FEDRO](#) terminology, don't have official English translations; not even the department itself was able to

⁵an example is the SIMATIC S7-1500 by Siemens AG, <http://w3.siemens.com/mcms/programmable-logic-controller/en/advanced-controller/s7-1500/Pages/default.aspx>

⁶an example is the NIS ComNode by Trans Data Management AG, <http://trans-data.com/en/k2-categories/item/149-niscomnode>

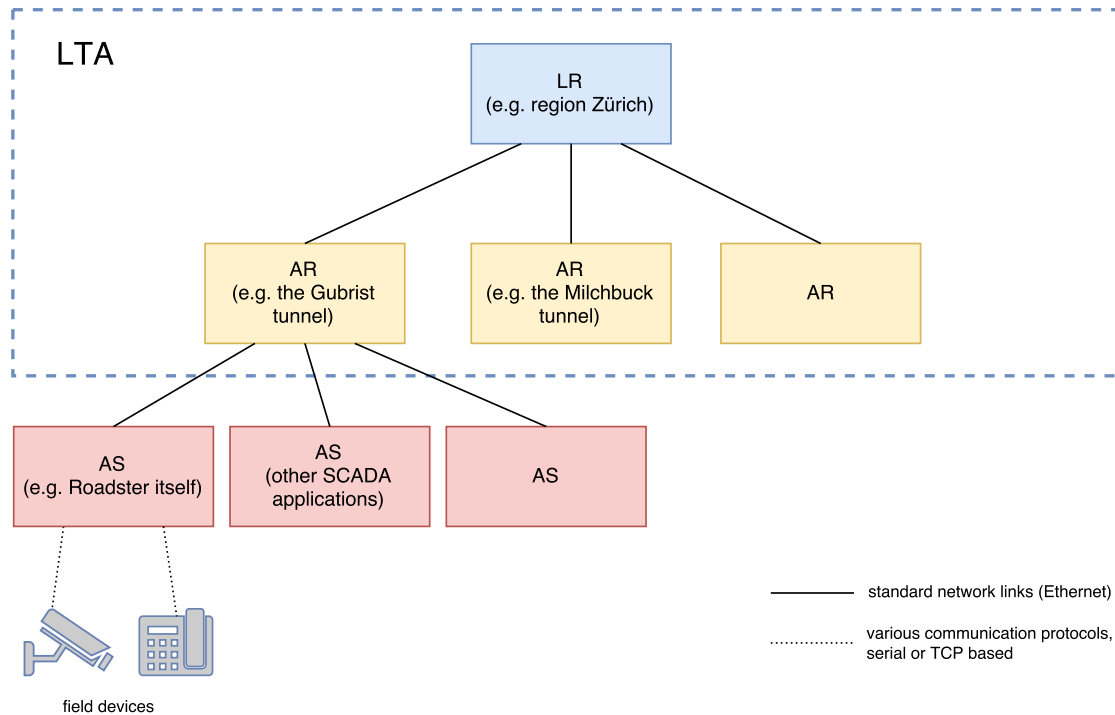


Figure 4.1: Roadster's place within the overall system

help with translations on request, so they were left unchanged. A brief description follows, from the bottom up:

Field devices:

Field devices are various kinds of [PLCs](#) and other subordinate systems used to supervise and control industrial processes. Communication with them happens over protocols such as [Modbus TCP](#), [IEC 60870-5-104](#), and [OPC UA](#).

Anlagesystem (AS):

This is Roadster's domain, or of course the one of another product with similar functionality. An AS is responsible for one facility (e.g. emergency call system, lighting system, fire alarm system, video monitoring system, ventilation system, power supply system, train signaling system).

Abschnittsrechner (AR):

The higher level system of the collection of all AS found in one larger facility such as a tunnel. With the results of this bachelor thesis, this *could* be Roadster's domain as well.

Leitrechner (LR):

The higher level system of the collection of all AR found in a region such as Zürich. As with AR, this *could* be Roadster's domain as well.

Leittechnikanlage (LTA):

This collective term comprises both the levels of AR and LR. For simplicity and readability's sake, this will be referred to as *client* for the remainder of this document.

4.2.2.2 Typical hardware

Roadster typically runs on entry-level rack server hardware powered by an Intel® Xeon® processor, or industrial box PCs for smaller systems commonly used for [Internet of Things \(IoT\)](#) which are powered by more energy efficient processors such as Intel® Core® and Intel® Atom™. The machines are usually equipped with 4 – 6 GiB of main memory and Gigabit Ethernet. For

reliable systems without any moving parts, an industrial grade [Solid State Disk \(SSD\)](#) or two (in a software [redundant array of independent disks \(RAID\)](#) level 1 setup) are used.

4.2.3 ØMQ

To understand Roadster’s architecture and the rest of this document, it’s helpful to understand the basics of ØMQ first. This is a brief introduction to ØMQ for the unfamiliar reader. What follows is a quote from the [Zguide](#) which does a fairly good job at describing ØMQ in a 100 words:

“ZeroMQ (also known as ØMQ, 0MQ, or zmq) looks like an embeddable networking library but acts like a concurrency framework. It gives you sockets that carry atomic messages across various transports like in-process, inter-process, TCP, and multicast. You can connect sockets N-to-N with patterns like fan-out, pub-sub, task distribution, and request-reply. It’s fast enough to be the fabric for clustered products. Its asynchronous I/O model gives you scalable multicore applications, built as asynchronous message-processing tasks. It has a score of language APIs and runs on most operating systems. ZeroMQ is from iMatix and is LGPLv3 open source.”

For a more detailed introduction, see [Appendix E](#).

Roadster uses ØMQ to carry messages between its processes. The library⁷ it uses to interface with ØMQ is unmaintained and doesn’t support recent versions of ØMQ (namely the ones supporting encryption).

4.2.4 Software architecture

As mentioned earlier, Roadster is event-driven⁸ and built on the Actor model, meaning it exhibits a shared-nothing architecture. Each Roadster node runs a number of Ruby processes which communicate via ØMQ sockets. The key here is communication:

“Don’t communicate by sharing state; share state by communicating.”

Running multiple, loosely coupled processes (actors) allows leveraging the full potential of modern multi-core processors, while avoiding a whole class of traditional concurrency problems.

Every Roadster node runs a group of actors:

CORE:

It is responsible to start the other actors. It also plays a key role in keeping state in all actors synchronized, being the source of truth.

COMM:

A bunch of COMM actors communicate with the outside world of a node. It typically either acts as a client of various kinds of field devices, or as a server to the client. To do so, it utilizes the appropriate adapter which is defined by the static configuration of a Roadster application. These adapters implement communication protocols used to communication with field devices, but also to client systems (e.g. over [OPC UA](#)). The webserver⁹ for the web UI also runs in a COMM actor.

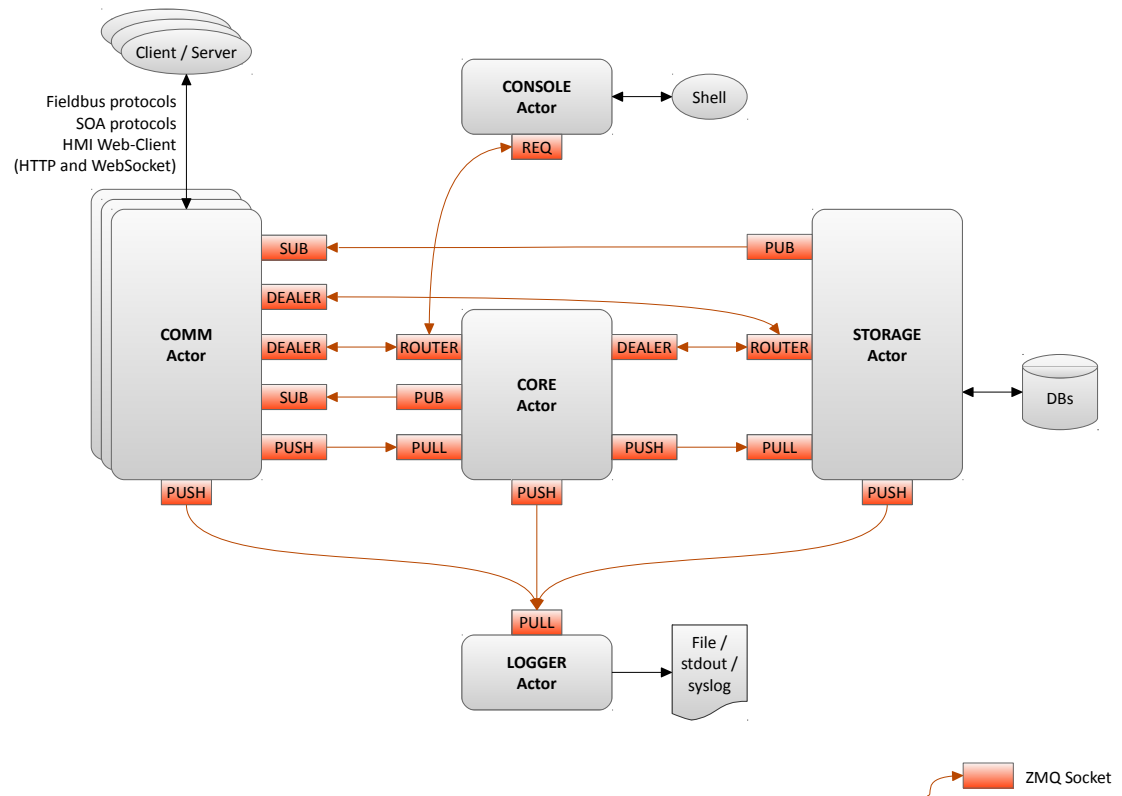
STORAGE:

This actor is used when information needs to be persisted, such as time series or event journals. It’s the interface to a key-value store.

⁷The library is called *ffi-rzmq* and is hosted on <https://github.com/chuckremes/ffi-rzmq>

⁸https://en.wikipedia.org/wiki/Event-driven_programming

⁹The webserver *Thin* is utilized, see <http://code.macournoyer.com/thin/>



Source: Andy Rohr

Figure 4.2: Roadster's software architecture

LOGGER:

This actor collects logging data and sends it to whatever target is configured, be it **STDOUT**, a file, or a syslog server.

Figure 4.2 illustrates Roadster's architecture.

4.2.4.1 Communication Layers

The communication architecture in Roadster consists of three layers, as illustrated in Figure 4.3. The following list briefly explains the layers from top (most abstracted) to bottom:

Engine layer:

Here is the business logic of Roadster, e.g. the **Domain Information Model (DIM)**, user authentication, adapters for different devices, the web **user interface (UI)**, etc.

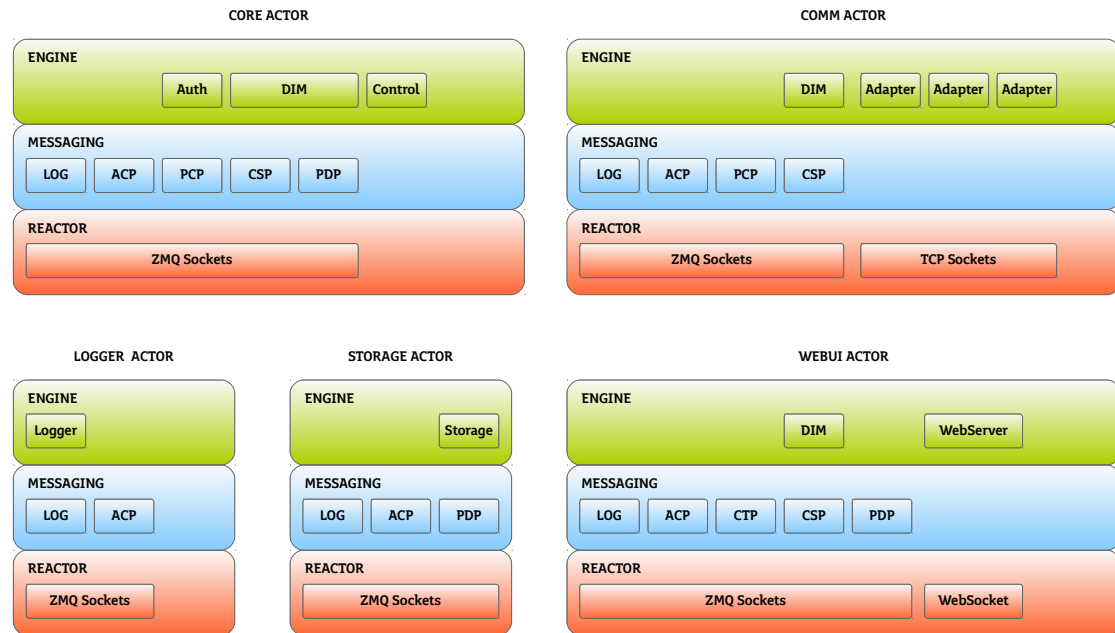
Messaging layer:

The **Roadster Messaging Protocols (RMP)** reside here and implement essential protocols used for logging, state synchronization, commands, application controlling, and storage. They're explained below in subsection 4.2.4.2.

Reactor layer:

This layer forms the base, which is where the **ØMQ** sockets and **WebSockets** are utilized. It is powered by an event-loop¹⁰. Sockets used by **COMM** actors to communicate with various field devices are also integrated into this event-loop.

¹⁰EventMachine is used as a high-performance event-loop to manage large numbers of sockets and timers, <https://github.com/eventmachine/eventmachine>



Source: Andy Rohr

Figure 4.3: Roadster's communication layers

4.2.4.2 RMP

The **RMP** are a collection of protocols implemented and used by Roadster internally. They reside in the messaging communication layer, and include:

Clone State Protocol (CSP):

Used to synchronize state between the actors.

Application Control Protocol (ACP):

Used to control the application state, e.g. shutdown.

Persistent Data Protocol (PDP):

Used when data needs to be persisted.

Supress Management Protocol (SMP):

Used to suppress the generation of certain **cases**, e.g. when a sensor is defect and repeatedly causes cases.

Peer Control Protocol (PCP):

Used for asynchronous command execution via COMM peers with feedback.

LOG protocol:

Used for system logging.

Every actor in Roadster uses a subset of these protocols to perform its job.

Passing messages from actor to actor, which are nothing but serialized Ruby objects, happens in one of two modes:

Fire & Forget:

No guarantee of correct processing, e.g. DIM updates from COMM to CORE. This doesn't mean there are no other mechanisms in place to ensure reliability.

Dialog:

An immediate answer is expected, e.g. when creating a user session. Sending a message like this looks like it's a synchronous call, even though it's handled asynchronously under the hood¹¹ Any protocol can make use of this primitive.

4.2.4.3 DIM

The [Domain Information Model \(DIM\)](#) is a tree data structure that lives inside every actor of a Roadster node. Every actor builds it when starting up by reading the configuration files available to all actors.

The DIM consists of static objects and dynamic objects. The dynamic objects¹² can be updated. The updates are then replicated across all actors to keep the DIM synchronized. This works by marking the updated object dirty¹³ so it is subsequently replicated via the [CSP](#). [Figure 4.4](#) illustrates the class diagram of all the classes whose instances constitute the DIM.

4.2.4.4 Existing CSP in a nutshell

This is a brief introduction/refresher for the Clone State Pattern implemented by Roadster, which is used for the DIM synchronization. Although Roadster actually sends serialized instances of CSP message classes to fulfill this protocol, for better readability the [Zguide](#)'s canonical nomenclature of [Clone Pattern](#) messages will be used.

The existing [Clone State Protocol \(CSP\)](#) is closely related to the [Clone Pattern](#) from the [Zguide](#). Its goal is to keep a state (a list of key-value pairs) in sync across a set of participants. To greatly reduce the complexity, it's not decentralized: There's a server part which serves as the single source of truth.

The server uses a ROUTER, a PULL, and a PUB socket; each client a DEALER, a PUSH, and a SUB socket. The protocol consists of three distinct messages flows:

Snapshots: Requesting and receiving the complete, current snapshot of the state (all key-value pairs). This happens via a ROUTER/DEALER pair of sockets. The request message consists solely of the humorously named ICANHAZ command. The response is the complete set of KVSET messages so a late-joining (or previously disconnected) client can rebuild the current snapshot.

Upstream updates: Updates always originate from clients and are sent to the server via a PUSH/PULL pair of sockets. These are KVSET messages.

Downstream updates: After being applied to the server's copy of the state, updates get a sequence number and are published back to all clients. This happens via the PUB socket and uses KVPUB messages.

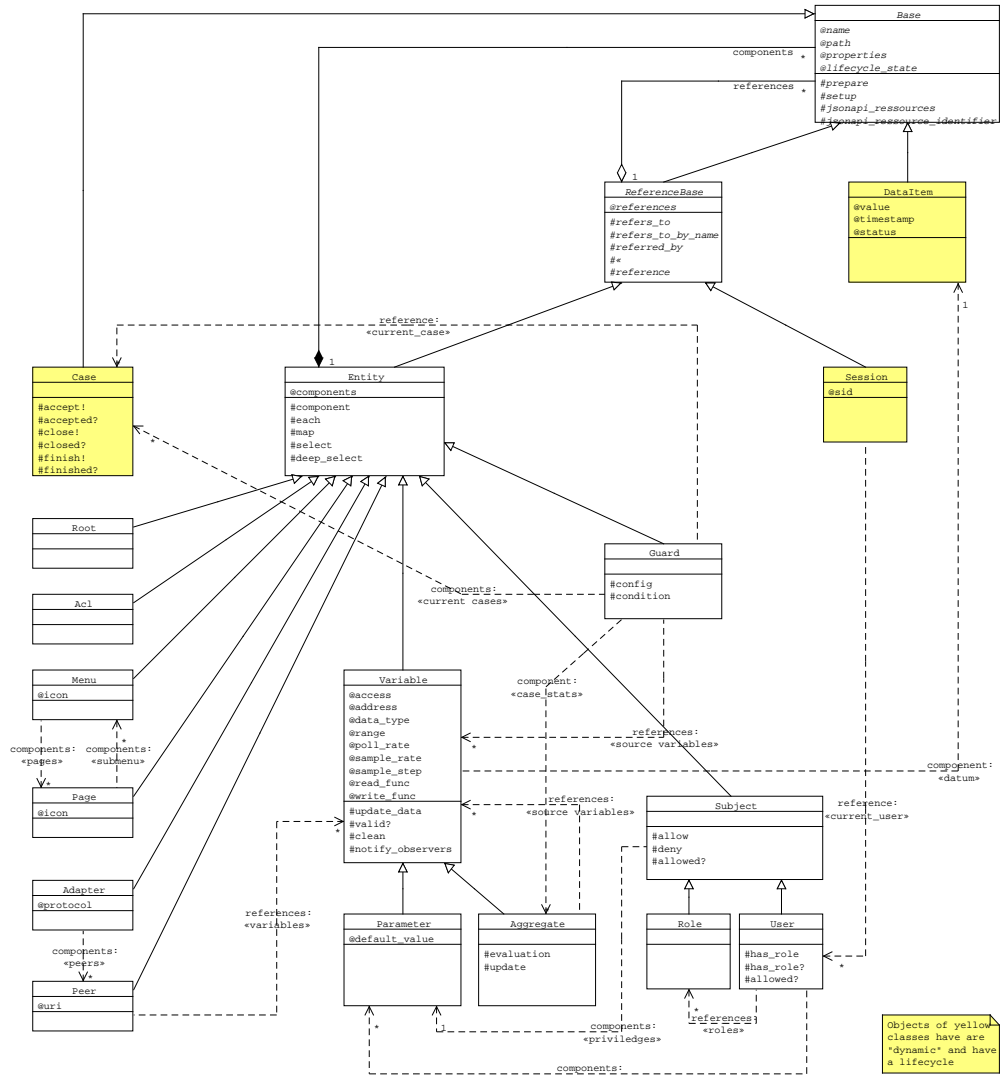
By making all updates go through the server, a total order is enforced, which is crucial to keep the state consistent across all clients.

To avoid risking a gap between requesting the current snapshot and subscribing to updates, a client actually subscribes to the updates first, then gets the snapshot, and then starts reading the updates from the socket (which has been queueing updates in the meantime, if any). Updates that are older or the same age as the received snapshot are skipped, and only successive updates are applied (tested by comparing the sequence numbers).

¹¹This is done by wrapping the affected code in a Ruby `Fiber`, which is similar to a thread but allows for cooperative scheduling as opposed to preemptive.

¹²Namely instances of the meta-model classes `Case`, `DataItem`, and `Session`

¹³It is marked dirty by setting its `@lifecycle_state = "updated"`



Source: Andy Rohr

Figure 4.4: Class diagram for Roadster's meta model used in the DIM

Because message loss via the third message flow (PUB-SUB) is unlikely but theoretically possible, the client checks for gaps in the sequence number of each KVPUB message. If a gap is detected, the current state is discarded and a complete resynchronization happens. This is brutal, but is very simple and thus robust; there's no complexity that would leave room for nasty corner cases.

A feature described in the [Zguide](#), but not implemented in Roadster as of this writing, are subtrees. Keys can be treated hierarchically (e.g. `topic.subtopic.key`) and thus, a client can optionally subscribe to only a particular subtree. This is useful when the number of client grows and not all of the state needs to be on every client. In that case, the topic of interest is sent by the client along with the ICANHAZ message.

4.2.4.5 Persisted data

Certain data on a Roadster node is persisted, which is done by the STORAGE actor. Different data goes into different [TokyoCabinet](#) database files, including:

Event journal:

This is a history of all cases (alarms), including the ones that have been confirmed and thus removed from the DIM. It resides in a [TokyoCabinet](#) *table* database, where the key is a [Universally unique identifier \(UUID\)](#), one of the columns (attributes) is the timestamp of the case, and another one is the actual, serialized [Case](#) object. Objects in this database can be modified, e.g. when a pending case is confirmed.

Time series:

This is a pure key-value store that stores samples of sensor data from field devices. Only the most recent value of these are actually kept in the DIM. One file per series is used. The timestamp is part of the key.

Parameters:

Parameters are the third and last kind of persisted data in Roadster. Parameters are typically changed by a user. Each parameter is backed by a default value from the configuration, which is used as long as there is no actual corresponding value set or read from the database. Credentials for users of the UI are an example parameters.

4.3 Goals

To summarize the mandatory goals from the Task Description in [Appendix B](#):

1. Getting familiar with Roadster
2. Extending the communication protocols to support federation of multiple nodes
3. Extending the communication protocols to allow high availability clusters of two peer nodes

The optional goals are:

1. Encryption of the communication
2. Providing of the highly available [OPC UA](#) server interface

4.3.1 Security concerns of SCADA applications

Secure inter-node communication within a Roadster federation is important to mitigate common security concerns with SCADA systems which are becoming more and more open due to standardization. To quote [\[8, Security issues\]](#) wikipedia:

“In particular, security researchers are concerned about:

- the lack of concern about security and authentication in the design, deployment and operation of some existing SCADA networks
- the belief that SCADA systems have the benefit of security through obscurity through the use of specialized protocols and proprietary interfaces
- the belief that SCADA networks are secure because they are physically secured
- the belief that SCADA networks are secure because they are disconnected from the Internet.”

The goal is to provide a framework that makes it trivial to provide reasonably secure SCADA applications which are assumed to be running on insecure networks.

Chapter 5

Requirements

The requirements gathered during the first and second meeting with the client are explained in this chapter. These are more concrete than the ones listed in the Task Description in [Appendix B](#).

First of all, these are the client's priorities:

1. Federation functionality, namely:
 - (a) Topology definition
 - (b) DIM synchronization
 - (c) Message routing
2. [High availability \(HA\)](#)
3. Persistence synchronization
4. Encryption (optional)
5. DIM access control
6. OPC UA [HA](#) (optional)

The following sections explain the requirements in greater detail, starting with the functional ones. The non-functional requirements are described in [section 5.4](#).

The functional requirements are each described in prose as understood by the students first and then formally in [Gherkin](#) style features. Those feature specifications will later be useful to deduce test cases.

5.1 Federation

Roadster will need to be run on multiple nodes in a hierarchical topology, forming a distributed computing architecture. The root node would then act as the client-facing server. Typical node topologies include:

Single level, single node

This is the legacy setup and is what Roadster is already able to do. It consists of a single node. This is illustrated in [Figure 5.1](#).

Multi level

This is the most basic federation setup. There is a root node, and two subnodes. Each

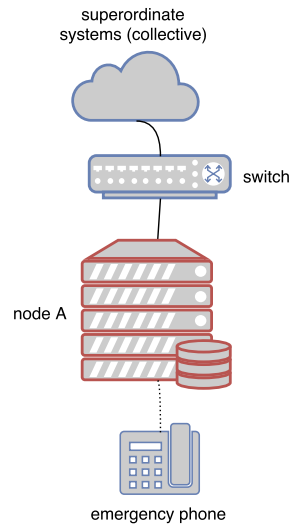


Figure 5.1: Physical legacy example: a single node and a field device each

subnode is directly connected to a number of field devices such as PLCs or emergency phones. This is illustrated in Figure 5.2.

This is formally specified in Listing 5.1

Feature: Federation

```
In order to communicate with the rest of the federation
As a Roadster federation node
I need to be able to deduce who I am, who are the others,
and how can I contact my neighbors
```

Scenario: Topology loading

```
Given a static configuration file defining the federation topology
When a node starts
Then configuration will be loaded
```

Scenario: Role loading

```
Given a static configuration file defining the role for each node
When a node starts
Then the CORE actor starts the appropriate other actors
```

Listing 5.1: Formal federation feature

5.1.1 DIM extension

5.1.1.1 Synchronization

Extending the CSP to keep the DIM in sync across all nodes is a central part of the federation functionality. This means replicating modifications to the dynamic DIM objects to all other actors on the node (legacy behavior) and to all other nodes.

According to the meta model in Figure 4.4, these dynamic objects are exactly the instances of one of the following three classes (marked yellow in the class diagram):

- `DataItem`

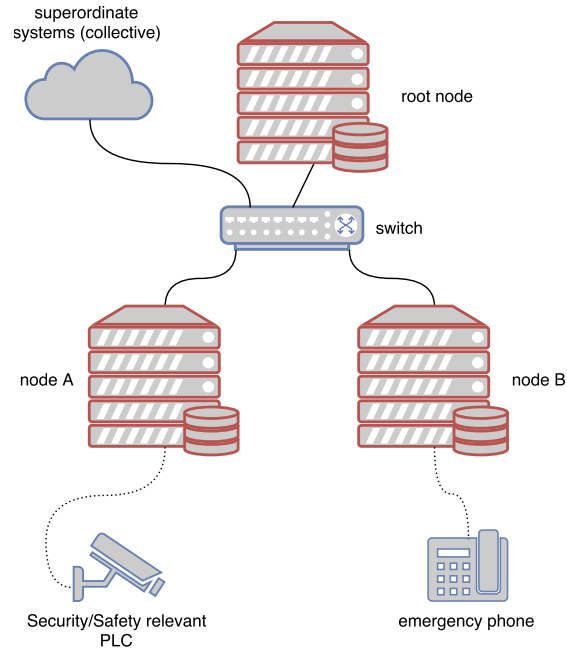


Figure 5.2: Physical federation topology example: supernode, two subnodes, a field device each

- **Session**
- **Case**

As described earlier, instances of these entity classes are marked "dirty" when modified until they are replicated.

5.1.1.2 Access control

The above requirements imply that modifications to the **DIM** can only be done by the owning node. A node must not modify objects owned by other nodes directly. This is to ensure that each node is its own source of truth to all other nodes in the federation. Only the owner node can enforce a single sequence of updates to its part of the DIM, which is necessary to guarantee eventual consistency [9, Chapter 5, Reliable Pub-Sub (Clone Pattern), Republishing Updates from Clients] across all actors of all nodes.

These two aspects to the DIM extension are formally specified in [Listing 5.2](#)

5.1.2 Autonomy

It's important that every node (and its subnodes) can keep up the operation autonomously even if the link to its supernode fails or the supernode itself fails. This means that updates to the **DIM** must be possible even when neighboring nodes (including the supernode) are unavailable. After the recovery from the outage, the **DIM** synchronization shall be reinitiated so all pending updates are shared to all other nodes as per normal operation.

This is formally specified in [Listing 5.3](#)

Feature: DIM extension

In order to keep the DIM synchronized across the federation
As a Roadster federation node
I need to replicate DIM updates across the federation and vice versa

Background:

Given a federation topology configuration available on all nodes
And subnodes S1, S2 have a network link to root node R
And nodes S1, S2 are running

Scenario: Initial DIM synchronization

Given root node R starts
When connections between subnodes and the root node have been established
Then synchronization between the pairs (S1,R), (S2,R) happens bidirectionally
And the DIM is eventually synchronized across all nodes S1, S2, R

Scenario: Continuous DIM synchronization

Given root node R is running
And connections between subnodes and root node are established
When the DIM is updated on node S1
Then the update is eventually replicated across the remaining nodes S2, R

Scenario: Access control

Given that every element in the DIM is owned by exactly one node
When replicating an update to a DIM element originating from its owning node
Then all other nodes verify that the update originates from its owning node
And unverifiable updates are discarded

Listing 5.2: Formal DIM synchronization feature

5.1.3 Message routing

There needs to be a message routing mechanism so a user of one node's web UI can send a command (passed as a message) to another node where it will be executed. An example for this is a forced value in the DIM to ignore the actually measured value reported by a device in case the device is known to be wrong. The common case where the command is issued at a higher level in the node topology is priority. E.g. in a setup with a root node and two subnodes A and B, issuing a command on A for B has a low priority.

This is formally specified in [Listing 5.4](#)

5.2 High availability

Roadster must be able to run in certain high availability setups. Achieving this is done by adding redundant Roadster nodes.

The following additional federation topologies must be supported:

Single level HA

This is when there are exactly two nodes, both of them connected to the same set of field devices. The difference to a non-redundant case is the added backup node, forming a HA cluster. The field devices are thus connected to the HA cluster using two redundant paths. Both HA peers are able to interact with the field devices to perform operation tasks (e.g. reading sensor data, writing down configurations), but only one of them (the active one)

Feature: Autonomy

In order for each node to run independently
As a Roadster federation node
I need to be able to queue DIM updates and replicate them after recovery

Background:

Given a federation topology configuration available on all nodes
And subnodes S1 and S2 have a network link to root node R
And nodes S1, S2, R are running

Scenario: Unreachable neighbor nodes

When link to neighboring node fails or neighboring node crashes
And neighboring node becomes unreachable
Then node continues to perform its assigned tasks
And field devices, if any, are still being monitored
And updates to the DIM are still possible
And data to persist, if any, is still being persisted
And the node's web UI stays accessible and functional

Scenario: DIM synchronization after link failure recovery

Given root node is running
And connection between subnode S1 and root node R are established
And network link between pair (S2, R) is down
When network link between pair (S2, R) recovers
Then synchronization between the pairs (S2,R) happens bidirectionally
And eventually the DIM is synced across all nodes

Listing 5.3: Formal autonomy feature

must do so. This is illustrated in [Figure 5.3](#).

Multi level, HA at root only

There can be multiple hierarchy levels within a Roadster federation, such as two or three (anything else is considered exotic). Introducing redundancy is done at the root level in the form of a HA cluster, consisting of a primary and a backup node. An example of this setup is illustrated in [Figure 5.4](#).

The following cases are more exotic and are outside the scope of this thesis; however, they should be kept in mind so a future extension to support them is feasible:

Multi level, HA at bottom

A single root node at the top and a subordinate HA pair each connected to the same field device.

Multi level, HA in the middle

A single root node at the top, a subordinate HA pair, which in turn has a subordinate node connected to some field device.

At initialization, the two peers must automatically find consensus on which one becomes active first. At any time, only one of the two HA peers must be active (i.e. serving clients), and the other one must stay passive (i.e. ignore client requests and only keep its DIM up-to-date). The passive HA peer shall take over in the event that the currently active peer becomes unavailable. Measures must be taken to avoid the dreaded split-brain syndrome where both HA peers become active.

- Hardware/software failure on the primary node
- Failure of one of the redundant networking paths connecting the subsystem to the two nodes

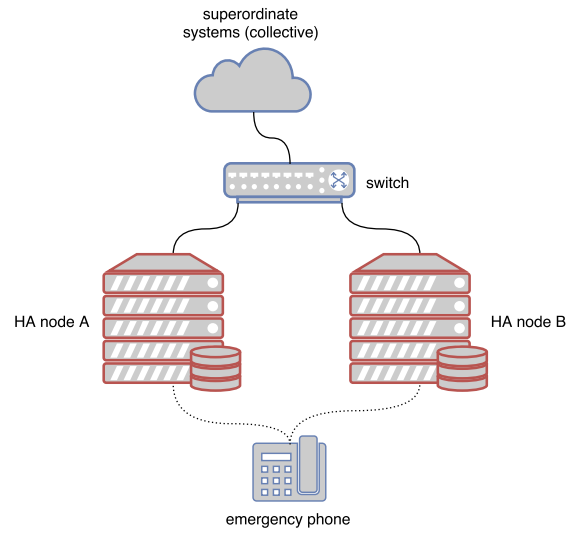


Figure 5.3: Federation example: a HA cluster and redundantly connected field devices

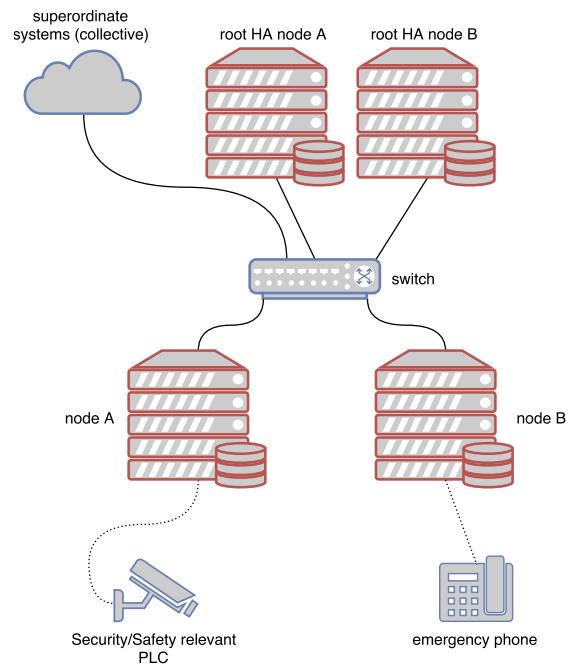


Figure 5.4: Federation example: HA cluster at root, two subnodes, each with field devices

Feature: Message routing

In order to make updates to objects of the DIM owned by other federation nodes, or interact with their field devices
As a Roadster federation node
I need to be able to send messages to other nodes in the federation

Background:

Given a federation topology configuration available on all nodes
And subnode S1 has a network link to root node R
And nodes S1, R are running

Scenario: Sending command from R to S1

Given a user is logged in on the UI of node R
When user changes a setting affecting node S1 directly
Then a command is generated in the UI
And the generated command is routed as a message to node S1
And is executed there

Listing 5.4: Formal message routing feature

The types of failures that need to be handled include:

- Software failure on the primary node, like an application or OS crash
- Hardware failure on the primary node, like a defect power supply
- Failure of a network link, completely disconnecting a HA peer from the federation

All three failure types listed above can collectively be called *crash*, as their effects are the same from the point of view of the whole federation.

The high availability requirement is formally specified in [Listing 5.5](#)

5.3 Persistence synchronization

This is about the synchronization of persisted data, which is currently stored in [TokyoCabinet](#) databases on a Roadster node, one for each kind of data. With the federation functionality, this is still true: Every node will have its own set of key-value stores. Changes to the persisted data must flow from south to north (towards the root node), so the root node can collect and maintain a replication of the persisted data of all nodes within the federation, recursively.

Again, it's important that every node and its subnodes form an autonomous subsystem. So in case the link to its supernode fails, it has to continue working. As soon as the link is repaired, synchronization of the delta (changes to the data) can be initiated.

This is different from [DIM](#) synchronization, as the DIM is shared across all nodes and is a relatively small data structure holding merely the current state. The [TokyoCabinet](#) databases can possibly contain large amounts of data (in the hundreds of megabytes) and are shared only towards the root node (thus "bubbling up").

100% consistency is not an absolute requirement for persistence synchronization. Nor is zero data loss an absolute requirement. However, it is mandatory that updates make it to the root node within 30 seconds.

This is formally specified in [Listing 5.6](#)

Feature: High availability

Background:

Given two peer nodes P1 (primary) and P2 (backup)
And network equipment connecting P1 and P2
And a static configuration file defining the HA cluster

Scenario: Healthy startup

When starting the nodes P1 and P2
Then primary node becomes active
And backup node becomes passive

Scenario: Unhealthy startup

When starting the nodes P1 and P2
And either P1 or P2 crashes or is otherwise unreachable
Then the other node does not simply take over
But generates an alarm

Scenario: Active node crashes

Given a healthily running HA cluster
And a client connected to the active peer
When active node crashes or becomes otherwise unreachable
And client connects to passive node (fallback)
Then the failover is initiated
And the passive node becomes the active node
And an alarm is generated

Scenario: Passive node crashes

Given a healthily running HA cluster
When passive node crashes or becomes otherwise unreachable
Then active node continues to operate normally
And generates an alarm

Listing 5.5: Formal high availability feature

5.4 Non-functional requirements

The following subsections elaborate on the non-functional requirements.

5.4.1 Simplicity

The two reoccurring patterns that surfaced during the requirements gathering meeting were:

1. **KISS** principle. Simplicity is favored, as experience shows that simpler systems are more stable, so complexity should be avoided if not absolutely necessary.
2. No premature optimization since it's the root of all evil.¹

5.4.2 Testing

Regarding testing, the following requirements exist:

¹Quote by Donald Knuth: "Premature optimization is the root of all evil."

Feature: Persistence synchronization

In order to collection and allow aggregation of persisted data

As a Roadster federation node

I need to be able to inform my supernode about changes to persisted data

Background:

Given a federation of multiple levels

Scenario: Initialization persistence synchronization

TODO

Given the current node has a supernode

Scenario: Continuous persistence synchronization

TODO

* up to 30 seconds!

* close to 100% consistency and 0 data loss

* towards root

Scenario: Persistence synchronization after link failure recovery

TODO

Given

When

Then

And

Listing 5.6: Formal persistence synchronization feature

- the student's contributions are verified with unit tests
- use cases shall be integration tested in a close-to-reality setup, either automatically or manually

5.4.3 High availability for OPC UA

The high availability feature shall be design with [OPC UA](#) in mind. It should be easy to adapt it to the various kinds of server redundancy specified in OPC-UA, including the transparent and non-transparent variants.

5.4.4 Encryption

This is optional. Also, this requirement has the lowest priority not because it's insignificant, but because it's easy to enable transport level security on ØMQ sockets later on.

The inter-node communication of a Roadster federation must be secured using encryption. Recent versions of ØMQ offers modern, authenticated encryption, including server and client authentication (the latter is optional). The client favors a solution where every communication partner (a node) authenticates all its communication partners, and vice-versa.

Since the ØMQ binding used in the legacy version is unmaintained and doesn't allow encryption, it has to be exchanged with a more appropriate library.

5.4.5 Coding Guidelines

The coding guidelines desired by the client are basically the ones written down in the popular Ruby style guide [1], with the following differences or special remarks:

- method calls: only use parenthesis when needed, even with arguments (as opposed to ²)
- 2 blank lines before method definition (slightly extending ³)
- YARD API doc, 1 blank comment line before param documentation, one blank comment line before code (ignoring ⁴)
- Ruby 1.9 symbol keys are wanted
 - e.g. `foo: "bar", baz: 42` instead of `:foo => "bar", :baz => 42`, just like ⁵
- align multiple assignments so there's a column of equal signs

²<https://github.com/bbatsov/ruby-style-guide#method-invocation-parens>

³<https://github.com/bbatsov/ruby-style-guide#empty-lines-between-methods>

⁴<https://github.com/bbatsov/ruby-style-guide#rdoc-conventions>

⁵<https://github.com/bbatsov/ruby-style-guide#hash-literals>

Chapter 6

Approach

This chapter describes the approach taken by the students to fulfill the requirements. Possible variants as well as the final choices are explained here.

6.1 Getting familiar with Roadster

The client gave a short introduction into Roadster's code base during the meeting in the first week of this thesis. Although quite overwhelming, the first impression was that the code is clean, makes good use of abstractions and has loosely coupled classes. API documentation is scarce though.

TODO: describe approach of getting more familiar with Roadster, e.g. during prototyping

6.2 Testing

This section describes the test methods we used to check particular methods, the integration of multiple components, as well as the behavior of the whole application. All test results can be found in [chapter 7](#).

6.2.1 Setup

Due to the fact that Roadster's Github repository is private, online [continuous integration \(CI\)](#) services such as *Travis CI*¹ can't be used without payment². Fortunately, Gitlab CI is free and can be installed on one's own infrastructure, such as the [virtual machine \(VM\)](#) provided by HSR, where it was installed and configured. Unit and integration tests are run every time new commits are checked in. This is useful to get informed proactively when something breaks.

¹[urlhttps://travis-ci.com/](https://travis-ci.com/)

²Payment is required after the first 100 builds.

6.2.2 Unit tests

To ensure the correctness of the implementations, unit tests are written using RSpec. 100% coverage of the students' contributions can be achieved by adhering to [test-driven development \(TDD\)](#). Naturally, this also simplifies refactoring the code without risking things breaking silently. Unit tests reside under the `spec` directory of Roadster's code base.

6.2.3 Integration tests

Integration tests verify the interaction between the individual components. To test core features like federation, high availability, and persistence synchronization, integration tests have been written. Multiple nodes can be simulated easily by starting them as different process groups. This is possible since ØMQ completely abstracts the transport away.

In order to test the failover or synchronization functionality, individual processes can simply be killed and restarted later if the scenario defines this.

More details about the integration test scenarios can be found in [chapter 7](#).

6.2.4 Continuous integration

[CI](#) helps us prevent integration problems also known as *integration hell*. Each push to the repository will trigger a CI build using a predefined build script. This will install Roadster's dependencies, an example app built with Roadster, and run finally Roadster's test suites.

6.2.5 System test

System tests are designed to test the application under close-to-reality conditions. To create a realistic environment, Mininet and fake [PLCs](#) are used.

Mininet:

Mininet allows creating virtual networks instantaneously. It relies on Linux cgroups and network namespaces to isolate the [VMs](#), the very same primitives used by Docker. Mininet can also simulate link/connectivity problems between nodes.

Fake:

Fake [PLCs](#) are simple scripts that respond to requests.

All events are stored in log files which are subsequently used for the evaluation of the test result. A Ruby program checks these log files for correctness. All system tests are performed manually after each construction iteration.

6.2.5.1 Test scenarios

The test scenarios primarily test situations encountered in the real world. Out of scientific interest, more test scenarios, which reflect more extreme cases, have been added.

The test scenarios contain configurations for mininet and the particular Roadster nodes. The procedure of each scenario is described in detail below to make the results reproducible.

6.3 Port to new ØMQ library

Porting Roadster to a new ØMQ library early on makes sense for the following reasons:

- to exclude possible failures from faults in the unmaintained `ffi-rmq`³ library
- encryption is needed later anyway, which is not supported by the currently used library
- all other tasks involve ØMQ communication anyway

There is currently only a single Ruby library that is maintained, supports encryption, and freely available, which is [CZTop](#). Technically it's a binding for the [CZMQ](#) library, which is the modern and recommended way of using ØMQ. More info about CZMQ can be found in [Appendix E](#).

As stated in the Task Description already, Roadster's event loop makes use of the ØMQ options `ZMQ_FD` and `ZMQ_EVENTS`. Getters for these had to be added to in CZTop, which was a matter of minutes.

6.3.1 Actual port

Due to Roadster's beautiful software architecture, code that actually made use of the `ffi-rmq` library directly was located in a single file. The following things needed to be done:

- Tell Ruby to load CZTop instead of `ffi-rmq`.
- Remove code to send and receive multi-part messages. This has been simplified in CZMQ and thus is a single method call using CZTop.
- Remove error checking code. CZTop always checks error codes, and raises an appropriate exception if needed.
- Simplify code that reads option values such as `ZMQ_FD` and `ZMQ_EVENTS`.
- Rewrite library calls to use CZTop instead of `ffi-rmq`.

This was about an hour's work.

6.4 Federation

A Roadster federation is illustrated in [Figure 6.1](#). Adding federation functionality to Roadster involves the following aspects:

- node topology DSL
This DSL also has to provide means to define the roles/functionality of each node, e.g. the set of COMM actors running on a particular node
- DIM synchronization
- message routing
- What needs to be done if a WebUI user wants to e.g. change some value on a PLC, possibly on a remote node? Is it completely handled via DIM or do we need message routing?

³<https://github.com/chuckremes/ffi-rmq>

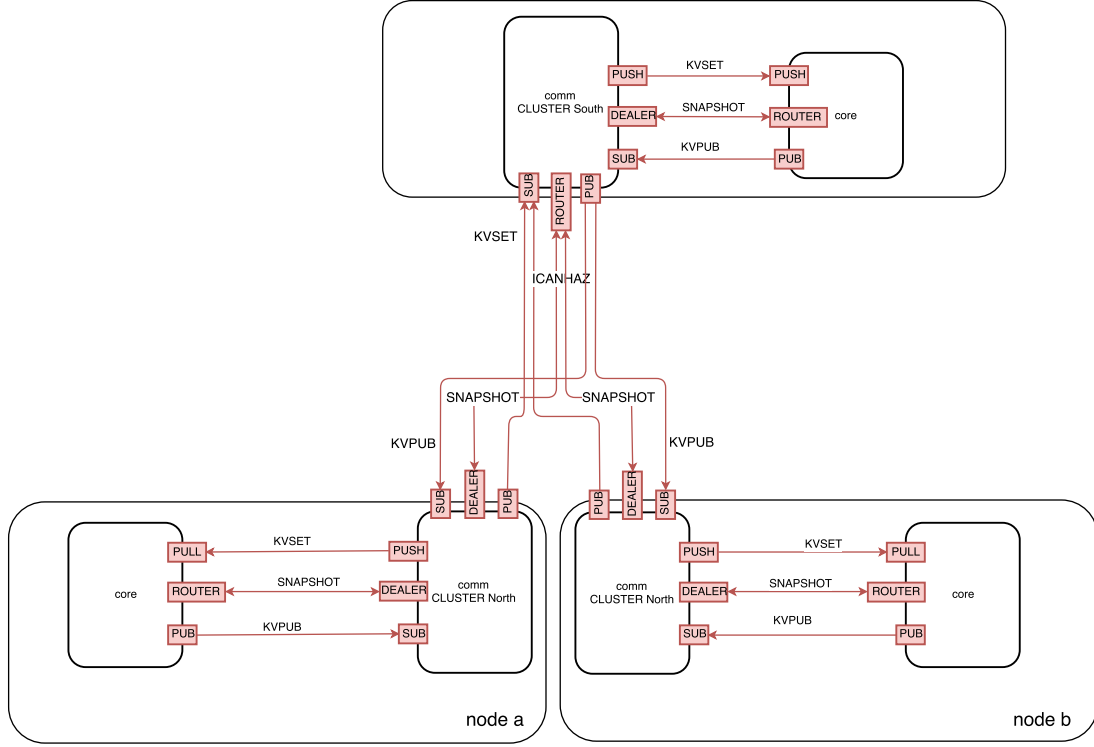


Figure 6.1: Federation between a supernode and two subnodes

6.4.0.1 Fallacies of distributed computing

At this place, it is worth noting the common fallacies encountered in distributed computing, as explained on [7].

6.4.1 DIM synchronization

The following is a list of things that are missing before the requirements can be fulfilled:

- it has to work across several nodes
- it has to be able to handle HA supernodes, which affects DIM synchronization and message routing

There are multiple choices when it comes to what exactly of the DIM should be synchronized:

Variant 1. Sync self-subtree only

Always sync on subtree only, which means a node only knows the DIM part of itself. The big disadvantage is that it won't have a copy of the rest of the DIM, which can be useful to inspect variables on neighboring nodes, especially when they're unreachable.

Variant 2. Sync complete tree

Always sync on complete tree, which means getting the snapshot from the supernode and merge the own subtree into it, replacing whatever subtree is already there. This works because of the autonomy requirement for nodes and their subnodes. This variant is very easy to implement at first.

Variant 3. Either sync on subtree or complete tree

Make it configurable: Either sync on subtree or on complete tree. The topology DSL would

allow to specify this property for each node. This is the best of both worlds, but more effort.

Variant 2 will be the first step. Variant 3 will be the second step, if at all.

Since COMM actors are used to communicate with things outside of a node, new COMM actors will have to be introduced: One kind that is south-facing to communicate with subnodes, and another one that is north-facing for communication with supernodes. They will be named COMM FEDNF and COMM FEDSF.

Their responsibility is the inter-node synchronization of the DIM, similarly to what's happening in the existing CSP within a single node.

To send KVSET updates from one actor to the CORE actor, PUSH-PULL sockets are used. However, similar to the [Clustered Hashmap Protocol \(CHP\)](#) described in the [Zguide](#), this mechanism here needs to be able to handle replication to one or two (in case of HA) supernodes. That means using PUB-SUB messaging for is more appropriate, so all direct supernodes hear the updates.

6.4.1.1 CAP theorem

The CAP theorem [3] states that it is impossible for a distributed computer system to simultaneously provide consistency, availability, and network partition tolerance. In the face of a network partition, one has to chose between availability and consistency. Because subsystems of a Roadster federation must be autonomous, availability is chosen.

Eventual consistency is guaranteed by restricting write access to the owning node, and recovering from a network partition when communication is restored is done by simply reinitiating the DIM synchronization process.

6.4.2 Node topology definition

The federation topology has to be defined somewhere. This can be done using a [Domain Specific Language \(DSL\)](#) and then put into a static file (e.g. `topology_conf.rb`) shared on all nodes of a Roadster federation. Each actor could then read the file at startup, just like it's done for other configuration pieces of a Roadster node. [Listing 6.1](#) shows how such a configuration snippet might look.

To let the actors of a node know which node they belong to, an additoinal line has to be added to the specific configuration file (`conf.rb`), e.g. `conf.system_id = "nodes.root"`. Using that information, the topology created using the DSL can be walked like a tree to find the correct node and important information like its neighbor nodes.

A HA node pair could be one DIM object which has one name but two IP addresses (primary and backup, in that order). Direct subnodes can use that information to connect to the correct supernode during normal operation and also when the primary node is unavailable. The respective DSL snippet is shown in [Listing 6.2](#).

Not every node in a federation has the same role: Some are only connected with other nodes, some directly communicate with field devices. To define different node roles, a syntax as shown in [Listing 6.3](#) is possible.

Listing 6.1: Federation DSL example without HA

```
# * basic method to add a node: #add_node(ID, south_facing_bind_endpoint)
# * it takes a block for defining subnodes

conf.nodes do |map|
  map.add_node("root", "tcp://10.0.0.1:5000") do |map|
    map.add_node("subnode_a", "tcp://10.0.0.10:5000")
    map.add_node("subnode_b", "tcp://10.0.0.11:5000")
  end
end

# subnode_a can infer its endpoints from its position in the tree:
conf.system_id = "nodes.root.subnode_a"
#=> this node is "subnode_a"
#=> its IP address is 10.0.0.10
#=> north facing COMM actor's bind port is 5001
#=> south facing COMM actor's bind port is 5000
#=> north facing COMM actor will connect to "root" node on "tcp://10.0.0.1:5000"
```

6.4.3 Message routing

Messages need to be sent from an actor on one node to an actor on another node. The best place to put this logic is the CORE actor which already does this for messages exchanged within a node. It needs to be extended to know about nodes and their actors, not only actors on the current node. Then messages can be passed around hop-by-hop.

In case a message is sent in *Dialog* mode, this implicates Russian doll routing: At every hop, a new dialog is started which expects an immediate response, which will subsequently be passed back and complete the open dialogs.

6.4.3.1 Example

When a user of the root node's web UI wants to change a value on a field device connected to the root node's subordinate node, a command is sent from the browser to the web UI's COMM actor. From there it's sent via the CORE actor out on the FEDSF actor to the subordinate node. There it's routed via the CORE actor to the correct COMM actor, where the command can actually be executed on the field device.

6.5 High availability

If Roadster is going to be run in a federation, measures need to be taken to mitigate the risk of failure, since many nodes are more likely to fail than a single node (unless they add redundancy). Availability shall be ensured by adding redundancy on certain levels of the node hierarchy (e.g. at the bottom of the topology, right above the PLC, or at the root level), in the form of a fully functional backup node in addition to the primary one.

Run together in a hot-standby cluster, the passive node's responsibility is to take over in case the active one goes down.

Listing 6.2: Fedreation DSL example with HA

```

conf.nodes do |map|
  map.add_ha_pair("root", "tcp://10.0.0.1:5000", "tcp://10.0.0.2:5000") do |map|
    map.add_node("subnode_a", "tcp://10.0.0.10:5000")
    map.add_node("subnode_b", "tcp://10.0.0.11:5000")
  end
end

# subnodeA can infer its endpoints from its position in the tree:
conf.system_id = "nodes.root.subnode_a"
#=> this node is "subnode_a"
#=> its IP address is 10.0.0.10
#=> north facing COMM actor's bind port is 5001
#=> south facing COMM actor's bind port is 5000
#=> north facing COMM actor will connect to "root" HA pair on "tcp://10.0.0.1:5000" OR "tcp://10.0.0.2:5000"

# for primary root:
conf.system_id = "nodes.root[primary]"

```

6.5.1 Defining reliability

When speaking about reliability, it's worth listing the failures we want to be able to handle. According to the requirements, these are exactly:

Hardware or software failure on the primary node: This could be one of the actors crashing, the whole OS crashing, or a fatal disk failure, irrecoverable memory error, or even just someone accidentally pulling the power plug.

Network failure This only includes the failure of the link connecting a HA node to the rest of the federation. Interestingly, this limitation applies to both single level and multi level HA.

Failures that won't be covered include:

Failure of the link between a subnode and one of its supernodes: This can't be handled since the two HA peers would have to continually share the number of subnodes connected to them, and based on that, make a decision on which one should be active or passive. Since the link between them could fail as well, this decision can't be done reliably, which could lead to the dreaded split brain syndrome.

Failure of the link between a HA peer and the field device The [Binary Star Pattern](#) algorithm won't initiate a failover since the active is still alive and is able to tell the passive node so. The missing life signs via the field device could cause an alarm, but no failover, since they're only half of the conditions that have to be met for a failover.

The [Zguide](#) describes a very simple mechanism to achieve this kind of high availability with exactly two redundant nodes: The [Binary Star Pattern](#). It provides a set of clients a highly available service by running two server nodes in a hot-standby setup. It is simple and thus very robust, avoids the split-brain syndrome, and is fairly easy to implement, even as reusable code. The implementation could be contained within a new kind of COMM actor called BSTAR. This makes sense since it talks to the outside world.

6.5.2 Binary Star in a nutshell

Two HA peers are started either as primary or as backup. After an initial handshake, the primary one becomes active, the backup node becomes passive. The two continually exchange heartbeats.

Listing 6.3: Federation DSL example with HA and roles

```
# Idea for node topology definition and assigning roles (features/adapters) to  
# diffent kinds of nodes.
```

```
module Roadster  
  module Domain::Model  
  
    build do  
      nodes do  
        node "root" do # or maybe ha_node or bstar_node  
          endpoint "tcp://10.0.0.1:5000", "tcp://10.0.0.2:5000"  
          label 'BA Roadster App'  
          desc 'Sample application for experimenting and developing the new features within the  
  
          load_conf ::Conf::AccessControl  
          load_conf ::Conf::Objects  
          load_conf ::Conf::Navigation  
  
          node "subnode_a" do  
            endpoint "tcp://10.0.0.1:5000"  
            load_conf ::Conf::Adapters  
            # load_conf ...  
          end  
        end  
      end  
    end  
  
  end # Domain::Model  
end # Roadster
```

Clients always connect to the primary's endpoint first.

The passive node takes over when the following two conditions are met:

1. no life signs from the active node
2. connection requests from clients

The second condition is to prevent the split-brain syndrome and thus can be thought of as an external vote for the node to actually initiate the failover. This works because clients will always try to connect to the primary node's endpoint first, then move on to the backup node's endpoint. This algorithm is explained in [9, Chapter 4 - Reliable Request-Reply Patterns, Client-Side Reliability (Lazy Pirate Pattern)].

6.5.3 Failover

In case the currently active peer goes down, the two conditions will be met. This means that the passive node starts accepting snapshot requests (ICANHAZ messages) and updates the DIM, so every other node will know about the new, active node. This is needed for the message routing to work.

It's important to mention that a dedicated, direct link from one HA node to its peer actually worsens high availability. In case the non-dedicated link from the primary HA node goes down, meaning the HA node is effectively offline and unavailable for subnodes, the failover won't happen since heartbeats are still exchanged with the HA peer over the dedicated link.

6.5.3.1 Alarm generation

When a failover happens, it makes sense to create a **Case** (alarm) in the DIM, so the outage is visible to operational personnel in one of the web UIs. The same applies to the case where the passive node goes down, although it doesn't have an immediate effect on availability. This is so the operational personnel can act upon the alarm and e.g. initiate field forces to inspect the failed node and repair it.

Once repaired, it's restarted with the exact same configuration — either primary or backup. Since there's already an active node (either the primary one, or the backup one), the newly repaired node will become the new passive node.

6.5.3.2 Failover from backup to primary node

Once failed over, the newly active backup node stays active. It does so until it fails itself or is manually stopped. It never automatically switches back to make the primary node the new active one without a failure. This is key. If a node becomes unreachable, the failover happens automatically, but anything else will require human interaction.

Subsequently, when the previously broken, primary node has been repaired, it rejoins the **Binary Star Pattern** cluster as the passive node. At that point, the backup node can be killed if need be, and the primary node will take over again. This works because the **Binary Star Pattern** operates symmetrically after a successful handshake during initialization.

6.5.4 Side benefit: Rolling upgrades

6.5.5 A note on dedicated links

The **Zguide** mentions that a dedicated link between the two HA peers (traditionally done with a crossover cable) is the best solution to prevent the split-brain syndrome. This is true. But in some cases, it could also prevent a failover from happening.

Imagine the currently active peer's other network equipment fails, i.e. its NIC connected to the switch fails, or its switch port fails, then it's unreachable to the rest of the federation. In that case, a failover would be wanted. But it can't happen, since heartbeats are still exchanged with the passive node over the dedicated link. So it's a trade-off between risking the split-brain syndrome and not being able to perform a failover.

6.5.5.1 Extending **Binary Star Pattern**

Of course the **Binary Star Pattern** mechanism could be extended to communicate the number of fully connected clients. That way, a passive node could recognize the situation correctly, since it will get requests from clients which are trying to fallback to it since the active peer is unreachable. Knowing its active peer has zero connected clients, it could actually promote itself to the new active peer.

For this to work, a way of counting fully connected (not just requesting) clients has to be introduced. Since **OMQ** abstracts connection handling completely away, this needs to be done using heartbeats, i.e. from the clients that are currently fully connected (i.e. registered to receive DIM updates). This number, or list of clients, could be communicated privately just between the two HA peers, along with the heartbeats, or it could be published via the DIM.

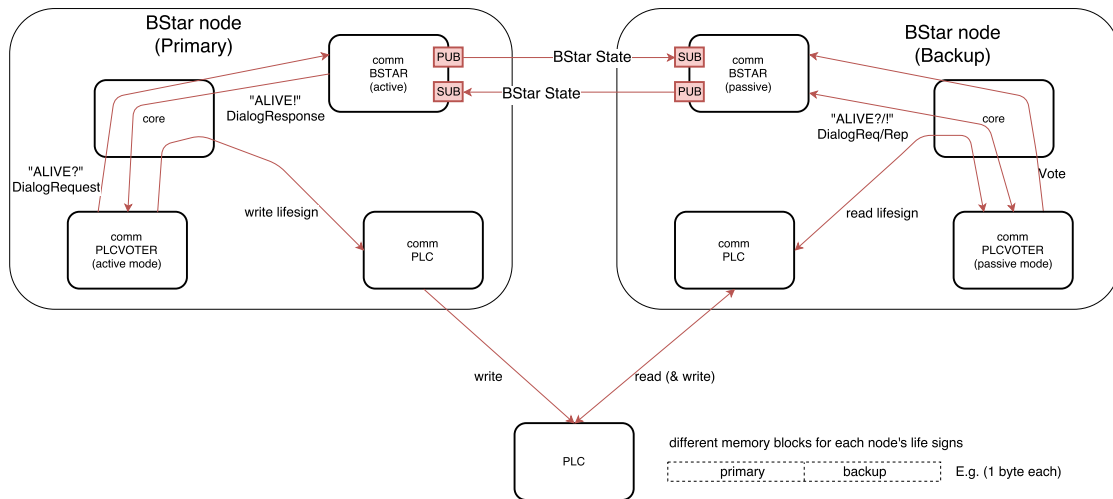


Figure 6.2: Single level HA setup between a HA pair and a field device (PLC)

Another thing that's needed is a HA peer's ability to step down from being the active node as soon as its peer promotes itself to the newly active peer. In the standard [Binary Star Pattern](#) mechanism, this would be recognized as the split-brain syndrome and handled fatally.

6.5.5.2 Dangerous corner case

6.5.6 Single level

This is different from what's described in the [Zguide](#) because the concept of client requests is missing here (field devices don't request anything from Roadster nodes). What can be done instead is periodically sending life signs from one node to the other through the field device by updating some designated memory block. This can actually be done by both the active and the passive node, which reduces code complexity.

The passive node will check the active node's life signs periodically as well. In case the life signs cease, it can give its vote to the COMM BSTAR actor. This would satisfy the second condition of the [Binary Star Pattern](#) for a failover to take place. The first condition would be the missing heartbeats which are normally transmitted through the network link.

TODO new actor: COMM BSTARVOTER (or maybe directly into CORE actor)

6.5.6.1 Caveats

Special attention needs to be paid when it comes to writing these life signs. A naïve developer might implement the COMM BSTARVOTER so it autonomously causes life signs to be written on the field device. This works as long as the failures only affect the hardware. But what if a software error happens in the CORE or BSTAR actor? They'd crash or hang, while the BSTARVOTER happily sends out life signs, which it obviously shouldn't be doing at that moment.

A better implementation would have the BSTARVOTER poll the BSTAR via the CORE router whether it's still alive, and only send out a life sign in case it gets an answer. This way, the BSTAR and the CORE actor are being tested for responsiveness. We'll call the two messages being sent back and forth **"DEAD?"** and **"ALIVE!"**.

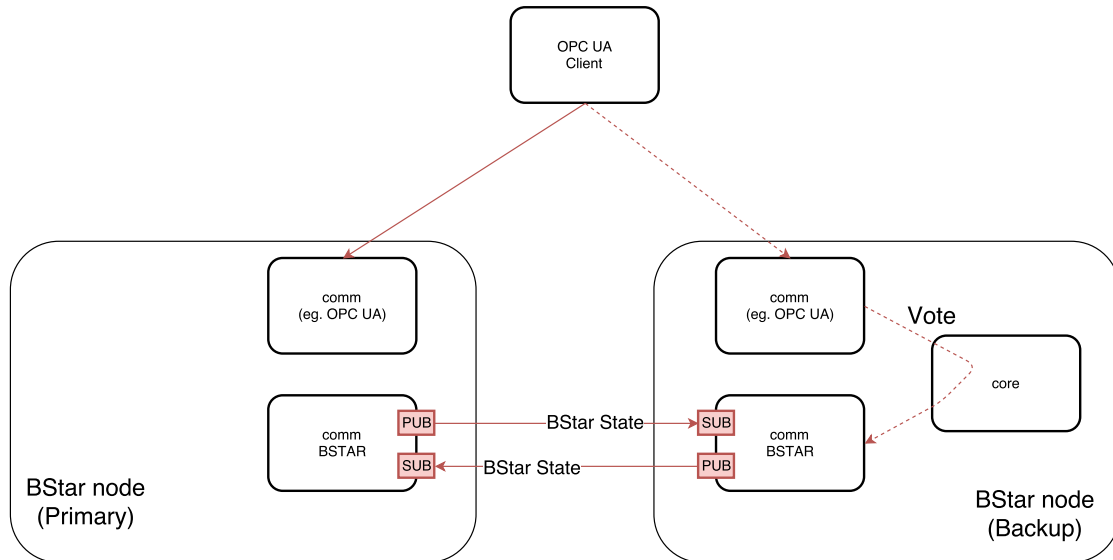


Figure 6.3: Multi level HA setup between a HA pair and a number of client nodes

6.5.6.2 Link failure between Roadster node and field device

TODO describe why this won't cause a failover and thus can't be handled, as mentioned above

6.5.6.3 Supporting different field devices

TODO: adapter

6.5.7 Multi Level

This kind of HA setup is closely related to the [Binary Star Pattern](#) described in [9, Chapter 4 - Reliable Request-Reply Patterns, High-Availability Pair (Binary Star Pattern)]. This means that the passive node would actually receive requests from clients in case the active node fails, which simplifies the implementation. These requests will count as votes to fulfill the second condition that has to be met for a failover to be initiated. This is illustrated in [Figure 6.3](#).

6.6 Persistence synchronization

The persisted data and updates to it, handled by the STORAGE actor, need to bubble up and collected in the root node.

6.6.1 Aspects

There are multiple aspects involved in persistence synchronization:

Initial synchronization: How does one get the initial delta of updates since the last synchronization?

Continuous synchronization: Further updates, one-by-one. This is only needed in case the solution aims for event-driven (meaning close to instantaneous) synchronization.

HA peer sync: How is the passive HA peer updated? This not only matters when the supernode is a HA pair (multi level), but also when it's at the bottom of the node hierarchy (single level).

What about that last case? We need to synchronize east-west. Something similar as with Binary Star, where updates go into a *pending* queue on the passive node until confirmed by the active node?

6.6.2 Variants

There are multiple variants to achieve the needed functionality.

6.6.2.1 Polling only

The supernode just periodically request persistence deltas. This would be handled over a DEALER/ROUTER pair of sockets. The nice thing about this variant is that the subnode only has to do one thing, which is responding to requests from the supernode(s); it doesn't have to proactively send any updates after sending the an initial delta.

A big drawback is that the synchronization only happens periodically. This doesn't seem to fit well into the overall Roadster architecture, which is completely event-driven (no polls or "sleeps").

Another drawback is efficiency. This variant will periodically cause the subnode's database to be searched for all keys. Depending on the size of the database and the efficiency of searching through keys, this could be a lot of wasted resources or even cause bottle necks when interacting with the STOR actor.

In case the supernode is a HA pair, this variant would generate duplicated traffic. To avoid this, another pair of sockets has to be introduced to synchronize persistence between a HA pair. This also means designing another protocol, and more moving parts overall.

Overall, this variant is very simple, but doesn't offer some features we'd normally expect from a framework like Roadster. The fruits are hanging low; achieving event-driven synchronization and better efficiency is easy.

6.6.2.2 PUSH-PULL

This variant avoids the delays introduced by the polling mechanism of the first variant.

Procedure (for each subnode):

1. via a ROUTER/DEALER socket pair:
 - (a) supernode tells subnode its most recent timestamp in an ICANHAZ request
 - (b) subnode sends delta
 - (c) supernode receives and processes the complete delta
2. subnode sends updates to supernode via PUSH-PULL
3. during low-traffic times, we can send HUGZ as heartbeats

This seems nice at first, but the PUSH socket's send buffer will fill up when the connection is interrupted. This isn't bad in and of itself, because when it's full (and writes start to block), we can just destroy the socket and reinitialize and start syncing anew (from ICANHAZ) after a certain timeout. But the problem is that, in case the delta is large, it will inevitably fill the PUSH socket's send buffer, temporarily reaching its high water mark, which is part of its normal operation.

So we'd have to introduce logic to recognize whether the PUSH socket is just temporarily full (e.g. during delta transmission), or permanently full (e.g. the supernode or the link to it is down).

Another disadvantage is that there needs to be another channel to synchronize persistence updates to the other HA peer, if there is one. This means another pair of sockets, another protocol to be designed, and more moving parts overall.

6.6.2.3 PUB/SUB

This is similar to [CSP/CHP](#). It's not 100% reliable, but even with unstable links, no data loss will occur if the client (the supernode) is able to reconnect within a specific amount of time. ØMQ's default for that amount is 10 seconds. As the requirements specify, 100% consistency is not mandatory for the persistent data.

A possible drawback is that the traffic is duplicated in case the supernode is a HA pair. However, there are numerous opportunities to mitigate this.

Procedure (for each subnode):

1. supernode subscribes to updates from subnode
2. via a ROUTER/DEALER socket pair:
 - (a) supernode tells subnode its most recent timestamp in an ICANHAZ request
 - (b) subnode sends delta
 - (c) supernode receives and processes the complete delta
3. supernode starts reading updates, possibly skipping the first few (based on timestamp)

6.6.3 Chosen Variant

We'll most likely go with the PUB-SUB variant, since it's simple and is similar to what's used for the new [CSP](#) in conjunction with multi-node [HA](#). It provides the best opportunities to improve efficiency later on.

Its possible performance issues can be ignored right now, as trying to fix them is arguably considered premature optimization. If this turns out to be an issue in a productive deployment, like over a cellular network link, a future version can switch to multicast. ØMQ supports PGM, which is a reliable multicast protocol. (Pragmatic General Multicast, standardized, directly on top of IP, requires access to raw sockets and thus may require additional privileges) and EPGM (Encapsulated Pragmatic General Multicast, encapsulated in a series of UDP datagrams, doesn't require additional privileges, useful in a ØMQ-only setup).

If its reliability turn out to be an issue, one the socket option `ZMQ_RECOVERY_IVL` can be increased from 10 seconds to, say, 60 seconds, which gives an unstable link more time to recover before any data loss happens.

TODO: describe reasonable default setting, in case we change ZMQ's default.

6.7 Encryption

It's fairly easy to enable transport security to secure the communication between ØMQ sockets over an unsecure network. The simplest variant, as described in [section E.1](#), is to do server authentication only and allow any client to connect.

Because ØMQ allows any order of the bind and connect actions to take place, and doesn't impose a specific action on each socket type, the meanings of "server" and "client" become blurry. Still, with respect to the CURVE security mechanism in ØMQ, one of the two involved sockets needs to be designated as the CURVE server, and the other one as a CURVE client⁴. Based on the autonomy aspect of the federation, meaning the lower-level systems act as servers for the higher-level systems, the same schema is applied to the CURVE.

Since client side authentication is wanted in addition to the server authentication always performed, both sides need to be in possession of the other side's public key. Due to the nature of public keys, they can be distributed conveniently and safely through the federation topology configuration file. Curve25519 public keys are only 40 characters long in [Z85 armor](#) notation.

Of course the public keys and the private keys are generated in pairs. However, distributing the private keys is less convenient, since they must be kept private. This can be done via SSH.

6.7.1 Key generation and distribution procedure

The following procedure can be followed to generate the required keys in advance and make them available through the shared configuration file:

1. for all nodes that will have sockets that will act as CURVE servers (i.e. all non-leaf nodes)
 - (a) generate key pair and save it on the respective node
 - (b) for each direct subnode
 - i. put a copy of the public key (in Z85 notation) into that node's block within the shared topology configuration file

This procedure could be implemented as a script that generates the keys and distributes them via [Secure Shell \(SSH\)](#).

6.7.1.1 Client authentication

In case client authentication is actually desired, the clients' keys also have to be generated in advance, and the public key files for each subnode of a supernode in question have to be stored into a directory. That directory can then be specified to the `CZTop::Authenticator` actor.

6.7.2 In code

[Listing 6.4](#) shows how to start an authentication handler which allows any client to connect, effectively only allowing authenticated encryption, but not a true client authentication. What it actually does is start an in-process thread to which can be communicated via a PAIR-PAIR

⁴This is a simple method call on each socket where the involved keys are passed as well.

socket pair (it's a CZMQ actor⁵). The thread calls a function⁶ (passed as a function pointer at actor creation) provided by CZMQ, which understands the [ZMQ Authentication Protocol \(ZAP\)](http://api.zeromq.org/czm3-0:zap), different security mechanisms (NULL, PLAIN, and CURVE) and is capable of reading client public keys from a directory on the file system and even listens to changes in that directory.

Listing 6.4: Starting an authentication handler that allows any clients

```
##
# on the supernode:

authenticator = CZTop::Authenticator.new
authenticator.verbose!
authenticator.curve # use CURVE mechanism, but allow any

##
# start the server sockets
# ...
server_cert = CZTop::Certificate.load("/path/to/private_key")
pub = CZTop::Socket::PUB.new
pub.CURVE_server!(server_cert)
pub.bind("tcp://*:1234")

#####

##
# on the subnode:

server_cert = CZTop::Certificate.load("/path/to/server_public_key")
client_cert = begin
  CZTop::Certificate.load("/path/to/client_private_key")
rescue Errno::ENOENT # file doesn't exist
  CZTop::Certificate.new.save("/path/to/client_private_key") # generate
  retry
end

sub = CZTop::Socket::SUB.new
sub.CURVE_client!(client_cert, server_cert)
sub.bind("tcp://*:1234")
```

6.8 OPC UA Interface: High availability

The OPC UA standard seems pretty complicated. And given that the requirement isn't concrete yet, no possible solution has been worked out yet.

In case it's non-transparent server redundancy, [5, 6.4.2.4 Non-transparent Redundancy, p. 96] describes the exact behavior.

⁵CZMQ provides a very simple actor framework based on threads communicating over ØMQ sockets, <http://api.zeromq.org/czm3-0:zactor>

⁶The function is `zauth()` and is described along with its capabilities here: <http://api.zeromq.org/czm3-0:zauth>

Chapter 7

Results

7.1 Port

7.2 Federation

7.3 High Availability

7.4 Persistence synchronization

7.5 Encryption

7.6 OPC UA Interface: High availability

Chapter 8

Discussion

8.1 Value Added

8.2 Limitations

8.3 Business Benefits

8.4 Ideas for Improvement

- HA within a node: kill and respawn an actor when it's unresponsive
- switch to Moneta for a unified key-value store interface, then eventually away from [TokyoCabinet](#) to something more modern and maintained, like LMDB (it's super fast and crash-proof)
- TIPC: high performance cluster communication protocol, suitable because Roadster nodes are Linux and there are direct links to peers (required for TIPC)
- client authentication
- key management in a DB (instead of files), with GUI to accept new clients
- dynamic node topology (maybe via DSL-file in Etcd, or DIM-only, or Zookeeper)
- other method for data serialization (like MessagePack), would allow adding other programming languages to the cluster
- fast compression for messages, like LZ4 or Snappy
- SERVER/CLIENT sockets from ZMQ 4.2 for simplified message routing

Chapter 9

Conclusion

TODO write conclusion, overall experience and opinion of product
TODO they should teach the actor model in APF, because ...

Bibliography

- [1] B. Batsov. *Ruby Style Guide*. URL: <https://github.com/bbatsov/ruby-style-guide> (visited on 10/06/2016).
- [2] Daniel J. Bernstein, Niels Duif, Tanja Lange, Peter Schwabe, and Bo-Yin Yang. *High-speed high-security signatures*. URL: <http://ed25519.cr.yp.to/ed25519-20110926.pdf> (visited on 10/12/2016).
- [3] *CAP theorem*. URL: https://en.wikipedia.org/wiki/CAP_theorem (visited on 10/10/2016).
- [4] A. Dworak, F. Ehm, P. Charrue, and W. Sliwinski. „The new CERN Controls Middleware“. In: *Journal of Physics: Conference Series* 396.012017 (2012). URL: <http://iopscience.iop.org/article/10.1088/1742-6596/396/1/012017/pdf> (visited on 10/06/2016).
- [5] OPC Foundation. *OPC Unified Architecture*. Part 4: Services. July 2015. URL: <https://opcfoundation.org/developer-tools/specifications-unified-architecture/part-4-services/> (visited on 10/06/2016).
- [6] Pieter Hintjens. *CurveZMQ*. URL: <https://rfc.zeromq.org/spec:26/CURVEZMQ/> (visited on 10/12/2016).
- [7] Arnon Rotem-Gal-Oz. *Fallacies of Distributed Computing Explained*. URL: <http://www.rgoarchitects.com/Files/fallacies.pdf> (visited on 10/10/2016).
- [8] *SCADA*. URL: <https://en.wikipedia.org/wiki/SCADA> (visited on 10/17/2016).
- [9] *Zguide*. URL: <http://zguide.zeromq.org/page:all> (visited on 10/11/2016).

Glossary

ØMQ High-performance socket library and concurrency framework for advanced messaging. [7](#), [8](#)

ACP Application Control Protocol. [13](#)

Actor Model A mathematical model for concurrent computation where there's no shared state and all communication between actors happens through messages. [7](#), [8](#), [63](#)

AR Abschnittsrechner. [10](#)

AS Anlagesystem. [10](#)

ASTRA Bundesamt für Strassen. [9](#)

Binary Star Pattern A fairly simple hot-standby and failover mechanism to achieve high availability between two servers, described as a reliable request-reply pattern in the [Zguide](#). [34](#), [36–38](#)

BSD Berkeley Software Distribution. [63](#)

C A compiled, imperative, very influential low-level programming language, invented in the early 1970s as a Unix system programming language. Compared to other languages, it very simple, knows only a handful of primitives and keywords. [7](#)

case An alarm in a Roadster application that needs to be confirmed. [13](#)

CHP Clustered Hashmap Protocol. [32](#), [40](#)

CI continuous integration. [28](#), [29](#)

Clone Pattern A client-server protocol to share state (a list of key-value pairs) across multiple clients, described as a reliable pub-sub pattern in the [Zguide](#). [14](#)

CSP Clone State Protocol. [13](#), [14](#), [19](#), [40](#)

CZMQ A thin abstraction layer ([wrapper façade](#)) for [ØMQ](#) with some additional functionality, written in clean and elegant C. [7](#), [30](#), [64](#)

CZTop A modern, [Foreign Function Interface \(FFI\)](#) based [Ruby](#) binding for [CZMQ](#), written by Patrik Wenger. [7](#), [8](#), [30](#)

DIM Domain Information Model. [12](#), [14](#), [19](#), [20](#), [24](#)

distributed denial-of-service attack is an attempt to interrupt the availability of a service by flooding it with forged requests using a large number of source systems. [63](#)

DSL Domain Specific Language. [32](#)

ECC elliptic curve cryptography. [64](#)

FEDRO Federal Roads Office. [9](#)

Gherkin A simple language to specify feature specifications in steps such as Given, When, Then¹. 18

HA high availability. 18, 21, 22, 38, 40

IEC 60870-5-104 A [International Electrotechnical Commission \(IEC\)](#) transmission protocol used by SCADA applications in power system automation that enables communication via standard networks. 10

IoT Internet of Things. 10

ISA-95 An international standard² for developing an automated interface between enterprise and control systems. 8

KISS The design principle “Keep it simple, stupid”, which favors simplicity over complexity. 25

libsodium A portable and installable variant of [NaCl](#)³. 7, 63

LOG protocol Used within Roadster for system logging. 13

LR Leitrechner. 10

LTA Leittechnikanlage. 10

Modbus TCP The TCP-based variant of Modbus, a *de facto* standard serial communication protocol used to connect electronic devices. 10

MOM Message Oriented Middleware. 63

NaCl Networking and Cryptography Library⁴. Modern, state-of-the-art cryptography library, created by the Daniel J. Bernstein. 7

OPC UA [Open Platform Communications \(OPC\)](#) glsUA: A set of modern standards for industrial control systems, based on cross platform webservice and other modern technology. 9–11, 16, 26

PCP Peer Control Protocol. 13

PDP Persistent Data Protocol. 13

PGM Pragmatic General Multicast. 63

PLC Programmable Logic Controller. 9, 10, 18, 29

RAID redundant array of independent disks. 11

RMP Roadster Messaging Protocols. 12, 13

Ruby A modern and expressive scripting language from Japan. 7, 8

RUP Rational Unified Process. 2

SCADA Supervisory Control and Data Acquisition. 8

SMP Supress Management Protocol. 13

SOAP Service Oriented Application Protocol. 9

SSD Solid State Disk. 11

SSH Secure Shell. 41

STDOUT the standard output channel of a Unix process (file descriptor 1). 12

¹<https://cucumber.io/docs/reference>

²<https://en.wikipedia.org/wiki/ANSI/ISA-95>

³<https://libsodium.org>

⁴<https://nacl.cr.yp.to>

TCP Transmission Control Protocol. [63](#)

TDD test-driven development. [29](#)

TIPC Transparent Inter-Process Communication. [63](#)

TokyoCabinet a library to manage a key-value store in a single file (no server involved). [16](#), [24](#), [44](#)

TweetNaCl A compact, portable reimplementation⁵ of the NaCl in the form of 100 tweets, suited to be included it into one's trusted code base (as opposed to an external dependency). Implemented Daniel J. Bernstein et al. [63](#)

UI user interface. [12](#)

Unix Domain Sockets Named pipes for extremely performant, duplex inter-process communication on Unix systems. [63](#)

UUID Universally unique identifier. [16](#)

VM virtual machine. [28](#), [29](#)

WebSocket A protocol for full-duplex communication between web browsers and web servers, standardized by the [Internet Engineering Task Force \(IETF\)](#) as [Request for Comments \(RFC\)](#) 6455 in 2011. [12](#)

Z85 armor a space efficient, [American Standard Code for Information Interchange \(ASCII\)](#) based, string-safe variant of the Base85 binary-to-text encoding. [41](#)

ZAP ZMQ Authentication Protocol. [42](#), [64](#)

Zguide An extensive online document⁶ describing best-practice patterns for [ØMQ](#). [11](#), [14](#), [16](#), [32](#), [34](#), [36](#), [37](#), [63](#)

⁵<http://tweetnacl.cr.yp.to>

⁶<http://zguide.zeromq.org/page:all>

Part III

Appendix

Appendix A

Self Reflection

TODO how did we perform, completion of goals, accuracy of estimated efforts, efficiency, resourcefulness

Appendix B

Task Description

The following five pages are the original task description, signed by Prof. Dr. F. Mehta.

Bachelor Thesis: Extending a SCADA framework to support high availability

1 Client and Supervisor

Client: mindclue GmbH

Client Contact: Andy Rohr, andy.rohr@mindclue.ch

Supervisor: Prof. Dr. Farhad Mehta, HSR Rapperswil

2 Students

- Patrik Wenger, pwenger@hsr.ch
- Manuel Schuler, mschuler@hsr.ch

3 Setting

The company mindclue GmbH, located in Ziegelbrücke, develops SCADA¹ applications for controlling systems used in traffic systems, energy, and water supply. For that purpose, the company developed the *Roadster* framework, which provides the basis for project specific applications.

Roadster is implemented in Ruby and is architecturally based on the Actor model [1], which means that multiple parallel running, single-threaded processes (actors) are coupled via messaging (shared-nothing architecture). The messaging layer is based on ZeroMQ (ZMQ [2]) and has an asynchronous/non-blocking nature. Several different messaging patterns/messaging protocols are used. Additionally, the system includes a web UI which is based on ember.js and connected to the messaging via WebSocket. Fundamentally, the system follows the Reactive Manifesto [3].

4 Goals

The main aim of this thesis is to extend the *Roadster* framework to support high availability.

Roadster currently lacks the following features:

1. A *Roadster* application is currently limited to one node (one instance). The goal is to be able to build systems which consist of multiple nodes. Example: A master node forms together with multiple subordinate nodes a system (basically a distributed system). The subordinate nodes are responsible for their respective subtask of a facility and communicate with their components (e.g. PLCs). The subsystems are integrated into the master node to form an overall view of the facility, which is visualized in the web UI.

This requirement implies:

¹Supervisory Control And Data Acquisition, see <https://en.wikipedia.org/wiki/SCADA>

- Extension of the messaging protocols to allow the communication between nodes across levels in the hierarchy.
 - Encryption of the communication.
2. A *Roadster* application has to have the ability to be run as a highly available active/passive cluster. Two nodes (primary and backup) at the same level in the hierarchy form a hot-standby cluster, where the two nodes stay in constant connection with each other. In case the active node fails, the passive node immediately takes over and becomes the new active node.

This requirement implies:

- Extension of the messaging protocols to allow the communication between nodes within the same level in the hierarchy.
 - Implementation of resilient failover mechanisms.
 - Encryption of the communication.
3. With (2), it is possible to implement a highly available OPC UA server. OPC UA [4] includes a concept for redundant UA servers. *Roadster* already implements a OPC UA server, although not highly available.

This requirement implies:

- Extension of the OPC UA implementation to support OPC UA HA mechanisms.

The client essentially wants *Roadster* to be extended by the three features described above, whereas the **third one is optional** and is only to be approached in case there is time for it. The **same applies to the encrypted communication requirement**.

5 Tasks

Here is an overview of the currently planned tasks that need to be performed:

1. Getting familiar with the concepts and implementation of *Roadster*, particularly the messaging layer. For that, Andy Rohr (mindclue GmbH) will provide an extensive introduction.
2. Elaboration of a subnode concept. This includes the design, implementation, and testing of extensions of the existing messaging protocols for the communication between nodes of different levels in the hierarchy.

One of the most important *Roadster* messaging protocols is called *Clone State Protocol* and is based on the *Clone Pattern* described in the zguide [5]. It provides means to replicate the current state of the domain model into the different actors within an application, as those actors behave according to the following principle [6]:

“Don’t communicate by sharing state; share state by communicating.”

For the communication between nodes, the protocol has to be extended accordingly. The messages are basically Ruby objects serialized using `Marshal.dump` and transported over ZMQ sockets.

3. Elaboration of a HA concept. This includes the design, implementation, and testing of extensions of the existing messaging protocols for the communication between nodes within the same level in the hierarchy, including resilient failover mechanisms.

The *Binary Star Pattern* [7][8] forms the basis of the HA concept. However, the concept will have to be adapted to fit *Roadster*'s needs. Availability has to be ensured under the following scenarios:

- hardware or software failure of the primary node
- network failure

A more detailed definition will have to be worked out during the elaboration phase of the thesis.

4. Implementation of encrypted communication. The current implementation is based on ZMQ 3 and the Ruby binding ffi-rzmq [9]. However, encryption has been introduced in ZMQ 4, which isn't supported by ffi-rzmq. On top of that, ffi-rzmq is not being maintained anymore. A possible solution is CZTop [10], which is based on CZMQ [11] and authored by Patrik Wenger.

In case CZTop is used, it would have to be extended to allow the watching of ZMQ sockets by EventMachine [12], e.g. `EM.watch(socket_file_descriptor)`. *Roadster* uses EventMachine as a reactor implementation [13].

5. Extensions of the *Roadster* OPC UA server implementation to support HA mechanisms. This part of *Roadster* is a Ruby extension, which is implemented based on the Unified Automation C++ SDK [14]. The extension is written in C++ and uses rbplusplus [15].

6 License

To grant mindclue GmbH unrestricted usage of the student's contributions, the student's code changes and additions shall be protected under the ISC License [16], which is functionally equivalent to the MIT license and the Simplified BSD license, but uses simpler language.

7 Guidelines

The students and the supervisor will plan weekly meetings to check and discuss progress. The student will schedule meetings with the client as and when required (recommendation: 1 meeting per week of 1 hour duration).

All meetings are to be prepared by the students with an agenda. The agenda will be sent at least 24h prior to the meeting. The results will be documented in meeting minutes that will be sent to the supervisor.

A project plan must be developed at the beginning of the thesis to promote continuous and visible work progress. For every milestone defined in the project plan, the temporary versions of all artefacts need to be submitted. The students will receive a provisional feedback for the submitted milestone results. The definitive grading is however only based on the final results of the formally submitted report.

8 Documentation

The project must be documented according to the regulations of the Computer Science Department at HSR [17]. All required documents are to be listed in the project plan. All documents must be continuously updated, and should document the project results in a consistent form upon final submission. All documentation and work artefacts have to be completely submitted

in three copies on CD/DVD (one copy each for the client, university, and supervisor). Three printed copies of the report need to be submitted (one copy each for the client, external examiner, and supervisor).

9 Important Dates

Please refer to <https://www.hsr.ch/Termine-Diplom-Bachelor-und.5142.0.html>.

10 Workload

A successful Bachelor thesis project results in 12 ECTS credit points per student. One ECTS point corresponds to a work effort of 30 hours. All time spent on the project must be recorded and documented.

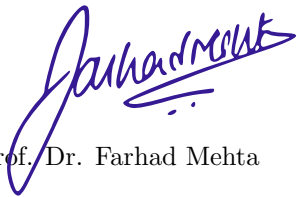
11 Grading

The HSR supervisor is responsible for grading the master thesis. The following table gives an overview of the weights used for grading.

Facet	Weight
1. Organisation, Execution	1/6
2. Report	1/6
3. Content	3/6
4. Final Presentation & Examination	1/6

The effective regulations of the HSR and Department of Computer Science [18] apply.

Rapperswil, Wednesday 28th September, 2016



Prof. Dr. Farhad Mehta

References

- [1] URL: https://en.wikipedia.org/wiki/Actor_model.
- [2] URL: <http://zeromq.org/>.
- [3] URL: <http://www.reactivemanifesto.org/>.
- [4] URL: <https://opcfoundation.org/about/opc-technologies/opc-ua/>.
- [5] URL: <http://zguide.zeromq.org/page:all#Reliable-Pub-Sub-Clone-Pattern>.
- [6] URL: <https://www.igvita.com/2010/12/02/concurrency-with-actors-goroutines-ruby/>.
- [7] URL: <http://zguide.zeromq.org/page:all#High-Availability-Pair-Binary-Star-Pattern>.
- [8] URL: <http://zguide.zeromq.org/page:all#Adding-the-Binary-Star-Pattern-for-Reliability>.
- [9] URL: <https://github.com/chuckremes/ffi-rmq>.
- [10] URL: <https://github.com/paddor/cztop>.
- [11] URL: <http://czmq.zeromq.org>.
- [12] URL: <http://www.rubydoc.info/gems/eventmachine>.
- [13] URL: https://en.wikipedia.org/wiki/Reactor_pattern.
- [14] URL: <https://www.unified-automation.com/products/server-sdk/c-ua-server-sdk.html>.
- [15] URL: <https://github.com/jasonroelofs/rbplusplus>.
- [16] URL: https://en.wikipedia.org/wiki/ISC_license.
- [17] URL: <https://www.hsr.ch/Allgemeine-Infos-Bachelor-und.4418.0.html>.
- [18] URL: <https://www.hsr.ch/Ablaeufe-und-Regelungen-Studie.7479.0.html>.

Appendix C

License

As stated in the task description, all of our code contributions underlie the ISC license, which is functionally equivalent to the MIT license and the Simplified BSD license, but uses simpler language. In addition to that, we hereby explicitly grant mindclue GmbH unrestricted usage of all our code contributions.

Appendix D

Project Plan

D.1 Timetable

D.1.1 Estimated time

The projekt span is from 19.09 until 23.12. We expect 26 hours work per week which is a total of 320 hours per group member.

Table D.1: Timetable

Project duration	14 weeks
count of workers	2
time per worker	26 hours per week
Total estimated time (without weighted damage)	XX
Total estimated time (inclusive weighted damage)	XX
Project start	19.09.2016
Project end	23.12.2016

D.1.2 Time Tracking

All schedulings are done with Everhour and with git. Git includes all issues and Everhour is used for the time capturing (should / is).

D.1.3 Infrastructure

Table D.2: Timetable

Servename	sinv-56092.edu.hsr.ch
ip address	152.96.56.92
os	Ubuntu 14.04 LTS
VM EndofLife	03.03.2017
CPU	1 vCPU max. 2.2 GHz
RAM	1 GB
Disk drive	15 GB

D.1.4 Tools

Table D.3: Timetable

Use	Name	Version
IDE	vim, ruby mine	
version control system	git	2.*
project management	git, ever- hoour	

D.1.5 Quality Measure

D.1.5.1 Documentation Review

The project partner reads important sentences as well it will be discussed in the weekly stand up meetings.

D.1.5.2 Meeting Minutes

Meeting minutes include the participant, agenda items, decisions and todos. Manuel Schuler take over this task and logs everything. Every meeting minutes is saved on the github wiki site 24 hours before the meeting start and after the meeting it will be updated with the new information.

D.1.5.3 Git Policy

It's only allowed to push if all ruby spec tests pass.

D.1.6 Meeting

Meetings are agreed in consultation with the lecturer or the customer. The meeting minutes is kept, the individual agenda items, decisions and todos. The project team makes one Weekly standup meeting - place in early / mid-week and serves only as a short exchange of information, review last week, recap the week goals, problems and plenary for questions or suggestions.

D.2 Risks

There were X risks that have been identified by the end of the Inception phase. The risks have a total damage of xxx hours. The total damaged hours multiplied with the probability of admission get a total of xx hours. The weighted damage hours are included in the project planning.

D.2.1 Handling Risks

Due to the nature of the risks, it is only natural they change during the course of a project. To mitigate this, the risks are checked regularly (in weekly meetings) using the table below.

Changes to existing risks are possible. This usually means that either the likelihood or the unweighted damage must be adapted immediately. Moreover, it's possible for a risk to be completely ruled out, or that a new risk arises. All these points need to be discussed in the team and tracked accordingly.

D.3 Listed Risks

P = Probability

1. Unlikely
2. Very rare
3. Rare
4. Possible
5. Common

D = Damage potential / R = Risk

1. Insignificant
2. Low
3. Significant
4. Critical
5. Project Threatening

The delay is specified in days. One day equals 16 man-hours.

Table D.4: Initial Risks

ID	Description	P	DP	Prevention	Measures to be taken upon event
R01	Roadster requires different ZMQ contexts to function (not possible with CZTop because CZMQ hides contexts)	1	3	check with client (done)	extract and run affected ZMQ sockets in their own process delay: 1-2 days
R02	wrong protocols chosen / protocol design flaw	3	5	architecture reviews prototypes	fix (reevaluate reengineer, redesign) delay: 8-12 days
R03	ZMQ communication patterns (such as Binary Star) are difficult to implement as clean, reusable code	2	4	use software engineering knowhow to aim for clean, reusable prototypes	nice solution: build more iteratively, step by step delay: 2-4 days dirty solution: customized solution built right into Roadster, not as a public gem delay: 1-2 days
R04	CZTop design flaws/limitations	2	2	check functionality in the elaboration phase	adapt CZtop delay: 1-2 days
R05	CZMQ changes API	1	3	(hope)	adapt CZTop, change CZTop adapter in Roadster, or just don't upgrade CZMQ (use a commit before the breaking change) delay: 1-2 days
R06	wrong time estimations	4	3	use time well during planning, and define clear milestones	If possible, change the duration of the individual project phases. Otherwise, drop planned features (starting with the optional goal) delay: 4-5 days

Continues on the next page

Table D.4 – continues

ID	Description	P	DP	Prevention	Measures to be taken upon event
R07	managing multiple Projects (at least one per repo) on Github too painful	4	1	setup project structure in the elaboration phase	<p>partial solution: CodeTree (can't seem to be used for private repos like Roadster itself (maybe yes! see mindclue/roadster#5))</p> <p>complete solution: Use a single Project which just has cards that link to issues from other repos. Linking to "foreign" issues is additional effort but should be straight forward using Github syntax (https://github.com/org/repo/issues/42) delay: 1 day</p>
R08	Prolonged loss of a team member	2	3	Track absences in meeting minutes.	In a prolonged absence, move milestones and, if necessary, change the project scope. delay: 3-10 days
R09	Failure to achieve the defined task in time	1	4	Continuous monitoring whether we are on schedule and whether all requirements are met.	Meeting convened as we still can transpose a large part of the required task within the prescribed period. delay: 1-5 days

Table D.5: Initial Risk Matrix

Propability / Damage	1-Insignificant.	2-Low	3-Significant	4-Critical	5-Project Threatening
5-Common					
4-Possible	R07		R06		
3-Rare					R02
2-Very rare		R04	R08	R03	
1-Unlikely			R01, R05	R09	

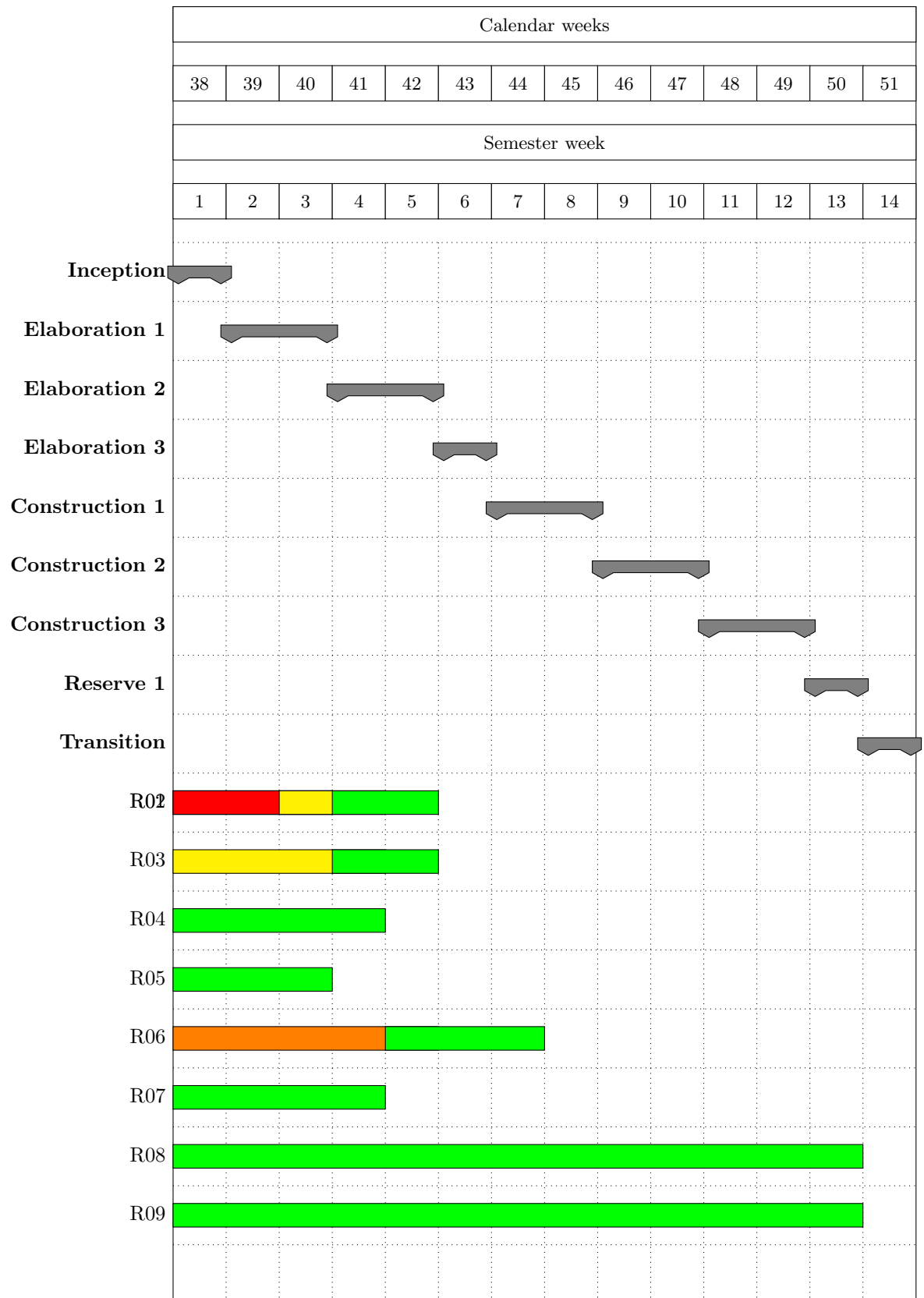


Table D.6: Risk-Timeline change protocol

Week	Risk	Description
3	R01
3	R06

D.4 Iterations & Phases

Table D.7: Phase and Iterations

Iteration	Description	
Inception	setup proj mgmt, init documentation, define scope, understand requirements, set priorities, assess & analyze risks, estimate schedule, get familiar with Roadster	S
MS Inception	Date 25th Sept 2016 Description Inceptionphase ended Workproducts project plan risk matrix project mgmt infrastrucur	
Elaboration 1	write use cases, fundamental thoughts on testing, roughly design protocols (cluster single & multi level HA, persistence, key distribution, OPC-UA HA interface)	S
MS E1 Protocol Designs	Date 9th Oct 2016 Description Protocol designs are defined. Workproducts requirements + use cases protocol designs	
Elaboration 2	implement prototypes (cluster, single & multi level HA, persistence, secure socket, communication, OPC-UA HA interface)	S
MS E2 Prototypes	Date 23rd Oct 2016 Description Prototypes are implemented and tested. Workproducts Runnable prototypes	
Elaboration 3	revise risks, finish bulk of documentation, (reserve)	S
Construction 1	port CZTop, cluster (refactor & integrate prototype)	S
MS C1 Cluster	Date 13th Nov 2016 Description Runnable cluster functionality on top of CZTop. Workproducts Cluster functionality CZTOP integration	
Construction 2	refactor, integrate and verify HA prototypes, persistence synchronization	S
MS C2 HA	Date 27th Nov 2016 Description Working HA functionality and persistence synchronization Workproducts HA functionality persistence synchronization	
Construction 3	security (implement prototype, test), OPC UA HA (implement prototype, verify)	S
MS C3 Security	Date 11th Dec 2016 Description secure communication between nodes Workproducts Security	

Continues on the next

Table D.7 – continues

Iteration	Description	
Transition 1	polish documentation, write abstract, create poster, print documentation & burn CDs	S
MS T1 Delivery	Date 16th Dec 2016 Description Complete handover of thesis Workproducts Thesis in paperform	S
Reserve	in case things go south	S

Appendix E

ØMQ

ØMQ is a [Message Oriented Middleware \(MOM\)](#) implemented as an open source library, that is, it doesn't require a dedicated broker. Instead, it offers sockets with an abstract interface similar to [BSD](#) sockets. Different types of sockets are used for different messaging patterns such as request-reply, publish-subscribe, and push-pull.

A single socket can bind/connect to multiple endpoints, which allows ØMQ to use round-robin on the sender side, and fair-queueing on the receiver side, where applicable. It doesn't matter whether the communication happens in-process (between threads), inter-process (e.g. over [Unix Domain Socketss](#)), or inter-node (e.g. over [TCP/PGM/TIPC](#)), since the transport is completely abstracted away. The same goes for connection handling; an arbitrary amount of connections is handled over a single socket and reconnecting after short network failures is done transparently.

ØMQ is lightweight and allows for extremely low latencies, which means it can also be used as the fabric of concurrent applications, e.g. for the [Actor Model](#). In case of the TCP transport, it incorporates advanced techniques such as smart message batching to achieve significantly higher throughputs than with raw TCP or other [MOM](#) solutions [4, Figure 2, Middleware evaluation and prototyping, p. 4].

To build a solution with ØMQ, its sockets are used as building blocks to design custom message flows. Certain patterns are used to achieve reliability with respect to the failure types that need to be addressed in particular. The [Zguide](#) explains best practices, including commonly needed, resilient messaging patterns.

The above characteristics make ØMQ a valuable asset when it comes to building robust, distributed high-performance systems.

E.1 Transport security

Since version 4.0 (released in October 2013), ØMQ boasts strong encryption and authentication, based on the excellent and highly renown [libsodium](#)¹. The protocol used (CurveZMQ) is described in [6].

Transport encryption is completely transparent to the application. The security handshake is designed to be highly resilient against [distributed denial-of-service attacks](#) using encrypted and authenticated cookies, so the server doesn't actually allocate any memory before

¹There's also the possibility to do it using [TweetNaCl](#) which avoids the additional dependency.

the handshake is completed, and key generation is very cheap [2, p. 2] with Curve25519² elliptic curve cryptography (ECC).

To enable, one socket is designated as the server, the other as the client. Server authentication is always performed. This means that nodes running the client sockets must be in possession of the server's public key.

Client authentication is optional. If desired — meaning not just any client is allowed to connect — the clients' public keys (either shared by all clients, or a unique key for each client) must be already available to the node running the server sockets. The server socket actually talks to another, designated socket to do authentication. The protocol used between them is ZAP. Completely abstracting the authentication in another socket allows any kind of authentication service to be easily plugged in. It's also trivial to write such a ZAP handler to register new client public keys somewhere to be checked and confirmed by a human, which simplifies the process of adding new client public keys.

E.2 Data serialization

Data serialization is outside the scope of ØMQ. To fill the gap, one typically uses another library such as MsgPack³, Protocol Buffers⁴, or even a programming language's built-in object serialization support⁵.

E.3 Language availability

Bindings or full-blown reimplementations of ØMQ exist for a plethora of programming languages, including C, C++, Java, Ruby, Perl, Erlang, Python, Lua, Go, Tcl, PHP, and .Net. This allows for building distributed software systems using modern approaches like service-oriented architectures, where different parts may be implemented in different languages.

E.4 CZMQ

CZMQ is a high-level abstraction layer for ØMQ. It makes working with the ØMQ library more expressive and allows for better portability. It also provides additional functionality such as a reactor, a simple actor implementation, as well as utilities for certificate and authentication handling, and LAN node discovery. This is the recommended way of using ØMQ nowadays, as it allows for much cleaner C code and also simplifies bindings for other languages.

²<https://en.wikipedia.org/wiki/Curve25519>

³<http://msgpack.org>

⁴<https://developers.google.com/protocol-buffers/>

⁵such as Ruby's marshalling support: <http://ruby-doc.org/core/Marshal.html>

Appendix F

Infrastructural Problems

TODO describe serious problems here, if any

F.1 Project Management Software

TODO Github/Trello/Harvest/Everhour/Elegantt/Ganttify/Redmine