Utilising Edge Computing in a 3-Tier system between Client Devices and a Data Centre

A dissertation submitted in partial fulfilment of the requirements for the degree of BACHELOR OF ENGINEERING in Computer Science

in

The Queens University of Belfast

by

Connor Dickson

9th May 2017

# Acknowledgements

* Hans
* Automated Intelligence
* Cassio

# Abstract

* 100 word outline of subject matter/findings

Table of Contents

[Introduction and Problem Specification 3](#_Toc477258767)

[System Requirements Specification 3](#_Toc477258768)

[Design 4](#_Toc477258769)

[Architectural Description 4](#_Toc477258770)

[User Interface Design 4](#_Toc477258771)

[Software System Design 5](#_Toc477258772)

[Client 5](#_Toc477258773)

[Edge 5](#_Toc477258774)

[Data Centre 5](#_Toc477258775)

[Implementation and Testing 5](#_Toc477258776)

[Choice of implementation languages/dev environments 5](#_Toc477258777)

[Use of software libraries 6](#_Toc477258778)

[Key implementation details 6](#_Toc477258779)

[How each component was implemented 6](#_Toc477258780)

[Discussion of test approach 6](#_Toc477258781)

[Caching System Tests 6](#_Toc477258782)

[Client Application 6](#_Toc477258783)

[Edge Node Application 7](#_Toc477258784)

[Voice Recognition System Tests 7](#_Toc477258785)

[Client Application 7](#_Toc477258786)

[Edge Node Application 9](#_Toc477258787)

[Data Centre Application 9](#_Toc477258788)

[Machine Learning System Tests 10](#_Toc477258789)

[Client Application 10](#_Toc477258790)

[Edge Node Application 10](#_Toc477258791)

[Data Centre Application 10](#_Toc477258792)

[Load Balancing System Tests 10](#_Toc477258793)

[Client Application 10](#_Toc477258794)

[Edge Node Application 10](#_Toc477258795)

[Data Centre Application 10](#_Toc477258796)

[System Evaluation 10](#_Toc477258797)

[Experimentation 11](#_Toc477258798)

[Caching Experimentation 11](#_Toc477258799)

[Does utilising Edge Computing reduce the latency of requests for the Client Device? 11](#_Toc477258800)

[Voice Recognition Experimentation 12](#_Toc477258801)

[Does utilising Edge Computing reduce the computational load on the Client Device? 12](#_Toc477258802)

[Does utilising Edge Computing increase the latency of the request for the Client Device? 13](#_Toc477258803)

[Machine Learning Experimentation 14](#_Toc477258804)

[Does utilising Edge Computing reduce the Network Utilisation? (not sure what to do with this one so far) 14](#_Toc477258805)

[Load Balancing Experimentation 14](#_Toc477258806)

[Do the benefits of utilising Edge Computing deteriorate when multiple requests are made concurrently? Can this be improved with a custom load balancing aspect? 14](#_Toc477258807)

[Conclusion 15](#_Toc477258808)

[References 16](#_Toc477258809)

[Copyright 16](#_Toc477258810)

[Appendices 16](#_Toc477258811)

[User manual 16](#_Toc477258812)

[Test Results 16](#_Toc477258813)

[Minutes of meetings 16](#_Toc477258814)

# Introduction and Problem Specification

Background material which introduces the problem area, context, and background.

Identify my problem

Systematically researched and fully analysed the problem

# System Requirements Specification

Precise description of the system developed. Should be updated for the final system delivered. List assumptions made about the problem and any system constraints(RAM?)

Functional and non-functional requirements should be complete, clear, accurate, feasible and objectively verifiable.

Read req’s for list of what can be included in this section

# Design

## Architectural Description



The entry point for the system is the Electron Application which runs on the Client Raspberry Pi. There are 4 pages that the user can access from the homepage. Traffic from each application is colour coded.

The Edge is comprised of 3 Raspberry Pi’s which form a Docker swarm. The applications run in containers that can be running on any of the Raspberry Pi worker nodes and have multiple instances depending on the deployment.

There are multiple data centres depending on what application is making the request. The WebAPI project is hosted in IIS and has multiple endpoints.

## User Interface Design

I tried to keep a consistent theme throughout my UI design. The main purpose of it was to demonstrate the functionality of the Edge Nodes.



(Other pages might need to redo the homepage one also)

## Software System Design

### Client

The role of the Client Application is to run tests and get results from the Edge and the Data Centre. It was a requirement of the system to provide an application that allows a user to interact with the Edge Node and Data Centre and to display data to the user. It does this by running different applications for each of the services running on the Edge Node to keep everything distantly separate but accessible.

(Read notes and discuss key design decisions)

Caching

Notes

Voice Recognition

Notes

Machine Learning

Notes

### Edge

The role of the Edge Service is to pre-process requests for the user. It was a requirement of the system to pre-process requests quickly before passing on Data to the Data Centre or returning it to the Client. Another requirement of the Edge Node is the ability to run multiple application on the same Edge Node to allow for extensibility. This is achieved by running the applications in Docker. This is It’s a service running in a container. I also created a reusable nodejs libraries

(Read notes and discuss key design decisions)

Caching

Redis on all pi’s (global) and access load balance style

Voice Recognition

No continuous listening

Machine Learning

Multiple iterations of development after meetings

Need a decent sized section for refactor of the way the machine learning was going to happen

### Data Centre

The role of the Data Centre is to be the endpoint that the Client wants an operation to occur on. For example the Client wants Voice Recognition to occur when they click a button so the fact it was pre-processed before reaching the Data Centre does not matter to them, all they want is the result of the request. The Data Centre was designed and implemented with dependency injection, this makes it easier to maintain and control dependencies and code reuse.

(Read notes and discuss key design decisions)

Voice Recognition

Slightly more challenging to get it working on windows

Machine Learning

Consistent code style. DI in C#. Choice of languages

* Role of each component and the interface between the components
* Clear correlation between design and specs

Design should be linked to requirements and give a critical discussion of key design decisions/styles/patterns. Read more in notes

# Implementation and Testing

## Choice of implementation languages/dev environments

Client

The Client Application was written using Electron. Electron utilises NodeJS for the backend. This decision was made as it would allow the same application to run cross platform. The Electron framework can run in Windows, MacOS and Linux.

Edge

The Edge Node applications are written in JavaScript and run in a NodeJS environment. This was chosen because the Client was written in JavaScript and NodeJS it was a familiar language. It was also easier to reuse code between the two applications such as the CPU library.

Data Centre

The Data Centre was written utilising WebAPI and hosted in IIS as this was readily available on a Windows environment.

## Use of software libraries

* Pocketsphinx for voice recognition (mention that a recording was used rather than continuous listening)
* Nodejs packages
  + NodeJS proxy
  + NodeRedis
  + Request
* JSON
* Redis

## Key implementation details

Docker… and deployment scripts. The UI that was written for ARM. How the docker file and base images work.

* Once I got the setup done.
* Deciding on redis because of docker image
* Deciding on node/webapi
* Read notebook
* Proxy instead of regular web server. This allows for extraction of URL’s on edge node and makes it seem more like real life

## How each component was implemented

Client

The Client was implemented by

Edge

Docker Scripts or base images. The Edge was implemented by

Data Centre

The Data Centre was implemented by

## Discussion of test approach

Because of the nature of the code it could prove difficult to test as it has to run in a nodejs server and made actual requests so there was a lot of manual testing. I tried to run tests using QUnit but the test would throw an exception as soon as the first NodeJS keyword (“require”) came up. This meant that manual testing was the approach that would be taken. This would be adequate as the system is for research purposes. Unit tests in WebAPI also proved difficult as a lot of the code deals with actual requests. The tests that were able to be written did not deal with request code.

## Caching System Tests

### Client Application

|  |  |  |  |
| --- | --- | --- | --- |
| Test Name | Method | Expected Results | Passed |
| Can reach Caching Application page | Click on the Client Application button on the home page | The Caching Application page is now launched | True |
| Home button | Click home button | Should be returned to the home page |  |
| URL bar is present | On the caching page ensure there is a URL bar | The URL bar is present |  |
| Can type into URL bar | Click into the URL bar and type an address | The address types will show up in the URL bar |  |
| Go button without URL typed | Click the Go button when there is no URL typed out | Nothing, there will be no error |  |
| Go button with an invalid URL typed | Click the Go button when an invalid URL is typed into the URL bar | A message will appear warning the user that the request failed |  |
| Go button with a valid URL typed | Click the Go button when a valid URL is typed | The requested webpage will be fetched and displayed to the user |  |
| The timer | Click the go button with a valid URL | The timer will start and when the request is finished will record the total time for the request to finish |  |
| Clear cache button | Click the clear cache button | Information will appear to indicate that the cache on the edge node was cleared successfully |  |
| Can connect to the proxy | Make a web request | Ensure that the proxy is utilised |  |

### Edge Node Application

|  |  |  |  |
| --- | --- | --- | --- |
| Test Name | Method | Expected Results | Passed |
| Can connect to the Redis instance | Deploy the web server | The server will automatically connect to Redis |  |
| Can make request to Caching Web Server | Make a curl request to the Edge Node using the host name and port number | Receive Data back |  |
| Receive clear cache request | Call the Edge Node with a ClearCache request | The clear cache command is executed using the redis-cli and the cache is cleared. This information is then returned to the user |  |
| Receive new request | Call the Edge Node with a new URL | The content is retrieved, stored in Redis and returned to the client |  |
| Receive a request for the second time | Call the Edge Node with a URL that has already been requested | The content is retrieved from Redis and returned to the user |  |

## Voice Recognition System Tests

### Client Application

|  |  |  |  |
| --- | --- | --- | --- |
| Test Name | Method | Expected Results | Passed |
| Can reach the Voice Recognition page | Click on the Voice Recognition button on the home page | The Voice Recognition Application page is now launched |  |
| Home button | Click home button | Should be returned to the home page |  |
| Recording status after page launch | Launch the Voice Recognition application | The recording status should be “Not Recording” |  |
| Recording status after clicking record | Click the record button | The recording status should change to “Recording” |  |
| Recording status after stopping recording | Click the stop recording button after starting a recording | The recording status should change to “Finished recording” |  |
| Recording button when not recording | Stop a recording or don’t start a recording | The recording button should read to “Start Recording” |  |
| Recording button when recording | Start a recording or don’t stop a recording | The recording button should read “Stop Recording” |  |
| Execute local recognition button without recording | Press the Execute Local Recognition button without recording a phrase | Nothing should happen, no error should be thrown |  |
| Execute local recognition button after recording | Press the Execute Local Recognition button after recording a phrase | The execution of the recording should be performed locally |  |
| Execute remote recognition button with no recording | Press the Execute Remote Recognition button without recording a phrase | Nothing should happen, no error should be thrown |  |
| Execute remote recognition button after recording without edge processing | Press the Execute Remote Recognition button after recording a phrase and don’t tick the pre-process checkbox | The Execution of the recording should be performed remotely on the Data Centre |  |
| Execute remote recognition button after recording with edge processing | Press the Execute Remote Recognition button after recording a phrase and tick the pre-process checkbox | The Execution of the recording should be performed remotely on the Edge Node and the results sent to the Data Centre |  |
| Local Results section when no local evaluation has occurred | Launch the Voice Recognition Application and don’t perform local evaluation | The results section is empty |  |
| Local Results section when the local evaluation has occurred | Perform local evaluation | The local results section should show statistics |  |
| Remote Results section when no remote evaluation has occurred | Launch the Voice Recognition Application and don’t perform remote evaluation | The results section is empty |  |
| Remote Results section when no remote evaluation has occurred | Perform remote evaluation | The remote results section should show statisticss |  |

### Edge Node Application

|  |  |  |  |
| --- | --- | --- | --- |
| Test Name | Method | Expected Results | Passed |
| Can make request to Voice Recognition Web Server | Make a curl request to the Edge Node using the hostname and port number | Receive Data back |  |
| Receive Get Request | Make a get request to the Edge Node | No error but no data returned |  |
| Receive Put Request | Make a put request to the Edge Node | No error but no data returned |  |
| Receive post request with invalid data | Make a post request with an invalid audio file or incorrect data | No error but no data is returned |  |
| Receive post request with valid recording | Make a request with a valid voice recording | A JSON object is returned with information about the request and the audio spoken |  |
| Receive valid request with the pre-process header | Make request with the pre-process request header set | The processing of the voice file happens on the Edge Node |  |
| Receive valid request without the pre-process header | Make request without the pre-process request header set | The voice recording is sent to the data centre for processing and the results returned to the user |  |
| Request information from the Data Centre when the WebAPI is not published | Remove the WebAPI and make a request | The service should record an error |  |
| Request information from the Data Centre when the WebAPI is published | Publish the WebAPI and make a request | The information should be returned correctly |  |
| Measure CPU use | Make a regular request | The CPU use on the edge node should be returned as part of the request |  |
| Measure Time of request | Make a regular request | The Time taken to process the request should be recorded |  |

### Data Centre Application

|  |  |  |  |
| --- | --- | --- | --- |
| Test Name | Method | Expected Results | Passed |
| Receive invalid request to process voice | Make a valid POST request with binary voice data | The voice is processed correctly and a valid response is produced |  |
| Receive valid request to process voice | Make an invalid POST request with invalid voice data | The voice request should be handled correctly and gracefully respond |  |
| Receive invalid request to record information | Make a valid POST request with textual data from a pre-processed request | The request should be processed correctly and a response generated |  |
| Receive valid request to record information | Make an invalid POST request without valid textual data | The request should be handled gracefully and a response sent |  |

## Machine Learning System Tests

### Client Application

|  |  |  |  |
| --- | --- | --- | --- |
| Test Name | Method | Expected Results | Passed |
|  |  |  |  |

### Edge Node Application

|  |  |  |  |
| --- | --- | --- | --- |
| Test Name | Method | Expected Results | Passed |
|  |  |  |  |

### Data Centre Application

|  |  |  |  |
| --- | --- | --- | --- |
| Test Name | Method | Expected Results | Passed |
|  |  |  |  |

## Load Balancing System Tests

### Client Application

|  |  |  |  |
| --- | --- | --- | --- |
| Test Name | Method | Expected Results | Passed |
|  |  |  |  |

### Edge Node Application

|  |  |  |  |
| --- | --- | --- | --- |
| Test Name | Method | Expected Results | Passed |
|  |  |  |  |

### Data Centre Application

|  |  |  |  |
| --- | --- | --- | --- |
| Test Name | Method | Expected Results | Passed |
|  |  |  |  |

# System Evaluation

Not sure what’s included here

# Experimentation

A lot of the emphasis in this project will be here.

Provide empirical results and draw conclusions.

Describe methodology (use experiment methodology like goals and hypothesis, what’s measured/controlled)

When is it good to use edge node. When is it good to use DC? Not always good to use both

Read more in notes

## Caching Experimentation

### Does utilising Edge Computing reduce the latency of requests for the Client Device?

Setup;

* Run the Caching Application on the Client Device
* Deploy the Caching Service to one of the Edge Devices
* Deploy a Redis Instance to each of the Edge Devices

Isolate Variables;

Constants:

* Interface connecting Raspberry Pi to the router
* The web page being loaded
* The other applications deployed to the Edge Nodes

Variables

* Whether a webpage is cached or not

Method;

Measurements recorded will be the time between when the WebView element in the Client Application starts a request and ends a request.

After the Caching Service and Redis Service have been deployed, 10 warmup requests will performed without wiping the cache. The experiment will be performed with 10 iterations and an average will be taken.

An iteration will consist of an initial request to “http://www.bbc.co.uk” and then a subsequent request when the information is already cached on the edge node. The two times will then be recorded in seconds and the cache will be cleared ready for the next request

Hypothesis;

The hypothesis is that the cached request should take less time to execute than the initial request.

Results;

<Graph>

Analyse;

<Analysis of graph>

Conclusion;

## Voice Recognition Experimentation

### Does utilising Edge Computing reduce the computational load on the Client Device?

Setup;

* Run the Voice Recognition Application on the Client Device
* Save a Voice Recording using the Client Application
* Deploy the Voice Recognition Service to one of the Edge Device
* Publish the Data Centre WebAPI with the Voice Recognition endpoint

Isolate Variables;

Constants:

* The Voice Recording
* The Language Model and Dictionary being used for the processing
* The applications deployed to the Edge Nodes

Variables

* Where the voice processing is occurring

The variable in this experiment is whether I process the file locally on the Client Device, remotely pre-process it on the Edge Device or remotely process it on the Data Centre.

Method;

Measurements recorded will be:

* CPU load in a percentage of the total CPU on the Client Device
* CPU load in a percentage of the total CPU on the Edge Node
* Processing time in seconds
* File size
* Length of voice recording

After the Voice Recognition Service has been deployed 10 warmup requests will be performed on Client Device, a further 10 will be performed on the Edge Node and a final 10 on the Data Centre.

The measure of time will be recorded on the Client Device from as soon as the request starts to when the request finishes. The experiment will be repeated 10 times and an average will be taken.

The experiment will consist of a voice being recorded. The same recording will be used throughout the experiment to allow for a fair comparison of computational load.

There will be a round of 10 experiments run on the Client Device, then a round of 10 where the device is pre-processed on the Edge Node, and a final 10 where the request is processed on the Edge Device.

Hypothesis;

The hypothesis is that the Data Centre should process the request the fastest as it has the most powerful CPU but the CPU strain on the Client Device should be reduced regardless of the request being pre-processed on the Edge Device or the Data Centre.

Results;

<Graph>

Analyse;

<Analyse graph>

Conclusion;

### Does utilising Edge Computing increase the latency of the request for the Client Device?

Setup;

* Run the Voice Recognition Application on the Client Device
* Save a Voice Recording using the Client Application
* Deploy the Voice Recognition Service to one of the Edge Device
* Publish the Data Centre WebAPI with the Voice Recognition endpoint

Isolate Variables;

Constants:

* The Voice Recording
* The Language Model and Dictionary being used for the processing
* The applications deployed to the Edge Nodes

Variables

* Where the voice processing is occurring

The variable in this experiment is whether I process the file locally on the Client Device, remotely pre-process it on the Edge Device or remotely process it on the Data Centre.

Method;

Measurements recorded will be:

* CPU load in a percentage of the total CPU on the Client Device
* CPU load in a percentage of the total CPU on the Edge Node
* Processing time in seconds
* File size
* Length of voice recording

After the Voice Recognition Service has been deployed 10 warmup requests will be performed on Client Device, a further 10 will be performed on the Edge Node and a final 10 on the Data Centre.

The measure of time will be recorded on the Client Device from as soon as the request starts to when the request finishes. The experiment will be repeated 10 times and an average will be taken.

The experiment will consist of a voice being recorded. The same recording will be used throughout the experiment to allow for a fair comparison of computational load.

There will be a round of 10 experiments run on the Client Device, then a round of 10 where the device is pre-processed on the Edge Node, and a final 10 where the request is processed on the Edge Device.

Hypothesis;

The latency of the request should not increase much if the request is pre-processed on the Edge Device as both devices are Raspberry Pi’s however the request should be processed much quicker on the Data Centre as it has a faster processor although this means that the Data Centre is performing more work.

Results;

<Graphs>

Analyse;

<Analysis of graphs>

Conclusion;

## Machine Learning Experimentation

### Does utilising Edge Computing reduce the Network Utilisation? (not sure what to do with this one so far)

Setup;

Isolate Variables;

Constants:

Variables

Method;

Hypothesis;

Results;

<Graph>

Analyse;

<Analysis of graph>

Conclusion;

## Load Balancing Experimentation

### Do the benefits of utilising Edge Computing deteriorate when multiple requests are made concurrently? Can this be improved with a custom load balancing aspect?

Setup;

* Run the Load Balancing application on the Client Device
* Deploy the Voice Recognition Service to one of the Edge Nodes
* Publish the Data Centre WebAPI with the Voice Recognition endpoint

Isolate Variables;

Constants;

* The Voice Recording
* The software package being used
* The Language Model and Dictionary being used for Processing
* The CPU use from other applications

Variables;

* Number of requests
* Whether the deployed Edge Node service is utilising custom Load Balancing

Method;

Measurements recorded will be:

* Average CPU load in a percentage of the total CPU on the Client Device
* Average CPU load in a percentage of the total CPU on the Edge Node
* Processing time in seconds for all requests to finish

The experiment will consist of a voice being recorded. The same recording will be used throughout the experiment to allow for a fair comparison of computational load.

Deploy the Voice Recognition Service to the Edge Node without custom load balancing enabled and perform 2 sets of 10 warm up requests.

Perform the experiment 10 times with varying number of requests, from 10 to 50 in increments of 10. Repeat all 5 experiments after deploying the Service to the Edge Node with load balancing enabled and performing the warmup again.

Hypothesis;

Adding load balancing will improve response times (over 70% CPU utilisation the request gets sent to the Data Centre)

Results;

Analyse;

Conclusion

# Conclusion

General summary of the success of the project with respect to criteria identified in the intro

Discussion of significance of experimental results. Agree with others work?

Strengths and weaknesses

Evaluation of hardware/software

Looking for critical appraisal and significance of contribution in the context of wider work.

# References

[1]"Final Year Project Repository", *GitLab*, 2017. [Online]. Available: https://gitlab.eeecs.qub.ac.uk/40103631/FinalYearProject. [Accessed: 21- Mar- 2017]

# Copyright

* Electron
* Nodejs/ Nodejs libraries
  + Node JS proxy
* Pocketsphinx
* Testing Suites
* Redis/Docker src images
* JSON
* Check licence file in Client src folder

# Appendices

## User manual

## Test Results

Excel results?

## Minutes of meetings