

Alexsi Pekkala

Migrating a web application to serverless architecture

Master's Thesis in Information Technology

February 6, 2018

University of Jyväskylä

Department of Mathematical Information Technology

Author: Aleksi Pekkala

Contact information: alvianpe@student.jyu.fi

Supervisor: Oleksiy Khriyenko

Title: Migrating a web application to serverless architecture

Työn nimi: Web-sovelluksen siirtäminen serverless-arkkitehtuuriin

Project: Master's Thesis

Study line: Master's Thesis in Information Technology

Page count: 20+0

Abstract: This document is a sample gradu3 thesis document class document. It also functions as a user manual and supplies guidelines for structuring a thesis document.

The abstract is typically short and discusses the background, the aims, the research methods, the obtained results, the interpretation of the results and the conclusions of the thesis. It should be so short that it, the Finnish translation, and all other meta information fit on the same page.

The Finnish tiivistelmä of a thesis should usually say exactly the same things as the abstract.

Keywords: serverless, FaaS, architecture, cloud computing, web applications

Suomenkielinen tiivistelmä: Tämä kirjoitelma on esimerkki siitä, kuinka gradu3-tutkielmapohjaa käytetään. Se sisältää myös käyttöohjeet ja tutkielman rakennetta koskevia ohjeita.

Tutkielman tiivistelmä on tyypillisesti lyhyt esitys, jossa kerrotaan tutkielman taustoista, tavoitteesta, tutkimusmenetelmästä, saavutetuista tuloksista, tulosten tulkinnasta ja johtopäätöksistä. Tiivistelmän tulee olla niin lyhyt, että se, englanninkielinen abstrakti ja muut metatiedot mahtuvat kaikki samalle sivulle.

Sen tulee kertoa täsmälleen samat asiat kuin englanninkielinen abstrakti.

Avainsanat: serverless, FaaS, arkkitehtuuri, pilvilaskenta, web-sovellukset

Contents

1	INTRODUCTION	1
1.1	Research questions	2
1.2	Outline	2
2	SERVERLESS COMPUTING	4
2.1	What is serverless?	4
2.2	Use cases.....	4
2.3	Execution models and triggers	5
2.4	Runtime	5
2.5	Service providers	6
2.6	Economics of serverless	6
2.7	Security	6
2.8	Drawbacks and limitations	6
3	SERVERLESS DESIGN PATTERNS	8
3.1	Serverless patterns.....	8
3.2	Enterprise Integration Patterns	9
3.3	FaaSification	9
4	MIGRATION PROCESS	10
5	EVALUATION.....	11
6	CONCLUSION	12
	BIBLIOGRAPHY	13

1 Introduction

It is a good idea to start the Introduction with the main thesis statement or research question of the thesis. After that, it is a good idea to clarify things by defining any necessary terms. Definitions after the thesis statement! Also, don't babble in the introduction. The introduction is also a good place to discuss why your thesis statement is scientifically or practically relevant and interesting. Ideally, it would be relevant and interesting from both the scientific and the practical point of view. It would also be excellent if you explained, in the introduction, what your contribution is; that is, what such knowledge your thesis contains that you have investigated personally instead of reading it from somewhere. The contribution could well be, that you have personally checked the truth of a claim you found in a book or article. At the end of the Introduction, it is customary to briefly explain the structure of the thesis – what each chapter is about.

Buyya et al. (2017) go as far as to describe cloud computing as the backbone of modern economy.

In a nutshell: serverless computing is a relatively new form of cloud computing that enables efficient resource utilization and elasticity, thus bringing about potential savings in hosting costs and increases in developer productivity (Roberts 2006). As a rather new development there's a lack of research in serverless computing. Multiple authors, e.g. Baldini, Castro, et al. (2017), Fox et al. (2017) and Eyk et al. (2017) note the identification of FaaS-specific patterns (composing functions into larger systems) as an important research topic. The object of this thesis is to fill that gap by a) surveying existing patterns in literature and b) evaluate, extend and propose new patterns by migrating parts of a web application into serverless architecture.

Balalaie, Heydarnoori, and Jamshidi (2016) talk about the reasons behind migrating systems to cloud-native architectures.

1.1 Research questions

Introduce the following research questions:

1. Why should a web app be migrated to FaaS?
2. What kind of patterns are there for building serverless web application backends?
3. Do the existing patterns have gaps or missing parts, and if so, can we come up with improvements or alternative solutions?
4. How does migrating a web app to FaaS effect its quality?

1.2 Outline

The thesis is structured as follows: the second chapter serves as an introduction to the concept of serverless computing. The chapter describes the main benefits and drawbacks of the platform, as well as touching upon its internal mechanisms and briefly comparing the main service providers. Extra emphasis is placed on how the platform's limitations should be taken into account when designing web application backends.

The third chapter consists of a survey into existing serverless design patterns and recommendations. Applicability of other cloud computing, distributed computing and enterprise integration patterns is also evaluated.

The fourth chapter describes the process of migrating an existing web application to serverless architecture. The patterns discovered in the previous chapter are utilized to implemented various typical web application features on a serverless platform. In cases where existing patterns prove insufficient or unsuitable as per the target application's characteristics, modifications or new patterns are proposed. This chapter forms the exploratory or constructive part of the thesis.

The outcome of the migration process is evaluated in the fifth chapter. The potential benefits and drawbacks of the serverless platform outlined in chapter 2 are used to reflect on the final artifact. The chapter includes approximations on measurable attributes such as hosting costs

and performance as well as discussion on the more subjective attributes like maintainability and developer experience.

The final chapter of the thesis aims to draw conclusions on the migration process and the resulting artifacts. The chapter contains a summary of the research outcomes and ends with recommendations for further research topics.

2 Serverless computing

The goal of the theoretical part of a thesis is to develop the theoretical background required in the thesis. The idea is that a reader of the thesis should, based on just the thesis itself, be able to understand all the special concepts and methods used in the thesis. A good thesis also gives well-argued reasons for why exactly these concepts and methods are in use in the thesis (with the main alternatives given in the literature mentioned).

2.1 What is serverless?

Definition and implications of serverless computing. Main features or tenets. Benefits, use cases, notable users? Roberts (2006) has a thorough description of the platform. Eyk et al. (2017), Varghese and Buyya (2018) and Buyya et al. (2017) provide a more historical context and future research directions. Albuquerque Jr et al. (2017) compare FaaS to PaaS.

Two historical/evolutionary views: on-premises infra -> grid -> IaaS -> PaaS -> serverless and on-prem -> VMs -> containers -> serverless.

How does serverless relate to concepts such as FaaS, SOA, cloud-native, microservices, containers, ...? Cloud-ready/cloud-native as defined by Pozdniakova and Mazeika (2017).

2.2 Use cases

Malawski et al. (2017) find that while serverless infrastructures are designed mainly for processing background tasks of Web and IoT applications, the simple mode of operation makes this approach easy to use and promising in scientific workflows too. Jonas et al. (2017) argue that a serverless execution model with stateless functions can enable radically-simpler, fundamentally elastic, and more user-friendly distributed data processing systems. Spillner, Mateos, and Monge (2018) also find that in many domains of scientific and high-performance computing, solutions can be engineered based on simple functions which are executed on commercially offered or self-hosted FaaS platforms.

Glikson, Nastic, and Dustdar (2017) propose the novel paradigm of Deviceless Edge Computing that extends the serverless paradigm to the edge of the network, enabling IoT and Edge devices to be seamlessly integrated as application execution infrastructure. Nastic et al. (2017) present a novel approach implementing cloud-supported, real-time data analytics in serverless edge-computing applications. Baresi, Filgueira Mendonça, and Garriga (2017) propose a serverless architecture at the edge, bringing a highly scalable, intelligent and cost-effective use of edge infrastructure's resources with minimal configuration and operation efforts.

Fouladi et al. (2017) present a serverless video-processing framework. Yan et al. (2016) present the architecture and prototype of a chatbot using a serverless platform, where developers compose stateless functions together to perform useful actions. Ishakian, Muthusamy, and Slominski (2017) evaluate the suitability of a serverless computing environment for the inferencing of large neural network models. Ast and Gaedke (2017) describe an approach of how to utilize serverless computing to enable self-contained web components by deploying Web Component business logic as cloud-hosted functions.

2.3 Execution models and triggers

Describe the two supported execution models, synchronous and asynchronous, and how they relate to application design. The former is used to build a typical request-response flow, e.g. a REST API endpoint, whereas the latter relates to pub-sub and other event-driven flows. Give examples on the kind of triggers supported by serverless platforms (HTTP calls, messaging, database events, ...).

2.4 Runtime

Briefly describe the inner workings of a FaaS runtime.

Spillner (2017b) presents the design and implementation of a research-friendly FaaS runtime. McGrath and Brenner (2017) present the design of a novel performance-oriented serverless computing platform, discussing implementation challenges such as function scaling and con-

tainer discovery, lifecycle, and reuse.

2.5 Service providers

Lynn et al. (2017) provide an overview and multi-level feature analysis of seven enterprise serverless computing platforms.

2.6 Economics of serverless

Eivy (2017) and Villamizar et al. (2016) both focus on the economic aspects of serverless. Adzic and Chatley (2017) explain how novel design patterns are used to significantly optimize costs – just running traditional web apps inside Lambda containers doesn't necessarily equate to savings. Adzic and Chatley (2017) also report savings between 66 and 95% in two case studies, and present a handy table comparing hosting prices for intermittent service tasks. Spillner (2017a) exploits the control plane of AWS Lambda to implement services practically for free. Leitner, Cito, and Stöckli (2016) present an approach to model deployment costs of AWS Lambda applications in real-time. Kuhlenkamp and Klems (2017) present another cost-tracing system that enables per-request cost-tracing for cloud-based software services, noting that cost testing should not only rely on isolated tests of single services but consider comprehensive end-to-end cost traces.

2.7 Security

Address the security implications of serverless.

2.8 Drawbacks and limitations

What to take into consideration when migrating to serverless?

Lloyd et al. (2018) analyze serverless performance and elasticity, identifying the cold start phenomenon.

Baldini, Cheng, et al. (2017) identify three competing constraints in serverless function com-

position: functions should be considered as black boxes; function composition should obey a substitution principle with respect to synchronous invocation; and invocations should not be double-billed.

Roberts (2006) and Adzic and Chatley (2017) respectively list a number of limitations, including lack of strong SLA, vendor lock-in, short life-span and immature local development tools.

Kuhlenkamp and Klems (2017) discover two serverless cost tradeoffs: the retry cost effect and the cost ripple effect.

The need for circuit breakers (risk of DDoSing yourself) when interacting with non-serverless components like a database. Mention novel cloud-native database services like Google's Cloud Spanner and AWS Aurora. Figure out a source for this – Hohpe and Woolf (2004) might have a relevant pattern.

3 Serverless design patterns

Survey of serverless design patterns. Baldini, Castro, et al. (2017) put the question as follows:

Will there be patterns for building serverless solutions? How do we combine low granularity basic building blocks of serverless into bigger solutions? How are we going to decompose apps into functions so that they optimize resource usage? For example how do we identify CPU-bound parts of applications built to run in serverless services? Can we use well-defined patterns for composing functions and external APIs? What should be done on the server vs. client (e.g., are thicker clients more appropriate here)? Are there lessons learned that can be applied from OOP design patterns, Enterprise Integration Patterns, etc.?

Composing serverless functions is not like composing regular functions. All the difficulties of distributed computing – message loss, timeouts and others – apply and have to be handled. Possible solutions include retry policies, dead-letter queues and idempotent functions.

3.1 Serverless patterns

Sbarski and Kroonenburg (2017) introduce the following five patterns: Command, Messaging Priority queue, Fan-out and Pipes and filters.

Adzic and Chatley (2017) suggest 3 methods for optimizing resource usage: use distributed authorization, let clients orchestrate workflows and allow clients to directly connect to AWS resources. The authors also discuss how paying only for actual utilization has two additional benefits of 1) removing incentives for bundling and 2) removing barriers to versioning.

Ast and Gaedke (2017) describe self-contained web components with serverless backends.

3.2 Enterprise Integration Patterns

Hohpe and Woolf (2004) present a number of asynchronous messaging architectures in the classic Enterprise Integration Patterns. While predating the whole serverless phenomenon the patterns are still relevant. Hohpe even demonstrated implementing one of his patterns on top of Google's serverless platform in a blog post.

3.3 FaaSification

Spillner (2017c) describes an automated approach to transform monolithic Python code into modular FaaS units by partially automated decomposition. Doesn't really seem suitable for the web application migration process covered in this thesis but worth mentioning.

4 Migration process

Implement a subset of the target app in a serverless style, utilizing the surveyed patterns and keeping a of the tricky parts. In case the patterns prove unsuitable for the given problem, try to come up with an alternative solution.

Decide on the parts to migrate. Should demonstrate the features and limitations of serverless as outlined above.

A simple REST API endpoint to showcase API Gateway and synchronous invocation. Shouldn't require any big changes to application code.

Interaction between multiple services to demonstrate distributed transactions.

A scheduled (cron) event.

Interacting with an external SaaS service like Twilio, Auth0. Demonstrate event-driven invocation.

5 Evaluation

Evaluation the outcome of migration process. Estimate the effects on performance and hosting costs. Weigh in on maintainability, testability, developer experience etc.

6 Conclusion

The last chapter of a thesis is the Conclusion (some authors use Conclusions, instead). Keep it short, and discuss what one can conclude about the thesis statement or research question given in the Introduction, in light of all that has been written in the thesis. The Conclusion is also the place to discuss any limitations and weaknesses of the thesis (especially those that cast doubt on the reliability of the results given in the thesis), if they have not been already discussed, for example in a Discussion chapter. It is also customary to state, what further research might be beneficial in light of this thesis.

If the Conclusion threatens to become too long, it is a good idea to split the interpretation of the results into its own chapter, often called Discussion, making Conclusion short and sweet.

Bibliography

Adzic, Gojko, and Robert Chatley. 2017. “Serverless computing: economic and architectural impact”. In *Proceedings of the 2017 11th Joint Meeting on Foundations of Software Engineering*, 884–889. ACM.

Albuquerque Jr, Lucas F, Felipe Silva Ferraz, Rodrigo FAP Oliveira, and Sergio ML Galdino. 2017. “Function-as-a-Service X Platform-as-a-Service: Towards a Comparative Study on FaaS and PaaS”. *ICSEA 2017*: 217.

Ast, Markus, and Martin Gaedke. 2017. “Self-contained Web Components Through Serverless Computing”. In *Proceedings of the 2Nd International Workshop on Serverless Computing*, 28–33. WoSC '17. Las Vegas, Nevada: ACM. ISBN: 978-1-4503-5434-9. doi:10.1145/3154847.3154849. <http://doi.acm.org/10.1145/3154847.3154849>.

Balalaie, Armin, Abbas Heydarnoori, and Pooyan Jamshidi. 2016. “Migrating to Cloud-Native Architectures Using Microservices: An Experience Report”. In *Advances in Service-Oriented and Cloud Computing*, edited by Antonio Celesti and Philipp Leitner, 201–215. Cham: Springer International Publishing. ISBN: 978-3-319-33313-7.

Baldini, Ioana, Paul C. Castro, Kerry Shih-Ping Chang, Perry Cheng, Stephen J. Fink, Vatche Ishakian, Nick Mitchell, et al. 2017. “Serverless Computing: Current Trends and Open Problems”. *CoRR* abs/1706.03178. arXiv: 1706.03178. <http://arxiv.org/abs/1706.03178>.

Baldini, Ioana, Perry Cheng, Stephen J. Fink, Nick Mitchell, Vinod Muthusamy, Rodric Rabbah, Philippe Suter, and Olivier Tardieu. 2017. “The Serverless Trilemma: Function Composition for Serverless Computing”. In *Proceedings of the 2017 ACM SIGPLAN International Symposium on New Ideas, New Paradigms, and Reflections on Programming and Software*, 89–103. Onward! 2017. Vancouver, BC, Canada: ACM. ISBN: 978-1-4503-5530-8. doi:10.1145/3133850.3133855. <http://doi.acm.org/10.1145/3133850.3133855>.

- Baresi, Luciano, Danilo Filgueira Mendonça, and Martin Garriga. 2017. "Empowering Low-Latency Applications Through a Serverless Edge Computing Architecture". In *Service-Oriented and Cloud Computing*, edited by Flavio De Paoli, Stefan Schulte, and Einar Broch Johnsen, 196–210. Cham: Springer International Publishing. ISBN: 978-3-319-67262-5.
- Buyya, Rajkumar, Satish Narayana Srirama, Giuliano Casale, Rodrigo N. Calheiros, Yogesh Simmhan, Blesson Varghese, Erol Gelenbe, et al. 2017. "A Manifesto for Future Generation Cloud Computing: Research Directions for the Next Decade". *CoRR* abs/1711.09123. arXiv: 1711.09123. <http://arxiv.org/abs/1711.09123>.
- Eivy, Adam. 2017. "Be Wary of the Economics of" Serverless" Cloud Computing". *IEEE Cloud Computing* 4 (2): 6–12.
- Eyk, Erwin van, Alexandru Iosup, Simon Seif, and Markus Thömmes. 2017. "The SPEC cloud group's research vision on FaaS and serverless architectures". In *Proceedings of the 2nd International Workshop on Serverless Computing*, 1–4. ACM.
- Fouladi, Sadjad, Riad S Wahby, Brennan Shacklett, Karthikeyan Balasubramaniam, William Zeng, Rahul Bhalerao, Anirudh Sivaraman, George Porter, and Keith Winstein. 2017. "Encoding, Fast and Slow: Low-Latency Video Processing Using Thousands of Tiny Threads." In *NSDI*, 363–376.
- Fox, Geoffrey C., Vatche Ishakian, Vinod Muthusamy, and Aleksander Slominski. 2017. "Status of Serverless Computing and Function-as-a-Service (FaaS) in Industry and Research". *CoRR* abs/1708.08028. arXiv: 1708.08028. <http://arxiv.org/abs/1708.08028>.
- Glikson, Alex, Stefan Nastic, and Schahram Dustdar. 2017. "Deviceless Edge Computing: Extending Serverless Computing to the Edge of the Network". In *Proceedings of the 10th ACM International Systems and Storage Conference*, 28:1–28:1. SYSTOR '17. Haifa, Israel: ACM. ISBN: 978-1-4503-5035-8. doi:10.1145/3078468.3078497. <http://doi.acm.org/10.1145/3078468.3078497>.
- Hohpe, Gregor, and Bobby Woolf. 2004. *Enterprise integration patterns: Designing, building, and deploying messaging solutions*. Addison-Wesley Professional.

Ishakian, Vatche, Vinod Muthusamy, and Aleksander Slominski. 2017. “Serving deep learning models in a serverless platform”. *CoRR* abs/1710.08460. arXiv: 1710.08460. <http://arxiv.org/abs/1710.08460>.

Jonas, Eric, Shivaram Venkataraman, Ion Stoica, and Benjamin Recht. 2017. “Occupy the Cloud: Distributed Computing for the 99%”. *CoRR* abs/1702.04024. arXiv: 1702.04024. <http://arxiv.org/abs/1702.04024>.

Kuhlenkamp, Jörn, and Markus Klems. 2017. “Costradamus: A Cost-Tracing System for Cloud-Based Software Services”. In *Service-Oriented Computing*, edited by Michael Maximilien, Antonio Vallecillo, Jianmin Wang, and Marc Oriol, 657–672. Cham: Springer International Publishing. ISBN: 978-3-319-69035-3.

Leitner, P., J. Cito, and E. Stöckli. 2016. “Modelling and Managing Deployment Costs of Microservice-Based Cloud Applications”. In *2016 IEEE/ACM 9th International Conference on Utility and Cloud Computing (UCC)*, 165–174.

Lloyd, Wes, Shruti Ramesh, Swetha Chinthalapati, Lan Ly, and Shrideep Pallickara. 2018. “Serverless Computing: An Investigation of Factors Influencing Microservice Performance”. *The IEEE International Conference on Cloud Engineering (IC2E)*. Forthcoming.

Lynn, Theo, Pierangelo Rosati, Arnaud Lejeune, and Vincent Emeakaroha. 2017. “A Preliminary Review of Enterprise Serverless Cloud Computing (Function-as-a-Service) Platforms”. In *2017 IEEE International Conference on Cloud Computing Technology and Science (CloudCom)*, 162–169. IEEE.

Malawski, Maciej, Adam Gajek, Adam Zima, Bartosz Balis, and Kamil Figiela. 2017. “Serverless execution of scientific workflows: Experiments with HyperFlow, AWS Lambda and Google Cloud Functions”. *Future Generation Computer Systems*. ISSN: 0167-739X. doi:<https://doi.org/10.1016/j.future.2017.10.029>. <http://www.sciencedirect.com/science/article/pii/S0167739X1730047X>.

McGrath, G., and P. R. Brenner. 2017. “Serverless Computing: Design, Implementation, and Performance”. In *2017 IEEE 37th International Conference on Distributed Computing Systems Workshops (ICDCSW)*, 405–410. doi:10.1109/ICDCSW.2017.36.

- Nastic, S., T. Rausch, O. Scekcic, S. Dustdar, M. Gusev, B. Koteska, M. Kostoska, B. Jakimovski, S. Ristov, and R. Prodan. 2017. "A Serverless Real-Time Data Analytics Platform for Edge Computing". *IEEE Internet Computing* 21 (4): 64–71. ISSN: 1089-7801. doi:10.1109/MIC.2017.2911430.
- Pozdniakova, Olesia, and Dalius Mazeika. 2017. "Systematic Literature Review of the Cloud-ready Software Architecture". *Baltic Journal of Modern Computing* 5 (1): 124.
- Roberts, Mike. 2006. "Serverless Architectures". Visited on February 1, 2017. <https://martinfowler.com/articles/serverless.html>.
- Sbarski, Peter, and S Kroonenburg. 2017. *Serverless Architectures on AWS: With examples using AWS Lambda*. Manning Publications, Shelter Island.
- Spillner, Josef. 2017a. "Exploiting the Cloud Control Plane for Fun and Profit". *CoRR* abs/1701.05945. arXiv: 1701.05945. <http://arxiv.org/abs/1701.05945>.
- . 2017b. "Snafu: Function-as-a-Service (FaaS) Runtime Design and Implementation". *CoRR* abs/1703.07562. arXiv: 1703.07562. <http://arxiv.org/abs/1703.07562>.
- . 2017c. "Transformation of Python Applications into Function-as-a-Service Deployments". *CoRR* abs/1705.08169. arXiv: 1705.08169. <http://arxiv.org/abs/1705.08169>.
- Spillner, Josef, Cristian Mateos, and David A. Monge. 2018. "FaaSter, Better, Cheaper: The Prospect of Serverless Scientific Computing and HPC". In *High Performance Computing*, edited by Esteban Mocoskos and Sergio Nesmachnow, 154–168. Cham: Springer International Publishing. ISBN: 978-3-319-73353-1.
- Varghese, Blessen, and Rajkumar Buyya. 2018. "Next generation cloud computing: New trends and research directions". *Future Generation Computer Systems* 79:849–861. ISSN: 0167-739X. doi:<https://doi.org/10.1016/j.future.2017.09.020>. <http://www.sciencedirect.com/science/article/pii/S0167739X17302224>.

Villamizar, Mario, Oscar Garces, Lina Ochoa, Harold Castro, Lorena Salamanca, Mauricio Verano, Rubby Casallas, Santiago Gil, Carlos Valencia, Angee Zambrano, et al. 2016. “Infrastructure cost comparison of running web applications in the cloud using AWS lambda and monolithic and microservice architectures”. In *Cluster, Cloud and Grid Computing (CC-Grid), 2016 16th IEEE/ACM International Symposium on*, 179–182. IEEE.

Yan, Mengting, Paul Castro, Perry Cheng, and Vatche Ishakian. 2016. “Building a Chatbot with Serverless Computing”. In *Proceedings of the 1st International Workshop on Mashups of Things and APIs*, 5:1–5:4. MOTA '16. Trento, Italy: ACM. ISBN: 978-1-4503-4669-6. doi:10.1145/3007203.3007217. <http://doi.acm.org/10.1145/3007203.3007217>.