



UNIVERSITY OF APPLIED SCIENCES

DEPARTMENT OF COMPUTER SCIENCE

Bachelor Thesis

Vergleich eines Usecases mit Serverless Technologie gegenüber Spring Boot Technologie am Beispiel von Instant Payments

Eingereicht am:

25. August 2021

Eingereicht von:

Silas Hoffmann

Traberweg 52

22159 Hamburg

Tel.: (040) 643 94 73

E-mail: inf103088@stud.fh-wedel.de

Referent:

Prof. Dr. Dennis Säring

Fachhochschule Wedel

Feldstraße 143

22880 Wedel

Phone: (041 03) 80 48-43

E-mail: dennis.saering@fh-wedel.de

Betreut von:

Kai Roßdeutscher

DPS Engineering GmbH

Eiffestraße 78

20537 Hamburg

Phone: (040) 25 15 41-44

E-mail: kai.rossdeutscher@dps.de

Inhaltsverzeichnis

Abbildungsverzeichnis	IV
List of Listings	V
1 Einleitung	1
1.1 Überblick	1
1.2 Motivation	1
2 Zielsetzung	3
2.1 Problemstellung	3
2.2 Lösungsweg	3
3 Ist-Analyse	5
3.1 JBoss (Microprofile)	5
3.2 Probleme	5
4 Vorgehensmodell	7
4.1 Anforderungen an Daten zur Messung des Startup-Verhaltens von Containern	7
4.2 Anforderungen an Prototypen	11
4.2.1 Festlegung fiktiver Workflow	11
4.2.2 Serverless	12
4.3 Anforderungen an Containerisierungsplattform	12
4.4 Anforderungen an Lasttest	12
4.5 Anforderungen Visualisierung und Monitoring zur Unterstützung der Auswertung . .	12
5 Problemlösung	13
5.1 Bestimmung von Daten zur Messung des Startup-Verhaltens von Containern	13
5.1.1 Kriterienkatalog	13
5.2 Implementierung Prototyp	13
5.2.1 Node.js	13
5.2.2 Spring Boot	13
5.3 Implementierung mittels Containerisierungsplattform	13
5.3.1 Container Lifecycle	13
5.3.2 Docker Swarm	14
5.4 Implementierung Lasttest	15
5.4.1 Timeline	15
5.4.2 Testbedingungen	15
5.5 Implementierung Visualisierung und Monitoring zur Unterstützung der Auswertung .	16
6 Ergebnisanalyse	18
6.1 Ergebnisse	18
6.2 Analyse	18
6.3 Diskussion	18
6.3.1 Begründung Startupzeit	18
7 Zusammenfassung	19
8 Ausblick	20

Inhaltsverzeichnis

9 Literaturverzeichnis	21
Literaturverzeichnis	22
10 Eidesstattliche Erklärung	23

Abbildungsverzeichnis

4.1	Agile Testing Quadrants	8
4.2	iso25010	8
5.1	Komponenten-Stack im Überblick	14

List of Listings

1

Einleitung

1.1 Überblick

- Was gibt es bisher, nur kurz anreissen
- wird im Grundlagen Kapitel tiefergehender behandelt...

Google, Spotify oder Netflix: Immer mehr Unternehmen setzen heutzutage auf Virtualisierungsplattformen. Aber warum sind Technologien wie zum Beispiel Docker als Containerisierungsplattform, oder Kubernetes als Orchestrator, so beliebt? In der folgenden Thesis möchte ich hierauf eingehen. Ich werde einen Überblick über die verfügbaren Technologien geben und auf diverse Eigenarten eingehen (was heißt es serverless zu deployen etc.).

1.2 Motivation

- Problemstellung kurz beschreiben
- DPS möchte untersuchen ...
- portierung der anwendung in die cloud (amazon) Problem ist: spring boot anwendung benötigt viel startup-zeit, heute ist die maximale skalierung in produktion vorbestimmt und muss immer vorgehalten werden, alle server sind 24x7x365 aktive und werden eigentlich nur in 1-2h pro tag in den spitzenzeiten ausgelastet... also nicht nur ineffiziente nutzung von hardware und energie sondern auch von kapital. -> hyperscaler geschäftsmodell verspricht abhilfe...
- hier ruhig auch einen Satz zur neuen Sau, die jetzt durchs dorf getrieben wird „new green economy“.. ressourcen sparen um eisbären und gretha thunberg zu retten... energieeffizienz...etc bla bla bla sofort-ready vollautomatische Skalierung etc..das sind dann die gganzen Versprechungen der hyperscaler zu der cloud...kannst du hier alle auflisten. Kunde möchte vorbereitet sein, um ggf. Auflagen der Regulierer zu „green“ erfüllen zu können.
- Welche Erkenntnisse sollen gewonnen werden?
- Kriterienkatalog nennen

Mit den neuen Möglichkeiten die diese neuartigen Technologien mit sich bringen, möchte sich die DPS GmbH in Zukunft ebenfalls in diese Richtung orientieren, denn die Ansprüche an eine moderne IT-Infrastruktur ändern sich. Man möchte effizienter arbeiten. Nicht nur bezogen auf verbrauchte Energie, sondern zum Beispiel auch auf nötiges Fachpersonal zur Wartung der Artifacts in Produktion. Denn diese Technologien ermöglichen ein Abstraktionslevel, welches es dem Anwendungsentwickler ermöglicht direkt Arbeitspakete für das Deployment zu generieren. Die Grenze zwischen der Operation- und der Entwicklungsabteilung verschwimmen immer mehr und es kommt zu weniger Tradeoffs zwischen mehreren Teams. Auch ermöglicht diese Abstraktion, falls gewünscht, ein komplettes Auslagern der Infrastruktur, wie es zum Beispiel mit Cloud Technologien der Fall ist.

1 Einleitung

Diese Abstraktion ermöglicht den Entwicklern sich mehr auf das Schreiben der Businesslogik zu konzentrieren und können durch das Nutzen von standardisierten Schnittstellen ein vollautomatisiertes Deployment gewährleisten. Dieses Prinzip wird je nach Abstraktionstiefe auch als *Infrastructure as Code* (IaC) beziehungsweise *Infrastructure as a Service* (IaaS) bezeichnet. Gerade IaC erlaubt den Entwicklern ihre Programmierkenntnisse auf die Konfiguration von Infrastruktur zu übertragen. Auch die DPS plant in geraumer Zukunft auf Cloud Technologie zu setzen, langfristig ist hierbei eine Portierung in die Cloud von Amazon (AWS) geplant.

Ein Fokus, welchen ich mir in dieser Thesis im Detail anschauen möchte betrifft das Startup-Verhalten einer Anwendung in genau solch einer virtualisierten Umgebung. Denn unter den genannten Effizienz Gesichtspunkten möchte man heutzutage auch nötige Ressourcen vor allem auf Anfrage verwenden und nicht mehr rund um die Uhr laufen lassen, selbst wenn dies zu einem gegebenen Zeitpunkt eigentlich gar nicht nötig wäre. Gerade im Banking Bereich gibt es Zeiträume in denen ein relativ geringer Geldfluss festzustellen ist, während es zu anderen Zeitpunkten zu regelrechten Bursts kommen kann, wenn zum Beispiel zu Feiertagen relativ viel Geld den Besitzer wechselt. Um dieses Prinzip der Ressourcennutzung auf Anfrage etwas anschaulicher zu gestalten, werde ich eine vereinfachte Kopie einer realen Anwendung vom Unternehmen nachbauen und hinsichtlich der Startzeiten von Containern untersuchen. Der Prototyp beinhaltet mehrere verschiedene Komponenten zur Abarbeitung der Logik, um im Nachhinein auf die Performanz hinsichtlich der genannten Effizienz Gesichtspunkte zu untersuchen und zwischen den verwendeten Technologien zu vergleichen.

2

Zielsetzung

Nachdem im letzten Kapitel ein kurzer Einblick in die Motivation für die Portierung alter JEE Technologie in neuere IT-Infrastruktur beschrieben wurde, wird im folgenden darauf eingegangen, welcher wesentlichen Problemstellung sich die DPS gegenüber konfrontiert sieht und wie ich versuchen werde dies zu untersuchen.

2.1 Problemstellung

- Vergleich der Startup-Zeit einer containerbasierten Cloudanwendung bei steigender Last. Es soll auf Unterschiede zwischen Spring und Node eingegangen werden.
- insbesondere auf Durchsatz der Instant Payments eingehen
- da anwendung in spring boot ist nicht cloud-fähig weil zu lange für startup, dadurch riskieren wir timeouts also rejects – das ist die besonderheit der anwendung. ...neuer container muss sofort verfügbar sein! Eine alternative wird gesucht: serverless verspricht minimale startup zeiten!....

Da das Unternehmen im Banking Bereich tätig ist, stellt eine hohe Verfügbarkeit der Komponenten eines der wesentlichen Kernziele dar. Auf die neuen Technologien bezogen, bedeutet dies, dass es zu ermitteln gilt wie viel Zeit ein Skalierungsprozess in einem orchestrierten Komponentenstack im Detail braucht. Dazu müssen diverse Metriken erhoben werden, welche im Kapitel ... beschrieben werden. Wichtig an dieser Stelle, es soll vorallem untersucht werden, inwiefern eine Java-Anwendung als cloudfähig bezeichnet werden kann. Denn wenn sie zu lange zum Starten braucht, werden timeouts riskiert, was es unter allen Umständen zu vermeiden gilt. Eine Alternative versprechen serverless Technologien, welche sich vorallem durch ihre minimale Startzeit für diesen Usecase qualifizieren. Um einen konkreten Vergleich durchführen zu können, wird im zu implementierenden Prototypen jeweils eine Komponente mit dem Java Backend Framework „*Spring Boot*“ und eine weitere Komponente mit dem Javascript Framework „*Node.js*“ implementiert.

Kapitel
Referenz
einbinde

2.2 Lösungsweg

- Prinzip / Ablauf erklären
- Grundlegende Struktur des Prototypen erklären
- vergleich der startupzeiten zwischen serverless und aktueller springboot variante anhand eines jeweils protypen und fiktiven workflows und fiktiven lastszenarios in einer cloud umgebung und messung des startup verhaltens bei plötzlich auftetenen spitzen oder so...

Um festzustellen wie gut eine Spring- beziehungsweise eine Node.js-Komponente tatsächlich geeignet ist, wird ein fiktiver Workflow mittels eines Komponentenstacks implementiert, der in der Lage ist verschiedene Lastszenarien abarbeiten zu können. Hierbei ist es dem Benutzer möglich dynamisch zwischen der Spring- und Node-Komponente zu wechseln um einen entsprechenden Vergleich anstellen zu können. Es werden diverse Skripte bereitgestellt die es dem Benutzer erlauben beliebig komplexe Lasten zu simulieren. Um eine Auswertung auch grafisch aufarbeiten zu können, wird der Stack ebenfalls mit Komponenten zur Visualisierung ausgestattet. Diese Elemente zur Visualisierung greifen dabei nicht in den Ablauf des restlichen Stacks ein, sondern greifen lediglich auf bereits bestehende Schnittstellen zurück um einen möglichst ungestörten Ablauf und damit unverfälschte Datensätze zu garantieren.

Der wesentliche Usecase des Systems wird die Verarbeitung von generierten Payments sein, welche nach mehreren fiktiven Arbeitsschritten Datensätze generieren. Der genaue Gegenstand der Bearbeitung ist allerdings eher nebensächlich, da er so gut wie keinen Einfluss auf die für diesen Usecase interessanten Metriken hat.

[Knu98]

3

Ist-Analyse

Nachdem im letzten Kapitel ein grober Überblick über die bestehende Problemstellung sowie den geplanten Lösungsweg gegeben wurde, folgt ein kurzer Einblick in die bestehenden Technologien um die Sinnhaftigkeit neuerer Technologien zu beleuchten.

3.1 JBoss (Microprofile)

- aktuelle Architektur beschreiben
- Hinweis darauf geben, dass Prototyp in der Thesis vereinfacht mit Spring dargestellt wird
- jetzt ja fiktiv SpringBoot, hier rein technischer Ist-Stand, Systemarchitektur. . .

Die aktuelle Verarbeitung von Payments innerhalb der Anwendung läuft in Produktion auf vier Instanzen des kommerziellen Applikation-Servers „JBoss“. Im Development wird hierbei eine Open-Source Variante namens „Wildfly“ verwendet. In diesen Application Servern werden entsprechende .war Dateien deployed welche den ausführbaren Code der Anwendung beinhalten. Der Application Server bietet dem Server-Teil der Client-Server-Anwendung eine Laufzeitumgebung, in der dieser ausgeführt werden kann. JBoss bietet nun standardisierte Schnittstellen nach dem jeweiligen Java Enterprise Standard um zum Beispiel die Kommunikation mit der Außenwelt zu ermöglichen oder um der Anwendung eine Persistenzschicht zur Verfügung zu stellen. Außerdem laufen die Instanzen im *Microprofile* Modus, der es ermöglicht, bestimmte Konfigurationsparameter auszulagern. So können Applikationen auf unterschiedlichen Systemen deployed werden ohne komplett neu gebaut zu werden ([mic]). Um eine gleichmäßige Aufteilung der Last zu gewährleisten, teilt ein so genannter „Load Balancer“ die eingehenden Nachrichten den entsprechenden Instanzen zu. Jede der vorhandenen Instanzen besitzt eine minimale sowie maximale Anzahl an parallel ausführbaren Prozessen. Diese Angaben werden auch „*max. / min. Poolsize*“ genannt. Eine minimale Poolsize muss gegeben sein, um sicherzustellen, dass eine gewisse Grundlast falls nötig sofort bearbeitet werden kann, daher darf diese minimale Anzahl auch nicht Null betragen. Die maximale Poolsize stellt sicher, dass es zu keiner Überlastung des Systems kommt. Wenn eine Instanz bereits mit der maximale Anzahl an Prozessen arbeitet, wird dies dem Loadbalancer signalisiert und dieser teilt der entsprechenden Komponente in diesem Zeitraum keine weiteren Nachrichten mehr zu. Um zu gewährleisten, dass die Nachrichten nicht verloren gehen, werden sie in eine „*Request Queue*“ geschrieben, welche lediglich dazu gedacht ist den Overhead abzuspeichern. Wie die Daten im Detail verarbeitet werden, ist für die weitere Betrachtung irrelevant und wird daher nicht weiter erläutert.

3.2 Probleme

- Probleme mit aktuellem System (Stichwort Deployment, Wartbarkeit)

Simple
Schaubil
einfügen

Quelle
einfügen

3 Ist-Analyse

- Prof. hat extra darauf hingewiesen, dass es nicht nur um die Vorteile der Cloud gehen soll
- „Erwartete Probleme in einer Cloud Umgebung“ à aktuell „...starre Hardware und Software-Skalierung...“ unerwartete Lastspitzen könnten zu Problemen führen, wenn sie die erwartete und verfügbare Obergrenze an Kapazität übersteigt und wie oben gesagt ineffiziente Nutzung von Kapital... nochmal mit anderen Worten aus
- 1.2 à die Auflistung der Probleme hier, müssen dann in der Zusammenfassung wieder auftauchen und abgehakt werden! Die wollen wir ja auch lösen...

Mit der monolithischen Struktur des aktuellen Systems kommt es im Laufe der Zeit zu unterschiedlichen Problemen. Über die Zeit werden Komponenten immer weiter verstrickt, sodass es in einem solchen System auch nach Personalwechsel nicht einfacher wird den Überblick über die gesamte Applikation zu behalten. Dies führt zu fehleranfälligerem Code, welcher sich wiederum negativ auf das Kapital einer Firma niederschlägt. Diese Fehler führen in einem monolithischen System zu einem großräumigen Ausfall der gesamten Applikation, da es keine klare Abgrenzung der einzelnen Komponenten gibt.

Außerdem laufen die Application Server auf Servern, die (wenn überhaupt) eine sehr dünne Abstraktionsschicht bieten („starrer Hardware“). Es ist nicht möglich diese schnell und flexibel zu ersetzen, falls es zu Fehlern in der Produktionsphase kommen sollte. Application Server bringen einen großen Konfigurationsaufwand mit sich, der sich schlecht automatisieren lässt. Man braucht geschultes Fachpersonal um ein solches System zum Laufen zu bekommen.

Eine Skalierung ist mit monolithischen Strukturen nur dadurch möglich, den ausführbaren Code auf zusätzlichen Servern zu deployen, dies wird auch *horizontale Skalierung* genannt. Jede dieser Kopien nutzt die gleiche Ressourcenanzahl, was es zu einem ineffizienten Design macht, da sie sich nicht dynamisch der gegebenen Last anpassen. Tatsächlich ist es jedoch ein viel größeres Problem, wenn ein bereits aufgeteiltes System an seine Kapazitätsgrenze stößt, denn wie bereits erwähnt, besitzen Java-Anwendungen eine relativ lange Initialisierungsphase. Um den heutigen Anforderungen der dynamischen Skalierung gerecht zu werden, werden Ressourcen auf Abruf gebraucht. Die Systeme sollen so schnell wie möglich verfügbar sein.

4

Vorgehensmodell

Nachdem im letzten Kapitel der aktuelle Aufbau der Payment-Anwendungen, sowie die durch die monolithische Struktur hervorgebrachten Probleme, erläutert wurden, werde ich im Folgenden beschreiben wie ich vorgehen werde um den Problemfokus der Startzeit zu untersuchen.

4.1 Anforderungen an Daten zur Messung des Startup-Verhaltens von Containern

- Problem reduzieren
- Messkriterien festlegen
- Quellen finden
- hier nur theoretisch schwafeln über einen Kriterienkatalog und die möglichen „Messwerte“, die man ggf. braucht -> in Kap. 5 dann den Kriterienkatalog konkret aufstellen

Der Fokus des Projekts liegt allerdings auf der Ermittlung der Cloudfähigkeit der unterschiedlichen System-Komponenten. Um dies festzustellen, muss im Vorwege eine Klassifizierung der Anforderungen stattfinden um festzustellen welche Messdaten für dieses Projekt überhaupt von Relevanz sind. Hierfür hat die ISO bereits einen ausführlichen Kriterienkatalog aufgestellt (siehe ISO/IEC 25010:2011 [iso11], fig. 4.2). Die ISO 25010 besteht aus acht Hauptkategorien und 31 Unterkriterien, wobei lediglich die erste Hauptkategorie der „*Funktionalität*“ dabei als *funktionale Qualitätseigenschaft* (engl. „*functional requirement*“) auftritt. Die restlichen Hauptkriterien lassen sich der Kategorie der *Non-Functional requirements* zuordnen. Diese Unterscheidungen dienen als eine erste Abstraktionsschicht. Functional requirements beschreiben hierbei die Korrektheit des Produktes [BA18]. „Sie spezifizieren die Funktion, die ein System oder Systemkomponente erfüllen muss“ [Eid05]. So lässt sich feststellen ob die Applikation den inhaltlichen Ansprüchen, sowohl aus technischer als auch aus Unternehmenssicht, genügen. Hierbei wird auf eine Vielzahl unterschiedlicher Tests zurückgegriffen, welche sich auf breit gefächerte Abstraktionstiefen etc. beziehen lassen (siehe 4.1). Je nach Teilbereich lassen sich hierbei einzelne Gebiete automatisiert oder mittels anderweitiger Werkzeuge testen. Diese Tests sind in einem System in Produktion unabdingbar, für die Messdatenerhebung zur Beurteilung der Cloudfähigkeit allerdings vernachlässigbar.

Zur Eingrenzung der Anforderungen an die Messdaten sind viel mehr die *Non-Functional Requirements* von Bedeutung. Sie beschreiben nicht *was* vom System geleistet werden muss, sondern *wie* dies geschehen soll, deshalb werden die Kriterien unter anderem auch *system-quality attributes* genannt. Im Folgenden werden die für das Projekt relevanten Kriterien einmal zusammengefasst dargestellt.

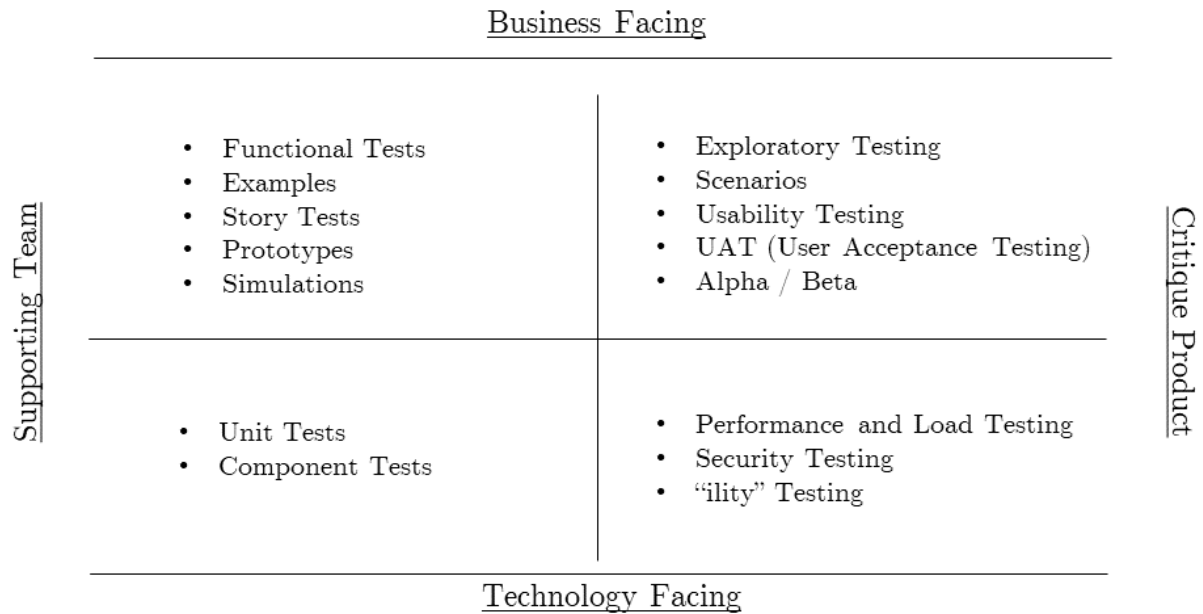


Abbildung 4.1: Agile Testing Quadrants

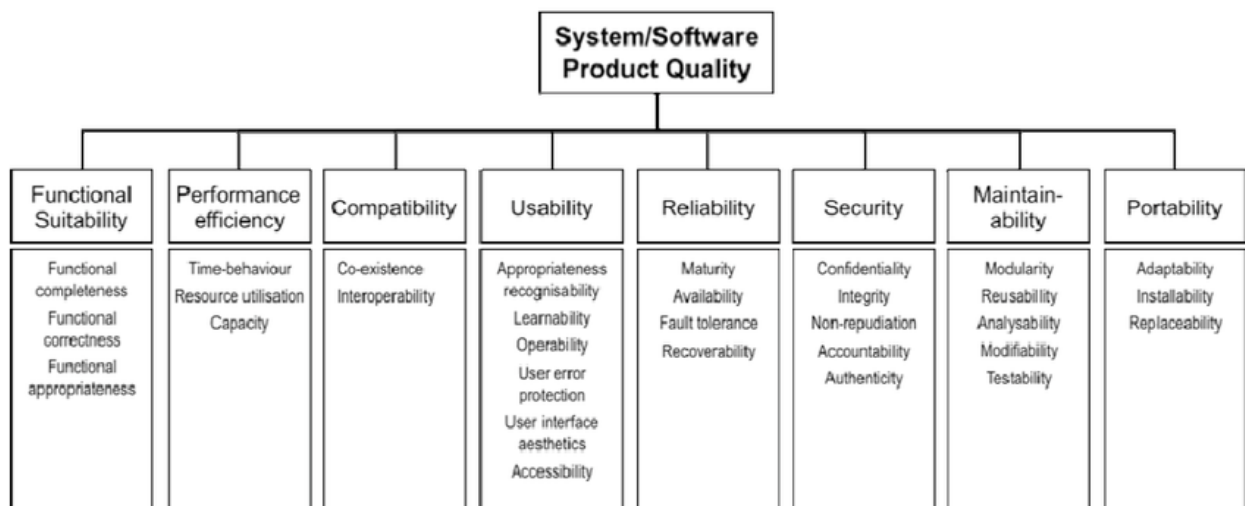


Abbildung 4.2: ISO 25010

Leistungsfähigkeit

- „Performance of a product or an app defines how a product/app is performing or behaving as compared to its expected behavior“
- Reaktionszeiten, wie schnell ist app wenn User interagiert
 - eng verwand *Latenz* / *Delay*, wie lange darf ein User max. warten müssen (Threshold ist die Reaktionszeit)
 - typischerweise pro Screen gesetzt, kann aber auch auf die Masse bezogen werden (bspw. 80 Prozent der Interaktionen haben eine Reaktionszeit von max. 3 Sekunden)

4 Vorgehensmodell

- Durchsatz: Anzahl der Daten, welche die Applikation handeln kann
 - kann sich sowohl auf Benutzeranfragen, db Anfragen, API calls etc. beziehen
 - Durchsatzziele als Teil der Anforderungsanalyse
 - werden mittels Performance / Load Testing validiert
- tools wie gatling / Jmeter fuer diesen Usecase unpassend, da nicht die Api Anfrage sondern nur die Hochfahrzeit gemessen werden soll
-

„Performance of a product or an app defines how a product/app is performing or behaving as compared to its expected behavior“ [nfr]. Die Leistungsfähigkeit einer Anwendung setzt sich aus den Punkten *Zeitverhalten*, *Ressourcennutzung* sowie *Kapazität* zusammen. Das Zeitverhalten wird vorallem durch die Reaktionszeit der Applikation geprägt. Es wird gemessen wie schnell auf Benutzereingaben eingegangen werden kann. Eng verwand ist hierbei auch der Begriff der *Latenzzeit*, da dieser beschreibt wie lange eine Beantwortung der Anfrage in Praxis letztendlich dauert. Die Reaktionszeit dient hierbei als ein Threshold der nicht überschritten werden sollte. Typischerweise wird die Reaktionszeit pro Client festgelegt, es ist jedoch auch möglich, dies für eine Menge von mehreren Clients festzulegen. Ein Beispiel wäre festzulegen, dass ein bestimmter Anteil der Interaktionen eine maximale Reaktionszeit von n Sekunden besitzt. Aber auch das zweite Unterkriterium der *Ressourcennutzung* darf nicht vernachlässigt werden, so dürfen zum Beispiel bestimmte Kommunikationsschnittstellen nicht überlastet werden oder diverse Connectionspools nicht überlaufen, da es sonst zu Timeouts kommen könnte. Die Kapazität beschreibt zum Beispiel die Anzahl von möglichen parallelen Anfragen oder wie viele Nachrichten übermittelt beziehungsweise vom System gespeichert werden können. Ein weiterer Unterpunkt, welchen man der Leistungsfähigkeit zuschreiben kann, ist der des *Durchsatzes*. Dieser Begriff wird in der ISO Norm allerdings nicht erwähnt. Im Idealfall möchte man nicht nur feststellen, wie lange Nachrichten verarbeitet werden, sondern auch wie viele im Schnitt verarbeitet werden können. Der Datendurchsatz ist also durchaus auch von größerem Interesse und sollte definitiv gemessen werden. Die verarbeiteten Daten können sich hierbei auch auf verschiedenes beziehen. Es können zum Beispiel die Anzahl von Anfragen an eine API sowie deren Rückmeldung sein, oder diverse Messages an Broker, welche durch das System gereicht werden. Es kann sich aber auch ganz banal auf die Anzahl der Datenbankanfragen etc. beziehen. Diese Art der Performanz wird in der Regel mittels spezieller Tools, wie zum Beispiel „Gatling“ oder „JMeter“ in sogenannten *Performance* oder auch *Load Tests* validiert. Alles in allem stellt die Leistungsfähigkeit mit ihren Subkategorien eine der wohl wichtigsten Kriterien dar.

Skalierbarkeit

Hardware

- je nach Anforderungen hoch oder herunterfahren
- *concurrent users* : Wie viele Benutzer verwenden zeitgleich die Applikation. I.d.R. 20 - 30 Prozent der Userbase. Kann auf 100 Prozent bei besonderem Event hochgehen
- Zeitfaktor nicht vergessen, Applikation Skalierung sollte auch zeitlich gesteuert werden
- *vertikale Skalierung / scaling up* : Mehr CPUs / RAM fuer existierende Infrastruktur
- *hori Skalierung / scaling out* : mehr Server
- vielleicht Bild einfüegen scale out vs scale up

- cloud provider verwenden generell scale out

Software

- Bsp. Ladezeiten einer single page application zu hoch (z.B. Laden von Daten aus DB), scale out bringt hierbei nichts...
- Performanz sollte von Anfang an bedacht werden
- Frameworks sollten auf multithreading etc. ausgelegt sein, vor Verwendung einmal überprüfen
- Bottlenecks früh bestimmen, Legacy Libs eventuell austauschen
- Load Balancer sollte in irgendeiner Art vorhanden sein um Hardware Skalierung mit Software in Verbindung zu bringen
- Ohne Session Management ist die Performance schlecht, gilt es ebenfalls im Vorwege zu prüfen

Ein weiteres wichtiges Kriterium welches allerdings nicht direkt in der ISO 25010 Norm auftritt, ist die Skalierbarkeit des Systems. Bei der Beurteilung kommt es in erster Linie zwischen dem Skalieren der *Hardware* sowie der *Software* zu unterscheiden [nfr]. Die Hardwareskalierung könnte man grob dem Unterkriterium „*Ressourcennutzung*“ zuordnen. Denn in einer modernen containerisierten Cloud-Umgebung soll es möglich sein, die einzelnen Komponenten nicht nur auf einem Gerät beliebig zu skalieren, sondern falls die aktuelle Hardware an ihre Kapazitätsgrenzen stößt auch noch weitere Hardwarekomponenten hinzuziehen können. Inwiefern dies automatisiert oder manuell geschieht lässt sich nur fallabhängig bestimmen, es sollte jedoch evaluiert werden ob dies überhaupt generell möglich ist. Metriken zum Ermitteln des Skalierungszeitpunkt könnten zum Beispiel die Anzahl der parallel arbeitenden Benutzer der Applikation (engl. „*concurrent users*“), die Latenzzeit oder die CPU Auslastung der beteiligten Maschinen sein. Es ist allerdings auch wichtig, dass eine Applikation nicht nur von aktiv eingehenden Ereignissen skaliert wird, sondern auch ein Zeitverhalten mit sich bringt und sich selbstständig zeitabhängig skalieren kann. Dies macht den Skalierungsprozess robuster, da sich die Komponente dadurch wieder etwas von seiner Umwelt abgrenzen kann (welche in der Regel durchaus fehlerbehaftet sein kann). Bei der Hardwareskalierung gibt es zwei wesentliche Arten, die *vertikale* und die *horizontale Skalierung*. Der Begriff der *vertikalen Skalierung* (auch *scale up* Skalierung) beschreibt das modifizieren der Servereigenen Ressourcen. Dazu können zum Beispiel die Anzahl der CPU Kerne oder der verfügbare Speicher gehören. Dieses Verfahren wird in monolithischen Strukturen mit älteren Systemen verwendet, die nicht darauf ausgelegt sind, sich auf mehrere Server zu verteilen. Dagegen steht der Begriff der *horizontalen Skalierung* (auch *scale out* Skalierung), welcher gerade dieses Auslagern bestimmter Komponenten beziehungsweise Prozessinstanzen auf neue Hardware beschreibt. Dieser Ansatz wird in der Cloud-Technologie verfolgt und nicht nur auf *on premise*, sondern auch von den großen *public cloud vendors* wie zum Beispiel Amazon AWS oder Microsoft Azure, verwendet.

Dennoch muss die Software auf eine Skalierung der Hardware ausgelegt sein. Manche Probleme lassen sich mit mehr Rechenperformance nicht lösen, wenn zum Beispiel die Ladezeiten bestimmter Kommunikationsschnittstellen einen Bottleneck darstellen. Alter Legacy Code könnte ebenfalls Performanceeinbußen darstellen und ohne ein vernünftiges Session Management bringt auch das Skalieren der Hardware nichts. Diese Grundfunktionalitäten werden heutzutage jedoch von fast allen Frameworks standardmäßig mitgeliefert. Dennoch gilt es dies im Vorwege noch einmal zu validieren.

Footnote
on premise

footnote
public
cloud vendor

Kompatibilität

Dieses Hauptkriterium beschreibt, wie nebenläufig das System im Endeffekt ist. Ist es möglich, dass mehrere Komponenten gleichzeitig arbeiten können, ohne sich gegenseitig zu behindern? Gerade in einer Containerisierten Umgebung ist dies von fundamentaler Wichtigkeit. Der Unterbegriff *Ko-Existenz* beschreibt gerade dieses Verhalten. Der zweite Unterbegriff der „*Interoperabilität*“ beschreibt inwiefern die verschiedenen Komponenten in der Lage sind Daten untereinander auszutauschen. Wie genau dieser Datenaustausch aussieht kann sich von System zu System unterscheiden. Zusammenfassend lässt sich sagen, dass diese Kriterien in einer cloudbasierten Anwendung zwar gelten müssen, sie allerdings dennoch derartig abstrakt gehalten sind, dass es schwieriger ist hier messbare Metriken zu generieren. Entweder das System funktioniert oder es ist fehlerhaft, einen messbaren Mittelweg gibt es nicht direkt.

Zuverlässigkeit

- High Availability
- Angabe wie lange app garantiert laeuft (bspw. Prozentangabe pro Jahr)
- Design for failure: Komponenten darauf ausgelegt, dass Ausfall einzelner Komponenten zu keinem Systemausfall fuehrt
- accessibility: app ist nicht für die Allgemeinheit konzipiert worden

Das Hauptkriterium der *Zuverlässigkeit* beschreibt die Angabe wie lange eine Applikation garantiert läuft. Dies kann auf verschiedene Weisen festgehalten werden, eine typische wäre zum Beispiel den Anteil der Tage im Jahr, Tage pro Monat, oder Stunden am Tag die ein Service ansprechbar ist. Allgemein lässt sich diese Angabe als Metrik der *Verfügbarkeit* charakterisieren. In einer Containerisierten Umgebung kommt es immer wieder zu Ausfällen einzelner Systeme, was es bei diesem Punkt besonders zu beachten gilt. Das Gesamtsystem soll hiervon jedoch unberührt bleiben. Dieses Prinzip wird in heutigen Systeme auch „*Design for failure*“ genannt. Das Hauptkriterium setzt sich aus unter anderem aus dem Unterkriterium der *Reife* zusammen. Hierbei sind diejenigen Anforderungen gemeint, welche in einer normalen Prozessabarbeitungsphase ohne besondere Ereignisse hinsichtlich zum Beispiel der Anzahl der zu verarbeitenden Nachrichten, oder der Komplexität der zu bewältigenden Aufgaben, gelten müssen. Auch eine gewisse *Fehlertoleranz* wird vom System erwartet. Um dies zu testen genügt es bereits dokumentierte allerdings fehlerbehaftete Benutzereingaben in einer Testumgebung zu tätigen und das Systemverhalten zu analysieren. Eine standardisierte Metrik gibt es auch nicht. Das letzte Unterkriterium der *Zuverlässigkeit* ist die *Wiederherstellbarkeit* nach einem beschriebenen Fehlerfall, falls es doch einmal zu einem Ausfall einzelner Komponenten kommen sollte. Generell gilt, auch dieses Hauptkriterium gibt Aufschluss über die Eignung gewisser technologischer Ansätze zum Arbeiten als Cloudkomponente, dennoch beschreibt sie die relevanten Umstände in manchen Teilen zu abstrakt als das man dort genaue Metriken ermitteln könnte. Im nächsten Kapitel wird festgehalten welche Metriken dennoch relevant erscheinen.

irrelevante Themenfelder

- Security (Injection, Broken Authentication, Cross Site Scripting etc.)
- Nutzbarkeit (Benutzerfreundlichkeit, Browser Kompatibilität, Internationalisierung - Eingabesprachen)

Ref einb
den bzw
schauen
ob über-
haupt
aufgenom
men...

Weitere Hauptkriterien der ISO Norm betreffen die *Nutzbarkeit*, sowie die *Sicherheit* der Anwendung. Da der Fokus des Projekts aber darin liegt die Skalierbarkeit sowie die Performance des Systems zu testen, können all diese Punkte jedoch vernachlässigt werden. Die Themengebiete der *Wartbarkeit* sowie der *Übertragbarkeit* sind in einer containerisierten Plattform zwar ebenfalls unabdingbar, sind allerdings derartig abstrakt gehalten, dass es den Rahmen dieser Thesis sprengen würde ebenfalls näher auf diese einzugehen.

4.2 Anforderungen an Prototypen

- einmal als spring boot und einer weiteren variante mit einer cloud-native technologie -> hier wurde vorgegeben mit serverless zu arbeiten.

4.2.1 Festlegung fiktiver Workflow

Da der Fokus auf der Untersuchung der Startupzeit der Komponenten liegt, wird lediglich eine minimale beispielhafte Implementierung erfolgen, welche die Arbeitsschritte der eigentlichen Applikation in vereinfacht darstellen soll. Allerdings werden im System konkrete Nachrichten im XML Format vermittelt, welche einer XSD-Spezifikation folgen wie sie im realen Umfeld ebenfalls genutzt wird.

Sobald eine neue Nachricht eingetroffen ist, soll drei Arbeitsschritte ausgeführt werden:

1. Es soll geprüft werden, ob das eingegangene XML der XSD-Spezifikation folgt oder nicht. Wenn dies nicht der Fall sein sollte, wird die Nachricht zwar derartig acknowledged, dass sie zwar aus der Eingangsqueue im Message Broker entfernt wird, allerdings bei der Verarbeitung ignoriert wird.
2. Falls es sich um valides XML handelt, wird ein Feld aus dem XML-Inhalt ausgelesen.
3. in einem letzten Schritt wird dieses Element in eine Datenbank geschrieben damit auch eine Persistenz-Operation in die Verarbeitungszeit einfließt.

Auszug
aus XML
einbin-
den???

4.2.2 Serverless

- 4.2.2 spring boot 4.2.3 serverless → so bekommt der prof schon beim draufgucken auf das inhaltsverzeichnis die story mit. . .

4.3 Anforderungen an Containerisierungsplattform

4.4 Anforderungen an Lasttest

4.5 Anforderungen Visualisierung und Monitoring zur Unterstützung der Auswertung

5

Problemlösung

5.1 Bestimmung von Daten zur Messung des Startup-Verhaltens von Containern

5.1.1 Kriterienkatalog

- Vorallem Skalierbarkeit und Performance
- Transaktionen - Durchsatz wichtig
- Deployment egal, da bereits im Container deployed wird
- Generell sagen, warum einige Aspekte egal sind
- Wartbarkeit nicht so wichtig
- Ressourcennutzung
- Störungsfälle (Chaos Monkey)
- Non Functionals - Kai hat PDF geschickt

5.2 Implementierung Prototyp

5.2.1 Node.js

5.2.2 Spring Boot

5.3 Implementierung mittels Containerisierungsplattform

5.3.1 Container Lifecycle

- Auf verschiedene Schichten eingehen
- Auf Ergebnisse beziehen

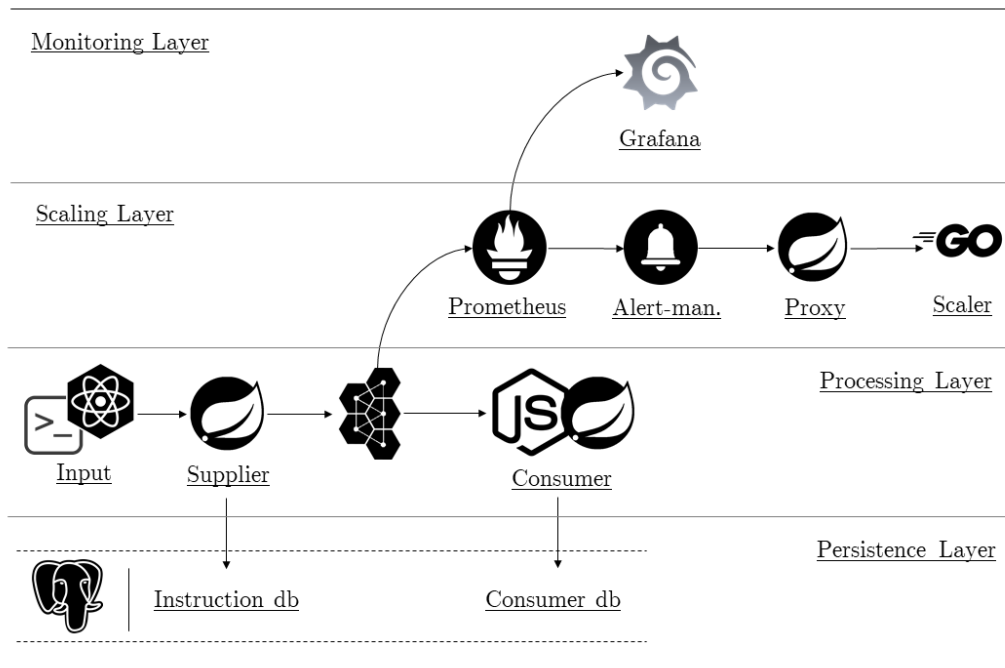


Abbildung 5.1: Komponenten-Stack im Überblick

5.3.2 Docker Swarm

- Prototypen im Detail erläutern

```
cat .env

# qbn: queue bound (level) n
# cbn: container bound (level) n

QB0=15
QB1=30
QB2=100
CB0=1
CB1=5
CB2=10
CB3=30
```

Tabelle 5.1: Server Specs

Prozessor	Intel(R) Xeon(R) Gold 6226R CPU @ 2.90GHz
Kerne	6 Prozessoren á 16 Kerne
RAM	16 GB
Storage	150 GB

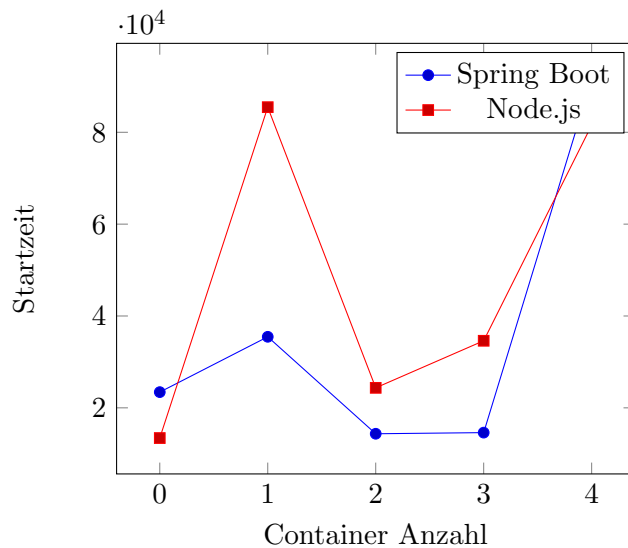
$\underline{QL3}$ $QB2 < \overline{MC}$	UP $abs(CB0 - CB3)$	UP $abs(CB1 - CB3)$	UP $abs(CB2 - CB3)$	OK –
$\underline{QL2}$ $QB1 < \overline{MC} \leq QB2$	UP $abs(CB0 - CB2)$	UP $abs(CB1 - CB2)$	OK –	DOWN $abs(CB2 - CB3)$
$\underline{QL1}$ $QB0 < \overline{MC} \leq QB1$	UP $abs(CB0 - CB1)$	OK –	DOWN $abs(CB1 - CB2)$	DOWN $abs(CB1 - CB3)$
$\underline{QL1}$ $QB0 < \overline{MC} \leq QB1$	UP $abs(CB0 - CB1)$	OK –	DOWN $abs(CB1 - CB2)$	DOWN $abs(CB1 - CB3)$
$\underline{QL0}$ $QB0 == \overline{MC}$	OK –	DOWN $abs(CB0 - CB1)$	DOWN $abs(CB0 - CB2)$	DOWN $abs(CB0 - CB3)$
	$\underline{CL0}$ $CB0 == \overline{MC}$	$\underline{CL1}$ $CB0 < \overline{MC} \leq CB1$	$\underline{CL1}$ $CB1 < \overline{MC} \leq CB2$	$\underline{CL1}$ $CB2 < \overline{MC} \leq CB3$

5.4 Implementierung Lasttest

5.4.1 Timeline

5.4.2 Testbedingungen

- Kommt in den Anhang
- hat Prof. zwar als eigenes Kapitel erwahnt, bin mir aber nicht sicher ob das wirklich noetig ist
- auf welcher Hardware werden Tests durchgefuehrt?
- chaos monkey / Stoerfaelle erlaeutern



5.5 Implementierung Visualisierung und Monitoring zur Unterstützung der Auswertung

```

1  // standard constructor
2  public Bank(int playerCnt) {
3      this.entries = new Entry[playerCnt];
4      this.bankSize = playerCnt;
5      this.rand = new Random();
6  }
7
8  // testing constructor - no fileIO
9  public Bank(Entry[] entries, Random pseudoRandom) {
10     this.entries = entries;
11     this.rand = pseudoRandom;
12     this.bankSize = entries.length;
13 }
14
15 // testing constructor - with fileIO
16 public Bank(String preallocation, List<Player> players, Random rand) {
17     assert null != preallocation && null != players && null != rand;
18     this.bankSize = players.size();
19     this.entries = new Entry[this.bankSize];
20     if (0 < preallocation.length()) {
21         String[] singleEntries = preallocation.split(SEPERATOR_STRING_REPRESENTATION);
22         int offset = this.bankSize - singleEntries.length;
23         for (int i = singleEntries.length - 1; i >= 0; i--) {
24             this.entries[i + offset] = new Entry(singleEntries[i], players);
25         }
26     }
27 }

```

```

1  #!/bin/bash
2
3  # Add two numeric value
4  ((sum=25+35))
5
6  #Print the result

```

7 | `echo $sum`

6

Ergebnisanalyse

6.1 Ergebnisse

- Vorstellung der erhaltenen Daten
- Interpretation / Analyse allerdings Teil vom naechsten Kapitel?

6.2 Analyse

- Interpretation / Analyse der Daten
- Begrueendung fuer Verhalten suchen

6.3 Diskussion

6.3.1 Begründung Startupzeit

- Warum Node.js schneller ist
- Erlaeutern warum die erhaltenen Ergebnisse in einem real-life Szenario vielleicht nicht aussagekraeftig sein koennten

7

Zusammenfassung

8

Ausblick

9

Literaturverzeichnis

Literaturverzeichnis

- [BA18] Daniel Bryant and Marin-Perez Abraham. *Continuous delivery in Java essential tools and best practices for deploying code to production*. O'Reilly, 2018.
- [Eid05] Petter L. H. Eide. Quantification and traceability of requirements. <https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.95.6464&rep=rep1&type=pdf>, 2005.
- [iso11] Iso/iec 25010:2011. <https://www.iso.org/obp/ui/#iso:std:iso-iec:25010:ed-1:v1:en>, 2011.
- [Knu98] Donald E. Knuth. *The Art of Computer Programming, Volume 3: (2Nd Ed.) Sorting and Searching*. Addison Wesley Longman Publishing Co., Inc., Redwood City, CA, USA, 1998.
- [mic] microprofile Spezifikation. https://access.redhat.com/documentation/en-us/jboss_enterprise_application_platform_continuous_delivery/15/html/configuration_guide/eclipse_microprofile. Aufgerufen am: 25-06-2021.
- [nfr] non functional requirements - Evaluierung. <https://www.dotnetcurry.com/project-management/1462/non-functional-requirements-nfrs>. Aufgerufen am: 26-06-2021.

10

Eidesstattliche Erklärung

Ich erkläre hiermit an Eides statt, dass ich die vorliegende Arbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe; die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sind als solche kenntlich gemacht. Die Arbeit wurde bisher in ähnlicher Form keiner anderen Prüfungskommission vorgelegt und auch nicht veröffentlicht.

Ort, Datum

Silas Hoffmann