Methoden und Messverfahren für Mechanismen des automatischen Skalierens in elastischen Cloudumgebungen[1]

Nikolas Herbst[2]

Abstract: Auto-Skalierungsmechanismen für Cloud-Umgebungen versprechen stabile Servicequalität bei niedrigen Kosten und wechselnder Auslastung. Die großen, öffentlichen Cloud-Anbieter bieten regelbasierte Auto-Skalierer auf Basis von Schwellenwerten an. Diese Art des Auto-Skalierens hat jedoch Reaktionszeiten in der Größenordnung von Minuten. Neuartige Auto-Skalierungsmechanismen aus der Literatur versuchen, die Grenzen reaktiver Mechanismen durch den Einsatz proaktiver Vorhersagemethoden zu überwinden. Allerdings ist die Akzeptanz von proaktivem, automatischem Skalieren in der Produktion immer noch sehr gering, da das Risiko hoch ist, sich auf eine einzelne proaktive Methode zu verlassen. Diese Doktorarbeit befasst sich mit der Herausforderung, dieses Risiko zu reduzieren, indem sie unter anderem einen neuen hybriden automatischen Skalierungsmechanismus vorschlägt, der mehrere verschiedene proaktive Methoden kombiniert, die wiederum mit einem reaktiven Rückfallmechanismus gekoppelt sind. Hierbei werden bedarfsgesteuerte, automatisierte Prognoseverfahren zur Vorhersage der ankommenden Lastintensität in Kombination mit einer Servicebedarfsschätzung genutzt, um den erforderlichen Ressourcenverbrauch pro Arbeitseinheit zu berechnen, ohne dass eine Anwendungsinstrumentierung erforderlich ist. Der vorgeschlagene Ansatz wird mit fünf aktuellen proaktiven und reaktiven Auto-Skalierungsmechanismen in drei sowohl privaten wie öffentlichen Cloud-Umgebungen unter eigens entwickelten Wettbewerbsbedingungen fair verglichen. Dabei werden jeweils fünf repräsentative Arbeitslastverläufe generiert, die jeweils aus verschiedenen realen Aufzeichnungen entnommen sind. Insgesamt erreicht der in dieser Arbeit vorgeschlagene Ansatz das beste dynamische Skalierungsverhalten basierend auf Benutzer- und Elastizitätsmetriken von Ergebnissen aus 400 Stunden aggregierter Experimentierzeit.

1 Synopsis

Diese Zusammenfassung ist wie folgt strukturiert: Der Abschnitt 2 motiviert das Thema und setzt den Kontext für die hier zusammengefasste Doktorarbeit [He18a]. Im Anschluss fasst der Abschnitt 3 den aktuellen Stand der Technik zusammen und formuliert die Problemstellung. Schließlich hebt der Abschnitt 4 die zwei Leitziele dieser Doktorarbeit hervor, welche jeweils von vier Forschungsfragen zur Definition der Beitragspunkte begleitet werden. Die Dissertation selbst gliedert sich anhand vier separater Beiträge. Darauf aufbauend fasst der Abschnitt 5 die einzelnen Beiträge zusammen, während erste Einblicke in den Entwurf und die Ergebnisse der umfassenden Evaluation gegeben werden. Abschließend soll der Abschnitt 6 einen Ausblick vermitteln.

[1] Englischer Titel der Dissertation: “Methods and Benchmarks for Auto-Scaling Mechanisms in Elastic Cloud Environments”

[2] Universität Würzburg, Fakultät für Mathematik und Informatik, nikolas.herbst@uni-wuerzburg.de

2 Motivation und Kontext

Vor etwas mehr als einem Jahrzehnt im Jahr 2006 stellte Amazon Web Services (AWS) als erster kommerzieller Anbieter Cloud-Dienstleistungen der allgemeinen Öffentlichkeit bereit und löste damit einen Hype Rund um Themen des Cloud Computing in Wissenschaft und Industrie aus. Erst zwei Jahre später begann der Wettbewerb im Bereich des Cloud Computing, nachdem Microsoft, Google und IBM mit eigenen Cloud Diensten auf den Markt kamen. Während AWS als Marktführer ein Drittel des Marktes hält, teilen sich Microsoft, Google und IBM ein weiteres Drittel des Marktes. In den folgenden Jahren verzeichnete der Cloud-Computing Markt überwältigende Wachstumsraten, welche sich laut einem Gartner Report[3] in einem Marktvolumen von 247 Milliarden US-Dollar im Jahr 2017 widerspiegeln. Begleitet von einer Reihe neu gegründeter, hochkarätiger Konferenzen (z.B. IEEE Cloud, ACM Symposium on Cloud Computing SoCC) und Fachzeitschriften (z.B. IEEE Transactions on Cloud Computing), hat die Forschergemeinschaft in den letzten zehn Jahren unzählige Publikationen im Bereich des Cloud Computing veröffentlicht.

Nunmehr gehört diese Phase des ausgeprägten Wachstums der Vergangenheit an und befindet sich im Übergang zu einer Stabilisierungs- und Reifephase mit Wachstumsraten von unter 18%, wie von Gartner prognostiziert wird. Dennoch verändert das Cloud Computing Paradigma den Betrieb von Rechenzentren weiter. Chief-Executive-Officer Marc Hurd[4] der Oracle Corporation prognostiziert, dass bis 2025 80% der klassischen Rechenzentren verschwinden werden, da Anwendungen in der Produktion zunehmend in Cloud-Umgebungen betrieben werden. Insbesondere in der anhaltenden Stabilisierungs- und Reifephase von Cloud Computing-Angeboten hängen der wissenschaftliche Fortschritt und das Branchenwachstum von etablierten Messverfahren und einer standardisierten Berichterstattung der Qualitätsmerkmale von Cloud-Systemen ab, wie in einem kürzlich erschienenen Gigaom-Analystenbericht[5] aufgezeigt wird.

Laut einem Gartner-Bericht von 2009[6], ist das wichtigste Verkaufsargument von Cloud Computing-Angeboten ihr Pay-per-Use-Modell ohne langfristige Investitionen und Betriebskosten für den Nutzer. In Kombination mit der Basistechnologie der Hardwarevirtualisierung bietet das Pay-per-Use-Servicemodell die Möglichkeit, die zugewiesenen Rechenressourcen elastisch an die aktuelle Nachfrage anzupassen. Cloud-Betreiber können ihre physischen Ressourcen - zumindest in der Theorie - so verwalten, dass die Effizienz optimiert wird. Dabei geht es in einigen Fällen auch darum, mehr virtuelle Ressourcen zu verkaufen als physisch verfügbar sind - auch bekannt als Überbuchung. Gleichzeitig versucht der Betreiber das Betriebsrisiko für den Kunden auf ein maßgeschneidertes Minimum zu beschränken, indem man ihm die Möglichkeit gibt, Ressourcenprioritäten zu definieren.

---

[3] Gartner Cloud Report 2017: `https://www.gartner.com/newsroom/id/3616417`
[4] Oracle CEO Marc Hurd: `https://markhurd.com/about-mark-hurd/`
[5] Gigaom Analyst Report: Die Bedeutung von Benchmarking Clouds:
`https://gigaom.com/report/the-importance-of-benchmarking-clouds/`
[6] Gartner Highlights Fünf Attribute von Cloud Computing:
`https://www.gartner.com/doc/965212/refining-attributes-public-private-cloud`

3 Stand der Technik und Problemstellung

Die elastische Skalierung der zugewiesenen Rechenressourcen erfolgt durch so genannte Auto-Skalierungsmechanismen, die überwachten Leistungskennzahlen analysieren. Dabei haben die Mechansimen die Aufgabe die Ressourcenzuweisung dem aktuellen Bedarf derart dynamisch anzupassen, dass im Optimalfall die Leistung stabil bleibt und die Ressourcen effizient genutzt werden. Gängige Praxis ist die Verwendung simpler, Schwellwertbasierter Mechanismen, die aufgrund ihrer reaktiven Natur zu Leistungseinbußen während der Zeiten von Bereitstellungsverzögerungen führen. Im Gegensatz dazu sind die Verantwortlichkeiten von Cloud-Infrastrukturbetreibern hochkomplex um arbeitslastabhängige Ressourcenplatzierung, Wechselwirkungen, Lastverteilung, Dimensionierungs- und Routingfragen kontinuierlich zu optimieren. Als Folge dieser komplexen und miteinander verflochtenen Wechselwirkungen und dynamischer Optimierungspotentiale, erleben die Cloud-Kunden eine hohe Leistungsvariabilität. Das stellt für unternehmenskritische Anwendungen immer noch einen Hinderungsgrund dar, Cloud-Lösungen zu nutzen [IYE11].

Im Laufe des letzten Jahrzehnts wurde in der Literatur eine große Anzahl von Auto-Skalierungsmechanismen vorgeschlagen, die versuchen die Grenzen reaktiver Mechanismen zu überwinden, indem sie proaktive Prognosemethoden anwenden. Lorido-Botran et al. [LBMAL14] untersuchten diese Mechanismen systematisch. Sie schlagen vor, Auto-Skalierungsmechanismen in Ansätze aus der Warteschlangentheorie, der Kontrolltheorie, der Zeitreihenanalyse und dem maschinellen Lernen zu gruppieren.

Proaktive Autoskalierungsverfahren auf Basis der Zeitreihenanalyse schätzen die Ressourcennutzung, die Reaktionszeiten oder die Systemlast unter Verwendung einfacher Regressionsverfahren, Histogrammanalysen oder basierend auf Black-Box-Methoden wie den auto-regressiven integrierten gleitenden Durchschnitten (ARIMA-Modelle). Letztere haben bekannte Defizite in Bezug auf Laufzeit und Genauigkeit in Szenarien mit komplexen saisonalen Mustern und bei feiner als halbstündlich aufgelösten Zeitreihenwerten. Andere Ansätze, z.B. nicht quelloffene Autoskalierer mit Beteiligung von Google oder Entwicklungen von Netflix, nutzen Signalverarbeitungsmethoden, um das Frequenzspektrum über Fourier- oder Wavelet-Transformationen zu charakterisieren, ohne die Fähigkeit zur Erfassung von Trends zu unterstützen. Auto-Skalierungsmechanismen, welche die Theorie der Warteschlangenbildung nutzen, werden in den meisten Fällen mit einem der anderen Ansätze kombiniert. Kontrolltheoretische Ansätze teilen die Einschränkung kurzer Vorhersagehorizonte, während auf maschinellem Lernen basierende Methoden auf Trainingsphasen beziehungsweise bei Systemanpassungen auf Rekalibrierungsperioden angewiesen sind, die in Produktionsumgebungen nicht realisierbar sind. Darüber hinaus teilt die Mehrheit der vorgeschlagenen Auto-Skalierungsmechanismen die Annahme einer linearen und endlosen Skalierbarkeit der von der Cloud zur Verfügung gestellten Ressourcen. In der Praxis ist diese Annahme aufgrund von Kommunikationsaufwänden, Überbuchungspraktiken und zustandsbehafteten Anwendungsdiensten nicht realistisch.

Mit wenigen Ausnahmen bleiben die in der Literatur vorgeschlagenen proaktiven Auto-Skalierungsmechanismen ohne offengelegte Code-Artefakte. Diese Tatsache reduziert die Reproduzierbarkeit der Versuchsergebnisse und die Vergleichbarkeit der Alternativen er-

heblich. Infolgedessen ist der Einsatz von proaktiven Auto-Skalierern in der Produktion immer noch sehr gering, da das Risiko hoch ist, wenn man sich auf eine einzige proaktive Methode verlässt, auf deren Grundlage automatische Skalierungsentscheidungen getroffen werden. Laut einer eigens durchgeführten systematischen Literaturanalyse werden etwa 40% der Publikationen der Cloud-Forschung mittels Simulation evaluiert. Da diese Literaturanalsyse auch Auto-Skalierungsmechanismen umfasst, kann gesagt werden, dass es gängige Praxis ist, Auto-Skalierungsansätze mit simulativen Werkzeugen zu bewerten. Experimentelle Auto-Skalierer-Evaluationen werden in der Regel in mehr oder weniger ähnlicher Art und Weise durchgefuhrt. Es wird in einer Fallsstudie veranschaulicht, dass die vorgeschlagene Methode in der Lage ist, die Einhaltung des Dienstgüte im Vergleich zu einer beliebigen statischen Ressourcenzuweisung zu verbessern. Die Auswertungsszenarien dabei werden oft von synthetischen Lastintensitätsprofilen wie Sinussignalen oder Sägezahnmustern ohne wirkliche Repräsentativität getrieben, zumal sie von einer proaktiven Autoskalierung leicht vorhersehbar sind. Reale Lastprofile weisen eine Mischung aus Trend-, Saison-, Burst- und Rauschkomponenten auf und sind daher komplex zu erfassen, zu teilen, zu modifizieren und skaliert zu erzeugen.

Der Bewertungsprozess von Cloud-Infrastrukturangeboten in Bezug auf die Qualität der realisierten Elastizität bleibt unspezifiziert. Es fehlt an präzise definierten, aussagekräftigen Metriken gibt, welche die Qualität der tatsächlich erzielten Anpassungen elastischer Ressourcen unter Einhaltung von spezifizierten Messablaufregeln erfassen. Daher ist keine klare Anleitung für die Auswahl und Konfiguration eines Auto-Skalierers für einen gegebenen Kontext verfügbar. Die wenigen bestehenden proaktiven Auto-Skalierer werden auf eine sehr anwendungsspezifische Weise optimiert und in der Regel unter Verschluss gehalten, während im Gegensatz dazu viele andere Artefakte von Cloud-Software inzwischen quelloffen sind.

Zusammenfassend lässt sich sagen, dass das in dieser Arbeit behandelte Problem sich wie folgt formulieren lässt: Das Risiko ist nach wie vor hoch, einen proaktiven Auto-Skalierungsalgorithmus in einer Produktionsumgebung einzusetzen und hat eine geringe Akzeptanz zur Folge. Bestehende Lösungen sind entweder nicht offengelegt oder maßgeschneidert. In der Literatur beschriebene Ansätze werden nicht mit Hilfe eines standardisierten Benchmarks getestet, der faire Vergleiche ermöglichen würde.

4 Leitziele und Forschungsfragen

Nachdem eine Reihe von Defiziten im aktuellen Stand der Technik und bei der Bewertung von Auto-Skalierern identifiziert wurde, formuliert die Arbeit nun zwei übergeordnete Leitziele. Diese wiederum werden jeweils von vier Forschungsfragen begleitet. Die Doktorarbeit selbst ist in drei Teile gegliedert. Teil I stellt die Hintergründe und Grundlagen vor, die zum Verständnis der Beiträge der Arbeit erforderlich sind, sowie eine umfassende Zusammenfassung des Technikstandes. Teil II konzentriert sich auf das erste Leitziel A mit seinen Forschungsfragen. Schließlich behandelt Teil III das zweite Leitziel mit seinen vier weiteren Forschungsfragen, während Teil IV die Evaluation der einzelnen Beiträge enthält.

Leitziel A: Es gilt einen Benchmark für moderne Auto-Skalierer zu entwickeln, um das Vertrauen in neuartige proaktive Mechanismen zu stärken.

Um dieses Leitziel zu erreichen, werden die jeweiligen Herausforderungen in mehrere Teilziele aufgeteilt. Zunächst formulieren zwei Forschungsfragen die Notwendigkeit repräsentative Lastintensitätsprofile flexibel definieren, modifizieren und generieren zu können, um eine realistische Menge an Ressourcenanpassungen auszulösen. Zweitens erfassen zwei weitere Forschungsfragen die Notwendigkeit einer fundierten Definition von Metriken und Messmethodik als Bausteine für einen Elastizitätsbenchmark.

A.1: Wie können Lastintensitätsprofile aus realen Aufzeichnungen auf anschauliche, kompakte, flexible und intuitive Weise definiert werden?

A.2: Wie lassen sich automatisch Modelle von Lastintensitätsprofilen aus bestehenden Aufzeichnungen mit einer angemessenen Genauigkeit und Rechenzeit extrahieren?

A.3: Was sind sinnvolle und intuitive Metriken zur Quantifizierung von Genauigkeit, Timing und Stabilität als Qualitätsaspekte bei der Anpassung elastischer Ressourcen?

A.4: Wie können vorgeschlagene Elastizitätsmetriken zuverlässig und wiederholbar gemessen werden, um faire Vergleiche und auch konsistente Rangordnungen über Systeme mit unterschiedlicher Leistungsfähigkeit zu ermöglichen?

Leitziel B: Es gilt das Risiko der Verwendung neuartiger Auto-Skalierer im Betrieb zu reduzieren, indem mehrere proaktive Mechanismen genutzt werden, die auch in Kombination mit einem herkömmlichen Reaktionsmechanismus nutzbar sind.

Die Herausforderung dieses Ziels werden adressiert, indem zunächst ein neuartiger hybrider Auto-Skalierungsmechanismus vorschlagen und dann seine Leistung im Detail mit den bestehenden modernen Auto-Skalierern verglichen wird. Die umfassende Auto-Skalierer-Bewertung wird durch die Ergebnisse von Leitziel 1 ermöglicht. Zweitens werden Defizite der derzeitigen Zeitreihenprognosemethoden behoben, indem ein hybrider Prognosemechanismus vorschlagen und seine Vorteile im Kontext der automatischen Skalierung aufzeigt wird.

B.1: Wie können widersprüchliche Auto-Skalierungsentscheidungen aus unabhängigen reaktiven und proaktiven Entscheidungsschleifen kombiniert werden, um die Gesamtqualität der Anpassungsentscheidungen zu verbessern?

B.2: Wie gut schneidet der vorgeschlagene, hybride Auto-Skalierungsansatz im Vergleich zu modernsten Mechanismen in realistischen Umgebungen und Anwendungsszenarien ab?

B.3: Wie kann ein hybrider Prognosemechanismus auf der Grundlage der Zerlegung so konzipiert werden, dass er in der Lage ist, genaue und schnelle Vorhersagen komplexer saisonaler Zeitreihen zu liefern?

B.4: Ist ein solcher hybrider Prognoseansatz in der Lage, die Leistung und Zuverlässigkeit von Auto-Skalierungsmechanismen zu verbessern?

5 Beiträge und Zusammenfassung der Evaluation

Nachdem zwei Leitziele und je vier Forschungsfragen definiert wurden, werden nun die vier Kernbeiträge dieser Arbeit zusammengefasst. Jeder Beitrag greift zwei der Forschungsfragen und damit je einen Teil von Ziel A oder B auf. Die Beiträge haben gemeinsam, dass sie aufeinander aufbauen und im Endergebnis integriert sind.

Beitrag I:

Zur Adressierung von Leitziel A und Beantwortung der Forschungsfragen A.1 und A.2, schlägt die Arbeit ein beschreibendes Lastprofil-Modellierungssystem zusammen mit einer automatisierten Modellextraktion aus Aufzeichnungen vor, um eine reproduzierbare Erzeugung von Arbeitslasten mit realistischen Lastintensitätsschwankungen zu ermöglichen. Der Modellentwurf folgt dem Ansatz der Zerlegung aufgezeichneter Daten in ihre deterministischen Komponenten aus stückweise definierten Trends und wiederkehrenden oder übergreifenden Saisonmustern, wobei gleichzeitig stochastische Rauschverteilungen und explizite Spitzen modelliert werden können. Die Komponenten sind im Prinzip stückweise definierte mathematische Funktionen, die verschachtelt und mit mathematischen Operationen in einem Funktionsbaum kombiniert werden können. Automatisierte Extraktionsprozesse erkennen Frequenzen und zerlegen aufgezeichnete Daten basierend auf einer effizienten Heuristik. Das vorgeschlagene Modell nennt sich “Descartes Load Intensity Model (DLIM)” mit seiner Limbo Werkzeugkette, die wichtige Funktionen zum Benchmarking von Ressourcenmanagementansätzen auf repräsentative und faire Weise bereitstellt.

Die Ausdrucksmächtigkeit des DLIM-Modells wird bewertet, indem unterschiedliche Konfigurationen der Extraktionsprozesse anwendet und auf zehn verschiedenen realen Aufzeichnungen verglichen werden, die zwischen zwei Wochen und sieben Monaten an Anfrageraten umfassen. Automatisch extrahierte DLIM-Modellinstanzen weisen einen durchschnittlichen Modellierungsfehler von 15,2% auf. In Bezug auf Genauigkeit und Verarbeitungsgeschwindigkeit liefern die vorgeschlagenen Extraktionsmethoden, die auf deskriptiven Modellen basieren, bessere oder ähnliche Ergebnisse im Vergleich zu bestehenden nichtdeskriptiven Zeitreihenzerlegungsmethoden. Im Gegensatz zu DLIM-Modellen liefern klassische Zeitreihenzerlegungsansätze drei Reihen von Datenpunkten als Ausgabe im Gegensatz zu einem kompakten und flexiblen deskriptiven Modell. Dieser Beitrag führte zu einem Fachzeitschriftenartikel in den ACM Transactions on Autonomous and Adaptive Systems (TAAS) [Ki17], der 2017 veröffentlicht wurde.

Beitrag II:

Um die Forschungsfragen A.3 und A.4 anzugehen, wird zunächst eine klare Definition des Begriffs "**Elastizität**" im Cloud Computing[7] erarbeitet. Die Kernaspekte der Elastizität werden beschrieben und von verwandten Begriffen wie Effizienz und Skalierbarkeit abgegrenzt. Darüber hinaus wird eine Reihe von neuen, intuitiv verständlichen Metriken für die Quantifizierung von Timings-, Stabilitäts- und Genauigkeitsaspekten der Elastizität definiert. Basierend auf diesen Metriken, die auch von der Forschergruppe der Standard Performance Evaluation Corporation SPEC [8] befürwortet wurden, wird einen neuartigen Ansatz für das Benchmarking der Elastizität von Auto-Skalierern vorgeschlagen. Dabei können die praktisch erzielte Elastizität von "Infrastructure-as-a-Service" (IaaS)-Cloud-Plattformen unabhängig von der Leistungsfähigkeit der zugrunde liegenden Ressourcen bewertet und verglichen werden. Das vorgeschlagene Bungee Elastizitätsbenchmarking-Werkzeug nutzt die Modellierungsfunktionen von DLIM, um realistische Lastintensitätsprofile zu erzeugen.

In der zugehörigen Evaluation wird gezeigt, dass für jede der vorgeschlagenen Metriken ein konsistentes Ranking der elastischen Systeme auf einer Ordinalskala geliefert wird. Die Bungee-Messmethodik kann sowohl reproduzierbare Ergebnisse in einer kontrollierten Umgebung als auch Ergebnisse mit einer akzeptablen Variation in unkontrollierten Umgebungen wie öffentlichen Clouds erzielen. Schließlich wird eine umfangreiche Fallstudie von realer Komplexität präsentiert, die zeigt, dass der vorgeschlagene Ansatz in realistischen Szenarien anwendbar ist und mit unterschiedlichen Leistungsniveaus der zugrunde liegenden Ressourcen umgehen kann. Die Definitionen der Elastizitätsmetriken wurden zu einem wesentlichen Bestandteil eines Artikels in den ACM Transactions on Modeling and Performance Evaluation of Computing Systems (ToMPECS) [He18b].

Beitrag III:

In diesem Beitrag wird nun das zweite Leitziel angegangen: Das Risiko, sich auf einen einzigen proaktiven Auto-Skalierer zu verlassen, wird reduziert durch einen neuartigen hybriden Auto-Skalierungsmechanismus namens Chameleon. Dabei werden mehrere proaktive Methoden kombiniert, die wiederum mit einem reaktiven Rückfallmechanismus gekoppelt sind. Chameleon verwendet bedarfsgesteuerte, automatisierte, Zeitreihen-basierte Prognoseverfahren, um die ankommende Lastintensität in Kombination mit Schätzverfahren für den Ressourcenbedarf vorherzusagen. Es ist erforderlich den Ressourcenverbrauch pro Arbeitseinheit zu berechnen, ohne dass eine Anwendungsinstrumentierung vorgenommen wurde. Der Ansatz kann auch strukturelles Anwendungswissen nut-

[7] Die in dieser Arbeit vorgeschlagene Definition der Elastizität im Cloud Computing [HKR13] wurde von Wikipedia in einem entsprechenden enzyklopädischen Artikel aufgegriffen (c.f. `https://en.wikipedia.org/wiki/Elasticity_(cloud_computing)`).

[8] Standard Performance Evaluation Corporation SPEC Forschergruppe `http://research.spec.org`

zen, indem es Warteschlangennetzwerke in Produktform löst, aus denen dann optimierte Skalierungsaktionen abgeleitet werden. Der Chameleon-Ansatz löst Konflikte zwischen reaktiven und proaktiven Skalierungsentscheidungen auf intelligente Weise und nutzt als Bausteine die wichtigsten Entwicklungen der Descartes-Forschungsgruppe wie die Descartes Modellierungssprache (DML) [Hu17] zur Erfassung von Anwendungsstrukturen in Rechenzentren sowie die Bibliothek für Schätzverfahren von Ressourcenbedarfen (LibReDE) [Sp15].

In der Arbeit wird ein umfangreicher Auto-Skalierer-Wettbewerb durchgeführt unter Nutzung der Ergebnisse aus den Beiträgen I und II: Der Chameleon Ansatz wird systematisch mit vier verschiedenen modernen proaktiven Auto-Skalierern sowie einem herkömmlichen Schwellenwert-basierten in drei unterschiedlichen Cloud-Umgebungen verglichen: (i) eine private CloudStack-basierte Cloud-Umgebung, (ii) die öffentliche AWS EC2 Cloud, sowie (iii) eine OpenNebula-basierte geteilte IaaS-Cloud. Es werden insgesamt fünf repräsentative Lastprofile generiert, die jeweils aus verschiedenen realen Aufzeichnungen stammen. Die Funktionalität der Werkzeuge Limbo (Beitrag I) und Bungee (Beitrag II) werden genutzt, um eine variierende, CPU-intensive Systemauslastung zu erreichen. Die Beispielanwendung wird dem SPEC Server Efficiency Rating (SERT) Werkzeug entnommen und berechnet Matrixdekompositionen. Insgesamt erreicht Chameleon das beste und stabilste Skalierungsverhalten basierend auf Benutzer- und Elastizitätsmetriken. Dabei werden die Ergebnisse von 400 Stunden aggregierter Experimentierzeit analysiert. Es wird gezeigt, dass durch die Kombination von Skalierungsentscheidungen aus reaktiven und proaktiven Zyklen, basierend auf der vorgeschlagenen Konfliktlösungsheuristik, die Auto-Skalierungsleistung von Chameleon verbessert wird. Dieser Beitrag führte zu einem Artikel zu den IEEE Transactions on Parallel and Distributed Systems (TPDS) [Ba19].

Beitrag IV:

Als weiterer Beitrag dieser Arbeit werden Forschungsfragen B.3 und B.4 beantwortet, indem einen ein neuartiges Prognoseverfahren für Zeitreihen namens Telescope vorschlagen wird. Telescope integriert mehrere individuelle Prognosemethoden, indem es die univariaten Zeitreihen in die Komponenten Trend, Saison und Rest zerlegt. Zunächst wird automatisch die Frequenz bestimmt sowie Anomalien erkannt und beseitigt. Danach wird die Art der Zerlegung (multiplikativ oder additiv) basierend auf einer Mehrheitsentscheidung von maßgeschneiderten Tests ermittelt. Nach der Zerlegung wird die ARIMA-Methode (autoregressive integrierte gleitende Durchschnitte) ohne Saisonalität auf das Trendmuster angewendet, wobei die ermittelten saisonalen Muster einfach fortgesetzt werden. Darüber hinaus werden die einzelnen Perioden gruppiert, um kategorische Informationen zu erhalten. Die Cluster-Labels werden durch den Einsatz künstlicher neuronaler Netze prognostiziert. Dies hilft, automatisch zwischen verschiedenen Arten von Tagen zu unterscheiden. Schließlich wird eXtreme Gradient Boosting (XGBoost), eine neuartige und vielversprechende Methode, die 2016 veröffentlicht wurde, verwendet, um die Abhängigkeit zwischen allen zuvor extrahierten Kovariablen zu ermitteln und die Prognosen der einzelnen Komponenten zu kombinieren.

Die Bewertung zeigt anhand von zwei Zeitreihen, dass eine prototypische Implementierung des Telescope-Ansatzes sechs aktuelle Prognosemethoden in Bezug auf die Genauigkeit übertrifft. Telescope verbessert auch die Berechnungszeiten im Vergleich zu den drei wettbewerbsfähigsten Prognosemethoden um das bis zu 19-fache. In einer Fallstudie wird gezeigt, dass Telescope in der Lage ist, die Auto-Skalierungsleistung von Chameleon im Vergleich zu der früher verwendeten Prognosemethode tBATS oder saisonalem ARIMA weiter zu verbessern.

6 Ausblick

Die vier Kernbeiträge dieser Arbeit bringen das Potenzial, die Art und Weise zu verändern, wie Cloud-Ressourcenmanagement-Ansätze bewertet werden. Das wiederum dürfte eine Verbesserung der Qualität von autonomen Managementalgorithmen als Ergebnis mit sich bringen. Um eine solche Entwicklung zu unterstützen, wurden Code-Artefakte aller vier Beiträge dieser Arbeit als quelloffene Software-Werkzeuge veröffentlicht, die aktiv gepflegt und von Benutzer- und Entwicklerleitfäden begleitet werden[9].

Über die in den einzelnen Kapiteln explizit genannten Einschränkungen und Annahmen hinaus sehen gibt es eine Reihe von Herausforderungen für die zukünftige Forschung im Cloud-Ressourcenmanagement und dessen Bewertungsmethoden: (I) Die Einführung der Containerisierung auf virtuellen Maschineninstanzen führt zu einer weiteren Ebene der Indirektion. Infolgedessen erhöht die Verschachtelung virtueller Ressourcen die Ressourcenfragmentierung und verursacht unzuverlässige Bereitstellungsverzögerungen. (II) Des Weiteren neigen virtualisierte Rechenressourcen dazu, immer inhomogener zu werden, verbunden mit verschiedenen Prioritäten und Kompromissen. (III) Durch DevOps-Praktiken werden Updates für Cloud-gehostete Dienste mit einer höheren Frequenz veröffentlicht, was sich auf die Dynamik des Nutzerverhaltens auswirkt. Auto-Skalierungsmechanismen müssen sich zunehmend selbst an sich ändernde Serviceanforderungen und Ankunftsmuster anpassen.

Literaturverzeichnis

[Ba19] Bauer, André; Herbst, Nikolas; Spinner, Simon; Ali-Eldin, Ahmed; Kounev, Samuel: Chameleon: A Hybrid, Proactive Auto-Scaling Mechanism on a Level-Playing Field. IEEE Transactions on Parallel and Distributed Systems, 30(4):800–813, September 2019.

[He18a] Herbst, Nikolas: Methods and Benchmarks for Auto-Scaling Mechanisms in Elastic Cloud Environments. Dissertation, Universität Würzburg, Deutschland, Juli 2018. SPEC Kaivalya Dixit Distinguished Dissertation Award 2018.

[He18b] Herbst, Nikolas; Bauer, André; Kounev, Samuel; Oikonomou, Giorgos; van Eyk, Erwin; Kousiouris, George; Evangelinou, Athanasia; Krebs, Rouven; Brecht, Tim; Abad, Cristina L.; Iosup, Alexandru: Quantifying Cloud Performance and Dependability: Taxonomy, Metric Design, and Emerging Challenges. ACM Transactions on Modeling

[9] Descartes Tools: `https://descartes.tools`

and Performance Evaluation of Computing Systems (ToMPECS), 3(4):19:1–19:36, August 2018.

[HKR13] Herbst, Nikolas; Kounev, Samuel; Reussner, Ralf: Elasticity in Cloud Computing: What it is, and What it is Not. In: Proceedings of the 10th International Conference on Autonomic Computing (ICAC 2013). USENIX, June 2013. Top 1 most cited ICAC paper (according to Google Scholar).

[Hu17] Huber, Nikolaus; Brosig, Fabian; Spinner, Simon; Kounev, Samuel; Bähr, Manuel: Model-Based Self-Aware Performance and Resource Management Using the Descartes Modeling Language. IEEE Transactions on Software Engineering (TSE), 43(5), 2017.

[IYE11] Iosup, A.; Yigitbasi, N.; Epema, D.: On the Performance Variability of Production Cloud Services. In: CCGrid 2011. S. 104–113, 2011.

[Ki17] von Kistowski, Jóakim; Herbst, Nikolas; Kounev, Samuel; Groenda, Henning; Stier, Christian; Lehrig, Sebastian: Modeling and Extracting Load Intensity Profiles. ACM Transactions on Autonomous and Adaptive Systems (TAAS), 11(4):23:1–23:28, Januar 2017.

[LBMAL14] Lorido-Botran, Tania; Miguel-Alonso, Jose; Lozano, Jose A: A Review of Auto-scaling Techniques for Elastic Applications in Cloud Environments. Journal of Grid Computing, 12(4):559–592, 2014.

[Sp15] Spinner, Simon; Casale, Giuliano; Brosig, Fabian; Kounev, Samuel: Evaluating Approaches to Resource Demand Estimation. Elsevier Performance Evaluation, 92:51 – 71, October 2015.

Nikolas Herbst leitet die Forschergruppe für prädiktive Datenanalyse am Lehrstuhl für Software Engineering der Universität Würzburg. Er promovierte 2018 an derselben Universität. Bevor er im Jahr 2014 vom Forschungszentrum für Informatik (FZI) am Karlsruher Institut für Technologie (KIT) zusammen mit seinem Doktorvater Samuel Kounev an die Würzburger Universität wechselte, erwarb er 2012 am KIT ein Diplom der Informatik. Nikolas is gewählter, stellvertretender Vorsitzender der SPEC Research Cloud Gruppe. Zu seinen Forschungsthemen gehören neben Elastizität im Cloud Computing, Autoskalierung und Ressourcemanagement auch die Leistungsbewertung von virtualisierten Umgebungen, Techniken des Autonomic und Self-Aware Computing, sowie der Datenanalyse und Modellbildung mittels Kombinationen von maschinellem Lernen und stochastischen Verfahren.