Technische Universität
KAISERSLAUTERN

Young Researcher
Symposium

(YRS) 2011

15. Februar 2011
Fraunhofer Zentrum
Kaiserslautern

Proceedings

$(CM)^2$

Center for Mathematical +
Computational Modelling

Preface

First, we want to thank our sponsor *(CM)^2 – Center for Mathematical and Computational Modelling* for financial support and to host this event.

The aim of this symposium is to bring together young researchers from various fields. Its interdisciplinary character is underlined by the support of the University of Kaiserslautern, DFKI, Fraunhofer ITWM, IVW, and the Rhineland-Palatinate research initiative. With more than 40 participants and 14 submissions, the first YRS demonstrated the strength and interest in such an event. In a relaxed atmosphere, young researchers were able to discuss new ideas for cooperations and support.

These proceedings contain peer-reviewed and accepted contributions from the domains *materials characterization*, *production processes*, *economics*, and *system-on-chip design (SOC)*. The close relationship to mathematics and computer science, which is essential to engineers and natural scientists, was always visible and will help to advance progress in relevant areas.

The top three contributions were awarded by the committee and the audience. The winners are: "Hardware Accelerators for Financial Mathematics – Methodology Results and Benchmarking" by Christian De Schryver, Henning Marxen, and Daniel Schmidt (first place), "Mathematical Optimization Based Channel Coding: Current Achievements and Future Challenges" by Michael Helmling, Stefan Scholl, and Akin Tanatmis (second place), and "Zur Untersuchung von stahlfaserverstärkten Betonen auf Basis der 3D-Computer-Tomographie" by Frank Schuler and Katja Schladitz.

We thank the authors for their submissions and the program committee for their hard work.

Nachwuchsring des (CM)^2

February 2011   Frank Balle, Nicole Marheineke, Frank Kienle,
                                         Karsten Schmidt, Nicole Tschauder

Organizing Committee

Prof. Dr. Nicole Marheineke, University of Erlangen-Nürnberg
Dipl.-Math. Nicole Tschauder, University of Kaiserslautern
Dr.-Ing. Frank Kienle, University of Kaiserslautern
Dipl.-Inf. Karsten Schmidt, University of Kaiserslautern
Dr.-Ing. Frank Balle, University of Kaiserslautern

Program Committee

Prof. Dr. Ralf Korn (Chairman)
Prof. Dr.-Ing. Norbert Wehn
Prof. Dr. Arnd Poetzsch-Heffter
Prof. Dr.-Ing. Dietmar Eifler
Prof. Dr. Nicole Marheineke
Dipl.-Math. Nicole Tschauder
Dr.-Ing. Frank Kienle
Dipl.-Inf. Karsten Schmidt
Dr.-Ing. Frank Balle

Attendees

1st Young Researcher Symposium
15th February 2011
Fraunhofer Center
Kaiserslautern (Germany)

Contents

7

Zuverlässige und Schnelle Erzeugung von Zufallsnetzwerken für Evaluationszwecke

Darko Obradović, Wolfgang Schlauch

AG Wissensbasierte Systeme, Fachbereich Informatik, TU Kaiserslautern
Kaiserslautern, Germany
darko.obradovic@dfki.uni-kl.de
wolfgang.schlauch@dfki.uni-kl.de

Abstract—Die Analyse von Netzwerken hat in der Soziologie seit dem 20. Jahrhundert einen hohen Stellenwert, ist aber auch für andere Disziplinen interessant. Zum Beispiel bei der Untersuchung von neuronale Netzen oder DNA-Transkriptionen in der Biologie. Für die Evaluation von Mustern gibt es verschiedene Methoden, wobei wir in dieser Arbeit den Fokus auf den statistischen Vergleich des realen Netzwerks mit gleichartigen Zufallsnetzwerken legen, da diese in besonderem Maße objektive Aussagen liefern kann. Für diese Methode existieren zwei Verfahren mit verschiedenen Unzulänglichkeiten, welche beide nicht in allen Fällen zuverlässig machen. Für das Markov-Ketten Monte-Carlo Verfahren versuchen wir, das offene Problem einer effizienten, aber auch zuverlässigen Schrittzahl mit einem neuen Ansatz anzugehen. Das Idee besteht darin, eine möglichst exakte Berechnung dieser Schrittzahl für bestimmte KLassen von Netzwerken über ein Modell der Markov-Kette zu erreichen. Sollte uns das in unserer zukünftigen Arbeit fundiert gelingen, wäre dies der bisherigen empirischen Methode deutlich überlegen.

I. Einführung

In diesem Papier beschäftigen wir uns mit der Analyse von Netzwerkstrukturen, welche theoretisch stark in der Graphentheorie der Mathematik verwurzelt ist, praktisch aber vor allem durch die Untersuchung von sozialen Netzwerken vorangetrieben wird [1]. Trotz der Fokussierung dieser Disziplin auf soziale Netzwerke, woher sie auch ihren Namen *Soziale Netzwerkanalyse* (SNA) erhielt, lassen sich die Methoden auf jede andere Disziplin übertragen, welche ihre zu analysierenden Strukturen in Netzwerken repräsentieren kann. Folglich zeigten sich auch schon erste Anwendungen der Biologie in Neuronalen- oder Gen-Netzwerken.

A. Historie der Sozialen Netzwerkanalyse [2]

Ende der 1930er Jahre tauchte bei John Moreno erstmals das Konzept des *Soziogramms* auf, mit dem er soziale Beziehungen formalisierte, um sie danach methodisch zu untersuchen. Im Jahre 1954 wurde der Term „Soziales Netzwerk"von J. A. Barnes als ein feststehender Ausdruck verwendet, um sich wiederholende Strukturen in verschiedenen Gruppierungen zu beschreiben, zum Beispiel in Stämmen, Familien oder auch in sozialen Kategorien.

Seitdem ging es mit der Untersuchung von sozialen Netzwerken langsam, aber stetig, weiter. Einerseits untersuchte Harrison White mit seiner Studiengruppe an der Harvard University, Department of Social Relations, dieses Themengebiet auf einer sozialen Ebene, während Charles Tilly sich mit

sozialen Bewegungen und politischen Netzwerken beschäftigte. Besondere Erwähnung verdient insbesondere Stanley Milgram, der mit der These des *Kleinen-Welt-Phänomens* die Untersuchung der sozialen Netzwerke vorantrieb und durch die These, dass in den USA jeder Mensch mit jedem anderen über durchschnittlich sechs Verbindungen bekannt sei, großes Interesse erweckte, auch wenn dieses Experiment methodisch stark umstritten ist.

Durch Duncan J. Watts und Steven H. Strogatz wurde das Kleine-Welt-Phänomen 1998 weiter untersucht [3]. Sie brachten interessante neue Ergebnisse hervor im Bezug auf bestimmte Effekte wie den Abstand zweier Knoten voneinander. Sie nahmen in der realen Welt vorkommende Netze, zum Beispiel den Graphen der Zusammenarbeit von Schauspielern, und untersuchten sie, schlugen aber auch die genauere Untersuchung von anderen sozialen Netzwerken vor.

B. Evaluation von Netzwerkeigenschaften

Bei der Analyse von realen Netzwerken sind zwei grundlegende Methoden stark verbreitet.

Zum Einen die *explorative Analyse*, bei welcher der Forscher eine graphische Darstellung des Netzwerks interaktiv durchsieht, oder spezielle Layouts betrachtet, um auffällige Muster oder Eigenschaften zu entdecken. Diese Entdeckungen sind dann das Ergebnis der Analyse. Hierfür existieren eine Vielzahl von Tools wie bspw. **GUESS**.

Zum Anderen gibt es noch die *Metriken-basierte Analyse*, bei welcher Metriken auf dem Netzwerk (z.B. Dichte, Durchmesser, ...) oder Metriken für einzelne Knoten (z.B. Zentralität, Clustering-Koeffizient, ...) im Durchschnitt oder in ihrer Verteilung betrachtet werden. Hier können aber auch Algorithmen zur Suche bestimmter quantifizierbarer Muster zum Einsatz kommen (z.B. Clustering, Motive, ...).

Die explorative Methode ist bei vielen Forschern, die anwendungsorientiert arbeiten, sehr beliebt, da man mit Standardwerkzeugen sehr schnell und einfach zu Ergebnissen kommen kann. Diesen Ergebnissen haftet jedoch stets ein Anschein der Subjektivität an, wo man in der Wissenschaft doch aber stets um Objektivierungen bemüht ist. Folglich ist die Metriken-basierte Analyse nach allgemeiner Ansicht die aussagekräftigere und stärkere Methode. Für viele Standardmetriken gibt es heute auch schon mächtige Werkzeuge wie z.B. **Pajek**. Diese stoßen zwar an ihre Grenzen, wenn das

Repertoire der Standardmetriken zur Beantwortung der Forschungsfragen nicht ausreicht, und erfordern dann weitreichende Programmierfähigkeiten. Dennoch hat sich die Metriken-basierte Methode heute für wissenschaftliche Untersuchungen weitgehend durchgesetzt, auch in Kombination mit der explorativen Methode.

Eine dritte, recht neue Methode ist die *statistisch-stochastische Analyse*, welche die Metriken-basierte Analyse erweitert. Die Objektivität der Metriken-basierten Analyse ist nämlich an jener Stelle nicht mehr gegeben, an welcher die numerischen Ergebnisse durch den Forscher interpretiert werden. Problematisch sind hier Aussagen, welche die Besonderheit von Mustern, bzw. deren erwartetes Auftreten subjektiv bewerten. Dies wurde in einem sehr bedeutenden Artikel von Watts und Strogatz 1998 erstmals eindrücklich durch die Untersuchung von Zufallsnetzwerken demonstriert [3]. In diesem Licht erscheinen z.B. auch die „Six Degrees of Separation" weit weniger spektakulär als ursprünglich angenommen, da das Gegenteil in einem nicht perfekt-regulären Netzwerk objektiv gesehen eine viel größere Sensation gewesen wäre.

Diese Beobachtung machten sich Forscher dann auch erstmals zu Nutze, um Interpretationen von Metriken zu objektivieren. Hierfür wird eine Anzahl vergleichbarer[1] Zufallsnetzwerke generiert, und die gewünschte Metrik für diese ermittelt. Somit erhält man einen Anhaltspunkt, welchen Wert der Metrik man statistisch in gleichartigen Netzwerken erwarten kann.

So wies beispielsweise Alon 2007 nach, dass bestimmte *Motive* in biologischen Netzen um die vielfache Standardabweichung öfter auftauchen, als im Durchschnitt von gleichartigen Zufallsnetzwerken. Erst jetzt ist also die Aussage, dass die gemessene Ausprägung einer Metrik ungewöhnlich sei, tatsächlich objektiv nachgewiesen.

C. Motivation

Die statistisch-stochastische metrik-basierte Untersuchungsmethode für reale Netzwerke ist eine recht junge, aber sehr vielversprechende Methode, wie die Rückmeldungen aus der Forschergemeinschaft zeigen. So erregte damals die Arbeit von Watts and Strogatz schon viel Aufmerksamkeit, die Ergebnisse Alons aus der Praxis wurden ebefalls renommiert publiziert. Und auch wir haben für Analysen der Blogosphäre mit Hilfe dieser Methode viel Anerkennung erhalten [4], weshalb wir diese Methode intensiv weiterverfolgen möchten.

Als kritischer Punkt hat sich hierbei jedoch die Wahl des Algorithmus zur Generierung dieser gleichartigen Zufallsnetzwerke erwiesen. Die Gleichartigkeit wird zuallererst natürlich über die Anzahl an Knoten und Kanten definiert, nach allgemeinem Konsens aber auch über die Knotengrade der einzelnen Knoten, welche eine vorgegebene Knotengradsequenz für das Netzwerk zur Folge hat. Nun gibt es zwei Algorithmen zur Generierung eines Zufallsnetzwerkes mit einer solchen vorgeschriebenen Sequenz, welche in [5] erläutert und verglichen werden. Zum Einen das *Konfigurationsmodell*

als direkt generierender Algorithmus, welcher sehr einfach, und in der Praxis am weitesten verbreitet ist, jedoch die ihm gestellte Aufgabe nur bis zu einem bestimmten Grad erfüllt, und Zweifel an seiner Zuverlässigkeit offen lässt. Zum Anderen gibt es den *Markov-Ketten Monte-Carlo Algorithmus*, welcher deutlich aufwändiger ist, und sich in der Praxis noch nicht etablieren konnte. Er erfüllt zwar die Anforderungen an die Zufallsnetzwerke voll, hinterlässt beim Anwender aber, wie viele andere Markov-Ketten-basierte Algorithmen auch, die Frage nach der richtigen Schrittzahl, welche einerseits für die Geschwindigkeit und andererseits für die Uniformität des Verfahrens, und somit auch für seine stochastische Aussagekraft von entscheidender Bedeutung ist.

Eine lineare Größenordnung dieser Schrittzahl, welche für die Praktikabilität des Verfahrens von entscheidender Bedeutung ist, wird bisher nur vermutet, und für den Vorfaktor gibt es nur indirekte empirische Untersuchungen. Tatsächlich bewiesen wurde bisher nur eine quadratische Größenordnung abhängig von der zu generierenden Kantenzahl. Diese Situation führt bei uns, und möglicherweise auch bei anderen Forschern, noch zu einer Zurückhaltung vor der Anwendung.

In diesem Papier präsentieren wir unsere Idee und erste Experimente, um die Schrittzahl in Einzelfällen exakt zu bestimmen. Indem wir dies für verschieden große Netzwerke einer bestimmten, sehr weit verbreiteten Klasse von Knotengradsequenzen, den sogenannten *skalenfreien Netzen* [6] ermitteln, erhoffen wir uns, allgemeine und verlässliche Rückschlüsse auf die notwendige Schrittzahl für die Uniformität in dieser Klasse zu ermitteln.

II. ZUFALLSNETZWERKE

In diesem Kapitel gehen wir im Detail auf die bereits erwähnte statistisch-stochastische Untersuchungsmethode für Netzwerke ein, und beschreiben die daraus resultierenden Anforderungen an entsprechende Generierungsverfahren für Zufallsnetzwerke. Danach werden die bekannten Verfahren erläutert, kritisch betrachtet und verglichen.

A. Die Statistisch-Stochastische Methode im Detail

In diesem Abschnitt wollen wir genauer auf die in Kapitel I-B beschriebene statistisch-stochastische metrik-basierte Untersuchungsmethode für reale Netzwerke eingehen. Ausgangspunkt sind hierbei zum Einen das reale Netzwerk als Gegenstand der Untersuchung, und zum Anderen die vom Forscher ausgewählte Metrik, über welche er Aussagen treffen möchte.

Neben der Berechnung der Metriken auf dem realen Netzwerk, wird diese auch mit ihren erwarteten Ausprägungen bei gleichartigen Zufallsnetzwerken verglichen. Diese Gleichartigkeit definiert sich über die Anzahl der Knoten und Kanten, und zusätzlich auch über die Knotengradsequenz des Netzwerkes. D.h. dass zu jedem Knoten des realen Netzwerkes im Zufallsnetzwerk ein Pendant mit exakt der gleichen Zahl an ein- und ausgehenden Kanten existiert.

Verfügt man gedanklich über eine Urne, welche alle möglichen Netzwerke enthält, die in Größe und Gradsequenz

---

[1]mehr zur Vergleichbarkeit in den folgenden Abschnitten

dem realen Netzwerk entsprechen, so kann man nun eine hinreichend große Anzahl s dieser Netzwerke mit jeweils gleicher Wahrscheinlichkeit ziehen, und erhält somit eine repräsentative Teilmenge. Üblich sind hier alles zwischen 30 und 1000 Ziehungen. Für jedes gezogene Netzwerk wird nun die Metrik berechnet, und für die Menge insgesamt kann man nun die deskriptive statistische Verteilung der Ausprägung ermitteln. Dies resultiert in einem Durchschnittswert mit einer Standardabweichung.

Für das reale Netzwerk kann man nun den absoluten Abstand der Metrik zum statistischen Mittel der Zufallsnetzwerke berechnen, und gibt diesen geteilt durch die Standardabweichung als z-Wert an. Hat die Metrik auf dem realen Netzwerk ein $z < 1$ so ist die Eigenschaft im Rahmen des zu Erwartenden, weist sie allerdings ein $z > 3$ auf, so kann man davon ausgehen, eine Besonderheit des realen Netzwerks im Vergleich zu gleichartigen Zufallsnetzwerken entdeckt zu haben. Die Interpretation der Bedeutung dieser Besonderheit ist nun natürlich wieder subjektiv, stützt sich mittlerweile aber auf schon zwei objektivierte Aussagen. Im Normalfall wird damit eine vorher entsprechend aufgestellte Hypothese überprüft.

B. Anforderungen an die Verfahren

Ein konkretes Netzwerk, welches eine gegebene Knotengradsequenz realisiert, nennt man eine *Konfiguration* dieser Sequenz. Nun übersteigt schon bei kleinen Sequenzen die Anzahl der Konfigurationen aufgrund der kombinatorischen Explosion alle Speichermöglichkeiten [7], weshalb eine entsprechende Urne durch ein uniformes Generierungsverfahren für diese Konfigurationen simuliert werden muss; uniform insofern, dass jede mögliche Konfiguration exakt die gleiche Wahrscheinlichkeit haben muss, durch ein solches Verfahren generiert zu werden.

Aufgrund der Beschaffenheit fast aller realer Netzwerke kommen noch folgende Einschränkungen an die Menge der Konfigurationen hinzu. Erstens soll die Gesamtmenge aller Konfigurationen nur einfache Graphen enthalten, d.h. sie sollen weder parallele Kanten noch Kanten zwischen ein und demselben Knoten, welche wir hier Selbst-Kanten nennen, enthalten. Zweitens sind oftmals auch nur zusammenhängende Graphen gewünscht, also Graphen ohne vollständig isolierte Subgraphen. Diese Anforderungen sollen die statistisch-stochastischen Werte auf Grundlage der real überhaupt nur in Frage kommenden Netzwerke ermitteln.

C. Das Konfigurationsmodell

Das Konfigurationsmodell ist ein sehr einfaches, direkt generierendes Verfahren (siehe [8] Abschnitt IV.B.1). Die Idee besteht darin, alle Knoten zu Beginn mit der entsprechenden Anzahl an Kantenstummeln zu versehen, und danach iterativ jeweils zwei zufällig ausgewählte Stummel mittels einer Kante miteinander zu verbinden. Dies funktioniert in linearer Zeit, generiert aber zum Einen auch nicht-einfache Graphen, und kann auch die Verbundenheit des Netzwerkes nicht garantieren.
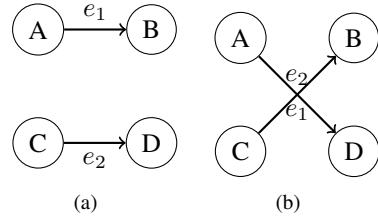

Fig. 1. gültiger Kantentausch

Um dennoch einfache Graphen zu erhalten, werden in der Praxis parallele Kanten und Selbst-Kanten einfach gelöscht, was aber sowohl die Anzahl der Kanten als auch die Gradsequenz des damit generierten Netzwerkes verfälscht. Dies wird in der Praxis bei großen Netzwerken gerne als vernachlässigbar dargestellt. Zwar mag dies oftmals stimmen, dennoch gibt es Erkenntnisse, welche Probleme bei der Uniformität des Verfahrens nachweisen [5]. Insgesmat sind wir der Meinung, dass das Verfahren unter strengsten wissenschaftlichen Kriterien als nicht zuverlässig betrachtet werden muss.

Die Verbundenheit der Graphen kann in keinster Weise garantiert werden, aber auch der Versuch, alle nicht verbundenen generierten Graphen zu verwerfen ist für fast alle real vorkommenden Gradsequenzen hoffnungslos, wie fundierte Abschätzungen über den Anteil nicht verbundener Graphen für gegebene Sequenzen zeigen [9]. Benötigt man in der Praxis dennoch diese Einschränkung, so ist es üblich, nur den größten zusammenhängenden Teilgraphen des Netzwerkes zu nehmen. Eine solche *giant component*, welche nahe an die vorgegebene Größe des Netzwerkes herankommt, wird es zwar meistens mit hoher Wahrscheinlichkeit geben [10], das Resultat verliert aber noch weiter an Zuverlässigkeit.

Von Uniformität kann man bei diesem Verfahren kaum noch sprechen, da die generierten Netzwerke gar nicht in der vorgegebenen Menge der Konfigurationen liegen. Hier werden stattdessen die statistischen Werte bestimmter Metriken als Vergleich herangezogen, die Uniformität also rein empirisch betrachtet.

D. Der MCMC-Algorithmus

Der Markov-Ketten Monte-Carlo Algorithmus ist ebenfalls ein Verfahren für die Generierung entsprechender Konfigurationen. Die Idee ist hierbei, aus einer Startkonfiguration heraus solange zwei zufällig ausgewählte Kanten miteinander zu vertauschen, bis jede Konfiguration mit der gleichen Wahrscheinlichkeit erreicht wird.

Als Startkonfiguration für das Verfahren nimmt man entweder das reale Netzwerk, oder, falls nur eine Gradsequenz gegeben ist, generiert man eine Konfiguration mit dem Havel-Hakimi-Algorithmus [11]. Beim Tausch zweier zufällig ausgewählter Kanten gibt es gültige und ungültige Folgekonfigurationen. Ungültig ist ein Kantentausch dann, wenn die Folgekonfiguration kein einfacher Graph mehr wäre, dies lässt sich jedoch sehr leicht anhand der ausgewählten Kanten vorab überprüfen, wie in den Abbildungen 1 und 2 zu sehen ist.
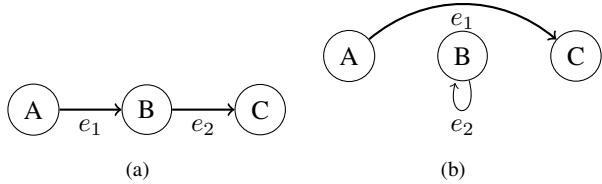
Fig. 2.   ungültiger Kantentausch wegen Schleife

Fig. 3.   Empirische Schätzung des Vorfaktors f_q aus [5]

Aufgrund theoretischer Gegebenheiten konvergiert dieses Verfahren bei einer Schrittzahl q von erfolgreichen oder versuchten Kantentauschen gegen unendlich zu einer perfekt uniformen Generierung aller Konfigurationen. Doch genau bei der Bestimmung dieser Schrittzahl q liegt das Problem dieses Verfahrens, wie bei vielen anderen Markov-Ketten-Verfahren auch. Nachgewiesen ist, dass eine Schrittzahl in der Größenordnung von $O(m^2)$ für die Uniformität ausreichend ist, jedoch gibt es keine verlässliche Aussage über den Vorfaktor. Empirische Tests legen gar eine lineare Größenordnung nahe, und schlagen einen Vorfaktor f_q von 100 dazu vor [5], [12], also insgesamt eine Schrittzahl von $q = 100 \cdot m$. Wir werden in Kapitel III genauer auf diese Problematik und Überlegungen zu möglichen Auswegen eingehen.

Vorab wollen wir noch kurz auf die Generierung von ausschließlich verbundenen Graphen eingehen, welche mit dem MCMC-Algorithmus unter Wahrung der Uniformität möglich ist, was das Verfahren deswegen auch so attraktiv macht. Nehmen wir an, die Schrittzahl liege tatsächlich in $O(m)$. Man kann nun nach jedem Kantentausch überprüfen, ob die Nachfolgekonfiguration ein verbundener Graph ist oder nicht. Dies kann mittels einer einfachen Tiefensuche in m Schritten entschieden werden. Davon abhängig gilt der Kantentausch dann entweder als gültig, oder aber man geht wieder zur ursprünglichen Konfiguration zurück und versucht erneut einen Kantentausch, ohne den vorangegangenen mitzuzählen. Dies führt insgesamt zu einer Laufzeit des Verfahrens in $O(m^2)$. Viger und Latapy haben diese Methode weiterentwickelt, und nachweislich eine Erhöhung um nur O(log m) für die Verbundenheitsprüfung erreicht, so dass ihr optimierter MCMC-Algorithmus verbundene Netzwerke mit einer Laufzeit in $O(m \cdot log\ m)$ generiert.

Bestünde nun Gewissheit über die für die Uniformität tatsächlich notwendige Schrittzahl, so würde dieser Algorithmus zuverlässig und schnell die gewünschten verbundenen Netzwerke generieren. Tatsächlich wird dieser Algorithmus aber bis heute in keinem der Standardwerkzeuge und -bibliotheken angeboten. Stattdessen findet man stets eine Implementierung des Konfigurationsmpdells vor, trotz der oben beschriebenen Unzulänglichkeiten. Im nächsten Kapitel werden wir unsere Ideen für eine bessere Bestimmung der Schrittzahl vorstellen, um so einen Beitrag für die Etablierung der MCMC-Algorithmen zu leisten.

III. Die Schrittzahl des MCMC-Algorithmus

In diesem Kapitel stellen wir den Stand der Wissenschft bezüglich der Bestimmung der Schrittzahl für den MCMC-Algorithmus vor, so dass dieser möglichst uniforme Ergebnisse liefert, aber nicht übermäßig viel Zeit braucht. Wir hinterfragen diesen kritisch und zeigen anhand einiger Beispiele Unzulänglichkeiten und mögliche Lösungswege hierfür auf.
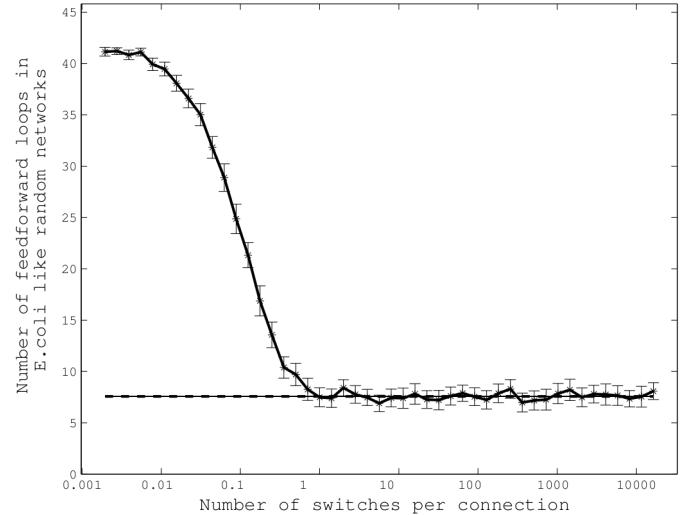
A. Bisherige Festlegung

Dass die Schrittzahl q für den MCMC-Algorithmus linear von der Anzahl der Kanten m abhängt ist bisher wie gesagt nur eine Vermutung. Um unter dieser Voraussetzung den passenden Vorfaktor f_q zu ermitteln, wurde in vorangegangenen Arbeiten [5], [12] stets eine indirekte empirische Methode verwendet. Das Prinzip beruht darauf, sich eine Startkonfiguration mit einer sehr ungewöhnlich ausgeprägten Metrik zu nehmen, davon ausgehend wiederholt Kantentausche durchzuführen, und zu messen, ab welcher Schrittzahl die Metrik sich beim erwarteten Durchschnittswert einpendelt. Hieraus wird dann eine ausreichende Unifomität des Verfahrens geschlussfolgert.

Abbildung 3 aus [5] zeigt eine solche Bestimmung anhand des Transkriptionsnetzwerkes des E.Coli-Bakteriums. Als Metrik wurde die gefundene Anzahl des Motivs „Feed-Forward Loop" gewählt. Man sieht recht gut, dass der Erwartungswert gleichartiger Zufallsnetzwerke in diesem Beispiel schon mit $f_q = 1$ erreicht wird, dennoch ist es die Schlussfolgerung der Autoren, im allgemeinen Fall lieber $f_q = 100$ zu wählen.

Das Problem dieser Art der Bestimmung sind die Auswahl der Metrik und einer dazugehörigen außergewöhnlichen Startkonfiguration. Nun haben schon Watts und Strogatz 1998 gezeigt, dass verschiedene Metriken verschieden schnell auf Kantentausche reagieren [3], eine unter allen Umständen „trägste" Metrik ist jedoch nicht bekannt. Dennoch erscheint es fragwürdig, eine besonders dynamische Metrik wie den Durchmesser eines Netzwerkes für solch einen Test heranzuziehen, wie in Vorarbeiten geschehen [12].

Das andere Problem betrifft die Startkonfiguration. Hat das Transkriptionsnetzwerk zwar mit 42 Motiven einen z-Wert von ca. 10 bei der Abweichung von der erwarteten Motivanzahl,

so lassen sich bei der gegebenen Sequenz dennoch auch leicht Konfigurationen mit über 100 Motiven realisieren, was das Konvergieren aus einer solchen Konfiguration heraus weiter verlangsamen würde. Auch die Größen der Netzwerke bleiben unzureichend berücksichtigt, da diese Art der Bestimmung sehr zeitaufwändig ist. Das kommunizierte Ergebnis ist hier immer, dass „für eine Vielzahl verschiedener Netzwerke und verschiedener Metriken sich $f_q = 100$ immer als ausreichend groß herausgestellt hat".

Wir sind mit diesem Ergebnis bezüglich der Zuverlässigkeit des Verfahrens nicht zufrieden, und die Zurückhaltung bei der Akzeptanz dieses Verfahrens scheint dies zu bestätigen. Wir stellen in den folgenden Abschnitten unseren alternative Ansatz für die Bestimmung der Schrittzahl q vor.

B. Die Markov-Kette

Zu Beginn unserer Überlegungen beschäftigen wir uns mit der dem MCMC-Algorithmus zu Grunde liegenden Markov-Kette. Jede mögliche Konfiguration einer Gradsequenz ist ein Knoten der Markov-Kette, den wir zur besseren Unterscheidung *Konfigurationsknoten* nennen werden. Jeder gültige Kantentausch führt zu einem anderen Konfigurationsknoten, und die Umkehroperation auch immer wieder zurück, so dass immer bidirektionale Verbindungen zwischen zwei Konfigurationsknoten bestehen. Alle ungültigen Kantentausche zählen ebenfalls als Schritt des Algorithmus, da sie aber zu keiner neuen Konfiguration führen, werden sie deshalb in einer Selbst-Kante mit entsprechender Gewichtung zusammengefasst.

Es gibt insgesamt $\frac{m^2+m}{2}$ ungeordnete Kantenpaare, so dass jeder Konfigurationsknoten diesen Wert als Ein- und Ausgangsgrad besitzt. Da auch die Verbundenheit der Markov-Kette bewiesen ist, führt dies theoretisch zu einer perfekten Uniformität. Entscheidend hierfür ist in der Praxis aber die notwendige Schrittzahl, bis zum sogenannten *Mixing* der Markov-Kette.

C. Exakte Berechnung im Einzelfall

Für gegebene Gradsequenzen und sehr kleine Netzwerke lässt sich die Markov-Kette mit normaler Hardware noch vollständig erzeugen. Verfügt man über diese Markov-Kette, so kann man die für deren Mixing notwendige Schrittzahl mit einer Variante des einfachen ungedämpften PageRank-Algorithmus [13] exakt berechnen. Hierfür wird initial nicht jedem Knoten der Wert 1 gegeben, sondern der Startkonfigurationsknoten erhält den Wert m, und alle anderen Konfigurationsknoten den Wert 0. Nun werden soviele Iterationen q des ungedämpften PageRank-Algorithmus durchgeführt, bis der Wert aller Knoten $1 \pm \delta$ beträgt.

Die Werte entsprechen der Wahrscheinlichkeit geteilt durch m, nach dem jeweiligen Schritt des MCMC-Algorithmus bei dieser Konfiguration zu sein. Der Wert δ gibt die maximal auftretende bzw. tolerierte Abweichung einer Konfiguration von der geforderten Wahrscheinlichkeit an.

Wir untersuchen dies für vier Gradsequenzen mit 12 Kanten. Zum Einen die des zur Messung der Uniformität in [5]

TABLE I
EXAKT BERECHNETE SCHRITTZAHLEN

Netzwerk	q	f_q	$\pm \delta$
negativ korreliert	61	~ 5	0.009
positiv korreliert	89	~ 7	0.02
zufällig	78	~ 7	0.15
Toy-Netzwerk	21	~ 2	0

benutzten „toy network", welches über eine sehr spezielle, für die Realität untypische Sequenz verfügt, aber nur 91 Konfigurationen erlaubt. Zum anderen haben wir uns, unserem Interesse an real vorkommenden Sequenzen folgend, eine kleine heterogene Powerlaw-Verteilung (siehe [8] Abschnitt III.C.1) mit 12 Kanten generiert, wie sie in einem skalenfreien Netzwerk vorkommen würde. Wir haben $\alpha = 1,7$ gewählt und damit die Knotengrade $(3, 2, 2, 1, 1, 1, 1, 1)$ für 8 Knoten erhalten. Um eine Gradsequenz für einen gerichteten Graphen mit Ein- und Ausgangsgraden zu erhalten, haben wir diese einmal positiv, einmal negativ, und einmal nicht korreliert, also zufällig, auf die acht Knoten verteilt.

Diese vier Sequenzen haben wir mit dem Havel-Hakimi-Algorithmus in einem Netzwerk realisiert, von diesen Konfigurationen aus jeweils den Zustandsraum der Markov-Kette durch Verfolgung aller jeweils möglichen 66 Kantenpaare vollständig generiert und als Netzwerk gespeichert. Die Markov-Kette des Toy-Netzwerks verfügt wie bekannt über 91 Konfigurationsknoten, die Markov-Ketten der drei skalenfreien Netzwerke über 90.000 bis 120.000 Konfigurationsknoten.

Tabelle I listet die errechneten Schrittzahlen q für die vier Beispielnetzwerke auf. Abbruchbedingung war hier, dass sich die Werte nicht mehr ändern, was bei der 32-Bit-Darstellung schon vor Erreichen der Gleichverteilung eintreten kann. Der δ-Wert gibt hierzu die maximale Abweichung eines Konfigurationsknotens vom Wert 1 an. Unter Annahme einer von m linear abhängigen Schrittzahl des MCMC-Algorithmus würde der entsprechend angegebene Vorfaktor f_q gelten.

In allen Fällen liegt f_q weit unter 100, aber auch höher als 1, wie es in Abbildung 3 möglicherweise suggeriert wurde. Dies sind aber Einzelfallbetrachtungen, aus denen man noch keine Verallgemeinerungen folgern kann.

D. Wege zur Verallgemeinerung

Die exakte Berechnung der Schrittweite ist schon bei kleinen Werten von m aufgrund der exponentiell steigenden Anzahl an Konfigurationen unpraktikabel. Um mit dieser Methode dennoch verallgemeinerbare Aussagen, zumindest für bestimmte Klassen von Netzwerken, wie z.B. die skalenfreien Netze, treffen zu können, ist noch ein weiterer Schritt notwendig. Mit einem Verständnis für die Struktur der zugrundeliegenden Markov-Ketten ließe sich die Schrittzahl für das Mixing dieser Ketten über das modifizierte PageRank-Verfahren direkt abschätzen und simulieren, anstatt sie indirekt über das beschriebene empirische Verfahren zu bestimmen.

Hierbei indentifizieren wir 3 wichtige Punkte, welche für die Geschwindigkeit des Kovergieren des PageRank-Verfahrens

von entscheidender Bedeutung sind.

1) die Anzahl der externen Verbindungen eines Konfigurationsknotens, d.h. die Anzahl der Nachbarkonfigurationen und deren Verhältnis zum Gewicht der Selbst-Kante
2) den Clustering-Koeffizienten der Markov-Kette
3) den Durchmesser der Markov-Kette

Wären diese drei Ausprägungen abhängig von der zu Grunde liegenden Gradsequenz abschätzbar, so ließe sich für eine Klasse von Gradsequenzen ein Modell der Markov-Kette aufstellen, und anhand von Simulationen auf verschieden großen Instanzen dieses Modells die notwendige Schrittzahl abschätzen oder gar ohne Simulation direkt berechnen.

Aus unseren bisherigen Erkenntnissen können wir bereits erste Hypothesen aufstellen, deren genauere Untersuchung unsere zukünftige Arbeit sein wird.

Die Topologie der in Abschnitt III-C untersuchten Markov-Ketten der skalenfreien Netze bilden ein mehrdimensionales relativ regelmäßiges Gitter, welches entlang der Dimensionen rundum verbunden ist, ähnlich wie die Oberfläche einer Kugel in zweidimensionaler Sicht. Dies ist offensichlich im allgemeinen Fall so, bedarf jedoch noch eines Nachweises. Definiert wird diese regelmäßige Struktur durch die oben genannten 3 Eigenschaften.

Beispielsweise zeigt Abbildung 4 die Verteilung der externen Verbindungen der Konfigurationsknoten der Markov-Kette der nicht-korrelierten skalenfreien Gradsequenz, welche eine nur geringe Schwankung um den Wert 32 aufweist. Die Anzahl ungültiger Kantentausche ließe sich allgemein gut aus der Sequenz abschätzen, da Kantenpaare, welche nur 3 Knoten involvieren, nie getauscht werden können. Der theoretisch maximal mögliche Durchmesser von $m - 1$ scheint von den skalenfreien Netzen voll ausgereizt zu werden, unsere Beispielsequenzen hatten hier alle den Wert 10. Der Clustering-Koeffizient ist mit ca. $0,1$ vergleichsweise hoch, was durch den hohen Durchmesser bedingt ist, da benachbarte Konfigurationen viele Nachbarn gemeinsam haben.

Diese beiden Metriken lassen sich, die gewisse Regelmäßigkeit vorausgesetzt, durch Stichproben-Generierunge einzelner Konfigurationen und deren Nachbarschaft ermitteln und statistisch abschätzen. Dies wäre in vertretbarer Zeit berechenbar, und würde ebenfalls eine Abschätzung erlauben, und somit eine empirische Ermittlung der notwendigen Schrittzahl auf dem Modell erlauben.

IV. Zusammenfassung

In diesem Papier haben wir den Stand der Wissenschaft bezüglich der Evaluationsmethoden von konkreten realen Netzwerken zusammengefasst, und die unserer Meinung nach vorzuziehende Methode, der Evaluation mittels gleichartiger Zufallsnetzwerke kritisch analysiert. Die aufgezeigten Probleme bei der Zuverlässigkeit der gängigen Methoden wurden bewertet, und für das unserer Meinung nach überlegene MCMC-Verfahren haben wir einen Ansatz vorgestellt, mit welchem das letzte offene Problem, der zu wählenden Schrittzahl q zumindest für bestimmte Klassen von Netzwerken deutlich
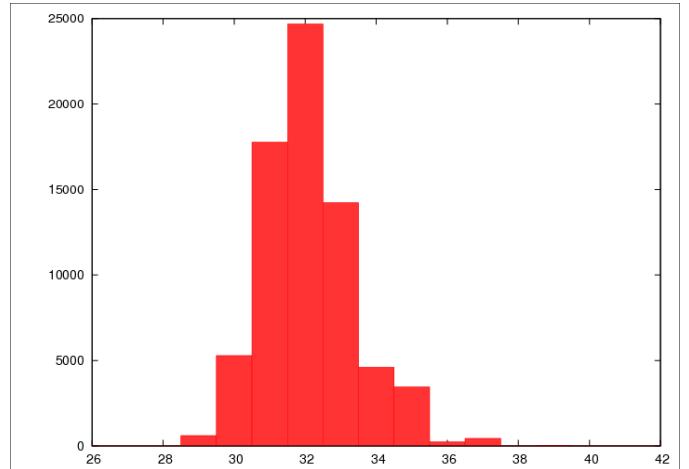

Fig. 4. Gradverteilung in der Markov-Kette des nicht-korrelierten skalenfreien Gradsequenz

reduziert werden könnte. Wir werden versuchen, die dafür notwendigen Kenngrößen in Zukunft in ausgiebigen Simulationen abschätzen zu können, und damit eine für jede Netzwerkgröße dieser Klasse sichere und vor allem effiziente Wahl von q ermöglichen zu können.

References

[1] S. Wasserman, K. Faust, and D. Iacobucci, *Social Network Analysis : Methods and Applications (Structural Analysis in the Social Sciences)*. Cambridge University Press, 1994.

[2] L. C. Freeman, *The Development of Social Network Analysis: A Study in the Sociology of Science*. Empirical Press, 2004.

[3] D. Watts and S. Strogatz, "Collective dynamics of small-world networks," *Nature*, no. 393, pp. 440–442, 1998.

[4] D. Obradovic and S. Baumann, "A journey to the core of the blogosphere," in *Proceedings of the International Conference on Advances in Social Network Analysis and Mining (ASONAM 2009)*. IEEE, 2009, pp. 1–6, best paper award.

[5] R. Milo, N. Kashtan, S. Itzkovitz, M. E. J. Newman, and U. Alon, "On the uniform generation of random graphs with prescribed degree sequences," *Arxiv preprint cond-mat/0312028*, 2003.

[6] A. L. Barabasi and R. Albert, "Emergence of scaling in random networks," *Science*, vol. 286, pp. 509–512, 1999.

[7] T. Snijders, "Enumeration and simulation methods for 0-1 matrices with given marginals," *Psychometrika*, vol. 56, no. 3, pp. 397–417, September 1991.

[8] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, pp. 167–256, 2003.

[9] M. Molloy and B. Reed, "A critical point for random graphs with a given degree sequence," *Random Structures and Algorithms*, vol. 6, pp. 161–179, 1995.

[10] ——, "The size of the giant component of a random graph with a given degree sequence," vol. 7, p. 295–305, November 1998.

[11] P. L. Erdős, I. Miklós, and Z. Toroczkai, "A simple Havel-Hakimi type algorithm to realize graphical degree sequences of directed graphs," *Electronic Journal of Combinatorics*, vol. 17, no. 1, p. R66, 2010.

[12] F. Viger and M. Latapy, "Efficient and simple generation of random simple connected graphs with prescribed degree sequence." in *Proceedings of the 11th international conference on Computing and Combinatorics*, ser. Lecture Notes in Computer Science, L. Wang, Ed., vol. 3595. Springer, 2005, pp. 440–449.

[13] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," Stanford University," Technical Report, 1998.

Mathematical Optimization Based Channel Coding: Current Achievements and Future Challenges

Michael Helmling[#], Stefan Scholl[*], Akın Tanatmis[#]

[*]*Department of Electrical and Computer Engineering*
[#]*Department of Mathematics*
University of Kaiserslautern, 67653 Kaiserslautern, Germany
{helmling, tanatmis}@mathematik.uni-kl.de
scholl@eit.uni-kl.de

Abstract— **Channel coding and mathematical optimization— two omnipresent branches of science which heavily influence our everyday life, which is certainly unimaginable without the epochal achievements of each of the two disciplines since they affect nearly every communication system as well as every transportation, manufacturing, and organization process. The following report is dedicated to some of the achievements of a research project decisively influenced by a cooperative interplay of these two disciplines.**

I. CHANNEL CODING

Channel coding is an important technique for the correction of transmission errors that is used in nearly every communication system today. It is used in mobile phones as well as in satellite communication, navigation systems, storage devices like hard disks, CDs and also in internet or broadcast applications. Let us first review what channel coding is and why we need it. To understand the problems channel coding solves, we will have a look at the old days. As an example we take a communication system without channel coding that becomes more and more obsolete: analog TV.

Television broadcasting is a form of one way communication. On the one hand we have a transmitter, which is usually placed on top of a mountain and transmits radio waves with high power to reach an area as big as possible. On the other hand we have the receivers, ordinary TV sets, in our households, that pick up the radio waves.

If the connection between transmitting antenna and receiving antenna is of good quality, we a see clear picture and hear a comfortable sound. But what happens if the quality of the connection is not so good? Let us assume our receiver is placed far away from the transmitter, behind a mountain, in the cellar or simply in bad weather, so that the reception quality drops. Then the picture is distorted or, even worse, the whole picture bounces up and down. Also the sound might be clicking and crackling, disturbing the viewers enjoyment. The TV signal is noisy or distorted.

Figure 1 shows such a situation where we have two TV watchers, A and B. Whereas A is close to the transmitter and receives a clear picture, B is not in such a good position. He has worse receiving conditions due to serveral obstacles.

The goal of new digital TV broadcasting systems is to make picture and sound more resistant to distortion and noise. In
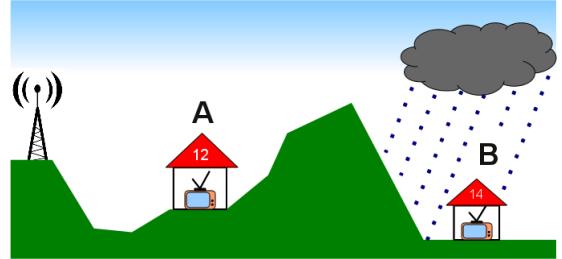
Fig. 1.   TV transmitter and two receivers

Fig. 2.   TV pictures from A and B

contrast to the presented analog system, there are additional possibilities for techniques to improve the link from the transmitter to the TV set at home. One of them is channel coding. Channel coding is a technique to detect and correct occuring errors. But how is that possible?

Whenever a signal is transmitted in modern digital communication systems, the data is transmitted in form of bits or bitstreams, respectively. To allow error correction on the receiver side, we have to make some preprocessing before transmission, called *encoding*. To encode the data, the bitstream is first cut into small blocks. At the end of every block some extra bits, called parity bits, are added as shown in Figure 3. The k data bits together with the r parity bits form a vector of n bits which is called codeword the **c**.
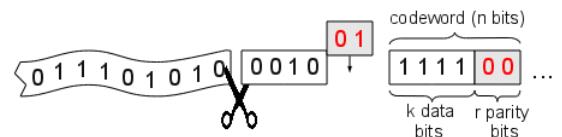
Fig. 3.   A bitstream is being encoded

During encoding the additional parity bits are calculated

from the data bits. The "way" we calculate the parity bits is called *channel code* and determines the performance of the channel coding system. Since data is processed block by block, the channel codes described here are called *linear block codes*. A linear block code is defined by a binary $r \times n$ matrix H, so that every bitvector $\mathbf{c}$ that fullfills

$$H \cdot \mathbf{c} = 0 \qquad (1)$$

(where the calculation takes place modulo 2) is a codeword. H is called *parity-check matrix*. By $\mathcal{C} = \{\mathbf{c} \in \{0,1\}^n : H \cdot \mathbf{c} \equiv 0 \bmod 2\}$ we denote the set of all codewords [1]. Finding a good channel code with its matrix H is non-trivial and subject to research since decades.

After encoding the data it is sent to the receiver. During transmission some bits may be distorted and change their values.

The actual correction of corrupted data takes place in the receiver by a device called channel decoder, that tries to "repair" the incoming blocks. For this purpose the decoder uses the previously added parity bits. In many cases this technique works well and many errors can be successfully corrected.
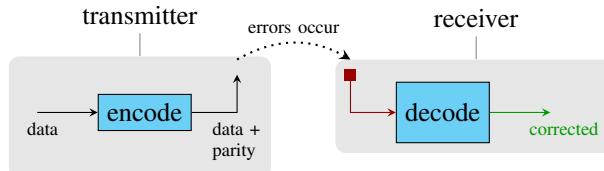

Fig. 4.   Encoding, transmission, and decoding of data

However there is a probability that the decoder cannot correct an erroneous block. This remaining error probability is called frame error rate (FER) and is dependent on the signal-to-noise ratio (SNR), which is the signal power divided by the noise power at the receiver. The FER is an indicator for the quality of the channel coding system. It is measured with Monte-Carlo simulation on a computer and visualized with a special chart, shown in Figure 5. This type of chart has a typical shape. For weak signals, that is, low SNR, the FER is very high (close to 1) and drops with increasing signal strength. In Figure 5 there are two curves, the blue one represents a system without channel coding, the red one corresponds to the same system with channel coding. One can easily see that the channel coding provides significantly less errors. The lower a curve lies in the performance chart, the better the coding system works. In fact there are many parameters that influence the performance of channel coding, but the performance chart allows for a good comparison.

In practical applications different error probabilities are necessary. For audio signals, like in a mobile, a minimal error probability 0.01 is required, otherwise crackling or interruptions make a call uncomfortable. For communications through optical fiber error rates of less than 10^{-12} may be necessary for flawless operation.
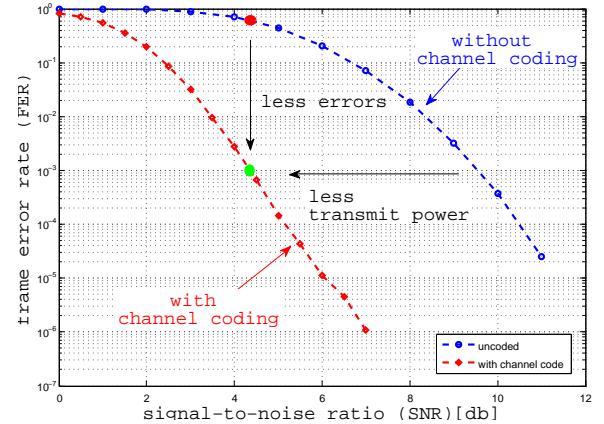

Fig. 5.   FER for a system without channel coding and one using channel coding

Using channel coding we can handle more weak signals without any loss of quality or provide much more reliable data for a given signal strength. This advantage can be translated into equipment cost reduction, for example smaller antennas, cheaper electronics or smaller batteries.

We now have seen which processing steps in channel coding are necessary, namely encoding and decoding. The decoder is much more complex than the encoder and is the main part of a channel coding system. So now we will have a closer look at the decoding process.

In channel coding the data is processed in blocks of n bits consisting of k data bits and $r = n - k$ parity bits. Because the parity bits are calculated from the data bits, they are fully dependent on the data bits. Therefore there are only 2^k possible codewords, although there are 2^n possible bit combinations. When a erroneous codeword is received, the channel decoder checks which one of the 2^k valid codewords is most similar to the received block. This process is called maximum likelihood (ML) decoding and is the optimal way of decoding [1]. However, because all 2^k codewords need to be tested, this procedure is applicable only for small values of k. With increasing k the complexity grows exponentially. Since, in practice, we deal with pretty large blocks (for example $k = 100$ or $k = 10000$), the ML decoding procedure of going through all 2^{100} or even 2^{10000} codewords is simply not possible to calculate in an imaginable amount of time.

To overcome this problem engineers often use heuristics or suboptimal algorithms for the decoding procedure. These algorithms are very often capable of finding the correct solution, but with much less complexity. In some cases, however, the suboptimal algorithms decide for a wrong solution. This leads to a small degradation in performance, which one has to accept if the problem should be feasible.

Only if the complexity is low enough, the decoding algorithm can be implemented on a chip for use in practical applications. For many of modern codes, algorithms with a good tradeoff between performance and complexity are known. Moreover, for many used channel codes only the

performance with suboptimal decoding algorithms is known—not the one of optimal ML decoding.

From a mathematicans point of view there exist other means of solving the ML decoding problem than comparing all 2^k codewords to the received one. The decoding can be formulated as follows.

Assume that at the receiver a vector $\tilde{\mathbf{c}}$ was received. It will check all possible sent codewords $\mathbf{c}$ out of $\mathcal{C}$ and decide for the codeword $\mathbf{c}^*$ which was most likely sent. Thus, the ML codeword $\mathbf{c}^*$ is the codeword which maximizes the so called conditional probability $p(\tilde{\mathbf{c}} \text{ received}|\mathbf{c} \text{ sent})$:

$$\mathbf{c}^* = \arg\max_{\mathbf{c} \in \mathcal{C}} p(\tilde{\mathbf{c}} \text{ received}|\mathbf{c} \text{ sent}) \qquad (2)$$

This problem is an optimization problem. In the following, we will present how decoding can be done with methods of integer optimization.

II. Optimization Methods for ML Decoding

Mathematical optimization is the task of finding an optimal solution for and the optimal value of an objective function under the condition that some constraints are satisfied. Objective function and side constraints typically model some real-world scenario, where the variables of the function are the abstraction of parameters like the per-day work time of different machines or the path on which a truck drives from location A to location B. Depending on the task, "optimal" may be either minimal (if the objective function represents some sort of cost or time) or maximal (e.g., if the function value stands for the profit gained). The constraints are used to model limitations of resources, working time, money, etc., or to avoid solutions that would be mathematically correct, but do not make sense in the real world—for example, a machine cannot work -2 hours per day, and one cannot build 1.72 facilities.

Linear programming is the best-studied discipline of mathematical optimization and it deals with optimization of a *linear* objective function which is constrained by linear (in)equalities. The domain of the variables is $\mathbb{R}^n$ for some $n \in \mathbb{N}$ and a *linear program (LP)* is then defined as

$$\min (\text{or } \max) \ w^T x = \sum_{i=1}^n w_i \cdot x_i$$
$$\text{such that} \quad Ax \leq b$$
$$x \in \mathbb{R}^n$$

where $w \in \mathbb{R}^n$ is the cost (benefit) vector, A the $m \times n$ constraint matrix, $b \in \mathbb{R}^m$ the right hand side vector of the constraints, and the variables are real-valued. Linear programming has become famous since George Dantzig in 1947 proposed an algorithm to solve LPs as the above one, called the Simplex algorithm. Despite the fact that the Simplex algorithm was later proven to have exponential worst-case complexity, it is still the algorithm most used in practice. For the far most problem instances, the Simplex method is very fast, and outperforms polynomial-time algorithms that were developed in the 1980's. Since its invention, linear programming has

become a standard tool in business for a wide range of applications such as planning, scheduling, routing or facility location.

Yet, this is not the end of the story—for many applications, linear programs lack some important capabilities, such as the restriction to *integral* or *binary* decision variables.

From this requests, the field of *integer linear programming* evolved. An integer program (IP) is defined in the same way as a linear program, but additionally requires that some (or all) of the variables take values in $\mathbb{Z}^n$ instead of $\mathbb{R}^n$. The increased power obtained by this extension vastly enlarges the number of practical problems that can be modelled by IPs—at the cost, however, of efficiency, since it was shown that the general integer program is NP-hard to solve. Nevertheless, since this way of modelling is so powerful, a lot of research has been done to find and improve algorithms that tackle integer programs and make them solvable at least for moderate problem sizes or specific well-understood classes of programs.

How can integer linear programming techniques be used to decode binary linear codes and to add some surplus to the state-of-the-art knowledge in channel coding?

Jon Feldman was one of the first researchers who addressed this question intensively. To this end, he reformulated the ML problem (2) and interpreted it as an integer programming problem: Under the assumption of a *memoryless* channel, i.e. the error distributions are independent for subsequent bits, it can be calculated as

$$\mathbf{c}^* = \arg\max_{\mathbf{c} \in \mathcal{C}} \left(\prod_{i=1}^n p(\tilde{\mathbf{c}}_\mathbf{i}|\mathbf{c}_\mathbf{i}) \right)$$
$$= \arg\min_{\mathbf{c} \in \mathcal{C}} \left(\sum_{i:c_i=1} \ln\left(\frac{p(\tilde{\mathbf{c}}_\mathbf{i}|0)}{p(\tilde{\mathbf{c}}_\mathbf{i}|1)} \right) \right).$$

The quotient

$$y_i = \ln\left(\frac{p(\tilde{\mathbf{c}}_\mathbf{i}|0)}{p(\tilde{\mathbf{c}}_\mathbf{i}|1)} \right) \qquad (3)$$

is called the *log likelihood ratio* (LLR) of the i-th input bit. It only depends on the parameters of the channel, which are assumed to be known, and the signal $\tilde{\mathbf{c}}$ that was received. So, y_i is known to the decoder. From the above calculation it follows that $\mathbf{c}^* = \arg\min_{\mathbf{c} \in \mathcal{C}} \left(\sum_{i=1}^n y_i \mathbf{c}_\mathbf{i} \right)$.

ML decoding can be achieved by minimizing this linear objective function over all codewords $\mathbf{c}$ that satisfy (1); in other words,

$$\min \sum_{i=1}^n y_i \mathbf{c}_\mathbf{i}$$
$$\text{such that} \quad H \cdot \mathbf{c} \equiv 0 \bmod 2$$
$$\mathbf{c} \in \{0, 1\}^n.$$

Due to the modulo constraints, this is not exactly an integer programming model of the ML decoding problem. However, there exist several integer programming re-formulations of this optimization problem. For example, Jon Feldman introduced a formulation in which the number of both variables and constraints grows exponentially in the number of ones per

row [2]. Among others, this exponential growth led us to the development and investigation of the following improved integer programming model with less constraints and variables by the introduction of artificial variables $z \in \mathbb{Z}^m$ (one for each row of H)

$$\min \sum_{i=1}^{n} y_i \mathbf{c_i}$$
$$\text{such that} \quad H \cdot \mathbf{c} - 2z = 0 \qquad (4)$$
$$\mathbf{c} \in \{0,1\}^n, \ z \in \mathbb{Z}^m$$

The j-th row of the constraint system reads $\sum_{i=1}^{n} H_{i,j} \cdot \mathbf{c_i} - 2z_j = 0$. Since z_j is required to be integral, this equation ensures that an even number of entries of c_i is set to 1. So, indeed this is a simple, general, and sparse integer programming problem for the ML decoding problem [3]. For any given code specified by a parity check matrix H, the solution to this problem is the maximum likelihood codeword.

It should be emphasized that this problem is difficult to solve (it is NP-hard in general). Therefore, the integrality requirements on the variables $\mathbf{c}$ and z are dropped in order to get the so-called LP-relaxation of this IP which can be considered an approximation of the original problem.

Our IP formulation proved to be beneficial in several ways.

First, in [4], we solved this IP formulation by a commercial all-purpose solver and compared its results with those of five other IP formulations existing in the literature. To sum it up, our formulation had the fewest number of variables as well as constraints and the solution time was the least.

Second, this formulation allows for an efficient procedure to approximately solve this problem (this is important if commercial solvers are not efficient enough or cannot be used due to some constraints). In [3], a separation algorithm was proposed utilizing the structural properties of this IP formulation. This algorithm works as follows: First, the LP solution is computed and checked for integrality. If it is integral, it is obviously also the optimal IP solution and the algorithm terminates. Otherwise, the IP-formulation is improved by adding additional linear constraints $\lambda^T \mathbf{c} \leq \lambda_0, (\lambda, \lambda_0) \in \Lambda$ where $\Lambda \subset \mathbb{R}^{n+1}$ is a set depending on some polyhedral properties of the code (which shall not be specified due to intended brevity of this article). However, these constraints ensure that if the LP obtained by adding this constraint is solved, then the previous LP solution is infeasible and the new LP solution is closer to the IP solution. Applying this idea iteratively yields an algorithm which approximates the IP solution. Consequently, the LP to be solved repeatedly is of the form

$$\min \sum_{i=1}^{n} y_i \mathbf{c_i}$$
$$\text{such that} \quad H \cdot \mathbf{c} - 2z = 0$$
$$\lambda^T \mathbf{c} \leq \lambda_0 \quad (\lambda, \lambda_0) \in \Lambda$$
$$0 \leq \mathbf{c_i} \leq 1, \quad i = 1, \dots, n.$$
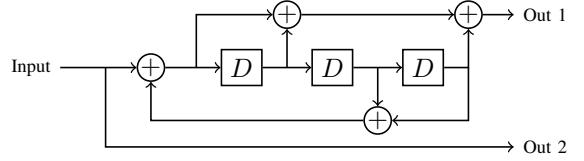
Fig. 6. A convolutional encoder with 3 memory registers.

The constraints $\lambda^T \mathbf{c} \leq \lambda_0, (\lambda, \lambda_0) \in \Lambda$, are derived from the current LP solution based on some combinatorial reasoning and the indicator variables z allow for a very fast generation of these constraints. Moreover, these constraints can be classified with respect to optimization theory: although they are derived in a combinatorial manner, they correspond to what is known as *Gomory cuts*. Additionally, it turned out that the formulation (4) offers another possibility which leads to an algorithm with enhanced performance: The representation of the code in terms of a parity check matrix H is not unique. In our separation algorithm, alternative parity check matrices of the code are generated during the execution as soon as the current one does not allow for further improvement of the solution.

To sum it up, this formulation enabled the development of an algorithm which

- is applicable to *any* binary linear block code in contrast to existing heuristics which are tailored to one specific type of code,
- yields a better error-correcting performance than other algorithms, and
- employs properties of binary linear codes to improve the algorithmic treatment.

Besides this general purpose formulation, we have also invented and investigated specialized approaches for an important subclass of codes, as described below.

III. LINEAR PROGRAMMING FOR TRELLIS-BASED CODES

A class of codes that have gained tremendous interest since their invention in 1993 are the so-called *turbo codes*. A turbo code is built by the parallel concatenation of two convolutional codes that are separated by an interleaver. Convolutional codes have an inherent graph structure by their *trellis representation* which faciliates the processes both of encoding and decoding. This structure is partly inherited by the more complex turbo codes, which allowed us to invent a powerful integer programming formulation, as well as several approximation algorithms, for this special set of codes.

The encoder of a convolutional code is built by a number of delay shift registers (D) and several XOR (binary addition) gates, as shown in Figure 6. The registers comprise the memory of the encoder, and the (binary) content of the registers define its state. Thus, an encoder with k registers has 2^k possible states. By the XOR gates, the current state of the encoder, together with the current input bit, determines both the two current output bits at Out 1 and Out 2 and new state for the next step. This process can be visualized by the *trellis graph* of the code, where each vertex represents a state at a
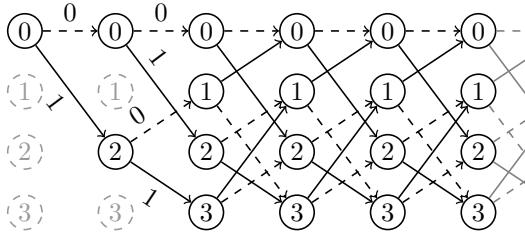
Fig. 7.   Excerpt from a trellis graph with 4 states.
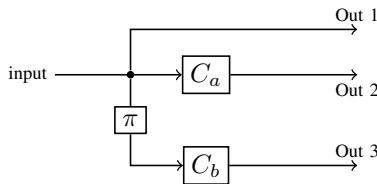

Fig. 8.   Turbo encoder with two convolutional encoders C_a, C_b and interleaver π.

specific step, and there are two types of arcs for zero and one inputs, respectively, that model the transition of one state to another (see Figure 7 for a trellis with 4 states). One usually assumes that initially all registers contain a zero, therefore at the first step there is only one possible state. Arcs representing a zero input bit are drawn in a dashed style, where solid arcs correspond to a one at the input.

In a trellis, a codeword corresponds to a *path* from the initial state to the left to the end state at the right (which is not visible in the figure). It is possible to assign a cost attribute to each arc, based on the LLR values from (3), in a way that the ML codeword can be found by computing the path of minimum cost—a task for which very fast algorithms exist. Therefore, for a simple convolutional code, ML decoding can be accomplished efficiently.

The encoder of a turbo code uses two convolutional encoders. The first one processes the input bits as described above. A copy of the input is sent into the second encoder after passing an *interleaver* π. The interleaver permutes the bits in a pseudo-random way before they enter the second encoder. This way, the input (and thus also the output) is different for both decoders. By this additional complexity, the theoretical error-correcting performance of a turbo code is vastly increased compared to single convolutional encoding, as vital research over the past decades has revealed. The downside of this is that also the *decoding* complexity, the effort to compute the ML codeword for a given received signal, increases. Speaking in terms of trellis graphs, a turbo code can be modelled by two trellises which are synchronized by the constraint that in a given step, the path in the first trellis must use the same type of arc (zero or one) as the corresponding step in the second trellis which is determined by the interleaver function. The ML codeword is now represented by two paths (one for each trellis) of minimum cost that match this additional constraints defined by the interleaver. Unfortunately, it turns out that this problem is much harder than the single, unconstrained min-

cost path in the case of a convolutional code. However, we can formulate the synchronized trellises as an integer linear program, where the interleaver constraints are of the form

$$\sum_{a \in S_{1,1}^i} x_a = \sum_{a \in S_{2,1}^{\pi(i)}} x_a. \tag{5}$$

Here, $S_{1,1}^i$ are the one-arcs in the first trellis for step i, and $S_{2,1}^{\pi(i)}$ are the one-arcs in the second trellis for the step to which the interleaver maps i.

We have run numerical simulations, showing that this formulation leads to greatly reduced computation time compared to the generic one given in (4). This is because only a rather small part of the problem definition, namely the interleaver constraints (5), separate the problem from the easy problem of unconstrained shortest path computation. *Lagrangian relaxation* can be used in such a situation of several "complicating constraints". We have adapted this technique to the case of turbo codes [5]. The idea of Lagrangian relaxation is to remove those constraints, but penalize their violation in the objective function. In the context of turbo coding, this means that we allow the two paths to use different types of arcs in the steps that are interconnected by the interleaver, but each of this violations increases the total cost of the path—the constraint given in (5) leads to the term

$$\eta_i \left(\sum_{a \in S_{1,1}^i} x_a - \sum_{a \in S_{2,1}^{\pi(i)}} x_a \right)$$

in the objective function which becomes larger the more the two sums differ.

A central result of Lagrangian relaxation is that, for an appropriate choice of penalty factors η_i, the optimal paths with respect to the Lagrangian relaxation will converge to the LP relaxation of the problem, where all additional constraints are satisfied. However, in general no integral solution is obtained.

One of our approaches to finding integral solutions is based on the following idea: Given the shortest path on one of the trellises, it is rather unlikely that the corresponding path on the other trellis (determined by the interleaver) is also the shortest path for that trellis. However, the optimal global solution corresponds to the k-th shortest path on one of the trellises, and under certain assumtions k will be relatively small k. We have adapted an efficient algorithm to compute the k shortest paths on a graph to the trellis scenario, and showed that, for low noise and not too large trellises, this can be an efficient approach—especially in conjunction with the modified objective function by the Lagrangian relaxation.

IV. RESULTS, IMPACT, & OUTLOOK

Remember that in practical communication systems the use of an optimal decoder is very often not feasible, because of its large complexity. So when implementing decoders on a chip suboptimal algorithms are used. They are much less complex than the optimal decoder, but also perform worse. The algorithms are often iterative heuristics, that usually run
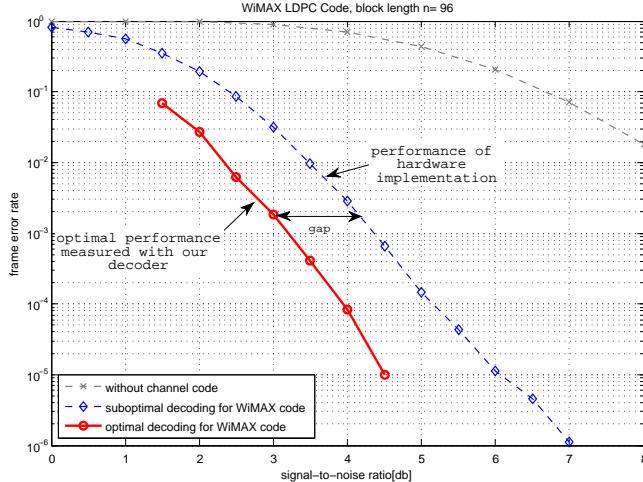
Fig. 9.   Comparison of suboptimal and optimal decoding

a fixed number of iterations or abort earlier according to some stopping criterion. For example the decoding of LDPC codes in practical applications is usually done with a special iterative algorithm called belief-propagation, that is also used in artificial intelligence [6].

Finding decoding algorithms with a good tradeoff between hardware complexity and communications performance is an important research area for hardware designers. To judge an algorithm's performance it can be compared to the optimal decoding. This comparison also helps to estimate if further research on new algorithms is promising or not. However the problem is that for most linear block codes used in practice the optimal decoding performance is unknown.

Our aim is to provide ML simulation results for practical channel codes with larger blocklengths. Furthermore with mathematical techniques we like to get further insight in channel codes and their decoding procedure.

In chapter I we presented the optimal decoding problem as integer optimization problem. With help of the presented techniques we built an optimal decoder, that is able to perform ML decoding efficiently for larger block lengths. With this decoder we measured the optimal performance of some short channel codes used in the recent communication standard WiMAX [7], which were previously unknown.

Figure 9 shows a LDPC code taken from WiMAX with a block lengths of 96 bits. The suboptimal performance as it may be implemented on a chip is shown. Additionally a second curve is shown, which was simulated with the use of our integer optimization decoder and shows the optimal performance.

One can see that there is a gap of about 1 db between the algorithm used in hardware and what is theoretically achievable. This gap is quite large, if you consider that sometimes a few tenths of a db are a good sales argument.

The work on ML decoding as an optimization problem did not only lead to a better understanding of the decoding performance of the suboptimal algorithms and new simulation results. The study of different formulations also led to a new stopping criterion for a iterative decoding algorithm. Stopping criterions abort the iterative process early, if sufficient iterations were made. Here we used a cost function to indicate, when the iterations can be stopped. This leads to a faster computation and a more power efficient decoding algorithm in hardware.

The study of linear block codes gives rise to a lot of new types of mathematical problems and structures that have not been considered before, as well as interesting links to different mathematical fields (like integer programming, graph and network theory, finite algebra, matroid theory). If we use integer programming for ML decoding, we are in the situation that, for a given code, we solve the same program over and over again, where only the cost vector varies. So far, only little attention has been given to this special case, which might allow for more efficient ways to solve the problem by re-using information that was collected in previous problem instances. As another example, by studying certain types of cyclic codes we encountered a new type of network flow problem, where the supply or demand of vertices is variable and thus a part of the problem itself, but restricted to be a multiple of 2. Also, the application of coding theory might arouse interest in constraints of the form $Ax \equiv b \pmod{q}$ in integer programs, which have not yet been studied extensively.

The recently developed 3-dimensional turbo codes, which are an extension of turbo codes that use a third convolutional encoder in conjunction with a second interleaver, show that the development of good Lagrangian relaxation techniques for trellis-like graphs remains an important task for the area of convolutional coding.

REFERENCES

[1] S. Lin and D. J. Costello, *Error Control Coding, Second Edition*.   Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 2004.

[2] J. Feldman, "Decoding error-correcting codes via linear programming," Ph.D. dissertation, Massachusetts Institute of Technology, 2003.

[3] A. Tanatmis, S. Ruzika, H. Hamacher, M. Punekar, F. Kienle, and N. Wehn, "A separation algorithm for improved LP-decoding of linear block codes," *Information Theory, IEEE Transactions on*, vol. 56, no. 7, pp. 3277–3289, 2010.

[4] A. Tanatmis, S. Ruzika, M. Punekar, and F. Kienle, "Numerical Comparison of IP Formulations as ML Decoders," in *Communications (ICC), 2010 IEEE International Conference on*.   IEEE, 2010, pp. 1–5.

[5] A. Tanatmis, S. Ruzika, and F. Kienle, "A lagrangian relaxation based decoding algorithm for LTE turbo codes," in *Turbo Codes and Iterative Information Processing (ISTC), 2010 6th International Symposium on*. IEEE, 2010, pp. 369–373.

[6] T. Richardson and R. Urbanke, "The Renaissance of Gallager's Low-Density Pariy-Check Codes," *IEEE Communications Magazine*, vol. 41, pp. 126–131, Aug. 2003.

[7] *IEEE 802.16*, Local and metropolitan area networks; Part 16: Air Interface for Fixed and Mobile Broadband Wireless Access Systems; Amendment 2:Physical and Medium Access Control Layers for Combined Fixed and Mobile Operation in Licensed Bands.

Advanced Prediction of Pulsed Extraction Column Performance using LLECMOD

Moutasem JARADAT[1,2], Menwer ATTARAKIH[1,3], and Hans-Jörg BART[1,2]

[1] *Chair of Separation Science and Technology, TU Kaiserslautern, POB 3049, 67653 Kaiserslautern, Germany*

[2] *Centre of Mathematical and Computational Modelling, TU Kaiserslautern, Germany*

jaradat@mv.uni-kl.de, bart@mv.uni-kl.de

[3] *Faculty of Eng. Tech., Chem. Eng. Dept., Al-Balqa Applied University, POB 15008, 11134 Amman, Jordan*

attarakih@yahoo.com

Abstract— **A bivariate population balance model (the base of LLECMOD program) for the dynamic simulation of liquid extraction columns is extended to simulate pulsed and sieve extraction columns. The model is programmed using visual digital FORTRAN and then integrated into the *LLECMOD* program. In addition pulsed and sieve tray columns, *LLECMOD* simulates stirred two types of agitated columns (RDC & Kühni). As a case study, *LLECMOD* is used to simulate the steady state performance of a pulsed extraction column. Two chemical test systems recommended by the EFCE are used in the simulation. Model predictions are successfully validated against steady state and dynamic experimental data, where good agreements with the experimental data are achieved.**

I. INTRODUCTION

Liquid-liquid extraction is an important separation processes encountered in many chemical process industries [2]. Different types of liquid-liquid columns are in use nowadays, which can be classified into two main categories: agitated (RDC & Kühni) and pulsed (packed & sieve plate) columns. The latter are frequently used in liquid–liquid extraction operations due to their high throughput, high separation efficiency and insensitivity towards contamination of the interface. These columns have found wide applications in nuclear fuel reprocessing and chemical industry. They have a clear advantage over other mechanical contactors when processing corrosive or radioactive solutions. The absence of moving mechanical parts in such columns obviates the need for frequent repair and servicing. The internals (packing, sieve plates) reduces axial mixing; increases drop coalescence and breakage rates resulting in increased mass transfer rates, and affect the mean residence time of the dispersed phase. The performance of these columns is markedly dependent on the mechanical pulsation of the continuous phase. This is a result of an increase in shear forces and consequent reduction in size of dispersed droplets so that the interfacial area, and hence the mass transfer rate, is increased [3].

To shed more light on the extraction behaviour in pulsed (packed & sieve plate) columns, the hydrodynamics and mass transfer characteristics must be well understood. Our present knowledge of the design and performance of extraction columns is still far from satisfactory. The reason is mainly due to the complex interactions of the hydrodynamics and mass transfer [4]. It is obvious that the changes in the characteristics (holdup, Sauter diameter, etc.) of the drop population along the column have to be considered in order to describe conveniently the behaviour of the column. The dispersed phase in the case of liquid-liquid extraction undergoes changes and loses its identity continuously as the drops break and coalesce. Accordingly, detailed modelling on a discrete level is needed using the population balance equation as a mathematical framework. The multivariate non-equilibrium population balance models have emerged as an effective tool for the study of the complex coupled hydrodynamics and mass transfer in liquid-liquid extraction columns.

The simulation of modern industrial chemical processes is becoming extremely important as an economical tool in the integration of steady state and dynamic design as well as the simulation of the existing plants. The development of computational tools to model industrial processes has increased in the last decades. However; to the best of the authors' knowledge, there are no comprehensive non-equilibrium population balance models to describe in sufficient detail the behaviour of extraction columns. One of the recent approaches in modelling liquid-liquid extraction columns is by adequately describing the complex behaviour of the dispersed phase using the population balances equations (PBE). However, even the numerical solution of the resulting system of PBEs is still not efficiently developed particularly when coupled hydrodynamics and mass transfer take place simultaneously.

The main objective of this work is to develop a model that is capable of describing the dynamic and steady state behaviour of pulsed (packed & sieve plate) extraction columns. The models of both columns are integrated into the existing program: *LLECMOD* [1], which can also simulate agitated extraction columns (RDC & Kühni). *LLECMOD* can simulate the steady state and dynamic behaviour of extraction columns taking into account the effect of dispersed phase inlet (light or heavy phase is dispersed) and the direction of mass transfer (from continuous to dispersed phase and vice versa). So, scale-up and simulation of agitated and pulsed extraction columns based on population balance modelling can now be carried out successfully [1, 5].

II. MATHEMATICAL MODELLING

The modelling of extraction columns still demands improvement. Simulating liquid-liquid extraction columns is a

challenging task due to the discrete character of the dispersed phase. This results from random breakage and coalescence of droplets, which are highly coupled to the mass transport of solutes between the two existing phases. Modelling of such extremely important and complex transport phenomena is resolved by using a multivariate population balance equation. The population of droplets is modelled by a multivariate number concentration function, which takes into account droplet size and solute concentrations. Understanding of dynamic behaviour of liquid-liquid extraction columns can be notably used in the design of process control strategy or the start-up and shutdown procedures [6-9]. The attempts to model this dynamic behaviour are not well established because of the discrete nature of the dispersed phase. This is due to the complex nature of the macroscopic dispersed phase interactions as well as the microscopic interphase mass transfer occurring in the continuously turbulent flow field. These macroscopic interactions such as droplet breakage and coalescence coupled to the interphase mass transfer result in a distributed population of droplets. This population is distributed not only in the spatial domain of the contacting equipment, but also randomly distributed with respect to the droplet state (properties) such as size, concentration and age. Several attempts have been done to propose the proper mathematical model for liquid-liquid extraction columns [10]. An empirical model for predicting the hydrodynamics in a pulsed sieve plate column was proposed by Kumar and Hartland [11]. A stagewise model for the transient behaviour of a sieve-plate extraction column taking into account the back flow and assuming constant hold-up, was developed by Blass and Zimmerman [10]. Reference [6] evaluated a differential model in a Kühni column. Hufnagl et al. [12] modelled a packed column using differential contact model without axial mixing. Flow models such as the dispersion or backmixing model; describe the non-ideal flow, where one parameter accounts for all deviations from the ideal plug flow behaviour [13]. These models are too simple to describe the real hydrodynamics, where one of the liquid phases is normally dispersed as droplets in the second continuous phase [14]. Therefore, the influences of droplet movement, droplet interaction (breakage and coalescence), energy input (agitation, pulsation) and mass transfer cannot be described satisfactorily. Weinstein et al. [7] evaluated the differential model for a Kühni column. An improved dynamic combined model considering the influence of drop size distribution was developed by Xiaojin et al. [9]. Several population balance models have been proposed by various authors. Garg and Pratt [15] developed a population balance model for a pulsed sieve-plate extraction taking into account experimentally determined values for drop breakage and coalescence. Casamatta and Vogelpohl [16] proposed a population balance model for which a good review is found in [17]. Al Khani et al. [18] have applied this model for dynamic and steady-state simulations of a pulsed sieve-plate extraction column. Recently much works have been done in the population balance modelling for extraction columns [1, 17, 19-23].

A. multivariate non-equilibrium population balance model

The general spatially distributed population balance (SDPBE) for describing the coupled hydrodynamics and mass transfer in liquid extraction columns in a one spatial domain can be written as:

$$\frac{\partial f_{d,c_y}(\psi)}{\partial t} + \frac{\partial[u_y f_{d,c_y}(\psi)]}{\partial z} + \sum_{i=1}^{2} \frac{\partial[\dot{\zeta}_i f_{d,c_y}(\psi)]}{\partial \zeta_i} =$$

$$\frac{\partial}{\partial z}\left[D_y \frac{\partial f_{d,c_y}(\psi)}{\partial z}\right] + \frac{Q_y^{in}}{A_c}\frac{f_y^{in}}{\overline{v}_{in}}(d,c_y;t)\delta(z-z_y) + \Upsilon\{\psi\}$$

.. (1)

In this equation the components of the vector $\psi = [d\ c_y z\ t]$ are those for the droplet internal coordinates (diameter and solute concentration), the external coordinate (column height), z, and the time, t, where the velocity along the concentration coordinate (c_y) is $\dot{c}_y$. The source term $\Upsilon.\partial\zeta$ represents the net number of droplets produced by breakage and coalescence per unit volume and unit time in the coordinates range $[\zeta, \zeta+\partial\zeta]$. The left hand side is the continuity operator in both the external and internal coordinates, while the first part on the right hand side is the droplets axial dispersion characterized by the dispersion coefficient, D_y, which might be dependent on the energy dissipation and the droplet rising velocity [24]. The second term on the right hand side is the rate at which the droplets entering the LLEC with volumetric flow rate, $Q_{y,in}$, that is perpendicular to the column cross-sectional area, A_c, at a location z_y with an inlet number density, f_y^{in}, and is treated as a point source in space. The dispersed phase velocity, u_y, relative to the walls of the column is determined in terms of the relative (slip) velocity with respect to the continuous phase and the continuous phase velocity, u_x, with respect to the walls of the column [22, 23].

The solute concentration in the continuous phase, c_x, is predicted using a component solute balance on the continuous phase [22, 23]:

$$\frac{\partial(\phi_x c_x)}{\partial t} - \frac{\partial}{\partial z}\left(u_x \phi_x c_x + D_x \frac{\partial(\phi_x c_x)}{\partial z}\right) =$$

$$\frac{Q_x^{in} c_x^{in}}{A_c}\delta(z-z_y) - \int_0^\infty \int_0^{c_{y,max}} \dot{c}_y v(d) f_{d,c_y}(\psi)\partial d\partial c_y$$

(2)

Note that the volume fraction of the continuous phase, Φ_x, satisfies the physical constraint: $\Phi_x + \Phi_y = 1$, where y denotes the droplet phase. The left hand side of Eq.(2) as well as the first term on the right hand side have the same interpretations as those given in Eq.(1); however, with respect the continuous phase. The last term appearing in Eq.(2) is the total rate of solute transferred from the continuous to the dispersed phase, where the liquid droplets are treated as point sources. Note that Eq.(1) is coupled to the solute balance in the continuous phase given by Eq.(2) through the convective and the source terms.

B. Model parameters

The SDPBE is general for any type of extraction column. However, what makes the equation specific is the internal geometry of the column as reflected by the required correlations for hydrodynamics and mass transfer. Experimental correlations are used for the estimation of the turbulent energy dissipation and the slip velocities of the moving droplets along with interaction frequencies of breakage and coalescence [1, 22, 23].

C. Breakage probability and daughter droplet distribution

For pulsed packed extraction column the daughter droplet distribution is assumed to follow the beta distribution β, which is given by Bahmanyar and Slater [25]. In Eq.(3) below v is the average number of daughter drops per breakage event.

$$\nu = 2 + 0.34\left(\left(d'/d_{stab}\right) - 1\right)^{1.96} \quad (3)$$

$$\beta_n(d \mid d') = 3\,\nu\,(\nu - 1)\left(1 - \left(d/d'\right)^3\right)^{\nu-2}\left(d^5/d'^6\right) \quad (4)$$

Here, d_{stab} is the stable droplet diameter, where droplets having diameter less than d_{stab} are not expected to break. The data for the stable drop diameters for pulsed packed and sieve trays columns under different operating conditions for the two standard EFCE test systems (water-acetone-toluene and water-acetone-butyl acetate) are given in [26]. The droplet breakage frequency and the daughter droplet distribution are correlated based on single droplet experiments. The droplet breakage frequency used in the simulation is given by [26]:

$$P_B\left(d\right) = C_1 \pi_{af}^{C_2} \frac{\left[\left(d - d_{stab}\right)/\left(d_{100} - d_{stab}\right)\right]^{C_3}}{C_4 + \left[\left(d - d_{stab}\right)/\left(d_{100} - d_{stab}\right)\right]^{C_3}} \quad (5)$$

Where π_{af} is a dimensionless number taking into account the influence of the pulsation intensity on the breakage probability and is given by [26]:

$$\pi_{af} = a.f.\left(\rho_c^2 / \mu_c.\Delta\rho.g\right)^{1/3} \quad (6)$$

This breakage frequency describes the breakage in a packed and sieve tray compartment with only one set of constant parameters for a liquid/liquid-system. The constants C_i appearing in Eq.(5) for pulsed (packed & sieve plate) columns are listed in [26]. To predict the breakage probability in pulsed columns with Eq.(5), the characteristic drop diameters d_{100} & d_{stab} have to be experimentally determined for each pulsation intensity. Data of the characteristic drop diameter for pulsed columns investigated here are listed in [26].

D. Droplet coalescence probability

In this work droplet coalescence probability for a pulsed extraction column is given by [27]:

$$P_C\left(d\right) = \frac{\varepsilon\,H_{cd}^{1/6}\,\left(\Delta\rho g\right)^{1/2}\,\Delta t\,\sigma^{1/3}}{\xi_8\,\mu_c\,d^{1/3}} \quad (7)$$

The parameter ξ_8 was fitted to experimental data for two standard EFCE test systems, for t-a-w ξ_8 is 2500, and for b-a-w ξ_8 is 1500 [27], Hamaker constant H_{cd} values used in the simulation are listed in table I & II.

E. The terminal droplet velocity

The terminal droplet velocity used in the simulation is given by [27]:

$$v_t = \frac{v_{os,de}v_{sph}}{\left(v_{os,de}^{a_{15}} + v_{sph}^{a_{16}}\right)^{1/a_{16}}} \quad (8)$$

In this velocity model, a_i the adjustable parameters can be fitted to the experimental data. v_{sph} is the spherical droplet velocity, $v_{os,de}$ is a smooth transition from oscillating to deformed droplets [27].

For the description of the drop motion in the structured packing, the correlation for the slowing factor developed by Garthe [26] is applied:

$$k_v = 0.077\pi_{H_{Pk}}^{0.138}\pi_{a_{Pk}}^{-0.566}\pi_d^{-0.769}\pi_\sigma^{0.184}\left(1 + \pi_{af}\right)^{0.08} \quad (9)$$

where π_{Hpk}, π_{apk}, π_d, π_σ, π_{af} are dimensionless numbers [26].

F. Mass –transfer

The mass transfer fluxes in the *LLECMOD* program are calculated based on the two-film theory, where the individual mass transfer coefficients are defined separately for the continuous (k_x) and the dispersed (k_y) phases. The mass transfer model used here in this simulation is taken from the work of Henschke [27].

G. Numerical solution

The mathematical model, which consists of integro-partial differential and algebraic equations is solved using an optimized and efficient numerical algorithms developed in [20-22]. These are utilized successfully for the simulation of coupled hydrodynamics and mass transfer for general liquid-liquid extraction columns.

III. LLECMOD PROGRAM

The complete mathematical model described above is programmed using Visual Digital FORTRAN. To facilitate the data input and output, a graphical user interface is designed. The graphical interface of the *LLECMOD* program contains the main input window and sub-windows for parameters and correlations input. The basic feature of this program [1] is to provide an easy tool for the simulation of coupled hydrodynamics and mass transfer in liquid-liquid extraction columns based on the population balance approach for both transient and steady states conditions through an interactive windows input dialog. Note that *LLECMOD* is not restricted to a certain type of liquid-liquid extraction column since it is built in the most general form that allows the user to input the various droplet interaction functions. These functions include droplet terminal velocity (taking into account the swarm effect) and the slowing factor due to column geometry, the breakage frequency and daughter droplet distribution, the coalescence frequency and the axial dispersion coefficients. Using *LLECMOD* simulations can now be carried out successfully for different types of extraction columns including agitated (RDC & Kühni) and pulsed (sieve plate & packed) columns. The design of *LLECMOD* is flexible in such a way that allows the user to

define droplet terminal velocity, energy dissipation, axial dispersion, breakage and coalescence frequencies and the other internal geometrical details of the column. The correlation parameters that are obtained based on single droplet and droplet swarm experiments, are considered in a modularized structure for the simulation program.

A. Coalescence parameter optimization package

As a new feature of *LLECMOD*, the program is reinforced by a parameter estimation package for the droplet coalescence models, which enables the user to fit the column hydrodynamics (droplet size distribution, holdup and mean droplet diameter) to the available experimental data. Prediction of the mass transfer profiles is then performed based on correlations obtained only from single droplet experiments. Simulation results are compared with data from pilot plant columns, such as agitated and pulsed/ un-pulsed columns. Different case studies showing columns' performance are presented using only optimized coalescence parameters. Additional results will be published in separate publications.

IV. RESULTS AND DISCUSSION

In this section a sample problem is considered to illustrate the basic features of the *LLECMOD* and the coalescence parameters estimation package. For this purpose, a pilot plant laboratory scale pulsed (packed & sieve plate) columns are considered whose dimensions are: column height $H = 4.4 \ m$, inlet of the dispersed phase $z_y = 0.85 \ m$, inlet of the continuous phase $z_x = 3.8 \ m$, column diameter $d = 0.08 \ m$. The two EFCE test systems toluene-acetone-water (t-a-w) and butyl acetate-acetone-water (b-a-w) are used whose physical properties are available online (http//dechema.de/extraktion). To completely specify the model, the inlet feed is fitted to a normal distribution with mean equals to *3.2 mm* and standard deviation of *0.5 mm*. The direction of mass transfer is from the continuous to the dispersed phase. All the operating conditions for the packed column are listed Table (I), and for sieve plate column in Table (II).

TABLE I
OPERATING CONDITIONS: PACKED COLUMN

Test System	Qc lit/h	Qd lit/h	$C_{y,in}$ %	$C_{x,in}$ %	af cm/sec	H_{cd} Nm
t-a-w	61.3	74.5	0.6	5.89	2	10e-10
b-a-w	62.0	72.0	0.0	5.21	2	10e-10

TABLE III
OPERATING CONDITIONS: SIEVE PLATE COLUMN

Test system	Qc lit/h	Qd lit/h	$C_{y,in}$ %	$C_{x,in}$ %	af cm/sec	H_{cd} Nm
b-a-w	62.0	72.0	0.0	5.46	1	10e-10
t-a-w	61.3	74.5	0.15	5.52	1	10e-12

Fig.(1) shows the variation of the mean droplet diameter along the height pulsed packed column and compared to the experimental data for both chemical systems. The agreement

between the simulation and experiment is excellent for both test systems.
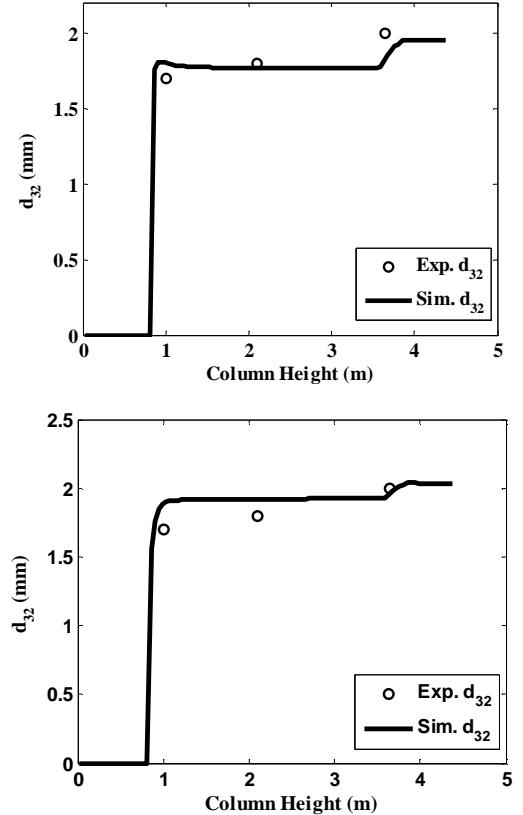
Fig. 1 Simulated mean droplet diameter along the column height compared to the experimental data [26]. Upper panel the test system is (b–a-w) and the lower panel is (t-a-w) in packed column.

Fig.(2) depicts the variation of the mean droplet diameter along the height of the pulsed sieve plate column and compared to the experimental data for both chemical systems. A fairly good agreement between the experimental and simulated profiles is achieved for both systems.
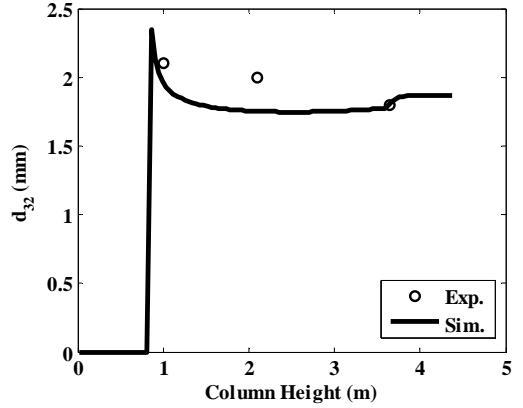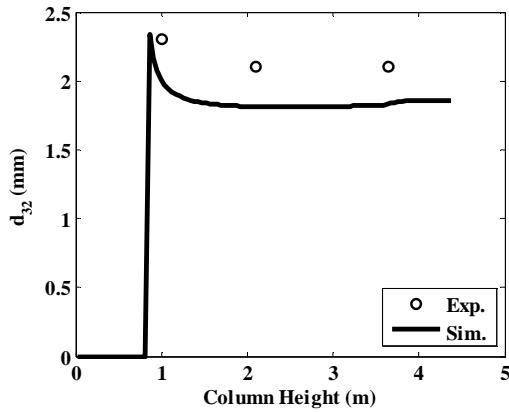
Fig. 2 Simulated mean droplet diameter along the column height compared to the experimental data [26]. Upper panel the test system is (b–a-w) and the lower panel is (t-a-w) in sieve plate column.

A comparison between the simulated holdup profiles along the height of the pulsed packed column and the experimental data [26] is shown in Fig.(3). Again, an excellent agreement is achieved for both test systems.
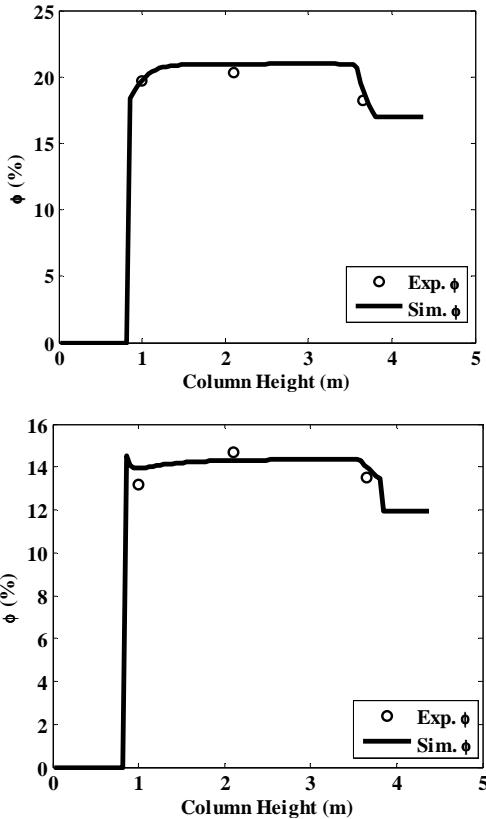

Fig. 3 Simulated holdup profiles along the column height compared to the experimental data [26]. Upper panel the test system is (t-a-w) and the lower panel is (t-a-w) in packed column.

Fig.(4) depicts the variation of the holdup profiles along the height of pulsed sieve plate column compared to the experimental data for both chemical systems. A fairly good agreement between the experimental and simulated profiles is achieved for both systems.
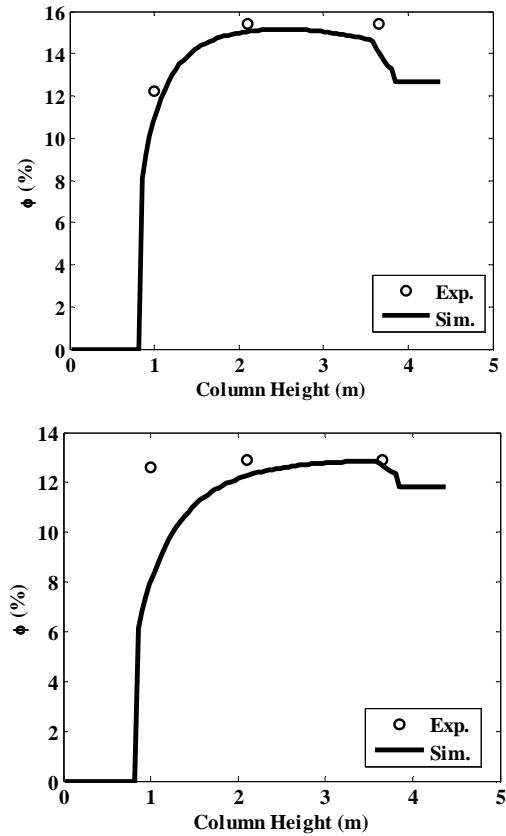
Fig. 4 Simulated holdup profiles along the column height compared to the experimental data [26]. Upper panel the test system is (b–a-w) and the lower panel is (t-a-w) in sieve plate column.

Fig.(5) shows the simulated and experimental solute concentration profiles as function of column height in both phases. The agreement between the simulation and experiment is very good for both test systems.

As a conclusion, the very good agreement between the simulated profiles and the experimental data appearing in Figs.(1-5) for different columns and chemical systems shows that LLECMOD capability of predicting the actual steady state performance of pulsed (packed and sieve plate) extraction columns.

25
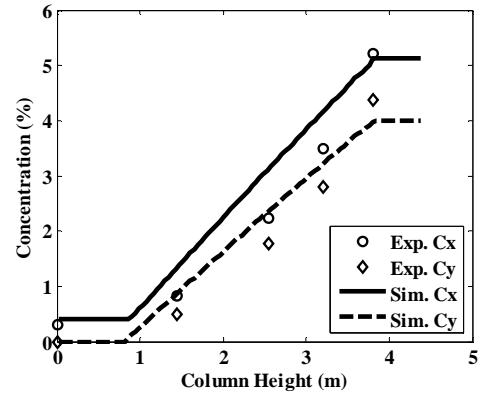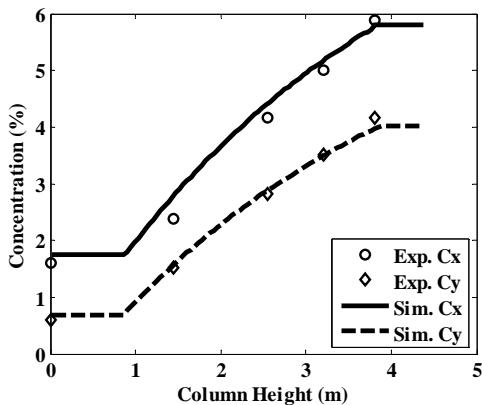
Fig. 5 Simulated solute concentration profiles in both phases along the column height compared to the experimental data [26]. Upper panel the test system is (b -a-w) and the lower panel is (t-a-w) in packed column.

Moreover, the dynamic behaviour of liquid extraction columns can be predicted using LLECMOD. This is necessary to improve the knowledge of the dynamic behaviour of extraction columns under the effect of different disturbances. Knowing how a system responds to disturbances is a prerequisite for controller design and optimization of column start up and shut down procedures. Further detailed research is underway to validate *LLECMOD* against available dynamic experimental data.

V. CONCLUSIONS

The LLECMOD non-equilibrium bivariate population balance model is found capable of simulating new types of extraction columns; namely, pulsed and un-pulsed packed and sieve tray columns in addition to the agitated columns (Kühni & RDC types). The rapid steady state hydrodynamics column behaviour can be efficiently predicted by only adjusting few parameters in the droplet coalescence model. These parameters can then be used to predict the very slow mass transfer process independently. In doing this, the model is validated successfully against the available experimental data.

REFERENCES

[1] Attarakih M.M., Bart H.-J., Steinmetz T., Dietzena M. and Faqir N.M., 2008, LLECMOD: A Bivariate Population Balance Simulation Tool for Liquid- Liquid Extraction Columns. Open Chem. Eng. J., 2, 10-34.

[2] Lo T.C., Baird M.H.I. and Hanson C., Eds., 1983, Handbook of Solvent Extraction, J. Wiley & Sons, New York.

[3] Pratt H.R.C. and Stevens G.W., 1992, Selection, Design, Pilot-Testing and Scale-up of Extraction Equipment," In: Thornton J.D., Ed., "Science and Practice in Liquid-Liquid Extraction," Oxford University Press, New York, 491–589.

[4] Luo G., Li H., Fei W. and Wang J., 1998, A simplified correlation of mass transfer in a pulsed sieve plate extraction column, Chem. Eng. Technol., 21, 10, 823–827.

[5] Jaradat M., Attarakih M. and Bart H.-J., 2010, Effect of phase dispersion and mass transfer direction on steady state RDC performance using population balance modelling , Chem. Eng. J., 165, 2, 379-387.

[6] Hufnagl H., McIntyre M. and Blaß E., 1991, Dynamic behaviour and simulation of a liquid–liquid extraction column, Chem. Eng. Technol., 14, 301–306.

[7] Weinstein O., Semiat R. and Lewin D., 1998, Modelling, simulation and control of liquid–liquid extraction columns. Chem. Eng. Sci., 53, 2, 325–339.

[8] Mjalli F.S., 2005, Neural network model-based predictive control of liquid–liquid extraction contactors, Chem. Eng. Sci., 60, 239–253.

[9] Xiaojin T., Guangsheng L. and Jiading W., 2005, An improved dynamic combined model for evaluating the mass transfer performances in extraction columns, Chem. Eng. Sci., 60, 4409–4421.

[10] Blass E. and Zimmerman H., 1982, Mathematische Simulation und experimentelle Bestimmung des instationären Verhaltens einer flüssigkeitspulsierten Siebbodenkolonne zur Flüssig-Flüssig-Extraktion, Verfahrenstechnik, 16, 9, 652-690.

[11] Kumar A., Hartland S., 1995, A unified correlation for the prediction of dispersed-phase hold-up in liquid-liquid extraction columns, Ind. Eng. Chem. Res., 34, 3925–3932.

[12] Steiner L., Bertschmann H. and Hartland S., 1995, A Model for Simulation of Hydrodynamics and Mass-Transfer in Extraction Columns Filled with a Regular Packing, Chem. Eng. Res. Des., 73, 5, 542-550.

[13] Thornton, D., Science and Practice of Liquid-Liquid Extraction: Volume 2: Process Chemistry and Extraction Operations in the Hydrometallurgical, Nuclear, Pharmaceutical,... (Oxford Engineering Science Series), Oxford: Clarendon Press (1992).

[14] Bart H.-J. and Stevens G., 2004, Reactive Solvent Extraction, In: Ion Exchange and Solvent Extraction, Eds. Kertes M., Sengupta A.K., 17, 37-82, Marcel Dekker, New York.

[15] Garg M.O. and Pratt H.R.C., 1984, Measurement and modelling of droplet coalescence and breakage in a pulsed-plate extraction column, AIChE J., 30, 3, 432–441.

[16] Casamatta G. and Vogelpohl A., 1985, Modelling of fluid dynamics and mass transfer in extraction columns, Ger. Chem. Eng., 8, 96-103.

[17] Gourdon C., Casamatta G. and Muratet G., 1994, in: J.C. Godfrey and M.J. Slater (Eds.), Liquid-Liquid Extraction Equipment, Wiley, Chichester, 137-226.

[18] Al Khani S.D., Gourdon C. and Casamatta G., 1989, Dynamic and steady state simulation of hydrodynamics and mass transfer in liquid-liquid extraction column, Chem. Eng. Sci., 44, 6, 1295-1305.

[19] Kronberger T., Ortner A., Zulehner W. and Bart H.-J., 1995, Numerical Simulation of Extraction Columns Using a Drop Population Model, Comput. Chem. Eng., 19, 639-644.

[20] Attarakih M.M., Bart H.-J., and Faqir N.M., 2004a, Numerical solution of the spatially distributed population balance equation describing the hydrodynamics of interacting liquid-liquid dispersions, Chem. Eng. Sci., 59, 2567-2592.

[21] Attarakih M.M., Bart H.-J., and Faqir N.M., 2004b, Solution of the droplet breakage equation for interacting liquid-liquid dispersions: a conservative discretization approach, Chem. Eng. Sci., 59, 2547-2565.

[22] Attarakih M.M., Bart H.-J., and Faqir N.M., 2006a, Numerical Solution of the Bivariate Population Balanced Equation for the Interacting Hydrodynamics and Mass Transfer in Liquid-Liquid Extraction Columns, Chem. Eng. Sci., 61, 113-123.

[23] Attarakih M.M., Bart H.-J., and Faqir N.M., 2006b,.A Hybrid Scheme for the Solution of the Bivariate Spatially Distributed Population Balanced Equation, Chem. Eng. Tech., 29, 435-441.

[24] Modes, G., Bart, H.-J., Rodrigues-Perancho, D., Bröder, D., 1999, Simulation of the fluid dynamics of solvent extraction columns from single droplet parameters, Chem. Eng. Tech., 22 (3), 231-236.

[25] Bahmanyar, H. and M.J. Slater, 1991. Studies of Drop Break up in Liquid-liquid Systems in a Rotary Disc Contactor, Part I: Conditions of No Mass Transfer, Chemical Engineering and Technology, 14: 79-89.

[26] Garthe G., 2006, Fluid Dynamics and Mass Transfer of Single Particles and Swarms of Particles in Extraction Column, Dissertation, TU München, Germany.

[27] Henschke M., 2004, Auslegung pulsierter Siebboden-Extraktionskolonnen, Shaker Verlag Aachen.

CFD Simulation und verbesserte Datenauswertung einer Extraktionskolonne vom Typ Kühni

Mark W. Hlawitschka [#1], Fang Chen [*2], Hans-Jörg Bart [#3], Hans Hagen [*4]

[#]*Chair of Seperation Science and Technology, Gebäude 44, Raum 476, Gottlieb-Daimler Straße, 67663 Kaiserslautern*
[*]*AG Computergrafik, Gebäude 36, Raum 22, Gottlieb-Daimler Straße, 67663 Kaiserslautern*
[1]mark.hlawitschka@mv.uni-kl.de
[2]chen@informatik.uni-kl.de
[3]bart@mv.uni-kl.de
[4]hagen@informatik.uni-kl.de

Zusammenfassung—**Computational Fluid Dynamic (CFD) Simulationen wurden gekoppelt mit dem One-Group Populationsbilanzenmodell durchgeführt, um die Tropfengröße in einer Kühni Miniplant Extraktionskolonne zu bestimmen. Zur Strömungssimulation wurde aufgrund der geringeren Rechenzeit auf das Euler-Euler Modell zurückgegriffen, welches die Tropfen als interpenetrierende Phase behandelt. Die Tropfeninteraktion, wie Tropfenkoaleszenz und Tropfenteilung werden durch das Modell von Prince und Blanch [1] und von Martínez-Bazán et al. [2] beschrieben. Eine verbesserte Visualisierungstechnik wird angewendet, um mehrere Daten in einem Bild kombiniert zu visualisieren. Dies bildet die Basis für Visualisierungserweiterungen, wie z.B. der Visualisierung der Tropfenteilung.**

Abb. 1.   Kooperation zwischen den Instituten

I. Kooperation

Die hier vorgestellten Ergebnisse stammen aus der Kooperation *Multiphysics, Modelling and feature based Visualization* mit Schwerpunkt Tropfenpopulationsdynamik. Die Kooperation besteht aus Forschern der AG Technomathematik sowie des ITWM, des Lehrstuhls für Thermische Verfahrenstechnik sowie der AG Computergrafik und HCI Gruppe der TU Kaiserslautern. Jede Gruppe, wie in Abb. 1 gezeigt, trägt zur Kooperation bei. Der Bereich der Mathematik ist auf die Modellentwicklung (Finite Pointset Method, FPM) sowie Modellbeurteilung spezialisiert. Die Verfahrenstechnik beschäftigt sich mit der Modellierung von Apparaten, u.a. gerührten Extraktionskolonnen, mit kommerziellen und Open-Source Software. Ebenso liefert diese experimentelle Daten zur Validierung. Die Ergebnisse der Simulationen dienen zur verbesserten Auslegung der Kolonnen und zur gegenseitigen Validierung der mathematischen Modelle sowie deren Parameter. Die erzeugten Simulationsdaten sind hauptsächlich Ensight formatierte Binärdaten. Die AG Computergrafik und HCI beschäftigt sich daher mit der Entwicklung verbesserter bzw. angepasster Visualisierungstools, die es ermöglichen, u.a. mehrere Parameter gleichzeitig sichtbar zu machen und den Auswerteprozess somit beträchtlich vereinfachen.

II. Motivation

Die Auslegung und Modellierung von flüssig-flüssig Extraktionskolonnen basiert noch immer auf vereinfachten Modellen, die nur teilweise die realen Geschehnisse wie Tropfenkoaleszenz und Tropfenteilung in der Kolonne abbilden. Das einfachste verwendete Modell ist das Stufenmodell, welches auf einer idealisierten Propfenströmung basiert. Dieses Modell berücksichtigt eine Änderung des Zustands in Strömungsrichtung, jedoch keine radiale Änderung senkrecht zu dieser. Bei Abweichungen von der idealen Propfenströmung wird daher auf das Backmixing oder Dispersionsmodell zurückgegriffen, welche die Abweichungen in radialer Richtung über einen einzelnen Parameter berücksichtigen. Kombiniert mit Tropfenpopulationsbilanzmodellen (TPBM) wird Koaleszenz und Teilung der Tropfen mit in Betracht gezogen und es ergibt sich eine gute Modellierungsbasis der Kolonne. Allerdings muss man hierbei auf Korrelationen zurückgreifen, die über große Datenpools zeitaufwändig verifiziert wurden. In den letzten Jahrzenten ergab sich durch die steigende Rechenleistung eine detailliertere, geometrieunabhängige Auflösung des Strömungsproblems, meist auf Basis der Navier-Stokes-Gleichungen oder Euler-Gleichungen, die heute als Computational Fluid Dynamics (CFD) bezeichnet wird. So wurden auch in den letzten Jahren verstärkt in der Verfahrenstechnik fluiddynamische Simulationen von Extraktionskolonnen vom Typ Rotating Disc Contactor (RDC) sowie vom Typ Kühni, dargestellt in Abb. 2, durchgeführt. Die Simulation der RDC Kolonne kann dabei aufgrund der Rotationssymmetrie zweidimensional dargestellt werden, während eine Simulation der Kühni Kolonne dreidimensional bzw. dreidimensional rotationssymmetrisch erfolgen muss.

Einphasige Simulationen einer RDC Kolonne stammen u.a.
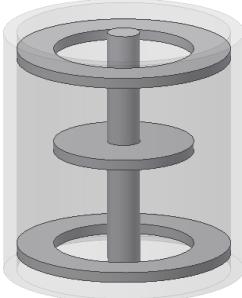
(a) RDC Kolonne
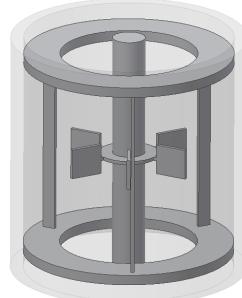
(b) Kühni Miniplant Kolonne

Abb. 2.   Unterschiedliche Extraktionskolonnengeometrien

(a) Gitter eines Kompartments

(b) Rührerzone mit Rührerblättern

Abb. 3.   Erstelltes Gitter

von Modes und Bart [3], wobei Rieger [4] die Zweiphasenströmung in einer RDC Kolonne simulierte. Haderer et al. [5] verwendete ein Euler-Euler Modell zur Simulation der Zweiphasenströmung in RDC Kolonnen. Kolb [6] untersuchte die Strömung in Kühni Kolonnen und validierte seine Ergebnisse mit einer Lasermesstechnik, der Particle Image Velocimetry. Wang und Mao [7] simulierten die Zweiphasenströmung in einem Kühni Miniplant Kompartment unter Verwendung periodischer Randbedingungen in axialer Richtung. Vikhansky und Kraft [8] verwendeten Euler-Lagrange Modell zur Simulierung der Zweiphasenströmung in einer RDC Kolonne gekoppelt mit der Monte-Carlo Methode zur Berücksichtigung von Koaleszenz und Zerfall. Drumm et al. [9] untersuchte die Zweiphasenströmung auf Basis des Euler-Euler Modells und koppelte dieses mit stochastischen Populationsmodellen. CFD Untersuchungen an Extraktionskolonnen vom Typ Kühni gekoppelt mit Populationsbilanzen sind in der Literatur neu und erfordern eine Anpassung der bestehenden Zerfalls- sowie Koaleszenzmodelle an experimentelle Ergebnisse. Eine effektive Datenauswertung ist dabei für den Anwender wichtig, wobei zur schnellen und benutzerfreundlichen Datenauswertung existieren jedoch keine entsprechende Tools. Kommerzielle Post-Processing Programme können jeweils eine fluide Eigenschaft darstellen, z.B. Phasenanteil, Geschwindigkeit oder Tropfengröße. Um aussagekräftige Informationen zu erhalten, benötigt es einer Überlagerung der Informationen z.B. der Tropfengröße und des Phasenanteils. Ebenso können Tropfen, die aufgrund des zu benötigenden hohen Rechenaufwands nicht aufgelöst dargestellt werden, sondern nur als Datenpunkt in jeder einzelnen numerischen Zelle, nicht wirklichkeitsgetreu visualisiert werden. Zudem kann der zeitliche Verlauf der Tropfen nur durch Stromlinien dargestellt werden, wobei diese aufgrund der angenommen Randbedingung an allen Wänden und Einbauten enden und eine Darstellung nahezu unmöglich machen. In dieser Arbeit werden die vorangegangenen Arbeiten zum Thema Populationsbilanzen von den 2D-Modellen (RDC-Kolonne) auf einen 3D-Anwendungsfall (Kühni-Kolonne) transferiert. Ein besonderes Augenmerk wird hierbei auf eine verbesserte Auswertemöglichkeit gelegt.
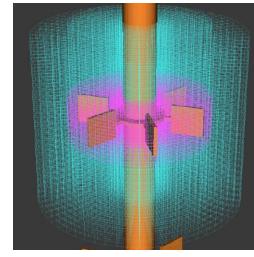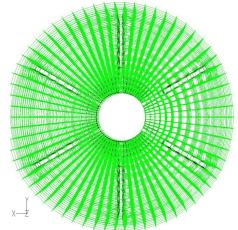
III. METHODE

Die hier beschriebenen Methoden gliedern sich in drei Teile. Der erste Teil beschäftigt sich mit der numerischen Simulation. Der zweite Teil befasst sich mit den Populationsbilanzen und deren Implementierung. Der dritte Teil zeigt die Vorgehensweise bei der verbesserten Datenauswertung und Visualisierung der Simulationsergebnisse auf.

A. Methode der Simulation

Der Kühni Miniplant Extraktor hat einen Nenndurchmesser von 32 mm und einer Kompartmenthöhe von 28 mm und Strombrecher in Form von flachen Blechen sind in einem Winkel von 120° zueinander angeordnet. Ein Sechsblattrührer ist in der Mitte jedes Kompartments mit der Welle verbunden. In dieser Arbeit wurden sieben Kompartments für die Simulation als Gitter mit dem Pre-Prozessor Gambit generiert. Für die CFD-Simulation mit dem kommerziellen CFD Tool FLUENT wurde das Gitter in zwei Regionen geteilt, eine statische und eine drehende Region, die den Rührer umgibt. Am Boden und am Kopf des Gitters wurde zusätzlich eine Zone für den Einlauf und den Auslauf definiert. Die Randbedingung am Boden wurde als Geschwindigkeitseinlass definiert, wobei der Auslass durch einen Austrittsdruck definiert wird. Als numerischer Rahmen wurde das Euler-Euler Modell verwendet, welches die beiden Phasen als interpenetrierende Kontinua behandelt. Ein Populationsbilanzmodell wurde mit der CFD über benutzerdefinierte Funktionen ("user defined functions", UDF) gekoppelt, um Koaleszenz- sowie Zerfallsereignisse zu berücksichtigen. Das verwendete flüssig-flüssig-System besteht aus Wasser und n-Butylacetat, wobei die organische Phase in Form von Tropfen am Boden und die kontinuierliche Phase (Wasser) am Kopf in die Kolonne fließt, wodurch eine Gegenströmung der beiden Phasen entsteht. Das für die Simulation erstellte Gitter besteht aus über 500 000 Zellen, welches als Ausschnitt in Abbildung 3 dargestellt ist.

Die Kolonnenwand sowie die Rührer, Welle und Strombrecher sind als *no-slip* Randbedingung definiert, d.h. dass die Geschwindigkeit der Fluide an der Wand gleich Null ist. Der *first-order implicit solver* wurde für die Zeitdiskretisierung verwendet, während für die Diskretisierung im Raum das *first order upwind scheme* verwendet wurde. Die Druck-Geschwindigkeitskopplung erfolgt über die Verwendung des Simple Algorithmus. Standard Relaxationsfaktoren wurden für

die Simulation verwendet. Im Folgenden werden die mathematischen Modelle des Zweiphasenmodells zusammengefasst.

1) Zweiphasenmodell: Das verwendete Zweiphasenmodell ist das Euler-Euler Modell, bei dem die jeweilige Phase i durch den Volumenanteil α_i beschrieben wird. Die Kontinuitätsgleichung für die kontinuierliche Phase c ist durch

$$\frac{\partial (\alpha_l \rho_c)}{\partial t} + \nabla \cdot (\alpha_c \rho_c \vec{u_c}) = 0 \tag{1}$$

gegeben. In dieser Gleichung ist der Phasenanteil durch α_c, die Dichte der kontinuierlichen Phase durch ρ_c und die Geschwindigkei der Phase durch $\vec{u_c}$ gekennzeichnet. Die Erhaltungsgleichung wird durch Gleichung 2 beschrieben.

$$\begin{aligned}\frac{\partial}{\partial t} \left(\rho_i \vec{u_i} \alpha_i\right) + \nabla \left(\rho \vec{u_i} \vec{u_i} \alpha\right) &= -\alpha_i \nabla p \\ &+ \nabla \tau_i + \alpha_i \rho_i \vec{g} + F_{d,i}\end{aligned} \tag{2}$$

Hierbei ist der Druck durch p und der Stresstensor der Phase i durch τ_i gekennzeichnet. Zusätzlich zu diesen Gleichungen muss für die Zweiphasenströmung die Erhaltung des Phasenanteils gewährleistet sein:

$$\alpha_c + \alpha_d = 1. \tag{3}$$

Die Phaseninteraktion wurde durch den Widerstandsterm berücksichtigt. Die Widerstandskraft wird durch Gleichung 4 beschrieben:

$$F_{d,i} = \frac{3\rho_c \alpha_c \alpha_d C_d |\vec{u}_d - \vec{u}_c| (u_{d,i}^{\vec{}} - u_{c,i}^{\vec{}})}{4 d_d}, \tag{4}$$

wobei C_d über das Modell von Schiller-Naumann [10] berechnet wird.

B. Populationsbilanzgleichung

Die Populationsbilanzgleichung wird generell in Termen der Anzahlkonzentration N ausgedrückt :

$$\begin{aligned}\frac{\partial}{\partial t} \left[\rho_d N\left(V,t\right)\right] + \nabla_e \cdot \left[u \rho_d N\left(V,t\right)\right] \\ + \nabla_i \cdot \left[G \rho_d N\left(V,t\right)\right] = \rho_d S\left(V,t\right),\end{aligned} \tag{5}$$

wobei V das Tropfenvolumen und $S(V,t)$ der Quellterm zur Berücksichtigung von Teilung und Koaleszenz durch die Interaktion der Tropfen ist. Der Wachstumsterm G wird bei Massentransferberechnungen benötigt und wird hier vernachlässigt. Der Quellterm besteht aus vier Termen, jeweils zwei für den Zerfall von Tropfen und zwei für die Koaleszenz der Tropfen. Die zwei Terme berücksichtigen jeweils die Entstehung neuer Tropfen mit einer bestimmten Größe und die damit einhergehende Auslöschung von Tropfen einer bestimmten Größe.

$$\begin{aligned}S\left(V,t\right) &= B^C\left(V,t\right) - D^C\left(V,t\right) \\ &+ B^B\left(V,t\right) - D^B\left(V,t\right)\end{aligned} \tag{6}$$

Die Zerfalls- und Koaleszenzkernel sind Funktionen der Tropfengröße, systemtypische Parameter wie Viskosität und Oberflächenspannung sowie der turbulenten Energie. Die One

Primary One Secondary Particle Methode (OPOSPM), entwickelt von Attarakih et al. [11], ist ist ein Speziallfall der Sectional Quadrature Method of Moments (SQMOM), wobei hier nur ein primäres und ein sekundäres Partikel verwendet wird. Die primären Partikel stellen hierbei die die Klassen dar, wobei bei der OPOSP Methode die Tropfenverteilung durch eine Tropfengröße in jeder Zelle ausgedrückt wird. Die sekundären Partikel können als Lagrange fluide Partikel verstanden werden, die Informationen über die Tropfenpopulation transportieren. Daher wird die Populationsverteilung durch ein einzelnes Partikel repräsentiert, welches eine Kugelform besitzt und deren Position (Größe) durch den mittleren Massendurchmesser gegeben ist:

$$d_{30} = \sqrt[3]{\frac{\pi \alpha_d}{6 N_d}} = \sqrt[3]{\frac{m_3}{m_0}}, \tag{7}$$

wobei die totale Anzahl N und die Volumenkonzentration α_d dem nullten Moment (m_0) und dem dritten Moment (m_3) der Distribution entsprechen. In der Literatur wird häufig der Sauter Durchmesser verwendet. Aufgrund des nicht vorhandenen zweiten Moments in der CFD erscheint die Verwendung des d_{30} zielführender für die Simulation. Durch die Verwendung des nullten Moments, d.h. der Anzahl der Tropfen ist das Modell mathematisch konsistent. Das dritte Moment wird für die Zweiwegekopplung auf Eins normiert und führt zu:

$$d_{30} = \sqrt[3]{\frac{1}{m_0}} \tag{8}$$

Da das dritte Moment der Kontinuitätsgleichung der dispersen Phase entspricht (welche im CFD Solver vorhanden ist), muss nur das nullte Moment durch eine UDF beschrieben werden. Die Quellterme für Zerfall und Koaleszenz werden von jeder Zelle vom CFD Solver zurückgegeben. Der Quellterm für die Tropfenzerfall $g(d_{30})$ Koaleszenz $a(d_{30}, d_{30})$ ist wie folgt definiert:

$$S = \left(n_d\left(d_{30}\right) - 1\right) N_d - \frac{1}{2} a\left(d_{30}, d_{30}\right) N_d^2, \tag{9}$$

wobei n_d die Anzahl der Tochtertropfen, N_d die Anzahl der Partikel ist. Mehr Details über die Kopplung von CFD in Zusammenhang mit dem OPOSPM Modell und dem verwendeten Zweiphasenmodell sind in [9] gegeben. Die Koaleszenzrate $a(d_{30}, d_{30})$ setzt sich aus der Koaleszenzeffizienz $\lambda(d,d)$ und der Kollisionsrate $h(d,d)$ zusammen und wird durch das Modell von Prince und Blanch [1] beschrieben:

$$\lambda(d_{30}, d_{30}) = exp\left(-\frac{(d_{30}/4))^{5/6} \rho_c^{1/2} \epsilon^{1/3}}{4\sigma^{1/2}} ln\left(\frac{h_0}{h_c}\right)\right) \tag{10}$$

$$h(d_{30}, d_{30}) = 0.089\pi(d_{30} + d_{30})^2 \epsilon^{1/3}(d_{30}^{2/3} + d_{30}^{2/3})^{1/2} \tag{11}$$

Hierbei ist ϵ die Energiedissipation, σ die Grenzflächenspannung und h_0 die Filmdicke und h_c die kritische Filmdicke, wobei das Verhältniss der Filmdicken nach Drumm et al. [12] auf 10^4 gesetzt wurde.

Zur Beschreibung des Tropfenzerfalls wird das Modell Martínez-Bazán et al. [2] verwendet:

$$g(d_{30}) = \frac{K\sqrt{\beta_0(\epsilon d)^{2/3} - 12(\sigma/(\rho d))}}{d_{30}} \qquad (12)$$

K ist eine Konstante, die vom ursprünglichen Wert von 0.25 auf einen Wert von 1 gesetzt wurde, um die anpassbaren Parameter zu reduzieren (Drumm et al. [12]).

Die Kenngröße β_0 wurde in vorangegangenen Arbeiten angepasst.

Die Anzahl der Tochtertropfen wurde mit Hilfe einer Hochgeschwindigkeitskamera durch Einzeltropfenversuche in der Miniplant Kühni Kolonne (s. Abb. 4) bestimmt. Einzeltropfen wurden über eine Zweistoffdüse erzeugt, womit die Ablösezeit des Tropfens von der Kapillare sowie die Tropfengröße durch ein um die Kapillare strömendes Fluid beeinflusst werden kann. Der Abstand zwischen den erzeugten Tropfen wurde so groß gewählt, dass sich immer nur ein Tropfen im Kompartment befand. Die entstehenden Tropfen bei einem Zerfallsprozess wurden anschließend über die Videoaufnahmen ausgewertet.
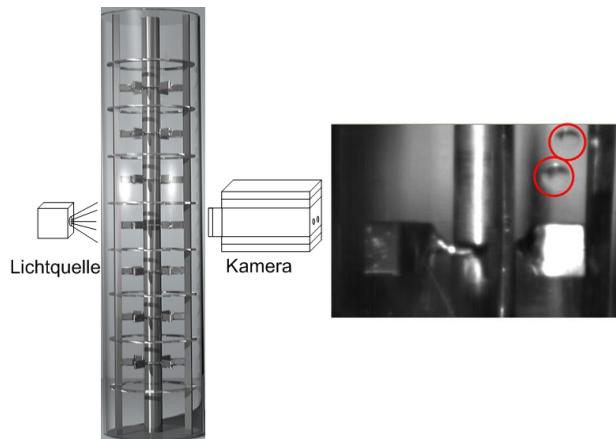
Abb. 4.   Tropfenzerfallsbestimmung

C. Visualisierung

Durch die Verwendung von modernen Visualisierungstools wird es dem Anwender bzw. dem Forscher ermöglicht, die Simulation besser zu analysieren. In einem ersten Schritt wurde dazu die Geometrie auf Basis der Simulationsdaten visualisiert und ist bereits in der Einleitung dargestellt. Dabei wurde auf eine möglichst detailgetreue Darstellung der metallenen Rührer und des Glasschusses wert gelegt (vgl. Abb. 4).

Die Tropfengröße und damit auch die Anzahl der Tropfen stellt eine kritische Größe in der Kolonne dar. Da diese durch das benutzte Modell als Mittelwerte im Gitter berechnet werden, repräsentieren diese eine Wahrscheinlichkeit. Die Schwierigkeit bei der Visualisierung der Tropfen in der Kolonne ist die Bestimmung der Tropfenposition und der Tropfengröße. Stochastische Modellierung ist daher einer der möglichen Wege, um die Position der visualisierten Partikel zu bestimmen. Für jeden Zeitschritt kann eine Anzahl von N Tropfen mit einem Durchmesser von d in einer Zelle

durch eine homogene Verteilungsfunktion für die Tropfenmitte $[p_x, p_y, p_z]$ dargestellt werden:

$$p(x) \begin{cases} \frac{1}{x_{Zellmax.} - x_{Zellmin.}}, & \text{wenn } x_{Zellmin.} \leq x \leq x_{Zellmax.} \\ 0 & \text{ansonsten.} \end{cases}$$

Adäquate Formeln gelten für die Komponenten p_y, p_z. Die Anzahl der Tropfen N kann durch folgende Gleichung beschrieben werden:

$$\sum N \times V_{\text{jeder Tropfen}} = \alpha \times V_{\text{Zellenvolumen,}}$$

wobei α den Phasenanteil der dispersen Phase darstellt und durch die Simulationsdaten zur Verfügung gestellt wird.

Neben der Tropfenvisualisierung soll durch eine Kombination der CFD mit der Tropfenvisualisierung die Verfolgung einzelner Tropfen sowie durch Erweiterungen die Visualisierung des Tropfenzerfalls und der Tropfenkoaleszenz ermöglicht werden. Hierzu erfolgte in einem ersten Schritt die Visualisierung des Vektorfeldes in der Auswertesoftware und um diese um Partikelbahnen zu erweitern. Die Partikelbahnen zeigen den Weg der Tropfen durch die Kolonne über die Zeit. Der Start einer solchen Partikelbahn wird durch den Einlass der Kolonne definiert und endet am Auslass einer Kolonne bzw. durch einen Zerfall des Tropfen. Die mathematische Beschreibung der Tropfen erfolgt durch:

$$\begin{cases} \dfrac{d}{dt}\vec{p}(x,y,z) = \vec{v}(p,t) \\ \vec{p}(t_0) = \vec{p_0} \end{cases}.$$

Die Partikelbahnen der Tropfen können für eine kontinuierliche Visualisierung der Tropfenverteilung verwendet werden. Überschneidungen zweier Bahnen können als Indikator einer Tropfenkoaleszenz verwedendet werden.

IV. Ergebnisse und Diskussion

In vorangegangenen Forschungsarbeiten wurde der Zerfallskernel aus dem Modell an die Ergebnisse von Steinmetz et al. [13] angepasst. Die Tochtertropfenanzahl ergab sich aus den Einzeltropfenexperimenten. Es wurden im Mittel 4 Tochtertropfen festgestellt. Durch die Kopplung von CFD mit PBM wurde eine Aussage über die Tropfengröße ermöglicht. Die simulierte Tropfengröße mit dem gekoppelten CFD-PBM Tool ist in Abb. 5 für die gesamte Kolonne sowie vergrößert für ein einzelnes Kompartment dargestellt. Die Tropfen koaleszieren unterhalb der Statoren und werden durch den Energieeintrag des Rührers zerteilt. Die Tropfengröße im Bereich des Rührers bei 2,6 mm und sinkt aufgrund der steigenden Turbulenz auf 2,4 mm direkt an der Rührerspitze ab. Durch Koaleszenz steigt die Tropfengröße auf 3,1 mm unterhalb der Statoren wieder an.

Durch eine erweiterte Visualisierung wird eine Kombination des Tropfendurchmessers mit dem Phasenanteil erzielt, womit Aussagen über die Tropfenverteilung im dreidimensionalen Raum durch eine einzelne Abbildung, die den real sichtbaren Verhältnissen entspricht, getroffen werden können. Dies ist ist Abb. 6 dargestellt. Die Tropfen treten durch den unteren
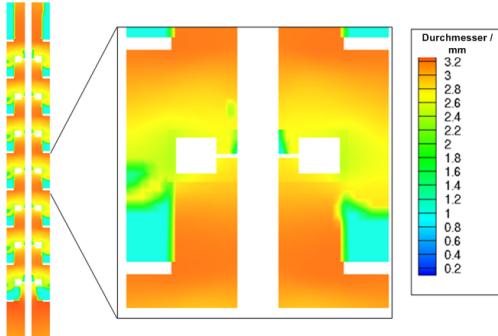
Abb. 5. Tropfengrößenverteilung dargestellt als Schnitt durch die Kolonne

Statorring in das Kompartment ein. Im unteren Bereich des Kompartments sind die Tropfen auf die Mitte des Kompartments hin konzentriert, während sie im oberen Teil durch die Rührerwirkung verteilt werden. Das Vektorfeld ist in Abb. 7 gezeigt, wobei zur Übersichtlichkeit die Anzahl der Vektoren reduziert wurde.

Die Visualisierung bietet die Möglichkeit der Filterung von Daten, z.B. extrem großer bzw. kleiner Tropfen und wird durch die Visualisierung der Zerfallswahrscheinlichkeit sowie der Koaleszenzwahrscheinlichkeit ergänzt.
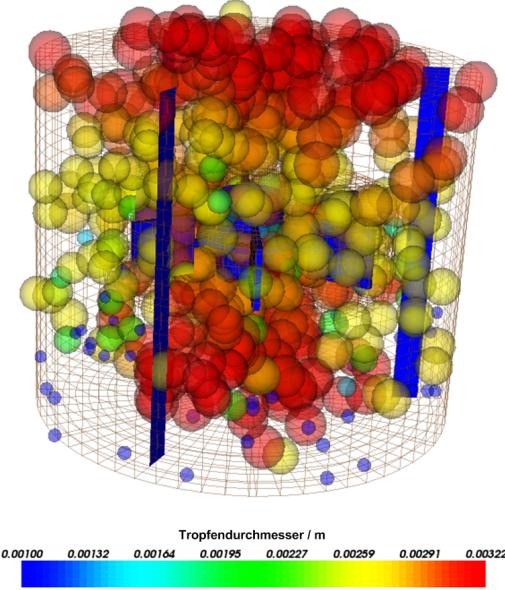
Abb. 6. Darstellung der Tropfen sowie der Tropfengrößenverteilung

V. Zusammenfassung

Die Simulationen der sich im Gegenstrom befindlichen Zweiphasenströmung für die Miniplant Extraktionskolonne vom Typ Kühni wurde mit dem CFD solver (FLUENT) gekoppelt mit Populationsbilanzmodellen durchgeführt. Hierfür wurde ein Eingruppenmodell, das One Primary One Secondary Particle Modell, für die Zweiwegekopplung verwendet. Zur Berücksichtigung von Koaleszenz und Zerfall der Tropfen
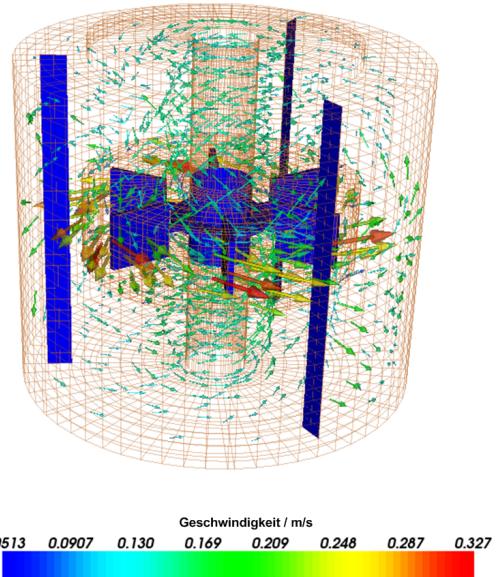
Abb. 7. Vektorfeld der kontinuierlichen Phase im Kompartment

wurde ein kombiniertes Modell, bestehend aus dem Zerfallskernel von Martínez-Bazán [2] und dem Koaleszenzkernel von Prince und Blanch [1] verwendet. Durch gemittelte Daten auf Basis stochastischer Modelle erfolgte eine Visualisierung der Tropfen, womit erstmals zwei Parameter, der Phasenanteil und die Tropfengröße kombiniert in einer Abbildung dargestellt werden können.

VI. Ausblick

Das bestehende CFD Modell wird durch weitere Koaleszenz- und Teilungskernel erweitert werden, um eine umfangreiche Basis für verschiedene Stoffeigenschaften sicherzustellen. Eine Implementierung von Massentransfer wird zunächst anhand des zweidimensional rotationssymmetrischen Modells der RDC Kolonne erfolgen und mit Literaturdaten sowie eigenen Messungen validiert werden. Eine Übertragung der Ergebnisse auf unterschiedliche Kolonnengrößen sowie ist in naher Zukunft geplant. Die Visualisierung wird durch die Möglichkeit zur Erstellung von Animationen bewegter Tropfen erweitert werden. 3D Partikelbahnen können durch eine Standard-Integrationsmethode eingebunden werden. Die Implementierung von Tropfenkollisionen und Tropfenteilungen in die Visualisierungsumgebung erfordern zusätzliche Modellierung sowie Implementierungen. Topologische Analysen der transienten Tropfengrößenverteilung sind dazu mit dem Morse-Smale Komplex sowie mit Reeb Graphs geplant.

Danksagung

LITERATUR

[1] M. J. Prince and H. W. Blanch, "*Bubble Coalescence and Break-up in Air-Sparged Bubble Columns*," *AIChE J.*, vol. 36, no. 10, pp. 1485–1499, 1990.

[2] C. Martínez-Bazán, J. L.Monatenes, and J. C. Lasheras, "*On the Breakup of an Air Bubble injected into a Fully Developed Turbulent Flow. Part 1. Breakup frequency*," *Journal of Fluid Mechanics*, vol. 401, pp. 157–182, 1999.

[3] G. Modes and H.-J. Bart, "*CFD Simulation of Nonideal Dispersed Phase Flow in Stirred Extraction Columns*," *Chem. Eng. Technol.*, vol. 24, no. 12, pp. 1242–1245, 2001.

[4] R. Rieger, C. Weiss, G. Wigley, H. J. Bart, and R. Marr, "*Investigating the Process of Liquid-Liquid Extraction by means of Computational Fluid Dynamics*," *Computers & Chemical Engineering*, vol. 20, no. 12, pp. 1467 – 1475, 1996.

[5] T. Haderer, R. Marr, S. Martens, and M. Siebenhofer, "*Bestimmung auslegungsrelevanter hydrodynamischer Kenngrößen einer RDC-Extraktionskolonne mit CFD*," *Chemie Ingenieur Technik*, vol. 77, no. 8, pp. 1055–1055, 2005.

[6] P. Kolb, "*Hydrodynamik und Stoffaustausch in einem gerührten Miniplantextraktor der Bauart Kühni*," Dissertation, TU Kaiserslautern, Germany, 2005.

[7] F. Wang and Z.-S. Mao, "*Numerical and Experimental Investigation of LiquidâLiquid Two-Phase Flow in Stirred Tanks*," *Industrial & Engineering Chemistry Research*, vol. 44, no. 15, pp. 5776–5787, 2005.

[8] A. Vikhansky and M. Kraft, "*Modelling of a RDC using a combined CFD-population balance approach*," *Chemical Engineering Science*, vol. 59, no. 13, pp. 2597 – 2606, 2004.

[9] C. Drumm, M. Attarakih, M. W. Hlawitschka, and H.-J. Bart, "*One-Group Reduced Population Balance Model for CFD Simulation of a Pilot-Plant Extraction Column*," *Industrial & Engineering Chemistry Research*, vol. 49, no. 7, pp. 3442–3451, 2010.

[10] L. Schiller and Z. Naumann, "*Über die grundlegenden Berechnungen bei der Schwerkraftaufbereitung.*" *Zeitschrift des Vereines Deutscher Ingenieure*, vol. 77, p. 143, 1935.

[11] M. M. Attarakih, M. Jaradat, C. Drumm, H.-J. Bart, S. Tiwari, V. K. Sharma, J. Kuhnert, and A. Klar, "*Solution of the Population Balance Equation using the One Primary and One Secondary Particle Method (OPOSPM)*," *Proc. ESCAPE, Cracow, Poland, June 14-17.*, vol. 19, 2009.

[12] C. Drumm, M. M. Attarakih, and H.-J. Bart, "*Coupling of CFD with DPBM for an RDC extractor*," *Chemical Engineering Science*, vol. 64, pp. 721–732, 2009.

[13] T. Steinmetz, *Tropfenpopulationsbilanzgestütztes Auslegungsverfahren zur Skalierung einer gerührten Miniplant-Extraktionskolonne*. Fortschr.-Ber. VDI, Reihe 3, Nr. 885, Düsseldorf: VDI Verlag, 2007.

Nonlinear Finite Element Simulation of Thin Dielectric Elastomer Structures

Sandro Zwecker [#1], Sven Klinkel [#2], R. Müller [*3]

[#] *Statik und Dynamik der Tragwerke, TU Kaiserslautern*
Paul-Ehrlich-Str. 14, 67663 Kaiserslautern, Germany
[1] zwecker@rhrk.uni-kl.de
[2] klinkel@rhrk.uni-kl.de

[*] *Lehrstuhl für Technische Mechanik, TU Kaiserslautern*
Postfach 30 49, 67653 Kaiserslautern, Germany
[3] ram@rhrk.uni-kl.de

Abstract— **To simulate the behavior of thin dielectric elastomer structures a finite solid shell element formulation is presented. Dielectric elastomers belong to the group of electroactive polymers and their use as actuators is caused by the efficient coupling between electrical energy input and mechanical energy output. Also the large elongation strain of 120-380% of the dielectric elastomer actuators and their light weight are advantages that make the material very attractive. Regarding the electro-mechanical coupling a constitutive model is expounded. For the definition of an electric stress tensor and a total stress tensor the electrical body force and couple are considered in the balance of linear momentum and angular momentum, respectively. The governing constitutive equations are derived and incorporated in a solid shell element formulation based on a Hu-Washizu mixed variational principle considering six fields: displacements, electric potential, strains, electric field, mechanical stresses, and dielectric displacements. This formulation allows large deformations and accounts for physical nonlinearities to capture the main characteristics of dielectric elastomers.**

I. INTRODUCTION

In recent years dielectric elastomers (DE) have become popular for the usage in actuators. The efficient transformation of electrical energy in mechanical energy and the ability to maintain large strains makes them very attractive. A Constitutive model, which describes the specific material behavior, is introduced by Dorfmann and Ogden [1], Steigmann [2] and the references therein. The numerical treatment in the context of the finite element method is presented by Vu et al. [3], who provide a brick element formulation, which incorporates the nonlinear constitutive model. In the present work a shell finite element formulation for DE is presented. It is motivated by the fact that the most DE devices are thin structures, which have a very high length to thickness ratio. The electromechanical coupling is considered in the body force and the couple density, see e.g. Eringen and Maugin [4] or Müller et al. [5]. The angular momentum equation is fulfilled by assuming a Maxwell stress tensor. The linear momentum equation is approximately fulfilled by the finite element method. A mixed solid shell element formulation is introduced. It incorporates specific interpolations for the displacements, strains, stresses as well as

for the electric potential, electric field and dielectric displacements. As nodal degrees of freedom three displacements and the electric potential are assumed. The mixed formulation allows for a consistent finite element approximation to avoid electromechanical locking effects. The element formulation is able to simulate large deformations. Some numerical examples show the applicability of the proposed solid shell element.

II. KINEMATICS

Let Φ be a deformation that point maps $\vec{X}$ of the reference configuration $\mathcal{B}$ to $\vec{x}$ of the current configuration $\mathcal{B}_t$ at time t, see Figure 1. The deformation gradient $\boldsymbol{F}$ is declared as the tangent to Φ and given by

$$\boldsymbol{F} = \frac{\partial \vec{x}}{\partial \vec{X}} . \tag{1}$$

With this definition for the deformation gradient the right Cauchy-Green deformation tensor $\boldsymbol{C}$ reads,

$$\boldsymbol{C} = \boldsymbol{F}^{\mathrm{T}} \boldsymbol{F} . \tag{2}$$

Followed by the Green-Lagrange strain tensor $\boldsymbol{E}$ defined as

$$\boldsymbol{E} = \frac{1}{2} \left(\boldsymbol{C} - \boldsymbol{1} \right) . \tag{3}$$
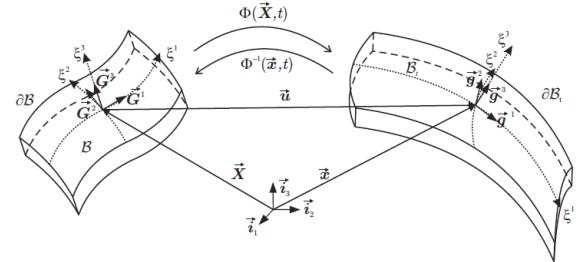

Fig. 1 Reference and current configuration with position vectors $\vec{X}$ and $\vec{x}$

To describe the shell formulation convective coordinates ξ^i are introduced, where ξ^1 and ξ^2 are the in-plane coordinates and ξ^3 is the coordinate in thickness direction. The covariant

tangent vectors are obtained in the reference and current configuration as $\vec{G}_i = \frac{\partial \vec{X}}{\partial \xi^i}$, $\vec{g}_i = \frac{\partial \vec{x}}{\partial \xi^i}$ respectively. The contravariant basis vectors are defined in a standard manner by $\vec{G}_i \cdot \vec{G}^j = \delta_i^j$ and $\vec{g}_i \cdot \vec{g}^j = \delta_i^j$, where δ_i^j is the Kronecker-delta. With this convective description the deformation gradient also reads $\boldsymbol{F} = \vec{g}_i \otimes \vec{G}^i$.

A potential φ is used to describe the electric field $\vec{E}$. With φ satisfying the Laplace's equation $\vec{E}$ reads

$$\vec{E} = -\frac{\partial \varphi}{\partial \xi^i} \vec{G}^i . \tag{4}$$

Applying the pull-back operation the physical electric field $\vec{e} = \vec{E}\boldsymbol{F}^{-1}$ is observed. The displacement vector $\vec{u}$ is defined by

$$\vec{u} = \vec{x} - \vec{X} . \tag{5}$$

Boundary conditions for $\vec{u}$ and φ are given on $\partial_u B$ and $\partial_\varphi B$, respectively.

III. FORCE AND COUPLE DENSITY

To get a macroscopic representation of force and couple density a close look on the microscopic level is presented. Therefore the physical model will start with a particle P^k of the current configuration $\mathcal{B}_t$ to deduce electric force in the presence of electromagnetic matter and in absence of a magnetic field.
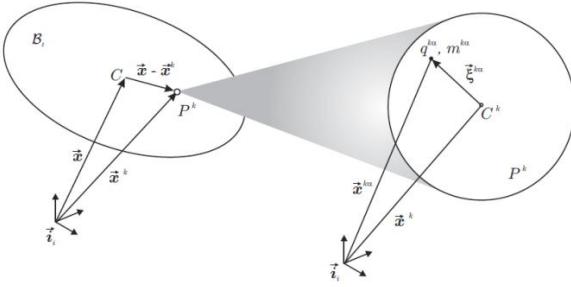
Fig. 2 Physical model for the microscopic description; left: the current configuration $\mathcal{B}_t$ with a particle P^k, right: the zoomed inner structure of the particle.

Relative to the mass center C^k eccentric by $\vec{\xi}^{k\alpha}$ not to be confused with the convective coordinates ξ^i are the point charges $q^{k\alpha}$ and the point masses $m^{k\alpha}$, as shown in Figure 2. In an electric field a force is acting on each point charge $q^{k\alpha}$ called Lorentz force and it is defined by $q^{k\alpha} \vec{e}$. The gravitation field $\vec{a}$ is acting on each point mass producing the Newton force given by $m^{k\alpha} \vec{a}$. Summing over α leads to the resultant force on the particle:

$$\vec{b}^k = \sum_\alpha m^{k\alpha} \vec{a} + q^{k\alpha} \vec{e} . \tag{6}$$

The position of a point within the particle P^k can be described by $\vec{x}^{k\alpha} = \vec{x}^k + \vec{\xi}^{k\alpha}$. Assuming the gravitation to be constant it follows: $\vec{a}(\vec{x}^{k\alpha}) = \vec{a}(\vec{x}^k)$. This assumption and expanding the external field in a Taylor series

$\vec{e}(\vec{x}^{k\alpha}) = \vec{e}(\vec{x}^k) + [\nabla_x \otimes \vec{e}] \cdot \vec{\xi}^{k\alpha} + \cdots$ by neglecting higher order terms results in

$$\vec{b}^k = (\sum_\alpha m^{k\alpha}) \vec{a} + (\sum_\alpha q^{k\alpha}) \vec{e} + [\nabla_x \otimes \vec{e}] \cdot (\sum_\alpha q^{k\alpha} \vec{\xi}^{k\alpha}) . \tag{7}$$

After defining the resultant mass $\mathfrak{m}^k = \sum_\alpha m^{k\alpha}$, charge $\mathfrak{q}^k = \sum_\alpha q^{k\alpha}$, and polarization $\vec{\mathfrak{p}}^k = \sum_\alpha q^{k\alpha} \vec{\xi}^{k\alpha}$ of the particle P^k and performing a simple space averaging $\rho_t dv = \sum_k \mathfrak{m}^k$, $\varrho_t dv = \sum_k \mathfrak{q}^k$, $\vec{p} \, dv = \sum_k \vec{\mathfrak{p}}^k$ leads to the macroscopic body force density

$$\vec{b} = \rho_t \vec{a} + \varrho_t \vec{e} + \vec{p} \cdot [\nabla_x \otimes \vec{e}] = \rho_t \vec{a} + \vec{f} , \tag{8}$$

where $\vec{f}$ combines the electric contribution to the body force density. Deriving the corresponding couple density with the same arguments the resultant couple of the particle P^k with respect to the mass center C^k reads

$$\vec{c}^k = \sum_\alpha \vec{\xi}^{k\alpha} \times m^{k\alpha} \vec{a} + \sum_\alpha \vec{\xi}^{k\alpha} \times q^{k\alpha} \vec{e} . \tag{9}$$

Here it is assumed that the mass dipole moment $\sum_\alpha \vec{\xi}^{k\alpha} m^{k\alpha} = \vec{0}$. With the resultant polarization $\vec{\mathfrak{p}}^k = \sum_\alpha q^{k\alpha} \vec{\xi}^{k\alpha}$ and a space averaging it follows the macroscopic couple density

$$\vec{c} = \vec{p} \times \vec{e} . \tag{10}$$

IV. BALANCE LAWS AND STRESS TENSOR

The balance laws incorporating the electric force, couple, and power densities are summarized. After that the global integral forms and the local field equations are presented. The Cauchy stress tensor and an electric stress tensor are introduced. Conservation of mass $\frac{d}{dt} \int_\mathcal{B} \rho_t J dV = 0$, with $J = \det \boldsymbol{F}$ is assumed. Localization of the material description results in $\rho = J \rho_t$. In quasi static processes the integral form of the balance of linear momentum is given as

$$0 = \int_{\mathcal{B}_t} \vec{b} \, dv + \int_{\partial \mathcal{B}_t} \vec{t} \, da , \tag{11}$$

where $\vec{b}$ is the body force density (8) and $\vec{t}$ the traction vector on $\partial \mathcal{B}_t$. With the Cauchy stress tensor $\boldsymbol{\sigma}$ the traction is determined by a linear map of the normal vector $\vec{t} = \boldsymbol{\sigma}\vec{n}$, according to Cauchy's stress theorem. Considering conservation of mass, applying the divergence theorem and the localization theorem, the field equation along with the jump condition are observed as

$$\nabla_x \cdot \boldsymbol{\sigma} + \vec{b} = \vec{0} \qquad \text{in } \mathcal{B}_t , \tag{12}$$

$$\boldsymbol{\sigma} \cdot \vec{n} = \vec{\bar{t}} \qquad \text{on } \partial_\sigma \mathcal{B}_t . \tag{13}$$

With $\vec{\bar{t}}$ as a prescribed traction on the boundary $\partial_\sigma B_t$. The boundaries with a given traction and a given displacement satisfy $\partial B_t = \partial_\sigma B_t \cup \partial_u B_t$ and $\emptyset = \partial_\sigma B_t \cap \partial_u B_t$. The global form of the balance of angular momentum reads

$$0 = \int_{\mathcal{B}_t} \vec{x} \times \vec{b} + \vec{c} \, dv + \int_{\partial \mathcal{B}_t} \vec{x} \times \vec{t} \, da , \tag{14}$$

where $\vec{c}$ is the couple density (10). Using the integral theorem, considering conservation of mass and the linear momentum balance along with the localization theorem results in the field equation

$$\mathbf{1} \times \boldsymbol{\sigma} + \vec{c} = \vec{\mathbf{0}} \qquad \text{in } \mathcal{B}_t . \qquad (15)$$

For dielectric materials the conservation of charge in integral form with the surface charge density ς_t on $\partial_d \mathcal{B}_t$ reads

$$0 = \int_{\mathcal{B}_t} \varrho_t \, \mathrm{d}v + \int_{\partial \mathcal{B}_t} \varsigma_t \, \mathrm{d}a . \qquad (16)$$

The dielectric displacement vector is denoted by $\vec{d}$ and determined by Gauss' law $\vec{d} \cdot \vec{n} = \varsigma_t$. Applying the divergence theorem results in the field equation along with the jump condition

$$\nabla_x \cdot \vec{d} - \varrho_t = 0 \qquad \text{in } \mathcal{B}_t , \qquad (17)$$

$$\vec{d} \cdot \vec{n} = \overline{\varsigma}_t \qquad \text{on } \partial_d \mathcal{B}_t . \qquad (18)$$

Here, $\overline{\varsigma}_t$ is a prescribed surface charge. For the total boundaries with a given electric potential and a given surface charge it holds $\partial \mathcal{B}_t = \partial_\varphi \mathcal{B}_t \cup \partial_d \mathcal{B}_t$ and $\emptyset = \partial_\varphi \mathcal{B}_t \cap \partial_d \mathcal{B}_t$. With the constitutive equations in matter the dielectric displacements are given as $\vec{d} = \epsilon_0 \vec{e} + \vec{p}$, where ϵ_0 is the permittivity in vacuum and the polarization depends on the considered material. An electric stress tensor is introduced as

$$\boldsymbol{\tau} = \vec{e} \otimes \vec{p} + \epsilon_0 \vec{e} \otimes \vec{e} - \frac{1}{2} \epsilon_0 \vec{e} \cdot \vec{e} \, \mathbf{1} , \qquad (19)$$

such that $\nabla_x \cdot \boldsymbol{\tau} = \vec{f}$ and $\mathbf{1} \times \boldsymbol{\tau} = \vec{c}$. It is remarked that $\epsilon_0 \vec{e} \otimes \vec{e} - \frac{1}{2} \epsilon_0 \vec{e} \cdot \vec{e} \, \mathbf{1}$ is also known as Maxwell stress tensor. With (15) it follows that the total stress tensor $\boldsymbol{\sigma} + \boldsymbol{\tau}$ has to be symmetric. The remaining field equations and boundary conditions are

$$\nabla_x \cdot (\boldsymbol{\sigma} + \boldsymbol{\tau}) + \rho_t \vec{a} = \vec{\mathbf{0}} \qquad \text{in } \mathcal{B}_t , \qquad (20)$$

$$\boldsymbol{\sigma} \cdot \vec{n} = \vec{\bar{t}} \qquad \text{on } \partial_\sigma \mathcal{B}_t , \qquad (21)$$

$$\nabla_x \cdot \vec{d} - \varrho_t = 0 \qquad \text{in } \mathcal{B}_t , \qquad (22)$$

$$\vec{d} \cdot \vec{n} = \overline{\varsigma}_t \qquad \text{on } \partial_d \mathcal{B}_t . \qquad (23)$$

Considering the 2$^{\text{nd}}$ Piola-Kirchhoff stress tensor $\boldsymbol{S} = J \boldsymbol{F}^{-1} \boldsymbol{\sigma} \boldsymbol{F}$ and $\boldsymbol{T} = J \boldsymbol{F}^{-1} \boldsymbol{\tau} \boldsymbol{F}$, the pull-back of the dielectric displacement $\vec{D} = J \boldsymbol{F}^{-1} \vec{d}$ and the transformation of the densities by J leads to the material description

$$\nabla_X \cdot [\boldsymbol{F}(\boldsymbol{S} + \boldsymbol{T})] + \rho_t \vec{a} = \vec{\mathbf{0}} \qquad \text{in } \mathcal{B} , \qquad (24)$$

$$\boldsymbol{F} \boldsymbol{S} \cdot \vec{N} = \vec{\bar{T}} \qquad \text{on } \partial_S \mathcal{B} , \qquad (25)$$

$$\nabla_X \cdot \vec{D} - \varrho = 0 \qquad \text{in } \mathcal{B} , \qquad (26)$$

$$\vec{D} \cdot \vec{N} = \overline{\varsigma} \qquad \text{on } \partial_D \mathcal{B} . \qquad (27)$$

Where $\vec{\bar{T}}$ is the traction with respect to the reference configuration.

V. CONSTITUTIVE EQUATIONS

Introducing the energy function

$$\Psi = \psi - \frac{1}{2} \chi \epsilon_0 (\vec{E} \otimes \vec{E}) : \boldsymbol{C}^{-1} - \frac{1}{2} J \epsilon_0 [(\vec{E} \otimes \vec{E}) : \boldsymbol{C}^{-1}] \boldsymbol{C}^{-1} , \qquad (28)$$

where ψ is a function of $\boldsymbol{C}$ to fulfill material objectivity, here an Ogden-type material is chosen, and the susceptibility of the material is denoted by χ, the total stress and the dielectric displacements are derived as

$$\boldsymbol{S} + \boldsymbol{T} = \frac{\partial \Psi}{\partial \boldsymbol{E}} , \qquad (29)$$

$$\vec{D} = -\frac{\partial \Psi}{\partial \vec{E}} . \qquad (30)$$

VI. WEAK FORMULATION

In this section a mixed variational formulation is introduced. Let the set $\{\delta \vec{x} \in [H^1(\mathcal{B})]^3 \, \| \, \delta \vec{x} = \vec{\mathbf{0}} \text{ on } \partial_u \mathcal{B}\}$ be the space of admissible displacement variations and $\{\delta \varphi \in [H^1(\mathcal{B})] \, \| \, \delta \varphi = 0 \text{ on } \partial_\varphi \mathcal{B}\}$ be the space of admissible electric potential variations. Further let $\{\delta \boldsymbol{S} + \delta \boldsymbol{T} \in [L_2(\mathcal{B})]^6\}$, $\{\delta \boldsymbol{E} \in [L^2(\mathcal{B})]^6\}$ the spaces of admissible variations of the total stresses and strains and $\{\delta \vec{D} \in [L_2(\mathcal{B})]^3\}$, $\{\delta \vec{E} \in [L_2(\mathcal{B})]^3\}$ the spaces of admissible variations of the dielectric displacement and the electric field. Since the variations are arbitrary the field equations (24)-(27), the constitutive equations (29), (20), and the kinetic field equations (3), (4) are rewritten as

$$\int_{\mathcal{B}} (\nabla_X \cdot [\boldsymbol{F}(\boldsymbol{S} + \boldsymbol{T})] + \rho \vec{a}) \cdot \delta \vec{x} \, \mathrm{d}V = 0 , \qquad (31)$$

$$\int_{\mathcal{B}} (\nabla_X \cdot \vec{D} - \varrho) \, \delta \varphi \, \mathrm{d}V = 0 , \qquad (32)$$

$$\int_{\mathcal{B}} \left(\frac{\partial \Psi}{\partial \boldsymbol{E}} - (\boldsymbol{S} + \boldsymbol{T}) \right) : \delta \boldsymbol{E} \, \mathrm{d}V = 0 , \qquad (33)$$

$$\int_{\mathcal{B}} \left(\frac{\partial \Psi}{\partial \vec{E}} + \vec{D} \right) \cdot \delta \vec{E} \, \mathrm{d}V = 0 , \qquad (34)$$

$$\int_{\mathcal{B}} \left(\boldsymbol{E} - \frac{1}{2} (\boldsymbol{C} - \mathbf{1}) \right) : (\delta \boldsymbol{S} + \delta \boldsymbol{T}) \, \mathrm{d}V = 0 , \qquad (35)$$

$$\int_{\mathcal{B}} (\vec{E} + \nabla_X \varphi) \cdot \delta \vec{D} \, \mathrm{d}V = 0 . \qquad (36)$$

Applying integration by part, using the divergence theorem and considering the boundary conditions the weak formulation reads

$$\delta \pi = \int_{\mathcal{B}} \frac{1}{2} (\delta \boldsymbol{F}^{\text{T}} \boldsymbol{F} + \boldsymbol{F}^{\text{T}} \delta \boldsymbol{F}) : (\boldsymbol{S} + \boldsymbol{T}) + \nabla_X \delta \varphi \cdot \vec{D} - \\ \delta \vec{x} \cdot \vec{a} \rho - \varrho \delta \varphi \, \mathrm{d}V - \int_{\partial \mathcal{B}} \delta \vec{x} \cdot \vec{\bar{T}} + \delta \varphi \overline{\varsigma} \, \mathrm{d}A + \\ \int_{\mathcal{B}} (\delta \boldsymbol{S} + \delta \boldsymbol{T}) : \frac{1}{2} (\boldsymbol{F}^{\text{T}} \boldsymbol{F} - \mathbf{1}) + \delta \vec{D} \cdot \nabla_X \varphi \, \mathrm{d}V + \\ \int_{\mathcal{B}} \delta \boldsymbol{E} : \frac{\partial \Psi}{\partial \boldsymbol{E}} + \delta \vec{E} \cdot \frac{\partial \Psi}{\partial \vec{E}} - \delta \boldsymbol{E} : (\boldsymbol{S} + \boldsymbol{T}) + \delta \vec{E} \cdot \vec{D} - \\ (\delta \boldsymbol{S} + \delta \boldsymbol{T}) : \boldsymbol{E} + \delta \vec{D} \cdot \vec{E} \, \mathrm{d}V = 0 , \qquad (37)$$

with $\delta \boldsymbol{F} = \boldsymbol{\nabla}_X \otimes \delta \vec{x}$, $\delta \boldsymbol{E} = \frac{1}{2}\left(\delta \boldsymbol{F}^{\mathrm{T}} \boldsymbol{F} + \boldsymbol{F}^{\mathrm{T}} \delta \boldsymbol{F}\right)$, and $\delta \vec{E} = -\boldsymbol{\nabla}_X \delta \varphi$.

VII. FINITE ELEMENT APPROXIMATION

In this section a solid shell element is introduced. The finite element approximation is constructed in the sense that the whole domain is divided in element domains with $\mathcal{B} = \bigcup_{\mathcal{B}e=1}^{nelem} \mathcal{B}_e$, where $nelem$ is the total number of elements. The geometry, displacements, and electric potential are approximated as $\vec{X}_e^h = \sum_{I=1}^{8} N_I \vec{X}_I$, $\vec{u}_e^h = \sum_{I=1}^{8} N_I \vec{u}_I$, and $\varphi_e^h = \sum_{I=1}^{8} N_I \varphi_I$ with the same interpolation functions $N_I = \frac{1}{2}\left(1 + \xi_I^1 \xi^1\right)\left(1 + \xi_I^2 \xi^2\right)\left(1 + \xi_I^3 \xi^3\right)$, $-1 \leq \xi^i \leq +1$ at nodes $I = 1,2,3,...,8$. The vectors $\vec{X}_I$ and $\vec{u}_I$ contain the nodal coordinates and displacements, respectively. Arranging N_I in the matrix $\boldsymbol{N} = [N_1, N_2, N_3, N_4, N_5, N_6, N_7, N_8]$ with $\boldsymbol{N}_I = \mathrm{diag}[N_I, N_I, N_I, N_I]$, the virtual quantities are interpolated as

$$\delta \vec{v}^h = \begin{bmatrix} \delta \vec{u}_e^h \\ \delta \varphi_e^h \end{bmatrix} = \boldsymbol{N} \delta \vec{v}_e \ , \tag{38}$$

where $\vec{v}_e^{\mathrm{T}} = [\vec{v}_1^{\mathrm{T}}, \ \vec{v}_2^{\mathrm{T}}, \ \vec{v}_{3,}^{\mathrm{T}} \ \cdots, \ \vec{v}_8^{\mathrm{T}}]$ is the vector of nodal degrees of freedom $\vec{v}_I^{\mathrm{T}} = [\vec{u}_1, \ \vec{u}_2, \ \vec{u}_3 \ \ \varphi]_I$. Accordingly, $\delta \vec{v}_e$ is the vector of the virtual values.

The gradient fields are defined with respect to the curvilinear coordinates ξ^i. The constitutive equations (29) and (30) will be given with respect to a local orthonormal coordinate system $\vec{r}_i$. This necessitates a transformation of the strains and the electric field. Introducing the transformation matrix

$$\boldsymbol{Q} = \begin{bmatrix} (J_{11})^2 & (J_{12})^2 & (J_{13})^2 & aJ_{11}J_{12} & aJ_{11}J_{13} & aJ_{12}J_{13} \\ (J_{21})^2 & (J_{22})^2 & (J_{23})^2 & aJ_{21}J_{22} & aJ_{21}J_{23} & aJ_{22}J_{23} \\ (J_{31})^2 & (J_{32})^2 & (J_{33})^2 & aJ_{31}J_{32} & aJ_{31}J_{33} & aJ_{32}J_{33} \\ bJ_{11}J_{21} & bJ_{12}J_{22} & bJ_{13}J_{23} & J_{11}J_{22}+J_{12}J_{21} & J_{11}J_{23}+J_{13}J_{21} & J_{12}J_{23}+J_{13}J_{22} \\ bJ_{11}J_{31} & bJ_{12}J_{32} & bJ_{13}J_{33} & J_{11}J_{32}+J_{12}J_{31} & J_{11}J_{33}+J_{13}J_{31} & J_{12}J_{33}+J_{13}J_{32} \\ bJ_{21}J_{31} & bJ_{22}J_{32} & bJ_{23}J_{33} & J_{11}J_{32}+J_{22}J_{31} & J_{21}J_{33}+J_{23}J_{31} & J_{22}J_{33}+J_{23}J_{32} \end{bmatrix}, \tag{39}$$

where $\boldsymbol{Q}_S = \boldsymbol{Q}$ defined with $a = 1$, $b = 2$ and $J_{ik} = \vec{G}_i \cdot \vec{r}_k$ and $\vec{r}_1 = \frac{\vec{G}_1}{\|\vec{G}_1\|}$, $\vec{r}_2 = \frac{\vec{G}_3 \times \vec{G}_1}{\|\vec{G}_3 \times \vec{G}_1\|}$, $\vec{r}_3 = \vec{r}_1 \times \vec{r}_2$. The Jacobian matrix is denoted by $\boldsymbol{J} = \vec{G}_i \otimes \vec{r}_i$. For the sake of a compact notation the contravariant components of the virtual strain tensor and the virtual electric field vector are arranged in a generalized vector

$$\delta \vec{\varepsilon}_G = \begin{bmatrix} \boldsymbol{Q}_S^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{J}^{-1} \end{bmatrix} \cdot \begin{bmatrix} \delta \vec{g}_1 \cdot \vec{g}_1 \\ \delta \vec{g}_2 \cdot \vec{g}_2 \\ \delta \vec{g}_3 \cdot \vec{g}_3 \\ \delta \vec{g}_1 \cdot \vec{g}_2 + \vec{g}_1 \cdot \delta \vec{g}_2 \\ \delta \vec{g}_1 \cdot \vec{g}_3 + \vec{g}_1 \cdot \delta \vec{g}_3 \\ \delta \vec{g}_2 \cdot \vec{g}_3 + \vec{g}_2 \cdot \delta \vec{g}_3 \\ -\frac{\partial \delta \varphi}{\partial \vec{\xi}^1} \\ -\frac{\partial \delta \varphi}{\partial \vec{\xi}^2} \\ -\frac{\partial \delta \varphi}{\partial \vec{\xi}^3} \end{bmatrix} . \tag{40}$$

The approximation on element level of the virtual gradient field $\delta \vec{\varepsilon}_G$ reads

$$\delta \vec{\varepsilon}_G = \boldsymbol{B} \delta \vec{v}_e \ , \tag{41}$$

with $\boldsymbol{B} = [\boldsymbol{B}_1, \ \boldsymbol{B}_2, \ \boldsymbol{B}_3, \ \cdots, \ \boldsymbol{B}_8]$ and $\boldsymbol{B}_I = \begin{bmatrix} \boldsymbol{B}_I^u & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B}_I^\varphi \end{bmatrix}$. The matrix $\boldsymbol{B}_I^u$ is defined with some ANS interpolations in [6] and reads

$$\boldsymbol{B}_I^u = \boldsymbol{Q}_S^{-1} \begin{bmatrix} N_{I,\xi^1} \vec{g}_1^{h\,\mathrm{T}} \\ N_{I,\xi^2} \vec{g}_2^{h\,\mathrm{T}} \\ \sum_{L=1}^{iv} \frac{1}{4}(1 + \xi^{1L}\xi^1)(1 + \xi^{2L}\xi^2) N_{I,\xi^3} \vec{g}_3^{L\,\mathrm{T}} \\ N_{I,\xi^1} \vec{g}_2^{h\,\mathrm{T}} + N_{I,\xi^2} \vec{g}_1^{h\,\mathrm{T}} \\ \frac{1}{2}(1 - \xi^2)\left(N_{I,\xi^1}^B \vec{g}_3^{B\,\mathrm{T}} + N_{I,\xi^3}^B \vec{g}_1^{B\,\mathrm{T}}\right) + \frac{1}{2}(1 + \xi^2)\left(N_{I,\xi^1}^D \vec{g}_3^{D\,\mathrm{T}} + N_{I,\xi^3}^D \vec{g}_1^{D\,\mathrm{T}}\right) \\ \frac{1}{2}(1 - \xi^1)\left(N_{I,\xi^2}^A \vec{g}_3^{A\,\mathrm{T}} + N_{I,\xi^3}^A \vec{g}_2^{A\,\mathrm{T}}\right) + \frac{1}{2}(1 + \xi^1)\left(N_{I,\xi^2}^C \vec{g}_3^{C\,\mathrm{T}} + N_{I,\xi^3}^C \vec{g}_2^{C\,\mathrm{T}}\right) \end{bmatrix} , \tag{42}$$

where N_{I,ξ^i} denotes the partial derivative of the shape function with respect to the curvilinear coordinates. The matrix $\boldsymbol{B}_I^\varphi$ at the node I is determined as

$$\boldsymbol{B}_I^\varphi = \boldsymbol{J}^{-1}[N_{I,\xi^1} \ \ N_{I,\xi^2} \ \ N_{I,\xi^3}] . \tag{43}$$

The physical stresses and dielectric displacements are derived from the potential function and are arranged in the vector $\left[\frac{\partial \Psi}{\partial \vec{\varepsilon}_1}, \ \frac{\partial \Psi}{\partial \vec{\varepsilon}_2}, \ \frac{\partial \Psi}{\partial \vec{\varepsilon}_3}, \ \cdots, \ \frac{\partial \Psi}{\partial \vec{\varepsilon}_9}\right]^{\mathrm{T}}$. Here, $\vec{\varepsilon}_i$ are independently assumed quantities for strain and electric field components and are approximated with the following interpolations, see Klinkel and Wagner [7]:

$$\vec{\varepsilon}_e^h = \boldsymbol{M}_\alpha^1 \vec{\alpha}^1 + \boldsymbol{M}_\alpha^2 \vec{\alpha}^2 \ , \tag{44}$$

with $\boldsymbol{M}_\alpha^1 = \begin{bmatrix} \boldsymbol{N}_E & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{N}_{\vec{E}} \end{bmatrix}$, $\boldsymbol{M}_\alpha^2 = \begin{bmatrix} \boldsymbol{M}_E & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{M}_{\vec{E}} \end{bmatrix}$, $\vec{\alpha}^1 \in \mathbb{R}^{30}$, and $\vec{\alpha}^2 \in \mathbb{R}^{10}$. The matrices $\boldsymbol{N}_E$ and $\boldsymbol{N}_{\vec{E}}$ are given as

$$\boldsymbol{N}_E = (\boldsymbol{Q}_E^0)^{\mathrm{T}} \mathbf{1}, \begin{bmatrix} \xi^3 & \xi^2\xi^3 & 0 & 0 & 0 & \xi^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \xi^3 & \xi^1\xi^3 & 0 & 0 & \xi^1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \xi^1 & \xi^2 & \xi^1\xi^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & \xi^3 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \xi^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \xi^1 \end{bmatrix}, \tag{45}$$

$$\boldsymbol{N}_{\vec{E}} = \boldsymbol{J}_0^{\mathrm{T}} \begin{bmatrix} 1 & 0 & 0 & \xi^2 & \xi^3 & \xi^2\xi^3 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & \xi^1 & \xi^3 & \xi^1\xi^3 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & \xi^1 & \xi^2 & \xi^1\xi^2 \end{bmatrix} . \tag{46}$$

Quantities, which are evaluated at the element center are denoted with the index 0 and $\mathbf{1}$ is a 6×6 identity matrix. The transformation matrix $\boldsymbol{Q}_E$ is obtained by (39) considering $a = 2$ and $b = 1$. The matrices $\boldsymbol{M}_E$ and $\boldsymbol{M}_{\vec{E}}$ are defined as

$$\boldsymbol{M}_E = \frac{\det J_0}{\det J}(\boldsymbol{Q}_S^0)^{-1} \begin{bmatrix} \xi^1 & 0 & 0 & \xi^1\xi^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \xi^2 & 0 & 0 & \xi^1\xi^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \xi^3 & 0 & 0 & \xi^1\xi^2 & \xi^1\xi^2 & \xi^1\xi^2\xi^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \xi^1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \xi^2 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \tag{47}$$

$$M_{\vec{E}} = \frac{\det J_0}{\det J} J_0^{-1} \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ \xi^3 & \xi^1\xi^3 & \xi^2\xi^3 \end{bmatrix}. \tag{48}$$

The approximation of the independent field $\boldsymbol{\sigma}$ is defined as

$$\vec{\boldsymbol{\sigma}}_e^h = M_\beta \vec{\boldsymbol{\beta}}, \tag{49}$$

with $M_\beta = \begin{bmatrix} N_S & \mathbf{0} \\ \mathbf{0} & N_{\vec{D}} \end{bmatrix}$ and $\vec{\boldsymbol{\beta}} \in \mathbb{R}^{30}$.

Here the matrix N_S is equivalent to N_E of (45), where instead of $(Q_E^0)^T$ the transformation matrix $(Q_S^0)^T$ is used. The interpolation $N_{\vec{D}}$ is identical to $N_{\vec{E}}$.

The weak formulation of (37) will be approximated on element level as following:

$$\delta\pi_e^h = \int_{\mathcal{B}_e} \delta\vec{\boldsymbol{\varepsilon}}_G^T \vec{\boldsymbol{\sigma}}_e^h - \delta\vec{\boldsymbol{v}}_e^T \cdot \begin{bmatrix} \vec{a}\rho \\ \varrho \end{bmatrix} dV_e$$

$$- \int_{\partial\mathcal{B}_e} \delta\vec{\boldsymbol{v}}_e^T \cdot \begin{bmatrix} \vec{\bar{\bar{T}}} \\ \bar{\varsigma} \end{bmatrix} dA_e + \int_{\mathcal{B}_e} \delta\vec{\boldsymbol{\sigma}}_e^{hT} \vec{\boldsymbol{\varepsilon}}_G \, dV_e$$

$$+ \int_{\mathcal{B}_e} \delta\vec{\boldsymbol{\varepsilon}}^T \frac{\partial\Psi}{\partial\vec{\varepsilon}} - \delta\vec{\boldsymbol{\varepsilon}}^T \vec{\boldsymbol{\sigma}}_e^h - \delta\vec{\boldsymbol{\sigma}}_e^{hT} \vec{\boldsymbol{\varepsilon}} \, dV_e, \tag{50}$$

with $\delta\vec{\boldsymbol{v}}_e^T = [\delta\vec{\boldsymbol{x}}_e^T \quad \delta\varphi_e]$ and $\vec{\boldsymbol{\sigma}}$ contains the components of $\boldsymbol{S}$, $\boldsymbol{T}$, and $\vec{\boldsymbol{D}}$ accordingly to the vector notation. The weak form is solved iteratively by employing Newton-Raphson's method. This requires the linearization

$$D[\delta\pi_e^h] = \int_{\mathcal{B}_e} \Delta\delta\vec{\boldsymbol{\varepsilon}}_G^T \vec{\boldsymbol{\sigma}}_e^h \, dV_e + \int_{\mathcal{B}_e} \delta\vec{\boldsymbol{\varepsilon}}_G^T \Delta\vec{\boldsymbol{\sigma}}_e^h \, dV_e$$

$$+ \int_{\mathcal{B}_e} \delta\vec{\boldsymbol{\varepsilon}}_e^{hT} \frac{\partial^2\Psi}{\partial\vec{\varepsilon}\partial\vec{\varepsilon}} \Delta\vec{\boldsymbol{\varepsilon}}_e^h - \delta\vec{\boldsymbol{\varepsilon}}_e^{hT} \Delta\vec{\boldsymbol{\sigma}}_e^h - \delta\vec{\boldsymbol{\sigma}}_e^{hT} \Delta\vec{\boldsymbol{\varepsilon}}_e^h \, dV_e$$

$$+ \int_{\mathcal{B}_e} \Delta\vec{\boldsymbol{\sigma}}_e^{hT} \delta\vec{\boldsymbol{\varepsilon}}_G \tag{51}$$

Considering the above interpolations (50) and (51) one obtains the following matrices:

$$A_e^{ij} = \int_{\mathcal{B}_e} (M_\alpha^i)^T \frac{\partial^2\Psi}{\partial\vec{\varepsilon}\partial\vec{\varepsilon}} M_\alpha^j \, dV_e, \tag{52}$$

$$C_e = \int_{\mathcal{B}_e} (M_\alpha^1)^T M_\beta \, dV_e, \tag{53}$$

$$L_e = \int_{\mathcal{B}_e} B^T M_\beta \, dV_e, \tag{54}$$

$$K_e = \int_{\mathcal{B}_e} G \, dV_e, \tag{55}$$

where G is the matrix of [7] and vectors

$$\vec{\boldsymbol{a}}_e^i = \int_{\mathcal{B}_e} (M_\alpha^i)^T \left(\frac{\partial\Psi}{\partial\vec{\varepsilon}} - \vec{\boldsymbol{\sigma}}_e^h \right) dV_e, \tag{56}$$

$$\vec{\boldsymbol{b}}_e = \int_{\mathcal{B}_e} M_\beta^T (\vec{\boldsymbol{\varepsilon}}_G - \vec{\boldsymbol{\varepsilon}}_e^h) \, dV_e, \tag{57}$$

$$\vec{\boldsymbol{f}}_e^{int} = \int_{\mathcal{B}_e} B^T \vec{\boldsymbol{\sigma}}_e^h \, dV_e, \tag{58}$$

$$\vec{\boldsymbol{f}}_e^{ext} = \int_{\mathcal{B}_e} N^T \vec{\tilde{\boldsymbol{p}}} \, dV_e + \int_{\partial\mathcal{B}_e} N^T \vec{\tilde{\boldsymbol{t}}} \, dA_e, \tag{59}$$

with $i = 1,2$ and $j = 1,2$. In Eq. (59) the body and surface loads are determined by $\vec{\tilde{\boldsymbol{p}}}^T = [\rho\vec{\boldsymbol{a}}^T, \quad \varrho]$ and $\vec{\tilde{\boldsymbol{t}}}^T = \left[\vec{\bar{\boldsymbol{T}}}^T, \quad \bar{\varsigma} \right]$. Having in mind that (50) is solved iteratively the following approximation on element level is obtained

$$[\delta\pi + D[\delta\pi] \cdot (\Delta\vec{\boldsymbol{u}}, \Delta\varphi, \Delta\vec{\boldsymbol{\varepsilon}}, \Delta\vec{\boldsymbol{\sigma}})]_e^h$$

$$\Rightarrow \begin{bmatrix} \delta\vec{\boldsymbol{v}}_e \\ \delta\vec{\boldsymbol{\alpha}}_e^1 \\ \delta\vec{\boldsymbol{\alpha}}_e^2 \\ \delta\vec{\boldsymbol{\beta}}_e \end{bmatrix}^T \left(\begin{bmatrix} \vec{\boldsymbol{f}}_e^{int} - \vec{\boldsymbol{f}}_e^{ext} \\ \vec{\boldsymbol{a}}_e^1 \\ \vec{\boldsymbol{a}}_e^2 \\ \vec{\boldsymbol{b}}_e \end{bmatrix} + \begin{bmatrix} K_e & 0 & 0 & L_e \\ 0 & A_e^{11} & A_e^{12} & -C_e \\ 0 & A_e^{21} & A_e^{22} & 0 \\ L_e^T & -C_e^T & 0 & 0 \end{bmatrix} \begin{bmatrix} \Delta\vec{\boldsymbol{v}}_e \\ \Delta\vec{\boldsymbol{\alpha}}_e^1 \\ \Delta\vec{\boldsymbol{\alpha}}_e^2 \\ \Delta\vec{\boldsymbol{\beta}}_e \end{bmatrix} \right).$$

$$\tag{60}$$

Taking into account that the finite element interpolations for the fields $\vec{\boldsymbol{\sigma}}$ and $\vec{\boldsymbol{\varepsilon}}$ are discontinuous across the element boundaries a condensation on element level yields the element stiffness matrix and the right hand side vector

$$K_{Te} = K_e + L_e C_e^{-1} A_e C_e^{-T} L_e^T, \tag{61}$$

$$\vec{\boldsymbol{f}}_e = \vec{\boldsymbol{f}}_e^{ext} - \vec{\boldsymbol{f}}_e^{int} + L_e C_e^{-1} A_e C_e^{-T} \vec{\boldsymbol{b}}_e - L_e C_e^{-1} \vec{\boldsymbol{a}}_e, \tag{62}$$

with $A_e = A_e^{11} - A_e^{12}(A_e^{11})^{-1}A_e^{21}$ and $\vec{\boldsymbol{a}}_e = \vec{\boldsymbol{a}}_e^1 - A_e^{12}(A_e^{11})^{-1}\vec{\boldsymbol{a}}_e^2$. After assembly over all elements $K_T = \bigcup_{e=1}^{nelem} K_{Te}$, $\Delta\vec{\boldsymbol{V}} = \bigcup_{e=1}^{nelem} \Delta\vec{\boldsymbol{v}}_e$ and $\vec{\boldsymbol{F}} = \bigcup_{e=1}^{nelem} \vec{\boldsymbol{f}}_e$ one obtains

$$K_T \Delta\vec{\boldsymbol{V}} = \vec{\boldsymbol{F}}, \tag{63}$$

with the unknown incremental nodal displacements and the electric potential. The update of the internal degrees of freedom reads

$$\Delta\vec{\boldsymbol{\alpha}}_e^1 = C_e^{-T} L_e^T \Delta\vec{\boldsymbol{v}}_e + C_e^{-T} \vec{\boldsymbol{b}}_e, \tag{64}$$

$$\Delta\vec{\boldsymbol{\alpha}}_e^2 = -(A_e^{22})^{-1}\vec{\boldsymbol{a}}_e^2 - (A_e^{22})^{-1}A_e^{21}\Delta\vec{\boldsymbol{\alpha}}_e^1, \tag{65}$$

$$\Delta\vec{\boldsymbol{\beta}}_e = C_e^{-1} A_e \Delta\vec{\boldsymbol{\alpha}}_e^1 + C_e^{-1} \vec{\boldsymbol{a}}_e. \tag{66}$$

VIII. NUMERICAL EXAMPLES

Embedding the solid shell element formulation in a modified version of the program FEAP demonstrates the capturing of the characteristics of the DE material behavior. The ability of the shell element is shown in the following examples.

A. Eigenvalue Problem

To analyze the locking modes of the solid shell element an eigenvalue problem is solved. In this first example a cube form DE material with an edge length of 2cm × 2cm × 2cm is examined. The cube is zero-stress supported. The data for the Ogden-type material are according to [8] $\alpha_1 = 1.293$,

$\alpha_2 = 2.325$, $\alpha_3 = 2.561$, $\mu_1 = 0.00592\text{MPa}$, $\mu_2 = 0.0582\text{MPa}$, $\mu_3 = -0.0161\text{MPa}$ and $\Lambda = 2.49256777\text{MPa}$. The relative permittivity is given as $\varepsilon_r = (1 + \chi) = 4$.

In Table 1 it is shown that there is only eigenvalue number 18 much greater than zero. This eigenvalue accounts for the volumetric locking mode which arises by incompressible materials like DE. At this point no other locking modes are observed. The result is also the same when taking an irregular non-cube element.

TABLE I
EIGENVALUES OF A DIELECTRIC ELASTOMER CUBE

No.	Eigenvalue	No.	Eigenvalue
1	0.11E-02	10	0.50E-01
2	0.31E-02	11	0.62E-01
3	0.72E-02	12	0.75E-01
4	0.11E-01	13	0.85E-01
5	0.18E-01	14	0.87E-01
6	0.20E-01	15	0.97E-01
7	0.21E-01	16	0.10E+00
8	0.27E-01	17	0.10E+00
9	0.34E-01	18	0.57E+01

B. Bending Actuator

The second example is presented to demonstrate the valid reflection of the electro-mechanical coupling phenomenon of DEs. Therefore a square plate with the dimension of 100mm $\times$ 100mm is investigated. It is clamped on one side and consists of two layers with the thickness of 0.5mm each, see Figure 3. The dataset of the material is the same as given in the first example.
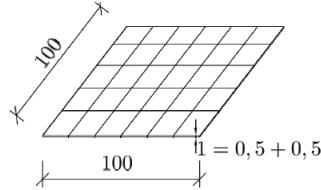
Fig. 3 Dimensions of the bending actuator

To get a bending answer of the thin structure either the upper or lower layer is loaded by an electric field applied in thickness direction of the structure. Due to the electro-mechanical coupling the loaded layer responds with an in-plane extension. Because of the eccentricity the bending effect of the whole structure occurs. Figure 4 shows several deformed configurations.
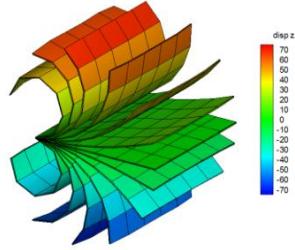
Fig. 4 Bending actuator - several deformed configurations

In the simulation the surface charge is increased from 0V to 7.35×10^6 V and decreases back to 0V again. Then the other layer is loaded in the same way. The tip deflection versus the applied voltage is shown in Figure 5.
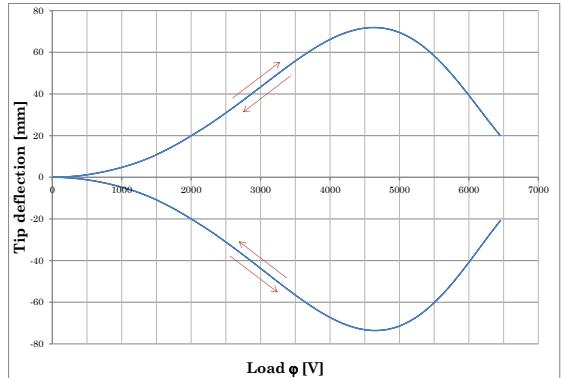
Fig. 5 Tip deflection versus voltage relations

IX. SUMMARY AND OUTLOOK

The presented element formulation is based on a mixed variational approach. It results in an independent interpolation of the displacement, electric potential, strains, electric field, stresses and dielectric displacements. The element possesses only four nodal degrees of freedom, displacements and the electric potential. It allows a consistent approximation of the electromechanical coupled problem. The governing field equations and boundary conditions are presented. The formulation accounts for geometric and material nonlinear behavior. Further tasks will be to embed the material formulation into an improved stabilized element formulation. For more accurate simulations other material models for the mechanical part should be discussed.

REFERENCES

[1] A. Dorfmann and R.W.Ogden, "Nonlinear electroelasticity", Acta Mechanica 174, pp. 167-183, 2005.
[2] D.J. Steigmann, "Equilibrium theory for magnetic elastomers and magnetoelastic membranes", Int. J. Nonlinear Mech. 39(7), pp. 1193-1216, 2004.
[3] D.K. Vu, P. Steinmann and G. Possart, "Numerical modelling of non-linear electroelasticity", Int. J. Numer. Meth. Engng 70, pp.685-704, 2007.
[4] A.C. Eringen and G.A. Maugin, *Electrodynamics of continua I, Foundations and Solid Media*, New York : Springer-Verlag, 1990.
[5] R. Müller, B.X. Xu, D. Gross, M. Lyschik, D. Schrade, S. Klinkel, „Deformable dielectrics – optimization of heterogeneities", International Journal of Engineering Science 48, pp. 647–657, 2010.
[6] S. Klinkel, "A thermodynamic consistent 1d model for ferroelastic and ferroelectric hysteresis effects in piezoceramics", Communications in Numerical Methods in Engineering, 22(7), pp.727-739, 2006.
[7] S. Klinkel and W. Wagner, "A piezoelectric solid shell element based on a mixed variational formulation for geometrically linear and nonlinear applications", Computers and Structures 86, pp. 38-46, 2008.
[8] M.T. Wissler, *Modeling dielectric elastomer actuators*, PhD. Thesis, Swiss Federal Institute of Technology Zürich, 2007.

A Software Language Approach to Derivative Contracts in Finance

Dipl.-Inf. Jean-Marie Gaillourdet

Software Technology Group, Department of Computer Science, University of Kaiserslautern, Germany

Abstract— **Financial derivatives are an important tool of to-day's finance sector. Despite their often negative public perception after the crisis of 2008, they are a tool originally developed to reduce risks in trading real goods. Understanding and evaluating derivatives is an important problem in practice. We present a declarative language of derivative contracts which is independent of pricing models. We also present a denotational semantics, enabling a calculus of contracts. The given denotational semantics enables also the application of abstract interpretation and static analysis techniques developed in the programming language community.**

I. Introduction

Financial derivatives are an important tool of today's finance sector. Despite their often negative public perception after the crisis of 2008, they are a tool — originally — developed to reduce risks in trading real goods such as wheat or copper. But at the same time financial derivatives introduce possibilities to speculate over the value of something without owning the actual something. Both uses lead eventually to their important role in banking and investment. Naturally, financial derivatives are a well studied topic in economics and mathematics, e.g. [1], [2]. Both mathematics and economics often use extensive software systems to analyze financial derivatives. Therefore, the apparent lack of intensive research in informatics seems surprising, notable exceptions are: [3], [4], [5], [6], [7].

But the interest to investigate financial contracts from a language perspective seems to be growing. In 2010 the U.S. Securities and Exchange Commission published a "concept release" with a request for comments, which included a proposal to require the publication of Python programs which represent the contract of Asset Backed Securities, a special kind of financial derivative contracts.

But what are financial derivatives? At first they are contracts, i.e. legal documents defining rights and obligations of two or more parties. Second, they are derived, i.e. the rights and obligations defined by the contract depend on some external variables. We call these external variables: *Observables*. They differ from values defined in a contract by their ability to change over time. E.g. possible observables are the temperature at Kaiserslautern or the course of some shares at some stock exchange. This notion of observables was already used in [3].

Henceforth, financial derivative contracts are contracts which are based upon observables describing rights and obligations in terms of payed money from one party to another.

We will further restrict this notion to contracts between two parties. One of them is the holder, he becomes the holder by acquiring the derivative contract. The other party remains anonymous as we will always analyze contracts from the view point of the holder.

The last restriction we place on the notion of derivatives, is that the contracts under consideration may grant the right to choose between alternatives only to the holder of the contract.

The contributions of this article are:

1) An overview of work done in informatics, especially from a software language point of view, on financial derivatives.
2) A formal definition of a language — based on [3] — which is capable of expressing financial derivatives themselves and not some derived notion.
3) A denotational semantics of this language, which enables to a notion of equality of derivative contracts, which strives to model the ideal of a contract and not its syntactic representation.

Section II covers the sparse related work in informatics. Section III defines our language of derivative contracts, which is an extension of the language of [3]. Section IV presents our denotational semantics of contracts and an equality notion of contracts. In Section V we will discuss shortly possible applications of and future work on our language and semantics.

II. Related Work

The earliest attempt at defining a software language for financial products, as far as we know, is a joint academic and industrial project resulting in the definition and implementation of RISLA. It is based on the insight, "that a product can be characterized by describing its cash flows." [4, p. 2]. Its purpose was to simplify the integration of new financial products into the existing business software systems of banks. Therefore, the project developed a compiler from RISLA to COBOL, which generated the necessary routines to integrate a financial product into the existing software infrastructure. The focus on integrating financial products into business software systems lead to a language design containing information about the expected user interface in the business software system. RISLA was not used to analyze the contract or perform pricing[1] on it. From the very sparse publicized documentation of it, it is not clear whether choices of the holder of a derivative are expressible or not.

---

[1]Pricing is a mathematical analysis which defines a notion of a *fair price*, see e.g. [2] for some recent results on pricing options.

In 2000, Peyton-Jones and Eber [3] presented a domain-specific language to declare derivatives in Haskell. Their language was implemented as combinator library in Haskell. They sketched a *valuation semantics* of derivatives and gave some hints on their implementation of a generic pricing mechanism. The valuation semantics of [3] is tied to the pricing or valuation approach of [1].

In [5] Mogensen introduces a language for cashflow reengineering, which was adopted by a non-disclosed "major Danish bank". The author presents a language, which is similar to data flow languages in programming, and a linear type system, which ensures that every amount of cash is spent only once.

Andersen, et al. present in [6] a language for compositional specification of contracts. That work is also based on [3]. Their focus is on specifying the exchange of resources between any number of involved parties. While our work represents every resource by its monetary value. Additionally, they give denotational and operational semantics for that language, which allows e.g. to decide whether a trace of steps in the real world conforms to a specified contract or not. In comparison to our language, this language is broader as it allows not only to specify derivative contracts, but general contracts, e.g. as they occur when you buy goods in store which provides the right to return the good under certain circumstances.

In [7] Reitz and Nögel present their work in the COMDECO project. That includes an XML based language for derivative contracts. They argue for the use of active documents instead of other approaches to develop applications, which perform pricing of derivatives.

III. A LANGUAGE OF DERIVATIVE CONTRACTS

A. Design Goals

The language for contracts we are going to present has been designed to enable standard mathematical analysis, like option valuation, on contracts written in that language. But at the same time, we did not want to include any assumptions which are introduced by a certain approach of analysis. The language should also allow to resolve any ambiguities about contracts, such might occur e.g. in a text document describing a contract, but must not be possible in this language.

Such a formally defined language should also be amenable to a mathematical treatment of contracts. Such a treatment should include a notion of equality of contracts, that corresponds to our intuition about contracts in real life. Nevertheless, it has to be a mathematical equality, i.e. enabling the substitution of equals and the formulation of mathematical laws on derivative contracts.

Last but not least, a language of contracts should be composable and support abstraction, i.e. support the naming of recurring patterns of contracts. The goal to support composition and abstraction is compatible with the design, we present later in this section, but we will not discuss it in this paper. Instead we will focus on the goals mentioned before.

B. Examples

Before we come to the language definition of derivative contracts, we will have a look at several examples of derivatives specified in it.

1) Zero Coupon Discount Bond: A *zero coupon discount bond* is a contract granting the holder the right to receive an amount of money k at a specified time t in the future. This example has been taken from [3]. This is not a *derivative* contract as it does not depend on external unknown variables.

When (At t) (Scale (Const k) One)

The meaning of this contract is the following. At time t, the holder acquires the contract Scale (Const k) One). Which is a contract granting the right to immediately receive k units of money. For a more detailed description of the keywords, see Section III-C.

2) European Put Option: A *european put option* is a derivative contract granting the holder the right but not the obligation to sell one item of e.g. a share of Siemens at a previously fixed price, here 100, at a predefined time, in this example 5.

When (At 5)
  (Or
    (Scale (Lift2 (−) (Unknown "Siemens") (Const 100)))
    Zero)

3) Barriers: The previous derivative contains the possibility to achieve very high pay-offs when the value of a Siemens share drops to almost zero. Therefore, it is common to include, barriers into derivatives. E.g. a derivative contract could state, when the value of one Siemens share gets below 50, the complete option is void. This is easily expressed with Until:

Until (Comp < (Unknown "Siemens") (Const 50))
  (When (At 5)
    (Or
      (Scale (Lift2 (−) (Unknown "Siemens") (Const 100)))
      Zero))

4) History dependent observables: So called asian variants of options use the mean of price of a share to determine the payoff. In order to construct an observable expression which represents the average of some observable, we have to add the values of observable at different times. In order to allow this, our language contains the constructor Reduce.

Lift2 /
  (Reduce + (Between (5,10)) (Unknown "Siemens"))
  (Const 5)

The observable above computes the average price of one Siemens share between time 5 and 10. Reduce collects all values of the third argument at times at which the second argument evaluates to true. These values are summed — the first argument. The initial value is the left neutral element of the given operation. The operation here is addition, therefore it is zero in this example.

These history dependent observables are not expressible in the language in [3].

$$
\begin{aligned}
m &\in \{<, >, >=, <=, ==, \neq\} \\
s &\in \{+, -, *, \min, \max\} \\
u &\in \{-, abs, \log, ...\} \\
b &\in \{+, -, *, /\} \\
o &::= \quad \text{Unknown n} \\
&\quad | \quad \text{Const k} \\
&\quad | \quad \text{Time t} \\
&\quad | \quad \text{Lift u o} \\
&\quad | \quad \text{D2C o} \\
&\quad | \quad \text{C2D o} \\
&\quad | \quad \text{Lift2 b } o_1\ o_2 \\
&\quad | \quad \text{Comp m } o_1\ o_2 \\
&\quad | \quad \text{At t} \\
&\quad | \quad \text{Between } (t_1, t_2) \\
&\quad | \quad \text{Reduce s } o_1\ o_2
\end{aligned}
$$

Fig. 1.   Syntax of observables

C. Syntax Definition and Language Description

Our language for derivatives is based on the language in [3]. Their language is given as a set of Haskell functions with type signatures. [2] The language itself remains very similar with the exception of some simplifications and one addition, which we will discuss later in this section.

As usual $\mathcal{N}$ denotes the natural numbers including 0, $\mathcal{Z}$ denotes the positive and negative integers, $\mathcal{R}$ denotes the real numbers and $\mathcal{B}$ denotes the set of boolean values. We use $\mathcal{V} = \mathcal{Z} \cup \mathcal{R} \cup \mathcal{B}$ as the values of observables. Every time t is element of $\mathcal{N}$. And $\mathcal{S}$ denotes the set of names of unknown observables. These names are strings. We use the following convention for variable names: $t \in \mathcal{N}$, $k, v \in \mathcal{V}$, and $n \in \mathcal{S}$.

Fig. 1 contains the syntax definition of observable expressions in a style which is common in the programming language community. The single elements of the definitions are either mathematical statements $m \in \{...\}$ or are algebraic data type declaration. In the following, o denotes, depending on the context, the set of all observable expressions, or one concrete expression.

- Unknown n is an observable representing an external variable varying over time. In this work we will restrict the language to unknown observables with $\mathcal{R}$ as value domain.
- Const k is an observable of value $k \in \mathcal{V}$ for all times.
- Time is an observable with domain $\mathcal{N}$. Its value is t at every time t.
- Lift u o is the application of a unary operator u on o.
- Lift2 b o_1 o_2 is the application of a binary operator b on o_1 and o_2.

[2] Although, we don't present the language as an Haskell-embedded domain specific language our implementation is done as an Haskell-embedded domain specific language. We deliberately choose not to discuss such an embedding and the resulting benefits in this paper. Such benefits would include added expressiveness, modularization and abstraction capabilities, and a simplified implementation.

- Comp m o_1 o_2 is the comparison of o_1 and o_2 with the comparison operator m at every point in time.
- C2D o is the conversion of a continuous-valued observable o, i.e. an observable with domain $\mathcal{R}$, to a discrete-valued observable, i.e. an observable with domain $\mathcal{N}$.
- D2C o is the conversion of a discrete-valued observable o, i.e. an observable with domain $\mathcal{N}$, to a continuous-valued observable, i.e. an observable with domain $\mathcal{R}$.
- Reduce s o_1 o_2 is the fold with the binary function s over the sequence of values of o_2 at, when o_1 evaluates to true. The start value of the fold is the left neutral element of the function s.
- At t is syntactic sugar for: Comp (==) Time (Const t).
- Between (t_1, t_2) is syntactic sugar for:

  Lift2 (&&)
      (Comp ($<=$) (Const t_1) Time)
      (Comp ($<$) Time (Const t_2))

Fig. 2 defines the syntax of contracts. We'll give an informal description of the meaning of the constructs here:

- Zero is the contract which grants no rights and no obligations.
- One is the contract which grants the holder the right to receive one unit of money at the time of acquisition.
- Scale o c is a contract which is the same as c with every payment multiplied by the value of the observable o at the time of the payment.
- And c_1 c_2 is the contract which grants all rights and obligations of both c_1 and c_2 at the same time.
- Or c_1 c_2 is the contract which grants the holder the right to choose at time of acquisition to acquire either c_1 or c_2.
- Cond o c_1 c_2 is the contract which grants depending on the value of the boolean observable o at time of acquisition either all rights and obligations of c_1, if o is true, or c_2, if o is false.
- When o c is the contract which acquires the contract c at the first time the boolean observable o becomes true.
- Anytime o c is the contract which grants the holder at every time the boolean observable o is true the right to decide whether he wants to acquire c and release the complete contract Anytime o c or not.
- Until o c is the contract which grants the holder the right to choose between keeping his contract or acquiring c and releasing the complete contract. But this right is only granted as long as the boolean observable o is false. When o becomes true the whole contract becomes equivalent to Zero identical.

The language for derivative contracts is typed. But, because it has a completely standard type system, we omit a formal definition. We describe it only informally:

Every observable expression o has a one of three types: discrete, continuous, and boolean. Their value sets are $\mathcal{Z}$, $\mathcal{R}$, and $\mathcal{B}$. All unary operators take arguments of one type and

$$
\begin{array}{rcl}
c & ::= & \text{Zero} \\
& | & \text{One} \\
& | & \text{And } c_1\ c_2 \\
& | & \text{Or } c_1\ c_2 \\
& | & \text{Cond } c\ c_1\ c_2 \\
& | & \text{Scale o c} \\
& | & \text{When o c} \\
& | & \text{Anytime o c} \\
& | & \text{Until o c}
\end{array}
$$

Fig. 2.   Syntax of contracts

return the same type. All binary and comparison operators take two arguments of the same type. Binary operators return the same type as their arguments and comparison operators always return boolean. C2D and D2C are the only ways to convert or cast observables. Contracts c have no type, the observable arguments of the contract constructors are of type continuous, if not given otherwise in the description of the contract constructors above.

IV. Semantics of Contracts

A. Introduction

In this section, we'll give a formal denotational semantics for our language of contracts, for an overview on denotational semantics see e.g. [8]. A formal semantics allows to study what it means to execute a contract, or in more natural terms: to fulfill a contract. By establishing formal semantics and removing all ambiguities — which are all too often present in natural language texts —, we gain insight into the contract itself. We can answer questions like: Is a sequence of payments between the holder and its contract counter party consistent with a certain contract? Is, for a given contract, every consistent sequence of payments between two contract parties of finite length? What is the maximal length of such a sequence? Is the holder able to receive a payment at all under a given contract? And so forth.

Semantics of formal language haven been studied for a long time. The largest fraction of the literature on this topic probably covers semantics of formally specified logics and programming languages. Applying these techniques to other kinds of languages is not new but certainly not widespread. In the context of contract languages, we are aware of only one previous article which gave formal denotational semantics for a contract language: [6].

We choose to give a denotational semantics, because a denotational semantics provides a straight forward notion of equality, and because it is quite naturally defined for our derivative contracts.

B. Denotational Semantics of Observables

A denotation of unknown observables is a function from time to values, $\mathcal{N} \to \mathcal{V}$. Therefore, an environment, i.e. a mapping of names to their denotation, is a function from $e : \mathcal{S} \to \mathcal{N} \to \mathcal{V}$. An environment represents the values of all

unknown observables of a contract for all times. Therefore, the denotational semantics of observable expression is a function:

$$
[\![\cdot]\!] : o \to (\mathcal{S} \to \mathcal{N} \to \mathcal{V}) \to (\mathcal{N} \to \mathcal{V})
$$

We write $[\![o]\!]_e$ to denote the semantics of an observable expression o applied to an environment e. With an additional time argument t we write $[\![o]\!]_e^t$.

$[\![\cdot]\!]$ is defined as follows[3]:

$$
\begin{array}{rcl}
[\![\text{Unknown } n]\!]_e^t & = & e\ n\ t \\
[\![\text{Const } k]\!]_e^t & = & k \\
[\![\text{Time}]\!]_e^t & = & t \\
[\![\text{Lift u o}]\!]_e^t & = & u\ [\![o]\!]_e^t \\
[\![\text{D2C o}]\!]_e^t & = & [\![o]\!]_e^t \\
[\![\text{C2D o}]\!]_e^t & = & \lfloor [\![o]\!]_e^t \rfloor \\
[\![\text{Lift2 } b\ o_1\ o_2]\!]_e^t & = & [\![o]\!]_e^t\ b\ [\![o]\!]_e^t \\
[\![\text{Comp } m\ o_1\ o_2]\!]_e^t & = & [\![o]\!]_e^t\ m\ [\![o]\!]_e^t \\
\underbrace{[\![\text{Reduce } s\ o_1 o_2]\!]_e^t}_{d} & = & \left\{ \begin{array}{ll} init(s) & \text{if } t < 0 \\ [\![d]\!]_e^{t-1} & \text{if } [\![o_1]\!]_e^t = false \\ [\![d]\!]_e^{t-1}\ s\ [\![o_2]\!]_e^t & \text{if } [\![o_1]\!]_e^t = true \end{array} \right.
\end{array}
$$

$init(s)$ denotes the left neutral element of s.

C. Denotational Semantics of Contracts

In the following the notation $\widehat{x}$ will denote a set of x. We define the domain $\mathcal{D}$ of the denotation of contracts as the largest fixed point of the following recursive algebraic data type equation.

$$
p \quad = \quad o \vartriangleleft \widehat{p}
$$

$\vartriangleleft$ is the only constructor of this data type, it has two arguments, the first is an element of $\mathcal{V}$, the second is a finite set of application of this constructor. Therefore, this data type co-inductively defines the domain of all denotation of contracts. It is a forest of infinite height n-ary trees which are marked with elements of $\mathcal{V}$. The depths of a node of the forest is the time at which the mark of the node is payed to the holder of a contract. E.g. $\{17 \vartriangleleft 0 \vartriangleleft \dots \}$ is the denotation of a contract which grants the holder the right to receive a payment of 17 money units at time 0 and a payment of zero money units at time 1.

The branching of the trees denotes the choices the *holder* of the contract has.

The empty denotation which consists only of payments of 0 units of money is abbreviated as $\bullet$. It is largest fixpoint of the following equation:

$$
\bullet \quad = \quad \{0 \vartriangleleft \bullet\}
$$

We need the following auxiliary functions on $\mathcal{D}$, before we are able to define the denotational semantics.

$$
\begin{array}{rcl}
\widehat{p} \times \widehat{q} & = & \left\{ (p+q) \vartriangleleft (\widehat{p'} \times \widehat{q'}) \middle| (p \vartriangleleft \widehat{p'}) \in \widehat{p},\ (q \vartriangleleft \widehat{q'}) \in \widehat{q} \right\} \\
o *_e^t \widehat{p} & = & \left\{ ([\![o]\!]_e^t * v) \vartriangleleft (o *_e^{t+1} \widehat{p'}) \middle| (v' \vartriangleleft \widehat{p'}) \in \widehat{p} \right\}
\end{array}
$$

$\times$ adds two elements of $\mathcal{D}$ such that both arguments are combined into one, as if they were both executed.

---

[3]We use a curried function application notation which does not use parenthesis

$*_e^t$ multiplies every payment of an element of $\mathcal{D}$ with the current value of an observable expression.

Now, we are able to define the denotation of contracts, which is the following function:

$$\llbracket c \rrbracket : (\mathcal{S} \to \mathcal{N} \to \mathcal{V}) \to \mathcal{N} \to \mathcal{D}$$

$$
\begin{aligned}
\llbracket \text{Zero} \rrbracket_e^t &= \bullet \\
\llbracket \text{One} \rrbracket_e^t &= \{1 \lhd \bullet\} \\
\llbracket \text{And a b} \rrbracket_e^t &= \llbracket a \rrbracket_e^t \times \llbracket b \rrbracket_e^t \\
\llbracket \text{Or a b} \rrbracket_e^t &= \llbracket a \rrbracket_e^t \cup \llbracket b \rrbracket_e^t \\
\llbracket \text{Scale o c} \rrbracket_e^t &= o *_e^t \llbracket c \rrbracket_e^t \\
\llbracket \text{Cond c a b} \rrbracket_e^t &= \begin{cases} \llbracket b \rrbracket_e^t & \text{if } \llbracket c \rrbracket_e^t = false \\ \llbracket a \rrbracket_e^t & \text{if } \llbracket c \rrbracket_e^t = true \end{cases} \\
\underbrace{\llbracket \text{When b c} \rrbracket_e^t}_{d} &= \begin{cases} \{0 \lhd \llbracket d \rrbracket_e^{t+1}\} & \text{if } \llbracket b \rrbracket_e^t = false \\ \llbracket c \rrbracket_e^t & \text{if } \llbracket b \rrbracket_e^t = true \end{cases} \\
\underbrace{\llbracket \text{Anytime b c} \rrbracket_e^t}_{d} &= \begin{cases} \{0 \lhd \llbracket d \rrbracket_e^{t+1}\} & \text{if } \llbracket b \rrbracket_e^t = false \\ \llbracket c \rrbracket_e^t \cup \{0 \lhd \llbracket d \rrbracket_e^{t+1}\} & \text{if } \llbracket b \rrbracket_e^t = true \end{cases} \\
\underbrace{\llbracket \text{Until b c} \rrbracket_e^t}_{d} &= \begin{cases} \llbracket c \rrbracket_e^t \cup \{0 \lhd \llbracket d \rrbracket_e^{t+1}\} & \text{if } \llbracket b \rrbracket_e^t = false \\ \bullet & \text{if } \llbracket b \rrbracket_e^t = true \end{cases}
\end{aligned}
$$

When we return to the example of Sec. III-B.2, we can now determine the following: Let x be the price of one Siemens share at time 5 as determined by e, i.e. $\llbracket \text{Unknown "Siemens"} \rrbracket_e^5 = x$, and let c be the referenced example contract, then

$$\llbracket c \rrbracket_e^t = \{0 \lhd \{0 \lhd \{0 \lhd \{0 \lhd \{0 \lhd (\bullet \cup \{(x-100) \lhd \bullet\})\}\}\}\}\}$$

The denotation tells us that this European put option is a contract, which can be described as a sequence of 5 payments of zero amounts of money, followed either by an infinite sequence of zero-valued payments or by a payment of value $x - 100$ followed by an infinite sequence of zero-valued payments.

D. Derived Notions

We define the equality of two contracts a and b, now as follows:

$$a = b \text{ if and only if } \forall e. \llbracket a \rrbracket_e^0 = \llbracket b \rrbracket_e^0$$

Now, we have a notion of equality which is not tied to the syntax of contracts or observables, instead we claim it is tied to the intended meaning, we associate with a contract. Yet, we have a mathematical definition of this equality notion.

Most contracts in real life have a maximal life time. The same is true for derivative contracts in finance. We are going to define a notion of a horizon, which will be the earliest time after which nothing interesting may happen.

The smallest time t for which the following equation holds is called the horizon of a contract c.

$$\forall t'. t' > t \implies \forall e. \exists \widehat{p}_{i \in \{1 \ldots n\}} . \llbracket c \rrbracket_e^{t'} = \{0 \lhd \widehat{p_1}, \ldots, 0 \lhd \widehat{p_n}\}$$

Note: Such a horizon t doesn't have to exist for every contract c, while most practical contracts do have a defined horizon. The following contract does not have a defined horizon, because there is no time at which we can guaranty, that the price of a Siemens share has been at least once smaller than the price of a Daimler share.

When
(Comp < (Unknown "Siemens") (Unknown "Daimler"))
One

V. APPLICATIONS, FUTURE WORK AND CONCLUSIONS

As we said in Sec. III-A, we wanted to design a language which can be used to do option pricing, but which is not tied to one approach. So far, we have implemented option pricing for a subset with help of Qian Liang and Stefanie Müller based on [3],[2], and [1]. We plan to add Monte-Carlo based option pricing to our implementation.

Our language and denotational semantics could also be used to apply abstract interpretation or some specialized variants of it to derivative contracts. That could allow to compute worst or best case scenarios. Or it could allow to determine whether there is horizon, at which times decision have to be made by the holder, and much more.

The semantics itself allows to develop a calculus of contracts, which allows to transform contracts in order to simplify other analysises for example.

We have designed a language to express financial derivatives, which is more general than the language of [3] and more specific than the language of [6]. We believe this is a useful middle ground to work with derivatives and not with more general contracts. We have given a formal semantics which captures the intuitive meaning in a formal definition, which allows to apply the substitution principle. We don't have to resort to bisimulation of processes or similar approaches in order to show the equivalence of two contracts.

REFERENCES

[1] J. C. Cox, S. A. Ross, and M. Rubinstein, "Option pricing: A simplified approach," *Journal of Financial Economics*, vol. 7, no. 3, pp. 229–263, 1979.

[2] S. Müller, "The binomial approach to option valuation." [Online]. Available: http://kluedo.ub.uni-kl.de/volltexte/2010/2462/

[3] S. Peyton Jones, J.-M. Eber, and J. Seward, "Composing contracts: an adventure in financial engineering (functional pearl)," in *ICFP '00: Proceedings of the fifth ACM SIGPLAN international conference on Functional programming*. New York, NY, USA: ACM, 2000, pp. 280–292.

[4] B. Arnold, A. V. Deursen, and M. Res, "An algebraic specification of a language for describing financial products," in *ICSE-17 Workshop on Formal Methods Application in Software Engineering*. IEEE, 1995, pp. 6–13.

[5] T.Æ. Mogensen, "Linear types for cashflow reengineering," pp. 823–845, 2003.

[6] J. Andersen, E. Elsborg, F. Henglein, J. G. Simonsen, and C. Stefansen, "Compositional specification of commercial contracts," *International Journal on Software Tools for Technology Transfer (STTT)*, vol. 8, no. 6, pp. 485–516, November 2006.

[7] M. Reitz and U. Nögel, "Components: A valuable investment for financial engineering," in *PPPJ '06: Proceedings of the 4th international symposium on Principles and practice of programming in Java*. New York, NY, USA: ACM, 2006, pp. 153–162.

[8] D. A. Schmidt, *Denotational Semantics*. Allyn And Bacon, Inc., 1986.

An Inverse Problem Method for RDC Simulation

Hanin B. Jildeh[#1], Menwer Attarakih[*2], Hans-Jörg Bart[#3]

[#] *Chair of Separation Science and Technology, TU Kaiserslautern, Postfach 3049 - 67653 Kaiserslautern, Germany*

[*] *Faculty of Engineering Technology, Chemical Engineering Department, Al-Balqa Applied University, Postfach 15008 - 11134 Amman, Jordan*

[1]`hanin.jildeh@mv.uni-kl.de`
[2]`attarakih@yahoo.com`
[3]`bart@mv.uni-kl.de`

Abstract— **The hydrodynamic and mass transfer behavior of a Rotating Disk Contactor (RDC) extraction column is investigated for two different liquid-liquid systems recommended by the European Federation of Chemical Engineering (EFCE). An inverse problems method is applied to estimate the coalescence parameters in a RDC extraction column. Single-droplet studies in a small lab scale RDC is used to evaluate the coalescence parameters necessary for column simulations, which were obtained by an inverse solution of the population balance model using the generalized fixed-pivot technique for the discretization of the droplet internal coordinate. The coalescence parameters resulted from solving the inverse problem are dependent on the chemical test system used for the simulation. Then the resulted coalescence parameters values were used in LLECMOD to study the hydrodynamic and mass transfer behavior of pilot plant of RDC extraction column. The simulated Sauter mean droplet diameter, hold-up values and concentration profiles for organic and aqueous phase were found to be well predicted compared to the experimental data.**

I. INTRODUCTION

Liquid-liquid extraction is a powerful separation technique after distillation and it is widely used in many industrial fields such as petrochemical and biochemical industries. Liquid-liquid extraction columns are classified into three main categories: stirred columns (Rotating Disc Contactor (RDC) and Kühni columns), non-agitated columns and pulsed columns (sieve plate and packed columns).

Liquid–liquid extraction column (LLEC) simulation based on population balance modeling has difficulties due to contiguous change in the droplet properties such as Sauter mean droplet diameter, hold-up values and concentration profiles for organic and aqueous phase. The most common mechanisms effects on the population balance model are breakage, coalescence and mass transfer [1].

Inverse problem is an ill-posed in general and need some stabilization techniques to get reliable optimized parameters. Solving it is a necessary part for reliable modelling strategy. However, this is complicated due to the increase in the size of the differential-algebraic system when dealing with inverse problems using population balance models.

The objective of this paper is to evaluate the coalescence parameter in a small lab scale RDC for two EFCE systems as is butylacetate-water and toluene-water. We then use these estimated constants to predict the Sauter mean droplet diameter, holdup and concentration profiles for a pilot plant scale column.

II. THE MATHEMATICAL MODEL

The general spatially distributed population balance equation (SDPBE) for describing the coupled hydrodynamics and mass transfer in liquid-liquid extraction columns (LLECs) in a one spatial domain could be written as [2]:

$$\frac{\partial f_{d,c_y}(\psi)}{\partial t} + \frac{\partial\left[u_y f_{d,c_y}(\psi)\right]}{\partial z} + \frac{\partial\left[\dot{c}_y f_{d,c_y}(\psi)\right]}{\partial c_y} =$$
$$\frac{\partial}{\partial z}\left[D_y \frac{f_{d,c_y}(\psi)}{\partial z}\right] + \frac{Q_y^{in}}{A_c} f_y^{in}(d,c_y;t)\delta(z - z_y) + \Upsilon\{\psi\}$$

In this equation the components of the vector $\psi = [d\ c_y\ z\ t]$ are those for the droplet internal coordinates (diameter and solute concentration), the external coordinate is (z) and for time is (t). The velocity vector along the internal coordinates is given by (c_y) is ($\dot{c}_y$) and (Υ) represent the source term. (ζ) represent the net number of droplets produced by breakage and coalescence per unit volume and unit time in the coordinates range [ζ, $\zeta+\partial\zeta$]. The dispersed phase velocity (u_y) relative to the walls of the column is determined in terms of the slip velocity (u_s) with respect to the continuous phase velocity (u_x) with respect to the walls of the column [3], [4].

The source term is given by the following equation:

$$\underbrace{\Upsilon\{\psi\}}_{Source\ Term} = \underbrace{B^b(d,c_y;t,z)}_{Birth\ by\ Breakage} - \underbrace{D^b(d,c_y;t,z)}_{Death\ by\ Breakage}$$
$$+ \underbrace{B^c(d,c_y;t,z)}_{Birth\ by\ Coalescence} - \underbrace{D^c(d,c_y;t,z)}_{Death\ by\ Coalescence}$$

The bivariate source term for the four rates of droplets birth and death due to droplet breakage and coalescence are quite complicated they are expressed in details [2].

Simulating liquid-liquid extraction columns is a challenging task due to the discrete character of the dispersed phase. In technical geometries the population balance model has no general analytical solution and hence a special numerical technique is required to solve the PBE [4]. Several numerical approach are proposed to solve the PBE which are

classified into two categories namely classes method Kumar and Ramkrishna [5] and method of moments [6]. Attarakih has invented a new technique for the solution of the population balance equation using the sectional quadrature method of moments (SQMOM) which combines the advantages of classes method and method of moments [7].

III. INVERSE PROBLEM APPROACH

Inverse problems for population balances have attracted many researchers in the recent years [3], [8]–[10]. Inverse problem is considered an ill-posed in general and need some stabilization techniques to get reliable optimized parameters. In more general cases a mathematical programming procedure has to be applied in order to find the best fit of the experimental data. Since the inverse problems are highly sensitive to errors in the experimental data, the experimental data has been fitted to a normal distribution to give a smother curve and remove some noise, by regularization.

By solving the population balance equations the optimization algorithm derives the unknown coalescence parameters by fitting the simulated outlet distribution to the experimental one using an objective function.

In this work an example of the inverse problem program has been used to obtain the coalescence parameters of a lab scale RDC for two cases of two EFCE systems as is butylacetate-water and toluene-water.

IV. ESTIMATION PARAMETERS

Despite the rich literature on population balance equation modeling, there are few studies on optimization packages for coalescence models that helps to choose the best coalescence model for the experimental data. These optimization packages must fully describe the system by solving the population balance problem inversely to get the best values of constants in the coalescence model. In addition the objective function should be wisely chosen to minimize the error and to have a unique solution for a given data. Since droplet coalescence is highly sensitive to the hydrodynamics and physic-chemical properties, interfacial dynamics and mass transfer [3].

Coulaloglou & Tavlarides (1977) developed a model for stirred vessel, which is based on the kinetic theory of gases and drainage film theory; and calculate the coalescence frequency from the collision rate (frequency) (h) and the coalescence efficiency (λ) [11].

$$\omega(d,d',\phi_d) = h(d,d',\phi_d) \ \lambda(d,d',\phi_d)$$

The Coulaloglou & Tavlarides coalescence model is given by the following equation [12]:

$$\omega(d,d',\phi_d) = \left[c_1 \frac{\varepsilon^{1/3}}{1+\phi}(d+d')^2(d^{2/3}+d'^{2/3})^{1/2} \right] \times$$

$$\left[\exp\left(-\frac{c_2\eta_x\rho_x\varepsilon}{\sigma^2(1+\phi)^3} \right) \left(\frac{(dd')}{(d+d')} \right)^4 \right]$$

An alternative to this model is developed by Sovova (1981) introducing a different formula for the efficiency of collisions, the mechanism of coalescence of drops by film drainage is replaced by a mechanism based on the effect of collision impact [13]:

$$\omega(d,d',\phi_d) = \left[c_3 \frac{\varepsilon^{1/3}}{1+\phi}(d+d')^2(d^{2/3}+d'^{2/3})^{1/2} \right] \times$$

$$\left[\exp\left(-\frac{c_4\eta_x\rho_x\varepsilon}{\sigma^2(1+\phi)^3} \right) \left(\frac{(d^2+d'^2)\times(d^3+d'^3)}{(dd')\times(d^{2/3}+d'^{2/3})} \right) \right]$$

The coalescence frequency model proposed by Laso (1986) is expressed by the following equation [14]:

$$\omega(d,d',\phi_d) = c_7 \left(\frac{\pi d^3}{6D_R^3} \right)^{-0.49} We^{-0.51}\phi_d^{0.9}Oh^{-0.05}$$

A semi-empirical model proposed by Casamatta and Vogelpohl in (1985) in which the first term at the right hand of the equation is the coalescence frequency caused by random motion, which is assumed to be proportional to the drop volumes. The second term represents the coalescence induced by different rise velocities are expressed by the following equation [11], [15]:

$$\omega(d,d',\phi_d) = \left[c_5 \frac{\pi^2}{36}(d\times d')^3 \right] + \left[c_6 \frac{\pi^2}{6}\left(\frac{d+d'}{2} \right)^2 |d^3-d'^3| \right]$$

The inverse problem program can estimate the constants c_1……c_7 of the previous coalescence models along with the ability to choose one of the breakage models available to run the simulation or with no breakage mode. Additional results will be discussed in further publications.

V. LLECMOD PROGRAM

The mathematical model described above was programmed using Visual Digital FORTRAN and to facilitate the data input and output a LLECMOD program was designed. A graphical user interface of LLECMOD program (Liquid-Liquid Extraction Column MODule) simulates liquid-liquid extraction columns based on the population balance model to predict the column performance at steady state and transient conditions. Furthermore the design of LLECMOD is flexible in such a way it enables the user to define the input operating condition, the coalescence and breakage parameters and other different options according to the simulated case studied. Also LLECMOD enables to input the droplet terminal velocity, energy dissipation, axial dispersion, breakage and coalescence frequencies and the other internal geometrical details of the column.

LLECMOD is considered a user-friendly and flexible windows-based program that contains the main input window and other sub-windows for parameters and correlations input. Using LLECMOD simulations can now be carried out successfully for different types of extraction columns

including agitated (Rotating Disc Contactor (RDC) and Kühni columns) and non-agitated or pulsed columns (sieve plate and packed columns) with different internal geometry.

VI. RESULTS AND DISCUSSION

In this study we simulate two cases of two phase liquid-liquid system for butylacetate-water (B-W) and toluene-water (T-W) in a lab scale RDC [16]. The RDC extraction column has five compartments, its internal diameter is 150 mm, column height 150 mm, compartment height is 30 mm, internal stator diameter is 105 mm, rotator diameter 90 mm and diameter of the rotating shaft is 54 mm. The angular velocity of the shafts is constant 300 rpm and volumetric flow rate for continuous and dispersed phase is $Q_c = Q_d = 100$ l/hr for both systems.

The values of coalescence parameters for Coulaloglou and Tavlarides coalescence model (c_1 and c_2) were estimated using the inverse problem for both systems shown in Table 1.

TABLE I
COALESCENCE PARAMETERS FOR COULALOGLOU AND TAVLARIDES

Chemical System	c_1 (-)	c_2 (m^{-2})	Objective Function
B-W	8.19×10^{-02}	$2.37\times10^{+11}$	1.76×10^{-04}
T-W	2.17×10^{-02}	$1.19\times10^{+11}$	1.69×10^{-04}

The inlet and outlet cumulative volume distribution are used from the work of Simon [16]. The unknown coalescence parameters values were obtained by minimizing the square sum of errors according to the following equation:

$$Objective\ Function = \sum_{k=1}^{N}(Q_{3,k}^{sim} - Q_{3,k}^{exp})$$

where; Q3 is the cumulative volume distribution.

$$Q_3^{sim} = \int_0^{d_{max}} q_3(d)\,\partial d$$

where; d: droplet diameter [mm]. Fig. 1 shows the resulted curve form simulating the cumulative volume distribution at the outlet to the experimental data, where a very good agreement is achieved.

Then the estimated constants were used to simulate a pilot plant RDC extraction column at steady state using the LLECMOD program; for the column height 4.4 m, column diameter 0.08 m, inlet of the dispersed and continuous phase are 0.85 m and 3.8 m, respectively. The two EFCE test systems butylacetate-acetone-water (B-A-W) and toluene-acetone-water (T-A-W) are used with inlet feed that is normally distributed with mean value equal to 2.1 mm and standard deviation of 0.6 mm.
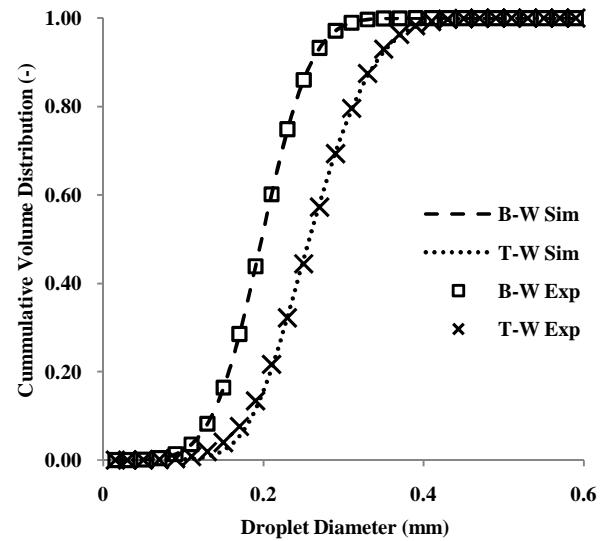
Fig. 1 Simulated and experimental cumulative volume distribution for butylacetate-water (B-W) and toluene-water (T-W) in a lab scale RDC
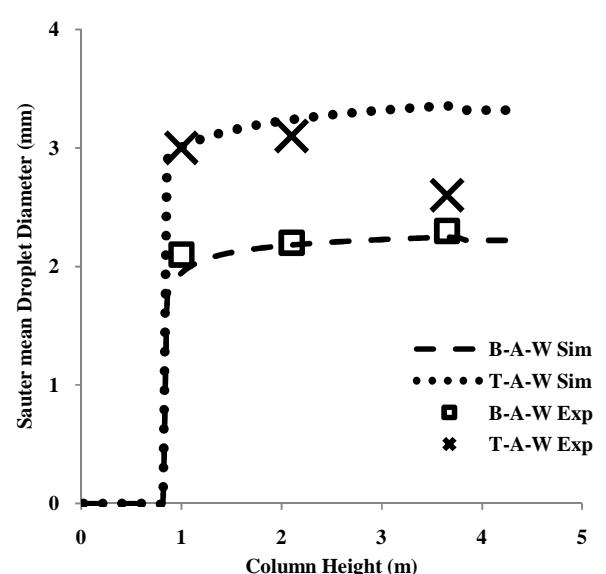
Fig. 2 Simulated Sauter mean droplet diameter along the column height compared to the experimental data [17] for two EFCE test system B-A-W & T-A-W

The pilot plant RDC extraction column geometric internal data: its compartment height is 50 mm, internal stator diameter is 50 mm, rotator diameter 45 mm and diameter of the rotating shaft is 10 mm. The angular velocity of the shafts is constant 200 rpm and volumetric flow rate for continuous and dispersed phase is 40 l/hr and 48 l/hr, respectively for both systems. The direction of mass transfer is from the continuous to the dispersed phase.
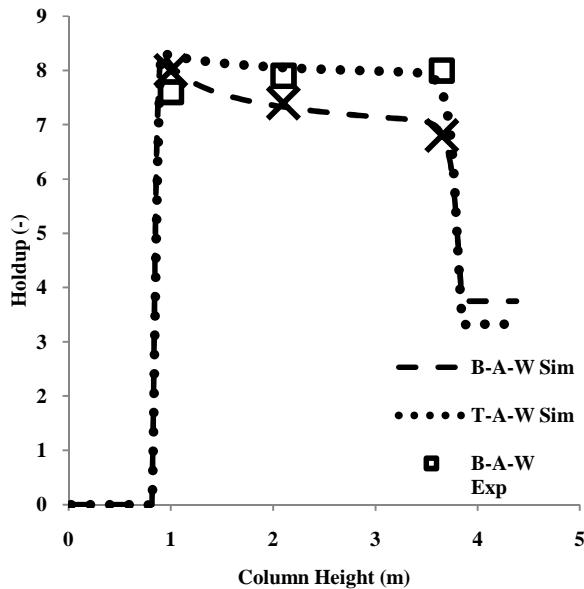
Fig. 3 Simulated holdup along the column height compared to the experimental data [17] for two EFCE test system B-A-W & T-A-W
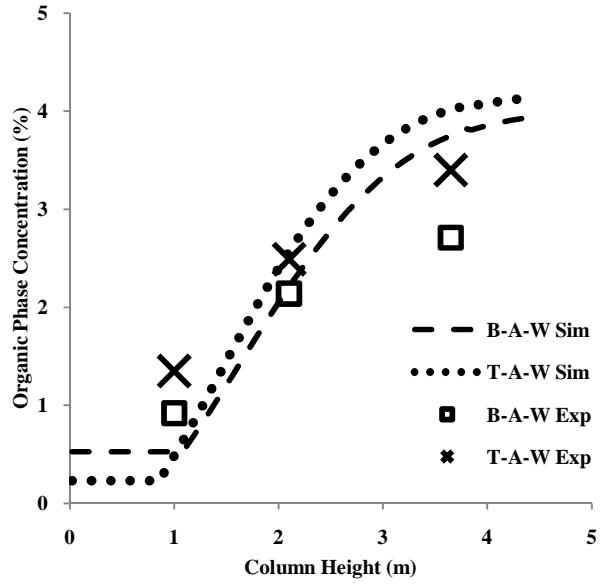
Fig. 5 Simulated organic solute concentration profile along the column height compared to the experimental data [17] for two EFCE test system B-A-W & T-A-W

Fig. 4 & 5 shows the simulated curve and experimental data for solute concentration profiles for aqueous and organic phase respectively as a function of column height, the agreement between the simulation and experiment is good for both test systems.

Results shown in Fig. 2 to 4; prove that the obtained values from the solution of the inverse problem gave a very good agreement to express the experimental data used from a pilot RDC plant in a simulation when inserting the values of Table 1 into the LLECMOD program.

VII. CONCLUSION

In this paper the inverse problem method was used for parameter estimation for different droplet coalescence models for lab scale RDC extraction column. The resulted parameters derived are valid for pilot plant extraction column but it is dependent on the chemical system used. The user friendly windows-based program LLECMOD was used for the hydrodynamics and mass transfer simulation of RDC extraction columns for steady state simulation of pilot plant RDC using the derived values of coalescence parameters resulted from the inverse problem program. The resulted simulation for the Sauter mean droplet diameter, hold-up and concentration organic and aqueous phase profiles showed a good agreement to the experimental data.

Fig. 4 Simulated aqueous solute concentration profile along the column height compared to the experimental data [17] for two EFCE test system B-A-W & T-A-W

A comparison between the simulated Sauter mean droplet diameter along the column height and the experimental data taken from the work of Garthe (2006) [17] is shown in Fig. 2 where a good agreement is achieved for both tested chemical systems.

Fig. 3 shows the simulated holdup along the column height compared to the experimental data for both chemical systems; where a very good agreement is achieved for both tested systems.

NOMENCLATURE

A_c: column cross-sectional area [m^2]

C: solute concentration [kg/m^3]

B, D: birth and death source terms [$1/m^3/s$]

$c_1 \ldots \ldots c_7$: coalescence parameters

d: droplet diameter [m]

d': mother droplet diameter [m]

D_R: column diameter[m]

f: Flux vector

Oh: Ohnesorge no. $= Oh = \dfrac{\eta_d}{(\sigma d \rho_d)^{0.5}}$

Q: volumetric flow rate [m^3/s]

Q_3: cumulative volume distribution

u: velocity [m/s]

t: time [s]

We: Weber no. $= we = \dfrac{\rho \varepsilon^{2/3} d^{5/3}}{\sigma}$

z: space coordinate [m]

GREEK SYMBOLS

ε: energy dissipation [m^2/sec^3]

η: dynamic viscosity [kg/m sec]

λ: coalescence efficiency

ρ: Density [kg/m^3]

σ: surface tension [N/m]

Υ: source term that represents the net number of droplet produced by breakage and coalescence [1/s]

φ: volume fraction

φ_d: hold up

ψ: internal and external coordinates vector ([d c_y z t])

ω: coalescence rate [m^3/sec]

SUBSCRIPTS

x, y: continuous and dispersed phases respectively

s: slip velocity

SUPERSCRIPTS

$\bullet$: derivative with respect to time

B: breakage

C: coalescence

exp: experimental

in: inlet

max: maximum

out: outlet

sim: simulated

REFERENCES

[1] J. Kumar, G. Warnecke, M. Peglow and S. Heinrich," Comparison of numerical methods for solving population balance equations incorporating aggregation and breakage," Powder Technology, vol. 189, issue 2, pp. 218-229, 2009.

[2] M. M. Attarakih, H.-J Bart, T. Steinmetz, M. Dietzen and N. M. Faqir, "LLECMOD: A Bivariate Population Balance Simulation Tool for Liquid- Liquid Extraction Columns," Open Chemical Engineering Journal, vol. 2, pp. 10-34, 2008.

[3] M. M. Attarakih H.-J Bart. L. Lagar G., and N. M. Faqir, "LLECMOD: A Window-based program for hydrodynamics simulation of liquid-liquid extraction columns," Chemical Engineering Processing, vol. 45, pp. 113-123, 2006.

[4] M. M. Attarakih H.-J Bart. and N. M. Faqir, "Numerical Solution of the Bivariate Population Balanced Equation for the Interacting Hydrodynamics and Mass Transfer in Liquid-Liquid Extraction Columns," Chemical Engineering Science, vol. 61, pp. 113-123, 2006.

[5] S. Kumar and D. Ramkrishna, "On the solution of population balance equations by discretization—II. A moving pivot technique," Chemical Engineering Science, vol. 51, issue 8, pp. 1333–1342, 1996.

[6] R. McGraw, "Description of aerosol dynamics by the quadrature method of moments," Aerosol Science and Technology, vol. 27, pp. 255–265, 1997.

[7] M. M. Attarakih H.-J Bart and N. M. Faqir, "Solution of the population balance equation using the sectional quadrature method of moments (SQMOM)," Computer Aided Chemical Engineering, vol. 21, pp. 209-214, 2009.

[8] A. M. O'Rourke and P.F. MacLoughlin, "A study of drop breakage in lean dispersions using the inverse-problem method," Chemical Engineering Science, vol. 65, issue 11, pp. 3681-3694, 2010

[9] A. Vikhansky, M. Kraft, M. Simon, S. Schmidt and H.-J. Bart, "Droplets Population Balance in a Rotating Disc Contactor: An Inverse Problem Approach," AIChE Journal, vol. 52, issue 4, pp. 1441- 1450, 2006.

[10] A. Braumann, P. L. W. Man and M. Kraft, "Statistical approximation of the inverse problem in multivariate population balance modeling," pp. 1473 – 4273, August 2009.

[11] Y. Liao and D. Lucas, "A literature review on mechanisms and models for the coalescence process of fluid particles," Chemical Engineering Science, vol. 65, issue 10, pp. 2851-2864, 2010.

[12] C. A. Coulaloglou and L. L. Tavlarides, "Description of interaction processes in agitated liquid–liquid dispersions," Chemical Engineering Science, vol. 32, pp. 1289–1297, 1977. H.

[13] Sovova, "Breakage and coalescence of drops in a batch stirred vessel-II Comparison of model and experiments," Chemical Engineering Science, vol. 36, pp. 1567–1573, 1981.

[14] T. Steinmetz, "Tropfenpopulationsbilanzgestütztes Auslegungs-verfahren zur Skalierung einer gerührten Miniplant Extraktions-kolonne," Forchr.-Ber. VDI, Reihe 3, Nr. 885, Düsseldorf: VDI Verlag 2007.

[15] G. Casamatta and A. Vogelpohl, "Modelling of fluid dynamics and mass transfer in extraction columns," German Chemical Engineering, vol. 8, pp. 96-103, 1985.

[16] M. Simon, "Koaleszenz von Tropfen und Tropfenschwärmen," Dissertation, Technischen Universität Kaiserslautern, Germany, April 2004.

[17] D. Garthe, "Fluid Dynamics and Mass Transfer of Single Particles and Swarms of Particles in Extraction Column," Dissertation, Technischen Universität München, Germany, April 2006.

Hardware Accelerators for Financial Mathematics - Methodology, Results and Benchmarking

Christian de Schryver #, Henning Marxen *, Daniel Schmidt #

#*Micrelectronic Systems Design Department, University of Kaiserslautern*
Erwin-Schroedinger-Strasse, 67663 Kaiserslautern, Germany
schryver@eit.uni-kl.de
schmidt@eit.uni-kl.de

*Stochastics and Financial Mathematics Department, University of Kaiserslautern
Erwin-Schroedinger-Strasse, 67663 Kaiserslautern, Germany
marxen@mathematik.uni-kl.de

Abstract—**Modern financial mathematics consume more and more computational power and energy. Finding efficient algorithms and implementations to accelerate calculations is therefore a very active area of research. We show why interdisciplinary cooperation such as (CM)^2 are key in order to build optimal designs.**

For option pricing based on the state-of-the-art Heston model, no implementation on dedicated hardware is known, yet. We are currently designing a highly parallel architecture for field programmable gate arrays based on the multi-level Monte Carlo method. It is optimized for high throughput and low energy consumption, compared to GPGPUs. In order to be able to evaluate different algorithms and their implementations, we present a benchmark set for this application. We will show a very promising outlook on future work, including dedicated ASIPs, fixed-point research and real-time applications.

Index Terms—**finance, benchmarking, hardware acceleration**

I. INTRODUCTION

Nowadays, financial markets are as vivid as never before. In modern electronic markets, stock prices may change several times within a few milliseconds. Participating traders (that can also be computers) have to evaluate the prices and react very quickly in order to get the highest profit, which requires a lot of computational effort.

In general, running these computations on servers or clusters with standard CPUs is not feasible due to either long run times or high energy consumption. Using general purpose graphics processing units (GPGPUs) as accelerators helps to increase the speed, but still requires a lot of energy. Besides, at the moment energy efficiency becomes more and more crucial for the reason of high energy costs and - even more critical - a limited supply of energy that can be provided. For example, in [16] it is stated that the City of London (with its new financial center Canary Wharf where a lot of leading institutes are located) does not provide additional energy until after the Olympic winter games in 2012, that have higher priority. Financial institutes are currently outsourcing all computing systems not used for pricing computations (such as storage or backup) out of the critical area.

This leads to the dilemma of needing faster computations on the one hand and limited energy resources on the other hand.
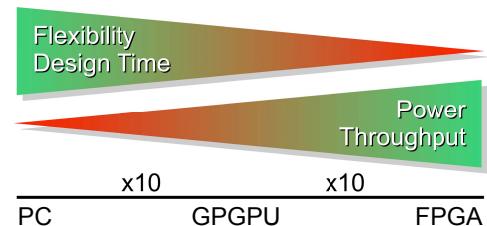
Fig. 1. PC vs. GPGPU vs. FPGA

Moving away from GPGPUs to dedicated hardware accelerators can help to drastically reduce the power consumption at the same or even higher throughput. For different application domains, some comparisons between CPU, GPGPU and programmable hardware units (field programmable gate arrays, FPGAs) have already been shown in [13] and [5], highlighting the enormous potential of energy savings for FPGAs. Figure 1 shows that standard software implementations require the least effort for implementation and can provide the highest flexibility, while dedicated hardware solutions on FPGAs are hard to design and - once finished - not easy to be changed again. From a different view, FPGAs can save up to about 99% of energy compared to a software implementation on a standard PC and allow a much higher throughput. GPGPUs are located between standard PCs and FPGAs. Between each neighboring architectures, one can expect a difference of about one order of magnitude on average for power consumption and throughput [13]. Although most financial institutes are relying on GPGPUs at the moment for the reason of standardized software development toolkits and their flexibility, FPGAs are an interesting alternative because of their higher energy efficiency.

A big challenge is the complexity of many models used to estimate the future price behavior of financial products. In many cases no mathematical closed-form solution exists so that approximation methods like Monte Carlo simulations or the finite difference method must be employed. Though, it is necessary to precisely specify a solution right at the beginning of the design process. Re-designing a nearly finished hardware implementation can require a very high amount of effort. The Center of Mathematical and Computational Modeling (CM)^2

of the University of Kaiserslautern is a perfect forum for an interdisciplinary cooperation to tackle this issue.

For this project, we have developed a design methodology that helps to select a feasible parameter set for a hardware accelerator in question that we present in Section III. In order to make implementations transparently comparable, we propose to use standardized benchmark sets - we elaborate on this in Section IV. By applying this methodology and our benchmark, we have developed some reference implementation designs that we show in Section V, together with the status quo of our research and our contributions up to now. In Section VI we give an overview of open issues and what we plan to examine in the future. Section VII concludes the paper.

II. State-of-the-Art and Related Work

Mathematical finance basically has two different directions. One is concerned with the evaluation of optimal investment strategies under certain market conditions and the other direction is the pricing of derivatives. The basic idea of pricing options is to assume some sort of model for the underlying price process and take the discounted expected value - under a certain measure - of an option as the option price.

A very common problem treated is the calculation of option prices based on the Black-Scholes model from 1973. This model relies on one stochastic differential equation and describes the price development of an option over the time, depending on market parameters such as riskless interest rate, long term drift and a *constant* volatility.

Accelerator design for financial mathematics is a very active research area, and several FPGA implementations have been published in the past. At the FPL 2008 Woods and VanCourt [17] presented a hardware accelerator for multiple, quasi-random, standard Brownian motions suitable for the acceleration of quasi-Monte Carlo simulation of financial derivatives. For credit risk modelling, Thomas and Luk could gain a speed-up of more than 90 times compared to a 2.4 GHz Pentium-4 Core2 [14]. An accelerator for Monte Carlo based credit derivative pricing was developed by Kaganov, Chow and Lakhany [10] in 2008 and showed to be 63 times faster than their software model. Wynnyk and Magdon-Ismail [18] presented an FPGA accelerator for American option pricing based on the Black-Scholes model in 2009 and could achieve a speedup of eleven up to 73 times compared to a software implementation running on a standard PC.

However, nowadays the Black-Scholes model is no longer up to date and does not provide an accurate reflection of modern financial market behaviors, mostly because of the volatility not being constant in reality. Furthermore, closed-form solutions for the Black-Scholes model exist and it only has demonstration purposes to apply stochastic solution methods such as Monte Carlo simulations or the finite difference method [1]. Nevertheless, it is still very common to publish accelerator implementations based on that model, at least in the electrical engineering community.

In 1993, Steven L. Heston presented a more accurate model [9] that extends the model from Black and Scholes by a second stochastic differential equation for stochastic volatility variations. This significantly increases the complexity of the calculation and of the implementation thereof. Nevertheless, the Heston model reflects the real behavior of current stock markets much better and is nowadays widely accepted in the financial mathematics community. But - to the best of our knowledge - no hardware accelerator for that model has been published up to now. For GPGPUs, the first implementations have been presented in the last year. Bernemann, Schreyer and Spanderen from the german bank WestLB [3] showed that they could achieve a speeup of 50 times over CPU by using GPGPUs for simulating the Heston model. Zhang and Oosterlee published a technical report [19] in March 2010 where they even showed speedups of more than 100 times.

The presented speed-ups look very impressive. However, unfortunately we were not able to fairly decide which solution seems to be the most promising for further research and refinements. We will go a bit more into the details of that problem in Section IV.

III. How To Choose The Right Design

For many fields of applications, finding the most efficient design under certain constraints is a difficult job. The main reason for this is a large *design space*. The design space is made up of all possible parameter choices for the design, that means all possible implementation instances.

Most parameters are not adjustable independently, since they are mutually linked. For example, fixing the target architecture to FPGAs one the one hand has a large impact on the selection of suitable algorithms and number systems, and on the other hand affects many performance metrics such as energy consumption, throughput and numerical precision.

Furthermore, the parameters within the design space are in many cases not limited to a single domain of expertise, but require interdisciplinary know-how and decisions. This makes not only the choice of the right values a challenge, but also the evaluation and comparison of different implementations. Besides speedup, more characteristics such as energy efficiency, convergence rate or numerical precision may be very important. This especially holds true for financial mathematics accelerators.

During our research we have seen a lot of papers that show elaborate implementations of a specific algorithm (see Section II) that is not questioned in the papers anymore. However, we claim that the algorithm itself is in fact not the most important selection. An accelerator should be designed to solve a specific problem - it does not matter which algorithm is used, as long as the result is calculated correctly. We therefore propose to distinguish clearly between three terms:

- the *problem* that is tackled (what to solve)
- the employed *model* (how to solve)
- the *solution* (how to build)

To clarify the situation, we use the *problem* "calculate the price of an option with two barriers for a given duration" as an example. European knock-out barrier options pay a certain amount of money at a fixed maturity time depending on the
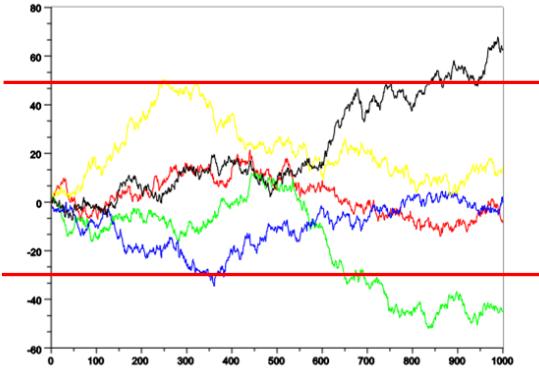
Fig. 2. Barrier testing for a Brownian motion

value of the underlying asset. This amount is only paid when the barrier is not crossed up to the maturity time. If one of the barriers is hit, the option becomes worthless. Thus it needs to be checked whether the barrier was ever hit or not. Figure 2 illustrates the typical random behavior of different realizations of an asset price over the time.

It is obvious that the problem description itself does not yet give any suggestions to the solution. Since the price of an option is tightly coupled to the price of a certain stock at the market, we need a *model* that provides the stock price behavior. For our chosen example, suitable models are for example the Black-Scholes model (outdated nowadays) or the Heston model. The model in general gives a formal and abstract view of (a certain aspect of) the problem.

The *solution* finally is a dedicated approach for solving a (modeled) problem. It is characterized by a specific *algorithm* and its *implementation*. For evaluating the Heston model, for example, finite difference methods or stochastic Monte Carlo simulations can be used. They may be implemented for example on standard PC clusters, GPGPUs or on FPGAs.

The parameters of the design space can basically be divided into two groups: the algorithmic parameters that are mostly selected by mathematicians, and the implementation parameters determined by the hardware designers. However, as mentioned before correlations exist between several parameters, so that the selection should be optimally made by having a generative exchange between experts of both groups.

For the rest of the paper, we will use the problem-model-solution triple "calculate the price of an option with two barriers for a given duration with the Heston model by using Monte Carlo methods" as a showcase. Even for that specific selection, the design space is still very large. An extract of the related design space is shown in Figure 3[1]. We cannot explain every parameter here (for details see [11] [8]), but in a nutshell it is obvious that even for a very specific task a huge amount of possible accelerator implementations may exist. In Figure 3 we see two different views of the same tree. These trees are symbolic for the design space. The left one is the view of a mathematician that has mostly to do with algorithmic and numerical aspects. The right view is the

one of an electrical engineer who may not understand the mathematician's concerns in detail, but is wondering what the best decisions with respect to hardware efficiency might be.

IV. BENCHMARKING - FAIRLY COMPARING IMPLEMENTATIONS

Comparing different implementations is a non-trivial task. Many attributes can be considered, including speed, accuracy and energy consumption. This becomes even more difficult when it is not clear which algorithm was used. Furthermore, in many cases it is not possible to distinct whether a presented algorithm or implementation has the displayed behavior only with the employed example or in a more general setting. Nevertheless, it is important to be able to compare various algorithms and different implementations, also over various target architectures.

Therefore the need for a benchmark set arises. This set should be independent of the algorithm and implementation used. For option pricing in financial mathematics, this need has already been claimed by Morris and Aubury in 2007 [12]. We are not aware of any progress made since that paper was published. We therefore have decided to develop a completely new benchmark that will enable us to fairly compare different algorithms, e.g. multi-level and single-level Monte Carlo, on different hardware. Thus we propose a benchmark based on the *problem/model* combination. In our case it is the pricing of double barrier options in the Heston model. It is clear that independently of the used algorithm and implementation the result must be the same. Therefore the final prices of the different options in the benchmark set have to be provided.

With the benchmark set it is possible to use different metrics, like speed (that is now the real time until the results are available), accuracy and power consumption, for the calculations leading to the right (or approximate) result without actually looking at the implementation and the algorithm itself. This allows a fair and publicly traceable comparison of the *solution* part of the *problem/model/solution* triplet.

The benchmark itself consists of different combinations of parameters for the Heston model and for barrier options, including the prices. The data for the Heston parameters is taken from different recent publications ([2], [11], [20]) and are enlarged by an extremer case. The benchmark parameters span a wide range of possible combinations used in this field. For some options of the benchmark closed form solutions exist that allow to obtain the exact results. This is important to verify that simulations converge to the correct values and makes it easier to compare the results. For the other cases the exact prices are not known and are therefore provided as close approximations.

For further publications we not only encourage the authors to use the presented benchmark but also give details of the algorithm and the implementation used. Thus it is possible to see where an increase in performance comes from. This is essential in order to evaluate the contribution of a certain result and to find ways to improve it even further. To achieve a higher transparency we will publish the code we used to

---

[1] http://www.sxc.hu/photo/285734 We thank sxc user vxdigital for sharing this image of the oak tree and allowing the use of it in this paper. He holds all copyrights to this image.

(a) Algorithmic parameters

(b) Implementation parameters
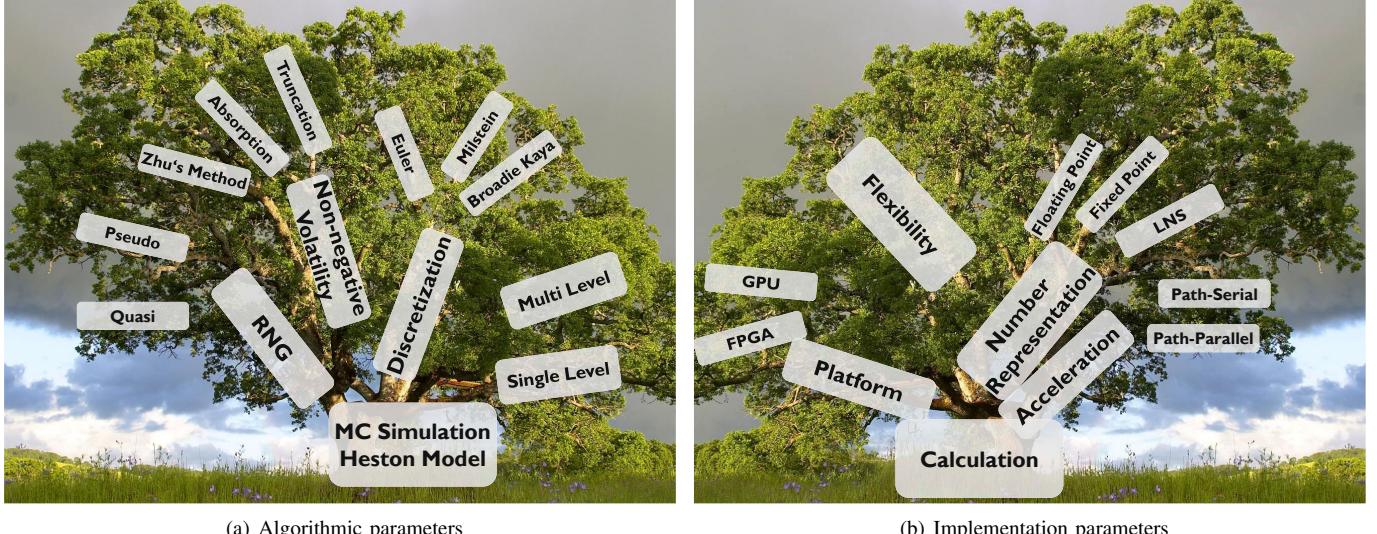
Fig. 3. Two views of the same solution tree

analyze the algorithms and the one we implemented on the FPGA.

The benchmark was directly used when comparing different Monte Carlo algorithms with the metric of computational complexity. In our special *problem/model* combination there are a lot more adjustments to the algorithms than seen in figure 3(a). Many of them can be combined, what leads to a huge design space that now can be handled by applying the benchmark set, so that we haven been able to choose a specific algorithm to be efficiently implemented on dedicated hardware. The multi-level Monte Carlo method provides a better asymptotic convergence behavior, using our benchmark we checked whether this method is beneficial for our application. We will show our first results in the next section.

V. STATUS QUO AND FIRST RESULTS

We have started this cooperation within (CM)^2 about one year ago now. Both participating chairs have experience of more than ten years in their respective field of research, so that we can profit from a lot of knowledge in the areas of efficient hardware design respectively stochastics and financial mathematics.

After evaluating the state-of-the-art, we decided to focus on accelerators for option pricing based on the Heston model - it seems to be a very promising topic since no implementations (either on GPGPUs or FPGAs) have been available one year ago. In contrast to that, the Heston model is already widely spread within the financial community. From former research done in the group of Prof. Korn, multi-level Monte Carlo methods [7] seemed to provide a better convergence behavior than standard single-level Monte Carlo or finite difference methods. Monte Carlo methods also have the advantage of being very flexible. A barrier that is only relevant on a certain time interval to evaluate an option price for example can be easily implemented. Furthermore, multi-dimensional problems can also be solved. This is needed in the case that an option has

more than one underlying asset. Nevertheless, for our project we will stick to one asset.

A. A New Random Number Generator for Non-Uniform Distributions

Inherently, Monte Carlo simulations always consume a huge amount of random numbers. To obtain the maximum hardware efficiency for our implementation, we have developed a new random number generator for non-uniform distributions tailored to our application.

For our option pricing accelerators, we need two independent, normally distributed random numbers for each time step of a single simulated stock price path. In general, non-uniformly distributed random numbers are generated in two steps: First, a uniform random number generator creates uniformly distributed values, and in a second step this number is transformed into the desired target distribution.

For the uniform random number generation, a lot of research has already been made leading to efficient and well-proven implementations, such as the Mersenne Twister MT19937 that we use. The three main approaches for obtaining non-uniform distributions are transformation, rejection, and inversion methods [15].

For FPGAs, inversion methods are the usual way to go. They combine many desireable properties: by applying the respective inverse cumulative distribution function (ICDF), they transform every input sample $x \in (0, 1)$ from a uniform distribution to one output sample $y = icdf(x)$ of the desired output distribution by using piecewise polynomial approximation of the ICDF. The works of Woods and VanCourt [17] and Cheung et al. [4] show FPGA implementations of the inversion method.

However, both implementations use fixed-point number representations at the input. This leads to a loss of precision in the tail regions where the probability of a value lying there is very low. But these extreme events can have a large impact, for example for options with barriers it is crucial to know if a

barrier was hit or not, since it completely changes the refund conditions. We have therefore developed a new implementation based on floating-point representation that provides the same precision over the whole ICDF implementation at much lower hardware costs. This work has been presented at the 2010 International Conference on ReConFigurable Computing and FPGAs (ReConFig) in December in Cancún, Mexico [6]. Our random number converter unit requires only about half of the area compared to other state-of-the-art implementations by even higher numerical precision.

To validate our work, it was crucial to develop a new testing methodology, since standardized test suites do not exist for non-uniform distributions. This work has been carried out in cooperation with Elke Korn who has a lot of knowledge in the field of random numbers. The methodology and the validation results for our implementation are also presented in the paper. Our random numbers did not show any noticeable problems in the stochastic tests and also perfectly passed two different application simulations.

B. Fully Parallel Hardware Accelerator

To the best of our knowlege, no hardware implementations of option price accelerators based on the Heston model exist at the moment. We have therefore started with the first implementation, that is nearly finished now. The hardware is fully-parallel, fully-pipelined and designed for high throughput.
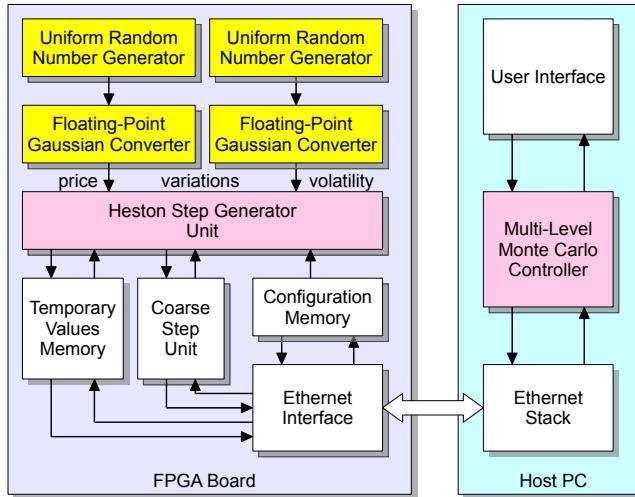
Fig. 4.   Fully Parallel Accelerator Structure

Figure 4 shows the structure of one pipelined accelerator circuit. In each clock cycle, our unit consumes two normally distributed random numbers, one for the stock price variation and one for the volatility variation. The Heston step generator unit calculates the price and volatility values for the next time step based on a multi-level Monte Carlo algorithm. The pipeline depth is about 60 stages. In order to maximally utilize the pipelined hardware, it computates one time step for 60 assets in parallel, before moving to the next time step. The values for the respective time steps are stored in a memory temporarily. The coarse step memory holds interim

values for a higher step width, this means for a lower Monte Carlo simulation level. In the configuration memory, all model parameters are stored.

Due to the inherent parallelism of Monte Carlo simulations, it is not only feasible but self-evident to instantiate as many of these circuits as possible on an FPGA in order to increase the simulation throughput.

The accelerator is implemented on a Xilinx ML-605 evaluation board equipped with a Xilinx Virtex-6 FPGA. The board is connected via Gigabit ethernet to a host PC running the user interface and the program that calculates and sets the configuration values for the accelerator based on the retrieved simulation results.

Synthesis and benchmarking results will be available soon. We are currently supporting single- and double-precision floating-point computations and are working on a fixed-point implementation as well.

VI. Outlook and Future Work

Also for the intended future work a close cooperation between financial mathematics and electrical engineering will be mandatory, since we are planning to research aspects out of both fields.

One characteristic of the Monte Carlo method is the inherent capability to parallelize the calculation. It therefore makes no difference whether several calculations are done in parallel but slowly or only one calculation is done at high speed, as long as the number of calculations remains equal altogether. At the moment, our parallel implementation allows simulating one time step on many assets simultaneously. The whole procedure of calculating one time step is fixed on the FPGA and limited to a single algorithm that is set at design time.

One possible way to improve this is to sequentially compute the basic calculations needed for one time step on a Application-Specific Instruction-Set Processor (ASIP) within the FPGA, i.e. it is runtime-programmable. This procedure can reduce the required area and allows to calculate various algorithms since the functionality is defined in a corresponding program. It is therefore sufficient to load a different program without changing the hardware. The ASIP will occupy much less area than the parallel implementation presented in Section V-B, therefore many ASIPs can be instantiated in parallel. We are currently investigating the necessary instruction set.

To increase the speed of the implementation even more, the floating-point computation can be replaced by fixed point computations. In order to do so, errors resulting from the use of fixed-point calculations have to be approximated. This task will also require both theoretical and practical expertise.

Besides working on the implementation, the benchmark is very important to evaluate the designs. It will also allow to research fixed-point solutions with respect to necessary precision. Further steps will be to publish the analysis of the algorithms with the benchmark in a journal. To increase the transparency even more, we are currently setting up a web site offering all the program code used to create the analysis. It will contain an implementation of the multi-level and the

crude Monte Carlo algorithm with a focus on the calculation complexity rather than the implementation.

To provide more benchmarking results, also for different architectures, we are working on a GPGPU implementation. It provides more flexibility and less implementation work than hardware designs do, but also requires higher energy consumption. In order to verify this assumption the implementation is required. The Monte Carlo simulation algorithm for the Heston model is currently carried out as a student work.

Moreover, we are currently researching on real-time acceleration of financial calculations. This means that hardware or GPGPU accelerators are linked to real-time data streams. This approach seems to be very promising for keeping track with the prices changing quickly in high-frequency trading.

VII. Conclusion

The financial world is running faster and faster and the importance of energy consumption increases drastically. To address this challenge the question of pricing double barrier options in the Heston setting is faced. As the model is more complex than the famous Black-Scholes model and these types of options are path dependent, the algorithms for the calculations are more distinct and also the implementation thereof. In order to be able to cope with the strong connection between the algorithm and the implementation, a combined mathematical and electrical engineering view is needed. $(CM)^2$ provides a perfect framework to do so.

To approximate the pricing process Monte Carlo simulations are used. For a good implementation a fast algorithm with an adjusted implementation thereof is needed. In order to distinguish the different algorithms we have created a benchmark set for double barrier options. This benchmark allows to fairly analyze and compare the diverse algorithms and designs, which is a very important issue due to the big differences in the convergence speed of these algorithms.

For the target architecture, using FPGAs is the hardware of choice if implementation time is not considered. It allows fast computation with low energy consumption. Nevertheless, optimal FPGA designs require deep understanding of the FPGA characteristics and the calculations need to be optimized for it. One commonality of all the Monte Carlo algorithms is the use of (pseudo-)random numbers. In the Heston setting standard normal random numbers are used. There are procedures to create these running efficiently on GPGPUs and CPUs. A new method was presented which is very efficient for an implementation on a FPGA.

The detailed analysis of the diverse algorithms was used to make an efficient implementation. Therefore the algorithm is implemented on an FPGA. This should allow a fast computation with low energy consumption. As far as we know, this will be the first implementation of a Monte Carlo simulation in the Heston model on a FPGA. Furthermore it will be the first implementation of the multi-level Monte Carlo method on this hardware. Thus this work expands the field of implementations of financial mathematical problems on dedicated hardware in several ways as new concepts are taken into consideration.

References

[1] R. L. Akers, I. Bica, E. Kant, C. Randall, and R. L. Young. SciFinance: A Program Synthesis Tool for Financial Modeling. *AI Magazine*, 22(2):27, 2001.

[2] L. Andersen. Efficient simulation of the heston stochastic volatility model. 2006.

[3] A. Bernemann, R. Schreyer, and K. Spanderen. Pricing Structured Equity Products on GPUs. In *High Performance Computational Finance (WHPCF), 2010 IEEE Workshop on*, pages 1 –7, Nov. 2010.

[4] R. C. C. Cheung, D.-U. Lee, W. Luk, and J. D. Villasenor. Hardware Generation of Arbitrary Random Number Distributions From Uniform Distributions Via the Inversion Method. *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 15(8):952–962, Aug. 2007.

[5] B. Cope, P. Y. Cheung, W. Luk, and S. Witt. Have GPUs made FPGAs redundant in the field of Video Processing? In *Field-Programmable Technology, 2005. Proceedings. 2005 IEEE International Conference on*, pages 111 –118, Dec. 2005.

[6] C. de Schryver, D. Schmidt, N. Wehn, E. Korn, H. Marxen, and R. Korn. A New Hardware Efficient Inversion Based Random Number Generator for Non-Uniform Distributions. In *Proceedings of the International Conference on ReConFigurable Computing and FPGAs (ReConFig)*, pages 190–195, Dec. 2010.

[7] M. B. Giles. Multilevel Monte Carlo path simulation. *Operations Research-Baltimore*, 56(3):607–617, 2008.

[8] A. V. Haastrecht and A. Pelsser. Efficient, almost exact simulation of the Heston stochastic volatility model. *International Journal of Theoretical and Applied Finance*, 13(1):1–43, 2010.

[9] S. L. Heston. A Closed-Form Solution for Options with Stochastic Volatility with Applications to Bond and Currency Options. *Review of Financial Studies*, 6(2):327, 1993.

[10] A. Kaganov, P. Chow, and A. Lakhany. FPGA Acceleration of Monte-Carlo based Credit Derivative Pricing. In *Proc. Int. Conf. Field Programmable Logic and Applications FPL 2008*, pages 329–334, Sept. 2008.

[11] R. Lord, R. Koekkoek, and D. van Dijk. A comparison of biased simulation schemes for stochastic volatility models. *Quantitative Finance*, 10(2):177–194, 2010.

[12] G. W. Morris and M. Aubury. Design Space Exploration of the European Option Benchmark using Hyperstreams. In *Field Programmable Logic and Applications, 2007. FPL 2007. International Conference on*, pages 5 –10, Aug. 2007.

[13] D. B. Thomas, L. Howes, and W. Luk. A Comparison of CPUs, GPUs, FPGAs, and Massively Parallel Processor Arrays for Random Number Generation. In *Proceeding of the ACM/SIGDA international symposium on Field programmable gate arrays*, FPGA '09, pages 63–72, New York, NY, USA, 2009. ACM.

[14] D. B. Thomas and W. Luk. Credit Risk Modelling using Hardware Accelerated Monte-Carlo Simulation. In *Proc. 16th Int. Symp. Field-Programmable Custom Computing Machines FCCM '08*, pages 229–238, Apr. 2008.

[15] D. B. Thomas, W. Luk, P. H. Leong, and J. D. Villasenor. Gaussian Random Number Generators. *ACM Comput. Surv.*, 39(4):11, Oct. 2007.

[16] P. Warren. City business races the Games for power. *The Guardian*, May 2008.

[17] N. A. Woods and T. VanCourt. FPGA Acceleration of Quasi-Monte Carlo in Finance. In *Proc. Int. Conf. Field Programmable Logic and Applications FPL 2008*, pages 335–340, 2008.

[18] C. Wynnyk and M. Magdon-Ismail. Pricing the American Option using Reconfigurable Hardware. In *Computational Science and Engineering, 2009. CSE '09. International Conference on*, volume 2, pages 532 –536, Aug. 2009.

[19] B. Zhang and C. W. Oosterlee. Acceleration of Option Pricing Technique on Graphics Processing Units. Technical Report 10-03, Delft University of Technology, Feb. 2010.

[20] J. E. Zhang and J. Shu. Pricing s&p 500 index options with heston's model. In *Proc. IEEE Int Computational Intelligence for Financial Engineering Conf*, pages 85–92, 2003.

Characterisation of the Cyclic Deformation Behaviour of the Nodular Cast Iron ASTM 80-55-06 on the Basis of Physical Measurement Methods

Holger Germann[#1], Peter Starke[#2], Dietmar Eifler[#3]

[#]*Institute of Materials Science and Engineering;University of Kaiserslautern*
P.O. Box 3046, 67653 Kaiserslautern, Germany

[1]germann@mv.uni-kl.de
[2]starke@mv.uni-kl.de
[3]eifler@mv.uni-kl.de

Abstract—**The current paper is focussed on the nodular cast iron ASTM 80-55-06 (EN-GJS-600). The individual microstructure was investigated by light and scanning electron microscopy as well as micro-hardness measurements. Stress-controlled fatigue tests were performed at ambient temperature with testing frequencies of f = 5 Hz and f = 150 Hz. The cyclic deformation behaviour was characterised by means of mechanical stress-strain hysteresis (f = 5 Hz) as well as the change in temperature, electrical resistance and frequency (f = 150 Hz) measurements. Increasing testing frequencies result in higher values of the change in temperature caused by the increasing heat dissipation per second.**

I. Introduction

Good mechanical, physical and manufacturing properties as well as the relatively low production costs lead to a common interest in cast irons for various technical applications [1]. Due to the combination of a relatively high tensile strength and an adequate ductility nodular cast iron is often used for highly stressed components in the automotive and commercial vehicle industry, e.g. for crankcases, crankshafts or exhaust manifolds, as well as in the wind power industry, e.g. for rotor hubs or nacelles [2-5]. An appropriate material selection and a weight-optimised dimensioning are essential for safe and economic operation conditions of such cast iron components. Therefore, the characterisation of the fatigue behaviour is of major importance. In the literature, fracture-mechanic investigations [6-8] as well as fatigue tests with mechanical stress-strain hysteresis [9] and in some cases temperature measurements [10] are described for nodular cast iron [11].

In the scope of the current paper stress-controlled fatigue tests were performed with specimens of the nodular cast iron ASTM 80-55-06 (EN-GJS-600) at ambient temperature on servohydraulic testing systems with a testing frequency of 5 Hz and on resonators with a testing frequency of 150 Hz. For a detailed microstructural-based characterisation of the fatigue behaviour, high-precision mechanical stress-strain hysteresis (f = 5 Hz), temperature, electrical resistance and frequency (f = 150 Hz) measurements were applied. The plastic strain amplitude ($\varepsilon_{a,p}$) as well as the change in temperature (ΔT), electrical resistance (ΔR) and frequency (Δf) were plotted versus the number of cycles in cyclic

deformation, cyclic temperature, cyclic electrical resistance and cyclic frequency curves. All measured quantities are based on microstructural changes due to fatigue processes in the plastically deformed volume of the gauge length of the specimens, e.g. deformation-induced matrix debonding of the graphite. On the basis of the measured temperature data a linear relation between the ΔT values at f = 5 Hz and f = 150 Hz was found. For both testing frequencies the Woehler (S-N) curves were determined. In addition, fracture surfaces were investigated by using scanning electron microscopy (SEM).

II. Investigated Cast Iron

The cast iron ASTM 80-55-06 (EN-GJS-600) was provided by the Daimler AG in round bars with a diameter of 36 mm and a length of 300 mm. Brinell hardness measurements yield values of 235±6 HBW30. Figure 1 shows characteristic light (a) and scanning electron (b) micrographs of the investigated material. The microstructure consists of a pearlitic matrix with a ferrite content of 14.6±2 area-% and a graphite content of 9.8±0.8 area-%. As can be seen in Figure 1, the formation of ferrite is predominantly observed in ferrite zones located in the surrounding of the graphite nodules. The mean diameter of the graphite nodules was determined to 19.1±6.8 μm. Table I summarises the microstructural parameters of the cast iron.
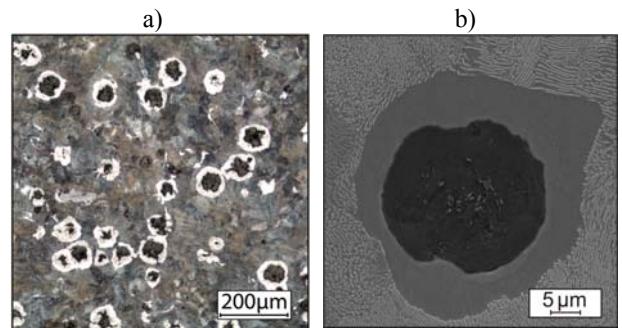

Fig. 1 Light (a) and scanning electron (b) micrographs of the microstructure of the cast iron ASTM 80-55-06 (**EN-GJS-600**)

TABLE I
MICROSTRUCTURAL PARAMETERS OF THE CAST IRON ASTM 80-55-06
(EN-GJS-600)

Ferrite fraction [area-%]	Graphite fraction [area-%]	Mean diameter of nodules [μm]
14.6±2	9.8±0.8	19.1±6.8

In order to evaluate the hardness of the pearlite, ferrite and graphite fraction a Martens micro-hardness pattern with a width of 200 x 200 μm was measured on a polished cross-section (Figure 2). The load of 0.1 N was applied without a hold time within 10 s. Afterwards, the microstructure was etched and correlated with the results of the micro-hardness measurements. The micro-hardness values significantly decrease in the sequence pearlite (3095 HM0.1), ferrite (1757 HM0.1) and graphite (597 HM0.1). Corresponding to the phase distribution, a mean micro-hardness value of 2654 HM0.1 can be calculated by multiplying the pearlite (ferrite, graphite) content with the appropriate micro-hardness value of 3095 (1757, 597) HM0.1. It is possible to revalue the Martens hardness to Vickers hardness and a revaluation according to [12] yields about 240 HV0.1. This value shows a good accordance with the Brinell hardness determined to be 235±6 HBW30.

HM0.1 [N/mm²]

4000 – 5000	
2500 – 4000	Pearlite
1000 – 2500	Ferrite
0 – 1000	Graphite

Fig. 2 Martens micro-hardness pattern for the cast iron ASTM 80-55-06 (EN-GJS-600)
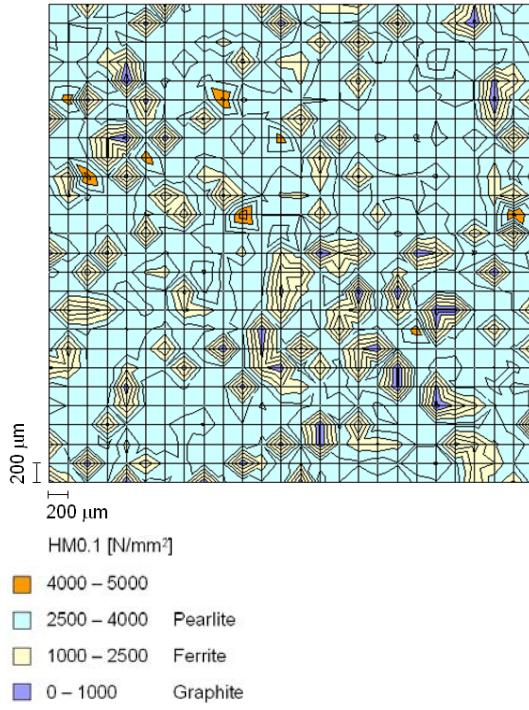
All fatigue tests were performed with the specimen geometry shown in Figure 3.
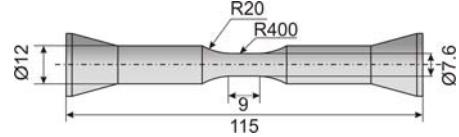
Fig. 3 Specimen geometry

III. EXPERIMENTAL SETUP

Stress-controlled load increase tests (LITs) and constant amplitude tests (CATs) were carried out at ambient temperature at a testing frequency of 5 Hz on servohydraulic testing systems using a triangular load-time function at a load ratio of R = -1. In the LITs, the stress amplitude σ_a was increased from $\sigma_{a, start}$ continuously with the rate $d\sigma_a/dt = 11.1 \cdot 10^{-3}$ MPa/s until specimen failure. The CATs were performed until failure or to a maximum number of cycles N_{max} of $2 \cdot 10^6$. In Figure 4, the experimental setup is shown schematically.
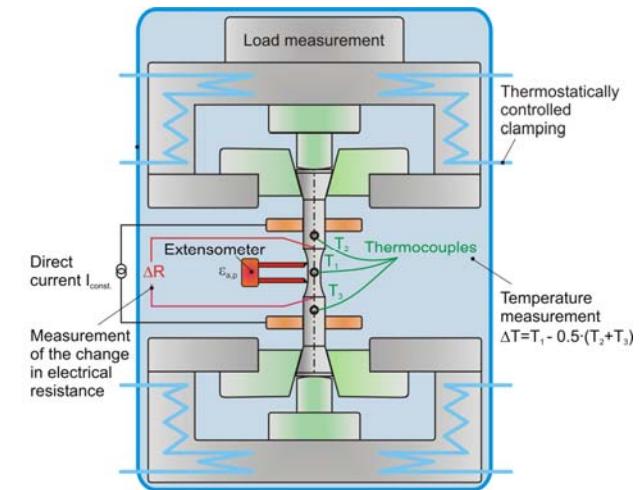
Fig. 4 Experimental setup for mechanical hysteresis, temperature, and electrical resistance measurements during fatigue loading schematically

During the fatigue tests, the plastic strain amplitude $\varepsilon_{a,p}$ [13], the change in temperature ΔT [14] and the change in electrical resistance ΔR [15] were measured to characterise the microstructure-based fatigue behaviour in detail. The physical quantities $\varepsilon_{a,p}$, ΔT and ΔR are a function of the deformation induced change in microstructure in the bulk of the specimen and represent the actual fatigue state [16-19]. For the measurement of $\varepsilon_{a,p}$ an extensometer was fixed in the middle of the gauge length. The change in temperature ΔT was detected with one thermocouple in the middle of the gauge length (T_1) and two thermocouples at the elastically loaded specimen shafts (T_2 and T_3). For electrical resistance measurements a DC-power supply was fixed at both shafts and ΔR was measured with two wires spot welded at the transition of the gauge length and the shafts (Figure 4). Apart from the geometry, the change in electrical resistance ΔR strongly depends on the resistivity ρ^* which is directly related to deformation induced changes in the microstructure, e.g. dislocation density and arrangement, vacancies, micro-

pinholes, micro-shrinkage cavities or micro-cracks. In the case of cast irons the measured value ΔR is of major importance, in particular to get detailed information about the actual fatigue state with respect to damage mechanisms like graphite-matrix debonding.

Furthermore, stress-controlled CATs were performed at ambient temperature at a testing frequency of 150 Hz on a resonator using a sinusoidal load-time function at a load ratio of R = -1 until specimen failure or to a maximum number of cycles N_{max} of $2 \cdot 10^7$. In addition to the above mentioned physical quantities, the change in frequency Δf of the electromagnetical resonance device can be used for the characterisation of the fatigue behaviour. Caused by the functional principle of a resonator the specimen is part of the spring-mass-system of the testing setup and the measured value Δf depends on the damping behaviour of the specimen which depends on the deformation induced changes in the microstructure during fatigue loading [20].

IV. RESULTS

A. Load Increase Tests

Load increase tests (LITs) allow a reliable estimation of the endurance limit with one single specimen related to a maximum number of cycles $N_{max} = 2 \cdot 10^6$. In Figure 5, besides the stress amplitude σ_a, the plastic strain amplitude $\varepsilon_{a,p}$, the change in temperature ΔT and the change in electrical resistance ΔR are plotted versus the number of cycles N for a LIT with the cast iron ASTM 80-55-06 (EN-GJS-600). The ΔR-N curve indicates an initial decrease, among others caused by closing micro-cracks during the compression half cycles. Then, the courses of the change in electrical resistance ΔR are characterised by a saturation state between $6 \cdot 10^4$ and $1 \cdot 10^5$ cycles, followed by an increase indicating cumulative graphite-matrix debonding and micro-crack growth. A significant change in the slope of the $\varepsilon_{a,p}$-N, ΔT-N and ΔR-N curve of the LIT occurs at $\sigma_{RW, LIT} = 220$ MPa. This stress amplitude can be used for the estimation of the endurance limit [14]. The stress amplitude 398 MPa leads to specimen failure.
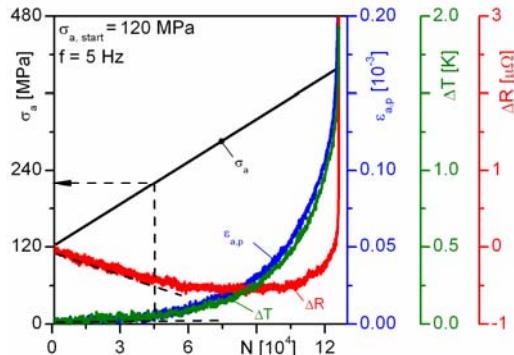
Fig. 5 Plastic strain amplitude, change in temperature and change in electrical resistance in a continuous load increase test for the cast iron ASTM 80-55-06 (**EN-GJS-600**)

B. Constant Amplitude Tests

Constant amplitude tests (CATs) were performed with stress amplitudes in the range of $220 \leq \sigma_a \leq 340$ MPa at a testing frequency of 5 Hz. In Figure 6a, the plastic strain amplitude is plotted versus the numbers of cycles. The $\varepsilon_{a,p}$-N curves immediately indicate plastic strain amplitudes between $0.015 \cdot 10^{-3}$ for $\sigma_a = 220$ MPa and $0.12 \cdot 10^{-3}$ for $\sigma_a = 340$ MPa followed by cyclic hardening processes in the matrix until macroscopic crack growth. After an initial increase caused by thermal conduction, the ΔT-N curves shown in Figure 6b also describe the cyclic hardening as a consequence of reduced plastic deformation. In Figure 6c, the change in electrical resistance is plotted versus the numbers of cycles. At first, the ΔR-N curves are characterised by a decrease leading to minimum ΔR values between -2.81 μΩ for $\sigma_a = 220$ MPa and -1.32 μΩ for $\sigma_a = 340$ MPa. Then, with increasing numbers of cycles an enforced graphite-matrix debonding and micro-crack growth result in increasing ΔR-N values.
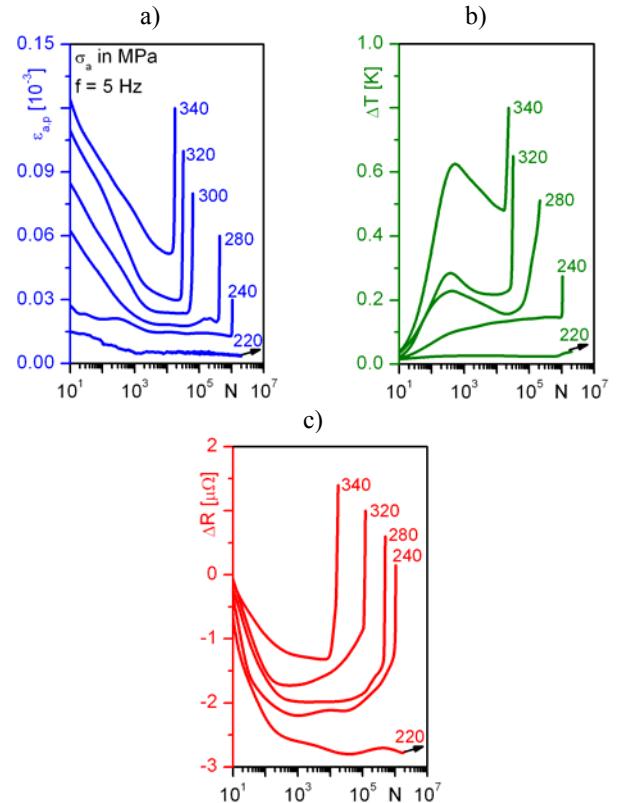
Fig. 6 Cyclic deformation curves (a), cyclic temperature curves (b) and cyclic electrical resistance curves (c) for constant amplitude tests with a testing frequency of 5 Hz for the cast iron ASTM 80-55-06 (**EN-GJS-600**)

The CATs performed with stress amplitudes in the range $240 \leq \sigma_a \leq 340$ MPa lead to numbers of cycles to failure between $2 \cdot 10^4$ and $1 \cdot 10^6$ whereas the CAT performed with $\sigma_a = 220$ MPa reaches $2 \cdot 10^6$ cycles without failure. The stress amplitude of 220 MPa corresponds very well to $\sigma_{RW, LIT} = 220$ MPa estimated in the LIT. This underlines the high capability of LITs for a reliable estimation of the

endurance limit of cast irons with one single specimen. Thus, this test procedure yields large economic advantages.

To evaluate the influence of the testing frequency on the cyclic deformation behaviour, CATs were performed in the range of $190 \leq \sigma_a \leq 280$ MPa at a starting testing frequency of 150 Hz. In Figure 7, cyclic deformation curves are shown on the basis of the change in temperature ΔT (a) and the change in frequency Δf (b). The ΔT-N curves are characterised by an initial increase due to thermal conduction leading to maximum ΔT values which are located between 1.12 K for $\sigma_a = 190$ MPa and 5.48 K for $\sigma_a = 280$ MPa. Then, a decrease of the ΔT-N curves indicates cyclic hardening processes in the matrix until macroscopic crack growth. The change in frequency depends on the before mentioned cyclic deformation behaviour of the investigated material, whereby a decrease (increase) of the Δf values is caused by an increase (decrease) of the damping capacity. At first, a slight increase of the Δf values probably indicating micro-crack closure occurs. Then, with increasing numbers of cycles the Δf-N curves decrease due to a cumulative graphite-matrix debonding and micro-crack growth. The CATs performed with stress amplitudes in the range $200 \leq \sigma_a \leq 280$ MPa result in numbers of cycles to failure between $3.5 \cdot 10^5$ and $3.5 \cdot 10^6$. The CAT performed with $\sigma_a = 190$ MPa reaches $2 \cdot 10^7$ cycles without failure.

Fig. 8 Linear relation between the change in frequency and the change in temperature for f = 150 Hz for the cast iron ASTM 80-55-06 (**EN-GJS-600**)

Figure 9 shows the comparison of the ΔT-N curves for constant amplitude loading with $\sigma_a = 280$ MPa at the testing frequencies of 5 Hz and 150 Hz. An increase in the testing frequency from f = 5 Hz to f = 150 Hz leads to significantly higher ΔT values because of a higher heat dissipation per second in the plastically deformed gauge length of the specimen.

Fig. 7 Cyclic temperature curves (a) and cyclic frequency curves (b)) for constant amplitude tests with a testing frequency of 150 Hz at the range $190 \leq \sigma_a \leq 280$ MPa for the cast iron ASTM 80-55-06 (**EN-GJS-600**)

In Figure 8, the change in frequency Δf is plotted as a function of the change in temperature ΔT at $N = 1 \cdot 10^4$ cycles at stress amplitudes in the range $190 \leq \sigma_a \leq 280$ MPa. As can be seen, there is a linear relation between both quantities.


Fig. 9 Comparison of the cyclic temperature curves at f = 5 Hz and f = 150 Hz for constant amplitude loading with $\sigma_a = 280$ MPa for the cast iron ASTM 80-55-06 (**EN-GJS-600**)

At stress amplitudes of 220 MPa, 230 MPa and 280 MPa CATs were performed at testing frequencies of 5 Hz and 150 Hz. On the basis of temperature data measured at $1 \cdot 10^4$ cycles a linear relation between the ΔT values for f = 5 Hz $\Delta T_{5\,Hz}$ and the ΔT values for f = 150 Hz $\Delta T_{150\,Hz}$ can be calculated. The $\Delta T_{5\,Hz}$-$\Delta T_{150\,Hz}$ relation is presented in Figure 10. The ratio $\Delta T_{5\,Hz}/\Delta T_{150\,Hz}$ determined to 0.039 corresponds very well with the ratio between the testing frequencies of 0.033. This illustrates that a higher testing frequency results in higher heat dissipation per second and unit of the plastically deformed volume in the gauge length of the specimen. The increase of the temperature is a function of decreasing values of the plastic strain amplitude [22,23] and the increased heat dissipation per second for increasing testing frequencies.
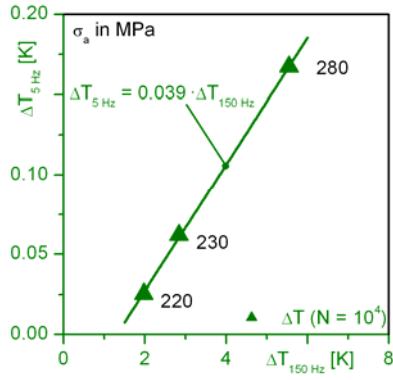
Fig. 10 Linear $\Delta T_{5\,Hz}$-$\Delta T_{150\,Hz}$ relation determined on the basis of temperature data measured at $1 \cdot 10^4$ cycles in constant amplitude tests with a frequency of f = 5 Hz and f = 150 Hz for the cast iron ASTM 80-55-06 **(EN-GJS-600)**

Figure 11 shows the Woehler (S-N) curves for both testing frequencies. As can be seen e.g. in [22], provided that the testing conditions are identical, an increase in the testing frequency generally results in longer lifetimes. Nevertheless, with regard to the CATs performed in the scope of this paper, due to the different load-time functions, at identical stress amplitudes the lifetimes are slightly longer at f = 5 Hz (triangular load-time function) in comparison to f = 150 Hz (sinusoidal load-time function), see e.g. [23]. The endurance limit was not determined statistically in the scope of this paper. At f = 5 Hz (f = 150 Hz) the specimen loaded with 230 MPa (190 MPa) reaches $2 \cdot 10^6$ ($2 \cdot 10^7$) cycles without failure.
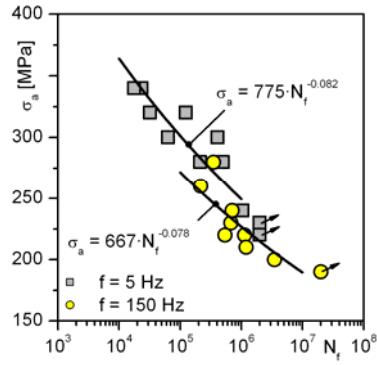
Fig. 11 Woehler curves for the testing frequency f = 5 Hz and f = 150 Hz for the cast iron ASTM 80-55-06 **(EN-GJS-600)**

To get more information about the fracture mechanisms fracture surfaces were investigated with scanning electron microscopy (SEM). Independent of the stress amplitude and the testing frequency the initiation of fatigue cracks can be often observed at defects, like micro-pinholes or micro-shrinkage cavities. Figure 12a shows the fatigue crack initiation at a defect in the centre of the fracture surface of a specimen, which was loaded with $\sigma_a = 300$ MPa at a testing frequency of 5 Hz and reached a number of cycles to failure of $4.1 \cdot 10^5$. In Figure 12b, the striation area is shown in detail. The final fracture surface, which is shown in Figure 12c in detail, is characterised by characteristic dimple structures.
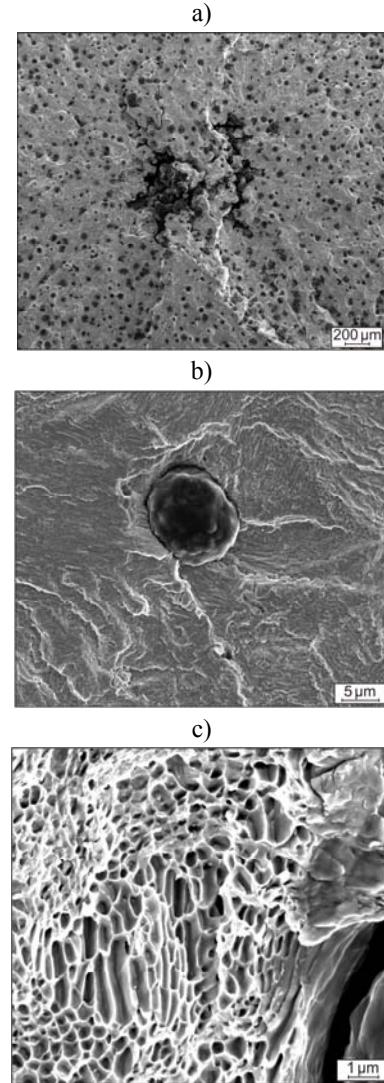

Fig. 12 Crack initiation at a defect in the centre of the specimen (a), striation area (b) and final fracture surface (c) for a constant amplitude test with $\sigma_a = 300$ MPa at a testing frequency of 5 Hz for the cast iron ASTM 80-55-06 **(EN-GJS-600)**

V. SUMMARY

The present paper is focused on the cast iron ASTM 80-55-06 (EN-GJS-600). The microstructure predominantly consists of a pearlitic matrix with nodular graphite. In load increase tests at a testing frequency of 5 Hz the endurance limit of the investigated cast iron can be estimated with one single specimen. The cyclic deformation behaviour under constant amplitude loading was characterised at the testing frequencies f = 5 Hz and f = 150 Hz on the basis of the plastic strain amplitude (f=5Hz) as well as the change in temperature, electrical resistance and frequency (f=150 Hz). At both testing frequencies the cyclic deformation behaviour is dominated by cyclic hardening of the matrix and graphite-matrix debonding until macroscopic crack initiation. In the constant amplitude tests at f = 150 Hz higher values for the change in temperature were measured caused by the higher energy dissipation per

second in the plastically deformed gauge length compared to fatigue tests with $f = 5$ Hz. With respect to the change in temperature ΔT, a liner relation between the ΔT values at $f = 150$ Hz and $f = 5$ Hz can be determined. The fatigue strength at $f = 5$ Hz is slightly higher than at $f = 150$ Hz because of the different load time functions, in particular triangular for $f = 5$ Hz and sinusoidal for $f = 150$ Hz.

REFERENCES

[1] R. O'Rourke, "Cast iron: The engineered metal," *Advanced Materials & Processes*, vol. 159, pp. 65-68, Jan. 2001.

[2] M Hafiz, "Mechanical properties of SG-iron with different matrix structure," *Journal of Materials Science*, vol. 36, pp. 1293-1300, 2001.

[3] M. A. Kenawy, A. M. Abdel-Fattah, N. Okasha and M. EL-Gazery, "Mechanical and Structural Properties of Ductile Cast Iron," *Egyptian Journal of Solids*, vol. 24, pp. 151-159, 2001.

[4] F. Iacoviello, O. Di Bartolomeo, V. Di Cocco and V. Piacente, "Damaging micromechanisms in ferritic-pearlitic ductile cast irons," *Materials Science and Engineering A* , vol. 478, pp. 181-186, 2008.

[5] D. Veldkamp, "A probabilistic evaluation of wind turbine fatigue design rules," *Wind Energy*, vol. 11, pp. 655-672, 2008.

[6] T. Ogawa and H. Kobayashi, "Near threshold fatigue crack growth and crack closure in a nodular cast iron," *Fatigue & Fracture of Engineering Materials & Structures*, vol. 10, pp. 273-280, 1987.

[7] H. Yaacoub Agha, A.-S. Béranger, R. Billardon, and F. Hild, "High-cycle fatigue behaviour of spheroidal graphite cast iron," *Fatigue & Fracture of Engineering Materials & Structures*, vol. 21, pp. 287-296, 1997.

[8] P. Hübner, H. Schlosser, G. Pusch, and H. Biermann, "Load history effects in ductile cast iron for wind turbine components," *International Journal of Fatigue*, vol. 29, pp. 1788-1796, 2007.

[9] C. Guillemer-Neel, V. Bobet, and M. Clavel, "Cyclic deformation behaviour and Bauschinger effect in ductile cast iron," *Materials Science and Engineering A, vol.* 272, pp. 431-442, 1999.

[10] H. Q. Xue, E. Bayraktar, and C. Bathias, "Damage mechanism of a nodular cast iron under the very high cycle fatigue regime," *Journal of Materials Processing Technology*, vol. 202, pp. 216-223, 2008.

[11] P. Hübner, G. Pusch and L. Krodel, "Fatigue Behaviour of Cast iron with Globular Graphite," *Advanced Engineering Materials*, vol. 6, pp. 541-544, 2004.

[12] C. Heermant and D. Dengel, "Klassische Werkstoffkennwerte abschätzen," *MP Material Testing*, vol. 38, pp. 374-378, 1996.

[13] P. Lukáš and M. Klesnil, "Cyclic stress-strain response and fatigue life of metals in low amplitude region," *Materials Science and Engineering*, vol. 11, pp. 345-356, 1973.

[14] D. Dengel, H. Harig, "Estimation of the fatigue limit by progressively increasing load tests," *Fatigue & Fracture of Engineering Materials & Structures*, vol. 3, pp. 113-128, 1980.

[15] J. Polák, "Electrical resistivity of cyclically deformed copper," *Czechoslovak Journal of Physics*, vol. 18, pp. 315-322, 1969.

[16] P. Starke, F. Walther, and D. Eifler, "PHYBAL – A new method for lifetime prediction based on strain, temperature and resistance measurements," *International Journal of Fatigue*, vol. 28, pp. 1028-1036, 2006.

[17] P. Starke, F. Walther, and D. Eifler; "Fatigue assessment and fatigue life calculation of quenched and tempered SAE 4140 steel based on stress-strain hysteresis, temperature and electrical resistance measurements," *Fatigue & Fracture of Engineering Materials & Structures*, vol. 30, pp. 1044-1051, 2007.

[18] P. Starke and D. Eifler, "Fatigue assessment and fatigue life calculation of metals on the basis of mechanical hysteresis, temperature, and resistance Data," *MP Materials Testing*, vol. 51, pp. 261-268, 2009.

[19] H. Germann, P. Starke, E. Kerscher, and D. Eifler, "Fatigue behaviour and lifetime calculation of the cast irons EN-GJL-250, EN-GJS-600 and EN-GJV-400," *Procedia Engineering*, vol. 2, pp. 1087-1094, 2010.

[20] V. Wagner, P. Starke, E. Kerscher, and D. Eifler, „Cyclic deformation behaviour of railway wheel steels in the very high cycle fatigue (VHCF) regime," *International Journal of Fatigue*, vol. 33, pp. 69-74, Jan. 2011.

[21] D. Eifler, *Zusammenhang zwischen Mikrostruktur und Schwingfestigkeitsverhalten bei Stählen*, D. Munz, Ed.. Oberursel, Germany: DGM Informationsgesellschaft mbH, 1985.

[22] W. J. Harris, Ed., *Metallic Fatigue*, ser. International Series of Monographs in Aeronautics and Astronautics, Division VIII: Materials Science & Engineering. Oxford, United Kingdom: Pergamon Press, 1961, vol. 1.

[23] K. Koibuchi and M. Yamane, "Wave-form effect on fatigue strength and its rheological analysis", *Bulletin of JSME*, vol. 11, pp.761-770, 1968.

Zur Untersuchung von stahlfaserverstärkten Betonen auf Basis der 3D-Computer-Tomographie

Frank Schuler [#1], Katja Schladitz [*2],

[#]*Fachgebiet Massivbau und Baukonstruktion, Technische Universität Kaiserslautern*
Paul-Ehrlich-Str. Geb. 14, 67663 Kaiserslautern
[1]`fschuler@rhrk.uni-kl.de`

[*]*Fraunhofer Institut für Techno- und Witschaftsmathematik*
Fraunhofer-Platz 1, 67663 Kaiserslautern
[2]`katja.schladitz@itwm.fraunhofer.de`

Abstract—In einem gemeinsamen Forschungsprojekt haben die TU Kaiserslautern und das Fraunhofer Institut für Techno- und Wirtschaftsmathematik ITWM nachgewiesen, dass die Computer-Tomographie (CT) kombiniert mit moderner Bildverarbeitungs- und Analysesoftware geeignet ist, die Faserorientierung und Faserverteilung in faserverstärkten Hochleistungsbetonen zu untersuchen. Gegenüber bisher angewendeten Methoden eröffnet die Computer-Tomographie die Möglichkeit, für Betone mit Fasern und Gelegen aller Art die Faserorientierung und Faserverteilung im gesamten Volumen eines Probekörpers zu betrachten und zu analysieren.

Nach Abschluss dieser Machbarkeitsstudie stehen nun Versuche im Bereich stahlfaserverstärkter Betone im Vordergrund. Dabei sollen neue Anwendungsfelder in der Praxis und im Bereich der Forschung erschlossen werden.

Ein neuartiger Versuchsaufbau der die CT-Aufnahme eines Probekörpers während der Belastung erlaubt, ermöglicht erstmalig die in-situ Beobachtung der Betonmikrostruktur und des Faserverhaltens bei mechanischer Beanspruchung.

I. EINLEITUNG

Betone unterschiedlichster Art sind hervorragend geeignet, Druckspannungen aufzunehmen, jedoch ist die Zugfestigkeit dieser Betone wesentlich niedriger und von vielen äußeren Parametern abhängig. Gerade bei steigender Druckfestigkeit der Betone ist das Nachbruchverhalten, sprich die Duktilität des Baustoffes, nicht zufriedenstellend. Um die geforderte Duktilität zu erreichen und dem spröden Bruchverhalten entgegenzuwirken, werden diesen Betonen Fasern beigemischt. Für die Verbesserung der Festbetoneigenschaften ist die Fasermenge, Faserorientierung und Faserverteilung ausschlaggebend. Dies macht eine Überwachung dieser Eigenschaften notwendig, sei es zur Qualitätssicherung auf der Baustelle oder im Bereich der Forschung zur Weiterentwicklung von Hochleistungsbetonen.

Zu Beginn des Forschungsvorhabens wurde daher die Möglichkeit untersucht, Fasern in faserverstärkten Betonen mithilfe der Computer-Tomographie (CT) zu detektieren und hinsichtlich ihrer Orientierung und Verteilung zu analysieren. Bei diesen Vorversuchen kamen unterschiedliche Faserarten in Betonen mit unterschiedlicher Probengeometrie zum Einsatz. Diese Studie hat gezeigt, dass die CT geeignet ist, die Faserorientierung und -verteilung zu quantifizieren. Insbesondere kann der Orientierungsbeiwert mit Methoden der 3D-Bildanalyse robust gemessen werden, siehe [1] und [2].

Weiterführende Arbeiten umfassten insbesondere Versuche zur Beurteilung der Wiederholstreuung der Messergebnisse sowie die Entwicklung vereinfachter bzw. teilweise automatisierter Analysetools.

Aufgrund der Anwendbarkeit in der Praxis und der Mitarbeit des Fachgebiets Massivbau und Baukonstruktion der TU Kaiserslautern im Schwerpunktprogram SPP 1182 „Nachhaltiges Bauen mit ultra-hochfestem Beton" der Deutschen Forschungsgemeinschaft (DFG) konzentrieren sich die aktuellen Arbeiten einerseits auf den Schwerpunkt „stahlfaserverstärkte Ultrahochfeste Betone" und andererseits auf die Untersuchung von „normal festen" Stahlfaserbetonen.

Neben der Darlegung von möglichen Anwendungsgebieten der im Folgenden vorgestellten CT-Analyse-Verfahren, liegt der Fokus dieses Aufsatzes auf der Beobachtung und Beschreibung des Nachbruchverhaltens dieser Betone mithilfe dieser Verfahren.

Fig. 1 gibt einen strukturierten Überblick über die bisher durchgeführten und geplanten Anwendungen der CT-Analyse-Methoden.

II. COMPUTER-TOMOGRAPHIE

A. Grundlagen

Mithilfe der Computer-Tomographie können zerstörungsfrei alle Arten von Fehlstellen in Bau-/Werkstoffen (Poren, Lunker, Einschlüsse, etc.) detektiert und als virtuelle, dreidimensionale Körper dargestellt werden. Hierbei können Größe und Lage solcher Fehlstellen genau vermessen werden.

Die 3D-Computer-Tomographie ist ein bildgebendes, zerstörungsfreies Prüfverfahren, bei dem das Objekt aus unterschiedlichen Richtungen mit Röntgenstrahlen durchstrahlt wird. Dabei entstehen Projektionsbilder, welche die Stärke der Absorption an verschiedenen Stellen abhängig von der Materialart und Materialstärke widerspiegeln. Aus den Projektionen wird die dreidimensionale Aufnahme tomographische rekonstruiert.

Zur Erzeugung eines tomographischen 3D-Bildes wird die Betonprobe schrittweise gedreht. Fig. 2 zeigt eine schemati-
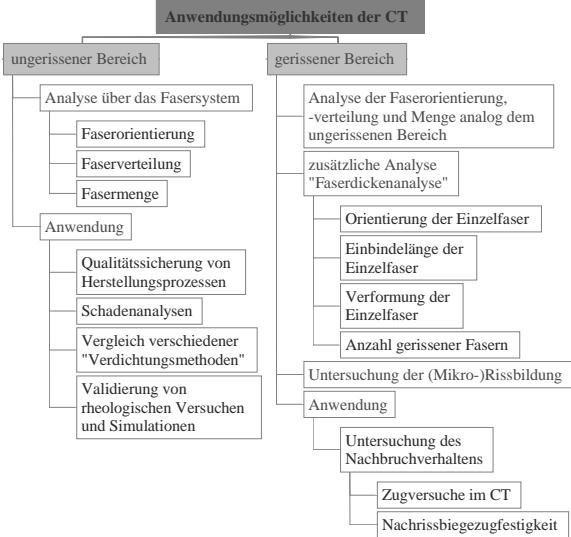
Fig. 1. Anwendungsmöglichkeiten der Computer-Tomographie

sche Darstellung einer CT-Aufnahme. Für jede Winkelposition wird dabei jeweils ein Projektionsbild erzeugt, bis ein Winkelbereich von 360 Grad abgedeckt ist. Die tomographische Rekonstruktion erzeugt dann ein räumliches Bild des untersuchten Probekörperteils. Aus diesen dreidimensionalen Bildern können nun Bestandteile mit gleichen physikalischen Eigenschaften herausgefiltert werden.

Verschiedene Komponenten der untersuchten Probe werden durch unterschiedlichen Absorptionskontrast auf entsprechend verschiedene Grauwerte abgebildet.
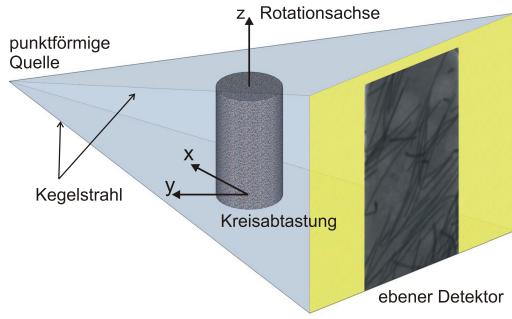
Fig. 2. Schematische Darstellung einer 3D-Computer-Tomographie-Aufnahme

B. Segmentierung des Fasersystems

Auf der Basis des segmentierten Fasersystems können wichtige Kennwerte wie der Faservolumengehalt sowohl der gesamten Probe als auch in Teilvolumina bestimmt werden. So kann die Homogenität der Struktur beurteilt werden.

Die rekonstruierten mikrocomputertomographischen Bilder (μCT) liegen in einem 16-bit Graustufenformat vor, d. h. in Abstufung von 0 (schwarz) bis 65.536 (weiß).

Segmentierung bedeutet im vorliegenden Fall, die Fasern von der Betonmatrix und eventuell vorhandenen anderen

Objekten zu trennen. Alle Pixel, die zur Faserkomponente gehören, bekommen den Wert „1", allen übrigen Pixel (Betonmatrix, Poren, Einschlüsse, etc.) wird der Wert „0" zugewiesen. Da es jetzt im Bild nur noch zwei Werte gibt, wird dieses Verfahren "Binarisierung"genannt.

Zum Einsatz kommt ein Schwellwertverfahren, bei dem nur einzelne Pixel betrachtet werden. Der Grauwert eines Pixels stellt einen reinen Helligkeitswert dar, nach dem jedes Pixel der Gruppe (Segment) zugeordnet wird (vgl. [3]). Fig. 3 stellt an zwei Beispielen die Segmentierungsergebnisse den 16-bit Schnittbildern gegenüber.
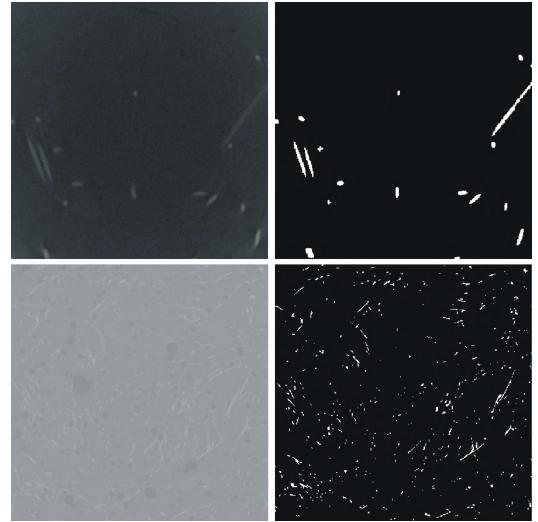

Fig. 3. Zweidimensionale Schnitte der 3D-Aufnahmen vor und nach der Segmentierung; Oben: Graustufenbild (links) und segmentiertes Bild (rechts) eines Stahlfaserbetons (Stahlfaser: l=60 mm, d=1 mm, Kantenlänge des betrachteten Bereichs: 68 mm); Unten: Graustufenbild (links) und segmentiertes Bild (rechts) eines UHPC (Stahlfaser: l=6 mm, d= 0,175 mm, Kantenlänge des betrachteten Bereichs: 48 mm)

C. Faserorientierungsbeiwert

Der Faserorientierungsbeiwert beschreibt den Einfluss der Faserorientierung auf die Wirksamkeit der Faser. Bisher wurde bei der experimentellen Bestimmung des Faserorientierungsbeiwertes die Orientierung der Faser zum Beispiel an einer Schnittfläche oder Bruchfläche ermittelt. Ein Überblick über diese Analysemethoden wird in [1], [4] und [5] gegeben.

Im Gegensatz hierzu wird bei der Analyse der dreidimensionalen Bilder der Computer-Tomographie gleich das gesamte Volumen betrachtet. Dieser Faserorientierungsbeiwert η_ϕ beschreibt das Verhältnis der projizierten Faserlänge $L_{p,\phi}$ zur tatsächlichen Länge L_V innerhalb des betrachteten Volumens V und kann so nach Gleichung (1) berechnet werden. Der berechnete Faserorientierungsbeiwert liegt zwischen „0" und „1" (Wert „1", alle Fasern richten sich in Richtung ϕ aus; Wert „0", alle Fasern richten sich orthogonal zur Richtung ϕ aus).

$$\eta_\phi = \frac{L_{p,\phi}}{L_V \cdot V} \tag{1}$$

Sowohl die Faserlängendichte L_V als auch die verallgemeinerte Projektionslänge $L_{p,\phi}$ lassen sich leicht aus den Dichten

der inneren Volumina ableiten bzw. als Nebenprodukt bei deren Bestimmung anhand digitaler Bilddaten gewinnen. Für Details siehe [6]. L_V und $L_{p,\phi}$ für 13 Raumrichtungen lassen sich mithilfe des Softwaresystems MAVI [7] effizient aus 3D Bilddaten bestimmen.

Zur Beurteilung der Wiederholbarkeit der Messergebnisse wurde ein Probekörper mehrfach tomographiert. Neben der Variation des Startwinkels der CT-Aufnahmen wurde der Probekörper auch überkopf aufgenommen.

Die Ergebnisse bestätigen eine hohe Aussagegenauigkeit der ermittelten Faserorientierungsbeiwerte. Lediglich bei der Ermittlung des Faservolumens kann durch untereinander abweichende „Bildqualitäten" die Wahl eines geeigneten Schwellwertes erschwert werden, was bei den durchgeführten Versuchen zu Abweichungen von max. 1 % führte.

D. Faserdickenanalyse

Für die quantitative Analyse des Faserauszugsverhaltens ist u.a. die Faserdicke interessant. Lägen die Fasern einzeln und sich nicht berührend in der Probe, so könnte jede Faser im Bild als einzelnes Objekt segmentiert und vermessen werden. Trotz des im Vergleich z. B. zu faserverstärkten Kunststoffen geringen Faservolumengehalts, berühren sich viele Fasern. Die Bildverarbeitung bietet verschiedene Methoden zur Trennung von Objektes, jedoch ist keine robust auf langgestreckte Objekte mit großem Berührungsbereich anwendbar.

Eine Alternative bietet der in [8] entwickelte Algorithmus für die Bestimmung sogenannter Radienkarten. In jedem Pixel des Fasersystems werden die Sehnenlängen in 26 Raumrichtungen gemessen und anschließend aus den Sehnenlängen Trägheitsachse und -momente berechnet. Die korrigierte Trägheitsachse ist gerade die lokale Richtung der Faser. Mit ihrer Hilfe kann dann aus den Sehnenlängen der Radius abgeleitet werden:

$$r = \frac{1}{8} \sum_{i=12}^{19} d_i \sin(\angle(\nu_i, \nu)) \tag{2}$$

wobei d_i die nach ihrer Größe sortierten Sehnenlängen sind, ν_i die zugehörigen Richtungen und ν die lokale Richtung der Faser. Die kleinsten und größten Sehnenlängen werden nicht in die Mittelung einbezogen, um den Schätzer robuster zu machen.

Die Korrektheit dieses Alogorithmus wurde in [8] anhand computergenerierter Strukturen nachgewiesen. Die Ergebnisse für reale Daten hängen jedoch stark von der Qualität der Binarisierung (vgl. Abschnitt II-B) ab. Bei weichen Grauwertübergängen zwischen Matrix und Fasern kann die Faserdicke daher leicht unterschätzt werden. Der Einfluss der Binarisierung kann jedoch durch direktes Arbeiten auf dem Grauwertbild vermieden werden. Statt der Länge von Sehnen im Vordergrund werden nun die (gerichteten) Abstände zu einem starken Grauwertabfall bestimmt, siehe [8].

III. ANWENDUNGSMÖGLICHKEITEN DER COMPUTER-TOMOGRAPHIE

A. Analyse der Faserorientierung, -verteilung und -menge mithilfe der CT

1) Sonderbauteile: Bei der Entwicklung von Formteilen aus mit Fasern hochbewehrten ultrahochfesten Betonen sind Faserorientierung, -verteilung und -menge für die Festbetoneigenschaften die ausschlaggebenden Faktoren. Fig. 4 zeigt einen ausgewählten Bildausschnitt einer 3D-Visualisierung eines Druck-Schublagers der Fa. Halfen, dessen Zulassungsversuche an der TU Kaiserslautern durchgeführt wurden. Die Kenntnis über die Verteilung, Ausrichtung und Menge der Fasern gestattet die Qualitätskontrolle der Herstellung und die Beurteilung des Tragverhaltens.
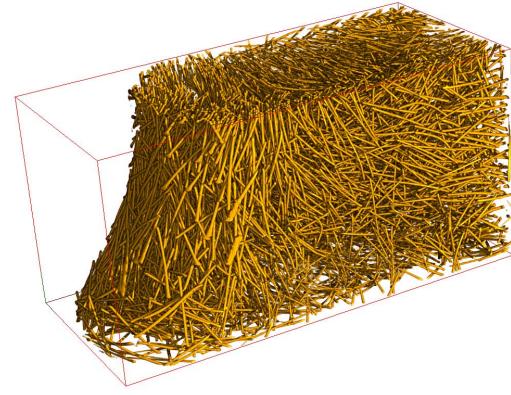

Fig. 4. Ausgewählter Bildausschnitt einer 3D-Visualisierung eines Druck-Schublagers der Fa. Halfen

2) Untersuchung von Stahlfaser-Industriefußböden: Zur Qualitätskontrolle der Herstellung und der Verarbeitung von Stahlfaserbetonen gibt das CT-Analyseverfahren Auskunft über die Orientierung, die Menge und die Verteilung der eingebauten Fasern.

Mit Unterstützung der ortsansässigen Transportbetonfirma TRAPOBET wird eine Reihenuntersuchung an ausgeführten Stahlfaserindustriefußböden bzw. -bodenplatten durchgeführt.

Aufgrund der immer gleichen Materialstärke bei der Durchstrahlung, stellen zylindrische Probekörper die ideale Probengeometrie dar. Daher werden für die Versuche Bohrkerne entnommen.

Um eine ausreichend genaue Aussage über Fasergehalt und Fasermenge treffen zu können, müssen an jeder Messstelle mehrere Bohrkerne ($n > 3$) entnommen werden. Gegenüber einer „anonymen" Entnahme von Proben, stehen in dieser Prüfreihe alle relevanten Parameter wie der Mischungsentwurf, die Art des Beimischens der Stahlfasern sowie das Einbringund Verdichtungsverfahren zur Bewertung zur Verfügung.

Ziel ist es den Einfluss verschiedener Herstellverfahren auf die Ausrichtung und Verteilung der Fasern zu untersuchen.

Der in Fig. 5 dargestellte Probekörper, zeigt einen Mix von geraden Fasern mit gekröpften ①, abgeplatteten ②

und als Köpfchen ③ ausgebildeten Enden, sowie gewellten Fasern ④.
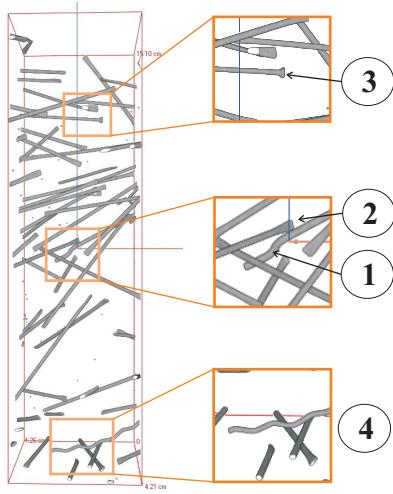
Fig. 5.   Visualisierung eines Bohrkerns, Fasermix, betrachteter Probebereich $40 \times 40 \times 150\,\mathrm{mm}$

B. Untersuchung des Nachbruchverhaltens

1) Zugversuche im Computer-Tomographen: Zur experimentellen Untersuchung der lokalen Faserverbundeigenschaften und der Mikrorissbildung von faserverstärkten Hochleistungsbetonen unter Last wurde ein Verfahren entwickelt, das es ermöglicht, Computer-Tomographie-Aufnahmen während der Belastung des Probekörpers durchzuführen.

Die Aufnahme der CT-Bilder wird bei unterschiedlichen Laststufen durchgeführt, so dass zwischen den dreidimensionalen Bildern der einzelnen Laststufen Differenzbilder erstellt werden können. Unter einem Differenzbild versteht man ein Bild, das genau die Unterschiede zweier Bilder enthält. So kann man in dreidimensionalen Bildern die Veränderung der Mikrostruktur sichtbar machen.

Mit dem Versuchsaufbau (vgl. Fig. 6) werden zur Zeit CT-Aufnahmen mit Pixelgrößen kleiner $15\,\mu\mathrm{m}$ erreicht. Aufgrund dieser hohen Auflösung darf sich zur „verwacklungsfreien" Aufnahme der aufzunehmende Bereich nicht bewegen. Gerade dies macht die in-situ Beobachtung schwierig, da der Probekörper beim Halten einer konstanten Spannung kriecht. Daher muss das Aufnahmeverfahren dem Kriechverhalten des Baustoffes angepasst werden. Dies lässt sich nur über eine Verkürzung der Aufnahmedauer erreichen.

Um trotz einer verkürzten Aufnahmezeit Bilder mit möglichst hoher Auflösung und gutem Kontrast zu erhalten, können bei der Tomographie verschiedene Einstellungen verändert werden. Beispielsweise können die Winkelpositionen und die zu mittelnden Bilder (je Winkelposition werden mehrere Bilder aufgenommen und gemittelt) reduziert werden, was bei einer zu starken Reduzierung zum Rauschen des Bildes und einem schlechteren Kontrast führt. Eine Verkürzung der Belichtungszeit pro Bild führt zu dunkleren Bildern, da weniger Energie am Detektor ankommt. Bei den momentan


Fig. 6.   Lastzelle im Computer-Tomographen

stattfindenden Versuchen werden diese und weitere Parameter aufeinander abgestimmt. So konnte die Aufnahmedauer je Laststufe auf ca. 20 Minuten reduziert werden.

Entgegen der üblichen „Knochen"-Form eines Probekörpers für zentrische Zugversuche, werden zylindrische Probekörper verwendet. Bei jeder Winkelposition wird die gleiche Materialstärke durchstrahlt. Diese Probeform reduziert Strahlungsschatten und verbessert den Grauwertverlauf.

Die Verwendung von eingeschnürten Zylindern ermöglicht die Beobachtung der Mehrfachrissbildung, die typisch für die angestrebte Duktilität von faserverstärkten ultrahochfesten Betonen ist. Für gekerbte Probekörper spricht die Möglichkeit, Untersuchungen an einem definierten Einzelriss durchzuführen. Der vorgegebene Entstehungsort des Risses ermöglicht eine nähere Positionierung der Röntgenröhre am Probekörper, so dass Aufnahmen mit höherer Auflösung (ca. $8\,\mu\mathrm{m}$/Pixel) erreicht werden können.

Zur Zeit finden Vorversuche statt, um trotz der angesprochenen „Probleme" wie das Kriechen der Probekörper, CT-Aufnahmen mit ausreichender Schärfe und Kontrast für eine Analyse zu erhalten. Die dann folgenden experimentellen Untersuchungen konzentrieren sich sowohl auf die Beobachtung und Beschreibung des Faserauszugverhaltens von Einzelfasern und Fasergruppen als auch die topographische Beschreibung und Vermessung der (Mikro-)Rissbildung.

2) Nachrissbiegezugfestigkeit an miniaturisierten Balken: In Anlehnung an die Nachrissbiegezugprüfung nach Vorgabe des DFG Schwerpunktprogramms SPP 1182 „Nachhaltiges Bauen mit ultra-hochfestem Beton" werden zurzeit verformungsgesteuerte Biegezugprüfungen an miniaturisierten Balken (Prismen) durchgeführt.

Fig. 7 zeigt beispielhaft die Auswertung der Nachrissbiegezugprüfung zweier Prismen der gleichen Herstellungscharge. Die Grafik zeigt in der Last-Verformungs-Kurve bei Probe „UHPC 1" ein etwas duktileres Entfestigungsverhalten als bei Probe „UHPC 2". Dies lässt sich durch die unterschiedliche Orientierung und Verteilung der Fasern im Probekörper begründen.

Aufgrund der Probekörperabmessungen kann der gesamte Probekörper nach der Ermittlung der Nachrissbiegezugfes-

Fig. 7. Last-Verformungs-Kurve, Einfluss der Faserorientierung, -verteilung und -menge

| | Fasermenge | Faserorientierungsbeiwert η_φ | | | $f^f_{cfl,L1}$ | $f^f_{cfl,L2}$ |
	[kg/m³]	x	y	z	[N/mm²]	[N/mm²]
Probe 1	21,4	0,57	0,31	0,60	2,20	0,84
Probe 2	18,1	0,71	0,29	0,45	3,41	2,13

der Fasern im Probequerschnitt beruhen, wodurch sich eine unterschiedliche Anzahl von Fasern im Betonquerschnitt des gerissenen Bereichs befinden. Jedoch ist nicht nur die Anzahl der Fasern im enstehenden Rissbereich ausschlaggebend für das Nachrissverhalten, sondern auch die Orientierung und Verteilung der Fasern im Querschnitt.

Um die Einflüsse der Orientierung und der Menge der Fasern im Bereich des Risses auf die beschriebene Streuung der Nachrissbiegezugfestigkeiten innerhalb einer Serie zu untersuchen, wurden Vorversuche an 6 Stahlfaserbetonbalken mit einem Fasergehalt von $20\,\mathrm{kg/m^3}$ durchgeführt. Diese Vorversuche unterstützen zudem die Entwicklungsarbeit des neuen Analysetools.

Nach der Ermittlung der Nachrissbiegezugfestigkeit wurden die Rissbereiche (ca. $100 \times 150 \times 150\,\mathrm{mm}$) herausgesägt und tomographiert. Dabei ergibt sich eine zu durchstrahlende Diagonale von mindestens $180\,\mathrm{mm}$. Daher wurden die Probekörper nochmals mittig geteilt, um optimale CT-Aufnahmen zu garantieren.

Tab. I zeigt anhand des Beispiels der beiden am stärksten von einander abweichenden Einzelprüfergebnisse einer Serie die Abhängigkeit des Nachrissverhaltens von der Faserorientierung. Trotz höherer Fasermenge in Probe 1 (Fig. 8) werden wesentlich geringere Nachrissbiegezugfestigkeiten gegenüber Probe 2 (Fig. 9) gemessen. Es zeigt sich, dass durch die stärkere Ausrichtung der Fasern in Zugkraftrichtung ($\varphi = \mathrm{x}$) bei Probe 2 höhere Nachrisszugfestigkeiten erreicht werden können.

Neben der automatischen Analyse der Faserorientierung und der Fasermenge, besteht die Möglichkeit aus den visualisierten Daten Parameter wie die Anzahl der gerissenen Fasern, der Einbindelänge der Fasern und der Verformung der Fasern zu bestimmen.

Für eine genauere Bewertung können die zuvor genannten Parameter mit einbezogen werden. So sind in Fig. 8 (Probe 1) Faseransammlungen und „stehende" Fasern zu erkennen, die eine optimale Kraftübertragung über den Riss hinweg behindern.

Weitere Arbeiten beschäftigen sich mit der Entwicklung von Methoden für die genaue Darstellung der Topographie von Rissen und deren Vermessung. Diese sollen dann im Bereich der stahlfaserverstärkten Ultahochfesten Betone zur Untersuchung der Mikrorissbildung eingesetzt werden.

tigkeit tomographiert werden. So ist es möglich, nicht nur einen einzelnen Rissbereich zu untersuchten, sondern es wird die Möglichkeit geboten bei einer Mehrfachrissbildung die gesamte Zugzone bezüglich Faserorientierung, Faserverteilung und Rissausbildung zu analysieren und zu bewerten.

Das für die Analyse zur Verfügung stehende Programm betrachtet für die Bestimmung der Faserorientierung und der Fasermenge die Fasergesamtheit. Schon rein visuell kann hier die Verteilung der Fasern in rekonstruierten 3D-Modellen betrachtet und bewertet werden.

Mit einem in der Entwicklung befindlichen Analysetool können zukünftig Fasern einzeln betrachtet werden (vgl. Abschnitt II-D). So wird es unter anderem möglich sein, die genauen Winkel zu ermitteln, unter denen eine Faser die Rissufer schneidet. Entstehende Faserauszugskanäle und Verformungen von Fasern können ebenfalls analysiert werden.

Gegenüber anderen Methoden, wie z. B. dem Aufbrechen des Risses, bleibt der zu untersuchende Bereich unbeschädigt, so dass auch eventuell gerissene Fasern detektiert werden können.

3) Nachrissbiegezugfestigkeit von Stahlfaserbeton: Zur Klassifizierung von Stahlfaserbetonen werden nach der Richtlinie „Stahlfaserbeton" des DAfStb [9] diese über die Nachrisszugfestigkeit, also die Leistungsfähigkeit im Nachrissbereich, in Leistungsklassen eingestuft.

Die Nachrisszugfestigkeit lässt sich unter zu Hilfenahme von Umrechnungsfaktoren aus der Nachrissbiegezugfestigkeit bestimmen. Diese Nachrissbiegezugfestigkeit wird in einem verformungsgesteuerten 4-Punkt-Biegeversuch an einer Serie aus mindestens 6 Balken der Abmessung $150 \times 150 \times 700\,\mathrm{mm}$ ermittelt. Die Eingangswerte für die Berechnung der Nachrissbiegezugfestigkeit bilden die zu den Durchbiegungen $0,5\,\mathrm{mm}$ und $3,5\,\mathrm{mm}$ zugehörigen Lasten. Bei der Bestimmung der Nachrissbiegezugfestigkeit werden in der Praxis starke Streuungen innerhalb von Prüfserien mit Variationskoeffizienten von über $25\,\%$ beobachtet.

In [10] wird die große Streuung zwischen den Messkurven der gleichen Betonzusammensetzung auf den großen Einfluss einzelner Fasern auf das Tragverhalten bei den vorliegenden kleinen Querschnittsabmessungen der Balken ($150 \times 150\,\mathrm{mm}$) zurückgeführt. Auch in [11] wird darauf hingewiesen, dass die großen Abweichungen auf eine inhomogene Verteilung

IV. FAZIT

Gerade bei steigender Druckfestigkeit der Betone ist das Nachbruchverhalten, sprich die Duktilität des Baustoffes nicht

Fig. 8. Probe 1, gekröpfte Fasern, betrachteter Probebereich: $30 \times 70 \times 70$ mm


Fig. 9. Probe 2, gekröpfte Fasern, betrachteter Probebereich: $30 \times 70 \times 70$ mm

um bestehende Bemessungsmodelle für aus diesen Betonen hergestellte Bauteile anzupassen bzw. erweitern zu können.

V. Danksagung

References

[1] F. Schuler and T. Sych, *Analyse der Faserorientierung in Betonen mit Hilfe der Computer-Tomographie*. Fraunhofer IRB Verlag, 2009.

[2] J. Schnell, K. Schladitz, and F. Schuler, *Richtungsanalyse von Fasern in Betonen auf Basis der Computer-Tomographie*, 2010, vol. 105, no. 2.

[3] R. Steinbrecher, *Bildverarbeitung in der Praxis*. München: Oldenbourg, 1993.

[4] Y.-Z. Lin, *Tragverhalten von Stahlfaserbeton*, ser. Deutscher Ausschuss für Stahlbeton. Berlin: Beuth, 1999, vol. 494.

[5] N. V. Tue, S. Henze, M. Küchler, G. Schenk, and K. Wille, *Ein optoanalytisches Verfahren zur Bestimmung der Faserverteilung und -orientierung in stahlfaserverstärktem UHFB*, 2007, vol. 102, no. 10.

[6] J. Ohser and K. Schladitz, *3d Images of Materials Structures – Processing and Analysis*. Wiley VCH, 2009.

[7] Fraunhofer Institut für Techno- und Wirtschaftsmathematik, *MAVI-Modular Algorithms for Volume Images*, 2005.

[8] H. Altendorf and D. Jeulin, *3d Directional Mathematical Morphology for Analysis of Fiber Orientations*, 2009, vol. 28, no. 3.

[9] Deutscher Ausschuss für Stahlbeton-DAfStb, *Stahlfaserbeton, Ergänzungen und Änderungen zu DIN 1045, Teile 1 bis 3 und DIN EN 206-1*. Berlin: Beuth, März 2010.

[10] M. Strack, *Modellbildung zum rissbreitenabhängigen Tragverhalten von Stahlfaserbeton unter Biegebeanspruchung: Univ., Diss.–Bochum, 2007.*, ser. Schriftenreihe des Instituts für Konstruktiven Ingenieurbau, Ruhr-Universität Bochum. Aachen: Shaker, 2007, vol. 2007,5.

[11] M. Teutsch, U. Wiens, and C. Alfes, *Stahlfaserbeton nach DAfStb-Richtlinie "Stahlfaserbeton"*, 2010, vol. 105, no. 8.

zufriedenstellend. Um diese geforderte Duktiliät bei ultrahochfesten Betonen (UHPC) zu erreichen, werden diesen Betonen zumeist Stahlfasern beigemischt.

Die Fähigkeit eines stahlfaserverstärkten Betons, Zugkräfte über einen Riss hinweg übertragen zu können, ist u.a. abhängig von der Orientierung, der Verteilung und der Anzahl der „rissüberbrückenden" Fasern im Bereich der sich ausbildenden Zugzone bzw. des sich einstellenden Risses bei Belastung.

Um weitere, das Nachbruchverhalten bestimmenden Parameter, wie die Verbundeigenschaften der Fasern mit der Matrix genauer bestimmen zu können, wurde ein Versuchsaufbau entwickelt, der die in-situ Beobachtung der Betonmikrostruktur und des Faserverhaltens bei Zugbeanspruchung des Probekörpers mithilfe der Computer-Tomographie ermöglicht. Die Ergebnisse dieser Versuche sollen neue Erkenntnisse über das Faserverbundverhalten und die (Mikro-) Rissbildung liefern,