



Stiftung
Werner-von-
Siemens-Ring

Zukunftsthemen · Emerging Technologies 2026

Zukunftsthemen · Emerging Technologies 2026

Neue Horizonte in Technik und Naturwissenschaft

Zukunftsthemen · Emerging Technologies 2026

Stiftung Werner-von-Siemens-Ring (Hrsg.)

Impressum

Stiftung Werner-von-Siemens-Ring
Steinplatz 1 – 10623 Berlin
www.siemens-ring.de

Vorsitzende des Stiftungsrats: Prof. Dr. Cornelia Denz
Geschäftsführer: Dr. Jan Fischer-Wolfarth
Stiftungsregister Berlin: AZ 3416 / 208 – II.2

Das Werk ist urheberrechtlich geschützt. Jede Verwertung außerhalb der engen Grenzen des Urheberrechtsgesetzes ist ohne Zustimmung des Verlags unzulässig und strafbar. Die Wiedergabe von Gebrauchsnamen, Handelsnamen, Warenbeschreibungen etc. berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Markenschutz-Gesetzgebung als frei zu betrachten waren und von jedermann benutzt werden dürfen. Aus der Veröffentlichung kann nicht geschlossen werden, dass die beschriebenen Lösungen frei von gewerblichen Schutzrechten (z. B. Patente, Gebrauchsmuster) sind. Eine Haftung des Verlags für die Richtigkeit und Brauchbarkeit der veröffentlichten Programme, Schaltungen und sonstigen Anordnungen oder Anleitungen sowie für die Richtigkeit des technischen Inhalts des Werks ist ausgeschlossen. Die gesetzlichen und behördlichen Vorschriften sowie die technischen Regeln (z. B. das VDE-Vorschriftenwerk) in ihren jeweils geltenden Fassungen sind unbedingt zu beachten.

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.dnb.de> abrufbar.

ISBN 978-3-8007-6755-7

© 2026 VDE VERLAG GMBH · Berlin · Offenbach
Bismarckstr. 33, 10625 Berlin

Alle Rechte vorbehalten.

Titelbild: © hichako/istock
Layout: Burga Fillery, bf-design.myportfolio.com
Druck: Salzland Druck GmbH & Co. KG, Staßfurt
Printed in Germany



Wer neu und anders denkt,
kann die Welt verändern.

WERNER VON SIEMENS, 1816–1892



W. Siemens



Inhalt

Vorwort	8
Einleitung	
Auszeichnung und Wertschätzung für herausragende Forschungstalente	10
Franziska Boenisch	
Privatsphäre und Künstliche Intelligenz (KI): Wie die KI lernt, was sie nicht verraten darf	14
Elena Jordan	
Skalierbare Ionenfallen mit integrierter Nanophotonik	26
Martha Kalina	
Ein Phasenfeldmodell für Ermüdungsrisse	42
Taha Soliman	
Architecting Efficiency for Sustainable Edge AI	54
Johannes Tiedau	
Von Atom- und Kernuhren	70
Das Netzwerk der Werner-von-Siemens-Fellows	
Impressionen	84

Vorwort

Eine innovationsstarke und zukunftsfähige Gesellschaft braucht exzellente Forschung und technologische Stärke. Diese Eigenschaften zeigen auch die fünf Werner-von-Siemens-Fellows des Jahres 2025.

In diesem Sammelband geben sie Einblicke in die Vielfalt ihrer unterschiedlichen, aber allesamt hoch zukunftsweisenden Forschungsbereiche. Sie entwickeln neue Ansätze für mehr Privatsphäre im maschinellen Lernen, bessere Prognosen von Ermüdungsbrüchen und innovative Anwendungen von Ionenfallen in der Nanophotonik. Sie arbeiten am Co-Design von Hard- und Software für alltagstaugliche KI und treiben die Präzision von Atomuhren immer weiter voran.

Seit mehr als einem Jahrhundert hat sich die Stiftung Werner-von-Siemens-Ring der Förderung herausragender Forschungstalente verschrieben. Seit 1916, dem hundertsten Geburtstag ihres Namensgebers Werner von Siemens, zeichnet sie mit dem Werner-von-Siemens-Ring außergewöhnliche Leistungen in den angewandten Natur- und Technikwissenschaften aus. Die Ausgezeichneten verbindet eine besondere Fähigkeit: Sie denken weiter, gehen neue Wege und transformieren bahnbrechende Forschung in unternehmerisches Engagement. Ganz im Sinne des erfolgreichen Unternehmers Werner von Siemens entstehen so Innovationen, die unseren Alltag nachhaltig verändern – vom Kühlschrank über moderne Fahrzeugtechnik und Computerchips bis hin zu höchstauflösenden Mikroskopen, Krebstherapien und Impfstoffen. Die Bandbreite dieser Beispiele zeigt eindrucksvoll, wie unmittelbar forschungsgetriebenes Unternehmertum unser Leben prägen kann.

1977 erweiterte die Stiftung ihr Wirken, um neben den großen Namen aus Forschung und Unternehmertum fortan auch junge, vielversprechende Talente zu würdigen – als Werner-von-Siemens-Fellows. Sie werden in ein starkes Netzwerk aufgenommen, und profitieren so vom Austausch mit anderen Fellows sowie mit den Preisträgerinnen und Preisträgern des

Werner-von-Siemens-Rings. Seit 2023 ist die Auszeichnung zudem mit einem Preisgeld von 10.000 Euro verbunden.

Im Mittelpunkt steht immer der Netzwerkgedanke. Gegenseitige Besuche ermöglichen den Werner-von-Siemens-Fellows einen engen und im Rahmen der Stiftung exklusiven Austausch. Preisverleihungen und Diskussionsformate bringen Forschende verschiedener Disziplinen und Institutionen generationenübergreifend zusammen. Die beste Inspiration liefert dabei Werner von Siemens selbst mit seinem Gedanken: „Wer neu und anders denkt, kann die Welt verändern.“ Einen Einblick in dieses interdisziplinäre Netzwerken gibt die kleine Fotostrecke am Ende dieses Bandes.

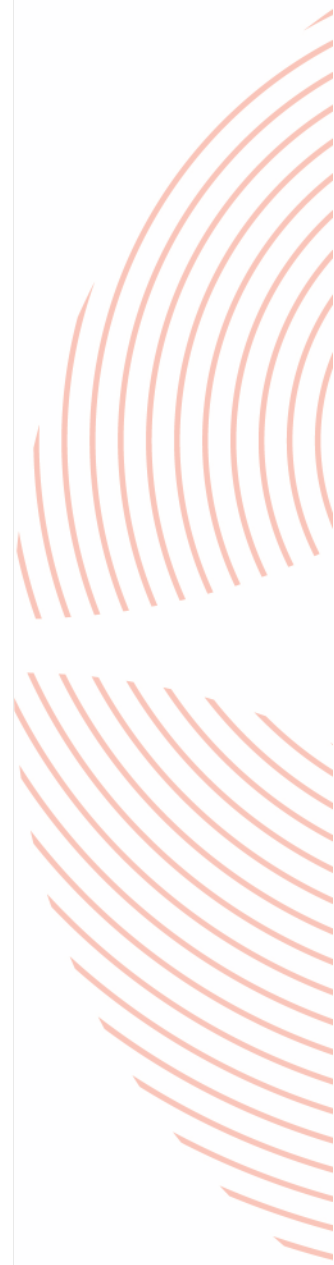
Das vielfältige Wirken der Stiftung Werner-von-Siemens-Ring lebt vom Engagement vieler Menschen. Die Mitglieder des Stiftungsrats bringen ihre Zeit, ihre Erfahrung und ihre Netzwerke ein und tragen die Ziele der Stiftung mit großer Überzeugung nach außen. Ebenso wichtig sind unsere Unterstützerinnen und Unterstützer, die die Arbeit der Stiftung mit großzügigen Spenden ermöglichen und dazu beitragen, jungen Forschungstalenten eine Bühne und ein starkes Netzwerk zu geben. Ihnen allen gilt mein herzlicher Dank.

Ich wünsche Ihnen eine anregende Lektüre mit spannenden Einblicken in die Ideen, die die Werner-von-Siemens-Fellows 2025 bewegen. Lassen Sie sich gemeinsam mit uns von ihrer Neugier, ihrem Forschergeist und ihrem Mut zu neuen Wegen inspirieren!

Prof. Dr. Cornelia Denz

Vorsitzende des Stiftungsrats

Präsidentin der Physikalisch-Technischen Bundesanstalt



Einleitung

Ausgezeichnete Forschung sichert unsere Zukunft

Auch wenn es aktuell kaum vorstellbar scheint: Die technologische Entwicklung beschleunigt sich immer weiter – basierend vor allem auf künstlicher Intelligenz und ihren Anwendungen in nahezu allen Bereichen des beruflichen und privaten Alltags. Technik verändert unsere Welt rasant, und die Menschen verändern sich mit. In Zeiten vieler Unsicherheiten ist es umso wichtiger, nicht nur handlungsfähig zu bleiben, sondern den Wandel zu gestalten. Dies ist nicht zuletzt die Aufgabe von Forschenden, die ihre Erkenntnisse aus der Wissenschaft teilen, Entscheidungsgrundlagen liefern und so Fortschritt mit Nutzen für die Menschen in die breite Anwendung bringen.

Nur mit Hilfe von exzellenter Forschung können wir technologische Antworten auf die großen Fragen unserer Zeit finden: Wie schützen wir unsere Privatsphäre und unsere digitale Souveränität im Zeitalter lernender Maschinen? Wie nutzen wir das Potenzial der Quantentechnologie für Anwendungen im Alltag? Wie entwickeln wir Werkstoffe, die langlebiger und sicherer sind und gleichzeitig Ressourcen schonen? Wie schaffen wir KI-Systeme, die energieeffizient und damit umweltverträglich sind?

Die Stiftung Werner-von-Siemens-Ring ehrt seit fast fünfzig Jahren Forschungstalente, die sich zwar gerade erst aufmachen, ihre Forschungswege zu gehen und dabei bereits herausragende und vielversprechende Ergebnisse für künftige Anwendungen erzielt haben. Seit 2023 trägt die Auszeichnung den Namen Werner-von-Siemens-Fellows – ein Begriff, der die Zugehörigkeit zu einem wachsenden Netzwerk exzellenter Wissenschaftlerinnen und Wissenschaftler unterstreicht.

Die drängenden Fragen unserer Zeit erfordern Forschende, die nicht nur Ideen entwickeln, sondern diese auch mit Ausdauer und unternehmerischem Mut in die Praxis überführen.

Der aktuelle Jahrgang der Werner-von-Siemens-Fellows, ausgezeichnet von der Stiftung Werner-von-Siemens-Ring, zeigt die Vielfalt möglicher Antworten auf einige dieser Herausforderungen. Was sie verbindet, ist ihre Leidenschaft für Forschung und die Überzeugung, dass sie mit Technologien wirklich etwas verändern können.

Das Wirken der Werner-von-Siemens-Fellows zeigt: Innovation aus Deutschland bleibt Maßstab und Gütesiegel. Die Werner-von-Siemens-Fellows leisten damit in doppelter Hinsicht wichtige Beiträge: zum wissenschaftlichen Fortschritt einerseits und andererseits zur Stärkung Deutschlands als wettbewerbsfähiger Wirtschaftsstandort.



Wandel gestalten heißt
Verantwortung übernehmen.

JAN FISCHER-WOLFARTH



Doch der Weg in die Spitzenforschung ist nicht immer einfach. Umso wichtiger ist es, diese Talente kontinuierlich zu fördern, von der Schule über das Studium bis in die wissenschaftliche Laufbahn. Genau dieses Potenzial zu entfalten, ist ein zentrales Anliegen der Stiftung Werner-von-Siemens-Ring.

Die Erfolgsgeschichten der Werner-von-Siemens-Fellows 2025

Die Stiftung zeichnete am 12. Dezember 2025 bei TRUMPF in Ditzingen – einem derjenigen Unternehmen, die zuletzt einen Preisträger des Werner-von-Siemens-Ringes aus ihren Reihen stellten – die Arbeiten der folgenden fünf Forscher:innen aus:

Dr. Franziska Boenisch: Individualisierte Privatsphäre im maschinellen Lernen

CISPA Helmholtz-Zentrum für Informationssicherheit, Saarbrücken

Dr. Franziska Boenisch wurde als Werner-von-Siemens-Fellow ausgezeichnet für die Entwicklung eines neuen Paradigmas im Bereich des Datenschutzes für KI-Modelle: der individualisierten Privatsphäre. Mit ihrem Ansatz eines flexiblen, nutzerzentrierten Verständnisses von Differential Privacy – einem mathematischen Konzept zum Schutz der Privatsphäre bei der Veröffentlichung



Unser Ziel ist nicht, ein riesiges System zu programmieren, sondern das System bestmöglich zu nutzen, um unsere Fragestellungen zu bearbeiten.

FRANZISKA BOENISCH



statistischer Informationen – ermöglicht sie es erstmals, individuelle Datenschutzpräferenzen von Patient:innen in das Training großer Sprachmodelle (LLMs) einzubeziehen. Ihre Arbeit stellt einen bedeutenden Fortschritt insbesondere für den Einsatz von KI im Gesundheitswesen dar, wo Vertrauen und Datensouveränität entscheidend sind.



Dr. Elena Jordan:

Ionenfallen mit integrierter Nanophotonik
 Physikalisch-Technische Bundesanstalt (PTB),
 Institut für Experimentelle
 Quantenmetrologie (QUEST), Braunschweig

Dr. Elena Jordan wurde für ihre Pionierarbeit auf dem Gebiet der mobilen Quantensensorik ausgezeichnet. Sie entwickelt skalierbare Ionenfallen mit integrierter Nanophotonik – einer Schlüsseltechnologie für Quantensensoren und Quantencomputer. Ihr innovatives Fallendesign wurde bereits patentiert und erfüllt höchste Anforderungen in der fundamentalen Physik und der Frequenzmetrologie. Damit schafft Elena Jordan die Voraussetzungen für portable Quantensensoren und den Einsatz von Quantentechnologien außerhalb spezialisierter Labore. Dies markiert einen entscheidenden Schritt hin zu einer breiten gesellschaftlichen Nutzung der Quantentechnologie.

Dr.-Ing. Martha Kalina:

Simulation von Ermüdungsbrüchen
 Technische Universität Dresden,
 Fakultät Maschinenwesen

Dr.-Ing. Martha Kalina wurde für ihre Forschung an der Schnittstelle von Maschinenbau, Werkstoffmechanik und numerischer Mathematik ausgezeichnet. Mit ihrem Phasenfeldmodell zur Simulation von Rissinitiierung und -fortschritt entwickelt sie neue Ansätze zur Lebensdauerprognose moderner Werkstoffe. Die Methode erlaubt erstmals eine skalierbare und mikroskopisch präzise Analyse von Ermüdungsschäden – ein wichtiger Beitrag zur Entwicklung nachhaltiger, langlebiger Materialien für Maschinen und Bauteile. Ihre Arbeit verbindet grundlegende Wissenschaft mit konkretem Nutzen für Industrie und Gesellschaft.

Dr. Johannes Tiedau:

Grundlagen für die erste optische Kernuhr
 Physikalisch-Technische Bundesanstalt (PTB),
 Fachbereich Zeit und Frequenz, Braunschweig

Dr. Johannes Tiedau erhielt die Auszeichnung für einen wissenschaftlichen Meilenstein auf dem Weg zur ersten optischen Kernuhr: Gemeinsam mit Kolleg:innen der Technischen Universität Wien gelang es ihm, mithilfe eines Lasers den Kernzustand im Isotop Thorium-229 zu verändern. Die neuartige Uhren-



Mich treibt die Frage an:
Was kann man Neues finden,
das noch nie jemand heraus-
gefunden hat?"

JOHANNES TIEDAU



generation hat das Potenzial, die Genauigkeit aktueller optischer Atomuhren nochmals deutlich zu übertreffen und eröffnet neue Wege für fundamentale Tests der Physik sowie hochpräzise Zeitmessung. Seine Forschung verbindet fundamentale Grundlagenforschung mit konkreten Anwendungsperspektiven in Metrologie und Navigation.

Dr. Taha Soliman:
KI mit Co-Design von Hard- und Software
Bosch Research, Renningen

Dr. Taha Soliman wurde als Werner-von-Siemens-Fellow für seine Entwicklung innovativer Hardware-Architekturen und Optimierungsmethoden für effiziente eingebettete KI-Systeme ausgezeichnet. Durch die kombinierte Weiterentwicklung von Hard- und Software reduziert er Latenz und Energieverbrauch, wodurch KI-Anwendungen alltagstauglich werden. Seine Forschung ermöglicht den ressourcenschonenden Betrieb komplexer neuronaler Netze und fördert so die Umweltverträglichkeit und Energieeffizienz. Damit macht er KI zu einer Technologie, die auch im eingebetteten Kontext zuverlässig und nachhaltig eingesetzt werden kann.

Diese fünf Persönlichkeiten, die Werner-von-Siemens-Fellows 2025, die auf diesem Weg ihre Forschungsarbeit vorstellen, repräsentieren eine Generation von Forschenden, die sich den großen Fragen unserer Zeit stellt und Wandel nicht als Bedrohung, sondern als Gestaltungsaufgabe begreift.

Welche Umbrüche noch vor uns liegen und wie wir ihnen begegnen, lässt sich zwar nicht abschließend sagen. Doch eines steht fest: Wir leben in einer Zeit voller Möglichkeiten. Und die Technikwissenschaften werden diesen Weg maßgeblich mitprägen.

Dr. Jan Fischer-Wolfarth
Geschäftsführer
der Stiftung Werner-von-Siemens-Ring



Privatsphäre und Künstliche Intelligenz



Dr. Franziska Boenisch ist Tenure-Track-Faculty am CISPA Helmholtz Center for Information Security in Saarbrücken und leitet das Sprint-ML-Lab für Trustworthy Machine Learning. Zuvor war sie Postdoctoral Fellow an der University of Toronto und am Vector Institute bei Prof. Nicolas Papernot. Ihre Forschung konzentriert sich auf private und vertrauenswürdige KI, insbesondere auf Privatsphäre-Risiken in Foundation Models und auf Schutzmechanismen mit formalen Garantien. Promoviert wurde sie an der Freien Universität Berlin (summa cum laude); während der Promotion arbeitete sie zudem als wissenschaftliche Mitarbeiterin am Fraunhofer AISEC. Für ihre Arbeiten erhielt sie u. a. den Fraunhofer TALENTA Grant, den Preis der Deutschen Industrieforschungsgemeinschaft, den Fraunhofer ICT Dissertation Award (2023), wurde GI-Junior-Fellow (2024) sowie Werner-von-Siemens-Fellow (2025) und erhielt 2025 einen ERC Starting Grant zum Thema „Privacy in Foundation Models“.

Franziska Boenisch

Privatsphäre und Künstliche Intelligenz (KI): Wie die KI lernt, was sie nicht verraten darf

Als ich zum ersten Mal erlebt habe, wie sich aus einem trainierten KI-Modell konkrete Trainingsdaten rekonstruieren lassen, hatte ich zwei Gedanken zugleich. Der erste war technisch: Das kann nicht nur ein Randphänomen sein, dafür war der Effekt zu klar. Der zweite war gesellschaftlich: Wenn wir solche Systeme in Medizin, Finanzen, Verwaltung oder Bildung einsetzen, dann reden wir nicht über abstrakte Risiken, sondern über das reale Leben einzelner Menschen. Moderne KI-Systeme sind heute in der Lage, Texte zu schreiben, Bilder zu erzeugen und Entscheidungen zu unterstützen, die früher mensch-

licher Expertise vorbehalten waren. Dieser Fortschritt ist beeindruckend und nützlich. Aber er basiert auf Daten, die oft zutiefst persönlich sind. Genau hier beginnt meine Forschung: bei der Frage, wie wir KI, und insbesondere maschinelles Lernen, so gestalten, dass es leistungsfähig bleibt und gleichzeitig die Privatsphäre der Individuen schützt, deren Daten für das Training benutzt werden.



Viele denken: Wenn die Daten nicht im Output stehen, ist alles sicher. Aber Modelle können Informationen indirekt verraten – und genau da wird's gefährlich.

FRANZISKA BOENISCH



Zusammenfassung:

Moderne KI-Systeme wie Sprachmodelle oder Bildgeneratoren werden auf riesigen Datenmengen trainiert und können dabei unbeabsichtigt persönliche Informationen „mitlernen“. In ungünstigen Fällen leaken diese Trainingsdaten später wieder aus dem Modell heraus. Franziska Boenisch erforscht, warum solche Datenlecks entstehen, wie Angriffe in realistischen Szenarien funktionieren und wie sich Datenschutz mit messbaren Garantien umsetzen lässt. Eine Pionierleistung ihrer Forschung ist dabei die „individualisierte Privatsphäre für KI“: Statt alle Daten gleich zu behandeln, wie es bei vorherigen Garantien der Fall war, werden besonders sensible Datenpunkte stärker geschützt. Diese Fragestellungen und dieser Ansatz bilden auch die Grundlage für Franziska Boenischs ERC Starting Grant-Projekt Privacy4FMs, in dem sie Methoden entwickelt, die Privatsphäre über den gesamten Lebenszyklus von Foundation Models absichern und audittierbar machen.

Abbildung 1

Heutzutage generieren wir alle große Mengen an sensiblen Daten, z. B. Fotos, die wir mit unseren Handys machen oder Bilder, die wir an Freunde und Familie verschicken. Viele dieser Daten enden in KI-Modellen von Firmen, die uns damit bessere Services anbieten wollen, wie Eingabehilfen beim Schreiben oder Gesichtserkennung in Fotos.

Alltägliche Generierung von Daten, die in KI-Modellen landen**Die falsche Hoffnung auf Privatsphäre**

Privatsphäre in maschinellern Lernen wird häufig auf eine einfache Vorstellung reduziert. Die Daten liegen im Training, das Modell wird anschließend veröffentlicht oder als Dienst angeboten, und solange niemand direkt in die Trainingsdaten schauen kann, sei alles in Ordnung. In der Praxis ist Privatsphäre jedoch viel komplizierter: Ein trainiertes KI-Modell kann Informationen über seine Trainingsdaten auf verschiedene Weise nach außen tragen: über seine Antworten, über sein Verhalten bei bestimmten Eingaben, oder darüber, was im Modell passiert, wenn wir es auf bestimmte Weise abfragen. Und das kann ein signifikantes Risiko für die

Privatsphäre der Trainingsdaten darstellen. Der Alltag liefert dafür greifbare Bilder. Man stelle sich ein Modell vor, das im Krankenhaus eingesetzt wird, etwa um Ärztinnen und Ärzte bei der Befundung zu unterstützen. Es wird auf historischen Daten von bisherigen Patient:innen trainiert und später für Voraussagen auf neuen Patient:innen genutzt. Wenn dieses Modell, direkt oder über Umwege, Informationen aus einzelnen Patientenakten preisgibt, dann ist der Schaden nicht auf einen technischen Metrikwert begrenzt. Es geht um die Offenlegung von Diagnosen, um Stigmatisierung, um Diskriminierung im Job oder bei Versicherungen – und um das Gefühl, die eigene Geschichte nicht mehr kontrollieren zu können. Ähn-

Beispiele für stark und weniger stark memorierte Trainingsdaten

a: stark memoriert	b: nicht memoriert
 An item of a vase with something inside of it.	 A girl in pink sweater putting a blue umbrella over a yellow fire hydrant
 Two males and a female in a red top holding some flowers.	 Close-up of two street signs reading "Airport Pkwy" and "Kamill Ave"
 A woman opening the trunk of her car.	 A sign with oriental writing on the words saying "Hyatt" on the Bund.
 A sleeping toddler lying on a woman's shoulder.	 A pink Hello Kitty microwave on a store shelf.
 An open point of view of a room with various things all around.	 A sign saying "Don't Honk, \$350 Penalty" on a pole.

Abbildung 2

Beispiele von den Trainingsdaten, die wir in meiner Forschung als stark memoriert (links) und wenig memoriert (rechts) identifiziert haben. Im Gegensatz zu vorheriger Forschung, die zeigte, dass immer ungewöhnlich aussehende Datenpunkte memoriert werden, hat unsere Forschung gezeigt, dass in neuen multi-modalen Modellen auch „normal“ aussehende Daten memoriert werden, wenn es ein Mismatch zwischen den Modalitäten gibt, also z. B. der Text nicht zum Bild passt. Das birgt ein neues Risiko für Datenpunkte.

liche Szenarien entstehen bei Sprachmodellen, die mit Kommunikationsdaten mit Individuen oder mit privaten E-Mails in Berührung kommen, oder bei Bildgeneratoren, die aus großen Sammlungen privater Fotos lernen. In der Forschung ist dieses Problem inzwischen bekannt, etwa durch Beispiele, bei denen Modelle Trainingsdaten nahezu wörtlich ausgeben oder Bilder erzeugen, die auffällig nah an einzelnen Trainingsbeispielen liegen.

Doch bekannt heißt noch nicht gelöst. Denn sobald man versucht, das Problem in der Breite zu adressieren, stößt man auf drei Engpässe, die meine Arbeit strukturieren: Wir müssen die Ursachen verstehen, wir

müssen den Bedrohungsraum realistisch erfassen, und wir müssen Schutzmechanismen bauen, die starke Garantien geben und trotzdem nützlich bleiben.

Die drei Säulen meiner Forschung: Memorisation, Angreifbarkeit und lieferbarer Schutz

Der erste Teil beginnt mit einer scheinbar einfachen Frage: Wie kann es dazu kommen, dass Modelle einzelne Informationen ihrer individuellen Trainingsdaten genau behalten und was können wir dagegen machen? Wir nennen dieses Phänomen in der Forschung Memorisation. In klassischen,

Ansatzpunkte für Leaks von privaten Daten im Modelltraining

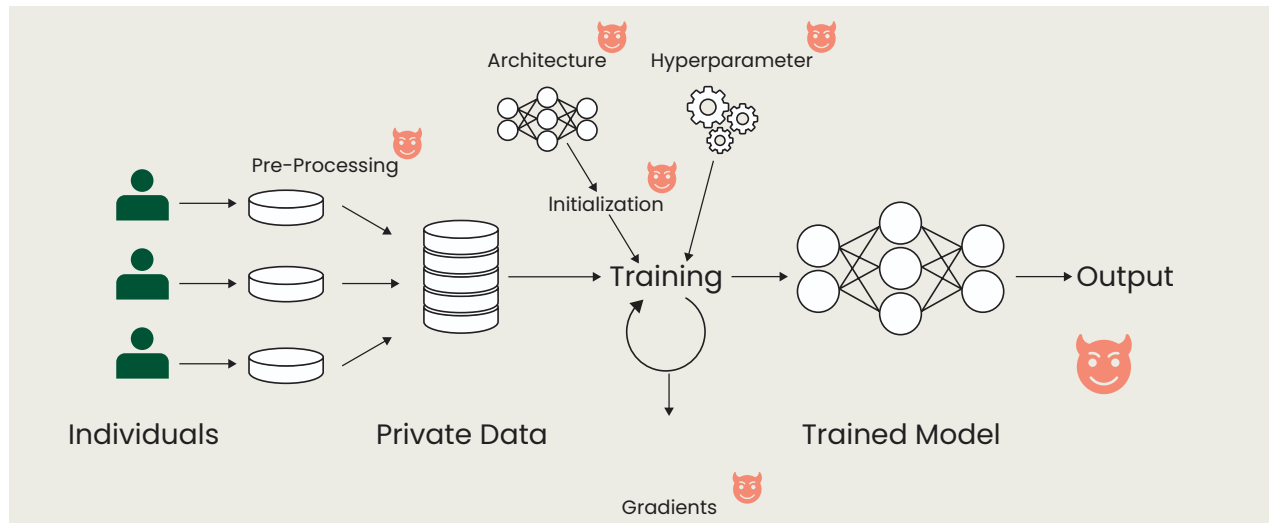


Abbildung 3

In den Trainingspipelines von KI-Modellen können private Informationen über viele Kanäle leaken, nicht nur über die Modell-Ausgaben. Auch Entscheidungen zum Setup des Modells, wie der Modellarchitektur und -parameterwahl können private Informationen preisgeben. Darüber hinaus können private Daten schon über die Gradienten im Training leaken.

überwachten Lernsettings, also Modellen, die Vorhersagen in Form eines Labels machen, z.B. „krank“ oder „gesund“, konnte man Memorisation oft daran festmachen, ob ein Modell ein Trainingsbeispiel besonders sicher korrekt klassifiziert. Bei modernen und insbesondere generativen Systemen ist das zu kurz gegriffen. Foundation Models, also große General-Purpose-Modelle wie z.B. GPT oder Claude erzeugen Texte, Bilder oder multimodale Ausgaben. Sie werden selbstüberwacht trainiert, sie optimieren komplexe Ziele, und sie lernen aus Daten, deren Struktur sich fundamental von kuratierten Datensätzen unterscheidet. In diesem Umfeld mit den komplexen Ausgaben reicht diese Betrachtung nicht mehr aus. Man braucht neue Metriken, die fein genug

sind, um Unterschiede zwischen Beispielen sichtbar zu machen, und formal genug, um daraus Mechanismen und Gegenmaßnahmen abzuleiten.

Ein prägnantes Beispiel dafür sind multimodale Vision Language Modelle wie CLIP, die Bilder und Texte in einem gemeinsamen Raum verankern. Hier habe ich mit entwickelt, wie sich Memorisation in einer Weise messen lässt, die über die Einzelmodalitäten hinausgeht^[6]. Dabei zeigte sich ein Muster, das man ohne eine passende Messung leicht übersehen hätte: Besonders stark memoriert werden nicht unbedingt die offensichtlich ungewöhnlichen Beispiele, sondern oft die falsch gepaarten, also solche, bei denen Bild und Text jeweils für sich unauffällig sind,

aber die Zuordnung nicht stimmt^[6]. Diese Beobachtung ist aus zwei Gründen wichtig. Erstens zeigt sie, dass Memorisation in solchen Modellen durch das Zusammenspiel der Modalitäten getrieben sein kann. Zweitens eröffnet sie eine Form von Mitigation, die im Datenschutz selten ist: Wenn man diese Fehlpaarungen erkennt und korrigiert, kann man Memorisation reduzieren und zugleich die Modellqualität verbessern, weil das Modell weniger gezwungen ist, inkonsistente Beispiele durch Speichern statt Verstehen zu kompensieren^[6]. Für mich war das ein Schlüsselmoment, weil er zeigt, dass Privatsphäre nicht immer nur ein Bremsklotz sein muss. Sie kann – richtig verstanden – auch ein Weg zu besserer Modellqualität sein. (Abb. 2 → S. 17)

Ein weiteres Beispiel für Ursachenforschung betrifft generative Bildmodelle, insbesondere Diffusionsmodelle. In der Praxis gelten diese Modelle als robust und schwer zu durchschauen, weil ihre Generierung iterativ ist und viele Parameter umfasst. Klassische Gegenmaßnahmen gegen Datenlecks setzen daher oft am Großen an: Modell neu trainieren oder Ausgaben filtern. Beides ist teuer oder unpraktisch, besonders wenn Modelle offen veröffentlicht werden. In unserer Arbeit haben wir einen anderen Ansatz verfolgt und gezeigt, dass sich Memorisation in Diffusionsmodellen überraschend lokal messen lässt^[7]. Es gibt Konstellationen, in denen einzelne Neuronen oder kleine Parametergruppen der Gesamtmodelle eine überproportionale Rolle dabei spielen, dass ein bestimmter Trainingsdatenpunkt reproduzierbar wird^[7].

Wenn man diese Teile identifizieren kann, lässt sich die Leckage gezielt reduzieren, ohne das gesamte Modell zu zerstören oder für immer auf Laufzeitfilter angewiesen zu sein. Solche Ergebnisse sind nicht nur technisch interessant. Sie verändern auch die Art, wie man über Verantwortung nachdenkt. Denn sie legen nahe, dass man nicht immer zwischen kompletter Veröffentlichung und kompletter Kontrolle wählen muss, sondern dass es Zwischenräume gibt, in denen man Risiken konkret adressieren kann.

Der zweite Teil meiner Forschung rückt die Frage in den Mittelpunkt, wie angreifbar die Trainingsdaten in KI-Modellen sind und über welche Wege sie leaken können. In der Sicherheitsforschung gilt: Ein System ist nur so sicher wie das realistischste Bedrohungsmodell, das man ernst nimmt. In der Praxis werden Bedrohungen jedoch oft unterschätzt, nicht aus bösem Willen, sondern weil maschinelles Lernen neue Angriffsflächen bietet, die weniger bekannt sind. In einer Studie habe ich mit gezeigt, dass in der Praxis Sicherheits- und Datenschutzaspekte nicht systematisch berücksichtigt werden, selbst wenn ihre Systeme in sensiblen Kontexten stehen^[1]. Das ist verständlich, weil die Werkzeuge fehlen und weil viele Risiken kontraintuitiv sind. Genau deshalb arbeite ich daran, den Bedrohungsraum zu kartieren und greifbar zu machen.

(Abb. 3 → links)



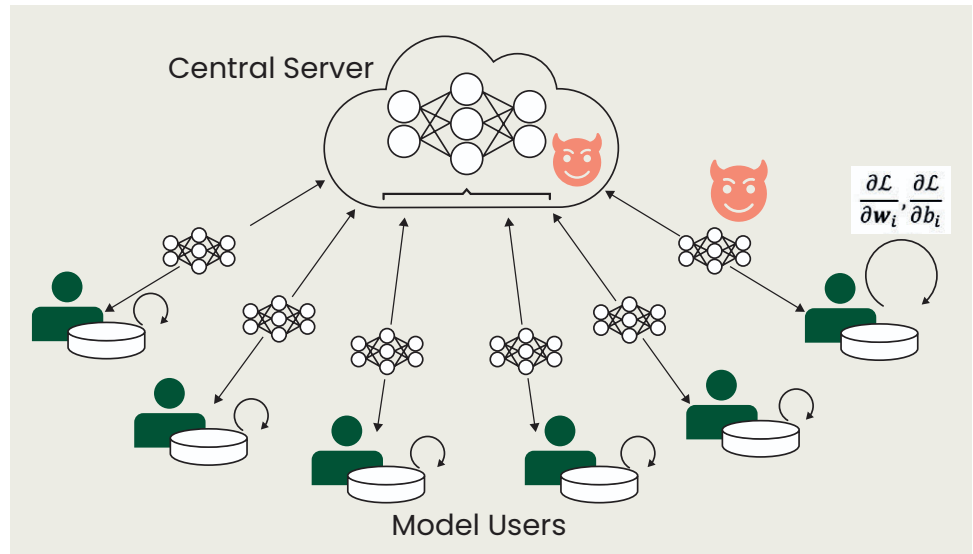
Foundation Models sind nicht nur größer. Sie werden auch anders trainiert – und deshalb funktionieren viele klassische Datenschutzrezepte nicht mehr.

FRANZISKA BOENISCH



Abbildung 4a

Das Setup für Federated Learning: ein zentraler Server initiiert ein KI-Modell und sendet es zum Training an verschiedene Nutzende. Diese trainieren das Modell lokal auf ihren Daten und senden nur Modellupdates an den Server. Weil die Daten die Nutzenden nie verlassen, hielt sich lange Zeit das Narrativ, dass Federated Learning Privatsphäre schafft. Modellupdates können jedoch individuelle Trainingsdatenpunkte direkt preisgeben.

Setup für Federated Learning Placement

Ein besonders prägnanter Fall ist Federated Learning. Die Grundidee klingt attraktiv: Daten bleiben auf dem Gerät, nur Modellupdates werden ausgetauscht. Daraus wurde in der Praxis häufig ein implizites Privatsphäreversprechen abgeleitet. In unserer Arbeit haben wir gezeigt, dass dieses Versprechen in vielen Fällen nicht hält^[3]. Gradienten und Updates können so viel Information enthalten, dass sich einzelne Trainingsdatenpunkte direkt rekonstruieren lassen, teilweise mit sehr geringem Rechenaufwand und ohne die Notwendigkeit, komplexe Modelle zu invertieren^[3]. Das allein ist bereits problematisch. Noch gravierender wird es, wenn man berücksichtigt, dass der koordi-

nierende Server nicht nur passiv empfängt, sondern aktiv das Training beeinflussen kann. Wir haben gezeigt, dass ein Server durch unauffällige Manipulationen das Ausmaß der Leakage erhöhen kann, ohne dass die Teilnehmenden es bemerken^[3]. Diese Ergebnisse haben eine klare praktische Botschaft: Federated Learning ist in vielen realen Konfigurationen eher datenminimierend als privat. Und wer es als Privatheitsgarantie verkauft, baut Vertrauen auf einem wackligen Fundament. (Abb. 4a → oben)

In Folgearbeit haben wir außerdem gezeigt, dass selbst gehärtete Varianten, etwa solche, die verteilte Differential Privacy und Secure Aggregation kombinieren, individuelle Ga-

Bei Attacke wiederhergestellte sowie nicht wiederhergestellte Datenpunkte



Abbildung 4b

Die Datenpunkte, die aus den Gradienten wiederhergestellt werden konnten. Alle Datenpunkte, über denen eine Null (**grün**) steht, sind nicht extrahiert worden von der Attacke. Alle anderen Datenpunkte (**koralle-rot**) sind direkt extrahierbar und die angegebene Anzahl zeigt an, wie oft sie im selben Experiment extrahiert wurden. Je höher die Anzahl, umso einfacher ist der Datenpunkt anzugreifen.

rantien verlieren können, wenn der Server zusätzliche kontrollierte Teilnehmende einschleust oder das Protokoll strategisch ausnutzt^[2]. Für mich ist daran die strukturelle Erkenntnis besonders wichtig: Privatsphäre ist nicht nur Mathematik, sie ist auch Machtbalance. Wenn eine Seite das Protokoll steuert und die andere Seite weder Einsicht noch Kontrolle hat, dann können formale Annahmen in der Praxis kippen. Diese Einsicht fließt heute in viele Debatten ein, in denen man sich fragt, welche Vertrauensmodelle wir in verteilten KI-Systemen überhaupt akzeptieren wollen. (Abb. 4b → oben)

Der dritte Teil meiner Forschung widmet sich der Frage, wie man Schutz nicht nur ver-

spricht, sondern wirklich liefert. Differential Privacy ist dabei ein zentraler Baustein, weil es als mathematischer Goldstandard eine klare Aussage ermöglicht: Der Beitrag eines einzelnen Datenpunkts zum Modellverhalten ist begrenzt. In der Anwendung ist Differential Privacy jedoch oft mit einem hohen Preis verbunden, weil der Schutz durch das Hinzufügen von statistischem Rauschen ins Training erkaufte wird und Rauschen Leistung kostet. Sprich, die privaten Modelle sind oft schlechter, was insbesondere in kritischen Domänen wie der Medizin ein Hindernis darstellt, da wir gerne weiterhin korrekte Vorhersagen machen wollen. Dieses Spannungsfeld wird häufig als unvermeidlich dargestellt. Ich glaube, dass es in vielen

Fällen weniger fundamental ist, als es wirkt, weil wir die falsche Abstraktion wählen. Ich habe dafür in meiner Forschung beobachtet, dass alle bisherigen Verfahren zum privaten Lernen alle Trainingsdaten gleich schützen mit einem einheitlichen Schutzniveau. Da dieses Schutzniveau alle Datenpunkte abdecken muss, richtete man sich automatisch nach dem strengsten, sensibelsten Fall und verschlechtert das Modell für alle, selbst wenn viele Datenpunkte weniger Schutz benötigen oder wenn die Risiken sehr ungleich verteilt sind.

Mein Ansatz: individualisierte Privatsphäre im maschinellen Lernen

Aus dieser Beobachtung heraus habe ich das Konzept individualisierter Privatsphäre in maschinellern Lernen mit geprägt. Die Idee ist einfach zu erklären: Nicht jedes Datum ist gleich sensibel, und nicht jede Person hat die gleichen Anforderungen. Ein medizinisches Bild kann eine andere Schutzstufe benötigen als ein generisches Foto. Ein Eintrag in einem Kommunikationsprotokoll kann höheres Risiko tragen als ein öffentlich verfügbares Faktum. Wenn man das ernst nimmt, ergibt sich ein neuer Gestaltungsraum: Man kann starke Garantien dort einsetzen, wo sie nötig sind, ohne das gesamte System auf den strengsten Wert zu zwingen. In meiner Arbeit habe ich dafür zentrale private Lernmechanismen erweitert, unter anderem

PATE und DP SGD, sodass sie pro Datenpunkt unterschiedliche Garantien ermöglichen und trotzdem formal beweisbar bleiben^{[5], [4]}. Das Ergebnis ist ein besserer Privacy Utility Ausgleich, weil das System Daten gezielter nutzen kann, ohne die sensibelsten Beiträge preiszugeben. (Abb. 5 → rechts)

Individualisierte Privatsphäre ermöglicht verschiedene Schutzniveaus für Trainingsdatenpunkte (Abb. 5): Durch meine individualisierte Privatsphäre im maschinellen Lernen können erstmalig alle Trainingsdatenpunkte auf einem individuellen Schutzniveau geschützt werden, welches ihren Ansprüchen oder Präferenzen entspricht. Dies setzt nicht nur die digitale Selbstbestimmung um, sondern ermöglicht es, bessere Modelle zu trainieren.

Mein Ziel: Privacy-by-Design in der KI

Wenn man diese drei Stränge zusammenführt, ergibt sich eine einheitliche Geschichte. Wir können Privatsphäre nicht durch einen einzigen Trick retten, weil die Ursachen vielfältig sind und die Systeme sich schnell verändern. Wir müssen erstens verstehen, wann und warum Modelle konkrete Datenpunkte behalten, also memorieren, und wie sich dieses Behalten in verschiedenen Modellfamilien äußert. Wir müssen zweitens realistisch modellieren, wie Angriffe tatsächlich stattfinden, und welche Daten besondere Risiken erleiden. Und wir müssen drittens Schutzmechanismen bauen, die nicht nur gut klingen, sondern messbare, überprüfbare Garantien geben, und die den

 Mich interessiert besonders, warum ein Modell etwas memoriert. Wenn wir die Mechanismen verstehen, können wir gezielter und günstiger schützen.

FRANZISKA BOENISCH



Individualisierte Privatsphäre ermöglicht verschiedene Schutzniveaus für Trainingsdatenpunkte

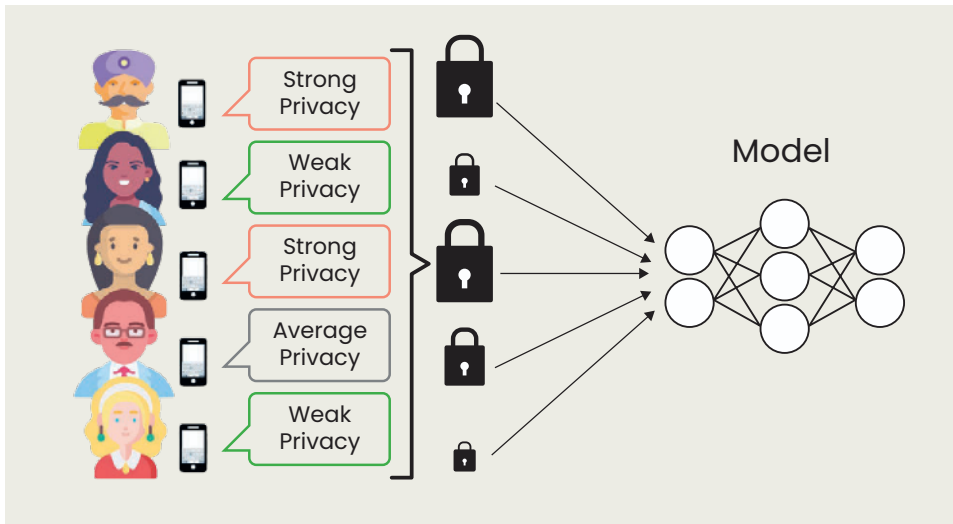


Abbildung 5

Durch meine individualisierte Privatsphäre im maschinellen Lernen können erstmalig alle Trainingsdatenpunkte auf einem individuellen Schutzniveau geschützt werden, welches ihren Ansprüchen oder Präferenzen entspricht. Dies setzt nicht nur die digitale Selbstbestimmung um, sondern ermöglicht es, bessere Modelle zu trainieren.

praktischen Anforderungen standhalten, etwa begrenzter Rechenleistung, begrenztem Zugriff auf Modelle oder heterogenen Schutzbedarfen.

Diese Überzeugung prägt auch meinen Blick auf die aktuelle Regulierungslandschaft. Die DSGVO und der EU AI Act formulieren Schutzziele, aber sie sagen nicht, welche Metriken, Audits und technischen Verfahren in komplexen KI Pipelines genügen. Genau an dieser Schnittstelle sehe ich eine zentrale Aufgabe der Forschung: Wir brauchen Methoden, die nicht nur in Laborumgebungen funktionieren, sondern als Grundlage für verantwortliche Entwicklung, Zertifizierung und Auditierung dienen können. In meiner

Arbeit an Privacy Risiken synthetischer Daten oder an Audits in anderen Kontexten zeigt sich immer wieder, wie wichtig es ist, technische Signale so aufzubereiten, dass sie in Entscheidungsprozessen nutzbar werden, und zwar ohne ihre Aussagekraft zu verlieren. Denn ein Audit, das nur Spezialistinnen verstehen, schützt niemanden. Ein Audit, das leicht verständlich ist, aber die falschen Dinge misst, schützt auch niemanden. Die Kunst liegt darin, beides zusammenzubringen.

Meine nächsten Schritte

Aus dieser Motivation heraus ist auch mein nächster Forschungsschritt entstanden, der

Privatsphäre ist nicht für alle gleich:
Ein MRT-Bild ist etwas anderes als
ein Katzenfoto. Individualisierte
Privatsphäregarantien sind für
mich der Schlüssel zu besseren
Modellen und mehr digitaler Selbst-
bestimmung.

FRANZISKA BOENISCH

sich in meinem ERC Starting Grant Privacy-4FMs widerspiegelt, den ich 2025 erhalten habe und der seit 2026 läuft. Der Kern dieses Projekts ist nicht, einen weiteren privaten Trainingsalgorithmus zu bauen, sondern das Problem entlang des Lebenszyklus von Foundation Models zu denken. Diese Modelle entstehen in einem Pretrain Adapt Paradigma, in dem Vortraining und Anpassung eng miteinander verwoben sind. Daten können sich überlappen, Verteilungen können sich ähneln, und Effekte aus dem Vortraining können die Anpassung so dominieren, dass formale Garantien für die Anpassungsdaten in der Praxis wenig aussagen, wenn man das Vortraining ignoriert. Das Projekt nimmt daher die

Interdependenzen ernst und verbindet Ursachenanalyse, Garantien und Audits zu einem Rahmen, der die reale Entwicklungspraxis widerspiegelt. Dabei spielt individualisierte Privatsphäre eine natürliche Rolle, weil in solchen Pipelines die Schutzbedarfe nicht nur von individuellen Präferenzen abhän-

gen, sondern auch vom Kontext, etwa davon, wie nahe ein Beispiel an der Vortrainingsverteilung liegt oder wie stark es durch Anpassungsschritte exponiert wird.

Am Ende läuft alles auf eine Haltung hinaus, die ich für zentral halte: Privatsphäre ist kein Feature, das man am Schluss dazu schaltet. Sie ist eine Eigenschaft, die man in die Architektur, die Trainingsprotokolle und die Evaluationskultur in der KI einbauen muss. Wenn wir das nicht tun, werden wir immer wieder überrascht sein von Datenlecks, die wir uns im Rückblick hätten erklären können. Wenn wir es tun, eröffnen sich neue Möglichkeiten, die über reines Risikomanagement hinausgehen. Dann können wir Systeme entwickeln, die leistungsfähig sind, weil sie sauberer lernen, die vertrauenswürdiger sind, weil sie überprüfbar bleiben, und die gesellschaftlich akzeptabler sind, weil sie den Schutz einzelner Menschen nicht als Kollateralschaden behandeln. Ich sehe meine Forschung als Beitrag zu genau diesem Ziel: eine KI, die nicht nur beeindruckend ist, sondern vertrauenswürdig, und eine Privatsphäre, die nicht nur ein Recht auf dem Papier ist, sondern eine technische Realität in den Modellen, die unseren Alltag prägen.

Literatur

- [1] Boenisch, Franziska, Verena Battis, Nicolas Buchmann, and Maija Poikela. “i never thought about securing my machine learning systems”: A study of security and privacy awareness of machine learning practitioners. In *Proceedings of Mensch und Computer 2021*, pp. 520–546. 2021.
- [2] Boenisch, Franziska, Adam Dziedzic, Roei Schuster, Ali Shahin Shamsabadi, Ilia Shumailov, and Nicolas Papernot. “Reconstructing individual data points in federated learning hardened with differential privacy and secure aggregation.” In *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*, pp. 241–257. IEEE, 2023.
- [3] Boenisch, Franziska, Adam Dziedzic, Roei Schuster, Ali Shahin Shamsabadi, Ilia Shumailov, and Nicolas Papernot. “When the curious abandon honesty: Federated learning is not private.” In *2023 IEEE 8th European Symposium on Security and Privacy (EuroS&P)*, pp. 175–199. IEEE, 2023.
- [4] Boenisch, Franziska, Christopher Mühl, Roy Rinberg, Jannis Ihrig, and Adam Dziedzic. “Individualized PATE: Differentially Private Machine Learning with Individual Privacy Guarantees.” *Proceedings on Privacy Enhancing Technologies* 1 (2023): 158–176.
- [5] Boenisch, Franziska, Christopher Mühl, Adam Dziedzic, Roy Rinberg, and Nicolas Papernot. “Have it your way: Individualized privacy assignment for DP-SGD.” *Advances in Neural Information Processing Systems* 36 (2023): 19073–19103.
- [6] Wang, Wenhao, Adam Dziedzic, Grace C. Kim, Michael Backes, and Franziska Boenisch. “Captured by captions: On memorization and its mitigation in CLIP models. In *The Thirteenth International Conference on Learning Representations* (2025)”.
- [7] Hintersdorf, Dominik, Lukas Struppek, Kristian Kersting, Adam Dziedzic, and Franziska Boenisch. “Finding nemo: Localizing neurons responsible for memorization in diffusion models.” *Advances in Neural Information Processing Systems* 37 (2024): 88236–88278.

Skalierbare Ionenfallen mit integrierter Nanophotonik



Dr. Elena Jordan ist Projekt- und Teamleiterin am QUEST Institut der Physikalisch-Technischen Bundesanstalt in Braunschweig. Sie forscht an der Integration von Nanophotonik in Ionenfallen und an Multi-Ensemble-Spektroskopie für optische Uhren. Aktuell betreut sie mehrere nationale und internationale Projekte. Ihre Doktorarbeit im Bereich Laserspektroskopie von negativen Lanthan-Ionen führte sie von 2012 bis 2015 am Max-Planck-Institut für Kernphysik in Heidelberg durch. Währenddessen war sie Mitglied der AEGIS-Kollaboration und arbeitete bei Strahlzeiten mit Antiprotonen am CERN mit. Anschließend forschte sie als Leopoldina Fellow am NIST in Boulder/Colorado, USA, an Quantensimulationen des Dicke-Modells und Quantensensing für ultra-schwache elektrische Felder mit großen zweidimensionalen Ionenkristallen in der Penningfalle. In dieser Zeit gelang es ihr, die Schwingungsbewegung großer Kristalle in der Penningfalle mit über 100 Ionen erstmalig nahe an den Grundzustand der Bewegung zu kühlen.

Elena Jordan

Skalierbare Ionenfallen mit integrierter Nanophotonik

Ionenfallen werden heute sehr erfolgreich für eine Vielzahl an Präzisionsmessungen und Quantentechnologien eingesetzt. Bereits lange vor der Erfindung der Ionenfallen in den 1950er-Jahren träumten Wissenschaftler:innen davon, Materieteilchen freischwebend, isoliert von der Umgebung einzufangen zu können, um sie ganz ungestört zu beobachten^[1]. Genau dazu werden Ionenfallen eingesetzt. In einer weitestgehend ungestörten Umgebung können präzise Messungen über Zeiträume von vielen Monaten durchgeführt werden. Quantenzustände können präpariert, gespeichert und manipuliert werden.

Das Gebiet der Nanophotonik entstand mit der Erfindung der Laser, optischer Fasern und lithografischer Methoden^[2]. Sie beschäftigt sich mit Licht in Nanostrukturen, das darin erzeugt, geleitet, manipuliert und detektiert werden kann. Die Anfänge der Photonik waren durch die Telekommunikation motiviert, aus der sie heute nicht mehr wegzudenken ist. Während man für die Telekommunikation auch mit Infrarot-Wellenlängen auskommt, benötigt man für die Quantentechnologien mit Ionen oder Atomen verschiedene Wellenlängen bis in den blauen und ultravioletten Spektralbereich.

Die Photonik ist aktuell ein sehr aktives Forschungsgebiet.

Ionenfallen mit integrierter Nanophotonik bieten die Möglichkeit, kompakte portable Systeme zu bauen, welche als Quantensensor im Feld angewendet werden können. Für die Skalierung der Quantentechnologien von wenigen auf hunderte oder tausende von Ionen oder Atomen mit individueller optischer Kontrolle ist integrierte Photonik unerlässlich.

Zusammenfassung:

Ionenfallen ermöglichen es, Ionen über lange Zeit isoliert von der Umgebung zu speichern und zu beobachten. Dies ermöglicht Messungen mit höchster Präzision. Gefangene Ionen werden daher in optischen Uhren als Frequenzreferenz und in ionenbasierten Quantencomputern für die Realisierung von Qubits eingesetzt. Zur Kühlung und zur gezielten Manipulation der Ionen werden Laser und Mikrowellen verwendet, deren Optiken bislang viel Platz benötigen und oft nachjustiert werden müssen. An der PTB entwickeln wir Ionenfallen mit integrierter Nanophotonik, die eine kompakte und robuste Alternative darstellen. Da die meisten Ionenspezies Übergänge im blauen oder UV-Wellenlängenbereich besitzen, muss die Photonik für diesen Wellenlängenbereich unter Verwendung von neuen Materialien entwickelt werden. Diese Technologie bietet neue Möglichkeiten für portable optische Uhren, die in der Erdbeobachtung eingesetzt werden können, oder für Quantencomputer mit einer großen Anzahl an Qubits.

Chip-basierte Ionenfalle der Physikalisch-Technischen Bundesanstalt aus AlN-Chips mit Goldbeschichtung

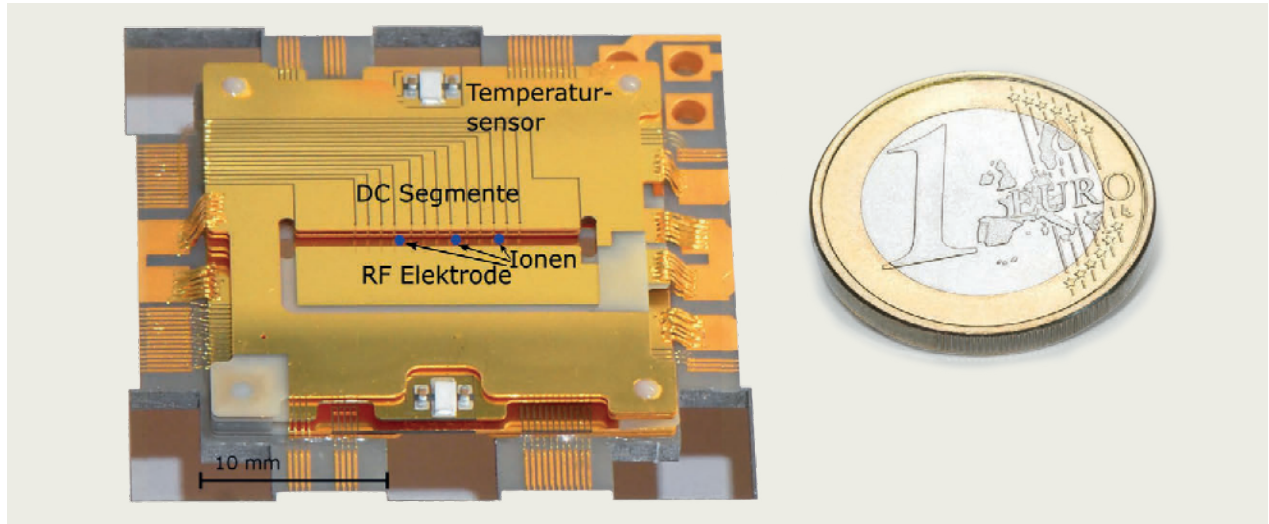


Abbildung 1

Die Ionen werden im Fallenschlitz gefangen, indem an der Radiofrequenz (RF)-Elektrode eine elektrische Wechselspannung angelegt wird.

An den Segmenten der Direct Current (DC)-Elektrode können Spannungen für das Einfangen entlang der Achse und zur Kompensation von Streufeldern angelegt werden.

Quelle: E. Jordan, M. Brinkmann, A. Didier, et al. *Quantum Sci. Technol.* 10 045005 (2025), Beschriftung hinzugefügt. CC-BY 4.0

Dabei bietet die Photonik viel mehr als nur eine Größenreduktion. Integrierte photonische Strukturen bieten vielfältige Möglichkeiten für die Gestaltung, die Eigenschaften und die Effizienz der Optik. Hinzu kommt, dass eine stabile Ausrichtung der Laserstrahlen auf die Ionen bereits mit der Fabrikation gegeben ist.

Im folgenden Text werden zunächst einige Grundlagen erklärt, z.B. wie Ionenfallen funktionieren und welche Rolle Laser im Zusammenhang mit gefangenen Ionen spielen, sowie die resultierenden Anforderungen an die Nanophotonik erläutert. Anschließend wird umrissen, wie wir an der Physikalisch-Technischen Bundesanstalt (PTB) Ionenfallen mit integrierter Photonik entwi-

ckeln. Zum Abschluss werden einige Anwendungen von Ionenfallen vorgestellt, wie optische Uhren und Quantencomputer. Daraus wird sicherlich schnell klar, weshalb skalierbare photonisch integrierte Ionenfallen ein Zukunftsthema sind.

Ionenfallen

Eine Ionenfalle ist, wie der Name schon sagt, eine Vorrichtung, mit der man ein einzelnes, mehrere oder viele Ionen einfangen kann. In der Falle sitzen die Ionen in einem tiefen Potenzial, gut isoliert von der Umgebung, und können so über viele Monate gespeichert werden. Voraussetzung ist ein guter

Größenvergleich einer Oberflächenfalle mit integrierter Photonik



Abbildung 2a

*Eine Oberflächenfalle
mit integrierter Photonik
im Vergleich zu einem
1-Cent-Stück*

Vakuumdruck in der Falle, also eine sehr geringe Stoßrate mit dem Restgas, und eine lange Halbwertszeit bei radioaktiven Ionen. Ein Beispiel einer Ionenfalle ist in *Abbildung 1* → links gezeigt.

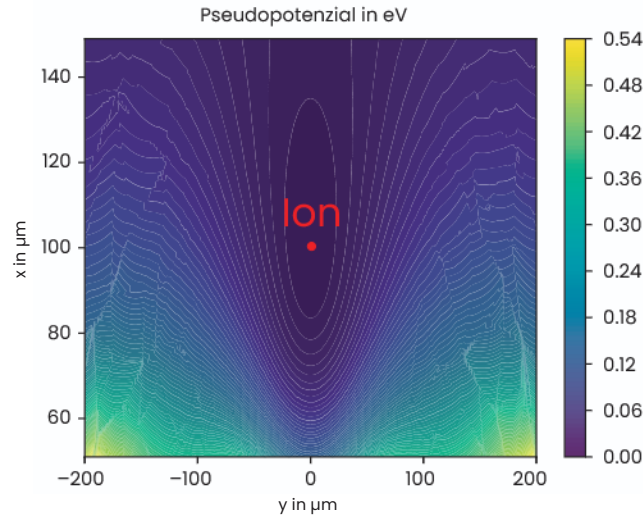
Da Ionen eine Ladung besitzen, erfahren sie eine Kraft in elektromagnetischen Feldern. Möchte man die Ionen an einer Koordinate oder auf einer Achse einschließen, benötigt man eine Kraft, die das Ion zur Einfangposition zurückbringt, also mit dem Abstand zur Fallenmitte zunimmt. Solch eine Kraft, die z.B. linear mit dem Abstand zunimmt, bekommt man mit einem harmonischen Potenzial in drei Dimensionen. Das entspricht elektrischen oder magnetischen Multipol-

feldern^[1]. Für das Fallenfeld findet man verschiedene Lösungen. Ein dreidimensionaler Einschluss der Ionen ist nicht mit statischen elektrischen Feldern möglich: Man benötigt hierfür ein zeitabhängiges Potenzial, wie in einer Paulfalle, oder ein magnetisches Feld wie in der Penningfalle.

In einer linearen Paulfalle werden die Ionen auf der Fallenachse gefangen. Dorthin werden sie mit elektrischen Wechselfeldern immer wieder zurückgeschoben. Entlang der Fallenachse sind sie durch ein statisches elektrisches Potenzial eingeschlossen. Für Präzisionsmessungen haben sich Chip-basierte Ionenfallen bewährt, die mit hoher Präzision gefertigt werden können^[3].

Abbildung 2b

FEM-Simulation des Fallenpotenzials der Oberflächenfalle in radialer Richtung; x ist die Höhe über der Oberfläche und y die Richtung senkrecht zur Fallenachse
eV: Elektronenvolt
 μm : Mikrometer
FEM: Finite-Elemente-Methode

FEM-Simulation des Fallenpotenzials

Dabei wird ein elektrisches Wechselfeld an den Radiofrequenzelektroden angelegt. Zusammen mit den geerdeten Elektroden erzeugen sie das Feld für den radialen Einschluss. Für den axialen Einschluss werden Spannungen auf den Segmenten entlang der Fallenachse angelegt. Für Ionen-basierte Quantencomputer werden häufig Oberflächenfallen eingesetzt, bei denen alle Elektroden in einer Ebene liegen. Sie sind einfach herzustellen und sehr flexibel im Design, was auch die Integration von Photonik erleichtert. Das mithilfe von Simulationen der Finite-Elemente-Methode (FEM) berechnete elektrische Feld einer Oberflächenfalle (Paulfalle) wird in *Abbildung 2b oben* gezeigt. Hier liegt das Ion in einer Mulde und ist nach oben nur schwach

eingeschlossen. Die Ionen bewegen sich im Fallenpotenzial – man könnte sagen, sie schwingen darin. Sitzen mehrere Ionen im gleichen Potenzial, sind ihre Bewegungen durch die Coulomb-Wechselwirkung gekoppelt. Sie spüren, wenn die „Nachbarn“ sich bewegen, und bewegen sich mit. Diese Kopplung ist sehr wichtig für Quantenlogikanwendungen. Wird die Bewegung der Ionen stark gekühlt, so ordnen sie sich aufgrund der Coulomb-Wechselwirkung in einem regelmäßigen Coulomb-Kristall an. Die Form des Kristalls hängt vom Fallenpotenzial ab. (*Abb. 2b oben*)

In der Penningfalle erfolgt der radiale Einschluss durch ein starkes homogenes Mag-

netfeld entlang der Fallenachse. Bewegen sich die Ionen weg von der Achse, erfahren sie eine Lorentzkraft senkrecht zur Bewegungsrichtung, die sie auf die Kreisbahn zwingt. Durch diese Zyklotronbewegung sind die Ionen radial eingeschlossen. Für den Einschluss entlang der Achse legt man ein statisches elektrisches Feld an. Die Penningfalle wurde von Hans Dehmelt ebenfalls in den 1950er-Jahren erfunden. Penningfallen werden heutzutage für hochpräzise Atommassenvergleiche und Antimaterieexperimente eingesetzt sowie für Quantensimulationen. Da für das starke und homogene Magnetfeld große, meist supraleitende Magnete eingesetzt werden, sind Paulfallen einfacher und kostengünstiger aufzubauen und werden öfter eingesetzt.

Ionenfallen geben den Physiker:innen viel Zeit, um ein oder mehrere Ionen zu beobachten, zu untersuchen und zu manipulieren. Hierfür werden oft Laser und Mikrowellen eingesetzt. Man kann Ionen durchaus auch ohne Laser fangen, jedoch spielen Laser und Mikrowellen für die meisten Anwendungen von Ionenfallen eine wichtige Rolle.

Das Laserkühlen ist eine Routine für Experimente mit gefangenen Ionen. Im Prinzip ist es das Abbremsen der Ionen mit dem Impuls der Laserphotonen. Die Ionen werden dabei aus einer Richtung angeregt, durch den Impulsübertrag abgebremst und emittieren die Photonen dann in eine zufällige Richtung. Mit dieser Technik lassen sich die Ionen in einer Falle bei Raumtemperatur innerhalb von wenigen Millisekunden auf Temperaturen weniger als 1°C über dem absoluten

Nullpunkt kühlen. Auch für das Kühlen nahe an den Grundzustand der Bewegung gibt es etablierte Methoden. Die vom Ion ausgesandten Photonen, die Fluoreszenz, kann man beobachten und als Indikator dafür nutzen, in welchem elektronischen Zustand sich das Ion befindet.

Mithilfe von Lasern oder Mikrowellen können optische Übergänge zwischen zwei elektronischen Zuständen in den Ionen angeregt werden. Dabei hängt die Übergangswahrscheinlichkeit von dem Abstand der Laserfrequenz zur Übergangsfrequenz, der Intensität des Laserstrahls und der Dauer des Einstrahlens ab.

Die meisten Experimente und Messaufbauten füllen optische Tische mit Lasern und Optik (Abb. 3 → S. 33). Die Laser werden über optische Fasern und Spiegel zum Experiment gebracht. Auf dem Weg dorthin durchlaufen sie Wellenplatten, Polarisationsfilter, Modulatoren und optische Schalter, die die Polarisation anpassen und die Phase oder Amplitude modulieren. All das nimmt Platz ein und muss immer wieder, normalerweise in Handarbeit, nachjustiert werden. Daher ist es nicht verwunderlich, dass es großes Interesse an Ionenfallen gibt, in welche die Laseroptik direkt integriert ist.



Das Einsatzgebiet von Ionenfallen umfasst optische Uhren, Quantencomputer, Quantenrepeater und die Grundlagenforschung.

ELENA JORDAN



Integrierte Nanophotonik

Nanophotonik gibt es bereits seit Mitte der 1980er-Jahre^[4]. Für Telekomwellenlängen, insbesondere 1550 nm, im Infrarot-Bereich sind neben günstigen Single-Mode-Fasern Faser-Arrays, Wellenleiter mit geringen Verlusten $< 0,1$ dB/cm, integrierte Fotodioden, integrierte Schalter, Polarisationsoptik und gute optische Gitter erhältlich. Mit einer Bandlücke von 1,12 Elektronenvolt ist Silizium transparent bis zu einer Wellenlänge von $1,1 \mu\text{m}$. Der hohe Brechungsindex von $n_{\text{Si}} = 3,47$ für 1550 nm macht effiziente und kompakte Strukturen möglich und durch die Synergie mit der Complementary-Metal-Oxide-Semiconductor-Technologie gibt es etablierte Prozesse für die Herstellung.

Für gefangene Ionen benötigt man Laser bei kürzeren Wellenlängen, für die meisten Spezies liegt eine Wellenlänge im blauen oder ultraviolett (UV)-Bereich. Hierfür gibt es keine so etablierte Materialplattform wie Silizium. Wellenleitermaterialien, wie Siliziumnitrid (Si_3N_4), Lithiumniobat (LiNbO_3), Alumina (Al_2O_3), Tantalpentoxid (Ta_2O_5) und Aluminiumnitrid (AlN), sind transparent für sichtbare und teils auch für UV-Wellenlängen. Die Herstellungsprozesse für verlustarme Wellenleiter sind teils noch in der Entwicklung und die Optiken, die in Silizium realisiert wurden, lassen sich nicht direkt auf diese Plattformen übertragen. Gründe dafür sind neben den unterschiedlichen

thermischen und optischen Eigenschaften auch die höheren Anforderungen an die photonischen Strukturen. Zum Beispiel nimmt der Streuverlust an rauen Seitenwänden mit kürzer werdender Wellenlänge stark zu (in der 4. Potenz $1/\lambda^4$), die benötigten Strukturgrößen sind kleiner und bringen so lithografische Prozesse an ihre Grenzen. Die Photonik in Si_3N_4 und LiNbO_3 sind am weitesten entwickelt und werden von einigen Chipherstellern angeboten. Mit beiden Materialien wurden optische Wellenleiter mit geringen Propagationsverlusten $< 0,1$ dB/cm (Dezibel pro cm) und optische Modulatoren demonstriert. Die Eigenschaften beider Materialien sind sehr unterschiedlich: Während Lithiumniobat wegen seiner nicht-linearen optischen Eigenschaften sehr beliebt ist, hat Siliziumnitrid andere Vorzüge, wie eine hohe Zerstörungsschwelle bei hoher Laserleistung und sehr geringen Propagationsverlusten, was optische Resonatoren mit einigen Metern Länge und hoher Güte aufgerollt auf einem Chip ermöglicht. Schalter werden in Siliziumnitrid meist durch Temperaturänderung realisiert und sind sehr langsam. Die aktiven Komponenten mit Si_3N_4 sind zum Teil noch in der Entwicklung^[5]. Auf die Frage, bis zu welcher Wellenlänge diese beiden Plattformen transparent sind, bekommt man von verschiedenen Herstellern sehr unterschiedliche Antworten. Einige sind sich die meisten, dass die Verluste unter 400 nm Wellenlänge sehr hoch sind.

Für UV-Photonik sind daher Tantalpentoxid, Alumina und Aluminiumnitrid von Interesse. Die Photonik für sichtbare und UV-Wellenlängen ist ebenso wie die Silizium-Photonik ein sehr aktuelles Forschungsgebiet.



Die Integration von Photonik wird die Skalierbarkeit von Ionenfallen entscheidend verbessern, daher wird auch in der Industrie intensiv daran geforscht.

ELENA JORDAN



Optischer Tisch im Uhrenlabor

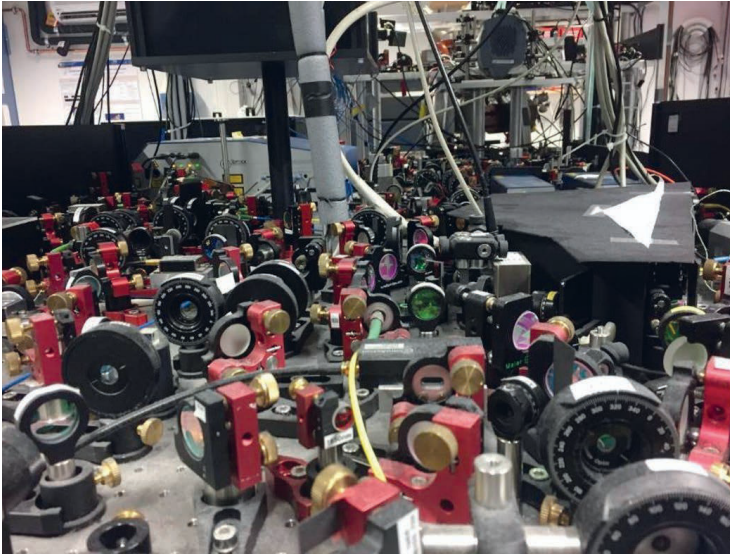


Abbildung 3

Optischer Tisch im Uhrenlabor von Prof. Tanja E. Mehlstäubler. Im Bild sind ein Laser, Spiegel und Polarisationsoptiken zu sehen.

Quelle: Tanja Mehlstäubler

Integrierte Ionenfallen

Integrierte Ionenfallen sind keine neue Erfindung. Vorreiter sind die MIT Lincoln Labs in Lexington/Massachusetts USA und die Sandia National Laboratories in Albuquerque/New Mexico USA. Sie stellen bereits seit Jahrzehnten integrierte Ionenfallen her, mit integrierten Wellenleitern, Auskopplern und Einzelphotonen-Detektoren. Kürzlich haben Forscher der Sandia National Laboratories eine optische Uhr basierend auf einer photonisch integrierten Ionenfalle demonstriert^[7].

Für eine integrierte Ionenfalle benötigt man einen Photonik-Chip, der auch als Photonic Integrated Circuit (PIC) bezeichnet wird. Für

eine monolithisch integrierte Falle werden die Fallenelektroden direkt auf die Oberfläche des PICs aufgebracht und strukturiert. Meist werden Elektroden aus Gold verwendet, da es keine Oxidschicht bildet und eine hohe elektrische Leitfähigkeit besitzt. Elektroden aus Aluminium, Titannitrid oder Niob werden auch verwendet. Neben dem Fallenchip an sich sind auch optische Fasern für die Zuleitung und elektrische Anschlüsse nötig.

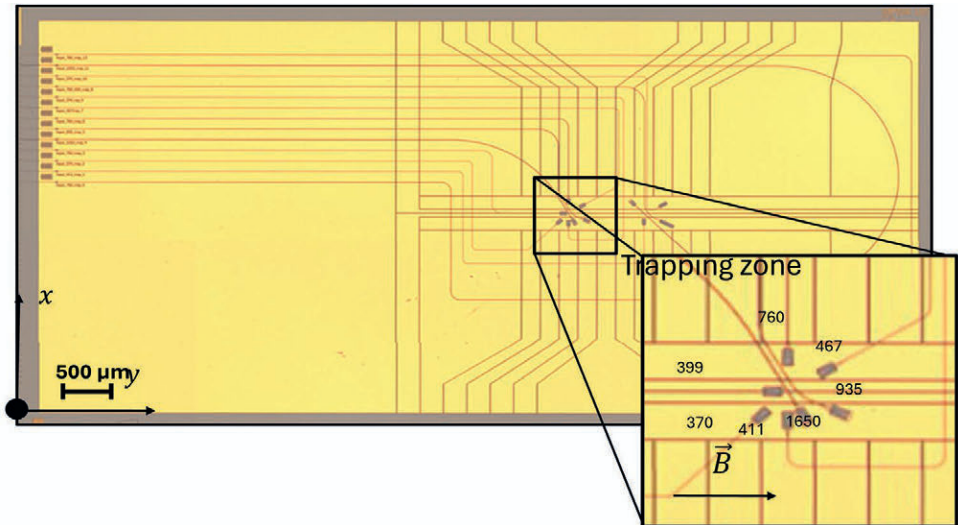
Unsere Entwicklungsarbeit ist Teamarbeit und besteht aus mehreren Teilen. Der erste Teil ist der Entwurf des Designs der Falle mit integrierter Nanophotonik. Wir beginnen mit Überlegungen, welche Anforderungen unsere Anwendung diktiert und welche

Abbildung 4

Das rechte Quadrat zeigt vergrößert die optischen Auskoppler um den Einfangort der Ionen unter Angabe der Laserwellenlängen.

μm : Mikrometer

Quelle: ^[10] Carl Frederik Grimpe, Guochun Du, Fatemeh Salahshoori, Elena Jordan, Michiel De Goede, Sonia M. Garcia-Blanco, Tanja E. Mehlstäubler, A fully integrated ion trap with a single layer of Al₂O₃ nanophotonics supporting light delivery from UV-NIR, EPJ Web of Conferences 335, 03003 (2025)
CC-BY 4.0

Ionenfalle mit integrierter Photonik

Plattformen zur Auswahl haben. Da wir unsere Projekte mit einem Konsortium von akademischen und Industriepartnern durchführen, sind wir in der Auswahl der Photonik-Plattform meist festgelegt durch die Partner, die die Photonik-Chips produzieren. Sie informieren uns über die spezifischen Eigenschaften und Materialparameter.

In das Design gehen auch Überlegungen zur Ionenspezies und zur Atomphysik ein. Das Design der integrierten Falle beinhaltet ein Photonikdesign zusammen mit einem Elektrodendesign. Für das Photonikdesign führen wir Simulationen mit der Methode der Finite-Difference Time-Domain (FDTD) und Simulationen der optischen Mode durch.

Für das Falldesign simulieren wir die Fallenfelder mit der Finite-Elemente-Methode (FEM). Meist werden mehrere Photonik- und Elektrodenlagen benötigt. Das Design wird an die Fabrikationspartner im bevorzugten Format übergeben und mit ihnen diskutiert und optimiert für die Fabrikation. Unsere bisherigen integrierten Fallen sind Oberflächen, z. B. *Abbildung 4 (oben)*, da sich hier die Optik einfacher integrieren lässt.

Meistens kombinieren wir Photonik-Testchips und Ionenfallenchips auf dem gleichen Wafer. Auf den Testchips können neue Photonik-Komponenten ausprobiert, Designstudien durchgeführt und Teststrukturen für das Material, wie Ausbreitungsverlust, Verluste

Strahlrekonstruktion ausgekoppelter Laserstrahlen

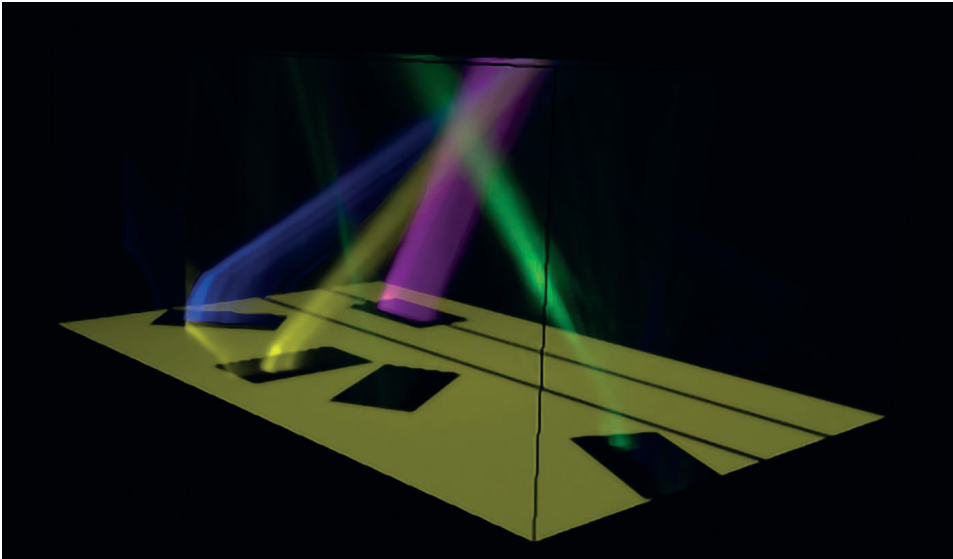


Abbildung 5

Strahlrekonstruktion der aus dem Chip ausgekoppelten Laserstrahlen aus der Falle in Abbildung 4. Die verschiedenen Farben entsprechen verschiedenen Wellenlängen, die Farben entsprechen aber nicht der Farbe der Laser.

Quelle: ^[10] Carl Frederik Grimpe, Guochun Du, Fatemeh Salahshoori, Elena Jordan, Michiel De Goede, Sonia M. García-Blanco, Tanja E. Mehlstäubler, A fully integrated ion trap with a single layer of Al₂O₃ nanophotonics supporting light delivery from UV-NIR, EPJ Web of Conferences 335, 03003 (2025)
CC-BY 4.0

in Biegungen und beim Einkoppeln, untergebracht werden. Für die Ionenfallenchips kommen bewährte Strukturen zum Einsatz.

Für die vertikale Auskopplung der Laserstrahlen aus dem Chip werden optische Gitter mit einer Größe von einigen Mikrometern eingesetzt. Der Auskoppelwinkel ergibt sich aus der Bragg-Bedingung, also der Wellenlänge und dem Abstand der Gitterlinien. Für die Form des Strahls können die Gitterperiode variiert sowie der Öffnungswinkel und Bogenradius des Gitters gewählt werden, wodurch das Gitter fächerförmig aussieht. Dadurch kann man einen Strahl mit Fokus am Ort der Ionen erhalten.

Die Ionenhöhe wird durch die Elektrodengeometrie festgelegt. Daher muss das Photonikdesign auf das Elektrodendesign abgestimmt sein.

Der Photonikchip besteht aus einem Substratmaterial. Meistens wird in der Fabrikation Silizium bevorzugt, wenn auch für Ionenfallen Glas oder besonders Saphir die bessere Wahl wären, da die elektrischen Verluste für die Radiofrequenz geringer sind. Auf das Substrat kommt ein sogenanntes Cladding-Material, welches das Wellenleitermaterial umgibt. Standardmäßig wird hier Siliziumdioxid (SiO₂) verwendet, da es sich mit seinen hervorragenden optischen Eigenschaften im sichtbaren und UV-Bereich

gut eignet. Das Wellenleitermaterial wird mit einer Schichtdicke zwischen ca. 50 nm und 400 nm aufgebracht und mithilfe von Lithografie strukturiert. Dabei wird für feine Strukturen zwischen ca. 50 nm bis 500 nm oft Elektronenstrahlolithografie eingesetzt^[9].

Wenn wir die fertigen Chips erhalten, inspizieren wir sie unter einem Mikroskop und charakterisieren die Photonik mithilfe von Messungen. An der PTB haben wir dafür ein spezialisiertes Photonik-Testlabor mit Lasern aufgebaut, die den UV- bis Infrarot-Bereich (370 nm bis 1650 nm) abdecken. Um

die Strahlqualität der Strahlen aus dem PIC zu beurteilen, haben wir ein Verfahren für 3D-Strahltomografie entwickelt. Dafür wird das Mikroskop beginnend mit dem Fokus auf der Chipoberfläche z.B. in Schritten von 1 μm nach oben gefahren und in jeder Höhe ein Bild aufgenommen. Die

aufgenommenen Bildpunkte werden dann von der Auswertungssoftware dreidimensional übereinander angeordnet, sodass eine 3D-Rekonstruktion des Laserstrahls entsteht. (Abb. 5 $\rightarrow$ S. 35)

Im Rahmen des Projekts ATIQ (Quantencomputer mit gespeicherten Ionen für Anwendungen), das vom Bundesministerium für Forschung, Technologie und Raumfahrt (BM-FTR) gefördert wird, haben wir Si_3N_4 -Nanophotonik für IR-Wellenlängen getestet und dabei zusammen mit unseren Fabrikationspartnern von der AMO GmbH in Aachen die

Auswirkungen der Variation der Fabrikationsparameter auf die Funktion der Nanophotonik untersucht. Dabei ist beispielsweise aufgefallen, dass die Strukturen ein robustes Verhalten mit der Variation der Dosis bei der Lithografie gezeigt haben^[9].

Im EU-Projekt QU-PIC (**Q**uantum **U**niversal **P**hotonics **I**ntegrated **C**ircuit platform) arbeiten wir mit akademischen und Industriepartnern aus mehreren europäischen Staaten an der Entwicklung von photonischen Bausteinen für optische Uhren^[10, 11]. Hier untersuchen wir Alumina-PICs, die es ermöglichen, Licht bis in den UV-Bereich zu verarbeiten.

Ein weiterer Teil der Entwicklung ist, geeignete Materialien für den Anschluss und den Einbau der Falle in die Vakuumkammer auszuwählen, falls möglich einen Anbieter zu finden und die Materialien auf die Eignung für Ionenfallen zu testen.

Da die Ionenfallen in einer Ultra-Hochvakuumumgebung betrieben werden müssen (vorzugsweise $p < 10^{-10}$ mbar), dürfen die Materialien nicht oder nur wenig ausgasen. Die Ausgasrate wird mit einem Restgasanalysator gemessen, einem Massenspektrometer für das Restgas. Für die Messung stehen uns verschiedene Verfahren mit unterschiedlicher Sensitivität zur Verfügung.

Wir betreiben unsere Ionenfallen bei Raumtemperatur, damit sie auch für Anwendungen außerhalb des Labors einsetzbar sind. Hierzu sind Chip-basierte dreidimensionale Fallen mit tiefem Einfangpotenzial vorteilhaft, um die Ionen für lange Zeit in der


Ionenfallen mit Lasern bieten, was man für Präzisionsexperimente braucht: Eine ungestörte Umgebung, lange Beobachtungszeiten und eine präzise Kontrolle.

ELENA JORDAN



Übersicht der besten optischen Uhren basierend auf gefangenen Ionen

Isotop	Uhrenübergang	Linienbreite	Wellenlänge	Relative Unsicherheit
²⁷ Al ⁺ (Aluminium)	¹ S ₀ ↔ ³ P ₀	8 mHz	267 nm	5,5 × 10 ⁻¹⁹
⁴⁰ Ca ⁺ (Calcium)	⁴ S _{1/2} ↔ ³ D _{5/2}	0,1 Hz	729 nm	4,4 × 10 ⁻¹⁹
¹¹⁵ In ⁺ (Indium)	¹ S ₀ ↔ ³ P ₀	0,8 Hz	237 nm	2,5 × 10 ⁻¹⁸
⁸⁸ Sr ⁺ (Strontium)	⁵ S _{1/2} ↔ ⁴ D _{5/2}	0,4 Hz	674 nm	5,3 × 10 ⁻¹⁹
¹⁷¹ Yb ⁺ (Ytterbium)	¹ S ₀ ↔ ³ F ₁	3 nHz	467 nm	3,2 × 10 ⁻¹⁸
¹⁷⁶ Lu ⁺ (Luthetium)	¹ S ₀ ↔ ³ D ₁	0,26 μHz	848 nm	< 8 × 10 ⁻¹⁹

Falle zu halten. Darin überstehen sie auch Stöße mit dem Restgas. Ein Beispiel dafür ist unser patentiertes Zwei-Chip-Design für eine 3D-Ionenfalle, das zusätzliche Elektroden für die Kompensation der Fallenfelder auf den zwei Chips bereits enthält. Das Design enthält an sich keine Optik, eignet sich aber für die Integration. Häufig werden Ionenfallen bei kryogenen Temperaturen betrieben. Das ist nicht notwendig, hat aber die Vorteile, dass viele Stoffe an den Wänden der Vakuumkammer kondensieren und der Druck stark sinkt. Auch in der Auswahl der Materialien ist man dadurch freier.

Optische Uhren

Was kann man am genauesten messen mit der geringsten Messunsicherheit? Die Antwort: Den Genauigkeitsrekord halten Frequenzmessungen. Frequenz *f* ist heutzutage die am genauesten bestimmbare Größe mit erreichten relativen Unsicherheiten $\Delta f/f < 10^{-18} = 0,000.000.000.000.000.001$ ^[12,13,15,18,19].

Mit solch einer stabilen Frequenzreferenz kann eine Uhr realisiert werden, indem man ein Gerät baut, das die Schwingungen zählt^[20]. Um zu vermitteln, wie unvorstellbar diese Genauigkeit ist, wird oft in etwa folgendes Beispiel genannt: Die derzeit beste Uhr ginge über eine Zeit drei Mal so lang wie das Alter des Universums, das auf 13,8 Milliarden Jahre geschätzt wird^[21], um weniger als eine Sekunde falsch. Vorausgesetzt sie wäre so lange gelaufen. In der Realität ist diese Uhr nur einige Wochen gelaufen.

Als Frequenzreferenz verwenden diese Uhren atomare Übergänge von einem elektronischen Zustand in einen anderen eines gefangenen Ions in einer Paulfalle^[12, 14, 15, 16, 18, 19] oder Atomen in einem optischen Gitter^[13].

Dabei dient das gefangene Ion als Langzeitreferenz für die Frequenz. Verwendet man das gleiche Isotop mit dem gleichen Ladungszustand, dann sind diese identisch und die Elektronen schwingen bei der gleichen Frequenz. Beispiele für die in Uhren verwen-

Tabelle 1:
Übersicht der besten optischen Uhren (Stand März 2026) basierend auf gefangenen Ionen. Die erste Spalte gibt das verwendete Isotop an, die zweite Spalte den atomaren Uhrenübergang, die dritte Spalte die Linienbreite des Übergangs, die vierte Spalte die Wellenlänge des Uhrenlasers und die fünfte Spalte die erreichte relative Frequenzunsicherheit ^[12, 14–19].

Hz: Hertz
mHz: Millihertz
μHz: Mikrohertz
nm: Nanometer

deten Übergänge, Linienbreiten und Übergangswellenlängen und -frequenzen sowie damit erreichte relative Unsicherheiten sind in *Tabelle 1* → *S. 37* aufgeführt.

In einer optischen Uhr wird ein Laser mit der Frequenz des atomaren Übergangs synchronisiert. Ist der Laser nahe der Resonanz, wird der Übergang mit hoher Wahrscheinlichkeit angeregt. Man erhält ein Signal, z.B. ein verringertes Fluoreszenzsignal auf einem anderen Übergang, der stark fluoresziert. Das Fluoreszenzsignal wird mit einem Einzelphotonendetektor gemessen. Die Frequenz des Lasers wird durch Rückkopplung immer nahe der atomaren Übergangsfrequenz gehalten. Um solch eine genaue Messung durchführen zu können, muss man auch alle Einflüsse aus der Umgebung sehr genau kennen und daraus bestimmen, welchen Einfluss sie auf die Frequenz des Übergangs haben.

Je geringer das Rauschen auf den Komponenten der Messung und je stärker das Signal ist, desto geringer ist die Unsicherheit der Messung, die man erreichen kann. Das Signal skaliert mit der Anzahl der Ionen, deshalb führen Messungen mit mehreren oder vielen Atomen, wie in einer optischen Gitteruhr oder in einer Multi-Ionen-Uhr, zu einem besseren Signal-zu-Rausch-Verhältnis.

Die Anforderungen an den Uhrenlaser sind sehr hoch. Er muss eine sehr hohe spektrale Reinheit und eine sehr stabile Phase, also ein geringes Phasenrauschen, aufweisen. Um das zu erreichen, wird der Uhrenlaser auf einen ultra-stabilen Resonator stabilisiert.

Da man relative Unsicherheiten angibt ($\Delta f/f$, wobei f die Frequenz des Lasers ist), spielt auch die Laserfrequenz eine wichtige Rolle. Je höher, desto besser und da hohe Frequenz kurze Wellenlänge bedeutet, sind blaue und UV-Wellenlängen vorteilhaft.

Neben dem Uhrenlaser werden auch Laser benötigt für das Laserkühlen und weitere Laser, die die Ionen zurück in die gewollten Zustände bringen, sogenannte Rückpumper. Diese Uhren sind so genau, dass die Rotverschiebung im Gravitationspotenzial der Erde messbar wird. Uhren in einem höheren Gravitationspotenzial gehen langsamer. So kann man einen Unterschied in der Frequenz beobachten, wenn man die Höhe der Uhr um einen Zentimeter verändert oder zwei Uhren auf unterschiedlichen Höhen vergleicht. Die Methode nennt man relativistische Geodäsie oder chronometrische Höhenbestimmung. Mit einer hochgenauen portablen Uhr kann man die Methode nutzen, um Höhenveränderungen nach Erdbeben oder das Füllen von Magmakammern bei Vulkanen zu beobachten^[22] – und das deutlich genauer, als es heutzutage mit Satelliten möglich ist.

Optische Uhren sind so genau, dass man damit Änderungen der relativistischen Rotverschiebung im Gravitationspotenzial der Erde verfolgen kann.

ELENA JORDAN

Ionenfallen für Quantencomputer

Auch für Ionen-basierte Quantencomputer benötigt man eine Ionenfalle und Laser für das Laserkühlen, Rückpumpen, die Zustandspräparation und für die Quantenlogikgatter, wobei je nach Energiedifferenz der Zustände auch Mikrowellen eingesetzt werden.

Die Ionen haben elektronische Zustände, von denen man zwei als Qubit-Zustände (oder mehrere als Qudits) definiert. Nun kann man Logikgatter ausführen, indem man die Zustände mit einer bestimmten Wahrscheinlichkeit anregt, mehrere Quantenzustände simultan über mehrere Pfade anregt etc.

Ionen sind eine sehr vielversprechende Plattform für Quantencomputer mit einer sehr

langen Kohärenzzeit. Die Kohärenzzeit gibt an, wie lange das Quantensystem im präparierten Zustand bleibt. Ionen können eine sehr lange Kohärenzzeit haben, bis zu elf Stunden wurden beobachtet^[23]. Zum Vergleich liegt der Rekord für die Kohärenzzeiten in Supraleitenden Quantencomputern bei 0,001 s^[24] und in Rydbergatomen bei einigen μ s.

Im Gegensatz zur Uhr ist es hier sehr wichtig, genau ein einzelnes Ion ansprechen zu können. Dies erfordert viele optische Kanäle auf engem Raum. Deswegen ist hier die Motivation, integrierte Ionenfallen zu entwickeln, besonders hoch.



Da Quantenzustände in gefangenen Ionen sehr langlebig sein können, werden sie oft als Quantenspeicher eingesetzt.

ELENA JORDAN



Referenzen

- [1] W. Paul, Nobel Lecture, (1998)
- [2] Wikipedia, Photonics. Stand (27.03.2026), <https://en.wikipedia.org/wiki/Photonics> (2026)
- [3] E. Jordan, M. Brinkmann, A. Didier, et al. Quantum Sci. Technol. 10 045005 (2025)
- [4] S. Shekhar, W. Bogaerts, L. Chrostowski, et al. Nat Commun 15, 751 (2024)
- [5] C. Xiang, W. Jin, J. E. Bowers, Photonics Res. 10, A82 (2022)
- [6] K. K. Mehta, C. Bruzewicz, R. McConnell, et al. Nature Nanotech 11, 1066 (2016)
- [7] T. D. Morrison, J. Kwon, M. A. Delaney, et al. arXiv:2512.08921 [quant-ph] (2026)
- [8] G. Du, E. Jordan, T. E. Mehlstäubler, arXiv:2503.20387v2 [quant-ph] (2025)
- [9] C.-F. Grimpe, A. Sorokina, G. Du, et al. arXiv:2603.22578 [physics.optics] (2026)
- [10] C.-F. Grimpe, G. Du, F. Salahshoori, et al., EPJ Web of Conferences 335, 03003 (2025)
- [11] R. Banerjee Chaudhuri, C.-F. Grimpe, G. Du, et al. Proc. SPIE 13900, Integrated Optics: Devices, Materials, and Technologies XXX, 139000F (2026)
- [12] M. C. Marshall, D. A. Rodriguez Castillo, W. J. Arthur-Dworschak et al., Phys. Rev. Lett. 135, 033201 (2025)
- [13] Y. A. Yang, M. Miklos and Y. M. Tso, et al. Phys. Rev. Lett. 135, 193202 (2025)
- [14] J. Keller, N.H. Hauser, I.M. Richter, et al. 2024 J. Phys.: Conf. Ser. 2889 012050 (2024)
- [15] M. Filzinger, M. R. Steinel, J. Jiang, et al., arXiv:2603.23446 [physics.atom-ph], (2026)
- [16] N. Huntemann, C Sanner, B Lipphardt, et al. Phys. Rev. Lett., 116(6):063001, (2016)
- [17] R. Lange, A. A. Peshkov, N. Huntemann, et al. Phys. Rev. Lett. 127, 213001, (2021)
- [18] Z. Zhiqiang, K. J. Arnold, R. Kaewuam, et al. Sci. Adv. 9, 1971 (2023)
- [19] B.-L. Zhang¹, Z.-X. Ma, Y. Huang, et al., Phys. Rev. Lett. 136, 053202 (2026)
- [20] A. D. Ludlow, M. M. Boyd, J. Ye, et al. Planck, Rev. Mod. Phys. 87, 637 (2015)
- [21] Collaboration, A&A 641, (2020)
- [22] R. Bondarescu, A. Schäfer, A. Lundgren, et al. Geophys. J. Int. 202, 1 (2015)
- [23] J. Pi, X. Liu, J. Cao, et al. arXiv:2603.19631v1 [quant-ph] (2026)
- [24] Tuokkola, M., Sunada, Y., Kivijärvi, H. et al. Nat Commun 16, 5421 (2025).

Danksagungen

Ich bedanke mich bei meinem Ionenfallen- und Photonik-Team und der gesamten Forschungsgruppe von Prof. Tanja E. Mehlstäubler am QUEST Institut der PTB für die Zusammenarbeit. Diese Arbeit wird gefördert durch das BMFTR in den Projekten ATIQ Projekt (FKZ: 13N16126), EVAQS (FKZ: 13N15937) und IDEAL (13N14962), die VolkswagenStiftung und das Niedersächsische Ministerium für Wissenschaft und Kultur im Projekt Quantum Valley Lower Saxony Q1, das EU Horizon Project QU-PIC (Grant Agreement No #101135845), die DFG im Excellence Cluster Quantum Frontiers (EXC-2123-390837967), das Exzellenz Cluster PhoenixD (EXC-2122) und den Sonderforschungsbereich DQ-mat (SFB 1227).

Ein Phasenfeldmodell für Ermüdungsrisse



Dr. Martha Kalina ist studierte Bauingenieurin und absolvierte ihre Promotion im Maschinenbau an der Technischen Universität Dresden. Im Forschungsfeld Computational Mechanics beschäftigt sie sich mit der Modellierung von Materialverhalten. Inhalt ihrer Dissertation ist die Entwicklung eines Phasenfeldmodells zur Beschreibung von Ermüdungsrisen. Das entwickelte Modell fand praktische Anwendung in einem Kooperationsprojekt mit dem Deutschen Zentrum für Luft- und Raumfahrt, in dem Risswachstum in Flugzeugrümpfen untersucht wurden. Im Jahr 2023 verbrachte Martha Kalina einen Forschungsaufenthalt an der Columbia University in New York, wo sie den Fokus ihrer Modellierung auf elastisch-plastisches Materialverhalten legte. Für ihre wissenschaftlichen Leistungen wurde sie von der GAMM – der Gesellschaft für Angewandte Mathematik und Mechanik – als eine der GAMM Juniors ausgezeichnet. Seit 2025 ist sie in der Softwareentwicklung für die Halbleiterindustrie tätig.

Martha Kalina

Ein Phasenfeldmodell für Ermüdungsrisse

Ermüdungsbrüche gehören zu den häufigsten Ursachen für das Versagen technischer Strukturen. Ob in Flugzeugen, Fahrzeugen, Brücken oder Maschinenbauteilen – überall dort, wo Materialien über lange Zeiträume zyklischen Belastungen ausgesetzt sind, entstehen schleichend Schäden. Diese beginnen auf mikroskopischer Ebene und können sich über Millionen von Lastwechseln hinweg zu makroskopisch sichtbaren Rissen entwickeln, die schließlich zum Versagen des Bauteils führen. Die besondere Herausforderung der Ermüdung liegt darin, dass sie nicht durch eine einzelne Überlast ausgelöst wird, sondern durch die wiederholte Beanspruchung weit unterhalb der statischen Festigkeit eines Materials. Ein Bauteil kann also scheinbar sicher betrieben werden und dennoch ohne Vorwarnung versagen. Aus diesem Grund spielen präzise Modelle zur Vorhersage von Ermüdungsrisssinitiierung und -wachstum eine zentrale Rolle für die Sicherheit und Effizienz vielgestaltiger Strukturen.

Eine besondere Herausforderung stellen Flugzeuge dar. Als ultra-leichte Strukturen können sie nicht so stark ausgelegt werden, dass Risse komplett verhindert würden. Stattdessen akzeptiert man, dass z. B. Flugzeuggrümpfe Risse aufweisen, und stellt

durch regelmäßige Checks sicher, dass sich diese Risse nicht über eine kritische Länge hinausentwickeln^[SW194]. Für die Auslegung der Struktur sowie die Festlegung von Wartungsintervallen ist die genaue Kenntnis der Ausbreitungsgeschwindigkeit eventueller Risse von entscheidender Bedeutung: um Material möglichst sparsam und effizient einsetzen zu können, aber auch um ermüdungsbedingtes Versagen zu verhindern – und damit Unglücke wie den Absturz einer Maschine der



Ermüdungsschäden sind für einen großen Teil von Schäden in Bauteilen und Strukturen des Maschinenbaus und anderer Ingenieurdisziplinen verantwortlich.

MARTHA KALINA



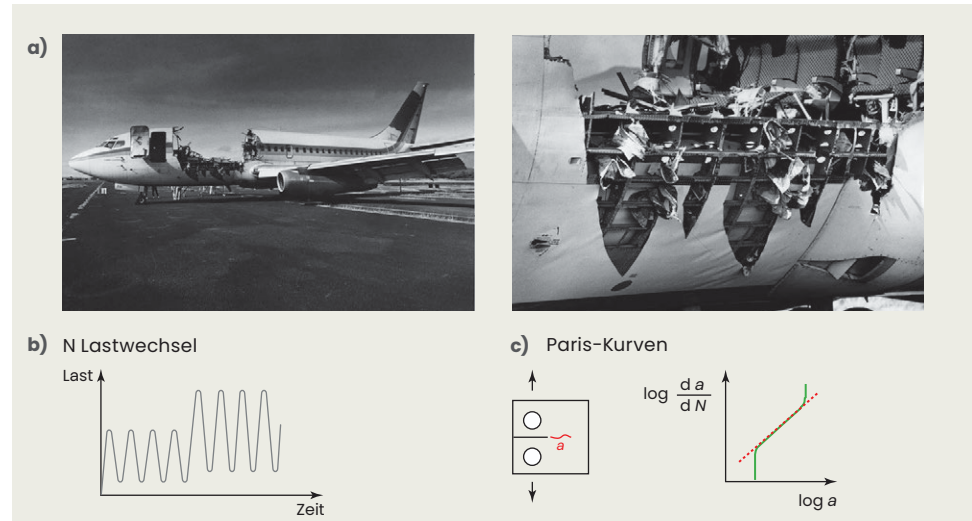
Zusammenfassung:

Viele Bauteile leiden unter Ermüdungslasten, d. h. sich wiederholenden Lasten. Das können bei einer Brücke die unzähligen Überfahrten von Autos sein oder bei Flugzeugen die immerwährenden Starts und Landungen. Diese führen zu Rissen, die dann Stück für Stück voranschreiten. Diesen Rissfortschritt abschätzen zu können ist entscheidend für eine sichere und materialsparende Auslegung der Bauteile. Abschätzen kann man dies durch Simulationen. Eine Schwierigkeit bei der Simulation von Rissfortschritt ist, dass sich die simulierte Geometrie permanent ändert, wenn neue Rissoberfläche entsteht. Eine weitere Herausforderung sind die oft langen Rechenzeiten der Modelle, da Tausende oder gar Millionen von Lastwechseln simuliert werden müssen. Unser Phasenfeldmodell für Ermüdungsrisssfortschritt vereint Ansätze für beide Herausforderungen.

Abbildung 1 (a, b, c)

(a) Ermüdungsversagen einer Boeing 737-200 der Aloha Airlines nach 90.000 Starts im Jahr 1988
 (b) Ermüdungsversagen entsteht durch sich wiederholende Lastwechsel.
 (c) Die **Rissfortschrittsrate** da/dN über der **Risslänge** a kann durch das Paris-Gesetz^[PAR63] beschrieben werden.

Alle Bilder^[NTS88].

Ermüdungsrisse

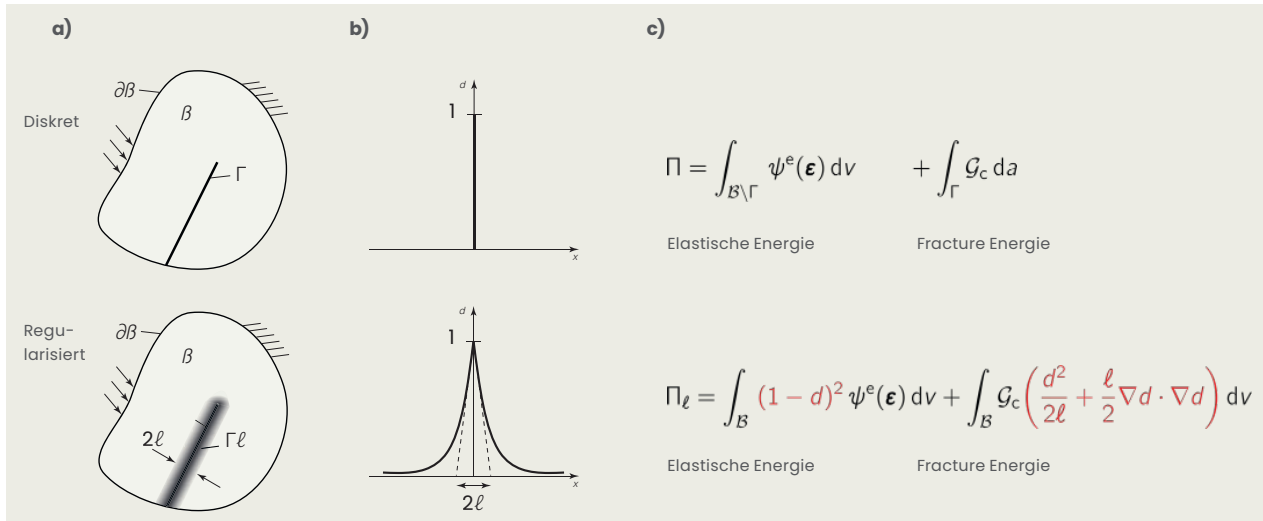
Aloha Airlines^[NTS88] im Jahr 1988 (Abb. 1a). Herkömmliche Methoden der Abschätzung des Rissfortschritts basieren auf standardisierten Rissfortschrittsexperimenten. Dabei wird eine Probe mit einer zyklischen Last (Abb. 1b) belastet und der gemessene Rissfortschritt a über der Lastwechselzahl N mithilfe von Paris-Formeln^[PAR63] beschrieben. Basierend darauf können gerade Risse in einfachen Blechen gut abgeschätzt werden. Doch Flugzeugrümpfe weisen häufig eine komplexe Geometrie auf und auch die sich dadurch ergebenden Rissmuster sind vielschichtig: Phänomene wie Rissablenkung und Verzweigung sind zu beobachten. Hier stoßen herkömmliche Methoden an ihre Grenzen. Im Folgenden beschreiben wir, wie Ermüdungsrisse stattdessen besonders flexibel

und gleichzeitig effizient modelliert und simuliert werden können – mit der sogenannten Phasenfeldmethode^[SEI20].

Die Idee der Phasenfeldmethode

Die zentrale Herausforderung bei der Simulation von Rissen ist die neu entstehende Oberfläche, wenn der Riss sich durch das Material voran bewegt. Das Bauteil verändert gewissermaßen seine Geometrie. Die Finite-Elemente-Methode (FEM) als die verbreitetste Methode zur Simulation des mechanischen Verhaltens von Bauteilen verwendet aber ein festes Berechnungsnetz, um die Geometrie eines Bauteils zu beschreiben, man könnte auch sagen, zu triangulieren.

Regularisierung des Risses



Genau hier setzt die Phasenfeldmethode an: Anstelle das Berechnungsnetz zu verändern, sobald sich eine neue Rissoberfläche auftut, führt sie eine zusätzliche Feldvariable ein, um den Riss zu beschreiben^[MIE10]. Diese Feldvariable d dient als Indikator für den Riss. Man definiert $d = 0$ als intaktes Material und $d = 1$ als gebrochenes Material. Das Konzept der Beschreibung der Geometrie unterschiedlicher Phasen eines Materials mithilfe von Indikatorvariablen ist auf viele verschiedene Systeme anwendbar, seien es verschiedene Aggregatzustände, Gefüge oder eben der Riss. Die grundlegende Formulierung der Phasenfeldmodelle ist immer ähnlich.

Abbildung 2 veranschaulicht das Prinzip am Beispiel eines gebrochenen Körpers B mit ei-

ner Rissfläche Γ (Abb. 2a). Der Verlauf der Variable d quer über den Riss (Abb. 2b) für den Fall eines diskreten (d. h. scharfen) Risses ergibt dabei einen Peak mit $d = 1$ am Ort des Risses. Die Gesamtenergie eines solchen Körpers (Abb. 2c) setzt sich in diesem Fall zusammen aus der elastischen Energie und der Rissenergie des Körpers. Letztere ergibt sich aus dem Integral der Bruchzähigkeit $\mathcal{G}_c$ des Materials über die Rissfläche Γ . Die Bruchzähigkeit ist der wichtigste Parameter eines Materials zur Beschreibung seines Bruchverhaltens.

Im nächsten Schritt wird der Riss regularisiert, d. h., über eine gewisse charakteristische Breite ℓ verschmiert (Abb. 2, zweite Zeile). Im Profil des Phasenfelds ergibt sich dabei

Abbildung 2 (a, b, c)

Risse können durch die riss-anzeigende **Phasenfeldvariable d** beschrieben werden. Bei $d = 1$ liegt ein Riss vor. Die Phasenfeldmethode regularisiert (verschmiert) den eigentlich diskreten Riss über eine **Breite 2ℓ** .

(c) Die Energie eines gerissenen **Körpers B** besteht aus elastischer und Fracture-Energie. Regularisierung führt zu einer **Degradation der elastischen Energie um $(1-d)^2$** und einer neuen Beschreibung der **Bruchfläche Γ** ^[MIE10].

ein exponentieller Verlauf statt eines Sprunges. Warum sich diese Formulierung für die numerische Lösung des Problems mit der FEM besonders gut eignet, lässt sich an der Gesamtenergie erkennen: Durch die Regularisierung kann die Bruchfläche nun durch das Phasenfeld d und seinen Gradienten ∇d beschrieben werden. Integrationsgebiet der Bruchenergie ist dadurch nicht mehr die Bruchfläche, sondern der ganze Körper B . Im Algorithmus der FEM kann die Integration so unabhängig vom Riss über das gesamte Gebiet durchgeführt werden. Des Weiteren

wird auch in der elastischen Energie eine Anpassung gemacht: Die elastische Energie wird mit $(1-d)^2$ degradiert. Im Bereich des Risses wird also keine elastische Energie gespeichert.

Durch ihre elegante Formulierung kann die Phasenfeldmethode auf viele Probleme angewendet

und einfach in gängige FEM-Software implementiert werden. Herkömmliche Methoden zur Simulation von Rissen definieren richtungsabhängige Rissfortschrittskriterien und komplexe Regeln für Rissinitiierung und -verzweigung. All dies liefert die Phasenfeldmethode „out of the box“. Erstmals in ihrer hier dargestellten Form wurde die Phasenfeldmethode von Miehe et al.^[MIE10] beschrieben. Im Folgenden stellen wir die Anwendung der Methode auf den Spezialfall der Ermüdungsrisse vor.


Die Phasenfeldmethode eignet sich in besonderem Maße zur Simulation von Rissen, da sie beliebige Rissgeometrien, Rissinitiierung und -verzweigung ohne zusätzliche Bedingungen abbilden kann.

MARTHA KALINA



Ein pseudo-plastisches Modell

Aluminium, welches zur Konstruktion von Flugzeugen verwendet wird, ist wie viele Metalle ein Material, das sich duktil verhält: Rund um die Risssspitze von Rissen unter statischer Last, aber auch von Ermüdungsrisse, bildet sich eine plastische Zone^[BES16]. Das bedeutet, dass sich das Material in einem häufig nierenförmigen Gebiet um die Risssspitze plastisch verhält (Abb. 3 → rechts). Durch Versetzungen in der Kristallstruktur deformiert sich das Material irreversibel, d. h. unumkehrbar. Dabei wird Energie umgewandelt. Eine gute Beschreibung dieses plastischen Verhaltens sowie des Übergangs zwischen elastischer (reversibler) und plastischer Verformung ist essenziell, um Ermüdungsrissfortschritt in Metallen beschreiben und schließlich vorhersagen zu können.

Zur Beschreibung elastisch-plastischen Materialverhaltens existieren Materialmodelle, auch kombiniert mit der Phasenfeldmethode^[KAL25]. Das Problem bei herkömmlichen, im Folgenden als „ausführliche“ bezeichnete, elastisch-plastischen Modellen ist, dass sie den Berechnungsablauf während der Simulation gegenüber einer rein elastischen Berechnung verkomplizieren. Zur Lösung der Modellgleichungen ist eine weitere Iterationsschleife in der Materialroutine erforderlich. Diese muss in jedem berechneten Materialpunkt durchlaufen werden, was die Berechnungszeit erheblich steigert. Gerade für die Simulation von Ermüdungslasten mit Tausenden oder gar Millionen von Lastwechseln, die ohnehin schon sehr aufwendig sind, macht dies die Simulation realistischer

Plastizität an der Rissspitze

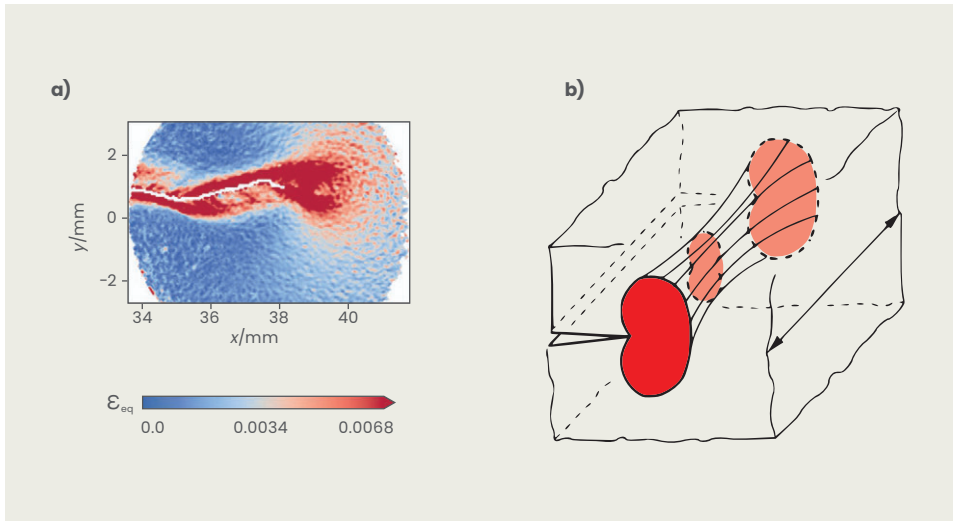


Abbildung 3

(a) Mikroskopische Darstellung der plastischen Zone an der Spitze eines Ermüdungsrisses. Experiment des DLR.

(b) Schematische Darstellung der plastischen Zone über die Dicke des Bauteils.

Bild nach ^[GRO11].

Bauteile fast unmöglich. Die Reduktion der Rechenzeit war daher oberstes Ziel bei der Entwicklung unseres Phasenfeldmodells für Ermüdungsrissfortschritt.

Kern des Modells ist eine vereinfachte Formulierung der Plastizität. Diese erlaubt eine ausreichend gute Näherung der plastischen Prozesse bei gleichzeitig drastischer Reduktion der Rechenzeit. Die Vereinfachung vom ausführlichen, elastisch-plastischen Modell zum vereinfachten, pseudo-plastischen Modell erfolgt dabei in drei Schritten^[vgl. KAL25]:

1. Zum einen wird der Spannungs-Dehnungs-Pfad approximiert. Herkömmliche Modelle lösen den Pfad am Materialpunkt komplett auf, in der Regel mit mindestens

zehn Inkrementen, d. h. Berechnungsschritten. Dem pseudo-plastischen Modell genügt die Simulation der Umkehrpunkte im Spannungs-Dehnungs-Pfad. Grundlage dieser Näherung ist die Kenntnis des Spannungs-Dehnungs-Verhaltens unter zyklischer Last dieses Materials. Dieses wird mit zyklischen Experimenten an Materialproben untersucht und geht als sogenannte Zyklische Spannungs-Dehnungs-Kurve in das Modell ein.

2. Soll mit Phasenfeldmodellen Ermüdungsschädigung simuliert werden, kommen sogenannte Ermüdungsvariablen zum Einsatz. Diese beschreiben die fortschreitende Schädigung gegenüber dem Ausgangsmaterial, aufgelöst je Material-

Abbildung 4

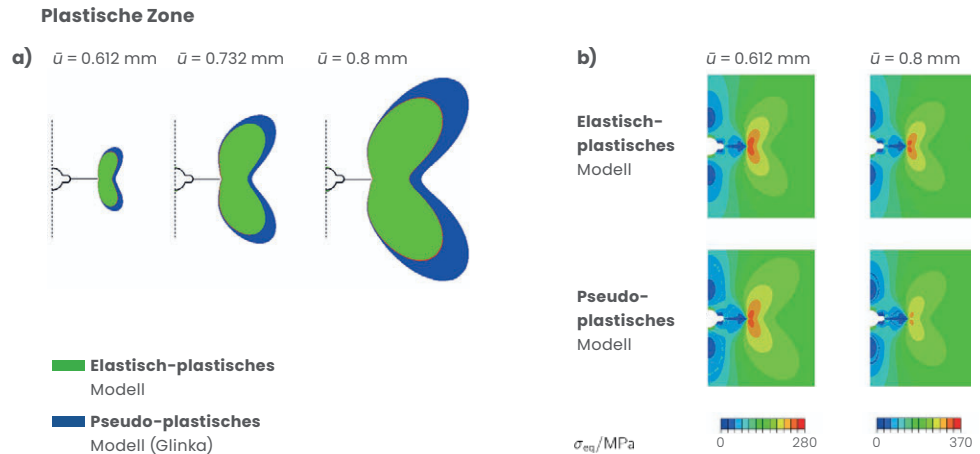
Vergleich des pseudo-plastischen Modells mit dem Elastisch-Plastischen

(a) Die plastische Zone ist wenig kleiner.

(b) Die Spannung an der Rissspitze (oben: Vergleichsspannung, unten: unbewertete Spannung nach^[GL185])

wird leicht unterschätzt.

Bei geringen Amplituden ist die Näherung gut.

Vergleich pseudo-plastisches und elastisch-plastisches Modell

punkt über der Zeit. Während herkömmliche Modelle die Ermüdungsvariable aus Integralen über dem Spannungs-Dehnungs-Pfad berechnen, verwendet das pseudo-plastische Modell eine Näherung auf Basis der Umkehrpunkte. Hier kommt u. a. der Schädigungsparameter nach Heitmann et al.^[HE184] zum Einsatz. Die Beschreibung der schädigenden Wirkung der Last wird erneut durch experimentelle Daten gestützt: Die Dehnungs-Wöhlerlinie nimmt die Anzahl der Lastwechsel bis Bruch auf, abhängig von der Lastamplitude.

- Die Risstriebkraft ist der Teil des Phasenfeldmodells, der den Rissfortschritt infolge statischer Lasten beschreibt – sozusagen alle Effekte, die nicht nur auf

Ermüdung zurückzuführen sind. Teil der Risstriebkraft ist neben der elastischen Energiedichte normalerweise auch die plastische Energiedichte. Wie wir in Untersuchungen zeigen konnten, kann Letztere aber für niedrige bis mittlere Lastniveaus vernachlässigt werden.

Diese drei Schritte bewirken eine deutliche Verkürzung der Rechenzeit um bis zu Faktor 189^[KAL25]. Diese Reduktion lässt sich durch die folgenden Effekte erklären:

- Statt einer elastisch-plastischen Materialroutine kann eine elastische verwendet werden, die keine zusätzlichen Iterationen erfordert.
- Es werden weniger Berechnungsschritte pro vollständig berechnetem Lastwechsel

benötigt: statt z. B. zehn Schritten nur die zwei Umkehrpunkte.

- Sind die Lasten moderat und schreitet der Riss nur langsam voran, kann die Berechnung sogar noch stärker gerafft werden: Da sich die Rissgeometrie über viele Lastwechsel kaum ändert, bleibt auch die Risstriebkraft gleich und muss daher nicht neu berechnet werden. Der Rissfortschritt lässt sich so über viele Lastwechsel extrapolieren. Eine explizite Berechnung ist nur in größeren Abständen nötig. Wie viele Lastwechsel so „übersprungen“ werden können, richtet sich dabei nicht nach deren Anzahl, sondern nach dem Rissfortschritt und der Lasthöhe. Für niedrige Lasten und hohe Lastwechselzahlen kann so besonders viel Rechenzeit eingespart werden.

Dass die eingeführten Vereinfachungen des pseudo-plastischen Modells nicht nur Rechenzeit einsparen, sondern den Prozess auch noch ausreichend genau abbilden, wurde durch Simulationen untersucht^[KAL25].

Abbildung 4a (→ links) zeigt die simulierte plastische Zone für das pseudo-plastische Modell, verglichen mit dem ausführlichen, elastisch-plastischen Modell. Form und Größe der plastischen Zone stimmen mit guter Näherung überein. Bei höheren Lastamplituden wird die Näherung etwas schlechter. Dass die pseudo-plastische Zone stets kleiner ist als die aus der Vergleichsrechnung, liegt daran, dass plastische Umverteilungsmechanismen nicht berücksichtigt werden können. Abbildung 4b (→ links) zeigt den Vergleich hinsichtlich des Spannungszustandes

an der Rissspitze: Sowohl qualitativ als auch quantitativ wird der Zustand gut beschrieben. Große Amplituden führen ebenfalls wieder zu größeren Abweichungen.

Grundsätzlich kann also festgestellt werden: Das pseudo-plastische Modell ist eine gute Wahl für kleine und mittlere Lastamplituden. Für hohe Lasten und kleine Lastwechselzahlen sollte auf ein elastisch-plastisches Modell zurückgegriffen werden.

Ein Modell für anisotrope Werkstoffe

Aluminiumbleche, die zur Herstellung von Flugzeugrümpfen verwendet werden, werden gewalzt. Schaut man sich ihre Struktur genauer an, stellt man fest, dass der Walzprozess zu einem Gefüge aus länglichen Kristallen führt. Entnimmt man Proben in verschiedenen Richtungen zur Walzrichtung, so kann man unterschiedliches Rissverhalten beobachten: Risse breiten sich in Walzrichtung schneller aus, weil sie entlang der länglichen Körner wachsen können und weniger Korngrenzen überwinden müssen als Risse quer zur Walzrichtung^[JES22]. Das Blech verhält sich also hinsichtlich seines Rissverhaltens anisotrop, d. h. richtungsabhängig. Da eine einzige Vorzugsrichtung – die Walzrichtung – existiert, spricht man von transversaler Isotropie.

Diese Richtungsabhängigkeit kann man durch Messung



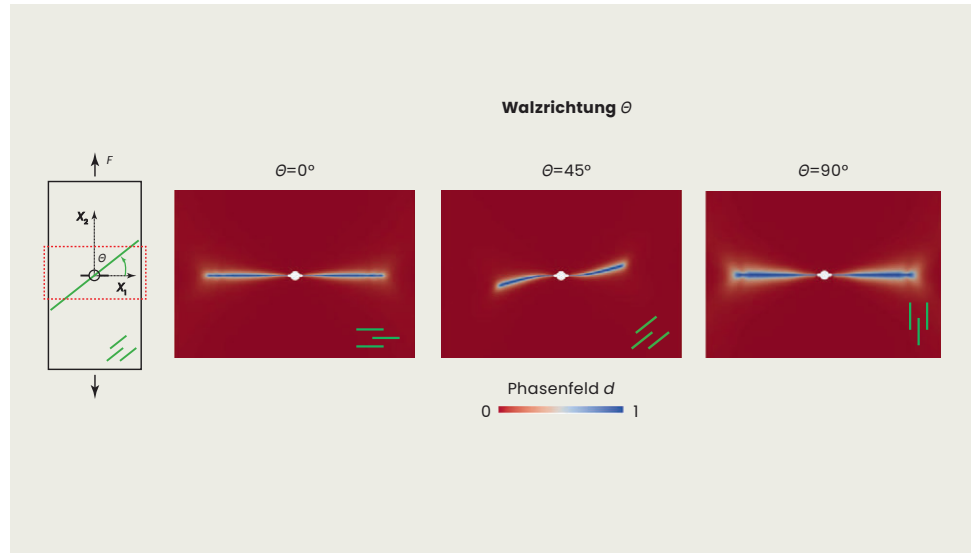
Duktile Materialien wie Aluminium zeigen plastische Zonen an der Rissspitze. Plastizität sollte daher in der Modellierung von Ermüdungsrisse berücksichtigt werden.

MARTHA KALINA



Abbildung 5

Simulation des Rissfortschritts in Middle-Tension-Proben für verschiedene Ausrichtungen der Probe zur Walzrichtung; Riss dargestellt durch Phasenfeldvariable $d = 1$. Die Walzrichtung 45° führt zu schrägem Riss trotz probenparalleler Belastung. Quelle^[KAL23].

Simulation unterschiedlicher Risspfade

der Bruchzähigkeit in verschiedene Richtungen charakterisieren. Aus den drei gemessenen Werten in Winkeln von 0° , 45° und 90° zur Walzrichtung lässt sich dann ein Modell der Bruchzähigkeit für sämtliche Richtungen konstruieren. Dieses Modell lässt sich

nach Teichtmeister et al.^[TE117] wiederum in einen Strukturtenor A übersetzen. Fügt man diesen in den Gradiententerm der Bruchenergie des Phasenfeldmodells ein, entsteht eine anisotrope Phasenfeldformulierung: Auch im Modell wird Neigung zum Rissfortschritt richtungsabhängig.

Simuliert man mit diesem Modell eine Middle-Tension-Probe, d. h. eine Probe mit mittigem Loch und horizontalen Anrissen unter vertikaler Belastung, ergeben sich die in *Abbildung 5* dargestellten Rissmuster^[KAL23]. Der Riss ist dabei durch den in blau dargestellten Bereich von $d = 1$ markiert. Wird die Probe parallel oder orthogonal zur Walzrichtung aus dem Blech „geschnitten“, wächst der Riss erwartungsgemäß orthogonal zur aufgebrachten Last. Bei einem Winkel von 45° zur Walzrichtung wird der Riss stattdessen schräg abgelenkt.

Gleiches beobachtete das DLR auch im realen Experiment^[KAL23]. *Abbildung 6* ($\rightarrow$ rechts) zeigt den Vergleich von Simulation und

Für die Auslegung dünnwandiger Bauteile, wie z. B. Flugzeugrümpfe, ist es entscheidend zu wissen, wie schnell Ermüdungsrisse voranschreiten.

MARTHA KALINA

Vergleich von Simulationen und experimentellen Ergebnissen

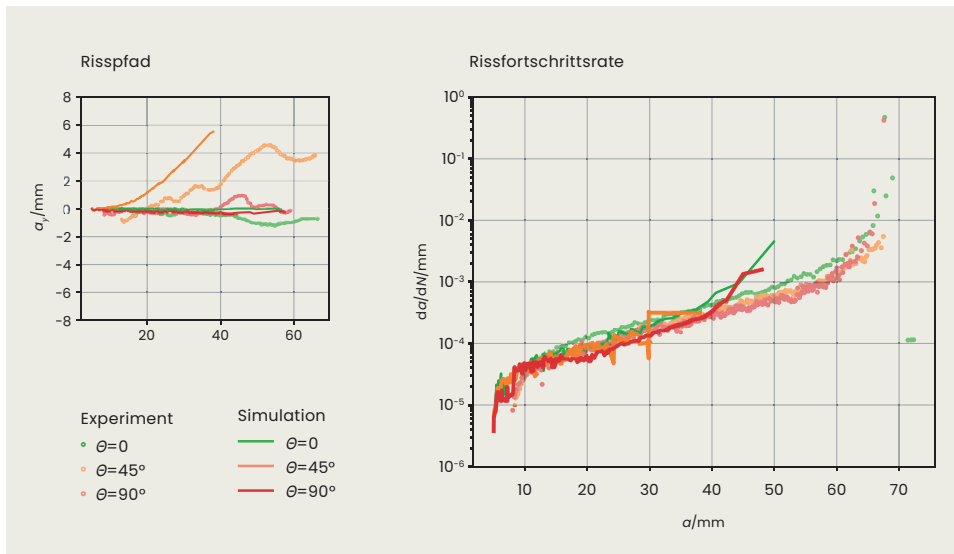


Abbildung 6

Vergleich Simulationen und experimenteller Ergebnisse des DLR, Ermüdungsrissfortschritt in einer Middle-Tension-Probe unter drei Ausrichtungen zur Walzrichtung.
Quelle^[KAL23].

Experiment: Links ist der Risspfad dargestellt. Die Ablenkung des Risses in der 45°-Probe ist zu beobachten. Auch für die Rissfortschrittsraten stimmt der Trend in Experiment und Simulation überein. Rechts dargestellt ist der Rissfortschritt pro Lastwechsel da/dN über der Risslänge a . Proben mit Walzrichtung parallel zur Kraftrichtung

haben die geringste Rissfortschrittsrate, orthogonale Proben die höchste. Dass eine Modellierung dieses komplexen Verhaltens möglich ist, kann potenziell die Durchführung vieler Experimente vermeiden und die Auslegung von Flugzeugen, aber auch anderen Strukturen, sicherer und materialsparender machen.

Referenzen

- [BES16] Besel, M., Breitbarth, E., 2016. Advanced analysis of crack tip plastic zone under cyclic loading. *International Journal of Fatigue* 93, 92–108. <https://doi.org/10.1016/j.ijfatigue.2016.08.013>
- [GLI85] Glinka, G., 1985. Energy density approach to calculation of inelastic strain-stress near notches and cracks. *Engineering Fracture Mechanics* 22, 485–508. [https://doi.org/10.1016/0013-7944\(85\)90148-1](https://doi.org/10.1016/0013-7944(85)90148-1)
- [GRO11] Gross, D., Seelig, T., 2011. *Fracture Mechanics: With an Introduction to Micromechanics*, Mechanical Engineering Series. Springer Berlin Heidelberg, Berlin, Heidelberg. <https://doi.org/10.1007/978-3-642-19240-1>
- [HEI84] Heitmann, H.H., Vehoff, H., Neumann, P., 1984. Life prediction for random load fatigue based on the growth behavior of microcracks, in: *Fracture 84*. Elsevier, pp. 3599–3606. <https://doi.org/10.1016/B978-1-4832-8440-8.50388-4>
- [HOF13] Hoffmann, J., Sathish, S., Shell, E.B., Fassbender, S., Meyendorf, N., 2004. Acoustic Imaging Techniques for Characterization of Corrosion, Corrosion Protective Coatings, and Surface Cracks, in: Meyendorf, N.G.H., Nagy, P.B., Rokhlin, S.I. (Eds.), *Nondestructive Materials Characterization*, Springer Series in Materials Science. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 294–322. https://doi.org/10.1007/978-3-662-08988-0_11
- [JES22] Jesus, J.S., Antunes, F.V., Prates, P., Branco, R., Antunes, P.V., Borrego, L.P., Neto, D.M., 2022. Influence of specimen orientation on fatigue crack growth in 7050-T7451 and 2050-T8 aluminium alloys. *International Journal of Fatigue* 164, 107136. <https://doi.org/10.1016/j.ijfatigue.2022.107136>
- [KAL23] Kalina, M., Schöne, V., Spak, B., Paysan, F., Breitbarth, E., Kästner, M., 2023. Fatigue crack growth in anisotropic aluminium sheets — phase-field modelling and experimental validation. *International Journal of Fatigue* 176, 107874. <https://doi.org/10.1016/j.ijfatigue.2023.107874>

- [KAL25] Kalina, M., Schneider, T., Waisman, H., Kästner, M., 2025. Phase-field models for ductile fatigue fracture. *Theoretical and Applied Fracture Mechanics* 136, 104842. <https://doi.org/10.1016/j.tafmec.2024.104842>
- [MIE10] Miehe, C., Welschinger, F., Hofacker, M., 2010. Thermodynamically consistent phase-field models of fracture: Variational principles and multi-field FE implementations. *International Journal for Numerical Methods in Engineering* 83, 1273–1311. <https://doi.org/10.1002/nme.2861>
- [NTS88] NTSB, 1988. Aloha Airlines, Flight 243 (No. NTSB/AAR-89/03). National Transportation Safety Board.
- [PAR63] Paris, P., Erdogan, F., 1963. A Critical Analysis of Crack Propagation Laws. *Journal of Basic Engineering* 85, 528. <https://doi.org/10.1115/1.3656900>
- [SEI20] Seiler, M., Linse, T., Hantschke, P., Kästner, M., 2020. An efficient phase-field model for fatigue fracture in ductile materials. *Engineering Fracture Mechanics* 224, 106807. <https://doi.org/10.1016/j.engfracmech.2019.106807>
- [SWI94] Swift, T., 1994. Damage tolerance capability. *International Journal of Fatigue* 16, 75–94. [https://doi.org/10.1016/0142-1123\(94\)90446-4](https://doi.org/10.1016/0142-1123(94)90446-4)
- [TEI17] Teichtmeister, S., Kienle, D., Aldakheel, F., Keip, M.-A., 2017. Phase field modeling of fracture in anisotropic brittle solids. *International Journal of Non-Linear Mechanics* 97, 1–21. <https://doi.org/10.1016/j.ijnonlinmec.2017.06.018>



Dr. Taha Soliman (31) lives in Cambridge and works at Arm, a leading semiconductor and software design company, on edge AI performance topics. He has an international academic background in embedded systems, having studied at a number of prestigious universities. Prior to Arm he joined Bosch Research in 2019 where he completed his PhD with summa cum laude on energy-efficient hardware for AI, and advanced to an expert role in AI acceleration and hardware/software co-design. Before joining Bosch, he gained extensive research and industry experience at various research institutes.

Architecting Efficiency for Sustainable Edge AI

Taha Soliman

Architecting Efficiency for Sustainable Edge AI

AI is actively reshaping how we design and interact with systems across a vast spectrum of domains, from consumer electronics and smart homes to industrial automation and next-generation automotive systems. However, this success has also illuminated a new frontier: on-device, or “embedded,” AI. As we seek to build more autonomous, responsive, and private systems, the reliance on a constant connection to a distant cloud server becomes a critical liability.

The memory wall as basic challenge: a widening performance gap between the processing units and the memory systems that feed them

It is at this edge where the harsh realities of embedded deployment come to the forefront. Engineers are no longer working with the infinite resources of a data center, but are instead confronted with tight power budgets, constrained on-chip memory, and strict real-time latency requirements in many cases in the order of milliseconds. Furthermore, in domains like medical devices and autonomous vehicles, these systems must operate under stringent safety and security constraints. (Figure 1 → p. 56)

These demanding embedded requirements are on a collision course with a fundamental challenge that has plagued hardware design for decades: the “memory wall” Bavikadi et al. (2020). This term describes the ever-widening performance gap between the processing units and the memory systems that feed them. At its core, the problem is one of imbalanced progress. For decades, the performance of microprocessors, driven by Moore’s Law, has improved at a staggering rate of nearly double every 18 months. In stark contrast, the bandwidth of the commodity Dynamic Random Access Memory (DRAM) memory they rely on has improved by a mere 7 percent per year according to Lu et al. (2012).

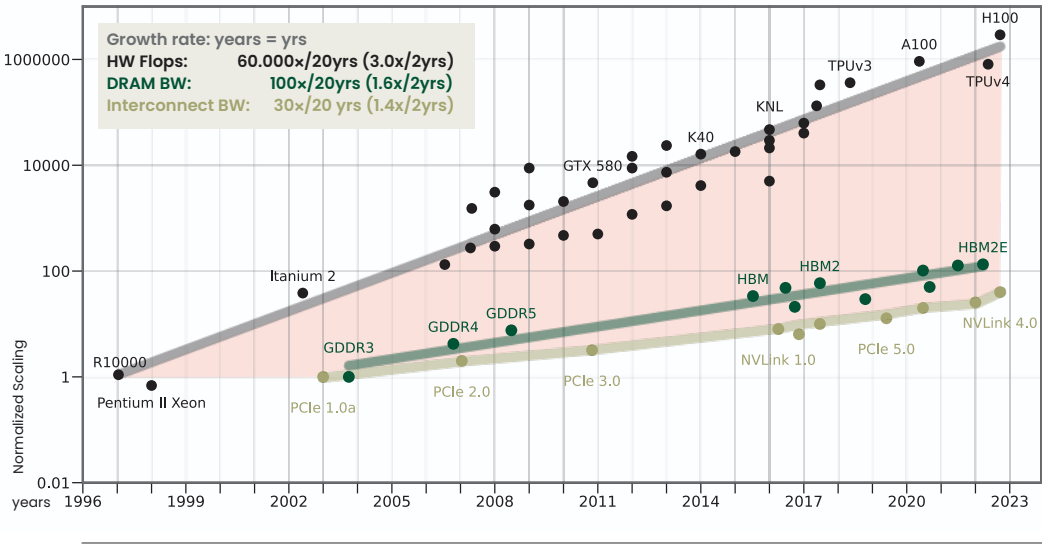
Resume:

While the foundational concepts of Artificial Intelligence (AI) date back to the pioneering work of researchers like Rosenblatt in the 1950s Rosenblatt (1957, 1958), the last decade has marked a pivotal shift. AI has transitioned from a specialized academic pursuit into a dependable, mainstream technology. This remarkable transformation was not accidental; it was propelled by a confluence of two critical factors. First, the dawn of the “big data” era provided an unprecedented volume of information, creating massive datasets essential for training and refining sophisticated Machine Learning (ML) algorithms. Second, the parallel evolution of cloud computing delivered seemingly limitless computational power and storage, allowing these data-hungry models to be trained and deployed at a global scale. As a result of this cloud-powered revolution, AI has successfully moved from laboratory prototypes into the fabric of our everyday products.

Figure 1 Growth Rate of Computing Power vs. Growth Rate of Data Movement

Computing power (black) is scaling far faster than data movement – memory bandwidth and chip-to-chip interconnect bandwidth (green and olive) – creating a growing gap where processors can compute much more than memory systems and interconnects can feed them.

Abbreviations:
BW: Bandwidth
HW: Hardware
Flops: Floating Point Operations Per Second
DRAM: Dynamic Random Access Memory



Arithmetic Logic Unit (ALU), Register File, PE, Buffer and DRAM

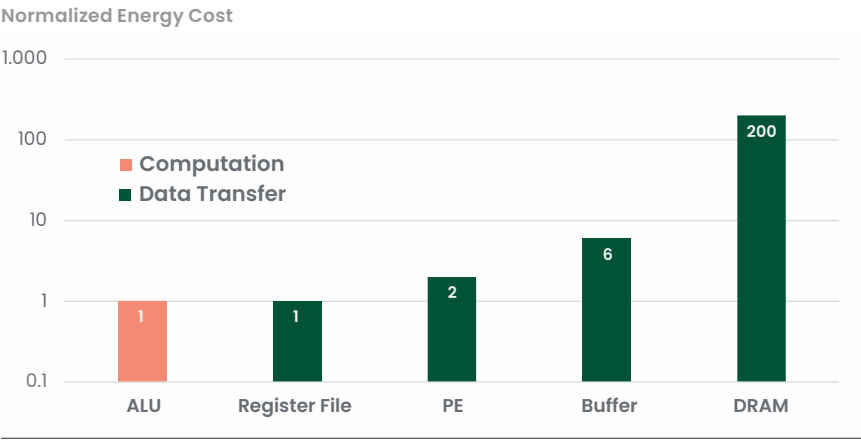
In a traditional computing paradigm, this means the processor spends an increasing amount of its time idle, waiting for data to arrive from slow, off-chip memory. For the massively parallel and data-intensive workloads of modern AI, this is not just a bottleneck; it is a catastrophic impediment to efficiency.

The theoretical problem of the memory wall becomes a tangible energy crisis when deploying Deep Neural Networks (DNNs). The sheer volume of weights and intermediate activations in a typical DNN far exceeds the capacity of small, fast on-chip memories.

This forces the processor to constantly fetch data from power-hungry off-chip DRAM. As illustrated in Figure 2 → opposite page, the energy cost of the actual computation – a single Multiply-Accumulate (MAC) operation – is minuscule. The real cost is in the data movement. Loading the operands for that single MAC from external memory can consume up to 200 times more energy than the operation itself. While a memory hierarchy exists – with faster, more energy-efficient caches and SRAM closer to the Arithmetic Logic Unit (ALU) – these memories are, by nature, significantly smaller. This physical size limitation restricts opportunities for data reuse, forcing the system back into the costly cycle of off-chip access and undermining overall utilization.

Data Movement Dominates Energy Consumption

Accessing off-chip memory is far more energy-intensive than doing arithmetic operations.



The solution: A new holistic approach to the relationship between hardware and software

Consequently, the incredible potential of DNNs remains largely to be explored for battery-powered devices and autonomous systems where offloading to the cloud is not an option due to latency, security, or connectivity limitations. Overcoming this profound efficiency barrier is not a task that can be solved by a single silver-bullet technique. A slightly better algorithm or a marginally faster chip will only yield incremental gains. The path forward requires a holistic, integrated approach that re-imagines the relationship between hardware and software.

In this article, we will explore this challenge and its solutions through three complementary and deeply interconnected lenses:

1. Emerging Hardware Architectures:

The first pillar is the reinvention of the hardware itself. We will delve into the rise of AI computing, where general-purpose CPUs are augmented with specialized accelerators like Neural Processing Units (NPUs) and dataflow engines. A key focus of these new architectures is the development of deep on-chip memory hierarchies and tiling capabilities to maximize data reuse and minimize costly off-chip transfers. The main focus of this section will be on exploring paradigm-shifting approaches like In-Memory Computing

Figure 2

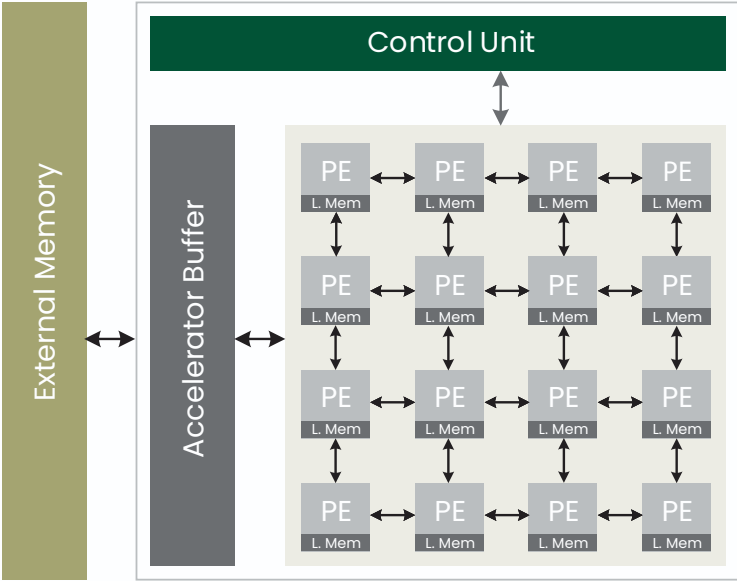
The cost of data movement for the targeted Multiply-Accumulate (MAC) operation across memory hierarchy compared to the mac operation itself. Chen et al. (2016).

Abbreviations:

Arithmetic Logic Unit (ALU) computation is the cheapest operation (normalized to 1, as on the right). Moving data inside the register file (RF) costs about the same as computation. Transferring data to a processing element (PE) costs roughly 2× more energy. Moving data through an on-chip buffer costs about 6× more energy. Accessing dynamic random-access memory is extremely expensive, costing about 200× more energy than a computation.

Figure 3
Neural Processing Unit (NPU) systolic architecture depends on parallelizing the Multiply-Accumulate (MAC) operations using interconnected Processing Elements (PE) and leveraging memory hierarchy for maximizing data reuse.

Neural Processing Unit (NPU) Systolic Architecture



(IMC), which seeks to obliterate the memory wall by performing computations directly where the data is stored, promising orders-of-magnitude gains in energy efficiency.

2. Hardware-Aware Software

Optimization:

The second pillar involves sculpting the software to perfectly fit the contours of the hardware. It is no longer enough to simply design an accurate model; it must be an efficient model. We will discuss a suite of powerful optimization techniques, including quantization to reduce the numerical precision of the

model, and structured pruning to remove redundant components in a hardware-friendly manner. Furthermore, we will examine how advanced techniques like Hardware-Aware Neural Architecture Search (NAS) can automate the discovery of novel AI models that are inherently optimized for the arithmetic units, memory layout, and dataflow of a specific target device.

3. Hardware-Software Co-design:

The third and most powerful pillar is the synthesis of the first two. Here, we explore how hardware and software can influence

each other leading to the ultimate goal of co-design. This is an iterative process that begins with system-level constraints – latency, power, accuracy – and jointly optimizes the AI model and the hardware architecture. This could mean modifying tensor shapes in the algorithm to fit perfectly within on-chip memory tiles, or adding a small, specialized unit to the hardware to accelerate a critical operator in the model. In this case, hardware and software evolve together to create a system that is far more efficient and performant than what could ever be achieved by optimizing each in isolation.

Emerging Hardware Architectures

As previously highlighted, given the requirements across diverse edge use cases, the migration from general-purpose computing on CPUs and GPUs to specialized accelerators has been inevitable. Spatial dataflow accelerators—often branded as Neural Processing Units (NPUs) push computation into regular arrays of MAC units with carefully orchestrated movement of weights and activations, adopting systolic or near-systolic execution to amortize data movement over many operations as shown in *Figure 3 → left page*. This architectural shift is not about chasing marketed peak Tera Operations per Seconds (TOPS); it is about sustaining throughput under tight memory and energy budgets.

As AI models have grown, their memory footprints have expanded accordingly, placing memory-centric design at the heart of modern accelerators. Because off-chip DRAM

accesses dominate energy and latency, edge-class NPUs build deep on-chip hierarchies composed of register files, multi-banked scratchpad Static Random Access Memory (SRAM), and local buffers attached to compute arrays. Unlike cache-based systems that rely on probabilistic reuse, scratchpads are explicitly managed by the compiler/runtime, enabling deterministic latency and avoiding tag lookup energy.

Two techniques recur across efficient designs: tiling and double buffering. Tiling partitions tensors into sub-blocks that fit within SRAM, maximizing local reuse of weights and partial sums before committing results to higher levels. Double buffering overlaps Direct Memory Access (DMA) transfers with compute, ensuring the array remains fed with data without stalls. Also, the Network on Chip (NoC) connecting these memories to compute arrays is equally critical: multicast which refers to using a single input across different computational units, reduces bandwidth. These techniques deliberately exploit the deterministic nature of inference graphs, which are fully known at compilation time.

Another heavily explored design optimization is maximizing data reuse, which can strongly influence memory traffic whether on weights, inputs, or outputs. Weight-stationary schedules pin weights near MAC arrays



As AI models have grown, their memory footprints have expanded accordingly, placing memory-centric design at the heart of modern accelerators.

TAHA SOLIMAN

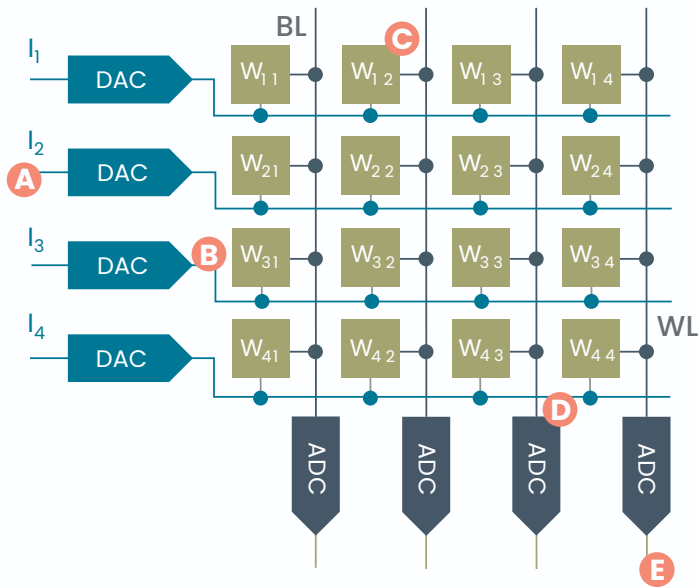


Figure 4

A: The digital activation of the computed layer.
B: DAC convert the digital input into an analog signal to be applied to the memory cell.
C: A memory cell stores the kernel value of the currently computed layer.
D: The summation line accumulates the resulting signal out of the memory cell representing the operation results.
E: ADC convert back the result into the digital domain for further processing.

Abbreviations:
ADC: Analog to Digital Converter
BL: Bit Line
DAC: Digital to Analog Converter
I: Input
W: Weight
WL: Word Line

The General Concept of the In-memory Crossbar



to exploit their reuse across many input tiles. Output-stationary keeps partial sums local, minimizing costly writes and reads of accumulators. Row-stationary, as exemplified by Eyeriss, balances reuse of inputs, weights, and outputs with flexible mapping to multiple array dimensions Chen et al. (2016).

Nevertheless, there is no universally optimal dataflow: convolutions with large kernels and channels favor one strategy, depth-wise or group convolutions favor another, and General Matrix Multiply (GEMM)-like layers can be tuned based on tile aspect ratios. Modern compilers expose dataflow as a

scheduling decision, selecting the mapping that minimizes total bytes moved across all hierarchy levels for the given operator mix and SRAM sizes.

Despite all these advances and optimization ideas, the memory wall remains formidable whenever tensors exceed on-chip capacity or when operator shapes prevent efficient reuse. This reality motivates moving compute closer to memory. Processing near/in memory augments memory with simple compute near banks, offering massive internal bandwidth for operations like reductions and bit-wise transforms. This can offload memory-

Benchmark of a proposed Ferroelectric Field-Effect Transistor (FeFET)

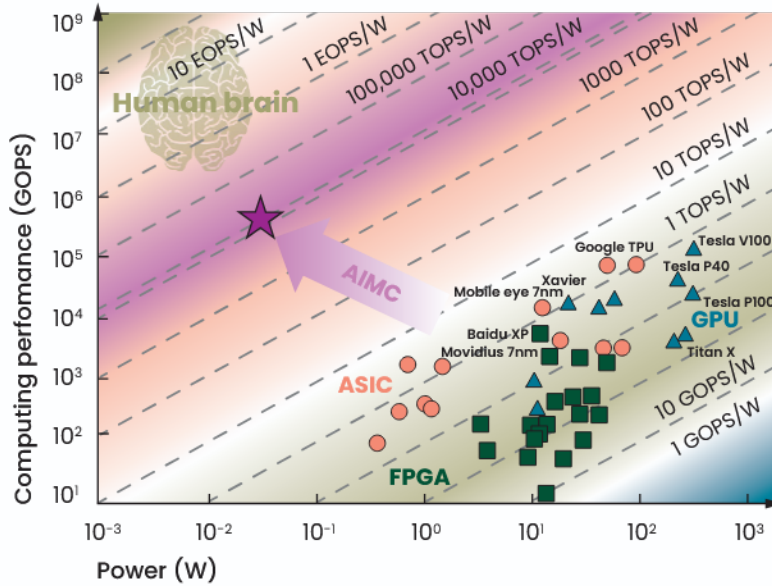


Figure 5

Benchmark of a proposed Ferroelectric Field-Effect Transistor (**FeFET**) based IMC Soliman et al. (2020) with existing solutions with respect to power efficiency and performance. A possible performance can be reached based on such accelerator concept reaching a power efficiency of > 10000 TOPS/W.

Abbreviations:

AIMC: Analog In-memory Computing

ASIC: Application-Specific Integrated Circuit

EOPS/W: Exa Operations per Second per Watt

FeFET Ferroelectric Field-Effect Transistor

FPGA: Field-Programmable Gate Array

GOPS: Giga Operations per Second

GPU: Graphics Processing Unit

TOPS/W: Tera Operations per Second per Watt

W: Watt

bound kernels and reduce host-memory traffic, but programmability, consistency models, and integration into the existing toolchain are active challenges; moreover, gains depend on workloads that align with the limited arithmetic supported in-memory. Several commercial and research efforts already explore this model, including High Bandwidth Memory (HBM)-Processing-in-Memory (PIM)-style solutions and Dual Inline Memory Module (DIMM)-based DRAM-Processing-in-Memory (PIM), as in Kim et al. (2023); Lv et al. (2025).

In-memory computing (IMC) goes further, attempting to perform arithmetic directly within memory arrays. SRAM-based IMC leverages bitline operations and sense amplifier behavior to implement low-precision vector-matrix multiplications and logic, while resistive memory (ReRAM/PCM) arrays exploit analog dot-product properties via conductance states. IMC offers compelling locality: weights reside where computation occurs, theoretically collapsing the memory wall. However, practical embedded deployments face several hurdles: Analog-to-Digital Converter (ADC)/DAC interfaces dominate energy and area at higher

Sparsity pattern matrix multiplication

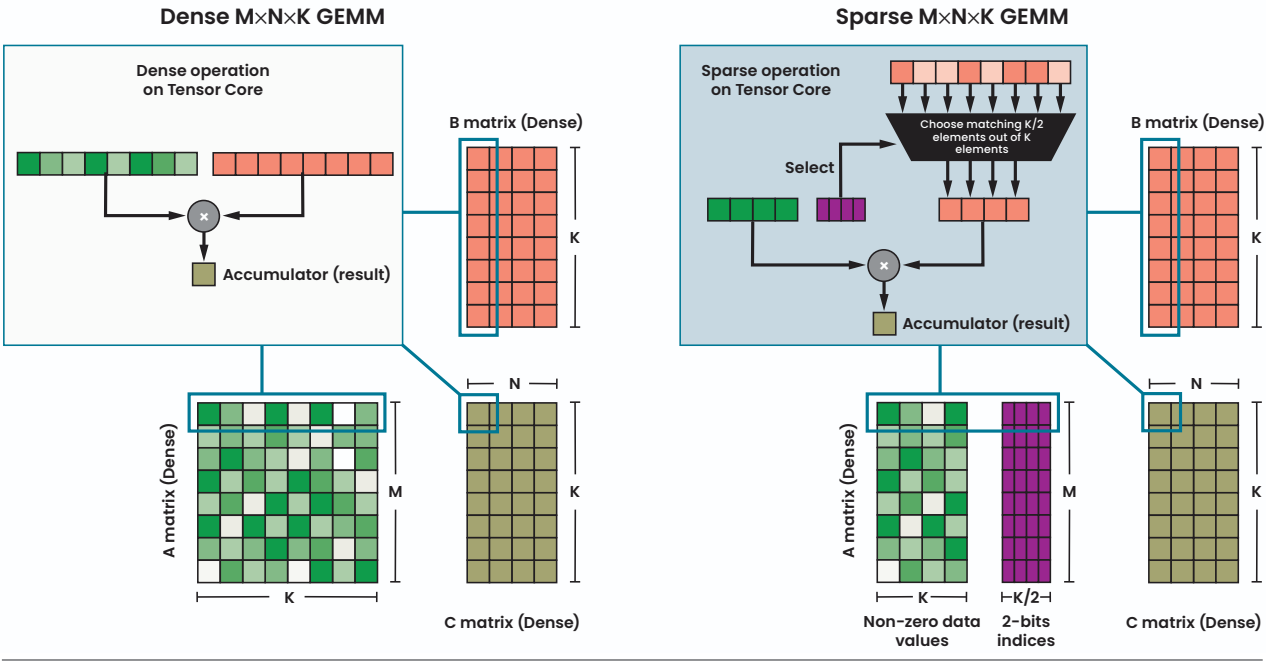


Figure 6 above
Illustration of 2:4 sparsity pattern matrix multiplication on NVIDIA GPUs. Adapted from Mishra et al. (2021).

precisions; device variability, nonlinearity, and drift require calibration and often constrain usable precision; endurance and write energy limit usage for training or frequent updates; and mapping realistic DNNs with diverse operators to IMC-friendly kernels demands nontrivial compiler support. As a result, near-term IMC wins are concentrated in inference workloads with static weights, moderate precision, and carefully selected layers, often paired with conventional digital logic for control and non-linearities. (Figure 4 → p. 60)

Even in these regimes, the key metric remains end-to-end energy per inference; any benefit from “free” MACs must outweigh converter overheads and control complexity. IMC aligns similarly with innovations in emerging technologies for example the work that utilizes emerging memory technologies such as FeFET leads to unprecedented performance as shown in Figure 5 → p. 61.

Another increasingly important architectural theme is explicit support for approximate computing. Error-bounded approximations to MAC operations – via inexact multipliers/

adders, operand truncation, or voltage underscaling with lightweight error detection – can reduce energy and ease the design of computational units, while sometimes easing memory requirements through compressed or downscaled representations. Real benefits, however, depend on algorithmic tolerance and careful placement: approximations must be applied to layers and tensors with demonstrated robustness. Consistent speed-ups and savings require compiler awareness of hardware contracts so that kernels are scheduled where approximations are acceptable and fused appropriately with quantization and sparsity mechanisms, a preview of the co-design arguments we develop later. Soliman et al. (2022); De la Parra et al. (2021).

Taken together, emerging architectures for embedded AI converge on several design principles. First, prioritize locality: maximize reuse through dataflow scheduling, tiling, and scratchpad management before adding compute units. Second, expose predictable behavior: deterministic latencies and explicit memory control beat opaque caches in safety-critical and hard real-time contexts. Third, embrace precision diversity: offer efficient numerical representations and mixed-precision paths aligned with realistic quantization strategies, including per-channel scaling and fused requantization. Fourth, consider memory-centric accelerations which are promising, but their true value is measured at the system level, accounting for converter overheads, control, and toolchain complexity. Finally, design the compiler and runtime as first-class citizens: the best hardware is ineffective without software that can

articulate dataflows, select tile sizes, orchestrate DMA, and align operator implementations with the physical constraints of the memory hierarchy.

Hardware-Aware Software Optimization

On embedded devices, the most effective software optimizations are those that explicitly track what the hardware can execute efficiently. In this section, the main focus will lie on three main directions, quantization, pruning, and hardware-aware neural architecture search (NAS), which all reduce the memory and computational requirements allowing for reshaping the models to fit the arithmetic units, memory hierarchy, and dataflow supported by the target. Their benefits are not abstract; they materialize only when formats, sparsity patterns, and architectures align with integer MAC pipelines, accumulator widths, scratchpad sizes, DMA behavior, and compiler lowering on the device.

Quantization interchanges floating-point representation with a reduced precision one, cutting compute energy and memory traffic. The main idea here is how to represent the values of the features and weights using the lower representation without any significant accuracy loss. In that sense, two directions are usually followed. The First one is Post-training quantization (PTQ) which is at-



The most effective software optimizations are those that explicitly track what the hardware can execute efficiently.

TAHA SOLIMAN



tractive for fast deployment when activation statistics are well behaved.

Nevertheless, it can be difficult to map the values directly to the lower representation without significant accuracy loss. The second technique is quantization-aware training (QAT) which recovers accuracy in more sensitive layers and enables lower precisions e.g., Integer 4 (INT4) by letting the model adapt to quantization noise. Mixed-precision regimes that allows lower representation e.g., Floating Point 4 (FP4) INT4 for most layers, Brain Floating Point 16 (bfloat16)/FP16/FP8/INT8 for fragile blocks is driven by per-layer sensitivity and constrained by what the device implements natively. All these optimizations here are dependent on the compiler that can propagate scales, choose layouts that match vector widths and banking, and fuse dequant/requant with adjacent ops; otherwise, conversions and unfused activations erase much of the theoretical savings.

Pruning aims at nullifying weights on the premise that hardware can skip those computations accordingly. Unstructured pruning can yield high sparsity in weights, yet its irregular patterns rarely map to embedded NPU efficiently while only a limited numbers of architecture currently support it without significant overhead that makes the gains obsolete. The other possibility is structured pruning which drops entire channels, filters, or blocks and accordingly produces complete blocks that can be skipped; nevertheless, to maximize the gains, those pruned parts need to be studied in the context of fitting SRAM tiles and sustaining array uti-

lization Liu et al. (2022). Another option is when the device supports N:M sparsity (e.g., 2:4) (*Figure 6* → p. 62) with dedicated decode hardware Mishra et al. (2021), enforcing that pattern during training or fine-tuning unlocks real throughput and bandwidth gains. Compressed formats and block encodings must be selected to keep index overhead low and memory accesses coalesced; otherwise, sparse data degenerates into scattered loads. In practice, pruning is paired with QAT to recover accuracy and with quantization to compound bandwidth savings.

Hardware-aware NAS pushes these ideas upstream by searching model spaces with hardware in the loop. (*Figure 6* → p. 62) In this case, objectives include not just accuracy but latency, peak SRAM footprint, and energy per inference, measured on the actual device or via calibrated proxies. Search spaces are constrained to operators and channel/kernel configurations that the hardware and compiler can lower without spills, emphasizing architectures that tile cleanly into the memory and align with the hardware dimensions. Hardware in the loop workflows ensure that fusion, layout, and tiling decisions applied during lowering are reflected in the search evaluation, avoiding models that appear efficient in isolation but trigger off-chip traffic after deployment.

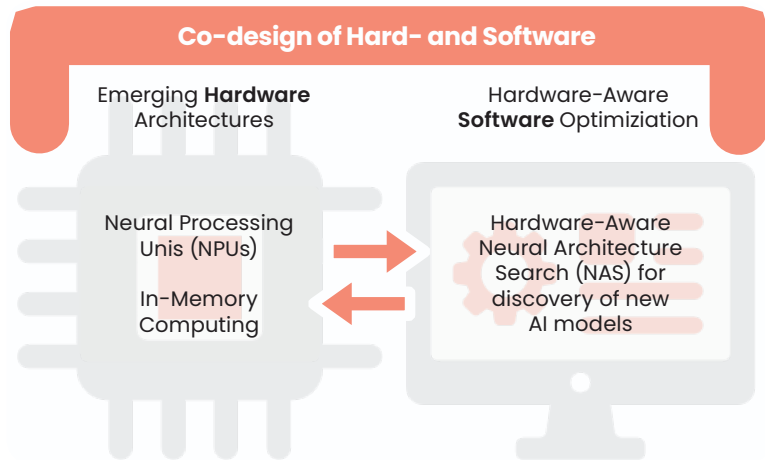
There have been many techniques presented starting from two stage HW-NAS as well as single stage Benmeziane et al. (2021) but in all these methods the main goal and technique remain the same, which is how to find the best variation of networks that maintains

the targeted accuracy while matching different hardware preferences in terms of parallelization schemes, size and distribution across the different layers.

In combination, these hardware-aware techniques offer a practical recipe: quantify device constraints, apply the data representation that maximizes the gain, prune with patterns the accelerator can exploit, and, when warranted, use NAS with measurement feedback to converge on architectures that fit tiles and sustain utilization. The common thread is alignment with what the hardware actually does well – integer arithmetic with fused data paths, structured sparsity with efficient decode, and operators that respect on-chip memory boundaries. When formats, sparsity, and architectures are chosen with those realities in mind, software optimizations reliably translate into fewer data movement, better utilization, and consistent real-time performance on embedded targets.

Co-design of Hardware and Software

True efficiency at the embedded edge emerges when models and machines are shaped together. In a co-design flow, we begin from system constraints – latency budgets, peak power, SRAM limits, accuracy floors – and jointly refine the neural network, the compiler schedule, and the accelerator micro-architecture so that data reuse, precision, and execution order align with what the hardware can do at high utilization and low energy. Rather than treating hardware as fixed and software as an afterthought, co-design



elevates hardware parameters (array dimensions, buffer sizes, dataflow choices, sparsity engines, numeric formats) to tunable variables, and gives software the authority to suggest which knobs matter.

Hardware-aware NAS is a natural vehicle for this joint exploration. Beyond searching operator choices and channel/kernel configurations, the search space can include hardware parameters that shape memory traffic and utilization: MAC array height/width (which set preferred blocking factors for channels and tiles), the size and banking of on-chip SRAM scratchpads (which dictate feasible tile sizes and double-buffering depth), supported dataflows (weight-/output-/row-stationary), DMA burst sizes, and even which precisions and sparsity patterns are implemented in the datapath Rai et al. (2024). Each candidate architecture is compiled with the real lowering pipeline, scheduled into tiles that fit the

Edge AI race will not be won
by raw TOPS (Tera Operations
per Seconds).

TAHA SOLIMAN



proposed buffers, and evaluated with cycle- and byte-accurate cost models (periodically validated on silicon or an instruction-accurate simulator). The outcome is a Pareto set of model–hardware pairs that meet accuracy while minimizing latency and energy under explicit memory constraints, rather than a single model contorted to an arbitrary fixed array.

Quantization likewise benefits from co-design. The software side identifies the minimal precisions each layer can tolerate (via PTQ/QAT and sensitivity analysis) and proposes mixed-precision boundaries; the hardware side answers with the most area- and energy-efficient way to implement them. Concretely, this often yields: per-channel weight scaling paths, accumulators with configurable saturation, and fused dequant/requant instructions that combine bias addition, scaling, and activation in a single pass Jacob et al. (2017).

Pruning is equally a joint exercise. The model proposes which structures to remove; the hardware specifies which sparsity contracts deliver most speedups. If the accelerator implements N:M sparsity (e.g., 2:4) with dedicated decode near the MACs, the training loop enforces that pattern and the compiler keeps weights in the native compressed format end-to-end. For activation sparsity, hardware can add lightweight run-length or block compression units at producer edges and corresponding consumers that maintain coalesced accesses.

Without these micro-architectural accommodations and compiler awareness, pruning loses its advantages as irregular memory access compromises the gains.

A practical co-design loop ties these threads together. Starting from a baseline accelerator template and a candidate model, the compiler maps each layer to a dataflow schedule and tile splitting that fits the current SRAM partitions, inserting double buffering and fusing ops where profitable. This architecture is then profiled to expose the dominant memory transfers and stalls. The architecture and the model are then refactored to relieve the bottleneck: either by adjusting the model, such as channel counts to multiples of the array height, splitting or fusing layers to keep activations on-chip, or shifting a sensitive block to a different precision with QAT to enable more aggressive fusion. On the other hand, small hardware tweaks are often decisive: a slightly larger partial-sum buffer to keep accumulators local; a new fused requant activation instruction; wider DMA bursts to match tile strides; or a sparsity decode unit tailored to the trained N:M pattern. Because inference graphs are static, these changes yield deterministic schedules and predictable worst-case latency—critical for safety and real-time guarantees.

The same philosophy extends to memory-centric architectures. If IMC features are under consideration, the compiler identifies kernels whose arithmetic and precision align with near-/in-memory operators, while the hardware exposes converter resolution, batch sizes, and placement that make the

offload worthwhile. For IMC specifically, co-design may reduce layer precision to match ADC/DAC sweet spots, restructure operators to maximize on-array reuse, and confine analog compute to static-weight, linear blocks, leaving control and nonlinearities to digital logic.

Conclusion

Edge AI race will not be won by raw TOPS. This is an apparent conclusion for many technology giants in the market. The most promising path is an understanding how the complete stack where hardware and software can be designed together: accelerators

with deep on-chip hierarchies and optimized dataflows; software that embraces hardware realities through quantization and structured pruning; and hardware-aware NAS that searches only what tiles cleanly and sustains utilization. Co-design then closes the loop, letting mixed-precision boundaries, sparsity patterns, and even array and buffer dimensions co-evolve with the network and compiler schedule to meet explicit latency, energy, and cost targets. Measured at the system level – data moved, joules per inference, and worst-case deadlines – this integrated approach turns memory from a wall into a collaborator, delivering reliable, real-time performance on embedded devices while preserving the accuracy that makes modern AI useful.

References

- Bavikadi, S., P. R. Sutradhar, K. N. Khasawneh, A. Ganguly, and S. M. Pudukotai Dinakarrao, 2020: A review of in-memory computing architectures for machine learning applications. *Proceedings of the 2020 on Great Lakes Symposium on VLSI*, Association for Computing Machinery, New York, NY, USA, 89–94, GLSVLSI '20, doi:10.1145/3386263.3407649, URL <https://doi.org/10.1145/3386263.3407649>.
- Benmeziame, H., K. E. Maghraoui, H. Ouarnoughi, S. Niar, M. Wistuba, and N. Wang, 2021: A comprehensive survey on hardware-aware neural architecture search. <https://arxiv.org/abs/2101.09336>, 2101.09336.
- Chen, Y.-H., J. Emer, and V. Sze, 2016: Eyeriss: A spatial architecture for energy-efficient dataflow for convolutional neural networks. *2016 ACM/IEEE 43rd Annual International Symposium on Computer Architecture (ISCA)*, 367–379, doi:10.1109/ISCA.2016.40.
- De la Parra, C., A. El-Yamany, T. Soliman, A. Kumar, N. Wehn, and A. Guntoro, 2021: Exploiting resiliency for kernel-wise cnn approximation enabled by adaptive hardware design. *2021 IEEE International Symposium on Circuits and Systems (ISCAS)*, 1–5, doi: 10.1109/ISCAS51556.2021.9401517.
- Gholami, A., Z. Yao, S. Kim, C. Hooper, M. W. Mahoney, and K. Keutzer, 2024: Ai and memory wall. *IEEE Micro Journal*.
- Jacob, B., S. Kligys, B. Chen, M. Zhu, M. Tang, A. Howard, H. Adam, and D. Kalenichenko, 2017: Quantization and training of neural networks for efficient integer-arithmetic-only inference. URL <https://arxiv.org/abs/1712.05877>, 1712.05877.
- Kim, J. H., and Coauthors, 2023: Samsung pim/pnm for transfrmer based ai : Energy efficiency on pim/pnm cluster. *2023 IEEE Hot Chips 35 Symposium (HCS)*, 1–31, doi: 10.1109/HCS59251.2023.10254711.
- Liu, J., B. Zhuang, Z. Zhuang, Y. Guo, J. Huang, J. Zhu, and M. Tan, 2022: Discrimination-aware network pruning for deep model compression. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44 (8), 4035–4051, doi:10.1109/TPAMI.2021.3066410.
- Lu, S.-L., T. Karnik, G. Srinivasa, K.-Y. Chao, D. Carmean, and J. Held, 2012: Scaling the “memory wall”. *Proceedings of the International Conference on Computer-Aided Design*, Association for Computing Machinery, New York, NY, USA, 271–272, ICCAD '12, doi: 10.1145/2429384.2429437, URL <https://doi.org/10.1145/2429384.2429437>.

- Lv, Z., T. Bu, and Q. Yang, 2025: Fast, transparent and accurate simulation of thousand processing-in-memory cores. Proceedings of the Great Lakes Symposium on VLSI 2025, Association for Computing Machinery, New York, NY, USA, 29–34, GLSVLSI '25, doi: 10.1145/3716368.3735197, URL <https://doi.org/10.1145/3716368.3735197>.
- Mishra, A., J. A. Latorre, J. Pool, D. Stosic, D. Stosic, G. Venkatesh, C. Yu, and P. Micikevicius, 2021: Accelerating sparse deep neural networks. URL <https://arxiv.org/abs/2104.08378>, 2104.08378.
- Rai, S., C. De La Parra, M. Rapp, J. M. Borrmann, N. Bretz, S. Metzloff, T. Soliman, and C. Schorn, 2024: Accelerating automated driving and adas using hw/sw codesign. 2024 IEEE 37th International System-on-Chip Conference (SOCC), 1–6, doi:10.1109/SOCC62300.2024.10737753.
- Rosenblatt, F., 1957: The perceptron - a perceiving and recognizing automaton. Tech. Rep. 85-460-1, Cornell Aeronautical Laboratory, Ithaca, New York.
- Rosenblatt, F., 1958: The perceptron: a probabilistic model for information storage and organization in the brain. Psychological review, 65 (6), 386.
- Soliman, T., A. Eldebiky, C. De La Parra, A. Guntoro, and N. Wehn, 2022: Efficient hardware approximation for Bit-Decomposition based deep neural network accelerators. 2022 IEEE 35th International System-on-Chip Conference (SOCC) (SOCC 2022), Belfast, United Kingdom (Great Britain).
- Soliman, T., and Coauthors, 2020: Ultra-low power flexible precision fefet based analog in-memory computing. 2020 IEEE International Electron Devices Meeting (IEDM), 29.2.1–29.2.4, doi:10.1109/IEDM13553.2020.9372124.



***Dr. Johannes Tiedau** arbeitet als wissenschaftlicher Mitarbeiter im Team von Ekkehard Peik an der Physikalisch-Technischen Bundesanstalt in Braunschweig. In seiner Forschung beschäftigt er sich mit der Entwicklung einer neuartigen Atomuhr, der sogenannten Thorium-Kernuhr. Nach seinem Studium an der Leibniz Universität Hannover mit Station an der University of Oxford wurde er 2020 an der Universität Paderborn mit einer Arbeit zur Quantenoptik promoviert. Seine Laufbahn ist durch zahlreiche Auszeichnungen geprägt, darunter der Preis des Präsidiums der Leibniz Universität, Stipendien der Studienstiftung und des Deutschen Akademischen Austauschdienstes (DAAD).*

Die Kernuhr

Johannes Tiedau

Von Atom- und Kernuhren

Die Zeit ist von allen physikalischen Größen diejenige, die am genauesten gemessen werden kann. Mittlerweile können unvorstellbare relative Unsicherheiten im Bereich von 10^{-19} erreicht werden.^[1, 2] Das bedeutet, diese Uhr würde während der gesamten Lebensdauer des Universums weniger als eine 1 Sekunde nachgehen. Bei dieser Betrachtung zwingt sich förmlich die Frage auf, wofür noch bessere Uhren überhaupt benötigt werden. Wenn Sie sich mit einem Freund auf einen Kaffee treffen möchten, ist eine Uhrzeitangabe wie 15:30 Uhr und 3,31293847982293953 Sekunden jedenfalls weder praktikabel noch hilfreich. Dennoch sind moderne Uhren aus unserem Alltag fast nicht mehr wegzudenken, obwohl die wenigsten über sie nachdenken werden. Stellen Sie sich vor, dass Sie sich für 15:30 Uhr auf einen Kaffee verabredet haben, aber nicht wissen, wie Sie zu dem Ort kommen. Die meisten Menschen nutzen mittlerweile für diese Fragestellung ihr Smartphone und lassen sich zum gewünschten Ort leiten. Aber wie weiß es, wo wir uns befinden? Dafür spielt wiederum die Zeitmessung eine entscheidende Rolle. Mithilfe zahlreicher Satelliten, die unsere Erde umrunden und hochgenaue Zeitsignale aussenden, können unsere Mobiltelefone berechnen, wie weit sie von den Satelliten entfernt sind. Da die Position der Satelliten bekannt ist, kann unser Standort daraus berechnet werden. Der zentrale Punkt ist, dass die Zeitsignale hier hinreichend genau sein müssen. Eine Ortsangabe eines Navigationsgerätes

in Form von „Sie befinden sich irgendwo in Deutschland“ wäre wohl in den wenigsten Fällen hilfreich. Für eine Standortbestimmung mit wenigen Metern Messunsicherheit sollten die Uhren eine Unsicherheit von kleiner als 10^{-13} aufweisen. In einer ähnlichen Größenordnung muss die Zeit auch für die Synchronisation von Kraftwerken und Telekommunikationsnetzen bereitgestellt werden. Noch bessere Uhren benötigt man für die Grundlagenforschung, und sogar Höhenunterschiede auf der Erde können mittlerweile mit Uhren gemessen werden.

Zusammenfassung

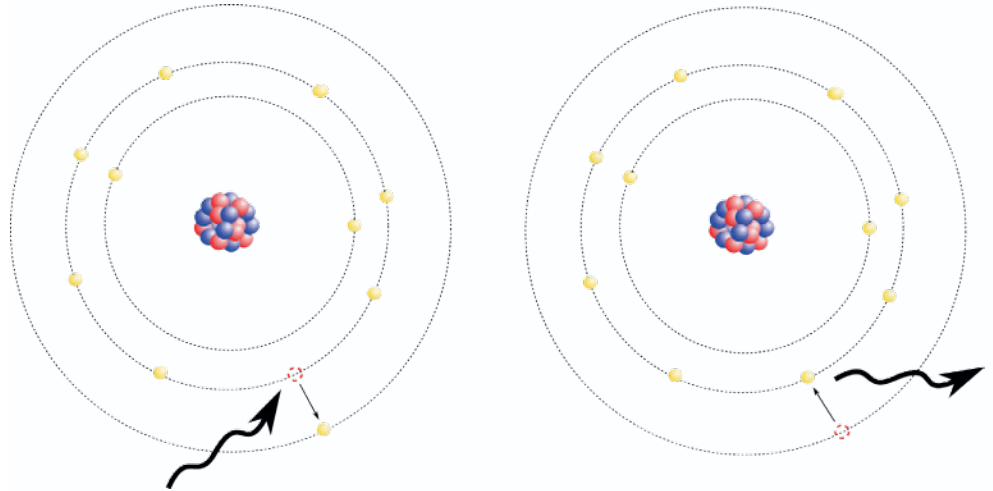
Eine hochgenaue Zeitmessung ist für viele Bereiche wie die Satellitennavigation (z.B. GPS), aber auch bei der Synchronisierung von Daten und Stromnetzen nicht mehr wegzudenken. Die derzeit besten Uhren basieren auf Übergängen in der Elektronenhülle von Atomen. Ein neuer Ansatz für die Entwicklung noch besserer und möglicherweise kompakterer Uhren ist es, einen Übergang in einem Atomkern als Referenz zu verwenden. Von allen bekannten Elementen bietet nur Thorium-229 die Möglichkeit, eine solche Kernuhr zu bauen. Seit der ersten Laseranregung des Atomkerns 2023 an der Physikalisch-Technischen Bundesanstalt (PTB) in Zusammenarbeit mit der Technischen Universität (TU) Wien erfährt dieses Forschungsfeld große Aufmerksamkeit und einen rasanten Fortschritt. Zukünftig könnte der Kernübergang in Thorium helfen, neue physikalische Zusammenhänge zu entdecken.

Abbildung 1

Links: Licht wird von einem Atom absorbiert. Ein **Elektron** (gelb) wird dadurch in einen höheren Energiezustand befördert. Die Protonen und Neutronen des Atomkerns sind als blaue und rote Kugeln illustriert.

Rechts: Durch die Abgabe von Energie sendet ein Atom Licht aus. Dieser Vorgang wird als Lumineszenz bezeichnet. Bei einer Kernuhr absorbiert und emittiert der Atomkern das Licht.

Quelle: Grafik PTB

Aufnahme und Abgabe von Licht durch ein Atom

Durch die allgemeine Relativitätstheorie ist bekannt, dass Uhren in der Nähe von schweren Körpern wie unserer Erde langsamer gehen. Dieser Effekt ist sehr klein und bedeutet in etwa eine relative Verschiebung von 10^{-16} pro Meter Höhenunterschied. Dennoch sind diese Effekte mittlerweile messbar und könnten bald zur Kartierung verwendet werden. In Deutschland ist die Physikalisch-Technische Bundesanstalt (PTB) für die Bereitstellung der physikalischen Einheiten wie Ampere, Meter, Kilogramm usw. (SI-Einheiten) verantwortlich – und damit insbesondere der Sekunde.

Wie kann man diese Genauigkeit technisch realisieren?

Die Menschheit versucht schon seit Jahrtausenden, Uhren zu entwickeln, die besonders Genauigkeit, Verfügbarkeit und Praktikabilität miteinander verbinden. Bis vor ziemlich genau 100 Jahren war man dafür auf verschiedene mechanische Konstruktionen oder die Sonne angewiesen. Diese Uhren waren für viele Bereiche schon hinreichend und hatten eine Abweichung von wenigen Sekunden pro Monat. Allerdings waren sie immer noch sehr stark von äußeren Faktoren abhängig. Beispielsweise hatten Temperaturänderungen oder Alterungsprozesse einen starken Einfluss auf die Uhr. Zusätzlich war

die Reproduzierbarkeit nicht gewährleistet, wodurch zwei gleiche Uhren Abweichungen aufweisen konnten. Wenn man sich auf der ganzen Welt auf eine Definition der Sekunde einigen möchte, ist es sehr unbefriedigend, ein Einzelstück per Definition als Referenz für alle anderen Uhren zu bestimmen. Hier gab es eine der Sternstunden der Atomphysik: Der Übergang von verschiedenen Energiezuständen in einem Atom ist eine ausgezeichnete Referenzquelle, die sehr viel unabhängiger von externen Störeinflüssen ist als mechanische Uhren. Des Weiteren ist diese Referenzquelle auch sehr gut reproduzierbar, denn beispielsweise ein Cäsium-Atom hat überall auf der Welt die gleichen Eigenschaften. Wie man die eine Uhr auf dieser Grundlage eines Atomübergangs beschreiben kann, werden wir uns im Folgenden ansehen.

Was sind Atomzustände?

Atome können in verschiedenen Zuständen vorkommen und auch von einem Zustand in einen anderen übergehen. Meistens sind damit verschiedene Elektronenkonfigurationen gemeint. Sehr vereinfacht kann man sich einen Atomkern vorstellen, der von diversen Elektronen umkreist wird. Die Elektronen können ihn auf verschiedenen Bahnen umkreisen. Wenn sie ihre Bahn wechseln, geht dies mit einer Energieänderung einher. Beispielsweise kann dies durch Aufnahme oder Abgabe einer elektromagnetischen Welle geschehen. Man sagt, das Atom verändert seinen Zustand und somit sein Energieniveau. Dieses vereinfachte Bild, das Elekt-

ronen auf Bahnen beschreibt, lässt sich zwar quantenmechanisch nicht halten, vermittelt aber eine erste Intuition für den Sachverhalt (vgl. Abb. 1 → links). Im Hinblick auf Uhren ist es sehr spannend, spezielle Zustandsänderungen im Atom zu betrachten, die eine besonders kleine Energieunsicherheit haben. Wenn nun das Atom eine solche Zustandsänderung vollzieht, wird Licht einer genau dieser Energie entsprechenden Frequenz abgegeben oder aufgenommen.

Der zweite physikalische Aspekt, der zur Realisierung einer Atomuhr eine genauere Betrachtung verdient, ist der Übergangsprozess mittels einer elektromagnetischen Welle. Eine elektromagnetische Welle ist eine Schwingung des elektrischen und magnetischen Feldes. Das bekannteste Beispiel ist wahrscheinlich Licht. Je nach Frequenz bezeichnet man diese Welle auch als Röntgenstrahlung, Radio- oder Mikrowelle. Spannend im Hinblick auf den Uhrenbetrieb ist, dass die Energie dieser Schwingung direkt mit der Frequenz gekoppelt ist. Eine eindeutige Energie bedeutet gleichzeitig eine eindeutige Frequenz und damit eine Möglichkeit, diese als Taktgeber für eine Uhr zu verwenden. Die heutige Definition der Sekunde basiert auf einem Mikrowellenübergang im Cäsium-Atom. Die zugehörige Frequenz von 9.192.631.770 Schwingungen pro Sekunde konnte bereits in den 1950er Jahren, als diese Art von Uhren aufkam, mithilfe von schneller Elektronik in ein „Ticken“



Atomuhren sind für viele alltägliche Bereiche wie Stromnetze, Datenkommunikation und für die Grundlagenforschung wichtig.

JOHANNES TIEDAU





Die Kernuhr ist ein neues Konzept
für eine hochpräzise Uhr,
die auf einem Kernübergang
beruht.

JOHANNES TIEDAU



umgewandelt werden.^[3] In den vergangenen Jahren wurden optische Übergänge in Atomen deutlich wichtiger, weil diese eine noch genauere Uhr ermöglichen. Rein elektronisch lässt sich allerdings eine Frequenz im optischen Bereich von ca. 400 bis 750 THz nicht

mehr verarbeiten. Hier half die Entwicklung des Frequenzkammes (für den 2005 der Physik-Nobelpreis verliehen wurde), wodurch optische Frequenzen in langsamere elektrische Schwingungen im Gigahertz-Bereich übersetzt werden können. Diese Messmethode ist mittlerweile eine Standardtechnik der Atomphysik geworden.

Zusammenfassend kann man mithilfe einer elektromagnetischen Welle die unübertroffene Stabilität von atomaren Übergängen nutzen und daraus einen Sekundentakt für unsere Uhren schaffen.

Die eben beschriebenen Verfahren werden seit Jahren zur Messung der Zeit verwendet und legen mit dem Cäsium-Übergang auch die Definition einer Sekunde fest. Der neue Aspekt einer Kernuhr ist, dass ein Übergang im Atomkern selbst anstatt in der Elektronenhülle verwendet wird. Auch im Atomkern befinden sich verschiedene Energiezustände, und die Protonen und Neutronen, aus denen der Atomkern besteht, können in verschiedenen Konfigurationen vorkommen. Dieses Konzept, also den Übergang in einem Atomkern für die Zeitmessung zu verwenden, wurde erstmalig von Peik und

Tamm 2003 beschrieben.^[4] Der Atomkern bietet gegenüber bisherigen Übergängen in der Elektronenhülle viele Vorteile. Als Erstes koppelt der im Verhältnis zur Elektronenhülle sehr kleine Atomkern viel schwächer an externe Störeinflüsse. Bestehende Atomuhren, die auf Übergängen in der Elektronenhülle basieren, sind beispielsweise durch externe elektrische Störfelder limitiert. Intuitiv kann man dies erwarten, da das Störfeld die Elektronenhülle verzerrt und damit auch die Energieniveaus verschiebt. Im Atomkern ist die Sensitivität auf externe Felder sehr viel schwächer ausgeprägt. Dies geht so weit, dass man erstmalig die Atome mit einem interessanten Uhrenübergang in einen Festkörper einbetten kann. Durch Wechselwirkungen im Festkörper wären normale Elektronenübergänge vollständig verändert. Der Atomkern hingegen ist so unempfindlich auf Störeinflüsse, dass er in erster Näherung von diesen Effekten nicht betroffen ist. Dies führt zum zweiten großen Vorteil des Kernübergangs: Eine hochgenaue Uhr basierend auf einem Festkörper ist erstmalig möglich. In Festkörpern befindet sich eine unvorstellbar große Anzahl von Atomen (1 Gramm Eisen enthält beispielsweise 10^{22} Atome). In einer Festkörperuhr kann nur ein kleiner Prozentanteil der Atome für die Uhr genutzt werden. Dennoch enthalten diese vielen Größenordnungen mehr Atome, als bisherige Atom- oder Ionenfallen es können. Dies hat zur Folge, dass ein starkes Signal von den Atomkernen erzeugt werden kann. Bei herkömmlichen Atomuhren muss über eine längere Zeit gemittelt werden, um die statistischen Schwankungen aufgrund des niedrigen Signals auszugleichen. Eine

höhere Anzahl von verfügbaren Atomen führt dazu, dass man für die gleiche relative Unsicherheit weniger lange mitteln muss. Trotz dieses Vorteils des Festkörpers bleiben auch Kernübergänge in Ionenfallen (Ionen sind geladenen Atome) ein interessantes Forschungsfeld, da hier systematische Verschiebungen besonders klein und gut kontrollierbar sind. Der dritte Vorteil einer Kernuhr ist, dass diese bei der Suche nach bislang noch unbekannten Gesetzmäßigkeiten in der Physik helfen könnte. Viele Bereiche unserer heutigen Beobachtungen können Physiker sehr präzise beschreiben. Dennoch entziehen sich einige Phänomene gegenwärtig den theoretischen Erklärungen. Beispielsweise können die Rotationseigenschaften von Galaxien nicht gut beschrieben werden, wodurch eine noch unbekannte Masse postuliert wurde. Da diese Form von Materie nicht sichtbar ist, wird sie als Dunkle Materie bezeichnet. Mithilfe einer Kernuhr könnte es möglich sein, weitere Einsichten über Dunkle Materie zu bekommen. Der Vorteil einer sehr genauen Kernuhr wäre, dass diese neben der Coulombkraft, die zwischen geladenen Teilchen wirkt, auch auf die starke Kernkraft reagiert. Falls die Dunkle Materie also doch auf minimale Weise unsere „normale“ Welt jenseits der Gravitation verändert, könnte man dies möglicherweise im Vergleich von Atomuhren mit Kernuhren sehen.

Wenn der Atomkern eine solch gute Grundlage für eine Uhr bietet, warum basieren nicht alle Uhren auf Kernübergängen?

Die Besonderheit von Thorium-229

Das Hauptproblem der Kernübergänge liegt in ihren hohen Energien. Dies ist von der Gamma-Strahlung beim Kernzerfall bekannt. Durch die sehr hohe Energie dieser Strahlung kann sie nicht vollständig geblockt, sondern nur abgeschwächt werden, was ein Problem im Umgang mit radioaktiven Stoffen darstellt. Viele dieser Kernübergänge sind sehr schmalbandig, was bedeutet, dass sie Strahlung einer besonders gut definierten Wellenlänge ausstrahlen. Mithilfe von Laserlicht, das bei hohen Energien durch große Teilchenbeschleuniger erzeugt wird, ist es möglich, Atomkerne anzuregen.^[5] Beide Faktoren wären eigentlich eine sehr gute Grundlage für eine Zeitdefinition. Das große Problem ist allerdings, dass gegenwärtig keine Technologie bekannt ist, die diese hohen Frequenzen in den Bereich von zählbarem Ticken (1 pro Sekunde) zurückführen kann. Ein Frequenzkamm für Röntgenlicht ist gegenwärtig nicht realisierbar.

Die einzige Ausnahme, bei der die erste Anregungsenergie eines Atomkerns von einem kompakten Lasersystem erreicht werden kann, ist Thorium-229. Nur bei dieser Atomsorte und diesem besonderen Isotop – als Isotop bezeichnet man Atome, die sich nur durch eine Neutronenanzahl im Kern unterscheiden – heben sich ungewöhnlicherweise

die starke Kernkraft und die elektrische Coulombkraft so auf, dass ein angeregter Kernzustand mit einer sehr niedrigen Energie entsteht.^[6]

Thorium ist ähnlich wie Uran ein radioaktives Schwermetall. Das benötigte Thorium-Isotop 229 kommt auf der Erde aber nicht natürlich vor. Es besitzt eine Halbwertszeit von etwas weniger als 8.000 Jahren und geht durch einen Alpha-Zerfall (Abgabe von je zwei Protonen und Neutronen) in Radium-225 über. Diese Halbwertszeit ist allerdings für praktische Anwendungen ausreichend lang,

um den Uhrenbetrieb nicht zu behindern. Das Isomer wiederum besitzt eine Lebensdauer von einigen Minuten. Dies ist die Zeit, in der der Atomkern von einem angeregten Zustand in den Grundzustand zurückkehrt, ohne dass sich die Atomsorte selbst verändert.

Auch die Energieunschärfe des Übergangs, typischerweise als Linienbreite angegeben, ist mit etwas weniger als 1 mHz sehr gut für einen Uhrenbetrieb geeignet.

Der Weg zur Kernuhr

Die verschiedenen Übergänge (Gamma-Strahlung) in Atomkernen zu untersuchen, ist ein etabliertes Gebiet in der Physik und die zugrunde liegenden Prozesse lassen sich gut theoretisch beschreiben.^[7] Bei der Klassifikation dieser Spektren fiel allerdings in

den 1980er-Jahren auf, dass ein Bandenkopf, also ein Schwingungsgrundzustand, der zwingend vorhanden sein musste, bei Thorium-229 nicht beobachtet wurde. Die völlig korrekte Erklärung war: Dieser Bandenkopf befindet sich so nah an dem Grundzustand, dass dieser mit den typischen Methoden der Gamma-Spektroskopie nicht aufgelöst werden konnte.^[8] Dieser besonders niedrige Übergang wurde bald von anderen aufgegriffen^[9] und für einen Uhrenbetrieb vorgeschlagen.^[4] Immer bessere Messungen konnten den Energiebereich immer weiter eingrenzen. Allerdings gingen die frühen Veröffentlichungen von einer Energie von etwa 3,5 eV^[10] aus, was Licht im UV-Bereich entspricht. Jedoch zeigten spätere Messungen, dass dieser Wert deutlich zu niedrig angenommen worden war. Verschiedene Veröffentlichungen ab 2007 konnten den Wert allerdings korrekt bei einer Energie um 8 eV einordnen.^[11, 12, 13]

Diese deutlich höhere Energie ist indes so hoch, dass sie als elektromagnetische Welle in unserem Alltag nicht vorkommt, da sie vollständig durch den Sauerstoff der Luft absorbiert wird (dieser entspricht einer Frequenz um 2.000 THz und einer Wellenlänge von 150 nm). Daher muss der gesamte Versuchsaufbau im Vakuum gehalten werden, wodurch dieser Frequenzbereich auch als Vakuum-Ultraviolett (VUV) Bereich bekannt ist. Durch die Absorption in Luft wurden auch zahlreiche anfängliche Detektionsversuche verhindert, da noch die falsche Energie von 3,5 eV angenommen worden war.

„Mithilfe der Kernuhr
könnten sich neue physikalische
Gesetzmäßigkeiten
finden lassen.“

JOHANNES TIEDAU

Ein mit Thorium dotierter CaF_2 -Kristall auf einem Aluminiumhalter



Abbildung 2

Mit Thorium dotierter CaF_2 -Kristall der TU Wien, auf einem Aluminiumhalter montiert. Zu Anschauungszwecken ist der Kristall mit einem blauen Laserlicht beleuchtet. Der Kristall hat einen Durchmesser von 3 mm.

Quelle: Bild PTB

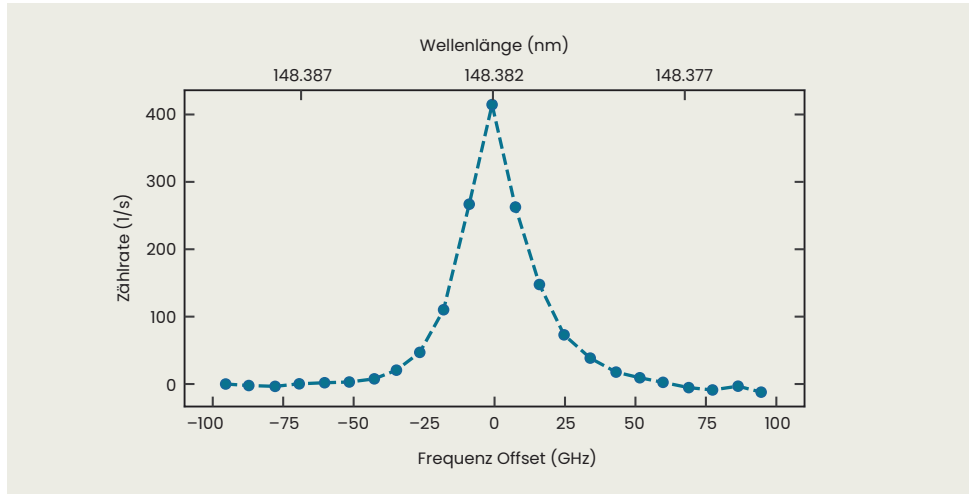
Ein wichtiger Schritt zur Laseranregung des Thorium-Übergangs war die erstmalige Vermessung der Fluoreszenz¹⁴ des Isomers. Das sehr kurzlebige Isotop Actinium-229 zerfällt mit einer sehr hohen Wahrscheinlichkeit in den angeregten Thorium-229-Zustand. Dieser angeregte Zustand wiederum verliert nach einiger Zeit seine zusätzliche Energie und geht in den Grundzustand über. Die Energiedifferenz wird durch Licht bei exakt der gesuchten Übergangsfrequenz emittiert. Dies ist zwar nicht hilfreich, um die Frequenz des Lichtes auf 18 Nachkommastellen zu bestimmen, aber die Energieunsicherheit konnte mithilfe eines Spektrometers stark eingeschränkt werden.

Bereits vor dieser eben beschriebenen Arbeit wurde an der PTB mit dem Aufbau eines Lasersystems für die Kernanregung begonnen. Wie bereits erwähnt, muss der gesamte Laser-Strahlengang im Vakuum gehalten werden, um eine Absorption des Lichts von Luftmolekülen zu verhindern. Weiterhin gibt es auch kein Lasermedium, das eine Verstärkung bei dieser Frequenz ermöglichen würde. Daher muss die Laserstrahlung durch Frequenzmischung von niedrigeren Frequenzen erzeugt werden. An der PTB wird dieses Licht in einem sogenannten Vierwellenmischprozess generiert, der drei eingestrahlte Lichtwellen in eine vierte konvertieren kann. Dieser Prozess ist lange bekannt und auch in verschiedenen Versionen

Abbildung 3

Eine der ersten Messungen des Thorium-Kernübergangs. Durch die Laseranregung sieht man einen Anstieg des emittierten Lichtes. Die Messung ist hintergrundkorrigiert.

Quelle: Grafik PTB, Daten aus [22]

Erste Messungen des Thorium-Kernübergangs

erprobt.^[15] Dennoch bietet diese Lichtquelle einige technische Herausforderungen, wie die Laserstrahljustage im Vakuumaufbau. Der Aufbau basiert auf zwei Farbstoffverstärkern, die sich so einstellen lassen, dass die VUV-Frequenz von 2.025 THz bis 1.934 THz abgestimmt werden kann.^[16] Glücklicherweise konnte dieser Bereich aber durch die Arbeiten am CERN bis auf 6 THz eingegrenzt und somit der für die Suche benötigte Zeitraum verringert werden.^[14]

Zusätzlich zur Laserquelle benötigt man schließlich noch Thoriumatome. Hierzu werden gegenwärtig zwei Ansätze in der Wissenschaft erprobt. Zum einen kann man Thoriumionen in einem elektrischen Wechselfeld fangen. Dies ist eine Standardmethode im Bereich der optischen Uhren¹⁷ und

eignet sich besonders gut dazu, die Thoriumatome von möglichen äußeren Störfaktoren abzuschirmen. Der zweite Ansatz ist, die Thoriumatome in einen Festkörper einzubetten – dies ist als Dotieren bekannt. Diese neue Methode ist für einen Uhrenübergang wie oben beschrieben erstmalig möglich, da die Atomkerne deutlich weniger auf äußere Störeinflüsse reagieren als Übergänge in der Elektronenhülle.

Gerade bei der initialen Suche mit einem Laser war ein hohes und eindeutiges Signal mit niedrigem Hintergrundrauschen wichtig. Hierbei boten die dotierten Festkörperkristalle einen klaren Vorteil, da durch das größere erwartete Signal der Energiebereich viel schneller abgesucht werden konnte. Für den Ausgangskristall, der zum Dotieren verwen-

Sensitivität des Thorium-Übergangs auf Dunkle Materie

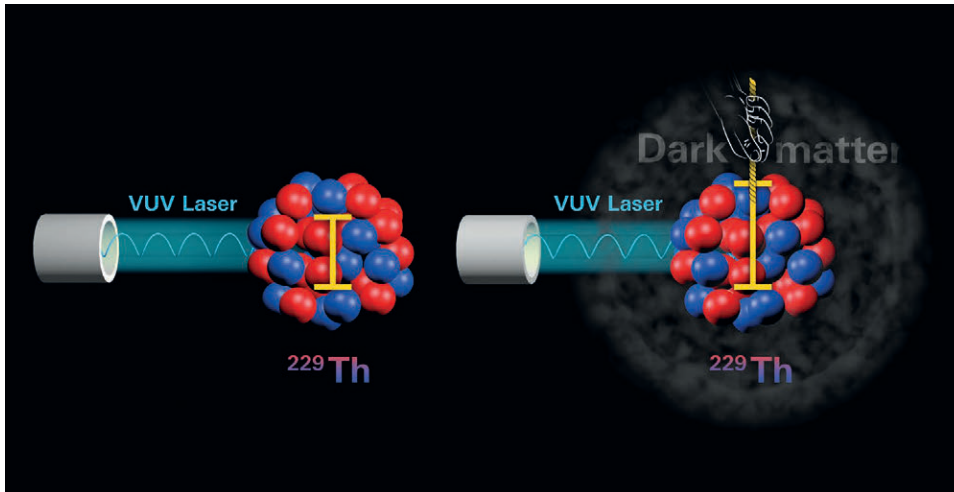


Abbildung 4

Der Thorium-Übergang des Atomkerns ist besonders empfindlich für noch unbekannte Physik. Beispielsweise könnte die Dunkle Materie die Energie des Kernübergangs verändern und somit messbar werden.

Quelle: Bild PTB

det wird, wurden bis 2023 zwei Materialien näher betrachtet: LiSrAlF_6 an der UCLA und CaF_2 an der TU Wien. Diese Kristalle müssen bereits über eine gute Transmission für das Licht im VUV-Bereich verfügen, wodurch nur wenige Materialien zur Auswahl stehen. In Zusammenarbeit mit der Gruppe von Thorsten Schumm von der TU Wien wurden ab 2021 mit Thorium-229 dotierte CaF_2 -Kristalle mit dem VUV-Licht an der PTB untersucht (Abb. 2 → S. 77). Diese Kristalle mit einer hohen VUV-Transmission herzustellen, erforderte eine jahrelange Optimierung, da die Transmission trotz der Dotierung des radioaktiven Thoriums erhalten bleiben muss.^[18, 19]

Zusätzlich können auch die hohen Spitzenintensitäten des gepulsten VUV-Lasers zu Defekten im CaF_2 führen. Daher wurden die

Kristalle auf etwa -120°C heruntergekühlt, wodurch der Kristall deutlich robuster gegenüber dem Laserlicht wird. Allerdings führten diese niedrigen Temperaturen auch dazu, dass dünne Lagen von Restgasen aus der Vakuumkammer am kalten Kristall kondensierten. Diese Schichten besitzen eine besonders schlechte VUV-Transmission, wodurch keine Laseranregung möglich war. Durch bessere Vakuumbedingungen in der Kammer um den Kristall und ein regelmäßiges Aufwärmen des Kristalls über Nacht konnten diese Effekte aber kontrolliert werden. Der durch das CERN-Experiment vorgegebene Bereich wurde abgescannt. Dazu wurde der Kristall für



Die weltweit erste Laseranregung von Th-229 erfolgte 2023 an der PTB in Zusammenarbeit mit der TU Wien.

JOHANNES TIEDAU



Seit der ersten Laseranregung
des Th-229-Kerns gibt
es viele Fortschritte
auf dem Weg zur Kernuhr.“
JOHANNES TIEDAU



120 s der VUV-Strahlung ausgesetzt. Danach wurde das Laserlicht ausgeschaltet, um ein niedrigeres Hintergrundrauschen zu erhalten. Mithilfe eines Einzelphotonendetektors wurde nach der charakteristischen Fluoreszenz des Kristalls gesucht. Diese bleibt aufgrund des schmalbandigen Übergangs für einige Minuten sichtbar, sodass die Detektion für 150 s gemittelt wurde. Falls kein Signal festgestellt wurde, wurde der Laser

etwas in der Frequenz weiter verschoben, wieder angeschaltet und damit der nächste Bereich abgesucht.

Ende 2023 fand sich der erste klare Hinweis auf einen Zerfall des Isomers in den Messdaten der PTB/TU Wien (Abb. 3 → S. 78). Bei dieser Art von Detektion ist besondere

Vorsicht geboten. In der Vergangenheit wurden bereits mehrere Veröffentlichungen publiziert, die den Übergang vermeintlich gefunden hatten.^[20, 21] Eines der zugrunde liegenden Probleme des Isomers im Festkörper ist, dass durch die radioaktiven Zerfälle im Kristall Lichtblitze emittiert werden, die fälschlicherweise als Isomersignal interpretiert werden können. Das detektierte Signal an der PTB zeigte allerdings in allen weiteren Tests die erwarteten Eigenschaften, wie die korrekte Abhängigkeit von Laser-Linienbreite und Intensität. Zusätzlich wurden auch Thorium-232-Kristalle untersucht, die durch den gleichen Prozess an der TU Wien hergestellt worden waren. Thorium-232 ist das langlebige, natürliche vorkommende Isotop, das nicht den bei Thorium-229 interessanten

Übergang besitzt. Bei diesen Kristallen zeigte sich kein Signal, welches die tatsächliche Laseranregung des Thorium-229 bestärkte.^[22]

Bereits kurz nach der Veröffentlichung dieser Daten konnten die Messungen in einem weiteren Material nachgewiesen werden, das auch eine hohe VUV-Transmission besitzt (LiSrAlF₆).^[23] Obwohl die direkte Umgebung der Thoriumatome im Kristallgitter völlig anders aussieht, stimmten die gemessenen Übergangsfrequenzen im Rahmen der verwendeten Laserlinienbreite überein. Bereits kurze Zeit später konnte der Isomerübergang in Th:CaF₂ mithilfe eines fs-Lasers um einige Größenordnungen genauer untersucht werden.²⁴ Es folgten Anregungen in weiteren thoriumdotierten Kristallen^{25, 26} und viele weitere Publikationen, die hier nicht alle diskutiert werden können.

Ein weiterer wichtiger Schritt zum tatsächlichen Uhrenbetrieb einer Kernuhr war die kürzliche Entwicklung der ersten Dauerstrichlasersysteme bei der Thoriumwellenlänge von 148 nm an der PTB in Zusammenarbeit mit dem Max-Born-Institut in Berlin²⁷ sowie eines weiteren Systems an der Tsinghua-Universität in China.²⁸ Die Ansätze hierbei waren völlig unterschiedlich. Während das PTB-Lasersystem vollständig auf Nichtlinearitäten in Festkörpern zur Wellenlängenerzeugung in VUV setzte, verwendete das chinesische System den Vierwellenmischprozess in einem heißen Cadmiumdampf als Konversionsmedium. Mit beiden Methoden sollten aber in Zukunft Anregungsexperimente mit Thorium möglich sein. Dabei bieten die Dauerstrichsysteme

den Vorteil, dass die Frequenzbreite schmaler ist und somit einen höheren Überlapp mit dem schmalbandigen Thorium-Übergang bietet. In den thoriumdotierten Kristallen kommt es daher zu keinen zusätzlichen Störfaktoren, die beispielsweise den Isomerezustand abregen können (dies wird auch als Quenching bezeichnet)^[29, 30] oder die Kristallstruktur verändern.

Parallel zu diesen Arbeiten werden an verschiedenen Orten weltweit gerade Experimente mit gespeicherten Thoriumionen durchgeführt oder aufgebaut. Durch die in einem hohen Maße bekannten Rahmenbedingungen in einem solchen Aufbau versprechen diese Messungen deutlich kleinere Frequenzunsicherheiten. Eine Forschungsrichtung, die dazu gerade untersucht wird, hat Ähnlichkeiten mit dem Actinium-Experiment am CERN. Hierbei wird eine sogenannte Rückstoßquelle verwendet, die Thorium-229 aus Uran-233 durch einen Alpha-Zerfall produziert. Da bei diesem Prozess 2 Prozent der Thoriumkerne im Isomerezustand produziert werden, kann dieser auch ohne Laseranregung untersucht werden.^[31, 32, 33, 34]

Zusammen mit Arbeiten am Grundzustand^[35] können damit Änderungen am Kern durch den Zustandswechsel sehr präzise gemessen werden. Beispielsweise ändert sich bei dem Kernübergang die Größe des Ladungsradius im Kern, was insofern erstaunlich ist, als für besagten Übergang ein Neutron (also ein ungeladenes Teilchen) verantwortlich ist.^[33]

Ausblick

Was könnte man also mit einer Kernuhr messen, wenn diese als fertiges Produkt verfügbar ist? Eine Motivation für die Kernuhr ist, dass diese möglicherweise kompakter gebaut werden kann. Gerade durch die Einbettung der Thoriumatome in einen Festkörper kann die Notwendigkeit einer Vakuumkammer vollständig entfallen. Diese Form der Kernuhr bietet also die Möglichkeit einer Miniaturisierung, auch wenn bis dahin einige technische Probleme, wie ein kompaktes Lasersystem, gelöst werden müssten.

Auch wenn die Festkörperuhren einige Vorteile bieten, werden die besten Kernuhren höchstwahrscheinlich auf Thoriumionen in Ionenfallen beruhen, da sich äußere Störeinflüsse hier besonders gut kontrollieren lassen. Bislang wurde eine Anregung von derartig isolierten Ionen noch nicht gezeigt, auch wenn es dazu weltweit einige Arbeiten und Fortschritte gibt. Gerade in Hinblick auf die Suche nach neuen, noch unbekannten physikalischen Gesetzmäßigkeiten könnten diese Uhren einen Durchbruch schaffen. (Abb. 4 → S. 79)

Die weltweiten Aktivitäten zur Erforschung der Thoriumuhr nehmen gerade in all diesen Bereichen zu und es bleibt spannend abzuwarten, wohin sich das Feld in den nächsten Jahren bewegt.

Referenzen

- [1] Brewer, Samuel M., et al. “Al⁺ 27 quantum-logic clock with a systematic uncertainty below 10⁻¹⁸.” *Physical review letters* 123.3 (2019): 033201.
- [2] Aeppli, Alexander, et al. “Clock with 8 × 10⁻¹⁹ systematic uncertainty.” *Physical Review Letters* 133.2 (2024): 023401.
- [3] Essen, Louis, and J. V. L. Parry. “The caesium resonator as a standard of frequency and time.” *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences* 250.973 (1957): 45–69.
- [4] Peik, E., and Tamm C. “Nuclear laser spectroscopy of the 3.5 eV transition in Th-229.” *EPL (Europhysics Letters)* 61.2 (2003): 181–186.
- [5] Shvyd'ko, Yuri, et al. “Resonant X-ray excitation of the nuclear clock isomer 45Sc.” *Nature* 622.7983 (2023): 471–475.
- [6] Fadeev, Pavel, Julian C. Berengut, and Victor V. Flambaum. “Sensitivity of Th 229 nuclear clock transition to variation of the fine-structure constant.” *Physical Review A* 102.5 (2020): 052833.
- [7] Bohr, Aage Niels, and Benjamin Roy Mottelson. “Collective and individual-particle aspects of nuclear structure.” *Dan. mat. fys. Medd.* 27.CERN-57-38 (1953): 1–174.
- [8] Kroger, L. A., and C. W. Reich. “Features of the low-energy level scheme of 229Th as observed in the α-decay of 233U.” *Nuclear Physics A* 259.1 (1976): 29–60.
- [9] Tkalya, E. V., et al. “Processes of the nuclear isomer 229mTh (3/2⁺, 3.5±1.0 eV) resonant excitation by optical photons.” *Physica Scripta* 53.3 (1996): 296–299.
- [10] Helmer, R. G., and C. W. Reich. “An excited state of Th 229 at 3.5 eV.” *Physical Review C* 49.4 (1994): 1845.
- [11] Beck, B. R., et al. “Energy splitting of the ground-state doublet in the nucleus Th 229.” *Physical review letters* 98.14 (2007): 142501.
- [12] Seiferle, Benedict, et al. “Energy of the 229Th nuclear clock transition.” *Nature* 573.7773 (2019): 243–246.
- [13] Sikorsky, Tomas, et al. “Measurement of the th 229 isomer energy with a magnetic microcalorimeter.” *Physical Review Letters* 125.14 (2020): 142503.
- [14] Kraemer, Sandro, et al. “Observation of the radiative decay of the 229Th nuclear clock isomer.” *Nature* 617.7962 (2023): 706–710.
- [15] Hanna, S. J., et al. “A new broadly tunable (7.4–10.2 eV) laser based VUV light source and its first application to aerosol mass spectrometry.” *International Journal of Mass Spectrometry* 279.2–3 (2009): 134–146.
- [16] Thielking, J., et al. “Vacuum-ultraviolet laser source for spectroscopy of trapped thorium ions.” *New journal of physics* 25.8 (2023): 083026.
- [17] Ludlow, Andrew D., et al. “Optical atomic clocks.” *Reviews of Modern Physics* 87.2 (2015): 637–701.

- [18] Beeks, Kjeld. The nuclear excitation of thorium-229 in the CaF_2 environment: development of a crystalline nuclear clock. Diss. Technische Universität Wien, 2022.
- [19] Beeks, Kjeld, et al. "Optical transmission enhancement of ionic crystals via superionic fluoride transfer: growing VUV-transparent radioactive crystals." *Physical Review B* 109.9 (2024): 094111.
- [20] Irwin, G. M., and Ki-Hyon Kim. "Observation of electromagnetic radiation from deexcitation of the 229 Th isomer." *Physical review letters* 79.6 (1997): 990.
- [21] Zhao, Xinxin, et al. "Observation of the Deexcitation of the Th 229 m Nuclear Isomer." *Physical review letters* 109.16 (2012): 160801.
- [22] Tiedau, J., et al. "Laser excitation of the Th-229 nucleus." *Physical Review Letters* 132.18 (2024): 182501.
- [23] Elwell, R., et al. "Laser excitation of the Th 229 nuclear isomeric transition in a solid-state host." *Physical Review Letters* 133.1 (2024): 013201.
- [24] Zhang, Chuankun, et al. "Frequency ratio of the 229mTh nuclear isomeric transition and the 87Sr atomic clock." *Nature* 633.8028 (2024): 63–70.
- [25] Zhang, Chuankun, et al. "229ThF4 thin films for solid-state nuclear clocks." *Nature* 636.8043 (2024): 603–608.
- [26] Elwell, Ricky, et al. "Laser-based conversion electron Mössbauer spectroscopy of 229ThO2." *Nature* 648.8093 (2025): 300–305.
- [27] Lal, Vishal, et al. "Continuous-wave laser source at the 148 nm nuclear transition of Th-229." *Optica* 12.12 (2025): 1971–1974.
- [28] Xiao, Qi, et al. "Continuous-wave narrow-linewidth vacuum ultraviolet laser source." *Nature* (2026): 1–5.
- [29] Schaden, Fabian, et al. "Laser-induced quenching of the Th-229 nuclear clock isomer in calcium fluoride." *Physical Review Research* 7.2 (2025): L022036.
- [30] Terhune, James Emory Squiers, et al. "Photoinduced quenching of the Th 229 isomer in a solid-state host." *Physical Review Research* 7.2 (2025): L022062.
- [31] Von Der Wense, Lars, et al. "Direct detection of the 229Th nuclear clock transition." *Nature* 533.7601 (2016): 47–51.
- [32] Thielking, Johannes, et al. "Laser spectroscopic characterization of the nuclear-clock isomer 229mTh." *Nature* 556.7701 (2018): 321–325.
- [33] Yamaguchi, Atsushi, et al. "Laser spectroscopy of triply charged 229Th isomer for a nuclear clock." *Nature* 629.8010 (2024): 62–66.
- [34] Zitzer, G., et al. "Laser spectroscopy on the hyperfine structure and isotope shift of sympathetically cooled Th 3+ 229 ions." *Physical Review A* 111.5 (2025): L050802.
- [35] Campbell, C. J., A. G. Radnaev, and A. Kuzmich. "Wigner crystals of Th 229 for optical excitation of the nuclear isomer." *Physical review letters* 106.22 (2011): 223001.

Das Netzwerk der Werner-von-Siemens-Fellows

Auszeichnung der Werner-von-Siemens-Fellows 2025 bei TRUMPF in Ditzingen

Am 12. Dezember haben wir die Werner-von-Siemens-Fellows 2025 ausgezeichnet – traditionell vor Ort bei einem Preisträger des Werner-von-Siemens-Rings des Vorjahres, dieses Mal TRUMPF. Die Werner-von-Siemens-Fellows wurden gewürdigt unter anderem von der Ministerin für Wissenschaft, Forschung und Kunst des Landes Baden-Württemberg Petra Olschowski, dem Rektor der Universität Stuttgart Prof. Dr.-Ing. Peter Middendorf sowie Dr. Nicola Leibinger-Kammüller, Vorstandsvorsitzende bei TRUMPF. Prof. Dr. Katharina Hölzle vom Fraunhofer IAO führte als Moderatorin durch den Abend (im Bild S. 85 links).



von links nach rechts:
Elena Jordan,
Franziska Boenisch,
Taha Soliman,
Johannes Tiedau und
Martha Kalina



Fotos: Verena Mueller und Astis Krause

Forschung der Werner-von-Siemens-Fellows vor Ort: Magdeburg 2026

Beim Netzwerktreffen an der Otto-von-Guericke-Universität Magdeburg lud Prof. Dr. Dieter Schinzer, Werner-von-Siemens-Fellow seit 1986 (mittig im Bild rechts unten, damals ausgezeichnet als „Jungwissenschaftler“), ein, interdisziplinäre Forschungsprojekte kennenzulernen – unter an-

derem Lösungsansätze zur Bekämpfung von lebensbedrohlichen Infektionen. Außerdem durften wir einen Blick hinter die Kulissen des Max Planck Institute for Dynamics of Complex Technical Systems werfen und lernten bei einer Führung den Dom und das historische Magdeburg kennen.



Fotos: Max Gierth für die Otto-von-Guericke-Universität Magdeburg

Ein herzlicher Dank gilt allen, die ihre Türen für uns geöffnet haben und das Netzwerk mit Leben füllen. Wir freuen uns auf viele weitere gemein-

same Erlebnisse im Umfeld von exzellenter Forschung, Technik und Innovation!

Die Werner-von-Siemens-Fellows eröffnen neue Horizonte in Technik und Naturwissenschaften

Individualisierte Privatsphäre im Maschinellen Lernen, Co-Design von Hard- und Software für alltagstaugliche KI, die genauesten Messungen mit Atom-Uhren, mehr Sicherheit über die Prognose von Ermüdungsbrüchen und Ionenfallen in Nanophotonik: In fünf Beiträgen gewähren die Werner-von-Siemens-Fellows 2025 Einblick in ihre zukunftsweisenden Forschungsgebiete.

The Werner von Siemens Fellows open up new horizons in technology and natural sciences

Individualized privacy in machine learning, co-design of hardware and software for everyday AI, the most accurate measurements with atomic clocks, greater certainty in predicting fatigue fractures, ion traps in nanophotonics: In five articles, the Werner von Siemens Fellows 2025 provide insights into their research areas, thus shaping emerging technologies.