Energieeffizientes KI-System

AI goes Ultra-Low-Power – Teil 1

(Bild: PopTika | Shutterstock)

Einer der Sieger im Wettbewerb um die Entwicklung eines energieeffizienten KI-Systems für die EKG-Analyse konnte die mittlere Leistung auf lediglich 1 µW senken. Im 1. Teil wird das Gesamtkonzept eines Ultra-Low-Power-Beschleunigers für die EKG- oder eine allgemeine Zeitreihen-Analyse vorgestellt.

Von Dr. Marco Breiling, Dr. Peter Reichel und Dr. Marc Reichenbach

»Kann Deutschland noch Sprunginnovationen?« – diese Frage stellte sich die Bundesregierung und schob gleich hinterher: »Und wie können wir diese fördern?« Von diesen Überlegungen ausgehend entwickelte das Bundesministerium für Bildung und Forschung (BMBF) das Konzept der neuen Bundesagentur für Sprunginnovationen (SPRIN-D) und – um Erfahrungen dafür zu sammeln – dreier Pilot-Innovationswettbewerbe. Einer davon stellte den Teilnehmern die Aufgabe, ein möglichst »energieeffizientes KI-System« als Hardware auf einem ASIC bzw. FPGA zu entwickeln. Mit diesem sollte mit minimaler Energie ein Stapel von Hunderten jeweils zwei Minuten langen EKG-Signalen durch einen Machine-Learning- (ML) Algorithmus analysiert werden: Ist der Patient gesund oder zeigt die Aufnahme Vorhofflimmern? Es ist nämlich bekannt, dass Vorhofflimmern eine häufige Ursache von Schlaganfällen ist – eine energieeffiziente und kostengünstige Analyse könnte daher vielfach Schlaganfälle verhindern.

27 Teams bewarben sich um eine Teilnahme am Wettbewerb, elf Teams wurden dafür ausgewählt, vier Teams erreichten einen 1. Platz. Das Team Lo3-ML (Low-Power Low-Memory Low-Cost ECG Signal Analysis using Machine Learning Algorithms) des Fraunhofer-Instituts für Integrierte Schaltungen IIS und der Friedrich-Alexander-Universität Erlangen ist einer der Sieger des Pilot-Innovationswettbewerbs „Energieeffizientes KI-System“ (**Bild 1**). Ihm gelang es, die für die EKG-Analyse benötigte durchschnittliche Leistungs-

Bild 1. Der Siegerpokal für das Team Lo3-ML (Low-Power Low-Memory Low-Cost ECG Signal Analysis using Machine Learning Algorithms) im Pilot-Innovationswettbewerb der Bundesagentur für Sprunginnovationen (SPRIN-D). (Bild: Fraunhofer IIS)

aufnahme auf lediglich 1 µW senken – das ist so wenig Leistung, wie eine 4 cm × 4 cm große Solarzelle bei Mondschein liefert.

Erste Überlegungen

Generell kann eine hohe Energieeffizienz durch eine starke Spezialisierung auf eine einzige Anwendung erreicht werden – hier die EKG-Analyse. Dies geht allerdings generell zulasten der Flexibilität, um das System auch für andere ML-Anwendungen einsetzen zu können oder zumindest Änderungen wie ein Update der Parameter zu ermöglichen. Neben der geringeren Energieeffizienz ist eine programmierbare, universelle Schaltung in der Regel auch weniger leistungsfähig. Aus diesem Grund wurde im Lo3-ML-Projekt eine Schaltung für das größere Anwendungsfeld „Zeitreihensignale“ erstellt, die einen Kompromiss zwischen beiden Randfällen darstellt.

Neben der EKG-Analyse erlaubt der umgesetzte Ansatz die Analyse weiterer Zeitreihensignale, z.B. aus den Bereichen Predictive Maintenance, Audio/Hörgeräte usw., für die Neuronale Netze (NN) ähnliche Strukturen haben – mehrere 1D-Convolutional Layer gefolgt von Fully Connected (FC) Layern für die finale Klassifikation. Damit darüber hinaus weitere Anwendungen vom entwickelten Chip profitieren können, hat das Team in diesem Wettbewerb großes Gewicht auf einen möglichst effizienten und teilautomatisierten Entwurfsablauf gelegt. Dieser wird es für zukünftige Anwendungen ermöglichen, mit relativ geringen Fixkosten (Non-recurring Engineering Costs) höchst energieeffiziente, anwendungsspezifische Hardwarebeschleuniger zu entwickeln.

Keine energiehungrige Steuerlogik

Der entwickelte Deep-Learning-Beschleuniger basiert auf einer digitalen Schaltung. Hier tragen vor allem drei Aspekte erheblich zum Energiebedarf bei: Die für den Datenpfad benötigte Steuerlogik, der Zugriff auf externe Speicher – z.B. Flash und DRAM – und die Berechnungen. Daher setzt die Optimierung genau an diesen Punkten an.

In frei programmierbaren Prozessoren braucht die Steuerlogik – Instruktionsdecoder, Sprungvorhersage etc. – erheblich Energie. Gleiches gilt für die verwendeten festen Wortbreiten, z.B. 8 bit oder 32 bit. Beides ist bei anwendungsspezifischen HW-Beschleunigern unnötig. Auch bei stark programmierbaren Arrays von einfachen Processing Elements wie der Eyeriss-V2-Architektur [1] beansprucht die Steuerlogik einen Großteil der insgesamt benötigten Energie. Aus diesem Grund entschied sich das Lo3-ML-Team für einen datenflussorientierten Ansatz, bei dem die Steuerlogik implizit durch die Struktur der Rechenschaltung gegeben ist.

Wenn ein Schaltungsteil nicht benötigt wird: Abschalten

In aktuellen Beschleunigerschaltungen werden meist flüchtige Speicher eingesetzt. Dadurch kann der Chip nicht „stromlos“ geschaltet werden – um Energie zu sparen, wenn er nicht benötigt wird. In der Initialisierungsphase und bei jedem Wiedereinschalten wird viel Energie benötigt, um viele Daten – insbesondere die Gewichte des NNs – wieder aus dem externen Speicher zu laden. Deswegen setzte das Lo3-ML-Team nichtvolatile Speicher, konkret resistive RAMs (RRAMs), zur Speicherung dieser Daten ein. Solche Speicherelemente behalten ihren Inhalt auch nach Abschaltung der Energieversorgung [2]. Das bedeutet, dass der Beschleuniger „schlafen gelegt“ werden kann, wenn nicht genügend Arbeitslast anfällt, ohne beim „Aufwecken“ die energieintensive Initialphase erneut durchlaufen zu müssen.

Dieser Vorteil ist besonders groß, wenn Sensordaten nur relativ langsam in den Beschleuniger fließen – was bei EKG-Signalen im praktischen Einsatz am Patienten und allgemein bei fast

allen Zeitreihensignalen gegeben ist: Ist der Beschleuniger z.B. in ein Wearable integriert, um den Patienten laufend zu überwachen, so ist es sinnvoll, das mit niedriger Frequenz abgetastete EKG-Signal über mehrere Sekunden zu sammeln, während der Beschleuniger schläft, danach den Beschleuniger kurz zu aktivieren und die neuen Daten schnell abzuarbeiten, um anschließend den Beschleuniger wieder abzuschalten.

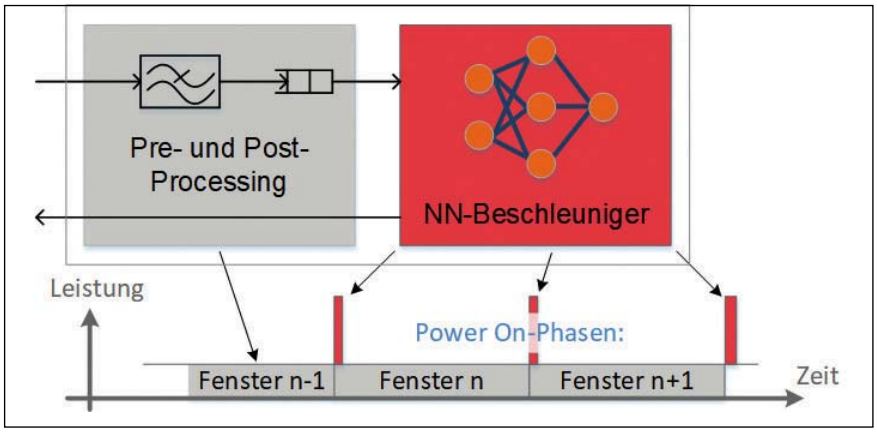


Bild 2. Aufteilung des Beschleunigerchips auf zwei Blöcke, einen permanent aktiven Bereich (links) und einen abschaltbaren NN-Beschleuniger (rechts). (Bild: Fraunhofer IIS)

Aller guten Dinge sind drei

Der dritte und letzte Optimierungsansatz ist die Energieeinsparung bei der Berechnung selbst: Die Grundidee ist hierbei die Einsparung von Energie durch eine bewusste Verringerung der Rechengenauigkeit. Anders als allgemeine Prozessoren, die mit einer festen Wortbreite – z. B. 32 bit Fließkomma – arbeiten, können in einer anwendungsspezifischen Schaltung Wortbreiten an den Algorithmus angepasst werden. Das Lo3-ML-Team entschied sich aus zwei Gründen, ternäre Gewichte zu verwenden, d.h. sie auf die drei Werte (–1, 0, +1) zu beschränken [3].

➔ Erster Grund: Man kann damit innerhalb einer Schicht des NNs Differenzen, Summen und Negationen beliebiger Eingangsaktivierungen berechnen. Über mehrere Filter und Schichten hinweg kann somit auch eine Gewichtung mit anderen Faktoren realisiert werden, womit das NN Aktivierungen betonen und andere dämpfen kann. Mit binären Gewichten wäre das dagegen so nicht möglich und quaternäre Gewichte sind dafür nicht unbedingt nötig. Zudem sind Multiplikationen mit –1 und +1 faktisch Additionen und somit sehr energieeffizient realisierbar.

➔ Zweiter Grund: RRAMs sind in der Lage, mehr als zwei Zustände zu speichern, sodass jedes ternäre Gewicht komplett in einer RRAM-Zelle abgelegt werden kann.

Zweiteilung spart Energie

Wie **Bild 2** zeigt, besteht das gesamte KI-System zur EKG-Analyse aus einer Hauptschaltung, die u.a. ein Preprocessing der Eingangsdaten enthält, und einem NN-Beschleuniger samt RRAM-Speicherblöcken, der dank einer eigenen Stromversorgung abgeschaltet werden kann. Preprocessing und Beschleuniger sind über einen doppelten Datenpuffer gekoppelt.

Zur Analyse von EKG-Daten ist die Betrachtung von Signalen ausreichender Länge erforderlich, weshalb jede zweiminütige EKG-Aufnahme in neun Fenster zu je 12,7 s unterteilt wird. Zunächst reduziert eine Bandpass-Filterung mit Down-Sampling die Datenrate im Preprocessing. Die neuronale Verarbeitung kann entsprechend dieser Datenratenreduktion deutlich langsamer erfolgen als die Eingabe. Dies wird zum Energiesparen genutzt, indem der NN-Beschleuniger zeitweise schlafengelegt wird.

Entwurfsablauf Ein wichtiger Aspekt des Projekts war der in **Bild 3** dargestellte, teilautomatisierte Entwurfsablauf. Dieser erlaubt eine frühzeitige Evaluierung des Energiebedarfs und die Berücksichtigung von HW-Eigenschaften wie z.B. der sehr kleinen Wortbreiten. Bereits bei der Spezifikation der NN-Hyperparameter – u. a. Schichttypen und -größen – wird ein grober Kompromiss der benötigten Energie anhand von simulierten Energiebedarfen für die benötigten Additions- und Multiplikationsoperationen für verschiedene Quantisierungsoptionen durchgeführt. Weiterhin wird die sehr starke Quantisierung von Gewichten (ternär) und Aktivierungen bereits im Trainingsalgorithmus berücksichtigt, wofür spezielle Bibliotheken zur Anbindung an TensorFlow geschaffen wurden.

Das gewählte und optimierte NN wird im ONNX-Format ausgegeben. Zum einen wird daraus automatisch – auf der Basis vorher entwickelter HW-Komponenten – die RTL-Schaltung in VHDL erstellt. Zum anderen wird die ONNX-Datei auch von einer Bit-True-Simulation in Python mit quantisierten Datentypen gelesen, die eine sehr schnelle und zu hundert Prozent genaue Simulation der Schaltung erlaubt, z.B. für automatisierte Rastersuchen (Grid

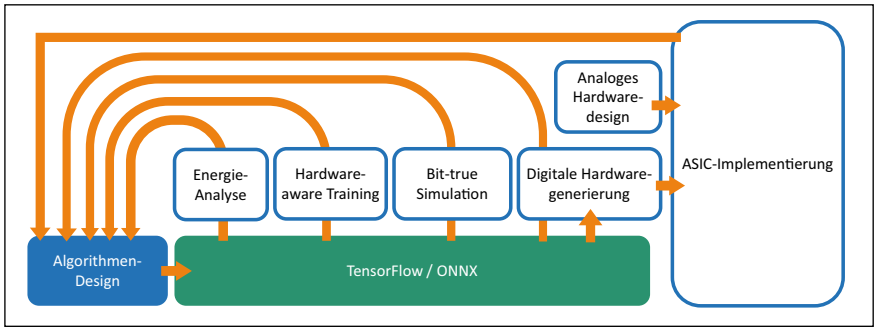


Bild 3. Der Entwurfsablauf für das neuronale Netzwerk und die digitale Schaltung besteht aus mehreren Iterationsschleifen. (Bild: Fraunhofer IIS)

Searches). Sämtliche Quantisierungseffekte können damit frühzeitig und vollständig erfasst werden und in die NN-Spezifikation zurückfließen. Zusätzlich dient die Bit-True-Simulation auch zur Verifikation der implementierten Schaltung.

Mit diesem Ansatz konnte sogar auf einem kostengünstigen 130-nm-Halbleiter-Prozess ein Beschleuniger mit 578 GOPS/W und 42 MOPS/mm² realisiert werden.

Wie die verschiedenen Aspekte, Entwurfsschritte und Aufgaben im Entwurfsablauf (siehe Bild 3) ineinandergreifen, ist in **Bild 4** dargestellt. Genau deshalb wurde ein hochgradig iterativer Prozess gewählt. So wurden zur Vorbereitung des Entwicklungsprozesses erste NNs und ASICs als „Pipe-Cleaner" erstellt, diese genau auf ihre Schwachpunkte u.a. hinsichtlich ihrer Energieaufnahme untersucht und damit gezielt Verbesserungen eingebracht.

Einsatzmöglichkeiten für energieeffiziente KI-Chips

Der beschriebene KI-Beschleuniger wurde besonders für Zeitreihensignale und den Einsatz in batteriebetriebenen Geräten entwickelt. Ein Beispiel dafür sind Wearables zur Gesundheitsüberwachung wie das Fitness-Shirt [4]. Dabei erlaubt die Aufteilung des Beschleunigers auf zwei Blöcke neben dem gezielten Abschalten des NN-Beschleunigers auch eine getrennte Wahl und Optimierung der Taktfrequenzen von beiden Blöcken für maximale Energieeffizienz abhängig von der jeweiligen Eingangsabtastrate und der tolerierbaren Latenz der Anwendung. Auch bei einer Verarbeitung vieler Sensorsignale auf einem Server kann eine solche sensornahe NN-Berechnung als Vorverarbeitung genutzt werden, um die zum Server zu übertragende Datenmenge massiv zu senken.

Nachdem im 1. Teil die grundsätzlichen Überlegungen und Entscheidungen dargestellt wurden, geht es in einer folgenden *Elektronik*-Ausgabe im 2. Teil in die Details der Umsetzung. hs

Literatur

[1] *Chen, Y.; Yang, T.; Emer J. und Sze, V.:* Eyeriss v2: A Flexible Accelerator for Emerging Deep Neural Networks on Mobile Devices. IEEE Journal on Emerging and Selected Topics in Circuits and Systems, 2019, Nr. 2, S. 292–308, DOI: 10.1109/JETCAS.2019.2910232.

[2] *Su, F.; Ma, K.; Li, X.; Wu, T.; Liu, Y. und Narayanan, V.:* Nonvolatile processors: Why is it trending? Design, Automation & Test in Europe Conference & Exhibition (DATE), 2017, S. 966–971, DOI: 10.23919/DATE.2017.7927131.

[3] *Li, F.; Zhang, B. und Liu, B.:* Ternary weight networks. arXiv, 1605.04711, 16. Mai 2016, https://arxiv.org/abs/1605.04711v1.

[4] CardioTEXTIL – Textiles Mehrkanal-EKG für den mobile Einsatz. Fraunhofer-Institut für Integrierte Schaltungen IIS, Website, www.iis.fraunhofer.de/de/ff/sse/health/medical-sensors-and-analytics/prod/cardiotextil.html.

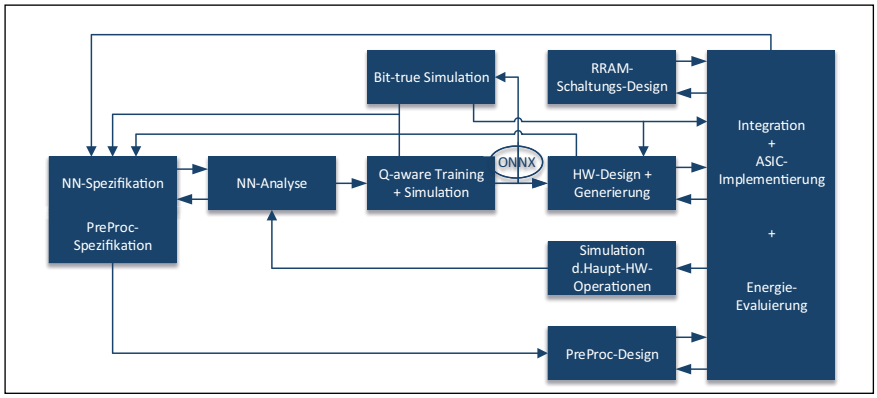


Bild 4. Im Projekt Lo3-ML wurde der Entwurfsablauf (vergleiche Bild 3) auf mehrere Schritte und Aufgaben aufgeteilt. (Bild: Fraunhofer IIS)

Dr. Marco Breiling

studierte Elektrotechnik in Karlsruhe, Trondheim, Paris und Southampton und promovierte in Erlangen. Er ist seit 2001 am Fraunhofer-Institut für Integrierte Schaltungen IIS in Erlangen tätig. Dort arbeitet er als Chief Scientist an den neuromorphen Hardwareentwicklungen.
marco.breiling@iis.fraunhofer.de

Dr. Peter Reichel

studierte in Dresden Informationssystemtechnik und promovierte in technischer Informatik. Er ist seit 2011 am Fraunhofer-Institut für Integrierte Schaltungen IIS, Institutsteil Entwicklung Adaptiver Systeme EAS in Dresden, tätig und arbeitet als wissenschaftlicher Mitarbeiter im Bereich intelligenter Sensorik.
peter.reichel@eas.iis.fraunhofer.de

Dr. Marc Reichenbach

studierte Informatik in Jena und promovierte in Erlangen. Seit 2010 ist er an der Friedrich-Alexander-Universität (FAU) am Lehrstuhl Rechnerarchitektur tätig. Als Post-Doktorand forscht er dort an neuen energieeffizienten Rechnerarchitekturen u.a. für Anwendungen aus dem Bereich der künstlichen Intelligenz.
marc.reichenbach@fau.de

Energieeffizientes KI-System

AI goes Ultra-Low-Power – Teil 2

(Bild: PopTika | Shutterstock)

Nachdem im 1. Teil [5] das Gesamtkonzept eines Ultra-Low-Power-Beschleunigers für die EKG- oder allgemeine Zeitreihenanalyse vorgestellt wurde, werden nun einzelne Aspekte des Chipentwurfs genauer beleuchtet.

Von Dr. Marco Breiling, Dr. Peter Reichel und Dr. Marc Reichenbach

Das Preprocessing (vergleiche Bild 2 im 1. Teil dieses Aufsatzes [5]) besteht hauptsächlich aus einem Bandpassfilter mit anschließendem Downsampling. Obwohl dieser Teil der Signalverarbeitung simpel im Vergleich zum Neuronalen Netzwerk (NN) erscheint, musste dieser Schaltungsteil dennoch stark optimiert werden; diese Teilschaltung läuft nämlich mit einer höheren Abtastfrequenz als der Rest und verwendet die große Wortbreite der digitalisierten Abtastwerte. Deshalb verwendete das Lo3-ML-Team an dieser Stelle verschachtelte, multiplikationsfreie CIC-Filter (Cascaded Integrator Comb, kaskadierte Integrator-Differentiator-Filter).

Neuronales Netzwerk mit speziellen Schichten

Der verwendete ML-Algorithmus hat einen maßgeblichen Einfluss auf die Energieaufnahme. Neben der Anzahl von Operationen bestimmt er auch die geforderte Quantisierung. Für maximale Energieeffizienz bei vorgegebener Klassifikationsgenauigkeit kombinierte das Team mehrere Methoden aus der aktuellen Forschung, um Aktivierungen und Gewichte möglichst grob quantisieren zu können. Dafür wurden spezielle Schichttypen eingeführt: Convolutional »Lo3Conv«- und Fully-Connected »Lo3FC«-Layer mit dem in **Bild 5** gezeigten Aufbau. Zuerst erfolgt eine quantisierte Faltung bzw. FC-Berechnung (QConv/QFC). Ihr folgt eine Binary-Shift Batch Normalization (BSBN); dies ist eine Batch Normalization (BN) y=Sx+B, bei der der Bias B ein vorzeichenbehafteter 6-bit-Festkommawert ist und die Skalierung S eine Zweierpotenz im Bereich (inklusive) 1/64 ... 4, d. h. mit neun möglichen Werten. Dadurch muss die Schaltung nur sehr energieeffiziente, binäre Shift-Operationen ohne »echte« Multiplikationen durchführen. Da Bias und Skalierung konfigurierbar sind und vom Training bestimmt werden, bietet eine BSBN die nötige Flexibilität für zukünftige Anwendungen und sichert weiterhin eine hohe Energieeffizienz,

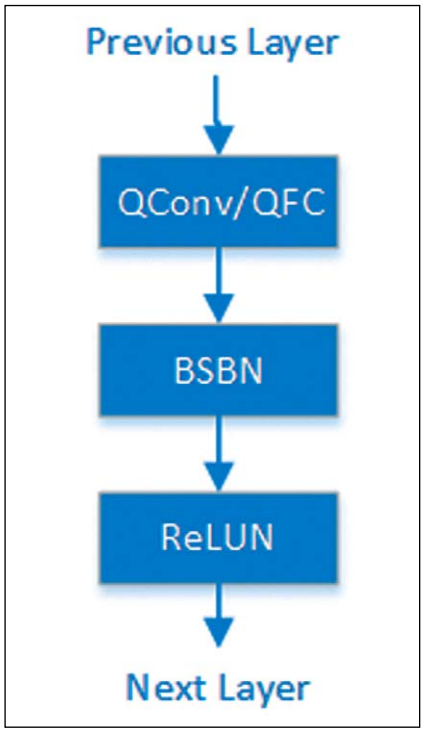


Bild 5. Um für eine hohe Energieeffizienz die Aktivierungen und Gewichte möglichst grob quantisieren zu können, wurden spezielle Layertypen eingeführt: Convolutional »Lo3Conv«- und Fully-Connected »Lo3FC«-Layer. (Bild: Fraunhofer IIS)

verglichen zur Standard-BN. Die letzte Operation in Lo3Conv-/Lo3FC-Layern ist ReLUN – eine Rectified Linear Unit (ReLU) mit Saturation im Wert N: ReLU(x,N)=min(max(0,x),N).

Wie **Bild 6** zeigt, besteht das NN aus vier Lo3Conv-Layern, einem Max-Pooling- und schließlich zwei Lo3FC-Layern. Das ganze NN ist völlig frei von »echten« Multiplikationen, weil ausschließlich ternäre Gewichte und die beschriebene BSBN eingesetzt werden.

Wegen der Entscheidung für eine ternäre, d.h. äußerst grobe Quantisierung der Gewichte, können diese keine größeren Wertebereiche darstellen, was normalerweise das NN stark degradiert. Die BSBN wurde zur Lösung dieses Problems eingeführt, damit jeder Ausgangskanal jeder Schicht individuell skaliert werden kann. Zur Bestimmung der Skalierungsfaktoren gibt es mehrere Möglichkeiten [6], um die Parameter B und S in der Fehlerrückführung (Backpropagation) zu optimieren. Obwohl die BSBN nur energiearme Shift-Operationen nutzt und ihr nur 5 % der Parameter und 3 % der Operationen angehören, braucht sie relativ viel Speicher und Energie bei der Berechnung, da hier im Gegensatz zu den Faltungs- und FC-Schichten nicht nur ternäre Parameter verwendet werden.

Der beschriebene Entwurfsablauf bezieht die Quantisierung aller Operationen und der für sie nötigen Energie bereits während der NN-Spezifikation ein. Für jeden betrachteten Kandidaten für das NN wird eine schnelle und grobe Energieabschätzung durchgeführt, indem einfach die für die quantisierten Operationen benötigte Energie mit der Zahl dieser Operationen im Kandidaten-NN multipliziert wird. Dies erlaubt sehr kurze Durchlaufzeiten im Entwicklungsprozess und entsprechend schnelle Zyklen während der NN-Entwurfsraum-Exploration.

Eine solch grobe (ternäre) Quantisierung erfordert ein Quantization-Aware-Training. Dafür hat das Lo3-ML-Team den Trainingsalgorithmus speziell angepasst, indem es Standard-TensorFlow (TF) um die eigenen Elemente QConv, QFC Layer und BSBN erweitert hat. Ohne diese Herangehensweise, nur mit Post-Training-Quantisierung, wäre die Genauigkeit des NNs extrem verschlechtert worden.

Durch die skizzierte Vorgehensweise wird zwar die Quantisierung bereits während des Trainings berücksichtigt, aber die Simulation des NNs in TensorFlow ist nicht zu 100 % identisch zur (quantisierten) Schaltung, weil alle TF-internen Berechnungen – Multiplikationen und Teilsummen – mit TF-Kernels erfolgen, die float32 und float64-Werte nutzen. Daher wurde die Inferenz als spezielle BT-Simulation (Bit True) in Python implementiert. Alle BT-Operationen entsprechen in ihrem Verhalten zu 100 % den Hardwareoperationen mit Festkommawerten, inkl. wahlweise Saturation oder Wrap-Around bei Overflows. Dies erlaubt:

➔ 1. die Verifikation der HW-Implementierung, da in jedem Punkt die Zwischenergebnisse verglichen werden können und

➔ 2. eine sehr frühzeitige und sehr schnelle Messung der exakten finalen Genauigkeit, die auch der Hardwarebeschleuniger so ausgeben würde. Auch dies führt zu kurzen Durchlaufzeiten und entsprechend schnellen Zyklen in der Entwurfsraum-Exploration des quantisierten NNs und erlaubt eine automatische Rastersuche über die NN-Hyperparameter.

Energieaufnahme der Schaltung bewerten und optimieren

Die Architektur des Rechenkerns für das NN, die im Zuge dieses Projekts entwickelt wurde, soll eine hinreichende Flexibilität aufweisen, damit dasselbe Konzept für unterschiedliche Applikationen angewendet werden kann. Entsprechend lässt sich dieser Beschleuniger auch für verschiedene ähnliche Problemstellungen unter Einsatz geänderter Parameter und Gewichte nutzen. Gleichzeitig soll die Umsetzung nur minimale Energie für die eigentliche Berechnung benötigen. Zudem wird die korrespondierende Schaltung besonders auf den Einsatz von ternären nicht-volatilen RRAM-Zellen hin optimiert. Ausgehend von der Grundidee datenflussgetriebener Parallelrechner wurde ein neuartiges, hochoptimiertes Schaltungskonzept entwickelt, welches besonders den Aufwand in der Steuerlogik reduziert sowie unnötige Operationen vermeidet.

Für eine möglichst geringe dynamische Verlustleistung der arithmetischen Operationen fließen bei der Entwicklung des NNs verschiedene Optimierungen ein, z.B. die Quantisierung aller Operanden auf eine geringe Bitbreite oder die Optimierung der Filterlängen zur Reduktion der notwendigen MAC-Operationen (Multiply Accumulate). Um diese Optimierungen durchführen zu können, wurde ein analytisches Energiemodell entwickelt sowie eine (Back-) Annotation von geschätzten Energiewerten der im NN durchzuführenden arithmetischen Operationen in Abhängigkeit der Datenbreite der Parameter vorgenommen. Damit konnte schon frühzeitig, während der NN-Spezifikation, eine energetische Bewertung verschiedener NN-Modelle durchgeführt werden, ohne diese komplett in Hardware umsetzen und simulieren zu müssen. Diese Optimierungen

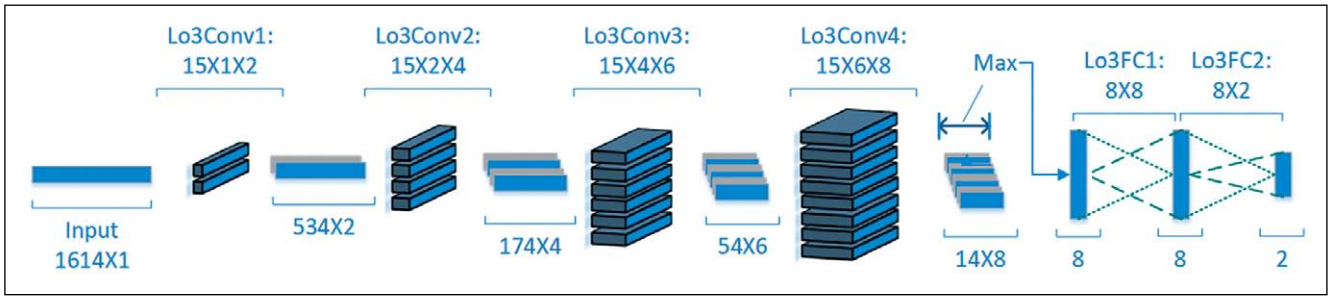


Bild 6. Das finale NN-Modell besteht aus vier Lo3Conv-Layern, einem Max-Pooling- und zwei Lo3FC-Layern. (Bild: Fraunhofer IIS)

wirken sich stark positiv auf die gesamte Verlustleistung aus und reduzieren die Gesamtenergieaufnahme um einen erheblichen Teil. Schon mit geringen Optimierungen an den Schichten des NNs kann so bis zu 50 % der Energie eingespart werden.

Datenübertragung ohne komplizierte Steuerung

Die Architektur ist so konzipiert, dass jede Schicht des NNs durch ein entsprechendes Schaltungsmodul realisiert wird. Der Datenfluss zwischen den Schichten erfordert zudem weder eine Pufferung noch eine aufwendige Steuerung. Er stimmt zudem jeweils am Ausgang eines Moduls mit dem am Eingang des nächsten Moduls überein, um bei der Verknüpfung dieser Module Steueraufwand einzusparen. Hierdurch wird ferner eine flexible Verbindung zwischen Komponenten geschaffen, um andere Netzwerkstrukturen schnell realisieren zu können. **Bild 7** zeigt eine schematische Darstellung der Architektur. Die verschiedenen Kanäle einer NN-Schicht werden dabei parallel berechnet, haben jedoch unterschiedliche Verzögerungen, um eine Verarbeitung ohne Pufferung im nächsten Modul zu erleichtern. Die Schaltung ist damit außerdem genau auf die Datenrate der einzelnen Schichten abgestimmt. Durch die Verwendung von Strides, also Datenreduktion, in den Filtern der Convolutional- und FC-Schichten, ist die Datenrate am Eingang jeder Schicht anders als am Ausgang. Diese nimmt damit in den hinteren Schichten des NNs im Vergleich zu den vorderen Schichten immer stärker ab. Die Register im Datenpfad werden daher mittels Clock-Gating und dem Einsatz von Valid-Signalen gesteuert. Die so entstandene Logik zur Steuerung des Datenflusses ist klein und sehr effizient implementierbar.

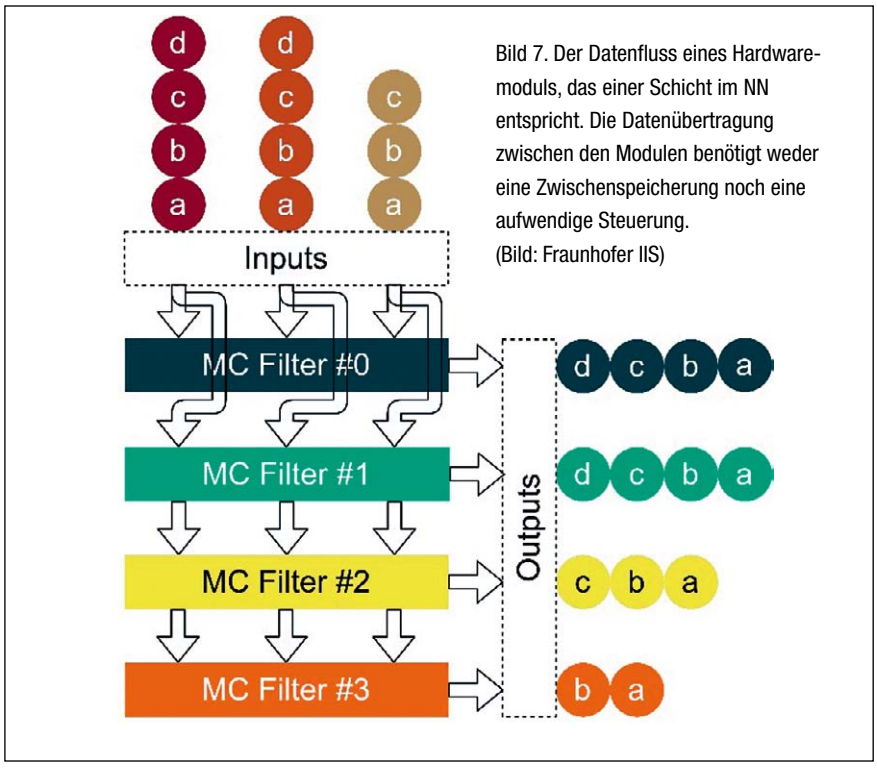


Bild 7. Der Datenfluss eines Hardwaremoduls, das einer Schicht im NN entspricht. Die Datenübertragung zwischen den Modulen benötigt weder eine Zwischenspeicherung noch eine aufwendige Steuerung. (Bild: Fraunhofer IIS)

Die HW-Einheiten zur Realisierung arithmetischer Operationen bei der Umsetzung der rechenintensiven Schichten des NN werden als Processing Elemente (PE) bezeichnet und sind in **Bild 8** schematisch dargestellt. Die lokal erforderlichen Gewichte sind in den zuvor genannten nicht-volatilen Parameterspeicher abgelegt und den jeweiligen PE zugeordnet. Ein PE führt eine Multiplikation eines Eingangswertes mit einem Gewichtswert durch und gibt den Eingangswert zusätzlich an das nächste PE weiter. Zur Steuerung des Datenflusses haben alle PEs Clock-Enable-Eingänge, um mithilfe eines einfachen Datenflusszählers die Berechnung nicht benötigter Teilergebnisse zu vermeiden.

Optimum bei drei Zuständen

Eine maßgebliche Innovation des hier skizzierten Entwurfs ist die Anpassung des NNs und der Schaltung an die verfügbaren Logikzustände des eingesetzten RRAM. Um eine aus Energieeffizienzsicht optimale Anzahl an Zuständen zu erhalten, wurden dazu Voruntersuchungen durchgeführt. Neben der Abschätzung der Energie, die für eine Lese- und Schreibansteuerung der RRAM-Zellen notwendig ist, wurde auch die oben beschriebene Energie-Annotation für verschiedene neuronale Netze durchgeführt. Das Resultat der Untersuchung zeigt ein Optimum bei drei Zuständen in den RRAM-Zellen,

wofür eine effiziente und platzsparende, analoge Ausleseschaltung und ein bezüglich der Energie effizienteres NN umsetzbar sind. Beispielsweise fällt die Energie für ein NN mit binären Gewichtszuständen 1,6-mal höher aus als die eines NNs mit ternären Gewichten. Um mit wenig Aufwand eine Entwurfsraum-Exploration durchführen zu können, wurde ein eigenes Hardwaregenerierungs-Tool entwickelt. Dieses setzt aus den selbst entworfenen Hardwarekomponenten den Beschleunigerkern zusammen und parametrisiert automatisch alle Komponenten.

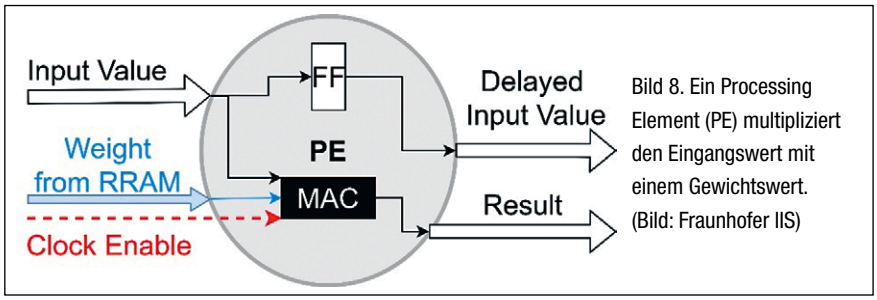


Bild 8. Ein Processing Element (PE) multipliziert den Eingangswert mit einem Gewichtswert. (Bild: Fraunhofer IIS)

RRAM-Zellen mit analoger Peripherie

Das integrierte RRAM [7, 8] ist in Speicherblöcke mit je 32 Zellen eingeteilt. Jede dieser Speicherzellen kann drei verschiedene Werte annehmen: Einen »High Resistive State« (HRS) und jeweils zwei »Low Resistive States« (LRS1 und LRS2). Um ein paralleles Auslesen zu erlauben, ist jeder Speicherblock mit der benötigten, peripheren Analogschaltung ausgestattet. Diese umfasst je einen Operationsverstärker (OPV) und einen Referenzenblock zur Erzeugung der benötigten Spannungspegel. Zusätzlich ist jede 1T1R-Zelle (Transistor + RRAM) in eine Speicherzelle mit peripherer Beschaltung nach **Bild 9** eingebettet. Die Ansteuerung der RRAM-Zellen erfolgt über Spannungspulse, die entweder an den Bitline-Anschluss (BL) oder an den Sourceline-Anschluss (SL) angelegt werden. Der Wordline-Anschluss (WL) der ausgewählten Zellen wird in dieser Phase auf eine definierte Spannung gelegt. Generell sind für jede Operation unterschiedliche Spannungspulse sowie WL-Spannungen notwendig.

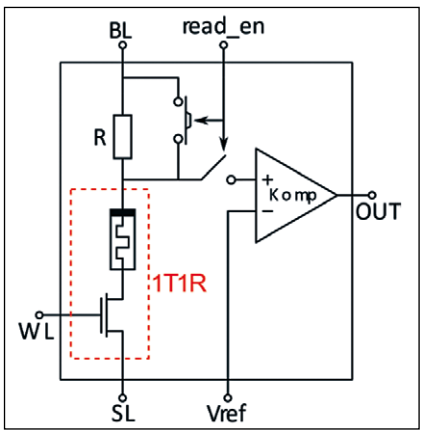
Die Methode zum Auslesen des Widerstandswertes einer 1T1R-Zelle basiert auf der Auswertung der Spannung am Knoten eines Spannungsteilers, bestehend aus dem Vergleichswiderstand R sowie dem Widerstand der Speicherzelle (Bild 9). Nach dem Aktivieren der WL wird ein Spannungspuls von 0,5 V am BL-Anschluss angelegt, während der SL-Anschluss auf Massepotenzial gelegt ist. Die Spannung am Knoten zwischen dem Vergleichswiderstand R und der 1T1R-Zelle hängt somit vom Zellenwiderstand ab. Diese Spannung wird vom Komparator mit einer Referenzspannung U_{Ref} verglichen.

Der Reset-Prozess, bei dem die Speicherzelle zunächst in den HRS gesetzt wird, setzt voraus, dass der Programmierstrom durch die 1T1R-Zelle in die entgegengesetzte Richtung, also vom SL-Anschluss zum BL-Anschluss, fließt. Daher muss während des Resets der Spannungspuls am SL-Anschluss angelegt werden, während der BL-Anschluss auf Massepotenzial liegt. Das Programmieren von LRS1 und LRS2 (Set) ausgehend von HRS unterscheidet sich bezüglich der Signalverläufe nicht vom Lesevorgang: Der Spannungspuls wird an die BL angelegt, während die SL mit Masse verbunden ist. Die WL-Spannung dient hier zur Begrenzung des Schreibstromes.

Arbeitsteilung und Ruhepausen sparen Energie

Abschließend wurden die implementierten Komponenten in ein Gesamtsystem zusammengefasst und dieses dazu in zwei separate ASIC-Blöcke aufgeteilt: Einen Data Control Core zur Kommunikation mit der Außenwelt für Dateneingabe (»Recording Interface«) und Konfiguration der RRAMs (»External RRAM Interface«) sowie zur Vorverarbeitung und Zwischenspeicherung von Daten, und einen Processing Core zur Ausführung der eigentlichen Berechnungen des NNs (**Bild 10**). Sämtliche digitalen Komponenten sind in VHDL implementiert und werden unter Verwendung der zugrundeliegenden Halbleiterfertigungstechnik durch Synthese- und Place&Route-Werkzeuge in Netzlisten bzw. in entsprechende Layouts überführt, wobei beide Blöcke separat und unabhängig voneinander betrachtet werden. Zur Adaption der RRAM-Technik an diesen Entwurfsablauf werden jeweils 32 RRAM-Zellen mit der zugehörigen analogen und digitalen Beschaltung zu einem Makroblock »RRAMBlock32« zusammengefasst. Von diesem können beliebig viele Instanzen in einer Schiebekette miteinander verbunden werden, was eine Ansteuerung von außerhalb zur ein- oder mehrmaligen Programmierung der Gewichtswerte erlaubt.

Anders als der Data Control Core, der die gesamte Zeit über aktiv ist und die vom Sensor mit geringer Abtastrate aufgenommenen Daten permanent mit einem Abtastwert pro Takt entgegennimmt und aufbereitet, ist die Versorgungsspannung des Processing Cores die meiste Zeit über abgeschaltet. Auf feingranulares Power-Gating wurde dabei auch aufgrund mangelnder Unterstützung der Bibliotheken der zugrundeliegenden Halbleiterfertigungstechnik verzichtet – stattdessen werden sämtliche Versorgungs-



(Bild: Fraunhofer IIS)

Bild 9. Steuerschaltung für eine RRAM-Zelle (1T1R) mit den drei Steuerleitungen BL (Bitline), SL (Sourceline) und WL (Wordline).

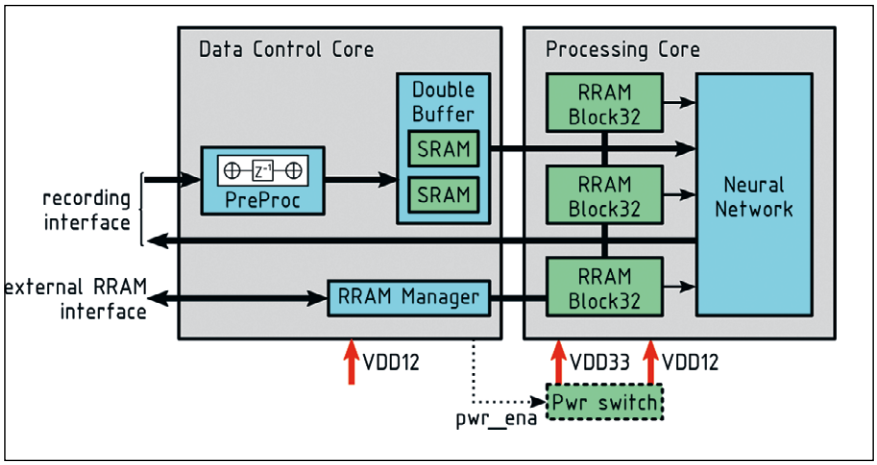


Bild 10. Aufbau des ASICs mit zwei Bereichen, dem Data Control Core für die Steuerung und Kommunikation nach außen und dem Processing Core für die Berechnungen des NN. Beide Bereiche haben eigene Stromversorgungen. (Bild: Fraunhofer IIS)

spannungen durch Schalttransistoren (»Pwr Switch« in Bild 9) abgeschaltet, die neben dem Processing Core platziert sind. Sobald seitens des Data Control Core genügend EKG-Daten bereitstehen, d.h. ein 12,7-s-Fenster vollständig eingelesen wurde, wird die Versorgungsspannung des Processing Cores aktiviert und unmittelbar das parallele Lesen der Gewichte aus den RRAMs veranlasst. Nach Abschluss der neuronalen Verarbeitung wird die Versorgungsspannung des Processing Cores zur Vermeidung unnötiger statischer Verluste sofort wieder deaktiviert.

Die Abschätzung der zur Verarbeitung der EKG-Aufnahmen erforderlichen Energie erfolgt durch eine Simulation auf Netzlistenebene anhand der tatsächlichen, zeitlichen Verläufe aller vorhandenen Signale. Die Aufzeichnungen verschiedener Simulationsläufe werden dafür, unter Hinzunahme der für den jeweiligen Halbleiterfertigungsprozess spezifischen Leistungsbeschreibung der Standardzellenbibliothek, mithilfe des Werkzeugs Synopsys PrimePower analysiert.

Sowohl für die Abtastrate des Sensors als auch für die Vorverarbeitung im Data Control Core wird eine Frequenz von 512 Hz angenommen. Zur Minimierung des Energiebedarfs im Taktbaum wird hingegen eine Verarbeitungsfrequenz von 70 kHz für den Processing Core gewählt. Bedingt durch das Frequenzverhältnis sowie das Downsampling durch die Vorverarbeitung erfordert die Verarbeitung eines 12,7 s umfassenden Fensters lediglich ca. 25 ms. Die neuronale Verarbeitung ist also lediglich in ca. 0,2 % der Zeit aktiv, in der verbleibenden Zeit ist die Versorgungsspannung abgeschaltet, wodurch auch keine statische Verlustleistung anfällt.

In der **Tabelle** sind Simulationsergebnisse dargestellt, wobei der Anteil der statischen Verluste dominiert. Durch die Abschaltung der Versorgungsspannung des Processing Cores und die korrespondierende Einsparung der statischen Verluste können deshalb im Vergleich zu einem »Always-on«-Ansatz ca. 94,7 % der Energie eingespart werden.

Erfolg durch Systemansatz und ganzheitliche Optimierung

Wenn der gesamte Entwurfsprozess – Algorithmus plus Tools, plus ASIC-Entwicklung, plus ASIC-Implementierung – gemeinsam optimiert wird, so kann tatsächlich einiges an KI-Signalverarbeitung in Geräten und Sensoren eingebaut werden, die sparsam mit Energie haushalten müssen. Eine gewisse Flexibilität sollte dabei jedoch vorhanden sein – allein schon, um spätere Updates zu ermöglichen – selbst wenn dies immer auch auf Kosten der Energieeffizienz geht. Gerade bei Zeitreihensignalen tröpfeln die Abtastwerte viel langsamer ein, als sie verarbeitet werden können. Hier lohnt sich der Einsatz von eingebetteten, nicht-flüchtigen Speichern, die das überwiegende Schlafenlegen des KI-Beschleunigers ermöglichen – hier 99,8 % der Zeit. Im gezeigten Fall spart dies ca. 95 % der Energie verglichen mit demselben Beschleuniger im »Always-on«-Modus.

Wer nun denkt, dass diese moderate Rechenlast genauso gut durch eine CPU, die dann allerdings nicht schlafen würde, abzuarbeiten wäre, der irrt jedoch: Durch den Wegfall der allermeisten Steuerlogik sowie die kleinen Wortbreiten und den Einsatz von für die Anwendung ausreichenden, ternären Gewichten, ist der beschriebene KI-Beschleuniger ein Vielfaches energieeffizienter als eine Standard-CPU. Das zeigt, dass spezielle KI-Beschleuniger in Zukunft häufig in ASICs integriert werden dürften, die damit gänzlich neue Anwendungen ermöglichen werden. hs

Block	dynamische Energieaufnahme (E_{Dyn}) [µJ]	statische Energieaufnahme [µJ]		Gesamtenergieaufnahme [µJ] im geschalteten Betrieb	gegenüber dem Dauerbetrieb eingesparte Energie [µJ]	Relative Energieersparnis [%]
		Dauerbetrieb (E_{Stat})	geschalteter Betrieb mit Ruhepausen ($E_{Stat,ein}$)			
Data Control Core	0,566	11,183	11,183	11,749	–	0
Processing Core	0,127	220,371	0,423	0,55	219,948	99,8
47 RRAM-Blöcke (analog)	0,053	1,188	0,002	0,055	1,186	95,5
Summe	0,746	232,742	11,608	12,354	221,134	94,7

Tabelle. Die Energieaufnahme der im Lo3-ML-ASIC integrierten Blöcke wurde per Simulation für ein 12,7 s langes Fenster bei einer Eingabefrequenz von 512 Abtastwerten pro Sekunde ermittelt. Durch die Ruhephase des HW-Beschleunigers und des RRAMs lässt sich deutlich Energie einsparen. (Bild: Fraunhofer IIS)

Die Autoren danken dem ganzen Lo3-ML-Team, besonders Daniel Reiser, Maen Mallah und Stefan Pechmann sowie Prof. Dr. Dietmar Fey, für die hervorragende Zusammenarbeit und ihre Beiträge zu diesem Aufsatz, dem Bundesministerium für Bildung und Forschung für die Förderung (Kennzeichen 16ES1143 und 16ES1142K), dem Projektträger VDI/VDE, dem DFKI Kaiserslautern und dem IHP Frankfurt/Oder für ihre Unterstützung während des Projektes.

Literatur:

[5] *Breiling, M.; Reichel, P. und Reichenbach, M.:* Energieeffizientes KI-System – Teil 1: AI Goes Ultra-Low-Power. *Elektronik* 2021, H. 19, S. 22–25.

[6] *Guo, Y.:* A survey on methods and theories of quantized neural networks. arXiv, 1808.04752, 13. August 2018, https://arxiv.org/abs/1808.04752

[7] *Grossi, A.; et al.:* Impact of the precursor chemistry and process conditions on the cell-to-cell variability in 1T-1R based HfO2 RRAM devices. Scientific Reports, 24. Juli 2018, https://www.nature.com/articles/s41598-018-29548-7

[8] *Reuben, J.; Fey, D. und Wenger, C.:* A Modeling Methodology for Resistive RAM Based on Stanford-PKU Model With Extended Multilevel Capability. IEEE Transactions on Nanotechnology, 2019, S. 647–656, DOI: 10.1109/TNANO.2019.2922838.

Dr. Marco Breiling

studierte Elektrotechnik in Karlsruhe, Trondheim, Paris und Southampton und promovierte in Erlangen. Er ist seit 2001 am Fraunhofer-Institut für Integrierte Schaltungen IIS in Erlangen tätig. Dort arbeitet er als Chief Scientist an den neuromorphen Hardwareentwicklungen.
marco.breiling@iis.fraunhofer.de

Dr. Peter Reichel

studierte in Dresden Informationssystemtechnik und promovierte in technischer Informatik. Er ist seit 2011 am Fraunhofer-Institut für Integrierte Schaltungen IIS, Institutsteil Entwicklung Adaptiver Systeme EAS in Dresden, tätig und arbeitet als wissenschaftlicher Mitarbeiter im Bereich intelligenter Sensorik.
peter.reichel@eas.iis.fraunhofer.de

Dr. Marc Reichenbach

studierte Informatik in Jena und promovierte in Erlangen. Seit 2010 ist er an der Friedrich-Alexander-Universität (FAU) am Lehrstuhl Rechnerarchitektur tätig. Als Post-Doktorand forscht er dort an neuen energieeffizienten Rechnerarchitekturen u.a. für Anwendungen aus dem Bereich der künstlichen Intelligenz.
marc.reichenbach@fau.de

Director Content Electronics: Dr. Ingo Kuss
Markenteam: Joachim Kroll (jk/1335), Chefredakteur (verantwortlich für den Inhalt), Markus Kien, Chef vom Dienst (mk/1333)
Redaktionsteam: Melanie Erhardt (me/1346), Markus Haller (mha/1371), Ralf Higgelke (rh/1341), Engelbert Hopf, Chefreporter (eg/1320), Ute Häußler (uh/1369), Irina Hübner (ih/1339), Andreas Knoll, Ltd. Red. (ak/1319), Corinna Puhlmann-Hespen (cp/1316), Corinne Schindlbeck, Ltd. Red. (sc/1311), Tobias Schlichtmeier (ts/1368), Harry Schubert (hs/1338), Iris Stroh, Ltd. Red. (st/1326), Kathrin Veigel (kv/1746), Nicole Wörner (nw/1325), Karin Zühlke, Ltd. Red. (zü/1329)
Die Ressortverteilung entnehmen Sie bitte der Internetseite elektroniknet.de/electronics-redaktion
Redaktionsassistenz: Andrea Seidel (sei/1332)
Layoutteam: Wolfgang Bachmaier (Ltg.), Andreas Geyh, Norbert Preiss, Bernhard Süßbauer, Alexander Zach

So erreichen Sie die Redaktion: Tel.: 089 25556-1332; Fax: 089 25556-1670
redaktion@elektronik.de, www.elektronik.de

Director New Business: Marc Adelberg (1572)
Sales Director New Business: Carolin Schlüter (1570)
Sales Director New Electronics: Christian Stadler (1375)
Regional Sales Managers: Petra Beck (1378), Burkhard Bock (1305), Tanja Lewin (1386), Konrad Nadler (1382), Martina Niekrawietz (1309)
Sales Operations Specialist: Simone Schiller (1383)
Assistenz: Rosi Böhm (1307), Michaela Stolka (1376)
Anzeigenverwaltung und Disposition: Astrid Brück (1471), Teresa Manuri (1482)
International Account Managers: Konrad Nadler (1382), Martina Niekrawietz (1309)
Auslandsrepräsentanz (Foreign Representation):
USA: Véronique Lamarque, E&Tech Media, llc, 80 Kendrick Street, Brighton, MA 02135, Phone/Fax: +1 860-536-6677, E-Mail: veroniquelamarque@gmail.com, Skype: E&Tech Media
Anzeigenpreise: Es gilt die Anzeigenpreisliste Nr. 56 vom 1. Januar 2021

So erreichen Sie die Anzeigenabteilung: Tel.: 089 25556-1376; Fax: 089 25556-1651
media@elektronik.de, www.elektroniknet.de/media

Vertriebsleiter: Marc Schneider (1509, mschneider@weka-fachmedien.de)
Bestell- und Abonnement-Service: WEKA Fachmedien GmbH, c/o Zenit Pressevertrieb GmbH, Postfach 810640, 70523 Stuttgart, Tel.: 0711 7252-210, Fax: 0711 7252-333, abo@weka-fachmedien.de
Bestellungen Schweiz: Thali AG, Industriestr. 14, CH-6285 Hitzkirch, Tel.: 041 9196611, Fax: 041 9196677, abo@thali.ch, www.thali.ch
Organschaft: Die Elektronik ist Organ der VDE/VDI-Gesellschaft Mikroelektronik, Mikrosystem- und Feinwerktechnik (GMM). Die Mitglieder der GMM erhalten die Elektronik im Rahmen ihrer Mitgliedschaft.

Erscheinungsweise: 26 Ausgaben

Jahresabonnement Print Inland	179,00 €, davon 115,30 € Heft, 63,70 € Versand
Jahresabonnement Print Ausland	201,10 €, davon 115,30 € Heft, 85,80 € Versand inkl. der aktuellen MwSt.
Einzelausgabe Print	8,00 € inkl. der aktuellen MwSt., zzgl. 3,00 Euro Versandkosten
Jahresbezug digitales E-Paper	69,99 € inkl. der aktuellen MwSt., ohne Versandkosten (Inland/Ausland)
Einzelausgabe digitales E-Paper	2,99 € inkl. der aktuellen MwSt., ohne Versandkosten (Inland/Ausland)

shop.weka-business-communication.com
70. Jahrgang, ISSN 0013-5658, Vertriebskennzeichen ZKZ 2594

Leitung Herstellung: Marion Stephan (1442)
Sonderdruck-Dienst: Alle in dieser Ausgabe erschienenen Beiträge können für Werbezwecke als Sonderdrucke hergestellt werden. Anfragen an Melanie Griesbach, Tel. 089 25556-1440, mgriesbach@wekanet.de
Druck: L.N. Schaffrath GmbH & Co. KG DruckMedien, Marktweg 42–50, 47608 Geldern, auch Anschrift für Beihefter und Beilagen.

Geschäftsführer: Kurt Skupin, Matthäus Hose

Anschrift für Verlag, Redaktion, Vertrieb, Anzeigenverwaltung und alle Verantwortlichen:
WEKA Fachmedien GmbH, Richard-Reitzner-Allee 2, 85540 Haar
Tel. 089 25556-1000, Fax 089 25556-1399, www.weka-fachmedien.de
Telefon-Durchwahl im Verlag: Sie wählen 089 25556 und dann die Nummer, die in Klammern hinter dem jeweiligen Namen angegeben ist.

Mitglied der Informationsgemeinschaft zur Feststellung der Verbreitung von Werbeträgern e.V. (IVW), Bad Godesberg