Schlussbericht Teil I: Kurzbericht

ZE:	Ferroelectric Memory GmbH Charlotte-Bühler-Str. 12 01099 Dresden	Förderkennzeichen:	16MEE0112S
Vorhabenbezeichnung:	KI für neue Elektroniksysteme und Edge-Computing-Technologien **-ANDANTE-**		
Laufzeit des Vorhabens:	07/2020 – 11/2023		
Berichtszeitraum:	07/2022 – 11/2023		

1 Aufgabenstellung

Im Projekt Andante plante FMC ein auf ferroelektrischen Feldeffekt-Transistoren (FeFET) basierendes Speichermakro zu entwickeln, welches gleichzeitig die Anforderungen des Anwendungsfalls „Speicher für Programmcode und Daten“ als auch die Anforderungen des Anwendungsfalls „Beschleuniger für KI-Anwendungen“ erfüllt. Die Zieltechnologie war eine 28nm Bulk Technologie, da die FeFET-Prozessentwicklung bisher so gut wie ausschließlich auf dieser Technologie erfolgte.

Um die Grundlage für den Schaltkreisentwurf zu schaffen, sollte ein „Physical Design Kit“ (PDK) entwickelt werden, welches den Schaltkreisentwurf des Makros ermöglicht, erleichtert und möglichst automatisiert.

2 Wissenschaftlicher und technischer Stand

FMC arbeitet an der Entwicklung der FeFET-Technologie seit seiner Gründung 2016. Das FMC-Team entwarf mehr als sechs Testchips für die Entwicklung des FeFET-Herstellungsprozesses bei Foundry-Partnern. FMC entwickelt außerdem ein auf die Technologie zugeschnittenes Charakterisierungssystem, welches fortlaufend erweitert wird. FeFET-Speicherzellen erreichen mehr als 100M Schreibyklen und mehr als 24h Datenhaltung bei 200°C.

3 Ablauf des Vorhabens

FMC war in den Arbeitspaketen 2 und 3 involviert. Im Arbeitspaket 2 war FMC an der Arbeitsaufgabe T2.1 beteiligt, in der es insbesondere um die Demonstration der Skalierbarkeit von FeFET-basierten Speichern hin zu größeren Anwendungen künstlicher Intelligenz ging. Dafür wurden FeFET Speicherzellen und Speicherarrays charakterisiert und auf verschiedene, in Hinblick für Speicher wichtige, Merkmale untersucht. Weiterhin wurde in diesem Arbeitspaket ein „Physical Design Kit“ (PDK) entwickelt, welches eine nahtlose Integration in die Chip-Design-Software Cadence Virtuoso ermöglicht.

Sowohl die gewonnenen Erkenntnisse aus den elektrischen Messdaten, wie auch das entwickelte PDK flossen in die Tätigkeiten in Arbeitspaket 3 ein. In diesem Arbeitspaket arbeitet FMC an der Arbeitsaufgabe T3.4. Dabei ging es um die Entwicklung von sogenannter Foundation IP, also Baublöcke, die in größeren Designs verwendet werden können, um komplexere Neuronale Netze und Anwendungen zu realisieren. FMC entwickelte ein Speichermakro, welches für die Verarbeitung künstlicher neuronaler Netze geeignet ist und auf ferroelektrischen Feldeffekttransistoren als Speicherelement basiert. Dafür wurden die Notwendigen Arbeiten wie z.B. Architektur-Entwurf, Schaltkreisentwurf des Analog- und Digitalteils und Verifikation durchgeführt.

Ein ursprünglich geplantes Tapeout wurde im Laufe des Projektes durch eine FPGA-Implentierung ersetzt. Dadurch waren weitere Tätigkeiten, wie die Portierung der digitalen Hardwarebeschreibung auf die FPGA-Plattform, die Entwicklung eines PCBs, diverse Programmier- und Verifikationstätigkeiten notwendig.

4 Ergebnisse

FMC konnte im Projekt ANDANTE eine Speichermakro entwickeln, welches für die Anwendung im Bereich Edge-AI geeignet ist und zudem einen eingebauten Beschleunigerteil für neuronale Faltungsnetzwerke besitzt. Es wurde eine FPGA-Plattform genutzt, um den Digitalteil des Makrodesigns zu testen, wobei die Daten auf einem bereits existierenden, von FMC entwickeltem, FeFET-Testchip gespeichert wurden und vom FPGA-Design gelesen und verarbeitet wurden. Mit diesem System konnten Kamerabilder eines PCs verarbeitet werden, wobei damit die Echtzeitfähigkeit des entwickelten Makros demonstriert werden konnte.

Der Entwurf des Makros setzte voraus, dass ferroelektrische Feldeffekt-Transistoren im Design integriert werden können. Dafür wurde erfolgreich ein Physical Design Kit (PDK) entwickelt, welches sich nahtlos in die Designsoftware Cadence Virtuoso integriert, sowie im Zusammenspiel mit dem von GlobalFoundries bereitgestellten PDK für die CMOS-Technologie reibungslos funktioniert. Dabei wurden alle benötigten Bestandteile, die für die Erstellung von Schaltplänen und Layouts, sowie für die Durchführung von Simulationen und der Physikalischen Verifikation notwendig sind, entwickelt.

Durch die Charakterisierung der Speicherzellen konnte FMC FeFET-Parameter in Abhängigkeit der durchgeführten Prozessierung bei der Fertigung bewerten. Es konnten Erkenntnisse über die Verteilung der Parameter in größeren Speicherarrays gewonnen werden, die wiederum für den Schaltkreisentwurf genutzt werden konnten. Trotz den erreichten Fortschritten in der Analyse der FeFET-Zellen und der gewonnenen Erkenntnisse zu Lösch- und Programmieralgorithmen und deren Parameter ist die FeFET-Entwicklung nicht weit genug fortgeschritten. Besonders die erreichten Bitfehlerraten für produktrelevante Speicherfeldgrößen genügen noch nicht den Ansprüchen für ein Produkt.

Schlussbericht Teil II: Eingehende Darstellung

ZE:	Ferroelectric Memory GmbH Charlotte-Bühler-Str. 12 01099 Dresden	Förderkennzeichen:	16MEE0112S
Vorhabenbezeichnung:	KI für neue Elektroniksysteme und Edge-Computing-Technologien **-ANDANTE-**		
Laufzeit des Vorhabens:	07/2020 – 11/2023		
Berichtszeitraum:	07/2020 – 11/2023		

1 Einleitung

FMC beschäftigt sich mit der Entwicklung, Charakterisierung, Vermarktung von Lösungen für eingebettete nichtflüchtige Speicher auf Basis von FMC's FeFET-Technologie. Das Ziel im Projekt ANDANTE war die Erforschung und Entwicklung neuartiger Anwendungen für eingebettete künstliche Intelligenz. Dafür entwickelte FMC ein „Physical Design Kit" (PDK) für die FeFET-Technologie, um die Technologie in einem Schaltkreisdesign nutzen zu können. Neben der Weiterentwicklung der Prozesstechnologie stand die Entwicklung eines Speichermakros im Fokus, welches gleichzeitig die Anforderungen des Anwendungsfalls „Speicher für Programmcode und Daten" als auch die Anforderungen des Anwendungsfalls „Beschleuniger für KI-Anwendungen" befriedigt. Im Projekt ANDANTE war FMC an den Arbeitspaketen 2 und 3 beteiligt.

2 Arbeitspaket 2

Arbeitspaket 2 beschäftigt sich mit der Untersuchung neuer Speichertechnologien für die Anwendung künstlicher Intelligenz, sowohl für Künstliche Neuronale Netze (KNN) als auch Pulsende Neuronale Netze (SNN), wobei FMC sich im Rahmen des Projektes auf KNNs konzentriert hat. Dabei war FMC Teil der Arbeitsaufgabe T2.1. Ein Projektziel, welches FMC verfolgt hat, war die Demonstration der Skalierbarkeit hin zu größeren Anwendungen künstlicher Intelligenz. Dabei ging es insbesondere um die Integration ferroelektrische Feldeffekttransistoren, die FMC für den 28nm-Technologieknoten bei Globalfoundries untersuchte. Ein weiteres Ziel war die Erstellung eines Physical Design Kits (PDK), welches die Entwicklung von Schaltkreisen ermöglicht und somit eine Voraussetzung für Arbeitspaket 3 ist.

2.1 Charakterisierung der FeFET Speicherzellen

Im Rahmen des Arbeitspakets 2 wurden die FeFET Speicherzellen charakterisiert und auf verschiedene, in Hinblick für Speicher wichtige, Merkmale untersucht. Dabei kam unter anderem ein bereits existierender Testchip zum Einsatz, der es ermöglichte, Prozessfehler und Fehlermechanismen zu erkennen und das Wafermaterial zu bewerten.

Eine wichtige Kenngröße der FeFET Speicherzelle ist die Schwellspannung Vth des ferroelektrischen Transistors nach dem Programmieren oder Löschen der Zelle. Mit dem Testchip war es möglich große Datenmengen der Vth-Verteilung zu sammeln und hinsichtlich intrinsischer und extrinsischer Faktoren zu untersuchen. Der Testchip stellte dafür zwei verschiedene Leseoperationen bereit.

Es wurden Messungen im sogenannten DMA-Modus, dem Direct-Memory-Access, durchgeführt. D.h. es wurden direkt die Drain-Ströme der ferroelektrischen Feldeffekt-Transistoren im Speicherfeld gemessen. So konnten Lesestrom-Verteilungen und FeFET-Transferkennlinien aufgenommen werden, indem die Wortleitungsspannung während der

Messung kontinuierlich hochgefahren wurde. Allerdings erfolgten diese Messungen mit einem externen Oszilloskop und benötigen eine entsprechende Integrationszeit, die den Zellzustand verfälschen können. Deshalb wurde eine weitere Methode angewendet, um die Zellzustände auszulesen.

Der Testchip verfügte weiterhin über eingebaute Leseverstärkerschaltkreise. Diese liefern eine digitale Ausgabe des gelesenen Zellzustands, d.h. 1 oder 0. Mit dieser Methode wurden Verteilungen der Schwellspannung Vth gemessen, indem bei einem konstanten Referenzstrom die Wortleitungsspannung erhöht wurde. Die Anzahl der gelesenen Einsen oder Nullen gibt dann Auskunft über die Verteilung der Schwellspannungen. Die beiden möglichen binären Zustände der Ferroelektrischen Transistoren sind dementsprechend der HVT- (High Vth) und der LVT-Zustand (Low Vth), wobei Zellen im HVT-Zustand bei gleicher Gate-Spannung einen niedrigeren Drainstrom liefern, als Zellen im LVT-Zustand. Die Differenz zwischen den Schwellspannungen in den verschiedenen Zuständen wird als Speicherfenster bezeichnet, welches eingehend untersucht wurde. Da die Zellzustände normalverteilt sind, ist auch das minimale Speicherfenster abhängig von der Anzahl der gemessenen Zellen. Für sehr große Statistiken beginnt sich das Speicherfenster irgendwann zu schließen, da sich LVT und HVT Zustände überlappen können. Um zu untersuchen, ab welchen Statistiken sich das Speicherfenster zu schließen beginnt, wird die sogenannte Bitfehlerrate herangezogen.

Mit den gewonnenen Erkenntnissen aus der Schwellspannungsverteilung wurden weitere Messungen zur Bit-Fehler-Rate durchgeführt. Dazu wurden dedizierte Muster in die Speicherfelder programmiert und wieder ausgelesen und mit dem erwarteten Ergebnis verglichen. Aus der Anzahl der falsch zurück gelesenen Bits und der Gesamtzahl der Speicherzellen wurden dann die Bit-Fehler-Raten berechnet. Die Bit-Fehler-Rate wurde zudem als Kennzahl zur Bewertung der Materialqualität der Wafer herangezogen.

Das untersuchte Wafermaterial zeigte zudem einen Effekt, der als Wakeup-Effekt bezeichnet wird. Dabei kann man feststellen, dass sich Eigenschaften, wie z.B. das Speicherfenster, verbessern, wenn die FeFET-Zellen für einer bestimmten Anzahl an Programmier- und Löschvorgängen unterzogen werden. Es wurde untersucht, wann das Speicherfenster am größten ist, bevor es sich nach weiteren Schreibzyklen wieder verkleinert.

Auch die Datenhaltung, als ein wichtiger Parameter zur Bewertung von Speicherzellen, wurde untersucht, sowie die Zyklenfestigkeit, also wie viele Programmier- und Löschoperationen durchgeführt werden können, bis eine gewisse Alterung eintritt, die zu einer unzuverlässigen Operation des Speichers führt.

Durch das Testen einer großen Anzahl von FeFET-Speicherzellen in einem Speicherfeld konnte der Einfluss von extrinsischen Faktoren quantifiziert werden, die bei der ausschließlichen Untersuchung von Einzelzellen wahrscheinlich übersehen worden wären.

Die durchgeführten Untersuchungen zeigten, dass die Datenhaltung in einigen Prozess-Splits sehr vielversprechend waren. Es konnten zudem Bitfehlerraten von weniger als 10^{-5} in Sektoren mit der Größe von 131 KBits erreicht werden. Die Zyklenfestigkeit erreichte 10000 bis 100000 Schreibzyklen.

Die Ergebnisse wurden im Arbeitsergebnis D2.7 im Detail dargestellt.

2.2 Entwicklung des FeFET PDKs

Für das Schaltkreisdesign in Arbeitspaket 3 ist ein sogenanntes „PDK“, also ein Physical Design Kit notwendig. Dieses wurde in Arbeitspaket 2 entwickelt und enthält für den Schaltkreisentwurf wichtige Informationen in Form von verschiedenen Dateien, die von den zahlreichen Entwicklungswerkzeugen eingelesen und verarbeitet werden können.

Das entwickelte PDK erlaubt eine nahtlose Integration in das Software-Werkzeug „Virtuoso“ von Cadence und wird zusätzlich zu dem von Globalfoundries ausgelieferten PDK geladen.

Für die Nutzung der FeFET-Zelle in einem Schaltplan wurde ein Transistor-Symbol entwickelt. Es beschreibt neben dem Aussehen des Schaltelements, welche Ein- und Ausgänge das Element hat und welche Parameter im Schaltplan annotiert werden können.

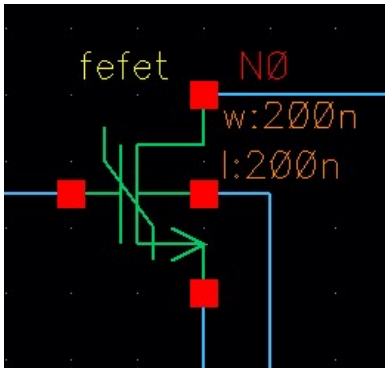


Abbildung 1: FeFET Schaltkreis-Symbol des entwickelten PDKs

Für den Schaltkreisentwurf werden weitere Informationen des Symbols benötigt, die im Hintergrund abgefragt werden, die sogenannten CDF-Parameter, die auch im Rahmen dieser Arbeit erstellt wurden. Diese beinhalten unter anderem Informationen zu den erlaubten Wertebereichen der Parameter, wie z.B. Weite W und Länge L, aber auch sogenannte Callbacks, also Funktionen, die aufgerufen werden, wenn der Nutzer einen Parameter ändert. Dann wird z.B. geprüft, ob der eingebene Wert akzeptiert wird, oder ggf. durch einen anderen ersetzt wird.

Auch eine Layout-Zelle wurde für das PDK entwickelt. Dadurch, dass die Parameter der FeFET-Zelle flexibel sind, kann auch das Layout der Zelle nicht starr sein. Dafür existieren sogenannte PCells, also programmierbare Zellen. Diese wurden mit der in Cadence Virtuoso integrierten Skriptsprache „Skill“ programmiert. Die Layoutzelle berechnet die Koordinaten und Abmessungen verschiedener für den FeFET benötigten Design-Layer auf Grundlage der eingegebenen Parameter.

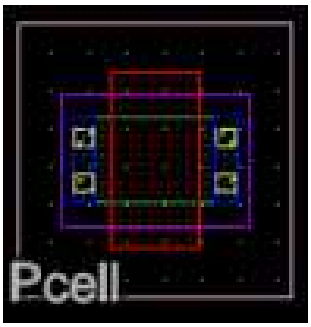


Abbildung 2: Layout-PCell des FeFET des entwickelten PDKs

Ein Teil der physischen Verifikation des Schaltkreisentwurfs ist der Design Rule Check oder kurz DRC, also eine automatische Kontrolle der eingehaltenen Entwurfsregeln. Es werden beispielsweise die Breite von Formen oder Abstände zwischen benachbarten Formen überprüft. Dies geschieht üblicherweise durch eine Art Skript-Datei, in der die einzelnen Entwurfsregeln mithilfe einer Skriptsprache beschrieben werden. In unserem Fall wurden die Layout-Entwurfsregeln für die Software Cadence PVS-DRC geschrieben.

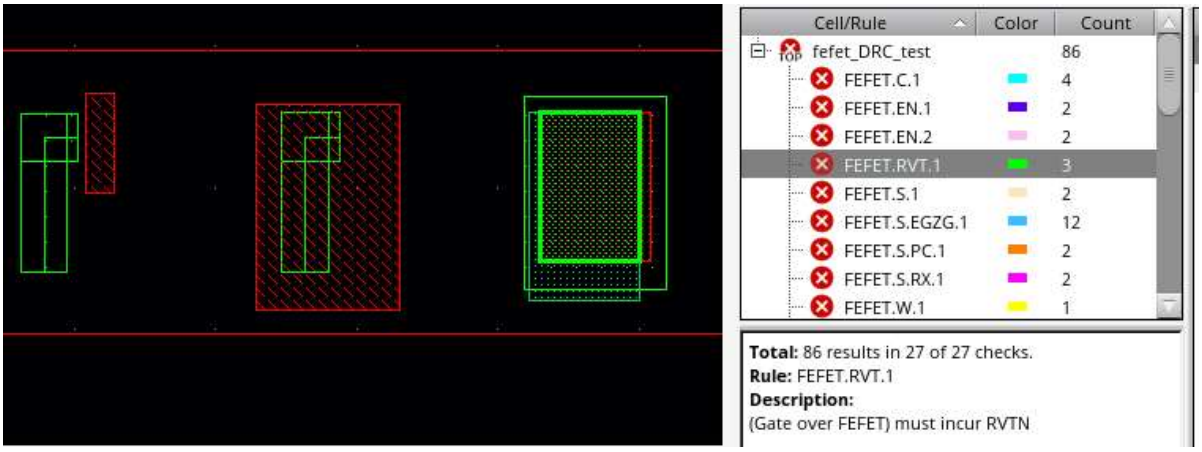


Abbildung 3: Testlayout und DRC Test-Ergebnis des entwickelten PDKs

Ein weiterer Teil der physischen Verifikation ist der LVS Check, also die Überprüfung, dass das erstellte Layout genau dem Schaltplan entspricht. Dabei wird ein Schaltplan aus dem Layout zurückextrahiert, um ihn mit dem ursprünglichen Schaltplan vergleichen zu können. Die Regeln, wie ein Schaltungselement aus dem Layout erkannt wird auch hier in einer Skript-Datei hinterlegt. Die LVS-Regeln für den FeFET wurden für die Software Cadence PVS-LVS geschrieben.

Um die implementierten Regel-Dateien für LVS und DRC zu überprüfen, wurden Testlayouts und -schaltpläne erstellt, die zum einen die Designregeln einhalten und zum anderen bewusst

verletzen, um auch erwartete Fehler aufzuwerfen. Diese Testdesigns wurden dann der DRC- und LVS-Prüfung unterzogen.

3 Arbeitspaket 3

Arbeitspaket 3 ist darauf ausgerichtet Werkzeuge und Architekturen für die verschiedenen Anwendungen künstlicher Intelligenz zu entwickeln und sogenannte Foundation IP zu entwickeln, also Baublöcke, die in größeren Designs verwendet werden können, um komplexere Neuronale Netze und Anwendungen zu realisieren. Das Ziel von FMC war es ein Speichermakro zu entwickeln, welches für die Verarbeitung künstlicher neuronaler Netze geeignet ist und auf ferroelektrischen Feldeffekttransistoren als Speicherelement basiert. Die Schaltkreisimplementierung sollte entweder auf dem 28nm SLP oder 22FDX Technologieknoten des Halbleiterfertigers Globalfoundries erfolgen, wobei die Wahl letztendlich auf den 28nm Knoten gefallen ist. Die Arbeiten waren Teil der Arbeitsaufgabe T3.4.

3.1 Architektur-Entwurf

Der Architekturentwurf des Speichermakros umfasste Überlegungen und Recherchen zur Schnittstelle des Schaltkreises, also mit welchem Datenprotokoll auf einfache Weise, aber mit hohem Datendurchsatz, die Stimuli und Gewichte für Neuronale Netze übertragen werden können. Weiterhin wurde die Speicherorganisierung, der Datenfluss und das Datenformat definiert, mit dem das Speichermakro am besten für die Bearbeitung künstlicher neuronaler Netze eingesetzt werden kann. Der Entwurf des Makros gliedert sich in Analog- und Digitalteil, die jeweils mit unterschiedlichen Entwurfsmethoden entwickelt wurden.

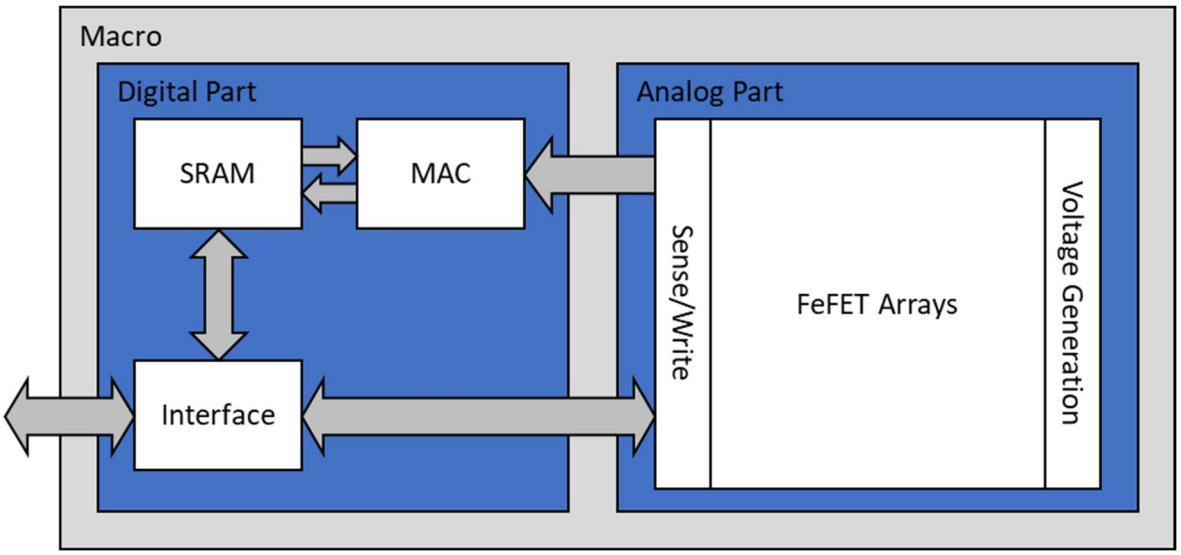


Abbildung 4: Architektur des in Arbeitspaket 3 entwickelten Makros

3.2 Schaltkreisentwurf Analogteil

Im Schaltkreisentwurf des Analogteils wurden zunächst verschiedene Unterbaublöcke entwickelt, die für das Funktionieren des Makros essenziell sind. Diese wurden im sogenannten Full-Custom-Design entwickelt, d.h. die Schaltpläne und Layouts wurden manuell erstellt, wobei wenig bis keine Automatisierung genutzt werden kann. Der Kern des Designs ist das Speicherzellenfeld. Dafür wurde zunächst eine Speicherzelle entwickelt, für die die Topologie und die Dimensionen und das Layout festgelegt wurden. In unserem Entwurf wurde zur Verbesserung der Zuverlässigkeit ein differenzielles Zellkonzept gewählt, welches das effektive Speicherfenster vergrößern kann und welches wiederum auch ein robusteres Referenzschema für das Auslesen der Zellzustände ermöglicht. Die Speicherzellen wurden dann in einem Speicherfeld zusammengeschaltet, sodass eine bestimmte Anzahl and Speicherzellen simultan ausgewählt, ausgelesen oder geschrieben werden können.

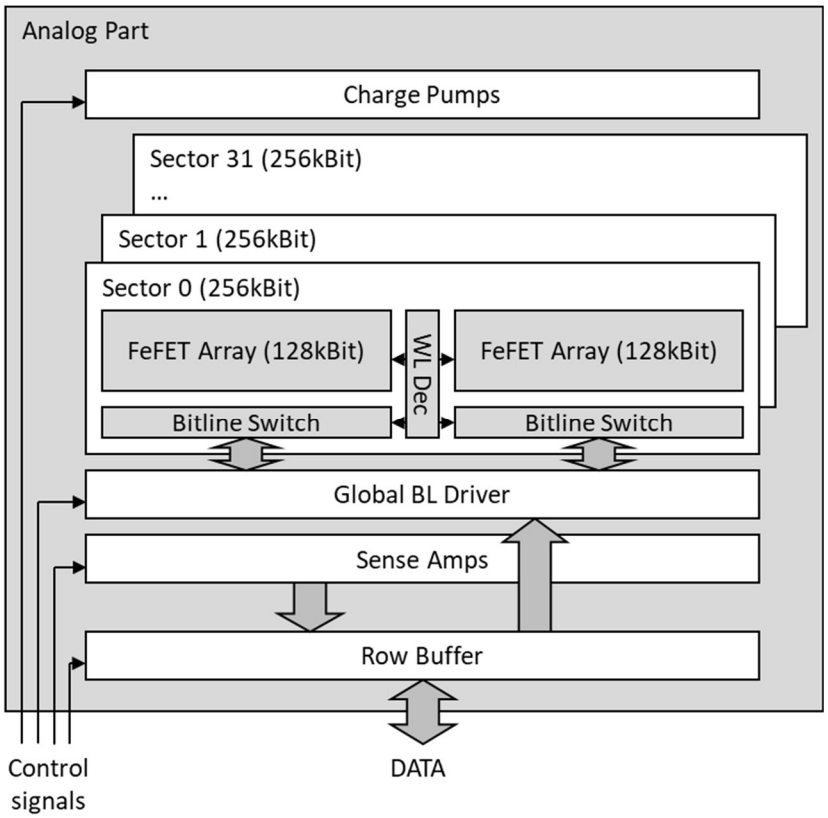


Abbildung 5: Blockschaltbild des entwickelten Analogteils

Um nun eine Adresse im Speicherfeld auszuwählen, wurde eine Wortleitungs-Dekoder-Schaltung implementiert und Bitleitungsschalter zum Verbinden der Bitleitungen des Speicherfeldes des ausgewählten Sektors mit den Leseverstärkern (Senseamps), sowie Bitleitungs-Treiber zum Programmieren und Löschen der Zellen entworfen.

Es wurde weiterhin ein Pufferspeicher für die gelesenen Daten implementiert, der es ermöglicht, bereits intern die Daten der folgenden Adresse zu lesen, um die Latenzzeit beim Lesen zu minimieren.

Für das Auslesen der Zellzustände wurden Ausleseverstärkerschaltkreise implementiert, die hinsichtlich Flächeneffizienz optimiert wurden, um ein Multiplexing zu vermeiden, und eine komplette Wortleitung auslesen zu können. Es wurde ein Source-seitiges spannungs-basiertes Auslesen verwendet, welches ein möglichst robustes Auslesen ermöglicht und zudem mit sehr wenigen und kleinen Transistoren auskommt.

Die FeFET-Speicherzelle benötigt zum Lesen und Schreiben verschiedene Spannungen, die teilweise die verwendeten Versorgungsspannungen überschreiten. Um diese Spannungen aus den vorhandenen Versorgungsspannungen zu generieren und bereitzustellen, wurden spezielle Ladungspumpen (Charge pumps) entwickelt, die auf die Lastkapazitäten der Speicherfelder abgestimmt sind und ein schnelles Hoch- und Herunterfahren der Spannungen ermöglichen. Diese Ladungspumpen wurden in einer „Latched Charge pump“-Topologie ausgeführt, welche effizient und zugleich einfach in der Ansteuerung ist.

Es wurden Simulationen durchgeführt, um die Funktionalität des Analogteils zu verifizieren. Dabei konnten auch die Geschwindigkeit der Leseoperation extrahiert werden, sowie der Stromverbrauch verschiedener Komponenten bei verschiedenen Operationen.

Alle Schaltungsblöcke wurden in ein Layout überführt und schließlich zusammengefügt. Dabei wurde Wert auf eine besonders flächeneffiziente Implementierung gelegt.

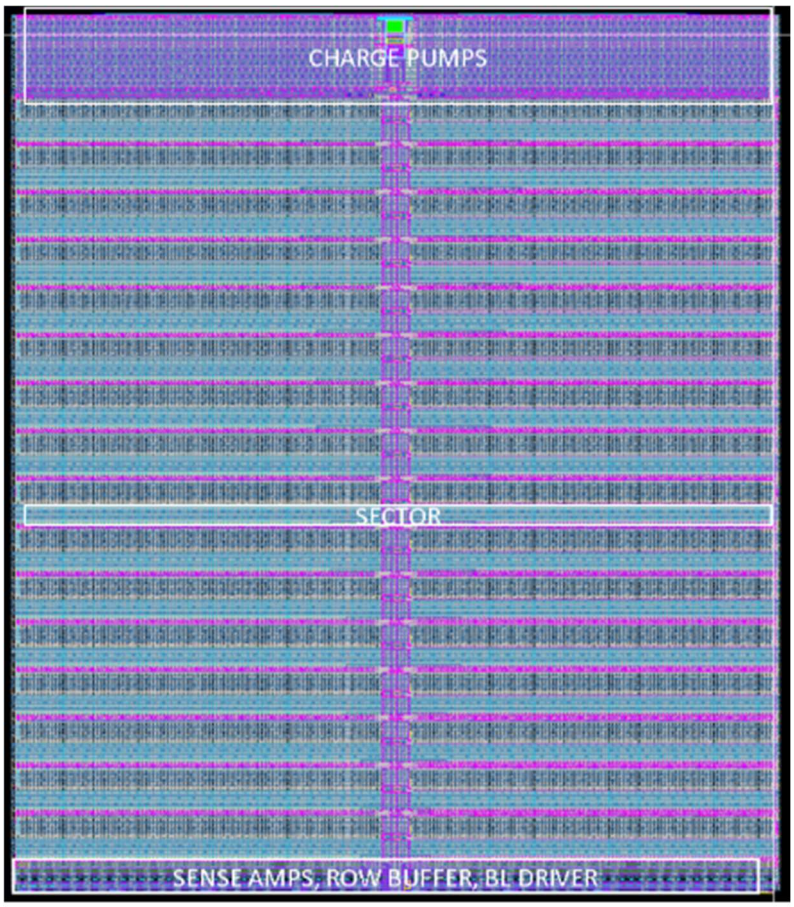


Abbildung 6: Layout des entwickelten Analogteils

Mithilfe des in Arbeitspakets 2 entwickelten FeFET-PDKs in Kombination mit dem von Globalfoundries für die 28nm-Technologie ausgelieferten PDKs wurde die physikalische Verifikation durchgeführt. Dabei wurden sowohl die Entwurfsregeln geprüft (DRC, design rule check), als auch, dass das entworfene Layout dem Schaltplan entspricht (LVS, layout vs. Schematic).

3.3 Schaltkreisentwurf Digitalteil

Im Gegensatz zum Full-Custom-Entwurf des Analogteils wurde der Digitalteil mit der Hardware-Beschreibungssprache Verilog implementiert. Dieser dient im Wesentlichen zur Ansteuerung des Analogteils, der Kommunikation mit der Außenwelt und der Beschleunigung von Operationen für Faltungsnetzwerke. D.h. der Digitalteil ermöglicht es, mit einem hohen Datendurchsatz sogenannte MAC-Operationen durchzuführen, bei denen also Multiplikationen und Additionen hoch-parallel und effizient durchgeführt werden. Dafür werden die Eingabedaten in einer Art und Weise sortiert, dass ein fast unterbrechungsfreier Strom von Daten verarbeitet werden kann.

Ein Teil der Arbeit entfiel dabei auf die Entwicklung der Steuerlogik des Analogteils. Dieser besteht im Wesentlichen aus einem endlichen Zustandsautomaten, der die Algorithmen zum Schreiben der FeFET-Speicherzellen ausführt und die Lesepufferspeicher steuert. Die Steuerlogik übernimmt zudem eine Vorkodierung der Adresssignale, die dann weiter zum Wortleitungs-Dekoder geleitet werden. Eine Besonderheit stellt hier die Steuerung des Lesepuffers dar. Dieser wird so angesteuert, dass bereits neue Daten aus dem FeFET-Array gelesen werden können, während die zuvor gelesenen Daten noch vom Datenpfad verarbeitet werden. Das ermöglicht eine sehr schnelle Ausführung der Neuronalen Netzen, da die benötigten Daten gewöhnlich seriell im Speicher abgelegt werden können und keine Sprünge in der Adressierung notwendig sind.

Für die Kommunikation mit der Außenwelt wurde eine FIFO-Schnittstelle implementiert, die mit handelsüblichen FTDI-USB-Chips kompatibel ist. Das Interface ermöglicht eine schnelle parallele Datenübertragung und eine einfache Softwareanbindung, da dafür benutzerfreundliche Bibliotheken für die Programmierung zur Verfügung stehen. Auch hierfür wurde ein Zustandsautomat entwickelt, der die Spezifikation des Kommunikationsprotokolls umsetzt und außerdem bereits Befehle und Daten unterscheiden kann und somit eine effiziente Übertragung und interne Weiterleitung der Daten ermöglicht.

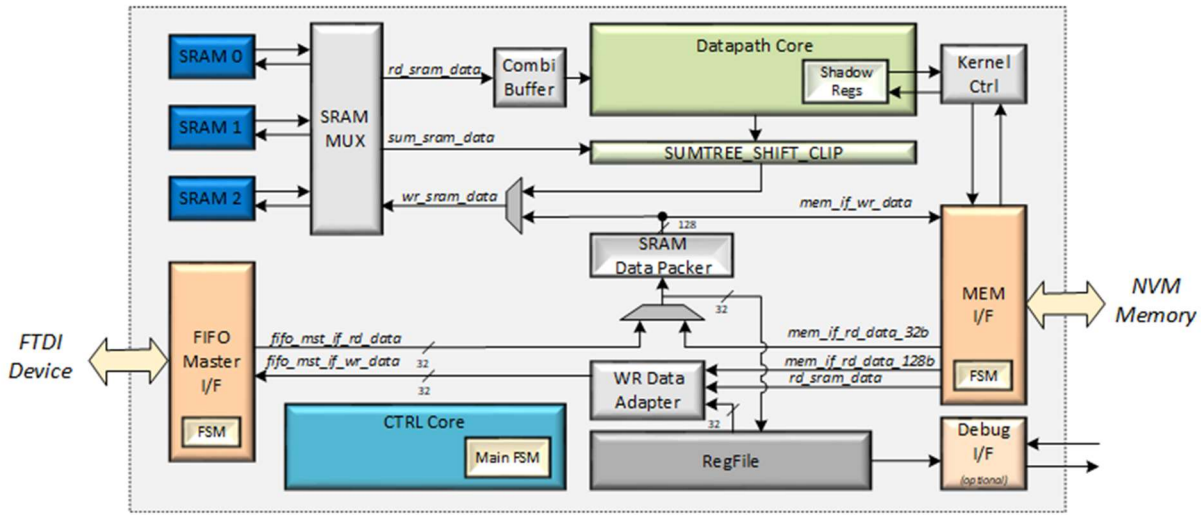


Abbildung 7: Blockschaltbild des entwickelten Digitalteils

Das Makro ist hinsichtlich der unterstützten Faltungsoperationen höchst flexibel. Es können verschieden große Tensoren für die Ein- und Ausgabedaten (Input/output feature maps) verarbeitet werden und auch verschiedene Kernel-Größen werden unterstützt. Weiterhin ermöglicht das Design verschiedene Padding-Konfigurationen, also das Auffüllen der Umrandung der Eingangsdaten mit Nullen, verschiedene Aktivierungsfunktionen und verschiedene Parameter für die Max-Pooling-Operation. All diese Konfigurationsdaten müssen für den Betrieb des Makros gespeichert werden. Dafür wurde ein sogenanntes

Register-File implementiert, in welchem die Einstellungen abgelegt werden, aber auch für eine Fehleranalyse ausgelesen werden können.

Den Kern des Beschleunigers bildet die MAC-Einheit im Datenpfad. In diesem werden die Multiplikations- und Akkumulationsoperationen ausgeführt. Dafür wurden Multiplizierer-Zellen für 8-bit vorzeichenbehaftete Daten implementiert. Die Ergebnisse werden dann in einen Summierungs-Baum eingespeist, der die einzelnen Multiplikationsergebnisse aufsummiert. Diese Struktur wurde in einer Matrix angeordnet, um eine parallele Prozessierung der Daten zu ermöglichen. Weiterhin wurden für diesen Schaltungsblock Schattenregister implementiert, die Kernel-Parameter für die Faltungsoperationen enthalten, und bereits während der Ausführung einer Operation mit neuen Werten im Hintergrund geladen werden kann, was die Ausführungszeit beschleunigt und vorteilhaft für die Ausführung von neuronalen Netzen in Echtzeit ist.

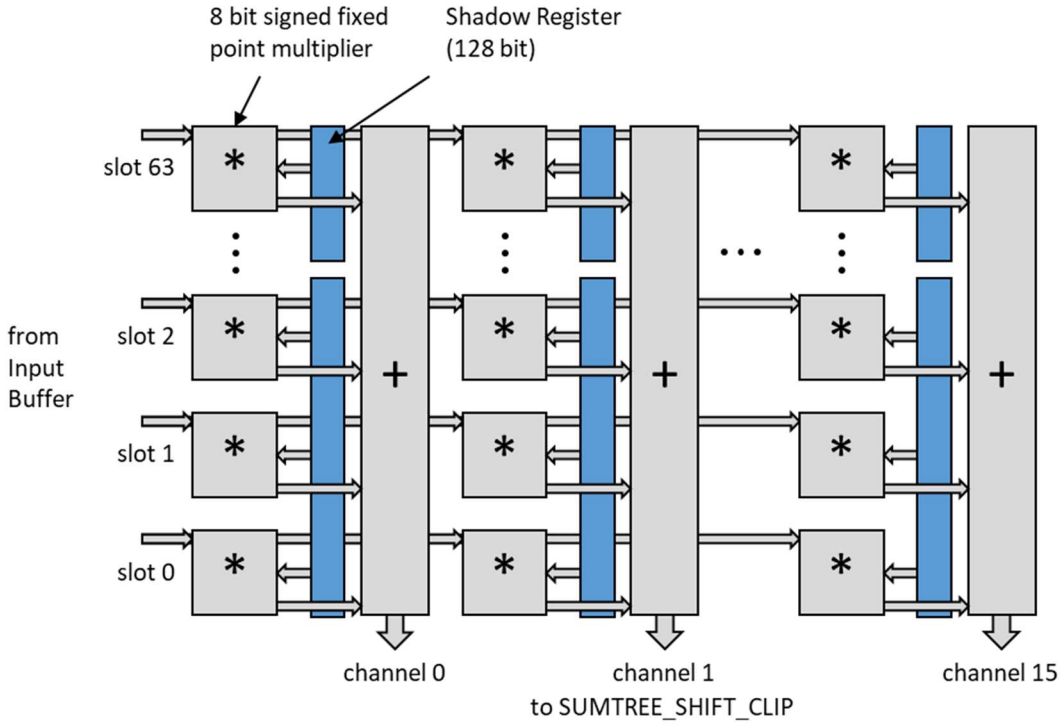


Abbildung 8: Blockschaltbild der MAC-Einheit

Da die Ausführung von neuronalen Netzen in einzelnen Schichten erfolgt, müssen die Ausgabedaten für die nächste Schicht zwischengespeichert werden. Das geschieht in flüchtigem SRAM, da der Speicherinhalt hier sehr oft geändert wird und eine nicht-flüchtige Speicherung in diesem Fall nicht geeignet ist.

Für die effiziente Ausführung von Faltungsoperationen müssen die Eingangsdaten ein einer bestimmten Reihenfolge zur MAC-Einheit geführt werden, um einen kontinuierlichen Datenfluss zu ermöglichen. Für diese Umsortierung der Eingangsdaten wurde ein

Schaltungsteil entwickelt, der auf verschiedene Kernel-Größen und verschiedene Größen der Eingangstensoren konfiguriert werden kann.

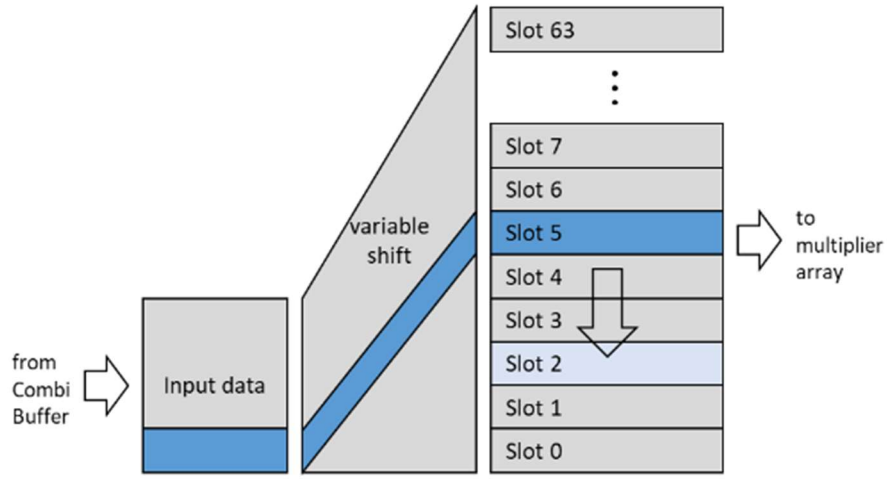


Abbildung 9: Prinzip der Umsortierung der Eingangsdaten

Um den kompletten Ablauf des Datenflusses zu steuern, wurde eine Steuerlogik implementiert. Diese unterscheidet zwischen Befehlen und Daten und bestimmt, ob die Ausführung vom Interface her oder aus den Daten aus dem FeFET-Array gesteuert wird. Dafür wurde ein Set von Befehlen definiert, mit dem alle benötigten Operationen abgedeckt werden können. Es können somit z.B. Kernel-Parameter in den FeFET-Arrays abgelegt, eine bestimmte Konfiguration ins Registerfile geschrieben, neue Eingangsdaten im SRAM abgelegt oder Ausgangsdaten aus dem SRAM ausgelesen werden.

3.4 Verifikation des Makro Designs

Um sicherzustellen, dass der Schaltkreisentwurf die spezifizierten Funktionen ausführt und keine ungewünschten Zustände auftreten, wurde eine Verifikationsumgebung implementiert. Es wurden zunächst Verifikationsmodule entwickelt, die die Funktion einzelner Baugruppen, wie z.B. der MAC-Einheit verifiziert.

Um den realen Betrieb des Makros zu simulieren, war es notwendig, Stimuli zu erzeugen, die alle für den Betrieb benötigten Schritte enthalten, wie das Generieren eines Speicherabbildes, welches im FeFET-Speicher abgelegt werden soll. Weiterhin musste eine Befehlsfolge erzeugt werden, um das Speicherabbild über das FIFO-Interface zu übertragen. Und schließlich wurden Eingabe-Daten erzeugt, die im Makro verarbeitet werden sollten.

Dafür war es notwendig Modelle neuronaler Netze zu erstellen bzw. anzupassen und Skripte zu entwickeln, die diese Netze auf die Hardwarekonfiguration abbilden. Das Trainieren der Neuronalen Netze geschah durch die Anwendung der quelloffenen Software Pytorch für die Programmiersprache Python. Nach dem Trainieren wird eine ONNX-Datei erzeugt, also ein

Dateiformat zur Beschreibung neuronaler Netze, welches dann zur Weiterverarbeitung herangezogen werden kann.

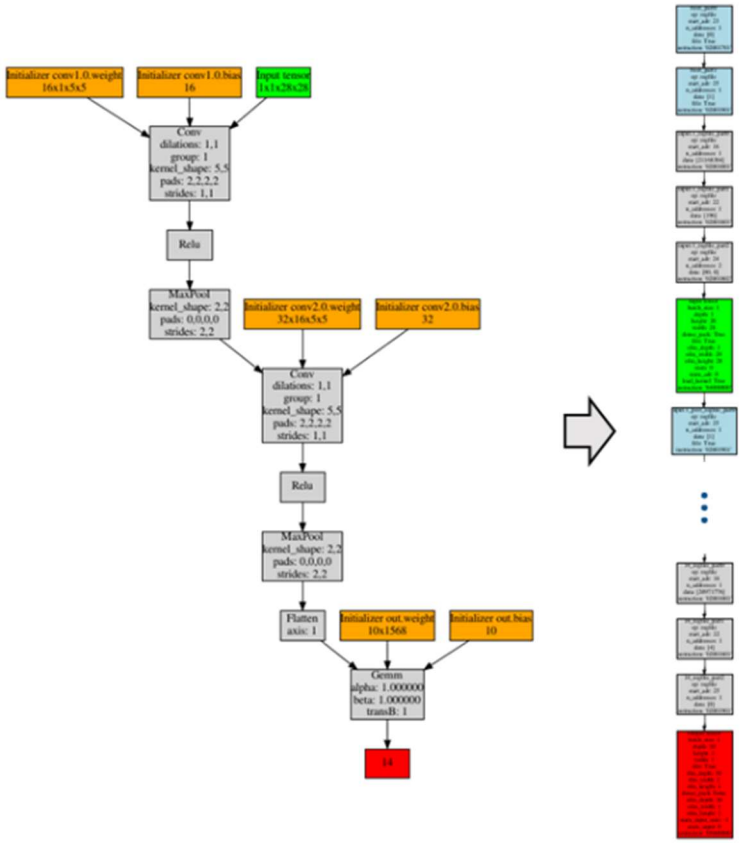


Abbildung 10: Umwandlung des Graphen eines neuronalen Netzes für die Ausführung im Makro

Um nun die korrekte Funktion zu verifizieren wurden Softwaresimulation der neuronalen Netze durchgeführt und die daraus resultierenden Ergebnisse zum Vergleich mit der Hardwaresimulation herangezogen. Dabei mussten auch hardwarespezifische Eigenschaften, wie z.B. eine Quantisierung der Daten berücksichtigt werden.

Die durchgeführten Tests beinhalteten zunächst einfacherer Testfälle, wie z.B. einzelne Faltungs-, oder MaxPooling-Operationen. Um eine möglichst weite Abdeckung des Parameterraumes zu erreichen, wurden dabei verschiedene Parameter, wie z.B. die Größe der Eingabe-Tensoren, Kernel-Größen oder Padding-Konfigurationen zufällig generiert. Das ermöglichte auch das Aufspüren und Beheben unerwarteter Fehler, die sich bei bestimmten Kombinationen von Parametern ergeben konnten.

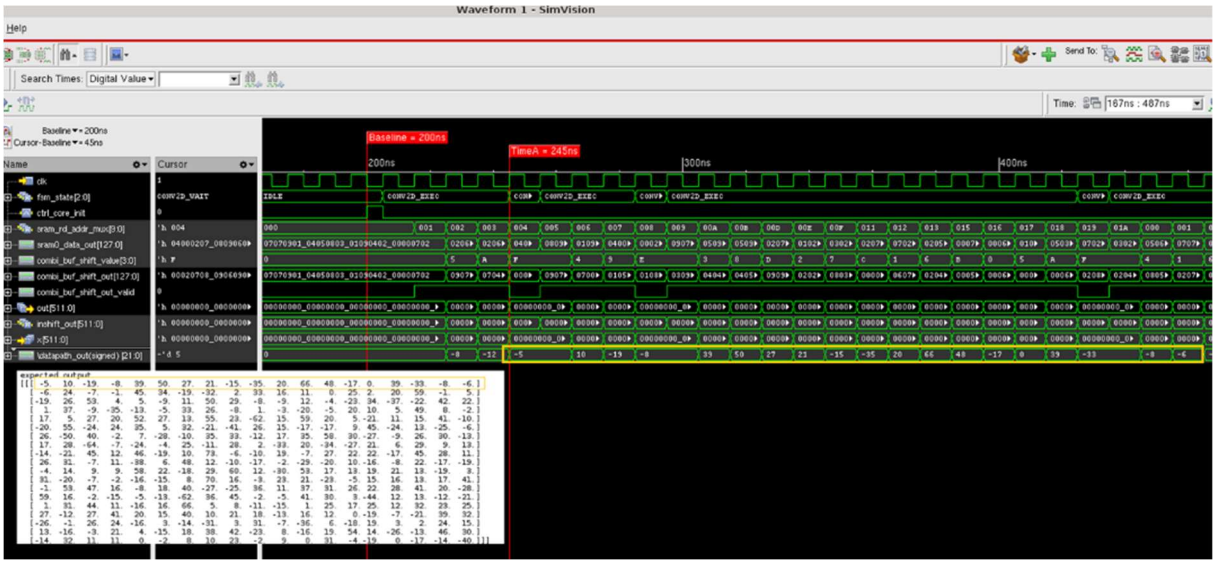


Abbildung 11: Simulation einer Faltungsoperation

Weiterhin wurden auch komplexere Testfälle, wie die Klassifizierung des MNIST-Datensatzes durchgeführt. Dieser beinhaltet handgeschriebene Ziffern, die von einem neuronalen Netz erkannt werden können. In diesem Test werden verschiedene Operationen, wie Faltung, MaxPooling und Aktivierungsfunktionen durchgeführt, wodurch das korrekte Zusammenspiel aller Komponenten verifiziert werden konnte. Als Eingabe-Daten werden die Bilder der handgeschriebenen Ziffern in Graustufen im Format 28x28 Pixel übertragen. Als Ausgabedaten werden nach der vollständigen Bearbeitung des neuronalen Netzes 10 Bytes zurückgegeben, die jeweils die Wahrscheinlichkeit der Erkennung der Ziffern von 0 bis 9 angeben.

Der Entwurf des Makroschaltkreises und die daraus resultierenden Ergebnisse wurden im Arbeitsergebnis D3.7 im Detail berichtet.

Abbildung 12: Beispiel für Eingangsdaten für den MNIST-Testfall

```
########## Compare simulation result ###############
Batch: 0
    abs err                 : 1.769502
    rel err (1-norm)        : 0.146641
    rel err (infinity norm) : 0.095909
    top-1                   : pass      (reported)
    softmax err             : 0.001313
Batch: 1
    abs err                 : 1.925420
    rel err (1-norm)        : 0.088656
    rel err (infinity norm) : 0.123197
    top-1                   : pass      (reported)
    softmax err             : 0.012464
Batch: 2
```

Abbildung 13: Auszug aus dem Verifikationsergebnis des MNIST-Testfalls

3.5 FPGA Implementierung

Ursprünglich war im Arbeitspaket die Fertigung eines Testchips für das entwickelte Speicher-Makro vorgesehen. Da die abgeschätzte Fertigungszeit für FeFET auf 28SLP zum Zeitpunkt des Fertigstellens des Makros auf bis zu 9 Monate geschätzt wurde, wurde die Entscheidung getroffen, zur Risikominimierung, eine Variante des Schaltkreises für FPGA zu portieren um den Beschleunigerteil testen zu können. Diese Änderung war Teil des genehmigten Änderungsantrages auf EU-Ebene. Die Befehle, Kernel-Parameter und Konfigurationseinstellungen sollten dabei von einem bereits gefertigten FeFET-Chip geliefert werden und mit der FPGA über ein PCB verbunden werden. Da der der digitale Beschleunigerschaltkreis einen wichtigen Teil des Makros darstellt, ist eine FPGA Implementierung sehr wertvoll für die Verifizierung, die Fehlerfindung- und Behebung und Demonstration des Systems.

Für die Umsetzung waren verschiedene Arbeiten notwendig. Der Schaltkreis wurde zunächst auf Xilinx Zynq FPGA Plattform portiert. Dazu waren einige Änderungen im RTL-Code notwendig. Da sich die Schnittstelle von der ursprünglichen Anbindung an die FeFET-Speicherfelder unterscheidet, musste eine Schnittstelle zum bestehenden FeFET-Test-Chip implementiert werden.

Die FPGA-Entwicklungsumgebung „Vivado“ stellt auch eine Simulationsumgebung bereit, mit der das angepasste Design anschließend verifiziert wurde.

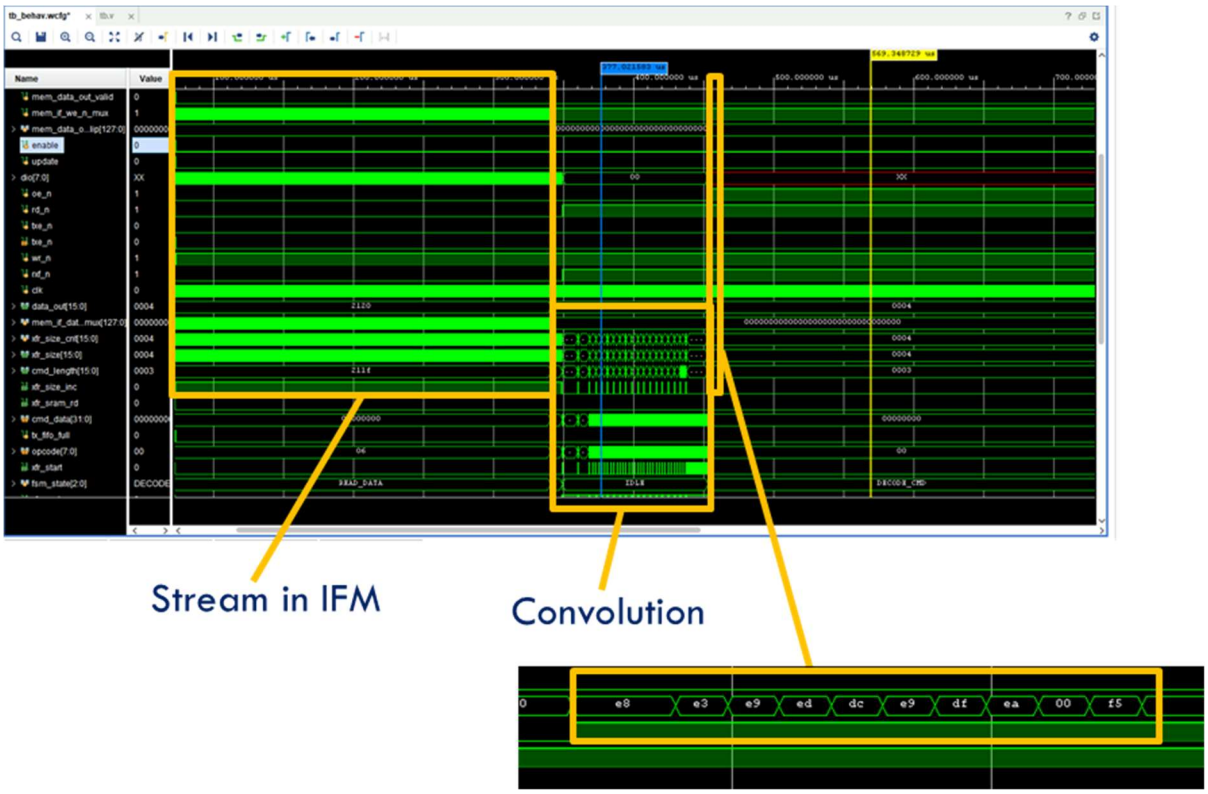


Abbildung 14: Simulation der FPGA-Implementierung

Die Anbindung des FPGAs an den FeFET-Testchip erfolgte mittels einem eigens entwickeltem PCB. Weitere Komponenten, wie Digital-Analog-Wandler zur Versorgung des FeFET-Testchips und Stecker, sowie ein USB-zu-FIFO-Umsetzer mussten in dem Entwurf berücksichtigt werden. Das PCB konnte dann mit einer USB-Schnittstelle an einen PC angeschlossen werden.

Um die Funktionalität der FPGA-Implementierung zu demonstrieren, wurde ein Software-Programm entwickelt, welches die Ein- und Ausgabedaten per USB-Schnittstelle übermittelt. Dafür konnte eine Programmierbibliothek für die Programmiersprache Python des USB-zu-FIFO-Chipherstellers genutzt werden. Zur Validierung der Echtzeitfähigkeit des Systems wurde ein Video-Stream der eingebauten Kamera des PCs als Eingabe genutzt.

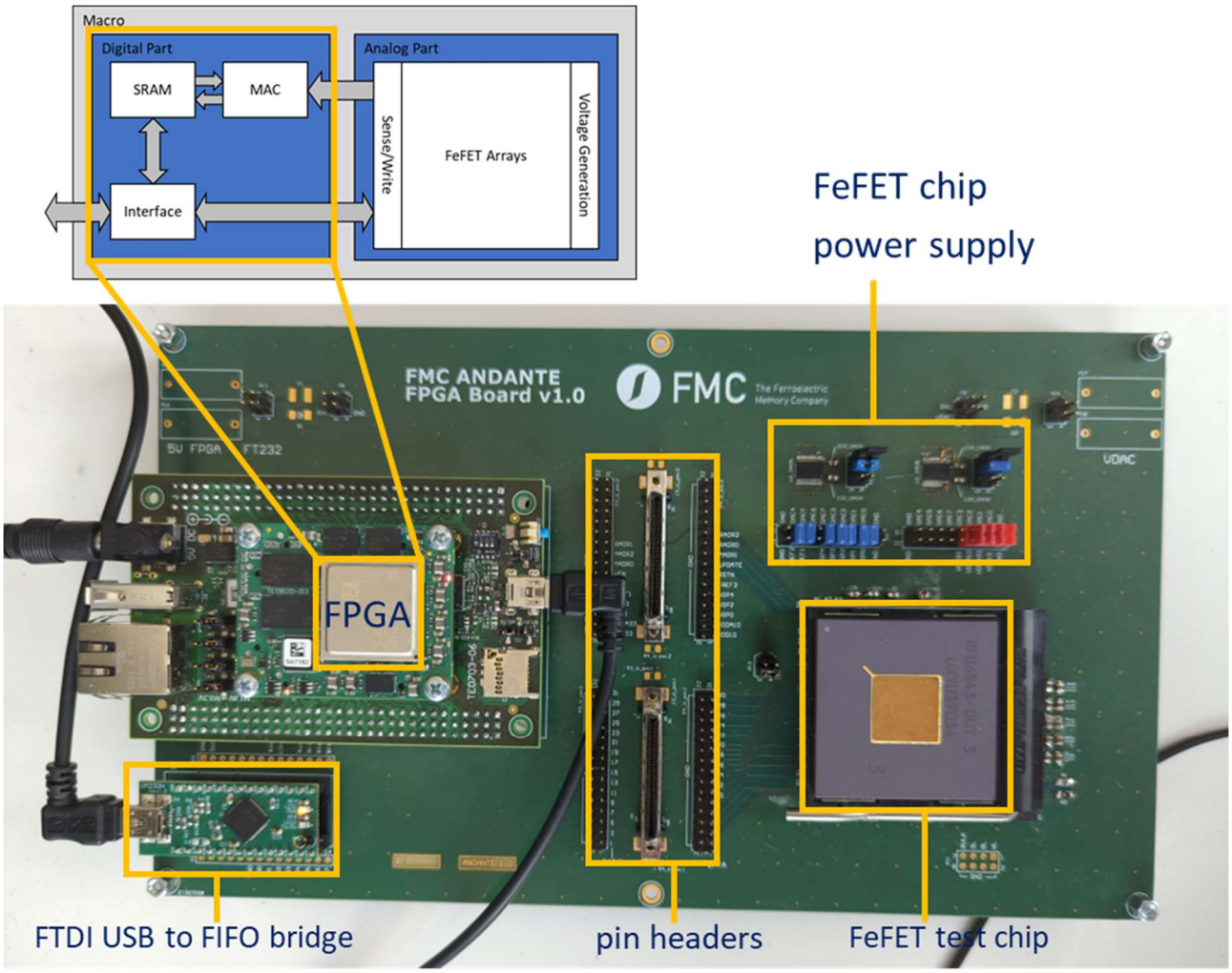


Abbildung 15: FPGA-PCB

Zur Veranschaulichung wurde ein Sobel-Filter zur Kantenerkennung getestet. Dabei werden Faltungsoperationen durchgeführt, um bestimmte Merkmale aus Bildern zu extrahieren. Dazu wurde ein Ausschnitt des Kamerabildes kontinuierlich zur Verarbeitung an das FPGA geschickt. Das Ergebnis der Filterung wurde dann zurückgelesen und mit dem ursprünglichen Bild überlagert. In unserem Beispiel werden horizontale Linien verstärkt. Wie in Abbildung Abbildung 12 zu erkennen ist, werden im linken Bild die Waagerechten Linien der Finger verstärkt (schwarz), während sie bei senkrechter Haltung verschwinden. Diese Art von Filterung ist Bestandteil vieler neuronaler Netzwerkmodelle zur Erkennung und Klassifizierung von Objekten. Dieses Beispiel eines Anwendungsfalls stellt aber nur einen kleinen Teil der Möglichkeiten des Systems dar.

Alle Daten zur Ausführung des Netzwerkmodells, also Befehle, Konfigurationseinstellungen und Kernel-Parameter kamen dafür aus dem FeFET-Testchip an das FPGA angeschlossenen FeFET-Testchip. Mit dem FPGA-PCB-Aufbau konnte die Funktionalität des Schaltkreisentwurfs verifiziert und eine Echtzeitfähigkeit bei der Ausführung von Faltungsnetzwerken gezeigt werden.

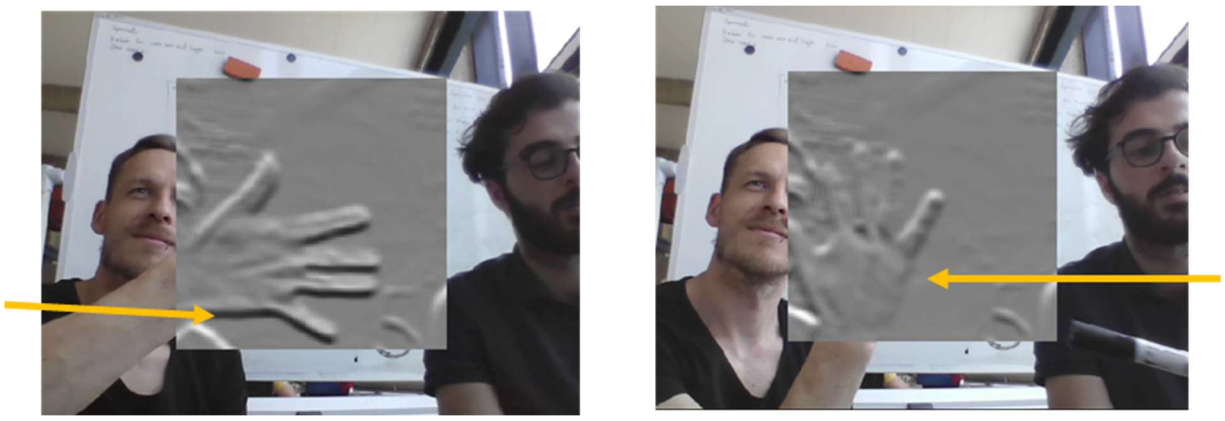


Abbildung 16: Demonstration des FPGA-Systems am Beispiel Kantenerkennung

4 Positionen des zahlenmäßigen Nachweises

In der Gesamtvorkalkulation wurden Selbstkosten in Höhe von 1.168T EUR ermittelt, aufgrund der ursprünglich geplanten Zusammenarbeit mit Global Foundries. Ursprünglich sollte die genutzte Fe-FET Technologie bei dem Fertigungspartner hergestellt werden. Allerdings wurden Fertigungszeiten von bis zu 9 Monaten angekündigt. Um die Zielsetzung sowie den Projektzeitraum einzuhalten, musste FMC reagieren. Es wurde beantragt das ASIC Design des Schaltkreises und dem damit verbundenen Tapeout durch ein FPGA Setup zu ersetzen. Somit wurde ein FPGA-Design hauseigens entwickelt, im Testlabor getestet und in Betrieb genommen. Das führte jedoch zu dem Ergebnis, dass deutlich mehr Projektmonate und damit Personalkosten auf dem Projekt aufgelaufen sind, als ursprünglich geplant. Mit den höheren Projektdesignstunden gehen auch höhere Kosten für EDA Tools von Cadence einher. Für die Inbetriebnahme der FPGA-Plattform wurde eine Leiterplatte (PCB) extern entwickelt. Somit sind Kosten in der Kategorie Fremdleistungen angefallen. Die Materialkosten sind geringer gegenüber der Vorkalkulation aufgrund der Eigenaktivitäten im Testlabor und Nutzung der bereits vorhandenen Materialien. Es gab keine Reiskosten während des Projektzeitraumes, aufgrund der Pandemie.

Das Projekt weist insgesamt Kosten in Höhe von 1.590T EUR auf. Folglich ist der Eigenanteil höher als in der Vorkalkulation. Dennoch konnten alle Projektziele und Meilensteine erfolgreich umgesetzt werden.

		Gesamt-vorkalkulation	Kosten 2020	Kosten 2021	Kosten 2022	Kosten 2023	Gesamt-nachkalkulation
0813	**Materialkosten**	39.679,00				849,22	849,22
0823	**Fremdleistung**	0				2.775,00	2.775,00
0837	**Personalkosten**	631.052,00	49.847,46	220.986,83	672.038,30	76.700,03	1.019.572,62
0838	**Reise**	19.750,00					
0847	**Vorhabensspezifische Afa**	19.392,00					
0850	**Sonstige Kosten**	459.000,00	120.000,00	232.230,00	215.000,00		567.230,00
0881	**Selbstkosten**	**1.168.873,00**	**169.847,46**	**453.216,83**	**887.038,30**	**83.948,47**	**1.590.426,84**
0882	**Eigenmittel**	**467.507,00**					**889.060,84**

Abbildung 17: Vergleich Vor- und Nachkalkulation des Projektes

5 Notwendigkeit und Angemessenheit der geleisteten Projektarbeiten

Um das Ziel der Erstellung eines FeFET-basierten Speichermakros für KI-Anwendungen in Arbeitspaket 3 zu erreichen, waren gewisse Schritte und Arbeiten notwendig. Speziell im Schaltkreisentwurf gibt es eine Reihe verschiedener Entwurfsschritte, die in Abhängigkeit zueinanderstehen. So bedarf es eines Architekturentwurfes am Anfang des Entwicklungsprozesses, um überhaupt die zu entwickelnden Komponenten identifizieren zu können. Darauf aufbauend kann ein Schaltkreislayout, also das Endprodukt des Entwurfes, welches schließlich zur Fertigung von Chips benötigt wird, nur entstehen, wenn es ein entsprechender Schaltplan vorliegt. Um zu überprüfen, dass das Layout vorgegebene Entwurfsregeln einhält und mit dem Schaltplan übereinstimmt, muss eine physikalische Verifikation durchgeführt werden. Auch die durchgeführten Simulationen sind für die Verifikation der Funktionalität des Schaltkreises notwendig.

Für das Design des FeFET-Speicherfeldes waren die Arbeiten in Arbeitspaket 2 essenziell. Zum einen wurde das erstellte PDK mit all seinen Komponenten, wie Layout-Zellen, Schematic-Symbolen und Entwurfsregeln, also DRC- und LVS-Checks benötigt, um die FeFET-Zellen in den Schaltkreis integrieren zu können. Zum anderen gaben die Messergebnisse Aufschluss über die zu erwartenden Parameter, die wiederum in den Entwurf der Leseverstärkerschaltungen eingegangen sind.

Durch die Änderung der Zielimplementierung in einem FPGA-Aufbau ließ sich der entworfene Digitalteil im realen Betrieb testen. Dabei wurde festgestellt, dass die implementierte Schnittstelle nicht im ersten Entwurf mit dem USB-zu-FIFO-Chip ordnungsgemäß kommunizieren konnte. Das war einer unzureichenden Dokumentation hinsichtlich des Herstellers zuzuschreiben. Mithilfe der FPGA-Implementierung konnte dieser Fehler leicht behoben werden, was mit einem Chipdesign nicht mehr möglich gewesen wäre. Weiterhin ließ sich mit diesem System auch die Funktionalität verifizieren.

6 Nutzen und Verwertbarkeit des Ergebnisses

Die im Arbeitspaket 2 gewonnenen Ergebnisse haben erheblich dazu beigetragen, Entscheidungen im Herstellungsprozess treffen zu können, um die Eigenschaften der Speicherzellen zu verbessern. Die durchgeführten Experimente haben zudem essenzielle Informationen für die besten Betriebsbedingungen für große Speicherfelder geliefert. Zudem wurden tiefgehende Erkenntnisse für Design-Methoden und Algorithmen für den Betrieb von FeFET-basierten Speichern gewonnen.

Das ANDANTE-Projekt hat weiterhin zu einem wertvollen Erkenntnisgewinn beigetragen. Dabei wurde FMC durch die Arbeiten in Arbeitspaket 3 in die Lage versetzt, Speicherschaltkreise zu entwickeln, die den Anforderungen von Edge-AI-Systemen genügen. Darüber hinaus wurde Know-how in der Hardware-Umsetzung von Operationen für neuronale Netze, wie z.B. Faltung, Max-Pooling oder Aktivierungsfunktionen gewonnen, sowie umfangreiches Wissen über verschiedene Netzwerkmodelle und die Verarbeitung neuronaler Netze allgemein angeeignet.

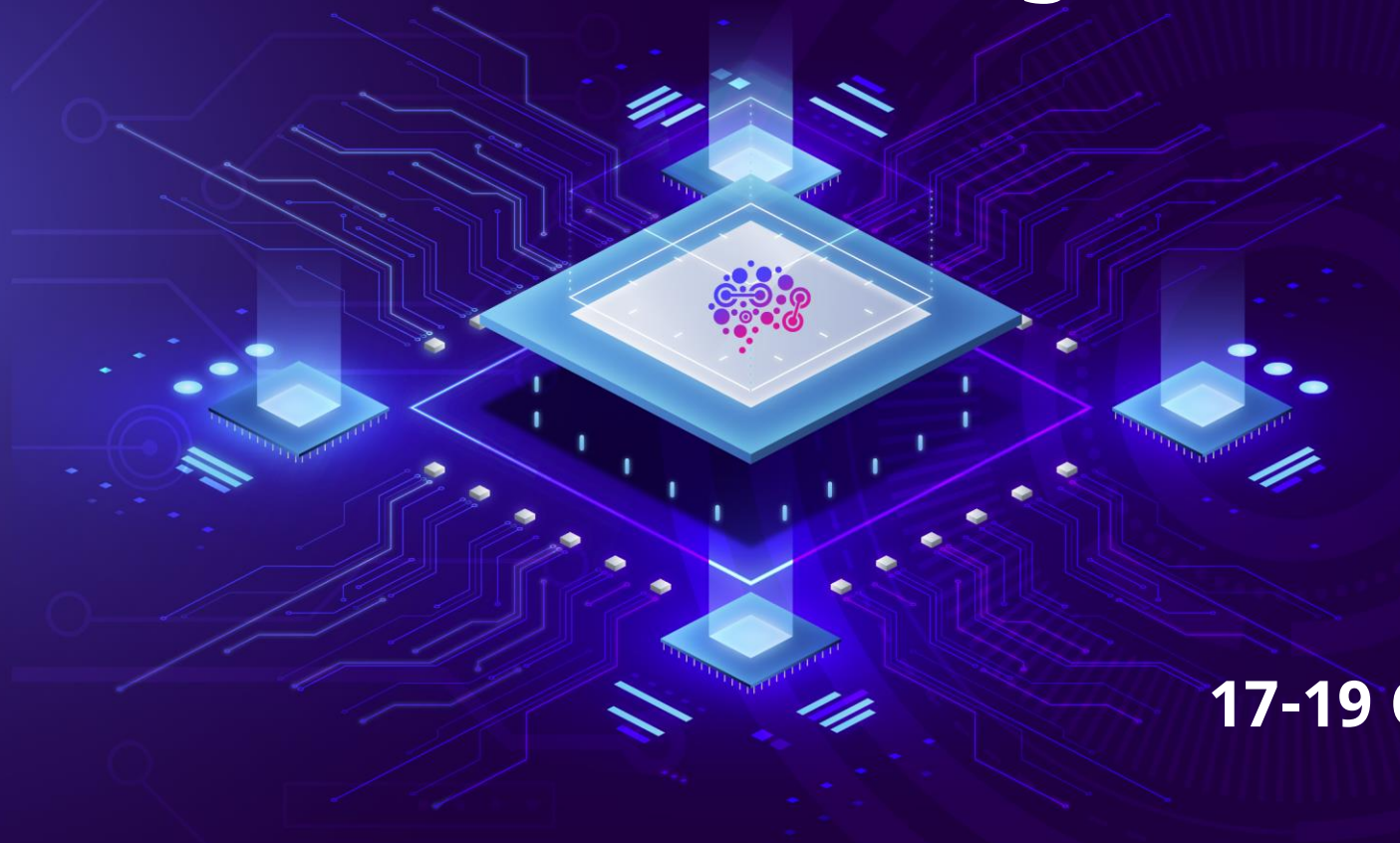
Es wurden Kompetenzen zur Implementierung und der Simulation von Schaltkreisen auf FPGA-Plattformen entwickelt, die in zukünftigen Arbeiten nützlich sein werden.

Trotz den erreichten Fortschritten in der Analyse der FeFET-Zellen und der gewonnenen Erkenntnisse zu Lösch- und Programmieralgorithmen und deren Parameter ist die FeFET-Entwicklung nicht weit genug fortgeschritten. Besonders die erreichten Bitfehlerraten für produktrelevante Speicherfeldgrößen genügen nicht den Ansprüchen für ein Produkt. Zudem sind konkurrierende Speicherlösungen auch in der Entwicklung weiter vorangeschritten und wie im Fall von MRAM oder RRAM auch bei verschiedenen Auftragsfertigern in Produktion. Es ist somit eine weitere Verbesserung der Wettbewerbs-Situation, insbesondere hinsichtlich Zellgröße, bzw. -effizienz und den damit verbundenen Chipkosten, notwendig. Eine direkte Verwertung der im Projekt erzielten Ergebnisse muss daher aufgeschoben werden.

7 Veröffentlichungen

Das in Arbeitspaket 3 entwickelte FeFET-basierte Speichermakro mit Beschleuniger für Faltungsnetzwerke, sowie die Umsetzung des Digitalteils in der FPGA-Plattform wurden in einem Beitrag mit dem Titel „A FeFET-based non-volatile memory and AI accelerator macro" auf der Edge-AI-Konferenz 2023, die in Athen ausgetragen wurde, vorgestellt.

European Conference on EDGE AI Technologies and Applications - EEAI



Advancing emerging edge AI technologies and driving next generation intelligent applications.

17-19 October 2023, Athens, Greece

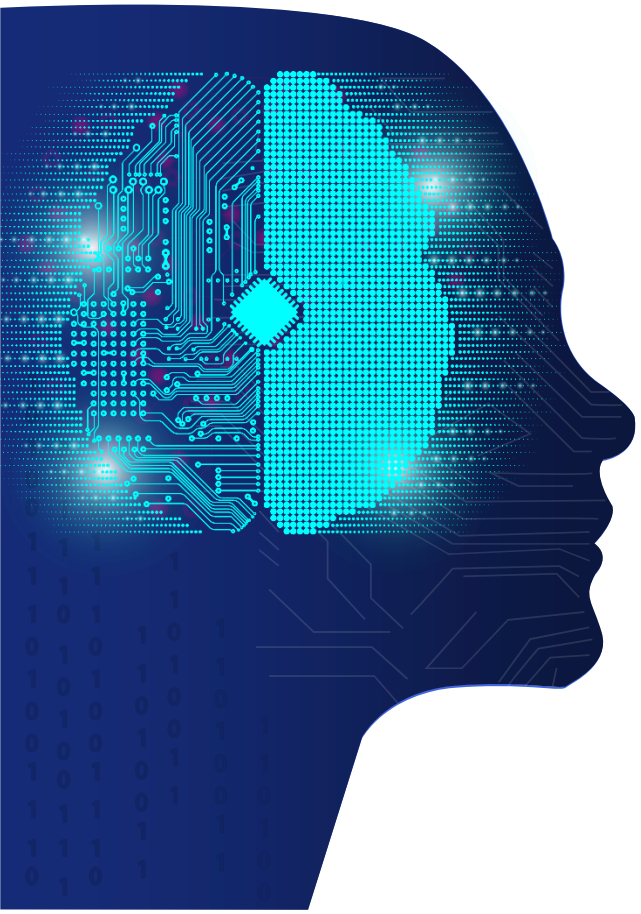
European Conference on EDGE AI Technologies and Applications - EEAI

A FeFET-based non-volatile memory and AI accelerator macro

Marko Noack, Ferroelectric Memory GmbH, Germany

17-19 October 2023 Athens, Greece

Presentation Outline



- Introduction
- Goal
- Methodology
- Design and Implementation
- Results
- Future Research
- Conclusion

Introduction

FMC The Ferroelectric Memory Company

- Edge AI offers
  - Reduced latency
  - Independency of network connections
  - Scalability
  - Enhanced Security
- But we need solutions with
  - Low power
  - Low cost
  - Non-volatile storage of network weights
- Our goal
  - Demonstrate viability of FeFET memory in Edge AI systems

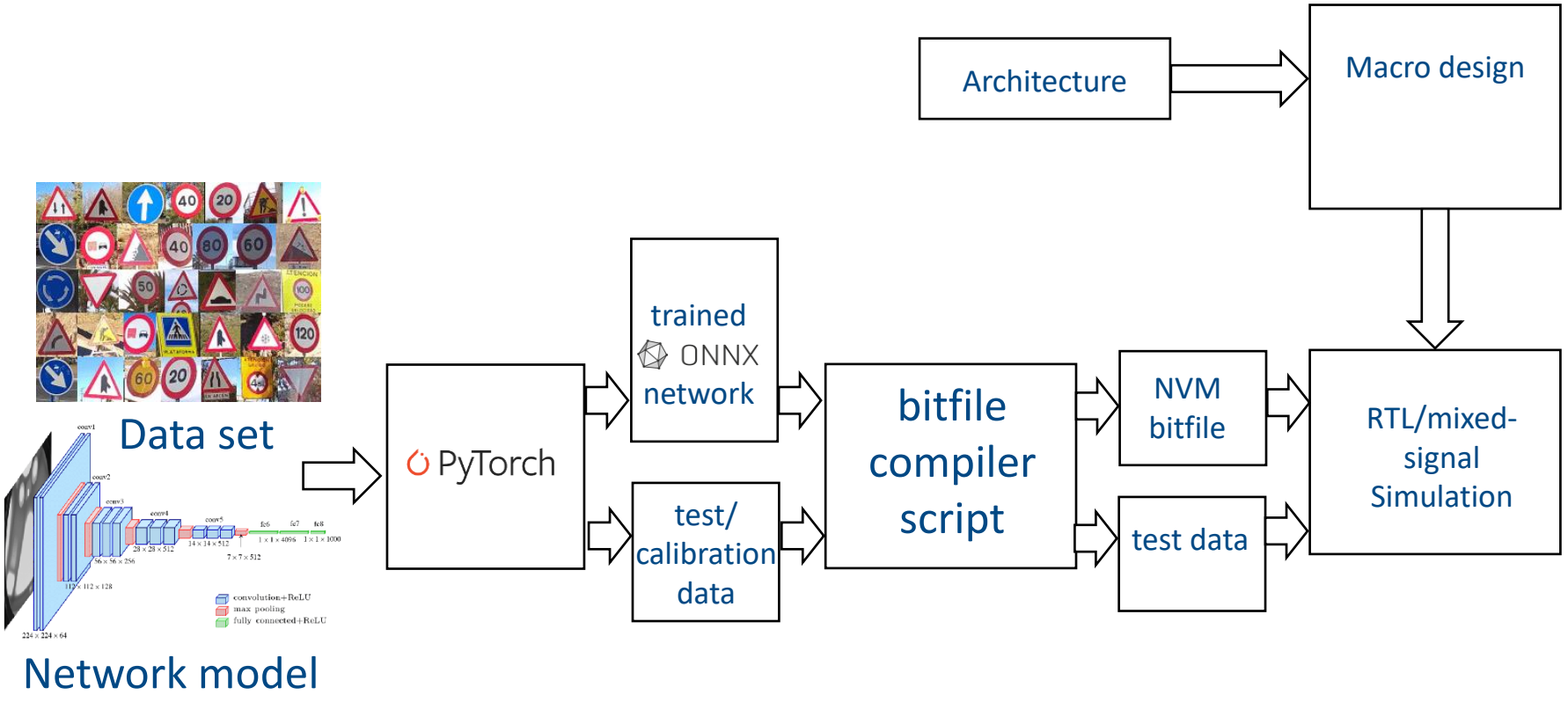
Goal

FMC The Ferroelectric Memory Company

- Development of a FeFET-based NVM macro
  - Capable of data storage, but also with
  - Built-in accelerator circuit for CNNs
  - Maximize speed and utilization
  - Maintain Flexibility for different kinds of network models

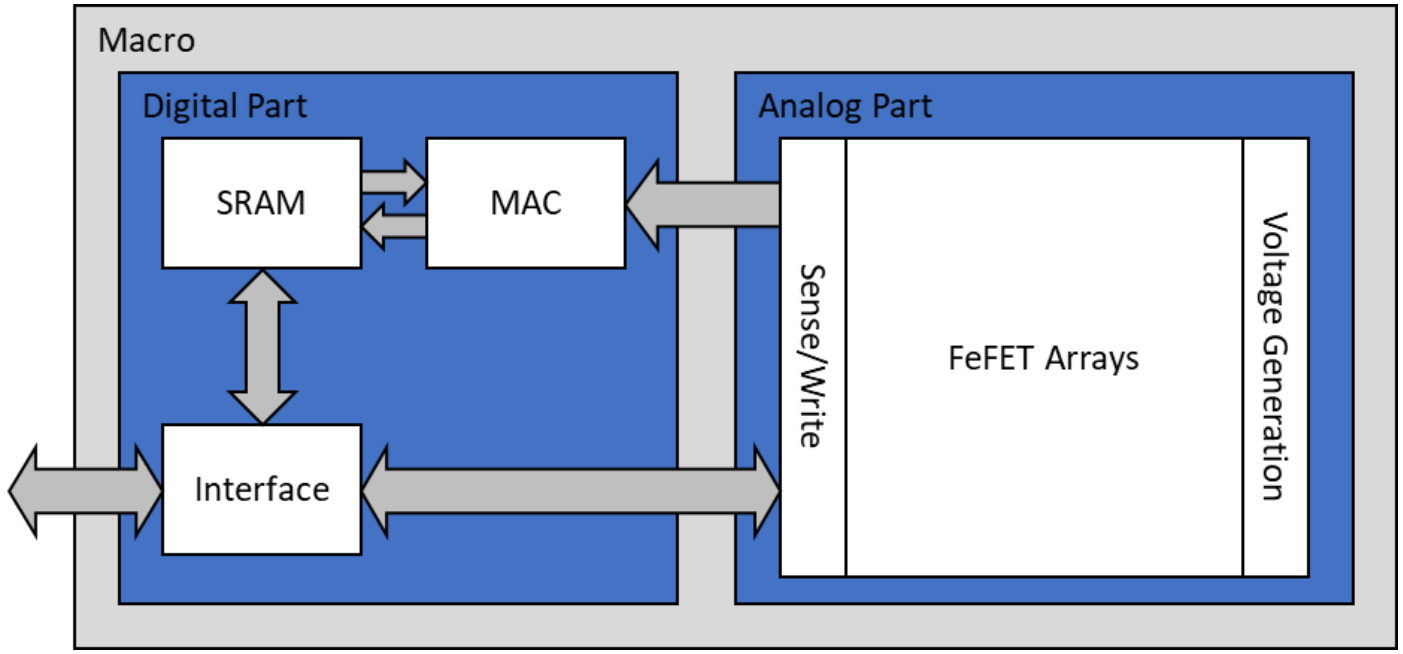
Methodology

- Implement macro design
- Train network model with data set
- Generate NVM bitfile and testdata from trained network and test data



Design and Implementation

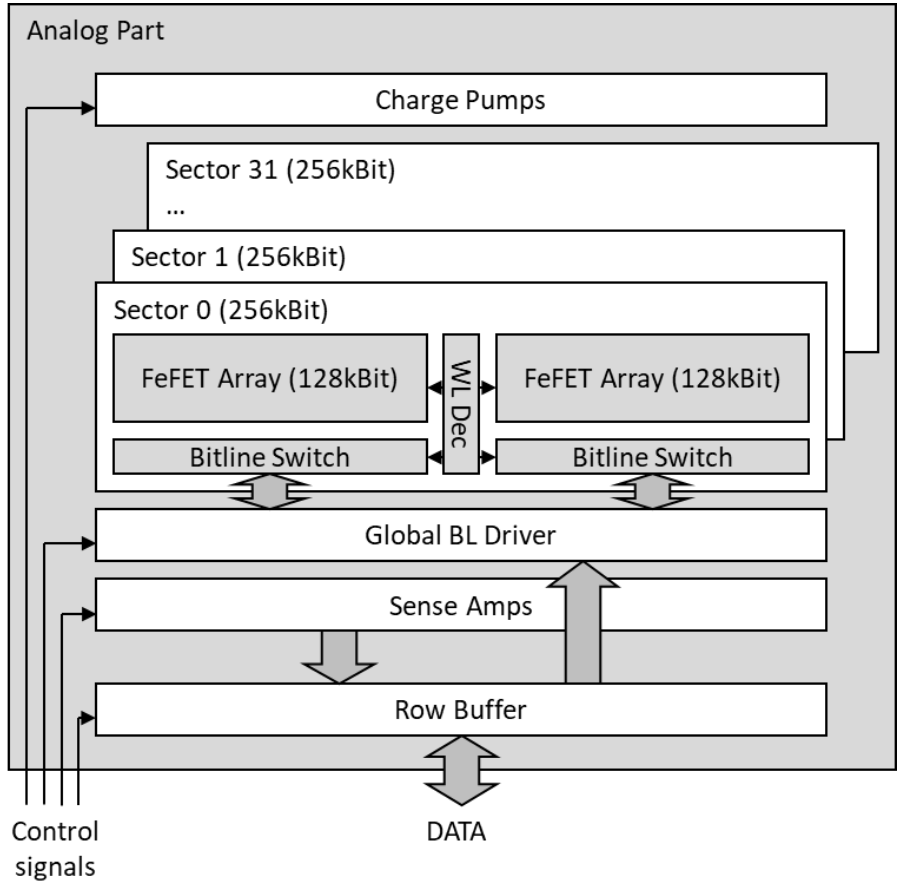
- Design split in full-custom analog and RTL-based digital part
- Analog part contains:
  - memory arrays
  - charge pumps
  - Sense amps
- Digital part contains
  - Interface
  - SRAM to store IFMs/OFMs
  - MAC unit
  - State machines



FMC The Ferroelectric Memory Company

Analog Part

- FeFET Arrays
  - 8 Mbit organized in 32 sectors
  - Stores commands, configuration settings and kernel parameters
  - 1024 bits read at a time and multiplexed out with row buffer while next sensing operation can already be performed

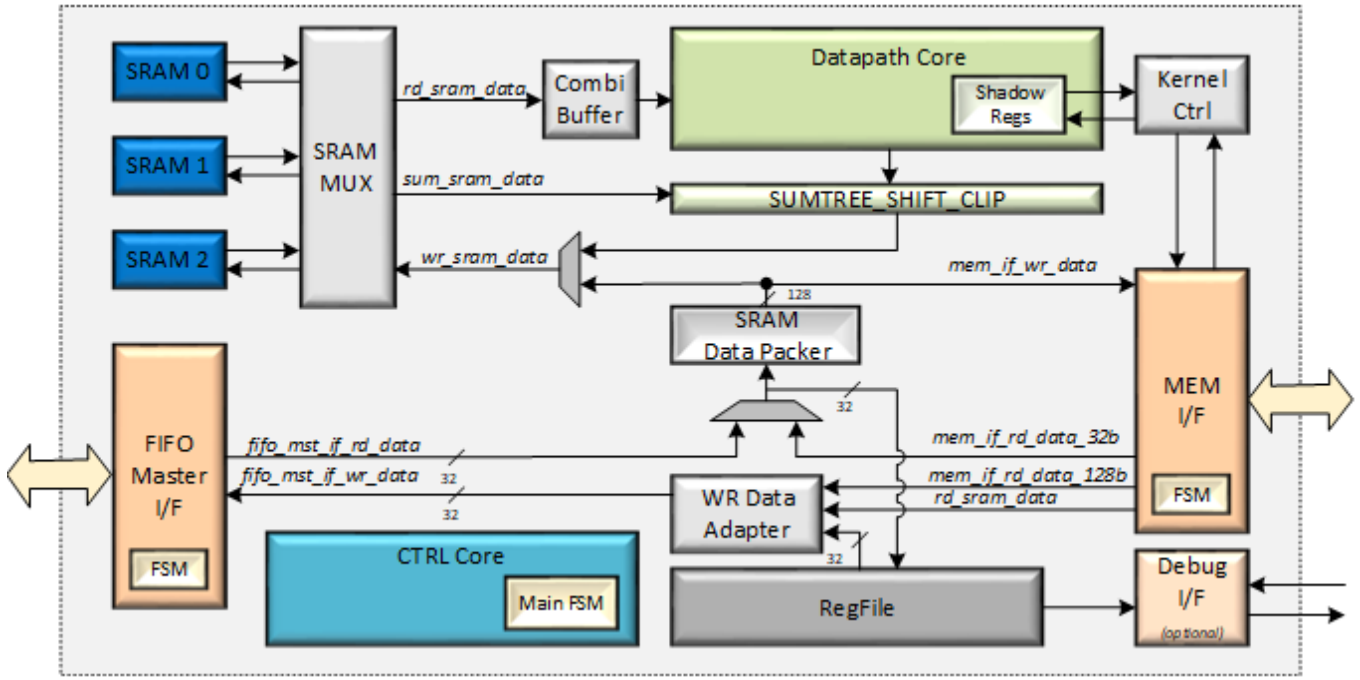


Design and Implementation

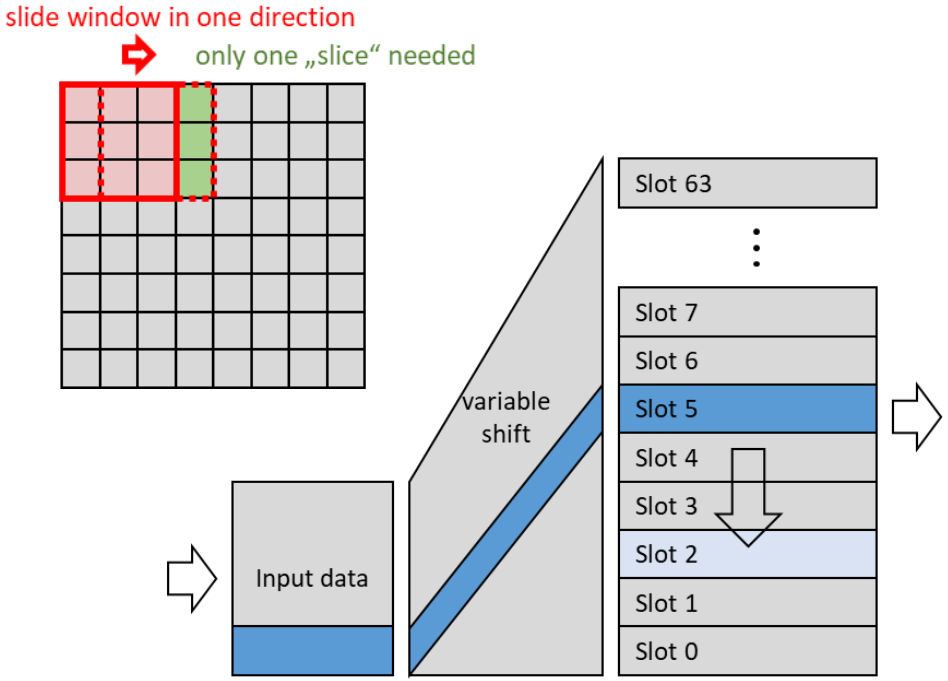
FMC The Ferroelectric Memory Company

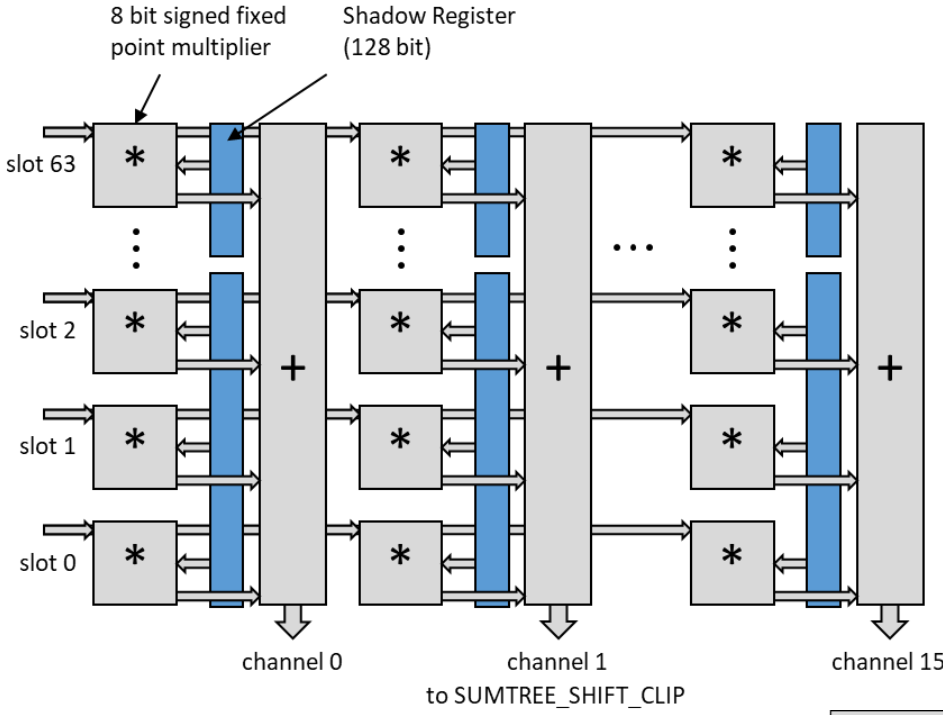
Digital Part

- Input feature map is streamed in to one of the SRAMs via FIFO interface
- Convolution/MaxPool
- Operations performed layer by layer
- Layers can be split into parts, partial sums can be added through second SRAM
- Output feature map stored in SRAM → streamed out via FIFO interface

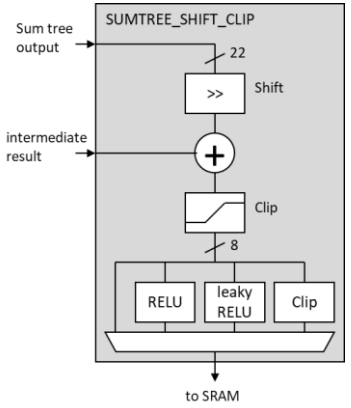


Design and Implementation





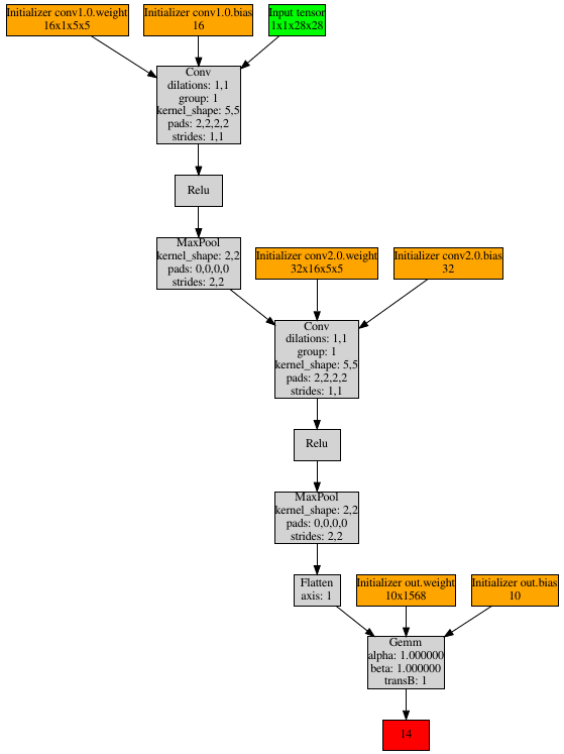
- For convolution operation a sliding window over the IFM is generated
- Only one input „slice“ is needed per clock cycle
- Input buffer arranges IFM data and sends it to multiplier array
- IFM data shared by row
- Each multiplier stores kernel weight in shadow register
- 16 colums/channels generates OFM data which fits again in SRAM width
- Clip/Relu function can be applied to OFM directly



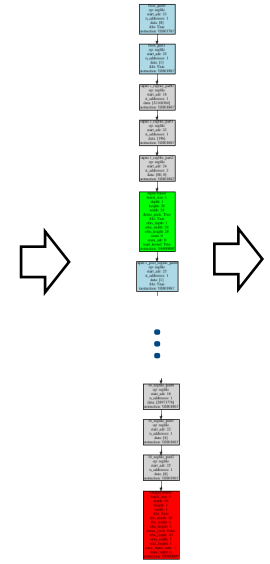
Simulation Results

- 100ns FeFET read spead (1024 bits)
- Multiplexed to 16 x 128-bit words @ 100MHz
- Up to 0.1 TOPs (signed 8-bit fixed-point multiplications)

MNIST used as input and training data set



network graph

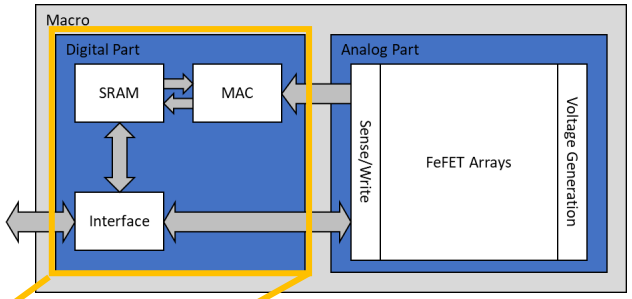


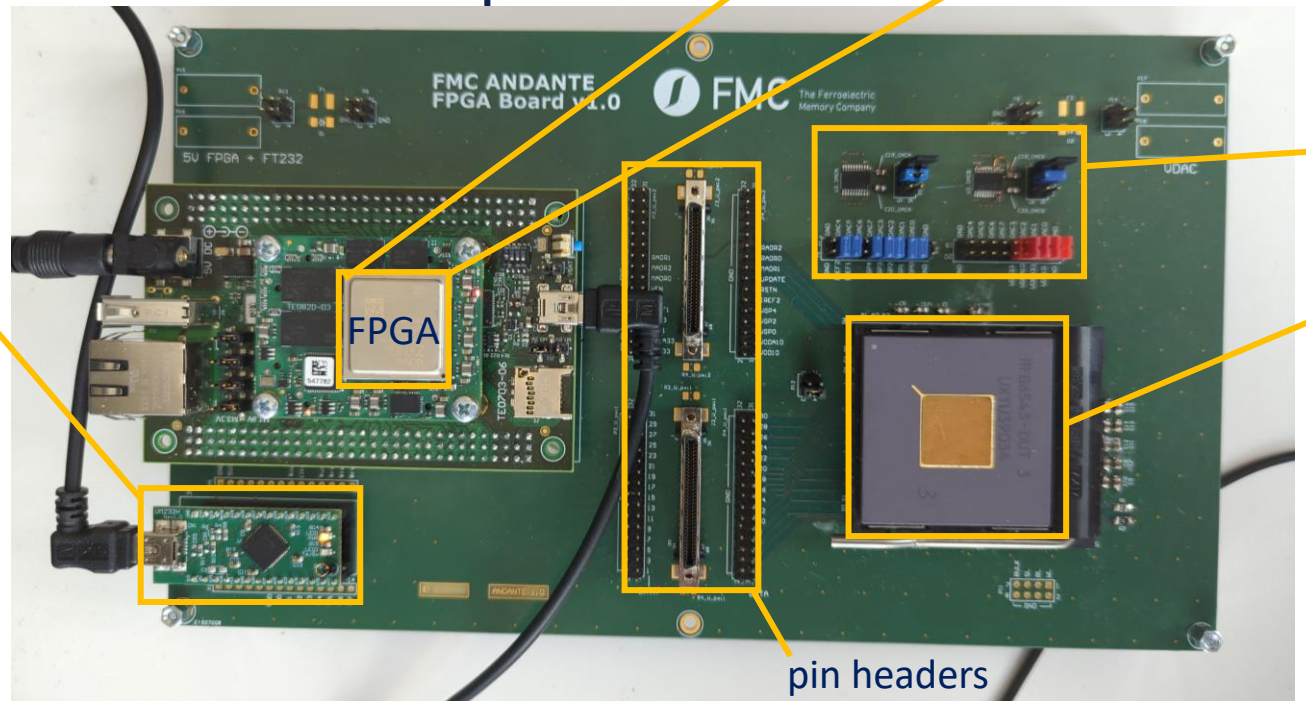
translated to list of operations

Simulation verified correct classification results

Experimental Results

- Digital part has been implemented on a Xilinx FPGA for testing
- FeFET memory data comes from an external test chip





FTDI USB to FIFO bridge

FeFET chip power supply

FeFET test chip

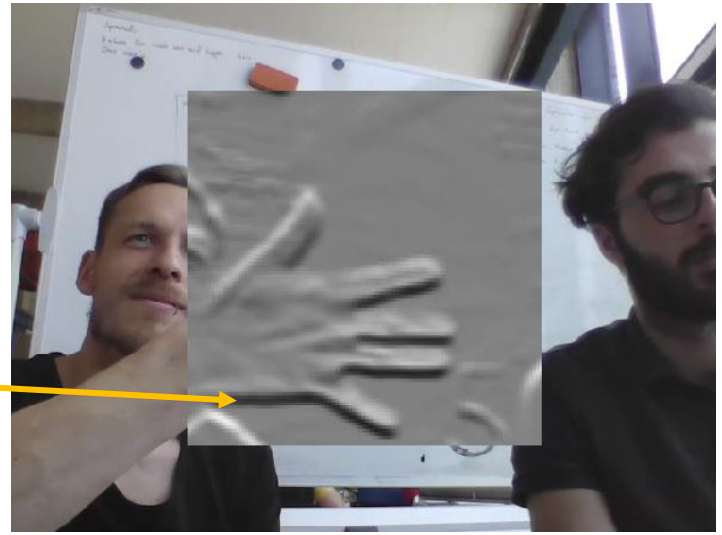
pin headers

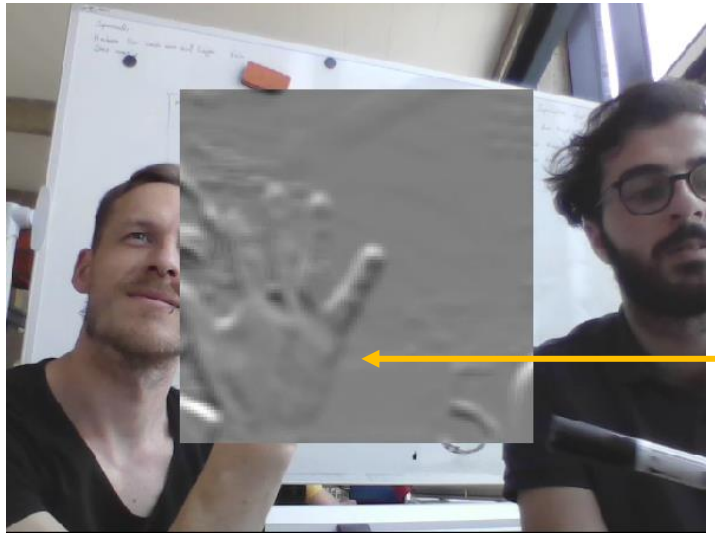
Experimental Results

FMC The Ferroelectric Memory Company

- Limitation in FPGA setup allowed for small network only
- Example: Sobel edge detection filter to detect horizontal features

strong horizontal features





no horizontal features

Future Research

- Looking for future cooperations
- Explore possible optimizations

Conclusion

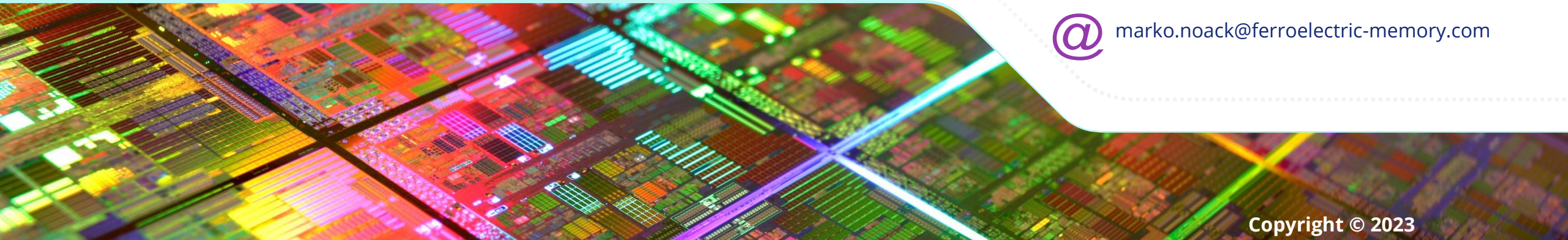
FMC The Ferroelectric Memory Company

- FeFET-based NVM very suitable for Edge AI applications
- Low cost (low mask count, easy CMOS integration)
- With the presented architecture, high throughputs can be achieved while maintaining high flexibility

EDGE AI

Thank You

For your attention



@ marko.noack@ferroelectric-memory.com