Hardwareimplementierung Neuronaler Netze

U. Rückert

Technische Universität Hamburg-Harburg
Technische Elektronik
Hamburg

Einleitung

Neuronale Netze bedienen sich intern einer massiv parallelen Informationsverarbeitung. Im Gegensatz dazu ist die derzeit häufigste Implementierungsvariante die relativ zeitaufwendige Simulation auf konventionellen, weitgehend seriellen Rechnern, wie z.B. Personalcomputern (PCs), Workstations oder auch Vektor-Rechnern. Dies führt für die Simulation künstlicher neuronaler Netze (KNN) bei realistischer Netzgröße zu Rechenzeiten, die für eine Echtzeitverarbeitung inakzeptabel sind und eine Untersuchung des dynamischen Verhaltens neuronaler Netze sowie die Optimierung ihrer Topologie unmöglich machen. Aus diesem Grunde kann der Nutzen von KNN heute nur für genügend kleine Anwendungen festgestellt oder durch Simulationsexperimente mit miniaturisierten Szenarien von anspruchsvollen Anwendungen (z.B. Bilderkennung, Sprachverarbeitung) abgeschätzt werden. Von der Anwenderforschung wird daher immer wieder der Wunsch nach effizienteren Realisierungsmöglichkeiten geäußert. Die gewünschte Erhöhung der Rechenleistung ist aber nicht primär durch den Technologiefortschritt in der Mikroelektronik und die dadurch bedingte jährliche Verdoppelung der Rechnerleistung konventioneller Rechner zu erreichen. Ferner sind neue architektonische Lösungen mit massiver Zeit- und Raumparallelisierung bei ökonomischem Umgang mit der Verlustleistung gefordert. Einen Beitrag hierzu leisten sogenannte *Neurocomputer* und *Neuro-Chips* , auf denen KNN entsprechend dem Stand der Technik effizient implementiert werden können.

In der einschlägigen Literatur finden sich inzwischen eine Vielzahl von unterschiedlichen Vorschlägen für die Hardwareimplementierung neuronaler Architekturen, von denen sich der überwiegende Teil allerdings noch in der Entwicklungs- bzw. Testphase befindet [1-8]. Die unterschiedlichen Ansätze lassen sich grob in zwei Gruppen unterteilen (Bild 1). Zum einen in Architekturen auf der Basis von handelsüblichen Standard-VLSI-Bausteinen (VLSI=Very Large Scale Integration), die sich weiter in Zusatzkarten (Add-On-Boards) für konventionelle Arbeitsplatzrechner (PC, Workstation, etc.) und spezielle Parallelrechnersysteme aufteilen. Zum anderen in Architekturen auf der Basis von anwendungsspezifischen VLSI-Bausteinen (ASICs, application specific integrated circuits), die in digitaler oder analoger Schaltungstechnik realisiert sein können. Der Vollständigkeit halber sei erwähnt, daß in Zukunft neben den hier genannten rein mikroelektronischen Techniken sicherlich auch opto-elektronische bzw. molekular-elektronische Realisierungstechniken interessante Perspektiven bieten werden. In naher Zukunft spielen sie für Anwendungen der KNN noch keine Rolle und werden daher hier auch nicht näher erläutert.

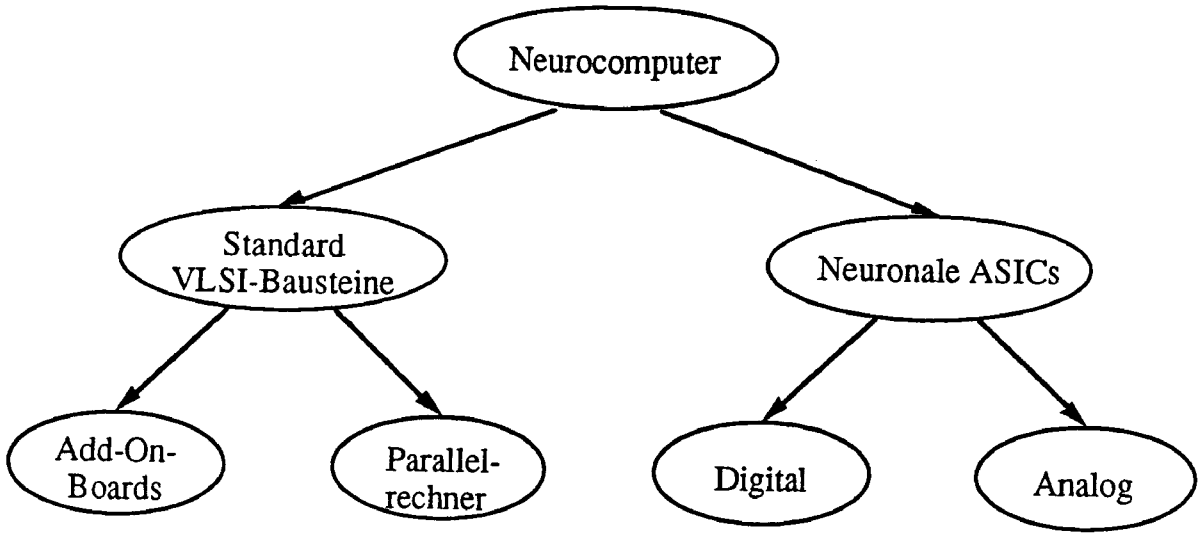

Bild 1 Realisierungsalternativen für Neurocomputer

Die folgenden Abschnitte enthalten eine Übersicht der aktuellen Ansätze der Implementierung von KNN in Hardware entsprechend der in Bild 1 gezeigten Gliederung. Aufgrund der Vielzahl von bekannten Realisierungsvorschlägen für Neurocomputer und Neuro-Chips liegt der Schwerpunkt dieser Übersicht auf ausgewählten Realisierungen, die entweder bereits kommerziell erhältlich sind oder ein breites internationales Interesse erweckt haben. Die Auswahl ist sicherlich subjektiv und erhebt nicht den Anspruch repräsentativ zu sein.

Vergleichskriterien

Für die Leistungsbeurteilung einer neuronalen Netzwerkimplementierung haben sich derzeit noch keine standardisierten Vergleichsverfahren (benchmarks) etabliert. Damit aber dennoch eine grobe Aussage über die Leistungsfähigkeit einer bestimmten Realisierung gemacht werden kann, finden sich in der Literatur häufig die Maßzahlen *Gewichte pro Sekunde* ((inter-) connections per second, *CPS* (bzw. IPS)) für die Anwendungsphase und *Gewichtsänderungen pro Sekunde* ((inter-)connection updates per second, *CUPS* (bzw. IUPS)) für die Lern- bzw. Trainingsphase eines KNN. Für die Anwendungsphase ist mit CPS gemeint, wie häufig deren zentrale Rechenoperation, meistens die Multiplikation eines Eingabewertes x_i mit dem zugeordneten Gewicht w_{ij} (x_i*w_{ij}) und Addition des Ergebnisses zum Aktivierungswert S ($S:=S+x_i*w_{ij}$), pro Sekunde ausgeführt werden kann (multiply&adds per second). Entsprechendes gilt für CUPS in der Lernphase, wobei hier zusätzliche Rechenoperationen gemäß dem verwendeten Lerngesetz durchgeführt werden müssen. Die Angabe 1MCPS (MegaCPS) bedeutet somit, daß eine Million Multiplikationen mit anschließender Addition pro Sekunde durchgeführt werden können; d.h. ein einschichtiges Netz mit 1000 Prozessorelementen (PE, künstliche Neuronen) mit jeweils 1000 Gewichten kann die Multiplikation eines Eingabevektors (1000 Komponenten) mit der Gewichtsmatrix in einer Sekunde durchführen.

Neben der Anzahl und der Art der Rechenoperationen ist die geforderte bzw. implementierte Rechengenauigkeit beim Vergleich zu berücksichtigen. Hier finden sich Modelle mit binären, diskreten (ganzzahligen) und kontinuierlichen (reellen) Ein-/ Ausgaben bzw. Gewichtswerten.

Sowohl die Aktivierungsfunktionen als auch die Lernfunktionen dieser Modelle unterscheiden sich zum Teil erheblich hinsichtlich des erforderlichen Rechenaufwandes. Für Modelle mit binären Ein-/ Ausgaben entfällt z.B. die zeit- bzw. flächenintensive Multiplikation bei der Berechnung des Aktivierungswertes (s.o.). Hier wird deutlich, daß die Leistungsangaben CPS bzw. CUPS erst im Kontext eines bestimmten neuronalen Netzwerkmodells und dessen Anwendung aussagekräftig werden.

Weitere, nicht minder wichtige Aspekte bei der Beurteilung einer bestimmten Realisierung eines neuronalen Netzes sind die Flexibilität und die Skalierbarkeit einer Implementierung. Eine Realisierung ist flexibel, wenn unterschiedliche neuronale Netzwerkmodelle implementiert und deren Parameter verändert werden können. Sie ist skalierbar, wenn die Erweiterung der Hardwarearchitektur (z.B. Anzahl der Prozessoren, Größe des Gewichtsspeichers) und damit die Steigerung der Leistung (CPS, CUPS) und der Netzwerkgröße einfach möglich ist.

Eines der wichtigsten und häufig ausschlaggebenden Beurteilungskriterien ist letztlich der Preis eines Neurocomputers. Hier ist zum einen der Preis für die eigentliche Hardware zu berücksichtigen als auch der Preis für die im allgemeinen unerläßliche Softwareunterstützung in Form von angepaßten KNN-Entwicklungsumgebungen. Anhand der vorgestellten Vergleichskriterien werden in den folgenden Abschnitten verschiedene Realisierungsvarianten für Neurocomputer diskutiert.

Neurocomputer auf der Basis von Standard-VLSI-Bausteinen

Der Begriff Neurocomputer ist, trotz seiner häufigen Verwendung in der einschlägigen Literatur, bisher noch nicht genauer definiert worden. Im allgemeinen versteht man unter dem Begriff Neurocomputer eine (Rechner-)Hardware, auf der sich KNN effizient implementieren lassen. Unter effizient versteht man wiederum eine schnellere (z.B. in CPS und CUPS ausgedrückt) oder eine kostengünstigere Implementierung als dies zum Beispiel mit konventionellen (Parallel-)Rechnern möglich ist. Konventionelle Rechner sind in diesem Zusammenhang Rechner, die nicht speziell für die Implementierung von KNN konzipiert wurden. Hierzu gehören sowohl Arbeitsplatzrechner (z.B. PCs), Großrechner, Supercomputer (z.B. der CM5-Rechner von Cray Research) als auch massiv parallele Rechnerarchitekturen (z.B. die Connection Machine, Tab. 1), die man im allgemeinen nicht als konventionelle Rechner bezeichnet. Der Preis und die Leistungsfähigkeit dieser konventionellen Rechner bilden somit untere Schranken für eine effiziente Implementierung von KNN auf Neurocomputern.

Zusatzkarten (Add-On-Boards)

Die einfachste und kostengünstigste Art die Leistungsfähigkeit eines Arbeitsplatzrechners zu erhöhen, bilden sogenannte Beschleuniger- oder Zusatzkarten. Sie haben alle den prinzipiell gleichen Aufbau. Basierend auf einem leistungsfähigen Prozessor, hier kommen sowohl CISC (complex instruction set computer, z.B. der Mikroprozessor MC68040 von Motorola)), RISC (reduced instruction set computer, z.B. der Transputer von INMOS oder der Prozessor i960 von Intel) als auch insbesondere DSP (digital signal processor, z.B. der TMS320C30 von Texas Instruments oder der DSP-32 von AT&T) Prozessoren zum Einsatz, verfügen die Karten über einen möglichst großen und schnellen Datenspeicher sowie die notwendige Ansteuer-

bzw. Schnittstellenlogik. Zielsysteme sind im wesentlichen IBM-PCs und Kompatible sowie gängige Workstations (z.B. von SUN Microsystems oder Hewlett Packard).

Die Größe der bearbeitbaren KNN und letztlich die Leistungsfähigkeit der Beschleunigerkarten hängt einerseits von der Größe des Gewichtsspeichers und andererseits von der Rechengeschwindigkeit der verwendeten Prozessoren ab. Die Anzahl der speicherbaren Gewichte ergibt sich einfach aus dem Quotienten des Gewichtspeichers und der geforderten Darstellungsgenauigkeit eines einzelnen Gewichtes. Die Rechengeschwindigkeit ist durch den verwendeten Prozessor weitgehend festgelegt, wobei zu beachten ist, daß CISC-Prozessoren i.a. durch einen mathematischen Coprozessor unterstützt werden können. Einen wichtigen Einfluß auf die Rechengeschwindigkeit hat auch die Art der verwendeten Speicherbausteine (statische RAMs (random access memories) versus dynamische RAMs). Aus Kostengründen, der Preis einer Zusatzkarte wird weitgehend durch die Speichergröße bestimmt, werden in kommerziellen Produkten fast ausschließlich die langsameren, aber preisgünstigeren dynamischen RAMs (DRAMs) verwendet.

Aufgrund der derzeitigen Verfügbarkeit und den relativ niedrigen Anschaffungskosten dominieren zur Zeit die Beschleunigerkarten den Markt für neuronale Hardware. Die in Tabelle 1 genannten Produkte ANZA+ und Delta II gehören zu den ersten am Markt verfügbaren Zusatzboards für PCs zur schnellen Simulation von neuronalen Netzen. Für die schnelle Multiplikation und Addition werden hier keine Standard-Prozessoren sondern für diese Operationen spezialisierte VLSI-Bausteinsätze verwendet. Beide Produkte gehören aber nicht mehr zu den neuesten Entwicklungen der am Markt vertretenen Firmen. Die steigende Leistungsfähigkeit der Mikroprozessoren, die sich der 100 MFLOP (Millionen Floating Point Operations per Second) Leistungsmarke nähert, und die steigende Dichte bei den integrierten Speicherbausteinen führt auch zu leistungsfähigeren Beschleunigerkarten, sodaß für Add-On-Boards Leistungswerte von 100 MCPS bzw. 50 MCUPS möglich werden.

Parallelrechner

Parallelrechner mit vielen einzelnen Prozessoren sind heute im Bereich der Rechnersysteme nichts neues mehr. Auf allen bekannten Parallelrechnersystemen wurden bereits neuronale Netze implementiert. Bekannte Vertreter sind beispielsweise die Connection Machine von Thinking Machines (Tab. 1) mit bis zu 65.536 bitseriellen Rechenwerken oder der iPSC von Intel mit maximal 128 Mikroprozessoren vom Typ Intel 80386, unterstützt vom 80387-Coprozessor. Die Connection Machine legt eine Implementierung von neuronalen Netzen nahe, in welcher jeder Prozessor mit der Simulation eines PEs beauftragt ist. Im allgemeinen haben Parallelrechner aber weniger Prozessoren als das zu simulierende neuronale Netz Neuronen (PEs) hat (wie z.B. beim iPSC), sodaß ein Prozessor mehrere PEs simulieren muß. Man spricht hier auch von einer *virtuellen KNN-Implementierung*. Dies muß nicht notwendigerweise eine geringere Leistungsfähigkeit nach sich ziehen, da entsprechend leistungsfähigere Prozessoren die geringere Parallelität wieder ausgleichen können.

Produkt	Hersteller	Hardware	Gewichts-speicher[MB]	CPS/CUPS	Preis x1000$
Connection Machine CM-2	Thinking Machines Corp.	Bit-serielle Prozessoren		100M/40M	
ANZA+	Hecht-Nielsen Corp.	Weitek-XL-Chipsatz	10MByte	6M/1.5M	
Delta II	SAIC	Cipsatz	12Mbyte	11M/ 3M	
N6400	Adaptive Solution Inc.	digitaler ASIC, 64 PE	256KByte SRAM	1.6G/256M	ca. 1
MA 16	Siemens AG	digitaler ASIC, 4 PE	off chip	800M/--	ca. 2
CNAPS	Adaptive Solution Inc.	4xN6400	1MB	5.1G/1.1G	
Synapse-1	Siemens AG	8xMA-16	128M	5.1G/ca.1G	ca. 200
NAM256	AT&T	analoger ASIC, max. 256PE	32KBit	80G/--	ca. 1
ETANN 80170NX	INTEL	analoger ASIC, 64 PE	analoge Gewichte	2G/--	

Tabelle 1 Beispiele für Neurocomputer und Neuro-Chips

Das eigentliche Problem bei der Realisierung von neuronalen Netzen auf Parallelrechnersystemen resultiert aus dem hohen Vernetzungsgrad neuronaler Netze. Die Prozessorknoten eines Parallelrechners sind aus Aufwandsgründen nur mit einer Teilmenge der übrigen Knoten verbunden. Bei der Versendung von Aktivierungswerten von PEs innerhalb des neuronalen Netzwerkes müssen daher mehrere Prozessoren durchlaufen werden. Der nötige Kommunikationsaufwand steigt schnell über die Kapazität der Kanäle, so daß die Rechenleistung des Systems nicht mehr vollständig ausgeschöpft wird. Zur Ausnutzung der maximalen Systemleistung stellt sich die nicht triviale Aufgabe, eine Ausgewogenheit zwischen Parallelisierungsgrad und Kommunikationsaufwand zu erreichen.

Parallelrechnersysteme des o.g. Typs sind kommerziell erhältlich und ermöglichen eine schnelle Simulation von sehr großen neuronalen Netzen. Allerdings stellen sie für viele Anwendungen aufgrund der hohen Anschaffungskosten sowie des großen räumlichen Volumens keine adäquate Lösung dar. Die Lücke zwischen Einprozessorsystemen und hochgradig parallelen Rechnersystemen bilden erwartungsgemäß Systeme mit wenigen (<100) Prozessoren. Insbesondere sind hier auch die Systeme interessant, die als Zusatzkarten für Arbeitsplatzrechner zur Verfügung stehen; z.B. Transputerkarten für den PC. Der überwiegende Anteil der bekannten Systeme verwendet derzeit noch eine lineare Anordnung der Prozessoren (bus-oriented architecture), die relativ einfach zu programmieren ist. Ansonsten muß auch hier auf eine Ausgewogenheit zwischen Parallelisierungsgrad und Kommunikationsaufwand geachtet werden.

Neurocomputer mit Multiprozessorarchitektur können derzeit eine Leistungsfähigkeit von bis zu 1 GCPS (Giga=10^9) bzw. 500 MCUPs erreichen. Auch bei den Parallelrechnern wird die ständige Weiterentwicklung von Mikroprozessoren und von speziellen, aber KNN-unspezifischen VLSI-Bausteinen unmittelbar auch zu einer Erhöhung dieser Leistungswerte führen. Hier liegt ein nicht zu vernachlässigender Vorteil von Neurocomputern auf der Basis von handelsüblichen VLSI-Bausteinen.

Neurocomputer auf der Basis neuronaler ASICs

Die bisher betrachteten Hardwarerealisierungen von KNN bewegen sich auf sehr konventionellen Bahnen. Der nächste Schritt ist die Entwicklung von speziellen, auf den Einsatz zugeschnittenen VLSI-Bausteinen, die hier mit *neuronalen ASICs* bezeichnet werden (Bild 1). Neuronale ASICs können sowohl in analoger als auch digitaler Schaltungstechnik realisiert werden. Ferner findet man in der Literatur die Unterscheidung in *Neuro-Chips* und *neuronale Signalprozessoren* (neural signal processor, NSP).

Neuro-Chips werden für konkrete Anwendungen entwickelt, welche mit kleinen Netzen eines bestimmten KNN-Modells auskommen. Bei geringer Zahl von Neuronen bzw. Gewichten lassen sich diese zur Ermittlung einer anwendungsgerechten Topologie und der optimalen synaptischen Werte (Gewichte) auf heutigen Rechnern bzw. den im vorhergehenden Abschnitt erwähnten Neurocomputern noch in erträglicher Zeit simulieren. Aufgrund der bekannten Erfolge der Mikroelektronik ist es bereits heute möglich, kleine bis mittelgroße neuronale Netze (bis zu 1000 PEs, 1 Million Gewichte) auf einem bzw. wenigen Bausteinen zu integrieren. Der modulare und oftmals reguläre Aufbau neuronaler Netze kommt den Chip-Entwicklern sehr entgegen, weil sich dadurch die Entwurfskomplexität deutlich verringert. Aus diesem Grunde gibt es eine Vielzahl von Entwurfsstudien zu Neuro-Chips, von denen aber nur wenige kommerziell erhältlich sind oder bereits angewendet werden. Die Diskrepanz zwischen dem heutigen Potential der Mikroelektronik und dem Realisierungsstand neuronaler ASICs verringert sich aber zunehmend mit der Verbreitung und Anwendung von neuronalen Netzen. Mit einer stärkeren Nutzung von KNN in der Informationstechnik wird auch der Bedarf an Spezialbausteinen wachsen. Insbesondere dann, wenn Lösungen mit sehr geringem Platzbedarf und Echtzeitverhalten gefordert sind.

Die Analyse und Erforschung von neuen KNN-Modellen und Lernverfahren für Netze mit mehreren tausend Neuronen erfordern allerdings leistungsfähigere Spezialrechner. Um die Forderung der Anwendungsforschung nach diesen leistungsfähigen *General-Purpose-Neurocomputern* zu erfüllen und den Neuro-Chip-Entwicklern eine geeignete Simulationsplattform bereitstellen zu können, muß die Architektur eines derartigen Neurocomputers auf den rechenintensiven Operationen aufbauen, die einer möglichst großen Palette von KNN-Modellen gemeinsam ist. Zu den rechenintensiven Operationen gehören die Multiplikation und Addition von Matrizen, die Transposition und Skalierung von Matrizen, die Normierung von Vektoren und die Bestimmung von Minima/Maxima [5]. Um die Leistungsfähigkeit von Neurocomputern auf der Basis von kommerziellen Mikroprozessoren zu erhöhen, sind hier spezielle VLSI-Bausteine gefordert (Neural Signal Processor, NSP). Diese sind optimiert für eine schnelle Ausführung der genannten rechenintensiven KNN-Operationen, wohingegen die nicht-rechenintensiven Operationen von einem Mikroprozessor bearbeitet

werden können [5]. Die Programmierung solcher Neurocomputer erfolgt ebenfalls auf der Grundlage der neuronalen Elementaroperationen.

Offensichtlich weisen Neuro-Chips und NSPs unterschiedliche Merkmale auf. Um einen Einblick in den Stand der Technik zu geben, werden im folgenden einige ausgewählte, bereits realisierte VLSI-Bausteine vorgestellt.

Digitale neuronale ASICs

Die Digitaltechnik ist nach wie vor die dominierende Schaltungstechnik im Bereich der Rechnerhardware. Der Vorteil der Digitaltechnik liegt darin, daß die Modellierung der Netzwerkeigenschaften weitgehend unabhängig von der schaltungstechnischen Realisierung ist. Ferner ist es leichter zwischen verschiedenen KNN-Modellvarianten umzuschalten. Hauptaufgabe des digitalen Schaltungstechnikers ist eine für die geforderte Reaktionszeit geeignete Multiplizier- Akkumulatorarchitektur mit minimalem Flächenbedarf zu entwickeln sowie die Einstellbarkeit von Charakteristiken und Parametern eines KNN-Modells in genügend weitem Bereich zu ermöglichen [5].

In Tabelle 1 sind zwei Beispiele für digitale neuronale ASICs aufgeführt (N6400, MA16), die bereits realisiert und kommerziell erhältlich sind. Für diese Bausteine ist auch die notwendige Entwicklungsumgebung verfügbar. Der N6400 und der MA16 sind NSPs und daher ausgelegt für den Aufbau eines General-Purpose-Neurocomputers, mit dem eine Vielzahl von neuronalen Netzwerkmodellen simuliert werden kann. Beide Bausteine gehören zu den derzeit neuesten Entwicklungen im Bereich der kommerziellen neuronalen ASICs. Sie unterstützen insbesondere auch das Lernen von KNN.

Der N6400 (Adaptive Solutions [9]) enthält 64 nutzbare Prozessoreinheiten (16 Bit Rechenwerke mit einem 8x16 Bit Multiplizierer) mit jeweils 4KByte (K=1024) statischem Speicher (SRAM) auf dem Chip. Die Auflösung der Gewichte kann zwischen 1 und 16 Bit variieren. Damit kann ein N6400 maximal 128K 16Bit Gewichte bzw. 2048K 1Bit Gewichte speichern. Für die Simulation von größeren KNN müssen entsprechend mehr Bausteine verwendet werden. Die Ausgabe eines Neurons ist auf maximal 8 Bit beschränkt.

Der MA16 (Siemens AG [10]) hat eine systolische Architektur mit 4 Verarbeitungseinheiten, die unter anderem vier 16x16 Bit Multiplizierer mit nachfolgender Addierereinheit enthalten. Die Ergebnisse können bis zu 47 Bit aufakkumuliert und ausgegeben werden. Neben dieser Grundoperation sind eine Reihe weiterer Operationen auf dem Chip realisiert, mit denen man alle gängigen KNN Algorithmen rechnen kann (Abstände zwischen Vektoren, min/max-Suche, etc.). Im Gegensatz zum N6400 ist der Gewichtsspeicher nicht auf dem Baustein integriert. Damit wird wertvolle Chipfläche zur Steigerung der Rechenleistung verwendet, während kostengünstige DRAM Bausteine zur Speicherung der Daten dienen. Ferner kann so der Gewichtsspeicher und damit auch die Netzwerkgröße einfacher variiert werden.

Digitale neuronale ASICs erreichen bereits als einzelne VLSI-Bausteine eine höhere Leistung als die unter 3.2 erwähnten Parallelrechner. Für den N6400 und MA16 sind bereits Neurocomputer mit mehreren solcher Bausteine und entsprechenden KNN-Entwicklungsumgebungen entwickelt worden (Tab. 1: CNAPS, Synapse-1). Je nach Ausstattung dieser Neurocomputer sind Leistungen von 10 GCPS und 3 GCUPS realisierbar.

Insbesondere die hohe Lernleistung und die Flexibilität machen diese beiden Systeme geeignet sowohl für die Analyse der Eigenschaften und der Anwendungsfelder unterschiedlicher KNN-Modelle, als auch für die Simulation von Neuro-Chips.
Beispiel digitaler Neuro-Chips für die Realisierung spezieller KNN-Modelle finden sich ebenfalls in der einschlägigen Literatur [1-8]. Digitale Neuro-Chips haben aber kein nennenswertes kommerzielles Interesse geweckt, weil sie gegenüber NSPs zu unflexibel und gegenüber analogen Neuro-Chips sowohl langsamer als auch flächenintensiver sind.

Analoge Neuro-Chips

Eine der Stärken der digitalen Schaltungstechnik ist die hohe Rechengenauigkeit gegenüber der Analogtechnik. Dadurch ist sie in den letzten Jahrzehnten in immer mehr Bereiche eingedrungen, die vormals von der analogen Schaltungstechnik beherrscht wurden. Durch die neuronalen Netze hat die Analogtechnik jedoch in den letzten Jahren eine gewisse Renaissance erfahren. Die Auffassung, daß viele neuronale Modelle auch mit geringer Rechengenauigkeit arbeiten können, hat analogen Realisierungen neue Möglichkeiten eingeräumt. Die zentralen Operationen Multiplikation und Addition können durch physikalische Gesetze, z.B. das Ohmsche und Kirchhofsche Gesetz, mit geringem Aufwand realisiert werden. Zudem sind die Schnittstellen von elektronischen Systemen zur Außenwelt oftmals ebenfalls analog, sodaß eine Analog-Digital- bzw. Digital-Analog-Wandlung der Ein-/Ausgabesignale entfallen kann.

Hauptproblem der analogen Realisierung ist derzeit die analoge Speicherung der Gewichte eines neuronalen Netzes. Umgeht man dieses Problem, indem man die Gewichte digital speichert, gibt man einen wesentlichen Vorteil einer analogen Realisierung auf, die Kompaktheit. Viele Realisierungsvarianten entschärfen das Problem insofern, daß sie nur binäre bzw. ternäre Gewichte (-1,0,+1) zulassen (Tab. 1, NAM256). Eine interessante Alternative bietet die Verwendung von sog. Floating-Gate-Transistoren, die auch bei digitalen EEPROM-Speichern (elctrically eraseable programmable read only memory) Anwendung finden. Dieses Prinzip wird beispielsweise im ETANN-Chip (electrically trainable artificial neural network, Tab. 1) von Intel [11] ausgenutzt, dem zur Zeit einzigen kommerziell erhältlichen analogen neuronalen ASIC. Der ETANN-Chip, auf dem 64 PEs mit insgesamt 10240 Gewichten integriert sind, kann 2 GCPS berechnen. Die Ein-/Ausgaben des Bausteins sind analog, was sehr interessant für sensorische Anwendungen ist. Das Lernen wird allerdings aufgrund der relativ langen Programmierzeiten für die EEPROM-Zellen (im Millisekundenbereich) nicht direkt unterstützt. Die Gewichte werden vorher berechnet, z.B. von einem KNN-Simulator, und anschließend in die Speicherzellen einprogrammiert. Ein anschließendes Nachlernen (fine tuning) zur Leistungsverbesserung wird von der ebenfalls kommerziell erhältlichen Entwicklungsumgebung allerdings unterstützt.

Neben der modellorientierten Umsetzung von KNN in Silizium hat sich der sogenannte *neuromorphische Ansatz* (neuromorphic approach [12]) entwickelt. Bei diesem Ansatz, der von der analogen Schaltungstechnik dominiert wird, findet eine stärkere Einbeziehung der physikalischen Eigenschaften mikroelektronischer Bauelemente und Schaltungen statt, in dem diese Eigenschaften direkt in das KNN-Modell mit einfließen. Bekannte und eindrucksvolle Beispiele sind die von C. Mead und seiner Forschergruppe entwickelte Silizium-Retina und -Cochlea [12]. In diesen Bausteinen sind die Sensorelemente und die Signalverarbeitungsstufen eng miteinander verzahnt. Für diese Integration zeigt sich die analoge Schaltungstechnik der digitalen Technik deutlich überlegen. Zum einen, weil die Eingangssignale analog sind. Zum

anderen, können die Signalvorverarbeitungsalgorithmen (z.B. logarithmische Kompression, Mittelwertbildung, etc.) sehr effizient mit analogen Bauelementen realisiert werden. Der neuromorphische Ansatz ist zur Zeit ausgesprochen forschungsorientiert, daher sind kommerzielle Produkte noch nicht verfügbar.

Aufgrund der eingeschränkten Rechengenauigkeit analoger Schaltungen und dem Problem bei der adaptiven Gewichtsspeicherung eignen sich analoge neuronale ASICs nur zur Realisierung von neuronalen Netzen mittlerer Größe, die mit einer Rechengenauigkeit von weniger als 8 Bit auskommen. In diesem Bereich können analoge Realisierungen durchaus interessante Alternativen zu digitalen Lösungen bieten, die sowohl wesentlich schneller arbeiten als auch mit weniger Fläche und geringerem Leistungsverbrauch auskommen. Allerdings erkauft man sich diese Vorteile mit einem Verlust an Flexibilität, da i.a. nur ein bestimmtes neuronales Modell realisiert wird, und an Skalierbarkeit, da zumindest die Anzahl der Eingänge pro PE festgelegt ist. Hat man für eine spezielle Anwendung den Nutzen eines bestimmten KNNs nachgewiesen, sind diese Freiheitsgrade allerdings nicht mehr entscheidend. Analoge neuronale ASICs zielen daher besonders auf sogenannte eingebettete Anwendungen (embedded applications), wie z.B. die Erkennung von handgeschriebenen Ziffern oder Buchstaben auf Briefen oder Schecks.

Diskussion und Ausblick

Die Realisierung von Künstlichen Neuronalen Netzen in Hardware stellt bereits heute aufgrund der revolutionären Entwicklung der Mikroelektronik kein prinzipielles Problem mehr dar. Fast alle namhaften Chip-Hersteller zeigen Aktivitäten auf dem Gebiet der Hardwareentwicklung für KNN, so daß sich die Anzahl der kommerziell verfügbaren Neurocomputer bzw. Neuro-Chips auch in Zukunft steigern wird. Darüber hinaus gibt es vielfältige Forschungs- und Entwicklungsprojekte an Universitäten und Forschungsinstituten. Die Leistungsfähigkeit der verschiedenen Neurocomputerkonzepte (Bild 1) wird sich ebenfalls mit der technologischen Weiterentwicklung ständig erhöhen. Ähnlich wie die Zielsetzung der TeraFLOP-(10^{12}FLOP=10^{6}MFLOP)-Initiativen im Bereich der Hochleistungsrechner werden auch im Bereich der Neurocomputer 10^{12}CPS (TeraCPS) bis zum Jahre 2000 angestrebt. Diese Leistungsstufe wird noch erreichbar sein, ohne auf optoelektronische und optische Konzepte zurückgreifen zu müssen. Die Mikroelektronik wird somit noch für mindestens eine Dekade die Hardware-Realisierung von KNN dominieren.

Die Entwicklungslinien für Neurocomputer, wie sie in Bild 1 aufgezeigt und hier diskutiert worden sind, werden in Zukunft weiterhin Bestand haben. In Bild 2 werden die Rechenleistung und Speicherkapazität der hier diskutierten vier Neurocomputervarianten bezüglich wichtiger KNN-Anwendungsfelder gegenübergestellt. Arbeitsplatzrechner, gegebenenfalls mit Beschleunigerkarten aufgerüstet, werden aufgrund ihrer großen Verbreitung, ihrer Flexibilität und der Verfügbarkeit von unterschiedlichen KNN-Simulatoren weiterhin das untere Leistungsspektrum abdecken (Bild 2). Parallelrechner und insbesondere Neurocomputer auf der Basis von digitalen neuronalen ASICs ermöglichen gegenüber diesen Systemen eine Leistungssteigerung um mindestens zwei Größenordnungen. Eine wichtige Voraussetzung für die Nutzung dieser Leistung und die Akzeptanz dieser Systeme wird die Verfügbarkeit adäquater KNN-Entwicklungsumgebungen sein. Das gilt insbesondere für General-Purpose-Neurocomputer, die für eine flexible Simulation und eine Analyse verschiedener KNN-Modelle

ausgelegt sind. Die Entwicklungsunterstützung und die Benutzerschnittstelle spielen eine bedeutende Rolle für den Erfolg dieser Systeme.

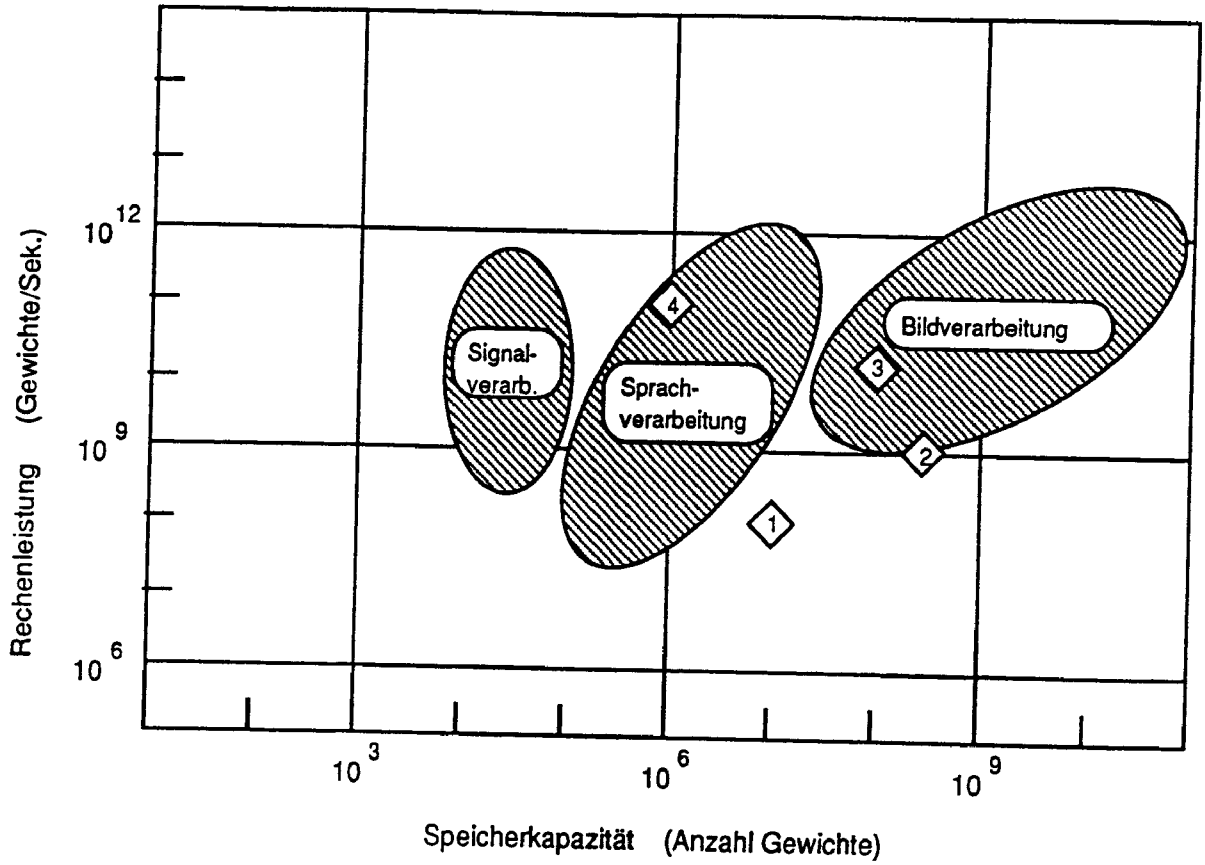

Bild 2 Einordnung der Neurocomputervarianten sowie wichtiger KNN-Anwendungsfelder bezüglich der Speicherkapazität und Rechenleistung (nach [5] und [8]): 1: Add-On-Boards; 2: Parallelrechner; 3: NSP-Systeme; 4: Neuro-Chips.

Anwendungsspezifische und modellspezifische Neuro-Chips bilden das obere Ende des Leistungsspektrums neuronaler Hardware. Hier treten analoge Realisierungen in den Vordergrund, die allerdings in der Gewichtsauflösung (< 8 Bit) und Flexibilität sehr eingeschränkt sind. Spezialisierte neuronale ASICs eignen sich insbesondere für eingebettete, zeitkritische Anwendungen, wie z.B. die Erkennung einfacher Muster im sensorischen Bereich (Bild- oder Sprachvorverarbeitung), die zudem noch eine kompakte Realisierung fordern. Analoge neuronale ASICs unterstützen derzeit noch nicht das Lernen, d.h. hier benötigt man leistungsfähige digitale Neurocomputer für die Lernphase und die Simulation der Anwendungen.

Die in diesem Artikel häufig zitierten Leistungsmaße CPS und CUPS sind, wie bereits erwähnt, nur grobe Anhaltspunkte bei der Beurteilung von Neurocomputern. Signifikant sind diese Angaben streng genommen erst dann, wenn allgemein akzeptierte Bestimmungsverfahren (benchmarks) angewendet werden, was zur Zeit aber nicht der Fall ist. Geringe Unterschiede in den Leistungsangaben von Neurocomputern sind daher mit Vorsicht zu genießen, während Unterschiede von einer Größenordnung und mehr als Entscheidungskriterium durchaus herangezogen werden können. Insbesondere sollte die Angabe von CPS bzw. CUPS auf die realisierbare Gewichtsgenauigkeit bezogen werden. Umso geringer die Gewichtsgenauigkeit, umso höher ist die prinzipielle Leistungsfähigkeit einer Hardwareimplementierung. Zum

Beispiel ist die hohe Leistungsfähigkeit der neuronalen digitalen ASICs N6400 und MA16 (Tab. 1) gegenüber konventionellen Parallelrechnern auch darin begründet, daß die maximale Gewichtsgenauigkeit auf 16Bit beschränkt wurde. Dadurch nehmen die integrierten Rechenwerke (PEs) weniger Siliziumfläche ein, werden schneller und es können mehr PEs auf einem Baustein integriert werden (höherer Parallelitätsgrad). Entsprechendes gilt für die analogen neuronalen ASICs. Inwieweit diese Einschränkungen signifikant werden, hängt letztlich von dem KNN-Modell und der Anwendung ab.

	MCPS	Preis (DM)	DM/MCPS	Antwortzeit (Millisek.)	Größe
Add-On-Board	100	5.000	50	100	10
Parallelrechner	1.000	1.000.000	1.000	100	1.000
Neurocomputer	10.000	200.000	20	10	100
Neurochip	100.000	1.000	0,01	0,01	1

Tabelle 2 Grober Vergleich verschiedener KNN-Realisierungen

Mindestens genauso wichtig wie das Auswahlkriterium Geschwindigkeit ist der Preis bzw. das Preisleistungsverhältnis eines Neurocomputers. In Tab. 2 ist eine Gegenüberstellung der Anschaffungskosten (soweit bekannt) und der CPS-Angaben der in Bild 2 aufgeführten Neurocomputervarianten aufgezeigt. Hier wird deutlich, daß die Entwicklung spezieller Hardware für Neurocomputer neben reinen Geschwindigkeitsaspekten auch aufgrund eines besseren Preisleistungsverhältnisses (DM/MCPS) sinnvoll ist. Neben dem Preis für die Hardware ist hier auch der Preis für die notwendige KNN-Entwicklungsumgebung zu berücksichtigen. Für Neuro-Chips ergeben sich ferner Größenvorteile für die Realisierung von kleinen und mittelgroßen KNN, wobei die in Tab. 2 angegebenen Zahlen nur eine grobe Abschätzung darstellen. Die aufgeführten Antwortzeiten ergeben sich aus dem Quotienten von Speicherkapazität und Rechenleistung bezogen auf die Angaben aus Bild 2.

Zusammenfassend sind die wichtigsten Entscheidungskriterien bei der Auswahl eines Neurocomputers der Preis, die Verfügbarkeit einer komfortablen Entwicklungsumgebung, die Geschwindigkeit (CPS; CUPS), die Skalierbarkeit und die Flexibilität. Die Rangfolge dieser Kriterien hängt letztlich vom Anwendungsprofil ab. Befindet sich der Anwender noch im Experimentierstadium, dann sind neben dem Preis und der Verfügbarkeit einer komfortablen Benutzerschnittstelle die Flexibilität und die Lernleistung eines Neurocomputers entscheidend. Steht für eine spezielle Anwendung das zu verwendende KNN-Modell inklusive der Modellparameter fest, so treten neben dem Preis die Geschwindigkeit (CPS), gegebenenfalls die Kompaktheit und die Einbettung, d.h. die Integration des Neurocomputers in eine vorgegebene (Hardware-) Umgebung, in den Vordergrund. Aus diesen Gründen wird es auch in Zukunft nicht den allgemeinen Neurocomputer für alle Fälle geben, sondern in Abhängigkeit von dem Anwendungs- bzw. Anforderungsprofil werden die verschiedenen, hier skizzierten Konzepte für Neurocomputer auch weiterhin ihr Einsatzgebiet finden.

Literatur

Goser, K., Ramacher, R., Rückert, U.: Tagungsband zum 1. Int. Workshop "Microelectronics for Neural Networks". Universität Dortmund, 1990.

Ramacher, U., Rückert, U.: VLSI Design of Neural Networks. Kluwer Academic Publ., Boston 1990.

Ramacher, U., Rückert, U., Nossek, J.A.: Tagungsband zum 2. Int. Workshop "Microelectronics for Neural Networks". Kyrill&Method Verlag München, 1991.

Ramacher, U., Rückert, U., Nossek, J.A.: Tagungsband zum 3. Int. Workshop "Microelectronics for Neural Networks". Verlag Edinghburg, 1993.

Goser, K., Ramacher, U.: Mikroelektronische Realisierung von künstlichen neuronalen Netzen. Informationstechnik it (1992), Heft 4, S. 241-247.

Klar, H., Ramacher, U.: Microelectronics for Artificial Neural Nets. Fortschritt-Berichte, Reihe 21: Elektrotechnik, Nr. 42, VDI Verlag Düsseldorf, 1989.

Pochmueller, W., Koenig, A., Halgamuge, S.K., Glesner, M.: Neurocomputers. Report ANN91R10, Darmstadt 1992.

DAPRA Neural Network Study. AFCEA Int. Press, 1988.

Hammerstrom, D.: A VLSI Architecture for High-Performance, Low-Cost, On-Chip Learning. Proc. of the Int. Joint Conf. on Neural Networks, Vol. II, 1990, S. 537-544.

Ramacher, U.: SYNAPSE - A Neurocomputer that Synthesizes Neural Algorithms on a Parallel Systolic Engine. Journ. of Parallel and Distributed Computing 14, 1992, S. 306-318.

Holler, M., Tam, S., Castro, H., Benson, R.: An Electrically Trainable Artificial Neural Network (ETANN) with 10240 Floating Gate Synapses. Proc. of the Int. Joint Conf. on Neural Networks, 1989, S. 191-196.

Mead, C.: Analog VLSI and Neural Systems. Addison-Wesley, 1989.