SambaNova SYSTEMS

WHITEPAPER

KI verändert alles, was Sie über Hardware und Software wissen. Hier erfahren Sie den Grund dafür

So ändert die RDA von SambaNova Systems die Regeln für KI und HPC

SITUATION PUBLISHING

Die Anforderungen immer raffinierterer KI-Anwendungsfälle haben die klassische Computerhardware und -software an ihre Grenzen gebracht. Aber werden die Architekturen, die für die nächste Generation von KI eingesetzt werden, auch unseren Ansatz für herkömmliches HPC und Computing im Allgemeinen verändern?

Genau darauf arbeitet das Start-up SambaNova Systems für softwaredefinierte KI-Hardware mit seiner RDA (Reconfigurable Dataflow Architecture) hin, die eine neue Art der KI-Freiheit von den Einschränkungen herkömmlicher Software und Hardware ermöglicht.

Dieses erweiterte Profil erklärt, wie die Gründer von SambaNova Systems ihre jahrzehntelange Erfahrung in einigen der bekanntesten Unternehmen und Institutionen des Silicon Valley genutzt haben, um herauszufinden, warum herkömmliche Architekturen bei der Förderung von maschinellem Lernen und KI an ihre Grenzen stoßen und wie dies zur Entwicklung von RDA geführt hat.

Es zeigt auch, wie die RDA in Einrichtungen wie dem Lawrence Livermore National Laboratory, wo sie zur Lösung grundlegender physikalischer Probleme eingesetzt wird, den Ansatz für die Lösung klassischer Rechenprobleme völlig neu definieren kann.

Unternehmen müssen KI-Technologien einführen – nicht nur, weil sie es können, sondern, weil sie es müssen. KI ist die Technologie, die Unternehmen dabei unterstützt, agil, innovativ und skalierbar zu sein. Dieser Meinung ist IDC, ein Anbieter von Technologie-Analysen. Das Unternehmen prognostiziert, dass sich die globalen Ausgaben für KI-Systeme in den nächsten vier Jahren von 50,1 Milliarden USD in diesem Jahr auf über 110 Milliarden USD im Jahr 2024 verdoppeln werden.

– IDC-Bericht 2020: „Worldwide Spending on Artificial Intelligence Is Expected to Double in Four Years, Reaching $110 Billion in 2024, According to New IDC Spending Guide“

Zu den treibenden Faktoren für die Einführung von KI gehören die „Bereitstellung einer besseren Kundenerfahrung und die Unterstützung von Mitarbeitern bei der Verbesserung ihrer Arbeit“, so IDC. „Dies spiegelt sich in den führenden Anwendungsfällen für KI wider, darunter automatisierte Kundenservicemitarbeiter, Empfehlungen für und Automatisierung von Vertriebsprozessen, automatisierte Bedrohungsinformationen und -prävention sowie IT-Automatisierung.“ Zu den am schnellsten wachsenden Anwendungsfällen gehören automatisierte Personalprozesse und pharmazeutische Forschungs- und Entdeckungsprozesse, fügt das Forschungsunternehmen hinzu.

Die Vorteile dieser technologischen Revolution sind jedoch sehr ungleichmäßig verteilt, so Kunle Olukotun, Mitbegründer und Chief Technologist von SambaNova Systems, dem Start-up für softwaredefinierte KI-Hardware. „Wenn man sich die Menschen ansieht, die in der Lage sind, diese Art von Systemen zu entwickeln, findet man nur einige wenige große Unternehmen, die über die Daten, die Rechenleistung und das Talent verfügen, um diese Art von Algorithmen zu entwickeln. Und natürlich haben sie diese Systeme genutzt, um zu den reichsten Unternehmen der Welt zu werden – Google, Apple, Amazon, Facebook und andere", sagt er.

Die grundlegende Herausforderung liegt in der enormen Rechenleistung, die für den Aufbau und das Training vieler fortschrittlicher Modelle erforderlich ist, die derzeit entwickelt werden. Die Modelle werden immer größer. Bei einigen Anwendungen wachsen auch die Datenvolumen, die für das Training erforderlich sind, unkontrolliert. Verschärft wird dies durch die Verlangsamung der Performance-Steigerungen von nachfolgenden Generationen von Prozessorchips – ein Trend, der laut Marshall Choy, Vice President of Product bei SambaNova, von einigen bereits als das Ende des mooreschen Gesetzes proklamiert wird.

„Die Multicore-Technik hat ihre Grenzen erreicht und einzelne Kerne sind ineffizient. Wenn man viele davon einfach auf einem Chip zusammenführt, nimmt die Ineffizienz offensichtlich zu", sagt Choy. „Wir benötigen also eine viel effizientere Architektur als Plattform für zukünftige Innovationen in den Bereichen KI und maschinelles Lernen, um diese völlig neue Klasse von KI-Anwendungen zu unterstützen."

Heute werden KI-Workloads in der Regel von Racks mit Systemen verarbeitet, die eine Kombination aus CPUs und GPUs verwenden. Letztere verfügen über eine Architektur, die auf eine viel höhere Parallelität ausgelegt ist, da sie Hunderte von relativ einfachen Kernen enthalten, die für einen höheren Gleitkomma-Durchsatz optimiert sind, was sich für Aufgaben wie das Training von maschinellen Lernfunktionen als wesentlich besser geeignet erwiesen hat als CPUs.

DIE GOLDLÖCKCHEN-ZONE DER KI

Dieser Erfolg verschleiert jedoch bisher die Tatsache, dass GPUs ursprünglich nicht für maschinelles Lernen entwickelt wurden und möglicherweise nicht für jede KI-Workload geeignet sind, so Choy.

„Tatsächliche bietet die GPU nur in einem kleinen Teil des Forschungsbereichs für Algorithmen und Anwendungen für maschinelles Lernen eine gute Performance. Die GPU fügt sich in diese Goldlöckchen-Zone für KI ein, da sie im Grunde sehr gut Modelle ausführt, die an die Größe des GPU-Speichers angepasst sind und innerhalb der Grenzen der Architektur liegen", sagt er.

Forscher lassen diese Goldlöckchen-Zone jetzt hinter sich und wenden sich einerseits kleineren, detaillierteren Modellen mit Transformatoren zu, die auf Effizienz auf der einen Seite ausgerichtet sind, und andererseits größeren Modellen mit größeren Datensätzen und höheren Parameterzahlen, wie BERT und GPT im Bereich Natural Language Processing, bei denen Arbeitsspeicherüberlastungen auftreten können oder die Tausende von GPUs für die Bereitstellung erfordern.

Diese Einschränkungen sind von Bedeutung, da viele Unternehmen heute stark in Infrastruktur investieren, die möglicherweise zu unflexibel ist, um sich an die sich schnell verändernden wirtschaftlichen und geschäftlichen Bedingungen anzupassen.

Abgesehen von dem potenziellen Mangel an frei verfügbarer Verarbeitungsleistung für neuere Modelle des maschinellen Lernens gibt es weitere Trends, die den Bedarf an neuen Ansätzen und neuen Architekturen für KI hervorheben. Der erste ist laut Choi, dass die Trainings- und Inferenzprozesse traditionell getrennt gehalten wurden. In der Regel wird das Training eines Modells mit der schieren Kraft von GPUs durchgeprügelt, während die Inferenz, die mithilfe des trainierten Algorithmus für maschinelles Lernen eine Vorhersage erstellt, häufiger mithilfe einer ASIC oder CPU durchgeführt wird.

„In Bezug auf reale Anwendungen beobachten wir eine Notwendigkeit, Training und Inferenz zusammenzuführen. Schließlich möchten Sie Dinge wie die Feinabstimmung von Modellen auf bestimmte Anwendungsfälle und kontinuierliche Lernvorgänge für kleine Modelle ermöglichen, um beispielsweise die Übertragung von Lerninhalten und ein inkrementelles neues Training auf dem Inferenzknoten zu unterstützen“, erklärt er.

Wenn dies mit verschiedenen Systemen erfolgt, sodass die Ergebnisse von einem System zum anderen hin und her verschoben werden müssen, drohen hohe Kosten und eine hohe Latenz im Rechenzentrum. Daher ist der Umstieg auf eine Architektur, die beide Aufgaben auf demselben System ausführen kann, sinnvoller.

SOFTWARE 2.0

Dieses Datenflusskonzept bildet für SambaNova den Kern der Funktionsweise von Computerarchitekturen der nächsten Generation. Diese Veränderung ist so tiefgreifend, dass sie nach der Auffassung des Unternehmens eine neue Ära der Computertechnik einläuten wird.

Die Datenflussberechnung ist das Endergebnis der Annäherung an Anwendungen auf die „Software 2.0“-Art – ein von Andrej Karpathie geprägter Begriff, der sich auf die Art und Weise bezieht, wie Algorithmen für maschinelles Lernen entwickelt werden.

„Vor dem maschinellen Lernen hatten wir die Software 1.0. Ihr Code ist in C++ oder einer anderen Sprache auf hoher Ebene geschrieben. Er erfordert Fachkenntnisse, um das Problem zu lösen und Algorithmen für die verschiedenen Komponenten zu entwickeln und diese dann wieder zusammenzufügen“, so Olukotun.

„Vergleichen Sie dies mit Software 2.0, wo die Idee darin besteht, neuronale Netzwerke mithilfe von Trainingsdaten zu trainieren, und das Programm im Kontext des neuronalen Netzwerks geschrieben wird. Dies bietet eine Reihe von Vorteilen. Der wichtigste ist, dass Sie eine geringere Anzahl von Codezeilen haben, die explizit vom Programmierer entwickelt werden müssen“, erklärt Olukotun.

Olukotun nennt beispielsweise den Google Übersetzer-Dienst, den Google von 500.000 C-Codezeilen in TensorFlow, einem domänenspezifischen Framework für maschinelles Lernen, das von Google entwickelt wurde, aber auch anderweitig verbreitet ist, auf nur 500 Dataflow-Codezeilen reduziert hat.

„Wir beobachten, dass bei der Entwicklung von Anwendungen für maschinelles Lernen übergeordnete Frameworks wie TensorFlow und PyTorch verwendet werden. Und diese Frameworks erzeugen ein Datenflussdiagramm von Operatoren für maschinelles Lernen wie Faltung, Matrix-Multiplikation, Batchnormalisierung und dergleichen", sagt Olukotun.

Diese domänenspezifischen Operatoren für maschinelles Lernen können dann in „parallele Muster" umgewandelt werden, die Parallelität und Lokalität in der Anwendung ausdrücken und für eine höhere Performance optimiert werden können.

„Und wir beobachten, dass diese parallelen Muster nicht nur Anwendungen für maschinelles Lernen, sondern auch die Operatoren in SQL darstellen können, die für die Datenverarbeitung verwendet werden. Und diese können wiederum mithilfe paralleler Muster effizient dargestellt werden", fügt Olukotun hinzu.

Das Rezept für die neue Ära der Computertechnik von SambaNova lautet wie folgt: Unterstützung für einen hierarchischen Datenfluss mit Parallelmuster als natürliches Ausführungsmodell für maschinelles Lernen, Unterstützung für sehr große Terabyte-Modelle, die eine viel höhere Genauigkeit bieten, Unterstützung für die flexible Zuordnung dieser maschinellen Lerndiagramme zur zugrunde liegenden Hardware und die Notwendigkeit, die Datenverarbeitung zu unterstützen, insbesondere SQL-Vorgänge, da diese einen wichtigen Teil des maschinellen Lernens bilden.

Um dieses Rezept umzusetzen, hat SambaNova Systems eine neue Rechenarchitektur namens „Reconfigurable Dataflow Architecture" (RDA) entwickelt, die Software 2.0 unterstützt und maschinelles Lernen für alle Arten von Datenfluss-Rechenproblemen ermöglicht.

Der Name „Reconfigurable Dataflow" bezieht sich auf die Art und Weise, wie die Architektur auf den Fluss von Daten- und Datenflussdiagrammen ausgerichtet wurde, anstatt herkömmliche Computeransätze zu verwenden, und wie sie für die Ausführung von Datenflussdiagrammen neu konfiguriert werden kann. SambaNova kündigte kürzlich die Verfügbarkeit von DataScale an, einer Plattform, die auf der RDA basiert.

HOUSTON, WIR HABEN EIN DATENFLUSSPROBLEM

Der Datenflussprozess ist jedem bekannt, der weiß, wie KI und maschinelles Lernen funktionieren. SambaNova argumentiert jedoch, dass die Architektur, die das Unternehmen zur Beschleunigung dieser Datenflussdiagramme entwickelt hat, auf Probleme jenseits des maschinellen Lernens anwendbar ist, einschließlich vieler Anwendungen im Bereich HPC. Dies liegt daran, dass viele dieser HPC-Anwendungen auch auf Datenflussprobleme zurückgeführt werden können und sehr große Datensätze umfassen, wie es bei maschinellem Lernen der Fall ist.

Herkömmliche Rechenarchitekturen sind jedoch nicht für diese Anwendungsfälle optimiert. Sowohl CPUs als auch GPUs lesen die Daten und Gewichtungen aus dem Arbeitsspeicher aus, führen Berechnungen durch und schreiben die Ausgabeergebnisse dann wieder in den Arbeitsspeicher. Der Prozess muss für jede Phase des Datenflussdiagramms erneut wiederholt werden, was bedeutet, dass eine enorme Speicherbandbreite erforderlich ist, um die Daten ständig hin- und herzubewegen.

Wenn bei Ihnen also ein Datenflussproblem auftritt, benötigen Sie ein System, das auf den Datenfluss ausgelegt ist, um dieses zu lösen – ein System, das von Grund auf als integrierte Software- und Hardwareplattform entwickelt wurde. Hier kommt die Reconfigurable Dataflow Architecture von SambaNova ins Spiel.

„Reconfigurable Dataflow definiert Rechenvorgänge grundlegend neu und konzentriert sich darauf, welche Algorithmen des maschinellen Lernens benötigt werden“, sagt Olukotun. „Es ist ein Meer aus Rechenleistung und Arbeitsspeicher, das durch ein programmierbares Netzwerk eng miteinander verbunden ist. Und der Schlüssel zu diesem Netzwerk besteht darin, dass Sie den Datenfluss zwischen den verschiedenen Rechen- und Arbeitsspeicherelementen so programmieren können, dass er den Anforderungen Ihrer Anwendung entspricht.“

„CARDINAL“-REGELN

Dies zeigt sich am Cardinal SN10, dem vom Unternehmen entwickelten Prozessorchip, den es als „Reconfigurable Dataflow Unit“ (RDU) bezeichnet, um ihn von einer CPU oder GPU zu unterscheiden. Der Cardinal SN10 besteht aus einem Raster konfigurierbarer Elemente, Pattern Compute Units (PCUs) und Pattern Memory Units (PMUs), die über den Chip verteilt und durch eine flexible On-Chip-Kommunikationsstruktur miteinander verbunden sind.

Um einen Algorithmus oder eine Workload zu implementieren, werden die funktionalen Komponenten des Algorithmus auf die Rechner- und Speichereinheiten abgebildet und der Datenfluss zwischen ihnen wird über die Konfiguration der Kommunikationsstruktur implementiert. So wird sichergestellt, dass der Datenfluss, der in diesem Algorithmus oder neuronalen Netzwerk zu sehen ist, in der Konfiguration der Chip-Elemente elegant widergespiegelt wird.

Die On-Chip-Elemente von SambaNova lassen sich nicht mit CPU- oder GPU-Kernen vergleichen, so Choy. „Sie sollten sich diesen Chip als eine Tile-basierte Architektur mit neu konfigurierbaren SIMD-Pipelines vorstellen. Das interessantere Element sind jedoch die Arbeitsspeichereinheiten auf dem Chip, wo wir eine Reihe von SRAM-Banken sowie Speicherverschränkung und -partitionierung haben, d. h. ein Großteil der „Magie" im Chip findet tatsächlich auf der Arbeitsspeicherseite statt."

Es gibt mehrere Hundert Elementtypen (PCUs und PMUs), deren kombinierter Arbeitsspeicher sich auf „Hunderte von Megabyte" an On-Chip-Speicher summiert, was laut Choy im Vergleich zu GPUs ein besseres Ergebnis liefert.

Das Ergebnis ist eine deutliche Reduzierung der Bandbreitenanforderungen außerhalb des Chips, die zu einer sehr hohen Auslastung der Rechenfunktionen führt und es ermöglicht, Teraflop-Performance beim maschinellen Lernen zu erreichen, anstatt Rechenleistung durch das Hin- und Herbewegen von Daten zu verschwenden, wie es in herkömmlichen Architekturen der Fall ist, so Choy.

Das kleinste DataScale-System (SN10-8) benötigt nur ein Viertel des Platzes im Rack und umfasst acht Cardinal SN10-RDUs und einen Arbeitsspeicher von 3 TB, 6 TB oder 12 TB. Dank der Tile-basierten Architektur der RDU können Sie sicher mehrere mandantenfähige, leistungsstarke, gemischte Workloads oder einfach eine große Anwendung über alle RDUs hinweg auf dem gesamten DataScale-System ausführen.

SambaFlow ist im DataScale-System ebenso wichtig. Die Software lässt sich in gängige Frameworks für maschinelles Lernen wie PyTorch und TensorFlow integrieren und optimiert das Datenflussdiagramm für jedes Modell in den RDUs für die Ausführung.

„Nehmen wir zum Beispiel ein PyTorch-Diagramm. Dieses besteht aus einen Diagramm-Analyzer, der die gemeinsamen parallelen Muster aus diesem Diagramm herauszieht. Anschließend durchlaufen die Muster eine Reihe von Zwischendarstellungsebenen, während unser Datenflussoptimierer, Compiler und Assembler die Runtime aufbauen, die wir dann auf dem Chip ausführen", erklärt Choy.

Die Vorbereitungsarbeiten werden von einem x86-Subsystem mit Linux durchgeführt, können jedoch mithilfe der RDU Direct-Technologie von SambaNova umgangen werden, die eine direkte Verbindung zu den RDU-Modulen gestattet.

LAWRENCE LIVERMORE

Auf diese Weise wurde die SambaNova DataScale-Plattform bereits bei einem frühen Kunden, dem Lawrence Livermore National Laboratory (LLNL), eingesetzt. Hier wurde das Corona-Supercomputing-Cluster, das eine Spitzenleistung von über 11 Petaflops bietet, in ein SambaNova DataScale SN10-8R-System integriert.

LLNL-Forscher verwenden die Plattform für kognitive Simulationsansätze, die eine Kombination aus High Performance Computing und KI beinhalten. Die Fähigkeit von SambaNova DataScale, Dutzende von Inferenzmodellen und wissenschaftliche Berechnungen zu Corona gleichzeitig auszuführen, wird dazu beitragen, maschinelles Lernen zur Verbesserung der Forschungsbemühungen für Kernfusion zu nutzen, so das LLNL in einer Pressemitteilung. Forscher haben bereits berichtet, dass das DataScale-System eine fünffache Verbesserung gegenüber einer vergleichbaren GPU mit denselben Modellen zeigt. SambaNova erklärte, dass ein einziges DataScale SN10-8R-System Modelle im Terabyte-Format trainieren kann, wofür ansonsten acht Racks mit Nvidia DGX A100-Systemen auf Basis von GPUs benötigt werden würden.

Diese Konvergenz von HPC und KI unterstützt die Überzeugung von SambaNova, dass seine Architektur, die auf dem Datenfluss basiert, KI-Funktionen nicht nur beschleunigt, sondern auch die nächste Generation der Computertechnik darstellt. (Das Unternehmen geht jedoch nicht davon aus, dass diese neue Computertechnik CPU-basierte Systeme für eher transaktionsorientierte Anwendungen ersetzen wird.)

„Dieser Software 2.0-Ansatz eignet sich nicht nur für die High-End-Probleme bei der Übersetzung und Bilderkennung, sondern auch für klassische Probleme“, so Olukotun.

„Es hat sich herausgestellt, dass viele klassische Probleme, von der Datenbereinigung über Networking bis zu Datenbanken, eine beträchtliche Menge an Heuristik enthalten. Sie werden bemerken, dass es sich lohnt, diese Heuristik durch Modelle auf der Grundlage von Daten zu ersetzen, da Sie so sowohl eine höhere Genauigkeit als auch eine bessere Leistung erreichen.“

Er nennt das Beispiel für parallele Muster, die Anwendungen des maschinellen Lernens darstellen, und betont, dass diese ebenso gut geeignet sind, um die für die Datenverarbeitung verwendeten Operatoren in SQL darzustellen.

KI-FORTSCHRITT FÜR ALLE

Gleichzeitig möchte SambaNova zeigen, dass die Technologie des Unternehmens für Kunden aller Größen verfügbar und nicht nur High-End-Forschungslaboren oder großen Unternehmen vorbehalten ist.

So hat das Unternehmen beispielsweise ein „Dataflow-as-a-Service“-Angebot eingeführt, das DataScale-Systeme über einen monatlichen Abonnementservice zur Verfügung stellt. Das Abonnement wird von SambaNova verwaltet und bietet Unternehmen eine risikofreie Möglichkeit, ihre KI-Projekte mit einem OpEx-Modell zu starten.

KI mag für viele ein obskures Thema sein, aber maschinelles Lernen ist ein wichtiger und wachsender Teil des Computings in einem breiten Spektrum alltäglicher Anwendungen. Alles, was KI zugänglicher und effizienter machen kann, ist unverzichtbar, um zukünftige Fortschritte voranzutreiben. SambaNova argumentiert, dass sein systemweiter Ansatz einer vollständigen Software- und Hardwareplattform mit Fokus auf den Datenfluss die richtige Antwort ist.

Auf diese breitere Anwendung von KI haben die Gründer von SambaNova Systems in ihrer gesamten Karriere hingearbeitet. Das Start-up wurde 2017 von einer Gruppe vorausdenkender Ingenieure und Datenwissenschaftler gegründet, die erkannt hatten, dass den aktuellen Ansätzen für KI und maschinelles Lernen langsam die Puste ausging und eine komplett neue Architektur erforderlich wäre, um KI für alle zugänglich zu machen und die Skalierbarkeit, Leistung, Genauigkeit und Benutzerfreundlichkeit zu bieten, die zukünftige Anwendungen benötigen.

Insbesondere KI und maschinelles Lernen haben sich in den letzten zehn Jahren zu wichtigen Tools für die Verarbeitung und Interpretation großer und komplexer Datasets entwickelt.

Dieser Trend wird sich fortsetzen, wobei IDC prognostiziert, dass der Markt für KI-Software im Jahr 2024 einen Umsatz von 240 Milliarden USD erreichen wird (gegenüber 156 Milliarden USD im Jahr 2020).

– IDC-Bericht 2020: „Worldwide Spending on Artificial Intelligence Is Expected to Double in Four Years, Reaching $110 Billion in 2024, According to New IDC Spending Guide“

KI ist jedoch nicht mehr nur für Supercomputing geeignet. Die von Unternehmen gesammelten Datenmengen haben sich zu großen und komplexen Datensätzen entwickelt, was dazu führt, dass maschinelles Lernen in alle Arten von Anwendungen integriert wird, von Natural Language Processing (NLP) über hochauflösende Computer Vision, Empfehlungen und High Performance Computing (HPC) bis hin zu alltäglichen Geschäftsprozessen.

Wie wir gesehen haben, hat der Erfolg dieses Ansatzes dazu geführt, dass Modelle des maschinellen Lernens immer größer wurden, zum Teil, um die Genauigkeit zu erhöhen. Dies erfordert immer mehr Rechenleistung. Wenn sich dieser Trend fortsetzt, könnte er die Entwicklung von moderneren Modellen für maschinelles Lernen behindern und dazu führen, dass fortschrittliche KI bald nur noch den größten Unternehmen zugänglich ist.

Die Gründer von SambaNova waren sich der Einschränkungen bestehender Architekturen bei der Verarbeitung von KI wahrscheinlich besser bewusst als viele andere in der IT-Branche. Kunle Olukotun, Mitbegründer und Chief Technologist des Unternehmens, hat als Professor an der Stanford University einige Pionierarbeit an Multicore-Prozessorarchitekturen geleistet und ein Unternehmen namens Afara WebSystems mitbegründet, um die Technologie auf den Markt zu bringen.

Afara wurde von Sun Microsystems übernommen, wo seine Technologie die Grundlage für die SPARC T1-Prozessorreihe „Niagara" von Sun bildete. Olukotun kehrte nach Stanford zurück, wo er begann, Software zu entwickeln, die die Funktionen von Multicore-Prozessoren voll ausschöpft.

EIN SCHRITT ZURÜCK, EIN SPRUNG NACH VORN

Dies führte zum Konzept von domänenspezifischen Sprachen – eine Idee, die heute weithin für maschinelle Lernaufgaben eingesetzt wird. Diese Idee wurde schließlich in die Konzepte integriert, die SambaNova zur Kommerzialisierung entwickelt hat, d. h. Methoden zur Entwicklung von Hardware und Software, die KI-Technologie deutlich zugänglicher gestalten.

Rodrigo Liang, Mitbegründer und CEO von SambaNova, arbeitete ebenfalls für Afara und blieb nach der Übernahme von Sun bis 2017 dort, wo er der SPARC-Prozessorentwicklung vorstand. Seine Kombination aus geschäftlicher und technischer Erfahrung prädestinierten ihn zum CEO, als SambaNova gegründet wurde, um eine Plattform für maschinelles Lernen und Analysen von Grund auf zu entwickeln.

Der dritte Mitbegründer und Gewinner des „Genius Grant" von MacArthur 2015, Christopher Ré, war ebenfalls an der Stanford University tätig, nämlich als Associate Professor im Department of Computer Science, das eng mit der Statistical Machine Learning Group und dem Stanford AI Lab verbunden ist. Auf der Grundlage seiner Forschungen zu Systemen für maschinelles Lernen gründete Ré Lattice.io, ein Start-up-Unternehmen für Data Mining und maschinelles Lernen, das 2017 von Apple übernommen wurde. Anschließend half er, SambaNova Systems zu gründen, indem er seine Arbeit an der Beschleunigung des maschinellen Lernens beisteuerte.

Die Hintergründe und das Fachwissen in den Bereichen Hardware-, Software- und Chip-Design, Scale-Out-Architektur und maschinelles Lernen haben es den Gründern ermöglicht, von vertrauten bestehenden Rechnerarchitekturen aus einen Schritt zurückzutreten. Von Grund auf haben sie ein integriertes System aus Software und Hardware entwickelt, das sich auf die Anforderungen der Datenverarbeitung aktueller und neuer Anwendungen konzentriert.

Ihr Angebot scheint die Anleger sichtlich beeindruckt zu haben. Bis 2018 hatte SambaNova sich unter der Leitung von Walden International und Google Ventures mit Beteiligung von Redline Capital und Atlantic Bridge Ventures Serie-A-Finanzierungen in Höhe von 56 Millionen USD gesichert.

Im Jahr 2019 folgte eine Finanzierungsrunde der Serie B in Höhe von 150 Millionen USD, dieses Mal unter der Leitung von Intel Capital mit zusätzlicher Beteiligung von Google Ventures, Walden International, Atlantic Bridge Ventures und Redline Capital. Eine Finanzierungsrunde der Serie C folgte im Jahr 2020 und brachte dem Unternehmen 250 Millionen USD unter der Leitung von Fonds und Konten, die von BlackRock verwaltet werden, unter Beteiligung bestehender Anleger.

Die Technologie von SambaNova basiert größtenteils auf Studien von Olukotun und Ré, die sich auf den Workflow und insbesondere den Datenfluss konzentrierten, anstatt auf iterative Anweisungen herkömmlicher Prozessoren.

Muster aus rekonfigurierbarem Arbeitsspeicher und Recheneinheiten, die über eine programmierbare Kommunikationsstruktur verbunden sind, die zur Darstellung des Datenflusses paralleler Muster programmiert werden kann.

SKALIERBARKEIT FÜR DIE ZUKUNFT

Doch bei der vollständigen SambaNova DataScale-Plattform, die sowohl „as-a-Service" als auch als lokale Lösung angeboten wird, ist die Software ein ebenso wichtiger Bestandteil des Puzzles. SambaFlow ist ein kompletter Software-Stack, der Eingaben von standardmäßigen maschinellen Lernstrukturen wie PyTorch und TensorFlow übernimmt und die Kompilierung, Optimierung und Ausführung der Modelle auf allen RDUs im System weitgehend automatisiert.

Dieser Ansatz verspricht bereits heute, komplexe Probleme des maschinellen Lernens effizient zu verarbeiten, einschließlich Modelle mit 100 Milliarden Parametern, und lässt sich einfach skalieren, um Trainingsdaten im Terabyte-Bereich oder mehrere Modelle gleichzeitig zu verarbeiten. Dabei kommt dasselbe Programmiermodell zum Einsatz, das auch auf einer einzigen RDU ausgeführt würde.

Es gibt Hinweise darauf, dass Sprachmodelle jedes Jahr um den Faktor 10 wachsen. SambaNova behauptet sogar, dass die Vorarbeit und die bisher erreichten Ergebnisse des Unternehmens zeigen, dass ein Modell mit einer Trillion Parameter denkbar sei. Dieser Spielraum ist nach Einschätzung des Unternehmens erforderlich, da insbesondere Trends für einen umfassenderen Kontext und größere Einbettungen im Bereich Natural Language Processing die Infrastrukturanforderungen über die aktuellen Grenzen hinaus verschieben werden.

Dieser Datenflussansatz zur Verarbeitung von Workloads bietet laut SambaNova auch eine breitere allgemeine Anwendbarkeit über das maschinelle Lernen hinaus, da parallele Muster verwendet werden können, um die Operatoren in SQL darzustellen, die für Tasks wie Datenvorbereitung und Datenanalyse verwendet werden.

Laut Bronis de Supinski, Chief Technology Officer bei LLNL, nutzt das Labor die DataScale-Plattform, die es in seinen Corona-Supercomputing-Cluster integriert hat, um eine Technik zu untersuchen, die von den Wissenschaftlern als „kognitive Simulation" bezeichnet wird. Dabei wird maschinelles Lernen verwendet, um die Verarbeitung von Teilen der Simulationen zu beschleunigen.

Diese Arbeit, bei der LLNL Pionierleistungen erbringt, dürfte in Zukunft zahlreichen Branchen zugutekommen, die ebenfalls Physiksimulationen im Rahmen ihres Betriebs ausführen, wie z. B. Öl- und Gasexploration, Flugzeugherstellung und Engineering.

Tatsächlich wird maschinelles Lernen in Zukunft in fast allen Aspekten der Computerbranche eine größere Rolle spielen. Arti Garg, Head of Advanced AI Solutions & Technology bei HPE, einem Unternehmen, das SambaNova als strategischen Partner betrachtet, sieht die Technologie an der Schwelle zu einer viel breiteren Akzeptanz. Basierend auf dieser Entwicklung wird KI seiner Meinung nach viel mehr Menschen beeinflussen, als es derzeit der Fall ist. Und die Erwartungen an die Möglichkeiten von KI-Technologien werden sich verändern.

Liang von SambaNova sagt: „Wir stehen vor einer recht großen Veränderung in der Computerbranche. Sie wurde durch KI vorangetrieben, aber auf Makroebene wird die Veränderung in den nächsten 20-30 Jahren größer sein als KI und maschinelles Lernen."

ÜBER SAMBANOVA SYSTEMS

SambaNova Systems entwickelt die fortschrittlichste Systemplattform der Branche, um KI-Anwendungen vom Rechenzentrum über die Cloud bis zum Edge auszuführen. SambaNova Systems wurde im November 2017 von Branchenexperten, Hardware- und Software-Designexperten sowie erstklassigen Innovatoren von Sun/Oracle und der Stanford University gegründet und hat es sich zum Ziel gesetzt, Unternehmen auf der ganzen Welt KI-Innovationen bereitzustellen, die in der fortgeschrittenen Forschung entwickelt wurden, um einen universellen Zugang zu KI zu erreichen. Zu den Investoren des Unternehmens mit Sitz in Palo Alto, Kalifornien, zählen BlackRock, Walden International, GV, Intel Capital, Redline Capital, Atlantic Bridge Ventures, WRVI Capital und mehrere andere. Wenn Sie weitere Informationen wünschen, besuchen Sie uns unter sambanova.ai oder nehmen Sie direkt Kontakt mit uns auf.