Ein neuer Stand der Technik für Empfehlungsmodelle

Empfehlungsdienste sind ein allgegenwärtiger Bestandteil vieler gängiger und weit verbreiteter Internet- und Verbraucherdienste in allen Branchen – von Einzelhandels- und E-Commerce-Anwendungen, die Cross- und Up-Selling von Produkten und Dienstleistungen ermöglichen, bis zu Online-Verbraucherdiensten wie Mitfahrdiensten oder Peer-Reviews. Sie sind auch im Bank-, Versicherungs- und Gesundheitswesen sowie anderen Branchen zu finden, wo sie schnelle und effiziente Kundenempfehlungen und -erfahrungen liefern.

Es gibt zahlreiche alltägliche Beispiele für Empfehlungssysteme, die Nutzern über Newsfeeds, Beiträge in sozialen Medien, Vorschläge zu Streaming-Medien, Reisen und Hotels sowie Anzeigen mit der höchsten emotionalen Bindung Tipps bieten. Und das aus gutem Grund. Schließlich können selbst kleine Verbesserungen bei den Konversionsraten große finanzielle Vorteile bedeuten. Darüber hinaus müssen für die Fähigkeit eines Unternehmens, umfangreichere, aussagekräftigere Empfehlungen zu liefern, weitaus mehr Attribute in ein Empfehlungssystem integriert werden als nur die Browser- oder Kaufhistorie eines Benutzers.

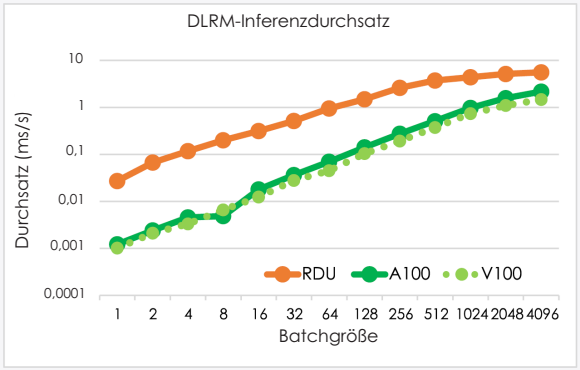
Das scheint ziemlich einfach und intuitiv zu sein. Reale Implementierungen mit veralteten Technologiekomponenten können jedoch die Bemühungen zur Erzielung hochmoderner Genauigkeit, die für die Verbesserung der Geschäftsergebnisse von entscheidender Bedeutung ist, zunichtemachen. Dies gilt für beide wichtigen Phasen der Implementierung eines Empfehlungssystems: Training und Inferenz. Die DataScale-Architektur von SambaNova bietet jedoch die Möglichkeit einer einzigen Plattform, die in beiden Phasen erhebliche Vorteile liefert.

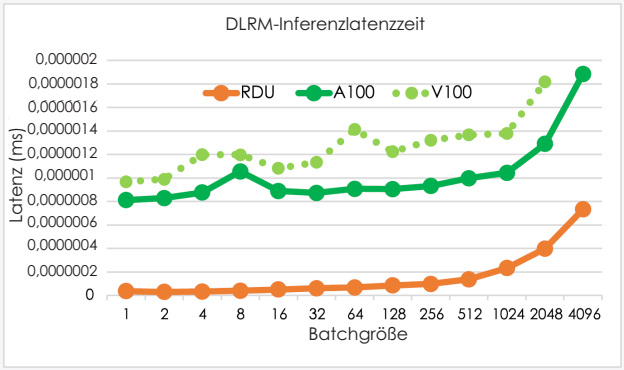
Die Inferenz für Empfehlungsdienste ist einer der weltweit am weitesten verbreiteten Workloads für maschinelles Lernen. Herkömmliche Architekturen stoßen dabei jedoch an ihre Grenzen.

Um aktuelle Einschränkungen zu überwinden, demonstrierte SambaNova Systems, dass das Unternehmen durch die Verwendung des SambaNova DataScale-Systems Empfehlungsinferenzen auf einer Ebene durchführen kann, die neue Funktionen und Geschäftsmöglichkeiten über 20-mal schneller als die führende GPU auf einem branchenüblichen Benchmark-Modell ermöglicht. Diese Art der Verbesserung in Form eines höheren Durchsatzes bei gleichzeitiger Reduzierung der Latenz sorgt für eine deutlich bessere Nutzerbindung und damit für eine schnellere Umsatzgenerierung.

Die Auswirkungen sind sowohl aus technologischer Sicht als auch aus der geschäftlichen Perspektive enorm. Laut Facebook entfallen 79 % der KI-Inferenzzyklen in ihren Produktionsrechenzentren auf Empfehlungen (Quelle). Diese Engines dienen als Hauptfaktoren für die Nutzerbindung und die Gewinngenerierung zahlreicher anderer Fortune-100-Unternehmen. So basieren etwa 35 % der Käufe bei Amazon und 75 % der gesehenen Netflix-Sendungen auf Empfehlungen (Quelle).

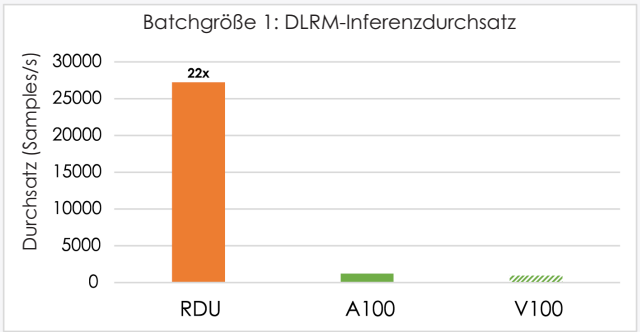
BAHNBRECHENDE EMPFEHLUNGSGESCHWINDIGKEIT

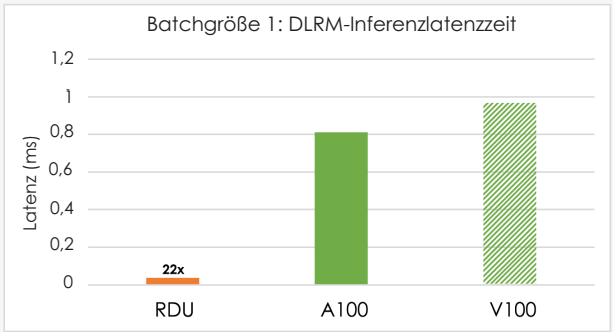




Um die Performance des SambaNova DataScale-Systems zu messen, verwendete SambaNova das Empfehlungsmodell von MLPerf, dem maßgebenden Benchmark für ML-Forscher und -Anwender. Die Task zur Messung der Empfehlungsleistung verwendet das DLRM-Modell im Terabyte-großen Clickthrough-Dataset. Da keine Zahlen für die A100-GPU von Nvidia vorliegen, hat SambaNova eine für Nvidia optimierte Version dieses Modells (Quelle) gemessen. Das Modell wurde auf einer einzigen A100 ausgeführt, die mit einem Triton-Server (Version 20.06) mit FP16-Präzision bereitgestellt wurde. SambaNova führte diesen Test mit einer Vielzahl von Batchgrößen aus, um eine Reihe realistisch bereitgestellter Inferenzszenarien zu simulieren. Als V100-Werte verwendete SambaNova die Performance-Ergebnisse von Nvidia für FP16 (Quelle).

In Bereitstellungsszenarien, in denen Abfragen in Echtzeit gestreamt werden und die Latenz entscheidend ist, sind häufig geringe Batchgrößen erforderlich. Bei diesen geringen Batchgrößen wird der Vorteil der Datenflussarchitektur deutlich erkennbar. Bei einer Batchgröße von 1 bietet das SambaNova DataScale-System eine 20-mal schnellere Performance als eine einzelne A100.





Während Online-Inferenzen bei einer Batchgröße von 1 in bereitgestellten Systemen ein häufiges Anwendungsbeispiel darstellen, möchten Kunden manchmal auch einen Teil ihrer Daten in Batchverarbeitung erfassen, um den Gesamtdurchsatz des Systems zu verbessern. Um die Vorteile des SambaNova DataScale-Systems zu demonstrieren, zeigte SambaNova auch die gleiche DLRM-Benchmark bei einer Batchgröße von 4K. Bei dieser höheren Batchgröße erzielte das DataScale-System für Durchsatz und Latenz mehr als doppelt so gute Performancewerte wie eine A100.

DIE KOMBINIERTE LÖSUNG: TRAINING UND INFERENZ HAND IN HAND

Während viele dieser Messungen auf die Inferenz-Task von MLPerf ausgerichtet sind, zeichnet sich das DataScale-System sowohl bei der Inferenz als auch beim Training durch eine hervorragende Leistung aus. Durch ein neues Training desselben DLRM-Modells von Grund auf und die Nutzung von Variationen, die auf GPU-Hardware gar nicht erst möglich sind, übertrifft die Reconfigurable Dataflow Unit von SambaNova Systems locker den aktuellen Stand der Technik. Lesen Sie diesen Artikel, um mehr zu erfahren.

Um den Anforderungen globaler Unternehmen gerecht zu werden, bietet SambaNova Systems jetzt mehrere DaaS-Abonnements (Dataflow-as-a-Service) zur Unterstützung von Empfehlungen, Natural Language Processing und hochauflösenden Vision-Workloads an. Diese „Quick-Start"-OpEx-Abonnements ermöglichen Unternehmen die schnelle Erstellung von KI-Lösungen und die bedarfsgerechte Skalierung innerhalb eines einfach bereitzustellenden Frameworks, das basierend auf einem Cloud-Verbrauchsmodell verwaltet und abgerechnet wird. Weitere Vorteile dieser „as-a-Service"-Angebote sind die Möglichkeit, das interne ML-Know-how durch mehr Leistung zu erweitern. Nutzen Sie eine Komplettlösung mit zusätzlicher Anbieterexpertise, um Ihren Wechsel zu KI beschleunigen, während Sie dank fortlaufender DaaS-Updates mit aktuellen Modellen und Algorithmustechniken von SambaNova stets von den neuesten F&E-Trends profitieren.

ÜBER DIE BENCHMARK HINAUS: EMPFEHLUNGSMODELLE IN DER PRODUKTION

Die MLPerf-DLRM-Benchmark simuliert eine realistische Empfehlungsaufgabe, kann jedoch nicht den Umfang eines tatsächlich bereitgestellten Workloads erfassen. In einer Analyse von Empfehlungssystemen schreibt Facebook, dass „Empfehlungsmodelle im Vergleich zu Benchmarks in der Produktion mehr Einbettungen aufweisen" (Quelle). Mit dem Wachstum der Modelle geraten CPUs und GPUs zunehmend ins Stocken. Das SambaNova DataScale-System hat hingegen kein Problem damit, diese größeren Rechen- und Speicheranforderungen zu erfüllen, und bietet zudem eine langfristige Lösung, die auf Skalierbarkeit ausgelegt ist.

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.