Myrtle Software Limited

01.06.2023 – 19:03 Uhr

Myrtle.ai erreicht 5,1 Mikrosekunden Latenzzeit im Financial LSTM Inference Benchmark mit VOLLO

Cambridge, England (ots/PRNewswire) -

Myrtle.ai, ein anerkannter Marktführer bei der Beschleunigung maschineller Lerninferenzen, gab heute bekannt, dass ein Stack mit seinem Produkt VOLLO™ kürzlich von STAC®, einer führenden Benchmark-Autorität für die Finanzbranche, geprüft wurde.[1] Dies ist die erste FPGA-based Lösung, deren Ergebnisse für die Tacana Suite von STAC-ML™ veröffentlicht wurden, und sie erzielte unglaubliche Latenzwerte. Die STAC-ML Markets (Inference) Benchmark entspricht den Anforderungen von Kapitalmarktunternehmen, die maschinelles Lernen einsetzen, um schnell auf Marktveränderungen zu reagieren.

STAC-ML™ Markets (Inference) Benchmarks wurden vom STAC Benchmark Council™ entwickelt, dem die weltweit größten Banken, Broker, Börsen, Hedge-Fonds, Eigenhandelsgeschäfte und Vermögensverwalter sowie mehr als 50 führende Technologieanbieter angehören. Sie sind ein Instrument zum objektiven Vergleich verschiedener Plattformen in Bezug auf Latenz, Durchsatz, Effizienz und Qualität bei maschinellen Lerninferenzen (ML).

VOLLO erreichte Latenzzeiten von nur 5,08 Mikrosekunden bei einem Durchsatz von über 800k Inferenzen/Sekunde.[2] Aufgrund der geringen Latenz und des hohen Durchsatzes können Nutzer intelligentere Entscheidungen mit komplexeren Modellen schneller als bisher treffen, was ihnen einen Wettbewerbsvorteil beim Handel, der Risikoanalyse, bei Kursen und vielen anderen handelsbezogenen Aktivitäten verschafft. VOLLO wurde für eine einfache und schnelle Installation auf gemeinsam genutzten Servern entwickelt und ermöglicht zudem eine Reduzierung des Platzbedarfs im Rack und des Energieverbrauchs – zwei wertvolle Ressourcen an solchen Standorten.

Auf dem STAC Summit London am 4. Mai, auf dem die Ergebnisse von STAC bekannt gegeben wurden, hielt Myrtle.ai einen kurzen Vortrag darüber, wie Finanzunternehmen ohne jegliche FPGA-Fähigkeiten von den extrem niedrigen Latenzen profitieren können, die mit einem FPGA-based Produkt wie VOLLO erreichbar sind.

VOLLO läuft auf einer PCIe-Beschleunigerkarte mit Standardformfaktor und Intel Agilex® 7 FPGAs, von denen bis zu 4 in einem 1U-Server installiert werden können. VOLLO kann mit vom Kunden trainierten Modellen konfiguriert werden, die Modellarchitekturen aus dem LSTM Model Zoo verwenden. Dies ermöglicht es Nutzern via Exportprozess aus Standard-ML-Tool-Flows, eine Reihe von Arbeitslasten einzusetzen, die speziell auf ihre Anwendungsanforderungen zugeschnitten sind. Es kann bis zu 12 parallele Modelle pro im System installierter FPGA-Beschleunigerkarte unterstützen, was bei einem System mit 4 installierten Karten ein Maximum von 48 parallelen Modellen ermöglicht.

Der FPGA, der VOLLO zugrunde liegt, ist die Intel Agilex 7 FPGA F-Serie – die erste FPGA-Familie der Branche mit gehärteter, nativer bfloat16-Unterstützung. Für KI-Applikationen, die eine niedrige Latenzzeit benötigen, wie die in der STAC-ML-Benchmark, verringert die Verwendung von gehärtetem bfloat16 die Latenzzeit und erhöht den Durchsatz. Peter Baldwin, CEO von Myrtle.ai, sagte: „Seit der Bekanntgabe unserer ersten STAC-ML-Benchmark-Ergebnisse im Dezember arbeiten wir nun mit Unternehmen zusammen, die sich einen Wettbewerbsvorteil verschaffen wollen, indem sie schnell auf Echtzeit-Marktdaten reagieren. Diese Unternehmen wissen, dass sie den Latenzvorteil, den die FPGAs von Intel bieten, nutzen wollen, verfügen aber oft nicht über die FPGA-Designer, die sie für die Entwicklung ihrer eigenen Lösungen benötigen würden. VOLLO löst dieses Problem."

„Myrtle.ai's herausragende Ergebnisse bei der STAC-ML Markets (Inference) Benchmark zeugen von seiner kombinierten FPGA- und ML-Expertise, und wir freuen uns, dass sie sich für Intel Agilex 7 FPGAs entschieden haben, um den Wert des VOLLO-Produkts zu demonstrieren", sagt Jim Dworkin, Vice President und General Manager der Cloud & Enterprise Acceleration Division in Intel's Programmable Solutions Group. „Alle Finanzdienstleister in dem riesigen Ökosystem, das STAC kultiviert, werden von den niedrigen Latenzzeiten und der hohen Durchsatzleistung profitieren, die VOLLO bietet und die es diesen Unternehmen ermöglichen, intelligentere Handelsentscheidungen zu treffen."

„STAC-Benchmarks werden von Finanzunternehmen auf der Grundlage ihrer geschäftlichen Anforderungen spezifiziert, und sie haben dieses Benchmark entwickelt, um die Inferenzleistung über mehrere Architekturen hinweg zu vergleichen", kommentierte STAC President Peter Nabicht. „Myrtle.ai hat nun in beiden STAC-ML Inferenz-Benchmark-Suiten Ergebnisse mit niedriger Latenz geliefert, was die Fähigkeiten der FPGA-Technologie

unterstreicht und Finanzunternehmen, die mit sich schnell entwickelnden Märkten zu tun haben, wertvolle Leistungsdaten liefert."

VOLLO ist ab heute verfügbar, und Bewertungen können vereinbart werden. Weitere Informationen finden Sie auf myrtle.ai/ fintech oder kontaktieren Sie Myrtle.ai noch heute unter fintech@myrtle.ai.

Weitere Informationen zu STAC und den STAC Benchmark Council finden Sie auf www.STACresearch.com. Die vollständigen Benchmark-Ergebnisse finden Sie im STAC Report (SUT ID MRTL230426) auf www.STACresearch.com/MRTL230426.

Informationen zu Myrtle.ai

Myrtle.ai ist ein KI/ML-Softwareunternehmen, das erstklassige Inferenzbeschleuniger auf FPGA-based Plattformen aller führenden FPGA-Anbieter liefert. Dank seiner Erfahrung mit neuronalen Netzen über das gesamte Spektrum von ML-Netzwerken hat Myrtle Beschleuniger für Fintech, Sprachverarbeitung und Empfehlungen entwickelt.

[1] www.STACresearch.com/MRTL230426

[2] STAC-ML.Markets.Inf.T.LSTM_A.4.LAT.v1 und STAC-ML.Markets.Inf.T.LSTM_A.4.TPUT.v1

Foto: https://mma.prnewswire.com/media/2090072/Vollo_Graphic.jpg

Logo: https://mma.prnewswire.com/media/1969391/Myrtle_Software_Limited_Logo.jpg

View original content:https://www.prnewswire.com/news-releases/myrtleai-erreicht-5-1-mikrosekunden-latenzzeit-im-financial-lstm-inference-benchmark-mit-vollo-301840441.html

Pressekontakt:

Giles Peckham,
+44 7785 278478,
giles@myrtle.ai