HOCHSCHULE HEILBRONN

Proseminar (282136)

XaaS - Anything as a Service

Suphi Pembe (207617),
Andreas Würzer (207258),
Christian Nguyen (207613)

Sommersemester 2022

Vorgelegt bei Claudia Pittel

Management Summary

Hier sollte ziemlich genau bzw. maximal 1 Seite Text stehen (ziemlich genau bedeutet, man sollte so nah wie möglich an 1 Seite herankommen). Text für Test commit haha

Inhaltsverzeichnis

Abkürzungsverzeichnis

GPU Graphics-Processing-Unit oder Grafikkarte

HPC High-Performance-Computing

RAID redundant array of independent disk

Abbildungsverzeichnis

Tabellenverzeichnis

1 Einleitung

Einleitungstext mit Motivation, Ziel der Arbeit (d.h. Erläuterung der Forschungsfrage) und Beschreibung der Vorgehensweise bzw. Aufbau der Arbeit[1]

1.1 Motivation

Durch den aktuell anhaltende Halbleitermangel besteht ein Engpass an Ressourcen von die meisten Wirtschaftszweige betroffen sind. Einer dieser Wirtschaftszweige ist die Produktion von GPUs (graphics processing unit. Diese werden für diverse Anwendung von Computern verwendet, im betrieblichen wie auch im privaten Bereich. Primär in dieser Arbeit werden die Bereiche High-Performance-Computing (HPC) und Gaming haben. Beide diese Bereiche benötigen GPU-Rechenleistung, welche im konventionell von einer lokal verbauten GPU zur Verfügung gestellt wird. Als Langfristige Lösung soll analysiert werden ob es möglich ist durch zentrale Services, welche GPU as a Service anbieten. Durch die zentralen Ressourcenteilung dem Mangel entgegenzuwirken mit einer alternative für den Bedarf zu schaffen.

1.2 Ziel der Arbeit

...

1.3 Vorgehensweise

...

[1] vgl. Theisen, *Wissenschaftliches Arbeiten: Technik – Methodik – Form*, S. 38.

2 Anything as a Service - Cloud Computing

Zwischen den Gliederungspunkten sollten jeweils kurze Überleitungssätze stehen, damit man weiß, um was es inhaltlich in den folgenden Unterkapiteln geht.[2]

2.1 Definition

Bei den Gliederungspunkten immer auf eine Ausgewogenheit achten, damit eine gleichmäßige Gliederung gefördert werden kann. Sofern Abbildungen (wie Abbildung 1: Beispielbild) verwendet werden, müssen diese auch inhaltlich im Text erwähnt und erläutert werden, sowie ein Abbildungsverzeichnis erstellt werden.[3]

2.2 Typische Servicemodelle

Untergliederungen nur in der Mehrzahl erstellen, d.h. nie 1 Unterkapitel alleine stehen lassen. In gleicher Art und Weise wie Abbildungen dargestellt und beschriftet werden, verhält es sich mit Tabellen.

2.2.1 IaaS: Infrastructure as a Service

Inhalt

2.2.2 SaaS: Software as a Service

Inhalt

2.2.3 PaaS: Plattform as a Service

2.3 Vor- und Nachteile

[2] Vgl. Free Software Foundation, *GNU General Public License*.
[3] Vgl. Hochschule Heilbronn, *Wirtschaftsinformatik (B.Sc.) im Überblick*.

3 Knappheit von Grafikkarten

Die Knappheit von Grafikkarten hat den aktuellen Markt durch neue Branchen die GPU-Leistung nutzen nachhaltig verändert. Diese Knappheit entsteht nicht nur durch den Mangel des Rohstoffes, sondern auch durch die Weiterentwicklung von verwendeten Computern in allen Einsatzgebieten.[4]

Im Vergleich zu 2011 wurden für die Automobilindustrie im Jahr 2021 fast drei mal so viele Halbleiter geliefert. Ebenfalls mit der Weiterentwicklung von Internet of Things Produkten wird in Zukunft der Bedarf an Halbleitern weiter steigen.[5][6]
In diesem Kapitel soll die Preisentwicklung von GPUs betrachtet werden, dabei wird ein Zusammenhang geschaffen mit den Ursachen die diese Preisentwicklung verursacht haben.

3.1 Preisentwicklung

Die rapide steigende Preisentwicklung von GPUs ist auf zwei Kernfaktoren reduzierbar.
- Größerer Bedarf an GPUs und Halbleitern, dem Kernbestandteil von GPUs
- Mangelnde Kapazitäten zur Produktion von Halbleitern

Der Bedarf an Halbleitern und GPUs ist konstant im Anstieg. Besonders durch die Corona Pandemie, hat sich im Vergleich zu 2019 im Jahr 2020 ein Umsatzanstieg von 5,4% aufgezeigt.[7]
Wie in Abbildung 3.2 zu sehen ist der Umsatzanstieg größtenteils durch Erlöse von Computersystemen entstanden. Im Vergleich dazu sind Umsätze, die durch Abnehmer in der Automobilbranche entstanden sind gesunken. Das lässt sich auf den steigenden Bedarf an Computersystemen zurückführen. Während der Pandemie mussten viele Menschen Home-Schooling und Home-Office aneignen, um weiter den Alltagsbetrieb ausführen zu können. Ein Nebenläufiger Effekt ist damit, dass durch die Digitalisierung weniger Mobilität benötigt wird. Damit lässt sich der reduzierte Bedarf an Halbleitern erklären. Dennoch ist damit insgesamt der Bedarf an Halbleiter gestiegen.[8]

[4] Vgl. Voas, Kshetri und DeFranco, "Scarcity and Global Insecurity: The Semiconductor Shortage".
[5] Vgl. McClean, *The 2022 McClean Report*.
[6] Vgl. Voas, Kshetri und DeFranco, "Scarcity and Global Insecurity: The Semiconductor Shortage".
[7] Vgl. ebd.
[8] Vgl. ebd.

Der steigende Bedarf allein ist aber wie angeführt nicht der einzige Faktor. Die Produktion von Halbleitern stagniert. Das lässt sich auf verschiedene Ursachen zurückführen. Da die meisten Halbleiter in asiatischen Ländern produziert werden und diese den eigenen Bedarf zuerst decken, ist für den Export weniger Verfügbar. Ebenfalls haben sin in den letzten Jahren Naturdesaster ereignet die Ebenfalls durch z.B. Dürre die Produktion Lahmgelegt haben.[9]
Durch mangelnde Produktion und steigender Bedarf hat sich nun ein sehr hoher Marktpreis entwickelt. Um diesem Entgegenzuwirken ist nicht nur Ressourcenverfügbarkeit zu schaffen, sondern auch eine effizientere Methode die Ressourcen zu nutzen.

3.2 Ursache Halbleitermangel und KryptoMining

Um den Halbleitermangel besser zu verstehen, sollten eine neue Branche die primär GPU-Leistung nutzt thematisiert werden. Seit dem Jahr 2021 haben Kryptowährungen ein mehr als fünffaches Investitionsvolumen im Vergleich zum Vorjahr.[10]
Um den Zusammenhang zu erläutern, Kryptowährungen validieren Ihre Transaktionen durch die Nutzung der Blockchain-Technologie. Bei diesem Validierungsprozess werden neue Datenblöcke in einer Datenbank gespeichert und von anderen Nutzern überprüft durch eine Prüfsumme die dabei gebildet werden können muss. Um diese Prüfsumme berechnen zu können Nutzen sogenannte "Krypto-Miner"primär GPU-Leistung.[11]
Da große Investitionen für das Kryptomining getätigt werden hat sich ein neuer Markt mit einer großen Nachfrage gebildet, der GPUs benötigt.

[9] Vgl. Voas, Kshetri und DeFranco, "Scarcity and Global Insecurity: The Semiconductor Shortage".
[10] Statista-Research-Department, *Volumen der weltweiten Investitionen in Blockchain-Technologien und Kryptowährungen von 2018 bis 2021*.
[11] Vgl. Arslanian, *The Book of Crypto: The Complete Guide to Understanding Bitcoin, Cryptocurrencies and Digital Assets*, S.259-273.

4 Gaming as a Service

Cloud Computing umfasst die Bereitstellung von Rechenleistung und Anwendungen als Dienst über das Internet und soll daher mit Gaming-as-a-Service als Beispiel vertieft werden. Neben der Funktionsweise, werden auch die aktuell verfügbaren Angebote verglichen und damit dem Kauf eines eigenen Computers gegenübergestellt.

4.1 Funktionsweise

Die Ausführung der Spiele, einschließlich der Spielelogik und Wiedergabe der Szenen findet innerhalb der Cloud bzw. Server statt. In Verbindung steht das Gerät des Endnutzers, oder auch Thin-Client genannt. Dieser empfängt die komprimiert gestreamten Audio- und Videosignale über das Internet und gibt sie auf dem Thin-Client wieder. Bei eingehenden Befehlen des Endnutzern, werden diese erfasst und an die Cloud übertragen. Durch die Leistung des Netzwerks zwischen dem Client und der Cloud sind die Prozesse eingeschränkt.

4.2 Anbietervergleich

Inhalt

4.2.1 Voraussetzung

Inhalt

4.2.2 Angebot

Inhalt

4.2.3 Preis

Inhalt

4.3 Hardwarevorraussetzung um Usability zu gewährleisten

Inhalt

5 GPU as a Service

Außerhalb von Gaming as a Service gibt es weitere Service Möglichkeiten für GPU as a Service. Wachsende Märkte dafür sind das rendern von 3D-Modellen und animierten Videos, wie auch deep learning model training für KIs. Ein übliches Monetarisierungsmodel für GPU as a Service ist ön-demand pay-per-use".[12]
Im folgenden Kapitel soll erklärt werden wie GPUs funktionieren, lokal als Hardware und als Service aus der Cloud. Wie auch die Einsatzgebiete feststellen und identifizieren in welchen von diesen Einsatzgebiete sich eine Cloud-Lösung eignet.

5.1 Funktionsweise

Bei der Dienstleistung Inanspruchnahme werden von die von einem geforderten Prozesse, wie z.B. das rendern von 3D-Modellen, durch die Rechenleistung des Anbieters verarbeitet. Im Gegensatz zu der Privatnutzung bei der nur eine GPU genutzt wird, verwenden GPU a as Service Anbieter mehrere GPUs. Allerdings da GPUs nur als Co-Prozessoren in solchen Systemen genutzt werden, können diese nicht eigenständig betrieben werden sondern benötigen einen zentrales Betriebssystem, auch Kernel bezeichnet. Üblicherweise wird das skalierbar angewendet mit einer Vielzahl an Kernels, welche eine Vielzahl an GPUs besitzen.[13]

Es gibt verschiedene Methoden diese Prozesse darauf zu verarbeiten. Eine Methode davon ist das Zeit-Multiplexverfahren. Bei diesem Verfahren werden mehrere Prozesse auf die Kernels sequenziell aufgeteilt und verarbeitet, wie in Abbildung 5.1 zu sehen. Dieses Methode hat keinen direkten Mehrwert in Bezug zur Verarbeitungseffizienz, allerdings verhindert sie das Prozesse Kernels unnötig blockieren.[14]

Eine weitere Methode ist das das man die Kernels über Software als einen Kernel verwendet. Das ist vergleichbar mit dem für Festplatten verwendete "redundant array of independent diskSystem auch abgekürzt genannt RAID, welches gängiger bekannt ist. Durch das fusionieren der Kernels über die Software kann eine konstante Leistung gewährleistet werden. Ebenfalls auch eine fairness zwischen allen Leistungsabnehmern, da es keine

[12] Lattuada u. a., “Performance prediction of deep learning applications training in GPU as a service systems”.
[13] Wang u. a., “Quality of Service Support for Fine-Grained Sharing on GPUs”.
[14] Ebd.

Situation geben kann einen ßchlechten Kernelërwischt zu haben.[15]
[16]

5.2 Einsatzgebiete

[17]

5.3 Vergleich eigene GPU und GPU in der Cloud

Inhalt

[15] Wang u. a., “Quality of Service Support for Fine-Grained Sharing on GPUs”.
[16] Ebd.
[17] Loop und Blinn, “Real-time GPU rendering of piecewise algebraic surfaces”.

6 Marktvorhersage

Inhalt

7 Fazit und Ausblick

Kritische Begutachtung inklusive Zusammenfassung der Arbeit sowie eventuelle Zukunftsperspektiven zum Thema können hier im Fazit und im Ausblick eingebracht werden.[18]

[18] Vgl. Han u. a., "High Expression of Human Homologue of Murine Double Minute 4 and the Short Splicing Variant, HDM4-S, in Bone Marrow in Patients With Acute Myeloid Leukemia or Myelodysplastic Syndrome".

Anhang

Der Anhang soll den eigentlichen Hauptteil nicht ergänzen, sondern darüber hinaus weitere möglicherweise interessante Informationen liefern, die aber nicht zwangsläufig notwendig sind, um den Hauptinhalt zu verstehen

Quellenverzeichnis

Arslanian, Henri. *The Book of Crypto: The Complete Guide to Understanding Bitcoin, Cryptocurrencies and Digital Assets*. Springer eBook Collection. Cham: Springer International Publishing und Imprint Palgrave Macmillan, 2022.

Free Software Foundation, Inc. *GNU General Public License*. Aus Vorlage. 2007. URL: http://www.gnu.org/licenses/gpl.html (besucht am 05. 12. 2015).

Han, Xin u. a. "High Expression of Human Homologue of Murine Double Minute 4 and the Short Splicing Variant, HDM4-S, in Bone Marrow in Patients With Acute Myeloid Leukemia or Myelodysplastic Syndrome". In: *Clinical Lymphoma Myeloma and Leukemia* 16 (2016). Proceedings of the Society of Hematologic Oncology 2015 Annual Meeting, S30 –S38.

Hochschule Heilbronn. *Wirtschaftsinformatik (B.Sc.) im Überblick*. Aus Vorlage. URL: https://www.hs-heilbronn.de/win (besucht am 08. 03. 2018).

Lattuada, Marco u. a. "Performance prediction of deep learning applications training in GPU as a service systems". In: *Cluster Computing* 25.2 (2022), S. 1279–1302.

Loop, Charles und Jim Blinn. "Real-time GPU rendering of piecewise algebraic surfaces". In: *ACM SIGGRAPH 2006 Papers on - SIGGRAPH '06*. Hrsg. von John Finnegan und Julie Dorsey. New York, New York, USA: ACM Press, 2006, S. 664.

McClean, Bill. *The 2022 McClean Report*. URL: https://www.icinsights.com/news/bulletins/The-Real-Reason-Behind-The-Automotive-Industry-IC-ShortageA-StepFunction-Surge-In-Demand/ (besucht am 02. 06. 2022).

Statista-Research-Department. *Volumen der weltweiten Investitionen in Blockchain-Technologien und Kryptowährungen von 2018 bis 2021*.

Theisen, Manuel René. *Wissenschaftliches Arbeiten: Technik – Methodik – Form*. 17., aktualisierte und bearbeitete Auflage. Aus Vorlage. München: Vahlen, 2017.

Voas, Jeffrey, Nir Kshetri und Joanna F. DeFranco. "Scarcity and Global Insecurity: The Semiconductor Shortage". In: *IT Professional* 23.5 (2021), S. 78–82.

Wang, Zhenning u. a. "Quality of Service Support for Fine-Grained Sharing on GPUs". In: *Proceedings of the 44th Annual International Symposium on Computer Architecture*. New York, NY, USA: ACM, 2017, S. 269–281.

Ehrenwörtliche Erklärung

„Wir versichern, dass die vorliegende Arbeit von uns selbständig und ausschließlich unter Verwendung der angegebenen Quellen und Hilfsmittel angefertigt wurde. Alle Stellen, die wörtlich oder annähernd aus Veröffentlichungen entnommen sind, haben wir als solche kenntlich gemacht. Die Arbeit wurde bisher in gleicher oder ähnlicher Form, auch nicht in Teilen, keiner anderen Prüfungsbehörde vorgelegt und auch nicht veröffentlicht.“

SEITEN oder KAPITEL VON BIS 1 wurden von Suphi Pembe verfasst.
SEITEN oder KAPITEL VON BIS 2 wurden von Andreas Würzer verfasst.
SEITEN oder KAPITEL VON BIS 3 wurden von Christian Nguyen verfasst.

Ort, Datum Unterschrift

Ort, Datum Unterschrift

Ort, Datum Unterschrift