



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

Inhaltsverzeichnis

Management Summary	ii
Abkürzungsverzeichnis	v
Abbildungsverzeichnis	vi
1 Einleitung	1
1.1 Motivation	1
1.2 Ziel der Arbeit	1
1.3 Vorgehensweise	1
2 Anything as a Service - Cloud Computing	2
2.1 Definition	2
2.2 Typische Servicemodelle	2
2.2.1 IaaS: Infrastructure as a Service	2
2.2.2 SaaS: Software as a Service	2
2.2.3 PaaS: Plattform as a Service	2
2.3 Vor- und Nachteile	2
3 Knappheit von Grafikkarten	3
3.1 Preisentwicklung	4
3.2 Ursache Halbleitermangel und KryptoMining	5
4 Gaming as a Service	6
4.1 Funktionsweise	6
4.2 Anbietervergleich	6
4.2.1 Voraussetzung	6
4.2.2 Angebot	6
4.2.3 Preis	6
4.3 Hardwarevoraussetzung um Usability zu gewährleisten	6
5 GPU as a Service	7
5.1 Funktionsweise	7
5.2 Einsatzgebiete	8
5.3 Vergleich eigene GPU und GPU in der Cloud	9
6 Marktvorhersage	10

7 Fazit und Ausblick	11
Quellenverzeichnis	vii
Ehrenwörtliche Erklärung	viii

Abkürzungsverzeichnis

GPU Graphics-Processing-Unit oder Grafikkarte

HPC High-Performance-Computing

RAID redundant array of independent disk

Abbildungsverzeichnis

3.1	M. Kords (2022)	3
3.2	Voas, Kshetri und DeFranco (2021)	4
5.1	Wang u. a. (2017)	7

1 Einleitung

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

...

2 Anything as a Service - Cloud Computing

2.1 Definition

2.2 Typische Servicemodelle

2.2.1 IaaS: Infrastructure as a Service

2.2.2 SaaS: Software as a Service

2.2.3 PaaS: Plattform as a Service

2.3 Vor- und Nachteile

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.¹



Abbildung 3.1: Weltweite Lieferung von Halbleiterprodukten für die Automobilindustrie von 2011 bis 2021

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.²³⁴

In diesem Kapitel soll die Preisentwicklung von GPUs betrachtet werden, dabei wird ein Zusammenhang geschaffen mit den Ursachen die diese Preisentwicklung verursacht haben.

¹Vgl. Voas, Kshetri und DeFranco, "Scarcity and Global Insecurity: The Semiconductor Shortage".

²Vgl. McClean, *The 2022 McClean Report*.

³Vgl. Voas, Kshetri und DeFranco, "Scarcity and Global Insecurity: The Semiconductor Shortage".

⁴Kords, *Weltweite Lieferung von Halbleiterprodukten (integrated circuit) für die Automobilindustrie von 2011 bis 2021*, Vgl.

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 Umsatzzanstieg von 5,4% aufgezeigt.⁵

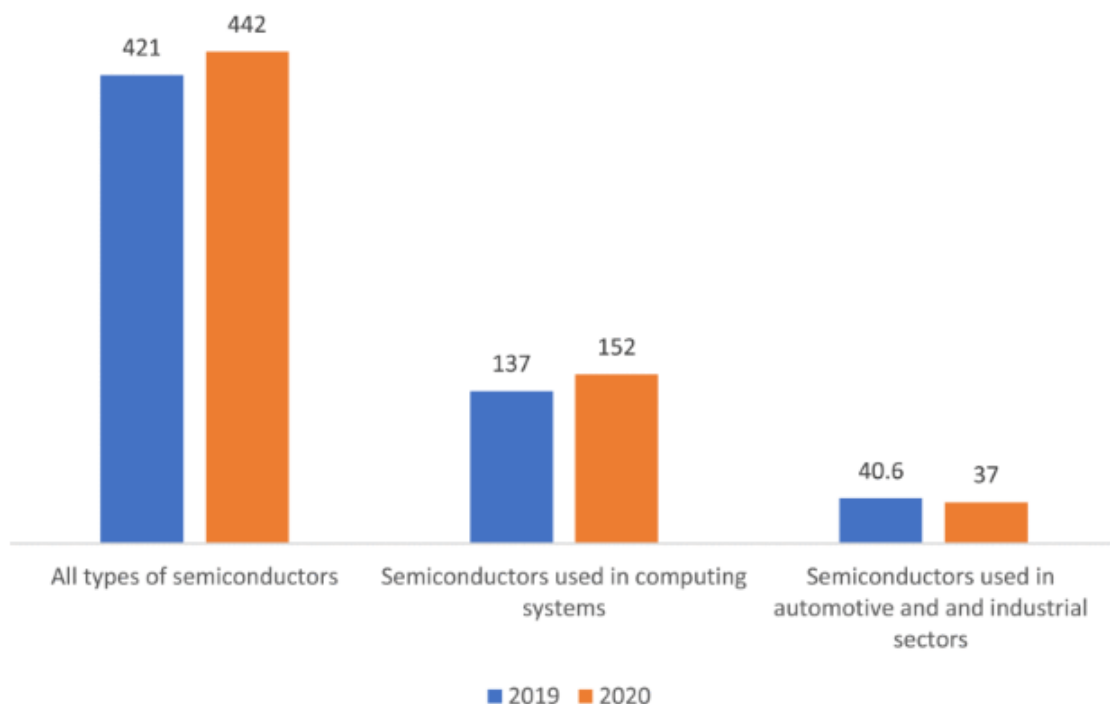


Abbildung 3.2: Worldwide semiconductor revenues in 2019 and 2020 (dollar, billions)

Wie in Abbildung 3.2 zu sehen ist der Umsatzzanstieg 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.⁶

⁵Vgl. Voas, Kshetri und DeFranco, "[Scarcity and Global Insecurity: The Semiconductor Shortage](#)".

⁶Vgl. [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 sich in den letzten Jahren Naturdesaster ereignet die Ebenfalls durch z.B. Dürre die Produktion Lahmgelegt haben.⁷

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.⁸

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.⁹

Da große Investitionen für das Kryptomining getätigt werden hat sich ein neuer Markt mit einer großen Nachfrage gebildet, der GPUs benötigt.

Allerdings gibt es Faktengestützte Prognosen, welche behaupten das der Halbleitermangel nicht zu lange anhalten wird. Nach Gartner war die Prognose auf das zweite Quartal 2022 geschätzt, auch wenn dies nicht im vollen Umfang eingetreten ist sind die Argumente für die weitere Zukunft nicht irrelevant.

Angeführt wird das durch den Halbleitermangel nun die Lieferkette enger überwacht wird. Daraus resultierend wird mehr Transparenz und Vorinvestitionen um die Lieferung garantieren zu können. Ebenfalls möchte man auch die Abnahme von mehreren Lieferanten bevorzugen, anstatt von einem Lieferanten abhängig zu sein. Diese Faktoren sollen langfristig Sicherheit in der Lieferkette schaffen.¹⁰

⁷Vgl. Voas, Kshetri und DeFranco, "[Scarcity and Global Insecurity: The Semiconductor Shortage](#)".

⁸Statista-Research-Department, [Volumen der weltweiten Investitionen in Blockchain-Technologien und Kryptowährungen von 2018 bis 2021](#).

⁹Vgl. Arslanian, [The Book of Crypto: The Complete Guide to Understanding Bitcoin, Cryptocurrencies and Digital Assets](#), S.259-273.

¹⁰Vgl. Rimol, [Gartner Says Global Chip Shortage Expected to Persist Until Second Quarter of 2022](#).

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 Endnutzers, 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

4.2.1 Voraussetzung

4.2.2 Angebot

4.2.3 Preis

4.3 Hardwarevoraussetzung um Usability zu gewährleisten

5 GPU as a Service

Außerhalb von Gaming as a Service gibt es weitere Service Möglichkeiten die ebenfalls GPU as a Service beanspruchen bzw. inkludieren. Wachsende Märkte dafür sind das rendern von 3D-Modellen und animierten Videos, wie auch deep learning model training für KIs.

Im folgenden Kapitel wird erläutert werden wie GPU as a Service Anbieter Ihren Service realisieren können und Qualität schaffen mit verschiedenen Methoden. 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 Privatsnutzung 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.¹¹

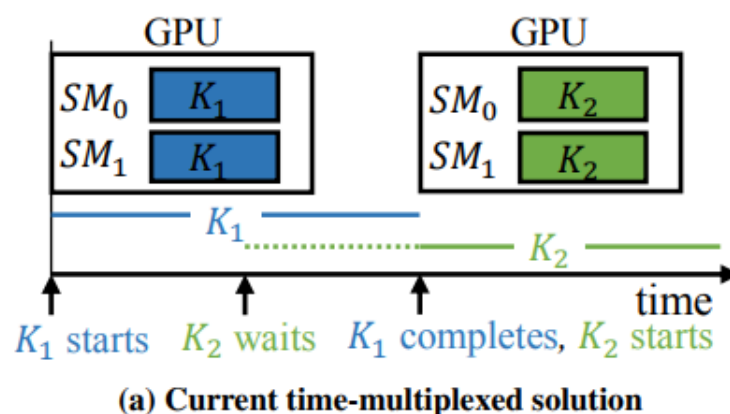


Abbildung 5.1: Current time-multiplexed solution

¹¹Vgl. Wang u. a., "Quality of Service Support for Fine-Grained Sharing on GPUs".

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.¹²

Ein anderer Ansatz um Konstanz zu schaffen ist das das man die Kernels über Software zu einen Kernel fusioniert. 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 Situation geben kann einen schlechten Kernel erwisch zu haben.¹³

Eine quasi gegenteilige Methode die Kernels in verschiedene Partitionen aufzuteilen, anstatt Sie zu fusionieren. Wiedermal vergleichbar wie man auch mit Festplatten umgehen kann. Die Leistung des Kernel wird in verschiedene Partitionen aufgeteilt. Diese können dann je nach Monetarisierungsmodell vermietet werden für einen Zeitraum oder auf einer Bedarfsbasis. Daraus besteht der Vorteil, dass Nutzer des Services Ihren Bedarf selbst definieren können und garantiert diese Leistung in Anspruch nehmen können.¹⁴

Es gibt neben diesen Methoden weitere Maßnahmen für die Qualitätssicherung des GPU as a Service Modells. Allerdings würden diese sich zu weit vertiefen und über dem Ziel dieser Arbeit hinaus gehen. Für eine weitere vertiefung ist die Arbeit von Wang u. a., "Quality of Service Support for Fine-Grained Sharing on GPUs". empfehlen.

5.2 Einsatzgebiete

Die Einsatzgebiete sind weitreichend und können auf zwei Bedarfsmethoden eingeteilt werden. Einmal in Schüben, in denen ein festes Ziel und der Aufwand definiert wird, welcher der Prozess erreichen soll. Beispiele dafür sind rendern von 3D-Modellen und animierten Videos.

Um dieses Beispiele auszuführen, wenn ein 3D-Modell dargestellt wird. Das 3D-Modell besteht aus Matrixoperationen die von GPUs in der Regel berechnet werden. Meist mit dem Ziel diese realitätsgeträu und logisch in einer Umgebung darzustellen. Dieses Beispiel lässt

¹²Vgl. Wang u. a., "Quality of Service Support for Fine-Grained Sharing on GPUs".

¹³Vgl. ebd.

¹⁴Vgl. ebd.

erweitern auf animierte Videos, bei diesen dann eine veränderte Version des Vorgängermodells oder ein vollkommen anderes Modell für jedes Bild das gezeigt wird.¹⁵

Ebenso ist es möglich, dass optimal mit einem Ziel aber unbekannten Aufwand, so ein Prozess ausgeführt wird. In diesem Fall wird eine konstante Rechenleistung benötigt. Beispiele dafür sind deep learning modelle für KIs oder Gaming as a Service.

Um hier Ebenfalls ein Beispiel anzuführen, wenn Gaming as a Service Dienste in Anspruch genommen. Bei diesem Service ist kein Ziel festgelegt, da auch nicht ein konkreter Prozess vorgegeben wird. Anhand der Eingaben des Nutzers variiert die benötigte Visualisierung und Spielelogik. Da die Eingaben nicht vorher konkret bekannt sind, wird besteht ein konstanter Bedarf an Rechenleistung um auf die Eingaben reagieren zu können. Die Rechenleistung wird benötigt bis der Nutzer das Spiel beendet.¹⁶¹⁷¹⁸

5.3 Vergleich eigene GPU und GPU in der Cloud

Nach den angeführten Informationen lässt sich folgendes feststellen. Der Bedarf an GPU Rechenleistung ist im konstanten Wachstum für verschiedene Märkte. Die GPU as a Service Anbieter können Ihre Systeme entsprechend dem Monetarisierungsmodell und den Leistungsprioritäten aufstellen. Dabei sind das keine neuen Methoden und vergleichbar mit der Massenspeicherverwaltung. Je nach geplanter Nutzung werden GPUs entweder in Schüben oder konstante Rechenleistung abverlangt.

Nach unserem Ermessen ist besonders die Entscheidung zwischen der eigenen GPU oder GPU as a Service abhängig von der benötigten Rechenleistung und der Verfügbarkeit. In der Annahme dass der Dienstleister des GPU as a Service permanent verfügbar ist. Sollten für Prozesse welche in Schüben erfolgen GPU as a Service ein attraktives Angebot sein, da diese mit der großen Rechenleistung diese Prozesse in kurzer Zeit abgeschlossen werden können.

Eine eigene GPU ist attraktiv eher im Fall von konstanter benötigter Rechenleistung. Wenn eine verfügbare GPU in der Lage die benötigte Rechenleistung zu erbringen. Andernfalls ist GPU as a Service ebenfalls eine Lösung, besonders in Zeiten in denen die Preise von GPUs mit dem Halbleitermangel gestiegen sind.

¹⁵Vgl. Loop und Blinn, "Real-time GPU rendering of piecewise algebraic surfaces".

¹⁶Vgl. Lattuada u. a., "Performance prediction of deep learning applications training in GPU as a service systems".

¹⁷Vgl. Wang u. a., "Quality of Service Support for Fine-Grained Sharing on GPUs".

¹⁸Vgl. Loop und Blinn, "Real-time GPU rendering of piecewise algebraic surfaces".

6 Marktvorhersage

7 Fazit und Ausblick

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.
- Kords, Martin. *Weltweite Lieferung von Halbleiterprodukten (integrated circuit) für die Automobilindustrie von 2011 bis 2021*. URL: <https://de.statista.com/statistik/daten/studie/1288183/umfrage/halbleiterlieferungen-fuer-kraftfahrzeuge/> (besucht am 03. 06. 2022).
- 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).
- Rimol, Meghan. *Gartner Says Global Chip Shortage Expected to Persist Until Second Quarter of 2022*. URL: <https://www.gartner.com/en/newsroom/press-releases/2021-05-12-gartner-says-global-chip-shortage-expected-to-persist-until-second-quarter-of-2022> (besucht am 09. 06. 2022).
- Statista-Research-Department. *Volumen der weltweiten Investitionen in Blockchain-Technologien und Kryptowährungen von 2018 bis 2021*. URL: <https://de.statista.com/statistik/daten/studie/1198230/umfrage/weltweite-investitionen-in-blockchain-technologien-und-kryptowaehrungen/> (besucht am 03. 06. 2022).
- 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.“

Ort, Datum

Unterschrift

Ort, Datum

Unterschrift

Ort, Datum

Unterschrift