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

Abkürzungsverzeichnis

GPU Graphics-Processing-Unit oder Grafikkarte

HPC High-Performance-Computing

RAID redundant array of independent disk

Abbildungsverzeichnis

1 Einleitung

1.1 Motivation

Durch den aktuell anhaltende Halbleitermangel besteht ein Engpass an Ressourcen von dem die meisten Wirtschaftszweige betroffen sind. Einer dieser Wirtschaftszweige ist die Produktion von GPUs (graphics processing unit). Diese werden für diverse Prozesse in Computern verwendet, im betrieblichen wie auch im privaten Bereich. Primär in dieser Arbeit werden die Bereiche High-Performance-Computing (HPC) und Gaming betrachtet. Beide diese Bereiche benötigen GPU-Rechenleistung, welche 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 wird dem Mangel entgegengewirkt durch die Schaffung einer Alternative für den Bedarf.

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]

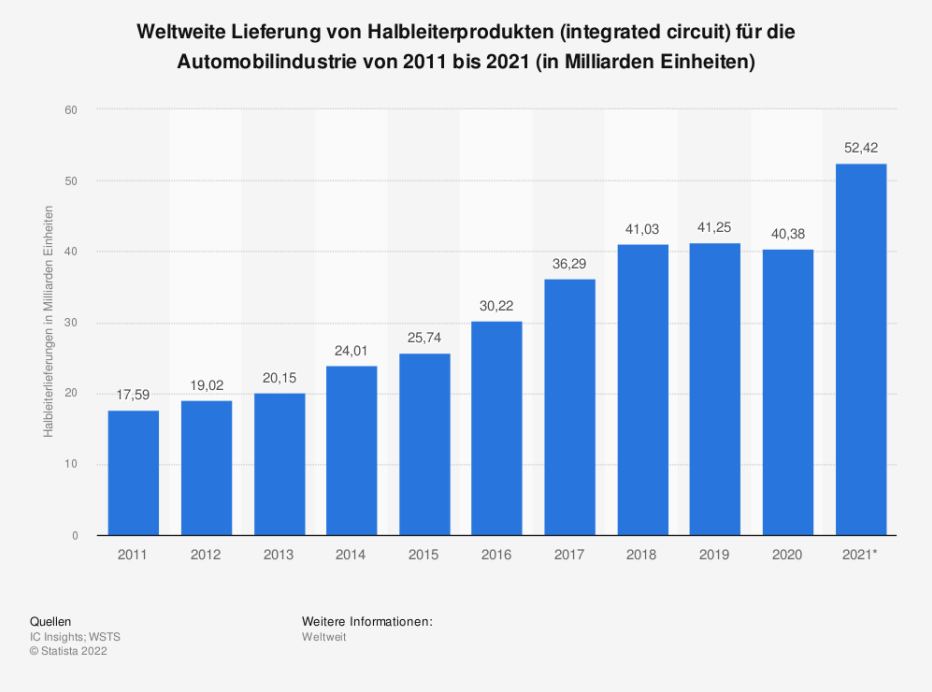


Abbildung 3.1: Weltweite Lieferung von Halbleiterprodukten für die Automobilindustrie von 2011 bis 2021

Im Vergleich zur Lage im Jahr 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.[2][3][4]
In diesem Kapitel soll die Preisentwicklung von GPUs betrachtet werden, dabei wird ein Zusammenhang geschaffen mit den Ursachen die diese Preisentwicklung verursacht haben.

[1] Vgl. Voas, Kshetri und DeFranco, "Scarcity and Global Insecurity: The Semiconductor Shortage".
[2] Vgl. McClean, *The 2022 McClean Report*.
[3] Vgl. Voas, Kshetri und DeFranco, "Scarcity and Global Insecurity: The Semiconductor Shortage".
[4] Kords, *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 Umsatzanstieg von 5,4 % aufgezeigt.[5]

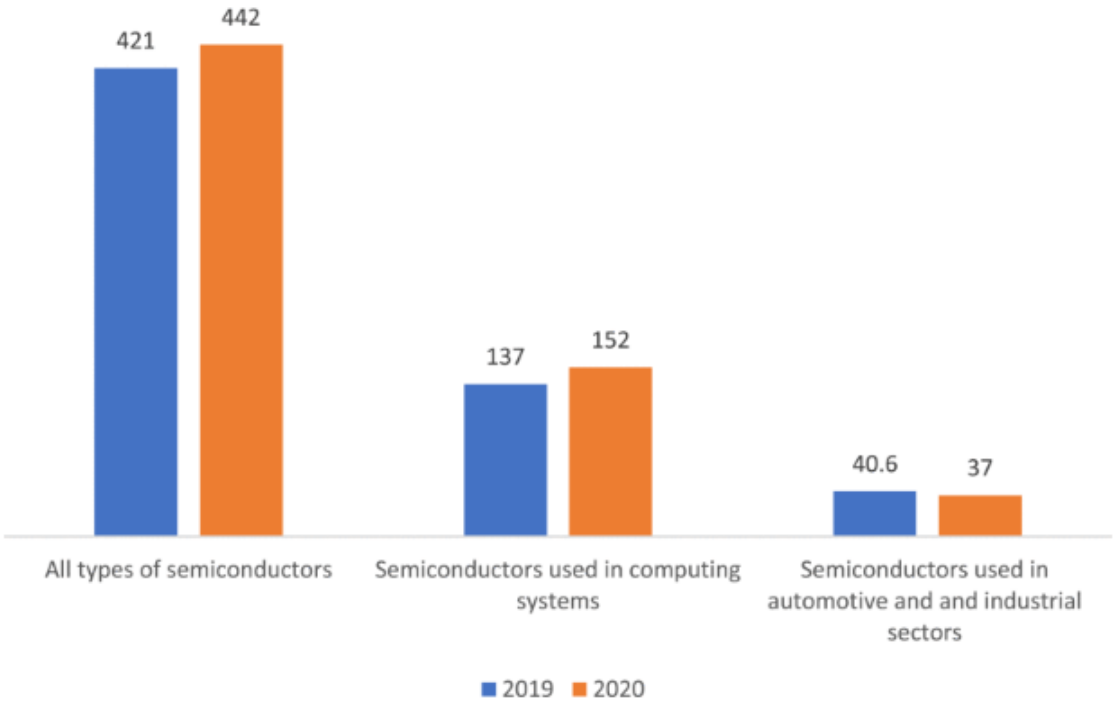


Abbildung 3.2: Worldwide semiconductor revenues in 2019 and 2020 (dollar, billions)

Wie in Abbildung 3.2 zu sehen ist, 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 sich viele Menschen an Home-Schooling und Home-Office anpassen, 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 in der Automobilindustrie erklären. Dennoch ist damit insgesamt der Bedarf an Halbleitern gestiegen.[6]

[5] Vgl. Voas, Kshetri und DeFranco, "Scarcity and Global Insecurity: The Semiconductor Shortage".
[6] 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 sich in den letzten Jahren Naturkatastrophen ereignet, die z.B. durch Dürre, die Produktion lahmgelegten.[7]
Durch mangelnde Produktion und steigenden Bedarf hat sich nun ein sehr hoher Marktpreis entwickelt. Um diesem entgegenzuwirken ist nicht nur Ressourcenverfügbarkeit zu schaffen, sondern auch eine effizientere Methode zur Nutzung der Ressourcen.

3.2 Ursache Halbleitermangel und Krypto-Mining

Um den Halbleitermangel besser zu verstehen, sollte 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 zu verzeichnen.[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 faktengestütze Prognosen, welche behaupten, dass der Halbleitermangel nicht zu lange anhalten wird. Nach Gartners Prognose für das zweite Quartal 2022 sollte sich der Halbleitermangel verringern, auch wenn dies nicht im vollen Umfang eingetreten ist, sind die Argumente für die weitere Zukunft nicht irrelevant.
Angeführt wird, dass durch den Halbleitermangel nun die Lieferkette enger überwacht wird. Daraus resultierend werden mehr Transparenz und Vorinverstionen geschaffen, um die Lieferungen 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 bieten.[10]

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

4.2.1 Voraussetzung

4.2.2 Angebot

4.2.3 Preis

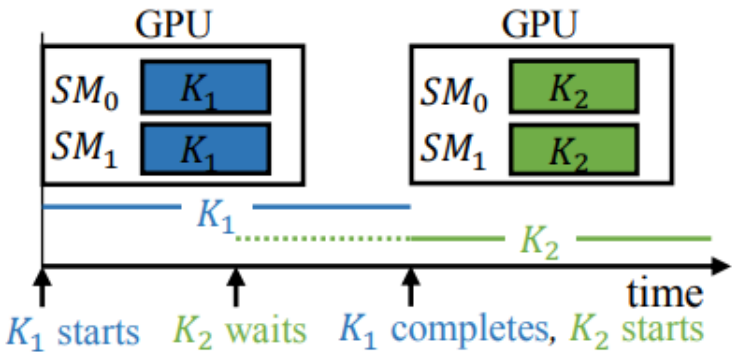
4.3 Hardwarevorraussetzung um Usability zu gewährleisten

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 mit verschiedenen Methoden schaffen. Wie auch Einsatzgebiete für Cloud-Lösungen festgestellt und identifiziert werden.

5.1 Funktionsweise

Bei der Dienstleistungsinanspruchnahme werden die von einem Nutzer 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 ein zentrales Betriebssystem, welches auch als Kernel bezeichnet. Üblicherweise wird dies skalierbar angewendet mit einer Vielzahl an Kernels, welche eine Vielzahl an GPUs besitzen.[11]



(a) Current time-multiplexed solution

Abbildung 5.1: Current time-multiplexed solution

Es gibt verschiedene Methoden, diese Prozesse zu verarbeiten. Eine Methode davon ist

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

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, dass Prozesse Kernels unnötig blockieren können.[12]

Ein anderer Ansatz, um Konstanz zu schaffen, ist, dass man die Kernels über Software zu einem Kernel fusioniert. Das ist vergleichbar mit dem für Festplatten verwendeten "redundant array of independent diskSystem, auch abgekürzt RAID genannt, welches gängiger bekannt ist. Durch das Fusionieren der Kernels über die Software kann eine konstante Leistung gewährleistet werden. Ebenfalls, um eine Fairness zwischen allen Leistungsabnehmern zu schaffen, da es keine Situation geben kann, einen leistungsschwächeren Kernel zugewiesen zu bekommen.[13]

Eine quasi gegenteilige Methode, die Kernels in verschiedene Partitionen aufzuteilen, anstatt Sie zu fusionieren. Wieder mal 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 entsteht 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 das Ziel dieser Arbeit hinausgehen. Für eine weitere Vertiefung ist die Arbeit von Wang u.a., “Quality of Service Support for Fine-Grained Sharing on GPUs”. zu 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 werden, welche der Prozess erreichen soll. Beispiele dafür sind Rendern von 3D-Modellen und animierten Videos.
Um dieses Beispiel auszuführen: Wenn ein 3D-Modell dargestellt wird, besteht es aus Matrixoperationen, die von GPUs in der Regel berechnet werden. Dies hat meist das Ziel, diese realitätsgetreu und logisch in einer Umgebung darzustellen. Dieses Beispiel lässt

[12] Vgl. Wang u. a., “Quality of Service Support for Fine-Grained Sharing on GPUs”.
[13] Vgl. ebd.
[14] Vgl. ebd.

sich auf animierte Videos erweitern, bei denen dann eine veränderte Version des Vorgängermodells oder ein vollkommen anderes Modell für jedes Bild im Video dargestellt wird.[15]

Ebenso ist es möglich, dass mit einem optionalen 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 auch hier ein Beispiel anzuführen, falls Gaming as a Service-Dienste in Anspruch genommen wird. Bei diesem Service ist kein Ziel festgelegt, da auch kein konkreter Prozess vorgegeben ist. Anhand der Eingaben des Nutzers variiert die benötige Visualsierung und Spielelogik. Da die Eingaben nicht vorher konkret bekannt sind, besteht ein konstanter Bedarf an Rechenleistung um auf die Eingaben reagieren zu können. Die Rechenleistung wird benötigt bis der Nutzer das Spiel beendet.[16][17][18]

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 handelt es sich nicht um neue Methoden, sondern um etwas mit der Massenspeicherverwaltung Vergleichbaren. Je nach geplanter Nutzung werden GPUs entweder in Schüben oder mit konstanter Rechenleistung beansprucht.

Nach unserem Ermessen spielt besonders bei der Entscheidung zwischen der eigenen GPU und GPU as a Service die benötigte Rechenleistung und die Verfügbarkeit des Services eine Rolle. Unter der Annahme, dass der Dienstleister des GPU as a Serive permanent verfügbar ist. Für Prozesse, welche in Schüben erfolgen und große Rechenleistung benötigen, ist GPU as a Service ein attraktives Angebot. Diese können dadurch in kurzer Zeit abgeschlossen werden.

Eine eigene GPU ist attraktiver im Fall von konstant benötigter Rechenleistung, wenn eine verfügbare GPU in der Lage ist, 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.

[15] Vgl. Loop und Blinn, “Real-time GPU rendering of piecewise algebraic surfaces”.
[16] Vgl. Lattuada u. a., “Performance prediction of deep learning applications training in GPU as a service systems”.
[17] Vgl. Wang u. a., “Quality of Service Support for Fine-Grained Sharing on GPUs”.
[18] Vgl. 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