(19) Deutsches Patent- und Markenamt

(10) **DE 10 2021 206 331 A1** 2021.12.23

(12) # Offenlegungsschrift

(21) Aktenzeichen: **10 2021 206 331.4**
(22) Anmeldetag: **21.06.2021**
(43) Offenlegungstag: **23.12.2021**

(51) Int Cl.: ***G06T 7/33*** (2017.01)
G06T 1/40 (2006.01)

(30) Unionspriorität:
16/908,397 **22.06.2020** **US**

(71) Anmelder:
NVIDIA Corporation, Santa Clara, CA, US

(74) Vertreter:
Kraus & Weisert Patentanwälte PartGmbB, 80539 München, DE

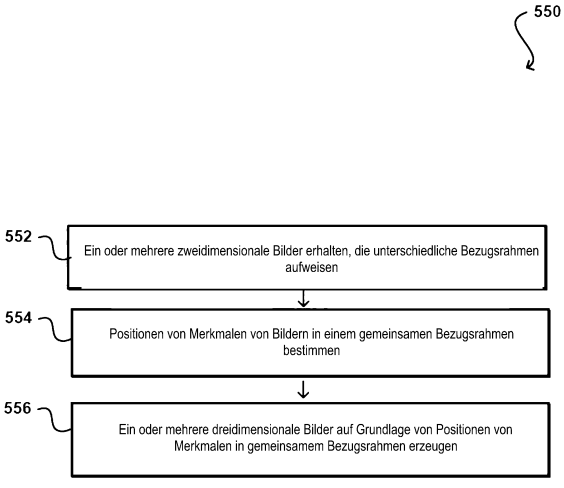
(72) Erfinder:
Philion, Jonah, Santa Clara, CA, US; Fidler, Sanja, Santa Clara, CA, US

Prüfungsantrag gemäß § 44 PatG ist gestellt.

Die folgenden Angaben sind den vom Anmelder eingereichten Unterlagen entnommen.

(54) Bezeichnung: **BILDERZEUGUNG UNTER VERWENDUNG VON EINEM ODER MEHREREN NEURONALEN NETZWERKEN**

(57) Zusammenfassung: Es werden Einrichtungen, Systeme und Methoden vorgestellt, um ansichtspezifische Darstellungen eines Objektes oder einer Umgebung zu erzeugen. In mindestens einer Ausführungsform werden ein oder mehrere neuronale Netzwerke verwendet, um ein oder mehrere Bilder mindestens teilweise auf Grundlage von zwei- oder mehr zweidimensionalen (2D-)Bildern zu erzeugen, die unterschiedliche Bezugsrahmen aufweisen.



Beschreibung

GEBIET

[0001] Mindestens eine Ausführungsform betrifft Verarbeitungsressourcen, die verwendet werden, um künstliche Intelligenz auszuführen und zu unterstützen. Zum Beispiel bezieht sich mindestens eine Ausführungsform auf Prozessoren oder Computersysteme, die verwendet werden, um neuronale Netzwerke gemäß verschiedenen in dieser Schrift beschriebenen neuartigen Methoden zu trainieren.

ALLGEMEINER STAND DER TECHNIK

[0002] Automation wird verwendet, um eine stetig wachsende Vielfalt von Aufgaben durchzuführen. In vielen Fällen stützt sich diese Automation auf Kameras und Sensoren, um Informationen über seine Umgebung zu erhalten, was dieser Automation dabei helfen kann, Bewegungsbahnen zu bestimmen und Zusammenstöße zu vermeiden. Bestehende Systeme nutzen typischerweise eine feststehende Anzahl und Ausrichtung von Kameras und Sensoren, die außerdem Tiefeninformationen bereitstellen können. Eine stetig wachsende Vielfalt von Automation geht mit einer stetig wachsenden Vielfalt der Anzahlen und Platzierungen von Kameras und Sensoren einher, die außerdem Tiefeninformationen bereitstellen können oder nicht, was dann erfordern würde, dass ein individualisiertes System für jede Art von Automation entwickelt werden müsste. Für viele Anwendungen kann eine derartige Individualisierung kostspielig sein oder kann mindestens eine Umsetzung verzögern und zusätzliche Instandhaltung und zusätzlichen Aufwand erfordern. Ein derartiges System müsste für jegliche Änderungen an den Kameras oder Sensoren aktualisiert werden.

Figurenliste

[0003] Ausführungsformen gemäß der vorliegenden Offenbarung werden unter Bezugnahme auf die Zeichnungen beschrieben, in denen Folgendes gilt:

Die **Fig. 1A**, **Fig. 1B**, **Fig. 1C**, **Fig. 1D**, **Fig. 1E** und **Fig. 1F** veranschaulichen Bilder gemäß mindestens einer Ausführungsform;

Fig. 2 veranschaulicht ein Automationssystem gemäß mindestens einer Ausführungsform;

Fig. 3 veranschaulicht Komponenten eines Bildansichtsgenerators gemäß mindestens einer Ausführungsform;

die **Fig. 4A**, **Fig. 4B**, **Fig. 4C** und **Fig. 4D** veranschaulichen Daten für Bilder in einem gemeinsamen Bezugsrahmen gemäß mindestens einer Ausführungsform;

die **Fig. 5A** und **Fig. 5B** veranschaulichen Prozesses zum Erzeugen einer Bildansicht gemäß mindestens einer Ausführungsform;

Fig. 6A veranschaulicht Ableitungs- und/oder Trainingslogik gemäß mindestens einer Ausführungsform;

Fig. 6B veranschaulicht Ableitungs- und/oder Trainingslogik gemäß mindestens einer Ausführungsform;

Fig. 7 veranschaulicht ein beispielhaftes Rechenzentrumssystem gemäß mindestens einer Ausführungsform;

Fig. 8 veranschaulicht ein Computersystem gemäß mindestens einer Ausführungsform;

Fig. 9 veranschaulicht ein Computersystem gemäß mindestens einer Ausführungsform;

Fig. 10 veranschaulicht ein Computersystem gemäß mindestens einer Ausführungsform;

Fig. 11 veranschaulicht ein Computersystem gemäß mindestens einer Ausführungsform;

Fig. 12A veranschaulicht ein Computersystem gemäß mindestens einer Ausführungsform;

Fig. 12B veranschaulicht ein Computersystem gemäß mindestens einer Ausführungsform;

Fig. 12C veranschaulicht ein Computersystem gemäß mindestens einer Ausführungsform;

Fig. 12D veranschaulicht ein Computersystem gemäß mindestens einer Ausführungsform;

die **Fig. 12E** und **Fig. 12F** veranschaulichen ein gemeinsam genutztes Programmiermodell gemäß mindestens einer Ausführungsform;

Fig. 13 veranschaulicht beispielhafte integrierte Schaltungen und verbundene Grafikprozessoren gemäß mindestens einer Ausführungsform;

die **Fig. 14A-14B** veranschaulichen beispielhafte integrierte Schaltungen und verbundene Grafikprozessoren gemäß mindestens einer Ausführungsform;

die **Fig. 15A-15B** veranschaulichen zusätzliche beispielhafte Grafikprozessorlogik gemäß mindestens einer Ausführungsform;

Fig. 16 veranschaulicht ein Computersystem gemäß mindestens einer Ausführungsform;

Fig. 17A veranschaulicht einen Parallelprozessor gemäß mindestens einer Ausführungsform;

Fig. 17B veranschaulicht eine Partitionseinheit gemäß mindestens einer Ausführungsform;

Fig. 17C veranschaulicht einen Verarbeitungscluster gemäß mindestens einer Ausführungsform;

Fig. 17D veranschaulicht einen Grafikmultiprozessor gemäß mindestens einer Ausführungsform;

Fig. 18 veranschaulicht ein System mit Multigrafikverarbeitungseinheit (graphics processing unit - GPU) gemäß mindestens einer Ausführungsform;

Fig. 19 veranschaulicht einen Grafikprozessor gemäß mindestens einer Ausführungsform;

Fig. 20 veranschaulicht eine Mikroarchitektur eines Prozessors gemäß mindestens einer Ausführungsform;

Fig. 21 veranschaulicht einen Deep-Learning-Anwendungsprozessor gemäß mindestens einer Ausführungsform;

Fig. 22 veranschaulicht einen beispielhaften neuromorphen Prozessor gemäß mindestens einer Ausführungsform;

die **Fig. 23** und **Fig. 24** veranschaulichen mindestens Teile eines Grafikprozessors gemäß mindestens einer Ausführungsform;

Fig. 25 veranschaulicht mindestens Teile eines Grafikprozessorkerns gemäß mindestens einer Ausführungsform;

die **Fig. 26A-26B** veranschaulichen mindestens Teile eines Grafikprozessorkerns gemäß mindestens einer Ausführungsform;

Fig. 27 veranschaulicht eine Parallelverarbeitungseinheit (parallel processing unit - „PPU“) gemäß mindestens einer Ausführungsform;

Fig. 28 veranschaulicht einen allgemeinen Verarbeitungscluster (general processing cluster - „GPC“) gemäß mindestens einer Ausführungsform;

Fig. 29 veranschaulicht eine Speicherpartitionseinheit einer Parallelverarbeitungseinheit („PPU“) gemäß mindestens einer Ausführungsform;

Fig. 30 veranschaulicht einen Streaming-Multiprozessor gemäß mindestens einer Ausführungsform;

Fig. 31 ist ein beispielhaftes Datenflussdiagramm für eine fortschrittliche Rechenpipeline gemäß mindestens einer Ausführungsform;

Fig. 32 ist ein Systemdiagramm für ein beispielhaftes System zum Trainieren, Anpassen, Instanziieren und Umsetzen von Modellen des maschinellen Lernens in einer fortschrittlichen Rechenpipeline gemäß mindestens einer Ausführungsform;

Fig. 33A veranschaulicht ein Datenflussdiagramm für einen Prozess zum Trainieren eines Modells des maschinellen Lernens gemäß mindestens einer Ausführungsform;

Fig. 33B ist eine beispielhafte Darstellung einer Client-Server-Architektur zum Verbessern von Anmerkungswerkzeugen mit vortrainierten Anmerkungsmodellen gemäß mindestens einer Ausführungsform;

Fig. 34A veranschaulicht ein Beispiel eines autonomen Fahrzeugs gemäß mindestens einer Ausführungsform;

Fig. 34B veranschaulicht ein Beispiel von Kamerapositionen und -sichtfeldern für das autonome Fahrzeug aus **Fig. 34A** gemäß mindestens einer Ausführungsform;

Fig. 34C veranschaulicht eine beispielhafte Systemarchitektur für das autonome Fahrzeug aus **Fig. 34A** gemäß mindestens einer Ausführungsform; und

Fig. 34D veranschaulicht ein System zur Kommunikation zwischen einem oder mehreren cloudbasierten Servern und dem autonomen Fahrzeug aus **Fig. 34A** gemäß mindestens einer Ausführungsform; und

DETAILLIERTE BESCHREIBUNG

[0004] In mindestens einer Ausführungsform können Informationen zu einem Schauplatz, einem Objekt oder einer Umgebung unter Verwendung von einem oder mehreren aufgenommenen Bildern bestimmt werden, die verschiedene Ansichten dieses Schauplatzes, dieses Objektes oder dieser Umgebung zeigen. In mindestens einer Ausführungsform können diese Bilder zweidimensionale Bilder sein, die durch eine(n) oder mehrere Kameras oder Sensoren aufgenommen wurden. In mindestens einer Ausführungsform kann eine Vorrichtung, wie etwa ein Kamerarigg, eine willkürliche oder veränderliche Anzahl von Kameras (oder Sensoren oder andere bildaufnehmende Elemente) an verschiedenen Positionen und in verschiedenen Ausrichtungen beinhalten. In mindestens einer Ausführungsform nehmen diese Kameras jeweils eine entsprechende Ansicht auf, wie in den aufgenommenen Bildern **100**, **110**, **120** veranschaulicht, die in den **Fig. 1A**, **Fig. 1B** und **Fig. 1C** veranschaulicht sind. In mindestens einer Ausführungsform veranschaulichen diese Bilder jeweils eine andere zweidimensionale Ansicht einer dreidimensionalen Umgebung, in der sich die Vorrichtung befindet. In mindestens einer Ausführungsform kann es wünschenswert sein, dreidimensionale Informationen über diese Umgebung unter Verwendung dieser zweidimensionalen Bilder zu bestimmen. In mindestens einer Ausführungsform können diese Bilder **100**, **110**, **120** unter Verwendung eines oder mehrerer neuronaler Netzwerke verarbeitet werden, um mindestens ein Bild **130** zu erzeugen, das Objekte in dieser Umgebung von einem einzelnen oder gemeinsamen Koordinatensystem oder Bezugsrahmen zeigt. In mindestens einer Ausführungsform können ein oder mehrere neuronale Netzwerke Merkmale dieser Bilder nutzen, um außerdem mindestens ein Bild **140** zu erzeugen, das Tiefeninformationen beinhaltet, die von diesen zweidimensionalen Bildern abgeleitet wurden, und dieser Ansicht in einem einzelnen Bezugsrahmen entspricht. In mindestens einer Ausführungsform können diese Tiefeninformationen für einen einzelnen Bezugsrahmen verwendet werden, um eine dreidimensionale Darstellung dieser Umgebung zu erzeugen. In mindestens einer Ausführungsform kann dies eine Punktwolke in einem gemeinsamen Koordinatensystem beinhalten, bei dem die Positionen dieser Punkte durch das Platzieren dieser zweidimensionalen Bilder in dem dreidimensionalen Koordinatensystem bestimmt werden. In mindestens einer Ausführungsform können diese dreidimensionalen Informationen verwendet werden, um eine oder mehrere Ansichten dieser Umgebung zu erzeugen, wie etwa eine Draufsicht **150** oder „Vogelperspektive“, welche die Positionen verschiedener Objekte (z. B. die Fahrzeuge **152**) in Bezug auf diese Vorrichtung oder diesen Satz von Kameras zeigt, die diese Bilder aufnehmen. In mindestens einer Ausführungsform kann einer Vorrichtung oder einem System, das steuerbar oder mindestens teilweise automatisiert ist, wie etwa einem Roboter, einem unbemannten Luftfahrzeug oder einem autonomen Fahrzeug, eine derartige Ansicht bereitgestellt werden.

[0005] In mindestens einer Ausführungsform kann es unterschiedliche Arten von Kameras **206**, **208**, **212** oder Bildgebungssensoren **210** in einem Kamerarigg **204** oder einer Baugruppe eines Automationssystems **202** geben, wie etwa mit einer Vorrichtung oder einem System beinhaltet oder mit diesem in Verbindung stehen kann, die/das durch dieses Automationssystem gesteuert werden soll. In mindestens einer Ausführungsform können diese Kameras eine willkürliche Anzahl aufweisen oder können pro Rigg, Baugruppe oder eine andere derartige Gruppierung variieren. In mindestens einer Ausführungsform kann dies Kameras und Sensoren mit unterschiedlichen Auflösungen, Farbtiefen, optischen Fähigkeiten oder anderen derartigen Aspekten beinhalten. In mindestens einer Ausführungsform können eine oder mehrere dieser Kameras oder Sensoren außerdem dreidimensionale Daten oder Tiefeninformationen bereitstellen. In mindestens einer Ausführungsform können diese Kameras und Sensoren an einer Reihe von möglichen Positionen und mit einer Reihe von möglichen Ausrichtungen positioniert sein. In kann dies dazu führen, dass unterschiedliche Kameras oder Sensoren Daten in unterschiedlichen Koordinatensystemen oder Bezugsrahmen bereitstellen. In mindestens einer Ausführungsform können Bilder (oder Videos oder andere Bilddaten), die durch diese Bilder aufgenommen wurden, einem Bildverarbeitungssystem **214** bereitgestellt werden. In kann dieses Bildverarbeitungssystem **214** ein oder mehrere neuronale Netzwerke zum Analysieren dieser Bilder und zum Erzeugen eines oder mehrerer Bilder oder einer oder mehreren Darstellungen beinhalten. In mindestens einer Ausführungsform können diese Darstellungen in einem einzelnen oder einem gemeinsamen Koordinatensystem oder Bezugsrahmen erfolgen. In mindestens einer Ausführungsform kann dies durch das Bestimmen von Positionen dieser zweidimensionalen Bilder in diesem gemeinsamen Bezugsrahmen und dann das Ableiten von Projektionen von Merkmalen dieser Bilder in drei Dimensionen in diesem Bezugsrahmen erreicht werden, wie unter Verwendung von einem oder mehreren neuronalen Netzwerken **216** erreicht werden kann. In mindestens einer Ausführungsform können Kalibrierungsparameter, wie etwa die Intrinsik und die Extrinsik, die für jede dieser Kameras und Senso-

ren **206**, **208**, **210**, **212** bekannt sind, durch diese neuronalen Netzwerke bei dem Platzieren und Projizieren dieser Punkte in einem gemeinsamen Bezugsrahmen genutzt werden. In mindestens einer Ausführungsform stellen die Kameraextrinsik, oder extrinsischen Parameter, eine Position einer Kamera in einem dreidimensionalen Schauplatz dar, und können die Drehung und Verschiebung beinhalten, während die Kameraintrinsik, oder intrinsischen Parameter, Aspekte, wie etwa einen optischen Mittelpunkt und Brennweite einer Kamera, darstellen können.

[0006] In mindestens einer Ausführungsform können Bilder oder Daten, die in diesem Bezugsrahmen erzeugt wurden, einem Steuersystem **220** bereitgestellt werden, um Maßnahmen zu bestimmen, die durch ein oder mehrere steuerbare Teilsysteme **220** dieses Automationssystems **202** unternommen werden sollen. In mindestens einer Ausführungsform kann dies das Bestimmen von Positionen von Objekten in einer Umgebung und das Bestimmen einer oder mehrerer Bewegungen eines Systems beinhalten, um einem oder mehreren dieser Objekte auszuweichen oder mit diesen zu interagieren. In mindestens einer Ausführungsform können Vorrichtungsdaten **224** verwendet werden, um dem Steuersystem **220** zu ermöglichen, diese Objektpositionsdaten zu verwenden, um diesen steuerbaren Teilsystemen **220** geeignete Anweisungen bereitzustellen, um eine oder mehrere Aufgaben durchzuführen. In kann mindestens ein Teil dieser Bildverarbeitung oder Steuerungsbestimmung durch einen Cloudserver **280** oder ein anderes derartiges System oder einen anderen derartigen Dienst durchgeführt werden, wie unter Verwendung mindestens eines Netzwerks **240**, wie etwa einem drahtgebundenen oder drahtlosen Netzwerk kommuniziert werden kann. In mindestens einer Ausführungsform kann ein Cloudserver eine Steueranwendung **284** beinhalten, die mindestens einen Teil dieser Aufgaben durchführen kann, was einen Sitzungsverwalter zum Verwalten einer Steuersitzung für das Automationssystem **202**, einen Vorrichtungsverwalter **290**, der Informationen für das Automationssystem **202** speichert, eine Bildverarbeitungseinrichtung **292** zum Verarbeiten von Bilddaten von dem Automationssystem **202** und einen Steuerverwalter **288** zum Erzeugen von Steueranweisungen auf Grundlage dieser aufgenommenen und verarbeiteten Bilddaten beinhalten kann. In mindestens einer Ausführungsform kann der Cloudserver **280** außerdem Vorrichtungsdaten **294** und Benutzerdaten **292** speichern, um diese und andere derartige Aufgaben zu unterstützen. In mindestens einer Ausführungsform nutzt das Automationssystem **202** möglicherweise den Cloudserver **280** für keine dieser Bestimmungen, kann aber aus anderen Gründen, wie etwa zum Protokollieren, Speichern von Verlaufsdaten oder zur Problembehandlung, Daten über das Netzwerk **240** übermitteln.

[0007] In mindestens einer Ausführungsform ist ein wichtiges Anliegen eines Automationssystems, wie etwa eines autonomen Fahrzeugs, ist das Verschmelzen semantischer Darstellungen, die aus Sensorrohdaten gewonnen wurden, zu einer einzelnen „Vogelperspektiven“-Koordinatenrahmen zum Verbrauch durch Bewegungsplanung. In mindestens einer Ausführungsform kann die Ende-zu-Ende-Architektur **300** wie in **Fig. 3** veranschaulicht genutzt werden, die eine Draufsicht oder Vogelperspektivendarstellung von durch einen Schauplatz vorgegebenen Bilddaten aus einer willkürlichen Anzahl von Kameras oder Sensoren erzeugen kann. In mindestens einer Ausführungsform kann dies das einzelne Heben jedes Bildes in einen Kegelstumpf von Merkmalen für jede Kamera beinhalten. In mindestens einer Ausführungsform können diese Kegelstümpfe dann in einem gerasterten Vogelperspektivenraster platziert oder in diesen projiziert werden. In mindestens einer Ausführungsform kann das Trainieren einer gesamten Kamerabaugruppe ermöglichen, dass nicht nur das Darstellen von Bildern erlernt wird, sondern, wie man Vorhersagen von einer beliebigen oder allen verfügbaren Kameras in eine einzelne zusammenhängende Darstellung eines Schauplatzes verschmilzt, während es widerstandsfähig gegenüber Kalibrierungsfehlern ist. In mindestens einer Ausführungsform kann eine Draufsicht für Aufgaben, wie etwa Objektsegmentierung und Kartensegmentierung, nützlich sein. In mindestens einer Ausführungsform ermöglichen derartige Darstellungen, die von einem Modell abgeleitet wurden, wie etwa einem Deep-Learning-Modell, eine Ende-zu-Ende-Bewegung, die unter Verwendung von Expertenfahrbahnverläufe trainiert werden kann.

[0008] In mindestens einer Ausführungsform kann die Architektur des neuronalen Netzwerks **300** eines Ansichtsgenerators **308** genutzt werden, der zwei große Netzwerkbackbones beinhaltet, wie in **Fig. 3** veranschaulicht. In mindestens einer Ausführungsform können diese Backbones Merkmalsextrahierernetzwerke umfassen, wie etwa neuronale Faltungsnetzwerke (convolutional neural networks - CNNs), die Merkmale aus einer willkürlichen Anzahl n von Eingabebildern **302**, **304**, **306** berechnen können, wie sie etwa durch einen Dekodierer hochgerechnet werden können, um Masken oder Darstellungen zu erzeugen. In mindestens einer Ausführungsform kann ein erstes Backbone ein Merkmalsextrahierernetzwerk **310** beinhalten, das an jedem Bild einzeln arbeiten kann, um eine Punktwolke, die aus jedem Eingabebild **302** erzeugt wurde, mit Merkmalen auszustatten. In mindestens einer Ausführungsform kann ein zweites Backbone an dieser Punktwolke arbeiten, sobald diese Wolke in Säulen in einem Bezugsrahmen abgebildet wird. In mindestens einer Ausführungsform kann zu diesen beiden Netzwerken eine gemeinsame Schicht hinzukommen, wie an anderer Stelle in dieser Schrift erörtert. In mindestens einer Ausführungsform kann ein erstes Merkmalsextrahierernetzwerk **310**

an jedem Bild arbeiten und kann Schichten von einem Netzwerk nutzen, wie etwa einem auf ImageNet für Modelle, die Ausgangswerte beinhalten, vortrainierten EfficientNet-B0. In mindestens einer Ausführungsform kann ein ResNet oder ein ähnliches Netzwerk anstelle eines EfficientNet verwendet werden. In mindestens einer Ausführungsform kann ein zweites Netzwerk, das in dieser Schrift als ein Vogelperspektivennetzwerk bezeichnet ist, eine Kombination aus ResNet-Blöcken nutzen. In mindestens einer Ausführungsform kann nach einer Faltung mit dem Kernel 7 und der Schrittbreite 2, gefolgt von einer Batchnormalisierung und einer ReLu oder einer anderen Aktivierung, erste 3 Metaschichten von ResNet-18 durchquert werden, um ein oder mehrere Bilder **314** mit konkreten Ansichten zu erhalten, die drei unterschiedliche Vogelperspektivendarstellungen bei unterschiedlichen Auflösungen x_1, x_2, x_3 beinhalten können. In mindestens einer Ausführungsform kann x_3 durch einen Skalierungsfaktor von 4 hochskaliert werden, mit x_1 verkettet werden, es kann ein ResNet-Block angewendet werden und schließlich um 2 hochgerechnet werden, um eine Auflösung eines Pseudobildes aus Vogelperspektive einer ursprünglichen Eingabe auszugeben. In mindestens einer Ausführungsform kann es eine größere Anzahl von trainierbaren Parametern, wie etwa 14,3 Millionen trainierbare Parameter, in einem endgültigen Netzwerk geben.

[0009] In mindestens einer Ausführungsform kann es mehrere Hyperparameter geben, die eine Auflösung eines Modells bestimmen. In mindestens einer Ausführungsform kann ein Hyperparameter einer Größe eines Eingabebilds H × W entsprechen. In können Eingabebilder mit Extrinsik und Intrinsik **316**, die für diese Kameras und Sensoren, die dementsprechend umgewandelt wurden, gespeichert sind, in der Größe verändert und zugeschnitten werden, wie etwa auf die Größe 128 × 352 Bildpunkte. In mindestens einer Ausführungsform betrifft ein weiterer wichtiger Hyperparameter dieses Netzwerks eine Größe und Auflösung eines Vogelperspektivenrasters X × Y. In mindestens einer Ausführungsform können Behälter in sowohl x als auch y von -50 Metern bis 50 Meter mit Zellen der Größe 0,5 Meter × 0,5 Meter eingestellt werden. In mindestens einer Ausführungsform kann ein sich daraus ergebendes Raster 200 × 200 betragen. In mindestens einer Ausführungsform kann eine Auswahl von D eine Auflösung von Tiefe bestimmen, die durch ein Netzwerk bestimmt wurde. In mindestens einer Ausführungsform kann D im Wert eingeschränkt sein, wie etwa zwischen 4,0 Metern und 45,0 Metern, im Abstand von 1,0 Meter. In mindestens einer Ausführungsform kann mit diesen Hyperparametern und architektonischen Ausgestaltungsentscheidungen eine Vorwärtsweitergabe eines Modells bei 35 Hz auf einer Titan-X-GPU laufen.

[0010] In mindestens einer Ausführungsform kann ein derartiges System genutzt werden, um eine willkürliche Anordnung von Kameras, was einem Rigg von Kameras entsprechen kann, in eine einzelne Darstellung eines Schauplatzes oder einer Umgebung zu kodieren. In mindestens einer Ausführungsform kann dies eine „Vogelperspektive“, die von einem Blickwinkel über Objekten dieser Umgebung genommen wurde, oder eine Ansicht sein, die einem weiteren bestimmten Blickwinkel entspricht. In mindestens einer Ausführungsform kann dies unter Verwendung von Bildern oder anderen Arten von aufgenommenen Daten für ein Objekt oder eine Umgebung durchgeführt werden, die lediglich zweidimensionale Daten beinhalten, obwohl in mindestens einer Ausführungsform ein oder mehrere Bilder oder Arten von Daten zweidimensionale oder dreidimensionale Daten beinhalten können. In mindestens einer Ausführungsform kann ein derartiger Ansatz Schwierigkeiten durch das Umwandeln von zweidimensionalen Bildern von einer willkürlichen Anzahl von Kameras in eine einzelne Darstellung, wie etwa eine Vogelperspektivendarstellung überwinden, angesichts der Tatsache, dass zweidimensionale Kameras und Sensoren keine Tiefeninformationen bereitstellen. In mindestens einer Ausführungsform kann eine mit Merkmalen ausgestattete Punktwolke erzeugt werden, die Punkte in drei Dimensionen in einem Koordinatenrahmen beinhaltet, der zu jedem Bildpunkt jeder Kamera ortsnah ist. In mindestens einer Ausführungsform können diese Punktwolken dann in einem Weltrahmen unter Verwendung von Extrinsik und Intrinsik für entsprechende Kameras oder Sensoren akkumuliert werden. In mindestens einer Ausführungsform ist ein derartiger Ansatz auf eine beliebige Anzahl von Kameras in einer beliebigen Anzahl von Stellungen und Ausrichtungen skalierbar. In mindestens einer Ausführungsform kann ein derartiger Ansatz das Ableiten einer Stellung von 3D-Strukturen für Anwendungen ermöglichen, wie etwa einer autonomen Navigation oder Robotik unter Verwendung von lediglich zweidimensionalen Kamera- oder Sensordaten.

[0011] In mindestens einer Ausführungsform kann eine einzelne Darstellung eines Schauplatzes aus Bilddaten erzeugt werden, die durch eine angemessene Anzahl von Kameras aufgenommen wurden. In mindestens einer Ausführungsform sind Mengen in einem Bezugsrahmenkoordinatensystem zu Zwecken der Erörterung mit Hutsymbolen bezeichnet. In mindestens einer Ausführungsform können Bilder unterschiedliche Dimensionalitäten aufweisen, aber zur Einfachheit der Erläuterung geht mindestens ein Teil der Beispiele und der Erörterung in dieser Schrift von Bildern derselben Dimensionalität aus. In mindestens einer Ausführungsform kann ein System mit einer Anzahl von Kameras, n, jeweils ein Bild für einen einzelnen Zeitpunkt bereitstellen. In mindestens einer Ausführungsform führt dies zum Empfangen von n Bildern $(X_k \in \mathbb{R}^{3 \times H \times W}\}_n$, jeweils mit einer extrinsischen Matrix $E_k \in \mathbb{R}^{3 \times 4}$ und einer intrinsischen Matrix $I_k \in \mathbb{R}^{3 \times 3}$. In mindestens einer Ausführungsform

wird ein Versuch unternommen, um eine gerasterte Darstellung eines Schauplatzes in einem Bezugskoordinatenrahmen zu bestimmen, wie etwa einem Vogelperspektivenkoordinatenrahmen $\hat{y} \in \mathbb{R}^{C \times X \times Y}$. In mindestens einer Ausführungsform definieren extrinsische und intrinsische Matrizen für jede dieser n Kameras zusammen eine Zuordnung von Bezugskoordinaten ($\hat{x}$, $\hat{y}$, $\hat{z}$) zu örtlichen Bildpunktkoordinaten (h, w, d). In mindestens einer Ausführungsform gibt es kein Erfordernis für Zugriff auf jegliche Art von Tiefensensor oder -daten während des Trainings oder der Ableitung.

[0012] In mindestens einer Ausführungsform kann eine erste Stufe eines Deep-Learning-Modells an jedem Bild getrennt arbeiten. In mindestens einer Ausführungsform kann dies ermöglichen, dass jedes Bild von einer Position im zweidimensionalen Koordinatensystem in einen dreidimensionalen Bezugsrahmen oder ein Koordinatensystem gehoben wird, das über alle Kameras gemeinsam genutzt wird, die Bilddaten für ein relevantes Objekt oder eine relevante Umgebung bereitstellen. In mindestens einer Ausführungsform sind Tiefeninformationen, die den einzelnen Bildpunkten dieser zweidimensionalen Bilder zugeordnet sind, mehrdeutig und ein derartiges Modell kann verwendet werden, um Darstellungen bei einer beliebigen oder allen möglichen Tiefen für jeden Bildpunkt zu erzeugen. In mindestens einer Ausführungsform ist $X \in \mathbb{R}^{3 \times H \times W}$ ein Bild mit Extrinsik E und Intrinsik I und wobei $p_i \in X$ ein Bildpunkt in diesem Bild mit den Bildkoordinaten (h, w) darstellt. In mindestens einer Ausführungsform ist jedes Bildpunkt |D|-Punkten auf einem Strahl zugeordnet, der einen Ursprung dieser Kamera durchläuft, und ein Bildpunkt {(h, w, d) |d ∈ D}, bei dem D ein Satz diskreter Tiefen ist, kann durch $\{d_0 + \Delta, ..., d_0 + |D|\Delta\}$ definiert sein. In mindestens einer Ausführungsform gibt es keine erlernbaren Parameter in dieser Umwandlung, wobei eine große Punktwolke für ein bestimmtes Bild der Größe D x H x W erzeugt wird.

[0013] In mindestens einer Ausführungsform ist diese Punktwolke in einem gewünschten Bezugsrahmen definiert, ist aber nicht in irgendeiner Form abhängig vom Inhalt dieses Bildes. In mindestens einer Ausführungsform kann eine Abhängigkeit von Inhalt durch das Trainieren eines Netzwerks f_θ modelliert werden, um einen Kontextvektor $\hat{c}_d \in \mathbb{R}^C$ für jeden Punkt in dieser Punktwolke $p_{i,d}$ abzuleiten. In mindestens einer Ausführungsform ist ein Kontextvektor mit Parametern versehen, um mit einem Konzept der Aufmerksamkeit und diskreten Tiefenableitung übereinzustimmen. Bei jedem Bildpunkt sagt dieses Netzwerk einen Kontext $c \in \mathbb{R}^C$ und eine Verteilung über die Tiefe $1 \in \mathbb{R}^{|D|}$ für jeden Bildpunkt voraus. In mindestens einer Ausführungsform kann für einen Punkt bei der Tiefe d, der durch p_o erzeugt wurde, der Inhalt durch Folgendes angegeben werden:

$$\hat{\mathbf{c}}_d = \mathbf{c} l_d$$

[0014] In mindestens einer Ausführungsform würde, wenn dieses Netzwerk einen One-Hot-Vektor für I vorhersagen würde, der Kontext aus dem Bildpunkt p_i ausschließlich in einer einzigen Tiefe ungleich null sein, wie bei Pseudolidar. In mindestens einer Ausführungsform würde, wenn dieses Netzwerk eine gleichförmige Verteilung über die Tiefe vorhersagen würde, dieses Netzwerk eine selbe Darstellung für den Bildpunkt p_i unabhängig der Tiefe vorhersagen. In mindestens einer Ausführungsform kann dieses Netzwerk dann in der Lage sein, zwischen dem Platzieren von Kontext aus einem Bild an einer konkreten Position einer einzelnen Vogelperspektivendarstellung im Vergleich zum Verbreiten von Kontext über einen gesamten Strahl des Raumes wählen, wie etwa bei mehrdeutiger Tiefe. In mindestens einer Ausführungsform kann eine Funktion g_c: (x, y, z) $\in \mathbb{R}^2 \rightarrow c \in \mathbb{R}^C$ für jedes Bild erzeugt werden, das an einer beliebigen räumlichen Position abgefragt und einen Kontextvektor zurückgeben kann. In mindestens einer Ausführungsform kann ein System stattdessen Raum diskretisieren, um sich separate Faltungen zunutze zu machen. In mindestens einer Ausführungsform kann ein Volumen des Raums, das für eine bestimmte Kamera sichtbar ist, einem Kegelstumpf entsprechen.

[0015] In mindestens einer Ausführungsform kann diese anfängliche Netzwerkstufe eine Punktwolke mit n × H × W × D Punkten produzieren, wobei jeder davon einen Kontextvektor der Dimensionalität C diesem zugeordnet aufweist. In mindestens einer Ausführungsform ist diese Punktwolke grundsätzlich dahingehend unterschiedlich zu einer LIDAR-Punktwolke, dass eine räumliche Dichte von Punkt unabhängig von der Semantik ist. In mindestens einer Ausführungsform kann diese Punktwolke in eine Darstellung umgewandelt werden, die für effiziente neuronale Faltungsnetzwerke (CNNs) in 2D bestens geeignet ist. In mindestens einer Ausführungsform können Säulen verwendet werden, um sich auf Voxel mit unendlicher Höhe zu beziehen. In mindestens einer Ausführungsform kann jeder Punkt seiner nächsten Säule zugewiesen werden und eine Summenbündelung durchgeführt werden, um einen C × H × W - Tensor zu erzeugen, der durch ein standardmäßiges CNN für Vogelperspektivenableitung verarbeitet werden kann.

[0016] In mindestens einer Ausführungsform können Daten für diese Schritte wie in den **Fig. 4A** bis **Fig. 4D** veranschaulicht sichtbar gemacht werden. In mindestens einer Ausführungsform veranschaulicht der Verlauf **400** aus **Fig. 4A** eine Darstellung von sechs zweidimensionalen Bildern, die ein einem einzelnen Koordinaten-

rahmen oder Bezugsrahmen positioniert oder in dieses gehoben werden. In mindestens einer Ausführungsform kann dies das Verwenden von Extrinsik und Intrinsik einer Kamera mit einem neuronalen Netzwerk beinhalten, um eine geeignete Platzierung von jedem dieser Bilder in diesen Bezugsrahmen zu bestimmen. In mindestens einer Ausführungsform wird derartig eine kegelstumpfförmige Punktwolke für jedes Bild vorhergesagt, dass die Dichte lediglich von einer Kalibrierung einer entsprechenden Kalibrierung und unabhängig von der Semantik dieses Bildes ist. In mindestens einer Ausführungsform ist ein Zusammenschluss von Punktwolken für alle sechs Kameras in drei Dimensionen in **Fig. 4A** sichtbar gemacht. In mindestens einer Ausführungsform können dann unterschiedliche Projektionen, die Projektionen dieser Punktwolke in eine X-Y-Ebene **420** in **Fig. 4B** und in eine X-Z-Ebene **440** in **Fig. 4C** entsprechen, erzeugt werden. In mindestens einer Ausführungsform veranschaulicht der Verlauf **460** aus **Fig. 4D** ein Ergebnis eines Projizierens dieser Merkmale in einen gemeinsamen Bezugsrahmen. In mindestens einer Ausführungsform entspricht dies dem Diskretisieren einer ersten Ebene, wie etwa einer x-y-Ebene, eines Bezugsrahmens und dem Durchführen einer Summenbündelung von Merkmalen aller Punkte in dieser Punktwolke innerhalb jeder Rasterzelle. In mindestens einer Ausführungsform gehen Höheninformationen in dieser Umwandlung möglicherweise nicht verloren, was eine große Punktwolke für ein bestimmtes Bild der Größe D × H × W erzeugt.

[0017] In mindestens einer Ausführungsform kann ein Ansatz verwendet werden, um die Summenbündelung zu beschleunigen. In mindestens einer Ausführungsform kann Effizienz für das Trainieren eines Modells wichtig sein, vorausgesetzt die Größen von Punktwolken erzeugten (B × N × D × H × W für Batchgröße B). In mindestens einer Ausführungsform können bei einer vorgegebenen Punktwolke von N Punkten, von denen jeder eine D-dimensionale geometrische Position $G \in \mathbb{R}^{N \times D}$, eine C-dimensionale Kontextkodierung $O \in \mathbb{R}^{N \times C}$ und eine Gruppierungs-ID $P \in \mathbb{R}^{N}$ aufweist, diese Punkte geordnet werden, sodass alle Punkte mit einer selben Gruppierungs-ID aneinander angrenzen. In mindestens einer Ausführungsform kann die Summenbündelung dann durch das Berechnen einer Gesamtsumme über alle Kontextvektoren in dieser Punktwolke und das Auswerten von Unterschieden zwischen einer Gesamtsumme an Endpunkten jeder Gruppierungs-ID durchgeführt werden. In mindestens einer Ausführungsform weist dieser Vorgang einen analytischen Gradienten auf, der effizient berechnet werden kann, um eine automatische Unterscheidung zu beschleunigen, wie etwa durch das Verwenden von Autograd für PyTorch.

[0018] In mindestens einer Ausführungsform kann ein derartiges Modell Ende-zu-Ende-Kostenzuordnungserlemen für die Bewegungsplanung aus der Kameraeingabe bereitstellen. In mindestens einer Ausführungsform kann das Planen einer Ableitungszeit durch das Projizieren verschiedener Bahnverläufe, das Bewerten relativer Kosten dieser Bahnverläufe, dann das Handeln nach einem Bahnverlauf mit den geringsten Kosten erreicht werden. In mindestens einer Ausführungsform kann dies unter Verwendung eines Bahnverlaufsbestimmungsnetzwerks **318** durchgeführt werden, das in der Lage ist, einen optimalen Bahnverlauf aus einem Satz möglicher Bahnverläufe abzuleiten und entsprechende Bahnverlaufsdaten **320** auszugeben. In mindestens einer Ausführungsform kann dies eine exakte Ende-zu-Ende auslegbare Bewegungsplanung bereitstellen. In mindestens einer Ausführungsform kann das Planen das Vorhersagen einer Verteilung über Bahnverläufe für ein Egofahrzeug einbeziehen, wie durch Folgendes vorgegeben werden kann:

$$\tau = \{\tau_i\}_K = \{\{x_j, y_j, t_j\} T\} k$$

[0019] In mindestens einer Ausführungsform kann dies von Sensorbeobachtungen p(το) abhängen. In mindestens einer Ausführungsform kann das Planen als Zuordnung über einen Satz von K Vorlagenbahnverläufen konzipiert sein. In mindestens einer Ausführungsform kann, um eine Kostenvolumeneigenschaft eines Planungsproblems zu nutzen, eine Verteilung über K Vorlagenbahnverläufe gezwungen werden, eine festgelegte Form anzunehmen, die durch Folgendes angegeben werden kann:

$$p(\tau_i|o) = \frac{exp\left(-\sum_{x_i, y_i \in \tau_i,} c(x_i, y_i|o)\right)}{\sum_{T \in T} - exp\left(\left(-\sum_{x_i, y_i \in \tau_i \tau} c\ x_i, y_i|o\right)\right)}$$

wobei c(.o) aus Daten durch das Optimieren für eine log-Wahrscheinlichkeit von fachgerechten Bahnverläufen erlernt werden kann.

[0020] In mindestens einer Ausführungsform kann, für Markierungen und bei einem vorgegebenen Ground-Truth-Bahnverlauf, ein nächster Nachbar in L2-Entfernung zu den Vorlagenbahnverläufen τ berechnet werden, dann mit einem Kreuzentropieverlust trainiert werden. In mindestens einer Ausführungsform ermöglicht diese Definition von $p(\tau_i|o)$ das Erlernen einer auslegbaren Raumkostenfunktion, ohne einen Hard-Margin-Verlust zu definieren. In mindestens einer Ausführungsform kann ein Satz von Vorlagenbahnverläufen unter Verwendung von K-Means bei allen Egobahnverläufen in einem Trainingssatz bestimmt werden.

[0021] In mindestens einer Ausführungsform kann das Trainieren der Effizienz wichtig für das Erlernen von Daten aus einem Satz von Kameras sein, wie etwa für ein gesamtes Sensorrigg. In mindestens einer Ausführungsform kann die säulenübergreifende Summenbündelung anstatt einer maximalen Bündelung genutzt werden, da sowohl ein Vorwärtsdurchlauf als ein Rückwärtsdurchlauf der Summenbündelung effizient in einem derartigen System durchgeführt werden können. In mindestens einer Ausführungsform kann eine Kegelstumpfbündelungsschicht das Umwandeln von Kegelstümpfen, die durch n Bilder produziert wurden, in einen feststehenden dimensionalen Tensor unabhängig von einer Anzahl von Kameras n erledigen.

[0022] In mindestens einer Ausführungsform können extrinsische Matrizen eines Satzes von Kameras relative Ausrichtungen von jeder dieser Kameras durch das Vornehmen einer willkürlichen Auswahl des Bezugsrahmens definieren. In mindestens einer Ausführungsform kann eine affine Umwandlung, die gemeinschaftlich auf die gesamte Extrinsik angewendet wird, zu einem zweiten gültigen Bezugsrahmen führen. In mindestens einer Ausführungsform können, da ein derartiges Netzwerk die Extrinsik dieser Kameras bedingen kann, Ground-Truth-Ziele gedreht und verschoben werden, während diese Extrinsik ebenfalls gedreht und verschoben wird. In mindestens einer Ausführungsform kann mindestens eine Art von Regulierung genutzt werden, die eine dreidimensionale Struktur eines derartigen Modells berücksichtigt. In mindestens einer Ausführungsform können Kameras zufällig aus jedem Minibatch gelöscht werden, wodurch ein Netzwerk Korrelationen von Sensoren erlernen kann.

[0023] In mindestens einer Ausführungsform kann ein CNN zu Beginn mit einer dreidimensionalen Struktur ausgestattet sein. In mindestens einer Ausführungsform kann diese Struktur wichtig für die Leistung sein. In mindestens einer Ausführungsform kann ein CNN trainiert sein, Vogelperspektivenmarkierungen von Bildern auszugeben, während es induktives Denken bei der Architektur für mehrere Kameras nutzt. In mindestens einer Ausführungsform kann eine CNN-Architektur ein EfficientNet-BO-Backbone aufweisen, das unabhängig Merkmale aus allen Bildern extrahiert. In mindestens einer Ausführungsform können andere Arten von Backbones verwendet werden, die ResNet, DeepLap oder einem Gate-Form-CNN (Gate-Shape CNN - GCCNN) für semantische Segmentierung entsprechen. In mindestens einer Ausführungsform können diese Darstellungen dann verkettet und eine bilineare Interpolation durchgeführt werden um sie in einen $\mathbb{R}^{X \times Y}$ -Tensor hochzurechnen, wie durch dieses Modell ausgegeben wird. In mindestens einer Ausführungsform kann ein derartiges CNN semantische 3D-Darstellungen erlernen, die eine Überwachung in einem Vogelperspektivenrahmen vorgeben.

[0024] In mindestens einer Ausführungsform kann, wenn die Fusion der Kameras durch ein Vogelperspektiven-CNN erlernt wurde, dieses Modell trainiert werden, um gegenüber einfachen Rauschmodellen widerstandsfähig zu sein, die in Anwendungen auftreten, wie etwa selbstfahrenden Fahrzeugen, was sich darauf beziehen kann, das Extrinsik verzerrt ist oder Kameras kaum noch Energie haben. In mindestens einer Ausführungsform kann das Löschen von Kameras während des Trainings die Leistung dieses Modells verbessern. In mindestens einer Ausführungsform kann das Löschen einer Kamera oder eines Sensors ein derartiges Modell zwingen, Korrelationen zwischen Bildern auf unterschiedlichen Kameras zu erlernen. In mindestens einer Ausführungsform kann das Trainieren eines derartigen Modells mit rauschender Extrinsik außerdem zu einer besseren Leitung zum Testzeitpunkt führen. In mindestens einer Ausführungsform kann eine Bedeutsamkeit jeder Kamera für die Leistung jeder Objektsegmentierung für neue Schauplätze bestimmt werden. In mindestens einer Ausführungsform kann das Verlieren von Kameras für neue Schauplätze bedeuten, dass ein oder mehrere Bereiche um eine Kamera oder Sensorbaugruppe keine Sensormessungen aufweisen werden, und folglich kann die Leistung durch die Leistung mit einem kompletten Sensorrigg streng nach oben begrenzt sein. In mindestens einer Ausführungsform kann die Bedeutsamkeit jeder Kamera gemessen werden, was andeuten kann, wo in diesem System Redundanz am wichtigsten sein könnte. In mindestens einer Ausführungsform kann ein derartiges Modell verschiedene Verallgemeinerungen unterstützen, die sich auf Unterschiede bei verfügbaren Kameras oder unterschiedlichen Anzahlen von Kameras, als für das Training verwendet wurden, beziehen können. In mindestens einer Ausführungsform kann ein derartiges Netzwerk ebenfalls derartig trainiert werden, dass es widerstandsfähig gegenüber üblichen Quellen von Sensorfehlern ist. In mindestens einer Ausführungsform kann das Trainieren mit einer großen Menge an Rauschen in der Extrinsik zu einem Netzwerk führen, dass gegenüber extrinsischem Rauschen zum Ableitungszeitpunkt widerstandsfähiger wird.

In mindestens einer Ausführungsform kann das zufällige Löschen von Kameras von jedem Batch während des Trainings die Widerstandsfähigkeit gegenüber dem Sensorausfall zum Ableitungszeitpunkt erhöhen.

[0025] In mindestens einer Ausführungsform kann ein Modell trainiert werden, um in der Lage zu sein, beim Planen der Leistung unter Verwendung einer angemessenen Kostenfunktion zu unterstützen. In mindestens einer Ausführungsform kann eine Vielzahl von möglichen Wegen oder Bahnverläufen für eine steuerbare Vorrichtung erzeugt werden. In mindestens einer Ausführungsform können diese erzeugten Bahnverläufe eine gewünschte Länge und Beabstandung aufweisen, wie etwa Bahnverläufe von fünf Sekunden Länge, die ungefähr 0,25 Sekunden beabstandet sind. In mindestens einer Ausführungsform kann ein Clusteringansatz, wie etwa K-Means, auf alle Egobahnverläufe eines Trainingssatzes neuer Schauplätze angepasst werden, um eine oder mehrere Vorlagen zu erlangen. In mindestens einer Ausführungsform können Ableitungszeitpunktmessungen vorgenommen werden, wie gut ein Netzwerk in der Lage ist, eine Vorlage vorherzusagen, die einem Ground-Truth-Bahnverlauf unter einer Metrik, wie etwa einer L2-Metrik am nächsten ist. In mindestens einer Ausführungsform kann diese Aufgabe ein deutlich geringeres Überwachungssignal als die Segmentierungsaufgaben aufweisen, die vorher erörtert wurden. In mindestens einer Ausführungsform ist diese Aufgabe jedoch ein wichtiges Experiment für das Selbstfahren, da Ground-Truth-Ziele für ein derartiges Experiment von Größenordnungen sein können, die deutlich kostengünstiger als Ground-Truth-3D-Begrenzungsfelder in der Anschaffung sind, da geschicktes Fahren für die meisten Menschen intuitiv ist. In mindestens einer Ausführungsform kann diese Aufgabe für die Benchmarkingleistung kamerabasierter Ansätze wichtig sein.

[0026] In mindestens einer Ausführungsform kann eine derartige Architektur trainiert werden, um verschiedene Darstellungen, wie etwa Vogelperspektivendarstellungen, aus einer willkürlichen Anzahl von Kameras oder einer willkürlichen Kamerabaugruppe zu erzeugen. In mindestens einer Ausführungsform kann ein Modell in dieser Architektur widerstandsfähig gegenüber einfachen Modellen von Kalibrierungsrauschen sein. In mindestens einer Ausführungsform kann ein derartiges Modell eine Ende-zu-Ende-Bewegungsplanung ermöglichen, die einem Bahnverlaufsfotoaufhahmenbezugssystem folgt. In mindestens einer Ausführungsform wird ein Aufpreis jedes Vorlagenbahnverlaufs berechnet und als eine Boltzmann-Verteilung über einem Satz von Vorlagen ausgelegt. In mindestens einer Ausführungsform kann ein Argmax dieser Verteilung gemäß einer ausgewählten Vorlage aus werden.

[0027] In mindestens einer Ausführungsform kann ein Prozess **500** zum Erzeugen eines Bildes einer Umgebung mit einer konkreten Ansicht wie in **Fig. 5A** veranschaulicht durchgeführt werden. In mindestens einer Ausführungsform kann ein Satz von zweidimensionalen Bildern empfangen werden **520**, die alle zu oder nahe einem einzelnen Zeitpunkt aufgenommen wurden. In mindestens einer Ausführungsform stellen diese Bilder Ansichten einer Umgebung in verschiedenen Bezugsrahmen dar. In mindestens einer Ausführungsform kann mindestens eine dieser Kameras (oder Bildgebungssensoren) ebenfalls Tiefen- oder dreidimensionale Daten beinhalten. In mindestens einer Ausführungsform können diese Bilder verarbeitet werden **504**, wie etwa durch das Verwenden eines trainierten CNN, um einen Satz kontextueller oder semantischer Bildmerkmale zu extrahieren. In können diese Merkmale beliebige Merkmale sein, die durch dieses CNN oder einen weiteren Merkmalsextrahierer bestimmt wurden, um ein entsprechendes Bild darzustellen. In mindestens einer Ausführungsform können diese Merkmale verwendet werden, um eine mit Merkmalen ausgestattete kegelstumpfförmige Punktwolke in einem örtlichen Bezugsrahmen für jede Kamera zu erzeugen **506**. In mindestens einer Ausführungsform können diese einzelnen Punktwolken in einem dreidimensionalen Weltbezugsrahmen positioniert sein **508**, der verschiedene Intrinsik und Extrinsik nutzen kann, die bekannt für eigene Kameras bekannt sind, von denen diese Bilder empfangen wurden. In mindestens einer Ausführungsform kann diese Platzierung zu einer einzelnen Punktwolke in drei Dimensionen führen, die für diese Umgebung steht. In mindestens einer Ausführungsform kann ein Bild oder Datensatz erzeugt werden **510**, das/der einen bestimmten Blickwinkel für diese Umgebung darstellt. In mindestens einer Ausführungsform schließt dies das Projizieren von Punkt in diesem gemeinsamen Weltbezugsrahmen auf eine bestimmte Ebene im Hinblick auf einen konkreten Ausgangspunkt ein. In mindestens einer Ausführungsform können mehrere Bilder erzeugt werden, die unterschiedliche Auflösungen aufweisen oder unterschiedlichen Blickwinkeln entsprechen können. In mindestens einer Ausführungsform können diese Bilder dann zur Verwendung beim Ausführen von Aufgaben im Hinblick auf diese Umgebung bereitgestellt sein **512**, wie etwa das Steuern oder Navigieren eines Fahrzeugs, Roboters oder einer Vorrichtung. In mindestens einer Ausführungsform kann die Projektion dieser Punktkegelstümpfe auf eine Referenzebene für Aufgaben nützlich sein, wie etwa die Bewegungsplanung, wobei eine einzelne zusammenhängende Darstellung unter Verwendung zweidimensionaler Bilddaten von einer willkürlichen Anzahl von Kameras erzeugt wird.

[0028] In mindestens einer Ausführungsform kann ein Prozess **550** zum Erzeugen eines dreidimensionalen Bildes durchgeführt werden **556**, wie in **Fig. 5B** veranschaulicht. In mindestens einer Ausführungsform werden

ein oder mehrere zweidimensionale Bilder erlangt **552**, wobei mindestens einige dieser Bilder unterschiedliche Bezugsrahmen aufweisen können. In mindestens einer Ausführungsform können Positionen von Merkmalen dieser Bilder in einem gemeinsamen Bezugsrahmen bestimmt werden 554. In mindestens einer Ausführungsform können ein oder mehrere dreidimensionale Bilder mindestens zum Teil auf Grundlage von Positionen dieser Bildmerkmale in diesem gemeinsamen Bezugsrahmen erzeugt werden.

ABLEITUNGS- UND TRAININGSLOGIK

[0029] Fig. 6A veranschaulicht Ableitungs- und/oder Trainingslogik **615**, die verwendet wird, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. Details bezüglich der Ableitungs- und/oder Trainingslogik **615** werden nachfolgend in Verbindung mit den **Fig. 6A** und/oder 6B bereitgestellt.

[0030] In mindestens einer Ausführungsform kann die Ableitungs- und/oder Trainingslogik **615** ohne Einschränkung Folgendes beinhalten: Code- und/oder Datenspeicher **601**, um Gewichtung und/oder Eingabe-/Ausgabedaten und/oder andere Parameter vorwärtsgerichtet zu speichern und/oder auszugeben, um Neuronen oder Schichten eines neuronalen Netzwerks zu konfigurieren, das in Aspekten einer oder mehrerer Ausführungsformen trainiert und/oder zum Ableiten verwendet wird. In mindestens einer Ausführungsform kann die Trainingslogik **615** einen Code- und/oder Datenspeicher **601** beinhalten oder mit diesem gekoppelt sein, um Diagrammcode oder andere Software zu speichern, um die Zeitsteuerung und/oder die Reihenfolge zu steuern, in der Gewichtungsinformationen und/oder Informationen zu anderen Parametern gespeichert werden sollen, um Logik zu konfigurieren, die Ganzzahl- und/oder Gleitkommaeinheiten (zusammen arithmetische Logikeinheiten (arithmetic logic units - ALUs)) beinhalten. In mindestens einer Ausführungsform lädt Code, wie etwa Diagrammcode, Gewichtungsinformationen oder Informationen zu anderen Parametern in Prozessor-ALUs auf Grundlage der Architektur eines neuronalen Netzwerks, dem dieser Code entspricht. In mindestens einer Ausführungsform speichert der Code- und/oder Datenspeicher **601** Gewichtungsparameter und/oder Eingabe-/Ausgabedaten jeder Schicht eines neuronalen Netzwerks, das trainiert oder in Verbindung mit einer oder mehreren Ausführungsformen während der Vorwärtspropagation von Eingabe-/Ausgabedaten und/oder Gewichtsparametern während des Trainings und/oder dem Ableiten unter Verwendung von Aspekten einer oder mehrerer Ausführungsformen verwendet wird. In mindestens einer Ausführungsform kann ein beliebiger Teil des Code- und/oder Datenspeichers **601** in einem anderen chipinternen oder chipexternen Datenspeicher beinhaltet sein, was einen L1 - , L2- oder L3-Zwischenspeicher oder Systemspeicher eines Prozessors beinhaltet.

[0031] In mindestens einer Ausführungsform kann ein beliebiger Teil des Code- und/oder Datenspeichers **601** intern oder extern von einem oder mehreren Prozessoren oder anderen Hardwarelogikvorrichtungen oder -schaltungen sein. In mindestens einer Ausführungsform kann der Code- und/oder Datenspeicher **601** ein schneller Pufferspeicher, ein dynamischer zufällig adressierbarer Speicher (dynamic randomly addressable memory - „DRAM“), ein statischer zufällig adressierbarer Speicher (static randomly addressable memory - „SRAM“), ein nichtflüchtiger Speicher (z. B. Flash-Speicher) oder ein anderer Speicher sein. In mindestens einer Ausführungsform kann die Wahl, ob der Code- und/oder Datenspeicher **601** zum Beispiel intern oder extern von einem Prozessor ist oder DRAM, SRAM, Flash oder eine andere Speicherart umfasst, von verfügbarem chipinternen im Vergleich zu chipexternen Speicher, Latenzanforderungen von ausgeführten Trainings- und/oder Ableitungsfunktionen, Batchgröße von Daten, die beim Ableitungen und/oder Trainieren eines neuronalen Netzwerkes verwendet werden, oder einer Kombination dieser Faktoren abhängen.

[0032] In mindestens einer Ausführungsform kann die Ableitungs- und/oder Trainingslogik **615** ohne Einschränkung Folgendes beinhalten: einem Code- und/oder Datenspeicher **605**, um Gewichtung und/oder Eingabe-/Ausgabedaten rückwärtsgerichtet zu speichern und/oder auszugeben, die Neuronen oder Schichten eines neuronalen Netzwerks entsprechen, das in Aspekten einer oder mehrerer Ausführungsformen trainiert und/oder zum Ableiten verwendet wird. In mindestens einer Ausführungsform speichert der Code- und/oder Datenspeicher **605** Gewichtungsparameter und/oder Eingabe-/Ausgabedaten jeder Schicht eines neuronalen Netzwerks, das trainiert oder in Verbindung mit einer oder mehreren Ausführungsformen während der Rückwärtspropagation von Eingabe-/Ausgabedaten und/oder Gewichtsparametern während des Trainings und/oder dem Ableiten unter Verwendung von Aspekten einer oder mehrerer Ausführungsformen verwendet wird. In mindestens einer Ausführungsform kann die Trainingslogik **615** einen Code- und/oder Datenspeicher **605** beinhalten oder mit diesem gekoppelt sein, um Diagrammcode oder andere Software zu speichern, um die Zeitsteuerung und/oder die Reihenfolge zu steuern, in der Gewichtungsinformationen und/oder Informationen zu anderen Parametern gespeichert werden sollen, um Logik zu konfigurieren, die Ganzzahl- und/oder Gleitkommaeinheiten (zusammen arithmetische Logikeinheiten (ALUs)) beinhalten. In mindestens einer Ausführungsform lädt Code, wie etwa Diagrammcode, Gewichtungsinformationen oder Informationen zu anderen Parametern in

Prozessor-ALUs auf Grundlage einer Architektur eines neuronalen Netzwerks, dem dieser Code entspricht. In mindestens einer Ausführungsform kann ein beliebiger Teil des Code- und/oder Datenspeichers **605** in einem anderen chipinternen oder chipexternen Datenspeicher beinhaltet sein, was einen L1-, L2- oder L3-Zwischenspeicher oder Systemspeicher eines Prozessors beinhaltet. In mindestens einer Ausführungsform kann ein beliebiger Teil des Code- und/oder Datenspeichers **605** auf einem oder mehreren Prozessoren oder anderen Hardwarelogikvorrichtungen oder -schaltungen intern oder extern sein. In mindestens einer Ausführungsform kann der Code- und/oder Datenspeicher **605** ein schneller Pufferspeicher, DRAM, SRAM, ein nichtflüchtiger Speicher (z. B. Flash-Speicher) oder ein anderer Speicher sein. In mindestens einer Ausführungsform kann die Wahl, ob der Code- und/oder Datenspeicher **605** zum Beispiel intern oder extern von einem Prozessor ist oder DRAM, SRAM, Flash oder eine andere Speicherart umfasst, von verfügbarem chipinternen im Vergleich zu chipexternen Speicher, Latenzanforderungen von ausgeführten Trainings- und/oder Ableitungsfunktionen, Batchgröße von Daten, die beim Ableitungen und/oder Trainieren eines neuronalen Netzwerkes verwendet werden, oder einer Kombination dieser Faktoren abhängen.

[0033] In mindestens einer Ausführungsform können der Code- und/oder Datenspeicher **601** und der Code- und/oder Datenspeicher **605** getrennte Speicherstrukturen sein. In mindestens einer Ausführungsform können der Code- und/oder Datenspeicher **601** und der Code- und/oder Datenspeicher **605** dieselbe Speicherstruktur sein. In mindestens einer Ausführungsform können der Code- und/oder Datenspeicher **601** und der Code- und/oder Datenspeicher **605** teilweise dieselbe Speicherstruktur und teilweise getrennte Speicherstrukturen sein. In mindestens einer Ausführungsform kann ein beliebiger Teil des Code- und/oder Datenspeichers **601** und des Code- und/oder Datenspeichers **605** in einem anderen chipinternen oder chipexternen Datenspeicher beinhaltet sein, was einen L1-, L2- oder L3-Zwischenspeicher oder Systemspeicher eines Prozessors beinhaltet.

[0034] In mindestens einer Ausführungsform kann die Ableitungs- und/oder Trainingslogik **615** ohne Einschränkung Folgendes beinhalten: eine oder mehrere arithmetische Logikeinheit(en) („ALU(s)“) **610**, die Ganzzahl- und/oder Gleitkommaeinheiten beinhalten, um logische und/oder mathematische Operationen durchführen, die mindestens teilweise auf Trainings- und/oder Ableitungscode (z. B. Diagrammcode) basieren oder durch diesen angezeigt werden, deren Ergebnis Aktivierungen (z. B. Ausgabewerte von Schichten oder Neuronen innerhalb eines neuronales Netzwerk) erzeugen, die in einem Aktivierungsspeicher **620** gespeichert sind, die Funktionen von Eingabe/Ausgabe- und/oder Gewichtsparameterdaten sind, die in dem Code- und/oder Datenspeicher **601** und/oder Code- und/oder Datenspeicher **605** gespeichert sind. In mindestens einer Ausführungsform werden im Aktivierungsspeicher **620** gespeicherte Aktivierungen gemäß linearer algebraischer und/oder matrixbasierter Mathematik erzeugt, die von ALU(s) **610** als Reaktion auf das Ausführen von Anweisungen oder anderem Code ausgeführt wird, wobei in dem Code und/oder Datenspeicher **605** und/oder dem Code- und/oder Datenspeicher **601** gespeicherte Gewichtungswerte als Operanden zusammen mit anderen Werten verwendet werden, wie etwa Verzerrungswerten, Gradienteninformationen, Impulswerten oder anderen Parametern oder Hyperparametem, von denen einige oder alle in dem Code und/oder Datenspeicher **605** oder dem Code- und/oder Datenspeicher **601** oder einem anderen Speicher chipintern oder -extern gespeichert werden können.

[0035] In mindestens einer Ausführungsform sind die ALU(s) **610** in einem oder mehreren Prozessoren oder anderen Hardwarelogikvorrichtungen oder -schaltungen beinhaltet, wohingegen in einer weiteren Ausführungsform die ALU(s) **610** außerhalb eines Prozessors oder einer anderen Hardwarelogikvorrichtung oder -schaltung liegen können, die sie verwendet (z. B. ein Co-Prozessor). In mindestens einer Ausführungsform können die ALUs **610** in den Ausführungseinheiten eines Prozessors oder anderweitig in einer Bank von ALUs beinhaltet sein, auf welche die Ausführungseinheiten eines Prozessors, entweder innerhalb desselben Prozessors oder zwischen verschiedenen Prozessoren unterschiedlicher Arten (z. B. Zentraleinheiten, Grafikverarbeitungseinheiten, Einheiten mit fester Funktion usw.) verteilt, zugreifen können. In mindestens einer Ausführungsform können sich der Code- und/oder Datenspeicher **601**, der Code- und/oder Datenspeicher **605** und der Aktivierungsspeicher **620** auf demselben Prozessor oder einer anderen Hardwarelogikvorrichtung oder -schaltung befinden, wohingegen sie sich in einer weiteren Ausführungsform in unterschiedlichen Prozessoren oder anderen Hardwarelogikvorrichtungen oder -schaltungen oder einer Kombination von gleichen und unterschiedlichen Prozessoren oder anderen Hardwarelogikvorrichtungen oder -schaltungen befinden können. In mindestens einer Ausführungsform kann ein beliebiger Teil des Aktivierungsspeichers **620** in einem anderen chipinternen oder chipexternen Datenspeicher beinhaltet sein, was einen L1-, L2- oder L3-Zwischenspeicher oder Systemspeicher eines Prozessors beinhaltet. Darüber hinaus kann Ableitungs- und/oder Trainingscode mit anderem Code gespeichert werden, auf den ein Prozessor oder eine andere Hardwarelogik oder -schaltung zugreifen kann, und unter Verwendung der Abruf-, Dekodier-, Planungs-, Ausführungs-, Stilllegungs- und/oder anderen logischen Schaltungen eines Prozessors abgerufen und/oder verarbeitet werden.

[0036] In mindestens einer Ausführungsform kann der Aktivierungsspeicher **620** ein schneller Pufferspeicher, DRAM, SRAM, ein nichtflüchtiger Speicher (z. B. Flash-Speicher) oder ein anderer Speicher sein. In mindestens einer Ausführungsform kann sich der Aktivierungsspeicher **620** vollständig oder teilweise innerhalb oder außerhalb von einem oder mehreren Prozessoren oder anderen logischen Schaltungen befinden. In mindestens einer Ausführungsform kann die Wahl, ob der Aktivierungsspeicher **620** zum Beispiel intern oder extern von einem Prozessor ist oder DRAM, SRAM, Flash oder eine andere Speicherart umfasst, von verfügbarem chipinternen im Vergleich zu chipexternen Speicher, Latenzanforderungen von ausgeführten Trainings- und/oder Ableitungsfunktionen, Batchgröße von Daten, die beim Ableitungen und/oder Trainieren eines neuronalen Netzwerkes verwendet werden, oder einer Kombination dieser Faktoren abhängen. In mindestens einer Ausführungsform kann die Ableitungs- und/oder Trainingslogik **615**, die in **Fig. 6A** veranschaulicht ist, in Verbindung mit einer anwendungsspezifischen integrierten Schaltung (application-specific integrated circuit - „ASIC“) verwendet werden, wie etwa einer Tensorflow® Processing Unit von Google, einer Ableitungsverarbeitungseinheit (inference processing unit - IPU) von Graphcore™ oder einem Nervana®-(z. B. „Lake Crest“-)Prozessor von Intel Corp. In mindestens einer Ausführungsform kann die Ableitungs- und/oder Trainingslogik **615**, die in **Fig. 6A** veranschaulicht ist, in Verbindung mit Hardware einer Zentralverarbeitungseinheit (central processing unit - „CPU“), Hardware einer Grafikverarbeitungseinheit (graphics processing unit - „GPU“) oder anderer Hardware, wie etwa feldprogrammierbaren Gatearrays („FPGAs“) verwendet werden.

[0037] **Fig. 6B** veranschaulicht Ableitungs- und/oder Trainingslogik **615** gemäß mindestens einer oder mehreren Ausführungsformen. In mindestens einer Ausführungsform kann die Ableitungs- und/oder Trainingslogik **615** ohne Einschränkung Folgendes beinhalten: Hardwarelogik, bei der Berechnungsressourcen dediziert oder anderweitig ausschließlich in Verbindung mit Gewichtungswerten oder anderen Informationen verwendet werden, die einer oder mehreren Schichten von Neuronen innerhalb eines neuronalen Netzwerks entsprechen. In mindestens einer Ausführungsform kann die Ableitungs- und/oder Trainingslogik **615**, die in **Fig. 6B** veranschaulicht ist, in Verbindung mit einer anwendungsspezifischen integrierten Schaltung (ASIC) verwendet werden, wie etwa einer Tensorflow® Processing Unit von Google, einer Ableitungsverarbeitungseinheit (IPU) von Graphcore™ oder einem Nervana®-(z. B. „Lake Crest“-)Prozessor von Intel Corp. In mindestens einer Ausführungsform kann die Ableitungs- und/oder Trainingslogik **615**, die in **Fig. 6B** veranschaulicht ist, in Verbindung mit Hardware einer Zentralverarbeitungseinheit (CPU), Hardware einer Grafikverarbeitungseinheit (GPU) oder anderer Hardware, wie etwa feldprogrammierbaren Gatearrays (FPGAs) verwendet werden. In mindestens einer Ausführungsform beinhaltet die Ableitungs- und/oder Trainingslogik **615** ohne Einschränkung den Code- und/oder Datenspeicher **601** und den Code- und/oder Datenspeicher **605**, die verwendet werden können, um Code (z. B. Diagrammcode), Gewichtungswerte und/oder andere Informationen zu speichern, was Verzerrungswerte, Gradienteninformationen, Impulswerte und/oder andere Parameter- oder Hyperparameterinformationen beinhaltet. In mindestens einer in **Fig. 6B** veranschaulichten Ausführungsform ist jeder von dem Code- und/oder Datenspeicher **601** und dem Code- und/oder Datenspeicher **605** mit einer dedizierten Berechnungsressource verknüpft, wie etwa einer Berechnungshardware **602** bzw. einer Berechnungshardware **606**. In mindestens einer Ausführungsform umfasst jede von der Berechnungshardware **602** und der Berechnungshardware **606** eine oder mehrere ALUs, die mathematische Funktionen, wie etwa lineare algebraische Funktionen, nur an Informationen ausführen, die im Code- und/oder Datenspeicher **601** und Code- und/oder Datenspeicher **605** gespeichert sind, deren Ergebnis im Aktivierungsspeicher **620** gespeichert ist.

[0038] In mindestens einer Ausführungsform entspricht jeder von dem Code- und/oder Datenspeicher **601** und **605** und die entsprechende Berechnungshardware **602** bzw. **606** derartig unterschiedlichen Schichten eines neuronalen Netzwerks, dass die sich ergebende Aktivierung von einem „Speicher-/Berechnungspaar **601/602**“ des Code- und/oder Datenspeichers **601** und der Berechnungshardware **602** als eine Eingabe für das „Speicher-/Berechnungspaar **605/606**“ des Code- und/oder Datenspeichers **605** und der Berechnungshardware **606** bereitgestellt wird, um die konzeptionelle Organisation eines neuronalen Netzwerk widerzuspiegeln. In mindestens einer Ausführungsform kann jedes der Speicher-/Berechnungspaare **601/602** und **605/606** mehr als einer Schicht des neuronalen Netzwerks entsprechen. In mindestens einer Ausführungsform können zusätzliche Speicher-/Berechnungspaare (nicht gezeigt) nach oder parallel zu den Speicherberechnungspaaren **601/602** und **605/606** in der Ableitungs- und/oder Trainingslogik **615** beinhaltet sein.

RECHENZENTRUM

[0039] **Fig. 7** veranschaulicht ein beispielhaftes Rechenzentrum **700**, in dem mindestens eine Ausführungsform verwendet werden kann. In mindestens einer Ausführungsform beinhaltet das Rechenzentrum **700** eine Rechenzentrumsinfrastrukturschicht **710**, eine Frameworkschicht **720**, eine Softwareschicht **730** und eine Anwendungsschicht **740**.

[0040] In mindestens einer Ausführungsform kann, wie in **Fig. 7** gezeigt, die Rechenzentrumsinfrastrukturschicht **710** einen Ressourcenorchestrierer **712**, gruppierte Berechnungsressourcen **714** und Knotenberechnungsressourcen („Knoten-CRs“) **716(1)-716(N)** beinhalten, wobei „N“ eine beliebige ganze positive Zahl darstellt. In mindestens einer Ausführungsform können die Knoten-CRs **716(1)-716(N)** eine beliebige Anzahl von Zentralverarbeitungseinheiten („CPUs“) oder andere Prozessoren (die Beschleuniger, feldprogrammierbare Gatearrays (FPGAs), Grafikprozessoren usw.), Arbeitsspeichervorrichtungen (z. B. dynamischer Nur-Lese-Speicher), Datenspeichervorrichtungen (z. B. Festkörper- oder Festplattenlaufwerke), Netzwerk-Ein- /Ausgabevorrichtungen („NW E/A“), Netzwerk-Switches, virtuellen Maschinen („VMs“), Leistungsmodulen und Kühlmodule usw. beinhalten, ohne darauf beschränkt zu sein. In mindestens einer Ausführungsform können ein oder mehrere Knoten-CRs von den Knoten-CRs **716(1)-716(N)** ein Server sein, der eine oder mehrere der vorstehend erwähnten Berechnungsressourcen aufweist.

[0041] In mindestens einer Ausführungsform können gruppierte Berechnungsressourcen **714** getrennte Gruppierungen von Knoten-CRs beinhalten, die in einem oder mehreren Racks (nicht gezeigt) untergebracht sind, oder vielen Racks, die in Rechenzentren an verschiedenen geografischen Standorten (ebenfalls nicht gezeigt) untergebracht sind. Getrennte Gruppierungen von Knoten-CRs innerhalb gruppierter Berechnungsressourcen **714** können gruppierte Rechen-, Netzwerk-, Arbeitsspeicher- oder Datenspeicherressourcen beinhalten, die konfiguriert oder zugewiesen sein können, um eine oder mehrere Arbeitslasten zu unterstützen. In mindestens einer Ausführungsform können mehrere Knoten-CRs, die CPUs oder Prozessoren beinhalten, in einem oder mehreren Racks gruppiert sein, um Berechnungsressourcen bereitzustellen, um eine oder mehrere Arbeitslasten zu unterstützen. In mindestens einer Ausführungsform können ein oder mehrere Racks auch eine beliebige Anzahl von Leistungsmodulen, Kühlmodulen und Netzwerk-Switches in beliebiger Kombination beinhalten.

[0042] In mindestens einer Ausführungsform kann der Ressourcenorchestrierer **712** einen oder mehrere Knoten-CRs **716(1)-716(N)** und/oder gruppierte Berechnungsressourcen **714** konfigurieren oder anderweitig steuern. In mindestens einer Ausführungsform kann der Ressourcenorchestrierer **712** eine Softwaredesigninfrastruktur-(„SDI“-)Verwaltungsinstanz für das Rechenzentrum **700** beinhalten. In mindestens einer Ausführungsform kann der Ressourcenorchestrierer Hardware, Software oder eine Kombination davon beinhalten.

[0043] In mindestens einer Ausführungsform beinhaltet, wie in **Fig. 7** gezeigt, die Frameworkschicht **720** einen Aufgabenplaner **722**, einen Konfigurationsverwalter **724**, einen Ressourcenverwalter **726** und ein verteiltes Dateisystem **728**. In mindestens einer Ausführungsform kann die Frameworkschicht **720** ein Framework beinhalten, um Software **732** der Softwareschicht **730** und/oder eine oder mehrere Anwendungen(en) **742** der Anwendungsschicht **740** zu unterstützen. In mindestens einer Ausführungsform kann/können die Software **732** oder die Anwendung(en) **742** jeweils webbasierte Dienstsoftware oder - anwendungen beinhalten, wie etwa diejenigen, die von Amazon Web Services, Google Cloud und Microsoft Azure bereitgestellt sind. In mindestens einer Ausführungsform kann die Frameworkschicht **720** eine Art von freiem und Open-Source-Software-Webanwendungs-Framework sein, ohne darauf beschränkt zu sein, wie etwa Apache Spark™ (im Folgenden „Spark“), welches das verteilte Dateisystem **728** für umfangreiche Datenverarbeitungen (z.B. „Big Data“) nutzen kann. In mindestens einer Ausführungsform kann der Aufgabenplaner **722** einen Spark-Treiber beinhalten, um die Planung von Arbeitslasten zu erleichtern, die von verschiedenen Schichten des Rechenzentrums **700** unterstützt werden. In mindestens einer Ausführungsform kann der Konfigurationsverwalter **724** in der Lage sein, unterschiedliche Schichten zu konfigurieren, wie etwa die Softwareschicht **730** und die Frameworkschicht **720**, was Spark und das verteilte Dateisystem **728** zum Unterstützen einer umfangreichen Datenverarbeitung beinhaltet. In mindestens einer Ausführungsform kann der Ressourcenverwalter **726** in der Lage sein, geclusterte oder gruppierte Berechnungsressourcen zu verwalten, die dem verteilten Dateisystem **728** und dem Aufgabenplaner **722** zur Unterstützung zugeordnet oder zugewiesen sind. In mindestens einer Ausführungsform können geclusterte oder gruppierte Berechnungsressourcen eine gruppierte Berechnungsressource **714** auf der Rechenzentrumsinfrastrukturschicht **710** beinhalten. In mindestens einer Ausführungsform kann sich der Ressourcenverwalter **726** mit dem Ressourcenorchestrierer **712** koordinieren, um diese zugeordneten oder zugewiesenen Berechnungsressourcen zu verwalten.

[0044] In mindestens einer Ausführungsform kann die in der Softwareschicht **730** beinhaltete Software **732** Software beinhalten, die von mindestens Teilen der Knoten-CRs **716(1)-716(N)**, gruppierten Berechnungsressourcen **714** und/oder dem verteilten Dateisystem **728** der Frameworkschicht **720** verwendet wird Eine oder mehrere Arten von Software können Internet-Webseiten-Suchsoftware, E-Mail-Virenscan-Software, Datenbanksoftware und Streaming-Video-Content-Software beinhalten, ohne darauf beschränkt zu sein.

[0045] In mindestens einer Ausführungsform kann/können die Anwendung(en) **742**, die in der Anwendungsschicht **740** enthalten ist/sind, eine oder mehrere Arten von Anwendungen beinhalten, die von mindestens Tei-

len der Knoten-CRs **716(1)-716(N)**, gruppierten Berechnungsressourcen **714** und/oder dem verteilten Dateisystem **728** der Frameworkschicht **720** verwendet werden. Eine oder mehrere Arten von Anwendungen können eine beliebige Anzahl einer Genomikanwendung, einer kognitiven Rechenanwendung und einer maschinellen Lemanwendung umfassen, die Trainings- oder Ableitungssoftware beinhaltet, Framework-Software des maschinellen Lernens (z. B. PyTorch, TensorFlow, Caffe usw.) oder andere maschinelle Lernanwendungen beinhalten, ohne darauf beschränkt zu sein, die in Verbindung mit einer oder mehreren Ausführungsformen verwendet werden.

[0046] In mindestens einer Ausführungsform können Konfigurationsverwalter **724**, Ressourcenverwalter **726** und Ressourcenorchestrierer **712** eine beliebige Anzahl und Art von selbstmodifizierenden Handlungen auf Grundlage einer beliebigen Menge und Art von Daten umsetzen, die auf jede technisch machbare Weise erfasst werden. In mindestens einer Ausführungsform können selbstmodifizierende Handlungen einen Rechenzentrumsbetreiber des Rechenzentrums **700** davon befreien, möglicherweise schlechte Konfigurationsentscheidungen zu treffen, und möglicherweise vermeiden, dass Teile eines Rechenzentrums nicht ausgelastet und/oder leistungsschwach sind.

[0047] In mindestens einer Ausführungsform kann das Rechenzentrum **700** Werkzeuge, Dienste, Software oder andere Ressourcen beinhalten, um ein oder mehrere Modelle des maschinellen Lernens zu trainieren oder Informationen unter Verwendung eines oder mehrerer Modelle des maschinellen Lernens gemäß einer oder mehreren in dieser Schrift beschriebenen Ausführungsformen vorherzusagen oder abzuleiten. Zum Beispiel kann in mindestens einer Ausführungsform ein Modell des maschinellen Lernens durch das Berechnen von Gewichtungsparametern gemäß einer neuronalen Netzwerkarchitektur unter Verwendung von Software und Berechnungsressourcen trainiert werden, die vorstehend im Hinblick auf das Rechenzentrum **700** beschrieben wurden. In mindestens einer Ausführungsform können trainierte Modelle maschinellen Lernens verwendet werden, die einem oder mehreren neuronalen Netzwerken entsprechen, um Informationen unter Verwendung der vorstehend beschriebenen Ressourcen in Bezug auf das Rechenzentrum **700** durch das Verwenden von Gewichtungsparameters abzuleiten oder vorherzusagen, die durch eine oder mehrere in dieser Schrift beschriebene Trainingsmethoden berechnet wurden.

[0048] In mindestens einer Ausführungsform kann das Rechenzentrum CPUs, anwendungsspezifische integrierte Schaltkreise (ASICs), GPUs, FPGAs oder andere Hardware verwenden, um das Training und/oder die Ableitung unter Verwendung der vorstehend beschriebenen Ressourcen durchzuführen. Darüber hinaus können eine oder mehrere vorstehend beschriebene Software- und/oder Hardwareressourcen als ein Dienst konfiguriert sein, um Benutzern zu ermöglichen, zu trainieren oder das Ableiten von Informationen durchzuführen, wie etwa Bilderfassung, Spracherfassung oder andere Dienste der künstlichen Intelligenz.

[0049] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. Details bezüglich der Ableitungs- und/oder Trainingslogik **615** werden nachfolgend in Verbindung mit den **Fig. 6A** und/oder 6B bereitgestellt. In mindestens einer Ausführungsform kann eine Ableitungs- und/oder Trainingslogik **615** im System der **Fig. 7** zum Ableiten oder Vorhersagen von Operationen mindestens zum Teil auf Grundlage von Gewichtungsparametern verwendet werden, die unter Verwendung von neuronalen Netzwerktrainingsoperationen, Funktionen und/oder Architekturen von neuronalen Netzwerken oder in dieser Schrift beschriebenen Anwendungsfällen von neuronalen Netzwerken berechnet wurden.

[0050] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. In mindestens einer Ausführungsform kann diese Logik mit Komponenten dieser Figuren verwendet werden, um ein dreidimensionales Bild einer bestimmten Ansicht aus zweidimensionalen Bildern aus einer willkürlichen Anzahl von Kameras zu erzeugen.

COMPUTERSYSTEME

[0051] Fig. 8 ist ein Blockdiagramm, das ein beispielhaftes Computersystem veranschaulicht, das ein System mit miteinander verbundenen Vorrichtungen und Komponenten, ein System-on-a-Chip (SOC) oder eine Kombination davon sein kann **800**, die mit einem Prozessor gebildet ist, der Ausführungseinheiten beinhalten kann, um eine Anweisung auszuführen, gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform kann das Computersystem **800** ohne Einschränkung eine Komponente umfassen, wie etwa einen Prozessor **802**, um Ausführungseinheiten zu verwenden, die eine Logik beinhalten, um Algorithmen für Prozessdaten gemäß der vorliegenden Offenbarung durchzuführen, wie etwa in der in dieser Schrift beschrie-

benen Ausführungsform. In mindestens einer Ausführungsform kann das Computersystem **800** Prozessoren beinhalten, wie etwa PENTIUM®-Prozessorfamilie, Mikroprozessoren von Xeon™, Itanium®, XScale™ und/oder StrongARM™, Intel® Core™ oder Intel® Nervana™, die von Intel® Corporation of Santa Clara, Kalifornien, erhältlich sind, obwohl auch andere Systeme (welche PCs mit anderen Mikroprozessoren, Engineering-Workstations, Set-Top-Boxen und dergleichen beinhalten) verwendet werden können. In mindestens einer Ausführungsform kann das Computersystem **800** eine Version des WINDOWS-Betriebssystems ausführen, das von der Microsoft Corporation in Redmond, Washington, erhältlich ist, obwohl auch andere Betriebssysteme (zum Beispiel UNIX und Linux), eingebettete Software und/oder grafische Benutzeroberflächen verwendet werden können.

[0052] Ausführungsformen können in anderen Vorrichtungen verwendet werden, wie etwa Handheld-Vorrichtungen und eingebetteten Anwendungen. Einige Beispiele für Handheld-Vorrichtungen beinhalten Mobiltelefone, Internetprotokoll-Vorrichtungen, Digitalkameras, persönliche digitale Assistenten („PDAs“) und Handheld-PCs. In mindestens einer Ausführungsform können eingebettete Anwendungen eine Mikrosteuerung, einen digitalen Signalprozessor („DSP“), ein System-on-a-Chip, Netzwerkcomputer („NetPCs“), Set-Top-Boxen, Netzwerk-Hubs, ein Weitverkehrsnetzwerk-(wide area network - „WAN“-)Switches oder ein beliebiges anderes System beinhalten, das eine oder mehrere Anweisungen gemäß mindestens einer Ausführungsform ausführen kann.

[0053] In mindestens einer Ausführungsform kann das Computersystem **800** ohne Einschränkung einen Prozessor **802** beinhalten, der ohne Einschränkung eine oder mehrere Ausführungseinheiten **808** beinhalten kann, um ein Training und/oder eine Ableitung für ein Modell maschinellen Lernens gemäß in dieser Schrift beschriebenen Methoden durchzuführen. In mindestens einer Ausführungsform ist das Computersystem **800** ein Desktop mit einem einzigen Prozessor oder ein Serversystem, aber in einer weiteren Ausführungsform kann das Computersystem **800** ein Mehrprozessorsystem sein. In mindestens einer Ausführungsform kann der Prozessor **802** ohne Einschränkung Folgendes beinhalten: einen Mikroprozessor eines Computers mit komplexem Anweisungssatz (complex instruction set computer - „CISC“), einen Mikroprozessor zum Berechnen mit reduziertem Anweisungssatz (reduced instruction set computing - „RISC“), einen Mikroprozessor mit sehr langen Anweisungsworten (very long instruction word - „VLIW“), einen Prozessor, der eine Kombination von Anweisungssätzen umsetzt, oder eine beliebige andere Prozessorvorrichtung, wie etwa ein digitaler Signalprozessor. In mindestens einer Ausführungsform kann der Prozessor **802** an einen Prozessorbus **810** gekoppelt sein, der Datensignale zwischen dem Prozessor **802** und anderen Komponenten im Computersystem **800** übermitteln kann.

[0054] In mindestens einer Ausführungsform kann der Prozessor **802** ohne Einschränkung einen internen schnellen Pufferspeicher(„Zwischenspeicher“) der Ebene 1 („L1") **804** beinhalten. In mindestens einer Ausführungsform kann der Prozessor **802** einen einzelnen internen Zwischenspeicher oder mehrere Ebenen des internen Zwischenspeichers aufweisen. In mindestens einer Ausführungsform kann sich der schnelle Pufferspeicher außerhalb des Prozessors **802** befinden. Andere Ausführungsformen können auch eine Kombination von sowohl internen als auch externen Zwischenspeichern abhängig von einer bestimmten Umsetzung und Anforderungen beinhalten. In mindestens einer Ausführungsform kann die Registerdatei **806** verschiedene Arten von Daten in verschiedenen Registern speichern, die ohne Einschränkung ein Ganzzahlregister, Gleitkommaregister, Statusregister und Anweisungsverweisregister beinhalten.

[0055] In mindestens einer Ausführungsform befindet sich die Ausführungseinheit **808**, die ohne Einschränkung eine Logik zum Durchführen von Ganzzahl- und Gleitkommaoperationen beinhaltet, ebenfalls im Prozessor **802**. In mindestens einer Ausführungsform kann der Prozessor **802** außerdem einen Mikrocode-(„ucode“-)Nur-Lese-Speicher („ROM“) beinhalten, der Mikrocode für bestimmte Makrobefehle speichert. In mindestens einer Ausführungsform kann die Ausführungseinheit **808** Logik beinhalten, um einen gepackten Anweisungssatz **809** zu handhaben. In mindestens einer Ausführungsform können Operationen, die von vielen Multimediaanwendungen verwendet werden, unter Verwendung von gepackten Daten in einem Allzweckprozessor **802** durch das Beinhalten des gepackten Anweisungssatzes **809** in einen Anweisungssatz eines Allzweckprozessors **802** durchgeführt werden, zusammen mit einer zugehörigen Schaltung, um Befehle auszuführen. In einer oder mehreren Ausführungsformen können viele Multimediaanwendungen beschleunigt und effizienter durch das Verwenden der vollen Breite des Datenbusses eines Prozessors zum Ausführen von Operationen an gepackten Daten ausgeführt werden, wodurch die Notwendigkeit beseitigt werden kann, kleinere Dateneinheiten über den Datenbus des Prozessors zu übertragen, um eine oder mehrere Operationen ein Datenelement nach dem anderen durchzuführen.

[0056] In mindestens einer Ausführungsform kann die Ausführungseinheit **808** auch in Mikrosteuerungen, eingebetteten Prozessoren, Grafikvorrichtungen, DSPs und anderen Arten von Logikschaltungen verwendet werden. In mindestens einer Ausführungsform kann das Computersystem **800** ohne Einschränkung einen Speicher **820** beinhalten. In mindestens einer Ausführungsform kann der Speicher **820** als ein dynamischer Direktzugriffsspeicher (Dynamic Random Access Memory - „DRAM“), ein statischer Direktzugriffsspeicher (Static Random Access Memory - „SRAM“), eine Flash-Speichervorrichtung oder eine andere Speichervorrichtung umgesetzt sein. In mindestens einer Ausführungsform kann der Speicher **820** Anweisung(en) **819** und/oder Daten **821** speichern, die durch Datensignale dargestellt sind, die durch den Prozessor **802** ausgeführt werden können.

[0057] In mindestens einer Ausführungsform kann der Systemlogikchip an den Prozessorbus **810** und den Speicher **820** gekoppelt sein. In mindestens einer Ausführungsform kann der Systemlogikchip ohne Einschränkung einen Speichersteuerungshub (memory controller hub - „MCH“) **816** beinhalten, und der Prozessor **802** kann mit dem MCH **816** über den Prozessorbus **810** kommunizieren. In mindestens einer Ausführungsform kann der MCH **816** einen Speicherpfad mit hoher Bandbreite **818** zum Speicher **820** zur Anweisungs- und Datenspeicherung und zur Speicherung von Grafikbefehlen, Daten und Texturen bereitstellen. In mindestens einer Ausführungsform kann der MCH **816** Datensignale zwischen dem Prozessor **802**, dem Speicher **820** und anderen Komponenten im Computersystem **800** leiten und Datensignale zwischen dem Prozessorbus **810**, dem Speicher **820** und einer System-E/A **822** überbrücken. In mindestens einer Ausführungsform kann der Systemlogikchip einen Grafikport zum Koppeln an eine Grafiksteuerung bereitstellen. In mindestens einer Ausführungsform kann der MCH **816** über einen Speicherpfad mit hoher Bandbreite **818** an den Speicher **820** gekoppelt sein und die Grafik- /Videokarte **812** kann an den MCH **816** über eine Verbindung eines beschleunigten Grafikports (Accelerated Graphics Port - „AGP“) **814** gekoppelt sein.

[0058] In mindestens einer Ausführungsform kann das Computersystem **800** die System-E/A 822 verwenden, die ein proprietärer Hubschnittstellenbus ist, um den MCH **816** an den E/A-Steuerungshub (I/O controller hub - „ICH“) **830** zu koppeln. In mindestens einer Ausführungsform kann der ICH **830** über einen lokalen E/A-Bus direkte Verbindungen zu einigen E/A-Vorrichtungen bereitstellen. In mindestens einer Ausführungsform kann der lokale E/A-Bus ohne Einschränkung einen Hochgeschwindigkeits-E/A-Bus zum Verbinden von Peripheriegeräten mit dem Speicher **820**, dem Chipsatz und dem Prozessor **802** beinhalten. Beispiele können ohne Einschränkung eine Audiosteuerung **829**, einen Firmware-Hub („Flash-BIOS“) **828**, einen drahtlosen Sendeempfänger **826**, einen Datenspeicher **824**, eine Alt-E/A-Steuerung **823**, die Benutzereingabe- und Tastaturschnittstellen **825** beinhaltet, einen seriellen Erweiterungsport **827**, wie etwa universellen seriellen Bus („USB“), und eine Netzwerksteuerung **834** beinhalten. Der Datenspeicher **824** kann ein Festplattenlaufwerk, ein Diskettenlaufwerk, eine CD-ROM-Vorrichtung, ein Flash-Speichervorrichtung oder eine andere Massenspeichervorrichtung umfassen.

[0059] In mindestens einer Ausführungsform veranschaulicht **Fig. 8** ein System, das miteinander verbundene Hardwarevorrichtungen oder „Chips“ beinhaltet, wohingegen in anderen Ausführungsformen **Fig. 8** ein beispielhaftes System auf einem Chip („SoC“) veranschaulichen kann. In mindestens einer Ausführungsform können die in **Fig. cc** veranschaulichten Vorrichtungen mit proprietären Verbindungen, standardisierten Verbindungen (z. B. PCIe) oder einer Kombination davon verbunden sein. In mindestens einer Ausführungsform sind eine oder mehrere Komponenten des Computersystems **800** unter Verwendung von Compute-Express-Link-(CXL-)Verbindungen miteinander verbunden.

[0060] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. Details bezüglich der Ableitungs- und/oder Trainingslogik **615** werden nachfolgend in Verbindung mit den **Fig. 6A** und/oder 6B bereitgestellt. In mindestens einer Ausführungsform kann eine Ableitungs- und/oder Trainingslogik **615** im System der **Fig. 8** zum Ableiten oder Vorhersagen von Operationen mindestens zum Teil auf Grundlage von Gewichtungsparametern verwendet werden, die unter Verwendung von neuronalen Netzwerktrainingsoperationen, Funktionen und/oder Architekturen von neuronalen Netzwerken oder in dieser Schrift beschriebenen Anwendungsfällen von neuronalen Netzwerken berechnet wurden.

[0061] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. In mindestens einer Ausführungsform kann diese Logik mit Komponenten dieser Figuren verwendet werden, um ein dreidimensionales Bild einer bestimmten Ansicht aus zweidimensionalen Bildern aus einer willkürlichen Anzahl von Kameras zu erzeugen.

[0062] **Fig. 9** ist ein Blockdiagramm, das eine elektronische Vorrichtung **900** zum Nutzen eines Prozessors **910** gemäß mindestens einer Ausführungsform veranschaulicht. In mindestens einer Ausführungsform kann die elektronische Vorrichtung **900** zum Beispiel und ohne Einschränkung ein Notebook, ein Tower-Server, ein Rack-Server, ein Blade-Server, ein Laptop, ein Desktop, ein Tablet, eine Mobilvorrichtung, ein Telefon, ein eingebetteter Computer oder eine beliebige andere geeignete elektronische Vorrichtung sein.

[0063] In mindestens einer Ausführungsform kann das System **900** ohne Einschränkung einen Prozessor **910** beinhalten, der kommunikativ an eine beliebige geeignete Anzahl oder Art von Komponenten, Peripheriegeräten, Modulen oder Vorrichtungen gekoppelt ist. In mindestens einer Ausführungsform ist der Prozessor **910** unter Verwendung eines Busses oder einer Schnittstelle gekoppelt, wie etwa eines 1°C-Busses, eines Systemverwaltungsbusses (System Management Bus - „SMBus“), eines Busses mit geringer Pin-Anzahl (Low Pin Count - LPC), einer seriellen Peripherieschnittstelle (Serial Peripheral Interface - „SPI“), eines High-Definition-Audio-(„HDA“-)Busses, eines Serial-Advance-Technology-Attachment-(„SATA“-)Busses, eines universellen seriellen Busses („USB“) (Versionen 1, 2, 3) oder eines Busses eines universellen asynchronen Empfänger/Senders (Universal Asynchronous Receiver/Transmitter - „UART“). In mindestens einer Ausführungsform veranschaulicht **Fig. 9** ein System, das miteinander verbundene Hardwarevorrichtungen oder „Chips“ beinhaltet, wohingegen in anderen Ausführungsformen **Fig. 9** ein beispielhaftes System auf einem Chip („SoC“) veranschaulichen kann. In mindestens einer Ausführungsform können die in **Fig. 9** veranschaulichten Vorrichtungen mit proprietären Verbindungen, standardisierten Verbindungen (z. B. PCIe) oder einer Kombination davon verbunden sein. In mindestens einer Ausführungsform sind eine oder mehrere Komponenten der **Fig. 9** unter Verwendung von Compute-Express-Link-(CXL-)Verbindungen miteinander verbunden.

[0064] In mindestens einer Ausführungsform kann **Fig. 9** Folgendes beinhalten: ein Display **924**, einen Touchscreen **925**, ein Touchpad **930**, eine Nahfeldkommunikationseinheit (Near Field Communications - „NFC“) **945**, einen Sensorhub **940**, einen Wärmesensor **946**, einen Express-Chipsatz („EC“) **935**, ein Trusted-Platform-Modul („TPM“) **938**, BIOS-/Firmware-/Flash-Speicher („BIOS, FW-Flash“) **922**, ein DSP **960**, ein Laufwerk **920**, wie etwa ein Festkörperlaufwerk (Solid State Disk - „SSD“) oder ein Festplattenlaufwerk (Hard Disk Drive - „HDD“), eine drahtlose lokale Netzwerkeinheit (local area network - „WLAN“) **950**, eine Bluetooth-Einheit **952**, eine drahtlose Weitbereichsnetzwerkeinheit (Wireless Wide Area Network - „WWAN“) **956**, ein globales Positionsbestimmungssystem (GPS) **955**, eine Kamera („USB 3.0-Kamera“) **954**, wie etwa eine USB 3.0-Kamera und/oder eine Speichereinheit mit doppelter Datenrate bei niedriger Leistung (Low Power Double Data Rate - „LPDDR“) („LPDDR3“) **915**, die zum Beispiel im LPDDR3-Standard umgesetzt ist. Diese Komponenten können jeweils auf beliebige geeignete Weise umgesetzt sein.

[0065] In mindestens einer Ausführungsform können andere Komponenten kommunikativ an den Prozessor **910** durch die vorstehend erörterten Komponenten gekoppelt sein. In mindestens einer Ausführungsform können ein Beschleunigungsmesser **941**, Umgebungslichtsensor (Ambient Light Sensor - „ALS“) **942**, ein Kompass **943** und ein Gyroskop **944** kommunikativ an den Sensorhub **940** gekoppelt sein. In mindestens einer Ausführungsform können der Wärmesensor **939**, ein Lüfter **937**, eine Tastatur **946** und ein Touchpad **930** kommunikativ an den EC **935** gekoppelt sein. In mindestens einer Ausführungsform können der Lautsprecher **963**, die Kopfhörer **964** und das Mikrofon („mic“) **965** kommunikativ an eine Audioeinheit („Audiocodec und Klasse-D-Verstärker“) **962** gekoppelt sein, die wiederum kommunikativ an den DSP **960** gekoppelt sein kann. In mindestens einer Ausführungsform kann die Audioeinheit **964** zum Beispiel und ohne Einschränkung einen Audiokodierer/-dekodierer („Codec“) und einen Klasse-D-Verstärker beinhalten. In mindestens einer Ausführungsform kann die SIM-Karte („SIM“) **957** kommunikativ an die WWAN-Einheit **956** gekoppelt sein. In mindestens einer Ausführungsform können Komponenten wie die WLAN-Einheit **950** und die Bluetooth-Einheit **952** sowie die WWAN-Einheit **956** in einem Formfaktor der nächsten Generation (Next Generation Form Factor - „NGFF“) umgesetzt sein.

[0066] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. Details bezüglich der Ableitungs- und/oder Trainingslogik **615** werden nachfolgend in Verbindung mit den **Fig. 6A** und/oder 6B bereitgestellt. In mindestens einer Ausführungsform kann eine Ableitungs- und/oder Trainingslogik **615** im System der **Fig. 9** zum Ableiten oder Vorhersagen von Operationen mindestens zum Teil auf Grundlage von Gewichtungsparametern verwendet werden, die unter Verwendung von neuronalen Netzwerktrainingsoperationen, Funktionen und/oder Architekturen von neuronalen Netzwerken oder in dieser Schrift beschriebenen Anwendungsfällen von neuronalen Netzwerken berechnet wurden.

[0067] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. In mindestens einer Aus-

führungsform kann diese Logik mit Komponenten dieser Figuren verwendet werden, um ein dreidimensionales Bild einer bestimmten Ansicht aus zweidimensionalen Bildern aus einer willkürlichen Anzahl von Kameras zu erzeugen.

[0068] **Fig. 10** veranschaulicht ein Computersystem **1000** gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform ist das Computersystem **1000** dazu konfiguriert, verschiedene Prozesse und Verfahren umzusetzen, die in dieser Offenbarung beschrieben sind.

[0069] In mindestens einer Ausführungsform umfasst das Computersystem **1000** ohne Einschränkung mindestens eine Zentralverarbeitungseinheit („CPU“) **1002**, die mit einem Kommunikationsbus **1010** verbunden ist, der unter Verwendung eines beliebigen geeigneten Protokolls umgesetzt ist, wie etwa PCI („Peripheral Component Interconnect“), Peripheral Component Interconnect Express („PCI-Express“), AGP („Accelerated Graphics Port“), HyperTransport oder einem beliebigen anderen Bus- oder Punkt-zu-Punkt-Kommunikationsprotokoll. In mindestens einer Ausführungsform beinhaltet das Computersystem **1000** ohne Einschränkung einen Hauptspeicher **1004** und eine Steuerlogik (z. B. umgesetzt als Hardware, Software oder eine Kombination davon) und Daten werden im Hauptspeicher **1004** gespeichert, der die Form eines Direktzugriffsspeichers (random access memory - „RAM“) annehmen kann. In mindestens einer Ausführungsform stellt ein Netzwerkschnittstellenteilsystem („Netzwerkschnittstelle“) **1022** eine Schnittstelle zu anderen Computervorrichtungen und Netzwerken zum Empfangen von Daten von anderen Systemen und Übermitteln von Daten von dem Computersystem **1000** an anderen Systemen bereit.

[0070] In mindestens einer Ausführungsform beinhaltet das Computersystem **1000** in mindestens einer Ausführungsform ohne Einschränkung Eingabevorrichtungen **1008**, ein Parallelverarbeitungssystem **1012** und Anzeigevorrichtungen **1006**, die unter Verwendung einer herkömmlichen Kathodenstrahlröhre (cathode ray tube - „CRT“), Flüssigkristallanzeige (liquid crystal display - „LCD“), Leuchtdiode (light emitting diode - „LED“), Plasmaanzeige oder anderen geeigneten Anzeigetechnologien umgesetzt sein können. In mindestens einer Ausführungsform wird eine Benutzereingabe von Eingabevorrichtungen **1008** empfangen, wie etwa Tastatur, Maus, Touchpad, Mikrofon und mehr. In mindestens einer Ausführungsform kann sich jedes der vorstehenden Module auf einer einzigen Halbleiterplattform befinden, um ein Verarbeitungssystem zu bilden.

[0071] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. Details bezüglich der Ableitungs- und/oder Trainingslogik **615** werden nachfolgend in Verbindung mit den **Fig. 6A** und/oder 6B bereitgestellt. In mindestens einer Ausführungsform kann eine Ableitungs- und/oder Trainingslogik **615** im System der **Fig. 10** zum Ableiten oder Vorhersagen von Operationen mindestens zum Teil auf Grundlage von Gewichtungsparametern verwendet werden, die unter Verwendung von neuronalen Netzwerktrainingsoperationen, Funktionen und/oder Architekturen von neuronalen Netzwerken oder in dieser Schrift beschriebenen Anwendungsfällen von neuronalen Netzwerken berechnet wurden.

[0072] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. In mindestens einer Ausführungsform kann diese Logik mit Komponenten dieser Figuren verwendet werden, um ein dreidimensionales Bild einer bestimmten Ansicht aus zweidimensionalen Bildern aus einer willkürlichen Anzahl von Kameras zu erzeugen.

[0073] **Fig. 11** veranschaulicht ein Computersystem **1100** gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform beinhaltet das Computersystem **1100** ohne Einschränkung einen Computer **1110** und einen USB-Stick **1120**. In mindestens einer Ausführungsform kann der Computer **1110** ohne Einschränkung eine beliebige Anzahl und Art von Prozessor(en) (nicht gezeigt) und einen Speicher (nicht gezeigt) beinhalten. In mindestens einer Ausführungsform beinhaltet der Computer **1110** ohne Einschränkung einen Server, eine Cloud-Instanz, einen Laptop und einen Desktop-Computer.

[0074] In mindestens einer Ausführungsform beinhaltet der USB-Stick **1120** ohne Einschränkung eine Verarbeitungseinheit **1130**, eine USB-Schnittstelle **1140** und eine USB-Schnittstellenlogik **1150**. In mindestens einer Ausführungsform kann die Verarbeitungseinheit **1130** ein beliebiges Befehlsausführungssystem, eine Einrichtung oder eine Vorrichtung sein, die Befehle ausführen kann. In mindestens einer Ausführungsform kann die Verarbeitungseinheit **1130** ohne Einschränkung eine beliebige Anzahl und Art von Verarbeitungskernen (nicht gezeigt) beinhalten. In mindestens einer Ausführungsform umfasst der Verarbeitungskern **1130** eine anwendungsspezifische integrierte Schaltung (application specific integrated circuit - „ASIC“), die optimiert ist, um jede Menge und Art von Operationen durchzuführen, die mit maschinellem Lernen verbunden sind. Zum Bei-

spiel ist in mindestens einer Ausführungsform der Verarbeitungskern **1130** eine Tensor-Verarbeitungseinheit (tensor processing unit - „TPC“), die optimiert ist, um Ableitungsoperationen des maschinellen Lernens durchzuführen. In mindestens einer Ausführungsform ist der Verarbeitungskern **1130** eine Bildverarbeitungseinheit (vision processing unit - „VPU“), die optimiert ist, um Operationen maschinelles Sehen und der Ableitung durch maschinelles Lernen durchzuführen.

[0075] In mindestens einer Ausführungsform kann die USB-Schnittstelle **1140** eine beliebige Art von USB-Stecker oder USB-Buchse sein. Zum Beispiel ist in mindestens einer Ausführungsform die USB-Schnittstelle **1140** eine USB-3.0-Typ-C-Buchse für Daten und Leistung. In mindestens einer Ausführungsform ist die USB-Schnittstelle **1140** ein USB-3.0-Typ-A-Anschluss. In mindestens einer Ausführungsform kann die USB-Schnittstellenlogik **1150** eine beliebige Menge und eine beliebige Art von Logik beinhalten, die es der Verarbeitungseinheit **1130** ermöglicht, mit Vorrichtungen (z. B. dem Computer **1110**) über den USB-Anschluss **1140** eine Schnittstelle zu bilden.

[0076] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. Details bezüglich der Ableitungs- und/oder Trainingslogik **615** werden nachfolgend in Verbindung mit den **Fig. 6A** und/oder 6B bereitgestellt. In mindestens einer Ausführungsform kann eine Ableitungs- und/oder Trainingslogik **615** im System der **Fig. 11** zum Ableiten oder Vorhersagen von Operationen mindestens zum Teil auf Grundlage von Gewichtungsparametern verwendet werden, die unter Verwendung von neuronalen Netzwerktrainingsoperationen, Funktionen und/oder Architekturen von neuronalen Netzwerken oder in dieser Schrift beschriebenen Anwendungsfällen von neuronalen Netzwerken berechnet wurden.

[0077] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. In mindestens einer Ausführungsform kann diese Logik mit Komponenten dieser Figuren verwendet werden, um ein dreidimensionales Bild einer bestimmten Ansicht aus zweidimensionalen Bildern aus einer willkürlichen Anzahl von Kameras zu erzeugen.

[0078] **Fig. 12A** veranschaulicht eine beispielhafte Architektur, in der eine Vielzahl von GPUs **1210-1213** kommunikativ an eine Vielzahl von Mehrkernprozessoren **1205-1206** über Hochgeschwindigkeitsverbindungen **1240-1243** (z. B. Busse, Punkt-zu-Punkt-Verbindungen usw.) gekoppelt ist. In einer Ausführungsform unterstützen die Hochgeschwindigkeitsverbindungen **1240-1243** einen Kommunikationsdurchsatz von 4 GB/s, 30 GB/s, 80 GB/s oder mehr. Es können verschiedene Verbindungsprotokolle verwendet werden, die PCIe 4.0 oder 5.0 und NVLink 2.0 beinhalten, ohne darauf beschränkt zu sein.

[0079] Zusätzlich und in einer Ausführungsform sind zwei oder mehr der GPUs **1210-1213** über Hochgeschwindigkeitsverbindungen **1229-1230** miteinander verbunden, die unter Verwendung derselben oder anderer Protokolle/Verbindungen als denen, die für die Hochgeschwindigkeitsverbindungen **1240-1243** verwendet werden, umgesetzt sein können. Gleichermaßen können zwei oder mehr Mehrkernprozessoren **1205-1206** über eine Hochgeschwindigkeitsverbindung **1228** verbunden sein, die symmetrische Mehrprozessor-(symmetric multi-processor - SMP-)Busse sein können, die mit 20 GB/s, 30 GB/s, 120 GB/s oder mehr betrieben werden. Alternativ kann die gesamte Kommunikation zwischen verschiedenen Systemkomponenten, die in **Fig. 12A** gezeigt sind, unter Verwendung derselben Protokolle/Verbindungen (z. B. über eine gemeinsame Verbindungsstruktur) erreicht werden.

[0080] In einer Ausführungsform ist jeder Mehrkernprozessor **1205-1206** über Speicherverbindungen **1226-1227** jeweils kommunikativ an einen Prozessorspeicher **1201-1202** gekoppelt, und jede GPU **1210-1213** ist jeweils über GPU-Speicherverbindungen **1250-1253** kommunikativ an den GPU-Speicher **1220-1223** gekoppelt. Die Speicherverbindungen **1226-1227** und **1250-1253** können selbe oder unterschiedliche Speicherzugriffstechnologien nutzen. Als Beispiel und nicht als Einschränkung können Prozessorspeicher **1201-1202** und GPU-Speicher **1220-1223** flüchtige Speicher sein, wie etwa dynamische Direktzugriffsspeicher (DRAMs) (die gestapelte DRAMs beinhalten), Grafik-DDR-SDRAM (GDDR) (z. B. GDDR5, GDDR6) oder High Bandwidth Memory (HBM) und/oder können nichtflüchtige Speicher, wie etwa 3D XPoint oder Nano-Ram, sein. In einer Ausführungsform kann ein Teil der Prozessorspeicher **1201-1202** flüchtiger Speicher sein und ein anderer Teil kann nichtflüchtiger Speicher sein (z. B. unter Verwendung einer Zwei-Ebenen-Speicher-(2LM-)Hierarchie).

[0081] Wie nachfolgend beschrieben, kann, obwohl verschiedene Prozessoren **1205-1206** und GPUs **1210-1213** jeweils physisch an einen bestimmten Speicher **1201-1202**, **1220-1223**, gekoppelt sind, ein einheit-

liche Speicherarchitektur umgesetzt sein, bei der ein selber Adressbereich des virtuellen Systems (auch als „effektiver Adressbereich“ bezeichnet) auf verschiedene physische Speicher verteilt ist. Zum Beispiel können die Prozessorspeicher **1201- 1202** jeweils 64 GB Adressbereich des Systemspeichers umfassen und die GPU-Speicher **1220-1223** können jeweils 32 GB Adressbereich des Systemspeichers umfassen (was in diesem Beispiel zu insgesamt **256** GB adressierbarem Speicher führt).

[0082] Fig. 12B veranschaulicht zusätzliche Details für eine Verbindung zwischen einem Mehrkernprozessor **1207** und einem Grafikbeschleunigungsmodul **1246** gemäß einer beispielhaften Ausführungsform. Das Grafikbeschleunigungsmodul **1246** kann einen oder mehrere GPU-Chips beinhalten, die auf einer Leitungskarte integriert sind, die über eine Hochgeschwindigkeitsverbindung **1240** an den Prozessor **1207** gekoppelt ist. Alternativ kann das Grafikbeschleunigungsmodul **1246** auf einem selben Gehäuse oder Chip wie der Prozessor **1207** integriert sein.

[0083] In mindestens einer Ausführungsform beinhaltet der veranschaulichte Prozessor **1207** eine Vielzahl von Kernen **1260A-1260D**, jeder mit einem Adressenübersetzungspuffer **1261A-1261D** und einem oder mehreren Zwischenspeichern **1262A-1262D**. In mindestens einer Ausführungsform können die Kerne **1260A-1260D** verschiedene andere Komponenten zum Ausführen von Anweisungen und zum Verarbeiten von Daten beinhalten, die nicht veranschaulicht sind. Die Zwischenspeicher **1262A-1262D** können Zwischenspeicher der Ebene 1 (L1) und der Ebene 2 (L2) umfassen. Zusätzlich können ein oder mehrere gemeinsam genutzte Zwischenspeicher **1256** in den Zwischenspeichern **1262A-1262D** beinhaltet sein und von den Sätzen von Kernen **1260A-1260D** gemeinsam genutzt werden. Zum Beispiel beinhaltet eine Ausführungsform des Prozessors **1207** 24 Kerne, jeder mit seinem eigenen L1-Zwischenspeicher, zwölf gemeinsam genutzten L2-Zwischenspeichern und zwölf gemeinsam genutzten L3-Zwischenspeichern. In dieser Ausführungsform werden ein oder mehrere L2- und L3-Zwischenspeicher von zwei benachbarten Kernen gemeinsam genutzt. Der Prozessor **1207** und das Grafikbeschleunigungsmodul **1246** sind mit dem Systemspeicher **1214** verbunden, der die Prozessorspeicher **1201-1202** der **Fig. 12A** beinhalten kann.

[0084] Die Kohärenz für Daten und Anweisungen, die in verschiedenen Zwischenspeichern **1262A-1262D**, **1256** und dem Systemspeicher **1214** gespeichert sind, wird über eine Kommunikation zwischen den Kernen über einen Kohärenzbus **1264** aufrechterhalten. Zum Beispiel kann jeder Zwischenspeicher eine Zwischenspeicher-Kohärenz-Logik/-Schaltung aufweisen, die damit verbunden ist, um als Reaktion auf erfasste Lese- oder Schreibvorgänge in bestimmte Zwischenspeicher-Zeilen über den Kohärenzbus **1264** damit zu kommunizieren. In einer Umsetzung wird ein Zwischenspeicher-Abhörprotokoll über den Kohärenzbus **1264** umgesetzt, um Zwischenspeicher-Zugriffe abzuhören.

[0085] In einer Ausführungsform koppelt eine Proxy-Schaltung **1225** das Grafikbeschleunigungsmodul **1246** kommunikativ an den Kohärenzbus **1264**, was es dem Grafikbeschleunigungsmodul **1246** ermöglicht, an einem Zwischenspeicher-Kohärenzprotokoll als eine Gegenstelle der Kerne **1260A-1260D** teilzunehmen. Insbesondere stellt eine Schnittstelle **1235** eine Anbindung zur Proxy-Schaltung **1225** über eine Hochgeschwindigkeitsverbindung **1240** (z. B. einen PCIe-Bus, NVLink usw.) bereit, und eine Schnittstelle **1237** verbindet das Grafikbeschleunigungsmodul **1246** mit der Verbindung **1240**.

[0086] In einer Umsetzung stellt eine Beschleunigerintegrationsschaltung **1236** Zwischenspeicherverwaltungs-, Speicherzugriffs-, Kontextverwaltungs- und Unterbrechungsverwaltungsdienste im Namen einer Vielzahl von Grafikverarbeitungsengines **1231**, **1232**, N des Grafikbeschleunigungsmoduls **1246** bereit. Die Grafikverarbeitungsengines **1231**, **1232**, N können jeweils eine getrennte Grafikverarbeitungseinheit (GPU) umfassen. Alternativ können die Grafikverarbeitungsengines **1231**, **1232**, N verschiedene Arten von Grafikverarbeitungsengines innerhalb einer GPU umfassen, wie etwa Grafikausführungseinheiten, Medienverarbeitungsengines (z. B. Videokodierer/-dekodierer), Abtaster und Blitengines. In mindestens einer Ausführungsform kann das Grafikbeschleunigungsmodul **1246** eine GPU mit einer Vielzahl von Grafikverarbeitungsengines **1231-1232**, N oder Grafikverarbeitungsengines **1231-1232** sein, N können einzelne GPUs sein, die auf einem gemeinsamen Gehäuse, einer gemeinsamen Leitungskarte oder einem gemeinsamen Chip integriert sind.

[0087] In einer Ausführungsform beinhaltet die Beschleunigerintegrationsschaltung **1236** eine Speicherverwaltungseinheit (memory management unit - MMU) **1239** zum Durchführen verschiedener Speicherverwaltungsfunktionen, wie etwa Übersetzungen von virtuellem in physischen Speicher (auch als Übersetzungen von effektivem in realen Speicher bezeichnet) und Speicherzugriffsprotokolle zum Zugreifen auf den Systemspeicher **1214**. Die MMU **1239** kann auch einen Adressenübersetzungspuffer (TLB) (nicht gezeigt) zum Zwischenspeichern von Übersetzungen von virtuellen/effektiven in physische/reale Adressen beinhalten. In einer Umsetzung speichert ein Zwischenspeicher **1238** Befehle und Daten für einen effizienten Zugriff durch die

Grafikverarbeitungsengines **1231-1232**, N. In einer Ausführungsform werden die im Zwischenspeicher **1238** und den Grafikspeichern **1233-1234**, M gespeicherten Daten mit den Kernzwischenspeichern **1262A-1262D**, **1256** und dem Systemspeicher **1214** kohärent gehalten. Wie vorstehend erwähnt kann dies über die Proxy-Schaltung **1225** im Namen des Zwischenspeichers **1238** und der Speicher **1233-1234**, M erreicht werden (z. B. das Senden von Aktualisierungen an den Zwischenspeicher **1238** in Bezug auf Modifikationen/Zugriffe von Zwischenspeicher-Zeilen auf die Prozessor-Zwischenspeicher **1262A-1262D**, **1256** und das Empfangen von Aktualisierungen vom Zwischenspeicher **1238**).

[0088] Ein Satz von Registern **1245** speichert Kontextdaten für Threads, die von den Grafikverarbeitungsengines **1231-1232**, N ausgeführt werden, und eine Kontextverwaltungsschaltung **1248** verwaltet Thread-Kontexte. Zum Beispiel kann die Kontextverwaltungsschaltung **1248** Sicherungs- und Wiederherstellungsoperationen ausführen, um Kontexte verschiedener Threads während Kontextwechseln zu sichern und wiederherzustellen (z. B. wenn ein erster Thread gespeichert wird und ein zweiter Thread gespeichert wird, so dass ein zweiter Thread durch eine Grafikverarbeitungsengine ausgeführt werden kann). Zum Beispiel kann die Kontextverwaltungsschaltung **1248** bei einem Kontextwechsel aktuelle Registerwerte in einem bestimmten Bereich im Speicher speichern (z. B. durch einen Kontextverweis identifiziert). Er kann dann Registerwerte wiederherstellen, wenn er zu einem Kontext zurückkehrt. In einer Ausführungsform empfängt und verarbeitet eine Unterbrechungsverwaltungsschaltung **1247** von Systemvorrichtungen empfangene Unterbrechungen.

[0089] In einer Umsetzung werden virtuelle/effektive Adressen von einer Grafikverarbeitungsengine **1231** durch die MMU **1239** in reale/physische Adressen im Systemspeicher **1214** übersetzt. Eine Ausführungsform der Beschleunigerintegrationsschaltung **1236** unterstützt mehrere (z. B. 4, 8, 16) Grafikbeschleunigermodule **1246** und/oder andere Beschleunigervorrichtungen. Das Grafikbeschleunigermodul **1246** kann einer einzelnen Anwendung gewidmet sein, die auf dem Prozessor **1207** ausgeführt wird, oder kann von mehreren Anwendungen gemeinsam genutzt werden. In einer Ausführungsform wird eine virtualisierte Grafikausführungsumgebung dargestellt, in der Ressourcen der Grafikverarbeitungsengines **1231-1232**, N mit mehreren Anwendungen oder virtuellen Maschinen (VMs) gemeinsam genutzt werden. In mindestens einer Ausführungsform können Ressourcen in „Slices“ unterteilt werden, die unterschiedlichen VMs und/oder Anwendungen auf Grundlage Verarbeitungsanforderungen und -prioritäten zugewiesen sind, die mit VMs und/oder Anwendungen verbunden sind.

[0090] In mindestens einer Ausführungsform dient die Beschleunigerintegrationsschaltung **1236** als eine Brücke zu einem System für das Grafikbeschleunigungsmodul **1246** und stellt Adressenübersetzungs- und Systemspeicher-Zwischenspeicherdienste bereit. Außerdem kann die Beschleunigerintegrationsschaltung **1236** Virtualisierungseinrichtungen für einen Hostprozessor bereitstellen, um die Virtualisierung der Grafikverarbeitungsengines **1231-1232**, N, Unterbrechungen und Speicherverwaltung zu verwalten.

[0091] Da Hardwareressourcen der Grafikverarbeitungsengines **1231-1232**, N ausdrücklich einem realen Adressbereich zugeordnet sind, den der Hostprozessor **1207** sieht, kann ein beliebiger Hostprozessor diese Ressourcen direkt unter Verwendung eines effektiven Adresswerts ansteuern. Eine Funktion der Beschleunigerintegrationsschaltung **1236** ist in einer Ausführungsform die physische Trennung der Grafikverarbeitungsengines **1231-1232**, N, so dass sie einem System als unabhängige Einheiten erscheinen.

[0092] In mindestens einer Ausführungsform sind ein oder mehrere Grafikspeicher **1233-1234**, M an jede der Grafikverarbeitungsengines **1231-1232** bzw. N gekoppelt. Grafikspeicher **1233- 1234**, M speichern Anweisungen und Daten, die von jeder der Grafikverarbeitungsengines **1231-1232**, N verarbeitet werden. Die Grafikspeicher **1233-1234**, M können flüchtige Speicher sein, wie etwa DRAMs (was gestapelte DRAMs beinhaltet), GDDR Speicher (z. B. GDDR5, GDDR6) oder HBM und/oder können nichtflüchtige Speicher sein, wie etwa 3D XPoint oder Nano-Ram.

[0093] Um den Datenverkehr über die Verbindung **1240** zu verringern, werden Verzerrungsmethoden verwendet, um sicherzustellen, dass die in den Grafikspeichern **1233-1234**, M gespeicherten Daten Daten sind, die am häufigsten von den Grafikverarbeitungsengines **1231-1232**, N verwendet und vorzugsweise nicht von den Kernen **1260A-1260D** (zumindest nicht häufig) verwendet werden. Gleichermaßen versucht ein Verzerrungsmechanismus, Daten, die von Kernen (und vorzugsweise nicht von den Grafikverarbeitungsengines **1231-1232**, N) benötigt werden, innerhalb der Zwischenspeicher **1262A-1262D**, **1256** der Kerne und des Systemspeichers **1214** zu halten.

[0094] **Fig. 12C** veranschaulicht eine weitere beispielhafte Ausführungsform, in der die Beschleunigerintegrationsschaltung **1236** in den Prozessor **1207** integriert ist. Zumindest in dieser Ausführungsform kommunizieren die Grafikverarbeitungsengines **1231-1232**, N über die Schnittstelle **1237** und die Schnittstelle **1235** (die wie-

derum jede Form von Bus oder Schnittstellenprotokoll nutzen können) direkt über die Hochgeschwindigkeitsverbindung **1240** mit der Beschleunigerintegrationsschaltung **1236**. Die Beschleunigerintegrationsschaltung **1236** kann die gleichen Operationen wie die im Hinblick auf **Fig. 12B** beschriebenen durchführen, jedoch möglicherweise mit einem höheren Durchsatz aufgrund seiner Nähe zum Kohärenzbus **1264** und den Zwischenspeichern **1262A-1262D**, **1256**. Mindestens eine Ausführungsform unterstützt unterschiedliche Programmiermodelle, die ein Programmiermodell für dedizierte Prozesse (keine Virtualisierung des Grafikbeschleunigungsmoduls) und gemeinsam genutzte Programmiermodelle (mit Virtualisierung) beinhalten, die Programmiermodelle, die durch die Beschleunigerintegrationsschaltung **1236** gesteuert werden, und Programmiermodelle beinhalten können, die durch das Grafikbeschleunigungsmodul **1246** gesteuert werden.

[0095] In mindestens einer Ausführungsform sind die Grafikverarbeitungsengines **1231-1232**, N für eine einzelne Anwendung oder einen einzelnen Prozess unter einem einzigen Betriebssystem bestimmt. In mindestens einer Ausführungsform kann eine einzelne Anwendung andere Anwendungsanforderungen an die Grafikverarbeitungsengines **1231-1232**, N weiterleiten, wodurch eine Virtualisierung innerhalb einer VM/Partition bereitgestellt wird.

[0096] In mindestens einer Ausführungsform können die Grafikverarbeitungsengines **1231-1232**, N von mehreren VM-/Anwendungspartitionen gemeinsam genutzt werden. In mindestens einer Ausführungsform können gemeinsam genutzte Modelle einen Systemhypervisor verwenden, um die Grafikverarbeitungsengines **1231-1232**, N zu virtualisieren, um den Zugriff durch jedes Betriebssystem zu ermöglichen. Bei Einzelpartitionssystemen ohne einen Hypervisor gehören die Grafikverarbeitungsengines **1231-1232**, N einem Betriebssystem. In mindestens einer Ausführungsform kann ein Betriebssystem die Grafikverarbeitungsengines **1231-1232**, N virtualisieren, um Zugriff auf jeden Prozess oder jede Anwendung bereitzustellen.

[0097] In mindestens einer Ausführungsform wählt das Grafikbeschleunigungsmodul **1246** oder eine einzelne Grafikverarbeitungsengine **1231-1232**, N ein Prozesselement unter Verwendung eines Prozesshandles aus. In mindestens einer Ausführungsform werden Prozesselemente im Systemspeicher **1214** gespeichert und sind unter Verwendung von in dieser Schrift beschriebenen effektiven Adress-zu-Real-Adressübersetzungsmethoden ansteuerbar. In mindestens einer Ausführungsform kann ein Prozesshandle ein umsetzungsspezifischer Wert sein, der einem Hostprozess beim Registrieren seines Kontexts bei der Grafikverarbeitungsengine **1231-1232**, N (d. h. das Aufrufen der Systemsoftware, um ein Prozesselement zu einer verlinkten Prozesselementliste hinzuzufügen) bereitgestellt wird. In mindestens einer Ausführungsform können niedrigere 16 Bits eines Prozesshandles ein Versatz eines Prozesselements innerhalb einer verlinkten Prozesselementliste sein.

[0098] **Fig. 12D** veranschaulicht einen beispielhaften Beschleunigerintegrations-Slice **1290**. Wie in dieser Schrift verwendet, umfasst ein „Slice“ einen spezifizierten Teil von Verarbeitungsressourcen der Beschleunigerintegrationsschaltung **1236**. Der effektive Adressbereich **1282** der Anwendung innerhalb des Systemspeichers **1214** speichert Prozesselemente **1283**. In einer Ausführungsform werden die Prozesselemente **1283** als Reaktion auf GPU-Aufrufe **1281** von Anwendungen **1280** gespeichert, die auf dem Prozessor **1207** ausgeführt werden. Ein Prozesselement **1283** enthält einen Prozesszustand für die entsprechende Anwendung **1280**. Ein im Prozesselement **1283** enthaltener Arbeitsdeskriptor (work descriptor - WD) **1284** kann eine einzelne Aufgabe sein, die von einer Anwendung angefordert wird, oder kann einen Verweis auf eine Warteschlange von Aufgaben enthalten. In mindestens einer Ausführungsform ist der WD **1284** ein Verweis auf eine Aufgabenanforderungswarteschlange im Adressbereich **1282** einer Anwendung.

[0099] Das Grafikbeschleunigungsmodul **1246** und/oder einzelne Grafikverarbeitungsengines **1231-1232**, N können von allen oder einer Teilmenge von Prozessen in einem System gemeinsam genutzt werden. In mindestens einer Ausführungsform kann eine Infrastruktur zum Einrichten des Prozesszustands und zum Senden eines WD **1284** an ein Grafikbeschleunigungsmodul **1246** zum Starten einer Aufgabe in einer virtualisierten Umgebung beinhaltet sein.

[0100] In mindestens einer Ausführungsform ist ein Programmiermodell für einen dedizierten Prozess umsetzungsspezifisch. In diesem Modell besitzt ein einzelner Prozess das Grafikbeschleunigungsmodul **1246** oder eine einzelne Grafikverarbeitungsengine **1231**. Da das Grafikbeschleunigungsmodul **1246** einem einzelnen Prozess gehört, initialisiert ein Hypervisor die Beschleunigerintegrationsschaltung **1236** für eine besitzende Partition und ein Betriebssystem initialisiert die Beschleunigerintegrationsschaltung **1236** für einen besitzenden Prozess, wenn das Grafikbeschleunigungsmodul **1246** zugewiesen ist.

[0101] Im Betrieb ruft eine WD-Abrufeinheit **1291** im Beschleunigerintegrations-Slice **1290** den nächsten WD **1284** ab, der eine Angabe der Arbeit beinhaltet, die von einer oder mehreren Grafikverarbeitungsengines des

Grafikbeschleunigungsmoduls **1246** zu erledigen ist. Daten von dem WD **1284** können in den Registern **1245** gespeichert und von der MMU **1239**, der Unterbrechungsverwaltungsschaltung **1247** und/oder der Kontextverwaltungsschaltung **1248** wie veranschaulicht verwendet werden. Zum Beispiel beinhaltet eine Ausführungsform der MMU **1239** eine Segment-/Page-Walk-Schaltung zum Zugreifen auf Segment-/Page-Tabellen **1286** innerhalb des virtuellen Adressbereichs **1285** des Betriebssystems. Die Unterbrechungsverwaltungsschaltung **1247** kann Unterbrechungsereignisse **1292** verarbeiten, die von dem Grafikbeschleunigungsmodul **1246** empfangen wurden. Beim Durchführen von Grafikoperationen wird eine effektive Adresse **1293**, die von einer Grafikverarbeitungsengine **1231-1232** erzeugt wird, N von der MMU **1239** in eine reale Adresse übersetzt.

[0102] In einer Ausführungsform wird ein selber Satz von Registern **1245** für jede Grafikverarbeitungsengine **1231-1232**, N und/oder Grafikbeschleunigungsmodul **1246** dupliziert und kann durch einen Hypervisor oder ein Betriebssystem initialisiert werden. Jedes dieser duplizierten Register kann in einem Beschleunigerintegrations-Slice **1290** beinhaltet sein. Tabelle 1 zeigt beispielhafte Register, die von einem Hypervisor initialisiert werden können.

Tabelle 1 - durch Hypervisor initialisierte Register

1	Slice-Steuerregister
2	Reale Adresse (RA) des Verweises geplanter Prozessbereiche
3	Berechtigungsmaskenüberschreibungsregister
4	Unterbrechungsvektortabellen-Eintragsversatz
5	Eingabelimit für Unterbrechungsvektortabellen
6	Staatliches Register
7	Logische Partitions-ID
8	Reale Adresse (RA) des Verweises des Hypervisor-Beschleunigungsnutzungsdatensatzes
9	Registerspeicherbeschreibung

[0103] Tabelle 2 zeigt beispielhafte Register, die von einem Betriebssystem initialisiert werden können.

Tabelle 2 - durch Betriebssystem initialisierte Register

1	Prozess- und Threadidentifizierung
2	Effektive Adresse (EA) des Verweises der Kontextspeicherung/- wiederherstellung
3	Virtuelle Adresse (VA) des Verweises des Beschleunigungsnutzungsdatensatzes
4	Virtuelle Adresse (VA) des Verweises der Speichersegmenttabelle
5	Berechtigungsmaske
6	Arbeitsdeskriptor

[0104] In einer Ausführungsform ist jeder WD **1284** spezifisch für ein bestimmtes Grafikbeschleunigungsmodul **1246** und/oder bestimmte Grafikverarbeitungsengines **1231-1232**, N. Er enthält alle Informationen, die von einer Grafikverarbeitungsengine **1231-1232**, N benötigt werden, um Arbeit zu erledigen, oder es kann ein Verweis auf einen Speicherort sein, an dem eine Anwendung eine Befehlswarteschlange mit zu erledigender Arbeit eingerichtet hat.

[0105] **Fig. 12E** veranschaulicht zusätzliche Details für eine beispielhafte Ausführungsform eines gemeinsam genutzten Modells. Diese Ausführungsform beinhaltet einen realen Adressbereich **1298** des Hypervisors, in dem eine Prozesselementliste **1299** gespeichert ist. Auf den realen Adressbereich **1298** des Hypervisors kann über einen Hypervisor **1296** zugegriffen werden, der Grafikbeschleunigungsmodulengines für das Betriebssystem **1295** virtualisiert.

[0106] In mindestens einer Ausführungsform ermöglichen gemeinsam genutzte Programmiermodelle, dass alle oder eine Teilmenge von Prozessen aus allen oder einer Teilmenge von Partitionen in einem System ein Grafikbeschleunigungsmodul **1246** verwenden. Es gibt zwei Programmiermodelle, bei denen das Grafikbe-

schleunigungsmodul **1246** von mehreren Prozessen und Partitionen gemeinsam genutzt wird: zeitlich geslicte gemeinsame Nutzung und grafikgeleitete gemeinsame Nutzung.

[0107] In diesem Modell gehört dem Systemhypervisor **1296** das Grafikbeschleunigungsmodul **1246** und er stellt seine Funktion allen Betriebssystemen **1295** zur Verfügung. Damit ein Grafikbeschleunigungsmodul **1246** die Virtualisierung durch den Systemhypervisor **1296** unterstützt, kann das Grafikbeschleunigungsmodul **1246** Folgendes beachten: 1) Die Aufgabenanforderung einer Anwendung muss autonom sein (das heißt, der Status muss zwischen den Aufträgen nicht aufrechterhalten werden) oder das Grafikbeschleunigungsmodul **1246** muss einen Kontextsicherungs- und -wiederherstellungsmechanismus bereitstellen. 2) Das Grafikbeschleunigungsmodul **1246** garantiert, dass eine Aufgabenanforderung einer Anwendung in einer bestimmten Zeit abgeschlossen wird, was etwaige Übersetzungsfehler beinhaltet, oder das Grafikbeschleunigungsmodul **1246** stellt eine Fähigkeit bereit, einer Verarbeitung einer Aufgabe zuvorzukommen. 3) Dem Grafikbeschleunigungsmodul **1246** muss die Fairness zwischen den Prozessen garantiert werden, wenn es in einem geleiteten gemeinsam genutzten Programmiermodell betrieben wird.

[0108] In mindestens einer Ausführungsform ist die Anwendung **1280** erforderlich, um einen Systemaufruf des Betriebssystems **1295** mit einer Art eines Grafikbeschleunigungsmoduls **1246**, einem Arbeitsdeskriptor (WD), einem Berechtigungsmaskenregisterwert (authority mask register - AMR) und einem Kontextsicherungs-/-wiederherstellungsbereichsverweis (context save/restore area pointer - CSRP) durchzuführen. In mindestens einer Ausführungsform beschreibt die Art des Grafikbeschleunigungsmoduls **1246** eine gezielte Beschleunigungsfunktion für einen Systemaufruf. In mindestens einer Ausführungsform kann die Art des Grafikbeschleunigungsmoduls **1246** ein systemspezifischer Wert sein. In mindestens einer Ausführungsform ist der WD speziell für das Grafikbeschleunigungsmodul **1246** formatiert und kann in Form eines Befehls des Grafikbeschleunigungsmoduls **1246**, eines effektiven Adressverweises auf eine benutzerdefinierte Struktur, eines effektiven Adressverweises auf eine Warteschlange von Befehlen, oder eine beliebige andere Datenstruktur vorliegen, um die vom Grafikbeschleunigungsmodul **1246** auszuführende Arbeit zu beschreiben. In einer Ausführungsform ist ein AMR-Wert ein AMR-Zustand, der für einen aktuellen Prozess zu verwenden ist. In mindestens einer Ausführungsform ähnelt ein an ein Betriebssystem übergebener Wert einer Anwendung, die einen AMR festlegt. Wenn Umsetzungen der Beschleunigerintegrationsschaltung **1236** und des Grafikbeschleunigungsmoduls **1246** ein Benutzerberechtigungsmaskenüberschreibungsregister (User Authority Mask Override Register-UAMOR) nicht unterstützen, kann ein Betriebssystem einen aktuellen UAMOR-Wert auf einen AMR-Wert anwenden, bevor ein AMR in einem Hypervisor-Aufruf übergeben wird. Der Hypervisor **1296** kann wahlweise einen aktuellen Wert des Berechtigungsmaskenüberschreibungsregisters (Authority Mask Override Register - AMOR) anwenden, bevor ein AMR in dem Prozesselement **1283** platziert wird. In mindestens einer Ausführungsform ist der CSRP eines der Register **1245**, das eine effektive Adresse eines Bereichs im effektiven Adressbereich **1282** einer Anwendung enthält, damit das Grafikbeschleunigungsmodul **1246** den Kontextzustand speichert und wiederherstellt. Dieser Verweis ist optional, wenn kein Status zwischen Aufgaben gespeichert werden muss oder wenn eine Aufgabe vorzeitig beendet wird. In mindestens einer Ausführungsform kann der Kontextsicherungs-/-wiederherstellungsbereich ein fixierter Systemspeicher sein.

[0109] Beim Empfang eines Systemaufrufs kann das Betriebssystem **1295** überprüfen, ob die Anwendung **1280** registriert wurde und die Berechtigung erhalten hat, das Grafikbeschleunigungsmodul **1246** zu verwenden. Das Betriebssystem **1295** ruft dann den Hypervisor **1296** mit den in Tabelle 3 gezeigten Informationen auf.

Tabelle 3 - Parameter für Aufrufe vom Betriebssystem an den Hypervisor

1	Ein Arbeitsdeskriptor (WD)
2	Ein Wert des Berechtigungsmaskenregisters (Authority Mask Register - AMR) (potenziell maskiert)
3	Eine effektive Adresse (EA) des Verweises des Kontextspeicherungs-/- wiederherstellungsbereichs (Context Save/Restore Area Pointer - CSRP)
4	Eine Prozess-ID (PID) und eine optionale Thread-ID (TID)
5	Eine Virtuelle Adresse (VA) des Verweises des Beschleunigungsnutzungsdatensatzes (accelerator utilization record pointer - AURP)
6	Virtuelle Adresse des Speichersegmenttabellenverweises (storage segment table pointer - SSTP)
7	Eine logische Unterbrechungsdienstnummer (logical interrupt service number - LISN)

[0110] Beim Empfang eines Hypervisor-Aufrufs überprüft der Hypervisor **1296**, ob sich das Betriebssystem **1295** registriert hat und ihm die Berechtigung erteilt wurde, das Grafikbeschleunigungsmodul **1246** zu verwenden. Der Hypervisor **1296** setzt dann das Prozesselement **1283** in eine verlinkte Prozesselementliste für eine entsprechende Art des Grafikbeschleunigungsmoduls **1246**. Ein Prozesselement kann die in Tabelle 4 gezeigten Informationen beinhalten.

Tabelle 4 - Prozesselementeinformationen

1	Ein Arbeitsdeskriptor (WD)
2	Ein Wert des Berechtigungsmaskenregisters (Authority Mask Register - AMR) (potenziell maskiert).
3	Eine effektive Adresse (EA) des Verweises des Kontextspeicherungs-/- wiederherstellungsbereichs (Context Save/Restore Area Pointer - CSRP)
4	Eine Prozess-ID (PID) und eine optionale Thread-ID (TID)
5	Eine Virtuelle Adresse (VA) des Verweises des Beschleunigungsnutzungsdatensatzes (accelerator utilization record pointer -
	AURP)
6	Virtuelle Adresse des Speichersegmenttabellenverweises (storage segment table pointer - SSTP)
7	Eine logische Unterbrechungsdienstnummer (logical interrupt service number - LISN)
8	Unterbrechungsvektortabelle, abgeleitet von Hypervisor-Aufrufparametern
9	Ein Wert des Zustandsregisters (state register - SR)
10	Eine logische Partitions-ID (LPID)
11	Eine Reale Adresse (RA) des Verweises des Hypervisor-Beschleunigungsnutzungsdatensatzes
12	Speicherdeskriptorregister (SDR)

[0111] In mindestens einer Ausführungsform initialisiert der Hypervisor eine Vielzahl von Registern **1245** von Beschleunigerintegrations-Slices **1290**.

[0112] Wie in **Fig. 12F** veranschaulicht, wird in mindestens einer Ausführungsform ein einheitlicher Speicher verwendet, der über einen gemeinsamen virtuellen Speicheradressbereich ansteuerbar ist, der verwendet wird, um auf physische Prozessorspeicher **1201-1202** und GPU-Speicher **1220-1223** zuzugreifen. In dieser Umsetzung verwenden Operationen, die auf den GPUs **1210-1213** ausgeführt werden, denselben virtuellen/effektiven Speicheradressbereich, um auf die Prozessorspeicher **1201-1202** zuzugreifen und umgekehrt, wodurch die Programmierbarkeit vereinfacht wird. In einer Ausführungsform wird ein erster Teil eines virtuellen/effektiven Adressbereichs dem Prozessorspeicher **1201**, ein zweiter Teil dem zweiten Prozessorspeicher **1202**, ein dritter Teil dem GPU-Speicher **1220** usw. zugewiesen. In mindestens einer Ausführungsform wird dadurch ein gesamter virtueller/effektiver Speicherbereich (manchmal als effektiver Adressbereich bezeichnet) auf jeden der Prozessorspeicher **1201-1202** und GPU-Speicher **1220-1223** verteilt, wodurch es jedem Prozessor oder jeder GPU ermöglicht wird, auf beliebige physische Speicher mit einer virtuellen Adresse zuzugreifen, die diesem Speicher zugeordnet ist.

[0113] In einer Ausführungsform stellt die Verzerrungs-/Kohärenzverwaltungsschaltung **1294A-1294E** in einer oder mehreren der MMUs **1239A**-**1239E** die Zwischenspeicherkohärenz zwischen Zwischenspeichern von einem oder mehreren Hostprozessoren (z. B. **1205**) und den GPUs **1210-1213** sicher und setzt Verzerrungsmethoden um, die physische Speicher angeben, in denen bestimmte Arten von Daten gespeichert werden sollen. Obwohl mehrere Instanzen der Verzerrungs-/Kohärenzverwaltungsschaltungen **1294A-1294E** in **Fig. 12F** veranschaulicht sind, kann eine Verzerrungs-/Kohärenzschaltung in einer MMU eines oder mehrerer Hostprozessoren **1205** und/oder in der Beschleunigerintegrationsschaltung **1236** umgesetzt sein.

[0114] Eine Ausführungsform ermöglicht, dass der an die GPU angeschlossene Speicher **1220-1223** als Teil des Systemspeichers zugeordnet ist und auf ihn unter Verwendung der Technologie des gemeinsam genutzten virtuellen Speichers (shared virtual memory - SVM) zugegriffen wird, ohne jedoch Leistungsnachteile zu erleiden, die mit der vollständigen Systemzwischenspeicherkohärenz verbunden sind. In mindestens einer Ausführungsform stellt eine Fähigkeit für den GPU-angeschlossenen Speicher **1220-1223**, auf den Systemspeicher ohne lästigen Zwischenspeicherkohärenzaufwand zuzugreifen, eine vorteilhafte Betriebsumgebung für die GPU-Auslagerung bereit. Diese Anordnung ermöglicht es der Software des Hostprozessors **1205**, Ope-

randen einzurichten und auf Berechnungsergebnisse zuzugreifen, ohne Aufwand von herkömmlichen E/A-DMA-Datenkopien. Derartige traditionelle Kopien beziehen Treiberaufrufe, Unterbrechungen und speicherzugeordnete E/A-(memory mapped I/O - MMIO-)Zugriffe ein, die alle im Vergleich zu einfachen Speicherzugriffen ineffizient sind. In mindestens einer Ausführungsform kann eine Fähigkeit, ohne Zwischenspeicherkohärenzaufwände auf den GPUangehängten Speicher **1220-1223** zuzugreifen, für die Ausführungszeit einer ausgelagerten Berechnung entscheidend sein. In Fällen mit erheblichem Streaming-Schreibspeicherverkehr kann zum Beispiel der Zwischenspeicherkohärenzaufwand eine effektive Schreibbandbreite erheblich verringern, die von einer GPU **1210-1213** gesehen wird. In mindestens einer Ausführungsform können die Effizienz der Operandeneinrichtung, die Effizienz des Ergebniszugriffs und die Effizienz der GPU-Berechnung eine Rolle bei der Bestimmung der Effektivität einer GPU-Auslagerung spielen.

[0115] In mindestens einer Ausführungsform wird die Auswahl von GPU-Verzerrung und Hostprozessor-Verzerrung durch eine Verzerrungsortungsdatenstruktur bestimmt. Es kann zum Beispiel eine Verzerrungstabelle verwendet werden, die eine seitengranulare Struktur sein kann (d. h. mit einer Granularität einer Speicherseite gesteuert), die 1 oder 2 Bits pro GPUangeschlossener Speicherseite beinhaltet. In mindestens einer Ausführungsform kann eine Verzerrungstabelle in einem gestohlenen Speicherbereich eines oder mehrerer GPUangeschlossener Speicher **1220-1223** mit oder ohne Verzerrungszwischenspeicher in der GPU **1210-1213** umgesetzt sein (z. B. um häufig/zuletzt verwendete Einträge einer Verzerrungstabelle zwischenzuspeichern). Alternativ kann eine gesamte Verzerrungstabelle innerhalb einer GPU aufrechterhalten werden.

[0116] In mindestens einer Ausführungsform wird auf einen Verzerrungstabelleneintrag, der mit jedem Zugriff auf den GPU-angeschlossenen Speicher **1220-1223** verbunden ist, vor dem tatsächlichen Zugriff auf einen GPU-Speicher zugegriffen, was die folgenden Operationen verursacht. Zuerst werden lokale Anfragen von der GPU **1210-1213**, die ihre Seite in der GPU-Verzerrung finden, direkt an einen entsprechenden GPU-Speicher **1220-1223** weitergeleitet. Lokale Anforderungen von einer GPU, die ihre Seite in der Host-Verzerrung finden, werden an den Prozessor **1205** weitergeleitet (z. B. über eine Hochgeschwindigkeitsverbindung, vorstehend erörtert). In einer Ausführungsform vervollständigen Anforderungen vom Prozessor **1205**, die eine angeforderte Seite in der Hostprozessor-Verzerrung finden, eine Anforderung wie ein normaler Speicherlesevorgang. Alternativ können Anfragen, die an eine GPU-verzerrte Seite gerichtet sind, an die GPU **1210-1213** weitergeleitet werden. In mindestens einer Ausführungsform kann eine GPU dann eine Seite auf eine Hostprozessor-Verzerrung umstellen, wenn sie derzeit keine Seite verwendet. In mindestens einer Ausführungsform kann der Verzerrungszustand einer Seite entweder durch einen softwarebasierten Mechanismus, einen hardwareunterstützten softwarebasierten Mechanismus oder für eine begrenzte Anzahl von Fällen durch einen rein hardwarebasierten Mechanismus geändert werden.

[0117] Ein Mechanismus zum Ändern des Verzerrungszustands verwendet einen API-Aufruf (z. B. OpenCL), der wiederum den Vorrichtungstreiber einer GPU aufruft, der wiederum eine Mitteilung an eine GPU sendet (oder einen Befehlsdeskriptor einreiht), die sie anweist, eine Verzerrungszustand zu ändern und bei einigen Übergängen eine Zwischenspeicherleerungsoperation in einem Host durchzuführen. In mindestens einer Ausführungsform wird die Zwischenspeicherleerungsoperation für einen Übergang von der Verzerrung des Hostprozessors **1205** zur GPU-Verzerrung verwendet, jedoch nicht für einen entgegengesetzten Übergang.

[0118] In einer Ausführungsform wird die Zwischenspeicherkohärenz durch das vorübergehende Rendern von GPU-verzerrten Seiten aufrechterhalten, die vom Hostprozessor **1205** nicht zwischengespeichert werden können. Um auf diese Seiten zuzugreifen, kann der Prozessor **1205** Zugriff von der GPU **1210** anfordern, die den Zugriff sofort gewähren kann oder nicht. Um die Kommunikation zwischen dem Prozessor **1205** und der GPU **1210** zu verringern, ist es daher vorteilhaft sicherzustellen, dass GPU-verzerrte Seiten diejenigen sind, die von einer GPU aber nicht von dem Hostprozessor **1205** benötigt werden und umgekehrt.

[0119] Die Ableitungs- und/oder Trainingslogik **615** wird verwendet, um eine oder mehrere Ausführungsformen durchzuführen. Details bezüglich der Ableitungs- und/oder Trainingslogik **615** werden nachfolgend in Verbindung mit den **Fig. 6A** und/oder 6B bereitgestellt.

[0120] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. In mindestens einer Ausführungsform kann diese Logik mit Komponenten dieser Figuren verwendet werden, um ein dreidimensionales Bild einer bestimmten Ansicht aus zweidimensionalen Bildern aus einer willkürlichen Anzahl von Kameras zu erzeugen.

[0121] **Fig. 13** veranschaulicht beispielhafte integrierte Schaltungen und verbundene Grafikprozessoren, die unter Verwendung von einem oder mehreren IP-Kernen gemäß verschiedener in dieser Schrift beschriebener Ausführungsformen gefertigt werden können. Zusätzlich zu dem, was veranschaulicht ist, können andere Logiken und Schaltungen in mindestens einer Ausführungsform beinhaltet sein, was zusätzliche Grafikprozessoren/-kerne, Peripherieschnittstellensteuerungen oder Allzweck-Prozessorkerne beinhaltet.

[0122] **Fig. 13** ist ein Blockdiagramm, das ein beispielhaftes System auf einer integrierten Chipschaltung **1300** veranschaulicht, die unter Verwendung eines oder mehrerer IP-Kerne gemäß mindestens einer Ausführungsform gefertigt werden kann. In mindestens einer Ausführungsform beinhaltet die integrierte Schaltung **1300** einen oder mehrere Anwendungsprozessor(en) **1305** (z. B. CPUs), mindestens einen Grafikprozessor **1310** und kann zusätzlich einen Bildprozessor **1315** und/oder einen Videoprozessor **1320** beinhalten, die ein modularer IP-Kern sein kann. In mindestens einer Ausführungsform beinhaltet die integrierte Schaltung **1300** Peripherie- oder Buslogik, was eine USB-Steuerung **1325**, eine UART-Steuerung **1330**, eine SPI/SDIO-Steuerung **1335** und eine I^2S/I^2C-Steuerung **1340** beinhaltet. In mindestens einer Ausführungsform kann die integrierte Schaltung **1300** eine Anzeigevorrichtung **1345** beinhalten, die an einen oder mehrere von einer Steuerung einer Multimediaschnittstelle mit hoher Auflösung (high-definition multimedia interface - HDMI) **1350** und eine Anzeigeschnittstelle für eine mobile Industrieprozessorschnittstelle (mobile industry processor interface - MIPI) **1355** gekoppelt ist. In mindestens einer Ausführungsform kann der Speicher durch ein Flash-Speicherteilsystem **1360** bereitgestellt sein, das Flash-Speicher und eine Flash-Speichersteuerung beinhaltet. In mindestens einer Ausführungsform kann die Speicherschnittstelle über eine Speichersteuerung **1365** für den Zugriff auf SDRAM- oder SRAM-Speichervorrichtungen bereitgestellt sein. In mindestens einer Ausführungsform beinhalten einige integrierte Schaltungen zusätzlich eine eingebettete Sicherheitsengine **1370**.

[0123] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. Details bezüglich der Ableitungs- und/oder Trainingslogik **615** werden nachfolgend in Verbindung mit den **Fig. 6A** und/oder 6B bereitgestellt. In mindestens einer Ausführungsform kann eine Ableitungs- und/oder Trainingslogik **615** in der integrierten Schaltung **1300** zum Ableiten oder Vorhersagen von Operationen mindestens zum Teil auf Grundlage von Gewichtungsparametern verwendet werden, die unter Verwendung von neuronalen Netzwerktrainingsoperationen, Funktionen und/oder Architekturen von neuronalen Netzwerken oder in dieser Schrift beschriebenen Anwendungsfällen von neuronalen Netzwerken berechnet wurden.

[0124] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. In mindestens einer Ausführungsform kann diese Logik mit Komponenten dieser Figuren verwendet werden, um ein dreidimensionales Bild einer bestimmten Ansicht aus zweidimensionalen Bildern aus einer willkürlichen Anzahl von Kameras zu erzeugen.

[0125] Die **Fig. 14A-14B** veranschaulichen beispielhafte integrierte Schaltungen und verbundene Grafikprozessoren, die unter Verwendung von einem oder mehreren IP-Kernen gemäß verschiedener in dieser Schrift beschriebener Ausführungsformen gefertigt werden können. Zusätzlich zu dem, was veranschaulicht ist, können andere Logiken und Schaltungen in mindestens einer Ausführungsform beinhaltet sein, was zusätzliche Grafikprozessoren/-kerne, Peripherieschnittstellensteuerungen oder Allzweck-Prozessorkerne beinhaltet.

[0126] Die **Fig. 14A-14B** sind Blockdiagramme, die beispielhafte Grafikprozessoren zur Verwendung in einem SoC gemäß in dieser Schrift beschriebenen Ausführungsformen veranschaulichen. **Fig. 14A** veranschaulicht einen beispielhaften Grafikprozessor **1410** eines Systems auf einer integrierten Chipschaltung, die unter Verwendung eines oder mehrerer IP-Kerne gemäß mindestens einer Ausführungsform gefertigt werden kann. **Fig. 14B** veranschaulicht einen zusätzlichen beispielhaften Grafikprozessor **1440** eines Systems auf einer integrierten Chipschaltung, die unter Verwendung eines oder mehrerer IP-Kerne gemäß mindestens einer Ausführungsform gefertigt werden kann. In mindestens einer Ausführungsform ist der Grafikprozessor **1410** aus **Fig. 14A** ist ein Grafikprozessorkern mit niedriger Leistung. In mindestens einer Ausführungsform ist der Grafikprozessor **1440** aus **Fig. 14B** ein Grafikprozessorkern mit höherer Leistung. In mindestens einer Ausführungsform kann jeder der Grafikprozessoren **1410**, **1440** eine Variante des Grafikprozessors **1310** aus **Fig. 13** sein.

[0127] In mindestens einer Ausführungsform beinhaltet der Grafikprozessor **1410** einen Scheitelpunktprozessor **1405** und einen oder mehrere Fragmentprozessoren **1415A-1415N** (z. B. **1415A**, **1415B**, **1415C**, **1415D** bis **1415N-1** und **1415N**). In mindestens einer Ausführungsform kann der Grafikprozessor **1410** derartig unterschiedliche Shader-Programme über getrennte Logik ausführen, dass der Scheitelpunktprozessor **1405**

optimiert ist, um Operationen für Scheitelpunkt-Shader-Programme auszuführen, während ein oder mehrere Fragmentprozessoren **1415A-1415N** Shading-Operationen für Fragmente (z. B. Bildpunkt) für Fragment- oder Bildpunkt-Shader-Programme ausführen. In mindestens einer Ausführungsform führt der Scheitelpunktprozessor **1405** eine Scheitelpunktverarbeitungsstufe einer 3D-Grafikpipeline durch und erzeugt Primitive und Scheitelpunkt-Daten. In mindestens einer Ausführungsform verwenden ein oder mehrere Fragmentprozessoren **1415A-1415N** Primitiv- und Scheitelpunkt-Daten, die vom Scheitelpunktprozessor **1405** erzeugt werden, um einen Bildspeicher zu erzeugen, der auf einer Anzeigevorrichtung angezeigt wird. In mindestens einer Ausführungsform sind ein oder mehrere Fragmentprozessoren **1415A-1415N** optimiert, um Fragment-Shader-Programme auszuführen, wie in einer OpenGL-API bereitgestellt, die verwendet werden können, um ähnliche Operationen wie ein Bildpunkt-Shader-Programm durchzuführen, wie es in einer Direct 3D-API vorgesehen ist.

[0128] In mindestens einer Ausführungsform beinhaltet der Grafikprozessor **1410** zusätzlich eine oder mehrere Speicherverwaltungseinheiten (memory management units - MMUs) **1420A-1420B**, Zwischenspeicher **1425A-1425B** und Schaltungsverbindung(en) **1430A-1430B**. In mindestens einer Ausführungsform stellen eine oder mehrere MMU(s) **1420A-1420B** eine virtuell-zu-physische Adresszuordnung für den Grafikprozessor **1410** bereit, was für den Scheitelpunktprozessor **1405** und/oder Fragmentprozessor(en) **1415A-1415N** beinhaltet, die sich auf Scheitelpunkt- oder im Speicher gespeicherte Bild-/Texturdaten zusätzlich zu in einem oder mehreren Zwischenspeichern **1425A-1425B** gespeicherten Scheitelpunkt- oder Bild-/Texturdaten beziehen. In mindestens einer Ausführungsform können eine oder mehrere MMU(s) **1420A-1420B** mit anderen MMUs innerhalb des Systems synchronisiert werden, was eine oder mehrere MMUs beinhaltet, die derartig mit einem oder mehreren Anwendungsprozessoren **1305**, Bildprozessoren **1315** und/oder Videoprozessoren **1320** aus **Fig. 13** verbunden sind, dass jeder Prozessor **1305-1320** an einem gemeinsam genutzten oder einheitlichen virtuellen Speichersystem teilnehmen kann. In mindestens einer Ausführungsform ermöglichen eine oder mehrere Schaltungsverbindungen **1430A-1430B** dem Grafikprozessor **1410**, mit anderen IP-Kernen innerhalb des SoC Schnittstellen zu bilden, entweder über einen internen Bus des SoC oder über eine direkte Verbindung.

[0129] In mindestens einer Ausführungsform beinhaltet der Grafikprozessor **1440** eine oder mehrere MMU(s) **1420A-1420B**, Zwischenspeicher **1425A-1425B** und Schaltungsverbindung(en) **1430A-1430B** des Grafikprozessors **1410** aus **Fig. 14A**. In mindestens einer Ausführungsform beinhaltet der Grafikprozessor **1440** einen oder mehrere Shader-Kerne **1455A-1455N** (z. B. **1455A**, **1455B**, **1455C**, **1455D**, **1455E**, **1455F** bis **1455N-1** und **1455N**), was eine einheitliche Shader-Kernarchitektur bereitstellt, in der ein(e) einzelne(r) Kern oder Art oder Kern alle Arten von programmierbarem Shader-Code ausführen kann, was Shader-Programmcode beinhaltet, um Scheitelpunkt-Shader, Fragment-Shader und/oder Compute-Shader umzusetzen. In mindestens einer Ausführungsform kann eine Anzahl von Shader-Kernen variieren. In mindestens einer Ausführungsform beinhaltet der Grafikprozessor **1440** eine Aufgabenverwalter innerhalb des Kerns **1445**, die als Thread-Verteiler fungiert, um Ausführungsthreads an einen oder mehrere Shader-Kerne **1455A-1455N** zu verteilen, und eine Kachelungseinheit **1458**, um Kachelungsoperationen für kachelbasiertes Rendering zu beschleunigen, bei dem Rendering-Operationen für eine Szene im Bildraum unterteilt werden, um zum Beispiel lokale räumliche Kohärenz innerhalb einer Szene auszunutzen oder um die Nutzung interner Zwischenspeicher zu optimieren.

[0130] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. Details bezüglich der Ableitungs- und/oder Trainingslogik **615** werden nachfolgend in Verbindung mit den **Fig. 6A** und/oder 6B bereitgestellt. In mindestens einer Ausführungsform kann eine Ableitungs- und/oder Trainingslogik **615** in der integrierten Schaltung **14A** und/oder **14B** zum Ableiten oder Vorhersagen von Operationen mindestens zum Teil auf Grundlage von Gewichtungsparametern verwendet werden, die unter Verwendung von neuronalen Netzwerktrainingsoperationen, Funktionen und/oder Architekturen von neuronalen Netzwerken oder in dieser Schrift beschriebenen Anwendungsfällen von neuronalen Netzwerken berechnet wurden. Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. In mindestens einer Ausführungsform kann diese Logik mit Komponenten dieser Figuren verwendet werden, um ein dreidimensionales Bild einer bestimmten Ansicht aus zweidimensionalen Bildern aus einer willkürlichen Anzahl von Kameras zu erzeugen.

[0131] Die **Fig. 15A-15B** veranschaulichen zusätzliche beispielhafte Grafikprozessorlogik gemäß in dieser Schrift beschriebenen Ausführungsformen. **Fig. 15A** veranschaulicht einen Grafikkern **1500**, der in dem Grafikprozessor **1310** aus **Fig. 13** in mindestens einer Ausführungsform beinhaltet sein kann, und kann ein einheitlicher Shader-Kern **1455A-1455N** wie in **Fig. 14B** in mindestens einer Ausführungsform sein. **Fig. 15B** veranschaulicht eine hochparallele Allzweck-Grafikverarbeitungseinheit **1530**, die in mindestens einer Ausführungsform zum Einsatz auf einem Mehrchipmodul geeignet ist.

[0132] In mindestens einer Ausführungsform beinhaltet der Grafikkern **1500** einen gemeinsam genutzten Anweisungszwischenspeicher **1502**, eine Textureinheit **1518** und einen Zwischenspeicher/gemeinsam genutzten Speicher **1520**, die Ausführungsressourcen innerhalb des Grafikkerns **1500** gemeinsam sind. In mindestens einer Ausführungsform kann der Grafikkern **1500** mehrere Slices **1501A-1501N** oder eine Partition für jeden Kern beinhalten, und ein Grafikprozessor kann mehrere Instanzen des Grafikkerns **1500** beinhalten. Die Slices **1501A-1501N** können eine Unterstützungslogik beinhalten, die einen lokalen Anweisungszwischenspeicher **1504A-1504N**, einen Thread-Planer **1506A-1506N**, einen Thread-Verteiler **1508A-1508N** und einen Satz von Registern **1510A-1510N** beinhaltet. In mindestens einer Ausführungsform können die Slices **1501A-1501N** einen Satz zusätzlicher Funktionseinheiten (AFUs (additional function units) **1512A-1512N**), Gleitkommaeinheiten (FPU (floating-point units) **1514A-1514N**), ganzzahlige arithmetische Logikeinheiten (ALUs (arithmetic logic units) **1516-1516N**), Adressberechnungseinheiten (ACU (address computational unit) **1513A-1513N**), Gleitkommaeinheiten mit doppelter Genauigkeit (DPFPU (double-precision floating-point unit) **1515A-1515N**) und Matrixverarbeitungseinheiten (MPU (matrix processing unit) **1517A-1517N**) beinhalten.

[0133] In mindestens einer Ausführungsform können die FPUs **1514A-1514N** Gleitkommaoperationen mit einfacher Genauigkeit (32 Bit) und mit halber Genauigkeit (16 Bit) ausführen, während die DPFPUs **1515A-1515N** Gleitkommaoperationen mit doppelter Genauigkeit (64 Bit) ausführen. In mindestens einer Ausführungsform können die ALUs **1516A-1516N** ganzzahlige Operationen mit variabler Präzision mit einer Genauigkeit von 8-Bit, 16-Bit und 32-Bit ausführen und können für Operationen mit gemischter Präzision konfiguriert sein. In mindestens einer Ausführungsform können die MPUs **1517A-1517N** auch für Matrixoperationen mit gemischter Genauigkeit konfiguriert sein, die Gleitkomma- und 8-Bit-Ganzzahloperationen mit halber Genauigkeit beinhalten. In mindestens einer Ausführungsform können die MPUs **1517A-1517N** eine Reihe von Matrixoperationen ausführen, um Anwendungsframeworks des maschinellen Lernens zu beschleunigen, welche die Ermöglichung der Unterstützung für beschleunigte allgemeine Matrix-zu-Matrix-Multiplikation (general matrix to matrix multiplication - GEMM) beinhalten. In mindestens einer Ausführungsform können die AFUs **1512A-1512N** zusätzliche logische Operationen ausführen, die nicht von Gleitkomma- oder Ganzzahleinheiten unterstützt werden, die trigonometrische Operationen (z. B. Sinus, Cosinus usw.) beinhalten.

[0134] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. Details bezüglich der Ableitungs- und/oder Trainingslogik **615** werden nachfolgend in Verbindung mit den **Fig. 6A** und/oder 6B bereitgestellt. In mindestens einer Ausführungsform kann eine Ableitungs- und/oder Trainingslogik **615** in dem Grafikkern **1500** zum Ableiten oder Vorhersagen von Operationen mindestens zum Teil auf Grundlage von Gewichtungsparametern verwendet werden, die unter Verwendung von neuronalen Netzwerktrainingsoperationen, Funktionen und/oder Architekturen von neuronalen Netzwerken oder in dieser Schrift beschriebenen Anwendungsfällen von neuronalen Netzwerken berechnet wurden.

[0135] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. In mindestens einer Ausführungsform kann diese Logik mit Komponenten dieser Figuren verwendet werden, um ein dreidimensionales Bild einer bestimmten Ansicht aus zweidimensionalen Bildern aus einer willkürlichen Anzahl von Kameras zu erzeugen.

[0136] **Fig. 15B** veranschaulicht eine Universalverarbeitungseinheit (general-purpose processing unit - GPGPU) **1530**, die konfiguriert werden kann, um hochparallele Rechenoperationen zu ermöglichen, die von einem Array von Grafikverarbeitungseinheiten ausgeführt werden sollen, in mindestens einer Ausführungsform. In mindestens einer Ausführungsform kann die GPGPU **1530** direkt mit anderen Instanzen der GPGPU **1530** verbunden sein, um einen Multi-GPU-Cluster zu erzeugen, um die Trainingsgeschwindigkeit für tiefe neuronale Netzwerke zu verbessern. In mindestens einer Ausführungsform beinhaltet die GPGPU **1530** eine Hostschnittstelle **1532**, um eine Verbindung mit einem Hostprozessor zu ermöglichen. In mindestens einer Ausführungsform ist die Hostschnittstelle **1532** eine PCI-Express-Schnittstelle. In mindestens einer Ausführungsform kann die Hostschnittstelle **1532** eine herstellerspezifische Kommunikationsschnittstelle oder eine Kommunikationsstruktur sein. In mindestens einer Ausführungsform empfängt die GPGPU **1530** Befehle von einem Hostprozessor und verwendet einen globalen Planer **1534**, um Ausführungsthreads, die diesen Befehlen zugeordnet sind, an einen Satz von Rechenclustern **1536A-1536H** zu verteilen. In mindestens einer Ausführungsform teilen sich die Rechencluster **1536A-1536H** einen schnellen Pufferspeicher **1538**. In mindestens einer Ausführungsform kann der schnelle Pufferspeicher **1538** als ein Zwischenspeicher höherer Ebene für schnelle Pufferspeicher innerhalb der Rechencluster **1536A-1536H** dienen.

[0137] In mindestens einer Ausführungsform beinhaltet die GPGPU **1530** einen Speicher **1544A-1544B**, der über einen Satz von Speichersteuerungen **1542A-1542B** mit Rechenclustern **1536A-1536H** gekoppelt ist. In mindestens einer Ausführungsform kann der Speicher **1544A-1544B** verschiedene Arten von Speichervorrichtungen umfassen, die dynamischen Direktzugriffsspeicher (DRAM) oder Grafik-Direktzugriffsspeicher, wie etwa synchronen Grafik-Direktzugriffsspeicher (SGRAM) beinhalten, der Grafik-Doppeldatenraten-(GDDR-) Speicher beinhaltet.

[0138] In mindestens einer Ausführungsform beinhalten die Rechencluster **1536A-1536H** jeweils einen Satz von Grafikkernen, wie etwa den Grafikkern **1500** aus **Fig. 15A**, die mehrere Arten von Ganzzahl- und Gleitkomma-Logikeinheiten beinhalten kann, die Rechenoperationen mit einer Reihe von Genauigkeiten ausführen können, die für maschinelle Lemberechnungen geeignete beinhalten. Zum Beispiel kann in mindestens einer Ausführungsform mindestens eine Teilmenge von Gleitkommaeinheiten in jedem der Rechencluster **1536A-1536H** dazu konfiguriert sein, 16-Bit- oder 32-Bit-Gleitkommaoperationen auszuführen, während eine andere Teilmenge von Gleitkommaeinheiten dazu konfiguriert sein können, 64-Bit-Gleitkommaoperationen auszuführen.

[0139] In mindestens einer Ausführungsform können mehrere Instanzen der GPGPU **1530** dazu konfiguriert sein, als ein Rechencluster betrieben zu werden. In mindestens einer Ausführungsform variiert die Kommunikation, die von den Rechenclustern **1536A-1536H** für die Synchronisation und den Datenaustausch verwendet wird, zwischen den Ausführungsformen. In mindestens einer Ausführungsform kommunizieren mehrere Instanzen der GPGPU **1530** über die Hostschnittstelle **1532**. In mindestens einer Ausführungsform beinhaltet die GPGPU **1530** einen E/A-Hub **1539**, der die GPGPU **1530** mit einer GPU-Verbindung **1540** koppelt, der eine direkte Verbindung zu anderen Instanzen der GPGPU **1530** ermöglicht. In mindestens einer Ausführungsform ist der GPU-Link **1540** an eine dedizierte GPU-zu-GPU-Brücke gekoppelt, welche die Kommunikation und Synchronisation zwischen mehreren Instanzen der GPGPU **1530** ermöglicht. In mindestens einer Ausführungsform ist die GPU-Verbindung **1540** mit einer Hochgeschwindigkeitsverbindung gekoppelt, um Daten an andere GPGPUs oder Parallelprozessoren zu übermitteln und zu empfangen. In mindestens einer Ausführungsform befinden sich mehrere Instanzen der GPGPU **1530** in getrennten Datenverarbeitungssystemen und kommunizieren über eine Netzwerkvorrichtung, auf die über die Hostschnittstelle **1532** zugegriffen werden kann. In mindestens einer Ausführungsform der GPU kann die Verbindung **1540** dazu konfiguriert sein, eine Verbindung zu einem Hostprozessor zusätzlich zu oder als Alternative zu der Hostschnittstelle **1532** zu ermöglichen.

[0140] In mindestens einer Ausführungsform kann die GPGPU **1530** dazu konfiguriert sein, neuronale Netzwerke zu trainieren. In mindestens einer Ausführungsform kann die GPGPU **1530** innerhalb einer Ableitungsplattform verwendet werden. In mindestens einer Ausführungsform, in der die GPGPU **1530** für die Ableitung verwendet wird, kann die GPGPU weniger Rechencluster **1536A-1536H** beinhalten, als wenn die GPGPU zum Trainieren eines neuronalen Netzwerks verwendet wird. In mindestens einer Ausführungsform kann sich die dem Speicher **1544A-1544B** zugeordnete Speichertechnologie zwischen Ableitung- und Trainingskonfigurationen unterscheiden, wobei Speichertechnologien mit höherer Bandbreite Trainingskonfigurationen gewidmet sind. In mindestens einer Ausführungsform kann die Ableitungskonfiguration der GPGPU **1530** die Ableitung von spezifischen Anweisungen unterstützen. Zum Beispiel kann in mindestens einer Ausführungsform eine Ableitungskonfiguration Unterstützung für eine oder mehrere 8-Bit-Ganzahlpunktproduktanweisungen bereitstellen, die während den Ableitungsoperationen für eingesetzte neuronale Netzwerke verwendet werden können.

[0141] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. Details bezüglich der Ableitungs- und/oder Trainingslogik **615** werden nachfolgend in Verbindung mit den **Fig. 6A** und/oder 6B bereitgestellt. In mindestens einer Ausführungsform kann eine Ableitungs- und/oder Trainingslogik **615** in der GPGPU **1530** zum Ableiten oder Vorhersagen von Operationen mindestens zum Teil auf Grundlage von Gewichtungsparametern verwendet werden, die unter Verwendung von neuronalen Netzwerktrainingsoperationen, Funktionen und/oder Architekturen von neuronalen Netzwerken oder in dieser Schrift beschriebenen Anwendungsfällen von neuronalen Netzwerken berechnet wurden.

[0142] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. In mindestens einer Ausführungsform kann diese Logik mit Komponenten dieser Figuren verwendet werden, um ein dreidimensionales Bild einer bestimmten Ansicht aus zweidimensionalen Bildern aus einer willkürlichen Anzahl von Kameras zu erzeugen.

[0143] **Fig. 16** ist ein Blockdiagramm, das ein Computersystem **1600** gemäß mindestens einer Ausführungsform veranschaulicht. In mindestens einer Ausführungsform beinhaltet das Computersystem **1600** ein Verar-

beitungsteilsystem **1601** mit einem oder mehreren Prozessoren **1602** und einem Systemspeicher **1604**, der über einen Verbindungspfad kommuniziert, der einen Speicher-Hub **1605** beinhalten kann. In mindestens einer Ausführungsform kann der Speicher-Hub **1605** eine getrennte Komponente innerhalb einer Chipsatzkomponente sein oder kann in einen oder mehrere Prozessoren **1602** integriert sein. In mindestens einer Ausführungsform ist der Speicher-Hub **1605** über eine Kommunikationsverbindung **1606** mit einem E/A-Teilsystem **1611** gekoppelt. In mindestens einer Ausführungsform beinhaltet das E/A-Teilsystem **1611** einen E/A-Hub **1607**, der es dem Computersystem **1600** ermöglichen kann, Eingaben von einem oder mehreren Eingabevorrichtungen **1608** zu empfangen. In mindestens einer Ausführungsform kann der E/A-Hub **1607** einer Anzeigesteuerung, die in einem oder mehreren Prozessoren **1602** beinhaltet sein kann, ermöglichen, einer oder mehreren Anzeigevorrichtungen **1610A** Ausgaben bereitzustellen. In mindestens einer Ausführungsform können ein oder mehrere Anzeigevorrichtungen **1610A**, die mit dem E/A-Hub **1607** gekoppelt sind, eine lokale, interne oder eingebettete Anzeigevorrichtung beinhalten.

[0144] In mindestens einer Ausführungsform beinhaltet das Verarbeitungsteilsystem **1601** einen oder mehrere Parallelprozessoren **1612**, die über einen Bus oder eine andere Kommunikationsverbindung **1613** mit dem Speicher-Hub **1605** gekoppelt sind. In mindestens einer Ausführungsform kann die Kommunikationsverbindung **1613** eine von einer beliebigen Anzahl von auf Standards basierenden Kommunikationsverbindungstechnologien oder - protokollen sein, wie etwa, aber nicht beschränkt auf PCI Express, oder kann eine herstellerspezifische Kommunikationsschnittstelle oder Kommunikationsstruktur sein. In mindestens einer Ausführungsform bilden ein oder mehrere Parallelprozessoren **1612** ein rechenfokussiertes Parallel- oder Vektorverarbeitungssystem, das eine große Anzahl von Verarbeitungskernen und/oder Verarbeitungsclustern beinhalten kann, wie etwa einen Prozessor mit vielen integrierten Kernen (many integrated core - MIC). In mindestens einer Ausführungsform bilden ein oder mehrere Parallelprozessoren **1612** ein Grafikverarbeitungsteilsystem, das Bildpunkt an eine oder mehrere Anzeigevorrichtungen **1610A** ausgeben kann, die über den E/A-Hub **1607** gekoppelt sind. In mindestens einer Ausführungsform können ein oder mehrere Parallelprozessoren **1612** auch einen Anzeigesteuerung und eine Anzeigeschnittstelle (nicht gezeigt) beinhalten, um eine direkte Verbindung zu einer oder mehreren Anzeigevorrichtungen **1610B** zu ermöglichen.

[0145] In mindestens einer Ausführungsform kann sich eine Systemspeichereinheit **1614** mit dem E/A-Hub **1607** verbinden, um einen Speichermechanismus für das Computersystem **1600** bereitzustellen. In mindestens einer Ausführungsform kann ein E/A-Switch **1616** verwendet werden, um einen Schnittstellenmechanismus bereitzustellen, um Verbindungen zwischen dem E/A-Hub **1607** und anderen Komponenten zu ermöglichen, wie etwa einem Netzwerkadapter **1618** und/oder einem drahtlosen Netzwerkadapter **1619**, die in eine oder mehrere Plattformen integriert sein können, und verschiedenen anderen Vorrichtungen, die über eine oder mehrere Erweiterungsvorrichtungen **1620** hinzugefügt werden können. In mindestens einer Ausführungsform kann der Netzwerkadapter **1618** ein Ethemet-Adapter oder ein anderer verdrahteter Netzwerkadapter sein. In mindestens einer Ausführungsform kann der drahtlose Netzwerkadapter **1619** eine oder mehrere von einer Wi-Fi-, Bluetooth-, Nahfeldkommunikations-(near field communication - NFC-) oder einer anderen Netzwerkvorrichtung beinhalten, die ein oder mehrere drahtlose Funkvorrichtungen beinhaltet.

[0146] In mindestens einer Ausführungsform kann das Computersystem **1600** andere Komponenten beinhalten, die nicht ausdrücklich dargestellt sind, die USB oder andere Port-Verbindungen, optische Speicherlaufwerke, Videoaufnahmevorrichtungen und dergleichen beinhalten, und auch mit dem E/A-Hub **1607** verbunden sein können. In mindestens einer Ausführungsform können Kommunikationswege, die verschiedene Komponenten in **Fig. 16** verbinden, unter Verwendung beliebiger geeigneter Protokolle umgesetzt sein, wie etwa auf PCI (Peripheral Component Interconnect) basierte Protokolle (z. B. PCI-Express) oder andere Bus- oder Punkt-zu-Punkt-Kommunikationsschnittstellen und/oder -protokolle, wie etwa NV-Link-Hochgeschwindigkeitsverbindung oder Verbindungsprotokolle.

[0147] In mindestens einer Ausführungsform schließen ein oder mehrere Parallelprozessoren **1612** eine Schaltung ein, die für die Grafik- und Videoverarbeitung optimiert ist, was zum Beispiel eine Videoausgabeschaltung beinhaltet, und bilden eine Grafikverarbeitungseinheit (GPU). In mindestens einer Ausführungsform schließen ein oder mehrere Parallelprozessoren **1612** eine Schaltung ein, die für eine universelle Verarbeitung optimiert ist. In mindestens einer Ausführungsform können Komponenten des Computersystems **1600** mit einem oder mehreren anderen Systemelementen auf einer einzigen integrierten Schaltung integriert sein. Zum Beispiel können in mindestens einer Ausführungsform ein oder mehrere Parallelprozessoren **1612**, Speicher-Hub **1605**, Prozessoren **1602** und E/A-Hub **1607** in eine integrierte System-on-Chip-(SoC)-Schaltung integriert sein. In mindestens einer Ausführungsform können Komponenten des Computersystems **1600** in ein einzelnes Paket integriert werden, um eine System-in-Gehäuse-(system in package - SIP-)Konfiguration zu bilden. In mindestens einer Ausführungsform kann mindestens ein Teil der Komponenten des Computersystems **1600**

in ein Mehrchipmodul (MCM) integriert werden, das mit anderen Mehrchipmodulen zu einem modularen Computersystem verbunden werden kann.

[0148] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. Details bezüglich der Ableitungs- und/oder Trainingslogik **615** werden nachfolgend in Verbindung mit den **Fig. 6A** und/oder 6B bereitgestellt. In mindestens einer Ausführungsform kann eine Ableitungs- und/oder Trainingslogik **615** im System der FIG. 1600 zum Ableiten oder Vorhersagen von Operationen mindestens zum Teil auf Grundlage von Gewichtungsparametern verwendet werden, die unter Verwendung von neuronalen Netzwerktrainingsoperationen, Funktionen und/oder Architekturen von neuronalen Netzwerken oder in dieser Schrift beschriebenen Anwendungsfällen von neuronalen Netzwerken berechnet wurden.

[0149] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. In mindestens einer Ausführungsform kann diese Logik mit Komponenten dieser Figuren verwendet werden, um ein dreidimensionales Bild einer bestimmten Ansicht aus zweidimensionalen Bildern aus einer willkürlichen Anzahl von Kameras zu erzeugen.

PROZESSOREN

[0150] **Fig. 17A** veranschaulicht einen Parallelprozessor **1700** gemäß mindestens einer Ausführungsform; In mindestens einer Ausführungsform können verschiedene Komponenten des Parallelprozessors **1700** unter Verwendung einer oder mehrerer integrierten Schaltungsvorrichtungen umgesetzt sein, wie etwa programmierbare Prozessoren, anwendungsspezifische integrierte Schaltungen (ASICs) oder feldprogrammierbare Gatearrays (FPGA). In mindestens einer Ausführungsform ist der veranschaulichte Parallelprozessor **1700** eine Variante eines oder mehrerer Parallelprozessoren **1612**, die in **Fig. 16** gezeigt sind, gemäß einer beispielhaften Ausführungsform.

[0151] In mindestens einer Ausführungsform beinhaltet der Parallelprozessor **1700** eine Parallelverarbeitungseinheit **1702**. In mindestens einer Ausführungsform beinhaltet die Parallelverarbeitungseinheit **1702** eine E/A-Einheit **1704**, welche die Kommunikation mit anderen Vorrichtungen ermöglicht, was andere Instanzen der Parallelverarbeitungseinheit **1702** beinhaltet. In mindestens einer Ausführungsform kann die E/A-Einheit **1704** direkt mit anderen Vorrichtungen verbunden sein. In mindestens einer Ausführungsform verbindet sich die E/A-Einheit **1704** mit anderen Vorrichtungen über die Verwendung einer Hub- oder Switch-Schnittstelle, wie etwa dem Speicher-Hub **1605**. In mindestens einer Ausführungsform bilden Verbindungen zwischen dem Speicher-Hub **1605** und der E/A-Einheit **1704** eine Kommunikationsverbindung **1613**. In mindestens einer Ausführungsform ist die E/A-Einheit **1704** mit einer Hostschnittstelle **1706** und einer Speicher-Crossbar **1716** verbunden, wobei die Hostschnittstelle **1706** Befehle empfängt, die auf die Ausführung von Verarbeitungsoperationen gerichtet sind, und die Speicher-Crossbar **1716** Befehle empfängt, die auf die Ausführung von Speicheroperationen gerichtet sind.

[0152] In mindestens einer Ausführungsform kann die Hostschnittstelle **1706**, wenn die Hostschnittstelle **1706** einen Befehlspuffer über die E/A-Einheit **1704** empfängt, Arbeitsoperationen anweisen, um diese Befehle an einem Front-End **1708** auszuführen. In mindestens einer Ausführungsform ist das Front-End **1708** mit einem Planer **1710** gekoppelt, der dazu konfiguriert ist, Befehle oder andere Arbeitselemente an ein Verarbeitungsclusterarray **1712** zu verteilen. In mindestens einer Ausführungsform stellt der Planer **1710** sicher, dass das Verarbeitungsclusterarray **1712** richtig konfiguriert ist und sich in einem gültigen Zustand befindet, bevor Aufgaben an das Verarbeitungsclusterarray **1712** verteilt werden. In mindestens einer Ausführungsform wird der Planer **1710** über eine Firmware-Logik umgesetzt, die auf einer Mikrosteuerung ausgeführt wird. In mindestens einer Ausführungsform ist der durch eine Mikrosteuerung umgesetzte Planer **1710** konfigurierbar, um komplexe Planungs- und Arbeitsverteilungsoperationen mit grober und feiner Granularität durchzuführen, was eine schnelle Vorbelegung und Kontextumschaltung von Threads ermöglicht, die auf dem Verarbeitungsarray **1712** ausgeführt werden. In mindestens einer Ausführungsform kann die Host-Software Arbeitslasten zum Planen auf dem Verarbeitungsarray **1712** über eine von mehreren Doorbells für die Grafikverarbeitung nachweisen. In mindestens einer Ausführungsform können Arbeitslasten dann durch die Logik des Planers **1710** innerhalb einer Mikrosteuerung, die den Planer **1710** beinhaltet, automatisch über das Verarbeitungsarray **1712** verteilt werden.

[0153] In mindestens einer Ausführungsform kann das Verarbeitungsclusterarray **1712** bis zu „N“ Verarbeitungscluster beinhalten (z. B. Cluster **1714A**, Cluster **1714B** bis Cluster **1714N**). In mindestens einer

Ausführungsform kann jeder Cluster **1714A-1714N** des Verarbeitungsclusterarrays **1712** eine große Anzahl gleichzeitiger Threads ausführen. In mindestens einer Ausführungsform kann der Planer **1710** den Clustern **1714A-1714N** des Verarbeitungsclusterarrays **1712** unter Verwendung verschiedener Planungs- und/oder Arbeitsverteilungsalgorithmen Arbeit zuweisen, die abhängig von der für jede Art von Programm oder Berechnung entstehenden Arbeitslasten variieren können. In mindestens einer Ausführungsform kann die Planung dynamisch durch den Planer **1710** gehandhabt werden oder kann teilweise durch die Compilerlogik während der Kompilierung der Programmlogik unterstützt werden, die für die Ausführung durch das Verarbeitungsclusterarray **1712** konfiguriert ist. In mindestens einer Ausführungsform können unterschiedliche Cluster **1714A-1714N** des Verarbeitungsclusterarrays **1712** zum Verarbeiten unterschiedlicher Arten von Programmen oder zum Durchführen unterschiedlicher Arten von Berechnungen zugewiesen werden.

[0154] In mindestens einer Ausführungsform kann das Verarbeitungsclusterarray **1712** dazu konfiguriert sein, verschiedene Arten von Parallelverarbeitungsoperationen durchzuführen. In mindestens einer Ausführungsform ist das Verarbeitungsclusterarray **1712** dazu konfiguriert, universelle Parallelrechenoperationen durchzuführen. Zum Beispiel kann das Verarbeitungsclusterarray **1712** in mindestens einer Ausführungsform Logik zum Ausführen von Verarbeitungsaufgaben beinhalten, was das Filtern von Video- und/oder Audiodaten, das Durchführen von Modellierungsoperationen, was Physikoperationen beinhaltet, und das Durchführen von Datentransformationen beinhaltet.

[0155] In mindestens einer Ausführungsform ist das Verarbeitungsclusterarray **1712** dazu konfiguriert, parallele Grafikverarbeitungsoperationen durchzuführen. In mindestens einer Ausführungsform kann das Verarbeitungsclusterarray **1712** zusätzliche Logik beinhalten, um die Ausführung derartiger Grafikverarbeitungsoperationen zu unterstützen, was Texturabtastlogik, um Texturoperationen durchzuführen, sowie Tessellationslogik und andere Scheitelpunktverarbeitungslogik beinhaltet, ohne darauf beschränkt zu sein. In mindestens einer Ausführungsform kann das Verarbeitungsclusterarray **1712** dazu konfiguriert sein, grafikverarbeitungsbezogene Shader-Programme auszuführen, wie etwa aber jedoch nicht beschränkt auf Scheitelpunkt-Shader, Tessellation-Shader, Geometrie-Shader und Bildpunkt-Shader. In mindestens einer Ausführungsform kann die Parallelverarbeitungseinheit **1702** Daten vom Systemspeicher zur Verarbeitung über die E/A-Einheit **1704** übertragen. In mindestens einer Ausführungsform können während der Verarbeitung übertragene Daten während der Verarbeitung im On-Chip-Speicher (z. B. Parallelprozessorspeicher **1722**) gespeichert und dann in den Systemspeicher zurückgeschrieben werden.

[0156] In mindestens einer Ausführungsform kann, wenn die Parallelverarbeitungseinheit **1702** verwendet wird, um eine Grafikverarbeitung durchzuführen, der Planer **1710** dazu konfiguriert sein, eine Verarbeitungsarbeitslast in ungefähr gleich große Aufgaben aufzuteilen, um bessere eine Verteilung von Grafikverarbeitungsoperationen auf mehrere Cluster **1714A-1714N** des Verarbeitungsclusterarrays **1712** zu ermöglichen. In mindestens einer Ausführungsform können Teile des Verarbeitungsclusterarrays **1712** dazu konfiguriert sein, unterschiedliche Verarbeitungsarten durchzuführen. Zum Beispiel kann in mindestens einer Ausführungsform ein erster Teil dazu konfiguriert sein, Scheitelpunkt-Shading und Topologieerzeugung durchzuführen, ein zweiter Teil kann dazu konfiguriert sein, Tessellation und Geometrie-Shading durchzuführen, und ein dritter Teil dazu kann konfiguriert sein, Bildpunkt-Shading oder andere Bildschirmraumoperationen durchzuführen, um ein gerendertes Bild zur Anzeige zu erzeugen. In mindestens einer Ausführungsform können Zwischendaten, die von einem oder mehreren der Cluster **1714A-1714N** erzeugt werden, in Puffern gespeichert werden, um zu ermöglichen, dass Zwischendaten zwischen den Clustern **1714A-1714N** zur weiteren Verarbeitung übermittelt werden.

[0157] In mindestens einer Ausführungsform kann das Verarbeitungsclusterarray **1712** über den Planer **1710** auszuführende Verarbeitungsaufgaben empfangen, der Verarbeitungsaufgaben definierende Befehle vom Front-End **1708** empfängt. In mindestens einer Ausführungsform können Verarbeitungsaufgaben Indizes von zu verarbeitenden Daten beinhalten, z.B. Oberflächen-(Patch-)Daten, Primitivdaten, Scheitelpunkt-Daten und/oder Bildpunktdaten sowie Zustandsparameter und Befehle, die definieren, wie Daten verarbeitet werden sollen (z. B. welches Programm ausgeführt werden soll). In mindestens einer Ausführungsform kann der Planer **1710** dazu konfiguriert sein, Indizes abzurufen, die Aufgaben entsprechen, oder kann Indizes vom Front-End **1708** empfangen. In mindestens einer Ausführungsform kann das Front-End **1708** dazu konfiguriert sein, sicherzustellen, dass das Verarbeitungsclusterarray **1712** in einen gültigen Zustand konfiguriert ist, bevor eine durch eingehende Befehlspuffer (z. B. Stapelpuffer, Push-Puffer usw.) spezifizierte Arbeitslast initiiert wird.

[0158] In mindestens einer Ausführungsform kann jede von einer oder mehreren Instanzen der Parallelverarbeitungseinheit **1702** mit dem Parallelprozessorspeicher **1722** gekoppelt sein. In mindestens einer Ausführungsform kann auf den Parallelprozessorspeicher **1722** über die Speicher-Crossbar **1716** zugegriffen werden,

die Speicheranforderungen von dem Verarbeitungsclusterarray **1712** sowie der E/A-Einheit **1704** empfangen kann. In mindestens einer Ausführungsform kann die Speicher-Crossbar **1716** über eine Speicherschnittstelle **1718** auf den Parallelprozessorspeicher **1722** zugreifen. In mindestens einer Ausführungsform kann die Speicherschnittstelle **1718** mehrere Partitionseinheiten (z. B. Partitionseinheit **1720A**, Partitionseinheit **1720B** bis Partitionseinheit **1720N**) beinhalten, die jeweils an einen Teil (z. B. Speichereinheit) des Parallelprozessorspeichers **1722** koppeln können. In mindestens einer Ausführungsform ist eine Anzahl von Partitionseinheiten **1720A-1720N** dazu konfiguriert, derartig gleich einer Anzahl von Speichereinheiten zu sein, dass eine erste Partitionseinheit **1720A** eine entsprechende erste Speichereinheit **1724A** aufweist, eine zweite Partitionseinheit **1720B** eine entsprechende Speichereinheit **1724B** aufweist und eine N-te Partitionseinheit **1720N** eine entsprechende N-te Speichereinheit **1724N** aufweist. In mindestens einer Ausführungsform ist eine Anzahl von Partitionseinheiten **1720A-1720N** möglicherweise nicht gleich einer Anzahl von Speichervorrichtungen.

[0159] In mindestens einer Ausführungsform können die Speichereinheiten **1724A-1724N** verschiedene Arten von Speichervorrichtungen umfassen, die dynamischen Direktzugriffsspeicher (DRAM) oder Grafik-Direktzugriffsspeicher, wie etwa synchronen Grafik-Direktzugriffsspeicher (SGRAM) beinhalten, der Grafik-Doppeldatenraten-(GDDR-)Speicher beinhaltet. In mindestens einer Ausführungsform können die Speichereinheiten **1724A-1724N** auch 3D-Stapelspeicher beinhalten, was Speicher mit hoher Bandbreite (high bandwidth memory - HBM) beinhaltet, ohne darauf beschränkt zu sein. In mindestens einer Ausführungsform können Renderziele, wie etwa Bildspeicher oder Texturkarten, über die Speichereinheiten **1724A-1724N** gespeichert werden, was es den Partitionseinheiten **1720A-1720N** ermöglicht, Teile jedes Renderziels parallel zu schreiben, um die verfügbare Bandbreite des Parallelprozessorspeichers **1722** effizient zu nutzen. In mindestens einer Ausführungsform kann eine lokale Instanz des Parallelprozessorspeichers **1722** zugunsten eines einheitlichen Speicherdesigns ausgeschlossen werden, das Systemspeicher in Verbindung mit lokalem schnellem Pufferspeicher nutzt.

[0160] In mindestens einer Ausführungsform kann jeder der Cluster **1714A-1714N** des Verarbeitungsclusterarrays **1712** Daten verarbeiten, die in eine der Speichereinheiten **1724A-1724N** innerhalb des Parallelprozessorspeichers **1722** geschrieben werden. In mindestens einer Ausführungsform kann die Speicher-Crossbar **1716** dazu konfiguriert sein, eine Ausgabe jedes Clusters **1714A-1714N** an eine beliebige Partitionseinheit **1720A-1720N** oder an einen anderen Cluster **1714A-1714N** zu übertragen, die zusätzliche Verarbeitungsoperationen an einer Ausgabe durchführen können. In mindestens einer Ausführungsform kann jeder Cluster **1714A-1714N** mit der Speicherschnittstelle **1718** über die Speicher-Crossbar **1716** kommunizieren, um von verschiedenen externen Speichervorrichtungen zu lesen oder in sie zu schreiben. In mindestens einer Ausführungsform weist die Speicher-Crossbar **1716** eine Verbindung zur Speicherschnittstelle **1718** auf, um mit der E/A-Einheit **1704** zu kommunizieren, sowie eine Verbindung zu einer lokalen Instanz des Parallelprozessorspeichers **1722**, wodurch Verarbeitungseinheiten innerhalb unterschiedlicher Verarbeitungscluster **1714A-1714N** ermöglicht werden, um mit dem Systemspeicher oder einem anderen Speicher zu kommunizieren, der nicht lokal für die Parallelverarbeitungseinheit **1702** ist. In mindestens einer Ausführungsform kann die Speicher-Crossbar **1716** virtuelle Kanäle verwenden, um Verkehrsströme zwischen den Clustern **1714A-1714N** und den Partitionseinheiten **1720A-1720N** zu trennen.

[0161] In mindestens einer Ausführungsform können mehrere Instanzen der Parallelverarbeitungseinheit **1702** auf einer einzelnen Erweiterungskarte bereitgestellt sein oder es können mehrere Erweiterungskarten miteinander verbunden sein. In mindestens einer Ausführungsform können unterschiedliche Instanzen der Parallelverarbeitungseinheit **1702** dazu konfiguriert sein, zusammenzuarbeiten, selbst wenn unterschiedliche Instanzen unterschiedliche Anzahlen von Verarbeitungskernen, unterschiedliche Mengen an lokalem Parallelprozessorspeicher und/oder andere Konfigurationsunterschiede aufweisen. Zum Beispiel können in mindestens einer Ausführungsform einige Instanzen der Parallelverarbeitungseinheit **1702** Gleitkommaeinheiten mit höherer Genauigkeit in Bezug auf andere Instanzen beinhalten. In mindestens einer Ausführungsform können Systeme, die eine oder mehrere Instanzen der Parallelverarbeitungseinheit **1702** oder des parallelen Prozessors **1700** einschließen, in einer Reihe von Konfigurationen und Formfaktoren umgesetzt sein, was Desktop-, Laptop- oder Handheld-Personalcomputer, Server, Workstations, Spielekonsolen und/oder eingebettete Systeme beinhaltet, ohne darauf beschränkt zu sein.

[0162] **Fig. 17B** ist ein Blockdiagramm einer Partitionseinheit **1720** gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform ist die Partitionseinheit **1720** eine Instanz einer der Partitionseinheiten **1720A-1720N** aus **Fig. 17A**. In mindestens einer Ausführungsform beinhaltet die Partitionseinheit **1720** einen L2-Zwischenspeicher **1721**, eine Bildspeicherschnittstelle **1725** und eine Rasteroperationseinheit (raster operations unit - „ROP“) **1726**. Der L2-Zwischenspeicher **1721** ist ein Lese-/Schreib-Zwischenspeicher, der dazu konfiguriert ist, Lade- und Speicheroperationen auszuführen, die von der Speicher-Crossbar **1716** und

der ROP **1726** empfangen werden. In mindestens einer Ausführungsform werden Lesefehler und dringende Rückschreibanforderungen vom L2-Zwischenspeicher **1721** zur Verarbeitung an die Bildspeicherschnittstelle **1725** ausgegeben. In mindestens einer Ausführungsform können Aktualisierungen auch über die Bildspeicherschnittstelle **1725** zur Verarbeitung an einen Bildspeicher gesendet werden. In mindestens einer Ausführungsform ist die Bildspeicherschnittstelle **1725** mit einer der Speichereinheiten im Parallelprozessorspeicher, wie etwa den Speichereinheiten **1724A-1724N** von aus **Fig. 17** (z. B. innerhalb des Parallelprozessorspeichers **1722**) verbunden.

[0163] In mindestens einer Ausführungsform ist die ROP **1726** eine Verarbeitungseinheit, die Rasteroperationen, wie etwa Schablone, Z-Test, Mischen und so weiter, durchführt. In mindestens einer Ausführungsform gibt die ROP **1726** dann verarbeitete Grafikdaten aus, die im Grafikspeicher gespeichert sind. In mindestens einer Ausführungsform beinhaltet die ROP **1726** Komprimierungslogik zum Komprimieren von Tiefen- oder Farbdaten, die in den Speicher geschrieben werden, und zum Dekomprimieren von Tiefen- oder Farbdaten, die aus dem Speicher gelesen werden. In mindestens einer Ausführungsform kann die Komprimierungslogik eine verlustfreie Komprimierungslogik sein, die einen oder mehrere von mehreren Komprimierungsalgorithmen verwendet. Die von der ROP **1726** durchgeführte Komprimierungslogik kann auf Grundlage von statistischen Eigenschaften der zu komprimierenden Daten variieren. Zum Beispiel wird in mindestens einer Ausführungsform eine Delta-Farbkompression an Tiefen- und Farbdaten pro Kachel durchgeführt.

[0164] In mindestens einer Ausführungsform ist die ROP **1726** in jedem Verarbeitungscluster (z. B. Cluster **1714A-1714N** aus **Fig. 17A)** anstatt in der Partitionseinheit **1720** beinhaltet. In mindestens einer Ausführungsform werden Lese- und Schreibanforderungen für Bildpunktdaten anstelle von Bildpunktfragmentdaten über die Speicher-Crossbar **1716** übermittelt. In mindestens einer Ausführungsform können verarbeitete Grafikdaten auf einer Anzeigevorrichtung angezeigt werden, wie etwa einer oder mehreren Anzeigevorrichtungen **1610** aus **Fig. 16**, die zur Weiterverarbeitung durch den/die Prozessor(en) **1602** weitergeleitet oder zur Weiterverarbeitung durch eine von den Verarbeitungseinheiten innerhalb des Parallelprozessors **1700** aus **Fig. 17A** weitergeleitet wurde.

[0165] Fig. 17C ist ein Blockdiagramm eines Verarbeitungsclusters **1714** innerhalb einer Parallelverarbeitungseinheit gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform ist ein Verarbeitungscluster eine Instanz von einem der Verarbeitungscluster **1714A-1714N** aus **Fig. 17A**. In mindestens einer Ausführungsform können einer oder mehrere Verarbeitungscluster **1714** dazu konfiguriert sein, viele Threads parallel auszuführen, wobei sich „Thread“ auf eine Instanz eines bestimmten Programms bezieht, das mit einem bestimmten Satz von Eingabedaten ausgeführt wird. In mindestens einer Ausführungsform werden Einzelbefehls-Mehrfachdaten-(single-instruction, multiple-data - SIMD-)Befehlsausgabemethoden verwendet, um die parallele Ausführung einer großen Anzahl von Threads zu unterstützen, ohne mehrere unabhängige Anweisungseinheiten bereitzustellen. In mindestens einer Ausführungsform werden Einzelbefehls-Mehrfachthread-(SIMT)-Methoden verwendet, um die parallele Ausführung einer großen Anzahl von im Allgemeinen synchronisierten Threads zu unterstützen, die eine gemeinsame Befehlseinheit verwenden, die dazu konfiguriert ist, Befehle an einen Satz von Verarbeitungsengines innerhalb von jedem der Verarbeitungscluster auszugeben.

[0166] In mindestens einer Ausführungsform kann der Betrieb des Verarbeitungsclusters **1714** über einen Pipelineverwalter **1732** gesteuert werden, der Verarbeitungsaufgaben an SIMT-Parallelprozessoren verteilt. In mindestens einer Ausführungsform empfängt der Pipeline-Verwalter **1732** Anweisungen von dem Planer **1710** aus **Fig. 17A** und verwaltet die Ausführung dieser Anweisungen über einen Grafikmultiprozessor **1734** und/oder eine Textureinheit **1736**. In mindestens einer Ausführungsform ist der Grafikmultiprozessor **1734** eine beispielhafte Instanz eines SIMT-Parallelprozessors. In mindestens einer Ausführungsform können jedoch verschiedene Arten von SIMT-Parallelprozessoren unterschiedlicher Architekturen innerhalb des Verarbeitungsclusters **1714** beinhaltet sein. In mindestens einer Ausführungsform können eine oder mehrere Instanzen des Grafikmultiprozessors **1734** innerhalb eines Verarbeitungsclusters **1714** beinhaltet sein. In mindestens einer Ausführungsform kann der Grafikmultiprozessor **1734** Daten verarbeiten und eine Daten-Crossbar **1740** kann verwendet werden, um verarbeitete Daten an eines von mehreren möglichen Zielen zu verteilen, was andere Shader-Einheiten beinhaltet. In mindestens einer Ausführungsform kann der Pipeline-Verwalter **1732** die Verteilung verarbeiteter Daten durch das Festlegen der Ziele für zu verteilende verarbeitete Daten über die Daten-Crossbar **1740** erleichtern.

[0167] In mindestens einer Ausführungsform kann jeder Grafikmultiprozessor **1734** innerhalb des Verarbeitungsclusters **1714** einen identischen Satz von funktioneller Ausführungslogik (z. B. arithmetische Logikeinheiten, Lade-Speicher-Einheiten usw.) beinhalten. In mindestens einer Ausführungsform kann eine funktionelle Ausführungslogik in einer Pipeline-Weise konfiguriert sein, in der neue Anweisungen ausgegeben werden

können, bevor vorherige Anweisungen abgeschlossen sind. In mindestens einer Ausführungsform unterstützt die funktionelle Ausführungslogik eine Reihe von Operationen, die Ganzzahl- und Gleitkommaarithmetik, Vergleichsoperationen, Boolesche Operationen, Bitverschiebung und Berechnung verschiedener algebraischer Funktionen beinhalten. In mindestens einer Ausführungsform kann dieselbe Hardware der funktionellen Einheit genutzt werden, um unterschiedliche Operationen durchzuführen, und es kann eine beliebige Kombination von funktionellen Einheiten vorhanden sein.

[0168] In mindestens einer Ausführungsform bilden Anweisungen, die an den Verarbeitungscluster **1714** übermittelt werden, einen Thread. In mindestens einer Ausführungsform ist ein Satz von Threads, die über einen Satz von Parallelverarbeitungsengines ausgeführt werden, eine Thread-Gruppe. In mindestens einer Ausführungsform führt die Thread-Gruppe ein Programm mit unterschiedlichen Eingabedaten aus. In mindestens einer Ausführungsform kann jeder Thread innerhalb einer Thread-Gruppe einer anderen Verarbeitungsengine innerhalb eines Grafikmultiprozessors **1734** zugewiesen werden. In mindestens einer Ausführungsform kann eine Thread-Gruppe weniger Threads als eine Anzahl von Verarbeitungsengines innerhalb des Grafikmultiprozessors **1734** beinhalten. In mindestens einer Ausführungsform können, wenn eine Thread-Gruppe weniger Threads als eine Anzahl von Verarbeitungsengines beinhaltet, eine oder mehrere Verarbeitungsengines während Zyklen, in denen diese Thread-Gruppe verarbeitet wird, inaktiv sein. In mindestens einer Ausführungsform kann eine Thread-Gruppe außerdem mehr Threads als eine Anzahl von Verarbeitungsengines innerhalb des Grafikmultiprozessors **1734** beinhalten. In mindestens einer Ausführungsform kann die Verarbeitung, wenn eine Thread-Gruppe mehr Threads als Verarbeitungsengines innerhalb des Grafikmultiprozessors **1734** beinhaltet, über aufeinanderfolgende Taktzyklen durchgeführt werden. In mindestens einer Ausführungsform können mehrere Thread-Gruppen gleichzeitig auf einem Grafikmultiprozessor **1734** ausgeführt werden.

[0169] In mindestens einer Ausführungsform beinhaltet der Grafikmultiprozessor **1734** einen internen schnellen Pufferspeicher, um Lade- und Speicheroperationen durchzuführen. In mindestens einer Ausführungsform kann der Grafikmultiprozessor **1734** auf einen internen Zwischenspeicher verzichten und einen schnellen Pufferspeicher (z. B. L1-Zwischenspeicher **1748**) innerhalb des Verarbeitungsclusters **1714** verwenden. In mindestens einer Ausführungsform hat jeder Grafikmultiprozessor **1734** auch Zugriff auf L2-Zwischenspeicher innerhalb von Partitionseinheiten (z. B. die Partitionseinheiten **1720A-1720N** aus **Fig. 17A)**, die von allen Verarbeitungsclustem **1714** gemeinsam genutzt werden, und kann verwendet werden, um Daten zwischen Threads zu übertragen. In mindestens einer Ausführungsform kann der Grafikmultiprozessor **1734** auch auf den chipexternen globalen Speicher zugreifen, der einen oder mehrere von dem lokalen Parallelprozessorspeicher und/oder Systemspeicher beinhalten kann. In mindestens einer Ausführungsform kann jeder beliebige Speicher außerhalb der Parallelverarbeitungseinheit **1702** als globaler Speicher verwendet werden. In mindestens einer Ausführungsform beinhaltet der Verarbeitungscluster **1714** mehrere Instanzen des Grafikmultiprozessors **1734**, der gemeinsame Anweisungen und Daten gemeinsam nutzen kann, die im L1-Zwischenspeicher **1748** gespeichert sein können.

[0170] In mindestens einer Ausführungsform kann jeder Verarbeitungscluster **1714** eine Speicherverwaltungseinheit („MMU“) **1745** beinhalten, die dazu konfiguriert ist, virtuelle Adressen physischen Adressen zuzuordnen. In mindestens einer Ausführungsform können sich eine oder mehrere Instanzen der MMU **1745** innerhalb der Speicherschnittstelle **1718** aus **Fig. 17A** befinden. In mindestens einer Ausführungsform beinhaltet die MMU **1745** einen Satz von Seitentabelleneinträgen (page table entries - PTEs), die verwendet werden, um eine virtuelle Adresse einer physischen Adresse einer Kachel und optional einem Zwischenspeicherzeilenindex zuzuordnen. In mindestens einer Ausführungsform kann die MMU **1745** Adressübersetzungspuffer (TLB) oder Zwischenspeicher beinhalten, die sich innerhalb des Grafikmultiprozessors **1734** oder des L1-Zwischenspeichers oder des Verarbeitungsclusters **1714** befinden können. In mindestens einer Ausführungsform wird die physische Adresse verarbeitet, um den Oberflächendatenzugriffsstandort zu verteilen, um eine effiziente Anforderungsverschachtelung zwischen Partitionseinheiten zu ermöglichen. In mindestens einer Ausführungsform kann der Zwischenspeicherzeilenindex verwendet werden, um zu bestimmen, ob eine Anforderung nach einer Zwischenspeicherzeile ein Treffer oder ein Fehler ist.

[0171] In mindestens einer Ausführungsform kann ein Verarbeitungscluster **1714** derartig konfiguriert sein, dass jeder Grafikmultiprozessor **1734** an eine Textureinheit **1736** zum Durchführen von Texturzuordnungsoperationen gekoppelt ist, z. B. zum Bestimmen von Texturabtastpositionen, Lesen von Texturdaten und Filtern von Texturdaten. In mindestens einer Ausführungsform werden Texturdaten aus einem internen Textur-L1-Zwischenspeicher (nicht gezeigt) oder aus einem L1-Zwischenspeicher innerhalb des Grafikmultiprozessors **1734** gelesen und nach Bedarf aus einem L2-Zwischenspeicher, einem lokalen Parallelprozessorspeicher oder Systemspeicher abgerufen. In mindestens einer Ausführungsform gibt jeder Grafikmultiprozessor **1734** verarbeitete Aufgaben an die Daten-Crossbar **1740** aus, um verarbeitete Aufgaben einem anderen Ver-

arbeitungs-Cluster **1714** zur weiteren Verarbeitung bereitzustellen oder verarbeitete Aufgaben in einem L2-Zwischenspeicher, einem lokalen Parallelprozessorspeicher oder Systemspeicher über die Speicher-Crossbar **1716** zu speichern. In mindestens einer Ausführungsform ist preROP **1742** (pre-raster operations unit - Vorrasteroperationseinheit) dazu konfiguriert, Daten vom Grafikmultiprozessor **1734** zu empfangen, um Daten an ROP-Einheiten zu leiten, die sich bei Partitionseinheiten befinden können, wie in dieser Schrift beschrieben (z. B. die Partitionseinheiten **1720A-1720N** aus **Fig. 17A)**. In mindestens einer Ausführungsform kann die PreROP-1742-Einheit Optimierungen für die Farbmischung durchführen, Bildpunktfarbdaten organisieren und Adressübersetzungen durchführen.

[0172] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. Details bezüglich der Ableitungs- und/oder Trainingslogik **615** werden nachfolgend in Verbindung mit den **Fig. 6A** und/oder 6B bereitgestellt. In mindestens einer Ausführungsform kann eine Ableitungs- und/oder Trainingslogik **615** in dem Grafikverarbeitungscluster **1714** zum Ableiten oder Vorhersagen von Operationen mindestens zum Teil auf Grundlage von Gewichtungsparametern verwendet werden, die unter Verwendung von neuronalen Netzwerktrainingsoperationen, Funktionen und/oder Architekturen von neuronalen Netzwerken oder in dieser Schrift beschriebenen Anwendungsfällen von neuronalen Netzwerken berechnet wurden.

[0173] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. In mindestens einer Ausführungsform kann diese Logik mit Komponenten dieser Figuren verwendet werden, um ein dreidimensionales Bild einer bestimmten Ansicht aus zweidimensionalen Bildern aus einer willkürlichen Anzahl von Kameras zu erzeugen.

[0174] **Fig. 17D** zeigt einen Grafikmultiprozessor **1734** gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform ist der Grafikmultiprozessor **1734** mit dem Pipeline-Verwalter **1732** des Verarbeitungsclusters **1714** gekoppelt. In mindestens einer Ausführungsform weist der Grafikmultiprozessor **1734** eine Ausführungspipeline auf, die einen Anweisungszwischenspeicher **1752**, eine Anweisungseinheit **1754**, eine Adresszuordnungseinheit **1756**, eine Registerdatei **1758**, eine oder mehrere Allzweck-Grafikverarbeitungseinheits-(GPGPU-)Kerne **1762** und eine oder mehrere Lade-/Speichereinheiten **1766** beinhaltet, ohne darauf beschränkt zu sein. Ein oder mehrere GPGPU-Kerne **1762** und Lade-/Speichereinheiten **1766** sind mit dem schnellen Pufferspeicher **1772** und dem gemeinsam genutzten Speicher **1770** über eine Speicher- und Zwischenspeicher-Verbindung **1768** gekoppelt.

[0175] In mindestens einer Ausführungsform empfängt der Anweisungszwischenspeicher **1752** einen Strom von Anweisungen zur Ausführung vom Pipeline-Verwalter **1732**. In mindestens einer Ausführungsform werden Anweisungen im Anweisungszwischenspeicher **1752** zwischengespeichert und zur Ausführung durch die Anweisungseinheit **1754** gesendet. In mindestens einer Ausführungsform kann die Anweisungseinheit **1754** Anweisungen als Thread-Gruppen (z. B. Warps) versenden, wobei jede Thread-Gruppe einer anderen Ausführungseinheit innerhalb des einen oder der mehreren GPGPU-Kerne **1762** zugewiesen ist. In mindestens einer Ausführungsform kann ein Befehl auf einen lokalen, gemeinsam genutzten oder globalen Adressbereich durch das Festlegen einer Adresse innerhalb eines einheitlichen Adressbereichs zugreifen. In mindestens einer Ausführungsform kann die Adresszuordnungseinheit **1756** verwendet werden, um Adressen in einem einheitlichen Adressbereich in eine eindeutige Speicheradresse zu übersetzen, auf die durch die Lade-/Speichereinheit(en) **1766** zugegriffen werden kann.

[0176] In mindestens einer Ausführungsform stellt die Registerdatei **1758** einen Satz von Registern für die Funktionseinheiten des Grafikmultiprozessors **1734** bereit. In mindestens einer Ausführungsform stellt die Registerdatei **1758** einen temporären Speicher für Operanden bereit, die mit Datenpfaden von funktionellen Einheiten (z. B. GPGPU-Kernen **1762**, Lade-/Speichereinheit(en) **1766**) des Grafikmultiprozessors **1734** verbunden sind. In mindestens einer Ausführungsform wird die Registerdatei **1758** derartig zwischen jeder der funktionellen Einheiten aufgeteilt, dass jeder funktionellen Einheit ein dedizierter Teil der Registerdatei **1758** zugewiesen wird. In mindestens einer Ausführungsform ist die Registerdatei **1758** auf unterschiedliche Warps aufgeteilt, die vom Grafikmultiprozessor **1734** ausgeführt werden.

[0177] In mindestens einer Ausführungsform können die GPGPU-Kerne **1762** jeweils Gleitkommaeinheiten (FPUs) und/oder Ganzzahlarithmetiklogikeinheiten (ALUs) beinhalten, die verwendet werden, um Anweisungen des Grafikmultiprozessors **1734** auszuführen. GPGPU-Kerne **1762** können eine ähnliche Architektur aufweisen oder sich in der Architektur unterscheiden. In mindestens einer Ausführungsform beinhaltet ein erster Teil von GPGPU-Kernen **1762** eine FPU mit einfacher Genauigkeit und eine ganzzahlige ALU, während ein

zweiter Teil von GPGPU-Kernen eine FPU mit doppelter Genauigkeit beinhaltet. In mindestens einer Ausführungsform können FPUs den Standard IEEE 754-2008 für Gleitkommaarithmetik umsetzen oder Gleitkommaarithmetik mit variabler Genauigkeit ermöglichen. In mindestens einer Ausführungsform kann der Grafikmultiprozessor **1734** zusätzlich eine oder mehrere feste Funktions- oder Spezialfunktionseinheiten beinhalten, um konkrete Funktionen, wie etwa Kopierrechteck- oder Bildpunktmischoperationen, auszuführen. In mindestens einer Ausführungsform können einer oder mehrere der GPGPU-Kerne auch feste oder Spezialfunktionslogik beinhalten.

[0178] In mindestens einer Ausführungsform beinhalten die GPGPU-Kerne **1762** eine SIMD-Logik, die in der Lage ist, eine einzelne Anweisung an mehreren Datensätzen auszuführen. In mindestens einer Ausführungsform können die GPGPU-Kerne **1762** SIMD4-, SIMD8- und SIMD16-Anweisungen physisch ausführen und SIMD1-, SIMD2- und SIMD32-Anweisungen logisch ausführen. In mindestens einer Ausführungsform können SIMD-Anweisungen für GPGPU-Kerne zur Kompilierzeit von einem Shader-Compiler erzeugt werden oder automatisch erzeugt werden, wenn Programme ausgeführt und kompiliert werden, die für Einzelprogramm-Mehrfachdaten-(SPMD-) oder SIMT-Architekturen geschrieben und kompiliert wurden. In mindestens einer Ausführungsform können mehrere Threads eines Programms, das für ein SIMT-Ausführungsmodell konfiguriert ist, über eine einzige SIMD-Anweisung ausgeführt werden. Zum Beispiel können in mindestens einer Ausführungsform acht SIMT-Threads, die gleiche oder ähnliche Operationen ausführen, über eine einzige SIMD8-Logikeinheit parallel ausgeführt werden.

[0179] In mindestens einer Ausführungsform ist die Speicher- und Zwischenspeicher-Verbindung **1768** ein Verbindungsnetzwerk, das jede funktionelle Einheit des Grafikmultiprozessors **1734** mit der Registerdatei **1758** und mit dem gemeinsam genutzten Speicher **1770** verbindet. In mindestens einer Ausführungsform ist die Speicher- und Zwischenspeicher-Verbindung **1768** eine Crossbar-Verbindung, die es der Lade-/Speichereinheit **1766** ermöglicht, Lade- und Speicheroperationen zwischen dem gemeinsam genutzten Speicher **1770** und der Registerdatei **1758** umzusetzen. In mindestens einer Ausführungsform kann die Registerdatei **1758** mit derselben Frequenz wie die GPGPU-Kerne **1762** arbeiten, wodurch die Datenübertragung zwischen den GPGPU-Kernen **1762** und der Registerdatei **1758** eine sehr geringe Latenzzeit hat. In mindestens einer Ausführungsform kann der gemeinsam genutzte Speicher **1770** verwendet werden, um eine Kommunikation zwischen Threads zu ermöglichen, die auf funktionellen Einheiten innerhalb des Grafikmultiprozessors **1734** ausgeführt werden. In mindestens einer Ausführungsform kann der schnelle Pufferspeicher **1772** zum Beispiel als Datenzwischenspeicher verwendet werden, um zwischen funktionellen Einheiten und der Textureinheit **1736** kommunizierte Texturdaten zwischenzuspeichern. In mindestens einer Ausführungsform kann der gemeinsam genutzte Speicher **1770** auch als programmverwalteter Zwischenspeicher verwendet werden. In mindestens einer Ausführungsform können Threads, die auf GPGPU-Kernen **1762** ausgeführt werden, zusätzlich zu automatisch zwischengespeicherten Daten, die im schnellen Pufferspeicher **1772** gespeichert sind, programmgesteuert Daten innerhalb des gemeinsam genutzten Speichers speichern.

[0180] In mindestens einer Ausführungsform ist ein Parallelprozessor oder eine GPGPU, wie in dieser Schrift beschrieben, kommunikativ an Host-/Prozessorkerne gekoppelt, um Grafikoperationen, Maschinenlemoperationen, Musteranalyseoperationen und verschiedene Allzweck-GPU-(GPGPU)-Funktionen zu beschleunigen. In mindestens einer Ausführungsform kann die GPU kommunikativ über einen Bus oder eine andere Verbindung (z. B. eine Hochgeschwindigkeitsverbindung, wie etwa PCIe oder NVLink) an Hostprozessor/Kerne gekoppelt sein. In mindestens einer Ausführungsform kann die GPU in demselben Gehäuse oder Chip wie Kerne integriert sein und kommunikativ über einen internen Prozessorbus/eine interne Verbindung (d. h. innerhalb des Gehäuses oder Chips) an die Kerne gekoppelt sein. In mindestens einer Ausführungsform können Prozessorkerne ungeachtet der Art und Weise, in der die GPU verbunden ist, der GPU Arbeit in Form von Sequenzen von Befehlen/Anweisungen zuweisen, die in einem Arbeitsdeskriptor enthalten sind. In mindestens einer Ausführungsform verwendet die GPU dann eine dedizierte Schaltung/Logik zum effizienten Verarbeiten dieser Befehle/Anweisungen.

[0181] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. Details bezüglich der Ableitungs- und/oder Trainingslogik **615** werden nachfolgend in Verbindung mit den **Fig. 6A** und/oder 6B bereitgestellt. In mindestens einer Ausführungsform kann eine Ableitungs- und/oder Trainingslogik **615** in dem Grafikmultiprozessor **1734** zum Ableiten oder Vorhersagen von Operationen mindestens zum Teil auf Grundlage von Gewichtungsparametern verwendet werden, die unter Verwendung von neuronalen Netzwerktrainingsoperationen, Funktionen und/oder Architekturen von neuronalen Netzwerken oder in dieser Schrift beschriebenen Anwendungsfällen von neuronalen Netzwerken berechnet wurden.

[0182] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. In mindestens einer Ausführungsform kann diese Logik mit Komponenten dieser Figuren verwendet werden, um ein dreidimensionales Bild einer bestimmten Ansicht aus zweidimensionalen Bildern aus einer willkürlichen Anzahl von Kameras zu erzeugen.

[0183] **Fig. 18** veranschaulicht ein Multi-GPU-Computersystem **1800** gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform kann das Multi-GPU-Computersystem **1800** einen Prozessor **1802** beinhalten, der über einen Hostschnittstellen-Switch **1804** an mehrere Allzweck-Grafikverarbeitungseinheiten (GPGPUs) **1806A-D** gekoppelt ist. In mindestens einer Ausführungsform ist der Hostschnittstellen-Switch **1804** eine PCI-Express-Switch-Vorrichtung, die den Prozessor **1802** mit einem PCI-Express-Bus koppelt, über den der Prozessor **1802** mit den GPGPUs **1806A-D** kommunizieren kann. GPGPUs **1806A-D** können über einen Satz von Hochgeschwindigkeits-Punkt-zu-Punkt-GPU-zu-GPU-Links **1816** miteinander verbunden werden. In mindestens einer Ausführungsform sind GPU-zu-GPU-Links **1816** über einen dedizierten GPU-Link mit jeder der GPGPUs **1806A-D** verbunden. In mindestens einer Ausführungsform ermöglichen P2P-GPU-Links **1816** eine direkte Kommunikation zwischen jeder der GPGPUs **1806A-D**, ohne dass eine Kommunikation über den Hostschnittstellenbus **1804** erforderlich ist, mit dem der Prozessor **1802** verbunden ist. In mindestens einer Ausführungsform bleibt der Hostschnittstellenbus **1804** mit GPU-zu-GPU-Verkehr, der an P2P-GPU-Links **1816** gerichtet ist, für den Systemspeicherzugriff verfügbar oder um mit anderen Instanzen des Multi-GPU-Computersystems **1800** zu kommunizieren, zum Beispiel über eine oder mehr Netzwerkvorrichtungen. Während in mindestens einer Ausführungsform die GPGPUs **1806A-D** über den Hostschnittstellen-Switch **1804** mit dem Prozessor **1802** verbunden sind, beinhaltet der Prozessor **1802** in mindestens einer Ausführungsform eine direkte Unterstützung für P2P-GPU-Links 1816 und kann sich direkt mit den GPGPUs **1806A-D** verbinden.

[0184] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. Details bezüglich der Ableitungs- und/oder Trainingslogik **615** werden nachfolgend in Verbindung mit den **Fig. 6A** und/oder 6B bereitgestellt. In mindestens einer Ausführungsform kann eine Ableitungs- und/oder Trainingslogik **615** in dem Multi-GPU-Computersystem **1800** zum Ableiten oder Vorhersagen von Operationen mindestens zum Teil auf Grundlage von Gewichtungsparametern verwendet werden, die unter Verwendung von neuronalen Netzwerktrainingsoperationen, Funktionen und/oder Architekturen von neuronalen Netzwerken oder in dieser Schrift beschriebenen Anwendungsfällen von neuronalen Netzwerken berechnet wurden.

[0185] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. In mindestens einer Ausführungsform kann diese Logik mit Komponenten dieser Figuren verwendet werden, um ein dreidimensionales Bild einer bestimmten Ansicht aus zweidimensionalen Bildern aus einer willkürlichen Anzahl von Kameras zu erzeugen.

[0186] **Fig. 19** ist ein Blockdiagramm eines Grafikprozessors **1900** gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform beinhaltet der Grafikprozessor **1900** eine Ringverbindung **1902**, ein Pipeline-Front-End **1904**, eine Medienengine **1937** und Grafikkerne **1980A-1980N**. In mindestens einer Ausführungsform koppelt die Ringverbindung **1902** den Grafikprozessor **1900** an andere Verarbeitungseinheiten, die andere Grafikprozessoren oder einen oder mehrere Allzweck-Prozessorkerne beinhalten. In mindestens einer Ausführungsform ist der Grafikprozessor **1900** einer von vielen Prozessoren, die in ein Mehrkernverarbeitungssystem integriert sind.

[0187] In mindestens einer Ausführungsform empfängt der Grafikprozessor **1900** Batches von Befehlen über die Ringverbindung **1902**. In mindestens einer Ausführungsform werden eingehende Befehle von einem Befehlsstreamer **1903** im Pipeline-Front-End **1904** ausgelegt. In mindestens einer Ausführungsform beinhaltet der Grafikprozessor **1900** eine skalierbare Ausführungslogik, um eine 3D-Geometrieverarbeitung und eine Medienverarbeitung über einen oder mehrere Grafikkerne **1980A-1980N** durchzuführen. In mindestens einer Ausführungsform liefert der Befehlsstreamer **1903** für 3D-Geometrieverarbeitungsbefehle Befehle an die Geometriepipeline **1936**. In mindestens einer Ausführungsform liefert der Befehlsstreamer **1903** für mindestens einige Medienverarbeitungsbefehle Befehle an ein Video-Front-End **1934**, das mit einer Medienengine **1937** gekoppelt ist. In mindestens einer Ausführungsform beinhaltet die Medienengine **1937** eine Videoqualitätsengine (VQE) **1930** für die Video- und Bildnachbearbeitung und eine Mehrformat-Kodierungs-/Dekodierungs-(MFX) 1933engine, um eine hardwarebeschleunigte Mediendatenkodierung und -dekodierung bereitzustellen. In mindestens einer Ausführungsform erzeugen die Geometriepipeline **1936** und die Medienengine **1937** je-

weils Ausführungs-Threads für Thread-Ausführungsressourcen, die von mindestens einem Grafikkern **1980A** bereitgestellt werden.

[0188] In mindestens einer Ausführungsform beinhaltet der Grafikprozessor **1900** skalierbare Thread-Ausführungsressourcen mit modularen Kernen **1980A-1980N** (manchmal als Kern-Slices bezeichnet), die jeweils mehrere Teilkerne **1950A-1950N**, **1960A-1960N** (manchmal als KernTeil-Slices bezeichnet) aufweisen. In mindestens einer Ausführungsform kann der Grafikprozessor **1900** eine beliebige Anzahl von Grafikkernen **1980A** bis **1980N** aufweisen. In mindestens einer Ausführungsform beinhaltet der Grafikprozessor **1900** einen Grafikkern **1980A**, der mindestens einen ersten Teilkern **1950A** und einen zweiten Teilkern **1960A** aufweist. In mindestens einer Ausführungsform ist der Grafikprozessor **1900** ein Niedrigenergieprozessor mit einem einzelnen Teilkern (z. B. **1950A**). In mindestens einer Ausführungsform beinhaltet der Grafikprozessor **1900** mehrere Grafikkerne **1980A-1980N**, von denen jeder einen Satz von ersten Teilkernen **1950A-1950N** und einen Satz von zweiten Teilkernen **1960A-1960N** beinhaltet. In mindestens einer Ausführungsform beinhaltet jeder Teilkern in den ersten Teilkernen **1950A-1950N** mindestens einen ersten Satz von Ausführungseinheiten **1952A-1952N** und Medien-/Textur-Abtaster **1954A-1954N**. In mindestens einer Ausführungsform beinhaltet jeder Teilkern in den zweiten Teilkernen **1960A-1960N** mindestens einen zweiten Satz von Ausführungseinheiten **1962A-1962N** und Abtastern **1964A-1964N**. In mindestens einer Ausführungsform teilt sich jeder Teilkern **1950A-1950N**, **1960A-1960N** einen Satz gemeinsam genutzter Ressourcen **1970A-1970N**. In mindestens einer Ausführungsform beinhalten gemeinsam genutzte Ressourcen gemeinsam genutzte(n) schnellen Pufferspeicher und Bildpunktoperationslogik.

[0189] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. Details bezüglich der Ableitungs- und/oder Trainingslogik **615** werden nachfolgend in Verbindung mit den **Fig. 6A** und/oder 6B bereitgestellt. In mindestens einer Ausführungsform kann eine Ableitungs- und/oder Trainingslogik **615** in dem Prozessor **1900** zum Ableiten oder Vorhersagen von Operationen mindestens zum Teil auf Grundlage von Gewichtungsparametern verwendet werden, die unter Verwendung von neuronalen Netzwerktrainingsoperationen, Funktionen und/oder Architekturen von neuronalen Netzwerken oder in dieser Schrift beschriebenen Anwendungsfällen von neuronalen Netzwerken berechnet wurden.

[0190] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. In mindestens einer Ausführungsform kann diese Logik mit Komponenten dieser Figuren verwendet werden, um ein dreidimensionales Bild einer bestimmten Ansicht aus zweidimensionalen Bildern aus einer willkürlichen Anzahl von Kameras zu erzeugen.

[0191] **Fig. 20** ist ein Blockdiagramm, das eine Mikroarchitektur für einen Prozessor **2000** veranschaulicht, der Logikschaltungen beinhalten kann, um Anweisungen gemäß mindestens einer Ausführungsform auszuführen. In mindestens einer Ausführungsform kann der Prozessor **2000** Anweisungen ausführen, die x86-Anweisungen, ARM-Anweisungen, spezialisierte Anweisungen für anwendungsspezifische integrierte Schaltkreise (ASICs) usw. beinhalten In mindestens einer Ausführungsform kann der Prozessor **2000** Register beinhalten, um gepackte Daten zu speichern, wie etwa 64 Bit breite MMX™-Register in Mikroprozessoren, die mit MMX-Technologie der Intel Corporation aus Santa Clara, Kalifornien, aktiviert sind. In mindestens einer Ausführungsform können MMX-Register, die sowohl in ganzzahliger als auch in Gleitkommaform verfügbar sind, mit gepackten Datenelementen arbeiten, die Anweisungen mit Einzelbefehlen, Mehrfachdaten („SIMD“) und Streaming-SIMD-Erweiterungen (streaming SIMD extensions - „SSE“) begleiten. In mindestens einer Ausführungsform können 128 Bit breite XMM-Register, die sich auf SSE2, SSE3, SSE4, AVX oder eine darüberhinausgehende (allgemein als „SSEx“ bezeichnete) Technologie beziehen, derartige gepackte Datenoperanden halten. In mindestens einer Ausführungsform kann der Prozessor **2000** Anweisungen ausführen, um Algorithmen des maschinellen Lernens oder von Deep Learning, das Trainieren oder das Ableiten zu beschleunigen.

[0192] In mindestens einer Ausführungsform beinhaltet der Prozessor **2000** ein reihenfolgengetreues Front-End („Front-End“) **2001**, um auszuführende Anweisungen abzurufen und Anweisungen vorzubereiten, die später in der Prozessorpipeline verwendet werden sollen. In mindestens einer Ausführungsform kann das Front-End **2001** mehrere Einheiten beinhalten. In mindestens einer Ausführungsform ruft ein Anweisungsvorabrufer **2026** Anweisungen aus dem Speicher ab und führt einem Anweisungsdekodierer **2028** Anweisungen zu, der wiederum Anweisungen dekodiert oder auslegt. Zum Beispiel dekodiert der Anweisungsdekodierer **2028** in mindestens einer Ausführungsform eine empfangene Anweisung in eine oder mehrere Operationen, die als „Mikroanweisungen“ oder „Mikrooperationen“ (auch als „Mikroops“ oder „uops“ bezeichnet) bezeichnet werden, welche diese Maschine ausführen kann. In mindestens einer Ausführungsform zerlegt der Anweisungs-

dekodierer **2028** die Anweisung in einen Operationscode und entsprechende Daten- und Steuerfelder, die möglicherweise von der Mikroarchitektur verwendet werden, um Operationen gemäß mindestens einer Ausführungsform durchzuführen. In mindestens einer Ausführungsform kann ein Ablaufverfolgungszwischenspeicher **2030** dekodierte uops in programmgeordnete Sequenzen oder Ablaufverfolgungen in einer uop-Warteschlange **2034** zur Ausführung zusammenstellen. In mindestens einer Ausführungsform stellt ein Mikrocode-ROM **2032** uops bereit, die benötigt werden, um die Operation abzuschließen, wenn der Ablaufverfolgungszwischenspeicher **2030** auf eine komplexe Anweisung trifft.

[0193] In mindestens einer Ausführungsform können einige Befehle in eine einzelne Mikroop umgewandelt werden, wohingegen andere mehrere Mikroops benötigen, um den vollständigen Betrieb abzuschließen. In mindestens einer Ausführungsform kann der Anweisungsdekodierer **2028** auf den Mikrocode-ROM **2032** zugreifen, um eine Anweisung auszuführen, wenn mehr als vier Mikroops benötigt werden, um eine Anweisung abzuschließen. In mindestens einer Ausführungsform kann eine Anweisung in eine kleine Anzahl von Mikroops zur Verarbeitung am Anweisungsdekodierer **2028** dekodiert werden. In mindestens einer Ausführungsform kann eine Anweisung im Mikrocode-ROM **2032** gespeichert werden, falls eine Anzahl von Mikroops benötigt wird, um den Betrieb durchzuführen. In mindestens einer Ausführungsform bezieht sich der Ablaufverfolgungszwischenspeicher **2030** auf ein programmierbares Logikarray („PLA“) für den Eintrittspunkt, um einen korrekten Mikroanweisungsverweis zum Lesen von Mikrocodesequenzen zu bestimmen, um eine oder mehrere Anweisungen aus dem Mikrocode-ROM **2032** gemäß mindestens einer Ausführungsform abzuschließen. In mindestens einer Ausführungsform kann das Front-End **2001** der Maschine, nachdem der Mikrocode-ROM **2032** die Sequenzierung von Mikroops für eine Anweisung beendet hat, das Abrufen von Mikroops aus dem Ablaufverfolgungszwischenspeicher **2030** wieder aufnehmen.

[0194] In mindestens einer Ausführungsform kann die Engine zur Ausführung außerhalb der Reihenfolge („out-of-order-Engine“) **2003** Anweisungen zur Ausführung vorbereiten. In mindestens einer Ausführungsform weist die Logik für die Ausführung außerhalb der Reihenfolge eine Anzahl von Puffern auf, um den Ablauf von Anweisungen zu glätten und neu zu ordnen, um die Leistung zu optimieren, wenn sie in die Pipeline übergehen und für die Ausführung geplant werden. In mindestens einer Ausführungsform beinhaltet die Engine zur Ausführung außerhalb der Reihenfolge **2003** ohne Einschränkung einen Zuteiler/Registerumbenenner **2040**, eine Speicher-uop-Warteschlange **2042**, eine Ganzzahl-/Gleitkomma-uop-Warteschlange **2044**, einen Speicherplaner **2046**, einen schnellen Planer **2002**, einen langsamen/allgemeinen Gleitkomma-Planer („langsamer/allgemeiner FP-Planer“) **2004** und einen einfachen Gleitkomma-Planer („einfacher FP-Planer“) **2006**. In mindestens einer Ausführungsform werden der schnelle Planer **2002**, der langsame/allgemeine Gleitkomma-Planer **2004** und der einfache Gleitkomma-Planer **2006** in dieser Schrift auch gemeinsam als „uop-Planer **2002**, **2004**, **2006**“ bezeichnet. In mindestens einer Ausführungsform weist der Zuteiler/Registerumbenenner **2040** Maschinenpuffer und Ressourcen zu, die jede uop benötigt, um ausgeführt zu werden. In mindestens einer Ausführungsform benennt der Zuteiler/Registerumbenenner **2040** logische Register auf Einträge in einer Registerdatei um. In mindestens einer Ausführungsform weist der Zuteiler/Registerumbenenner **2040** auch einen Eintrag für jede uop in einer von zwei uop-Warteschlangen, der Speicher-uop-Warteschlange **2042** für Speicheroperationen und der Ganzzahl-/Fließkomma-uop-Warteschlange **2044** für Nicht-Speicheroperationen vor dem Speicherplaner **2046** und den uop-Planern **2002**, **2004**, **2006** zu. In mindestens einer Ausführungsform bestimmen die uop-Planer **2002**, **2004**, **2006** auf Grundlage der Bereitschaft ihrer abhängigen Eingangsregister-Operandenquellen und der Verfügbarkeit von Ausführungsressourcen, die uops benötigen, um ihre Operation abzuschließen, wann eine uop zur Ausführung bereit ist. In mindestens einer Ausführungsform kann der schnelle Planer **2002** von mindestens einer Ausführungsform auf jede Hälfte des Haupttaktzyklus planen, während der langsame/allgemeine Gleitkomma-Planer **2004** und der einfache Gleitkomma-Planer **2006** einmal pro Hauptprozessortaktzyklus planen können. In mindestens einer Ausführungsform vermitteln die uop-Planer **2002**, **2004**, **2006** für Versandports, um uops für die Ausführung zu planen.

[0195] In mindestens einer Ausführungsform beinhaltet der Ausführungsblock **2011** ohne Einschränkung eine Ganzzahlregisterdatei/ein Umgehungsnetzwerk **2008**, eine Gleitkommaregisterdatei/ein Umgehungsnetzwerk („FP-Registerdatei/Umgehungsnetzwerk“) **2010**, Adresserzeugungseinheiten (address generation units - „AGUs“) **2012** und **2014**, schnelle arithmetische Logikeinheiten (ALUs) („schnelle ALUs“) **2016** und **2018**, eine langsame arithmetische Logikeinheit („langsame ALU“) **2020**, eine Gleitkomma-ALU („FP“) **2022** und eine Gleitkommabewegungseinheit („FP-Bewegung“) **2024**. In mindestens einer Ausführungsform werden die Ganzzahlregisterdatei/das Umgehungsnetzwerk **2008** und Gleitkommaregisterdatei/Umgehungsnetzwerk **2010** in dieser Schrift auch als „Registerdateien **2008**, **2010**“ bezeichnet. In mindestens einer Ausführungsform werden AGUs **2012** und **2014**, schnelle ALUs **2016** und **2018**, die langsame ALU **2020**, die Gleitkomma-ALU **2022** und die Gleitkommabewegungseinheit **2024** in dieser Schrift auch als „Ausführungseinheiten **2012**, **2014**, **2016**, **2018**, **2020**, **2022** und **2024**“ bezeichnet. In mindestens einer Ausführungsform kann der Ausführungs-

block b11 ohne Einschränkung eine beliebige Anzahl (die Null beinhaltet) und eine beliebige Art von Registerdatei, Umgehungsnetzwerk, Adresserzeugungseinheit und Ausführungseinheit in beliebiger Kombination beinhalten.

[0196] In mindestens einer Ausführungsform können die Registerdateien **2008**, **2010** zwischen den uop-Planern **2002**, **2004**, **2006** und den Ausführungseinheiten **2012**, **2014**, **2016**, **2018**, **2020**, **2022** und **2024** angeordnet sein. In mindestens einer Ausführungsform führt das Ganzzahlregisterdatei/Umgehungsnetzwerk **2008** Ganzzahloperationen durch. In mindestens einer Ausführungsform führt das Gleitkommaregisterdatei/Umgehungsnetzwerk **2010** Gleitkommaoperationen durch. In mindestens einer Ausführungsform kann jede der Registerdateien **2008**, **2010** ohne Einschränkung ein Umgehungsnetzwerk beinhalten, das gerade abgeschlossene Ergebnisse, die noch nicht in die Registerdatei geschrieben wurden, an neue abhängige uops umleiten oder weiterleiten kann. In mindestens einer Ausführungsform können die Registerdateien **2008**, **2010** Daten miteinander austauschen. In mindestens einer Ausführungsform kann das Ganzzahlregisterdatei/Umgehungsnetzwerk **2008** ohne Einschränkung zwei getrennte Registerdateien beinhalten, eine Registerdatei für niederwertige zweiunddreißig Datenbits und eine zweite Registerdatei für höherwertige zweiunddreißig Datenbits. In mindestens einer Ausführungsform kann das Gleitkommaregisterdatei/Umgehungsnetzwerk **2010** ohne Einschränkung 128 Bit breite Einträge beinhalten, da Gleitkommabefehle typischerweise Operanden von 64 bis 128 Bit Breite aufweisen.

[0197] In mindestens einer Ausführungsform können die Ausführungseinheiten **2012**, **2014**, **2016**, **2018**, **2020**, **2022**, **2024** Anweisungen ausführen. In mindestens einer Ausführungsform speichern die Registerdateien **2008**, **2010** Ganzzahl- und Gleitkomma-Datenoperandenwerte, die Mikroanweisungen ausführen müssen. In mindestens einer Ausführungsform kann der Prozessor **2000** ohne Einschränkung eine beliebige Anzahl und Kombination von Ausführungseinheiten **2012**, **2014**, **2016**, **2018**, **2020**, **2022**, **2024** beinhalten. In mindestens einer Ausführungsform können die Gleitkomma-ALU **2022** und die Gleitkomma-Bewegungseinheit **2024** Gleitkomma-, MMX-, SIMD-, AVX- und SSE-Operationen oder andere Operationen ausführen, was spezialisierte Anweisungen zum maschinellen Lernen beinhaltet. In mindestens einer Ausführungsform kann die Gleitkomma-ALU **2022** ohne Einschränkung einen 64-Bit-mal-64-Bit-Gleitkommateiler beinhalten, um Divisions-, Quadratwurzel- und Rest-Mikroops auszuführen. In mindestens einer Ausführungsform können Anweisungen, die einen Gleitkommawert einschließen, mit Gleitkomma-Hardware gehandhabt werden. In mindestens einer Ausführungsform können ALU-Operationen an schnelle ALUs **2016**, **2018** weitergegeben werden. In mindestens einer Ausführungsform können schnelle ALUs **2016**, **2018** schnelle Operationen mit einer effektiven Latenz von einem halben Taktzyklus ausführen. In mindestens einer Ausführungsform gehen die meisten komplexen Ganzzahloperationen zur langsamen ALU **2020**, da die langsame ALU **2020** ohne Einschränkung Ganzzahlausführungshardware für Operationen mit langer Latenzzeit beinhalten kann, wie etwa eine Multiplikation, Verschiebungen, Kennzeichenlogik und Verzweigungsverarbeitung. In mindestens einer Ausführungsform können Speicherlade-/-speicheroperationen von AGUS **2012**, **2014** ausgeführt werden. In mindestens einer Ausführungsform können die schnelle ALU **2016**, die schnelle ALU **2018** und die langsame ALU **2020** Ganzzahloperationen an 64-Bit-Datenoperanden ausführen. In mindestens eine Ausführungsform können die schnelle ALU **2016**, die schnelle ALU **2018** und die langsame ALU **2020** umgesetzt sein, um eine Reihe von Datenbitgrößen zu unterstützen, die sechzehn, zweiunddreißig, 128, 256, usw. beinhalten. In mindestens einer Ausführungsform können die Gleitkomma-ALU **2022** und die Gleitkomma-Bewegungseinheit **2024** umgesetzt sein, um einen Bereich von Operanden mit Bits verschiedener Breiten zu unterstützen. In mindestens einer Ausführungsform können die Gleitkomma-ALU **2022** und die Gleitkomma-Bewegungseinheit **2024** an 128 Bit breiten gepackten Datenoperanden in Verbindung mit SIMD- und Multimedia-Anweisungen arbeiten.

[0198] In mindestens einer Ausführungsform verteilen die uop-Planer **2002**, **2004**, **2006** abhängige Operationen, bevor die Ausführung der übergeordneten Last beendet ist. In mindestens einer Ausführungsform kann der Prozessor **2000**, da uops spekulativ geplant und im Prozessor **2000** ausgeführt werden können, auch eine Logik beinhalten, um Speicherfehler zu handhaben. In mindestens einer Ausführungsform kann es, wenn ein Datenladen im Datenzwischenspeicher fehlschlägt, abhängige Operationen im Flug in der Pipeline geben, die den Planer mit vorübergehend falschen Daten verlassen haben. In mindestens einer Ausführungsform verfolgt ein Wiedergabemechanismus Anweisungen, die falsche Daten verwenden, und führt sie erneut aus. In mindestens einer Ausführungsform müssen abhängige Operationen möglicherweise wiederholt werden und es unabhängige wird möglicherweise ermöglicht, dass sie abgeschlossen werden. In mindestens einer Ausführungsform können die Planer und der Wiedergabemechanismus mindestens einer Ausführungsform eines Prozessors auch so ausgelegt sein, dass sie Anweisungssequenzen für Zeichenkettenvergleichsoperationen abfangen.

[0199] In mindestens einer Ausführungsform kann sich der Ausdruck „Register“ auf Speicherorte des integrierten Prozessors beziehen, die als Teil von Anweisungen verwendet werden können, um Operanden zu identifizieren. In mindestens einer Ausführungsform können Register derartige sein, die von außerhalb des Prozessors (aus der Perspektive eines Programmierers) verwendet werden können. In mindestens einer Ausführungsform sind Register möglicherweise nicht auf eine bestimmte Schaltungsart beschränkt. Vielmehr kann ein Register in mindestens einer Ausführungsform Daten speichern, Daten bereitstellen und in dieser Schrift beschriebene Funktionen ausführen. In mindestens einer Ausführungsform können in dieser Schrift beschriebene Register durch eine Schaltung innerhalb eines Prozessors unter Verwendung einer beliebigen Anzahl verschiedener Methoden umgesetzt sein, wie etwa dedizierte physische Register, dynamisch zugewiesene physische Register unter Verwendung von Registerumbenennung, Kombinationen von dedizierten und dynamisch zugewiesenen physischen Registern usw. In mindestens einer Ausführungsform speichern Ganzzahlregister 32-Bit-Ganzzahldaten. Eine Registerdatei von mindestens einer Ausführungsform enthält außerdem acht Multimedia-SIMD-Register für gepackte Daten.

[0200] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. Details bezüglich der Ableitungs- und/oder Trainingslogik **615** werden nachfolgend in Verbindung mit den **Fig. 6A** und/oder 6B bereitgestellt. In mindestens einer Ausführungsform können Teile der oder die gesamte Ableitungs- und/oder Trainingslogik **615** in den Ausführungsblock **2011** und andere gezeigte oder nichtgezeigte Speicher oder Register integriert sein. Zum Beispiel können in mindestens einer Ausführungsform in dieser Schrift beschriebene Trainings- und/oder Ableitungsmethoden eine oder mehrere der im Ausführungsblock **2011** veranschaulichten ALUs verwenden. Darüber hinaus können Gewichtungsparameter in chipinternen oder chipexternen Speicher und/oder Registern (gezeigt oder nichtgezeigt) gespeichert werden, die ALUs des Ausführungsblocks **2011** konfigurieren, um einen oder mehrere Algorithmen maschinellen Lernens, neuronale Netzwerkarchitekturen, Anwendungsfälle oder in dieser Schrift beschriebene Trainingsmethoden auszuführen.

[0201] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. In mindestens einer Ausführungsform kann diese Logik mit Komponenten dieser Figuren verwendet werden, um ein dreidimensionales Bild einer bestimmten Ansicht aus zweidimensionalen Bildern aus einer willkürlichen Anzahl von Kameras zu erzeugen.

[0202] **Fig. 21** veranschaulicht einen Deep-Learning-Anwendungsprozessor **2100** gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform verwendet der Deep-Learning-Anwendungsprozessor **2100** Anweisungen, die bei Ausführung durch den Deep-Learning-Anwendungsprozessor **2100** bewirken, dass der Deep-Learning-Anwendungsprozessor **2100** einige oder alle der in dieser Offenbarung beschriebenen Prozesse und Methoden ausführt. In mindestens einer Ausführungsform ist der Deep-Learning-Anwendungsprozessor **2100** eine anwendungsspezifische integrierte Schaltung (ASIC). In mindestens einer Ausführungsform führt der Anwendungsprozessor **2100** Matrixmultiplikationsoperationen durch, die entweder als Ergebnis der Ausführung einer oder mehrerer Anweisungen oder beider in Hardware „festverdrahtet“ sind. In mindestens einer Ausführungsform beinhaltet der Deep-Learning-Anwendungsprozessor **2100** ohne Einschränkung Verarbeitungscluster **2110(1)-2110(12)**, Inter-Chip-Links („ICLs“) **2120(1)-2120(12)**, Inter-Chip-Steuerungen (Inter-Chip Controllers - „ICCs“) **2130(1)-2130(2)**, Speichersteuerungen (memory controllers - „Mem Ctrlrs“) **2142(1)-2142(4)**, eine physische Speicherschicht mit hoher Bandbreite (high bandwidth memory physical layer - „HBM PHY“) **2144(1)-2144(4)**, eine Verwaltungssteuerungszentralverarbeitungseinheit („Verwaltungssteuerungs-CPU“) **2150**, eine Steuerung zur schnellen Verbindung von Peripheriegerätekomponenten und einen Block zum direkten Speicherzugriff (peripheral component interconnect express controller and direct memory access block - „PCIe Controller und DMA“) **2170** und einen sechzehnspuriges Anschluss zur schnellen Verbindung von Peripheriegerätekomponenten (peripheral component interconnect express port - „PCI Express x 16“) **2180**.

[0203] In mindestens einer Ausführungsform können Verarbeitungscluster **2110** Deep-Learning-Operationen, was Ableitungs- oder Vorhersageoperationen beinhaltet, auf Grundlage von Gewichtungsparametern durchführen, die eine oder mehrere Trainingsmethoden berechnet haben, was die in dieser Schrift beschriebenen beinhaltet. In mindestens einer Ausführungsform kann jeder Verarbeitungscluster **2110** ohne Einschränkung eine beliebige Anzahl und Art von Prozessoren beinhalten. In mindestens einer Ausführungsform kann der Deep-Learning-Anwendungsprozessor **2100** eine beliebige Anzahl und Art von Verarbeitungsclustem **2100** beinhalten. In mindestens einer Ausführungsform sind die Inter-Chip-Links **2120** bidirektional. In mindestens einer Ausführungsform ermöglichen die Inter-Chip-Links **2120** und die Inter-Chip-Steuerung **2130** mehreren Deep-Learning-Anwendungsprozessoren **2100** den Austausch von Informationen, die Aktivierungsinformatio-

nen beinhalten, die sich aus der Ausführung eines oder mehrerer Algorithmen des maschinellen Lernens ergeben, die in einem oder mehreren neuronalen Netzwerken verkörpert sind. In mindestens einer Ausführungsform kann der Deep-Learning-Anwendungsprozessor **2100** eine beliebige Anzahl (die Null beinhaltet) und Art von ICLs **2120** und ICCs **2130** beinhalten.

[0204] In mindestens einer Ausführungsform stellen HBM2s **2140** insgesamt 32 Gigabyte (GB) Speicher bereit. HBM2 **2140(i)** ist sowohl der Speichersteuerung **2142(i)** als auch HBM PHY **2144(i)** zugeordnet. In mindestens einer Ausführungsform kann eine beliebige Anzahl von HBM2s **2140** eine beliebige Art und Gesamtmenge von Speicher mit hoher Bandbreite bereitstellen und kann einer beliebigen Anzahl (die Null beinhaltet) und Art von Speichersteuerungen **2142** und HBM-PHYs **2144** zugeordnet sein. In mindestens einer Ausführungsform können SPI, I2C, GPIO **2160**, PCIe-Steuerung und DMA **2170** und/oder PCIe **2180** durch eine beliebige Anzahl und Art von Blöcken ersetzt werden, die eine beliebige Anzahl und Art von Kommunikationsstandards auf eine beliebige technisch machbare Weise ermöglichen.

[0205] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. Details bezüglich der Ableitungs- und/oder Trainingslogik **615** werden nachfolgend in Verbindung mit den **Fig. 6A** und/oder 6B bereitgestellt. In mindestens einer Ausführungsform wird der Deep-Learning-Anwendungsprozessor **2100** verwendet, um ein Modell maschinellen Lernens, wie etwa ein neuronales Netzwerk, zu trainieren, um dem Deep-Learning-Anwendungsprozessor **2100** bereitgestellte Informationen vorherzusagen oder abzuleiten. In mindestens einer Ausführungsform wird der Deep-Learning-Anwendungsprozessor **2100** verwendet, um Informationen auf Grundlage eines trainierten Modells maschinellen Lernens (z. B. neuronales Netzwerk) abzuleiten oder vorherzusagen, das von einem anderen Prozessor oder System oder von dem Deep-Learning-Anwendungsprozessor **2100** trainiert wurde. In mindestens einer Ausführungsform kann der Prozessor **2100** verwendet werden, um einen oder mehrere in dieser Schrift beschriebene Anwendungsfälle für neuronale Netzwerke durchzuführen.

[0206] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. In mindestens einer Ausführungsform kann diese Logik mit Komponenten dieser Figuren verwendet werden, um ein dreidimensionales Bild einer bestimmten Ansicht aus zweidimensionalen Bildern aus einer willkürlichen Anzahl von Kameras zu erzeugen.

[0207] **Fig. 22** ist ein Blockdiagramm eines neuromorphen Prozessors **2200** gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform kann der neuromorphe Prozessor **2200** eine oder mehrere Eingaben von Quellen außerhalb des neuromorphen Prozessors **2200** empfangen. In mindestens einer Ausführungsform können diese Eingaben an ein oder mehrere Neuronen **2202** innerhalb des neuromorphen Prozessors **2200** übermittelt werden. In mindestens einer Ausführungsform können Neuronen **2202** und Komponenten davon unter Verwendung von Schaltungen oder Logik umgesetzt sein, die eine oder mehrere arithmetische Logikeinheiten (ALUs) beinhalten. In mindestens einer Ausführungsform kann der neuromorphe Prozessor **2200** ohne Einschränkung Tausende oder Millionen von Instanzen von Neuronen **2202** beinhalten, aber es kann eine beliebige geeignete Anzahl von Neuronen **2202** verwendet werden. In mindestens einer Ausführungsform kann jede Instanz des Neurons **2202** einen Neuroneneingang **2204** und einen Neuronenausgang **2206** beinhalten. In mindestens einer Ausführungsform können die Neuronen **2202** Ausgaben erzeugen, die an Eingänge anderer Instanzen von Neuronen **2202** übermittelt werden können. Zum Beispiel können in mindestens einer Ausführungsform Neuroneneingänge **2204** und Neuronenausgänge **2206** über Synapsen **2208** miteinander verbunden sein.

[0208] In mindestens einer Ausführungsform können die Neuronen **2202** und die Synapsen **2208** derartig miteinander verbunden sein, dass der neuromorphe Prozessor **2200** betrieben wird, um von dem neuromorphen Prozessor **2200** empfangene Informationen zu verarbeiten oder zu analysieren. In mindestens einer Ausführungsform können die Neuronen **2202** einen Ausgangsimpuls (oder „Feuer“ oder „Spitze“) übermitteln, wenn durch den Neuroneneingang **2204** empfangene Eingaben einen Schwellenwert überschreiten. In mindestens einer Ausführungsform können die Neuronen **2202** an den Neuroneneingängen **2204** empfangene Signale summieren oder integrieren. Zum Beispiel können die Neuronen **2202** in mindestens einer Ausführungsform als undichte integrate-and-fire-Neuronen umgesetzt sein, wobei, wenn eine Summe (als ein „Membranpotential“ bezeichnet) einen Schwellenwert überschreitet, das Neuron **2202** eine Ausgabe (oder „Feuer“) unter Verwendung einer Übertragungsfunktion überschreitet, wie etwa einer Sigmoid- oder Schwellenfunktion. In mindestens einer Ausführungsform kann ein undichtes integrate-and-fire-Neuron Signale, die an Neuroneneingängen **2204** empfangen werden, zu einem Membranpotential summieren und kann auch einen Abklingfaktor

(oder Leck) anwenden, um ein Membranpotential zu verringern. In mindestens einer Ausführungsform kann ein undichtes integrate-and-fire-Neuron feuern, wenn mehrere Eingangssignale an den Neuroneneingängen **2204** schnell genug empfangen werden, um einen Schwellenwert zu überschreiten (d. h. bevor ein Membranpotential zu weit abfällt, um zu feuern). In mindestens einer Ausführungsform können die Neuronen **2202** unter Verwendung von Schaltungen oder Logik umgesetzt sein, die Eingaben empfangen, Eingaben in ein Membranpotential integrieren und ein Membranpotential abklingen lassen. In mindestens einer Ausführungsform können Eingaben gemittelt werden oder es kann eine beliebige andere geeignete Übertragungsfunktion verwendet werden. Darüber hinaus können die Neuronen **2202** in mindestens einer Ausführungsform ohne Einschränkung Vergleicherschaltungen oder Logik beinhalten, die eine Ausgangsspitze am Neuronenausgang **2206** erzeugen, wenn das Ergebnis des Anwendens einer Übertragungsfunktion auf den Neuroneneingang **2204** einen Schwellenwert überschreitet. In mindestens einer Ausführungsform kann das Neuron **2202**, sobald es feuert, zuvor empfangene Eingabeinformationen ignorieren, indem es zum Beispiel ein Membranpotential auf 0 oder einen anderen geeigneten Standardwert zurücksetzt. In mindestens einer Ausführungsform kann das Neuron **2202** nach einem geeigneten Zeitraum (oder Refraktärzeitraum) den normalen Betrieb wieder aufnehmen, sobald das Membranpotential auf 0 zurückgesetzt ist.

[0209] In mindestens einer Ausführungsform können die Neuronen **2202** durch Synapsen **2208** miteinander verbunden sein. In mindestens einer Ausführungsform können die Synapsen **2208** arbeiten, um Signale von einem Ausgang eines ersten Neurons **2202** an einen Eingang eines zweiten Neurons **2202** zu übermitteln. In mindestens einer Ausführungsform können die Neuronen **2202** Informationen über mehr als eine Instanz der Synapse **2208** übermitteln. In mindestens einer Ausführungsform können eine oder mehrere Instanzen des Neuronenausgangs **2206** über eine Instanz der Synapse **2208** mit einer Instanz des Neuroneneingangs **2204** in demselben Neuron **2202** verbunden sein. In mindestens einer Ausführungsform kann eine Instanz des Neurons **2202**, die eine über eine Instanz der Synapse **2208** zu übermittelnde Ausgabe erzeugt, in Bezug auf diese Instanz der Synapse **2208** als ein „präsynaptisches Neuron“ bezeichnet werden. In mindestens einer Ausführungsform kann eine Instanz des Neurons **2202**, die eine über eine Instanz der Synapse **2208** übermittelte Eingabe empfängt, in Bezug auf diese Instanz der Synapse **2208** als ein „postsynaptisches Neuron“ bezeichnet werden. Da eine Instanz des Neurons **2202** Eingaben von einer oder mehreren Instanzen der Synapse **2208** empfangen kann und auch Ausgaben über eine oder mehrere Instanzen der Synapse **2208** übermitteln kann, kann eine einzelne Instanz des Neurons **2202** daher sowohl ein „präsynaptisches Neuron“ als auch sein „postsynaptisches Neuron“ in Bezug auf verschiedene Instanzen von Synapsen **2208** in mindestens einer Ausführungsform sein.

[0210] In mindestens einer Ausführungsform können die Neuronen **2202** in einer oder mehreren Schichten organisiert sein. Jede Instanz des Neurons **2202** kann einen Neuronenausgang **2206** aufweisen, der durch eine oder mehrere Synapsen **2208** zu einem oder mehreren Neuroneneingängen **2204** auffächern kann. In mindestens einer Ausführungsform können Neuronenausgänge **2206** von Neuronen **2202** in einer ersten Schicht **2210** mit Neuroneneingängen **2204** von Neuronen **2202** in einer zweiten Schicht **2212** verbunden sein. In mindestens einer Ausführungsform kann die Schicht **2210** als eine „vorwärtsgekoppelte Schicht“ bezeichnet werden. In mindestens einer Ausführungsform kann jede Instanz des Neurons **2202** in einer Instanz der ersten Schicht **2210** zu jeder Instanz des Neurons **2202** in der zweiten Schicht **2212** auffächern. In mindestens einer Ausführungsform kann die erste Schicht **2210** als eine „vollständig verbundene vorwärtsgekoppelte Schicht“ bezeichnet werden. In mindestens einer Ausführungsform kann jede Instanz des Neurons **2202** in einer Instanz der zweiten Schicht **2212** auf weniger als alle Instanzen des Neurons **2202** in einer dritten Schicht **2214** auffächern. In mindestens einer Ausführungsform kann die zweite Schicht **2212** als „kaum verbundene vorwärtsgekoppelte Schicht“ bezeichnet sein. In mindestens einer Ausführungsform können sich Neuronen **2202** in der zweiten Schicht **2212** zu Neuronen **2202** in mehreren anderen Schichten auffächern, was zu Neuronen **2202** in (derselben) zweiten Schicht **2212** beinhaltet. In mindestens einer Ausführungsform kann die zweite Schicht **2212** als „wiederkehrende Schicht“ bezeichnet sein. In mindestens einer Ausführungsform kann der neuromorphe Prozessor **2200** ohne Einschränkung eine beliebige geeignete Kombination von wiederkehrenden Schichten und vorwärtsgekoppelten Schichten beinhalten, die ohne Einschränkung sowohl kaum verbundene vorwärtsgekoppelte Schichten als auch vollständig verbundene vorwärtsgekoppelte Schichten beinhalten.

[0211] In mindestens einer Ausführungsform kann der neuromorphe Prozessor **2200** ohne Einschränkung eine rekonfigurierbare Verbindungsarchitektur oder dedizierte fest verdrahtete Verbindungen beinhalten, um die Synapse **2208** mit den Neuronen **2202** zu verbinden. In mindestens einer Ausführungsform kann der neuromorphe Prozessor **2200** ohne Einschränkung eine Schaltung oder Logik beinhalten, die es Synapsen ermöglicht, unterschiedlichen Neuronen **2202** nach Bedarf auf Grundlage der Topologie des neuronalen Netzwerks und dem Eingangs-/Ausgangslastfaktor von Neuronen zugewiesen zu werden. Zum Beispiel können die Synapsen **2208** in mindestens einer Ausführungsform mit Neuronen **2202** unter Verwendung einer Verbindungs-

struktur, wie etwa Network-on-Chip, oder mit dedizierten Verbindungen verbunden sein. In mindestens einer Ausführungsform können Synapsenverbindungen und Komponenten davon unter Verwendung von Schaltungen oder Logik umgesetzt sein.

[0212] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. In mindestens einer Ausführungsform kann diese Logik mit Komponenten dieser Figuren verwendet werden, um ein dreidimensionales Bild einer bestimmten Ansicht aus zweidimensionalen Bildern aus einer willkürlichen Anzahl von Kameras zu erzeugen.

[0213] **Fig. 23** ist ein Blockdiagramm eines Verarbeitungssystems gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform beinhaltet das System **2300** einen oder mehrere Prozessoren **2302** und einen oder mehrere Grafikprozessoren **2308** und kann ein Einzelprozessor-Desktop-System, ein Mehrprozessor-Workstation-System oder ein Server-System sein, dass eine große Anzahl von Prozessoren **2302** oder Prozessorkernen **2307** aufweist. In mindestens einer Ausführungsform ist das System **2300** eine Verarbeitungsplattform, die in eine integrierte System-on-a-Chip-(SoC)-Schaltung zur Verwendung in mobilen, tragbaren oder eingebetteten Vorrichtungen integriert ist.

[0214] In mindestens einer Ausführungsform kann das System **2300** eine serverbasierte Spielplattform, eine Spielkonsole, die eine Spiel- und Medienkonsole beinhaltet, eine mobile Spielkonsole, eine Handheld-Spielekonsole oder eine Online-Spielekonsole beinhalten oder darin integriert sein. In mindestens einer Ausführungsform ist das System **2300** ein Mobiltelefon, ein Smartphone, eine Tablet-Rechenvorrichtung oder eine mobile Internetvorrichtung. In mindestens einer Ausführungsform kann das Verarbeitungssystem **2300** auch eine tragbare Vorrichtung, wie etwa eine tragbare Smartwatch-Vorrichtung, eine intelligente Brillenvorrichtung, eine Augmented-Reality-Vorrichtung oder eine Virtual-Reality-Vorrichtung beinhalten, mit dieser gekoppelt oder darin integriert sein. In mindestens einer Ausführungsform ist das Verarbeitungssystem **2300** ein Fernsehgerät oder eine Set-Top-Box-Vorrichtung, die einen oder mehrere Prozessoren **2302** beinhaltet, und eine grafische Schnittstelle, die von einem oder mehreren Grafikprozessoren **2308** erzeugt wird.

[0215] In mindestens einer Ausführungsform beinhalten ein oder mehrere Prozessoren **2302** jeweils einen oder mehrere Prozessorkerne **2307**, um Anweisungen zu verarbeiten, die, wenn sie ausgeführt werden, Operationen für System- und Benutzersoftware ausführen. In mindestens einer Ausführungsform ist jeder von einem oder mehreren Prozessorkernen **2307** dazu konfiguriert, einen konkreten Anweisungssatz **2309** zu verarbeiten. In mindestens einer Ausführungsform kann der Anweisungssatz **2309** das Berechnen mit komplexem Anweisungssatz (Complex Instruction Set Computing - CISC), das Berechnen mit verringertem Anweisungssatz (Reduced Instruction Set Computing - RISC) oder das Berechnen über ein sehr langes Anweisungswort (Very Long Instruction Word - VLIW) erleichtern. In mindestens einer Ausführungsform können die Prozessorkerne **2307** jeweils einen anderen Anweisungssatz **2309** verarbeiten, der Anweisungen beinhalten kann, um die Emulation anderer Anweisungssätze zu erleichtern. In mindestens einer Ausführungsform kann der Prozessorkern **2307** auch andere Verarbeitungsvorrichtungen beinhalten, wie etwa einen digitalen Signalprozessor (DSP).

[0216] In mindestens einer Ausführungsform beinhaltet der Prozessor **2302** einen schnellen Pufferspeicher **2304**. In mindestens einer Ausführungsform kann der Prozessor **2302** einen einzelnen internen Zwischenspeicher oder mehrere Ebenen des internen Zwischenspeichers aufweisen. In mindestens einer Ausführungsform wird der schnelle Pufferspeicher von verschiedenen Komponenten des Prozessors **2302** gemeinsam genutzt. In mindestens einer Ausführungsform verwendet der Prozessor **2302** außerdem einen externen Zwischenspeicher (z. B. einen Level-3-(L3-)Zwischenspeicher oder Last-Level-Zwischenspeicher (LLC)) (nicht gezeigt), der von den Prozessorkernen **2307** unter Verwendung bekannter Zwischenspeicher-Kohärenzmethoden gemeinsam genutzt werden kann. In mindestens einer Ausführungsform ist die Registerdatei **2306** zusätzlich im Prozessor **2302** beinhaltet, der unterschiedliche Arten von Registern zum Speichern verschiedener Arten von Daten (z. B. Ganzzahlregister, Gleitkommaregister, Statusregister und ein Anweisungsverweisregister) beinhalten kann. In mindestens einer Ausführungsform kann die Registerdatei **2306** Allzweckregister oder andere Register beinhalten.

[0217] In mindestens einer Ausführungsform sind ein oder mehrere Prozessoren **2302** mit einem oder mehreren Schnittstellenbussen **2310** gekoppelt, um Kommunikationssignale, wie etwa Adress-, Daten- oder Steuersignale, zwischen dem Prozessor **2302** und anderen Komponenten im System **2300** zu übermitteln. In mindestens einer Ausführungsform kann der Schnittstellenbus **2310** in einer Ausführungsform ein Prozessorbus sein, wie etwa eine Version eines Mediendirektsschnittstellen-(Direct Media Interface - DMI-)Busses. In min-

destens einer Ausführungsform ist die Schnittstelle **2310** nicht auf einen DMI-Bus beschränkt und kann einen oder mehrere Peripheriegerätekomponentenverbindungsbusse (z. B. PCI, PCI Express), Speicherbusse oder andere Arten von Schnittstellenbussen beinhalten. In mindestens einer Ausführungsform beinhalten Prozessor(en) **2302** eine integrierte Speichersteuerung **2316** und einen Plattformsteuerungshub **2330**. In mindestens einer Ausführungsform erleichtert der Speichersteuerung **2316** die Kommunikation zwischen einer Speichervorrichtung und anderen Komponenten des Systems **2300**, während der Plattformsteuerungshub (platform controller hub - PCH) **2330** Verbindungen zu E/A-Vorrichtungen über einen lokalen E/A-Bus bereitstellt.

[0218] In mindestens einer Ausführungsform kann die Speichervorrichtung **2320** ein dynamischer Direktzugriffsspeicher (DRAM), ein statischer Direktzugriffsspeicher (SRAM), eine Flash-Speichervorrichtung, eine Phasenänderungsspeichervorrichtung oder eine andere Speichervorrichtung sein, die eine geeignete Leistung aufweist, um als Prozessspeicher zu dienen. In mindestens einer Ausführungsform kann die Speichervorrichtung **2320** als Systemspeicher für das System **2300** arbeiten, um Daten **2322** und Anweisungen **2321** zur Verwendung zu speichern, wenn ein oder mehrere Prozessoren **2302** eine Anwendung oder einen Prozess ausführen. In mindestens einer Ausführungsform ist die Speichersteuerung **2316** auch mit einem optionalen externen Grafikprozessor **2312** gekoppelt, der mit einem oder mehreren Grafikprozessoren **2308** in den Prozessoren **2302** kommunizieren kann, um Grafik- und Medienoperationen durchzuführen. In mindestens einer Ausführungsform kann eine Anzeigevorrichtung **2311** mit Prozessor(en) **2302** verbunden sein. In mindestens einer Ausführungsform kann die Anzeigevorrichtung **2311** eine oder mehrere von einer internen Anzeigevorrichtung, wie in einer mobilen elektronischen Vorrichtung oder einer Laptopvorrichtung, oder einer externen Anzeigevorrichtung beinhalten, die über eine Anzeigeschnittstelle (z. B. DisplayPort usw.) angeschlossen ist. In mindestens einer Ausführungsform kann die Anzeigevorrichtung **2311** eine am Kopf befestigte Anzeige (head mounted display - HMD) beinhalten, wie etwa eine stereoskopische Anzeigevorrichtung zur Verwendung in Virtual-Reality-(VR-)Anwendungen oder Augmented-Reality-(AR-)Anwendungen.

[0219] In mindestens einer Ausführungsform ermöglicht der Plattformsteuerungshub **2330** den Peripheriegeräten, sich über einen Hochgeschwindigkeits-E/A-Bus mit der Speichervorrichtung **2320** und dem Prozessor **2302** zu verbinden. In mindestens einer Ausführungsform beinhalten E/A-Peripheriegeräte, ohne darauf beschränkt zu sein, eine Audiosteuerung **2346**, eine Netzwerksteuerung **2334**, eine Firmwareschnittstelle **2328**, einen drahtlosen Sendeempfänger **2326**, Berührungssensoren **2325**, eine Datenspeichervorrichtung **2324** (z.B. Festplatte, Flash-Speicher usw.). In mindestens einer Ausführungsform kann sich die Datenspeichervorrichtung **2324** über eine Speicherschnittstelle (z. B. SATA) oder über einen Peripheriegerätebus verbinden, wie etwa einen Peripheriegerätekomponentenverbindungsbus (z. B. PCI, PCI Express). In mindestens einer Ausführungsform können die Berührungssensoren **2325** Berührungsbildschirmsensoren, Drucksensoren oder Fingerabdrucksensoren beinhalten. In mindestens einer Ausführungsform kann der drahtlose Sendeempfänger **2326** ein Wi-Fi-Sendeempfänger, ein Bluetooth-Sendeempfänger oder ein Mobilfunknetz-Sendeempfänger wie etwa ein 3G-, 4G- oder Long Term Evolution (LTE)-Sendeempfänger sein. In mindestens einer Ausführungsform ermöglicht die Firmwareschnittstelle **2328** die Kommunikation mit der System-Firmware und kann zum Beispiel eine einheitliche erweiterbare Firmwareschnittstelle (unified extensible firmware interface - UEFI) sein. In mindestens einer Ausführungsform kann der Netzwerksteuerung **2334** eine Netzwerkverbindung zu einem drahtgebundenen Netzwerk ermöglichen. In mindestens einer Ausführungsform ist eine Hochleistungsnetzwerksteuerung (nicht gezeigt) mit dem Schnittstellenbus **2310** gekoppelt. In mindestens einer Ausführungsform ist die Audiosteuerung **2346** eine Mehrkanal-Audiosteuerung mit hoher Auflösung. In mindestens einer Ausführungsform beinhaltet das System **2300** eine optionale Alt-E/A-Steuerung **2340** zum Koppeln von Alt-Vorrichtungen (z. B. Personal System 2 (PS/2)) an das System. In mindestens einer Ausführungsform kann der Plattformsteuerungshub **2330** auch mit einer oder mehreren universellen seriellen Bus-(USB-)Steuerungen **2342** verbunden sein, um Eingabevorrichtungen zu verbinden, wie etwa Kombinationen aus Tastatur und Maus **2343**, eine Kamera **2344** oder andere USB-Eingabevorrichtungen.

[0220] In mindestens einer Ausführungsform kann eine Instanz der Speichersteuerung **2316** und des Plattformsteuerungshubs **2330** in einen diskreten externen Grafikprozessor, wie etwa den externen Grafikprozessor **2312**, integriert sein. In mindestens einer Ausführungsform können sich der Plattformsteuerungshub **2330** und/oder die Speichersteuerung **2316** außerhalb eines oder mehrerer Prozessoren **2302** befinden. Zum Beispiel kann das System **2300** in mindestens einer Ausführungsform eine externe Speichersteuerung **2316** und einen Plattformsteuerungshub **2330** beinhalten, die als Speichersteuerungshub und Peripheriegerätesteuerungshub innerhalb eines Systemchipsatzes konfiguriert sein können, der mit Prozessor(en) **2302** kommuniziert.

[0221] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. Details bezüglich der Ableitungs- und/oder Trainingslogik **615** werden nachfolgend in Verbindung mit den **Fig. 6A** und/oder 6B bereit-

gestellt. In mindestens einer Ausführungsform können Teile der oder die gesamte Ableitungs- und/oder Trainingslogik **615** in den Grafikprozessor **2300** integriert sein. Zum Beispiel können in mindestens einer Ausführungsform in dieser Schrift beschriebene Trainings- und/oder Ableitungsmethoden eine oder mehrere der im Grafikprozessor **2312** realisierten ALUs verwenden. Darüber hinaus können in mindestens einer Ausführungsform in dieser Schrift beschriebene Ableitung- und/oder Trainingsoperationen unter Verwendung einer anderen Logik als der in den **Fig. 6A** oder **Fig. 6B** veranschaulichten Logik durchgeführt werden. In mindestens einer Ausführungsform können Gewichtungsparameter in chipinternen oder chipexternen Speicher und/oder Registern (gezeigt oder nichtgezeigt) gespeichert werden, die ALUs des Grafikprozessors **2300** konfigurieren, um einen oder mehrere Algorithmen maschinellen Lernens, neuronale Netzwerkarchitekturen, Anwendungsfälle oder in dieser Schrift beschriebene Trainingsmethoden auszuführen.

[0222] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. In mindestens einer Ausführungsform kann diese Logik mit Komponenten dieser Figuren verwendet werden, um ein dreidimensionales Bild einer bestimmten Ansicht aus zweidimensionalen Bildern aus einer willkürlichen Anzahl von Kameras zu erzeugen.

[0223] Fig. 24 ist ein Blockdiagramm eines Prozessors **2400**, der einen oder mehrere Prozessorkerne **2402A - 2402N**, eine integrierte Speichersteuerung **2414** und einen integrierten Grafikprozessor **2408** gemäß mindestens einer Ausführungsform aufweist. In mindestens einer Ausführungsform kann der Prozessor **2400** zusätzliche Kerne beinhalten und den zusätzlichen Kern **2402N** beinhalten, der durch gestrichelte Kästen dargestellt ist. In mindestens einer Ausführungsform beinhaltet jeder der Prozessorkerne **2402A-2402N** eine oder mehrere Einheiten des internen Zwischenspeichers **2404A-2404N**. In mindestens einer Ausführungsform hat jeder Prozessorkern auch Zugriff auf eine oder mehrere gemeinsam genutzte zwischengespeicherte Einheiten **2406**.

[0224] In mindestens einer Ausführungsform stellen die Einheiten des internen Zwischenspeichers **2404A-2404N** und die Einheiten des gemeinsam genutzten Zwischenspeichers **2406** eine Zwischenspeicherspeicherhierarchie innerhalb des Prozessors **2400** dar. In mindestens einer Ausführungsform können die Einheiten des schnellen Pufferspeichers **2404A-2404N** mindestens eine Ebene des Anweisungs- und Datenzwischenspeichers innerhalb jedes Prozessorkerns und eine oder mehrere Ebenen eines gemeinsam genutzten Zwischenspeichers der mittleren Ebene beinhalten, wie etwa eine Ebene 2 (L2), Ebene 3 (L3), Ebene 4 (L4) oder andere Zwischenspeicherebenen, wobei die höchste Zwischenspeicherebene vor dem externen Speicher als eine LLC klassifiziert ist. In mindestens einer Ausführungsform hält die Zwischenspeicher-Kohärenzlogik die Kohärenz zwischen verschiedenen Zwischenspeichereinheiten **2406** und **2404A-2404N** aufrecht.

[0225] In mindestens einer Ausführungsform kann der Prozessor **2400** auch einen Satz von einer oder mehreren Bussteuerungseinheiten **2416** und einem Systemagentenkern **2410** beinhalten. In mindestens einer Ausführungsform verwalten eine oder mehrere Bussteuerungseinheiten **2416** einen Satz von Peripheriegerätebussen, wie etwa einen oder mehrere PCI- oder PCI-Express-Busse. In mindestens einer Ausführungsform stellt der Systemagentenkern **2410** Verwaltungsfunktionen für verschiedene Prozessorkomponenten bereit. In mindestens einer Ausführungsform beinhaltet der Systemagentenkern **2410** eine oder mehrere integrierte Speichersteuerungen **2414**, um den Zugriff auf verschiedene externe Speichervorrichtungen (nicht gezeigt) zu verwalten.

[0226] In mindestens einer Ausführungsform beinhalten einer oder mehrere der Prozessorkerne **2402A-2402N** Unterstützung für simultanes Multithreading. In mindestens einer Ausführungsform beinhaltet der Systemagentenkern **2410** Komponenten zum Koordinieren und Betreiben der Kerne **2402A - 2402N** während der Multithread-Verarbeitung. In mindestens einer Ausführungsform kann der Systemagentenkern **2410** zusätzlich eine Leistungssteuereinheit (PCU) beinhalten, die Logik und Komponenten beinhaltet, um einen oder mehrere Leistungszustände der Prozessorkerne **2402A-2402N** und des Grafikprozessors **2408** zu regulieren.

[0227] In mindestens einer Ausführungsform beinhaltet der Prozessor **2400** zusätzlich den Grafikprozessor **2408**, um Grafikverarbeitungsoperationen auszuführen. In mindestens einer Ausführungsform ist der Grafikprozessor **2408** mit Einheiten des gemeinsam genutzten Zwischenspeichers **2406** und dem Systemagentenkern **2410** gekoppelt, was eine oder mehrere integrierte Speichersteuerungen **2414** beinhaltet. In mindestens einer Ausführungsform beinhaltet der Systemagentenkern **2410** außerdem eine Anzeigesteuerung **2411**, um die Grafikprozessorausgabe an eine oder mehrere gekoppelte Anzeigen zu lenken. In mindestens einer Ausführungsform kann die Anzeigesteuerung **2411** auch ein getrenntes Modul sein, das über mindestens eine

Verbindung mit dem Grafikprozessor **2408** gekoppelt ist, oder kann innerhalb des Grafikprozessors **2408** integriert sein.

[0228] In mindestens einer Ausführungsform wird eine ringbasierte Verbindungseinheit **2412** verwendet, um interne Komponenten des Prozessors **2400** zu koppeln. In mindestens einer Ausführungsform kann eine alternative Verbindungseinheit verwendet werden, wie etwa eine Punkt-zu-Punkt-Verbindung, eine geschaltete Verbindung oder andere Methoden. In mindestens einer Ausführungsform ist der Grafikprozessor **2408** über eine E/A-Verbindung **2413** mit der Ringverbindung **2412** gekoppelt.

[0229] In mindestens einer Ausführungsform stellt die E/A-Verbindung **2413** mindestens eine von mehreren Arten von E/A-Verbindungen dar, was eine E/A-Verbindung auf dem Gehäuse beinhaltet, welche die Kommunikation zwischen verschiedenen Prozessorkomponenten und einem eingebetteten Hochleistungsspeichermodul **2418** erleichtert, wie etwa als eDRAM-Modul. In mindestens einer Ausführungsform verwenden jeder der Prozessorkerne **2402A-2402N** und der Grafikprozessor **2408** eingebettete Speichermodule **2418** als gemeinsam genutzten Zwischenspeicher der letzten Ebene.

[0230] In mindestens einer Ausführungsform sind die Prozessorkerne **2402A-2402N** homogene Kerne, die eine gemeinsame Anweisungssatzarchitektur ausführen. In mindestens einer Ausführungsform sind die Prozessorkerne **2402A-2402N** im Hinblick auf die Anweisungssatzarchitektur (instruction set architecture - ISA) heterogen, wobei einer oder mehrere der Prozessorkerne **2402A-2402N** einen gemeinsamen Anweisungssatz ausführen, während ein oder mehrere andere Kerne der Prozessorkerne **2402A-2402N** eine Teilmenge eines gemeinsamen Anweisungssatzes oder einen anderen Anweisungssatz ausführen. In mindestens einer Ausführungsform sind die Prozessorkerne **2402A-2402N** hinsichtlich der Mikroarchitektur heterogen, wobei ein oder mehrere Kerne, die einen verhältnismäßig höheren Leistungsverbrauch aufweisen, mit einem oder mehreren Leistungskernen gekoppelt sind, die einen geringeren Leistungsverbrauch aufweisen. In mindestens einer Ausführungsform kann der Prozessor **2400** auf einem oder mehreren Chips oder als integrierter SoC-Schaltkreis umgesetzt sein.

[0231] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. Details bezüglich der Ableitungs- und/oder Trainingslogik **615** werden nachfolgend in Verbindung mit den **Fig. 6A** und/oder 6B bereitgestellt. In mindestens einer Ausführungsform können Teile der oder die gesamte Ableitungs- und/oder Trainingslogik **615** in den Prozessor **2400** integriert sein. Zum Beispiel können in mindestens einer Ausführungsform in dieser Schrift beschriebene Trainings- und/oder Ableitungsmethoden eine oder mehrere der im Grafikprozessor **2312**, den Grafikkernen **2402A-2402N** oder anderen Komponenten in **Fig. 24** realisierten ALUs verwenden. Darüber hinaus können in mindestens einer Ausführungsform in dieser Schrift beschriebene Ableitung- und/oder Trainingsoperationen unter Verwendung einer anderen Logik als der in den **Fig. 6A** oder **Fig. 6B** veranschaulichten Logik durchgeführt werden. In mindestens einer Ausführungsform können Gewichtungsparameter in chipinternen oder chipexternen Speicher und/oder Registern (gezeigt oder nichtgezeigt) gespeichert werden, die ALUs des Grafikprozessors **2400** konfigurieren, um einen oder mehrere Algorithmen maschinellen Lernens, neuronale Netzwerkarchitekturen, Anwendungsfälle oder in dieser Schrift beschriebene Trainingsmethoden auszuführen.

[0232] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. In mindestens einer Ausführungsform kann diese Logik mit Komponenten dieser Figuren verwendet werden, um ein dreidimensionales Bild einer bestimmten Ansicht aus zweidimensionalen Bildern aus einer willkürlichen Anzahl von Kameras zu erzeugen.

[0233] **Fig. 25** ist ein Blockdiagramm einer Hardwarelogik eines Grafikprozessorkerns **2500** gemäß mindestens einer in dieser Schrift beschriebenen Ausführungsform. In mindestens einer Ausführungsform ist der Grafikprozessorkern **2500** in einem Grafikkernarray beinhalten. In mindestens einer Ausführungsform kann der Grafikprozessorkern **2500**, der manchmal als Core-Slice bezeichnet wird, ein oder mehrere Grafikkerne innerhalb eines modularen Grafikprozessors sein. In mindestens einer Ausführungsform ist der Grafikprozessorkern **2500** beispielhaft für einen Grafikkern-Slice, und ein wie in dieser Schrift beschriebener Grafikprozessor kann mehrere Grafikkern-Slices auf Grundlage von Zielleistungs- und Leistungshüllkurven beinhalten. In mindestens einer Ausführungsform kann jeder Grafikkern **2500** einen Festfunktionsblock **2530** beinhalten, der mit mehreren Teilkernen **2501A-2501F** gekoppelt ist, die auch als Teil-Slices bezeichnet werden, die modulare Blöcke von Allzweck- und Festfunktionslogik beinhalten.

[0234] In mindestens einer Ausführungsform beinhaltet der Festfunktionsblock **2530** eine Geometrie-/Festfunktionspipeline **2536**, die von allen Teilkernen im Grafikprozessor **2500** geteilt werden kann, zum Beispiel in Umsetzungen von Grafikprozessoren mit geringerer Leistung und/oder niedrigerer Leistung. In mindestens einer Ausführungsform beinhaltet die Geometrie-/Festfunktionspipeline **2536** eine 3D-Festfunktionspipeline, eine Video-Front-End-Einheit, einen Thread-Schaffer und Thread-Versender und einen einheitlichen Rückgabepufferverwalter, der einheitliche Rückgabepuffer verwaltet.

[0235] In mindestens einer festen Ausführungsform beinhaltet der Funktionsblock **2530** außerdem eine Grafik-SoC-Schnittstelle **2537**, eine Grafik-Mikrosteuerung **2538** und eine Medienpipeline **2539**. In mindestens einer festen Ausführungsform stellt die Grafik-SoC-Schnittstelle **2537** eine Schnittstelle zwischen dem Grafikkern **2500** und anderen Prozessorkernen innerhalb eines Systems auf einem integrierten Chip-Schaltkreis bereit. In mindestens einer Ausführungsform ist die Grafikmikrosteuerung **2538** ein programmierbarer Teilprozessor, der dazu konfiguriert werden kann, verschiedene Funktionen des Grafikprozessors **2500** zu verwalten, was Thread-Versendung, -Planung und -Bevorrechtigung beinhaltet. In mindestens einer Ausführungsform beinhaltet die Medienpipeline **2539** eine Logik, um das Dekodieren, Codieren, Vorverarbeiten und/oder Nachbearbeiten von Multimediadaten zu erleichtern, was Bild- und Videodaten beinhaltet. In mindestens einer Ausführungsform setzt die Medienpipeline **2539** Medienoperationen über Anforderungen an die Rechen- oder Abtastlogik innerhalb der Teilkerne **2501-2501F** um.

[0236] In mindestens einer Ausführungsform ermöglicht die SoC-Schnittstelle **2537** dem Grafikkern **2500**, mit Allzweck-Anwendungsprozessorkernen (z. B. CPUs) und/oder anderen Komponenten innerhalb eines SoC zu kommunizieren, was Speicherhierarchieelemente beinhaltet, wie etwa einem gemeinsam genutzten schnellen Pufferspeicher der letzten Ebene, System, RAM und/oder eingebetteten DRAM auf dem Chip oder auf dem Gehäuse. In mindestens einer Ausführungsform kann die SoC-Schnittstelle **2537** auch die Kommunikation mit Vorrichtungen mit fester Funktion innerhalb eines SoC ermöglichen, wie etwa Kamerabildgebungspipelines, und ermöglicht die Verwendung globaler Speicheratomare, die zwischen dem Grafikkern **2500** und den CPUs innerhalb einer SoC gemeinsam genutzt werden können, und/oder setzt diese um. In mindestens einer Ausführungsform kann die SoC-Schnittstelle **2537** auch Leistungsverwaltungssteuerungen für den Grafikkern **2500** umsetzen und eine Schnittstelle zwischen einer Taktdomäne des Grafikkerns **2500** und anderen Taktdomänen innerhalb eines SoC ermöglichen. In mindestens einer Ausführungsform ermöglicht die SoC-Schnittstelle **2537** den Empfang von Befehlspuffern von einem Befehlsstreamer und einem globalen Thread-Versender, die dazu konfiguriert sind, jedem von einem oder mehreren Grafikkernen innerhalb eines Grafikprozessors Befehle und Anweisungen bereitzustellen. In mindestens einer Ausführungsform können Befehle und Anweisungen an die Medienpipeline **2539** gesendet werden, wenn Medienoperationen ausgeführt werden sollen, oder an eine Geometrie- und Festfunktionspipeline (z. B. Geometrie- und Festfunktionspipeline **2536**, Geometrie- und Festfunktionspipeline **2514**), wenn Grafikverarbeitungsoperationen durchgeführt werden.

[0237] In mindestens einer Ausführungsform kann die Grafikmikrosteuerung **2538** dazu konfiguriert sein, verschiedene Planungs- und Verwaltungsaufgaben für den Grafikkern **2500** durchzuführen. In mindestens einer Ausführungsform kann die Grafikmikrosteuerung **2538** Grafik- und/oder Rechenarbeitslast-Planung auf verschiedenen Grafik-Parallelengines innerhalb der Ausführungseinheits-(EU-)Arrays **2502A-2502F**, **2504A-2504F** innerhalb der Teilkerne **2501A-2501F** durchführen. In mindestens einer Ausführungsform kann Hostsoftware, die auf einem CPU-Kern eines SoCs ausgeführt wird, der den Grafikkern **2500** beinhaltet, Arbeitslasten an eine von mehreren Grafikprozessor-Doorbells senden, die eine Planungsoperation auf einer geeigneten Grafikengine aufruft. In mindestens einer Ausführungsform beinhaltet das Planen von Operationen das Bestimmen, welche Arbeitslast als nächstes ausgeführt werden soll, das Übermitteln einer Arbeitslast an einen Befehlsstreamer, das Bevorrechtigen vorhandener Arbeitslasten, die auf einer Engine ausgeführt werden, das Überwachen des Fortschritts einer Arbeitslast und das Benachrichtigen der Hostsoftware, wenn eine Arbeitslast abgeschlossen ist. In mindestens einer Ausführungsform kann die Grafikmikrosteuerung **2538** auch Niedrigenergie- oder Ruhezustände für den Grafikkern **2500** ermöglichen, wodurch dem Grafikkern **2500** eine Fähigkeit bereitgestellt wird, Register innerhalb des Grafikkerns **2500** über Zustandsübergänge mit geringem Leistungsverbrauch unabhängig von einem Betriebssystem und/oder Grafiktreibersoftware auf einem System zu speichern und wiederherzustellen.

[0238] In mindestens einer Ausführungsform kann der Grafikkern **2500** mehr oder weniger als die dargestellten Teilkerne **2501A-2501F**, bis zu N modulare Teilkerne aufweisen. Für jeden Satz von N Teilkernen kann der Grafikkern **2500** in mindestens einer Ausführungsform auch eine Logik für gemeinsam genutzte Funktionen **2510**, einen gemeinsam genutzten und/oder schnellen Pufferspeicher **2512**, eine Geometrie-/Festfunktionspipeline **2514** sowie eine zusätzliche Festfunktionslogik **2516** beinhalten, um verschiedene Grafik- und Rechenverarbeitungsvorgänge zu beschleunigen. In mindestens einer Ausführungsform kann die Logik für

gemeinsam genutzte Funktionen **2510** Logikeinheiten (z. B. Abtaster-, Mathematik- und/oder Zwischen-Thread-Kommunikationslogik) beinhalten, die von allen N Teilkernen innerhalb des Grafikkerns **2500** gemeinsam genutzt werden können. In mindestens einer Ausführungsform kann der feste, gemeinsam genutzte und/oder schnelle Pufferspeicher **2512** einen Zwischenspeicher der letzten Ebene für N Teilkerne **2501A-2501F** innerhalb des Grafikkerns **2500** sein und kann außerdem als gemeinsam genutzter Speicher dienen, auf den mehrere Teilkerne zugreifen können. In mindestens einer Ausführungsform kann die Geometrie-/Festfunktionspipeline **2514** anstelle der Geometrie-/Festfunktionspipeline **2536** innerhalb des Festfunktionsblocks **2530** beinhaltet sein und kann selbe oder ähnliche Logikeinheiten beinhalten.

[0239] In mindestens einer Ausführungsform beinhaltet der Grafikkern **2500** zusätzliche Festfunktionslogik **2516**, die verschiedene Festfunktionsbeschleunigungslogik zur Verwendung durch den Grafikkern **2500** beinhalten kann. In mindestens einer Ausführungsform beinhaltet die zusätzliche Festfunktionslogik **2516** eine zusätzliche Geometriepipeline zur Verwendung bei der Schattierung von lediglich der Position. Bei einer Schattierung von lediglich der Position existieren mindestens zwei Geometriepipelines, wohingegen in einer Vollgeometriepipeline innerhalb der Geometrie-/Fixfunktionspipeline **2516**, **2536** und eine Auslesepipeline, die eine zusätzliche Geometriepipeline ist, die innerhalb der zusätzlichen Festfunktionslogik **2516** beinhaltet sein kann. In mindestens einer Ausführungsform ist die Auslesepipeline eine gekürzte Version einer Vollgeometriepipeline. In mindestens einer Ausführungsform können eine vollständige Pipeline und eine Auslesepipeline verschiedene Instanzen einer Anwendung ausführen, wobei jede Instanz einen getrennten Kontext aufweist. In mindestens einer Ausführungsform kann eine Schattierung von lediglich der Position lange Ausleseläufe von verworfenen Dreiecken verbergen, wodurch die Schattierung in einigen Fällen früher abgeschlossen werden kann. Zum Beispiel kann in mindestens einer Ausführungsform die Auslesepipeline-Logik innerhalb der zusätzlichen Festfunktionslogik **2516** Positions-Shader parallel zu einer Hauptanwendung ausführen und erzeugt im Allgemeinen kritische Ergebnisse schneller als eine vollständige Pipeline, da die Auslesepipeline das Positionsattribut von Scheitelpunkten abruft und schattiert, ohne eine Rasterisierung und ein Rendern von Bildpunkten in einen Bildspeicher durchzuführen. In mindestens einer Ausführungsform kann die Auslesepipeline erzeugte kritische Ergebnisse verwenden, um Sichtbarkeitsinformationen für alle Dreiecke zu berechnen, unabhängig davon, ob diese Dreiecke aussortiert wurden. In mindestens einer Ausführungsform kann eine vollständige Pipeline (die in diesem Fall als eine Wiedergabepipeline bezeichnet werden kann) Sichtbarkeitsinformationen verbrauchen, um aussortierte Dreiecke zu überspringen, um nur sichtbare Dreiecke zu schattieren, die schließlich zu einer Rasterisierungsphase weitergeleitet werden.

[0240] In mindestens einer Ausführungsform kann die zusätzliche Logik der feststehenden Funktion **2516** auch eine Logik der Beschleunigung des maschinellen Lernens, wie etwa eine Matrixmultiplikationslogik mit feststehender Funktion, für Umsetzungen beinhalten, die Optimierungen für das Training oder Ableitungen maschinellen Lernens beinhalten.

[0241] In mindestens einer Ausführungsform beinhaltet jeder Grafik-Teilkern **2501A-2501F** einen Satz von Ausführungsressourcen, die verwendet werden können, um Grafik-, Medien- und Rechenoperationen als Reaktion auf Anforderungen durch die Grafik-Pipeline, Medien-Pipeline oder Shader-Programme durchzuführen. In mindestens einer Ausführungsform beinhalten die Grafik-Teilkerne **2501A-2501F** mehrere EU-Arrays **2502A-2502F**, **2504A-2504F**, eine Thread-Versende- und Zwischen-Thread-Kommunikations-(TD/IC-)Logik **2503A-2503F**, einen 3D-(z. B. Textur-)Abtaster **2505A-2505F**, ein Medien-Abtaster **2506A-2506F**, ein Shader-Prozessor **2507A-2507F** und einen gemeinsam genutzten lokalen Speicher (shared local memory - SLM) **2508A-2508F**. Die EU-Arrays **2502A-2502F**, **2504A-2504F** beinhalten jeweils mehrere Ausführungseinheiten, die Allzweck-Grafikverarbeitungseinheiten sind, die Gleitkomma- und Ganzzahl-/Festkomma-Logikoperationen im Dienst einer Grafik-, Medien- oder Rechenoperation ausführen können, was Grafik-, Medien- oder Rechen-Shader-Programme beinhaltet. In mindestens einer Ausführungsform führt die TD/IC-Logik **2503A-2503F** lokale Thread-Versende- und Thread-Steuerungsoperationen für Ausführungseinheiten innerhalb eines Teilkerns durch und erleichtert die Kommunikation zwischen Threads, die auf Ausführungseinheiten eines Teilkerns ausgeführt werden. In mindestens einer Ausführungsform kann der 3D-Abtaster **2505A-2505F** Daten mit Bezug zu Textur- oder anderer 3D-Grafik in den Speicher einlesen. In mindestens einer Ausführungsform kann der 3D-Abtaster Texturdaten auf Grundlage eines konfigurierten Sample-Zustands und eines Texturformats unterschiedlich lesen, das einer bestimmten Textur zugeordnet ist. In mindestens einer Ausführungsform kann der Medien-Abtaster **2506A-2506F** ähnliche Leseoperationen auf Grundlage eines Typs und eines Formats durchführen, die Mediendaten zugeordnet sind. In mindestens einer Ausführungsform kann jeder Grafik-Teilkern **2501A-2501F** alternativ einen einheitlichen 3D- und Medien-Abtaster beinhalten. In mindestens einer Ausführungsform können Threads, die auf Ausführungseinheiten in jedem der Teilkerne **2501A-2501F** ausgeführt werden, einen gemeinsam genutzten lokalen Speicher **2508A-2508F** innerhalb jedes Teilkerns nutzen,

um Threads, die in einer Thread-Gruppe ausgeführt werden, die Ausführung unter Verwendung eines gemeinsamen Pools des Speichers auf dem Chip zu ermöglichen.

[0242] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. Details bezüglich der Ableitungs- und/oder Trainingslogik **615** werden nachfolgend in Verbindung mit den **Fig. 6A** und/oder 6B bereitgestellt. In mindestens einer Ausführungsform können Teile der oder die gesamte Ableitungs- und/oder Trainingslogik **615** in den Grafikprozessor **2510** integriert sein. Zum Beispiel können in mindestens einer Ausführungsform in dieser Schrift beschriebene Trainings- und/oder Ableitungsmethoden eine oder mehrere der in dem Grafikprozessor **2312**, der Grafikmikrosteuerung **2538**, der Geometrie- und Festfunktionspipeline **2514** und **2536** oder einer anderen Logik in **Fig. 24** realisierten ALUs verwenden. Darüber hinaus können in mindestens einer Ausführungsform in dieser Schrift beschriebene Ableitung- und/oder Trainingsoperationen unter Verwendung einer anderen Logik als der in den **Fig. 6A** oder **Fig. 6B** veranschaulichten Logik durchgeführt werden. In mindestens einer Ausführungsform können Gewichtungsparameter in chipinternen oder chipexternen Speicher und/oder Registern (gezeigt oder nichtgezeigt) gespeichert werden, die ALUs des Grafikprozessors **2500** konfigurieren, um einen oder mehrere Algorithmen maschinellen Lernens, neuronale Netzwerkarchitekturen, Anwendungsfälle oder in dieser Schrift beschriebene Trainingsmethoden auszuführen.

[0243] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. In mindestens einer Ausführungsform kann diese Logik mit Komponenten dieser Figuren verwendet werden, um ein dreidimensionales Bild einer bestimmten Ansicht aus zweidimensionalen Bildern aus einer willkürlichen Anzahl von Kameras zu erzeugen.

[0244] Die **Fig. 26A-26B** veranschaulichen die Thread-Ausfiihrungslogik **2600**, die ein Array von Verarbeitungselementen eines Grafikprozessorkerns gemäß mindestens einer Ausführungsform beinhaltet. **Fig. 26A** veranschaulicht mindestens eine Ausführungsform, in der die Thread-Ausführungslogik **2600** verwendet wird. **Fig. 26B** veranschaulicht beispielhafte interne Details einer Ausführungseinheit gemäß mindestens einer Ausführungsform.

[0245] Wie in **Fig. 26A** veranschaulicht, beinhaltet die Thread-Ausfiihrungslogik **2600** in mindestens einer Ausführungsform einen Shader-Prozessor **2602**, einen Thread-Versender **2604**, einen Anweisungszwischenspeicher **2606**, ein skalierbares Ausführungseinheitenarray, das eine Vielzahl von Ausführungseinheiten **2608A-2608N** beinhaltet, Abtaster **2610**, einen Datenzwischenspeicher **2612** und einen Datenanschluss **2614**. In mindestens einer Ausführungsform kann ein skalierbares Ausführungseinheitenarray dynamisch skalieren, indem eine oder mehrere Ausführungseinheiten (z. B. eine der Ausführungseinheiten **2608A**, **2608B**, **2608C**, **2608D** bis **2608N-1** und **2608N**) zum Beispiel auf Grundlage der Rechenanforderungen von einer Arbeitslast aktiviert oder deaktiviert werden. In mindestens einer Ausführungsform sind skalierbare Ausführungseinheiten über eine Verbindungsstruktur miteinander verbunden, die mit jeder der Ausführungseinheiten verbunden ist. In mindestens einer Ausführungsform beinhaltet die Thread-Ausfiihrungslogik **2600** eine oder mehrere Verbindungen zu einem Speicher, wie etwa einem Systemspeicher oder schnellen Pufferspeicher, über einen oder mehrere von Anweisungszwischenspeicher **2606**, Datenanschluss **2614**, Abtaster **2610** und Ausführungseinheiten **2608A-2608N**. In mindestens einer Ausführungsform ist jede Ausführungseinheit (z. B. **2608A**) eine eigenständige programmierbare Allzweck-Recheneinheit, die in der Lage ist, mehrere gleichzeitige Hardware-Threads auszuführen, während mehrere Datenelemente parallel für jeden Thread verarbeitet werden. In mindestens einer Ausführungsform ist das Array von Ausführungseinheiten **2608A-2608N** skalierbar, um eine beliebige Anzahl einzelner Ausführungseinheiten zu beinhalten.

[0246] In mindestens einer Ausführungsform werden die Ausführungseinheiten **2608A-2608N** hauptsächlich verwendet, um Shader-Programme auszuführen. In mindestens einer Ausführungsform kann der Shader-Prozessor **2602** verschiedene Shader-Programme verarbeiten und Ausführungsthreads, die Shader-Programmen zugeordnet sind, über einen Thread-Versender **2604** verteilen. In mindestens einer Ausführungsform beinhaltet der Thread-Versender **2604** eine Logik, um Thread-Initiierungsanforderungen von Grafik- und Medienpipelines zu vermitteln und angeforderte Threads auf einer oder mehreren Ausführungseinheiten in den Ausführungseinheiten **2608A-2608N** zu instanziieren. Zum Beispiel kann eine Geometrie-Pipeline in mindestens einer Ausführungsform Scheitelpunkt-, Tessellations- oder Geometrie-Shader an die Thread-Ausführungslogik zur Verarbeitung senden. In mindestens einer Ausführungsform kann der Thread-Versender **2604** außerdem Laufzeitthreadschaffungsanforderungen von der Ausführung von Shader-Programmen verarbeiten.

[0247] In mindestens einer Ausführungsform unterstützen die Ausführungseinheiten **2608A-2608N** einen Anweisungssatz, der derartig native Unterstützung für viele standardmäßige 3D-Grafik-Shader-Befehle beinhaltet, dass Shader-Programme von Grafikbibliotheken (z. B. Direct 3D und OpenGL) mit einer minimalen Übersetzung ausgeführt werden. In mindestens einer Ausführungsform unterstützen Ausführungseinheiten Scheitelpunkt- und Geometrieverarbeitung (z. B. Scheitelpunktprogramme, Geometrieprogramme, Scheitelpunkt-Shader), Bildpunktverarbeitung (z. B. Bildpunkt-Shader, Fragment-Shader) und Allzweckverarbeitung (z. B. Rechen- und Medien-Shader). In mindestens einer Ausführungsform ist jede der Ausführungseinheiten **2608A-2608N**, die eine oder mehrere arithmetische Logikeinheiten (ALUs) beinhalten, zu einer Mehrfachausgabe-Einzelbefehl-Mehrfachdaten-(SIMD-)Ausführung in der Lage, und eine Multithread-Operation ermöglicht eine effiziente Ausführungsumgebung trotz höherer Latenzspeicherzugriffe. In mindestens einer Ausführungsform weist jeder Hardware-Thread innerhalb jeder Ausführungseinheit eine dedizierte Registerdatei mit hoher Bandbreite und einen zugeordneten unabhängigen Thread-Zustand auf In mindestens einer Ausführungsform erfolgt die Ausführung mehrfach pro Takt an Pipelines, die zu Ganzzahl- und Gleitkommaoperationen mit einfacher und doppelter Genauigkeit, SIMD-Verzweigungsfähigkeit, logischen Operationen, transzendenten Operationen und anderen verschiedenartigen Operationen in der Lage sind. In mindestens einer Ausführungsform veranlasst die Abhängigkeitslogik innerhalb der Ausführungseinheiten **2608A-2608N**, während auf Daten aus dem Speicher oder einer von gemeinsam genutzten Funktionen gewartet wird, dass ein wartender Thread im Ruhezustand bleibt, bis angeforderte Daten zurückgegeben wurden. In mindestens einer Ausführungsform können Hardwareressourcen für die Verarbeitung anderer Threads verwendet werden, während sich ein wartender Thread im Ruhezustand befindet. Zum Beispiel kann in mindestens einer Ausführungsform während einer Verzögerung, die mit einer Scheitelpunkt-Shader-Operation verbunden ist, eine Ausführungseinheit Operationen für einen Bildpunkt-Shader, Fragment-Shader oder eine andere Art von Shader-Programm ausführen, was einen anderen Scheitelpunkt-Shader beinhaltet.

[0248] In mindestens einer Ausführungsform arbeitet jede Ausführungseinheit in den Ausführungseinheiten **2608A-2608N** an Arrays von Datenelementen. In mindestens einer Ausführungsform ist eine Anzahl von Datenelementen die „Ausfiihrungsgröße“ oder die Anzahl von Kanälen für eine Anweisung. In mindestens einer Ausführungsform ist ein Ausführungskanal eine logische Ausführungseinheit für den Datenelementzugriff, die Maskierung und die Ablaufsteuerung innerhalb von Anweisungen. In mindestens einer Ausführungsform kann eine Anzahl von Kanälen unabhängig von einer Anzahl von physischen arithmetischen Logikeinheiten (ALUs) oder Gleitkommaeinheiten (FPUs) für einen bestimmten Grafikprozessor sein. In mindestens einer Ausführungsform unterstützen die Ausführungseinheiten **2608A-2608N** Ganzzahl- und Gleitkomma-Datenarten.

[0249] In mindestens einer Ausführungsform beinhaltet ein Ausführungseinheitanweisungssatz SIMD-Anweisungen. In mindestens einer Ausführungsform können verschiedene Datenelemente als eine gepackte Datenart in einem Register gespeichert werden und die Ausführungseinheit verarbeitet verschiedene Elemente auf Grundlage der Datengröße von Elementen. Zum Beispiel werden in mindestens einer Ausführungsform beim Arbeiten mit einem 256 Bit breiten Vektor 256 Bits eines Vektors in einem Register gespeichert und eine Ausführungseinheit arbeitet mit einem Vektor als vier getrennte gepackte 64-Bit-Datenelemente (Datenelemente mit der Größe Quad-Word (QW)), acht getrennte 32-Bit-gepackte Datenelemente (Datenelemente mit der Größe Double-Word (DW)), sechzehn getrennte 16-Bit-gepackte Datenelemente (Datenelemente mit der Größe Word (W)) oder zweiunddreißig getrennte 8-Bit-Datenelemente (Datenelemente mit der Größe Byte (B). Bei mindestens einer Ausführungsform sind jedoch unterschiedliche Vektorbreiten und Registergrößen möglich.

[0250] In mindestens einer Ausführungsform können eine oder mehrere Ausführungseinheiten zu einer verschmolzenen Ausführungseinheit **2609A-2609N** mit einer Thread-Steuerungslogik (**2607A-2607N**) kombiniert werden, die verschmolzenen EUs gemeinsam ist. In mindestens einer Ausführungsform können mehrere EUs zu einer EU-Gruppe verschmolzen werden. In mindestens einer Ausführungsform kann jede EU in einer verschmolzenen EU-Gruppe dazu konfiguriert sein, einen getrennten SIMD-Hardware-Thread auszuführen. Die Anzahl von EUs in einer verschmolzenen EU-Gruppe kann gemäß verschiedenen Ausführungsformen variieren. In mindestens einer Ausführungsform können verschiedene SIMD-Breiten pro EU durchgeführt werden, was SIMD8, SIMD16 und SIMD32 beinhaltet, ohne darauf beschränkt zu sein. In mindestens einer Ausführungsform beinhaltet jede verschmolzene Grafikausführungseinheit **2609A-2609N** mindestens zwei Ausführungseinheiten. Zum Beispiel beinhaltet die verschmolzene Ausführungseinheit **2609A** in mindestens einer Ausführungsform eine erste EU **2608A**, eine zweite EU **2608B** und eine Thread-Steuerungslogik **2607A**, die der ersten EU **2608A** und der zweiten EU **2608B** gemeinsam ist. In mindestens einer Ausführungsform steuert die Thread-Steuerungslogik **2607A** Threads, die auf der verschmolzenen Grafikausführungseinheit **2609A** ausgeführt werden, was es jeder EU innerhalb der verschmolzenen Ausführungseinheiten **2609A-2609N** ermöglicht, unter Verwendung eines gemeinsamen Anweisungsverweisregisters auszuführen.

[0251] In mindestens einer Ausführungsform sind ein oder mehrere interne Anweisungszwischenspeicher (z. B. **2606**) in der Thread-Ausführungslogik **2600** beinhaltet, um Thread-Anweisungen für Ausführungseinheiten zwischenzuspeichern. In mindestens einer Ausführungsform sind ein oder mehrere Datenzwischenspeicher (z. B. **2612**) beinhaltet, um Thread-Daten während der Thread-Ausführung zwischenzuspeichern. In mindestens einer Ausführungsform ist ein Abtaster **2610** beinhaltet, um eine Texturabtastung für 3D-Operationen und eine Medienabtastung für Medienoperationen bereitzustellen. In mindestens einer Ausführungsform beinhaltet der Abtaster **2610** eine spezielle Textur- oder Medien-Abtastungsfunktion, um Textur- oder Mediendaten während eines Abtastungsprozesses zu verarbeiten, bevor abgetastete Daten an eine Ausführungseinheit geliefert werden.

[0252] Während der Ausführung senden in mindestens einer Ausführungsform Grafik- und Medienpipelines Thread-Initiierungsanforderungen an die Thread-Ausführungslogik **2600** über Thread-Schaffungs- und -Versende-Logik. In mindestens einer Ausführungsform wird, sobald eine Gruppe geometrischer Objekte verarbeitet und in Bildpunktdaten gerastert wurde, eine Bildpunktprozessorlogik (z. B. Bildpunktshaderlogik, Fragmentshaderlogik usw.) innerhalb des Shaderprozessors **2602** aufgerufen, um weitere Ausgabeinformationen zu berechnen und zu veranlassen, dass Ergebnisse auf Ausgabeoberflächen (z. B. Farbpuffer, Tiefenpuffer, Schablonenpuffer usw.) geschrieben werden. In mindestens einer Ausführungsform berechnet ein Bildpunkt-Shader oder Fragment-Shader Werte verschiedener Scheitelpunkt-Attribute, die über ein gerastertes Objekt zu interpolieren sind. In mindestens einer Ausführungsform führt die Bildpunktprozessorlogik innerhalb des Shader-Prozessors **2602** dann ein von der Anwendungsprogrammierschnittstelle (API) geliefertes Bildpunkt- oder Fragment-Shader-Programm aus. Um ein Shader-Programm auszuführen, verteilt der Shader-Prozessor **2602** in mindestens einer Ausführungsform Threads über den Thread-Versender **2604** an eine Ausführungseinheit (z. B. **2608A**). In mindestens einer Ausführungsform verwendet der Shader-Prozessor **2602** Textur-Abtastungs-Logik im Abtaster **2610**, um auf Texturdaten in Texturkarten zuzugreifen, die im Speicher gespeichert sind. In mindestens einer Ausführungsform berechnen arithmetische Operationen an Texturdaten und eingegebenen Geometriedaten Bildpunktfarbdaten für jedes geometrische Fragment oder verwerfen ein oder mehrere Bildpunkte aus der weiteren Verarbeitung.

[0253] In mindestens einer Ausführungsform stellt der Datenanschluss **2614** einen Speicherzugriffsmechanismus für die Thread-Ausführungslogik **2600** bereit, um verarbeitete Daten an den Speicher zur weiteren Verarbeitung auf einer Grafikprozessor-Ausgabepipeline auszugeben. In mindestens einer Ausführungsform beinhaltet der Datenanschluss **2614** einen oder mehrere schnelle Pufferspeicher (z. B. den Datenzwischenspeicher **2612**) oder ist daran gekoppelt, um Daten für den Speicherzugriff über einen Datenanschluss zwischenzuspeichern.

[0254] Wie in **Fig. 26B** veranschaulicht, kann in mindestens einer Ausführungsform eine Grafikausführungseinheit **2608** eine Anweisungsabrufeinheit **2637**, ein allgemeines Registerdateiarray (general register file - GRF) **2624**, ein Architekturregisterdateiarray (architectural register file - ARF) **2626**, einen Thread-Arbiter **2622**, eine Sendeeinheit **2630**, eine Verzweigungseinheit **2632**, einen Satz von SIMD-Gleitkommaeinheiten (FPUs) **2634** und in mindestens einer Ausführungsform einen Satz von dedizierten ganzzahligen SIMD-ALUs **2635** beinhalten. In mindestens einer Ausführungsform beinhalten GRF **2624** und ARF **2626** einen Satz von allgemeinen Registerdateien und Architekturregisterdateien, die jedem simultanen Hardware-Thread zugeordnet sind, der in der Grafikausführungseinheit **2608** aktiv sein kann. In mindestens einer Ausführungsform wird der Architekturzustand pro Thread in ARF **2626** aufrechterhalten, während Daten, die während der Thread-Ausführung verwendet werden, in GRF **2624** gespeichert werden. In mindestens einer Ausführungsform kann der Ausführungszustand jedes Threads, was den Anweisungsverweis für jeden Thread beinhaltet, in Thread-spezifischen Registern in ARF **2626** gehalten werden.

[0255] In mindestens einer Ausführungsform weist die Grafikausführungseinheit **2608** eine Architektur auf, die eine Kombination aus gleichzeitigem Multi-Threading (SMT) und feinkörnigem verschachtelten Multi-Threading (IMT) ist. In mindestens einer Ausführungsform weist die Architektur eine modulare Konfiguration auf, die zur Entwurfszeit auf Grundlage einer Zielanzahl gleichzeitiger Threads und einer Anzahl von Registern pro Ausführungseinheit fein abgestimmt werden kann, wobei die Ressourcen der Ausführungseinheit auf die Logik aufgeteilt sind, die zum Ausführen mehrerer gleichzeitiger Threads verwendet wird.

[0256] In mindestens einer Ausführungsform kann die Grafikausführungseinheit **2608** mehrere Anweisungen gemeinsam ausgeben, die jeweils unterschiedliche Anweisungen sein können. In mindestens einer Ausführungsform kann der Thread-Arbiter **2622** des Grafikausführungseinheit-Threads **2608** Anweisungen zur Ausführung an eine von der Sendeeinheit **2630**, der Verzweigungseinheit **2642** oder der SIMD-FPU(s) **2634** versenden. In mindestens einer Ausführungsform kann jeder Ausführungsthread auf **128** Allzweckregister inner-

halb von GRF **2624** zugreifen, wobei jedes Register **32** Byte speichern kann, auf die als ein SIMD-8-Element-Vektor von 32-Bit-Datenelementen zugegriffen werden kann. In mindestens einer Ausführungsform hat jeder Ausführungseinheits-Thread Zugriff auf 4 KByte innerhalb des GRF **2624**, obwohl Ausführungsformen nicht darauf beschränkt sind und in anderen Ausführungsformen mehr oder weniger Registerressourcen bereitgestellt sein können. In mindestens einer Ausführungsform können bis zu sieben Threads gleichzeitig ausgeführt werden, obwohl auch eine Anzahl von Threads pro Ausführungseinheit gemäß Ausführungsformen variieren kann. In mindestens einer Ausführungsform, in der sieben Threads auf 4 Kbyte zugreifen können, kann das GRF **2624** insgesamt 28 Kbyte speichern. In mindestens einer Ausführungsform können flexible Adressierungsmodi zulassen, dass Register gemeinsam adressiert werden, um effektiv breitere Register aufzubauen oder gestaffelte rechteckige Blockdatenstrukturen darzustellen.

[0257] In mindestens einer Ausführungsform werden Speicheroperationen, Abtastoperationen und andere Systemkommunikationen mit längerer Latenz über „Sende“-Anweisungen versendet, die von der Mitteilungsweiterleitungssendeeinheit **2630** ausgeführt werden. In mindestens einer Ausführungsform werden Verzweigungsanweisungen an eine dedizierte Verzweigungseinheit **2632** versendet, um die SIMD-Abweichung und eventuelle Annäherung zu erleichtern.

[0258] In mindestens einer Ausführungsform beinhaltet die Grafikausführungseinheit **2608** eine oder mehrere SIMD-Gleitkommaeinheiten (FPU(s)) **2634**, um Gleitkommaoperationen durchzuführen. In mindestens einer Ausführungsform unterstützen die FPU(s) **2634** außerdem die Ganzzahlberechnung. In mindestens einer Ausführungsform der FPU(s) **2634** kann SIMD bis zu einer Anzahl von M 32-Bit-Gleitkomma-(oder Ganzzahl-)Operationen ausführen, oder SIMD kann bis zu 2M 16-Bit-Ganzkomma- oder 16-Bit-Gleitkomma-Operationen ausführen. In mindestens einer Ausführungsform stellt mindestens eine von den FPU(s) erweiterte mathematische Fähigkeiten bereit, um transzendente mathematische Funktionen mit hohem Durchsatz und 64-Bit-Gleitkomma mit doppelter Genauigkeit zu unterstützen. In mindestens einer Ausführungsform ist ebenfalls ein Satz von 8-Bit-Ganzzahl-SIMD-ALUs **2635** vorhanden und kann speziell optimiert sein, um Operationen durchzuführen, die mit Berechnungen maschinellen Lernens verbunden sind.

[0259] In mindestens einer Ausführungsform können Arrays von mehreren Instanzen der Grafikausführungseinheit **2608** in einer Grafikteilkerngruppierung (z. B. einem Teil-Slice) instanziiert werden. In mindestens einer Ausführungsform kann die Ausführungseinheit **2608** Anweisungen über eine Vielzahl von Ausführungskanälen ausführen. In mindestens einer Ausführungsform wird jeder auf der Grafikausführungseinheit **2608** ausgeführte Thread auf einem anderen Kanal ausgeführt.

[0260] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. Details bezüglich der Ableitungs- und/oder Trainingslogik **615** werden nachfolgend in Verbindung mit den **Fig. 6A** und/oder 6B bereitgestellt. In mindestens einer Ausführungsform können Teile der oder die gesamte Ableitungs- und/oder Trainingslogik **615** in die Ausführungslogik **2600** integriert sein. Darüber hinaus können in mindestens einer Ausführungsform in dieser Schrift beschriebene Ableitung- und/oder Trainingsoperationen unter Verwendung einer anderen Logik als der in den **Fig. 6A** oder **Fig. 6B** veranschaulichten Logik durchgeführt werden. In mindestens einer Ausführungsform können Gewichtungsparameter in chipinternen oder chip externen Speicher und/oder Registern (gezeigt oder nichtgezeigt) gespeichert werden, die ALUs der Ausführungslogik **2600** konfigurieren, um einen oder mehrere Algorithmen maschinellen Lernens, neuronale Netzwerkarchitekturen, Anwendungsfälle oder in dieser Schrift beschriebene Trainingsmethoden auszuführen.

[0261] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. In mindestens einer Ausführungsform kann diese Logik mit Komponenten dieser Figuren verwendet werden, um ein dreidimensionales Bild einer bestimmten Ansicht aus zweidimensionalen Bildern aus einer willkürlichen Anzahl von Kameras zu erzeugen.

[0262] **Fig. 27** veranschaulicht eine Parallelverarbeitungseinheit („PPU“) **2700** gemäß mindestens einer Ausführungsform; In mindestens einer Ausführungsform ist die PPU **2700** mit engineslesbarem Code konfiguriert, der, wenn er von der PPU **2700** ausgeführt wird, die PPU **2700** veranlasst, einige oder alle der in dieser Offenbarung beschriebenen Prozesse und Methoden durchzuführen. In mindestens einer Ausführungsform ist die PPU **2700** ein Multithread-Prozessor, der auf einer oder mehreren integrierten Schaltungsvorrichtungen umgesetzt ist und Multithreading als Latenzverbergungsmethode nutzt, die entwickelt wurde, um computerlesbare Anweisungen (auch als engineslesbare Anweisungen bezeichnet oder einfach Anweisungen) in mehreren Threads parallel zu verarbeiten. In mindestens einer Ausführungsform bezieht sich ein Thread auf einen Aus-

führungs-Thread und ist eine Instanziierung eines Satzes von Anweisungen, die dazu konfiguriert sind, von der PPU **2700** ausgeführt zu werden. In mindestens einer Ausführungsform ist die PPU **2700** eine Grafikverarbeitungseinheit („GPU“), die dazu konfiguriert ist, eine Grafik-Rendering-Pipeline zum Verarbeiten von dreidimensionalen („3D“) Grafikdaten umzusetzen, um zweidimensionale („2D“) Bilddaten zur Anzeige auf einer Anzeigevorrichtung, wie etwa einer Flüssigkristallanzeige (liquid crystal display - „LCD“), zu erzeugen. In mindestens einer Ausführungsform wird die PPU **2700** genutzt, um Berechnungen durchzuführen, wie etwa Operationen der linearen Algebra und Operationen des maschinellen Lernens. **Fig. 27** veranschaulicht einen beispielhaften Parallelprozessor lediglich zu Veranschaulichungszwecken und sollte als nichteinschränkendes Beispiel von Prozessorarchitekturen verstanden werden, die im Schutzumfang dieser Offenbarung vorgesehen sind, und dass jeder geeignete Prozessor eingesetzt werden kann, um diesen zu ergänzen und/oder zu ersetzen.

[0263] In mindestens einer Ausführungsform sind eine oder mehrere PPUs **2700** dazu konfiguriert, Hochleistungsberechnungen (High Performance Computing - „HPC“), Rechenzentrumsanwendungen und Anwendungen maschinellen Lernens zu beschleunigen. In mindestens einer Ausführungsform ist die PPU **2700** dazu konfiguriert, Deep-Learning-Systeme und -Anwendungen zu beschleunigen, was die folgenden nicht einschränkenden Beispiele beinhaltet: autonome Fahrzeugplattformen, Deep Learning, hochpräzise Sprache, Bilder, Texterfassungssysteme, intelligente Videoanalyse, molekulare Simulationen, Arzneimittelentdeckung, Krankheitsdiagnose, Wettervorhersage, Big-Data-Analyse, Astronomie, Molekulardynamiksimulation, Finanzmodellierung, Robotik, Fabrikautomatisierung, Echtzeit-Sprachübersetzung, Online-Suchoptimierung und personalisierte Benutzerempfehlungen und mehr.

[0264] In mindestens einer Ausführungsform beinhaltet die PPU **2700** ohne Einschränkung eine Eingabe/Ausgabe-(„E/A“)-Einheit **2706**, eine Front-End-Einheit **2710**, eine Planungseinheit **2712**, eine Arbeitsverteilungseinheit **2714**, einen Hub **2716**, eine Crossbar („Xbar“) **2720**, ein oder mehrere allgemeine Verarbeitungscluster (general processing clusters - „GPCs“) **2718** und eine oder mehrere Partitionseinheiten („Speicherpartitionseinheiten“) **2722**. In mindestens einer Ausführungsform ist die PPU **2700** mit einem Hostprozessor oder anderen PPUs **2700** über eine oder mehrere Hochgeschwindigkeits-GPU-Verbindungen („GPU-Verbindungen“) **2708** verbunden. In mindestens einer Ausführungsform ist die PPU **2700** über eine Verbindung **2702** mit einem Hostprozessor oder anderen Peripherievorrichtungen verbunden. In mindestens einer Ausführungsform ist die PPU **2700** mit einem lokalen Speicher verbunden, der eine oder mehrere Speichervorrichtungen („Speicher“) **2704** umfasst. In mindestens einer Ausführungsform beinhalten die Speichervorrichtungen **2704** ohne Einschränkung eine oder mehrere dynamische Direktzugriffsspeicher („DRAM“). In mindestens einer Ausführungsform sind ein oder mehrere DRAM-Vorrichtungen als Teilsysteme mit hohem Bandbreitenspeicher („HBM“) konfiguriert und/oder konfigurierbar, wobei mehrere DRAM-Matrizen in jeder Vorrichtung gestapelt sind.

[0265] In mindestens einer Ausführungsform kann sich die Hochgeschwindigkeits-GPU-Verbindung **2708** auf eine drahtgebundene mehrspurige Kommunikationsverbindung beziehen, die von Systemen verwendet wird, um zu skalieren und eine oder mehrere PPUs **2700** in Kombination mit einer oder mehreren Zentralverarbeitungseinheiten (central processing units - „CPUs“) zu beinhalten, unterstützt Zwischenspeicher-Kohärenz zwischen PPUs **2700** und CPUs sowie CPU-Mastering. In mindestens einer Ausführungsform werden Daten und/oder Befehle durch die Hochgeschwindigkeits-GPU-Verbindung **2708** über den Hub **2716** an/von anderen Einheiten der PPU **2700**, wie etwa einer oder mehreren Kopierengines, Videokodierern, Videodekodierern, Leistungsverwaltungseinheiten und andere Komponenten übermittelt, die möglicherweise nicht ausdrücklich in **Fig. 27** veranschaulicht sind.

[0266] In mindestens einer Ausführungsform ist die E/A-Einheit **2706** dazu konfiguriert, Kommunikationen (z. B. Befehle, Daten) von einem Hostprozessor (in **Fig. 27** nicht veranschaulicht) über den Systembus **2702** zu übermitteln und zu empfangen. In mindestens einer Ausführungsform kommuniziert die E/A-Einheit **2706** mit dem Hostprozessor direkt über den Systembus **2702** oder über eine oder mehrere Zwischenvorrichtungen, wie etwa eine Speicherbrücke. In mindestens einer Ausführungsform kann die E/A-Einheit **2706** mit einem oder mehreren anderen Prozessoren, wie etwa einer oder mehreren der PPUs **2700**, über den Systembus **2702** kommunizieren. In mindestens einer Ausführungsform setzt die E/A-Einheit **2706** eine Peripheriegerätekomponentenverbindungsexpress-(„PCIe“-)Schnittstelle für die Kommunikation über einen PCIe-Bus um. In mindestens einer Ausführungsform setzt die E/A-Einheit **2706** Schnittstellen zum Kommunizieren mit externen Vorrichtungen um.

[0267] In mindestens einer Ausführungsform dekodiert die E/A-Einheit **2706** über den Systembus **2702** empfangene Pakete. In mindestens einer Ausführungsform stellen mindestens einige Pakete Befehle dar, die dazu konfiguriert sind, die PPU **2700** zu veranlassen, verschiedene Operationen durchzuführen. In mindestens einer Ausführungsform übermittelt die E/A-Einheit **2706** dekodierte Befehle an verschiedene andere Einheiten der

PPU **2700**, wie durch Befehle festgelegt. In mindestens einer Ausführungsform werden Befehle an die Front-End-Einheit **2710** übermittelt und/oder an den Hub **2716** oder andere Einheiten der PPU **2700** übermittelt, wie etwa eine oder mehrere Kopierengines, einen Videokodierer, einen Videodekodierer, eine Leistungsverwaltungseinheit usw. (nicht ausdrücklich in **Fig. 27** veranschaulicht). In mindestens einer Ausführungsform ist die E/A-Einheit **2706** dazu konfiguriert, Kommunikationen zwischen und unter verschiedenen logischen Einheiten der PPU **2700** weiterzuleiten.

[0268] In mindestens einer Ausführungsform kodiert ein vom Hostprozessor ausgeführtes Programm einen Befehlsstrom in einem Puffer, welcher der PPU **2700** Arbeitslasten zur Verarbeitung bereitstellt. In mindestens einer Ausführungsform umfasst eine Arbeitslast Anweisungen und Daten, die von diesen Anweisungen verarbeitet werden sollen. In mindestens einer Ausführungsform ist der Puffer ein Bereich in einem Speicher, auf den sowohl der Hostprozessor als auch die PPU **2700** zugreifen (z. B. lesen/schreiben) können - eine Host-Schnittstelleneinheit kann dazu konfiguriert sein, auf Puffer in einem Systemspeicher zuzugreifen, der mit dem Systembus **2702** über Speicheranforderungen verbunden ist, die über den Systembus **2702** durch die E/A-Einheit **2706** übermittelt werden. In mindestens einer Ausführungsform schreibt der Hostprozessor den Befehlsstrom in den Puffer und übermittelt dann derartig einen Verweis zum Anfang des Befehlsstroms an die PPU **2700**, dass die Front-End-Einheit **2710** Verweise auf einen oder mehrere Befehlsströme empfängt und einen oder mehrere Befehlsströme verwaltet, wobei Befehle aus Befehlsströmen gelesen und Befehle an verschiedene Einheiten der PPU **2700** weitergeleitet werden.

[0269] In mindestens einer Ausführungsform ist die Front-End-Einheit **2710** an die Planereinheit **2712** gekoppelt, die verschiedene GPCs **2718** dazu konfiguriert, Aufgaben zu verarbeiten, die durch einen oder mehrere Befehlsströme definiert sind. In mindestens einer Ausführungsform ist die Planereinheit **2712** dazu konfiguriert, Zustandsinformationen in Bezug auf verschiedene Aufgaben zu verfolgen, die von der Planereinheit **2712** verwaltet werden, wobei Zustandsinformationen angeben können, welchem der GPCs **2718** eine Aufgabe zugewiesen ist, ob die Aufgabe aktiv oder inaktiv ist, eine der Aufgabe zugeordnete Prioritätsebene usw. In mindestens einer Ausführungsform verwaltet die Planereinheit **2712** die Ausführung einer Vielzahl von Aufgaben auf einem oder mehreren der GPCs **2718**.

[0270] In mindestens einer Ausführungsform ist die Planereinheit **2712** an eine Arbeitsverteilungseinheit **2714** gekoppelt, die dazu konfiguriert ist, Aufgaben zur Ausführung auf GPCs **2718** zu versenden. In mindestens einer Ausführungsform verfolgt die Arbeitsverteilungseinheit **2714** eine Anzahl geplanter Aufgaben, die von der Planereinheit **2712** empfangen wurden, und verwaltet die Arbeitsverteilungseinheit **2714** einen Pool anstehender Aufgaben und einen Pool aktiver Aufgaben für jeden der GPCs **2718**. In mindestens einer Ausführungsform umfasst der Pool anstehender Aufgaben eine Anzahl von Steckplätzen (z. B. 32 Steckplätze), die Aufgaben enthalten, die einem bestimmten GPC **2718** zur Verarbeitung zugewiesen sind; der Pool aktiver Aufgaben kann eine Anzahl von Steckplätzen (z. B. 4 Steckplätze) für Aufgaben umfassen, die derartig aktiv von den GPCs **2718** verarbeitet werden, dass, wenn einer der GPCs **2718** die Ausführung einer Aufgabe abschließt, diese Aufgabe aus dem aktiven Aufgabenpool für den GPC **2718** entfernt wird und eine der anderen Aufgaben aus dem Pool anstehender Aufgaben ausgewählt und zur Ausführung auf dem GPC **2718** geplant wird. In mindestens einer Ausführungsform wird, wenn sich eine aktive Aufgabe auf dem GPC **2718** im Ruhezustand befindet, wie etwa während auf die Auflösung einer Datenabhängigkeit gewartet wird, dann die aktive Aufgabe aus dem GPC **2718** entfernt und zum Pool anstehender Aufgaben zurückgegeben, während eine andere Aufgabe im Pool anstehender Aufgaben ausgewählt und zur Ausführung auf GPC **2718** geplant wird.

[0271] In mindestens einer Ausführungsform kommuniziert die Arbeitsverteilungseinheit **2714** mit einem oder mehreren GPCs **2718** über die XBar **2720**. In mindestens einer Ausführungsform ist die XBar **2720** ein Verbindungsnetzwerk, das viele der Einheiten der PPU **2700** mit anderen Einheiten der PPU **2700** koppelt und dazu konfiguriert sein kann, die Arbeitsverteilungseinheit **2714** an einen bestimmten GPC **2718** zu koppeln. In mindestens einer Ausführungsform können eine oder mehrere andere Einheiten der PPU **2700** außerdem über den Hub **2716** mit der XBar **2720** verbunden sein.

[0272] In mindestens einer Ausführungsform werden Aufgaben von der Planereinheit **2712** verwaltet und von der Arbeitsverteilungseinheit **2714** an eine der GPCs **2718** versendet. Der GPC **2718** ist dazu konfiguriert, Aufgaben zu verarbeiten und Ergebnisse zu erzeugen. In mindestens einer Ausführungsform können Ergebnisse von anderen Aufgaben innerhalb des GPC **2718** verbraucht, über die XBar **2720** an einen anderen GPC **2718** geleitet oder im Speicher **2704** gespeichert werden. In mindestens einer Ausführungsform können Ergebnisse über Partitionseinheiten **2722**, die eine Speicherschnittstelle zum Lesen und Schreiben von Daten in den/aus dem Speicher **2704** umsetzen, in den Speicher **2704** geschrieben werden. In mindestens einer Ausführungsform können Ergebnisse über eine Hochgeschwindigkeits-GPU-Verbindung **2708** an eine andere PPU **2704**

oder CPU übermittelt werden. In mindestens einer Ausführungsform beinhaltet die PPU **2700** ohne Einschränkung eine Anzahl U von Partitionseinheiten **2722**, die gleich der Anzahl von getrennten und unterschiedlichen Speichervorrichtungen **2704** ist, die an die PPU **2700** gekoppelt sind. In mindestens einer Ausführungsform wird die Partitionseinheit **2722** nachfolgend in Verbindung mit **Fig. 29** ausführlicher beschrieben.

[0273] In mindestens einer Ausführungsform führt ein Hostprozessor einen Treiber-Kernel aus, der eine Anwendungsprogrammierschnittstelle („API“) umsetzt, die es einer oder mehreren Anwendungen ermöglicht, die auf dem Hostprozessor ausgeführt werden, Operationen zur Ausführung auf der PPU **2700** zu planen. In mindestens einer Ausführungsform werden mehrere Rechenanwendungen gleichzeitig von der PPU **2700** ausgeführt und die PPU **2700** stellt Isolation, Dienstgüte (quality of service - „QoS“) und unabhängige Adressbereiche für mehrere Rechenanwendungen bereit. In mindestens einer Ausführungsform erzeugt eine Anwendung Anweisungen (z. B. in Form von API-Aufrufen), die den Treiberkernel veranlassen, eine oder mehrere Aufgaben zur Ausführung durch die PPU **2700** zu erzeugen, und der Treiberkernel gibt Aufgaben an einen oder mehrere Streams aus, die von der PPU **2700** verarbeitet werden. In mindestens einer Ausführungsform umfasst jede Aufgabe eine oder mehrere Gruppen zusammengehöriger Threads, die als ein Warp bezeichnet sein können. In mindestens einer Ausführungsform umfasst ein Warp mehrere zusammengehörige Threads (z. B. 32 Threads), die parallel ausgeführt werden können. In mindestens einer Ausführungsform können sich zusammenwirkende Threads auf eine Vielzahl von Threads beziehen, die Anweisungen beinhalten, um eine Aufgabe durchzuführen, und die Daten über einen gemeinsam genutzten Speicher austauschen. In mindestens einer Ausführungsform werden Threads und zusammenwirkende Threads gemäß mindestens einer Ausführungsform in Verbindung mit **Fig. 29** ausführlicher beschrieben.

[0274] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. Details bezüglich der Ableitungs- und/oder Trainingslogik **615** werden nachfolgend in Verbindung mit den **Fig. 6A** und/oder 6B bereitgestellt. In mindestens einer Ausführungsform wird der Deep-Learning-Anwendungsprozessor verwendet, um ein Modell maschinellen Lernens, wie etwa ein neuronales Netzwerk, zu trainieren, um der PPU **2700** bereitgestellte Informationen vorherzusagen oder abzuleiten. In mindestens einer Ausführungsform wird die PPU **2700** verwendet, um Informationen auf Grundlage eines trainierten Modells maschinellen Lernens (z. B. neuronales Netzwerk) abzuleiten oder vorherzusagen, das von einem anderen Prozessor oder System oder von der PPU **2700** trainiert wurde. In mindestens einer Ausführungsform kann die PPU **2700** verwendet werden, um einen oder mehrere in dieser Schrift beschriebene Anwendungsfälle für neuronale Netzwerke durchzuführen.

[0275] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. In mindestens einer Ausführungsform kann diese Logik mit Komponenten dieser Figuren verwendet werden, um ein dreidimensionales Bild einer bestimmten Ansicht aus zweidimensionalen Bildern aus einer willkürlichen Anzahl von Kameras zu erzeugen.

[0276] **Fig. 28** veranschaulicht einen allgemeinen Verarbeitungscluster („GPC“) **2800** gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform ist der GPC **2800** der GPC **2718** aus **Fig. 27**. In mindestens einer Ausführungsform beinhaltet jeder GPC **2800** ohne Einschränkung eine Anzahl von Hardwareeinheiten zum Verarbeiten von Aufgaben und jeder GPC **2800** beinhaltet ohne Einschränkung einen Pipeline-Verwalter **2802**, eine Vorrasteroperationseinheit („PROP“) **2804**, eine Rasterengine **2808**, eine Arbeitsverteilungs-Crossbar (work distribution crossbar - „WDX“) **2816**, eine Speicherverwaltungseinheit (memory management unit - „MMU“) **2818**, ein oder mehrere Datenverarbeitungs-Cluster (Data Processing Clusters - „DPCs“) **2806** und eine beliebige geeignete Kombination von Teilen.

[0277] In mindestens einer Ausführungsform wird der Betrieb des GPC **2800** durch den Pipeline-Verwalter **2802** gesteuert. In mindestens einer Ausführungsform verwaltet der Pipeline-Verwalter **2802** die Konfiguration einer oder mehrerer DPCs **2806** zum Verarbeiten von Aufgaben, die dem GPC **2800** zugewiesen sind. In mindestens einer Ausführungsform konfiguriert der Pipeline-Verwalter **2802** mindestens einen von einem oder mehreren DPCs **2806**, um mindestens einen Teil einer Grafik-Rendering-Pipeline umzusetzen. In mindestens einer Ausführungsform ist der DPC **2806** dazu konfiguriert, ein Scheitelpunkt-Shader-Programm auf einem programmierbaren Streaming-Multiprozessor („SM“) **2814** auszuführen. In mindestens einer Ausführungsform ist der Pipeline-Verwalter **2802** dazu konfiguriert, Pakete, die von einer Arbeitsverteilungseinheit empfangen werden, in mindestens einer Ausführungsform an geeignete logische Einheiten innerhalb des GPC **2800** zu leiten, und einige Pakete können an Hardwareeinheiten mit feststehender Funktion in der PROP **2804** und/oder Rasterengine **2808** geleitet werden, während andere Pakete zu den DPCs **2806** zur Verarbeitung durch eine primitive Engine **2812** oder SM **2814** geleitet werden können. In mindestens einer Ausführungsform konfiguriert

der Pipeline-Verwalter **2802** mindestens einen der DPCs **2806**, um ein Modell des neuronalen Netzwerks und/oder eine Rechenpipeline umzusetzen.

[0278] In mindestens einer Ausführungsform ist die PROP-Einheit **2804** dazu konfiguriert, in mindestens einer Ausführungsform Daten, die von der Rasterengine **2808** und den DPCs **2806** erzeugt werden, an eine Rasteroperations-(„ROP“-)Einheit in der Partitionseinheit **2722** weiterzuleiten, die vorstehend in Verbindung mit **Fig. 27** ausführlicher beschrieben wurde. In mindestens einer Ausführungsform ist die PROP-Einheit **2804** dazu konfiguriert, Optimierungen für die Farbmischung durchzuführen, Bildpunktdaten zu organisieren, Adressübersetzungen durchzuführen und mehr. In mindestens einer Ausführungsform beinhaltet die Rasterengine **2808** ohne Einschränkung eine Anzahl von Hardwareeinheiten mit feststehender Funktion, die dazu konfiguriert sind, verschiedene Raster-Operationen in mindestens einer Ausführungsform durchzuführen, und die Rasterengine **2808** beinhaltet ohne Einschränkung eine Einrichtungsengine, ein Grobrasterengine, eine Ausleseengine, eine Zuschneideengine, eine Feinrasterengine, eine Kachelverschmelzengine und eine beliebige geeignete Kombination davon. In mindestens einer Ausführungsform empfängt die Einrichtungsengine transformierte Scheitelpunkte und erzeugt Ebenengleichungen, die mit geometrischen Primitiven verbunden sind, die durch Scheitelpunkte definiert sind; werden Ebenengleichungen an eine Grobrasterengine übermittelt, um Abdeckungsinformationen (z. B. eine x, y-Abdeckungsmaske für eine Kachel) für Primitive zu erzeugen; wird die Ausgabe der Grobrasterengine an eine Ausleseengine übermittelt, wo Fragmente, die mit Primitiven verbunden sind, die einen Z-Test nicht bestehen, aussortiert werden, und an eine Zuschneideengine übermittelt werden, wo Fragmente, die außerhalb eines Sichtkegels liegen, abgeschnitten werden. In mindestens einer Ausführungsform werden Fragmente, die das Ausschneiden und die Auslese überleben, an eine Feinrasterengine übergeben, um Attribute für Bildpunktfragmente auf Grundlage von Ebenengleichungen zu erzeugen, die von der Einrichtungsengine erzeugt werden. In mindestens einer Ausführungsform umfasst die Ausgabe der Rasterengine **2808** Fragmente, die von einer beliebigen geeigneten Entität verarbeitet werden sollen, wie etwa von einem innerhalb des DPC **2806** umgesetzten Fragment-Shader.

[0279] In mindestens einer Ausführungsform umfasst jeder DPC **2806**, der im GPC **2800** beinhaltet ist, ohne Einschränkung eine M-Pipe-Steuerung (M-Pipe Controller - „MPC“) **2810**; eine Primitivengine **2812**; eine oder mehrere SMs **2814**; und eine beliebige geeignete Kombination davon. In mindestens einer Ausführungsform steuert die MPC **2810** den Betrieb des DPC **2806**, der vom Pipeline-Verwalter **2802** empfangene Pakete an geeignete Einheiten im DPC **2806** weiterleitet. In mindestens einer Ausführungsform werden Pakete, die einem Scheitelpunkt zugeordnet sind, zu einer Primitivengine **2812** weitergeleitet, die dazu konfiguriert ist, sie Scheitelpunktattribute, die einem Scheitelpunkt zugeordnet sind, aus dem Speicher abzuholen; im Gegensatz dazu können Pakete, die einem Shader-Programm zugeordnet sind, an den SM **2814** übermittelt werden.

[0280] In mindestens einer Ausführungsform umfasst der SM **2814** ohne Einschränkung einen programmierbaren Streaming-Prozessor, der dazu konfiguriert ist, Aufgaben zu verarbeiten, die durch eine Anzahl von Threads dargestellt sind. In mindestens einer Ausführungsform ist der SM **2814** multithreaded und dazu konfiguriert, mehrere Threads (z. B. 32 Threads) aus einer bestimmten Gruppe von Threads gleichzeitig auszuführen, und setzt eine Einzelbefehls-Mehrfachdaten-(„SIMD“-)Architektur um, wobei jeder Thread in einer Gruppe von Threads (z. B. einem Warp) dazu konfiguriert ist, einen anderen Datensatz auf Grundlage desselben Anweisungssatzes zu verarbeiten. In mindestens einer Ausführungsform führen alle Threads in der Gruppe von Threads dieselben Anweisungen aus. In mindestens einer Ausführungsform setzt der SM **2814** eine Einzelbefehls-Mehrfachthread-(„SIMT“)-Architektur um, wobei jeder Thread in einer Gruppe von Threads dazu konfiguriert ist, einen anderen Datensatz auf Grundlage desselben Anweisungssatzes zu verarbeiten, wobei jedoch einzelne Threads in einer Gruppe von Threads während der Ausführung abweichen dürfen. In mindestens einer Ausführungsform werden ein Programmzähler, ein Aufrufstapel und ein Ausführungszustand für jeden Warp aufrechterhalten, was eine Gleichzeitigkeit zwischen Warps und eine serielle Ausführung innerhalb von Warps ermöglicht, wenn Threads innerhalb des Warps abweichen. In einer anderen Ausführungsform werden ein Programmzähler, ein Aufrufstapel und ein Ausführungszustand für jeden einzelnen Thread aufrechterhalten, was eine gleichberechtigte Gleichzeitigkeit zwischen allen Threads innerhalb und zwischen Warps ermöglicht. In mindestens einer Ausführungsform wird der Ausführungszustand für jeden einzelnen Thread aufrechterhalten, und Threads, die dieselben Anweisungen ausführen, können zusammengeführt und für eine bessere Effizienz parallel ausgeführt werden. Mindestens eine Ausführungsform des SM **2814** ist nachfolgend ausführlicher beschrieben.

[0281] In mindestens einer Ausführungsform stellt die MMU **2818** eine Schnittstelle zwischen dem GPC **2800** und der Speicherpartitionseinheit (z. B. der Partitionseinheit **2722** aus **Fig. 27**) bereit, und die MMU **2818** stellt eine Übersetzung virtueller Adressen in physische Adressen, einen Speicherschutz und eine Arbitrierung von Speicheranforderungen bereit. In mindestens einer Ausführungsform stellt die MMU **2818** einen oder mehrere

Adressenübersetzungspuffer („TLBs“) zum Durchführen einer Übersetzung von virtuellen Adressen in physische Adressen im Speicher bereit.

[0282] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. Details bezüglich der Ableitungs- und/oder Trainingslogik **615** werden nachfolgend in Verbindung mit den **Fig. 6A** und/oder 6B bereitgestellt. In mindestens einer Ausführungsform wird der Deep-Learning-Anwendungsprozessor verwendet, um ein Modell maschinellen Lernens, wie etwa ein neuronales Netzwerk, zu trainieren, um der GPC **2800** bereitgestellte Informationen vorherzusagen oder abzuleiten. In mindestens einer Ausführungsform wird die GPC **2800** verwendet, um Informationen auf Grundlage eines trainierten Modells maschinellen Lernens (z. B. neuronales Netzwerk) abzuleiten oder vorherzusagen, das von einem anderen Prozessor oder System oder von der GPC **2800** trainiert wurde. In mindestens einer Ausführungsform kann die GPC **2800** verwendet werden, um einen oder mehrere in dieser Schrift beschriebene Anwendungsfälle für neuronale Netzwerke durchzuführen.

[0283] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. In mindestens einer Ausführungsform kann diese Logik mit Komponenten dieser Figuren verwendet werden, um ein dreidimensionales Bild einer bestimmten Ansicht aus zweidimensionalen Bildern aus einer willkürlichen Anzahl von Kameras zu erzeugen.

[0284] **Fig. 29** veranschaulicht eine Speicherpartitionseinheit **2900** einer Parallelverarbeitungseinheit („PPU“) gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform beinhaltet die Speicherpartitionseinheit **2900** ohne Einschränkung eine Rasteroperations-(„ROP“-)Einheit **2902**; einen Zwischenspeicher **2904** der Ebene zwei („L2“); eine Speicherschnittstelle **2906**; und eine beliebige geeignete Kombination davon. In mindestens einer Ausführungsform ist die Speicherschnittstelle **2906** an den Speicher gekoppelt. In mindestens einer Ausführungsform kann die Speicherschnittstelle **2906** 32-, 64-, 128-, 1024-Bit-Datenbusse oder ähnliche Umsetzungen für eine Hochgeschwindigkeitsdatenübertragung umsetzen. In mindestens einer Ausführungsform integriert die PPU U-Speicherschnittstellen **2906**, eine Speicherschnittstelle **2906** pro Paar von Partitionseinheiten **2900**, wobei jedes Paar von Partitionseinheiten **2900** mit einer entsprechenden Speichervorrichtung verbunden ist. Zum Beispiel kann die PPU in mindestens einer Ausführungsform mit bis zu Y Speichervorrichtungen verbunden sein, wie etwa Speicherstapeln mit hoher Bandbreite oder synchronem dynamischen Direktzugriffsspeicher mit doppelter Grafikdatenrate, Version 5 (graphics double-data-rate, version 5, synchronous dynamic random access memory - „GDDR5 SDRAM“).

[0285] In mindestens einer Ausführungsform setzt die Speicherschnittstelle **2906** eine Speicherschnittstelle der zweiten Generation mit hoher Bandbreite (high bandwidth memory second generation - „HBM2“) um und Y ist gleich der Hälfte von U. In mindestens einer Ausführungsform befinden sich die HBM2-Speicherstapel auf demselben physischen Gehäuse wie die PPU, was erhebliche Leistungs- und Flächeneinsparungen im Vergleich zu herkömmlichen GDDR5-SDRAM-Systemen bereitstellt. In mindestens einer Ausführungsform beinhaltet jeder HBM2-Stapel ohne Einschränkung vier Speichermatrizen und Y ist gleich 4, wobei jeder HBM2-Stapel zwei 128-Bit-Kanäle pro Matrize für insgesamt 8 Kanäle und eine Datenbusbreite von 1024 Bit beinhaltet. In mindestens einer Ausführungsform unterstützt der Speicher Einzelfehlerkorrektur-Doppelfehlererfassungs-(Single-Error Correcting Double-Error Detecting - „SECDED“-)Fehlerkorrekturcode (Error Correction Code - „ECC“), um Daten zu schützen. In mindestens einer Ausführungsform bietet ECC eine höhere Zuverlässigkeit für Rechenanwendungen, die empfindlich gegenüber Datenkorruption sind.

[0286] In mindestens einer Ausführungsform setzt die PPU eine mehrstufige Speicherhierarchie um. In mindestens einer Ausführungsform unterstützt die Speicherpartitionseinheit **2900** einen einheitlichen Speicher, um einen einzelnen einheitlichen virtuellen Adressbereich für die Zentralverarbeitungseinheit („CPU“) und den PPU-Speicher bereitzustellen, wodurch die gemeinsame Nutzung von Daten zwischen virtuellen Speichersystemen ermöglicht wird. In mindestens einer Ausführungsform wird die Häufigkeit von Zugriffen durch eine PPU auf Speicher verfolgt, der sich auf anderen Prozessoren befindet, um sicherzustellen, dass Speicherseiten in den physischen Speicher der PPU verschoben werden, die häufiger auf Seiten zugreift. In mindestens einer Ausführungsform unterstützt die Hochgeschwindigkeits-GPU-Verbindung **2708** Adressübersetzungsdienste, die es der PPU ermöglichen, direkt auf die Seitentabellen einer CPU zuzugreifen, und den vollen Zugriff auf den CPU-Speicher durch die PPU bereitzustellen.

[0287] In mindestens einer Ausführungsform übertragen Kopierengines Daten zwischen mehreren PPUs oder zwischen PPUs und CPUs. In mindestens einer Ausführungsform können Kopierengines Seitenfehler für Adressen erzeugen, die nicht Seitentabellen zugeordnet sind, und die Speicherpartitionseinheit **2900** bedient

dann Seitenfehler, indem sie Adressen Seitentabellen zuordnet, wonach die Kopierengine die Übertragung durchführt. In mindestens einer Ausführungsform ist der Speicher für mehrere Kopierengineoperationen zwischen mehreren Prozessoren gepinnt (d. h. nicht auslagerbar), wodurch der verfügbare Speicher erheblich verringert wird. In mindestens einer Ausführungsform können bei Hardwareseitenfehlern Adressen an Kopierengines weitergegeben werden, ohne Rücksicht darauf, ob Speicherseiten resident sind, und der Kopierprozess ist transparent.

[0288] Daten aus dem Speicher **2704** aus **Fig. 27** oder ein anderer Systemspeicher wird von der Speicherpartitionseinheit **2900** abgerufen und im L2-Zwischenspeicher **2904** gespeichert, der sich auf dem Chip befindet und gemäß mindestens einer Ausführungsform von verschiedenen GPCs gemeinsam genutzt wird. Jede Speicherpartitionseinheit **2900** beinhaltet in mindestens einer Ausführungsform ohne Einschränkung mindestens einen Teil des L2-Zwischenspeichers, der einer entsprechenden Speichervorrichtung zugeordnet ist. In mindestens einer Ausführungsform werden Zwischenspeicher niedrigerer Ebene in verschiedenen Einheiten innerhalb der GPCs umgesetzt. In mindestens einer Ausführungsform kann jeder der SMs **2814** einen Zwischenspeicher der Ebene eins („LI“) umsetzen, wobei der L1-Zwischenspeicher ein privater Speicher ist, der einem bestimmten SM **2814** gewidmet ist, und Daten aus dem L2-Zwischenspeicher **2904** werden abgerufen und in jedem der L1-Zwischenspeicher zur Verarbeitung in Funktionseinheiten der SMs **2814** gespeichert. In mindestens einer Ausführungsform ist der L2-Zwischenspeicher **2904** an die Speicherschnittstelle **2906** und XBar **2720** gekoppelt.

[0289] Die ROP-Einheit **2902** führt in mindestens einer Ausführungsform Grafikrasteroperationen in Bezug auf die Bildpunktfarbe durch, wie etwa Farbkompression, Bildpunktmischung und mehr. Die ROP-Einheit **2902** setzt in mindestens einer Ausführungsform Tiefentests in Verbindung mit der Rasterengine **2808** um und empfängt eine Tiefe für einen Abtastort, der einem Bildpunktfragment von der Ausleseengine der Rasterengine **2808** zugeordnet ist. In mindestens einer Ausführungsform wird die Tiefe gegen eine entsprechende Tiefe in einem Tiefenpuffer für einen dem Fragment zugeordneten Abtastort getestet. Wenn in mindestens einer Ausführungsform das Fragment den Tiefentest für den Abtastort besteht, dann aktualisiert die ROP-Einheit **2902** den Tiefenpuffer und übermittelt ein Ergebnis des Tiefentests an die Rasterengine **2808**. Es versteht sich, dass die Anzahl der Partitionseinheiten **2900** von der Anzahl der GPCs verschieden sein kann und daher jede ROP-Einheit **2902** in mindestens einer Ausführungsform an jede der GPCs gekoppelt sein kann. In mindestens einer Ausführungsform verfolgt die ROP-Einheit **2902** Pakete, die von verschiedenen GPCs empfangen werden, und bestimmt, an welche ein von der ROP-Einheit **2902** erzeugtes Ergebnis durch die XBar **2720** weitergeleitet wird.

[0290] **Fig. 30** veranschaulicht einen Streaming-Multiprozessor („SM“) **3000** gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform ist der SM **3000** der SM **2814** aus **Fig. 28**. In mindestens einer Ausführungsform beinhaltet der SM **3000** ohne Einschränkung einen Anweisungszwischenspeicher **3002**; eine oder mehrere Planereinheiten **3004**; eine Registerdatei **3008**; einen oder mehrere Verarbeitungskerne („Kerne“) **3010**; eine oder mehrere Sonderfunktionseinheiten (special function units - „SFUs“) **3012**; eine oder mehrere Lade-/Speichereinheiten („LSUs“) **3014**; ein Verbindungsnetzwerk **3016**; einen gemeinsam genutzten Speicher/Ebene-Eins-Zwischenspeicher („L1“) **3018**; und eine beliebige geeignete Kombination davon. In mindestens einer Ausführungsform versendet eine Arbeitsverteilungseinheit Aufgaben zur Ausführung auf allgemeinen Verarbeitungsclustern („GPCs“) von Parallelverarbeitungseinheiten („PPUs“) und jede Aufgabe wird einem bestimmten Datenverarbeitungscluster („DPC“) innerhalb eines GPC zugewiesen und, wenn die Aufgabe einem Shader-Programm verbunden ist, wird die Aufgabe einem der SMs **3000** zugewiesen. In mindestens einer Ausführungsform empfängt die Planereinheit **3004** Aufgaben von der Arbeitsverteilungseinheit und verwaltet die Anweisungsplanung für einen oder mehrere Thread-Blöcke, die dem SM **3000** zugewiesen sind. In mindestens einer Ausführungsform plant die Planereinheit **3004** Thread-Blöcke zur Ausführung als Warps von parallelen Threads, wobei jedem Thread-Block mindestens ein Warp zugewiesen ist. In mindestens einer Ausführungsform führt jeder Warp Threads aus. In mindestens einer Ausführungsform verwaltet die Planereinheit **3004** eine Vielzahl von unterschiedlichen Thread-Blöcken, weist Warps unterschiedlichen Thread-Blöcken zu und versendet dann Anweisungen von einer Vielzahl von unterschiedlichen kooperativen Gruppen an verschiedene funktionelle Einheiten (z. B. die Verarbeitungskerne **3010**, SFUs **3012** und LSUs **3014**) während jedes Taktzyklus.

[0291] In mindestens einer Ausführungsform können sich kooperative Gruppen auf ein Programmiermodell zum Organisieren von Gruppen von kommunizierenden Threads beziehen, das es Entwicklern ermöglicht, die Granularität auszudrücken, mit der Threads kommunizieren, was den Ausdruck reicherer, effizienterer paralleler Zerlegungen ermöglicht. In mindestens einer Ausführungsform unterstützen kooperative Start-APIs die Synchronisation zwischen Thread-Blöcken zur Ausführung paralleler Algorithmen. In mindestens einer Aus-

führungsform stellen Anwendungen herkömmlicher Programmiermodelle ein einzelnes, einfaches Konstrukt zum Synchronisieren kooperierender Threads bereit: eine Barriere über alle Threads eines Threadblocks (z. B. syncthreads()-Funktion). In mindestens einer Ausführungsform können Programmierer jedoch Gruppen von Threads mit kleineren als Thread-Block-Granularitäten definieren und innerhalb definierter Gruppen synchronisieren, um eine größere Leistung, Ausgestaltungsflexibilität und Softwarewiederverwendung in Form von gemeinschaftlichen gruppenweiten Funktionsschnittstellen zu ermöglichen. In mindestens einer Ausführungsform ermöglichen es kooperative Gruppen Programmierern, Gruppen von Threads explizit bei Teilblock- (d. h. so klein wie ein einzelner Thread) und Multiblock-Granularitäten zu definieren und kollektive Operationen, wie etwa die Synchronisierung von Threads, in einer kooperativen Gruppe durchzuführen. In mindestens einer Ausführungsform unterstützt das Programmiermodell eine saubere Anordnung über Softwaregrenzen hinweg, so dass Bibliotheken und Dienstprogrammfunktionen innerhalb ihres lokalen Kontexts sicher synchronisieren können, ohne Annahmen über Annäherung treffen zu müssen. In mindestens einer Ausführungsform ermöglichen Primitive kooperativer Gruppen neue Muster kooperativer Parallelität, was ohne Einschränkung Erzeuger-Verbraucher-Parallelität, opportunistische Parallelität und globale Synchronisation über ein gesamtes Netz von Thread-Blöcken beinhaltet.

[0292] In mindestens einer Ausführungsform ist eine Versendeeinheit **3006** dazu konfiguriert, Anweisungen an eine oder mehrere der funktionellen Einheiten zu übermitteln, und die Planereinheit **3004** beinhaltet ohne Einschränkung zwei Versendeeinheiten **3006**, die es zwei unterschiedlichen Anweisungen desselben Warps ermöglichen, während jedes Taktzyklus versendet zu werden. In mindestens einer Ausführungsform beinhaltet jede Planereinheit **3004** eine einzelne Versendeeinheit **3006** oder zusätzliche Versendeeinheiten **3006**.

[0293] In mindestens einer Ausführungsform beinhaltet jeder SM **3000** in mindestens einer Ausführungsform ohne Einschränkung die Registerdatei **3008**, die einen Satz von Registern für funktionelle Einheiten des SM **3000** bereitstellt. In mindestens einer Ausführungsform wird die Registerdatei **3008** derartig zwischen jeder der funktionellen Einheiten aufgeteilt, dass jeder funktionellen Einheit ein dedizierter Teil der Registerdatei **3008** zugewiesen wird. In mindestens einer Ausführungsform ist die Registerdatei **3008** zwischen unterschiedlichen Warps aufgeteilt, die von SM **3000** ausgeführt werden, und die Registerdatei **3008** stellt einen temporären Speicher für Operanden bereit, die mit Datenpfaden von funktionellen Einheiten verbunden sind. In mindestens einer Ausführungsform umfasst jeder SM **3000** ohne Einschränkung eine Vielzahl von L Verarbeitungskernen **3010**. In mindestens einer Ausführungsform beinhaltet der SM **3000** ohne Einschränkung eine große Anzahl (z. B. **128** oder mehr) unterschiedlicher Verarbeitungskerne **3010**. In mindestens einer Ausführungsform beinhaltet jeder Verarbeitungskem **3010** in mindestens einer Ausführungsform ohne Einschränkung eine vollständig gepipelinete Verarbeitungseinheit mit einfacher Genauigkeit, doppelter Genauigkeit und/oder gemischter Genauigkeit, die ohne Einschränkung eine arithmetische Gleitkommalogikeinheit und eine arithmetische Ganzzahllogikeinheit beinhaltet. In mindestens einer Ausführungsform setzen arithmetische Gleitkommalogikeinheiten die Norm IEEE 754-2008 für Gleitkomma-Arithmetik um. In mindestens einer Ausführungsform beinhalten die Verarbeitungskerne **3010** ohne Einschränkung 64 Gleitkommakeme mit einfacher Genauigkeit (32 Bit), 64 Ganzzahlkerne, 32 Gleitkommakeme mit doppelter Genauigkeit (64 Bit) und 8 Tensorkerne.

[0294] Tensorkerne sind dazu konfiguriert, Matrixoperationen gemäß mindestens einer Ausführungsform durchzuführen. In mindestens einer Ausführungsform sind ein oder mehrere Tensorkerne in den Verarbeitungskernen **3010** beinhaltet. In mindestens einer Ausführungsform sind Tensorkerne dazu konfiguriert, eine Deep-Learning-Matrixarithmetik durchzuführen, wie etwa Faltungsoperationen für das Training und die Ableitung von neuronalen Netzen. In mindestens einer Ausführungsform arbeitet jeder Tensorkern auf einer 4x4-Matrix und führt eine Matrixmultiplikations- und -akkumulationsoperation D = A X B + C durch, wobei A, B, C und D 4x4-Matrizen sind.

[0295] In mindestens einer Ausführungsform sind die Matrixmultiplikationseingänge A und B 16-Bit-Gleitkomma-Matrizen und die Akkumulationsmatrizen C und D sind 16-Bit-Gleitkomma- oder 32-Bit-Gleitkomma-Matrizen. In mindestens einer Ausführungsform arbeiten Tensorkerne mit 16-Bit-Gleitkomma-Eingangsdaten mit 32-Bit-Gleitkomma-Akkumulation. In mindestens einer Ausführungsform verwendet die 16-Bit-Gleitkommamultiplikation 64 Operationen und führt zu einem Produkt mit voller Genauigkeit, das dann unter Verwendung einer 32-Bit-Gleitkommaaddition mit anderen Zwischenprodukten für eine 4x4x4-Matrixmultiplikation akkumuliert wird. In mindestens einer Ausführungsform werden Tensorkerne verwendet, um viel größere zweidimensionale oder höherdimensionale Matrixoperationen durchzuführen, die aus diesen kleineren Elementen aufgebaut sind. In mindestens einer Ausführungsform legt eine API, wie etwa die CUDA 9 C++-API, spezielle Matrixlade-, Matrixmultiplizier- und -akkumulations- und Matrixspeicheroperationen offen, um Tensorkerne aus einem CUDA-C++-Programm effizient zu verwenden. In mindestens einer Ausführungsform geht die Warp-Ebenen-Schnittstelle auf CUDA-Ebene von 16×16 großen Matrizen aus, die alle 32 Warp-Threads überspannen.

[0296] In mindestens einer Ausführungsform umfasst jeder SM **3000** ohne Einschränkung M SFUs **3012**, die spezielle Funktionen (z. B. Attributbewertung, reziproke Quadratwurzel usw.) ausführen. In mindestens einer Ausführungsform beinhalten die SFUs **3012** ohne Einschränkung eine Baumdurchquerungseinheit, die dazu konfiguriert ist, eine hierarchische Baumdatenstruktur zu durchqueren. In mindestens einer Ausführungsform beinhalten die SFUs **3012** ohne Einschränkung eine Textureinheit, die dazu konfiguriert ist, Texturzuordnungsfilteroperationen durchzuführen. In mindestens einer Ausführungsform sind Textureinheiten dazu konfiguriert, Texturkarten (z. B. ein 2D-Array von Texeln) aus dem Speicher zu laden und Texturkarten abzutasten, um abgetastete Texturwerte zur Verwendung in von dem SM **3000** ausgeführten Shaderprogrammen zu erzeugen. In mindestens einer Ausführungsform werden Texturkarten im gemeinsam genutzten Speicher/L1-Zwischenspeicher **3018** gespeichert. In mindestens einer Ausführungsform setzen Textureinheiten gemäß mindestens einer Ausführungsform Texturoperationen, wie etwa Filteroperationen, unter Verwendung von Mip-Karten (z.B. Texturkarten mit veränderlichen Detailgraden) um. In mindestens einer Ausführungsform beinhaltet jeder SM **3000** ohne Einschränkung zwei Textureinheiten.

[0297] Jeder SM **3000** umfasst ohne Einschränkung N LSUs **3014**, die in mindestens einer Ausführungsform Lade- und Speicheroperationen zwischen dem gemeinsam genutzten Speicher/L1-Zwischenspeicher **3018** und der Registerdatei **3008** umsetzen. Jeder SM **3000** beinhaltet ohne Einschränkung ein Verbindungsnetzwerk **3016**, das in mindestens einer Ausführungsform jede der Funktionseinheiten mit der Registerdatei **3008** und die LSU **3014** mit der Registerdatei **3008** und dem gemeinsam genutzten Speicher/L1-Zwischenspeicher **3018** verbindet. In mindestens einer Ausführungsform ist das Verbindungsnetzwerk **3016** eine Crossbar, die dazu konfiguriert sein kann, eine beliebige der funktionellen Einheiten mit einem beliebigen der Register in der Registerdatei **3008** zu verbinden und die LSUs **3014** mit der Registerdatei **3008** und Speicherstellen im gemeinsam genutzten Speicher/L1-Zwischenspeicher **3018** zu verbinden.

[0298] In mindestens einer Ausführungsform ist der gemeinsam genutzte Speicher/L1-Zwischenspeicher **3018** ein Array von Speicher auf dem Chip, der in mindestens einer Ausführungsform Datenspeicherung und Kommunikation zwischen dem SM **3000** und der Primitivengine und zwischen den Threads in dem SM **3000** ermöglicht. In mindestens einer Ausführungsform umfasst der gemeinsam genutzte Speicher/L1-Zwischenspeicher **3018** ohne Einschränkung eine Speicherkapazität von **128** KB und befindet sich im Pfad vom SM **3000** zur Partitionseinheit. In mindestens einer Ausführungsform wird der gemeinsam genutzte Speicher/L1-Zwischenspeicher **3018** in mindestens einer Ausführungsform verwendet, um Lese- und Schreibvorgänge zwischenzuspeichern. In mindestens einer Ausführungsform sind einer oder mehrere des gemeinsam genutzten Speichers/L1-Zwischenspeichers **3018**, L2-Zwischenspeichers und des Speichers Sicherungsspeicher.

[0299] Das Kombinieren von Datenzwischenspeicher- und der Funktion des gemeinsam genutzten Speichers in einem einzelnen Speicherblock stellt in mindestens einer Ausführungsform eine verbesserte Leistung für beide Arten von Speicherzugriffen bereit. In mindestens einer Ausführungsform wird die Kapazität von Programmen, die keinen gemeinsam genutzten Speicher verwenden, als ein Zwischenspeicher verwendet oder kann als ein Zwischenspeicher verwendet werden, wie etwa, wenn der gemeinsam genutzte Speicher dazu konfiguriert ist, die Hälfte der Kapazität zu verwenden, und Textur- und Lade-/Speicheroperationen können die verbleibende Kapazität verwenden. Die Integration innerhalb des gemeinsam genutzten Speichers/L1-Zwischenspeichers **3018** ermöglicht es gemäß mindestens einer Ausführungsform dem gemeinsam genutzten Speicher/L1-Zwischenspeicher **3018**, als ein Hochdurchsatzkanal für das Streamen von Daten zu fungieren, während häufig wiederverwendeten Daten gleichzeitig ein Zugriff mit hoher Bandbreite und geringer Latenz bereitgestellt wird. In mindestens einer Ausführungsform kann, wenn sie für eine parallele Berechnung für allgemeine Zwecke konfiguriert ist, eine einfachere Konfiguration im Vergleich zur Grafikverarbeitung verwendet werden. In mindestens einer Ausführungsform werden Grafikverarbeitungseinheiten mit feststehender Funktion umgangen, wodurch ein viel einfacheres Programmiermodell erzeugt wird. In einer Konfiguration für parallele Berechnungen für allgemeine Zwecke weist die Arbeitsverteilungseinheit in mindestens einer Ausführungsform Blöcke von Threads direkt zu DPCs zu und verteilt sie. In mindestens einer Ausführungsform führen Threads in einem Block dasselbe Programm unter Verwendung einer eindeutigen Thread-ID bei der Berechnung aus, um sicherzustellen, dass jeder Thread eindeutige Ergebnisse erzeugt, unter Verwendung von SM **3000**, um Programme auszuführen und Berechnungen durchzuführen, gemeinsam genutztem Speicher/L1-Zwischenspeicher **3018**, um zwischen Threads zu kommunizieren, und LSU **3014**, um globalen Speicher über den gemeinsam genutzten Speicher/L1-Zwischenspeicher **3018** und die Speicherpartitionseinheit auszulesen und zu schreiben. In mindestens einer Ausführungsform schreibt der SM **3000**, wenn er für eine Parallelberechnung für allgemeine Zwecke konfiguriert ist, Befehle, welche die Planereinheit **3004** verwenden kann, um neue Arbeit auf den DPCs zu starten.

[0300] In mindestens einer Ausführungsform ist die PPU in Folgendem beinhaltet: einem Desktop-Computer, einem Laptop-Computer, einem Tablet-Computer, Servern, Supercomputern, einem Smartphone (z. B. einer drahtlosen tragbaren Vorrichtung), einem persönlichen digitalen Assistenten („PDA“), einer Digitalkamera, einem Fahrzeug, einer am Kopf befestigten Anzeige, einer tragbaren elektronischen Vorrichtung und mehr. In mindestens einer Ausführungsform ist die PPU auf einem einzelnen Halbleitersubstrat realisiert. In mindestens einer Ausführungsform ist die PPU in einem System auf einem Chip („SoC“) zusammen mit einer oder mehreren anderen Vorrichtungen beinhaltet, wie etwa zusätzliche PPUs, Speicher, einer CPU mit verringertem Anweisungssatz („RISC“), einer Speicherverwaltungseinheit („MMU“), einem Digital-AnalogWandler („DAC“) und dergleichen.

[0301] In mindestens einer Ausführungsform kann die PPU auf einer Grafikkarte beinhaltet sein, die ein oder mehrere Speichervorrichtungen beinhaltet. Eine Grafikkarte kann dazu konfiguriert sein, mit einem PCIe-Steckplatz auf einem Motherboard eines Desktop-Computers verbunden zu sein. In mindestens einer Ausführungsform kann die PPU eine integrierte Grafikverarbeitungseinheit (integrated graphics processing unit - „iGPU“) sein, die im Chipsatz des Motherboards beinhaltet ist.

[0302] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. Details bezüglich der Ableitungs- und/oder Trainingslogik **615** werden nachfolgend in Verbindung mit den **Fig. 6A** und/oder 6B bereitgestellt. In mindestens einer Ausführungsform wird der Deep-Learning-Anwendungsprozessor verwendet, um ein Modell maschinellen Lernens, wie etwa ein neuronales Netzwerk, zu trainieren, um dem SM **3000** bereitgestellte Informationen vorherzusagen oder abzuleiten. In mindestens einer Ausführungsform wird der SM **3000** verwendet, um Informationen auf Grundlage eines trainierten Modells maschinellen Lernens (z. B. neuronales Netzwerk) abzuleiten oder vorherzusagen, das von einem anderen Prozessor oder System oder von dem SM **3000** trainiert wurde. In mindestens einer Ausführungsform kann der SM **3000** verwendet werden, um einen oder mehrere in dieser Schrift beschriebene Anwendungsfälle für neuronale Netzwerke durchzuführen.

[0303] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. In mindestens einer Ausführungsform kann diese Logik mit Komponenten dieser Figuren verwendet werden, um ein dreidimensionales Bild einer bestimmten Ansicht aus zweidimensionalen Bildern aus einer willkürlichen Anzahl von Kameras zu erzeugen.

[0304] In mindestens einer Ausführungsform kann sich eine einzelne Halbleiterplattform auf eine(n) einzige(n) einheitliche(n) integrierte(n) Schaltung oder Chip auf Halbleiterbasis beziehen. In mindestens einer Ausführungsform können Mehrchipmodule mit erhöhter Anbindung verwendet werden, die einen Betrieb auf dem Chip simulieren und wesentliche Verbesserungen gegenüber der Nutzung einer herkömmlichen Zentralverarbeitungseinheit („CPU“) und einer Busumsetzung bewirken. In mindestens einer Ausführungsform können verschiedene Module auch getrennt oder in verschiedenen Kombinationen von Halbleiterplattformen nach den Wünschen des Benutzers angeordnet sein.

[0305] In mindestens einer Ausführungsform werden Computerprogramme in Form von maschinenlesbarem ausführbarem Code oder Computersteuerlogikalgorithmen im Hauptspeicher **1004** und/oder im Sekundärspeicher gespeichert. Computerprogramme ermöglichen, wenn sie von einem oder mehreren Prozessoren ausgeführt werden, dem System **1000** gemäß mindestens einer Ausführungsform, verschiedene Funktionen auszuführen. In mindestens einer Ausführungsform sind der Speicher **1004**, der Speicher und/oder jeder andere Speicher mögliche Beispiele für computerlesbare Medien. In mindestens einer Ausführungsform kann sich sekundärer Speicher auf ein(e) beliebige(s) geeignete(s) Speichervorrichtung oder -system beziehen, wie etwa ein Festplattenlaufwerk und/oder ein Wechselspeicherlaufwerk, das ein Diskettenlaufwerk, ein Magnetbandlaufwerk, ein Compact-Disk-Laufwerk, digitales vielseitiges Plattenlaufwerk („DVD“), Aufzeichnungsvorrichtung, universeller serieller Bus („USB“) Flash-Speicher usw. darstellt. In mindestens einer Ausführungsform werden Architektur und/oder Funktion verschiedener früherer Figuren im folgenden Kontext umgesetzt: die CPU **1002**; das Parallelverarbeitungssystem **1012**; eine integrierte Schaltung, die mindestens einen Teil der Fähigkeiten beider CPU **1002** erfüllen kann; ein Parallelverarbeitungssystem **1012**; ein Chipsatz (z. B. eine Gruppe integrierter Schaltungen, die ausgestaltet wurden, um zum Durchzuführen zusammengehöriger Funktionen zu funktionieren und als eine Einheit verkauft zu werden usw.); und jede geeignete Kombination von einer oder mehreren integrierten Schaltungen.

[0306] In mindestens einer Ausführungsform werden Architektur und/oder Funktion verschiedener vorheriger Figuren im Kontext eines allgemeinen Computersystems, eines Leiterplattensystems, eines Spielkonsolensys-

tems für Unterhaltungszwecke, eines anwendungsspezifischen Systems und mehr umgesetzt. In mindestens einer Ausführungsform kann das Computersystem **1000** die folgende Form annehmen: eines Desktop-Computers, eines Laptop-Computers, eines Tablet-Computers, Servers, Supercomputers, eines Smartphones (z. B. einer drahtlosen tragbaren Vorrichtung), eines persönlichen digitalen Assistenten („PDA"), einer Digitalkamera, eines Fahrzeugs, einer am Kopf befestigten Anzeige, einer tragbaren elektronischen Vorrichtung, eines Mobiltelefons, eines Fernsehers, einer Workstation, von Spielkonsolen, eines eingebetteten Systems und/oder jeder anderen Art von Logik.

[0307] In mindestens einer Ausführungsform beinhaltet das Parallelverarbeitungssystem **1012** ohne Einschränkung eine Vielzahl von Parallelverarbeitungseinheiten („PPUs") **1014** und zugehörige Speicher **1016**. In mindestens einer Ausführungsform sind die PPUs **1014** mit einem Hostprozessor oder anderen Peripherievorrichtungen über eine Verbindung **1018** und einen Switch **1020** oder einen Multiplexer verbunden. In mindestens einer Ausführungsform verteilt das Parallelverarbeitungssystem **1012** Berechnungsaufgaben über die PPUs **1014**, die parallelisierbar sein können - zum Beispiel als Teil der Verteilung von Berechnungsaufgaben über mehrere Thread-Blöcke von Grafikverarbeitungseinheiten („GPUs"). In mindestens einer Ausführungsform wird der Speicher von einigen oder allen PPUs **1014** gemeinsam genutzt und ist für diese zugänglich (z. B. für den Lese- und/oder Schreibzugriff), obwohl ein derartiger gemeinsam genutzter Speicher Leistungseinbußen in Bezug auf die Verwendung von lokalem Speicher und in einer PPU **1014** residenten Registern nach sich ziehen kann. In mindestens einer Ausführungsform wird der Betrieb der PPUs **1014** durch die Verwendung eines Befehls wie etwa _syncthreads() synchronisiert, wobei alle Threads in einem Block (z. B. über mehrere PPUs **1014** ausgeführt) einen bestimmten Punkt der Codeausführung erreichen, bevor sie fortfahren.

VIRTUALISIERTE RECHENPLATTFORM

[0308] Es werden Ausführungsformen in Bezug auf eine virtualisierte Rechenplattform für fortschrittliches Rechnen offenbart, wie etwa Bildableitung und Bildverarbeitung. Unter Bezugnahme auf **Fig. 31** ist ein beispielhaftes Datenflussdiagramm für einen Prozess **3100** zum Erzeugen und Einsetzen einer Bildverarbeitungs- und Ableitungspipeline gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform kann der Prozess **3100** zur Verwendung mit Bildgebungsvorrichtungen, Verarbeitungsvorrichtungen, Genomikvorrichtungen, Gensequenzierungsvorrichtungen, Radiologievorrichtungen und/oder anderen Vorrichtungsarten in einer oder mehreren Einrichtungen **3102** eingesetzt werden, wie etwa medizinischen Einrichtungen, Krankenhäusern, Gesundheitsinstituten, Kliniken, Forschungs- oder Diagnoselaboren usw. In mindestens einer Ausführungsform kann der Prozess **3100** eingesetzt werden, um eine Genomanalyse und Ableitung bei Sequenzierungsdaten durchzuführen. Beispiele für genomische Analysen, die unter Verwendung der in dieser Schrift beschriebenen Systeme und Verfahren durchgeführt werden können, beinhalten ohne Einschränkung Variantenbestimmung, Mutationserfassung und Genexpressionsquantifizierung. Der Prozess **3100** kann innerhalb eines Trainingssystems **3104** und/oder eines Einsatzsystems **3106** ausgeführt werden. In mindestens einer Ausführungsform kann das Trainingssystem **3104** verwendet werden, um das Training, den Einsatz und die Umsetzung von Modellen des maschinellen Lernens (z. B. neuronale Netzwerke, Objekterfassungsalgorithmen, Algorithmen des maschinellen Sehens usw.) zur Verwendung im Einsatzsystem **3106** durchzuführen. In mindestens einer Ausführungsform kann das Einsatzsystem **3106** dazu konfiguriert sein, Verarbeitungs- und Berechnungsressourcen zwischen einer verteilten Computerumgebung auszulagern, um die Infrastrukturanforderungen in der Einrichtung **3102** zu verringern. In mindestens einer Ausführungsform kann das Einsatzsystem **3106** eine optimierte Plattform zum Auswählen, Anpassen und Umsetzen virtueller Instrumente zur Verwendung mit Bildgebungsvorrichtungen (z. B. MRT, CT, Röntgen, Ultraschall usw.) oder Sequenziervorrichtungen in der Einrichtung **3102** bereitstellen. In mindestens einer Ausführungsform können virtuelle Instrumente softwaredefinierte Anwendungen zum Durchführen einer oder mehrerer Verarbeitungsoperationen in Bezug auf Bildgebungsdaten beinhalten, die von Bildgebungsvorrichtungen, Sequenziervorrichtungen, Radiologievorrichtungen und/oder anderen Vorrichtungsarten erzeugt werden. In mindestens einer Ausführungsform können eine oder mehrere Anwendungen in einer Pipeline Dienste (z.B. Ableitung, Visualisierung, Berechnung, KI usw.) des Einsatzsystems **3106** während der Ausführung von Anwendungen verwenden oder aufrufen.

[0309] In mindestens einer Ausführungsform können einige der Anwendungen, die in fortgeschrittenen Verarbeitungs- und Ableitungspipelines verwendet werden, Modelle des maschinellen Lernens oder andere KI verwenden, um einen oder mehrere Verarbeitungsschritte auszuführen. In mindestens einer Ausführungsform können Modelle des maschinellen Lernens in der Einrichtung **3102** unter Verwendung von Daten **3108** (wie etwa Bildgebungsdaten), die in der Einrichtung **3102** erzeugt wurden (und auf einem oder mehreren Bildarchivierungs- und Kommunikationssystem-(picture archiving and communication system - PACS-)Servern in der Einrichtung **3102** gespeichert sind), trainiert werden, können unter Verwendung von Bildgebungs- oder Sequenzierungsdaten **3108** von (einer) anderen Einrichtung(en) (z. B. einem/r anderen Krankenhaus, Labor,

Klinik usw.) oder einer Kombination davon trainiert werden. In mindestens einer Ausführungsform kann das Trainingssystem **3104** verwendet werden, um Anwendungen, Dienste und/oder andere Ressourcen zum Erzeugen von funktionierenden, einsetzbaren Modellen des maschinellen Lernens für das Einsatzsystem **3106** bereitzustellen.

[0310] In mindestens einer Ausführungsform kann die Modellregistrierungsdatenbank **3124** durch einen Objektspeicher unterstützt werden, der Versionsverwaltung und Objektmetadaten unterstützen kann. In mindestens einer Ausführungsform kann auf den Objektspeicher beispielsweise über eine mit Cloud-Speicher (z. B. die Cloud **3226** aus **Fig. 32**) kompatible Anwendungsprogrammierschnittstelle (API) innerhalb einer Cloud-Plattform zugegriffen werden. In mindestens einer Ausführungsform können Modelle des maschinellen Lernens in der Modellregistrierungsdatenbank **3124** von Entwicklern oder Partnern eines Systems, das mit einer API interagiert, hochgeladen, aufgelistet, modifiziert oder gelöscht werden. In mindestens einer Ausführungsform kann eine API Zugriff auf Verfahren bereitstellen, die es Benutzern mit entsprechenden Berechtigungsnachweisen ermöglichen, Modelle derartig Anwendungen zuzuordnen, dass Modelle als Teil der Ausführung von containerisierten Instanziierungen von Anwendungen ausgeführt werden können.

[0311] In mindestens einer Ausführungsform kann die Trainingspipeline **3204** (**Fig. 32**) ein Szenario beinhalten, in dem die Einrichtung **3102** ihr eigenes Modell des maschinellen Lernens trainiert oder über ein bestehendes Modell des maschinellen Lernens verfügt, das optimiert oder aktualisiert werden muss. In mindestens einer Ausführungsform können Bildgebungsdaten **3108** empfangen werden, die von (einer) Bildgebungsvorrichtung(en), Sequenziervorrichtungen und/oder anderen Vorrichtungsarten erzeugt werden. In mindestens einer Ausführungsform kann, sobald die Bildgebungsdaten **3108** empfangen wurden, die KI-unterstützte Anmerkung **3110** verwendet werden, um das Erzeugen von Anmerkungen entsprechend den Bildgebungsdaten **3108** zu unterstützen, die als Ground-Truth-Daten für ein Modell des maschinellen Lernens verwendet werden sollen. In mindestens einer Ausführungsform kann die KI-unterstützte Anmerkung **3110** ein oder mehrere Modelle des maschinellen Lernens (z. B. neuronale Faltungsnetzwerke (convolutional neural networks - CNNs)) beinhalten, die trainiert werden können, um Anmerkungen zu erzeugen, die bestimmten Arten von Bildgebungsdaten **3108** (z. B. von bestimmten Vorrichtungen) und/oder bestimmten Arten von Anomalien in Bildgebungsdaten **3108** entsprechen. In mindestens einer Ausführungsform können die KI-unterstützten Anmerkungen **3110** dann direkt verwendet werden oder können unter Verwendung eines Anmerkungswerkzeugs (z. B. von einem Forscher, einem Kliniker, einem Arzt, einem Wissenschaftler usw.) angepasst oder fein abgestimmt werden, um Ground-Truth-Daten zu erzeugen. In mindestens einer Ausführungsform können in einigen Beispielen gekennzeichnete Klinikdaten **3112** (z.B. Anmerkungen, die von einem Kliniker, Arzt, Wissenschaftler, Techniker usw. bereitgestellt wurden) als Ground-Truth-Daten zum Trainieren eines Modells des maschinellen Lernens verwendet werden. In mindestens einer Ausführungsform können KI-unterstützte Anmerkungen **3110**, gekennzeichnete Klinikdaten **3112** oder eine Kombination davon als Ground-Truth-Daten zum Trainieren eines Modells des maschinellen Lernens verwendet werden. In mindestens einer Ausführungsform kann ein trainiertes Modell des maschinellen Lernens als Ausgabemodell **3116** bezeichnet werden und kann von dem Einsatzsystem **3106** wie in dieser Schrift beschrieben verwendet werden.

[0312] In mindestens einer Ausführungsform kann die Trainingspipeline **3204** (**Fig. 32**) ein Szenario beinhalten, bei dem die Einrichtung **3102** ein Modell des maschinellen Lernens zur Verwendung beim Ausführen einer oder mehrerer Verarbeitungsaufgaben für eine oder mehrere Anwendungen im Einsatzsystem **3106** benötigt, die Einrichtung **3102** aber derzeit möglicherweise nicht über ein derartiges Modell des maschinellen Lernens verfügt (oder möglicherweise kein Modell aufweist, das für derartige Zwecke optimiert, effizient oder effektiv ist). In mindestens einer Ausführungsform kann ein vorhandenes Modell des maschinellen Lernens aus einer Modellregistrierungsdatenbank **3124** ausgewählt werden. In mindestens einer Ausführungsform kann die Modellregistrierungsdatenbank **3124** Modelle des maschinellen Lernens beinhalten, die trainiert sind, um eine Reihe von unterschiedlichen Ableitungsaufgaben an Bildgebungsdaten durchzuführen. In mindestens einer Ausführungsform können die Modelle des maschinellen Lernens in der Modellregistrierungsdatenbank **3124** an Bildgebungsdaten von anderen Einrichtungen als der Einrichtung **3102** (z. B. entfernt gelegenen Einrichtungen) trainiert worden sein. In mindestens einer Ausführungsform können Modelle des maschinellen Lernens an Bilddaten von einem Standort, zwei Standorten oder einer beliebigen Anzahl von Standorten trainiert worden sein. In mindestens einer Ausführungsform kann das Training beim Training mit Bilddaten von einem konkreten Standort an diesem Standort stattfinden oder mindestens auf eine Weise, welche die Vertraulichkeit von Bilddaten schützt oder die Übertragung von Bilddaten außerhalb von Räumlichkeiten einschränkt (z. B. um HIPAA-Vorschriften, Datenschutzbestimmungen usw. einzuhalten). In mindestens einer Ausführungsform kann ein Modell des maschinellen Lernens, sobald ein Modell an einem Ort trainiert - oder teilweise trainiert wird, der Modellregistrierungsdatenbank **3124** hinzugefügt werden. In mindestens einer Ausführungsform kann ein Modell des maschinellen Lernens dann in einer beliebigen Anzahl anderer Einrichtungen neu trainiert oder

aktualisiert werden, und ein neu trainiertes oder aktualisiertes Modell kann in der Modellregistrierungsdatenbank **3124** verfügbar gemacht werden. In mindestens einer Ausführungsform kann dann ein Modell des maschinellen Lernens aus der Modellregistrierungsdatenbank **3124** ausgewählt - und als Ausgabemodell **3116** bezeichnet - und im Einsatzsystem **3106** verwendet werden, um eine oder mehrere Verarbeitungsaufgaben für eine oder mehrere Anwendungen eines Einsatzsystems durchzuführen.

[0313] In mindestens einer Ausführungsform kann die Trainingspipeline **3204** (**Fig. 32**) ein Szenario beinhalten, bei dem die Einrichtung **3102** ein Modell des maschinellen Lernens zur Verwendung beim Ausführen einer oder mehrerer Verarbeitungsaufgaben für eine oder mehrere Anwendungen im Einsatzsystem **3106** erfordert, die Einrichtung **3102** aber derzeit möglicherweise nicht über ein derartiges Modell des maschinellen Lernens verfügt (oder möglicherweise kein Modell aufweist, das für derartige Zwecke optimiert, effizient oder effektiv ist). In mindestens einer Ausführungsform kann ein aus der Modellregistrierungsdatenbank **3124** ausgewähltes Modell des maschinellen Lernens aufgrund von Unterschieden in Populationen, genetischen Variationen, Robustheit von Trainingsdaten, die zum Trainieren eines Modells des maschinellen Lernens verwendet werden, Diversität bei Anomalien von Trainingsdaten und/oder anderen Probleme mit Trainingsdaten nicht für Bildgebungsdaten **3108**, die in der Einrichtung **3102** erzeugt werden, fein abgestimmt oder optimiert werden. In mindestens einer Ausführungsform kann die KI-unterstützte Anmerkung **3110** verwendet werden, um das Erzeugen von Anmerkungen entsprechend den Bildgebungsdaten **3108** zu unterstützen, die als Ground-Truth-Daten für das Neutrainieren oder Aktualisieren eines Modells des maschinellen Lernens verwendet werden sollen. In mindestens einer Ausführungsform können gekennzeichnete Klinikdaten **3112** (z.B. Anmerkungen, die von einem Kliniker, Arzt, Wissenschaftler usw. bereitgestellt wurden) als Ground-Truth-Daten zum Trainieren eines Modells des maschinellen Lernens verwendet werden. In mindestens einer Ausführungsform kann das Neutrainieren oder Aktualisieren eines Modells des maschinellen Lernens als Modelltraining **3114** bezeichnet werden. In mindestens einer Ausführungsform kann das Modelltraining **3114** - z.B. KI-unterstützte Anmerkungen **3110**, gekennzeichnete Klinikdaten **3112** oder eine Kombination davon als Ground-Truth-Daten zum Neutrainieren und Aktualisieren eines Modells des maschinellen Lernens verwenden. In mindestens einer Ausführungsform kann ein trainiertes Modell des maschinellen Lernens als Ausgabemodell **3116** bezeichnet werden und kann von dem Einsatzsystem **3106** wie in dieser Schrift beschrieben verwendet werden.

[0314] In mindestens einer Ausführungsform kann das Einsatzsystem **3106** Software **3118**, Dienste **3120**, Hardware **3122** und/oder andere Komponenten, Merkmale und Funktion beinhalten In mindestens einer Ausführungsform kann das Einsatzsystem **3106** derartig einen Software-„Stapel“ beinhalten, dass die Software **3118** auf den Diensten **3120** aufgebaut sein kann und die Dienste **3120** verwenden kann, um einige oder alle Verarbeitungsaufgaben auszuführen, und die Dienste **3120** und die Software **3118** können auf der Hardware **3122** aufgebaut sein und die Hardware **3122** verwenden, um Verarbeitungs-, Speicher- und/oder andere Berechnungsaufgaben des Einsatzsystems **3106** auszuführen. In mindestens einer Ausführungsform kann die Software **3118** eine beliebige Anzahl unterschiedlicher Container umfassen, wobei jeder Container eine Instanziierung einer Anwendung ausführen kann. In mindestens einer Ausführungsform kann jede Anwendung eine oder mehrere Verarbeitungsaufgaben in einer fortgeschrittenen Verarbeitungs- und Ableitungspipeline ausführen (z. B. Ableitung, Objekterfassung, Merkmalserfassung, Segmentierung, Bildverbesserung, Kalibrierung usw.). In mindestens einer Ausführungsform kann es für jede Art von Bildgebungsvorrichtung (z. B. CT, MRT, Röntgen, Ultraschall, Sonographie, Echokardiographie usw.), Sequenziervorrichtung, Radiologievorrichtung, Genomikvorrichtung usw. eine beliebige Anzahl von Containern geben, die eine Datenverarbeitungsaufgabe in Bezug auf Bildgebungsdaten **3108** (oder andere Datenarten, wie die in dieser Schrift beschriebenen) ausführen können, die von einer Vorrichtung erzeugt werden. In mindestens einer Ausführungsform kann eine erweiterte Verarbeitungs- und Ableitungspipeline auf Grundlage von Auswahlen von unterschiedlichen Containern definiert werden, die für die Verarbeitung von Bildgebungsdaten **3108** zusätzlich zu Containern gewünscht oder erforderlich sind, die Bildgebungsdaten zur Verwendung durch jeden Container und/oder zur Verwendung durch die Einrichtung **3102** nach der Verarbeitung durch eine Pipeline empfangen und konfigurieren (z. B. um Ausgaben zurück in eine verwendbare Datenart umzuwandeln, wie etwa Daten der digitalen Bildgebung und Kommunikation in der Medizin (digital imaging and communications in medicine - DICOM), Daten des Radiologieinformationssystems (RIS), Daten des klinischen Informationssystems (clinical information system - CIS), Daten des Fernprozessaufrufs (remote procedure call - RPC), Daten, die im Wesentlichen einer Darstellungszustandsübertragungs-(representation state transfer-REST-)Schnittstelle entsprechen, Daten, die im Wesentlichen einer dateibasierten Schnittstelle und/oder Rohdaten zur Speicherung und Anzeige in der Einrichtung **3102** entsprechen). In mindestens einer Ausführungsform kann eine Kombination von Containern innerhalb der Software **3118** (die z. B. eine Pipeline bilden) als ein virtuelles Instrument (wie in dieser Schrift ausführlicher beschrieben) bezeichnet sein, und ein virtuelles Instrument kann Dienste **3120** und Hardware **3122** nutzen, um einige oder alle Verarbeitungsaufgaben von in Containern instanziierten Anwendungen auszuführen.

[0315] In mindestens einer Ausführungsform kann eine Datenverarbeitungspipeline Eingabedaten (z. B. Bilddaten **3108**) in einem DICOM-, RIS-, CIS-, REST-konformen, RPC-, Roh- und/oder anderen Format als Reaktion auf eine Ableitungsanfrage (z. B. eine Anfrage von einem Benutzer des Einsatzsystems **3106**, wie etwa einem Kliniker, einem Arzt, einem Radiologen usw.) empfangen. In mindestens einer Ausführungsform können Eingabedaten repräsentativ für ein oder mehrere Bilder, Videos und/oder andere Datendarstellungen sein, die von einer oder mehreren Bildgebungsvorrichtungen, Sequenziervorrichtungen, Radiologievorrichtungen, Genomikvorrichtungen und/oder anderen Vorrichtungsarten erzeugt werden. In mindestens einer Ausführungsform können Daten als Teil einer Datenverarbeitungspipeline einer Vorverarbeitung unterzogen werden, um Daten für die Verarbeitung durch eine oder mehrere Anwendungen vorzubereiten. In mindestens einer Ausführungsform kann eine Nachbearbeitung an einer Ausgabe einer oder mehrerer Ableitungsaufgaben oder anderer Verarbeitungsaufgaben einer Pipeline durchgeführt werden, um Ausgabedaten für eine nächste Anwendung vorzubereiten und/oder Ausgabedaten für die Übermittlung und/oder Verwendung durch einen Benutzer vorzubereiten (z. B. als eine Antwort auf eine Ableitungsanfrage). In mindestens einer Ausführungsform können Ableitungsaufgaben durch ein oder mehrere Modelle des maschinellen Lernens durchgeführt werden, wie etwa trainierte oder eingesetzte neuronale Netzwerke, die Ausgabemodelle **3116** des Trainingssystems **3104** beinhalten können.

[0316] In mindestens einer Ausführungsform können Aufgaben der Datenverarbeitungspipeline in einem oder mehreren Containern gekapselt sein, die jeweils eine diskrete, voll funktionsfähige Instanziierung einer Anwendung und einer virtualisierten Computerumgebung darstellen, die in der Lage ist, auf Modelle des maschinellen Lernens zu verweisen. In mindestens einer Ausführungsform können Container oder Anwendungen in einem privaten (z. B. mit eingeschränktem Zugriff) Bereich einer Containerregistrierungsdatenbank (in dieser Schrift ausführlicher beschrieben) veröffentlicht werden, und trainierte oder eingesetzte Modelle können in der Modellregistrierungsdatenbank **3124** gespeichert und einer oder mehreren Anwendungen zugeordnet werden. In mindestens einer Ausführungsform können Bilder von Anwendungen (z.B. Containerbilder) in einer Containerregistrierungsdatenbank verfügbar sein, und sobald sie von einem Benutzer aus einer Containerregistrierungsdatenbank zum Einsatz in einer Pipeline ausgewählt wurden, kann ein Bild verwendet werden, um einen Container für eine Instanziierung einer Anwendung zur Verwendung durch das System eines Benutzers zu erzeugen.

[0317] In mindestens einer Ausführungsform können Entwickler (z. B. Softwareentwickler, Kliniker, Ärzte usw.) Anwendungen (z. B. als Container) entwickeln, veröffentlichen und speichern, um eine Bildverarbeitung und/oder Ableitungen auf gelieferte Daten durchzuführen. In mindestens einer Ausführungsform können die Entwicklung, Veröffentlichung und/oder Speicherung unter Verwendung eines Softwareentwicklungskits (software development kit - SDK) durchgeführt werden, das einem System zugeordnet ist (z. B. um sicherzustellen, dass eine Anwendung und/oder ein entwickelter Container mit einem System konform oder kompatibel ist). In mindestens einer Ausführungsform kann eine entwickelte Anwendung lokal (z. B. in einer ersten Einrichtung bei Daten von einer ersten Einrichtung) mit einem SDK getestet werden, das mindestens einige der Dienste **3120** als ein System (z. B. das System **3200** aus **Fig. 32**) unterstützen kann. In mindestens einer Ausführungsform kann ein Entwickler, da DICOM-Objekte irgendwo von einem bis zu Hunderten von Bildern oder anderen Datenarten enthalten können, und aufgrund von Datenvariationen, für die Verwaltung (z. B. das Einstellen von Konstrukten zum Einbauen von Vorverarbeitung in eine Anwendung usw.), Extraktion und Aufbereitung eingehender DICOM-Daten verantwortlich sein. In mindestens einer Ausführungsform kann eine Anwendung nach der Validierung durch das System **3200** (z. B. für Genauigkeit, Sicherheit, Privatsphäre des Patienten usw.) in einer Containerregistrierungsdatenbank zur Auswahl und/oder Umsetzung durch einen Benutzer (z. B. ein (e) Krankenhaus, Klinik, Labor, Gesundheitsdienstleister usw.) verfügbar sein, um eine oder mehrere Verarbeitungsaufgaben in Bezug auf Daten in einer Einrichtung (z. B. einer zweiten Einrichtung) eines Benutzers durchzuführen.

[0318] In mindestens einer Ausführungsform können Entwickler dann Anwendungen oder Container über ein Netzwerk für den Zugriff und die Verwendung durch Benutzer eines Systems (z. B. System **3200** aus **Fig. 32**) gemeinsam nutzen. In mindestens einer Ausführungsform können abgeschlossene und validierte Anwendungen oder Container in einer Containerregistrierungsdatenbank gespeichert werden und können zugehörige Modelle des maschinellen Lernens in der Modellregistrierungsdatenbank **3124** gespeichert werden. In mindestens einer Ausführungsform kann eine anfragende Entität (z. B. ein Benutzer in einer medizinischen Einrichtung) - die eine Ableitungs- oder Bildverarbeitungsanfrage stellt - eine Containerregistrierungsdatenbank und/oder Modellregistrierungsdatenbank **3124** nach einer Anwendung, einem Container, einem Datensatz, einem Modell des maschinellen Lernens usw. durchsuchen, eine gewünschte Kombination von Elementen zur Aufnahme in die Datenverarbeitungspipeline auswählen und eine Bildverarbeitungsanfrage absenden. In mindestens einer Ausführungsform kann eine Anfrage Eingabedaten (und in einigen Beispielen zugehörige Patien-

tendaten) beinhalten, die zum Ausführen einer Anfrage erforderlich sind, und/oder kann eine Auswahl der Anwendung(en) und/oder Modelle des maschinellen Lernens beinhalten, die bei der Verarbeitung einer Anfrage ausgeführt werden sollen. In mindestens einer Ausführungsform kann dann eine Anfrage an eine oder mehrere Komponenten des Einsatzsystems **3106** (z. B. eine Cloud) weitergeleitet werden, um eine Verarbeitung der Datenverarbeitungspipeline durchzuführen. In mindestens einer Ausführungsform kann die Verarbeitung durch das Einsatzsystem **3106** das Referenzieren ausgewählter Elemente (z.B. Anwendungen, Container, Modelle usw.) aus einer Containerregistrierungsdatenbank und/oder Modellregistrierungsdatenbank **3124** beinhalten. In mindestens einer Ausführungsform können, sobald Ergebnisse durch eine Pipeline erzeugt werden, Ergebnisse an einen Benutzer als Referenz (z. B. zum Betrachten in einer Betrachtungsanwendungssuite, die auf einer lokalen Workstation vor Ort oder einem Terminal ausgeführt wird) zurückgegeben werden. In mindestens einer Ausführungsform kann ein Radiologe Ergebnisse von einer Datenverarbeitungspipeline empfangen, die eine beliebige Anzahl von Anwendungen und/oder Containern beinhaltet, wobei die Ergebnisse eine Anomalieerfassung in Röntgenstrahlen, CTs, MRTs usw. beinhalten können.

[0319] In mindestens einer Ausführungsform können Dienste **3120** genutzt werden, um die Verarbeitung oder Ausführung von Anwendungen oder Containern in Pipelines zu unterstützen. In mindestens einer Ausführungsform können die Dienste **3120** Berechnungsdienste, Dienste der künstlichen Intelligenz (KI), Visualisierungsdienste und/oder andere Dienstarten beinhalten. In mindestens einer Ausführungsform können die Dienste **3120** eine Funktion bereitstellen, die eine oder mehrere Anwendungen in der Software **3118** gemeinsam haben, so dass die Funktion auf einen Dienst abstrahiert werden kann, der von Anwendungen aufgerufen oder genutzt werden kann. In mindestens einer Ausführungsform kann die von den Diensten **3120** bereitgestellte Funktion dynamisch und effizienter ausgeführt werden, wobei sie auch gut skaliert werden kann, indem es Anwendungen ermöglicht wird, Daten parallel zu verarbeiten (z. B. unter Verwendung einer Parallelrechenplattform **3230** (**Fig. 32**)). In mindestens einer Ausführungsform kann, anstatt dass für jede Anwendung erforderlich ist, die dieselbe Funktion teilt, die von einem Dienst **3120** angeboten wird, eine jeweilige Instanz des Dienstes **3120** aufzuweisen, der Dienst **3120** zwischen und unter verschiedenen Anwendungen geteilt werden. In mindestens einer Ausführungsform können die Dienste als nichteinschränkende Beispiele einen Ableitungsserver oder eine Ableitungsengine beinhalten, die zum Ausführen von Erfassungs- oder Segmentierungsaufgaben verwendet werden können. In mindestens einer Ausführungsform kann ein Modelltrainingsdienst beinhaltet sein, der Trainings- und/oder Umschulungsfähigkeiten für Modelle des maschinellen Lernens bereitstellen kann. In mindestens einer Ausführungsform kann ferner ein Datenerweiterungsdienst beinhaltet sein, der GPU-beschleunigte Extraktion, Größenänderung, Skalierung und/oder andere Erweiterungen von Daten (z. B. DICOM, RIS, CIS, REST-kompatibel, RPC, Rohdaten usw.) bereitstellen kann. In mindestens einer Ausführungsform kann ein Visualisierungsdienst verwendet werden, der Bildrendering-Effekte - wie etwa Raytracing, Rasterisierung, Rauschunterdrückung, Schärfung usw. - hinzufügen kann, um zweidimensionalen (2D) und/oder dreidimensionalen (3D) Modellen Realismus zu verleihen. In mindestens einer Ausführungsform können Dienste für virtuelle Instrumenten beinhaltet sein, die Strahlformung, Segmentierung, Ableitung, Bildgebung und/oder Unterstützung für andere Anwendungen innerhalb von Pipelines von virtuellen Instrumenten bereitstellen.

[0320] In mindestens einer Ausführungsform, in der ein Dienst **3120** einen KI-Dienst (z. B. einen Ableitungsdienst) beinhaltet, können ein oder mehrere Modelle des maschinellen Lernens, die mit einer Anwendung zur Anomalieerfassung (z. B. Tumoren, Wachstumsanomalien, Narbenbildung usw.) verknüpft sind, durch das Aufrufen (z. B. als API-Aufruf) eines Ableitungsdienstes (z. B. eines Ableitungsservers) ausgeführt werden, um ein oder mehrere Modelle des maschinellen Lernens oder deren Verarbeitung als Teil der Anwendungsausführung auszuführen. In mindestens einer Ausführungsform, in der eine andere Anwendung ein oder mehrere Modelle des maschinellen Lernens für Segmentierungsaufgaben beinhaltet, kann eine Anwendung einen Ableitungsdienst aufrufen, um Modelle des maschinellen Lernens zum Durchführen einer oder mehrerer Verarbeitungsoperationen auszuführen, die mit Segmentierungsaufgaben verknüpft sind. In mindestens einer Ausführungsform kann die Software **3118**, die eine fortschrittliche Verarbeitungs- und Ableitungspipeline umsetzt, die eine Segmentierungsanwendung und eine Anomalieerfassungsanwendung beinhaltet, optimiert werden, da jede Anwendung denselben Ableitungsdienst aufrufen kann, um eine oder mehrere Ableitungsaufgaben auszuführen.

[0321] In mindestens einer Ausführungsform kann die Hardware **3122** GPUs, CPUs, Grafikkarten, ein KI-/Deep-Learning-System (z. B. einen KI-Supercomputer, wie etwa DGX von NVIDIA), eine Cloud-Plattform oder eine Kombination davon beinhalten. In mindestens einer Ausführungsform können unterschiedliche Arten von Hardware **3122** verwendet werden, um eine effiziente, zweckgerichtete Unterstützung für Software **3118** und Dienste **3120** im Einsatzsystem **3106** bereitzustellen. In mindestens einer Ausführungsform kann die Verwendung der GPU-Verarbeitung für die lokale Verarbeitung (z. B. in der Einrichtung **3102**), innerhalb eines KI-/Deep-Learning-Systems, in einem Cloud-System und/oder in anderen Verarbeitungskomponenten des Ein-

satzsystems **3106** umgesetzt werden, um die Effizienz, Genauigkeit und Effektivität von Bildverarbeitung, Bildrekonstruktion, Segmentierung, MRT-Untersuchungen, Schlaganfall- oder Herzinfarkterfassung (z. B. in Echtzeit), Bildqualität beim Rendering usw. zu verbessern. In mindestens einer Ausführungsform kann eine Einrichtung Bildgebungsvorrichtungen, Genomikvorrichtungen, Sequenziervorrichtungen und/oder andere Vorrichtungsarten vor Ort umfassen, die GPUs nutzen können, um Bildgebungsdaten zu erzeugen, die eine Anatomie eines Subjekts darstellen. In mindestens einer Ausführungsform können die Software **3118** und/oder die Dienste **3120** als nichteinschränkende Beispiele für die GPU-Verarbeitung in Bezug auf Deep Learning, maschinelles Lernen und/oder Hochleistungsberechnungen optimiert sein. In mindestens einer Ausführungsform kann mindestens ein Teil der Rechenumgebung des Einsatzsystems **3106** und/oder des Trainingssystems **3104** in einem Rechenzentrum, einem oder mehreren Supercomputern oder Hochleistungsrechensystemen mit GPU-optimierter Software (z. B. einer Hardware- und Softwarekombination des DGX-Systems von NVIDIA). In mindestens einer Ausführungsform können Rechenzentren derartig den Bestimmungen von HIPAA entsprechen, dass der Empfang, die Verarbeitung und die Übermittlung von Bildgebungsdaten und/oder anderen Patientendaten in Bezug auf die Vertraulichkeit von Patientendaten sicher gehandhabt werden. In mindestens einer Ausführungsform kann die Hardware **3122** eine beliebige Anzahl von GPUs beinhalten, die aufgerufen werden können, um die Verarbeitung von Daten parallel durchzuführen, wie in dieser Schrift beschrieben. In mindestens einer Ausführungsform kann die Cloud-Plattform ferner GPU-Verarbeitung für eine GPU-optimierte Ausführung von Deep-Learning-Aufgaben, Aufgaben des maschinellen Lernens oder andere Berechnungsaufgaben beinhalten. In mindestens einer Ausführungsform kann die Cloud-Plattform (z. B. NVIDIAs NGC) unter Verwendung eines oder mehrerer KI/Deep-Learning-Supercomputer und/oder GPU-optimierter Software (z. B. wie auf NVIDIAs DGX-Systemen bereitgestellt) als eine Hardwareabstraktion und -skalierungsplattform ausgeführt werden. In mindestens einer Ausführungsform kann die Cloud-Plattform ein Anwendungscontainer-Clusteringsystem oder ein Orchestrierungssystem (z. B. KUBERNETES) auf mehreren GPUs integrieren, um eine nahtlose Skalierung und Lastverteilung zu ermöglichen.

[0322] Fig. 32 ist ein Systemdiagramm für ein beispielhaftes System **3200** zum Erzeugen und Bereitstellen einer Bildgebungsbereitstellungspipeline gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform kann das System **3200** verwendet werden, um den Prozess **3100** aus **Fig. 31** und/oder andere Prozesse umzusetzen, was fortschrittliche Verarbeitungs- und Ableitungspipelines beinhaltet. In mindestens einer Ausführungsform kann das System **3200** ein Trainingssystem **3104** und ein Einsatzsystem **3106** beinhalten. In mindestens einer Ausführungsform können das Trainingssystem **3104** und das Einsatzsystem **3106** unter Verwendung von Software **3118**, Diensten **3120** und/oder Hardware **3122**, wie in dieser Schrift beschrieben, umgesetzt sein.

[0323] In mindestens einer Ausführungsform kann das System **3200** (z. B. das Trainingssystem **3104** und/oder das Einsatzsystem **3106**) in einer Cloud-Computing-Umgebung (z. B. unter Verwendung der Cloud **3226**) umgesetzt sein. In mindestens einer Ausführungsform kann das System **3200** lokal in Bezug auf eine Gesundheitsdienstleistungseinrichtung oder als Kombination von sowohl Cloud- als auch lokalen Berechnungsressourcen umgesetzt sein. In mindestens einer Ausführungsform, können in Ausführungsformen, in denen Cloud-Computing umgesetzt ist, Patientendaten von einer oder mehreren Komponenten des Systems **3200** getrennt oder von diesen unverarbeitet werden, was die Verarbeitung nichtkonform mit HIPAA und/oder anderen Datenhandhabungen und Datenschutzvorschriften oder -gesetzen machen würde. In mindestens einer Ausführungsform kann der Zugriff auf APIs in der Cloud **3226** durch beschlossene Sicherheitsmaßnahmen oder -protokolle auf autorisierte Benutzer beschränkt sein. In mindestens einer Ausführungsform kann ein Sicherheitsprotokoll Web-Token beinhalten, die von einem Authentifizierungsdienst (z. B. AuthN, AuthZ, Gluecon usw.) signiert sein können und eine geeignete Autorisierung tragen können. In mindestens einer Ausführungsform können APIs von virtuellen Instrumenten (in dieser Schrift beschrieben) oder andere Instanziierungen des Systems **3200** auf einen Satz öffentlicher IPs beschränkt sein, die überprüft oder für eine Interaktion autorisiert wurden.

[0324] In mindestens einer Ausführungsform können verschiedene Komponenten des Systems **3200** zwischen einander und untereinander unter Verwendung einer Reihe von unterschiedlichen Netzwerkarten kommunizieren, was lokale Netzwerke (local area networks - LANs) und/oder Weitverkehrsnetzwerke (wide area networks - WANs) über drahtgebundene Netzwerke und/oder drahtlose Kommunikationsprotokolle beinhaltet, ohne darauf beschränkt zu sein. In mindestens einer Ausführungsform kann die Kommunikation zwischen Einrichtungen und Komponenten des Systems **3200** (z. B. zum Übermitteln von Ableitungsanfragen, zum Empfangen von Ergebnissen von Ableitungsanfragen usw.) über (einen) Datenbus(se), drahtlose Datenprotokolle (Wi-Fi), drahtgebundene Datenprotokolle (z. B. Ethernet) usw. kommuniziert werden.

[0325] In mindestens einer Ausführungsform kann das Trainingssystem **3104** Trainingspipelines **3204** ausführen, die den in dieser Schrift mit Bezug auf **Fig. 31** beschriebenen ähneln. In mindestens einer Ausführungsform, in der ein oder mehrere Modelle des maschinellen Lernens in den Einsatzpipelines **3210** durch das Einsatzsystem **3106** verwendet werden sollen, können die Trainingspipelines **3204** verwendet werden, um ein oder mehrere (z. B. vortrainierte) Modelle zu trainieren oder neu zu trainieren und/oder ein oder mehrere vortrainierte Modelle **3206** umzusetzen (z. B. ohne eine Notwendigkeit für erneutes Training oder Aktualisierung). In mindestens einer Ausführungsform können als ein Ergebnis der Trainingspipelines **3204** Ausgabemodell(e) **3116** erzeugt werden. In mindestens einer Ausführungsform können die Trainingspipelines **3204** eine beliebige Anzahl von Verarbeitungsschritten beinhalten, wie etwa aber jedoch nicht beschränkt auf eine Konvertierung oder Anpassung von Bilddaten (oder anderen Eingabedaten) (z. B. Verwenden des DICOM-Adapters **3202A** zum Konvertieren von DICOM-Bildern in ein anderes Format, das für die Verarbeitung durch entsprechende Modelle des maschinellen Lernens geeignet ist, wie z. B. das Format der Neuroimaging Informatics Technology Initiative (NIfTI)), KI-unterstützte Anmerkung **3110**, Kennzeichnung oder Anmerkung von Bildgebungsdaten **3108**, um gekennzeichnete Klinikdaten **3112** zu erzeugen, Modellauswahl aus einer Modellregistrierungsdatenbank, Modelltraining **3114**, Training, Neutraining oder Aktualisieren von Modellen und/oder andere Verarbeitungsschritten. In mindestens einer Ausführungsform können für unterschiedliche Modelle des maschinellen Lernens, die vom Einsatzsystem **3106** verwendet werden, unterschiedliche Trainingspipelines **3204** verwendet werden. In mindestens einer Ausführungsform kann die Trainingspipeline **3204**, ähnlich zu einem ersten Beispiel, das mit Bezug auf **Fig. 31** beschrieben wurde, für ein erstes Modell des maschinellen Lernens verwendet werden, die Trainingspipeline **3204** kann, ähnlich zu einem zweiten Beispiel, das mit Bezug auf **Fig. 31** beschrieben wurde, für ein zweites Modell des maschinellen Lernens verwendet werden, und eine Trainingspipeline **3204** kann, ähnlich zu einem dritten Beispiel, das mit Bezug auf **Fig. 31** beschrieben wurde, für ein drittes Modell des maschinellen Lernens verwendet werden. In mindestens einer Ausführungsform kann eine beliebige Kombination von Aufgaben innerhalb des Trainingssystems **3104** verwendet werden, abhängig davon, was für jedes jeweilige Modell des maschinellen Lernens erforderlich ist. In mindestens einer Ausführungsform können ein oder mehrere Modelle des maschinellen Lernens bereits trainiert und einsatzbereit sein, sodass Modelle des maschinellen Lernens möglicherweise keiner Verarbeitung durch das Trainingssystem **3104** unterzogen werden und durch das Einsatzsystem **3106** umgesetzt werden können.

[0326] In mindestens einer Ausführungsform können (ein) Ausgabemodell(e) **3116** und/oder (ein) vortrainierte(s) Modell(e) **3206** abhängig von der Umsetzung oder Ausführungsform beliebige Arten von Modellen des maschinellen Lernens beinhalten. In mindestens einer Ausführungsform und ohne Einschränkung können vom System **3200** verwendete Modelle des maschinellen Lernens (ein) Modell(e) des maschinellen Lernens unter Verwendung von folgendem beinhalten: lineare Regression, logistische Regression, Entscheidungsbäume, Support-Vektor-Maschinen (SVM), Naive Bayes, k-nächster Nachbar (Knn), K bedeutet Clustering, Random Forest, Dimensionsverringerungsalgorithmen, Gradient-Boosting-Algorithmen, neuronale Netzwerke (z. B. Autokodierer, Convolutional, Recurrent, Perceptrons, Long/Short Term Memory (LSTM), Hopfield, Boltzmann, Deep Belief, Dekonvolutional, Generative Adversarial, Liquid State Machine usw.) und/oder andere Arten von Modellen des maschinellen Lernens.

[0327] In mindestens einer Ausführungsform können die Trainingspipelines **3204** eine KI-unterstützte Anmerkung beinhalten, wie in dieser Schrift in Bezug auf mindestens **Fig. 35B** ausführlicher beschrieben wird. In mindestens einer Ausführungsform können gekennzeichnete Klinikdaten **3112** (z. B. traditionelle Anmerkung) durch eine beliebige Anzahl von Methoden erzeugt werden. In mindestens einer Ausführungsform können Beschriftungen oder andere Anmerkungen innerhalb eines Zeichenprogramms (z. B. eines Anmerkungsprogramms), eines computergestützten Ausgestaltungs-(computer aided design - CAD-)Programms, eines Beschriftungsprogramms, einer anderen Programmart, die zum Erzeugen von Anmerkungen oder Beschriftungen für Ground-Truth geeignet ist, erzeugt werden und/oder können in einigen Beispielen handgezeichnet werden. In mindestens einer Ausführungsform können Ground-Truth-Daten synthetisch erzeugt (z. B. aus Computermodellen oder Renderings erzeugt), real erzeugt (z. B. aus realen Daten ausgestaltet und erzeugt), maschinenautomatisiert (z. B. unter Verwendung von Merkmalsanalyse und Lernen, um Merkmale aus Daten zu extrahieren und dann Kennzeichen zu erzeugen), von Menschen kommentiert werden (z. B. Kennzeichner oder Anmerkungsexperte, definiert die Position von Kennzeichen) und/oder eine Kombination davon. In mindestens einer Ausführungsform kann es für jede Instanz von Bildgebungsdaten **3108** (oder einer anderen Datenart, die von Modellen des maschinellen Lernens verwendet wird) entsprechende Ground-Truth-Daten geben, die durch das Trainingssystem **3104** erzeugt werden. In mindestens einer Ausführungsform kann die KI-unterstützte Anmerkung als Teil der Einsatzpipelines **3210** durchgeführt werden; entweder zusätzlich oder anstelle der KI-unterstützten Anmerkung, die in den Trainingspipelines **3204** beinhaltet ist. In mindestens einer Ausführungsform kann das System **3200** eine mehrschichtige Plattform beinhalten, die eine Softwareschicht (z. B. die Software **3118**) von Diagnoseanwendungen (oder anderen Anwendungsarten) beinhalten kann, die eine oder mehrere

medizinische Bildgebungs- und Diagnosefunktionen ausführen können. In mindestens einer Ausführungsform kann das System **3200** kommunikativ (z. B. über verschlüsselte Verbindungen) mit PACS-Servernetzwerken einer oder mehrerer Einrichtungen verbunden sein. In mindestens einer Ausführungsform kann das System **3200** dazu konfiguriert sein, auf Daten (z. B. DICOM-Daten, RIS-Daten, Rohdaten, CIS-Daten, REST-konforme Daten, RPC-Daten, Rohdaten usw.) von PACS-Servern (z. B. über einen DICOM-Adapter **3202** oder einen Adapter einer anderen Datenart, wie etwa RIS, CIS, REST-kompatibel, RPC, Raw usw.), um Operationen durchzuführen, wie etwa das Trainieren von Modellen des maschinellen Lernens, das Einsetzen von Modellen des maschinellen Lernens, Bildverarbeitung, Ableitung und/ oder andere Operationen.

[0328] In mindestens einer Ausführungsform kann eine Softwareschicht als eine sichere, verschlüsselte und/ oder authentifizierte API umgesetzt sein, durch die Anwendungen oder Container von (einer) externen Umgebung(en) (z. B. Einrichtung **3102**) gestartet (z. B. aufgerufen) werden können. In mindestens einer Ausführungsform können Anwendungen dann einen oder mehrere Dienste **3120** zum Durchführen von Rechen-, KI- oder Visualisierungsaufgaben im Zusammenhang mit entsprechenden Anwendungen aufrufen oder ausführen, und die Software **3118** und/oder Dienste **3120** können die Hardware **3122** nutzen, um Verarbeitungsaufgaben auf effektive und effiziente Weise auszuführen.

[0329] In mindestens einer Ausführungsform kann das Einsatzsystem **3106** Einsatzpipelines **3210** ausführen. In mindestens einer Ausführungsform können die Einsatzpipelines **3210** eine beliebige Anzahl von Anwendungen beinhalten, die aufeinanderfolgend, nichtaufeinanderfolgend oder anderweitig auf Bildgebungsdaten (und/oder andere Datenarten) angewendet werden können, die von Bildgebungsvorrichtungen, Sequenziervorrichtungen, Genomikvorrichtungen usw. erzeugt werden - was KI-unterstützte Anmerkung wie vorstehend beschrieben beinhaltet. In mindestens einer Ausführungsform kann, wie in dieser Schrift beschrieben, eine Einsatzpipeline **3210** für eine einzelne Vorrichtung als ein virtuelles Instrument für eine Vorrichtung (z. B. ein virtuelles Ultraschallinstrument, ein virtuelles CT-Instrument, ein virtuelles Sequenzierungsinstrument usw.) bezeichnet werden. In mindestens einer Ausführungsform kann es für eine einzelne Vorrichtung mehr als eine Einsatzpipeline **3210** geben, abhängig von Informationen, die aus Daten gewünscht sind, die von einer Vorrichtung erzeugt werden. In mindestens einer Ausführungsform kann, wenn Erfassungen von Anomalien von einem MRT-Gerät gewünscht sind, eine erste Einsatzpipeline **3210** vorhanden sein, und kann, wenn Bildverbesserung von der Ausgabe eines MRT-Geräts gewünscht ist, eine zweite Einsatzpipeline **3210** vorhanden sein.

[0330] In mindestens einer Ausführungsform können Anwendungen, die für Einsatzpipelines **3210** verfügbar sind, eine beliebige Anwendung beinhalten, die zum Ausführen von Verarbeitungsaufgaben bei Bildgebungsdaten oder anderen Daten von Vorrichtungen verwendet werden kann. In mindestens einer Ausführungsform können unterschiedliche Anwendungen für Bildverbesserung, Segmentierung, Rekonstruktion, Anomalieerfassung, Objekterfassung, Merkmalserfassung, Behandlungsplanung, Dosimetrie, Strahlplanung (oder andere Bestrahlungsbehandlungsprozesse) und/oder andere Analyse-, Bildverarbeitungs- oder Ableitungsaufgaben verantwortlich sein. In mindestens einer Ausführungsform kann das Einsatzsystem **3106** derartig Konstrukte für jede der Anwendungen definieren, dass Benutzer des Einsatzsystems **3106** (z. B. medizinische Einrichtungen, Labore, Kliniken usw.) Konstrukte verstehen und Anwendungen zur Umsetzung in ihrer jeweiligen Einrichtung anpassen können. In mindestens einer Ausführungsform kann eine Anwendung zur Bildrekonstruktion zur Aufnahme in die Einsatzpipeline **3210** ausgewählt werden, aber die von einer Bildgebungsvorrichtung erzeugte Datenart kann sich von einer in einer Anwendung verwendeten Datenart unterscheiden. In mindestens einer Ausführungsform kann der DICOM-Adapter **3202B** (und/oder ein DICOM-Lesevorrichtung) oder ein Adapter oder eine Lesevorrichtung einer anderen Datenart (z. B. RIS, CIS, REST-kompatibel, RPC, Rohdaten usw.) innerhalb der Einsatzpipeline **3210** zum Umwandeln von Daten in eine Form verwendet werden, die von einer Anwendung innerhalb des Einsatzsystems **3106** verwendet werden kann. In mindestens einer Ausführungsform kann der Zugriff auf DICOM-, RIS-, CIS-, REST-konforme, RPC-, Rohdaten- und/oder andere Datenartbibliotheken akkumuliert und vorverarbeitet werden, was das Dekodieren, Extrahieren und/oder Durchführen jeglicher Faltungen, Farbkorrekturen, Schärfe, Gamma und/oder andere Erweiterungen zu Daten beinhaltet. In mindestens einer Ausführungsform können DICOM-, RIS- , CIS-, REST-konforme, RPC- und/oder Rohdaten ungeordnet sein und es kann ein Vorlauf ausgeführt werden, um gesammelte Daten zu organisieren oder zu sortieren. Da in mindestens einer Ausführungsform verschiedene Anwendungen gemeinsame Bildoperationen gemeinsam nutzen können, kann in einigen Ausführungsformen eine Datenerweiterungsbibliothek (z. B. als einer der Dienste **3120**) verwendet werden, um diese Operationen zu beschleunigen. Um Engpässe herkömmlicher Verarbeitungsansätze zu vermeiden, die auf CPU-Verarbeitung beruhen, kann in mindestens einer Ausführungsform die parallele Rechenplattform **3230** für die GPU-Beschleunigung dieser Verarbeitungsaufgaben verwendet werden.

[0331] In mindestens einer Ausführungsform kann eine Bildrekonstruktionsanwendung eine Verarbeitungsaufgabe beinhalten, welche die Verwendung eines Modells des maschinellen Lernens beinhaltet. In mindestens einer Ausführungsform möchte ein Benutzer möglicherweise sein eigenes Modell des maschinellen Lernens verwenden oder ein Modell des maschinellen Lernens aus der Modellregistrierungsdatenbank **3124** auswählen. In mindestens einer Ausführungsform kann ein Benutzer sein eigenes Modell des maschinellen Lernens umsetzen oder ein Modell des maschinellen Lernens zur Aufnahme in eine Anwendung zum Ausführen einer Verarbeitungsaufgabe auswählen. In mindestens einer Ausführungsform können Anwendungen auswählbar und anpassbar sein, und durch das Definieren von Anwendungskonstrukten werden der Einsatz und die Umsetzung von Anwendungen für einen bestimmten Benutzer als eine nahtlosere Benutzererfahrung dargestellt. In mindestens einer Ausführungsform können die Einsatzpipelines **3210** durch Nutzung anderer Merkmale des Systems **3200** - wie etwa Dienste **3120** und Hardware **3122** - noch benutzerfreundlicher sein, eine einfachere Integration ermöglichen und genauere, effizientere und zeitgerechtere Ergebnisse bereitstellen.

[0332] In mindestens einer Ausführungsform kann das Einsatzsystem **3106** eine Benutzeroberfläche **3214** (z. B. eine grafische Benutzeroberfläche, eine Webschnittstelle usw.) beinhalten, die verwendet werden kann, um Anwendungen zur Aufnahme in die Einsatzpipeline(s) **3210** auszuwählen, Anwendungen anzuordnen, Anwendungen oder Parameter oder Konstrukte davon zu modifizieren oder zu ändern, Einsatzpipeline(s) **3210** während des Aufbaus und/oder Einsatzes zu verwenden und damit zu interagieren und/oder auf andere Weise mit dem Einsatzsystem **3106** zu interagieren. In mindestens einer Ausführungsform, obwohl in Bezug auf das Trainingssystem **3104** nicht veranschaulicht, kann die Benutzerschnittstelle **3214** (oder eine andere Benutzerschnittstelle) zum Auswählen von Modellen zur Verwendung im Einsatzsystem **3106**, zum Auswählen von Modellen für das Training oder das erneute Training im Trainingssystem **3104** und/oder zur anderweitigen Interaktion mit dem Trainingssystem **3104** verwendet werden.

[0333] In mindestens einer Ausführungsform kann der Pipeline-Verwalter **3212** zusätzlich zu einem Anwendungsorchestrierungssystem **3228** verwendet werden, um die Interaktion zwischen Anwendungen oder Containern der Einsatzpipeline(s) **3210** und Diensten **3120** und/oder Hardware **3122** zu verwalten. In mindestens einer Ausführungsform kann der Pipeline-Verwalter **3212** dazu konfiguriert sein, Interaktionen von Anwendung zu Anwendung, von Anwendung zu Dienst **3120** und/oder von Anwendung oder Dienst zur Hardware **3122** zu erleichtern. In mindestens einer Ausführungsform soll dies, obwohl es als in der Software **3118** beinhaltet veranschaulicht ist, nicht einschränkend sein, und in einigen Beispielen kann der Pipeline-Verwalter **3212** in den Diensten **3120** beinhaltet sein. In mindestens einer Ausführungsform kann das Anwendungsorchestrierungssystem **3228** (z. B. Kubernetes, DOCKER usw.) ein Containerorchestrierungssystem beinhalten, das Anwendungen in Containern als logische Einheiten zur Koordination, Verwaltung, Skalierung und zum Einsatz gruppieren kann. In mindestens einer Ausführungsform kann jede Anwendung in einer eigenständigen Umgebung (z. B. auf Ebene eines Kernel) durch das Verknüpfen von Anwendungen aus der/den Einsatzpipeline(s) **3210** (z. B. einer Rekonstruktionsanwendung, einer Segmentierungsanwendung usw.) mit einzelnen Containern ausgeführt werden, um Geschwindigkeit und Effizienz zu erhöhen.

[0334] In mindestens einer Ausführungsform kann jede Anwendung und/oder jeder Container (oder jedes Bild davon) individuell entwickelt, modifiziert und eingesetzt werden (z. B. kann ein(e) erste(r) Benutzer oder Entwickler eine erste Anwendung entwickeln, modifizieren und einsetzen und ein zweiter Benutzer oder Entwickler kann eine zweite Anwendung unabhängig von einem ersten Benutzer oder Entwickler entwickeln, modifizieren und bereitstellen), was die Konzentration und Aufmerksamkeit auf eine Aufgabe einer einzelnen Anwendung und/oder Container ermöglichen kann, ohne durch Aufgaben anderer Anwendung(en) oder Container behindert zu werden. In mindestens einer Ausführungsform kann die Kommunikation und Zusammenarbeit zwischen verschiedenen Containern oder Anwendungen durch den Pipeline-Verwalter **3212** und das Anwendungsorchestrierungssystem **3228** unterstützt werden. In mindestens einer Ausführungsform können, solange eine erwartete Eingabe und/oder Ausgabe jedes Containers oder jeder Anwendung einem System bekannt ist (z. B. auf Grundlage von Konstrukten von Anwendungen oder Containern), das Anwendungsorchestrierungssystem **3228** und/oder der Pipeline-Verwalter **3212** die Kommunikation unter und zwischen und die gemeinsame Nutzung von Ressourcen unter und zwischen jeder der Anwendungen oder Container erleichtern. In mindestens einer Ausführungsform kann, da eine oder mehrere Anwendungen oder Container in der/den Einsatzpipeline(s) **3210** dieselben Dienste und Ressourcen gemeinsam nutzen können, das Anwendungsorchestrierungssystem **3228** orchestrieren, einen Lastausgleich durchführen und die gemeinsame Nutzung von Diensten oder Ressourcen zwischen und unter verschiedenen Anwendungen oder Containern bestimmen. In mindestens einer Ausführungsform kann ein Planer verwendet werden, um Ressourcenanforderungen von Anwendungen oder Containern, die aktuelle Nutzung oder geplante Nutzung dieser Ressourcen und die Ressourcenverfügbarkeit zu verfolgen. In mindestens einer Ausführungsform kann ein Planer somit Ressourcen unterschiedlichen Anwendungen zuweisen und Ressourcen im Hinblick auf Anforderungen und Verfügbarkeit

eines Systems zwischen und unter Anwendungen verteilen. In einigen Beispielen kann ein Planer (und/oder eine andere Komponente des Anwendungsorchestrierungssystems **3228**) die Ressourcenverfügbarkeit und -verteilung auf Grundlage von Einschränkungen bestimmen, die einem System auferlegt werden (z.B. Benutzereinschränkungen), wie etwa Dienstqualität (quality of service - QoS), Dringlichkeit der Notwendigkeit für Datenausgaben (z. B. um zu bestimmen, ob eine Echtzeitverarbeitung oder eine verzögerte Verarbeitung ausgeführt werden soll) usw.

[0335] In mindestens einer Ausführungsform können Dienste **3120**, die von Anwendungen oder Containern im Einsatzsystem **3106** genutzt und von diesen gemeinsam genutzt werden, Rechendienste **3216**, KI-Dienste **3218**, Visualisierungsdienste **3220** und/oder andere Dienstarten beinhalten. In mindestens einer Ausführungsform können Anwendungen einen oder mehrere der Dienste **3120** aufrufen (z. B. ausführen), um Verarbeitungsvorgänge für eine Anwendung durchzuführen. In mindestens einer Ausführungsform können die Rechendienste **3216** von Anwendungen genutzt werden, um Super-Computing- oder andere High-Performance-Computing-(HPC-)Aufgaben durchzuführen. In mindestens einer Ausführungsform können der/die Rechendienst(e) **3216** genutzt werden, um eine parallele Verarbeitung (z. B. unter Verwendung einer parallelen Rechenplattform **3230**) zum Verarbeiten von Daten durch eine oder mehrere Anwendungen und/oder eine oder mehrere Aufgaben einer einzelnen Anwendung im Wesentlichen gleichzeitig durchzuführen. In mindestens einer Ausführungsform kann die parallele Rechenplattform **3230** (z. B. CUDA von NVIDIA) Allzweckberechnungen auf GPUs (GPGPU) (z. B. GPUs **3222**) ermöglichen. In mindestens einer Ausführungsform kann eine Softwareschicht der parallelen Rechenplattform **3230** Zugriff auf virtuelle Anweisungssätze und parallele Rechenelemente von GPUs zur Ausführung von Rechenkemen bereitstellen. In mindestens einer Ausführungsform kann die parallele Rechenplattform **3230** Speicher beinhalten, und in einigen Ausführungsformen kann ein Speicher zwischen mehreren Containern und/oder zwischen und unter verschiedenen Verarbeitungsaufgaben innerhalb eines einzelnen Containers genutzt werden. In mindestens einer Ausführungsform können Zwischenprozesskommunikations-(IPC-)Aufrufe für mehrere Container und/oder für mehrere Prozesse innerhalb eines Containers erzeugt werden, um dieselben Daten aus einem gemeinsam genutzten Speichersegment der parallelen Rechenplattform **3230** zu verwenden (bei denen z. B. mehrere unterschiedliche Phasen einer Anwendung oder mehrere Anwendungen dieselben Informationen verarbeiten). In mindestens einer Ausführungsform können, anstatt eine Kopie von Daten zu erstellen und Daten an unterschiedliche Speicherorte zu verschieben (z. B. eine Lese-/Schreiboperation), dieselben Daten an demselben Speicherort für eine beliebige Anzahl von Verarbeitungsaufgaben verwendet werden (z. B, gleichzeitig, zu unterschiedlichen Zeiten usw.). In mindestens einer Ausführungsform können, wenn Daten verwendet werden, um neue Daten als ein Ergebnis der Verarbeitung zu erzeugen, diese Informationen einer neuen Position von Daten gespeichert und zwischen verschiedenen Anwendungen geteilt werden. In mindestens einer Ausführungsform können die Position von Daten und eine Position von aktualisierten oder modifizierten Daten Teil einer Definition dafür sein, wie eine Nutzlast innerhalb von Containern verstanden wird.

[0336] In mindestens einer Ausführungsform können die KI-Dienste **3218** genutzt werden, um Ableitungsdienste zum Ausführen von (einem) Modell(en) des maschinellen Lernens auszuführen, die mit Anwendungen verknüpft sind (z. B. mit der Ausführung einer oder mehrerer Verarbeitungsaufgaben einer Anwendung beauftragt). In mindestens einer Ausführungsform können die KI-Dienste **3218** das KI-System **3224** nutzen, um ein oder mehrere Modelle des maschinellen Lernens (z. B. neuronale Netzwerke, wie etwa CNNs) zur Segmentierung, Rekonstruktion, Objekterfassung, Merkmalserfassung, Klassifizierung und/oder anderen Ableitungsaufgaben auszuführen. In mindestens einer Ausführungsform können Anwendungen der Einsatzpipeline(s) **3210** eines oder mehrere der Ausgabemodelle **3116** vom Trainingssystem **3104** und/oder andere Anwendungsmodelle verwenden, um Ableitungen auf Bildgebungsdaten (z. B. DICOM-Daten, RIS-Daten, CIS-Daten, REST-konforme Daten, RPC-Daten, Rohdaten usw.) durchzuführen. In mindestens einer Ausführungsform können zwei oder mehr Beispiele für die Ableitung unter Verwendung des Anwendungsorchestrierungssystems **3228** (z. B. eines Planers) verfügbar sein. In mindestens einer Ausführungsform kann eine erste Kategorie einen Pfad mit hoher Priorität/niedriger Latenz beinhalten, der höhere Dienstgütevereinbarungen erreichen kann, wie etwa zum Durchführen von Ableitungen auf dringende Anforderungen während eines Notfalls oder für einen Radiologen während einer Diagnose. In mindestens einer Ausführungsform kann eine zweite Kategorie einen Standardprioritätspfad beinhalten, der für Anforderungen verwendet werden kann, die möglicherweise nicht dringend sind oder bei denen eine Analyse zu einem späteren Zeitpunkt durchgeführt werden kann. In mindestens einer Ausführungsform kann das Anwendungsorchestrierungssystem **3228** Ressourcen (z. B. Dienste **3120** und/oder Hardware **3122**) auf Grundlage von Prioritätspfaden für unterschiedliche Ableitungsaufgaben der KI-Dienste **3218** verteilen.

[0337] In mindestens einer Ausführungsform kann gemeinsam genutzter Speicher an KI-Diensten **3218** innerhalb des Systems **3200** angebracht sein. In mindestens einer Ausführungsform kann gemeinsam genutz-

ter Speicher als Zwischenspeicher (oder eine andere Speichervorrichtungsart) arbeiten und kann verwendet werden, um Ableitungsanforderungen von Anwendungen zu verarbeiten. In mindestens einer Ausführungsform kann, wenn eine Ableitungsanfrage übermittelt wird, eine Anfrage von einem Satz von API-Instanzen des Einsatzsystems **3106** empfangen werden, und eine oder mehrere Instanzen können ausgewählt werden (z. B. zur besten Anpassung, zum Lastausgleich usw.), um eine Anfrage zu bearbeiten. In mindestens einer Ausführungsform kann, um eine Anfrage zu verarbeiten, eine Anfrage in eine Datenbank eingetragen werden, kann ein Modell des maschinellen Lernens von der Modellregistrierungsdatenbank **3124** entfernt sein, wenn es sich nicht bereits in einem Zwischenspeicher befindet, kann ein Validierungsschritt sicherstellen, dass ein geeignetes Modell des maschinellen Lernens in einen Zwischenspeicher (z. B. gemeinsam genutzter Speicher) geladen wird und/oder eine Kopie eines Modells können in einen Zwischenspeicher gespeichert werden. In mindestens einer Ausführungsform kann ein Planer (z. B. des Pipeline-Verwalters **3212**) verwendet werden, um eine Anwendung zu starten, auf die in einer Anforderung verwiesen wird, wenn eine Anwendung nicht bereits ausgeführt wird oder wenn nicht genügend Instanzen einer Anwendung vorhanden sind. In mindestens einer Ausführungsform kann ein Ableitungsserver gestartet werden, wenn noch kein Ableitungsserver gestartet wurde, um ein Modell auszuführen. Pro Modell kann eine beliebige Anzahl von Ableitungsservern gestartet werden. In mindestens einer Ausführungsform können in einem Pull-Modell, in dem Ableitungsserver geclustert werden, Modelle zwischengespeichert werden, immer wenn ein Lastausgleich vorteilhaft ist. In mindestens einer Ausführungsform können Ableitungsserver statisch in entsprechende verteilte Server geladen werden.

[0338] In mindestens einer Ausführungsform kann die Ableitung unter Verwendung eines Ableitungsservers durchgeführt werden, der in einem Container läuft. In mindestens einer Ausführungsform kann eine Instanz eines Ableitungsservers mit einem Modell (und optional einer Vielzahl von Versionen eines Modells) verknüpft sein. In mindestens einer Ausführungsform kann, wenn eine Instanz eines Ableitungsservers nicht existiert, wenn eine Anforderung zum Ausführen einer Ableitung an einem Modell empfangen wird, eine neue Instanz geladen werden. In mindestens einer Ausführungsform kann, wenn ein Ableitungsservers gestartet wird, ein Modell derartig an einen Ableitungsserver übergeben werden, dass ein selber Container verwendet werden kann, um unterschiedliche Modelle zu bedienen, solange der Ableitungsserver als eine andere Instanz läuft.

[0339] In mindestens einer Ausführungsform kann während der Anwendungsausführung eine Ableitungsanfrage für eine gegebene Anwendung empfangen werden und ein Container (z. B. der eine Instanz eines Ableitungsservers hostet) kann geladen werden (wenn nicht bereits geschehen) und eine Startprozedur kann aufgerufen werden. In mindestens einer Ausführungsform kann die Vorverarbeitungslogik in einem Container eine beliebige zusätzliche Vorverarbeitung an eingehenden Daten laden, dekodieren und/oder durchführen (z. B. unter Verwendung von (einer) CPU(s) und/oder GPU(s)). In mindestens einer Ausführungsform kann ein Container, sobald Daten für die Ableitung vorbereitet sind, die Ableitung nach Bedarf an Daten durchführen. In mindestens einer Ausführungsform kann dies einen einzelnen Ableitungsabruf auf einem Bild (z. B. eine Handröntgenaufnahme) beinhalten oder kann eine Ableitung auf Hunderten von Bildern erfordern (z. B. ein Thorax-CT). In mindestens einer Ausführungsform kann eine Anwendung Ergebnisse vor dem Vervollständigen zusammenfassen, was ohne Einschränkung einen einzelnen Konfidenzwert, eine Segmentierung auf Bildpunktebene, eine Segmentierung auf Voxelebene, das Erzeugen einer Visualisierung oder das Erzeugen von Text zum Zusammenfassen von Ergebnissen beinhalten kann. In mindestens einer Ausführungsform können unterschiedlichen Modellen oder Anwendungen unterschiedliche Prioritäten zugewiesen werden. Zum Beispiel können einige Modelle eine Echtzeitpriorität (TAT < 1 min) aufweisen, während andere eine niedrigere Priorität (z. B. TAT < 10 min) aufweisen können. In mindestens einer Ausführungsform können Modellausführungszeiten von der anfordernden Institution oder Entität gemessen werden und können die Partnernetzwerkdurchquerungszeit sowie die Ausführung auf einem Ableitungsdienst beinhalten.

[0340] In mindestens einer Ausführungsform kann die Übertragung von Anforderungen zwischen Diensten **3120** und Ableitungsanwendungen hinter einem Softwareentwicklungskit (SDK) verborgen sein und ein robuster Transport kann über eine Warteschlange bereitgestellt sein. In mindestens einer Ausführungsform wird eine Anfrage über eine API in eine Warteschlange für eine individuelle Anwendungs-/Mandanten-ID-Kombination gestellt, und ein SDK zieht eine Anfrage aus einer Warteschlange und gibt eine Anfrage an eine Anwendung. In mindestens einer Ausführungsform kann ein Name einer Warteschlange in einer Umgebung bereitgestellt werden, von der ein SDK ihn abholt. In mindestens einer Ausführungsform kann eine asynchrone Kommunikation über eine Warteschlange nützlich sein, da sie es jeder Instanz einer Anwendung ermöglichen kann, Arbeit aufzunehmen, sobald sie verfügbar wird. Ergebnisse können über eine Warteschlange zurückübertragen werden, um sicherzustellen, dass keine Daten verloren gehen. In mindestens einer Ausführungsform können Warteschlangen auch die Möglichkeit bieten, Arbeit zu segmentieren, da die Arbeit mit der höchsten Priorität an eine Warteschlange mit den meisten Instanzen einer damit verbundenen Anwendung gehen kann, während die Arbeit mit der niedrigsten Priorität an eine Warteschlange mit einer damit verbundenen einzelnen Instanz

gehen kann, die Aufgaben in einer empfangenen Reihenfolge verarbeitet. In mindestens einer Ausführungsform kann eine Anwendung auf einer GPU-beschleunigten Instanz ausgeführt werden, die in der Cloud **3226** erzeugt wird, und ein Ableitungsdienst kann eine Ableitung auf einer GPU durchführen.

[0341] In mindestens einer Ausführungsform können die Visualisierungsdienste **3220** genutzt werden, um Visualisierungen zum Anzeigen von Ausgaben von Anwendungen und/oder Einsatzpipeline(s) **3210** zu erzeugen. In mindestens einer Ausführungsform können die GPUs **3222** von den Visualisierungsdiensten **3220** genutzt werden, um Visualisierungen zu erzeugen. In mindestens einer Ausführungsform können Rendering-Effekte, wie etwa Raytracing, durch die Visualisierungsdienste **3220** umgesetzt werden, um Visualisierungen höherer Qualität zu erzeugen. In mindestens einer Ausführungsform können Visualisierungen ohne Einschränkung 2D-Bildrendering, 3D-Volumenrendering, 3D-Volumenrekonstruktion, 2D-tomografische Schnitte, Virtual-Reality-Anzeigen, Augmented-Reality-Anzeigen usw. beinhalten. In mindestens einer Ausführungsform können virtualisierte Umgebungen verwendet werden, um eine virtuell interaktive Anzeige oder Umgebung (z. B. eine virtuelle Umgebung) zur Interaktion durch Benutzer eines Systems (z. B. Ärzte, Krankenschwestern, Radiologen usw.) zu erzeugen. In mindestens einer Ausführungsform können die Visualisierungsdienste **3220** einen internen Visualisierer, Kinematik und/oder andere Rendering- oder Bildverarbeitungsfähigkeiten oder -funktionen (z. B. Raytracing, Rasterisierung, interne Optik usw.) beinhalten.

[0342] In mindestens einer Ausführungsform kann die Hardware **3122** GPUs **3222**, das KI-System **3224**, die Cloud **3226** und/oder jede andere Hardware beinhalten, die zum Ausführen des Trainingssystems **3104** und/oder des Einsatzsystems **3106** verwendet wird. In mindestens einer Ausführungsform können GPUs **3222** (z. B. TESLA- und/oder QUADRO-GPUs von NVIDIA) eine beliebige Anzahl von GPUs beinhalten, die zum Ausführen von Verarbeitungsaufgaben von Rechendiensten **3216**, KI-Diensten **3218**, Visualisierungsdiensten **3220**, anderen Diensten und /oder eines der Merkmale oder Funktionen der Software **3118** verwendet werden können. Zum Beispiel können in Bezug auf die KI-Dienste **3218** die GPUs **3222** verwendet werden, um eine Vorverarbeitung von Bilddaten (oder anderen Datenarten, die von Modellen des maschinellen Lernens verwendet werden), eine Nachverarbeitung von Ausgaben von Modellen des maschinellen Lernens durchzuführen und/oder Ableitung (z. B. um Modelle des maschinellen Lernens auszuführen) durchzuführen. In mindestens einer Ausführungsform können die Cloud **3226**, das KI-System **3224** und/oder andere Komponenten des Systems **3200** die GPUs **3222** verwenden. In mindestens einer Ausführungsform kann die Cloud **3226** eine GPU-optimierte Plattform für Deep-Learning-Aufgaben beinhalten. In mindestens einer Ausführungsform kann das KI-System **3224** GPUs verwenden, und die Cloud **3226** - oder mindestens ein mit Deep Learning oder Ableitung beauftragter Teil - kann unter Verwendung eines oder mehrerer KI-Systeme **3224** ausgeführt werden. Obwohl die Hardware **3122** als diskrete Komponenten veranschaulicht ist, soll dies nicht einschränkend sein, und beliebige Komponenten der Hardware **3122** können mit beliebigen anderen Komponenten der Hardware **3122** kombiniert oder von diesen genutzt werden.

[0343] In mindestens einer Ausführungsform kann das KI-System **3224** ein zweckorientiertes Computersystem (z. B. einen Supercomputer oder einen HPC) beinhalten, das für Ableitung, Deep Learning, maschinelles Lernen und/oder andere Aufgaben der künstlichen Intelligenz konfiguriert ist. In mindestens einer Ausführungsform kann das KI-System **3224** (z. B. DGX von NVIDIA) GPU-optimierte Software (z. B. einen Software-Stapel) beinhalten, die unter Verwendung einer Vielzahl von GPUs **3222** zusätzlich zu CPUs, RAM, Speicher und/oder andere Komponenten, Merkmale oder Funktionen ausgeführt werden kann. In mindestens einer Ausführungsform können ein oder mehrere KI-Systeme **3224** in der Cloud **3226** (z. B. in einem Rechenzentrum) zum Durchführen einiger oder aller KI-basierten Verarbeitungsaufgaben des Systems **3200** umgesetzt sein.

[0344] In mindestens einer Ausführungsform kann die Cloud **3226** eine GPU-beschleunigte Infrastruktur (z. B. NGC von NVIDIA) beinhalten, die eine GPU-optimierte Plattform zum Ausführen von Verarbeitungsaufgaben des Systems **3200** bereitstellen kann. In mindestens einer Ausführungsform kann die Cloud **3226** ein oder mehrere KI-Systeme **3224** zum Durchführen einer oder mehrerer KI-basierter Aufgaben des Systems **3200** (z. B. als Hardware-Abstraktions- und Skalierungsplattform) beinhalten. In mindestens einer Ausführungsform kann die Cloud **3226** in das Anwendungsorchestrierungssystem **3228** integriert sein, das mehrere GPUs nutzt, um eine(n) nahtlose(n) Skalierung und Lastausgleich zwischen und unter den Anwendungen und Diensten **3120** zu ermöglichen. In mindestens einer Ausführungsform kann die Cloud **3226** die Aufgabe haben, mindestens einige der Dienste **3120** des Systems **3200** auszuführen, was Rechendienste **3216**, KI-Dienste **3218** und/oder Visualisierungsdienste **3220** beinhaltet, wie in dieser Schrift beschrieben. In mindestens einer Ausführungsform kann die Cloud **3226** Ableitungen kleiner und großer Batches durchführen (z. B. das Ausführen von TENSOR RT von NVIDIA), eine beschleunigte Parallelberechnungs-API und -Plattform **3230** (z. B. CUDA von NVIDIA) bereitstellen, das Anwendungsorchestrierungssystem **3228** (z. B. KUBERNETES) ausführen, eine Grafik-Rendering-API und -Plattform bereitstellen (z. B. für Raytracing, 2D-Grafiken, 3D-Grafiken und/

oder andere Rendering-Methoden, um Filmsequenzen höherer Qualität zu erzeugen) und/oder kann andere Funktionen für das System **3200** bereitstellen.

[0345] In mindestens einer Ausführungsform kann die Cloud **3226**, in dem Bemühen, die Vertraulichkeit des Patienten zu wahren (z. B. wenn Patientendaten oder -aufzeichnungen außerhalb von Geschäftsräumen verwendet werden sollen), eine Registrierungsdatenbank beinhalten - wie etwa eine Deep-Learning-Container-Registrierungsdatenbank. In mindestens einer Ausführungsform kann eine Registrierungsdatenbank Container für Instanziierungen von Anwendungen speichern, die Vorverarbeitungs-, Nachbearbeitungs- oder andere Verarbeitungsaufgaben an Patientendaten ausführen können. In mindestens einer Ausführungsform kann die Cloud **3226** Daten empfangen, die sowohl Patientendaten als auch Sensordaten in Containern beinhalten, eine angeforderte Verarbeitung nur für Sensordaten in diesen Containern durchführen und dann eine sich ergebende Ausgabe und/oder Visualisierungen an geeignete Parteien und/oder Vorrichtungen (z. B. medizinische Vorrichtungen vor Ort, die zur Visualisierung oder Diagnose verwendet werden) weiterleiten, alles ohne dass Patientendaten extrahiert, gespeichert oder anderweitig darauf zugegriffen werden muss. In mindestens einer Ausführungsform wird die Vertraulichkeit von Patientendaten in Übereinstimmung mit HIPAA und/oder anderen Datenvorschriften gewahrt.

[0346] **Fig. 33A** veranschaulicht ein Datenflussdiagramm für einen Prozess **3300** zum Trainieren, erneuten Trainieren oder Aktualisieren eines Modells für maschinelles Lernen gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform kann der Prozess **3300** unter Verwendung des Systems **3200** aus **Fig. 32** als ein nichteinschränkendes Beispiel ausgeführt werden. In mindestens einer Ausführungsform kann der Prozess **3300** Dienste **3120** und/oder Hardware **3122** des Systems **3200** nutzen, wie in dieser Schrift beschrieben. In mindestens einer Ausführungsform können durch den Prozess **3300** erzeugte präzisierte Modelle **3312** durch das Einsatzsystem **3106** für eine oder mehrere containerisierte Anwendungen in Einsatzpipelines **3210** ausgeführt werden.

[0347] In mindestens einer Ausführungsform kann das Modelltraining **3114** ein erneutes Trainieren oder Aktualisieren eines anfänglichen Modells **3304** (z. B. eines vortrainierten Modells) unter Verwendung neuer Trainingsdaten (z. B. neuer Eingabedaten, wie etwa der Kundendatensatz **3306**, und/oder neuer mit Eingabedaten verknüpften Ground Truth-Daten) beinhalten. In mindestens einer Ausführungsform können, um das Anfangsmodell **3304** neu zu trainieren oder zu aktualisieren, Ausgabe- oder Verlustschicht(en) des Anfangsmodells **3304** zurückgesetzt oder gelöscht und/oder durch (eine) aktualisierte oder neue Ausgabe- oder Verlustschicht (en) ersetzt werden. In mindestens einer Ausführungsform kann das Anfangsmodell **3304** zuvor feinabgestimmte Parameter (z. B. Gewichtungen und/oder Verzerrungen) aufweisen, die vom vorherigen Training übrig bleiben, sodass das Training oder erneute Training **3114** möglicherweise nicht so lange dauert oder so viel Verarbeitung erfordert wie das Trainieren eines Modells von Grund auf. In mindestens einer Ausführungsform können Parameter, während des Modelltrainings **3114** durch das Zurücksetzen oder Ersetzen der Ausgabe- oder Verlustschicht(en) des Anfangsmodells **3304** aktualisiert und für einen neuen Datensatz auf Grundlage von Verlustberechnungen neu abgestimmt werden, die mit der Genauigkeit der Ausgabe oder Verlustschicht (en) beim Erzeugen von Vorhersagen für einen neuen Kundendatensatz **3306** (z. B. die Bilddaten **3108** aus **Fig. 31**) verknüpft sind.

[0348] In mindestens einer Ausführungsform können vortrainierte Modelle **3206** in einem Datenspeicher oder einer Registrierungsdatenbank (z. B. die Modellregistrierungsdatenbank **3124** aus **Fig. 31**) gespeichert sein. In mindestens einer Ausführungsform können vortrainierte Modelle **3206** mindestens teilweise an einer oder mehreren Einrichtungen außer einem Einrichtungsausführungsprozess **3300** trainiert worden sein. In mindestens einer Ausführungsform können zum Schutz der Privatsphäre und der Rechte von Patienten, Subjekten oder Klienten unterschiedlicher Einrichtungen vortrainierte Modelle **3206** vor Ort unter Verwendung von vor Ort erzeugten Kunden- oder Patientendaten trainiert worden sein. In mindestens einer Ausführungsform können die vortrainierten Modelle **3206** unter Verwendung der Cloud **3226** und/oder anderer Hardware **3122** trainiert werden, aber vertrauliche Patientendaten mit gewährleistetem Datenschutz können nicht an beliebige Komponenten der Cloud **3226**(oder anderer Hardware außerhalb der Geschäftsräume) übertragen, von diesen verwendet oder diesen zugänglich gemacht werden. In mindestens einer Ausführungsform, in der ein vortrainiertes Modell **3206** darin trainiert wird, Patientendaten von mehr als einer Einrichtung zu verwenden, kann das vortrainierte Modell **3206** für jede Einrichtung individuell trainiert worden sein, bevor es mit Patienten- oder Kundendaten von einer anderen Einrichtung trainiert wird. In mindestens einer Ausführungsform, wie etwa wenn Kunden- oder Patientendaten wegen Datenschutzbedenken (z. B. durch Verzicht, für experimentelle Zwecke usw.) freigegeben wurden, oder wenn Kunden- oder Patientendaten in einem öffentlichen Datensatz beinhaltet sind, können Kunden- oder Patientendaten von einer beliebigen Anzahl von Einrichtungen verwen-

det werden, um das vortrainierte Modell **3206** vor Ort und/oder außerhalb, wie etwa in einem Rechenzentrum oder einer anderen Cloud-Computing-Infrastruktur zu trainieren.

[0349] In mindestens einer Ausführungsform kann ein Benutzer beim Auswählen von Anwendungen zur Verwendung in Einsatzpipelines **3210** auch Modelle des maschinellen Lernens auswählen, die für konkrete Anwendungen verwendet werden sollen. In mindestens einer Ausführungsform hat ein Benutzer möglicherweise kein Modell zur Verwendung, so dass ein Benutzer ein vortrainiertes Modell **3206** zur Verwendung mit einer Anwendung auswählen kann. In mindestens einer Ausführungsform ist das vortrainierte Modell **3206** möglicherweise nicht für das Erzeugen genauer Ergebnisse des Kundendatensatzes **3306** einer Einrichtung eines Benutzers optimiert (z. B. auf Grundlage von Patientendiversität, Demografie, Arten von verwendeten medizinischen Bildgebungsvorrichtungen usw.). In mindestens einer Ausführungsform kann das vortrainierte Modell **3206** vor dem Einsetzen des vortrainierten Modells **3206** in der Einsatzpipeline **3210** zur Verwendung mit einer oder mehreren Anwendungen für die Verwendung in einer jeweiligen Einrichtung aktualisiert, neu trainiert und/oder feinabgestimmt werden.

[0350] In mindestens einer Ausführungsform kann ein Benutzer ein vortrainiertes Modell **3206** auswählen, das aktualisiert, neu trainiert und/oder feinabgestimmt werden soll, und das vortrainierte Modell **3206** kann als Anfangsmodell **3304** für das Trainingssystem **3104** innerhalb des Prozesses **3300** bezeichnet werden. In mindestens einer Ausführungsform kann der Kundendatensatz **3306** (z. B. Bildgebungsdaten, Genomdaten, Sequenzierungsdaten oder andere Datenarten, die von Vorrichtungen in einer Einrichtung erzeugt werden) verwendet werden, um das Modelltraining **3114** (das ohne Einschränkung das Übertragungslernen beinhalten kann) auf dem anfänglichen Modell **3304** durchzuführen, um das präzisierte Modell **3312** zu erzeugen. In mindestens einer Ausführungsform können Ground-Truth-Daten, die dem Kundendatensatz **3306** entsprechen, durch das Trainingssystem **3104** erzeugt werden. In mindestens einer Ausführungsform können Ground-Truth-Daten mindestens teilweise von Klinikern, Wissenschaftlern, Ärzten, Praktizierenden, in einer Einrichtung (z. B. als gekennzeichnete Klinikdaten **3112** aus **Fig. 31**) erzeugt werden.

[0351] In mindestens einer Ausführungsform kann die KI-unterstützte Anmerkung **3110** in einigen Beispielen verwendet werden, um Ground-Truth-Daten zu erzeugen. In mindestens einer Ausführungsform kann die KI-unterstützte Anmerkung **3110** (z. B. umgesetzt unter Verwendung eines KI-unterstützten Anmerkungs-SDK) Modelle des maschinellen Lernens (z. B. neuronale Netzwerke) nutzen, um vorgeschlagene oder vorhergesagte Ground-Truth-Daten für einen Kundendatensatz zu erzeugen. In mindestens einer Ausführungsform kann der Benutzer **3310** Anmerkungswerkzeuge innerhalb einer Benutzeroberfläche (einer grafischen Benutzeroberfläche (GUI)) auf der Rechenvorrichtung **3308** verwenden.

[0352] In mindestens einer Ausführungsform kann der Benutzer **3310** über die Rechenvorrichtung **3308** mit einer GUI interagieren, um (automatische) Anmerkungen zu bearbeiten oder feinabzustimmen. In mindestens einer Ausführungsform kann ein Polygonbearbeitungsmerkmal verwendet werden, um Scheitelpunkte eines Polygons an genauere oder feinabgestimmte Positionen zu verschieben.

[0353] In mindestens einer Ausführungsform können, sobald der Kundendatensatz **3306** verknüpfte Ground-Truth-Daten aufweist, Ground-Truth-Daten (z. B. aus KI-unterstützter Anmerkung, manueller Kennzeichnung usw.) während des Modelltrainings **3114** verwendet werden, um ein präzisiertes Modell **3312** zu erzeugen. In mindestens einer Ausführungsform kann der Kundendatensatz **3306** beliebig oft auf das Anfangsmodell **3304** angewendet werden und Ground-Truth-Daten können verwendet werden, um Parameter des Anfangsmodells **3304** zu aktualisieren, bis ein zulässiges Genauigkeitsniveau für das präzisierte Modell **3312** erreicht wird. In mindestens einer Ausführungsform kann, sobald das präzisierte Modell **3312** erzeugt ist, das präzisierte Modell **3312** in einer oder mehreren Einsatzpipelines **3210** in einer Einrichtung zum Durchführen einer oder mehrerer Verarbeitungsaufgaben in Bezug auf medizinische Bildgebungsdaten eingesetzt werden.

[0354] In mindestens einer Ausführungsform kann das präzisierte Modell **3312** in das vortrainierte Modelle **3206** in der Modellregistrierungsdatenbank **3124** hochgeladen werden, um von einer anderen Einrichtung ausgewählt zu werden. In mindestens einer Ausführungsform kann dieser Prozess derartig durch eine beliebige Anzahl von Einrichtungen abgeschlossen werden, dass das präzisierte Modell **3312** an neuen Datensätzen eine beliebige Anzahl von Malen weiter präzisiert werden kann, um ein allgemeineres Modell zu erzeugen.

[0355] **Fig. 33B** ist eine beispielhafte Darstellung einer Client-Server-Architektur **3332** zum Verbessern von Anmerkungswerkzeugen mit vortrainierten Anmerkungsmodellen gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform können KI-unterstützte Anmerkungswerkzeuge **3336** auf Grundlage einer Client-Server-Architektur **3332** instanziiert werden. In mindestens einer Ausführungsform können Anmer-

kungswerkzeuge **3336** in Bildgebungsanwendungen Radiologen zum Beispiel dabei helfen, Organe und Anomalien zu identifizieren. In mindestens einer Ausführungsform können Bildgebungsanwendungen Softwarewerkzeuge beinhalten, die dem Benutzer **3310** helfen, als ein nichteinschränkendes Beispiel einige Extrempunkte auf einem bestimmten Organ von Interesse in Rohbildem **3334** (z. B. in einem 3D-MRT oder CT) zu identifizieren, und automatisch kommentierte Ergebnisse für alle 2D-Schnitte eines bestimmten Organs empfangen. In mindestens einer Ausführungsform können Ergebnisse in einem Datenspeicher als Trainingsdaten **3338** gespeichert und als (zum Beispiel und ohne Einschränkung) Ground-Truth-Daten für das Training verwendet werden. In mindestens einer Ausführungsform kann ein Deep-Learning-Modell, wenn die Rechenvorrichtung **3308** Extrempunkte für die KI-unterstützte Anmerkung **3110** sendet, diese Daten beispielsweise als Eingabe empfangen und Ableitungsergebnisse eines segmentierten Organs oder einer segmentierten Anomalie zurückgeben. In mindestens einer Ausführungsform können vorinstanziierte Anmerkungswerkzeuge, wie etwa das AI-unterstützte Anmerkungswerkzeug **3336B** in **Fig. 33B**, durch das Ausführen von API-Aufrufen (z. B. API-Aufruf **3344**) an einen Server, wie etwa einen Anmerkungsunterstützungsserver **3340**, der zum Beispiel einen Satz vortrainierter Modelle **3342** beinhalten kann, die in einer Anmerkungsmodellregistrierungsdatenbank gespeichert sind, verbessert werden. In mindestens einer Ausführungsform kann eine Anmerkungsmodellregistrierungsdatenbank vortrainierte Modelle **3342** (z. B. Modelle des maschinellen Lernens, wie etwa Deep-Learning-Modelle) speichern, die vortrainiert sind, um eine KI-unterstützte Anmerkung an einem bestimmten Organ oder einer bestimmten Anomalie durchzuführen. Diese Modelle können unter Verwendung von Trainingspipelines **3204** weiter aktualisiert werden. In mindestens einer Ausführungsform können vorinstallierte Anmerkungswerkzeuge im Laufe der Zeit verbessert werden, wenn neue gekennzeichnete Klinikdaten **3112** hinzugefügt werden.

[0356] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. In mindestens einer Ausführungsform kann diese Logik mit Komponenten dieser Figuren verwendet werden, um ein dreidimensionales Bild einer bestimmten Ansicht aus zweidimensionalen Bildern aus einer willkürlichen Anzahl von Kameras zu erzeugen.

AUTONOMES FAHRZEUG

[0357] **Fig. 34A** veranschaulicht ein Beispiel eines autonomen Fahrzeugs **3400** gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform kann das autonome Fahrzeug **3400** (alternativ in dieser Schrift als „Fahrzeug **3400**“ bezeichnet) ohne Einschränkung ein Personenkraftwagen sein, wie etwa ein Auto, ein Truck, ein Bus und/oder eine weitere Art von Fahrzeug, die einen oder mehrere Fahrgäste unterbringt. In mindestens einer Ausführungsform kann das Fahrzeug **3400** ein Sattelschlepper sein, der zum Befördern von Fracht verwendet wird. In mindestens einer Ausführungsform kann das Fahrzeug **3400** ein Flugzeug, ein Roboterfahrzeug oder eine andere Art von Fahrzeug sein.

[0358] Autonome Fahrzeuge können hinsichtlich der Automationsstufen beschrieben sein, die durch „Taxonomy and Definitions for Terms Related to Driving Automation Systems for On-Road Motor Vehicles“ (z. B. die Norm Nr. J3016-201806, veröffentlicht am 15. Juni 2018, die Norm Nr. J3016-201609, veröffentlicht am 30. September 2016, und vorherige und zukünftige Versionen dieser Norm) der National Highway Traffic Safety Administration („NHTSA“), einer Abteilung des US-amerikanischen Department of Transportation and Society of Automotive Engineers („SAE“), definiert sind. In einer oder mehreren Ausführungsformen kann das Fahrzeug **3400** zu der Funktion gemäß einer oder mehreren der Stufe 1 - Stufe 5 der autonomen Fahrstufen in der Lage sein. Zum Beispiel kann das Fahrzeug **3400** in mindestens einer Ausführungsform zu bedingter Automation (Stufe 3), hoher Automation (Stufe 4) und/oder vollständiger Automation (Stufe 5) abhängig von der Ausführungsform in der Lage sein.

[0359] In mindestens einer Ausführungsform kann das Fahrzeug **3400** ohne Einschränkung Komponenten beinhalten, wie etwa ein Fahrgestell, eine Fahrzeugkarosserie, Räder (z. B. 2, 4, 6, 8, 18 usw.), Reifen, Achsen und andere Komponenten eines Fahrzeugs. In mindestens einer Ausführungsform kann das Fahrzeug **3400** ohne Einschränkung ein Antriebssystem **3450** beinhalten, wie etwa einen internen Verbrennungsmotor, eine hybridelektrische Kraftanlage, einen vollektrischen Motor und/oder eine weitere Art von Antriebssystem. In mindestens einer Ausführungsform kann das Antriebssystem **3450** mit einem Antriebsstrang des Fahrzeugs **3400** verbunden sein, was ohne Einschränkung ein Getriebe, beinhalten kann, um den Antrieb des Fahrzeugs **3400** zu ermöglichen. In mindestens einer Ausführungsform kann das Antriebssystem **3450** als Reaktion auf das Empfangen von Signalen von einer Drossel/Gaspedal(en) **3452** gesteuert werden.

[0360] In mindestens einer Ausführungsform wird ein Lenksystem **3454**, das ohne Einschränkung ein Lenkrad beinhalten kann, verwendet, um ein Fahrzeug **3400** zu lenken (z. B. entlang einem gewünschten Weg oder einer gewünschten Strecke), wenn ein Antriebssystem **3450** betrieben wird (z. B. wenn sich das Fahrzeug bewegt). In mindestens einer Ausführungsform kann ein Lenksystem **3454** Signale von (einem) Lenkaktor(en) **3456** empfangen. Ein Lenkrad kann optional für die Funktion der vollständigen Automation (Stufe 5) sein. In mindestens einer Ausführungsform kann ein Bremssensorsystem **3446** verwendet werden, um Fahrzeugbremsen als Reaktion auf das Empfangen von Signalen von (einem) Bremsaktor(en) **3448** und/oder Bremssensoren zu betreiben.

[0361] In mindestens einer Ausführungsform stellen eine oder mehrere Steuerungen **3436**, die ohne Einschränkung einen oder mehrere Systeme auf Chips (system on chips - „SoCs“) (in **Fig. 34A** nicht gezeigt) und/oder eine oder mehrere Grafikverarbeitungseinheiten (graphics processing units - „GPUs“) Signale (z. B. für Befehle stehend) einer oder mehreren Komponenten und/oder Systemen des Fahrzeugs **3400** bereit. Zum Beispiel können in mindestens einer Ausführungsform die eine oder die mehreren Steuerungen **3436** Signale senden, um Fahrzeugbremsen über einen oder mehrere Bremsaktoren **3448** zu betreiben, das Lenksystem **3454** über einen oder mehrere Lenkaktoren **3456** zu betreiben und/oder das Antriebssystem **3450** über ein(e) oder mehrere Drosseln/Gaspedale **3452** zu betreiben. Die eine oder die mehreren Steuerungen **3436** können eine oder mehrere fahrzeuginterne (z. B. integrierte) Rechenvorrichtungen (z. B. Supercomputer) beinhalten, die Sensorsignale verarbeiten und Betriebsbefehle (z. B. Signale, die für Befehle stehen) ausgeben, um das autonome Fahren zu ermöglichen und/oder einen menschlichen Fahrer beim Fahren des Fahrzeugs **3400** zu unterstützen. In mindestens einer Ausführungsform können die eine oder die mehreren Steuerungen **3436** eine erste Steuerung **3436** für Funktionen autonomen Fahrens, eine zweite Steuerung **3436** für Funktionen der funktionellen Sicherheit, eine dritte Steuerung **3436** für die Funktion der künstlichen Intelligenz (z. B. Maschinelles Sehen), eine vierte Steuerung **3436** für die Funktion des Infotainment, eine fünfte Steuerung **3436** für die Redundanz in Notfallzustände und/oder andere Steuerungen beinhalten. In mindestens einer Ausführungsform kann eine einzelne Steuerung **3436** zwei oder mehrere der vorstehenden Funktionen handhaben, zwei oder mehrere Steuerungen **3436** können eine einzelne Funktion handhaben und/oder eine beliebige Kombination daraus.

[0362] In mindestens einer Ausführungsform können eine oder mehrere Steuerungen **3436** Signale zum Steuern einer oder mehrerer Komponenten und/oder Systeme des Fahrzeugs **3400** als Reaktion auf Sensordaten bereitstellen, die von einem oder mehreren Sensoren empfangen wurden (z. B. Sensoreingaben). In mindestens einer Ausführungsform können Sensordaten zum Beispiel und ohne Einschränkung von Folgenden empfangen werden: Sensoren von globalen Navigationssatellitensystemen („GNSS“) **3458** (z. B. einen oder mehrere Sensoren von globalen Positionsbestimmungssystemen), einen oder mehrere Radarsensorsen **3460**, einen oder mehrere Ultraschallsensoren **3462**, einen oder mehrere Lidarsensoren **3464**, einen oder mehrere Sensoren von Trägheitsmesseinheiten (inertial measurement unit - „IMU“) **3466** (z. B. einen oder mehrere Beschleunigungssensoren, ein oder mehrere Gyroskope, einen oder mehrere Magnetkompasse, ein oder mehrere Magnetometer usw.), ein oder mehrere Mikrofone **3496**, eine oder mehrere Stereokameras **3468**, eine oder mehrere Weitsichtkameras **3470** (z. B. Fischaugenkameras), eine oder mehrere Infrarotkameras **3472**, eine oder mehrere Rundumkameras **3474** (z. B. 360-Grad-Kameras), Langstreckenkameras (in **Fig. 34A** nicht gezeigt), eine oder mehrere Mittelstreckenkameras (in **Fig. 34A** nicht gezeigt), einen oder mehrere Drehzahlsensoren **3444** (z. B. zum Messen der Drehzahl des Fahrzeugs **3400**), einen oder mehrere Vibrationssensoren **3442**, einen oder mehrere Lenksensoren **3440**, einen oder mehrere Bremssensoren (z. B. als Teil des Bremssensorsystems **3446**) und/oder andere Sensorarten.

[0363] In mindestens einer Ausführungsform können eine oder mehrere der einen oder mehreren Steuerungen **3436** Eingaben (z. B. durch Eingangsdaten dargestellt) von einem Kombiinstrument **3432** des Fahrzeugs **3400** empfangen und Ausgaben (z. B. durch Ausgabedaten, Anzeigedaten usw. dargestellt) über eine Anzeige **3434** der Mensch-Maschine-Schnittstelle (human-machine interface - „HMI“), eine akustische Meldeeinrichtung, einen Lautsprecher und/oder über andere Komponenten des Fahrzeugs **3400** bereitstellen. In mindestens einer Ausführungsform können Ausgaben Informationen beinhalten, wie etwa Fahrzeuggeschwindigkeit, Drehzahl, Zeit, Kartendaten (z. B. eine hochauflösende Karte (in **Fig. 34A** nicht gezeigt), Standortdaten (z. B. den Standort des Fahrzeugs **3400**, wie etwa auf einer Karte), Richtung, Standort von anderen Fahrzeugen (z. B. ein Belegungsraster), Informationen zu Objekten und dem Zustand von Objekten, wie durch die eine oder mehreren Steuerungen **3436** wahrgenommen usw. Zum Beispiel kann in mindestens einer Ausführungsform die HMI-Anzeige **3434** Informationen zur Anwesenheit von einem oder mehreren Objekten (z. B. einem Straßenschild, einem Warnschild, einer Ampeländerung usw.) und/oder Informationen zu Fahrmanövern anzeigen, die ein Fahrzeug vorgenommen hat, vornimmt oder vornehmen wird (z. B. das sofortige Ändern von Spuren, das Nehmen der Ausfahrt **34B** in zwei Meilen usw.).

[0364] In mindestens einer Ausführungsform beinhaltet das Fahrzeug **3400** ferner eine Netzwerkschnittstelle **3424**, wie eine oder mehrere drahtlose Antennen **3426** und/oder ein oder mehrere Modems verwendet, um über ein oder mehrere Netzwerke zu kommunizieren. Zum Beispiel kann in mindestens einer Ausführungsform die Netzwerkschnittstelle **3424** zur Kommunikation über Folgendes in der Lage sein: langfristige Entwicklung (Long-Term Evolution - „LTE“), Breitbandcodeteilungsmehrfachzugriff (Wideband Code Division Multiple Access - „WCDMA“), das universelle Mobilfunktelekommunikationssystem (Universal Mobile Telecommunications System - „UMTS“), das globale System für Mobilfunkkommunikation (Global System for Mobile communication - „GSM“), IMT-CDMA Mehrfachträger („CDMA2000“) usw. In mindestens einer Ausführungsform können eine oder mehrere drahtlose Antennen **3426** außerdem die Kommunikation zwischen Objekten in der Umgebung (z.B. Fahrzeugen, mobilen Vorrichtungen usw.) unter Verwendung von einem oder mehreren lokalen Netzwerken, wie etwa Bluetooth, Bluetooth Low Energy („LE“), Z-Wave, ZigBee usw. und/oder einem oder mehreren Weitbereichsnetzwerken mit niedriger Leistung (low power wide-area networks - „LPWANs“) ermöglichen, wie etwa LoRaWAN, SigFox usw.

[0365] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. Details bezüglich der Ableitungs- und/oder Trainingslogik **615** werden nachfolgend in Verbindung mit den **Fig. 6A** und/oder 6B bereitgestellt. In mindestens einer Ausführungsform kann eine Ableitungs- und/oder Trainingslogik **615** im System der **Fig. 34A** zum Ableiten oder Vorhersagen von Operationen mindestens zum Teil auf Grundlage von Gewichtungsparametern verwendet werden, die unter Verwendung von neuronalen Netzwerktrainingsoperationen, Funktionen und/oder Architekturen von neuronalen Netzwerken oder in dieser Schrift beschriebenen Anwendungsfällen von neuronalen Netzwerken berechnet wurden.

[0366] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. In mindestens einer Ausführungsform kann diese Logik mit Komponenten dieser Figuren verwendet werden, um ein dreidimensionales Bild einer bestimmten Ansicht aus zweidimensionalen Bildern aus einer willkürlichen Anzahl von Kameras zu erzeugen. **Fig. 34B** veranschaulicht ein Beispiel von Kamerapostionen und -sichtfeldern für das autonome Fahrzeug **3400** aus **Fig. 34A** gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform sind Kameras und entsprechende Sichtfelder eine beispielhafte Ausführungsform und nicht dazu gedacht, einschränkend zu sein. Zum Beispiel können in mindestens einer Ausführungsform zusätzlich oder alternativ Kameras beinhaltet sein und/oder können sich Kameras an verschiedenen Positionen am Fahrzeug **3400** befinden.

[0367] In mindestens einer Ausführungsform können Kameraarten für Kameras Digitalkameras beinhalten, ohne darauf beschränkt zu sein, die zur Verwendung mit Komponenten und/oder Systemen des Fahrzeugs **3400** angepasst sein können. In mindestens einer Ausführungsform können eine oder mehrere der einen oder mehreren Kameras auf einer Automobilsicherheitsintegritätsstufe (automotive safety integrity level - „ASIL“) B und/oder auf einer weiteren ASIL betrieben werden. In mindestens einer Ausführungsform können Kameraarten zu einer beliebigen Bildaufnahmegeschwindigkeit, wie etwa 60 Bilder pro Sekunde (frames per second - fps), 120 fps, 240 fps usw. abhängig von der Ausführungsform in der Lage sein. In mindestens einer Ausführungsform können Kameras zum Verwenden von rollenden Verschlüssen, globalen Verschlüssen, einer weiteren Art von Verschluss oder einer Kombination daraus in der Lage sein. In mindestens einer Ausführungsform kann eine Farbfilteranordnung Folgendes beinhalten: eine Rot-Klar-Klar-Klar-(red clear clear clear - „RCCC“-)Farbfilteranordnung, eine Rot-Klar-Klar-Blau-(red clear clear blue - „RCCB“-)Farbfilteranordnung, eine Rot-Blau-Grün-Klar-(red blue green clear - „RBGC“-)Farbfilteranordnung, eine Foveon-X3-Farbfilteranordnung, eine Bayer-Sensoren-(„RGGB“-)Farbfilteranordnung, eine Monochromsensor-Farbfilteranordnung und/oder eine weitere Art von Farbfilteranordnung. In mindestens einer Ausführungsform können Klar-Bildpunkt-Kameras, wie etwa Kameras mit einer RCCC-, einer RCCB- und/oder einer RBGC-Farbfilteranordnung, mit dem Ziel verwendet werden, die Lichtempfindlichkeit zu erhöhen.

[0368] In mindestens einer Ausführungsform können eine oder mehrere der einen oder mehreren Kameras verwendet werden, um Funktionen von Systemen zur erweiterten Fahrerunterstützung (advanced driver assistance systems - „ADAS“) (z. B. als Teil einer redundanten oder ausfallsicheren Ausgestaltung) verwendet werden. Zum Beispiel kann in mindestens einer Ausführungsform eine Mehrfunktionen-Monokamera eingebaut sein, um Funktionen bereitzustellen, die eine Spurverlassenswarnung, eine Verkehrschildunterstützung und eine intelligente Scheinwerfersteuerung beinhalten. In mindestens einer Ausführungsform können eine oder mehrere der einen oder mehreren Kameras (z. B. alle der Kameras) Bilddaten (z. B. Video) gleichzeitig aufzeichnen und bereitstellen.

[0369] In mindestens einer Ausführungsform können eine oder mehrere der Kameras in einer Befestigungsbaugruppe befestigt sein, wie etwa einer individuell ausgestalteten (dreidimensionalen („3D“-)gedruckten) Baugruppe, um Streulicht und Reflexionen von innerhalb des Autos zu verdrängen (z. B. Reflexionen vom Armaturenbrett, die in den Windschutzscheibenspiegeln reflektiert werden), welche die Bilddatenaufnahmefähigkeiten von Kameras stören können. Unter Bezugnahme auf Seitenspiegelbefestigungsbaugruppen können in mindestens einer Ausführungsform Seitenspiegelbaugruppen individuell 3D-gedruckt sein, sodass die Kamerabefestigungsplatte mit der Form des Seitenspiegels übereinstimmt. In mindestens einer Ausführungsform können eine oder mehrere Kameras in den Seitenspiegel integriert sein. Für Außenspiegelkameras können außerdem eine oder mehrere Kameras innerhalb von vier Säulen an jeder Ecke der Kabine in mindestens einer Ausführungsform integriert sein.

[0370] In mindestens einer Ausführungsform können Kameras mit einem Sichtfeld, das Abschnitte der Umgebung vor dem Fahrzeug **3400** beinhaltet (z. B. nach vorn gerichtete Kameras) für eine Rundumsicht verwendet werden, um dabei zu helfen, nach vorn gerichtete Wege und Hindernisse zu identifizieren sowie dabei zu unterstützen, mit Hilfe von einer oder mehreren Steuerungen **3436** und/oder Steuerungs-SoCs, Informationen bereitzustellen, die ausschlaggebend für das Erzeugen eines Belegungsrasters und/oder das Bestimmen bevorzugter Fahrzeugwege sind. In mindestens einer Ausführungsform können nach vom gerichtete Kameras verwendet werden, um viele selbe ADAS-Funktionen, wie LIDAR, durchzuführen, was ohne Einschränkung Notfallbremsen, Fußgängererfassung und Zusammenstoßvermeidung beinhaltet. In mindestens einer Ausführungsform können nach vorn gerichtete Kameras außerdem für ADAS-Funktionen und Systeme verwendet werden, die ohne Einschränkung Spurverlassenswarnungen (Lane Departure Warnings - „LDW“), autonome Geschwindigkeitssteuerung (Autonomous Cruise Control - „ACC“) und/oder andere Funktionen beinhaltet, wie etwa Verkehrsschilderkennung.

[0371] In mindestens einer Ausführungsform kann eine Reihe von Kameras bei einer nach vom gerichteten Konfiguration verwendet werden, was zum Beispiel eine monokulare Kameraplattform beinhaltet, die einen CMOS-(„complementary metal oxide semiconductor“ - komplementären Metalloxidhalbleiter) Farbbildgeber beinhaltet. In mindestens einer Ausführungsform kann eine Weitsichtkamera **3470** verwendet werden, um Objekte wahrzunehmen, die von einem Außenrand ins Blickfeld gelangen (z. B. Fußgänger, Querverkehr oder Fahrräder). Obwohl in **Fig. 34B** lediglich eine Weitsichtkamera **3470** veranschaulicht ist, kann es in anderen Ausführungsformen eine beliebige Anzahl (die Null beinhaltet) von einer oder mehreren Weitsichtkameras **3470** am Fahrzeug **3400** geben. In mindestens einer Ausführungsform kann eine beliebige Anzahl von einer oder mehreren Langstreckenkameras **3498** (z. B. ein Weitsichtstereokamerapaar) für tiefenbasierte Objekterfassung verwendet werden, insbesondere für Objekte, für die ein neuronales Netzwerk noch nicht trainiert wurde. In mindestens einer Ausführungsform können eine oder mehrere Langstreckenkameras **3498** außerdem für die Objekterfassung und -zuordnung sowie als einfache Objektverfolgung verwendet werden.

[0372] In mindestens einer Ausführungsform kann eine beliebige Anzahl von einer oder mehreren Stereokameras **3468** außerdem in einer nach vorn gerichteten Konfiguration beinhaltet sein. In mindestens einer Ausführungsform können eine oder mehrere der einen oder mehreren Stereokameras **3468** eine integrierte Steuereinheit, die eine skalierbare Verarbeitungseinheit umfasst, die eine programmierbare Logik („FPGA“) bereitstellen kann, und einen Mehrkernmikroprozessor mit einer integrierten Controller-Area-Network- („CAN“-) oder Ethernet-Schnittstelle auf einem einzelnen Chip beinhalten. In mindestens einer Ausführungsform kann eine derartige Einheit verwendet werden, um eine 3D-Karte der Umgebung des Fahrzeugs **3400** zu erzeugen, was eine Entfernungsschätzung für alle Punkte in dem Bild beinhaltet. In mindestens einer Ausführungsform können eine oder mehrere der einen oder mehreren Stereokameras **3468** ohne Einschränkung einen oder mehrere Kompaktstereobildsensoren beinhalten, die ohne Einschränkung zwei Kameralinsen (jeweils einen rechts und links) und einen Bildverarbeitungschip beinhalten können, die eine Entfernung vom Fahrzeug **3400** zum Zielobjekt messen und erzeugte Informationen (z. B. Metadaten) verwenden können, um autonome Notfallbrems- und Spurverlassenswarnfunktionen einzuschalten. In mindestens einer Ausführungsform können andere Arten der einen oder der mehreren Stereokameras **3468** zusätzlich zu oder alternativ zu jenen in dieser Schrift beschriebenen verwendet werden.

[0373] In mindestens einer Ausführungsform können Kameras mit einem Sichtfeld, das Abschnitte der Umgebung seitlich von dem Fahrzeug **3400** beinhaltet (z.B. Seitenansichtkameras), für eine Rundumsicht verwendet werden, die Informationen bereitstellt, die verwendet werden, um ein Belegungsraster zu erzeugen und zu aktualisieren sowie um Seitenaufprallzusammenstoßwarnungen zu erzeugen. Zum Beispiel könnten in mindestens einer Ausführungsform eine oder mehrere Rundumkameras **3474** (z. B. vier Rundumkameras **3474**, wie in **Fig. 34B** veranschaulicht) am Fahrzeug **3400** positioniert sein. In mindestens einer Ausführungsform können eine oder mehrere Rundumkameras **3474** ohne Einschränkung eine beliebige Anzahl und Kombination von

einer oder mehreren Weitsichtkameras **3470**, einer oder mehreren Fischaugenkameras, einer oder mehreren 360-Grad-Kameras und/oder ähnliches beinhalten. Zum Beispiel können in mindestens einer Ausführungsform vier Fischaugenkameras vorn, hinten und an den Seiten des Fahrzeugs **3400** positioniert sein. In mindestens einer Ausführungsform kann das Fahrzeug **3400** drei Rundumkameras **3474** (z. B. links, rechts und hinten) verwenden und kann eine oder mehrere andere Kameras (z. B. eine nach vorn gerichtete Kamera) als eine vierte Rundumsichtkamera nutzen.

[0374] In mindestens einer Ausführungsform können Kameras mit einem Sichtfeld, das Abschnitte der Umgebung hinter dem Fahrzeug **3400** beinhaltet (z. B. Rückfahrkameras), zur Parkunterstützung, Rundumsicht, rückwärtigen Zusammenstoßwarnungen und zum Erzeugen und Aktualisieren eines Belegungsrasters verwendet werden. In mindestens einer Ausführungsform kann ein breites Spektrum von Kameras verwendet werden, was Kameras beinhaltet, ohne darauf beschränkt zu sein, die ebenfalls als eine oder mehrere nach vorn gerichtete Kameras (z.B. Langstreckenkameras **3498** und/oder eine oder mehrere Mittelstreckenkameras **3476**, eine oder mehrere Stereokameras **3468**), eine oder mehrere Infrarotkameras **3472** usw.), verwendet werden können, wie in dieser Schrift beschrieben.

[0375] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. Details bezüglich der Ableitungs- und/oder Trainingslogik **615** werden nachfolgend in Verbindung mit den **Fig. 6A** und/oder 6B bereitgestellt. In mindestens einer Ausführungsform kann eine Ableitungs- und/oder Trainingslogik **615** im System der **Fig. 34B** zum Ableiten oder Vorhersagen von Operationen mindestens zum Teil auf Grundlage von Gewichtungsparametern verwendet werden, die unter Verwendung von neuronalen Netzwerktrainingsoperationen, Funktionen und/oder Architekturen von neuronalen Netzwerken oder in dieser Schrift beschriebenen Anwendungsfällen von neuronalen Netzwerken berechnet wurden.

[0376] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. In mindestens einer Ausführungsform kann diese Logik mit Komponenten dieser Figuren verwendet werden, um ein dreidimensionales Bild einer bestimmten Ansicht aus zweidimensionalen Bildern aus einer willkürlichen Anzahl von Kameras zu erzeugen.

[0377] **Fig. 34C** ist ein Blockdiagramm, das eine beispielhafte Systemarchitektur für das autonome Fahrzeug **3400** aus **Fig. 34A** gemäß mindestens einer Ausführungsform veranschaulicht. In mindestens einer Ausführungsform ist jeder der Komponenten, jedes der Merkmale und jedes der Systeme des Fahrzeugs **3400** in **Fig. 34C** als über einen Bus **3402** verbunden veranschaulicht. In mindestens einer Ausführungsform kann der Bus **3402** ohne Einschränkung eine CAN-Datenschnittstelle (alternativ in dieser Schrift als ein „CAN-Bus“ bezeichnet) beinhalten. In mindestens einer Ausführungsform kann ein CAN-Bus ein Netzwerk in dem Fahrzeug **3400** sein, das verwendet wird, um bei der Steuerung verschiedener Merkmale und Funktionen des Fahrzeugs **3400** zu unterstützen, wie etwa die Betätigung der Bremsen, der Beschleunigung, des Bremsens, des Lenkens, der Windschutzscheibenwischer usw. In mindestens einer Ausführungsform kann der Bus **3402** dazu konfiguriert sein, dutzende oder sogar hunderte von Knoten aufzuweisen, jeder mit seiner eigenen einzigartigen Kennung (z. B. eine CAN-ID). In mindestens einer Ausführungsform kann der Bus **3402** ausgelesen werden, um den Lenkradwinkel, die Geschwindigkeit über Grund, Motorumdrehungen pro Minute („RPMs“), Schalterstellungen und/oder andere Zustandsanzeiger zu finden. In mindestens einer Ausführungsform kann der Bus **3402** ein CAN-Bus sein, der mit ASIL B kompatibel ist.

[0378] In mindestens einer Ausführungsform kann zusätzlich zu oder alternativ zu CAN, FlexRay und/oder Ethernet verwendet werden. In mindestens einer Ausführungsform kann es eine beliebige Anzahl von Bussen **3402** geben, die ohne Einschränkung null oder mehr CAN-Busse, null oder mehr FlexRay-Busse, null oder mehr Ethernet-Busse und/oder null oder mehr andere Arten von Bussen unter Verwendung eines unterschiedlichen Protokolls beinhalten. In mindestens einer Ausführungsform können zwei oder mehr Busse **3402** verwendet werden, um unterschiedliche Funktionen durchzuführen und/oder können zu Redundanzzwecken verwendet werden. Zum Beispiel kann ein erster Bus **3402** zur Zusammenstoßvermeidungsfunktion verwendet werden und ein zweiter Bus **3402** kann zur Betätigungssteuerung verwendet werden. In mindestens einer Ausführungsform kann jeder Bus **3402** mit einer beliebigen der Komponenten des Fahrzeugs **3400** kommunizieren und zwei oder mehr Busse **3402** können mit selben Komponenten kommunizieren. In mindestens einer Ausführungsform können eine beliebige Anzahl von einem oder mehreren Systemen auf einem oder mehreren Chips (system(s) on chip(s) - „SOC(s)“) **3404**, jede der einen oder mehreren Steuerungen **3436** und/oder jeder Computer innerhalb des Fahrzeugs Zugriff auf dieselben Eingabedaten (z. B. Eingaben von Sensoren des

Fahrzeugs **3400**) aufweisen und können mit einem gemeinsamen Bus, wie etwa dem CAN-Bus, verbunden sein.

[0379] In mindestens einer Ausführungsform kann das Fahrzeug **3400** eine oder mehrere Steuerungen **3436** beinhalten, wie etwa die in dieser Schrift im Hinblick auf **Fig. 34A** beschriebenen. Eine oder mehrere Steuerungen **3436** können für eine Reihe von Funktionen verwendet werden. In mindestens einer Ausführungsform können die eine oder die mehreren Steuerungen **3436** an eine beliebige von verschiedenen anderen Komponenten und Systemen des Fahrzeugs **3400** gekoppelt sein und können zur Steuerung des Fahrzeugs **3400**, der künstlichen Intelligenz des Fahrzeugs **3400**, das Infotainment des Fahrzeugs **3400** und/oder ähnliches verwendet werden.

[0380] In mindestens einer Ausführungsform kann das Fahrzeug **3400** eine beliebige Anzahl von SoCs **3404** beinhalten. Jedes der SoCs **3404** kann ohne Einschränkung Zentralverarbeitungseinheiten (central processing units - „CPU(s)“) **3406**, Grafikverarbeitungseinheiten (graphics processing units - „GPU(s)“ **3408**, einen oder mehrere Prozessoren **3410**, einen oder mehrere Zwischenspeicher **3412**, ein oder mehrere Gaspedale **3414**, einen oder mehrere Datenspeicher **3416** und/oder andere Komponenten und Merkmale beinhalten, die nicht veranschaulicht sind. In mindestens einer Ausführungsform können das eine oder die mehreren SOCs **3404** verwendet werden, um das Fahrzeug **3400** in einer Reihe von Plattformen und Systemen zu steuern. Zum Beispiel können in mindestens einer Ausführungsform das eine oder die mehreren SOCs **3404** in einem System (z. B. ein System des Fahrzeugs **3400**) mit einer hochauflösenden (High Definition - „HD“) Karte **3422** kombiniert werden, die Kartenauffrischungen und/oder -aktualisierungen über die Netzwerkschnittstelle **3424** von einem oder mehreren Servern (in **Fig. 34C** nicht gezeigt) erhalten können.

[0381] In mindestens einer Ausführungsform können eine oder mehrere CPUs **3406** einen CPU-Cluster oder einen CPU-Komplex (alternativ in dieser Schrift als ein „CCPLEX“ bezeichnet) beinhalten. In mindestens einer Ausführungsform können die eine oder die mehreren CPUs **3406** mehrere Kerne und/oder Zwischenspeicher der Stufe zwei („L2“) beinhalten. Zum Beispiel können in mindestens einer Ausführungsform die eine oder die mehreren CPUs **3406** acht Kerne in einer zusammenhängenden Mehrprozessorkonfiguration beinhalten. In mindestens einer Ausführungsform können die eine oder die mehreren CPUs **3406** vier Dualkerncluster beinhalten, bei denen jeder Cluster einen dedizierten L2-Zwischenspeicher (z. B. einen L2-Zwischenspeicher mit 2 MB) aufweist. In mindestens einer Ausführungsform können die eine oder die mehreren CPUs **3406** (z. B. CCPLEX) dazu konfiguriert sein, gleichzeitigen Clusterbetrieb zu unterstützen, was ermöglicht, dass eine beliebige Kombination von Clustern der einen oder der mehreren CPUs **3406** zu einem beliebigen bestimmten Zeitpunkt eingeschaltet ist.

[0382] In mindestens einer Ausführungsform können eine oder mehrere der einen oder der mehreren CPUs **3406** Leistungsverwaltungsfähigkeiten umsetzen, die ohne Einschränkung eines oder mehrere der folgenden Merkmale beinhalten: individuelle Hardwareblöcke können automatisch taktgesteuert sein, wenn sie sich im Ruhezustand befinden, um dynamische Leistung zu sparen; jeder Kerntakt kann gesteuert sein, wenn der Kern nicht aufgrund einer Ausführung von Anweisungen zum Warten auf eine Unterbrechung (Wait for Interrupt - „WFI“)/Warten auf ein Ereignis (Wait for Event - „WFE“) Anweisungen aktiv steuert; jeder Kern kann unabhängig leistungsgesteuert sein; jeder Kern kann unabhängig taktgesteuert sein, wenn alle Kerne taktgesteuert oder leistungsgesteuert sind; und/oder jeder Kerncluster kann unabhängig leistungsgesteuert sein, wenn alle Kerne leistungsgesteuert sind. In mindestens einer Ausführungsform können die eine oder die mehreren CPUs **3406** ferner einen verbesserten Algorithmus zum Verwalten von Leistungszuständen umsetzen, wobei zulässige Leistungszustände und erwartete Weckzeitpunkte festgelegt sind und Hardware/Mikrocode den besten Leistungszustand bestimmt, in den der Kern, der Cluster oder der CCPLEX übergeht. In mindestens einer Ausführungsform können Verarbeitungskerne vereinfachte Leistungszustandsübergangssequenzen in Software unterstützen, wobei Arbeit an Mikrocode ausgelagert wird.

[0383] In mindestens einer Ausführungsform können die eine oder die mehreren GPUs **3408** eine integrierte GPU (alternativ in dieser Schrift als eine „iGPU“ bezeichnet) beinhalten. In mindestens einer Ausführungsform können die eine oder die mehreren GPUs **3408** programmierbar sein und können effizient für eine parallele Arbeitslast sein In mindestens einer Ausführungsform können die eine oder die mehreren GPUs **3408** in mindestens einer Ausführungsform einen verbesserten Tensoranweisungssatz verwenden. In mindestens einer Ausführungsform können die eine oder die mehreren GPUs **3408** einen oder mehrere Streaming-Mikroprozessoren beinhalten, wobei jeder Streaming-Mikroprozessor einen Zwischenspeicher der Stufe eins („L1“) (z. B. einen L1-Zwischenspeicher mit mindestens 96 KB Speicherkapazität) beinhalten kann, und zwei oder mehr der Streaming-Mikroprozessoren können einen L2-Zwischenspeicher (z. B. einen L2-Zwischenspeicher mit einer Speicherkapazität von 512 KB) gemeinsam nutzen. In mindestens einer Ausführungsform können die

eine oder die mehreren GPUs **3408** mindestens acht Streaming-Mikroprozessoren beinhalten. In mindestens einer Ausführungsform können die eine oder die mehreren GPUs **3408** eine oder mehrere Anwendungsprogrammierschnittstellen (application programming interface(s) - API(s)) berechnen. In mindestens einer Ausführungsform können die eine oder die mehreren GPUs **3408** eine oder mehrere parallele Rechenplattformen und/oder Programmierungsmodelle (z. B. CUDA von NVIDIA) verwenden.

[0384] In mindestens einer Ausführungsform können eine oder mehrere der einen oder der mehreren GPUs **3408** für die beste Leistung in Automobil- und eingebetteten Anwendungsfällen leistungsoptimiert sein. Zum Beispiel könnten in einer Ausführungsform die eine oder die mehreren GPUs **3408** auf einem Steg-Feldeffekttansistor (Fin field-effect transistor - „FinFET“) gefertigt sein. In mindestens einer Ausführungsform kann jeder Streaming-Mikroprozessor eine Anzahl von Verarbeitungskernen mit unterschiedlicher Genauigkeit enthalten, die in mehrere Blöcke unterteilt sind. Zum Beispiel und ohne Einschränkung könnten 64 PF32-Keme und 32 PF64-Kerne in vier Verarbeitungsblöcke unterteilt sein. In mindestens einer Ausführungsform könnte jedem Verarbeitungsblock 16 FP32-Kerne, 8 FP64-Kerne, 16 INT32-Kerne, zwei NVIDIA-TENSOR-KERNE mit unterschiedlicher Genauigkeit zur Deep-Learning-Matrix-Arithmetik, ein Anweisungszwischenspeicher der Stufe null („L0“), ein Warp-Planer, eine Versandeinheit und/oder eine Registerdatei mit 64 KB zugewiesen sein. In mindestens einer Ausführungsform können Streaming-Mikroprozessoren unabhängige parallele Ganzzahl- und Gleitkomma-Datenpfade beinhalten, um eine effiziente Ausführung von Arbeitslasten mit einer Mischung aus Berechnungs- und Adressierungsberechnungen bereitzustellen. In mindestens einer Ausführungsform können Streaming-Mikroprozessoren unabhängige Thread-Planungsfähigkeiten beinhalten, um eine feinkörnigere Synchronisation und Zusammenarbeit zwischen parallelen Threads zu ermöglichen. In mindestens einer Ausführungsform können Streaming-Mikroprozessoren eine kombinierte Einheit aus L1-Datenzwischenspeicher und gemeinsam genutztem Speicher beinhalten, um die Leistung zu verbessern bei gleichzeitiger Vereinfachung der Programmierung.

[0385] In mindestens einer Ausführungsform können eine oder mehrere der einen oder der mehreren GPUs **3408** einen Speicher mit hoher Bandbreite (high bandwidth memory - „HBM“) und/oder ein HBM2-Speicherteilsystem mit 16 GB beinhalten, um in einigen Beispielen ungefähr 900 GB/Sekunde an Spitzenspeicherbandbreite bereitzustellen. In mindestens einer Ausführungsform kann zusätzlich zum oder alternativ zum HBM-Speicher ein synchroner Grafikdirektzugriffsspeicher (synchronous graphics random-access memory - „SGRAM“) verwendet werden, wie etwa ein synchroner Grafikdirektzugriffsspeicher mit doppelter Übertragungsrate vom Typ 5 (graphics double data rate type five synchronous random-access memory - „GDDR5“).

[0386] In mindestens einer Ausführungsform können die eine oder die mehreren GPUs **3408** eine einheitliche Speichertechnik beinhalten. In mindestens einer Ausführungsform kann die Unterstützung durch Adressübersetzungsdienste (address translation services - „ATS“) verwendet werden, um der einen oder den mehreren GPUs **3408** zu ermöglichen, direkt auf die Seitentabellen der einen oder der mehreren CPUs **3406** zuzugreifen. In mindestens einer Ausführungsform kann, wenn die Speicherverwaltungseinheit (memory management unit - „MMU“) der einen oder der mehreren GPUs **3408** einen Speicherfehler erlebt, eine Adressübersetzungsanforderung an die eine oder die mehreren CPUs **3406** übermittelt werden. Als Reaktion darauf kann die eine oder die mehreren CPUs **3406** in mindestens einer Ausführungsform in ihren Seitentabellen nach einer virtuell-zu-physisch-Zuordnung für die Adresse suchen und die Übersetzung zurück zu der einen oder den mehreren GPUs **3408** übermitteln. In mindestens einer Ausführungsform kann die einheitliche Speichertechnik einem einzelnen einheitlichen virtuellen Adressbereich für sowohl die eine oder die mehreren CPUs **3406** als auch die eine oder die mehreren GPUs **3408** zulassen, wodurch die Programmierung und Portierung von Anwendungen der einen oder der mehreren GPUs **3408** zu der einen oder den mehreren GPUs **3408** vereinfacht wird.

[0387] In mindestens einer Ausführungsform können die eine oder die mehreren GPUs **3408** eine beliebige Anzahl von Zugriffszählern beinhalten, die den Überblick über eine Frequenz des Zugriffs der einen oder der mehreren GPUs **3408** auf Speicher von anderen Prozessoren behält. In mindestens einer Ausführungsform können der/die Zugriffszähler dabei helfen sicherzustellen, dass Speicherseiten in physischen Speicher des Prozessors verschoben werden, der am häufigsten auf die Seiten zugreift, wodurch die Leistungsfähigkeit für Speicherbereichen verbessert wird, die gemeinsam von Prozessoren genutzt werden.

[0388] In mindestens einer Ausführungsform können einer oder mehrere des einen oder der mehreren SoCs **3404** eine beliebige Anzahl von Zwischenspeichern **3412** beinhalten, was jene in dieser Schrift beschriebenen beinhaltet. Zum Beispiel könnten in mindestens einer Ausführungsform der eine oder die mehreren Zwischenspeicher **3412** einen Zwischenspeicher der Stufe drei („L3“) beinhalten, der sowohl für die eine oder die mehreren CPUs **3406** als auch für die eine oder die mehreren GPUs **3408** verfügbar ist (der z. B. mit sowohl der einen oder den mehreren CPUs **3406** als auch mit der einen oder den mehreren GPUs **3408** verbunden

ist). In mindestens einer Ausführungsform können der eine oder die mehreren Zwischenspeicher **3412** einen Rückschreibzwischenspeicher beinhalten, der den Überblick über die Zustände von Zeilen behalten kann, wie etwa durch das Verwenden eines Zwischenspeicherkohärenzprotokolls (z. B. MEI, MESI, MSI usw.). In mindestens einer Ausführungsform kann L3-Zwischenspeicher 4 MB oder mehr beinhalten, abhängig von der Ausführungsform, obwohl kleinere Zwischenspeichergrößen verwendet werden können.

[0389] In mindestens einer Ausführungsform können ein oder mehrere des einen oder der mehreren SoCs **3404** einen oder mehrere Beschleuniger **3414** (z. B. Hardwarebeschleuniger, Softwarebeschleuniger oder eine Kombination daraus) beinhalten. In mindestens einer Ausführungsform können das eine oder die mehreren SoCs **3404** einen Hardwarebeschleunigungscluster beinhalten, der optimierte Hardwarebeschleuniger und/oder großen Speicher auf dem Chip beinhaltet. In mindestens einer Ausführungsform kann ein großer Speicher auf dem Chip (z. B. 4 MB SRAM) ermöglichen, dass ein Hardwarebeschleunigungscluster neuronale Netzwerke und andere Berechnungen beschleunigt. In mindestens einer Ausführungsform kann der Hardwarebeschleunigungscluster verwendet werden, um die eine oder die mehreren GPUs **3408** zu ergänzen und um einige Aufgaben der einen oder der mehreren GPUs **3408** auszulagern (z. B. um mehr Takte der einen oder der mehreren GPUs **3408** zum Durchführen anderer Aufgaben freizumachen). In mindestens einer Ausführungsform können ein oder mehrere Beschleuniger **3414** für gezielte Arbeitslasten (z. B. Wahrnehmung, neuronale Faltungsnetzwerke (convolutional neural networks - „CNNs“), rekurrente neuronale Netzwerke („RNNs“) usw.) verwendet werden, die stabil genug sind, um offen für Beschleunigung zu sein. In mindestens einer Ausführungsform kann ein CNN ein bereichsbasiertes oder regionales neuronales Faltungsnetzwerk („RCNNs“) und Schnelle RCNNs (z. B. wie zur Objekterfassung verwendet) oder eine andere Art von CNN beinhalten.

[0390] In mindestens einer Ausführungsform können der eine oder die mehreren Beschleuniger **3414** (z. B. der Hardwarebeschleunigungscluster) einen oder mehrere Deep-Learning-Beschleuniger (deep learning accelerator(s) - „DLA(s)“) beinhalten. Ein oder mehrere DLAs können ohne Einschränkung einen oder mehrere Tensorverarbeitungseinheiten (one or more Tensor processing units - „TPU(s)“) beinhalten, die dazu konfiguriert sein können, zusätzliche zehn Billionen Vorgänge pro Sekunde für Deep-Learning-Anwendungen und -Ableitung bereitzustellen. In mindestens einer Ausführungsform können die eine oder die mehreren TPUs Beschleuniger sein, die dazu konfiguriert und optimiert sind, Bildverarbeitungsfunktionen (z. B. für CNNs, RCNNs usw.) durchzuführen. Der eine oder die mehreren DLAs können ferner für einen konkreten Satz von Arten von neuronalen Netzwerken und Gleitkommaoperationen sowie für das Ableiten optimiert sein. In mindestens einer Ausführungsform kann die Ausgestaltung des einen oder der mehreren DLAs mehr Leistung pro Millimeter als eine typische Allzweck-GPU bereitstellen und übersteigt typisch deutlich die Leistung einer CPU. In mindestens einer Ausführungsform können die eine oder die mehreren TPUs mehrere Funktionen, die eine Einzelinstanzfaltungsfunktion beinhalten, das Unterstützen von zum Beispiel den Datenarten INT8, INT 16 und FP16 für sowohl Merkmale als auch Gewichtungen sowie Nachbearbeitungsfunktionen durchführen. In mindestens einer Ausführungsform können der eine oder die mehreren DLAs schnell und effizient neuronale Netzwerke, insbesondere CNNs, auf verarbeiteten und unverarbeiteten Daten für eine beliebige einer Reihe von Funktion ausführen, die zum Beispiel und ohne Einschränkung Folgendes beinhalten: ein CNN zur Objektidentifizierung und -erfassung unter Verwendung von Daten von Kamerasensoren; ein CNN zur Entfernungsschätzung unter Verwendung von Daten von Kamerasensoren; ein CNN zur Einsatzfahrzeugerfassung und -identifizierung und -erfassung unter Verwendung von Daten von den Mikrofonen **3496**; ein CNN zur Gesichtserkennung und Fahrzeugbesitzeridentifizierung unter Verwendung von Daten von Kamerasensoren; und/oder ein CNN für die Sicherheit und/oder sicherheitsrelevante Ereignisse.

[0391] In mindestens einer Ausführungsform können der eine oder die mehreren DLAs eine beliebige Funktion der einen oder der mehreren GPUs **3408** durchführen und durch das Verwenden eines Ableitungsbeschleunigers kann ein Ausgestalter zum Beispiel entweder den einen oder die mehreren DLAs oder die eine oder die mehreren GPUs **3408** für eine beliebige Funktion zur Vorgabe machen. In mindestens einer Ausführungsform kann der Ausgestalter auf das Verarbeiten von CNNs und Gleitkommaoperationen auf dem einen oder den mehreren DLAs konzentrieren und andere Funktionen der einen oder den mehreren GPUs **3408** und/oder dem einen oder den mehreren anderen Beschleunigern **3414** überlassen.

[0392] In mindestens einer Ausführungsform können der eine oder die mehreren Beschleuniger **3414** (z. B. der Hardwarebeschleunigungscluster) einen oder mehrere programmierbare Sichtbeschleuniger (programmable vision accelerator - „PVA“) beinhalten, die alternativ in dieser Schrift als ein Beschleuniger maschinellen Sehens bezeichnet sein können. In mindestens einer Ausführungsform können der eine oder die mehreren PVAs ausgestaltet und dazu konfiguriert sein, Algorithmen des maschinellen Sehens für ein erweitertes Fahrerunterstützungssystem (advanced driver assistance system - „ADAS“) **3438**, autonomes Fahren, Anwendungen der erweiterten Realität (augmented reality - „AR“) und/oder Anwendungen der virtuellen Realität („VR“) zu

beschleunigen. Der eine oder die mehreren PVAs können ein Gleichgewicht zwischen Leistung und Flexibilität bereitstellen. Zum Beispiel kann in mindestens einer Ausführungsform jeder PVA zum Beispiel und ohne Einschränkung eine beliebige Anzahl von Kernen von Computern mit verringertem Anweisungssatz (reduced instruction set Computer - „RISC“), direkten Speicherzugriff (direct memory access - „DMA“) und/oder eine beliebige Anzahl von Vektorprozessoren beinhalten.

[0393] In mindestens einer Ausführungsform können RISC-Kerne mit Bildsensoren (z.B. Bildsensoren einer beliebigen von in dieser Schrift beschriebenen Kameras), einem oder mehreren Bildsignalprozessoren und/oder ähnlichem interagieren. In mindestens einer Ausführungsform kann jeder der RISC-Kerne eine beliebige Speicherkapazität beinhalten. In mindestens einer Ausführungsform können RISC-Kerne eine beliebige Anzahl von Protokollen abhängig von der Ausführungsform verwenden. In mindestens einer Ausführungsform können RISC-Kerne ein Echtzeitbetriebssystem (real-time operating system - „RTOS“) ausführen. In mindestens einer Ausführungsform können RISC-Kerne unter Verwendung von einer oder mehreren Vorrichtungen mit integrierter Schaltung, anwendungsspezifischen integrierten Schaltungen (application specific integrated circuits - „ASICs“) und/oder Speichervorrichtungen umgesetzt sein. Zum Beispiel könnten in mindestens einer Ausführungsform die RISC-Kerne einen Anweisungszwischenspeicher und/oder eng gekoppelten RAM beinhalten.

[0394] In mindestens einer Ausführungsform kann DMA Komponenten von einem oder mehreren PVAs ermöglichen, unabhängig von einer oder mehreren CPUs **3406** auf Systemspeicher zuzugreifen. In mindestens einer Ausführungsform kann DMA eine beliebige Anzahl von Merkmalen unterstützen, die verwendet werden, um einem PVA Optimierung bereitzustellen, was das Unterstützen mehrdimensionaler Adressierung und/oder kreisförmiger Adressierung beinhaltet, ohne darauf beschränkt zu sein. In mindestens einer Ausführungsform kann DMA bis zu sechs weitere Dimensionen der Adressierung unterstützen, was Blockbreite, Blockhöhe, Blocktiefe, waagerechtes Blockstepping, senkrechtes Blockstepping und/oder Tiefenstepping ohne Einschränkung beinhalten kann.

[0395] In mindestens einer Ausführungsform können Vektorprozessoren programmierbare Prozessoren sein, die ausgestaltet sein können, Programmierung für Algorithmen des maschinellen Sehens effizient und flexibel auszuführen und Signalverarbeitungsfähigkeiten bereitzustellen. In mindestens einer Ausführungsform kann der PVA einen PVA-Kern und zwei Vektorverarbeitungsteilsystemunterteilungen beinhalten. In mindestens einer Ausführungsform kann der PVA-Kern ein Prozessorteilsystem, eine oder mehrere DMA-Engines (z. B. zwei DMA-Engines) und/oder andere Peripheriegeräte beinhalten. In mindestens einer Ausführungsform kann das Vektorverarbeitungsteilsystem als Hauptverarbeitungsengine des PVA betrieben werden und kann eine Vektorverarbeitungseinheit (vector processing unit - „VPU“), einen Anweisungszwischenspeicher und/oder einen Vektorspeicher (z. B. „VMEM“) beinhalten. In mindestens einer Ausführungsform kann die VPU einen Digitalsignalprozessor beinhalten, wie etwa zum Beispiel eine Digitalsignalprozessor mit einzelner Anweisung und mehreren Daten (single instruction, multiple data - „SIMD“) und sehr langen Anweisungsworten (very long instruction word - „VLIW“). In mindestens einer Ausführungsform kann eine Kombination aus SIMD und VLIW den Durchsatz und die Geschwindigkeit verbessern.

[0396] In mindestens einer Ausführungsform kann jeder der Vektorprozessoren einen Anweisungszwischenspeicher beinhalten und kann an dedizierten Speicher gekoppelt sein. Folglich kann in mindestens einer Ausführungsform jeder der Vektorprozessoren dazu konfiguriert sein, unabhängig von anderen Vektorprozessoren ausgeführt zu werden. In mindestens einer Ausführungsform können Vektorprozessoren, die in einem bestimmten PVA beinhaltet sind, dazu konfiguriert sein, Datenparallelität einzusetzen. Zum Beispiel kann in mindestens einer Ausführungsform eine Vielzahl von Vektorprozessoren, die in einem einzelnen PVA beinhaltet sind, denselben Algorithmus des maschinellen Sehens ausführen, aber auf unterschiedliche Bereiche eines Bildes. In mindestens einer Ausführungsform können Vektorprozessoren, die in einem bestimmten PVA beinhaltet sind, gleichzeitig unterschiedliche Algorithmen des maschinellen Sehens an demselben Bild ausführen oder sogar unterschiedliche Algorithmen an aufeinanderfolgenden Bildern oder Abschnitten eines Bildes ausführen. In mindestens einer Ausführungsform kann unter anderem eine beliebige Anzahl von PVAs in einem Hardwarebeschleunigungscluster beinhaltet sein und eine beliebige Anzahl von Vektorprozessoren kann in jedem der PVAs beinhaltet sein. In mindestens einer Ausführungsform können der eine oder die mehreren PVAs zusätzlichen Speicher mit Fehlerkorrekturcode (error correcting code - „ECC“) beinhalten, um die Sicherheit des gesamten Systems zu verbessern.

[0397] In mindestens einer Ausführungsform können der eine oder die mehreren Beschleuniger **3414** (z. B. der Hardwarebeschleunigungscluster) ein Netzwerk auf einem Chip des maschinellen Sehens und statischen Direktzugriffsspeicher (static random-access memory - „SRAM“) zum Bereitstellen von SRAM mit hoher Bandbreite und geringer Latenz für den einen oder die mehreren Beschleuniger **3414** beinhalten. In mindestens einer

Ausführungsform kann der Speicher auf dem Chip mindestens SRAM mit 4 MB beinhalten, der zum Beispiel und ohne Einschränkung aus acht feldkonfigurierbaren Speicherblöcken besteht, auf die sowohl durch den PVA als auch den DLA zugegriffen werden kann. In mindestens einer Ausführungsform kann jedes Paar von Speicherblöcken eine Schnittstelle mit verbessertem Peripheriegerätebus (advanced peripheral bus - „APB"), Konfigurationsschaltung, eine Steuerung und einen Mehrfachkoppler beinhalten. In mindestens einer Ausführungsform kann eine beliebige Art von Speicher verwendet sein. In mindestens einer Ausführungsform können der PVA und der DLA über eine Backbone, die dem PVA und dem DLA Schnellzugriff auf Speicher bereitstellt, auf den Speicher zugreifen. In mindestens einer Ausführungsform kann die Backbone ein Netzwerk auf einem Chip des maschinellen Sehens beinhalten, das den PVA und den DLA mit dem Speicher (z. B. unter Verwendung von einem APB) miteinander verbindet.

[0398] In mindestens einer Ausführungsform kann ein Netzwerk auf einem Chip des maschinellen Sehens eine Schnittstelle beinhalten, die vor der Übermittlung beliebiger Steuersignale/Adressen/Daten bestimmt, ob sowohl der PVA als auch der DLA bereite und gültige Signale bereitstellen. In mindestens einer Ausführungsform kann eine Schnittstelle getrennte Phasen und getrennte Kanäle zum Übermitteln von Steuersignalen/Adressen/Daten sowie stoßartiger Kommunikationen für durchgehende Datenübertragung bereitstellen. In mindestens einer Ausführungsform kann eine Schnittstelle den Normen der Internationalen Organisation für Normung (International Organization for Standardization - „ISO") 26262 oder der Internationalen Elektrotechnischen Kommission (International Electrotechnical Commission - „IEC") 61508 entsprechen, obwohl andere Normen und Protokolle verwendet werden können.

[0399] In mindestens einer Ausführungsform können einer oder mehrere des einen oder der mehreren SOCs **3404** einen Echtzeit-Raytracing-Hardwarebeschleuniger beinhalten. In mindestens einer Ausführungsform kann ein Echtzeit-Raytracing-Hardwarebeschleuniger verwendet werden, um schnell und effizient Positionen und Ausmaße von Objekten (z.B. innerhalb eines Weltmodells) zu bestimmen, um Echtzeit-Visualisierungssimulationen zur RADAR-Signalauslegung, zur Schallausbreitungssynthese und/oder -analyse, zur Simulation von SONAR-Systemen, zur allgemeinen Wellenausbreitungssimulation, zum Vergleich mit LIDAR-Daten zum Zweck der Ortung und/oder anderen Funktionen zu erzeugen, und/oder für andere Verwendungen.

[0400] In mindestens einer Ausführungsform weisen der eine oder die mehreren Beschleuniger **3414** (z. B. der Hardwarebeschleunigercluster) ein breites Spektrum von Verwendungen für autonomes Fahren auf. In mindestens einer Ausführungsform kann der PVA ein programmierbarer Sichtbeschleuniger sein, der für Schlüsselverarbeitungsstufen in ADAS und autonomen Fahrzeugen verwendet werden kann. In mindestens einer Ausführungsform passen die Fähigkeiten des PVA gut damit zusammen, dass algorithmische Domänen verlässliches Verarbeiten bei geringer Leistung und mit geringer Latenz benötigen. Mit anderen Worten funktioniert der PVA gut bei semidichten oder dichten regulären Berechnungen, selbst bei kleinen Datensätzen, die vorhersagbare Laufzeiten mit geringer Latenz und geringer Leistung benötigen. In mindestens einer Ausführungsform sind PVAs von autonomen Fahrzeugen, wie etwa dem Fahrzeug **3400**, ausgestaltet, um klassische Algorithmen des maschinellen Sehens auszuführen, da sie bei der Objekterfassung und beim Betreiben auf Ganzzahlmathematik effizient sind.

[0401] Zum Beispiel wird gemäß mindestens einer Ausführungsform der Technik der PVA verwendet, um Computerstereosehen durchzuführen. In mindestens einer Ausführungsform kann ein semiglobaler Algorithmus auf Übereinstimmungsbasis in einigen Beispielen verwendet werden, obwohl dies nicht einschränkend gedacht ist. In mindestens einer Ausführungsform verwenden Anwendungen für autonomes Fahren der Stufen 3-5 Bewegungsschätzung/Stereoübereinstimmung während der Verarbeitung (z. B. Struktur-von-Bewegung, Fußgängererkennung, Spurerfassung usw.). In mindestens einer Ausführungsform kann der PVA eine Computerstereosehenfunktion auf Eingaben von zwei monokularen Kameras durchführen.

[0402] In mindestens einer Ausführungsform kann der PVA verwendet werden, um dichten optischen Fluss durchzuführen. Zum Beispiel könnte der PVA in mindestens einer Ausführungsform RADAR-Rohdaten verarbeiten (z. B. unter Verwendung einer schnellen 4D-Fouriertransformation), um verarbeitete RADAR-Daten bereitzustellen. In mindestens einer Ausführungsform wird der PVA für Laufzeittiefenverarbeitung verwendet, indem zum Beispiel Laufzeitrohdaten verarbeitet werden, um verarbeitete Laufzeitdaten bereitzustellen.

[0403] In mindestens einer Ausführungsform kann der DLA verwendet werden, um eine beliebige Art von Netzwerk auszuführen, um die Steuerung und die Fahrsicherheit zu erhöhen, was zum Beispiel und ohne Einschränkung ein neuronales Netzwerk beinhaltet, das ein Vertrauensmaß für jede Objekterfassung ausgibt. In mindestens einer Ausführungsform kann das Vertrauen als eine Wahrscheinlichkeit dargestellt oder ausgelegt sein, oder als das Bereitstellen einer relativen „Gewichtung" jeder Erfassung im Vergleich zu anderen Erfas-

sungen. In mindestens einer Ausführungsform ermöglicht Vertrauen einem System, weitere Entscheidungen in Bezug darauf zu treffen, welche Erfassungen als echt positive Erfassungen statt als falsch positive Erfassungen betrachtet werden sollten. Zum Beispiel kann in mindestens einer Ausführungsform ein System einen Schwellenwert für Vertrauen einstellen und lediglich Erfassungen berücksichtigen, die den Schwellenwert als echt positive Erfassungen übersteigen. In einer Ausführungsform, in der ein System zur automatischen Notbremsung (automatic emergency braking - „AEB“) verwendet wird, würden falsch positive Erfassungen das Fahrzeug veranlassen, automatisch eine Notbremsung durchzuführen, was offensichtlich nicht wünschenswert ist. In mindestens einer Ausführungsform können äußerst zuversichtliche Erfassungen als Auslöser für AEB berücksichtigt werden. In mindestens einer Ausführungsform kann der DLA ein neuronales Netzwerk für einen rückläufigen Vertrauenswert ausführen. In mindestens einer Ausführungsform kann ein neuronales Netzwerk als seine Eingabe mindestens irgendeine Teilmenge von Parametern verwenden, wie etwa unter anderem Begrenzungsfeldabmessungen, erlangte (z. B. von einem anderen Teilsystem) Grundebenenschätzung, Ausgabe von einem oder mehreren IMU-Sensoren **3466**, die mit der Ausrichtung des Fahrzeugs **3400** zusammenhängen, Entfernung, 3D-Standortschätzungen des Objekts, die von neuronalen Netzwerk und/oder anderen Sensoren (z. B. LIDAR-Sensoren **3464** oder einem oder mehreren RADAR-Sensoren **3460**) erlangt wurden.

[0404] In mindestens einer Ausführungsform können ein oder mehrere des einen oder der mehreren SOCs **3404** einen oder mehrere Datenspeicher **3416** (z. B. Speicher) beinhalten. In mindestens einer Ausführungsform können der eine oder die mehreren Datenspeicher **3416** Speicher auf dem Chip von dem einen oder den mehreren SOCs **3404** sein, die neuronale Netzwerke speichern können, die auf der einen oder den mehreren GPUs **3408** und/oder dem DLA ausgeführt werden sollen. In mindestens einer Ausführungsform können der eine oder die mehreren Datenspeicher **3416** groß genug sein, was die Dichte betrifft, um mehrere Instanzen von neuronalen Netzwerken aus Gründen der Redundanz und Sicherheit zu speichern. In mindestens einer Ausführungsform können der eine oder die mehreren Datenspeicher **3416** einen oder mehrere L2- oder L3-Zwischenspeicher umfassen.

[0405] In mindestens einer Ausführungsform können ein oder mehrere des einen oder der mehreren SOCs **3404** eine beliebige Anzahl von einem oder mehreren Prozessoren **3410** (z. B. eingebettete Prozessoren) beinhalten. In mindestens einer Ausführungsform können der eine oder die mehreren Prozessoren **3410** einen Hochfahr- und Leistungsverwaltungsprozessor beinhalten, der ein dedizierter Prozessor und ein Teilsystem sein können, um die Hochfahrleistungs- und Verwaltungsfunktionen und die dazugehörige Durchsetzung von Sicherheitsrichtlinien zu handhaben. In mindestens einer Ausführungsform kann der Hochfahr- und Leistungsverwaltungsprozessor ein Teil der Hochfahrsequenz der einen oder der mehreren SOCs **3404** sein und kann Laufzeitleistungsverwaltungsdienste bereitstellen. In mindestens einer Ausführungsform kann der Hochfahrleistungs- und Verwaltungsprozessor eine Takt- und Spannungsprogrammierung, die Unterstützung in den Übergängen der Zustände mit niedriger Leistung des Systems, die Verwaltung von Wärmeabgaben des einen oder der mehreren SOCs **3404** und Temperatursensoren und/oder die Verwaltung der Leistungszustände des einen oder der mehreren SOCs **3404**. In mindestens einer Ausführungsform kann jeder Temperatursensor als ein Ringoszillator umgesetzt sein, dessen Ausgabefrequenz proportional zur Temperatur ist, und das eine oder die mehreren SoCs **3404** können Ringoszillatoren verwenden, um Temperaturen der einen oder der mehreren CPUs **3406**, GPUs **3408** und/oder des einen oder der mehreren Beschleuniger **3414** zu erfassen. In mindestens einer Ausführungsform kann, wenn bestimmt wurde, dass die Temperaturen einen Schwellenwert übersteigen, dann der Hochfahr- und Leistungsverwaltungsprozessor in ein Temperaturfehlerprogramm übergehen und das eine oder die mehreren SoCs **3404** in einen Zustand mit geringerer Leistung versetzen und/oder das Fahrzeug **3400** in einen Modus des Fahrens bis zu einem sicheren Stillstand (z. B. das Fahrzeug **3400** zu einem sicheren Stillstand bringen) versetzen.

[0406] In mindestens einer Ausführungsform können der eine oder die mehreren Prozessoren **3410** ferner einen Satz von eingebetteten Prozessoren beinhalten, die als eine Audioverarbeitungsengine dienen können. In mindestens einer Ausführungsform kann die Audioverarbeitungsengine ein Audioteilsystem sein, das eine vollständige Hardwareunterstützung für Mehrkanalaudio über mehrere Schnittstellen und einen breiten und flexiblen Bereich von Audio-E/A-Schnittstellen ermöglicht. In mindestens einer Ausführungsform ist die Audioverarbeitungsengine ein dedizierter Prozessorkern mit einem Digitalsignalprozessor mit dediziertem RAM.

[0407] In mindestens einer Ausführungsform können der eine oder die mehreren Prozessoren **3410** ferner eine Engine für dauerhaft eingeschaltete Prozessoren beinhalten, die erforderliche Hardwaremerkmale bereitstellt, um die Verwaltung von Sensoren mit niedriger Leistung und Weckanwendungsfälle zu unterstützen. In mindestens einer Ausführungsform kann die Engine für dauerhaft eingeschaltete Prozessoren ohne Einschränkung Folgendes beinhalten: einen Prozessorkern, einen eng gekoppelten RAM, unterstützende Peri-

pheriegeräte (z. B. Timer und Unterbrechungsteuerungen), verschiedene E/A-Steuerungsperipheriegeräte und Streckenführungslogik.

[0408] In mindestens einer Ausführungsform können der eine oder die mehreren Prozessoren **3410** ferner eine Sicherheitsclusterengine beinhalten, die ohne Einschränkung ein dediziertes Prozessorteilsystem beinhaltet, um die Sicherheitsverwaltung für Automobilanwendungen zu handhaben. In mindestens einer Ausführungsform kann die Sicherheitsclusterengine ohne Einschränkung Folgendes beinhalten: zwei oder mehr Prozessorkerne, einen eng gekoppelten RAM, Unterstützungsperipheriegeräte (z. B. Timer, eine Unterbrechungsteuerung usw.) und/oder Streckenführungslogik. In einem Sicherheitsmodus können zwei oder mehr Kerne in mindestens einer Ausführungsform in einem Gleichschrittmodus betrieben werden und funktionieren als ein einzelner Kern mit einer Vergleichslogik, um jegliche Unterschiede zwischen ihren Operationen zu erfassen. In mindestens einer Ausführungsform können der eine oder die mehreren Prozessoren **3410** ferner eine Echtzeitkameraengine beinhalten, die ohne Einschränkung ein dediziertes Prozessorteilsystem zum Handhaben der Echtzeitkameraverwaltung beinhalten kann. In mindestens einer Ausführungsform können der eine oder die mehreren Prozessoren **3410** ferner einen Prozessor für Signale mit hochdynamischem Bereich beinhalten, der ohne Einschränkung einen Bildsignalprozessor beinhalten kann, der eine Hardwareengine ist, die Teil einer Kameraverarbeitungspipeline ist.

[0409] In mindestens einer Ausführungsform können der eine oder die mehreren Prozessoren **3410** einen Videobildsetzer beinhalten, der ein Verarbeitungsblock (z. B. in einem Mikroprozessor umgesetzt) sein kann, der Videonachbearbeitungsfunktionen umsetzt, die durch eine Videowiedergabeanwendung benötigt werden, um das Fenster des finalen Bildes für die Abspieleinrichtung erzeugt. In mindestens einer Ausführungsform kann der Videobildsetzer Linsenverzerrungskorrektur an der einen oder den mehreren Weitsichtkameras **3470**, der einen oder den mehreren Rundumkameras **3474** und/oder an einer oder mehreren Überwachungskamerasensoren in der Kabine durchführen. In mindestens einer Ausführungsform werden der eine oder die mehreren Überwachungskamerasensoren in der Kabine vorzugsweise durch ein neuronales Netzwerk überwacht, das auf einer anderen Instanz des einen oder der mehreren SOCs **3404** ausgeführt wird, die dazu konfiguriert sind, Ereignisse in der Kabine zu identifizieren und dementsprechend zu reagieren. In mindestens einer Ausführungsform kann ein System in der Kabine ohne Einschränkung Folgendes durchführen: Lippenlesen, um den Mobilfunkdienst einzuschalten und einen Telefonanruf zu tätigen, das Diktieren von E-Mails, das Ändern des Ziels des Fahrzeugs, das Einschalten oder Ändern des Infotainmentsystems und dessen Einstellungen des Fahrzeugs oder das Bereitstellen von sprachgesteuertem Internetsurfen. In mindestens einer Ausführungsform sind bestimmte Funktionen für den Fahrer verfügbar, wenn das Fahrzeug in einem autonomen Modus betrieben wird, und anderweitig untauglich ist.

[0410] In mindestens einer Ausführungsform kann ein Videobildsetzer verbesserte zeitliche Rauschunterdrückung für sowohl räumliche als auch zeitliche Rauschunterdrückung beinhalten. Zum Beispiel, wägt in mindestens einer Ausführungsform, bei der Bewegung in einem Video auftritt, Rauschunterdrückung räumliche Informationen angemessen ab, was die Gewichtung von Informationen verringert, die durch angrenzende Einzelbilder bereitgestellt sind. In mindestens einer Ausführungsform, bei der ein Bild oder ein Abschnitt eines Bildes keine Bewegung beinhaltet, kann zeitliche Rauschunterdrückung, die durch den Videobildsetzer durchgeführt wird, Informationen aus dem vorherigen Bild verwenden, um das Rauschen in dem aktuellen Bild zu verringern.

[0411] In mindestens einer Ausführungsform kann der Videobildsetzer außerdem dazu konfiguriert sein, Stereokorrektur an Eingabestereolinseneinzelbildern durchzuführen. In mindestens einer Ausführungsform kann der Videobildsetzer ferner für die Benutzerschnittstellenzusammensetzung verwendet werden, wenn der Betriebssystem-Desktop verwendet wird und die eine oder die mehreren GPUs **3408** nicht benötigt werden, um durchgehend neue Oberflächen zu rendem. In mindestens einer Ausführungsform kann, wenn die eine oder die mehreren GPUs **3408** eingeschaltet und aktiv sind und 3D-Rendering durchführen, der Videobildsetzer verwendet werden, um die eine oder die mehreren GPUs **3408** zu entlasten, um die Leistung und Reaktionsgeschwindigkeit zu verbessern.

[0412] In mindestens einer Ausführungsform können eines oder mehrere des einen oder der mehreren SOCs **3404** ferner Folgendes beinhalten: eine serielle Schnittstelle einer Kamera einer Schnittstelle eines Prozessors der Mobilfunkbranche (mobile industry processor interface - „MIPI“) zum Empfangen von Video und zur Eingabe von Kameras, eine Hochgeschwindigkeitsschnittstelle und/oder einen Videoeingabeblock, der für Kameraeingabefunktionen und die Eingabefunktionen zusammengehöriger Bildpunkte verwendet werden kann. In mindestens einer Ausführungsform können eines oder mehrere des einen oder der mehreren SOCs **3404** ferner einen oder mehrere Eingabe-/Ausgabe-Steuerungen beinhalten, die durch Software gesteuert sein kön-

nen und für das Empfangen von E/A-Signalen verwendet werden können, die nicht an eine konkrete Rolle gebunden sind.

[0413] In mindestens einer Ausführungsform können eines oder mehrere des einen oder der mehreren SOCs **3404** ferner ein breites Spektrum von Peripheriegeräteschnittstellen beinhalten, um die Kommunikation mit Peripheriegeräten, Audiokodierern/-dekodierern („Codecs“, der Leistungsverwaltung und/oder anderen Vorrichtungen zu ermöglichen. Das eine oder die mehreren SoCs **3404** können verwendet werden, um Daten von Folgenden zu verarbeiten: Kameras (z. B. über eine serielle Gigabit-Multimedia-Verbindung oder Ethernet), Sensoren (z. B. der eine oder die mehreren LIDAR-Sensoren **3464**, der eine oder die mehreren RADAR-Sensoren **3460** usw., die über Ethernet verbunden sein können), Daten von dem Bus **3402** (z. B. die Geschwindigkeit des Fahrzeugs **3400**, die Lenkradstellung usw.), Daten von dem einen oder den mehreren GNSS-Sensoren **3458** (z. B. über Ethernet oder den CAN-Bus verbunden) usw. In mindestens einer Ausführungsform können eines oder mehrere des einen oder der mehreren SOCs **3404** ferner dedizierte Hochleistungsmassenspeichersteuerungen beinhalten, die ihre eigenen DMA-Engines beinhalten können und die verwendet werden können, um die eine oder die mehreren CPUs **3406** aus Programmdatenverwaltungsaufgaben zu befreien.

[0414] In mindestens einer Ausführungsform können eines oder mehrere des einen oder der mehreren SOCs **3404** eine Ende-zu-Ende-Plattform mit einer flexiblen Architektur sein, die sich über die Automationsstufen 3-5 spannt, wodurch eine umfassende funktionelle Sicherheitsarchitektur bereitgestellt wird, die maschinelles Sehen und ADAS-Methoden für Vielfalt und Redundanz nutzt und effizient daraus Nutzen zieht, eine Plattform für ein flexibles zuverlässiges Fahrsoftwarepaket, zusammen mit Deep-Learning-Werkzeugen, bereitstellt. In mindestens einer Ausführungsform können eines oder mehrere des einen oder der mehreren SOCs **3404** schneller, zuverlässiger und sogar energieeffizienter und platzsparender als herkömmliche Systeme sein. Zum Beispiel können in mindestens einer Ausführungsform der eine oder die mehreren Beschleuniger **3414**, wenn sie mit der einen oder den mehreren CPUs **3406**, der einen oder den mehreren GPUs **3408** und dem einen oder den mehreren Datenspeichern **3416** kombiniert werden, eine schnelle, effiziente Plattform für autonome Fahrzeuge der Stufe 3-5 bereitstellen.

[0415] In mindestens einer Ausführungsform können Algorithmen des maschinellen Sehens an CPUs ausgeführt werden, die dazu konfiguriert sein können, unter Verwendung von höherer Programmiersprache, wie etwa der Programmiersprache C, ein breites Spektrum von Verarbeitungsalgorithmen über ein breites Spektrum von optischen Daten auszuführen. In mindestens einer Ausführungsform sind die CPUs jedoch oftmals nicht in der Lage, Leistungsvoraussetzungen von vielen Anwendungen des maschinellen Sehens zu erfüllen, wie etwa zum Beispiel die in Bezug auf die Ausführungszeit und Leistungsaufnahme. In mindestens einer Ausführungsform sind viele CPUs nicht in der Lage, komplexe Objekterfassungsalgorithmen in Echtzeit auszuführen, die bei fahrzeuginternen ADAS-Anwendungen und bei zweckmäßigen autonomen Fahrzeugen der Stufen 3-5 verwendet werden.

[0416] In dieser Schrift beschriebene Ausführungsformen ermöglichen, dass mehrere neuronale Netzwerke gleichzeitig und/oder aufeinanderfolgend durchgeführt werden, und dass Ergebnisse miteinander kombiniert werden, um die Funktion des autonomen Fahrens der Stufen 3-5 zu ermöglichen. Zum Beispiel kann in mindestens einer Ausführungsform ein CNN, das auf einem DLA oder einer separaten GPU (z. B. die eine oder die mehreren GPUs **3420**) ausgeführt wird, Text- und Worterkennung beinhalten, was einem Supercomputer ermöglicht, Verkehrsschilder zu lesen und zu verstehen, was Schilder beinhaltet, für die das neuronale Netzwerk nicht konkret trainiert wurde. In mindestens einer Ausführungsform kann der DLA ferner ein neuronales Netzwerk beinhalten, das in der Lage ist, Schilder zu identifizieren, auszulegen und semantisches Verständnis bereitzustellen und dieses semantische Verständnis an Wegplanungsmodule weiterzugeben, die auf einem CPU-Komplex ausgeführt werden.

[0417] In mindestens einer Ausführungsform können mehrere neuronale Netzwerke gleichzeitig ausgeführt werden, wie für das Fahren der Stufen 3, 4 oder 5. Zum Beispiel kann in mindestens einer Ausführungsform ein Warnschild, das aus „Vorsicht! Blinkende Leuchten weisen auf Eisglätte hin“ zusammen mit einer elektrischen Leuchte besteht, unabhängig oder gemeinschaftlich durch mehrere neuronale Netzwerke ausgelegt werden. In mindestens einer Ausführungsform kann ein Schild selbst als ein Verkehrsschild durch ein erstes eingesetztes neuronales Netzwerk (z. B. ein neuronales Netzwerk, das trainiert wurde) identifiziert werden und ein Text „blinkende Leuchten weisen auf Eisglätte hin“ kann durch ein zweites eingesetztes neuronales Netzwerk ausgelegt werden, der Wegplanungssoftware des Fahrzeugs (die vorzugsweise auf einem CPU-Komplex ausgeführt wird), dass, wenn blinkende Leuchten erfasst werden, Eisglätte besteht. In mindestens einer Ausführungsform kann eine blinkende Leuchte durch das Betreiben eines dritten eingesetzten neuronalen Netzwerks über mehrere Einzelbilder identifiziert werden, welche die Wegplanungssoftware des Fahrzeugs zu

einer Anwesenheit (oder Abwesenheit) blinkender Leuchten informiert. In mindestens einer Ausführungsform können alle drei neuronalen Netzwerke gleichzeitig ausgeführt werden, wie etwa innerhalb des DLA und/oder auf der einen oder den mehreren GPUs **3408**.

[0418] In mindestens einer Ausführungsform kann ein CNN für Gesichtserkennung und Fahrzeugbesitzeridentifizierung Daten von Kamerasensoren verwenden, um die Anwesenheit eines autorisierten Fahrers und/oder Besitzers des Fahrzeugs **3400** zu identifizieren. In mindestens einer Ausführungsform kann eine Verarbeitungsengine für dauerhaft eingeschaltete Sensoren verwendet werden, um das Fahrzeug zu entriegeln, wenn sich der Besitzer der Fahrertür nähert, und die Leuchten anzuschalten, und im Sicherheitsmodus, um das Fahrzeug abzuschalten, wenn der Besitzer das Fahrzeug verlässt. Auf diese Art können das eine oder die mehreren SoCs **3404** Sicherheit gegenüber Diebstahl und/oder Autoraub bereitstellen.

[0419] In mindestens einer Ausführungsform kann das CNN zur Einsatzfahrzeugerfassung und -identifizierung Daten von Mikrofonen **3496** verwenden, um Einsatzfahrzeugsirenen zu erfassen und zu identifizieren. In mindestens einer Ausführungsform können das eine oder die mehreren SoCs **3404** das CNN zum Klassifizieren von Umgebungs- und städtischen Geräuschen sowie zum Klassifizieren von optischen Daten verwenden. In mindestens einer Ausführungsform ist das CNN, das auf einem DLA ausgeführt wird, trainiert, um eine verhältnismäßige Aufschließgeschwindigkeit des Einsatzfahrzeugs (z. B. durch das Verwenden des Doppler-Effekts) zu identifizieren. In mindestens einer Ausführungsform kann das CNN außerdem dazu trainiert sein, Einsatzfahrzeuge zu identifizieren, die charakteristisch für eine Gegend sind, in der das Fahrzeug betrieben wird, wie durch den einen oder die mehreren GNSS-Sensoren **3458** identifiziert. In mindestens einer Ausführungsform wird, wenn es in Europa betrieben wird, das CNN versuchen, europäische Sirenen zu erfassen, und wenn es in den Vereinigten Staaten von Amerika ist, wird das CNN versuchen, lediglich nordamerikanische Sirenen zu identifizieren. In mindestens einer Ausführungsform kann, sobald ein Einsatzfahrzeug erfasst wurde, ein Steuerprogramm verwendet werden, um ein Einsatzfahrzeugsicherheitsprogramm auszuführen, welches das Fahrzeug verlangsamt, an die Seite der Straße fährt, das Fahrzeug parkt und/oder das Fahrzeug in den Leerlauf versetzt, mit Unterstützung von einem oder mehreren Ultraschallsensoren **3462**, bis das eine oder die mehreren Einsatzfahrzeuge vorbeifahren.

[0420] In mindestens einer Ausführungsform kann das Fahrzeug **3400** die eine oder die mehreren CPUs **3418** (z. B. die eine oder die mehreren separaten CPUs oder die eine oder die mehreren dCPUs) beinhalten, die über eine Hochgeschwindigkeitsverbindung (z. B. PCIe) an das eine oder die mehreren SoCs **3404** gekoppelt ist. In mindestens einer Ausführungsform können die eine oder die mehreren CPUs **3418** zum Beispiel einen X86-Prozessor beinhalten. Die eine oder die mehreren CPUs **3418** können verwendet werden, um zum Beispiel eine beliebige einer Reihe von Funktionen durchzuführen, die möglicherweise uneinheitliche Ergebnisse zwischen ADAS-Sensoren und dem einen oder den mehreren SOCs **3404** vermitteln und/oder den Zustand des einen oder der mehreren Steuerungen **3436** und/oder eines Infotainmentsystems auf einem Chip („Infotainment-SoC) **3430** überwachen.

[0421] In mindestens einer Ausführungsform kann das Fahrzeug **3400** die eine oder die mehreren GPUs **3420** (z. B. die eine oder die mehreren separaten GPUs oder die eine oder die mehreren dGPUs) beinhalten, die über eine Hochgeschwindigkeitsverbindung (z. B. NVLINK von NVIDIA) an das eine oder die mehreren SoCs **3404** gekoppelt ist. In mindestens einer Ausführungsform können die eine oder die mehreren GPUs **3420** eine zusätzliche Funktion der künstlichen Intelligenz bereitstellen, wie etwa durch das Ausführen redundanter und/oder unterschiedlicher neuronaler Netzwerke, und kann verwendet werden, um neuronale Netzwerke mindestens zum Teil auf Grundlage von Eingaben (z. B. Sensordaten) von Sensoren des Fahrzeugs **3400** zu trainieren und/oder zu aktualisieren.

[0422] In mindestens einer Ausführungsform kann das Fahrzeug **3400** ferner die Netzwerkschnittstelle **3424** beinhalten, die ohne Einschränkung eine oder mehrere drahtlose Antennen **3426** (z. B. eine oder mehrere drahtlose Antennen **3426** für unterschiedliche Kommunikationsprotokolle, wie etwa eine Mobilfunkantenne, eine Bluetooth-Antenne usw.) beinhalten. In mindestens einer Ausführungsform kann die Netzwerkschnittstelle **3424** verwendet werden, um eine drahtlose Anbindung über das Internet an die Cloud (z. B. mit einem oder mehreren Servern und/oder anderen Netzwerkvorrichtungen), mit anderen Fahrzeugen und/oder mit Rechenvorrichtungen (z. B. Client-Vorrichtungen von Fahrgästen) zu ermöglichen. In mindestens einer Ausführungsform kann, um mit anderen Fahrzeugen zu kommunizieren, eine direkte Verbindung zwischen dem Fahrzeug **3400** und einem anderen Fahrzeug geschaffen werden und/oder es kann eine indirekte Verbindung geschaffen werden (z. B. über Netzwerke und über das Internet). In mindestens einer Ausführungsform können die direkten Verbindungen unter Verwendung einer Fahrzeug-zu-Fahrzeug-Kommunikationsverbindung bereitgestellt werden. Die Fahrzeug-zu-Fahrzeug-Kommunikationsverbindung kann dem Fahrzeug **3400** Informationen

über Fahrzeuge in der Nähe zu dem Fahrzeug **3400** bereitstellen (z. B. Fahrzeuge vor, seitlich von und/oder hinter dem Fahrzeug **3400**). In mindestens einer Ausführungsform kann die vorgenannte Funktion Teil einer kooperativen Abstandsregeltempomatfunktion des Fahrzeugs **3400** sein.

[0423] In mindestens einer Ausführungsform kann die Netzwerkschnittstelle **3424** ein SoC beinhalten, das die Funktion der Modulation und Demodulation bereitstellt und der einen oder den mehreren Steuerungen **3436** ermöglicht, über drahtlose Netzwerke zu kommunizieren. In mindestens einer Ausführungsform kann die Netzwerkschnittstelle **3424** ein Funkfrequenz-Front-End für Aufwärtsumwandlung vom Basisband zur Funkfrequenz und Abwärtsumwandlung von der Funkfrequenz zum Basisband beinhalten. In mindestens einer Ausführungsform können die Frequenzumwandlungen auf eine beliebige technisch machbare Weise durchgeführt werden. Zum Beispiel könnten die Frequenzumwandlungen über wohlbekannte Prozesse und/oder unter Verwendung von Überlagerungsprozessen durchgeführt werden. In mindestens einer Ausführungsform kann die Funkfrequenz-Front-End-Funktion durch einen getrennten Chip bereitgestellt sein. In mindestens einer Ausführungsform kann die Netzwerkschnittstelle drahtlose Funktionen für das Kommunizieren über LTE, WCDMA, UMTS, GSM, CDMA2000, Bluetooth, Bluetooth LE, Wi-Fi, Z-Wave, ZigBee, LoRaWAN und/oder andere drahtlose Protokolle beinhalten.

[0424] In mindestens einer Ausführungsform kann das Fahrzeug **3400** ferner den einen oder die mehrere Datenspeicher **3428** beinhalten, die ohne Einschränkung einen chipexternen (z.B. außerhalb des einen oder der mehreren SoCs **3404**) Speicher beinhalten. In mindestens einer Ausführungsform können der eine oder die mehreren Datenspeicher **3428** ohne Einschränkung ein oder mehrere Speicherelemente beinhalten, die RAM, SRAM, dynamischen Direktzugriffsspeicher („DRAM“), Videodirektzugriffsspeicher („VRAM“), Flash, Festplatten und/oder andere Komponenten und/oder Vorrichtungen beinhalten, die mindestens ein Datenbit speichern können.

[0425] In mindestens einer Ausführungsform kann das Fahrzeug **3400** ferner den einen oder die mehreren GNSS-Sensoren **3458** (z. B. GPS und/oder unterstützte GPS-Sensoren) beinhalten, um bei der Zuordnung, Wahrnehmung, Belegungsrastererzeugung und/oder Wegplanungsfunktionen zu unterstützen. In mindestens einer Ausführungsform kann eine beliebige Anzahl von dem einen oder den mehreren GNSS-Sensoren **3458** verwendet werden, was zum Beispiel und ohne Einschränkung ein GPS unter Verwendung eines USB-Anschlusses mit einer Ethernet-zu-Seriell-(z. B. RS-232-)Überbrückung beinhaltet.

[0426] In mindestens einer Ausführungsform kann das Fahrzeug **3400** ferner den einen oder die mehreren RADAR-Sensoren **3460** beinhalten. Der eine oder die mehreren RADAR-Sensoren **3460** können durch das Fahrzeug **3400** für Langstreckenfahrzeugerfassung verwendet werden, selbst in Dunkelheit und/oder extremen Wetterlagen. In mindestens einer Ausführungsform können RADAR-Funktionssicherheitsstufen ASIL B sein. Der eine oder die mehreren RADAR-Sensoren **3460** können den CAN und/oder Bus **3402** (z. B. um Daten zu übermitteln, die durch den einen oder die mehreren RADAR-Sensoren **3460** erzeugt wurden) zur Steuerung verwenden und um auf Objektverfolgungsdaten zuzugreifen, mit Zugriff zum Ethernet, um in einigen Beispielen auf Rohdaten zuzugreifen. In mindestens einer Ausführungsform kann eine breite Vielfalt von RADAR-Sensorarten verwendet sein. Zum Beispiel und ohne Einschränkung können der eine oder die mehreren RADAR-Sensoren **3460** für vordere, hintere und seitliche RADAR-Verwendung geeignet sein. In mindestens einer Ausführungsform sind einer oder mehrere des einen oder der mehreren RADAR-Sensoren **3460** ein oder mehrere Pulsdoppler-RADAR-Sensoren.

[0427] In mindestens einer Ausführungsform können der eine oder die mehreren RADAR-Sensoren **3460** unterschiedliche Konfigurationen beinhalten, wie etwa Langstrecke mit engem Sichtfeld, Kurzstrecke mit breitem Sichtfeld, Kurzstrecke mit seitlicher Abdeckung usw. In mindestens einer Ausführungsform kann Langstrecken-RADAR für die Funktion des adaptiven Abstandsregeltempomats verwendet sein. In mindestens einer Ausführungsform können Langstrecken-RADAR-Systeme ein breites Sichtfeld bereitstellen, das durch zwei oder mehr unabhängige Abtastungen umgesetzt ist, wie etwa innerhalb eines Bereichs von 250 m. In mindestens einer Ausführungsform können der eine oder die mehreren RADAR-Sensoren **3460** dabei helfen, zwischen feststehenden und sich bewegenden Objekten zu unterscheiden, und können durch ADAS-Systeme **3438** zur Notbremsunterstützung und nach vorn gerichteten Zusammenstoßwarnung verwendet werden. Der eine oder die mehreren Sensoren **3460**, die in einem Langstrecken-RADAR-System beinhaltet sind, können ohne Einschränkung monostatisches multimodales RADAR mit mehreren (z. B. sechs oder mehr) feststehenden RADAR-Antennen und einem Hochgeschwindigkeits-CAN-und-FlexRay-Schnittstelle beinhalten. In mindestens einer Ausführungsform mit sechs Antennen können die vier mittigen Antennen ein konzentriertes Strahlmuster erzeugen, das ausgestaltet ist, um die Umgebung des Fahrzeugs **3400** bei höheren Geschwindigkeiten mit minimaler Störung durch Verkehr in angrenzenden Spuren aufzuzeichnen. In mindestens einer Ausführungsform

können die anderen beiden Antennen das Sichtfeld erweitern, wodurch es möglich wird, Fahrzeuge schnell zu erfassen, welche in die Spur des Fahrzeugs **3400** einfahren oder diese verlassen.

[0428] In mindestens einer Ausführungsform können Mittelstrecken-RADAR-Systeme als ein Beispiel eine Reichweite bis zu 160 m (vorn) oder 80 m (hinten) und ein Sichtfeld von bis zu 42 Grad (vorn) oder 150 Grad (hinten) beinhalten. In mindestens einer Ausführungsform können Kurzstrecken-RADAR-Systeme ohne Einschränkung eine beliebige Anzahl des einen oder der mehreren RADAR-Sensoren **3460** beinhalten, die an beiden Enden eines hinteren Stoßfängers eingebaut werden sollen. Wenn sie an beiden Enden eines hinteren Stoßfängers eingebaut sind, kann ein RADAR-Sensorsystem in mindestens einer Ausführungsform zwei Strahlen erzeugen, die durchgehend den toten Winkel auf der Rückseite und neben dem Fahrzeug überwachen. In mindestens einer Ausführungsform können Kurzstrecken-RADAR-Systeme in dem ADAS-System **3438** zur Erfassung des toten Winkels und/oder Spurwechselunterstützung verwendet werden.

[0429] In mindestens einer Ausführungsform kann das Fahrzeug **3400** ferner den einen oder die mehreren Ultraschallsensoren **3462** beinhalten. Der eine oder die mehreren Ultraschallsensoren **3462**, die an der Vorderseite, der Rückseite und/oder den Seiten des Fahrzeugs **3400** positioniert sein können, können für die Einparkunterstützung verwendet werden und/oder um ein Belegungsraster zu erzeugen und zu aktualisieren. In mindestens einer Ausführungsform kann eine breite Vielfalt von dem einen oder den mehreren Ultraschallsensoren **3462** verwendet werden und unterschiedliche Ultraschallsensoren **3462** können für unterschiedliche Reichweiten der Erfassung (z. B. 2,5 m, 4 m) verwendet werden. In mindestens einer Ausführungsform können der eine oder die mehreren Ultraschallsensoren **3462** bei den Funktionssicherheitsstufen von ASIL B betrieben werden.

[0430] In mindestens einer Ausführungsform kann das Fahrzeug **3400** den einen oder die mehreren LIDAR-Sensoren **3464** beinhalten. Der eine oder die mehreren LIDAR-Sensoren **3464** können zur Objekt- und Fußgängererfassung, Notbremsung, Zusammenstoßvermeidung und/oder zu anderen Funktionen verwendet werden. In mindestens einer Ausführungsform können der eine oder die mehreren LIDAR-Sensoren **3464** die Funktionssicherheitsstufe ASIL B aufweisen. In mindestens einer Ausführungsform kann das Fahrzeug **3400** den einen oder die mehreren LIDAR-Sensoren **3464** (z. B. zwei, vier, sechs usw.) beinhalten, die Ethernet verwenden können (z. B. um einem Gigabit-Ethernet-Switch Daten bereitzustellen).

[0431] In mindestens einer Ausführungsform können der eine oder die mehreren LIDAR-Sensoren **3464** in der Lage sein, eine Liste von Objekten und ihren Entfernungen für ein 360-Grad-Sichtfeld bereitzustellen. In mindestens einer Ausführungsform können der eine oder die mehreren handelsüblichen LIDAR-Sensoren **3464** zum Beispiel eine ausgeschriebene Reichweite von ungefähr 100 m aufweisen, mit einer Genauigkeit von 2 cm-3 cm, und mit einer Unterstützung für eine 100 Mbps-Ethernet-Verbindung. In mindestens einer Ausführungsform können ein oder mehrere nichthervorstehende LIDAR-Sensoren **3464** verwendet werden. In einer derartigen Ausführungsform können der eine oder die mehreren LIDAR-Sensoren **3464** als eine kleine Vorrichtung umgesetzt sein, die in eine Vorderseite, Rückseite, die Seiten und/oder die Ecken des Fahrzeugs **3400** eingebettet sein können. In mindestens einer Ausführungsform können der eine oder die mehreren LIDAR-Sensoren **3464** in einer derartigen Ausführungsform ein Sichtfeld mit 120 Grad in der Waagerechten und 35 Grad in der Senkrechten mit einer Reichweite von 200 m bereitstellen, selbst für Objekte mit geringem Reflexionsvermögen. In mindestens einer Ausführungsform können der eine oder die mehreren LIDAR-Sensoren **3464** für ein waagerechtes Sichtfeld zwischen 45 Grad und 135 Grad konfiguriert sein.

[0432] In mindestens einer Ausführungsform können LIDAR-Techniken, wie etwa 3D-Flash-LIDAR, ebenfalls verwendet werden. 3D-Flash-LIDAR verwendet einen Blitz eines Lasers als eine Übermittlungsquelle, um die Umgebung des Fahrzeugs **3400** bis zu ungefähr 200 m zu beleuchten. In mindestens einer Ausführungsform beinhaltet eine Flash-LIDAR-Einheit ohne Einschränkung einen Empfänger, der die Laserimpulsübermittlungszeit und das reflektierte Licht an jedem Bildpunkt aufzeichnet, was wiederum der Reichweite von dem Fahrzeug **3400** zu Objekten entspricht. In mindestens einer Ausführungsform kann Flash-LIDAR hochgenaue und störungsfreie Bilder der Umgebung ermöglichen, die mit jedem Laserblitz erzeugt werden soll. In mindestens einer Ausführungsform können vier Flash-LIDAR-Sensoren eingesetzt werden, einer an jeder Seite des Fahrzeugs **3400**. In mindestens einer Ausführungsform beinhalten 3D-Flash-LIDAR-Systeme ohne Einschränkung eine Festkörper-3D-Starranordnungs-LIDAR-Kamera mit keinen beweglichen Teilen außer einem Lüfter (z. B. eine nichtabtastende LIDAR-Vorrichtung). In mindestens einer Ausführungsform können die eine oder die mehreren Flash-LIDAR-Vorrichtungen einen Laserimpuls von 5 Nanosekunden der Klasse I (augensicher) pro Einzelbild verwenden und können reflektiertes Laserlicht in Form von 3D-Reichweitenpunktwolken und koregistrierten Lichtstärkedaten aufnehmen.

[0433] In mindestens einer Ausführungsform kann das Fahrzeug ferner einen oder mehrere IMU-Sensoren **3466** beinhalten. In mindestens einer Ausführungsform können sich der eine oder die mehreren IMU-Sensoren **3466** in mindestens einer Ausführungsform in einem Mittelpunkt einer Hinterachse des Fahrzeugs **3400** befinden. In mindestens einer Ausführungsform können der eine oder die mehreren IMU-Sensoren **3466** zum Beispiel und ohne Einschränkung Folgendes beinhalten: einen oder mehrere Beschleunigungssensoren, ein oder mehrere Magnetometer, ein oder mehrere Gyroskope, einen oder mehrere Magnetkompasse und/oder andere Sensorarten. In mindestens einer Ausführungsform, wie etwa in Anwendungen mit sechs Achsen, können der eine oder die mehreren IMU-Sensoren **3466** ohne Einschränkung Beschleunigungssensoren und Gyroskope beinhalten. In mindestens einer Ausführungsform, wie etwa in Anwendungen mit neun Achsen, können der eine oder die mehreren IMU-Sensoren **3466** ohne Einschränkung Beschleunigungssensoren und Gyroskope und Magnetometer beinhalten.

[0434] In mindestens einer Ausführungsform können der eine oder die mehreren IMU-Sensoren **3466** als ein GPS-gestütztes Miniaturhochleistungsträgheitsnavigationssystem (GPS-Aided Inertial Navigation System - „GPS/INS“) umgesetzt sein, das Trägheitssensoren von mikroelektromechanischen Systemen („MEMS“), einen hochempfindlichen GPS-Empfänger und erweiterte Kalman-Filteralgorithmen kombiniert, um Schätzungen der Position, Geschwindigkeit und des Verhaltens bereitzustellen. In mindestens einer Ausführungsform können der eine oder die mehreren IMU-Sensoren **3466** ermöglichen, dass das Fahrzeug **3400** den Kurs durch das direkte Beobachten und Abgleichen von Änderungen bei der Geschwindigkeit von dem GPS mit dem einen oder den mehreren IMU-Sensoren **3466** schätzt, ohne eine Eingabe von einem Magnetsensor zu benötigen. In mindestens einer Ausführungsform können der eine oder die mehreren IMU-Sensoren **3466** und der eine oder die mehreren GNSS-Sensoren **3458** in einer einzelnen integrierten Einheit kombiniert sein.

[0435] In mindestens einer Ausführungsform kann das Fahrzeug **3400** ein oder mehrere Mikrofone **3496** beinhalten, die in und/oder um das Fahrzeug **3400** platziert sind. In mindestens einer Ausführungsform können das eine oder die mehreren Mikrofone **3496** unter anderem zur Einsatzfahrzeugerfassung und -identifizierung verwendet werden.

[0436] In mindestens einer Ausführungsform kann das Fahrzeug **3400** ferner eine beliebige Anzahl von Kameraarten beinhalten, die Folgende beinhalten: eine oder mehrere Stereokameras **3468**, eine oder mehrere Weitsichtkameras **3470**, eine oder mehrere Infrarotkameras **3472**, eine oder mehrere Rundumkameras **3474**, eine oder mehrere Langstreckenkameras **3498**, eine oder mehrere Mittelstreckenkameras **3476** und/oder andere Kameraarten. In mindestens einer Ausführungsform können Kameras verwendet werden, um Bilddaten um einen gesamten Umfang des Fahrzeugs **3400** aufzunehmen. In mindestens einer Ausführungsform hängen die verwendeten Arten von Kameras von dem Fahrzeug **3400** ab. In mindestens einer Ausführungsform kann eine beliebige Kombination von Kameraarten verwendet werden, um eine notwendige Abdeckung um das Fahrzeug **3400** bereitzustellen. In mindestens einer Ausführungsform kann sich eine Anzahl von Kameras abhängig von der Ausführungsform unterscheiden. Zum Beispiel könnte das Fahrzeug **3400** in mindestens einer Ausführungsform sechs Kameras, sieben Kameras, zehn Kameras, zwölf Kameras oder eine andere Anzahl von Kameras beinhalten. Kameras können als ein Beispiel und ohne Einschränkung eine serielle Gigabit-Multimedia-Verbindung (Gigabit Multimedia Serial Link - „GMSL“) und/oder Gigabit-Ethernet unterstützen. In mindestens einer Ausführungsform ist jede der einen oder der mehreren Kameras im Hinblick auf **Fig. 34A** und **Fig. 34B** vorher in dieser Schrift ausführlicher beschrieben.

[0437] In mindestens einer Ausführungsform kann das Fahrzeug **3400** ferner den einen oder die mehreren Vibrationssensoren **3442** beinhalten. In mindestens einer Ausführungsform können der eine oder die mehreren Vibrationssensoren **3442** Vibrationen der Komponenten des Fahrzeugs **3400** messen, wie etwa eine oder mehrere Achsen. Zum Beispiel können in mindestens einer Ausführungsform Änderungen bei den Vibrationen eine Änderung des Straßenbelags anzeigen. In mindestens einer Ausführungsform können, wenn zwei oder mehr Vibrationssensoren **3442** verwendet werden, Unterschiede zwischen Vibrationen verwendet werden, um Reibung oder Schlupf von Straßenbelag zu bestimmen (z. B. wenn der Unterschied bei der Vibration zwischen einer leistungsgetriebenen Achse und einer sich frei drehenden Achse besteht).

[0438] In mindestens einer Ausführungsform kann das Fahrzeug **3400** das ADAS-System **3438** beinhalten. Das ADAS-System **3438** kann ohne Einschränkung in einigen Beispielen ein SoC beinhalten. In mindestens einer Ausführungsform kann das ADAS-Systeme **3438** ohne Einschränkung eine beliebige Anzahl und Kombination der Folgenden beinhalten: ein System eines autonomen/adaptiven/automatischen Tempomats (autonomous/adaptive/automatic cruise control - „ACC“), ein System der vorwärtsgerichteten Unfallwarnung (forward crash warning - „FCW“), ein System der automatischen Notfallbremsung (automatic emergency braking - „AEB“), ein System der Spurverlassenswarnung (lane departure warning - „LDW“), ein System der Spurhalte-

unterstützung (lane keep assist - „LKA“), ein System der Warnung vor dem toten Winkel (blind spot warning - „BSW“), ein System der Warnung vor rückwärtigem Querverkehr (rear cross-traffic warning - „RCTW“), ein System der Zusammenstoßwarnung (collision warning - „CW“), ein System zur Spurzentrierung (lane centering - „LC“) und/oder andere Systeme, Merkmale und/oder Funktionen.

[0439] In mindestens einer Ausführungsform kann das ACC-System den einen oder die mehreren RADAR-Sensoren **3460**, den einen oder die mehreren LIDAR-Sensoren **3464** und/oder eine beliebige Anzahl der einen oder der mehreren Kameras verwenden. In mindestens einer Ausführungsform kann das ACC-System ein längsverlaufendes ACC-System und/oder ein querverlaufendes ACC-System beinhalten. In mindestens einer Ausführungsform überwacht und steuert das längsverlaufende ACC-System die Entfernung zum Fahrzeug direkt vor dem Fahrzeug **3400** und passt automatisch die Geschwindigkeit des Fahrzeugs **3400** an, um eine sichere Entfernung zu Fahrzeugen davor aufrechtzuerhalten. In mindestens einer Ausführungsform führt das querverlaufende ACC-System das Entfernunghalten durch und empfiehlt dem Fahrzeug **3400** bei Bedarf die Spuren zu wechseln. In mindestens einer Ausführungsform hängt das querverlaufende ACC mit anderen ADAS-Anwendungen zusammen, wie etwa LC und CW.

[0440] In mindestens einer Ausführungsform verwendet das CACC-System Informationen von anderen Fahrzeugen, die über die Netzwerkschnittstelle **3424** und/oder die eine oder die mehreren drahtlosen Antennen **3426** von anderen Fahrzeugen über eine drahtlose Verbindung oder indirekt über eine Netzwerkverbindung (z. B. über das Internet) empfangen werden können. In mindestens einer Ausführungsform können direkte Verbindungen über eine Fahrzeug-zu-Fahrzeug-(vehicleto-vehicle - „V2V“-)Kommunikationsverbindung bereitgestellt werden, während indirekte Verbindungen über eine Infrastruktur-zu-Fahrzeug-(infrastructure-to-vehicle - „I2V“-)Kommunikationsverbindung bereitgestellt werden können. Im Allgemeinen stellt das V2V-Kommunikationskonzept Informationen zu unmittelbar vorausfahrenden Fahrzeugen (z.B. Fahrzeug unmittelbar vor und in derselben Spur wie das Fahrzeug **3400**) bereit, während das I2V-Kommunikationskonzept Informationen zum Verkehr weiter vorn bereitstellt. In mindestens einer Ausführungsform kann das CACC-System eine oder beide von der I2V- und der V2V-Informationsquelle beinhalten. In mindestens einer Ausführungsform kann das CACC-System, angesichts der Informationen von Fahrzeugen vor dem Fahrzeug **3400**, zuverlässiger sein und weist größeres Potential auf, um die Gleichmäßigkeit des Verkehrsflusses zu verbessern und die Verkehrsbelastung auf der Straße zu verringern.

[0441] In mindestens einer Ausführungsform ist das FCW-System ausgestaltet, um den Fahrer auf eine Gefahr aufmerksam zu machen, sodass der Fahrer Korrekturmaßnahmen vornehmen kann. In mindestens einer Ausführungsform verwendet das FCW-System eine nach vorn gerichtete Kamera und/oder den einen oder die mehreren RADAR-Sensoren **3460**, die an einen dedizierten Prozessor, DSP, FPGA und/oder ASIC gekoppelt sind, das elektrisch an eine Fahrerrückmeldung gekoppelt ist, wie etwa eine Anzeige, Lautsprecher und/oder eine Vibrationskomponente. In mindestens einer Ausführungsform kann das FCW-System eine Warnung bereitstellen, wie etwa in Form eines Tons, einer optischen Warnung, einer Vibration und/oder einem schnellen Bremsimpuls.

[0442] In mindestens einer Ausführungsform erfasst das AEB-System einen bevorstehenden vorderen Zusammenstoß mit einem anderen Fahrzeug oder anderen Objekt und kann automatisch Bremsen anwenden, wenn der Fahrer keine Korrekturmaßnahmen innerhalb eines festgelegten Zeit- oder Entfernungsparameters unternimmt. In mindestens einer Ausführungsform kann das AEB-System eine oder mehrere nach vorn gerichtete Kameras und/oder den einen oder die mehreren RADAR-Sensoren **3460** verwenden, die an einen dedizierten Prozessor, DSP, FPGA und/oder ASIC gekoppelt sind. In mindestens einer Ausführungsform warnt, wenn das AEB-System eine Gefahr erfasst, das AEB-System typischerweise zuerst den Fahrer, um Korrekturmaßnahmen zu unternehmen, um den Zusammenstoß zu vermeiden und, wenn der Fahrer keine Korrekturmaßnahmen vornimmt, kann das AEB-System automatisch die Bremsen mit dem Ziel anwenden, den Aufprall des vorhergesagten Zusammenstoßes zu verhindern oder mindestens abzuschwächen. In mindestens einer Ausführungsform kann das AEB-System Methoden beinhalten, wie etwa dynamische Bremsunterstützung und/oder Bremsen bei unmittelbar bevorstehendem Unfall.

[0443] In mindestens einer Ausführungsform stellt das LDW-System optische, akustische und/oder fühlbare Warnungen bereit, wie etwa Lenkrad- oder Sitzvibrationen, um einen Fahrer zu warnen, wenn das Fahrzeug **3400** die Spurmarkierungen überquert. In mindestens einer Ausführungsform schaltet sich das LDW-System nicht ein, wenn der Fahrer ein absichtliches Spurverlassen durch das Einschalten eines Fahrtrichtungsanzeigers anzeigt. In mindestens einer Ausführungsform kann das LDW-System auf die Vorderseite gerichtete Kameras verwendet, die an einen dedizierten Prozessor, DSP, FPGA und/oder ASIC gekoppelt sind, das elektrisch an eine Fahrerrückmeldung gekoppelt ist, wie etwa eine Anzeige, Lautsprecher und/oder eine Vibrati-

onskomponente. In mindestens einer Ausführungsform ist das LKA-System eine Variante des LDW-Systems. Das LKA-System stellt Lenkeingabe oder Bremsen bereit, um das Fahrzeug **3400** zu korrigieren, wenn das Fahrzeug **3400** beginnt, die Spur zu verlassen.

[0444] In mindestens einer Ausführungsform erfasst das BSW-System Fahrzeuge in einem toten Winkel des Autos und warnt den Fahrer vor diesen. In mindestens einer Ausführungsform kann das BSW-System eine optische, akustische und/oder fühlbare Warnmeldung bereitstellen, um anzuzeigen, dass das Einscheren auf oder das Wechseln auf Spuren unsicher ist. In mindestens einer Ausführungsform kann das BSW-System eine zusätzliche Warnung bereitstellen, wenn der Fahrer einen Fahrtrichtungsanzeiger verwendet. In mindestens einer Ausführungsform kann das BSW-System eine oder mehrere nach hinten gerichtete Kameras und/oder den einen oder die mehreren RADAR-Sensoren **3460** verwenden, die an einen dedizierten Prozessor, DSP, FPGA und/oder ASIC gekoppelt sind, das elektrisch an eine Fahrerrückmeldung gekoppelt ist, wie etwa eine Anzeige, Lautsprecher und/oder eine Vibrationskomponente.

[0445] In mindestens einer Ausführungsform kann das RCTW-System eine optische, akustisch und/oder fühlbare Benachrichtigung bereitstellen, wenn ein Objekt außerhalb der Reichweite der rückwärtigen Kamera erfasst wird, wenn das Fahrzeug **3400** zurückfährt. In mindestens einer Ausführungsform beinhaltet das RCTW-System das AEB-System, um sicherzustellen, dass die Fahrzeugbremsen angewendet werden, um einen Unfall zu vermeiden. In mindestens einer Ausführungsform kann das RCTW-System den einen oder die mehreren RADAR-Sensoren **3460** verwenden, die an einen dedizierten Prozessor, DSP, FPGA und/oder ASIC gekoppelt sind, das elektrisch an eine Fahrerrückmeldung gekoppelt ist, wie etwa eine Anzeige, Lautsprecher und/oder eine Vibrationskomponente.

[0446] In mindestens einer Ausführungsform können herkömmliche ADAS-Systeme anfällig für falsch positive Ergebnisse sein, was ärgerlich und ablenkend für einen Fahrer sein kann, die aber typischerweise nicht verheerend sind, da herkömmlich ADAS-Systeme den Fahrer warnen und dem Fahrer ermöglichen, zu entscheiden, ob eine Sicherheitslage wirklich besteht, und dementsprechend zu handeln. In mindestens einer Ausführungsform entscheidet das Fahrzeug **3400** selbst im Falle von widersprüchlichen Ergebnissen, ob das Ergebnis von einem Primärcomputer oder einem Sekundärcomputer (z. B. der ersten Steuerung **3436** oder der zweiten Steuerung **3436**) gefolgt werden soll. Zum Beispiel kann in mindestens einer Ausführungsform das ADAS-System **3438** ein Sicherungs- und/oder Sekundärcomputer zum Bereitstellen von Wahrnehmungsinformationen für ein Sicherungscomputerzweckmäßigkeitsmodul sein. In mindestens einer Ausführungsform kann die Sicherungscomputerzweckmäßigkeitsüberwachung eine redundante vielfältige Software auf Hardwarekomponenten ausführen, um Fehler in der Wahrnehmung und dynamische Fahraufgaben zu erfassen. In mindestens einer Ausführungsform können Ausgaben vom ADAS-System **3438** einer Aufsichts-MCU bereitgestellt sein. In mindestens einer Ausführungsform bestimmt, wenn sich die Ausgaben von Primärcomputer und Sekundärcomputer widersprechen, die Aufsichts-MCU wie der Widerspruch gelöst werden soll, um einen sicheren Betrieb sicherzustellen.

[0447] In mindestens einer Ausführungsform kann der Primärcomputer dazu konfiguriert sein, der Aufsichts-MCU einen Vertrauenswert bereitzustellen, der das Vertrauen des Primärcomputers in ein ausgewähltes Ergebnis anzeigt. In mindestens einer Ausführungsform kann, wenn der Vertrauenswert einen Schwellenwert übersteigt, die Aufsichts-MCU der Führung des Primärcomputers folgen, ungeachtet dessen, ob der Sekundärcomputer ein widersprüchliches oder uneinheitliches Ergebnis bereitstellt. In mindestens einer Ausführungsform, bei welcher der Vertrauenswert den Schwellenwert nicht erreicht, und bei welcher der Primär- und der Sekundärcomputer unterschiedliche Ergebnisse (z. B. einen Widerspruch) anzeigen, kann die Aufsichts-MCU zwischen den Computern vermitteln, um das geeignete Endergebnis zu bestimmen.

[0448] In mindestens einer Ausführungsform kann die Aufsichts-MCU dazu konfiguriert sein, ein oder mehrere neuronale Netzwerke auszuführen, die trainiert und dazu konfiguriert sind, mindestens zum Teil auf Grundlage von Ausgaben vom Primärcomputer und Sekundärcomputer Zustände zu bestimmen, unter denen der Sekundärcomputer Falschmeldungen bereitstellt. In mindestens einer Ausführungsform können das eine oder die mehreren neuronalen Netzwerke in der Aufsichts-MCU lernen, wann einer Ausgabe des Sekundärcomputers vertraut werden kann und wann nicht. Zum Beispiel kann in mindestens einer Ausführungsform, wenn der Sekundärcomputer ein RADAR-basiertes FCW-System ist, ein neuronales Netzwerk in der Aufsichts-MCU lernen, wann ein FCW-System metallische Objekte identifiziert, die eigentlich keine Gefahren darstellen, wie etwa ein Abflussrost oder ein Kanaldeckel, die einen Alarm auslösen. In mindestens einer Ausführungsform kann, wenn der Sekundärcomputer ein kamerabasiertes LDW-System ist, ein neuronales Netzwerk in der Aufsichts-MCU lernen, LDW außer Kraft zu setzen, wenn Fahrradfahrer oder Fußgänger anwesend sind und ein Spurverlassen eigentlich das sicherste Manöver ist. In mindestens einer Ausführungsform kann die Aufsichts-

MCU mindestens eines von einem DLA oder einer GPU beinhalten, die für das Ausführen von dem einen oder den mehreren neuronalen Netzwerken mit zugeordnetem Speicher geeignet sind. In mindestens einer Ausführungsform kann die Aufsichts-MCU eine Komponente des einen oder der mehreren SoCs **3404** sein und/oder in dieser als eine solche beinhaltet sein.

[0449] In mindestens einer Ausführungsform kann das ADAS-System **3438** einen Sekundärcomputer beinhalten, der die ADAS-Funktion unter Verwendung traditioneller Regeln des maschinellen Sehens durchführt. In mindestens einer Ausführungsform kann der Sekundärcomputer klassische Regeln des maschinellen Sehens (if-then) verwenden und die Anwesenheit eines neuronalen Netzwerks in der Aufsichts-MCU kann die Zuverlässigkeit, Sicherheit und Leistung verbessern. Zum Beispiel macht in mindestens einer Ausführungsform eine vielfältige Umsetzung und beabsichtigte Nichtidentität ein Gesamtsystem fehlertoleranter, insbesondere für Fehler, die durch eine Funktion von Software (oder einer Software-Hardware-Schnittstelle) verursacht werden. Zum Beispiel kann in mindestens einer Ausführungsform, wenn es einen Software-Bug oder Fehler in der Software gibt, die auf dem Primärcomputer ausgeführt wird, und nichtidentischer Softwarecode, der auf einem Sekundärcomputer ausgeführt wird, dasselbe Gesamtergebnis bereitstellt, die Aufsichts-MCU dann ein größeres Vertrauen aufweisen, dass das Gesamtergebnis korrekt ist und der Bug in der Software oder Hardware auf dem Primärcomputer keinen erheblichen Fehler verursacht.

[0450] In mindestens einer Ausführungsform kann die Ausgabe des ADAS-Systems **3438** in den Wahrnehmungsblock des Primärcomputer und/oder den Aufgabenblock für dynamisches Fahren des Primärcomputers eingespeist werden. Zum Beispiel kann in mindestens einer Ausführungsform, wenn das ADAS-System **3438** eine nach vorn gerichtet Unfallwarnung aufgrund eines unmittelbar davor befindlichen Objekts anzeigt, der Wahrnehmungsblock diese Informationen beim Identifizieren von Objekten verwenden. In mindestens einer Ausführungsform kann der Sekundärcomputer sein eigenes neuronales Netzwerk aufweisen, das trainiert ist und somit das Risiko falscher Positivmeldungen verringert, wie in dieser Schrift beschrieben.

[0451] In mindestens einer Ausführungsform kann das Fahrzeug **3400** ferner ein Infotainment-SoC **3430** (z. B. ein fahrzeuginternes Infotainmentsystem (in-vehicle infotainment -IVI)) beinhalten. Obwohl es als ein SoC veranschaulicht und beschrieben ist, ist das Infotainmentsystem **3430** in mindestens einer Ausführungsform möglicherweise kein SoC und kann ohne Einschränkung zwei oder mehr separate Komponenten beinhalten. In mindestens einer Ausführungsform kann das Infotainment-SoC **3430** ohne Einschränkung Folgendes beinhalten: eine Kombination aus Hardware und Software, die verwendet werden kann, um dem Fahrzeug **3400** Folgendes bereitzustellen: Audio (z. B. Musik, einen persönlichen digitalen Assistenten, Navigationsanweisungen, Nachrichten, Funk usw.), Video (z. B. TV, Filme, Streaming usw.), Telefon (z. B. Freisprechen), Netzwerkanbindungstechnik (z. B. LTE, WiFi usw.) und/oder Informationsdienste (z. B. Navigationssysteme, Rückwärtseinparkunterstützung, ein Funkdatensystem, fahrzeugbezogene Informationen, wie etwa Kraftstoffpegel, insgesamt zurückgelegte Entfernung, Bremsflüssigkeitspegel, Ölpegel, Tür offen/geschlossen, Luftfilterinformationen usw.). Zum Beispiel könnte das Infotainment-SoC **3430** Folgendes beinhalten: Radios, Plattenabspieleinrichtungen, Navigationssysteme, Videoabspieleinrichtungen, USB- und Bluetooth-Anbindung, Carputer, Entertainment im Auto, WiFi, Lenkradaudiosteuerungen, freihändige Sprachsteuerung, eine exponierte Anzeige (heads-up display - „HUD“), die HMI-Anzeige **3434**, eine Telematikvorrichtung, ein Bedienfeld (z. B. zum Steuern und/oder Interagieren mit verschiedenen Komponenten, Merkmalen und/oder Systemen) und/oder andere Komponenten. In mindestens einer Ausführungsform kann das Infotainment-SoC **3430** ferner verwendet werden, um einem oder mehreren Benutzern des Fahrzeugs Informationen (z. B. optisch und/oder akustisch), wie etwa Informationen von dem ADAS-System **3438**, Informationen zum autonomen Fahren, wie etwa geplante Fahrzeugmanöver, Bahnverläufe, Umgebungsinformationen (z. B. Kreuzungsinformationen, Fahrzeuginformationen, Straßeninformationen) und/oder andere Informationen bereitzustellen.

[0452] In mindestens einer Ausführungsform kann das Infotainment-SoC **3430** ein beliebiges Ausmaß und eine beliebige Art von GPU-Funktionsumfang beinhalten. In mindestens einer Ausführungsform kann das Infotainment-SoC **3430** über den Bus **3402** (z. B. der CAN-Bus, Ethernet usw.) mit anderen Vorrichtungen, Systemen und/oder Komponenten des Fahrzeugs **3400** kommunizieren. In mindestens einer Ausführungsform kann das Infotainment-SoC **3430** derartig mit einer Aufsichts-MCU gekoppelt sein, dass die GPU des Infotainmentsystems einige Selbstfahrfunktionen in dem Fall durchführen kann, dass die eine oder die mehreren primären Steuerungen **3436** (z. B. Primär- und/oder Sicherungscomputer des Fahrzeugs **3400**) ausfallen. In mindestens einer Ausführungsform kann das Infotainment-SoC **3430** das Fahrzeug **3400** in einen Modus des Fahrens bis zum sicheren Stillstand versetzen, wie in dieser Schrift beschrieben.

[0453] In mindestens einer Ausführungsform kann das Fahrzeug **3400** ferner das Kombiinstrument **3432** (z. B. ein digitales Armaturenbrett, ein elektronisches Kombiinstrument, ein digitales Armaturenbrett usw.) beinhal-

ten. In mindestens einer Ausführungsform kann das Kombiinstrument **3432** ohne Einschränkung eine Steuerung und/oder einen Supercomputer (z. B. eine separate Steuerung oder einen separaten Supercomputer) beinhalten. In mindestens einer Ausführungsform kann das Kombiinstrument **3432** ohne Einschränkung eine beliebige Anzahl und Kombination eines Satzes von Messtechnik beinhalten, wie etwa ein Geschwindigkeitsmesser, Kraftstoffpegel, Öldruck, Drehzahlmesser, Kilometerzähler, Fahrtrichtungsanzeiger, Schalthebelstellungsanzeigeeinrichtung, eine oder mehrere Sicherheitsgurtwarnleuchten, eine oder mehrere Feststellbremsenwarnleuchten, eine oder mehrere Motorstörungsleuchten, Informationen zu einem zusätzlichen Rückhaltesystem (z. B. Airbag), Beleuchtungssteuerungen, Sicherheitssystemsteuerungen, Navigationsinformationen usw. In einigen Beispielen können Informationen auf dem Infotainment-SoC **3430** und dem Kombiinstrument **3432** angezeigt und/oder von diesen gemeinsam genutzt. In mindestens einer Ausführungsform kann das Kombiinstrument **3432** als Teil des Infotainment-SoC **3430** oder umgekehrt beinhaltet sein.

[0454] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. Details bezüglich der Ableitungs- und/oder Trainingslogik **615** werden nachfolgend in Verbindung mit den **Fig. 6A** und/oder 6B bereitgestellt. In mindestens einer Ausführungsform kann eine Ableitungs- und/oder Trainingslogik **615** im System der **Fig. 34C** zum Ableiten oder Vorhersagen von Operationen mindestens zum Teil auf Grundlage von Gewichtungsparametern verwendet werden, die unter Verwendung von neuronalen Netzwerktrainingsoperationen, Funktionen und/oder Architekturen von neuronalen Netzwerken oder in dieser Schrift beschriebenen Anwendungsfällen von neuronalen Netzwerken berechnet wurden.

[0455] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. In mindestens einer Ausführungsform kann diese Logik mit Komponenten dieser Figuren verwendet werden, um ein dreidimensionales Bild einer bestimmten Ansicht aus zweidimensionalen Bildern aus einer willkürlichen Anzahl von Kameras zu erzeugen.

[0456] **Fig. 34D** ist ein Schaubild eines Systems **3476** zur Kommunikation zwischen einem oder mehreren cloudbasierten Servern und dem autonomen Fahrzeug **3400** aus **Fig. 34A** gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform kann das System **3476** ohne Einschränkung einen oder mehrere Server **3478**, ein oder mehrere Netzwerke **3490** und eine beliebige Anzahl und Art von Fahrzeugen beinhalten, die das Fahrzeug **3400** beinhalten. In mindestens einer Ausführungsform können der eine oder die mehreren Server **3478** ohne Einschränkung eine Vielzahl von GPUs **3484(A)-3484(H)** (gemeinsam in dieser Schrift als GPUs **3484** bezeichnet), PCIe-Switches **3482(A)-3482(D)** (gemeinsam in dieser Schrift als PCIe-Switches **3482** bezeichnet) und/oder CPUs **3480(A)-3480(B)** (gemeinsam in dieser Schrift als CPUs **3480** bezeichnet) beinhalten. GPUs **3484**, CPUs **3480** und PCIe-Switches **3482** können mit Hochgeschwindigkeitsverbindungen miteinander verbunden sein, wie etwa zum Beispiel und ohne Einschränkung NVLink-Schnittstellen 3488, die durch NVIDIA entwickelt wurden, und/oder PCIe-Verbindungen **3486**. In mindestens einer Ausführungsform sind die GPUs **3484** über einen NVLink- und/oder NVSwitch-SoC verbunden und die GPUs **3484** und PCIe-Switches **3482** sind über PCIe-Verbindungen verbunden. Obwohl in mindestens einer Ausführungsform acht GPUs **3484**, zwei CPUs **3480** und vier PCIe-Switches **3482** veranschaulicht sind, ist dies nicht einschränkend gemeint. In mindestens einer Ausführungsform kann jeder des einen oder der mehreren Server **3478** ohne Einschränkung eine beliebige Anzahl von GPUs **3484** CPUs **3480** und/oder PCIe-Switches **3482** in einer beliebigen Kombination beinhalten. Zum Beispiel könnte in mindestens einer Ausführungsform der eine oder die mehreren Server **3478** acht, sechzehn, zweiunddreißig und/oder mehr GPUs **3484** beinhalten.

[0457] In mindestens einer Ausführungsform können der eine oder die mehreren Server **3478**, über das eine oder die mehreren Netzwerke **3490** und von Fahrzeugen, Bilddaten empfangen, die Bilder darstellen, die unerwartete oder geänderte Straßenzustände zeigen, wie etwa kürzlich begonnene Bauarbeiten. In mindestens einer Ausführungsform können der eine oder die mehreren Server **3478**, über das eine oder die mehreren Netzwerke **3490** und an Fahrzeuge, neuronale Netzwerke **3492**, aktualisierte neuronale Netzwerke **3492** und/oder Karteninformationen **3494** übermitteln, die ohne Einschränkung Informationen bezüglich der Verkehrs- und Straßenzustände beinhalten. In mindestens einer Ausführungsform können Aktualisierungen für Karteninformationen **3494** ohne Einschränkung Aktualisierungen für die HD-Karte **3422** beinhalten, wie etwa Informationen in Bezug auf Baustellen, Schlaglöcher, Umleitungen, Überschwemmungen und/oder andere Hindernisse. In mindestens einer Ausführungsform können die neuronalen Netzwerke **3492**, die aktualisierten neuronalen Netzwerke **3492** und/oder die Karteninformationen **3494** aus neuem Training und/oder neuen Erfahrungen, die in Daten dargestellt sind, die von einer beliebigen Anzahl von Fahrzeugen in seiner Umgebung empfangen wurden, und/oder mindestens zum Teil auf Grundlage Training ergeben haben, das an einem Rechenzentrum

(z. B. unter Verwendung von dem einen oder den mehreren Servern **3478** und/oder anderen Servern) durchgeführt wurden.

[0458] In mindestens einer Ausführungsform können der eine oder die mehreren Server **3478** verwendet werden, um Modelle des maschinellen Lernens (z. B. neuronale Netzwerke) mindestens zum Teil auf Grundlage von Trainingsdaten zu trainieren. In mindestens einer Ausführungsform können die Trainingsdaten durch Fahrzeuge erzeugt werden und/oder können in einer Simulation (z. B. unter Verwendung einer Spiel-Engine) erzeugt werden. In mindestens einer Ausführungsform ist eine beliebige Menge an Trainingsdaten gekennzeichnet (z. B. bei denen das zugeordnete neuronale Netzwerk von dem überwachten Lernen profitiert) und/oder anderes Vorverarbeiten durchläuft. In mindestens einer Ausführungsform ist eine beliebige Menge an Trainingsdaten nicht gekennzeichnet und/oder vorverarbeitet (z. B. bei denen das zugeordnete neuronale Netzwerk kein überwachtes Lernen erfordert). In mindestens einer Ausführungsform können, sobald die Modelle des maschinellen Lernens trainiert sind, Modelle des maschinellen Lernens durch Fahrzeuge verwendet werden (z. B. an Fahrzeuge über das eine oder die mehreren Netzwerke **3490** übermittelt) und/oder Modelle des maschinellen Lernens können durch den einen oder die mehreren Server **3478** verwendet werden, um Fahrzeuge per Fernzugriff zu überwachen.

[0459] In mindestens einer Ausführungsform können der eine oder die mehreren Server **3478** Daten von Fahrzeugen empfangen und Daten auf neuronale Echtzeitnetzwerke auf dem neuesten Stand für intelligentes Echtzeitableiten anwenden. In mindestens einer Ausführungsform können der eine oder die mehreren Server **3478** Deep-Learning-Supercomputer und/oder dedizierte KI-Computer beinhalten, die durch die eine oder die mehreren GPUs **3484** betrieben werden, wie etwa DGX- und DGX-Station-Geräte, die von NVIDIA entwickelt wurden. In mindestens einer Ausführungsform können jedoch der eine oder die mehreren Server **3478** Deep-Learning-Infrastruktur beinhalten, die CPU-betriebene Rechenzentren verwenden.

[0460] In mindestens einer Ausführungsform kann die Deep-Learning-Infrastruktur des einen oder der mehreren Server **3478** zu schnellem Echtzeitableiten in der Lage sein und kann diese Fähigkeit verwenden, um den Gesamtzustand von Prozessoren, Software und/oder zugeordneter Hardware im Fahrzeug **3400** zu bewerten und zu überprüfen. Zum Beispiel kann in mindestens einer Ausführungsform Deep-Learning-Infrastruktur regelmäßige Aktualisierungen von dem Fahrzeug **3400** empfangen, wie etwa eine Abfolge von Bildern und/oder Objekten, die das Fahrzeug **3400** in der Abfolge von Bildern gefunden hat (z. B. über Objektzuordnungsmethoden des maschinellen Sehens und/oder des maschinellen Lernens). In mindestens einer Ausführungsform kann die Deep-Learning-Infrastruktur ihr eigenes neuronales Netzwerk ausführen, um Objekte zu identifizieren, und sie mit Objekten vergleichen, die durch das Fahrzeug **3400** identifiziert wurden, und, wenn die Ergebnisse nicht übereinstimmen und die Deep-Learning-Infrastruktur feststellt, dass die KI im Fahrzeug **3400** nicht funktioniert, dann können der eine oder die mehreren Server **3478** ein Signal an das Fahrzeug **3400** übermitteln, dass einen ausfallsicheren Computer des Fahrzeugs **3400** anweist, die Steuerung zu übernehmen, Fahrgäste zu benachrichtigen und ein sicheres Parkmanöver abzuschließen.

[0461] In mindestens einer Ausführungsform können der eine oder die mehreren Server **3478** die eine oder die mehreren GPUs **3484** und einen oder mehrere programmierbare Ableitungsbeschleuniger (z. B. TensorRT 3 von NVIDIA) beinhalten. In mindestens einer Ausführungsform kann die Kombination aus GPU-betriebenen Servern und Ableitungsbeschleunigung eine Echtzeitreaktionsgeschwindigkeit möglich machen. In mindestens einer Ausführungsform, wie etwa, wenn die Leistung weniger entscheidend ist, können Server, die durch CPUs, FPGAs und andere Prozessoren betrieben werden, zum Ableiten verwendet werden. In mindestens einer Ausführungsform wird die Ableitungs- und/oder Trainingslogik **615** verwendet, um eine oder mehrere Ausführungsformen durchzuführen. Details bezüglich der Ableitungs- und/oder Trainingslogik **615** werden nachfolgend in Verbindung mit den **Fig. 6A** und/oder 6B bereitgestellt.

[0462] Ableitungs- und/oder Trainingslogik **615** wird verwendet, um Ableitungs- und/oder Trainingsoperationen durchzuführen, die mit einer oder mehreren Ausführungsformen verknüpft sind. In mindestens einer Ausführungsform kann diese Logik mit Komponenten dieser Figuren verwendet werden, um ein dreidimensionales Bild einer bestimmten Ansicht aus zweidimensionalen Bildern aus einer willkürlichen Anzahl von Kameras zu erzeugen.

[0463] Andere Varianten liegen innerhalb des Geistes der vorliegenden Offenbarung. Somit sind, obwohl die offenbarten Methoden für verschiedene Modifikationen und alternative Bauweisen anfällig sind, bestimmte veranschaulichte Ausführungsformen davon in den Zeichnungen gezeigt und wurden vorstehend ausführlich beschrieben. Es versteht sich jedoch, dass nicht die Absicht verfolgt wird, die Offenbarung auf die konkrete(n) offenbarte(n) Form oder Formen einzuschränken, sondern im Gegenteil die Absicht darin besteht, sämtliche

Modifikationen, alternativen Bauweisen und Äquivalente abzudecken, die, wie in den angefügten Patentansprüchen definiert, in den Geist und Schutzumfang der Offenbarung fallen.

[0464] Die Verwendung der Ausdrücke „ein“ und „eine“ und „der/die/das“ und ähnlicher Referenten im Zusammenhang der Beschreibung der offenbaren Ausführungsformen (besonders im Zusammenhang der folgenden Patentansprüche) soll so ausgelegt werden, dass sie sowohl den Singular als auch den Plural abdeckt, sofern in dieser Schrift nicht anderweitig angegeben oder im eindeutigen Widerspruch zum Kontext, und nicht als eine Definition eines Ausdrucks. Die Ausdrücke „umfassend“, „aufweisend“, „beinhaltend“ und „enthaltend“ sind als offene Ausdrücke auszulegen (d. h. in der Bedeutung „beinhaltend, ohne darauf beschränkt zu sein“), es sei denn, es ist etwas anderes angegeben. Der Ausdruck „verbunden“ ist als teilweise oder vollständig ineinander enthalten, aneinander befestigt oder aneinander angefügt auszulegen, wenn er unmodifiziert vorliegt und sich auf physische Verbindungen bezieht, selbst, wenn ein Element dazwischen eingefügt ist. Die Nennung von Wertebereichen in dieser Schrift soll lediglich als ein schnelles Verfahren des einzelnen Bezugnehmens auf jeden getrennten Wert dienen, der in den Bereich fällt, es sei denn, in dieser Schrift ist etwas anderes angegeben, und jeder getrennte Wert ist in die Beschreibung integriert, als ob er einzeln in dieser Schrift wiedergegeben wäre. Die Verwendung des Ausdrucks „Satz“ (z. B. „ein Satz von Objekten“) oder „Teilsatz“ ist als eine nichtleere Zusammenstellung auszulegen, die ein oder mehrere Elemente umfasst, es sei denn, es ist etwas anderes angemerkt oder dies widerspricht dem Kontext. Ferner bezeichnet der Ausdruck „Teilsatz“ eines entsprechenden Satzes nicht notwendigerweise einen tatsächlichen Teilsatz des entsprechenden Satzes; vielmehr können der Teilsatz und der entsprechende Satz gleich sein, es sei denn, es ist etwas anderes angemerkt oder dies widerspricht dem Kontext.

[0465] Verbindende Sprache, wie etwa Ausdrücke der Form „mindestens eines von A, B, und C“ oder „mindestens eines von A, B und C“, sind, sofern nicht etwas anderes angegeben ist oder etwas anderes eindeutig aus dem Kontext hervorgeht, andernfalls in dem Kontext zu verstehen, in dem sie allgemein verwendet werden, um darzustellen, dass ein Objekt, ein Ausdruck usw. entweder A oder B oder C oder eine beliebige nichtleerer Teilsatz der Gruppe aus A und B und C sein kann. Zum Beispiel beziehen sich die verbindenden Ausdrücke „mindestens eines von A, B, und C“ und „mindestens eines von A, B und C“ in dem veranschaulichenden Beispiel eines Satzes, der drei Elemente aufweist, auf eine beliebige der folgenden Gruppen: {A}, {B}, {C}, {A, B}, {A, C}, {B, C}, {A, B, C}. Somit soll derartige verbindende Sprache im Allgemeinen nicht ausdrücken, dass bestimmte Ausführungen erforderlich machen, dass mindestens eines von A, mindestens eines von B und mindestens eines von C vorhanden sind. Zusätzlich bezeichnet, sofern nicht anders angegeben oder durch den Kontext widersprochen, der Ausdruck „Vielzahl“ außerdem einen Zustand der Pluralität (z. B. „eine Vielzahl von Elementen“ bezeichnet mehrere Elemente). Eine Vielzahl besteht aus mindestens zwei Elementen, kann jedoch auch mehr sein, wenn dies entweder explizit oder durch den Kontext angegeben ist. Ferner bedeutet, sofern nicht anders angegeben oder aus dem Kontext nicht eindeutig hervorgeht, der Ausdruck „auf Grundlage von“ „mindestens zum Teil auf Grundlage von“ und nicht „ausschließlich auf Grundlage von“.

[0466] In dieser Schrift beschriebene Operationen von Prozessen können in einer beliebigen geeigneten Reihenfolge durchgeführt werden, sofern es in dieser Schrift nicht anders angegeben ist oder der Kontext dem anderweitig eindeutig widerspricht. In mindestens einer Ausführungsform ist ein Prozess, wie etwa die in dieser Schrift beschriebenen Prozesse (oder Varianten und/oder Kombinationen davon) unter der Steuerung von einem oder mehreren Computersystemen ausgeführt werden, die mit ausführbaren Anweisungen konfiguriert sind, und wird als Code (z. B. ausführbare Anweisungen, ein oder mehrere Computerprogramme oder eine oder mehrere Anwendungen), die zusammen einen oder mehrere Prozessoren ausführen, durch Hardware oder Kombinationen davon umgesetzt sein. In mindestens einer Ausführungsform ist der Code auf einem computerlesbaren Speichermedium gespeichert, zum Beispiel in Form eines Computerprogramms, das eine Vielzahl von Anweisungen umfasst, die durch einen oder mehrere Prozessoren ausgeführt werden können. In mindestens einer Ausführungsform ist ein computerlesbares Speichermedium ein nichtflüchtiges computerlesbares Speichermedium, das flüchtige Signale ausschließt (z. B. eine sich ausbreitende flüchtige elektrische oder elektromagnetische Übermittlung), aber nicht nichtflüchtige Datenspeicherschaltungen (z.B. Puffer, Zwischenspeicher und Warteschlangen) innerhalb von Sendeempfängern von flüchtigen Signalen beinhaltet. In einigen Ausführungsformen ist der Code (z. B. ausführbarer Code oder Quellcode) auf einem Satz von einem oder mehreren nichtflüchtigen computerlesbaren Speichermedien gespeichert, die darauf gespeicherte ausführbare Anweisungen (oder anderen Speicher, um ausführbare Anweisungen zu speichern) aufweisen, die bei Ausführung (d. h. als Folge der Ausführung) durch einen oder mehrere Prozessoren eines Computersystems das Computersystem veranlassen, in dieser Schrift beschriebene Operationen auszuführen. Ein Satz von nichtflüchtigen computerlesbaren Speichermedien kann in mindestens einer Ausführungsform mehrere nichtflüchtige computerlesbare Speichermedien umfassen und eines oder mehrere von einzelnen nichtflüchtigen Speichermedien der mehreren nichtflüchtigen computerlesbaren Speichermedien verfügen möglicher-

weise nicht über den gesamten Code, während mehrere nichtflüchtige computerlesbares Speichermedien gemeinschaftlich den gesamten Code speichern. In mindestens einer Ausführungsform werden ausführbare Anweisungen derartig ausgeführt, dass unterschiedliche Anweisungen durch unterschiedliche Prozessoren ausgeführt werden. Zum Beispiel speichert ein nichtflüchtiges computerlesbares Speichermedium Anweisungen und eine Hauptzentralverarbeitungseinheit („CPU“) führt einige der Anweisungen aus, während eine Grafikverarbeitungseinheit („GPU“) andere Anweisungen ausführt. Im Allgemeinen können unterschiedliche Komponenten eines Computersystems getrennte Prozessoren aufweisen und können unterschiedliche Prozessoren unterschiedliche Teilsätze der Anweisungen ausführen.

[0467] Dementsprechend sind in mindestens einer Ausführungsform Computersysteme dazu konfiguriert, einen oder mehrere Dienste umzusetzen, die einzeln oder gemeinsam Operationen der in dieser Schrift beschriebenen Prozesse ausführen, und derartige Computersysteme sind mit anwendbarer Hardware und/oder Software konfiguriert, welche die Durchführung von Operationen ermöglicht. Ferner ist ein Computersystem, das verschiedene Ausführungsformen der vorliegenden Offenbarung umsetzt, eine einzelne Vorrichtungen und in einer weiteren Ausführungsform ist ein verteiltes Computersystem, das mehrere Vorrichtungen umfasst, die derartig unterschiedlich betrieben werden, dass das verteilte Computersystem die in dieser Schrift beschriebenen Operationen ausführt, und derartig, dass eine einzelne Vorrichtung nicht alle Operationen ausführt.

[0468] Die Verwendung aller Beispiele oder eine beispielhafte Wortwahl (z. B. „wie etwa“), die in dieser Schrift bereitgestellt sind, soll lediglich die Ausführungsformen der Offenbarung besser veranschaulichen und stellt keine Einschränkung des Schutzumfangs der Offenbarung dar, es sei denn, es ist etwas anderes beansprucht. Keinerlei Formulierung in der Beschreibung sollte so ausgelegt werden, dass sie ein beliebiges nichtbeanspruchtes Element als für die Umsetzung der Erfindung wesentlich angibt.

[0469] Alle Referenzen, die Veröffentlichungen, Patentanmeldungen und Patente beinhalten, die in dieser Schrift erwähnt werden, sind hiermit durch Bezugnahme in demselben Maße aufgenommen, als wäre jede Referenz einzeln und konkret als durch Referenz integriert angegeben und in ihrer Gesamtheit in dieser Schrift dargestellt.

[0470] In der Beschreibung und den Patentansprüchen können die Ausdrücke „gekoppelt“ und „verbunden“ zusammen mit ihren Ableitungen verwendet werden. Es versteht sich, dass diese Ausdrücke nicht als Synonyme füreinander gedacht sind. Vielmehr kann in bestimmten Beispielen „verbunden“ oder „gekoppelt“ verwendet werden, um anzuzeigen, dass zwei oder mehr Elemente in direktem oder indirektem physischem oder elektrischem Kontakt miteinander stehen. „Gekoppelt“ kann auch bedeuten, dass zwei oder mehr Elemente nicht in direktem Kontakt miteinander stehen, aber dennoch kooperieren oder miteinander interagieren.

[0471] Sofern nicht ausdrücklich anders angegeben, versteht es sich, dass sich in der gesamten Beschreibung Ausdrücke wie „Verarbeiten“, „Rechnen“, „Berechnen“, „Bestimmen“ oder Ähnliches auf Handlungen und/oder Prozesse eines Computers oder Computersystems oder eine ähnliche elektronische Rechenvorrichtung beziehen, die als physische Größen dargestellte Daten, wie etwa Elektronik, Größen innerhalb der Register und/oder Speicher des Computersystems, manipuliert und/oder in andere Daten umwandelt, die ähnlich als physische Größen in den Speichern, Registern oder anderen derartigen Informationsspeicher-, -Übermittlung- oder - anzeigevorrichtungen des Computersystems dargestellt sind.

[0472] In ähnlicher Weise kann sich der Ausdruck „Prozessor“ auf jede Vorrichtung oder einen Teil einer Vorrichtung beziehen, die elektronische Daten aus Registern und/oder Speicher verarbeitet und diese elektronischen Daten in andere elektronische Daten umwandelt, die in Register und/oder Speicher gespeichert werden können. Als nichteinschränkende Beispiele kann „Prozessor“ eine CPU oder eine GPU sein. Eine „Rechenplattform“ kann einen oder mehrere Prozessoren umfassen. Wie in dieser Schrift verwendet, können „Software“-Prozesse zum Beispiel Software- und/oder Hardwareeinheiten umfassen, die im Laufe der Zeit Arbeit ausführen, wie etwa Aufgaben, Threads und intelligente Agenten. Außerdem kann sich jeder Prozess auf mehrere Prozesse zum Ausführen von Anweisungen nacheinander oder parallel, kontinuierlich oder intermittierend beziehen. Die Ausdrücke „System“ und „Verfahren“ werden in dieser Schrift insoweit austauschbar verwendet, dass das System ein oder mehrere Verfahren verkörpern kann und Verfahren als ein System betrachtet werden können.

[0473] Im vorliegenden Dokument kann Bezug genommen werden auf das Erhalten, Übernehmen, Empfangen oder Eingeben von analogen oder digitalen Daten in ein Teilsystem, ein Computersystem oder eine computerumgesetzte Maschine. Das Erhalten, Übernehmen, Empfangen oder Eingeben von analogen und digitalen Daten kann auf vielfältige Weise erfolgen, wie etwa durch das Empfangen von Daten als ein Parameter ei-

nes Funktionsaufrufs oder eines Aufrufs an eine Anwendungsprogrammierschnittstelle. In einigen Umsetzungen kann der Prozess des Erhaltens, Übernehmens, Empfangens oder Eingebens von analogen oder digitalen Daten durch das Übertragen von Daten über eine serielle oder parallele Schnittstelle erreicht werden. In einer weiteren Umsetzung kann der Prozess des Erhaltens, Übernehmens, Empfangens oder Eingebens von analogen oder digitalen Daten durch das Übertragen von Daten über ein Computernetzwerk von einer bereitstellenden Einheit zu einer übernehmenden Einheit erreicht werden. Es kann auch Bezug genommen werden auf das Bereitstellen, Ausgeben, Übermitteln, Senden oder Darstellen von analogen oder digitalen Daten. In verschiedenen Beispielen kann der Prozess des Bereitstellens, Ausgebens, Übermittelns, Sendens oder Darstellens von analogen oder digitalen Daten durch das Übertragen von Daten als ein Eingabe- oder Ausgabeparameter eines Funktionsaufrufs, eines Parameters einer Anwendungsprogrammierschnittstelle oder eines Zwischenprozesskommunikationsmechanismus erreicht werden.

[0474] Obwohl die vorstehende Erörterung beispielhafte Umsetzungen beschriebener Methoden darlegt, können andere Architekturen verwendet werden, um die beschriebene Funktion umzusetzen, und sollen im Schutzumfang dieser Offenbarung liegen. Darüber hinaus können, obwohl vorstehend zu Erörterungszwecken konkrete Verteilungen von Verantwortlichkeiten definiert wurden, verschiedene Funktionen und Verantwortlichkeiten in Abhängigkeit von den Umständen auf unterschiedliche Weise verteilt und aufgeteilt werden.

[0475] Darüber hinaus versteht es sich, obwohl der Gegenstand in für Strukturmerkmale oder methodische Handlungen konkreter Sprache beschrieben wurde, dass der in den beigefügten Patentansprüchen beanspruchte Gegenstand nicht notwendigerweise auf die konkreten Merkmale oder Handlungen beschränkt ist. Vielmehr sind die konkreten Merkmale und Handlungen als beispielhafte Formen der Umsetzung der Patentansprüche offenbart.

Patentansprüche

1. Prozessor, umfassend:
eine oder mehrere Schaltungen, um ein oder mehrere neuronale Netzwerke zu verwenden, um ein oder mehrere dreidimensionale, 3D-, Bilder mindestens teilweise auf Grundlage von zwei- oder mehr zweidimensionalen, 2D-, Bildern zu erzeugen, die unterschiedliche Bezugsrahmen aufweisen.

2. Prozessor nach Anspruch 1, wobei die zwei oder mehr 2D-Bilder unter Verwendung einer willkürlichen Anzahl von Kameras mit unterschiedlichen Bezugsrahmen in einer Umgebung aufgenommen werden sollen.

3. Prozessor nach Anspruch 1 oder 2, wobei das eine oder die mehreren neuronalen Netzwerke ein Merkmalsextraktionsnetzwerk beinhalten, um Sätze von semantischen Merkmalen aus den zwei oder mehr 2D-Bildern zu extrahieren.

4. Prozessor nach Anspruch 3, wobei das eine oder die mehreren neuronalen Netzwerke ferner dazu dienen, unter Verwendung der semantischen Merkmale kegelstumpfförmige Punktwolken für die zwei oder mehr 2D-Bilder zu erzeugen und die Punktwolken in einem dreidimensionalen Weltbezugsrahmen zu platzieren.

5. Prozessor nach Anspruch 4, wobei das eine oder die mehreren neuronalen Netzwerke ferner dazu dienen, das eine oder die mehreren 3D-Bilder durch das Projizieren einer Gesamtpunktwolke in dem dreidimensionalen Weltbezugsrahmen auf eine oder mehrere Bezugsebenen zu erzeugen.

6. Prozessor nach einem der vorhergehenden Ansprüche, wobei das eine oder die mehreren neuronalen Netzwerke ferner dazu dienen, einen oder mehrere Bahnverläufe zum Bewegen durch die Umgebung mindestens zum Teil auf Grundlage eines oder mehrerer Objekte zu bestimmen, die in dem einen oder den mehreren 3D-Bildem dargestellt sind.

7. System, umfassend:
eine oder mehrere Prozessoren, um ein oder mehrere neuronale Netzwerke zu verwenden, um ein oder mehrere dreidimensionale, 3D-, Bilder mindestens teilweise auf Grundlage von zwei- oder mehr zweidimensionalen, 2D-, Bildern zu erzeugen, die unterschiedliche Bezugsrahmen aufweisen.

8. System nach Anspruch 7, wobei die zwei oder mehr 2D-Bilder unter Verwendung einer willkürlichen Anzahl von Kameras mit unterschiedlichen Bezugsrahmen in einer Umgebung aufgenommen werden sollen.

9. System nach Anspruch 7 oder 8, wobei das eine oder die mehreren neuronalen Netzwerke ein Merkmalsextraktionsnetzwerk beinhalten, um Sätze von semantischen Merkmalen aus den zwei oder mehr 2D-Bildern zu extrahieren.

10. System nach Anspruch 9, wobei das eine oder die mehreren neuronalen Netzwerke ferner dazu dienen, unter Verwendung der semantischen Merkmale kegelstumpfförmige Punktwolken für die zwei oder mehr 2D-Bilder zu erzeugen und die Punktwolken in einem dreidimensionalen Weltbezugsrahmen zu platzieren.

11. System nach Anspruch 10, wobei das eine oder die mehreren neuronalen Netzwerke ferner dazu dienen, das eine oder die mehreren 3D-Bilder durch das Projizieren einer Gesamtpunktwolke in dem dreidimensionalen Weltbezugsrahmen auf eine oder mehrere Bezugsebenen zu erzeugen.

12. System nach einem der Ansprüche 7 bis 11, wobei das eine oder die mehreren neuronalen Netzwerke ferner dazu dienen, einen oder mehrere Bahnverläufe zum Bewegen durch die Umgebung mindestens zum Teil auf Grundlage eines oder mehrerer Objekte zu bestimmen, die in dem einen oder den mehreren 3D-Bildem dargestellt sind.

13. Verfahren, umfassend:
das Verwenden von einem oder mehreren neuronalen Netzwerken, um ein oder mehrere dreidimensionale, 3D-, Bilder mindestens teilweise auf Grundlage von zwei- oder mehr zweidimensionalen, 2D-, Bildern zu erzeugen, die unterschiedliche Bezugsrahmen aufweisen.

14. Verfahren nach Anspruch 13, wobei die zwei oder mehr 2D-Bilder unter Verwendung einer willkürlichen Anzahl von Kameras mit unterschiedlichen Bezugsrahmen in einer Umgebung aufgenommen werden sollen.

15. Verfahren nach Anspruch 13 oder 14, wobei das eine oder die mehreren neuronalen Netzwerke ein Merkmalsextraktionsnetzwerk beinhalten, um Sätze von semantischen Merkmalen aus den zwei oder mehr 2D-Bildern zu extrahieren.

16. Verfahren nach Anspruch 15, wobei das eine oder die mehreren neuronalen Netzwerke ferner dazu dienen, unter Verwendung der semantischen Merkmale kegelstumpfförmige Punktwolken für die zwei oder mehr 2D-Bilder zu erzeugen und die Punktwolken in einem dreidimensionalen Weltbezugsrahmen zu platzieren.

17. Verfahren nach Anspruch 16, wobei das eine oder die mehreren neuronalen Netzwerke ferner dazu dienen, das eine oder die mehreren 3D-Bilder durch das Projizieren einer Gesamtpunktwolke in dem dreidimensionalen Weltbezugsrahmen auf eine oder mehrere Bezugsebenen zu erzeugen.

18. Verfahren nach einem der Ansprüche 13 bis 17, wobei das eine oder die mehreren neuronalen Netzwerke ferner dazu dienen, einen oder mehrere Bahnverläufe zum Bewegen durch die Umgebung mindestens zum Teil auf Grundlage eines oder mehrerer Objekte zu bestimmen, die in dem einen oder den mehreren 3D-Bildem dargestellt sind.

19. Maschinenlesbares Medium, das einen darauf gespeicherten Satz von Anweisungen aufweist, die, wenn sie von einem oder mehreren Prozessoren ausgeführt werden, den einen oder die mehreren Prozessoren mindestens zu Folgendem veranlassen:
das Verwenden von einem oder mehreren neuronalen Netzwerken, um ein oder mehrere dreidimensionale, 3D-, Bilder mindestens teilweise auf Grundlage von zwei- oder mehr zweidimensionalen, 2D-, Bildern zu erzeugen, die unterschiedliche Bezugsrahmen aufweisen.

20. Maschinenlesbares Medium nach Anspruch 19, wobei die zwei oder mehr 2D-Bilder unter Verwendung einer willkürlichen Anzahl von Kameras mit unterschiedlichen Bezugsrahmen in einer Umgebung aufgenommen werden sollen.

21. Maschinenlesbares Medium nach Anspruch 19 oder 20, wobei das eine oder die mehreren neuronalen Netzwerke ein Merkmalsextraktionsnetzwerk beinhalten, um Sätze von semantischen Merkmalen aus den zwei oder mehr 2D-Bildern zu extrahieren.

22. Maschinenlesbares Medium nach Anspruch 21, wobei das eine oder die mehreren neuronalen Netzwerke ferner dazu dienen, unter Verwendung der semantischen Merkmale kegelstumpfförmige Punktwolken

für die zwei oder mehr 2D-Bilder zu erzeugen und die Punktwolken in einem dreidimensionalen Weltbezugsrahmen zu platzieren.

23. Maschinenlesbares Medium nach Anspruch 22, wobei das eine oder die mehreren neuronalen Netzwerke ferner dazu dienen, das eine oder die mehreren 3D-Bilder durch das Projizieren einer Gesamtpunktwolke in dem dreidimensionalen Weltbezugsrahmen auf eine oder mehrere Bezugsebenen zu erzeugen.

24. Maschinenlesbares Medium nach einem der Ansprüche 19 bis 23, wobei das eine oder die mehreren neuronalen Netzwerke ferner dazu dienen, einen oder mehrere Bahnverläufe zum Bewegen durch die Umgebung mindestens zum Teil auf Grundlage eines oder mehrerer Objekte zu bestimmen, die in dem einen oder den mehreren 3D-Bildern dargestellt sind.

25. Steuersystem, umfassend:
eine oder mehrere Prozessoren, um ein oder mehrere neuronale Netzwerke zu verwenden, um ein oder mehrere dreidimensionale, 3D-, Bilder mindestens teilweise auf Grundlage von zwei- oder mehr zweidimensionalen, 2D-, Bildern zu erzeugen, die unterschiedliche Bezugsrahmen aufweisen; und
Speicher zum Speichern von Netzwerkparametern für das eine oder die mehreren neuronalen Netzwerke.

26. Steuersystem nach Anspruch 25, wobei die zwei oder mehr 2D-Bilder unter Verwendung einer willkürlichen Anzahl von Kameras mit unterschiedlichen Bezugsrahmen in einer Umgebung aufgenommen werden sollen.

27. Steuersystem nach Anspruch 25 oder 26, wobei das eine oder die mehreren neuronalen Netzwerke ein Merkmalsextraktionsnetzwerk beinhalten, um Sätze von semantischen Merkmalen aus den zwei oder mehr 2D-Bildern zu extrahieren.

28. Steuersystem nach Anspruch 27, wobei das eine oder die mehreren neuronalen Netzwerke ferner dazu dienen, unter Verwendung der semantischen Merkmale kegelstumpfförmige Punktwolken für die zwei oder mehr 2D-Bilder zu erzeugen und die Punktwolken in einem dreidimensionalen Weltbezugsrahmen zu platzieren.

29. Steuersystem nach Anspruch 28, wobei das eine oder die mehreren neuronalen Netzwerke ferner dazu dienen, das eine oder die mehreren 3D-Bilder durch das Projizieren einer Gesamtpunktwolke in dem dreidimensionalen Weltbezugsrahmen auf eine oder mehrere Bezugsebenen zu erzeugen.

30. Steuersystem nach einem der Ansprüche 25 bis 29, wobei das eine oder die mehreren neuronalen Netzwerke ferner dazu dienen, einen oder mehrere Bahnverläufe zum Bewegen durch die Umgebung mindestens zum Teil auf Grundlage eines oder mehrerer Objekte zu bestimmen, die in dem einen oder den mehreren 3D-Bildem dargestellt sind.

Es folgen 51 Seiten Zeichnungen

Anhängende Zeichnungen

Erstes Kamerabild **100**

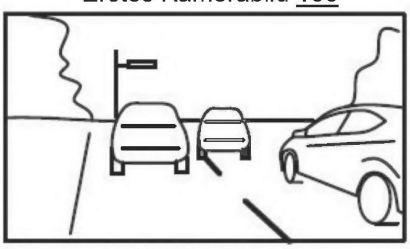


FIG. 1A

Zweites Kamerabild **110**

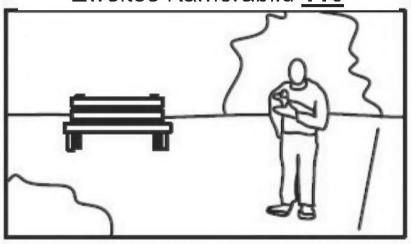


FIG. 1B

Drittes Kamerabild **120**

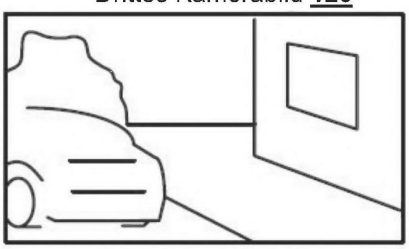


FIG. 1C

Erzeugtes Bild **130**

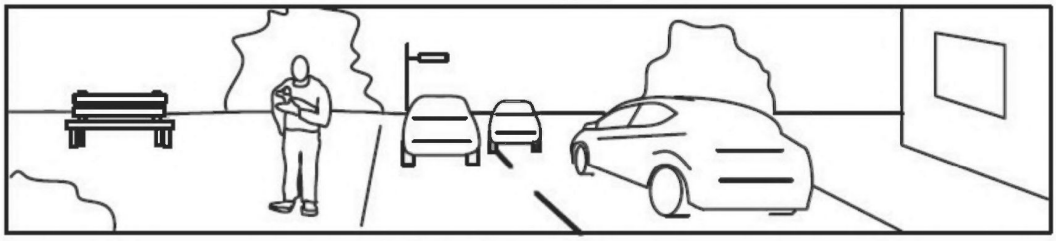


FIG. 1D

Erzeugtes Bild mit abgeleiteter Tiefe **140**

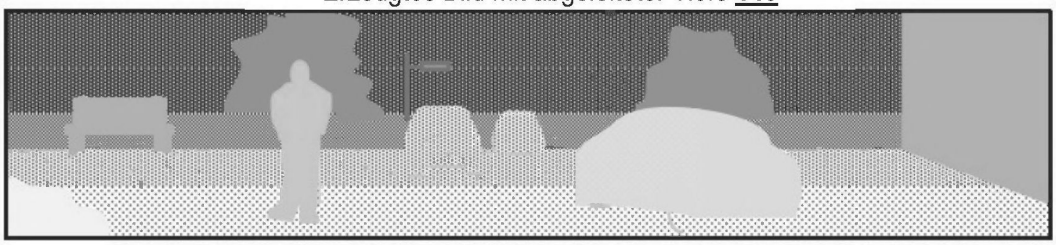


FIG. 1E

Erzeugte Draufsicht **150**

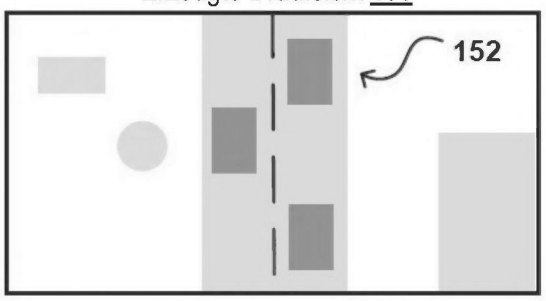


FIG. 1F

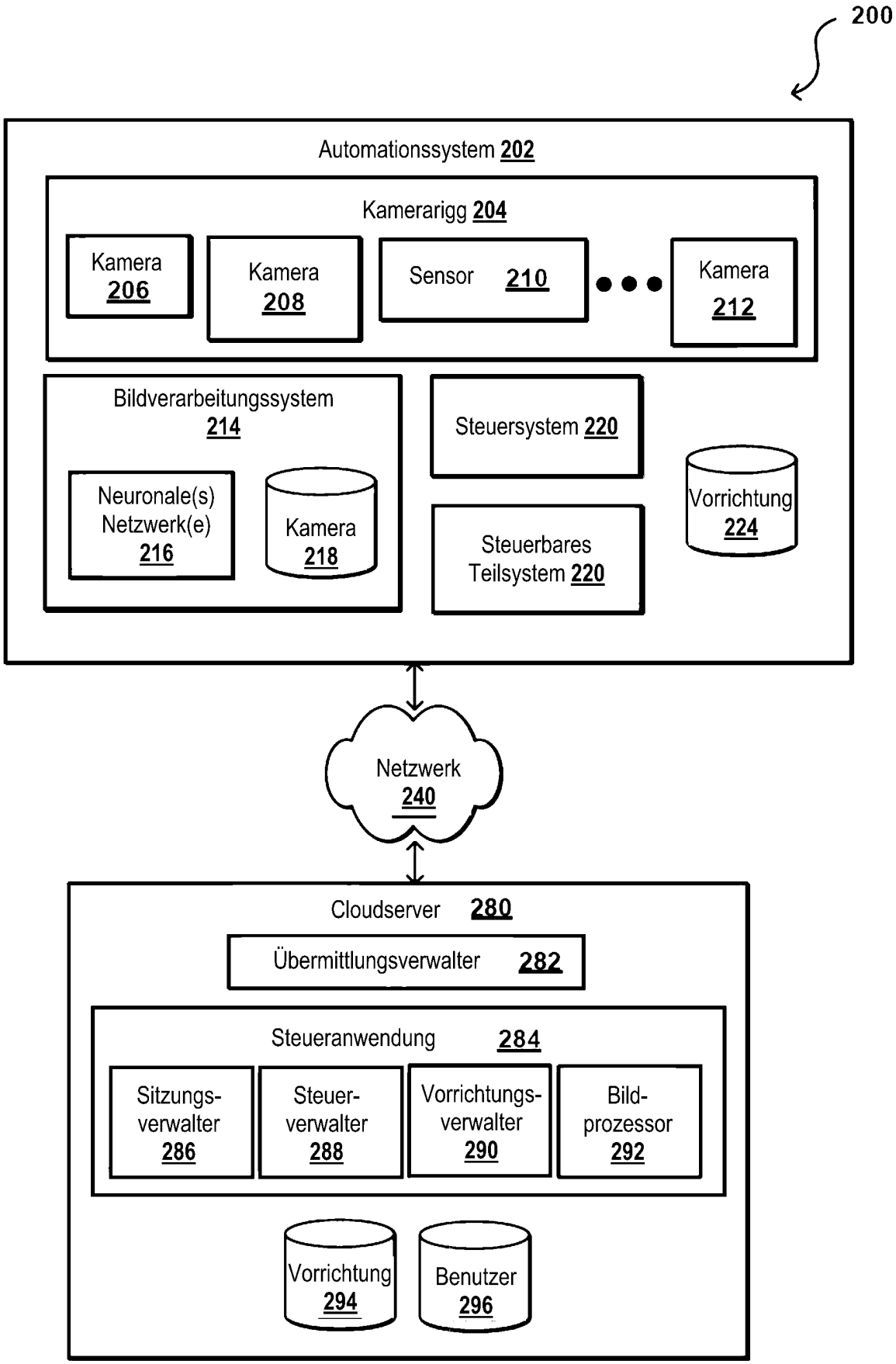

200
Automationssystem 202
Kamerarigg 204
Kamera 206
Kamera 208
Sensor 210
Kamera 212
Bildverarbeitungssystem 214
Neuronale(s) Netzwerk(e) 216
Kamera 218
Steuersystem 220
Steuerbares Teilsystem 220
Vorrichtung 224
Netzwerk 240
Cloudserver 280
Übermittlungsverwalter 282
Steueranwendung 284
Sitzungs-verwalter 286
Steuer-verwalter 288
Vorrichtungs-verwalter 290
Bild-prozessor 292
Vorrichtung 294
Benutzer 296

FIG. 2

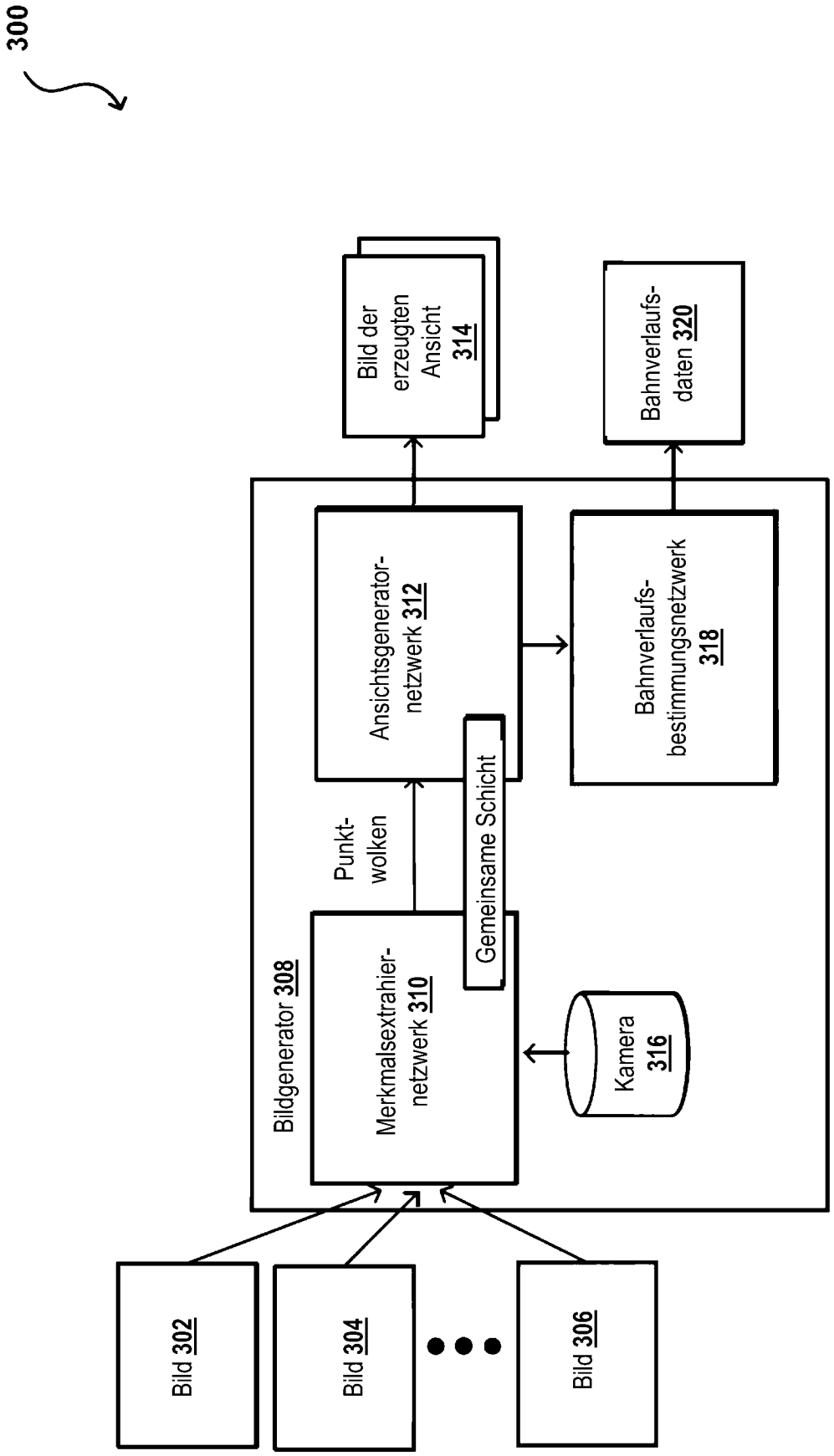

300
Bild der erzeugten Ansicht 314
Bahnverlaufs-daten 320
Ansichtsgenerator-netzwerk 312
Bahnverlaufs-bestimmungsnetzwerk 318
Gemeinsame Schicht
Punkt-wolken
Bildgenerator 308
Merkmalsextrahier-netzwerk 310
Kamera 316
Bild 302
Bild 304
Bild 306

FIG. 3

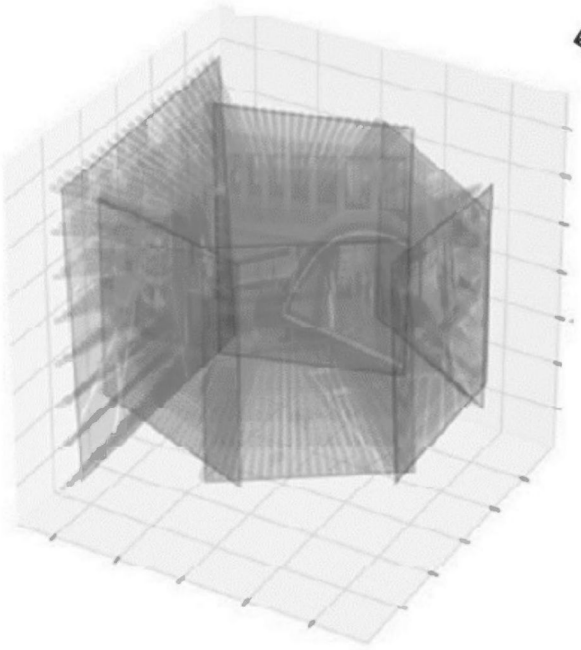


FIG. 4A

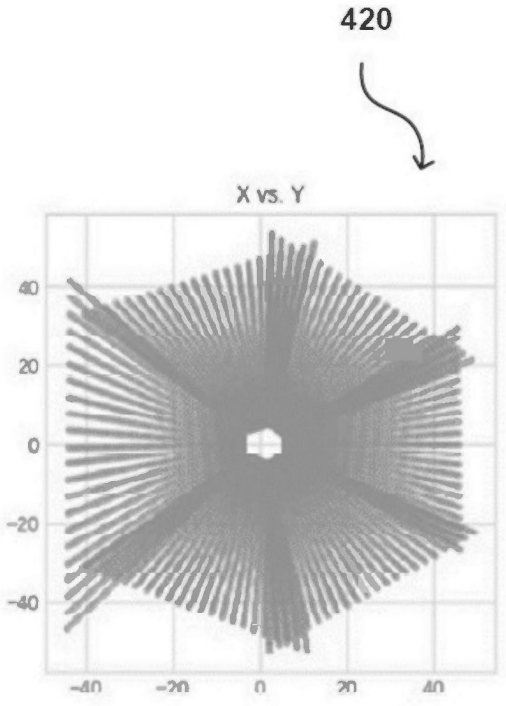


FIG. 4B

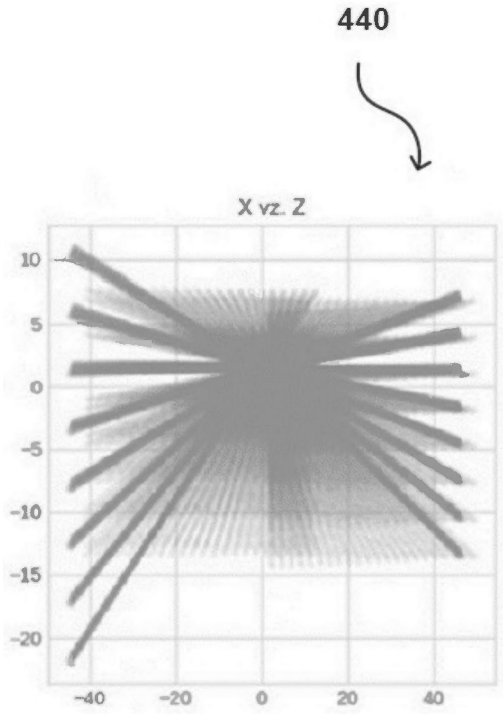


FIG. 4C

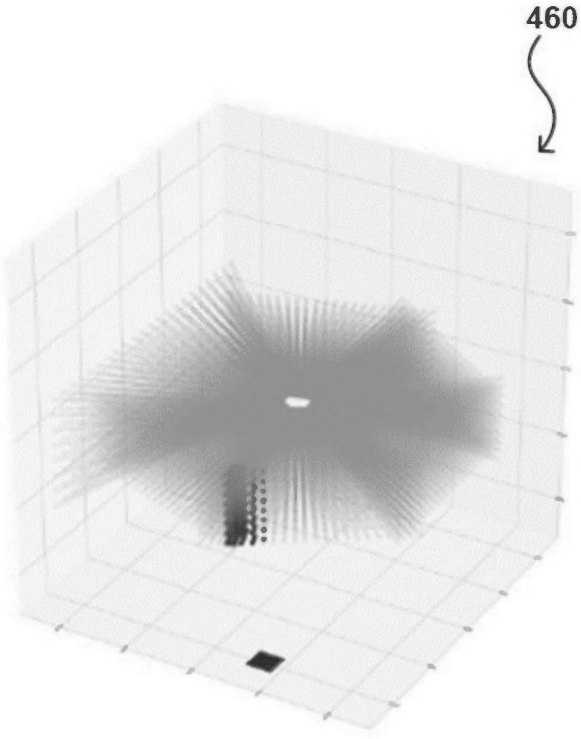


FIG. 4D

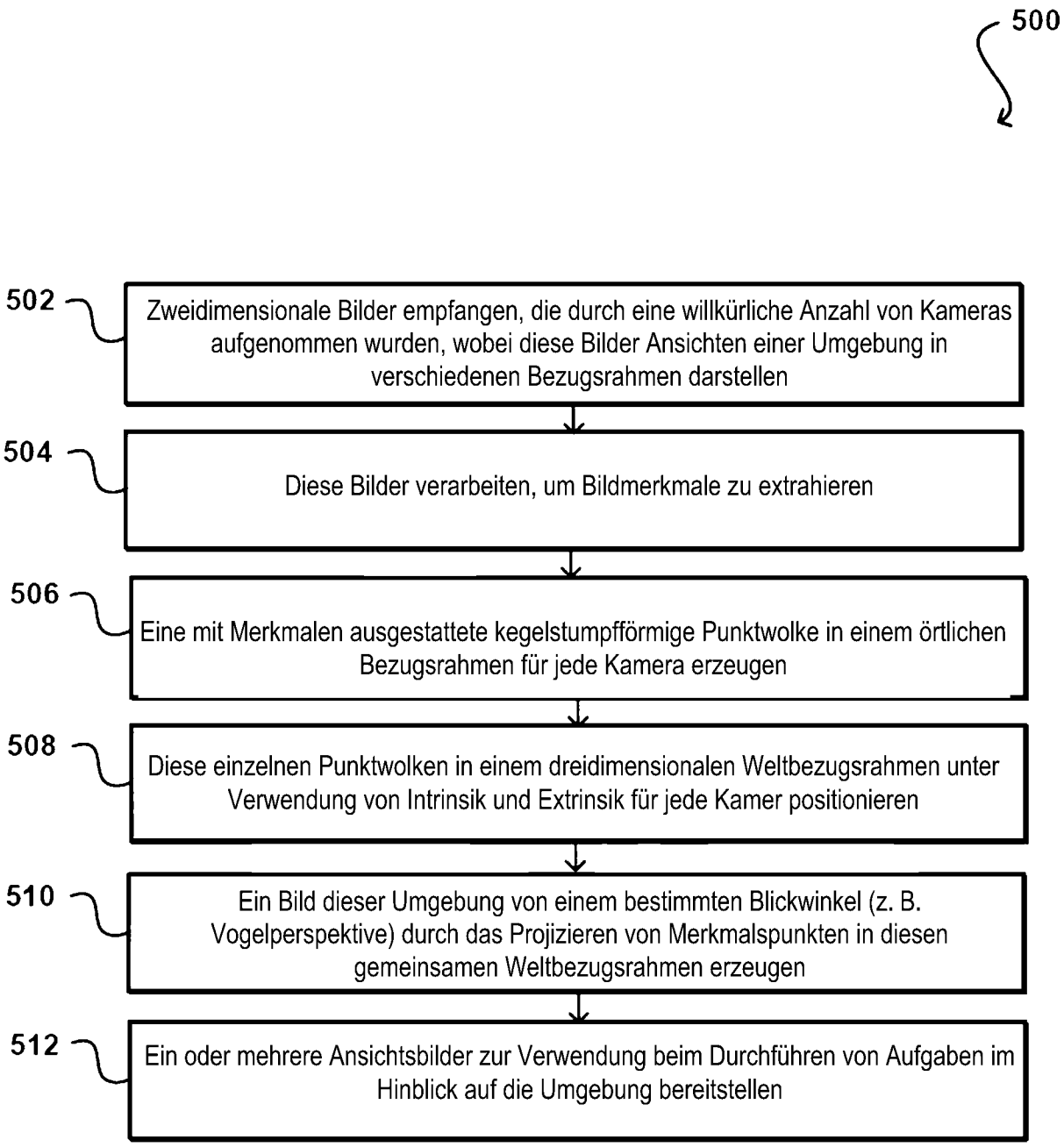

500
502
Zweidimensionale Bilder empfangen, die durch eine willkürliche Anzahl von Kameras aufgenommen wurden, wobei diese Bilder Ansichten einer Umgebung in verschiedenen Bezugsrahmen darstellen
504
Diese Bilder verarbeiten, um Bildmerkmale zu extrahieren
506
Eine mit Merkmalen ausgestattete kegelstumpfförmige Punktwolke in einem örtlichen Bezugsrahmen für jede Kamera erzeugen
508
Diese einzelnen Punktwolken in einem dreidimensionalen Weltbezugsrahmen unter Verwendung von Intrinsik und Extrinsik für jede Kamer positionieren
510
Ein Bild dieser Umgebung von einem bestimmten Blickwinkel (z. B. Vogelperspektive) durch das Projizieren von Merkmalspunkten in diesen gemeinsamen Weltbezugsrahmen erzeugen
512
Ein oder mehrere Ansichtsbilder zur Verwendung beim Durchführen von Aufgaben im Hinblick auf die Umgebung bereitstellen

FIG. 5A

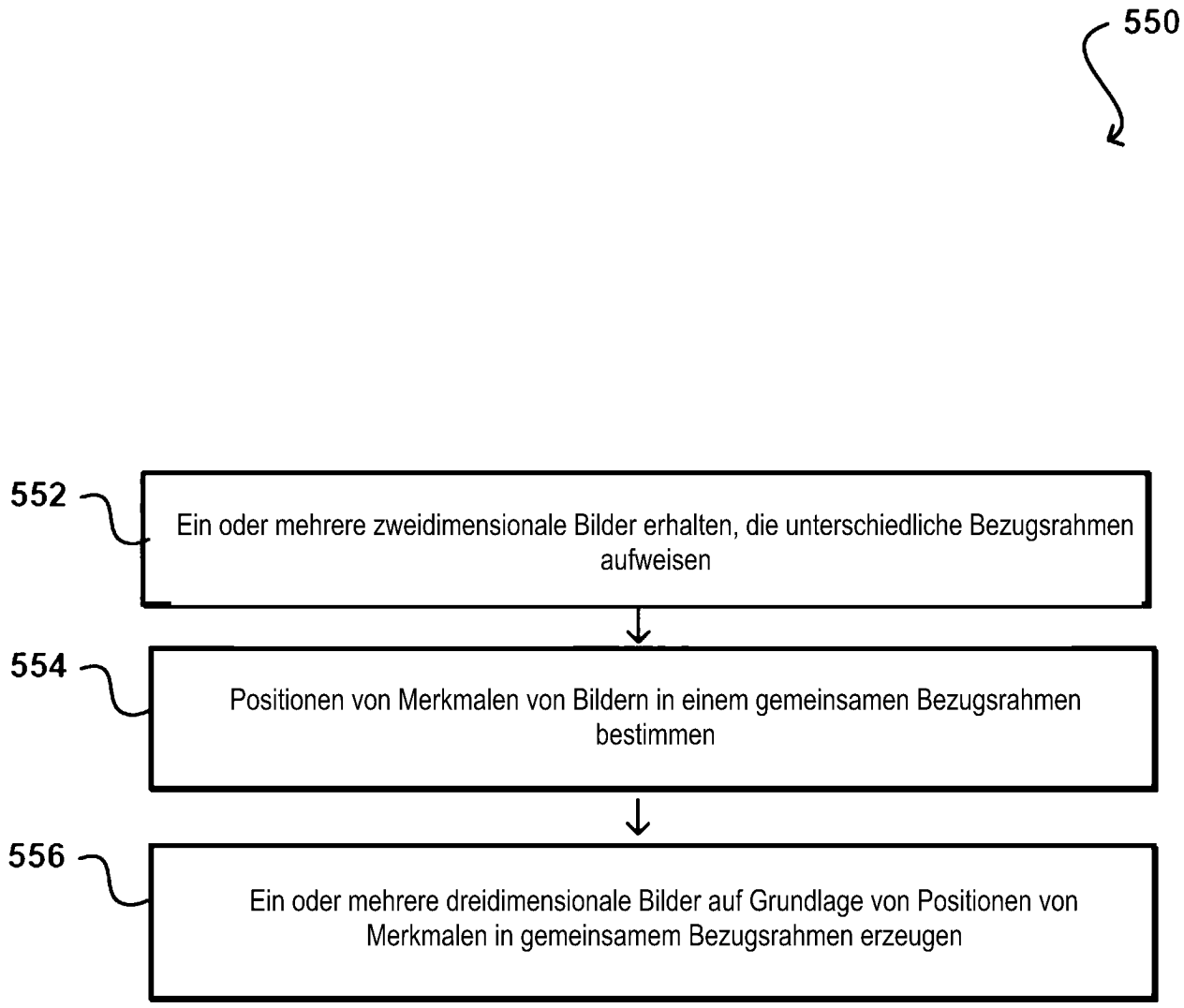

550
552
Ein oder mehrere zweidimensionale Bilder erhalten, die unterschiedliche Bezugsrahmen aufweisen
554
Positionen von Merkmalen von Bildern in einem gemeinsamen Bezugsrahmen bestimmen
556
Ein oder mehrere dreidimensionale Bilder auf Grundlage von Positionen von Merkmalen in gemeinsamem Bezugsrahmen erzeugen

FIG. 5B

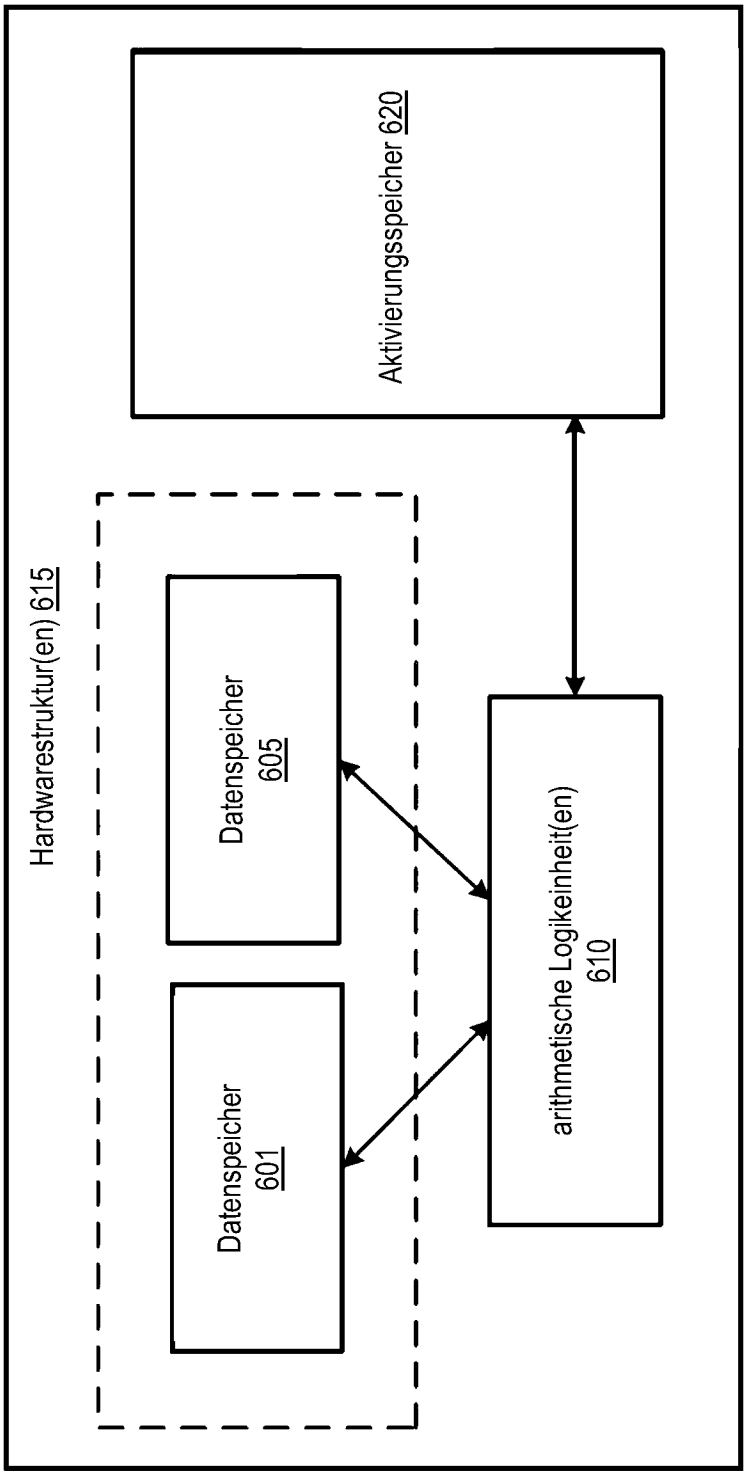

Aktivierungsspeicher 620
Hardwarestruktur(en) 615
Datenspeicher 605
Datenspeicher 601
arithmetische Logikeinheit(en) 610

FIG. 6A

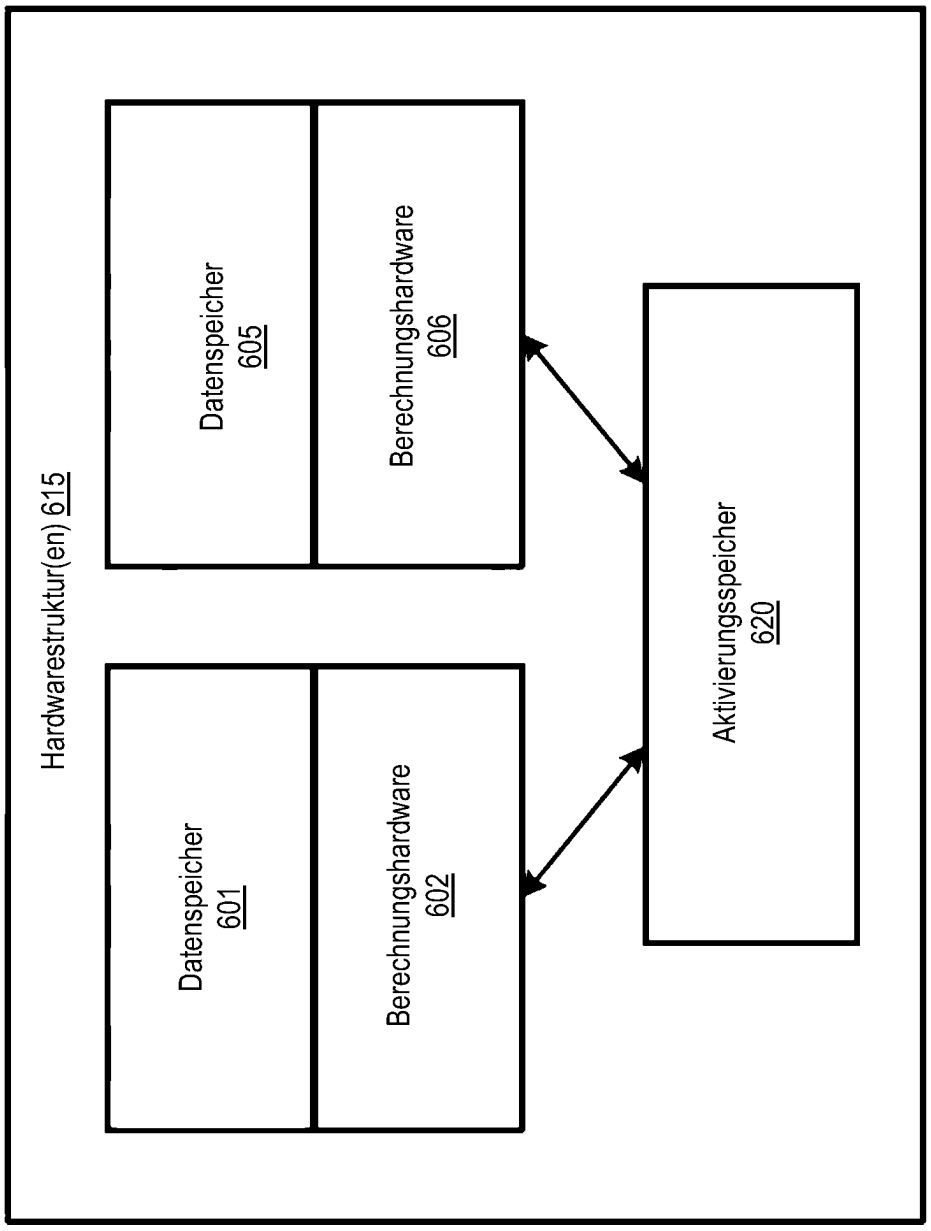

Hardwarestruktur(en) 615
Datenspeicher 605
Berechnungshardware 606
Datenspeicher 601
Berechnungshardware 602
Aktivierungsspeicher 620

FIG. 6B

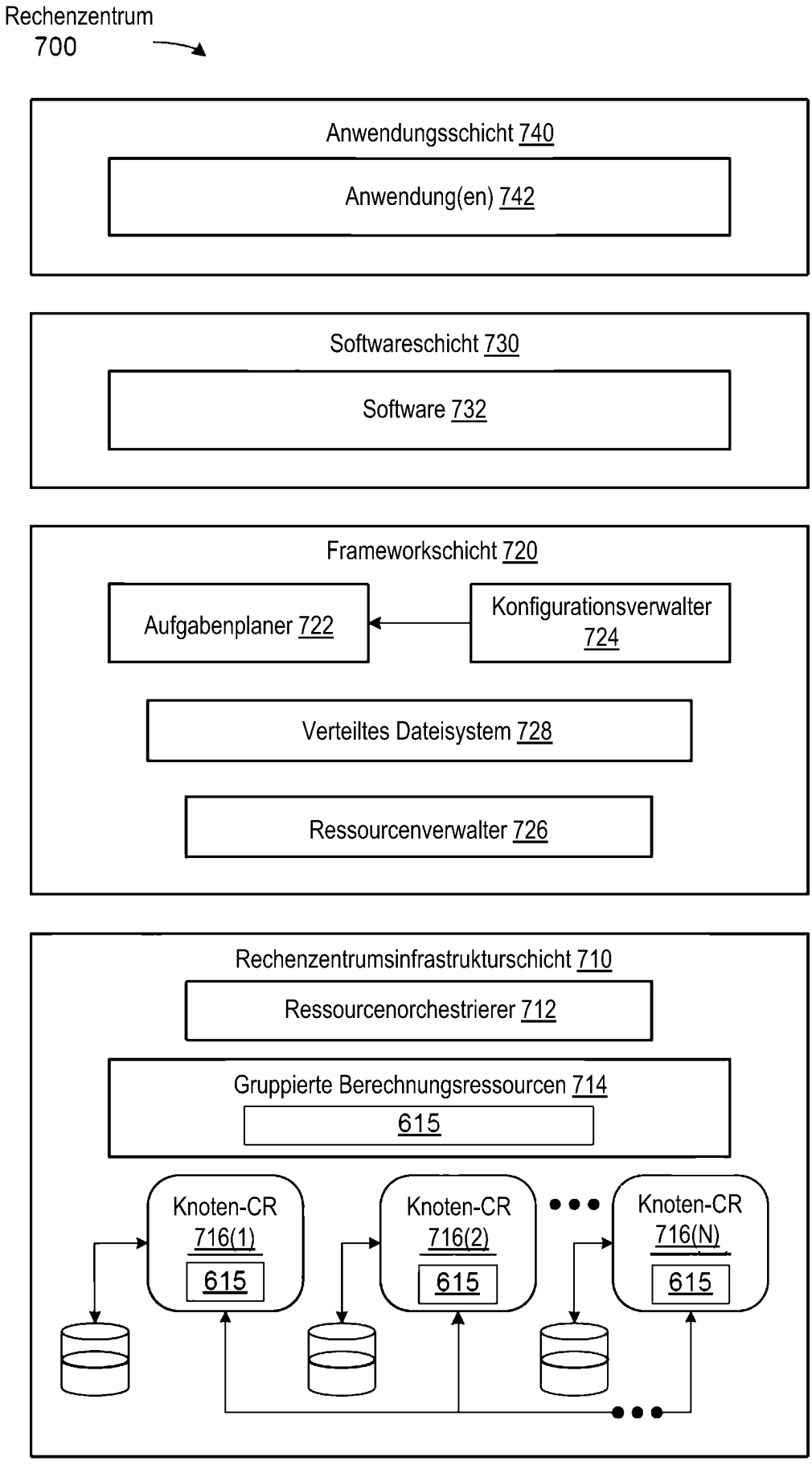

Rechenzentrum
700
Anwendungsschicht 740
Anwendung(en) 742
Softwareschicht 730
Software 732
Frameworkschicht 720
Aufgabenplaner 722
Konfigurationsverwalter 724
Verteiltes Dateisystem 728
Ressourcenverwalter 726
Rechenzentrumsinfrastrukturschicht 710
Ressourcenorchestrierer 712
Gruppierte Berechnungsressourcen 714
615
Knoten-CR 716(1)
615
Knoten-CR 716(2)
615
Knoten-CR 716(N)
615

FIG. 7

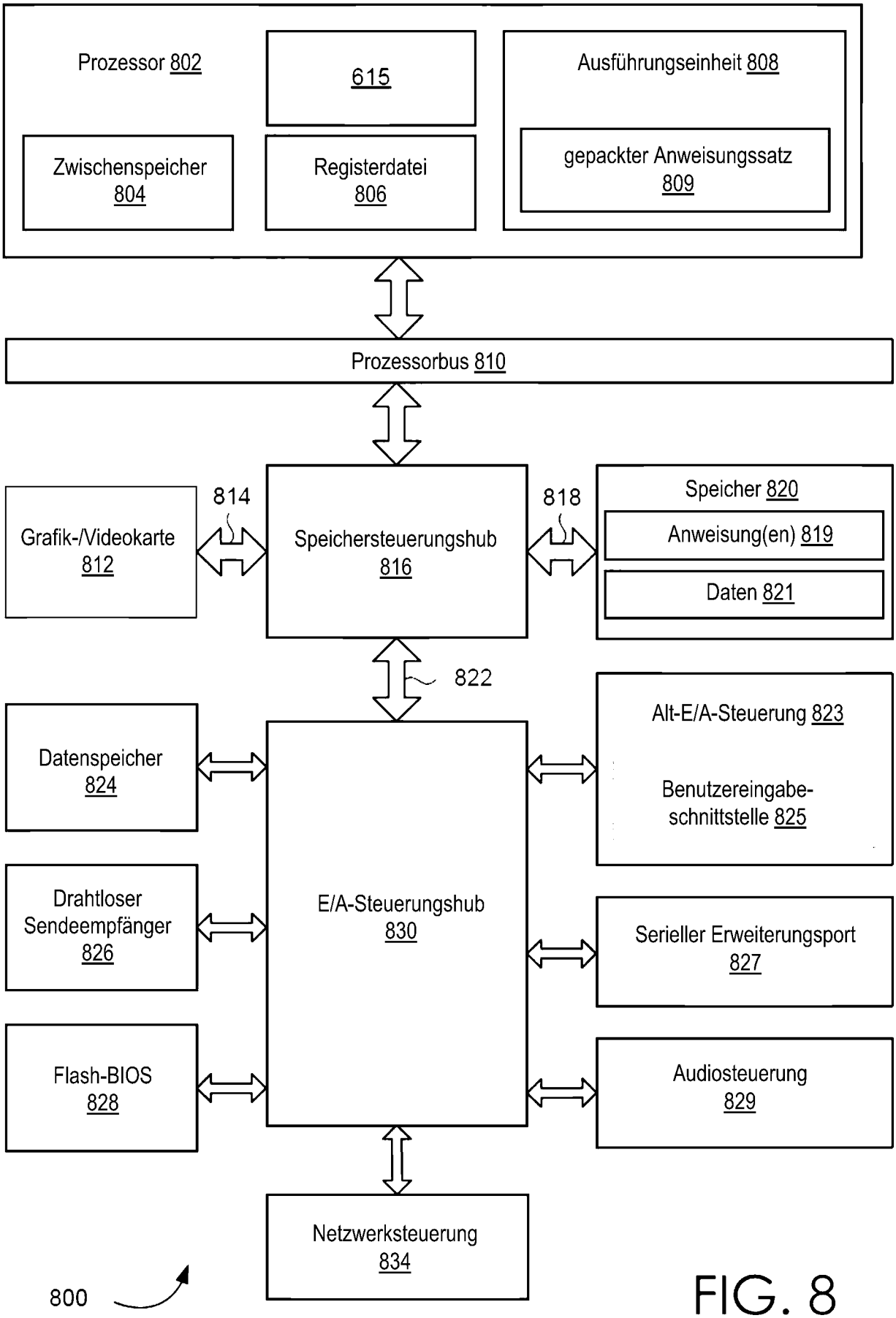

Prozessor 802
615
Ausführungseinheit 808
Zwischenspeicher 804
Registerdatei 806
gepackter Anweisungssatz 809
Prozessorbus 810
814
818
Grafik-/Videokarte 812
Speichersteuerungshub 816
Speicher 820
Anweisung(en) 819
Daten 821
822
Alt-E/A-Steuerung 823
Datenspeicher 824
Benutzereingabe-schnittstelle 825
Drahtloser Sendeempfänger 826
E/A-Steuerungshub 830
Serieller Erweiterungsport 827
Flash-BIOS 828
Audiosteuerung 829
Netzwerksteuerung 834
800

FIG. 8

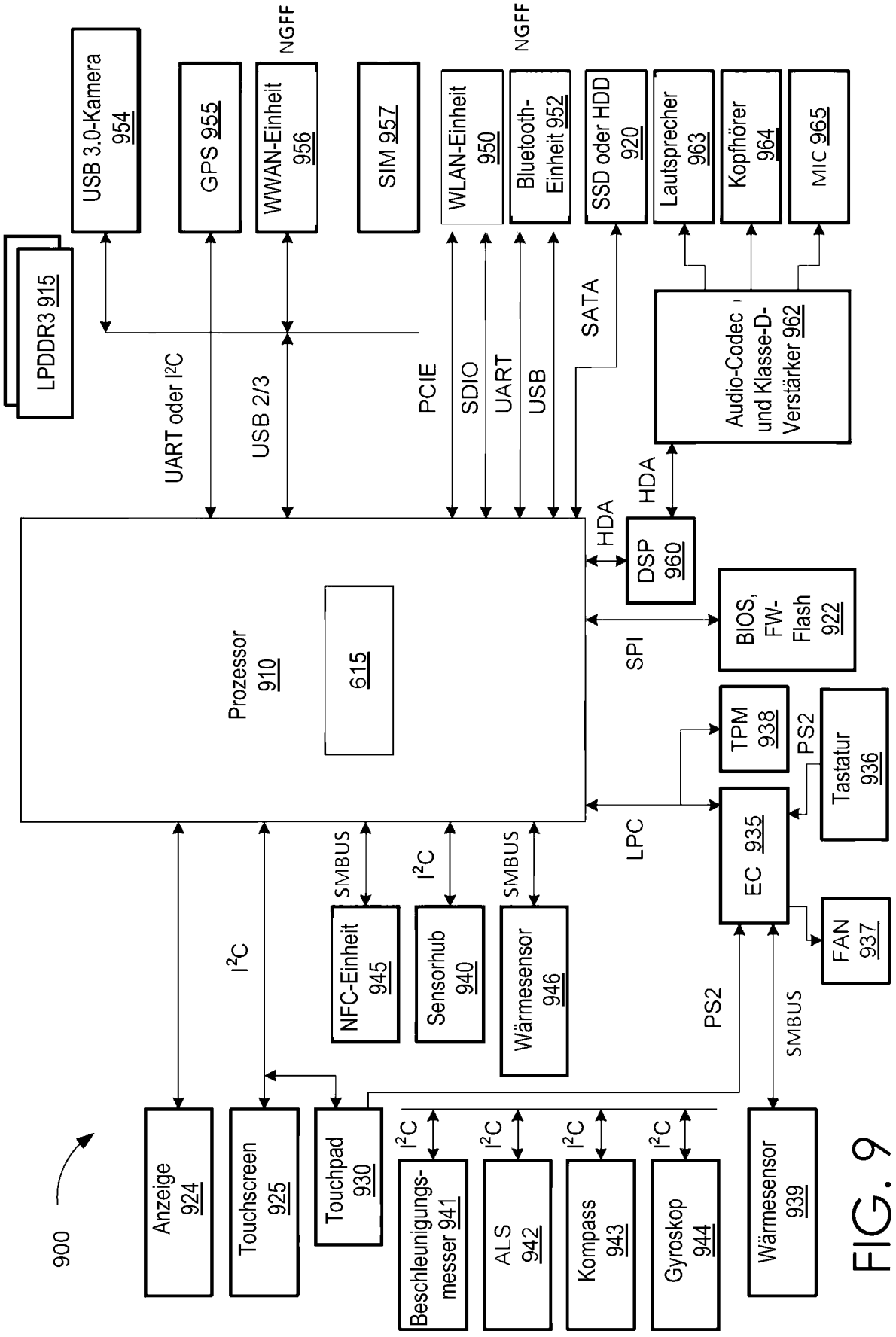

900
Anzeige 924
Touchscreen 925
Touchpad 930
Beschleunigungs-messer 941
ALS 942
Kompass 943
Gyroskop 944
Wärmesensor 939
NFC-Einheit 945
Sensorhub 940
Wärmesensor 946
EC 935
FAN 937
Tastatur 936
TPM 938
BIOS, FW-Flash 922
DSP 960
Prozessor 910
615
LPDDR3 915
USB 3.0-Kamera 954
GPS 955
WWAN-Einheit 956
NGFF
SIM 957
WLAN-Einheit 950
Bluetooth-Einheit 952
NGFF
SSD oder HDD 920
Lautsprecher 963
Kopfhörer 964
MIC 965
Audio-Codec und Klasse-D-Verstärker 962
UART oder I²C
USB 2/3
PCIE
SDIO
UART
USB
SATA
HDA
HDA
SPI
LPC
PS2
PS2
SMBUS
I²C
SMBUS
I²C
SMBUS
I²C
I²C
I²C
I²C

FIG. 9

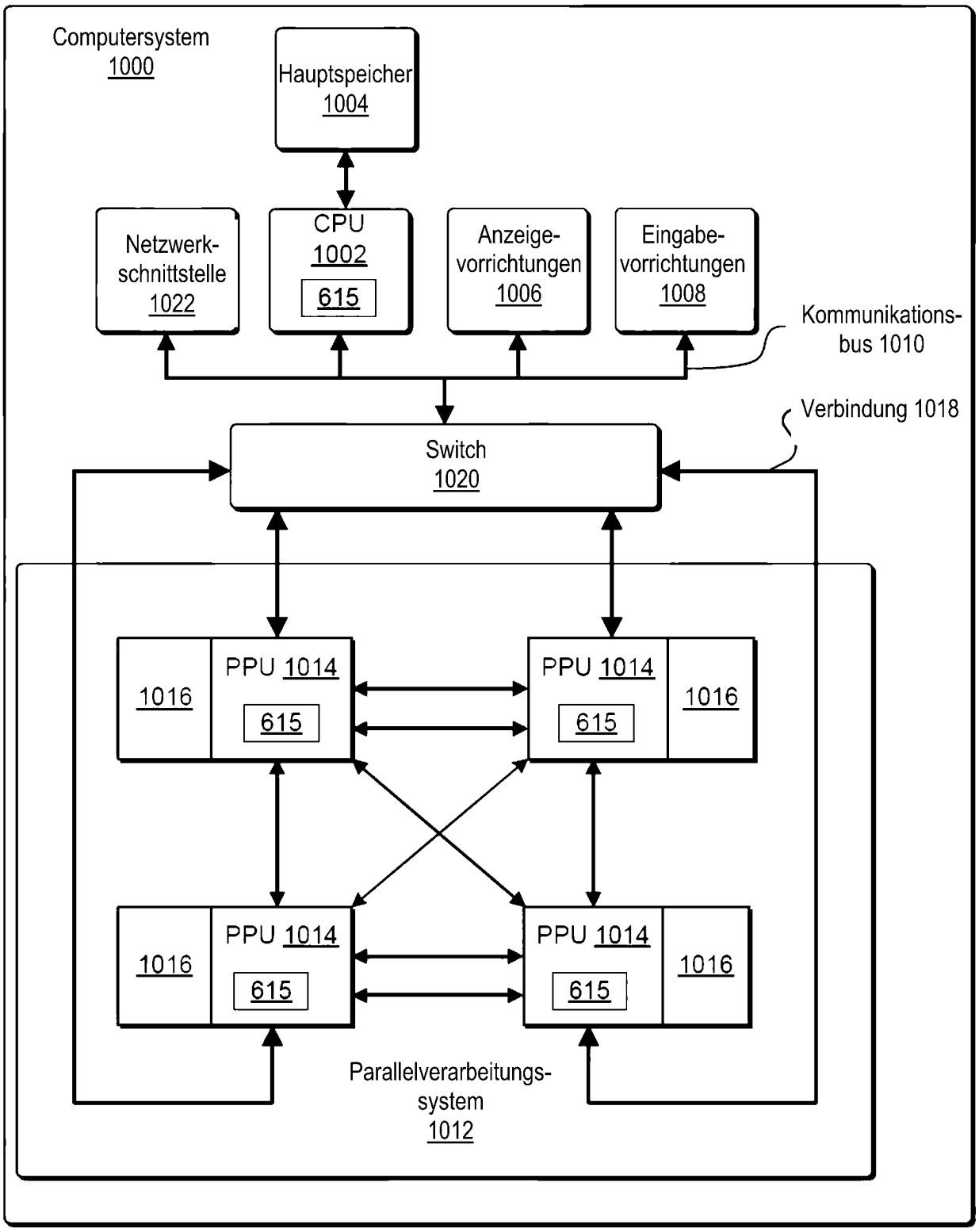

Computersystem
1000
Hauptspeicher
1004
Netzwerk-
schnittstelle
1022
CPU
1002
615
Anzeige-
vorrichtungen
1006
Eingabe-
vorrichtungen
1008
Kommunikations-
bus 1010
Verbindung 1018
Switch
1020
1016
PPU 1014
615
PPU 1014
615
1016
1016
PPU 1014
615
PPU 1014
615
1016
Parallelverarbeitungs-
system
1012

FIG. 10

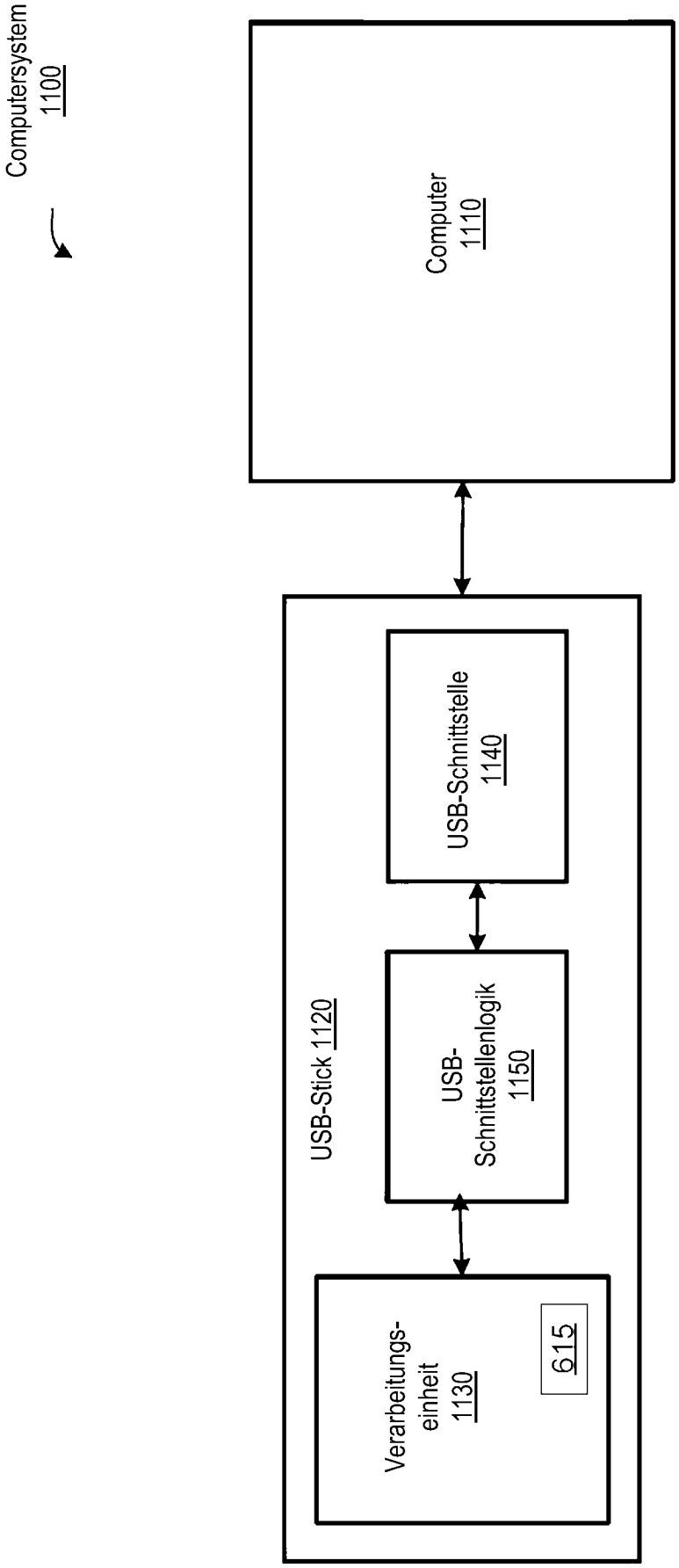

Computersystem
1100
Computer
1110
USB-Stick 1120
USB-Schnittstelle
1140
USB-
Schnittstellenlogik
1150
Verarbeitungs-
einheit
1130
615

FIG. 11

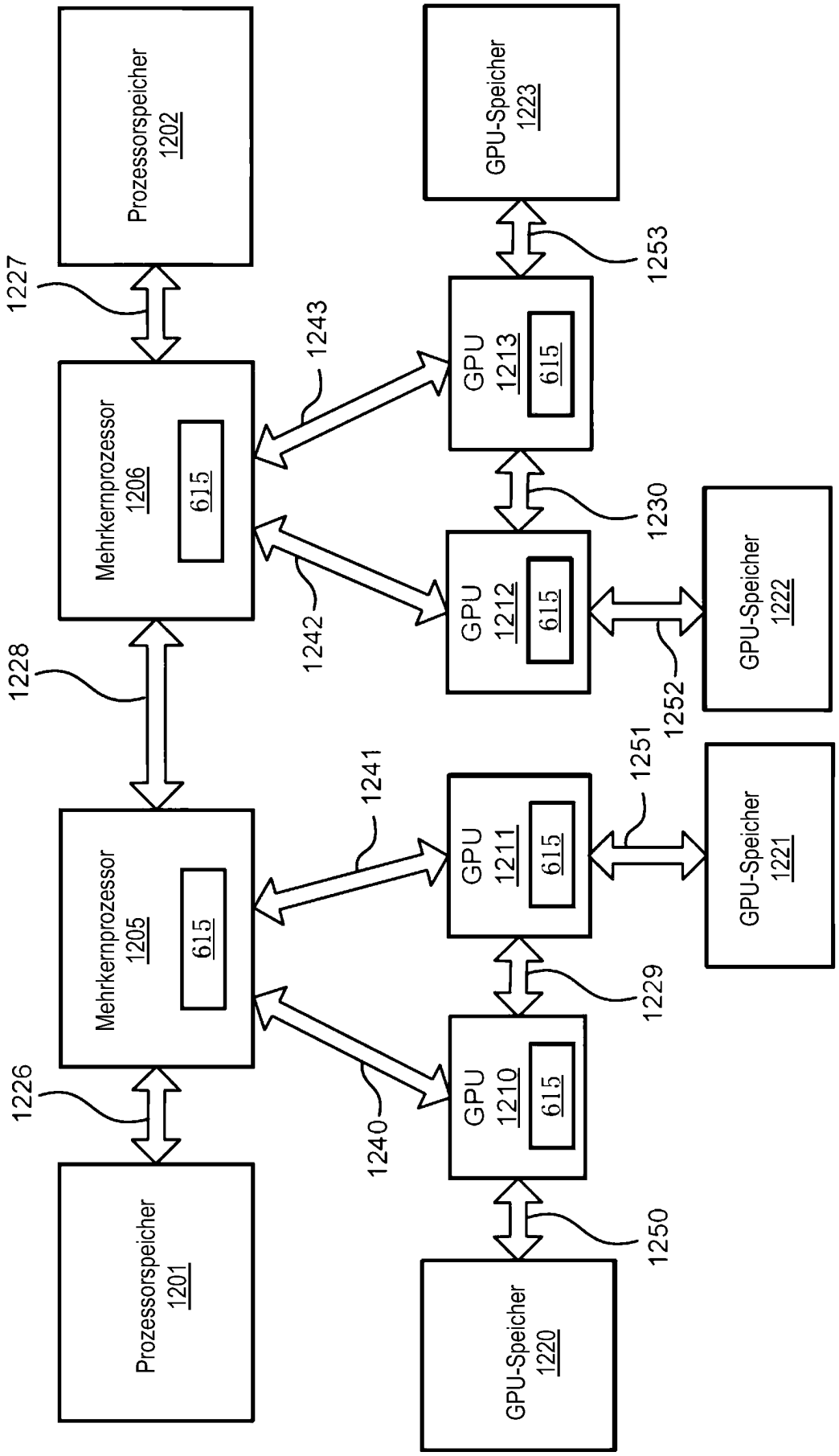

Prozessorspeicher
1202
GPU-Speicher
1223
1227
1253
1243
Mehrkernprozessor
1206
615
GPU
1213
615
1230
1228
1242
GPU
1212
615
GPU-Speicher
1222
1252
1241
1251
Mehrkernprozessor
1205
615
GPU
1211
615
GPU-Speicher
1221
1229
1226
1240
GPU
1210
615
1250
Prozessorspeicher
1201
GPU-Speicher
1220

FIG. 12A

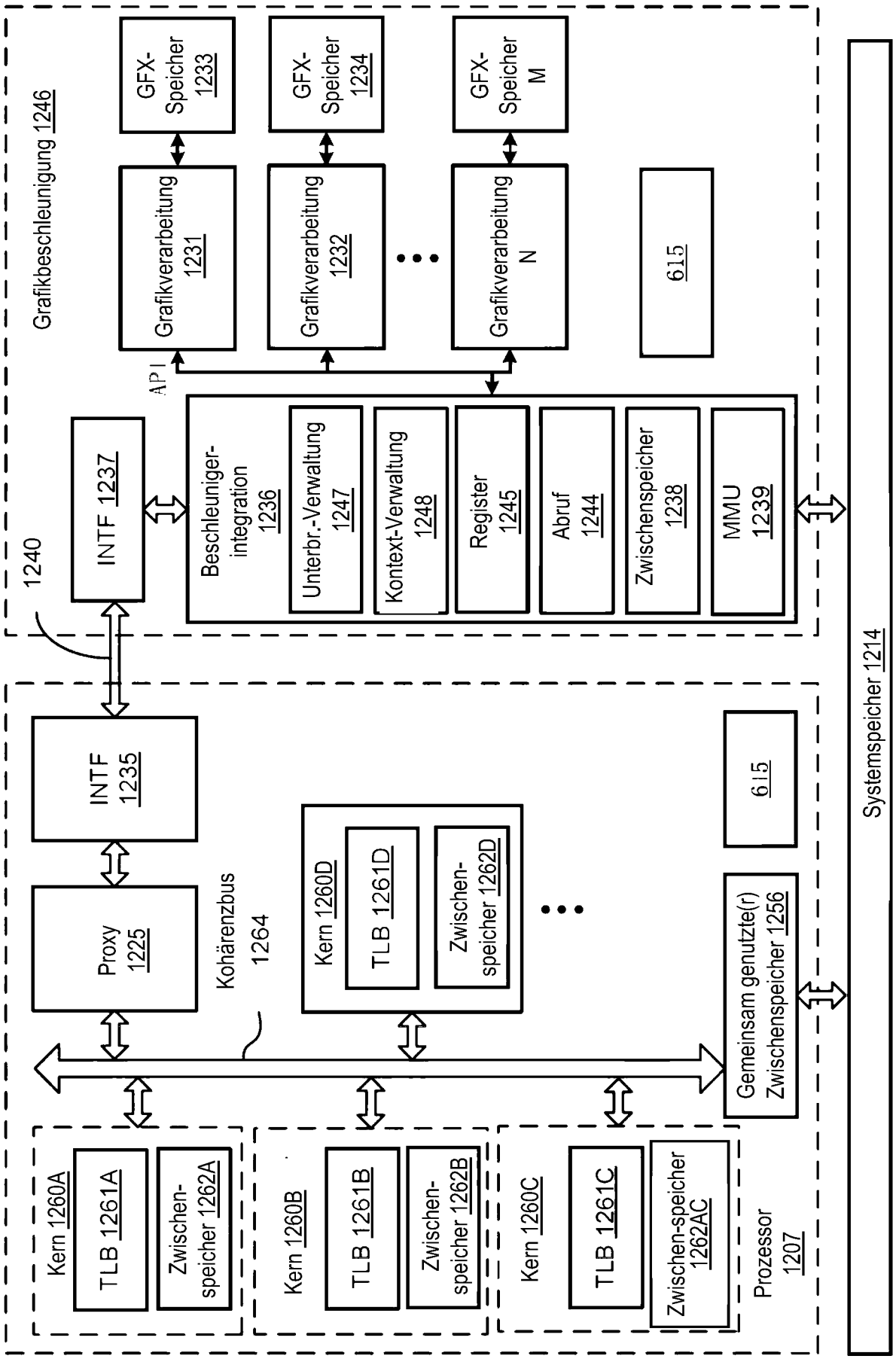

Grafikbeschleunigung 1246
GFX-Speicher 1233
GFX-Speicher 1234
GFX-Speicher M
Grafikverarbeitung 1231
Grafikverarbeitung 1232
Grafikverarbeitung N
615
API
1240
INTF 1237
Beschleuniger-integration 1236
Unterbr.-Verwaltung 1247
Kontext-Verwaltung 1248
Register 1245
Abruf 1244
Zwischenspeicher 1238
MMU 1239
INTF 1235
Proxy 1225
Kohärenzbus 1264
Kern 1260D
TLB 1261D
Zwischen-speicher 1262D
615
Gemeinsam genutzte(r) Zwischenspeicher 1256
Systemspeicher 1214
Kern 1260A
TLB 1261A
Zwischen-speicher 1262A
Kern 1260B
TLB 1261B
Zwischen-speicher 1262B
Kern 1260C
TLB 1261C
Zwischen-speicher 1262AC
Prozessor 1207

FIG. 12B

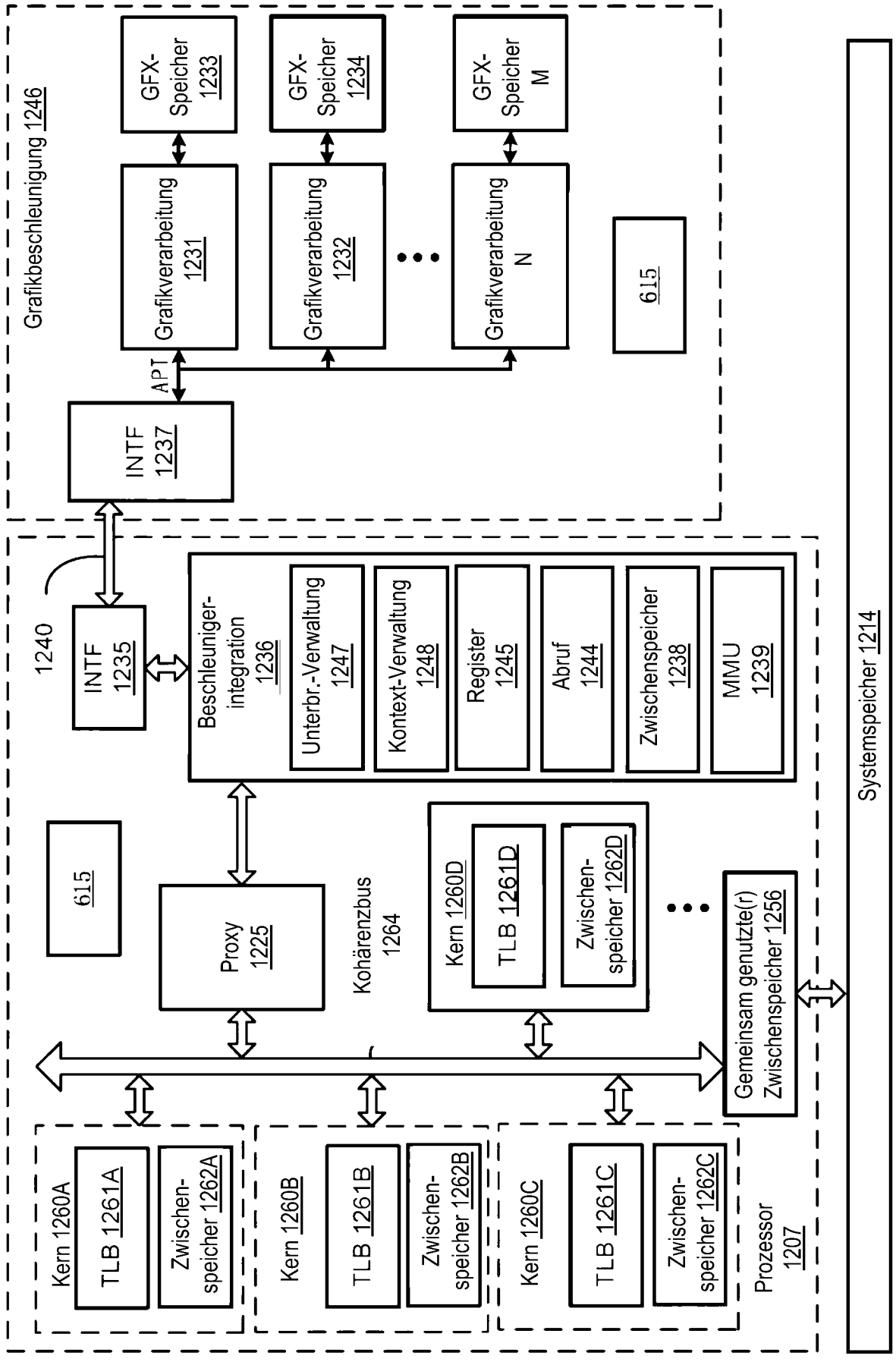

Grafikbeschleunigung 1246
GFX-Speicher 1233
GFX-Speicher 1234
GFX-Speicher M
Grafikverarbeitung 1231
Grafikverarbeitung 1232
Grafikverarbeitung N
615
APT
INTF 1237
1240
INTF 1235
Beschleuniger-integration 1236
Unterbr.-Verwaltung 1247
Kontext-Verwaltung 1248
Register 1245
Abruf 1244
Zwischenspeicher 1238
MMU 1239
615
Proxy 1225
Kohärenzbus 1264
Kern 1260D
TLB 1261D
Zwischen-speicher 1262D
Gemeinsam genutzte(r) Zwischenspeicher 1256
Systemspeicher 1214
Kern 1260A
TLB 1261A
Zwischen-speicher 1262A
Kern 1260B
TLB 1261B
Zwischen-speicher 1262B
Kern 1260C
TLB 1261C
Zwischen-speicher 1262C
Prozessor 1207

FIG. 12C

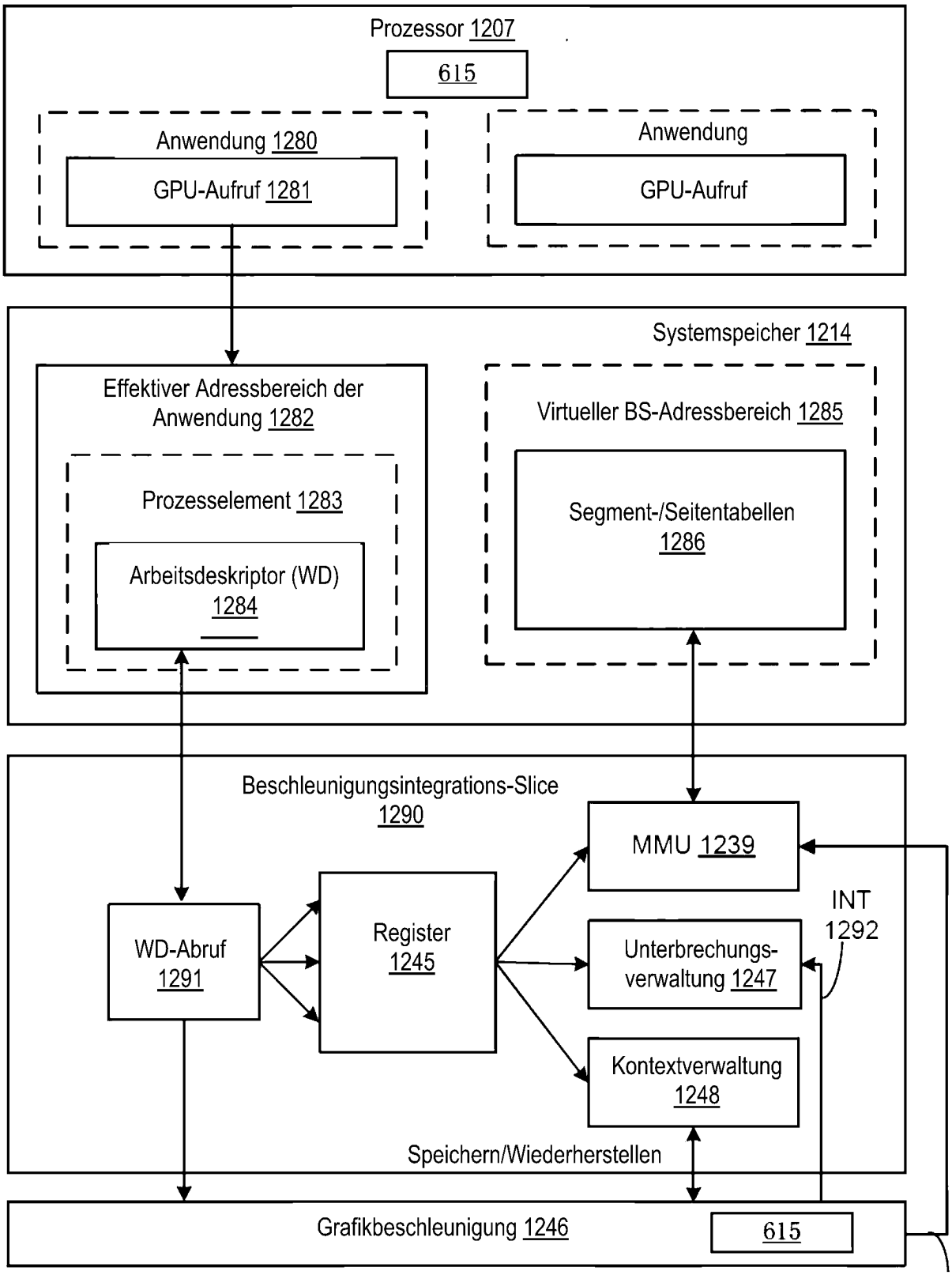

Prozessor 1207
615
Anwendung 1280
GPU-Aufruf 1281
Anwendung
GPU-Aufruf
Systemspeicher 1214
Effektiver Adressbereich der Anwendung 1282
Prozesselement 1283
Arbeitsdeskriptor (WD) 1284
Virtueller BS-Adressbereich 1285
Segment-/Seitentabellen 1286
Beschleunigungsintegrations-Slice 1290
MMU 1239
INT 1292
WD-Abruf 1291
Register 1245
Unterbrechungs-verwaltung 1247
Kontextverwaltung 1248
Speichern/Wiederherstellen
Grafikbeschleunigung 1246
615
Effektive Adresse 1293

FIG. 12D

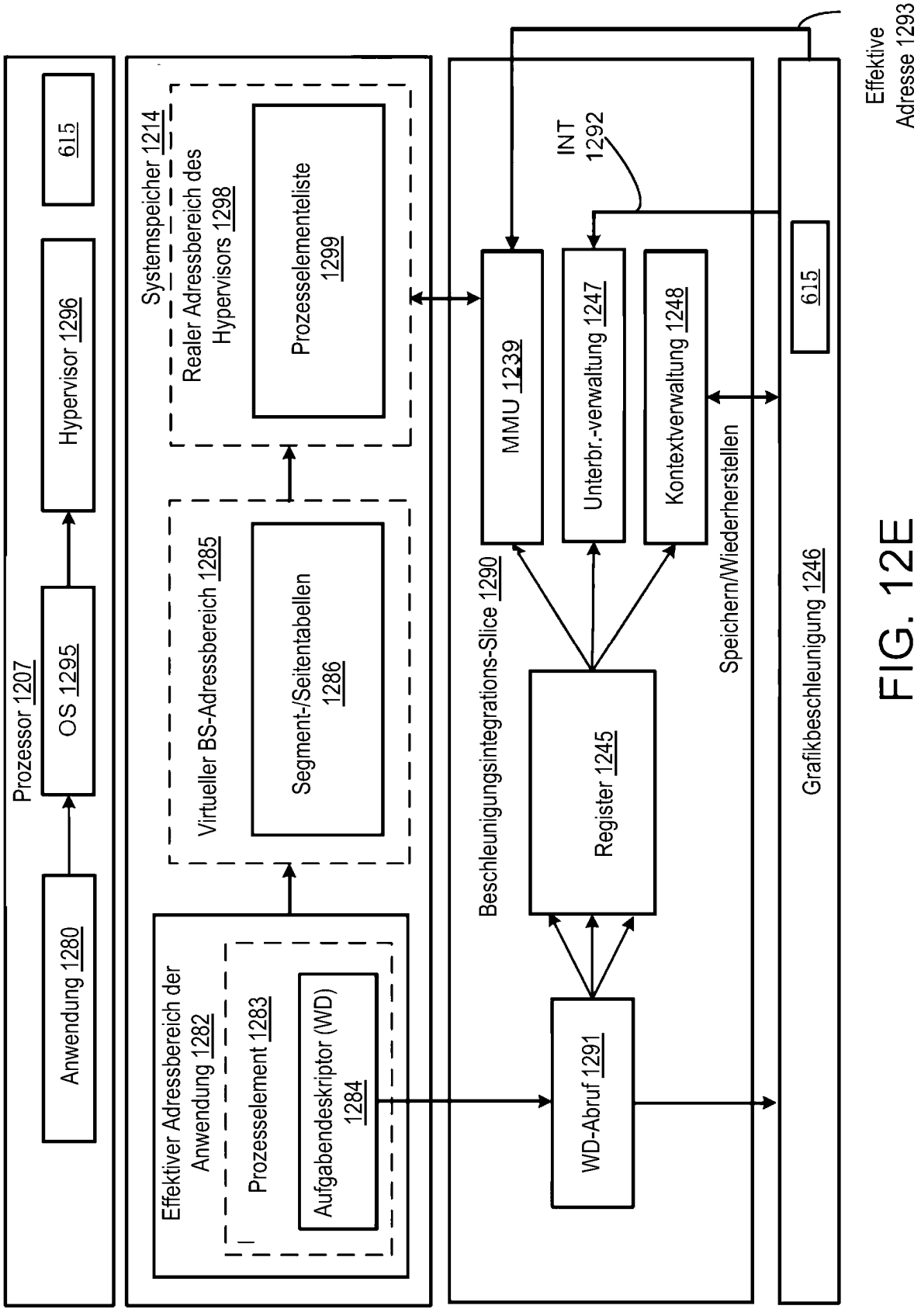

Prozessor 1207
Anwendung 1280
OS 1295
Hypervisor 1296
615
Systemspeicher 1214
Realer Adressbereich des Hypervisors 1298
Prozesselementeliste 1299
Virtueller BS-Adressbereich 1285
Segment-/Seitentabellen 1286
Effektiver Adressbereich der Anwendung 1282
Prozesselement 1283
Aufgabendeskriptor (WD) 1284
Beschleunigungsintegrations-Slice 1290
MMU 1239
Unterbr.-verwaltung 1247
Kontextverwaltung 1248
Register 1245
WD-Abruf 1291
INT 1292
Speichern/Wiederherstellen
Grafikbeschleunigung 1246
615
Effektive Adresse 1293

FIG. 12E

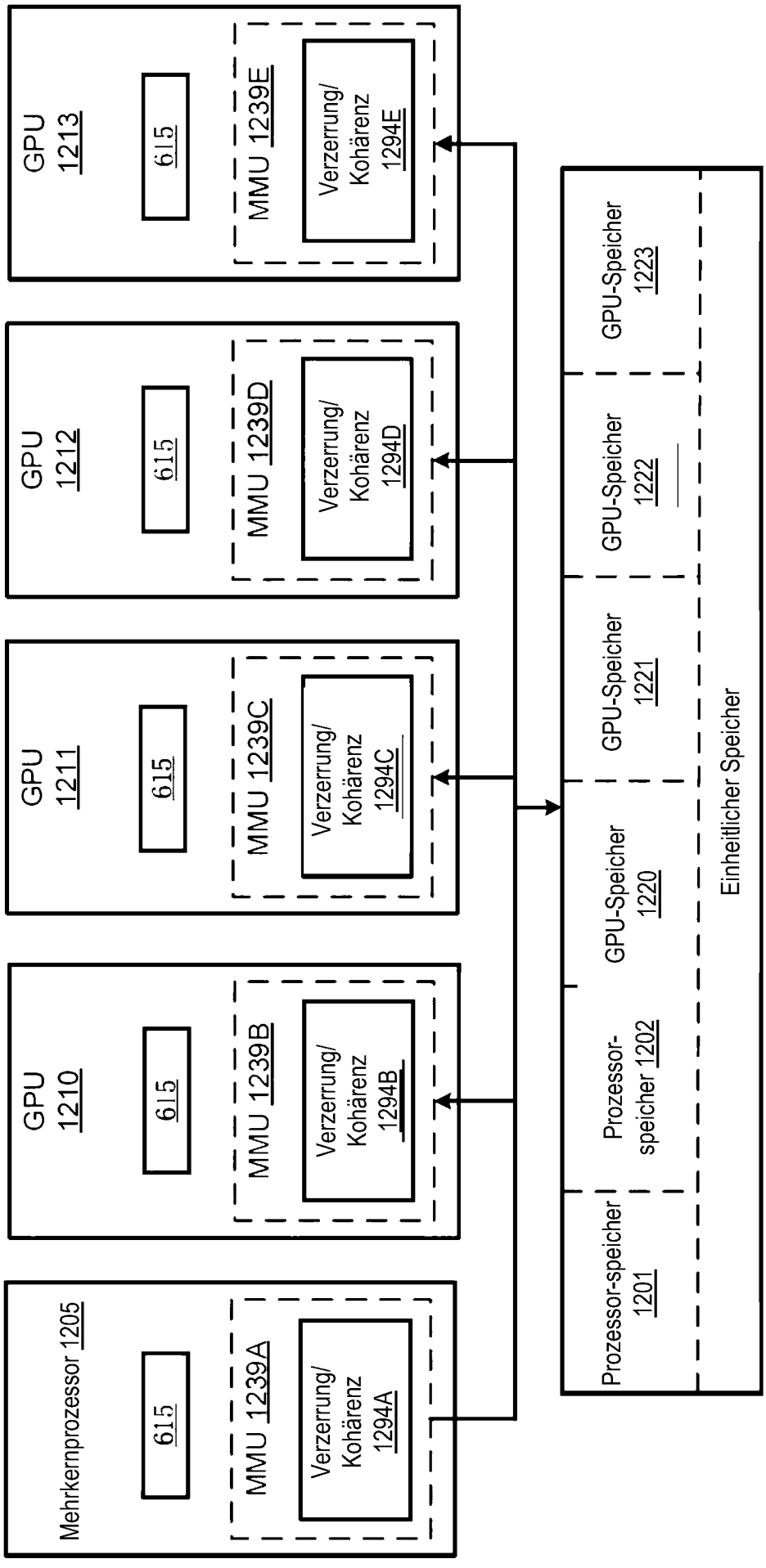

Mehrkernprozessor 1205
615
MMU 1239A
Verzerrung/ Kohärenz 1294A
GPU 1210
615
MMU 1239B
Verzerrung/ Kohärenz 1294B
GPU 1211
615
MMU 1239C
Verzerrung/ Kohärenz 1294C
GPU 1212
615
MMU 1239D
Verzerrung/ Kohärenz 1294D
GPU 1213
615
MMU 1239E
Verzerrung/ Kohärenz 1294E
Prozessor-speicher 1201
Prozessor-speicher 1202
GPU-Speicher 1220
GPU-Speicher 1221
GPU-Speicher 1222
GPU-Speicher 1223
Einheitlicher Speicher

FIG. 12F

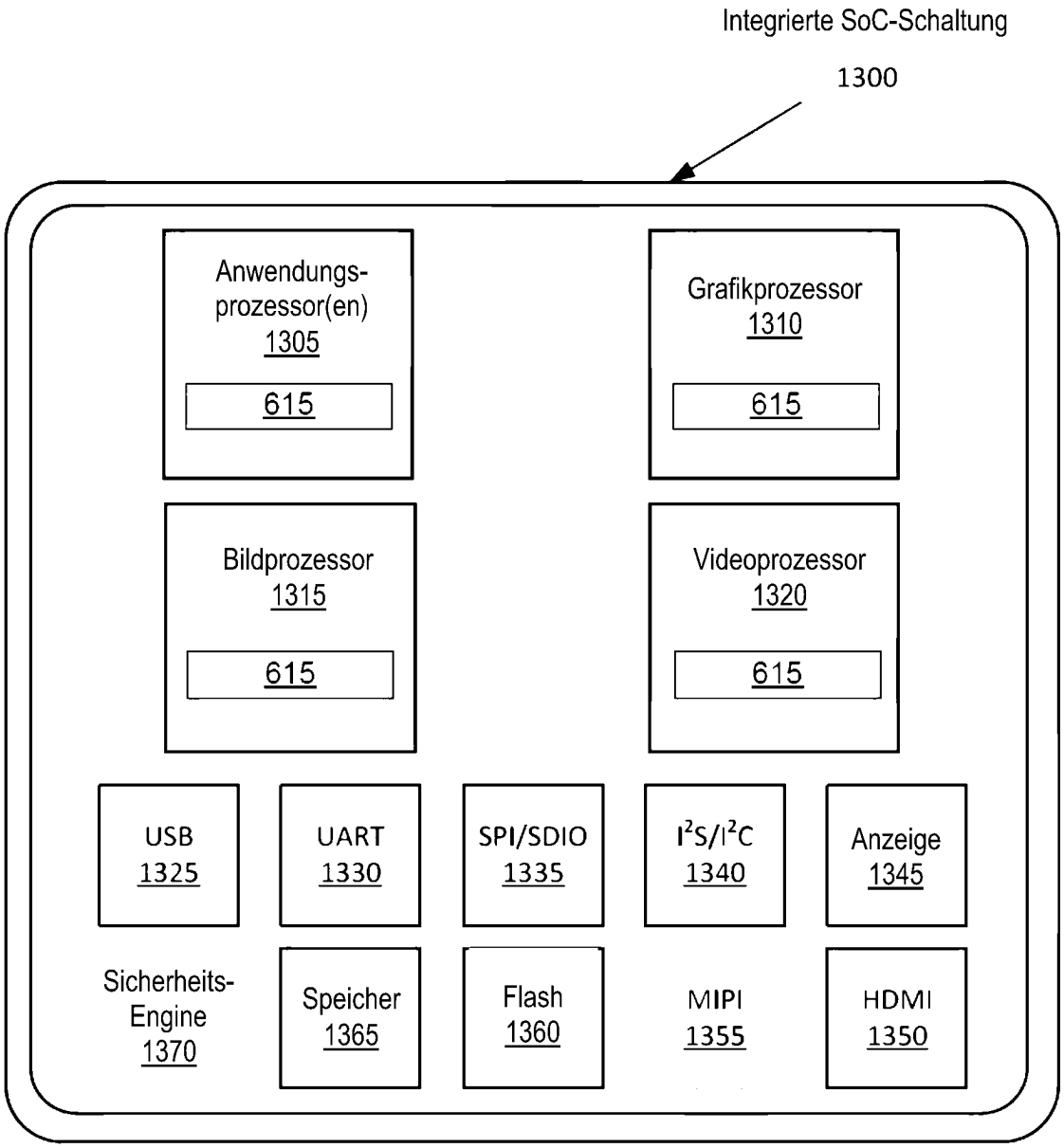

Integrierte SoC-Schaltung
1300
Anwendungs-prozessor(en) 1305
615
Grafikprozessor 1310
615
Bildprozessor 1315
615
Videoprozessor 1320
615
USB 1325
UART 1330
SPI/SDIO 1335
I²S/I²C 1340
Anzeige 1345
Sicherheits-Engine 1370
Speicher 1365
Flash 1360
MIPI 1355
HDMI 1350

FIG. 13

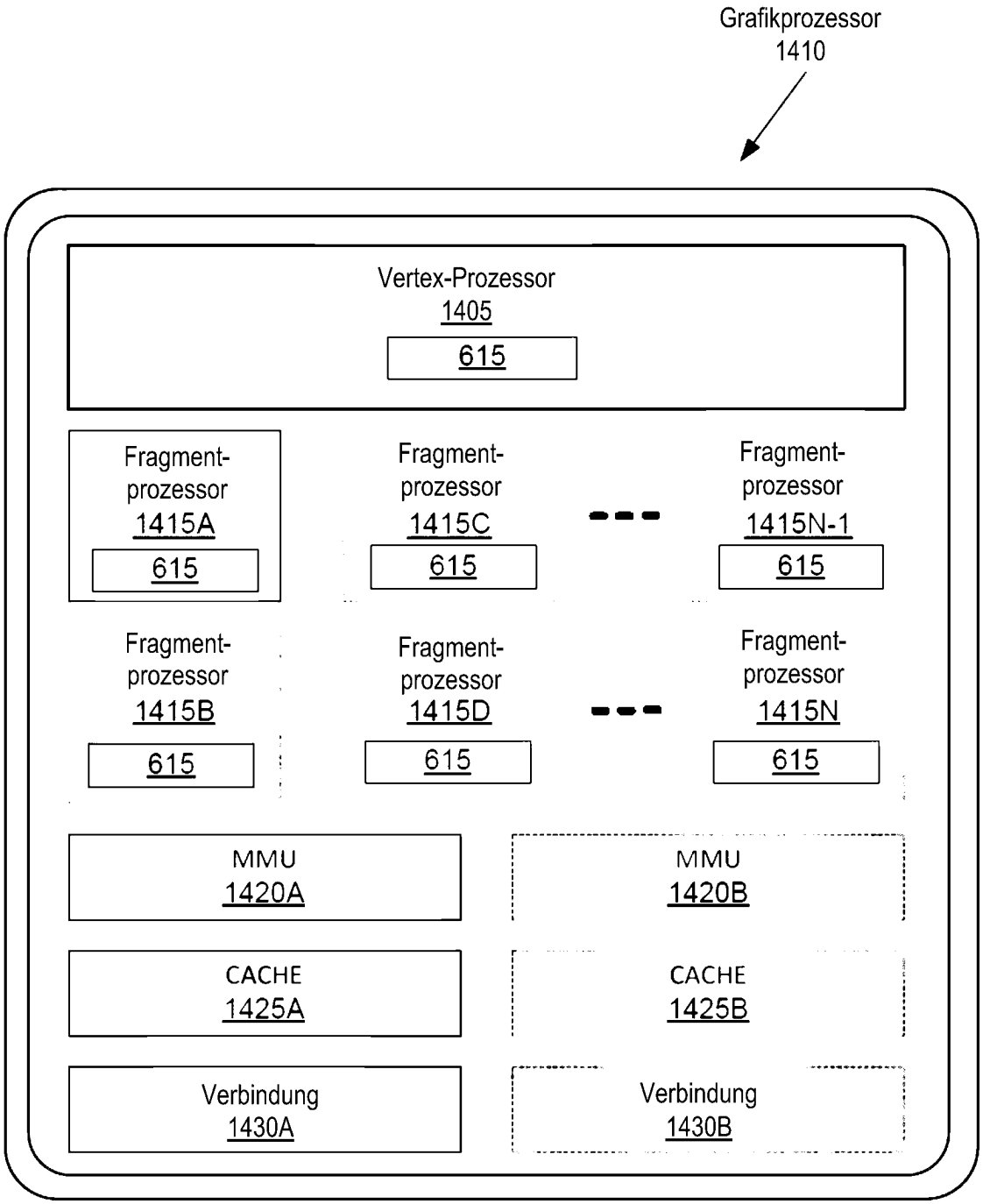

Grafikprozessor
1410
Vertex-Prozessor
1405
615
Fragment-
prozessor
1415A
615
Fragment-
prozessor
1415C
615
Fragment-
prozessor
1415N-1
615
Fragment-
prozessor
1415B
615
Fragment-
prozessor
1415D
615
Fragment-
prozessor
1415N
615
MMU
1420A
MMU
1420B
CACHE
1425A
CACHE
1425B
Verbindung
1430A
Verbindung
1430B

FIG. 14A

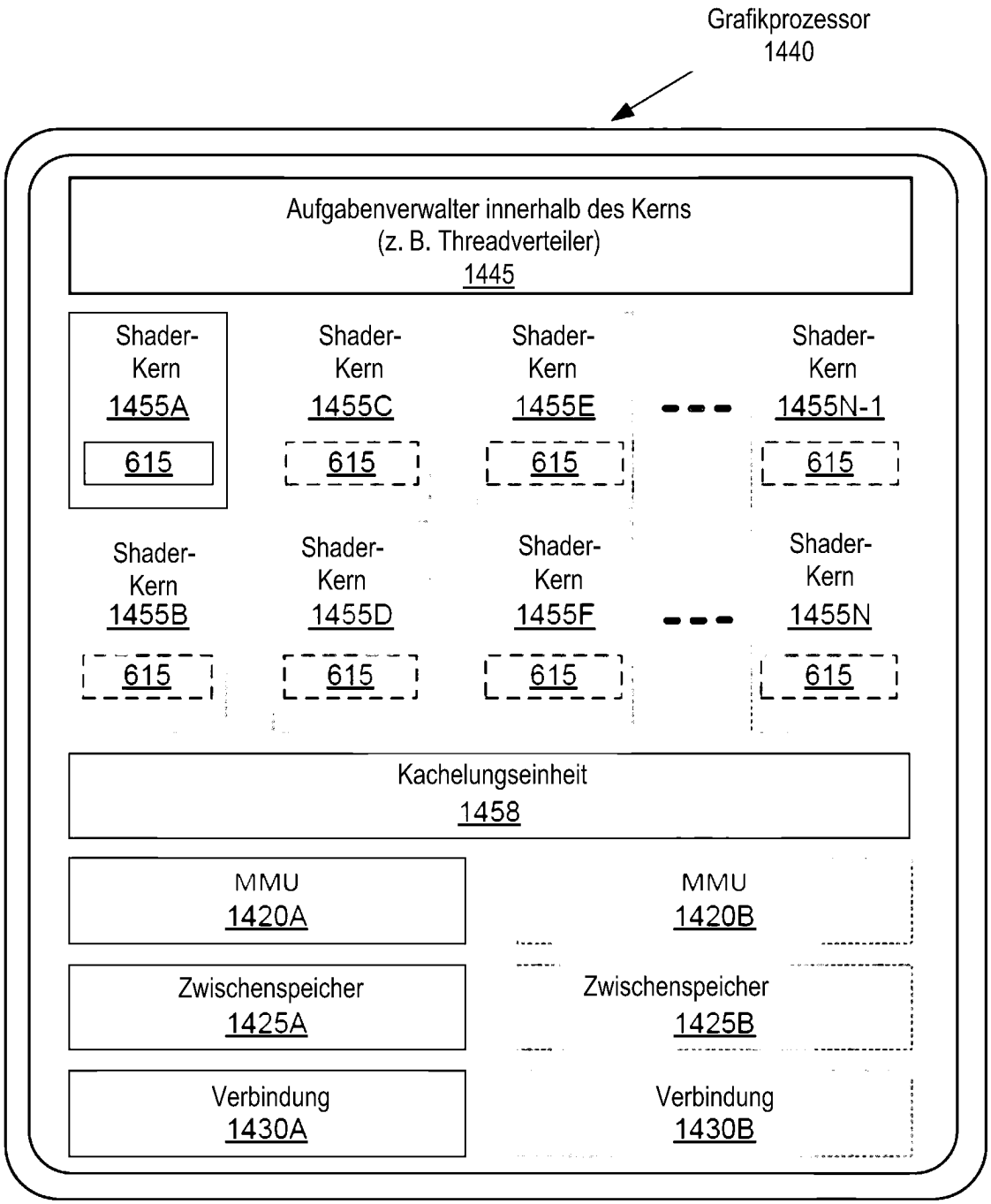

Grafikprozessor
1440
Aufgabenverwalter innerhalb des Kerns
(z. B. Threadverteiler)
1445
Shader-Kern 1455A
615
Shader-Kern 1455C
615
Shader-Kern 1455E
615
Shader-Kern 1455N-1
615
Shader-Kern 1455B
615
Shader-Kern 1455D
615
Shader-Kern 1455F
615
Shader-Kern 1455N
615
Kachelungseinheit
1458
MMU 1420A
MMU 1420B
Zwischenspeicher 1425A
Zwischenspeicher 1425B
Verbindung 1430A
Verbindung 1430B

FIG. 14B

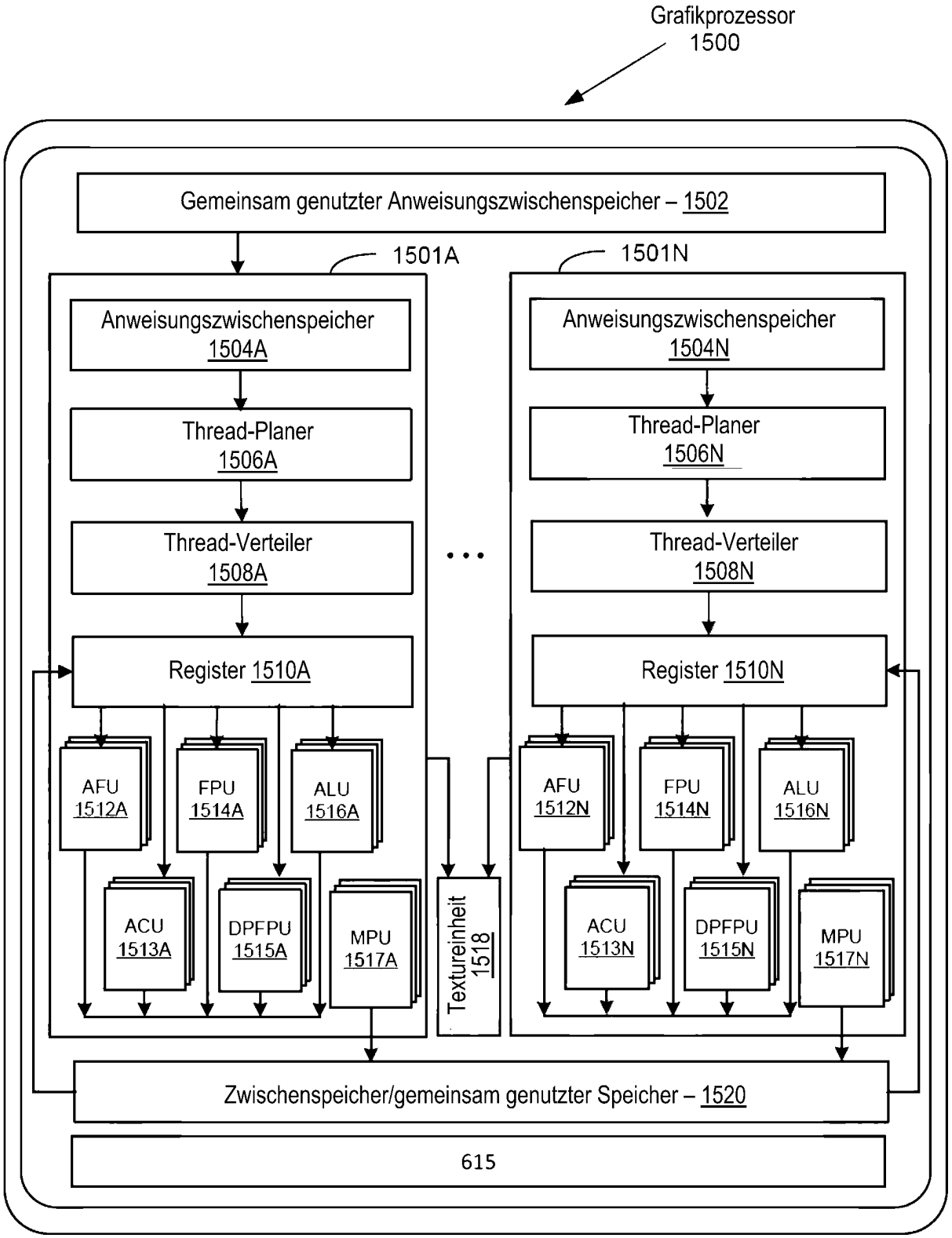

Grafikprozessor
1500
Gemeinsam genutzter Anweisungszwischenspeicher – 1502
1501A
1501N
Anweisungszwischenspeicher
1504A
Thread-Planer
1506A
Thread-Verteiler
1508A
Register 1510A
AFU
1512A
FPU
1514A
ALU
1516A
ACU
1513A
DPFPU
1515A
MPU
1517A
Anweisungszwischenspeicher
1504N
Thread-Planer
1506N
Thread-Verteiler
1508N
Register 1510N
AFU
1512N
FPU
1514N
ALU
1516N
ACU
1513N
DPFPU
1515N
MPU
1517N
Textureinheit
1518
Zwischenspeicher/gemeinsam genutzter Speicher – 1520
615

FIG. 15A

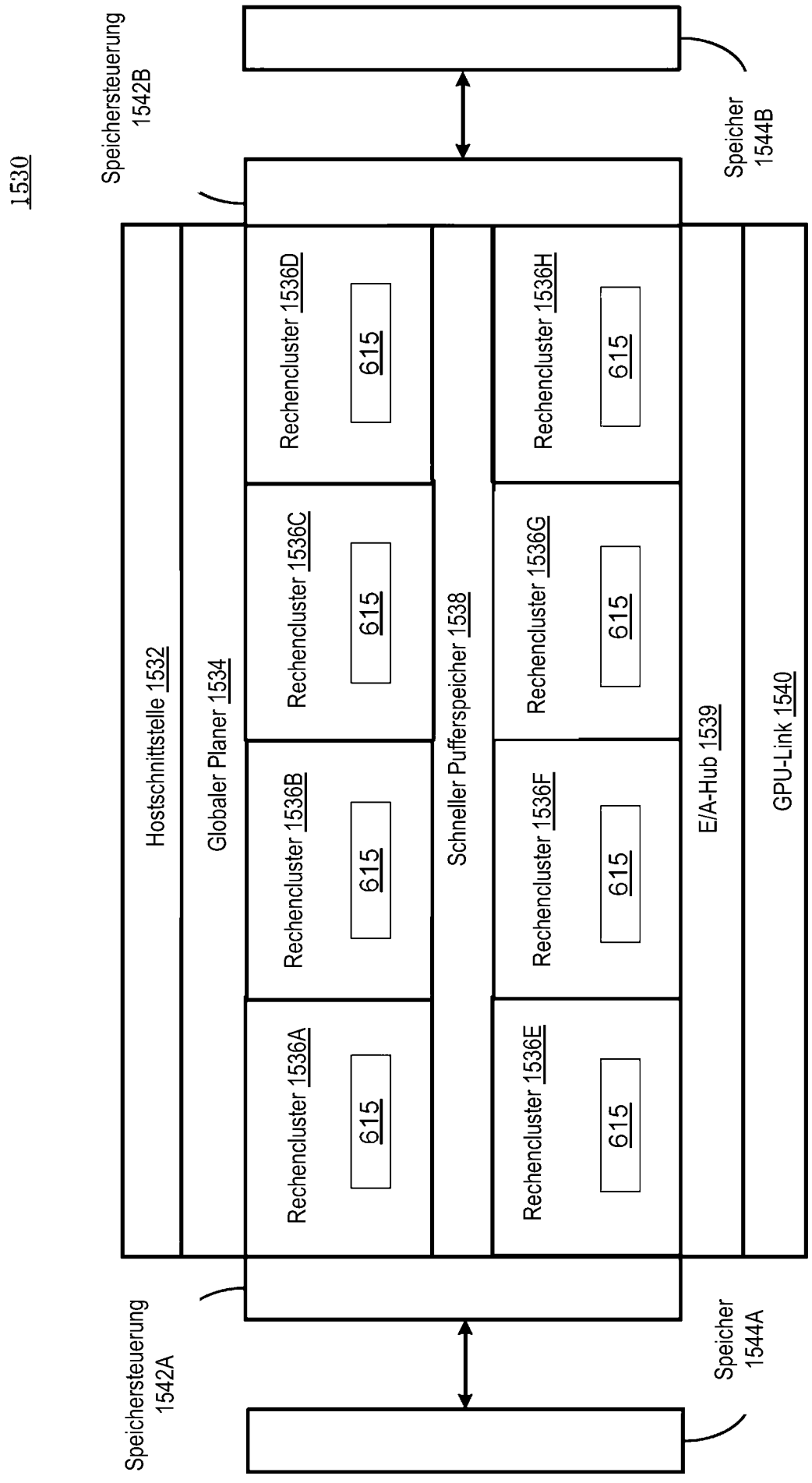

1530
Speichersteuerung 1542B
Speicher 1544B
Hostschnittstelle 1532
Globaler Planer 1534
Rechencluster 1536A
Rechencluster 1536B
Rechencluster 1536C
Rechencluster 1536D
Rechencluster 1536E
Rechencluster 1536F
Rechencluster 1536G
Rechencluster 1536H
615
Schneller Pufferspeicher 1538
E/A-Hub 1539
GPU-Link 1540
Speichersteuerung 1542A
Speicher 1544A

FIG. 15B

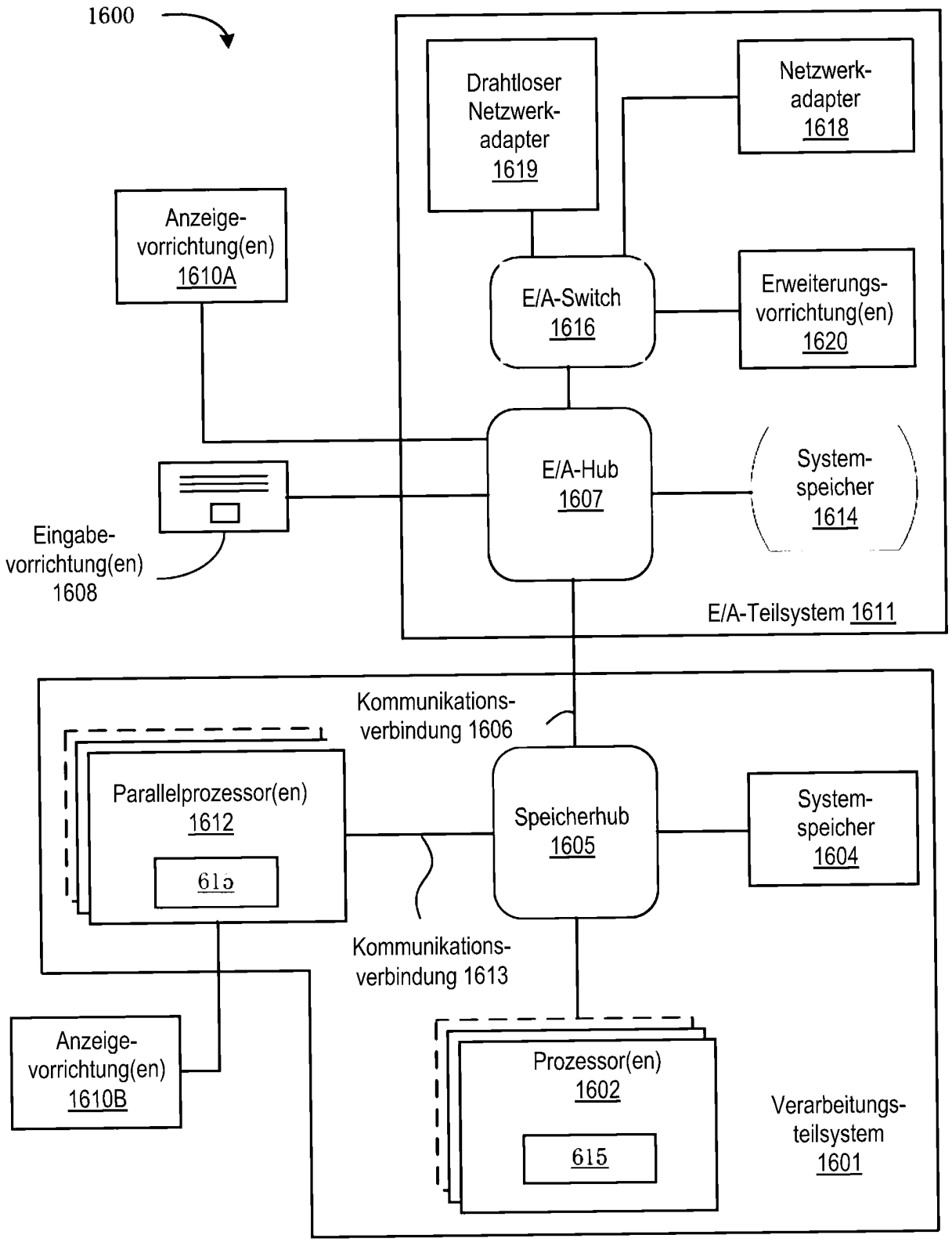

1600
Drahtloser Netzwerk-adapter 1619
Netzwerk-adapter 1618
Anzeige-vorrichtung(en) 1610A
E/A-Switch 1616
Erweiterungs-vorrichtung(en) 1620
E/A-Hub 1607
System-speicher 1614
Eingabe-vorrichtung(en) 1608
E/A-Teilsystem 1611
Kommunikations-verbindung 1606
Parallelprozessor(en) 1612
615
Speicherhub 1605
System-speicher 1604
Kommunikations-verbindung 1613
Anzeige-vorrichtung(en) 1610B
Prozessor(en) 1602
615
Verarbeitungs-teilsystem 1601

FIG. 16

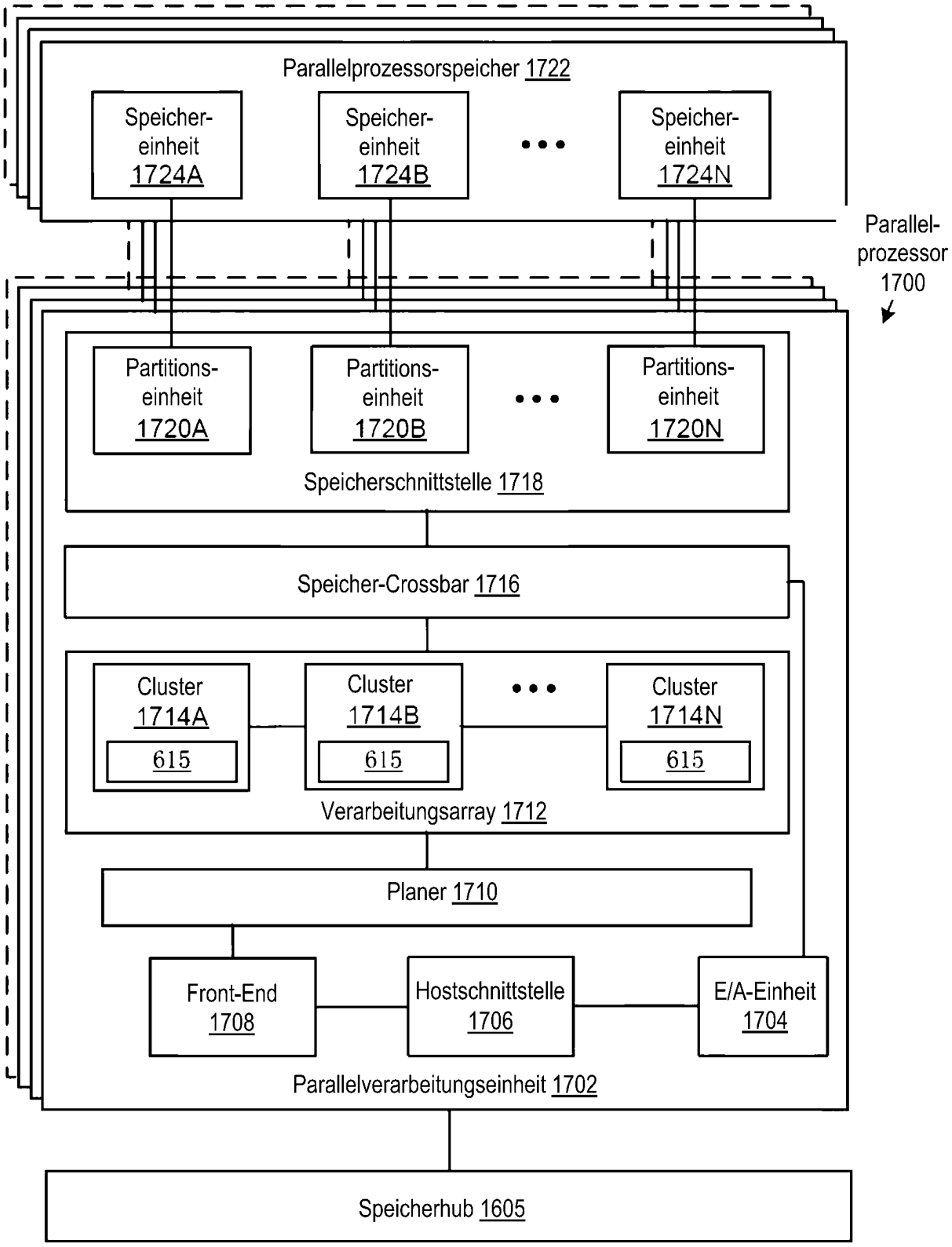

Parallelprozessorspeicher 1722
Speicher-einheit 1724A
Speicher-einheit 1724B
Speicher-einheit 1724N
Parallel-prozessor 1700
Partitions-einheit 1720A
Partitions-einheit 1720B
Partitions-einheit 1720N
Speicherschnittstelle 1718
Speicher-Crossbar 1716
Cluster 1714A
615
Cluster 1714B
615
Cluster 1714N
615
Verarbeitungsarray 1712
Planer 1710
Front-End 1708
Hostschnittstelle 1706
E/A-Einheit 1704
Parallelverarbeitungseinheit 1702
Speicherhub 1605

FIG. 17A

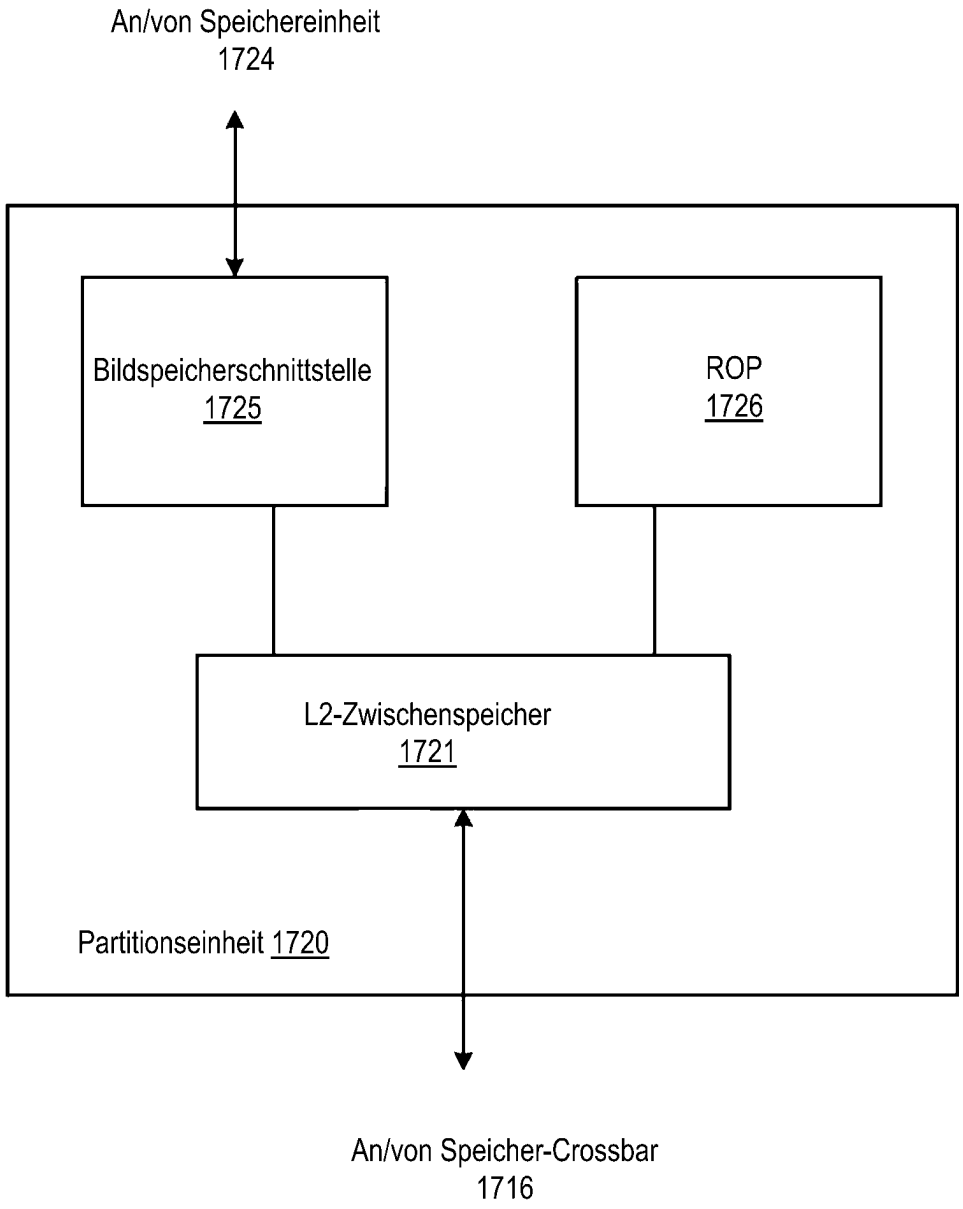

An/von Speichereinheit
1724
Bildspeicherschnittstelle
1725
ROP
1726
L2-Zwischenspeicher
1721
Partitionseinheit 1720
An/von Speicher-Crossbar
1716

FIG. 17B

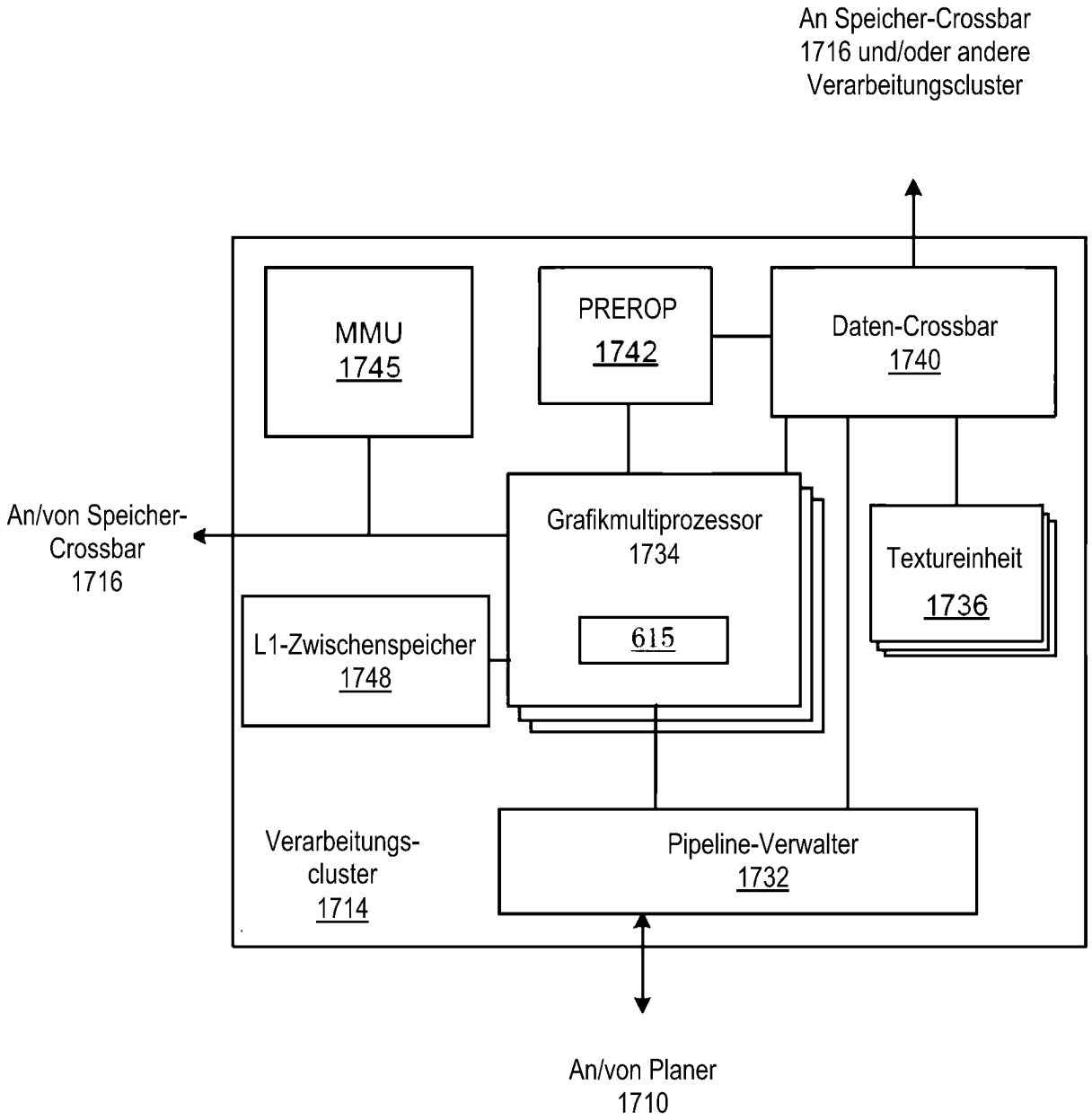

An Speicher-Crossbar
1716 und/oder andere
Verarbeitungscluster
MMU
1745
PREROP
1742
Daten-Crossbar
1740
An/von Speicher-
Crossbar
1716
Grafikmultiprozessor
1734
615
Textureinheit
1736
L1-Zwischenspeicher
1748
Verarbeitungs-
cluster
1714
Pipeline-Verwalter
1732
An/von Planer
1710

FIG. 17C

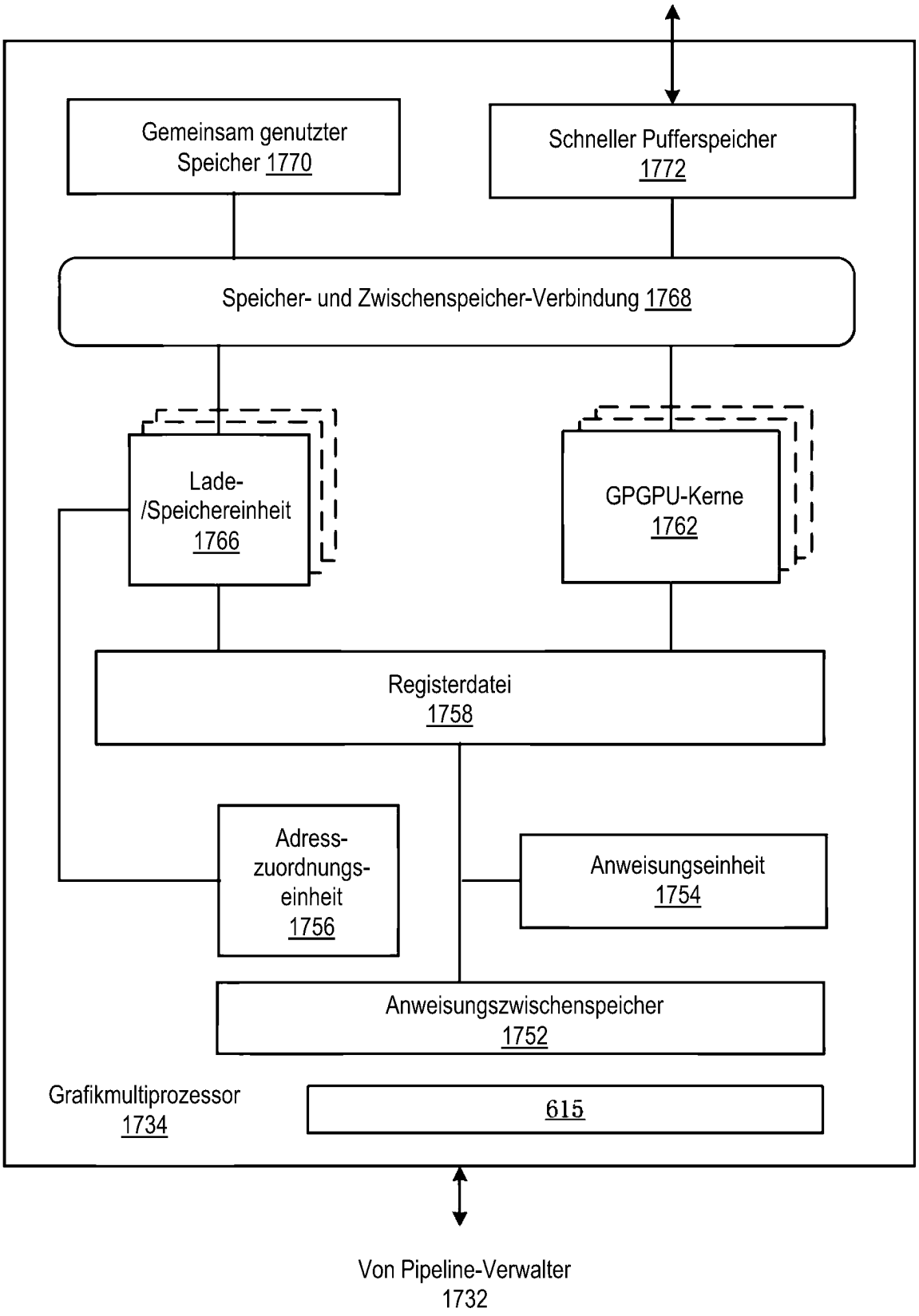

Gemeinsam genutzter Speicher 1770
Schneller Pufferspeicher 1772
Speicher- und Zwischenspeicher-Verbindung 1768
Lade-/Speichereinheit 1766
GPGPU-Kerne 1762
Registerdatei 1758
Adress-zuordnungs-einheit 1756
Anweisungseinheit 1754
Anweisungszwischenspeicher 1752
Grafikmultiprozessor 1734
615
Von Pipeline-Verwalter 1732

FIG. 17D

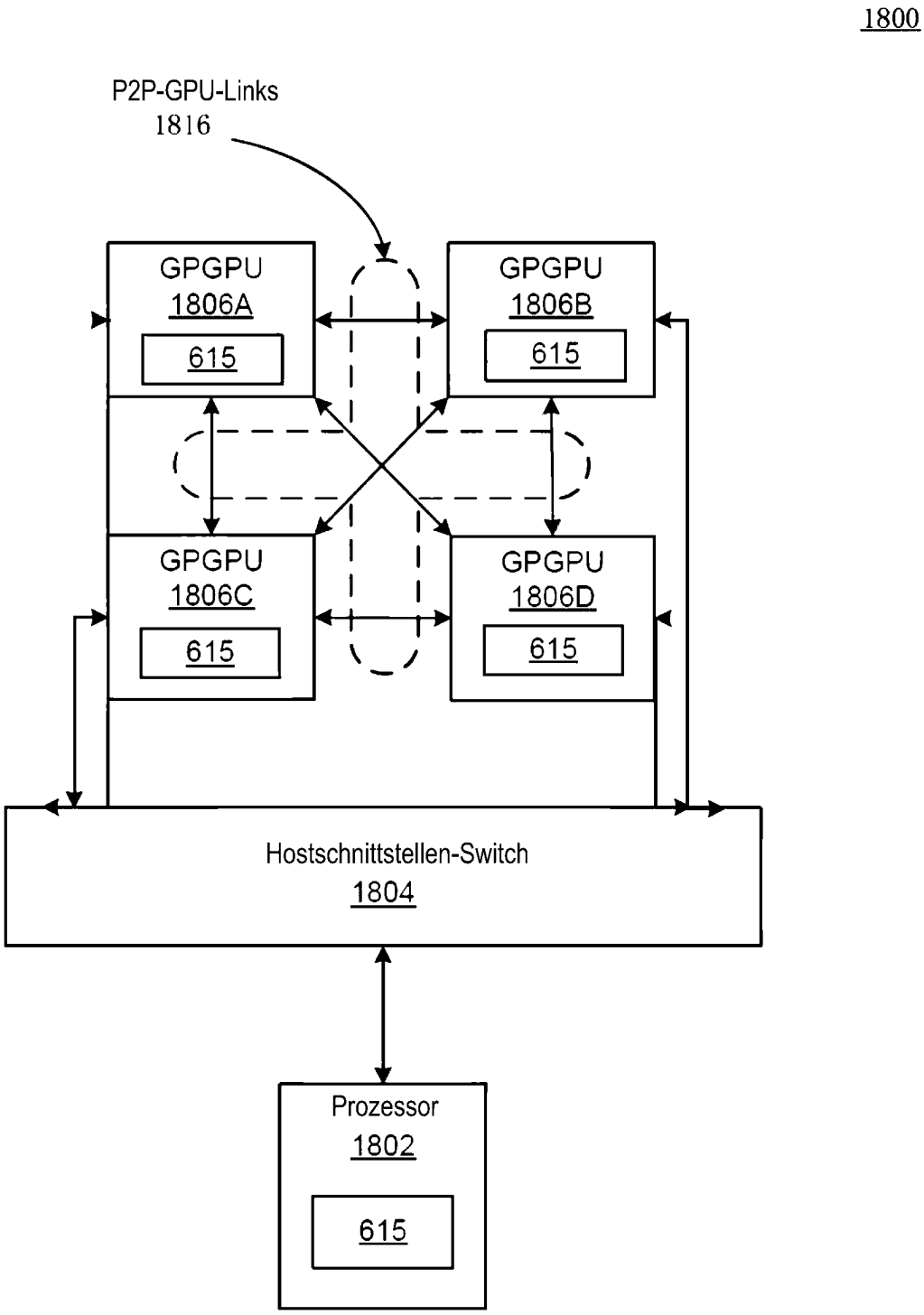

1800
P2P-GPU-Links
1816
GPGPU
1806A
615
GPGPU
1806B
615
GPGPU
1806C
615
GPGPU
1806D
615
Hostschnittstellen-Switch
1804
Prozessor
1802
615

FIG. 18

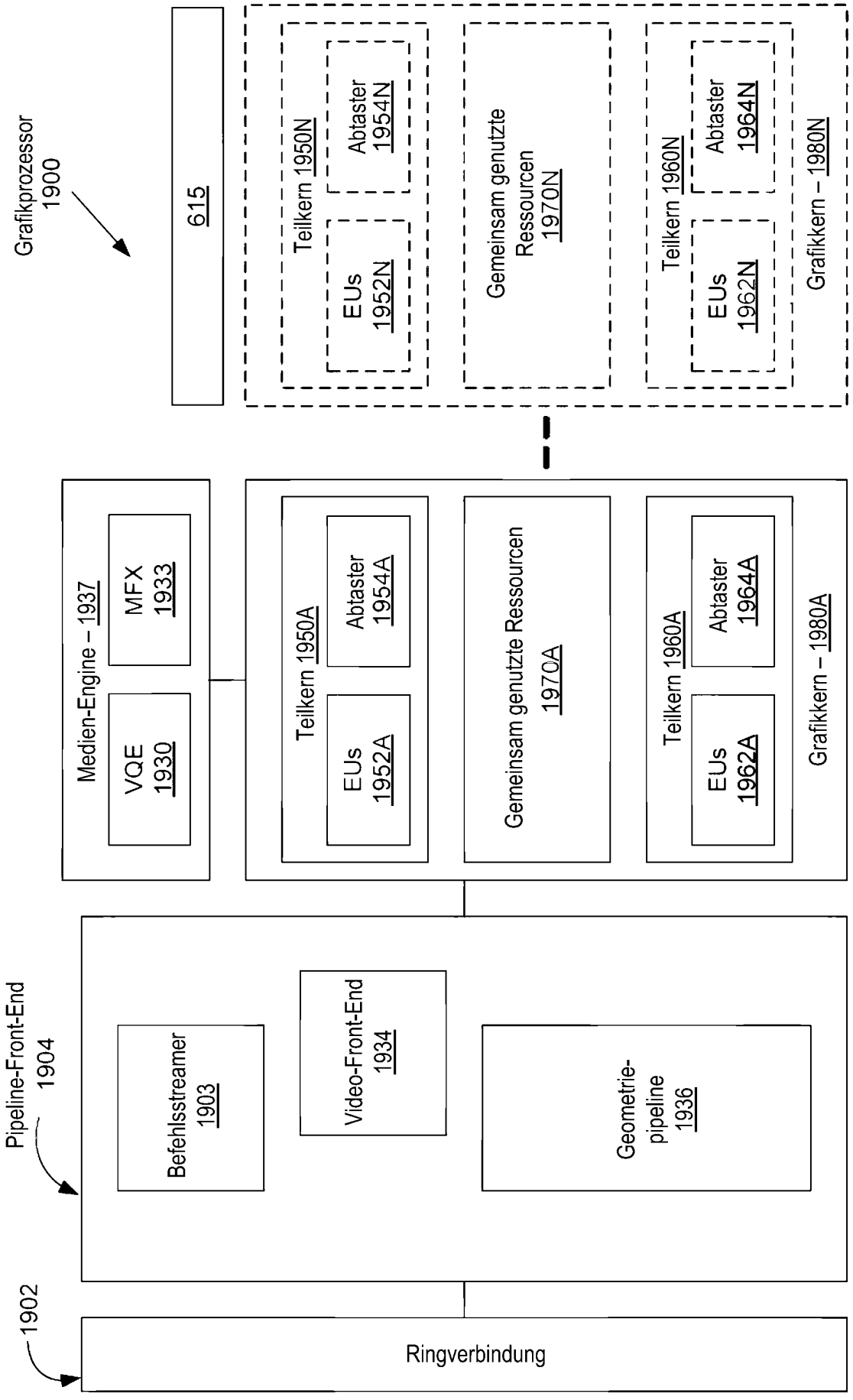

Grafikprozessor 1900
615
Teilkern 1950N
EUs 1952N
Abtaster 1954N
Gemeinsam genutzte Ressourcen 1970N
Teilkern 1960N
EUs 1962N
Abtaster 1964N
Grafikkern – 1980N
Medien-Engine – 1937
VQE 1930
MFX 1933
Teilkern 1950A
EUs 1952A
Abtaster 1954A
Gemeinsam genutzte Ressourcen 1970A
Teilkern 1960A
EUs 1962A
Abtaster 1964A
Grafikkern – 1980A
Pipeline-Front-End 1904
Befehlsstreamer 1903
Video-Front-End 1934
Geometrie-pipeline 1936
1902
Ringverbindung

FIG. 19

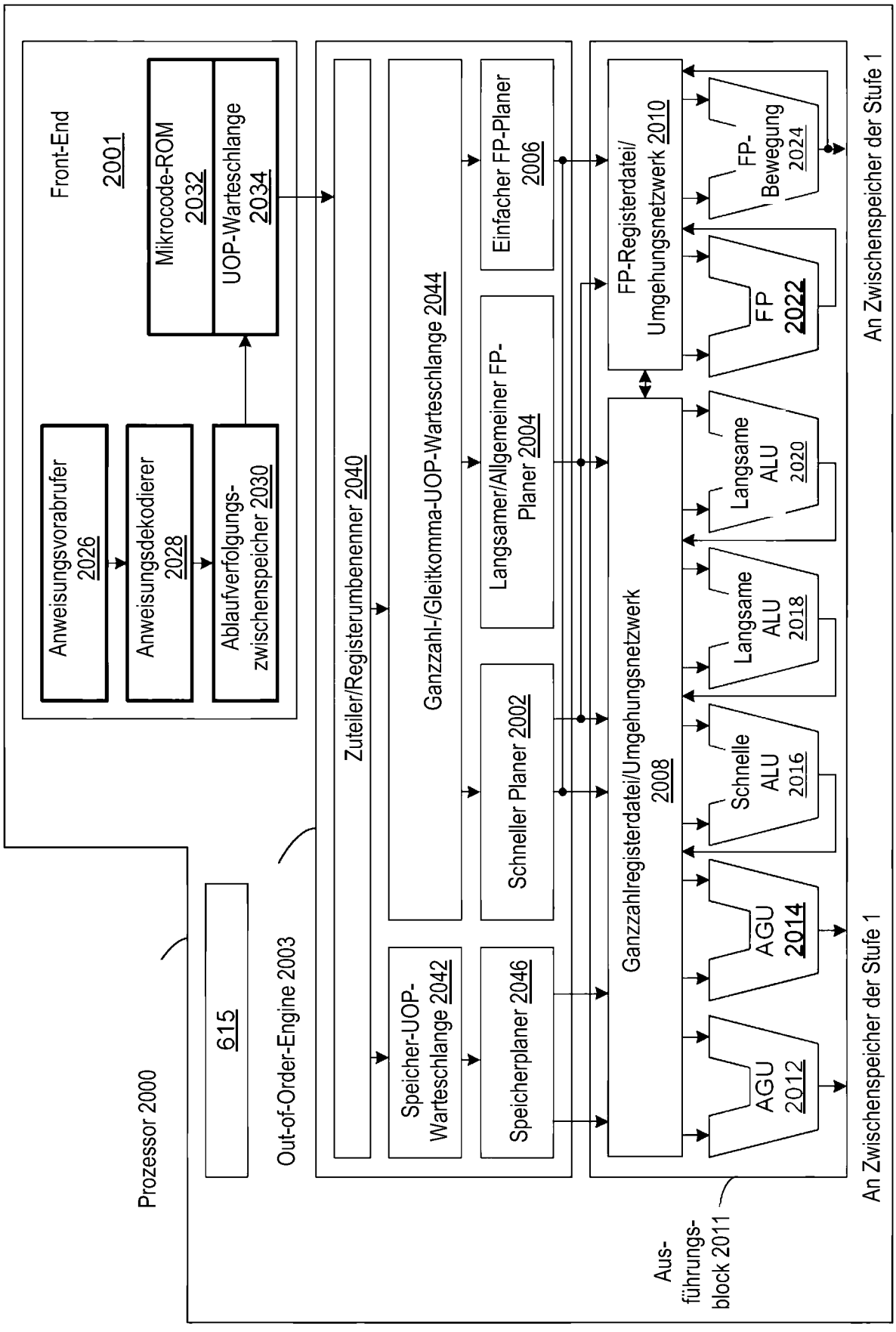

Prozessor 2000
615
Front-End
2001
Anweisungsvorabrufer 2026
Anweisungsdekodierer 2028
Ablaufverfolgungs-zwischenspeicher 2030
Mikrocode-ROM 2032
UOP-Warteschlange 2034
Out-of-Order-Engine 2003
Zuteiler/Registerumbenenner 2040
Speicher-UOP-Warteschlange 2042
Speicherplaner 2046
Ganzzahl-/Gleitkomma-UOP-Warteschlange 2044
Schneller Planer 2002
Langsamer/Allgemeiner FP-Planer 2004
Einfacher FP-Planer 2006
Ganzzahlregisterdatei/Umgehungsnetzwerk 2008
FP-Registerdatei/Umgehungsnetzwerk 2010
AGU 2012
AGU 2014
Schnelle ALU 2016
Langsame ALU 2018
Langsame ALU 2020
FP 2022
FP-Bewegung 2024
An Zwischenspeicher der Stufe 1
An Zwischenspeicher der Stufe 1
Aus-führungs-block 2011

FIG. 20

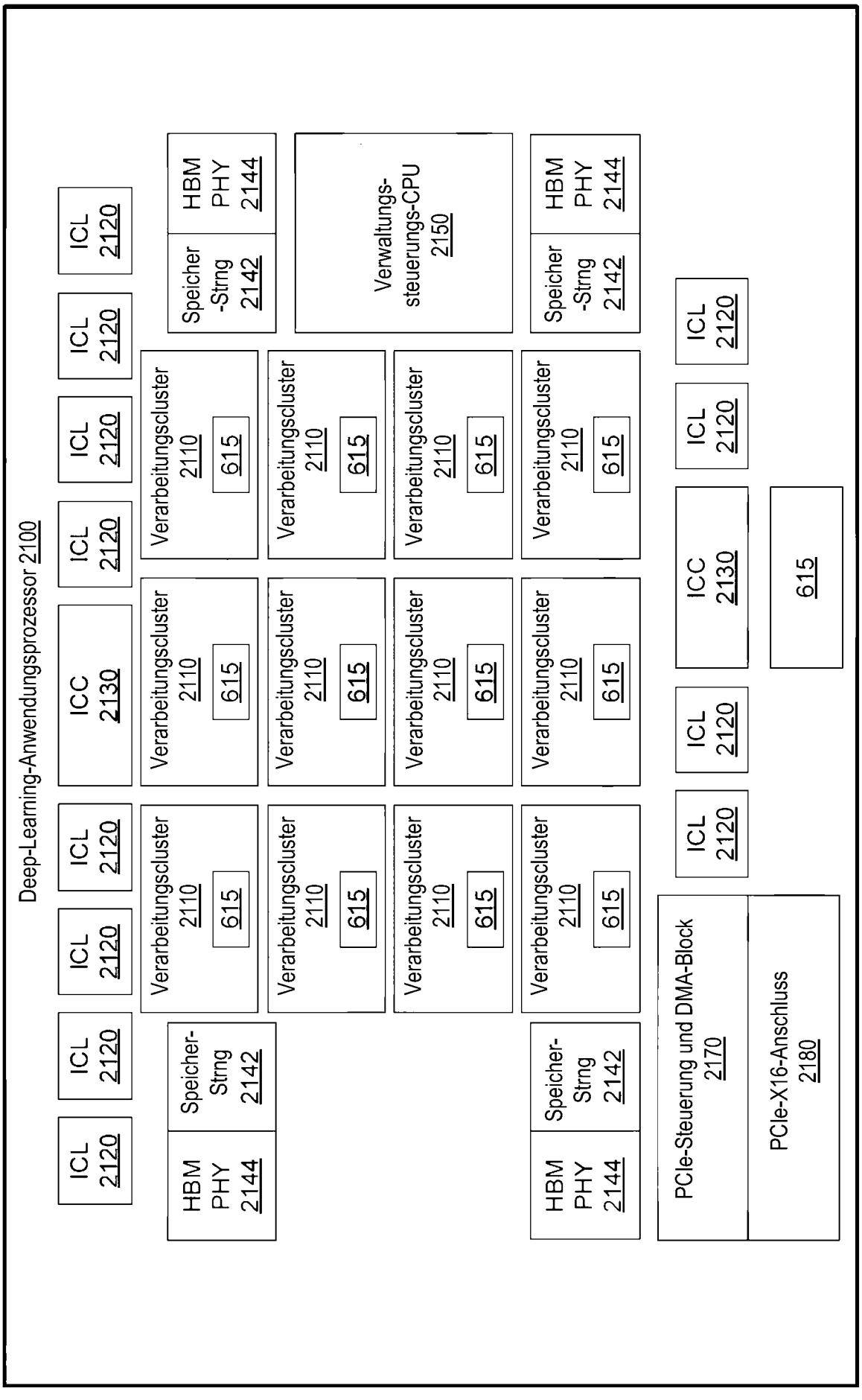

Deep-Learning-Anwendungsprozessor 2100
ICL 2120
ICC 2130
Verarbeitungscluster 2110
615
Speicher-Strng 2142
HBM PHY 2144
Verwaltungs-steuerungs-CPU 2150
PCIe-Steuerung und DMA-Block 2170
PCIe-X16-Anschluss 2180

FIG. 21

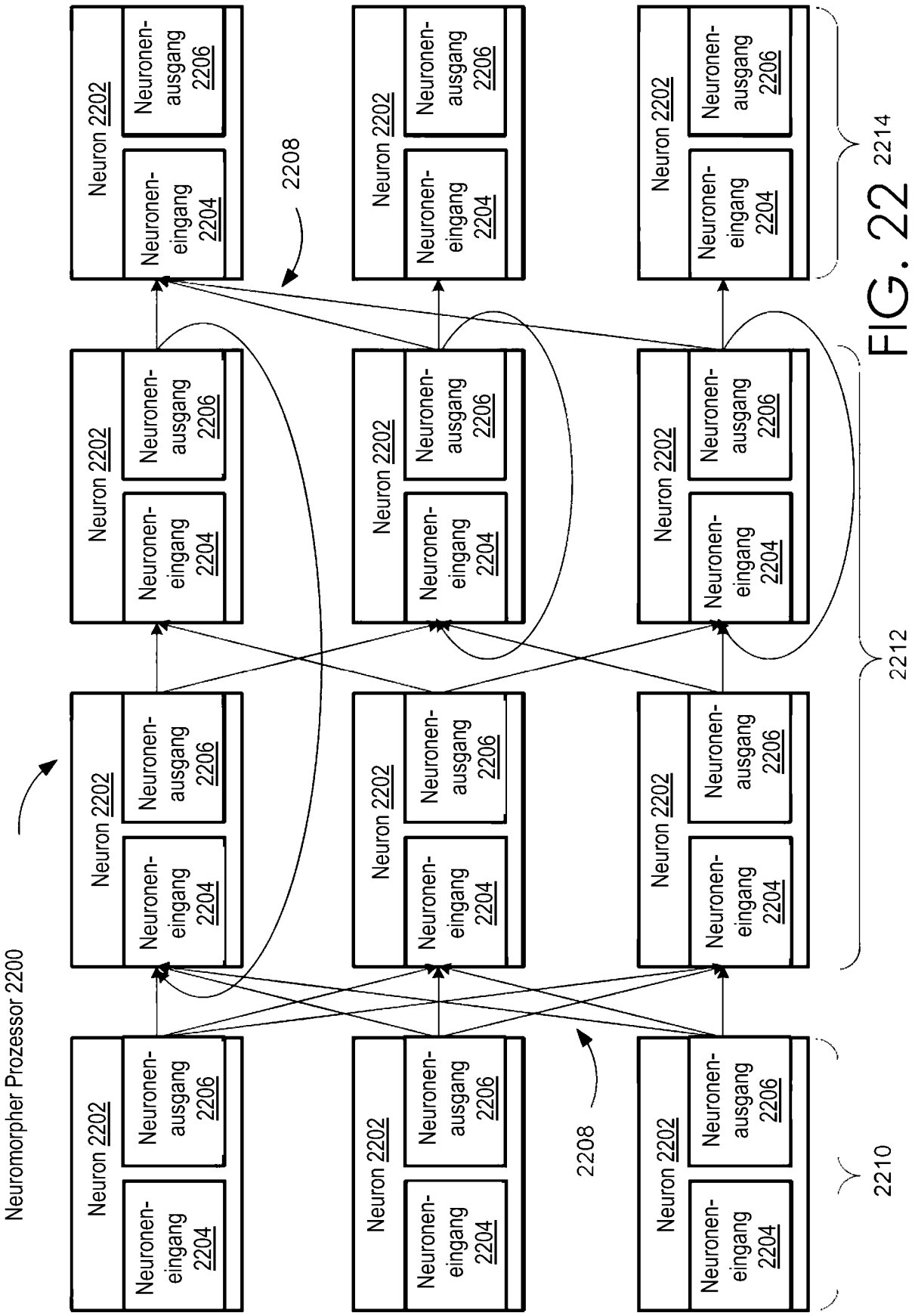

Neuromorpher Prozessor 2200
Neuron 2202
Neuronen-eingang 2204
Neuronen-ausgang 2206
2208
2210
2212
2214

FIG. 22

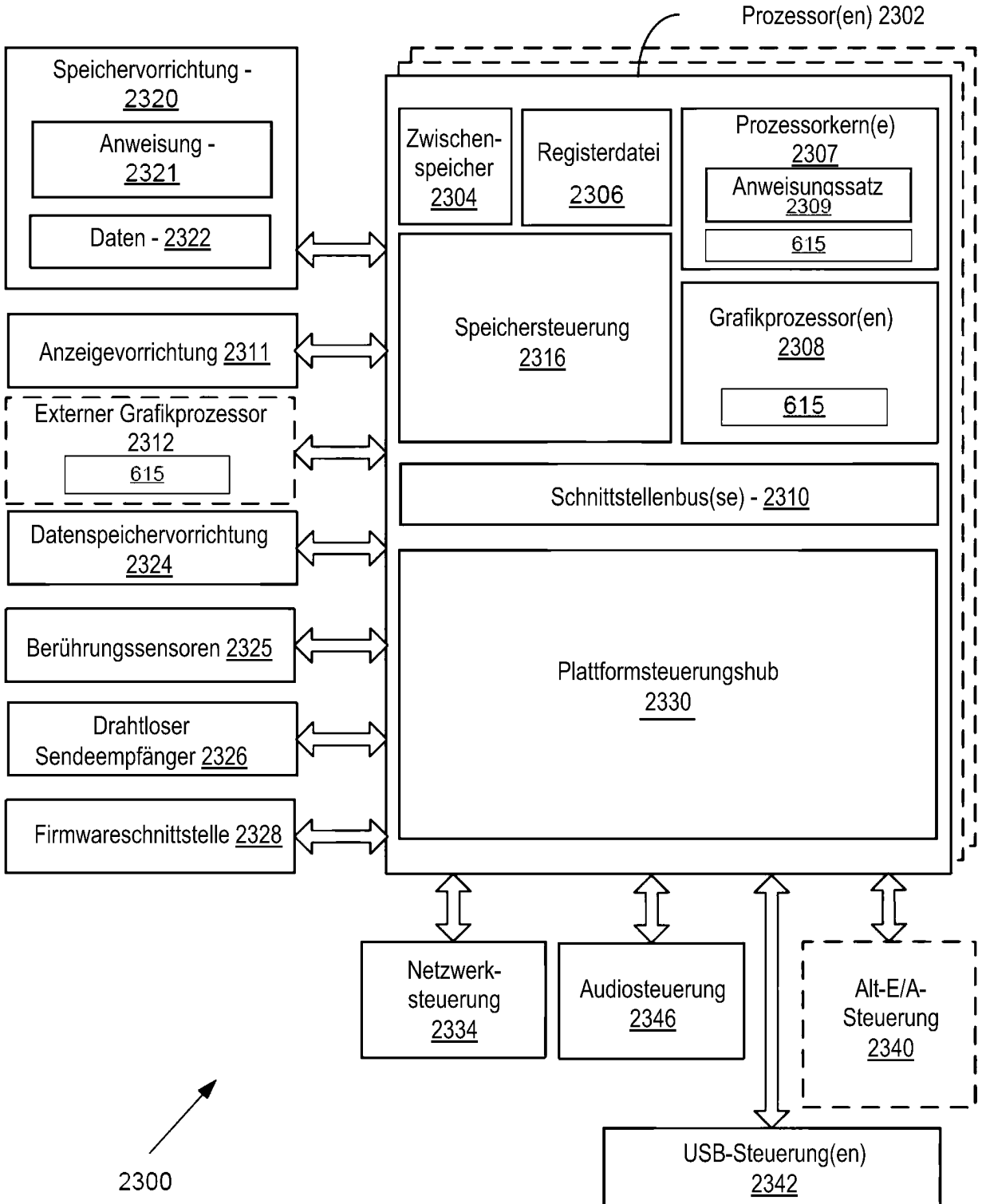

Prozessor(en) 2302
Speichervorrichtung - 2320
Anweisung - 2321
Daten - 2322
Zwischen-speicher 2304
Registerdatei 2306
Prozessorkern(e) 2307
Anweisungssatz 2309
615
Speichersteuerung 2316
Grafikprozessor(en) 2308
615
Anzeigevorrichtung 2311
Externer Grafikprozessor 2312
615
Schnittstellenbus(se) - 2310
Datenspeichervorrichtung 2324
Berührungssensoren 2325
Plattformsteuerungshub 2330
Drahtloser Sendeempfänger 2326
Firmwareschnittstelle 2328
Netzwerk-steuerung 2334
Audiosteuerung 2346
Alt-E/A-Steuerung 2340
USB-Steuerung(en) 2342
Tastatur/Maus 2343
Kamera 2344
2300

FIG. 23

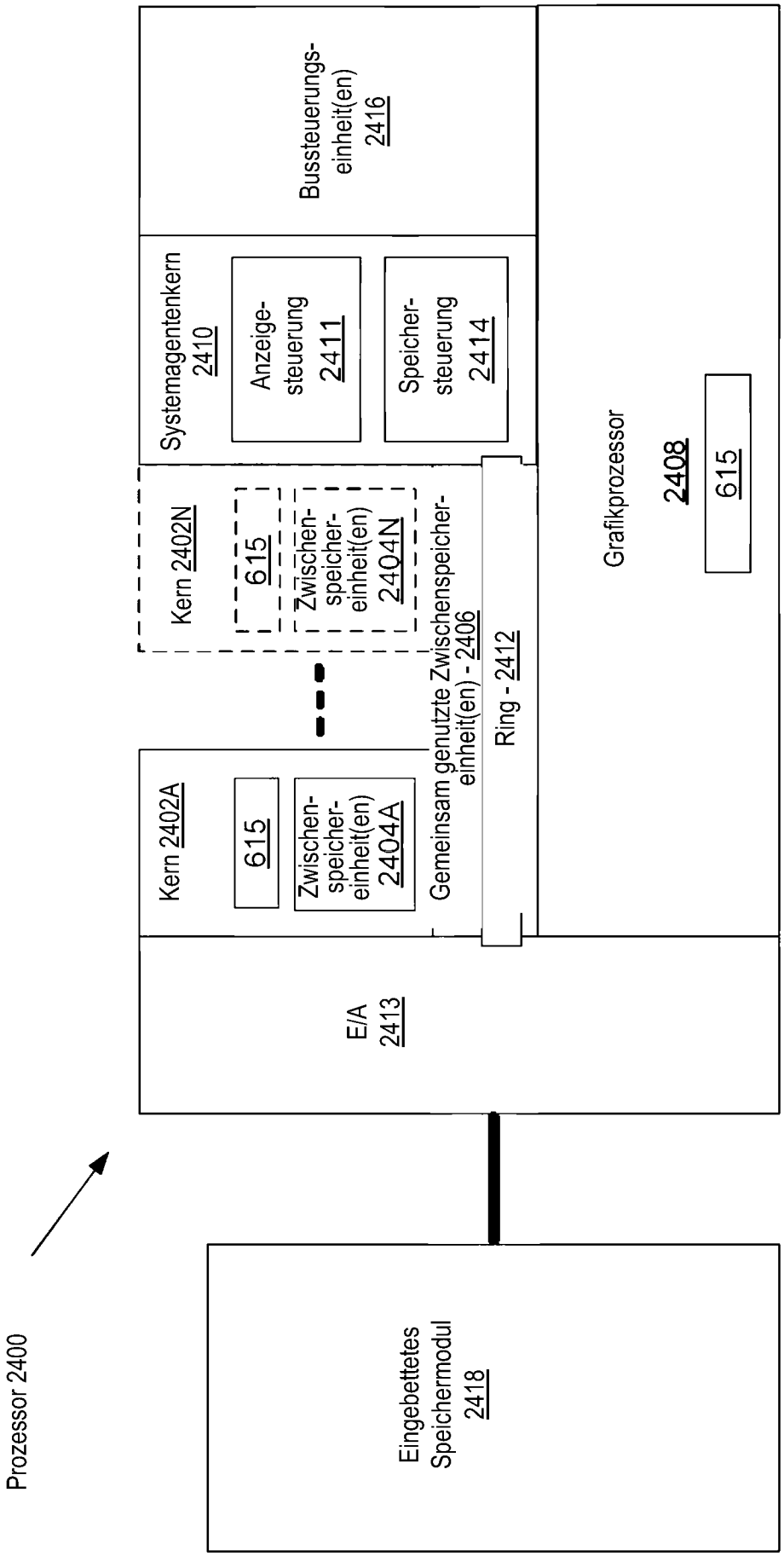

Prozessor 2400
Eingebettetes Speichermodul 2418
E/A 2413
Kern 2402A
615
Zwischen-speicher-einheit(en) 2404A
Kern 2402N
615
Zwischen-speicher-einheit(en) 2404N
Gemeinsam genutzte Zwischenspeicher-einheit(en) - 2406
Ring - 2412
Grafikprozessor 2408
615
Systemagentenkern 2410
Anzeige-steuerung 2411
Speicher-steuerung 2414
Bussteuerungs-einheit(en) 2416

FIG. 24

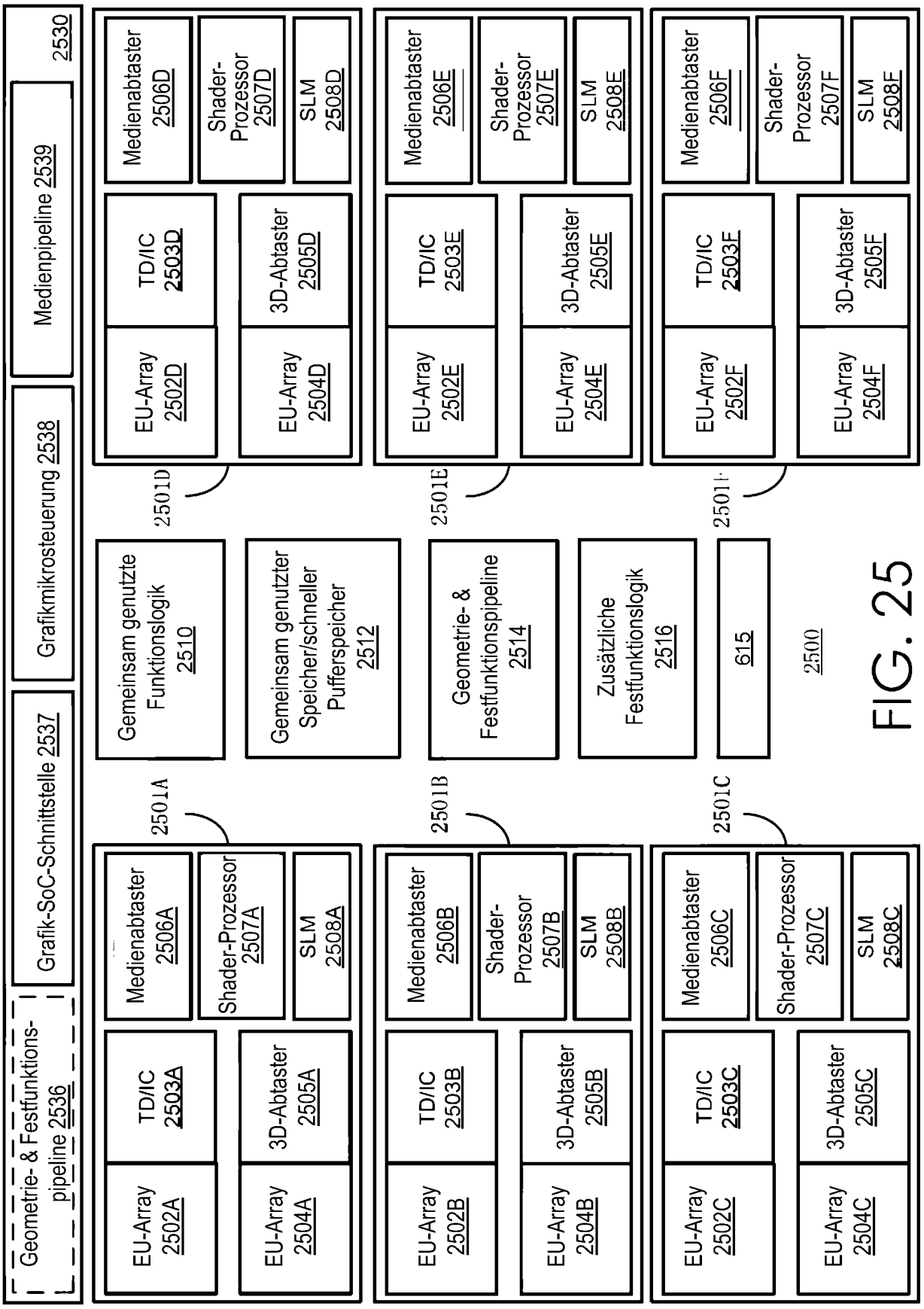

Geometrie- & Festfunktions-pipeline 2536
Grafik-SoC-Schnittstelle 2537
Grafikmikrosteuerung 2538
Medienpipeline 2539
2530
2501A
2501B
2501C
2501D
2501E
2501F
EU-Array 2502A
TD/IC 2503A
Medienabtaster 2506A
Shader-Prozessor 2507A
SLM 2508A
EU-Array 2504A
3D-Abtaster 2505A
EU-Array 2502B
TD/IC 2503B
Medienabtaster 2506B
Shader-Prozessor 2507B
SLM 2508B
EU-Array 2504B
3D-Abtaster 2505B
EU-Array 2502C
TD/IC 2503C
Medienabtaster 2506C
Shader-Prozessor 2507C
SLM 2508C
EU-Array 2504C
3D-Abtaster 2505C
EU-Array 2502D
TD/IC 2503D
Medienabtaster 2506D
Shader-Prozessor 2507D
SLM 2508D
EU-Array 2504D
3D-Abtaster 2505D
EU-Array 2502E
TD/IC 2503E
Medienabtaster 2506E
Shader-Prozessor 2507E
SLM 2508E
EU-Array 2504E
3D-Abtaster 2505E
EU-Array 2502F
TD/IC 2503F
Medienabtaster 2506F
Shader-Prozessor 2507F
SLM 2508F
EU-Array 2504F
3D-Abtaster 2505F
Gemeinsam genutzte Funktionslogik 2510
Gemeinsam genutzter Speicher/schneller Pufferspeicher 2512
Geometrie- & Festfunktionspipeline 2514
Zusätzliche Festfunktionslogik 2516
615
2500

FIG. 25

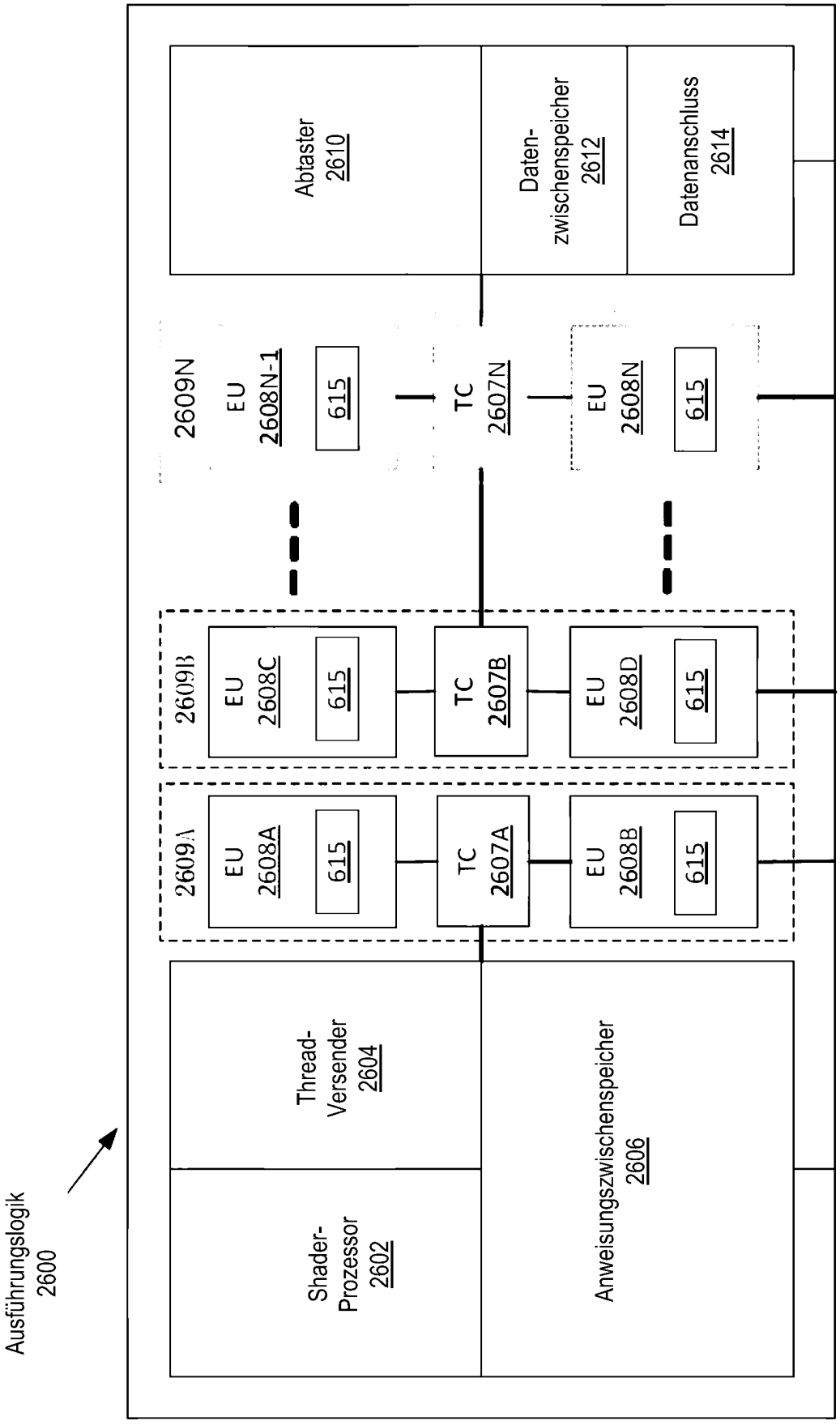

Ausführungslogik
2600
Shader-
Prozessor
2602
Thread-
Versender
2604
Anweisungszwischenspeicher
2606
2609A
EU
2608A
615
TC
2607A
EU
2608B
615
2609B
EU
2608C
615
TC
2607B
EU
2608D
615
2609N
EU
2608N-1
615
TC
2607N
EU
2608N
615
Abtaster
2610
Daten-
zwischenspeicher
2612
Datenanschluss
2614

FIG. 26A

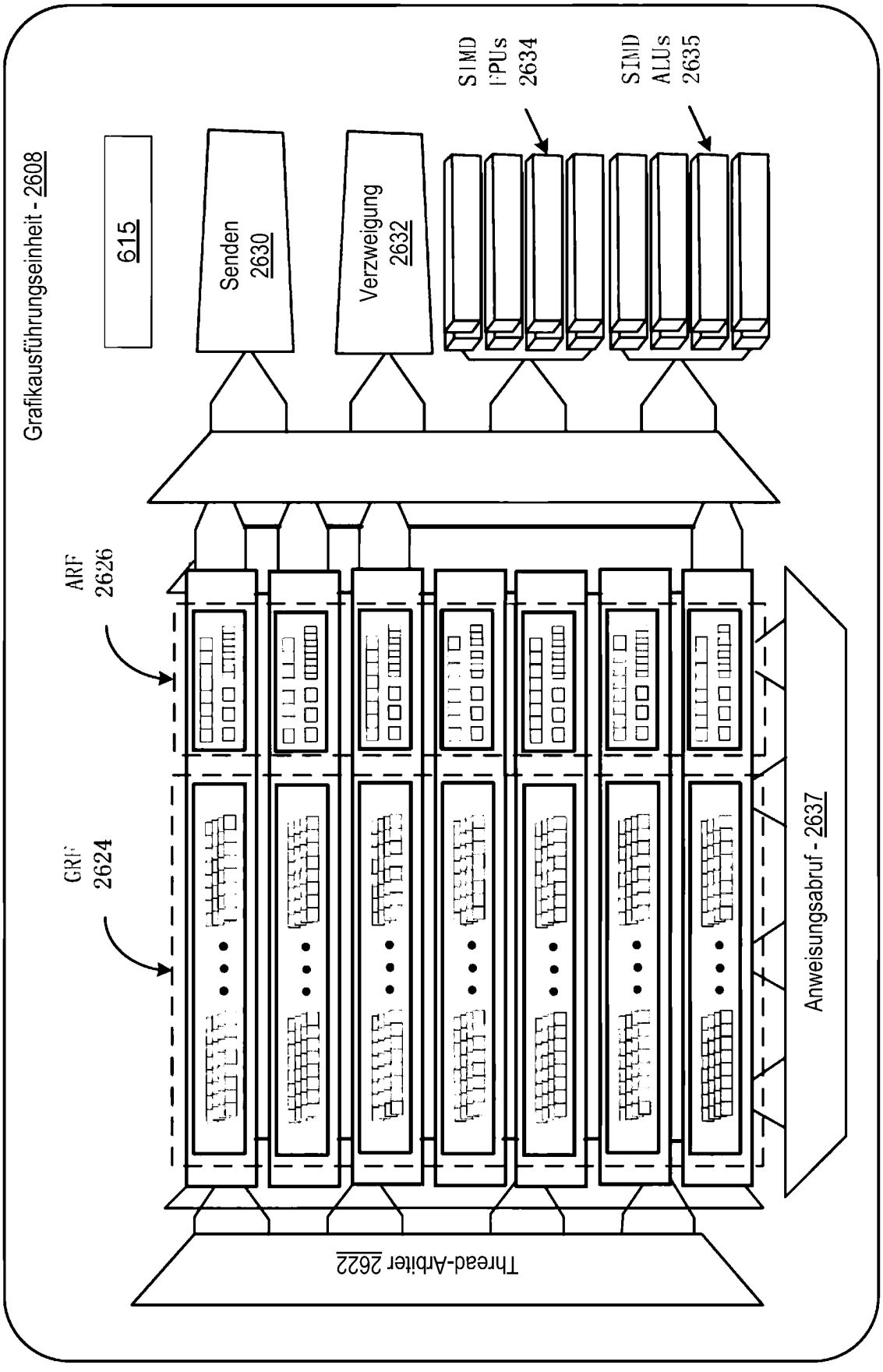

Grafikausführungseinheit - 2608
615
Senden 2630
Verzweigung 2632
SIMD FPUs 2634
SIMD ALUs 2635
ARF 2626
GRF 2624
Anweisungsabruf - 2637
Thread-Arbiter 2622

FIG. 26B

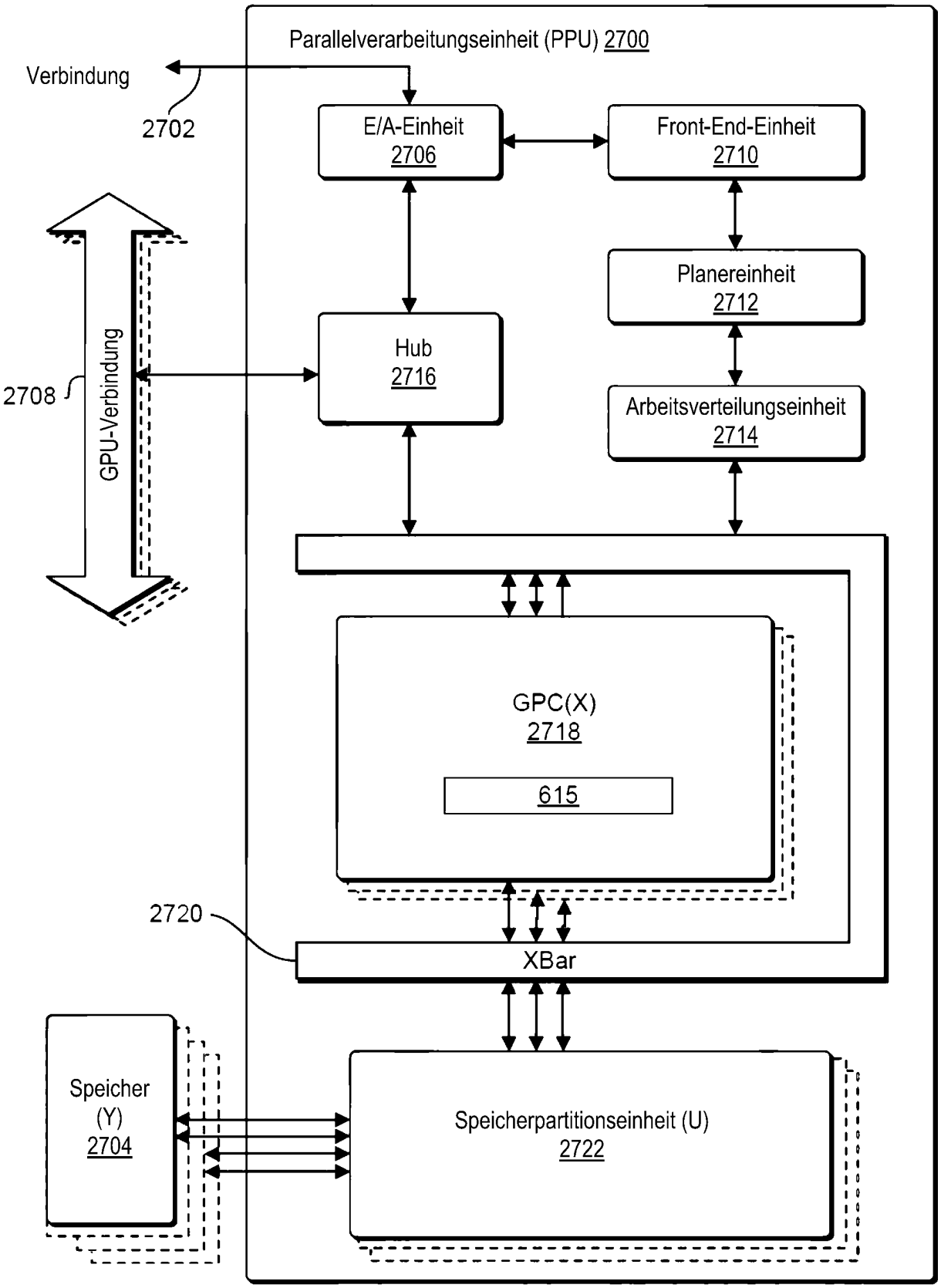

Parallelverarbeitungseinheit (PPU) 2700
Verbindung
2702
E/A-Einheit
2706
Front-End-Einheit
2710
Planereinheit
2712
Hub
2716
Arbeitsverteilungseinheit
2714
2708
GPU-Verbindung
GPC(X)
2718
615
2720
XBar
Speicher
(Y)
2704
Speicherpartitionseinheit (U)
2722

FIG. 27

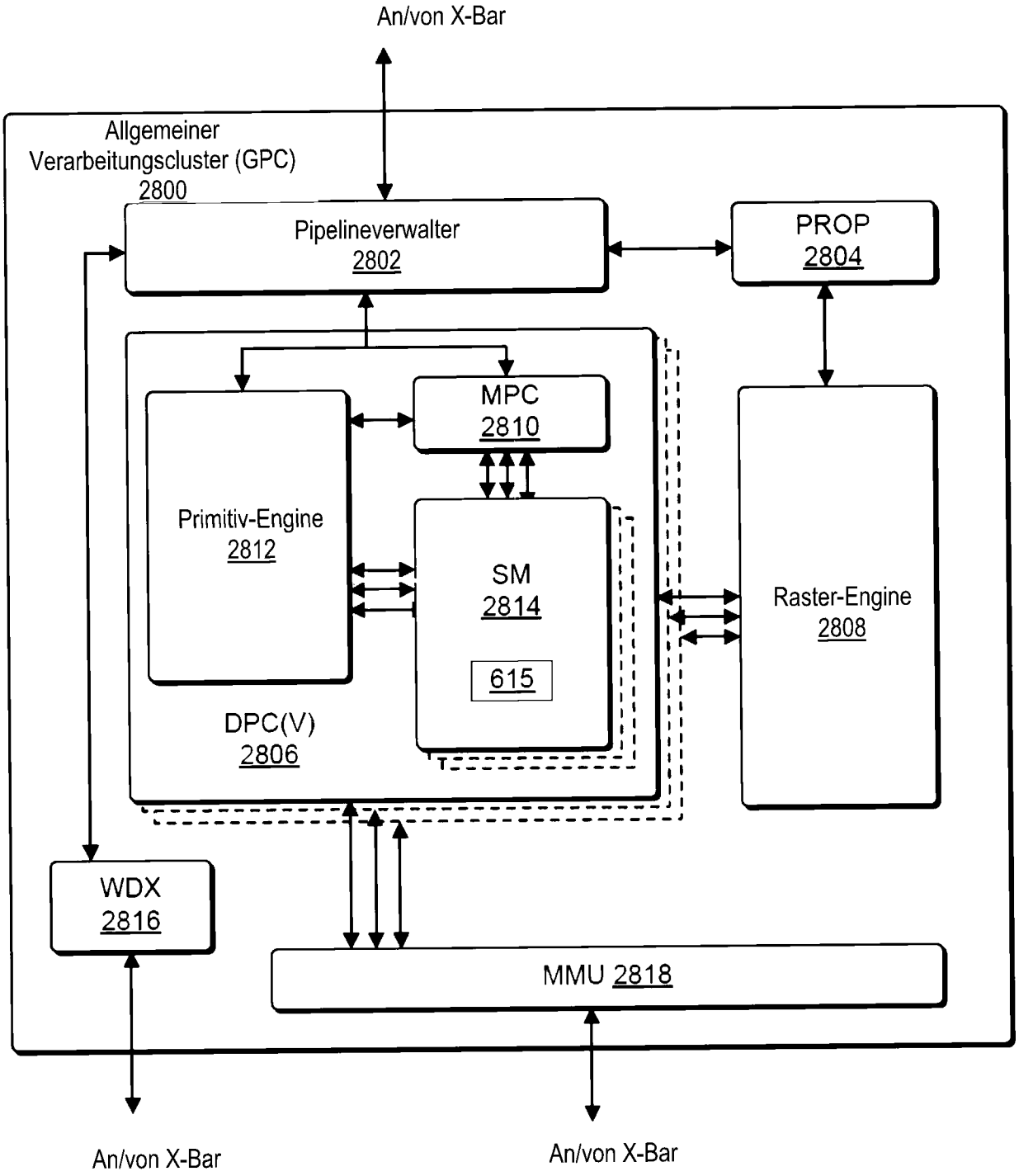

An/von X-Bar
Allgemeiner Verarbeitungscluster (GPC) 2800
Pipelineverwalter 2802
PROP 2804
MPC 2810
Primitiv-Engine 2812
SM 2814
615
Raster-Engine 2808
DPC(V) 2806
WDX 2816
MMU 2818
An/von X-Bar
An/von X-Bar

FIG. 28

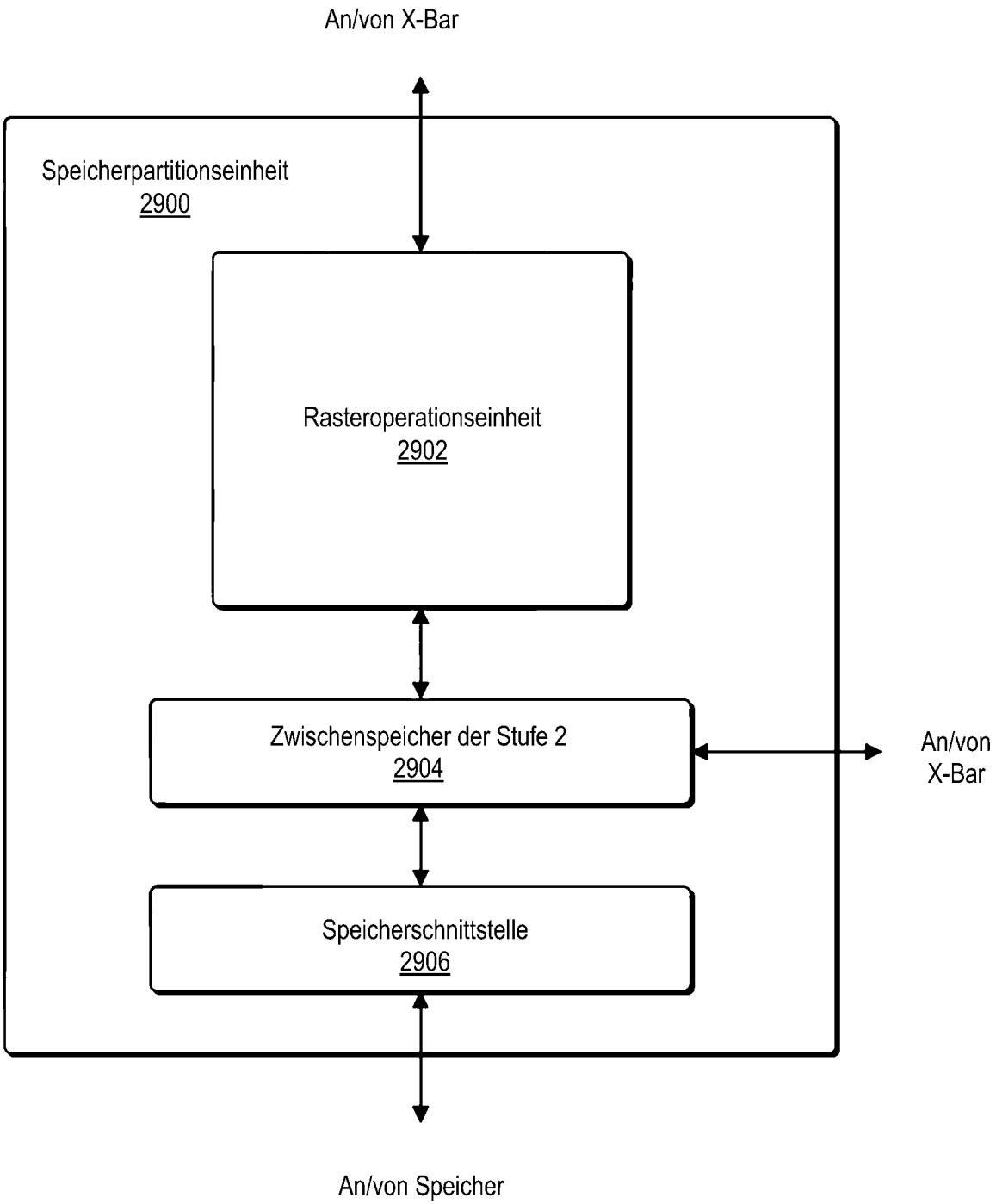

An/von X-Bar
Speicherpartitionseinheit
2900
Rasteroperationseinheit
2902
Zwischenspeicher der Stufe 2
2904
An/von
X-Bar
Speicherschnittstelle
2906
An/von Speicher

FIG. 29

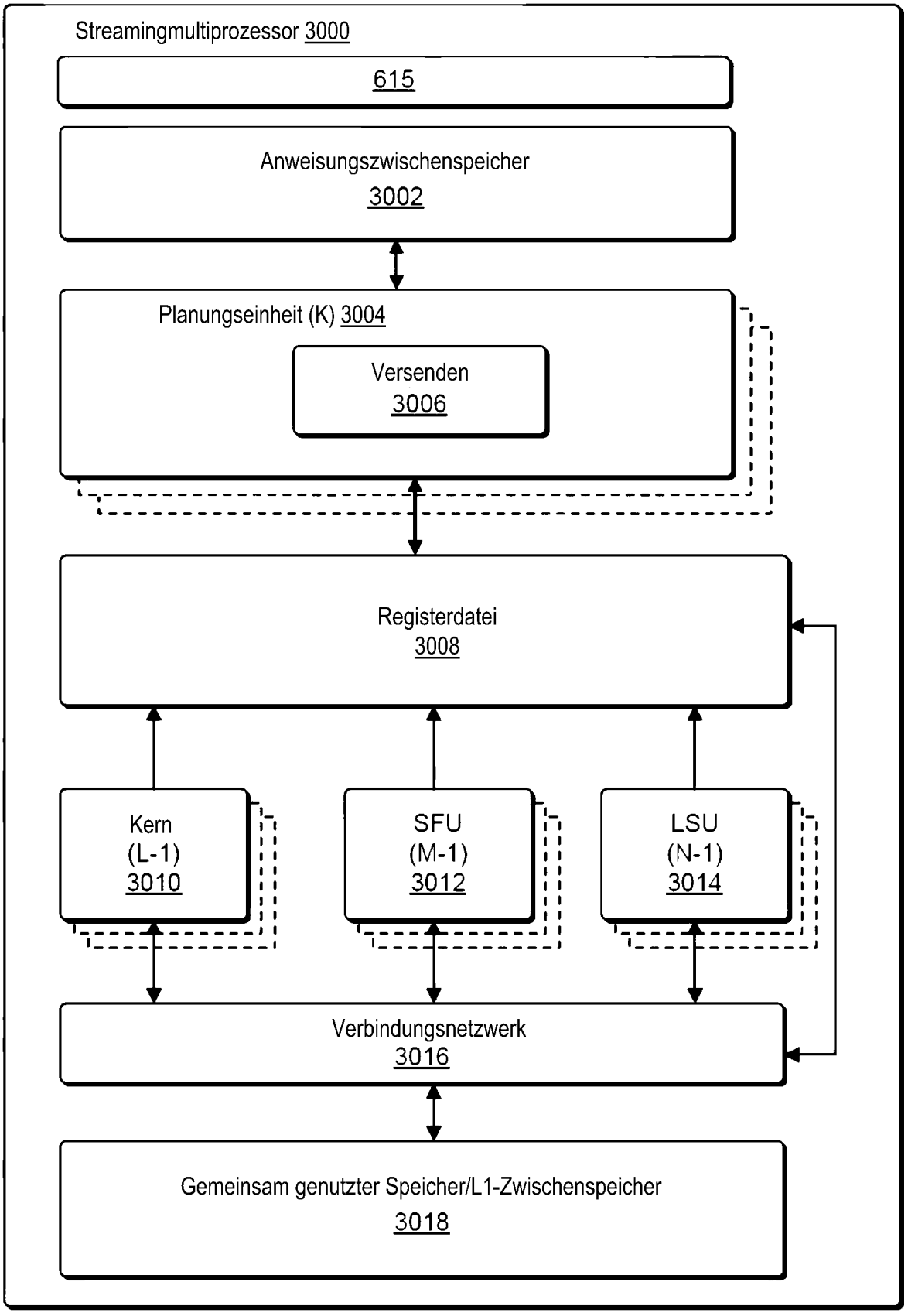

Streamingmultiprozessor 3000
615
Anweisungszwischenspeicher
3002
Planungseinheit (K) 3004
Versenden
3006
Registerdatei
3008
Kern
(L-1)
3010
SFU
(M-1)
3012
LSU
(N-1)
3014
Verbindungsnetzwerk
3016
Gemeinsam genutzter Speicher/L1-Zwischenspeicher
3018

FIG. 30

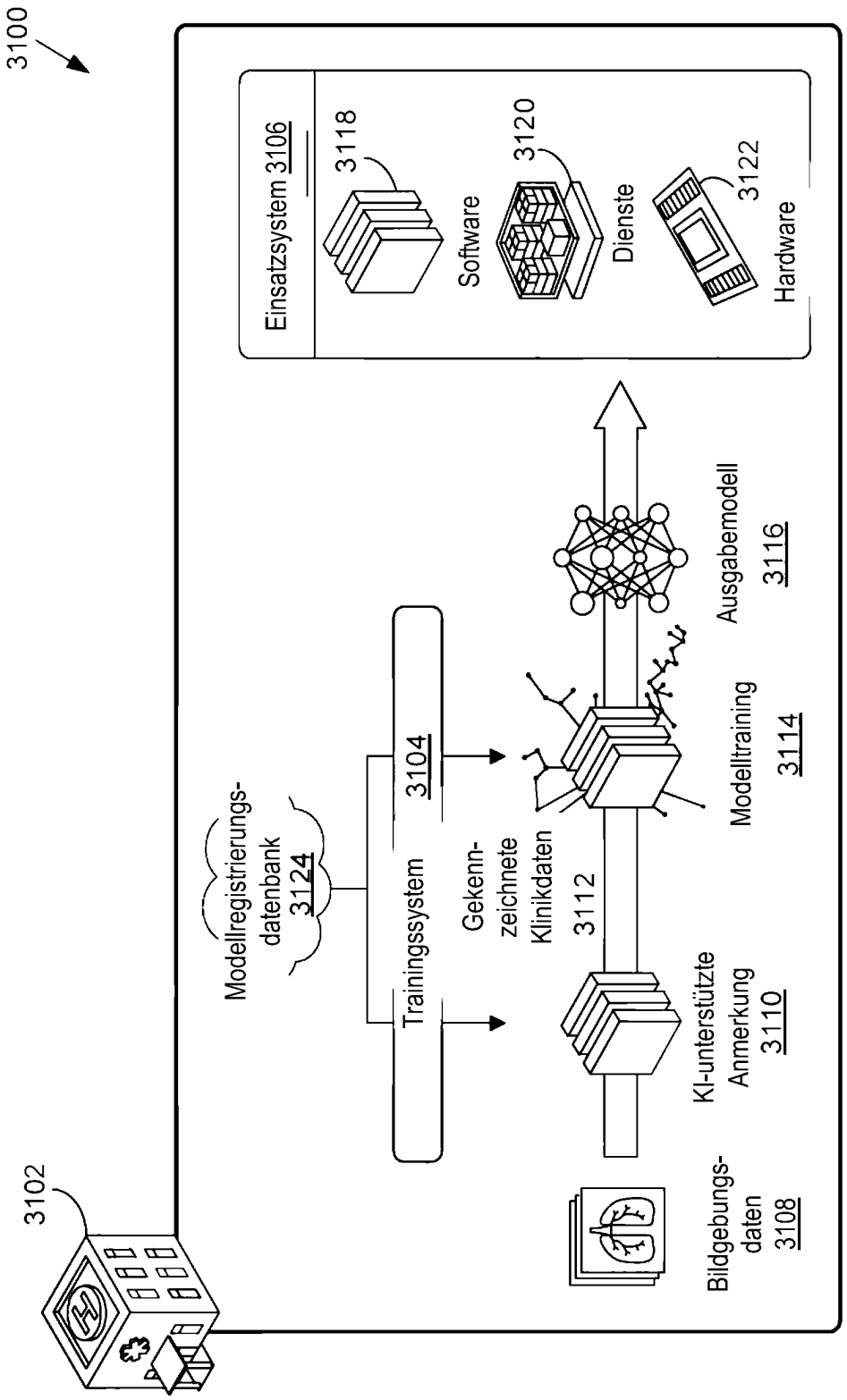

3100
3102
Einsatzsystem 3106
3118
Software
3120
Dienste
3122
Hardware
Modellregistrierungs-datenbank 3124
Trainingssystem 3104
Gekenn-zeichnete Klinikdaten 3112
Modelltraining 3114
Ausgabemodell 3116
KI-unterstützte Anmerkung 3110
Bildgebungs-daten 3108

FIG. 31

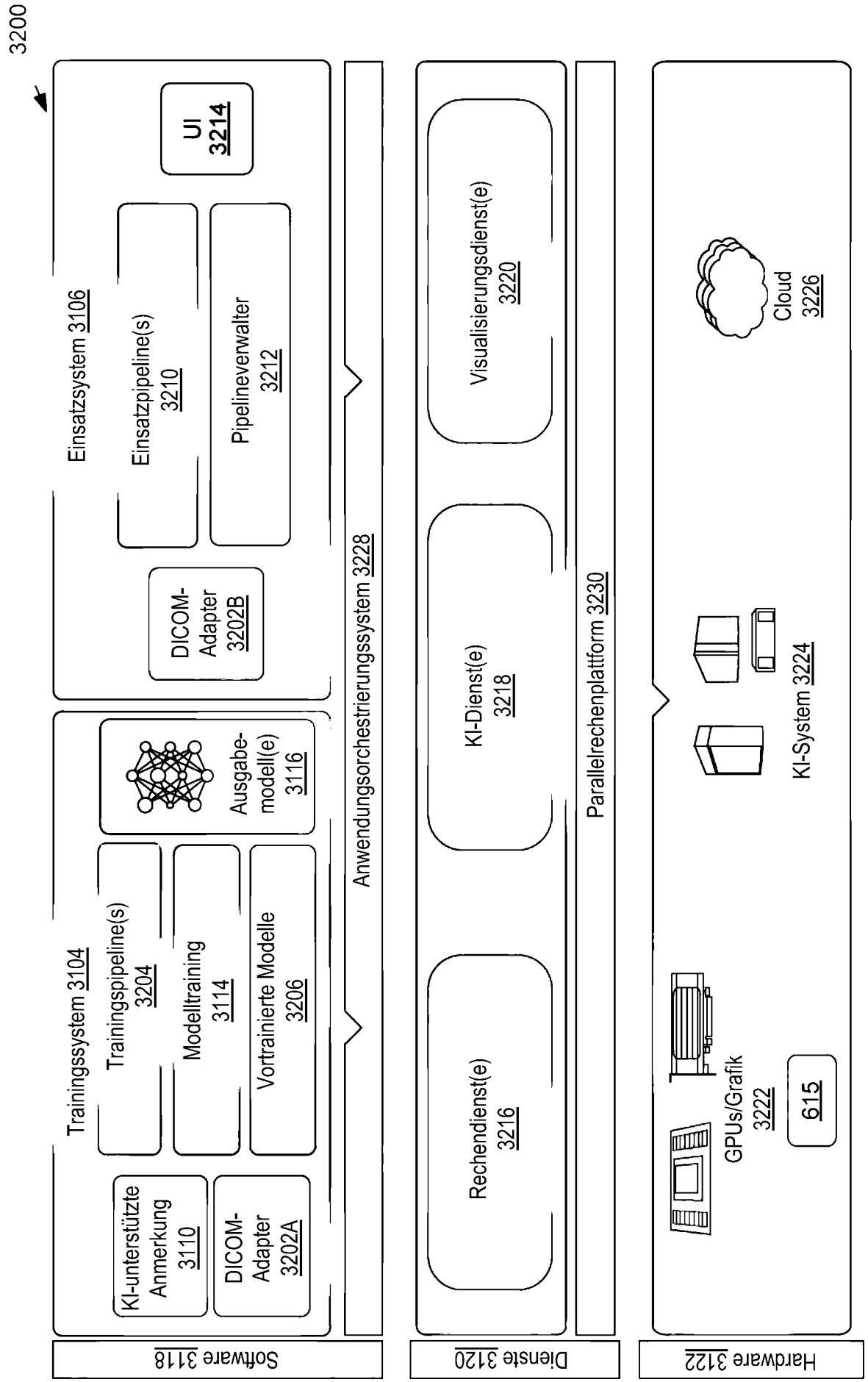

3200
Einsatzsystem 3106
UI 3214
Einsatzpipeline(s) 3210
Pipelineverwalter 3212
DICOM-Adapter 3202B
Trainingssystem 3104
Trainingspipeline(s) 3204
Modelltraining 3114
Vortrainierte Modelle 3206
Ausgabe-modell(e) 3116
KI-unterstützte Anmerkung 3110
DICOM-Adapter 3202A
Anwendungsorchestrierungssystem 3228
Visualisierungsdienst(e) 3220
KI-Dienst(e) 3218
Rechendienst(e) 3216
Parallelrechenplattform 3230
Cloud 3226
KI-System 3224
GPUs/Grafik 3222
615
Software 3118
Dienste 3120
Hardware 3122

FIG. 32

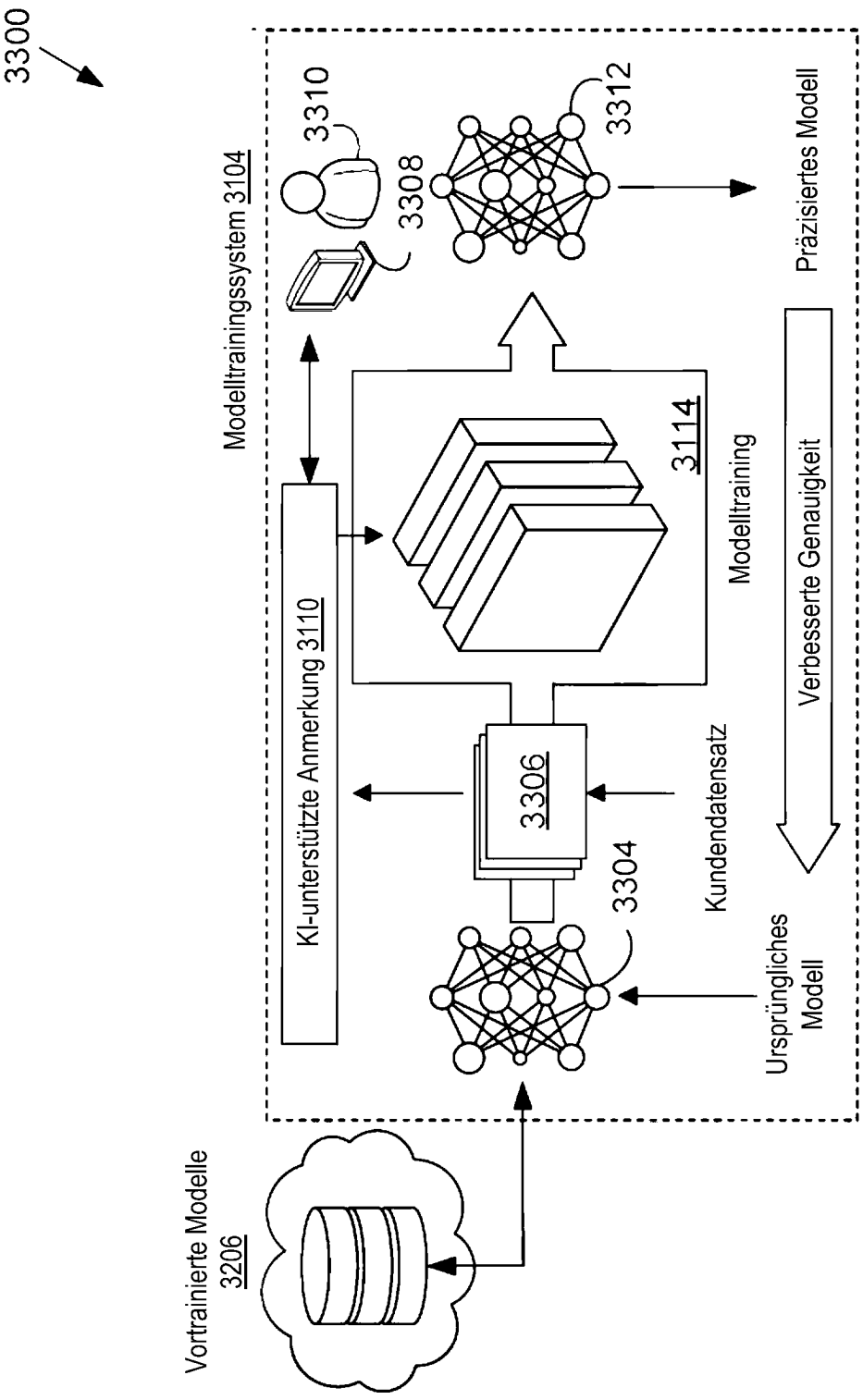

3300
Modelltrainingssystem 3104
3310
3308
3312
Präzisiertes Modell
3114
Modelltraining
KI-unterstützte Anmerkung 3110
Verbesserte Genauigkeit
3306
3304
Kundendatensatz
Ursprüngliches Modell
Vortrainierte Modelle 3206

FIG. 33A

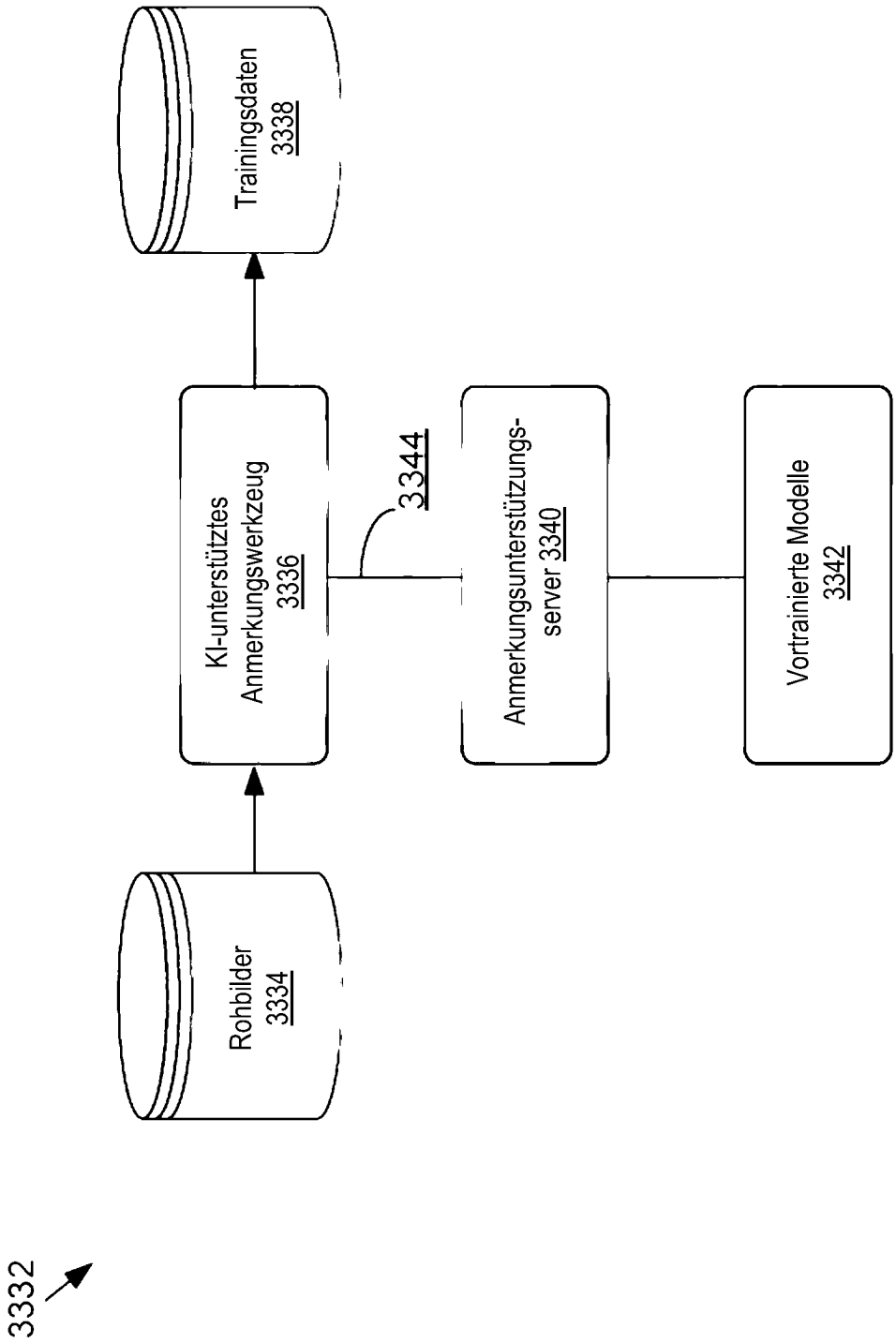

3332
Rohbilder 3334
KI-unterstütztes Anmerkungswerkzeug 3336
Trainingsdaten 3338
3344
Anmerkungsunterstützungs-server 3340
Vortrainierte Modelle 3342

FIG. 33B

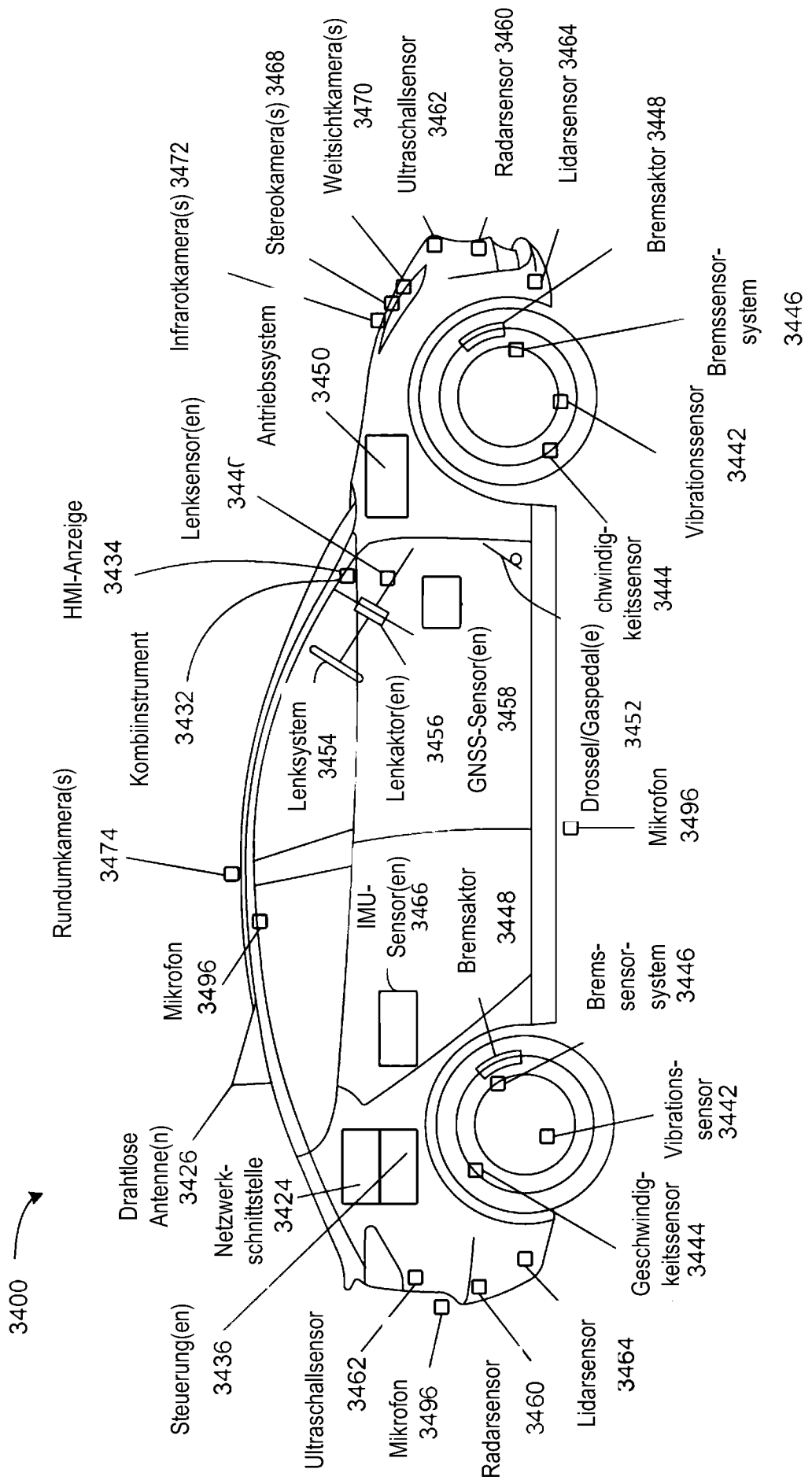

3400
Steuerung(en) 3436
Drahtlose Antenne(n) 3426
Netzwerk-schnittstelle 3424
Rundumkamera(s) 3474
Mikrofon 3496
HMI-Anzeige 3434
Kombiinstrument 3432
Lenksensor(en) 3447
Lenksystem 3454
Antriebssystem 3450
Infrarotkamera(s) 3472
Stereokamera(s) 3468
Weitsichtkamera(s) 3470
Ultraschallsensor 3462
Radarsensor 3460
Lidarsensor 3464
Bremsaktor 3448
Bremssensor-system 3446
Vibrationssensor 3442
Geschwindigkeitssensor 3444
Lenkaktor(en) 3456
GNSS-Sensor(en) 3458
IMU-Sensor(en) 3466
Drossel/Gaspedal(e) 3452
Mikrofon 3496
Ultraschallsensor 3462
Mikrofon 3496
Radarsensor 3460
Lidarsensor 3464
Geschwindigkeitssensor 3444
Vibrationssensor 3442
Bremssensor-system 3446
Bremsaktor 3448

FIG. 34A

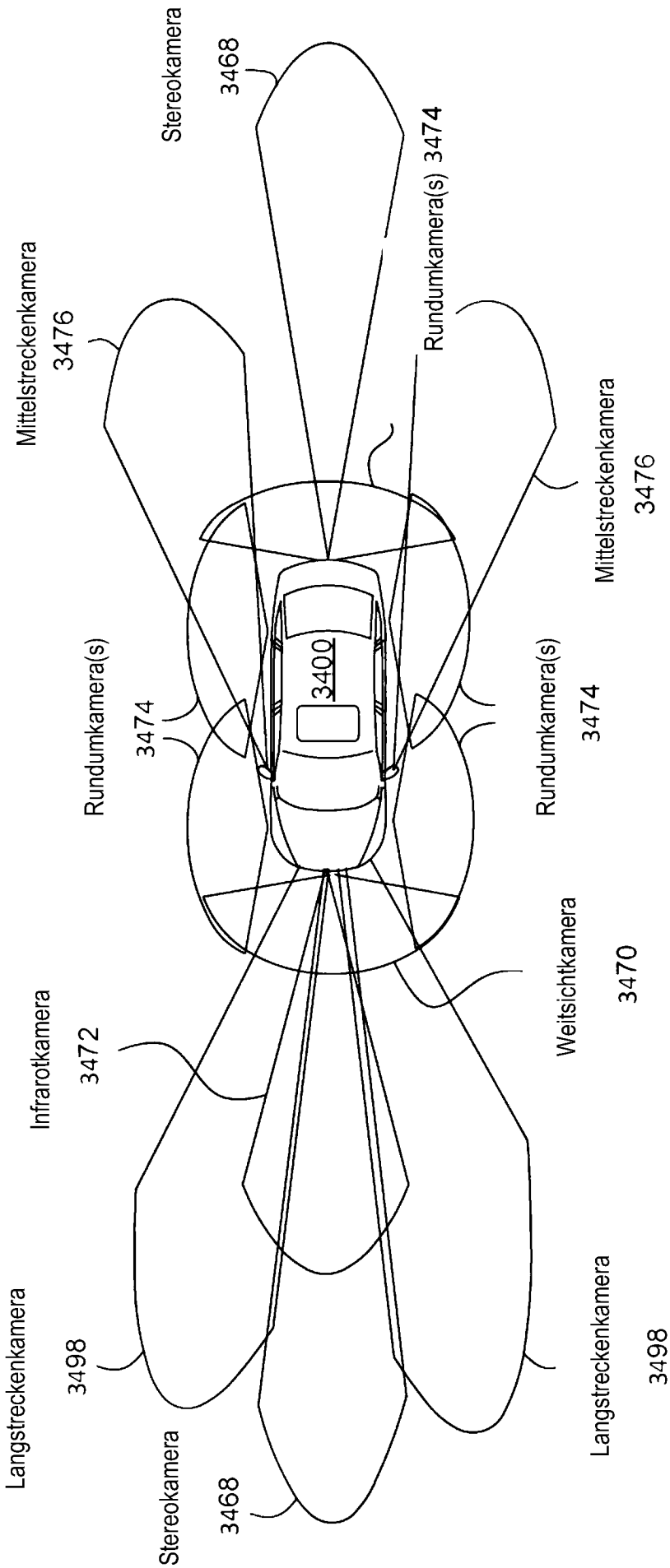

Stereokamera
3468
Mittelstreckenkamera
3476
Rundumkamera(s) 3474
3400
Rundumkamera(s)
3474
Mittelstreckenkamera
3476
Infrarotkamera
3472
Weitsichtkamera
3470
Langstreckenkamera
3498
Stereokamera
3468
Langstreckenkamera
3498

FIG. 34B

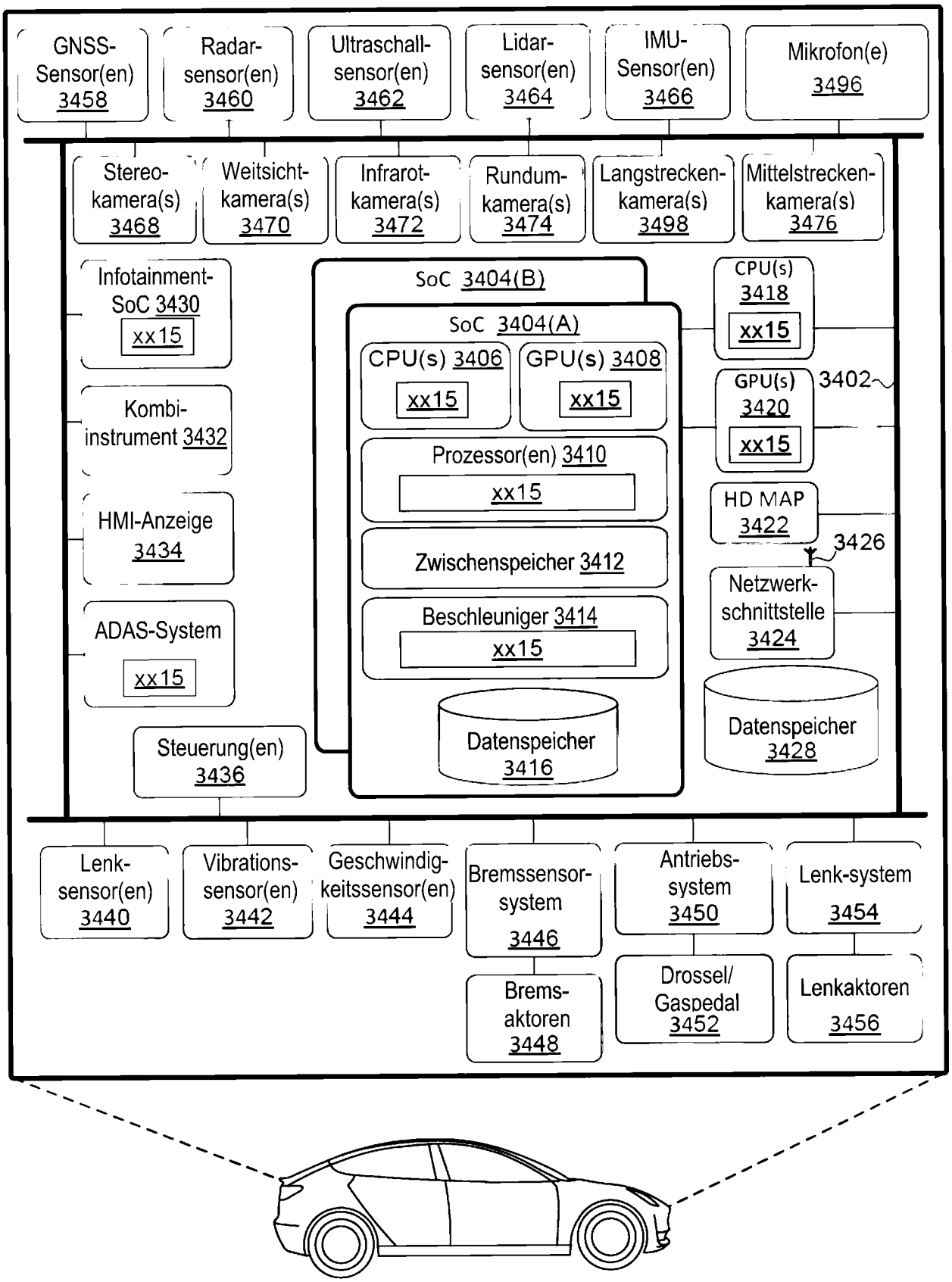

GNSS-Sensor(en) 3458
Radar-sensor(en) 3460
Ultraschall-sensor(en) 3462
Lidar-sensor(en) 3464
IMU-Sensor(en) 3466
Mikrofon(e) 3496
Stereo-kamera(s) 3468
Weitsicht-kamera(s) 3470
Infrarot-kamera(s) 3472
Rundum-kamera(s) 3474
Langstrecken-kamera(s) 3498
Mittelstrecken-kamera(s) 3476
Infotainment-SoC 3430
xx15
Kombi-instrument 3432
HMI-Anzeige 3434
ADAS-System
xx15
Steuerung(en) 3436
SoC 3404(B)
SoC 3404(A)
CPU(s) 3406
xx15
GPU(s) 3408
xx15
Prozessor(en) 3410
xx15
Zwischenspeicher 3412
Beschleuniger 3414
xx15
Datenspeicher 3416
CPU(s) 3418
xx15
GPU(s) 3420
xx15
3402
HD MAP 3422
3426
Netzwerk-schnittstelle 3424
Datenspeicher 3428
Lenk-sensor(en) 3440
Vibrations-sensor(en) 3442
Geschwindig-keitssensor(en) 3444
Bremssensor-system 3446
Brems-aktoren 3448
Antriebs-system 3450
Drossel/ Gaspedal 3452
Lenk-system 3454
Lenkaktoren 3456

FIG. 34C

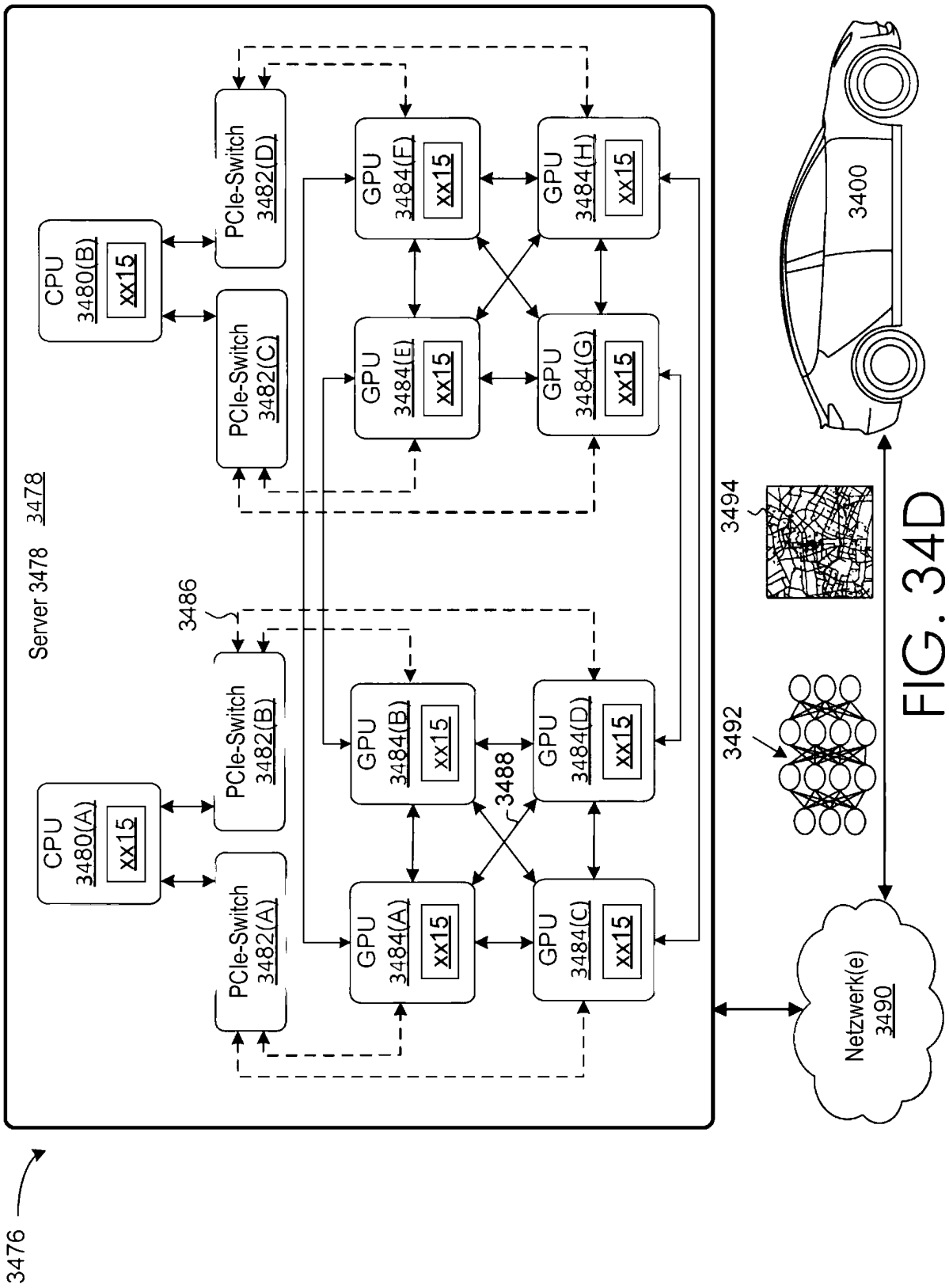

3476
Server 3478 3478
CPU 3480(A) XX15
CPU 3480(B) XX15
PCIe-Switch 3482(A)
PCIe-Switch 3482(B)
PCIe-Switch 3482(C)
PCIe-Switch 3482(D)
3486
3488
GPU 3484(A) XX15
GPU 3484(B) XX15
GPU 3484(C) XX15
GPU 3484(D) XX15
GPU 3484(E) XX15
GPU 3484(F) XX15
GPU 3484(G) XX15
GPU 3484(H) XX15
3400
3494
3492
Netzwerk(e) 3490

FIG. 34D