(19) Deutsches Patent- und Markenamt

(10) **DE 10 2022 120 593 A1** 2023.03.02

(12) # Offenlegungsschrift

(21) Aktenzeichen: **10 2022 120 593.2**
(22) Anmeldetag: **16.08.2022**
(43) Offenlegungstag: **02.03.2023**

(51) Int Cl.: ***G06F 1/3246*** (2019.01)
G06F 1/32 (2019.01)
G06F 9/50 (2006.01)

(30) Unionspriorität:
17/459,351 **27.08.2021** **US**

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

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

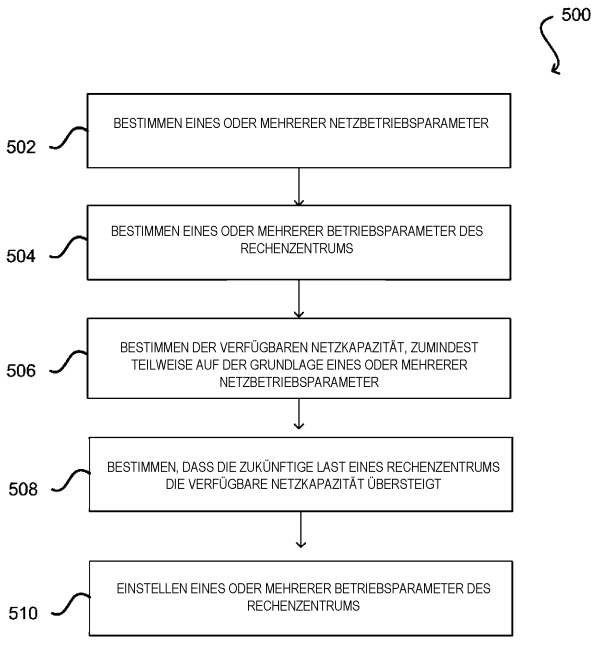
(72) Erfinder:
Goska, Benjamin, Santa Clara, CA, US; Albright, Ryan, Santa Clara, CA, US; Mecham, William Andrew, Santa Clara, CA, US; Weese, William Ryan, Santa Clara, CA, US; Carkin, Aaron Richard, Santa Clara, CA, US; Thompson, Michael, Santa Clara, CA, US

Prüfungsantrag gemäß § 44 PatG ist gestellt.

Die folgenden Angaben sind den vom Anmelder eingereichten Unterlagen entnommen.

(54) Bezeichnung: **LASTÜBERWACHUNG EINES RECHENZENTRUMS**

(57) Zusammenfassung: Es werden Systeme und Verfahren für die Verwaltung von Rechenzentrumslasten offengelegt. In mindestens einer Ausführungsform wird eine Netzkapazität bestimmt und ein oder mehrere Betriebsparameter für eine oder mehrere Rechenzentrumskomponenten werden zumindest teilweise basierend auf der Netzkapazität angepasst.



Beschreibung

GEBIET

[0001] Mindestens eine Ausführungsform betrifft Systeme zur Lastüberwachung eines Rechenzentrums. Mindestens eine Ausführungsform bezieht sich beispielsweise auf das Anpassen eines oder mehrerer Betriebsparameter eines Rechenzentrums basierend auf einer verfügbaren Last von einem Netz.

STAND DER TECHNIK

[0002] In Rechenumgebungen wie Rechenzentren werden verschiedene Komponenten in Racks installiert. Diese Komponenten können Serverkomponenten, Netzteile, Panels und andere umfassen. Komponenten können einem oder mehreren Sensoren zugeordnet sein, die Informationen bezüglich Betriebsbedingungen von Komponenten sammeln. Verschiedene Komponenten können auf unterschiedlichen Niveaus des Energieverbrauchs betrieben werden, der durch eine oder mehrere Verbindungen zu externen Stromquellen bereitgestellt wird, wie z. B. ein Netz, das Energie von verschiedenen Stromerzeugungsstandorten oder Stromerzeugungsmöglichkeiten vor Ort überträgt. Da die Netze Energie aufnehmen, bündeln und an eine Vielzahl verschiedener Endverbraucher verteilen, kann die Verfügbarkeit von Energie in Abhängigkeit von externen Faktoren zu unterschiedlichen Zeiten schwanken oder Spitzenwerte erreichen. Mit zunehmender Größe und Komplexität von Rechenzentren wird es immer schwieriger, eine ausreichende Stromversorgung sicherzustellen und gleichzeitig die Belastung der Netze zu reduzieren.

Figurenliste

Fig. 1 veranschaulicht ein Rechenzentrum gemäß mindestens einer Ausführungsform;

Fig. 2 veranschaulicht eine schematische Ansicht eines Stromnetzes und eines Rechenzentrums, gemäß mindestens einer Ausführungsform;

Fig. 3 veranschaulicht eine schematische Ansicht eines Stromnetzes und von Rechenzentren gemäß mindestens einer Ausführungsform;

Fig. 4A-4D veranschaulichen grafische Darstellungen von Rechenzentrumslasten gemäß mindestens einer Ausführungsform;

Fig. 5A veranschaulicht einen Prozess für das Lastmanagement von Rechenzentren gemäß mindestens einer Ausführungsform;

Fig. 5B zeigt einen Prozess für das Lastmanagement von Rechenzentren gemäß mindestens einer Ausführungsform;

Fig. 5C zeigt einen Prozess für das Lastmanagement von Rechenzentren gemäß mindestens einer Ausführungsform;

Fig. 6 veranschaulicht ein verteiltes System gemäß mindestens einer Ausführungsform;

Fig. 7 zeigt ein beispielhaftes Rechenzentrum gemäß mindestens einer;

Fig. 8 veranschaulicht ein Client-Server-Netzwerk gemäß mindestens einer Ausführungsform;

Fig. 9 veranschaulicht ein Computernetzwerk gemäß mindestens einer Ausführungsform;

Fig. 10A veranschaulicht ein vernetztes Computersystem gemäß mindestens einer Ausführungsform;

Fig. 10B veranschaulicht ein vernetztes Computersystem gemäß mindestens einer Ausführungsform;

Fig. 10C veranschaulicht ein vernetztes Computersystem gemäß mindestens einer Ausführungsform;

Fig. 11 zeigt eine oder mehrere Komponenten einer Systemumgebung gemäß mindestens einer Ausführungsform, in der Dienste als Drittanbieter-Netzwerkdienste angeboten;

Fig. 12 veranschaulicht eine Cloud-Computing-Umgebung gemäß mindestens einer Ausführungsform;

Fig. 13 veranschaulicht einen Satz funktionaler Abstraktionsschichten, der durch eine Cloud-Computing-Umgebung bereitgestellt wird, gemäß mindestens einer Ausführungsform;

Fig. 14 zeigt einen Supercomputer auf Chipebene gemäß mindestens einer;

Fig. 15 veranschaulicht einen Supercomputer auf Rackmodulebene gemäß mindestens einer Ausführungsform;

Fig. 16 zeigt einen Supercomputer auf Rackebene gemäß mindestens einer;

Fig. 17 zeigt einen Supercomputer auf Gesamtsystemebene gemäß mindestens;

Fig. 18A veranschaulicht Inferenz- und/oder Trainingslogik gemäß mindestens einer Ausführungsform;

Fig. 18B veranschaulicht Inferenz- und/oder Trainingslogik gemäß mindestens einer Ausführungsform;

Fig. 19 veranschaulicht das Training und den Einsatz eines neuronalen Netzes gemäß mindestens einer Ausführungsform;

Fig. 20 veranschaulicht eine Architektur eines Systems eines Netzwerks gemäß mindestens einer Ausführungsform;

Fig. 21 veranschaulicht eine Architektur eines Systems eines Netzwerks gemäß mindestens einer Ausführungsform;

Fig. 22 veranschaulicht einen Protokollstapel auf Steuerebene gemäß mindestens einer Ausführungsform;

Fig. 23 veranschaulicht einen Protokollstapel auf Benutzerebene gemäß mindestens einer Ausführungsform;

Fig. 24 veranschaulicht Komponenten eines Kernnetzes gemäß mindestens einer Ausführungsform;

Fig. 25 veranschaulicht Komponenten eines Systems zum Unterstützen von Netzwerkfunktionsvirtualisierung (NFV) gemäß mindestens einer Ausführungsform;

Fig. 26 veranschaulicht ein Verarbeitungssystem gemäß mindestens einer Ausführungsform;

Fig. 27 veranschaulicht ein Computersystem gemäß mindestens einer Ausführungsform;

Fig. 28 veranschaulicht ein System gemäß mindestens einer Ausführungsform;

Fig. 29 veranschaulicht eine beispielhafte integrierte Schaltung gemäß mindestens einer Ausführungsform;

Fig. 30 veranschaulicht ein Computersystem gemäß mindestens einer Ausführungsform;

Fig. 31 veranschaulicht eine APU gemäß mindestens einer Ausführungsform;

Fig. 32 veranschaulicht eine CPU gemäß mindestens einer Ausführungsform;

Fig. 33 zeigt ein beispielhaftes Beschleunigerintegrations-Slice gemäß mindestens einer Ausführungsform;

Fig. 34A-34B stellen beispielhafte Grafikprozessoren gemäß mindestens einer Ausführungsform dar;

Fig. 35A zeigt einen Grafikkern gemäß mindestens einer Ausführungsform;

Fig. 35B veranschaulicht eine GPGPU gemäß mindestens einer Ausführungsform;

Fig. 36A veranschaulicht einen Parallelprozessor gemäß mindestens einer Ausführungsform;

Fig. 36B veranschaulicht einen Verarbeitungs-Cluster gemäß mindestens einer Ausführungsform;

Fig. 36C veranschaulicht einen Grafikmultiprozessor gemäß mindestens einer Ausführungsform;

Fig. 37 veranschaulicht einen Softwarestapel einer Programmierplattform gemäß mindestens einer Ausführungsform;

Fig. 38 veranschaulicht eine CUDA-Implementierung eines Softwarestapels der **Fig. 37** gemäß mindestens einer Ausführungsform;

Fig. 39 veranschaulicht eine ROCm-Implementierung eines Softwarestapels der **Fig. 37** gemäß mindestens einer Ausführungsform;

Fig. 40 veranschaulicht eine OpenCL-Implementierung eines Softwarestapels der **Fig. 37** gemäß mindestens einer Ausführungsform;

Fig. 41 zeigt Software, die durch eine Programmierplattform gemäß mindestens einer Ausführungsform unterstützt wird;

Fig. 42 zeigt das Kompilieren von Code zum Ausführen auf Programmierplattformen der **Fig. 37-40** gemäß mindestens einer Ausführungsform.

DETAILLIERTE BESCHREIBUNG

[0003] In mindestens einer Ausführungsform kann eine Rechenumgebung eine Vielzahl von Rechenvorrichtungen und Steuersystemen umfassen, wie in dem Rechenzentrum 100 in **Fig. 1** dargestellt. In mindestens einer Ausführungsform kann das Rechenzentrum 100 einen oder mehrere Räume 102 mit Racks 104 und Hilfsausrüstung einschließen, die verwendet werden, um einen oder mehrere Server auf einem oder mehreren Serverfächern unterzubringen. In mindestens einer Ausführungsform wird das Rechenzentrum 100 von verschiedenen Kühlsystemen unterstützt, wie etwa Kühltürmen, Kühlschleifen, Pumpen und anderen Unterstützungssystemen. In mindestens einer Ausführungsform sind Server 106 in Racks 104 positioniert. In mindestens einer Ausführungsform erhalten die Server 106 in den Racks 104 Betriebsstrom von einer Quelle 108 und können auch mit verschiedenen Kommunikationsquellen gekoppelt sein, wie z. B. einer Verbindung mit einer Netzwerkleitung. In mindestens einer Ausführungsform können die Racks 104 ferner zusätzliche Rackkomponenten 110 einschließen, die Paneele, Router, Schalter, Luftströmungssysteme und verschiedene andere Optionen einschließen können. In mindestens einer Ausführungsform stellt die Quelle 108 Betriebsleistung für zusätzliche Rackkomponenten 110 bereit. In mindestens einer Ausführungsform sind mehrere Quellen 108 in Racks 104 angeordnet. In mindestens einer Ausführungsform erhalten Komponenten in bestimmten Racks 104 Betriebsleistung von Quellen 108 in bestimmten Racks 104. In mindestens einer Ausführungsform erhalten Komponenten in bestimmten Racks 104 Betriebsleistung von Quellen 108 in anderen Racks 104. In mindestens einer Ausführungsform sind ein oder mehrere Server 104, Stromquellen 108 und zusätzliche Rackkomponenten 110 miteinander gekoppelt oder verbunden. In mindestens einer Ausführungsform sind eine oder mehrere den Racks 104 zugeordnete Komponenten mit mehreren Ausrüstungsteilen gekoppelt und weisen eine Vielzahl von Kabeln 112 auf, die sich zwischen den Ausrüstungen erstrecken. In mindestens einer Ausführungsform sind die Kabel 112 von einer Vielzahl unterschiedlicher Typen mit unterschiedlichen Endverbindern, unterschiedlichen Größen und unterschiedlichen Führungskonfigurationen.

[0004] In mindestens einer Ausführungsform schließen Server 106 und zusätzliche Rackkomponenten 110 eine oder mehrere Stromversorgungseinheiten (PSUs) ein, die Strom für interne Komponenten von Servern 106 und/oder zusätzlichen Rackkomponenten 110 aufnehmen und verteilen können. In mindestens einer Ausführungsform wandeln PSUs Hauptwechselstrom (AC)-Leistung in niederspannungsgeregelten Gleichstrom (DC)-Leistung um. In mindestens einer Ausführungsform schließen die Server 106 und/oder die zusätzlichen Rackkomponenten 110 mehrere Netzteile ein, die die Stromversorgung zu verschiedenen Funktionen leiten können, die mit den Servern 106 und/oder den zusätzlichen Rackkomponenten 110 verbunden sind. In mindestens einer Ausführungsform erhalten PSUs Betriebsenergie von einer oder mehreren Stromverteilungseinheiten (PDUs), die in den Racks 104 installiert sein können oder nicht. In mindestens einer Ausführungsform schließen PDUs eine oder mehrere Steckdosen ein, um elektrische Energie zu verteilen, z. B. an Racks 104 und/oder einzelne Komponenten innerhalb der Racks 104.

[0005] In mindestens einer Ausführungsform sind verschiedene Sensoren oder Sensorarrays 114 an verschiedenen Orten verteilt, die dem Rechenzentrum 100 zugeordnet sind. In mindestens einer Ausführungsform können Sensoren oder Sensoranordnungen 114 verschiedene Betriebsaspekte von Racks 104 und zugehörigen Komponenten überwachen, wie z. B. Kühlsysteme, Umgebungstemperaturen, Konnektivität von Kabeln/Komponenten, Stromverbrauch und Betriebseffizienz, um nur einige zu nennen. In mindestens einer Ausführungsform können von Sensoren oder Sensorarray 114 gesammelte Informationen ausgewertet werden, um einen oder mehrere Betriebsmängel oder - fehler innerhalb der Rechenzentren 100 zu identifizieren. In mindestens einer Ausführungsform können die Sensoren oder Sensorarrays 114 Informationen bereitstellen, die Rückschlüsse oder Schätzungen bezüglich potenzieller Betriebsmängel oder -fehler ermöglichen.

[0006] In mindestens einer Ausführungsform erhalten die Rechenzentren 100 Betriebsstrom von einem oder mehreren Stromnetzen, die miteinander verbundenen Netzen zur Stromlieferung von Erzeugern zu Verbrauchern entsprechen können. In mindestens einer Ausführungsform umfassen elektrische Netze verschiedene Komponenten, wie beispielsweise Kraftwerke, elektrische Umspannwerke, Stromübertragungsleitungen und Stromverteilungsleitungen. In mindestens einer Ausführungsform können elektrische Netze eine große Anzahl von Verbrauchern mit elektrischem Strom versorgen, die industrielle und nicht-industrielle Einheiten umfassen können, wobei der Stromverbrauch im Laufe eines Tages schwankt. In mindestens einer Ausführungsform können unterschiedliche Endbenutzer unterschiedliche Anforderungen an den Energieverbrauch haben, wie z. B. einzelne Haushalte, die wahrscheinlich weniger Energie verbrauchen als eine Fabrik. In min-

destens einer Ausführungsform überwachen elektrische Netze den Stromverbrauch über einen Zeitraum zusammen mit der Stromerzeugung, um zu bestimmen, ob ausreichend Strom für alle Benutzer verfügbar ist.

[0007] In mindestens einer Ausführungsform können die Rechenzentren 100 den Energiebedarf unter Verwendung von Energiespeicherkomponenten, wie etwa Batterien, ergänzen, die unter Verwendung eines oder mehrerer vor Ort befindlicher oder zugehöriger Energiegeneratoren wieder aufgeladen werden können, darunter zumindest Sonnenkollektoren oder Windturbinen. In mindestens einer Ausführungsform können die Rechenzentren 100 zwischen dem Liefern von Energie an Batterien zur Speicherung und dem Verbrauch von Energie aus Batterien umschalten. In mindestens einer Ausführungsform können die Rechenzentren 100 ferner Energie an Netze zurückliefern, beispielsweise in Situationen, in denen Batterien vor Ort vollständig aufgeladen sind und zugehörige Energiegeneratoren über einer Schwellenkapazität produzieren. In mindestens einer Ausführungsform können unterschiedliche Komponenten innerhalb der Rechenzentren 100 ihre eigenen individuellen Batterie-Backups haben. In mindestens einer Ausführungsform können Backups unter Verwendung von Strom aus Netzen, von vor Ort zugeordneten Energiegeneratoren oder Kombinationen davon aufgeladen werden.

[0008] In mindestens einer Ausführungsform kann die Nachfrage oder Belastung eines Stromnetzes durch Hinzufügen oder Reduzieren von Nutzern und/oder durch Anpassung des Energieverbrauchs der Nutzer angepasst werden. In mindestens einer Ausführungsform können elektrische Netze eine Spitzenlast überwachen, die einer maximalen Last entsprechen kann, die elektrische Netze bereitstellen können. In mindestens einer Ausführungsform haben elektrische Netze eine maximale Kapazität, die im Wesentlichen äquivalent zu einer Ausgabe verschiedener Generatoren sein kann, die einem elektrischen Netz zugeordnet sind. In mindestens einer Ausführungsform können elektrische Netze ferner eine feste Nennkapazität aufweisen, die einer maximalen Ausgangsleistung entspricht, die über einen gegebenen Zeitraum sofort verfügbar ist. In mindestens einer Ausführungsform sind Stromnetze nicht in der Lage, Spitzenenergieanforderungen zu erfüllen, und können infolgedessen Last absenken, indem bestimmte Verbraucher aus Stromnetzen entfernt werden, wie etwa durch Abschalten von Diensten oder Absenken der Versorgungsspannung. In mindestens einer Ausführungsform kann die Lastabsenkung die Wahrscheinlichkeit eines Ausfalls verringern, wie z. B. eines Stromausfalls, der zu kostspieligen Reparaturen und Unterbrechungen für Benutzer führen kann.

[0009] In mindestens einer Ausführungsform können verschiedene Netzwerkverwaltungstechniken eingesetzt werden, um potenzielle Spitzen oder plötzliche Bedarfssteigerungen abzuschwächen, die Ausfälle verursachen können. In mindestens einer Ausführungsform kann ein plötzlicher Lastanstieg die Netze belasten, z. B. indem die Netze beim Anfahren plötzlich an ihre Kapazitätsgrenzen stoßen. In mindestens einer Ausführungsform kann die Verwaltung plötzlicher Bedarfsanstiege für Netze schwierig sein, da es schwierig sein kann, eine Quelle eines plötzlichen Anstiegs rechtzeitig zu identifizieren, um Schaden zu mindern. In mindestens einer Ausführungsform kann eine Gruppe von Akteuren in einer konzertierten Aktion erhöhte Lasten zu besonders ausgewählten Zeiten einsetzen, um ein oder mehrere Stromnetze absichtlich zu belasten oder zu überlasten. In mindestens einer Ausführungsform kann die Identifizierung dieser Akteure schwierig sein.

[0010] In mindestens einer Ausführungsform kann ein Lastverwaltungssystem 200 eine Last aus einem Netz 202 auswerten, um eine oder mehrere Betriebsanpassungen für Rechenzentren 100 zu bestimmen, wie in **Fig. 2** dargestellt. In mindestens einer Ausführungsform schließen die Netze 202 verschiedene Komponenten ein, z. B. Stromerzeuger 204, Übertragungssysteme 206 und Verteilungssysteme 208. In mindestens einer Ausführungsform können die Stromgeneratoren 204 einem oder mehreren Kraftwerken entsprechen, wie z. B. Erdgaskraftwerken, Kohlekraftwerken, Kernkraftwerken, Solarparks, Windturbinen, geothermischen Kraftwerken und anderen. In mindestens einer Ausführungsform schließen die Stromgeneratoren 204 ferner einzelne Generatoren oder andere Generatoren ein, die verwendet werden, um Wärme oder kinetische Energie, wie beispielsweise von Wasser oder Wind, in Energie umzuwandeln, beispielsweise durch Rotation einer Turbine. In mindestens einer Ausführungsform können die Netze 202 ein Netz oder eine Anzahl verschiedener Stromgeneratoren 204 einschließen, wobei eine Gesamtkapazität durch eine Ausgabe von kombinierten Stromgeneratoren 204 der Netze 202 bestimmt wird. In mindestens einer Ausführungsform können zusätzliche Energiegeneratoren zu Netzen hinzugefügt werden, um die Kapazität zu erhöhen, wie zum Beispiel das Einbeziehen einzelner eigenständiger Turbinen, wenn erwartet wird, dass die Nachfrage die Kapazität erhöht. In mindestens einer Ausführungsform können die Stromgeneratoren 204 Informationen an einen oder mehrere Netzverwalter 210 liefern, die verwendet werden können, um einen oder mehrere Aspekte des Netzzustands oder der Betriebskapazität zu überwachen.

[0011] In mindestens einer Ausführungsform werden die Übertragungssysteme 206 für die Massenbewegung elektrischer Energie von den Stromgeneratoren 204 zu den Verteilungssystemen 208 verwendet. In

mindestens einer Ausführungsform stellen die Übertragungssysteme 206 Hochspannungsübertragungssysteme dar, die mit niedrigeren Stromstärken arbeiten, um Elektrizität über große Entfernungen zu übertragen. In mindestens einer Ausführungsform senken die Verteilersysteme 208 die Spannungen für die Verteilung an die Verbraucher ab. In mindestens einer Ausführungsform können Verbraucher in der Lage sein, Hochspannungsenergie von den Übertragungssystemen 206 zu handhaben und können Elektrizität von den Übertragungssystemen 206 ohne oder mit weniger Zwischenkomponenten empfangen. In mindestens einer Ausführungsform kann das Übertragungssystem 206 mit Redundanzen und Backup-Betriebsverfahren konfiguriert sein, um einen oder mehrere Betriebsausfälle zu berücksichtigen, um zu verhindern, dass Ausfälle oder Fehler bei der Übertragung Endbenutzer oder Generatoren beeinträchtigen.

[0012] In mindestens einer Ausführungsform schließen die Verteilungssysteme 208 Komponenten wie Umspannwerke zum Transformieren von Spannung, wie etwa zum Heraufsetzen oder Heruntersetzen von Spannungen ein. In mindestens einer Ausführungsform können vor der Verteilung an Endbenutzer mehrere Umspannwerke oder andere Komponenten verwendet werden, die verschiedene Übertragungssysteme einschließen können, wie etwa Stromleitungen, Transformatoren, Busse und andere Systeme.

[0013] In mindestens einer Ausführungsform können Fehler auf verschiedenen Ebenen der Netze 202 zu Betriebsstörungen führen. In mindestens einer Ausführungsform sind die Netzsensoren 212 über die Netze 202 verteilt und können Informationen an einen oder mehrere Netzmanager 210 liefern. In mindestens einer Ausführungsform empfangen die Netzverwalter 210 Informationen von verschiedenen anderen einzelnen Steuerungen, die mit verschiedenen Teilen der Netze 202 verbunden sind, wie etwa einer Steuerung oder einem Verwalter für einzelne Generatoren, Teile von Übertragungssystemen und anderen Unterkomponenten der Netze 202. In mindestens einer Ausführungsform übertragen die Netze 202 elektrische Energie an Rechenzentren 100, die einen oder mehrere Sensoren oder Sensorarrays 114 einschließen können, um verschiedene Aspekte der an Rechenzentren gelieferten Energie zu überwachen. In mindestens einer Ausführungsform kann Eingangsenergie unter Verwendung von einem oder mehreren Sensoren oder Sensorarrays 114A überwacht werden. In mindestens einer Ausführungsform wird der Verbrauch an einzelnen Clustern 214 unter Verwendung eines oder mehrerer Sensoren oder Sensorarrays 114B überwacht. In mindestens einer Ausführungsform wird der Verbrauch an einzelnen Racks unter Verwendung eines oder mehrerer Sensoren oder Sensorarrays 114C überwacht. In mindestens einer Ausführungsform wird der Verbrauch an einzelnen Komponenten 106, 108, 110 unter Verwendung eines oder mehrerer Sensoren oder Sensorarrays 114D überwacht.

[0014] In mindestens einer Ausführungsform kann der Verbrauch einer Last entsprechen, die angefordert oder anderweitig auf die Netze 202 angewendet wird, wobei der Gesamtverbrauch für die Rechenzentren 100 von einem oder mehreren Rechenzentrumscontrollern 216 überwacht werden kann. In mindestens einer Ausführungsform können Rechenzentrumscontrollern 216 Anweisungen an verschiedene Geräte innerhalb von Rechenzentren liefern, um einen oder mehrere Betriebsparameter als Reaktion auf den Energieverbrauch anzupassen oder zu ändern. In mindestens einer Ausführungsform kann der Energieverbrauch in Form von Kosten reguliert werden, wie z. B. Kosten, die von einem Benutzer angegeben werden. In mindestens einer Ausführungsform kann der Energieverbrauch hinsichtlich der Betriebseffizienz reguliert werden. In mindestens einer Ausführungsform kann der Energieverbrauch reguliert werden, um eine an die Netze 202 angelegte Gesamtlast zu steuern. In mindestens einer Ausführungsform kann der Energieverbrauch in Erwartung einer Zunahme oder Abnahme aufgrund von Änderungen der Arbeitsbelastungsparameter für Geräte in den Rechenzentren 100 reguliert werden.

[0015] In mindestens einer Ausführungsform empfängt der Rechenzentrumscontroller 216 Informationen von einem oder mehreren Netzsensoren 212 und/oder einem oder mehreren Netzverwaltern 210, um die verfügbare Kapazität in den Netzen zu ermitteln. In mindestens einer Ausführungsform kann der Rechenzentrumscontroller 216 Operationen unter einem Schwellenbetrag regulieren oder anderweitig einschränken, der verfügbaren Kapazitäten oder einem Prozentsatz verfügbarer Kapazitäten entsprechen kann. In mindestens einer Ausführungsform kann der Rechenzentrumscontroller 216 Operationen regulieren, um zu verhindern, dass eine Schwellenmenge an verfügbarer Netzkapazität überschritten wird. In mindestens einer Ausführungsform kann der Rechenzentrumscontroller 216 Arbeitslasten beenden oder einschränken, um Betriebsparameter innerhalb einer oder mehrerer Grenzen zu halten. In mindestens einer Ausführungsform kann der Rechenzentrumscontroller 216 Start- oder Abschaltvorgänge zumindest teilweise basierend auf der Netzkapazität bestimmen. In mindestens einer Ausführungsform kann der Rechenzentrumscontroller 216 eine oder mehrere Operationen basierend auf Trends ausführen oder beenden, die der Netzkapazität zugeordnet sind.

[0016] In mindestens einer Ausführungsform erstrecken sich die Netze 202 über ein großes geografisches Gebiet und stellen einem oder mehreren Rechenzentren 100A-100N Strom bereit, wie in **Fig. 3** dargestellt. In mindestens einer Ausführungsform werden die Rechenzentren 100A-100N von einer gemeinsamen Einheit betrieben, die untereinander kommuniziert, so dass Informationen aus verschiedenen Rechenzentren gemeinsam genutzt und zusammengefasst werden können, um unterschiedliche Betriebsbedingungen zu bewerten. In mindestens einer Ausführungsform kann ein erstes Rechenzentrum 100A einen Rechenzentrumscontroller 216A einschließen, die als Betreiber oder Gesamtverwalter für Gruppen von Rechenzentren 100A - 100N agiert. In mindestens einer Ausführungsform empfängt der Rechenzentrumscontroller 216A Betriebsinformationen von anderen Rechenzentren und bestimmt die verfügbare Last für Gruppen von Rechenzentren 100A-100N zumindest teilweise basierend auf einer verfügbaren Kapazität für das Netz 202. In mindestens einer Ausführungsform kann die verfügbare Kapazität vom Netz 202 zumindest teilweise auf einem geografischen Standort der Rechenzentren 100A-100N in Bezug auf das Netz 202 basieren. In mindestens einer Ausführungsform kann die verfügbare Kapazität unabhängig vom geografischen Standort der gesamten verfügbaren Kapazität vom Netz 202 entsprechen, kann aber ferner Leitungsverluste und Übertragungsverluste berücksichtigen, wenn bestimmt wird, wie Kapazität verschiedenen an das Netz 202 gekoppelten Komponenten bereitgestellt werden kann. In mindestens einer Ausführungsform kann der Rechenzentrumscontroller 216A kommende oder zukünftige geplante Arbeitslasten auswerten sowie Trends bei den Arbeitslasten betrachten, um zu bestimmen, ob die Rechenzentren 100A-100N an einem bestimmten Punkt das Netz 202 überanstrengen oder anderweitig überlasten werden. In mindestens einer Ausführungsform kann der Rechenzentrumscontroller 216A Anweisungen zum Anpassen von Betriebsparametern in anderen Rechenzentren empfehlen oder senden, um die Wahrscheinlichkeit einer Überlastung des Netzes 202 zu verhindern oder zu verringern.

[0017] In mindestens einer Ausführungsform können die Rechenzentren 100A - 100N Jobs, Arbeitslasten oder Operationen haben, die für einen bestimmten Zeitraum geplant sind. In mindestens einer Ausführungsform können die Rechenzentren 100A - 100N den Energieverbrauch zumindest teilweise basierend auf geplanten Arbeitsbelastungen, Wetterberichten, Arbeitslasttrends und anderen Informationen vorhersagen. In mindestens einer Ausführungsform kann wärmeres Wetter beispielsweise zu höheren Energiekosten für die Kühlung führen, und infolgedessen können die Rechenzentren 100A - 100N den Energieverbrauch mit der Verfügbarkeit des Netzes 202 sowie Einschränkungen oder Vorlieben von Benutzern planen oder budgetieren. In mindestens einer Ausführungsform können die Rechenzentren 100A-100N das Netz 202 präventiv beim Lastabwurf unterstützen, indem sie einen oder mehrere Knoten oder Cluster abschalten oder herunterfahren, wenn sie feststellen, dass die Betriebslast der Rechenzentren 100A-100N oder eines Rechenzentrums die verfügbare Kapazität des Netzes 202 übersteigt. In mindestens einer Ausführungsform kann der Rechenzentrumscontroller 216A den Betrieb so einschränken oder regeln, dass die Rechenzentren 100A-100N einen Schwellenwert oder einen Prozentsatz der Gesamtlast des Netzes 202 nicht überschreiten, z. B. eine Gesamtlast zu einem bestimmten Zeitpunkt, einen Durchschnitt über einen bestimmten Zeitraum oder andere Zeiträume.

[0018] In mindestens einer Ausführungsform können ein oder mehrere Rechenzentrumscontroller 216 als Energieüberwacher verwendet werden, wie zum Beispiel der Rechenzentrumscontroller 216A. In mindestens einer Ausführungsform kann der Rechenzentrumscontroller 216A Informationen von dem Netz 202 empfangen, wie beispielsweise Informationen, die einem oder mehreren Netzsensoren 212 zugeordnet sind, die unter anderen Arten von Informationen Lastinformationen bereitstellen. In mindestens einer Ausführungsform kann der Rechenzentrumscontroller 216A Trends in der Energienutzung analysieren oder bestimmen, z. B. zusätzlichen Verbrauch während heißerer Tageszeiten als in der Nacht, um die verfügbare Kapazität zu verschiedenen Zeiten zu bestimmen oder vorherzusagen. In mindestens einer Ausführungsform kann der Rechenzentrumscontroller 216A unterschiedliche Arbeitslasten umplanen oder neu anordnen, um unterschiedliche Knoten oder Cluster zu unterschiedlichen Zeiten zu betreiben, um Zeiten mit höherer Kapazität zu entsprechen. In mindestens einer Ausführungsform kann der Rechenzentrumscontroller 216A unterschiedliche Zeiten für das Anfahren oder Herunterfahren festlegen, die zumindest teilweise auf der aktuell verfügbaren Kapazität und/oder der erwarteten zukünftigen verfügbaren Kapazität basieren. In mindestens einer Ausführungsform kann der Rechenzentrumscontroller 216A eine oder mehrere Leistungsgrenzen ändern, um den Stromverbrauch des Knotens zu erhöhen oder zu verringern. In mindestens einer Ausführungsform kann der Rechenzentrumscontroller 216A Knoten ein- oder ausschalten. In mindestens einer Ausführungsform kann der Rechenzentrumscontroller 216A Strom von einer Sicherungsstromversorgung absenken oder beziehen. In mindestens einer Ausführungsform kann der Rechenzentrumscontroller 216A Arbeitslasten ermitteln, die schädliche Leistungssignaturen erzeugen. In mindestens einer Ausführungsform kann der Rechenzentrumscontroller 216 Dummy-Arbeitslasten auf ungenutzten Knoten aktivieren, um ein schnelles Absinken des Stromverbrauchs zu verhindern. In mindestens einer Ausführungsform können die

von Rechenzentrumssteuerungen 216 ergriffenen Maßnahmen zumindest teilweise auf Netzinformationen basieren, um die Wahrscheinlichkeit eines Netzausfalls zu verhindern oder zu verringern.

[0019] In mindestens einer Ausführungsform kann der Energieverbrauch, der auch als Last bezeichnet werden kann, grafisch dargestellt werden, wie in den **Fig. 4A-4D** gezeigt. In mindestens einer Ausführungsform veranschaulicht eine erste Darstellung 400 eine Last 402 über die Zeit 404, wobei die Last 402 auf einer y-Achse dargestellt ist und die Zeit 404 auf einer z-Achse dargestellt ist. In mindestens einer Ausführungsform wird die Netzlast 406 durch eine durchgezogene Linie dargestellt. In mindestens einer Ausführungsform ist die Netzlast 406 eine vereinfachte grafische Darstellung, die keine Spitzen oder allmählichen Anstiege oder Abfälle im Laufe der Zeit einschließt, und sollte nicht als einschränkend, sondern als veranschaulichendes Beispiel interpretiert werden. In mindestens einer Ausführungsform kann die Netzlast 406 unterschiedliche Spitzen und Täler sowie allmähliche Änderungen umfassen, wenn unterschiedliche Lasten ans Netz gehen. In mindestens einer Ausführungsform wird die Netzlast 406 unter einer Netzgesamtkapazität 408 gehalten, die als unterbrochene Linie dargestellt ist. In mindestens einer Ausführungsform veranschaulichen unterschiedliche Zeitsegmente 410A - 410C unterschiedliche Nutzungsraten, wenn die Netzlast 408 zunimmt oder abnimmt. In mindestens einer Ausführungsform stellt das Zeitsegment 410A einen Morgenzeitraum dar, in dem die Temperaturen kühler sein können oder weniger Geschäfte oder Dienste in Betrieb sind, und infolgedessen ist die Netzlast 406 über das Zeitsegment 410A geringer als die Netzlast 406 über das Zeitsegment 410B, was an einem Nachmittag stattfinden kann, wenn die Temperaturen höher sein können und mehr Unternehmen in Betrieb sind. In mindestens einer Ausführungsform kann das Zeitsegment 410C einen Abend darstellen, an dem die Netzlast 406 erneut abnimmt, beispielsweise weil Unternehmen den Betrieb einstellen oder die Temperaturen fallen. In mindestens einer Ausführungsform können zusätzliche oder weniger Zeitabschnitte ausgewertet werden. In mindestens einer Ausführungsform weist jedes Segment 410A-410C eine Kapazitätsgrenze 412A-412C auf, die für die verfügbare Last im Vergleich zur Gesamtkapazität 408 des Netzes repräsentativ ist. In mindestens einer Ausführungsform weist jedes Segment 410A-410C eine
Betriebskapazitätsgrenze 414A-414C auf, die repräsentativ für eine Betriebslast oder einen Schwellenwert im Vergleich zu einer Betriebsnetzkapazität 416 ist. In mindestens einer Ausführungsform ist die Betriebsnetzkapazität 416 geringer als die Gesamtnetzkapazität 408, z. B. um einen Schwellenwert, um plötzliche Spannungsspitzen oder Änderungen des Energieverbrauchs aufzufangen, was dem Netz 202 einen Puffer oder Zeit verschafft, um zusätzliche Erzeugung hinzuzufügen, Verbraucher abzuschalten oder andere Verwaltungsoperationen durchzuführen, um Überlastungen des Netzes 202 zu verhindern.

[0020] In mindestens einer Ausführungsform wird die Betriebslast 418 des Rechenzentrums durch eine gestrichelte Linie dargestellt, wobei sich die Betriebslast 418 des Rechenzentrums im Laufe der Zeit ändern kann, z. B. innerhalb jedes Segments 410A-410C, wobei verschiedene Segmente mehr oder weniger Last ziehen können, die zumindest teilweise auf verschiedenen Betriebsparametern oder Benutzeroperationen basiert. In mindestens einer Ausführungsform kann die Betriebslast 418 des Rechenzentrums im Laufe der Zeit verwaltet und angepasst werden, um Änderungen der Netzlast 406 anzupassen, wie z. B. das Erhöhen des Betriebs während Perioden mit Überkapazität oder das Abschalten oder Entfernen des Betriebs während Perioden mit reduzierter Kapazität.

[0021] In mindestens einer Ausführungsform könnte der Rechenzentrumsbetrieb so geplant werden, dass er die Gesamtkapazität des Netzes 408 belastet oder anderweitig überschreitet, wie in **Fig. 4B** gezeigt. In mindestens einer Ausführungsform veranschaulicht eine zweite Darstellung 420 die Netzlast 406 und die Gesamtkapazität 408 des Netzes. In mindestens einer Ausführungsform kann die Rechenzentrumsbetriebslast 418 die Netzgesamtkapazität 408 übersteigen, wie etwa während des Zeitsegments 41 0B. In mindestens einer Ausführungsform kann ein solcher schneller Anstieg des Verbrauchs unter anderem mit der gleichzeitigen Aktivierung eines oder mehrerer Knoten verbunden sein, was auf einen schädlichen Angriff hinweisen kann. In mindestens einer Ausführungsform können ein oder mehrere schlechte Akteure Betriebskapazität in mehreren Rechenzentren erwerben, die auf einem einzigen Netz betrieben werden. In mindestens einer Ausführungsform können ein oder mehrere schlechte Akteure den Betrieb für einen bestimmten, sich überschneidenden Zeitraum konfigurieren, sodass der Betrieb die Gesamtkapazität des Netzes 408 belasten oder anderweitig überschreiten würde. In mindestens einer Ausführungsform könnten die Netze in einer konzertierten Aktion gezielt angegriffen werden, um die Netze zum Erliegen zu bringen oder um den Nutzern Unannehmlichkeiten zu bereiten, die bei der Steuerung und Stabilisierung der Netzlasten ausfallen.

[0022] In mindestens einer Ausführungsform kann der Betrieb des Rechenzentrums an eine erwartete Last angepasst werden, die die Gesamtkapazität des Netzes 408 übersteigt, wie in **Fig. 4C** dargestellt. In mindestens einer Ausführungsform veranschaulicht eine dritte Darstellung 420 die Netzlast 406 und die Gesamtka-

pazität 408 des Netzes. In mindestens einer Ausführungsform wird die Rechenzentrumsbetriebslast 418 als abgestimmte Rechenzentrumsbetriebslast 452 dargestellt, wobei die abgestimmte Rechenzentrumsbetriebslast 452 die Netzgesamtkapazität 408 nicht überschreitet. In mindestens einer Ausführungsform können eine oder mehrere Arbeitslasten oder Operationen modifiziert werden, um eine abgestimmte Rechenzentrumsbetriebslast 452 zu ermöglichen, wie etwa durch Abschalten einer oder mehrerer Arbeitslasten, Ändern von Leistungsgrenzen oder andere Optionen. In mindestens einer Ausführungsform kann ein starker Anstieg 454 beginnend bei dem Zeitsegment 410B erkannt werden, wie z. B. die Verwendung eines oder mehrerer Sensoren oder Sensorarrays, um Operationen in einem oder mehreren Rechenzentren zu überwachen. In mindestens einer Ausführungsform können ein oder mehrere Rechenzentrumscontroller feststellen, dass die erwartete Last die Gesamtkapazität 408 des Netzes übersteigt, und die Last drosseln, um den Betrieb auf oder unter der Gesamtkapazität 408 des Netzes zu halten, was genügend Zeit für den Netzbetrieb bieten kann, um die Betriebseinstellungen anzupassen, oder für das Rechenzentrum, um eine oder mehrere Maßnahmen zur Verringerung der Betriebslast einzuleiten. In mindestens einer Ausführungsform schließen die Minderungsmaßnahmen die Verringerung der Anzahl der gleichzeitig ans Netz gehenden Knoten, das Hoch- oder Herunterfahren des Betriebs oder die Aussetzung des Betriebs ein, z. B. bei Leistungssignaturen, die mit bösartigen Akteuren oder schädlichen Absichten in Verbindung gebracht werden.

[0023] In mindestens einer Ausführungsform können die Rechenzentrumscontroller auch Dummy-Arbeitslasten aktivieren, um den Stromverbrauch im Wesentlichen konstant zu halten oder zu verringern, wie in **Fig. 4D** dargestellt. In mindestens einer Ausführungsform enthält eine dritte Darstellung 470 eine Rechenzentrumsbetriebslast 472, die einen Abfall 474 während des Zeitsegments 410B enthält, bevor der Betrieb bei einer niedrigeren Last fortgesetzt wird. In mindestens einer Ausführungsform kann eine drastische oder plötzliche Reduzierung des Stromverbrauchs für Netzverwaltungszwecke unerwünscht sein, und infolgedessen können eine oder mehrere Dummy-Arbeitslasten aktiviert werden, wie durch die modifizierte Rechenzentrums-Betriebslast 476 gezeigt. In mindestens einer Ausführungsform erzeugt das Aktivieren von Dummy-Arbeitslasten eine schrittweise Verringerung der Betriebslast, was für die Netzverwaltung wünschenswert sein kann. In mindestens einer Ausführungsform können Dummy-Arbeitslasten die Last auf früheren Niveaus halten, wodurch die Last auf den Netzen aufrechterhalten wird und auch die Betriebsniveaus im Rechenzentrum für zukünftige Arbeitslasten aufrechterhalten werden, wodurch die Wahrscheinlichkeit verringert wird, dass die Last für zukünftige Arbeitslasten nicht verfügbar ist.

[0024] In mindestens einer Ausführungsform kann ein Prozess 500 zur Überwachung der Rechenzentrumslast wie in **Fig. 5A** dargestellt durchgeführt werden. In mindestens einer Ausführungsform werden ein oder mehrere Netzbetriebsparameter bestimmt 502. In mindestens einer Ausführungsform können Netzbetriebsparameter aus Sensordaten erfasst werden, die über Netze verteilt sind, und können zumindest teilweise verfügbare Kapazität, aktuelle Last, erwartete Last oder andere Informationen einschließen. In mindestens einer Ausführungsform können Netzbetriebsparameter ferner Informationen bezüglich verfügbarer Kapazität liefern. In mindestens einer Ausführungsform werden ein oder mehrere Rechenzentrumsbetriebsparameter bestimmt 504. In mindestens einer Ausführungsform können ein oder mehrere Rechenzentrumsbetriebsparameter einer Anzahl von laufenden Knoten oder Clustern, dem aktuellen Stromverbrauch, dem erwarteten Stromverbrauch, bestimmten GPU/CPU-Betriebsparametern, dem Kühlungsstromverbrauch und anderen Informationen entsprechen. In mindestens einer Ausführungsform kann eine verfügbare Netzkapazität bestimmt werden 506. In mindestens einer Ausführungsform kann die verfügbare Netzkapazität einer zum Verbrauch verfügbaren Menge elektrischer Energie entsprechen. In mindestens einer Ausführungsform kann die verfügbare Netzkapazität eine Differenz zwischen einer aktuellen Last und einer maximalen Last oder einer eingestellten Last sein. In mindestens einer Ausführungsform kann die verfügbare Netzkapazität zumindest teilweise auf den Kosten zum Empfangen von Energie basieren. In mindestens einer Ausführungsform wird bestimmt, dass eine zukünftige Rechenzentrumslast eine verfügbare Netzkapazität 508 übersteigt. In mindestens einer Ausführungsform kann die künftige Rechenzentrumslast durch geplante Arbeitslasten und/oder durch den Vergleich von Betriebsparametern mit der erwarteten oder bereitgestellten Kapazität in einem oder mehreren Netzen bestimmt werden. In mindestens einer Ausführungsform werden ein oder mehrere Rechenzentrumsbetriebsparameter angepasst 510. In mindestens einer Ausführungsform kann die Anpassung eines oder mehrerer Rechenzentrumsbetriebsparameter dem Ändern von Betriebseinstellungen, dem Herunterfahren von Knoten oder Clustern, dem Ändern eines Timings eines Betriebs oder anderen Anpassungen entsprechen.

[0025] In mindestens einer Ausführungsform kann ein Prozess 550 zur Verwaltung der Rechenzentrumslast wie in **Fig. 5B** dargestellt durchgeführt werden. In mindestens einer Ausführungsform wird eine verfügbare Netzkapazität bestimmt 552. In mindestens einer Ausführungsform kann die Netzkapazität berechnet werden oder Informationen können von einem Netzbetreiber bereitgestellt werden. In mindestens einer Ausführungs-

form kann eine erwartete Rechenzentrumslast bestimmt werden 554. In mindestens einer Ausführungsform können die erwarteten Rechenzentrumslasten einem oder mehreren Zeiträumen mit erwarteten Lasten entsprechen, die anstehenden Operationen entsprechen oder zumindest teilweise auf Schätzungen beruhen können, die durch Auswertungen gesammelter Daten erzeugt wurden. In mindestens einer Ausführungsform kann bestimmt werden, ob die erwartete Last innerhalb eines Schwellenwerts der verfügbaren Netzkapazität 556 liegt. In mindestens einer Ausführungsform kann eine solche Bestimmung den Rechenzentrumsoperationen Informationen liefern, ob ausreichend Kapazität für den Betrieb zukünftiger Arbeitslasten verfügbar ist. In mindestens einer Ausführungsform wird bestimmt, dass die erwartete Rechenzentrumslast nicht innerhalb eines Schwellenbetrags liegt, und eine oder mehrere zukünftige Rechenzentrumsoperationen werden angepasst 558. In mindestens einer Ausführungsform kann ein überschrittener Schwellenwert dem Vorhandensein einer unzureichenden Kapazität zum Unterstützen von Rechenzentrumsoperationen entsprechen. In mindestens einer Ausführungsform können Modifikationen an zukünftigen Rechenzentrumsoperationen eine erwartete Rechenzentrumslast ändern, so dass Netze eine ausreichende Kapazität haben. In mindestens einer Ausführungsform werden eine oder mehrere zukünftige Rechenzentrumsoperationen ausgeführt 560. In mindestens einer Ausführungsform wird die Ausführung zukünftiger Rechenzentrumsoperationen nach Anpassungen an einer oder mehreren Operationen oder nach der Bestimmung, dass verfügbare Kapazität vorhanden ist, durchgeführt.

[0026] In mindestens einer Ausführungsform kann ein Prozess 570 zur Verwaltung der Rechenzentrumslast wie in **Fig. 5C** dargestellt durchgeführt werden. In mindestens einer Ausführungsform werden ein oder mehrere Rechenzentrumsbetriebsparameter empfangen 572. In mindestens einer Ausführungsform können ein oder mehrere Rechenzentrumsbetriebsparameter einer aktuellen Last, einer erwarteten Last oder einer voraussichtlichen Last entsprechen, die zumindest teilweise auf einem oder mehreren maschinellen Lernmodellen basieren können, die trainiert wurden, um eine durchschnittliche oder erwartete Last während eines oder mehrerer unterschiedlicher Betriebszustände zu bestimmen. In mindestens einer Ausführungsform wird eine Leistungssignatur bestimmt 574, die zumindest teilweise auf einem oder mehreren Rechenzentrumsbetriebsparametern basieren kann. In mindestens einer Ausführungsform kann eine Leistungssignatur einer über einen Zeitraum verbrauchten Leistungsmenge entsprechen. In mindestens einer Ausführungsform kann eine Leistungssignatur angeben, wie schnell der Stromverbrauch ansteigt. In mindestens einer Ausführungsform kann eine Leistungssignatur einer Tageszeit entsprechen, die dem Stromverbrauch entspricht. In mindestens einer Ausführungsform können zahlreiche Parameter verwendet werden, um eine Leistungssignatur zu bestimmen. In mindestens einer Ausführungsform werden Leistungssignaturen für ein oder mehrere Merkmale evaluiert 576. In mindestens einer Ausführungsform können ein oder mehrere Merkmale Anomalien oder verschiedenen Verbrauchsverhalten entsprechen, die einen Schwellenwert überschreiten oder außerhalb eines erwarteten Bereichs liegen. In mindestens einer Ausführungsform können ein oder mehrere Merkmale dem Verbrauch in zusätzlichen Rechenzentren entsprechen, wie etwa einer Reihe von Operationen, die über mehrere Rechenzentren mit ähnlichen Eigenschaften verteilt sind.

[0027] In mindestens einer Ausführungsform werden Leistungssignaturen ausgewertet, um zu bestimmen, ob die Leistungssignatur einer schädlichen Leistungssignatur 578 entspricht oder nicht. In mindestens einer Ausführungsform kann eine schädliche Leistungssignatur mit bekannten schädlichen Signaturen in einer oder mehreren Datenbanken verglichen werden, wobei eine ermittelte Leistungssignatur mit einem oder mehreren Merkmalen einer bekannten schädlichen Leistungssignatur verglichen werden kann und eine Leistungssignatur bei ausreichender Übereinstimmung als schädlich eingestuft werden kann. In mindestens einer Ausführungsform werden, wenn festgestellt wird, dass eine Leistungssignatur schädlich ist, eine oder mehrere Arbeitslasten, die einer schädlichen Leistungssignatur zugeordnet sind, beendet 580. In mindestens einer Ausführungsform kann, wenn bestimmt wird, dass eine Leistungssignatur nicht schädlich ist, eine Bewertung durchgeführt werden, um zu bestimmen, ob die Last, die einer oder mehreren Arbeitslasten zugeordnet ist, innerhalb eines akzeptablen Schwellenwerts 582 liegt. In mindestens einer Ausführungsform können Bewertungen anhand eines Schwellenwerts der Bestimmung entsprechen, ob die Ausführung einer oder mehrerer Arbeitslasten eine verfügbare Kapazität eines Netzes überschreiten wird. In mindestens einer Ausführungsform wird bestimmt, dass eine Leistungssignatur einen Schwellenwert überschreitet, und eine oder mehrere Rechenzentrumsarbeitslasten werden modifiziert 584. In mindestens einer Ausführungsform wird bestimmt, dass eine Leistungssignatur einen Schwellenwert nicht überschreitet, und eine oder mehrere Arbeitslasten werden ausgeführt 586.

SERVER UND RECHENZENTREN

[0028] Die folgenden Figuren legen ohne Einschränkung beispielhafte Netzwerkserver und rechenzentrumbasierte Systeme dar, die verwendet werden können, um mindestens eine Ausführungsform zu implementieren.

[0029] **Fig. 6** zeigt ein verteiltes System 600 gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform beinhaltet das verteilte System 600 eine oder mehrere Client-Rechenvorrichtungen 602, 604, 606 und 608, die konfiguriert sind, um eine Client-Anwendung, wie etwa einen Webbrowser, einen proprietären Client und/oder Variationen davon, über ein oder mehrere Netzwerke 610 auszuführen und zu betreiben. In mindestens einer Ausführungsform kann der Server 612 über das Netzwerk 610 mit entfernten Client-Rechenvorrichtungen 602, 604, 606 und 608 kommunikativ gekoppelt sein.

[0030] In mindestens einer Ausführungsform kann der Server 612 ausgelegt sein, um einen oder mehrere Dienste oder eine oder mehrere Softwareanwendungen auszuführen, wie etwa Dienste und Anwendungen, die Sitzungsaktivitäten des Zugriffs mit einmaliger Anmeldung (single sign-on - SSO) über mehrere Rechenzentren hinweg verwalten können. In mindestens einer Ausführungsform kann der Server 612 auch andere Dienste bereitstellen oder können Softwareanwendungen nicht virtuelle und virtuelle Umgebungen beinhalten. In mindestens einer Ausführungsform können diese Dienste Benutzern der Client-Rechenvorrichtungen 602, 604, 606 und/oder 608 als webbasierte oder Cloud-Dienste oder im Rahmen eines Software-as-a-Service-(SaaS-)Modells angeboten werden. In mindestens einer Ausführungsform können Benutzer, die Client-Rechenvorrichtungen 602, 604, 606 und/oder 608 betreiben, wiederum eine oder mehrere Client-Anwendungen nutzen, um mit dem Server 612 zu interagieren, um durch diese Komponenten bereitgestellte Dienste zu nutzen.

[0031] In mindestens einer Ausführungsform sind die Softwarekomponenten 618, 620 und 622 des Systems 600 auf dem Server 612 implementiert. In mindestens einer Ausführungsform können eine oder mehrere Komponenten des Systems 600 und/oder durch diese Komponenten bereitgestellte Dienste auch durch eine oder mehrere der Client-Rechenvorrichtungen 602, 604, 606 und/oder 608 implementiert sein. In mindestens einer Ausführungsform können Benutzer, die Client-Rechenvorrichtungen betreiben, dann eine oder mehrere Client-Anwendungen nutzen, um durch diese Komponenten bereitgestellte Dienste zu verwenden. In mindestens einer Ausführungsform können diese Komponenten in Hardware, Software, Firmware oder Kombinationen davon implementiert sein. Es versteht sich, dass verschiedene unterschiedliche Systemkonfigurationen möglich sind, die sich von dem verteilten System 600 unterscheiden können. Die in **Fig. 6** gezeigte Ausführungsform ist somit mindestens eine Ausführungsform eines verteilten Systems zum Implementieren einer Ausführungsform eines Systems und soll nicht einschränkend sein.

[0032] In mindestens einer Ausführungsform können die Client-Rechenvorrichtungen 602, 604, 606 und/oder 608 verschiedene Arten von Computersystemen beinhalten. In mindestens einer Ausführungsform kann eine Client-Rechenvorrichtung transportable Vorrichtungen (z. B. ein iPhone®, Mobiltelefon, ein iPad®, ein Computertablet, einen persönlichen digitalen Assistenten (personal digital assistant - PDA)) oder tragbare Vorrichtungen (z. B. ein Google Glass® mit am Kopf montierter Anzeige) beinhalten, auf denen Software wie Microsoft Windows Mobile® und/oder eine Vielzahl von mobilen Betriebssystemen wie iOS, Windows Phone, Android, BlackBerry 10, Palm OS und/oder Variationen davon ausgeführt wird. In mindestens einer Ausführungsform können Vorrichtungen verschiedene Anwendungen unterstützen, wie z. B. verschiedene internetbezogene Anwendungen, E-Mail, Kurznachrichtendienst-(SMS)-Anwendungen und können verschiedene andere Kommunikationsprotokolle verwenden. In mindestens einer Ausführungsform können Client-Rechenvorrichtungen auch Allzweck-Personalcomputer beinhalten, darunter mittels mindestens einer Ausführungsform Personalcomputer und/oder Laptop-Computer, auf denen verschiedene Versionen von Microsoft Windows®, Apple Macintosh® und/oder Linux-Betriebssysteme laufen.

[0033] In mindestens einer Ausführungsform können Client-Rechenvorrichtungen Workstation-Computer sein, auf denen ein beliebiges von einer Vielfalt von handelsüblichen UNIX ® oder UNIX-ähnlichen Betriebssystemen läuft, darunter ohne Einschränkung eine Vielfalt von GNU/Linux-Betriebssystemen, wie beispielsweise Google Chrome OS. In mindestens einer Ausführungsform können Client-Rechenvorrichtungen auch elektronische Vorrichtungen beinhalten, wie etwa einen Thin-Client-Computer, ein internetfähiges Spielsystem (z. B. eine Microsoft Xbox-Spielekonsole mit oder ohne Kinect®-Gesteneingabevorrichtung) und/oder eine persönliche Nachrichtenvorrichtung, die in der Lage ist, über Netzwerk(e) 610 zu kommunizieren. Auch wenn das verteilte System 600 in **Fig. 6** mit vier Client-Rechenvorrichtungen gezeigt ist, kann eine beliebige

Anzahl von Client-Rechenvorrichtungen unterstützt werden. Andere Vorrichtungen, wie etwa Vorrichtungen mit Sensoren usw., können mit dem Server 612 interagieren.

[0034] In mindestens einer Ausführungsform kann/können das/die Netzwerk(e) 610 in dem verteilten System 600 jede Art von Netzwerk sein, das Datenkommunikationen unter Verwendung eines beliebigen einer Vielfalt von verfügbaren Protokollen unterstützen kann, einschließlich ohne Einschränkung TCP/IP (transmission control protocol/Internet protocol - Übertragungssteuerungsprotokoll/Internetprotokoll), SNA (Systemnetzwerkarchitektur), IPX (Internet Packet Exchange), AppleTalk und/oder Variationen davon. In mindestens einer Ausführungsform kann/können das/die Netzwerk(e) 610 ein lokales Netzwerk (LAN), Netzwerke basierend auf Ethernet, Token-Ring, ein Weitverkehrsnetzwerk, Internet, ein virtuelles Netzwerk, ein virtuelles privates Netzwerk (VPN), ein Intranet, ein Extranet, ein öffentliches Telefonnetz (PSTN), ein Infrarotnetz, ein drahtloses Netz (z. B. ein Netzwerk, das mit der IEEE 802.11 Protokollfamilie (Institute of Electrical and Electronics Engineers - IEEE), Bluetooth® und/oder jedem anderen drahtlosen Protokoll arbeitet) und/oder eine beliebige Kombination dieser und/oder anderer Netzwerke sein.

[0035] In mindestens einer Ausführungsform kann der Server 612 aus einem oder mehreren Allzweckcomputern, spezialisierten Servercomputern (einschließlich in mindestens einer Ausführungsform PC-(Personal-Computer)-Servern, UNIX®-Servern, Midrange-Servern, Großrechnern, Rack-montierten Servern usw.), Serverfarmen, Serverclustern oder jeder anderen geeigneten Anordnung und/oder Kombination bestehen. In mindestens einer Ausführungsform kann der Server 612 eine oder mehrere virtuelle Maschinen, auf denen virtuelle Betriebssysteme laufen, oder andere Rechenarchitekturen mit Virtualisierung beinhalten. In mindestens einer Ausführungsform können ein oder mehrere flexible Pools logischer Speichervorrichtungen virtualisiert werden, um virtuelle Speichervorrichtungen für einen Server zu verwalten. In mindestens einer Ausführungsform können virtuelle Netzwerke durch den Server 612 unter Verwendung von softwaredefinierten Netzwerken gesteuert werden. In mindestens einer Ausführungsform kann der Server 612 ausgelegt sein, um einen oder mehrere Dienste oder eine oder mehrere Softwareanwendungen auszuführen.

[0036] In mindestens einer Ausführungsform kann der Server 612 ein beliebiges Betriebssystem sowie ein beliebiges im Handel erhältliches Server-Betriebssystem ausführen. In mindestens einer Ausführungsform kann der Server 612 auch eine beliebige einer Vielfalt zusätzlicher Serveranwendungen und/oder Mid-Tier-Anwendungen ausführen, einschließlich HTTP-Server (Hypertext Transport Protocol), FTP-Server (File Transfer Protocol), CGI-Server (Common Gateway Interface), JAVAO-Server, Datenbankserver und/oder Variationen davon. In mindestens einer Ausführungsform beinhalten beispielhafte Datenbankserver ohne Einschränkung diejenigen, die im Handel von Oracle, Microsoft, Sybase, IBM (International Business Machines) erhältlich sind, und/oder Variationen davon.

[0037] In mindestens einer Ausführungsform kann der Server 612 eine oder mehrere Anwendungen beinhalten, um Datenfeeds und/oder Ereignisaktualisierungen zu analysieren und zu konsolidieren, die von Benutzern der Client-Rechenvorrichtungen 602, 604, 606 und 608 empfangen werden. In mindestens einer Ausführungsform können Datenfeeds und/oder Ereignisaktualisierungen ohne Einschränkung darauf Twitter®-feeds, Facebook®-Updates oder Echtzeit-Updates, die von einer oder mehreren Informationsquellen Dritter empfangen werden, und kontinuierliche Datenströme, die Echtzeitereignisse in Bezug auf Sensordatenanwendungen, Finanzticker, Netzwerkleistungsmesstools (z. B. Netzwerküberwachungs- und Verkehrsmanagementanwendungen), Clickstream-Analysetools, Automobilverkehrsüberwachung und/oder Variationen davon beinhalten können, beinhalten. In mindestens einer Ausführungsform kann der Server 612 auch eine oder mehrere Anwendungen beinhalten, um Datenfeeds und/oder Ereignisaktualisierungen über eine oder mehrere Anzeigevorrichtungen der Client-Rechenvorrichtungen 602, 604, 606 und 608 anzuzeigen.

[0038] In mindestens einer Ausführungsform kann das verteilte System 600 auch eine oder mehrere Datenbanken 614 und 616 beinhalten. In mindestens einer Ausführungsform können Datenbanken einen Mechanismus zum Speichern von Informationen bereitstellen, wie etwa Benutzerinteraktionsinformationen, Nutzungsmusterinformationen, Anpassungsregelinformationen und andere Informationen. In mindestens einer Ausführungsform können sich die Datenbanken 614 und 616 an einer Vielfalt von Stellen befinden. In mindestens einer Ausführungsform können sich eine oder mehrere der Datenbanken 614 und 616 auf einem nichttransitorischen Speichermedium lokal beim (und/oder resident im) Server 612 befinden. In mindestens einer Ausführungsform können die Datenbanken 614 und 616 vom Server 612 entfernt sein und mit dem Server 612 über eine netzwerkbasierte oder dedizierte Verbindung kommunizieren. In mindestens einer Ausführungsform können sich die Datenbanken 614 und 616 in einem Speicherbereichsnetzwerk (storage-area network - SAN) befinden. In mindestens einer Ausführungsform können alle erforderlichen Dateien zum Durchführen von Funktionen, die dem Server 612 zugeschrieben werden, je nach Bedarf lokal auf dem Ser-

ver 612 und/oder entfernt gespeichert sein. In mindestens einer Ausführungsform können die Datenbanken 614 und 616 relationale Datenbanken beinhalten, wie etwa Datenbanken, die ausgelegt sind, um Daten als Reaktion auf SQL-formatierte Befehle zu speichern, zu aktualisieren und abzurufen.

[0039] Fig. 7 zeigt ein beispielhaftes Rechenzentrum 700 gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform beinhaltet das Rechenzentrum 700 ohne Einschränkung eine Rechenzentrumsinfrastrukturschicht 710, eine Framework-Schicht 720, eine Softwareschicht 730 und eine Anwendungsschicht 740.

[0040] In mindestens einer Ausführungsform kann die Rechenzentrumsinfrastrukturschicht 710 wie in **Fig. 7** dargestellt einen Ressourcen-Orchestrator 712, gruppierte Rechenressourcen 714 und Knoten-Rechenressourcen („Knoten-C.R.s“) 716(1)-716(N) beinhalten, wobei „N“ eine beliebige ganze positive Zahl darstellt. In mindestens einer Ausführungsform können die Knoten-C.R.s 716(1)-716(N) eine beliebige Anzahl von Zentraleinheiten („CPU“) oder andere Prozessoren (die Beschleuniger, feldprogrammierbare Gate-Arrays („FPGA“), Grafikprozessoren usw.), Arbeitsspeichervorrichtungen (z. B.dynamischer Nur-LeseSpeicher), Datenspeichervorrichtungen (z. B.Festkörper- oder Festplattenlaufwerke), Netzwerk-Ein-/Ausgabevorrichtungen („NW E/A“), Netzwerk-Switches, virtuellen Maschinen („VM“), Leistungsmodule und Kühlmodule usw. beinhalten, ohne darauf beschränkt zu sein. In mindestens einer Ausführungsform können ein oder mehrere Knoten-C.R.s von den Knoten-C.R.s 716(1)-716(N) ein Server sein, der eine oder mehrere der vorstehend erwähnten Rechenressourcen aufweist.

[0041] In mindestens einer Ausführungsform können gruppierte Rechenressourcen 714 getrennte Gruppierungen von Knoten-C.R.s beinhalten, die in einem oder mehreren Racks (nicht dargestellt) untergebracht sind, oder vielen Racks, die in Rechenzentren an verschiedenen geografischen Standorten (ebenfalls nicht dargestellt) untergebracht sind. Getrennte Gruppierungen von Knoten-C.R.s 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-C.R.s, die CPUs oder Prozessoren beinhalten, innerhalb eines oder mehrerer Racks gruppiert sein, um Rechenressourcen zum Unterstützen einer oder mehrerer Arbeitslasten bereitzustellen. In mindestens einer Ausführungsform können ein oder mehrere Racks auch eine beliebige Anzahl von Leistungsmodulen, Kühlmodulen und Netz-Switches in beliebiger Kombination beinhalten.

[0042] In mindestens einer Ausführungsform kann der Ressourcen-Orchestrator 712 einen oder mehrere Knoten-C.R.s 716(1)-716(N) und/oder gruppierte Rechenressourcen 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 Ressourcen-Orchestrator 712 Hardware, Software oder eine Kombination davon beinhalten.

[0043] In mindestens einer Ausführungsform beinhaltet die Framework-Schicht 720, wie in **Fig. 7** dargestellt, ohne Einschränkung einen Aufgaben-Scheduler 732, einen Konfigurationsmanager 734, einen Ressourcenmanager 736 und ein verteiltes Dateisystem 738. In mindestens einer Ausführungsform kann die Framework-Schicht 720 ein Framework beinhalten, um Software 752 der Softwareschicht 730 und/oder eine oder mehrere Anwendung(en) 742 der Anwendungsschicht 740 zu unterstützen. In mindestens einer Ausführungsform kann/können die Software 752 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 SparkTM (im Folgenden „Spark“), welches das verteilte Dateisystem 738 für umfangreiche Datenverarbeitungen (z.B. „Big Data“) nutzen kann. In mindestens einer Ausführungsform kann der Aufgabenplaner 732 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 734 in der Lage sein, unterschiedliche Schichten zu konfigurieren, wie etwa die Softwareschicht 730 und die Frameworkschicht 720, was Spark und das verteilte Dateisystem 738 zum Unterstützen einer umfangreichen Datenverarbeitung beinhaltet. In mindestens einer Ausführungsform kann der Ressourcenverwalter 736 in der Lage sein, geclusterte oder gruppierte Rechenressourcen zu verwalten, die dem verteilten Dateisystem 738 und dem Aufgabenplaner 732 zur Unterstützung zugeordnet oder zugewiesen sind. In mindestens einer Ausführungsform können geclusterte oder gruppierte Rechenressourcen eine gruppierte Rechenressource 714 auf der Rechenzentrumsinfrastruktur-

schicht 710 beinhalten. In mindestens einer Ausführungsform kann sich der Ressourcenverwalter 736 mit dem Ressourcenorchestrierer 712 koordinieren, um diese zugeordneten oder zugewiesenen Rechenressourcen zu verwalten.

[0044] In mindestens einer Ausführungsform kann die in der Softwareschicht 730 enthaltene Software 752 Software beinhalten, die mindestens durch Abschnitte der Knoten-C.R.s 716(1)-716(N), der gruppierten Rechenressourcen 714 und/oder des verteilten Dateisystems 738 der Rahmenschicht 720 verwendet wird. Zu einem oder mehreren Typen von Software können Software zum Durchsuchen von Internet-Webseiten, Software zum Scannen von E-Mails auf Viren, Datenbank-Software und Software für Streaming-Videoinhalte gehören, ohne darauf beschränkt zu sein.

[0045] In mindestens einer Ausführungsform kann/können die Anwendung(en) 742, die in der Anwendungsschicht 740 beinhaltet ist/sind, eine oder mehrere Arten von Anwendungen beinhalten, die durch mindestens Teile der Knoten-C.R.s 716(1)-716(N), gruppierte Rechenressourcen 714 und/oder das verteilte Dateisystem 738 der Framework-Schicht 720 verwendet werden. In mindestens einem oder mehreren Anwendungstypen können ohne Einschränkung CUDA-Anwendungen, 5G-Netzwerkanwendungen, Anwendungen künstlicher Intelligenz, Rechenzentrumsanwendungen und/oder Variationen davon enthalten sein.

[0046] In mindestens einer Ausführungsform können beliebige von dem Konfigurationsmanager 734, Ressourcenmanager 736 und Ressourcen-Orchestrator 712 eine beliebige Anzahl und Art von selbstmodifizierenden Handlungen auf Grundlage einer beliebigen Menge und Art von Daten implementieren, die auf jede technisch machbare Weise erfasst werden. In mindestens einer Ausführungsform können selbstmodifizierende Handlungen einen Rechenzentrumsbetreiber des Rechenzentrums 700 davon entlasten, möglicherweise schlechte Konfigurationsentscheidungen zu treffen, und möglicherweise vermeiden, dass Teile eines Rechenzentrums nicht ausgelastet und/oder leistungsschwach sind.

[0047] Fig. 8 zeigt ein Client-Server-Netzwerk 804, das gemäß mindestens einer Ausführungsform durch eine Vielzahl von miteinander verbundenen Netzwerk-Server-Computern 802 gebildet wird. In mindestens einer Ausführungsform speichert jeder Netzwerk-Server-Computer 802 Daten, auf die andere Netzwerk-Server-Computer 802 und Client-Computer 806 und Netzwerke 808 zugreifen können, die mit einem Weitverkehrsnetzwerk 804 verbunden sind. In mindestens einer Ausführungsform kann sich die Konfiguration eines Client-Server-Netzwerks 804 im Laufe der Zeit ändern, wenn sich Client-Computer 806 und ein oder mehrere Netzwerke 808 mit einem Netzwerk 804 verbinden und von diesem trennen, und wenn ein oder mehrere Verbindungsleitungs-Server-Computer 802 einem Netzwerk 804 hinzugefügt oder daraus entfernt werden. In mindestens einer Ausführungsform, wenn ein Client-Computer 806 und ein Netzwerk 808 mit Netzwerk-Server-Computern 802 verbunden sind, beinhaltet das Client-Server-Netzwerk einen solchen Client-Computer 806 und ein Netzwerk 808. In mindestens einer Ausführungsform beinhaltet der Begriff Computer jede Vorrichtung oder jede Maschine, die in der Lage ist, Daten zu akzeptieren, vorgeschriebene Prozesse auf Daten anzuwenden und Ergebnisse von Prozessen bereitzustellen.

[0048] In mindestens einer Ausführungsform speichert das Client-Server-Netzwerk 804 Informationen, auf die Netzwerk-Server-Computer 802, entfernte Netzwerke 808 und Client-Computer 806 zugreifen können. In mindestens einer Ausführungsform werden die Netzwerk-Server-Computer 802 durch Mainframe-Computer, Minicomputer und/oder Mikrocomputer mit jeweils einem oder mehreren Prozessoren gebildet. In mindestens einer Ausführungsform sind die Server-Computer 802 durch drahtgebundene und/oder drahtlose Übertragungsmedien miteinander verbunden, wie etwa leitfähige Drähte, Glasfaserkabel und/oder Mikrowellenübertragungsmedien, Satellitenübertragungsmedien oder andere leitfähige, optische oder elektromagnetische Wellenübertragungsmedien. In mindestens einer Ausführungsform greifen Client-Computer 806 auf einen Netzwerk-Server-Computer 802 durch ein ähnliches drahtgebundenes oder ein drahtloses Übertragungsmedium zu. In mindestens einer Ausführungsform kann sich ein Client-Computer 806 mit einem Client-Server-Netzwerk 804 unter Verwendung eines Modems und eines standardmäßigen Telefonkommunikationsnetzwerks verbinden. In mindestens einer Ausführungsform können auch alternative Trägersysteme wie Kabel- und Satellitenkommunikationssysteme verwendet werden, um sich mit dem Client-Server-Netzwerk 804 zu verbinden. In mindestens einer Ausführungsform können andere private oder zeitverschachtelte Trägersysteme verwendet werden. In mindestens einer Ausführungsform ist das Netzwerk 804 ein globales Informationsnetzwerk, wie etwa das Internet. In mindestens einer Ausführungsform ist das Netzwerk ein privates Intranet, das ähnliche Protokolle wie das Internet verwendet, jedoch mit zusätzlichen Sicherheitsmaßnahmen und eingeschränkten Zugriffskontrollen. In mindestens einer Ausführungsform ist das Netzwerk 804 ein privates oder halbprivates Netzwerk, das proprietäre Kommunikationsprotokolle verwendet.

[0049] In mindestens einer Ausführungsform ist der Client-Computer 806 ein beliebiger Endbenutzer-Computer und kann auch ein Mainframe-Computer, ein Minicomputer oder ein Mikrocomputer mit einem oder mehreren Mikroprozessoren sein. In mindestens einer Ausführungsform kann der Server-Computer 802 manchmal als ein Client-Computer fungieren, der auf einen anderen Server-Computer 802 zugreift. In mindestens einer Ausführungsform kann das entfernte Netzwerk 808 ein lokales Netzwerk sein, ein Netzwerk, das durch einen unabhängigen Dienstanbieter (independent service provider - ISP) für das Internet zu einem Weitverkehrsnetzwerk hinzugefügt wird, oder eine andere Gruppe von Computern, die durch drahtgebundene oder drahtlose Übertragungsmedien mit einer Konfiguration, die entweder feststehend ist oder sich im Laufe der Zeit ändert, sein. In mindestens einer Ausführungsform können sich Client-Computer 806 unabhängig oder über ein entferntes Netzwerk 808 mit einem Netzwerk 804 verbinden und darauf zugreifen.

[0050] Fig. 9 zeigt ein Computernetzwerk 908, das gemäß mindestens einer Ausführungsform eine oder mehrere Rechenmaschinen verbindet. In mindestens einer Ausführungsform kann das Netzwerk 908 eine beliebige Art einer elektronisch verbundenen Gruppe von Computern sein, einschließlich beispielsweise der folgenden Netzwerke: Internet, Intranet, lokale Netzwerke (LAN), Weitverkehrsnetzwerke (WAN) oder eine miteinander verbundene Kombination dieser Netzwerkarten. In mindestens einer Ausführungsform kann die Konnektivität innerhalb eines Netzwerks 908 ein entferntes Modem, Ethernet (IEEE 802.3), Token Ring (IEEE 802.5), Fiber Distributed Datalink Interface (FDDI), Asynchronous Transfer Mode (ATM) oder ein beliebiges anderes Kommunikationsprotokoll sein. In mindestens einer Ausführungsform können mit einem Netzwerk verbundene Rechenvorrichtungen ein Desktop, ein Server, ein tragbares Gerät, ein Handgerät, eine Set-Top-Box, ein persönlicher digitaler Assistent (PDA), ein Terminal oder jede andere gewünschte Art oder Konfiguration sein. In mindestens einer Ausführungsform können netzwerkverbundene Vorrichtungen abhängig von ihrer Funktionalität in der Verarbeitungsleistung, dem internen Speicher und anderen Leistungsaspekten stark variieren.

[0051] In mindestens einer Ausführungsform kann die Kommunikation innerhalb eines Netzwerks und zu oder von Rechenvorrichtungen, die mit einem Netzwerk verbunden sind, entweder drahtgebunden oder drahtlos sein. In mindestens einer Ausführungsform kann das Netzwerk 908 zumindest teilweise das weltweite öffentliche Internet beinhalten, das im Allgemeinen eine Vielzahl von Benutzern gemäß einem Client-Server-Modell gemäß einer Übertragungssteuerungsprotokoll/Internetprotokoll (TCP/IP)-Spezifikation verbindet. In mindestens einer Ausführungsform ist das Client-Server-Netzwerk ein vorherrschendes Modell für die Kommunikation zwischen zwei Computern. In mindestens einer Ausführungsform gibt ein Client-Computer („Client“) einen oder mehrere Befehle an einen Server-Computer („Server“) aus. In mindestens einer Ausführungsform führt der Server Client-Befehle aus, indem er auf verfügbare Netzwerkressourcen zugreift und Informationen gemäß den Client-Befehlen an einen Client zurückgibt. In mindestens einer Ausführungsform wird Client-Computersystemen und Netzwerkressourcen, die auf Netzwerkservern resident sind, eine Netzwerkadresse zur Identifizierung während der Kommunikation zwischen Elementen eines Netzwerks zugewiesen. In mindestens einer Ausführungsform beinhalten Kommunikationen von anderen netzwerkverbundenen Systemen zu Servern eine Netzwerkadresse eines relevanten Servers/einer relevanten Netzwerkressource als Teil der Kommunikation, sodass ein geeignetes Ziel von Daten bzw. einer Anfrage als ein Empfänger identifiziert wird. In mindestens einer Ausführungsform, wenn ein Netzwerk 908 das globale Internet umfasst, ist eine Netzwerkadresse eine IP-Adresse in einem TCP/IP-Format, die Daten mindestens teilweise an ein E-Mail-Konto, eine Website oder ein anderes auf einem Server residentes Internet-Tool weiterleiten kann. In mindestens einer Ausführungsform können Informationen und Dienste, die auf Netzwerkservern resident sind, für einen Webbrowser eines Client-Computers über einen Domänennamen (z. B. www.site.com) verfügbar sein, der einer IP-Adresse eines Netzwerk-Servers zugeordnet ist.

[0052] In mindestens einer Ausführungsform ist eine Vielzahl von Clients 902, 904 und 906 über jeweilige Kommunikationsverbindungen mit einem Netzwerk 908 verbunden. In mindestens einer Ausführungsform kann jeder dieser Clients über jede gewünschte Kommunikationsform auf ein Netzwerk 908 zugreifen, wie beispielsweise über eine Einwahlmodemverbindung, eine Kabelverbindung, eine digitale Teilnehmerleitung (DSL), eine drahtlose oder Satellitenverbindung oder jede andere Form der Kommunikation. In mindestens einer Ausführungsform kann jeder Client unter Verwendung einer beliebigen Maschine kommunizieren, die mit einem Netzwerk 908 kompatibel ist, wie beispielsweise ein Personal Computer (PC), eine Arbeitsstation, ein dediziertes Terminal, ein persönlicher Datenassistent (PDA) oder eine andere ähnliche Ausrüstung. In mindestens einer Ausführungsform können sich die Clients 902, 904 und 906 in einem gleichen geografischen Gebiet befinden oder nicht.

[0053] In mindestens einer Ausführungsform ist eine Vielzahl von Servern 910, 912 und 914 mit einem Netzwerk 918 verbunden, um Clients zu bedienen, die mit einem Netzwerk 918 kommunizieren. In mindestens

einer Ausführungsform ist jeder Server typischerweise ein leistungsstarker Computer oder eine leistungsstarke Vorrichtung, die Netzwerkressourcen verwaltet und auf Client-Befehle reagiert. In mindestens einer Ausführungsform umfassen Server computerlesbare Datenspeichermedien wie Festplattenlaufwerke und RAM-Speicher, die Programmanweisungen und Daten speichern. In mindestens einer Ausführungsform führen die Server 910, 912, 914 Anwendungsprogramme aus, die auf Client-Befehle reagieren. In mindestens einer Ausführungsform kann der Server 910 eine Webserver-Anwendung zum Reagieren auf Client-Anfragen nach HTML-Seiten ausführen und kann auch eine Mail-Server-Anwendung zum Empfangen und Weiterleiten von elektronischer Post ausführen. In mindestens einer Ausführungsform können auch andere Anwendungsprogramme, wie etwa ein FTP-Server oder ein Medienserver zum Streamen von Audio-/Videodaten an Clients, auf einem Server 910 ausgeführt werden. In mindestens einer Ausführungsform können unterschiedliche Server dazu bestimmt sein, unterschiedliche Tasks auszuführen. In mindestens einer Ausführungsform kann der Server 910 ein dedizierter Webserver sein, der Ressourcen in Bezug auf Websites für verschiedene Benutzer verwaltet, während ein Server 912 dazu bestimmt sein kann, eine Verwaltung von elektronischer Post (E-Mail) bereitzustellen. In mindestens einer Ausführungsform können andere Server für Medien (Audio, Video usw.), ein Dateiübertragungsprotokoll (file transfer protocol - FTP) oder eine Kombination von beliebigen zwei oder mehr Diensten, die typischerweise verfügbar sind oder über ein Netzwerk bereitgestellt werden, bestimmt sein. In mindestens einer Ausführungsform kann sich jeder Server an einem Standort befinden, der mit dem anderen Server identisch ist oder sich davon unterscheidet. In mindestens einer Ausführungsform kann es mehrere Server geben, die gespiegelte Tasks für Benutzer ausführen, wodurch Datenstaus verringert werden oder Datenverkehr, der zu und von einem einzelnen Server geleitet wird, minimiert wird. In mindestens einer Ausführungsform stehen die Server 910, 912, 914 unter der Kontrolle eines Webhosting-Anbieters in einem Unternehmen der Pflege und Bereitstellung von Inhalten Dritter über ein Netzwerk 918.

[0054] In mindestens einer Ausführungsform liefern Webhosting-Anbieter Dienste an zwei unterschiedliche Arten von Clients. In mindestens einer Ausführungsform fordert eine Art, die als Browser bezeichnet werden kann, Inhalt von den Servern 910, 912, 914 an, wie etwa Webseiten, E-Mail-Nachrichten, Videoclips usw. In mindestens einer Ausführungsform beauftragt eine zweite Art, die als Benutzer bezeichnet werden kann, einen Webhosting-Anbieter eine Netzwerkressource, wie etwa eine Website, zu pflegen und für Browser verfügbar zu machen. In mindestens einer Ausführungsform schließen Benutzer einen Vertrag mit einem Webhosting-Anbieter ab, um Speicherplatz, Prozessorkapazität und Kommunikationsbandbreite für ihre gewünschte Netzwerkressource gemäß einer Menge von Serverressourcen verfügbar zu machen, die ein Benutzer nutzen möchte.

[0055] Damit ein Webhosting-Anbieter Dienste für diese beiden Clients bereitstellen kann, müssen in mindestens einer Ausführungsform Anwendungsprogramme, die durch Server gehostete Netzwerkressourcen verwalten, richtig konfiguriert sein. In mindestens einer Ausführungsform umfasst der Programmkonfigurationsprozess das Definieren eines Satzes von Parametern, die mindestens teilweise die Reaktion eines Anwendungsprogramms auf Browseranforderungen steuern und die auch mindestens teilweise Serverressourcen definieren, die einem bestimmten Benutzer zur Verfügung stehen.

[0056] In einer Ausführungsform steht ein Intranet-Server 916 mit einem Netzwerk 908 über eine Kommunikationsverbindung in Kommunikation. In mindestens einer Ausführungsform steht der Intranet-Server 916 mit einem Servermanager 918 in Kommunikation. In mindestens einer Ausführungsform umfasst der Servermanager 918 eine Datenbank mit Konfigurationsparametern eines Anwendungsprogramms, die in den Servern 910, 912, 914 genutzt werden. In mindestens einer Ausführungsform modifizieren Benutzer eine Datenbank 920 über ein Intranet 916 und interagiert ein Servermanager 918 mit Servern 910, 912, 914, um Anwendungsprogrammparameter so zu modifizieren, dass sie mit einem Inhalt einer Datenbank übereinstimmen. In mindestens einer Ausführungsform meldet sich ein Benutzer bei einem Intranet-Server 916 an, indem er über den Computer 902 eine Verbindung zu einem Intranet 916 herstellt und Authentifizierungsinformationen wie beispielsweise einen Benutzernamen und ein Passwort eingibt.

[0057] In mindestens einer Ausführungsform authentifiziert ein Intranet-Server 916 einen Benutzer, wenn sich ein Benutzer für einen neuen Dienst anmelden oder einen bestehenden Dienst modifizieren möchte, und stellt einem Benutzer eine interaktive Bildschirmanzeige/ein Bedienfeld bereit, die bzw. das einem Benutzer den Zugriff auf Konfigurationsparameter für ein konkretes Anwendungsprogramm erlaubt. In mindestens einer Ausführungsform wird einem Benutzer eine Reihe von modifizierbaren Textfeldern präsentiert, die Aspekte einer Konfiguration einer Benutzerwebsite oder einer anderen Netzwerkressource beschreiben. Wenn ein Benutzer in mindestens einer Ausführungsform den auf einem Server für seine Website reservierten Speicherplatz vergrößern möchte, wird einem Benutzer ein Feld bereitgestellt, in dem ein Benutzer einen

gewünschten Speicherplatz festlegt. In mindestens einer Ausführungsform aktualisiert ein Intranet-Server 916 als Reaktion auf den Empfang dieser Informationen eine Datenbank 920. In mindestens einer Ausführungsform leitet der Servermanager 918 diese Informationen an einen geeigneten Server weiter, und ein neuer Parameter wird während des Anwendungsprogrammbetriebs verwendet. In mindestens einer Ausführungsform ist ein Intranet-Server 916 konfiguriert, um Benutzern Zugriff auf Konfigurationsparameter von gehosteten Netzwerkressourcen (z. B. Webseiten, E-Mail, FTP-Sites, Mediensites usw.) bereitzustellen, für die ein Benutzer einen Vertrag mit einem Webhosting-Dienstanbieter abgeschlossen hat.

[0058] Fig. 10A zeigt ein vernetztes Computersystem 1000A gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform umfasst das vernetzte Computersystem 1000A eine Vielzahl von Knoten oder Personalcomputern („PCs“) 1002, 1018, 1020. In mindestens einer Ausführungsform umfasst der Personalcomputer oder Knoten 1002 einen Prozessor 1014, einen Speicher 1016, eine Videokamera 1004, ein Mikrofon 1006, eine Maus 1008, Lautsprecher 1010 und einen Monitor 1012. In mindestens einer Ausführungsform können die PCs 1002, 1018, 1020 beispielsweise jeweils einen oder mehrere Desktop-Server eines internen Netzwerks innerhalb einer gegebenen Firma ausführen oder können Server eines allgemeinen Netzwerks sein, das nicht auf eine spezifische Umgebung beschränkt ist. In mindestens einer Ausführungsform gibt es einen Server pro PC-Knoten eines Netzwerks, sodass jeder PC-Knoten eines Netzwerks einen konkreten Netzwerkserver mit einer konkreten Netzwerk-URL-Adresse darstellt. In mindestens einer Ausführungsform verwendet jeder Server standardmäßig eine Standardwebseite für den Benutzer dieses Servers, die selbst eingebettete URL enthalten kann, die auf weitere Unterseiten dieses Benutzers auf diesem Server oder auf andere Server oder Seiten auf anderen Servern in einem Netzwerk zeigen.

[0059] In mindestens einer Ausführungsform sind die Knoten 1002, 1018, 1020 und andere Knoten eines Netzwerks über ein Medium 1022 miteinander verbunden. In mindestens einer Ausführungsform kann das Medium 1022 ein Kommunikationskanal sein, wie etwa ein Integrated Services Digital Network („ISDN“). In mindestens einer Ausführungsform können verschiedene Knoten eines vernetzten Computersystems durch eine Vielzahl von Kommunikationsmedien verbunden sein, darunter lokale Netze („LAN“), einfache alte Telefonleitungen („POTS“), die manchmal als öffentlich vermittelte Telefonnetze („PSTN“) bezeichnet werden und/oder Variationen davon. In mindestens einer Ausführungsform können verschiedene Knoten eines Netzwerks auch Computersystembenutzer darstellen, die über ein Netzwerk wie das Internet miteinander verbunden sind. In mindestens einer Ausführungsform hat jeder Server in einem Netzwerk (der von einem bestimmten Knoten eines Netzwerks in einer gegebenen Instanz ausgeht) eine eindeutige Adresse oder Kennung innerhalb eines Netzwerks, die in Form eines URL spezifiziert werden kann.

[0060] In mindestens einer Ausführungsform kann somit eine Vielzahl von Mehrpunkt-Konferenzeinheiten (multi-point conferencing units - „MCU“) verwendet werden, um Daten zu und von verschiedenen Knoten oder „Endpunkten“ eines Konferenzsystems zu übertragen. In mindestens einer Ausführungsform können Knoten und/oder MCU zusätzlich zu verschiedenen anderen Kommunikationsmedien wie über das Internet verbundene Knoten über eine ISDN-Verbindung oder über ein lokales Netzwerk („LAN“) miteinander verbunden sein. In mindestens einer Ausführungsform können Knoten eines Konferenzsystems im Allgemeinen direkt mit einem Kommunikationsmedium wie beispielsweise einem LAN oder über eine MCU verbunden sein, und kann ein Konferenzsystem andere Knoten oder Elemente wie beispielsweise Router, Server usw. oder Variationen davon umfassen.

[0061] In mindestens einer Ausführungsform ist der Prozessor 1014 ein programmierbarer Allzweckprozessor. In mindestens einer Ausführungsform können Prozessoren von Knoten des vernetzten Computersystems 1000A auch Spezial-Videoprozessoren sein. In mindestens einer Ausführungsform können verschiedene Peripheriegeräte und Komponenten eines Knotens wie die des Knotens 1002 von denen anderer Knoten abweichen. In mindestens einer Ausführungsform können der Knoten 1018 und der Knoten 1020 genauso wie oder anders als der Knoten 1002 konfiguriert sein. In mindestens einer Ausführungsform kann ein Knoten zusätzlich zu PC-Systemen auf jedem geeigneten Computersystem implementiert sein.

[0062] Fig. 10B zeigt ein vernetztes Computersystem 1000B gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform veranschaulicht das System 1000B ein Netzwerk wie ein LAN 1024, das verwendet werden kann, um eine Vielzahl von Knoten zu verbinden, die miteinander kommunizieren kann. In mindestens einer Ausführungsform ist eine Vielzahl von Knoten an das LAN 1024 angeschlossen, wie etwa PC-Knoten 1026, 1028, 1030. In mindestens einer Ausführungsform kann ein Knoten auch über einen Netzwerkserver oder andere Mittel mit dem LAN verbunden sein. In mindestens einer Ausführungsform umfasst das System 1000B andere Arten von Knoten oder Elementen, wobei mindestens eine Ausführungsform Router, Server und Knoten beinhaltet.

[0063] **Fig. 10C** zeigt ein vernetztes Computersystem 1000C gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform veranschaulicht das System 1000C ein WWW-System, das Kommunikationen über ein Backbone-Kommunikationsnetzwerk wie etwa das Internet 1032 aufweist, das verwendet werden kann, um eine Vielzahl von Knoten eines Netzes miteinander zu verbinden. In mindestens einer Ausführungsform ist das WWW ein Satz von Protokollen, der auf dem Internet arbeitet und ermöglicht, dass ein grafisches Schnittstellensystem darauf arbeitet, um über das Internet auf Informationen zuzugreifen. In mindestens einer Ausführungsform ist eine Vielzahl von Knoten an das Internet 1032 angeschlossen, wie etwa PC-Knoten 1040, 1042, 1044. In mindestens einer Ausführungsform bildet ein Knoten über einen WWW-HTTP-Server, wie etwa die Server 1034, 1036, eine Schnittstelle mit anderen Knoten des WWW. In mindestens einer Ausführungsform kann der PC 1044 ein PC sein, der einen Knoten des Netzwerks 1032 bildet und selbst seinen Server 1036 betreibt, obwohl der PC 1044 und der Server 1036 in **Fig. 10C** zur Veranschaulichung getrennt veranschaulicht sind.

[0064] In mindestens einer Ausführungsform ist WWW ein verteilter Anwendungstyp, gekennzeichnet durch WWW HTTP, das WWW-Protokoll, das auf dem Übertragungssteuerungsprotokoll/Internetprotokoll („TCP/IP“) des Internets läuft. In mindestens einer Ausführungsform kann das WWW somit durch einen Satz von Protokollen (d. h. HTTP) gekennzeichnet sein, die im Internet als sein „Backbone“ laufen.

[0065] In mindestens einer Ausführungsform ist ein Webbrowser eine auf einem Knoten eines Netzwerks laufende Anwendung, die in WWW-kompatiblen Netzwerksystemen Benutzern eines bestimmten Servers oder Knotens ermöglicht, solche Informationen anzuzeigen und somit einem Benutzer ermöglicht, grafische und textbasierte Dateien, die über in Dokumente oder Dateien eingebettete Hypertext-Links, die von Servern in einem Netzwerk verfügbar sind, die HTTP verstehen, miteinander verknüpft sind. In mindestens einer Ausführungsform kann ein abgerufenes Dokument verschiedene darin eingebettete Hypertext-Links und eine lokale Kopie einer Seite aufweisen, die lokal für einen abrufenden Benutzer erstellt wird, wenn eine gegebene Webseite eines ersten Servers, der einem ersten Knoten zugeordnet ist, durch einen Benutzer unter Verwendung eines anderen Servers in einem Netzwerk wie dem Internet abgerufen wird. Wenn ein Benutzer auf einen Hypertext-Link klickt, reichen in mindestens einer Ausführungsform typischerweise lokal gespeicherte Informationen in Bezug auf einen ausgewählten Hypertext-Link aus, um es der Maschine eines Benutzers zu ermöglichen, eine Verbindung über das Internet zu einem Server zu öffnen, der durch einen Hypertext-Link angezeigt wird.

[0066] In mindestens einer Ausführungsform kann mehr als ein Benutzer mit jedem HTTP-Server über ein LAN, wie etwa das LAN 1038, gekoppelt sein, wie in Bezug auf den WWW-HTTP-Server 1034 dargestellt. In mindestens einer Ausführungsform kann das System 1000C auch andere Arten von Knoten oder Elementen umfassen. In mindestens einer Ausführungsform ist ein WWW-HTTP-Server eine Anwendung, die auf einer Maschine, wie etwa einem PC, läuft. In mindestens einer Ausführungsform kann davon ausgegangen werden, dass jeder Benutzer einen eindeutigen „Server“ aufweist, wie in Bezug auf den PC 1044 zeigt. In mindestens einer Ausführungsform kann ein Server als ein Server betrachtet werden, wie etwa der WWW-HTTP-Server 1034, der Zugriff auf ein Netzwerk für ein LAN oder eine Vielzahl von Knoten oder eine Vielzahl von LAN bereitstellt. In mindestens einer Ausführungsform gibt es eine Vielzahl von Benutzern, von der jeder einen Desktop-PC oder Knoten eines Netzwerks aufweist, wobei jeder Desktop-PC potentiell einen Server für einen Benutzer davon herstellt. In mindestens einer Ausführungsform ist jeder Server einer konkreten Netzwerkadresse oder URL zugeordnet, die, wenn darauf zugegriffen wird, eine Standardwebseite für diesen Benutzer bereitstellt. In mindestens einer Ausführungsform kann eine Webseite weitere Links (eingebettete URL) enthalten, die auf weitere Unterseiten dieses Benutzers auf diesem Server oder auf andere Server in einem Netzwerk oder Seiten auf anderen Servern in einem Netzwerk zeigen.

CLOUD-COMPUTING UND -DIENSTE

[0067] Die folgenden Figuren legen ohne Einschränkung beispielhafte Cloudbasierte Systeme dar, die zum Implementieren mindestens einer Ausführungsform verwendet werden können.

[0068] In mindestens einer Ausführungsform ist Cloud-Computing eine Art der Datenverarbeitung, bei der dynamisch skalierbare und oft virtualisierte Ressourcen als Dienst über das Internet bereitgestellt werden. In mindestens einer Ausführungsform brauchen die Benutzer keine Kenntnisse, kein Fachwissen oder keine Kontrolle über die sie unterstützende Technologieinfrastruktur zu haben, die als „in der Cloud“ bezeichnet werden kann. In mindestens einer Ausführungsform umfasst Cloud-Computing Infrastruktur-als-Dienst, Plattform-als-Dienst, Software-als-Dienst und andere Variationen, die ein gemeinsames Leitmotiv der Abhängigkeit vom Internet zum Erfüllen von Rechenbedürfnissen von Benutzern aufweisen. In mindestens einer Aus-

führungsform kann ein typischer Cloud-Einsatz, wie etwa in einer privaten Cloud (z. B. Unternehmensnetzwerk) oder einem Rechenzentrum (data center - DC) in einer öffentlichen Cloud (z. B. Internet) aus Tausenden von Servern (oder alternativ VM), Hunderten von Ethernet-, Fibre Channel- oder Fibre Channel over Ethernet (FCoE)-Ports, Switching- und Speicherinfrastruktur usw. bestehen. In mindestens einer Ausführungsform kann die Cloud auch aus einer Netzwerkdienstinfrastruktur wie IPsec-VPN-Hubs, Firewalls, Lastausgleichern, Weitverkehrsnetz (WAN)-Optimierern usw. bestehen. In mindestens einer Ausführungsform können entfernte Teilnehmer sicher auf Cloud-Anwendungen und - Dienste zugreifen, indem sie sich über einen VPN-Tunnel, wie etwa einen IPsec-VPN-Tunnel, verbinden.

[0069] In mindestens einer Ausführungsform ist Cloud-Computing ein Modell zum Ermöglichen eines bequemen On-Demand-Netzwerkzugriffs auf einen gemeinsam genutzten Pool konfigurierbarer Rechenressourcen (z. B. Netzwerke, Server, Speicher, Anwendungen und Dienste), die mit minimalem Verwaltungsaufwand oder Dienstanbieterinteraktion schnell bereitgestellt und freigegeben werden können.

[0070] In mindestens einer Ausführungsform ist Cloud-Computing durch On-Demand-Selbstbedienung gekennzeichnet, bei der ein Verbraucher einseitig Rechenfähigkeiten, wie etwa Serverzeit und Netzwerkspeicher, nach Bedarf automatisch bereitstellen kann, ohne dass eine menschliche Interaktion mit dem jeweiligen Dienstanbieter erforderlich ist. In mindestens einer Ausführungsform ist Cloud-Computing durch breiten Netzzugang gekennzeichnet, bei dem Fähigkeiten über ein Netzwerk verfügbar sind und durch Standardmechanismen darauf zugegriffen wird, die die Nutzung durch heterogene Thin- oder Thick-Client-Plattformen (z. B. Mobiltelefone, Laptops und PDA) fördern. In mindestens einer Ausführungsform ist Cloud-Computing gekennzeichnet durch Ressourcen-Pooling, bei dem die Rechenressourcen eines Anbieters gepoolt werden, um mehrere Verbraucher unter Verwendung eines mehrmandantenfähigen Modells zu bedienen, wobei verschiedene physische und virtuelle Ressourcen gemäß Verbrauchernachfrage dynamisch zugewiesen und neu zugewiesen werden. In mindestens einer Ausführungsform besteht ein Gefühl der Standortunabhängigkeit darin, dass ein Kunde im Allgemeinen keine Kontrolle oder Kenntnis über einen genauen Standort der bereitgestellten Ressourcen hat, aber in der Lage sein kann, den Standort auf einer höheren Abstraktionsebene (z. B. Land, Staat oder Rechenzentrum) festzulegen.

[0071] In mindestens einer Ausführungsform beinhalten Ressourcen Speicherplatz, Verarbeitung, Speicher, Netzwerkbandbreite und virtuelle Maschinen. In mindestens einer Ausführungsform ist Cloud-Computing durch eine schnelle Elastizität gekennzeichnet, bei der Fähigkeiten schnell und elastisch, in einigen Fällen automatisch, für ein schnelles Scale-Out bereitgestellt werden können und für ein schnelles Scale-In schnell freigegeben werden. In mindestens einer Ausführungsform erscheinen einem Verbraucher die zur Bereitstellung verfügbaren Fähigkeiten oft als unbegrenzt und können jederzeit in beliebiger Menge erworben werden. In mindestens einer Ausführungsform ist Cloud-Computing durch einen gemessenen Dienst gekennzeichnet, bei dem Cloud-Systeme die Ressourcennutzung automatisch steuern und optimieren, indem sie eine Messfähigkeit auf einer bestimmten Abstraktionsebene nutzen, die für eine Art von Dienst (z. B. Speicherung, Verarbeitung, Bandbreite und aktive Benutzerkonten) geeignet ist. In mindestens einer Ausführungsform kann die Ressourcennutzung überwacht, gesteuert und gemeldet werden, um Transparenz sowohl für einen Anbieter als auch für einen Verbraucher eines genutzten Dienstes bereitzustellen.

[0072] In mindestens einer Ausführungsform kann Cloud-Computing mit verschiedenen Diensten verbunden sein. In mindestens einer Ausführungsform kann sich Cloud Software as a Service (SaaS) als Dienst bezeichnen, bei dem eine einem Verbraucher bereitgestellte Fähigkeit darin besteht, die Anwendungen eines Anbieters zu verwenden, die auf einer Cloud-Infrastruktur laufen. In mindestens einer Ausführungsform sind Anwendungen von verschiedenen Client-Vorrichtungen über eine Thin-Client-Schnittstelle wie etwa einen Webbrowser (z. B. webbasierte E-Mail) zugänglich. In mindestens einer Ausführungsform verwaltet oder kontrolliert der Verbraucher die zugrunde liegende Cloud-Infrastruktur nicht, darunter Netzwerk, Server, Betriebssysteme, Speicher oder sogar individuelle Anwendungsfähigkeiten, mit einer möglichen Ausnahme von eingeschränkten benutzerspezifischen Anwendungskonfigurationseinstellungen.

[0073] In mindestens einer Ausführungsform kann sich Cloud Platform as a Service (PaaS) auf einen Dienst beziehen, bei dem eine einem Verbraucher bereitgestellte Fähigkeit darin besteht, vom Verbraucher erstellte oder erworbene Anwendungen, die unter Verwendung von Programmiersprachen und Tools erstellt wurden, die von einem Anbieter unterstützt werden, auf einer Cloud-Infrastruktur bereitzustellen. In mindestens einer Ausführungsform verwaltet oder steuert der Verbraucher die zugrunde liegende Cloud-Infrastruktur, darunter Netzwerke, Server, Betriebssysteme oder Speicherplatz nicht, hat aber die Kontrolle über installierte Anwendungen und möglicherweise über ausgewählte Netzwerkkomponenten zum Hosten von Anwendungen.

[0074] In mindestens einer Ausführungsform kann sich Cloud Infrastructure as a Service (IaaS) auf einen Dienst beziehen, bei dem eine einem Verbraucher bereitgestellte Fähigkeit darin besteht, Verarbeitungs-, Speicher-, Netzwerk- und andere grundlegende Computerressourcen bereitzustellen, wobei ein Verbraucher beliebige Software, die Betriebssysteme und Anwendungen beinhalten kann, installieren und ausführen kann. In mindestens einer Ausführungsform verwaltet oder kontrolliert der Verbraucher die zugrunde liegende Cloud-Infrastruktur nicht, hat aber die Kontrolle über Betriebssysteme, Speicher, eingesetzte Anwendungen und möglicherweise begrenzte Kontrolle über ausgewählte Netzwerkkomponenten (z. B. Host-Firewalls).

[0075] In mindestens einer Ausführungsform kann Cloud-Computing auf verschiedene Art und Weise eingesetzt werden. In mindestens einer Ausführungsform kann sich eine private Cloud auf eine Cloud-Infrastruktur beziehen, die ausschließlich für eine Organisation betrieben wird. In mindestens einer Ausführungsform kann eine private Cloud von einer Organisation oder einem Dritten verwaltet werden und kann vor Ort oder außerhalb existieren. In mindestens einer Ausführungsform kann sich eine gemeinschaftliche Cloud auf eine Cloud-Infrastruktur beziehen, die von mehreren Organisationen gemeinsam genutzt wird und eine bestimmte Gemeinschaft unterstützt, die gemeinsame Anliegen hat (z. B. Zielsetzung, Sicherheitsanforderungen, Richtlinien und Compliance-Überlegungen). In mindestens einer Ausführungsform kann eine Community Cloud von einer Organisation oder einem Dritten verwaltet werden und kann vor Ort oder außerhalb existieren. In mindestens einer Ausführungsform kann sich eine öffentliche Cloud auf eine Cloud-Infrastruktur beziehen, die einer breiten Öffentlichkeit oder einer großen Industriegruppe zur Verfügung gestellt wird und sich im Besitz einer Organisation befindet, die Cloud-Dienste bereitstellt. In mindestens einer Ausführungsform kann sich eine Hybrid-Cloud auf eine Cloud-Infrastruktur beziehen, die eine Zusammensetzung aus zwei oder mehr Clouds (privat, gemeinschaftlich oder öffentlich) ist, die einzigartige Einheiten bleiben, aber durch standardisierte oder proprietäre Technologie miteinander verbunden sind, die Daten- und Anwendungsportabilität (z. B. Cloud Bursting für den Lastenausgleich zwischen Clouds) ermöglicht. In mindestens einer Ausführungsform ist eine Cloud-Computing-Umgebung dienstorientiert mit einem Fokus auf Staatenlosigkeit, geringe Kopplung, Modularität und semantische Interoperabilität.

[0076] **Fig. 11** zeigt eine oder mehrere Komponenten einer Systemumgebung gemäß mindestens einer Ausführungsform, 1100 in der Dienste als Drittanbieter-Netzwerkdienste angeboten werden. In mindestens einer Ausführungsform kann ein Drittanbieter-Netzwerk als Cloud, Cloud-Netzwerk, Cloud-Computing-Netzwerk und/oder Variationen davon bezeichnet werden. In mindestens einer Ausführungsform beinhaltet die Systemumgebung 1100 eine oder mehrere Client-Rechenvorrichtungen 1104, 1106 und 1108, die durch Benutzer verwendet werden können, um mit einem Drittanbieter-Netzwerkinfrastruktursystem 1102 zu interagieren, das Drittanbieter-Netzwerkdienste bereitstellt, die als Cloud-Computing-Dienste bezeichnet werden können. In mindestens einer Ausführungsform kann das Drittanbieter-Netzwerkinfrastruktursystem 1102 einen oder mehrere Computer und/oder Server umfassen.

[0077] Es versteht sich, dass das in **Fig. 11** dargestellte Drittanbieter-Netzwerkinfrastruktursystem 1102 andere als die dargestellten Komponenten aufweisen kann. Ferner zeigt **Fig. 11** eine Ausführungsform eines Drittanbieter-Netzwerkinfrastruktursystems. In mindestens einer Ausführungsform kann das Drittanbieter-Netzwerkinfrastruktursystem 1102 mehr oder weniger Komponenten aufweisen als in **Fig. 11** dargestellt, kann zwei oder mehr Komponenten kombinieren oder kann eine andere Konfiguration oder Anordnung von Komponenten aufweisen.

[0078] In mindestens einer Ausführungsform können die Client-Rechenvorrichtungen 1104, 1106 und 1108 zum Betreiben einer Client-Anwendung, wie etwa einen Webbrowser, eine proprietäre Client-Anwendung oder eine andere Anwendung konfiguriert sein, die durch einen Benutzer einer Client-Rechenvorrichtung verwendet werden kann, um mit dem Drittanbieter-Netzwerkinfrastruktursystem 1102 zu zusammenzuwirken, um Dienste zu nutzen, die durch das Drittanbieter-Netzwerkinfrastruktursystem 1102 bereitgestellt werden. Auch wenn die beispielhafte Systemumgebung 1100 mit drei Client-Rechenvorrichtungen gezeigt ist, kann eine beliebige Anzahl von Client-Rechenvorrichtungen unterstützt werden. In mindestens einer Ausführungsform können andere Vorrichtungen, wie etwa Vorrichtungen mit Sensoren usw., mit dem Drittanbieter-Netzwerkinfrastruktursystem 1102 interagieren. In mindestens einer Ausführungsform können das/die Netzwerk(e) 1110 die Kommunikation und den Austausch von Daten zwischen den Client-Rechenvorrichtungen 1104, 1106 und 1108 und dem Drittanbieter-Netzwerkinfrastruktursystem 1102 ermöglichen.

[0079] In mindestens einer Ausführungsform können Dienste, die durch das Netzinfrastruktursystem 1102 eines Dritten bereitgestellt werden, eine Reihe von Diensten umfassen, die Benutzern eines Netzinfrastruktursystems eines Dritten auf Anfrage zur Verfügung gestellt werden. In mindestens einer Ausführungsform können auch verschiedene Dienste angeboten werden, einschließlich ohne Einschränkung Online-Daten-

speicher- und Sicherungslösungen, webbasierte E-Mail-Dienste, gehostete Office-Suiten und Dienste für die Zusammenarbeit von Dokumenten, Datenbankverwaltung und -verarbeitung, verwaltete technische Supportdienste und/oder Variationen davon. In mindestens einer Ausführungsform können Dienste, die durch ein Drittanbieter-Netzwerkinfrastruktursystem bereitgestellt werden, dynamisch skalieren, um die Bedürfnisse ihrer Benutzer zu erfüllen.

[0080] In mindestens einer Ausführungsform kann eine spezifische Instanziierung eines Dienstes, der durch das Drittanbieter-Netzwerkinfrastruktursystem 1102 bereitgestellt wird, als eine „Dienstinstanz“ bezeichnet werden. In mindestens einer Ausführungsform wird im Allgemeinen jeder Dienst, der einem Benutzer über ein Kommunikationsnetz, wie etwa das Internet, von einem System eines Drittanbieter-Netzdienstanbieters zur Verfügung gestellt wird, als „Netzdienst eines Drittanbieters“ bezeichnet. In mindestens einer Ausführungsform unterscheiden sich in einer öffentlichen Drittanbieter-Netzwerkumgebung Server und Systeme, die das System eines Drittanbieter-Netzwerkdienstanbieters bilden, von den eigenen lokalen Servern und Systemen eines Kunden. In mindestens einer Ausführungsform kann das System eines Drittanbieter-Netzwerkdienstanbieters eine Anwendung hosten, und ein Benutzer kann über ein Kommunikationsnetzwerk wie das Internet auf Anfrage eine Anwendung bestellen und verwenden.

[0081] In mindestens einer Ausführungsform kann ein Dienst in einer Computernetzwerk-Netzwerkinfrastruktur eines Drittanbieters einen geschützten Computernetzwerkzugriff auf Speicher, eine gehostete Datenbank, einen gehosteten Webserver, eine Softwareanwendung oder einen anderen Dienst beinhalten, die einem Benutzer durch einen Drittanbieter-Netzwerkhersteller bereitgestellt werden. In mindestens einer Ausführungsform kann ein Dienst einen passwortgeschützten Zugriff auf einen entfernten Speicher in einem Drittanbieter-Netzwerk über das Internet beinhalten. In mindestens einer Ausführungsform kann ein Dienst eine auf einem Webdienst basierende gehostete relationale Datenbank und eine Middleware-Engine in Skriptsprache zur privaten Verwendung durch einen vernetzten Entwickler beinhalten. In mindestens einer Ausführungsform kann ein Dienst Zugriff auf eine E-Mail-Softwareanwendung beinhalten, die auf der Website eines Drittanbieter-Netzwerkherstellers gehostet wird.

[0082] In mindestens einer Ausführungsform kann das Drittanbieter-Netzwerkinfrastruktursystem 1102 eine Suite von Anwendungen, Middleware und Datenbankdienstangeboten beinhalten, die einem Kunden in Selbstbedienung, abonnementbasiert, elastisch skalierbar, zuverlässig, hochverfügbar und sicher geliefert werden. In mindestens einer Ausführungsform kann das Drittanbieter-Netzwerkinfrastruktursystem 1102 auch „Big Data“-bezogene Berechnungs- und Analysedienste bereitstellen. In mindestens einer Ausführungsform wird der Begriff „Big Data“ allgemein verwendet, um sich auf extrem große Datensätze zu beziehen, die von Analysten und Forschern gespeichert und manipuliert werden können, um große Datenmengen zu visualisieren, Trends zu erkennen und/oder anderweitig mit Daten zu interagieren. In mindestens einer Ausführungsform können Big Data und zugehörige Anwendungen durch ein Infrastruktursystem auf vielen Ebenen und in unterschiedlichen Maßstäben gehostet und/oder manipuliert werden. In mindestens einer Ausführungsform können Dutzende, Hunderte oder Tausende parallel verbundener Prozessoren auf solche Daten einwirken, um sie darzustellen oder externe Kräfte auf Daten oder das, was sie darstellen, zu simulieren. In mindestens einer Ausführungsform können diese Datensätze strukturierte Daten, die beispielsweise in einer Datenbank oder anderweitig gemäß einem strukturierten Modell organisiert sind, und/oder unstrukturierte Daten (z. B. E-Mails, Bilder, Datenblobs (binäre große Objekte), Web-Seiten, komplexe Ereignisverarbeitung) umfassen. In mindestens einer Ausführungsform kann durch Nutzung der Fähigkeit einer Ausführungsform, relativ schnell mehr (oder weniger) Rechenressourcen auf ein Ziel zu konzentrieren, ein Drittanbieter-Netzwerkinfrastruktursystem besser verfügbar sein, um Tasks an großen Datensätzen basierend auf der Nachfrage von einem Unternehmen, einer Regierungsbehörde, einer Forschungseinrichtung, einer Privatperson, einer Gruppe gleichgesinnter Personen oder Organisationen oder einer anderen Einheit auszuführen.

[0083] In mindestens einer Ausführungsform kann das Drittanbieter-Netzwerkinfrastruktursystem 1102 angepasst sein, um das Abonnement eines Kunden für Dienste, die durch das Drittanbieter-Netzwerkinfrastruktursystem 1102 angeboten werden, automatisch bereitzustellen, zu verwalten und zu verfolgen. In mindestens einer Ausführungsform kann das Drittanbieter-Netzwerkinfrastruktursystem 1102 über verschiedene Einsatzmodelle Drittanbieter-Netzwerkdienste bereitstellen. In mindestens einer Ausführungsform können Dienste im Rahmen eines öffentlichen Drittanbieter-Netzwerkmodells bereitgestellt werden, bei dem das Drittanbieter-Netzwerkinfrastruktursystem 1102 im Besitz einer Organisation ist, die Drittanbieter-Netzwerkdienste verkauft und Dienste einer breiten Öffentlichkeit oder unterschiedlichen Industrieunternehmen zur Verfügung stellt. In mindestens einer Ausführungsform können Dienste im Rahmen eines privaten Netzmodells Dritter bereitgestellt werden, in dem das Netzinfrastruktursystem 1102 Dritter ausschließlich für eine einzelne Organisation betrieben wird und Dienste für eine oder mehrere Einheiten innerhalb einer Organisation bereitstel-

len kann. In mindestens einer Ausführungsform können Drittanbieter-Netzwerkdienste auch im Rahmen eines gemeinschaftlichen Drittanbieter-Netzwerkmodells bereitgestellt werden, bei dem das Drittanbieter-Netzwerkinfrastruktursystem 1102 und Dienste, die durch das Drittanbieter-Netzwerkinfrastruktursystem 1102 bereitgestellt werden, von mehreren Organisationen in einer zusammenhängenden Gemeinschaft gemeinsam genutzt werden. In mindestens einer Ausführungsform können Drittanbieter-Netzwerkdienste auch unter einem hybriden Drittanbieter-Netzwerkmodell bereitgestellt werden, das eine Kombination von zwei oder mehr unterschiedlichen Modellen ist.

[0084] In mindestens einer Ausführungsform können Dienste, die vom Netzwerkinfrastruktursystem 1102 eines Drittanbieters bereitgestellt werden, einen oder mehrere Dienste umfassen, die unter der Kategorie Software as a Service (SaaS), der Kategorie Platform as a Service (PaaS), der Kategorie Infrastructure as a Service (laaS) oder unter anderen Kategorien von Diensten, einschließlich Hybriddiensten, bereitgestellt werden. In mindestens einer Ausführungsform kann ein Kunde über eine Abonnementbestellung einen oder mehrere Dienste bestellen, die vom Netzinfrastruktursystem 1102 eines Dritten bereitgestellt werden. In mindestens einer Ausführungsform führt das Drittanbieter-Netzwerkinfrastruktursystem 1102 dann die Verarbeitung durch, um Dienste in einem Abonnementauftrag des Kunden bereitzustellen.

[0085] In mindestens einer Ausführungsform können Dienste, die durch das Drittanbieter-Netzwerkinfrastruktursystem 1102 bereitgestellt werden, ohne Einschränkung Anwendungsdienste, Plattformdienste und Infrastrukturdienste beinhalten. In mindestens einer Ausführungsform können Anwendungsdienste durch ein Drittanbieter-Netzwerkinfrastruktursystem über eine SaaS-Plattform bereitgestellt werden. In mindestens einer Ausführungsform kann die SaaS-Plattform konfiguriert sein, um Drittanbieter-Netzwerkdienste bereitzustellen, die in eine SaaS-Kategorie fallen. In mindestens einer Ausführungsform kann die SaaS-Plattform Fähigkeiten zum Erstellen und Liefern einer Suite von On-Demand-Anwendungen auf einer integrierten Entwicklungs- und Einsatzplattform bereitstellen. In mindestens einer Ausführungsform kann die SaaS-Plattform die zugrundeliegende Software und Infrastruktur zum Bereitstellen von SaaS-Diensten verwalten und steuern. In mindestens einer Ausführungsform können Kunden durch Nutzung von durch eine SaaS-Plattform bereitgestellten Diensten Anwendungen nutzen, die in einem Drittanbieter-Netzwerkinfrastruktursystem ausgeführt werden. In mindestens einer Ausführungsform können Kunden Anwendungsdienste erwerben, ohne dass Kunden separate Lizenzen und Support erwerben müssen. In mindestens einer Ausführungsform können verschiedene unterschiedliche SaaS-Dienste bereitgestellt werden. In mindestens einer Ausführungsform kann dies ohne Einschränkung Dienste beinhalten, die Lösungen für das Vertriebsleistungsmanagement, die Unternehmensintegration und die Geschäftsflexibilität für große Organisationen bereitstellen.

[0086] In mindestens einer Ausführungsform können Plattformdienste durch das Drittanbieter-Netzwerkinfrastruktursystem 1102 über eine PaaS-Plattform bereitgestellt werden. In mindestens einer Ausführungsform kann die PaaS-Plattform konfiguriert sein, um Drittanbieter-Netzwerkdienste bereitzustellen, die in eine PaaS-Kategorie fallen. In mindestens einer Ausführungsform können Plattformdienste ohne Einschränkung Dienste beinhalten, die es Organisationen ermöglichen, vorhandene Anwendungen auf einer geteilten, gemeinsamen Architektur zu konsolidieren, sowie die Fähigkeit, neue Anwendungen zu erstellen, die durch eine Plattform bereitgestellte geteilte Dienste einsetzen. In mindestens einer Ausführungsform kann die PaaS-Plattform die zugrundeliegende Software und Infrastruktur zum Bereitstellen von PaaS-Diensten verwalten und steuern. In mindestens einer Ausführungsform können Kunden PaaS-Dienste, die durch das Drittanbieter-Netzwerkinfrastruktursystem 1102 bereitgestellt werden, erwerben, ohne dass Kunden separate Lizenzen und Support erwerben müssen.

[0087] In mindestens einer Ausführungsform können Kunden durch die Nutzung von Diensten, die durch eine PaaS-Plattform bereitgestellt werden, Programmiersprachen und Tools einsetzen, die durch ein Drittanbieter-Netzwerkinfrastruktursystem unterstützt werden, und auch eingesetzte Dienste steuern. In mindestens einer Ausführungsform können Plattformdienste, die durch ein Drittanbieter-Netzwerkinfrastruktursystem bereitgestellt werden, Datenbank-Drittanbieter-Netzwerkdienste, Middleware-Drittanbieter-Netzwerkdienste und Drittanbieter-Netzwerkdienste beinhalten. In mindestens einer Ausführungsform können Datenbank-Drittanbieter-Netzwerkdienste Shared-Service-Bereitstellungsmodelle unterstützen, die es Organisationen ermöglichen, Datenbankressourcen zu bündeln und Kunden eine Database as a Service in Form eines Datenbank-Drittanbieternetzwerks anzubieten. In mindestens einer Ausführungsform können Middleware-Drittanbieter-Netzwerkdienste eine Plattform für Kunden bereitstellen, um verschiedene Geschäftsanwendungen zu entwickeln und bereitzustellen, und die Drittanbieter-Netzwerkdienste können eine Plattform für Kunden bereitstellen, um Anwendungen in einem Drittanbieter-Netzwerkinfrastruktursystem einzusetzen.

[0088] In mindestens einer Ausführungsform können verschiedene unterschiedliche Infrastrukturdienste durch eine IaaS-Plattform in einem Drittanbieter-Netzwerkinfrastruktursystem bereitgestellt werden. In mindestens einer Ausführungsform erleichtern Infrastrukturdienste die Verwaltung und Kontrolle von zugrunde liegenden Rechenressourcen, wie beispielsweise Speicher, Netzwerke und andere grundlegende Rechenressourcen für Kunden, die Dienste nutzen, die von einer SaaS-Plattform und einer PaaS-Plattform bereitgestellt werden.

[0089] In mindestens einer Ausführungsform kann das Drittanbieter-Netzwerkinfrastruktursystem 1102 auch Infrastrukturressourcen 1130 zum Bereitstellen von Ressourcen beinhalten, die verwendet werden, um Kunden eines Drittanbieter-Netzwerkinfrastruktursystems verschiedene Dienste bereitzustellen. In mindestens einer Ausführungsform können die Infrastrukturressourcen 1130 vorintegrierte und optimierte Kombinationen von Hardware beinhalten, wie beispielsweise Server, Speicher und Netzwerkressourcen, um Dienste auszuführen, die von einer Paas-Plattform und einer Saas-Plattform und anderen Ressourcen bereitgestellt werden.

[0090] In mindestens einer Ausführungsform können Ressourcen im Drittanbieter-Netzwerkinfrastruktursystem 1102 von mehreren Benutzern gemeinsam genutzt und je nach Bedarf dynamisch neu zugewiesen werden. In mindestens einer Ausführungsform können Benutzern in unterschiedlichen Zeitzonen Ressourcen zugewiesen werden. In mindestens einer Ausführungsform kann das Drittanbieter-Netzwerkinfrastruktursystem 1102 einer ersten Gruppe von Benutzern in einer ersten Zeitzone ermöglichen, Ressourcen eines Drittanbieter-Netzwerkinfrastruktursystems für eine festgelegte Anzahl von Stunden zu nutzen, und dann eine Neuzuweisung derselben Ressourcen für eine andere Gruppe von Benutzern, die sich in einer anderen Zeitzone befinden, ermöglichen, wodurch die Nutzung der Ressourcen maximiert wird.

[0091] In mindestens einer Ausführungsform kann eine Reihe interner gemeinsam genutzter Dienste 1132 bereitgestellt werden, die von verschiedenen Komponenten oder Modulen des Drittanbieter-Netzwerkinfrastruktursystems 1102 gemeinsam genutzt werden, um die Bereitstellung von Diensten durch das Drittanbieter-Netzwerkinfrastruktursystem 1102 zu ermöglichen. In mindestens einer Ausführungsform können diese internen gemeinsam genutzten Dienste ohne Einschränkung einen Sicherheits- und Identitätsdienst, einen Integrationsdienst, einen Unternehmens-Repository-Dienst, einen Unternehmensmanagerdienst, einen Virenscan- und Whitelist-Dienst, einen Hochverfügbarkeits-, Sicherungs- und Wiederherstellungsdienst, Dienst zum Ermöglichen von Drittanbieter-Netzwerkunterstützung, einen E-Mail-Dienst, einen Benachrichtigungsdienst, einen Dateiübertragungsdienst und/oder Variationen davon beinhalten.

[0092] In mindestens einer Ausführungsform kann das Netzinfrastruktursystem 1102 eines Dritten eine umfassende Verwaltung von Netzdiensten Dritter (z. B. SaaS-, PaaS- und IaaS-Dienste) in einem Drittanbieter-Netzinfrastruktursystem bereitstellen. In mindestens einer Ausführungsform kann die Drittanbieter-Netzwerkverwaltungsfunktionalität Fähigkeiten zum Bereitstellen, Verwalten und Verfolgen eines durch das Drittanbieter-Netzwerkinfrastruktursystem 1102 empfangenen Abonnements eines Kunden und/oder Variationen davon beinhalten.

[0093] In mindestens einer Ausführungsform kann die Drittanbieter-Netzwerkverwaltungsfunktionalität wie in **Fig. 11** dargestellt durch ein oder mehrere Module bereitgestellt werden, wie etwa ein Auftragsverwaltungsmodul 1120, ein Auftragsorchestrierungsmodul 1122, ein Auftragsbereitstellungsmodul 1124, ein Auftragsverwaltungs- und -überwachungsmodul 1126 und ein Identitätsverwaltungsmodul 1128. In mindestens einer Ausführungsform können diese Module einen oder mehrere Computer und/oder Server beinhalten oder unter Verwendung derselben bereitgestellt werden, die Allzweckcomputer, spezialisierte Server-Computer, Serverfarmen, Servercluster oder jede andere geeignete Anordnung und/oder Kombination sein können.

[0094] In mindestens einer Ausführungsform kann ein Kunde, der eine Client-Vorrichtung, wie etwa die Client-Rechenvorrichtungen 1104, 1106 oder 1108, verwendet, in Schritt 1134 mit dem Drittanbieter-Netzwerkinfrastruktursystem 1102 interagieren, indem er einen oder mehrere Dienste anfordert, die durch das Drittanbieter-Netzwerkinfrastruktursystem 1102 bereitgestellt werden, und einen Auftrag für ein Abonnement für einen oder mehrere Dienste erteilt, die durch das Drittanbieter-Netzwerkinfrastruktursystem 1102 angeboten werden. In mindestens einer Ausführungsform kann ein Kunde auf eine Benutzerschnittstelle (user interface - UI) eines Drittanbieter-Netzwerks zugreifen, wie etwa eine Drittanbieter-Netzwerk-UI 1112, eine Drittanbieter-Netzwerk-UI 1114 und/oder eine Drittanbieter-Netzwerk-UI 1116, und einen Abonnementauftrag über diese UI erteilen. In mindestens einer Ausführungsform können Auftragsinformationen, die durch das Drittanbieter-Netzwerkinfrastruktursystem 1102 als Reaktion darauf, dass ein Kunde einen Auftrag erteilt, empfangen werden, Informationen beinhalten, die einen Kunden und einen oder mehrere Dienste identifizieren, die

durch ein Drittanbieter-Netzwerkinfrastruktursystem 1102 angeboten werden, die ein Kunde abonnieren möchte.

[0095] In mindestens einer Ausführungsform können bei Schritt 1136 von einem Kunden empfangene Auftragsinformationen in einer Auftragsdatenbank 1118 gespeichert werden. In mindestens einer Ausführungsform kann, wenn es sich um einen neuen Auftrag handelt, ein neuer Datensatz für einen Auftrag erstellt werden. In mindestens einer Ausführungsform kann die Auftragsdatenbank 1118 eine von mehreren Datenbanken sein, die durch das Netzinfrastruktursystem 1118 eines Dritten und in Verbindung mit anderen Systemelementen betrieben werden.

[0096] In mindestens einer Ausführungsform können bei Schritt 1138 Auftragsinformation an ein Auftragsverwaltungsmodul 1120 weitergeleitet werden, das konfiguriert sein kann, um Abrechnungs- und Buchhaltungsfunktionen in Bezug auf einen Auftrag auszuführen, wie zum Beispiel das Verifizieren eines Auftrags, und nach der Verifizierung das Verbuchen eines Auftrags.

[0097] In mindestens einer Ausführungsform können bei Schritt 1140 Informationen bezüglich eines Auftrags an ein Auftragsorchestrierungsmodul 1122 übermittelt werden, das, zum Koordinieren der Bereitstellung von Diensten und Ressourcen für einen durch einen Kunden erteilten Auftrag konfiguriert ist. In mindestens einer Ausführungsform kann das Auftragsorchestrierungsmodul 1122 Dienste des Auftragsbereitstellungsmoduls 1124 für die Bereitstellung verwenden. In mindestens einer Ausführungsform ermöglicht das Auftragsorchestrierungsmodul 1122 die Verwaltung von Geschäftsprozessen, die jedem Auftrag zugeordnet sind, und wendet Geschäftslogik an, um zu bestimmen, ob ein Auftrag mit der Bereitstellung fortfahren sollte.

[0098] In mindestens einer Ausführungsform sendet das Auftragsorchestrierungsmodul 1122 bei Schritt 1142 nach Erhalt eines Auftrags für ein neues Abonnement eine Anforderung an das Auftragsbereitstellungsmodul 1124, Ressourcen zuzuweisen und Ressourcen zu konfigurieren, die zum Erfüllen einer Abonnementbestellung benötigt werden. In mindestens einer Ausführungsform ermöglicht das Auftragsbereitstellungsmodul 1124 eine Zuweisung von Ressourcen für durch einen Kunden beauftragte Dienste. In mindestens einer Ausführungsform stellt das Auftragsbereitstellungsmodul 1124 eine Abstraktionsebene zwischen Netzdiensten Dritter, die durch das Drittanbieter-Netzinfrastruktursystem 1100 bereitgestellt werden, und einer physikalischen Implementierungsschicht bereit, die zum Vorsehen von Ressourcen zum Bereitstellen angeforderter Dienste verwendet wird. In mindestens einer Ausführungsform ermöglicht dies, das Auftragsorchestrierungsmodul 1122 von Implementierungsdetails zu isolieren, beispielsweise ob Dienste und Ressourcen tatsächlich in Echtzeit bereitgestellt oder vorab bereitgestellt und nur auf Anfrage zugewiesen/zugewiesen werden.

[0099] In mindestens einer Ausführungsform kann bei Schritt 1144, sobald Dienste und Ressourcen bereitgestellt sind, eine Benachrichtigung an Abonnementkunden gesendet werden, die angibt, dass ein angeforderter Dienst jetzt einsatzbereit ist. In mindestens einer Ausführungsform können Informationen (z. B. ein Link) an einen Kunden gesendet werden, die es einem Kunden ermöglichen, mit der Verwendung der angeforderten Dienste zu beginnen.

[0100] In mindestens einer Ausführungsform kann bei Schritt 1146 ein Abonnementauftrag eines Kunden durch ein Auftragsverwaltungs- und - überwachungsmodul 1126 verwaltet und verfolgt werden. In mindestens einer Ausführungsform kann das Auftragsverwaltungs- und -überwachungsmodul 1126 konfiguriert sein, um Nutzungsstatistiken bezüglich einer Kundennutzung von abonnierten Diensten zu sammeln. In mindestens einer Ausführungsform können Statistiken für eine verwendete Speichermenge, eine übertragene Datenmenge, eine Anzahl von Benutzern und eine Menge an Systembetriebszeit und Systemausfallzeit und/oder Variationen davon gesammelt werden.

[0101] In mindestens einer Ausführungsform kann das Drittanbieter-Netzwerkinfrastruktursystem 1100 ein Identitätsverwaltungsmodul 1128 beinhalten, das zum Bereitstellen von Identitätsdiensten wie etwa Zugriffsverwaltungs- und Autorisierungsdiensten im Drittanbieter-Netzwerkinfrastruktursystem 1100 konfiguriert ist. In mindestens einer Ausführungsform kann das Identitätsverwaltungsmodul 1128 Informationen über Kunden steuern, die Dienste nutzen möchten, die durch das Drittanbieter-Netzwerkinfrastruktursystem 1102 bereitgestellt werden. In mindestens einer Ausführungsform können solche Informationen, die Identitäten solcher Kunden authentifizieren, und Informationen beinhalten, die beschreiben, zu welchen Handlungen diese Kunden in Bezug auf verschiedene Systemressourcen (z. B. Dateien, Verzeichnisse, Anwendungen, Kommunikationsports, Speichersegmente usw.) autorisiert sind. In mindestens einer Ausführungsform kann das Identitätsverwaltungsmodul 1128 auch die Verwaltung von beschreibenden Informationen über jeden Kunden und

darüber, wie und von wem auf diese beschreibenden Informationen zugegriffen und diese modifiziert werden können, beinhalten.

[0102] **Fig. 12** zeigt eine Cloud-Computing-Umgebung 1202 gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform umfasst die Cloud-Computing-Umgebung 1202 ein oder mehrere Computersysteme/Server 1204, mit denen Rechenvorrichtungen wie ein persönlicher digitaler Assistent (PDA) oder ein Mobiltelefon 1206A, ein Desktop-Computer 1206B, ein Laptop-Computer 1206C und/oder ein Autocomputersystem 1206N kommunizieren. In mindestens einer Ausführungsform ermöglicht dies, dass Infrastruktur, Plattformen und/oder Software als Dienste von der Cloud-Computing-Umgebung 1202 angeboten werden, sodass nicht jeder Client diese Ressourcen separat pflegen muss. Es versteht sich, dass die in **Fig. 12** gezeigten Arten von Rechenvorrichtungen 1206A-N nur veranschaulichend sein sollen und dass die Cloud-Computing-Umgebung 1202 mit jeder Art von computergestützter Vorrichtung über jede Art von Netzwerk und/oder Netzwerk-/adressierbare Verbindung (z. B. über einen Webbrowser) kommunizieren kann.

[0103] In mindestens einer Ausführungsform ist ein Computersystem/Server 1204, das bzw. der als ein Cloud-Computing-Knoten bezeichnet werden kann, mit zahlreichen anderen Allzweck- oder Spezialzweck-Rechensystemumgebungen oder - konfigurationen betriebsfähig. In mindestens einer Ausführungsform umfassen Computersysteme, Umgebungen und/oder Konfigurationen, die zur Verwendung mit dem Computersystem/Server 1204 geeignet sein können, Personalcomputersysteme, Servercomputersysteme, Thin Clients, Thick Clients, Handheld- oder Laptop-Geräte, Multiprozessorsysteme, mikroprozessorbasierte Systeme, Set-Top-Boxen, programmierbare Unterhaltungselektronik, Netzwerk-PCs, Minicomputersysteme, Mainframe-Computersysteme und verteilte Cloud-Computing-Umgebungen, die eines der oben genannten Systeme oder Geräte beinhalten, und/oder Variationen davon.

[0104] In mindestens einer Ausführungsform kann das Computersystem/der Server 1204 in einem allgemeinen Kontext von durch ein Computersystem ausführbaren Anweisungen, wie etwa Programmmodulen, die durch ein Computersystem ausgeführt werden, beschrieben werden. In mindestens einer Ausführungsform beinhalten Programmmodule Routinen, Programme, Objekte, Komponenten, Datenstrukturen usw., die bestimmte Aufgaben ausführen oder bestimmte abstrakte Datentypen implementieren. In mindestens einer Ausführungsform kann das beispielhafte Computersystem/der beispielhafte Server 1204 in verteilten Cloud-Computing-Umgebungen in die Praxis umgesetzt sein, in denen Tasks durch entfernte Verarbeitungsvorrichtungen durchgeführt werden, die über ein Kommunikationsnetzwerk miteinander verbunden sind. In mindestens einer Ausführungsform können sich in einer verteilten Cloud-Computing-Umgebung Programmmodule sowohl auf lokalen als auch entfernten Computersystemspeichermedien, einschließlich Speichervorrichtungen, befinden.

[0105] **Fig. 13** zeigt einen Satz von funktionalen Abstraktionsschichten, der gemäß mindestens einer Ausführungsform durch die Cloud-Computing-Umgebung 1202 (**Fig. 12**) bereitgestellt wird. Es sollte im Voraus verstanden werden, dass die in **Fig. 13** gezeigten Komponenten, Schichten und Funktionen nur zur Veranschaulichung gedacht sind und Komponenten, Schichten und Funktionen variieren können.

[0106] In mindestens einer Ausführungsform beinhaltet die Hardware- und Softwareschicht 1302 Hardware- und Softwarekomponenten. In mindestens einer Ausführungsform beinhalten die Hardwarekomponenten Mainframes, verschiedene Server, die auf einer RISC-Architektur (Reduced Instruction Set Computer) basieren, verschiedene Rechensysteme, Superrechensysteme, Speichervorrichtungen, Netzwerke, Netzwerkkomponenten und/oder Variationen davon. In mindestens einer Ausführungsform beinhalten die Softwarekomponenten Netzwerkanwendungsserversoftware, verschiedene Anwendungsserversoftware, verschiedene Datenbanksoftware und/oder Variationen davon.

[0107] In mindestens einer Ausführungsform stellt die Virtualisierungsschicht 1302 eine Abstraktionsschicht bereit, von der die folgenden beispielhaften virtuellen Einheiten bereitgestellt werden können: virtuelle Server, virtueller Speicher, virtuelle Netzwerke, einschließlich virtueller privater Netzwerke, virtuelle Anwendungen, virtuelle Clients und/oder Variationen davon.

[0108] In mindestens einer Ausführungsform stellt die Verwaltungsschicht 1306 verschiedene Funktionen bereit. In mindestens einer Ausführungsform stellt Ressourcenbereitstellung eine dynamische Beschaffung von Rechenressourcen und anderen Ressourcen bereit, die genutzt werden, um Tasks innerhalb einer Cloud-Computing-Umgebung durchzuführen. In mindestens einer Ausführungsform stellt eine Verbrauchserfassung eine Nutzungsverfolgung, wenn Ressourcen innerhalb einer Cloud-Computing-Umgebung genutzt werden, und eine Abrechnung oder Fakturierung für den Verbrauch dieser Ressourcen bereit. In mindestens

einer Ausführungsform können Ressourcen Anwendungssoftwarelizenzen umfassen. In mindestens einer Ausführungsform stellt eine Sicherheit eine Identitätsprüfung für Benutzer und Tasks sowie Schutz für Daten und andere Ressourcen bereit. In mindestens einer Ausführungsform stellt die Benutzerschnittstelle sowohl Benutzern als auch Systemadministratoren Zugriff auf eine Cloud-Computing-Umgebung bereit. In mindestens einer Ausführungsform stellt die Dienstebenenverwaltung die Zuweisung und Verwaltung von Cloud-Computing-Ressourcen bereit, sodass die erforderlichen Dienstebenen erfüllt werden. In mindestens einer Ausführungsform stellt die Verwaltung der Dienstleistungsvereinbarung (Service Level Agreement - SLA) eine Vorab-Anordnung für und Beschaffung von Cloud-Computing-Ressourcen bereit, für die gemäß einem SLA eine zukünftige Anforderung erwartet wird.

[0109] In mindestens einer Ausführungsform stellt die Arbeitslastenschicht 1308 Funktionalität bereit, für die eine Cloud-Computing-Umgebung verwendet wird. In mindestens einer Ausführungsform umfassen Arbeitslasten und Funktionen, die von dieser Schicht bereitgestellt werden können: Mapping und Navigation, Softwareentwicklung und -verwaltung, Bildungsdienste, Datenanalyse und - verarbeitung, Transaktionsverarbeitung und Dienstbereitstellung.

SUPERCOMPUTER

[0110] Die folgenden Figuren legen ohne Einschränkung Supercomputer-basierte Systeme dar, die zum Implementieren mindestens einer Ausführungsform verwendet werden können.

[0111] In mindestens einer Ausführungsform kann sich ein Supercomputer auf ein Hardwaresystem beziehen, das im Wesentlichen Parallelität aufweist und mindestens einen Chip umfasst, wobei Chips in einem System durch ein Netzwerk miteinander verbunden und in hierarchisch organisierten Gehäusen angeordnet sind. In mindestens einer Ausführungsform ist ein großes Hardwaresystem, das einen Maschinenraum mit mehreren Racks füllt, von denen jedes mehrere Platinen/Rack-Module enthält, von denen jedes mehrere Chips enthält, die alle durch ein skalierbares Netzwerk miteinander verbunden sind, mindestens eine Ausführungsform eines Supercomputers. In mindestens einer Ausführungsform ist ein einzelnes Rack eines solchen großen Hardwaresystems mindestens eine andere Ausführungsform eines Supercomputers. In mindestens einer Ausführungsform kann ein einzelner Chip, der eine beträchtliche Parallelität aufweist und mehrere Hardwarekomponenten enthält, gleichermaßen als Supercomputer angesehen werden, da mit abnehmenden Merkmalsgrößen auch eine Menge an Hardware, die in einen einzelnen Chip integriert werden kann, zunehmen kann.

[0112] **Fig. 14** zeigt einen Supercomputer auf Chipebene gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform wird innerhalb eines FPGA- oder ASIC-Chips der Hauptrechenaufwand in endlichen Zustandsmaschinen (SM) (1404) durchgeführt, die als Thread-Einheiten bezeichnet werden. In mindestens einer Ausführungsform verbinden Task- und Synchronisationsnetzwerke (1402) endliche Zustandsmaschinen und werden verwendet, um Threads abzuschicken und Vorgänge in der richtigen Reihenfolge auszuführen. In mindestens einer Ausführungsform wird unter Verwendung von Speichernetzwerken (1406, 1410) auf eine partitionierte Mehrebenen-Cache-Hierarchie (1408, 1412) auf einem Chip zugegriffen. In mindestens einer Ausführungsform wird auf den chipexternen Speicher unter Verwendung von Speichersteuerungen (1416) und einem chipexternen Speichernetzwerk (1414) zugegriffen. In mindestens einer Ausführungsform wird eine E/A-Steuerung (1418) für die chipübergreifende Kommunikation verwendet, wenn eine Ausgestaltung nicht in einen einzelnen Logikchip passt.

[0113] **Fig. 15** zeigt einen Supercomputer auf Rackmodulebene gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform gibt es innerhalb eines Rack-Moduls mehrere FPGA- oder ASIC-Chips (1502), die mit einer oder mehreren DRAM-Einheiten (1504) verbunden sind, die einen Hauptbeschleunigerspeicher bilden. In mindestens einer Ausführungsform ist jeder FPGA/ASIC-Chip mit seinem benachbarten FPGA/ASIC-Chip unter Verwendung von breiten Bussen auf einer Platine mit differentieller Hochgeschwindigkeitssignalisierung (1506) verbunden. In mindestens einer Ausführungsform ist jeder FPGA/ASIC-Chip auch mit mindestens einem seriellen Hochgeschwindigkeitskommunikationskabel verbunden.

[0114] **Fig. 16** zeigt einen Supercomputer auf Rackebene gemäß mindestens einer Ausführungsform. **Fig. 17** veranschaulicht einen Supercomputer auf Gesamtsystemebene gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform werden unter Bezugnahme auf **Fig. 16** und **Fig. 17** zwischen Rack-Modulen in einem Rack und über Racks hinweg durch ein gesamtes System serielle optische Hochgeschwindigkeits- oder Kupferkabel (1602, 1702) verwendet, um ein skalierbares, möglicherweise unvollständiges Hypercube-Netzwerk zu realisieren. In mindestens einer Ausführungsform ist einer der

FPGA/ASIC-Chips eines Beschleunigers über eine PCI-Express-Verbindung (1704) mit einem Host-System verbunden. In mindestens einer Ausführungsform umfasst das Host-System einen Host-Mikroprozessor (1708), auf dem ein Softwareteil einer Anwendung läuft, und einen Speicher, der aus einer oder mehreren Host-Speicher-DRAM-Einheiten (1706) besteht, der mit dem Speicher auf einem Beschleuniger kohärent gehalten wird. In mindestens einer Ausführungsform kann das Host-System ein separates Modul auf einem der Racks sein oder kann in eines der Module eines Supercomputers integriert sein. In mindestens einer Ausführungsform stellt eine Cube-verbundene Zyklen-Topologie Kommunikationsverbindungen bereit, um ein Hypercube-Netzwerk für einen großen Supercomputer zu schaffen. In mindestens einer Ausführungsform kann eine kleine Gruppe von FPGA/ASIC-Chips auf einem Rack-Modul als ein einzelner Hypercube-Knoten fungieren, sodass eine Gesamtzahl von externen Verbindungen jeder Gruppe im Vergleich zu einem einzelnen Chip erhöht ist. In mindestens einer Ausführungsform enthält eine Gruppe die Chips A, B, C und D auf einem Rack-Modul mit internen breiten differentiellen Bussen, die A, B, C und D in einer Torus-Organisation verbinden. In mindestens einer Ausführungsform gibt es 12 serielle Kommunikationskabel, die ein Rack-Modul mit der Außenwelt verbinden. In mindestens einer Ausführungsform ist Chip A auf einem Rack-Modul mit den seriellen Kommunikationskabeln 0, 1, 2 verbunden. In mindestens einer Ausführungsform ist Chip B mit den Kabeln 3, 4, 5 verbunden. In mindestens einer Ausführungsform ist Chip C mit 6, 7, 8 verbunden. In mindestens einer Ausführungsform ist Chip D mit 9, 10, 11 verbunden. In mindestens einer Ausführungsform kann eine gesamte Gruppe {A, B, C, D}, die ein Rack-Modul bildet, einen Hypercube-Knoten innerhalb eines Supercomputersystems mit bis zu 212 = 4096 Rack-Modulen (16384 FPGA/ASIC-Chips) bilden. In mindestens einer Ausführungsform muss eine Nachricht zuerst an Chip B mit einer On-Board-Differential-Wide-Bus-Verbindung geroutet werden, damit Chip A eine Nachricht auf Verbindung 4 der Gruppe {A, B, C, D} aussendet. In mindestens einer Ausführungsform muss eine Nachricht, die in einer Gruppe {A, B, C, D} auf Verbindung 4 ankommt (d. h. bei B ankommt), die für Chip A bestimmt ist, auch zuerst an einen korrekten Zielchip (A) innerhalb einer Gruppe {A, B, C, D} intern geroutet werden. In mindestens einer Ausführungsform können auch parallele Supercomputersysteme anderer Größen implementiert sein.

KÜNSTLICHE INTELLIGENZ

[0115] Die folgenden Figuren legen ohne Einschränkung beispielhafte Systeme auf Basis künstlicher Intelligenz dar, die verwendet werden können, um mindestens eine Ausführungsform zu implementieren.

[0116] **Fig. 18A** zeigt Inferenz- und/oder Trainingslogik 1815, die zum Durchführen von Inferenz- und/oder Trainingsvorgängen in Verbindung mit einer oder mehreren Ausführungsformen verwendet wird. Details bezüglich der Inferenz- und/oder Trainingslogik 1815 werden nachfolgend in Verbindung mit den **Fig. 18A** und/oder **Fig. 18B** bereitgestellt.

[0117] In mindestens einer Ausführungsform kann die Inferenz- und/oder Trainingslogik 1815 ohne Einschränkung einen Code- und/oder Datenspeicher 1801 beinhalten, um Vorwärts- und/oder Ausgabegewichtungs- und/oder Eingabe-/Ausgabedaten und/oder andere Parameter zu speichern, um Neuronen oder Schichten eines neuronalen Netzwerkes zu konfigurieren, das in Aspekten einer oder mehrerer Ausführungsformen trainiert und/oder zum Ableiten verwendet wird. In mindestens einer Ausführungsform kann die Trainingslogik 1815 einen Code- und/oder Datenspeicher 1801 beinhalten oder mit diesem gekoppelt sein, um Graphencode oder andere Software zu speichern, um die Zeitsteuerung und/oder die Reihenfolge zu steuern, in der Gewichtungs- und/oder andere Parameterinformationen geladen werden sollen, um Logik zu konfigurieren, die Ganzzahl- und/oder Gleitkommaeinheiten (zusammen arithmetisch-logische Einheiten (arithmetic logic units - ALU)) beinhalten. In mindestens einer Ausführungsform lädt Code, wie etwa Graphencode, Gewichtungs- oder andere Parameterinformationen in Prozessor-ALUs auf Grundlage einer Architektur eines neuronalen Netzes, dem derartiger Code entspricht. In mindestens einer Ausführungsform speichert der Code- und/oder Datenspeicher 1801 Gewichtungsparameter und/oder Ein-/Ausgabedaten jeder Schicht eines neuronalen Netzwerkes, die in Verbindung mit einer oder mehreren Ausführungsformen während der Vorwärtspropagierung von Ein-/Ausgabedaten und/oder Gewichtungsparametern beim Trainieren und/oder Ableiten unter Verwendung von Aspekten einer oder mehrerer Ausführungsformen trainiert oder verwendet wird. In mindestens einer Ausführungsform kann ein beliebiger Teil des Code- und/oder Datenspeichers 1801 in einem anderen chipinternen oder chipexternen Datenspeicher beinhaltet sein, darunter einem L1-, L2- oder L3-Cache oder Systemspeicher eines Prozessors.

[0118] In mindestens einer Ausführungsform kann ein beliebiger Abschnitt des Code- und/oder Datenspeichers 1801 zu einem oder mehreren Prozessoren oder anderen Hardware-Logikvorrichtungen oder -Schaltungen intern oder extern sein. In mindestens einer Ausführungsform kann Code und/oder Code- und/oder Datenspeicher 1801 ein Cache-Speicher, ein dynamischer zufällig adressierbarer Speicher (dynamic rand-

omly 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 eine Wahl, ob der Code und/oder Code- und/oder Datenspeicher 1801 zu einem Prozessor intern oder extern ist oder DRAM, SRAM, Flash oder einen anderen Speichertyp umfasst, von dem verfügbaren chipinternen oder chipexternen Speicher, den Latenzanforderungen der durchgeführten Trainings- und/oder Inferenzfunktionen, der Batch-Größe der Daten, die beim Ableiten und/oder Trainieren eines neuronalen Netzwerkes verwendet werden, oder einer Kombination dieser Faktoren abhängen.

[0119] In mindestens einer Ausführungsform kann die Inferenz- und/oder Trainingslogik 1815 ohne Einschränkung einen Code- und/oder Datenspeicher 1805 beinhalten, um Rückwärts- und/oder Ausgabegewichtungs- und/oder Eingabe-/Ausgabedaten zu speichern, die Neuronen oder Schichten eines neuronalen Netzwerkes entsprechen, das in Aspekten einer oder mehrerer Ausführungsformen trainiert und/oder zum Inferenzieren verwendet wird. In mindestens einer Ausführungsform speichert der Code- und/oder Datenspeicher 1805 Gewichtungsparameter und/oder Eingabe-/Ausgabedaten jeder Schicht eines neuronalen Netzwerkes, die in Verbindung mit einer oder mehreren Ausführungsformen während der Rückwärtspropagierung von Eingabe-/Ausgabedaten und/oder Gewichtungsparametern während des Trainierens und/oder Ableitens unter Verwendung von Aspekten einer oder mehrerer Ausführungsformen trainiert oder verwendet wird. In mindestens einer Ausführungsform kann die Trainingslogik 1815 einen Code- und/oder Datenspeicher 1805 beinhalten oder mit diesem gekoppelt sein, um Diagrammcode oder andere Software zu speichern, um den Zeitablauf und/oder die Reihenfolge zu steuern, in der Gewichtungs- und/oder andere Parameterinformationen geladen werden sollen, um Logik zu konfigurieren, die Ganzzahl- und/oder Gleitkommaeinheiten (zusammen arithmetisch-logische Einheiten (arithmetic logic units - ALU)) beinhalten.

[0120] In mindestens einer Ausführungsform bewirkt Code, wie etwa Graphencode, das Laden von Gewichtungs- oder anderen Parameterinformationen in Prozessor-ALU auf Grundlage einer Architektur eines neuronalen Netzes, dem derartiger Code entspricht. In mindestens einer Ausführungsform kann ein beliebiger Teil des Code- und/oder Datenspeichers 1805 in einem anderen chipinternen oder chipexternen Datenspeicher beinhaltet sein, darunter einem L1-, L2- oder L3-Cache oder Systemspeicher eines Prozessors. In mindestens einer Ausführungsform kann ein beliebiger Abschnitt des Code- und/oder Datenspeichers 1805 zu einem oder mehreren Prozessoren oder anderen Hardware-Logikvorrichtungen oder - Schaltungen intern oder extern sein. In mindestens einer Ausführungsform kann der Code- und/oder Datenspeicher 1805 Cache-Speicher, DRAM, SRAM, nichtflüchtiger Speicher (z. B. Flash-Speicher) oder anderer Speicher sein. In mindestens einer Ausführungsform kann die Wahl, ob Code- und/oder der Datenspeicher 1805 in mindestens einer Ausführungsform intern oder extern zu einem Prozessor ist oder DRAM, SRAM, Flash oder eine andere Speicherart umfasst, von verfügbarem chipinternen im Vergleich zu chipexternem Speicher, Latenzanforderungen von ausgeführten Trainings- und/oder Inferenzfunktionen, Batchgröße von Daten, die beim Ableiten und/oder Trainieren eines neuronalen Netzwerkes verwendet werden, oder einer Kombination dieser Faktoren abhängen.

[0121] In mindestens einer Ausführungsform können der Code- und/oder Datenspeicher 1801 und der Code- und/oder Datenspeicher 1805 separate Speicherstrukturen sein. In mindestens einer Ausführungsform können der Code- und/oder Datenspeicher 1801 und der Code- und/oder Datenspeicher 1805 eine kombinierte Speicherstruktur sein. In mindestens einer Ausführungsform können der Code- und/oder Datenspeicher 1801 und der Code- und/oder Datenspeicher 1805 teilweise kombiniert und teilweise separat sein. In mindestens einer Ausführungsform kann ein beliebiger Abschnitt des Code- und/oder Datenspeichers 1801 und des Code- und/oder Datenspeichers 1805 in einem anderen chipinternen oder chipexternen Datenspeicher, darunter des L1-, L2- oder L3-Caches oder Systemspeichers eines Prozessors, enthalten sein.

[0122] Bei mindestens einer Ausführungsform kann die Ableitungs- und/oder Trainingslogik 1815 ohne Einschränkung Folgendes beinhalten: eine oder mehrere arithmetische Logikeinheit(en) („ALU(s)“) 1810, die Ganzzahl- und/oder Gleitkommaeinheiten beinhalten, um logische und/oder mathematische Operationen auszuführen, die mindestens zum Teil auf Trainings- und/oder Ableitungscode (z. B. Graphcode) basieren oder durch diesen angezeigt werden, deren Ergebnis Aktivierungen (z. B. Ausgabewerte von Schichten oder Neuronen innerhalb eines neuronalen Netzwerks) erzeugen kann, die in einem Aktivierungsspeicher 1820 gespeichert sind, die Funktionen von Eingabe/Ausgabe- und/oder Gewichtsparameterdaten sind, die in dem Code- und/oder Datenspeicher 1801 und/oder Code- und/oder Datenspeicher 1805 gespeichert sind. In mindestens einer Ausführungsform werden in dem Aktivierungsspeicher 1820 gespeicherte Aktivierungen gemäß linearer algebraischer und/oder matrixbasierter Mathematik erzeugt, die durch die ALU(s) 1810 als Reaktion auf das Durchführen von Anweisungen oder anderem Code durchgeführt wird, wobei im Code-

und/oder Datenspeicher 1805 und/oder im Datenspeicher 1801 gespeicherte Gewichtungswerte als Operanden zusammen mit anderen Werten, wie etwa Verzerrungswerten, Gradienteninformationen, Momentwerten oder anderen Parametern oder Hyperparametern, verwendet werden, von denen beliebige oder alle im Code- und/oder Datenspeicher 1805 oder im Code- und/oder Datenspeicher 1801 oder einem anderen chipinternen oder -externen Speicher gespeichert sein können.

[0123] Bei mindestens einer Ausführungsform sind die ALU(s) 1810 in einem oder mehreren Prozessoren oder anderen Hardwarelogikvorrichtungen oder -schaltungen beinhaltet, wohingegen bei einer weiteren Ausführungsform die ALU(s) 1810 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 1810 innerhalb der Ausführungseinheiten eines Prozessors oder anderweitig innerhalb einer Bank von ALUs enthalten sein, worauf die Ausführungseinheiten eines Prozessors entweder innerhalb des gleichen Prozessors oder verteilt auf unterschiedliche Prozessoren unterschiedlichen Typs (z. B. zentrale Verarbeitungseinheiten, Grafikverarbeitungseinheiten, Festfunktionseinheiten usw.) zugreifen können. In mindestens einer Ausführungsform können sich der Code- und/oder Datenspeicher 1801, der Code- und/oder Datenspeicher 1805 und der Aktivierungsspeicher 1820 einen Prozessor oder eine andere Hardware-Logikvorrichtung oder -schaltung teilen, während sie sich in einer anderen Ausführungsform in unterschiedlichen Prozessoren oder anderen Hardware-Logikvorrichtungen oder -Schaltungen oder in einer Kombination aus gleichen und unterschiedlichen Prozessoren oder anderen Hardware-Logikvorrichtungen oder -Schaltungen befinden können. In mindestens einer Ausführungsform kann ein beliebiger Abschnitt des Aktivierungsspeichers 1820 in einem anderen chipinternen oder chipexternen Datenspeicher, darunter dem L1-, L2- oder L3-Cache oder Systemspeicher eines Prozessors, enthalten sein. Des Weiteren kann der Inferenz- und/oder Trainingscode mit anderem Code gespeichert sein, auf den ein Prozessor oder eine andere Hardware-Logik oder -Schaltung zugreifen kann und der unter Verwendung der Abruf-, Decodier-, Planungs-, Ausführungs-, Ausscheidungs- und/oder anderen Logikschaltungen eines Prozessors abgerufen und/oder verarbeitet wird.

[0124] In mindestens einer Ausführungsform kann der Aktivierungsspeicher 1820 Cache-Speicher, DRAM, SRAM, nichtflüchtiger Speicher (z. B. Flash-Speicher) oder anderer Speicher sein. In mindestens einer Ausführungsform kann sich der Aktivierungsspeicher 1820 ganz oder teilweise innerhalb oder außerhalb eines oder mehrerer Prozessoren oder anderer logischer Schaltungen befinden. In mindestens einer Ausführungsform kann die Wahl, ob der Aktivierungsspeicher 1820 in mindestens einer Ausführungsform 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 Inferenzfunktionen, Batchgröße von Daten, die beim Inferenzen und/oder Trainieren eines neuronalen Netzwerkes verwendet werden, oder einer Kombination dieser Faktoren abhängen.

[0125] In mindestens einer Ausführungsform kann die Inferenz- und/oder Trainingslogik 1815, die in **Fig. 18A** dargestellt ist, in Verbindung mit einer anwendungsspezifischen integrierten Schaltung („ASIC“) verwendet werden, wie etwa einer TensorFlow® Verarbeitungseinheit von Google, einer Inferenzverarbeitungseinheit (inference processing unit - IPU) von Graphcore™ oder einem Nervana@ (z. B. „Lake Crest“)-Prozessor von Intel Corp. In mindestens einer Ausführungsform kann die Inferenz-und/oder Trainingslogik 1815, die in **Fig. 18A** veranschaulicht ist, in Verbindung mit Hardware einer Zentraleinheit („CPU“), Hardware einer Grafikverarbeitungseinheit („GPU“) oder anderer Hardware, wie etwa feldprogrammierbaren Gatearrays („FPGA“) verwendet werden.

[0126] **Fig. 18B** zeigt die Inferenz- und/oder Trainingslogik 1815 gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform kann die Inferenz- und/oder Trainingslogik 1815 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 Netzes entsprechen. In mindestens einer Ausführungsform kann die Inferenz- und/oder Trainingslogik 1815, die in **Fig. 18B** veranschaulicht ist, in Verbindung mit einer anwendungsspezifischen integrierten Schaltung (ASIC) verwendet werden, wie etwa einer TensorFlow® Verarbeitungseinheit von Google, einer Inferenzverarbeitungseinheit (IPU) von Graphcore™ oder einem Nervana®
(z. B. „Lake Crest“)-Prozessor von Intel Corp. In mindestens einer Ausführungsform kann die Inferenz-und/oder Trainingslogik 1815, die in **Fig. 18B** veranschaulicht ist, in Verbindung mit Hardware einer Zentraleinheit (CPU), Hardware einer Grafikverarbeitungseinheit (GPU) oder anderer Hardware, wie etwa feldprogrammierbaren Gatearrays (FPGA) verwendet werden. In mindestens einer Ausführungsform beinhaltet die Inferenz- und/oder Trainingslogik 1815 ohne Einschränkung den Code- und/oder Datenspeicher 1801 und den Code-

und/oder Datenspeicher 1805, die zum Speichern von Code (z. B. Graphencode), Gewichtungswerten und/oder anderen Informationen, einschließlich Vorspannungswerten, Gradienteninformationen, Momentwerten und/oder anderer Parameter- oder Hyperparameterinformationen, verwendet werden können. In mindestens einer Ausführungsform, die in **Fig. 18B** veranschaulicht ist, ist jeder von dem Code- und/oder Datenspeicher 1801 und dem Code- und/oder Datenspeicher 1805 einer dedizierten Berechnungsressource, wie etwa der Berechnungshardware 1802 bzw. der Berechnungshardware 1806, zugeordnet. In mindestens einer Ausführungsform umfasst jede von der Berechnungshardware 1802 und der Berechnungshardware 1806 eine oder mehrere ALU, die mathematische Funktionen, wie etwa lineare algebraische Funktionen, nur an Informationen durchführen, die im Code- und/oder Datenspeicher 1801 und Code- und/oder Datenspeicher 1805 gespeichert sind, deren Ergebnis im Aktivierungsspeicher 1820 gespeichert ist.

[0127] In mindestens einer Ausführungsform entspricht sowohl der Code- und/oder Datenspeicher 1801 und 1805 als auch die entsprechende Berechnungshardware 1802 bzw. 1806 unterschiedlichen Schichten eines neuronalen Netzwerkes, sodass die resultierende Aktivierung von einem Speicher-/Rechenpaar 1801/1802 des Code- und/oder Datenspeichers 1801 und der Berechnungshardware 1802 als Eingabe einem nächsten Speicher-/Rechenpaar 1805/1806 des Code- und/oder Datenspeichers 1805 und der Berechnungshardware 1806 bereitgestellt wird, um eine konzeptionelle Organisation eines neuronalen Netzwerkes widerzuspiegeln. In mindestens einer Ausführungsform kann jedes der Speicher-/Rechenpaare 1801/1802 und 1805/1806 mehr als einer Schicht eines neuronalen Netzwerkes entsprechen. In mindestens einer Ausführungsform können zusätzliche Speicher/Berechnungspaare (nicht gezeigt) nach oder parallel zu den Speicher/Berechnungspaaren 1801/1802 und 1805/1806 in der Inferenz- und/oder Trainingslogik 1815 beinhaltet sein.

[0128] **Fig. 19** zeigt das Training und den Einsatz eines tiefen neuronalen Netzwerkes gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform wird das untrainierte neuronale Netz 1906 unter Verwendung eines Trainingsdatensatzes 1902 trainiert. In mindestens einer Ausführungsform ist der Trainingsrahmen 1904 ein PyTorch-Rahmen, wohingegen der Trainingsrahmen 1904 in anderen Ausführungsformen ein TensorFlow-, Boost-, Caffe-, Microsoft-Cognitive-Toolkit/CNTK-, MXNet-, Chainer-, Keras-, Deeplearning4j- oder ein anderer Trainingsrahmen ist. In mindestens einer Ausführungsform trainiert der Trainingsrahmen 1904 ein untrainiertes neuronales Netzwerk 1906 und ermöglicht, dass es unter Verwendung der hierin beschriebenen Verarbeitungsressourcen trainiert wird, um ein trainiertes neuronales Netzwerk 1908 zu erzeugen. In mindestens einer Ausführungsform können die Gewichtungen zufällig oder durch Vorabtraining unter Verwendung eines Deep-Belief-Netzes gewählt werden. In mindestens einer Ausführungsform kann das Training entweder auf überwachte, teilweise überwachte oder nicht überwachte Weise durchgeführt werden.

[0129] In mindestens einer Ausführungsform wird das untrainierte neuronale Netzwerk 1906 unter Verwendung von überwachtem Lernen trainiert, wobei der Trainingsdatensatz 1902 eine Eingabe beinhaltet, die mit einer gewünschten Ausgabe für eine Eingabe gepaart ist, oder wobei der Trainingsdatensatz 1902 eine Eingabe beinhaltet, die eine bekannte Ausgabe aufweist, und eine Ausgabe des neuronalen Netzwerkes 1906 manuell bewertet wird. In mindestens einer Ausführungsform wird das untrainierte neuronale Netzwerk 1906 auf überwachte Weise trainiert und verarbeitet Eingaben aus dem Trainingsdatensatz 1902 und vergleicht die resultierenden Ausgaben mit einem Satz von erwarteten oder gewünschten Ausgaben. In mindestens einer Ausführungsform werden Fehler dann durch das untrainierte neuronale Netzwerk 1906 rückpropagiert. In mindestens einer Ausführungsform stellt der Trainingsrahmen 1904 Gewichtungen ein, die das untrainierte neuronale Netzwerk 1906 steuern. In mindestens einer Ausführungsform beinhaltet das Trainings-Framework 1904 Tools, um zu überwachen, wie gut das untrainierte neuronale Netz 1906 zu einem Modell konvergiert, wie etwa einem trainierten neuronalen Netz 1908, das dazu geeignet ist, korrekte Antworten zu erzeugen, wie etwa in einem Ergebnis 1914, die auf Eingabedaten, wie etwa einem neuen Datensatz 1912, basieren. In mindestens einer Ausführungsform trainiert der Trainingsrahmen 1904 das untrainierte neuronale Netzwerk 1906 wiederholt, während Gewichtungen eingestellt werden, um eine Ausgabe des untrainierten neuronalen Netzwerkes 1906 unter Verwendung einer Verlustfunktion und eines Einstellungsalgorithmus, wie etwa des stochastischen Gradientenabstiegs, zu verfeinern. In mindestens einer Ausführungsform trainiert der Trainingsrahmen 1904 das untrainierte neuronale Netzwerk 1906, bis das untrainierte neuronale Netzwerk 1906 eine gewünschte Genauigkeit erreicht. In mindestens einer Ausführungsform kann das trainierte neuronale Netzwerk 1908 dann zum Implementieren einer beliebigen Anzahl von Operationen des maschinellen Lernens eingesetzt werden.

[0130] In mindestens einer Ausführungsform wird das untrainierte neuronale Netzwerk 1906 unter Verwendung von nicht überwachtem Lernen trainiert, wobei das untrainierte neuronale Netzwerk 1906 versucht, sich selbst unter Verwendung von unbeschrifteten Daten zu trainieren. In mindestens einer Ausführungsform

beinhaltet der Trainingsdatensatz 1902 für nicht überwachtes Lernen Eingabedaten ohne zugeordnete Ausgabedaten oder „Ground-Truth“-Daten. In mindestens einer Ausführungsform kann das untrainierte neuronale Netzwerk 1906 Gruppierungen innerhalb des Trainingsdatensatzes 1902 lernen und bestimmen, wie einzelne Eingaben mit dem untrainierten Datensatz 1902 in Bezug stehen. In mindestens einer Ausführungsform kann nicht überwachtes Training verwendet werden, um eine selbstorganisierende Karte im trainierten neuronalen Netzwerk 1908 zu erzeugen, die Operationen durchführen kann, die beim Reduzieren der Dimensionalität des neuen Datensatzes 1912 nützlich sind. In mindestens einer Ausführungsform kann nicht überwachtes Training auch zur Anomaliedetektion verwendet werden, was die Identifizierung von Datenpunkten im neuen Datensatz 1912 ermöglicht, die von normalen Mustern des neuen Datensatzes 1912 abweichen.

[0131] In mindestens einer Ausführungsform kann halbüberwachtes Lernen verwendet werden, wobei es sich um eine Technik handelt, bei der der Trainingsdatensatz 1902 eine Mischung aus beschrifteten und unbeschrifteten Daten beinhaltet. In mindestens einer Ausführungsform kann das Trainings-Framework 1904 verwendet werden, um inkrementelles Lernen durchzuführen, wie etwa durch Transferlerntechniken. In mindestens einer Ausführungsform ermöglicht das inkrementelle Lernen es dem trainierten neuronalen Netzwerk 1908, sich an den neuen Datensatz 1912 anzupassen, ohne das Wissen zu vergessen, das dem trainierten neuronalen Netzwerk 1408 während des anfänglichen Trainings beigebracht wurde.

5G-NETZE

[0132] Die folgenden Figuren legen ohne Einschränkung beispielhafte 5G-Netzbasierte Systeme dar, die verwendet werden können, um mindestens eine Ausführungsform zu implementieren.

[0133] **Fig. 20** zeigt eine Architektur eines Systems 2000 eines Netzwerks gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform ist gezeigt, dass das System 2000 eine Benutzerausrüstung (UE) 2002 und eine UE 2004 beinhaltet. In mindestens einer Ausführungsform sind die UEs 2002 und 2004 als Smartphones (z. B. tragbare mobile Touchscreen-Rechenvorrichtungen, die mit einem oder mehreren Mobilfunknetzen verbindbar sind) dargestellt, können aber auch eine beliebige mobile oder nicht mobile Rechenvorrichtung umfassen, wie beispielsweise persönliche Datenassistenten (PDAs), Pager, Laptop-Computer, Desktop-Computer, drahtlose Handgeräte oder eine beliebige Rechenvorrichtung mit einer drahtlosen Kommunikationsschnittstelle.

[0134] In mindestens einer Ausführungsform kann jedes der UEs 2002 und 2004 ein Internet der Dinge (IoT)-UE umfassen, das eine Netzwerkzugriffsschicht umfassen kann, die für IoT-Anwendungen mit geringem Stromverbrauch ausgelegt ist, die kurzlebige UE-Verbindungen verwenden. In mindestens einer Ausführungsform kann ein IoT-UE Technologien wie Machine-to-Machine (M2M) oder Machine-Type Communications (MTC) zum Austauschen von Daten mit einem MTC-Server oder - Gerät über ein öffentliches mobiles Landnetzwerk (PLMN), Proximity-Based Service (ProSe) oder Device-to-Device (D2D) Kommunikation, Sensornetzwerke oder IoT-Netzwerke verwenden. In mindestens einer Ausführungsform kann ein M2M- oder MTC-Datenaustausch ein maschineninitiierter Datenaustausch sein. In mindestens einer Ausführungsform beschreibt ein IoT-Netzwerk das Verbinden von IoT-UEs, die eindeutig identifizierbare eingebettete Rechenvorrichtungen (innerhalb der Internet-Infrastruktur) mit kurzlebigen Verbindungen beinhalten können. In mindestens einer Ausführungsform können IoT-UEs Hintergrundanwendungen ausführen (z. B. Keep-Alive-Nachrichten, Statusaktualisierungen usw.), um Verbindungen eines IoT-Netzwerks zu ermöglichen.

[0135] In mindestens einer Ausführungsform können die UEs 2002 und 2004 zum Herstellen einer Verbindung mit einem Funkzugangsnetz (RAN) 2016, z. B. zum kommunikativen Koppeln konfiguriert sein. In mindestens einer Ausführungsform kann das RAN 2016 in mindestens einer Ausführungsform ein Evolved Universal Mobile Telecommunications System (UMTS) Terrestrial Radio Access Network (E-UTRAN), ein NextGen RAN (NG RAN) oder eine andere Art von RAN sein. In mindestens einer Ausführungsform verwenden die UEs 2002 und 2004 Verbindungen 2012 bzw. 2014, von denen jede eine physikalische Kommunikationsschnittstelle oder -schicht umfasst. In mindestens einer Ausführungsform sind die Verbindungen 2012 und 2014 als Luftschnittstelle dargestellt, um eine kommunikative Kopplung zu ermöglichen, und können zellularen Kommunikationsprotokollen, wie etwa einem Global System for Mobile Communications (GSM) Protokoll, einem Code Division Multiple Access (CDMA) Netzwerkprotokoll, einem Push-to-Talk (PTT) Protokoll, einem PTT over Cellular (POC) Protokoll, einem Universal Mobile Telecommunications System (UMTS) Protokoll, einem 3GPP Long Term Evolution (LTE) Protokoll, einem 5G-Protokoll, einem New-Radio (NR) Protokoll und Variationen davon entsprechen.

[0136] In mindestens einer Ausführungsform können die UEs 2002 und 2004 ferner Kommunikationsdaten direkt über eine ProSe-Schnittstelle 2006 austauschen. In mindestens einer Ausführungsform kann die ProSe-Schnittstelle 2006 alternativ als Sidelink-Schnittstelle bezeichnet werden, die einen oder mehrere logische Kanäle umfasst, darunter einen Physical Sidelink Control Channel (PSCCH), Physical Sidelink Shared Channel (PSSCH), Physical Sidelink Discovery Channel (PSDCH) und Physical Sidelink Broadcast Channel (PSBCH), ohne darauf beschränkt zu sein.

[0137] In mindestens einer Ausführungsform ist das UE 2004 so konfiguriert dargestellt, um über die Verbindung 2008 auf einen Zugangspunkt (AP) 2010 zuzugreifen. In mindestens einer Ausführungsform kann die Verbindung 2008 eine lokale drahtlose Verbindung umfassen, wie beispielsweise eine Verbindung, die einem beliebigen IEEE 802.11-Protokoll entspricht, wobei der AP 2010 einen Wireless Fidelity (WiFi®)-Router umfassen würde. In mindestens einer Ausführungsform ist gezeigt, dass der AP 2010 mit einem Internet verbunden ist, ohne sich mit einem Kernnetzwerk eines drahtlosen Systems zu verbinden.

[0138] In mindestens einer Ausführungsform kann das RAN 2016 einen oder mehrere Zugangsknoten beinhalten, die Verbindungen 2012 und 2014 ermöglichen. In mindestens einer Ausführungsform können diese Zugangsknoten (ANs) als Basisstationen (BSs), NodeBs, weiterentwickelte NodeBs (eNBs), NodeBs der nächsten Generation (gNB), RAN-Knoten usw. bezeichnet werden und können Bodenstationen (z. B. terrestrische Zugangspunkte) oder Satellitenstationen umfassen, die eine Abdeckung innerhalb eines geografischen Gebiets (z. B. einer Zelle) bieten. In mindestens einer Ausführungsform kann RAN 2016 einen oder mehrere RAN-Knoten zum Bereitstellen von Makrozellen, z. B. Makro-RAN-Knoten 2018, und einen oder mehrere RAN-Knoten zum Bereitstellen von Femtozellen oder Picozellen (z. B. Zellen mit kleineren Abdeckungsbereichen, kleinerer Benutzerkapazität oder höherer Bandbreite im Vergleich zu Makrozellen), z. B. RAN-Knoten 2020 mit geringer Leistung (LP) umfassen.

[0139] In mindestens einer Ausführungsform kann jeder der RAN-Knoten 2018 und 2020 ein Luftschnittstellenprotokoll beenden und ein erster Kontaktpunkt für UEs 2002 und 2004 sein. In mindestens einer Ausführungsform kann jeder der RAN-Knoten 2018 und 2020 verschiedene logische Funktionen für RAN 2016 erfüllen, darunter Funknetzwerk-Controller-(RNC)-Funktionen wie Funkträgerverwaltung, dynamische Uplink- und Downlink-Funkressourcenverwaltung und Datenpaketplanung und Mobilitätsmanagement, ohne darauf beschränkt zu sein.

[0140] In mindestens einer Ausführungsform können die UEs 2002 und 2004 konfiguriert sein, unter Verwendung von Orthogonal Frequency-Division Multiplexing (OFDM)-Kommunikationssignalen miteinander oder mit einem der RAN-Knoten 2018 und 2020 über einen Mehrträger-Kommunikationskanal gemäß verschiedenen Kommunikationstechniken zu kommunizieren, darunter eine Kommunikationstechnik mit orthogonalem Frequenzmultiplex (OFDMA) (z. B. für Downlink-Kommunikation) oder eine Kommunikationstechnik mit einem Trägerfrequenzmultiplex (SC-FDMA) (z. B. für Uplink und ProSe oder Sidelink-Kommunikation) und/oder Variationen davon, ohne darauf beschränkt zu sein. In mindestens einer Ausführungsform können OFDM-Signale mehrere orthogonale Unterträger umfassen.

[0141] In mindestens einer Ausführungsform kann ein Downlink-Ressourcenraster für Downlink-Übertragungen von jedem der RAN-Knoten 2018 und 2020 zu UEs 2002 und 2004 verwendet werden, während Uplink-Übertragungen ähnliche Techniken verwenden können. In mindestens einer Ausführungsform kann ein Raster ein Zeit-Frequenz-Raster sein, das als Ressourcenraster oder Zeit-Frequenz-Ressourcenraster bezeichnet wird, das eine physikalische Ressource in einem Downlink in jedem Zeitfenster ist. In mindestens einer Ausführungsform ist eine solche Zeit-Frequenz-Ebenen-Darstellung eine gängige Praxis für OFDM-Systeme, was sie für die Funkressourcenzuweisung intuitiv macht. In mindestens einer Ausführungsform entspricht jede Spalte und jede Zeile eines Ressourcenrasters einem OFDM-Symbol bzw. einem OFDM-Unterträger. In mindestens einer Ausführungsform entspricht eine Dauer eines Ressourcenrasters in einer Zeitdomäne einem Schlitz in einem Funkrahmen. In mindestens einer Ausführungsform wird eine kleinste Zeit-Frequenz-Einheit in einem Ressourcenraster als Ressourcenelement bezeichnet. In mindestens einer Ausführungsform umfasst jedes Ressourcenraster eine Reihe von Ressourcenblöcken, die eine Zuordnung bestimmter physikalischer Kanäle zu Ressourcenelementen beschreiben. In mindestens einer Ausführungsform umfasst jeder Ressourcenblock eine Sammlung von Ressourcenelementen. In mindestens einer Ausführungsform kann dies in einem Frequenzbereich eine kleinste Menge an Ressourcen darstellen, die derzeit zugewiesen werden kann. In mindestens einer Ausführungsform gibt es mehrere unterschiedliche physikalische Downlink-Kanäle, die unter Verwendung solcher Ressourcenblöcke übermittelt werden.

[0142] In mindestens einer Ausführungsform kann ein gemeinsamer physischer Downlink-Kanal (PDSCH) Benutzerdaten und Signalisierung höherer Schicht zu den UEs 2002 und 2004 übertragen. In mindestens einer Ausführungsform kann ein physikalischer Downlink-Steuerkanal (PDCCH) unter anderem Informationen über ein Transportformat und Ressourcenzuweisungen in Bezug auf den PDSCH-Kanal tragen. In mindestens einer Ausführungsform kann er die UEs 2002 und 2004 auch über ein Transportformat, eine Ressourcenzuweisung und HARQ-(Hybrid Automatic Repeat Request)-Informationen bezüglich eines gemeinsam genutzten Uplink-Kanals informieren. In mindestens einer Ausführungsform kann typischerweise Downlink-Planung (Zuweisen von Kontroll- und Shared-Channel-Ressourcenblöcken zu UE 2002 innerhalb einer Zelle) an jedem der RAN-Knoten 2018 und 2020 basierend auf Kanalqualitätsinformationen durchgeführt werden, die von einem der UEs 2002 und 2004 zurückgekoppelt werden. In mindestens einer Ausführungsform können Downlink-Ressourcenzuweisungsinformationen auf einem PDCCH gesendet werden, der für jede der UEs 2002 und 2004 verwendet (z. B. zugewiesen) wird.

[0143] In mindestens einer Ausführungsform kann ein PDCCH Steuerkanalelemente (control channel elements - CCEs) verwenden, um Steuerinformationen zu übermitteln. In mindestens einer Ausführungsform können komplexwertige PDCCH-Symbole, bevor sie auf Ressourcenelemente abgebildet werden, zuerst in Quadruplets organisiert werden, die dann unter Verwendung eines Unterblock-Interleavers zur Ratenanpassung permutiert werden können. In mindestens einer Ausführungsform kann jeder PDCCH unter Verwendung einer oder mehrerer dieser CCEs übertragen werden, wobei jedes CCE neun Sätzen von vier physischen Ressourcenelementen entsprechen kann, die als Ressourcenelementgruppen (REGs) bekannt sind. In mindestens einer Ausführungsform können jedem REG vier Quadrature-Phase-Shift-Keying-(QPSK)-Symbole zugeordnet werden. In mindestens einer Ausführungsform kann PDCCH unter Verwendung eines oder mehrerer CCEs übertragen werden, was von einer Größe einer Downlink-Steuerinformation (DCI) und einer Kanalbedingung abhängt. In mindestens einer Ausführungsform können vier oder mehr unterschiedliche PDCCH-Formate in LTE mit unterschiedlichen Anzahlen von CCEs definiert sein (z. B. Aggregationsebene, L = 1, 2, 4 oder 8).

[0144] In mindestens einer Ausführungsform kann ein erweiterter physischer Downlink-Steuerkanal (EPDCCH), der PDSCH-Ressourcen verwendet, für die Übertragung von Steuerinformationen verwendet werden. In mindestens einer Ausführungsform kann der EPDCCH unter Verwendung eines oder mehrerer erweiterter Steuerkanalelemente (ECCEs) übertragen werden. In mindestens einer Ausführungsform kann jeder ECCE neun Sätzen von vier physischen Ressourcenelementen entsprechen, die als erweiterte Ressourcenelementgruppen (EREGs) bekannt sind. In mindestens einer Ausführungsform kann ein ECCE in einigen Situationen eine andere Anzahl von EREGs aufweisen.

[0145] In mindestens einer Ausführungsform ist gezeigt, dass das RAN 2016 über eine S1-Schnittstelle 2022 mit einem Kernnetzwerk (core network - CN) 2038 kommunikativ gekoppelt ist. In mindestens einer Ausführungsform kann das CN 2038 ein EPC-Netzwerk (Evolved Packet Core), ein NPC-Netzwerk (NextGen Packet Core) oder eine andere Art von CN sein. In mindestens einer Ausführungsform ist die S1-Schnittstelle 2022 in zwei Teile aufgeteilt: S1-U-Schnittstelle 2026, die Verkehrsdaten zwischen den RAN-Knoten 2018 und 2020 und dem Serving Gateway (S-GW) 2030 überträgt, und eine S1-Mobilitätsverwaltungseinheit (MME)-Schnittstelle 2024, die eine Signalisierungsschnittstelle zwischen den RAN-Knoten 2018 und 2020 und den MMEs 2028 ist.

[0146] In mindestens einer Ausführungsform umfasst CN 2038 MMEs 2028, S-GW 2030, Packet Data Network (PDN) Gateway (P-GW) 2034 und einen Home Subscriber Server (HSS) 2032. In mindestens einer Ausführungsform können die MMEs 2028 in ihrer Funktion einer Steuerebene von Legacy Serving General Packet Radio Service (GPRS) Support Nodes (SGSN) ähnlich sein. In mindestens einer Ausführungsform können die MMEs 2028 Mobilitätsaspekte beim Zugriff verwalten, wie z. B. die Gateway-Auswahl und die Verwaltung von Verfolgungsbereichslisten. In mindestens einer Ausführungsform kann der HSS 2032 eine Datenbank für Netzwerkbenutzer umfassen, einschließlich abonnementbezogener Informationen, um die Handhabung von Kommunikationssitzungen durch eine Netzwerkeinheit zu unterstützen. In mindestens einer Ausführungsform kann das CN 2038 einen oder mehrere HSS 2032 umfassen, abhängig von einer Anzahl von Mobilfunkteilnehmern, von einer Kapazität einer Ausrüstung, von einer Organisation eines Netzwerks usw. In mindestens einer Ausführungsform kann der HSS 2032 Unterstützung für Routing/Roaming, Authentifizierung, Autorisierung, Namens-/Adressierungsauflösung, Standortabhängigkeiten usw. bereitstellen.

[0147] In mindestens einer Ausführungsform kann das S-GW 2030 eine S1-Schnittstelle 2022 in Richtung des RAN 2016 beenden und Datenpakete zwischen RAN 2016 und CN 2038 routen. In mindestens einer

Ausführungsform kann das S-GW 2030 ein lokaler Mobilitätsankerpunkt für Inter-RAN-Knotenübergaben sein und kann auch einen Anker für Inter-3GPP-Mobilität bereitstellen. In mindestens einer Ausführungsform können andere Verantwortlichkeiten legales Abfangen, Gebühren und eine Durchsetzung gewisser Richtlinien beinhalten.

[0148] In mindestens einer Ausführungsform kann das P-GW 2034 eine SGi-Schnittstelle in Richtung eines PDN beenden. In mindestens einer Ausführungsform kann P-GW 2034 Datenpakete zwischen einem EPC-Netzwerk 2038 und externen Netzwerken, wie einem Netzwerk, das den Anwendungsserver 2040 (alternativ als Anwendungsfunktion (AF) bezeichnet) umfasst, über eine Internetprotokoll-(IP-)Schnittstelle 2042 routen. In mindestens einer Ausführungsform kann der Anwendungsserver 2040 ein Element sein, das Anwendungen anbietet, die IP-Trägerressourcen mit einem Kernnetzwerk verwenden (z. B. UMTS-Paketdienste-(PS-) Domäne, LTE-PS-Datendienste usw.). In mindestens einer Ausführungsform ist gezeigt, dass das P-GW 2034 über eine IP-Kommunikationsschnittstelle 2042 mit einem Anwendungsserver 2040 kommunikativ gekoppelt ist. In mindestens einer Ausführungsform kann der Anwendungsserver 2040 auch so konfiguriert sein, dass er einen oder mehrere Kommunikationsdienste (z. B. Voice-over-Internet Protocol (VoIP)-Sitzungen, PTT-Sitzungen, Gruppenkommunikationssitzungen, soziale Netzwerkdienste usw.) für UEs 2002 und 2004 über CN 2038 unterstützt.

[0149] In mindestens einer Ausführungsform kann das P-GW 2034 ferner ein Knoten zur Richtliniendurchsetzung und zur Erhebung von Gebührendaten sein. In mindestens einer Ausführungsform ist die Richtlinien- und Gebührendurchsetzungsfunktion (policy and charging enforcement function - PCRF) 2036 ein Richtlinien- und Gebührensteuerelement des CN 2038. In mindestens einer Ausführungsform kann es in einem Nicht-Roaming-Szenario ein einzelnes PCRF in einem Home Public Land Mobile Network (HPLMN) geben, das einer Sitzung des Internet Protocol Connectivity Access Network (IP-CAN) einer UE zugeordnet ist. In mindestens einer Ausführungsform kann es in einem Roaming-Szenario mit lokalem Breakout des Verkehrs zwei PCRFs geben, die einer IP-CAN-Sitzung einer UE zugeordnet sind: eine Heim-PCRF (H-PCRF) innerhalb eines HPLMN und eine besuchte PCRF (V-PCRF) innerhalb eines Visited Public Land Mobile Network (VPLMN). In mindestens einer Ausführungsform kann die PCRF 2036 über das P-GW 2034 kommunikativ an einen Anwendungsserver 2040 gekoppelt sein. In mindestens einer Ausführungsform kann der Anwendungsserver 2040 der PCRF 2036 signalisieren, einen neuen Dienstfluss anzugeben und eine geeignete Dienstgüte (Quality of Service - QoS) und Gebührenparameter auszuwählen. In mindestens einer Ausführungsform kann die PCRF 2036 diese Regel in einer Richtlinien- und Gebührendurchsetzungsfunktion (Policy and Charging Enforcement Function - PCEF) (nicht gezeigt) mit einer geeigneten Datenverkehrsschablone (Traffic Flow Template - TFT) und QoS-Kennungsklasse (QoS dass of identifier - QCI) bereitstellen, die eine QoS und eine Abrechnung, wie durch Anwendungsserver 2040 festgelegt, beginnt.

[0150] **Fig. 21** zeigt eine Architektur eines Systems 2100 eines Netzwerks gemäß einigen Ausführungsformen. In mindestens einer Ausführungsform ist gezeigt, dass das System 2100 eine UE 2102, einen 5G-Zugangsknoten oder RAN-Knoten (gezeigt als (R)AN-Knoten 2108), eine Benutzerebenenfunktion (gezeigt als UPF 2104), ein Datennetzwerk (DN 2106), das in mindestens einer Ausführungsform Betreiberdienste, Internetzugang oder Dienste von Drittanbietern sein können, und ein 5G-Kernnetzwerk (5GC) (als CN 2110 gezeigt) beinhaltet.

[0151] In mindestens einer Ausführungsform beinhaltet das CN 2110 eine Authentifizierungsserverfunktion (Authentification Server Function - AUSF 2114); eine Kernzugangs- und Mobilitätsverwaltungsfunktion (Access and Mobilty Function - AMF 2112); eine Sitzungsverwaltungsfunktion (Session Management Function - SMF 2118); eine Netzwerkoffenlegungsfunktion (Network Exposure Function - NEF 2116); eine Richtliniensteuerfunktion (Policy Control Function - PCF 2122); eine Netzwerkfunktions (NF)-Repository-Funktion (NRF 2120); eine vereinheitlichte Datenverwaltung (Unified Data Management - UDM 2124); und eine Anwendungsfunktion (AF 2126). In mindestens einer Ausführungsform kann das CN 2110 auch andere Elemente beinhalten, die nicht gezeigt sind, wie etwa eine strukturierte Datenspeicherungsnetzwerkfunktion (Structured Data Storage network function - SDSF), eine unstrukturierte Datenspeicherungsnetzwerkfunktion (Unstructured Data Storage network function - UDSF) und Variationen davon.

[0152] In mindestens einer Ausführungsform kann der UPF 2104 als Ankerpunkt für Intra-RAT- und Inter-RAT-Mobilität, als externer PDU-Sitzungspunkt der Verbindung mit DN 2106 und als Verzweigungspunkt zum Unterstützen einer mehrfach vernetzten PDU-Sitzung dienen. In mindestens einer Ausführungsform kann die UPF 2104 auch Paket-Routing und -Weiterleitung, Paket-Inspektion durchführen, einen Teil von Richtlinienregeln auf Benutzerebene durchsetzen, Pakete legal abfangen (UP-Sammlung); Datenverkehrsnutzungsberichte, QoS-Handling für die Benutzerebene durchführen (z. B. Paketfilterung, Gating, UL/DL-

Ratendurchsetzung), Uplink-Datenverkehrsverifizierung (z. B. SDF-zu-QoS-Flow-Mapping), Paketmarkierung auf Transportebene in Uplink und Downlink und Downlink-Paketpufferung und Downlink-Datenbenachrichtigungsauslösung durchführen. In mindestens einer Ausführungsform kann die UPF 2104 einen Uplink-Klassifizierer beinhalten, um das Routing von Datenverkehrsflüssen zu einem Datennetzwerk zu unterstützen. In mindestens einer Ausführungsform kann DN 2106 verschiedene Netzbetreiberdienste, Internetzugang oder Dienste Dritter darstellen.

[0153] In mindestens einer Ausführungsform kann die AUSF 2114 Daten zur Authentifizierung der UE 2102 speichern und authentifizierungsbezogene Funktionalität handhaben. In mindestens einer Ausführungsform kann die AUSF 2114 ein gemeinsames Authentifizierungsframework für verschiedene Zugriffsarten ermöglichen.

[0154] In mindestens einer Ausführungsform kann die AMF 2112 für die Registrierungsverwaltung (z. B. für die Registrierung der UE 2102 usw.), die Verbindungsverwaltung, die Erreichbarkeitsverwaltung, die Mobilitätsverwaltung und das legale Abfangen von AMF-bezogenen Ereignissen und die Zugriffsauthentifizierung und -autorisierung verantwortlich sein. In mindestens einer Ausführungsform kann die AMF 2112 den Transport von SM-Nachrichten für die SMF 2118 bereitstellen und als transparenter Proxy zum Routen von SM-Nachrichten fungieren. In mindestens einer Ausführungsform kann die AMF 2112 auch den Transport von Kurznachrichtendienst (SMS)-Nachrichten zwischen dem UE 2102 und einer SMS-Funktion (SMSF) (in **Fig. 21** nicht gezeigt) bereitstellen. In mindestens einer Ausführungsform kann die AMF 2112 als Sicherheitsankerfunktion (Security Anchor Function - SEA) fungieren, die eine Interaktion mit der AUSF 2114 und der UE 2102 und den Empfang eines Zwischenschlüssels beinhalten kann, der als Ergebnis des UE 2102-Authentifizierungsprozesses erstellt wurde. In mindestens einer Ausführungsform, bei der eine USIM-basierte Authentifizierung verwendet wird, kann die AMF 2112 Sicherheitsmaterial von der AUSF 2114 abrufen. In mindestens einer Ausführungsform kann die AMF 2112 auch eine Sicherheitskontextverwaltungsfunktion (security context management - SCM) umfassen, die einen Schlüssel von SEA empfängt, mit dem sie zugangsnetzspezifische Schlüssel ableitet. In mindestens einer Ausführungsform kann die AMF 2112 außerdem ein Abschlusspunkt der RAN-CP-Schnittstelle (N2-Referenzpunkt), ein Abschlusspunkt der NAS-(NI)-Signalisierung sein und NAS-Verschlüsselung und Integritätsschutz durchführen.

[0155] In mindestens einer Ausführungsform kann die AMF 2112 auch NAS-Signalisierung mit einer UE 2102 über eine N3-Interworking-Function-(IWF)-Schnittstelle unterstützen. In mindestens einer Ausführungsform kann die N3IWF verwendet werden, um Zugriff auf nicht vertrauenswürdige Einheiten bereitzustellen. In mindestens einer Ausführungsform kann die N3IWFein Abschlusspunkt für N2- und N3-Schnittstellen für die Steuerebene bzw. die Benutzerebene sein und kann somit N2-Signalisierung von der SMF und der AMF für PDU-Sitzungen und QoS handhaben, Datenpakete für IPSec- und N3-Tunneling einkapseln/entkapseln, N3-Benutzerebenen-Pakete im Uplink markieren und QoS entsprechend der N3-Paketmarkierung unter Berücksichtigung der QoS-Anforderungen durchsetzen, die einer solchen über N2 empfangenen Markierung zugeordnet sind. In mindestens einer Ausführungsform kann die N3IWFauch Uplink- und Downlink-Steuerebenen-NAS(NI)-Signalisierung zwischen der UE 2102 und der AMF 2112 weiterleiten und Uplink- und Downlink-Benutzerebenenpakete zwischen der UE 2102 und der UPF 2104 weiterleiten. In mindestens einer Ausführungsform stellt die N3IWFauch Mechanismen für den IPsec-Tunnelaufbau mit der UE 2102 bereit.

[0156] In mindestens einer Ausführungsform kann die SMF 2118 für das Sitzungsmanagement verantwortlich sein (z. B. Sitzungsaufbau, Modifizierung und Freigabe, einschließlich Tunnelwartung zwischen UPF und AN-Knoten); UE-IP-Adresszuweisung und -verwaltung (einschließlich optionaler Autorisierung); Auswahl und Steuerung der UP-Funktion; Konfigurieren der Verkehrssteuerung bei UPF, um den Verkehr an das richtige Ziel zu leiten; Beendigung von Schnittstellen zu Richtlinienkontrollfunktionen; Steuern eines Teil der Richtliniendurchsetzung und QoS; rechtmäßiges Abfangen (für SM-Ereignisse und Schnittstelle zum LI-System); Beendigung von SM-Teilen von NAS-Nachrichten; Downlink-Datenbenachrichtigung; Initiator von AN-spezifischen SM-Informationen, die über AMF über N2 an AN gesendet werden; Bestimmen des SSC-Modus einer Sitzung. In mindestens einer Ausführungsform kann die SMF 2118 die folgende Roaming-Funktionalität beinhalten: Handhaben der lokalen Durchsetzung, um QoS SLAB (VPLMN) anzuwenden; Gebührendatensammlung und Gebührenschnittstelle (VPLMN); legales Abfangen (im VPLMN für SM-Ereignisse und Schnittstelle zum LI-System); Unterstützung für die Interaktion mit dem externen DN für den Transport von Signalisierung für die PDU-Sitzungsautorisierung/-authentifizierung durch das externe DN.

[0157] In mindestens einer Ausführungsform kann die NEF 2116 Mittel zum sicheren Offenlegen von Diensten und Fähigkeiten bereitstellen, die durch 3GPP-Netzwerkfunktionen für Dritte bereitgestellt sind, interne Offenlegung/erneute Offenlegung, Anwendungsfunktionen (z. B. AF 2126), Edge-Computing- oder Fog-Com-

puting-Systeme usw. In mindestens einer Ausführungsform kann die NEF 2116 AF authentifizieren, autorisieren und/oder drosseln. In mindestens einer Ausführungsform kann die NEF 2116 auch mit der AF 2126 ausgetauschte Informationen und mit internen Netzwerkfunktionen ausgetauschte Informationen übersetzen. In mindestens einer Ausführungsform kann die NEF 2116 zwischen einem AF-Service-Identifier und einer internen 5GC-Information übersetzen. In mindestens einer Ausführungsform kann die NEF 2116 auch Informationen von anderen Netzwerkfunktionen (NF) basierend auf offengelegten Fähigkeiten anderer Netzwerkfunktionen empfangen. In mindestens einer Ausführungsform können diese Informationen als strukturierte Daten in der NEF 2116 oder in einer Datenspeicher-NF unter Verwendung standardisierter Schnittstellen gespeichert werden. In mindestens einer Ausführungsform können gespeicherte Informationen dann durch die NEF 2116 andere NFs und AFs erneut exponiert und/oder für andere Zwecke wie etwa Analysen verwendet werden.

[0158] In mindestens einer Ausführungsform kann die NRF 2120 Diensterkennungsfunktionen unterstützen, NF-Erkennungsanforderungen von NF-Instanzen empfangen und Informationen über erkannte NF-Instanzen an NF-Instanzen bereitstellen. In mindestens einer Ausführungsform pflegt die NRF 2120 auch Informationen über verfügbare NF-Instanzen und deren unterstützte Dienste.

[0159] In mindestens einer Ausführungsform kann die PCF 2122 Richtlinienregeln für die Funktion(en) auf Steuerebene bereitstellen, um diese durchzusetzen, und kann auch ein einheitliches Richtlinienframework unterstützen, um das Netzwerkverhalten zu regulieren. In mindestens einer Ausführungsform kann die PCF 2122 auch ein Frontend (FE) implementieren, um auf Abonnementinformationen zuzugreifen, die für Richtlinienentscheidungen in einem UDR der UDM 2124 relevant sind.

[0160] In mindestens einer Ausführungsform kann die UDM 2124 abonnementbezogene Informationen verarbeiten, um die Handhabung von Kommunikationssitzungen von Netzwerkeinheiten zu unterstützen, und kann Abonnementdaten der UE 2102 speichern. In mindestens einer Ausführungsform kann die UDM 2124 zwei Teile beinhalten, ein Anwendungs-FE und ein Benutzerdaten-Repository (User Data Repository - UDR). In mindestens einer Ausführungsform kann die UDM ein UDM FE umfassen, das für die Verarbeitung von Berechtigungsnachweisen, Standortverwaltung, Abonnementverwaltung usw. verantwortlich ist. In mindestens einer Ausführungsform können mehrere unterschiedliche Frontends einen gleichen Benutzer in unterschiedlichen Transaktionen bedienen. In mindestens einer Ausführungsform greift das UDM-FE auf Abonnementinformationen zu, die in einem UDR gespeichert sind, und führt Authentifizierungszugangsdatenverarbeitung; Handhabung der Benutzeridentifikation; Zugriffsberechtigung; Registrierung/Mobilitätsverwaltung; und Abonnementverwaltung durch. In mindestens einer Ausführungsform kann das UDR mit der PCF 2122 interagieren. In mindestens einer Ausführungsform kann die UDM 2124 auch SMS-Verwaltung unterstützen, wobei ein SMS-FE eine ähnliche Anwendungslogik wie zuvor erörtert implementiert.

[0161] In mindestens einer Ausführungsform kann die AF 2126 Anwendungseinfluss auf Datenverkehrsrouting, Zugriff auf eine Netzwerkleistungsfähigkeitsoffenlegung (Network Capability Exposure - NCE) bereitstellen und mit einem Richtlinienframework zur Richtliniensteuerung interagieren. In mindestens einer Ausführungsform kann die NCE ein Mechanismus sein, der es einem 5GC und einer AF 2126 erlaubt, einander über NEF 2116 Informationen bereitzustellen, die für Edge-Computing-Implementierungen verwendet werden können. In mindestens einer Ausführungsform können Dienste des Netzbetreibers und Dritter in der Nähe des Zugangspunkts der UE 2102 gehostet werden, um eine effiziente Dienstbereitstellung durch eine reduzierte Ende-zu-Ende-Latenz und Last auf einem Transportnetzwerk zu erreichen. In mindestens einer Ausführungsform kann das 5GC für Edge-Computing-Implementierungen eine UPF 2104 in der Nähe der UE 2102 auswählen und die Datenverkehrssteuerung von der UPF 2104 zu dem DN 2106 über die N6-Schnittstelle ausführen. In mindestens einer Ausführungsform kann dies auf UE-Abonnementdaten, dem UE-Standort und durch die AF 2126 bereitgestellten Informationen basieren. In mindestens einer Ausführungsform kann die AF 2126 die UPF-(Neu-)Auswahl und das Datenverkehrsrouting beeinflussen. In mindestens einer Ausführungsform kann ein Netzbetreiber basierend auf Bereitstellung durch den Betreiber der AF 2126 erlauben, direkt mit relevanten NFs zu interagieren, wenn die AF 2126 als eine vertrauenswürdige Einheit betrachtet wird.

[0162] In mindestens einer Ausführungsform kann das CN 2110 eine SMSF beinhalten, die für die SMS-Abonnementprüfung und -verifizierung und das Weiterleiten von SM-Nachrichten an/von der UE 2102 an/von anderen Einheiten, wie etwa einem SMS-GMSC/IWMSC/SMS-Router verantwortlich sein kann. In mindestens einer Ausführungsform kann SMS auch mit AMF 2112 und UDM 2124 für eine Benachrichti-

gungsprozedur interagieren, dass die UE 2102 für die SMS- Übertragung verfügbar ist (z. B. ein UE-Nichterreichbar-Flag setzen und die UDM 2124 benachrichtigen, wenn die UE 2102 für SMS verfügbar ist).

[0163] In mindestens einer Ausführungsform kann das System 2100 die folgenden dienstbasierten Schnittstellen beinhalten: Namf: durch die AMF gezeigte dienstbasierte Schnittstelle; Nsmf: durch die SMF gezeigte dienstbasierte Schnittstelle; Nnef: durch die NEF gezeigte dienstbasierte Schnittstelle; Npcf: durch die PCF gezeigte dienstbasierte Schnittstelle; Nudm: durch die UDM gezeigte dienstbasierte Schnittstelle; Naf: durch die AF gezeigte dienstbasierte Schnittstelle; Nnrf: durch die NRF gezeigte dienstbasierte Schnittstelle; und Nausf: durch die AUSF gezeigte dienstbasierte Schnittstelle.

[0164] In mindestens einer Ausführungsform kann das System 2100 die folgenden dienstbasierten Schnittstellen beinhalten: N1: Referenzpunkt zwischen UE und AMF; N2: Referenzpunkt zwischen (R)AN und AMF; N3: Referenzpunkt zwischen (R)AN und UPF; N4: Referenzpunkt zwischen SMF und UPF; und N6: Referenzpunkt zwischen UPF und einem Datennetzwerk. In mindestens einer Ausführungsform kann es viel mehr Referenzpunkte und/oder dienstbasierte Schnittstellen zwischen NF-Diensten in NF geben, jedoch wurden diese Schnittstellen und Referenzpunkte aus Gründen der Klarheit weggelassen. In mindestens einer Ausführungsform kann ein NS-Referenzpunkt zwischen der PCF und der AF liegen; kann ein N7-Referenzpunkt zwischen der PCF und der SMF liegen; kann ein N11-Referenzpunkt zwischen der AMF und derSMF; usw. In mindestens einer Ausführungsform kann das CN 2110 eine Nx-Schnittstelle beinhalten, die eine Inter-CN-Schnittstelle zwischen MME und AMF 2112 ist, um eine Zusammenarbeit zwischen dem CN 2110 und dem CN 7221 zu ermöglichen.

[0165] In mindestens einer Ausführungsform kann das System 2100 mehrere RAN-Knoten (wie etwa (R)AN-Knoten 2108) beinhalten, wobei eine Xn-Schnittstelle zwischen zwei oder mehr (R)AN-Knoten 2108 (z. B. gNB), die mit dem 5GC 410 verbunden sind, zwischen einem (R)AN-Knoten 2108 (z. B. gNB), der mit dem CN 2110 verbunden ist, und einem eNB (z. B. einem Makro-RAN-Knoten) und/oder zwischen zwei eNB, die mit dem CN 2110 verbunden sind, definiert ist.

[0166] In mindestens einer Ausführungsform kann die Xn-Schnittstelle eine Xn-Benutzerebenen-(Xn-U)-Schnittstelle und eine Xn-Steuerungsebenen-(Xn-C)-Schnittstelle umfassen. In mindestens einer Ausführungsform kann Xn-U eine nicht garantierte Zustellung von Benutzerebenen-PDUs bereitstellen und eine Datenweiterleitungs- und Durchflussreglerfunktionalität unterstützen/bereitstellen. In mindestens einer Ausführungsform kann Xn-C eine Verwaltungs- und Fehlerbehandlungsfunktionalität, eine Funktionalität zum Verwalten einer Xn-C-Schnittstelle; Mobilitätsunterstützung für das UE 2102 in einem verbundenen Modus (z. B. CM-CONNECTED) einschließlich der Funktionalität zum Verwalten der UE-Mobilität für den verbundenen Modus zwischen einem oder mehreren (R)AN-Knoten 2108 bereitstellen. In mindestens einer Ausführungsform kann die Mobilitätsunterstützung einen Kontexttransfer von einem alten bedienenden (R)AN-Knoten 2108 (Quelle) zu einem neuen bedienenden (R)AN-Knoten 2108 (Ziel); und Steuerung von Benutzerebenen-Tunneln zwischen dem alten bedienenden (R)AN-Knoten 2108 (Quelle) zu dem neuen bedienenden (R)AN-Knoten 2108 (Ziel) beinhalten.

[0167] In mindestens einer Ausführungsform kann ein Protokollstapel eines Xn-U eine Transportnetzwerkschicht umfassen, die auf einer Internetprotokoll(IP)-Transportschicht aufgebaut ist, und eine GTP-U-Schicht auf einer UDP- und/oder IP-Schicht(en) zu PDUs auf Benutzerebene tragen. In mindestens einer Ausführungsform kann der Xn-C-Protokollstapel ein Signalisierungsprotokoll der Anwendungsschicht (als Xn-Anwendungsprotokoll (Xn-AP) bezeichnet) und eine Transportnetzwerkschicht, die auf einer SCTP-Schicht aufgebaut ist, beinhalten. In mindestens einer Ausführungsform kann sich die SCTP-Schicht auf einer IP-Schicht befinden. In mindestens einer Ausführungsform stellt die SCTP-Schicht eine garantierte Zustellung von Nachrichten der Anwendungsschicht bereit. In mindestens einer Ausführungsform wird in einer Transport-IP-Schicht eine Punkt-zu-Punkt-Übertragung verwendet, um Signalisierungs-PDUs zu liefern. In mindestens einer Ausführungsform können der Xn-U-Protokollstapel und/oder ein Xn-C-Protokollstapel gleich oder ähnlich einem Benutzerebenen- und/oder Steuerebenen-Protokollstapel sein, die hierin gezeigt und beschrieben sind.

[0168] **Fig. 22** ist eine Veranschaulichung eines Protokollstapels auf Steuerebene gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform wird eine Steuerebene 2200 als Kommunikationsprotokollstapel zwischen der UE 2002 (oder alternativ der UE 2004), dem RAN 2016 und den MME 2028 gezeigt.

[0169] In mindestens einer Ausführungsform kann die PHY-Schicht 2202 Informationen, die durch die MAC-Schicht 2204 verwendet werden, über eine oder mehrere Luftschnittstellen senden oder empfangen. In mindestens einer Ausführungsform kann die PHY-Schicht 2202 ferner eine Verbindungsanpassung oder adaptive Modulation und Codierung (AMC), Leistungssteuerung, Zellensuche (z. B. für anfängliche Synchronisations- und Übergabezwecke) und andere Messungen durchführen, die durch höhere Schichten verwendet werden, wie etwa eine RRC-Schicht 2210. In mindestens einer Ausführungsform kann die PHY-Schicht 2202 weiterhin eine Fehlererkennung auf Transportkanälen, eine Vorwärtsfehlerkorrektur-(FEC)-Codierung/-Decodierung von Transportkanälen, eine Modulation/Demodulation von physikalischen Kanälen, eine Verschachtelung, eine Ratenanpassung, eine Abbildung auf physikalische Kanäle und Multiple Input Multiple Output (MIMO) Antennenverarbeitung durchführen.

[0170] In mindestens einer Ausführungsform kann die MAC-Schicht 2204 eine Zuordnung zwischen logischen Kanälen und Transportkanälen, ein Multiplexen von MAC-Dienstdateneinheiten (service data unit - SDU) von einem oder mehreren logischen Kanälen auf Transportblöcke (TB), die über Transportkanäle an PHY zugestellt werden sollen, De-Multiplexen von MAC-SDU zu einem oder mehreren logischen Kanälen von Transportblöcken (TB), die von PHY über Transportkanäle zugestellt werden, Multiplexen von MAC-SDU auf TB, Planen von Informationsberichten, Fehlerkorrektur durch hybride automatische Wiederholungsanforderung (hybrid automatic repeat request - HARD) und logische Kanalpriorisierung durchführen.

[0171] In mindestens einer Ausführungsform kann die RLC-Schicht 2206 in einer Vielzahl von Betriebsmodi arbeiten, einschließlich: Transparenter Modus (TM), Unbestätigter Modus (UM) und Bestätigter Modus (Acknowledged Mode - AM). In mindestens einer Ausführungsform kann die RLC-Schicht 2206 einen Transfer von Protokolldateneinheiten (PDU) der oberen Schicht, eine Fehlerkorrektur durch automatische Wiederholungsanforderung (automatic repeat request - ARQ) für AM-Datenübertragungen und eine Verkettung, Segmentierung und Wiederzusammensetzung von RLC-SDU für UM- und AM-Datentransfers ausführen. In mindestens einer Ausführungsform kann die RLC-Schicht 2206 auch eine Neusegmentierung von RLC-Daten-PDUs für AM-Datenübertragungen ausführen, RLC-Daten-PDUs für UM- und AM-Datenübertragungen neu anordnen, doppelte Daten für UM- und AM-Datenübertragungen erkennen, RLC-SDUs für UM und AM-Datenübertragungen verwerfen, Protokollfehler für AM-Datenübertragungen erkennen und RLC-Wiederherstellung durchführen.

[0172] In mindestens einer Ausführungsform kann die PDCP-Schicht 2208 eine Header-Komprimierung und -Dekomprimierung von IP-Daten ausführen, PDCP-Sequenznummern (SN) pflegen, eine sequentielle Zustellung von PDU der oberen Schicht bei der Wiederherstellung der unteren Schichten durchführen, Duplikate von SDU der unteren Schicht bei der Wiederherstellung niedrigerer Schichten für RLC AM zugeordnete Funkträger beseitigen, Steuerebenendaten verschlüsseln und entschlüsseln, Integritätsschutz und Integritätsverifizierung von Steuerebenendaten durchführen, zeitgeberbasiertes Verwerfen von Daten steuern und Sicherheitsoperationen (z. B. Verschlüsselung, Entschlüsselung, Integritätsschutz, Integritätsverifizierung usw.) durchführen.

[0173] In mindestens einer Ausführungsform können die Hauptdienste und - funktionen einer RRC-Schicht 2210 das Senden von Systeminformationen (z. B. in Master-Informationsblöcken (MIB) oder Systeminformationsblöcken (SIB) in Bezug auf eine Nicht-Zugriffsschicht (non-access stratum - NAS)); Senden von Systeminformationen in Bezug auf eine Zugangsschicht (access stratum - AS), Paging, Aufbau, Pflege und Freigabe einer RRC-Verbindung zwischen einer UE und E-UTRAN (z. B. RRC-Verbindungs-Paging, RRC-Verbindungsaufbau, RRC-Verbindungsmodifikation und RRC-Verbindungsfreigabe), Einrichtung, Konfiguration, Pflege und Freigabe von Punkt-zu-Punkt-Funkträgern, Sicherheitsfunktionen einschließlich Schlüsselverwaltung, Mobilität zwischen Funkzugangstechnologie (RAT) und Messkonfiguration für UE-Messberichte beinhalten. In mindestens einer Ausführungsform können die MIBs und SIBs ein oder mehrere Informationselemente (IEs) umfassen, die jeweils einzelne Datenfelder oder Datenstrukturen umfassen können.

[0174] In mindestens einer Ausführungsform können die UE 2002 und das RAN 2016 eine Uu-Schnittstelle (z. B. eine LTE-Uu-Schnittstelle) nutzen, um Steuerebenendaten über einen Protokollstapel auszutauschen, der die PHY-Schicht 2202, die MAC-Schicht 2204, die RLC-Schicht 2206, die PDCP-Schicht 2208 und die RRC-Schicht 2210 umfasst.

[0175] In mindestens einer Ausführungsform bilden Non-Access-Stratum-(NAS-)Protokolle (NAS-Protokolle 2212) eine höchste Schicht einer Steuerungsebene zwischen UE 2002 und MME(s) 2028. In mindestens einer Ausführungsform unterstützen die NAS-Protokolle 2212 die Mobilität der UE 2002 und Sitzungsverwal-

tungsprozeduren, um eine IP-Konnektivität zwischen der UE 2002 und dem P-GW 2034 einzurichten und aufrechtzuerhalten.

[0176] In mindestens einer Ausführungsform kann die Si-Anwendungsprotokoll-(S1-AP)-Schicht (Si-AP-Schicht 2222) Funktionen einer Si-Schnittstelle unterstützen und elementare Prozeduren (EPs) umfassen. In mindestens einer Ausführungsform ist eine EP eine Interaktionseinheit zwischen dem RAN 2016 und dem CN 2028. In mindestens einer Ausführungsform können S1-AP-Schichtdienste zwei Gruppen umfassen: UE-assoziierte Dienste und nicht UE-assoziierte Dienste. In mindestens einer Ausführungsform führen diese Dienste Funktionen durch, darunter, aber nicht beschränkt auf: E-UTRAN Radio Access Bearer (E-RAB)-Verwaltung, UE-Fähigkeitsanzeige, Mobilität, NAS-Signalisierungstransport, RAN-Informationsverwaltung (RIM) und Konfigurationsübertragung.

[0177] In mindestens einer Ausführungsform kann die Stream Control Transmission Protocol (SCTP)-Schicht (alternativ als Stream Control Transmission Protocol/Internet Protocol (SCTP/IP)-Schicht bezeichnet) (SCTP-Schicht 2220) eine zuverlässige Zustellung von Signalisierungsnachrichten zwischen RAN 2016 und MME(s) 2028, teilweise basierend auf einem IP-Protokoll, das von einer IP-Schicht 2218 unterstützt wird, sicherstellen. In mindestens einer Ausführungsform können sich die L2-Schicht 2216 und eine L1-Schicht 2214 auf Kommunikationsverbindungen (z. B. drahtgebunden oder drahtlos) beziehen, die durch einen RAN-Knoten und eine MME verwendet werden, um Informationen auszutauschen.

[0178] In mindestens einer Ausführungsform können das RAN 2016 und die MME 2028 eine S1-MME-Schnittstelle nutzen, um Steuerebenendaten über einen Protokollstapel auszutauschen, der eine L1-Schicht 2214, eine L2-Schicht 2216, eine IP-Schicht 2218, eine SCTP-Schicht 2220 und eine Si-AP-Schicht 2222 umfasst.

[0179] **Fig. 23** ist eine Darstellung eines Protokollstapels auf Benutzerebene gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform wird eine Benutzerebene 2300 als Kommunikationsprotokollstapel zwischen einer UE 2002, dem RAN 2016, dem S-GW 2030 und dem P-GW 2034 gezeigt. In mindestens einer Ausführungsform kann die Benutzerebene 2300 dieselben Protokollschichten wie die Steuerebene 2200 nutzen. In mindestens einer Ausführungsform können die UE 2002 und das RAN 2016 eine Uu-Schnittstelle (z. B. eine LTE-Uu-Schnittstelle) nutzen, um Benutzerebenendaten über einen Protokollstapel auszutauschen, der die PHY-Schicht 2202, die MAC-Schicht 2204, die RLC-Schicht 2206 und die PDCP-Schicht 2208 umfasst.

[0180] In mindestens einer Ausführungsform kann das Tunneling-Protokoll des General Packet Radio Service (GPRS) für eine Schicht der Benutzerebene (GTP-U) (GTP-U-Schicht 2302) verwendet werden, um Benutzerdaten innerhalb eines GPRS-Kernnetzwerks und zwischen einem Funkzugangsnetzwerk und einem Kernnetzwerk zu übertragen. In mindestens einer Ausführungsform können die transportierten Benutzerdaten Pakete in einem beliebigen der Formate IPv4, IPv6 oder PPP sein. In mindestens einer Ausführungsform kann die UDP- und IP-Sicherheits-(UDP/IP-)Schicht (UDP/IP-Schicht 2302) Prüfsummen für die Datenintegrität, Portnummern zum Adressieren unterschiedlicher Funktionen an einer Quelle und einem Ziel und Verschlüsselung und Authentifizierung für ausgewählte Datenflüsse bereitstellen. In mindestens einer Ausführungsform können das RAN 2016 und das S-GW 2030 eine S1-U-Schnittstelle nutzen, um Benutzerebenendaten über einen Protokollstapel auszutauschen, der die L1-Schicht 2214, die L2-Schicht 2216, die UDP/IP-Schicht 2302 und die GTP-U-Schicht 2302 umfasst. In mindestens einer Ausführungsform können das S-GW 2030 und das P-GW2034 eine S5/S8a-Schnittstelle nutzen, um Benutzerebenendaten über einen Protokollstapel auszutauschen, der die L1-Schicht 2214, die L2-Schicht 2216, die UDP/IP-Schicht 2302 und die GTP-U-Schicht 2302 umfasst. In mindestens einer Ausführungsform, wie vorstehend unter Bezugnahme auf die **Fig. 22** erörtert, unterstützen die NAS-Protokolle eine Mobilität der UE 2002 und Sitzungsverwaltungsprozeduren, um eine IP-Konnektivität zwischen der UE 2002 und dem P-GW 2034 einzurichten und aufrechtzuerhalten.

[0181] **Fig. 24** zeigt Komponenten 2400 eines Kernnetzwerks gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform können die Komponenten des CN 2038 in einem physikalischen Knoten oder separaten physikalischen Knoten implementiert sein, darunter Komponenten zum Lesen und Ausführen von Anweisungen von einem maschinenlesbaren oder computerlesbaren Medium (z. B. einem nichttransitorischen maschinenlesbaren Speichermedium). In mindestens einer Ausführungsform wird eine Netzwerkfunktionsvirtualisierung (Network Functions Virtualization - NFV) genutzt, um eine oder alle der vorstehend beschriebenen Netzwerkknotenfunktionen über ausführbare Anweisungen, die in einem oder mehreren computerlesbaren Speichermedien gespeichert sind (nachstehend ausführlicher beschrieben), zu virtualisieren.

In mindestens einer Ausführungsform kann eine logische Instanziierung von CN 2038 als Netzwerk-Slice 2402 bezeichnet werden (z. B. ist dargestellt, dass Netzwerk-Slice 2402 HSS 2032, MME(s) 2028 und S-GW2030 beinhaltet). In mindestens einer Ausführungsform kann eine logische Instanziierung eines Teils von CN 2038 als Netzwerk-Subslice 2404 bezeichnet werden (z. B. ist dargestellt, dass Netzwerk-Subslice in der Darstellung 2404 P-GW 2034 und PCRF 2036 beinhaltet).

[0182] In mindestens einer Ausführungsform können NFV-Architekturen und - Infrastrukturen verwendet werden, um eine oder mehrere Netzwerkfunktionen, die alternativ durch proprietäre Hardware durchgeführt werden, auf physischen Ressourcen zu virtualisieren, die eine Kombination aus Industriestandard-Serverhardware, Speicherhardware oder Switches umfassen. In mindestens einer Ausführungsform können NFV-Systeme verwendet werden, um virtuelle oder rekonfigurierbare Implementierungen einer oder mehrerer EPC-Komponenten/Funktionen auszuführen.

[0183] Fig. 25 ist ein Blockdiagramm, das Komponenten eines Systems 2500 zum Unterstützen von Netzwerkfunktionsvirtualisierung (NFV) gemäß mindestens einer Ausführungsform darstellt. In mindestens einer Ausführungsform ist das System 2500 so veranschaulicht, dass es einen virtualisierten Infrastrukturmanager (als VIM 2502 gezeigt), eine Netzwerkfunktionsvirtualisierungsinfrastruktur (als NFVI 2504 gezeigt), einen VNF-Manager (als VNFM 2506 gezeigt), virtualisierte Netzwerkfunktionen (als VNF 2508), einen Elementmanager (als EM 2510 gezeigt), einen NFV-Orchestrator (als NFVO 2512 gezeigt) und einen Netzwerkmanager (als NM 2514 gezeigt) beinhaltet.

[0184] In mindestens einer Ausführungsform verwaltet der VIM 2502 Ressourcen der NFVI 2504. In mindestens einer Ausführungsform kann die NFVI 2504 physische oder virtuelle Ressourcen und Anwendungen (einschließlich Hypervisoren) beinhalten, die verwendet werden, um das System 2500 auszuführen. In mindestens einer Ausführungsform kann VIM 2502 einen Lebenszyklus von virtuellen Ressourcen mit NFVI 2504 verwalten (z. B. Erstellung, Wartung und Abbau von virtuellen Maschinen (VMs), die mit einer oder mehreren physischen Ressourcen verbunden sind), Verfolgen von VM-Instanzen, Verfolgen von Leistung, Fehler und Sicherheit von VM-Instanzen und zugehörigen physischen Ressourcen, und Offenlegen von VM-Instanzen und zugehörigen physischen Ressourcen für andere Verwaltungssysteme.

[0185] In mindestens einer Ausführungsform kann der VNFM 2506 die VNF 2508 verwalten. In mindestens einer Ausführungsform kann die VNF 2508 verwendet werden, um EPC-Komponenten/-Funktionen auszuführen. In mindestens einer Ausführungsform kann der VNFM 2506 einen Lebenszyklus der VNF 2508 verwalten und Leistungsfähigkeit, Fehler und Sicherheit von virtuellen Aspekten der VNF 2508 verfolgen. In mindestens einer Ausführungsform kann der EM 2510 Leistungsfähigkeit, Fehler und Sicherheit von funktionalen Aspekten der VNF 2508 verfolgen. In mindestens einer Ausführungsform können Verfolgungsdaten des VNFM 2506 und des EM 2510 in mindestens einer Ausführungsform Daten einer Leistungsfähigkeitsmessung (performance measurement - PM) umfassen, die durch den VIM 2502 oder die NFVI 2504 verwendet werden. In mindestens einer Ausführungsform können sowohl der VNFM 2506 als auch der EM 2510 eine Menge der VNF des Systems 2500 nach oben/unten skalieren.

[0186] In mindestens einer Ausführungsform kann der NFVO 2512 Ressourcen der NFVI 2504 koordinieren, autorisieren, freigeben und einsetzen, um einen angeforderten Dienst bereitzustellen (z. B. um eine EPC-Funktion, -Komponente oder ein Slice auszuführen). In mindestens einer Ausführungsform kann der NM 2514 ein Paket von Endbenutzerfunktionen mit Verantwortung für die Verwaltung eines Netzwerks bereitstellen, das Netzwerkelemente mit VNF, nicht virtualisierte Netzwerkfunktionen oder beides beinhalten kann (die Verwaltung der VNF kann über den EM 2510 erfolgen).

COMPUTERBASIERTE SYSTEME

[0187] Die folgenden Figuren legen ohne Einschränkung beispielhafte computerbasierte Systeme dar, die verwendet werden können, um mindestens eine Ausführungsform zu implementieren.

[0188] Fig. 26 zeigt ein verteiltes System 2600 gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform beinhaltet das Verarbeitungssystem 2600 einen oder mehrere Prozessoren 2602 und einen oder mehrere Grafikprozessoren 2608 und kann ein Einzelprozessor-Desktop-System, ein Mehrprozessor-Workstation-System oder ein Server-System sein, dass eine große Anzahl von Prozessoren 2602 oder Prozessorkernen 2607 aufweist. In mindestens einer Ausführungsform ist das System 2600 eine Verarbeitungsplattform, die in eine integrierte Schaltung als System-on-a-Chip („SoC“) zur Verwendung in mobilen, tragbaren oder eingebetteten Vorrichtungen integriert ist.

[0189] In mindestens einer Ausführungsform kann das Verarbeitungssystem 2600 eine serverbasierte Spielplattform, eine Spielkonsole, eine Medienkonsole, eine mobile Spielkonsole, eine Handheld-Spielekonsole oder eine Online-Spielekonsole beinhalten oder darin integriert sein. In mindestens einer Ausführungsform ist das Verarbeitungssystem 2600 ein Mobiltelefon, ein Smartphone, eine Tablet-Rechenvorrichtung oder eine mobile Internetvorrichtung. In mindestens einer Ausführungsform kann das Verarbeitungssystem 2600 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 2600 eine Fernseh- oder Set-Top-Box-Vorrichtung, die einen oder mehrere Prozessoren 2602 und eine grafische Schnittstelle aufweist, die durch einen oder mehrere Grafikprozessoren 2608 erzeugt wird.

[0190] In mindestens einer Ausführungsform beinhalten ein oder mehrere Prozessoren 2602 jeweils einen oder mehrere Prozessorkerne 2607 zum Verarbeiten von Anweisungen, die bei ihrer Ausführung Operationen für System- und Benutzer-Software durchführen. In mindestens einer Ausführungsform ist jeder von einem oder mehreren Prozessorkernen 2607 dazu konfiguriert, einen konkreten Anweisungssatz 2609 zu verarbeiten. In mindestens einer Ausführungsform kann der Anweisungssatz 2609 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) ermöglichen. In mindestens einer Ausführungsform können die Prozessorkerne 2607 jeweils einen anderen Anweisungssatz 2609 verarbeiten, der Anweisungen beinhalten kann, um die Emulation anderer Anweisungssätze zu ermöglichen. In mindestens einer Ausführungsform kann der Prozessorkern 2607 auch andere Verarbeitungsvorrichtungen beinhalten, wie etwa einen digitalen Signalprozessor („DSP“).

[0191] In mindestens einer Ausführungsform beinhaltet der Prozessor 2602 einen Cache-Speicher („Cache“) 2604. In mindestens einer Ausführungsform kann der Prozessor 2602 einen einzelnen internen Cache oder mehrere Levels von internem Cache aufweisen. In mindestens einer Ausführungsform wird der Cache-Speicher von verschiedenen Komponenten des Prozessors 2602 gemeinsam genutzt. In mindestens einer Ausführungsform verwendet der Prozessor 2602 auch einen externen Cache (z. B. einen Level-3-(L3-)Cache oder Last-Level-Cache (LLC)) (nicht dargestellt), der unter Verwendung bekannter Cache-Kohärenztechniken von den Prozessorkernen 2607 gemeinsam genutzt werden kann. In mindestens einer Ausführungsform ist die Registerdatei 2606 zusätzlich im Prozessor 2602 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 Registerbank 2606 Allzweckregister oder andere Register beinhalten.

[0192] In mindestens einer Ausführungsform sind ein oder mehrere Prozessoren 2602 mit einem oder mehreren Schnittstellenbussen 2610 gekoppelt, um Kommunikationssignale, wie etwa Adress-, Daten- oder Steuersignale, zwischen dem Prozessor 2602 und anderen Komponenten im System 2600 zu übertragen. In mindestens einer Ausführungsform kann der Schnittstellenbus 2610 ein Prozessorbus sein, wie etwa eine Version eines Direct-Media-Interface-(DMI-)Busses. In mindestens einer Ausführungsform ist der Schnittstellenbus 2610 nicht auf einen DMI-Bus beschränkt und kann einen oder mehrere Peripheriekomponentenzwischenverbindungsbusse (z. B. Peripheral Component Interconnect - „PCI“, PCI Express („PCIe“)), Speicherbusse oder andere Arten von Schnittstellenbussen beinhalten. In mindestens einer Ausführungsform beinhalten die Prozessor(en) 2602 eine integrierte Speichersteuerung 2616 und einen Plattformsteuerungs-Hub 2630. In mindestens einer Ausführungsform ermöglicht die Speichersteuerung 2616 die Kommunikation zwischen einer Speichervorrichtung und anderen Komponenten des Systems 2600, während der Plattformsteuerungs-Hub (platform controller hub - PCH) 2630 Verbindungen mit E/A-Vorrichtungen über einen lokalen E/A-Bus bereitstellt.

[0193] In mindestens einer Ausführungsform kann die Speichervorrichtung 2620 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 2620 als Systemspeicher für das Verarbeitungssystem 2600 arbeiten, um Daten 2622 und Anweisungen 2621 zur Verwendung zu speichern, wenn ein oder mehrere Prozessoren 2602 eine Anwendung oder einen Prozess ausführen. In mindestens einer Ausführungsform ist die Speichersteuerung 2616 zudem mit einem optionalen externen Grafikprozessor 2612 gekoppelt, der mit einem oder mehreren Grafikprozessoren 2608 in den Prozessoren 2602 kommunizieren kann, um Grafik- und Medienoperationen durchzuführen. In mindestens einer Ausführungsform kann eine Anzeigevorrichtung 2611 mit dem/den Prozessor(en) 2602 ver-

bunden sein. In mindestens einer Ausführungsform kann die Anzeigevorrichtung 2611 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 Anzeigevorrichtung 2611 eine am Kopf befestigte Anzeige (head mounted display - HMD) beinhalten, wie etwa eine stereoskopische Anzeigevorrichtung zur Verwendung bei Virtual-Reality-(VR-)Anwendungen oder Augmented-Reality-(AR-)Anwendungen.

[0194] In mindestens einer Ausführungsform ermöglicht der Plattformsteuerungs-Hub 2630, dass Peripheriegeräte mit der Speichervorrichtung 2620 und dem Prozessor 2602 über einen Hochgeschwindigkeits-E/A-Bus verbunden werden. In mindestens einer Ausführungsform beinhalten die E/A-Peripheriegeräte eine Audiosteuerung 2646, eine Netzsteuerung 2634, eine Firmware-Schnittstelle 2628, einen drahtlosen Sendeempfänger 2626, Berührungssensoren 2625 und eine Datenspeichervorrichtung 2624 (z. B. Festplattenlaufwerk, Flash-Speicher usw.). In mindestens einer Ausführungsform kann die Datenspeichervorrichtung 2624 über eine Speicherschnittstelle (z. B. SATA) oder über einen Peripheriebus wie etwa einen PCI-Bus oder PCIe-Bus verbunden sein. In mindestens einer Ausführungsform können die Berührungssensoren 2625 Touchscreen-Sensoren, Drucksensoren oder Fingerabdrucksensoren beinhalten. In mindestens einer Ausführungsform kann der drahtlose Sendeempfänger 2626 ein Wi-Fi-Sendeempfänger, ein Bluetooth-Sendeempfänger oder ein Sendeempfänger für ein mobiles Netz, wie etwa ein 3G-, 4G- oder Long-Term-Evolution-(LTE-)Sendeempfänger sein. In mindestens einer Ausführungsform ermöglicht die Firmware-Schnittstelle 2628 die Kommunikation mit der System-Firmware und kann in mindestens einer Ausführungsform eine vereinheitlichte erweiterbare Firmware-Schnittstelle (unified extensible firmware interface - „UEFI“) sein. In mindestens einer Ausführungsform kann die Netzwerksteuerung 2634 eine Netzwerkverbindung zu einem drahtgebundenen Netzwerk ermöglichen. In mindestens einer Ausführungsform ist eine Hochleistungsnetzsteuerung (nicht gezeigt) mit dem Schnittstellenbus 2610 gekoppelt. In mindestens einer Ausführungsform ist die Audiosteuerung 2646 eine Mehrkanal-High-Definition-Audiosteuerung. In mindestens einer Ausführungsform beinhaltet das System 2600 eine optionale ältere E/A-Steuerung 2640 zum Koppeln von älteren Vorrichtungen (z. B. Personal System 2 (PS/2)) mit dem System 2600. In mindestens einer Ausführungsform kann der Plattformsteuerungs-Hub 2630 auch mit einer oder mehreren Universal-Serial-Bus-(USB-)Steuerungen 2642 verbunden sein, die mit Eingabevorrichtungen wie etwa Kombinationen aus Tastatur und Maus 2643, einer Kamera 2644 oder anderen USB-Eingabevorrichtungen verbunden sind.

[0195] In mindestens einer Ausführungsform kann eine Instanz der Speichersteuerung 2616 und des Plattformsteuerungs-Hubs 2630 in einen diskreten externen Grafikprozessor, wie etwa den externen Grafikprozessor 2612, integriert sein. In mindestens einer Ausführungsform können der Plattformsteuerungs-Hub 2630 und/oder die Speichersteuerung 2616 extern zu einem oder mehreren Prozessor(en) 2602 sein. Zum Beispiel kann das Verarbeitungssystem 2600 in mindestens einer Ausführungsform eine externe Speichersteuerung 2616 und einen Plattformsteuerungs-Hub 2630 beinhalten, die als Speichersteuerungs-Hub und Peripheriegerätesteuerungs-Hub innerhalb eines Systemchipsatzes konfiguriert sein können, der mit Prozessor(en) 2602 kommuniziert.

[0196] Fig. 27 zeigt ein verteiltes System 2700 gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform kann das Computersystem 2700 ein System mit verbundenen Vorrichtungen und Komponenten, ein SOC oder eine Kombination davon sein. In mindestens einer Ausführungsform ist das Computersystem 2700 mit einem Prozessor 2702 gebildet, der Ausführungseinheiten beinhalten kann, um eine Anweisung auszuführen. In mindestens einer Ausführungsform kann das Computersystem 2700 ohne Einschränkung eine Komponente beinhalten, wie etwa einen Prozessor 2702, um Ausführungseinheiten einzusetzen, die eine Logik beinhalten, um Algorithmen zum Verarbeiten von Daten durchzuführen. In mindestens einer Ausführungsform kann das Computersystem 2700 Prozessoren beinhalten, wie etwa PENTIUM®-Prozessorfamilie, Mikroprozessoren von XeonTM, Itanium®, XScaleTM und/oder StrongARMTM, Intel® Core™ oder Intel® Nervana™, die von Intel Corporation of Santa Clara, Kalifornien, erhältlich sind, obwohl auch andere Systeme (die PCs mit anderen Mikroprozessoren, Engineering-Workstations, Set-Top-Boxen und dergleichen beinhalten) verwendet werden können. In mindestens einer Ausführungsform kann das Computersystem 2700 eine Version des WINDOWS-Betriebssystems ausführen, das von der Microsoft Corporation in Redmond, Washington, erhältlich ist, obwohl auch andere Betriebssysteme (in mindestens einer Ausführungsform UNIX und Linux), eingebettete Software und/oder grafische Benutzeroberflächen verwendet werden können.

[0197] In mindestens einer Ausführungsform kann das Computersystem 2700 in anderen Vorrichtungen verwendet werden, wie etwa Handheld-Vorrichtungen und eingebetteten Anwendungen. Einige Beispiele für die mindestens eine Ausführungsform von tragbaren Vorrichtungen sind Mobiltelefone, Internetprotokoll-Vorrich-

tungen, Digitalkameras, persönliche digitale Assistenten (personal digital assistants - „PDAs“) und tragbare PCs. In mindestens einer Ausführungsform können eingebettete Anwendungen einen Mikrocontroller, einen digitalen Signalprozessor (digital signal processor- „DSP“), ein System auf einem Chip („SoC“), Netzcomputer („NetPCs“), Set-Top-Boxen, Netz-Hubs, Switches für ein Weitverkehrsnetz (wide area network - „WAN“) oder ein beliebiges anderes System beinhalten, das eine oder mehrere Anweisungen durchführen kann.

[0198] In mindestens einer Ausführungsform kann das Computersystem 2700 ohne Einschränkung einen Prozessor 2702 beinhalten, der ohne Einschränkung eine oder mehrere Ausführungseinheiten 2708 beinhalten kann, die konfiguriert sein können, um ein Compute Unified Device Architecture („CUDA“) (CUDA® wird von NVIDIA Corporation, Santa Clara, Kalifornien entwickelt) Programm auszuführen. In mindestens einer Ausführungsform ist ein CUDA-Programm mindestens ein Abschnitt einer Softwareanwendung, der in einer CUDA-Programmiersprache geschrieben ist. In mindestens einer Ausführungsform ist das Computersystem 2700 ein Desktop- oder Serversystem mit einem einzelnen Prozessor. In mindestens einer Ausführungsform kann das Computersystem 2700 ein Mehrprozessorsystem sein. In mindestens einer Ausführungsform kann der Prozessor 2702 ohne Einschränkung einen CISC-Mikroprozessor, einen RISC-Mikroprozessor, einen VLIW-Mikroprozessor, einen Prozessor, der eine Kombination von Anweisungssätzen implementiert, oder eine beliebige andere Prozessorvorrichtung, wie etwa ein digitaler Signalprozessor, in mindestens einer Ausführungsform beinhalten. In mindestens einer Ausführungsform kann der Prozessor 2702 an einen Prozessorbus 2710 gekoppelt sein, der Datensignale zwischen dem Prozessor 2702 und anderen Komponenten im Computersystem 2700 übertragen kann.

[0199] In mindestens einer Ausführungsform kann der Prozessor 2702 ohne Einschränkung einen internen Cachespeicher („Cache“) 2704 der Ebene 1 („L1“) beinhalten. In mindestens einer Ausführungsform kann der Prozessor 2702 einen einzelnen internen Cache oder mehrere Levels von internem Cache aufweisen. In mindestens einer Ausführungsform kann sich der Cache-Speicher extern zum Prozessor 2702 befinden. In mindestens einer Ausführungsform kann der Prozessor 2702 auch eine Kombination von sowohl internen als auch externen Caches beinhalten. In mindestens einer Ausführungsform kann eine Registerdatei 2706 unterschiedliche Typen von Daten in verschiedenen Registern speichern, darunter ohne Einschränkung in Integer-Registern, Gleitkommaregistern, Statusregistern und einem Anweisungszeigerregister.

[0200] In mindestens einer Ausführungsform befindet sich die Ausführungseinheit 2708, die ohne Einschränkung eine Logik zum Durchführen von Ganzzahl- und Gleitkommaoperationen beinhaltet, ebenfalls im Prozessor 2702. Der Prozessor 2702 kann auch Festwertspeicher (read only memory - „ROM“) für Mikrocode („µcode“) beinhalten, der Mikrocode für bestimmte Makroanweisungen speichert. In mindestens einer Ausführungsform kann die Ausführungseinheit 2708 Logik beinhalten, um einen gepackten Anweisungssatz 2709 zu handhaben. In mindestens einer Ausführungsform können Operationen, die von vielen Multimediaanwendungen verwendet werden, unter Verwendung von gepackten Daten in einem Allzweckprozessor 2702 durch das Beinhalten des gepackten Anweisungssatzes 2709 in einen Anweisungssatz eines Allzweckprozessors 2702 durchgeführt werden, zusammen mit einer zugehörigen Schaltung, um Befehle auszuführen. In mindestens einer Ausführungsform können viele Multimediaanwendungen beschleunigt und effizienter durch das Verwenden der vollen Breite des Datenbusses eines Prozessors zum Durchführen 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 an einem Datenelement nach dem anderen durchzuführen.

[0201] In mindestens einer Ausführungsform kann die Ausführungseinheit 2708 auch in Mikrocontrollern, eingebetteten Prozessoren, Grafikvorrichtungen, DSPs und anderen Typen von Logikschaltungen verwendet werden. In mindestens einer Ausführungsform kann das Computersystem 2700 ohne Einschränkung einen Speicher 2720 beinhalten. In mindestens einer Ausführungsform kann der Speicher 2720 als eine DRAM-Vorrichtung, eine SRAM-Vorrichtung, Flash-Speichervorrichtung oder andere Speichervorrichtung implementiert sein. Der Speicher 2720 kann Anweisung(en) 2719 und/oder Daten 2721 speichern, die durch Datensignale dargestellt sind, die durch den Prozessor 2702 ausgeführt werden können.

[0202] In mindestens einer Ausführungsform kann ein Systemlogikchip mit dem Prozessorbus 2710 und den Speicher 2720 gekoppelt sein. In mindestens einer Ausführungsform kann ein Systemlogikchip ohne Einschränkung einen Speichersteuerungs-Hub (memory controller hub - „MCH“) 2716 beinhalten und kann der Prozessor 2702 mit dem MCH 2716 über den Prozessorbus 2710 kommunizieren. In mindestens einer Ausführungsform kann der MCH 2716 einen Speicherpfad mit hoher Bandbreite 2718 zum Speicher 2720 zur Anweisungs- und Datenspeicherung und zur Speicherung von Grafikbefehlen, Daten und Texturen bereitstellen. In mindestens einer Ausführungsform kann der MCH 2716 Datensignale zwischen dem Prozessor 2702,

dem Speicher 2720 und anderen Komponenten im Computersystem 2700 leiten und Datensignale zwischen dem Prozessorbus 2710, dem Speicher 2720 und einer System-E/A 2722 überbrücken. In mindestens einer Ausführungsform kann der Systemlogikchip einen Grafikport zum Koppeln an eine Grafiksteuerung bereitstellen. In mindestens einer Ausführungsform kann der MCH 2716 über einen Speicherpfad mit hoher Bandbreite 2718 mit dem Speicher 2720 gekoppelt sein und die Grafik-/Videokarte 2712 kann mit dem MCH 2716 über eine Verbindung eines beschleunigten Grafikports (Accelerated Graphics Port - „AGP“) 2714 gekoppelt sein.

[0203] In mindestens einer Ausführungsform kann das Computersystem 2700 die System-E/A 2722 verwenden, die ein proprietärer Hubschnittstellenbus ist, um den MCH 2716 mit dem E/A-Steuerungshub (I/O controller hub - „ICH“) 2730 zu koppeln. In mindestens einer Ausführungsform kann der ICH 2730 direkte Verbindungen zu einigen E/A-Vorrichtungen über einen lokalen E/A-Bus bereitstellen. In mindestens einer Ausführungsform kann ein lokaler E/A-Bus ohne Einschränkung einen Hochgeschwindigkeits-E/A-Bus zum Verbinden von Peripheriegeräten mit dem Speicher 2720, einem Chipsatz und dem Prozessor 2702 beinhalten. Beispiele können ohne Einschränkung eine Audiosteuerung 2729, einen Firmware-Hub („Flash-BIOS“) 2728, einen drahtlosen Sendeempfänger 2726, einen Datenspeicher 2724, eine ältere E/A-Steuerung 2723, die Benutzereingabe- und Tastaturschnittstellen 2725 enthält, einen seriellen Erweiterungsport 2777, wie etwa einen Universal-Serial-Bus-(„USB“-)Port, und eine Netzwerksteuerung 2734 beinhalten. Der Datenspeicher 2724 kann ein Festplattenlaufwerk, ein Diskettenlaufwerk, eine CD-ROM-Vorrichtung, eine Flash-Speichervorrichtung oder eine andere Massenspeichervorrichtung umfassen.

[0204] In mindestens einer Ausführungsform zeigt **Fig. 27** ein System, das miteinander verbundene Hardware-Vorrichtungen oder „Chips“ beinhaltet. In mindestens einer Ausführungsform kann **Fig. 27** ein beispielhaftes SoC darstellen. In mindestens einer Ausführungsform können die in **Fig. 27** veranschaulichten Vorrichtungen mit proprietären Zwischenverbindungen, standardisierten Zwischenverbindungen (z. B. PCIe) oder einer Kombination davon miteinander verbunden sein. In mindestens einer Ausführungsform sind eine oder mehrere Komponenten des Systems 2700 unter Verwendung von Compute-Express-Link („CXL“)-Verbindungen miteinander verbunden.

[0205] **Fig. 28** zeigt ein verteiltes System 2800 gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform ist das System 2800 eine elektronische Vorrichtung, die einen Prozessor 2810 nutzt. In mindestens einer Ausführungsform kann das System 2800 in mindestens einer Ausführungsform 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.

[0206] In mindestens einer Ausführungsform kann das System 2800 ohne Einschränkung einen Prozessor 2810 beinhalten, der mit einer beliebigen geeigneten Anzahl oder Art von Komponenten, Peripheriegeräten, Modulen oder Vorrichtungen kommunikativ gekoppelt ist. In mindestens einer Ausführungsform ist der Prozessor 2810 unter Verwendung eines Busses oder einer Schnittstelle gekoppelt, wie etwa eines I2C-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 USB (Versionen 1, 2 oder 3) oder eines Busses eines universellen asynchronen Empfänger/Senders (Universal Asynchronous Receiver/Transmitter - „UART“). In mindestens einer Ausführungsform veranschaulicht **Fig. 28** ein System, das miteinander verbundene Hardware-Vorrichtungen oder „Chips“ beinhaltet. In mindestens einer Ausführungsform kann **Fig. 28** ein beispielhaftes SoC darstellen. In mindestens einer Ausführungsform können die in **Fig. 28** veranschaulichten Vorrichtungen mit proprietären Zusammenschaltungen, standardisierten Zusammenschaltungen (z. B. PCIe) oder einer Kombination davon zusammengeschaltet sein. In mindestens einer Ausführungsform sind eine oder mehrere Komponenten der **Fig. 28** unter Verwendung von CXL-Zwischenverbindungen miteinander verbunden.

[0207] In mindestens einer Ausführungsform kann **Fig. 28** Folgendes beinhalten: eine Anzeige 2824, einen Berührungsbildschirm 2825, ein Touchpad 2830, eine Nahfeldkommunikationseinheit („NFC“) 2845, einen Sensor-Hub 2840, einen Wärmesensor 2846, einen Express-Chipsatz („EC“) 2835, ein Trusted-Platform-Modul („TPM“) 2838, BIOS-/Firmware-/Flash-Speicher („BIOS, FW Flash“) 2822, einen DSP 2860, ein Festkörperlaufwerk (Solid State Disk - „SSD“) oder ein Festplattenlaufwerk (Hard Disk Drive - „HDD“) 2820, eine drahtlose lokale Netzwerkeinheit (Wireless Local Area Network - „WLAN“) 2850, eine Bluetooth-Einheit 2852, eine drahtlose Weitbereichsnetzwerkeinheit (Wireless Wide Area Network - „WWAN“) 2856, ein globales Positionsbestimmungssystem („GPS“) 2855, eine Kamera („USB 3.0-Kamera“) 2854, wie etwa eine USB 3.0-Kamera, und/oder eine Speichereinheit mit doppelter Datenrate bei niedriger Leistung (Low Power

Double Data Rate - „LPDDR“) („LPDDR3“) 2815, die in mindestens einer Ausführungsform im LPDDR3-Standard implementiert ist. Diese Komponenten können jeweils auf eine beliebige geeignete Weise implementiert sein.

[0208] In mindestens einer Ausführungsform können andere Komponenten durch die vorstehend erörterten Komponenten kommunikativ an den Prozessor 2810 gekoppelt sein. In mindestens einer Ausführungsform können ein Beschleunigungsmesser 2841, ein Umgebungslichtsensor (ambient light sensor - „ALS“) 2842, ein Kompass 2843 und ein Gyroskop 2844 mit dem Sensor-Hub 2840 kommunikativ gekoppelt sein. In mindestens einer Ausführungsform können der Wärmesensor 2839, ein Lüfter 2837, eine Tastatur 2846 und ein Touchpad 2830 mit dem EC 2835 kommunikativ gekoppelt sein. In mindestens einer Ausführungsform können der Lautsprecher 2863, die Kopfhörer 2864 und das Mikrofon („mic“) 2865 mit einer Audioeinheit („Audiocodec und Klasse-D-Verstärker“) 2864 kommunikativ gekoppelt sein, die wiederum mit dem DSP 2860 kommunikativ gekoppelt sein kann. In mindestens einer Ausführungsform kann die Audioeinheit 2864 zum Beispiel und ohne Einschränkung einen Audiokodierer/-dekodierer („Codec“) und einen Klasse-D-Verstärker beinhalten. In mindestens einer Ausführungsform kann eine SIM-Karte („SIM“) 2857 mit der WWAN-Einheit 2856 kommunikativ gekoppelt sein. In mindestens einer Ausführungsform können Komponenten wie die WLAN-Einheit 2850 und die Bluetooth-Einheit 2852 sowie die WWAN-Einheit 2856 in einem Formfaktor der nächsten Generation (Next Generation Form Factor - „NGFF“) implementiert sein.

[0209] Fig. 29 zeigt eine beispielhafte integrierte Schaltung 2900 gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform ist die beispielhafte integrierte Schaltung 2900 ein SoC, das unter Verwendung eines oder mehrerer IP-Kerne hergestellt werden kann. In mindestens einer Ausführungsform beinhaltet die integrierte Schaltung 2900 einen oder mehrere Anwendungsprozessor(en) 2905 (z. B. CPUs), mindestens einen Grafikprozessor 2910 und sie kann zusätzlich einen Bildprozessor 2915 und/oder einen Videoprozessor 2920 beinhalten, von denen jeder beliebige ein modularer IP-Kern sein kann. In mindestens einer Ausführungsform beinhaltet die integrierte Schaltung 2900 Peripherie- oder Buslogik, die eine USB-Steuerung 2925, eine UART-Steuerung 2930, eine SPI/SDIO-Steuerung 2935 und eine I2S/I2C-Steuerung 2940 beinhaltet. In mindestens einer Ausführungsform kann die integrierte Schaltung 2900 eine Anzeigevorrichtung 2945 beinhalten, die mit einer High-Definition-Multimedia-Interface-(HDMI-)Steuerung 2950 und/oder einer Mobile-Industry-Processor-Interface-(MIPI-)Anzeigeschnittstelle 2955 gekoppelt ist. In mindestens einer Ausführungsform kann die Speicherung durch ein Flash-Speicherteilsystem 2960 bereitgestellt sein, das Flash-Speicher und eine Flash-Speichersteuerung beinhaltet. In mindestens einer Ausführungsform kann eine Speicherschnittstelle über eine Speichersteuerung 2965 für den Zugriff auf SDRAM- oder SRAM-Speichervorrichtungen bereitgestellt sein. In mindestens einer Ausführungsform beinhalten einige integrierte Schaltungen zusätzlich eine eingebettete Sicherheits-Engine 2970.

[0210] Fig. 30 zeigt ein Rechensystem 3000 gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform beinhaltet das Rechensystem 3000 ein Verarbeitungsteilsystem 3001 mit einem oder mehreren Prozessoren 3002 und einem Systemspeicher 3004, der über einen Zwischenverbindungspfad kommuniziert, der einen Speicher-Hub 3005 beinhalten kann. In mindestens einer Ausführungsform kann der Speicher-Hub 3005 eine getrennte Komponente innerhalb einer Chipsatzkomponente sein oder kann in einen oder mehrere Prozessoren 3002 integriert sein. In mindestens einer Ausführungsform ist der Speicher-Hub 3005 über eine Kommunikationsverbindung 3006 mit einem E/A-Teilsystem 3011 gekoppelt. In mindestens einer Ausführungsform beinhaltet das E/A-Teilsystem 3011 einen E/A-Hub 3007, der es dem Rechensystem 3000 ermöglichen kann, Eingaben von einer oder mehreren Eingabevorrichtung(en) 3008 zu empfangen. In mindestens einer Ausführungsform kann der E/A-Hub 3007 es einer Anzeigesteuerung, die in einem oder mehreren Prozessor(en) 3002 beinhaltet sein kann, ermöglichen, einer oder mehreren Anzeigevorrichtung(en) 3010A Ausgaben bereitzustellen. In mindestens einer Ausführungsform können eine oder mehrere mit dem E/A-Hub 3007 gekoppelte Anzeigevorrichtung(en) 3010A eine lokale, interne oder eingebettete Anzeigevorrichtung beinhalten.

[0211] In mindestens einer Ausführungsform beinhaltet das Verarbeitungsteilsystem 3001 einen oder mehrere Parallelprozessor(en) 3012, die über einen Bus oder eine andere Kommunikationsverknüpfung 3013 mit dem Speicher-Hub 3005 gekoppelt sind. In mindestens einer Ausführungsform kann die Kommunikationsverbindung 3013 eine von einer beliebigen Anzahl von auf Standards basierenden Kommunikationsverbindungstechnologien oder -protokollen sein, wie etwa, aber nicht beschränkt auf PCIe, oder kann eine herstellerspezifische Kommunikationsschnittstelle oder Kommunikationsstruktur sein. In mindestens einer Ausführungsform bilden ein oder mehrere Parallelprozessor(en) 3012 ein rechnerisch fokussiertes Parallel- oder Vektorverarbeitungssystem, das eine große Anzahl von Verarbeitungskernen und/oder Verarbeitungs-Clustern beinhalten kann, wie etwa einen Prozessor mit vielen integrierten Kernen. In mindestens einer Aus-

führungsform bilden ein oder mehrere Parallelprozessoren 3012 ein Grafikverarbeitungsteilsystem, das Pixel an eine oder mehrere Anzeigevorrichtungen 3010A ausgeben kann, die über den E/A-Hub 3007 gekoppelt sind. In mindestens einer Ausführungsform können ein oder mehrere Parallelprozessoren 3012 auch einen Anzeigesteuerung und eine Anzeigeschnittstelle (nicht dargestellt) beinhalten, um eine direkte Verbindung zu einer oder mehreren Anzeigevorrichtungen 3010B zu ermöglichen.

[0212] In mindestens einer Ausführungsform kann eine Systemspeichereinheit 3014 mit dem E/A-Hub 3007 verbunden sein, um einen Speichermechanismus für das Rechensystem 3000 bereitzustellen. In mindestens einer Ausführungsform kann ein E/A-Switch 3016 verwendet werden, um einen Schnittstellenmechanismus bereitzustellen, der Verbindungen zwischen dem E/A-Hub 3007 und anderen Komponenten ermöglicht, wie etwa einem Netzadapter 3018 und/oder einem drahtlosen Netzadapter 3019, die in eine Plattform integriert werden können, sowie verschiedenen anderen Vorrichtungen, die über eine oder mehrere Erweiterungsvorrichtung(en) 3020 hinzugefügt werden können. In mindestens einer Ausführungsform kann der Netzwerkadapter 3018 ein Ethernet-Adapter oder ein anderer drahtgebundener Netzwerkadapter sein. In mindestens einer Ausführungsform kann der drahtlose Netzwerkadapter 3019 eine oder mehrere von einer Wi-Fi-, Bluetooth-, NFC- oder anderen Netzwerkvorrichtung beinhalten, die ein oder mehrere drahtlose Funkvorrichtungen beinhaltet.

[0213] In mindestens einer Ausführungsform kann das Rechensystem 3000 andere Komponenten beinhalten, die nicht ausdrücklich dargestellt sind, die USB oder andere Port-Verbindungen, optische Speicherlaufwerke, Videoaufnahmevorrichtungen und/oder Variationen davon beinhalten und auch mit dem E/A-Hub 3007 verbunden sein können. In mindestens einer Ausführungsform können Kommunikationswege, die verschiedene Komponenten in **Fig. 30** verbinden, unter Verwendung beliebiger geeigneter Protokolle implementiert sein, wie etwa auf PCI basierte Protokolle (z. B. PCIe) oder andere Bus- oder Punkt-zu-Punkt-Kommunikationsschnittstellen und/oder -protokolle, wie etwa eine NVLink-Hochgeschwindigkeitszwischenverbindung oder Zwischenverbindungsprotokolle.

[0214] In mindestens einer Ausführungsform schließen ein oder mehrere Parallelprozessoren 3012 eine Schaltung ein, die für die Grafik- und Videoverarbeitung optimiert ist, was in mindestens einer Ausführungsform eine Videoausgabeschaltung beinhaltet, und eine Grafikverarbeitungseinheit („GPU“) bildet. In mindestens einer Ausführungsform schließen ein oder mehrere Parallelprozessoren 3012 eine Schaltung ein, die für eine universelle Verarbeitung optimiert ist. In mindestens einer Ausführungsform können die Komponenten des Rechensystems 3000 in ein oder mehrere andere Systemelemente auf einer einzelnen integrierten Schaltung integriert sein. Zum Beispiel können in mindestens einer Ausführungsform ein oder mehrere Parallelprozessoren 3012, Speicher-Hub 3005, Prozessoren 3002 und E/A-Hub 3007 in eine integrierte System-on-Chip-(SoC)-Schaltung integriert sein. In mindestens einer Ausführungsform können die Komponenten des Rechensystems 3000 in ein einzelnes Gehäuse integriert sein, um eine Konfiguration mit einem System in einem Gehäuse (system in package - SIP) zu bilden. In mindestens einer Ausführungsform kann mindestens ein Abschnitt der Komponenten des Rechensystems 3000 in ein Multi-Chip-Modul (multi-chip module - „MCM“) integriert sein, das mit anderen Multi-Chip-Modulen zu einem modularen Rechensystem zusammengeschaltet sein kann. In mindestens einer Ausführungsform werden das E/A-Teilsystem 3011 und die Anzeigevorrichtungen 3010B aus dem Rechensystem 3000 weggelassen.

VERARBEITUNGSSYSTEME

[0215] Die folgenden Figuren legen ohne Einschränkung beispielhafte Verarbeitungssysteme dar, die verwendet werden können, um mindestens eine Ausführungsform zu implementieren.

[0216] Fig. 31 zeigt eine beschleunigte Verarbeitungseinheit (accelerated processing unit - „APU“) 3100 gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform wird die APU 3100 von der AMD Corporation, Santa Clara, Kalifornien, entwickelt. In mindestens einer Ausführungsform kann die APU 3100 zum Ausführen eines Anwendungsprogramms wie ein CUDA-Programm konfiguriert sein. In mindestens einer Ausführungsform beinhaltet die APU 3100 ohne Einschränkung einen Kernkomplex 3110, einen Grafikkomplex 3140, eine Struktur 3160, E/A-Schnittstellen 3170, Speichersteuerungen 3180, eine Anzeigesteuerung 3192 und eine Multimedia-Engine 3194. In mindestens einer Ausführungsform kann die APU 3100 ohne Einschränkung eine beliebige Anzahl von Kernkomplexen 3110, eine beliebige Anzahl von Grafikkomplexen 3150, eine beliebige Anzahl von Anzeigesteuerungen 3192 und eine beliebige Anzahl von Multimedia-Engines 3194 in einer beliebigen Kombination beinhalten. Zu Erläuterungszwecken werden hierin mehrere Instanzen von gleichen Objekten mit Bezugszeichen bezeichnet, die ein Objekt identifizieren, und in Klammern gesetzte Nummern identifizieren eine Instanz, wo dies erforderlich ist.

[0217] In mindestens einer Ausführungsform ist der Kernkomplex 3110 eine CPU, ist der Grafikkomplex 3140 eine GPU und ist die APU 3100 eine Verarbeitungseinheit, die ohne Einschränkung 3110 und 3140 auf einem einzelnen Chip integriert. In mindestens einer Ausführungsform können dem Kernkomplex 3110 einige Tasks zugewiesen werden und können dem Grafikkomplex 3140 andere Tasks zugewiesen werden. In mindestens einer Ausführungsform ist der Kernkomplex 3110 ein Masterprozessor der APU 3100, der Operationen anderer Prozessoren steuert und koordiniert. In mindestens einer Ausführungsform ist der Kernkomplex 3110 ein Masterprozessor der APU 3100, der Operationen anderer Prozessoren steuert und koordiniert. In mindestens einer Ausführungsform gibt der Kernkomplex 3110 Befehle aus, die eine Operation des Grafikkomplexes 3140 steuern. In mindestens einer Ausführungsform kann der Kernkomplex 3110 zum Ausführen vom Host ausführbaren Codes, der vom CUDA-Quellcode abgeleitet ist, konfiguriert sein, und der Grafikkomplex 3140 kann zum Ausführen von der Vorrichtung ausführbaren Codes, der vom CUDA-Quellcode abgeleitet ist, konfiguriert sein.

[0218] In mindestens einer Ausführungsform beinhaltet der Kernkomplex 3110 ohne Einschränkung Kerne 3120(1)-3120(4) und einen L3-Cache 3130. In mindestens einer Ausführungsform kann der Kernkomplex 3110 ohne Einschränkung eine beliebige Anzahl von Kernen 3120 und eine beliebige Anzahl und Art von Caches in beliebiger Kombination beinhalten. In mindestens einer Ausführungsform sind die Kerne 3120 konfiguriert, um Anweisungen einer konkreten Anweisungssatzarchitektur (instruction set architecture - „ISA“) auszuführen. In mindestens einer Ausführungsform ist jeder Kern 3120 ein CPU-Kern.

[0219] In mindestens einer Ausführungsform beinhaltet jeder Kern 3120 ohne Einschränkung eine Abruf-/Decodiereinheit 3122, eine Ganzzahl-Ausführungs-Engine 3124, eine Gleitkomma-Ausführungs-Engine 3126 und einen L2-Cache 3128. In mindestens einer Ausführungsform ruft die Abruf-/Decodiereinheit 3122 Anweisungen ab, decodiert solche Anweisungen, erzeugt Mikrooperationen und sendet separate Mikroanweisungen an die Ganzzahl-Ausführungs-Engine 3124 und die Gleitkomma-Ausführungs-Engine 3126 ab. In mindestens einer Ausführungsform kann die Abruf-/Decodiereinheit 3122 gleichzeitig eine Mikroanweisung an die Ganzzahl-Ausführungs-Engine 3124 und eine weitere Mikroanweisung an die Gleitkomma-Ausführungs-Engine 3126 absenden. In mindestens einer Ausführungsform führt die Ganzzahl-Ausführungs-Engine 3124 ohne Einschränkung Ganzzahl- und Speicheroperationen aus. In mindestens einer Ausführungsform führt die Gleitkomma-Engine 3126 ohne Einschränkung Gleitkomma- und Vektoroperationen aus. In mindestens einer Ausführungsform sendet die Abruf-/Decodiereinheit 3122 Mikroanweisungen an eine einzelne Ausführungs-Engine aus, die sowohl die Ganzzahl-Ausführungs-Engine 3124 als auch die Gleitkomma-Ausführungs-Engine 3126 ersetzt.

[0220] In mindestens einer Ausführungsform kann jeder Kern 3120(i), wobei i eine ganze Zahl ist, die eine konkrete Instanz des Kerns 3120 darstellt, auf den im Kern 3120(i) beinhalteten L2-Cache 3128(i) zugreifen. In mindestens einer Ausführungsform ist jeder im Kernkomplex 3110(j) beinhaltete Kern 3120, wobei j eine ganze Zahl ist, die eine konkrete Instanz des Kernkomplexes 3110 darstellt, mit anderen Kernen 3120 im Kernkomplex 3110(j) über den im Kernkomplex 3110(j) beinhalteten L3-Cache 3130(j) verbunden. In mindestens einer Ausführungsform können die im Kernkomplex 3110(j) beinhalteten Kerne 3120, wobei j eine ganze Zahl ist, die eine konkrete Instanz des Kernkomplexes 3110 darstellt, auf den gesamten L3-Cache 3130(j), der im Kernkomplex 3110(j) beinhaltet ist, zugreifen. In mindestens einer Ausführungsform kann der L3-Cache 3130 ohne Einschränkung eine beliebige Anzahl von Slices beinhalten.

[0221] In mindestens einer Ausführungsform kann der Grafikkomplex 3140 konfiguriert sein, um Rechenoperationen auf hochparallele Weise durchzuführen. In mindestens einer Ausführungsform ist der Grafikkomplex 3140 dazu konfiguriert, Grafikpipelineoperationen auszuführen, wie beispielsweise Zeichenbefehle, Pixeloperationen, geometrische Berechnungen und andere Operationen, die mit dem Rendern eines Bildes auf einer Anzeige verbunden sind. In mindestens einer Ausführungsform ist der Grafikkomplex 3140 konfiguriert, um Operationen auszuführen, die sich nicht auf Grafiken beziehen. In mindestens einer Ausführungsform ist der Grafikkomplex 3140 konfiguriert, um sowohl Operationen, die sich auf Grafiken beziehen, als auch Operationen, die sich nicht auf Grafiken beziehen, auszuführen.

[0222] In mindestens einer Ausführungsform beinhaltet der Grafikkomplex 3140 ohne Einschränkung eine beliebige Anzahl von Recheneinheiten 3150 und einen L2-Cache 3142. In mindestens einer Ausführungsform nutzen die Recheneinheiten 3150 den L2-Cache 3142 gemeinsam. In mindestens einer Ausführungsform ist der L2-Cache 3142 partitioniert. In mindestens einer Ausführungsform beinhaltet der Grafikkomplex 3140 ohne Einschränkung eine beliebige Anzahl von Recheneinheiten 3150 und eine beliebige Anzahl (einschließlich null) und Art von Caches. In mindestens einer Ausführungsform beinhaltet der Grafikkomplex 3140 ohne Einschränkung eine beliebige Menge an dedizierter Grafikhardware.

[0223] In mindestens einer Ausführungsform beinhaltet jede Recheneinheit 3150 ohne Einschränkung eine beliebige Anzahl von SIMD-Einheiten 3152 und einen gemeinsam genutzten Speicher 3154. In mindestens einer Ausführungsform implementiert jede SIMD-Einheit 3152 eine SIMD-Architektur und ist konfiguriert, um Operationen parallel durchzuführen. In mindestens einer Ausführungsform kann jede Recheneinheit 3150 eine beliebige Anzahl von Thread-Blöcken ausführen, aber jeder Thread-Block wird auf einer einzelnen Recheneinheit 3150 ausgeführt. In mindestens einer Ausführungsform beinhaltet ein Thread-Block ohne Einschränkung eine beliebige Anzahl von Ausführungs-Threads. In mindestens einer Ausführungsform ist eine Arbeitsgruppe ein Thread-Block. In mindestens einer Ausführungsform führt jede SIMD-Einheit 3152 einen anderen Warp aus. In mindestens einer Ausführungsform ist ein Warp eine Gruppe von Threads (z. B. 16 Threads), wobei jeder Thread in einem Warp zu einem einzelnen Thread-Block gehört und konfiguriert ist, um einen anderen Datensatz basierend auf einem einzelnen Satz von Anweisungen zu verarbeiten. In mindestens einer Ausführungsform kann Prädikation verwendet werden, um einen oder mehrere Threads in einem Warp zu deaktivieren. In mindestens einer Ausführungsform ist ein Pfad ein Thread. In mindestens einer Ausführungsform ist ein Arbeitselement ein Thread. In mindestens einer Ausführungsform ist eine Wellenfront ein Warp. In mindestens einer Ausführungsform können sich verschiedene Wellenfronten in einem Thread-Block miteinander synchronisieren und über den gemeinsam genutzten Speicher 3154 kommunizieren.

[0224] In mindestens einer Ausführungsform ist die Struktur 3160 eine Systemverbindung, die Daten- und Steuerübertragungen über den Kernkomplex 3110, Grafikkomplex 3140, E/A-Schnittstellen 3170, Speichersteuerungen 3180, eine Anzeigesteuerung 3192 und eine Multimedia-Engine 3194 ermöglicht. In mindestens einer Ausführungsform kann die APU 3100 ohne Einschränkung jede Menge und Art von Systemverbindung zusätzlich zu oder anstelle der Struktur 3160 umfassen, die Daten- und Steuerübertragungen über eine beliebige Anzahl und Art von direkt oder indirekt verbundenen Komponenten ermöglicht, die intern oder extern zu APU 3100 sein können. In mindestens einer Ausführungsform repräsentieren die E/A-Schnittstellen 3170 eine beliebige Anzahl und Art von E/A-Schnittstellen (z. B. PCI, PCI-Extended („PCI-X“), PCIe, Gigabit-Ethernet („GBE“), USB usw.). In mindestens einer Ausführungsform sind verschiedene Arten von Peripheriegeräten mit E/A-Schnittstellen 3170 gekoppelt. In mindestens einer Ausführungsform können Peripheriegeräte, die mit den E/A-Schnittstellen 3170 gekoppelt sind, ohne Einschränkung Tastaturen, Mäuse, Drucker, Scanner, Joysticks oder andere Arten von Spielesteuerungen, Medienaufzeichnungsvorrichtungen, externe Speichervorrichtungen, Netzwerkkarten usw. beinhalten.

[0225] In mindestens einer Ausführungsform zeigt die Anzeigesteuerung AMD92 Bilder auf einer oder mehreren Anzeigevorrichtungen an, wie etwa einer Flüssigkristallanzeige („LCD“). In mindestens einer Ausführungsform beinhaltet die Multimedia-Engine 240 ohne Einschränkung eine beliebige Menge und Art von Schaltungen, die sich auf Multimedia beziehen, wie etwa einen Video-Decodierer, einen Video-Codierer, einen Bildsignalprozessor usw. In mindestens einer Ausführungsform ermöglichen die Speichersteuerungen 3180 Datenübertragungen zwischen der APU 3100 und einem vereinheitlichten Systemspeicher 3190. In mindestens einer Ausführungsform nutzen der Kernkomplex 3110 und der Grafikkomplex 3140 den vereinheitlichten Systemspeicher 3190 gemeinsam.

[0226] In mindestens einer Ausführungsform implementiert die APU 3100 ein Speicherteilsystem, das ohne Einschränkung eine beliebige Menge und Art von Speichersteuerungen 3180 und Speichervorrichtungen (z. B. den gemeinsam genutzten Speicher 3154) beinhaltet, die einer Komponente zugeordnet oder von mehreren Komponenten gemeinsam genutzt werden können. In mindestens einer Ausführungsform implementiert die APU 3100 ein Cache-Teilsystem, das ohne Einschränkung einen oder mehrere Cache-Speicher (z. B. L2-Caches 2728, L3-Cache 3130 und L2-Cache 3142) beinhaltet, die jeweils für eine beliebige Anzahl von Komponenten reserviert sein können oder von diesen gemeinsam genutzt werden können (z. B. Kerne 3120, Kernkomplex 3110, SIMD-Einheiten 3152, Recheneinheiten 3150 und Grafikkomplex 3140).

[0227] **Fig. 32** zeigt eine CPU 3200 gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform wird die CPU 3200 von der AMD Corporation, Santa Clara, Kalifornien, entwickelt. In mindestens einer Ausführungsform ist die CPU 3200 zum Ausführen einer Hauptsteuersoftware, wie etwa eines Betriebssystems konfiguriert. In mindestens einer Ausführungsform ist die CPU 3200 zum Ausführen einer Hauptsteuersoftware, wie etwa eines Betriebssystems konfiguriert. In mindestens einer Ausführungsform gibt die CPU 3200 Befehle aus, die einen Betrieb einer externen GPU (nicht dargestellt) steuern. In mindestens einer Ausführungsform kann die CPU 3200 zum Ausführen vom Host ausführbarer Codes, der vom CUDA-Quellcode abgeleitet ist, konfiguriert sein, und eine externe GPU kann zum Ausführen von der Vorrichtung ausführbaren Codes, der diesem CUDA-Quellcode abgeleitet ist, konfiguriert sein. In mindestens einer Aus-

führungsform beinhaltet die CPU 3200 ohne Einschränkung eine beliebige Anzahl von Kernkomplexen 3210, eine Struktur 3260, E/A-Schnittstellen 3270 und Speichersteuerungen 3280.

[0228] In mindestens einer Ausführungsform beinhaltet der Kernkomplex 3210 ohne Einschränkung Kerne 3220(1)-3220(4) und einen L3-Cache 3230. In mindestens einer Ausführungsform kann der Kernkomplex 3210 ohne Einschränkung eine beliebige Anzahl von Kernen 3220 und eine beliebige Anzahl und Art von Caches in beliebiger Kombination beinhalten. In mindestens einer Ausführungsform sind die Kerne 3220 konfiguriert, um Anweisungen einer konkreten ISA auszuführen. In mindestens einer Ausführungsform ist jeder Kern 3220 ein CPU-Kern.

[0229] In mindestens einer Ausführungsform beinhaltet jeder Kern 3220 ohne Einschränkung eine Abruf-/Decodiereinheit 3222, eine Ganzzahl-Ausführungs-Engine 3224, eine Gleitkomma-Ausführungs-Engine 3226 und einen L2-Cache 3228. In mindestens einer Ausführungsform ruft die Abruf-/Decodiereinheit 3222 Anweisungen ab, decodiert solche Anweisungen, erzeugt Mikrooperationen und sendet separate Mikroanweisungen an die Ganzzahl-Ausführungs-Engine 3224 und die Gleitkomma-Ausführungs-Engine 3226 ab. In mindestens einer Ausführungsform kann die Abruf-/Decodiereinheit 3222 gleichzeitig eine Mikroanweisung an die Ganzzahl-Ausführungs-Engine 3224 und eine weitere Mikroanweisung an die Gleitkomma-Ausführungs-Engine 3226 absenden. In mindestens einer Ausführungsform führt die Ganzzahl-Ausführungs-Engine 3224 ohne Einschränkung Ganzzahl- und Speicheroperationen aus. In mindestens einer Ausführungsform führt die Gleitkomma-Engine 3226 ohne Einschränkung Gleitkomma- und Vektoroperationen aus. In mindestens einer Ausführungsform sendet die Abruf-/Decodiereinheit 3222 Mikroanweisungen an eine einzelne Ausführungs-Engine aus, die sowohl die Ganzzahl-Ausführungs-Engine 3224 als auch die Gleitkomma-Ausführungs-Engine 3226 ersetzt.

[0230] In mindestens einer Ausführungsform kann jeder Kern 3220(i), wobei i eine ganze Zahl ist, die eine konkrete Instanz des Kerns 3220 darstellt, auf den im Kern 3220(i) beinhalteten L2-Cache 3228(i) zugreifen. In mindestens einer Ausführungsform ist jeder im Kernkomplex 3210(j) beinhaltete Kern 3220, wobei j eine ganze Zahl ist, die eine konkrete Instanz des Kernkomplexes 3210 darstellt, mit anderen Kernen 3220 im Kernkomplex 3210(j) über den L3-Cache 3230(j), der im Kernkomplex 3210(j) beinhaltet ist, verbunden. In mindestens einer Ausführungsform können die im Kernkomplex 3210(j) beinhalteten Kerne 3220, wobei j eine ganze Zahl ist, die eine konkrete Instanz des Kernkomplexes 3210 darstellt, auf den gesamten L3-Cache 3230(j), der im Kernkomplex 3210(j) beinhaltet ist, zugreifen. In mindestens einer Ausführungsform kann der L3-Cache 3230 ohne Einschränkung eine beliebige Anzahl von Slices beinhalten.

[0231] In mindestens einer Ausführungsform ist die Struktur 3260 eine Systemverbindung, die Daten- und Steuerungsübertragungen über die Kernkomplexe 3210(1)-3210(N) (wobei N eine ganze Zahl größer als Null ist), E/A-Schnittstellen 3270 und Speichersteuerungen 3280 ermöglicht. In mindestens einer Ausführungsform kann die CPU 3200 ohne Einschränkung jede Menge und Art von Systemverbindung zusätzlich zu oder anstelle der Struktur 3260 umfassen, die Daten- und Steuerübertragungen über eine beliebige Anzahl und Art von direkt oder indirekt verbundenen Komponenten ermöglicht, die intern oder extern zu CPU 3200 sein können. In mindestens einer Ausführungsform repräsentieren die E/A-Schnittstellen 3270 eine beliebige Anzahl und Art von E/A-Schnittstellen (z. B. PCI, PCI-X, PCIe, GBE, USB usw.). In mindestens einer Ausführungsform sind verschiedene Arten von Peripheriegeräten mit E/A-Schnittstellen 3270 gekoppelt. In mindestens einer Ausführungsform können Peripheriegeräte, die mit den E/A-Schnittstellen 3270 gekoppelt sind, ohne Einschränkung Anzeigen, Tastaturen, Mäuse, Drucker, Scanner, Joysticks oder andere Arten von Spielesteuerungen, Medienaufzeichnungsvorrichtungen, externe Speichervorrichtungen, Netzwerkkarten usw. beinhalten.

[0232] In mindestens einer Ausführungsform ermöglichen die Speichersteuerungen 3280 Datenübertragungen zwischen der CPU 3200 und einem Systemspeicher 3290. In mindestens einer Ausführungsform nutzen der Kernkomplex 3210 und der Grafikkomplex 3240 den Systemspeicher 3290 gemeinsam. In mindestens einer Ausführungsform implementiert die CPU 3200 ein Speicherteilsystem, das ohne Einschränkung eine beliebige Menge und Art von Speichersteuerungen 3280 und Speichervorrichtungen beinhaltet, die einer Komponente zugeordnet oder von mehreren Komponenten gemeinsam genutzt werden können. In mindestens einer Ausführungsform implementiert die CPU 3200 ein Cache-Teilsystem, das ohne Einschränkung einen oder mehrere Cache-Speicher (z. B. L2-Caches 3228 und L3-Cachex 3230) beinhaltet, die jeweils für eine beliebige Anzahl von Komponenten reserviert sein können oder von diesen gemeinsam genutzt werden können (z. B. Kerne 3220 und Kernkomplexe 3210).

[0233] Fig. 33 zeigt ein beispielhaftes Beschleunigerintegrations-Slice 3390 gemäß mindestens einer Ausführungsform. Im hierin verwendeten Sinne umfasst ein „Slice“ einen spezifizierten Teil von Verarbeitungsressourcen der Beschleunigerintegrationsschaltung. In mindestens einer Ausführungsform stellt eine Beschleunigerintegrationsschaltung Cache-Verwaltungs-, Speicherzugriffs-, Kontextverwaltungs- und Unterbrechungsverwaltungsdienste im Namen einer Vielzahl von Grafikverarbeitungs-Engines, die in einem Grafikbeschleunigungsmodul beinhaltet ist, bereit. Die Grafikverarbeitungs-Engines können jeweils eine separate GPU umfassen. Alternativ können die Grafikverarbeitungsengines verschiedene Arten von Grafikverarbeitungs-Engines innerhalb einer GPU umfassen, wie etwa Grafikausführungseinheiten, Medienverarbeitungs-Engines (z. B. Videokodierer/- dekodierer), Abtaster und Blit-Engines. In mindestens einer Ausführungsform kann ein Grafikbeschleunigungsmodul eine GPU mit mehreren Grafikverarbeitungs-Engines sein. In mindestens einer Ausführungsform können Grafikverarbeitungs-Engines einzelne GPUs sein, die auf einem gemeinsamen Package, einer Linecard oder einem Chip integriert sind.

[0234] Ein effektiver Adressbereich 3382 der Anwendung innerhalb eines Systemspeichers 3314 speichert Prozesselemente 3383. In einer Ausführungsform werden die Prozesselemente 3383 als Reaktion auf GPU-Aufrufe 3381 von Anwendungen 3380 gespeichert, die auf dem Prozessor 3307 ausgeführt werden. Ein Prozesselement 3383 enthält einen Prozesszustand für die entsprechende Anwendung 3380. Ein im Prozesselement 3383 enthaltener Arbeitsdeskriptor (work descriptor - WD) 3384 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 3384 ein Zeiger auf eine Aufgabenanforderungswarteschlange im effektiven Adressbereich 3382 der Anwendung.

[0235] Das Grafikbeschleunigungsmodul 3346 und/oder einzelne Grafikverarbeitungs-Engines 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 3384 an ein Grafikbeschleunigungsmodul 3346 zum Starten einer Aufgabe in einer virtualisierten Umgebung beinhaltet sein.

[0236] In mindestens einer Ausführungsform ist ein Programmiermodell mit dediziertem Prozess implementationsspezifisch. In diesem Modell besitzt ein einzelner Prozess das Grafikbeschleunigungsmodul 3346 oder eine einzelne Grafikverarbeitungs-Engine. Da das Grafikbeschleunigungsmodul 3346 einem einzelnen Prozess gehört, initialisiert ein Hypervisor die Beschleunigerintegrationsschaltung für eine besitzende Partition und ein Betriebssystem initialisiert die Beschleunigerintegrationsschaltung für einen besitzenden Prozess, wenn das Grafikbeschleunigungsmodul 3346 zugewiesen ist.

[0237] Im Betrieb ruft eine WD-Abrufeinheit 3391 in einem Beschleunigerintegrations-Slice 3390 den nächsten WD 3384 ab, der eine Angabe der Arbeit beinhaltet, die durch eine oder mehrere Grafikverarbeitungs-Engines des Grafikbeschleunigungsmoduls 3346 zu erledigen ist. Daten vom WD 3384 können in den Registern 3345 gespeichert und von der Speicherverwaltungseinheit („MMU“) 3339, der Unterbrechungsverwaltungsschaltung 3347 und/oder der Kontextverwaltungsschaltung 3348 wie dargestellt verwendet werden. Eine Ausführungsform der MMU 3339 beinhaltet zum Beispiel eine Segment- /Seitenlaufschaltung zum Zugreifen auf Segment-/Seitentabellen 3386 innerhalb des virtuellen Adressraums 3385 eines OS. Die Unterbrechungsverwaltungsschaltung 3347 kann Unterbrechungsereignisse („INT“) 3392 verarbeiten, die von dem Grafikbeschleunigungsmodul 3346 empfangen wurden. Beim Durchführen von Grafikoperationen wird eine effektive Adresse 3393, die durch eine Grafikverarbeitungs-Engine erzeugt wird, durch die MMU 3339 in eine reale Adresse übersetzt.

[0238] In einer Ausführungsform wird ein gleicher Satz von Registern 3345 für jede Grafikverarbeitungs-Engine und/oder jedes Grafikbeschleunigungsmodul 3346 dupliziert und kann durch einen Hypervisor oder ein Betriebssystem initialisiert werden. Jedes dieser duplizierten Register kann in dem Beschleunigerintegrations-Slice 3390 beinhaltet sein. Beispielhafte Register, die durch einen Hypervisor initialisiert werden können, sind in Tabelle 1 gezeigt.

Tabelle 1 - durch Hypervisor initialisierte Register

1	Scheibensteuerregister
2	Bereichszeiger für geplante Prozesse für reale Adressen (RA)
3	Autoritätsmasken-Überschreibungsregister
4	Unterbrechungsvektor-Tabelleneintragsabweichung

5	Unterbrechungsvektor-Tabelleneintragsbegrenzung
6	Statusregister
7	ID der logischen Partition
8	Hypervisor-Beschleunigernutzungsaufzeichnungszeiger für reale Adressen (RA)
9	Speicherbeschreibungsregister

[0239] Beispielhafte Register, die von einem Betriebssystem initialisiert werden können, sind in Tabelle 2 gezeigt.

Tabelle 2- durch Betriebssystem initialisierte Register

1	Prozess- und Thread-Identifikation
2	Kontext-Sicherungs-/-Wiederherstellungszeiger für effektive Adressen (EA)
3	Beschleunigernutzungsaufzeichnungszeiger für virtuelle Adressen (VA)
4	Speichersegmenttabellenzeiger für virtuelle Adressen (VA)
5	Autoritätsmaske
6	Arbeitsdeskriptor

[0240] In einer Ausführungsform ist jeder WD 3384 spezifisch für ein konkretes Grafikbeschleunigungsmodul 3346 und/oder eine konkrete Grafikverarbeitungs-Engine. Er enthält alle Informationen, die von einer Grafikverarbeitungs-Engine benötigt werden, um Arbeit zu erledigen, oder er kann ein Zeiger auf einen Speicherort sein, an dem eine Anwendung eine Befehlswarteschlange mit zu erledigender Arbeit eingerichtet hat.

[0241] Fig. 34A-34B zeigen beispielhafte Grafikprozessoren gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform kann ein beliebiger der beispielhaften Grafikprozessoren unter Verwendung eines oder mehrerer IP-Kerne hergestellt sein. Zusätzlich zu dem, was veranschaulicht ist, können andere Logik und Schaltungen in mindestens einer Ausführungsform enthalten sein, einschließlich zusätzlicher Grafikprozessoren/-kerne, Peripherieschnittstellensteuerungen oder Universalprozessorkerne. In mindestens einer Ausführungsform dienen die beispielhaften Grafikprozessoren zur Verwendung in einem SoC.

[0242] Fig. 34A zeigt einen beispielhaften Grafikprozessor 3410 einer integrierten Schaltung eines SoC, die unter Verwendung eines oder mehrerer IP-Kerne gemäß mindestens einer Ausführungsform hergestellt sein kann. **Fig. 34B** veranschaulicht einen zusätzlichen beispielhaften Grafikprozessor 3440 einer integrierten Schaltung eines SoC, die unter Verwendung eines oder mehrerer IP-Kerne gemäß mindestens einer Ausführungsform hergestellt sein kann. In mindestens einer Ausführungsform ist der Grafikprozessor 3410 aus **Fig. 34A** ein Grafikprozessorkern mit niedriger Leistung. In mindestens einer Ausführungsform ist der Grafikprozessor 3440 aus **Fig. 34B** ein Grafikprozessorkern mit höherer Leistung. In mindestens einer Ausführungsform kann jeder der Grafikprozessoren 3410, 3440 eine Variante des Grafikprozessors 510 aus **Fig. 5** sein.

[0243] In mindestens einer Ausführungsform beinhaltet der Grafikprozessor 3410 einen Vertex-Prozessor 3405 und einen oder mehrere Fragmentprozessoren 3415A-3415N (z. B. 3415A, 3415B, 3415C, 3415D bis 3415N-1 und 3415N). In mindestens einer Ausführungsform kann der Grafikprozessor 3410 derartig unterschiedliche Shader-Programme über getrennte Logik ausführen, dass der Vertex-Prozessor 3405 optimiert ist, um Operationen für Vertex-Shader-Programme auszuführen, während ein oder mehrere Fragmentprozessoren 3415A-3415N Shading-Operationen für Fragmente (z. B. Pixel) für Fragment- oder Pixel-Shader-Programme ausführen. In mindestens einer Ausführungsform führt der Vertex-Prozessor 3405 eine Vertex-Verarbeitungsstufe einer 3D-Grafikpipeline durch und erzeugt Primitive und Vertex- Daten. In mindestens einer Ausführungsform verwenden ein oder mehrere Fragmentprozessoren 3415A-3415N Primitiv- und Vertex-Daten, die vom Vertex-Prozessor 3405 erzeugt werden, um einen Bildspeicher zu erzeugen, der auf einer Anzeigevorrichtung angezeigt wird. In mindestens einer Ausführungsform sind ein oder mehrere Fragmentprozessoren 3415A-3415N optimiert, um Fragment-Shader-Programme auszuführen, wie in einer OpenGL-API bereitgestellt, die verwendet werden können, um ähnliche Operationen wie ein Pixel-Shader-Programm durchzuführen, wie es in einer Direct 3D-API vorgesehen ist.

[0244] In mindestens einer Ausführungsform beinhaltet der Grafikprozessor 3410 eine oder mehrere MMU 3420A-3420B, Cache 3425A-3425B und Schaltungszwischenverbindungen 3430A-3430B. In mindestens einer Ausführungsform stellen eine oder mehrere MMU 3420A-3420B eine virtuellphysische Adresszuordnung für den Grafikprozessor 3410 bereit, was für den Vertex-Prozessor 3405 und/oder Fragmentprozessor(en) 3415A-3415N beinhaltet, die sich auf Vertex- oder im Speicher gespeicherte Bild-/Texturdaten zusätzlich zu in einem oder mehreren Cache 3425A-3425B gespeicherten Vertex- oder Bild-/Texturdaten beziehen. In mindestens einer Ausführungsform können eine oder mehrere MMU 3420A-3420B mit anderen MMU innerhalb des Systems synchronisiert werden, was beinhaltet, dass eine oder mehrere MMU einem oder mehreren Anwendungsprozessoren 505, Bildprozessoren 515 und/oder Videoprozessoren 520 aus **Fig. 5** zugeordnet sind, sodass jeder Prozessor 505-520 an einem gemeinsam genutzten oder vereinheitlichten virtuellen Speichersystem teilnehmen kann. In mindestens einer Ausführungsform ermöglichen eine oder mehrere Schaltungszwischenverbindungen 3430A-3430B dem Grafikprozessor 3410, mit anderen IP-Kernen innerhalb eines SoC Schnittstellen zu bilden, entweder über einen internen Bus eines SoC oder über eine direkte Verbindung.

[0245] In mindestens einer Ausführungsform beinhaltet der Grafikprozessor 3440 eine oder mehrere MMU 3420A-3420B, Cache 3425A-3425B und Schaltungszwischenverbindungen 3430A-3430B des Grafikprozessors 3410 aus **Fig. 34A**. In mindestens einer Ausführungsform beinhaltet der Grafikprozessor 3440 einen oder mehrere Shader-Kerne 3455A-3455N (z. B. 3455A, 3455B, 3455C, 3455D, 3455E, 3455F bis 3455N-1 und 3455N), was eine vereinheitlichte Shader-Kernarchitektur bereitstellt, in der ein(e) einzelne(r) Kern oder Typ oder Kern alle Arten von programmierbarem Shader-Code ausführen kann, was Shader-Programmcode beinhaltet, um Vertex-Shader, Fragment-Shader und/oder Compute-Shader zu implementieren. In mindestens einer Ausführungsform kann eine Anzahl der Shader-Kerne variieren. In mindestens einer Ausführungsform beinhaltet der Grafikprozessor 3440 einen Interkern-Taskmanager 3445, der als Thread-Dispatcher fungiert, um Ausführungs-Threads an einen oder mehrere Shader-Kerne 3455A-3455N abzusenden, und eine Kachelungseinheit 3458, um Kachelungsoperationen für kachelbasiertes Rendering zu beschleunigen, bei dem Rendering-Operationen für eine Szene im Bildraum unterteilt werden, um in mindestens einer Ausführungsform lokale räumliche Kohärenz innerhalb einer Szene auszunutzen oder um die Nutzung interner Caches zu optimieren.

[0246] **Fig. 35A** zeigt einen Grafikkern 3500 gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform kann der Grafikkern 3500 in dem Grafikprozessor 2410 aus **Fig. 24** beinhaltet sein. In mindestens einer Ausführungsform kann der Grafikkern 3500 ein vereinheitlichter Shader-Kern 3455A-3455N wie in **Fig. 34B** sein. In mindestens einer Ausführungsform beinhaltet der Grafikkern 3500 einen gemeinsam genutzten Anweisungscache 3502, eine Textureinheit 3518 und einen Cache/gemeinsam genutzten Speicher 3520, die Ausführungsressourcen innerhalb des Grafikkerns 3500 gemein sind. In mindestens einer Ausführungsform kann der Grafikkern 3500 mehrere Slices 3501A-3501 N oder eine Partition für jeden Kern beinhalten, und kann ein Grafikprozessor mehrere Instanzen des Grafikkerns 3500 beinhalten. Die Slices 3501A-3501 N können eine Unterstützungslogik beinhalten, die einen lokalen Anweisungscache 3504A-3504N, einen Thread-Scheduler 3506A-3506N, einen Thread-Dispatcher 3508A-3508N und einen Satz von Registern 3510A-3510N beinhaltet. In mindestens einer Ausführungsform können die Slices 3501A-3501 N einen Satz zusätzlicher Funktionseinheiten (additional function unit - „AFU“) 3512A-3512N, Gleitkommaeinheiten (floating-point units - „FPU“) 3514A-3514N, ganzzahlige arithmetisch-logische Einheiten („ALU“) 3516-3516N, Adressberechnungseinheiten (address computational unit - „ACU“) 3513A-3513N, Gleitkommaeinheiten mit doppelter Genauigkeit (double-precision floating-point unit - „DPFPU“) 3515A-3515N und Matrixverarbeitungseinheiten (matrix processing unit - „MPU“) 3517A-3517N beinhalten.

[0247] In mindestens einer Ausführungsform können die FPUs 3514A-3514N Gleitkommaoperationen mit einfacher Genauigkeit (32 Bit) und mit halber Genauigkeit (16 Bit) ausführen, während die DPFPUs 3515A-3515N Gleitkommaoperationen mit doppelter Genauigkeit (64 Bit) ausführen. In mindestens einer Ausführungsform können die ALUs 3516A-3516N 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 3517A-3517N 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 3517A-3517N 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 AFU 3512A-3512N zusätzliche logische Operationen ausführen, die nicht durch Gleitkomma- oder Ganzzahleinheiten unterstützt werden, einschließlich trigonometrischer Operationen (z. B. Sinus, Cosinus usw.).

[0248] Fig. 35B zeigt eine Allzweck-Grafikverarbeitungseinheit (general-purpose graphics processing unit - „GPGPU“) 3530 gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform ist die GPGPU 3530 hochparallel und für den Einsatz auf einem Multi-Chip-Modul geeignet. In mindestens einer Ausführungsform kann die GPGPU 3530 konfiguriert sein, um die Durchführung hochparalleler Rechenoperationen durch ein Array von GPU zu ermöglichen. In mindestens einer Ausführungsform kann die GPGPU 3530 direkt mit anderen Instanzen der GPGPU 3530 verbunden sein, um einen Multi-GPU-Cluster zu erzeugen, um die Ausführungszeit für CUDA-Programme zu verbessern. In mindestens einer Ausführungsform beinhaltet die GPGPU 3530 eine Host- Schnittstelle 3532, um eine Verbindung mit einem Host-Prozessor zu ermöglichen. In mindestens einer Ausführungsform ist die Host-Schnittstelle 3532 eine PCIe-Schnittstelle. In mindestens einer Ausführungsform kann die Host-Schnittstelle 3532 eine herstellerspezifische Kommunikationsschnittstelle oder Kommunikationsstruktur sein. In mindestens einer Ausführungsform empfängt die GPGPU 3530 Befehle von einem Host-Prozessor und verwendet einen globalen Scheduler 3534, um Ausführungs-Threads, die diesen Befehlen zugeordnet sind, an einen Satz von Rechenclustern 3536A-3536H zu verteilen. In mindestens einer Ausführungsform nutzen die Rechencluster 3536A-3536H einen Cache-Speicher 3538 gemeinsam. In mindestens einer Ausführungsform kann der Cache-Speicher 3538 als ein Cache höherer Ebene für Cache-Speicher innerhalb der Rechencluster 3536A-3536H dienen.

[0249] In mindestens einer Ausführungsform beinhaltet die GPGPU 3530 Speicher 3544A-3544B, der über einen Satz von Speichersteuerungen 3542A-3542B mit den Rechenclustern 3536A-3536H gekoppelt ist. In mindestens einer Ausführungsform kann der Speicher 3544A-3544B verschiedene Arten von Speichervorrichtungen beinhalten, die DRAM oder Grafik-Direktzugriffsspeicher, wie etwa synchronen Grafik-Direktzugriffsspeicher („SGRAM“) beinhalten, der Speicher mit Grafik-Doppeldatenraten („GDDR“) beinhaltet.

[0250] In mindestens einer Ausführungsform beinhalten die Rechencluster 3536A-3536H jeweils einen Satz von Grafikkernen, wie etwa den Grafikkern 3500 aus **Fig. 35A**, der mehrere Arten von Ganzzahl- und Gleitkomma-Logikeinheiten beinhalten kann, die Rechenoperationen mit einer Reihe von Genauigkeiten durchführen können, darunter solche, die für Berechnungen in Verbindung mit CUDA-Programmen geeignet sind. Zum Beispiel kann in mindestens einer Ausführungsform mindestens eine Teilmenge von Gleitkommaeinheiten in jedem der Rechencluster 3536A-3536H konfiguriert sein, um 16-Bit- oder 32-Bit-Gleitkommaoperationen durchzuführen, während eine andere Teilmenge von Gleitkommaeinheiten konfiguriert sein kann, um 64-Bit-Gleitkommaoperationen durchzuführen.

[0251] In mindestens einer Ausführungsform können mehrere Instanzen der GPGPU 3530 so konfiguriert sein, dass sie als Rechencluster arbeiten. In mindestens einer Ausführungsform können die Rechencluster 3536A - 3536H alle technisch machbaren Kommunikationstechniken für die Synchronisation und den Datenaustausch implementieren. In mindestens einer Ausführungsform kommunizieren mehrere Instanzen der GPGPU 3530 über die Host-Schnittstelle 3532. In mindestens einer Ausführungsform beinhaltet die GPGPU 3530 einen E/A-Hub 3539, der die GPGPU 3530 mit einem GPU-Link 3540 koppelt, die eine direkte Verbindung zu anderen Instanzen der GPGPU 3530 ermöglicht. In mindestens einer Ausführungsform ist der GPU-Link 3540 mit einer dedizierten GPU-zu-GPU-Brücke gekoppelt, welche die Kommunikation und Synchronisation zwischen mehreren Instanzen der GPGPU 3530 ermöglicht. In mindestens einer Ausführungsform ist der GPU-Link 3540 mit einer Hochgeschwindigkeitsverbindung gekoppelt, um Daten an andere GPGPUs 3530 oder Parallelprozessoren zu übermitteln und zu empfangen. In mindestens einer Ausführungsform befinden sich mehrere Instanzen der GPGPU 3530 in getrennten Datenverarbeitungssystemen und kommunizieren über eine Netzwerkvorrichtung, auf die über die Hostschnittstelle 3532 zugegriffen werden kann. In mindestens einer Ausführungsform kann der GPU-Link 3540 so konfiguriert sein, dass eine Verbindung zu einem Host-Prozessor zusätzlich zu oder alternativ zu der Host-Schnittstelle 3532 ermöglicht wird. In mindestens einer Ausführungsform kann die GPGPU 3530 konfiguriert sein, um ein CUDA-Programm auszuführen.

[0252] Fig. 36A zeigt einen Parallelprozessor 3600 gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform können verschiedene Komponenten des Parallelprozessors 3600 unter Verwendung einer oder mehrerer integrierter Schaltungsvorrichtungen implementiert sein, wie etwa programmierbare Prozessoren, anwendungsspezifische integrierte Schaltungen („ASIC“) oder feldprogrammierbare Gatearrays (FPGA).

[0253] In mindestens einer Ausführungsform beinhaltet der Parallelprozessor 3600 eine Parallelverarbeitungseinheit 3602. In mindestens einer Ausführungsform beinhaltet die Parallelverarbeitungseinheit 3602 eine E/A-Einheit 3604, die Kommunikation mit anderen Vorrichtungen ermöglicht, darunter mit anderen Instanzen der Parallelverarbeitungseinheit 3602. In mindestens einer Ausführungsform kann die E/A-Einheit 3604 direkt mit anderen Vorrichtungen verbunden sein. In mindestens einer Ausführungsform stellt die E/A-

Einheit 3604 eine Verbindung mit anderen Vorrichtungen über die Verwendung einer Hub- oder Switch-Schnittstelle, wie etwa dem Speicher-Hub 605 her. In mindestens einer Ausführungsform bilden Verbindungen zwischen dem Speicher-Hub 605 und der E/A-Einheit 3604 eine Kommunikationsverbindung. In mindestens einer Ausführungsform ist die E/A-Einheit 3604 mit einer Host-Schnittstelle 3606 und einem Crossbar-Speicher 3616 verbunden, wobei die Host-Schnittstelle 3606 Befehle empfängt, die auf das Durchführen von Verarbeitungsoperationen gerichtet sind, und der Crossbar-Speicher 3616 Befehle empfängt, die auf das Durchführen von Speicheroperationen gerichtet sind.

[0254] In mindestens einer Ausführungsform kann die Host-Schnittstelle 3606 Arbeitsoperationen zum Durchführen dieser Befehle an ein Frontend 3608 richten, wenn die Host-Schnittstelle 3606 einen Befehlspuffer über die E/A-Einheit 3604 empfängt. In mindestens einer Ausführungsform ist das Frontend 3608 mit einem Scheduler 3610 gekoppelt, der so konfiguriert ist, dass er Befehle oder andere Arbeitselemente an ein Verarbeitungsarray 3612 verteilt. In mindestens einer Ausführungsform stellt der Scheduler 3610 sicher, dass das Verarbeitungsarray 3612 richtig konfiguriert ist und sich in einem gültigen Zustand befindet, bevor Aufgaben an das Verarbeitungsarray 3612 verteilt werden. In mindestens einer Ausführungsform ist der Scheduler 3610 über Firmware-Logik implementiert, die auf einem Mikrocontroller ausgeführt wird. In mindestens einer Ausführungsform ist der durch eine Mikrosteuerung implementierte Scheduler 3610 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 3612 ausgeführt werden. In mindestens einer Ausführungsform kann die Host-Software Arbeitslasten zum Planen auf dem Verarbeitungsarray 3612 über eine von mehreren Doorbells für die Grafikverarbeitung nachweisen. In mindestens einer Ausführungsform können Arbeitslasten dann durch die Logik des Schedulers 3610 innerhalb einer Mikrosteuerung, die den Scheduler 3610 beinhaltet, automatisch über das Verarbeitungsarray 3612 verteilt werden.

[0255] In mindestens einer Ausführungsform kann das Verarbeitungsarray 3612 bis zu „N“ Cluster beinhalten (z. B. Cluster 3614A, Cluster 3614B bis Cluster 3614N). In mindestens einer Ausführungsform kann jeder Cluster 3614A-3614N des Verarbeitungsarrays 3612 eine große Anzahl gleichzeitiger Threads ausführen. In mindestens einer Ausführungsform kann der Scheduler 3610 den Clustern 3614A-3614N des Verarbeitungsarrays 3612 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 Scheduler 3610 gehandhabt werden oder kann teilweise durch die Compilerlogik während der Kompilierung der Programmlogik unterstützt werden, die für die Ausführung durch das Verarbeitungsarray 3612 konfiguriert ist. In mindestens einer Ausführungsform können unterschiedliche Cluster 3614A-3614N des Verarbeitungsarrays 3612 zum Verarbeiten unterschiedlicher Arten von Programmen oder zum Durchführen unterschiedlicher Arten von Berechnungen zugewiesen werden.

[0256] In mindestens einer Ausführungsform kann das Verarbeitungsarray 3612 konfiguriert sein, um verschiedene Arten von Parallelverarbeitungsoperationen durchzuführen. In mindestens einer Ausführungsform ist das Verarbeitungsarray 3612 konfiguriert, um Allzweck-Parallelrechenoperationen durchzuführen. In mindestens einer Ausführungsform kann das Verarbeitungsarray 3612 Logik zum Ausführen von Verarbeitungstasks beinhalten, was das Filtern von Video- und/oder Audiodaten, das Durchführen von Modellierungsoperationen, einschließlich Physikoperationen, und das Durchführen von Datentransformationen beinhaltet.

[0257] In mindestens einer Ausführungsform ist das Verarbeitungsarray 3612 konfiguriert, um parallele Grafikverarbeitungsoperationen durchzuführen. In mindestens einer Ausführungsform kann das Verarbeitungsarray 3612 zusätzliche Logik beinhalten, um die Ausführung derartiger Grafikverarbeitungsoperationen zu unterstützen, darunter unter anderem Texturabtastlogik, um Texturoperationen durchzuführen, sowie Tesselierungslogik und andere Vertexverarbeitungslogik, ohne darauf beschränkt zu sein. In mindestens einer Ausführungsform kann das Verarbeitungsarray 3612 konfiguriert sein, um grafikverarbeitungsbezogene Shader-Programme auszuführen, wie etwa Vertex-Shader, Tessellation-Shader, Geometrie-Shader und Pixel-Shader, ohne darauf beschränkt zu sein. In mindestens einer Ausführungsform kann die Parallelverarbeitungseinheit 3602 Daten aus dem Systemspeicher über die E/A-Einheit 3604 zur Verarbeitung übermitteln. In mindestens einer Ausführungsform können während der Verarbeitung die übermittelten Daten in einem chipinternen Speicher (z. B. dem Parallelprozessorspeicher 3622) während der Verarbeitung gespeichert und dann wieder in den Systemspeicher geschrieben werden.

[0258] In mindestens einer Ausführungsform kann, wenn die Parallelverarbeitungseinheit 3602 verwendet wird, um eine Grafikverarbeitung durchzuführen, der Scheduler 3610 konfiguriert sein, um eine Verarbei-

tungsarbeitslast in ungefähr gleich große Tasks aufzuteilen, um eine bessere Verteilung von Grafikverarbeitungsoperationen auf mehrere Cluster 3614A-3614N des Verarbeitungsarrays 3612 zu ermöglichen. In mindestens einer Ausführungsform können Teile des Verarbeitungsarrays 3612 konfiguriert sein, um unterschiedliche Verarbeitungsarten durchzuführen. Zum Beispiel kann in mindestens einer Ausführungsform ein erster Abschnitt so konfiguriert sein, dass er Vertex-Shading und Topologieerzeugung durchführt, ein zweiter Abschnitt kann so konfiguriert sein, dass er Tesselierungs- und Geometrie-Shading durchführt, und ein dritter Abschnitt kann so konfiguriert sein, dass er Pixel-Shading oder andere Bildschirmraumoperationen durchführt, um ein gerendertes Bild zur Anzeige zu erzeugen. In mindestens einer Ausführungsform können Zwischendaten, die durch einen oder mehrere der Cluster 3614A-3614N erzeugt werden, in Puffern gespeichert werden, um zu ermöglichen, dass Zwischendaten zwischen den Clustern 3614A-3614N zur weiteren Verarbeitung übertragen werden.

[0259] In mindestens einer Ausführungsform kann das Verarbeitungsarray 3612 auszuführende Verarbeitungstasks über den Scheduler 3610 empfangen, der Verarbeitungstasks definierende Befehle vom Frontend 3608 empfängt. In mindestens einer Ausführungsform können die Verarbeitungs-Tasks Indizes der zu verarbeitenden Daten beinhalten, z. B. Oberflächen-(Patch-)Daten, Primitivdaten, Vertexdaten und/oder Pixeldaten, sowie Zustandsparameter und Befehle, die definieren, wie die Daten verarbeitet werden sollen (z. B. welches Programm ausgeführt werden soll). In mindestens einer Ausführungsform kann der Scheduler 3610 konfiguriert sein, um Indizes abzurufen, die Tasks entsprechen, oder kann Indizes von dem Frontend 3608 empfangen. In mindestens einer Ausführungsform kann das Frontend 3608 konfiguriert sein, um sicherzustellen, dass das Verarbeitungsarray 3612 in einen gültigen Zustand konfiguriert ist, bevor eine durch eingehende Befehlspuffer (z. B. Stapelpuffer, Push-Puffer usw.) spezifizierte Arbeitslast initiiert wird.

[0260] In mindestens einer Ausführungsform kann jede von einer oder mehreren Instanzen der Parallelverarbeitungseinheit 3602 mit dem Parallelprozessorspeicher 3622 gekoppelt sein. In mindestens einer Ausführungsform kann auf den Parallelprozessorspeicher 3622 über den Crossbar-Speicher 3616 zugegriffen werden, die Speicheranforderungen von dem Verarbeitungsarray 3612 sowie von der E/A-Einheit 3604 empfangen kann. In mindestens einer Ausführungsform kann die Crossbar-Speicher 3616 über eine Speicherschnittstelle 3618 auf den Parallelprozessorspeicher 3622 zugreifen. In mindestens einer Ausführungsform kann die Speicherschnittstelle 3618 mehrere Partitionseinheiten (z. B. Partitionseinheit 3620A, Partitionseinheit 3620B bis Partitionseinheit 3620N) beinhalten, die jeweils mit einem Abschnitt (z. B. Speichereinheit) des Parallelprozessorspeichers 3622 gekoppelt werden können. In mindestens einer Ausführungsform ist eine Anzahl von Partitionseinheiten 3620A-3620N dazu konfiguriert, derartig gleich einer Anzahl von Speichereinheiten zu sein, dass eine erste Partitionseinheit 3620A eine entsprechende erste Speichereinheit 3624A aufweist, eine zweite Partitionseinheit 3620B eine entsprechende Speichereinheit 3624B aufweist und eine N-te Partitionseinheit 3620N eine entsprechende N-te Speichereinheit 3624N aufweist. In mindestens einer Ausführungsform ist eine Anzahl von Partitionseinheiten 3620A-3620N möglicherweise nicht gleich einer Anzahl von Speichervorrichtungen.

[0261] In mindestens einer Ausführungsform können die Speichereinheiten 3624A-3624N verschiedene Arten von Speichervorrichtungen beinhalten, die DRAM oder Grafik-Direktzugriffsspeicher, wie etwa SGRAM beinhalten, der GDDR-Speicher beinhaltet. In mindestens einer Ausführungsform können die Speichereinheiten 3624A-3624N auch 3D-Stapelspeicher beinhalten, einschließlich, aber nicht beschränkt auf, Speicher mit hoher Bandbreite (HBM). In mindestens einer Ausführungsform können Rendering-Ziele, wie etwa Bildspeicher oder Texturkarten, über die Speichereinheiten 3624A-3624N hinweg gespeichert werden, was es den Partitionseinheiten 3620A-3620N ermöglicht, Abschnitte jedes Rendering-Ziels parallel zu schreiben, um die verfügbare Bandbreite des Parallelprozessorspeichers 3622 effizient zu nutzen. In mindestens einer Ausführungsform kann eine lokale Instanz des Parallelprozessorspeichers 3622 zugunsten einer einheitlichen Speicherausgestaltung ausgeschlossen werden, die Systemspeicher in Verbindung mit lokalem Cache-Speicher nutzt.

[0262] In mindestens einer Ausführungsform kann jeder der Cluster 3614A-3614N des Verarbeitungsarrays 3612 Daten verarbeiten, die in eine der Speichereinheiten 3624A-3624N innerhalb des Parallelprozessorspeichers 3622 geschrieben werden. In mindestens einer Ausführungsform kann der Crossbar-Speicher 3616 so konfiguriert sein, dass er eine Ausgabe jedes Clusters 3614A-3614N an eine beliebige Partitionseinheit 3620A-3620N oder an einen anderen Cluster 3614A-3614N übermittelt, der zusätzliche Verarbeitungsoperationen an einer Ausgabe ausführen kann. In mindestens einer Ausführungsform kann jeder Cluster 3614A-3614N durch den Crossbar-Speicher 3616 mit der Speicherschnittstelle 3618 kommunizieren, um aus verschiedenen externen Speichervorrichtungen zu lesen oder in diese zu schreiben. In mindestens einer Ausführungsform weist der Crossbar-Speicher 3616 eine Verbindung mit der Speicherschnittstelle 3618 auf,

um mit der E/A-Einheit 3604 zu kommunizieren, sowie eine Verbindung mit einer lokalen Instanz des Parallelprozessorspeichers 3622, was es den Verarbeitungseinheiten innerhalb der unterschiedlichen Cluster 3614A-3614N ermöglicht, mit Systemspeicher oder anderem Speicher zu kommunizieren, der nicht lokal zur Parallelverarbeitungseinheit 3602 ist. In mindestens einer Ausführungsform kann der Crossbar-Speicher 3616 virtuelle Kanäle verwenden, um Verkehrsströme zwischen Clustern 3614A-3614N und Partitionseinheiten 3620A-3620N zu trennen.

[0263] In mindestens einer Ausführungsform können mehrere Instanzen der Parallelverarbeitungseinheit 3602 auf einer einzelnen Erweiterungskarte bereitgestellt sein oder mehrere Erweiterungskarten zusammengeschaltet sein. In mindestens einer Ausführungsform können unterschiedliche Instanzen der Parallelverarbeitungseinheit 3602 zum Zusammenarbeiten konfiguriert sein, 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 3602 Gleitkommaeinheiten mit höherer Genauigkeit in Bezug auf andere Instanzen beinhalten. In mindestens einer Ausführungsform können Systeme, die eine oder mehrere Instanzen der Parallelverarbeitungseinheit 3602 oder des Parallelprozessors 3600 einschließen, in einer Reihe von Konfigurationen und Formfaktoren implementiert sein, was Desktop-, Laptop- oder Handheld-Personalcomputer, Server, Workstations, Spielekonsolen und/oder eingebettete Systeme beinhaltet, ohne darauf beschränkt zu sein.

[0264] Fig. 36B veranschaulicht einen Verarbeitungs-Cluster 3694 gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform ist der Verarbeitungs-Cluster 3694 in einer Parallelverarbeitungseinheit beinhaltet. In mindestens einer Ausführungsform ist der Verarbeitungs-Cluster 3694 einer der Verarbeitungs-Cluster 3614A-3614N aus **Fig. 36**. In mindestens einer Ausführungsform kann der Verarbeitungs-Cluster 3694 konfiguriert sein, um viele Threads parallel auszuführen, wobei sich „Thread“ auf eine Instanz eines konkreten Programms bezieht, das mit einem konkreten Satz von Eingabedaten ausgeführt wird. In mindestens einer Ausführungsform werden Ausgabetechniken für Single-Instruction-Multiple-Data-(SIMD-) Anweisungen verwendet, um die parallele Ausführung einer großen Anzahl von Threads zu unterstützen, ohne mehrere unabhängige Anweisungseinheiten bereitzustellen. In mindestens einer Ausführungsform werden Single-Instruction-Multiple-Thread-(SIMT-)Techniken verwendet, um die parallele Ausführung einer großen Anzahl von im Allgemeinen synchronisierten Threads zu unterstützen, wobei eine gemeinsame Anweisungseinheit verwendet wird, die so konfiguriert ist, dass sie Anweisungen an einen Satz von Verarbeitungs-Engines innerhalb jedes der Verarbeitungs-Cluster 3694 ausgibt.

[0265] In mindestens einer Ausführungsform kann der Betrieb des Verarbeitungs-Clusters 3694 über einen Pipeline-Manager 3632 gesteuert werden, der Verarbeitungstasks an SIMT-Parallelprozessoren verteilt. In mindestens einer Ausführungsform empfängt der Pipeline-Manager 3632 Anweisungen vom Scheduler 3610 aus **Fig. 36** und verwaltet die Ausführung dieser Anweisungen über einen Grafikmultiprozessor 3634 und/oder eine Textureinheit 3636. In mindestens einer Ausführungsform ist der Grafikmultiprozessor 3634 eine beispielhafte Instanz eines SIMT-Parallelprozessors. In mindestens einer Ausführungsform können jedoch verschiedene Typen von SIMT-Parallelprozessoren mit unterschiedlichen Architekturen innerhalb des Verarbeitungs-Clusters 3694 enthalten sein. In mindestens einer Ausführungsform können eine oder mehrere Instanzen des Grafikmultiprozessors 3634 innerhalb eines Verarbeitungs-Clusters 3694 enthalten sein. In mindestens einer Ausführungsform kann der Grafikmultiprozessor 3634 Daten verarbeiten und eine Datenkreuzschiene 3640 kann verwendet werden, um verarbeitete Daten an eines von mehreren möglichen Zielen, darunter andere Shader-Einheiten, zu verteilen. In mindestens einer Ausführungsform kann der Pipeline-Manager 3632 die Verteilung von verarbeiteten Daten erleichtern, indem er Ziele für zu verteilende verarbeitete Daten über die Datenkreuzschiene 3640 vorgibt.

[0266] In mindestens einer Ausführungsform kann jeder Grafikmultiprozessor 3634 innerhalb des Verarbeitungs-Clusters 3694 einen identischen Satz von Funktionsausführungslogik (z. B. arithmetische Logikeinheiten, Lade-SpeicherEinheiten („LSU“) usw.) beinhalten. In mindestens einer Ausführungsform kann funktionelle Ausführungslogik pipelineartig konfiguriert sein, wobei neue Anweisungen ausgegeben werden können, bevor vorherige Anweisungen abgeschlossen sind. In mindestens einer Ausführungsform unterstützt die funktionelle Ausführungslogik eine Vielfalt von Vorgängen, darunter Integer- und Fließkommaarithmetik, Vergleichsvorgänge, boolesche Vorgänge, Bitverschiebung und Berechnung verschiedener algebraischer Funktionen. In mindestens einer Ausführungsform kann dieselbe Hardware einer funktionellen Einheit ausgenutzt werden, um unterschiedliche Operationen auszuführen, und eine beliebige Kombination von funktionellen Einheiten kann vorhanden sein.

[0267] In mindestens einer Ausführungsform bilden die an den Verarbeitungscluster 3694 übertragenen Anweisungen einen Thread. In mindestens einer Ausführungsform ist ein Satz von Threads, der über einen Satz von Parallelverarbeitungs-Engines ausgeführt wird, eine Thread-Gruppe. In mindestens einer Ausführungsform führt die Thread-Gruppe ein Programm anhand von unterschiedlichen Eingabedaten aus. In mindestens einer Ausführungsform kann jeder Thread innerhalb einer Thread-Gruppe einer anderen Verarbeitungs-Engine innerhalb eines Grafikmultiprozessors 3634 zugeordnet sein. In mindestens einer Ausführungsform kann eine Thread-Gruppe weniger Threads beinhalten als eine Anzahl von Verarbeitungs-Engines innerhalb des Grafikmultiprozessors 3634. Wenn eine Thread-Gruppe weniger Threads beinhaltet als eine Anzahl von Verarbeitungs-Engines, können in mindestens einer Ausführungsform eine oder mehrere der Verarbeitungs-Engines während der Zyklen, in denen diese Thread-Gruppe verarbeitet wird, inaktiv sein. In mindestens einer Ausführungsform kann eine Thread-Gruppe auch mehr Threads beinhalten als eine Anzahl von Verarbeitungs-Engines innerhalb des Grafikmultiprozessors 3634. Wenn eine Thread-Gruppe mehr Threads als eine Anzahl von Verarbeitungs-Engines innerhalb des Grafikmultiprozessors 3634 beinhaltet, kann das Verarbeiten in mindestens einer Ausführungsform über aufeinanderfolgende Taktzyklen durchgeführt werden. In mindestens einer Ausführungsform kann eine Thread-Gruppe auch mehr Threads beinhalten als eine Anzahl von Verarbeitungs-Engines innerhalb des Grafikmultiprozessors 3634.

[0268] In mindestens einer Ausführungsform beinhaltet der Grafikmultiprozessor 3634 einen internen Cache-Speicher, um Lade- und Speicheroperationen durchzuführen. In mindestens einer Ausführungsform kann der Grafikmultiprozessor 3634 auf einen internen Cache verzichten und einen Cache-Speicher (z. B. L1-Cache 3648) innerhalb des Verarbeitungs-Clusters 3694 verwenden. In mindestens einer Ausführungsform hat jeder Grafikmultiprozessor 3634 auch Zugriff auf Ebene-2-Caches („L2“) innerhalb von Partitionseinheiten (z. B. die Partitionseinheiten 3620A-3620N aus **Fig. 36A**), die von allen Verarbeitungs-Clustern 3694 gemeinsam genutzt werden, und kann verwendet werden, um Daten zwischen Threads zu übertragen. In mindestens einer Ausführungsform kann der Grafikmultiprozessor 3634 auch auf den chipexternen globalen Speicher zugreifen, der einen oder mehrere von dem lokalen Parallelprozessorspeicher und/oder dem Systemspeicher beinhalten kann. In mindestens einer Ausführungsform kann ein beliebiger Speicher, der extern zur Parallelverarbeitungseinheit 3602 ist, als globaler Speicher verwendet werden. In mindestens einer Ausführungsform beinhaltet der Verarbeitungs-Cluster 3694 mehrere Instanzen des Grafikmultiprozessors 3634, der gemeinsame Anweisungen und Daten gemeinsam nutzen kann, die im L1-Zwischenspeicher 3648 gespeichert sein können.

[0269] In mindestens einer Ausführungsform kann jeder Verarbeitungs-Cluster 3694 eine MMU 3645 beinhalten, die dazu konfiguriert ist, virtuelle Adressen physischen Adressen zuzuordnen. In mindestens einer Ausführungsform können sich eine oder mehrere Instanzen der MMU 3645 innerhalb der Speicherschnittstelle 3618 aus **Fig. 36** befinden. In mindestens einer Ausführungsform beinhaltet die MMU 3645 einen Satz von Seitentabelleneinträgen (page table entries - PTE), die verwendet werden, um eine virtuelle Adresse einer physischen Adresse einer Kachel und optional einem Cachezeilenindex zuzuordnen. In mindestens einer Ausführungsform kann die MMU 3645 Adressübersetzungspuffer (TLB) oder Cache-Speicher beinhalten, die sich innerhalb des Grafikmultiprozessors 3634 oder des L1-Cache 3648 oder des Verarbeitungs-Clusters 3694 befinden können. In mindestens einer Ausführungsform wird eine physische Adresse verarbeitet, um den Oberflächendatenzugriffsstandort zu verteilen, um eine effiziente Anforderungsverschachtelung zwischen Partitionseinheiten zu ermöglichen. In mindestens einer Ausführungsform kann ein Cache-Zeilenindex verwendet werden, um zu bestimmen, ob eine Anforderung für eine Cache-Zeile ein Treffer oder ein Fehler ist.

[0270] In mindestens einer Ausführungsform kann der Verarbeitungs-Cluster 3694 konfiguriert sein, sodass jeder Grafikmultiprozessor 3634 an eine Textureinheit 3636 zum Durchführen von Texturzuordnungsoperationen gekoppelt ist, z. B. zum Bestimmen von Texturabtastpositionen, Lesen von Texturdaten und Filtern von Texturdaten. In mindestens einer Ausführungsform werden die Texturdaten aus einem internen Textur-L1-Cache (nicht dargestellt) oder aus einem L1-Cache innerhalb des Grafikmultiprozessors 3634 gelesen und je nach Bedarf aus einem L2-Cache, dem lokalen Parallelprozessorspeicher oder dem Systemspeicher abgerufen. In mindestens einer Ausführungsform gibt jeder Grafikmultiprozessor 3634 verarbeitete Tasks an die Daten-Crossbar 3640 aus, um verarbeitete Aufgaben einem anderen Verarbeitungs-Cluster 3694 zur weiteren Verarbeitung bereitzustellen oder eine verarbeitete Task über die Speicher-Crossbar 3616 in einem L2-Cache, einem lokalen Parallelprozessorspeicher oder einem Systemspeicher zu speichern. In mindestens einer Ausführungsform ist eine Vorrasteroperationseinheit (pre-raster operations unit - „preROP“) 3642 konfiguriert, um Daten von dem Grafikmultiprozessor 3634 zu empfangen, um Daten an ROP-Einheiten zu leiten, die sich bei Partitionseinheiten befinden können, wie hierin beschrieben (z. B. die Partitionseinheiten

3620A-3620N aus **Fig. 36**). In mindestens einer Ausführungsform kann die PreROP 3642 Optimierungen für die Farbmischung durchführen, Pixelfarbdaten organisieren und Adressübersetzungen durchführen.

[0271] Fig. 36C zeigt einen Grafikmultiprozessor 3696 gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform ist der Grafikmultiprozessor 3696 der Grafikmultiprozessor 3634 aus **Fig. 36B**. In mindestens einer Ausführungsform ist der Grafikmultiprozessor 3696 mit dem Pipeline-Manager 3632 des Verarbeitungs-Clusters 3694 gekoppelt. In mindestens einer Ausführungsform weist der Grafikmultiprozessor 3696 eine Ausführungspipeline auf, die einen Anweisungs-Cache 3652, eine Anweisungseinheit 3654, eine Adressabbildungseinheit 3656, eine Registerdatei 3658, einen oder mehrere Kerne 3662 einer Universal-Grafikverarbeitungseinheit (GPGPU) und eine oder mehrere Lade-/Speichereinheiten 3666 beinhaltet, ohne darauf beschränkt zu sein. Ein oder mehrere GPGPU-Kerne 3662 und Lade-/Speichereinheiten 3666 sind mit dem schnellen Cache-Speicher 3672 und dem gemeinsam genutzten Speicher 3670 über eine Speicher- und Cache-Verbindung 3668 gekoppelt.

[0272] In mindestens einer Ausführungsform empfängt der Anweisungs-Cache 3652 einen Stream von auszuführenden Anweisungen vom Pipeline-Manager 3632. In mindestens einer Ausführungsform werden Anweisungen im Anweisungs-Cache 3652 zwischengespeichert und zur Ausführung durch die Anweisungseinheit 3654 gesendet. In mindestens einer Ausführungsform kann die Anweisungseinheit 3654 Anweisungen als Thread-Gruppen (z. B. Warps) zuteilen, wobei jeder Thread der Thread-Gruppe einer anderen Ausführungseinheit innerhalb der GPGPU-Kerne 3662 zugeordnet wird. In mindestens einer Ausführungsform kann eine Anweisung auf einen beliebigen von einem lokalen, gemeinsam genutzten oder globalen Adressraum zugreifen, indem sie eine Adresse innerhalb eines einheitlichen Adressraums vorgibt. In mindestens einer Ausführungsform kann die Adressabbildungseinheit 3656 verwendet werden, um Adressen in einem einheitlichen Adressraum in eine eindeutige Speicheradresse zu übersetzen, auf die durch die Lade-/Speichereinheiten 3666 zugegriffen werden kann.

[0273] In mindestens einer Ausführungsform stellt die Registerdatei 3658 einen Satz von Registern für funktionelle Einheiten des Grafikmultiprozessors 3696 bereit. In mindestens einer Ausführungsform stellt die Registerdatei 3658 einen temporären Datenspeicher für Operanden bereit, die mit Datenpfaden von funktionellen Einheiten (z. B. GPGPU-Kernen 3662, Lade-/Speichereinheiten 3666) des Grafikmultiprozessors 3696 verbunden sind. In mindestens einer Ausführungsform ist die Registerdatei 3658 derart zwischen den einzelnen funktionellen Einheiten aufgeteilt, dass jeder funktionellen Einheit ein dedizierter Abschnitt der Registerdatei 3658 zugewiesen ist. In mindestens einer Ausführungsform ist die Registerdatei 3658 auf unterschiedliche Thread-Gruppen aufgeteilt, die durch den Grafikmultiprozessor 3696 ausgeführt werden.

[0274] In mindestens einer Ausführungsform können die GPGPU-Kerne 3662 jeweils FPU und/oder Ganzzahl-ALU beinhalten, die verwendet werden, um Anweisungen des Grafikmultiprozessors 3696 auszuführen. GPGPU-Kerne 3662 können die gleiche Architektur aufweisen oder sich in der Architektur unterscheiden. In mindestens einer Ausführungsform beinhaltet ein erster Abschnitt der GPGPU-Kerne 3662 eine FPU mit einfacher Genauigkeit und eine Integer-ALU, während ein zweiter Abschnitt der GPGPU-Kerne 3662 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 3696 zusätzlich eine oder mehrere Festfunktions- oder Spezialfunktionseinheiten beinhalten, um spezifische Funktionen, wie etwa Operationen zum Kopieren von Rechtecken oder Pixel-Blending, durchzuführen. In mindestens einer Ausführungsform können einer oder mehrere der GPGPU-Kerne 3662 auch feste oder Spezialfunktionslogik beinhalten.

[0275] In mindestens einer Ausführungsform beinhalten die GPGPU-Kerne 3662 SIMD-Logik, die dazu in der Lage ist, eine einzelne Anweisung an mehreren Datensätzen durchzuführen. In mindestens einer Ausführungsform können die GPGPU-Kerne 3662 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 3662 zur Kompilierzeit durch einen Shader-Compiler oder automatisch erzeugt werden, wenn Programme ausgeführt werden, die für Single-Program-Multiple-Data-(SPMD-) oder SIMT-Architekturen geschrieben und kompiliert wurden. In mindestens einer Ausführungsform können mehrere Threads eines für ein SIMT-Ausführungsmodell konfigurierten Programms über eine einzelne SIMD-Anweisung ausgeführt werden. In mindestens einer Ausführungsform können in mindestens einer Ausführungsform acht SIMT-Threads, die gleiche oder ähnliche Operationen durchführen, parallel über eine einzelne SIMD8-Logikeinheit ausgeführt werden.

[0276] In mindestens einer Ausführungsform ist die Speicher- und Cache-Verbindung 3668 ein Verbindungsnetz, das jede funktionelle Einheit des Grafikmultiprozessors 3696 mit der Registerdatei 3658 und dem gemeinsam genutzten Speicher 3670 verbindet. In mindestens einer Ausführungsform ist die Speicher- und Cache-Verbindung 3668 eine Kreuzschienen-Verbindung, die es der Lade-/Speichereinheit 3666 ermöglicht, Lade- und Speicheroperationen zwischen dem gemeinsam genutzten Speicher 3670 und der Registerdatei 3658 zu implementieren. In mindestens einer Ausführungsform kann die Registerdatei 3658 mit derselben Frequenz wie die GPGPU-Kerne 3662 arbeiten, wodurch die Datenübertragung zwischen den GPGPU-Kernen 3662 und der Registerdatei 3658 eine sehr geringe Latenzzeit hat. In mindestens einer Ausführungsform kann der gemeinsam genutzte Speicher 3670 verwendet werden, um die Kommunikation zwischen Threads zu ermöglichen, die auf funktionellen Einheiten innerhalb des Grafikmultiprozessors 3696 ausgeführt werden. In mindestens einer Ausführungsform kann der Cache-Speicher 3672 zum Beispiel als Daten-Cache verwendet werden, um Texturdaten zwischenzuspeichern, die zwischen funktionellen Einheiten und der Textureinheit 3636 kommuniziert werden. In mindestens einer Ausführungsform kann der gemeinsam genutzte Speicher 3670 auch als programmverwalteter Cache verwendet werden. In mindestens einer Ausführungsform können Threads, die auf den GPGPU-Kernen 3662 ausgeführt werden, zusätzlich zu den automatisch zwischengespeicherten Daten, die innerhalb des Cache-Speichers 3672 gespeichert sind, programmatisch Daten innerhalb des gemeinsam genutzten Speichers speichern.

[0277] In mindestens einer Ausführungsform ist ein Parallelprozessor oder eine GPGPU, wie hierin beschrieben, kommunikativ an Host-/Prozessorkerne gekoppelt, um Grafikoperationen, Operationen des maschinellen Lernens, Musteranalyseoperationen und verschiedene Funktionen einer Universal-GPU (GPGPU) zu beschleunigen. In mindestens einer Ausführungsform kann eine GPU über einen Bus oder eine andere Verbindung (z. B. eine Hochgeschwindigkeitsverbindung wie etwa PCIe oder NVLink) mit dem Host-Prozessor/-den Kernen kommunikativ gekoppelt sein. In mindestens einer Ausführungsform kann eine GPU in einem Gehäuse oder Chip als Kerne integriert sein und über einen internen Prozessorbus/eine Verbindung, die intern zu einem Gehäuse oder Chip ist, mit Kernen kommunikativ gekoppelt sein. In mindestens einer Ausführungsform können die Prozessorkerne unabhängig von einer Weise, auf welche eine GPU verbunden ist, einer derartigen GPU Arbeit in Form von Sequenzen von Befehlen/Anweisungen zuweisen, die in einem WD enthalten sind. In mindestens einer Ausführungsform verwendet eine GPU dann eine dedizierte Schaltung/-Logik zum effizienten Verarbeiten dieser Befehle/Anweisungen.

ALLGEMEINE DATENVERARBEITUNG

[0278] Die folgenden Figuren legen ohne Einschränkung beispielhafte Softwarekonstrukte dar, die zum Implementieren mindestens einer Ausführungsform verwendet werden können.

[0279] Fig. 37 zeigt einen Softwarestapel einer Programmierplattform gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform ist eine Programmierplattform eine Plattform zum Einsetzen von Hardware auf einem Computersystem, um Berechnungstasks zu beschleunigen. In mindestens einer Ausführungsform kann eine Programmierplattform für Softwareentwickler durch Bibliotheken, Compilerdirektiven und/oder Erweiterungen von Programmiersprachen zugänglich sein. In mindestens einer Ausführungsform kann eine Programmierplattform CUDA, Radeon Open Compute Platform („ROCm“), OpenCL (OpenCL™ wird von der Khronos-Gruppe entwickelt), SYCL oder Intel One API sein, ohne darauf beschränkt zu sein.

[0280] In mindestens einer Ausführungsform stellt ein Softwarestapel 3700 einer Programmierplattform eine Ausführungsumgebung für eine Anwendung 3701 bereit. In mindestens einer Ausführungsform kann die Anwendung 3701 eine beliebige Computersoftware beinhalten, die auf dem Softwarestapel 3700 gestartet werden kann. In mindestens einer Ausführungsform kann die Anwendung 3701 eine Anwendung für künstliche Intelligenz („AI“)/Maschinenlernen („ML“), eine Hochleistungsrechenanwendung („HPC“), eine virtuelle Desktop-Infrastruktur umfassen, ist aber nicht darauf beschränkt („VDI“) oder eine Rechenzentrums-Arbeitslast.

[0281] In mindestens einer Ausführungsform laufen die Anwendung 3701 und der Softwarestapel 3700 auf Hardware 3707. Die Hardware 3707 kann in mindestens einer Ausführungsform eine oder mehrere GPUs, CPUs, FPGAs, KI-Engines und/oder andere Arten von Rechenvorrichtungen umfassen, die eine Programmierplattform unterstützen. In mindestens einer Ausführungsform, wie beispielsweise bei CUDA, kann der Softwarestapel 3700 herstellerspezifisch und nur mit Geräten von bestimmten Anbietern kompatibel sein. In mindestens einer Ausführungsform, wie etwa bei OpenCL, kann der Softwarestapel 3700 mit Vorrichtungen von verschiedenen Herstellern verwendet werden. In mindestens einer Ausführungsform beinhaltet die Hard-

ware 3707 einen Host, der mit einem oder mehreren Vorrichtungen verbunden ist, auf die zugegriffen werden kann, um Berechnungstasks über Aufrufe einer Anwendungsprogrammierschnittstelle (application programming interface - „API“) auszuführen. Eine Vorrichtung innerhalb der Hardware 3707 kann, ohne darauf beschränkt zu sein, eine GPU, ein FPGA, eine KI-Engine oder eine andere Rechenvorrichtung (aber kann auch eine CPU beinhalten) und seinen Speicher beinhalten, im Gegensatz zu einem Host innerhalb der Hardware 3707, der in mindestens einer Ausführungsform eine CPU (aber auch eine Rechenvorrichtung) und ihren Speicher beinhalten, ohne darauf beschränkt zu sein.

[0282] In mindestens einer Ausführungsform beinhaltet der Softwarestapel 3700 einer Programmierplattform ohne Einschränkung eine Anzahl von Bibliotheken 3703, eine Laufzeit 3705 und einen Vorrichtungs-Kerneltreiber 3706. Jede der Bibliotheken 3703 kann in mindestens einer Ausführungsform Daten und Programmiercode beinhalten, die durch Computerprogramme verwendet und während der Softwareentwicklung eingesetzt werden können. In mindestens einer Ausführungsform können die Bibliotheken 3703 vorgefertigten Code und Unterroutinen, Klassen, Werte, Typspezifikationen, Konfigurationsdaten, Dokumentation, Hilfedaten und/oder Nachrichtenvorlagen umfassen, sind aber nicht darauf beschränkt. In mindestens einer Ausführungsform beinhalten die Bibliotheken 3703 Funktionen, die für die Ausführung auf einer oder mehreren Arten von Vorrichtungen optimiert sind. In mindestens einer Ausführungsform können die Bibliotheken 3703 Funktionen zum Durchführen von mathematischen Operationen, tiefem Lernen und/oder anderen Arten von Operationen auf Vorrichtungen beinhalten, ohne darauf beschränkt zu sein. In mindestens einer Ausführungsform sind Bibliotheken 3803 entsprechenden APIs 3802 zugeordnet, die eine oder mehrere APIs beinhalten können, die in Bibliotheken 3803 implementierte Funktionen offenlegen.

[0283] In mindestens einer Ausführungsform ist die Anwendung 3701 als Quellcode geschrieben, der in ausführbaren Code kompiliert wird, wie nachstehend in Verbindung mit **Fig. 42** ausführlicher erörtert wird. In mindestens einer Ausführungsform kann der ausführbare Code der Anwendung 3701 mindestens teilweise auf einer Ausführungsumgebung laufen, die vom Softwarestapel 3700 bereitgestellt wird. In mindestens einer Ausführungsform kann während der Ausführung der Anwendung 3701 Code erreicht werden, der auf einer Vorrichtung im Gegensatz zu einem Host laufen muss. In einem solchen Fall kann die Laufzeit 3705 in mindestens einer Ausführungsform aufgerufen werden, um erforderlichen Code auf eine Vorrichtung zu laden und zu starten. In mindestens einer Ausführungsform kann die Laufzeit 3705 ein beliebiges technisch machbares Laufzeitsystem beinhalten, das die Ausführung der Anwendung S01 unterstützen kann.

[0284] In mindestens einer Ausführungsform ist die Laufzeit 3705 als eine oder mehrere Laufzeitbibliotheken implementiert, die mit entsprechenden APIs verknüpft sind, die als API(s) 3704 gezeigt sind. Eine oder mehrere solcher Laufzeitbibliotheken können in mindestens einer Ausführungsform unter anderem Funktionen für Speicherverwaltung, Ausführungssteuerung, Geräteverwaltung, Fehlerbehandlung und/oder Synchronisation umfassen. In mindestens einer Ausführungsform können die Speicherverwaltungsfunktionen Funktionen zum Zuweisen, Freigeben und Kopieren von Vorrichtungsspeicher sowie zum Übertragen von Daten zwischen Host-Speicher und Vorrichtungsspeicher beinhalten, ohne darauf beschränkt zu sein. In mindestens einer Ausführungsform können die Ausführungssteuerfunktionen Funktionen zum Starten einer Funktion (manchmal als „Kernel“ bezeichnet, wenn eine Funktion eine von einem Host aufrufbare globale Funktion ist) auf einer Vorrichtung beinhalten und Attributwerte in einem Puffer setzen, der durch eine Laufzeitbibliothek für eine bestimmte Funktion gepflegt wird, die auf einer Vorrichtung ausgeführt werden soll, ohne darauf beschränkt zu sein.

[0285] Laufzeitbibliotheken und entsprechende API(s) 3704 können in mindestens einer Ausführungsform auf jede technisch machbare Weise implementiert werden. In mindestens einer Ausführungsform kann eine (oder eine beliebige Anzahl von) API einen Satz von Funktionen einer niedrigen Ebene für eine feingranulare Steuerung einer Vorrichtung bereitstellen, während eine andere (oder eine beliebige Anzahl von) API einen Satz solcher Funktionen einer höheren Ebene offenlegen kann. In mindestens einer Ausführungsform kann eine Laufzeit-API einer hohen Ebene auf einer API einer niedrigen Ebene aufgebaut sein. In mindestens einer Ausführungsform können eine oder mehrere Laufzeit-APIs sprachspezifische APIs sein, die über einer sprachunabhängigen Laufzeit-API geschichtet sind.

[0286] In mindestens einer Ausführungsform ist der Vorrichtungs-Kerneltreiber 3706 konfiguriert, um die Kommunikation mit einem zugrunde liegenden Gerät zu ermöglichen. In mindestens einer Ausführungsform kann der Vorrichtungs-Kerneltreiber 3706 Low-Level-Funktionalitäten bereitstellen, auf die sich APIs, wie beispielsweise API(s) 3704 und/oder andere Software stützen. In mindestens einer Ausführungsform kann der Vorrichtungs-Kerneltreiber 3706 konfiguriert sein, um Intermediate-Repräsentations-(„IR“)-Code zur Laufzeit in Binärcode zu kompilieren. Für CUDA kann der Vorrichtungs-Kerneltreiber 3706 in mindestens einer Aus-

führungsform IR-Code zur parallelen Thread-Ausführung (Parallel Thread Execution - „PTX“), der nicht hardwarespezifisch ist, zur Laufzeit in Binärcode für eine bestimmtes Zielvorrichtung kompilieren (mit Zwischenspeichern des kompilierten Binärcodes), was manchmal auch als „Finalisieren“ von Code bezeichnet wird. Dies kann in mindestens einer Ausführungsform zulassen, dass finalisierter Code auf einer Zielvorrichtung läuft, der möglicherweise nicht existiert hat, als der Quellcode ursprünglich in PTX-Code kompiliert wurde. Alternativ kann in mindestens einer Ausführungsform der Vorrichtungsquellcode offline in Binärcode kompiliert werden, ohne dass es erforderlich ist, dass der Vorrichtungskerneltreiber 3706 den IR-Code während der Laufzeit kompiliert.

[0287] Fig. 38 veranschaulicht eine CUDA-Implementierung des Softwarestapels 3700 der **Fig. 37** gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform beinhaltet ein CUDA-Softwarestapel 3800, auf dem eine Anwendung 3801 gestartet werden kann, CUDA-Bibliotheken 3803, eine CUDA-Laufzeit 3805, einen CUDA-Treiber 3807 und einen Vorrichtungskerneltreiber 3808. In mindestens einer Ausführungsform wird der CUDA-Softwarestapel 3800 auf Hardware 3809 ausgeführt, die eine GPU beinhalten kann, die CUDA unterstützt und von NVIDIA Corporation, Santa Clara, Kalifornien entwickelt wird.

[0288] In mindestens einer Ausführungsform können die Anwendung 3801, die CUDA-Laufzeit 3805 und der Vorrichtungs-Kerneltreiber 3808 ähnliche Funktionalitäten wie die Anwendung 3701, die Laufzeit 3705 bzw. der Vorrichtungs-Kerneltreiber 3706 durchführen, die vorstehend in Verbindung mit **Fig. 37** beschrieben wurden. In mindestens einer Ausführungsform beinhaltet der CUDA-Treiber 3807 eine Bibliothek (libcuda.so), die eine CUDA-Treiber-API 3806 implementiert. Ähnlich einer CUDA-Laufzeit-API 3804, die durch eine CUDA-Laufzeitbibliothek (cudart) implementiert wird, kann die CUDA-Treiber-API 3806 in mindestens einer Ausführungsform ohne Einschränkung unter anderem Funktionen für Speicherverwaltung, Ausführungssteuerung, Vorrichtungsverwaltung, Fehlerbehandlung, Synchronisation und/oder Grafikinteroperabilität offenlegen. In mindestens einer Ausführungsform unterscheidet sich die CUDA-Treiber-API 3806 von der CUDA-Laufzeit-API 3804 darin, dass die CUDA-Laufzeit-API 3804 die Vorrichtungscodeverwaltung vereinfacht, indem sie implizite Initialisierung, Kontextverwaltung (analog zu einem Prozess) und Modulverwaltung (analog zu dynamisch geladenen Bibliotheken) bereitstellt. Im Gegensatz zur CUDA-Laufzeit-API 3804 auf hoher Ebene ist die CUDA-Treiber-API 3806 eine API auf niedriger Ebene, die in mindestens einer Ausführungsform eine feinkörnigere Steuerung einer Vorrichtung bereitstellt, insbesondere in Bezug auf Kontexte und Modulladen. In mindestens einer Ausführungsform kann die CUDA-Treiber-API 3806 Funktionen für die Kontextverwaltung offenlegen, die nicht durch die CUDA-Laufzeit-API 3804 offengelegt werden. In mindestens einer Ausführungsform ist die CUDA-Treiber-API 3806 auch sprachunabhängig und unterstützt z. B. OpenCL zusätzlich zur CUDA-Laufzeit-API 3804. Ferner können in mindestens einer Ausführungsform Entwicklungsbibliotheken, darunter die CUDA-Laufzeit 3805, als von Treiberkomponenten getrennt betrachtet werden, darunter der Benutzermodus-CUDA-Treiber 3807 und der Kernelmodus-Gerätetreibers 3808 (manchmal auch als „Anzeige“-Treiber bezeichnet).

[0289] In mindestens einer Ausführungsform können die CUDA-Bibliotheken 3803 mathematische Bibliotheken, Deep-Learning-Bibliotheken, parallele Algorithmusbibliotheken und/oder Signal-/Bild-/Videoverarbeitungsbibliotheken beinhalten, die parallele Rechenanwendungen wie die Anwendung 3801 nutzen können, ohne darauf beschränkt zu sein. In mindestens einer Ausführungsform können die CUDA-Bibliotheken 3803 mathematische Bibliotheken umfassen, wie etwa eine cuBLAS-Bibliothek, die eine Implementierung von Basic Linear Algebra Subprograms („BLAS“) zum Durchführen linearer Algebra-Operationen ist, eine cuFFT-Bibliothek zum Berechnen schneller Fourier-Transformationen („FFTs“) und unter anderem eine cuRAND-Bibliothek zur Generierung von Zufallszahlen. In mindestens einer Ausführungsform können die CUDA-Bibliotheken 3803 Deep-Learning-Bibliotheken beinhalten, wie etwa unter anderem eine cuDNN-Bibliothek von Primitiven für tiefe neuronale Netze und eine TensorRT-Plattform für Hochleistungs-Deep-Learning-Inferenz.

[0290] Fig. 39 veranschaulicht eine ROCm-Implementierung des Softwarestapels 3700 der **Fig. 37** gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform beinhaltet ein ROCm-Softwarestapel 3900, auf dem eine Anwendung 3901 gestartet werden kann, eine Sprachlaufzeit 3903, eine Systemlaufzeit 3905, einen Thunk 3907, einen ROCm-Kerneltreiber 3908 und einen Vorrichtungskerneltreiber 3909. In mindestens einer Ausführungsform wird der ROCm-Softwarestapel 3900 auf Hardware 3910 ausgeführt, die eine GPU beinhalten kann, die ROCm unterstützt und von AMD Corporation, Santa Clara, Kalifornien entwickelt wird.

[0291] In mindestens einer Ausführungsform kann die Anwendung 3901 ähnliche Funktionalitäten wie die Anwendung 3701 durchführen, die vorstehend in Verbindung mit **Fig. 37** erörtert wurde. Außerdem können die Sprachlaufzeit 3903 und die Systemlaufzeit 3905 in mindestens einer Ausführungsform ähnliche Funktio-

nalitäten wie die vorstehend in Verbindung mit **Fig. 37** erörterte Laufzeit 3705 durchführen. In mindestens einer Ausführungsform unterscheiden sich die Sprachlaufzeit 3903 und die Systemlaufzeit 3905 darin, dass die Systemlaufzeit 3905 eine sprachunabhängige Laufzeit ist, die eine ROCr-Systemlaufzeit-API 3904 implementiert und eine Laufzeit-API mit heterogener Systemarchitektur (Heterogeneous System Architecture - „HAS“) verwendet. Die HAS-Laufzeit-API ist in mindestens einer Ausführungsform eine schlanke Benutzermodus-API, die Schnittstellen für den Zugriff auf und die Interaktion mit einer AMD-GPU offenlegt, die unter anderem Funktionen für die Speicherverwaltung, die Ausführungssteuerung über den architekturdefinierten Versand von Kerneln, die Fehlerbehandlung, System- und Agenteninformationen sowie die Initialisierung und das Herunterfahren der Laufzeit beinhaltet. Im Gegensatz zur Systemlaufzeit 3905 ist die Sprachlaufzeit 3903 in mindestens einer Ausführungsform eine Implementierung einer sprachspezifischen Laufzeit-API 3902, die auf der ROCr-Systemlaufzeit-API 3904 geschichtet ist. In mindestens einer Ausführungsform kann die Sprachlaufzeit-API unter anderem eine Sprachlaufzeit-API für Heterogeneous Compute Interface for Portability („HIP“), eine Sprachlaufzeit-API für Heterogeneous Compute Compiler („HCC“) oder eine OpenCL-API beinhalten, ohne darauf beschränkt zu sein. Die HIP-Sprache ist insbesondere eine Erweiterung der Programmiersprache C++ mit funktional ähnlichen Versionen von CUDA-Mechanismen, und in mindestens einer Ausführungsform beinhaltet eine HIP-Sprachen-Laufzeit-API Funktionen, die denen der CUDA-Laufzeit-API 3804 ähneln, die vorstehend in Verbindung mit **Fig. 38** erörtert wurde, wie etwa unter anderem Funktionen zur Speicherverwaltung, Ausführungssteuerung, Vorrichtungsverwaltung, Fehlerbehandlung und Synchronisation.

[0292] In mindestens einer Ausführungsform ist der Thunk (ROCt) 3907 eine Schnittstelle, die verwendet werden kann, um mit dem zugrundeliegenden ROCm-Treiber 3908 zu interagieren. In mindestens einer Ausführungsform ist der ROCm-Treiber 3908 ein ROCk-Treiber, der eine Kombination aus einem AMDGPU-Treiber und einem HAS-Kernel-Treiber (amdkfd) ist. In mindestens einer Ausführungsform ist der AMDGPU-Treiber ein von AMD entwickelter Vorrichtungskerneltreiber für GPU, der ähnliche Funktionen wie der Vorrichtungskerneltreiber 3706 durchführt, der vorstehend in Verbindung mit **Fig. 37** erörtert wurde. In mindestens einer Ausführungsform ist der HAS-Kernel-Treiber ein Treiber, der es unterschiedlichen Arten von Prozessoren erlaubt, Systemressourcen effektiver über Hardware-Merkmale gemeinsam zu nutzen.

[0293] In mindestens einer Ausführungsform können verschiedene Bibliotheken (nicht dargestellt) im ROCm-Softwarestapel 3900 oberhalb der Sprachlaufzeit 3903 beinhaltet sein und eine Ähnlichkeit der Funktionalität zu den CUDA-Bibliotheken 3803 bereitstellen, die vorstehend in Verbindung mit **Fig. 38** erörtert wurden. In mindestens einer Ausführungsform können verschiedene Bibliotheken mathematische, Deep-Learning- und/oder andere Bibliotheken beinhalten, wie etwa unter anderem eine hipBLAS-Bibliothek, die Funktionen ähnlich denen von CUDA cuBLAS implementiert, eine rocFFT-Bibliothek zum Berechnen von FFT, die ähnlich CUDA cuFFT ist.

[0294] **Fig. 40** veranschaulicht eine OpenCL-Implementierung des Softwarestapels 3700 der **Fig. 37** gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform beinhaltet ein OpenCL-Softwarestapel 4000, auf dem eine Anwendung 4001 gestartet werden kann, ein OpenCL-Framework 4005, eine OpenCL-Laufzeit 4006 und einen Treiber 4007. In mindestens einer Ausführungsform wird der OpenCL-Softwarestapel 4000 auf Hardware 3809 ausgeführt, die nicht herstellerspezifisch ist. Da OpenCL durch Vorrichtungen unterstützt wird, die von verschiedenen Herstellern entwickelt wurden, können in mindestens einer Ausführungsform spezifische OpenCL-Treiber erforderlich sein, um mit Hardware von solchen Herstellern zusammenzuwirken.

[0295] In mindestens einer Ausführungsform können die Anwendung 4001, die OpenCL-Laufzeit 4006, der Vorrichtungs-Kerneltreiber 4007 und die Hardware 4008 ähnliche Funktionalitäten wie die Anwendung 3701, die Laufzeit 3705, der Vorrichtungs-Kerneltreiber 3706 bzw. die Hardware 3707 durchführen, die vorstehend in Verbindung mit **Fig. 37** beschrieben wurden. In mindestens einer Ausführungsform beinhaltet die Anwendung 4001 ferner einen OpenCL-Kernel 4002 mit Code, der auf einer Vorrichtung ausgeführt werden soll.

[0296] In mindestens einer Ausführungsform definiert OpenCL eine „Plattform“, die es einem Host erlaubt, mit einem Host verbundene Vorrichtungen zu steuern. In mindestens einer Ausführungsform stellt ein OpenCL-Framework eine Plattformschicht-API und eine Laufzeit-API bereit, die als Plattform-API 4003 und Laufzeit-API 4005 gezeigt sind. In mindestens einer Ausführungsform verwendet die Laufzeit-API 4005 Kontexte, um die Ausführung von Kerneln auf Vorrichtungen zu verwalten. In mindestens einer Ausführungsform kann jede identifizierte Vorrichtung einem jeweiligen Kontext zugeordnet sein, den die Laufzeit-API 4005 verwenden kann, für diese Vorrichtung um unter anderem Befehlswarteschlangen, Programmobjekte und Kernelobjekte zu verwalten und Speicherobjekte gemeinsam zu nutzen. In mindestens einer Ausführungsform

legt die Plattform-API 4003 Funktionen offen, die unter anderem die Verwendung von Vorrichtungskontexten zum Auswählen und Initialisieren von Vorrichtungen, zum Übermitteln von Arbeit an Vorrichtungen über Befehlswarteschlangen und zum Ermöglichen der Datenübertragung zu und von Vorrichtungen zulassen. Darüber hinaus stellt das OpenCL-Framework in mindestens einer Ausführungsform verschiedene integrierte Funktionen (nicht gezeigt) bereit, einschließlich unter anderem mathematischer Funktionen, relationaler Funktionen und Bildverarbeitungsfunktionen.

[0297] In mindestens einer Ausführungsform ist auch ein Compiler 4004 im OpenCL-Framework 4005 beinhaltet. Quellcode kann in mindestens einer Ausführungsform offline vor der Ausführung einer Anwendung oder online während der Ausführung einer Anwendung kompiliert werden. Im Gegensatz zu CUDA und ROCm können OpenCL-Anwendungen in mindestens einer Ausführungsform online durch den Compiler 4004 kompiliert werden, der beinhaltet ist, um repräsentativ für eine beliebige Anzahl von Compilern zu sein, die verwendet werden kann, um Quellcode und/oder IR-Code, wie etwa Portable Intermediate Representation („SPIR-V“)-Code, in Binärcode zu kompilieren. Alternativ können in mindestens einer Ausführungsform OpenCL-Anwendungen offline kompiliert werden, bevor solche Anwendungen ausgeführt werden.

[0298] **Fig. 41** zeigt Software, die durch eine Programmierplattform gemäß mindestens einer Ausführungsform unterstützt wird. In mindestens einer Ausführungsform ist eine Programmierplattform 4104 konfiguriert, um verschiedene Programmiermodelle 4103, Middlewares und/oder Bibliotheken 4102 und Frameworks 4101 zu unterstützen, auf die eine Anwendung 4100 zurückgreifen kann. In mindestens einer Ausführungsform kann die Anwendung 4100 eine KI/ML-Anwendung sein, die in mindestens einer Ausführungsform unter Verwendung eines Deep-Learning-Frameworks wie etwa MXNet, PyTorch oder TensorFlow implementiert ist, das auf Bibliotheken wie cuDNN, NVIDIA Collective Communications Library („NCCL“) und/oder NVIDA Developer Data Loading Library („DALI“) CUDA-Bibliotheken zurückgreifen kann, um beschleunigte Datenverarbeitung auf der zugrunde liegenden Hardware bereitzustellen.

[0299] In mindestens einer Ausführungsform kann die Programmierplattform 4104 eine der vorstehend in Verbindung mit **Fig. 33**, **Fig. 34** bzw. **Fig. 40** beschriebenen CUDA-, ROCm- oder OpenCL-Plattformen sein. In mindestens einer Ausführungsform unterstützt die Programmierplattform 4104 mehrere Programmiermodelle 4103, die Abstraktionen eines zugrunde liegenden Rechensystems sind, das Ausdrücke von Algorithmen und Datenstrukturen zulässt. Die Programmiermodelle 4103 können in mindestens einer Ausführungsform Elemente der zugrunde liegenden Hardware offenlegen, um die Leistungsfähigkeit zu verbessern. In mindestens einer Ausführungsform können die Programmiermodelle 4103 CUDA, HIP, OpenCL, C++ Accelerated Massive Parallelism („C++AMP“), Open Multi-Processing („OpenMP“), Open Accelerators („OpenACC“) und/oder Vulcan Compute beinhalten, ohne darauf beschränkt zu sein.

[0300] In mindestens einer Ausführungsform stellen Bibliotheken und/oder Middlewares 4102 Implementierungen von Abstraktionen von Programmiermodellen 4104 bereit. In mindestens einer Ausführungsform können derartige Bibliotheken Daten und Programmiercode beinhalten, die durch Computerprogramme verwendet und während der Softwareentwicklung eingesetzt werden können. In mindestens einer Ausführungsform beinhalten solche Middlewares Software, die Dienste für Anwendungen bereitstellt, die über diejenigen hinausgehen, die von der Programmierplattform 4104 verfügbar sind. In mindestens einer Ausführungsform können die Bibliotheken und/oder Middlewares 4102 cuBLAS-, cuFFT-, cuRAND- und andere CUDA-Bibliotheken oder rocBLAS-, rocFFT-, rocRAND- und andere ROCm-Bibliotheken beinhalten, ohne darauf beschränkt zu sein. Darüber hinaus können die Bibliotheken und/oder Middlewares 4102 in mindestens einer Ausführungsform NCCL- und ROCm Communication Collectives Library („RCCL“)-Bibliotheken umfassen, die Kommunikationsroutinen für GPUs bereitstellen, eine MIOpen-Bibliothek für Deep Learning-Beschleunigung und/oder eine Eigen-Bibliothek für lineare Algebra, Matrix- und Vektoroperationen, geometrische Transformationen, numerische Solver und verwandte Algorithmen.

[0301] In mindestens einer Ausführungsform hängen die Anwendungs-Frameworks 4101 von den Bibliotheken und/oder Middlewares 4102 ab. In mindestens einer Ausführungsform ist jedes der Anwendungs-Frameworks 4101 ein Software-Framework, das verwendet wird, um eine Standardstruktur von Anwendungssoftware zu implementieren. Eine KI/ML-Anwendung kann in mindestens einer Ausführungsform unter Verwendung eines Frameworks, wie etwa ein Caffe-, Caffe2-, TensorFlow-, Keras-, PyTorch- oder MxNet-Deep-Learning-Framework implementiert sein.

[0302] **Fig. 42** zeigt das Kompilieren von Code zum Ausführen auf einer der Programmierplattformen der **Fig. 37-40** gemäß mindestens einer Ausführungsform. In mindestens einer Ausführungsform empfängt ein Compiler 4201 Quellcode 4200, der sowohl Host-Code als auch Vorrichtungscode beinhaltet. In mindestens

einer Ausführungsform ist der Compiler 4201 konfiguriert, um den Quellcode 4200 in einen vom Host ausführbaren Code 4202 zur Ausführung auf einem Host und in einen von der Vorrichtung ausführbaren Code 4203 zur Ausführung auf einer Vorrichtung umzuwandeln. Der Quellcode 4200 kann in mindestens einer Ausführungsform entweder offline vor der Ausführung einer Anwendung oder online während der Ausführung einer Anwendung kompiliert werden.

[0303] In mindestens einer Ausführungsform kann der Quellcode 4200 Code in einer beliebigen Programmiersprache beinhalten, die durch den Compiler 4201 unterstützt wird, wie etwa C++, C, Fortran usw. In mindestens einer Ausführungsform kann der Quellcode 4200 in einer Einzelquelldatei mit einer Mischung aus Host-Code und Vorrichtungscode beinhaltet sein, wobei darin Stellen des Vorrichtungscodes angegeben sind. In mindestens einer Ausführungsform kann eine Einzelquelldatei eine .cu-Datei, die CUDA-Code beinhaltet, oder eine .hip.cpp-Datei, die HIP-Code beinhaltet, sein. Alternativ kann der Quellcode 4200 in mindestens einer Ausführungsform Mehrfachquellen-Codedateien anstelle einer Einzelquelldatei beinhalten, in die Host-Code und Vorrichtungscode getrennt sind.

[0304] In mindestens einer Ausführungsform ist der Compiler 4201 konfiguriert, um den Quellcode 4200 in den vom Host ausführbaren Code 4202 zur Ausführung auf einem Host und in den von der Vorrichtung ausführbaren Code 4203 zur Ausführung auf einer Vorrichtung zu kompilieren. In mindestens einer Ausführungsform führt der Compiler 4201 Operationen durch, die das Parsen des Quellcodes 4200 in einen abstrakten Systembaum (abstract system tree - AST), das Durchführen von Optimierungen und das Generieren von ausführbarem Code beinhalten. In mindestens einer Ausführungsform, in der der Quellcode 4200 eine Einzelquelldatei beinhaltet, kann der Compiler 4201 den Vorrichtungscode von dem Host-Code in einer solchen Einzelquelldatei trennen, den Vorrichtungscode und den Host-Code in einen von der Vorrichtung ausführbaren Code 4203 bzw. einen vom Host ausführbaren Code 4202 kompilieren und den von der Vorrichtung ausführbaren Code 4203 und den vom Host ausführbaren Code 4202 miteinander in einer Einzeldatei verlinken, wie nachstehend in Bezug auf **Fig. 26** ausführlicher erörtert wird.

[0305] In mindestens einer Ausführungsform können der vom Host ausführbare Code 4202 und der von der Vorrichtung ausführbare Code 4203 in einem beliebigen geeigneten Format vorliegen, wie etwa als Binärcode und/oder IR-Code. Im Falle von CUDA kann in mindestens einer Ausführungsform der vom Host ausführbare Code 4202 nativen Objektcode beinhalten und kann der von der Vorrichtung ausführbare Code 4203 Code in einer PTX-Zwischendarstellung beinhalten. Im Falle von ROCm kann sowohl der vom Host ausführbare Code 4202 als auch der von der Vorrichtung ausführbare Code 4203 in mindestens einer Ausführungsform Ziel-Binärcode beinhalten.

[0306] Andere Variationen sind im Sinne der vorliegenden Offenbarung. Somit können zwar bezüglich der offenbarten Techniken diverse Modifikationen und alternative Konstruktionen vorgenommen werden, bestimmte veranschaulichte Ausführungsformen davon werden jedoch in den Zeichnungen gezeigt und wurden vorangehend ausführlich beschrieben. Allerdings versteht es sich, dass nicht die Absicht verfolgt wird, die Offenbarung auf die spezifische(n) offenbarte(n) Form oder Formen einzuschränken, sondern die Absicht ganz im Gegenteil darin besteht, sämtliche Modifikationen, alternativen Konstruktionen und Äquivalente abzudecken, die in den Geist und Umfang der wie in den beigefügten Ansprüchen definierten Offenbarung fallen.

[0307] Die Verwendung der Ausdrücke „ein“ und „eine“ und „der/die/das“ und ähnlicher Referenten im Kontext des Beschreibens offenbarter Ausführungsformen (insbesondere im Kontext der folgenden Ansprüche) ist so auszulegen, dass sie sowohl den Singular als auch den Plural abdeckt, sofern hierin nichts anderes angegeben ist oder der Kontext dem eindeutig widerspricht, und nicht als Definition eines Ausdrucks. Die Begriffe „umfassend“, „aufweisend“, „beinhaltend“ und „enthaltend“ sollen als offene Begriffe ausgelegt sein (in der Bedeutung „beinhaltend, aber nicht darauf beschränkt“), sofern nichts anderes angegeben ist. Der Begriff „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 hierin soll lediglich als kurzes Verfahren zur einzelnen Bezugnahme auf jeden separaten Wert dienen, der in den Bereich fällt, es sei denn, hierin ist etwas anderes angegeben, und jeder separate Wert ist in die Beschreibung eingeschlossen, als ob er einzeln hierin wiedergegeben wäre. In mindestens einer Ausführungsform ist die Verwendung des Begriffs „Satz“ (z. B. „ein Satz von Objekten“) oder „Teilmenge“ als eine nicht leere Zusammenstellung auszulegen, die ein oder mehrere Elemente umfasst, es sei denn, es ist etwas anderes angemerkt oder dies widerspricht dem Kontext. Sofern nichts anderes angegeben ist oder der Kontext dem widerspricht, bezeichnet ferner der

Ausdruck „Teilmenge“ eines entsprechenden Satzes nicht notwendigerweise eine richtige Teilmenge des entsprechenden Satzes, sondern die Teilmenge und der entsprechende Satz können gleich sein.

[0308] Verbindende Sprache, wie etwa Ausdrücke der Form „wenigstens eines von A, B und C“ oder „mindestens eines von A, B und C“, sind andernfalls in dem Kontext zu verstehen, in dem sie allgemein verwendet werden, um darzustellen, dass ein Objekt, ein Begriff usw. entweder A oder B oder C oder eine beliebige nicht leere Untermenge der Gruppe aus A und B und C sein kann, es sei denn, es ist etwas anderes angegeben oder etwas anderes geht eindeutig aus dem Kontext hervor. In mindestens einer Ausführungsform beziehen sich die verbindenden Ausdrücke „wenigstens eines von A, B und C“ und „mindestens eines von A, B und C“ eines Satzes, der drei Elemente aufweist, auf einen der folgenden Gruppen: {A}, {B}, {C}, {A, B}, {A, C}, {B, C}, {A, B, C}. Somit soll derartige verbindende Sprache im Allgemeinen nicht implizieren, dass bestimmte Ausführungen es erforderlich machen, dass mindestens eines von A, mindestens eines von B und mindestens eines von C vorhanden ist. Sofern nichts anderes angemerkt ist oder der Kontext dem widerspricht, gibt der Ausdruck „Vielzahl“ einen Zustand der Pluralität an (z. B. gibt „eine Vielzahl von Gegenständen“ mehrere Gegenstände an). In mindestens einer Ausführungsform beträgt eine Anzahl der Gegenstände in einer Vielzahl mindestens zwei, es können aber auch mehr sein, wenn dies entweder explizit oder durch den Kontext angegeben ist. Sofern nichts anderes genannt ist oder es anderweitig aus dem Kontext ersichtlich ist, bedeutet die Formulierung „auf Grundlage von“ „mindestens zum Teil auf Grundlage von“ und nicht „ausschließlich auf Grundlage von“.

[0309] Hierin beschriebene Vorgänge von Prozessen können in einer beliebigen geeigneten Reihenfolge durchgeführt werden, sofern es hierin nicht anders angegeben ist oder der Kontext dem anderweitig eindeutig widerspricht. In mindestens einer Ausführungsform wird ein Prozess, wie etwa die hierin beschriebenen Prozesse (oder Variationen und/oder Kombinationen davon), unter der Steuerung von einem oder mehreren Computersystemen durchgeführt, die mit ausführbaren Anweisungen konfiguriert sind, und er ist als Code (z. B. ausführbare Anweisungen, ein oder mehrere Computerprogramme oder eine oder mehrere Anwendungen), der zusammen auf einem oder mehreren Prozessoren ausgeführt wird, durch Hardware oder Kombinationen davon implementiert. In mindestens einer Ausführungsform wird Code auf einem computerlesbaren Speichermedium gespeichert. In einer Ausführungsform in der 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 nichttransitorisches computerlesbares Speichermedium, das transitorische Signale (z. B. eine sich ausbreitende transiente elektrische oder elektromagnetische Übertragung) ausschließt, aber nichttransitorische Datenspeicherschaltungen (z. B. Puffer, Cache und Warteschlangen) innerhalb von Sendeempfängern von transitorischen Signalen einschließt. In mindestens einer Ausführungsform ist der Code (z. B. ausführbarer Code oder Quellcode) auf einem Satz von einem oder mehreren nichttransitorischen computerlesbaren Speichermedien gespeichert, auf denen ausführbare Anweisungen gespeichert sind (oder einem anderen Speicher zum Speichern ausführbarer Anweisungen), die bei Ausführung (d. h. als Ergebnis der Ausführung) durch einen oder mehrere Prozessoren eines Computersystems das Computersystem dazu veranlassen, hierin beschriebene Operationen durchzuführen. 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 möglicherweise nicht über den gesamten Code, während mehrere nichtflüchtige computerlesbares Speichermedien gemeinschaftlich den gesamten Code speichern. In mindestens einer Ausführungsform werden die ausführbaren Anweisungen so ausgeführt, dass unterschiedliche Anweisungen durch unterschiedliche Prozessoren ausgeführt werden - in mindestens einer Ausführungsform speichert ein nichttransitorisches computerlesbares Speichermedium Anweisungen und eine hauptsächliche zentrale Verarbeitungseinheit („CPU“) führt einige der Anweisungen aus, während eine Grafikverarbeitungseinheit („GPU“) andere Anweisungen ausführt. In mindestens einer Ausführungsform weisen unterschiedliche Komponenten eines Computersystems separate Prozessoren auf und unterschiedliche Prozessoren führen unterschiedliche Teilmengen von Anweisungen aus.

[0310] Dementsprechend sind in mindestens einer Ausführungsform Computersysteme so konfiguriert, dass sie einen oder mehrere Dienste implementieren, die einzeln oder kollektiv Vorgänge der hierin beschriebenen Prozesse durchführen, und sind solche Computersysteme mit geeigneter Hardware und/oder Software konfiguriert, welche die Durchführung der Vorgänge ermöglichen. Ferner ist ein Computersystem, das mindestens eine Ausführungsform der vorliegenden Offenbarung implementiert, eine einzelne Vorrichtung und in einer anderen Ausführungsform ein verteiltes Computersystem, das mehrere Vorrichtungen umfasst, die unterschiedlich arbeiten, sodass das verteilte Computersystem die hierin beschriebenen Operationen durchführt und sodass eine einzelne Vorrichtung nicht alle Operationen durchführt.

[0311] Die Verwendung jeglicher und aller der mindestens einen Ausführungsformen oder beispielhafter Wortwahl (z. B. „wie etwa“), die hierin bereitgestellt ist, soll lediglich die Ausführungsformen der Offenbarung besser verdeutlichen und stellt keine Einschränkung des Umfangs der Offenbarung dar, es sei denn, es ist etwas anderes beansprucht. Keinerlei Wortwahl in der Beschreibung sollte so ausgelegt werden, dass sie ein beliebiges nicht beanspruchtes Element als für die Umsetzung der Offenbarung wesentlich angibt.

[0312] Jegliche Referenzen, einschließlich Veröffentlichungen, Patentanmeldungen und Patenten, die hierin erwähnt werden, sind hiermit durch Bezugnahme in demselben Maße aufgenommen, als wäre jede Referenz einzeln und spezifisch als durch Referenz eingeschlossen angegeben und in ihrer Gesamtheit hierin ausgeführt.

[0313] In der Beschreibung und den Ansprüchen können die Begriffe „gekoppelt“ und „verbunden“ sowie deren Ableitungen verwendet werden. Es versteht sich, dass diese Ausdrücke nicht als Synonyme füreinander beabsichtigt sein können. Vielmehr kann in einer von mindestens einer Ausführungsform „verbunden“ oder „gekoppelt“ verwendet werden, um anzugeben, dass zwei oder mehr Elemente in direktem oder indirektem physischem oder elektrischem Kontakt miteinander stehen. Mit „gekoppelt“ kann auch gemeint sein, dass zwei oder mehr Elemente nicht in direktem Kontakt miteinander stehen, jedoch trotzdem miteinander zusammenwirken oder interagieren.

[0314] Sofern nicht ausdrücklich anders angegeben, versteht es sich, dass sich Begriffe wie „Verarbeitung“, „Berechnung“, „Berechnen“, „Bestimmen“ oder dergleichen in der gesamten Beschreibung auf Handlungen und/oder Prozesse eines Computers oder Rechensystems oder einer ähnlichen elektronischen Rechenvorrichtung, die Daten, die als physische, z. B. elektronische, Größen in den Registern und/oder Speichern des Rechensystems dargestellt sind, manipulieren und/oder in andere Daten umwandeln, die auf ähnliche Weise als physische Größen in den Speichern, Registern oder anderen derartigen Informationsspeicher-, - übertragungs- oder -anzeigevorrichtungen des Rechensystems dargestellt sind.

[0315] Auf ähnliche Weise kann sich der Begriff „Prozessor“ auf eine beliebige Vorrichtung oder einen Abschnitt einer Vorrichtung beziehen, die/der elektronische Daten aus Registern und/oder dem Speicher verarbeitet und diese elektronischen Daten in andere elektronische Daten umwandelt, die in Registern und/oder im Speicher gespeichert werden können. Als nicht einschränkende der mindestens einen Ausführungsform kann der „Prozessor“ eine CPU oder eine GPU sein. Eine „Rechenplattform“ kann einen oder mehrere Prozessoren umfassen. Im hierin verwendeten Sinne können „Software“-Prozesse zum Beispiel Software- und/oder Hardware-Entitäten beinhalten, die im Verlauf der Zeit Arbeit Durchführen, wie etwa Tasks, Threads und intelligente Agenten. Außerdem kann sich jeder Prozess auf mehrere Prozesse beziehen, um Anweisungen nacheinander oder parallel, kontinuierlich oder intermittierend auszuführen. Die Begriffe „System“ und „Verfahren“ werden hierin insofern austauschbar verwendet, dass ein System ein oder mehrere Verfahren verkörpern kann und die Verfahren als System betrachtet werden können.

[0316] Im vorliegenden Dokument kann auf das Erlangen, Erfassen, Empfangen oder Eingeben von analogen oder digitalen Daten in ein Teilsystem, ein Computersystem oder eine computerimplementierte Maschine Bezug genommen werden. In mindestens einer Ausführungsform kann der Prozess des Erlangens, Erfassens, Empfangens oder Eingebens von analogen und digitalen Daten auf eine Vielfalt von Weisen erzielt werden, wie etwa durch das Empfangen von Daten als Parameter eines Funktionsaufrufs oder eines Aufrufs einer 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 anderen Implementation kann der Prozess des Erhaltens, Erfassens, Empfangens oder Eingebens von analogen oder digitalen Daten durch die Übertragung von Daten über ein Computernetzwerk von der bereitstellenden Entität zur erfassenden Entität erfolgen. Es kann auch auf das Bereitstellen, Ausgeben, Übertragen, Senden oder Darstellen von analogen oder digitalen Daten Bezug genommen werden. In verschiedenen der mindestens einen Ausführungsform kann der Prozess des Bereitstellens, Ausgebens, Übertragens, Sendens oder Darstellens von analogen oder digitalen Daten durch die Übertragung von Daten als Eingabe- oder Ausgabeparameter eines Funktionsaufrufs, Parameter einer Anwendungsprogrammierschnittstelle oder eines Interprozess-Kommunikationsmechanismus erfolgen.

[0317] Auch wenn die vorstehende Erörterung einige der mindestens einen Ausführungsform mit Implementationen der beschriebenen Techniken darlegt, können auch andere Architekturen verwendet werden, um die beschriebene Funktionalität zu implementieren, und sie sollen im Umfang dieser Offenbarung liegen. Darüber hinaus könnten, obwohl spezifische Verteilungen von Zuständigkeiten vorstehend zum Zwecke der Erörte-

rung definiert sind, verschiedene Funktionen und Zuständigkeiten in Abhängigkeit von den Umständen anders verteilt und aufgeteilt werden.

[0318] Obwohl der Gegenstand in für Strukturmerkmale und/oder Verfahrenshandlungen spezifischer Sprache beschrieben wurde, versteht es sich ferner, dass der in den beigefügten Ansprüchen beanspruchte Gegenstand nicht unbedingt auf die beschriebenen konkreten Merkmale oder Handlungen beschränkt ist. Vielmehr werden spezifische Merkmale und Handlungen als beispielhafte Formen zum Implementieren der Ansprüche offenbart.

Patentansprüche

1. Prozessor, umfassend:
eine oder mehrere Schaltungen, um eine Netzkapazität zumindest teilweise basierend auf Netznutzungsinformationen zu bestimmen und einen oder mehrere Betriebsparameter für eine oder mehrere Rechenzentrumskomponenten zumindest teilweise basierend auf der Netzkapazität anzupassen.

2. Prozessor nach Anspruch 1, wobei der eine oder die mehreren Betriebsparameter mindestens einem von einer Prozessoreinheit-Leistungsgrenze, einem Knotenbetriebsstatus, einem Stromquellenstatus oder einem Arbeitslaststatus entsprechen.

3. Prozessor nach Anspruch 1 oder 2, wobei der eine oder die mehreren Schaltungen ferner dazu dienen, eine künftige Rechenzentrumslast zu bestimmen, die zumindest teilweise auf einer oder mehreren anstehenden Arbeitslasten basiert, und einen oder mehrere künftige Betriebsparameter zumindest teilweise auf der Grundlage der Netzkapazität anzupassen.

4. Prozessor nach einem der vorhergehenden Ansprüche, wobei der eine oder die mehreren Schaltungen ferner dazu dienen, eine Leistungssignatur für eine oder mehrere Arbeitslasten zu bestimmen, die dem einen oder den mehreren Betriebsparametern zugeordnet sind.

5. Prozessor nach Anspruch 4, wobei der eine oder die mehreren Schaltungen ferner dazu dienen, festzustellen, dass die Leistungssignatur schädlich ist.

6. Prozessor nach Anspruch 5, wobei die eine oder mehreren Schaltungen ferner dazu dienen, die Ausführung der einen oder mehreren Arbeitslasten zu beenden, die der Leistungssignatur zugeordnet sind.

7. Prozessor nach einem der vorhergehenden Ansprüche, wobei der eine oder die mehreren Betriebsparameter einem oder mehreren Betriebsparametern für mehrere Rechenzentren entsprechen, die auf einem gemeinsamen Netz arbeiten.

8. Verfahren, das Folgendes umfasst:
Bestimmen einer Netzkapazität, zumindest teilweise auf der Grundlage von Netznutzungsinformationen; und
Anpassen eines oder mehrerer Betriebsparameter von Rechenzentrumskomponenten zumindest teilweise basierend auf der Netzkapazität.

9. Verfahren nach Anspruch 8, das ferner Folgendes umfasst:
Bestimmen einer erwarteten Rechenzentrumslast; und
Bestimmen, dass die erwartete Rechenzentrumslast eine Schwellenlast in Bezug auf die Netzkapazität überschreitet.

10. Verfahren nach Anspruch 9, wobei der eine oder die mehreren Betriebsparameter der Rechenzentrumskomponente einer Zeit zum Ausführen einer Arbeitslast entsprechen, wobei das Verfahren ferner Folgendes umfasst:
Ändern einer Ausführungszeit der Arbeitslast.

11. Verfahren nach Anspruch 9 oder 10, wobei der eine oder die mehreren Rechenzentrumskomponenten-Betriebsparameter einem Knotenbetriebszustand entsprechen, wobei das Verfahren ferner Folgendes umfasst:
Beenden des Knotenbetriebs, wobei der Knoten eine oder mehrere Rechenzentrumskomponenten enthält.

12. Verfahren nach einem der Ansprüche 9 bis 11, ferner umfassend:
Bestimmen einer Leistungssignatur für eine oder mehrere Rechenzentrumskomponenten, die mit dem einen oder den mehreren Rechenzentrumskomponenten-Betriebsparametern assoziiert sind; und
Vergleichen der Leistungssignatur mit einer oder mehreren schädlichen Leistungssignaturen; und
Bestimmen, dass die Leistungssignatur der einen oder mehreren schädlichen Leistungssignaturen entspricht.

13. Verfahren nach einem der Ansprüche 8 bis 12, wobei der eine oder die mehreren Rechenzentrumsbetriebsparameter einer Komponentenleistungsgrenze entsprechen, wobei das Verfahren ferner Folgendes umfasst:
Verringern der Komponentenleistungsgrenze für eine oder mehrere Rechenzentrumskomponenten.

14. Verfahren nach einem der Ansprüche 8 bis 13, ferner umfassend:
Bewirken, dass mindestens ein Teil einer oder mehrerer Rechenzentrumskomponenten Strom von einer oder mehreren alternativen Stromquellen bezieht.

15. Verfahren nach einem der Ansprüche 8 bis 14, ferner umfassend:
Bestimmen einer erwarteten Rechenzentrumslast; und
Bestimmen, dass die erwartete Rechenzentrumslast niedriger als eine Schwellenlast in Bezug auf die Netzkapazität ist.

16. Verfahren nach Anspruch 15, das ferner Folgendes umfasst:
Ausführen einer oder mehrerer Dummy-Arbeitslasten, um eine Rechenzentrumsbetriebslast über einem Schwellenwert zu halten.

17. Prozessor, umfassend:
eine oder mehrere Schaltungen, um eine verfügbare Netzkapazität zumindest teilweise auf der Grundlage von Netznutzungsinformationen zu bestimmen und den Betrieb eines oder mehrerer Knoten zumindest teilweise auf der Grundlage der Feststellung zu beenden, dass der Betrieb des einen oder der mehreren Knoten die verfügbare Netzkapazität übersteigt.

18. Prozessor nach Anspruch 17, wobei eine Leistungssignatur des einen oder der mehreren Knoten einer schädlichen Leistungssignatur entspricht.

19. Prozessor nach Anspruch 17 oder 18, wobei der eine oder die mehreren Knoten in unterschiedlichen Rechenzentren angeordnet sind.

20. Prozessor nach einem der Ansprüche 17 bis 19, wobei die eine oder mehreren Schaltungen ferner einen oder mehrere zusätzliche Knoten aktivieren, um eine oder mehrere Dummy-Arbeitslasten nach Beendigung des einen oder der mehreren Knoten auszuführen.

Es folgen 51 Seiten Zeichnungen

Anhängende Zeichnungen

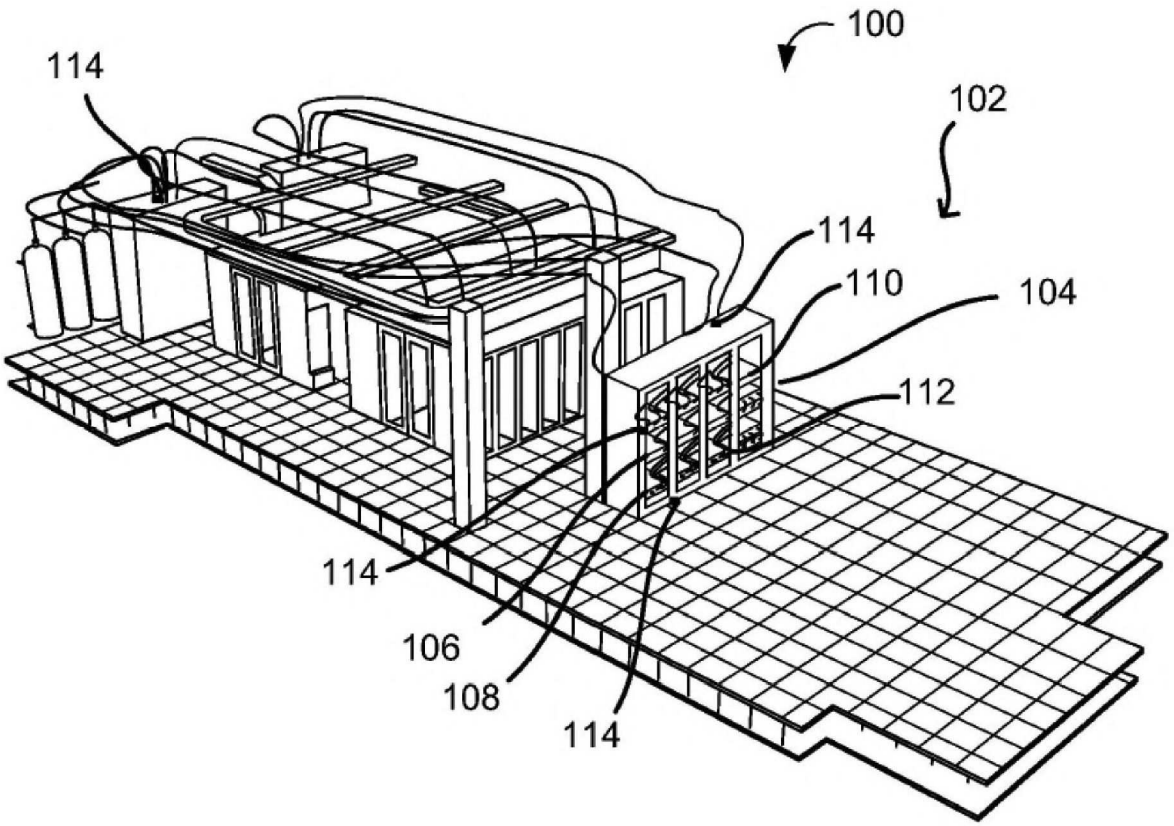


FIG. 1

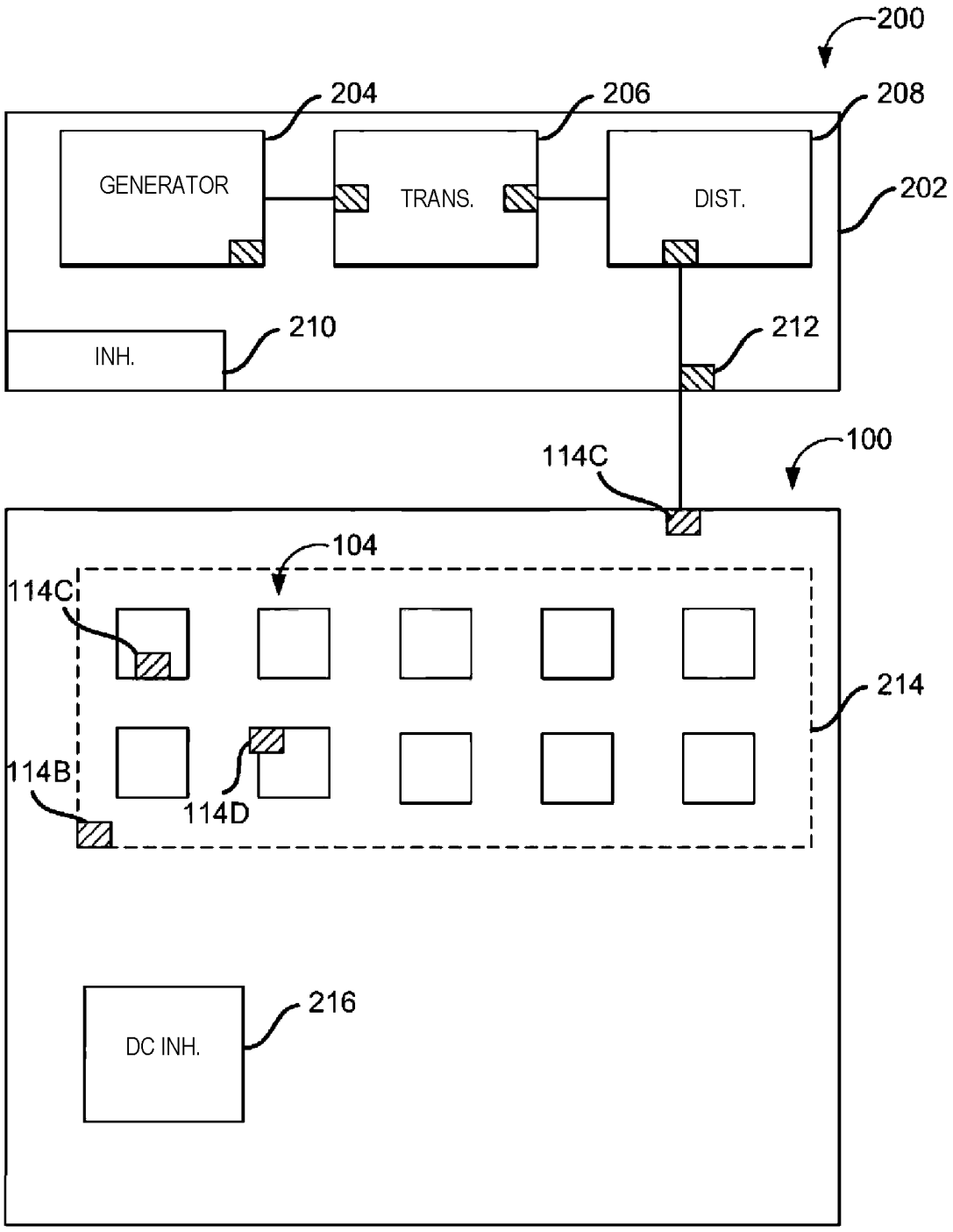

200
204
206
208
GENERATOR
TRANS.
DIST.
202
INH.
210
212
114C
100
104
114C
214
114B
114D
DC INH.
216

FIG. 2

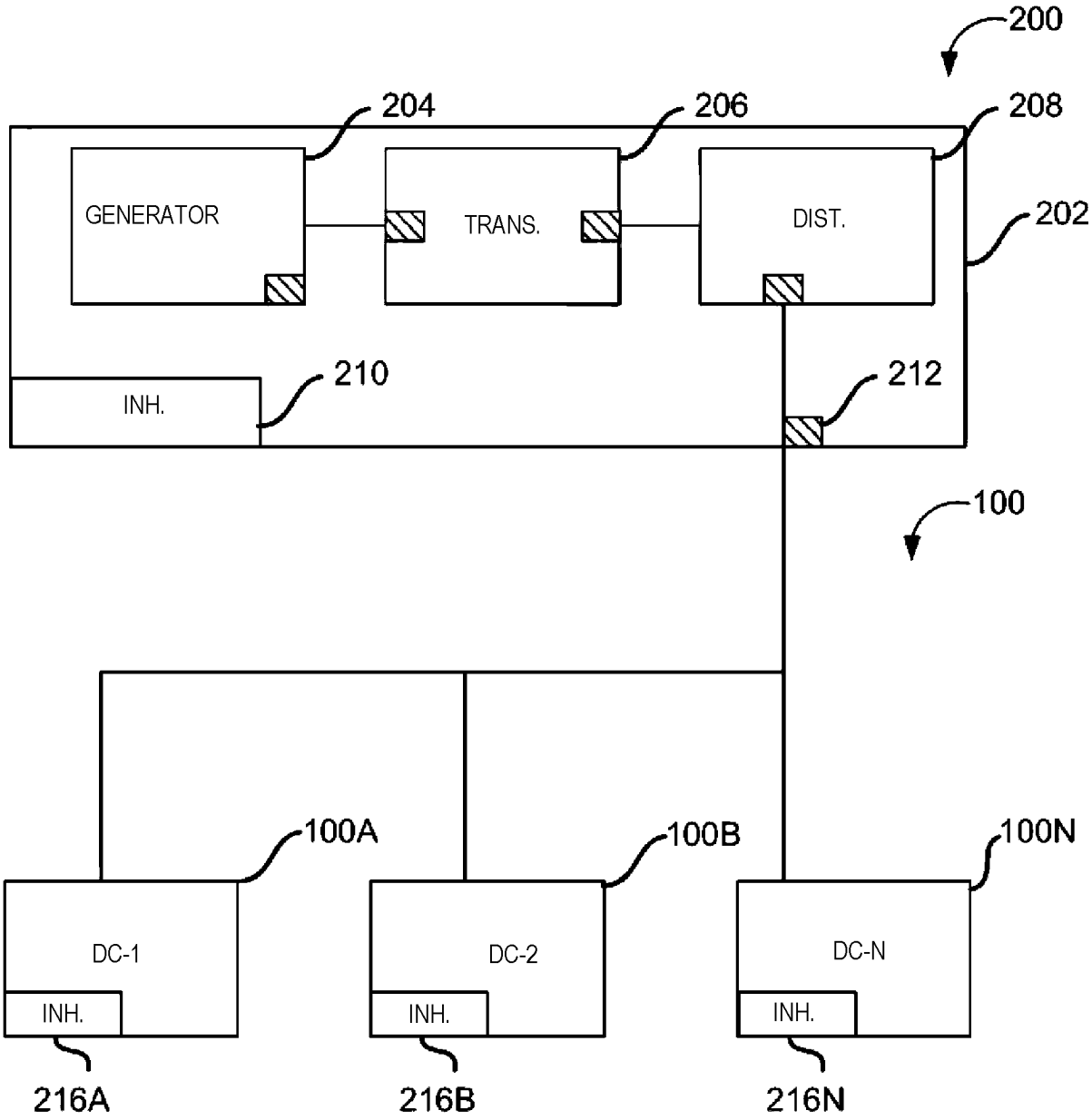

200
204
206
208
GENERATOR
TRANS.
DIST.
202
INH.
210
212
100
100A
100B
100N
DC-1
DC-2
DC-N
INH.
INH.
INH.
216A
216B
216N

FIG. 3

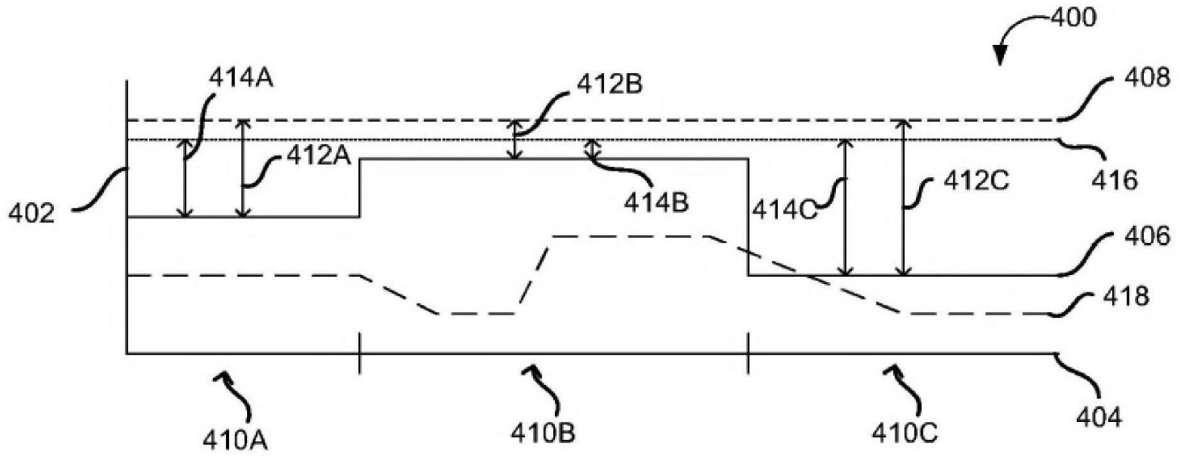

400
414A
412B
408
412A
416
402
414B
414C
412C
406
418
410A
410B
410C
404

FIG. 4A

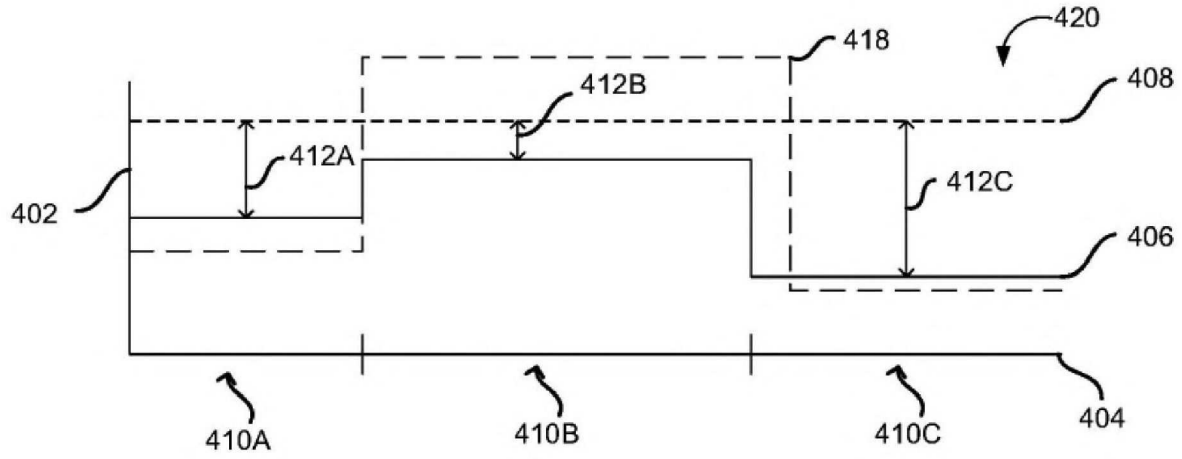

420
418
412B
408
412A
412C
402
406
410A
410B
410C
404

FIG. 4B

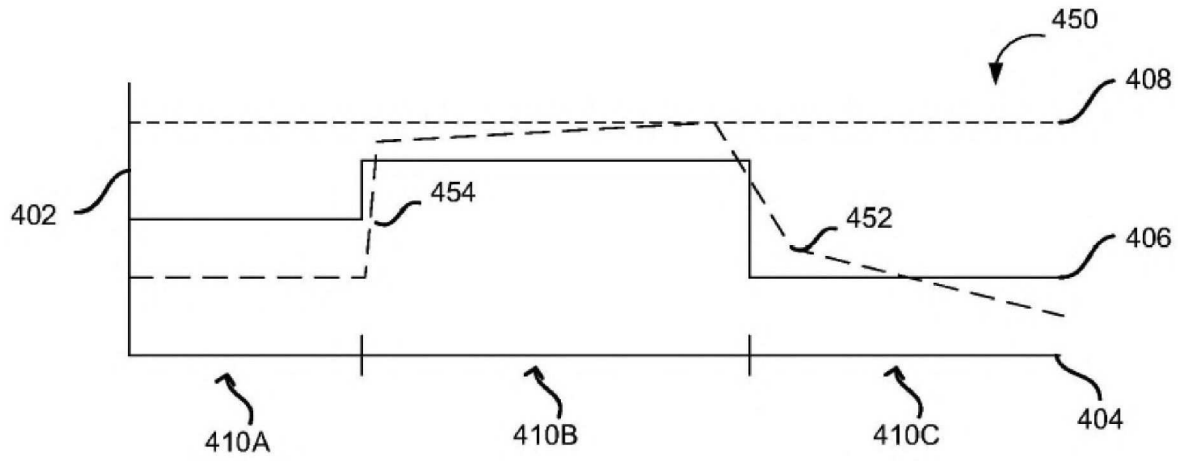

450
408
402
454
452
406
410A
410B
410C
404

FIG. 4C

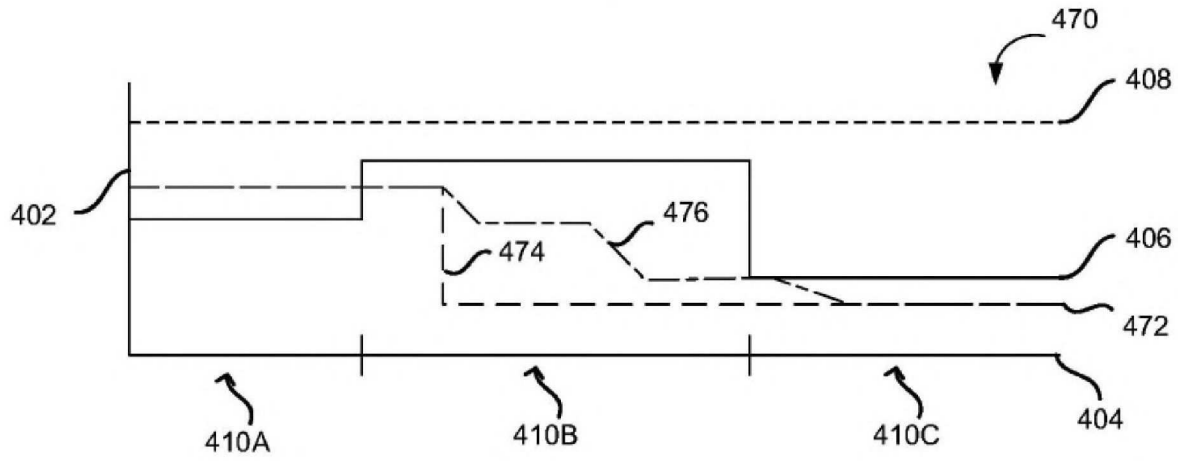

470
408
402
476
474
406
472
410A
410B
410C
404

FIG. 4D

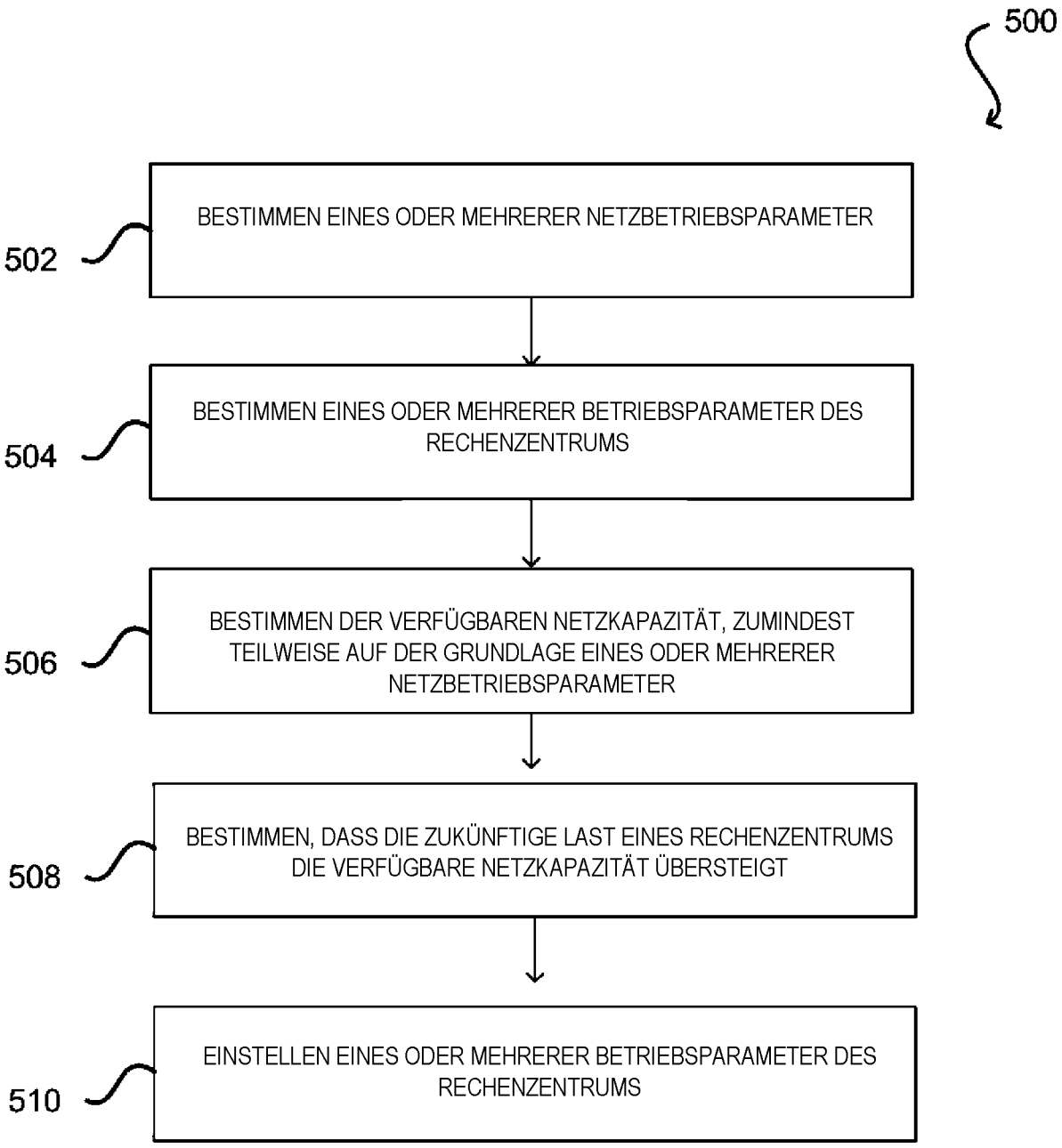

500
502
BESTIMMEN EINES ODER MEHRERER NETZBETRIEBSPARAMETER
504
BESTIMMEN EINES ODER MEHRERER BETRIEBSPARAMETER DES RECHENZENTRUMS
506
BESTIMMEN DER VERFÜGBAREN NETZKAPAZITÄT, ZUMINDEST TEILWEISE AUF DER GRUNDLAGE EINES ODER MEHRERER NETZBETRIEBSPARAMETER
508
BESTIMMEN, DASS DIE ZUKÜNFTIGE LAST EINES RECHENZENTRUMS DIE VERFÜGBARE NETZKAPAZITÄT ÜBERSTEIGT
510
EINSTELLEN EINES ODER MEHRERER BETRIEBSPARAMETER DES RECHENZENTRUMS

FIG. 5A

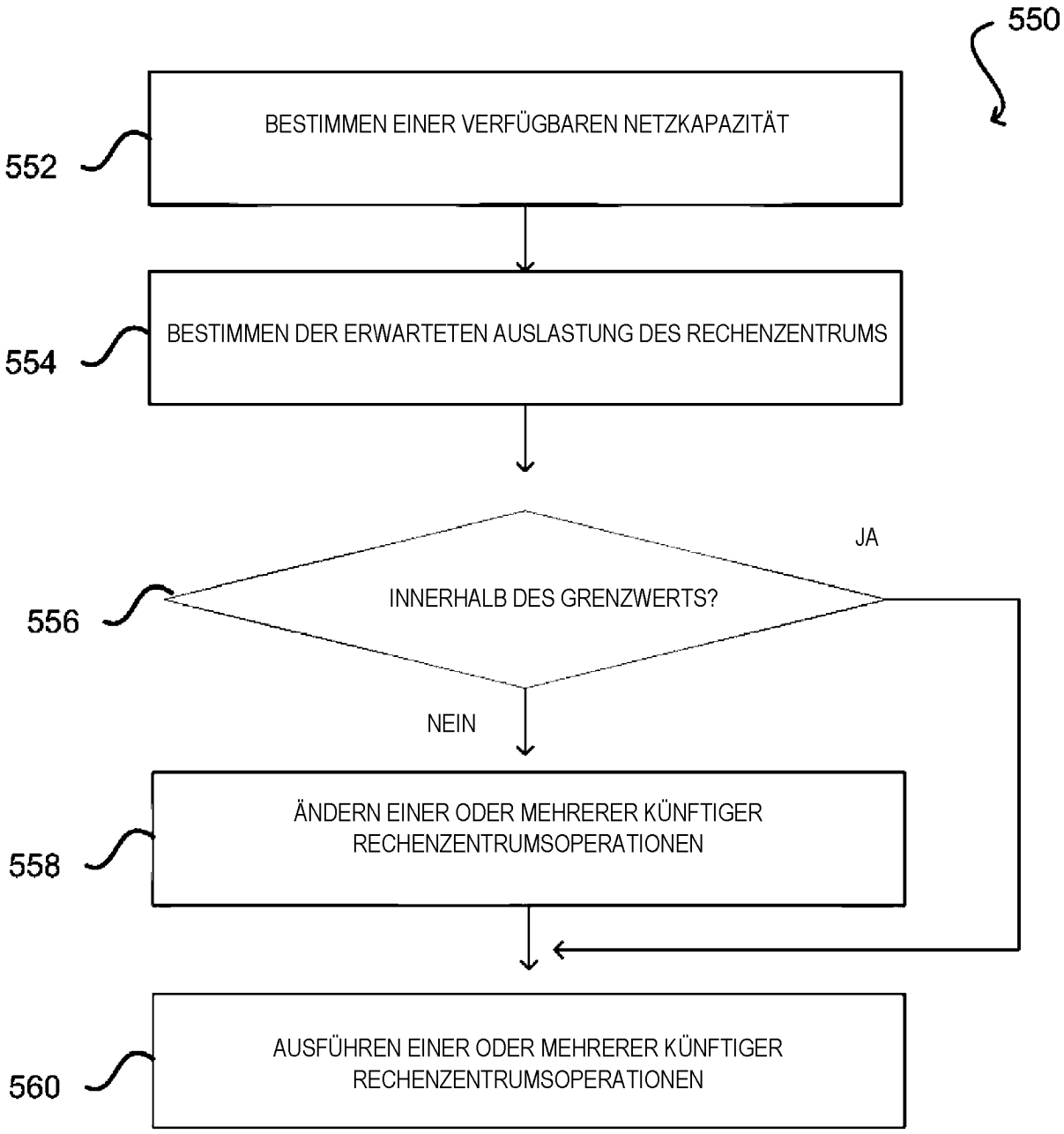

550
552
BESTIMMEN EINER VERFÜGBAREN NETZKAPAZITÄT
554
BESTIMMEN DER ERWARTETEN AUSLASTUNG DES RECHENZENTRUMS
556
INNERHALB DES GRENZWERTS?
JA
NEIN
558
ÄNDERN EINER ODER MEHRERER KÜNFTIGER RECHENZENTRUMSOPERATIONEN
560
AUSFÜHREN EINER ODER MEHRERER KÜNFTIGER RECHENZENTRUMSOPERATIONEN

FIG. 5B

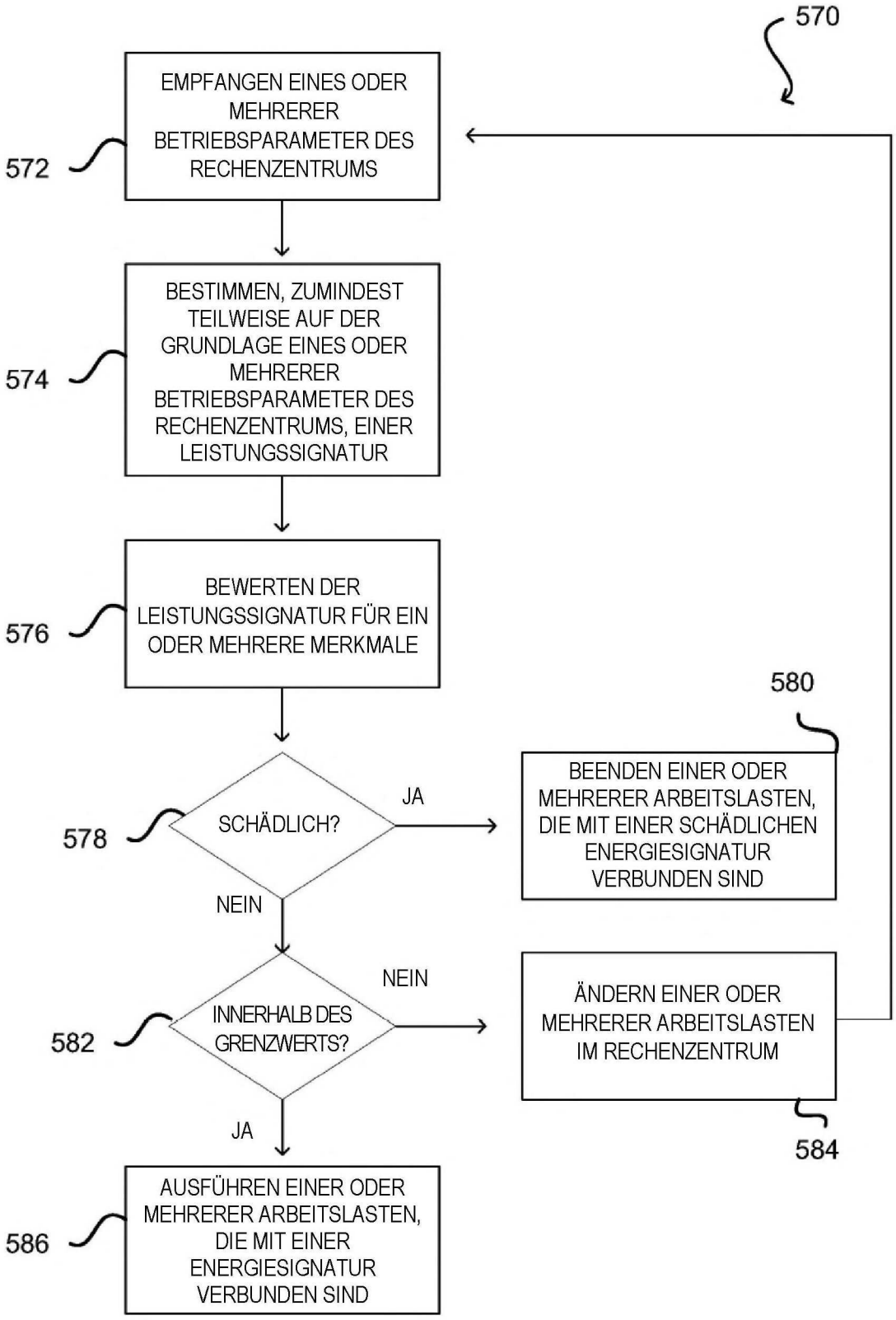

570
EMPFANGEN EINES ODER MEHRERER BETRIEBSPARAMETER DES RECHENZENTRUMS
572
BESTIMMEN, ZUMINDEST TEILWEISE AUF DER GRUNDLAGE EINES ODER MEHRERER BETRIEBSPARAMETER DES RECHENZENTRUMS, EINER LEISTUNGSSIGNATUR
574
BEWERTEN DER LEISTUNGSSIGNATUR FÜR EIN ODER MEHRERE MERKMALE
576
SCHÄDLICH?
578
JA
BEENDEN EINER ODER MEHRERER ARBEITSLASTEN, DIE MIT EINER SCHÄDLICHEN ENERGIESIGNATUR VERBUNDEN SIND
580
NEIN
INNERHALB DES GRENZWERTS?
582
NEIN
ÄNDERN EINER ODER MEHRERER ARBEITSLASTEN IM RECHENZENTRUM
584
JA
AUSFÜHREN EINER ODER MEHRERER ARBEITSLASTEN, DIE MIT EINER ENERGIESIGNATUR VERBUNDEN SIND
586

FIG. 5C

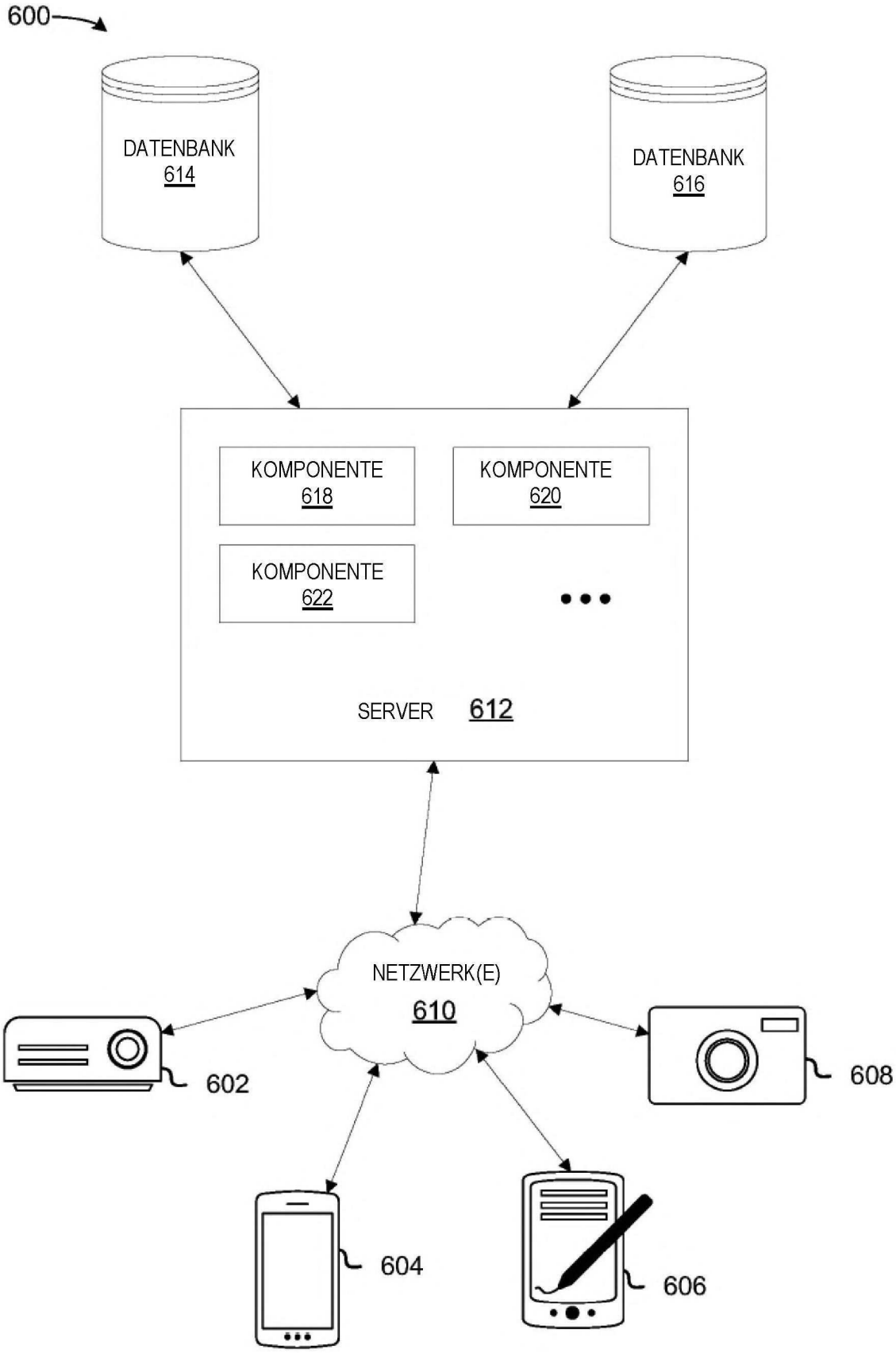

600
DATENBANK
614
DATENBANK
616
KOMPONENTE
618
KOMPONENTE
620
KOMPONENTE
622
SERVER
612
NETZWERK(E)
610
602
604
606
608

FIG. 6

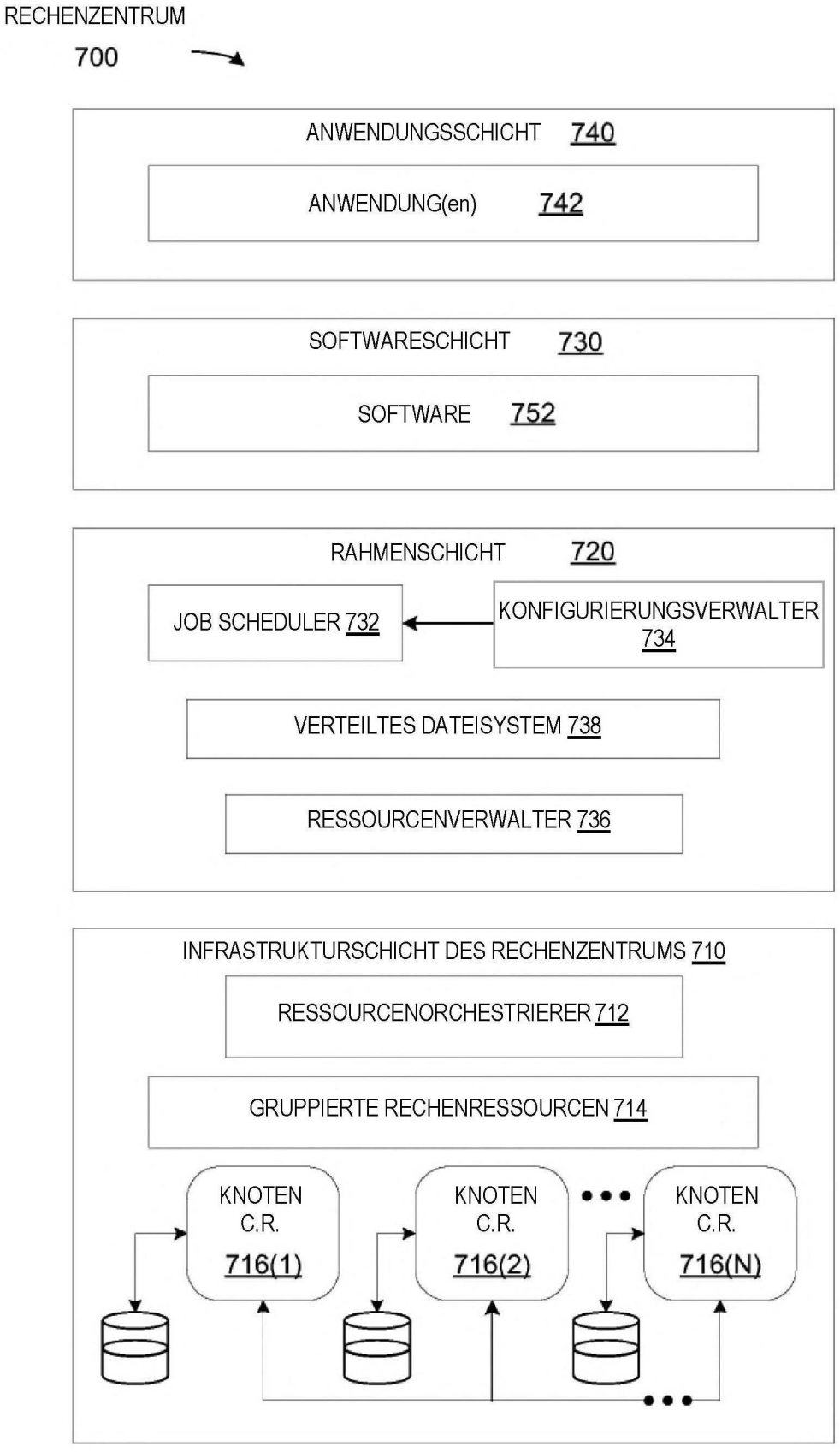

RECHENZENTRUM
700
ANWENDUNGSSCHICHT 740
ANWENDUNG(en) 742
SOFTWARESCHICHT 730
SOFTWARE 752
RAHMENSCHICHT 720
JOB SCHEDULER 732
KONFIGURIERUNGSVERWALTER 734
VERTEILTES DATEISYSTEM 738
RESSOURCENVERWALTER 736
INFRASTRUKTURSCHICHT DES RECHENZENTRUMS 710
RESSOURCENORCHESTRIERER 712
GRUPPIERTE RECHENRESSOURCEN 714
KNOTEN C.R. 716(1)
KNOTEN C.R. 716(2)
KNOTEN C.R. 716(N)

FIG. 7

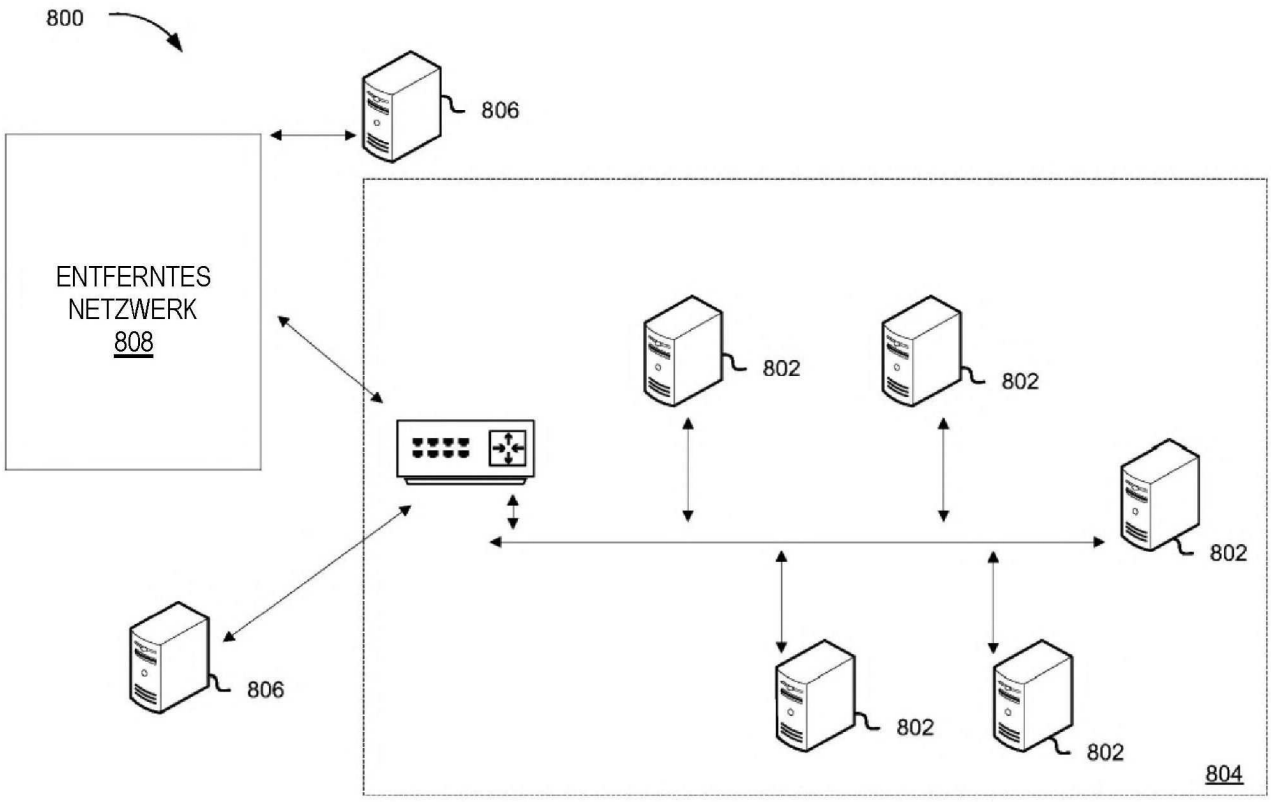


FIG. 8

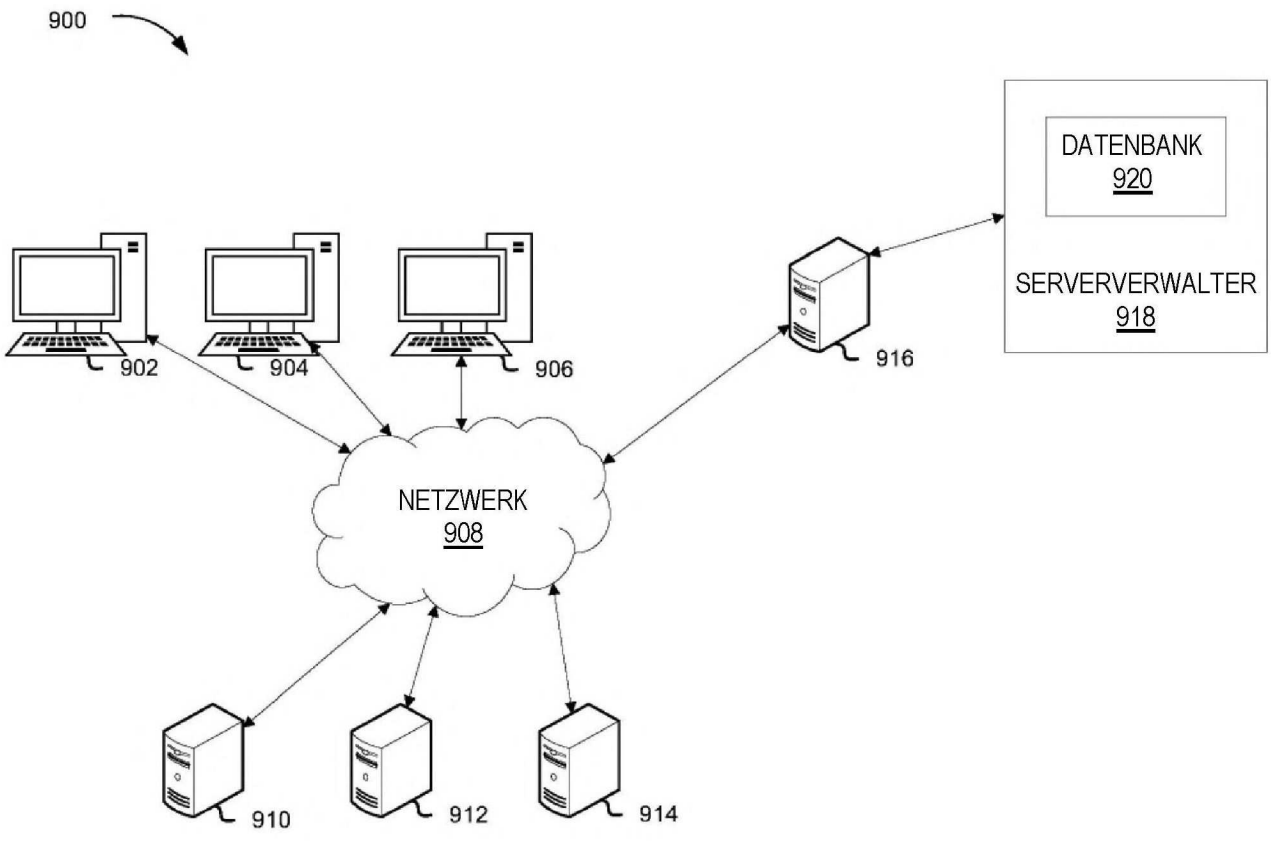

900
DATENBANK
920
SERVERVERWALTER
918
902
904
906
916
NETZWERK
908
910
912
914

FIG. 9

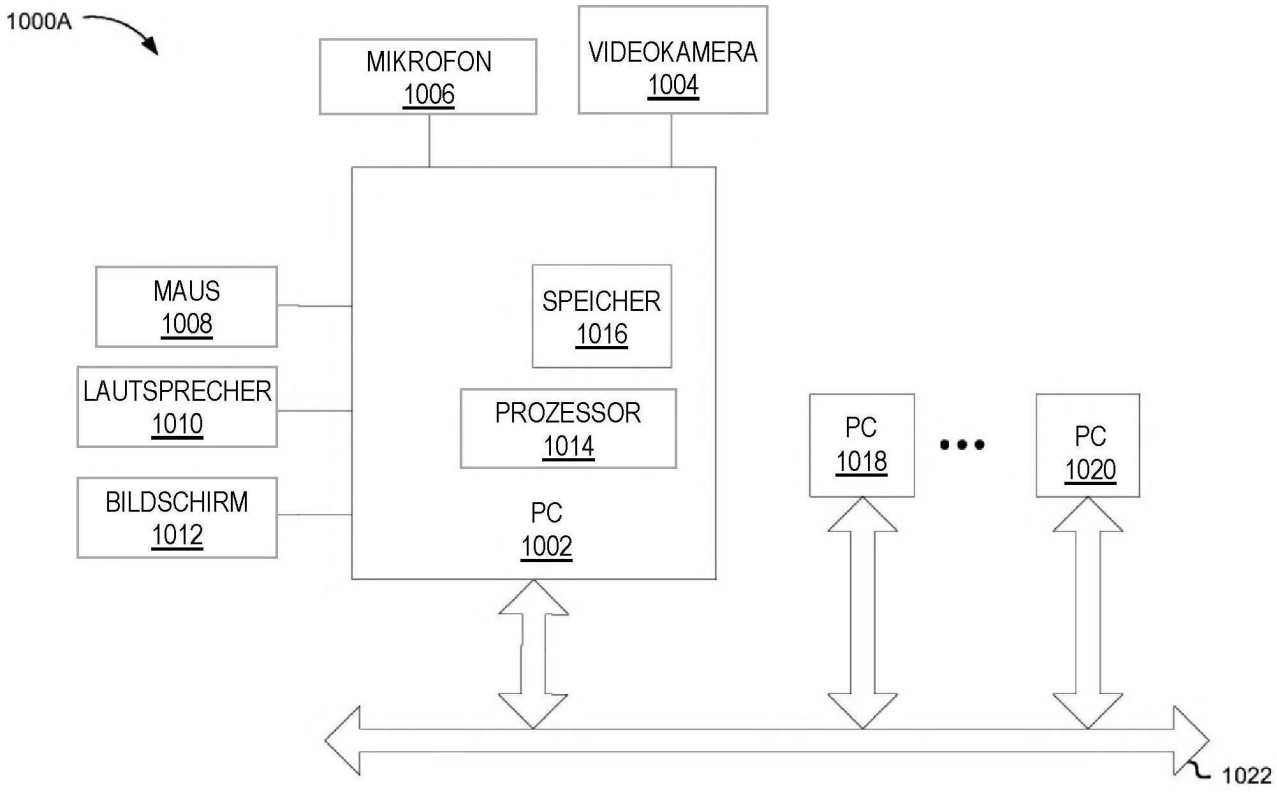

1000A
MIKROFON
1006
VIDEOKAMERA
1004
MAUS
1008
LAUTSPRECHER
1010
BILDSCHIRM
1012
SPEICHER
1016
PROZESSOR
1014
PC
1002
PC
1018
PC
1020
1022

FIG. 10A

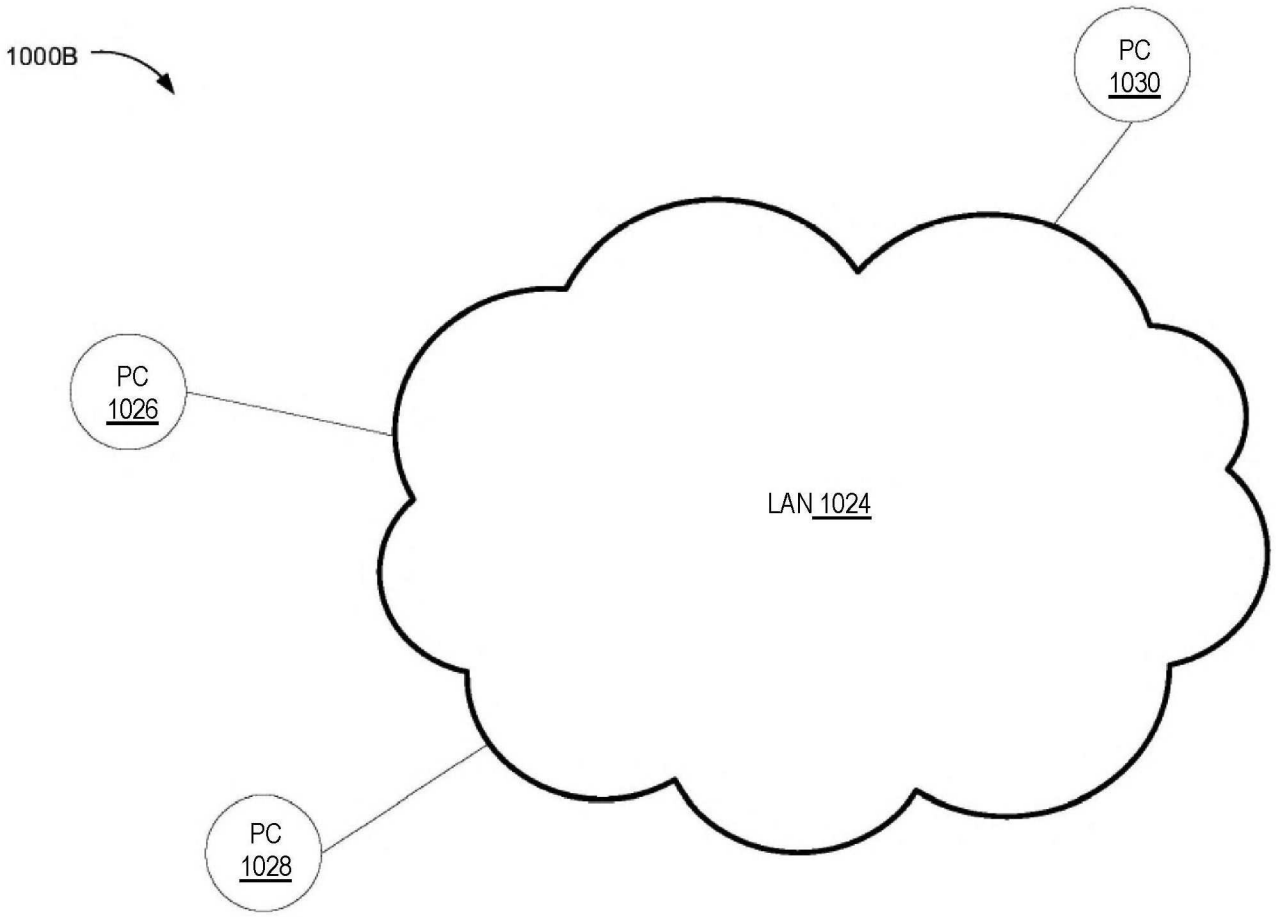


FIG. 10B

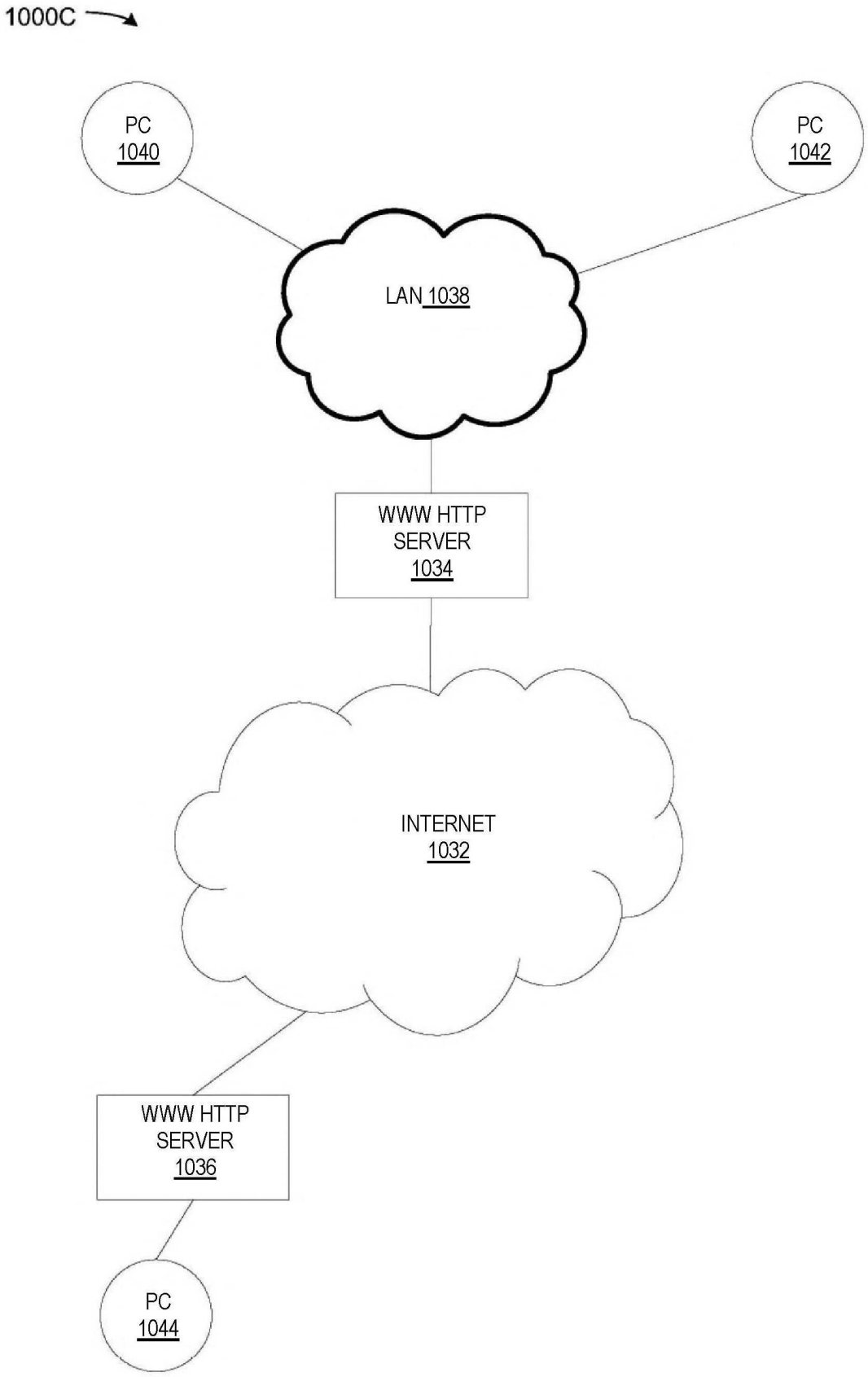

1000C
PC
1040
PC
1042
LAN 1038
WWW HTTP
SERVER
1034
INTERNET
1032
WWW HTTP
SERVER
1036
PC
1044

FIG. 10C

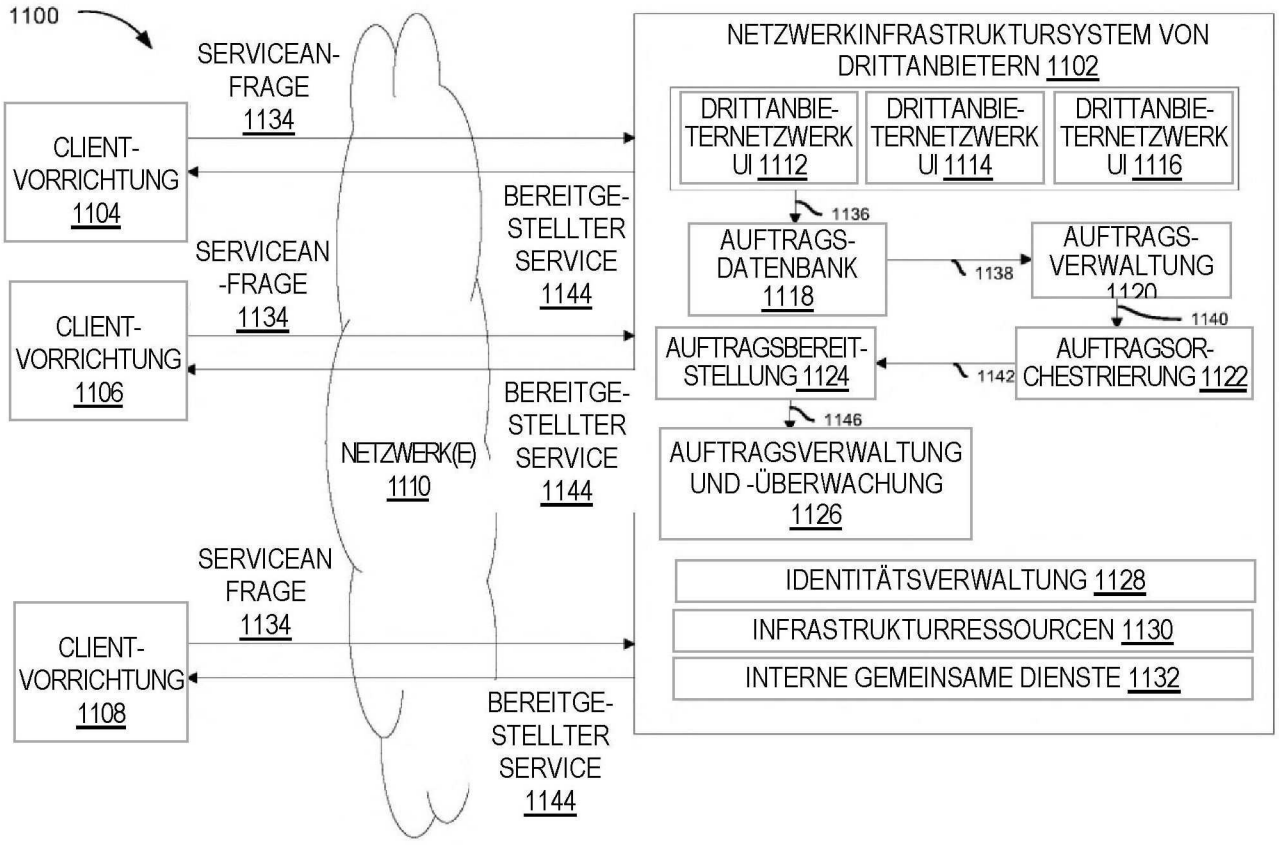

1100
SERVICEAN-FRAGE 1134
CLIENT-VORRICHTUNG 1104
SERVICEAN-FRAGE 1134
CLIENT-VORRICHTUNG 1106
SERVICEAN-FRAGE 1134
CLIENT-VORRICHTUNG 1108
NETZWERK(E) 1110
BEREITGE-STELLTER SERVICE 1144
BEREITGE-STELLTER SERVICE 1144
BEREITGE-STELLTER SERVICE 1144
NETZWERKINFRASTRUKTURSYSTEM VON DRITTANBIETERN 1102
DRITTANBIE-TERNETZWERK UI 1112
DRITTANBIE-TERNETZWERK UI 1114
DRITTANBIE-TERNETZWERK UI 1116
1136
AUFTRAGS-DATENBANK 1118
1138
AUFTRAGS-VERWALTUNG 1120
1140
AUFTRAGSBEREIT-STELLUNG 1124
1142
AUFTRAGSOR-CHESTRIERUNG 1122
1146
AUFTRAGSVERWALTUNG UND -ÜBERWACHUNG 1126
IDENTITÄTSVERWALTUNG 1128
INFRASTRUKTURRESSOURCEN 1130
INTERNE GEMEINSAME DIENSTE 1132

FIG. 11

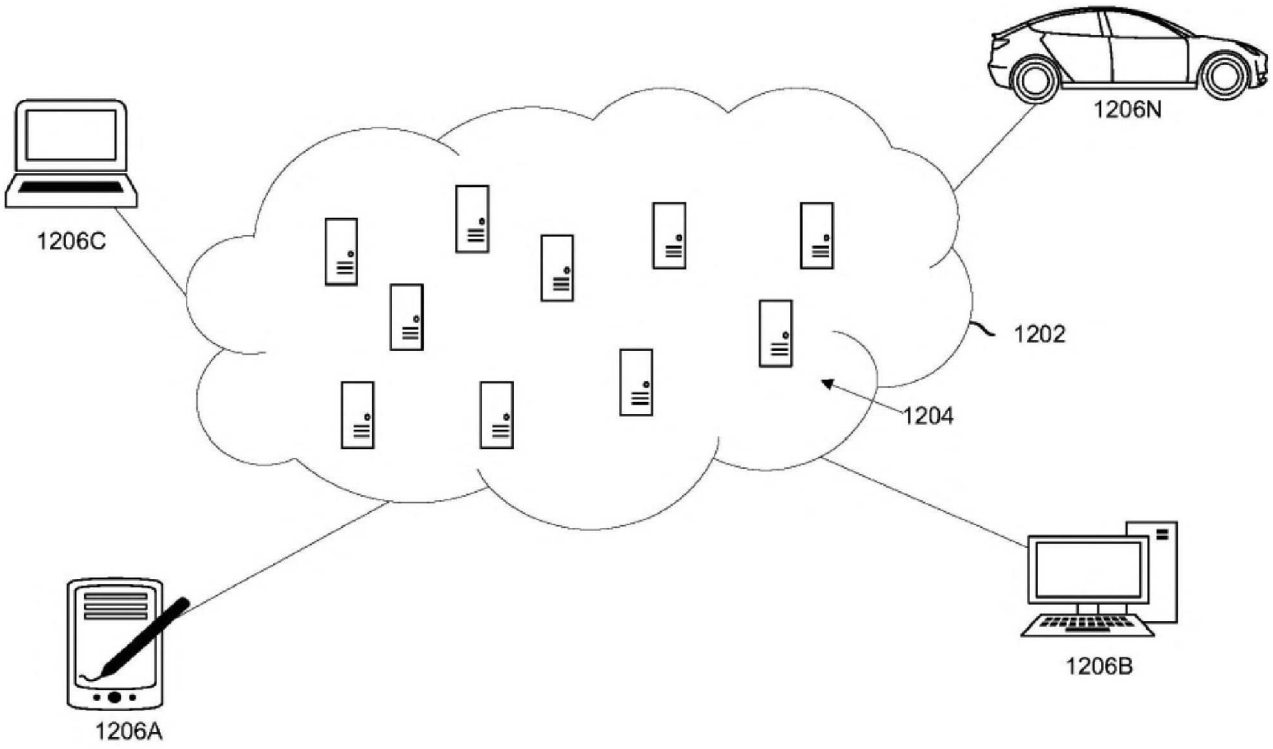


FIG. 12

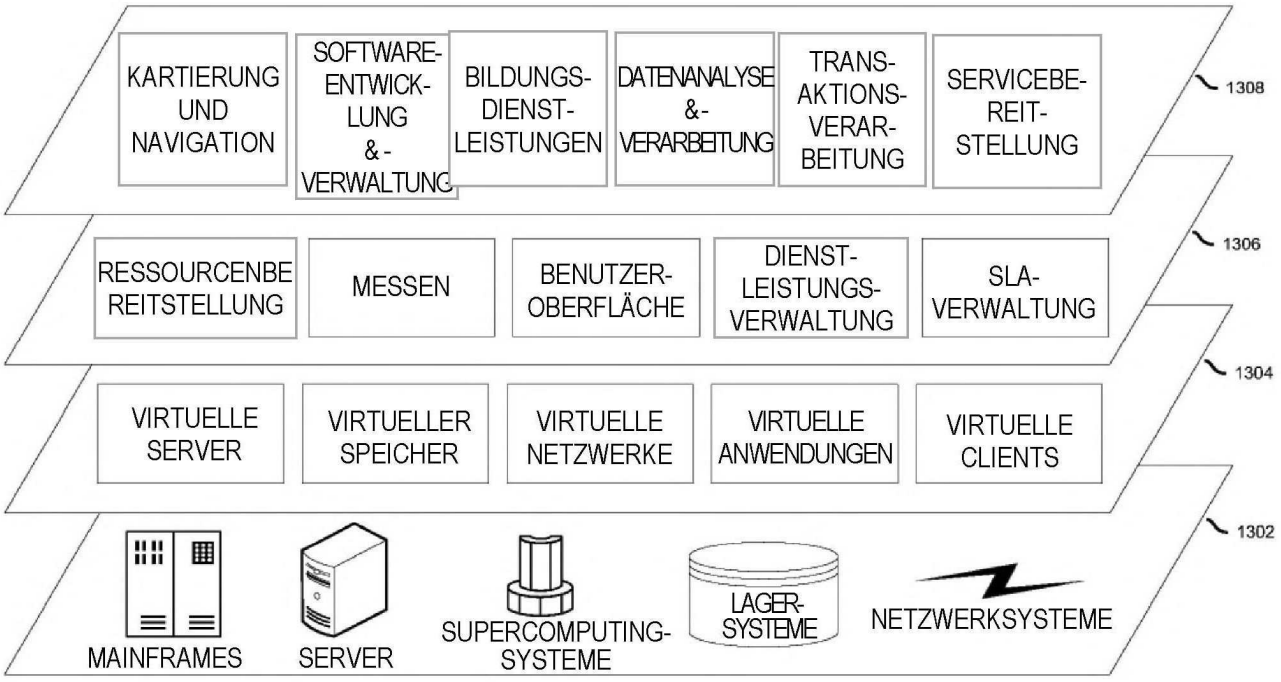

KARTIERUNG UND NAVIGATION
SOFTWARE-ENTWICK-LUNG &- VERWALTUNG
BILDUNGS-DIENST-LEISTUNGEN
DATENANALYSE &- VERARBEITUNG
TRANS-AKTIONS-VERAR-BEITUNG
SERVICEBE-REIT-STELLUNG
1308
RESSOURCENBE REITSTELLUNG
MESSEN
BENUTZER-OBERFLÄCHE
DIENST-LEISTUNGS-VERWALTUNG
SLA-VERWALTUNG
1306
VIRTUELLE SERVER
VIRTUELLER SPEICHER
VIRTUELLE NETZWERKE
VIRTUELLE ANWENDUNGEN
VIRTUELLE CLIENTS
1304
MAINFRAMES
SERVER
SUPERCOMPUTING-SYSTEME
LAGER-SYSTEME
NETZWERKSYSTEME
1302

FIG. 13

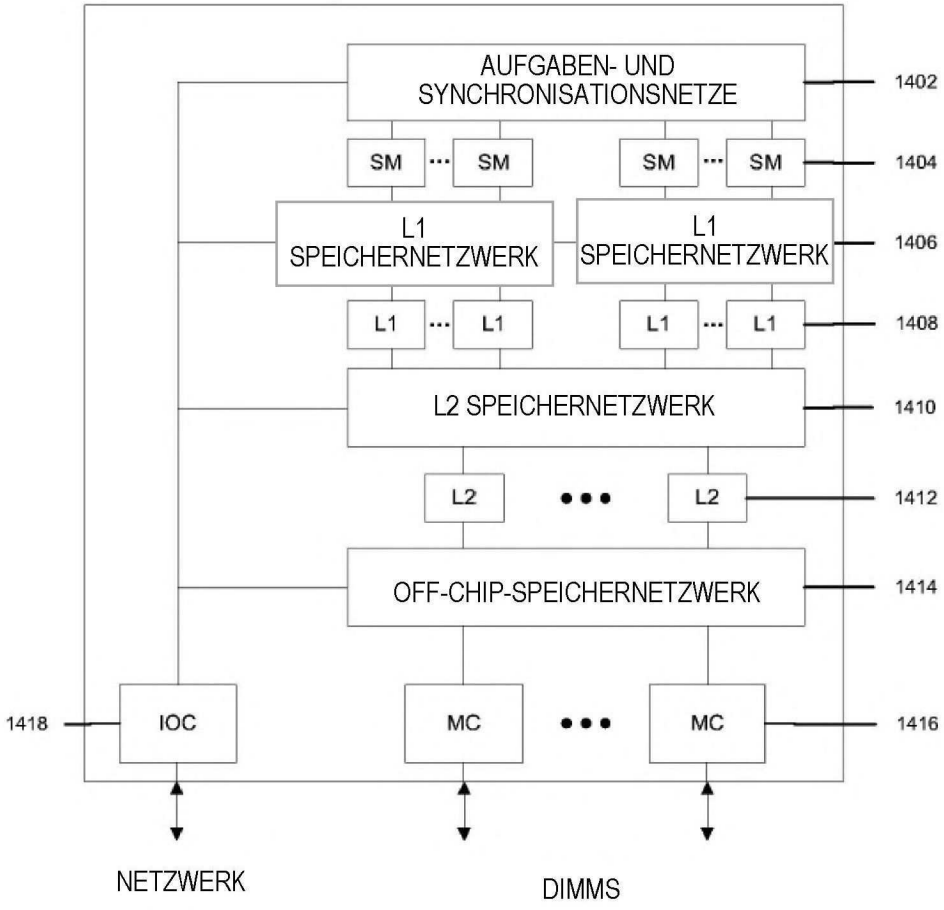

AUFGABEN- UND SYNCHRONISATIONSNETZE
1402
SM
SM
SM
SM
1404
L1 SPEICHERNETZWERK
L1 SPEICHERNETZWERK
1406
L1
L1
L1
L1
1408
L2 SPEICHERNETZWERK
1410
L2
L2
1412
OFF-CHIP-SPEICHERNETZWERK
1414
1418
IOC
MC
MC
1416
NETZWERK
DIMMS

FIG. 14

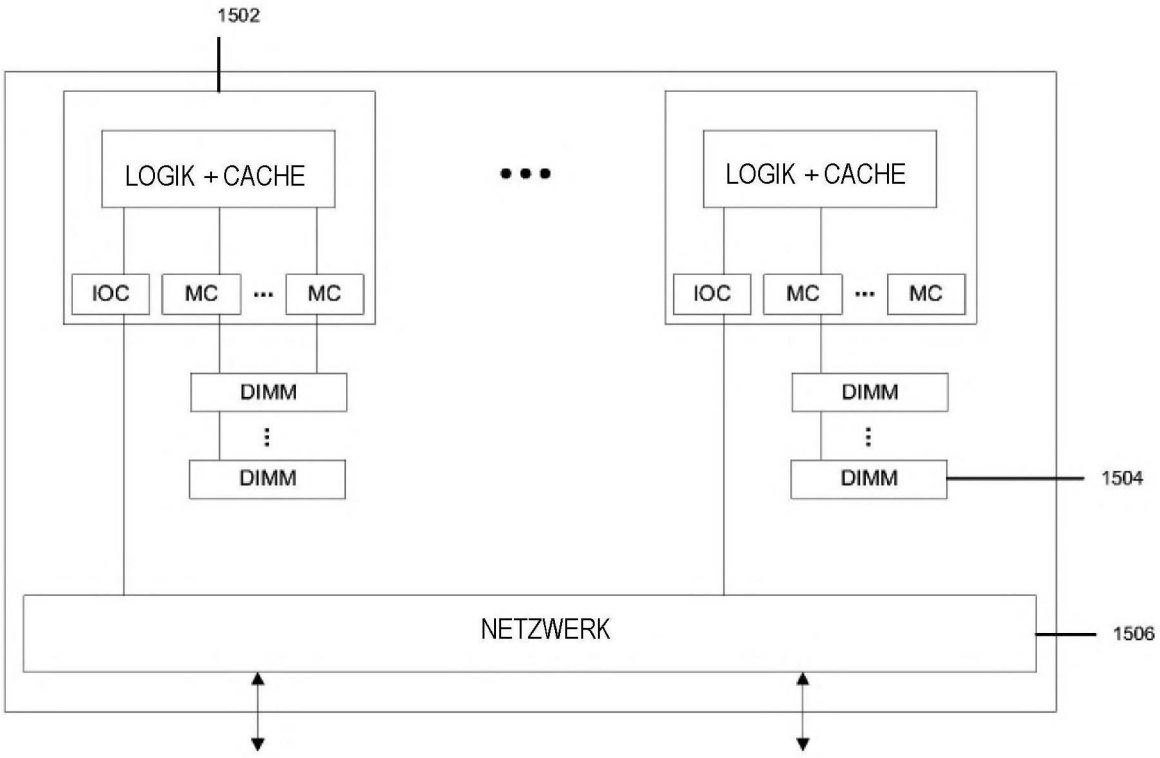

1502
LOGIK + CACHE
IOC
MC
MC
DIMM
DIMM
LOGIK + CACHE
IOC
MC
MC
DIMM
DIMM
1504
NETZWERK
1506

FIG. 15

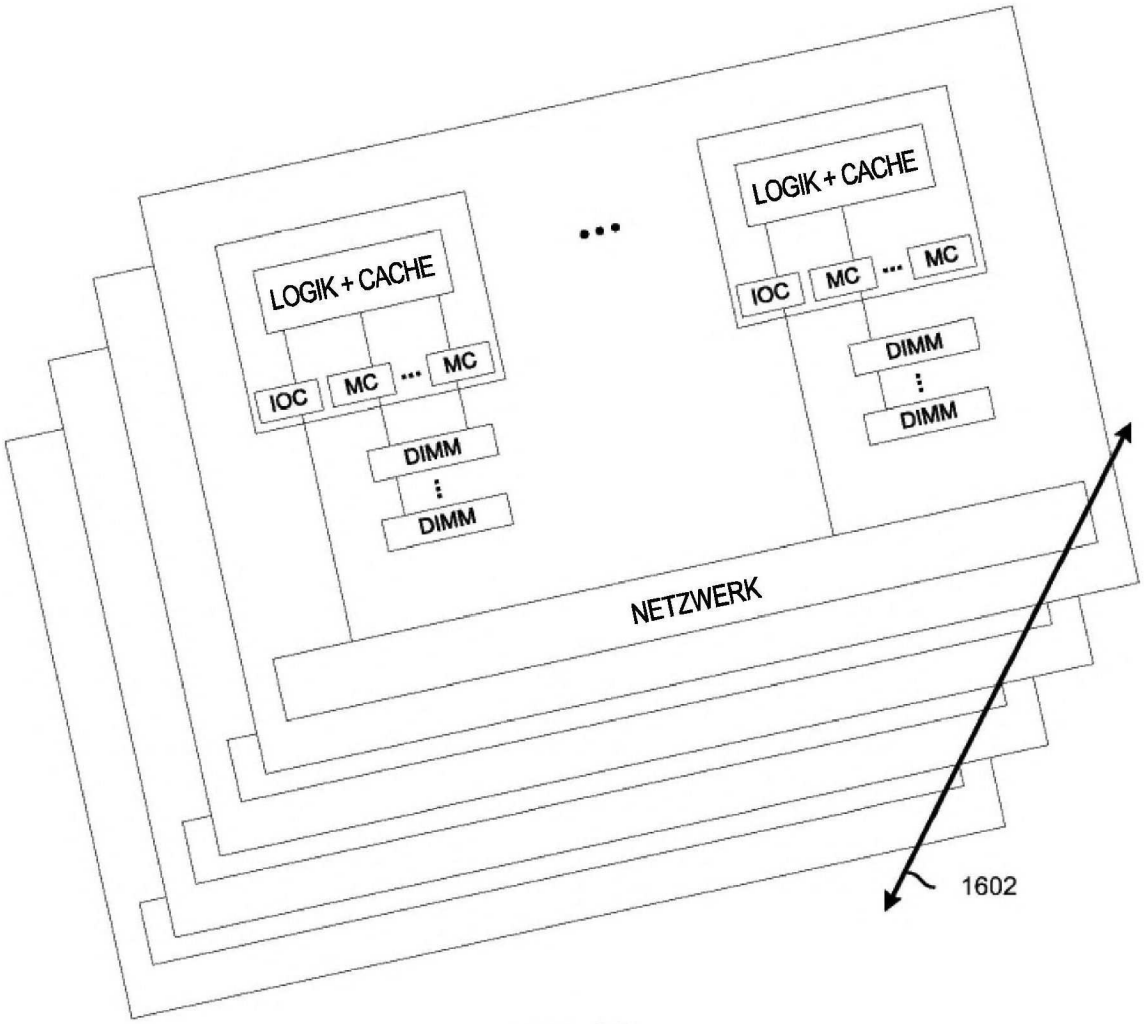

LOGIK + CACHE
IOC
MC
MC
DIMM
DIMM
LOGIK + CACHE
IOC
MC
MC
DIMM
DIMM
NETZWERK
1602

FIG. 16

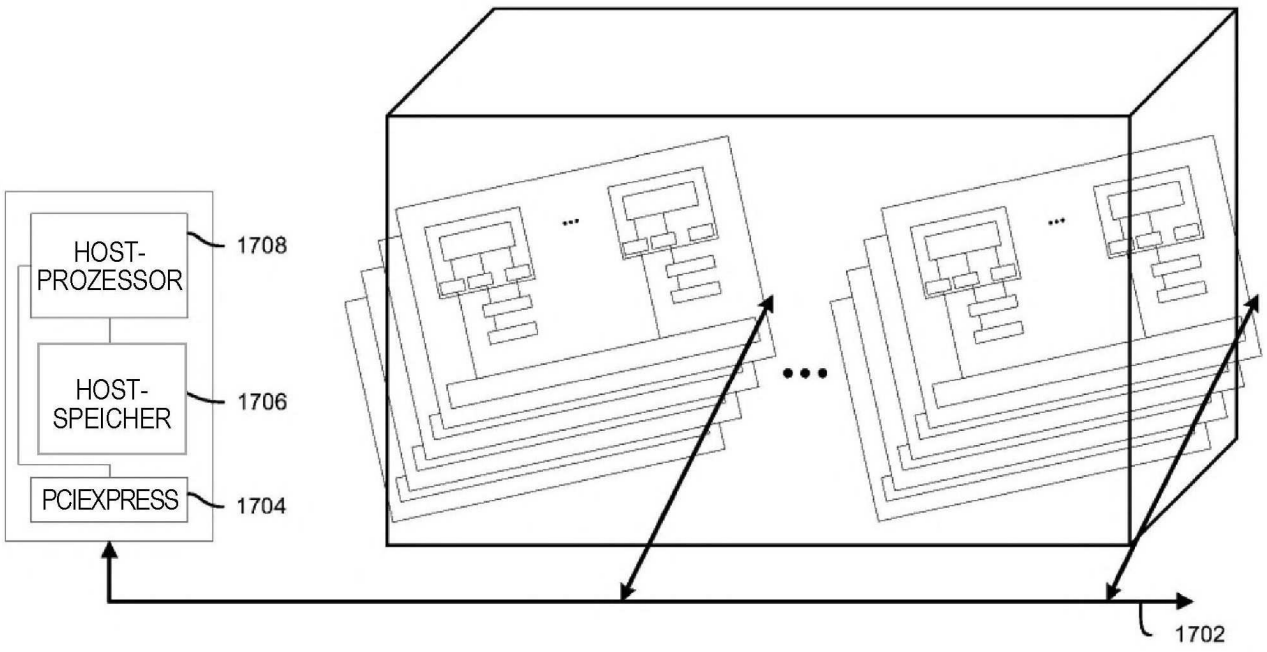

HOST-
PROZESSOR
1708
HOST-
SPEICHER
1706
PCIEXPRESS
1704
1702

FIG. 17

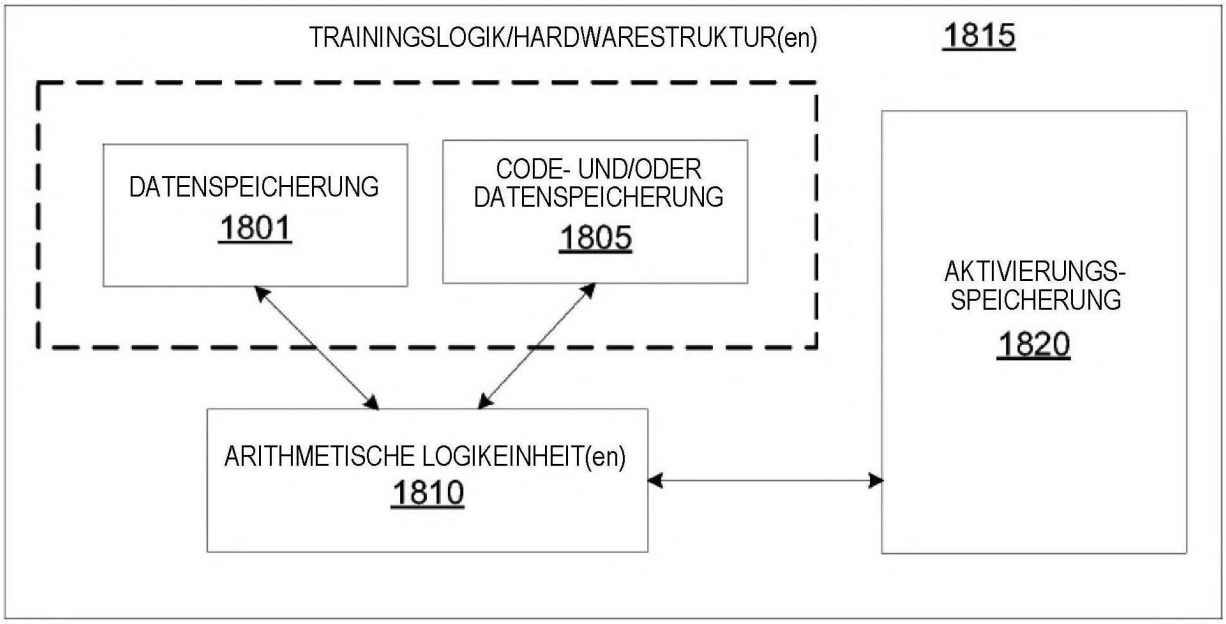


FIG. 18A

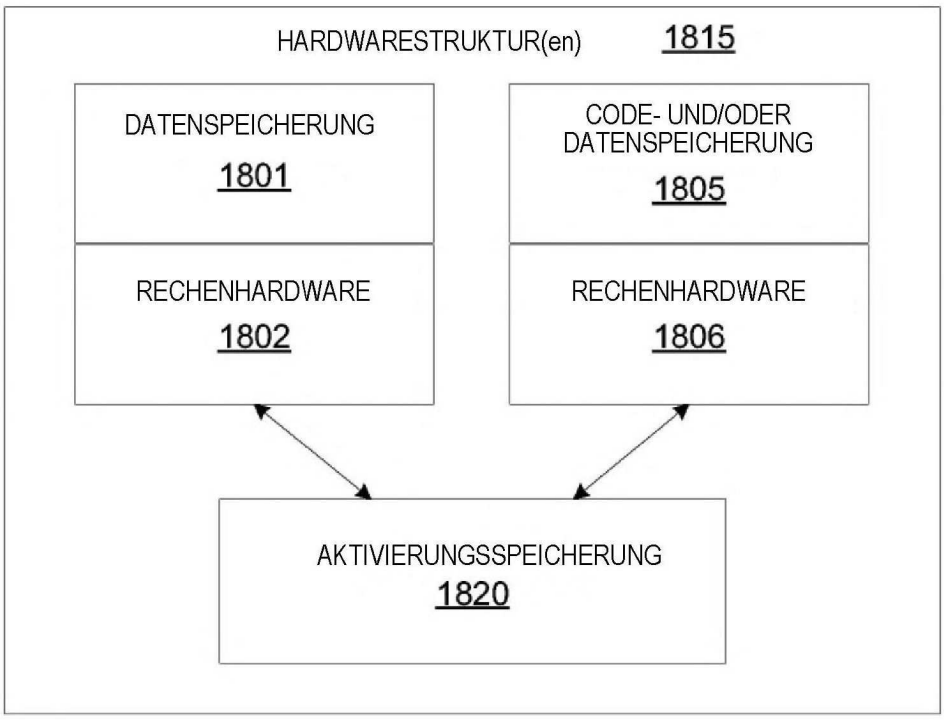


FIG. 18B

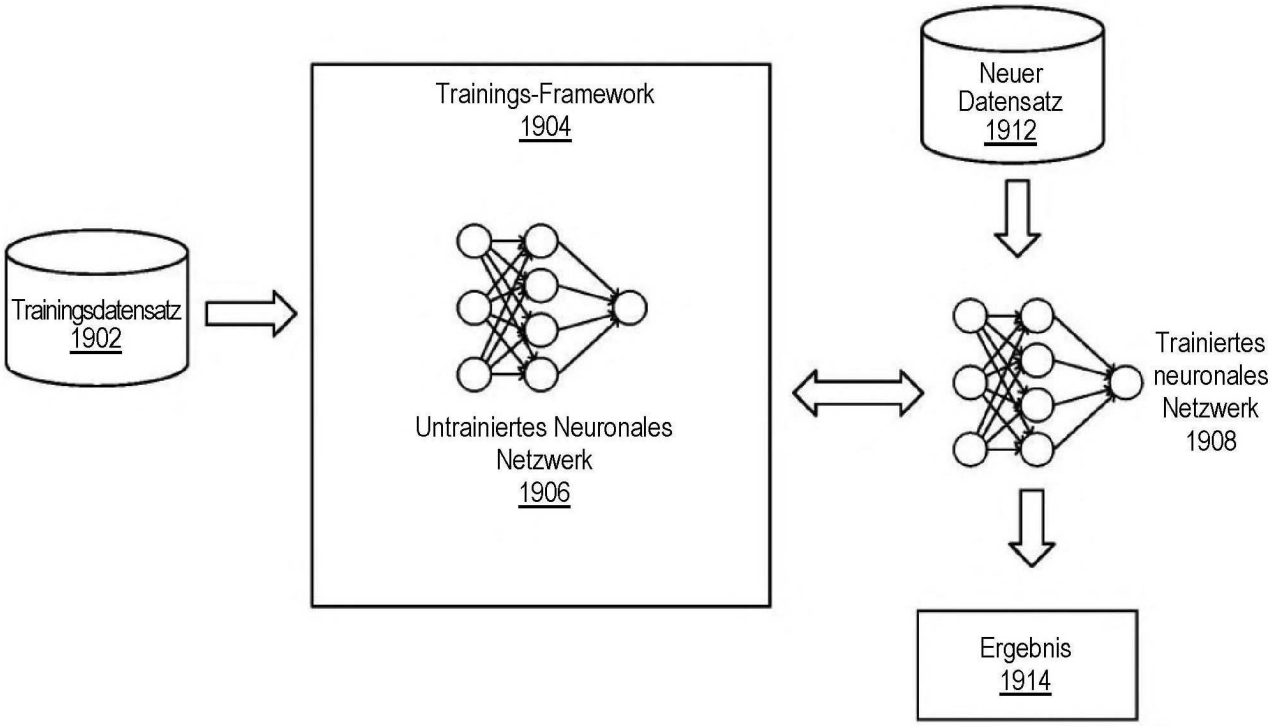

Trainings-Framework
1904
Neuer
Datensatz
1912
Trainingsdatensatz
1902
Untrainiertes Neuronales
Netzwerk
1906
Trainiertes
neuronales
Netzwerk
1908
Ergebnis
1914

FIG. 19

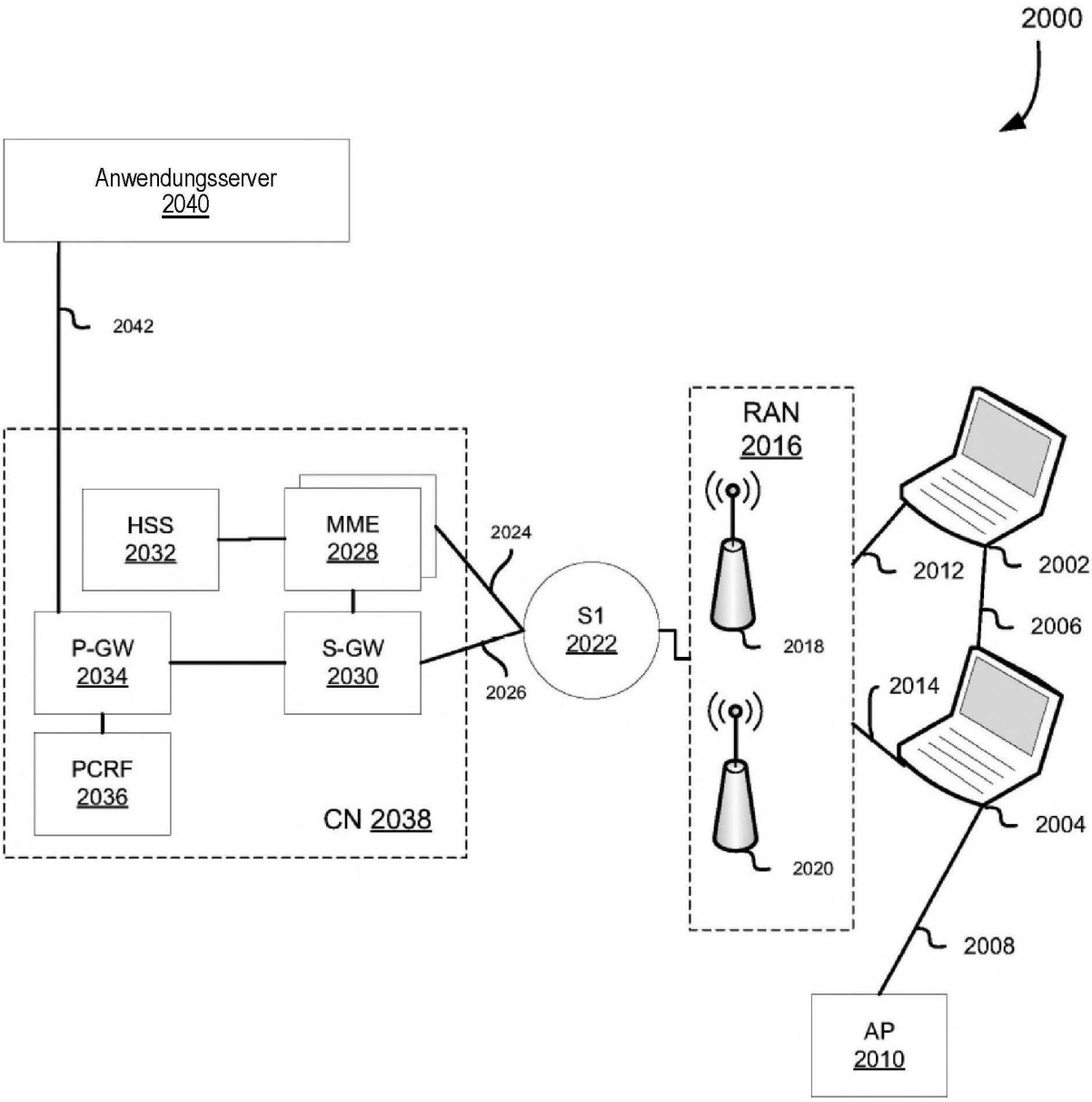

2000
Anwendungsserver
2040
2042
HSS
2032
MME
2028
P-GW
2034
S-GW
2030
PCRF
2036
CN 2038
2024
2026
S1
2022
RAN
2016
2018
2020
2012
2002
2006
2014
2004
2008
AP
2010

FIG. 20

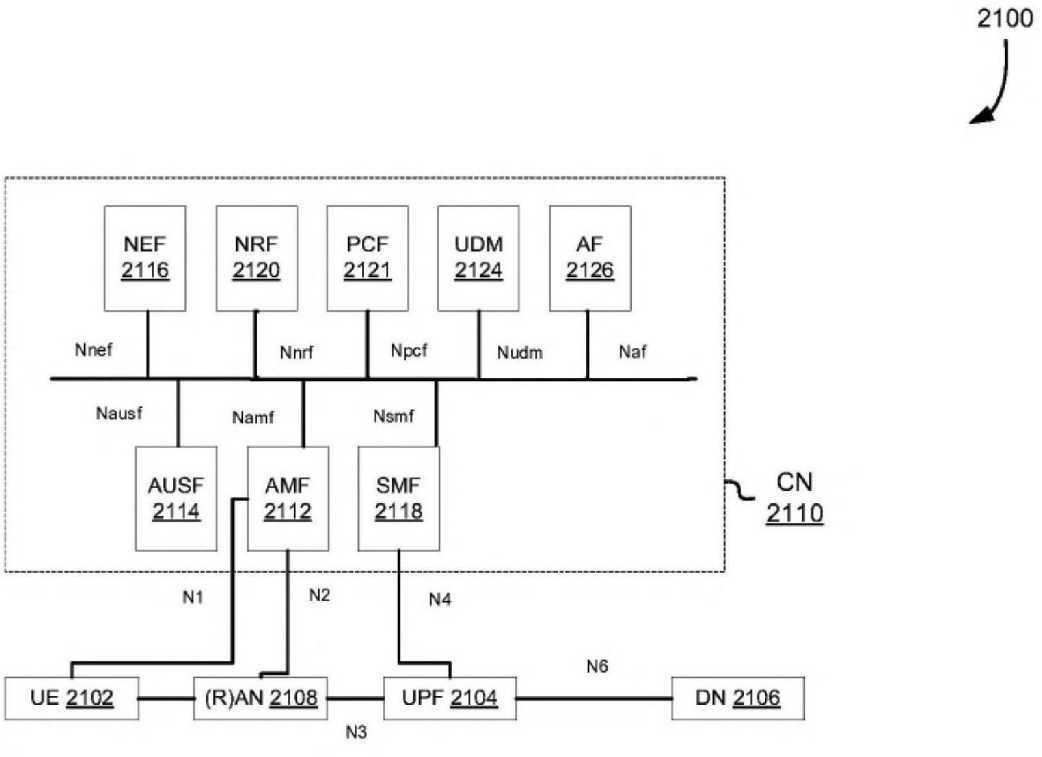

2100
NEF 2116
NRF 2120
PCF 2121
UDM 2124
AF 2126
Nnef
Nnrf
Npcf
Nudm
Naf
Nausf
Namf
Nsmf
AUSF 2114
AMF 2112
SMF 2118
CN 2110
N1
N2
N4
N6
UE 2102
(R)AN 2108
UPF 2104
DN 2106
N3

FIG. 21

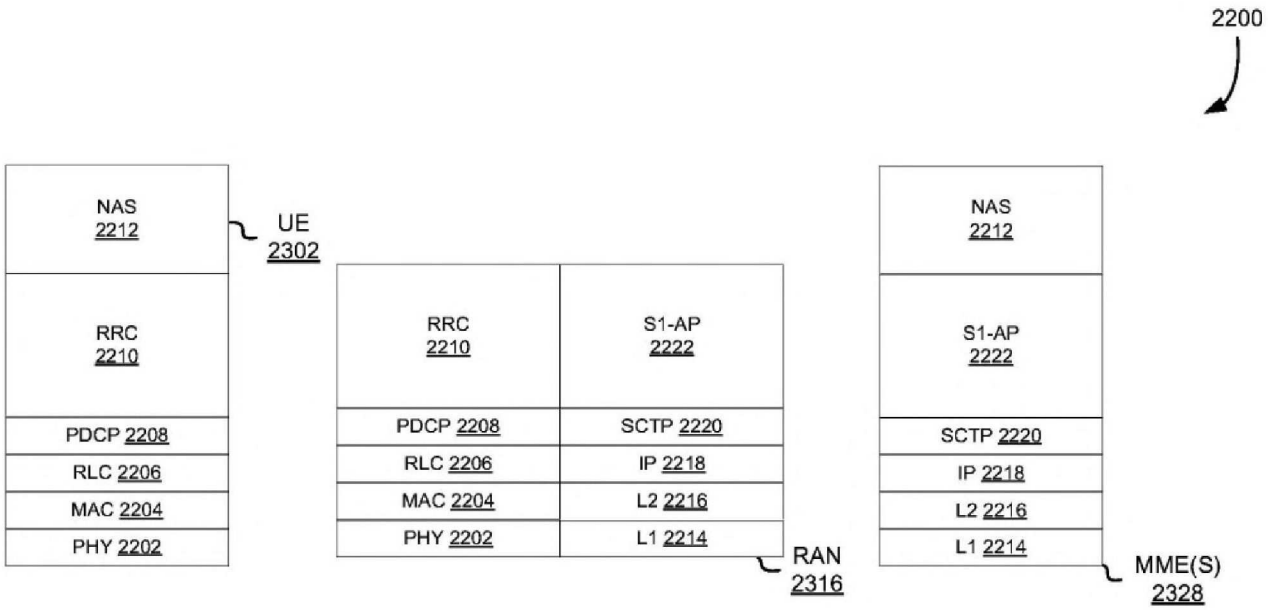

2200
NAS 2212
RRC 2210
PDCP 2208
RLC 2206
MAC 2204
PHY 2202
UE 2302
RRC 2210
S1-AP 2222
PDCP 2208
SCTP 2220
RLC 2206
IP 2218
MAC 2204
L2 2216
PHY 2202
L1 2214
RAN 2316
NAS 2212
S1-AP 2222
SCTP 2220
IP 2218
L2 2216
L1 2214
MME(S) 2328

FIG. 22

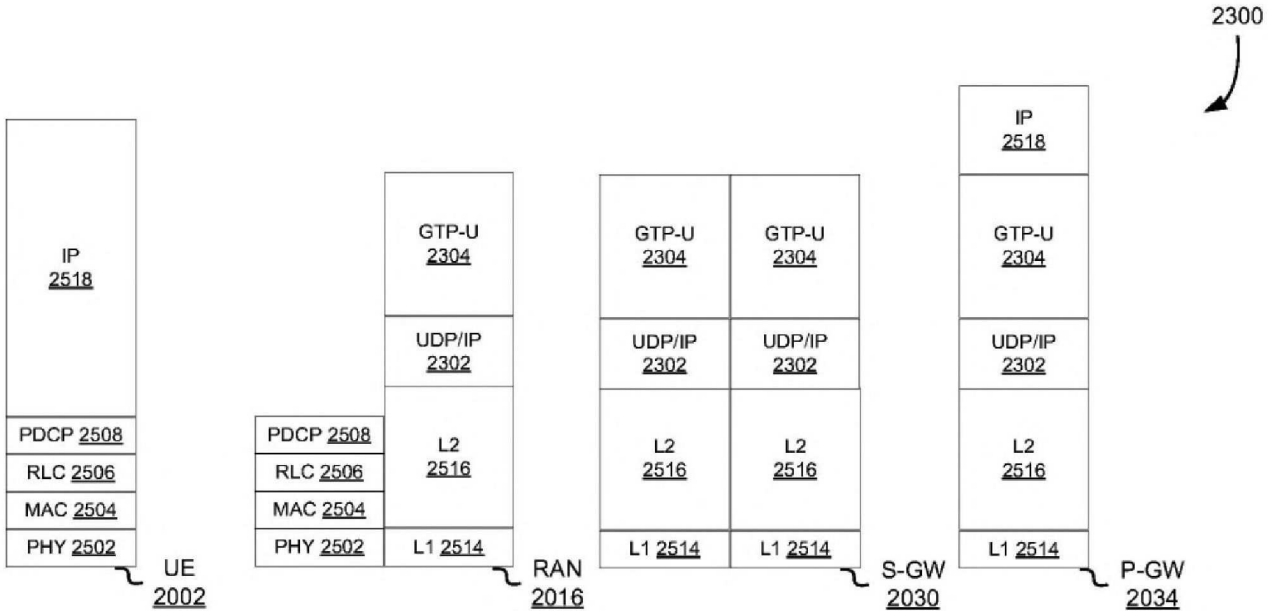

2300
IP
2518
PDCP 2508
RLC 2506
MAC 2504
PHY 2502
UE
2002
GTP-U
2304
UDP/IP
2302
L2
2516
PDCP 2508
RLC 2506
MAC 2504
PHY 2502
L1 2514
RAN
2016
GTP-U
2304
UDP/IP
2302
L2
2516
L1 2514
GTP-U
2304
UDP/IP
2302
L2
2516
L1 2514
S-GW
2030
IP
2518
GTP-U
2304
UDP/IP
2302
L2
2516
L1 2514
P-GW
2034

FIG. 23

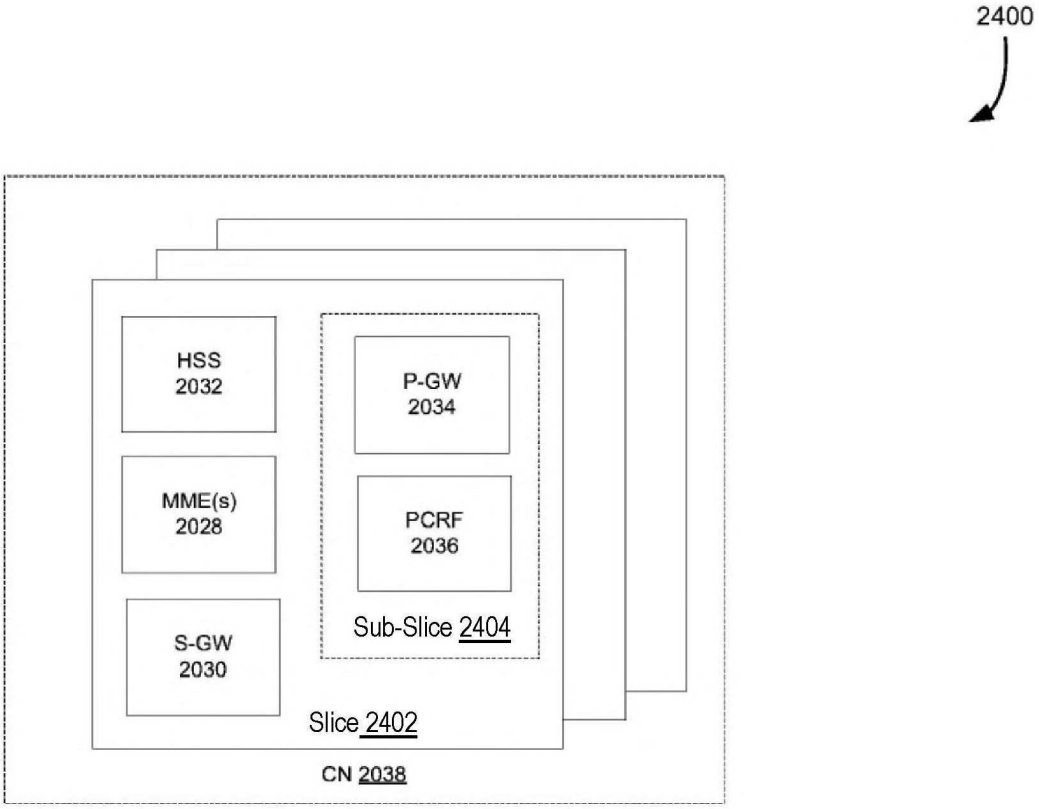

2400
HSS
2032
MME(s)
2028
S-GW
2030
P-GW
2034
PCRF
2036
Sub-Slice 2404
Slice 2402
CN 2038

FIG. 24

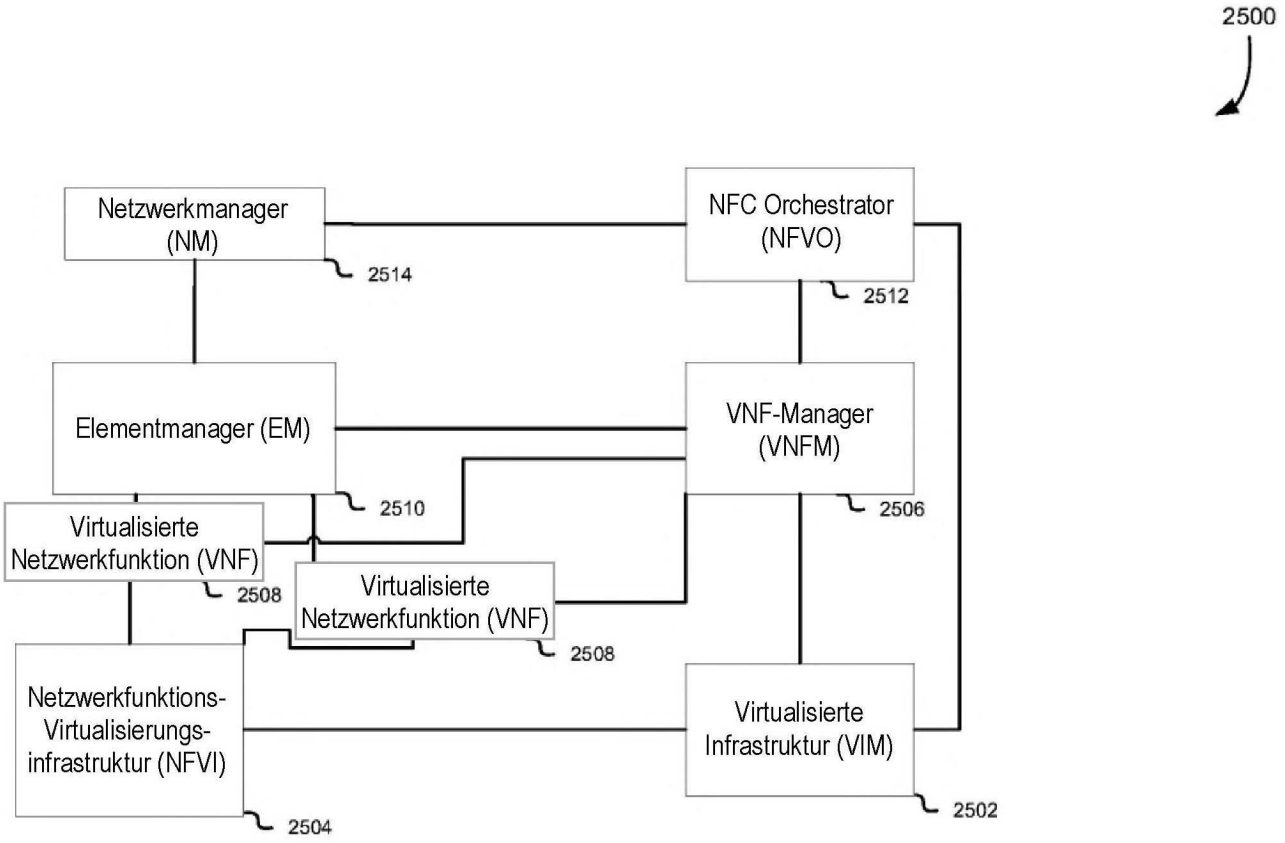

2500
Netzwerkmanager (NM)
2514
NFC Orchestrator (NFVO)
2512
Elementmanager (EM)
2510
VNF-Manager (VNFM)
2506
Virtualisierte Netzwerkfunktion (VNF)
2508
Virtualisierte Netzwerkfunktion (VNF)
2508
Netzwerkfunktions-Virtualisierungs-infrastruktur (NFVI)
2504
Virtualisierte Infrastruktur (VIM)
2502

FIG. 25

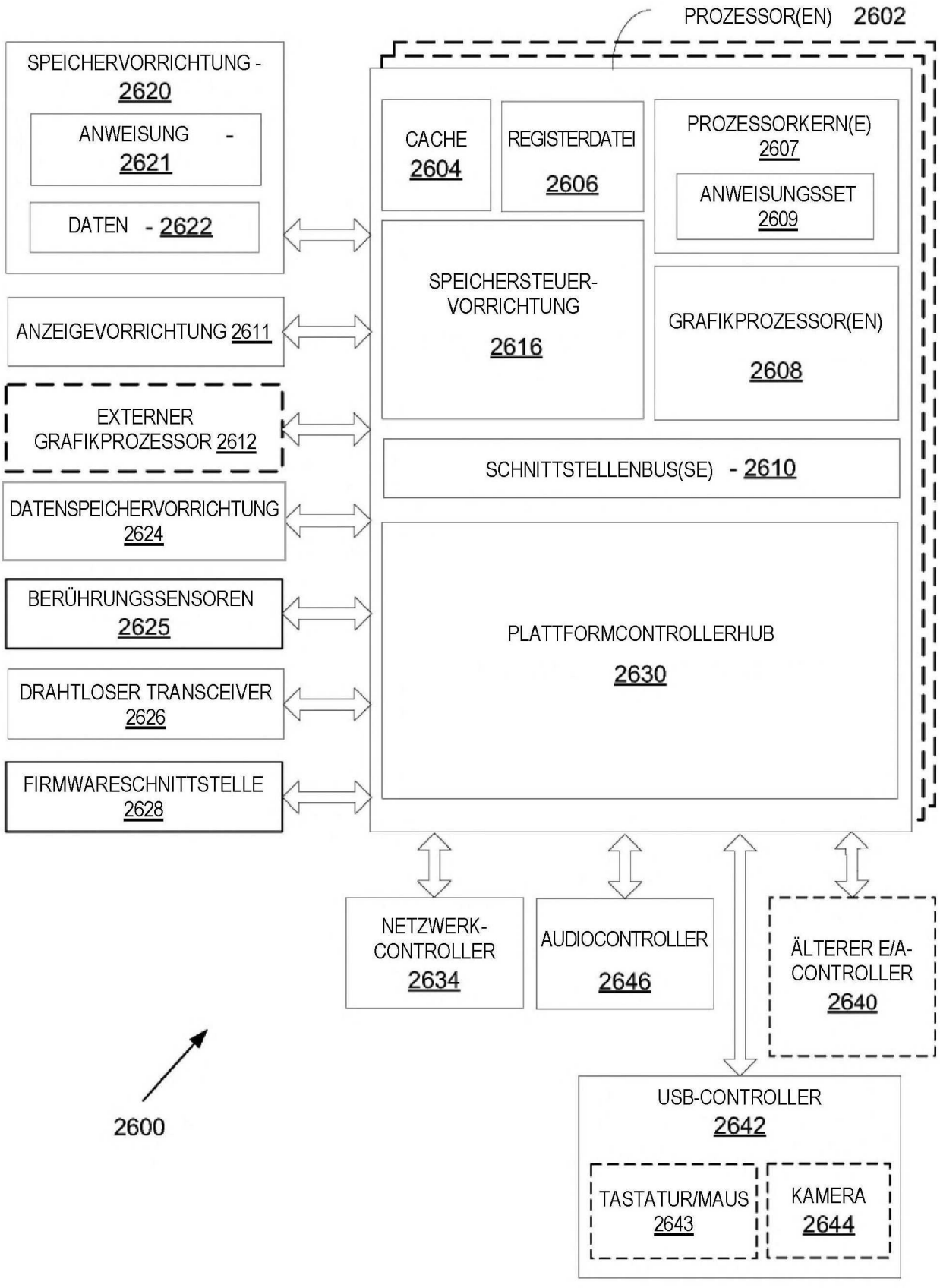

PROZESSOR(EN) 2602
SPEICHERVORRICHTUNG - 2620
ANWEISUNG - 2621
DATEN - 2622
CACHE 2604
REGISTERDATEI 2606
PROZESSORKERN(E) 2607
ANWEISUNGSSET 2609
SPEICHERSTEUER-VORRICHTUNG 2616
GRAFIKPROZESSOR(EN) 2608
ANZEIGEVORRICHTUNG 2611
EXTERNER GRAFIKPROZESSOR 2612
SCHNITTSTELLENBUS(SE) - 2610
DATENSPEICHERVORRICHTUNG 2624
BERÜHRUNGSSENSOREN 2625
PLATTFORMCONTROLLERHUB 2630
DRAHTLOSER TRANSCEIVER 2626
FIRMWARESCHNITTSTELLE 2628
NETZWERK-CONTROLLER 2634
AUDIOCONTROLLER 2646
ÄLTERER E/A-CONTROLLER 2640
USB-CONTROLLER 2642
TASTATUR/MAUS 2643
KAMERA 2644
2600

FIG. 26

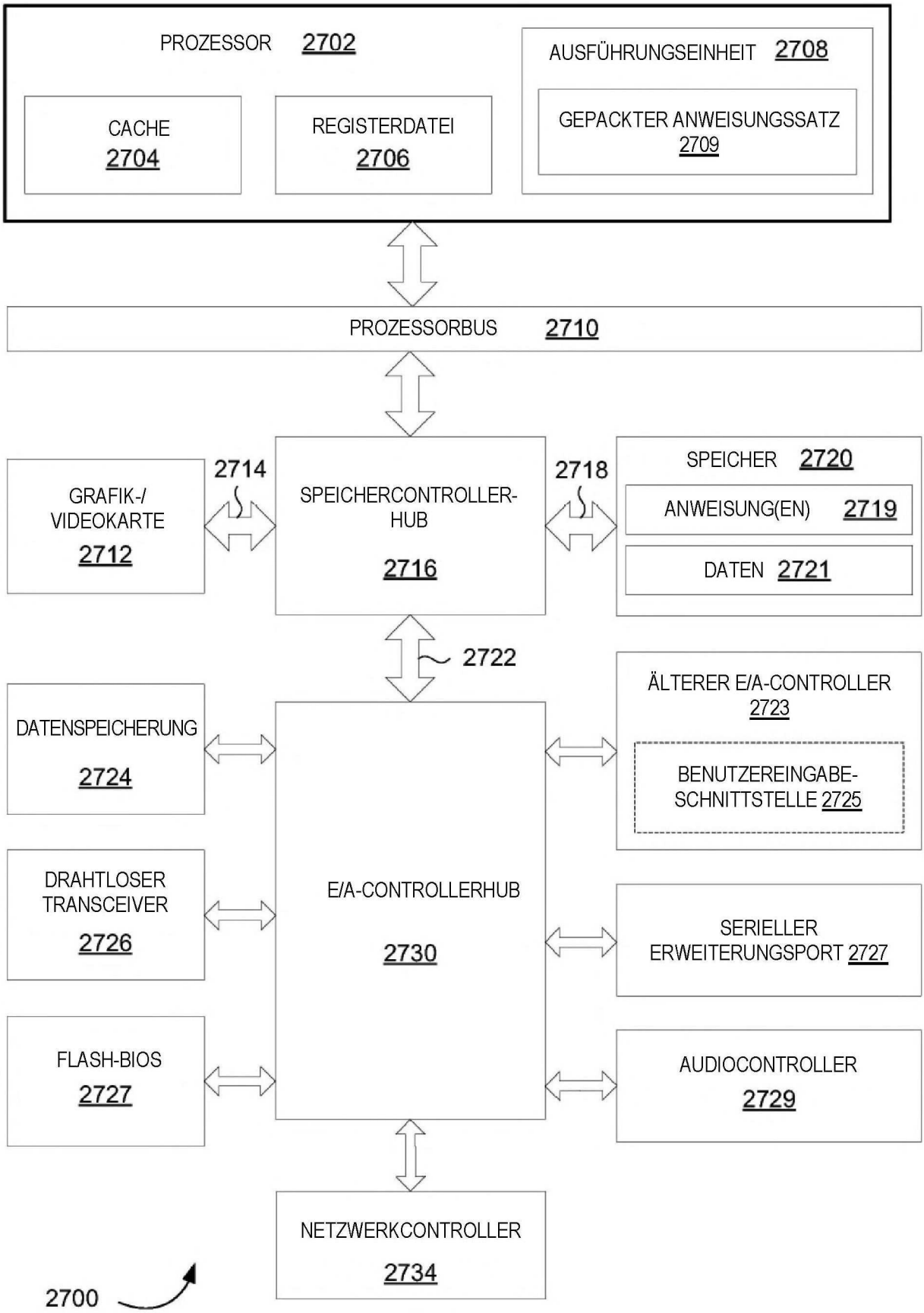

PROZESSOR 2702
CACHE 2704
REGISTERDATEI 2706
AUSFÜHRUNGSEINHEIT 2708
GEPACKTER ANWEISUNGSSATZ 2709
PROZESSORBUS 2710
GRAFIK-/ VIDEOKARTE 2712
2714
SPEICHERCONTROLLER-HUB 2716
2718
SPEICHER 2720
ANWEISUNG(EN) 2719
DATEN 2721
2722
DATENSPEICHERUNG 2724
ÄLTERER E/A-CONTROLLER 2723
BENUTZEREINGABE-SCHNITTSTELLE 2725
DRAHTLOSER TRANSCEIVER 2726
E/A-CONTROLLERHUB 2730
SERIELLER ERWEITERUNGSPORT 2727
FLASH-BIOS 2727
AUDIOCONTROLLER 2729
NETZWERKCONTROLLER 2734
2700

FIG. 27

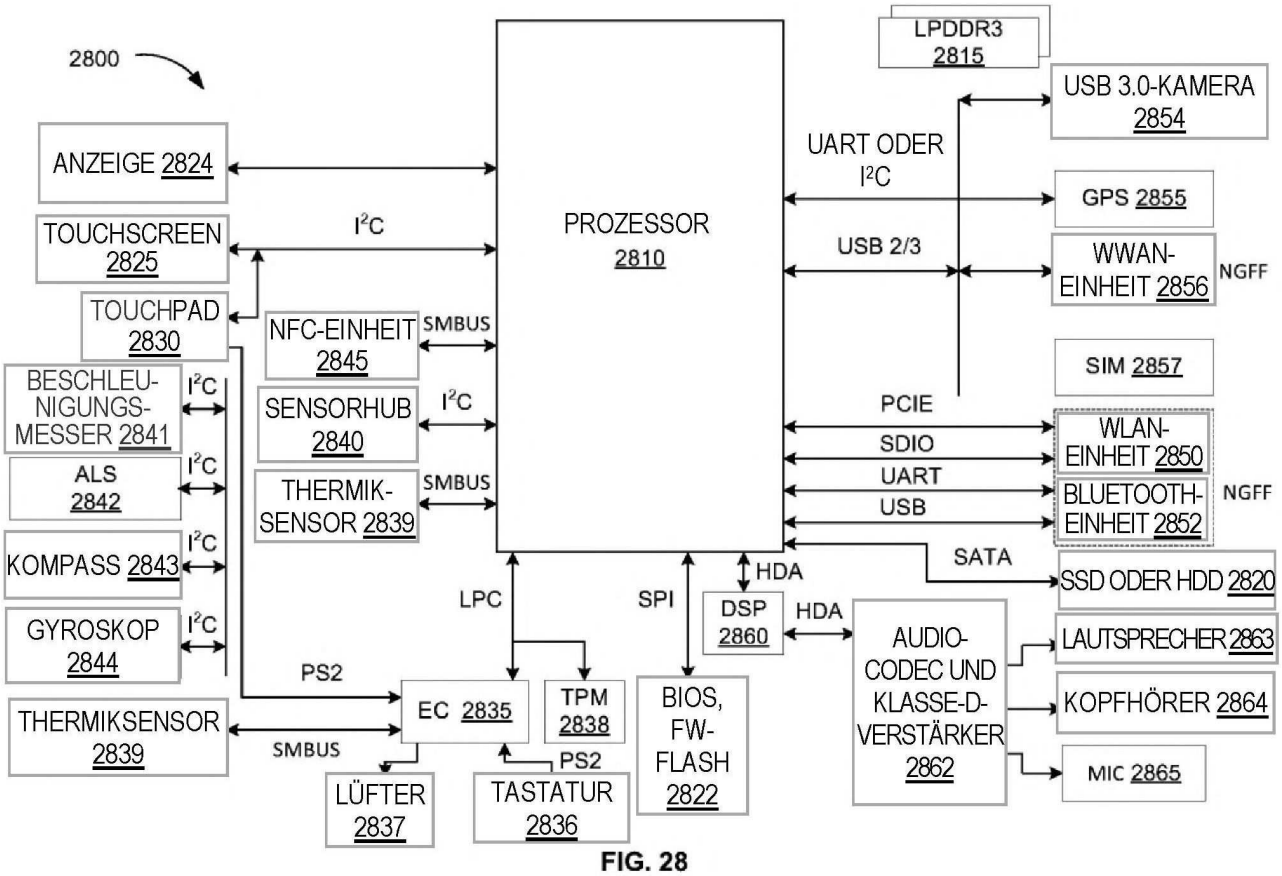

2800
ANZEIGE 2824
TOUCHSCREEN 2825
TOUCHPAD 2830
BESCHLEU-NIGUNGS-MESSER 2841
ALS 2842
KOMPASS 2843
GYROSKOP 2844
THERMIKSENSOR 2839
I²C
SMBUS
PS2
NFC-EINHEIT 2845
SENSORHUB 2840
THERMIK-SENSOR 2839
EC 2835
LÜFTER 2837
TASTATUR 2836
TPM 2838
PROZESSOR 2810
LPC
SPI
BIOS, FW-FLASH 2822
DSP 2860
HDA
AUDIO-CODEC UND KLASSE-D-VERSTÄRKER 2862
LPDDR3 2815
UART ODER I²C
USB 2/3
PCIE
SDIO
UART
USB
SATA
USB 3.0-KAMERA 2854
GPS 2855
WWAN-EINHEIT 2856
NGFF
SIM 2857
WLAN-EINHEIT 2850
BLUETOOTH-EINHEIT 2852
SSD ODER HDD 2820
LAUTSPRECHER 2863
KOPFHÖRER 2864
MIC 2865

FIG. 28

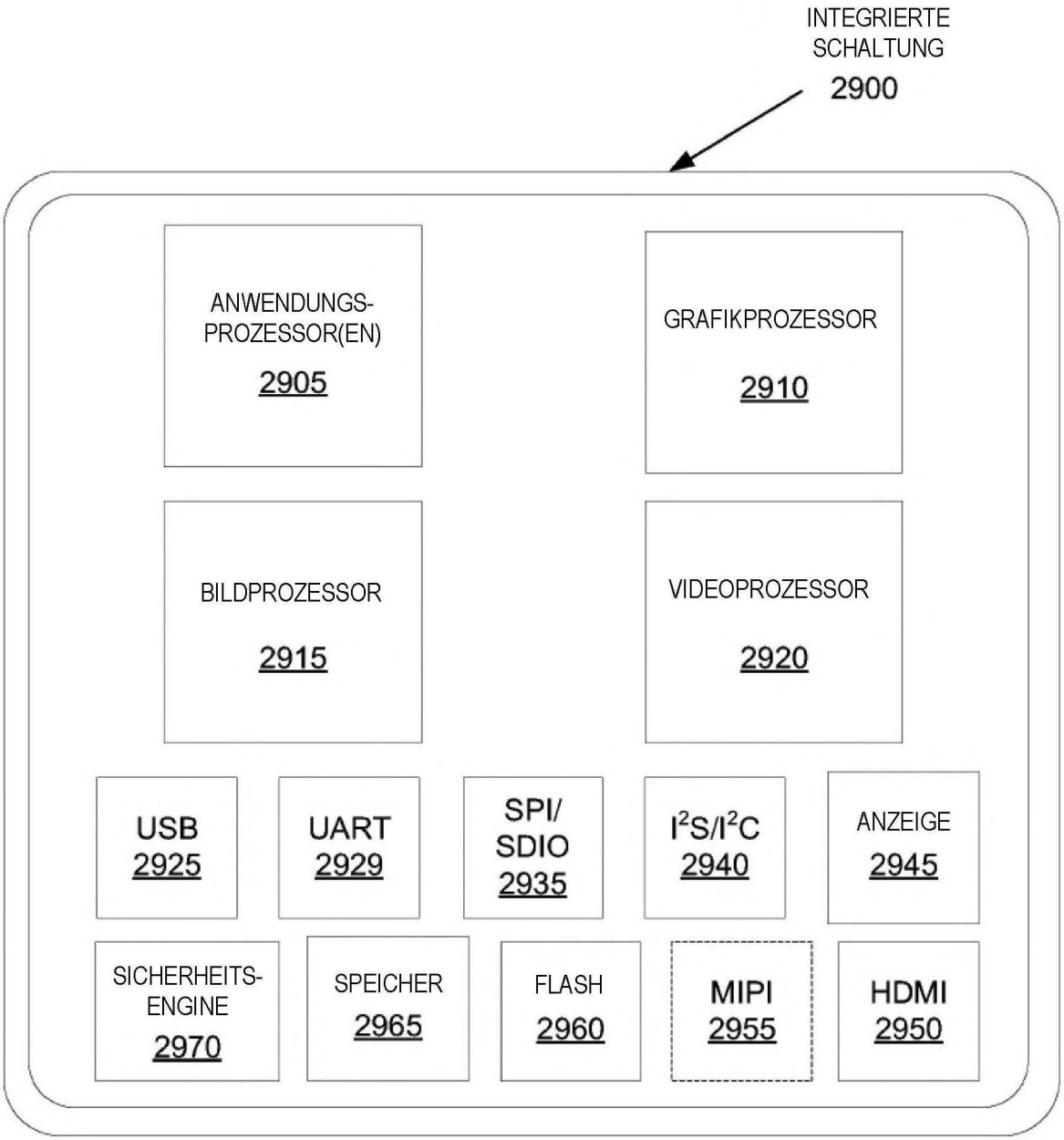

INTEGRIERTE SCHALTUNG
2900
ANWENDUNGS-PROZESSOR(EN)
2905
GRAFIKPROZESSOR
2910
BILDPROZESSOR
2915
VIDEOPROZESSOR
2920
USB
2925
UART
2929
SPI/ SDIO
2935
I²S/I²C
2940
ANZEIGE
2945
SICHERHEITS-ENGINE
2970
SPEICHER
2965
FLASH
2960
MIPI
2955
HDMI
2950

FIG. 29

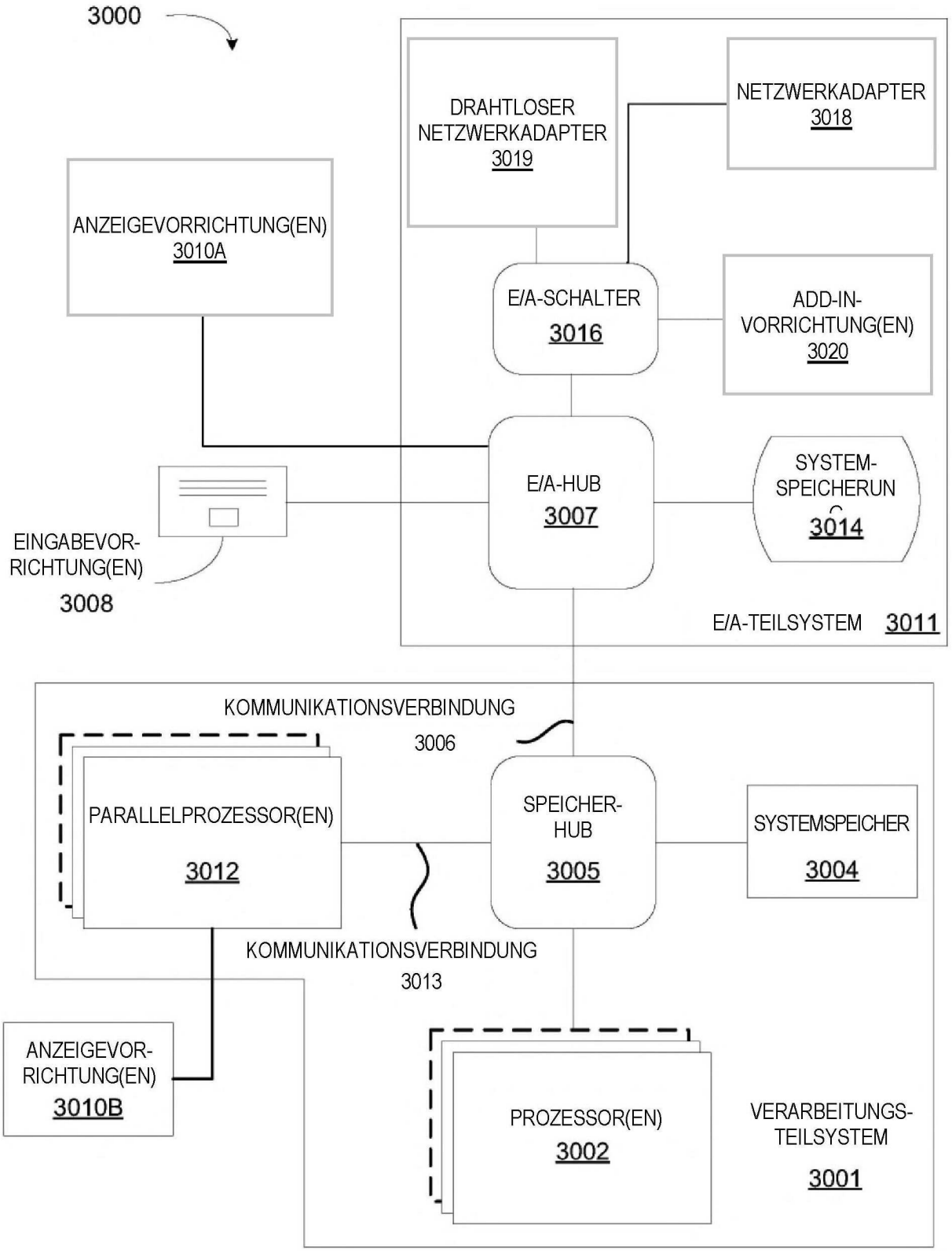

3000
DRAHTLOSER NETZWERKADAPTER 3019
NETZWERKADAPTER 3018
ANZEIGEVORRICHTUNG(EN) 3010A
E/A-SCHALTER 3016
ADD-IN-VORRICHTUNG(EN) 3020
E/A-HUB 3007
SYSTEM-SPEICHERUN 3014
EINGABEVOR-RICHTUNG(EN) 3008
E/A-TEILSYSTEM 3011
KOMMUNIKATIONSVERBINDUNG 3006
PARALLELPROZESSOR(EN) 3012
SPEICHER-HUB 3005
SYSTEMSPEICHER 3004
KOMMUNIKATIONSVERBINDUNG 3013
ANZEIGEVOR-RICHTUNG(EN) 3010B
PROZESSOR(EN) 3002
VERARBEITUNGS-TEILSYSTEM 3001

FIG. 30

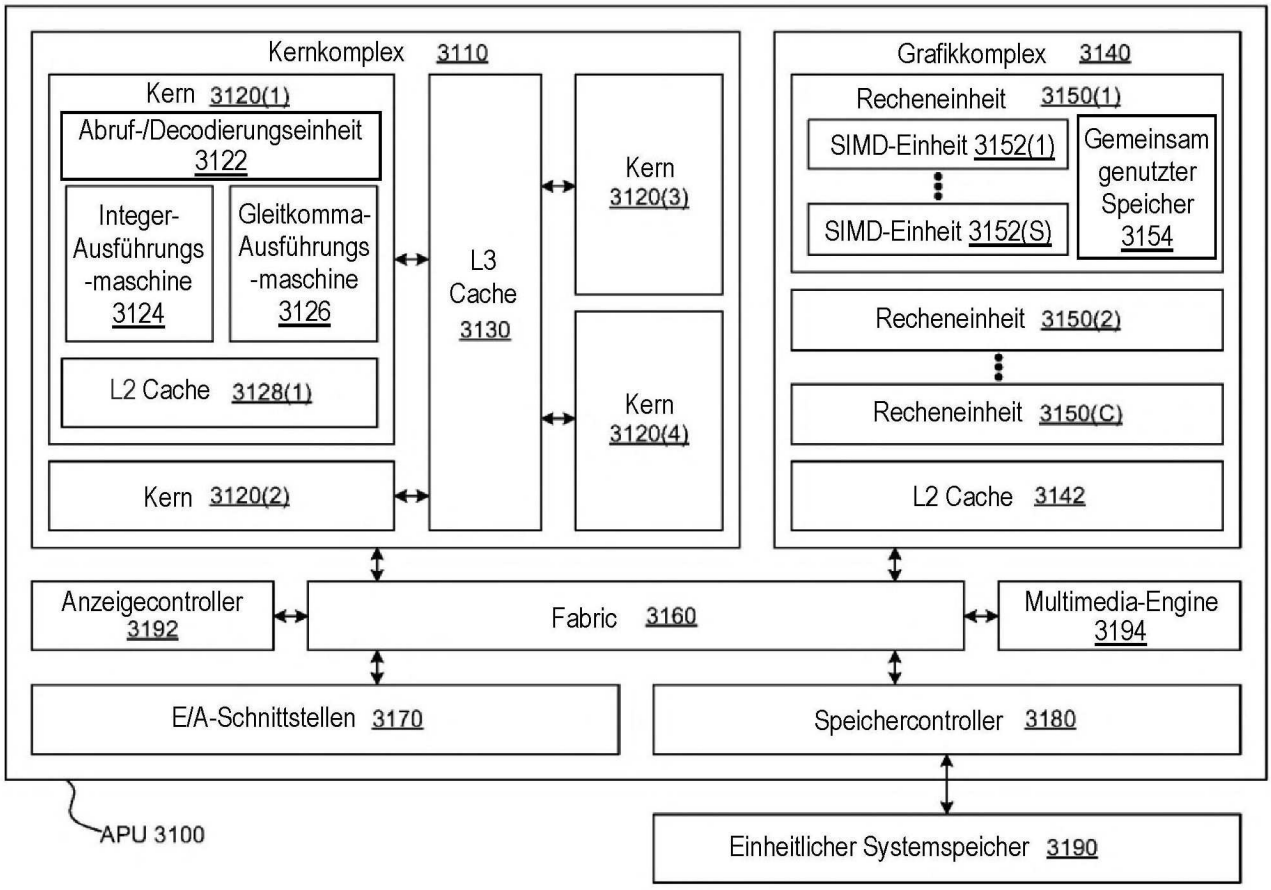

Kernkomplex 3110
Kern 3120(1)
Abruf-/Decodierungseinheit 3122
Integer-Ausführungs-maschine 3124
Gleitkomma-Ausführungs-maschine 3126
L2 Cache 3128(1)
Kern 3120(2)
L3 Cache 3130
Kern 3120(3)
Kern 3120(4)
Grafikkomplex 3140
Recheneinheit 3150(1)
SIMD-Einheit 3152(1)
SIMD-Einheit 3152(S)
Gemeinsam genutzter Speicher 3154
Recheneinheit 3150(2)
Recheneinheit 3150(C)
L2 Cache 3142
Anzeigecontroller 3192
Fabric 3160
Multimedia-Engine 3194
E/A-Schnittstellen 3170
Speichercontroller 3180
APU 3100
Einheitlicher Systemspeicher 3190

FIG. 31

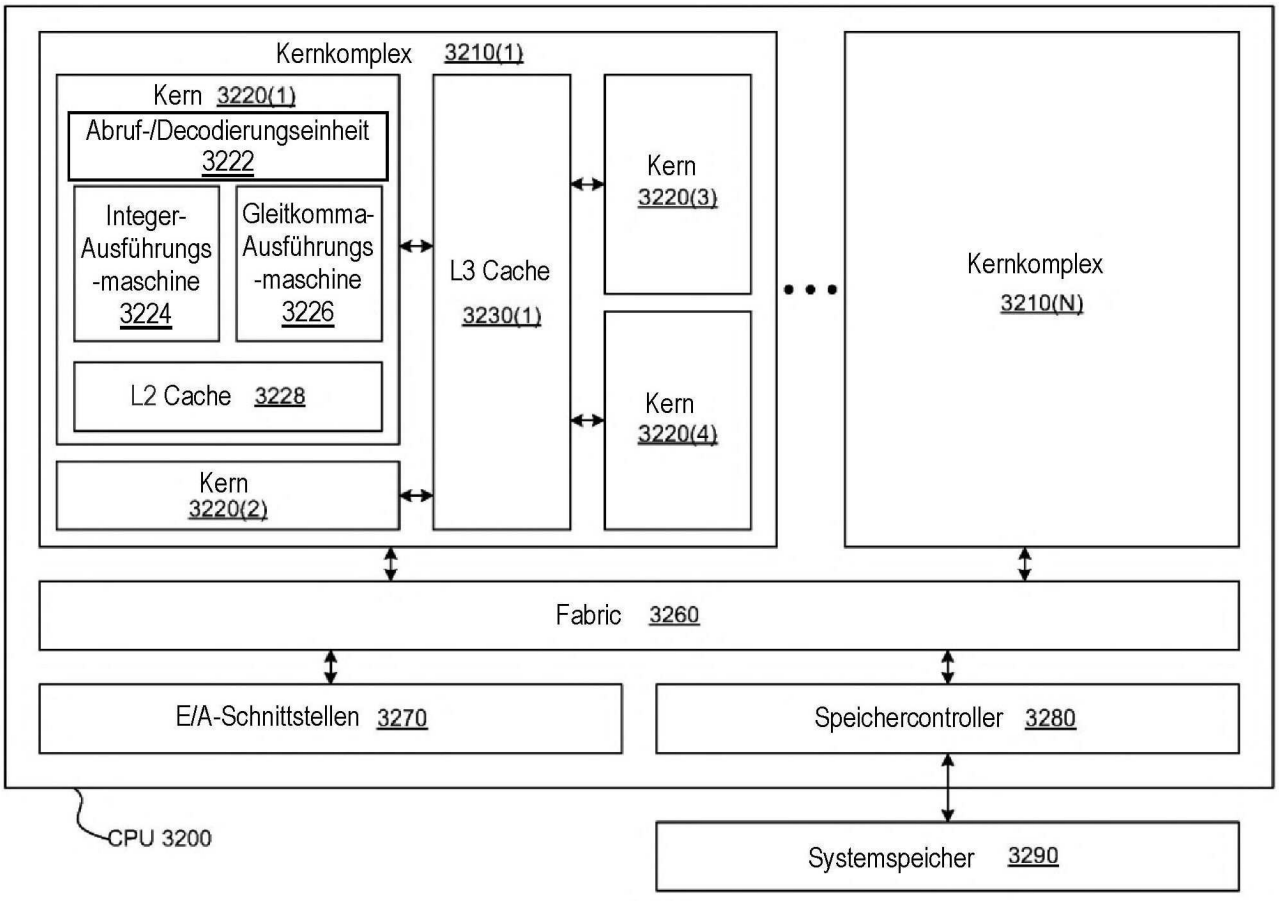

Kernkomplex 3210(1)
Kern 3220(1)
Abruf-/Decodierungseinheit 3222
Integer-Ausführungs-maschine 3224
Gleitkomma-Ausführungs-maschine 3226
L2 Cache 3228
Kern 3220(2)
L3 Cache 3230(1)
Kern 3220(3)
Kern 3220(4)
Kernkomplex 3210(N)
Fabric 3260
E/A-Schnittstellen 3270
Speichercontroller 3280
CPU 3200
Systemspeicher 3290

FIG. 32

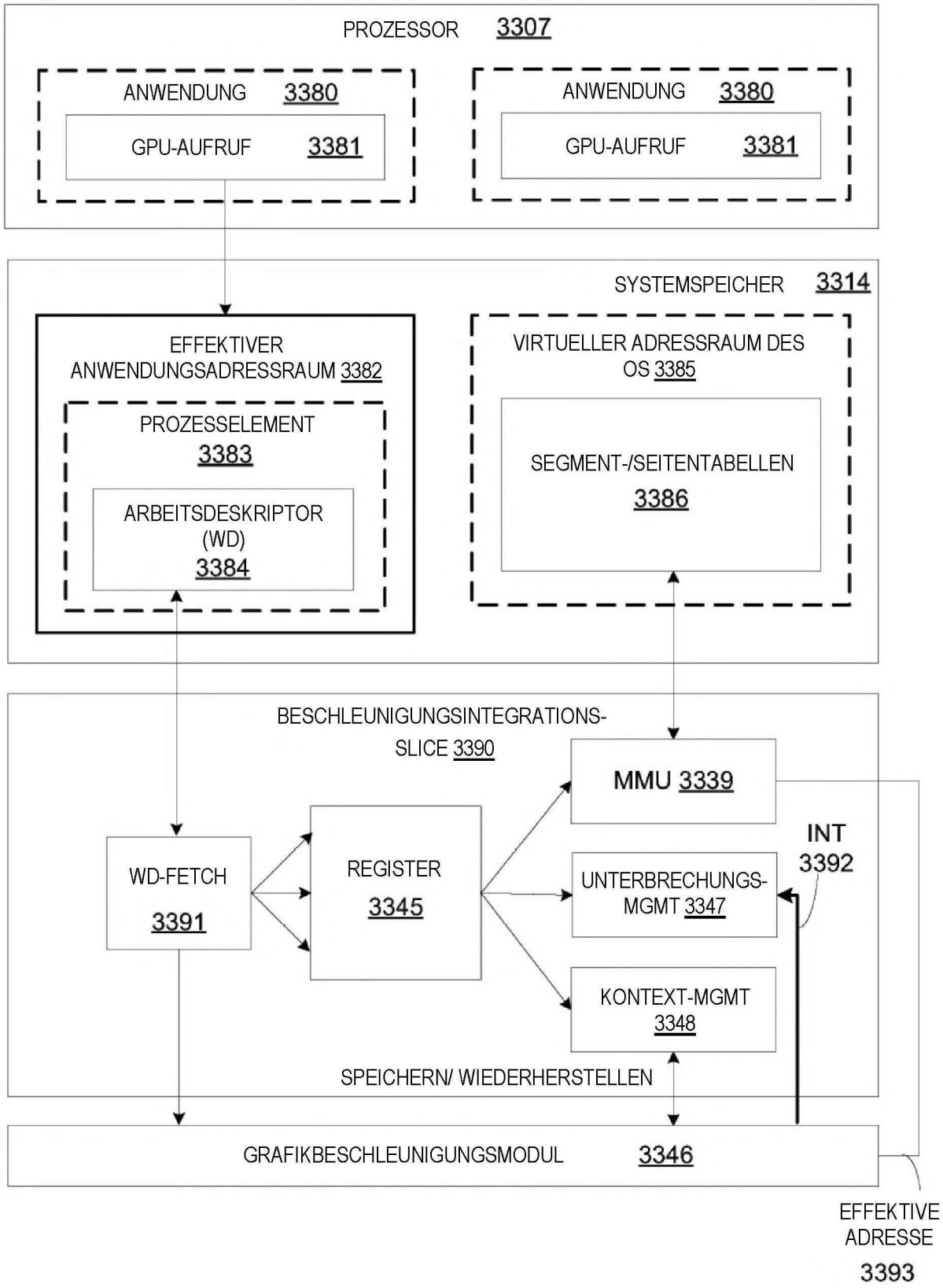

PROZESSOR 3307
ANWENDUNG 3380
GPU-AUFRUF 3381
ANWENDUNG 3380
GPU-AUFRUF 3381
SYSTEMSPEICHER 3314
EFFEKTIVER ANWENDUNGSADRESSRAUM 3382
PROZESSELEMENT 3383
ARBEITSDESKRIPTOR (WD) 3384
VIRTUELLER ADRESSRAUM DES OS 3385
SEGMENT-/SEITENTABELLEN 3386
BESCHLEUNIGUNGSINTEGRATIONS-SLICE 3390
MMU 3339
INT 3392
WD-FETCH 3391
REGISTER 3345
UNTERBRECHUNGS-MGMT 3347
KONTEXT-MGMT 3348
SPEICHERN/ WIEDERHERSTELLEN
GRAFIKBESCHLEUNIGUNGSMODUL 3346
EFFEKTIVE ADRESSE 3393

FIG. 33

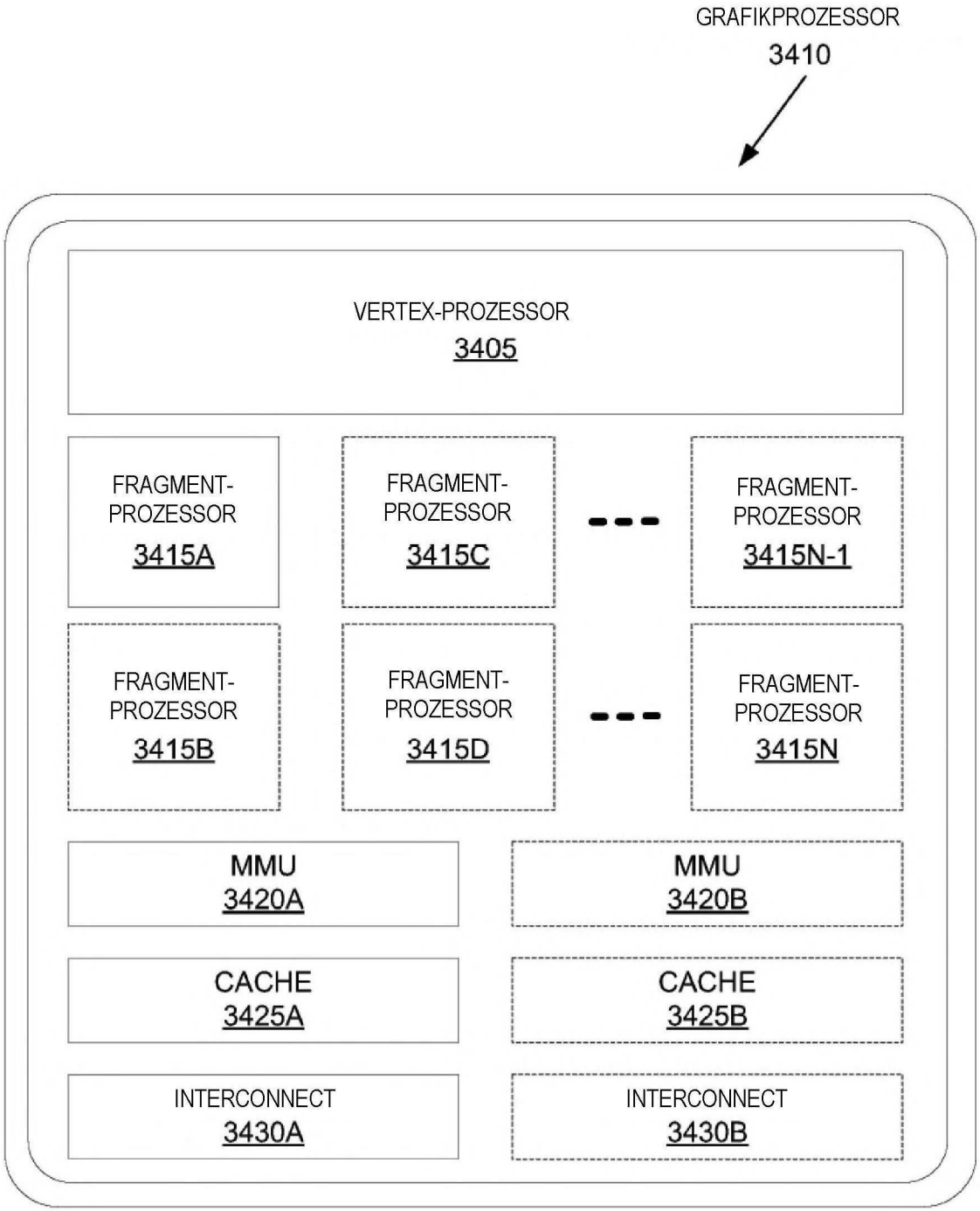

GRAFIKPROZESSOR
3410
VERTEX-PROZESSOR
3405
FRAGMENT-PROZESSOR
3415A
FRAGMENT-PROZESSOR
3415C
FRAGMENT-PROZESSOR
3415N-1
FRAGMENT-PROZESSOR
3415B
FRAGMENT-PROZESSOR
3415D
FRAGMENT-PROZESSOR
3415N
MMU
3420A
MMU
3420B
CACHE
3425A
CACHE
3425B
INTERCONNECT
3430A
INTERCONNECT
3430B

FIG. 34A

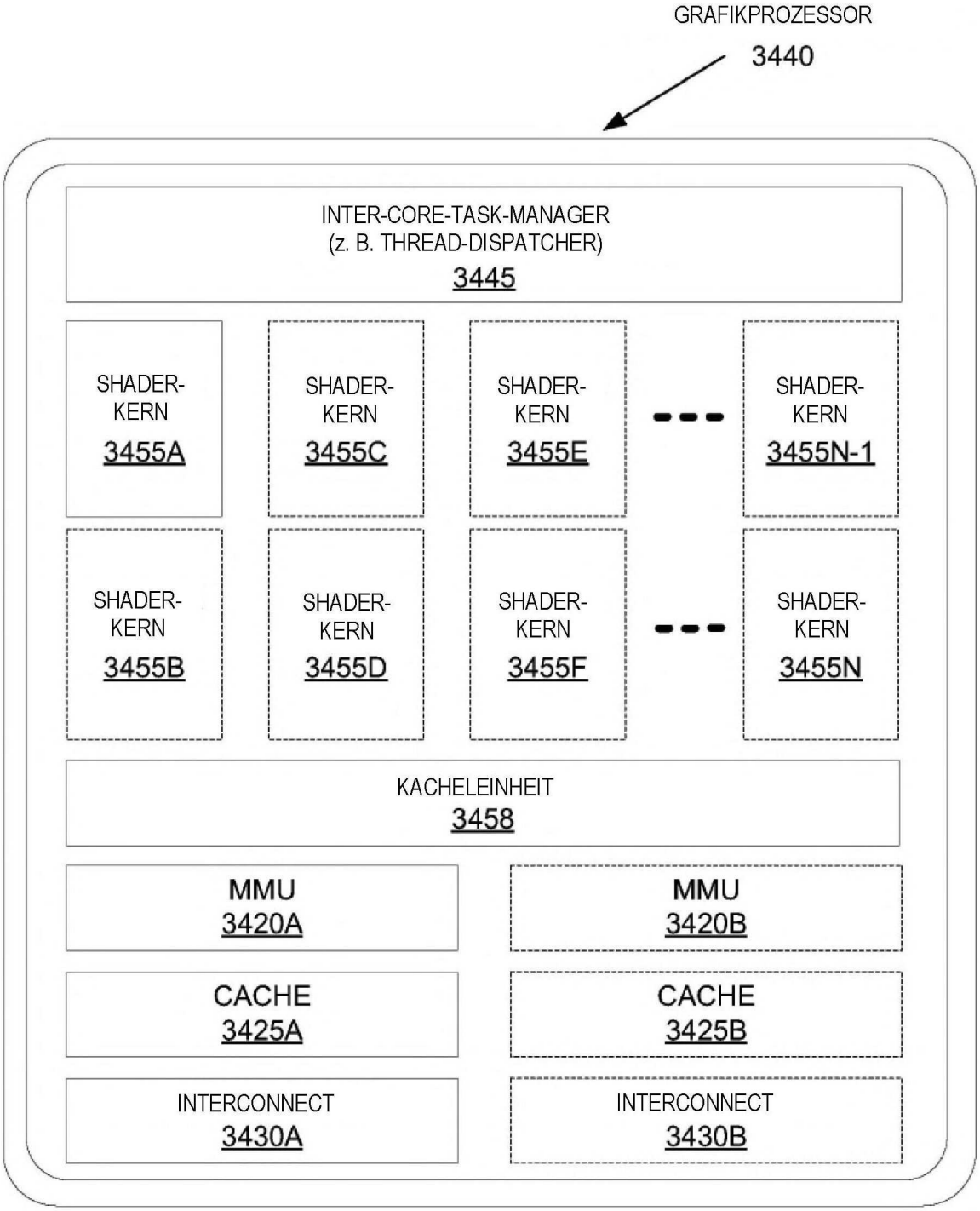

GRAFIKPROZESSOR
3440
INTER-CORE-TASK-MANAGER
(z. B. THREAD-DISPATCHER)
3445
SHADER-KERN
3455A
SHADER-KERN
3455C
SHADER-KERN
3455E
SHADER-KERN
3455N-1
SHADER-KERN
3455B
SHADER-KERN
3455D
SHADER-KERN
3455F
SHADER-KERN
3455N
KACHELEINHEIT
3458
MMU
3420A
MMU
3420B
CACHE
3425A
CACHE
3425B
INTERCONNECT
3430A
INTERCONNECT
3430B

FIG. 34B

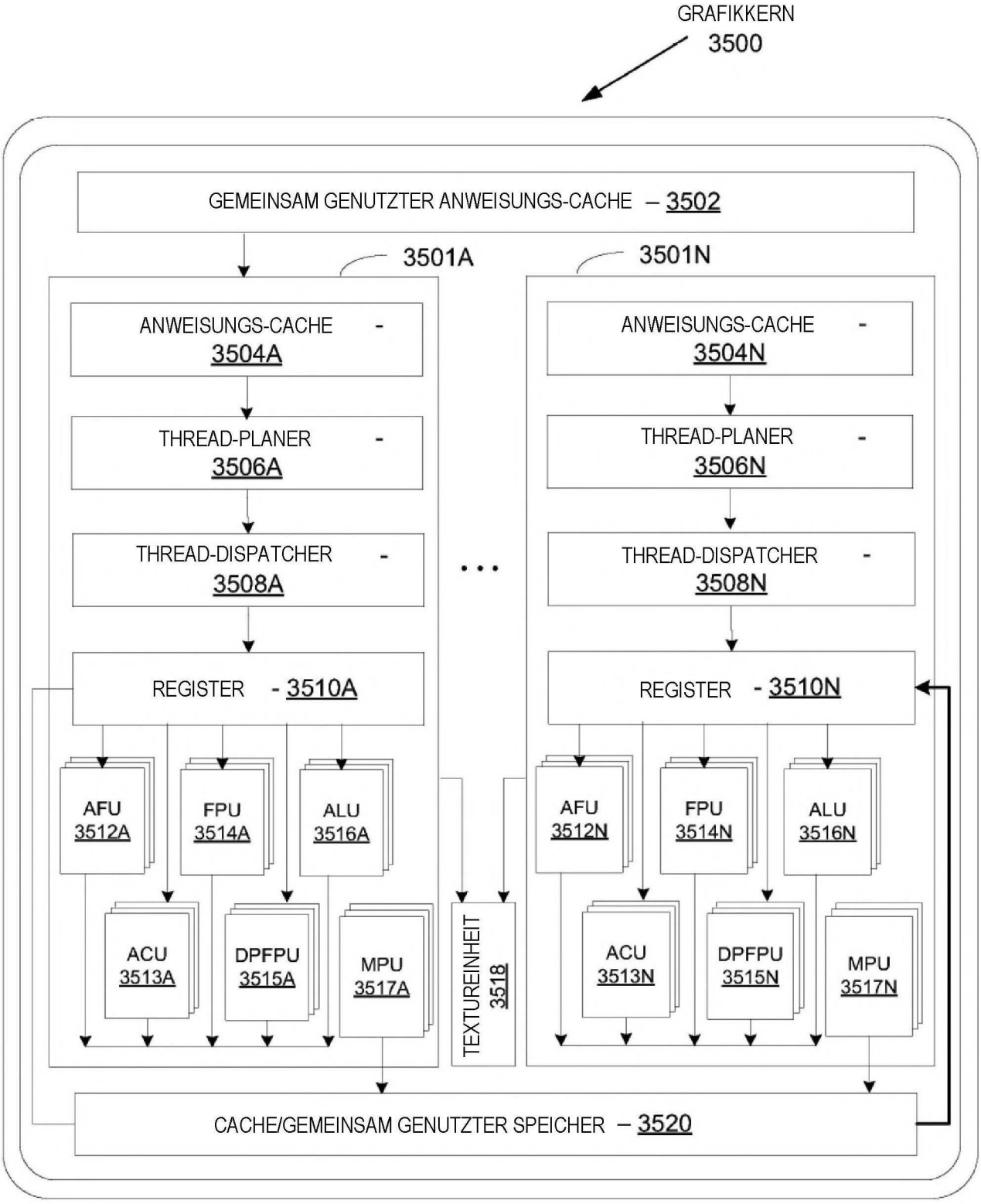

GRAFIKKERN
3500
GEMEINSAM GENUTZTER ANWEISUNGS-CACHE – 3502
3501A
3501N
ANWEISUNGS-CACHE - 3504A
THREAD-PLANER - 3506A
THREAD-DISPATCHER - 3508A
REGISTER - 3510A
AFU 3512A
FPU 3514A
ALU 3516A
ACU 3513A
DPFPU 3515A
MPU 3517A
TEXTUREINHEIT 3518
ANWEISUNGS-CACHE - 3504N
THREAD-PLANER - 3506N
THREAD-DISPATCHER - 3508N
REGISTER - 3510N
AFU 3512N
FPU 3514N
ALU 3516N
ACU 3513N
DPFPU 3515N
MPU 3517N
CACHE/GEMEINSAM GENUTZTER SPEICHER – 3520

FIG. 35A

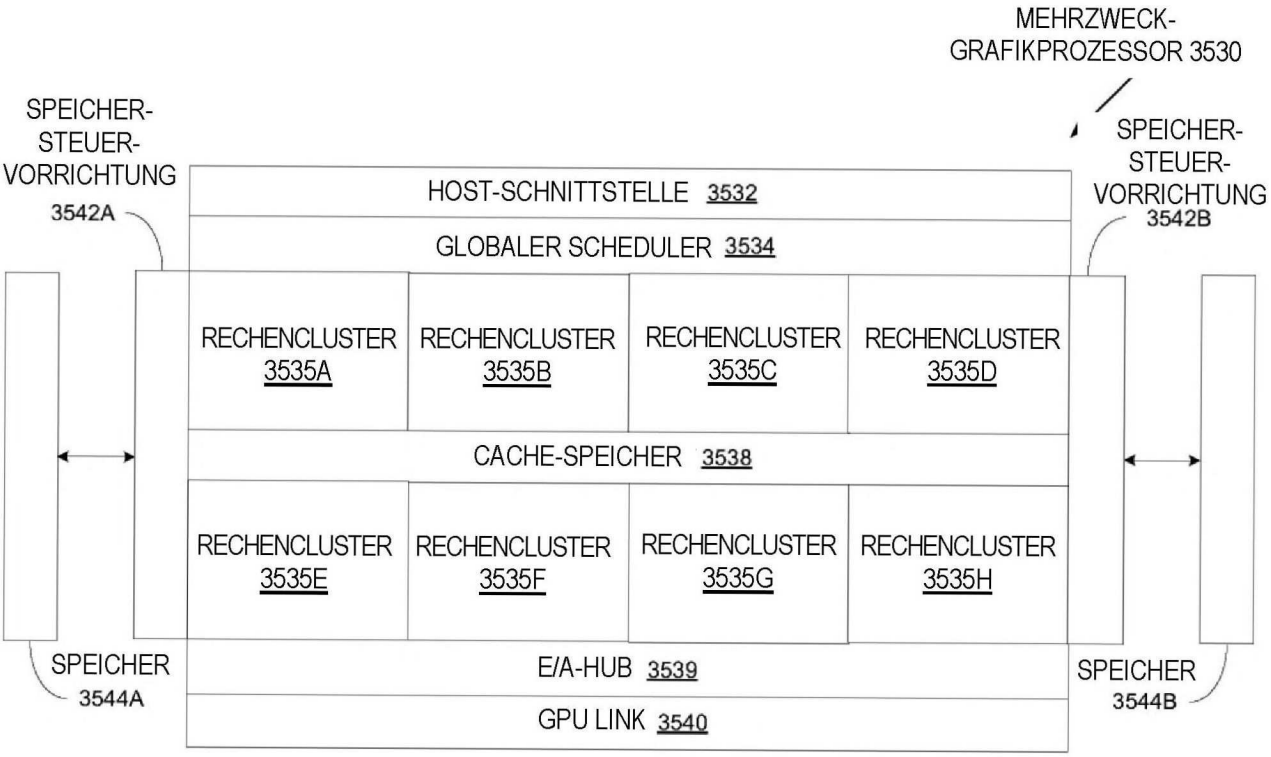

MEHRZWECK-
GRAFIKPROZESSOR 3530
SPEICHER-
STEUER-
VORRICHTUNG
3542A
SPEICHER-
STEUER-
VORRICHTUNG
3542B
HOST-SCHNITTSTELLE 3532
GLOBALER SCHEDULER 3534
RECHENCLUSTER 3535A
RECHENCLUSTER 3535B
RECHENCLUSTER 3535C
RECHENCLUSTER 3535D
CACHE-SPEICHER 3538
RECHENCLUSTER 3535E
RECHENCLUSTER 3535F
RECHENCLUSTER 3535G
RECHENCLUSTER 3535H
SPEICHER
3544A
E/A-HUB 3539
GPU LINK 3540
SPEICHER
3544B

FIG. 35B

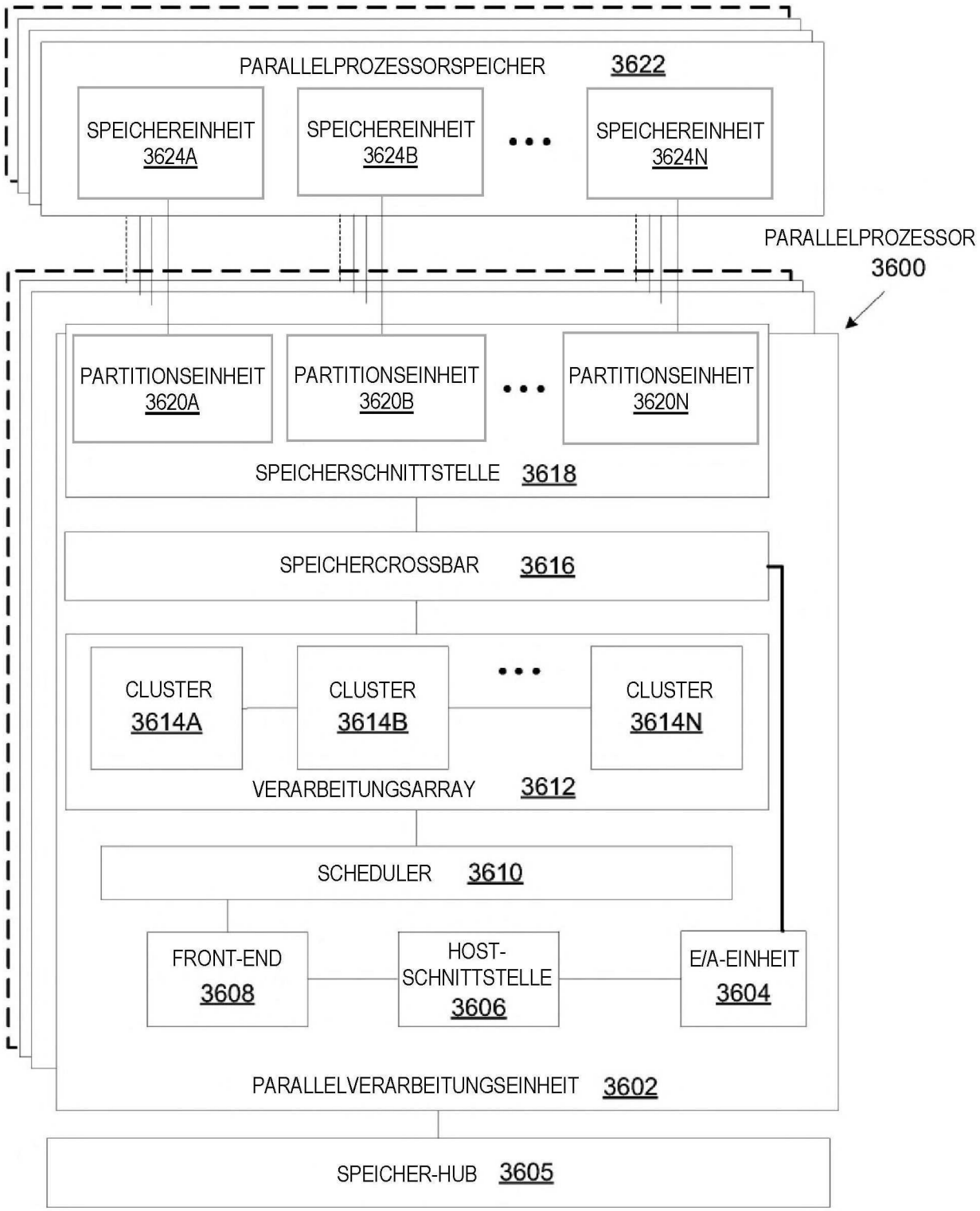

PARALLELPROZESSORSPEICHER 3622
SPEICHEREINHEIT 3624A
SPEICHEREINHEIT 3624B
SPEICHEREINHEIT 3624N
PARALLELPROZESSOR 3600
PARTITIONSEINHEIT 3620A
PARTITIONSEINHEIT 3620B
PARTITIONSEINHEIT 3620N
SPEICHERSCHNITTSTELLE 3618
SPEICHERCROSSBAR 3616
CLUSTER 3614A
CLUSTER 3614B
CLUSTER 3614N
VERARBEITUNGSARRAY 3612
SCHEDULER 3610
FRONT-END 3608
HOST-SCHNITTSTELLE 3606
E/A-EINHEIT 3604
PARALLELVERARBEITUNGSEINHEIT 3602
SPEICHER-HUB 3605

FIG. 36A

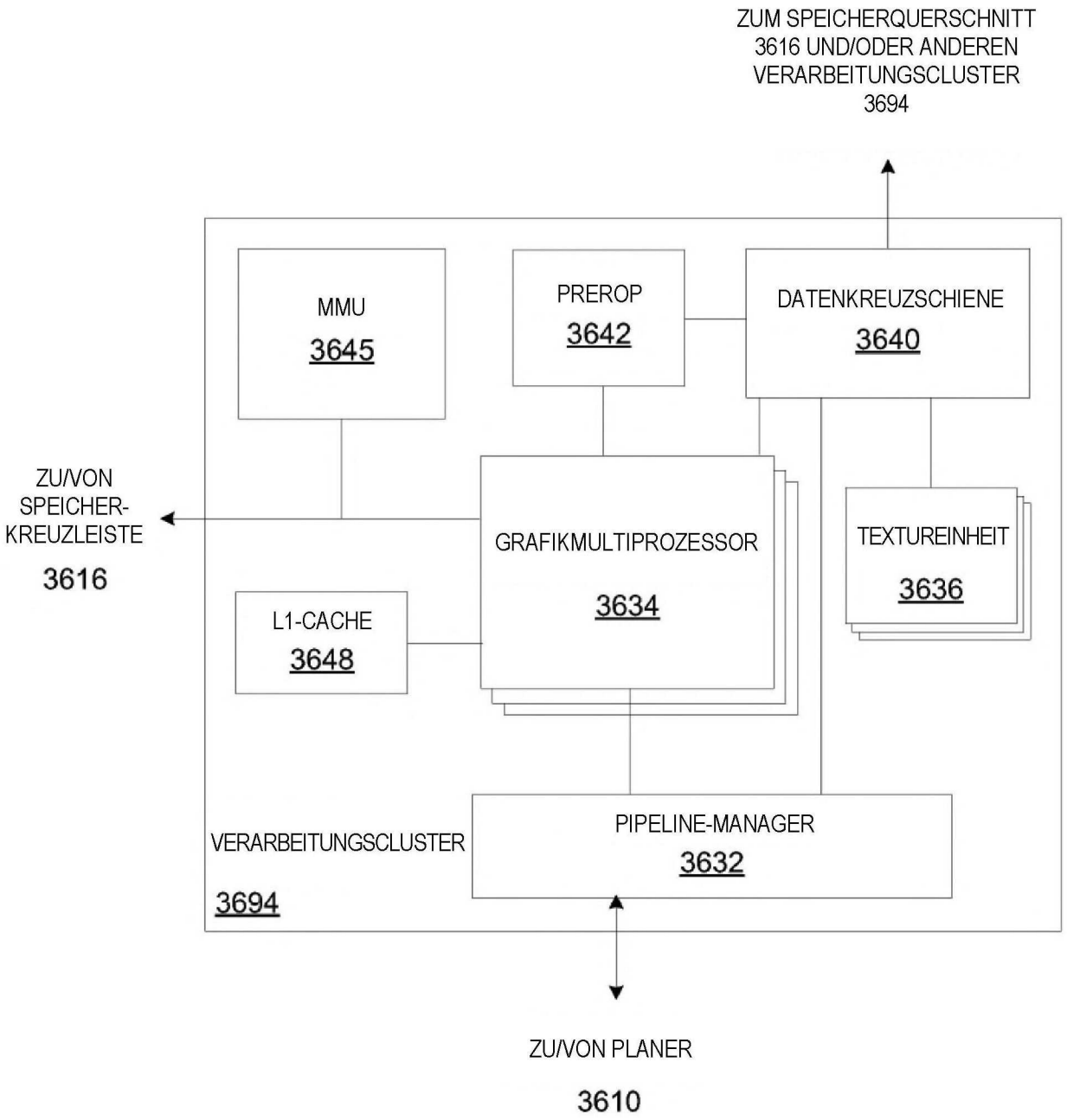

ZUM SPEICHERQUERSCHNITT
3616 UND/ODER ANDEREN
VERARBEITUNGSCLUSTER
3694
MMU
3645
PREROP
3642
DATENKREUZSCHIENE
3640
ZU/VON
SPEICHER-
KREUZLEISTE
3616
GRAFIKMULTIPROZESSOR
3634
TEXTUREINHEIT
3636
L1-CACHE
3648
VERARBEITUNGSCLUSTER
3694
PIPELINE-MANAGER
3632
ZU/VON PLANER
3610

FIG. 36B

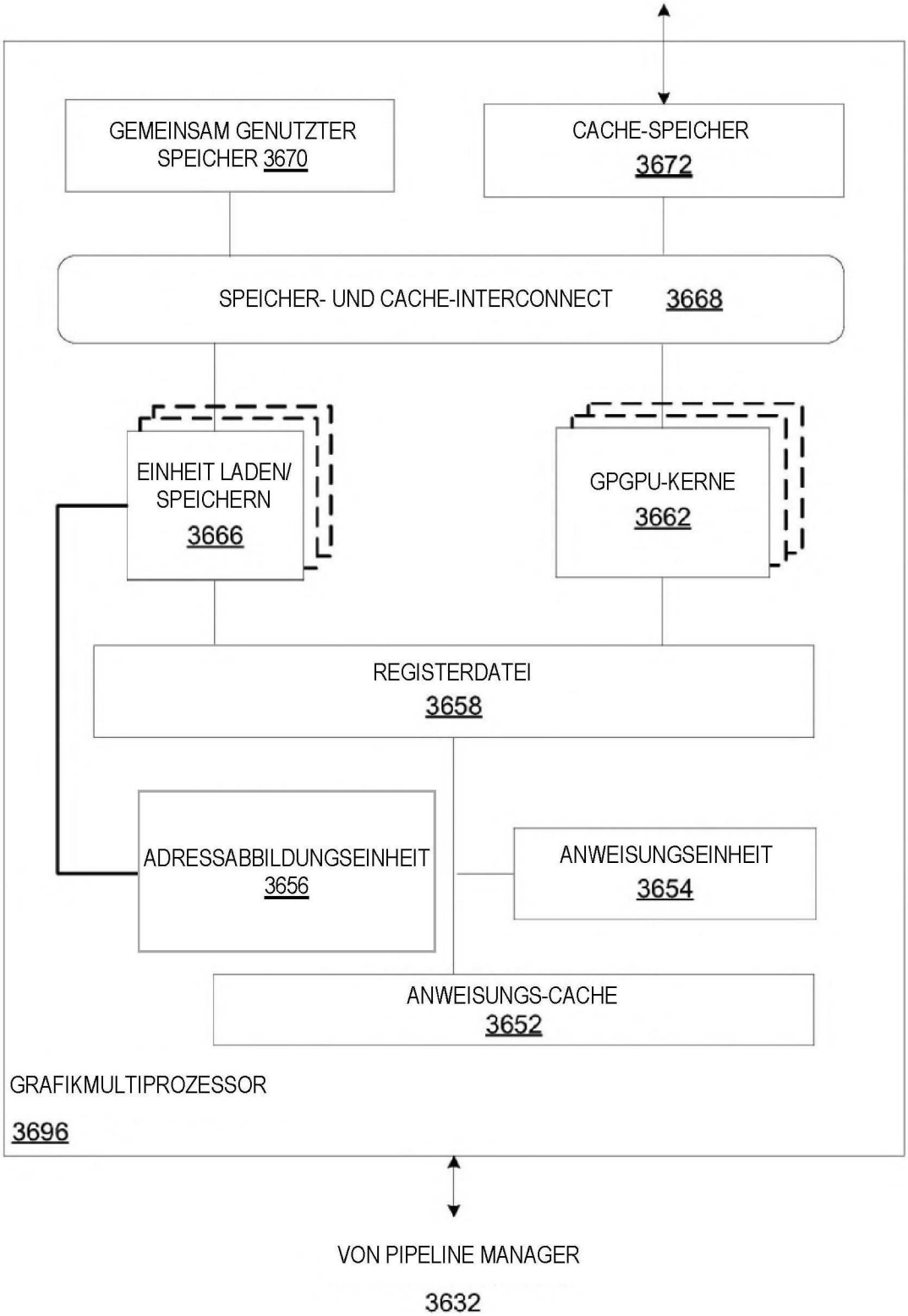

GEMEINSAM GENUTZTER SPEICHER 3670
CACHE-SPEICHER 3672
SPEICHER- UND CACHE-INTERCONNECT 3668
EINHEIT LADEN/ SPEICHERN 3666
GPGPU-KERNE 3662
REGISTERDATEI 3658
ADRESSABBILDUNGSEINHEIT 3656
ANWEISUNGSEINHEIT 3654
ANWEISUNGS-CACHE 3652
GRAFIKMULTIPROZESSOR 3696
VON PIPELINE MANAGER 3632

FIG. 36C

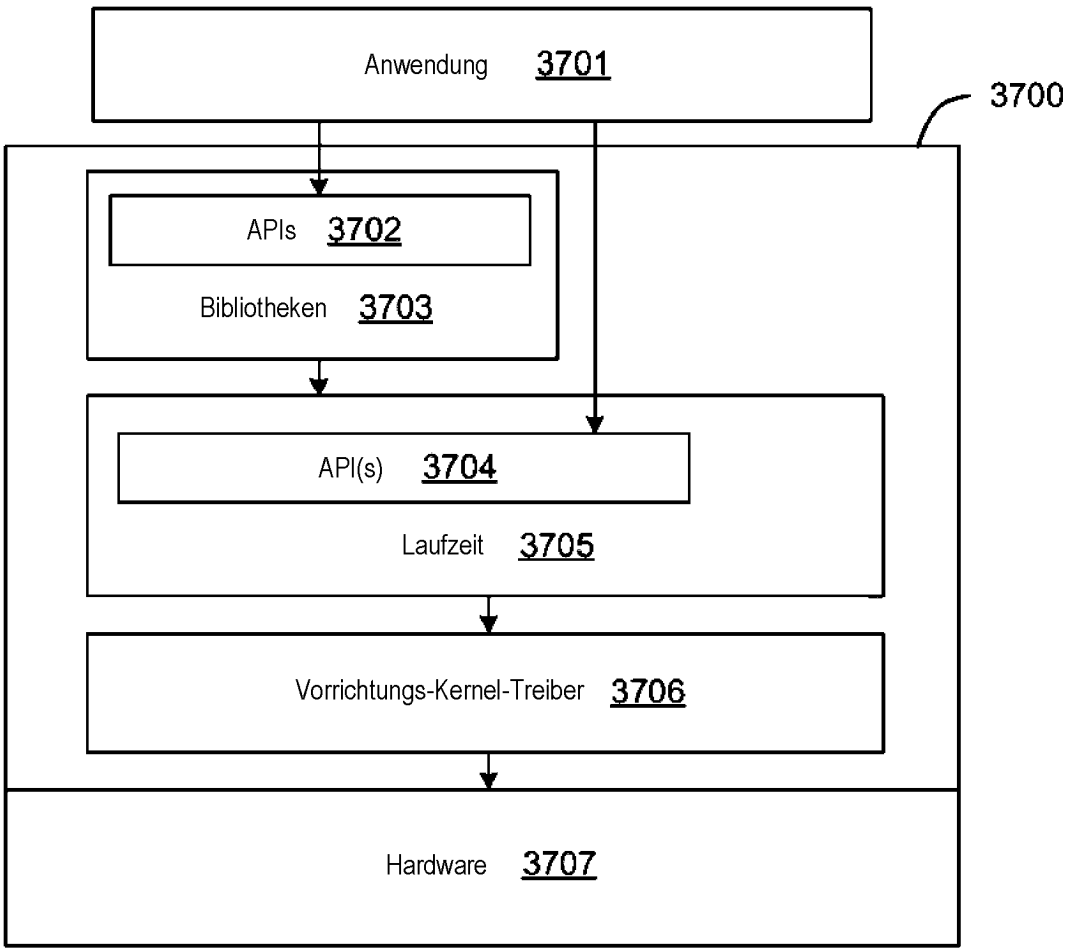

Anwendung 3701
3700
APIs 3702
Bibliotheken 3703
API(s) 3704
Laufzeit 3705
Vorrichtungs-Kernel-Treiber 3706
Hardware 3707

FIG. 37

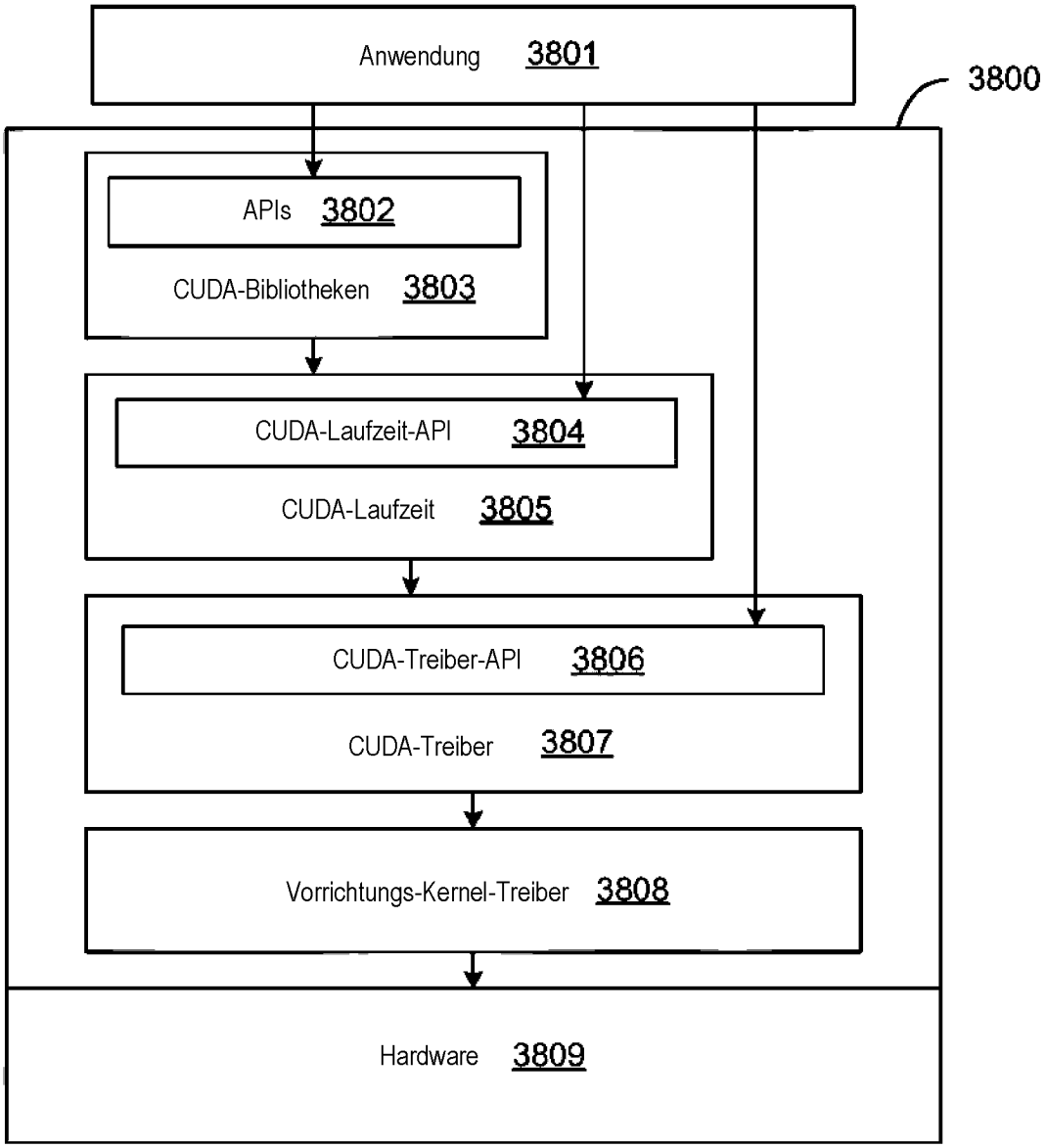

Anwendung 3801
3800
APIs 3802
CUDA-Bibliotheken 3803
CUDA-Laufzeit-API 3804
CUDA-Laufzeit 3805
CUDA-Treiber-API 3806
CUDA-Treiber 3807
Vorrichtungs-Kernel-Treiber 3808
Hardware 3809

FIG. 38

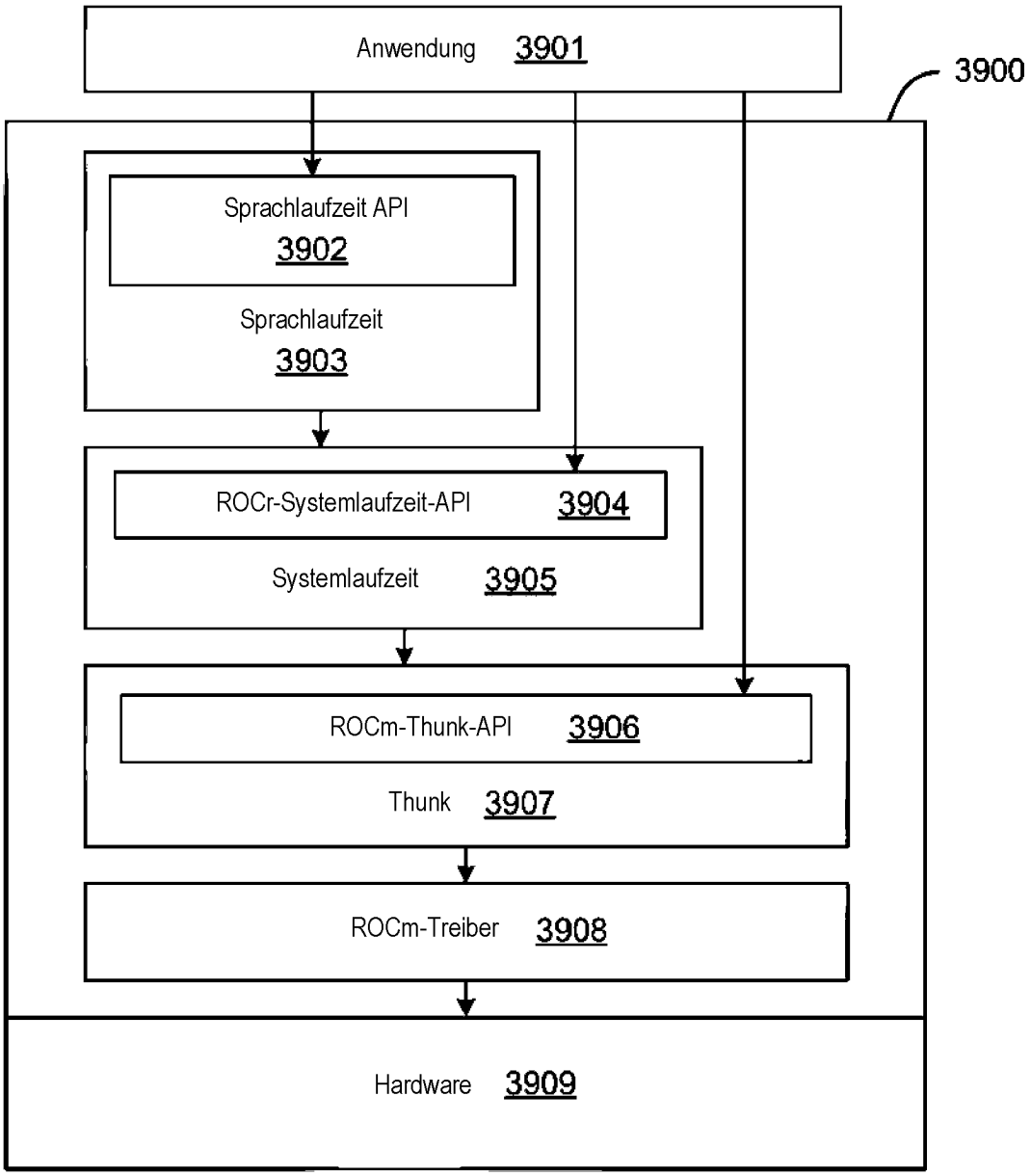

Anwendung 3901
3900
Sprachlaufzeit API
3902
Sprachlaufzeit
3903
ROCr-Systemlaufzeit-API 3904
Systemlaufzeit 3905
ROCm-Thunk-API 3906
Thunk 3907
ROCm-Treiber 3908
Hardware 3909

FIG. 39

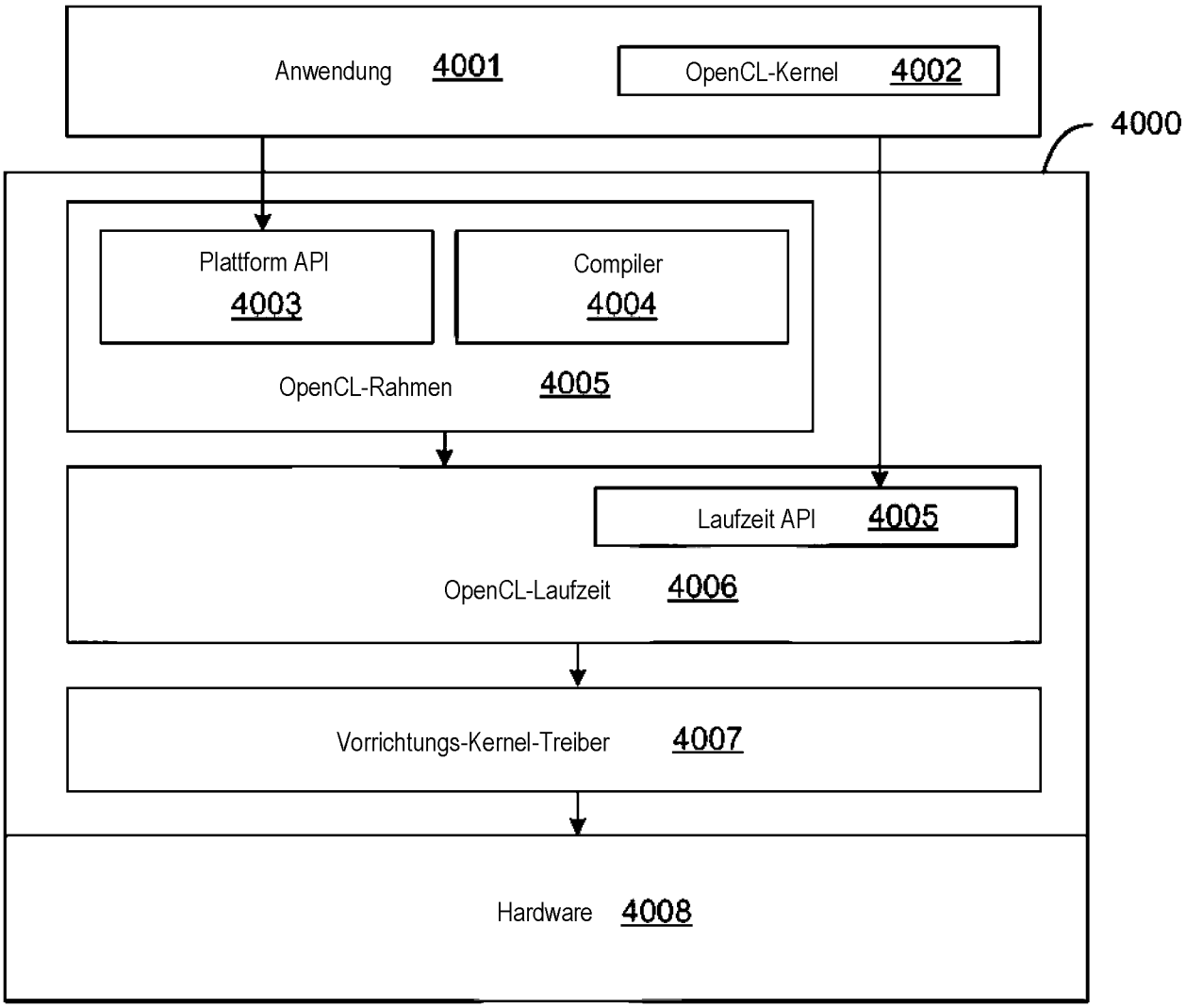

Anwendung 4001
OpenCL-Kernel 4002
4000
Plattform API 4003
Compiler 4004
OpenCL-Rahmen 4005
Laufzeit API 4005
OpenCL-Laufzeit 4006
Vorrichtungs-Kernel-Treiber 4007
Hardware 4008

FIG. 40

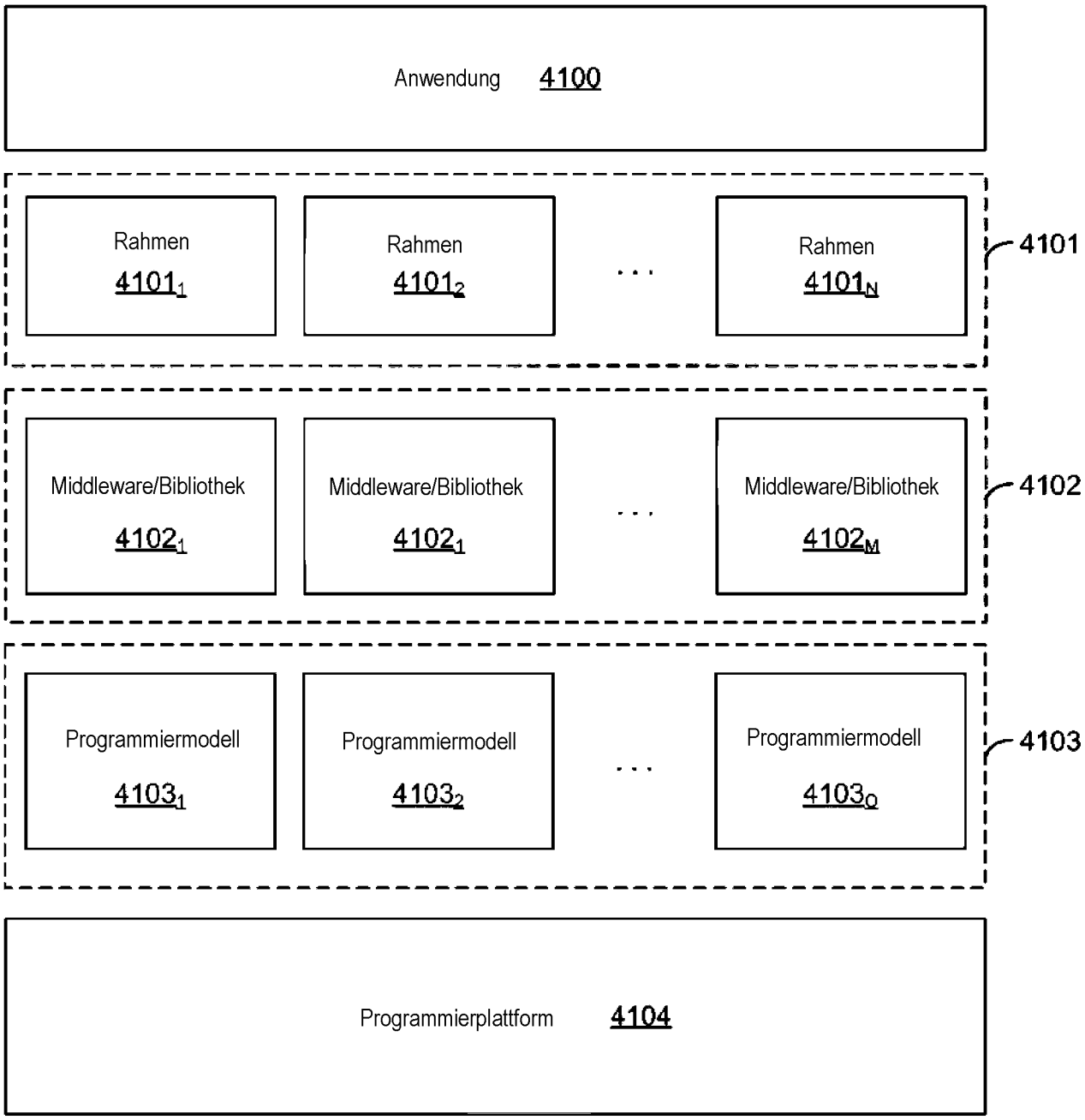

Anwendung 4100
Rahmen 4101$_1$
Rahmen 4101$_2$
Rahmen 4101$_N$
4101
Middleware/Bibliothek 4102$_1$
Middleware/Bibliothek 4102$_1$
Middleware/Bibliothek 4102$_M$
4102
Programmiermodell 4103$_1$
Programmiermodell 4103$_2$
Programmiermodell 4103$_O$
4103
Programmierplattform 4104

FIG. 41

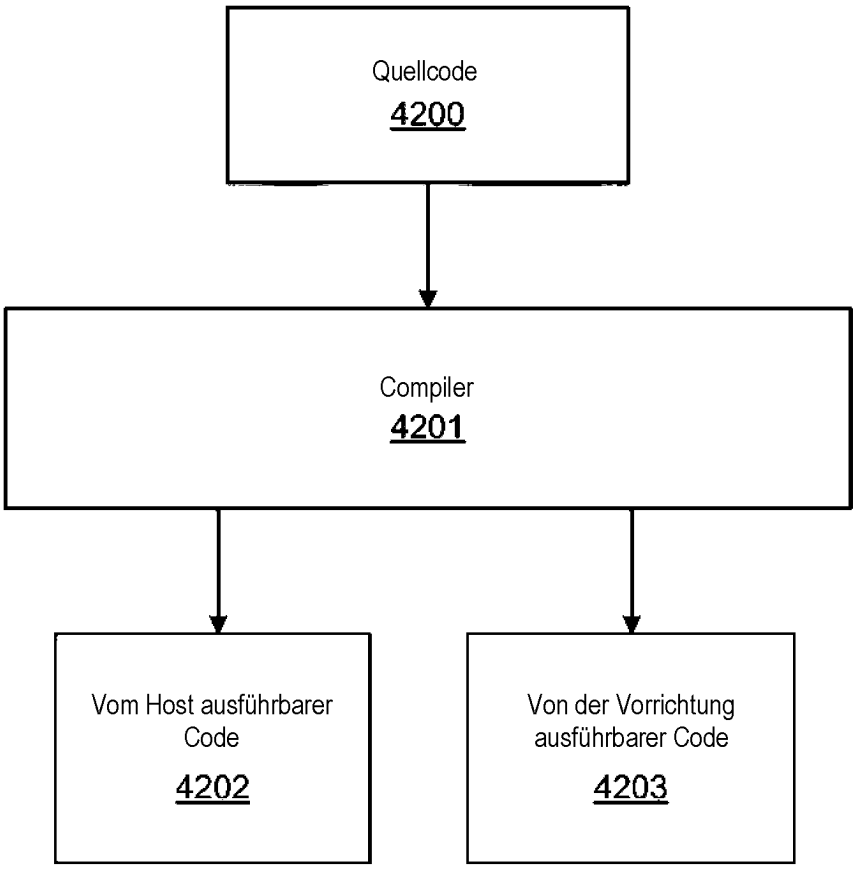


FIG. 42

 Das Dokument wurde durch die Firma Luminess hergestellt.