



(10) **DE 11 2012 004 336 B4** 2024.02.01

(12)

Patentschrift

(21) Deutsches Aktenzeichen: **11 2012 004 336.7**
(86) PCT-Aktenzeichen: **PCT/US2012/066164**
(87) PCT-Veröffentlichungs-Nr.: **WO 2013/078258**
(86) PCT-Anmeldetag: **20.11.2012**
(87) PCT-Veröffentlichungstag: **30.05.2013**
(43) Veröffentlichungstag der PCT Anmeldung
in deutscher Übersetzung: **14.08.2014**
(45) Veröffentlichungstag
der Patenterteilung: **01.02.2024**

(51) Int Cl.: **G06F 9/50 (2006.01)**
G06F 15/00 (2006.01)
G06Q 10/06 (2023.01)

Innerhalb von neun Monaten nach Veröffentlichung der Patenterteilung kann nach § 59 Patentgesetz gegen das Patent Einspruch erhoben werden. Der Einspruch ist schriftlich zu erklären und zu begründen. Innerhalb der Einspruchsfrist ist eine Einspruchsgebühr in Höhe von 200 Euro zu entrichten (§ 6 Patentkostengesetz in Verbindung mit der Anlage zu § 2 Abs. 1 Patentkostengesetz).

(30) Unionspriorität:
13/304,571 **25.11.2011** **US**

(73) Patentinhaber:
International Business Machines Corporation,
Armonk, NY, US

(74) Vertreter:
Richardt Patentanwälte PartG mbB, 65185
Wiesbaden, DE

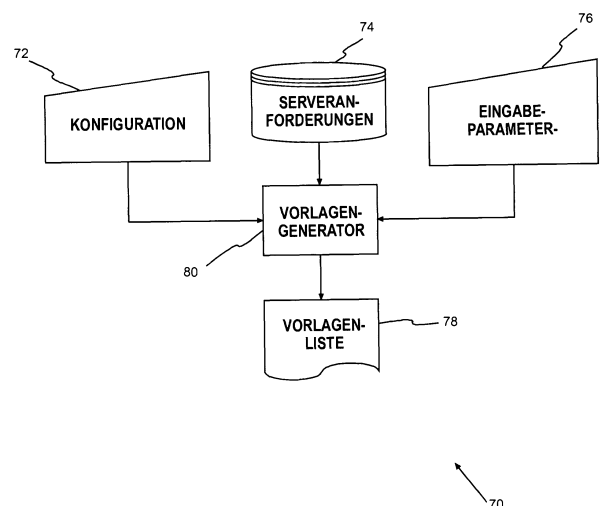
(72) Erfinder:
Assungao, Marcos Dias De, c/o IBM Corp., San
Paulo, BR; Netto, Marco Aurelio Stelmar, c/o IBM
Corp., Sao Paulo, BR; Peterson, Brian L., c/o IBM
Corp., Somers, N.Y., US; Renganarayana,
Lakshminarayanan, c/o IBM T.J. Wat. Res.,
Hawthorne, N.Y., US; Rofrano, John J., c/o IBM T.J.
Watson Res. Cent., Hawthorne, N.Y., US; Young,
Christopher C., c/o IBM T.J. Wats. Res.C.,
Hawthorne, N.Y., US

(56) Ermittelter Stand der Technik:
siehe Folgeseiten

(54) Bezeichnung: **System, Verfahren und Programmprodukt für kostenbewusste Auswahl von Vorlagen zum Bereitstellen von gemeinsam genutzten Ressourcen**

(57) Hauptanspruch: Gemeinsam genutztes Ressourcensystem, das aufweist:
eine Vielzahl von Client-Einheiten;
einen oder mehrere Dienstanbieter-Computer, die selektiv Ressourcenkapazität für die Vielzahl von Client-Einheiten verfügbar machen;
ein Netzwerk, wobei Ressourcenkapazität für die Vielzahl von Client-Einheiten von dem einen oder den mehreren Dienstanbieter-Computern über das Netzwerk bereitgestellt wird; und
eine Ressourcenbereitstellungseinheit, die Ressourcenvorlagen für alle zuordenbaren Ressourcenkapazitäts-Konfigurationen erzeugt, eine Untergruppe von erzeugten Ressourcenvorlagen auswählt und eine Anzahl von Ressourcenvorlagen verwaltet, die aus jeder Untergruppe ausgewählt wurden, wobei jede Vorlage eine zuordenbare Konfiguration von Ressourcenkapazität angibt, wobei jeder anfordernden Client-Einheit Ressourcen zugeordnet sind, die durch eine der ausgewählten Ressourcenvorlagen der Untergruppe bestimmt wurden, wobei die Ressourcenbereitstellungseinheit aufweist:
einen Speicher für Konfigurationen (72), der Kosten von zuordenbaren Ressourcen und zugehörige Attribute spei-

chert;
einen Speicher für Serveranforderungen (74), der vorher empfangene Anforderungen von Kapazität speichert;
einen ...



(19)



Deutsches
Patent- und Markenamt

(10) **DE 11 2012 004 336 B4** 2024.02.01

(56) Ermittelter Stand der Technik:

US	2007 / 0 271 203	A1
US	2008 / 0 126 406	A1
US	2008 / 0 294 777	A1
US	2009 / 0 171 705	A1

Beschreibung**HINTERGRUND DER ERFINDUNG****Gebiet der Erfindung**

[0001] Die vorliegende Erfindung betrifft ein Zuordnen von gemeinsam genutzten Ressourcen und insbesondere ein Auswählen einer optimalen Gruppe von Vorlagen, um Ressourcenanforderungen mit minimaler Über- und Unterbereitstellung zu erfüllen.

Beschreibung des Stands der Technik

[0002] Der Erwerb und die Verwaltung von Informationstechnologie (IT) ist für jede moderne Organisation ein größeres Budgetproblem. Darüber hinaus wird lokale IT-Hardware selten mit voller Kapazität genutzt. Um daher Kosten und Verschwendung von IT-Infrastruktur zu verringern, gehen Organisationen zunehmend dazu über, Arbeitslast auf virtuellen Maschinen (VMs) zu konsolidieren, die auf weniger Servern gehostet werden, anstatt physische Hardware zu erwerben. Ein entfernt angeordneter Server-Computer stellt jede VM als einen virtuellen Server mit virtuellen Ressourcen bereit, z.B. Verarbeitungsleistung, Arbeitsspeicher und Plattenspeicherplatz. Üblicherweise wird jede VM-Konfiguration aus einer Anzahl von virtuellen Ressourcenvorlagen (VRTs oder Vorlagen) ausgewählt. Jede VRT definiert vorgegebene virtuelle Ressourcenfunktionen, die zugewiesen werden können, um eine VM zu definieren. Der bzw. die Server-Computer ordnen jeder VM Kapazität zu (z.B. Plattenspeicherplatz, Verarbeitungsressourcen und Arbeitsspeicher), indem eine VRT zugewiesen wird, deren Konfiguration (Software-Stack und Lizenzen) dem beabsichtigten Zweck und den erwarteten Anforderungen für die VM am nächsten kommt.

[0003] Beim Verwalten dieser VMs war es schwierig, ihre optimale Kapazität und eine optimale Konfiguration zu bestimmen, d.h. die optimale VRT auszuwählen. Üblicherweise wählt ein Dienstanbieter eine VRT aus und ordnet jeder VM entsprechende physische Ressourcen zu, vorwiegend auf der Grundlage einer Anbietersystemoptimierung, von Arbeitslastvorhersagen und einer Ressourcennutzungshistorie, die aus einer ständigen Überwachung der VM-Ressourcennutzung erfasst wird. Jedoch können selbst gute Vorhersageergebnisse die Benutzererfahrung aufgrund von Über- oder Unterzuteilung beeinträchtigen. Eine Überzuteilung verschwendet Energie und Ressourcen, Kapazität, die ansonsten für andere Benutzer oder zum Unterstützen weiterer VMs verfügbar wäre. Da eine Unterzuteilung unzureichende Ressourcen zu einer oder mehreren VMs zuweist, wirkt sich dies auf die Servicequalität (Quality-of-Service (QoS)) dieser VMs aus, z.B. stockende Videodaten oder verstümmelte Audiodaten.

[0004] Benutzeranforderungen sind höchst unterschiedlich, was Anbieter zwingen kann, die Definition und Anzahl der von ihnen angebotenen Vorlagen zu verändern. Obwohl Ressourcenanbieter die Anzahl von angebotenen VRTs erhöhen können, um allen Anforderungen gerecht zu werden, kann ein Erhöhen der Anzahl eine Überbereitstellung von Ressourcen und eine Zersiedelung von Vorlagen (template sprawling) verursachen. Eine Ressourcen-Überbereitstellung, wie eine Überbuchung, nimmt mehr Ressourcen und Energie in Anspruch als für die bereitgestellte Kapazität notwendig ist; dieser Überschuss an Ressource und Energie könnte ansonsten für andere Benutzer oder weitere VMs verfügbar gemacht werden. Eine Vorlagenzersiedlung, die auch unter Bildzersiedlung (image sprawl) bekannt ist, tritt auf, wenn eine Vorlage oder ein Bild, die bzw. das den Anforderungen eines Benutzers entspricht, optimiert wird, um zu denen eines anderen zu passen, wodurch eine weitere Vorlage zu den angebotenen VRTs hinzugefügt wird. Letzten Endes wächst die Anzahl von Vorlagen auf eine nicht mehr verwaltbare Anzahl.

[0005] Mehrere Ansätze wurden ausprobiert, um Anforderungen von Anwendungsservern mit verfügbaren Ressourcen abzugleichen. Der typische Cloud-Computing-Ansatz bestand darin, eine VM-Vorlage auszuwählen, die den Anforderungen der Benutzer-Zielanwendungen am besten entsprach, wobei die Kosten der Auswahl unberücksichtigt blieben. Einige weitere vorherige Ansätze geben VM-Vorlagen und Netzwerkvorlagen an und speichern sie, um vor allem ein Installieren von VMs, zugehörigen Netzwerken und erforderlicher Software zu beschleunigen. In noch weiteren Ansätzen wurden Profile von in Betrieb befindlichen VMs und auswählenden Ziel-Migrationsservern erstellt. Statt den Schwerpunkt auf ein Auswählen von virtuellen Ressourcenvorlagen zu legen, stellen andere Techniken alternativ ein Skalieren von Anwendungen auf die Vorlagen in den Mittelpunkt, d.h. ein Anpassen der Anwendung an die VM statt umgekehrt. Schließlich gleicht ein technologisch ausgereifter Ansatz die vom Benutzer bereitgestellten Instanzen mit offenem Virtualisierungsformat (Open Virtualization Format (OVF)) mit entsprechenden Cloud-Angeboten ab.

[0006] Somit besteht ein Bedarf an einem Zuordnen von angemessenen IT-Ressourcen zu Mindestkosten und ohne Ressourcenverschwendung, während gleichzeitig auch eine Server-QoS aufrechterhalten wird, und insbesondere besteht ein Bedarf an einem Auswählen eines Satzes von VM-Vorlagen zum Bereitstellen von VMs in einer Cloud-Infrastruktur, von Vorlagen, die die meisten Benutzeranforderungen zu Mindestkosten einwandfrei erfüllen.

[0007] Im Stand der Technik offenbart US 2008 / 0 126 406 A1 ein objektbasiertes Model-

lierungssystem, -verfahren und -computergerät, welches einen realen Lebensprozess modelliert. US 2009 / 0 171 705 A1 offenbart Vorlagen, welche in einer IT-Umgebung bereitgestellt und verwendet werden. US 2008 / 0 294 777 A1 offenbart ein Verfahren zur automatischen Bereitstellung von Rechenressourcen in einem verteilten Rechnernetz. US 2007 / 0 271 203 A1 offenbart Systeme und Methoden zur Erstellung von Kostenvoranschlägen.

KURZDARSTELLUNG DER ERFINDUNG

[0008] Ein Merkmal der Erfindung ist ein Minimieren einer Überbereitstellung von Cloud-Ressourcen;

ein weiteres Merkmal der Erfindung ist eine verringerte Vorlagenzersiedelung/Bildzersiedelung in einer Cloud-Umgebung;

noch ein weiteres Merkmal der Erfindung ist ein Pool von VM-Vorlagen/-Bildern, die für ein kostenbewusstes Clustering ausgewählt wurden, um Kundenanforderungen besser zu entsprechen, wodurch Client-Kosten und Ressourcenverschwendung verringert werden.

[0009] Die vorliegende Erfindung betrifft einen Vorlagengenerator, der Vorlagen in einem kostenbewussten Clustering zusammenfasst, ein Verfahren zum Zuordnen von Ressourcen unter Verwendung eines kostenbewussten Clusterings und der dazugehörigen Computerprogrammprodukte. Eine Ressourcenbereitstellungseinheit erzeugt, wählt und verwaltet eine ausgewählte Anzahl von Ressourcenvorlagen. Jede Vorlage gibt eine zuordenbare Konfiguration von Ressourcenkapazität an. Jeder anfordernden Client-Einheit sind Ressourcen zugeordnet, die durch eine der ausgewählten Ressourcenvorlagen bestimmt wurden.

[0010] Eine Ressourcenbereitstellungseinheit enthält einen Speicher für Konfigurationen mit Kosten von zuordenbaren Ressourcen und zugehörigen Attributen, einen Speicher für Serveranforderungen mit vorher empfangenen Anforderungen und einen Speicher für Eingabeparameter mit Vorlagenlistenooptionen. Ein Vorlagengenerator bestimmt eine optimale Liste von Vorlagen, um vorher empfangene Anforderungen zu erfüllen. Ein Vorlagenspeicher speichert erzeugte Vorlagenlisten. Die sich daraus ergebenden Vorlagenlisten beruhen auf den jeweiligen Ressourcen-Gesamtkosten. Ressourcen können jeweils stark voneinander abweichende und unterschiedliche Kosten aufweisen, d.h. eine hohe Kostendifferenz bei Verarbeitungsleistung, Arbeitsspeicher und Plattenspeicherplatz.

KURZBESCHREIBUNG DER ZEICHNUNGEN

[0011] Die vorgenannten und weiteren Zielsetzungen, Gesichtspunkte und Vorteile sind besser

anhand der folgenden ausführlichen Beschreibung einer bevorzugten Ausführungsform der Erfindung unter Bezugnahme auf die Zeichnungen zu verstehen, wobei:

Fig. 1 einen Cloud-Computing-Knoten gemäß einer Ausführungsform der vorliegenden Erfindung veranschaulicht;

Fig. 2 eine Cloud-Computing-Umgebung gemäß einer Ausführungsform der vorliegenden Erfindung veranschaulicht;

Fig. 3 Abstraktionsmodellsschichten gemäß einer Ausführungsform der vorliegenden Erfindung veranschaulicht;

Fig. 4 ein Beispiel für eine Ressourcenbereitstellungseinheit für kostenbewusstes Clustering zum Abgleichen von Vorlagen mit Anforderungen für Ressourcen gemäß einer bevorzugten Ausführungsform der vorliegenden Erfindung zeigt;

Fig. 5 ein Beispiel für die Auswahl der Top-k-Vorlagen durch einen bevorzugten Vorlagengenerator zeigt;

Fig. 6A ein Beispiel für ein Auffinden von Mindestkosten-Übereinstimmungen zeigt;

Fig. 6B ein Beispiel für eine Ressourcen-Übereinstimmungstabelle zeigt, die für ein Auflisten von Mindestkosten-Übereinstimmungen erstellt wurde;

Fig. 7 ein Beispiel für ein Erstellen der endgültigen Vorlagenliste durch Umspannen der Top-k-Listen mit einem niedrigsten und einem höchsten Wert zeigt;

Fig. 8 ein Beispiel für eine Anwendung der vorliegenden Erfindung auf einen Satz von siebenhundertsiebenundvierzig (747) Serveranforderungen zeigt.

BESCHREIBUNG VON BEVORZUGTEN AUSFÜHRUNGSFORMEN

[0012] Es ist von vornherein zu verstehen, dass, obwohl diese Offenbarung eine detaillierte Beschreibung zu Cloud-Computing enthält, eine Umsetzung der hierin zitierten Lehren nicht auf eine Cloud-Computing-Umgebung begrenzt ist. Die Ausführungsformen der vorliegenden Erfindung sind stattdessen fähig, in Verbindung mit jedem anderen Typ einer jetzt bekannten oder später entwickelten Datenverarbeitungsumgebung und wie des Weiteren hierin im Folgenden angegeben umgesetzt zu werden.

[0013] Cloud-Computing ist ein Modell einer Dienstleistung zum Ermöglichen eines bequemen bedarfsabhängigen Netzwerkzugriffs (On-Demand-Netzwerkzugriffs) auf einen gemeinsam genutzten Pool von konfigurierbaren Datenverarbeitungsressourcen

(z.B. Netzwerke, Netzbandbreite, Server, Verarbeitung, Arbeitsspeicher, Speicher, Anwendungen, virtuelle Maschinen und Dienste), die mit minimalem Verwaltungsaufwand oder Interaktion mit einem Anbieter des Diensts rasch bereitgestellt und freigegeben werden können. Dieses Cloud-Modell kann mindestens fünf Merkmale, mindestens drei Dienstmodelle und mindestens vier Nutzungsmodelle enthalten.

[0014] Die Merkmale sind wie folgt:

On-Demand-Selfservice: Ein Cloud-Nutzer kann einseitig und je nach Bedarf automatisch für Datenverarbeitungs-Ressourcen, wie z.B. Serverzeit und Netzwerkspeicher, ohne erforderliche menschliche Interaktion mit dem Anbieter des Diensts Vorsorge treffen.

[0015] Breiter Netzwerkzugriff: Funktionen stehen über ein Netzwerk zur Verfügung, und auf sie wird über Standardmechanismen zugegriffen, die eine Verwendung über heterogene Thin Client- oder Thick Client-Plattformen unterstützen (z.B. Mobiltelefone, Laptops und PDAs).

[0016] Ressourcen-Pooling: Die Datenverarbeitungsressourcen des Anbieters sind in einem Pool zusammengeschlossen, um unter Verwendung eines Multi-Tenant-Modells mehreren Nutzern mit unterschiedlichen physischen und virtuellen Ressourcen zu dienen, deren Zuweisung oder Zuweisungsaufhebung je nach Bedarf dynamisch erfolgt. Es herrscht insofern ein Gefühl der Standortunabhängigkeit, als der Nutzer im Allgemeinen keine Kontrolle oder Kenntnis in Bezug auf den genauen Standort der bereitgestellten Ressourcen hat, aber in der Lage sein kann, einen Standort auf einer höheren Abstraktionsebene anzugeben (z.B. Land, Bundesstaat oder Rechenzentrum).

[0017] Schnelle Elastizität: Funktionen können rasch und elastisch, in einigen Fällen automatisch für eine schnelle Erweiterung bereitgestellt und für eine schnelle Reduzierung rasch freigegeben werden. Für den Nutzer scheinen die zur Bereitstellung verfügbaren Funktionen oft unbegrenzt zu sein und können jederzeit in beliebiger Menge erworben werden.

[0018] Richtig bemessener Dienst: Cloud-Systeme kontrollieren und optimieren Ressourcen automatisch durch die Nutzung einer Messfunktion auf einer Abstraktionsebene, die dem Diensttyp entspricht (z.B. Speicher, Verarbeitung, Bandbreite und aktive Benutzerkonten). Eine Inanspruchnahme von Ressourcen kann überwacht, kontrolliert und gemeldet werden, was sowohl für den Anbieter als auch den Nutzer des in Anspruch genommenen Diensts für Transparenz sorgt. Des Weiteren stellt die vorliegende Erfindung eine Client-Selbstüberwachung

zum Korrigieren von individueller Ressourcenzuordnung und -konfiguration während des Betriebs bereit für eine optimierte Ressourcenzuordnung in Echtzeit und mit minimierten Betriebskosten und minimierter Energie-Inanspruchnahme.

[0019] Es gibt folgende Dienstmodelle:

Software as a Service (SaaS): Die für den Nutzer bereitgestellte Funktion besteht darin, die auf einer Cloud-Infrastruktur ausgeführten Anwendungen des Anbieters zu nutzen. Auf die Anwendungen kann über eine Thin Client-Schnittstelle, wie beispielsweise einen Web-Browser (z.B. webberuhende eMail), von verschiedenen Client-Einheiten aus zugegriffen werden. Weder verwaltet noch kontrolliert der Nutzer die zu Grunde liegende Cloud-Infrastruktur, einschließlich Netzwerk, Server, Betriebssysteme, Speicher oder auch einzelne Anwendungsfunktionen, ausgenommen möglicherweise begrenzte benutzerspezifische Einstellungen einer Anwendungskonfiguration.

[0020] Platform as a Service (PaaS): Die für den Nutzer bereitgestellte Funktion besteht darin, auf der Cloud-Infrastruktur vom Nutzer erstellte oder erworbene Anwendungen zu nutzen, die unter Verwendung von Programmiersprachen und Tools erstellt wurden, die von dem Anbieter unterstützt werden. Weder verwaltet noch kontrolliert der Nutzer die zu Grunde liegende Cloud-Infrastruktur, einschließlich Netzwerk, Server, Betriebssysteme oder Speicher, aber er besitzt die Kontrolle über die umgesetzten Anwendungen und möglicherweise die Hosting-Umgebungskonfigurationen von Anwendungen.

[0021] Infrastructure as a Service (IaaS): Die für den Nutzer bereitgestellte Funktion besteht in der Bereitstellung von Verarbeitung, Speicher, Netzwerken und anderen grundlegenden Datenverarbeitungsressourcen, die manchmal als Hypervisor bezeichnet werden, wobei der Nutzer in der Lage ist, beliebige Software zu nutzen und auszuführen, zu der auch Betriebssysteme und Anwendungen gehören können. Der Nutzer verwaltet weder noch kontrolliert er die zu Grunde liegende Cloud-Infrastruktur, aber er hat die Kontrolle über Betriebssysteme, Speicher, genutzte Anwendungen und möglicherweise eine begrenzte Kontrolle über ausgewählte Netzwerkkomponenten (z.B. Host-Firewalls).

[0022] Es gibt folgende Nutzungsmodelle:

Private Cloud: Die Cloud-Infrastruktur wird ausschließlich für eine Organisation betrieben. Sie kann von der Organisation oder einer Drittpartei verwaltet werden und kann innerhalb oder außerhalb der Geschäftsräume vorhanden sein.

[0023] Community Cloud: Die Cloud-Infrastruktur wird von mehreren Organisationen gemeinsam genutzt und unterstützt eine spezielle Community, die gemeinsame Problemstellungen hat (z.B. Berücksichtigung von Zielsetzung, Sicherheitsanforderungen, Richtlinien und Konformität). Sie kann von den Organisationen oder einer Drittpartei verwaltet werden und kann innerhalb oder außerhalb der Geschäftsräume vorhanden sein.

[0024] Public Cloud: Die Cloud-Infrastruktur wird der allgemeinen Öffentlichkeit oder einer großen Industrie-Gruppe zur Verfügung gestellt und ist im Besitz einer Organisation, die Cloud-Dienste verkauft.

[0025] Hybrid Cloud: Die Cloud-Infrastruktur ist eine Zusammensetzung von zwei oder mehr Clouds (Private, Community oder Public), die eindeutige Entitäten bleiben, aber durch eine standardisierte oder proprietäre Technologie vereinigt sind, die eine Daten- und Anwendungsportabilität ermöglicht (z.B. Cloud Bursting für einen Lastausgleich zwischen Clouds).

[0026] Eine Cloud-Datenverarbeitungsumgebung ist dienstorientiert mit einem Schwerpunkt auf Statusungebundenheit, niedriger Kopplung, Modularität und semantischer Interoperabilität. Das Kernstück des Cloud-Computing ist eine Infrastruktur, die ein Netzwerk von miteinander verbundenen Knoten aufweist.

[0027] Unter Bezugnahme auf **Fig. 1** wird eine schematische Darstellung eines Beispiels für einen Cloud Computing-Knoten gezeigt. Der Cloud-Computing-Knoten 10 ist nur ein Beispiel für einen geeigneten Cloud-Computing-Knoten und soll keinerlei Einschränkung in Bezug auf den Schutzzumfang der Verwendung oder Funktionalität von Ausführungsformen der hierin beschriebenen Erfindung andeuten. Ungeachtet dessen lässt sich der Cloud-Computing-Knoten 10 umsetzen und/oder kann jede der im Vorgenannten erläuterten Funktionalitäten ausführen.

[0028] Im Cloud-Computing-Knoten 10 ist das Computersystem/der Server 12 vorhanden, das bzw. der mit zahlreichen anderen Umgebungen oder Konfigurationen für Universal- oder Sonder-Datenverarbeitungssysteme betriebsbereit ist. Zu Beispielen für bekannte Datenverarbeitungssysteme, Umgebungen und/oder Konfigurationen, die für eine Verwendung mit dem Computersystem/Server 12 geeignet sein können, zählen PC-Systeme, Servercomputersysteme, Thin Clients, Thick Clients, Handheld- oder Laptop-Einheiten, Multiprozessorsysteme, Systeme auf der Grundlage von Mikroprozessoren, Set-Top-Boxen, programmierbare Unterhaltungselektronik, Netzwerk-PCs, Minicomputersysteme, Mainframe-Computersysteme und verteilte Cloud-Computing-

Umgebungen, die beliebige der oben genannten Systeme oder Einheiten und dergleichen aufweisen, sie sind aber nicht darauf beschränkt.

[0029] Das Computersystem/der Server 12 kann in dem allgemeinen Kontext von durch ein Computersystem ausführbaren Anweisungen beschrieben werden, wie beispielsweise Programmmodule, die von einem Computersystem ausgeführt werden. Im Allgemeinen können Programmmodule Routinen, Programme, Objekte, Komponenten, Logik, Datenstrukturen usw. enthalten, die bestimmte Aufgaben ausführen oder bestimmte abstrakte Datentypen umsetzen. Das Computersystem/der Server 12 kann in verteilten Cloud-Computing-Umgebungen betrieben werden, wobei Aufgaben von entfernt angeordneten Verarbeitungseinheiten ausgeführt werden, die über ein Datenübertragungsnetzwerk verbunden sind. In einer verteilten Cloud-Computing-Umgebung können Programmmodule sich sowohl in lokalen als auch entfernt angeordneten Computersystem-Speichermedien befinden, die Speichereinheiten enthalten.

[0030] Wie in **Fig. 1** gezeigt, ist das Computersystem/der Server 12 im Cloud Computing-Knoten 10 in Form einer Universal-Datenverarbeitungseinheit gezeigt. Die Komponenten des Computersystems/-Servers 12 können einen oder mehrere Prozessoren oder Verarbeitungseinheiten 16, den Systemarbeitspeicher 28 und den Bus 18 enthalten, der verschiedene Systemkomponenten einschließlich des Systemarbeitspeichers 28 mit dem Prozessor 16 verbindet, sie sind aber nicht darauf beschränkt.

[0031] Der Bus 18 stellt einen oder mehrere von mehreren beliebigen Typen von Busstrukturen dar, einschließlich eines Arbeitsspeicherbusses oder Arbeitsspeicher-Controllers, eines Peripheriebusses, eines beschleunigten Grafikanchlusses und eines Prozessors oder lokalen Busses unter Verwendung von einer aus einer Vielfalt von Busarchitekturen. Als Beispiel und nicht einschränkend enthalten derartige Architekturen einen Industry Standard Architecture- (ISA) Bus, Micro Channel Architecture- (MCA) Bus, Enhanced ISA- (EISA) Bus, einen lokalen Video Electronics Standards Association- (VESA) Bus und Peripheral Component Interconnects- (PCI) Bus.

[0032] Das Computersystem/der Server 12 weist typischerweise eine Vielfalt von Medien auf, die von einem Computersystem gelesen werden können. Derartige Medien können alle verfügbaren Medien sein, auf die von dem Computersystem/Server 12 zugegriffen werden kann, und sie enthalten sowohl flüchtige als auch nicht flüchtige Medien, austauschbare und nicht austauschbare Medien.

[0033] Der Arbeitsspeicherspeicher 28 kann vom Computersystem lesbare Medien in Form von flüchtigem Arbeitsspeicher enthalten, wie beispielsweise den Direktzugriffsarbeitsspeicher (RAM) 30 und/oder den Cache-Zwischenarbeitsspeicher 32. Das Computersystem/der Server 12 kann ferner andere austauschbare/nicht austauschbare, flüchtige/nicht flüchtige Computersystem-Speichermedien enthalten. Nur als Beispiel kann das Speichersystem 34 für ein Lesen von und ein Schreiben auf nicht austauschbare, nicht flüchtige Magnetdatenträger bereitgestellt werden (nicht gezeigt und typischerweise als „Festplatte“ bezeichnet). Obwohl nicht gezeigt, können ein Magnetplattenlaufwerk für ein Auslesen und Beschreiben einer austauschbaren, nicht flüchtigen Magnetplatte (z.B. eine Diskette) und eine optisches Plattenlaufwerk für ein Auslesen oder Beschreiben einer austauschbaren, nicht flüchtigen optischen Platte wie einem CD-ROM, DVD-ROM oder andere optische Datenträger bereitgestellt werden. In solchen Fällen können alle über eine oder mehrere Datenträgerschnittstellen mit dem Bus 18 verbunden werden. Wie ferner im Folgenden dargestellt und beschrieben wird, kann der Arbeitsspeicher 28 mindestens ein Programmprodukt mit einem Satz (z.B. mindestens einem) von Programmmodulen enthalten, die so konfiguriert sind, dass sie die Funktionen von Ausführungsformen der Erfindung ausführen.

[0034] Das Programm/Dienstprogramm 40, das einen Satz von (mindestens einem) von Programmmodulen 42 aufweist, kann beispielsweise und nicht einschränkend im Arbeitsspeicher 28 gespeichert werden, ebenso wie ein Betriebssystem, ein oder mehrere Anwendungsprogramme, andere Programmmodule und Programmdateien. Das Betriebssystem, ein oder mehrere Anwendungsprogramme, weitere Programmmodule und Programmdateien oder eine Kombination davon können alle jeweils eine Umsetzung einer Netzwerkumgebung enthalten. Die Programmmodule 42 führen im Allgemeinen die Funktionen und/oder Methodiken von Ausführungsformen der Erfindung aus, wie hierin beschrieben.

[0035] Das Computersystem/der Server 12 kann auch mit einer oder mehreren externen Einheiten 14 Daten austauschen, wie beispielsweise einer Tastatur, einer Zeigeeinheit, der Anzeige 24 usw.; einer oder mehreren Einheiten, die es einem Benutzer ermöglichen, mit dem Computersystem/Server 12 zu interagieren; und/oder allen Einheiten (z.B. Netzwerkkarte, Modem usw.), die es dem Computersystem/Server 12 ermöglichen, mit einer oder mehreren anderen Datenverarbeitungseinheiten Daten auszutauschen. Ein derartiger Datenaustausch kann über Eingabe-/Ausgabe-(E/A) Schnittstellen 22 erfolgen. Außerdem kann das Computersystem/der Server 12 mit einem oder mehreren Netzwerken, wie beispielsweise einem lokalen Netz (LAN), einem allge-

meinen Weitverkehrsnetz (WAN) und/oder einem öffentlichen Netzwerk (z.B. dem Internet), über den Netzwerkadapter 20 Daten austauschen. Wie dargestellt, tauscht der Netzwerkadapter 20 mit den anderen Komponenten des Computersystems/Servers 12 über den Bus 18 Daten aus. Es sollte klar sein, dass, obwohl nicht gezeigt, andere Hardware- und/oder Software-Komponenten in Verbindung mit dem Computersystem/Server 12 verwendet werden könnten. Beispiele dafür, ohne darauf beschränkt zu sein, sind: Mikrocode, Einheits-treiber, redundante Verarbeitungseinheiten, externe Plattenlaufwerk-Arrays, RAID-Systeme, Bandlaufwerke und Speichersysteme zur Datenarchivierung usw.

[0036] Unter folgender Bezugnahme auf **Fig. 2** ist die erläuternde Cloud Computing-Umgebung 50 dargestellt. Wie gezeigt, weist die Cloud-Computing-Umgebung 50 einen oder mehrere Cloud-Computing-Knoten 10 auf, mit denen lokale Datenverarbeitungseinheiten Daten austauschen können, die von Nutzern der Cloud verwendet werden, wie beispielsweise dem Personal Digital Assistant (PDA) oder dem Mobiltelefon 54A, dem Desktop-Computer 54B, dem Laptop-Computer 54C und/oder dem Fahrzeug-Computersystem 54N. Die Knoten 10 können untereinander Daten austauschen. Sie können (nicht gezeigt) physisch oder virtuell in einem oder mehreren Netzwerken gruppiert werden, wie beispielsweise Private, Community, Public oder Hybrid Cloud, wie hierin oben beschrieben, oder eine Kombination davon sein. Damit hat die Cloud-Computing-Umgebung 50 die Möglichkeit, eine Infrastruktur, Plattformen und/oder Software als Dienste anzubieten, für die ein Cloud-Nutzer keinerlei Ressourcen auf einer lokalen Datenverarbeitungseinheit vorhalten muss. Es versteht sich, dass die in **Fig. 2** gezeigten Typen von Datenverarbeitungseinheiten 54A bis N nur zur Veranschaulichung dienen sollen, und dass die Cloud-Computing-Knoten 10 und die Cloud-Computing-Umgebung 50 mit jedem Typ einer computerisierten Einheit über jeden Typ von Netzwerk und/oder eine über ein Netzwerk adressierbare Verbindung (z.B. unter Verwendung eines Web-Browsers) Daten austauschen können.

[0037] Unter folgender Bezugnahme auf **Fig. 3** wird ein Set von funktionalen Abstraktionsschichten gezeigt, die von der Cloud Computing-Umgebung 50 (**Fig. 2**) bereitgestellt werden. Es sollte von vornherein verstanden werden, dass die in **Fig. 3** gezeigten Komponenten, Schichten und Funktionen nur zu Beispielszwecken dienen sollen und Ausführungsformen der Erfindung nicht darauf beschränkt sind. Wie dargestellt, werden die folgenden Schichten und entsprechenden Funktionen bereitgestellt:

Die Hardware- und Software-Schicht 60 enthält Hardware- und Software-Komponenten. Zu Beispielen für Hardware-Komponenten zählen Mainframes, in einem Beispiel zSeries®-Sys-

teme von IBM®; Server auf der Grundlage einer RISC-(Reduced Instruction Set Computer) Architektur, in einem Beispiel pSeries®-Systeme von IBM; xSeries®-Systeme von IBM; BladeCenter®-Systeme von IBM; Speichereinheiten; Netzwerke und Netzwerkkomponenten. Zu Beispielen für Software-Komponenten zählen Software für Netzwerkanwendungsserver, in einem Beispiel Software für WebSphere®-Anwendungsserver von IBM; und Datenbanksoftware, in einem Beispiel Software für DB2®-Datenbanken von IBM. (IBM, zSeries, pSeries, xSeries, BladeCenter, WebSphere, und DB2 sind Marken der International Business Machines Corporation, die weltweit unter zahlreichen Gerichtsbarkeiten registriert sind.)

[0038] Die Virtualisierungsschicht 62 stellt eine Abstraktionsschicht bereit, von der aus die folgenden Beispiele für virtuelle Entitäten bereitgestellt werden können: virtuelle Server; virtueller Speicher; virtuelle Netzwerke, einschließlich virtuelle private Netzwerke; virtuelle Anwendungen und Betriebssysteme; und virtuelle Clients.

[0039] In einem Beispiel kann die Verwaltungsschicht 64 die im Folgenden beschriebenen Funktionen bereitstellen. Die Ressourcenbereitstellung 70 sorgt für eine dynamische Beschaffung von Datenverarbeitungsressourcen und weiteren Ressourcen, die zum Ausführen von Aufgaben innerhalb der Cloud-Computing-Umgebung eingesetzt werden. Messung und Preisbestimmung ermöglichen beim Einsatz von Ressourcen innerhalb der Cloud-Computing-Umgebung eine Kostenverfolgung und eine Abrechnung oder Rechnungsstellung für die Inanspruchnahme dieser Ressourcen. In einem Beispiel können diese Ressourcen Lizenzen für Anwendungssoftware aufweisen. Die Sicherheit stellt für Nutzer der Cloud und Aufgaben eine Identitätsüberprüfung sowie Schutz für Daten und andere Ressourcen bereit. Ein Benutzerportal stellt den Zugang zur Cloud-Computing-Umgebung für Nutzer und Systemadministratoren bereit. Die Service-Level- (Dienstgüte) Verwaltung sorgt für die Zuordnung und Verwaltung von Cloud-Computing-Ressourcen, so dass erforderliche Service-Levels eingehalten werden. Planung und Vertragserfüllung des Service Level Agreement (SLA) (Dienstgütevereinbarung) stellen eine Vorab-Vereinbarung für und Beschaffung von Cloud-Computing-Ressourcen bereit, für die gemäß eines SLA eine zukünftige Anforderung erwartet wird.

[0040] Die Arbeitslastenschicht 66 stellt Beispiele für die Funktionalität bereit, für die die Cloud-Computing-Umgebung eingesetzt werden kann. Zu Beispielen für Arbeitslasten und Funktionen, die von dieser Schicht aus bereitgestellt werden können, zählen: Zuordnung und Navigation; Software-Entwicklung

und Lifecycle-Management; Bereitstellung von virtuellen Schulungen; Datenanalyseverarbeitung; Transaktionsverarbeitung; und Mobile Desktop.

[0041] Fig. 4 zeigt ein Beispiel der Ressourcenbereitstellungseinheit 70 für kostenbewusstes Clustering zum Abgleichen von Vorlagen mit Anforderungen für Ressourcen gemäß einer bevorzugten Ausführungsform der vorliegenden Erfindung. Insbesondere berücksichtigt die bevorzugte kostenbewusste Clustering-Technik die Gesamtkosten von Ressourcen, die jeweils stark voneinander abweichende und unterschiedliche Kosten haben können, z.B. hohe Kostenunterschiede für Verarbeitungsleistung, Arbeitsspeicher und Plattenspeicherplatz. Im Wesentlichen wird nach einem Bestimmen einer Mindestkostenübereinstimmung für jede Serveranforderung die Vorlagenliste gemäß der Anzahl von erfüllten Anforderungen einer vorgegebenen Vorlage optimiert. Die vorliegende Erfindung weist eine Anwendung auf Cloud-Computing auf, wodurch es Anbietern gestattet wird, die Kosten-Nutzen-Beziehungen zwischen der Anzahl von VM-Vorlagen, der Ebene der erfüllten Serveranforderungen und den potenziell weit voneinander abweichenden Kosten beim Erfüllen der Anforderungen zu analysieren.

[0042] Die bevorzugte Ressourcenbereitstellungseinheit 70 enthält den Speicher für Konfigurationen 72, den Speicher für Serveranforderungen 74, den Speicher für Eingabeparameter 76, den Vorlagenspeicher 78 und den Vorlagengenerator 80. Der Speicher für Konfigurationen 72, der Speicher für Serveranforderungen 74, der Speicher für Eingabeparameter 76 und der Vorlagenspeicher 78 können einzelne Speichereinheiten sein oder in einem einzigen Speicher gespeichert werden, z.B. im Speicher 34 von Fig. 1. Der bevorzugte Vorlagengenerator 80 setzt eine bevorzugte kostenbewusste Clustering-Technik um, die die Anzahl von Vorlagen auswählt und bestimmt, die zum Erfüllen der Datenverarbeitungsanforderungen für einen Satz von Serveranforderungen erforderlich sind.

[0043] Der Speicher für Konfigurationen 72 enthält Ressourcenkosten und zugehörige Attribute. Zum Beispiel kann ein Eintrag für eine Servervorlage mit einer virtuellen zentralen Verarbeitungseinheit (CPU) von 2,1 GHz, einem Plattenspeicher von 1 Terabyte (1 TB) und 2 GB Arbeitsspeicher Kosten von \$ 0,08 pro Stunde angeben. Vorzugsweise enthält der Speicher für Serveranforderungen 74 eine Datenbank für Serveranforderungen mit vorher empfangenen Serveranforderungen, die nach Ressourcentyp definiert sind, wie z.B. CPU (oder CPU-Entsprechung, z.B. Dual-Core 64-Bit i86), Arbeitsspeicher und Plattenspeicherplatz. Der Speicher für Eingabeparameter 76 speichert eine Angabe, ob eine Option „Beste Vorlagen“ ausgewählt wurde, sodass der Vorlagengenerator 80 alle Vorlagen auflistet oder die Werte von k

und einen Vorlagenprozentsatz speichert, die verwendet werden, wenn und falls eine Option für Top-k-Vorlagen ausgewählt wurde. Der bevorzugte Vorlagengenerator 80 erzeugt eine erste Gruppe von Vorlagen aus der Anzahl von verfügbaren Kombinationen. Danach bestimmt der Vorlagengenerator 80 durch Umsetzen einer bevorzugten kostenbewussten Clustering-Technik die Anzahl von Vorlagen, die zum Erfüllen der Erfordernisse der Serveranforderung in einer Mindestkostenübereinstimmung erforderlich sind. Nach dem Berechnen der Mindestkostenübereinstimmung für jede Serveranforderung optimiert der Vorlagengenerator 80 die Vorlagenliste gemäß der Anzahl von erfüllten Anforderungen einer bestimmten Vorlage und speichert optimierte Listen im Vorlagenspeicher 78.

[0044] Vorzugsweise werden Serveranforderungen in einer Datenbank für Anforderungen im Speicher für Serveranforderungen 74 gespeichert, und jede Anforderung wird nach Ressourcentyp definiert, z.B. CPU, Arbeitsspeicher und Platte. Der bevorzugte Vorlagengenerator 80 erzeugt eine vollständige Liste von Vorlagen, um alle Server-Bedarfsanforderungen zu erfüllen, z.B. auf der Grundlage einer Option „Beste Vorlagen“, die im Speicher für Eingabeparameter 76 angegeben sind. Alternativ erzeugt der bevorzugte Vorlagengenerator 80 die Top-k-Vorlagen, die die Anforderungen der meisten Serveranforderungen erfüllen, z.B. unter Verwendung der Option „Top-k-Vorlagen“, die auch im Speicher für Eingabeparameter 76 angegeben sind.

[0045] Der Vorlagengenerator 80 berechnet eine Mindestkostenübereinstimmung für jede Serveranforderung aus einer vollständigen Liste von Vorlagen, um alle Server-Bedarfsanforderungen zu erfüllen. Aus diesen Mindestkostenübereinstimmungen wählt der Vorlagengenerator 80 den k-Wert (k_1 , k_2 , ..., k_n) für die Top-k-Liste, die die Erfordernisse der meisten Serveranforderungen erfüllt, z.B. durch Erfüllen eines Prozentsatzes oder niedrigste Gesamtkosten. Der sich daraus ergebende Satz der am häufigsten angeforderten Vorlagen ist eine Vorlagen-Untergruppe, gibt die Gesamtkosten an und wird als der endgültige kleinere Satz von Vorlagen gewählt.

[0046] Fig. 5 zeigt ein Beispiel der Auswahl 80A der Top-k-Vorlagen durch den bevorzugten Vorlagengenerator 80. Zunächst werden die Konfigurationen 72 mit jeder Serveranforderung im Speicher für Anforderungen 74 verglichen, um die Mindestkostenübereinstimmung 86 für jede Anforderung zu bestimmen 82. Als Nächstes werden alle Vorlagen verworfen 88, die die Mindestanforderungen für mindestens eine der Anforderungen in dem Satz nicht erfüllen. Die restlichen Vorlagen werden gemäß der Anzahl von übereinstimmenden Anforderungen sortiert 90, und die sortierten Vorlagen werden gespeichert 92. Die

Top-k-Listen 94 und die sortierten Vorlagen 92 werden verwendet, um jedes k zu bestimmen 96. Die Anzahl von erfüllten Anforderungen für jedes k und die Kosten werden z.B. zur Anzeige an die Ausgabe 98 übergeben.

[0047] Zuerst werden also die Server-Mindestanforderungen für jede Anforderung bestimmt 82, indem zum Beispiel die Anforderungen 74 an CPU, Arbeitsspeicher und Platte mit der Vorlage mit den kleinsten Werten in den entsprechenden aufgelisteten Zielkonfigurationen 72 abgeglichen werden. In einem einfachen Beispiel kann eine Zielliste für CPU-Kapazität eine Auswahl von 1, 2, 4, 8, 12, 16, ..., $\max_{\text{CPU}}$ -Kernen gestatten, wobei $\max_{\text{CPU}}$ die maximal verfügbare Anzahl ist. Desgleichen können der auswählbare Arbeitsspeicher und Plattenspeicherplatz jeweils 1, 2, 4, 8, 12, 16, 32, ..., $\max_{\text{Arbeitsspeicher}}$ Gigabyte (GB) und 100, 200, 300, 400, 500, 600, ..., $\max_{\text{Festplatte}}$ GB sein. Obwohl CPU, Arbeitsspeicher und Plattenkapazität für dieses Beispiel verwendet werden, werden üblicherweise weitere Ressourcen in einer bevorzugten Lösung berücksichtigt, wie beispielsweise Netzwerkbandbreite.

[0048] Die Mindestkostenübereinstimmung 86 für jede Serveranforderung stellt eine untere Grenze für die Kosten zum Erfüllen von Anforderungen für jeden Server auf der Grundlage der aufgelisteten Ressourcen bereit, z.B. CPU, Arbeitsspeicher und Platte. Ferner verwendet der bevorzugte Vorlagengenerator 80 die Mindestkosten 86 zum Bestimmen, ob alle Anforderungen zu Mindestkosten erfüllt werden können. Jede Übereinstimmung mit einer Mindestkostenvorlage wird hierin als „Mindestkostenübereinstimmung“ bezeichnet.

[0049] Fig. 6A zeigt ein Beispiel zum Auffinden von Mindestkostenübereinstimmungen 82, und Fig. 6B zeigt ein Beispiel der Ressourcen-Übereinstimmungstabelle 86, die zum Auflisten von Mindestkostenübereinstimmungen gemäß einer bevorzugten Ausführungsform der vorliegenden Erfindung erstellt wird. Das Auffinden von Mindestkostenübereinstimmungen 82 beginnt bei 822 mit dem Erstellen einer Kostenliste für Mindestkonfigurationen 824 und der Ressourcen-Übereinstimmungstabelle 86, die in diesem Beispiel ein Feld für die Anzahl von Kernen 860, den erforderlichen Arbeitsspeicher 862 und den erforderlichen Plattenspeicherplatz 864 enthält. Die Anforderungen 74 werden in der angeforderten Reihenfolge ausgewählt 826, bis alle Anforderungen ausgewählt worden sind 828. Solange noch nicht ausgewählte Anforderungen übrig sind, erfolgt der Vergleich 830 der aktuell ausgewählten Anforderung mit der Zielressourcenliste 832, um eine Vorlage mit Mindestressourcen zu suchen, die mit den Erfordernissen der Anforderung übereinstimmt. Die Ressourcenübereinstimmung wird in die Ressourcen-Übereinstimmungstabelle 86 eingegeben, und unter

Verwendung der Liste von Ressourcenkosten 834 werden die Kosten der Übereinstimmung bestimmt 836. Die Kosten der übereinstimmenden Vorlage werden in die Kostenliste für Mindestkonfigurationen 824 eingegeben, und die nächste Anforderung wird ausgewählt 826. Sobald alle Anforderungen ausgewählt worden sind 828, werden die übereinstimmenden Anforderungen 86 und die dazugehörigen Kosten 824 zur anschließenden Filterung und Analyse zurückgegeben 838.

[0050] Jede Vorlage definiert eine zuordenbare Kombination von CPU, Arbeitsspeicher und Platten-speicherplatz. Daher werden nach dem Bestimmen der zugehörigen Kosten (824 in **Fig. 6A**) für jede übereinstimmende Anforderung (86 in **Fig. 5**) nicht verwendete Vorlagen (z.B. Duplikate) verworfen 88. Als Ergebnis dessen erfüllt der übrige Satz von eindeutigen Vorlagen mindestens eine Anforderung; und ferner kann jede Anforderung durch mindestens eine Vorlage erfüllt werden. Während ferner einige einzelnen Vorlagen für gewöhnlich zum Erfüllen von mehr als einer Anforderung verwendet werden können, können andere für gewöhnlich nicht verwendet werden und keinerlei Anforderung erfüllen. Der Satz von eindeutigen Vorlagen, die aus den Mindestkostenübereinstimmungen 86 ausgewählt wurden, enthält jedoch Vorlagen zum Erfüllen aller Anforderungen 88.

[0051] Als Nächstes wird der ausgewählte Satz von Vorlagen nach dem ausgewählten Kriterium sortiert 90, z.B. Speicher in aufsteigender Reihenfolge, Arbeitsspeicher in aufsteigender Reihenfolge, nach CPU-Kernen oder vorzugsweise nach der Anzahl von Anforderungen in absteigender Reihenfolge, die durch jede Vorlage erfüllt wurden. Wenn eine Liste von Vorlagen mit mindestens einer Übereinstimmung 88 gesammelt und sortiert worden ist, und die Liste 90 sortiert worden ist, filtert der bevorzugte Vorlagengenerator 80 die Liste von Anforderungen, um ein Auswählen 96 einer verschiedenen Anzahl von häufigsten oder Top-Anforderungen zu unterstützen, die n Vorlagenlisten, n Listen von Top-k-Vorlagen oder Top-k-Listen bereitstellt. Der bevorzugte Vorlagengenerator 80 verwendet die Top-k-Listen zum Identifizieren und Auflisten von Mindestressourcenübereinstimmungen für jeden Server.

[0052] Wie in **Fig. 7** gezeigt, erzeugt 96 der bevorzugte Vorlagengenerator (80 in **Fig. 4**) die endgültige Vorlagenliste 78 durch Umspannen von Anforderungen für die Top-k-Listen mit einem Mindest- und einem Höchstwert. Insbesondere bestimmt der bevorzugte Vorlagengenerator 80 die Anzahl von Anforderungen für jede Liste, die nicht erfüllt werden können, oder die zu Mindestkosten und über den Mindestkosten liegenden Kosten erfüllt werden können.

[0053] Zuerst 960 erstellt der bevorzugte Vorlagengenerator 80 eine Top-k-Anforderungstabelle 98 für die Anzahl von Anforderungen (z.B. 5, 10, 15, 20, 25), welche die Top-Vorlagen erfüllen. Der erste Wert von k (z.B. 5, 10, 15, 20, 25) wird aus der Top-k-Liste 94 ausgewählt 962. Die erste Anforderung wird aus dem Speicher für Anforderungen 74 abgerufen 108. Danach wird die ausgewählte Anforderung mit den Vorlagen in der aktuellen Top-k-Liste verglichen 966, um zu bestimmen, ob irgendeine der Vorlagen, entweder die für mindeste (Kosten) oder darüber liegende (höhere) die Anforderung erfüllt; oder ob die Anforderung von keiner der Top-k-Vorlagen erfüllt werden kann. Dies wird wiederholt 968, wobei die nächste Anforderung ausgewählt 964 und verglichen 966 wird, bis alle Anforderungen verglichen worden sind 968. Als nächstes wird ein Eintrag in die Top-k-Anforderungstabelle 98 für den aktuellen Wert von k vorgenommen 970, der die Anzahl von Anforderungen angibt, die durch diese Vorlagen erfüllt und nicht erfüllt worden sind. Wenn weitere Top-k-Werte 972 aufgelistet sind 94, wird beim Zurückkehren zu 962 der nächste Wert ausgewählt, und die Anzahl von Ergebnissen, die durch die Top-k-Vorlagen erfüllt und nicht erfüllt worden sind, wird bestimmt 962 bis 970. Andernfalls, wenn alle Werte von k ausgewählt worden sind 972, wird die Top-k-Anforderungstabelle 98 als die endgültige Vorlagenliste 78 zurückgegeben. Der bevorzugte Vorlagengenerator 80 verwendet die endgültige Vorlagenliste 78 in Analysen von Kosten-Nutzen-Beziehungen zum Bestimmen der besten Top-k-Liste, z.B. einen Punkt von verminderten Rückgaben, wie durch die Mindestkosten-k-Liste angegeben.

[0054] **Fig. 8** zeigt ein Beispiel einer Anwendung der vorliegenden Erfindung auf einen Satz von siebenhundredsiebenundvierzig (747) Serveranforderungen für k = 5, 10, 15, 20, 25 und 89 (d.h. alle Vorlagen) mit einem Prozentsatz, der bei Mindestkosten (unten), Höchstkosten (Mitte) und durch die Top-k-Vorlagen nicht erfüllt (oben) übereinstimmt.

[0055] Dementsprechend verringert die vorliegende Erfindung sowohl eine Über- als auch Unterbereitstellung von kostspieliger Ressourcenkapazität. z.B. Verarbeitungsleistung, Arbeitsspeicher und Platten-speicherplatz. Da zuordenbare Ressourcen jeweils stark voneinander abweichende und unterschiedliche Kosten haben können, wählt die bevorzugte kostenbewusste Clustering-Technik ferner Vorlagen auf der Grundlage der Ressourcen-Gesamtkosten in jeder Vorlage aus. Somit werden Vorlagen nach Gesamtkosten erzeugt und zusammengefasst, und eine Anzahl von Vorlagen mit niedrigsten Kosten wird ausgewählt, die Serveranforderungen bei optimalen Kosten erfüllen. Ferner erleichtern die erzeugten Vorlagenlisten die Analyse von Kosten-Nutzen-Beziehungen zwischen der Anzahl von verwalteten

Vorlagen, der Anzahl der für jeden Server erfüllten Anforderungen und den Kosten für das Erfüllen dieser Anforderungen.

[0056] Zwar wurde die Erfindung unter Bezugnahme auf bevorzugte Ausführungsformen beschrieben, doch wird der Fachmann erkennen, dass die Erfindung mit Modifizierungen innerhalb des Schutzedankens und Schutzzumfangs der Ansprüche im Anhang ausgeübt werden kann. Alle derartigen Variationen und Modifizierungen sollen unter den Schutzzumfang der Ansprüche im Anhang fallen. Beispiele und Zeichnungen sind dementsprechend als veranschaulichend und nicht als einschränkend zu betrachten.

Patentansprüche

1. Gemeinsam genutztes Ressourcensystem, das aufweist:
eine Vielzahl von Client-Einheiten;
einen oder mehrere Dienstanbieter-Computer, die selektiv Ressourcenkapazität für die Vielzahl von Client-Einheiten verfügbar machen;
ein Netzwerk, wobei Ressourcenkapazität für die Vielzahl von Client-Einheiten von dem einen oder den mehreren Dienstanbieter-Computern über das Netzwerk bereitgestellt wird; und
eine Ressourcenbereitstellungseinheit, die Ressourcenvorlagen für alle zuordenbaren Ressourcenkapazitäts-Konfigurationen erzeugt, eine Untergruppe von erzeugten Ressourcenvorlagen auswählt und eine Anzahl von Ressourcenvorlagen verwaltet, die aus jeder Untergruppe ausgewählt wurden, wobei jede Vorlage eine zuordenbare Konfiguration von Ressourcenkapazität angibt, wobei jeder anfordernden Client-Einheit Ressourcen zugeordnet sind, die durch eine der ausgewählten Ressourcenvorlagen der Untergruppe bestimmt wurden, wobei die Ressourcenbereitstellungseinheit aufweist:
einen Speicher für Konfigurationen (72), der Kosten von zuordenbaren Ressourcen und zugehörige Attribute speichert;
einen Speicher für Serveranforderungen (74), der vorher empfangene Anforderungen von Kapazität speichert;
einen Speicher für Eingabeparameter, der Optionen für Vorlagenlisten (78) speichert;
einen Vorlagengenerator (80), der eine Liste von Ressourcenvorlagen bestimmt, um die vorher empfangene Anforderungen zu optimalen Kosten zu erfüllen, wobei die Liste die Ressourcenvorlagen auflistet, die aus der Untergruppe ausgewählt wurden; und
einen Vorlagenspeicher, der erzeugte Vorlagenlisten speichert, wobei jede ausgewählte Vorlagenliste Mindestkosten für die Vorlagenlisten aufweist, die für den einen oder die mehreren Dienstanbieter-Computer erzeugt wurden.

2. Gemeinsam genutztes Ressourcensystem nach Anspruch 1, wobei der Vorlagengenerator eine Vorlagenliste für jeden Dienstanbieter-Computer erzeugt, wobei jede erzeugte Vorlagenliste in dem Vorlagenspeicher gespeichert wird.

3. Gemeinsam genutztes Ressourcensystem nach Anspruch 1, wobei die Optionen für Vorlagenlisten eine Option für eine vollständige Liste und eine Option für eine Top-k-Liste aufweisen.

4. Gemeinsam genutztes Ressourcensystem nach Anspruch 3, wobei die Option für die Top-k-Liste eine Vielzahl von Werten für k angibt, wobei der Vorlagengenerator Vorlagenlisten für alle Werte von k für jeden Dienstanbieter-Computer erzeugt, wobei eine der erzeugten Vorlagenlisten für jeden Dienstanbieter-Computer ausgewählt und in dem Vorlagenspeicher gespeichert wird.

5. Gemeinsam genutztes Ressourcensystem nach Anspruch 4, wobei jede ausgewählte Vorlagenliste Mindestkosten für die Vorlagenlisten aufweist, die für jeden Dienstanbieter-Computer erzeugt wurden.

6. Gemeinsam genutztes Ressourcensystem nach Anspruch 1, wobei jede Vorlage eine zuordenbare Konfiguration angibt, die eine Anzahl von Kernen, Arbeitsspeicherplatz und Plattenspeicherplatz aufweist.

7. Verfahren zum Zuordnen von Ressourcen, wobei das Verfahren aufweist:
Bereitstellen einer Sammlung von Anforderungen von Ressourcenkapazität von einer Vielzahl von Computern;
Zuordnen von Kosten zu zuordenbaren Computerressourcen;
Erzeugen von Ressourcenvorlagen aus den zuordenbaren Computerressourcen für alle zuordenbaren Konfigurationen;
Abgleichen jeder Anforderung mit mindestens einer Vorlage;
Auswählen der Vorlage mit Mindestkosten für jede Anforderung;
Sortieren der ausgewählten Mindestkostenvorlagen nach der Anzahl von übereinstimmenden Anforderungen; und
Bestimmen der Kosten von übereinstimmenden Anforderungen für eine ausgewählte Anzahl (k) von am häufigsten übereinstimmenden Vorlagen, wobei die k am häufigsten übereinstimmenden Vorlagen in einer Top-k-Liste aufgelistet werden.

8. Verfahren zum Zuordnen von Ressourcen nach Anspruch 7, wobei ein Abgleichen jeder Anforderung aufweist:
Erzeugen einer Liste von Ressourcenvorlagen-Übereinstimmungen, wobei die Liste jede Vorlage

aufflistet, die mit mindestens einer Anforderung übereinstimmt;
 Zuordnen von Kosten zu jeder aufgelisteten Vorlage, und
 Bestimmen einer übereinstimmenden Vorlage für jede Anforderung mit Mindestkosten aus allen Vorlagen mit zugehörigen Kosten.

9. Verfahren zum Zuordnen von Ressourcen nach Anspruch 8, wobei ein Auswählen der übereinstimmenden Vorlage mit Mindestkosten für jede Anforderung aufweist:
 Suchen einer Vorlage mit Mindest-Ressourcenübereinstimmung für jede Anforderung, wobei die Vorlage eine Mindestübereinstimmungsvorlage ist; und
 Bestimmen der Kosten für Ressourcen für die Mindestübereinstimmungsvorlage.

10. Verfahren zum Zuordnen von Ressourcen nach Anspruch 9, wobei ein Bestimmen von Top-k-Listen aufweist:
 Auflisten von sortierten übereinstimmenden Vorlagen nach der Anzahl von Anforderungen, die mit jeder Vorlage übereinstimmen;
 Bestimmen der Kostenspannen für jeden eines k-Werts; und
 Auflisten der Kostenspannen in einer Top-k-Vorlagentabelle.

11. Verfahren zum Zuordnen von Ressourcen nach Anspruch 10, wobei die Kostenspannen für jede Top-k-Liste die Anzahl von Anforderungen enthalten, die zu Mindestkosten und über den Mindestkosten liegenden Kosten erfüllt wurden sowie nicht erfüllt wurden.

12. Verfahren zum Zuordnen von Ressourcen in einer Cloud-Umgebung nach Anspruch 11, wobei Vorlagen in der Top-k-Liste mit den niedrigsten zugehörigen Kosten zum Bereitstellen von Ressourcen in einer Cloud-Umgebung verwendet werden.

13. Verfahren zum Zuordnen von Cloud-Umgebungsressourcen, wobei das Verfahren aufweist:
 Erzeugen von Ressourcenvorlagen aus zuordenbaren Computerressourcen für alle zuordenbaren Konfigurationen;
 Abgleichen jeder Anforderung von Ressourcenkapazität, die in einer Datenbank für Serveranforderungen mit einer oder mehreren Vorlagen enthalten ist, die zuordenbare Cloud-Ressourcen angeben;
 Sammeln von übereinstimmenden Vorlagen mit mindestens einer Übereinstimmung, wobei alle Vorlagen ohne Übereinstimmungen verworfen werden;
 Sortieren der übereinstimmenden Vorlagen nach der Anzahl von übereinstimmenden Anforderungen;
 Bestimmen der Anzahl von Übereinstimmungen für eine Vielzahl (k) von übereinstimmenden Vorlagen, wobei die Übereinstimmungen in einer Top-k-Liste aufgelistet werden; und

Erzeugen von Vorlagenlisten für alle Werte von k, wobei eine der generierten Vorlagenlisten so ausgewählt wird, dass die ausgewählte Vorlagenliste die Mindestkosten der Vorlagenlisten aufweist.

14. Verfahren zum Zuordnen von Cloud-Umgebungsressourcen nach Anspruch 13, wobei jede der Vorlagen eine zuordenbare Konfiguration angibt, die eine Anzahl von Kernen, Arbeitsspeicherplatz und Plattenspeicherplatz aufweist, und ein Abgleichen jeder Anforderung aufweist:
 Erzeugen einer Ressourcen-Übereinstimmungstabelle zum Auflisten jeder übereinstimmenden Anforderung und Vorlage;
 Erzeugen einer Mindestkostenliste zum Auflisten von Kosten, die jeder übereinstimmenden Vorlage zugehörig sind;
 Auswählen einer Anforderung aus der Datenbank für Serveranforderungen;
 Identifizieren einer Vorlage, die der ausgewählten Anforderung entspricht;
 Bestimmen von Kosten, die den zuordenbaren Ressourcen zugehörig sind, die durch die übereinstimmende Vorlage angegeben wurden;
 Hinzufügen der übereinstimmenden Vorlage und Anforderung zur Ressourcen-Übereinstimmungstabelle und der zugehörigen Kosten zu der Mindestkostenliste.

15. Verfahren zum Zuordnen von Cloud-Umgebungsressourcen nach Anspruch 14, wobei ein Abgleichen jeder Anforderung aufweist:
 Bestimmen, ob nicht übereinstimmende Anforderungen in der Datenbank für Serveranforderungen verbleiben; und bis keine nicht ausgewählten Anforderungen mehr übrig sind,
 Zurückkehren zum Auswählen einer nächsten Anforderung und Bestimmen einer übereinstimmenden Vorlage und von zugehörigen Kosten für die nächste Anforderung.

16. Verfahren zum Zuordnen von Cloud-Umgebungsressourcen nach Anspruch 15, wobei ein Bestimmen von Top-k-Listen aufweist:
 Sortieren einer Tabelle von übereinstimmenden Vorlagen mit der Anzahl von Anforderungen, die mit zuordenbaren Ressourcen für jede Vorlage übereinstimmen;
 Auswählen eines ersten k-Werts;
 Auswählen einer Anforderung aus der Datenbank für Serveranforderungen;
 Bestimmen von Kostenspannen für den k-Wert;
 Zurückkehren zum Auswählen einer nächsten Anforderung, bis alle Anforderungen ausgewählt worden sind; und wenn alle Anforderungen ausgewählt worden sind,
 Auflisten der Kostenspannen in einer Top-k-Vorlagentabelle.

17. Verfahren zum Zuordnen von Cloud-Umgebungsressourcen nach Anspruch 16, wobei die Kostenspannen für jede Top-k-Liste die Anzahl von Anforderungen enthalten, die zu Mindestkosten und über den Mindestkosten liegenden Kosten erfüllt wurden sowie nicht erfüllt wurden, und Vorlagen in der Top-k-Liste mit den niedrigsten zugehörigen Kosten zum Bereitstellen von Ressourcen verwendet werden.

18. Computerprogrammprodukt zum Zuordnen von Cloud-Umgebungsressourcen, wobei das Computerprogrammprodukt einen von einem Computer verwendbaren Datenträger mit einem darauf gespeicherten computerlesbaren Programmcode aufweist, wobei der computerlesbare Programmcode einen Computer veranlasst, den Code auszuführen zum:
Erzeugen von Ressourcenvorlagen aus zuordenbaren Cloud-Ressourcen für alle zuordenbaren Konfigurationen;
Erzeugen einer Ressourcen-Übereinstimmungstabelle zum Auflisten jeder übereinstimmenden Anforderung und Vorlage;
Erzeugen einer Mindestkostenliste zum Auflisten von Kosten, die jeder übereinstimmenden Vorlage zugehörig sind;
Auswählen einer Anforderung aus der Datenbank für Serveranforderungen;
Identifizieren einer Vorlage, die mit der ausgewählten Anforderung übereinstimmt;
Bestimmen von Kosten, die zuordenbaren Ressourcen zugehörig sind, die durch die übereinstimmende Vorlage angegeben werden;
Hinzufügen der übereinstimmenden Vorlage und Anforderung zur Ressourcen-Übereinstimmungstabelle und der zugehörigen Kosten für die erste Übereinstimmung zu der Mindestkostenliste; und bis alle Anforderungen ausgewählt sind,
Zurückkehren zum Auswählen einer nächsten Anforderung und Bestimmen einer übereinstimmenden Vorlage und von zugehörigen Kosten für die nächste Anforderung.

19. Computerprogrammprodukt zum Zuordnen von Cloud-Umgebungsressourcen nach Anspruch 18, wobei ein Bestimmen von Top-k-Listen aufweist, dass veranlasst wird, dass der Computer den Code ausführt zum
Sortieren der Tabelle von übereinstimmenden Vorlagen für die Anzahl von Anforderungen, die mit zuordenbaren Ressourcen für jede Vorlage übereinstimmen;
Auswählen eines k-Werts;
Auswählen einer Anforderung aus der Datenbank für Serveranforderungen; Bestimmen von Kostenspannen für den k-Wert;
Zurückkehren zum Auswählen einer nächsten Anforderung, bis alle Anforderungen ausgewählt worden sind;
Auflisten der Kostenspannen in einer Top-k-Vorla-

gentabelle, wenn alle Anforderungen ausgewählt worden sind; und
Zurückkehren zum Auswählen von nächsten k-Werten, bis alle k-Werte ausgewählt worden sind.

20. Computerprogrammprodukt zum Zuordnen von Cloud-Umgebungsressourcen nach Anspruch 19, wobei die Kostenspannen für jede Top-k-Liste die Anzahl von Anforderungen enthalten, die zu Mindestkosten und über den Mindestkosten liegenden Kosten erfüllt wurden sowie nicht erfüllt wurden, wobei der Computer den Code unter Verwendung der Vorlagen in der Top-k-Liste mit den niedrigsten zugehörigen Kosten zum Bereitstellen von Cloud-Ressourcen ausführt.

21. Computerprogrammprodukt zum Zuordnen von Cloud-Ressourcen, wobei das Computerprogrammprodukt einen von einem Computer verwendbaren Datenträger mit einem darauf gespeicherten computerlesbaren Programmcode aufweist, wobei der computerlesbare Programmcode aufweist:
computerlesbare Programmcode-Mittel, um selektiv Computer-Ressourcenkapazität von Dienst Anbietern zum Zuordnen zu anfordernden Client-Einheiten verfügbar zu machen;
computerlesbare Programmcode-Mittel, um Ressourcenvorlagen aus zuordenbarer Ressourcenkapazität für alle zuordenbaren Konfigurationen zu erzeugen;
computerlesbare Programmcode-Mittel, um Kosten und zugehörige Attribute von zuordenbaren Ressourcen zu speichern;
computerlesbare Programmcode-Mittel, um vorher empfangene Anforderungen von Kapazität zu speichern;
computerlesbare Programmcode-Mittel, um eine Anzahl der Ressourcenvorlagen auszuwählen und die ausgewählte Anzahl von Ressourcenvorlagen zu verwalten, wobei jede Vorlage eine zuordenbare Konfiguration von Ressourcenkapazität angibt, wobei jeder anfordernden Client-Einheit Ressourcen in einem kostenbewussten Clustering zugeordnet sind, das durch eine der ausgewählten Ressourcenvorlagen bestimmt wurde;
computerlesbare Programmcode-Mittel, um Kosten von zuordenbaren Ressourcen und zugehörige Attribute zu speichern;
computerlesbare Programmcode-Mittel, um Serveranforderungen der vorher empfangene Anforderungen von Kapazität zu speichern;
computerlesbare Programmcode-Mittel, um Eingabeparameter der Optionen für Vorlagenlisten zu speichern;
computerlesbare Programmcode-Mittel zum Bestimmen einer Liste von Ressourcenvorlagen, um die vorher empfangene Anforderungen zu optimalen Kosten zu erfüllen, wobei die Liste die Ressourcenvorlagen auflistet, die aus der Untergruppe ausgewählt wurden; und

computerlesbare Programmcode-Mittel, um erzeugte Vorlagenlisten zu speichern, wobei jede ausgewählte Vorlagenliste Mindestkosten für die Vorlagenlisten aufweist, die für den einen oder die mehreren Dienstanbieter-Computer erzeugt wurden.

22. Computerprogrammprodukt zum Zuordnen von Cloud-Ressourcen nach Anspruch 21, das ferner aufweist:

computerlesbare Programmcode-Mittel, um Optionen für Vorlagenlisten zu speichern, die eine Option für eine vollständige Liste und eine Option für eine Top-k-Liste aufweisen;

computerlesbare Programmcode-Mittel, um eine optimale Liste von Vorlagen zum Erfüllen der vorher empfangenen Anforderungen für jeden Dienstanbieter-Computer zu erzeugen; und

computerlesbare Programmcode-Mittel, um erzeugte Vorlagenlisten zu speichern.

23. Computerprogrammprodukt zum Zuordnen von Cloud-Ressourcen nach Anspruch 22, wobei die Option der Top-k-Liste eine Vielzahl von Werten für k angibt, wobei das computerlesbare Programmcode-Mittel zum Erzeugen der optimalen Liste Vorlagenlisten für alle Werte von k für jeden der Dienstanbieter-Computer erzeugt und eine der erzeugten Vorlagenlisten für jeden Dienstanbieter-Computer auswählt, wobei die erzeugten Vorlagenlisten in den Top-k-Listen gespeichert sind.

24. Computerprogrammprodukt zum Zuordnen von Cloud-Ressourcen nach Anspruch 22, wobei jede Vorlage eine zuordenbare Konfiguration angibt, die eine Anzahl von Kernen, Arbeitsspeicherplatz und Plattenspeicherplatz aufweist, und jede ausgewählte Vorlagenliste Mindestkosten für die Vorlagenlisten aufweist, die für jeden Dienstanbieter-Computer erzeugt worden sind.

Es folgen 8 Seiten Zeichnungen

10

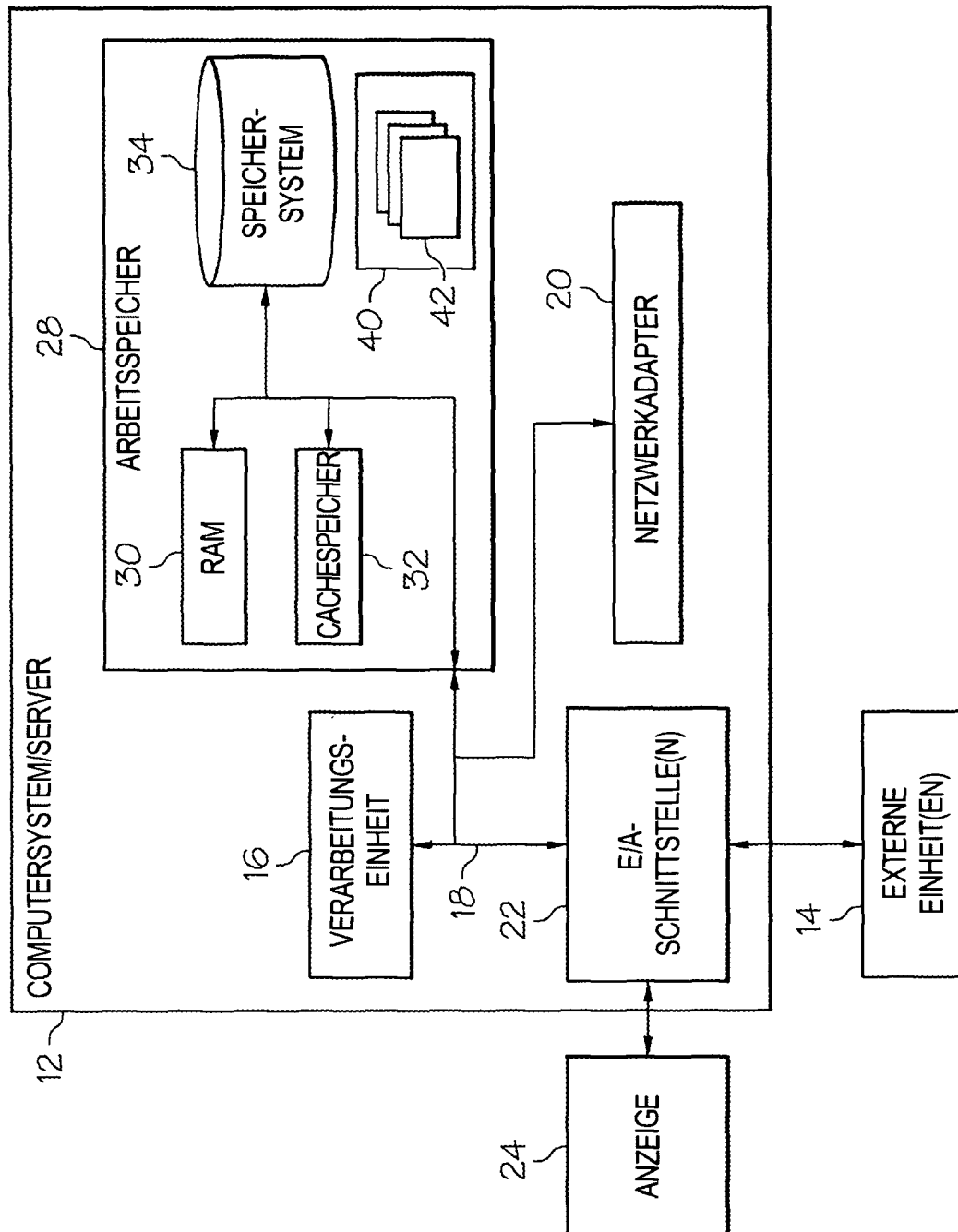


Fig. 1

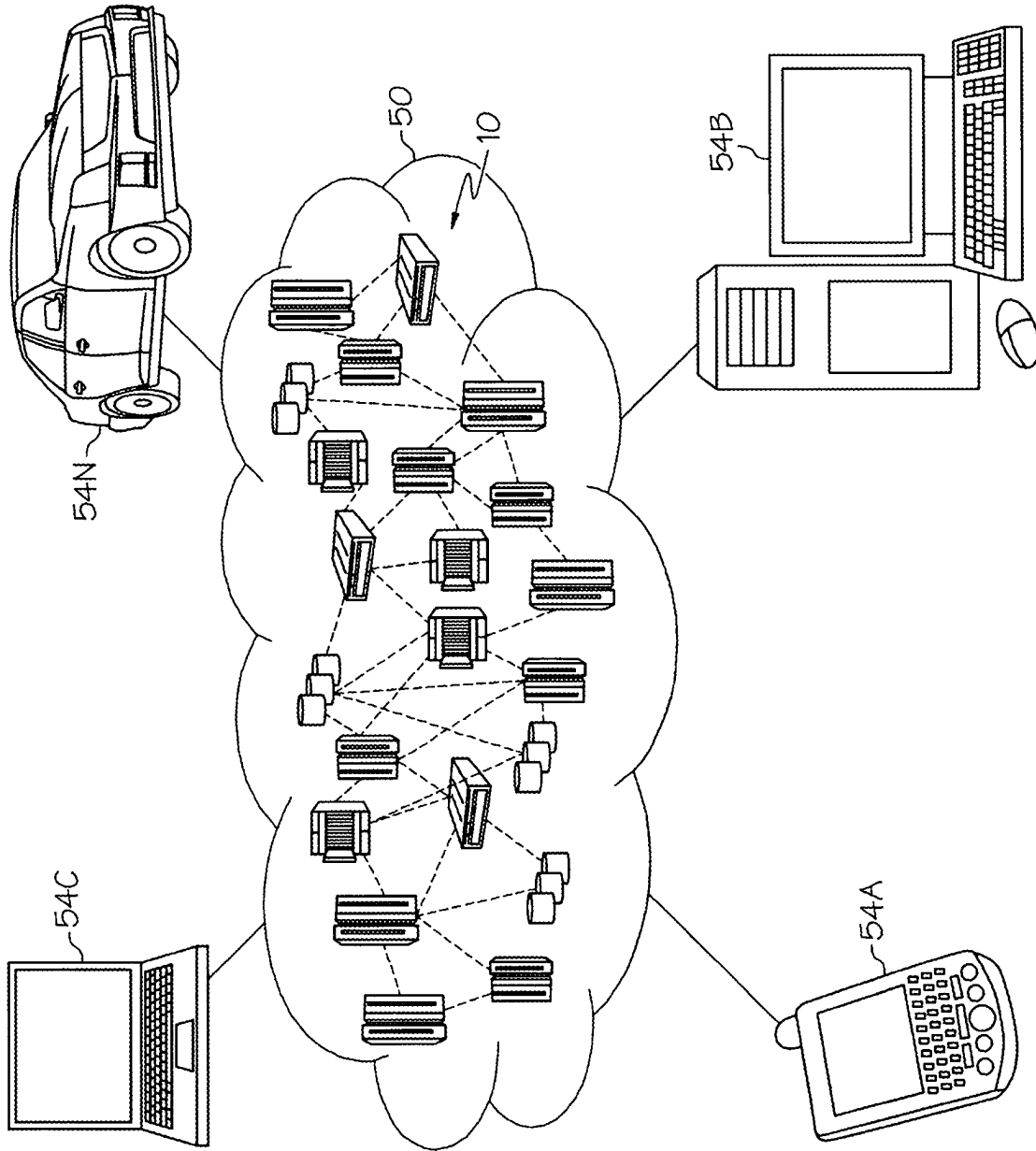


Fig. 2

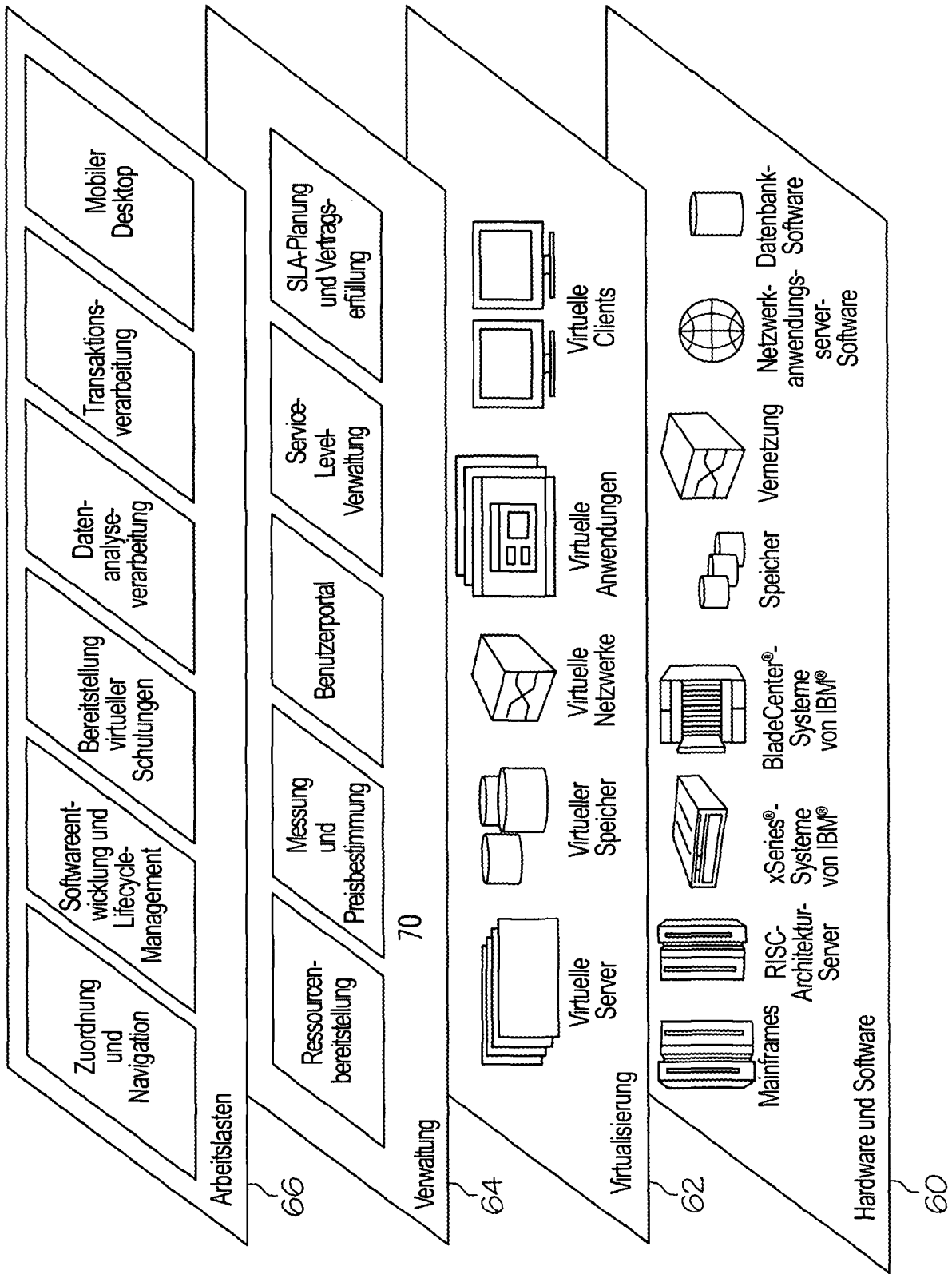


Fig. 3

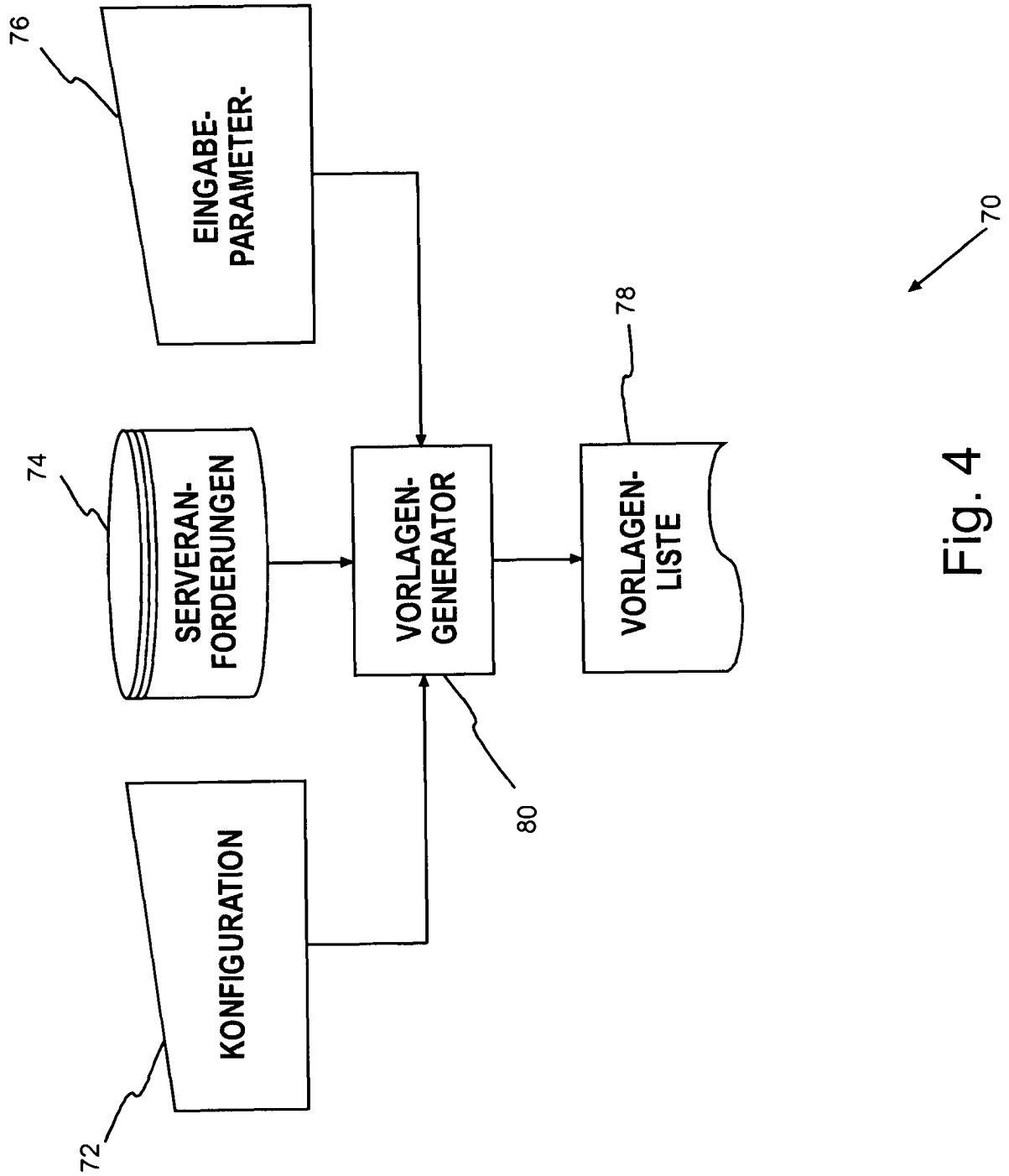


Fig. 4

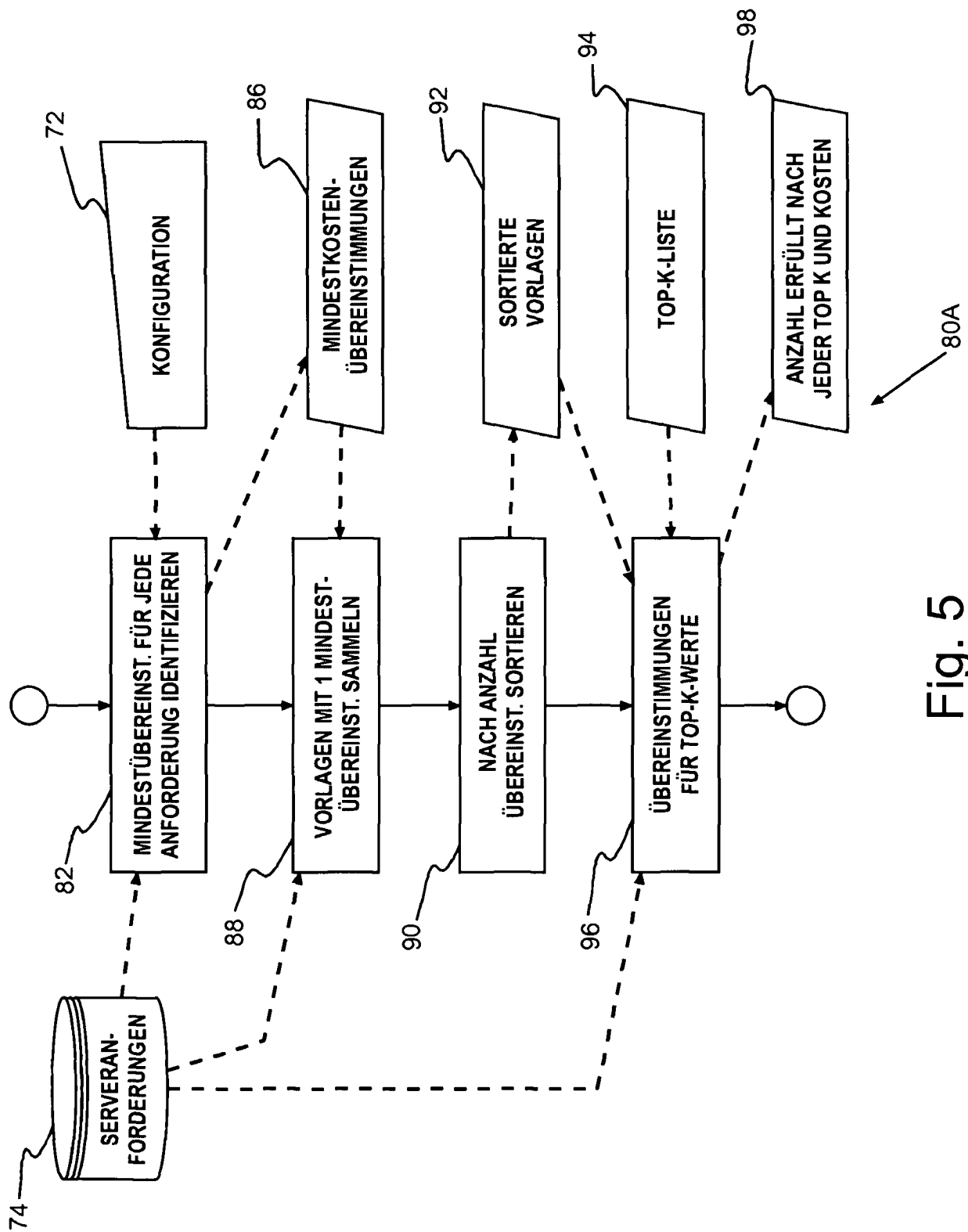


Fig. 5

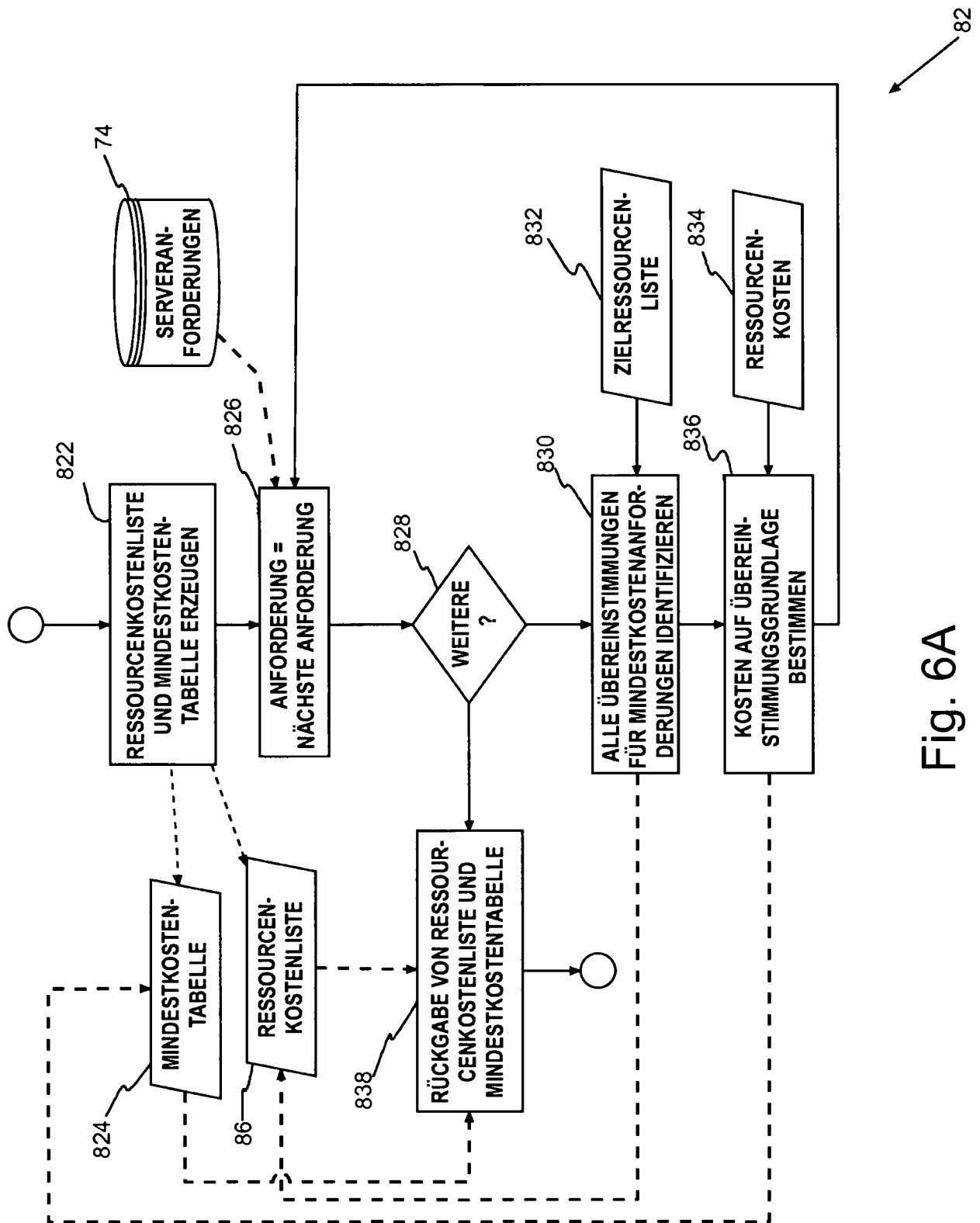


Fig. 6A

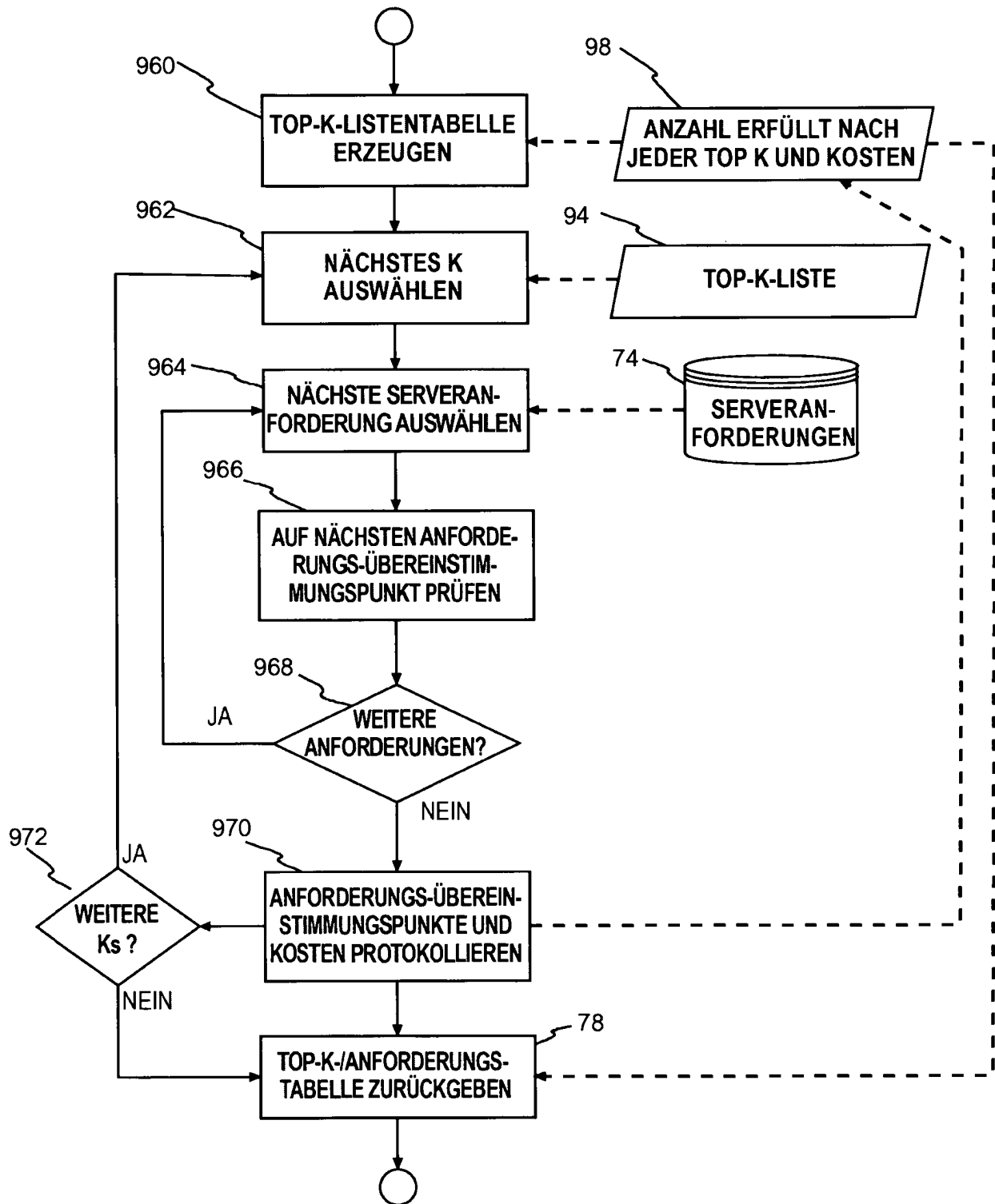


Fig. 7

96

	860	862	864
	↓	↓	↓
ANFORDERUNG	CPU	ARBEITSSPEICHER	PLATTE
01	1	2	300
02	2	12	500
03	1	8	600

Fig. 6B

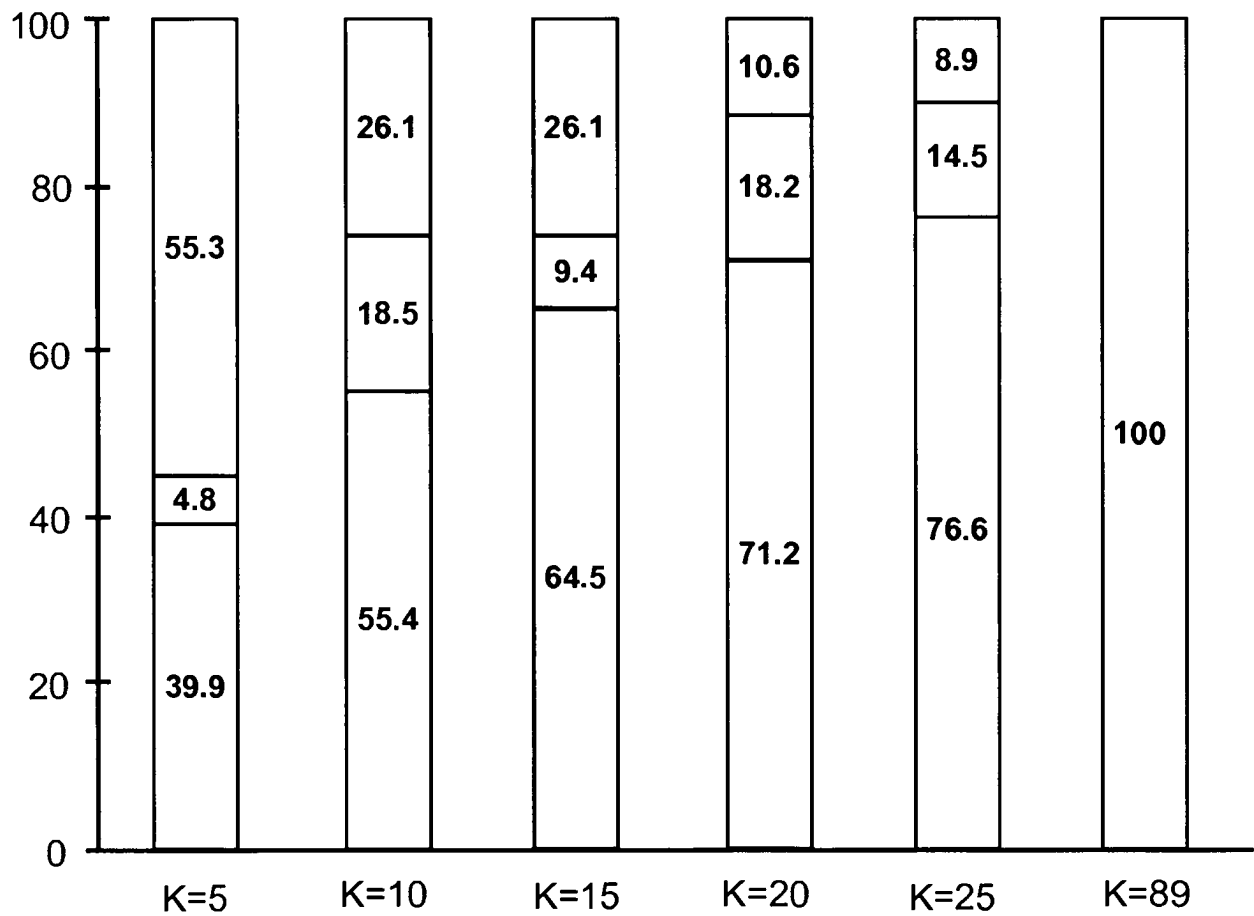


Fig. 8