(19) **Bundesrepublik Deutschland**
Deutsches Patent- und Markenamt

(10) **DE 698 37 938 T2** 2008.02.14

(12) # Übersetzung der europäischen Patentschrift

(97) **EP 1 048 145 B1**
(21) Deutsches Aktenzeichen: **698 37 938.1**
(86) PCT-Aktenzeichen: **PCT/US98/25688**
(96) Europäisches Aktenzeichen: **98 962 888.8**
(87) PCT-Veröffentlichungs-Nr.: **WO 1999/033227**
(86) PCT-Anmeldetag: **04.12.1998**
(87) Veröffentlichungstag der PCT-Anmeldung: **01.07.1999**
(97) Erstveröffentlichung durch das EPA: **02.11.2000**
(97) Veröffentlichungstag der Patenterteilung beim EPA: **13.06.2007**
(47) Veröffentlichungstag im Patentblatt: **14.02.2008**

(51) Int Cl.[8]: ***H04L 12/28*** (2006.01)
H04L 12/56 (2006.01)
H04L 29/06 (2006.01)
H04L 29/12 (2006.01)

(30) Unionspriorität:
994709 **19.12.1997** **US**

(84) Benannte Vertragsstaaten:
DE, FR, GB

(73) Patentinhaber:
Avaya Technology Corp., Basking Ridge, N.J., US

(72) Erfinder:
BHASKARAN, Sajit, Sunnyvale, CA 94087, US

(74) Vertreter:
Blumbach Zinngrebe, 65187 Wiesbaden

(54) Bezeichnung: **ÜBERGREIFENDE BILDUNG VON SERVER CLUSTERN MITTELS EINER NETZWERKFLUSSVERMITTLUNG**

Beschreibung

HINTERGRUND DER ERFINDUNG

Gebiet der Erfindung

[0001] Die vorliegende Erfindung betrifft allgemein Computernetze und spezieller Netzvermittlungseinrichtungen für hohe Bandbreite.

Beschreibung des Standes der Technik

[0002] Der zunehmende Verkehr über Computernetze wie beispielsweise das Internet sowie Intranets von Unternehmen, WANs und LANs erfordert oft die Verwendung mehrerer Server, um den Bedürfnissen eines einzelnen Diensteanbieters oder einer MIS-Abteilung zu entsprechen. Beispielsweise ist es möglich, dass ein Unternehmen, das eine Suchmaschine für das Internet bereitstellt, täglich 80 Millionen Treffer (d. h. Zugriffe auf die Web-Page des Unternehmens) behandelt. Ein einzelner Server kann ein solch großes Volumen an Diensteanforderungen nicht innerhalb einer akzeptablen Ansprechzeit abwickeln. Daher ist es für Diensteanbieter mit hohem Verkehrsvolumen wünschenswert, mehrere Server nutzen zu können, um Diensteanforderungen zu befriedigen.

[0003] Beispielsweise wird beim Internetprotokoll (IP), welches genutzt wird, um Computer zu identifizieren, die mit dem Internet und anderen globalen, Weiterverkehrs- oder lokalen Netzen verbunden sind, jedem Computer, der mit dem Netz verbunden ist, eine eindeutige IP-Adresse zugewiesen. Wenn also mehrere Server genutzt werden, muss auf jeden Server unter Verwendung der eigenen IP-Adresse des Servers zugegriffen werden.

[0004] Andererseits ist es wünschenswert, Nutzer in die Lage zu versetzen, auf alle Server eines Diensteanbieters unter Verwendung einer einzigen IP-Adresse zuzugreifen. Ansonsten müssten die Nutzer die Server verfolgen, die von dem Diensteanbieter unterhalten werden, sowie deren relative Arbeitslasten, um schnellere Ansprechzeiten zu erzielen. Durch Nutzung einer einzigen "virtuellen" IP-Adresse (d. h. einer IP-Adresse, die keinem der IP-Server entspricht, sondern vielmehr die gesamte Gruppe von IP-Servern bezeichnet), sind Diensteanbieter in der Lage, die Diensteanforderungen auf die Server aufzuteilen. Durch Nutzung dieses Schemas können IP-Server sogar zu der Gruppe von IP-Servern, die der virtuellen IP-Adresse entspricht, hinzugefügt werden oder aus dieser entfernt werden, um sich ändernde Verkehrsvolumina zu kompensieren. Mehrere Server, die auf diese Weise genutzt werden, werden bisweilen als ein "Cluster" bezeichnet.

[0005] Ein beispielhaftes System, das einen Hybriden der beiden vorstehend beschriebenen Schemata darstellt, ist in der EP 0 605 339 A2 offenbart. Während jeder Server (Knoten) in einem Cluster seine eigene eindeutige IP-Adresse besitzt, werden alle eingehenden Nachrichten an die IP-Adresse eines der Server adressiert, der als "Gateway" bezeichnet wird. Das Gateway verteilt die Nachrichten auf die Server des Clusters, indem es die IP-Adresse des Gateways in der Nachricht durch die eindeutige IP-Adresse des Zielservers ersetzt oder mit dieser ergänzt.

[0006] Fig. 1 stellt einen Cluster aus IP-Servern gemäß dem Stand der Technik dar. Ein Server-Lastausgleicher **100** leitet Pakete zwischen IP-Servern **110**, **120**, **130**, **140** sowie **150** und Netzroutern **160**, **170** sowie **180** weiter. Jeder der IP-Server **110**, **120**, **130**, **140** und **150** sowie Netzrouter **160**, **170** und **180** weist eine unterschiedliche IP-Adresse auf, jedoch kann von Netzwerken aus, die mit den Netzroutern **160**, **170** und **180** verbunden sind, auf jeden der IP-Server **110**, **120**, **130**, **140** und **150** über eine (nicht gezeigte) virtuelle IP-Adresse zugegriffen werden. Wenn ein Paket, das an die virtuelle IP-Adresse adressiert ist, von dem Server-Lastausgleicher **100** empfangen wird, wird die virtuelle IP-Adresse in die einzelnen IP-Adressen eines der IP-Server übersetzt und das Paket wird zu diesem IP-Server weitergeleitet. Mit der Übersetzung ist jedoch das Erzeugen einer neuen Prüfsumme für das Paket sowie das Umschreiben der Felder für die Quell/Ziel-IP-Adresse und die Prüfsumme, des IP-Header-Felds wie auch der TCP- und UDP-Header-Felder verbunden. Sowohl die IP-Header-Prüfsumme, welche die ISO-Schicht-3- oder Netzschicht-Header-Prüfsumme darstellt, als auch die TCP- oder UDP-Header-Prüfsummen, welche die ISO-Schicht-4- oder Transportschicht-Header-Prüfsummen darstellen, müssen für jedes Paket neu berechnet werden. Typischerweise erfordern diese Vorgänge die Intervention eines Prozessors des Server-Lastausgleichers.

[0007] Wenn ein großes Volumen an Anforderungen verarbeitet wird, hat der Overhead, der durch die Übersetzung entsteht, einen beträchtlichen Einfluss auf die Ansprechzeit der IP-Server. Wenn außerdem eine große Anzahl von IP-Servern genutzt wird, entsteht durch die Zeit, die zum Ausführen der Übersetzung erforderlich ist, ein Engpass im Leistungsverhalten des Server-Lastausgleichers, da die IP-Adresse für jedes Paket, das zu den IP-Servern oder von diesen weg übertragen wird, von der Vermittlungseinrichtung übersetzt werden muss. Daher besteht ein Bedarf an einem schnelleren Verfahren zum gemeinsamen Nutzen einer einzigen IP-Adresse durch mehrere IP-Server.

[0008] Es gibt andere Fälle, bei denen mehrere IP-Adressen genutzt werden und typischerweise ein Client versucht, auf einen primären IP-Server zuzugreifen. Wenn der primäre IP-Server nicht innerhalb

einer festgelegten Zeitspanne antwortet, versucht der Client, auf Reserve-IP-Server zuzugreifen, bis eine Antwort empfangen wird. Wenn also der primäre IP-Server nicht verfügbar ist, erlebt der Client eine schlechte Reaktionszeit. Derzeitige Serverreplikationssysteme wie solche, die bei DNS- und RADIUS-Servern genutzt werden, sind von diesem Problem betroffen. Es besteht also ein Bedarf an einem Verfahren zum Zugreifen auf mehrere IP-Server, bei dem keine schlechten Ansprechzeiten auftreten, wenn der primäre IP-Server nicht verfügbar ist.

[0009] Ein weiterer potenzieller Nachteil des Stands der Technik besteht darin, dass jeder replizierte Server eine physisch in dem Server konfigurierte eindeutige IP-Adresse erfordert. Da sämtliche IP-Netze Teilnetz-Maskierungsregeln unterliegen (welche oft durch einen externen Administrator bestimmt werden), ist die Skalierbarkeit der Replikation stark eingeschränkt. Wenn beispielsweise das Teilnetz-Präfix 28 Bits einer IP-Adresse mit 32 Bits ausmacht, beträgt die maximale Anzahl von replizierten Servern 16 ($2^{(32-28)}$). Es besteht ein Bedarf an einem Verfahren zum Replizieren von Servern, das eine Replikation von IP-Servern unabhängig von Teilnetz-Maskierungsregeln ermöglicht.

[0010] Adressen für IP Version 4 sind momentan im Internet knapp, somit ist jedes Verfahren der IP-Server-Replikation, das einen proportionalen Verbrauch an diesen knappen IP Adressen erfordert, von Natur aus verschwenderisch. Beispielsweise stellt ein Lastausgleich auf Basis des Domain Name Service (DNS) ein Beispiel für den Stand der Technik dar. DNS-Server werden zum Auflösen eines Servernamens (z. B. www.companyname.com) in eine global eindeutige IP-Adresse (z. B. 192.45.54.23) genutzt. Bei einem DNS-basierten Server-Lastausgleich werden viele eindeutige IP-Adressen pro Servername vorgehalten und sparsam ausgeteilt, um einen Lastausgleich zu ermöglichen. Dadurch reduziert sich jedoch die Anzahl der verfügbaren Adressen für IP Version 4. Es besteht daher ein Bedarf an einem Verfahren zur Clusterbildung mit IP-Servern, bei dem der Verbrauch des knappen IP-Adress-Raums minimiert wird.

[0011] Darüber hinaus kann, wenn die IP-Nutzlast eines Pakets verschlüsselt wird, um sichere Übertragungen über das Internet bereitzustellen, die Übersetzung der IP-Adresse nicht ausgeführt werden, ohne dass zunächst die IP-Nutzlast entschlüsselt wird, (welche die TCP- oder UDP-Header-Prüfsummen enthält). Bei den momentanen Rahmenbedingungen für IP-Sicherheit, die als IPSEC bezeichnet wird, stellt die Transportschicht einen Teil der Netzschicht-Nutzlast dar, die in einer Netzanwendung, welche IPSEC implementiert, vollständig verschlüsselt wird. IPSEC ist in RFCs 1825–1827, veröffentlicht von der Internet Engineering Taskforce, beschrieben. Die Verschlüsselung wird von dem Client ausgeführt, und die Entschlüsselung wird von dem Server ausgeführt, und zwar mit Hilfe geheimer Cryptoschlüssel, welche für jede Client-Server-Verbindung eindeutig sind. Wenn daher eine solche Verschlüsselung bei Client-Server-Kommunikationen ausgeführt wird, wie bei IPSEC, werden die Server-Lastausgleicher gemäß dem Stand der Technik nicht in der Lage sein, Lastausgleichsvorgänge auszuführen, ohne die IPSEC-Regeln zu verletzen. Dies ist der Fall, weil die Server-Lastausgleicher nicht auf die Transportschicht-Informationen zugreifen können (die als Teil der IP-Nutzlast verschlüsselt sind), ohne zunächst die IP-Nutzlast zu entschlüsseln. Da die zwischen dem Client und dem Server eingerichteten Cryptoschlüssel per Definition nicht öffentlich sind, kann die IP-Nutzlast von dem Server-Lastausgleicher nicht entsprechend IPSEC entschlüsselt werden (tatsächlich wird der Server-Lastausgleicher für sämtliche praktische Zwecke für verschlüsselte Pakete überhaupt nicht funktionieren).

[0012] Es besteht daher ein Bedarf an einem System, das nicht nur die Übertragungen von verschlüsselten Datenpaketen entsprechend dem IPSEC-Modell ermöglicht, sondern das auch Netzwerkadministratoren ermöglicht, sowohl einen Server-Lastausgleich als auch IPSEC in ihren Netzen auszuführen.

[0013] Darüber hinaus funktionieren derzeitige Server-Lastausgleicher typischerweise nur bei TCP-Paketen. Im Gegensatz dazu weisen IP-Header ein Protokollfeld von 8 Bit auf, das theoretisch bis zu 256 Transportprotokolle auf der Schicht 4 gemäß ISO unterstützt. Es besteht daher ein Bedarf an einem Server-Lastausgleichssystem, welches andere Transportprotokolle auf der Schicht 4 gemäß ISO als TCP unterstützt (z. B. UDP, IP_in_IP, usw.).

[0014] Systeme gemäß dem Stand der Technik ermöglichen einen Lastausgleich und bisweilen eine Fehlertoleranz für Netzverkehr nur in der Eingangsrichtung (d. h. Client-Router-Server). Ein Lastausgleich und eine Fehlertoleranz in der umgekehrten (abgehenden) Richtung (d. h. Server-Router-Client) wird nicht unterstützt. Speziell wird, wenn mehrere Router-Verbindungen für den Server bereitstehen, um Informationen an Clients zurückzusenden, kein Versuch unternommen, die Verkehrsflusslast über die Router-Verbindungen auszugleichen. Außerdem erfolgt, wenn ein spezieller IP-Server dafür konfiguriert ist, eine spezielle Standard-Router-IP-Adresse bei den abgehenden Übertragungen zu nutzen, keine Fehlertoleranz oder transparente Umleitung von Paketen, wenn der Router ausfällt. Es besteht daher ein Bedarf an einem System, das Clusterbildungsdienste für Verkehrsfluss in sowohl der Eingangs- als auch der Ausgangsrichtung ermöglicht.

[0015] Ein beispielhaftes System dieser Art ist be-

schrieben in "Load Balancing for Multiple Interfaces for Transmission Control Protocol/Internet Protocol for VM/MVS", IBM Technical Disclosure Bulletin (IBM Corp. New York, USA), Bd. 38, Nr. 9 (1. September 1995), S. 7–9, XP000540166, ISSN 0018–8689. Es offenbart das Zuordnen von eindeutigen MAC-Adressen und der gleichen IP-Adresse für mehrere Schnittstellen. Abgesetzte Hosts nutzen immer die gleiche IP-Adresse und kümmern sich nicht darum, welche Schnittstelle sie nutzen, wenn aber ein abgesetzter Host die IP-Adresse des Hosts über ARP anfragt, erhält er die MAC-Adresse einer der verfügbaren Schnittstellen. Um das Problem zu beheben, dass eine Schnittstelle ausfällt, müssen ARP-Zwischenspeicher in den abgesetzten Hosts, welche die ausgefallene Schnittstelle genutzt haben, aktualisiert werden.

[0016] Damit abgehend ein Lastausgleich stattfindet, muss die Möglichkeit gegeben sein, dem TCP/IP-Server zu sagen, dass es für einen gegebenen Bestimmungsort mehrere Schnittstellen gibt, und der TCP/IP-Server muss modifiziert werden, um den Verkehr zwischen allen verfügbaren Schnittstellen gleichmäßig auszugleichen. In dem Dokument sind zwei Alternativen zum Ausgleich des Verkehrs offenbart. Die eine besteht darin, mehrere Routen zu dem gleichen Bestimmungsort in der Wegelenkungstabelle des Gateways zuzulassen. Der TCP/IP muss eine davon herausgreifen, worauf für den abgesetzten Host über diese Schnittstelle ein ARP-Vorgang erfolgen muss, um die MAC-Adresse dieser Schnittstelle zu erhalten. Außerdem verfolgt ein neues Feld in dem ARP-Zwischenspeicher, welche Schnittstelle der TCP/IP-Server nutzt, um den abgesetzten Host zu kontaktieren. Auf diese Weise werden alle Daten, die zu diesem speziellen abgesetzten Host gesendet werden, immer von der gleichen MAC-Adresse ausgehen. Die andere Alternative komplettiert die erste, indem das Kombinieren sämtlicher Schnittstellen für ein gemeinsames Ziel zu einer Schnittstellengruppe ermöglicht wird. Der Vorteil besteht darin, dass nur eine einzige Route zu der Wegelenkungstabelle pro Ziel hinzugefügt zu werden braucht, anstatt eine Route pro Schnittstelle. Im Hinblick auf einen automatischen Lastausgleich überwacht der TPC/IP-Server die Verkehrslast, und wenn diese nicht gleich ist, kann der Server Verkehr von einem Adapter zu einem anderen verschieben, indem er seinen eigenen ARP-Zwischenspeicher aktualisiert und den abgesetzten Host über ARP anfragt, damit dieser seinen ARP-Zwischenspeicher aktualisiert.

[0017] Die Lösungen gemäß dem Stand der Technik stellen Hardware-Einrichtungen dar, die derart konfiguriert sind, dass sie für den Cluster von Servern, für den ein Lastausgleich erfolgt, als IP-Router erscheinen. Infolgedessen kommt zu der Domain des Administrators der Router für verwaltete IP-Router eine weitere Klasse von IP-Router-Einrichtungen hinzu. Dies stellt eine Einschränkung bezüglich einer zukünftigen Entwicklung des Routernetzes sowohl hinsichtlich des Hinzufügens von Routern neuer Anbieter in der Zukunft als auch in Bezug auf das Hinzufügen von neuen und weiterentwickelten Wegelenkungsmerkmalen dar. Das Suchen und Beheben von Wegelenkungsproblemen wird ebenfalls schwieriger. Es wäre somit vorzuziehen, ein vollständig transparentes Hardwareelement wie beispielsweise einen LAN-Switch oder -Hub als lastausgleichende Einrichtung anzuwenden. Gemäß dem Stand der Technik sind Server und etwaige externe Router mit der Lastausgleichseinrichtung unter Nutzung eines Shared Media Ethernets verbunden (d. h. einem Rundsendemedien-Netzwerk). Es besteht ein Bedarf an einer besseren Lösung, welche es ermöglicht, vermittelte Leitungen zu nutzen (z. B. Switched (vermitteltes) Ethernet, SONST), da vermittelte Leitungen von Natur aus (a) eine bestimmte Bandbreite bereitstellen und (b) Vollduplex bereitstellen (d. h. gleichzeitige Sende- und Empfangsvorgänge) für Geräte mit hergestellter Rufverbindung.

Zusammenfassung der Erfindung

[0018] Entsprechend der vorliegenden Erfindung wird eine Netzfluss-Vermittlungseinrichtung entsprechend Anspruch 1 zur Verfügung gestellt.

[0019] Entsprechend der vorliegenden Erfindung wird ein Verfahren entsprechend Anspruch 4 zur Verfügung gestellt.

[0020] Die vorliegende Erfindung stellt eine Netzfluss-Vermittlungseinrichtung (sowie ein Verfahren zum Betrieb selbiger) zum Verbinden eines Pools von IP-Routern mit einem Cluster von IP-Servern, die eine einzige IP-Adresse gemeinsam nutzen, zur Verfügung, ohne dass eine Übersetzung der IP-Adresse erforderlich ist und wobei eine bidirektionale Clusterbildung ermöglicht wird. Die Netzfluss-Vermittlungseinrichtung ermöglicht, indem sie transparent auf den ISO-Schichten 2 und 3 arbeitet, eine Cross-Plattform-Clusterbildung von Servern und Routern, wobei diese Router die so genannten "First-Hop-" Router darstellen, die von den Servern genutzt werden, um mit der Außenwelt zu kommunizieren. Das bedeutet, dass die Server in einem einzelnen Cluster von einem beliebigen Hersteller von Computer-Hardware stammen können und auf diesen ein beliebiges Betriebssystem laufen kann (z. B. WINDOWS NT von Microsoft, Unix, MACOS). WINDOWS NT stellt eine eingetragene Marke der Microsoft Corp. Redmond, Washington dar; MACOS ist eine eingetragene Marke der Apple Computer Inc., Cupertino, Kalifornien. Das bedeutet auch, die Router können von einem beliebigen Anbieter von Wegelenkungsausrüstung kommen. Die Netzfluss-Vermittlungseinrichtung ermöglicht daher den Kunden eine freie Auswahl bezüglich der Betriebssysteme der Server als auch der

Routersysteme bei der Gestaltung ihrer Server-Clusterbildungsstrukturen. Die einzigen Anforderungen für diese Server und Router bestehen darin, dass sie alle die standardmäßigen TCP/IP-Kommunikationsprotokolle oder einen gewissen anderen Protokollstapel entsprechend dem 7-Schichten-Modell der ISO/OSI für Computerkommunikation implementieren. Die Netzfluss-Vermittlungseinrichtung leitet Pakete zu einzelnen Servern weiter, indem sie die Sicherungsschicht(Data Link Layer)-Adresse des Ziel-IP-Servers in das Ziel-Adressfeld der Sicherungsschicht des Pakets schreibt. Pakete, die von den IP-Servern zu den IP-Routern übertragen werden, erfordern andererseits keine Modifikation des Sicherungsschicht-Adressfeldes.

[0021] Da in einer typischen Client-Server-Umgebung die Mehrzahl der Pakete, die durch die Netzfluss-Steuerungsvermittlungseinrichtung fließen, von dem Server zum Client übertragen wird, werden durch das Überflüssigmachen einer Prozessorintervention bei der Weiterleitung abgehender Pakete beträchtliche Verbesserungen im Leistungsverhalten möglich. Infolgedessen verringert sich die Wahrscheinlichkeit, dass die Netzfluss-Vermittlungseinrichtung zu einem Engpass wird, deutlich.

[0022] In einer einzigen Netzfluss-Vermittlungseinrichtung werden mehrere Cluster (ein oder mehrere IP-Server, die eine einzige IP-Adresse gemeinsam nutzen) unterstützt. Auf jeder einzelnen Verbindung, die an jeden der IP-Server angekoppelt ist, können mehrere Cluster unterstützt werden, wenn das Betriebssystem des IP-Servers mehrere IP-Adressen auf einer physischen Verbindung unterstützt.

[0023] Bei einigen Ausführungsformen führt die Netzfluss-Vermittlungseinrichtung zusätzlich dazu, dass sie die Pakete weiterleitet, einen Lastausgleich sowie eine Fehlertoleranzfunktion aus. Bei diesen Ausführungsformen führt ein Prozessor der Netzfluss-Vermittlungseinrichtung periodisch eine Lastausgleichsroutine aus, um die relative Arbeitslast für jeden der IP-Server festzustellen. Wenn die Netzfluss-Vermittlungseinrichtung ein Paket empfängt, das für den Cluster von IP-Servern bestimmt ist, wird das Paket zu demjenigen IP-Server mit einer optimalen Arbeitslast weitergeleitet, sodass sichergestellt wird, dass die Arbeitslast gleichmäßig auf die IP-Server verteilt wird. Außerdem wird, wenn ein Ausfall eines Netzrouters erkannt wird, ein Paket, das an diesen Netzrouter adressiert ist, zu einem anderen Netzrouter umgeleitet, indem die Sicherungsschicht-Zieladresse des Pakets umgeschrieben wird. Da die Netzfluss-Vermittlungseinrichtung kontinuierlich den Status der IP-Server überwacht, wird keine längere Zeitverzögerung in die Client-Server-Kommunikationen eingetragen, wenn ein IP-Server funktionsunfähig wird.

[0024] Da der IP-Header nicht modifiziert wird, arbeitet die Netzfluss-Vermittlungseinrichtung gemäß der vorliegenden Erfindung an Paketen, die gemäß einem beliebigen ISO-Schicht-4-Protokoll kodiert sind, und ist im Gegensatz zu Server-Lastausgleichern entsprechend dem Stand der Technik nicht auf TCP-kodierte Pakete beschränkt. Außerdem kann die Netzfluss-Vermittlungseinrichtung auch eine Umleitung, einen Lastausgleich und eine Fehlertoleranz für verschlüsselte Pakete transparent für sowohl den Server als auch den Client abwickeln.

[0025] Bei einigen Ausführungsformen wird der Lastausgleich auch für abgehende Pakete ausgeführt, sodass Pakete zu dem Router mit einer optimalen Arbeitslast geleitet werden.

[0026] Es werden somit ein Verfahren und eine Vorrichtung zum Ermöglichen einer bidirektionalen Clusterbildung im Hinblick aus Lastausgleich und Fehlertoleranz in der Eingangsrichtung (d. h. Client-Router-Server) als auch in der Abgangsrichtung (d. h. Server-Router-Client) bereitgestellt.

Kurze Beschreibung der Zeichnungen

[0027] Fig. 1 stellt einen Cluster von IP-Servern, die jeweils eine unterschiedliche IP-Adresse aufweisen, gemäß dem Stand der Technik dar, sowie eine Netzfluss-Vermittlungseinrichtung gemäß dem Stand der Technik zum Übersetzen einer virtuellen IP-Adresse, die von sämtlichen IP-Servern in dem Cluster gemeinsam genutzt wird, in die einzelnen IP-Adressen der IP-Server.

[0028] Fig. 2 stellt ein Cluster von IP-Servern sowie eine Netzfluss-Vermittlungseinrichtung entsprechend einer Ausführungsform der vorliegenden Erfindung dar. Jeder IP-Server weist die gleiche IP-Adresse auf. Eine Sicherungsschicht-Adresse wird genutzt, um jeden IP-Server in dem Cluster zu identifizieren.

[0029] Fig. 3A stellt das Format für ein Paket dar, das durch die Netzfluss-Vermittlungseinrichtung **205** aus Fig. 2 zu/von dem Cluster von IP-Servern geleitet wird.

[0030] Fig. 3B zeigt das Format des Verbindungsfelds **320** aus Fig. 3A.

[0031] Fig. 4A stellt die Struktur der Netzfluss-Vermittlungseinrichtung **205** aus Fig. 2 dar.

[0032] Fig. 4B stellt ein Flussdiagramm für den Prozess des Weiterleitens von Paketen von einem der Netz-Clients zu einem der IP-Server aus Fig. 2 über die Netzfluss-Vermittlungseinrichtung **205** aus Fig. 4A entsprechend einer Ausführungsform der vorliegenden Erfindung dar.

[0033] Fig. 4C stellt ein Flussdiagramm für den Prozess des Weiterleitens von Paketen von einem der IP-Server zu einem der Netz-Clients aus Fig. 2 über die Netzfluss-Vermittlungseinrichtung **205** aus Fig. 4A entsprechend einer Ausführungsform der Erfindung dar.

[0034] Fig. 5A stellt ein Blockdiagramm einer Netzfluss-Vermittlungseinrichtung, die mit Hilfe mehrerer Allzweck-Schaltungskarten implementiert ist, entsprechend einer Ausführungsform der Erfindung dar.

[0035] Fig. 5B stellt ein Blockdiagramm einer Netzfluss-Vermittlungseinrichtung, die mit Hilfe einer Allzweck-CPU-Karte und einer Spezial-Netzkarte implementiert ist, entsprechend einer Ausführungsform der Erfindung dar.

[0036] Fig. 5C stellt ein Blockdiagramm einer Netzfluss-Vermittlungseinrichtung, die mit Hilfe zweier Spezial-Schaltungskarten implementiert ist, entsprechend einer Ausführungsform der Erfindung dar.

[0037] Fig. 5D stellt ein Blockdiagramm einer Netzfluss-Vermittlungseinrichtung, die mit Hilfe einer einzigen Spezial-Schaltungskarte implementiert ist, entsprechend einer Ausführungsform der Erfindung dar.

[0038] Fig. 5E stellt ein Blockdiagramm einer Netzfluss-Vermittlungseinrichtung, die mit Hilfe einer Kombination aus Spezial- und Allzweck-Schaltungskarten implementiert ist, entsprechend einer Ausführungsform der vorliegenden Erfindung dar.

[0039] Fig. 5F stellt ein Blockdiagramm einer Netzfluss-Vermittlungseinrichtung, die mit Hilfe eines Crossbar-Switch implementiert ist, entsprechend einer Ausführungsform der Erfindung dar.

Detaillierte Beschreibung der Erfindung

[0040] Das Verfahren und die Vorrichtung gemäß der vorliegenden Erfindung ermöglichen, dass mehrere IP-Server eine gleiche IP-Adresse gemeinsam nutzen, und es wird eine Netzfluss-Vermittlungseinrichtung verwendet, um Pakete zwischen den IP-Servern basierend auf der Sicherungsschicht-Adresse der IP-Server weiterzuleiten (z. B. wird die Zieladresse der Pakete in die Sicherungsschicht-Adresse eines der IP-Server übersetzt). Da IP-Netze das Quell-Adressfeld der Sicherungsschicht der Pakete, die über das Netz übertragen werden, nicht kennen, wird eine Übersetzung der Sicherungsschicht-Adresse nur für Pakete ausgeführt, die von einem IP-Client zu einem IP-Server fließen. In der Rückflussrichtung, das heißt von einem IP-Server zu einem IP-Client, ist keine Übersetzung der Sicherungsschicht-Adresse erforderlich, somit wird ein sehr schneller Durchsatz durch die Netzfluss-Vermittlungseinrichtung ermöglicht.

[0041] Ein Cluster von IP-Servern **200** und eine Netzfluss-Vermittlungseinrichtung **205** entsprechend einer Ausführungsform der Erfindung sind in Fig. 2 gezeigt. Die Netzfluss-Vermittlungseinrichtung **205** leitet Pakete zwischen IP-Servern **210**, **220**, **230**, **240** sowie **250** und Netzroutern **260**, **270** und **280** weiter. Die IP-Server **210**, **220**, **230**, **240** und **250** sind identisch konfiguriert und besitzen eine virtuelle IP-Adresse **290**. Außerdem besitzt jeder der IP-Server **210**, **220**, **230**, **240** und **250** eine unterschiedliche Sicherungsschicht-Adresse und einen unterschiedlichen Verbindungsnamen. Der Verbindungsname wird genutzt, um den einzelnen Server in dem Cluster von Servern, welche die gleiche IP-Adresse gemeinsam nutzen, zu identifizieren. Wie noch später erklärt wird, wird die Sicherungsschicht-Adresse genutzt, um eine virtuelle Sicherungsschicht-Adresse in eine physische Sicherungsschicht-Adresse zu übersetzen, nachdem ein IP-Server von der Netzfluss-Vermittlungseinrichtung **205** zum Empfang des Pakets ausgewählt worden ist. Die IP-Adresse **290** ist für Einrichtungen, die mit dem Cluster **200** kommunizieren, sichtbar, während die einzelnen Sicherungsschicht-Adressen jedes IP-Servers nicht sichtbar sind. Die Netzfluss-Vermittlungseinrichtung **205** führt eigentlich eine stellvertretende Adress Resolution Protocol (ARP – Adressauflösungsprotokoll)-Funktion aus, durch welche eine "virtuelle" (nicht gezeigte) Sicherungsschicht-Adresse an eine mit dem Netz verbundene Einrichtung in Reaktion auf eine standardmäßige ARP-Anfrage zurückgesendet wird. Infolgedessen sehen an das Netz angeschlossene Einrichtungen den Cluster **200**, als hätte dieser eine einzige IP-Adresse **290** und eine einzige Sicherungsschicht-Adresse (nicht gezeigt).

[0042] Die Netzrouter **260**, **270** und **280** andererseits haben jeweils eine unterschiedliche IP-Adresse und eine unterschiedliche Sicherungsschicht-Adresse. Die Router werden genutzt, um den Cluster **200** über die Netzfluss-Vermittlungseinrichtung **205** mit (nicht gezeigten) externen Netzen zu verbinden. Somit wird eine Einrichtung, die mit einem der externen Netze verbunden ist (z. B. ein Router), um Informationspakete an den Cluster **200** zu übertragen, eine standardmäßige ARP-Anfrage an die Netzfluss-Vermittlungseinrichtung **205** ausgeben, um die virtuelle Sicherungsschicht-Adresse des Clusters **200** zu erhalten; die Netzfluss-Vermittlungseinrichtung **205** sendet eine Sicherungsschicht-Adresse der ausgewählten Empfangseinrichtung (z. B. eines der IP-Server) an die anfordernde Einrichtung (z. B. den Router) zurück. Die mit dem Netz verbundene Einrichtung sendet dann eine Reihe von Paketen an die Netzfluss-Vermittlungseinrichtung **205** (z. B. über einen der Netzrouter **260**, **270** oder **280**, die mit dem externen Netz verbunden sind). Die Pakete werden dann von der Netzfluss-Vermittlungseinrichtung **205** zu exakt einem der IP-Server **210**, **220**, **230**, **240** und **250** umgeleitet.

[0043] Da sämtliche Ausführungsformen der Netzfluss-Vermittlungseinrichtung sicherstellen, dass sich nicht zwei Server in demselben Cluster in dem demselben Flussvermittlungsteil befinden, wird eine Rundsendeisolation der replizierten Server möglich. Daher werden IP-Adresskonflikte vermieden, durch die aktive Intervention der Flussvermittlungseinrichtung für den Fall, dass wie vorstehend beschrieben ARP-Anforderungspakete von der Netzfluss-Vermittlungseinrichtung empfangen werden.

[0044] Das Format eines Pakets **300**, das über das externe Netz übertragen wird, ist in Fig. 3A dargestellt. Das Paket **300** weist ein Header-Feld **310**, ein Verbindungsfeld **320**, einen IP-Header **330**, einen TCP-Header **340**, eine Daten-Nutzlast **350**, ein CRC-Feld **360** und ein auch als Trailer bezeichnetes Endfeld **370** auf. Der Header **310** und der Trailer **370** stellen 8-Bit breite Felder mit "Private" Tag (Kennzeichnung "privat") dar: diese werden nicht über das externe Netz übertragen, sondern werden nur innerhalb der Netzfluss-Vermittlungseinrichtung genutzt. Der IP-Header **330** und der TCP-Header **340** stellen standardmäßige IP- und TCP-Header dar. Der IP-Header **330** umfasst neben anderen Informationen eine IP-Zieladresse und eine IP-Quelladresse für das Paket **300**. Das CRC-Feld **360** enthält einen Prüfsummen-Korrekturcode, der genutzt wird, um zu verifizieren, dass das Paket **300** ohne Fehler übertragen worden ist. Würde der IP-Header **330** modifiziert werden, wie es für Verfahren gemäß dem Stand der Technik zum gemeinsamen Nutzen einer einzigen IP-Adresse zwischen mehreren IP-Servern erforderlich ist, müsste die Prüfsumme für das CRC-Feld **360** neu berechnet werden, ein Vorgang, der die Intervention eines Prozessors erforderlich macht. Außerdem ist, wenn verschlüsselte Daten entsprechend den IPSEC-Sicherheitsrichtlinien übertragen werden, eine Entschlüsselung der IP-Nutzlast erforderlich. Es wird also dadurch, dass die Notwendigkeit der Neuberechnung der Prüfsumme für jedes Paket wegfällt, mit der Netzfluss-Vermittlungseinrichtung gemäß der vorliegenden Erfindung ein besserer Durchsatz als bei Einrichtungen gemäß dem Stand der Technik erreicht. Die Netzbesitzer können ferner IPSEC-Sicherheitsmechanismen transparent und ohne Befürchtung, dass die Kommunikation unterbrochen wird, anwenden.

[0045] Fig. 3B stellt das Format des Verbindungsfelds **320** dar. Das Verbindungsfeld **320** weist ein Sicherungsschicht-Quelladressfeld **380**, ein Sicherungsschicht-Zieladressfeld **390** und ein Typenfeld **395** auf. Da das Verbindungsfeld **320** nicht Teil des IP-Protokolls ist, besteht keine Notwendigkeit, die Prüfsumme für das CRC-Feld **360** neu zu berechnen, wenn das Verbindungsfeld **320** modifiziert wird. Dementsprechend wird eine Umleitung von Paketen entsprechend der vorliegenden Erfindung dadurch erreicht, dass die Sicherungsschicht-Zieladresse in dem Sicherungsschicht-Zieladressfeld **390** des Pakets **300** umgeschrieben wird. Weder der IP-Header **330** noch das CRC-Feld **360** werden modifiziert, wodurch sich die Verarbeitungszeit verringert, die erforderlich ist, um Pakete zu dem Cluster von IP-Servern und von diesem weg weiterzuleiten.

[0046] Eine Ausführungsform der Netzfluss-Vermittlungseinrichtung **205** (Fig. 2) wird durch das Blockdiagramm aus Fig. 4A veranschaulicht. Die Netzfluss-Vermittlungseinrichtung **205** weist eine CPU-Karte **400** sowie vier Ethernetkarten **415**, **416**, **417** und **418**, die durch einen PCI-Bus **410** verbunden sind, auf. Die CPU-Karte **400** wiederum umfasst eine CPU **402**, einen Speicher **404** und einen Speicher-Controller **406** zum Kontrollieren des Zugriffs auf den Speicher **404**. Jede der Ethernet-Karten **415**, **416**, **417** und **418** weist einen Ethernet-Controller und zwei Eingangs-/Ausgangsports **411** und **413** auf.

[0047] Eine Netzfluss-Vermittlungseinrichtung entsprechend einer Ausführungsform der Erfindung kann vollständig aus serienmäßig produzierten ASICs (anwendungsspezifischen integrierten Schaltungen) aufgebaut werden, die von einer Allzweck-CPU gesteuert werden, welche ein Softwareprogramm ausführt. Da viele kommerziell verfügbare Ethernet-Switche Allzweck-CPUs zum Vermittlungsmanagement bereitstellen (z. B. zum Ausführen SNMP- und IEEE 802.1D Spanning-Tree-Protokollen) kann eine Netzvermittlungseinrichtung entsprechend einer Ausführungsform der Erfindung in einfacher Weise auf solchen Hardwareplattformen implementiert werden. Die einzige Anforderung besteht darin, dass der ASIC in der Lage ist, eine bestimmte Art von "CPU-Intervention" zu unterstützen, die ausgelöst wird, wenn ein Paket mit einer bestimmten Sicherungsschicht-Zieladresse durch die Netzfluss-Vermittlungseinrichtung geleitet wird. ASICs, die diese Form von CPU-Intervention unterstützen, sind unter anderem erhältlich bei der Galileo Technology Ltd., Kormiel, Israel, der MMC Networks, Inc., Sunnyvale, Kalifornien und der I-Cube, Inc., Campbell, Kalifornien.

[0048] Der Prozess des Weiterleitens eines Pakets **300** (Fig. 3A), das von einem der Netzrouter **260**, **270** oder **280** empfangen wird, an einen der IP-Server **210**, **220**, **230**, **240** oder **250** aus Fig. 2, wird durch das Ablaufdiagramm aus Fig. 4B dargestellt. Zu Beginn wird ein Paket an einem Port einer der Ethernet-Karten **415**, **416**, **417** oder **418** empfangen, und zwar bei Stufe **420**. Bei Stufe **425** überprüft der Ethernet-Controller **412** dann ein CPU-Interventions-Bit, um festzustellen, ob das Paket zur Weiterverarbeitung an die CPU-Karte **400** gesendet werden muss. In einem solchen Fall wird das Paket über den PCI-Bus **410** an die CPU-Karte **400** übertragen und von dem Speicher-Controller **406** in dem Speicher **404** gespeichert, und zwar bei Stufe **430**. Wenn das

CPU-Interventions-Bit jedoch nicht gesetzt ist, wird die Verarbeitung mit Stufe **445** fortgesetzt. Bei Stufe **435** wird ein optionaler Lastausgleichsvorgang ausgeführt, um zu bestimmen, zu welchem IP-Server **210**, **220**, **230**, **240** oder **250** das Paket **300** weitergeleitet werden soll. Bei dem Lastausgleichsvorgang auf Stufe **435** wird versucht, die zu verarbeitenden Pakete entsprechend der Kapazität und der momentanen Nutzung jedes Servers auf die IP-Server aufzuteilen. Ein Lastausgleichsschema, das zur Nutzung in der vorliegenden Erfindung geeignet ist, ist beschrieben in einer verwandten Anmeldung mit dem Titel "Dynamic Load Balancer for Multiple Network Servers" von Sajit Bhaskaran und Abraham Matthews, mit dem Aktenzeichen 08/992,038 und dem Anwaltsaktenzeichen M-4969 US, welche hier vollumfänglich durch Bezugnahme einbezogen wird. Bei Stufe **440** wird dann das Sicherungsschicht-Zieladressenfeld des Pakets **300** umgeschrieben, sodass es angibt, zu welchem der IP-Server **210**, **220**, **230**, **240** oder **250** das Paket **300** weitergeleitet werden soll. Schließlich wird das Paket zu einer der Ethernet-Karten **415**, **416**, **417** oder **418**, mit welcher der durch das Sicherungsschicht-Zieladressenfeld des Pakets **300** spezifizierte IP-Server verbunden ist, übertragen, und zwar bei Stufe **445**.

[0049] Der Vorgang des Weiterleitens eines Pakets **300** (Fig. 3A) von einem der IP-Server **210**, **220**, **230**, **240** oder **250** zu einem der Netzrouter **260**, **270** oder **280** (Fig. 2) wird durch das Ablaufdiagramm aus Fig. 4C dargestellt. Zu Beginn wird ein Paket an einem Port einer der Ethernet-Karten **415**, **416**, **417** oder **418**, die mit einem der IP-Server **210**, **220**, **230**, **240** oder **250** verbunden ist, empfangen, und zwar bei Stufe **450**. Optional wird dann in Stufe **455** überprüft, ob der Netzrouter, zu welchem das Paket **300** weitergeleitet werden soll, in Betrieb ist, in welchem Fall die Verarbeitung mit Stufe **465** fortgesetzt wird. Ein Fehlertoleranzschema, das zur Verwendung bei der vorliegenden Erfindung geeignet ist, ist beschrieben in einer verwandten Patentanmeldung mit dem Titel "Router Pooling in a Network Flowswitch" von Sajit Bhaskaran, mit dem Aktenzeichen 08/994,405 und dem Anwaltsaktenzeichen M-4971 US, welche hier durch Bezugnahme vollumfänglich einbezogen wird. Ansonsten überträgt der Ethernet-Controller **412** in der optionalen Stufe **460** das Paket **300** über den PCI-Bus **410** zu der CPU-Karte **400**, und der Speicher-Controller **406** speichert das Paket **300** in dem Speicher **404**. Noch in Stufe **460** schreibt die CPU **402** das Sicherungsschicht-Zieladressfeld **390** des Pakets **300** um, sodass dieses angibt, zu welchem der Netzrouter **260**, **270** oder **280** das Paket **300** weitergeleitet werden soll. Schließlich überträgt der Speicher-Controller **406** das Paket **300** über den PCI-Bus **410** zu einer der Ethernet-Karten **415**, **416**, **417** oder **418**, in Abhängigkeit von dem Inhalt des Sicherungsschicht-Zieladressenfeldes **390** des Pakets **300**, und zwar in Stufe **465**.

[0050] Bei einigen Ausführungsformen ermöglicht die Netzfluss-Vermittlungseinrichtung einen Lastausgleich und eine Clusterbildung für abgehende Pakete. In einem solchen Fall werden die Netzrouter in "Router-Pools" gruppiert, genauso wie die IP-Server für die eingehende Verarbeitung in Clustern gruppiert wurden. Verkehr von den IP-Servern, der zu den IP-Clients läuft, erfährt einen Lastausgleich, wenn mehrere Netzrouter und/oder mehrere Netzrouterverbindungen vorhanden sind. Wenn beispielsweise vier Netzrouter, jeder mit einem 100 Mbit/s-Ethernetport, mit der Netzfluss-Vermittlungseinrichtung verbunden sind, wird die Verkehrslast ungefähr auf die vier Verbindungen aufgeteilt, was einen Durchsatz von nahezu 400 Mbit/s ermöglicht, selbst wenn alle IP-Server jeweils mit einer einzigen und identischen Standard-Router-IP-Adresse konfiguriert sind.

[0051] Dies wird erreicht, indem die Netzfluss-Vermittlungseinrichtung derart programmiert wird, dass sie auf ARP-Anforderungen von den IP-Servern in Hinsicht auf eine IP-Adresse eines speziellen Netzrouters wie folgt antwortet. Die Netzfluss-Vermittlungseinrichtung verfolgt die Last, die zu allen Netzroutern in einem Router-Pool geht (z. B. indem sie die Vektoren <Pakete ein, Pakete aus, Bytes ein, Bytes aus> verfolgt). Die IP-Server unterhalten ARP-Zwischenspeicher für die IP-Adresse der Netzrouter. Der ARP-Zwischenspeicher wird aktualisiert, indem periodisch eine ARP-Anforderung nach einer IP-Adresse eines Netzrouters ausgegeben wird. Die Netzfluss-Vermittlungseinrichtung fängt die Anforderung ab, untersucht die IP-Adresse des IP-Servers und antwortet auf die Anforderung, indem sie die Sicherungsschicht-Adresse desjenigen Netzrouters in dem Pool zuordnet, der am besten in der Lage ist, die Last, die von diesem speziellen Server kommt, zu bedienen (der "beste" wird durch Messwerte der Verkehrslast in Echtzeit oder mit Hilfe eines einfachen zyklischen Schemas basierend auf den Server-IP-Quelladressen bestimmt).

[0052] Für die Zwecke des Ausgleichs der abgehenden Last werden die Netzrouter im Gegensatz zum Ausgleich für die eingehende Last mit eindeutigen IP-Adressen anstatt einer einzigen IP-Adresse konfiguriert.

[0053] Bei einigen Ausführungsformen kann die Netzfluss-Vermittlungseinrichtung dafür konfiguriert sein, nur eine "Verfügbarkeits-Clusterbildung" auszuführen. Bei der Verfügbarkeits-Clusterbildung dient ein Server als der primäre IP-Server, während alle anderen IP-Server in dem Cluster jeweils als sekundäre IP-Server wirken ("sekundär – betriebsfähig" oder "sekundär – ausgefallen"). Der Verkehr wird immer zu dem primären IP-Server weitergeleitet. Wenn der primäre IP-Server ausfällt, wird der Ausfall von der Netzfluss-Vermittlungseinrichtung automatisch erkannt, und der Status des ausgefallenen IP-Ser-

vers wird in "sekundär – ausgefallen" umgewandelt. Einer der verfügbaren IP-Server mit dem Status "sekundär – betriebsfähig" wird dann in den Status "primär" umgewandelt. Die Netzfluss-Vermittlungseinrichtung fährt fort, den Status der Server mit dem Status "sekundär – ausgefallen" zu überwachen und erkennt automatisch, wenn diese wieder betriebsfähig werden. Wenn dies geschieht, wird deren Status in "sekundär – betriebsfähig" geändert. Daher wird ein ausgefallener primärer IP-Server, der wiederhergestellt wird, nachdem er für gewisse Zeit den Status "sekundär – ausgefallen" hatte, niemals den momentan primären Server verdrängen, sondern vielmehr in den Status "sekundär – betriebsfähig" kommen.

[0054] Außerdem wird der Status jedes Netzrouters in einem Router-Pool überwacht. Wenn der Netzrouter ausfällt, wird der gesamte Verkehr, der an diesen Netzrouter gerichtet ist, transparent zu einem anderen Netzrouter in dem Router-Pool umgeleitet, bis der Netzrouter wiederhergestellt ist. Es ist keine Intervention von den IP-Servern aus erforderlich, da die Umleitung vollständig durch die Netzfluss-Vermittlungseinrichtung abgewickelt wird.

[0055] Die Fig. 5A–Fig. 5C stellen mehrere mögliche Hardware-Implementierungen der Netzfluss-Vermittlungseinrichtung **205** (Fig. 2 und Fig. 4A) dar. Jede der Hardware-Implementierungen aus den Fig. 5A–Fig. 5C stellt einen anderen Kompromiss zwischen einer einfachen Implementierung und der Leistungsfähigkeit der Netzfluss-Vermittlungseinrichtung dar. Beispielsweise erfordert die Hardware-Implementierung aus Fig. 5A keinerlei spezielle Hardware und kann mit Hilfe von handelsüblichen Komponenten implementiert werden.

[0056] In den Fig. 5A–Fig. 5D stellt die CPU einen Prozessor Modell R-4700 dar, der bei der Integrated Devise Technology Inc., San Jose, Kalifornien erhältlich ist, der Speicher-Controller stellt einen Controller Modell GT-64010 dar, der von der Galileo Technologies Ltd., Karmiel, Israel erhältlich ist, und die Ethernet-Controller stellen Ethernet-Controller Modell GT-48002 dar, die ebenfalls bei der Galileo Technologies erhältlich sind. Wenngleich diese speziellen Hardware-Komponenten der Deutlichkeit halber beschrieben sind, ist die Erfindung nicht auf die speziellen Komponenten, Hersteller oder Modellnummern begrenzt. Anstatt der in den

[0057] Fig. 5A–Fig. 5C beschriebenen Komponenten können auch andere Komponenten, die von anderen Herstellern hergestellt werden, und mit anderen Modellnummern, verwendet werden.

[0058] Fig. 5A zeigt eine erste Hardware-Implementierung der Netzfluss-Vermittlungseinrichtung **205** mit einer CPU-Karte **500** und mehreren Ethernet-Karten **510**, **520**, **530** und **540**. Die CPU-Karte **500** weist einen Prozessor R-4700 auf, der mit einem asynchronen E/A-Controller 85C30 und mit einem Speicher-Controller GT-64010 verbunden ist. Der asynchrone Controller ist wiederum mit einem Paar Eingangs-/Ausgangs-Ports RS232/DB-25 verbunden, zur Kopplung mit anderen Einrichtungen. Der Speicher-Controller ist außer mit dem PCI-Bus **410** auch mit einem EPROM mit 512 kB, einem RAM mit 8 MB und einem Flash-Speicher mit 2MB verbunden. Die Ethernet-Karten **510**, **520**, **530** und **540** weisen einen Ethernet-Controller GT-48002, einen EDO RAM mit 1 MB und ein Paar Eingangs-/Ausgangs-Ports auf. Die CPU-Karte **500** und die Ethernet-Karten **510**, **520**, **530** und **540** stellen Allzweck-Schaltungskarten dar, die bei der Galileo Technologies erhältlich sind. Infolgedessen kann die Netzfluss-Vermittlungseinrichtung **205** nur mit Hilfe von Allzweck-Komponenten implementiert werden, wie in Fig. 5A dargestellt ist.

[0059] Fig. 5B stellt eine zweite Hardware-Implementierung der Netzfluss-Vermittlungseinrichtung **205** (Fig. 2 und Fig. 4A) dar. In Fig. 5B wird anstelle der Allzweck-Netzkarten aus Fig. 5A eine Spezial-Netzkarte **560** genutzt. Somit sind die Ethernet-Karten **510**, **520**, **530** und **540** durch eine einzige Netzkarte **560** ersetzt. Die Netzkarte **650** umfasst ihrerseits mehrere Ethernet-Controller, die jeweils mit einem Paar Eingangs-/Ausgangs-Ports sowie mit einem karteninternen PCI-Bus verbunden sind. Der externe PCI-Bus aus Fig. 5A fällt vollständig weg. Die Hardware-Implementierung aus Fig. 5B bietet ein verbessertes Leistungsverhalten sowie eine Kostenreduzierung gegenüber der Hardware-Implementierung aus Fig. 5A, auf Kosten des Hinzufügens von Spezial-Hardware.

[0060] Fig. 5C stellt eine dritte Hardware-Implementierung der Netzfluss-Vermittlungseinrichtung **205** (Fig. 2 und Fig. 4A) dar. In Fig. 5C werden anstelle der Allzweck-Schaltungskarten aus Fig. 5A zwei Spezial-Schaltungskarten genutzt. Die CPU-Karte **550** weist die gleichen Komponenten wie die CPU-Karte **500** aus Fig. 5A auf, außer dass ein FSRAM mit 4 MB hinzugefügt ist. Zusätzlich könnten ein inhaltsadressierbarer Speicher (CAM) sowie schnelle PLDs hinzugefügt werden, um das Leistungsverhalten der CPU-Karte **550** zu beschleunigen. Die Ethernet-Karten **510**, **520**, **530** und **540** sind jedoch durch eine einzige Netzkarte **560** ersetzt, wie mit Bezug auf Fig. 5B erklärt worden ist. Die Hardware-Implementierung aus Fig. 5C bietet ein besseres Leistungsverhalten gegenüber der Hardware-Implementierung aus den Fig. 5A und Fig. 5B (d. h. eine Unterstützung für Übertragungsraten von 100 Mbit/s und ein schnelleres CPU-Verhalten), auf Kosten des Hinzufügens von Spezial-Hardware.

[0061] Fig. 5D stellt noch eine dritte Hardware-Implementierung der Netzfluss-Vermittlungseinrichtung

205 (Fig. 2 und Fig. 4A) dar, bei welcher die gesamte Vermittlungseinrichtung auf einer einzigen Schaltungskarte **570** bereitgestellt wird. Die Schaltungskarte **570** weist sämtliche Komponenten der CPU-Karte **550** und der Netzkarte **560** aus Fig. 5C auf, außer dass der karteninterne PCI-Bus durch einen Pufferspeicher-Arbiter ersetzt ist. Das Eliminieren des PCI-Busses ermöglicht ein noch weiter verbessertes Leistungsverhalten (Übertragungsraten von mehr als 1 Gbit/s), auf Kosten einer teureren Spezial-Hardware.

[0062] Fig. 5E stellt eine weitere Hardware-Implementierung der Netzfluss-Vermittlungseinrichtung **205** (Fig. 2 und Fig. 4A) dar, bei welcher eine Spezial-Schaltungskarte **575** in Kombination mit Ethernet-Karten **510**, **520**, **530** und **540** (Fig. 5A) genutzt wird. Die Schaltungskarte **575** weist die gleichen Komponenten wie die Schaltungskarte **500** aus Fig. 5A auf, außer dass ein CPLD **585** und ein Doppelport-SRAM **580** hinzugefügt sind. Die Schaltungskarte **575** ist mit den Ethernet-Karten **510**, **520**, **530** und **540** über den PCI-Bus **410** verbunden. Bei dieser Ausführungsform erfolgen die Übersetzungen der Sicherungsschicht-Adresse durch den CPLD **585** anstatt durch die CPU R-4700, was eine schnellere Verarbeitung von Paketen ermöglicht. Die CPU R-4700 führt immer noch Verwaltungsaufgaben aus, beispielsweise das periodische Überprüfen der Lasten an jedem der IP-Server, das Erkennen von Ausfällen von IP-Servern und Netzroutern, usw.

[0063] Fig. 5F stellt eine weitere Hardware-Implementierung der Netzfluss-Vermittlungseinrichtung **205** (Fig. 2 und Fig. 4A) mit Hilfe eines Crossbar-Switch anstelle des PCI-Busses **410** dar. In Fig. 5F verbindet der Crossbar-Switch **594** die Verwaltungsprozessorkarten **590** und **592** mit Ethernet-Karten **582** und **584** sowie Schaltungskarten **586** und **588**. Jede der Schaltungskarten **586** und **588** umfasst einen ASIC **596**, welcher eine Nachschlagetabelle **598** mit einem Sicherungsschicht-Chip **595** verbindet. Bei dieser Ausführungsform werden die Verwaltungsprozessorkarten **590** und **592** genutzt, um Verwaltungsaufgaben auszuführen, wie zuvor im Zusammenhang mit Fig. 5E beschrieben, die Ethernet-Karten **582** und **584** werden für den abgehenden Paketfluss genutzt, wie zuvor im Zusammenhang mit Fig. 5A beschrieben, und die Schaltungskarten **586** und **588** werden genutzt, um die Sicherungsschicht-Adressfelder der eingehenden Pakete zu übersetzen. Dies wird erreicht, indem das Sicherungsschicht-Zieladressfeld des Pakets in dem Sicherungsschicht-Chip **595** extrahiert wird und ein schneller Suchvorgang in der Nachschlagtabelle **598** erfolgt, in welcher die Sicherungsschicht-Adresse des IP-Servers mit einer optimalen Last gespeichert ist. Sicherungsschicht-Chips, die zur Nutzung bei der vorliegenden Erfindung geeignet sind, sind unter anderem bei der Galileo Technologies, bei I-Cube und bei der MMC Networks erhältlich. Wenn eine Fehlertoleranz für die Netzrouter bereitgestellt wird, werden die Schaltungskarten **586** und **588** außerdem genutzt, um das Sicherungsschicht-Adressfeld abgehender Pakete, die aufgrund eines Netzrouter-Ausfalls umgeleitet werden, zu übersetzen.

[0064] Um das Leistungsverhalten zu verbessern, sollte jeder der IP-Server **210**, **220**, **230**, **240** und **250** sowie der Router **260**, **270** und **280** (entweder direkt oder über ein Netzwerk) mit der Netzfluss-Vermittlungseinrichtung **205** über einen Vermittlungsport (Switched Port) mit fest zugeordneter Vollduplex-Bandbreite verbunden sein. Die Netzfluss-Vermittlungseinrichtung **205** (Fig. 2 und Fig. 4A) funktioniert jedoch selbst für den Fall gut, dass sie mit einem der IP-Server über einen gemeinsam genutzten Medienport (Shared Media) verbunden ist. Jeder der IP-Server **210**, **220**, **230**, **240** und **250** ist somit anders konfiguriert, in Abhängigkeit davon, ob der Server mit der Netzfluss-Vermittlungseinrichtung **205** über einen gemeinsam genutzten anstatt einen Vermittlungsport verbunden ist. Jeder IP-Server wird automatisch im Moment des Hochfahrens konfiguriert, indem in dem Server ein Computerprogramm ausgeführt wird.

[0065] Bei einer Ausführungsform der Erfindung sind alle oder einige der Router und Server mit Hilfe von Vermittlungsleitungen auf der Sicherungsschicht verbunden. Damit wird für jede Einrichtung, die mit der Flussvermittlungseinrichtung verbunden ist, (a) eine dedizierte Bandbreite und (b) ein Vollduplex-Betrieb zur Verfügung gestellt. Fachleute auf dem Gebiet werden jedoch erkennen, dass die Netzfluss-Vermittlungseinrichtung gemäß der vorliegenden Erfindung auch auf nicht vermittelte Umgebungen angewendet werden kann (z. B. Ethernet-Hubs mit gemeinsam genutzten Medien (Shared Media) oder gemeinsam genutzte Ports unter Verwendung von kaskadierten Ethernet-Switchen.

[0066] Die vorstehend beschriebenen Ausführungsformen veranschaulichen die Erfindung, schränken diese aber nicht ein. Insbesondere ist die Erfindung nicht auf irgendwelche spezielle Hardware beschränkt, die genutzt wird, um die Netzfluss-Steuerungsvermittlungseinrichtung zu implementieren. Die Erfindung ist in jedem Fall nicht auf irgendeine spezielle Anzahl von Ethernet-Karten oder irgendeine spezielle Art von Prozessor, Speicher-Controller oder Bus eingeschränkt. Insbesondere kann eine beliebige Anzahl von Ethernet-Cards mit einer beliebig großen Anzahl von physischen Verbindungsports entsprechend der vorliegenden Erfindung genutzt werden. Erfindungsgemäß können auch andere Prozessoren als der R-4700 und der GT-64010 genutzt werden. Es können andere Ethernet-Vermittlungs-ASICs als der GT-48002A von Galileo genutzt werden, von Galileo oder von anderen Anbietern wie beispielswei-

se I-Cube oder MMC Networks. Darüber hinaus kann anstelle der CPU **402** und des Speicher-Controllers **406** (Fig. 4A) ein einziger Prozessor genutzt werden. Anstelle eines PCI-Busses **410** (Fig. 4A) können andere Busse als ein PCI-Bus (z. B. SCSI-Busse) oder sogar Crossbar-Switche genutzt werden. Schließlich können anstelle der Ethernet-Karten **415**, **416**, **417** und **418** (Fig. 4A) andere Netzkarten als Ethernet-Karten genutzt werden. Darüber hinaus ist die Erfindung nicht auf irgendeine Art oder Anzahl von Netzkarten eingeschränkt. Tatsächlich kann die Erfindung auf eine beliebige Anzahl von Netzkarten angewandt werden, die mit einer beliebigen Anzahl von Netzen verbunden sind. Auch andere Ausführungsformen und Varianten fallen in den Schutzumfang der Erfindung, wie er durch die folgenden Ansprüche definiert wird.

Patentansprüche

1. Netzfluss-Vermittlungseinrichtung (**205**) zum Weiterleiten von Paketen zu einer Mehrzahl (**200**) von IP-Servern (**210**–**250**) und von diesen weg, wobei jeder der IP-Server eine gleiche IP-Adresse (**290**) sowie eine Sicherungsschicht(Data Link Layer)-Adresse, die sich von der Sicherungsschicht-Adresse der anderen IP-Server unterscheidet, aufweist, und wobei die Netzfluss-Vermittlungseinrichtung umfasst:
einen Prozessor (**402**);
einen Speicher (**404**), der mit dem Prozessor verbunden ist; und
eine Mehrzahl von Netzports (**415**–**418**), die mit einem Netzwerk verbunden sind; **dadurch gekennzeichnet**, dass
die Netzfluss-Vermittlungseinrichtung dafür ausgelegt ist, ein Paket (**300**), das an einem ersten Netzport empfangen (**420**) wird, zu einem zweiten Netzport weiterzuleiten (**445**), indem sie eine Sicherungsschicht Adresse eines der IP-Server in das Paket schreibt (**440**).

2. Vermittlungseinrichtung nach Anspruch 1, bei welcher jeder Netzport (**510**–**540**) ferner einen Controller und einen Speicher umfasst und bei welcher das Weiterleiten von Paketen von einem der IP-Server zu einem Netzwerk-Ziel keine Intervention durch den Prozessor erfordert.

3. Vermittlungseinrichtung nach Anspruch 1, bei welcher das Paket entsprechend einem anderen ISO-Schicht-4-Transportprotokoll als TCP kodiert ist.

4. Verfahren zum Weiterleiten von Paketen zu einer Mehrzahl (**200**) von IP-Servern (**210**–**250**) und von diesen weg, wobei jeder der IP-Server eine gleiche IP-Adresse (**290**) sowie eine Sicherungsschicht-Adresse, die sich von der Sicherungsschicht-Adresse der anderen IP-Server unterscheidet, aufweist, wobei das Verfahren umfasst:
Empfangen eines Pakets (**300**) für die IP-Adresse der IP-Server; und
Weiterleiten des Pakets an zumindest einen der IP-Server;
dadurch gekennzeichnet, dass
das Empfangen umfasst:
Empfangen (**420**) des Pakets in einer Netzfluss-Vermittlungseinrichtung (**205**), welche der IP-Adresse der IP-Server entspricht; und
das Weiterleiten umfasst:
Weiterleiten (**445**) des Pakets an den zumindest einen der IP-Server durch Schreiben (**440**) der Ziel-Sicherungsschicht-Adresse des IP-Servers in das Paket in der Netzfluss-Vermittlungseinrichtung.

5. Verfahren nach Anspruch 4, ferner umfassend:
Empfangen (**450**) eines Pakets in der Netzfluss-Vermittlungseinrichtung von einem der IP-Server;
Extrahieren (**465**) einer Zieladresse aus dem Paket; und
Weiterleiten (**465**) des Pakets zu einem Netzwerk-Ziel basierend auf der Zieladresse des Pakets.

6. Verfahren nach Anspruch 5, bei welchem das Weiterleiten des Pakets keine Intervention durch einen Prozessor (**402**) der Netzfluss-Vermittlungseinrichtung erfordert.

7. Verfahren nach Anspruch 4, bei welchem das Paket entsprechend einem anderen ISO-Schicht-4-Transportprotokoll als TCP kodiert wird.

8. Verfahren nach Anspruch 4 zum Ausführen einer fehlertoleranten Lenkung von Paketen zu einem von einer Mehrzahl von IP-Servern und von diesem weg, ferner umfassend:
Senden eines oder mehrerer Pakete von einem Client aus, der mit einem Netzwerk mit einem Netzrouter (**260**–**280**) verbunden ist;
kontinuierliches Überwachen (**425**–**435**), und zwar in einer Netzfluss-Vermittlungseinrichtung, eines Status' jedes der Mehrzahl von IP-Servern, welche die gleiche IP-Adresse sowie eine Sicherungsschicht-Adresse, die sich von der Sicherungsschicht-Adresse der anderen IP-Server unterscheidet, aufweisen; und
Weiterleiten (**420**, **440**, **445**) der Pakete durch die Netzfluss-Vermittlungseinrichtung von dem Netzrouter zu einem der Mehrzahl von IP-Servern in einem Betriebsstatus.

9. Verfahren nach Anspruch 8, bei welchem die Pakete entsprechend einem anderen ISO-Schicht-4-Transportprotokoll als TCP kodiert werden.

10. Computerlesbares Medium (**404**), welches Anweisungen enthält, die, wenn sie in einem Computer (**205**) ausgeführt werden, bewirken, dass der

Computer alle Verfahrensschritte gemäß einem der Ansprüche 4 bis 9 ausführt.

Es folgen 11 Blatt Zeichnungen

Anhängende Zeichnungen

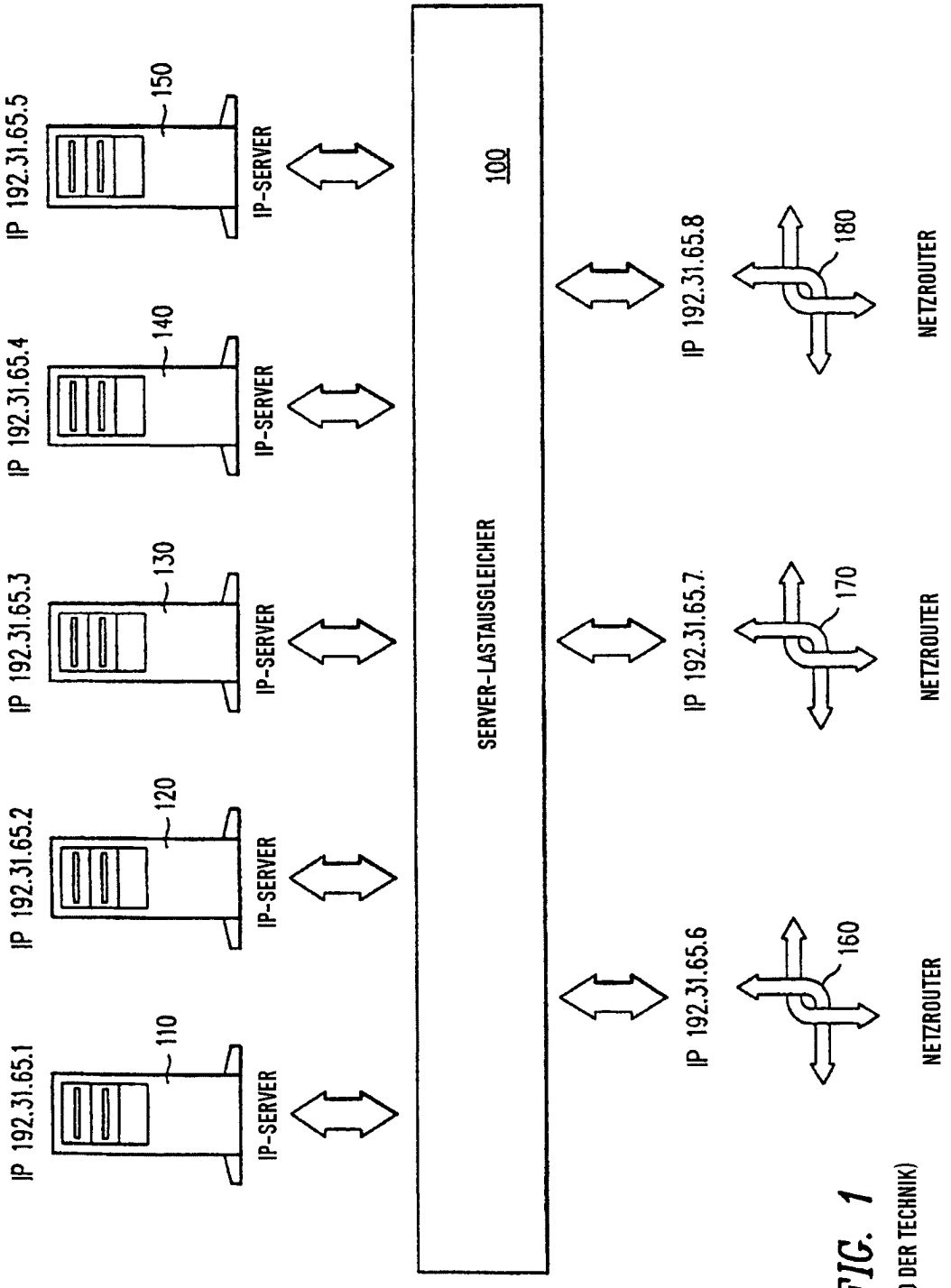

FIG. 1
(STAND DER TECHNIK)

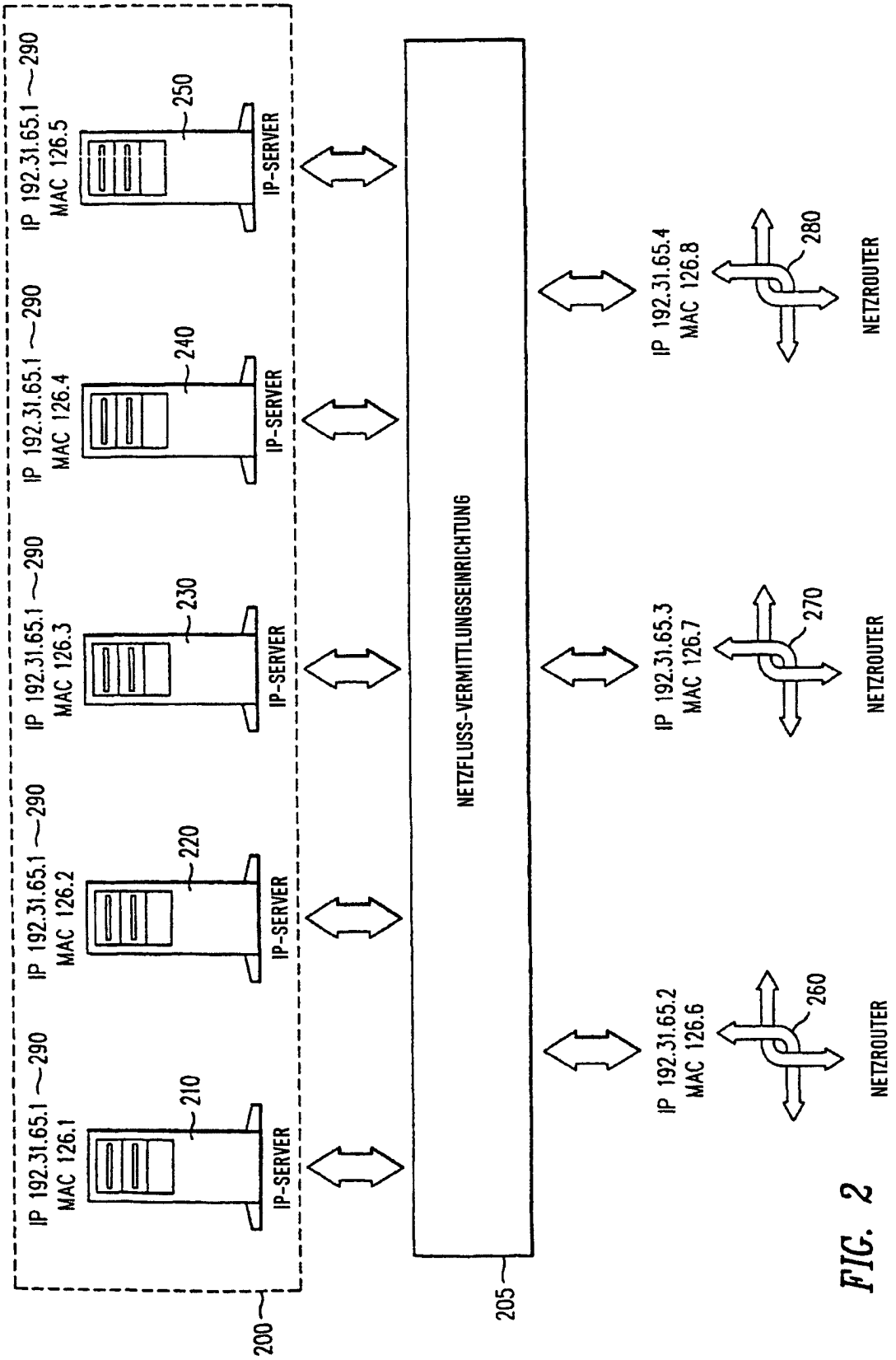
IP 192.31.65.1
MAC 126.1
290
210
IP-SERVER
IP 192.31.65.1
MAC 126.2
290
220
IP-SERVER
IP 192.31.65.1
MAC 126.3
290
230
IP-SERVER
IP 192.31.65.1
MAC 126.4
290
240
IP-SERVER
IP 192.31.65.1
MAC 126.5
290
250
IP-SERVER
200
NETZFLUSS-VERMITTLUNGSEINRICHTUNG
205
IP 192.31.65.2
MAC 126.6
260
NETZROUTER
IP 192.31.65.3
MAC 126.7
270
NETZROUTER
IP 192.31.65.4
MAC 126.8
280
NETZROUTER

FIG. 2

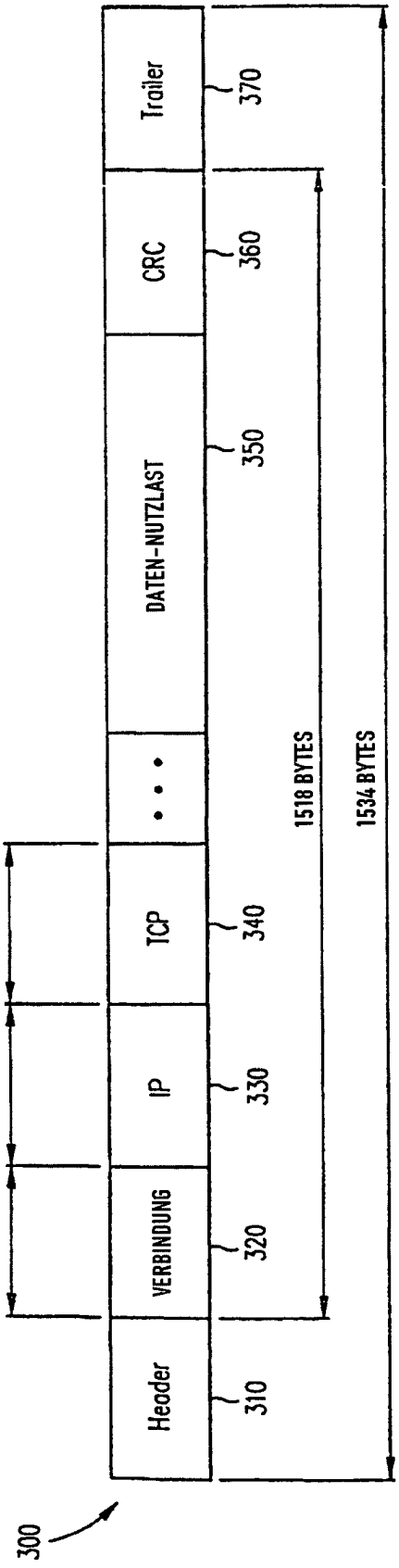
300
Header
VERBINDUNG
IP
TCP
• • •
DATEN-NUTZLAST
CRC
Trailer
310
320
330
340
350
360
370
14 BYTES
20 BYTES
20 BYTES
1518 BYTES
1534 BYTES

FIG. 3A

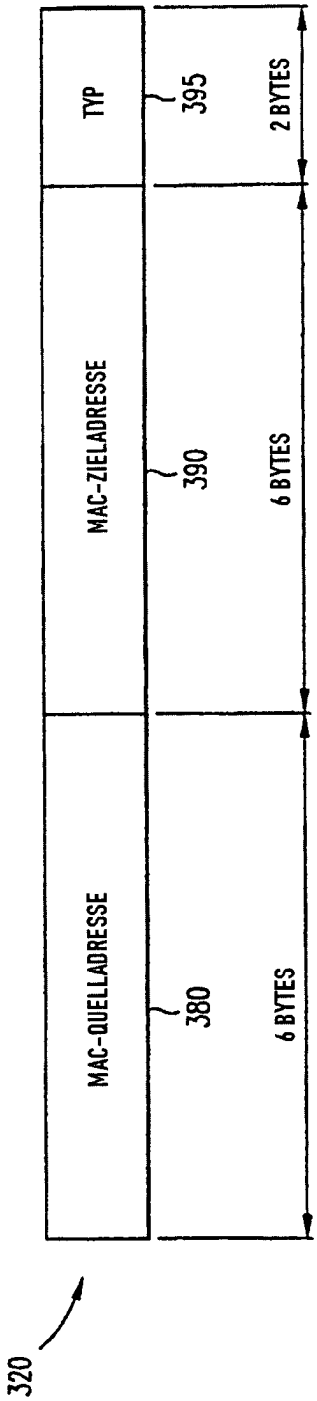
320
MAC-QUELLADRESSE
MAC-ZIELADRESSE
TYP
380
390
395
6 BYTES
6 BYTES
2 BYTES

FIG. 3B

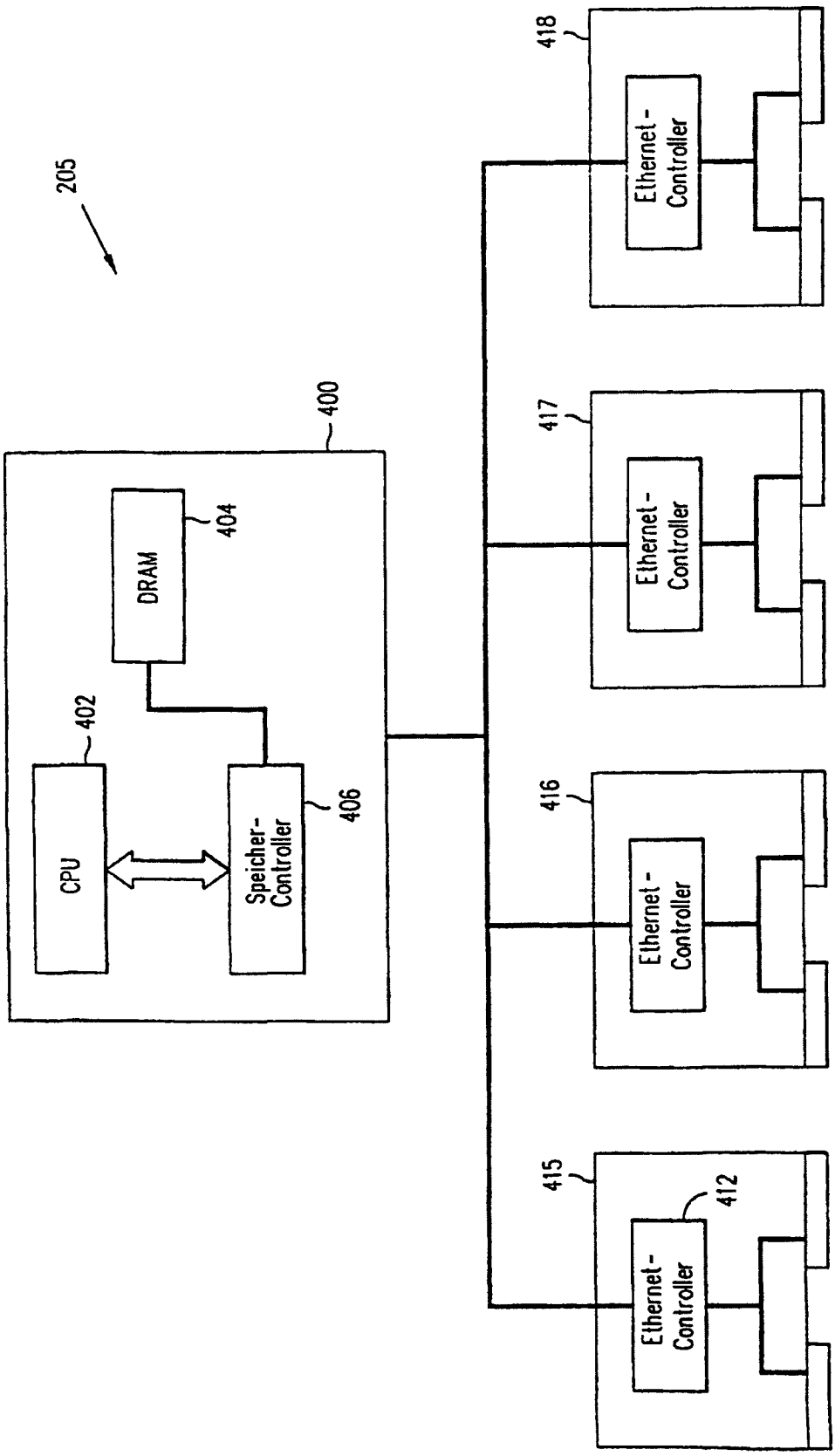
205
400
DRAM
404
402
CPU
Speicher-Controller
406
418
Ethernet-Controller
417
Ethernet-Controller
416
Ethernet-Controller
415
412
Ethernet-Controller

FIG. 4A

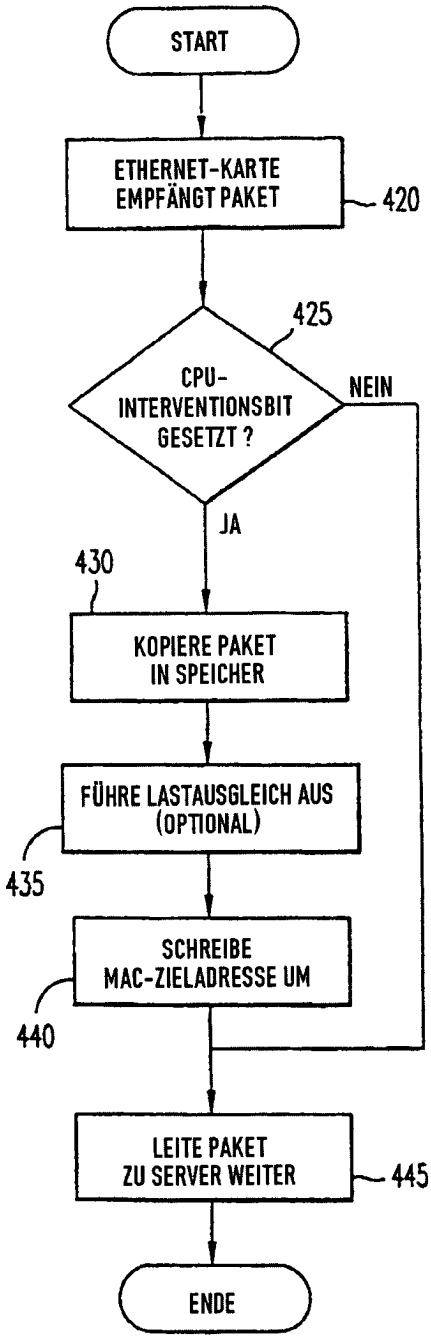
START
ETHERNET-KARTE EMPFÄNGT PAKET
420
425
CPU-INTERVENTIONSBIT GESETZT ?
NEIN
JA
430
KOPIERE PAKET IN SPEICHER
FÜHRE LASTAUSGLEICH AUS (OPTIONAL)
435
SCHREIBE MAC-ZIELADRESSE UM
440
LEITE PAKET ZU SERVER WEITER
445
ENDE

FIG. 4B

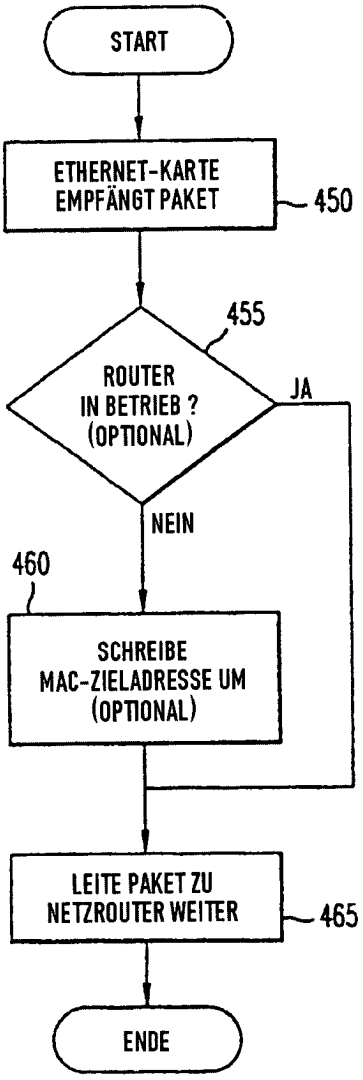
START
ETHERNET-KARTE EMPFÄNGT PAKET
450
455
ROUTER IN BETRIEB ? (OPTIONAL)
JA
NEIN
460
SCHREIBE MAC-ZIELADRESSE UM (OPTIONAL)
LEITE PAKET ZU NETZROUTER WEITER
465
ENDE

FIG. 4C

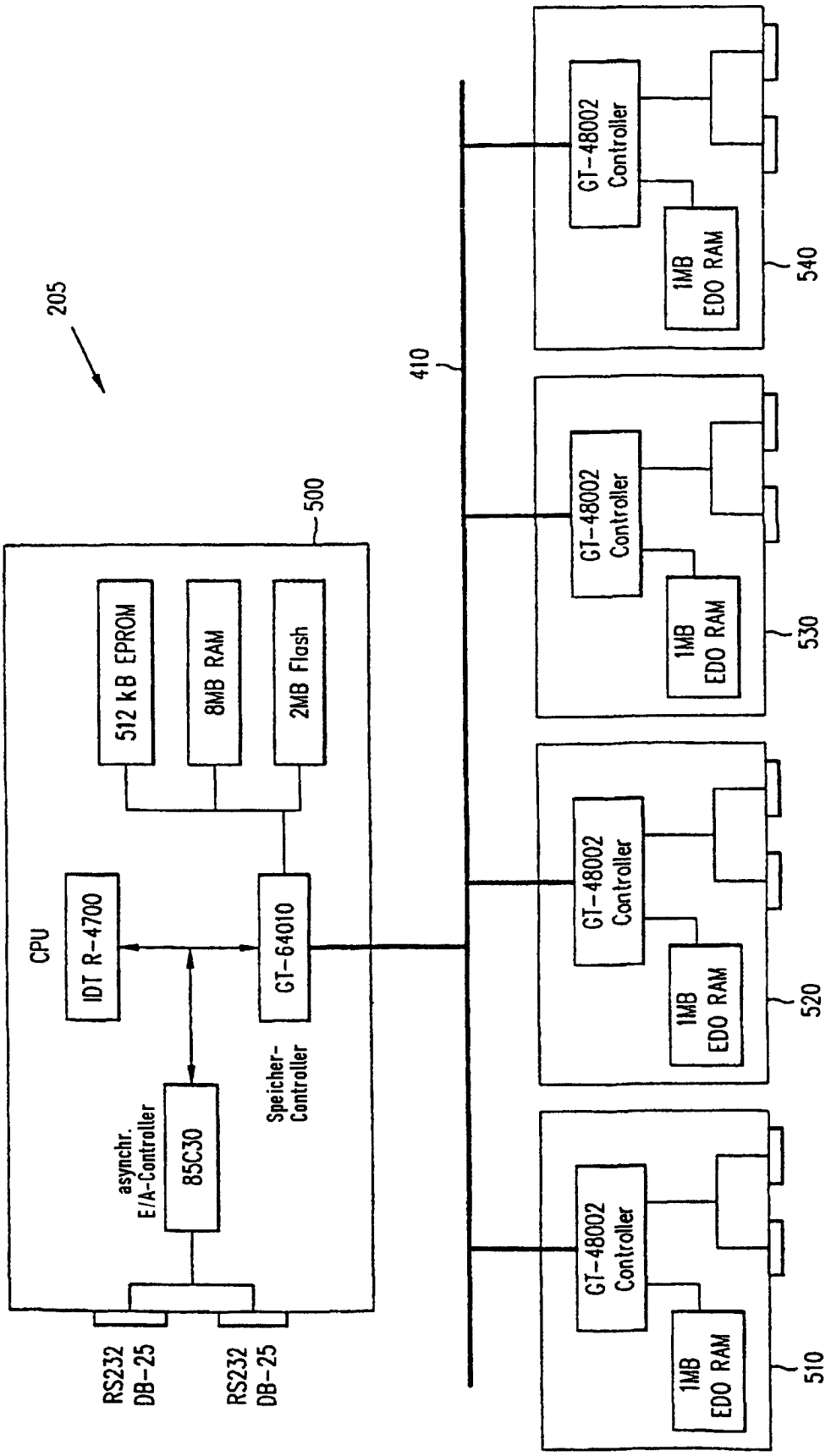
205
500
CPU
IDT R-4700
512 kB EPROM
8MB RAM
2MB Flash
GT-64010
Speicher-Controller
asynchr. E/A-Controller
85C30
RS232 DB-25
RS232 DB-25
410
GT-48002 Controller
1MB EDO RAM
510
520
530
540

FIG. 5A

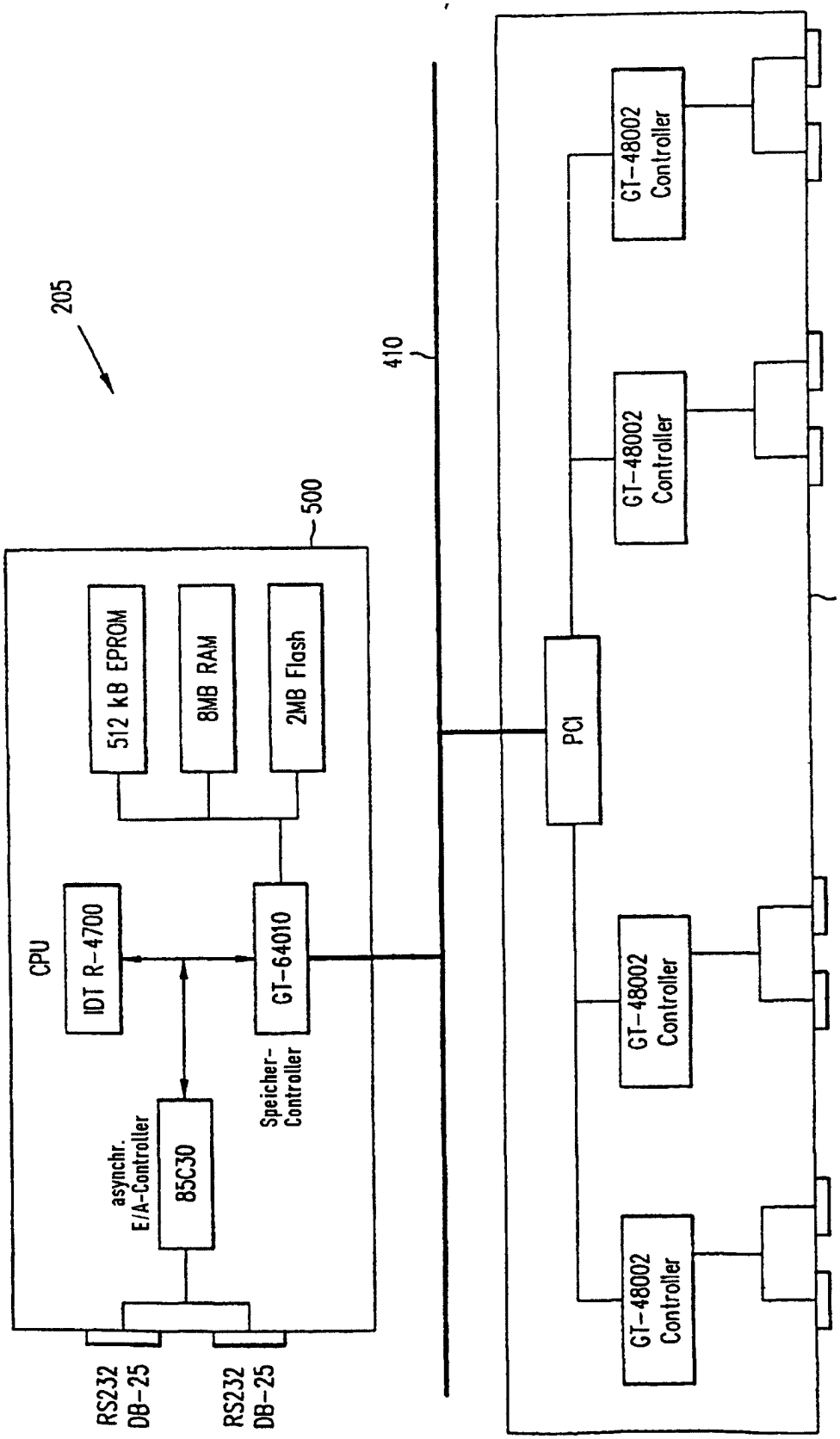
205
410
500
512 kB EPROM
8MB RAM
2MB Flash
CPU
IDT R-4700
GT-64010
Speicher-Controller
asynchr. E/A-Controller
85C30
RS232 DB-25
RS232 DB-25
PCI
560
GT-48002 Controller
GT-48002 Controller
GT-48002 Controller
GT-48002 Controller

FIG. 5B

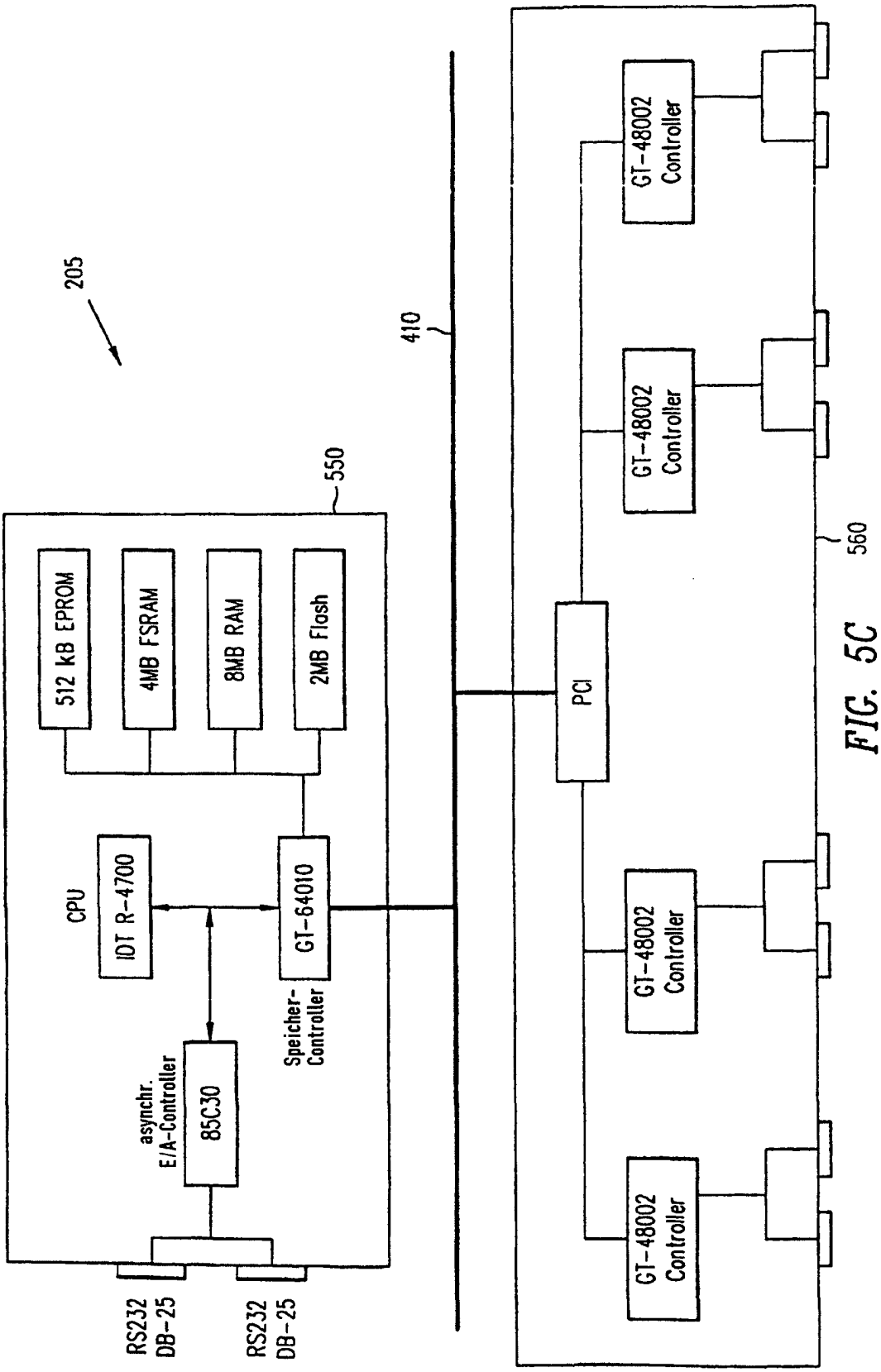
205
410
550
560
512 kB EPROM
4MB FSRAM
8MB RAM
2MB Flash
CPU
IDT R-4700
GT-64010
Speicher-Controller
asynchr. E/A-Controller
85C30
RS232 DB-25
RS232 DB-25
PCI
GT-48002 Controller
GT-48002 Controller
GT-48002 Controller
GT-48002 Controller

FIG. 5C

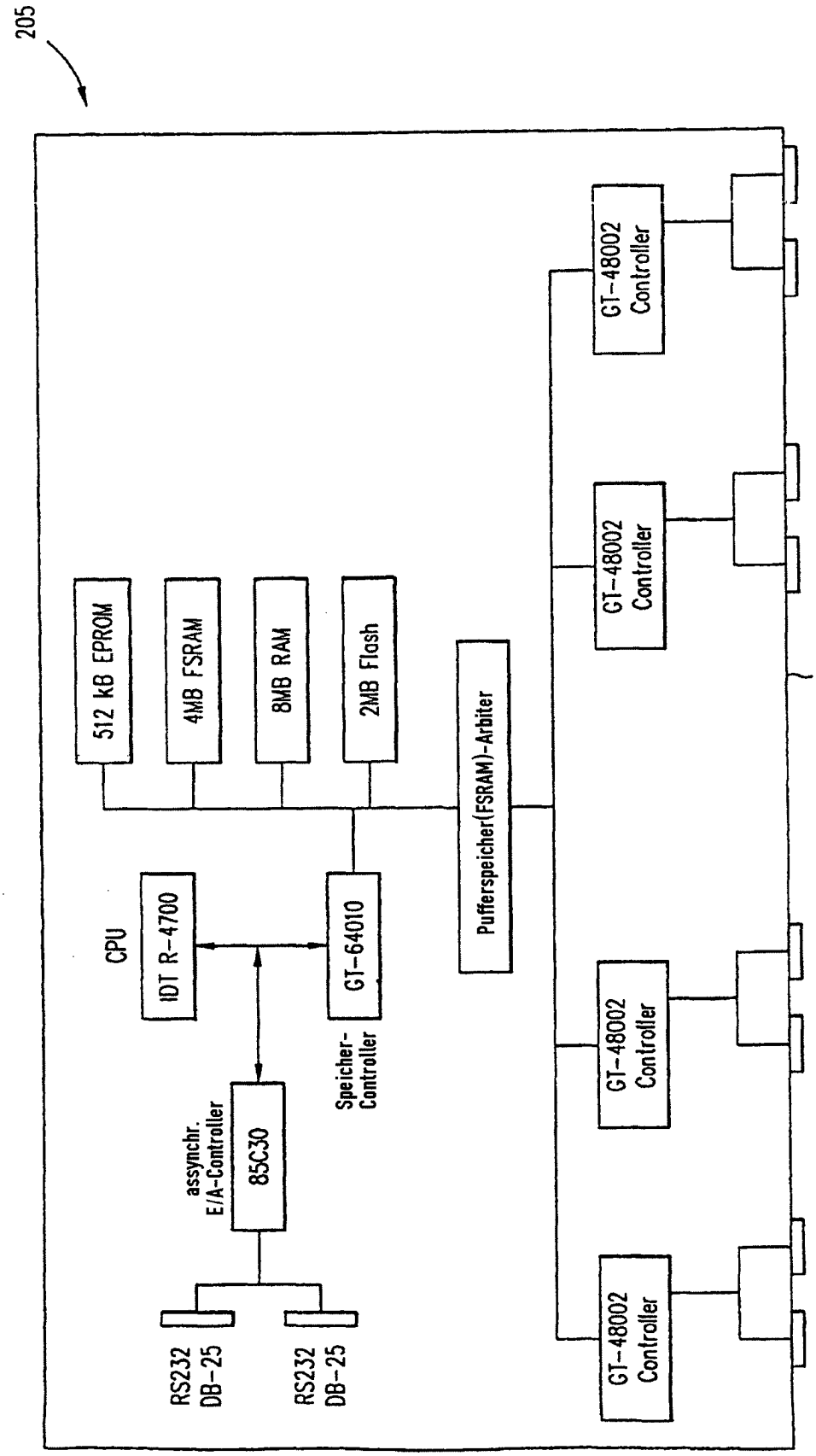
205
512 kB EPROM
4MB FSRAM
8MB RAM
2MB Flash
Pufferspeicher(FSRAM)-Arbiter
570
GT-48002 Controller
GT-48002 Controller
GT-48002 Controller
GT-48002 Controller
CPU
IDT R-4700
GT-64010
Speicher-Controller
assynchr. E/A-Controller
85C30
RS232 DB-25
RS232 DB-25

FIG. 5D

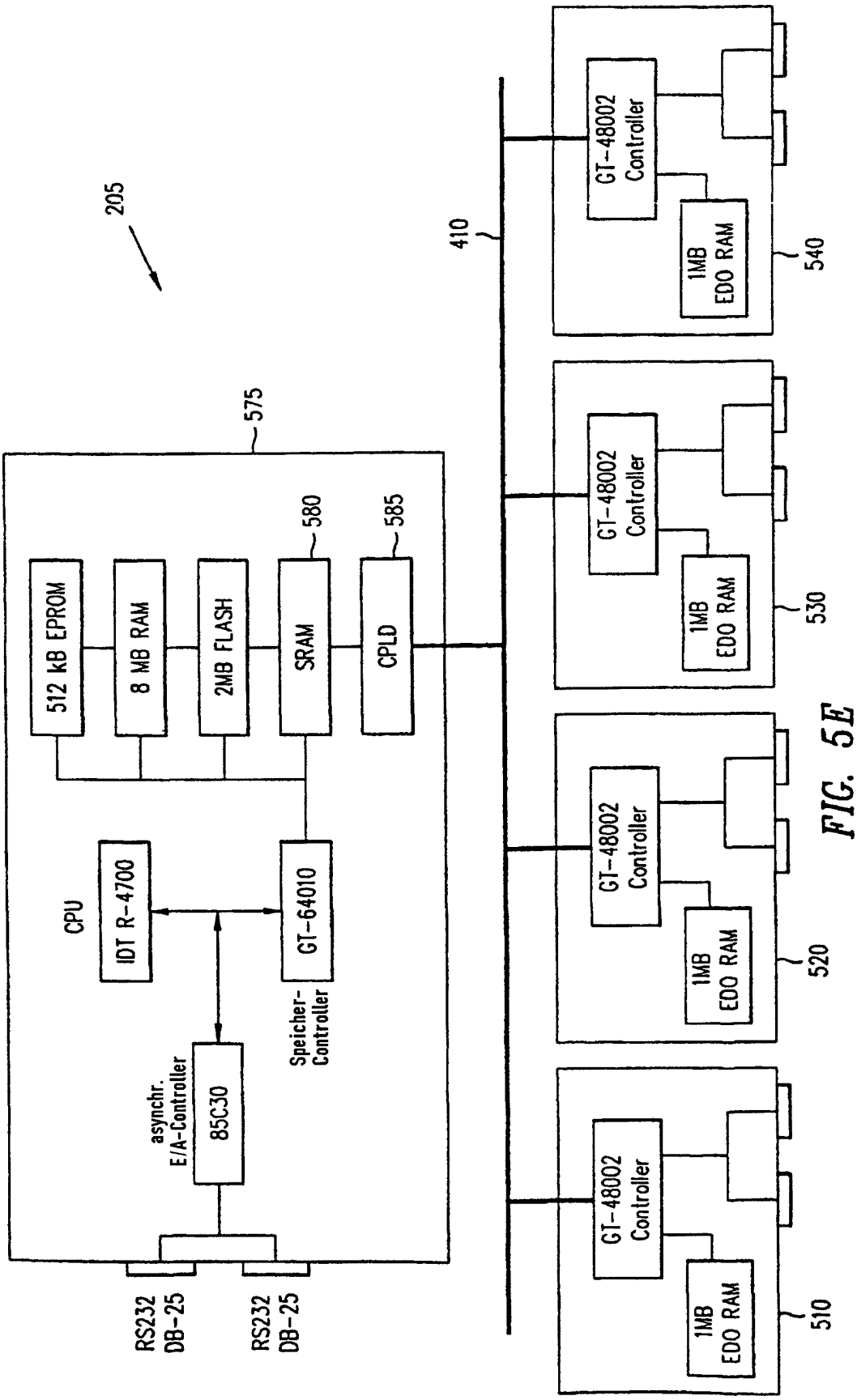
205
575
512 kB EPROM
8 MB RAM
2MB FLASH
SRAM
580
CPLD
585
CPU
IDT R-4700
GT-64010
Speicher-Controller
asynchr. E/A-Controller
85C30
RS232 DB-25
RS232 DB-25
410
GT-48002 Controller
1MB EDO RAM
510
520
530
540

FIG. 5E

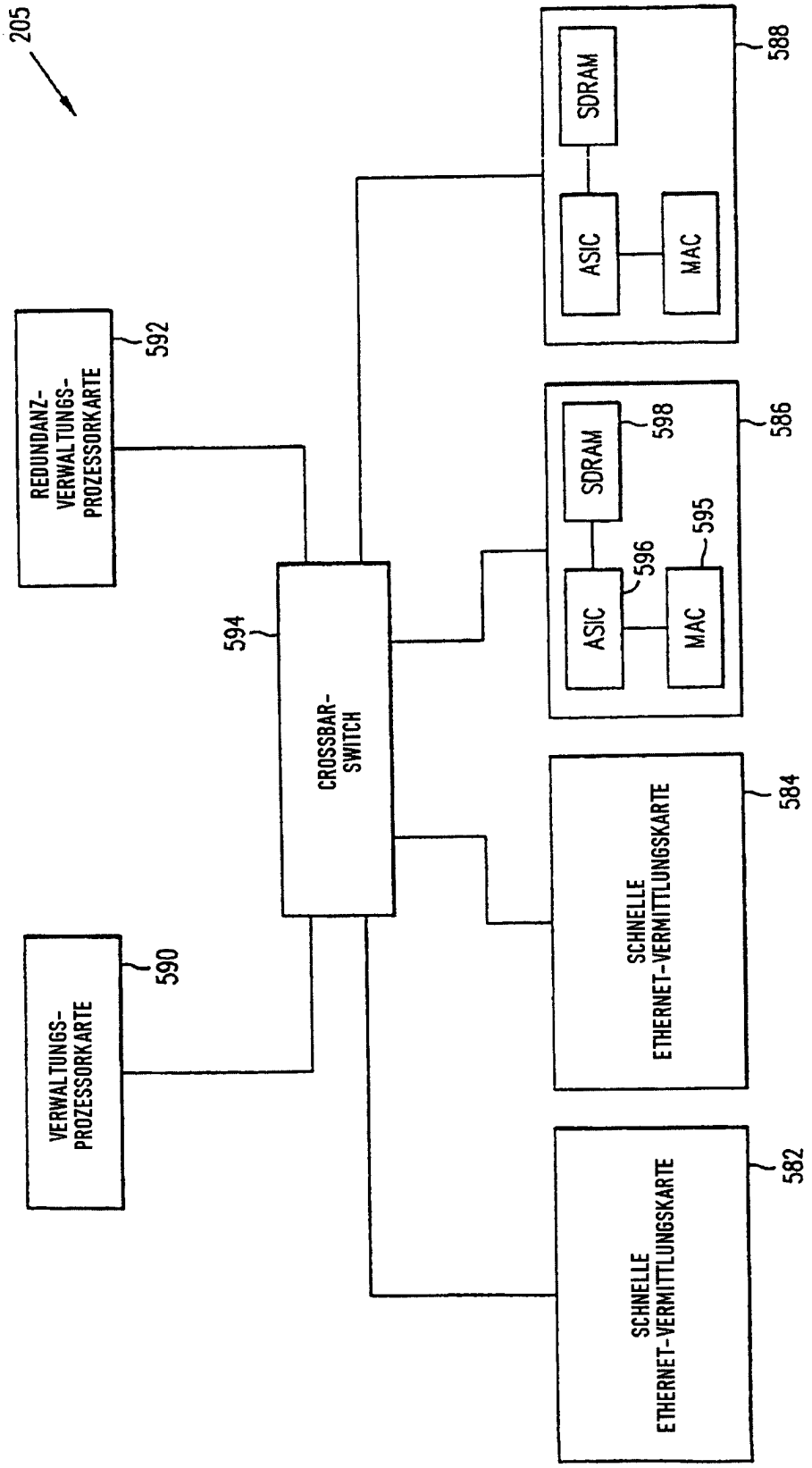
205
588
SDRAM
ASIC
MAC
592
REDUNDANZ-VERWALTUNGS-PROZESSORKARTE
586
598
SDRAM
596
595
ASIC
MAC
594
CROSSBAR-SWITCH
584
SCHNELLE ETHERNET-VERMITTLUNGSKARTE
590
VERWALTUNGS-PROZESSORKARTE
582
SCHNELLE ETHERNET-VERMITTLUNGSKARTE

FIG. 5F