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

(10) **DE 601 11 457 T2** 2006.05.11

(12) # Übersetzung der europäischen Patentschrift

(97) **EP 1 168 725 B1**
(21) Deutsches Aktenzeichen: **601 11 457.4**
(96) Europäisches Aktenzeichen: **01 305 335.0**
(96) Europäischer Anmeldetag: **19.06.2001**
(97) Erstveröffentlichung durch das EPA: **02.01.2002**
(97) Veröffentlichungstag der Patenterteilung beim EPA: **15.06.2005**
(47) Veröffentlichungstag im Patentblatt: **11.05.2006**

(51) Int Cl.[8]: ***H04L 12/56*** (2006.01)
H04L 12/46 (2006.01)

(30) Unionspriorität:

212592	**19.06.2000**	**US**
229305	**01.09.2000**	**US**

(73) Patentinhaber:
Broadcom Corp., Irvine, Calif., US

(74) Vertreter:
Bosch, Graf von Stosch, Jehle Patentanwaltsgesellschaft mbH, 80639 München

(84) Benannte Vertragsstaaten:
DE, FR, GB

(72) Erfinder:
Kalkunte, Mohan.c/o Broadcom Corporation, Irvine, US; Ambe,Shekhar.c/o Broadcom Corpration, Irvine, US; Pullela,Soma.c/o Broadcom Corporatio, Irvine, US; Lund, Martin.c/o Broadcom Corporation, Irvine, US; Kadambi,Shiri.c/o Broadcom Corporati, San Jose, US; Battle, Jim.c/o Broadcom Corporation, Irvine, US; Tai, Daniel.c/o Broadcom Corporation, Irvine, US

(54) Bezeichnung: **Vermittlungsanordnung mit redundanten Wegen**

Anmerkung: Innerhalb von neun Monaten nach der Bekanntmachung des Hinweises auf die Erteilung des europäischen Patents kann jedermann beim Europäischen Patentamt gegen das erteilte europäische Patent Einspruch einlegen. Der Einspruch ist schriftlich einzureichen und zu begründen. Er gilt erst als eingelegt, wenn die Einspruchsgebühr entrichtet worden ist (Art. 99 (1) Europäisches Patentübereinkommen).

Die Übersetzung ist gemäß Artikel II § 3 Abs. 1 IntPatÜG 1991 vom Patentinhaber eingereicht worden. Sie wurde vom Deutschen Patent- und Markenamt inhaltlich nicht geprüft.

Beschreibung

[0001] Die Erfindung bezieht sich auf eine Vorrichtung für Hochleistungs-Schalten in lokalen Kommunikationsnetzwerken, wie einem Token-Ring, ATM, Ethernet, Fast Ethernet und 1 und 10 Gigabit Ethernetumgebungen, allgemein bekannt als LANs. Insbesondere bezieht sich die Erfindung auf eine neue Schaltstruktur in einer integrierten, modularen Einzelchiplösung, welche auf einem Halbleitersubstrat, wie einem Siliziumchip, implementiert werden kann, und auf eine Vermittlungsanordnung, welche schnelle Kommunikation zwischen den Schaltern ermöglicht.

[0002] Da die Leistung von Computern in den letzten Jahren gestiegen ist, sind die Anforderungen an Computernetzwerke erheblich gestiegen; schnellere Computerprozessoren und höhere Speicherkapazitäten benötigen Netzwerke mit hohen Bandbreiten-Kapazitäten, um Hochgeschwindigkeitsübertragung von erheblichen Datenmengen zu ermöglichen. Die bekannte Ethernettechnologie, welche auf den zahlreichen IEEE Ethernetstandards basiert, ist ein Beispiel von Computernetzwerktechnologie, welche modifiziert und verbessert werden konnte, um als eine brauchbare Rechentechnik fortzubestehen. Eine sehr umfangreiche Erläuterung von Netzwerksystemen des Standes der Technik ist z.B. in „Switched and Fast Ethernet" von Breyer and Riley (Ziff-Davis, 1996) und zahlreichen IEEE Publikationen zu finden, welche sich auf IEEE 802 Standards beziehen. Basierend auf dem offenen Kommunikationssystem (OSI = Open Systems Interconnect) 7-Schicht Referenzmodell sind Netzwerkkapazitäten durch die Entwicklung von Repeatern, Brücken, Routern, und, in letzter Zeit, "Schaltern" gewachsen, welche mit verschiedenartigen Kommunikationsmedien arbeiten. Dickdraht (Thickwire), Dünndraht (Thinwire), verdrilltes Drahtpaar und optische Faser sind Beispiele von Medien, welche für Computernetzwerke verwendet worden sind. Schalter, sofern sie sich auf eine Vernetzung von Computern und auf Ethernet beziehen, sind Hardware-basierte Vorrichtungen, welche den Fluss von Datenpaketen oder Zellen basierend auf der Zieladress-Information steuern, welche in jedem Paket verfügbar ist. Ein korrekt entwickelter und implementierter Schalter sollte geeignet sein, ein Paket zu empfangen und das Paket an einen geeigneten Ausgangsanschluss mit einer sogenannten Leitungsgeschwindigkeit (wire speed oder line speed) zu verschieben, welche das maximale Geschwindigkeits-Leistungsvermögen des speziellen Netzwerkes darstellt.

[0003] Einfache Ethernet Leitungsgeschwindigkeit beträgt bis zu 10 Megabits pro Sekunde und Fast Ethernet beträgt bis zu 100 Megabits pro Sekunde. Das neueste Ethernet wird als 10 Gigabit Ethernet bezeichnet und ist geeignet, Daten über ein Netzwerk mit einer Rate von bis zu 1000 Megabits pro Sekunde zu übertragen. Da die Geschwindigkeit gestiegen ist, sind Konstruktionseinschränkungen und Konstruktionsanforderungen immer komplexer geworden, um geeigneten Konstruktions- und Protokollvorschriften zu entsprechen und eine billige, kommerziell nutzbare Lösung zu liefern. Z.B. erfordert Hochgeschwindigkeits-Schalten Hochgeschwindigkeits-Speicher, um angemessenes Zwischenspeichern von Paketdaten zu bieten; üblicher dynamischer Direktzugriffsspeicher (DRAM = Dynamic Random Access Memory) ist relativ langsam und benötigt Hardware-gesteuertes Auffrischen. Daher führt die Geschwindigkeit von DRAMs, als Pufferspeicher in Netzwerkschaltungen, zum Verlust wertvoller Zeit und es wird fast unmöglich, den Schalter oder das Netzwerk mit Leitungsgeschwindigkeit zu betreiben.

[0004] Darüber hinaus sollte externe CPU-Einbindung vermieden werden, da eine CPU-Einbindung es auch fast unmöglich macht, den Schalter mit Leitungsgeschwindigkeit zu betreiben. Da Netzwerkschalter hinsichtlich der Erfordernis von Regeltabellen und Speichersteuerung immer komplizierter geworden sind, ist zusätzlich eine komplexe Multi-Chip-Lösung notwendig, welche logische Schaltkreise, manchmal als Verknüpfungs-Schaltkreise bezeichnet, erfordert, um die verschiedenen Chips zu befähigen, miteinander zu kommunizieren. Zusätzlich können die Mittel, mit welchen die Elemente miteinander kommunizieren, die Betriebsgeschwindigkeit des Schalters begrenzen, wenn die Elemente gezwungen sind, auf jene Kommunikationen zu warten.

[0005] Bezugnehmend auf das zuvor erläuterte OSI 7-Schichtreferenzmodell, weisen die höheren Schichten typischerweise mehr Information auf. Verschiedene Produkttypen sind für ein Durchführen auf Schaltvorgänge bezogener Funktionen auf verschiedenen Stufen des OSI-Modells verfügbar. Netzknoten oder Repeater arbeiten auf Schicht 1 und kopieren und "übertragen" im Wesentlichen eintreffende Daten an eine Vielzahl von Verbindungen des Netzknotens. Auf Schaltvorgänge bezogene Schicht 2-Vorrichtungen werden üblicherweise als Multiportbrücken bezeichnet und sind geeignet, zwei getrennte Netzwerke zu verbinden. Brücken können eine Weiterleitungs-Regeltabelle darauf basierend erstellen, welche MAC (media access controller = Übertragungsmittelzugriff)-Adressen auf welchen Anschlüssen der Brücke vorhanden sind, und leiten Pakete weiter, welche für eine Adresse bestimmt sind, welche sich auf einer entgegengesetzten Seite der Brücke befindet. Brücken verwenden üblicherweise den als "Spanning tree"-Algorithmus bekannten Algorithmus, um potentielle Daten-

schleifen zu vermeiden; eine Datenschleife ist eine Situation, in welcher ein Paket auf der Suche nach einer spezielle Adresse Endlosschleifen in einem Netzwerk ausführt. Der Spanning Tree-Algorithmus definiert ein Protokoll zur Vermeidung von Datenschleifen. Schicht 3 Schalter, manchmal als Router bezeichnet, können Pakete basierend auf der Ziel-Netzwerkadresse weiterleiten. Schicht 3 Schalter sind geeignet, Adressen zu lernen und Aufstellungen davon zu führen, welche Anschlusszuordnungen entsprechen. Die Prozessgeschwindigkeit von Schicht 3 Schaltern kann verbessert werden, indem spezialisierte Hochleistungs-Hardware verwendet wird und die Haupt-CPU entlastet wird, so dass Anweisungsentscheidungen nicht das Weiterleiten eines Paketes verzögern.

[0006] Zusätzlich spielt die Vermittlungsanordnung auch eine wichtige Rolle für die Betriebsgeschwindigkeiten eines Netzwerkes. Zusammen mit Netzwerkschaltern verwendet, ermöglicht die Anordnung das Bilden von Schalteinheiten mit skalierbaren Anschlussdichten. Die Anordnung empfängt weitergeleitete Daten von Netzwerkschaltern und muss verschiedenartige Daten (d.h. sammelgesendet, einzelgesendet, rundgesendet, etc.) an andere verbundene Netzwerkschalter weiterleiten. Jedoch liefern Vermittlungsanordnungen des Standes der Technik nicht die benötigte Durchlaufleistung und können die gesamte Prozess-Leistungsfähigkeit von verbundenen Netzwerkschaltern begrenzen.

[0007] WO 99 56432 A beschreibt einen Router und ein Verfahren, welches alle Verbindungen oder Leitungsbündel zu einem bestimmten Ziel als ein einzelnes zusammengesetztes Leitungsbündel behandelt. Bei Verwendung von zusammengesetzten Leitungsbündeln würde jeder Verkehr zu einem vorbestimmten Ziel auf dem einzelnen zusammengesetzten Leitungsbündel zu jenem Ziel geleitet werden, anstatt in mehrere getrennte Teile aufgeteilt zu werden.

[0008] US-A-5 991 297 beschreibt ein Verfahren und eine Vorrichtung, welche ermöglichen, dass ein einer Ausgangs-Translations-Tabelle zugeordneter Speicherplatz für jeden Typ von Verbindungskennung innerhalb eines Netzwerkschalters in der Größe unabhängig von zu anderen Verbindungskennungen zugeordnetem Speicherplatz festgelegt wird.

[0009] Die vorliegende Erfindung ist auf eine Schalter-auf-Chip-Lösung für eine Selbstweiterleitende Anordnung gerichtet, welche geeignet ist, Ethernet, Fast Ethernet und 1 und 10 Gigabit Ethernetsysteme zu verwenden, wobei sämtliche Hardware auf einem einzelnen Mikrochip angeordnet ist. Die vorliegende Erfindung ist auch auf Verfahren gerichtet, die eingesetzt werden, um das gewünschte Verarbeiten und Weiterleiten von Daten zu erzielen. Die vorliegende Erfindung ist darauf ausgelegt, die Fähigkeit einer Paket-Weiterleitung mit Leitungsgeschwindigkeit zu maximieren, und auch eine modulare Anordnung bereitzustellen, wobei eine Vielzahl von getrennten Modulen auf einem gemeinsamen Chip angeordnet ist und worin einzelne Designänderungen an bestimmten Modulen nicht die Beziehung jenes bestimmten Moduls zu anderen Modulen in dem System beeinflusst.

[0010] Es ist ein Ziel der vorliegenden Erfindung, ein Verfahren zur Weiterleitung von Daten in einer Netzwerk-Vermittlungsanordnung bei Leitungsgeschwindigkeit bereitzustellen.

[0011] Dieses Ziel wird durch das Verfahren gemäß Anspruch 1 und die Netzwerk-Vermittlungsanordnung gemäß Anspruch 4 erzielt. Vorteilhafte Ausführungsformen der Erfindung sind in den abhängigen Ansprüchen definiert.

[0012] Die vorliegende Erfindung ist auf ein Verfahren zur Weiteleitung von Daten in einer Netzwerk-Vermittlungsanordnung gerichtet. Ein eintreffendes Datenpaket wird an einem ersten Anschluss der Anordnung empfangen und ein erster Teil des Paketes, kleiner als eine volle Paketlänge, wird gelesen, um eine bestimmte Paketinformation zu ermitteln, wobei die bestimmte Paketinformation eine Herkunftsadresse und eine Zieladresse einschließt. Eine Ausgangsanschluss-Bitmap wird basierend auf einer Suche in einer Weiterleitungsaufstellung ermittelt und es wird ermittelt, ob die Zieladresse zu einem Leitungsbündel von gebündelten Anschlüssen gehört. Das eintreffende Datenpaket wird basierend auf der Ausgangsanschluss-Bitmap weitergeleitet, wenn die Zieladresse nicht zu dem Leitungsbündel gehört. Wenn die Zieladresse zu dem Leitungsbündel gehört, wird ein bestimmter gebündelter Anschluss des Leitungsbündels ermittelt und das eintreffende Datenpaket wird dorthin weitergeleitet. Insbesondere kann der bestimmte gebündelte Anschluss des Leitungsbündels dadurch bestimmt werden, dass ein HASH-Wert basierend auf der Herkunftsadresse und der Zieladresse berechnet und der bestimmte gebündelte Anschluss basierend auf dem HASH-Wert ausgewählt wird. Zusätzlich wird auch eine Diensteklasse für das eintreffende Datenpaket aus der bestimmten Paketinformation ermittelt und eine Priorität für ein Weiterleiten wird basierend auf der Diensteklasse gesetzt.

[0013] Die vorliegende Erfindung ist in einem weiteren Ausführungsbeispiel auch auf ein weiteres Verfahren zum Weiterleiten von Daten in einer Netzwerk-Vermittlungsanordnung gerichtet. Ein eintreffendes Datenpaket wird an einem ersten Anschluss der Anordnung empfangen und ein erster Paketteil, weniger als eine volle Paketlänge, wird gelesen, um eine bestimmte Paketinformation zu ermitteln, wobei die bestimmte Paketinformation eine Herkunftsadresse und eine Zieladresse einschließt. Mindestens ein Ausgangsanschluss wird basierend auf einer Suche in einer Weiterleitungsaufstellung ermittelt und eine Diensteklasse wird für das eintreffende Datenpaket basierend auf der bestimmten Paketinformation ermittelt. Daten von dem eintreffenden Datenpaket werden basierend auf dem mindestens einem Ausgangsanschluss und der Diensteklasse einer Warteschlange zugeordnet. Die obigen Schritte werden für weitere eintreffende Datenpakete wiederholt und Daten aus den eintreffenden Datenpaketen werden zu einer Reihe von Warteschlangen zugeordnet und Datenpakete werden sequenziell von jeder Warteschlange aus der Reihe von Warteschlangen weitergeleitet.

[0014] Zusätzlich werden die Daten von dem eintreffenden Datenpaket zu einer Warteschlange zugeordnet, nachdem es in Zellen mit einer bestimmten Zelllänge gepackt werden. Das Verfahren kann auch ein Ermitteln einschließen, ob ein Ziel für das eintreffende Datenpaket ein gebündelter Anschluss ist, wobei aus einer Leitungsbündel-Aufstellung gelesen wird und die Ausgangsanschluss-Bitmap basierend auf Einträgen in der Leitungsbündel-Aufstellung erlangt wird. Die bestimmte Paketinformation kann auch einen Opcode-Wert einschließen, welcher festlegt, ob das eintreffende Datenpaket ein Einzelsende-Paket, ein Sammelsende-Paket, ein Rundsende-Paket ist oder zu einem Fehler bei der Zielsuche führt.

[0015] Die Ziele und Eigenschaften der Erfindung sind unter Bezugnahme auf die folgende Beschreibung und die beigefügten Zeichnungen einfacher zu verstehen, worin:

[0016] Fig. 1 ein Blockdiagramm zeigt, welches ein Ausführungsbeispiel der vorliegenden Erfindung erläutert, welches mit Netzwerk-Schaltern verwendet wird, um eine 64-Anschluss Gigabitlösung zu erzielen;

[0017] Fig. 2 ein Schema darstellt, welches Elemente der Anordnung der vorliegenden Erfindung zeigt;

[0018] Fig. 3 ein Schema darstellt, welches die interne Blockstruktur eines Ausführungsbeispiels der vorliegenden Erfindung zeigt;

[0019] Fig. 4 ein Flussdiagramm für die Eingangslogik für die vorliegenden Erfindung zeigt;

[0020] Fig. 5 ein Unter-Flussdiagramm für die Eingangslogik für die vorliegende Erfindung zeigt;

[0021] Fig. 6 ein Unter-Flussdiagramm für die Eingangslogik für die vorliegende Erfindung zeigt;

[0022] Fig. 7 ein weiteres Flussdiagramm für die Eingangslogik für ein Ausführungsbeispiel der vorliegenden Erfindung zeigt;

[0023] Fig. 8 die Topologie des Eingangs-Busringes erläutert;

[0024] Fig. 9 ein Anschluss-Zu-Anschluss kürzeste Pfad-Karte erläutert;

[0025] Fig. 10 ein Schema zeigt, welches die Warteschlangen-Struktur der Speicherverwaltungseinheit der vorliegenden Erfindung erläutert;

[0026] Fig. 11 ein Berechnen des Blockpausen-Verhaltens erläutert;

[0027] Fig. 12 ein Schema darstellt, welches eine Station des Eingangs-Busringes näher beschreibt;

[0028] Fig. 13 ein Schema der Ringkonnektivität des Eingangs-Busringes darstellt;

[0029] Fig. 14 ein Schema darstellt, welches die verschiedenen Paketgrenzen-Fälle innerhalb einer Zelle des Speichers erläutert;

[0030] Fig. 15 ein Flussdiagramm darstellt, welches ein Wiederherstellungsverfahren bei Speicherverfälschung der vorliegenden Erfindung erläutert;

[0031] Fig. 16 ein Blockdiagramm für den Entpacker der vorliegenden Erfindung darstellt.

[0032] Die vorliegende Erfindung ist auf eine selbst-weiterleitende Anordnung mit 4/8 10 Gigabit Schnittstellen gerichtet. Die externe Bandbreite der Anordnung beträgt gemäß einem Ausführungsbeispiel der vorliegenden Erfindung 80/16 Gbps. Die vorliegende Erfindung ermöglicht das Bilden von skalierbaren Gigabit Anschlussdichten zusammen mit anderen Netzwerk-Schaltvorrichtungen.

[0033] Die vorliegende Erfindung stellt eine Anordnung als Einzel-Chiplösung bereit, welche eine Durchlaufleistung von bis zu 160 Gbps unterstützt. Die vorliegende Erfindung unterstützt 8 Anschlüsse bei 10 Gbps vollständigem Duplexbetrieb und ermöglicht ein Weiterleiten mit voller Leitungsgeschwindigkeit auf jedem Anschluss. Die Anordnung unterstützt auch acht IEEE 802.1p Prioritätsklassen, strenge Priorität und gewichteten, zyklischen Warteschlangenbetrieb. Die Anordnung der vorliegenden Erfindung unterstützt 4096 VLANs für unbekannte Einzelsendungen/Rundsendungen und unterstützt IEEE 802.3x Ablaufsteuerung auf jedem Anschluss. Die vorliegende Erfindung unterstützt auch eine Vorrichtung zur Vermeidung von Zeilenkopf (HOL = Head of Line)-Blockieren auf einem Sendeanschluss und unterstützt Bündeln, Spiegeln und Redundanz. Schließlich stellt die vorliegende Erfindung in einem Ausführungsbeispiel eine 66 MHz, 32-bit PCIX Erweiterungsschnittstelle für CPU und andere PCI konforme Vorrichtungen bereit.

[0034] Fig. 1 erläutert ein Beispiel einer Anordnung der vorliegenden Erfindung, welches eine spezielle Anwendung darstellt. Fig. 1 erläutert ein Beispiel einer 64-Anschluss Gigabitlösung (nicht-blockierend), welche die Anordnung der vorliegenden Erfindung und Netzwerkschalter verwendet. Die allgemeine Struktur der Anordnung der vorliegenden Erfindung ist in Fig. 2 erläutert. Eine bevorzugte Ausführungsform der Anordnung weist acht Anschluss-Schnittstellen mit Datenraten von 10 Gbps und einen internen Ring auf, welcher die Übertragung von Information und Paketdaten zwischen den Anschluss-Schnittstellen ermöglicht.

[0035] Die vorliegende Erfindung stellt eine Hochgeschwindigkeitsanordnung dar, welche bezüglich der Logik von Entscheidungen für ein Weiterleiten von Datenübertragungsblöcken einfach ist. Jedes Paket, welches in die Anordnung eintritt, muss einen Modulkopf aufweisen, welcher Information für Einzelsende-Pakete über das/die Zielmodul(e) enthält, zu welchem/-n ein Datenübertragungsblock weitergeleitet werden muss. Der Modulkopf wird am Eingang von den Netzwerkschaltern vorangestellt.

[0036] In einer bevorzugten Ausführungsform besteht die Anordnung aus 8 Anschlüssen, von denen jeder mit einer Geschwindigkeit von 10 Gigabit arbeitet. An jeden Anschluss wird der Modulkopf untersucht und die Ausgangsanschluss-Bitmap wird darauf basierend ermittelt, ob das Paket ein bekanntes Einzelsende-, ein unbekanntes Einzelsende-, ein Rundsende-, ein Sammelsende-, oder ein IP Sammelsendepaket ist. Die obigen Pakettypen werden im folgenden behandelt.

Einzelsende-Paket

[0037] Wenn ein Datenübertragungsblock von dem Eingang der Anordnung empfangen wird, zeigt der Opcode-Wert von 1 in dem Kopf an, dass das Paket ein Einzelsende-Paket ist, die Ausgangsanschluss- und die Zielmodul-ID (DST_MODID)-Information in dem Modulkopf gültig ist. Die Anordnung leitet das Paket an einen Ausgangsanschluss der Anordnung weiter, welcher den Pfad zu dem Zielmodul darstellt. Alternativ kann es in einigen Ausgestaltungen mehr als einen Pfad zu dem Zielmodul in der Anordnung geben. Daher kann die Anordnung gezwungen sein, einen Ausgangsanschluss basierend auf dem Anordnungs-Eingangs-Anschluss und der Zielmodul-ID wählen. In Ausgestaltungen, in welchen die Zielmodule direkt mit der Anordnung verbunden sind, basiert die Auswahl des Anordnungs-Ausgangsanschlusses auf dem Zielmodul und ist unabhängig von dem Anordnungs-Eingangsanschluss.

[0038] Um ein Weiterleiten eines Datenübertragungsblockes eines Einzelsende-Paketes innerhalb der Anordnung in jeder Ausgestaltungsform zu unterstützen, wird eine einfache Weiterleitungs-Tabelle bereitgestellt. Das Format dieser Tabelle ist wie folgt:

Felder	# Bits	Beschreibung
Zielanschluss-Bitmap	9	Anschluss-Bitmap identifiziert alle Ausgangsanschlüsse, zu welchen das Paket gehen soll

Tabelle 1

[0039] Diese Tabelle ist 32 tief und wenn ein bekanntes Einzelsende-Paket an dem Anordnungseingang eintrifft, wird DST_MODID von dem Modulkopf herausgezogen und in der obigen Tabelle herausgesucht. Die resultierende Bitmap wird verwendet, um die geeigneten Anschlüsse, die zu den Bitfeldern gehören, weiterzuleiten.

Rundsende-Paket/DLF Weiterleiten

[0040] Wenn ein Paket an dem Anordnungseingang mit Opcode-Wert 2 eintrifft, zeigt er an, dass das Paket entweder ein Sammelsende-Paket oder ein unbekanntes (Domain Lookup Failure) Einzelsende-Paket ist. In diesem Fall wird die VLAN ID verwendet, um alle Anschlüsse anzuzeigen, zu denen das Paket geliefert werden soll. Eine Tabelle wird auf jedem Anschluss bereitgestellt:

Felder	# Bits	Beschreibung
Zielanschluss-Bitmap	9	Anschluss-Bitmap identifiziert alle Ausgangsanschlüsse, zu welchen das Paket gehen soll. Basierend auf VID

Tabelle 2

[0041] Die Tabelle ist 4096 Einträge tief und erlaubt alle Werte von VLAN Klassifizierung.

Weiterleiten von Sammelsende-Paketen

[0042] Wenn ein Paket an dem Anordnungseingang mit Opcode-Wert 3 oder 4 eintrifft, ist es eine Sammelsendung (MC = Multicast)- bzw. eine IP MC. Eine Tabelle ist für die Weiterleitung dieser Pakete implementiert. Der Index in dieser Tabelle ist eine Kombination aus der Zielanschluss-ID (DST_PORTID) und der Zielmodul-ID (DST_MODID), welche aus dem Modulkopf herausgezogen werden.

Felder	# Bits	Beschreibung
MC Anschluss-Bitmap	9	Anschluss-Bitmap identifiziert alle Ausgangsanschlüsse, zu welchen das Paket gehen soll. Basierend auf DST_PORT +
		DST_MODID
IPMC Anschluss-Bitmap	9	Anschluss-Bitmap identifiziert alle Ausgangs-Anschlüsse, zu welchen das Paket gehen soll. Basierend auf DST_PORT + DST_MODID

Tabelle 3

[0043] Es existieren acht Kopien aller obigen Tabellen, oder eine pro Anschluss. Die Ausgestaltung der Anordnung der vorliegenden Erfindung schließt ein Ausgangs-Maskenregister (EGRESS_MASK) ein. Dieses Register identifiziert die Gruppe von Anschlüssen, zu welchen das Paket nicht von einem Eingangsanschluss gesendet werden darf. Dieses Register ist 9 Bit breit und es gibt ein Register pro Anschluss.

[0044] Jeder Eingangsanschluss besitzt die folgenden Blöcke: einen Kern für physikalische Übertragung (SerDes), ein 10-Gigabit Ethernet Voll-Duplex MAC und ein Eingangslogik-Block, welcher das Weiterleiten eines Datenübertragungsblockes bestimmt (ING). Jeder Ausgangsanschluss besitzt die folgenden Blöcke: einen

Eingangs-Busring-Knoten (IBR = Ingress Bus Ring); eine Speicherverwaltungseinheit (MMU = Memory Management Unit) und ein Paket-Pool RAM.

[0045] Die vorliegende Erfindung unterstützt auch viele spezifische Merkmale. Die Anordnung unterstützt Verbindungs-Aggregation (Bündeln) ihrer acht 10 Gbps Anschlüsse. Bis zu 4 Leitungsbündel können unterstützt werden, jedes mit bis zu einem Maximum von vier Mitgliedern. Die Leitungsbündel-Tabelle wird verwendet, um den Ausgangsanschluss zu erlangen, wenn ein Paket über einen gebündelten Anschluss hinausgehen muss. Das RTAG wird der Leitungsbündel-Tabelle von der Leitungsbündel-Verteilungs-Logik entnommen, um den Verteilungsalgorithmus zu bestimmen.

Felder	# Bits	Beschreibung
RTAG	2	RTAG identifiziert das Bündel-Auswahlkriterium für dieses Leitungsbündel. 0 - basierend auf DA + SA hash 1 - vollständige Redundanz 2 - reserviert 3 - reserviert
Bündelanschluss-Bitmap	9	Bitmap stellt Anschluss-Elemente in diesem Leitungsbündel dar.

Tabelle 4

[0046] Es gibt vier Kopien der obigen Tabelle, welche vier Leitungsbündel erlauben.

[0047] Die Anordnung der vorliegenden Verbindung unterstützt auch ein dynamisches Abschalten der Anschlüsse für den Fall, dass die Verbindung abbricht. Die Anordnung unterbricht in einem derartigen Fall die CPU. Die CPU ist dann verantwortlich für das Programmieren der EPC Verbindungsregister, um Pakete aus dem deaktivierten Anschluss zurückzuweisen. Ein EPC Verbindungsregister ist eine Bitmap, welche die Anschlüsse darstellt, welche ein gültiges Verbindungssignal aufweisen. Zusätzlich wird auch Spiegelung unterstützt. Das Spiegel-zu-Anschluss-Register zeigt den Spiegel-zu-Anschluss in der Vorrichtung an und spezifiziert das Weiterleiten von Paketen, welche gespiegelt werden sollen. Die Spiegel-Information wird aus dem Modulkopf herausgezogen. Ein CPU-zu-Anschluss-Register wird auch unterstützt.

[0048] Eine Priorität auf COS Warteschlangen-Abbildungsregister (COS_SEL) wird verwendet, um die Priorität eines eintreffenden Paketes oder die erlangte Priorität eines Paketes (nach Adressbestimmung und Filtervorrichtung) auf die Ausgangs-COS-Warteschlange abzubilden. Das Abbilden der Priorität auf die COS Warteschlange wird durchgeführt, nachdem das Paket die Adressbestimmung und die Filtervorrichtung durchlaufen hat, direkt bevor das Paket auf dem CP Kanal gesendet wird. Dieses Abbilden ändert nicht das Prioritätsfeld in dem Kennzeichenkopf des Paketes, es bestimmt nur, aus welcher COS Warteschlange das Paket an den Ausgangsanschluss hinausgehen soll. Der Bedarf für dieses Register entsteht aufgrund empfohlener Abbildungen von Benutzer-Priorität auf eine in dem 802.1p Standard-definierte Verkehrsklasse.

[0049] Zusätzlich sind Konfigurationsregister in der Anordnung verfügbar. Jeder der folgenden besitzt 8 Kopien dieser Register, d.h., eine pro Anschluss. Ein MODE Register wird gesetzt, wenn alle Anschlüsse in der Anordnung im HiGig Modus arbeiten, andernfalls arbeitet die Vorrichtung in einem Modus mit niedriger Bit Rate. Register werden auch unterstützt, um Trigger für sowohl hohe als auch niedrige Wasserpegelstände für die Eingangs-Gegendruck-Flusssteuerung zu liefern. Ein weiteres Register spezifiziert auch den Prioritäts-Warteschlangen-Algorithmus, einschließlich einem strengen Prioritätsmodus, einem gerichteten zyklischen Modus und einem defizitären, zyklischen Modus. Register werden auch bereitgestellt, welche die Prioritätsgewichtungen für die Diensteklassen und das HOL-Blockierlimit für jeden speichern.

[0050] Die folgenden Zähler werden auch Anschluss-bezogen auf einer Send- und Empfangsseite bereitgestellt. Eine Eingangs-Wortzählung liefert die Zahl an Wörtern, die durch den MAC empfangen werden und eine Ausgangs-Wortzählung liefert die Zahl an Wörtern, die in dem Ausgang auf einer COS-Basis gespeichert sind. Zählungen fallengelassener Pakete werden für die Zahl von Paketen bestimmt, die von der Eingangslogik fallengelassen worden sind, und eine Zählung fallengelassener HOL Pakete liefert die Zahl von durch COS fallengelassenen Paketen. Eine Zählung bezüglich der Zahl von aufgrund von Alterung ausgesonderten Paketen

wird auch aufrechterhalten.

[0051] Zusätzlich liefert die Anordnung der vorliegenden Erfindung eine Unterstützung von gesichertem Weiterleiten (Assured Forwarding Support). Dieses Merkmal liefert ein begünstigtes Fallenlassen von Paketen in der Anordnung, wenn ein spezielles Bit in den Modulkopf gesetzt ist. Dieses Bit in dem Modulkopf wird von dem Netzwerkschalter gesetzt, wenn eine bestimmte Aussonderungspriorität gesetzt ist. Wenn ein Paket in einer MMU der Anordnung eintrifft, mit gesetztem Bit, wird die Zahl von Paketzeigern für die COS Warteschlange, welche dem Paket zugeordnet ist, mit einem CNGTHRESHOLD Register abgeglichen. Falls die Zahl der Einträge in der COS Warteschlange für den Anschluss den Wert in dem CNGTHRESHOLD Register übersteigt, wird das Paket fallengelassen.

[0052] Andernfalls wird das Paket in der COS Warteschlange in der MMU aufgenommen. Wenn das Paket fallengelassen wird, wird ein CNGDROPCOUNT Zähler aktualisiert. Falls das bestimmte Bit in dem Modulkopf nicht gesetzt ist, werden alle Pakete in der COS Warteschlange aufgenommen, bis der COS Warteschlangen-Schwellwert erreicht ist.

Logischer Fluss über den Eingang

[0053] Der logische Fluss über den Eingang in die Anordnung hinein wird nun erörtert. Fig. 4 zeigt ein Flussdiagramm, welches einen Teil der Logik bereitstellt. In dem ersten Schritt wird die Anschluss-Bitmap initialisiert und der COS wird von dem Modulkopf erlangt. Ein Opcode wird auch aus dem Modulkopf gelesen. Falls das Paket nur gespiegelt wird, wird keine weitere Auswertung des Kopfes benötigt. Andernfalls wird der Pakettyp aus dem Opcode bestimmt, wobei die Anschluss-Bitmap oder eine andere Bitmap festgelegt wird. Die unterstützten Typen schließen ein CPU Paket, welches zu der CPU gesendet wird, eine Einzelsendung, eine Rundsendung, eine Schicht-2-Sammelsendung und eine IP Sammelsendung ein. Sobald die geeigneten Variablen gesetzt sind, führt der logische Fluss zu dem Unter-Flussdiagramm M, außer die Logik schreibt vor, dass das Paket fallengelassen werden soll. In letzterem Fall wird ein Eingangszähler erhöht.

[0054] Der logische Fluss wird in M, Fig. 5, fortgesetzt, in welchem dann, falls das Paket nur gespiegelt wird, ein Register überprüft wird und falls das Paket noch nicht gespiegelt worden ist, wird das Spiegel-zu-Anschluss-Register festgelegt und die Anschluss-Bitmap wird festgelegt, um das Paket zu spiegeln. Wenn der Eingangsanschluss Teil eines Leitungsbündels ist, wird als nächstes die Anschluss-Bitmap dementsprechend festgelegt. Das Bearbeiten von Anschlüssen in einem Leitungsbündel wird speziell in Fig. 6 behandelt.

[0055] Fig. 7 erläutert einen unterschiedlichen Betriebsmodus für die Paketauswertung. Das alternative Verfahren untersucht Bits der Herkunfts- und Zielanschlüsse. In diesem Fall wird eine Vorrichtung-Anschluss-Abbildungstabelle verwendet, um den Schalter-Ausgangsanschluss zu bestimmen. Es wird angemerkt, dass die Betriebsmodi nicht vermischt werden sollten, letzterer Modus sendet jede Rundsendung, Sammelsendung und unbekannte Einzelsendung an alle Anschlüsse und Spiegeln wird in dem letzteren Modus nicht unterstützt.

Aufbau der Speicherverwaltungseinheit

[0056] Als nächstes wird der Aufbau und die Arbeitsweise der Speicherverwaltungseinheit der Anordnung der vorliegenden Erfindung ausführlicher erläutert.

[0057] Die Hauptfunktionen der Anordnung der vorliegenden Erfindung können in verschiedene Bereiche eingeteilt werden. Erstens, die Vermittlungsanordnung-Bandbreite betreffend, nimmt die Anordnung Paketströme von acht Eingangsanschlüssen mit in der Summe maximal 80 Gbps auf. Die Anordnung ermöglicht, dass Pakete an einem geeigneten Ausgang, für Einzelsendungen und Sammelsendungen mit insgesamt 80 Gbps, austreten, und berücksichtigt die Inanspruchnahme von Kapazität durch Eingang/Ausgang. Die Anordnung verwaltet effektiv Rundesendeverkehr und bearbeitet den zusätzlichen Modulkopf, welcher mit jedem Paket kommt. CPU Verkehr kann auch von oder zu jedem Anschluss kommen und weist eine maximale Datenblockrate von ~2 Gbps auf.

[0058] Die Anordnung unterstützt auch 802.1p Warteschlangenbildung. Die Anordnung priorisiert Pakete von COS und unterstützt bis zu 8 Warteschlangen. Die Anordnung unterstützt auch strenge Priorität und faire Warteschlangenbildung mit gewichteter Zuordnung durch Paketzählung. Die Anordnung gewährleistet ferner geeignete Ablaufsteuerung und Kapazitätsverwaltung. Wenn ein bestimmter Eingang einen Kapazitätsgrenzwert übersteigt, wird der verletzende Eingangsanschluss gemeldet. Der MAC sollte einen PAUSE Datenübertragungsblock an seinen Verbindungspartner senden, um den Paketfluss zu stoppen. Wenn ein bestimmter Aus-

gang einen Zell-Kapazitätsgrenzwert überschreitet, tritt ein HOL (Head Of Line = Zeilenkopf) Blockieren auf. Wenn sich ein bestimmter Ausgang in dem HOL Zustand befindet, wird jedes Paket von jedem Anschluss, welches für jenen Ausgang bestimmt ist, fallengelassen. Wenn eine Vorgangs-Warteschlange für eine bestimmte COS für einen bestimmten Ausgang voll wird, tritt sie in einen HOL (Head Of Line = Zeilenkopf) Blockierzustand ein. Sämtliche neue Pakete werden jedem Anschluss, welche für jenes COS/Ausgangs-Paar bestimmt sind, werden gelöscht.

[0059] Zusätzlich stellt das Folgende erweiterte Merkmale dar, welche in bestimmte bevorzugte Ausführungsformen integriert werden können. Die Anordnung unterstützt faire Warteschlangenbildung mit gewichteter Zuordnung durch Bytezählung und sieht Wiederherstellungsmaßnahmen bei Speicherfehlern vor.

Eingangsbus-Ring

[0060] Die Architektur der Anordnung schließt spezielle Teile ein, welche die Fähigkeit der Anordnung verbessern, Pakete weiterzuleiten. Ein erster Teil ist ein Eingangsbus-"Ring". Z.B. kann die Architektur für die MMU ein aufgeteiltes festes Kapazitätsschema sein, in welchem eine lokale Kopie eines RAMs (128 K Bytes) jedem Anschluss zugeordnet ist. Jeder Anschluss ist mit den benachbarten Anschlüssen über eine Sammlung von unidirektionalen Bussen verbunden, welche gewissermaßen einen Ring bilden, welcher alle neun (8 + CPU) Anschlüsse verbindet. Dieser Bus wird fortan Eingangsbus-Ring (IBR = Ingress Bus Ring) genannt. Die Busse sind 64 Bits breit und es gibt einen Bus für jeden Eingang (daher 64·8 = 512 Bits insgesamt). Die Busse sind verbunden und jeder Bus entsteht an dem Ausgang eines Flops eines Anschlusses und endet an dem Eingang eines Flops an seinen benachbarten Anschluss. Dies wird in **Fig. 3** erläutert. Dieses Punkt-zu-Punkt-Schema beschäftigt sich mit vielen technischen Designfragen, die auftreten, falls andererseits ein mehrfach benutzter globaler Speicher mit breiten (512 Bits) Datenleitungen vorliegt. Dieses Schema macht ferner jeden Anschluss zu einer getrennten logischen Einheit, welche wertvoll für die Testbarkeit ist.

[0061] Der IBR überbringt Paketdatenverkehr, welcher von dem Anschluss-Eingangs-Ausgangs-Block (PIE = Port Ingress-Egress) und tritt jeden Taktzyklus 64 Bit breit auf. Die Daten werden sofort auf lokalen Flops auf dem Bus gespeichert. Und bei jedem Takt wird dieses Wort in den Flops an dem benachbarten Anschluss/Anschlüssen gespeichert. Die MMU an jedem Anschluss überwacht dann die Wortströme auf dem Ring und ergreift das Paket, falls sie eine Zielanschluss-Übereinstimmung feststellt. Zielanschluss-Information wird über ein Steuerwort kommuniziert, welches mit dem Paket über ein Seitenband-Bus synchronisiert ist.

[0062] Eine Energieoptimierung, welche vorgenommen werden kann, ist ein Sperren der Datenweitergabe, falls keine Anschlüsse unten auf dem Ring Empfänger dieses Paketes sind. Ferner leitet in einem Ausführungsbeispiel jeder Buskanal das Wort in einer Richtung weiter, was zu einer maximalen Wartezeit von 8 Takten (oder 8 Sprüngen) führt, um den entferntesten Anschluss zu erreichen. Dies kann verbessert werden, wenn die Wörter in entgegengesetzte Richtungen (im Uhrzeigersinn und gegen den Uhrzeigersinn) weitergeleitet werden, sodass der maximale Sprung sich auf 4 erniedrigt. Das folgende in **Fig. 8** gezeigte Diagramm gilt für jeden Anschluss.

[0063] Effektiv gibt es 9 Busse in dem Anordnungschip (9 = 8 Anschlüsse + CPU Anschluss). Wenn jedoch ein Querschnitt zwischen zwei beliebigen Anschlüssen vorgenommen wird, wird Raum für nur 8 Busse benötigt, da für jeden Anschluss n seine Nachbarn n + 4 und n – 4 nicht verbunden sind. Kein Bus ist eine wirklich geschlossene Schleife. Eine kürzester-Weg-Karte von jedem Anschluss zu jedem anderen Anschluss in diesem Schema ist in **Fig. 9** erläutert.

Anschluss-COS Paare

[0064] Ein weiterer Teil der Anordnung befasst sich mit Anschluss-COS Paaren. Pakete treffen als 8 Byte-Wörter ein, aber das RAM hat eine Breite von 80 Byte-Zellen. Diese Inkongruenz wirft einige Verwendungsprobleme auf, besonders, wenn Paketlängen pathologische Fälle, wie Zellgröße + 1 (CELLSIZE + 1) (d.h., 65 Byte-Pakete) sind. Falls ein solcher Zustand fortdauert, wird eine RAM Bandbreite erheblich beeinträchtigt, was zu einer schwierigen 3,6 Lese-/Schreib-Erfordernis pro Takt führt.

[0065] Um dieses Problem zu lösen, werden Pakete zwischen RAM Grenzen gepackt und entpackt. Wörter werden in einer Registerdatei (oder SRAM) FIFO gesammelt, bis eine Zellgröße vor einem Schreiben fertig ist. Zellen werden aus einem Speicher gelesen, in ein FIFO eingebracht und dann langsam an den PIE als Wörter gegeben.

[0066] Dies jedoch wirft ein weiteres Problem auf, um SAP-zu-SAP Sequenzialisierung aufrechtzuerhalten, müssen Pakete an sehr spezifischen Plätzen im RAM angeordnet werden, sodass sein Eingang-zu-Ausgang-Pfad nicht durch Pakete von anderen Anschlüssen und Diensteklassen verfälscht wird. Daher wird eine Verwendung eines "Anschluss-COS" Paares eingeführt. Ein Anschluss-COS Paar besteht aus zwei Zahlen P:C. P bestimmt, von welchem Anschluss das Paket kam, und C bestimmt, welcher Diensteklasse dieses Paket angehört. Jedem Paket, welches in das Anordnungssystem eintritt, wird eine P:C Kennzeichnung gegeben (obwohl dies vielleicht nicht von jedem Speicherelement dargestellt wird), und sie müssen einem bestimmten P:C Strom in dem System folgen. In Bezug auf dieses System, P = 9 – 1 = 8, da keine Pakete für seinen eigenen Anschluss bestimmt sind, C = 4 und P·C – 32. Daher können bis zu 32 Ströme in der Anordnung auftreten. Für jeden Anschluss existieren logisch 8 Pack FIFOs, 8 Warteschlangen im Speicher und 8 Entpack-FIFOs.

Mitteilungsring

[0067] Ein weiterer Teil der Anordnung ist ein Mitteilungsring (MR = Message Ring). Um alle Stationen zur Steuerung zu verbinden, wird ein weiterer Ring, welcher unidirektional ist, an jede Station "gefloppt" ist und als eine geschlossene Schleife ausgebildet ist. Dieser Ring bietet ein allgemeines Verfahren der Mitteilungsübermittlung zwischen Stationen (einschließlich der CPU). Gegenwärtig dient er zwei Hauptzwecken: CPU Register/Speicher-konformer Zugriff zwischen allen Stationen und Berichten zwischen Eingangs- und Ausgangsstationen.

[0068] Das Mitteilungsübertragungs-Protokoll auf dem MR kann analog zu jenem eines IEEE 802.5 Token-Ring-Netzes, oder eines ISO 8802.7 geteilten-Ring-Netzes. Register-/Zähler-Lese-/Schreibvorgänge sowie Speicheranfragen und Bewilligungen auf dem MR werden unter Verwendung dieses Protokolls weitergeleitet. Es gibt zwei Anforderungen bei der Wahl eines Protokolls, es muss 1) der im ungünstigsten Fall benötigten Bandbreite genügen und 2) das Protokoll selbst muss robust und deterministisch (testbar) sein und niemals ein Blockieren auslösen.

[0069] Die Bandbreite im ungünstigsten Fall ist gegenwärtig durch Stationen-übergreifendes Berichten vorherbestimmt. Stationen-übergreifendes Berichten ist ein Verfahren, mit welchem ein Eingang berechnen kann, wie viele Bytes von jedem Ausgang für alle Pakete ausgesendet worden sind, welche von dem Anschluss aufgenommen wurden. Wenn nicht genügend Bytes gutgeschrieben wurden, wird er in den Gegendruck-(Back Pressure) Zustand eintreten, nachdem der Zähler einen programmierten Wasserpegelstand erreicht.

[0070] In diesem Mechanismus führt jeder Ausgang einen Zähler, welcher verfolgt, wie viele Bytes er für Pakete ausgesendet hat, welche von anderen Anschlüssen kamen. Daher muss er 8 Zähler führen. Nach einem programmierten Schwellwert muss jeder Ausgang diesen Zählwert zurück an den entsprechenden Ursprungseingang berichten. Der MR weist 9 Stationen auf und es wird ein Takt pro Station benötigt. Der ungünstigste Fall beträgt 9·8 = 72 Takte, bevor ein Ausgang all seine Gutschriften eliminieren kann.

Adaptiver, erweiterter Speicher

[0071] Die obige Speicherarchitektur hat jedoch eine Schwäche. Wenn nur 3 Anschlüsse aktiviert sind, sind nur 3·256 KB oder 768 KB für Paketspeicherung verfügbar. Tatsächlich können, wenn der einzige Datenverkehr darin besteht, dass zwei dieser Anschlüsse an den dritten Anschluss senden, nur 256 KB verwendet werden. Die RAMs in dem verbleibenden Teil des Chips sind vergeudet.

[0072] Die adaptive, erweiterte Speicherarchitektur der vorliegenden Erfindung erstreckt sich auf ein Einschließen eines adaptiven Protokolls, um Elastizität bei der Speicherverwendung zu bieten. Um diese Adaption zu übertragen, wird ein Mitteilungsübertragungs-Protokoll verwendet (über dem MR). Ein Anschluss tritt in den PANIC-Modus ein, wenn seine Anzahl freier Zellen einen niedrigen Wasserpegelstand erreicht. In diesem Zustand wird der Anschluss den nächsten verfügbaren Platz auf dem MR einnehmen und sendet eine Speicheranfrage-Mitteilung. Nur deaktivierte Anschlüsse können an der Gewährung von Speicherverwendung an "panische" Anschlüsse teilnehmen. Wenn die ursprüngliche Anfrage zu dem Anfordernden zurückkehrt, nachdem sie die Schleife durchlaufen hat, zeigt sie an, dass entweder kein Anschluss deaktiviert ist oder dass alle deaktivierten Anschlüsse bereits jemand anderem helfen.

[0073] Falls die Anfrage-Mitteilung bearbeitet wird und eine Speichergewährungs-Mitteilung zurücksendet, hört der anfragende Anschluss auf, an ihn gerichtete Pakete anzunehmen. Der gewährende Anschluss beginnt, an dessen Stelle Pakete anzunehmen. Da alle Pakete für alle Anschlüsse sichtbar sind, kann dieser Wechsel von Paket-Eigentum vorgenommen werden, aber nicht ohne Vorsicht. Es gibt verschiedene Probleme

bei einer Weitergabe- und Abgabe-Zeitsteuerung, die beachtet werden müssen.

[0074] Als Beispiel sind drei aktive Anschlüsse zu betrachten: Mit 0,4 und 8 nummeriert, und 5 ungenutzte Anschlüsse, nummeriert mit 1–3 und 5–7. Jeder aktive Anschluss benützt einen erweiterten Speicher von mehr als 256 KB Speicher. Hilfs-MMUs, wie in Anschlüssen 5, 6 und 7, akzeptieren und speichern Pakete im Auftrag von Anschluss 4, wodurch sie Anschluss 4 eine effektive Speicherverwendung von 1 Mbytes gestatten. Jede Hilfs-MMU muss Paar-COS Ströme aufrechterhalten sowie den in dem System vorhandenen Prioritätsalgorithmen folgen.

[0075] Pakete werden von Anschluss 4 abfließen, sobald seine freie-Zellen-Zählung einen niedrigen Wasserpegelstand erreicht, und er wird eine Trenn-Anfragemitteilung für seinen Hilfsanschluss aktivieren. Dann wird Anschluss 5, der nachgeschaltete Hilfsanschluss, "langsam" seine Ströme in einen Speicher zu Anschluss 4 ableiten. Dieser Effekt pflanzt sich entlang der Linie von Helfern fort. Bis ein Speicher von Anschluss 7 vollständig entleert ist, gibt Anschluss 7 eine Trennversuch-Mitteilung und eine Trenn-Bestätigungsmitteilung aus. Wenn getrennt, ist MMU 7 verfügbar für ein Gewähren von Speicheranfragen von jedem anderen Anschluss, einschließlich Anschluss 4. Wenn eine Hilfs-MMU zugeteilt worden ist, kann sie nur nicht mehr als einem weiteren Anschluss dienen. Jede MMU kann Daten von dem IBR mit 80 Gbps abgreifen. Jede MMU kann Daten mit 10 Gbps ableiten.

[0076] Die Absicht der Architektur ist Flexibilität beim Anschließen und Abtrennen einer beliebigen Anzahl von "Hilfs"-MMUs an jeden Anschluss. Folglich wird eine dynamische Zuordnung von eingebautem Speicher ermöglicht. Die Speicherarchitektur ermöglicht höhere unmittelbare Speicherkapazität pro Anschluss und besseres Puffern.

MMU-Theorie von Arbeitsvorgängen

[0077] Die Theorie eines Arbeitsvorgangs der MMU wird nun erörtert. Mit der beschriebenen MMUWarteschlangen-Architektur wird jedes Paket, welches an der Vermittlungsanordnung eintrifft, pauschal an jeden Anschluss über den IBR übertragen. Die Architektur wird in **Fig. 10** erläutert. Eine Kopie des Pakets wird nur gespeichert, wenn die lokale MMU so entscheidet. Im folgenden werden die lokalen Datenstrukturen beschrieben, wie Pakete gespeichert und weitergeleitet werden.

[0078] Das Pack-FIFO besteht aus 8 einzelnen, den 8 Eingangsanschlüssen zugeordneten RAMs, wodurch eine parallele Ankunft von Paketen ermöglicht wird. Jede RAM enthält Speicherplatz, welcher zwei Zellen tief ist. Zwei Zellen erlauben 20 Wörter oder 160 Bytes an Speicher. Jedes FIFO wird verwendet, um Pakete von dem gleichen Anschluss zu speichern. Wörter werden in dem FIFO gesammelt, bis eine Zelle angesammelt ist und sie dann in einen Speicher geschrieben wird. Der Packvorgang von Wörtern in Zellen ist unabhängig von Paketgrößen. Dies dient der Verringerung des Problems von verschwendeten "Löchern" im Speicher, wenn Paketgrößen von Zellgrenzen abweichen. Es gibt insgesamt 32 logische FIFOs, jeder gehört zu einem einzelnen Anschluss-COS Paar. Dies garantiert eine geordnete Paketlieferung und das korrekte Wiederzusammenfügen von Zellen an dem Ausgang des Speichers.

[0079] Der Paket-Pool-Zuteiler vermittelt die 8 Pack-FIFOs für Schreibzugriff auf den Hauptspeicher (Paket-Pool) in zyklischer Weise. Nur FIFOs mit einer vollständigen fertigen Zelle, oder ein FIFO, welches sich in einem Time-Out befindet (siehe Abschnitt über Time-Out-Mechanismus) wird vollständiger Zugriff erlaubt.

[0080] FreeQ ist ein Zeiger zu der aktuellen freien Zelle (oder Block, dies wird später erörtert werden), in welche eine neue Zelle geschrieben werden kann. Eine freie Warteschlange für alle verfügbaren Speicherzellen wird von dem LLA aufrechterhalten.

[0081] Vorgangs-Warteschlangen (XQ) sind ein Datenfeld, welches 8 Warteschlangen enthält, eine für jede COS. Die Größe von jeder Warteschlange ist programmierbar. Jeder Warteschlangeneintrag zeigt auf dem Kopf eines Paketes im Speicher und der Rest des Paketes wird durch eine Verbindungsliste in dem LLA gehalten. XQ führt einen Zeitstempel, welcher jedem Paket im Speicher eine Zeitgröße zuordnet. Pakete, welche gemäß einem programmierbaren Wert zu "alt" sind, werden fallengelassen. Das XQ hat eine Beschränkung auf 2048 Einträge. Daher kann jeder Ausgang nur bis zu 2048 Pakete speichern (siehe PP).

[0082] Das Verbindungslisten-Feld (LLA = Link List Array) ist ein Datenfeld, welches eine 1 zu 1 Abbildung auf den Paket-Pool-Speicher aufweist. Jeder Abstand in dem Datenfeld entspricht einem Zellplatz in dem Paket-Pool. In dem LLA werden Zeiger zu einer anderen Zelle gespeichert. Das LLA liefert eine komfortable Me-

thode zur Adress-Indirektion, wenn Datenstrukturen manipuliert werden. Das LLA führt n + 2 Verbindungslisten. Worin "n" die Anzahl von aktuell gespeicherten Paketen und die 2 die freie Warteschlange plus einer "Friedhofs"-Warteschlange darstellt. Das LLA hält auch einen Referenzzähler für jede Zelle. Dies ist notwendig, da die Zelle aktiv bleiben muss und nicht zu der freien Liste zurückgeführt wird, bis alle, die sich auf die Zelle beziehen, die Zelle nicht mehr nützen müssen.

[0083] Der Paket-Pool (PP) ist ein 128 Kbyte SRAM, welcher als der Hauptspeicher für Ausgangspakete für jenen Anschluss verwendet wird. Bei einer Breite von 640 Bits weist es 1600 Einträge auf. Die Größe von diesem RAM bestimmt letztendlich, wie viel gespeichert werden kann. Z.B. kann es wegen den XQ Begrenzungen bis zu 2048 Pakete minimaler Größe speichern, aber nur bis zu 82 Pakete maximaler Größe (1518 Bytes) und nur 14 Pakete einer "Jumbo"-Größe (9 Kbytes).

[0084] Die Ausgangsteuerung (EGS = Egress Scheduler) bestimmt das nächste Paket, welches aus dem PIE gesendet werden soll. Es folgt den in dem System programmierten Prioritätsregeln und ruft ein Paket ab, Zelle für Zelle, gemäß der von XQ und LLA gelieferten Information.

[0085] Der Entpacker (UPK = Unpacker) ist ein Zwilling zu den Pack-FIFO, in der Hinsicht, dass er die Inkongruenzen zwischen Wort und Zelle in diesem System auf dem Weg nach außen ausgleicht. Er ist jedoch unterschiedlich, da nur ein Anschluss von ihm zu einem Zeitpunkt lesen muss, mit 1/8 der Geschwindigkeit, somit wird nur ein RAM verwendet.

[0086] Die MMU-Ausführung ist eine reine Paket-Speicher- und Weiterleit-Maschine. Der Bedarf, im Paket nachzusehen, ist beseitigt worden, um das Unterstützen von verschiedenen Protokollen zu erleichtern. Die MMU unterstützt die folgenden Paketformate: Paket minimaler Größe von 64 Bytes, Pakete maximaler Größe von 9 K Bytes, ein Modulkopf und ein Einleitungskopf. Zusätzlich sind Bündelungs- und Spiegelungs-Unterstützung übergangslos, da die MMU nur auf ein Anschluss-Bitmap-Seitenband-Signal, welches auf dem IBR übertragen wird, reagiert.

[0087] Der grundlegende Fluss eines Paketes ist wie folgt: Das erste Wort des Paketes wird auf dem IBR für Anschluss m empfangen, durch den RXSTART für Anschluss m angezeigt und die COS des Paketes wird ermittelt, angezeigt durch das Feld in dem Wortkopfgebiet. Dieses Wort wird in dem Pack-RAM von Anschluss m in ein logisches FIFO gemäß der COS gespeichert. Nachfolgende Wörter werden in das gleiche COS FIFO gespeichert werden, bis RXEND für Anschluss m festgestellt wird.

[0088] Unterdessen, falls eine Zelle (10 Wörter) in eines der COS FIFOs für Anschluss m angesammelt worden ist, ist sie bereit, in das Paket-Pool-RAM zu gehen. Es wird angemerkt, dass alle anderen Anschlüsse das gleiche tun. Daher können eventuell alle 8 Anschlüsse eine Zelle aufweisen, welche bereit ist, zur gleichen Zeit in einen Speicher geschrieben zu werden. Der Paket-Pool-Zuteiler gewährt Schreibvorgänge an ein RAM in zyklischer Weise unter allen 8 Anschlüssen in jedem Takt und da 8 Takte benötigt werden, eine Zelle anzuhäufen, ist die Bandbreite ausreichend. Wenn eine Zelle bereit ist, zu gehen, verwendet der Paket-Pool-Zuteiler den FreeQ Zeiger und schreibt die Zelle in einen Speicher. Eine Verbindungsliste wird für das Paket gebildet (wenn nicht bereits geschehen). Dann wird das LLA mit der neuen freien Warteschlange (Free Queue) und der neuen Paket-Verbindungsliste aktualisiert. Dieser Vorgang wird für jede neue Zelle wiederholt.

[0089] Ein RXEND wird festgestellt und der Zeiger zu dem Zellkopf dieses Paketes wird zu der XQ geschoben, zu welcher er gehört. Die Ausgangssteuerung bemerkt, dass ein Paket in der XQ vorhanden ist, welches gemäß seinem Prioritätsalgorithmus bedient werden muss. Er setzt den UPK in Kenntnis, indem er die COS Nummer übermittelt. Der UPK hält es für bereit, übertragen zu werden, er ruft den Zeiger von dem Anfang des übermittelten COS von der XQ ab und verwendet ihn, um die erste Zelle aus dem Speicher von dem LLA zu lesen. Der UPK setzt die Zelle in das FIFO gemäß dem Anschluss-COS Paar, zu welchem es gehört. TXSTART wird dem PIE bestätigt und bei TXREADY werden Wörter an den PIE zur Übertragung getaktet. Alle Zellen vom RAM werden für jenes Paket bis zu EOP abgefragt (angezeigt durch die Feldgröße von XQ). Die Zeiger für jede Zelle werden von dem LLA geliefert, welches zur selben Zeit die Präferenzzählung für jene Zelle erniedrigt. Falls die neue Zählung 0 erreicht, wird die Zelle zurück in die freie Warteschlange gesetzt. TXEND wird mit dem letzten Wort aktiviert. Entpack-FIFO setzt die Zelle in das FIFO, gemäß dem Anschluss-COS Paar, zu welchem sie gehört.

[0090] Mehrere Szenarien sind möglich. Ein Szenario ist auf mehrere Pakete gerichtet, welche bei Nicht-Zellgrenzen enden. Als ein Beispiel trifft für das gleiche Anschluss-COS ein weiteres Paket, genannt B, direkt nach dem obigen Beispiel ein. Es weist 81 Bytes auf. Direkt danach treffen zwei weitere Pakete, genannt C und D,

ein, beide weisen auch 81 Bytes bzw. N Bytes auf.

[0091] Nach 81 Bytes (oder 10 Wörtern), die von B empfangen werden, wird es einer Zelle in dem PP übermittelt und ein Eintrag in dem LLA wird für es erzeugt. Nach 1 Byte von B und 72 Bytes von C wird eine weitere Zelle gewährt und sie werden zusammen in einen Speicher geschrieben. Ein zugehöriger Eintrag in dem LLA wird modifiziert, um an die in 1 verwendete Zelle zu koppeln. Da RXEND für Paket B empfangen worden ist, wird ein Eintrag für es in einer COS Warteschlange in der XQ geschaffen. EGS entscheidet, dass Paket B übertragen werden soll. Es ruft die erste Zelle aus einem Speicher ab und UPK setzt sie in seinen FIFO. 81 Bytes von B werden übertragen und EGS ruft die nächste Zelle für Paket B ab und setzt sie in das gleiche Anschluss-COS-Entpack-FIFO.

[0092] Darauf wird ein Byte von B übertragen. Nun, während all diesem, vollendet auch Paket C das Eintreffen. C wird als Zellen im Speicher gespeichert und die Einträge werden in der XQ mit einem Abstandswert von 1 gespeichert. Die Referenzzählung für die Zellen, in welchen B + D und C + D sich befinden, beträgt 2. Auf ein Bemerken des neuen Eintrags für C in dem XQ hin kann EGS/UPK den Rest von C (da ein Teil von ihm schon über B gelesen wurde) in das Entpack-FIFO abrufen, wenn der Übertragungsprozess das FIFO auf einen vorbestimmten Grenzwert entleert. C kann nun ausgesendet werden. Zuletzt sind Teile von D in dem Entpack-FIFO, in dem PP RAM und in dem Pack-FIFO übrig.

[0093] Ein zweites Szenario richtet sich auf Auszeit-Mechanismen. Nun wird angenommen, dass Paket D das letzte für jenen Anschluss-COS ist und nicht an einer schönen 80-Byte-Grenze endet. Ein paar Bytes von ihm sitzen in dem Pack-FIFO, einige Zellen in dem PP und 64 Bytes von ihm sitzen in dem Endpack-FIFO. Falls kein Paket an diesem Anschluss-COS nach einer Zeit T_{flush} eintrifft, wird den Inhalten von dem Pack-FIFO grünes Licht gegeben, in das RAM zu gehen. Und es wird eine Zelle im Speicher mit den übrigen Bytes zufällig einnehmen. Und ein Eintrag wird für D in der XQ geschaffen. Dieser "Flush"-Timer-Mechanismus wird verwendet, um stillstehende FIFO-Daten zu vermeiden. Der XQ Eintrag für D besitzt einen Abstand von 2 und sobald für D ein Eintrag in der XQ geschaffen ist, kann EGS dann das Paket aus einem RAM gemäß den früher erörterten Schritten abfragen.

[0094] Falls der Ausgangs-MAC überfüllt ist (d.h. irgendein Strom hoher Bandbreite nimmt den Anschluss ein, oder TXREADY wird zu keinem Zeitpunkt beobachtet), kann das Paket im Speicher festsitzen. Es gibt zwei Handlungsabläufe: 1) Im Falle einer Pack-FIFO-Überfüllung löst T_{flush} einen besonderen Zustand aus und ermöglicht den verbleibenden Bytes von Paket D in einen Speicher geschrieben zu werden. 2) Falls der Anschluss frei ist, wird nach einer Zeit T_{drop} das Paket für zu alt befunden und wird fallengelassen, sowohl von dem PP als auch möglicherweise vom Entpack-FIFO, falls es sich auch dort teilweise befindet. Das Alter des Pakets wird durch sein Zeit-„Tick"-Feld in der XQ bestimmt.

[0095] Ein drittes Szenario umfasst ein Aushungern oder eine Überzeichnung von Anschlüssen. Im Falle einer Überzeichnung oder einer schlechten Verbindung, häufen sich Pakete schnell in dem PP an und wenn ein Schwellwert in der XQ erreicht wird, wird Gegendruck an alle verletzenden Anschlüsse erklärt, um einen Gegendruck-Zustand anzuzeigen. Dies wird durch ein Gutschriftsystem auf dem MR vollzogen. Falls Pakete in dem XQ länger als T_{drop} verbleiben, werden sie fallengelassen.

[0096] Im Allgemeinen wird keinen Paketen ein Eintritt in die XQ erlaubt, falls es unvollständig ist, aufgrund eines Löschens fallengelassen wurde oder aufgrund eines Mangels an Puffern fallengelassen wird. Sobald ein Paket einem Anschluss-COS zugeordnet ist, verlässt es niemals jenen Anschluss-COS-Strom. Dies gilt für die Lebenszeit des Paketes in dem System, ungeachtet dessen, in welchem physikalischen RAM es sich befindet. Da jedes Paket einem Anschluss-COS zugeordnet ist, und jedes Schreiben in einen Speicher von nur einem Anschluss-COS erfolgt, enthält keine Zelle im Speicher zwei Pakete von verschiedenen Anschluss-COSs. Da Pakete eine Mindestgröße von nicht weniger als den 64 Bytes aufweisen müssen, können sich nicht mehr als 3 Pakete in derselben Zelle befinden, angesichts einer 80 Byte Zellgröße. $T_{drop} > T_{flush}$ und somit erfordert kein Paket-Aussonder-Ereignis das Leeren von Entpack-FIFOs.

Mitteilungs-Ring-Protokoll

[0097] Der Mitteilungs-Ring verwendet ein Token-Passing-Protokoll mit einigen Einschränkungen für eine Token-Haltezeit, um eine faire Bandbreitenzuweisung zu garantieren und die maximale Zeit festzulegen, in der einer Station ein Token gewährt wird, wenn sie einen benötigt. Der Ring selbst ist ein 33-Bit Bus. Bits [31:0] enthalten ein 32 Bit Mitteilungswort und Bit [32] ist das Token. Zu jedem beliebigen Zeitpunkt ist kein oder ein Token auf dem Ring. Mitteilungen bestehen aus einem bis drei Wörtern; das erste Wort der Mitteilung be-

schreibt den Mitteilungstyp, welcher auch die Mitteilungslänge einschließt. Ein Token-Bit ist stets nur an das letzte Wort einer Mitteilung angehängt.

[0098] Alle Mitteilungen beginnen mit einem üblichen Format, welches ein erstes Wort einer MR Mitteilung aufweist. Der 6-Bit Opcode spezifiziert den Mitteilungstyp und spezifiziert implizit die Mitteilungslänge. Die 5-Bit Zielstation kommt als nächstes, dann folgt die 5-Bit Ursprungsstation (diejenige, welche die Mitteilung erzeugt) und zuletzt ein 16-Bit mitteilungsabhängiger Teil. Einige Mitteilungen besitzen ein zweites und vielleicht ein drittes 32b Datenwort, welches Dinge wie Speicheradresse, Lesedaten und Schreibdaten enthält.

[0099] Bestimmte Mitteilungen werden als ein aufgeteilter Vorgang behandelt; dies bedeutet, dass eine Anfrage von einer Station erzeugt wird und einige Zeit später, vielleicht nachdem viele andere Mitteilungen auf dem Ring zirkuliert sind, sendet die antwortende Station eine Bestätigungsmitteilung zurück.

Kontierungsblock

[0100] Ein weiterer Teil der MMU ist ein Kontierungsblock (ACT = Accounting Block). Diese Logik nimmt einen Strom von 64b Wörtern mit der Kern-Taktfrequenz von dem MAC an, zusammen mit einiger Seitenband-Information, welche von dem PIE erzeugt wird. Es gibt keine direkte Fähigkeit, den Strom von Wörtern, welcher von dem MAC kommt, anzuhalten. Alle Pakete müssen angenommen werden (obwohl sie vielleicht später aus Kapazitätsmangel fallengelassen werden). Der Block ist auch verantwortlich für ein Überwachen von Ressourcen, welche von Paketen benutzt werden, die an jenem Eingang eintrafen, und für ein Anfragen, dass der MAC in einen PAUSE-Zustand eintritt oder austritt, wie angemessen.

[0101] Der ACT führt einen 16 Bit Zähler, welcher die Zahl von Oktbyte-Wörtern anzeigt, welche ein bestimmter Eingang in die MMU eingebracht hat und vermutlich an Ressourcen in Anspruch nimmt. Der Name des Registers lautet: MMU_REG_IngressWordCount. Es wird auf 0 zurückgesetzt und erhöht sich jedes Mal, wenn der PIE ein gültiges Wort auf dem IBR sendet (wie von dem PIE_mmu_vf bit angezeigt). Während Oktwörter ausgegeben werden oder aus irgendeinem Grund fallengelassen werden, wird die Zählung dieser Oktbyte-Wörter zeitweise an den Eingang über die MR IngressCredit-Mitteilung zurückgesendet und von der Zählung ausstehender Wörter abgezogen.

[0102] Somit steigt und fällt diese Anzahl über die Zeit. Falls die Anzahl zu groß ist, fordert der Eingang den MAC auf, eine PAUSE an seine Verbindungspartner zu senden, um den Datenverkehr, welcher in den Chip eintritt, zu verlangsamen. Wenn die Eingangsrate fällt und mäßiger wird, fordert der ACT den MAC auf, den PAUSE-Zustand zu verlassen. Dieses Verhalten wird in **Fig. 11** gezeigt. Obwohl der MAC ein Anfordern jedes PAUSE-Timer-Wertes von 0x0000 bis 0xFFFF erlaubt, verwendet der ACT Block stets nur zwei Werte: 0x0000 oder 0xFFFF. 0xFFFF wird verwendet, wenn eine PAUSE angefordert wird, und 0x0000 wird verwendet, um anzufragen, dass PAUSE aufgehoben wird. Es ist möglich, dass trotz des PAUSE-Zustands, in den er eingetreten ist, die untere Hysteresegrenze in 64 K Zyklen nicht erreicht wird. In diesem Fall fordert die ACT Vorrichtung den MAC auf, eine weitere PAUSE-Anfrage zu senden, um sicherzustellen, dass der PAUSE-Zustand aufrechterhalten wird. Auch dies ist in **Fig. 11** gezeigt.

Eingangs-Busring

[0103] Das Eingangs-Busring (IBR = Ingress Bus Ring)-Modul ist relativ einfach und weist nur ein paar Zweckbestimmungen auf. Erstens, werden die Eingangsbusse getaktet, bevor die Daten an die nächste Station weitergeleitet werden. Dies vereinfacht die oberste-Stufe-Zeitsteuerung, da der Pfad Punkt zu Punkt von benachbarten Stationen auf dem Ring verläuft. Zweitens, befindet sich der IBR dort, wo das Anschluss-Tauschen stattfindet. D.h., die Eingangsbusse werden um eine Position verschoben, bevor sie über die Ausgangsbusse ausgesendet werden. Dies ermöglicht den Stationen, eine einheitliche, nicht überlappende Busverdrahtung zu besitzen und dennoch eine angrenzende Anordnung auf dem obersten Level aufzuweisen. Drittens implementiert der IBR eine Energieoptimierungsstrategie. Während jedes Wort auf einem Eingangsbus eintrifft, wird seine Ausgangskarte überprüft. Wenn keine nachgeschaltete Station jenes Wort benötigt, wird der Ausgangsbus konstant gehalten, mit Ausnahme von dem Gültig-Bit, welches auf unwahr gesetzt wird.

[0104] Jedes der Bits der 8 Wege auf dem IBR hat eine zugeordnete Bedeutung. Obwohl es 9 Stationen auf dem IBR gibt, existieren nur 8 Wege zu jedem Querschnitt aufgrund der "Flügel"-Topologie des "Ringes". Bezüglich jeder Station sind 4 Stationen vorgeschaltet und 4 sind nachgeschaltet. Jede Station registriert ihre Ausgänge, wodurch vermieden wird, dass so viele Signale den ganzen Weg entlang des Chips in einem Takt gesendet werden müssen. Stattdessen wird es durch den Aufwand ersetzt, eine unterschiedliche Latenz von

einer Station zu jedem anderen Paar von Stationen aufzuweisen. Die verschiedenen Bits jeder Leitung sind exakt die gleiche Information, die von dem PIE Block erzeugt wird. Während einer Energieoptimierung kann eine Station all die Bits eines Busses konstant halten und ein UNWAHR "Gültig"-Bit verbreiten, falls entweder das eintreffende Wort nicht gültig ist, oder die Station feststellt, dass die Ausgangsanschluss-Karte keine nachgeschalteten Ziele aufweist. Jede Station auf dem Ring besitzt acht Eingangsbusse und acht Ausgangsbusse; vier verlaufen im Uhrzeigersinn, vier verlaufen gegen den Uhrzeigersinn.

[0105] Fig. 12 zeigt das Aussehen einer Station, während Fig. 13 zeigt, wie die Anschlüsse jeder Station miteinander verbunden sind. Es wird bemerkt, dass das logische Abbilden der Anschlüsse auf einen Eingangsbus an jeder Station sich verändert, aber die Topologie der Eingänge zu den Ausgängen konstant bleibt. Das bedeutet, dass nur ein Layout notwendig ist.

[0106] Die Fähigkeit, ein einzelnes Layout zu verwenden, ist wichtig für die Erfindung. Diese Topologie bedeutet, dass Stationen, welche auf dem Ring benachbart sind, auf dem physikalischen Chip benachbart sein können und angrenzen können, ohne irgendeinen verlorenen Raum zwischen ihnen, um diese zu verbinden. Jede andere Topologie würde verlorenen Raum zwischen physikalischen Blöcken erfordern, um die Ausgänge eines Blocks mit den geeigneten Eingängen der benachbarten Blöcke zu verbinden. Dies erleichtert auch ein Testen, da jede "Fliese" des IBR gleich ist. Es wird auch angemerkt, dass Anschluss 0 in Fig. 13 in beide Richtungen führt, während die anderen Anschlüsse alle durch die Station führen oder an der Station enden. Dies ist darauf zurückzuführen, dass Station 0 Eingang-0-Daten bezieht. Eine 4-Bit Identifizierung wird jeder Station auf dem Ring gegeben, so dass sie ihre Identität kennt.

Mitteilungsring

[0107] Der Mitteilungsring (MR = Message Ring) beruht auf dem folgenden Protokoll. Anfänglich, im zurückgesetzten Zustand, ist kein Token vorhanden. Nachdem ein paar Zyklen durchlaufen worden sind, prägt Station 0 ein Token aus und sendet es auf den Ring. Dieses Token-Wort zirkuliert weiter, bis eine Station eine Mitteilung senden muss. Eine derartige Station wartet, bis sie ein Token auf ihrem Eingangsbus eintreffen sieht. Da dieses Token mit dem letzten Wort der eintreffenden Mitteilung verbunden ist, leitet die Station Bits [31:0] zu ihrem MR Ausgangsanschluss weiter, entfernt aber das Token-Bit. In dem nächsten Zyklus beginnt die Station, welche gerade das Token aufgenommen hat, sämtliche Mitteilungen auszusenden, welche sie aussenden will, gemäß den unten vermerkten Anforderungen. Wenn die Station das Aussenden von Mitteilungen beendet hat, setzt sie das Token-Bit von ihrem Ausgangsbus auf "1" auf dem letzten Wort der letzten Mitteilung.

[0108] Es gibt drei Mitteilungsklassen: 1) ReadRegister, WriteRegister, ReadMemory, WriteMemory; 2) ReadRegisterAck, WriteRegisterAck, ReadMemoryAck, WriteMemoryAck; und 3) IngressCredit. Nur die Station, welche an die Verbindung zu einer CPU angegliedert ist, kann Typ (1) – Mitteilungen senden. Ferner kann nur eine derartige Mitteilung zu einem beliebigen Zeitpunkt ausstehen. "Ausstehen" bedeutet, dass die Typ (2) – Mitteilung, welche eine Typ (1) – Mitteilung vollendet, noch nicht von dem Sender der Typ (1) – Mitteilung empfangen worden ist. Eine Station sendet eine Typ (2) – Mitteilung nur in Erwiderung auf eine Typ (1) – Mitteilung.

[0109] Während einer Token-Eigentums-Zeit kann nur eine Mitteilung von jeder der drei Klassen gesendet werden. Dies hat die folgenden Konsequenzen. Die mit der CPU verbundene Station kann das Token höchstens vier Zyklen lang halten, da sie eine drei Zyklen-WriteMemory-Befehl und eine ein-Zyklus-IngressCredit-Mitteilung senden kann. Obwohl sie eine Typ (2) – Mitteilung in Erwiderung auf einen Typ (1) – Mitteilung erzeugen kann, geschieht dies nicht in derselben Token-Haltezeit. Andere Stationen halten das Token auch höchstens vier Zyklen lang, da sie eine drei Zyklen-ReadRegisterAck-Mitteilung und eine ein Zyklus-IngressCredit-Mitteilung senden können. Da neun Stationen auf dem Ring sind (mit CPU verbundene Station plus acht XAUI Anschlüsse), benötigt ein Token höchstens 15 Takte, um einen kompletten Kreis auszuführen. Dies ist darin begründet, dass nur eine Typ (1)- und eine Typ (2) – Mitteilung jemals während eines Zyklusses des Tokens erzeugt werden kann; daher benötigen zwei Stationen jeweils vier Zyklen und sieben Stationen benötigen jeweils einen Zyklus.

Packer

[0110] Der Zweck des Packer (PK)-Blockes oder der Packer-Einheit ist das Empfangen eines Stromes von 64-Bit-Wörtern von jeder der acht anderen Stationen. Die Ausgangsanschluss-Karte, welche mit jedem Paket verbunden ist, wird verwendet, um zu bestimmen, welche Pakete von einer bestimmten Station aus dem Ring auszulesen sind. Während die Datenwörter von einem bestimmten Eingang über den IBR eintreffen, wird jeder Strom zu 640b "Zellen" zusammengesetzt. Wenn eine Zelle vollständig ist, wird sie innerhalb von acht Taktzy-

klen zu dem PP (Packet Pool = Paket-Pool) übertragen. Die acht Packeinheiten (eine zu jedem Eingang gehörig) vermitteln untereinander, indem sie eine strenge Priorität verwenden, um Zugriff zu dem PP zu erhalten. Da jede Zelle 10 Wörter enthält und ein Paket minimaler Größe aus wenigstens acht Wörtern bestehen kann, ist es möglich, mehrere Paketfragmente in einer Zelle zu haben.

[0111] Fig. 14 zeigt einige mögliche Fälle, wie Pakete innerhalb einer Zelle angeordnet werden können. Jedes kleine Rechteck innerhalb der Zelle stellt ein 8-Byte-Wort dar. Die Pfeile mit der Beschriftung "A", "B", oder "C" darüber zeigen Pakete. Grauschattierte Rechtecke zeigen nicht verwendete Teile von Zellen; die Gründe hierfür werden später aufgeführt. Die großen Balken zeigen die Grenzen eines Paketes. Es ist zu beachten, dass eine Zelle Fragmente von bis zu drei verschiedenen Paketen enthalten kann und dass eine Zelle höchstens zwei Grenzen zwischen Zellen enthalten kann. Pakete sind nicht notwendigerweise zusammenhängend in einer Zelle, aufgrund der toten Wörter in den grauschattierten Rechtecken.

[0112] Die grauschattierten Rechtecke in Fig. 14 können aus einigen Gründen entstehen: Ein Fall wie #2 kann auftreten, wenn ein Eingang für eine Zeitspanne aufhört, Pakete zu senden; schließlich sendet die PK Vorrichtung trotzdem nur die unvollendete Zelle an die PP-Vorrichtung, um ein Stranden des Paketes "A" in der PK Vorrichtung zu verhindern. Weitere graue Rechtecke können auftreten, wenn die MAC Vorrichtung eine Lösch-Anforderung anzeigt, nachdem das Paket bereits begonnen hat. Anstatt alle Zeiger und dergleichen zurückzuspulen, zeigt die PK Vorrichtung nur jene betroffenen Wörter als tot an. Eine letzte Ursache für grauschattierte Rechtecke tritt auf, wenn die PK Vorrichtung versucht, ein Paket an das LLA zu schreiben und ein oder mehrere Fragmente aufgrund irgendeiner Art von Ressourceneinschränkung nicht erfolgreich geschrieben werden können.

[0113] Die Aufgabe des Packens ist der Erhalt der Bandbreite und die Rate der schmalen IGB Wege an die breiten PP Schnittstellen anzupassen. Falls die PK Vorrichtung nicht mehrere Paketfragmente in einer Zelle erlaubte, könnte eine unglaubliche Ineffizienz bei der Speicherverwendung und Bandbreite auftreten. Zum Beispiel, wenn der Verkehr ausschließlich aus 88-Byte-Paketen bestünde, würde ein Paket zwei gesamte Zellen benötigen, von welchen nur 11 der 20 Wörter belegt wären (55% Nutzung).

Verbindungslisten-Feld

[0114] Der Verbindungslisten-Feld-Block ist das Verbindungslisten-Gehirn der MMU. Er führt die folgenden Funktionen aus: Nimmt Schreib-Anforderungen von dem PK für jedes Paket an, bildet eine Verbindungsliste für jedes Paket, lenkt ein Einfügen seines XQ Eintrags, nimmt Lese-Anforderungen von dem UPK an und gibt Zellen frei, welche nicht mehr von Paketen benötigt werden. Das LLA führt auch eine freie Warteschlangen-Verbindungsliste, führt Referenzzähler für jede Zelle und führt ein Löschen von Paketen aufgrund expliziter oder impliziter Voraussetzungen durch und sendet die gelöschten Verbindungslisten zurück zu der freien Warteschlange.

[0115] Zur Erinnerung, es existieren acht getrennte Fälle, in welchen Pakete (A, B und C) in einer 80-Byte-Zelle sich befinden können (siehe Fig. 14).

	sof0	eof0	sof1	eof1		
1)	0	0	0	0	[A]	- DoMID
2)	0	1	0	0	[A >| xxxxxxxxxx]	- DoEOF
3)	1	0	0	0	[xxxxx | < A]	- DoSOF
4)	1	0	0	1	[xxx | < A > | xxx]	- DoONE
5)	1	0	1	1	[xx | < A > | < B]	- DoONESOF
6)	1	1	0	0	[A > | < B]	- DoEOFSOF
7)	1	1	0	1	[A > | < B > | xx]	- DoEOFONE
8)	1	1	1	1	[A > | < B > | < C]	- DoEOFONESOF

[0116] Die acht Fälle sind mit den 4 Signalen (sof0, sof1, eof0, eof1) von dem PK angemessen kodiert. Durch ein Dekodieren führt das LLA einen bestimmten Vorgang für jede Anweisung durch.

[0117] Es gibt zwei Arten von Löschvorgängen in dem LLA: explizite und implizite. Explizite Löschvorgänge: PK bestätigt/aktiviert ein „Lösch"-Bit am EOF, um ein schlechtes Paket anzuzeigen. LLA löscht demgemäß.

[0118] Implizite Löschvorgänge: PK unternimmt einen Schreibversuch, jedoch, da der vorherige Schreibvorgang bearbeitet wird, wird „voll" wahr. Das LLA hat folglich keinen Platz mehr, um das Paket zu speichern und lässt das Paket fallen. Im nächsten Zyklus muss PK erkennen, was sich ereignet hat. Er sollte die verbleibenden Bytes UND die Bytes, die gerade dem LLA gegeben wurden, löschen und zurückbuchen. Es ist anzumerken, dass der PK niemals einen expliziten Löschvorgang durchführt, auch wenn er ein voll-Signal von dem LLA abfragt, prüft der PK den nächsten Takt, um zu sehen, ob der Versuch erfolgreich war. Dies liegt darin begründet, dass während des vorhergehenden Zyklus ein Zelle freiwerden könnte.

[0119] Es gibt vier Auslöser für den „voll"-Zustand:

1) PP wird voll – kein weiterer freier Zell-Puffer im Speicher
2) COS Klasse in der XQ erreicht Paket-Verwendungs-Grenze
3) COS Klasse in der XQ erreicht Wort-Verwendungs-Grenze
4) XQ Anfrage-FIFO wird in dem LLA Block voll (selten)

[0120] Bedingungen 1), 2) und 4) sind in dem LLA Block implementiert, während 3) in dem XQ Block implementiert ist.

[0121] Wenn ein Löschvorgang, entweder implizit oder explizit, benötigt wird, muss das LLA die durch das verletzende Paket besetzte Verbindungsliste der freien Warteschlange überlassen. Da jede Zelle bis zu drei Pakete besitzen kann, die sich in ihr befinden, erzeugt dies einen ziemlich Resourcen-beanspruchenden Arbeitsvorgang. Im ungünstigsten Fall lautet der Arbeitsvorgang, in welchem dies durchgeführt wird:

1 LLA (Port.Tail) = LLA (FreeHead);
2 FreeHead = LLA (Port.Head);
3 UsedCellCount = UsedCellCount – Port.CellCount;
4 LLARefCnt (Port.Head) = LLARefCnt (PortHead) – 1;
#1,2: Verbindungslisten-Arbeitsvorgang, um die gelöschte Zelle auszuschneiden und in die freie Liste zurück einzusetzen.
#3: Zell-Zählung in dem System aktualisieren.
#4: Referenzzähler für die Kopfzelle des Paketes aktualisieren.

[0122] Da die Arbeitsvorgänge 1, 2, 3 und 4 Resourcenkonflikte verursachen, wird die folgende Logik entwickelt:

1 GraveYardHead = Port.HeadPtrPurge; LLA (Port.Tail) = GraveYardTail;
2 FreeHead = Port.HeadPtrPurge;
3 PurgedCellCount = PurgedCellCount + Port.CellCount;
4 LLARefCnt2 (Port.Head) = 1;
#1) GraveYard – Zeiger speichert eine einzelne Verbindungsliste für alle gelöschten Zellen. Diese Verbindungsliste wird (durch DoIncarnate) während eines UPK Lesevorgangs oder eines freien verfügbaren Zyklus zusammengesetzt. Dies vermeidet die Notwendigkeit, die gelöschte Verbindungsliste zur gleichen Zeit wie den Schreibvorgang neu zu verbinden.
#2) HeadPtrPurge verfolgt in angemessener Weise, wo die eingesetzte Verbindungsliste des Pakets beginnen soll, indem SOF und Löschvorgänge für die SOF Zelle und nachfolgende DoMID Zellen, welche als nächstes an der Reihe sind, betrachtet werden.
#3) PurgedCellCount ist ein getrennter Zähler, der genau das überwacht, was sein Name bezeichnet. Er wird auf einen DoIncarnate Zyklus hin mit UsedCellCount verschmolzen.
#4) LLARefCnt2 ist ein zusätzlicher Referenzzählungsspeicher, welcher auf DoReadCell hin verwendet wird, um die endgültige Zählung für jenen Zellort zu bestimmen. Dies ist nützlich, wenn die HeadPtr Zelle der gelöschten Verbindungsliste auch von einem anderen Paket verwendet wird; daher muss sein FragCount – 1 betragen.

[0123] Mit der obigen Implementierung können gelöschte Zellen unter schwierigen Schreibbedingungen eine Verzögerung der Verfügbarkeit des freien Pools bis zum nächsten freien Zyklus oder Lese-Zyklus darstellen. Da die Takt- und Speicherzugriffsbandbreite großzügig ausgelegt worden ist, steht ein freier Zyklus innerhalb von 8 Ticks bereit.

Wiederherstellen von verfälschtem Speicher

[0124] Um sich gegen mögliche Speicherfehler in einem riskanten 0,13um Prozess und einer RAM Verwendung zu schützen, hat die Anordnung der vorliegenden Erfindung einen Weg für Software entwickelt, Speicherfehler zu erkennen und sich von derartigen Fehlern zu erkennen und sich von derartigen Fehlern zu erholen, um weiter zu arbeiten. Die Eigenschaft einer Speicherwiederherstellung der MMU ist in **Fig. 15** erläutert. Die linke Seite der Figur erläutert die Hardwarezustände und die rechte Seite zeigt das Software-Flussdiagramm. Das Software-Flussdiagramm steuert den Zustandsübergang von Hardware und der Fluss ist horizontal ausgerichtet. Dieses Diagramm zeigt die Fähigkeit der Anordnung, verfälschte Adressen in dem Hauptspeicher dynamisch auszublenden, sowie eine Wiederherstellung durch eine Software-Rückstell-Sequenz. Es ist wichtig zu beachten, dass es zwei Typen von Speicherfehlern gibt, welche das System erkennt: #1ECC Fehler in dem Hauptspeicher-Pool und #2 Paritätsfehler in verschiedenen Nutz-SRAMs. Wie gezeigt, kann #1 dynamisch erkannt und durch Software ausgeblendet werden und #2 kann nur durch eine Software-Rückstell-Sequenz wiederhergestellt werden.

Paket-Pool-Speicher

[0125] Der Paket-Pool (PP)-Speicherblock ist eine Hülle für die Paket-Pool SRAM Makros, welche Paketdaten von dem PK Modul speichern. Nachdem die PK Vorrichtung eine Folge von Wörtern in eine Zelle gepackt hat, wird die Zelle automatisch an eine von dem LLA bestimmte Adresse in den PP geschrieben. Das Paket verbleibt in dem PP, bis der UPK Block alle Paketfragmente von dieser Zelle ausgelesen hat. Es können, abhängig von der Anordnung, 1, 2 oder 3 Paketfragmente in der Zelle vorhanden sein.

[0126] Dieses SRAM unterstützt ein Lesevorgang oder ein Schreibvorgang je Kern-Taktzyklus. Bei maximaler unmittelbarer Auslastung gibt es acht Schreibvorgänge (einen von jedem Eingang) und zwei Lesevorgänge (für Ausgang) je neun Zyklen. Diese maximale Auslastungssituation kann akzeptiert werden bis der PP voll wird. Üblicherweise (und in tragbarer Weise) gibt es jedoch einen Schreibvorgang und zwei Lesevorgänge je neun Zyklen.

Paket-Pool-Steuerung

[0127] Das Paket-Pool-Steuerungsmodul errechnet Fehlertest- und Korrektur-(ECC = Error Checking and Correction) Bits für Schreibdaten von dem PK, überprüft (und korrigiert eventuell) Lesedaten an den UPK und stellt Hauptcomputer-Lese/-Schreib-Zugriff (über das MR) zur Verfügung. ECC Fehler werden aufgezeichnet und gezählt und für den Hauptcomputer verfügbar gemacht, um über den MR zu lesen.

[0128] Zum Schutz gegen mögliche Fehler in dem Paket-Pool-Speicher werden zusätzliche ECC Bits an die Daten angefügt. Aufgrund der extrem breiten Schnittstelle zu dem RAM wäre es unpraktisch, eine einzelne ECC Paritätsgruppe für alle Bits zu besitzen. Stattdessen wird ECC auf einer Grundlage von vier 160 Bit-Wörtern errechnet. Jedes Wort wird von neun ECC Bits geschützt. Dies ist ausreichend, um volle SEC-DED (Single Error Correcddouble error detect = einfache Fehlerkorrektur/doppelte Fehlerdetektion) Abdeckung zu liefern. Zum weiteren Schutz gegen SRAM Fehler wird jede Gruppe von ECC Bits mit der an die Daten angefügten Adresse verrechnet. Dies hilft bei der Erkennung der Fälle, in welchen das SRAM die falsche Adresse gelesen haben könnte.

Vorgangs-Warteschlangen

[0129] Die Vorgangswarteschlangen (XQ = transaction queue) liefert die Auftragsinformation zu den Paketen. Die XQ implementiert eine first in first out-Warteschlange für acht COSes. Im wesentlichen ist der Eintrag ein Zeiger in dem PP, welcher anzeigt, wo das Paket gespeichert ist, zusammen mit einer Anzeige der Größe des Paketes. Diese Information wird von der PK Schnittstelle zu dem Zeitpunkt geliefert, an welchem die Zelle, welche das letzte Wort eines Paketes enthält, in den PP geschrieben wird. Die Information wird in der XQ gespeichert und schließt Felder für ein Tick, eine Paketgröße, einen Abstand, einen Eingangsanschluss# und einen Zeiger ein: Der Zeiger ist der Hauptzeiger zu dem Paket in einem Speicher. Die Eingangsanschlussnummer zeigt an, von welchem Anschluss dieses Paket kam und wird für den UPK verwendet. Der Abstand zeigt an, an welchem Ort in der Zelle dieses Paket tatsächlich beginnt (Folge des PK-Packens). Die Paketgröße unterstützt Bytebasierte gewichtete, faire Warteschlangenabarbeitung und wird auch von dem UPK verwendet. Der Tick ist ein Ersatz für einen Zeitstempel, welcher oben erläutert wurde.

[0130] Die 2K Einträge können in bis zu acht verschiedenen Warteschlangen für verschiedene COS Levels

unterteilt werden. Die Größe von jeder COS Klasse ist über Paketbegrenzung-Register programmierbar; jedoch muss die Summe aller definierten Klassen 2K oder weniger sein. Durch ein Sortieren von Paketen in getrennte Warteschlangen für unterschiedliche COS Klassen wird Paketen mit einer höheren Priorität ermöglicht, vor Paketen mit niedrigerer Priorität gesendet zu werden, auch wenn die Pakete mit niedrigerer Priorität zuerst eintrafen. Während das LLA Modul Daten für die XQ Einträge liefert, liest der Ausgangs-Steuerblock (EGS = Egress Scheduler) die vier ältesten Einträge von jeder der acht COS Klassen, um zu entscheiden, welches Paket als nächstes gesendet wird.

[0131] Die XQ implementiert eine besondere Methode zur Paket-Alterung, welches das Problem eines Speicherns großer Zeitstempel-Vektoren für jedes Paket sowie das Zeilenumbruch-Problem für den Vektorwert verringert. Der 3 Bit Tick-Wert stellt den "Zeitstempel" für ein Paket dar. Jeder Tick stellt eine Zeit dar, die von dem Register für die maximale Ausgangszeit spezifiziert wird, wobei das Register ein 24 Bit Register ist. Die Granularität beträgt 34us und legt fest, wie oft ein Tick auftritt. Der "Tick"-Wert sättigt bei 7 und für jedes Paket zeigt ein Tick-Wert von 7 an, dass das Paket zu alt ist und gelöscht werden wird.

[0132] Zum Beispiel tritt für einen Wert von EgrMaxtime = 24'h1E6928 (= 1,993·10^6 als Dezimalzahl) ein Tick alle 1,993E6·34us = 68 Sekunden auf. Tick sättigt nach 7 Ticks, was 68·7 = 480 s = 8 Minuten sind. Demzufolge werden alle Pakete, welche 8 Minuten oder älter sind, gelöscht.

Ausgangssteuerung

[0133] Während die XQ die Ordnung von Paketen innerhalb einer bestimmten COS Klasse enthält, liegt es in der Verantwortlichkeit der Ausgangssteuerung (EGS = Egress Scheduler), auszuwählen, welche der 8 COS Klassen als nächstes ein Paket senden kann. EGS kann programmiert sein, verschiedene Arten von Warteschlangen-Steuerungsalgorithmen zu aktivieren.

[0134] In einem Ausführungsbeispiel wird ein auf einer strengen Priorität basierender Steuerungsalgorithmus eingesetzt. Bei diesem Algorithmus sendet die Warteschlange mit der höchsten Priorität alle ihre ausstehenden Pakete aus, bevor irgendeine andere Warteschlange mit niedrigerer Priorität die Möglichkeit hierzu bekommt. Wenn die Warteschlange mit der höchsten Priorität leer ist, dann sendet die Warteschlange mit der nächsten niedrigeren Priorität ihre Pakete aus und so weiter. Falls ein Paket in eine beliebige Warteschlange mit höherer Priorität aufgenommen wird, wird das Aussenden des aktuellen Paketes vollendet und die Warteschlange mit höherer Priorität wird bedient. Der Hauptnachteil dieser Vorgehensweise ist ein mögliches Aushungern von Warteschlangen mit niedriger Priorität.

[0135] In einer bevorzugten Ausführungsform wird eine gewichtete zyklische (WRR = Weighted Round Robin) Steuerung eingesetzt. Diese Methode verringert den Nachteil der auf einer strengen Priorität basierenden Steuermethode, indem für alle Warteschlangen eine bestimmte minimale Bandbreite bereitgestellt wird, sodass keine der Warteschlangen ausgehungert wird. In gewisser Hinsicht ist die Bandbreite wirklich ein programmierbarer Parameter in der EGS und wird durch den Schaltereinsatz programmiert.

[0136] Jeder COS wird eine Gewichtung durch ein Register zugeordnet. Diese Gewichtung wird an ein Messregister weitergeleitet, welches auf jedes Paket-Austritts-Ereignis hin sich für jene COS verringert. Wenn alle COS Messvorrichtung Null erreichen, werden die Messvorrichtung mit den programmierten Gewichtungen neu geladen. Ein "Peg" wird behalten, um ein zyklisches Zuteilen zwischen dem acht COSes zu liefern, d.h., jeder Warteschlange wird erlaubt, ein Paket für jede Zuteilungsrunde zu senden, bis sein Gewichtungswert auf Null gesunken ist.

[0137] Wenn kein Paket für die COS, an welchem sich der Peg befindet, verfügbar ist, wird den anderen COS Warteschlangen ermöglicht, um den Platz zu konkurrieren, indem ein Kreis-Prioritätsverfahren verwendet wird, d.h., wenn ein Peg sich bei 2 befindet, dann wird 1 → 0 → 3 in dieser Reihenfolge evaluiert. Falls sich Peg bei 3 befindet, dann wird 2 → 1 → 0 in dieser Reihenfolge evaluiert. COSs, deren Gewichtungen zu dem Zeitpunkt Null betragen, sind nicht teilnahmeberechtigt. Falls jedoch keine weiteren COSs verfügbare Pakete besitzen, wird ihr erlaubt, zu gehen, sodass keine Bandbreite verschwendet wird (dies ist der arbeitssparende Aspekt).

[0138] Es wird angemerkt, dass in dem WRR Modus, obwohl der Zuteiler einer COS X gewähren kann, zu gehen, der tatsächlichen Übertragungslogik erlaubt wird, es vorzuziehen, eine andere COS Warteschlange herausgehen zu lassen. Dies ist tatsächlich erlaubt und beeinträchtigt nicht den internen WRR Betrieb. Jedoch weicht die Entkopplungseigenschaft eines derartigen Betriebs wahrscheinlich von der durch eine Programmierung ursprünglich beabsichtigten Fairness/Gewichtung ab.

[0139] Ein Nachteil von WRR besteht darin, dass es in pathologischen Fällen unfair wird. Wenn zum Beispiel ein Kanal viele lange Pakete maximaler Größe überträgt und ein anderer 64 Byte Pakete überträgt. Die Bandbreite des "mini-gram" Kanals wird beeinträchtigt, wenn Bandbreitenzuordnung auf einer Paketzahl basiert ist. Viele Studien wurden über faires Steuern durchgeführt. Während der theoretisch optimale Warteschlangen-Algorithmus, bekannt als General Processor Model (GPS) in der Umsetzung nicht realisierbar ist, kann eine bessere Nähring mit einem defizitären zyklischen (Deficit Round Robin) Algorithmus durchgeführt werden. Dieser Algorithmus kann in alternativen Ausführungsbeispielen unterstützt werden. Der Algorithmus kommt der min-max Erfordernis von gewichtiger Prioritätssteuerung nahe. Der Algorithmus ist arbeitssparend, d.h., Ressourcen sind nicht untätig, falls ein Paket auf einen Einsatz wartet. Es ist Byte-basiert, was ein genaueres Überwachen von tatsächlichem Datenverkehr-Durchsatz ermöglicht.

[0140] Die "Gewichtung" für jeden Kanal ist relativ zu einem "Maß"-Wert, welchen man dem Algorithmus zuordnet. Tatsächlich ist die Gewichtung von jedem Kanal ein ganzzahliges Vielfaches des Maßwertes. Der Maßwert sollte auf eine angemessene Byte-Länge der Datenverkehrsstruktur gesetzt werden. In der Welt des Ethernet besitzt das Datenverkehrsprofil eine bi-modale Verteilung, welche um eine Paketlänge von ca. 64 Byte und 1500 Byte zentriert ist.

Entpacker

[0141] Der Entpacker (UPK = Unpacker) liest Zellen für von dem EGS ausgewählten Paketen und formatiert sie neu in 64 Bitwörter für den MAC. Ein Blockdiagramm für den Entpacker ist in **Fig. 16** erläutert. Der Entpacker fordert ein neues Paket über ein Signal an und wenn dieses Signal und ein Bereit-Signal beide wahr sind, dann wird ein neuer Satz von Paketinformation von der XQ in dem nächsten Zyklus eintreffen.

[0142] Der Entpacker verwendet die Information von der XQ (Größe, Zeiger, Anschluss, etc.), um eine Sequenz von Leseanforderungen an das LLA für jedes Paket zu erzeugen. Das erste Adress-Lesen für jedes Paket ist der von dem XQ empfangene Zeiger. Nachfolgende Lesevorgänge verwenden den von dem LLA empfangenen Wert. Es ist zu beachten, dass die Schnittstelle ermöglicht, dass LLA Lesevorgänge in aufeinanderfolgenden Zyklen auftreten. Wenn der UPK dies auszuführen hat, aktiviert er ein Signal, welches das LLA veranlasst, aus dem nächsten Zell-Zeiger-Platz anstatt des Zell-Zeigers zu lesen. Dies vereinfacht eine Zeitsteuerung, indem die Notwendigkeit für den UPK beseitigt wird, den Zell-Zeiger aus dem nächsten Zell-Zeiger kombinatorisch zu erzeugen. Es ist anzumerken, dass das LLA UPK Lesevorgänge nach Bedarf blockieren kann.

[0143] Die Lese-Daten von dem PP-Speicher treffen an dem Eingang des Paket-Pool-Steuermoduls mit einer festgelegten Verzögerung (4 Zyklen) nach einer erfolgreichen Leseanforderung an das LLA ein. Die ECC Pipeline innerhalb des Paket-Pool-Steuerungsmoduls benötigt zwei Zyklen, um Fehler aus dem RAM zu überprüfen und eventuell zu korrigieren. Diese zwei "Pipe"-Stufen werden als Puffer von dem UPK Modul verwendet. Die geeigneten Wörter von den Zelldaten auf dem Paket-Pool-Steuermodul werden gebündelt und in das Ausgabe FIFO mit einer Rate von einem Wort pro Zyklus eingegeben.

[0144] Wenn Pakete innerhalb der XQ veralten, werden die Pakete aus dem PP Speicher gelöscht, jedoch nicht an den MAC gesendet. Veraltete Paketinformation wird in den Lösch-Puffer gelegt, sodass ein weiteres Paket hervorgeholt werden kann. Indem die gelöschte Paketinformation in den Lösch-Puffer gelegt wird, kann der UPK weiter nach guten Paketen suchen, wodurch jegliche Unterbrechungen in dem Datenfluss zu dem MAC minimiert werden. Der UPK kann Lesevorgänge für sowohl gute Pakete als auch gelöschte Pakete auf einer Zyklus-zu-Zyklus-Basis ausgeben. Wenn sowohl gute als auch gelöschte Pakete bedient werden, wird den guten Paketen Priorität gegeben. Gelöschte Pakete werden aus dem LLA genauso wie gute Pakete gelesen, mit der Ausnahme, dass ein Lösch-Signal aktiviert wird. Dies veranlasst das LLA, die indizierte Zelle freizugeben, aber zu vermeiden, einen Lesevorgang an den PP Speicher auszugeben (wodurch eine Verfälschung von Daten an den MAC vermieden wird).

[0145] Da die Paket-Pipeline zu dem MAC innerhalb des UPK ziemlich lang ist (bis zu 13 Pakete abhängig von Größe und Anordnung), ist es wahrscheinlich, dass innerhalb des UPK befindliche Pakete gelegentlich veralten. Um diesem Umstand Rechnung zu tragen, wird das Alter von jedem Paket innerhalb des Alter-Puffers behalten. Während Pakete von der XQ eintreffen, wird ihr Alter in dem Alter-Puffer aufgezeichnet (welcher als ein FIFO gestaltet ist). Immer wenn der Eingangszeit-Tick aktiviert wird, werden alle Alter um 1 erhöht (aber sättigen bei 7). Während jedes Paket zu dem MAC gesendet wird, wird sein Alter aus dem Alter-Puffer hervorgeholt. Bei Paketen, deren Alter 7 beträgt, wird ein Fehlersignal auf dem letzten Wort aktiviert.

[0146] Um dem ACT Modul zu ermöglichen, korrekt Pausen an den MAC auszugeben, wenn der PP Speicher

eines Anschlusses voll ist, sendet der UPK Gutschriften an das PK Modul über Signale, welche nach jedem erfolgreichen Lesen an das LLA gesendet werden (sowohl für gute Pakete als auch für jene, welche gelöscht werden). In jedem Zyklus gibt der UPK die Anzahl von veralteten Paketen aus, welche er von der XQ erhalten hat oder welche veralteten, wenn sie an den MAC ausgegeben wurden. Eine Gesamtanzahl von veralteten Paketen wird auch aufrechterhalten.

[0147] Die oben erläuterte Anordnung der Erfindung ist, in einem bevorzugten Ausführungsbeispiel, auf einem Halbleitersubstrat, wie Silizium, mit geeigneten Halbleiterherstellungsverfahren realisiert und basiert auf einer Schaltkreisanordnung, welche für Fachleute, basierend auf den oben erläuterten Ausführungsbeispielen, offensichtlich wäre. Ein Fachmann in Bezug auf Halbleiterdesign und -herstellung wäre in der Lage, die verschiedenen Module, Schnittstellen, Tabellen und Puffer, etc. der vorliegenden Erfindung auf ein einzelnes Halbleitersubstrat zu implementieren, basierend auf der oben erläuterten architektonischen Beschreibung. Es wäre auch innerhalb des Schutzumfangs der Erfindung, die beschriebenen Elemente der Erfindung in einzelnen elektronischen Komponenten zu implementieren, wobei die Vorteile der funktionalen Aspekte der Erfindung genutzt werden, ohne die Vorteile durch die Verwendung eines einzelnen Halbleitersubstrates zu maximieren.

Patentansprüche

1. Verfahren zur Weiterleitung von Daten in einer Netzwerk-Vermittlungsanordnung, umfassend:
– Empfangen eines eingehenden Datenpaketes an einem ersten Anschluss der Anordnung;
– Lesen eines ersten Paketteils, weniger als eine ganze Paketlänge, um bestimmte Paketinformation zu ermitteln, wobei die bestimmte Paketinformation eine Herkunftsadresse und eine Zieladresse einschließt;
– Ermitteln einer Ausgangsanschluss-Bitmap, basierend auf einer Suche in einer Weiterleitungsaufstellung;
– Ermitteln, ob die Zieladresse zu einem Leitungsbündel gebündelter Anschlüsse gehört;
– Weiterleiten des eingehenden Datenpaketes, basierend auf der Ausgangsanschluss-Bitmap, wenn die Zieladresse nicht zu dem Leitungsbündel gehört; und
– Ermitteln eines bestimmten gebündelten Anschlusses des Leitungsbündels, indem ein Hash-Wert, basierend auf der Herkunftsadresse und der Zieladresse, berechnet und der bestimmte gebündelte Anschluss, basierend auf dem Hash-Wert, ausgewählt wird, und Weiterleiten des eingehenden Datenpaketes dorthin, in Übereinstimmung mit dem Löschen alter anderen gebündelten Anschlüsse außer dem bestimmten gebündelten Anschluss des Leitungsbündels aus der Ausgangsanschluss-Bitmap und in Übereinstimmung mit dem Hinzufügen des bestimmten gebündelten Anschlusses zu der Ausgangsanschluss-Bitmap nach dem Ermitteln des bestimmten gebündelten Anschlusses des Leitungsbündels, wenn die Zieladresse zu dem Leitungsbündel gehört.

2. Verfahren gemäß Anspruch 1, wobei die bestimmte Paketinformation einen Operationscodewert einschließt, welcher aufzeigt, ob das eingehende Datenpaket ein Einzelsendepaket, ein Sammelsendepaket oder ein Rundsendepaket ist oder einen Fehler bei der Zielsuche ergeben hat.

3. Verfahren gemäß Anspruch 1, weiter umfassend Ermitteln einer Diensteklasse für das eingehende Datenpaket aus der bestimmten Paketinformation und Setzen einer Priorität für ein Weiterleiten, basierend auf der Diensteklasse.

4. Daten weiterleitende Netzwerk-Vermittlungsanordnung, umfassend:
– Mittel zum Empfangen eines eingehenden Datenpaketes an einem ersten Anschluss der Anordnung;
– Mittel zum Lesen eines ersten Paketteils, weniger als eine ganze Paketlänge, um eine bestimmte Paketinformation zu ermitteln, wobei die Paketinformation eine Herkunftsadresse und eine Zieladresse einschließt;
– Mittel zu Ermitteln einer Ausgangsanschluss-Bitmap, basierend auf einer Suche in einer Weiterleitungsaufstellung;
– Mittel zum Ermitteln, ob die Zieladresse zu einem Leitungsbündel gebündelter Anschlüsse gehört;
– Mittel zum Weiterleiten des eingehenden Datenpaketes, basierend auf der Ausgangsanschluss-Bitmap, wenn die Zieladresse nicht zu dem Leitungsbündel gehört;
– Mittel zum Ermitteln eines bestimmten gebündelten Anschlusses des Leitungsbündels, Mittel umfassend zum Berechnen eines Hash-Wertes, basierend auf der Herkunftsadresse und der Zieladresse, und Mittel zum Auswählen des bestimmten gebündelten Anschlusses, basierend auf dem Hash-Wert; und
– Mittel zum Weiterleiten des eingehenden Datenpaketes dorthin, in Übereinstimmung mit dem Löschen aller anderen gebündelten Anschlüsse außer dem bestimmten gebündelten Anschluss des Leitungsbündels aus der Ausgangsanschluss-Bitmap und in Übereinstimmung mit dem Hinzufügen des bestimmten gebündelten Anschlusses zu der Ausgangsanschluss-Bitmap nach dem Ermitteln des bestimmten gebündelten Anschlus-

ses des Leitungsbündels, wenn die Zieladresse zu dem Leitungsbündel gehört.

5. Netzwerk-Vermittlungsanordnung gemäß Anspruch 4, worin die bestimmte Paketinformation einen Operationscodewert einschließt, welcher aufzeigt, ob das eingehende Datenpaket ein Einzelsendepaket, ein Sammelsendepaket oder ein Rundsendepaket ist oder einen Fehler bei der Zielsuche ergeben hat.

6. Netzwerk-Vermittlungsanordnung gemäß Anspruch 4, ferner Mittel umfassend zum Ermitteln einer Diensteklasse für das eingehende Datenpaket aus der bestimmten Paketinformation und Mittel umfassend zum Setzen einer Priorität für ein Weiterleiten, basierend auf der Diensteklasse.

Es folgen 11 Blatt Zeichnungen

Anhängende Zeichnungen

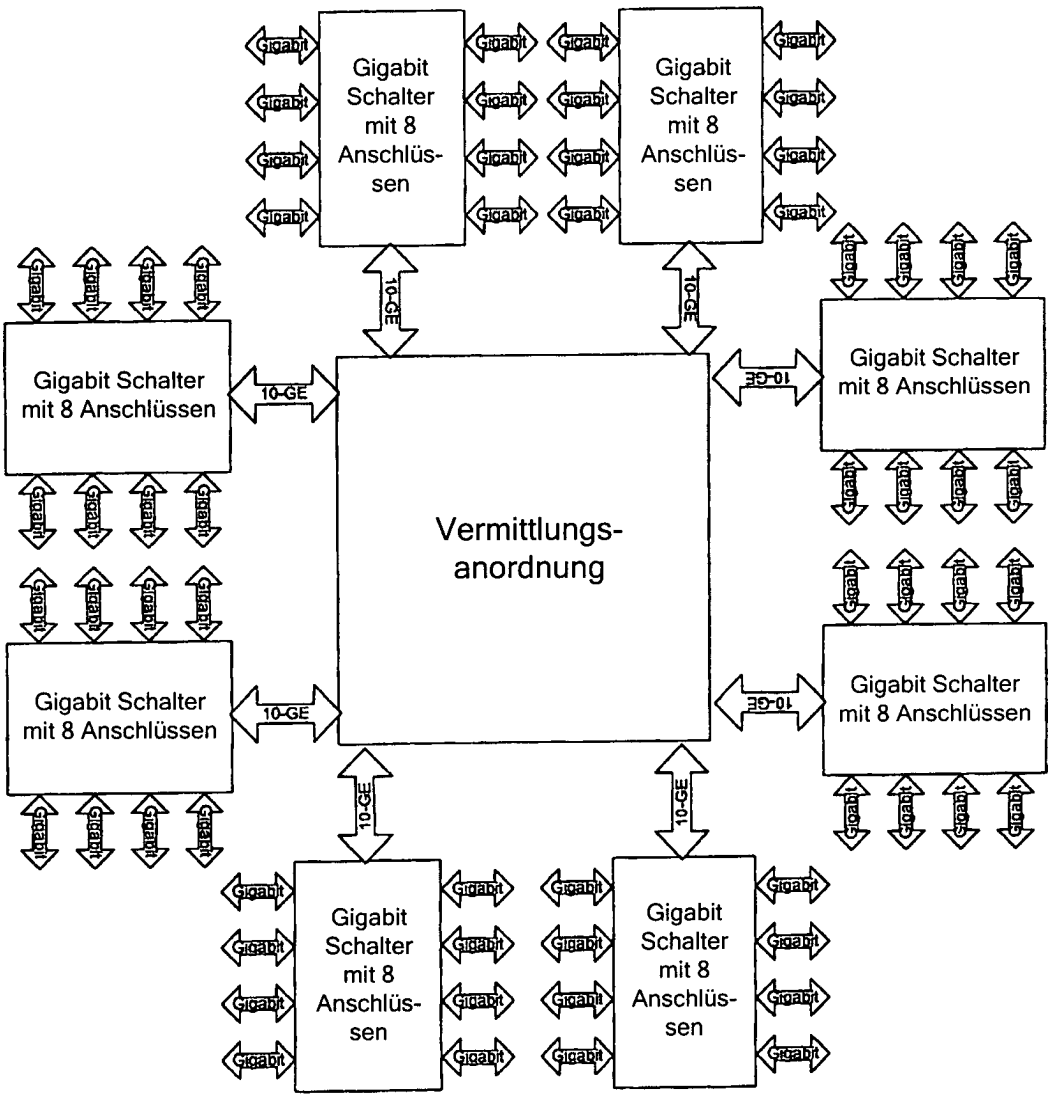

Fig. 1

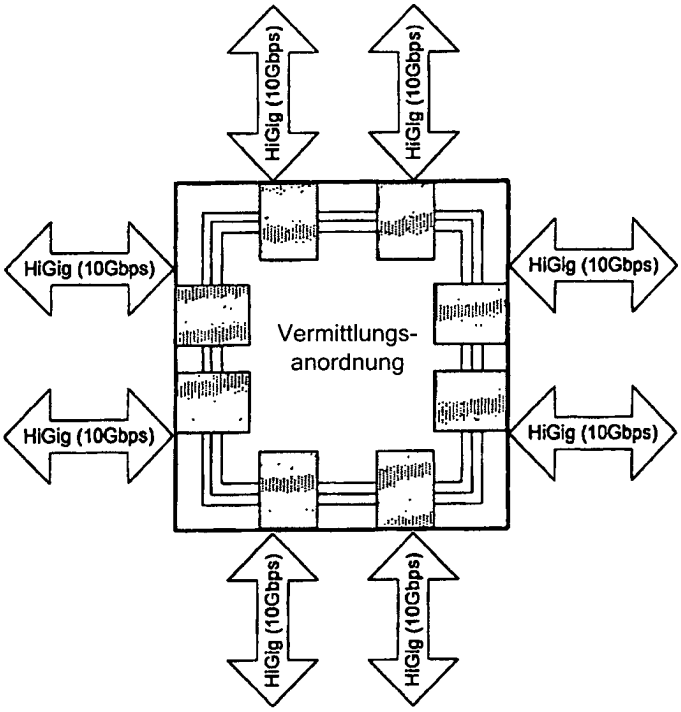
HiGig (10Gbps)
HiGig (10Gbps)
HiGig (10Gbps)
HiGig (10Gbps)
HiGig (10Gbps)
HiGig (10Gbps)
HiGig (10Gbps)
HiGig (10Gbps)
Vermittlungs-
anordnung

Fig. 2

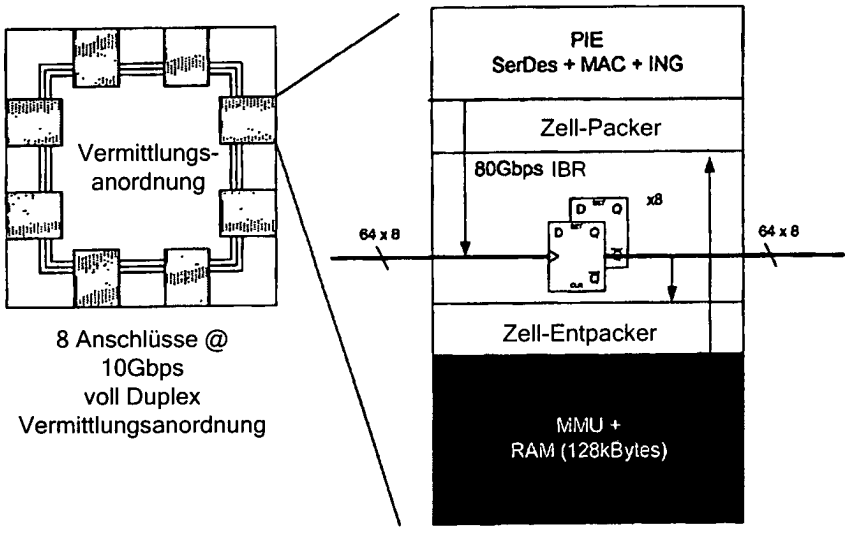
Vermittlungs-
anordnung
8 Anschlüsse @
10Gbps
voll Duplex
Vermittlungsanordnung
PIE
SerDes + MAC + ING
Zell-Packer
80Gbps IBR
x8
64 x 8
64 x 8
Zell-Entpacker
MMU +
RAM (128kBytes)

Fig. 3

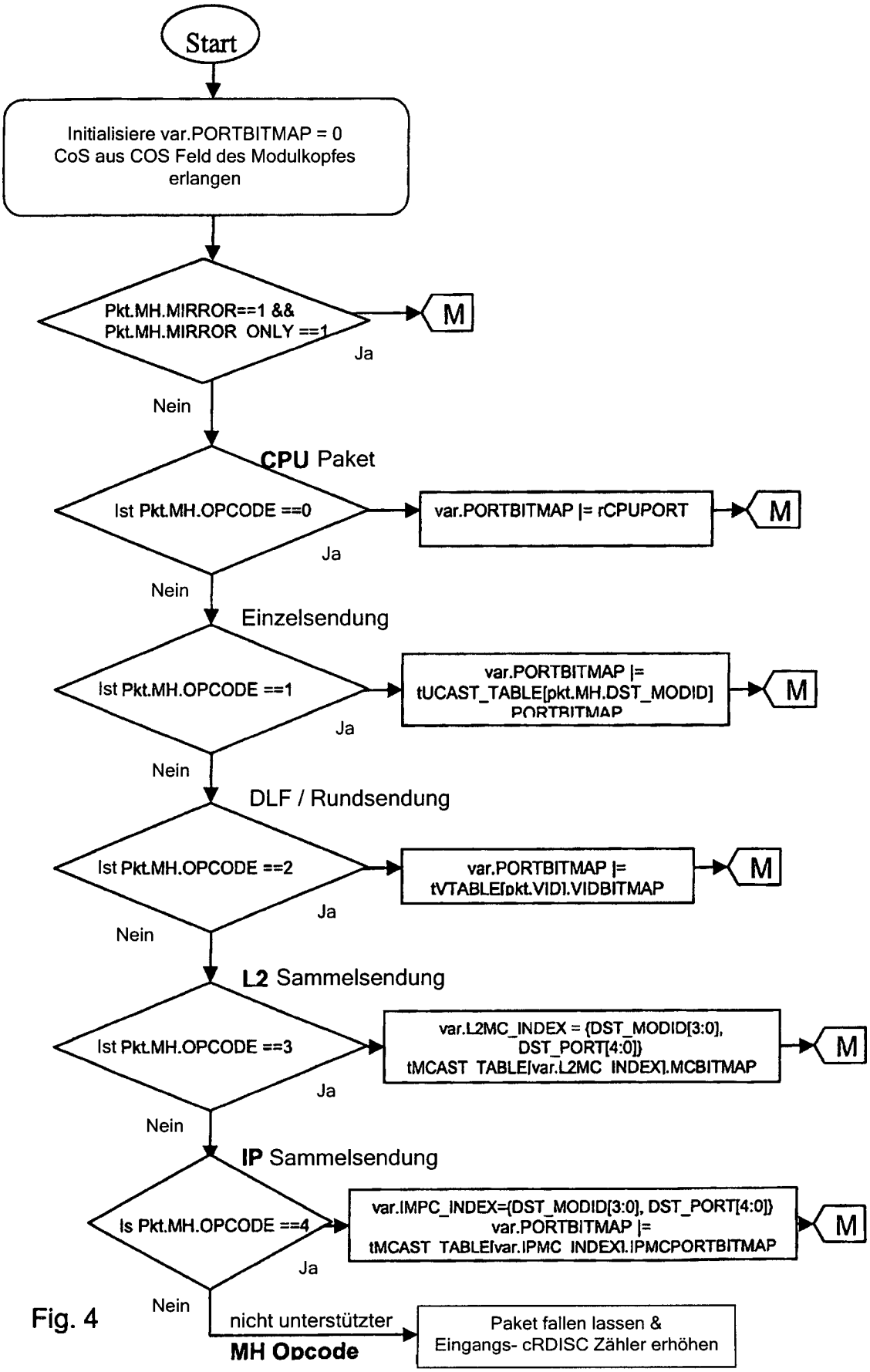
Start
Initialisiere var.PORTBITMAP = 0
CoS aus COS Feld des Modulkopfes erlangen
Pkt.MH.MIRROR==1 &&
Pkt.MH.MIRROR ONLY ==1
Ja
M
Nein
CPU Paket
Ist Pkt.MH.OPCODE ==0
Ja
var.PORTBITMAP |= rCPUPORT
M
Nein
Einzelsendung
Ist Pkt.MH.OPCODE ==1
Ja
var.PORTBITMAP |=
tUCAST_TABLE[pkt.MH.DST_MODID]
PORTBITMAP
M
Nein
DLF / Rundsendung
Ist Pkt.MH.OPCODE ==2
Ja
var.PORTBITMAP |=
tVTABLE[pkt.VID].VIDBITMAP
M
Nein
L2 Sammelsendung
Ist Pkt.MH.OPCODE ==3
Ja
var.L2MC_INDEX = {DST_MODID[3:0],
DST_PORT[4:0]}
tMCAST TABLE[var.L2MC INDEX].MCBITMAP
M
Nein
IP Sammelsendung
Is Pkt.MH.OPCODE ==4
Ja
var.IMPC_INDEX={DST_MODID[3:0], DST_PORT[4:0]}
var.PORTBITMAP |=
tMCAST TABLE[var.IPMC INDEX].IPMCPORTBITMAP
M
Nein
nicht unterstützter
MH Opcode
Paket fallen lassen &
Eingangs- cRDISC Zähler erhöhen

Fig. 4

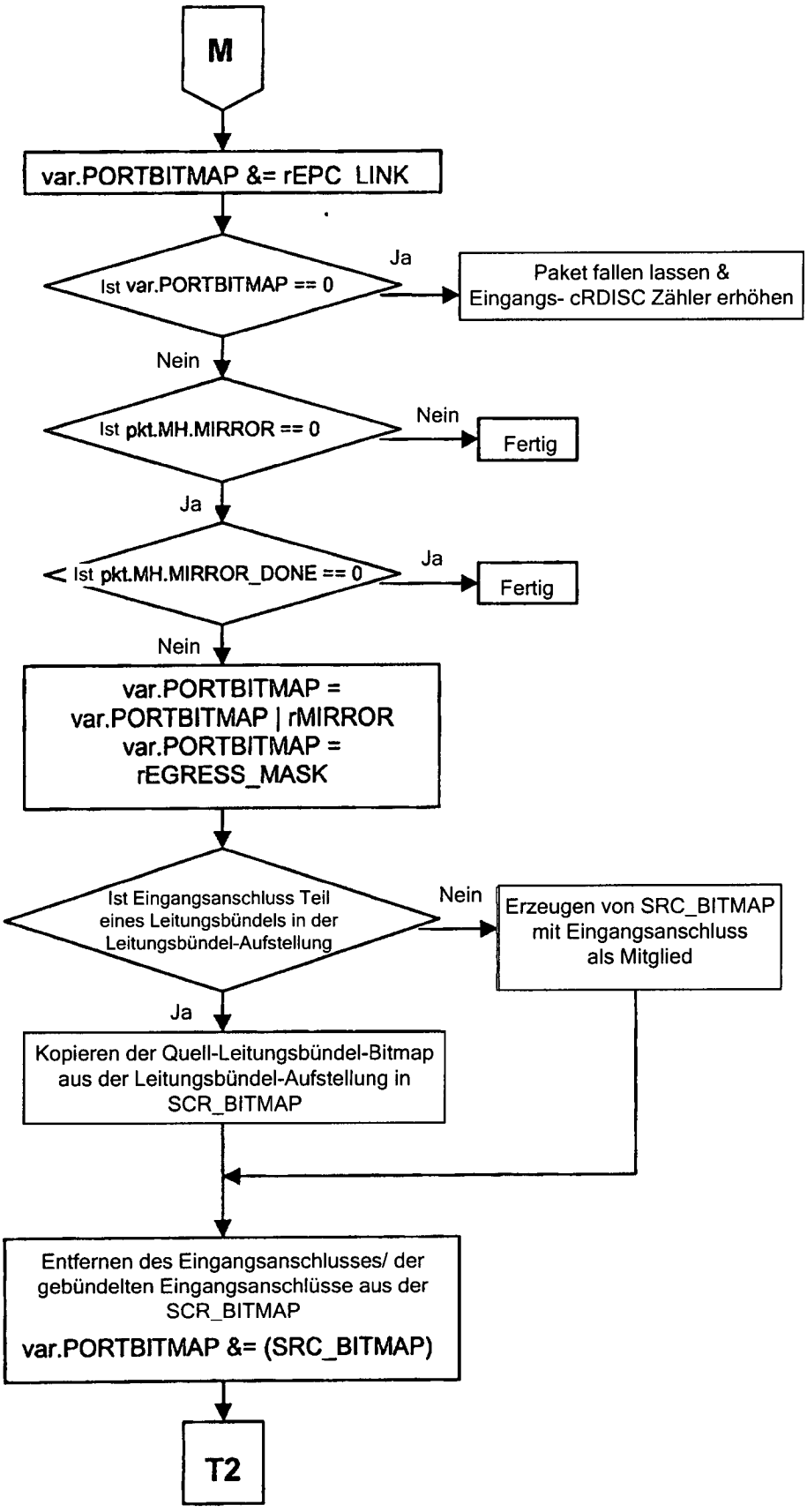
M
var.PORTBITMAP &= rEPC_LINK
Ist var.PORTBITMAP == 0
Ja
Paket fallen lassen & Eingangs- cRDISC Zähler erhöhen
Nein
Ist pkt.MH.MIRROR == 0
Nein
Fertig
Ja
Ist pkt.MH.MIRROR_DONE == 0
Ja
Fertig
Nein
var.PORTBITMAP = var.PORTBITMAP | rMIRROR var.PORTBITMAP = rEGRESS_MASK
Ist Eingangsanschluss Teil eines Leitungsbündels in der Leitungsbündel-Aufstellung
Nein
Erzeugen von SRC_BITMAP mit Eingangsanschluss als Mitglied
Ja
Kopieren der Quell-Leitungsbündel-Bitmap aus der Leitungsbündel-Aufstellung in SCR_BITMAP
Entfernen des Eingangsanschlusses/ der gebündelten Eingangsanschlüsse aus der SCR_BITMAP
var.PORTBITMAP &= (SRC_BITMAP)
T2

Fig. 5

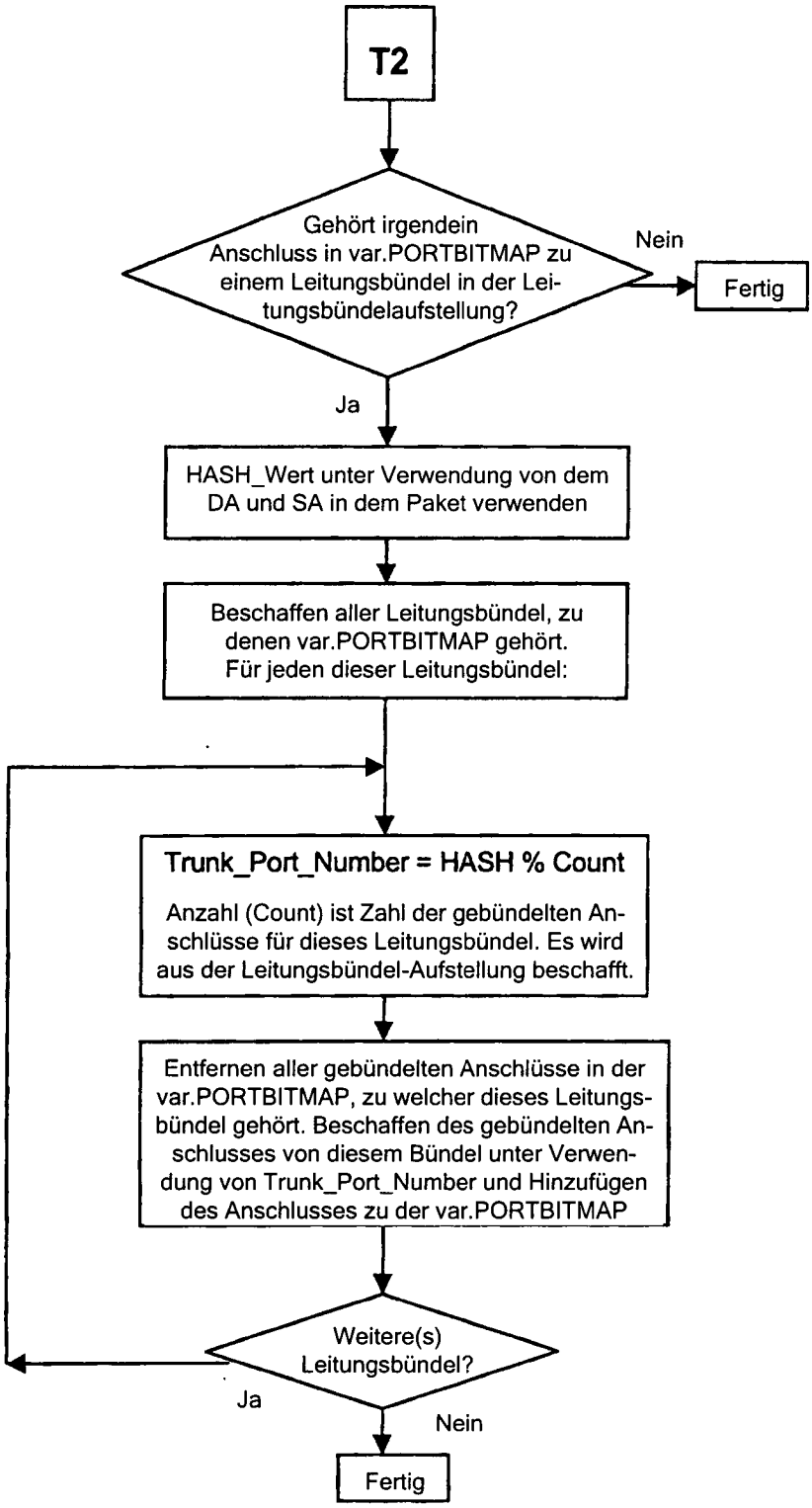
T2
Gehört irgendein Anschluss in var.PORTBITMAP zu einem Leitungsbündel in der Leitungsbündelaufstellung?
Nein
Fertig
Ja
HASH_Wert unter Verwendung von dem DA und SA in dem Paket verwenden
Beschaffen aller Leitungsbündel, zu denen var.PORTBITMAP gehört. Für jeden dieser Leitungsbündel:
Trunk_Port_Number = HASH % Count
Anzahl (Count) ist Zahl der gebündelten Anschlüsse für dieses Leitungsbündel. Es wird aus der Leitungsbündel-Aufstellung beschafft.
Entfernen aller gebündelten Anschlüsse in der var.PORTBITMAP, zu welcher dieses Leitungsbündel gehört. Beschaffen des gebündelten Anschlusses von diesem Bündel unter Verwendung von Trunk_Port_Number und Hinzufügen des Anschlusses zu der var.PORTBITMAP
Weitere(s) Leitungsbündel?
Ja
Nein
Fertig

Fig. 6

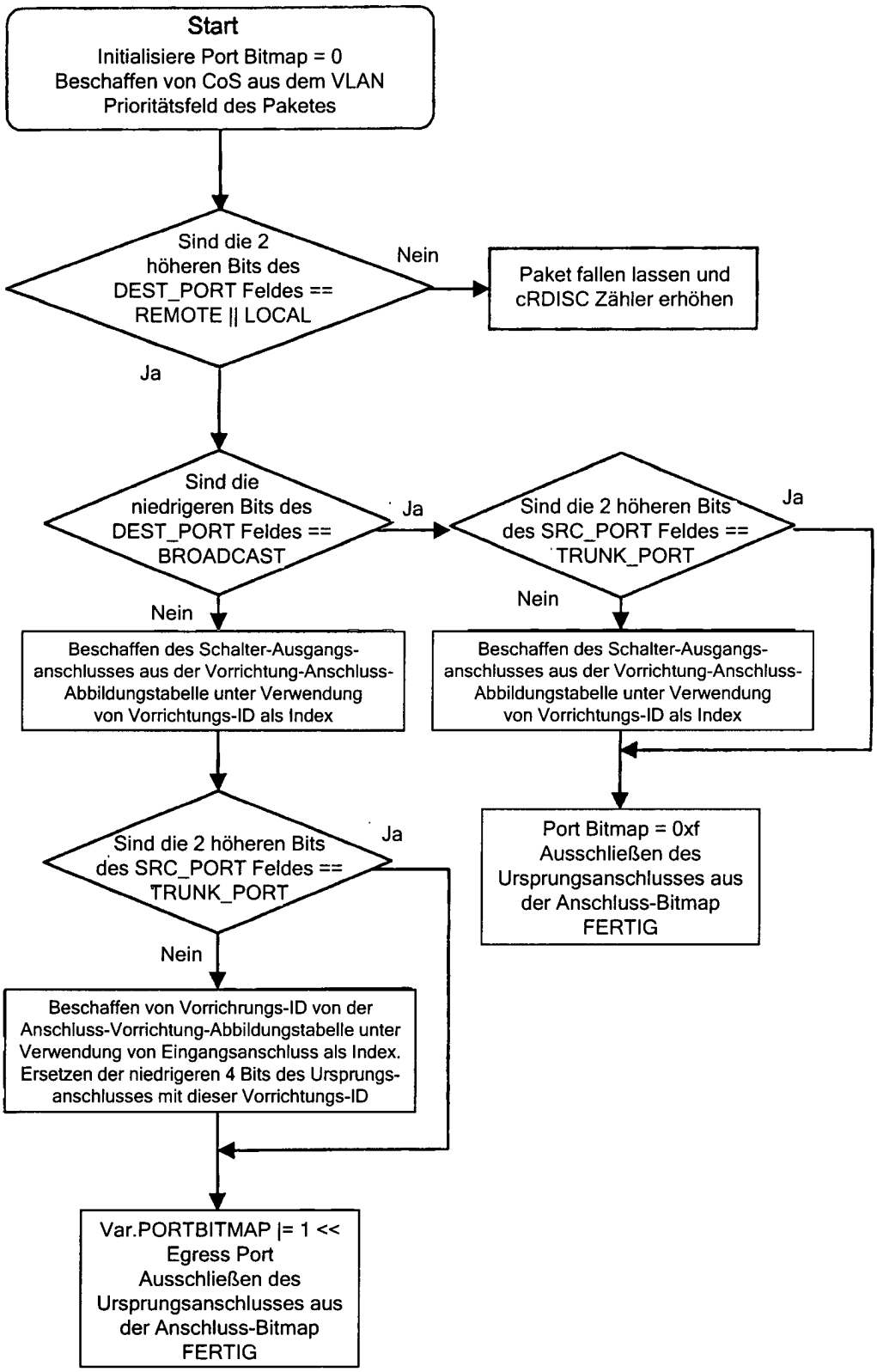
Start
Initialisiere Port Bitmap = 0
Beschaffen von CoS aus dem VLAN Prioritätsfeld des Paketes
Sind die 2 höheren Bits des DEST_PORT Feldes == REMOTE || LOCAL
Nein
Paket fallen lassen und cRDISC Zähler erhöhen
Ja
Sind die niedrigeren Bits des DEST_PORT Feldes == BROADCAST
Ja
Sind die 2 höheren Bits des SRC_PORT Feldes == TRUNK_PORT
Ja
Nein
Nein
Beschaffen des Schalter-Ausgangs-anschlusses aus der Vorrichtung-Anschluss-Abbildungstabelle unter Verwendung von Vorrichtungs-ID als Index
Beschaffen des Schalter-Ausgangs-anschlusses aus der Vorrichtung-Anschluss-Abbildungstabelle unter Verwendung von Vorrichtungs-ID als Index
Sind die 2 höheren Bits des SRC_PORT Feldes == TRUNK_PORT
Ja
Nein
Port Bitmap = 0xf
Ausschließen des Ursprungsanschlusses aus der Anschluss-Bitmap
FERTIG
Beschaffen von Vorrichrungs-ID von der Anschluss-Vorrichtung-Abbildungstabelle unter Verwendung von Eingangsanschluss als Index. Ersetzen der niedrigeren 4 Bits des Ursprungs-anschlusses mit dieser Vorrichtungs-ID
Var.PORTBITMAP |= 1 << Egress Port
Ausschließen des Ursprungsanschlusses aus der Anschluss-Bitmap
FERTIG

Fig. 7

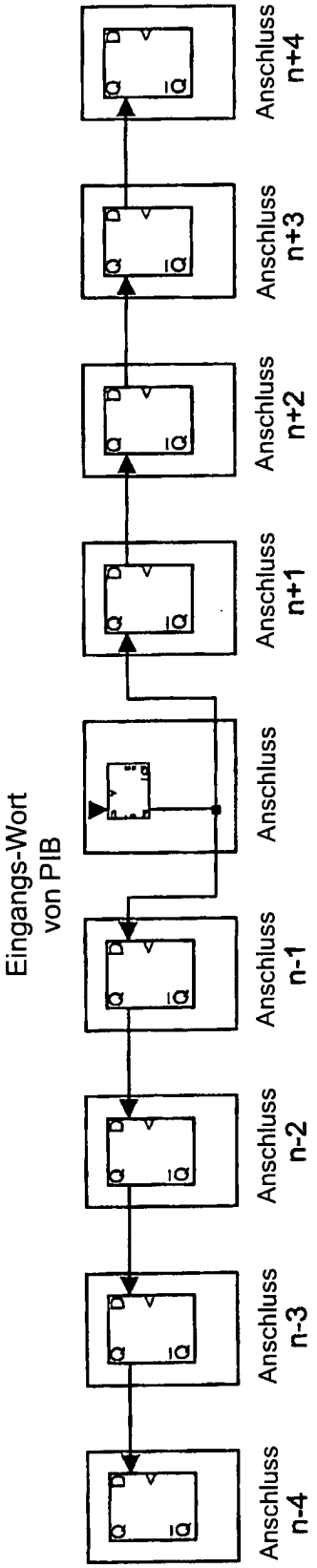
Eingangs-Wort von PIB
Anschluss n-4
Anschluss n-3
Anschluss n-2
Anschluss n-1
Anschluss
Anschluss n+1
Anschluss n+2
Anschluss n+3
Anschluss n+4

Fig. 8

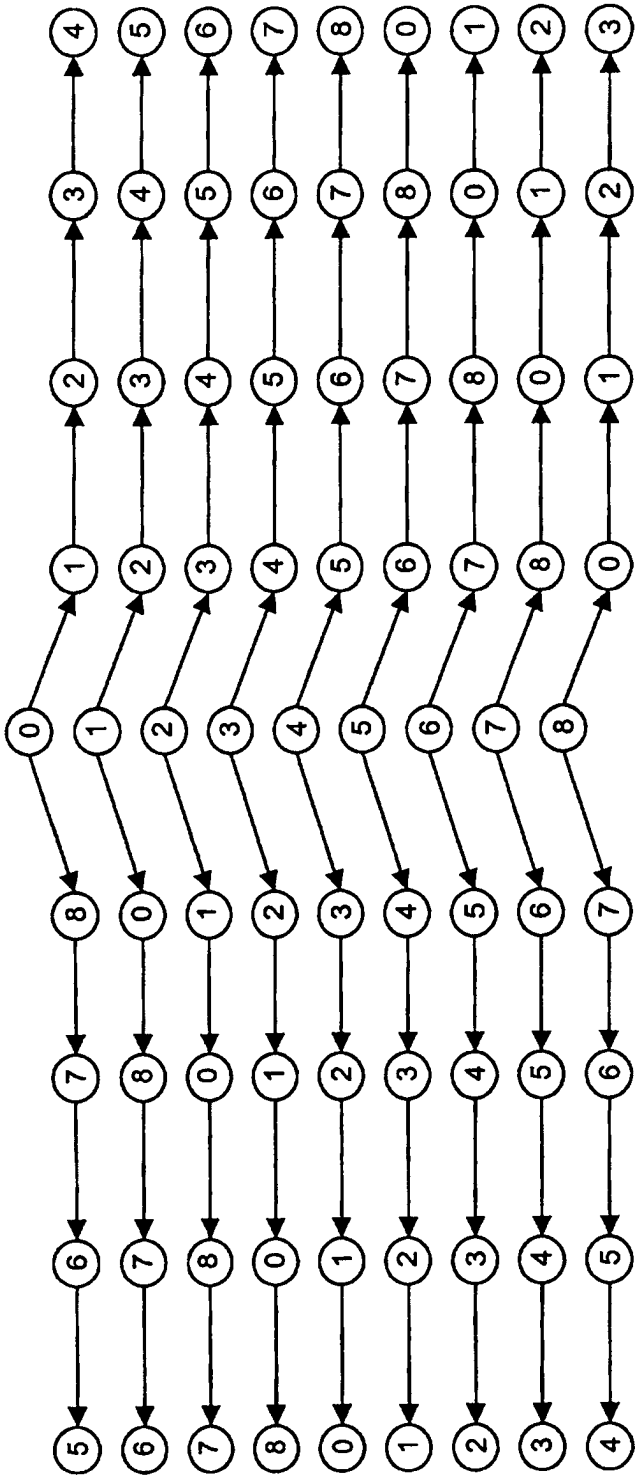

Fig. 9

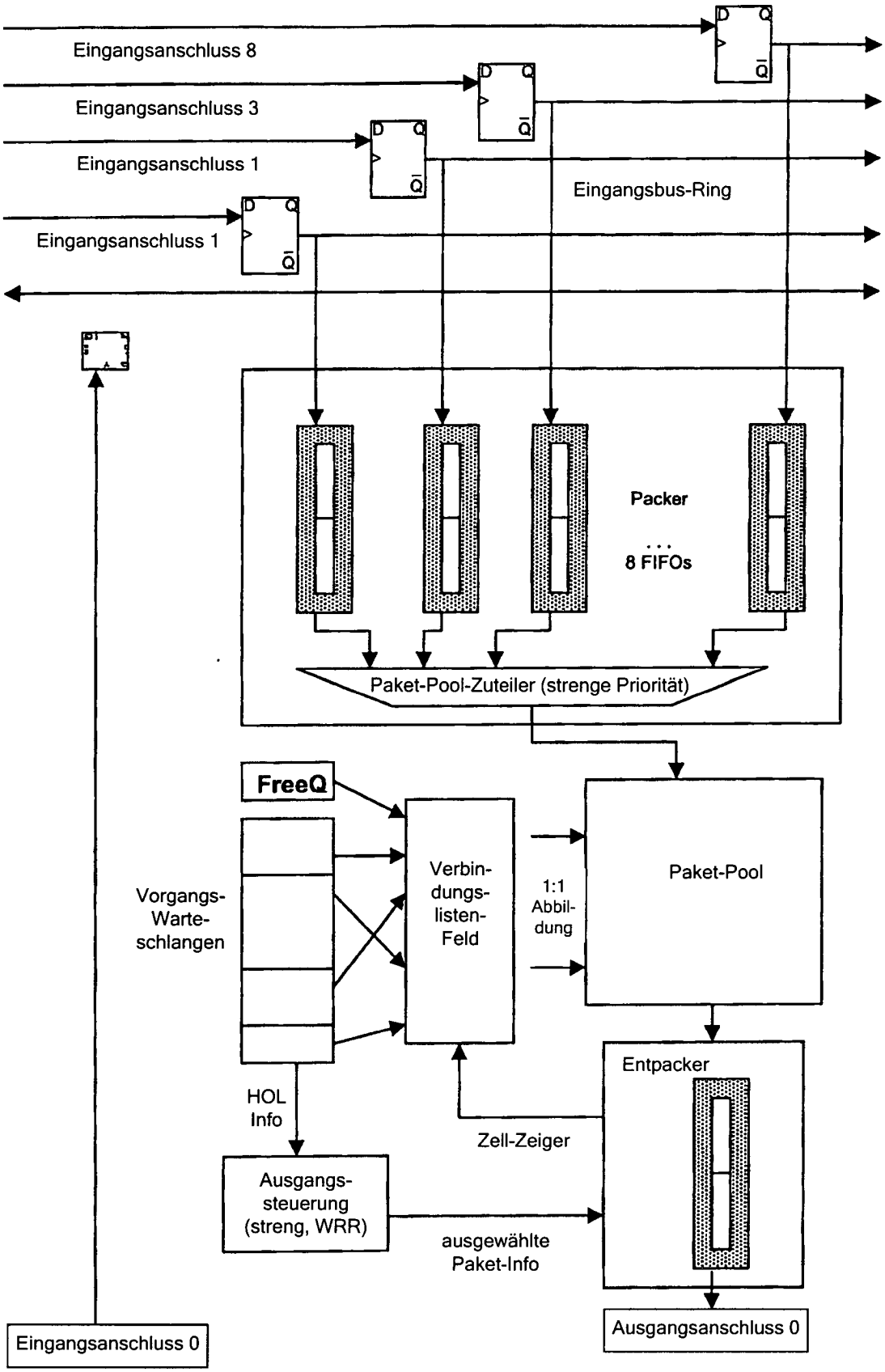
Eingangsanschluss 8
Eingangsanschluss 3
Eingangsanschluss 1
Eingangsanschluss 1
Eingangsbus-Ring
Packer
...
8 FIFOs
Paket-Pool-Zuteiler (strenge Priorität)
FreeQ
Vorgangs-Warte-schlangen
Verbin-dungs-listen-Feld
1:1 Abbil-dung
Paket-Pool
Entpacker
HOL Info
Zell-Zeiger
Ausgangs-steuerung (streng, WRR)
ausgewählte Paket-Info
Eingangsanschluss 0
Ausgangsanschluss 0

Fig. 10

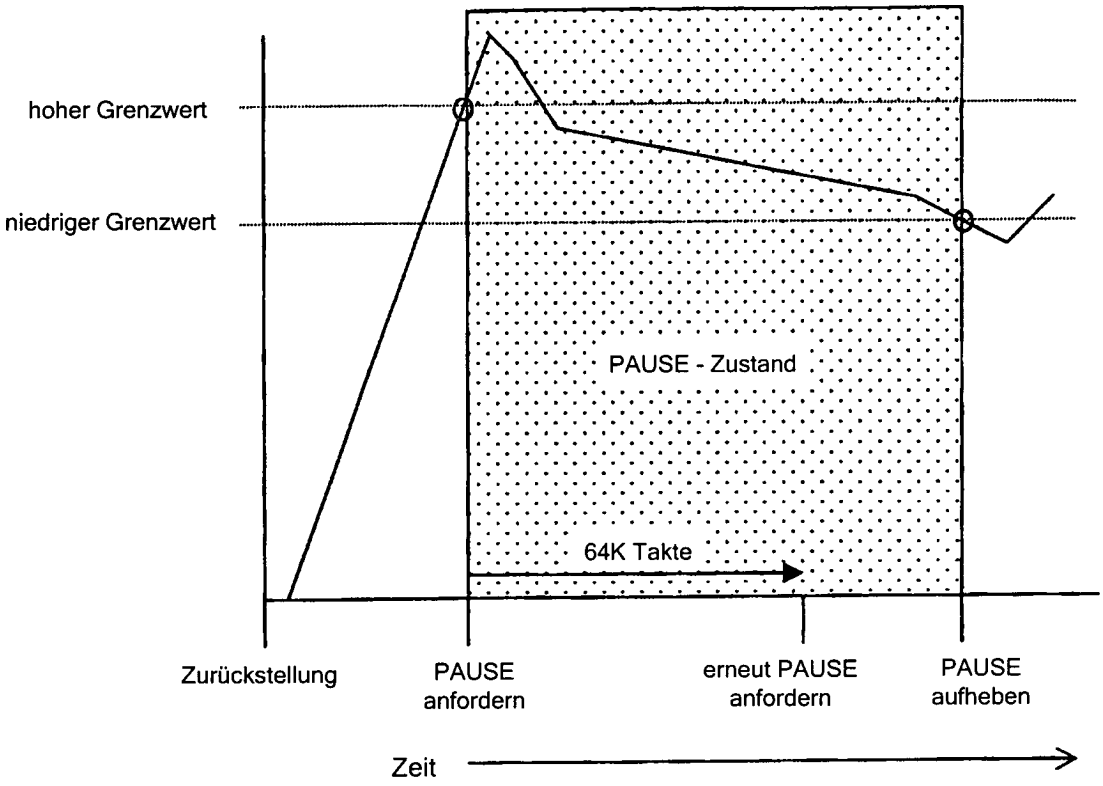

Fig. 11

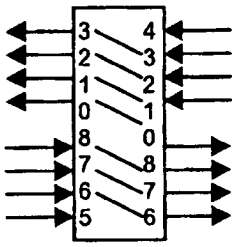

Fig. 12

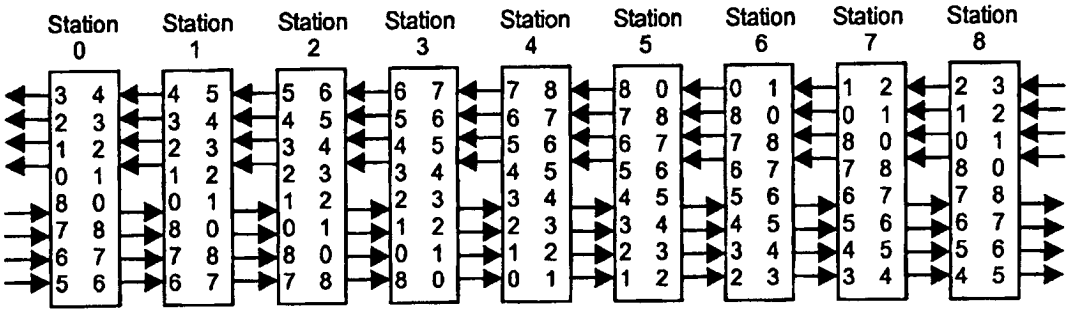

Fig. 13

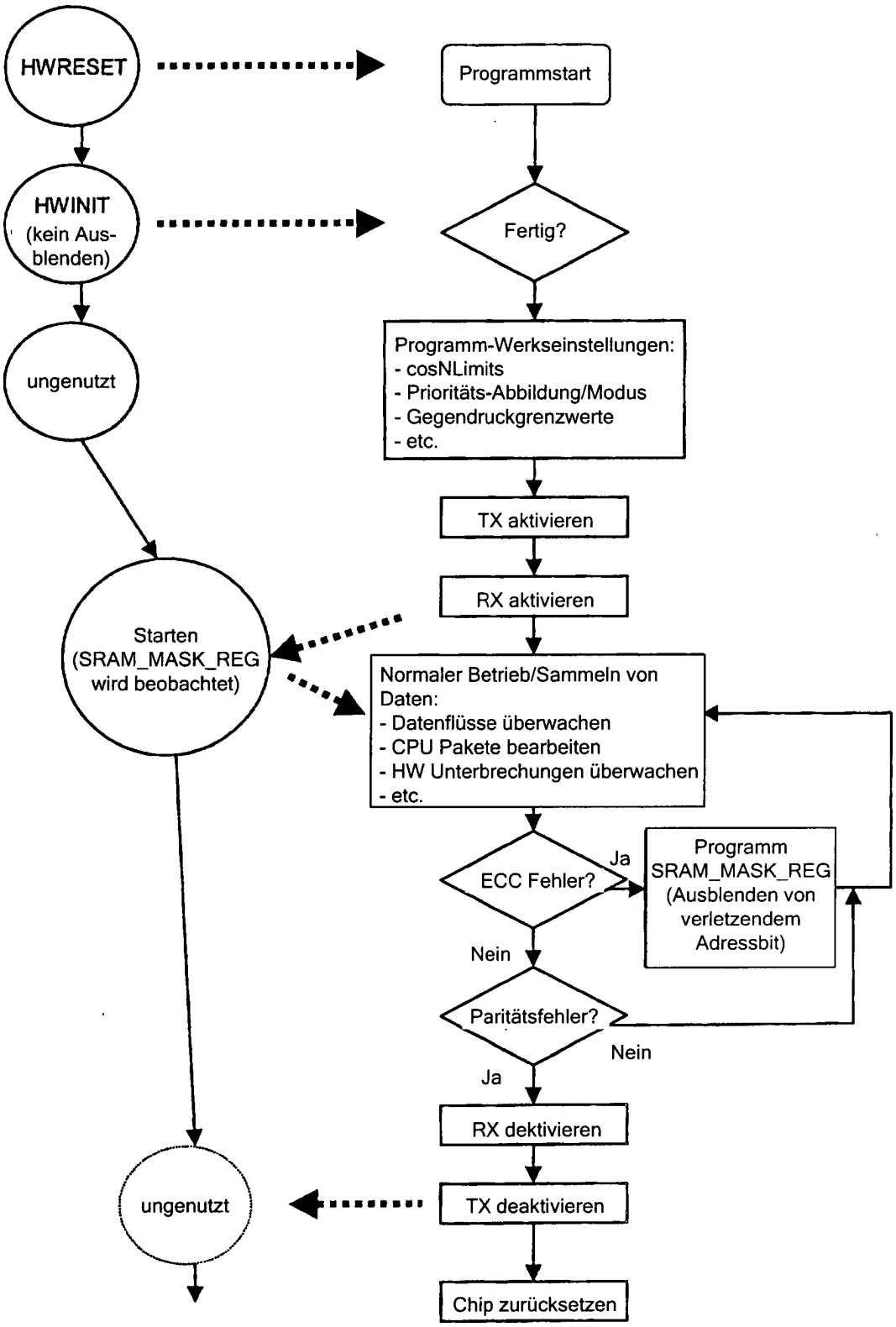
HWRESET
HWINIT
(kein Aus-
blenden)
ungenutzt
Starten
(SRAM_MASK_REG
wird beobachtet)
ungenutzt
Programmstart
Fertig?
Programm-Werkseinstellungen:
- cosNLimits
- Prioritäts-Abbildung/Modus
- Gegendruckgrenzwerte
- etc.
TX aktivieren
RX aktivieren
Normaler Betrieb/Sammeln von
Daten:
- Datenflüsse überwachen
- CPU Pakete bearbeiten
- HW Unterbrechungen überwachen
- etc.
ECC Fehler?
Ja
Programm
SRAM_MASK_REG
(Ausblenden von
verletzendem
Adressbit)
Nein
Paritätsfehler?
Nein
Ja
RX dektivieren
TX deaktivieren
Chip zurücksetzen

Fig. 15

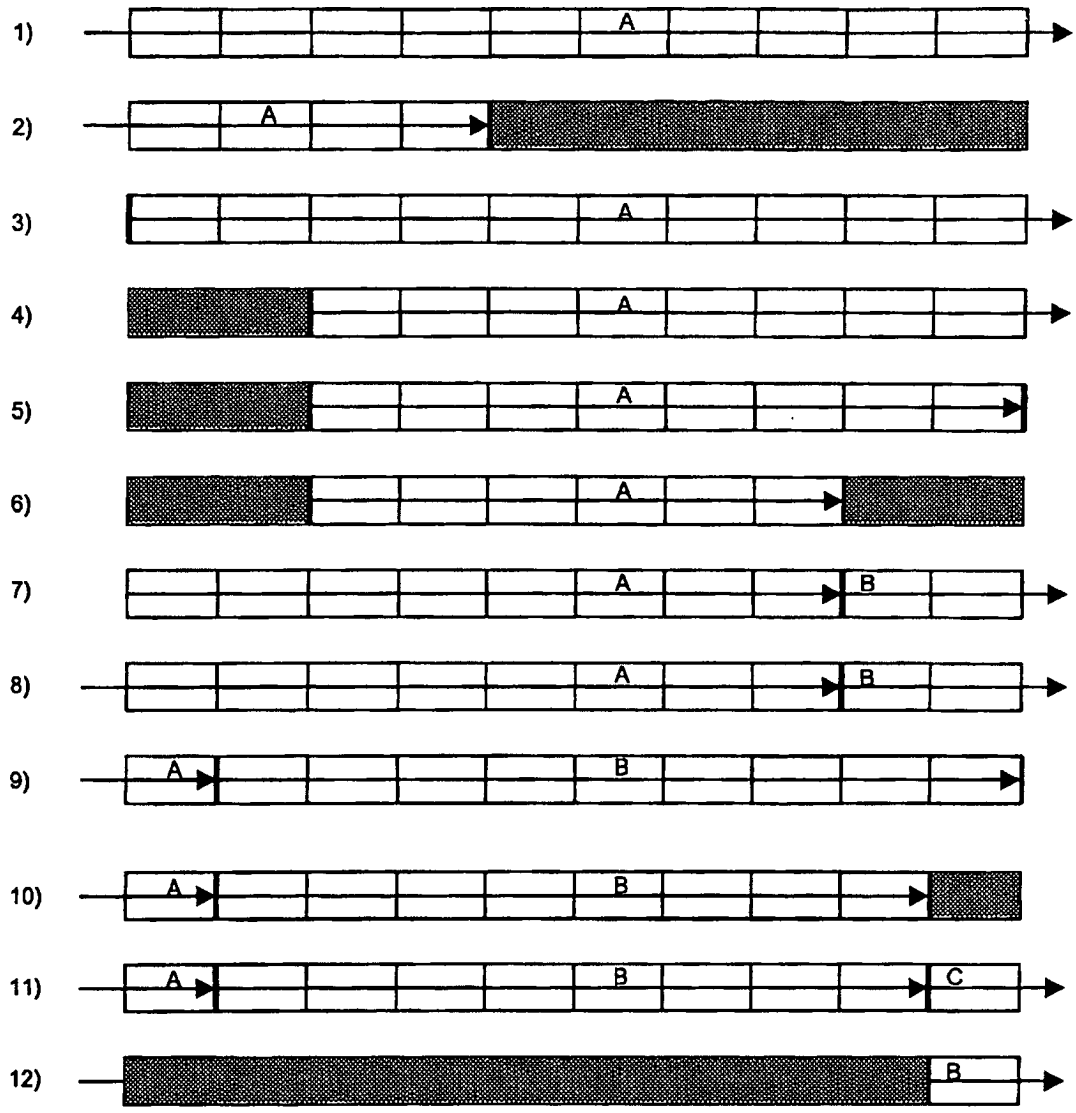
1)
A
2)
A
3)
A
4)
A
5)
A
6)
A
7)
A
B
8)
A
B
9)
A
B
10)
A
B
11)
A
B
C
12)
B

Fig. 14

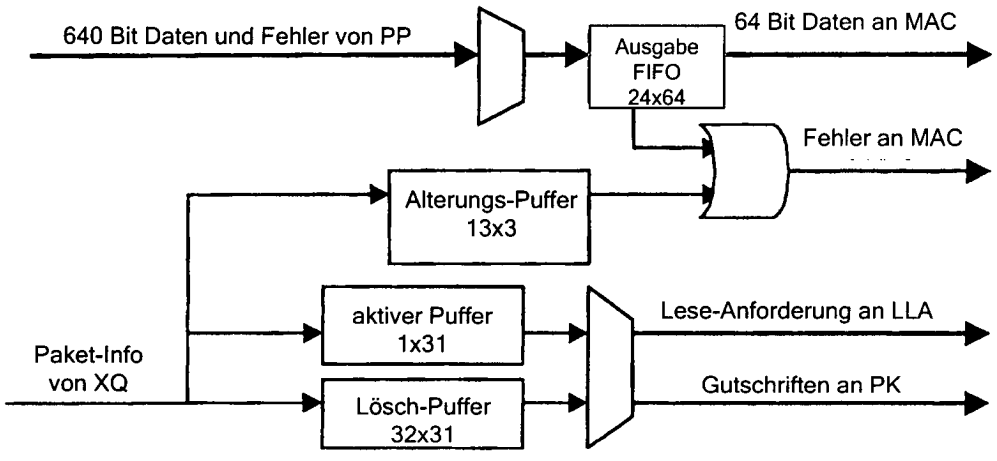
640 Bit Daten und Fehler von PP
Ausgabe FIFO 24x64
64 Bit Daten an MAC
Fehler an MAC
Alterungs-Puffer 13x3
aktiver Puffer 1x31
Lese-Anforderung an LLA
Paket-Info von XQ
Lösch-Puffer 32x31
Gutschriften an PK

Fig. 16