(19) Deutsches Patent- und Markenamt

(10) **DE 10 2011 014 588 A1** 2011.12.08

(12) # Offenlegungsschrift

(21) Aktenzeichen: **10 2011 014 588.5**
(22) Anmeldetag: **21.03.2011**
(43) Offenlegungstag: **08.12.2011**

(51) Int Cl.: ***G06F 12/16*** (2011.01)
G06F 13/00 (2011.01)
G06F 11/16 (2011.01)

(30) Unionspriorität:
12/748,764 **29.03.2010** **US**

(71) Anmelder:
Intel Corporation, Santa Clara, Calif., US

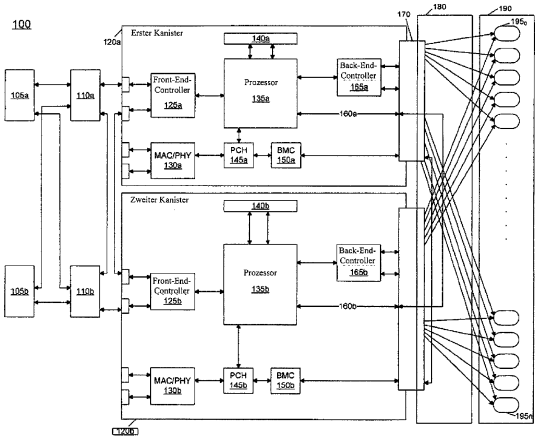
(74) Vertreter:
BOEHMERT & BOEHMERT, 28209, Bremen, DE

(72) Erfinder:
Kumar, Pankaj, Chandler, Ariz., US; Mitchell, James A., Chandler, Ariz., US

Die folgenden Angaben sind den vom Anmelder eingereichten Unterlagen entnommen

(54) Bezeichnung: **Multicasting-Schreibanforderungen an Mehrfachspeicher-Controller**

(57) Zusammenfassung: Bei einer Ausführungsform beinhaltet die vorliegende Erfindung ein Verfahren zum Ausführen von Multicasting, einschließlich Empfangen einer Schreibanforderung, einschließlich Schreibdaten, und einer Adresse von einem ersten Server in einem ersten Kanister, Bestimmen, ob die Adresse innerhalb einer Multicast-Region eines ersten Systemspeichers ist, und wenn ja, Senden der Schreibanforderung direkt an die Multicast-Region, um die Schreibdaten zu speichern, und ebenfalls an einen Spiegel-Port eines zweiten Kanisters, gekoppelt mit dem ersten Kanister, um die Schreibdaten an einen zweiten Systemspeicher des zweiten Kanisters zu spiegeln. Weitere Ausführungsformen sind beschrieben und werden beansprucht.



Beschreibung

Hintergrund

[0001] Speichersysteme, wie z. B. Datenspeichersysteme, beinhalten typischerweise eine externe Speicherplattform mit redundanten Speicher-Controllern, oft als Kanister bezeichnet, redundantem Netzteil, Kühllösung und einem Array an Platten. Die Plattformlösung ist so aufgebaut, dass sie eine einzelne Fehlerstelle mit voll-redundanten Eingabe-/Ausgabe-(input/output, I/O)-Pfaden und redundanten Controller toleriert, um Daten zugriffsbereit zu halten. Beide redundante Kanister in einem Gehäuse sind durch eine passive Backplane verbunden, um eine Cache-Spiegelungsfunktionalität zu ermöglichen. Wenn eins Kanister ausfällt, erhält der andere Kanister den Zugriff auf Festplatten, die mit dem ausfallenden Kanister verbunden sind, und fährt fort, I/O-Aufgaben bei den Platten auszuführen, bis der ausgefallene Kanister gewartet wird.

[0002] Um redundanten Betrieb zu ermöglichen, wird System-Cache-Spiegelung zwischen den Kanistern für alle ausstehenden plattengebundenen I/O-Transaktionen ausgeführt. Die Spiegelungsoperation beinhaltet primär Synchronisierung der System-Caches der Kanister. Obwohl der Ausfall eines einzelnen Knotenpunktes Verlust der Inhalte seines lokalen Cache bedeuten kann, wird immer noch eine zweite Kopie in dem Cache des redundanten Knotenpunktes beibehalten. Es existieren jedoch gewisse Komplexitäten in derzeitigen Systemen, einschließlich die Beschränkung der Bandbreite, die von den Spiegeloperationen verbraucht wird, und die Latenz, die erforderlich ist, um solche Operationen auszuführen.

Kurze Beschreibung der Zeichnungen

[0003] Fig. 1 ist ein Blockdiagramm eines Systems gemäß einer Ausführungsform der vorliegenden Erfindung.

[0004] Fig. 2 ist ein Blockdiagramm, das Details eines Kanisters gemäß einer anderen Ausführungsform der vorliegenden Erfindung zeigt.

[0005] Fig. 3 ist ein Datenfluss von Operationen gemäß einer Ausführungsform der vorliegenden Erfindung.

[0006] Fig. 4 ist ein Blockdiagramm von Komponenten, die bei direkter Adressübersetzung gemäß einer Ausführungsform der vorliegenden Erfindung verwendet werden.

Ausführliche Beschreibung

[0007] Bei verschiedenen Ausführungsformen können eingehende Schreiboperationen an einen Speicherkanister an mehrere Bestimmungsorte mehrfachgesendet werden. Bei einer Ausführungsform beinhalten diese Mehrfachorte Systemspeicher, der mit dem Speicherkanister und einem Spiegel-Port, z. B. entsprechend einem anderen Speicherkanister, verbunden ist. Auf diese Weise kann die Notwendigkeit für verschiedene Lese-/Schreiboperationen von Systemspeicher zu dem Spiegel-Port vermieden werden.

[0008] Obwohl der Umfang der vorliegenden Erfindung in dieser Hinsicht nicht begrenzt ist, kann Multicasting, das ein Dualcast an zwei Einheiten oder ein Multicast an mehr als zwei Einheiten sein kann, gemäß einer Peripheral Component Interconnect Express (PCI ExpressTM (PCIeTM))-Dualcasting-Funktionalität gemäß einer technischen Änderungsnotiz zu der PCIeTM-Basisspezifikation, Version 2.0 (veröffentlicht am 17. Januar 2007), ausgeführt werden. Es wird angenommen, dass hier ein erster Kanister eine eingehende gepostete Schreibanforderung, z. B. von einem Host, empfängt. Basierend auf einer Adresse der Anforderung kann das Schreibanforderungspaket an zwei Bestimmungsorte gerichtet sein, nämlich Systemspeicher des ersten Kanisters und den Spiegel-Port, z. B. ein zweiter Kanister, gekoppelt mit dem ersten Kanister, z. B. über einen PCIe™-NTB-(non-transparent bridge)-Port. Bei einer Ausführungsform kann die eingehende Adresse mit Basisadressregister (base address register, BAR) und Begrenzungsregistern des ersten Kanisters (z. B. verbunden mit dem PCIe™-I/O-Port des ersten Kanisters) und dem Spiegel-Port (PCIe™-NTB) verglichen werden, um sicherzustellen, dass die Pakete sowohl an den Systemspeicher als auch den Spiegel-Port weitergeleitet werden. Dieses Weiterleiten kann gleichzeitig ausgeführt werden, anstatt als eine serielle Implementierung, bei der Daten erst in den Systemspeicher geschrieben und sodann zu dem zweiten Kanister gespiegelt werden müssen.

[0009] Unter Verwendung von Ausführungsformen der vorliegenden Erfindung können fließende Spiegelschreibdatenflüsse für ein RAID-(redundant array of inexpensive disks)-System, wie z. B. ein RAID-5/6-System, verbessert werden. Da Speicherauslastungen in solch einem System sehr I/O-intensiv sein können und Systemspeicher mehrfach berühren, kann eine erhebliche Menge von Systemspeicherbandbreite verbraucht werden, besonders bei Entry-to-Mid-Range-Plattformen, die durch Systemspeicher leistungsbegrenzt sein können. Unter Verwendung einer Speicherbeschleunigungstechnologie kann gemäß einer Ausführungsform der vorliegenden Erfindung Speicherbandbreite verringert werden. Auf diese Weise kann Systemspeicher mit geringerer Leistung in einem System angenommen werden, wodurch die Systemkosten verringert werden. Beispielsweise können Bin-1-Speicherkomponenten (mit einer Frequenz mit geringerer Wertung als eine High-Bin-

Komponente) oder preiswerte DIMMs (dual inline memory modules) verwendet werden, um höhere RAID-5/6-Leistung zu erhalten.

[0010] Während Ausführungsformen eine PCIe™-Dualcast-Operation verwenden können, um eine eingehende Schreibanforderung hinsichtlich I/O-Schreiben zu Systemspeicher und PCIe™-zu-PCIe™-NTB in einer Operation auszuführen, können andere Implementierungen eine ähnliche Multicast- oder Broadcast-Operation verwenden, um eine Schreiboperation gleichzeitig an mehrere Bestimmungsorte zu leiten.

[0011] Indem nun Bezug genommen wird auf Fig. 1 wird ein Blockdiagramm eines Systems gemäß einer Ausführungsform der vorliegenden Erfindung veranschaulicht. Wie in Fig. 1 veranschaulicht, kann System **100** ein Speichersystem sein, in dem mehrere Server, z. B. Server **105a** und **105b** (allgemein Server **105**), mit einem Massenspeichersystem **190** verbunden sind, das eine Vielzahl von Plattenlaufwerken **1950**–**195n** (allgemein Plattenlaufwerke **195**) beinhalten kann, die ein RAID-System sein können, und gemäß einem Fibre Channel/SAS/SATA-Modell sein können. Bei RAID-5- oder RAID-6-Konfigurationen können Ausfälle einer Platte bzw. zweier Platten auf einer Speicherplattform toleriert werden.

[0012] Um Kommunikation zwischen Servern **105** und Speichersystem **190** zu realisieren, können Kommunikationen durch Schalter **110a** und **110b** (allgemein Schalter **110**) fließen, die Gigabit Ethernet (GigE)/Fibre Channel/SAS-Schalter sein können. Diese Schalter können wiederum mit einem Paar von Kanistern **120a** und **120b** (allgemein Kanister **120**) kommunizieren. Jeder dieser Kanister kann verschiedene Komponenten beinhalten, um Cache-Spiegelung gemäß einer Ausführungsform der vorliegenden Erfindung zu ermöglichen.

[0013] Speziell kann jeder Kanister einen Prozess **135** (allgemein) beinhalten. Zum Zwecke der Veranschaulichung wird erster Kanister **120a** erörtert, und somit kann Prozessor **135a** in Kommunikation mit einem Front-End-Controller-Gerät **125a** sein. Prozessor **135a** wiederum kann in Kommunikation mit einem peripheren Controller-Hub (peripheral controller hub, PCH) **145a** sein, der wiederum mit peripheren Geräten kommunizieren kann. Ebenfalls kann PCH **145** in Kommunikation mit einem MAC/PHY-(media access controller/physical)-Gerät **130a** sein, das in einer Ausführungsform ein Dual GigE MAC/PHY-Gerät sein kann, um Kommunikation von beispielsweise Managementinformationen zu ermöglichen. Es ist zu anzumerken, dass Prozessor **135a** weiter mit einem Baseboard-Management-Controller (baseboard management controller, BMC) **150a** gekoppelt sein kann, der wiederum mit einer Mid-Plane **180** über einen SM-(system management)-Bus kommunizieren kann.

[0014] Prozessor **135a** ist weiter gekoppelt mit einem Speicher **140a**, der bei einer Ausführungsform ein dynamischer Direktzugriffsspeicher (dynamic random access memory, DRAM), implementiert als DIMMs (dual inline memory modules), sein kann. Der Prozessor wiederum kann mit einem Back-End-Controller-Gerät **165a** gekoppelt sein, das durch Mid-Plane-Anschluss **170** ebenfalls mit Mid-Plane **180** gekoppelt ist.

[0015] Weiterhin kann eine PCIe™-NTB-Kopplungsstruktur **160** zwischen Prozessor **135a** und Mid-Plane-Anschluss **170** gekoppelt sein, um Spiegelung gemäß einer Ausführungsform der vorliegenden Erfindung zu ermöglichen. Wie ersichtlich, kann eine ähnliche Kopplungsstruktur Kommunikationen von diesem Link direkt an eine ähnliche PCIe™-NTB-Kopplungsstruktur **160b** weiterleiten, die an Prozessor **140b** des zweiten Kanisters **120b** ankoppelt. Diese Verbindung zwischen Prozessoren über die NTB-Kopplungsstruktur kann eine NTB-Domainadresse bilden. Es ist anzumerken, dass bei einigen Implementierungen die Kanister direkt ohne einen Mid-Plane-Anschluss ankoppeln können. Bei anderen Ausführungsformen kann, anstatt einer PCIe™-Kopplungsstruktur, eine andere PtP-(point-to-point)-Kopplungsstruktur vorhanden sein, wie z. B. gemäß dem Intel® Quick Path Interconnect(QPI)-Protokoll. Wie ersichtlich in Fig. 1, um redundante Operation zu ermöglichen, kann Mid-Plane **180** Kommunikation von jedem Kanister zu jedem entsprechenden Plattenlaufwerk **195** ermöglichen. Obwohl diese bestimmte Implementierung in der Ausführungsform von Fig. 1 veranschaulicht wird, ist der Umfang der vorliegenden Erfindung in dieser Hinsicht nicht eingeschränkt. Beispielsweise können mehr oder weniger Server und Plattenlaufwerke vorhanden sein, und bei einigen Ausführungsformen können ebenfalls zusätzliche Kanister bereitgestellt werden.

[0016] Indem nun Bezug genommen wird auf Fig. 2, wird ein Block-Diagramm gezeigt, das Details von Kanistern in Übereinstimmung mit einer anderen Ausführungsform der vorliegenden Erfindung zeigt. Es ist anzumerken, dass die Kanister von Fig. 2, nämlich ein erster Kanister **210a** und ein zweiter Kanister **210b**, Teil eines Systems **200** sein können, das einen oder mehr Server, ein Speichersystem, wie z. B. ein RAID-System, und Peripheriegeräte sowie andere solche Geräte beinhaltet. Bei zumindest einigen Implementierungen kann jedoch die Notwendigkeit für einen Schalter, um einen Server mit den Kanister zu koppeln, vermieden werden. Erster Kanister **210a** und zweiter Kanister **210b** sind über einen PCIe™-NTB-Link **250** gekoppelt, obwohl andere PtP-Verbindungen möglich sind. Über diesen Link kann System-Cache-Spiegelung zwischen den bei-

den Kanistern stattfinden. Eine NTB-Domainadresse **255** ist von beiden Kanistern **210** zugänglich. Bei der gezeigten Implementierung kann jeder Kanister **210** seine eigene Domainadresse haben, und kann einen Systemspeicher **240** beinhalten, der bei einer Ausführungsform unter Verwendung von preiswerten DIMMs implementiert werden kann, was durch die Speicherbeschleunigung ermöglicht wird, die unter Verwendung von Techniken gemäß einer Ausführungsform der vorliegenden Erfindung verfügbar ist.

[0017] Wie ersichtlich in Fig. 2, kann jeder Kanister I/O-Controller beinhalten, einschließlich einen oder mehr Host-I/O-Controller **212**, um Kommunikation mit Server und anderen Host-Geräten zu ermöglichen, sowie einen oder mehr Geräte-I/O-Controller **214**, um Kommunikation mit dem Plattensystem zu ermöglichen. Wie ersichtlich, können solche I/O-Controller mit einem entsprechenden Prozessor **220** über einen Root-Port **222** kommunizieren. Jeder Prozessor wiederum kann weiter einen NTB-Port **224** beinhalten, um Kommunikationen über NTB-Kopplungsstruktur **250** zu ermöglichen, die von NTB-Domainadresse **255** sein kann. Prozessor **220** kann weiter mit einem PCH **225** kommunizieren, der wiederum in Kommunikation mit einem MAC/PHY **230** sein kann. Es ist anzumerken, dass Prozessor **220** verschiedene interne Komponenten beinhalten kann, einschließlich einen integrierten Speicher-Controller, um Kommunikationen mit Systemspeicher zu ermöglichen, sowie eine integrierte DMA-(direct memory access)-Engine, und eine RAID-Prozessoreinheit, neben anderen solch spezialisierten Komponenten.

[0018] Unter Verwendung von Speicherbeschleunigung gemäß einer Ausführungsform der vorliegenden Erfindung kann eine Dualcasting-Technik verwendet werden, um Schreibdaten einer Schreibanforderung direkt zu Systemspeicher sowie zu einem verbundenen Gerät zu kommunizieren, wie z. B. ein PCIe™-verbundenes Gerät, wie z. B. ein anderer Kanister. Bezugnehmend auf Fig. 3, ist ein Datenfluss von Operationen gemäß einer Ausführungsform der vorliegenden Erfindung veranschaulicht. Wie in Fig. 3 veranschaulicht, wird der Datenfluss für einen RAID-5/6-Streaming-Spiegelschreibzugriff dargelegt. Allgemein kann ein Datenfluss, um eine Schreibanforderung zu empfangen und um Dualcasting-Spiegelung auszuführen, zwei Speicher-Leseoperationen und 2.25 Schreiboperationen beinhalten. Wie ersichtlich, kann eine eingehende Schreibanforderung, beispielsweise von einem Server, über einen Host-I/O-Controller **212a** von erstem Kanister **210a** empfangen werden. Abhängig von der Adresse der Schreibanforderung kann eine Dualcast-Operation initiiert werden. Spezifisch gesagt, wie nachstehend erörtert wird, wenn die Adresse innerhalb einer Dualcast-Region von Speicher ist, kann der Host-Controller die Daten in Systemspeicher **240a** gleichzeitig direkt schreiben sowie die Daten an Kanister **210b** über die NTB-Kopplungsstruktur spiegeln. Der Prozessor des zweiten Kanisters wiederum schreibt die Daten in seinen Systemspeicher als eine Spiegel-Schreiboperation.

[0019] Zu diesem Zeitpunkt können die Schreibdaten in beiden Systemspeichern vorliegen. Bei einer Implementierung kann sodann eine RAID-Prozessoreinheit, z. B. von Prozessor **220a** oder ein dedizierter RAID-Prozessor von Kanister **210a**, die Daten aus Speicher lesen, und RAID-5/6-Paritätsberechnungen ausführen, sowie die Paritätsdaten in den Systemspeicher **240a** schreiben, z. B. in Verbindung mit den Schreibdaten. Schließlich kann ein Geräte-I/O-Controller **214a** sowohl die Schreibdaten als auch die RAID-Paritätsdaten aus dem entsprechenden Systemspeicher **240a** lesen, und die Daten auf Platte schreiben, z. B. gemäß einer RAID-5/6-Operation, bei der die Daten über mehrere Platten gestreift sein können.

[0020] Es ist anzumerken, dass verschiedene Bestätigungen während der vorstehend beschriebenen Verarbeitung stattfinden können. Wenn die gespiegelten Schreibdaten beispielsweise in der geschützten Domain von Kanister **210b** erfolgreich empfangen wurden, um in Systemspeicher **240b** geschrieben zu werden, kann Kanister **210b** eine Bestätigung zurück an ersten Kanister **210a** kommunizieren. So wie diese Bestätigung anzeigt, dass die Schreibdaten jetzt erfolgreich in beide System-Caches, nämlich die beiden Systemspeicher, geschrieben wurden, kann zu diesem Zeitpunkt erster Kanister **210a** eine Bestätigung an den Anforderer zurücksenden, wie z. B. ein Server, um erfolgreiche Ausführung der Schreibanforderung zu bestätigen. Es ist anzumerken, dass diese Bestätigung gesendet werden kann, bevor die Schreibdaten in ihren finalen Bestimmungsort in dem RAID-System geschrieben werden, aufgrund der Redundanz, bereitgestellt von den Dualsystem-Caches. Dementsprechend kann das Schreiben von Systemspeicher **240a** auf Platte im Hintergrund stattfinden. Es ist anzumerken, dass die Systemspeicher der beiden Kanister durch Batterie-Backup gesichert sind. Zusätzlich, beim Schreiben der Daten in das Laufwerksystem, kann erster Kanister **210a** eine Nachricht an zweiten Kanister **210b** kommunizieren, um erfolgreiches Schreiben anzuzeigen. Zu diesem Zeitpunkt können die Schreibdaten, die in Systemspeicher **240b** (und Systemspeicher **240a**) gespeichert sind, in einen Dirty-Zustand versetzt werden, sodass der Platz für andere Daten wiederverwendet werden kann.

[0021] Damit kann die Notwendigkeit eingehende Daten von einem Host-I/O-Controller zuerst in Systemspeicher zu schreiben, und sodann eine DMA-Engine (wie z. B. des Prozessors) zu verwenden, um die Daten zwischen den beiden Kanistern zu spiegeln, vermieden werden. Stattdessen kann unter Verwen-

dung einer Ausführungsform der vorliegenden Erfindung das eingehende I/O-Schreibpaket gleichzeitig an zwei Bestimmungsorte, Systemspeicher und den Spiegel-Port, gesendet werden, wobei Speicher-Lese-/Schreiboperationen eliminiert werden und Speicher-Bandbreite eingespart wird, um höhere Leistung anzubieten. Oder kostengünstiger Speicher (wie z. B. Bin-Frequenz-1) kann verwendet werden, um Leistung anzubieten, die mit herkömmlichen RAID-Streaming-Operationen vergleichbar ist. Obwohl diese bestimmte Implementierung bei der Ausführungsform von Fig. 3 beschrieben ist, wird jedoch der Umfang der vorliegenden Erfindung in dieser Hinsicht nicht eingeschränkt.

[0022] Um eine Transaktion, die an einem Upstream-Port eines Root-Ports entsteht, mehrfachzusenden, das heißt, um sowohl auf Systemspeicher als auch ein gleichrangiges Gerät abzuzielen, kann ein Mechanismus verwendet werden, um es Transaktionen zu erlauben, die auf eine Teilmenge von Systemspeicher abzielen, ebenfalls transparent in den Spiegel-Port kopiert zu werden (z. B. der PCIe™-NTB-Port). Dazu kann Software in jedem Root-Port ein Multicast-Speicherfenster erzeugen, das zu Multicast-Operationen fähig ist. Beispielsweise kann ein Basis- und Begrenzungsregister bereitgestellt werden, um die Größe von einem der primären BARS der NTBs zu spiegeln, die dem gesamten BAR entsprechen können, das während Aufzählung für die NTB oder einer Teilmenge von diesem BAR definiert wurde.

[0023] Wenn eine Upstream-Schreibtransaktion an dem Root-Port gesehen wird, wird sie decodiert, um ihren Bestimmungsort zu bestimmen. Wenn die Adresse des Schreibzugriffs die Multicasting-Speicherregion trifft, wird sie sowohl an den Systemspeicher ohne Übersetzung als auch an das Speicherfenster der NTB nach Übersetzung gesendet. Bei einer Ausführungsform kann die Übersetzung eine direkte Adressübersetzung zwischen den beiden Seiten der NTB sein.

[0024] Bei einer Ausführungsform kann direkte Adressübersetzung nach angemessener Einrichtung von lokalen und entfernten Host-Adressabbildungen, die in jedem entsprechenden Host-Systemspeicher lokalisiert sein können, stattfinden. Bezugnehmend auf Fig. 4, wird ein Blockdiagramm von Komponenten, die in direkter Adressübersetzung gemäß einer Ausführungsform der vorliegenden Erfindung verwendet werden, veranschaulicht. Wie in Fig. 4 veranschaulicht, kann eine lokale Host-Adressabbildung **410** und eine entfernte Host-Adressabbildung **420** vorhanden sein. Wie ersichtlich, kann lokale Abbildung **410** einen Basisstandort **412** beinhalten, der einer Basisadresse für eine Dualcast-Speicherregion entsprechen kann. Zusätzlich kann ein Basis-Plus-Offset-Standort **414** verwendet werden, um eine übersetzte Basis- und Offset-Region **424** von entfernter Abbildung **420** zu erreichen. Zusätzlich kann ein Basis-Übersetzungsregister **422** in entfernter Abbildung **420** vorhanden sein. Verschiedene andere Register und Standorte können innerhalb dieser Adressabbildungen vorhanden sein.

[0025] Die folgenden Schritte skizzieren eine mögliche Implementierung. Zur Einrichtung liest Software Werte aus, die in der NTB für ein Basisadressregister (z. B. PBAR23SZ) gespeichert sind, und setzt eine Basisadresse für Dualcast-Operation (DUALCASTBASE) auf eine Größe eines Vielfachen von PBAR 23SZ. Das bedeutet, wenn PBAR23SZ 8 Gigabyte (GB) ist, dann wird DUALCASTBASE auf eine Größe eines Vielfachen von PBAR23SZ, z. B. 8 G, 16 G, 24 G und so weiter, gesetzt. Als nächstes kann eine Begrenzungsadresse für Dualcast-Operation gesetzt werden. Diese Begrenzungsadresse (DUALCASTLIMIT) kann auf weniger oder gleich DUALCASTBASE + PBAR23SZ gesetzt werden (beispielsweise wenn PBAR23SZ = 8 G und DUALCASTBASE = 24 G, dann kann DUALCASTLIMIT bis auf 32 G gesetzt werden). Dementsprechend kann die Dualcast-Region so gesetzt werden, dass sie die Region von Systemspeicher darstellt, die der Benutzer in entfernten Speicher spiegeln möchte. Diese Operationen können bei einer Ausführungsform von einem Betriebssystem (operating system, OS) gesetzt werden.

[0026] Während der Operation kann eine Upstream-Transaktion an dem Root-Port geprüft werden, um zu bestimmen, ob die empfangene Adresse in das von dem OS erzeugten Dualcast-Speicherfenster fällt. Diese Bestimmung kann gemäß der folgenden Gleichung sein: Gültige Dualcast-Adresse = ((DUALCASTLIMIT > Empfangene Adresse [63:0] >= DUALCASTBASE)).

[0027] Beispielsweise wird davon ausgegangen, dass Registerwerte von DUALCASTBASE = 0000 003A 0000 0000H, was die Dualcast-Basisadresse ist, die vom OS auf eine Größe eines Vielfachen der PBAR23SZ-Ausrichtung gesetzt ist, in diesem Fall 4 GB, und ein DUALCASTLIMIT = 0000 003A C000 0000H, was das Fenster auf 3 GB verringert. Es wird weiter davon ausgegangen, dass die Empfangene Adresse = 0000 003A 00A0 0000H ist. Gemäß vorstehender Gleichung entspricht dies einer gültigen Dualcast-Adresse, und somit kann eine Übersetzung stattfinden, die weiter nachstehend erörtert wird.

[0028] Wenn die empfangene Adresse außerhalb dieses Dualcast-Speicherfensters ist, kann die Transaktion basierend auf den Systemanforderungen decodiert werden. Die Transaktion kann beispielsweise an Systemspeicher decodiert, gleichrangig decodiert, subtraktiv an die Southbridge decodiert, oder masterabgebrochen werden.

[0029] Wenn wie vorstehend die Transaktion innerhalb der gültigen Dualcast-Region ist, kann sie an das definierte primärseitige NTB-Speicherfenster übersetzt werden. Diese Übersetzung kann wie folgt sein: Übersetzte Adresse
= ((Empfangene Adresse [63:0] & ~Sign_Extend (2^PBAR23SZ)|PBAR2XLAT [63:0])).

[0030] Um beispielsweise eine eingehende Adresse, die von einem 4 GB Fenster beansprucht wird, basierend auf 0000 003A 0000 0000H, an ein 4 GB Fenster, basierend auf 0000 0040 0000 0000H, zu übersetzen, kann die folgende Berechnung stattfinden.
Empfangene Adresse [63:0] = 0000 003A 00A0 0000H

[0031] PBAR23SZ = 32, was die Größe von Primär-BAR 2/3 = 4 GB in diesem Beispiel setzt. ~Sign_Extend(2^PBAR23SZ) = ~Sign_Extend (0000 0001 0000 0000H) = ~(FFFF FFFF 0000 0000H) = (0000 0000 FFFF FFFFH) PBAR2XLAT = 0000 0040 0000 0000H, was die Basisadresse in den primärseitigen NTB-Speicher (Größe mehrfach ausgerichtet) ist. Entsprechend ist die Übersetzte Adresse = 0000 003A 00A0 0000H & 0000 0000 FFFF FFFFH|0000 0040 0000 0000H = 0000 0040 00A0 0000H.

[0032] Es ist anzumerken, dass der Offset zu der Basis des 4 GB Fensters an der eingehenden Adresse in der übersetzten Adresse bewahrt wird.

[0033] Unter Verwendung der übersetzten Adresse kann eine Dualcast-Operation ausgeführt werden, um die eingehende Transaktion an Systemspeicher bei (0000 0030 00A0 0000H) und an die NTB bei (0000 0040 00A0 0000H) zu senden.

[0034] Implementierungen zur Handhabung einer eingehenden Multicast-Schreibanforderung können basierend auf der Mikroarchitektur, die verwendet wird, unterschiedlich ausgeführt werden. Eine Implementierung kann beispielsweise sein, eine Anforderung von einer receivergeposteten Reihe abzuziehen, und die Transaktion zeitweilig in einer Warteschlange zu halten. Sodann kann der Root-Port unabhängige Anforderung für Zugriff auf Systemspeicher und für Zugriff auf gleichrangige Speicher senden. Die Transaktion würde in der Warteschlange verbleiben, bis eine Kopie in sowohl Systemspeicher als auch gleichrangigem Speicher übernommen worden ist, und wird sodann aus der Warteschlange gelöscht. Eine alternative Implementierung kann warten, eine Anforderung von einer receivergeposteten Reihe abzuziehen, bis sowohl die Upstream-Ressourcen, die auf Systemspeicher abzielen, als auch gleichrangige Ressourcen verfügbar sind, und sie sodann an beide Pfade zur gleichen Zeit senden. Der Pfad zu Hauptspeicher kann beispielsweise die Anforderung mit der gleichen Adresse senden, die empfangen wurde, und der Pfad zu der gleichrangigen NTB kann die Anforderung nach Übersetzung an eines der primären NTB-Speicherfenster senden.

[0035] Ausführungsformen können als Code implementiert und auf einem Speichermedium gespeichert werden, das Befehle enthält, die zum Programmieren eines Systems für die Ausführung der Befehle verwendet werden können. Das Speichermedium kann beinhalten, ist aber nicht beschränkt auf, jede Art Disks, u. a. Floppy Disks, Optische Disks, Solid State-Laufwerke (SSDs), Compact Disk Read-Only Memories (CD-ROMs), Compact Disk Rewritables (CD-RWs) und magnetooptische Disks (MO), Halbleiter-Geräte, wie Read-Only Memories (ROMs), Random Access Memories (RAMs), wie Dynamic Random Access Memories (DRAMs), Static Raudom Access Memories (SRAMs), Erasable Programmable Read-Only Memories (EPROMs), Flash Memories, Electrically Erasable Programmable Read-Only Memories (EEPROMs), magnetische oder optische Karten oder jede andere Art Speichermedium, das sich für das Speichern von elektronischen Befehlen eignet.

[0036] Obwohl die vorliegende Erfindung im Hinblick auf eine begrenzte Anzahl von Ausführungsformen beschrieben wurde, sind sich Fachleute bewusst, dass viele weitere Modifikationen und Varianten davon möglich sind. Die beigefügten Ansprüche sollen alle solchen Modifikationen und Varianten abdecken, die dem Sinn und Schutzbereich der vorliegenden Erfindung entsprechen.

ZITATE ENTHALTEN IN DER BESCHREIBUNG

Diese Liste der vom Anmelder aufgeführten Dokumente wurde automatisiert erzeugt und ist ausschließlich zur besseren Information des Lesers aufgenommen. Die Liste ist nicht Bestandteil der deutschen Patent- bzw. Gebrauchsmusteranmeldung. Das DPMA übernimmt keinerlei Haftung für etwaige Fehler oder Auslassungen.

Zitierte Nicht-Patentliteratur

- technischen Änderungsnotiz zu der PCIeTM-Basisspezifikation, Version 2.0 (veröffentlicht am 17. Januar 2007) [**0008**]

Patentansprüche

1. Vorrichtung, umfassend:
einen ersten Kanister, um Speichern von Daten in einem Speichersystem zu steuern, das eine Vielzahl von Platten beinhaltet, wobei der erste Kanister einen ersten Prozessor, einen ersten Systemspeicher, um Daten in den Cache-Speicher aufzunehmen, die in dem Speichersystem gespeichert werden sollen, und einen ersten Spiegel-Port aufweist; und
einen zweiten Kanister, um Speichern von Daten in dem Speichersystem zu steuern, und gekoppelt an den ersten Kanister über eine PtP-(point-to-point)-Kopplungsstruktur, wobei der zweite Kanister einen zweiten Prozessor, einen zweiten Systemspeicher, um Daten in den Cache-Speicher aufzunehmen, die in dem Speichersystem gespeichert werden sollen, und einen zweiten Spiegel-Port beinhaltet, wobei die ersten und zweiten Systemspeicher eine gespiegelte Kopie der Daten speichern sollen, die in dem anderen Systemspeicher gespeichert ist, wobei die gespiegelte Kopie durch Dualcast-Transaktionen über die PtP-Kopplungsstruktur kommuniziert wird, wobei eingehende Daten an den ersten Kanister gleichzeitig in den ersten Systemspeicher geschrieben und an den zweiten Kanister über die ersten und zweiten Spiegel-Ports kommuniziert werden.

2. Vorrichtung nach Anspruch 1, wobei der erste Kanister direkt mit einem Server gekoppelt ist, der eine Schreibanforderung für die eingehenden Daten ohne einen Schalter hervorbringt.

3. Vorrichtung nach Anspruch 1, weiter umfassend einen Geräte-Controller, gekoppelt mit dem ersten Prozessor, wobei der Geräte-Controller die eingehenden Daten von dem ersten Systemspeicher empfangen soll, und die eingehenden Daten in zumindest ein Laufwerk eines Laufwerksystem des Speichersystems schreiben soll.

4. Vorrichtung nach Anspruch 1, weiter umfassend eine RAID-(redundant array of inexpensive disks)-Engine des ersten Prozessors, um die eingehenden Daten aus dem ersten Systemspeicher zu lesen, und um eine Paritätsoperation bei den eingehenden Daten auszuführen, und um ein Ergebnis der Paritätsoperation in dem ersten Systemspeicher zu speichern.

5. Vorrichtung nach Anspruch 1, weiter umfassend einen Root-Port des ersten Kanisters, wobei der Root-Port bestimmen soll, ob die eingehenden Daten über eine Dualcast-Transaktion, basierend auf einer Adresse einer Schreibanforderung, einschließlich der eingehenden Daten, gespiegelt werden sollen.

6. Vorrichtung nach Anspruch 5, wobei der Root-Port die Adresse der Schreibanforderung an ein Speicherfenster des zweiten Systemspeichers übersetzen soll, und die Dualcast-Transaktion an den ersten Systemspeicher mit der Adresse, und an den zweiten Kanister mit der übersetzten Adresse senden soll.

7. Vorrichtung nach Anspruch 2, wobei der zweite Prozessor eine Bestätigung nach Erhalt der gespiegelten Kopie der eingehenden Daten über die PtP-Kopplungsstruktur übertragen soll, und wobei in Antwort auf die Bestätigung der erste Prozessor eine zweite Bestätigung an den Server übertragen soll, um erfolgreiche Ausführung der Schreibanforderung für die eingehenden Daten anzuzeigen.

8. Verfahren, umfassend:
Empfangen einer Schreibanforderung, einschließlich Schreibdaten, und einer Adresse von einem ersten Server in einem ersten Kanister eines Speichersystems;
Bestimmen, ob die Adresse innerhalb einer Multicast-Region eines Systemspeichers des ersten Kanisters ist;
wenn ja, Senden der Schreibanforderung direkt an die Multicast-Region des Systemspeichers des ersten Kanisters, um die Schreibdaten in dem Systemspeicher des ersten Kanisters zu speichern, und an einen Spiegel-Port eines zweiten Kanisters, gekoppelt mit dem ersten Kanister über einen PtP-(point-to-point)-Link, um die Schreibdaten an einen Systemspeicher des zweiten Kanisters zu spiegeln; und
Empfangen einer Bestätigung des Erhalts der Schreibdaten in dem ersten Kanister von dem zweiten Kanister über den PtP-Link, und Kommunizieren einer zweiten Bestätigung von dem ersten Kanister an den ersten Server.

9. Verfahren nach Anspruch 8, weiter umfassend Lesen der Schreibdaten aus dem Systemspeicher des ersten Kanisters, und Ausführen einer Paritätsoperation bei den Schreibdaten sowie Speichern eines Ergebnisses der Paritätsoperation in dem Systemspeicher des ersten Kanisters.

10. Verfahren nach Anspruch 9, weiter umfassend Ausführen der Paritätsoperation unter Verwendung einer RAID-(redundant array of inexpensive disks)-Engine eines Prozessors des ersten Kanisters.

11. Verfahren nach Anspruch 10, weiter umfassend anschließendes Senden der Schreibdaten und des Ergebnisses der Paritätsoperation von dem Systemspeicher des ersten Kanisters an ein Laufwerksystem des Speichersystems über eine zweite Kopplungsstruktur.

12. Verfahren nach Anspruch 11, weiter umfassend Senden einer Nachricht von dem ersten Kanister an den zweiten Kanister, um erfolgreiches Schreiben der Schreibdaten und des Ergebnisses der Paritätsoperation auf das Laufwerksystem anzuzeigen.

13. Verfahren nach Anspruch 11, weiter umfassend Speichern der Schreibdaten und des Ergebnisses der Paritätsoperation über eine Vielzahl von Laufwerken des Laufwerksystems.

14. System, umfassend:
einen ersten Kanister, der einen ersten Prozessor, einen ersten Systemspeicher, um Daten in den Cache-Speicher aufzunehmen, einen ersten Eingabe-/Ausgabe-(input/output, I/O)-Controller, um mit einem ersten Server zu kommunizieren, einen ersten Geräte-Controller, um mit einem Plattenspeichersystem zu kommunizieren, und einen ersten Spiegel-Port beinhaltet;
einen zweiten Kanister, gekoppelt an den ersten Kanister über eine PtP-(point-to-point)-Kopplungsstruktur, wobei der zweite Kanister einen zweiten Prozessor, einen zweiten Systemspeicher, um Daten in den Cache-Speicher aufzunehmen, einen zweiten I/O-Controller, um mit einem zweiten Server zu kommunizieren, einen zweiten Geräte-Controller, um mit dem Plattenspeichersystem zu kommunizieren, und einen zweiten Spiegel-Port beinhaltet, wobei die ersten und zweiten Systemspeicher eine gespiegelte Kopie der Daten speichern sollen, die in dem anderen Systemspeicher gespeichert ist, wobei die gespiegelte Kopie durch Dualcast-Transaktionen über die PtP-Kopplungsstruktur kommuniziert wird, wobei eingehende Daten einer Schreibanforderung an den ersten Kanister gleichzeitig in den ersten Systemspeicher geschrieben und an den zweiten Kanister über die ersten und zweiten Spiegel-Ports kommuniziert werden; und
das Plattenlaufwerksystem, einschließlich einer Vielzahl von Plattenlaufwerken.

15. System nach Anspruch 14, weiter umfassend eine RAID-(redundant array of inexpensive disks)-Engine des ersten Prozessors, um die eingehenden Daten aus dem ersten Systemspeicher zu lesen, und um eine Paritätsoperation bei den eingehenden Daten auszuführen, und um ein Ergebnis der Paritätsoperation in dem ersten Systemspeicher zu speichern.

16. System nach Anspruch 15, wobei der erste Geräte-Controller die eingehenden Daten und das Ergebnis der Paritätsoperation aus dem ersten Systemspeicher in zumindest einige der Plattenlaufwerke des Plattenlaufwerksystems schreiben soll.

17. System nach Anspruch 16, wobei der erste Kanister eine Nachricht an den zweiten Kanister senden soll, um dem zweiten Kanister zu ermöglichen, eine Speicherregion freizugeben, die die gespiegelte Kopie der eingehenden Daten speichert.

18. System nach Anspruch 14, weiter umfassend einen Root-Port des ersten Kanisters, wobei der Root-Port bestimmen soll, ob die eingehenden Daten über eine Dualcast-Transaktion, basierend auf einer Adresse der Schreibanforderung, gespiegelt werden sollen.

19. System nach Anspruch 18, wobei der Root-Port die Adresse der Schreibanforderung an ein Speicherfenster des zweiten Systemspeichers übersetzen soll, und die Dualcast-Transaktion an den ersten Systemspeicher mit der Adresse, und an den zweiten Kanister mit der übersetzten Adresse senden soll.

20. System nach Anspruch 14, wobei der zweite Kanister eine Bestätigung nach Erhalt der gespiegelten Kopie der eingehenden Daten über die PtP-Kopplungsstruktur übertragen soll, und wobei in Antwort auf die Bestätigung der erste Kanister eine zweite Bestätigung an den Server übertragen soll, um erfolgreiche Ausführung der Schreibanforderung für die eingehenden Daten anzuzeigen.

Es folgen 4 Blatt Zeichnungen

Anhängende Zeichnungen

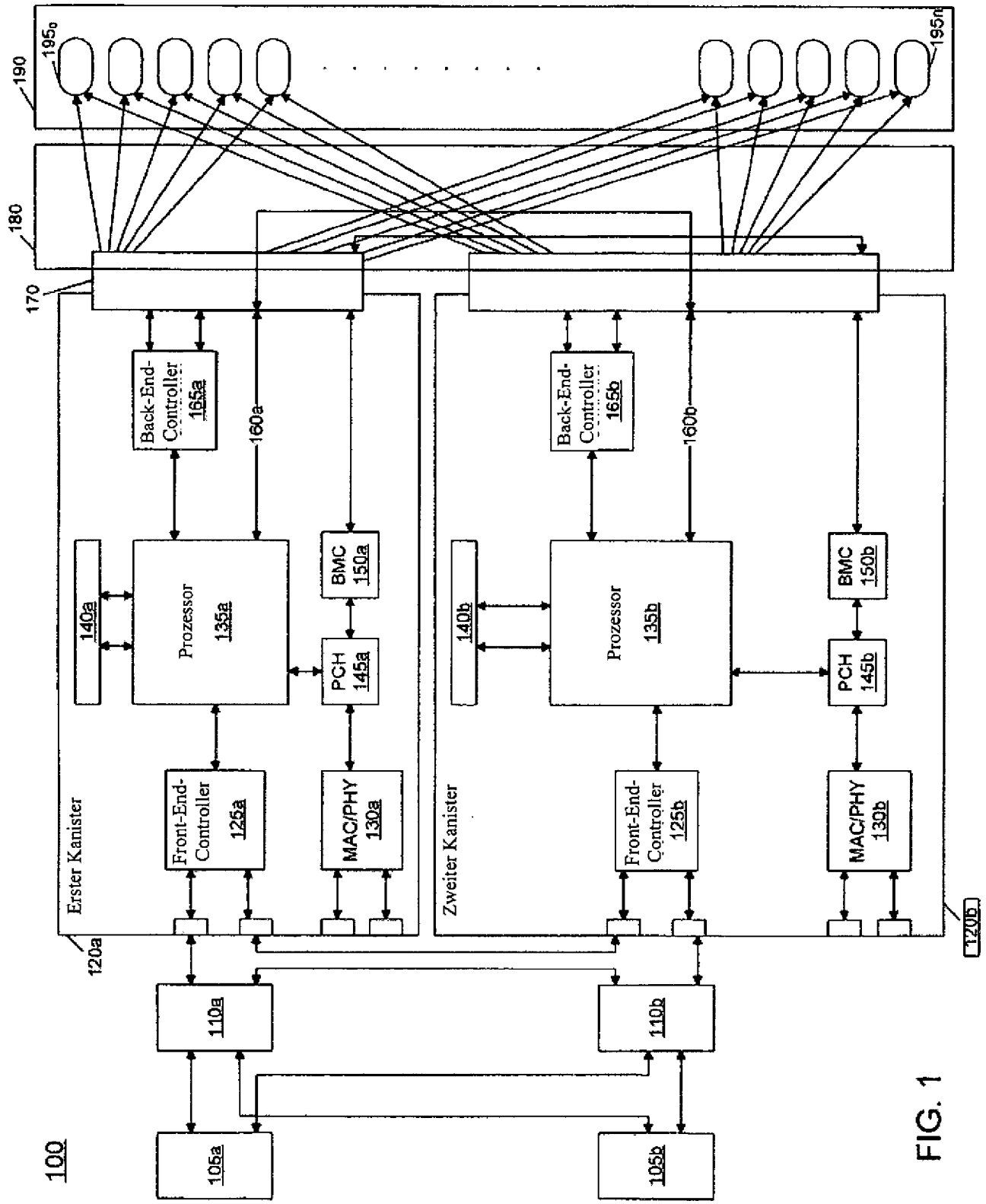


FIG. 1

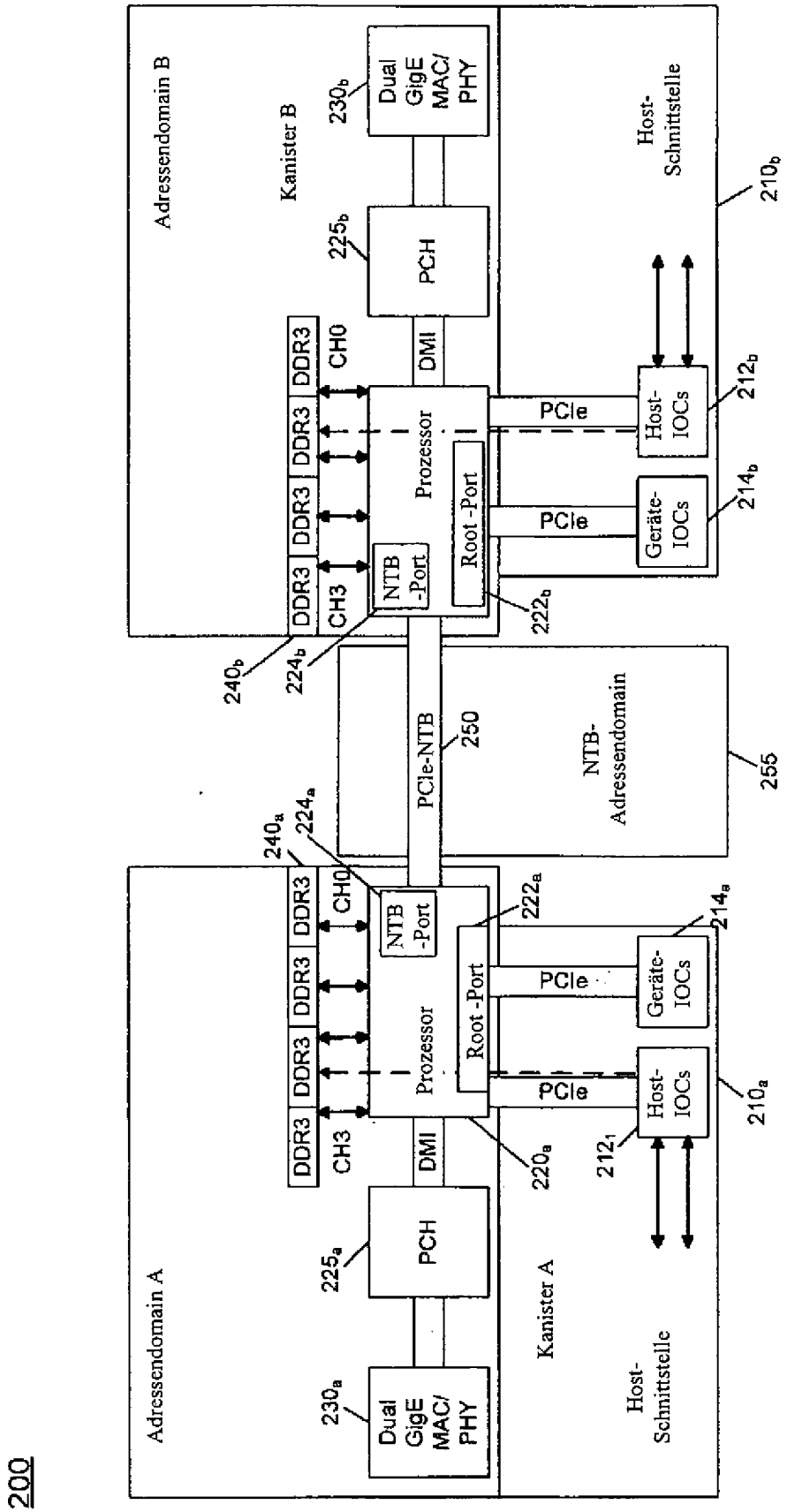

200
Adressendomain B
Kanister B
Dual GigE MAC/ PHY
230b
PCH
225b
DMI
DDR3
CH0
CH3
Prozessor
NTB -Port
Root -Port
PCIe
Host-IOCs
212b
Geräte-IOCs
214b
Host-Schnittstelle
210b
222b
240b
224b
PCIe-NTB
250
NTB-Adressendomain
255
240a
224a
222a
214a
212i
220a
210a
Adressendomain A
Kanister A
230a
225a

FIG. 2

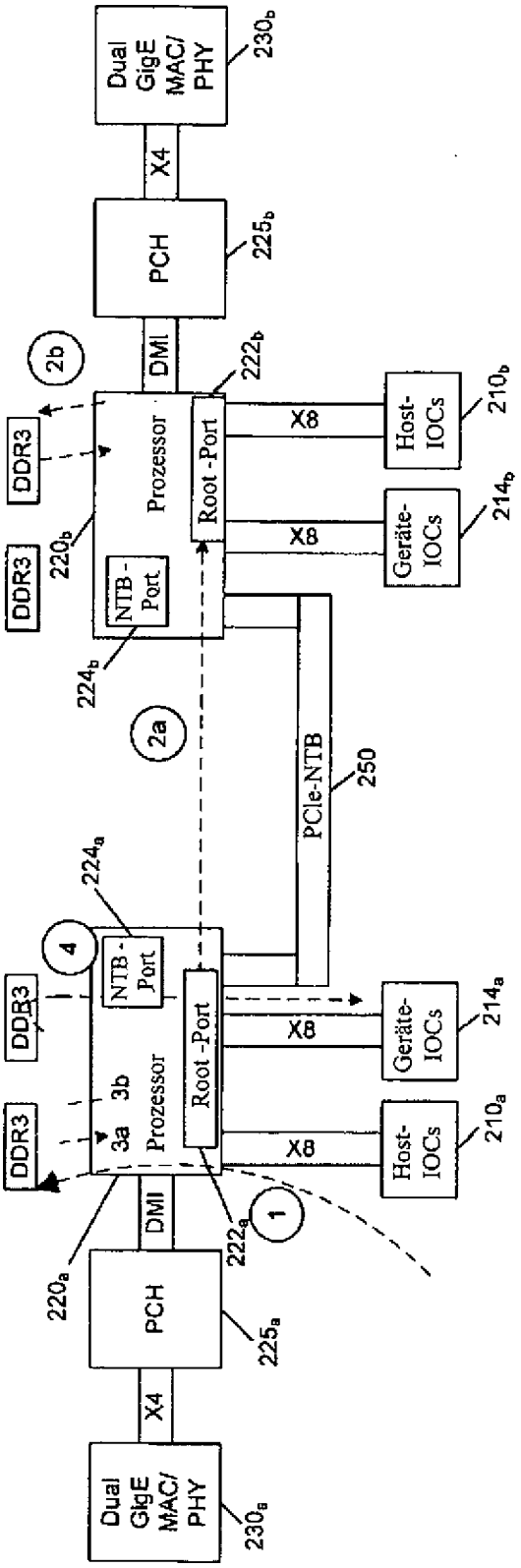

Dual GigE MAC/ PHY
230b
X4
PCH
225b
DMI
2b
222b
DDR3
Prozessor
Root -Port
X8
Host-IOCs
210b
DDR3
220b
NTB - Port
X8
Geräte-IOCs
214b
224b
2a
PCIe-NTB
250
224a
4
NTB - Port
DDR3
Root -Port
X8
Geräte-IOCs
214a
DDR3
3b
3a
Prozessor
X8
Host-IOCs
210a
DMI
1
222a
220a
PCH
225a
X4
Dual GigE MAC/ PHY
230a

FIG. 3

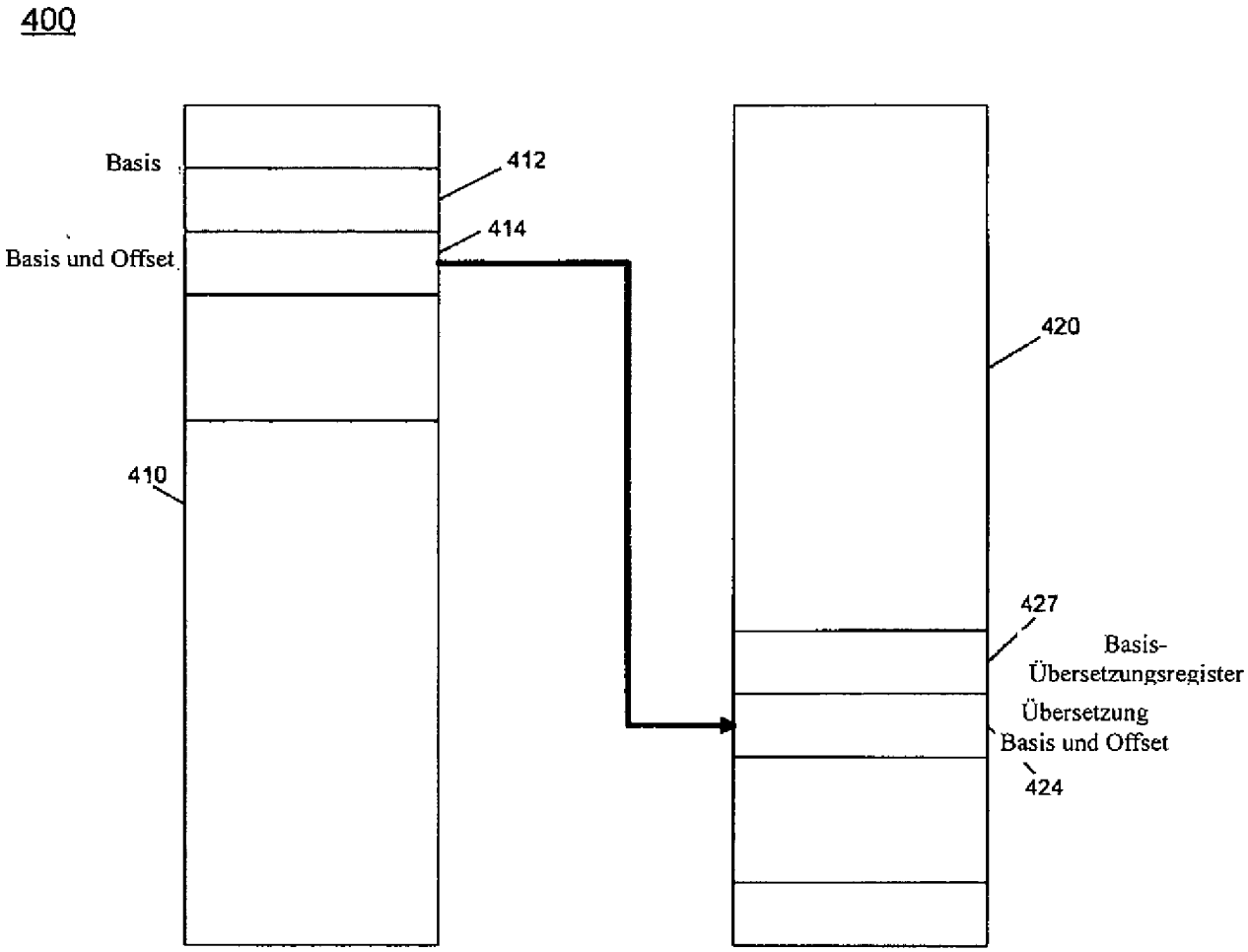

400
Basis
412
414
Basis und Offset
420
410
427
Basis-
Übersetzungsregister
Übersetzung
Basis und Offset
424

FIG. 4