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

(10) **DE 698 33 815 T2** 2006.12.07

(12) # Übersetzung der europäischen Patentschrift

(97) **EP 0 899 731 B1**
(21) Deutsches Aktenzeichen: **698 33 815.4**
(96) Europäisches Aktenzeichen: **98 105 564.3**
(96) Europäischer Anmeldetag: **26.03.1998**
(97) Erstveröffentlichung durch das EPA: **03.03.1999**
(97) Veröffentlichungstag der Patenterteilung beim EPA: **15.03.2006**
(47) Veröffentlichungstag im Patentblatt: **07.12.2006**

(51) Int Cl.[8]: ***G11B 19/02*** (2006.01)
G11B 27/00 (2006.01)
G11B 27/10 (2006.01)
G11B 20/12 (2006.01)
G06F 11/14 (2006.01)

(30) Unionspriorität:
920120 **26.08.1997** **US**

(84) Benannte Vertragsstaaten:
DE, GB

(73) Patentinhaber:
Hewlett-Packard Development Co., L.P., Houston, Tex., US

(74) Vertreter:
Schoppe, Zimmermann, Stöckeler & Zinkler, 82049 Pullach

(72) Erfinder:
Voigt, Douglas L., Boise, Idaho 83702, US; Burkes, Don L., Meridian, ID 83642, US; Hanson, Kirk A., Eagle, Idaho 83616, US

(54) Bezeichnung: **Verbesserter Disk-Log mit verteiltem Schreibsystem**

Beschreibung

Gebiet der Erfindung

[0001] Diese Erfindung bezieht sich allgemein auf Datenspeichersysteme und insbesondere auf eine Transaktionsprotokollverwaltung für Plattenarrayspeichersysteme.

Hintergrund der Erfindung

[0002] Computersysteme verbessern sich ständig hinsichtlich einer Geschwindigkeit, Zuverlässigkeit und Verarbeitungsfähigkeit. Folglich sind Computer in der Lage, komplexere und höherentwickelte Anwendungen zu handhaben. Wenn sich Computer verbessern, erhöhen sich Leistungsfähigkeitsforderungen, die an Massenspeicher- und Eingabe/Ausgabe-Vorrichtungen (I/O-Vorrichtungen; I/O = input/output) gestellt werden. Somit gibt es einen stetigen Bedarf, Massenspeichersysteme zu entwerfen, die hinsichtlich einer Leistungsfähigkeit mit sich entwickelnden Computersystemen Schritt halten.

[0003] Diese Erfindung betrifft insbesondere die Massenspeichersysteme des Plattenarraytyps. Plattenarraydatenspeichersysteme weisen mehrere Speicherplattenlaufwerksvorrichtungen auf, die angeordnet und koordiniert sind, um ein einziges Massenspeichersystem zu bilden. Es gibt drei primäre Entwurfskriterien für Massenspeichersysteme: Kosten, Leistungsfähigkeit und Verfügbarkeit. Es ist am erwünschtesten, Speichervorrichtungen zu erzeugen, die niedrige Kosten pro Megabyte, eine hohe Eingabe/Ausgabe-Leistungsfähigkeit und eine hohe Datenverfügbarkeit aufweisen. „Verfügbarkeit" ist die Fähigkeit, auf Daten zuzugreifen, die in dem Speichersystem gespeichert sind, und die Fähigkeit, einen fortlaufenden Betrieb in dem Fall eines gewissen Ausfalls sicherzustellen. Typischerweise ist eine Datenverfügbarkeit durch die Verwendung einer Redundanz vorgesehen, wobei Daten, oder Beziehungen zwischen Daten, bei mehreren Positionen gespeichert sind. Zwei häufige Verfahren zum Speichern redundanter Daten sind das „Spiegel"- und das „Paritäts"-Verfahren.

[0004] Ein Problem, das bei dem Entwurf von Plattenarraydatenspeichersystemen angetroffen wird, betrifft die Frage eines Behaltens genauer Abbildungsinformationen der Daten in einem Speicher in dem Fall eines Systemfehlers oder -ausfalls. Dies gilt für Systeme, die entweder eines oder beide Verfahren zum Speichern redundanter Daten einsetzen. Somit ist es im Zuge eines Verwaltens von Plattenarrayabbildungsinformationen häufig notwendig, sicherzustellen, dass jüngst geänderte Abbildungsinformationen zu Fehlererholungszwecken auf einer Platte gespeichert sind. Dieses Plattenschreiberfordernis kann aus mehreren Gründen auftreten, wie beispielsweise (i) einer zeitbasierten Frequenzstatusaktualisierung, (ii) einem Protokollseite-Voll-Status oder (iii) einer spezifischen Host-Anforderung. Allgemein sind jüngste Veränderungen bei zufälligen Positionen in Datenstrukturen gesammelt, die hinsichtlich einer Leistungsfähigkeit der Plattenarrayfunktion optimiert sind, und sind zusätzlich sequentiell in einem Protokoll gesammelt, das schneller als die anderen Datenstrukturen zu einer Platte geschrieben (abgelegt) werden kann. Diese Technik ist auf dem Gebiet einer Transaktionsverarbeitung üblich. Unvorteilhafterweise kann jedoch das Ablageerfordernis gleichzeitig mit einer anderen ablaufenden Plattenlese- oder Schreibaktivität auftreten, wodurch eine I/O-Konkurrenz erzeugt wird. Eine derartige I/O-Konkurrenz extrahiert häufig einen erheblichen Leistungsfähigkeitstreffer an dem System, insbesondere falls das Ablegen häufig auftritt, weil mehrere I/O-Ereignisse für eine einzige Ablage des Protokolls zu einer Platte auftreten müssen. Typischerweise ist die Protokollseite zuerst beispielsweise als ungültig markiert (d.h. dieselbe muss aktualisiert werden). Dann wird die Protokollseite zu einer Platte kopiert und nachfolgend als gültig markiert. Schließlich wird bei einem redundanten System die redundante Protokollseite zu einer Platte kopiert.

[0005] Angesichts des Vorhergehenden und der sich stets erhöhenden Rechengeschwindigkeiten, die angeboten werden, und massiver Informationsmengen, die verarbeitet werden, besteht ein ständiger Bedarf nach einer verbesserten Leistungsfähigkeit bei Plattenarraysystemen und dergleichen.

[0006] Die EP 0 426 354 A beschreibt ein Verfahren zum Ausgleichen der Lasten an Kanalwegen bei einem digitalen Computer während lange laufenden Anwendungen. Für jedes Datenvolumen, das für eine Verarbeitung in Frage kommt, wird die Auswahl von Volumen zuerst aus diesen vorgenommen, die eine Affinität zu dem aufrufenden Host aufweisen. Die Last über die entsprechenden verbundenen Kanalwege wird berechnet. Die Berechnung wird gewichtet, um die unterschiedlichen Beträge einer Last zu berücksichtigen, die aus unterschiedlichen Anwendungen resultieren, und um die Auswahl von Volumen zu bevorzugen, die mit den wenigsten unbenutzten Kanalwegen verbunden sind, derart, dass ein Verarbeiten des ausgewählten Volumens die maximale Anzahl von unbenutzten Kanalwegen beibehält.

[0007] Die US-A-5,551,003 beschreibt, wie eine Suchaffinität bei einem segmentausgerichteten, gecachten, protokollstrukturierten Array (LSA = Log Structured Array) von DASDs ansprechend auf Zugriffe bewahrt wird, die durch sequentielle Lesevorgänge und zufällige Schreibvorgänge logischer Spuren dominiert sind, die in den Segmenten gespeichert sind. Dies wird durch ein Sammeln aller schreibmodifizierten leseaktiven Spuren und reinen leseaktiven

Spuren, die entweder aus dem Cache ausgegeben oder abfallmäßig von dem LSA gesammelt werden, und ein Zurückschreiben derselben hinaus zu dem LSA als Segmente in Regionen von zusammenhängenden Segmenten von leseaktiven Spuren erreicht. Eine Abfallsammlung wird eingeleitet, wenn der erfasste freie Raum in einer Region unter eine Schwelle fällt, und geht weiter, bis die gesammelten Segmente eine zweite Schwelle überschreiten.

[0008] Die US-A-5,499,367, die den relevantesten Stand der Technik bildet, offenbart ein verteiltes Protokollsystem, bei dem die Protokolle auf einer Basis pro Client verteilt sind, wobei die Clients in Teilsätze partitioniert sind. Jedem Teilsatz von Clients ist ein Protokoll zugewiesen und die Wiedermachaufzeichnungen (Redo-Aufzeichnungen) dieser Clients sind in diesem Protokoll beibehalten. Dies reduziert Konkurrenzengpässe, weil die Anzahl von Clients, die zu irgendeinem Protokoll schreiben, begrenzt ist. Ferner sind alle Veränderungen eines Client in einem einzigen Protokoll gespeichert. Während eines Erholungsprozesses können die Veränderungen für einen Datenblock in unterschiedlichen Protokollen gespeichert sein.

Zusammenfassung der Erfindung

[0009] Es ist die Aufgabe der vorliegenden Erfindung, eine verbesserte Transaktionsprotokollverwaltung bei einem Speichersystem mit mehreren Speichermedien zu schaffen.

[0010] Diese Aufgabe wird durch ein Verfahren gemäß Anspruch 1 und durch ein System gemäß Anspruch 10 gelöst.

[0011] Gemäß Prinzipien der vorliegenden Erfindung wird bei einem bevorzugten Ausführungsbeispiel ein Transaktionsprotokoll, das in einem ersten Speicher gespeichert ist, zu einem von zwei getrennten Protokollbereichen an einem Speichersystem, das eine Mehrzahl von Speichermedien wie Plattenlaufwerken aufweist, selektiv abgelegt. Genauer gesagt tritt ein Ablegen (Posting) zu einem „Plattenprotokoll"-Bereich auf, wenn ein Seite-Voll-Status des ersten Speichertransaktionsprotokolls erfasst wird. Wenn eine Ablageanforderung auftritt, bevor ein Seite-Voll-Status des Transaktionsprotokolls erfasst wurde, dann tritt alternativ ein Ablegen vorteilhafterweise unmittelbar zu einer am wenigsten belegten Platte eines „Zwischenspeicherprotokoll"-Bereichs auf.

[0012] Gemäß weiteren Prinzipien tritt ein Ablegen zu dem „Plattenprotokoll" unter Verwendung von normalen Speichersystemverwaltungs- und Datenredundanztechniken auf. Jedoch tritt ein Ablegen zu dem „Zwischenspeicherprotokoll"-Bereich durch ein Umgehen normaler Speichersystemdatenverwaltungs- und Redundanztechniken auf, derart, dass Daten in dem Zwischenspeicherprotokollbereich nicht redundant an den Speichermedien gehalten werden. Anstelle dessen wird eine Redundanz durch die Tatsache beibehalten, dass das Transaktionsprotokoll zusätzlich zu einem Kopiertwerden zu dem Zwischenspeicherprotokollbereich bei dem ersten Speicher bleibt.

[0013] Der Zwischenspeicherprotokollbereich umfasst einen reservierten Raum auf jedem der Mehrzahl von Speichermedien und der reservierte Raum ist in logisch getrennte Abschnitte auf jedem der Speichermedien geteilt. Diese Konfiguration ermöglicht, dass Ablagen zu dem Zwischenspeicherprotokoll zwischen den reservierten Abschnitten „umgeschaltet" werden. An sich überschreiben zwei aufeinanderfolgende Zwischenspeicherprotokollablagen niemals den gleichen Abschnitt des reservierten Bereichs, ungeachtet dessen, welche Platte am wenigsten belegt ist.

[0014] Gemäß noch weiteren Prinzipien sind Sequenznummern und Plattensatznummern Datenaufzeichnungen zugewiesen, die zu den Protokollbereichen an den Speichermedien abgelegt werden. Eine Wiedergewinnung von Daten von den Plattenprotokoll- und Zwischenspeicherprotokollbereichen umfasst ein Referenzieren der Sequenznummern und Plattensatznummern, um das vollständige Transaktionsprotokoll ordnungsgemäß zu rekonstruieren.

[0015] Andere Aufgaben, Vorteile und Fähigkeiten der vorliegenden Erfindung werden ersichtlicher, wenn die Beschreibung weitergeht.

Beschreibung der Zeichnungen

[0016] Fig. 1 ist ein Blockdiagramm eines Plattenarraydatenspeichersystems, das die vorliegende Erfindung eines Plattenprotokollverfahrens mit verteiltem Schreibvorgang einsetzt.

[0017] Fig. 2 ist ein Blockdiagramm, das eine Transaktionsprotokollierung der vorliegenden Erfindung unter Verwendung verteilter Protokollschreibvorgänge zeigt.

[0018] Fig. 3–Fig. 6 sind Blockdiagramme, die das Plattenzwischenspeicherprotokoll der vorliegenden Erfindung bei verschiedenen Schnappschüssen mit der Zeit zeigen.

[0019] Fig. 7 ist ein Blockdiagramm einer Datenaufzeichnung.

[0020] Fig. 8 ist ein Blockdiagramm, das einen Prozessfluss für das Plattenprotokoll mit verteiltem Schreibvorgang zeigt.

Detaillierte Beschreibung der Erfindung

[0021] Fig. 1 zeigt ein Datenspeichersystem **10**, das das Plattenprotokollsystem und -verfahren mit verteiltem Schreibvorgang der vorliegenden Erfindung einsetzt. Bei dem gezeigten Beispiel ist das Datenspeichersystem **10** ein Plattenarraydatenspeichersystem, das ein hierarchisches Plattenarray **11** umfasst. Ein nicht hierarchisches Array (nicht gezeigt) ist gleichermaßen bei der vorliegenden Erfindung anwendbar. Das Plattenarray **11** umfasst eine Mehrzahl von Speicherplatten **12** zum Implementieren eines RAID-Speichersystems (RAID = Redundant Array of Independent Disks). Das Datenspeichersystem **10** umfasst eine Plattenarraysteuerung **14**, die mit dem Plattenarray **11** gekoppelt ist, um einen Datentransfer zu und von den Speicherplatten **12** zu koordinieren, und umfasst ferner ein RAID-Verwaltungssystem **16**. Das RAID-Verwaltungssystem **16** umfasst eine Einrichtung zum Bewirken des Plattenprotokollverfahrens mit verteiltem Schreibvorgang der vorliegenden Erfindung.

[0022] Zu Zwecken dieser Offenbarung ist eine „Platte" irgendeine nicht flüchtige, zufällig zugreifbare, wiederbeschreibbare Massenspeichervorrichtung, die die Fähigkeit zum Erfassen von eigenen Speicherausfällen derselben aufweist. Dieselbe umfasst sowohl sich drehende magnetische und optische Platten als auch Festkörperplatten oder nicht flüchtige elektronische Speicherelemente (wie beispielsweise PROMs, EPROMs und EEPROMs). Der Ausdruck „Plattenarray" ist eine Sammlung von Platten, der Hardware, die erforderlich ist, um dieselben mit einem oder mehreren Hostcomputern zu verbinden, und einer Verwaltungssoftware, die verwendet wird, um den Betrieb der physischen Platten zu steuern und dieselben der Hostbetriebsumgebung als eine oder mehrere virtuelle Platten zu präsentieren. Eine „virtuelle Platte" ist eine abstrakte Entität, die durch die Verwaltungssoftware in dem Plattenarray realisiert ist.

[0023] Der Ausdruck „RAID" bedeutet ein Plattenarray, bei dem ein Teil der physischen Speicherkapazität verwendet wird, um redundante Informationen über Benutzerdaten zu speichern, die auf dem Rest der Speicherkapazität gespeichert sind. Die redundanten Informationen ermöglichen eine Regeneration von Benutzerdaten in dem Fall, dass eine der Mitgliedsplatten des Arrays oder der Zugriffsweg zu derselben ausfällt. Eine detailliertere Erörterung von RAID-Systemen ist in einem Buch mit dem Titel The RAIDBook: A Source Book for RAID Technology, veröffentlicht am 9. Juni 1993 durch das RAID Advisory Board, Lino Lakes, Minnesota zu finden.

[0024] Obwohl ein RAID-System in Verbindung mit der vorliegenden Erfindung veranschaulicht ist, ist es offensichtlich, dass ein Nicht-RAID-System gleichermaßen anwendbar ist.

[0025] Die Plattenarraysteuerung **14** ist mit dem Plattenarray **11** über einen oder mehrere Schnittstellenbusse **13** gekoppelt, wie beispielsweise einer Kleincomputersystemschnittstelle (SCSI = Small Computer System Interface). Das RAID-Verwaltungssystem **16** ist wirksam mit der Plattenarraysteuerung **14** über ein Schnittstellenprotokoll **15** gekoppelt. Es ist zu beachten, dass das RAID-Verwaltungssystem **16** als eine getrennte Komponente verkörpert sein kann, wie es gezeigt ist (d.h. als eine Software oder Firmware), oder innerhalb der Plattenarraysteuerung **14** oder innerhalb des Hostcomputers konfiguriert sein kann, um eine Datenverwaltungseinrichtung zum Steuern von Plattenspeicher- und Zuverlässigkeitspegeln, zum Übertragen von Daten zwischen Speicherpegeln verschiedener Zuverlässigkeit und zum Implementieren einer Plattenprotokollierung mit verteiltem Schreibvorgang der vorliegenden Erfindung zu liefern. Das Datenspeichersystem **10** ist ferner mit einem Hostcomputer (nicht gezeigt) über einen I/O-Schnittstellenbus **17** gekoppelt.

[0026] Bei dem gezeigten System ist die Plattenarraysteuerung **14** als eine duale Steuerung implementiert, die aus einer Plattenarraysteuerung „A" **14a** und einer Plattenarraysteuerung „B" **14b** besteht. Die dualen Steuerungen **14a** und **14b** verbessern eine Zuverlässigkeit durch ein Liefern einer kontinuierlichen Sicherung und Redundanz in dem Fall, dass eine Steuerung funktionsunfähig wird. Die Verfahren dieser Erfindung können jedoch mit einer einzigen Steuerung oder anderen Architekturen praktiziert werden. In der Tat kann die vorliegende Erfindung bei einer Einzelsteuerungsarchitektur als besonders wertvoll betrachtet werden, bei der der Bedarf nach einem Beibehalten eines vollständigen und genauen Datenprotokolls sogar noch entscheidender als bei der Dualsteuerungsumgebung ist.

[0027] Das hierarchische Plattenarray **11** kann als unterschiedliche Speicherräume kennzeichenbar sein, einschließlich des physischen Speicherraums desselben und eines oder mehrerer virtueller Speicherräume. Die Speicherplatten **12** in dem Speicherarray **11** können beispielsweise als in einer Spiegelgruppe **18** von mehreren Platten **20** und einer Paritätsgruppe **22** von mehreren Platten **24** angeordnet konzeptioniert sein. Diese verschiedenen Speicheransichten sind durch Abbildungstechniken aufeinander bezogen. Zum Beispiel kann der physische Speicherraum des Plattenarrays in einen virtuellen Speicherraum abgebildet sein, der Speicherbereiche gemäß den verschiedenen Datenzuverlässigkeitspegeln entwirft. Einige Bereiche innerhalb des virtuellen Speicherraums können für einen Speicherpegel mit einer ersten Zuverlässigkeit zugeteilt sein, wie beispielsweise einen Spiegel- oder RAID-Pegel 1, und andere Bereiche können für einen Speicherpegel mit

einer zweiten Zuverlässigkeit zugeteilt sein, wie beispielsweise einem Paritäts- oder RAID-Pegel 5. Diese Bereiche können an der gleichen oder an getrennten Platten oder irgendeiner Kombination derselben konfiguriert sein.

[0028] Das Datenspeichersystem **10** umfasst einen Speicherabbildungsspeicher **21**, der eine beständige Speicherung der virtuellen Abbildungsinformationen liefert, die verwendet werden, um das Plattenarray **11** abzubilden. Der Speicherabbildungsspeicher befindet sich außerhalb des Plattenarrays und ist vorzugsweise in der Plattenarraysteuerung **14** resident. Die Speicherabbildungsinformationen können kontinuierlich oder periodisch durch die Steuerung **14** oder das RAID-Verwaltungssystem **16** aktualisiert werden, wenn sich die verschiedenen Abbildungskonfigurationen unter den unterschiedlichen Ansichten ändern.

[0029] Vorzugsweise ist der Speicherabbildungsspeicher **21** als zwei nicht flüchtige RAMs (RAM = Random Access Memory) **21a** und **21b** verkörpert, die in jeweiligen Steuerungen **14a** und **14b** positioniert sind. Die dualen NVRAMs **21a** und **21b** sorgen für eine redundante Speicherung der Speicherabbildungsinformationen. Die virtuellen Abbildungsinformationen werden dupliziert und in beiden NVRAMs **21a** und **21b** gemäß Spiegelredundanztechniken gespeichert. Auf diese Weise kann der NVRAM **21a** einem Speichern der ursprünglichen Abbildungsinformationen gewidmet sein und kann der NVRAM **21b** einem Speichern der redundanten Abbildungsinformationen gewidmet sein.

[0030] Wie es angegeben ist, weist das Plattenarray **11** mehrere Speicherplattenlaufwerksvorrichtungen **12** auf. Die Verwaltung einer Redundanz an den Vorrichtungen **12** ist durch das RAID-Verwaltungssystem **16** koordiniert. Wenn durch den Benutzer oder ein Hostanwendungsprogramm betrachtet, kann eine virtuelle Ansicht auf Anwendungsebenen eine einzige große Speicherkapazität darstellen, die den verfügbaren Speicherplatz auf den Speicherplatten **12** angibt. Das RAID-Verwaltungssystem **16** kann die Konfiguration der RAID-Bereiche über den physischen Speicherplatz dynamisch verändern. Folglich befinden sich die Abbildung der RAID-Bereiche in einer virtuellen Ansicht auf RAID-Ebene auf die Platten und die Abbildung einer virtuellen Vorderendansicht zu der RAID-Ansicht allgemein in einem Zustand einer Veränderung. Der Speicherabbildungsspeicher in den NVRAMs **21a** und **21b** behält die aktuellen Abbildungsinformationen, die durch das RAID-Verwaltungssystem **16** verwendet werden, um die RAID-Bereiche auf die Platten abzubilden, sowie die Informationen bei, die eingesetzt werden, um zwischen den zwei virtuellen Ansichten abzubilden. Wenn das RAID-Verwaltungssystem die RAID-Pegelabbildungen dynamisch verändert, aktualisiert dasselbe ferner die Abbildungsinformationen in dem Speicherabbildungsspeicher, um die Veränderungen wiederzuspiegeln.

[0031] Ungeachtet des RAID-Schemas oder des Datenspeicherschemas, das bei einem Plattenarray eingesetzt wird, ist jedoch klar, dass die Speicherabbildung **21** sich über eine gesamte Systemverwendung hinweg allgemein in einem ständigen Zustand einer Veränderung befindet. Somit werden Speicherabbildungsprotokollaufzeichnungen beibehalten und durch das RAID-Verwaltungssystem **16** ständig von einem Speicher zu einer Platte abgelegt, um eine Wiedergewinnung derselben in dem Fall eines Verlusts von NVRAMs **21** sicherzustellen. Folglich liefert die vorliegende Erfindung eine verbesserte Systemleistungsfähigkeit für ein Plattenprotokollschreiben durch ein Verwalten und Verteilen bestimmter der Protokollschreibvorgänge zu irgendeiner am wenigsten belegten Platte, die aus den mehreren verfügbaren Daten **12** ausgewählt ist, wobei so eine Konkurrenz für Plattenzugriffe zwischen Protokoll-I/Os und anderen im Ablauf befindlichen I/Os reduziert wird. Im Allgemeinen wird dies durch ein Reservieren eines „Zischenspeicherprotokoll"-Bereichs auf jeder Platte **12** für ein Halten des jüngsten Abschnitts des Protokolls erzielt. Falls dann eine Ablageanforderung auftritt, bevor eine Seite des Transaktionsprotokollspeichers voll ist, dann tritt die Ablage im Wesentlichen unmittelbar zu dem reservierten „Zwischenspeicherprotokoll"-Bereich der Platte auf, die am wenigsten belegt ist. Wenn nachfolgend eine Protokollwiedergewinnung erforderlich ist, werden die Fragmente von allen Platten **12** zu einem einzigen vollständigen Bild zusammengelegt.

[0032] Unter jetziger Bezugnahme auf Fig. 2 zeigt ein Blockdiagramm die Aspekte einer Plattenprotokollierung mit verteiltem Schreibvorgang der vorliegenden Erfindung. Eine NVRAM-Abbildung **45** stellt einen Teilsatz des nicht flüchtigen Speichers **21** an der Steuerung **14a**/**14b** (Fig. 1) dar, in dem Daten für eine Verwendung in Verbindung mit dem Plattenspeichersystem **10** gespeichert sind. Eine Plattenabbildung **50** ist an dem Array **11** (redundant) resident und ist ein herkömmliches Plattenabbildungsbild der NVRAM-Abbildung **45**. Ein regelmäßiges Speichern (Ablegen) der NVRAM-Abbildung **45** zu der Plattenabbildung **50** liefert eine Einrichtung zum Halten einer redundanten Kopie der Inhalte der NVRAM-Abbildung **45** auf einer Platte zu Fehlererholungszwecken. Allgemein tritt ein Ablegen der NVRAM-Abbildung **45** zu der Plattenabbildung **50** als ein Hintergrundprozess (durch das RAID-Verwaltungssystem **16** gesteuert) auf, wenn es eine normale Systemverarbeitung und I/O-Konkurrenzen erlauben. Als solches unterliegt das Ablegen von Daten der NVRAM-Abbildung **45** zu der Plattenabbildung **50** normalen Systemkonkurrenzen für einen I/O-Zugriff und einen Plattenraum und trägt deshalb ungeachtet dessen, wann ein Ablegen tatsächlich auftritt, ein Unsi-

cherheitselement.

[0033] Ein RAM-Protokollbild (RLI = RAM Log Image) **55** ist ebenfalls ein Teilsatz des Speichers **21** oder dasselbe kann einer getrennter Speicher sein (aber vorzugsweise nicht flüchtig). Das RLI **55** wird zum schnellen Speichern/Aufzeichnen von inkrementalen Veränderungen verwendet, die bei der NVRAM-Abbildung **45** auftreten. Bei einem bevorzugten Ausführungsbeispiel umfasst das RLI **55** sechzehn (als N gezeigt) adressierbare Seiten mit 64 Kbyte, obwohl andere Konfigurationen ebenfalls machbar sind. Die inkrementalen Veränderungen, die in dem RLI **55** gespeichert sind, werden nachfolgend zu einem Plattenprotokoll **60** oder einem Plattenzwischenspeicherprotokoll **65** abgelegt, wenn durch das RAID-Verwaltungssystem **16** angefordert wird, dies zu tun.

[0034] Mehrere Faktoren können bewirken, dass die RAID-Verwaltung **16** eine Ablageanforderung von Daten von dem RLI **55** zu einem Plattenprotokoll **60** oder einem Plattenzwischenspeicherprotokoll **65** einleitet. Beispielsweise wird bei einem bevorzugten Ausführungsbeispiel durch einen RLI-„Seite-Voll"-Status, der eingehalten wird, eine „Spülen"-Ablageanforderung zu dem Plattenprotokoll **60** eingeleitet. Wenn eine „Spülen"-Ablage auftritt, wird eine ganze Seite von Transaktionsprotokolldaten von dem RLI **55** zu dem Plattenprotokoll **60** geschrieben. Eine „Zwangs"-Ablageanforderung zu einem Plattenzwischenspeicherprotokoll **65** jedoch wird durch (i) ein zeitbasiertes Frequenzerfordernis, das eingehalten wird, oder (ii) eine spezifische Hostanforderung, die empfangen wird, eingeleitet. Wenn eine „Zwangs"-Ablage auftritt, wird einer oder werden mehrere ganze Blöcke von Transaktionsprotokolldaten von dem RLI **55** zu dem Plattenzwischenspeicherprotokoll **65** geschrieben. Der eine oder die mehreren Blöcke umfasst oder umfassen diese Blöcke (innerhalb der aktuellen Seite), die vorhergehend nicht vollständig geschrieben wurden und die eine oder mehrere Transaktionsprotokollaufzeichnungen enthalten, die nicht geschrieben wurden. Die Seite, die „spülmäßig" zu dem Plattenprotokoll **60** geschrieben wird, und die Blöcke, die „zwangsweise" zu dem Plattenzwischenspeicherprotokoll **65** geschrieben werden, werden hierin als die „ungeschriebenen" Daten des RLI **55** bezeichnet (obwohl eine „spülmäßig" geschriebene Seite einige Aufzeichnungen enthalten kann, die vorhergehend „zwangsweise" zu dem Zwischenspeicherprotokoll **65** geschrieben wurden). In jedem Fall stellen derartige Ablagen (entweder zu dem Plattenprotokoll **60** oder dem Plattenzwischenspeicherprotokoll **65**) sicher, dass Veränderungen bei der NVRAM-Abbildung **65** (in dem RLI **55** gefangen) zu Fehlererholungszwecken in dem Fall eines Verlusts eines NVRAM **21**, wenn die Plattenabbildung **50** nicht aktualisiert wurde, zu dem Plattenarray **11** gespeichert werden.

[0035] Das Plattenprotokoll **60** ist an dem Array **11** (Fig. 1) resident und ist ein herkömmliches Plattenbild des RLI **55**. Das Plattenprotokoll **60** ist vorzugsweise zum Speichern mehrerer Seiten von Daten in der Lage, ähnlich dem RLI **55**. Wie es gezeigt ist, ist das Plattenprotokoll **60** mit N Seiten zum Speichern von Daten etikettiert und kann, wie es auf dem Gebiet herkömmlich ist, zusammenhängend oder kreisförmig verbunden sein. Das Plattenprotokoll **60** ist an dem Plattenarray **11** unter Verwendung normaler Datenredundanzschemata des Plattenspeichersystems **10** (von Fig. 1) gespeichert und verwaltet. An sich tritt eine „Spül"-Ablage der „ungeschriebenen" Inhalte des RLI **55** zu dem Plattenprotokoll **60** unter normalen I/O-Umständen auf und unterliegt System-I/O-Konkurrenzen, um eine Plattenzugriff und einem Raum. Obwohl das Plattenprotokoll **60** allgemein häufiger als das Plattenprotokoll **50** aktualisiert wird, hält dasselbe lediglich die inkrementalen Veränderungen an der NVRAM-Abbildung **45** (in dem RLI **55** gefangen), während die Plattenabbildung **50** ein vollständiges Bild der NVRAM-Abbildung **45** (zum Zeitpunkt der letzten Aktualisierung) hält.

[0036] Das Plattenzwischenspeicherprotokoll **65** umfasst reservierte Zwischenspeicherbereiche **70**, **75**, **80**, **85**, **90**, **95**, **100** und **105** (als **70**–**105** bezeichnet), die Abschnitte der Platten des Plattenarrays **11** (Fig. 1) darstellen. Wie es erwähnt ist, wird bei einem bevorzugten Ausführungsbeispiel das Plattenzwischenspeicherprotokoll **65** verwendet, um die Inhalte des RLI **55** auf andere spezifizierte Ereignisse oder Zeiten als einen „Seite-Voll"-Status hin zu speichern. Dieses Kriterium zum Ablegen ist jedoch flexibel bei Systementwurfsveränderungen und/oder Benutzeraufhebungsbetrachtungen, wie es Durchschnittsfachleuten auf dem Gebiet offensichtlich wäre. Wenn ein spezifiziertes Ereignis auftritt (ein Anderes als ein „Seite-Voll"-Status), wie es durch das RAID-Verwaltungssystem **16** angefordert wird, legt das RLI **55** die „ungeschriebenen" Inhalte desselben „zwangsweise" zu irgendeiner der Platten 1–M des Plattenzwischenspeicherprotokolls **65** ab, je nachdem welche Platte am wenigsten belegt ist. Eine am wenigsten belegte Platte wird durch ein Überwachen einer I/O-Aktivität der Platten 1–M des Arrays **11** erfasst.

[0037] Ein „Zwang" des RLI **55** zu der am wenigsten belegten Platte bewirkt einen verteilten Schreibvorgang des Transaktionsprotokolls über das Plattenarrays mit der Zeit. Dies steht im Gegensatz zu einem Seite-Voll-„Spülen" des RLI **55** zu einem gegebenen, einzigen Plattenprotokoll **60**. Obwohl das Plattenprotokoll **60** tatsächlich in dem Fall, dass ein Paritätsredundanzschema verwendet wird, über mehrere Platten ausgebreitet wird, ist dasselbe im Wesentlichen ein „einziges" oder „nicht verteiltes" Plattenprotokoll, weil lediglich eine Basisadresse an einem einzigen Plattenlaufwerk benötigt/verwendet wird, um das gesamte Protokoll (ohne Betrachtung irgendeiner red-

undanten Kopie) zu adressieren/auf dasselbe zuzugreifen.

[0038] Vorteilhafterweise tritt eine „Zwangs"-Ablage bei einer reduzierten I/O-Konkurrenz (relativ zu einer anderen ablaufenden System-Lese/Schreib-I/O-Aktivität) auf, weil die am wenigsten belegte Platte ausgewählt ist. Ungleich einem Ablegen zu der Plattenabbildung **50** oder dem Plattenprotokoll **60** ist somit diesem Ablegen mit verteiltem Schreibvorgang zu dem Plattenzwischenspeicherprotokoll **65** allgemein ein unmittelbarer (oder zumindest schnellerer) Abschluss zugesichert. Zusätzlich ist eine „Zwangs"-Ablage allgemein schneller als eine „Spül"-Ablage, weil lediglich eine minimale Anzahl von ungeschriebenen Blöcken übertragen wird.

[0039] Im Gegensatz zu dem Plattenprotokoll **60** hält das Plattenzwischenspeicherprotokoll **65** die inkrementalen Veränderungen, die in dem RLI **55** notiert sind, in einer verteilten, nicht redundanten Weise über das Plattenarray **11**. Dasselbe ist nicht redundant, weil die Schreibvorgänge, die zu dem Plattenzwischenspeicherprotokoll **65** auftreten, von den normalen Redundanzschemata des RAID-Verwaltungssystems **16** ausgenommen sind. Somit tritt bei einem Ablegen zu dem Plattenzwischenspeicherprotokoll **65** zumindest ein I/O-Schritt weniger relativ zu dem Plattenprotokoll **60** auf. Zusätzlich zu einem Kopiertwerden zu dem Zwischenspeicherprotokollbereich, ist eine Redundanz nach einer „Zwangs"-Ablage durch die Tatsache beibehalten, dass das Transaktionsprotokoll bei dem ersten Speicher bleibt.

[0040] Bei einem bevorzugten Ausführungsbeispiel weist jede Platte 1–M in dem Array **11** eine zweckgebundene Menge an Raum auf, die zum Speichern des verteilten Protokolls reserviert ist. Zum Beispiel sind an jeder Platte in der gezeigten Darstellung zwei Seiten mit 64 Kbyte **70**/**75**, **80**/**85**, **90**/**95** und **100**/**105** reserviert. Zumindest zwei Seiten sind an jeder Platte reserviert, um ein mögliches Überschreiben (und einen Verlust) von gültigen Daten in dem Fall eines gewissen Ausfalls während des Plattenzwischenspeicherprotokollablageprozesses zu vermeiden. Genauer gesagt, schreibt (legt ab/zwingt) das RLI **55** zu Seiten in dem Plattenzwischenspeicherprotokoll **65** auf eine abwechselnd (austauschend oder umschaltend) gerade/ungerade Weise. Auf einen ersten Schreibvorgang hin kann beispielsweise das RLI **55** zu einer gerade nummerierten, reservierten Seite **70**, **80**, **90** oder **100** auf welcher Platte auch immer am wenigsten belegt ist ablegen. Dann legt bei einem nächsten Schreibvorgang, der auftritt, das RLI **55** zu der ungerade nummerierten Seite **75**, **85**, **95**, oder **105** von welcher Platte auch immer am wenigsten belegt ist ab. Auf diese Weise ist dem System ein weiterer Pegel einer Datenintegrität zugesichert und dasselbe vermeidet ein mögliches Überschreiben (d.h. in dem Fall, dass die gleiche am wenigsten belegte Platte aufeinanderfolgend ausgewählt wird) von jüngst abgelegten Daten während eines nächsten aufeinanderfolgenden Ablegens, das auftritt.

[0041] Unter jetziger Bezugnahme auf Fig. 3–Fig. 6 zeigen diese Blockdiagramme einen Abschnitt einer Seite **57** des RLI **55** und einen Abschnitt jeder Seite **70**–**105** des Plattenzwischenspeicherprotokolls **65**, um exemplarische Plattenzwischenspeicheraktivitäten mit verteiltem Schreibvorgang unter der vorliegenden Erfindung weiter zu detaillieren. Genauer gesagt, sind in jeder Fig. 3–Fig. 6 unterschiedliche zeitliche Schnappschüsse des Status des Plattenzwischenspeicherprotokolls **65** ansprechend auf getrennte Ablagen von dem RLI **55** gezeigt. Die Seite **57** des RLI **55** und jede Zwischenspeicherprotokollseite **70**–**105** sind als durch die gestrichelten Linien (logisch) in drei Blöcke (oder Sektoren) B1, B2 und B3 von 512 Byte geteilt gezeigt. Zu Klarheitszwecken und für eine einfache Erörterung sind lediglich drei Blöcke anstelle aller Blöcke in jeder Seite mit 64 Kbyte gezeigt. Die Protokollbildseite **57** (des RLI **55**) ist hierin und an dem Diagramm als „LI" (Log Image) bezeichnet. Zusätzlich ist jede Platte in dem Plattenzwischenspeicherprotokoll **65** jeweils als „D1–DM" (D = Disk) bezeichnet und jede der dualen Seiten, die in jeder Platte reserviert sind, ist als „P1" bzw. „P2" (P = Page) bezeichnet.

[0042] Unter jetziger Bezugnahme auf Fig. 3 spiegelt ein logischer Markierer T1 einen gegebenen Zeitpunkt wieder, wenn ein spezifiziertes Ereignis auftritt (wie es durch das RAID-Verwaltungssystem **16**, Fig. 1 angefordert wird), um eine „Zwangs"-Ablage der ungeschriebenen Daten der Seite **57** des RLI **55** zu dem Plattenzwischenspeicherprotokoll **65** einzuleiten. T1 identifiziert ferner eine Position, die angibt, wie „voll" die RLI-Seite **57** mit Protokolldaten zu diesem gegebenen Zeitpunkt ist. Wenn eine Ablage angefordert wird, werden volle Blöcke von ungeschriebenen Daten (in der Seite **57** des RLI **55**) abgelegt, wie es durch den logischen Markierer T1 identifiziert ist. Volle Blöcke werden abgelegt, da ein Datenblock mit 512 Byte die minimale Ablagegröße ist (gemäß beliebigen Systementwurfserfordernissen bei diesem Beispiel).

[0043] Auf das Auftreten des Ereignisses/der Zeit T1 hin legt somit beispielsweise das RLI **55** die „ungeschriebenen" Inhalte desselben (wie durch die Position T1 angegeben) in der Seite **57** zu einem der Plattenzwischenspeicherbereiche **70**–**105** der am wenigsten belegten Platte 1–M des Plattenzwischenspeicherprotokolls **65** ab – abhängig von dem hierin im Folgenden ausführlicher erörterten „Umschalt"-Aspekt. Genauer gesagt, wird ein Block „Eins" der Protokollbildseite **57** (LIB1) in seiner Gesamtheit abgelegt, weil derselbe „ungeschrieben" und vollständig voll ist, und ist ein Block „2" des Protokillbilds (LIB2) ebenfalls „ungeschrieben" und wird

somit ebenfalls in seiner Gesamtheit abgelegt (obwohl Protokolldaten LIB2 lediglich teilweise füllen, d.h. bis zu dem Ereignis/der Zeit T1). Unter (beliebiger) Annahme, dass die Platte 2 als am wenigsten belegt erfasst wird und dass eine Ablage mit einer geraden Seitennummer in dem Plattenzwischenspeicherprotokoll **65** beginnt, dann legt die Seite **57** die Blockinhalte LIB1 und LIB2 derselben zu entsprechenden Blöcken B1 und B2 der Seite P2 (**80**) der Platte D2 (d.h. D2P2B1 und D2P2B2) des Plattenzwischenspeicherprotokolls **65** ab. Der Block D2P2B1 enthält somit alle gültigen Daten (in horizontalen Umkehrdarstellungslinien gezeigt) und der Block D2P2B2 enthält teilweise gültige Daten bis zu der Zeit/dem Markierer T1 und der Rest des Blocks D2P2B2 enthält ungültige Daten oder „bedeutungslose" Daten (in Kreuzschraffierung gezeigt).

[0044] Mit jetziger Bezugname auf Fig. 4 identifiziert ein zweites Ereignis/eine zweite Zeit T2, zu der das RAID-Verwaltungssystem **16** erneut anfordert, dass das RLI **55** die Daten desselben ablegt. In dieser Instanz müssen die Daten der Protokollbildseite **57**, die zwischen der Zeit T1 und T2 gespeichert werden (d.h. die „ungeschriebenen" Daten), zu dem Plattenzwischenspeicherprotokoll **65** abgelegt werden (da noch kein Seite-Voll-Status erreicht wurde). (Falls jedoch die ganze Seite **57** vor dem Ereignis/der Zeit T2 mit Transaktionsdaten gefüllt wäre, dann würde das RLI **55** die gesamte Seite **57** zu dem Plattenprotokoll **60** ablegen, anstelle eines Ablegens lediglich eines Teils zu dem Plattenzwischenspeicherprotokoll **65**). Angenommen, dass die Platte 1 nun die am wenigsten belegte Platte ist, und mit der Kenntnis, dass eine Schreib-I/O lediglich bei vollen Blockgrößen auftritt, dann wird LIB2 ganz zu D1P1B2 geschrieben. Die ungerade Seite P1 (**75**) wird dieses Mal geschrieben, um die vorhergehend beschriebene Datenschutztechnik eines Seiten-„Umschaltens" (Vertauschens) aufzunehmen. Die ungültigen Daten (d.h. die Daten jenseits des spezifizierten Zeitmarkierers T2, die sich innerhalb der Blockgröße befinden) sind erneut in Kreuzschraffierung gezeigt.

[0045] Fig. 5 zeigt noch ein drittes Ereignis/eine dritte Zeit T3, bei der von dem RLI **55** erneut angefordert wird, die Daten desselben abzulegen, bevor ein Seite-Voll-Status erreicht wurde. In dieser Instanz müssen die Daten der Protokollbildseite **57**, die zwischen der Zeit T2 und T3 gespeichert werden (die „ungeschriebenen" Daten), abgelegt werden.

[0046] Angenommen, dass bei dieser Instanz die Platte 3 (D3) am wenigsten belegt ist, dann wird an sich LIB2 ganz zu D3P2B2 abgelegt und wird LIB3 ganz zu D3P2B3 abgelegt. Um ein Seitenvertauschen aufzunehmen, wird dieses Mal erneut die „gerade" Seite P2 (**90**) geschrieben.

[0047] Fig. 6 zeigt ein viertes Ereignis/eine vierte Zeit T4, bei der von dem RLI **55** erneut angefordert wird, die Daten desselben abzulegen, bevor ein Seite-Voll-Status erreicht ist. In dieser Instanz müssen die „ungeschriebenen" Daten der Protokollbildseite **57**, die zwischen der Zeit T3 und T4 gespeichert werden, abgelegt werden. Angenommen, dass bei dieser Instanz die Platte 1 (D1) erneut am wenigsten belegt ist, wird an sich LIB3 ganz zu D1P1B3 zwangsweise abgelegt.

[0048] Wie es mit Bezug auf Fig. 3–Fig. 6 zu sehen ist, ist die Gesamtsystem-I/O-Leistungsfähigkeitsauswirkung nicht nur reduziert, weil zu der am wenigsten belegten Platte geschrieben wird, sondern auch, weil keine redundanten Schreibvorgänge zu einer Platte oder zu Platten auftreten. Anstelle dessen ist eine Redundanz durch die Tatsache beibehalten, dass die Protokolldaten auf eine Platte (das Plattenzwischenspeicherprotokoll **65**) geschrieben werden und dennoch auch in dem RLI **55** bleiben. Zusätzlich ist erneut zu beachten, dass bei einem bevorzugten Ausführungsbeispiel ein „Zwang" zu dem Plattenzwischenspeicherprotokoll **65** für Ereignisse auftritt, die vor einem Seite-Voll-Status für das RLI **55** auftreten, und ein „Spülen" von dem RLI **55** zu dem Plattenprotokoll **60** (Fig. 2) in dem Fall auftritt, dass ein Seite-Voll-Status für das RLI **55** erfasst wird.

[0049] Unter jetziger Bezugnahme auf Fig. 7 ist eine Aufzeichnung **110** von Daten, die in Verbindung mit dem verteilten Schreiben der vorliegenden Erfindung verwendet werden, in einem Blockdiagramm gezeigt. Jeder 512-Byte-Block (Sektor) von Daten in dem RLI **55** (und dem Plattenprotokoll **60** und dem Plattenzwischenspeicherprotokoll **65**) weist eine oder mehrere Aufzeichnungen **110** auf und eine Aufzeichnung **110** kann Blockgrenzen übertreten. Zu Darstellungszwecken ist die Aufzeichnung **110** in vereinfachter Form gezeigt, d.h. es sind nicht alle Felder gezeigt, die bei der Aufzeichnung verwendet werden können. Dennoch umfasst die Aufzeichnung **110** zumindest einen Längenindikator **115** zum Identifizieren der Aufzeichnungslänge, eine Sequenznummer **120** zum Identifizieren einer Sequenzierung von Aufzeichnungen zum Zurückspeichern der Daten von dem Plattenzwischenspeicherprotokoll **65**, einen Plattensatzidentifizierer **125** zum Identifizieren des Plattensatzes, der dem Transaktionsprotokoll zugeordnet ist, einen Körper **30** zum Halten der tatsächlichen Protokolldaten, die gespeichert sind, und eine Prüfsumme **135** zu Datenvalidierungszwecken.

[0050] Die Sequenznummer **120** ist eine erzeugte Nummer, die sequentiell für jede neue Aufzeichnung inkrementiert wird, die zu der Transaktionsdatei hinzugefügt wird. Die Prüfsumme **135** ist eine Prüfsumme der ganzen Aufzeichnung **110** und wird verwendet, um den Status der Aufzeichnung während einer Transaktionsprotokollwiedergewinnung zu validieren. Der Plattensatzidentifizierer **125** ist ein beliebi-

ger Identifizierer der aktuellen Instanz des Plattensatzes, die dem RLI **55** zugeordnet ist, und wird verwendet, um sicherzustellen, dass „veraltete" (d.h. ungültige) Daten des Zwischenspeicherprotokolls **65** nicht während einer Transaktionsprotokollwiedergewinnung verwendet werden. Während einer Wiedergewinnung wird nämlich eine Aufzeichnung als gültig erkannt, falls der Plattensatzidentifizierer **125** derselben mit der aktuellen Instanz des Plattensatzes übereinstimmt, aber die Aufzeichnung wird als ungültig erkannt, falls der Plattensatzidentifizierer **125** derselben nicht mit aktuellen der Instanz des Plattensatzes übereinstimmt. Eine veraltete Aufzeichnung oder veraltete Aufzeichnungen können beispielsweise auftreten, falls ein Plattenlaufwerk von einem anderen Plattensatz eingetauscht wird. Falls dem so ist, ermöglicht der Plattensatzidentifizierer, der jeder Aufzeichnung zugeordnet ist, dass der Transaktionsprotokollwiedergewinnungsprozess irgendwelche veralteten Daten, die dieser neuen Platte zugeordnet sind, erkennt und nicht verwendet. Einfach gesagt, muss der Plattensatzidentifizierer der Aufzeichnung mit der aktuellen Plattensatzinstanz übereinstimmen.

[0051] Unter jetziger Bezugnahme auf Fig. 8 zeigt ein Blockflussdiagramm die Beziehung von Prozessen untereinander, die innerhalb des RAID-Verwaltungssystems **16** (von Fig. 1) verkörpert sind, zum Verwalten von Protokolltransaktionen der vorliegenden Erfindung. Diese Prozesse sind vorzugsweise in einer Firmware implementiert. Wenn eine Anwendung **150** (zum Beispiel das RAID-Verwaltungssystem **16** von Fig. 1) die NVRAM-Abbildung **45** (Fig. 2) manipuliert, wird eine Aufzeichnung **110**, die die Manipulationsaktivität identifiziert, zu der Steuerung der Protokollverwaltungseinrichtung **155** hinzugefügt und in dem RAM-Protokollbild **55** (Fig. 2) gespeichert. Aufzeichnungen werden kontinuierlich hinzugefügt, bis eines von mehreren Schlüsselereignissen auftritt. Falls die aktuelle Seite in dem RLI **55** voll wird, dann „spült" die Protokollverwaltungseinrichtung **155** die ganze Seite desselben durch ein Übertragen einer Steuerung zu einer Datenverwaltungseinrichtung **160**, die wiederum eine Schnittstelle mit einem Plattentreiber **165** zum redundanten Ablegen der vollständigen Seiteninhalte zu dem Plattenprotokoll **60** des Plattenarrays **11** bildet. Falls die aktuelle Seite in dem RLI **55** nicht voll ist, aber die Protokollverwaltungseinrichtung **155** erfasst, dass das Ereignis eine Anforderung ist, um eine Ablage zu dem Plattenzwischenspeicherbereich **65** des Arrays **11** zu „erzwingen", dann umgeht die Protokollverwaltungseinrichtung **155** die Datenverwaltungseinrichtung **160** und bildet direkt eine Schnittstelle mit dem Plattentreiber **165**, um die Daten „zwangsweise" zu dem Array abzulegen. Es wird keine redundante Kopie geschrieben, nachdem eine „Zwangs"-Ablage auftritt.

[0052] Unter jetziger Bezugnahme auf alle Figuren treten in dem Fall eines Bedarfs nach einer Protokollwiedergewinnung auf Grund eines gewissen Speicher- oder Systemausfalls mehrere Schritte auf, um die inkrementalen Protokolldaten, die in dem Plattenprotokoll **60** und dem Plattenzwischenspeicherprotokoll **65** gespeichert sind, wiederzugewinnen. Zuerst werden alle vollen Seiten des Plattenprotokolls **60** zu dem RLI **55** kopiert, um soviel wie möglich der Protokolldaten zu rekonstruieren. Es ist jedoch möglich, dass eine nicht volle Seite, wie beispielsweise die Seite **59**, Daten in dem Plattenzwischenspeicherprotokoll **65** aufweist, die ebenfalls zu dem RLI **55** kopiert werden müssen. Somit müssen die Fragmente von Protokolldaten von allen Seiten **70**–**105** des Plattenzwischenspeicherprotokolls **65** auf den Platten 1–M zu einem einzigen vollständigen Bild zum Kopieren zu dem RLI **55** zusammengelegt werden.

[0053] In Vorbereitung auf diese Wiedergewinnung von dem Plattenzwischenspeicherprotokoll **65** wird das RLI **55** abgetastet, um die Aufzeichnung mit der Sequenznummer **120** zu finden, die die letzte (jüngste) Aufzeichnung angibt, die zu dem Plattenprotokoll **60** geschrieben ist. Diese Abtastung muss sowohl eine Kreisförmigkeit des Protokolls und eine Umhüllung der Sequenznummern berücksichtigen. Die nächste aufeinanderfolgende Sequenznummer (Aufzeichnung) nach der als letztes Geschriebenen muss, falls überhaupt, in dem Plattenzwischenspeicherprotokoll **65** gefunden werden. Folglich wird dann das Plattenzwischenspeicherprotokoll **65** abgetastet, um die Aufzeichnung mit der nächsten aufeinanderfolgenden Sequenznummer zu finden (die die nächste Aufzeichnung angibt, die zu dem RLI **55** zurückgespeichert werden soll). Sobald die nächste Aufzeichnung (Sequenznummer) in dem Plattenzwischenspeicherprotokoll **65** gefunden ist, wird der Plattensatzidentifizierer **125** überprüft, um festzustellen, dass die Aufzeichnung zu der aktuellen Instanz des Plattensatzes gehört. Zusätzlich wird die Prüfsumme **135** der Aufzeichnung ausgewertet, um die Integrität der Aufzeichnung zu bestimmen. Falls alles gut ist, dann wird diese Aufzeichnung zu dem RLI **55** kopiert, um mit dem Transaktionsprotokollwiedergewinnungsprozess fortzufahren. Bei dem Beispiel von Fig. 6 genügt die erste Aufzeichnung in dem Block D2P2B1 diesen Wiedergewinnungskriterien des ersten Schritts.

[0054] Nachfolgend wird dann das Plattenzwischenspeicherprotokoll **65** erneut nach einer nächsten aufeinanderfolgenden Aufzeichnung abgetastet, die in einer inkrementalen Sequenznummernreihenfolge folgt. Die bekannte Länge **115** der vorhergehend gefundenen Aufzeichnung vorausgesetzt, ist bekannt, dass die nächste Aufzeichnung mit dem Versatz beginnt, der durch die Länge **115** der vorhergehend gefundenen Aufzeichnung beschrieben ist. An sich werden alle Plattenzwischenspeicherprotokollbereiche **65** bei dieser Versatzadresse durchsucht, bis die nächste Aufzeichnung gefunden ist, die

bei diesem Versatz beginnt, die die ordnungsgemäße nächste aufeinanderfolgende Sequenznummer **120** aufweist, die den ordnungsgemäßen Plattensatzidentifizierer **125** aufweist und die die Gültigkeitsanalyse der Prüfsumme **135** erfüllt. Bei dem Beispiel von Fig. 6 genügt die zweite Aufzeichnung in dem Block D2P2B1 (nicht sichtbar unterschieden) diesen Wiedergewinnungskriterien. Wenn dieselbe einmal gefunden ist, wird die Aufzeichnung zu dem RLI **55** kopiert.

[0055] Allgemein gesagt, wird dieser gesamte Prozess eines (i) Findens der nächsten aufeinanderfolgenden Aufzeichnung (durch die nächste aufeinanderfolgende inkrementale Sequenznummer identifiziert), (ii) eines Verifizierens des Plattensatzidentifizierers und (iii) eines Verifizierens der Prüfsumme durch das ganze Zwischenspeicherprotokoll **65** kontinuierlich wiederholt, bis alle Aufzeichnungen wiedergewonnen sind, die alle diese Wiedergewinnungskriterien erfüllen. Um beispielsweise mit Bezug auf Fig. 6 weiter zu veranschaulichen, wird jede der gültigen Aufzeichnungen, die in D2P2B1 und D2P2B2 identifiziert sind (in horizontalen Umkehrdarstellungslinien bezeichnet) zuerst in dieser Reihenfolge zu dem RLI **55** zurückgespeichert. Dann genügen irgendwelche der gültigen Aufzeichnungen, die in D1P1B2 oder D3P2D2 gefunden werden, den nächsten Wiedergewinnungsschritten. Falls beispielsweise die nächste gültige Aufzeichnung zuerst in D1P1B2 vor D3P2B2 gefunden würde, dann würden alle gültigen Aufzeichnungen in D1P1B2 (sequenziell eine zu einer Zeit) zu dem RLI **55** kopiert und der Rest der gültigen Aufzeichnungen, die nicht in D1P1B2 gefunden würden, würde nachfolgend in D3P2B2 gefunden und würde die gültigen Aufzeichnungen von D3P2B3 hindurch weitergehen. Falls jedoch die nächste gültige Aufzeichnung zuerst in D3P2B2 (und nicht D1P1B2) gefunden würde, würden alle diese gültigen Aufzeichnungen zu dem RLI **55** kopiert und würden dann die gültigen Aufzeichnungen von D3P2B3 zu dem RLI **55** kopiert. Erst dann würden nachfolgend die letzten gültigen Aufzeichnungen, die in D1P1B3 gefunden werden, verarbeitet und zu dem RLI **55** kopiert. Es ist bekannt, dass die letzte Aufzeichnung von dem Plattenzwischenspeicherprotokoll **65** wiedergewonnen wurde, wenn keine andere nächste aufeinanderfolgende Sequenznummer in irgendeiner anderen Aufzeichnung in dem Plattenzwischenspeicherprotokoll **65** gefunden wird.

[0056] Für die dargestellten Beispiele ist die Protokollwiedergewinnung für das Plattenprotokoll **60** und das Plattenzwischenspeicherprotokoll **65** zu dem RLI **55** nun abgeschlossen. Unter erneuter Bezugnahme auf Fig. 8 gibt als solches die Protokollverwaltungseinrichtung **155** die wiedergewonnenen Aufzeichnungen (nun in dem RLI **55** zu finden) zurück zu der Steuerung der Anwendung **150** (RAID-Verwaltungssystem **16**, Fig. 1), wobei angegeben wird, dass eine Protokollwiedergewinnung abgeschlossen ist und dass die Anwendung **150** nun fortfahren kann, die Protokollveränderungen zu bewirken, die in dem RLI **55** notiert sind, um die NVRAM-Abbildung **45** zurück in einen Zustand zu versetzen, der vor dem Systemfehler/-Ausfall existierte, der den Transaktionsprotokollwiedergewinnungsprozess einleitete.

[0057] Was oben beschrieben wurde, sind die bevorzugten Ausführungsbeispiele eines Verfahrens und einer Vorrichtung zum Verbessern einer Plattenprotokollschreibleistungsfähigkeit unter Verwendung verteilter Schreiboperationen über mehrere Platten in einem Plattenarray. Für einen Durchschnittsfachmann auf dem Gebiet ist offensichtlich, dass die vorliegende Erfindung ohne weiteres unter Verwendung von Irgendeiner von einer Vielfalt von Software, Firmware und/oder Hardware-Komponenten implementiert werden kann, die auf dem Gebiet existieren. Während die vorliegende Erfindung durch Bezugnahme auf spezifische Ausführungsbeispiele beschrieben wurde, ist zudem ersichtlich, dass andere alternative Ausführungsbeispiele und Verfahren einer Implementierung oder Modifikation eingesetzt werden können, ohne von dem echten Schutzbereich der Erfindung abzuweichen, wie derselbe durch die Ansprüche definiert ist.

Patentansprüche

1. Ein Verfahren zum Schreiben von Transaktionsprotokolldaten, die in einem ersten Speicher (**55**) eines Speichersystems (**10**) gespeichert sind, das mehrere Speichermedien (**12**) aufweist, zu getrennten Protokollbereichen (**60**, **65**) des Speichersystems (**10**), wobei die getrennten Protokollbereiche (**60**, **65**) einen Plattenprotokollbereich (**60**) in dem Speichersystem (**12**) und einen Zwischenspeicherprotokollbereich (**65**) aufweisen, der durch Zwischenspeicherbereiche (**70**–**105**) gebildet ist, die Abschnitte des Speichermediums (**12**) des Speichersystems (**10**) darstellen, wobei das Verfahren folgende Schritte aufweist:
(a) falls ein Seite-Voll-Status des ersten Speichers (**55**) erfasst wird, Schreiben von Transaktionsprotokolldaten von dem ersten Speicher (**55**) zu dem Plattenprotokollbereich (**60**); und
(b) falls eine Ablageanforderung für den ersten Speicher (**55**) erfasst wird, Schreiben von Transaktionsprotokolldaten von dem ersten Speicher (**55**) zu einem Zwischenspeicherbereich (**70**–**105**) des am wenigsten belegten Speichermediums (**12**).

2. Das Verfahren gemäß Anspruch 1, bei dem die mehreren Speichermedien (**12**) Direktzugriffsspeichermedien sind.

3. Das Verfahren gemäß Anspruch 1 oder 2, bei dem das am wenigsten belegte Speichermedium basierend auf einer Eingabe/Ausgabe-Aktivität (I/O-Ak-

tivität) ausgewählt wird.

4. Das Verfahren gemäß einem der Ansprüche 1 bis 3, bei dem eine volle Seite von Transaktionsprotokolldaten zu dem Plattenprotokollbereich (**60**) geschrieben wird, und wobei einer oder mehrere volle Blöcke von Transaktionsprotokolldaten zu dem Zwischenspeicherprotokollbereich (**65**) geschrieben werden.

5. Das Verfahren gemäß einem der Ansprüche 1 bis 4, bei dem die Ablageanforderung auftritt, bevor erfasst wird, dass ein bezeichneter Abschnitt (**57**) des ersten Speichers (**55**) mit den Daten voll ist.

6. Das Verfahren gemäß einem der Ansprüche 1 bis 5, bei dem die Transaktionsprotokolldaten nicht redundant zu dem ausgewählten, am wenigsten belegten Speichermedium geschrieben werden.

7. Das Verfahren gemäß einem der Ansprüche 1 bis 6, bei dem die Transaktionsprotokolldaten Vermerke umfassen, die eine Sequenzreihenfolge (**120**) der Daten angeben, die geschrieben werden.

8. Das Verfahren gemäß einem der Ansprüche 1 bis 7, bei dem jedes der Speichermedien (**12**) einen reservierten Bereich als einen Zwischenspeicherbereich (**65**) umfasst, der lediglich zum Schreiben verwendet wird, wenn das am wenigsten belegte Speichermedium ausgewählt ist.

9. Das Verfahren gemäß Anspruch 8, bei dem der reservierte Bereich zumindest zwei Teilbereiche (**70**, **75**, **80**, **85**, **90**, **95**, **100**, **105**) umfasst, und wobei einer der Teilbereiche auf ein erstes Ereignis eines Auswählens eines am wenigsten belegten Speichermediums hin beschrieben wird und der andere der Teilbereiche auf ein nächstes nachfolgendes Ereignis eines Auswählens eines am wenigsten belegten Speichermediums hin beschrieben wird, wodurch kein unmittelbar vorhergehend beschriebener Teilbereich bei einem nächsten nachfolgenden Schreibvorgang überschrieben wird, falls das gleiche am wenigsten belegte Speichermedium zweimal nacheinander ausgewählt wird.

10. Ein Speichersystem (**10**), das folgende Merkmale aufweist:
(a) einen ersten Speicher (**55**), der Transaktionsprotokolldaten speichert;
(b) mehrere Speichermedien (**12**), die mit dem ersten Speicher verbunden sind;
(c) getrennte Protokollbereiche (**60**, **65**), die einen Plattenprotokollbereich (**60**) und einen Zwischenspeicherprotokollbereich (**65**) umfassen, der durch Zwischenspeicherbereiche (**70**–**105**) gebildet ist, die Abschnitte der Speichermedien (**12**) darstellen;
(d) eine Einrichtung zum Schreiben von Transaktionsprotokolldaten von dem ersten Speicher (**55**) zu dem Plattenprotokollbereich (**60**), falls ein Seite-Voll-Status des ersten Speichers (**55**) erfasst wird; und
(e) eine Einrichtung zum Schreiben von Transaktionsprotokolldaten von dem ersten Speicher (**55**) zu einem Zwischenspeicherbereich (**70**–**105**) des am wenigsten belegten Speichermediums (**12**), falls eine Ablageanforderung für den ersten Speicher (**55**) erfasst wird.

11. Das Speichersystem gemäß Anspruch 10, bei dem eine volle Seite von Transaktionsprotokolldaten zu dem Plattenprotokollbereich (**60**) geschrieben wird, und wobei einer oder mehrere volle Blöcke von Transaktionsprotokolldaten zu dem Zwischenspeicherprotokollbereich (**65**) geschrieben werden.

12. Das Speichersystem gemäß Anspruch 10 oder 11, bei dem die Einrichtung zum Schreiben von Transaktionsprotokolldaten von dem ersten Speicher (**55**) zu dem Plattenprotokollbereich (**60**) eine Redundanz der Transaktionsprotokolldaten auf den mehreren Speichermedien (**12**) beibehält.

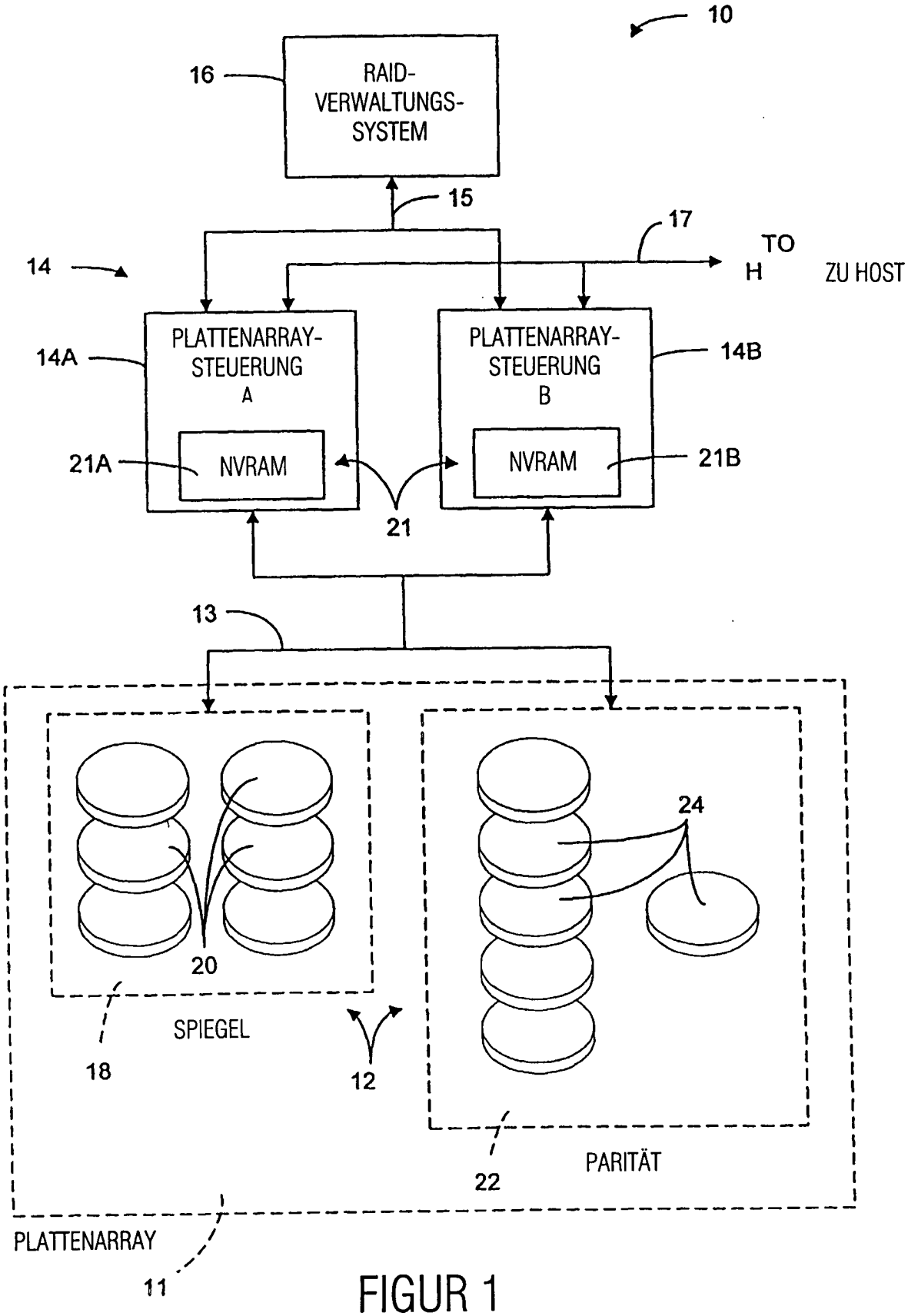
13. Das Speichersystem gemäß einem der Ansprüche 10 bis 12, bei dem die Ablageanforderung eine Anforderung umfasst, um die Transaktionsprotokolldaten in dem ersten Speicher (**55**) zwingend zu den mehreren Speichermedien (**12**) zu schreiben, bevor ein gegebener Abschnitt des ersten Speichers (**55**) voll mit den Transaktionsprotokolldaten ist.

14. Das Speichersystem gemäß einem der Ansprüche 10 bis 13, bei dem die Einrichtung zum Schreiben von Transaktionsprotokolldaten von dem ersten Speicher (**55**) zu einem Zwischenspeicherbereich (**70**–**105**) keine Redundanz der Transaktionsprotokolldaten auf den mehreren Speichermedien (**12**) beibehält.

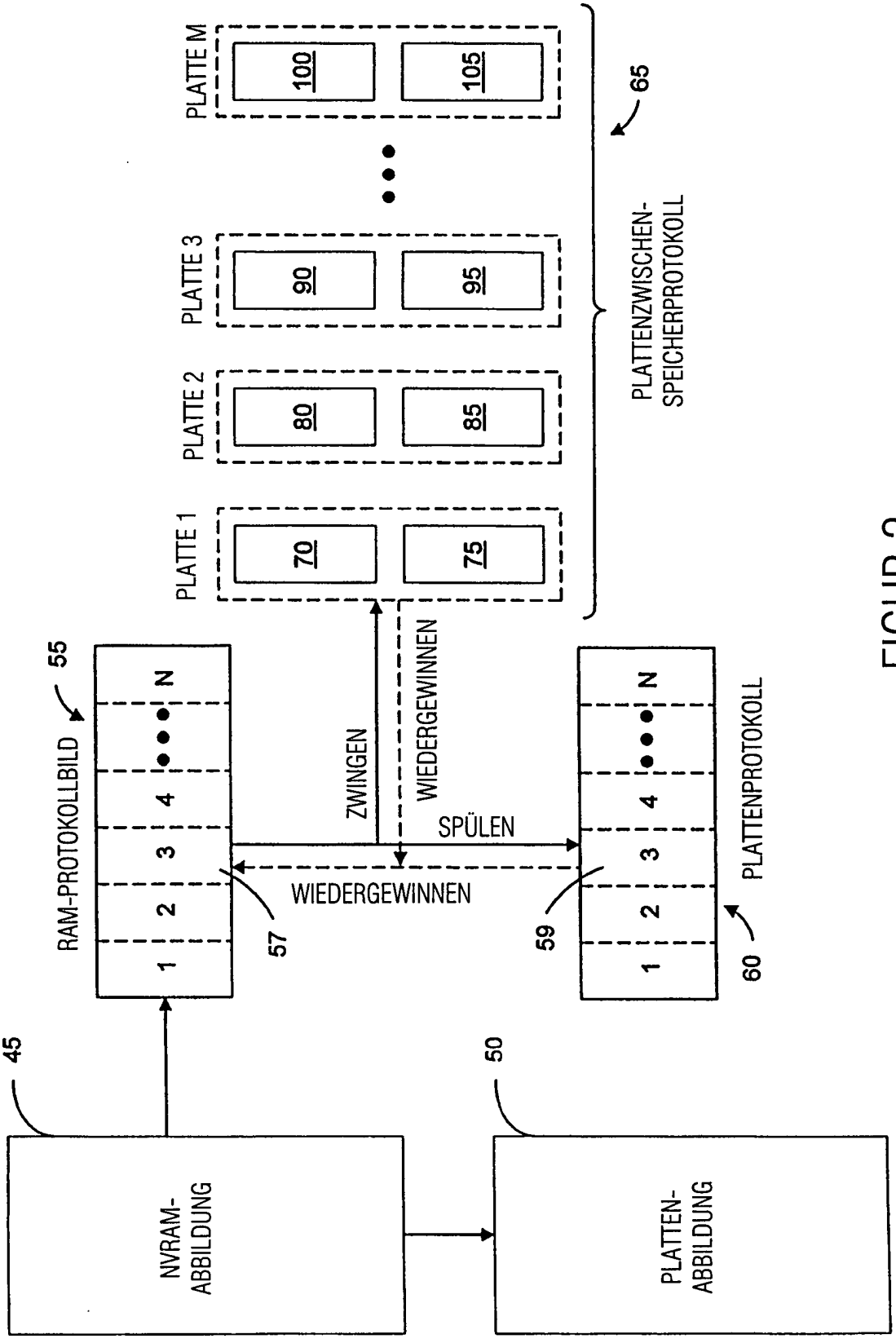
15. Das Speichersystem gemäß einem der Ansprüche 10 bis 14, bei dem das am wenigsten belegte Speichermedium zumindest zwei Teilbereiche (**70**, **75**, **80**, **85**, **90**, **95**, **100**, **105**) umfasst, die zum Schreiben der Transaktionsprotokolldaten reserviert sind.

Es folgen 7 Blatt Zeichnungen

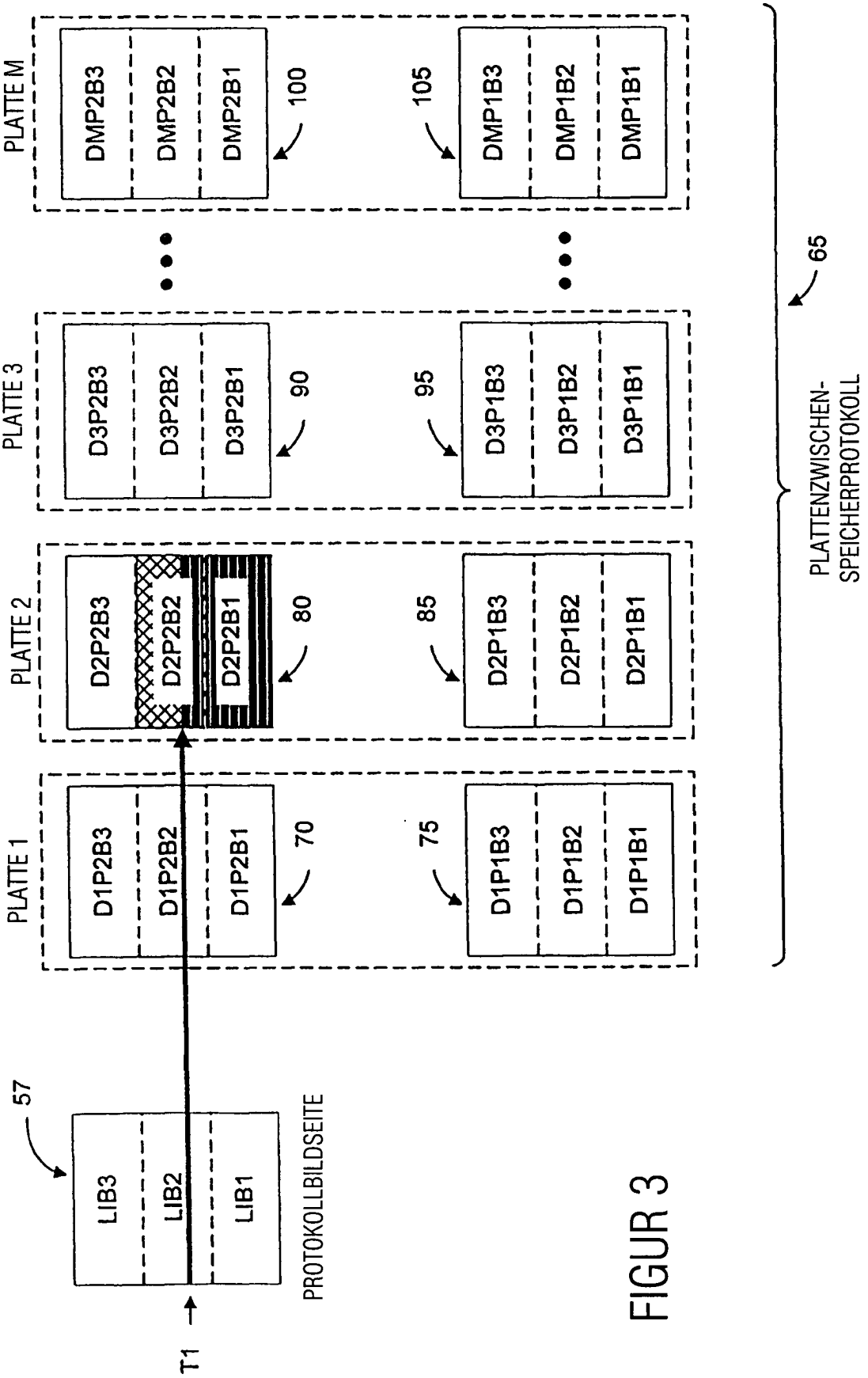
Anhängende Zeichnungen



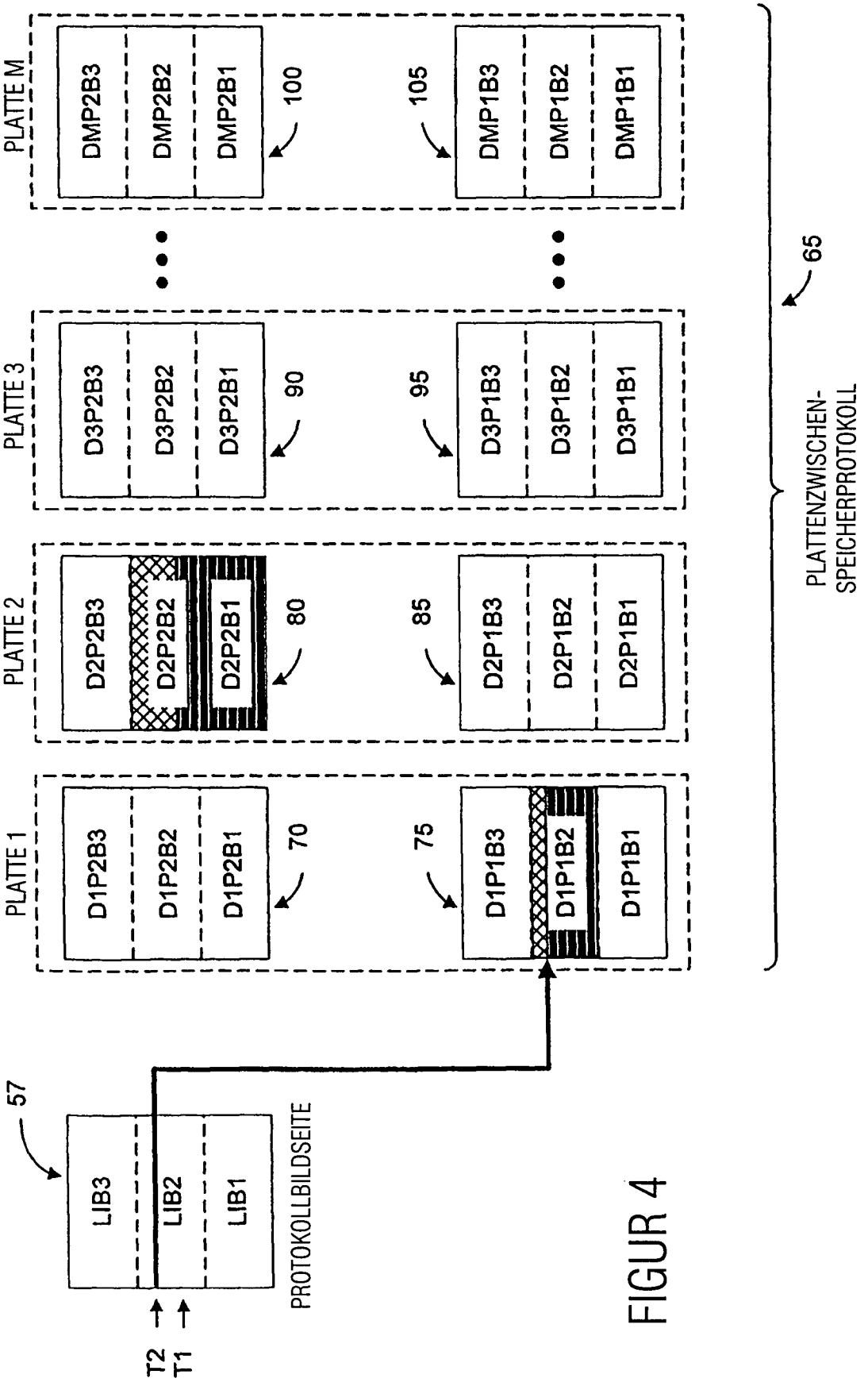
FIGUR 1


NVRAM-ABBILDUNG
PLATTEN-ABBILDUNG
45
50
RAM-PROTOKOLLBILD
55
1
2
3
4
N
57
59
60
PLATTENPROTOKOLL
ZWINGEN
WIEDERGEWINNEN
SPÜLEN
WIEDERGEWINNEN
PLATTE 1
PLATTE 2
PLATTE 3
PLATTE M
70
75
80
85
90
95
100
105
65
PLATTENZWISCHEN-SPEICHERPROTOKOLL

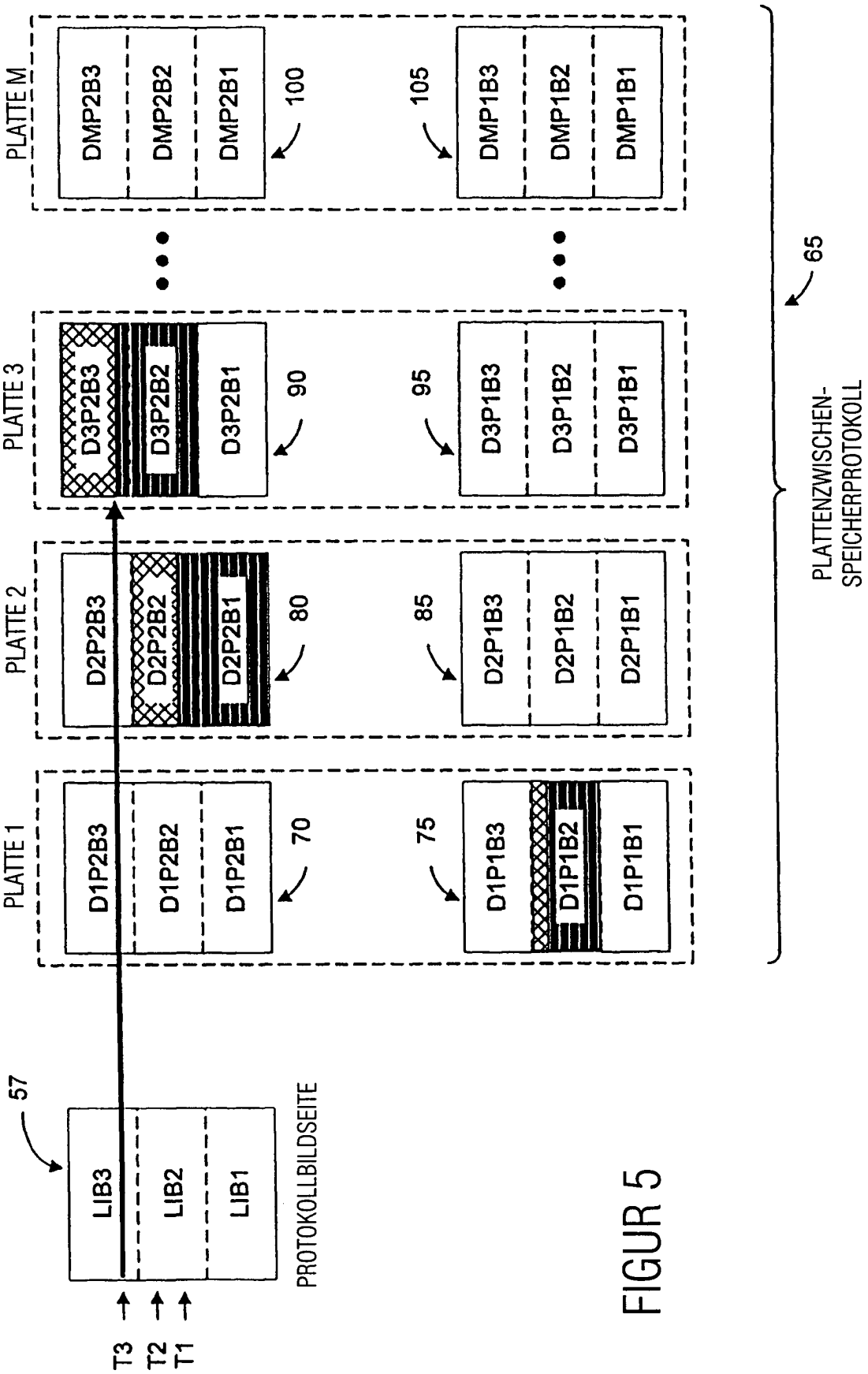
FIGUR 2


PLATTE M
DMP2B3
DMP2B2
DMP2B1
100
105
DMP1B3
DMP1B2
DMP1B1
PLATTE 3
D3P2B3
D3P2B2
D3P2B1
90
95
D3P1B3
D3P1B2
D3P1B1
PLATTE 2
D2P2B3
D2P2B2
D2P2B1
80
85
D2P1B3
D2P1B2
D2P1B1
PLATTE 1
D1P2B3
D1P2B2
D1P2B1
70
75
D1P1B3
D1P1B2
D1P1B1
65
PLATTENZWISCHEN-SPEICHERPROTOKOLL
57
LIB3
LIB2
LIB1
PROTOKOLLBILDSEITE
T1

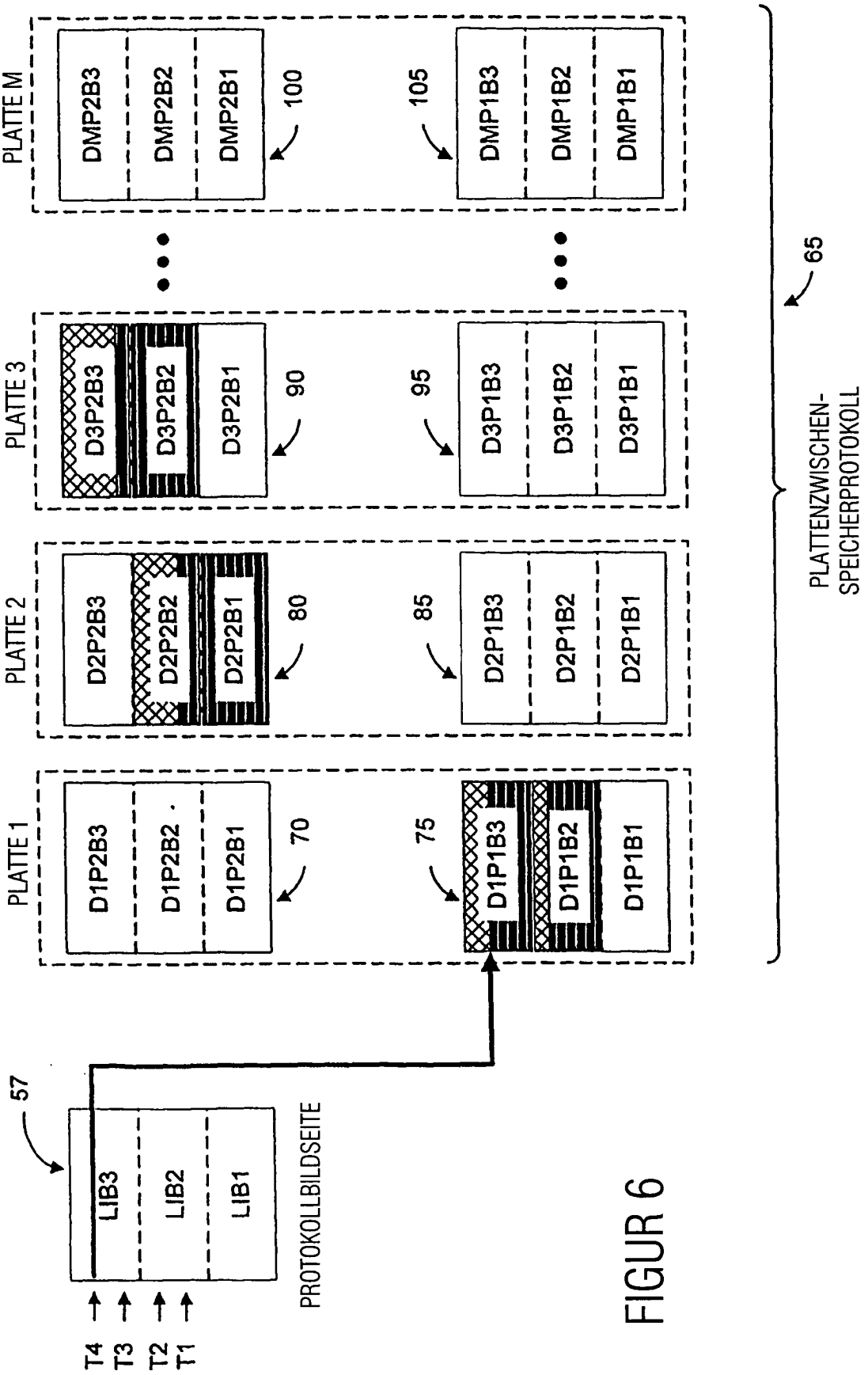
FIGUR 3


PLATTE M
DMP2B3
DMP2B2
DMP2B1
100
105
DMP1B3
DMP1B2
DMP1B1
PLATTE 3
D3P2B3
D3P2B2
D3P2B1
90
95
D3P1B3
D3P1B2
D3P1B1
PLATTE 2
D2P2B3
D2P2B2
D2P2B1
80
85
D2P1B3
D2P1B2
D2P1B1
PLATTE 1
D1P2B3
D1P2B2
D1P2B1
70
75
D1P1B3
D1P1B2
D1P1B1
65
PLATTENZWISCHEN-
SPEICHERPROTOKOLL
57
LIB3
LIB2
LIB1
PROTOKOLLBILDSEITE
T2
T1

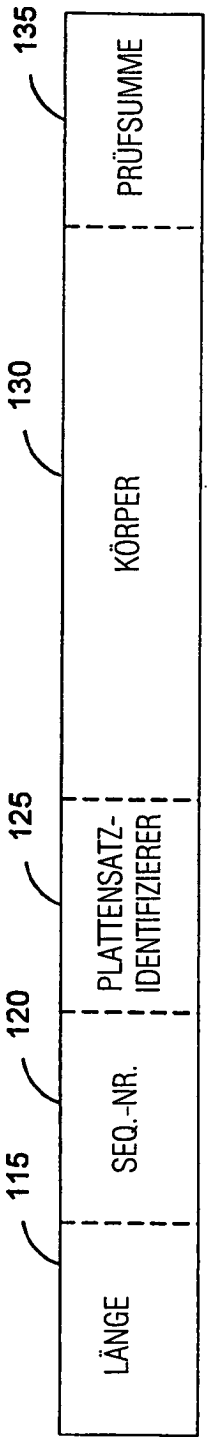
FIGUR 4


PLATTE M
DMP2B3
DMP2B2
DMP2B1
100
105
DMP1B3
DMP1B2
DMP1B1
PLATTE 3
D3P2B3
D3P2B2
D3P2B1
90
95
D3P1B3
D3P1B2
D3P1B1
PLATTE 2
D2P2B3
D2P2B2
D2P2B1
80
85
D2P1B3
D2P1B2
D2P1B1
PLATTE 1
D1P2B3
D1P2B2
D1P2B1
70
75
D1P1B3
D1P1B2
D1P1B1
65
PLATTENZWISCHEN-SPEICHERPROTOKOLL
57
LIB3
LIB2
LIB1
PROTOKOLLBILDSEITE
T3
T2
T1

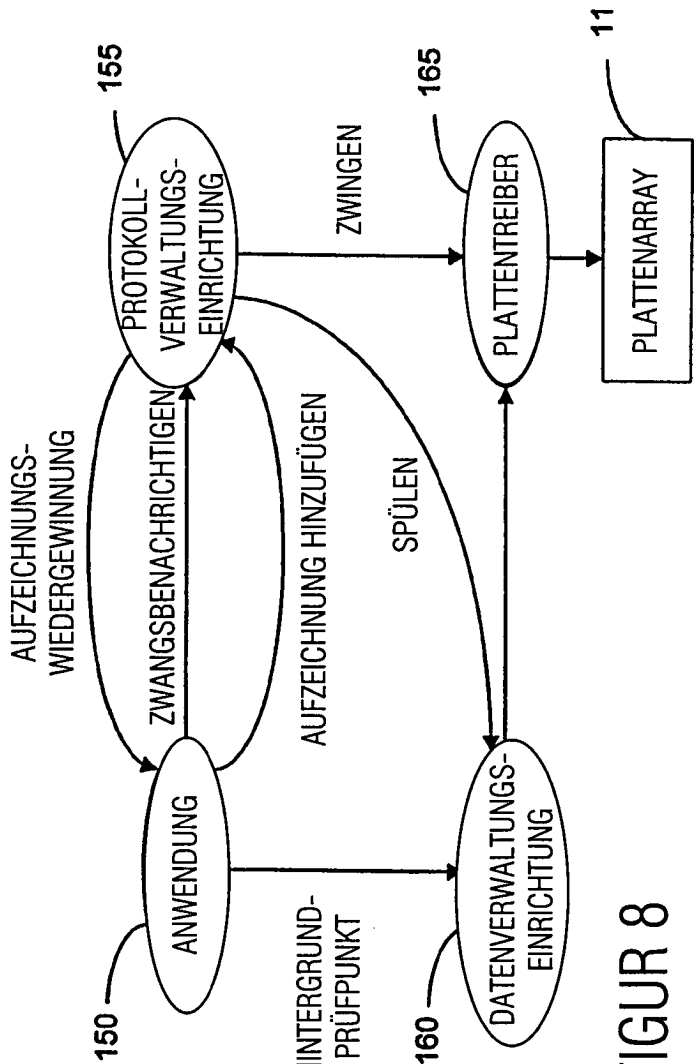
FIGUR 5


PLATTE M
DMP2B3
DMP2B2
DMP2B1
100
105
DMP1B3
DMP1B2
DMP1B1
PLATTE 3
D3P2B3
D3P2B2
D3P2B1
90
95
D3P1B3
D3P1B2
D3P1B1
PLATTE 2
D2P2B3
D2P2B2
D2P2B1
80
85
D2P1B3
D2P1B2
D2P1B1
PLATTE 1
D1P2B3
D1P2B2
D1P2B1
70
75
D1P1B3
D1P1B2
D1P1B1
65
PLATTENZWISCHEN-SPEICHERPROTOKOLL
57
LIB3
LIB2
LIB1
PROTOKOLLBILDSEITE
T4
T3
T2
T1

FIGUR 6



FIGUR 7



FIGUR 8