(12) NACH DEM VERTRAG ÜBER DIE INTERNATIONALE ZUSAMMENARBEIT AUF DEM GEBIET DES PATENTWESENS (PCT) VERÖFFENTLICHTE INTERNATIONALE ANMELDUNG

(19) Weltorganisation für geistiges Eigentum
Internationales Büro

(43) Internationales Veröffentlichungsdatum
6. März 2014 (06.03.2014)

WIPO | PCT

(10) Internationale Veröffentlichungsnummer
WO 2014/032910 A1

(51) Internationale Patentklassifikation:
G06F 3/06 (2006.01)

(21) Internationales Aktenzeichen: PCT/EP2013/066399

(22) Internationales Anmeldedatum:
5. August 2013 (05.08.2013)

(25) Einreichungssprache: Deutsch

(26) Veröffentlichungssprache: Deutsch

(30) Angaben zur Priorität:
10 2012 108 103.4
31. August 2012 (31.08.2012) DE

(71) Anmelder: **FUJITSU TECHNOLOGY SOLUTIONS INTELLECTUAL PROPERTY GMBH** [DE/DE]; Mies-van-der-Rohe-Straße 8, 80807 München (DE).

(72) Erfinder: **KASPER, Dieter**; Hornaustraße 8, 87640 Biessenhofen (DE).

(74) Anwalt: **EPPING HERMANN FISCHER, PATENTANWALTSGESELLSCHAFT MBH**; Zusammenschluss NR. 175, Schlossschmidstr. 5, 80639 Munich (DE).

(81) Bestimmungsstaaten *(soweit nicht anders angegeben, für jede verfügbare nationale Schutzrechtsart)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Bestimmungsstaaten *(soweit nicht anders angegeben, für jede verfügbare regionale Schutzrechtsart)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), eurasisches (AM, AZ, BY, KG, KZ, RU, TJ, TM), europäisches (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

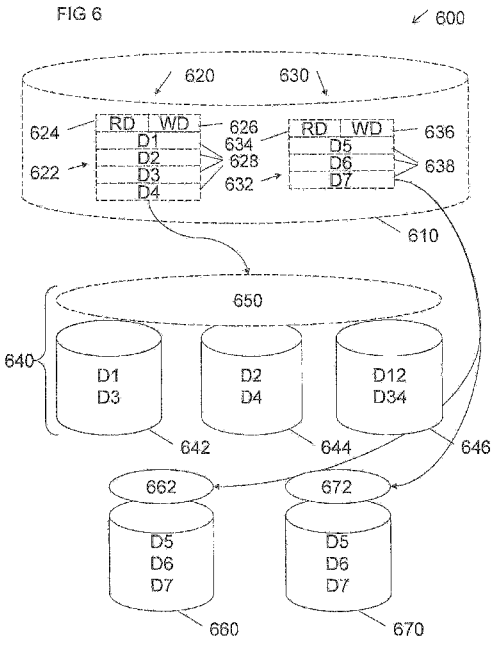
Veröffentlicht:

— *mit internationalem Recherchenbericht (Artikel 21 Absatz 3)*

(54) Title: WORKING METHOD FOR A MASS STORAGE SYSTEM, A MASS STORAGE SYSTEM, AND A COMPUTER PROGRAM PRODUCT

(54) Bezeichnung : ARBEITSVERFAHREN FÜR EIN MASSENSPEICHERSYSTEM, MASSENSPEICHERSYSTEM UND COMPUTERPROGRAMMPRODUKT

FIG 6



(57) Abstract: The invention relates to a working method for a mass system (100, 600), particularly a RAID system (500). The system comprises the steps of providing a virtual data file system (610) for at least one user of the mass system (100, 600), and determining an access probability for data (620, 630) that are logically stored in said virtual data file system (610). Data files (620) whose access probabilities lie above a predetermined limiting value (T_{REF}) are stored distributed in a plurality of first physical mass memories (642, 644, 646) which are independent of one another and have independent writer/reader units. Data files (630) whose access probabilities lie below the predetermined limiting value (T_{REF}) are stored together in at least one contiguous region of at least one second physical mass (660). Moreover, the invention also relates to a mass system (100, 600) which is configured to carry out the method, as well as to a computer program product that is suitable for implementing said method.

(57) Zusammenfassung: Die Erfindung betrifft ein Arbeitsverfahren für ein Massenspeichersystem (100, 600), insbesondere ein RAID-System (500). Das System umfasst den Schritt des Bereitstellens eines virtuellen Dateisystems (610) für wenigstens einen Benutzer des Massenspeichersystems (100, 600) sowie des Bestimmens einer Zugriffswahrscheinlichkeit für in dem virtuellen Dateisystem (610) logisch gespeicherte Daten (620, 630). Dateien (620), deren Zugriffswahrscheinlichkeit über einem

[Fortsetzung auf der nächsten Seite]

WO 2014/032910 A1

vorbestimmten Grenzwert (T_{REF}) liegt, werden verteilt in einer Mehrzahl von voneinander unabhängigen ersten physikalischen Massenspeichern (642, 644, 646) mit voneinander unabhängigen Schreib-/Leseeinheiten gespeichert. Dateien (630), deren Zugriffswahrscheinlichkeit unter dem vorbestimmten Grenzwert (T_{REF}) liegt, werden gemeinsam in wenigstens einem zusammenhängenden Speicherbereich wenigstens eines zweiten physikalischen Massenspeichers (660) gespeichert. Die Erfindung betrifft darüber hinaus auch ein zur Durchführung des Verfahrens eingerichtetes Massenspeichersystem (100, 600) sowie ein zur Ausführung des Verfahrens geeignetes Computerprogrammprodukt.

Beschreibung

Arbeitsverfahren für ein Massenspeichersystem, Massenspeichersystem und Computerprogrammprodukt

Die Erfindung betrifft ein Arbeitsverfahren für ein Massenspeichersystem, insbesondere ein RAID-System, das wenigstens ein virtuelles Dateisystem für wenigstens einen Benutzer des Massenspeichersystems bereitstellt. Die Erfindung betrifft des Weiteren ein zur Umsetzung des Verfahrens geeignetes Massenspeichersystem sowie ein Computerprogrammprodukt mit Programmanweisungen zum Ausführen auf einer Datenverarbeitungseinheit eines elektronischen Datenverarbeitungssystems.

Massenspeichersysteme, insbesondere RAID-Systeme, und zu ihrem Betrieb geeignete Arbeitsverfahren sind aus dem Stand der Technik bekannt. Insbesondere ist aus dem Artikel "A Case For Redundant Arrays of Inexpensive Disks (RAID)" von David A. Patterson, Garth Gibson und Randy H. Cats, veröffentlicht in der International Conference on Management of Data, 1988, Chicago, der inzwischen allgemein als RAID-System bekannte Ansatz bekannt, Daten verteilt über mehrere Festplatten, die physikalisch unabhängig voneinander sind, zu speichern. Dabei nimmt ein Benutzer des Massenspeichersystems die physikalische Aufteilung der einzelnen eingesetzten Massenspeicher nicht mehr wahr, sondern speichert seine Daten in einem logischen oder virtuellen Dateisystem, dessen zugehörige Daten physisch auf einem oder einer Mehrzahl von Massenspeichern abgelegt sind.

Die Verwendung von RAID-Systemen bietet im Allgemeinen eine Anzahl von Vorteilen. Insbesondere können umfangreiche Daten über mehrere physikalische Festplattenlaufwerke verteilt wer-

den. Dies besitzt unter anderem den Vorteil, dass Schreib-/Lese-Zugriffe beschleunigt werden können, da die Datentransferrate mehrerer physikalischer Massenspeichergeräte zur Verfügung steht. Des Weiteren kann auch eine Redundanz der Daten und somit eine Sicherheit gegenüber dem Ausfall eines einzigen Festplattenlaufwerks erreicht werden, indem Daten gleichzeitig auf mehreren physikalischen Massenspeichern abgelegt werden. Die genannten Vorteile lassen sich in unterschiedlichen Betriebsarten, die allgemein als RAID-Level bekannt sind, zumindest teilweise auch miteinander kombinieren.

Ein Nachteil, der mit dem oben genannten RAID-Ansatz einhergeht, liegt darin, dass durch die Vorsehung einer Vielzahl von voneinander unabhängig betriebenen Massenspeichern ein erhöhter Energiebedarf entsteht. Bereits bei der Vorsehung von nur wenigen Festplattenlaufwerken, beispielsweise vier Festplattenlaufwerken mit je 250 GB Kapazität, gegenüber der Vorsehung eines einzigen Festplattenlaufwerks mit beispielsweise 1 TB Kapazität entsteht durch die zusätzlich erforderliche Steuerelektronik und den Antrieb der jeweils vorhandenen Plattenstapel ein nicht unerheblicher Mehrbedarf an elektrischer Energie. Dieser zusätzliche Energiebedarf erhöht sich weiter, wenn, wie oben beschrieben, redundante Datenhaltung gewünscht ist und somit weitere, zusätzliche Festplattenlaufwerke eingesetzt werden. Insbesondere im gewerblichen Umfeld, in dem heutzutage sehr große Datenmengen in zentralen Rechenzentren vorgehalten werden, erhöht sich das Problem noch weiter. Denn bei einem Einsatz von einigen hundert oder gegebenenfalls sogar etlichen tausend Festplattenlaufwerken fällt zusätzlich zu deren Betriebsenergie ein weiterer Aufwand zu deren Kühlung im Betrieb an.

Aus dem Stand der Technik sind Maßnahmen bekannt, die Energieaufnahme einzelner Plattenlaufwerke zu reduzieren. Beispielsweise ist es bekannt, durch ein Betriebssystem oder einen internen Festplatten-Controller den Antrieb für einen Stapel mit rotierenden Speichermedien vorübergehend zu deaktivieren, wenn aktuell keine Zugriffe auf ein Laufwerk erfolgen. Die an sich bekannte Abschaltung birgt jedoch eine Reihe von Nachteilen. Zum einen muss bei einem erneuten Zugriff auf einen derartig deaktivierten Massenspeicher der Plattenstapel zunächst wieder beschleunigt werden, bevor der gewünschte Zugriff ausgeführt werden kann, was zu längeren Zugriffszeiten führt. Zum anderen führt das wiederholte Abschalten und erneute Beschleunigen der Plattenstapel zu einem erhöhten mechanischen Verschleiß und somit einer geringeren Lebensdauer des Massenspeichers. Somit werden derartige Energiesparoptionen zumindest in Rechenzentren, bei denen eine Vielzahl von Dateien für eine Vielzahl von Nutzern in umfangreichen Massenspeichersystemen vorgehalten werden, in der Regel nicht oder nur selten verwendet.

Eine Aufgabe der vorliegenden Erfindung liegt darin, ein Arbeitsverfahren für ein Massenspeichersystem bzw. ein Massenspeichersystem zu beschreiben, das bei im Wesentlichen gleicher Leistung eine gegenüber bekannten Massenspeichersystemen reduzierte Energieaufnahme aufweist oder das bei gleichem Energiebedarf eine höhere Leistung zur Verfügung stellen kann. Bevorzugt soll sich das Massenspeichersystem aus Sicht seiner Benutzer wie ein konventionelles Massenspeichersystem verhalten.

Gemäß einem ersten Aspekt der Erfindung wird ein Arbeitsverfahren für ein Massenspeichersystem, insbesondere ein RAID-System, beschrieben. Es umfasst die Schritte:

- Bereitstellen eines virtuellen Dateisystems für wenigstens einen Benutzer des Massenspeichersystems,
- Bestimmen einer Zugriffswahrscheinlichkeit für in dem virtuellen Dateisystem logisch gespeicherte Daten,
- verteiltes Speichern von Dateien, deren Zugriffswahrscheinlichkeit über einem vorbestimmten Grenzwert liegt, in einer Mehrzahl von voneinander unabhängigen ersten physikalischen Massenspeichern mit voneinander unabhängigen Schreib-/Leseeinheiten und
- gemeinsames Speichern von Dateien, deren Zugriffswahrscheinlichkeit unter dem vorbestimmten Grenzwert liegt, in wenigstens einem zusammenhängenden Speicherbereich wenigstens eines zweiten physikalischen Massenspeichers.

Das genannte Arbeitsverfahren unterteilt in einem virtuellen Dateisystem eines Massenspeichersystems abgelegten Dateien anhand einer Zugriffswahrscheinlichkeit in wenigstens zwei Gruppen mit unterschiedlichen Zugriffswahrscheinlichkeiten. Dabei werden Dateien, auf die mit einer verhältnismäßig hohen Wahrscheinlichkeit zugegriffen wird, verteilt in einer Mehrzahl voneinander unabhängiger erster physikalischer Massenspeicher abgelegt, um die Schreib-/Leseleistung der beteiligten Massenspeicher zum Vorteil des Benutzers wie bei konventionellen RAID-Systemen zu kombinieren. Dateien, deren Zugriffswahrscheinlichkeit unter einem vorbestimmten Grenzwert liegt, werden dagegen zusammenhängend in einem Speicherbereich wenigstens eines zweiten physikalischen Massenspeichers abgelegt. Zumindest zum Speichern der letztgenannten Gruppe von Dateien ist daher in der Regel nur der Betrieb eines einzelnen physikalischen Massenspeichers erforderlich, sodass die Energieaufnahme des Massenspeichersystems gegenüber einem konventionellen RAID-System mit einer verteilten Speicherung aller Daten insgesamt reduziert wird.

In einer bevorzugten Ausgestaltung wird der wenigstens eine zweite physikalische Massenspeicher in einen Betriebszustand mit gegenüber einem normalen Betriebszustand reduzierter Energieaufnahme geschaltet, wenn über einen vorbestimmten Zeitraum kein Zugriff auf den wenigstens einen zweiten physikalischen Massenspeicher durchgeführt wurde. Der wenigstens eine zweite physikalische Massenspeicher wird zurück in den normalen Betriebszustand geschaltet, wenn eine Zugriffsanforderung für wenigstens eine auf dem wenigstens einen zweiten physikalischen Massenspeicher gespeicherte Datei über das virtuelle Dateisystem erfasst wird. Die Vorsehung der oben genannten Schritte führt zu einer weiteren Energieeinsparung durch zumindest zeitweises Abschalten des zweiten physikalischen Massenspeichers. Die dabei grundsätzlich auftretenden Nachteile einer verzögerten Zugriffszeit sowie eines erhöhten Verschleißes fallen in dem genannten Arbeitsverfahren weniger ins Gewicht als bei bekannten Massenspeichersystemen, da die Zugriffswahrscheinlichkeit auf den zweiten physikalischen Massenspeicher aufgrund der Verteilung von Dateien besonders niedrig ist, so dass der zweite physikalische Massenspeicher auch nur sehr selten in den normalen Betriebszustand geschaltet werden muss.

Gemäß einer vorteilhaften Ausgestaltung werden die Dateien des virtuellen Dateisystems in Gruppen mit unterschiedlichen Zugriffswahrscheinlichkeitsbereichen aufgeteilt. Dabei können in einer weiteren Ausgestaltung Dateien einer ersten Gruppe mit einer verhältnismäßig hohen Zugriffswahrscheinlichkeit in Segmente vorbestimmter Größe aufgeteilt werden, wobei die Segmente einer Datei mit mehr als einem Segment verteilt auf der Mehrzahl von voneinander unabhängigen Massenspeichern gespeichert werden. Wird auf diese Weise der Inhalt großer Da-

teien, auf die häufig zugegriffen wird, auf mehrere physikalische Massenspeicher verteilt, ergibt sich insgesamt eine erhöhte Performance des Massenspeichersystems gegenüber der Ablage einer entsprechenden Datei auf einem einzelnen physikalischen Massenspeicher.

Die Zugriffswahrscheinlichkeit für eine Datei kann beispielsweise basierend auf wenigstens einem Zeitpunkt der Erstellung der Datei, einem Zeitpunkt des letzten Schreibzugriffs auf die Datei, einem Zeitpunkt des letzten Lesezugriffs auf die Datei und/oder einer Anzahl von Lese- und/oder Schreibzugriffen innerhalb eines vorbestimmten Zeitraums auf die Datei bestimmt werden.

Gemäß einem zweiten Aspekt der Erfindung wird ein Massenspeichersystem, insbesondere ein RAID-System, beschrieben. Das System umfasst eine Mehrzahl von voneinander unabhängigen ersten physikalischen Massenspeichern mit voneinander unabhängiger Schreib-/Leseeinheiten, wenigstens einen zweiten physikalischen Massenspeicher und wenigstens eine Schnittstellenvorrichtung zum Bereitstellen eines virtuellen Dateisystems für wenigstens einen Benutzer des Massenspeichersystems. Das System umfasst des Weiteren wenigstens eine Steuervorrichtung zum wahlweisen Speichern der in dem virtuellen Dateisystem logisch gespeicherten Dateien auf den ersten physikalischen Massenspeichern oder dem wenigstens einen zweiten physikalischen Massenspeicher. Dabei ist die Steuervorrichtung dazu eingerichtet, für die in dem virtuellen Dateisystem logisch gespeicherten Daten eine Zugriffswahrscheinlichkeit zu bestimmen, Dateien, deren Zugriffswahrscheinlichkeit über einem ersten vorbestimmten Grenzwert liegt, in einer Mehrzahl von voneinander unabhängigen ersten physikalischen Massenspeichern mit voneinander unabhängigen Schreib-/Leseeinheiten

verteilt zu speichern, und Dateien, deren Zugriffswahrscheinlichkeit unter dem vorbestimmten Grenzwert liegt, in wenigstens einem zusammenhängenden Speicherbereich des wenigstens einen zweiten physikalischen Massenspeichers gemeinsam zu speichern.

Ein derartiges Massenspeichersystem weist im Wesentlichen dieselben Vorteile auf wie das oben genannte Arbeitsverfahren gemäß dem ersten Aspekt.

Bevorzugt umfasst der wenigstens eine zweite physikalische Massenspeicher eine Antriebseinheit und wenigstens ein durch die Antriebseinheit rotatorisch antreibbares Speichermedium, wobei die Steuereinheit oder der wenigstens eine zweite physikalische Massenspeicher dazu eingerichtet ist, die Antriebseinheit abzuschalten, wenn über einen vorbestimmten Zeitraum kein Zugriff auf den wenigstens einen zweiten physikalischen Massenspeicher erfolgt. Durch das Abschalten der in konventionellen Festplatten enthaltenen Spindeln mit rotierenden Speichermedien kann deren Betriebsenergie in einem Bereitschaftszustand erheblich reduziert werden.

Gemäß vorteilhaften Ausgestaltungen werden die in dem Massenspeichersystem gespeicherten Daten redundant gespeichert, um eine Absicherung gegenüber dem Ausfall einzelner Massenspeicher zu schaffen. Zur redundanten Speicherung der Dateien mit niedriger Zugriffswahrscheinlichkeit umfasst das Massenspeichersystem wenigstens zwei zweite physikalische Massenspeicher, wobei auf einem ersten der wenigstens zwei zweiten physikalischen Massenspeicher gespeichert Daten auf wenigstens einem zweiten der wenigstens zwei zweiten physikalischen Massenspeicher redundant gespeichert werden. Dies entspricht im Wesentlichen der bekannten RAID-Betriebsart 1 und weist den

zusätzlichen Vorteil auf, dass beim Lesen von Daten nur eines der wenigstens zwei zweiten Massenspeicher in einen normalen Betriebszustand versetzt werden muss. Zur redundanten Speicherung der Dateien mit hoher Zugriffswahrscheinlichkeit werden auf einem ersten der Mehrzahl von ersten physikalischen Massenspeichern gespeicherte Dateien oder Segmente von Dateien auf wenigstens einem zweiten der Mehrzahl von ersten Massenspeichern redundant gespeichert. Dies entspricht im Wesentlichen den bekannten RAID-Betriebsarten 1, 3, 4, 5 oder 6 oder RAID-Kombinationen hiervon, wie beispielsweise ein RAID-15-Verbund, und kombiniert die Vorteile hoher Zugriffsbandbreite mit denen einer hohen Datensicherheit.

Gemäß einer weiteren vorteilhaften Ausgestaltung ist bei dem Massenspeichersystem die Mehrzahl von ersten physikalischen Massenspeichern über eine erste, lokale Verbindungsstruktur mit einer ersten Bandbreite mit der wenigstens einen Schnittstellenvorrichtung verbunden und der wenigstens eine zweite physikalische Massenspeicher über eine zweite Verbindungsstruktur mit einer gegenüber der ersten Bandbreite geringeren zweiten Bandbreite mit der wenigstens einen Schnittstellenvorrichtung verbunden. Durch den Einsatz unterschiedlicher Verbindungsstrukturen für die ersten physikalischen Massenspeicher einerseits und den wenigstens einen zweiten physikalischen Massenspeicher andererseits kann auch auf Ebene der Verbindungsstrukturen eine leistungsgerechte Entkopplung zwischen Dateien mit einer hohen Zugriffswahrscheinlichkeit und Dateien mit einer niedrigen Zugriffswahrscheinlichkeit vorgenommen werden. Beispielsweise können die Dateien der ersten Gruppe bevorzugt gemäß dem Fibre Channel Protokoll oder einem Remote Direct Memory Acess (RDMA) Protokoll, insbesondere einem RDMA over Infiniband Protokoll, wie zum Beispiel dem SCSI RDMA Protokoll (SRP), dem Socket Direct Protokoll (SDP) oder

dem nativen RDMA Protokoll, oder einem RDMA over Ethernet Protokoll, wie zum Beispiel dem RDMA over Converged Ethernet (RoCE) oder dem Internet Wide Area RDMA (iWARP) Protokoll, zum Zugriff auf lokale Massenspeicher mit einer besonders hohen Datenübertragungsrate gelesen und geschrieben werden, während Daten der zweiten Gruppe mit einem anderen, leistungsschwächeren Protokoll, beispielsweise einem TCP/IP-basierten Protokoll für den Zugriff auf entfernte Speichermedien, gelesen oder geschrieben werden können.

In einer bevorzugten Ausgestaltung wird die Mehrzahl von ersten physikalischen Massenspeichern von der Steuervorrichtung über ein gemeinsames, verteiltes erstes Dateisystem verwaltet und der wenigstens eine zweite physikalische Massenspeicher wird über ein für den Massenspeicher spezifisches zweites Dateisystem verwaltet. Die Vorsehung unterschiedlicher Dateisysteme ermöglicht eine einfache Implementierung und leistungsgerechte Verwaltung der unterschiedlichen Massenspeicher.

Gemäß einer weiteren vorteilhaften Ausgestaltung werden einzelne Untergruppen von Dateien des virtuellen Dateisystems unterschiedlichen physikalischen Massenspeichern zugeordnet, wobei die Zugriffsgeschwindigkeit der jeweils verwendeten Massenspeicher der Zugriffswahrscheinlichkeit der entsprechenden Gruppe angepasst ist. Beispielsweise ist es möglich, Dateien, auf die besonders häufig zugegriffen wird, auf besonders leistungsfähigen Halbleiter-Massenspeichern, insbesondere so genannte SSD-Laufwerken, abzulegen, während Dateien mit einer mittleren Zugriffswahrscheinlichkeit auf Massenspeicher mit rotierenden Speichermedien, wie insbesondere SCSI- und SATA-Festplattenlaufwerken abgelegt sind. Daten mit besonders niedriger Zugriffswahrscheinlichkeit können auf physikalisch extern angeordneten Massenspeichermedien, insbe-

sondere so genannten Network Attached Storage (NAS) oder Direct Attached Storage (DAS) Speicherlaufwerken abgelegt werden.

Gemäß einem dritten Aspekt der Erfindung wird ein Computerprogrammprodukt mit Programmanweisungen zur Ausführung auf einer Datenverarbeitungseinheit eines elektronischen Datenverarbeitungssystems gemäß Patentanspruch 15 beschrieben. Das Computerprogrammprodukt weist im Wesentlichen dieselben Vorteile auf wie das Arbeitsverfahren gemäß dem ersten Aspekt und das Massenspeichersystem gemäß dem zweiten Aspekt.

Weitere vorteilhafte Ausgestaltungen der Erfindung sind in den abhängigen Ansprüchen sowie der nachfolgenden ausführlichen Beschreibung von Ausführungsbeispielen dargestellt.

Die Erfindung wird nachfolgend anhand von Ausführungsbeispielen unter Bezugnahme auf die angehängten Figuren im Detail beschrieben. In den Figuren zeigen:

Figur 1 eine schematische Darstellung eines Massenspeichersystem gemäß einer Ausgestaltung der Erfindung,

Figur 2 ein Ablaufdiagramm eines ersten Arbeitsverfahrens zum Betrieb eines Massenspeichersystems,

Figur 3 ein Ablaufdiagramm eines zweiten Arbeitsverfahrens zum Betrieb eines Massenspeichersystems,

Figur 4 ein Ablaufdiagramm eines Bestimmungsverfahrens für und eine schematische Darstellung von unterschiedlicher Zugriffswahrscheinlichkeiten,

Figur 5 eine schematische Darstellung eines RAID-Systems gemäß einer Ausgestaltung der Erfindung und

Figur 6 eine schematische Darstellung von Dateisystemen eines erfindungsgemäßen Massenspeichersystems.

Figur 1 zeigt ein Massenspeichersystem 100 gemäß einer ersten Ausgestaltung der Erfindung. Das Massenspeichersystem 100 umfasst eine Netzwerkschnittstelle 110, eine Steuervorrichtung 120 und einen internen Speicher zum Ablegen von Steuerdaten, insbesondere von Metadaten über in dem Massenspeichersystem 100 gespeicherte Daten in Form einer Datenbank 130. Darüber hinaus umfasst das Massenspeichersystem 100 zwei erste physikalische Massenspeicher 142 und 144 sowie einen zweiten physikalischen Massenspeicher 150. Die ersten physikalischen Massenspeicher 142 und 144 bilden zusammen einen primären Speicherverbund 140 und sind über eine lokale primäre Verbindungsstruktur 160, insbesondere eine besonders leistungsfähige Fibre-Channel- oder RDMA-Verbindung, mit der Steuervorrichtung 120 verbunden. Der zweite physikalische Massenspeicher 150 ist über eine sekundäre Verbindungsstruktur 170, insbesondere eine verhältnismäßig preisgünstige Internet-Protokoll-Verbindung, z.B. eine TCP/IP-Verbindung über 1 GBit/s Ethernet, mit der Steuervorrichtung 120 verbunden. Des Weiteren ist die Steuervorrichtung 120 funktionsfähig mit der Netzwerkschnittstelle 110 verbunden, die das Massenspeichersystem 100 über ein Netzwerk 180 mit einem Client Computer 190 verbindet.

Im Betrieb greift der Client Computer 190 mittels eines geeigneten Netzwerkprotokolls über das Netzwerk 180 und die Netzwerkschnittstelle 110 auf die Daten des Massenspeichersystems 100 zu. Hierzu werden Schreib- und Leseanforderungen

der Client Computer 190 durch die Steuervorrichtung 120 analysiert und entsprechend den angeforderten Daten an die Massenspeicher 142, 144 und/oder 150 weitergeleitet. Für die Ermittlung, welche Daten auf welchen der Massenspeicher 142, 144, 150 angeordnet sind, dienen neben eventueller Datei-Informationen der jeweiligen eingesetzten Dateisysteme der Massenspeicher 142, 144 und 150 im Ausführungsbeispiel zusätzlich oder alternativ die in der Datenbank 130 enthaltenen Metadaten. Dabei kann die Datenbank 130 selbst auf den Massenspeichern 142 und/oder 144 des primären Speicherverbunds 140 oder einem sonstigen internen Massenspeicher oder dem Hauptspeicher des Massenspeichersystems 100 gespeichert sein.

Aus Gründen einer einfacheren Darstellung wird im Folgenden stets von auf den Massenspeichern 142, 144 und 150 gespeicherten "Dateien" gesprochen. Dabei kann es sich beispielsweise um in einem hierarchischen Dateisystem angeordnete und über einen vollständigen Pfad und Dateinamen identifizierbare Dateien handeln. An dieser Stelle wird jedoch darauf hingewiesen, dass sich das nachfolgend beschriebene System auch für andere Massenspeichersysteme eignet, die anstelle von Dateien andere Arten von Objekten und anstelle von Pfaden und Dateinamen andere Kennungen verwenden. Unter dem Begriff "Datei" ist daher jede inhaltlich zusammengehörige Dateneinheit zu verstehen, die mittels einer geeigneten Kennung aus einem Massenspeichersystem abrufbar ist. Beispiele solcher Dateneinheiten sind auch Objekte eines Objektspeichersystems oder vordefinierter Bereiche eines Massenspeichersystems. Unter dem Begriff "Dateisystem" ist entsprechend jede geeignete Zugriffsstruktur für derartige Dateneinheiten zu verstehen, also auch Tabellen oder andere Mechanismen zum Zuordnen einer Kennung zu einer Dateneinheit.

Im beschriebenen Ausführungsbeispiel werden sämtliche von dem Client Computer 190 empfangenen Schreibanforderungen über die Fibre-Channel- oder RDMA-Verbindung 150 auf die physikalischen Massenspeicher 142 und 144 verteilt. Dabei kann die Verteilung beispielsweise derart erfolgen, dass Schreibanforderungen abwechselnd an den physikalischen Massenspeicher 142 und an den physikalischen Massenspeicher 144 weitergeleitet werden, sodass eine erste Schreibanforderung für den physikalische Massenspeicher 142 eine nachfolgende Schreibanforderung für den weiteren physikalischen Massenspeicher 144 nicht blockiert.

Alternativ ist es jedoch auch möglich, die Bearbeitung einer einzigen Anfrage über die beiden physikalischen Massenspeicher 142 und 144 des primären Speicherverbunds zu verteilen. Dies bietet sich insbesondere beim Speichern von besonders umfangreichen Dateien an. Bei diesem so genannten "Data Slicing" werden in der Regel gleich große Blöcke von Daten einer umfangreichen Datei über physikalische Speichersektoren von mehreren Massenspeichern, in diesem Fall den physikalischen Massenspeichern 142 und 144, verteilt, sodass auch der Datendurchsatz beim Bearbeiten einer einzelnen Schreibanforderung erhöht wird.

Selbstverständlich können dieselben Daten oder die Daten und zugehörige Redundanzinformationen wie etwa Paritätswerte, auch auf mehreren physikalischen Massenspeichern 142 und 144 oder redundant an unterschiedlichen Stellen der physikalischen Massenspeicher 142 oder 144 des primären Speicherverbunds 140 gespeichert werden, um die gespeicherten Daten gegen den Ausfall eines gesamten Massenspeichers 142 oder 144 oder einzelner Sektoren des Massenspeichers 142 oder 144 zu schützen. Derartige Verfahren zur redundanten Datenspeiche-

rung sind aus dem Stand der Technik bekannt und werden daher hier nicht im Detail beschrieben.

Leseanforderungen an das Massenspeichersystem 100 werden im Ausführungsbeispiel in Abhängigkeit des Speicherortes der angeforderten Datei durchgeführt. Hierzu greift die Steuervorrichtung 120 zunächst auf die Datenbank 130 zu, in der im Ausführungsbeispiel für jede gespeicherte Datei-Ressource hinterlegt ist, ob sie in dem primären Speicherverbund 140 gespeichert ist oder auf dem zweiten physikalischen Massenspeicher 150 gespeichert ist. Die Datenbank 130 ermittelt eine logische oder physikalische Speicheradresse der angeforderten Daten in dem primären Speicherverbund 140 bzw. dem zweiten Massenspeicher 150 und liefert diese an die Steuervorrichtung 120 zurück.

Die Verteilung der Dateien über die Massenspeicher 142, 144, 150 wird nachfolgend anhand der Figuren 3 und 4 im Detail beschrieben. Zum Verständnis des hier dargestellten Massenspeichersystems 100 genügt es zunächst, dass Dateien, auf die besonders häufig zugegriffen wird, auf den physikalischen Massenspeichern 142, 144 abgelegt sind und weitere Dateien, auf die seltener zugegriffen wird, auf dem zweiten physikalischen Massenspeicher 150 abgelegt sind. Je nach ermitteltem Speicherort ruft die Steuervorrichtung 120 die entsprechenden gespeicherten Daten von den ersten physikalischen Massenspeichern 142 und 144 oder dem zweiten physikalischen Massenspeicher 150 ab. Beim Abrufen von Daten von den ersten physikalischen Massenspeichern 142, 144 kann dabei, wie oben bezüglich des Schreibens beschrieben, ein so genanntes "Data Slicing" Anwendung finden, sodass beispielsweise Datenblöcke abwechselnd von dem ersten physikalischen Massenspeicher 142 und dem ersten physikalischen Massenspeicher 144 gelesen werden,

um die zum Lesen zur Verfügung stehende Bandbreite über die Bandbreite eines einzelnen Massenspeichers 142 oder 144 zu erhöhen. Dateien, die auf dem zweiten physikalischen Massenspeicher 150 abgelegt sind, sind dagegen zusammenhängend gespeichert und werden von der Steuervorrichtung 120 unabhängig von ihrem Umfang als eine komplette Datei eingelesen und über die Netzwerkschnittstelle 110 an den anfragenden Client Computer 190 weitergeleitet.

Die Verarbeitung von Schreib- oder Leseanfragen des Client Computers 190 durch ein Arbeitsverfahren 200 für das Massenspeichersystem 100 ist im Ablaufdiagramm nach Figur 2 schematisch dargestellt.

In einem ersten Schritt 210 wird eine Anfrage bezüglich einer Datei durch die Steuervorrichtung 120 empfangen. Beispielsweise kann es sich um eine Anfrage gemäß einem Netzwerkspeicherprotokoll wie beispielsweise dem Server Message Block (SMB) Protokoll, dem Network File System (NFS) Protokoll, dem Amazon S3 (Simple Storage Service) Protokoll oder einem lokalen Dateizugriffsprotokoll handeln, die über die Netzwerkschnittstelle 110 von dem logischen oder physikalischen Client Computer 190 empfangen wird.

In einem nachfolgenden Schritt 220 überprüft die Steuervorrichtung 120 zunächst, ob es sich um eine Schreibanfrage oder Leseanfrage handelt, was sich in der Regel unmittelbar aus dem zur Kommunikation mit dem Client Computer 190 eingesetzten Protokoll ergibt.

Aktuell geschriebene Daten besitzen statistisch gesehen eine besonders hohe Zugriffswahrscheinlichkeit für einen nachfolgenden Lese- oder (Über-)Schreibvorgang, sodass die Daten ei-

nes Schreibzugriffs im Ausführungsbeispiel ohne weitere Überprüfung in dem primären Speicherverbund 140 gespeichert werden. Hierzu erfolgt in einem optionalen Schritt 230 zunächst eine Aufteilung der von dem Client Computer 190 übertragenen Daten in gleich große Dateisegmente, die in nachfolgenden Schritten 232 und 234 auf den Massenspeicher 142 bzw. den Massenspeicher 144 geschrieben werden.

Nachfolgend werden in einem Schritt 236 Metadaten der durchgeführten Schreibvorgänge in der Datenbank 130 gespeichert, um nachfolgende Leseanfragen auf die gespeicherten Daten zu ermöglichen. Beispielsweise kann eine physikalische Blockadresse der gespeicherten Dateisegmente auf den physikalischen Massenspeichern 142 und 144 oder eine logische Adresse der Datei in einem Clusterdateisystems des primären Speicherverbunds 140 insgesamt in der Datenbank 130 erfasst werden. Zusätzlich werden statistische Daten wie etwa der Zeitpunkt des Schreibvorgangs in der Datenbank 130 erfasst.

Mit einer optionalen Replizierung der Daten im Schritt 238 auf denselben oder weiteren Massenspeichern des primären Speicherverbunds 140 wird die Bearbeitung der Schreibanfrage beendet.

Wurde im Schritt 220 jedoch festgestellt, dass es sich bei der Anfrage um eine Leseanfrage handelt, werden in einem nachfolgenden Schritt 250 Metadaten zur Bestimmung einer Position der gespeicherten Daten innerhalb des Massenspeichersystems 100 von der Datenbank 130 angefordert.

Zur Positionsbestimmung wird in einer Entscheidung 260 anhand der Daten der Datenbank 130 zunächst überprüft, ob die angeforderte Datei auf dem primären Speicherverbund 140 oder dem

Massenspeicher 150 gespeichert ist. Dies kann entweder unmittelbar anhand von in der Datenbank 130 abgelegten Adressdaten oder mittelbar anhand der Ermittlung einer Zugriffswahrscheinlichkeit anhand von gespeicherten Metadaten erfolgen.

Handelt es sich um eine Datei mit einer hohen Zugriffswahrscheinlichkeit, wird eine Leseanforderung an die physikalischen Massenspeicher 142 und 144 weitergeleitet. Im optionalen Schritt 270 erfolgt dabei gegebenenfalls eine Aufteilung der Anfragen in Teilanfragen bezüglich einzelner über die physikalischen Massenspeicher 142 und 144 verteilt gespeicherte Dateisegmente. Diese werden in nachfolgenden Schritte 272 bzw. 274 von den physikalischen Massenspeichern 142 bzw. 144 eingelesen. Mit dem Übermitteln einer gegebenenfalls kombinierten Rückantwort in Schritt 276 endet die Leseanforderung.

Wurde in Schritt 260 dagegen festgestellt, dass die Anforderung auf eine Datei mit einer geringen Zugriffswahrscheinlichkeit gerichtet ist, die auf dem zweiten physikalischen Massenspeicher 150 abgelegt ist, wird die Datei in einem Schritt 280 von dem zweiten physikalischen Massenspeicher 150 angefordert. Sofern der Massenspeicher sich zu diesem Zeitpunkt in einem Energiesparzustand befindet, wird der Massenspeicher beispielsweise durch Hochfahren eines Antriebs für einen Speicherplattenstapel zunächst in einen normalen Betriebszustand versetzt. Nachfolgend kann die angeforderte Datei in der Regel aus einem zusammenhängen Speicherbereich des Massenspeichers 150 eingelesen und zurück an den Client Computer 190 übermittelt werden.

In einem nachfolgenden, optionalen Schritt 290 erfolgt gegebenenfalls eine Reorganisation der in dem Massenspeichersys-

tem 100 gespeicherten Daten. Grund dafür ist, dass durch den Zugriff auf die zuvor als Datei mit niedriger Zugriffswahrscheinlichkeit eingeordnete Datei die Wahrscheinlichkeit für einen weiteren Zugriff statistisch erhöht wird, sodass zumindest bezüglich der zugegriffenen Datei gegebenenfalls eine Reorganisation des Massenspeichersystems 100 erforderlich sein könnte.

In der Figur 3 ist schematisch ein mögliches Verfahren 300 zur Reorganisation der Daten des Massenspeichersystems 100 dargestellt.

In einem Schritt 310 wird zunächst überprüft, ob ein Anlass zur Reorganisation der in dem Massenspeichersystem 100 gespeicherten Dateien gegeben ist. Beispielsweise kann im Schritt 310 basierend auf einem Zeitgeber regelmäßig eine Reorganisation nach Ablauf einer vorbestimmten Zeitspanne ausgelöst werden. Beispielsweise kann eine Reorganisation zu einer festen Tageszeit, beispielsweise nachts wenn besonders wenige Zugriffe auf die in dem System 100 gespeicherten Daten zu erwarten sind, vorgenommen werden. Alternativ ist es auch möglich, die Reorganisation beim Eintreten eines vorbestimmten Ereignisses, beispielsweise beim Lesen einer Datei von dem zweiten physikalischen Massenspeicher 150 oder beim Erreichen einer vorbestimmten Speicherkapazitätsgrenze eines der Massenspeicher 144, 142 oder 150, auszulösen. Ebenfalls kann eine Regorganisation nach einer bestimmten Anzahl von Zugriffen auf das Massenspeichersystem 100 insgesamt oder den zweiten physikalischen Massenspeicher 150 ausgelöst oder manuell von einem Administrator des Systems 100 angefordert werden. Der Schritt 310 wird zyklisch wiederholt, bis ein entsprechender Anlass gegeben ist.

Wenn eine Reorganisation des Massenspeichersystems 100 gewünscht oder erforderlich ist, wird in einem nachfolgenden Schritt 320 zunächst überprüft, ob weitere Daten zur Reorganisation zur Verfügung stehen. Das Verfahren 300 zur Reorganisation kann sich grundsätzlich auf alle in dem Massenspeichersystem 100 gespeicherte Dateien, Dateien einzelner Massenspeicher 142, 144 und/oder 150 oder auch nur ausgewählte oder einzelne Dateien beziehen. Beispielsweise ist es möglich, die Zugriffswahrscheinlichkeit für eine einzelne Datei anlassbezogen, dass heißt bei einem bezüglich der Figur 2 beschriebenen Zugriff im Schritt 290 zu überprüfen.

Steht eine Datei zur Reorganisation an, wird in einem nachfolgenden Schritt 330 die Zugriffswahrscheinlichkeit für die in dem Massenspeichersystem 100 gespeicherte Datei bestimmt. Die Zugriffswahrscheinlichkeit wird in der weiteren Beschreibung anschaulich auch als "Temperatur" T der zu analysierenden Datei bezeichnet. Dabei steht eine "hohe Dateitemperatur" anschaulich für eine hohe Zugriffswahrscheinlichkeit, also für Dateien, auf die besonders häufig zugegriffen wird. Umgekehrt steht eine "niedrige Dateitemperatur" für solche Dateien, die nur selten oder überhaupt nicht mehr von dem Massenspeichersystem 100 abgerufen werden.

Zur Bestimmung der Temperatur T einer Datei sind unterschiedliche Mechanismen denkbar. Einer davon wird nachfolgend unter Bezugnahme auf die Figur 4 im Detail beschrieben. Allgemein gilt, dass die Wahrscheinlichkeit eines Zugriffs auf eine Datei höher ist, wenn auf die Datei bereits in der Vergangenheit häufig zugegriffen wurde oder der letzte Schreib- oder Lesezugriff nur eine verhältnismäßig kurze Zeitspanne zurückliegt. Umgekehrt gilt für Dateien, die in der Vergangenheit nicht oder in einem jüngeren Zeitraum nur sehr selten ange-

fordert wurden, dass sie auch in der voraussehbaren Zukunft wahrscheinlich nicht angefordert werden und daher eine niedrigere Dateitemperatur aufweisen.

Ist die Temperatur T einer Datei bestimmt, wird das Verfahren in Schritt 320 mit der Ermittlung fortgesetzt, ob die Temperatur für weitere Dateien bestimmt werden muss.

In einem nachfolgenden oder parallel ausgeführten Schritt 340 wird ein Grenzwert T_{REF} bezüglich der Abgrenzung zwischen Dateien mit hoher Zugriffswahrscheinlichkeit bzw. -temperatur und Dateien mit geringer Zugriffswahrscheinlichkeit bzw. -temperatur festgelegt. Dabei kann der Grenzwert T_{REF} entweder fest vorgegeben oder durch einen Systemadministrator frei wählbar sein.

Alternativ ist es auch möglich, den Grenzwert T_{REF} in Schritt 340 dynamisch, insbesondere anhand der Auslastung der Speicherkapazität der Massenspeicher 142, 144 und 150 festzulegen. Beispielsweise kann in Schritt 340 der Grenzwert T_{REF} so gewählt werden, dass jeweils die 20 % der gespeicherten "heißen" Daten auf dem ersten Massenspeicher 142 oder dem zweiten Massenspeicher 144 abgelegt sind, während die verbleibenden 80 % "kälterer" Daten auf dem zweiten physikalischen Massenspeicher 150 abgelegt sind. Neben einem festen Verhältnis kann das Verhältnis selbstverständlich auch dynamisch, insbesondere in Abhängigkeit der absoluten Speicherkapazität oder einer verbleibenden Speicherkapazität der Massenspeicher 142, 144 und/oder 150 gewählt werden.

Umgekehrt werden Dateien, deren in Schritt 330 bestimmte Temperatur unterhalb der in Schritt 340 bestimmten Grenztemperatur liegt, die jedoch in dem primären Speicherverbund 140 ge-

speichert sind, von dort im optionalen Schritt 360 auf den zweiten physikalischen Massenspeicher 150 verschoben.

In einem nachfolgenden Schritt 350 werden dann die Dateien, deren in Schritt 330 bestimmte Temperatur T über der in Schritt 340 bestimmten Grenztemperatur T_{REF} liegt, die jedoch auf dem zweiten physikalischen Massenspeicher 150 gespeichert sind, im Schritt 350 auf den ersten physikalischen Massenspeicher 142 und/oder 144 des Speicherverbunds 140 verschoben. Bezüglich des Schreibens der Daten wird auf die obigen Ausführungen zu den Schritten 230 bis 238 der Figur 2 verwiesen.

Ist die Reorganisation abgeschlossen, endet das Verfahren 300, wobei gegebenenfalls nicht mehr benötigte Speicherbereiche der Massenspeicher 142, 144 und 150 zur neuerlichen Verwendung freigegeben werden.

Figur 4 zeigt schematisch ein vereinfachtes Verfahren 400 zur Festlegung einer Temperatur einer Datei in vier unterschiedliche, vorgegebene Temperaturbereiche, die jeweils einer Gruppe von Dateien mit einer ähnlichen Zugriffswahrscheinlichkeit zugeordnet sind.

Gemäß dem Verfahren 400 wird in einem ersten Schritt 410 überprüft, ob auf die in Frage stehende Datei innerhalb eines vorbestimmten Zeitraums t1 schreibend zugegriffen wurde. Wurde die Datei beispielsweise innerhalb der letzten Minute erst angelegt oder überschrieben, handelt es sich um eine aktiv in Bearbeitung stehende Datei, die einer ersten Gruppe 420 mit "am heißesten" Dateien zugeordnet wird.

Wurde die Datei nicht aktuell geschrieben, wird in einem nachfolgenden Schritt 430 überprüft, wann der letzte Lesezugriff auf die Datei erfolgt ist. Erfolgte der letzte Lesezugriff innerhalb eines Zeitraums t2, beispielsweise innerhalb der letzten Stunde, wird sie einer weiteren Gruppe 440 mit "heißen Dateien" zugeordnet.

Ist auch dies nicht der Fall, wird in einer weiteren Überprüfung 450 festgestellt, wie oft insgesamt auf die Datei zugegriffen wurde. Liegt die festgestellte Anzahl von Zugriffen über einem vorbestimmten Grenzwert N, beispielsweise drei Lesezugriffen seit dem letzten Schreibzugriff, wird sie einer Gruppe 460 mit "warmen Dateien" zugeordnet. Anderenfalls wird sie einer letzten Gruppe 470 mit "kalten Dateien" zugeordnet.

Die Aufteilung gemäß dem Verfahren nach Figur 4 kann selbstverständlich in vielfältiger Weise abgewandelt oder verfeinert werden. Insbesondere können die für eine Datei in einem Dateisystem typischerweise erfassten Metadaten in vielfältiger Weise miteinander kombiniert werden, um eine mehr oder weniger kontinuierliche Verteilung von Dateitemperaturen bzw. Zugriffswahrscheinlichkeiten zu bestimmen.

Durch die pyramidenförmige Darstellung der Gruppen wird die übliche Verteilung von Zugriffswahrscheinlichkeiten angedeutet. Empirischen Studien zu Folge machen die am häufigsten verwendeten Dateien nur einen kleinen Teil der gesamten Speicherkapazität aus und umgekehrt. In der Figur 4 ist zusätzlich auch der Schwellwert T_{REF} angedeutet, der zur oben beschriebenen Ungleichbehandlung des oberen und unteren Teils der Pyramide führt.

Figur 5 zeigt eine mögliche Implementierung eines Massenspeichersystems in Form eines RAID-Systems 500 zum Abspeichern von Dateien, die vier unterschiedlichen Gruppen mit unterschiedlichen Zugriffswahrscheinlichkeiten zugeordnet sind.

Das RAID-System 500 umfasst eine Mehrzahl von so genannten NAS-Heads 510, die die in dem RAID-System 500 gespeicherten Dateien gemäß einem allgemein bekannten Protokoll, beispielsweise dem vom UNIX her bekannten Dateisystem NFS, über ein IP-Netzwerk 515 einer Mehrzahl von Clients 520 anbieten. Intern sind die NAS-Heads 510 im dargestellten Ausführungsbeispiel über eine so genannte Fibre Channel over Ethernet (FCoE) Interconnect-Struktur 530 mit einer Vielzahl von primären Massenspeichern 542, 544 und 546 eines primären Speicherverbunds 540 verbunden. Alternativ könnte die Interconnect-Struktur 530 auch als Infiniband und/oder Ethernet-Verbindung ausgestaltet sein.

Die primären Massenspeicher 542, 544 und 546 sind im Ausführungsbeispiel gemäß Figur 5 in drei Gruppen mit jeweils unterschiedlicher Leistungsfähigkeit aufgeteilt. Die erste Gruppe umfasst physikalische Massenspeicher 542 in Form von so genannten SSD-Halbleiter-Speicherlaufwerken. Die zweite Gruppe umfasst physikalische Massenspeicher 544 in Form von Massenspeicher mit rotierenden Speichermedien, die mit einer Schnittstelle gemäß dem Serial Attached SCSI (SAS) Standard ausgestattet sind. Die dritte Gruppe umfasst physikalische Massenspeicher 546 in Form von Massenspeichern mit rotierenden Speichermedien mit verhältnismäßig einfachen seriellen ATA (SATA) Schnittstellen. Die Massenspeicher 542, 544 und 546 weisen unterschiedliche Zugriffsgeschwindigkeiten auf, die beispielsweise den Gruppen 420, 440 und 460 von am heißesten, heißen und warmen Dateien zugeordnet sind. Auf diese

Weise kann für jede Art von Dateien ein bezüglich der Kosten und der Zugriffsleistung optimierter Ausgleich gefunden werden.

Das RAID-System 500 ist des Weiteren über eine so genannte IP-Interconnect-Schicht 560 mit einer Mehrzahl von externen Massenspeichern 550 verbunden. Bei den externen Massenspeichern 550 handelt es sich im Ausführungsbeispiel um so genannte Direct Attached Storage (DAS) Laufwerke, die jeweils mindestens eine Recheneinheit 570 aufweisen, die dem primären Speicherverbund 540 eine NAS-Schnittstelle anbieten. Optional sind die Recheneinheiten 570 auch dazu eingerichtet, die an die DAS-Laufwerke gerichteten Daten zu komprimieren und/oder zu deduplizieren und entsprechend zu speichern. Die externen Massenspeicher 550 sind zwar räumlich benachbart zu dem RAID-System 500 in demselben Rack angeordnet, werden jedoch davon unabhängig betrieben und über die IP-Interconnect-Schicht 560 angesprochen. Wegen der dazwischen liegenden IP-Verbindung ist die Übertragungsbandbreite zwischen den NAS-Heads 510 und den externen Massenspeichern 550 deutlich geringer als die Bandbreite zwischen den NAS-Heads 510 und den internen Massenspeichern 542, 544 oder 546. Dennoch können die in den externen Massenspeichern 550 abgelegten Daten innerhalb einer für den Benutzer kaum wahrnehmbaren Zeit über die NAS-Heads 510 bereitgestellt werden. Insbesondere erfolgt die Bereitstellung deutlich schneller als bei bekannten Archivsystemen, bei denen nicht länger in einem internen Massenspeicher zwischengespeicherte Daten erst über ein Bandlaufwerk von einem Magnetband oder von einem sonstigen Wechselspeichermedium eingelesen werden müssen.

Im beschriebnen Beispiel sind die einzelnen Komponenten des RAID-System 500 am selben Ort aufgestellt. Je nach eingesetz-

ten Interconnect-Strukturen 530 und/oder 560 können dessen Komponenten jedoch auch über mehrere Standorte verteilt werden, um eine geografische Redundanz und damit einen verbesserten Schutz gegenüber katastrophalen Ereignissen wie Bränden, Erdbeben etc. zu erreichen.

Da auf die externen Massenspeicher 550 nur sehr selten zugegriffen wird, werden diese Massenspeicher 550 bevorzugt in einen Energiesparzustand versetzt, wenn über einen vorbestimmten Zeitraum von beispielsweise mehreren Minuten oder Stunden kein Zugriff auf den jeweiligen Massenspeicher 550 mehr stattgefunden hat. Da statistisch gesehen die meisten Daten, die von einem Massenspeichersystem 100 vorgehalten werden, der Gruppe 470 von kalten Dateien zugeordnet werden können, empfiehlt es sich weiterhin, diese Daten gegebenenfalls in einem dedupliziertен und komprimierten oder zumindest komprimierten Zustand vorzuhalten. Zu diesem Zweck sind zwischen der IP-Interconnect-Schicht 560 und den eigentlichen physikalischen Massenspeichern 550 die Recheneinheiten 570 vorgesehen, die für eine Deduplizierung der Daten vor deren Speicherung auf den externen Massenspeichern 550 sorgen. Der mit der Duplizierung einhergehende zusätzliche Aufwand muss nur einmal oder zumindest nur verhältnismäßig selten erbracht werden, sofern die Daten erwartungsgemäß nicht wieder von dem RAID-System 500 abgerufen oder geändert werden.

Das RAID-System 500 gemäß der Figur 5 verteilt die in ihm gespeicherten Daten wie zuvor unter Bezugnahme auf die Figuren 1 bis 4 beschrieben, basierend auf einer statistisch zu erwartenden Zugriffswahrscheinlichkeit. Dabei werden besonders heiße Dateien durch den Einsatz besonders leistungsfähiger Hardware schnell verfügbar vorgehalten. Kalte Dateien können dagegen besonderes energiesparend auf einfacher Hardware vor-

gehalten werden, ohne dass die geschilderten Nachteile bekannter Archivlösungen für den Benutzer spürbar werden. Für den Betreiber des Systems ergibt sich der weitere Vorteil, dass sich mit moderatem wirtschaftlichem bzw. technischem Aufwand ein RAID-System mit einer aus Benutzersicht sehr hohen Gesamtleistung aufbauen lässt.

Figur 6 zeigt eine logische Ansicht von in einem Massenspeichersystem 600, beispielsweise dem Massenspeichersystem 100 oder dem RAID-System 500, abgelegten Daten. Darin ist insbesondere das Zusammenwirken unterschiedlicher Dateisysteme zu erkennen.

Für einen Benutzer des Massenspeichersystems 600 ist nach außen nur ein einziges Dateisystem 610 sichtbar. Bei dem Dateisystem 610 handelt es sich im technischen Sinne um ein virtuelles oder logisches Dateisystem, da in dem Massenspeichersystem 600 kein physikalischer Datenträger existiert, auf dem sämtliche in dem Dateisystem 610 erfassten Dateien abgelegt sind. Der Zugriff auf das Dateisystem 610 erfolgt in an sich bekannter Weise, beispielsweise über das NFS-Protokoll.

In der Figur 6 ist angedeutet, dass in dem Dateisystem 610 zwei Dateien 620 und 630 gespeichert sind. Die Dateien 620 bzw. 630 umfassen dabei neben ihrem eigentlichen Inhalt 622 bzw. 632 zusätzliche Metainformationen. Beispielsweise sind in dem virtuellen Dateisystem 610 für die Dateien 620 und 630 ein Zeitpunkt des letzten Zugriffs 624 bzw. 634 sowie ein Zeitpunkt des letzten Schreibens 626 bzw. 636 festgehalten. Darüber hinaus können weitere, in der Figur 6 nicht dargestellte Metainformationen wie Zugriffsrechte und Zugriffshäufigkeiten und weitere Attribute zu jeder Datei gespeichert werden. Die Metadateien müssen dabei nicht vollständig über

das virtuelle Dateisystem 610 für einen Benutzer zugänglich sein, sondern können auch nur der internen Organisation innerhalb des Massenspeichersystems 600 dienen.

Weiter ist in Figur 6 dargestellt, dass die eigentlichen Inhalte 622 bzw. 632 der ersten Datei 620 bzw. der zweiten Datei 630 logisch und physisch in mehrere Datensegmente 628 bzw. 638 mit Inhalten D1, D2, D3 und D4 bzw. D5, D6 und D7 aufgeteilt werden können. Bei den Dateisegmenten 628 und 638 kann es sich um Datenblöcke einer festen Größe, beispielsweise einer üblichen Blockgröße von 512 Byte, 4 kByte oder eine sonstige an die Struktur der verwendeten Massenspeicher angepasste Blockgröße handeln. Selbstverständlich können auch von einem Systemadministrator vorgegebene Blockgrößen oder relative Blockgrößen, beispielsweise jeweils ein bestimmter Bruchteil der Gesamtgröße einer Datei, gewählt werden.

Im in der Figur 6 dargestellten Zustand ergibt sich aus den Zeitpunkten 624 und 626 der ersten Datei 620, dass es sich um eine so genannte heiße Datei mit einer hohen Zugriffswahrscheinlichkeit handelt. Bildlich dargestellt liegt die Temperatur T1 der ersten Datei 620 über einer vorgegebenen Referenztemperatur T_{REF}. Deshalb wird die Datei 620 durch das Massenspeichersystem 600 auf ein verteiltes Clusterdateisystem 650 abgebildet, über das physikalische Massenspeicher 642, 644 und 646 eines primären Speicherverbunds 640 angesprochen werden.

Auch bei dem Cluster-Dateisystem 650 handelt es sich um ein logisches oder virtuelles Dateisystem, weil die darüber adressierbaren Daten nicht physikalisch auf einem einzigen Massenspeicher existieren, sondern über die zwei physikalischen Massenspeicher 642 und 644 verteilt gespeichert werden.

Beispielsweise kann ein erstes Dateisegment 628 mit Inhalt D1 der ersten Datei 620 auf dem Massenspeicher 642 und ein zweites Dateisegment 628 mit dem Inhalt D2 der Datei 620 auf dem physikalischen Massenspeicher 644 gespeichert werden. Weitere Dateisegmente 628 der Datei 620 werden in ähnlicher Weise ebenfalls abwechselnd über die physikalischen Massenspeicher 642 und 644 des primären Speicherverbunds 640 verteilt.

Zusätzlich ist in der Figur 6 angedeutet, dass weitere zugehörige Daten der Datei 620, insbesondere der mit den auf den Massenspeichern 642 und 644 gespeicherten Datensegmente 628 verknüpfte Paritätsdaten D12 und D34 auf einem weiteren Massenspeicher 646 des primären Speicherverbunds 140 abgelegt werden können. Alternativ können auch weitere Kopien der Segmente selbst in einem weiteren Massenspeicher gespeichert werden. Durch die zusätzlichen Daten auf dem zusätzlichen Massenspeicher 646 kann eine Redundanz der gespeicherten Datei 620 und somit eine Sicherheit gegen einen Ausfall eines der Massenspeicher 642 oder 644 geschaffen werden. Die dabei eingesetzten Verfahren zum redundanten Speichern von Daten und deren Wiederherstellung bei Ausfall eines oder mehrerer Massenspeicher sind aus dem Stand der Technik bekannt und werden daher hier nicht im Detail beschrieben.

In der Figur 6 ist weiter ersichtlich, dass basierend auf dem Lesezeitraum 634 und dem Schreibzeitraum 636 für die zweite Datei 630 ermittelt wurde, dass es sich um eine so genannte kalte Datei handelt, die höchstwahrscheinlich nicht oder nur noch sehr selten von dem Massenspeichersystem 600 abgerufen wird. Diese Datei 630 ist zusammenhängend auf einem sekundären Massenspeicher 660 gespeichert.

Zusammenhängend heißt in diesem Fall in erster Linie, dass sämtliche Dateisegmente 638 der zweiten Datei in einer gemeinsamen Partition des sekundären Massenspeichers 660 abgelegt sind und/oder über ein gemeinsames, für den Massenspeicher 660 spezifisches Dateisystem 662 abrufbar sind. Bevorzugt sind diese Dateisegmente in logisch aufeinander abfolgenden Blöcken gespeichert, wobei diese entsprechend dem Steueralgorithmus des Massenspeichers 660 durchaus über räumlich getrennte Speichersektoren verteilt sein können.

Um auch bezüglich der in dem Massenspeicher 660 gespeicherten Daten eine Redundanz und somit eine Sicherheit gegenüber dem Ausfall des Massenspeichers 660 sicherstellen zu können, sind die darin gespeicherten Daten auf einem zweiten sekundären Massenspeicher 670 mit einem dafür spezifischen Dateisystem 672 gespiegelt. Somit sind die Daten der zweiten Datei 630 auch auf dem zweiten Massenspeichers 670 zusammenhängend gespeichert, der beim Ausfall des ersten sekundären Massenspeichers 660 dessen Funktion übernehmen kann.

Anstelle der oben beschriebenen Spiegelung auf einem zweiten sekundären Massenspeicher 670 kann diese auch auf ein an sich aus der Archivierung bekanntes, relativ langsames tertiäres Speichermedium wie ein Magnetband oder mittels Cloud-Storage erfolgen. Der zum Wiedereinlesen des Speichermediums erforderliche Zeitaufwand ist dabei fast ohne praktische Bedeutung, da auf die Dateien des sekundären Massenspeichers 660 ohnehin nur sehr selten zugegriffen wird und selbst dann das Rücksichern vom Band beim ohnehin nur unwahrscheinlichen Ausfall des sekundären Massenspeichers 660 zur selben Zeit erforderlich wäre. Statistisch gesehen ist ein Aufeinandertreffen dieser unkorrelierten Ereignisse sehr unwahrscheinlich,

so dass es keine spürbare Auswirkung auf den Normalbetrieb des Massenspeichersystems 600 hat.

Zusammengefasst können die Massenspeichersysteme 100 und 600 bzw. das RAID-System 500 umfangreiche Daten zentral auf besonders energieeffiziente und leistungsfähige Weise speichern. Dabei dient eine Gruppierung von Dateien gemäß gleicher oder ähnlicher Zugriffswahrscheinlichkeit bzw. Dateitemperatur dazu, dass Dateien, auf die besonders häufig zugegriffen wird, in besonders schnellen, leistungsfähigen und damit teuren Datenspeichern abgelegt werden können. Umgekehrt werden solche Dateien, auf die nie oder nur selten zugegriffen wird, in verhältnismäßig großen und kostengünstigen, daher jedoch auch verhältnismäßig leistungsschwachen Massenspeichern abgelegt. Neben der bereits dadurch erreichten Performance-Verbesserung dient die Gruppierung von Daten auch dazu, dass Dateien mit besonders niedriger Dateitemperatur, also Dateien, auf die wahrscheinlich nie wieder zugegriffen wird, auf einem gemeinsamen physikalischen Massenspeicher gespeichert werden. Für diesen physikalischen Massenspeicher führt eine zumindest vorübergehende Abschaltung somit zu einer signifikanten Energieeinsparung, ohne dass dadurch die Leistung des Gesamtsystems für einen Benutzer erkennbar einbricht.

Bezugszeichenliste

100	Massenspeichersystem
110	Netzwerkschnittstelle
120	Steuervorrichtung
130	Datenbank
140	primärer Speicherverbund
142	physikalischer Massenspeicher
144	physikalischer Massenspeicher
150	physikalischer Massenspeicher
160	primäre Verbindungsstruktur
170	sekundäre Verbindungsstruktur
180	Netzwerk
190	Client Computer
200	Verfahren
210 - 290	Verfahrensschritte
300	Verfahren
310 - 360	Verfahrensschritte
400	Verfahren
410	Verfahrensschritt
420	Gruppe
430	Verfahrensschritt
440	Gruppe
450	Verfahrensschritt
460	Gruppe
470	Gruppe
500	RAID-System
510	NAS-Head
515	IP-Netzwerk

520 Client
530 Interconnect-Struktur
540 primärer Speicherverbund
542 primärer Massenspeicher
544 primärer Massenspeicher
546 primärer Massenspeicher
550 externer Massenspeicher
560 IP-Interconnect-Schicht
570 Recheneinheit

600 Massenspeichersystem
610 virtuelles Dateisystem
620 erste Datei
622 Dateninhalt
624 Lesedatum
626 Schreibdatum
628 Dateisegment
630 zweite Datei
632 Dateiinhalt
634 Lesedatum
636 Schreibdatum
638 Dateisegment
640 primärer Speicherverbund
642 physikalischer Massenspeicher
644 physikalischer Massenspeicher
646 physikalischer Massenspeicher
650 Clusterdateisystem
660 Massenspeicher
662 Dateisystem
670 Massenspeicher
672 Dateisystem

Patentansprüche

1. Arbeitsverfahren für ein Massenspeichersystem (100, 600), insbesondere ein RAID-System (500), mit den Schritten:
- Bereitstellen eines virtuellen Dateisystems (610) für wenigstens einen Benutzer des Massenspeichersystems (100, 600);
- Bestimmen einer Zugriffswahrscheinlichkeit für in dem virtuellen Dateisystem (610) logisch gespeicherte Dateien (620, 630);
- verteiltes Speichern von Dateien (620), deren Zugriffswahrscheinlichkeit über einem vorbestimmten Grenzwert (T_{REF}) liegt, in einer Mehrzahl von voneinander unabhängigen ersten physikalischen Massenspeichern (642, 644, 646) mit voneinander unabhängigen Schreib-/Leseeinheiten; und
- gemeinsames Speichern von Dateien (630), deren Zugriffswahrscheinlichkeit unter dem vorbestimmten Grenzwert (T_{REF}) liegt, in wenigstens einem zusammenhängenden Speicherbereich wenigstens eines zweiten physikalischen Massenspeichers (660).

2. Arbeitsverfahren nach Anspruch 1, mit den zusätzlichen Schritten:
- Schalten des wenigstens einen zweiten physikalischen Massenspeichers (660) in einen Betriebszustand mit gegenüber einem normalen Betriebszustand reduzierter Energieaufnahme, wenn über einen vorbestimmten Zeitraum kein Zugriff auf den wenigstens einen zweiten physikalischen Massenspeicher (660) durchgeführt wurde; und
- Schalten des wenigstens einen zweiten physikalischen Massenspeichers (660) in den normalen Betriebszustand, wenn eine Zugriffsanforderung für wenigstens eine auf dem wenigstens einen zweiten physikalischen Massenspeicher

(660) gespeicherte Datei (630) über das virtuelle Dateisystem (610) erfasst wird.

3. Arbeitsverfahren nach Anspruch 1 oder 2, bei dem die in dem virtuellen Dateisystem (610) logisch gespeicherten Dateien (620, 630) redundant auf unterschiedlichen Massenspeichern (642, 644, 646, 660, 670) gespeichert werden.

4. Arbeitsverfahren nach einem der Ansprüche 1 bis 3, bei dem im Schritt des Bestimmens der Zugriffswahrscheinlichkeit die in dem virtuellen Dateisystem (610) logisch gespeicherten Dateien (620, 630) in wenigstens zwei Gruppen (420, 470) mit unterschiedlichen Zugriffswahrscheinlichkeitsbereichen aufgeteilt werden, wobei Dateien (620), die einer ersten Gruppe (420) zugeordnet wurden, verteilt in der Mehrzahl von voneinander unabhängigen ersten physikalischen Massenspeichern (642, 644, 646) gespeichert werden, und Dateien (630), die einer zweiten Gruppe (470) zugeordnet wurden, gemeinsam in dem wenigstens einen zusammenhängenden Speicherbereich des wenigstens einen zweiten physikalischen Massenspeichers (660) gespeichert werden.

5. Arbeitsverfahren nach Anspruch 4, bei dem im Schritt des verteilten Speicherns Dateien (620) der ersten Gruppe in Segmente (628) einer vorbestimmten Größe aufgeteilt werden, wobei Segmente (628) einer Datei (620) mit mehr als einem Segment (628) verteilt auf der Mehrzahl von voneinander unabhängigen Massenspeichern (642, 644, 646) gespeichert werden.

6. Arbeitsverfahren nach Anspruch 4 oder 5, bei dem im Schritt des Bestimmens einer Zugriffswahrscheinlichkeit die in dem virtuellen Dateisystem (610) logisch gespeicherten Dateien (620, 630) in eine Mehrzahl von Gruppen (420, 440, 460, 470) mit unterschiedlichen Zugriffswahrscheinlichkeitsbereichen aufgeteilt werden, wobei jeder Gruppe (420, 440, 460), deren Zugriffswahrscheinlichkeitsbereich über dem vorbestimmten Grenzwert (T_{REF}) liegt, eine Mehrzahl von physikalischen Massenspeichern (542, 544, 546) zum verteilten Speichern der der Gruppe zugeordneten Dateien (620) zugeordnet ist und jeder Gruppe (470), deren Zugriffswahrscheinlichkeitsbereich unter dem vorbestimmten Grenzwert (T_{REF}) liegt, wenigstens ein physikalischer Massenspeicher (550) zum gemeinsamen Speichern der der Gruppe (470) zugeordneten Dateien (630) zugeordnet ist.

7. Arbeitsverfahren nach einem der Ansprüche 1 bis 6, bei dem die Zugriffswahrscheinlichkeit für eine Datei (620, 630), basierend auf wenigstens einem der folgenden Parameter bestimmt wird: Zeitpunkt der Erstellung der Datei, Zeitpunkt des letzten Schreibzugriffs auf die Datei, Zeitpunkt des letzten Lesezugriffs auf die Datei und Anzahl von Lese- und/oder Schreibzugriffen innerhalb eines vorbestimmten Zeitraums auf die Datei.

8. Massenspeichersystem (100, 600), insbesondere RAID-System (500), umfassend:

- eine Mehrzahl von voneinander unabhängigen ersten physikalischen Massenspeichern (642, 644, 646) mit voneinander unabhängigen Schreib-/Leseeinheiten;
- wenigstens einen zweiten physikalischen Massenspeicher (660);

- wenigstens eine Schnittstellenvorrichtung zum Bereitstellen eines virtuellen Dateisystems (610) für wenigstens einen Benutzer des Massenspeichersystems (100, 600); und
- wenigstens eine Steuervorrichtung (120) zum wahlweisen Speichern der in dem virtuellen Dateisystem (610) logisch gespeicherten Dateien (620, 630) auf den ersten physikalischen Massenspeichern (642, 644, 646) oder dem wenigstens einen zweiten physikalischen Massenspeicher (660), wobei die Steuervorrichtung (120) dazu eingerichtet ist, für die in dem virtuellen Dateisystem (610) logisch gespeicherten Dateien eine Zugriffswahrscheinlichkeit zu bestimmen, Dateien (620), deren Zugriffswahrscheinlichkeit über einem ersten vorbestimmten Grenzwert (T_{REF}) liegt, in einer Mehrzahl von voneinander unabhängigen ersten physikalischen Massenspeichern (642, 644, 646) mit voneinander unabhängigen Schreib-/Leseeinheiten verteilt zu speichern, und Dateien (630), deren Zugriffswahrscheinlichkeit unter dem ersten vorbestimmten Grenzwert (T_{REF}) liegt, in wenigstens einem zusammenhängenden Speicherbereich des wenigstens einen zweiten physikalischen Massenspeichers (660) gemeinsam zu speichern.

9. Massenspeichersystem (100, 600) nach Anspruch 8, bei dem der wenigstens eine zweite physikalische Massenspeicher (660) eine Antriebseinheit und wenigstens ein durch die Antriebseinheit rotatorisch antreibbares Speichermedium umfasst, wobei die Steuervorrichtung (120) oder der wenigstens eine zweite physikalische Massenspeicher (660) dazu eingerichtet ist, die Antriebseinheit abzuschalten, wenn über einen vorbestimmten Zeitraum kein Zugriff auf den wenigstens einen zweiten physikalischen Massenspeicher (660) erfolgt.

10. Massenspeichersystem (100, 600) nach Anspruch 8 oder 9, umfassend wenigstens zwei zweite physikalische Massenspeicher (660, 670), wobei auf einem ersten der wenigstens zwei zweiten physikalischen Massenspeicher (660) gespeicherte Dateien (630) auf wenigstens einem zweiten der wenigstens zwei zweiten physikalischen Massenspeicher (670) redundant gespeichert werden.

11. Massenspeichersystem (100, 600) nach einem der Ansprüche 8 bis 10, bei dem auf einem ersten der Mehrzahl von ersten physikalischen Massenspeichern (642) gespeicherte Dateien (620) oder Segmente (628) von Dateien (620) auf wenigstens einem zweiten der Mehrzahl von ersten Massenspeichern (642, 644, 646) redundant gespeichert werden.

12. Massenspeichersystem (100, 600) nach einem der Ansprüche 8 bis 11, bei dem die Mehrzahl von ersten physikalischen Massenspeichern (642, 644, 646) über eine erste, lokale Verbindungsstruktur (160) mit einer ersten Bandbreite, insbesondere gemäß dem Fibre Channel oder RDMA-Protokoll, mit der wenigstens einen Schnittstellenvorrichtung verbunden ist; und der wenigstens eine zweite physikalische Massenspeicher (660) über eine zweite Verbindungsstruktur (170) mit einer gegenüber der ersten Bandbreite geringeren zweiten Bandbreite, insbesondere gemäß dem TCP/IP-Protokoll, mit der wenigstens einen Schnittstellenvorrichtung verbunden ist.

13. Massenspeichersystem (100, 600) nach einem der Ansprüche 8 bis 12, bei dem die Mehrzahl von ersten physikalischen Massenspeichern (642, 644, 646) von der Steuervorrichtung (120) über ein gemeinsames, verteiltes erstes Dateisystem (650) verwaltet wird und der wenigstens eine zweite phy-

sikalische Massenspeicher (660) über ein für den Massenspeicher spezifisches zweites Dateisystem (662) verwaltet wird.

14. Massenspeichersystem (100, 600) nach einem der Ansprüche 8 bis 13, bei dem die Mehrzahl von ersten physikalischen Massenspeichern (542, 544, 546) wenigstens eine erste Untergruppe von Massenspeichern (542) mit einer ersten Zugriffsgeschwindigkeit, insbesondere Massenspeicher (542) mit wenigstens einem Halbleiter-Massenspeichermedium, und eine zweite Untergruppe von Massenspeichern (544, 546) mit einer gegenüber der ersten Zugriffsgeschwindigkeit geringeren zweiten Zugriffsgeschwindigkeit, insbesondere Massenspeicher (544, 546) mit wenigstens einem rotierenden Massenspeichermedium, umfasst; und bei dem die Steuervorrichtung (120) dazu eingerichtet ist, Dateien, deren Zugriffswahrscheinlichkeit über einem gegenüber dem ersten vorbestimmten Grenzwert (T_{REF}) größeren zweiten vorbestimmten Grenzwert liegt, in einer Mehrzahl von Massenspeichern (542) der ersten Untergruppe verteilt zu speichern und Dateien, deren Zugriffswahrscheinlichkeit unter dem zweiten vorbestimmten Grenzwert und über dem ersten vorbestimmten Grenzwert (T_{REF}) liegt, in einer Mehrzahl von Massenspeichern (544, 546) der zweiten Untergruppe verteilt zu speichern.

15. Computerprogrammprodukt, umfassend Programmanweisungen, wobei beim Ausführen der Programmanweisungen auf wenigstens einer Datenverarbeitungseinheit eines elektronischen Datenverarbeitungssystems die folgenden Schritte ausgeführt werden:

- Bestimmen einer Zugriffswahrscheinlichkeit für in einem virtuellen Dateisystem (610) wenigstens eines Benutzers logisch gespeicherte Dateien (620, 630);
- verteiltes Speichern von Dateien (620), deren Zugriffswahrscheinlichkeit über einem vorbestimmten Grenzwert (T_{REF}) liegt, in einer Mehrzahl von voneinander unabhängigen ersten physikalischen Massenspeichern (642, 644, 646) des elektronischen Datenverarbeitungssystems mit voneinander unabhängigen Schreib-/Leseeinheiten; und
- und gemeinsames Speichern von Dateien (630), deren Zugriffswahrscheinlichkeit unter dem vorbestimmten Grenzwert (T_{REF}) liegt, in wenigstens einem zusammenhängenden Speicherbereich wenigstens eines zweiten physikalischen Massenspeichers (660) des elektronischen Datenverarbeitungssystems.

FIG 1

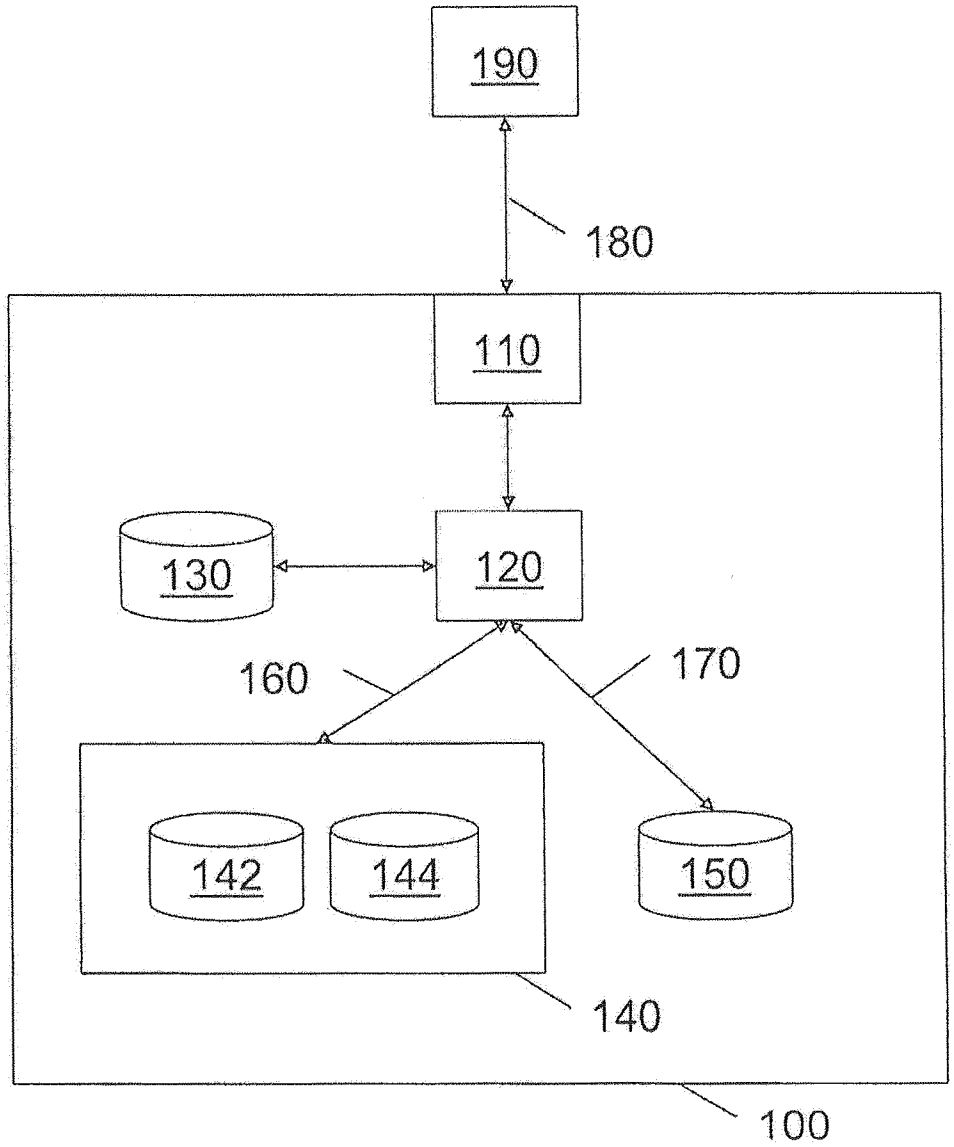


FIG 2

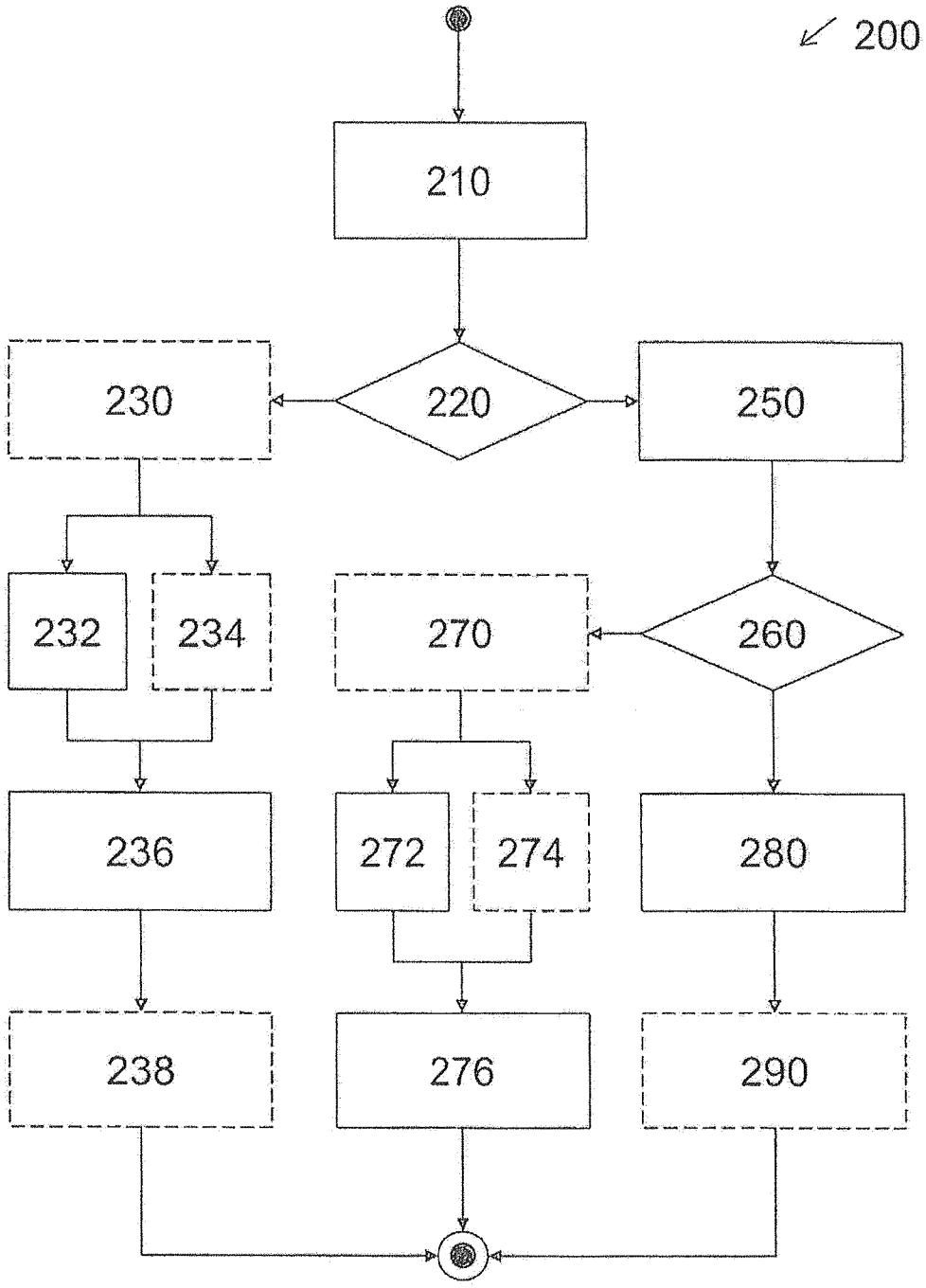
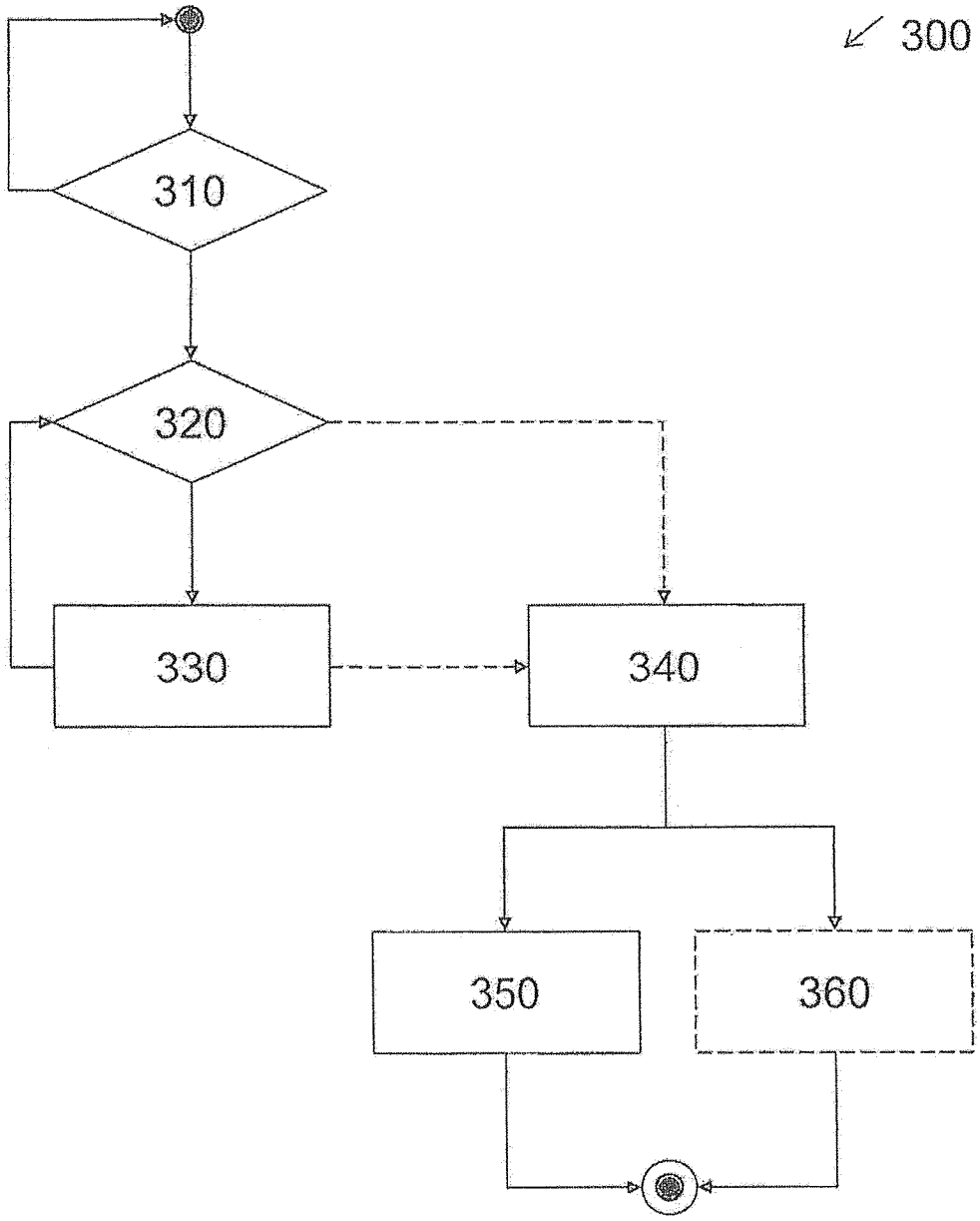


FIG 3



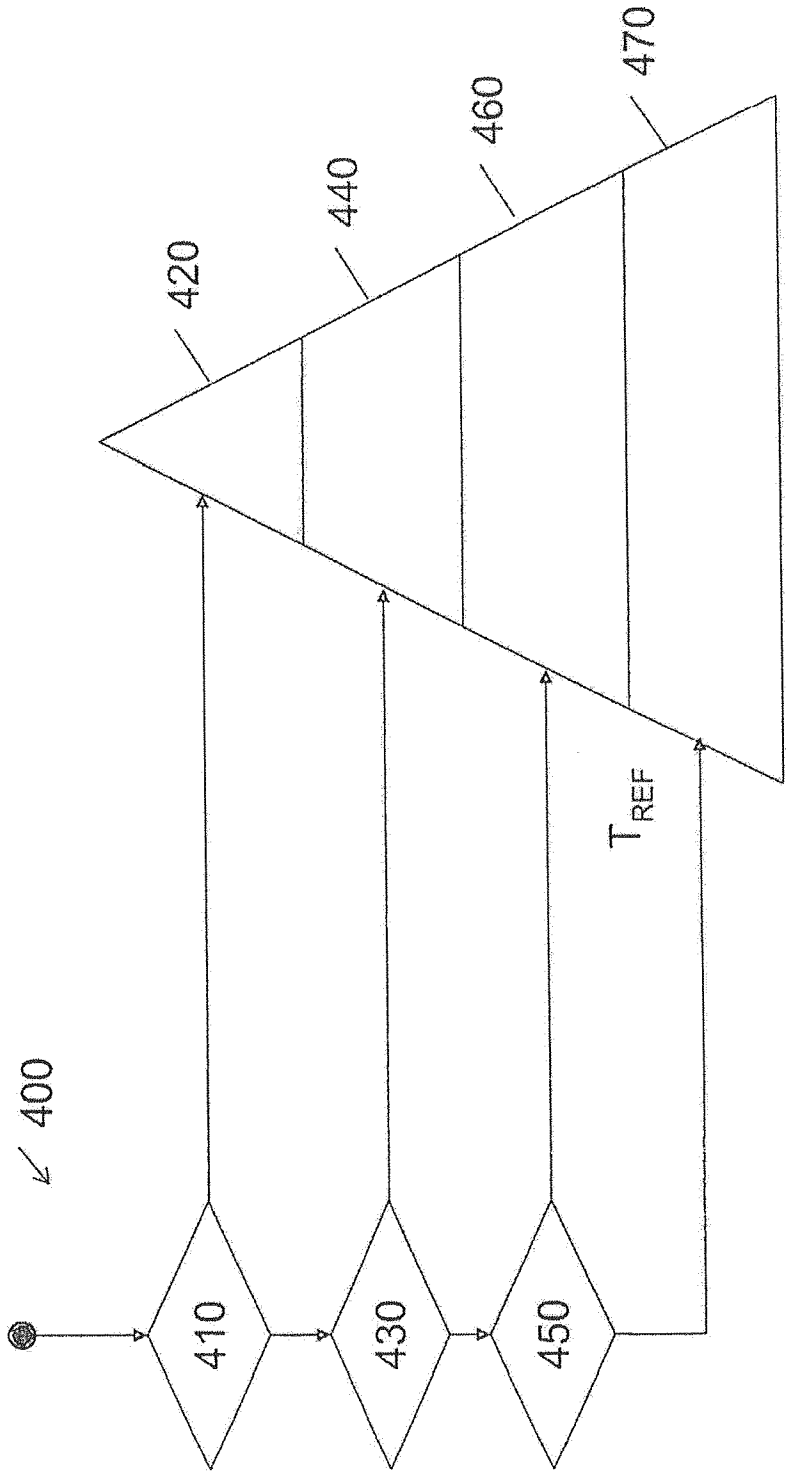


FIG 4

FIG 5

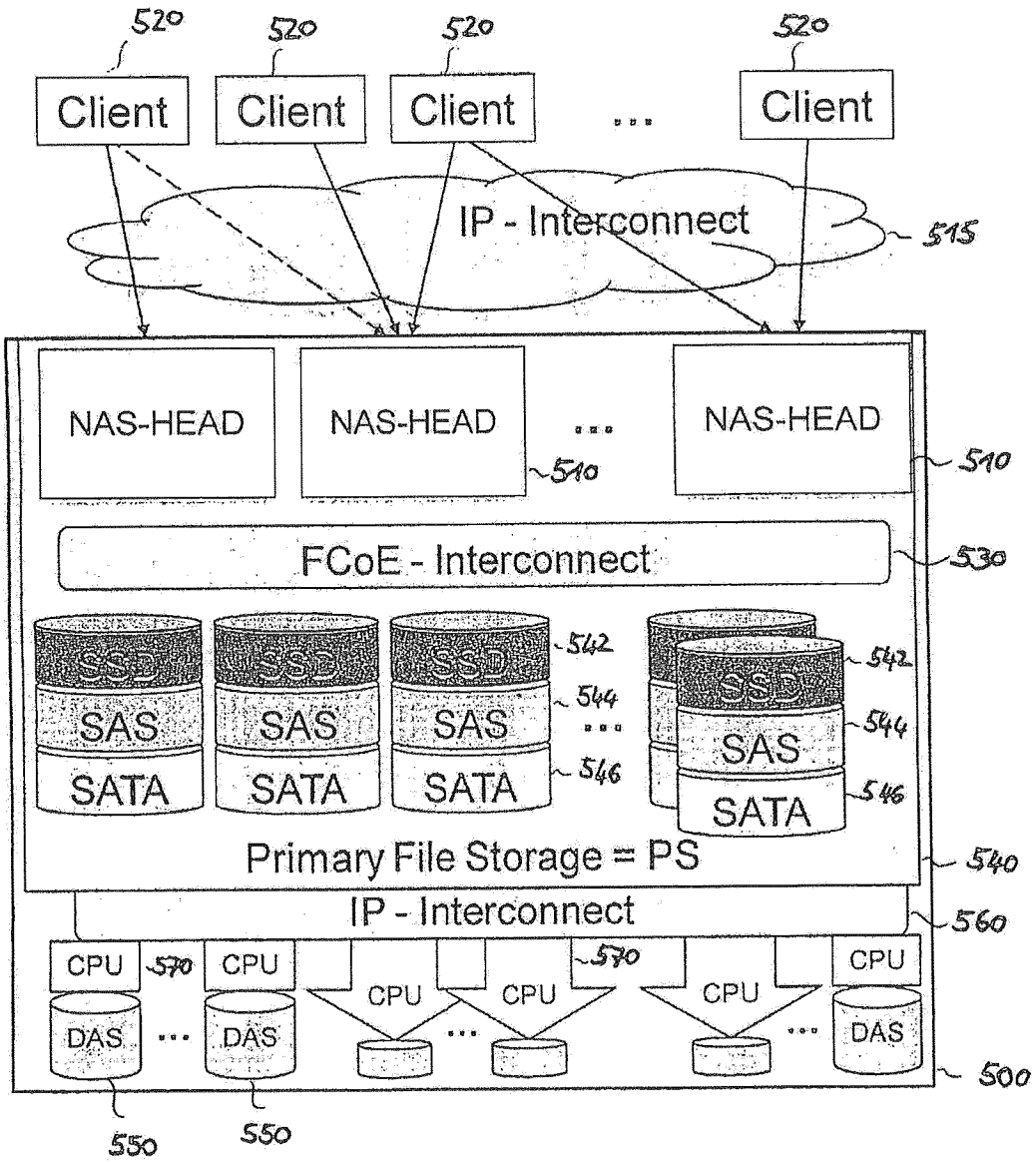
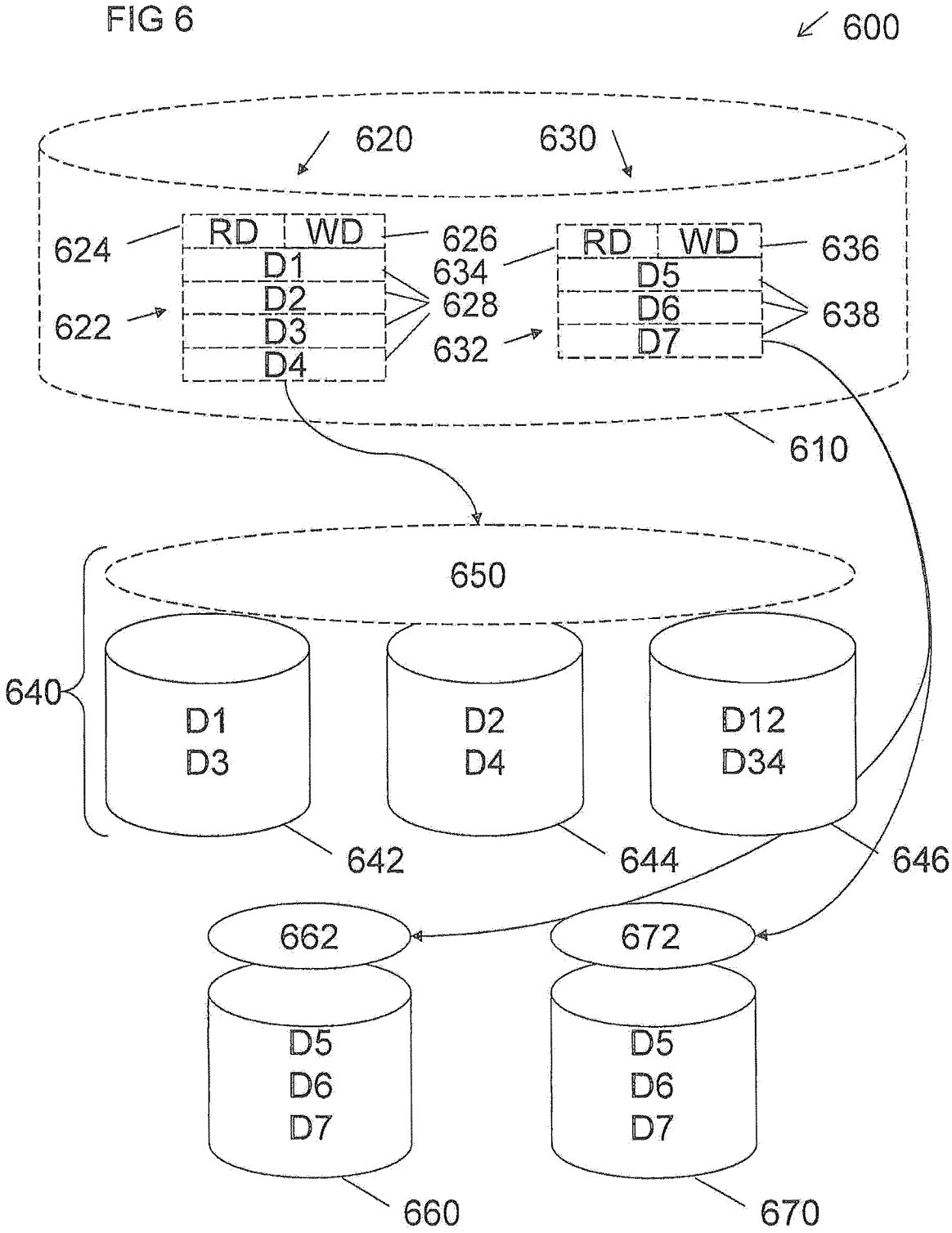

520
520
520
520
Client
Client
Client
...
Client
IP - Interconnect
515
NAS-HEAD
NAS-HEAD
...
NAS-HEAD
510
510
FCoE - Interconnect
530
SSD
SAS
SATA
542
544
546
Primary File Storage = PS
540
IP - Interconnect
560
CPU
570
DAS
550
500

FIG 6


600
620
630
624
RD
WD
626
D1
D2
D3
D4
622
634
628
632
RD
WD
636
D5
D6
D7
638
610
650
640
D1
D3
D2
D4
D12
D34
642
644
646
662
672
D5
D6
D7
D5
D6
D7
660
670

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2013/066399

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F3/06
ADD.

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06F

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2003/046270 A1 (LEUNG ALBERT [US] ET AL) 6 March 2003 (2003-03-06) paragraphs [0007] - [0012], [0023] - [0032], [0043], [0073] - [0075]; figures 1,3	1-15
X	WO 2004/021123 A2 (ARKIVIO INC [US]; MU YUEDONG [US]; LEUNG ALBERT [US]) 11 March 2004 (2004-03-11) paragraphs [0003] - [0005], [0025], [0039]; figure 1	1-15
X	EP 1 462 927 A2 (HITACHI LTD [JP]) 29 September 2004 (2004-09-29) paragraphs [0002] - [0003], [0020] - [0021], [0061], [0068], [0138], [0179]; figures 1,8,11	1-15

-/--

[X] Further documents are listed in the continuation of Box C.

[X] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
23 September 2013	02/10/2013

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Limacher, Rolf

1

Form PCT/ISA/210 (second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2013/066399

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 6 490 666 B1 (CABRERA LUIS FELIPE [US] ET AL) 3 December 2002 (2002-12-03) paragraphs [0003] - [0004], [0008], [0029]	1-15

1

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/EP2013/066399

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2003046270 A1	06-03-2003	NONE	
WO 2004021123 A2	11-03-2004	AU 2003262920 A1	19-03-2004
		US 2004083202 A1	29-04-2004
		US 2007288430 A1	13-12-2007
		WO 2004021123 A2	11-03-2004
EP 1462927 A2	29-09-2004	CN 1570842 A	26-01-2005
		CN 101034340 A	12-09-2007
		EP 1462927 A2	29-09-2004
		JP 4322031 B2	26-08-2009
		JP 2004295457 A	21-10-2004
		US 2004193760 A1	30-09-2004
		US 2005119994 A1	02-06-2005
		US 2005203964 A1	15-09-2005
		US 2008263277 A1	23-10-2008
		US 2011185123 A1	28-07-2011
US 6490666 B1	03-12-2002	US 6490666 B1	03-12-2002
		US 2003056069 A1	20-03-2003

INTERNATIONALER RECHERCHENBERICHT

Internationales Aktenzeichen
PCT/EP2013/066399

A. KLASSIFIZIERUNG DES ANMELDUNGSGEGENSTANDES
INV. G06F3/06
ADD.

Nach der Internationalen Patentklassifikation (IPC) oder nach der nationalen Klassifikation und der IPC

B. RECHERCHIERTE GEBIETE

Recherchierter Mindestprüfstoff (Klassifikationssystem und Klassifikationssymbole)
G06F

Recherchierte, aber nicht zum Mindestprüfstoff gehörende Veröffentlichungen, soweit diese unter die recherchierten Gebiete fallen

Während der internationalen Recherche konsultierte elektronische Datenbank (Name der Datenbank und evtl. verwendete Suchbegriffe)
EPO-Internal, WPI Data

C. ALS WESENTLICH ANGESEHENE UNTERLAGEN

Kategorie*	Bezeichnung der Veröffentlichung, soweit erforderlich unter Angabe der in Betracht kommenden Teile	Betr. Anspruch Nr.
X	US 2003/046270 A1 (LEUNG ALBERT [US] ET AL) 6. März 2003 (2003-03-06) Absätze [0007] - [0012], [0023] - [0032], [0043], [0073] - [0075]; Abbildungen 1,3	1-15
X	WO 2004/021123 A2 (ARKIVIO INC [US]; MU YUEDONG [US]; LEUNG ALBERT [US]) 11. März 2004 (2004-03-11) Absätze [0003] - [0005], [0025], [0039]; Abbildung 1	1-15
X	EP 1 462 927 A2 (HITACHI LTD [JP]) 29. September 2004 (2004-09-29) Absätze [0002] - [0003], [0020] - [0021], [0061], [0068], [0138], [0179]; Abbildungen 1,8,11	1-15

-/--

[X] Weitere Veröffentlichungen sind der Fortsetzung von Feld C zu entnehmen [X] Siehe Anhang Patentfamilie

* Besondere Kategorien von angegebenen Veröffentlichungen :
"A" Veröffentlichung, die den allgemeinen Stand der Technik definiert, aber nicht als besonders bedeutsam anzusehen ist
"E" frühere Anmeldung oder Patent, die bzw. das jedoch erst am oder nach dem internationalen Anmeldedatum veröffentlicht worden ist
"L" Veröffentlichung, die geeignet ist, einen Prioritätsanspruch zweifelhaft erscheinen zu lassen, oder durch die das Veröffentlichungsdatum einer anderen im Recherchenbericht genannten Veröffentlichung belegt werden soll oder die aus einem anderen besonderen Grund angegeben ist (wie ausgeführt)
"O" Veröffentlichung, die sich auf eine mündliche Offenbarung, eine Benutzung, eine Ausstellung oder andere Maßnahmen bezieht
"P" Veröffentlichung, die vor dem internationalen Anmeldedatum, aber nach dem beanspruchten Prioritätsdatum veröffentlicht worden ist
"T" Spätere Veröffentlichung, die nach dem internationalen Anmeldedatum oder dem Prioritätsdatum veröffentlicht worden ist und mit der Anmeldung nicht kollidiert, sondern nur zum Verständnis des der Erfindung zugrundeliegenden Prinzips oder der ihr zugrundeliegenden Theorie angegeben ist
"X" Veröffentlichung von besonderer Bedeutung; die beanspruchte Erfindung kann allein aufgrund dieser Veröffentlichung nicht als neu oder auf erfinderischer Tätigkeit beruhend betrachtet werden
"Y" Veröffentlichung von besonderer Bedeutung; die beanspruchte Erfindung kann nicht als auf erfinderischer Tätigkeit beruhend betrachtet werden, wenn die Veröffentlichung mit einer oder mehreren Veröffentlichungen dieser Kategorie in Verbindung gebracht wird und diese Verbindung für einen Fachmann naheliegend ist
"&" Veröffentlichung, die Mitglied derselben Patentfamilie ist

Datum des Abschlusses der internationalen Recherche	Absendedatum des internationalen Recherchenberichts
23. September 2013	02/10/2013

Name und Postanschrift der Internationalen Recherchenbehörde	Bevollmächtigter Bediensteter
Europäisches Patentamt, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Limacher, Rolf

1

INTERNATIONALER RECHERCHENBERICHT

Internationales Aktenzeichen
PCT/EP2013/066399

C. (Fortsetzung) ALS WESENTLICH ANGESEHENE UNTERLAGEN

Kategorie*	Bezeichnung der Veröffentlichung, soweit erforderlich unter Angabe der in Betracht kommenden Teile	Betr. Anspruch Nr.
X	US 6 490 666 B1 (CABRERA LUIS FELIPE [US] ET AL) 3. Dezember 2002 (2002-12-03) Absätze [0003] - [0004], [0008], [0029] -----	1-15

1

INTERNATIONALER RECHERCHENBERICHT

Angaben zu Veröffentlichungen, die zur selben Patentfamilie gehören

Internationales Aktenzeichen
PCT/EP2013/066399

Im Recherchenbericht angeführtes Patentdokument	Datum der Veröffentlichung	Mitglied(er) der Patentfamilie	Datum der Veröffentlichung
US 2003046270 A1	06-03-2003	KEINE	
WO 2004021123 A2	11-03-2004	AU 2003262920 A1	19-03-2004
		US 2004083202 A1	29-04-2004
		US 2007288430 A1	13-12-2007
		WO 2004021123 A2	11-03-2004
EP 1462927 A2	29-09-2004	CN 1570842 A	26-01-2005
		CN 101034340 A	12-09-2007
		EP 1462927 A2	29-09-2004
		JP 4322031 B2	26-08-2009
		JP 2004295457 A	21-10-2004
		US 2004193760 A1	30-09-2004
		US 2005119994 A1	02-06-2005
		US 2005203964 A1	15-09-2005
		US 2008263277 A1	23-10-2008
		US 2011185123 A1	28-07-2011
US 6490666 B1	03-12-2002	US 6490666 B1	03-12-2002
		US 2003056069 A1	20-03-2003