РОССИЙСКАЯ ФЕДЕРАЦИЯ

ФЕДЕРАЛЬНАЯ СЛУЖБА
ПО ИНТЕЛЛЕКТУАЛЬНОЙ СОБСТВЕННОСТИ

(19) **RU** (11) **2 646 312**(13) **C1**

(51) МПК
G06F 12/08 (2006.01)
G06F 15/16 (2006.01)

(12) ОПИСАНИЕ ИЗОБРЕТЕНИЯ К ПАТЕНТУ

(52) СПК
G06F 3/0607 (2006.01); G06F 3/0664 (2006.01); G06F 3/067 (2006.01)

(21)(22) Заявка: **2016144518, 14.11.2016**

(24) Дата начала отсчета срока действия патента:
14.11.2016

Дата регистрации:
02.03.2018

Приоритет(ы):
(22) Дата подачи заявки: **14.11.2016**

(45) Опубликовано: **02.03.2018** Бюл. № **7**

Адрес для переписки:
129090, Москва, пр-кт Мира, 6, ООО "Патентно-правовая фирма "ЮС"

(72) Автор(ы):
Игнатьев Александр Валерьевич (RU),
Сунгуров Андрей Борисович (RU)

(73) Патентообладатель(и):
Общество с ограниченной ответственностью "ИБС Экспертиза" (RU)

(56) Список документов, цитированных в отчете о поиске: **WO 2008/069811 A1, 12.06.2008. US 8990618 B2, 24.03.2015. US 2006/0112219 A1, 25.05.2006. RU 2302034 C9, 27.09.2007. US 2015/0106420 A1, 16.04.2015. WO 2014/039922 A2, 13.03.2014. US 7380039 B2, 27.05.2008.**

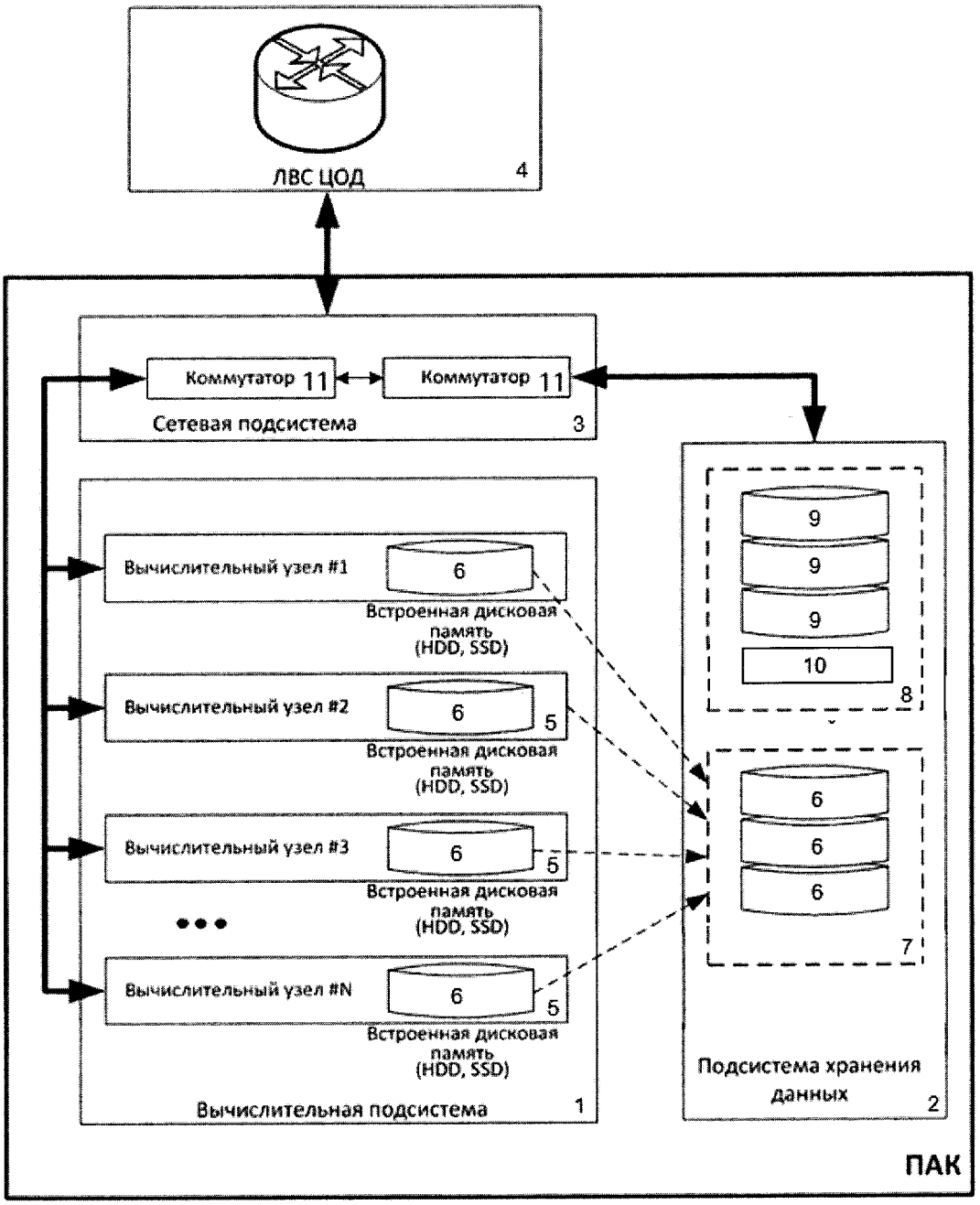
(54) Интегрированный программно-аппаратный комплекс

(57) Реферат:

Изобретение относится к области вычислительной техники, в частности к конвергентным (интегрированным) инфраструктурным программно-аппаратным комплексам. Техническим результатом является повышение общей производительности работы программно-аппаратного комплекса, а также его отказоустойчивости. Раскрыт интегрированный программно-аппаратный комплекс, содержащий: вычислительную подсистему, образованную по меньшей мере четырьмя вычислительными узлами, каждый из которых снабжен по меньшей мере одним процессором и встроенным диском для хранения данных, подсистему хранения данных, и сетевую подсистему, снабженную сетевыми коммутаторами для связи вычислительной подсистемы и подсистемы хранения данных между собой, а также с внешней сетью передачи данных, при этом подсистема хранения данных имеет независимые основной и дополнительный блоки хранения данных, основной блок хранения данных образован на базе встроенных дисков упомянутых вычислительных узлов с использованием установленных на вычислительные узлы программных средств, включающих средства для организации и управления хранением данных, средства для виртуализации вычислительных ресурсов и ресурсов хранения, а также средства мониторинга и управления, и дополнительный блок хранения данных включает по меньшей мере один отдельный, не входящий в состав вычислительных узлов дисковый массив, по меньшей мере один контроллерный узел, а также установленные на контроллерный узел программные средства для управления дисковым массивом, при этом указанные программные средства вычислительных узлов и контроллерного узла установлены с возможностью распределения данных приложений и системных сервисов по указанным блокам подсистемы хранения данных в зависимости от требований, характера и специфики работы приложений и сервисов. 2 з.п. ф-лы, 4 ил., 2 табл.

RU 2646312 C1

RU 2646312 C1



Фиг.1

RU 2646312 C1

RUSSIAN FEDERATION

FEDERAL SERVICE
FOR INTELLECTUAL PROPERTY

(19) **RU** (11) **2 646 312**(13) **C1**

(51) Int. Cl.
G06F 12/08 (2006.01)
G06F 15/16 (2006.01)

(12) ABSTRACT OF INVENTION

(52) CPC
G06F 3/0607 (2006.01); G06F 3/0664 (2006.01); G06F 3/067 (2006.01)

(21)(22) Application: **2016144518, 14.11.2016**

(24) Effective date for property rights:
14.11.2016

Registration date:
02.03.2018

Priority:
(22) Date of filing: **14.11.2016**

(45) Date of publication: **02.03.2018** Bull. № **7**

Mail address:
129090, Moskva, pr-kt Mira, 6, OOO "Patentno-pravovaya firma "YUS"

(72) Inventor(s):
Ignatev Aleksandr Valerevich (RU),
Sungurov Andrej Borisovich (RU)

(73) Proprietor(s):
Obshchestvo s ogranichennoj otvetstvennostyu "IBS Ekspertiza" (RU)

(54) INTEGRATED HARDWARE AND SOFTWARE SYSTEM

(57) Abstract:

FIELD: physics.

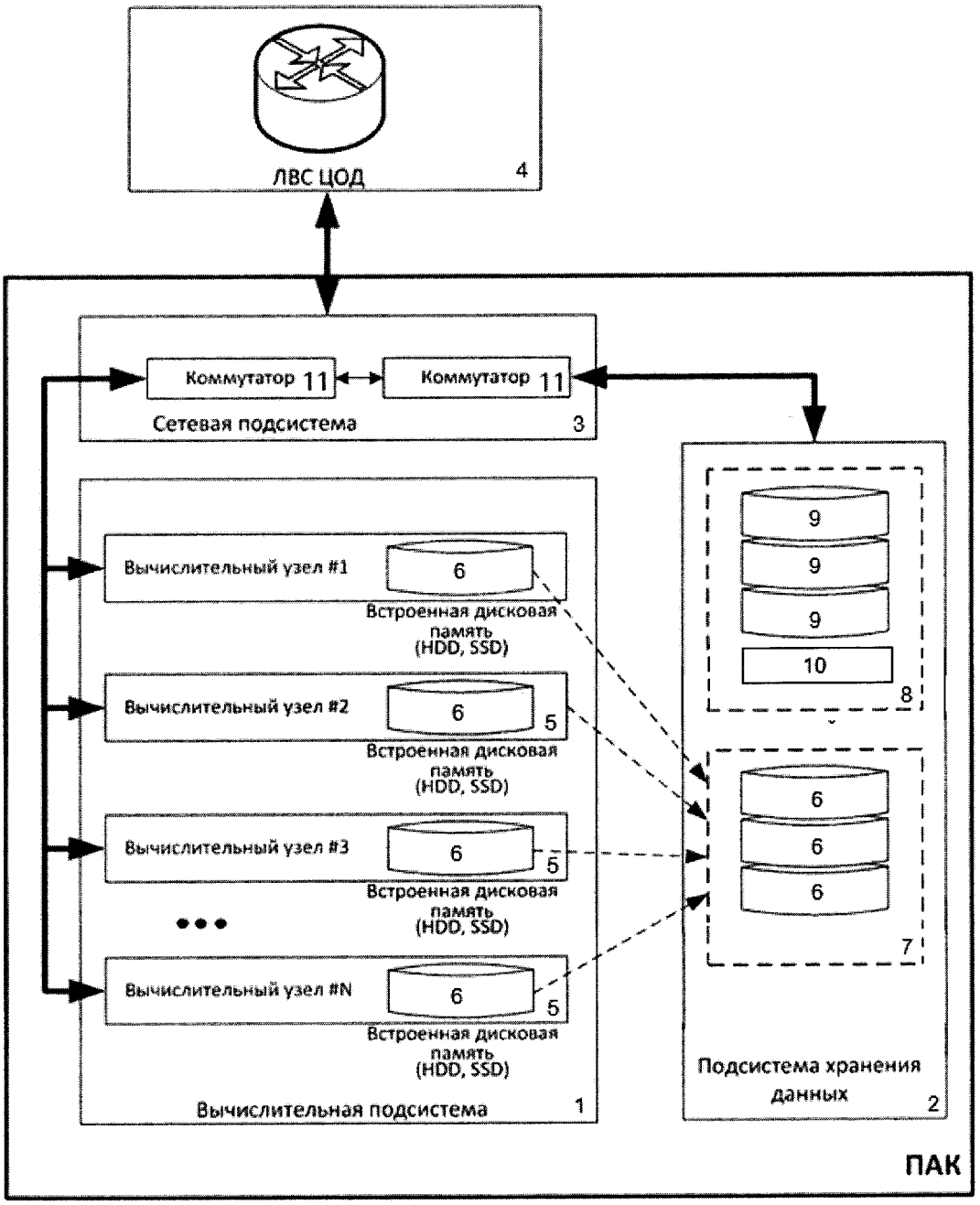
SUBSTANCE: integrated hardware and software system is disclosed, comprising: a computing subsystem formed by at least four computing nodes, each of which is provided with at least one processor and a built-in disk drive for the data storage, a data storage subsystem, and a network subsystem equipped with network switches for the communication of the computing subsystem and the storage subsystem among themselves and with the external data network. The data storage subsystem has independent primary and secondary data storage units, the primary data storage unit is formed on the basis of the mentioned built-in disks of the computing nodes using the software tools installed on the computing nodes, including the tools for the organization and control the data storage, the tools for the virtualization of the computing and storage resources, as well as monitoring and control tools, and the secondary data storing unit includes at least one separate disk array, being not a part of the computing node, at least one controller node, as well as software tools installed on the controller node to control the disk array. The software tools of the computing nodes and the controller node are installed with the ability to distribute the application and system service data to the specified units of the data storage subsystem, depending on the requirements, nature, and specificity of the applications and services.

EFFECT: increasing the overall performance of the hardware and software system, and its fault tolerance.

3 cl, 4 dwg, 2 tbl

RU 2646312 C1

RU 2646312 C1



Фиг.1

RU 2646312 C1

Область техники

Изобретение относится к области вычислительной техники, в частности к конвергентным (интегрированным) инфраструктурным программно-аппаратным комплексам (ПАК). Изобретение может быть использовано в инфраструктуре информационных технологий (ИТ-инфраструктуре) центров обработки данных (ЦОД), в том числе для развития ИТ-инфраструктуры ЦОД с использованием технологий виртуализации вычислительных ресурсов и ресурсов хранения данных.

Уровень техники

При реализации конвергентных (интегрированных) инфраструктурных комплексов, содержащих вычислительную подсистему и подсистему хранения данных, на практике используется один из двух походов: подсистема хранения данных организуется программными средствами вычислительных узлов с использованием механизмов виртуализации на базе внутренних дисковых накопителей вычислительных узлов, или подсистема хранения данных реализуется отдельными специализированными под задачи хранения компонентами.

Из уровня техники известны программно-аппаратные комплексы с подсистемой хранения данных на базе внутренних дисковых накопителей вычислительных узлов, в частности решение «HP ConvergedSystem 300» (опубликовано в сети Интернет http://www.mellanox.com/oem/hpe/rel_docs/HP%20PDW%20White%20Paper%204AA5-4088ENW.pdf, август 2014 г.), которое включает вычислительную подсистему в составе от трех до восьми вычислительных узлов, сетевую подсистему в составе двух сетевых коммутаторов и программно-определяемую подсистему хранения данных, реализованную на внутренней дисковой памяти вычислительных узлов специальным программным обеспечением.

Известно также решение «HC3 Systems» (https://www.scalecomputing.com/wp-content/uploads/2014/10/hc3-systems-product-specs.pdf, май 2016 года), описывающее ПАК, включающий вычислительную подсистему в составе от трех до восьми вычислительных узлов, сетевую подсистему в составе двух сетевых коммутаторов и программно-определяемую подсистему хранения данных, реализованную на внутренней дисковой памяти вычислительных узлов специальным программным обеспечением.

Из уровня техники известны ПАК с подсистемой хранения данных, реализованной отдельными специализированными под задачи хранения компонентами, в частности решение «Vblock System 100» (http://japan.vce.com/asset/documents/vblock-100-gen2-3-architecture-overview.pdf, ноябрь 2014 г.), являющееся наиболее близким аналогом настоящего изобретения и которое содержит вычислительную подсистему в составе от трех до восьми вычислительных узлов, сетевую подсистему в составе двух сетевых коммутаторов и специализированную под задачи хранения подсистему хранения данных, реализованную дополнительными программно-техническими средствами. Вычислительные узлы и оборудование подсистемы хранения данных подключены к сетевым коммутаторам, которые, в свою очередь, подключаются к локальной вычислительной сети (ЛВС) ЦОД. Подключения вычислительных узлов и оборудования подсистемы хранения данных к сетевым коммутаторам обеспечивают коммутацию и маршрутизацию трафика между вычислительной подсистемой и подсистемой хранения данных комплекса «Vblock System 100», а также между комплексом «Vblock System 100» и внешней сетью передачи данных.

Известен программно-аппаратный комплекс, снабженный двумя независимыми блоками хранения данных, реализованными на базе встроенных дисков вычислительных узлов (серверов приложений) (см. патент США US 8990618, 24.03.2015).

Основным недостатком описанных выше комплексов являются сравнительно невысокие производительность и отказоустойчивость.

Раскрытие изобретения

Задачей изобретения является устранение недостатков аналогов.

Техническим результатом изобретения является повышение общей производительности работы программно-аппаратного комплекса, а также его отказоустойчивости.

Указанный технический результат достигается в заявленном изобретении за счет того, что интегрированный программно-аппаратный комплекс содержит вычислительную подсистему, образованную по меньшей мере четырьмя вычислительными узлами, каждый из которых снабжен по меньшей мере одним процессором и встроенным диском для хранения данных, подсистему хранения данных и сетевую подсистему, снабженную сетевыми коммутаторами для связи вычислительной подсистемы и подсистемы хранения данных между собой, а также с внешней сетью передачи данных, при этом подсистема хранения данных имеет независимые основной и дополнительный блоки хранения данных, основной блок хранения данных образован на базе встроенных дисков упомянутых вычислительных узлов с использованием установленных на вычислительные узлы программных средств, включающих средства для организации и управления хранением данных, средства для виртуализации вычислительных ресурсов и ресурсов хранения, а также средства мониторинга и управления, а дополнительный блок хранения данных включает по меньшей мере один отдельный, не входящий в состав вычислительных узлов дисковый массив, по меньшей мере один контроллерный узел, а также установленные на контроллерный узел программные средства для управления дисковым массивом, причем указанные программные средства вычислительных узлов и контроллерного узла установлены с возможностью распределения данных приложений и системных сервисов по указанным блокам подсистемы хранения данных в зависимости от требований, характера и специфики работы приложений и сервисов.

Кроме того, согласно частным вариантам реализации изобретения:

- вычислительные узлы вычислительной подсистемы логически образуют серверы метаданных и связанные с ними серверы фрагментов, при этом серверы фрагментов выполнены с возможностью чтения и записи данных встроенных дисков вычислительных узлов, а серверы метаданных выполнены с возможностью хранения информации о серверах фрагментов и контроля количества копий каждого фрагмента данных;

- вычислительная подсистема содержит внутреннюю высокоскоростную сеть, обеспечивающую коммутацию вычислительных узлов и контроллерного узла с использованием первого набора Ethernet-адаптеров, а также связь указанных узлов с внешней сетью передачи данных, внешнюю клиентскую сеть, обеспечивающую коммутацию вычислительных узлов с внешней сетью передачи данных с использованием второго набора Ethernet-адаптеров, и сеть управления, обеспечивающую коммутацию вычислительных узлов и контроллерного узла через интерфейс IPMI.

Наличие двух независимых блоков хранения данных в заявленном комплексе позволяет распределить данные приложений (прикладных программ) и системных сервисов, использующих ПАК в качестве инфраструктурной основы для своей работы, по разным блокам хранения данных в зависимости от требований, характера и специфики работы приложений и сервисов, что обеспечивает повышение общей производительности работы приложений за счет оптимизации использования ресурсов комплекса. Так, основной блок хранения данных на базе встроенных дисков

вычислительных узлов используется в качестве файлового хранилища общего назначения и для хранения данных и файлов системы виртуализации, а дополнительный блок хранения данных, реализованный специализированными программно-техническими средствами, используется для данных ресурсоемких приложений, требующих высокой производительности системы хранения данных, например большой потоковой скорости передачи данных и/или выполнения большого количества операций ввода-вывода (IOPS). Такое разделение ресурсов хранения позволяет уменьшить нагрузку на вычислительные узлы подсистемы вычислительных ресурсов и тем самым повысить общую производительность ПАК.

Кроме того, используется комбинированный подход к виртуализации вычислительных ресурсов, включающий возможность одновременного использования виртуализации как гипервизорного, так и контейнерного типа, что позволяет оптимизировать вычислительные и дисковые ресурсы для достижения повышенной производительности и гибкости системы в целом. Таким образом, конвергентность данного ПАК как на уровне хранения данных, так и на уровне распределения вычислительных ресурсов и сетевого трафика позволяет относить данный программно-аппаратный комплекс к классу гиперконвергентных систем.

Краткое описание чертежей

Изобретение поясняется представленными фигурами, где:

на фиг. 1 показана принципиальная схема заявленного комплекса;

на фиг. 2 показана логическая схема организации хранения данных комплекса;

на фиг. 3 показана схема сетевых подключений комплекса;

на фиг. 4 показана схема организации потоков данных.

Осуществление изобретения

Заявленный программно-аппаратный комплекс (ПАК) (фиг. 1) содержит вычислительную подсистему (1), подсистему хранения данных (2) и сетевую подсистему (3) для коммутации подсистем (1) и (2) между собой и с внешней сетью (4) - локальной вычислительной сетью центров обработки данных (ЛВС ЦОД).

Вычислительная подсистема (1) содержит по меньшей мере четыре вычислительных узла (5) (сервера), в каждом из которых предусмотрен по меньшей мере один процессор (на фигурах не показан) и по меньшей мере один встроенный диск (6).

Подсистема хранения данных (2) имеет основной (7) и дополнительный (8) блоки хранения данных.

Основной блок хранения данных (7) организован программными средствами на базе встроенных дисков (6) вычислительных узлов (5). Дополнительный блок хранения данных (8) реализован специализированными под задачи хранения программно-техническими средствами, а именно одним или несколькими дисковыми массивами (9) (дисковыми полками), не входящими в состав вычислительных узлов (5), по меньшей мере одним контроллерным узлом (10), а также установленным на контроллерный узел (10) программным обеспечением для управления дисковым массивом (9).

Сетевая подсистема (3) содержит коммутаторы (11), которые связывают вычислительные узлы (5) подсистемы (1) и дополнительный блок хранения данных (8). Кроме того, коммутаторы (11) сетевой подсистемы (3) подключены к внешней сети (ЛВС ЦОД) (4), что обеспечивает коммутацию и маршрутизацию трафика между ПАК и внешней сетью передачи данных. На базе внутренней дисковой памяти - встроенных дисков (6) вычислительных узлов (5) вычислительной подсистемы (1) - программными средствами организуется логически единый функциональный блок хранения данных (7), представляемый как общий ресурс хранения в составе ПАК. При этом механизмы

обращения вычислительных узлов (5) ПАК (а также и внешних систем) к этому общему ресурсу хранения - блоку хранения данных (7) - унифицированы и не зависят от физической принадлежности дисковой памяти с требуемой информацией тому или иному вычислительному узлу. При этом обращение вычислительных узлов (5) к дополнительному блоку хранения данных (7) и взаимодействие вычислительных узлов (1) и подсистемы хранения данных (2) ПАК с внешней вычислительной сетью (ЛВС ЦОД (4)) осуществляется через коммутаторы (11) сетевой подсистемы (3).

Вычислительные узлы (5) (серверы) объединены в кластер. При этом на узлы (5) предустановлены программные средства, обеспечивающие исполнение виртуальных машин, предоставление необходимых для этого процессорных ресурсов и объемов оперативной памяти, а также взаимный обмен данными с использованием сетевой подсистемы (3).

Встроенные в вычислительные узлы (5) диски (6) могут представлять собой, например, HDD- или SSD-диски. Важной особенностью является то, что указанные встроенные диски (6) логически не являются частью вычислительной подсистемы (1), а относятся к подсистеме хранения данных (2), предоставляя общий пул ресурсов хранения для совместного использования всеми виртуальными машинами и приложениями.

Вычислительные узлы (5) логически образуют набор серверов, включающий серверы метаданных (12) и связанные с ними серверы фрагментов (13) (см. фиг. 2).

Серверы метаданных (MDS) (12) могут представлять собой виртуальные или физические машины, на которых хранится информация о серверах фрагментов (13), оперирующих данными, размещенными на встроенных дисках (6) основного блока хранения данных (7). Также серверы метаданных (12) контролируют количество копий каждого фрагмента данных для поддержания отказоустойчивости на уровне хранения. При этом для повышения надежности предусмотрена возможность создания нескольких серверов метаданных (12) на случай выхода из строя одного из них.

Серверы метаданных (12), а также дисковые массивы (9) дополнительного блока хранения (8) через сеть Ethernet (14) связаны с клиентами (15), представляющими собой все пользовательские приложения и виртуальные машины, которые обращаются к дисковым ресурсам распределенной подсистемы хранения данных (2).

Серверы фрагментов (13) являются агентами, входящими в состав каждой единицы оборудования, обладающей встроенными дисковыми ресурсами (6) в составе ПАК. Они отвечают за чтение и запись блоков данных подсистемы хранения данных (2).

Дополнительный блок хранения данных (8) представляет собой классическую систему хранения данных под управлением специализированного программного обеспечения. При обращении клиента к определенному блоку хранения данных этот запрос посредством сетевой подсистемы (3) поступает либо на сервер метаданных (12), в котором хранится информация о расположении всех блоков данных подсистемы хранения (7), после чего запрос переадресуется на соответствующий узел кластера, либо на внешнюю систему хранения данных (8).

Подсистема хранения данных (2) в составе двух блоков хранения (7) и (8) предоставляет единый унифицированный доступ к ресурсам хранения рассматриваемого ПАК по протоколам NFS/iSCSI. Обмен данными между ресурсами хранения, расположенными на разных физических хостах (узлах), производится посредством высокоскоростной сети передачи данных сетевой подсистемы (3), позволяя таким образом добиться низкой задержки и высокого показателя IOPS (операций ввода/вывода в секунду).

Сетевая подсистема (3) (фиг. 3) предпочтительно включает четыре коммутатора:

два коммутатора 1 Гбит/сек (11a) для связи с внешней сетью (4) и образования сети управления ПАК (18) и два высокоскоростных коммутатора 56 Гбит/сек (11b) для внутренней коммутации.

Логически сетевая подсистема (3) делится на три сети: внутреннюю высокоскоростную сеть (16), внешнюю клиентскую сеть (17) и сеть управления (18).

Внутренняя высокоскоростная сеть (16) (на фиг. 3 показана сплошной линией) обеспечивает коммутацию вычислительных узлов (5) подсистемы (1) и контроллерных узлов (10) дополнительного блока хранения данных (8) подсистемы (2), а также, опционально, их связь с внешней сетью (4) с использованием первого набора Ethernet-адаптеров (19), установленных в узлах (5) и (10).

Внешняя клиентская сеть (17) (на фиг. 3 показана точками) также обеспечивает соединение вычислительных узлов (5) с внешней сетью (4) при использовании второго набора Ethernet-адаптеров (20), установленных в узлах (5).

Сеть управления (18) (на фиг. 3 показана мелким пунктиром) обеспечивает коммутацию вычислительных (5) и контроллерных (10) узлов через интерфейс IPMI (21) и образует таким образом единую сеть мониторинга и управления ПАК. При этом контроллерные узлы (10) связаны с дисковыми массивами (9) посредством SAS-соединений (22) с использованием SAS адаптеров (23).

Использование высокоскоростной внутренней сети (16) продиктовано необходимостью выдерживать большие объемы трафика между узлами хранения данных, что позволяет добиться высокой производительности подсистемы хранения данных (2) в целом. Кроме того, дублирование коммутаторов (11) и сетевых адаптеров (19, 20) позволяет обеспечить отказоустойчивость на уровне сетевой инфраструктуры. Также отказоустойчивость ПАК достигается за счет дублирования всех сетевых соединений между вычислительными узлами (5) и коммутаторами (11), объединения вычислительных узлов в кластер, а также дублирования блоков данных на уровне ПО управления дисковыми ресурсами (по схеме, аналогичной RAID-1).

Обмен данными между хостами и виртуальными машинами осуществляется посредством внутренней высокоскоростной сети ПАК, с использованием протоколов TCP/IP и FCoE.

Алгоритм работы заявленного комплекса заключается в следующем (см. фиг. 4).

Приложением-клиентом инициируется запрос на чтение или запись данных, обращенный к логическому диску, предоставленному клиенту основным (7) или дополнительным (8) блоком данных (шаг «a»).

В зависимости от того, какой из блоков данных предоставил логический диск приложению (шаг «b»), запрос перенаправляется программному обеспечению (ПО) основного блока хранения данных (шаг «c») или ПО дополнительного блока (8) (шаг «d») хранения.

В случае использования дополнительного блока хранения (8) дальнейшая обработка осуществляется управляющим программным обеспечением (ПО) (например, RAIDIX), установленным на контроллерах (10) блока (8), после чего результат (запрошенные данные или отчет о записи) по пути «f»-«a» передается клиенту-инициатору.

В случае использования основного программно-реализованного блока хранения (7) запрос поступает управляющему ПО, установленному на вычислительных узлах (хостах) (например «P-Хранилище»), которое производит проверку доступности сервера метаданных (12) (шаг «e»). При подтверждении доступности запрос поступает на сервер метаданных (12), хранящий информацию о распределении данных по внутренним дискам вычислительных узлов. При отказе запрос принимает резервный сервер

метаданных (шаг «g»). Для этого вычислительные узлы и установленное на них ПО объединены в отказоустойчивый кластер не менее чем из четырех узлов.

После получения запроса сервером метаданных (шаг «h») происходит его выполнение на виртуальных серверах фрагментов (13) (шаг «i»), входящих в состав каждого вычислительного узла (5). При этом если данные подверглись изменению, то отчет об этом и информация о размещении записанных данных передается обратно на сервер метаданных.

Завершающим действием является передача запрошенной информации или отчета об успешной записи данных клиенту-инициатору (путь «i»-«a»).

Далее описаны варианты промышленной реализации заявленного комплекса, приведенные в качестве примеров, но не ограничивающие объем заявленного изобретения.

Программно-аппаратный комплекс состоит из промышленных компонентов со следующими характеристиками.

Вычислительные узлы (5) вычислительной подсистемы (1)

Используются вычислительные узлы DEPO Storm российского производителя, реализуются на основе следующих наборов опций, показанных в таблице 1.

Таблица 1. Технические характеристики вычислительных узлов ПАК.

Модельный ряд	**СКАЛА-Р 300**	**СКАЛА-Р 500**	**СКАЛА-Р 700**
Тип вычислительного узла	DEPO Storm 3Sk2U	DEPO Storm 5Sk1U	DEPO Storm 7Sk2U или DEPO Storm 7Sk4U
Тип процессора	Intel Xeon E5-2600	Intel Xeon E5-2600	Intel Xeon E7-4800 или Intel Xeon E7-8800
Количество ядер	Количество ядер определяется типом выбранного процессора, возможен любой тип процессора из соответствующей линейки.		
Объем оперативной памяти	Стандартно используются модули памяти по 32 ГБ, но могут использоваться и другие модули памяти – 8, 16 и 64 ГБ.		
Состав дисков используемых в системе хранения данных	В составе ПАК могут быть использованы следующие диски: • SSD 100 ГБ, 200ГБ, 400 ГБ • SAS 15K 600 ГБ; • SAS 10K 600 ГБ, 900 ГБ; • NL-SAS 7.2K 1000 ГБ, 2000 ГБ		
Оптимизация дисковой подсистемы	В рамках данной опции в конфигурацию каждого вычислительного узла добавляются SSD-диски, обеспечивающие кэширование запросов к дисковой подсистеме, что повышает производительность блока хранения данных, реализованного на базе встроенных дисков вычислительных узлов.		

Опция AllFlash	В рамках данной опции существует возможность замены всех жестких дисков твердотельными накопителями SSD, что позволяет значительно повысить производительность дисковой подсистемы, кратно увеличить число операций ввода/вывода (IOPS) и уменьшить время задержки системы. Также использование данной опции избавляет от необходимости использования кэша и журналов файловой системы распределенного хранилища, и позволяет работать с данными напрямую на носителях SSD.		
Базовый состав вычислительного узла	• два процессора серии Intel Xeon E5-2600; • минимум 64 ГБ оперативной памяти; • два порта 1000Base-T Ethernet; • два порта 56G Ethernet QSFP; • один порт IPMI RJ-45; • один диск SATA для размещения гипервизора; • один SSD диск 100 Гб для размещения служебной информации.		• четыре процессора серии Intel Xeon E7; • минимум 128 ГБ оперативной памяти; • два порта 1000Base-T Ethernet; • два порта 56G Ethernet QSFP; • один порт IPMI RJ-45; • один диск SATA для размещения гипервизора; • один SSD диск 200 Гб для размещения служебной информации.
Внешний дисковый массив	Опционально 5SKST	Опционально 5SKST	DEPO Storm 7SKST

На вычислительные узлы устанавливается платформенное программное обеспечение (ПО) «Росплатформа», включающее:

- ПО «Р-Виртуализация» - система виртуализации ресурсов, обеспечивающая возможность одновременного использования гипервизорной и контейнерной виртуализации;

- ПО «Р-Управление» - система оркестрации и управления виртуализацией;

- ПО «Р-Хранилище» - реализует программно-определяемую систему хранения данных.

Перечисленное программное обеспечение выполняет роль платформы виртуализации для ПАК, реализует подсистему хранения данных на базе дисковых накопителей вычислительных узлов, а также обеспечивает организацию вычислительных ресурсов ПАК в кластер, объединяющий от четырех и более вычислительных узлов. При этом использование четырех узлов является минимальной конфигурацией, при которой может быть обеспечена полноценная отказоустойчивость кластера. В случае необходимости увеличения вычислительной мощности или емкости для хранения данных можно добавлять по одному дополнительному вычислительному узлу и интегрировать их в единый комплекс с уже установленным оборудованием.

Основные функции и возможности установленного на вычислительные узлы программного обеспечения представлены ниже.

ПО «Р-Виртуализация»

Программное обеспечение «Р-Виртуализация» представляет собой классический гипервизор, инсталлируемый непосредственно на аппаратную платформу и не требующий дополнительной операционной системы для своего функционирования. Основные функциональные возможности ПО «Р-Виртуализация»:

- максимальное поддерживаемое количество виртуальных процессоров в виртуальных машинах (ВМ) Windows или Linux (максимальное количество для разных гостевых операционных систем (ОС) может сильно отличаться, что связано с ограничениями гостевой операционной системы) - 32 виртуальных процессора на ВМ;

- максимальное поддерживаемое количество памяти в виртуальных машинах (максимальное количество для разных гостевых ОС может сильно отличаться, что связано с ограничениями гостевой операционной системы) - 128 ГБ на ВМ;

- поддержка серийных портов для ВМ (могут быть привязаны к порту на физическом хосте, к именованным каналам или сетевым и портовым концентраторам) - максимально 16;

- поддержка USB-устройств в виртуальных машинах;

- возможность добавлять устройства к виртуальной машине в процессе ее работы: процессоры, память, диски, сетевые интерфейсы;

- возможность предоставлять виртуальным машинам больше памяти, чем доступно физически - осуществляется динамическим перераспределением памяти между виртуальными машинами и освобождением неиспользуемой памяти.

ПО «Р-Управление»

Программное обеспечение «Р-Управление» представляет собой гибкий инструмент управления группами физических вычислительных узлов и находящимися на них виртуальными средами. Программное обеспечение «Р-Управление» реализует следующие основные функции:

- осуществляет первичную регистрацию физических ресурсов;

- создает логическую структуру физических серверов и находящихся на них виртуальных сред;

- обеспечивает миграцию виртуальных сред между физическими и виртуальными вычислительными узлами (вычислительными машинами);

- создает и управляет шаблонами ОС и приложений;

- создает и управляет резервными копиями виртуальных сред;

- клонирует виртуальные машины;

- управляет ресурсами виртуальных сред;

- контролирует операции в виртуальных средах;

- выполняет групповые операции с виртуальными машинами;

- предоставляет средства настройки дискретного и ролевого доступа к функциям и ресурсам виртуальной среды;

- предоставляет средства настройки интерфейса «Р-Управление» и изменения личных настроек администраторов;

- создает резервные копии виртуальных машин;

- автоматизирует элементарные процессы формирования и отслеживания заявок на новые виртуальные машины и проблемы, возникающие в процессе эксплуатации виртуальной среды.

ПО «Р-Хранилище»

Программное обеспечение «Р-Хранилище» реализует основной блок хранения данных (7) на базе встроенных дисков вычислительных узлов ПАК. В каждом вычислительном узле ПАК может быть установлено до девяти (или двадцати шести для серий 300 и 700) встроенных дисков, при этом два из них используются для операционной системы и хранения метаданных. В случае выбора опции «оптимизация производительности дисковой системы» количество свободных слотов для установки дисков уменьшается до шести (или двадцати трех для серий 300 и 700) за счет добавления дополнительных SSD-дисков под функции кэша второго уровня для системы хранения данных. Основные функции (возможности) программного обеспечения «Р-Хранилище»:

- использование кэшей первого и второго уровня;

- автоматический перенос данных между носителями с разной скоростью доступа в зависимости от востребованности данных (tiering);

- обеспечение доступа к данным «Р-Хранилище» через NFS и iSCSI;

- обеспечение отказоустойчивости и увеличение производительности системы

хранения данных путем установления несколько путей от инициатора к источнику (multipath I/O);

- обеспечение защиты и сохранности данных за счет создания RAID-массивов с использованием технологий зеркалирования;

- формирование «мгновенных снимков» файловой системы (snapshot).

Дополнительный блок хранения данных (8)

Реализуется отдельным, внешним по отношению к вычислительным узлам дисковым массивом под управлением контроллера дискового массива, в качестве которого используется платформа DEPO Storm 5SKST либо 7SKST с установленным программным обеспечением RAIDIX. Программное обеспечение RAIDIX - специализированный под задачи хранения данных продукт, позволяющий создавать высокопроизводительную, надежную, отказоустойчивую систему хранения данных на стандартных аппаратных компонентах. Управляющее программное обеспечение RAIDIX обеспечивает управление массивами дисков и решает дополнительные задачи, такие как тонкая оптимизация, детальный мониторинг, запуск дополнительных приложений непосредственно на системе хранения. Высокая производительность системы хранения данных на базе продукта RAIDIX для ключевых и требовательных к ресурсам приложений обеспечивается:

- высокой скоростью обмена данными - до 8,0 ГБ/сек;

- механизмами приоритезации полосы пропускания для инициаторов запросов, обеспечивающими гарантированное время доступа к данным для ключевых приложений.

Внешний дисковый массив может использоваться как SAN- или NAS-устройство, при этом доступен широкий перечень протоколов взаимодействия: SMB/CIFS, NFS, FTP, AFP, iSCSI, FC.

Оборудование дополнительного блока хранения обладает характеристиками, приведенными в таблице 2.

Таблица 2. Технические характеристики компонентов дополнительного блока хранения ПАК.

Модельный ряд	**СКАЛА-Р 500**	**СКАЛА-Р 700**
Тип контроллерного узла	5SKST	7SKST
Кэш	32 ГБ на контроллер с возможностью увеличения до 192 ГБ	128 ГБ на контроллер с возможностью увеличения до 512 ГБ
Высокоскоростной интерфейс для синхронизаций кэша между контроллерами	Есть	Есть
Порты SAS	2 порта SAS 6 Гбит/сек	от 2 до 6 портов SAS 6 Гбит/сек
Количество подключаемых дисковых полок	До 4 шт. DEPO Storage 5SK24	До 12 шт. DEPO Storage 7SK24
Порты Ethernet	2 порта по 1 Гбит/с Ethernet	2 порта по 1 Гбит/с Ethernet
Порты высокоскоростного Ethernet	2 порта по 56 Гбит/с Ethernet	2 порта по 56 Гбит/с Ethernet
Высота в стойке	3U	3U
Количество мест для установки дисков	16 слотов 3,5" SATA/SAS	16 слотов 3,5" SATA/SAS
Количество блоков питания	2 блока питания 950W с резервированием	2 блока питания 750W с резервированием
Дисковая полка	**DEPO Storage 5SK24**	**DEPO Storage 7SK24**

Количество мест для установки дисков	24 слота 2.5" SATA/SAS	24 слота 2.5" SATA/SAS
Высота в стойке	2U	2U
Количество блоков питания	2 шт.	2 шт.
Количество адаптеров SAS	2 шт.	2 шт.
Скорость интерфейса SAS	6 Гбит/сек	6 Гбит/сек

Коммутаторы сетевой подсистемы (11)

Используются высокопроизводительные коммутаторы, например коммутаторы Mellanox, имеющие 12 или 36 портов (зависит от количества узлов ПАК). Каждый из портов коммутатора может работать на скоростях 10, 40 или 56 Гбит/сек, при этом скорость 56 Гбит/сек достигается только при совместной работе коммутатора и сетевой карты производства Mellanox. Отличительной особенностью коммутаторов Mellanox является чрезвычайно низкая латентность (задержка при передаче пакета) по сравнению с коммутаторами других производителей, что положительно сказывается на производительности распределенного дискового массива «Р-Хранилище».

Программно-аппаратный комплекс работает следующим образом.

Использование того или иного блока хранения данных ПАК при работе того или иного приложения определяется администратором при настройке и инициализации приложения. Основными параметрами, которые при этом нужно учитывать, являются требования приложений к объемам хранения данных и скорости (производительности) ввода/вывода, количество пользователей и др. При этом учитываются также данные мониторинга загрузки вычислительных ресурсов и ресурсов хранения данных ПАК. Кроме того, необходимо также учитывать, что активная работа приложений с большими массивами данных в блоке хранения данных на базе встроенных дисков вычислительных узлов (3) будет существенно загружать коммутаторы сетевой подсистемы (5), что может привести, в том числе, к деградации производительности сети передачи данных и тем самым снизить общую производительность ПАК. Это связано, в том числе, и с работой механизмов обеспечения резервирования данных в распределенном дисковом массиве за счет их хранения на физически различных вычислительных узлах ПАК. Использование приложениями ресурсов хранения данных ПАК в общем случае рекомендуется настраивать по следующей схеме.

1. Основной блок хранения данных на базе встроенных дисков вычислительных узлов (7).

Ресурсы данного блока назначаются приложениям и сервисам, не требующим высоких скоростей в режиме последовательного чтения/записи данных. К таким приложениям и сервисам могут быть отнесены, к примеру, общесистемные сервисы (DNS, DHCP), файловый сервис, внутренняя электронная почта, бухгалтерские программы, СУБД различного назначения.

2. Дополнительный блок хранения данных (8).

Ресурсы данного блока назначаются приложениям, работающим с большими объемами данных и требующим высокую производительность в режиме последовательного чтения/записи данных. Примеры такого рода приложений: хранение и обработка медиа-контента (видеозаписи, изображения с высоким разрешением, организация трансляций и вещания), специализированные приложения для больниц и медицинских центров (хранение и обработка историй болезней, результатов анализов), биллинговые системы, системы поддержки принятия решений и др.

Работа приложений с подсистемой хранения данных (2) осуществляется следующим

образом.

Приложение инициирует запрос на чтение или запись данных, который поступает в операционную систему (в данном случае в гипервизор Р-виртуализация), которая соотносит логическое наименование места нахождения данных с физическими ресурсами, в частности с контроллером системы хранения данных, который собственно и отвечает за выполнение физических операций записи/чтения данных. В зависимости от назначения конкретному приложению того или иного блока хранения данных запрос на чтение/запись будет перенаправлен либо в адрес «Р-Хранилище», либо в адрес внешнего контроллера, реализуемого средствами ПО RAIDIX. Далее операции записи/чтения данных на соответствующие дисковые накопители выполняются под управлением контроллеров системы хранения данных. При этом выполнение операций записи/чтения в блоке хранения данных на базе встроенных дисков вычислительных узлов осуществляется параллельно с дублированием на дисковые накопители разных вычислительных узлов, а данные в процессе дублирования передаются на другие вычислительные узлы через сетевые коммутаторы сетевой подсистемы, и частично при этом используются процессорные мощности вычислительных узлов. При большом количестве операций ввода/вывода и при больших потоках данных нагрузка на вычислительные узлы может быть существенной, что отрицательно скажется на общей производительности ПАК с точки зрения работы приложений (конечных пользователей).

Выполнение операций записи/чтения в дополнительном блоке хранения данных (8) также производится параллельно с дублированием на несколько дисковых накопителей внешнего дискового массива, но управление и контроль этих операций осуществляется самостоятельно контроллером (10) внешнего дискового массива блока (8) без задействования ресурсов вычислительной подсистемы. При этом обеспечивается высокая скорость обмена данными (как указано выше - до 8 ГБ/сек, что существенно выше скорости обмена данными для основного блока хранения данных (7)). Данные через сетевые коммутаторы сетевой подсистемы (3) передаются при этом один раз, то есть при работе приложений с дополнительным блоком хранения данных (8) обеспечивается высокая скорость обмена данными на уровне дисковых накопителей, и ресурсы вычислительной подсистемы, а также сетевой подсистемы дополнительно не задействуются.

Таким образом, операции записи/чтения данных для разных приложений будут распределяться по разным блокам хранения данных программно-аппаратного комплекса в зависимости от характера приложений и их требований к ресурсам хранения, а также с учетом загрузки вычислительных ресурсов и ресурсов хранения данных ПАК. Тем самым может быть оптимизировано использование общих ресурсов хранения данных программно-аппаратного комплекса, а его общая производительность с точки зрения работы приложений повышена.

(57) Формула изобретения

1. Интегрированный программно-аппаратный комплекс, содержащий:

вычислительную подсистему (1), образованную по меньшей мере четырьмя вычислительными узлами (5), каждый из которых снабжен по меньшей мере одним процессором и встроенным диском (6) для хранения данных,

подсистему хранения данных (2), и

сетевую подсистему (3), снабженную сетевыми коммутаторами (11) для связи вычислительной подсистемы (1) и подсистемы хранения данных (2) между собой, а также с внешней сетью передачи данных (4),

отличающийся тем, что

подсистема хранения данных (2) имеет независимые основной (7) и дополнительный (8) блоки хранения данных,

основной блок хранения данных (7) образован на базе встроенных дисков (6) упомянутых вычислительных узлов (5) с использованием установленных на вычислительные узлы (5) программных средств, включающих средства для организации и управления хранением данных, средства для виртуализации вычислительных ресурсов и ресурсов хранения, а также средства мониторинга и управления, и

дополнительный блок хранения данных (8) включает по меньшей мере один отдельный, не входящий в состав вычислительных узлов дисковый массив (9), по меньшей мере один контроллерный узел (10), а также установленные на контроллерный узел (10) программные средства для управления дисковым массивом (9),

при этом указанные программные средства вычислительных узлов (5) и контроллерного узла (10) установлены с возможностью распределения данных приложений и системных сервисов по указанным блокам подсистемы хранения данных в зависимости от требований, характера и специфики работы приложений и сервисов.

2. Комплекс по п. 1, отличающийся тем, что вычислительные узлы (5) вычислительной подсистемы (1) логически образуют серверы метаданных (12) и связанные с ними серверы фрагментов (13), при этом серверы фрагментов (13) выполнены с возможностью чтения и записи данных встроенных дисков (6) вычислительных узлов (5), а серверы метаданных (12) выполнены с возможностью хранения информации о серверах фрагментов (13) и контроля количества копий каждого фрагмента данных.

3. Комплекс по п. 1, отличающийся тем, что вычислительная подсистема (3) содержит:

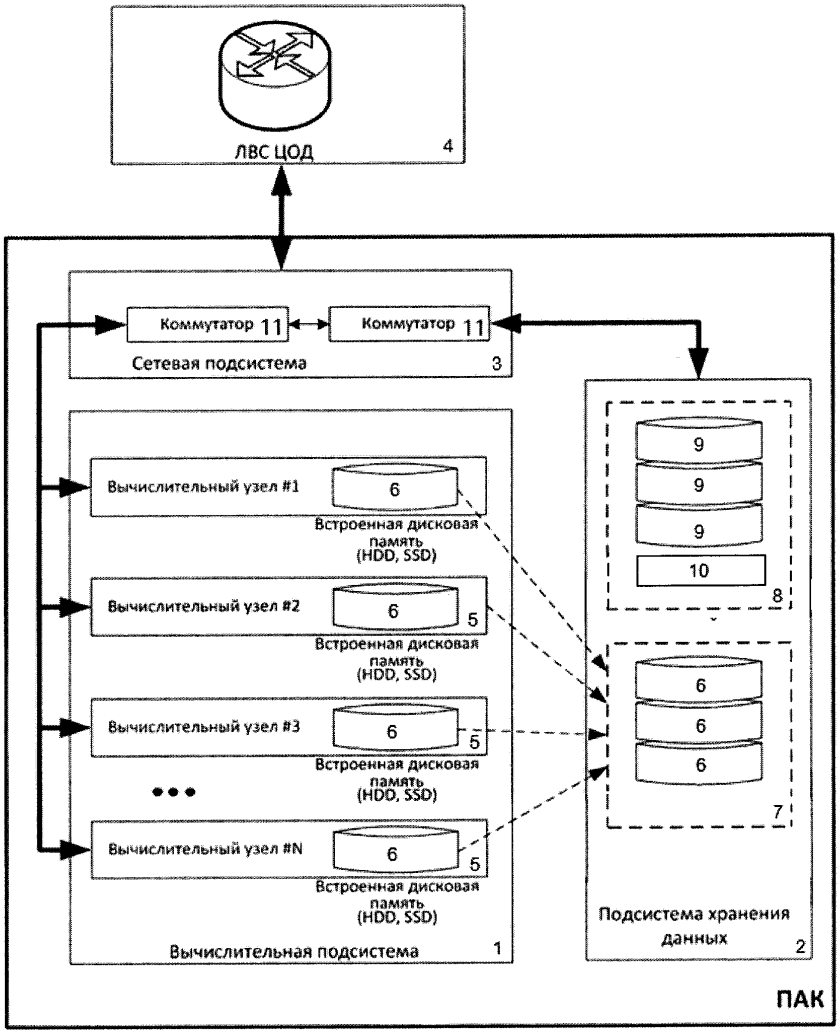
внутреннюю высокоскоростную сеть (16), обеспечивающую коммутацию вычислительных узлов (5) и контроллерного узла (10) с использованием первого набора Ethernet-адаптеров (19), а также связь указанных узлов с внешней сетью передачи данных (4),

внешнюю клиентскую сеть (17), обеспечивающую коммутацию вычислительных узлов (5) с внешней сетью передачи данных (4) с использованием второго набора Ethernet-адаптеров (20), и

сеть управления (18), обеспечивающую коммутацию вычислительных узлов (5) и контроллерного узла (10) через интерфейс IPMI (21).

1

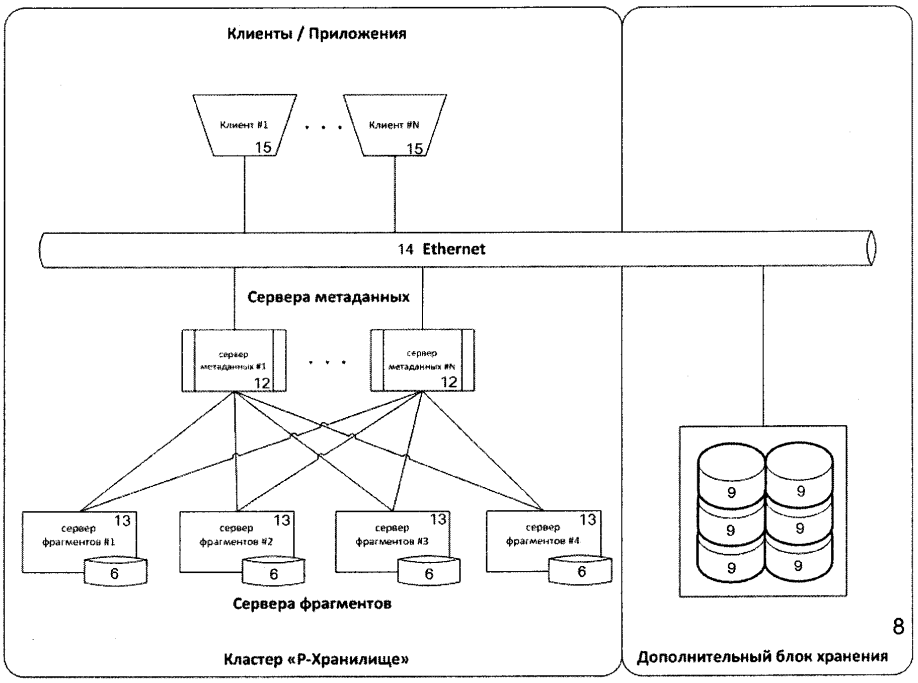
Интегрированный программно-аппаратный комплекс



Фиг.1

2

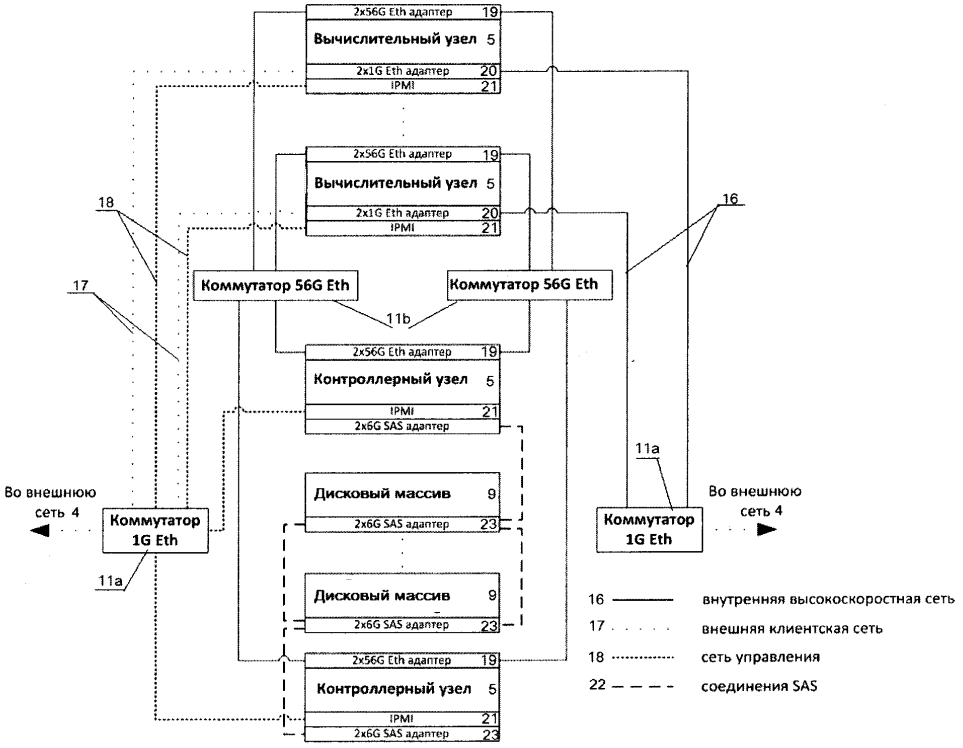
Интегрированный программно-аппаратный комплекс



Фиг.2

3

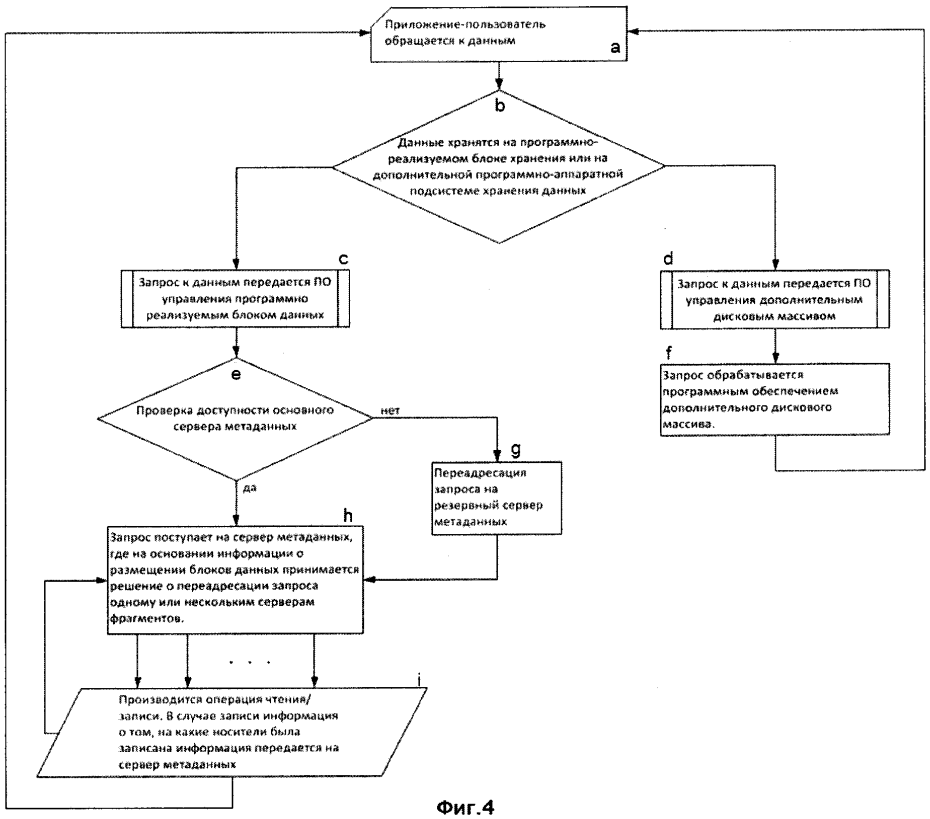
Интегрированный программно-аппаратный комплекс



Фиг.3

4

Интегрированный программно-аппаратный комплекс



Фиг.4