DOI: 10.15514/ISPRAS-2022-34(4)-9

Реализация функции долговременного хранения научных данных большого объема в вычислительном центре

Д.В. Иванков, ORCID: 0000-0003-4254-0104 <d.v.ivankov@yandex.ru>

Всероссийский НИИ технической физики имени академика Е.И. Забабахина, 456770, Россия, г. Снежинск, Челябинская область, ул. Васильева, 13

Аннотация. Длительное и целостное хранение объемных научных данных является одной из важных задач, стоящих перед многими вычислительными центрами. В целях снижения стоимости хранения информации, в некоторых решениях используется технология магнитно-ленточной памяти, а также специализированное программное обеспечение для управления носителями и данными. Ввиду инфраструктурной специфики и особенностей сложившихся техпроцессов генерации и обработки данных в научных лабораториях такие программно-аппаратные комплексы создаются и поддерживаются преимущественно собственными силами этих организаций. Разработка такой системы становится еще более востребованной в условиях стремления к обретению технологического суверенитета. В работе рассматриваются вопросы организации долговременного хранения цифровых научных данных в вычислительном центре ФГУП РФЯЦ-ВНИИТФ, полученных в ходе проведения расчетов задач математического моделирования. Приводится описание архитектуры и функционального состава разработанной архивной системы хранения данных. Описывается используемая модель данных и форматы группировки и записи. Рассматриваются предпринятые меры по обеспечению целостности архивных объектов, методы управления архивными носителями и вопросы технической организации архивного фонда. Приводится схема расчета аппаратной конфигурации типовой площадки архивной системы хранения данных, достаточной для обслуживания существующих потоков архивирования данных в вычислительном центре.

Ключевые слова: архивная система хранения; долговременное хранение; магнитные ленты; ленточные библиотеки; целостность информации

Для цитирования: Иванков Д.В. Реализация функции долговременного хранения научных данных большого объема в вычислительном центре. Труды ИСП РАН, том 34, вып. 4, 2022 г., стр. 117-134. DOI: 10.15514/ISPRAS-2022-34(4)-9

Large-scale scientific data and long-term data storage function in a computing center

D.V. Ivankov, ORCID: 0000-0003-4254-0104 <d.v.ivankov@yandex.ru>

E.I. Zababakhin All-Russian Scientific Research Institute of Technical Physics, 13, Vasilieva street, Chelyabinsk region, Snezhinsk, 456770, Russia

Abstract. Long-term data storing is an important task for many modern scientific laboratories and datacenters. In order to reduce cost of digital information ownership, some solutions use magnetic tape technology and special software to control medium and data. Considering the on-site infrastructure specifics and well-established workflows of data processing, these organizations build and support such systems mainly by their own efforts, what becomes an important task in seeking to acquire the technological sovereignty. This paper describes long-term data storage issues in the computing center of the Zababakhin All-Russia Research Institute of Technical Physics where mathematical modeling computations generate vast amount of scientific data. The

architecture and functional composition of the developed Archive Data Storage System are given as well as its internal data model, the chunk grouping rules, and the low-level tape format used. The measures taken to ensure an archived data consistency, methods of storage media management and issues of archival fund maintenance, are also considered. The calculation scheme of a typical archive system site's hardware configuration, sufficient to process archiving data flows existing in datacenter, is given.

Keywords: digital archive, long-term data storage, magnetic tapes, tape libraries, data consistency.

For citation: Ivankov D.V. Large-scale scientific data and long-term data storage function in a computing center. Trudy ISP RAN/Proc. ISP RAS, vol. 34, issue 4, 2022. pp. 117-134 (in Russian). DOI: 10.15514/ISPRAS-2022-34(4)-9

1. Введение

При проведении расчетов задач математического моделирования на высокопроизводительных вычислительных системах (далее по тексту ВВС) вычислительные процессы задействуют функции ввода-вывода главным образом для сохранения и считывания состояния своей оперативной памяти. Со стороны системы хранения (далее по тексту СХ) ВВС эта роль возложена на оперативные файловые системы (далее по тексту ФС), обладающие (по возможности) максимальной производительностью ввода-вывода и, вследствие их высокой стоимости, ограниченными объемами ресурсов хранения. Для непрерывного потока расчетов на ВВС необходимо поддержание достаточного объема свободных ресурсов оперативных ФС, что обычно достигается путем своевременного удаления «ненужных» данных и регулярного вытеснения устаревших, но не потерявших актуальности данных на специально выделенный сегмент СХ, предназначенный для долговременного хранения информации.

Многолетняя эксплуатация ВВС в ФГУП РФЯЦ-ВНИИТФ им. академ. Е.И. Забабахина (далее по тексту РФЯЦ-ВНИИТФ) показала важность управления расчетными данными в течение их жизненного цикла и необходимость организации системы долговременного хранения данных, способствующую объединению и эффективной эксплуатации информационно-вычислительных ресурсов института. Для решения этой задачи в РФЯЦ-ВНИИТФ была создана «архивная система хранения данных» (далее по тексту АСХД), опирающаяся на технологию магнитно-ленточной памяти. Целями ее создания были обозначены: создание единого ресурса долговременного хранения расчетных данных, сгенерированных в различных ВВС; обеспечение целостности и сохранности информации в течение длительного времени; а также снижение зависимости от внешних факторов и сторонних компонент, способных повлиять на возможность восстановления информации. Для управления этой системой было разработано одноименное программное обеспечение (далее по тексту ПО), которое затем вошло в состав операционной системы «СПО Супер-ЭВМ» [1].

Среди наиболее близких к АСХД по своему назначению программных систем следует выделить коммерческий продукт IBM HPSS [2], ставший стандартом де-факто во многих научных лабораториях и университетах благодаря своей зрелости и глобальной поддержке производителя. Некоторые центры обработки данных разработали и эксплуатируют собственное ПО для управления системами архивного хранения данных, например, Enstore [3] в Fermilab (США) и CTA [4] в европейском CERN.

2. АСХД

АСХД представляет собой аппаратно-программный комплекс, выполняющий загрузку, хранение, поиск и восстановление цифровой пользовательской информации. Высокая степень сохранности обрабатываемых данных обеспечивается благодаря широкому применению средств контроля целостности, встроенной репликации, использованию технологии магнитных лент, изоляции заполненных архивных носителей и самих архивных

объектов. Способность хранения значительного объема информации в течение длительного времени обеспечивается как за счет применения автоматизированных ленточных библиотек различных технологических поколений, так и благодаря разработанному регламенту обмена архивными носителями между библиотеками и стеллажным хранилищем.

В качестве источников информации для АСХД выступают стандартные файловые ресурсы хранения данных.

Программная часть системы реализована с применением клиент-серверного подхода и реляционной системы управления базами данных. АСХД обладает пользовательским и административным web-интерфейсом и интерфейсом командной строки.

3. Структура АСХД

Структура АСХД включает в себя следующий набор взаимодействующих компонентов (см. рис. 1):

- метасервер;
- сервер управления базой данных (СУБД);
- медиасервер;
- транспортный агент;
- консольный клиент (CLI);
- web-клиент;
- web-сервер;
- консоль администратора;
- автоматизированное рабочее место оператора хранилища (АРМ), оснащенное мобильным терминалом.

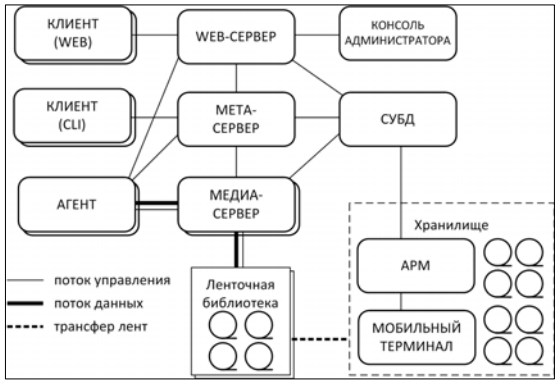

Рис. 1. Структурная схема АСХД
Fig. 1. Structural scheme of the Archive Data Storage System

3.1 Метасервер

Данный компонент является координирующим центром большей части процессов, циркулирующих в АСХД, и выполняет следующие основные функции:

- управление процессами загрузки/выгрузки архивных объектов;
- обработка (хранение/поиск/выдача) метаданных архивных объектов;
- управление очередью заданий;
- авторизация доступа к архивным объектам.

Программная реализация метасервера выполнена в виде многопоточной службы, взаимодействующей с остальными компонентами системы, а также с СУБД, обеспечивающей хранение всей метаинформации.

3.2 Медиасервер

Медиасервер является основным исполнительным компонентом АСХД, функцией которого является запись/чтение архивных объектов на архивные носители.

Программная реализация выполнена в виде многопоточной службы, в составе которой функционируют несколько компонентов:

- генератор планов записи;
- «финишер»;
- менеджер томов;
- главная коммуникационная нить;
- встроенный web-сервер;
- менеджеры ленточных накопителей.

В АСХД могут одновременно работать несколько медиасерверов, к каждому из которых подключаются одна или несколько ленточных библиотек. Медиасерверы конфигурируются таким образом, чтобы обеспечить параллельную работу множества ленточных накопителей. Каждый медиасервер эксклюзивно управляет накопителями и роботом ленточной библиотеки, поэтому текущая версия АСХД в большей степени ориентирована на использование библиотек midrange-класса. Медиасервер АСХД взаимодействует со SCSI устройствами (накопителями и библиотечными роботами), используя стандартные для Unix-подобных операционных систем средства [5, 6].

В связи с большой продолжительностью периодов записи/чтения ленточных накопителей взаимодействие метасервера с медиасервером реализовано в асинхронном режиме. Медиасервер функционирует в виде многопоточной службы, получающей задания от метасервера и обменивающейся данными с агентами.

Помимо ресурсов ленточной памяти каждый медиасервер оснащается дисковой памятью большого объёма, роль которой заключается в промежуточном хранении архивных объектов и обеспечении стабильности входного потока записываемой на ленточные накопители информации. Медиасервер хранит в этом дисковом кэше каждый полученный от агента архивный объект до тех пор, пока тот не будет полностью записан на ленточные носители. Аналогично, при восстановлении архивного объекта медиасервер сохраняет считываемый с ленточных носителей архивный объект в дисковом кэше и после успешного завершения этой операции отправляет восстановленный объект агенту.

3.3 Агент

Роль агента заключается в транспортировании пользовательских данных между ресурсами-источниками и АСХД, а также в поставке метаинформации web-клиентам. Разработанный для операционной системы Linux агент функционирует в виде системной службы, которая запущена на серверах, имеющих непосредственный доступ к ресурсам-источникам.

Являясь транспортным компонентом, агент обменивается архивируемыми (и восстанавливаемыми из архива) данными с медиасерверами под управлением метасервера.

3.4 Клиенты

К функциям клиента относятся:

- управление пользовательскими заявками (постановка в очередь, контроль, отмена) на архивирование/восстановление данных;
- назначение описательных атрибутов создаваемого архивного объекта;

- задание критериев поиска архивных объектов;
- удаление архивных объектов.

В настоящее время разработаны два вида клиента: консольный (для операционной системы Linux) и межплатформенный web-клиент.

На web-сервер АСХД возложена задача обслуживания как пользовательских запросов от web-клиента, так и выполнения административных запросов, поступающих от web-консоли администратора.

АСХД обладает собственным программным интерфейсом (REST API), позволяющим добавлять функции архивирования и восстановления данных в прикладные приложения.

3.5 АРМ оператора хранилища

К функциям АРМа оператора хранилища относятся:

- обслуживание заявок АСХД по набору и замене архивных носителей (картриджей) в библиотеках;
- инвентаризация хранилища ленточных картриджей;
- регистрация и последующее сопровождение ленточных картриджей.

АРМ оператора взаимодействует исключительно с базой данных метаинформации АСХД.

4. Модель данных

Типовая площадка АСХД может удовлетворять потребности пользователей нескольких ВВС, ресурсы хранения которых распределены между пользовательскими группами, сформированными по проектному принципу. Все существующие файловые ресурсы хранения, как и все проекты и пользователи, должны быть предварительно зарегистрированы в АСХД.

Операция архивирования данных инициируется постановкой соответствующего задания (заявки) в очередь АСХД. В результате успешной обработки задания метасервером формируется один или несколько архивных объектов, в которых инкапсулированы пользовательские данные, указанные в задании. В базу данных АСХД помещается метаинформация о каждом архивном объекте, которая включает в себя: числовой идентификатор; размер в байтах; информацию о владельце и времени создания; список доступа; текстовое описание и специфичный для организации, эксплуатирующей АСХД, набор определяемых пользователем тематических атрибутов, характеризующих содержимое архивного объекта.

Максимальный размер архивного объекта (параметр MAX_OBJECT_SIZE) ограничен объемом дискового кэша медиасервера, минимальный размер – 1КБ. В ходе выполнения архивирования объекты могут быть разрезаны на сегменты, назовем их чанками (от англ. chunk).

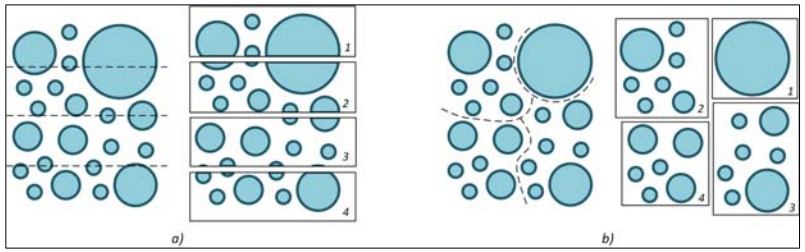

Рис. 2. Формирование чанков – (a) «распил» по размеру, (b) «распил» по границам
Fig. 2. Chunk generation – (a) splitting by size, (b) splitting along borders

Параметр CHUNK_SIZE определяет максимальный размер чанка. Если суммарный размер указанных в пользовательском задании данных больше CHUNK_SIZE, то соответствующий архивный объект разрезается на множество чанков одним из двух способов (см. рис. 2), иначе он состоит из одного чанка.

Каждый архивный объект обладает уникальным целочисленным идентификатором. Чанки архивного объекта также нумеруются по порядку. Под номером 0 регистрируется специальный чанк-дескриптор, хранящий подробную пользовательскую, взятую из заявки, и системную метаинформацию, описывающую архивный объект, а также контрольные суммы всех чанков данного архивного объекта. Чанки некоторых архивных объектов могут реплицироваться (см. п.5.3). Поэтому в ходе планирования операций записи/чтения медиасервер оперирует универсальными именами чанков, построенными по схеме:

<№ объекта> . <№ чанка> . <№ реплики>.

Например, в состав реплицируемого объекта 122 входят чанки «122.1.0», «122.2.0», «122.3.0», «122.4.0», «122.1.1», «122.2.1», «122.3.1», «122.4.1», а также два дескриптора «122.0.0» и «122.0.1» (см. рис. 3).

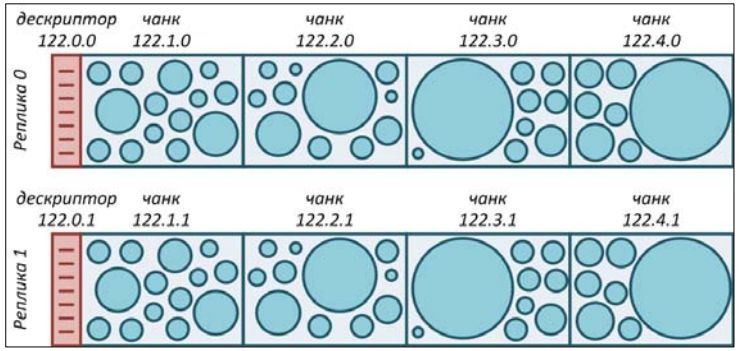

Рис. 3. Схема именования чанков
Fig. 3. Chunk naming scheme

Характерной проблемой систем хранения, использующих технологию магнитных лент, является достижение максимальной производительности ленточных накопителей, чаще всего связанное с недостаточностью входного потока информации. В этих условиях двигатель ленточного накопителя часто останавливается. Для продолжения записи он сначала перематывает ленту немного назад и затем, читая на низкой скорости записанные ранее данные, пытается найти метку конца предыдущей записи, чтобы продолжить запись вновь поступивших данных, постепенно увеличивая скорость протяжения ленты и, как следствие, производительность накопителя. К неприятным последствиям недостаточности входного потока также относится повышенный износ как самого двигателя, так и магнитных головок (эффект *shoe-shining*).

Поэтому для сокращения количества старт-стопных циклов накопителей разработчики систем хранения прибегают к обязательному предварительному кэшированию, контейнеризации предназначенных для записи данных и прочим организационно-техническим приемам.

Кэш медиасервера АСХД организован в виде локальной файловой системы достаточного объема, соизмеримого с максимальным объемом используемых типов архивных носителей. Для повышения согласованности реплик архивных объектов и для экономии потребляемых дисковых ресурсов репликация чанков реализована на этой файловой системе в виде *hard link*'ов.

Контейнеризация осуществляется путем группирования записываемых чанков в файле формата POSIX tar[7]. Тем самым медиасервер записывает группу чанков одним потоком, останавливая серводвигатель накопителя лишь один раз либо по причине выполнения работы (все чанки группы успешно записаны), либо при обнаружении «конца ленты» (носитель заполнен), либо в случае сбоя работы накопителя. В первом случае тот чанк, который оказался частично записанным, исключается из метаинформации текущей группы чанков и помещается в новую группу. Необходимым условием для старта записи группы является достаточный объем накопленных в кэше чанков, определяемый параметром MIN_DATA_SIZE_TO_WRITE.

В текущей версии АСХД используются три вида групп чанков (см. рис. 4). «Моно-группа» содержит чанки одного архивного объекта, размер которого превышает CHUNK_SIZE. Для записи множества небольших по размеру архивных объектов используется «ассорти-группа». В тех случаях, когда объем накопленных в кэше чанков недостаточен для формирования «ассорти-группы», а время ожидания соответствующих пользовательских заданий уже превышает некоторый лимит (параметр SMALL_TASK_WAITING), медиасервер использует «гибридную группу», в которую помещает наряду с накопленными чанками малоразмерных архивных объектов и один полноразмерный (CHUNK_SIZE) чанк какого-либо кэшированного объекта большего размера.

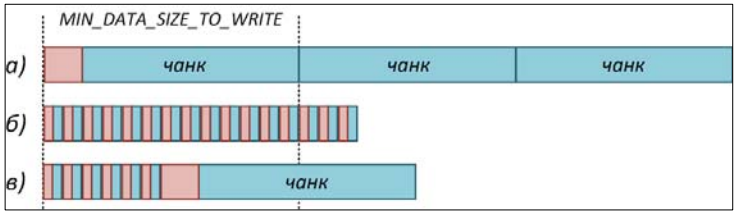

Рис. 4. Примеры трех видов групп чанков – (а) моно, (б) ассорти, (в) гибрид
Fig. 4. Chunk group samples – (a) mono group, (б) assorti group, (в) hybrid group

С целью повышения вероятности последующего восстановления архивного объекта разные реплики одного чанка никогда не помещаются в одну группу и на один носитель.

Первичная инициализация архивных носителей производится в момент их регистрации в АСХД, что снижает вероятность негативных последствий человеческих ошибок при работе с ленточными картриджами и облегчает сопровождение архивного фонда в долгосрочной перспективе. Низкоуровневый формат записи данных на архивный носитель (см. рис. 5) разработан с использованием стандартных меток [8].

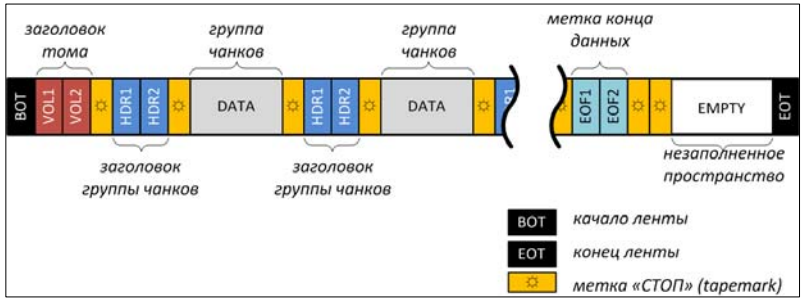

Рис. 5. Низкоуровневый формат тома
Fig. 5. Low-level volume format

5. Обеспечение целостности

Обеспечение целостности хранящихся архивных объектов в течение их срока жизни является первичной задачей АСХД. Она решается применением трех методов – использованием средств контроля целостности (контрольных сумм), введением избыточности (репликацией чанков), а также изоляцией архивных объектов.

5.1 Контроль целостности данных

Одной из болевых точек систем хранения, состоящих из большого количества компонентов, считается появление «скрытого повреждения данных», в борьбу с которым включены как производители дисков, дисковых подсистем, сетевого оборудования, так и разработчики промышленных файловых систем. Во многом эти усилия сосредоточены на внедрении механизмов сквозного контроля целостности по всей глубине стека блочного ввода/вывода – от серверного буфер-кэша до секторов на дисковых пластинах. Вместе с тем, контроль целостности информации в системах ленточной памяти по-прежнему лежит на плечах разработчиков прикладного ПО. В рамках АСХД контроль целостности данных осуществляется благодаря использованию контрольных сумм чанков на нескольких этапах их траектории движения (см. рис. 6). Согласно разработанному сетевому протоколу транспортной сессии, контрольная сумма передается непосредственно перед соответствующим чанком, что позволяет осуществлять однопроходной контроль целостности данных на приемной стороне.

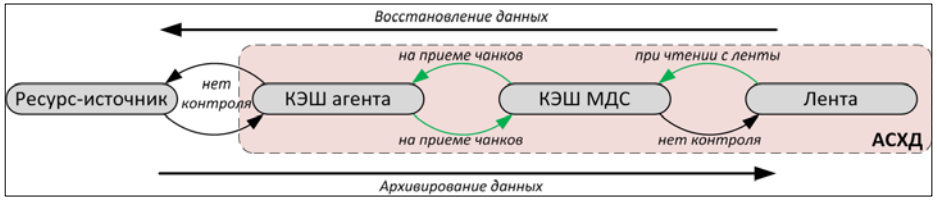

Рис. 6. Контроль целостности данных вдоль траектории их движения
Fig. 6. Data consistency control along the data path

Генерация контрольной суммы производится агентом на этапе считывания чанка из ресурса-источника, тогда как ее верификация производится:

- медиасервером при приеме архивируемых данных от агента;
- медиасервером при считывании данных с лент (в ходе выполнения операции восстановления или проверки архивных объектов);
- агентом при приеме восстанавливаемых данных от медиасервера.

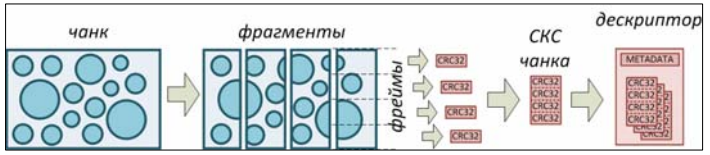

Рис. 7. Схема формирования «склеенной» контрольной суммы фрагмента
Fig. 7. Compound checksumming for a chunk fragment

Для обеспечения высокой скорости потокового контроля целостности в АСХД реализована идея «склеенных» контрольных сумм (далее по тексту СКС) чанков, заимствованная из проекта Apache HDFS [9]. В рамках транспортной сессии отправитель посылает очередной чанк в виде группы фрагментов фиксированного размера (параметр T_SESSION_BUFFER_SIZE). Непосредственно перед отправкой фрагмента рассчитывается его СКС как конкатенация 32-битных контрольных сумм более мелких порций данных (фреймов) размером T_SESSION_FRAME_SIZE (см. рис. 7).

Приемная сторона проводит однопроходной контроль целостности полученных фрагментов путем повторного расчета контрольных сумм фреймов и сравнения их с СКС, сформированной на стороне отправителя. По успешному завершению приема всех фрагментов чанка приемная сторона сохраняет все полученные СКС в виде цельной контрольной суммы чанка. После приема последнего байта архивного объекта, все контрольные суммы его чанков, наряду с подробной метаинформацией этого объекта, записываются в чанк-дескриптор, который затем помещается в «голове» каждой группы чанков на ленте.

На этапе восстановления архивного объекта с ленты считывание дескриптора (т.е. и всех контрольных сумм) непосредственно перед считыванием чанков данного архивного объекта позволяет таким же образом произвести однопроходной потоковый контроль целостности восстанавливаемых данных. Эта же схема применяется и далее, в ходе транспортирования данных от медиасервера к агенту.

5.2 Изоляция архивных данных

Для предотвращения случайного или намеренного повреждения архивных данных они должны быть максимально отдалены (изолированы) от наиболее вероятных источников проблем – от человеческой ошибки и от сбоев в программной или аппаратной части системы.

Изоляция архивного объекта в АСХД рассматривается и реализуется в нескольких аспектах:

- инкапсуляция архивного объекта – пользовательские данные преобразуются при архивировании в один из внутренних форматов АСХД и теряют связь с ресурсом-источником;
- ограничение пользовательского интерфейса – прямой доступ к архивному объекту запрещен, взаимодействие с АСХД осуществляется посредством заданий (заявок);
- организация оборота архивных носителей – заполненные пользовательскими данными ленты извлекаются из накопителей, а спустя определенный период времени они могут быть перемещены в хранилище.

Принятые меры по изоляции архивных данных позволяют, прямо или косвенно, снизить вероятность их повреждения со стороны возможных ошибочных действий пользователей, обсуживающего персонала и сбоев самой системы.

5.3 Опциональная репликация

Для улучшения сохранности информации в системах хранения широко применяется метод введения избыточности либо путем n-кратной ее репликации, либо добавлением кодов коррекции. В каждом отдельном случае выбор вида избыточности определяется, главным образом, скоростью кодирования, зависящей от сложности соответствующего алгоритма и имеющихся вычислительных ресурсов, а также объемом накладных расходов, т.е. объемом добавленной (избыточной) информации. Так как ленточная память обладает минимальной (по индустрии) удельной стоимостью гигабайта, а требования к обеспечению пропускной способности АСХД не предъявляются, то в этой системе применяется метод репликации.

Избыточность реализована на нижнем логическом уровне – на уровне чанков. Медиасервер компонует группы чанков таким образом, чтобы в рамках одной группы не было разных номеров реплик чанков одного архивного объекта. Тем самым гарантируется, что при включенной избыточности у архивного объекта повреждение одной группы его чанков не понизит шансы на целостное восстановление этого архивного объекта из другой группы.

Другая «линия обороны», повышающая вероятность восстановления архивированных данных в первоначальном состоянии, заключается в записи групп чанков, имеющих разные идентификаторы реплик, на разные носители. В случае потери или выхода из строя

отдельного носителя, на котором были сохранены чанки реплицируемого архивного объекта, медиасервер попробует восстановить недостающие чанки с другого носителя и собрать этот архивный объект.

Наконец идеальным вариантом является одновременное использование нескольких ленточных накопителей на этапе записи реплицируемого архивного объекта, когда подготовленные группы чанков, имеющих разные идентификаторы реплик, записываются параллельно. Тем самым, помимо сокращения времени записи, устраняется еще один фактор, потенциально влияющий на целостность записываемой информации – снижение качества работы отдельного накопителя, вызванное износом или загрязнением его магнитных головок.

6. Классы сервиса

Несмотря на достоинства, которые дает введение избыточности, для хранения некоторых типов данных применение репликации не всегда оправдано. Например, регулярное резервное копирование данных некоторой активной информационной системы порождает в СХ множество объектов хранения, в которых зафиксированы различные состояния одной и той же информационной системы в разные моменты времени. Если в период времени между двумя операциями резервного копирования состояние этой информационной системы изменяется лишь частично, то избыточность, в какой-то степени, появляется уже вследствие регулярности таких операций. Множественность таких сохраненных состояний информационной системы вскоре делает неактуальными наиболее старые из них, поэтому обычно они автоматически удаляются.

Так как потребность в резервном копировании возникает даже чаще, чем потребность в архивировании, а низкая удельная стоимость единицы хранения востребована в обоих этих случаях, и, учитывая, что в техническом плане эти функции не отличаются друг от друга, обе они были реализованы в рамках АСХД в виде так называемых классов сервиса.

Под классом сервиса здесь понимается именованная совокупность параметров сервиса АСХД по обработке пользовательской заявки, которую назначает ее владелец в момент постановки в очередь. В этот набор параметров входят:

- степень репликации порождаемого архивного объекта, т.е. количество реплик у каждого чанка;
- период времени, после которого архивный объект считается устаревшим и требующим удаления либо продления жизни;
- минимальный и максимальный объем объекта;
- размер чанка;
- приоритетность в обработке и пр.

Состав и значения параметров различных классов сервиса определяются техническими характеристиками площадки АСХД и особенностями технологических процессов обработки информации, существующих в организации, в которой установлена система. Настройку классов сервиса осуществляет администратор площадки АСХД.

Например, для сохранения данных в АСХД пользователю предлагаются два класса сервиса – A и B. Назначая класс A в качестве атрибута задания на архивирование, пользователь «говорит» архивной системе, чтобы сформированный на основе этой заявки архивный объект хранился в течение установленного в организации срока в нескольких репликах. Тогда как архивный объект, порожденный заданием класса B, будет храниться без репликации не более полугода, после чего будет автоматически удален системой.

Опциональная возможность включения репликации архивного объекта и задание других его свойств посредством назначения нужного класса сервиса позволяет пользователю выбирать в каждом отдельном случае наиболее подходящие способы обработки и хранения различных видов данных.

Другой важной особенностью классов сервиса является предоставляемая ими возможность реализации новых функций АСХД без необходимости изменять уже существующие отлаженные функции. Возможность добавления новых классов сервиса создает основание для дальнейшего развития системы при сохранении целевых свойств АСХД в отношении как новых, так и ранее заархивированных данных.

7. Типовой процесс

Для того чтобы заархивировать часть данных, расположенных на одном из ресурсов-источников, разработчик/пользователь соответствующего тематического проекта, зарегистрированный в АСХД, должен сформировать задание (заявку), включающее ресурс-источник, список архивируемых данных, их описание и прочую метаинформацию (см. рис. 8). Приняв в обработку это задание, АСХД считывает указанные в нем пользовательские данные, преобразует их и создает на их основе один или несколько архивных объектов, которые спустя некоторое время сохраняются на архивных носителях. В случае успешного окончания этой операции пользовательская заявка считается выполненной.

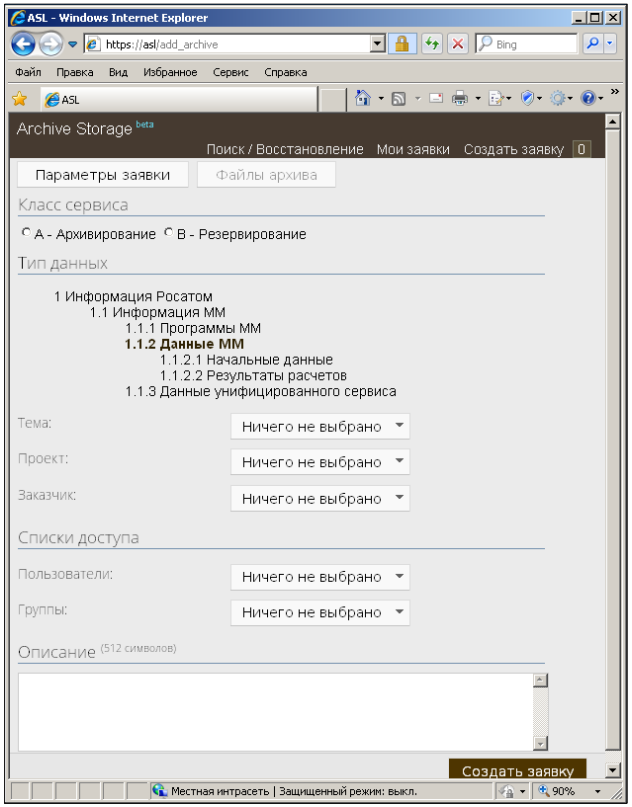

Рис. 8. Веб-форма создания пользовательской заявки
Fig. 8. Archival task creation web-form

Пользователь архивной системы может контролировать ход выполнения своих заданий и имеет доступ ко всей истории собственных заявок.

Найти необходимые архивные объекты в АСХД можно путем фильтрации всего списка архивных объектов по набору описательных атрибутов либо с помощью встроенной системы полнотекстового поиска, которая индексирует неформализованные текстовые описания каждой пользовательской заявки (см. рис. 9). Фильтрация списка объектов доступна как в консоли, так и в web-интерфейсе, а полнотекстовый поиск – только в web-интерфейсе.

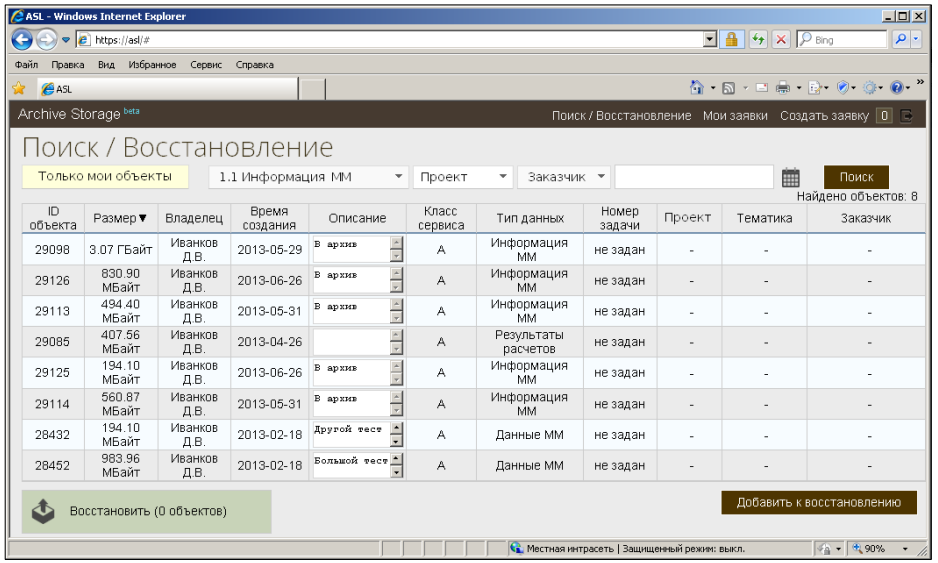

Рис. 9. Веб-форма поиска архивных объектов
Fig. 9. Archival object searching web-form

Для восстановления данных из АСХД пользователь должен знать идентификаторы соответствующих архивных объектов, которые он получает в результате выполнения операции поиска. Выбрав из результатов поиска релевантные архивные объекты, пользователь формирует задание, указывая в нем, помимо списка идентификаторов восстанавливаемых архивных объектов, ресурс-приемник, куда система должна будет поместить запрошенные пользователем данные. Система принимает это задание в обработку при условии, если (а) данный пользователь является владельцем запрошенных архивных объектов, или (б) он входит в списки доступа к этим объектам, или (в) он входит в группу «суперпользователей», имеющих доступ к любому архивному объекту.

Перед началом выполнения этого задания АСХД определяет набор архивных носителей, содержащих восстанавливаемые архивные объекты. В случае если соответствующие носители ранее были вытеснены в хранилище, АСХД формирует запрос оператору хранилища, который может либо самостоятельно (на АРМ оператора хранилища) считать запрошенные данные с носителей, либо инициировать их транспортировку и загрузку в ленточные библиотеки для последующей автоматической обработки. Считанные с архивных носителей пользовательские данные система помещает в указанное в заявке место на ресурсе-приемнике, преобразуя их в первоначальное, предшествующее архивированию, состояние.

8. Расчет площадки АСХД

При проектировании площадки АСХД для вычислительного центра необходимо провести расчет аппаратной составляющей системы, в частности определить количество и конфигурацию медиасерверов, тип, количество и конфигурацию автоматизированных

ленточных библиотек. Для этого построим простую линейную модель отдельного медиасервера, а на ее основе смоделируем работу всей АСХД (см. рис. 10).

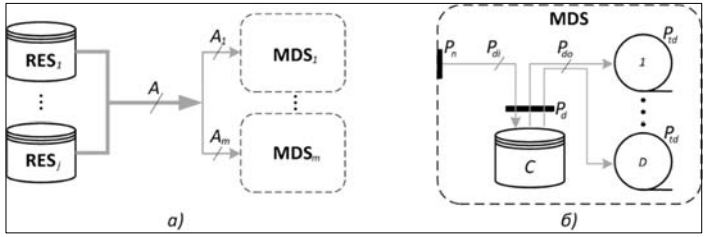

Рис. 10. (а) Модель архивной системы хранения и (б) модель медиасервера
Fig. 10. (a) The archive data system model and (б) the mediaserver model

Приведем основные положения этой модели.

- суммарный входной поток в АСХД (A) формируется из J ресурсов-источников данных (RES_j на рис. 10а), принадлежащих одной или нескольким ВВС;
- входной поток A равномерно распределяется между M однотипными медиасерверами (MDS_m на рис. 10а);
- считаем режим работы АСХД устоявшимся в течение определенного интервала времени, на протяжении которого суммарный объем регулярных пользовательских заявок на архивирование преобладает над объемом эпизодических заявок на восстановление данных; для удобства расчета длительность этого интервала выбираем равной 24 часам;
- рассматриваем медиасервер в виде накопительного устройства (дисковый кэш) и группы исполнительных устройств – ленточных драйвов (см. рис. 10б); поступающие на вход медиасервера данные сохраняются в кэше (C на рисунке 10б) и, по мере накопления достаточного суммарного объема, содержимое кэша записывается на ленточный носитель; для простоты считаем этот процесс линейным и бессбойным.

Параметрами модели медиасервера являются:

- пропускная способность сетевого адаптера (P_n), ограничивающая поток записи в кэш (P_{di});
- тип архивных (ленточных) носителей, определяющий в свою очередь ряд его технических характеристик: максимальную ёмкость носителя (V_t), максимальную производительность накопителя (P_t) и коэффициент аппаратного сжатия накопителя (K_z).

Также в модель вводятся коэффициенты, которые учитывают явления, имеющие место в работе реального медиасервера АСХД.

Реализация транспортного протокола (между агентом и медиасервером) вносит издержки, сужающие поток записи в кэш. Введем коэффициент снижения интенсивности сетевого потока K_n $(0 \le K_n \le 1)$

$$P_{di} = P_n \times K_n.$$

Производительность дискового контроллера в устоявшемся режиме обычно меньше пиковых «мгновенных» показателей, что связано с особенностями реализации буферизированного ввода/вывода операционной системы. Кроме того, полоса пропускания дискового контроллера может частично расходоваться и другими приложениями, запущенными на медиасервере. Обозначим возможное снижение КПД дискового контроллера через корректирующий коэффициент K_d $(0 \le K_d \le 1)$. Будем считать, что в устоявшемся режиме работы медиасервера потоки записи в кэш (P_{di}) и чтения из кэша (P_{do}) должны быть одинаковыми. Для выполнения этого условия достаточно оснастить медиасервер дисковым контроллером, имеющим производительность не менее P_d.

$$P_d = P_{di} * 2/K_d.$$

Ленточный накопитель достигает наибольшей производительности (P_t) только в условиях достаточного объема записываемых или считываемых данных, когда сервопривод работает с максимальной скоростью. В реальности эти условия не всегда достижимы ввиду как различий в размерах архивных объектов, так и особенности группировки чанков. Поэтому в модель медиасервера вводится корректирующий коэффициент K_t $(0 \le K_t \le 1)$, который отражает снижение производительности накопителя, вызванное разными скоростными режимами сервопривода и периодами вынужденного ожидания поступления данных в кэш.

Процесс записи накопленных в дисковом кэше данных на современные ленточные носители может длиться продолжительное время. Не редкой является ситуация, когда на фоне полной занятости накопителей (D_w) операциями записи на медиасервер поступает запрос на восстановление данных из архива. Для своевременной обработки таких запросов в модель медиасервера вводится параметр D_r, который определяет количество накопителей в медиасервере, зарезервированных исключительно под операции чтения. Таким образом, необходимое количество ленточных накопителей в медиасервере можно представить суммой:

$$D_m = D_w + D_r = \frac{P_d}{P_t \times K_t} + D_r$$

Периоды записи и чтения данных, выполняемые реальным ленточным накопителем, прерываются служебными операциями над картриджами – перемотка, загрузка/выгрузка, перемещение между накопителем и слотами хранения. Поэтому для расчета производительности медиасервера в устоявшемся режиме (P_m) вводится корректирующий коэффициент K_m $(0 \le K_m \le 1)$, отражающий влияние периодических остановок процесса записи данных.

$$P_m = P_t \times K_t \times D_w \times K_m$$

Непрерывный режим одновременной работы нескольких ленточных накопителей может быть обеспечен при наличии достаточного объема дискового кэша (C_m) в медиасервере. Для его оценки сверху учтем максимальную степень аппаратного сжатия, достижимую на выбранном типе драйва.

$$C_m = \frac{D \times V_t}{K_z}$$

Размещающиеся в голове каждой группы чанков контрольные суммы и метаинформацию соответствующего архивного объекта (см.п.4) следует отнести к накладным расходам емкости архивного носителя, объем которых описывается корректирующим коэффициентом Q (обычно $Q \le 0.01$).

Коэффициент компрессии (Z), описывающий степень сжатия записываемых данных, сильно зависит от самих этих данных. На практике он не превышает соответствующую паспортную характеристику накопителя $(1 \le Z \le K_z)$.

АСХД разрешает введение избыточности при записи архивируемых данных (см. разд. 6), поэтому при расчете потребления архивных носителей в устоявшемся режиме работы медиасервера добавим в модель коэффициент репликации $(1 \le R \le 2)$, описывающий среднюю степень избыточности среди множества пользовательских заявок.

Таким образом, количество медиасерверов (M), необходимое для обслуживания входного потока пользовательских данных в АСХД (A), можно оценить, отталкиваясь от вычисленной производительности отдельного медиасервера.

$$M = A/P_m$$

Количество носителей, которое потребляет АСХД для сохранения входного потока пользовательских данных, рассчитывается по следующей формуле.

$$N = \sum_{m=1}^{M} \frac{A_m \times R}{V_t \times (1 - Q) \times Z}$$

На основании вычисленного количества медиасерверов и ленточных накопителей можно спроектировать структуру аппаратной части площадки АСХД, а именно выбрать модель, конфигурацию и количество ленточных библиотек. Вычисленное количество потребляемых носителей за период позволяет определить общее количество слотов, которым должны обладать выбранные библиотеки.

Наличие долгосрочного контракта на поддержку ленточных библиотек со стороны производителя или интегратора создает наилучшие условия для сопровождения площадки АСХД. Если же такая поддержка недоступна, то в расчет аппаратной конфигурации площадки необходимо заложить некоторую степень избыточности.

Рассмотренный линейный подход к расчету аппаратной конфигурации площадки АСХД, безусловно, не претендует на полноту и точность, но позволяет провести быструю оценку объема необходимого оборудования для реализации функции архивного хранения данных в вычислительном центре.

9. Управление архивными носителями

Управление архивными носителями в АСХД может осуществляться в автоматическом и полуавтоматическом режиме.

Автоматический режим управления архивными носителями применим для площадки АСХД, весь архивный фонд которой помещается в слотах активных ленточных библиотек. Полуавтоматический режим актуален для организаций, чей архивный фонд превышает объем доступных ресурсов (количество слотов) активных библиотек. В таких ситуациях возникает необходимость организации off-site хранилища архивных носителей и разработки техпроцессов учета, розыска и транспортирования носителей между библиотеками и хранилищем. Участие в этих техпроцессах персонала (операторов хранилища) должно регламентироваться специальной политикой управления архивным фондом и разработанным регламентом обращения с архивными носителями.

Рассмотрим вопросы организации хранилища архивных носителей на примере типовой площадки АСХД, функционирующей в полуавтоматическом режиме.

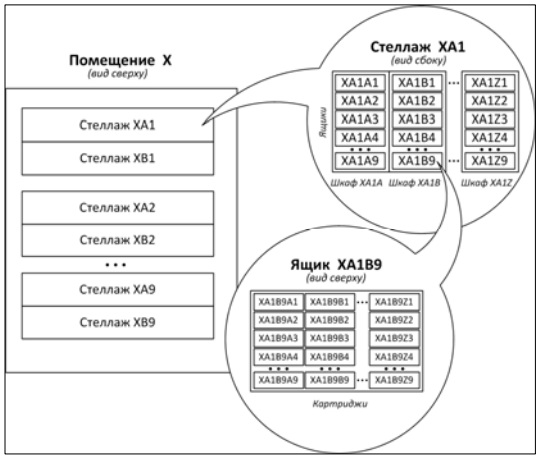

Рис. 11. Структура хранилища архивных носителей
Fig. 11. Structure of the vault

Структура хранилища (см. рис. 11) разработана так, чтобы облегчить оператору задачу поиска отдельных картриджей и упростить ориентирование в хранилище:

- хранилище может состоять из нескольких помещений;
- в каждом помещении хранилища устанавливаются стеллажи, состоящие из нескольких рядов соединенных друг с другом шкафов;
- в каждый шкаф входит несколько ящиков, в которых размещаются картриджи.

Картриджи уникально идентифицируются с помощью нанесенного на корпус штрих-кода. Каждый картридж поставляется и хранится в индивидуальной защитной пластиковой коробке. При регистрации нового картриджа в базе данных АСХД производится присвоение ему «адреса» в хранилище (см. рис. 12), который наносится на его коробку в символьной форме и в виде штрих-кода. Когда востребованные картриджи перемещаются в ленточные библиотеки соответствующие коробки остаются в архивном хранилище и никогда его не покидают. Транспортировка картриджей производится в специальных защитных контейнерах, обеспечивающих высокий уровень безопасности носителей. Задачи транспортирования носителей, поддержания порядка в хранилище и обеспечения соответствия метаинформации АСХД реальному состоянию архивного фонда возложены на службу операторов хранилища.

Рис. 12. Схема формирования «адреса» зарегистрированного картриджа
Fig. 12. «Address» of registered archived cartridge in the vault.

Работа оператора хранилища архивных носителей, вмещающего большое количество картриджей, построена вокруг АРМ оператора, который оснащен сканерами штрих-кодов и набором ленточных накопителей (всех используемых в АСХД типов носителей) в настольном исполнении. АРМ оператора управляется входящим в состав АСХД приложением «АРМ оператора хранилища» (см. рис. 13).

Использование сканеров штрих-кодов сокращает человеческие ошибки ввода информации и значительно облегчает работу оператора с множеством картриджей и коробок. Наличие ленточных накопителей в составе АРМ оператора позволяет оперативно выполнять пользовательские задания на восстановление отдельных архивных объектов, полностью или частично размещенных на носителях, которые были перемещены в хранилище, без необходимости транспортировки и загрузки этих носителей обратно в библиотеки. Тем самым время выполнения таких «точечных» заявок может быть сокращено.

В рамках АРМ регламентированы и реализованы следующие техпроцессы обращения с архивными носителями:

- передача носителей из хранилища для загрузки в ленточные библиотеки;
- приём в хранилище носителей, выгруженных из ленточных библиотек;
- инвентаризация хранилища архивных носителей;
- прием на учёт новых архивных носителей в установленном порядке и регистрация их на АРМ оператора хранилища АСХД.

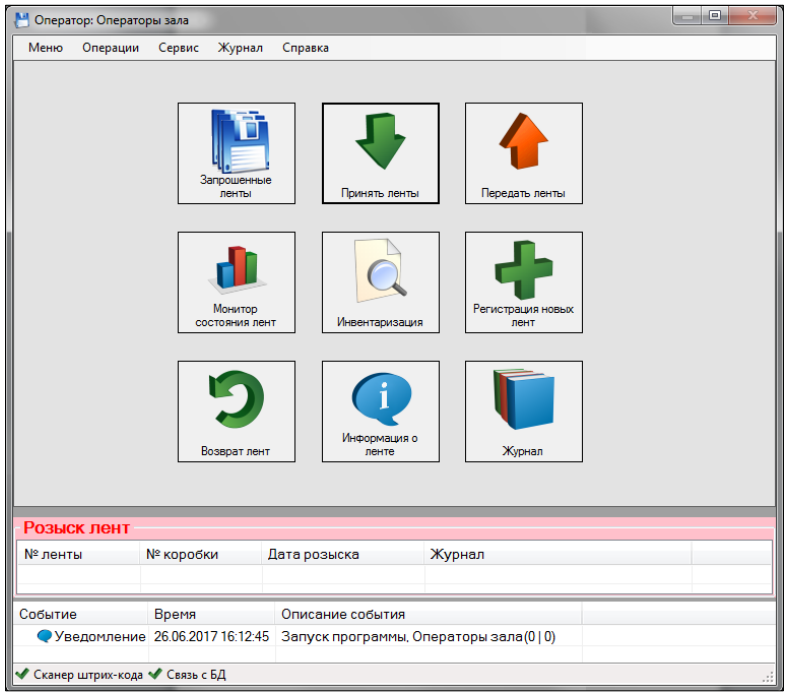

Рис. 13. Главное окно АРМа оператора хранилища
Fig. 13. Screenshot of the vault operator's desktop application

Для автоматизации техпроцесса инвентаризации хранилища используется мобильный сканер штрих-кодов с разработанным в рамках АСХД мобильным приложением, поддерживающим синхронизацию с АРМ и базой данных АСХД. Благодаря тому, что защитные коробки картриджей сделаны из полупрозрачного пластика, сканирование идентификаторов не требует изъятия картриджа из коробки, что существенно сокращает время инвентаризации хранилища.

11. Заключение

Использование архивной системы хранения данных в РФЯЦ-ВНИИТФ позволило решить проблему долговременного целостного хранения расчетных данных. За время ее эксплуатации были успешно обработаны десятки тысяч пользовательских заявок. Благодаря АСХД удалось реализовать методы управления расчетными данными в течение их жизненного цикла и тем самым повысить эффективность использования оперативных файловых систем ВВС.

Автор выражает надежду на то, что изложенный в данной статье подход мог бы быть полезен и другим специалистам, занимающимся организацией ресурсов долговременного хранения данных.

Список литературы / References

[1] СПО Супер-ЭВМ. Available at: http://vniitf.ru/article/spo-super-evm, accessed 28.08.2022 (in Russian).
[2] IBM HPSS. Available at: https://www.hpss-collaboration.org, accessed 28.08.2022.
[3] Enstore. Available at: https://www-stken.fnal.gov/enstore, accessed 28.08.2022.
[4] CERN Tape Archive. Available at: https://cta.web.cern.ch/cta, accessed 28.08.2022.

[5] Tape control program. Available at: https://github.com/iustin/mt-st, accessed 28.08.2022.
[6] Single or multi-drive SCSI media changer program. Available at: https://sourceforge.net/projects/mtx, accessed 28.08.2022.
[7] The Single UNIX Specification Version 3. Available at: https://unix.org/version3, accessed 28.08.2022.
[8] ANSI X3.27-1978. Available at: https://nulpubs.nist.gov/nistpub/Legacy/FIPS/fipspub79.pdf, accessed 28.08.2022.
[9] Apache Hadoop. Available at: https://hadoop.apache.org, accessed 28.08.2022.

Информация об авторе / Information about author

Дмитрий Владимирович ИВАНКОВ – начальник лаборатории. Сфера научных интересов: проектирование многоуровневых систем хранения данных, разработка высокопроизводительных систем хранения данных, исследования методов управления данными.

Dmitry Vladimirovich IVANKOV – Head of the Laboratory. Research interests: design of tiered data storage systems, development of high performance storage systems, research in data management methods.