КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2015 Т. 7 № 3 С. 621−630

Ки&М

СЕКЦИОННЫЕ ДОКЛАДЫ

УДК: 004.75

Ресурсный центр обработки данных уровня Tier-1 в национальном исследовательском центре «Курчатовский институт» для экспериментов ALICE, ATLAS и LHCb на Большом адронном коллайдере (БАК)

А. Я. Бережная[a], В. Е. Велихов, Ю. А. Лазин[b], И. Н. Лялин, Е. А. Рябинкин, И. А. Ткаченко

Национальный исследовательский центр «Курчатовский институт»,
Россия, 123182, г. Москва, пл. Академика Курчатова, д. 1

E-mail: [a]gridops@grid.kiae.ru, [b]Yury.Lazin@grid.kiae.ru

Получено 10 декабря 2014 г.

Представлен обзор распределенной вычислительной инфраструктуры ресурсных центров коллаборации WLCG для экспериментов БАК. Особое внимание уделено описанию решаемых задач и основным сервисам нового ресурсного центра уровня Tier-1, созданного в Национальном исследовательском центре «Курчатовский институт» для обслуживания ALICE, ATLAS и LHCb экспериментов (г. Москва).

Ключевые слова: высокопроизводительные вычислительные системы, системы распределенного массового хранения данных, системы распределенной обработки данных, грид

The Tier-1 resource center at the National Research Centre “Kurchatov Institute” for the experiments, ALICE, ATLAS and LHCb at the Large Hadron Collider (LHC)

A. Ya. Berezhnaya, V. E. Velikhov, Y. A. Lazin, I. N. Lyalin, E. A. Ryabinkin, I. A. Tkachenko

National Research Centre "Kurchatov Institute", 1 Kurchatov Sq., Moscow, 123182, Russia

The review of the distributed computing infrastructure of the Tier-1 sites for the Alice, ATLAS, LHCb experiments at the LHC is given. The special emphasis is placed on the main tasks and services of the Tier-1 site, which operates in the Kurchatov Institute in Moscow.

Keywords: high-performance computing systems, mass storage distributed system, distributed data processing, grid

Вычисления выполнялись на компьютерных ресурсах ЦКП «Комплекс моделирования и обработки данных исследовательских установок мегакласса».

Citation: *Computer Research and Modeling*, 2015, vol. 7, no. 3, pp. 621–630 (Russian).

Введение

Начиная с 2004 года вычислительные ресурсы НИЦ «Курчатовский институт» интегрированы в глобальную грид-систему — WLCG (Worldwide LHC Computing Grid, или Всемирный грид для Большого адронного коллайдера)[1] [LHC..., 2015] в Европейской организации ядерных исследований (ЦЕРН)[2] [CERN..., 2015].

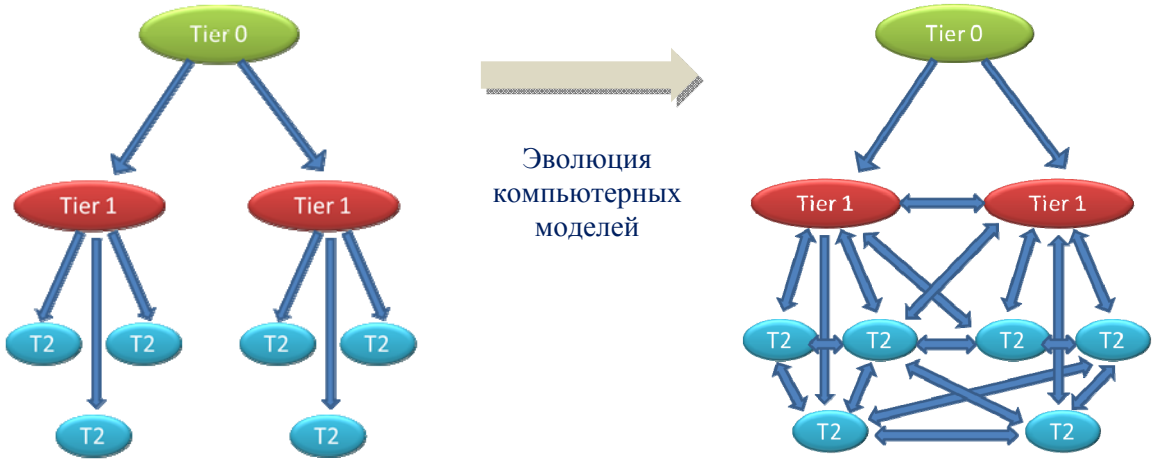
Ресурсный центр уровня Tier-1 НИЦ «Курчатовский институт» представляет собой высокоорганизованный вычислительный комплекс, включающий систему кондиционирования и систему резервированного питания с дизельной генераторной установкой (ДГУ).

Tier-1 НИЦ «Курчатовский институт» вносит вклад в распределенную обработку Больших Данных, получаемых в БАК-экспериментах ATLAS, ALICE и LHCb, и обеспечивает проведение полного цикла обработки экспериментальных и смоделированных событий, включающего в себя этапы приема/передачи исходных «сырых» данных, их последующую обработку, анализ и защищенное долговременное хранение. Доступ к ресурсам центра предоставляется всем участниками международной коллаборации WLCG, которая является самой большой академической распределенной вычислительной средой в мире, что обеспечивает полноправное участие российских исследователей в исследовательской и публикационной активности экспериментов БАК [The Large Hadron Collider, 2015].

Данная работа посвящена описанию инфраструктуры грид-компьютинга в международной коллаборации WLCG (разделы 1, 2). Раздел 3 включает описание ресурсного центра Tier-1 в НИЦ «Курчатовский Институт» (г. Москва).

1. Распределенная вычислительная инфраструктура грид-центров для БАК

Со времени создания (1998 год) по настоящий момент структура компьютерных моделей БАК-экспериментов претерпела эволюционные изменения (рис. 1).



(а) (б)

Рис. 1. Эволюция структуры компьютерных моделей БАК-экспериментов от строго иерархической до Mesh-топологии

Суть распределенной модели архитектуры компьютерной системы состоит в том, что первичная информации с детекторов БАК после обработки в реальном времени и первичной ее

[1] http://wlcg.web.cern.ch/
[2] http://www.cern.ch/

реконструкции в Tier-0 направляется для дальнейшей обработки, анализа и резервного хранения в региональные центры Tier-1, которые, в свою очередь, подключают к процессу распределенной обработки ресурсные центры уровня Tier-2.

Иерархия и соответствующие задачи центров каждого уровня определены в Меморандуме WLCG (Memorandum of Understanding for Collaboration in the Deployment and Exploitation of the Worldwide LHC Computing Grid) [Worldwide LHC..., 2014].

2. Центры уровня Tier-1

В настоящее время внутри WLCG-коллаборации функционируют следующие ресурсные центры Tier-1 (табл. 1).

Таблица 1

Ресурсный центр	***Обслуживаемые эксперименты***			
	ALICE	***ATLAS***	***CMS***	***LHCb***
Канада, TRIUMF		X		
Франция, CC-IN2P3	X	X	X	X
Германия, KIT	X	X	X	X
Италия, CNAF	X	X	X	X
Голландия LHC/Tier1	X	X		X
Скандинавские страны (NDGF)	X	X		
Республика Корея, GSDCatKISTI	X			
Россия, НИЦ «Курчатовский Институт»	X	X		X
Россия, ОИЯИ, г.Дубна			X	
Испания, PIC		X	X	X
Тайпей, ASGC		X	X	
Великобритания, RAL	X	X	X	X
США, BNL		X		
США, FNAL			X	

Для грид-пользователей ресурсные центры Tier-1 обеспечивают следующие сервисы, перечень которых официально определен в Меморандуме о взаимопонимании и согласован между коллаборацией WLCG и НИЦ «Курчатовский институт» [Memorandum of Understanding, 2014]:

1) предоставление управляемого дискового пространства, обеспечивающего постоянное и/или временное хранение данных для файлов и баз данных;
2) обеспечение доступа к хранимым данным со стороны других центров WLCG;
3) обеспечение работ конечных пользователей по анализу объектов данных;
4) предоставление других сервисов, например моделирования столкновений в соответствии с согласованными требованиями экспериментов;
5) обеспечение необходимой пропускной способности для обмена данными с центрами Tier-1 по согласованному между экспериментами и заинтересованными центрами Tier-1 плану.

Каждый из трех обслуживаемых в ресурсном центре Tier-1 НИЦ «КИ» экспериментов БАК имеет свою модель обработки экспериментальных данных, что, в свою очередь, определяет потребность в ресурсах центра (рис. 2, 3, 4) [LHC..., 2015]:

Обработка исходных данных состоит из автономной калибровки и обновления условий получения данных, а затем реконструкции и создания ESDs, AODs и обеспечения качества (QA — Quality Assurance) объектов. На рис. 2 показаны форматы данных и шаги по снижению их объема в процессе обработки. Типы данных и акронимы определены ниже.

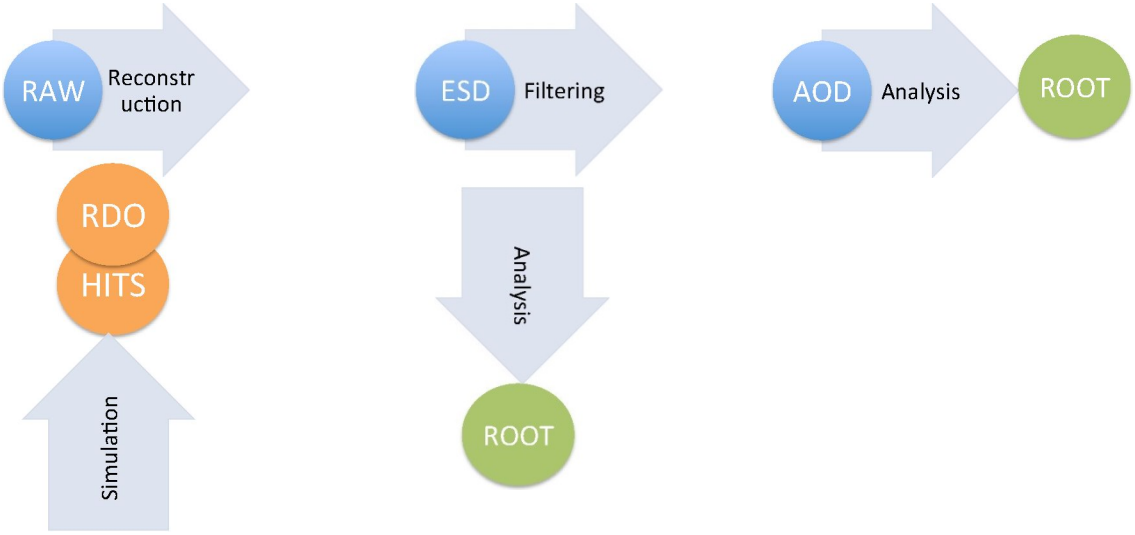


Рис. 2. Форматы данных и стадии их обработки в ALICE

Обработка смоделированных методом Монте-Карло событий и RAW-данных возвращает два анализа-ориентированных объекта: ESDs и AODs. ESDs содержит всю информацию, необходимую для любого анализа, последующие калибровки и проверки QA, а AODs содержит данные, подходящие для большинства задач анализа. Размер ESDs составляет 15–30 % от соответствующих исходных данных, в то время как AODs — около 30 % от соответствующего размера ESD. В основном для анализа физики предпочитают AODs-объекты. Некоторые очень специфические типы анализов выполняются на ESDs-данных. После обновления калибровки и программного обеспечения обновленные версии AODs получают через процедуру извлечения из ESDs с учетом новых условий.

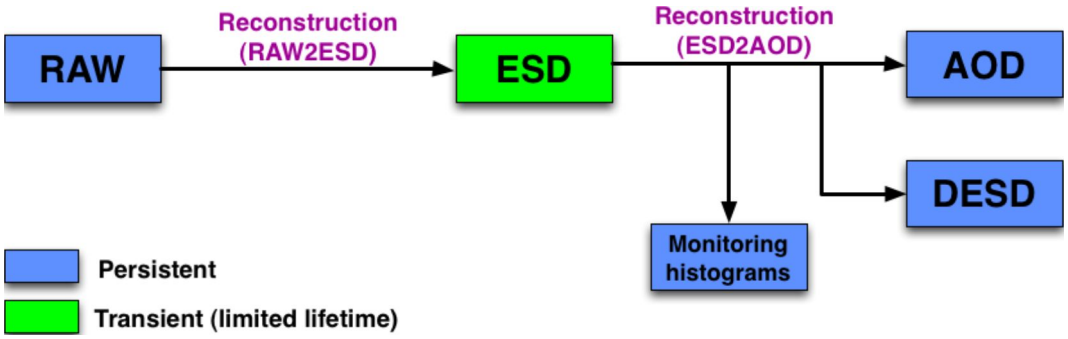


Рис. 3. Процесс реконструкции данных ATLAS

В ATLAS (рис. 3) первоначальная реконструкция данных выполняется на Tier-0. Исходные данные обрабатываются в два этапа в течение одного рабочего задания, вначале получается ESDs, а затем, на втором этапе, — AODs и DESDs. Данные RAW и выходы реконструкции экспортируется в ATLAS-грид-хранилище в соответствии с политикой репликации, то есть в Run-1 Tier-0 использовался для оперативной реконструкции данных вместе с быстрой калибровкой и оперативным определением качества данных и использовался как основное ленточное хранилище для RAW-данных. Ожидается, что в Run-2 роль Tier-0 останется без изменений, однако предполагается, что в Run-2 часть оперативной реконструкции данных может быть передана центрам Tier-1 в случает больших нагрузок на Tier-0.

Изменения в программном обеспечении ATLAS, а также изменения в распределенной вычислительной среде позволили центрам Tier-2 выполнять некоторые процессы, которые до это-

го были закреплены только для Tier-1 (репроцессинг, групповой анализ, реконструкция Монте-Карло) уже к концу Run-1. В Run-2 планируется совершенствовать это преимущество для оптимизации пропускной способности и нагрузки на сайты.

Центры Tier-1 по-прежнему будут иметь особую роль в качестве основного хранилища данных, а также предоставления второй копии RAW-данных с лент.

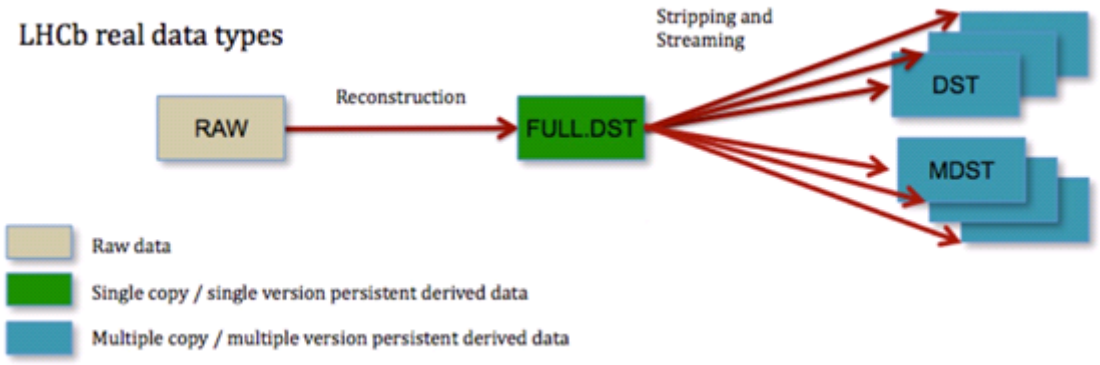


Рис. 4. Модель обработки данных LHCb

В LHCb (рис. 4) быстрая реконструкция и репроцессинг проходят идентично. На первом этапе необработанные данные восстанавливаются для получения FULL.DST, которые могут быть использованы в качестве входных данных для всех дальнейших действий по обработке данных. Второй этап состоит из выполнения приложения («зачистки»), которое выбирает события для физического анализа или калибровки; выполняется разборка нескольких сотен независимых потоков, каждый из которых связан с одним из десятков выходных потоков; в зависимости от анализа может использоваться, поток формата DST или MicroDST (MDST). Каждому MDST планируется также иметь одну копию DST (так называемую MDST.DST, не показанную на рис. 4), содержащую все события, соответствующие MDST-потокам, и от которой MDST может легко и быстро быть получен повторно. Это временная мера призванная стимулировать переход к использованию MDSTs путем предоставления «страховки», на случай если и когда будет найдена дополнительная информация, которая потребуется во время анализа. Это добавляет где-то 5–10 % к общему объему ленточного хранилища, необходимого для реальных данных, которые будут восстановлены, когда MDST-миграция завершится.

Поскольку шаг зачистки не зависит от шага восстановления, повторная зачистка может быть выполнена начиная с FULL.DST. Оба процесса — и реконструкция, и зачистка — организованы и планируются централизованно. Физические группы имеют доступ только к разбору выходных наборов данных. Ниже перечислены различные форматы данных:

RAW — необработанные («сырые») данные;

ESD — данные полной реконструкции событий (EventSummaryData);

AOD — данные, содержащие характеристики восстановленных физических объектов и используемые для физического анализа;

HITS — смоделированные данные чувствительности детектора;

RDO — смоделированные «сырые» данные;

DESD — сокращенные ESD для определенных целей;

FULLDST — полный выход реконструкции для всех физических событий после зачистки от шума;

DST — выход после зачистки: события, выбранные по критериям физики, полная копия восстановленного события плюс дерево(ья) распада частиц, которые вызвали отбор события;

MDST — данные, как DST, но содержащие только подмножество события (треки, PID, которое вызвало отбор события, и минимальные исходные данные).

3. Ресурсоемкость и функциональность центра уровня Tier-1 в НИЦ «Курчатовский институт»

В соответствии с требованиями, определенными соглашением по WLCG (MoU) [Memorandum of Understanding, 2014], ресурсный центр уровня Tier-1 в НИЦ «Курчатовский институт» в настоящий момент предоставляет для коллаборации WLCG ресурсы представленные в таблице 2.

Таблица 2

НИЦ КИ	2014
ЦПУ (HEP-SPEC06)	22700
Диски (ТБ)	2600
Ленты (ТБ)	2000

На данных ресурсах обеспечивается полная функциональность ресурсного центра уровня Tier-1, а именно:

– возможность управления большими объемами данных на высоких скоростях передачи данных;

– обеспечение необходимых характеристик для всех центральных процессоров и хранилищ;

– организация доступа к данным тысяч пользователей;

– обеспечение надежного долгосрочного хранения архивных данных.

Кроме того, вычислительная среда управляет обработкой реальных и моделируемых данных, а также предоставляет данные для пользовательских отчетов. При обработке данных для каждого эксперимента запускаются стандартные программы на все статистические наборы данных, в то время как для пользовательского анализа необходимый набор данных определяется только потребностями различных физических и аналитических команд и особенностями отчета, который будет выполняться.

Схема работы Tier-1 НИЦ «КИ» с указанием функций показана на рис. 5.

Tier-1 НИЦ «КИ» выполняет в автоматическом режиме следующие функции:

1) запись необработанных данных, получаемых из Tier-0 (ЦЕРН), и хранение их на ленточных накопителях;

2) запись обработанных данных, получаемых из Tier-0 (ЦЕРН), и хранение их на дисковых массивах;

3) предоставление хранимых данных другим центрам — Tier-1 и Tier-2;

4) обработку «сырых» (первичных) данных, получаемых на трех экспериментах — ATLAS, Alice, LHCb;

5) расчеты событий физического моделирования;

6) аутентифицированные и авторизованные запросы на загрузку и выгрузку экспериментальных данных;

7) аутентифицированные и авторизованные запросы на запуск вычислительных задач для обработки данных экспериментов БАК;

8) мониторинг загрузки всех структурных узлов системы и оперативное информирование обслуживающего персонала о приближении параметров мониторинга к критическим значениям.

Качество реализации функций обеспечивает полное выполнение входящих в их состав операций и задач и гарантирует корректную с точки зрения предметной области обработку данных и представление результатов.

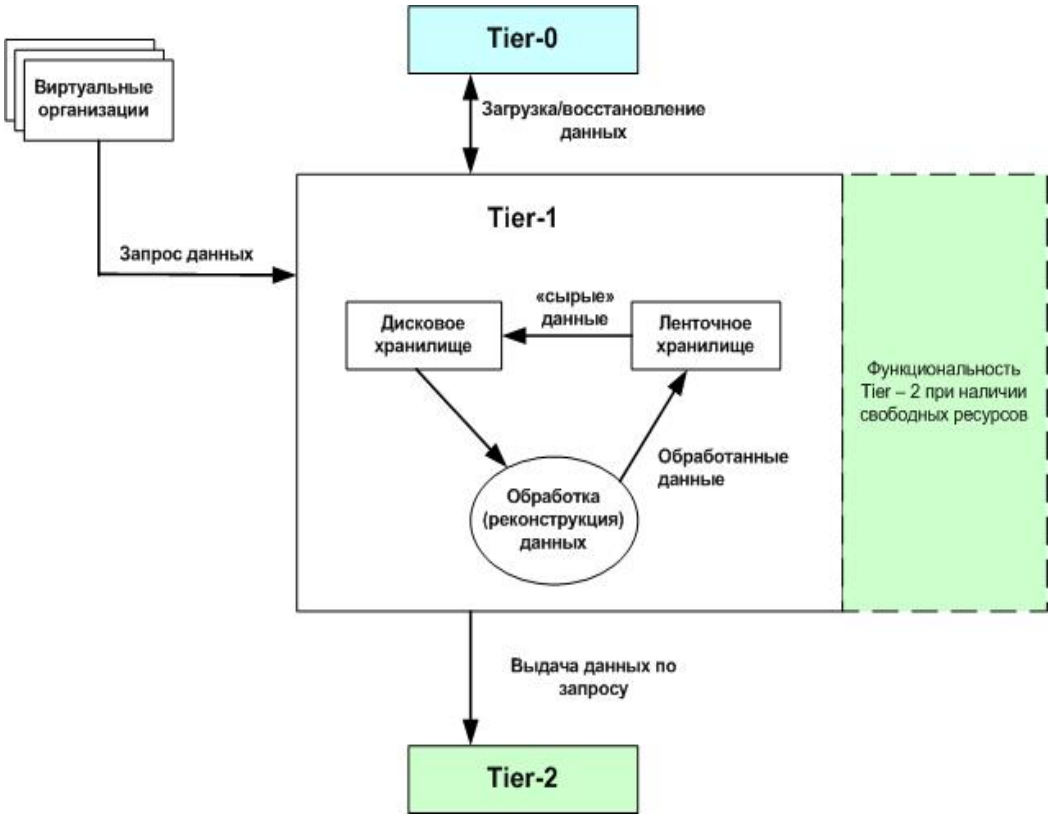


Рис. 5. Общая схема работы Tier-1 НИЦ «КИ»

Для надежного и эффективного управления инфраструктурой ресурсного центра осуществляется постоянный мониторинг его состояния средствами как внешнего, так и внутреннего мониторинга, как то:

– постоянное наблюдение за состоянием сервисов в грид-среде, как общих для всей инфраструктуры, так и сервисов в каждом ресурсном центре;

– сбор информации о количестве ресурсов (число процессоров, дисковое пространство) и их состоянии (свободные/занятые ресурсы);

– мониторинг выполнения заданий, передачи данных, запуск задач;

– отслеживание состояния каналов связи.

Центр работает в следующем режиме: в течение 24 часов, 7 дней в неделю, 365 дней в году.

Структурно комплекс технических средств ресурсного центра Tier-1 является слабосвязанным кластером на базе Intel-совместимых процессоров и сетей передачи данных, построенных с применением технологии Ethernet. Структура комплекса такова:

- **внешняя сеть кластера**, обслуживающая маршрутизируемые IP-адреса:

– устройства уровня ядра сети, обеспечивающие каналы передачи данных во внешние системы и являющиеся резервированными на уровне отказа единичных коммутаторов;

– устройства уровня доступа, подключающиеся к нескольким устройствам уровня ядра сети и обеспечивающие порты для подключения конечных устройств;

- **внутренняя сеть кластера**, обслуживающая немаршрутизируемые IP-адреса:

– устройства уровня ядра сети, обеспечивающие коммутацию пакетов между своими портами в неблокируемом режиме и являющиеся отказоустойчивыми на уровне выхода из строя отдельных компонентов коммутатора: блока питания, модуля портов, образа операционной системы;

– устройства уровня доступа, подключающиеся в ядро сети и обеспечивающие порты для доступа конечных устройств;

- **вычислительное поле**, состоящее из серверов, которые обслуживают вычислительные задачи, приходящие на Tier-1 через сервис CREAM CE;
- **сервисы хранения данных на дисках**, состоящие из серверов и дисковых массивов, подключенных к серверам по протоколам SAS и iSCSI;
- **сервисы хранения данных на лентах**, состоящие из:

– ленточных роботов и библиотеки, хранящей ленточные накопители и устройства считывания;

– серверов, обслуживающих устройства считывания, подсоединенные к машинам посредством интерфейсов FibreChannel;

– серверов и дисковых массивов, обеспечивающих дисковый буфер для ленточного хранилища;

- **сервисы gLite**/UMD-3: CREAM CE, site-BDII, APEL, Logging & Bookkeeping, top-BDII, VOBOX;
- **инфраструктурные сервисы**:

– сервисы фильтрации трафика и преобразования сетевых адресов,

– сервисы доменной системы имен,

– сервисы кеширования HTTP-запросов,

– сервисы локального мониторинга,

– сервисы терминального доступа для системных администраторов и обслуживающего персонала.

Схема структуры центра с указанием функциональных модулей показана на рис. 6:

Основные программные компоненты ресурсного центра Tier-1 В НИЦ «КИ»

С точки зрения программного обеспечения (ПО) структура ресурсного центра Tier-1 в НИЦ «КИ» состоит из следующих подсистем и сервисов.

- **Подсистема управления** компьютерными ресурсами на базе свободного ПО «Puppet [9, 10].
- **Подсистема передачи данных:**
  - сервисы передачи и хранения данных (dCache, EOS, Enstore).
- **Подсистема управления загрузкой:**
  - сервис управления счетными заданиями и пулом вычислительных узлов (CREAM CE);
  - сервис пакетного планирования запуска и выполнения заданий (Torque + MAUI);
  - информационные сервисы по ресурсам сайта (site-BDII, top-BDII);
  - сервис журналов (Logging и Bookeeping), хранящий информацию о задании.
- **Подсистема информационного обслуживания и мониторинга грид:**
  - сервисы сбора, хранения и предоставления информации этой подсистемы (Nagios, MonALISA, Panda, DIRAC).
- **Подсистема безопасности и контроля прав доступа:**
  - сервис выдачи и поддержки сертификатов.
- **Подсистема учета:**
  - сервис регистрации и учета вычислительных ресурсов;
  - сервис регистрации и учета ресурсов хранения данных.
- **Подсистема прикладного программного обеспечения экспериментов.**

Оперативное обслуживание и поддержка центров уровня Tier-2

Центр Tier-1 в НИЦ «КИ» функционирует в соответствии с вычислительными моделями БАК-экспериментов и согласно требованиям WLCG [LHC..., 2015]. С середины 2013 года центр участвует в процессе обработки полученных в Run-1 данных на поддерживаемых БАК-

экспериментах и осуществляет прием и хранение согласованных объемов экспериментальных данных и данных моделирования обеспечивая доступа к ним из других центров уровня Tier-1/ Tier-2 инфраструктуры WLCG.

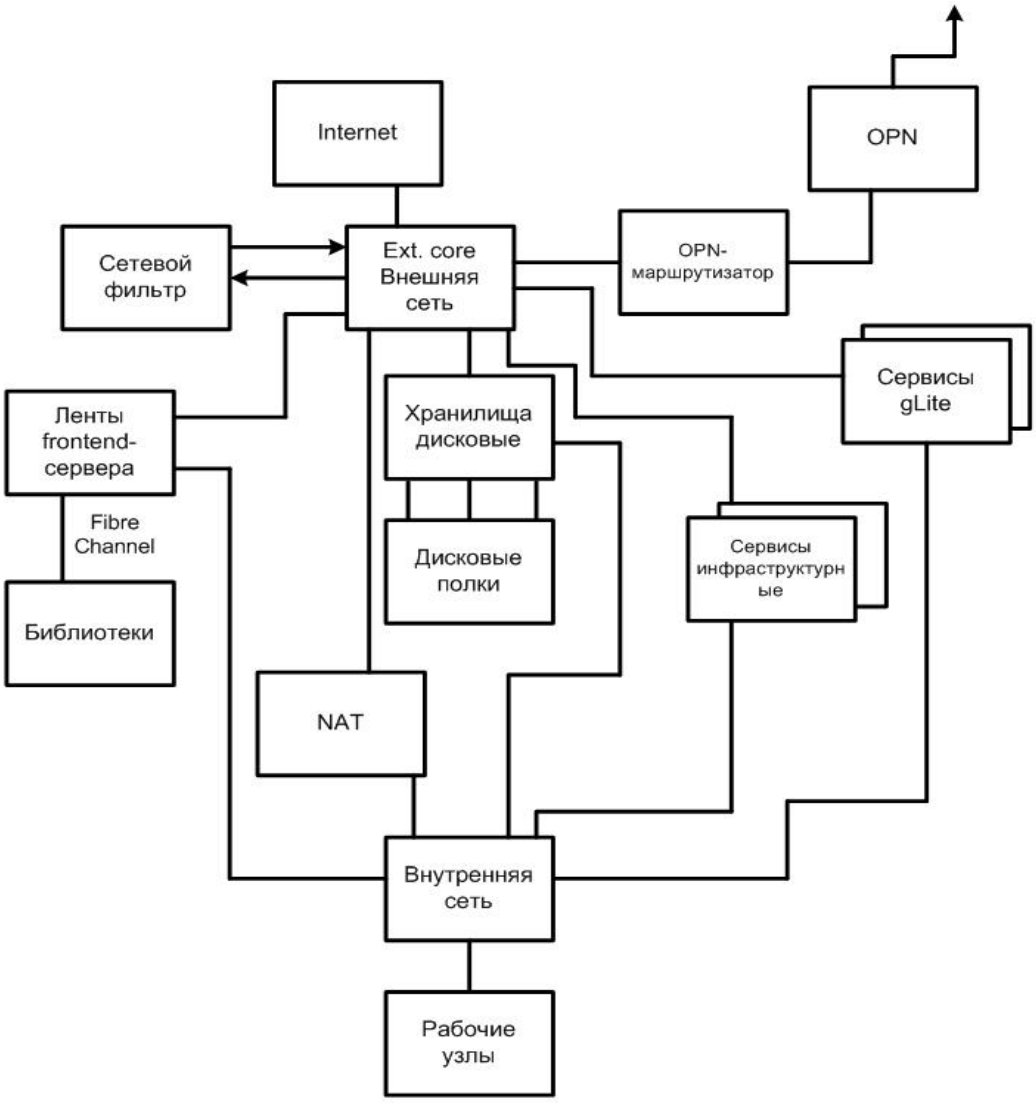


Рис. 6. Схема структуры центра Tier-1 НИЦ «КИ»

Во время Run-2 (с апреля 2015 года) на БАК Tier-1 НИЦ «КИ» будет выполнять основные функции центра, определенные в MoU [Worldwide LHC..., 2014], а также осуществлять:

- собственное оперативное обслуживание;
- поддержку региональных центров уровня Tier-2;
- поддержку грид-пользователей, включая консультации и помощь по специфическим проблемам грид-сервисов и выполнению счетных заданий;
- поддержку в разрешении инцидентов, связанных с безопасностью.

Список литературы

Климентов А., Кореньков В. Распределенные вычислительные системы и их роль в открытии новой частицы // Суперкомпьютеры. — 2012. — № 3 (11). — С. 7–11.

Ткаченко И. А. Опыт использования «Puppet» для управления вычислительным грид-кластером Tier-1 в НИЦ «Курчатовский Институт».

Aderholz M. et al. Models of Networked Analysis at Regional Centers for LHC Experiments (MONARC) — Phase 2 Report // CERN/LCB, 2000–2001 (2000).

ATLAS Collaboration: Observation of a new particle in the search for the Standard Model Higgs boson with the ATLAS detector at the LHC // Phys. Lett. B. — 2012. — Vol. 716. — P. 1–29.

CERN [электронный ресурс] // CERN, Switzerland. — 2015. — URL: http://public.web.cern.ch/public (дата обращения: 06.02.2015).

Dobre M., Stratan C. Monarc simulation framework // Proceedings of the RoEduNet International Conference, Buletinul Stiintific al Universitatii “Politehnica” din Timisoara, Romania, Seria Automatica si Calculatoare Periodica Politechnica, Transactions on Automatic Control and Computer Science. — 2004. — Vol. 49 (63). — P. 35–42. — ISSN 1224-600X.

LHC Computing Grid Technical Design Report. CERN-LHCC-2014-014 [электронный ресурс] // CERN, Switzerland. — 2015. — URL: http://lcg.web.cern.ch/LCG/public/default.htm (дата обращения: 17.01.2015).

Memorandum of Understanding [электронный ресурс] // CERN, Switzerland. — 2014. — URL: https://espace2013.cern.ch/WLCG-document-repository/MoU/countries/Russia/MoU-CERN-NRCKI_17JUL2014.pdf?Web=1 (дата обращения: 26.12.2014).

Puppet Labs [электронный ресурс] // London, United Kingdom. — 2014. — URL: http://puppetlabs.com/ (дата обращения: 26.12.2014).

The Large Hadron Collider [электронный ресурс] // CERN, Switzerland. — 2015. — URL: http://public.web.cern.ch/public/en/lhc/lhc-en.html (дата обращения: 06.02.2015).

Worldwide LHC Computing Grid Memorandum of Understanding [электронный ресурс] // CERN, Switzerland. — 2014. — URL: http://wlcg.web.cern.ch/collaboration/mou (дата обращения: 26.12.2014).