Письма в ЭЧАЯ. 2022. Т. 19, № 6(245). С. 660–668

КОМПЬЮТЕРНЫЕ ТЕХНОЛОГИИ В ФИЗИКЕ

ПРИНЦИПЫ РАБОТЫ СИСТЕМЫ АККАУНТИНГА ГРИД-САЙТОВ В ОИЯИ

И. А. Кашунин [1], *В. В. Мицын, Т. А. Стриж*

Объединенный институт ядерных исследований, Дубна

Раскрываются основные принципы функционирования новой системы аккаунтинга грид-сайтов Tier1 и Tier2 в ОИЯИ, являющихся компонентами Многофункционального информационно-вычислительного комплекса (МИВК) Лаборатории информационных технологий им. М. Г. Мещерякова. Новая система полностью повторяет функционал старой, значительно расширяет ее возможности. Также она интегрирована в общую систему мониторинга МИВК — LITMon.

This paper reveals the basic principles of the functioning of a new accounting system for the JINR Tier1 and Tier2 grid sites, which are components of the Multifunctional Information and Computing Complex (MICC) of the Meshcheryakov Laboratory of Information Technologies. The new system completely repeats the functionality of the old one, significantly expands its capabilities and is integrated into the general MICC monitoring system, i.e. LITMon.

PACS: 29.50.–v

ВВЕДЕНИЕ

В Лаборатории информационных технологий им. М. Г. Мещерякова (ЛИТ) Объединенного института ядерных исследований располагается Многофункциональный информационно-вычислительный комплекс [1]. Одними из его основных компонентов являются грид-сайты Tier1 и Tier2, входящие в распределенную вычислительную инфраструктуру Worldwide LHC Computing Grid (WLCG) [2], основная задача которой обеспечить обработку, хранение и анализ данных экспериментов на Большом адронном коллайдере. Для организации эффективной работы WLCG разработаны специальные системы учета ресурсов и их использования (системы аккаунтинга). В настоящее время учет ресурсов ЦПУ (центрального процессорного устройства) в WLCG основан на системе APEL [3] — для сбора и обработки данных — и портале учета EGI (European Grid Infrastructure) [4] — для визуализации. Портал обеспечивает пользователям создание отчетов, упорядочивая информацию, выбирая службы, ресурсные центры, группы пользователей за заданный период времени. Таким образом, можно получить информацию о работе собственных ресурсных центров в рамках общей распределенной инфраструктуры. Следует также отметить, что каждому ресурсному центру необходимо обеспечить его эффективную работу. Для оценки эффективности и получения статистических данных по различным параметрам, таким как время выполнения задач, их количество, эффективность использования ЦПУ, применяют специальные системы сбора и анализа данных — аккаунтинги.

[1] E-mail: miramir@jinr.ru

1. ОРИГИНАЛЬНАЯ СИСТЕМА АККАУНТИНГА

С начала 2000-х гг. в ЛИТ для организации вычислений на грид-сайте Tier2 применялось программное обеспечение (ПО) с открытым кодом: TORQUE (Terascale Open-Source Resource and QUEue Manager) [5] — менеджер ресурсов, отвечающий за отслеживание доступного количества ресурсов на узлах кластера и запуск задач. Он работал совместно с ПО Maui [6] — планировщиком заданий в параллельных и распределенных вычислительных системах (кластерах). Для оценки производительности данные системы предоставляли специальный скрипт для получения отчетов по параметрам, характеризующим функционирование вычислительных систем, таким как:

- #jobs — количество задач;
- ncore — количество ядер, требуемых для задачи;
- cpuclock — процессорное время (в часах), затраченное задачами с момента начала вычислений;
- wallclock — астрономическое время (в часах), затрачиваемое вычислительными задачами с момента поступления на выполнение;
- eff — эффективность — процентное соотношение процессорного времени к астрономическому (cpuclock/wallclock).

Отчеты хранились на специальном узле, доступном по веб-адресу, где можно было получить выписку данных за определенный период времени.

В 2015 г. в ЛИТ был введен в эксплуатацию вычислительный сегмент первого уровня для обработки данных эксперимента CMS на LHC — Tier1. Для его организации использовалось программное обеспечение, аналогичное Tier2. Система аккаунтинга Tier1 была также унаследована, однако в отличие от Tier2 определение процессорного и астрономического времени осуществлялось не по оценочному тесту specint2000 [7], а по HEPSpec06 (HS06) [8]. Таким образом, аккаунтинг оригинальной пакетной системы [9] представлял собой набор различных отчетов без каких-либо средств визуализации.

2. НОВАЯ СИСТЕМА АККАУНТИНГА

В 2021 г. TORQUE/Maui были заменены на новое программное обеспечение с открытым кодом — SLURM [10].

SLURM (Simple Linux Utility for Resource Management) является отказоустойчивой масштабируемой системой управления и планирования заданий для больших и малых кластеров Linux. Как менеджер рабочей нагрузки кластера, SLURM распределяет ресурсы, управляя очередью ожидающих запуска задач, предоставляет задачам доступ к вычислительным ресурсам на определенный период времени, обеспечивает выполнение и мониторинг вычислительных задач на выделенных элементах.

Разработка новой системы аккаунтинга. Для новой системы аккаунтинга был разработан алгоритм, принципы работы которого описываются 4-уровневой моделью, включающей в себя уровень сбора данных, уровень обработки данных, уровень хранения обработанных данных, уровень представления/визуализации данных (рис. 1).

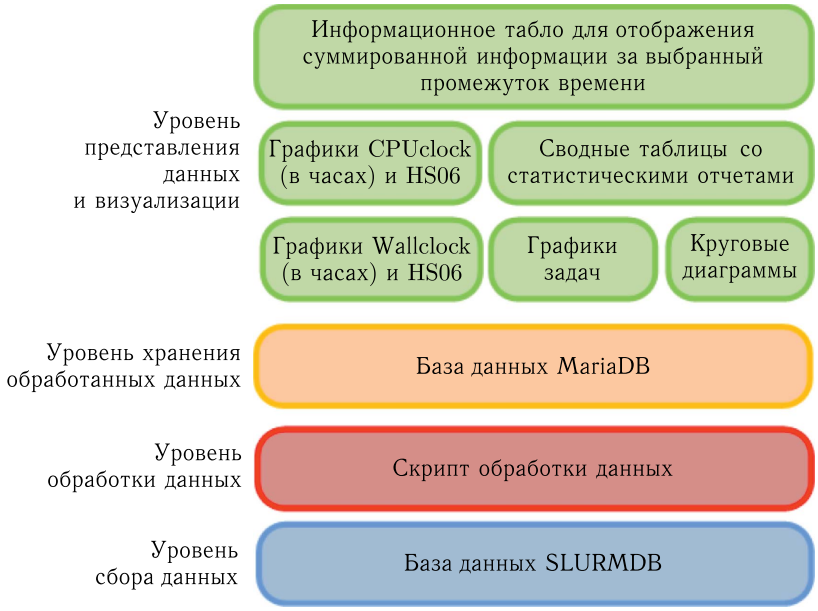

Рис. 1. Принципиальная схема работы системы аккаунтинга

Основная идея алгоритма заключается в том, чтобы получить данные из SLURM, обработать их, записать в специализированную базу данных и вывести на экран в виде графиков, таблиц, круговых диаграмм и информационных табло.

На первом уровне располагается база данных пакетной системы — SLURMDB. В ней помимо технических данных хранятся статистические данные о ее эксплуатации. В отличие от TORQUE/Maui, SLURM не предоставляет отчетов в привычном формате. Поэтому была поставлена задача о написании специального скрипта, который позволил бы, используя данные базы, генерировать отчеты, аналогичные отчетам из оригинальной системы, а также обрабатывать и выводить данные на консоль для записи в формате уже существующей структуры статистики.

На втором уровне представлен скрипт обработки данных. При сравнении этих двух пакетных систем были сформированы основные параметры для новой системы аккаунтинга, а также найдены их аналоги в SLURM (таблица).

Таким образом, в новой системе аккаунтинга необходимо провести расчет эффективности выполнения задач на вычислительном кластере (Eff). Суммы астроно-

Сопоставление параметров TORQUE/Maui и SLURM

Параметр	TORQUE/Maui	SLURM
Количество задач	#jobs	#jobs
Процессорное время	CPUclock	TotalCPU
Астрономическое время	Wallclock	Elapsed
Количество ядер для задачи	Average #nodes	NCPUS
Эффективность	Eff	Вычисляется

мического времени (HS06_Wallclock) и процессорного времени (HS06_CPUclock) за определенный промежуток в единицах HEPSpec06 рассчитываются как время работы всех задач (в часах), умноженное на HEPSpec06 — коэффициент K. Данный коэффициент применяется для перевода времени в единицы HEPSpec06, которые являются эталонными и применяются в том числе для сравнения временных показателей различных вычислительных кластеров, участвующих в обработке данных в физике высоких энергий, между собой. Для каждого из них коэффициент K вычисляется как усредненная величина, полученная на основе эталонных тестов, выполняемых на всех вычислительных машинах кластера.

Эффективность выполнения задач определяется процентным отношением суммарного процессорного времени к сумме произведения астрономического времени на количество ядер, используемого задачей (ncore).

Для работы скрипта необходимы следующие входные данные:

- время начала сбора данных — по умолчанию предыдущие сутки;
- время окончания сбора данных — по умолчанию текущее время;
- коэффициент перевода в единицы HEPSpec06 — по умолчанию на 2022 г. для JINR CMS Tier1 и JINR Tier2 он составляет 15,2;
- учетная запись пользователя — имя группы или название организации, выбранные по умолчанию, которые можно изменить.

При запуске скрипта с помощью специальных флагов можно активировать консольный вывод на экран или запись данных в базу. Для автоматизации заполнения статистики скрипт запускается специальной UNIX-службой Cron. Статистика пишется в базу за день, неделю, месяц и год.

Третий уровень представлен базой обработанных данных MariaDB [11]. Выбор данного ПО обусловлен тем, что это высокопроизводительное решение с необходимым функционалом, и оно применяется в системе мониторинга LITMon [10].

3. ВИЗУАЛИЗАЦИЯ ДАННЫХ

Отчеты в виде файлов имели ряд недостатков, основным из которых являлась необходимость последующей обработки с использованием дополнительного ПО (редактора) для визуализации и сравнения результатов. В настоящее время имеется достаточно много различных средств визуализации данных, одним из которых является Grafana [12] — мощная платформа с открытым исходным кодом для визуализации, мониторинга и анализа данных.

Грид сайт Tier1 в ОИЯИ используется не только для обработки данных эксперимента CMS, но и для моделирования эксперимента MPD проекта NICA [13]. Для системы аккаунтинга в Grafana был создан информационный дисплей с возможностью выбора различных параметров (метрик), перечисленных ниже.

- Учетная запись или группа пользователей эксперимента (account).
- Временной интервал статистики (time_period).

При расчете процессорного времени выполнения задач на вычислительном кластере данные рассчитываются за временные промежутки, такие как день, неделя, месяц и год. Для примера рассмотрим задачу, время работы которой превышает некоторый

time	jobs	cpuclock	wallclock	HS06_cpuclock	HS06_wallclock	cores	eff
2021-11-15	7942	1921465.67	238343.59	30224655	3749144.61	10	80.62%
2021-11-22	6430	1967215.89	245470.01	30944301.3	3861243.22	10	80.14%
2021-11-29	7083	1898790.18	229423.97	29867969.6	3608839.03	10	82.76%
2021-12-06	8500	1714744.34	238553.49	26972928.52	3752446.42	10	71.88%
2021-12-13	6271	1512060.80	183464.09	23784711.72	2885890.12	10	82.42%
2021-12-20	7120	2285426.37	260730.76	35949756.78	4101294.85	10	87.65%
2021-12-27	5740	2151781.28	260253.85	33847519.55	4093793.05	10	82.68%
2022-01-03	5357	2294026.26	262033.42	36085033.07	4121785.67	10	87.55%
2022-01-10	5354	2086921.87	230789.94	32827280.99	3629855.82	10	90.44%
2022-01-17	7447	2140156.91	247560.51	33664668.15	3894126.78	10	86.45%
2022-01-24	13763	2192285.62	259987.76	34484652.86	4089607.53	10	84.32%
2022-01-31	8009	2069921.95	251066.85	32559872.27	3949281.52	10	82.45%
2022-02-07	6397	2120472.88	251416.64	33355038.37	3954783.7	10	84.34%
2022-02-14	7504	1926905.68	243292	30310226.34	3826983.13	10	79.20%
2022-02-21	7984	1664339.56	212949.64	26180061.24	3349697.87	10	78.16%
2022-02-28	7954	2039451.25	256993.92	32080568.09	4042514.33	10	79.36%
2022-03-07	8315	2031912.82	248755.54	31961988.64	3912924.7	10	81.68%
2022-03-14	11668	1855880.58	250282.93	29193001.58	3936950.56	9.99	74.15%
2022-03-21	5906	2045531.11	245613.61	32176204.35	3863502.05	10	83.28%
2022-03-28	8546	1932418.02	239715.43	30711535.41	3770723.72	9.99	81.45%
2022-04-04	6275	1852616.00	234711.31	29141649.7	3692008.85	10	78.93%
2022-04-11	6213	2054747.40	254835.3	32321176.65	4008559.22	10	80.63%
2022-04-18	4791	1471779.90	200661.45	23151097.75	3156404.58	10	73.35%
2022-04-25	5681	1788557.70	227207.87	28134012.6	3573975.74	10	78.72%
2022-05-02	7473	1715687.44	216542.08	26987763.39	3406207	10	79.23%

Рис. 2. Визуализация данных для учетной записи cms_mcore Tier1

временной интервал. Задача, которая работает несколько дней и не умещается в интервал суток, не будет фигурировать в статистике за сутки, так как ее учет будет вестись на момент завершения. Следовательно, при наличии большого количества таких задач ежедневная статистика мало информативна. Если выбрать временной интервал в неделю, то таких задач уже будет значительно меньше и график будет более гладкий при тех же условиях. То есть чем больше временной отрезок, тем более корректно отображаются данные.

- Тип размерности (clock_type).
- Временной период статистики (дата).

В отличие от временного интервала, временной период является заданием времени отчета даты с определенного момента и по определенный момент. В консольном варианте данные параметры указываются с помощью флагов «–e» и «–s».

Ресурсы в Tier1 ОИЯИ активно используют 2 учетные записи — это cms_mcore и nica. Система визуализации Grafana позволила вывести на информационную панель в графическом и табличном представлении для каждой учетной записи следующие характеристики:

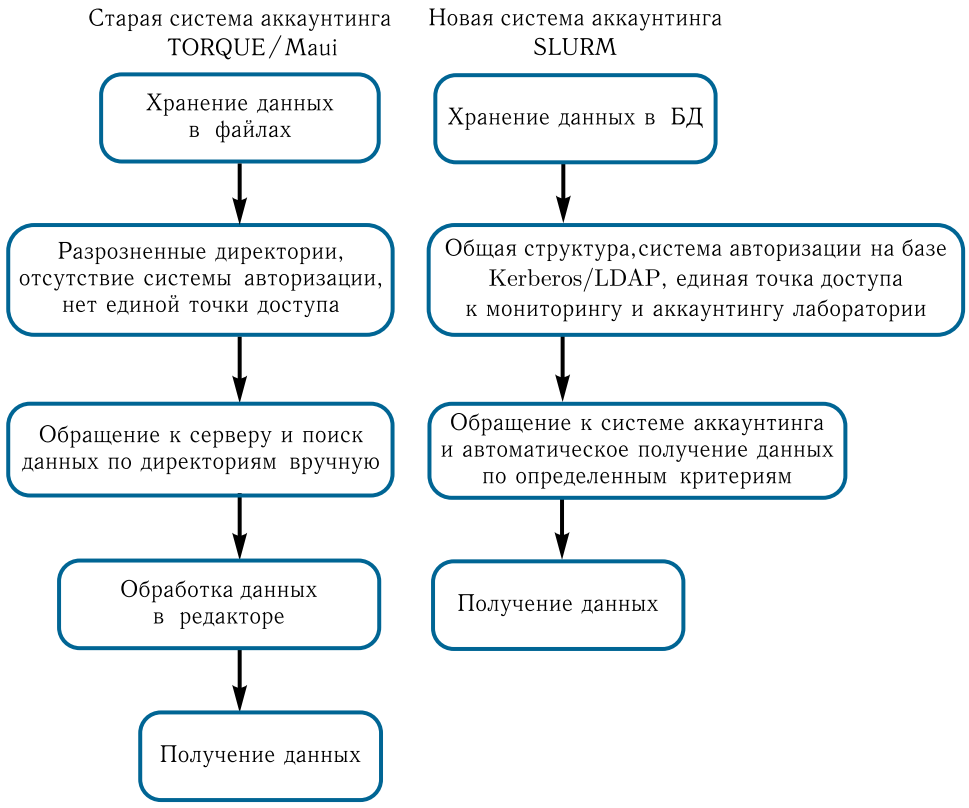

Рис. 3. Схема сравнения алгоритмов для обработки статистических данных системами аккаунтинга

Рис. 4. Визуализация данных для учетной записи cms_mcore Tier2

- Wallclock;
- CPUclock;
- HEPSpec06 Wallclock;
- HEPSpec06 CPUclock;
- суммы вышеперечисленных параметров за определенный промежуток времени;
- количество задач;
- эффективность использования вычислительного кластера.

Общий вид визуализации данных для Tier1 представлен на рис. 2.

Используя информационную панель системы аккаунтинга, пользователь имеет возможность получать различные распределения по интересующим его метрикам. Возможность построения графиков и диаграмм по различным параметрам позволяет сократить время обработки данных за счет автоматизации процесса, так как исключен процесс обработки в редакторе (рис. 3).

Для Tier2 был сделан аналогичный информационный дисплей. С учетом большего количества экспериментов в данном сегменте дополнительно были добавлены 2 круговые диаграммы (рис. 4). На них отображаются суммарное время использования ЦПУ в единицах HEPSpec06 CPUclock и количество выполненных задач на кластере Tier2 за определенный период времени, который задается в правом верхнем углу информационного дисплея.

ЗАКЛЮЧЕНИЕ

В рамках развития системы управления ресурсами грид-сайтов в ОИЯИ была создана новая система аккаунтинга, которая позволила значительно расширить функционал оригинальной системы, а также сократить время получения статистических данных за счет создания автоматической обработки данных системой визуализации. Реализованный подход обеспечивает отображение статистических данных напрямую из SLURM и позволяет осуществлять учет ресурсов и их использование как в рамках распределенной системы обработки данных, так и локально. Система визуализации предоставила мощный инструмент для анализа и составления различных отчетов, докладов и презентаций. Также стоит отметить интеграцию системы аккаунтинга в общую систему мониторинга LITMon [14]. Это позволило организовать единую точку входа и объединить разрозненные аккаунтинги в единую структуру.

Авторы благодарны Е. И. Лысенко за полезные замечания.

СПИСОК ЛИТЕРАТУРЫ

1. *Baginyan A., Balandin A., Balashov N., Dolbilov A., Gavrish A., Golunov A., Gromova N., Kashunin I., Korenkov V., Kutovskiy N., Mitsyn V., Pelevanyuk I., Podgainy D., Streltsova O., Strizh T., Trofimov V., Vorontsov A., Voytishin N., Zuev N.* Current Status of the MICC: An Overview // CEUR Workshop Proc. 2021. V. 3041. P. 1–8; http://ceur-ws.org/Vol-3041.
2. *Багинян А., Баландин А., Долбилов А., Голунов А., Громова Н., Кадочников И., Кашунин И., Кореньков В., Мицын В., Олейник Д., Пелеванюк И., Петросян А., Шматов С., Стриж Т., Воронцов А., Трофимов В., Войтишин Н., Жильцов В.* GRID at JINR // CEUR Workshop Proc. 2019; http://ceur-ws.org/Vol-2507/.

3. APEL. https://indico.cern.ch/event/55893/contributions/2041761/attachments/982855/1397352/APEL.pdf (accessed 14.02.2022).
4. EGI. https://accounting.egi.eu/ (accessed 14.02.2022).
5. Torque. http://dipc.ehu.es/cc/computing_resources/jobs/batch_systems/torque/ (accessed 15.07.2021).
6. Maui. http://dipc.ehu.es/cc/computing_resources/jobs/batch_systems/torque/ (accessed 15.07.2021).
7. Specint2000. https://en.wikipedia.org/wiki/SPECint (accessed 14.02.2022).
8. HepSpec06. https://www.gridpp.ac.uk/wiki/HEPSPEC06 (accessed 15.07.2021).
9. Batch. https://ru.wikipedia.org/wiki/%D0%9F%D0%B0%D0%BA%D0%B5%D1%82%D0%BD%D1%8B%D0%B9_%D1%84%D0%B0%D0%B9%D0%BB http://dipc.ehu.es/cc/computing_resources/jobs/batch_systems/torque/ (accessed 21.07.2021).
10. SLURM. https://slurm.schedmd.com/documentation.html (accessed 15.07.2021).
11. MariaDB. https://mariadb.org/ (accessed 15.07.2021).
12. Grafana. https://grafana.com/ (accessed 15.07.2021).
13. *Priakhina D., Trofimov V., Ososkov G., Gertsenberger K.* Data Center Simulation for the BM@N Experiment of the NICA Project // AIP Conf. Proc. 2021. V. 2377. P. 040007; https://aip.scitation.org/doi/10.1063/5.0063338.
14. *Кашунин И. А., Долбилов А. Г., Голунов А. О., Кореньков В. В., Мицын В. В., Стриж Т. А.* Система мониторинга многофункционального информационно-вычислительного комплекса // CEUR Workshop Proc. 2016. V. 1787. P. 235–240; http://ceur-ws.org/Vol-1787/256-263-paper-43.pdf.

Получено 19 апреля 2022 г.