Кластеры для ядерщиков

Вопросы построения вычислительных кластеров сегодня активно обсуждаются в ИТ-сообществе в целом и каждой предметной области в частности. Например, для обсуждения проблем создания и эксплуатации вычислительной инфраструктуры в ходе проведения экспериментов по физике высоких энергий регулярно проходят совещания HEPiX, анализ материалов которых позволяет понять текущее состояние дел в этой области.

Андрей Е. Шевель

Совещания HEPiX (High Energy Physics iT eXchange, wwwhepix.web.cern.ch) начали проводиться в 1991 году, и тогда их цель формулировалась следующим образом: HEPiX — международная группа сотрудничающих исследовательских институтов, которые экстенсивно расширяют использование ОС UNIX в экспериментах по физике высоких энергий. Усилия группы сфокусированы на: обмене опытом; информировании поставщиков вычислительной техники и организаций, готовящих стандарты; исследовании возможных решений проблем, которые имеют место в использовании UNIX в качестве основной ОС на главных вычислительных установках в институтах членах группы. Материалы данного совещания позволяют не только понять состояние дел в области компьютеризации лабораторий физики высоких энергий, но и выявить общие тенденции развития прикладных высокопроизводительных систем.

На первом совещании были рассмотрены следующие темы: пакетная обработка заданий; работа с магнитными лентами; управление программным обеспечением и распространение программ; переносимость программного окружения в физических вычислениях; поддержка оборудования и программного обеспечения распространенных персональных компьютеров. На этом совещании присутствовало 40 участников, среди которых были представлены организации: Argonne National Laboratory (ANL, USA), Brookhaven National Laboratory (BNL, USA), Fermi National Accelerator Laboratory (FNAL, USA), HEPnet, Istituto Nazionale di Fisica Nucleare (INFN, Italy), High Energy Accelerator Research Organization (KEK, Japan), Los Alamos Meson Physics Facility (LAMPF, USA), Lawrence Berkeley National Laboratory (LBL,USA), National Institute for Nuclear Physics and High Energy Physics in Amsterdam (NIKHEF, Netherlands), Stanford Linear Accelerator Center (SLAC, USA), Canada's National Laboratory for Particle and Nuclear Physics (TRIUMF, Canada). Сейчас такие совещания проводятся дважды в год, а число участников составляет 100-150 специалистов, представляющих ведущие научные учреждения мира и отвечающих за планирование развития кластеров и круглосуточное функционирование вычислительной инфраструктуры. Иными словами, информация на таких совещаниях поступает из первых рук. Аудитория состоит, как правило, из лиц непосредственно включённых в разработку компьютерной инфраструктуры. Все эти факторы позволяют отметить непредвзятость содержания сообщений.

Совещание HEPiX 2007 проходило по следующей программе:

- состояние/изменение основных вычислительных мощностей в различных институтах;
- решения и архитектуры;

• вторичная память и файловые системы;
• системный менеджмент;
• научный Linux (Scientific Linux) — информация о Scientific Linux 5 и о Scientific Linux в CERN;
• различные измерения в ЦОД и родственные проблемы;
• разное (на данном совещании были представлены соображения по увеличению надежности вычислительных установок с использованием тестирования и методики предсказания отказов).

На примерах данного совещания можно выявить основные проблемы построения и развития вычислительных кластеров.

СОСТОЯНИЕ ОСНОВНЫХ ВЫЧИСЛИТЕЛЬНЫХ МОЩНОСТЕЙ

В таблице приведены параметры нескольких вычислительных центров, развернутых в ведущих институтах.

Из таблицы виден масштаб некоторых кластеров уровня Tier 1 или T1 (деление кластеров, принятое в вычислительных распределенных grid-структурах для физики высоких энергий), хотя в ней упоминается не более трети кластеров уровня T1, которые планируются к использованию для обработки данных с ускорителя в ЦЕРН. Все кластеры имеют необходимое программное обеспечение для поддержки распределенных вычислений, программный инструментарий Globus и VDT (Virtual Data Toolkit, vdt.cs.wisc.edu) и массу дополнительных программ для реализации архитектуры grid.

В вычислительных grid-системах институтов физики высоких энергий данные хранятся в *элементах хранения* (Storage Elements, SE), представляющих собой сервер или кластер, содержащий необходимую вторичную память: дисковые массивы RAID5 или RAID6, а также роботизированную ленточную память. Общим местом для дисковых массивов является использование кластерных файловых систем: GPFS, Lustre и т.д. Передача информации между SE производится с помощью средств управления ресурсами хранения (Storage Resource Management, SRM). Данный термин употребляется в двух смыслах: SRM как стандарт на интерфейс к системе хранения/передачи данных и SRM как конкретная программная система. Сегодня используется стандарт SRM 2.2 (sdm.lbl.gov/srm-wg/doc/SRM.v2.2.pdf); более того, существует несколько реализаций систем типа SRM: CASTOR2 — иерархический сервер памяти на базе ленточной роботизированной библиотеки (совместная разработка ЦЕРН и RAL); dCache — иерархический сервер памяти, который может использовать различные роботизированные ленточные системы памяти (разработка DESY и FNAL); иерархический сервер памяти DPM на основе дисков (разработка ЦЕРН); StoRM — также иерархический сервер дисковой памяти, разработка INFN и ICTP, имеет интерфейс для работы с различными файловыми системами: GPFS, Lustre, XFS и др; BeStMan — иерархический сервер дисковой памяти, разработка LBNL. Все применяемые системы имеют интерфейс SRM 2.2, базирующийся на Grid Security Infrastructure (GSI) и других компонентах распределенной вычислительной среды.

КЛАСТЕРНЫЕ ФАЙЛОВЫЕ СИСТЕМЫ

Традиционные файловые системы, такие как ext3, xfs, reiserfs, обычно ассоциированы с отдельным физическим устройством (или его частью), например, дисковым накопителем, но в вычислительных кластерах, состоящих из сотен компьютеров, необходимы файловые системы, удовлетворяющие следующим требованиям:
• высокая пропускная способность по чтению/записи (главным образом за счет параллельного выполнения операций ввода/вывода);
• масштабируемость для большого числа узлов кластера;
• поддержка файлов большого размера (несколько терабайт);
• целостность данных, единое логическое пространство памяти и единое пространство имен файлов и каталогов;
• способность к продолжению работы после выхода из строя отдельных элементов.

Обычно кластерная файловая система содержит большой объем данных, распределенных по нескольким (или многим) серверам хранения, которыми пользуется множество клиентов для одновременного и независимого выполнения операций записи и/или чтения данных. Среди рассмотренных на совещании систем были PANASAS, GPFS и Lustre. К сожалению, никто не производил их сравнения, однако были

	Пакетная система выполнения заданий	Число машин в кластере	Емкость памяти на дисках (Тбайт)	Емкость памяти на лентах (Пбайт)	Пропускная способность внешних каналов связи (Гбит/с)
LAL/IN2P3	BQS	~900	~450	~1,6	23
TRIUMF	LSF	~200	~300	н/д	15
BNL	Condor/LSF	~2400	~900	~4	н\д
DESY	LSF	~1200	~430	~2	10
INFN-T1	LSF	~1600	~950	~0,75	30
RAL	н/д	н/д	~950	н/д	20
GridKa	н/д	н/д	~1550	н/д	10
SLAC	н/д	~1700	~750	н/д	н/д

обнародованы планы сокращения использования PANASAS и перехода на GPFS. Достигнутая средняя скорость (в течение недели) записи/чтения данных в кластерных файловых системах составляет более 1 Гбайт/с.

Всё больший интерес представляют попытки интеграции кластерной файловой системы и массовой долговременной памяти, например, GPFS и роботизированной библиотеки на магнитных лентах HPSS. В частности, планируется совместить эту интеграцию с процессом резервного копирования. Иными словами, если какие-то файлы уже записаны на ленточный робот, например, в связи с низкой частотой использования, то их не следует восстанавливать, например, после аварии с дискового накопителя.

ВИРТУАЛИЗАЦИЯ

В общем случае на каждом виртуальном узле может быть отдельная операционная система, отличная от ОС на соседнем виртуальном узле. Такого рода виртуализация может быть организована с использованием различных программных систем, например, VMware, Xen и др. Ряд ведущих институтов активно экспериментируют сегодня в области виртуализации вычислительных узлов больших кластеров, ориентируясь на такие системы как Xen, VMware, Solaris containers, User Mode Linux (UML) и Vserver. Наибольшей популярностью пользуются Solaris и Xen. Ожидается, что в недалеком будущем Xen и Solaris containers будут взаимозаменяемы.

УПРАВЛЕНИЕ СИСТЕМАМИ

Количество версий ОС и сложных сервисов растет на больших кластерах как снежный ком. Как следить за всем этим многообразием? Основной целью деятельности группы мониторинга сервисов в grid является повышение надежности работы grid-конфигурации путем сбора, обработки и представления корректной информации о состоянии различных сервисов. Важность работы определяется тем, что распределенная по разным континентам вычислительная сеть объединяет около 20 кластеров T1 (из разных институтов) примерно по тысяче машин и в среднем около 1 Пбайт дисковой памяти в каждом, не считая огромных роботизированных ленточных хранилищ. В этой сети

Окно в будущее

Казалось бы, сравнительно недавно суперкомпьютеры преодолели планку производительности в один TFLOPS, а уже не за горами эра PFLOPS (квадриллион, или 10^{15} операций над числами в формате с плавающей запятой в секунду). Первые такие суперкомпьютеры, по прогнозам экспертов, появятся в 2008-2009 гг., а в Европе уже стартовала подготовка к нескольким мета-проектам.

В среднем вычислительная мощь настольных ПК отстает от уровня производительности суперкомпьютеров на 13 лет, иными словами, по уровню производительности сегодняшние профессиональные ПК соответствуют суперкомпьютерам 13-летней давности, поэтому нынешнее положение дел в суперкомпьютерах можно считать ориентиром ситуации на компьютерном потребительском рынке в следующем десятилетии.

В свое время профессор и писатель Стив Чен попытался рассчитать, какая производительность необходима для решения различных задач будущего. По его мнению, для задач аэродинамики хватит производительности в несколько PFLOPS, для задач молекулярной динамики потребуется уже 20 PFLOPS, а для вычислительной космологии — производительность на уровне 10 экзафлопсов ($10x10^{18}$ FLOPS). Для задач вычислительной химии потребуются еще более мощные системы. По мнению Стива Павловского, главного директора по технологиям и генерального менеджера по архитектуре и планированию подразделения Digital Enterprise Group Intel, появление компьютеров с производительностью в секстиллион операций над числами в формате с плавающей точкой в секунду (10^{21}) можно ожидать уже к 2029 году.

В последнее десятилетие произошли заметные сдвиги в организации научного процесса — заметно усилилось направление компьютерного моделирования и эксперимента, что позволяет повысить эффективность процессов научного и технологического поиска. Становится возможным моделировать сложные физико-химические процессы и ядерные реакции, глобальные атмосферные явления, развитие экономики и промышленности.

Однажды Жан-Ив Блан из французской компании CGGVeritas, занимающейся геофизической разведкой, выразил формулу динамики роста потребностей человечества в высокопроизводительных вычислениях одной простой фразой: «Дайте мне в 100 раз большую вычислительную мощность, и я найду ей применение». Затем, выдержав паузу, добавил: «Дайте мне в 1000 раз большую мощность, я и ей найду применение».

Действительно, для разработки одного месторождения разведывательная компания должна обработать информацию объемом не менее 100 Тбайт, что в десять раз больше всего объема печатной информации, накопленной библиотекой Конгресса США. Стоимость бурения морской нефтяной или газовой скважины колеблется от 5 до 100 млн долл. и выбор бурения в неподходящем месте — не самое прибыльное вложение средств. Сегодня многие компании пересматривают свое отношение к данным, которые собирались в течение последних двадцати лет лишь для того, чтобы пылиться в архивах. Эти архивы содержат терабайты данных, которые в то время были слишком сложными для обработки и поэтому слишком рискованными для проведения разведывательных работ, стоимость которых может превышать 100 млн долл. Благодаря последним достижениям в сфере HPC «старые» данные предоставляют информацию о новых источниках энергии.

Год назад были завершены работы по созданию самого мощного суперкомпьютера в Европе — системы TERA-10, максимальная производительность которой составляет более 50 TFLOPS. Это суперкомпьютер создавался для научно-исследовательских целей в рамках оборонной программы Simulation Комиссариата по атомной энергии Франции (CEA) и находится в предместьях Парижа.

ежедневно выполняется несколько десятков тысяч заданий от сотен пользователей со всего мира и передаются десятки терабайт данных. Требуется постоянно следить за состоянием такой сети, чтобы иметь возможность предпринять контрмеры при возникновении неплановых ситуаций: снижение пропускной способности сети, отключение питания, выход из строя большого сегмента сети, выход из строя большого кластера и т.п. Имеется несколько прототипов решений по мониторингу, в частности, GridView, GridICE, GOCDB, Gstat, MonaLisa, Nagios, которые уже используются на конкретных кластерах, однако сложность задачи определяется рядом факторов (twiki.cern.ch/twiki/bin/view/LCG):

- методами сбора данных о работе отдельных компонентов grid (например, локально или удаленно);
- форматами данных о работе отдельных компонентов grid и протоколами передачи этих данных;
- организацией хранения собранных данных;
- формами отображения данных.

Системы мониторинга кластеров типа Nagios, Ganglia, Quattor и cfengine используются сегодня на большинстве кластеров в физике высоких энергий. Собранные данные мониторинга (в дополнение к основной массе физических данных) могут занимать немалый объем, особенно если планируется хранить сведения о состоянии вычислительной системы за длительное время (например, год или более), и здесь возникает проблема надежности обеспечения доступа к данным на дисках, причем в большинстве случаев молчаливо предполагается, что данные на дисках не изменяются со временем, но так ли это?

НАДЕЖНОСТЬ ХРАНЕНИЯ ДАННЫХ НА ДИСКАХ

При больших объемах памяти, исчисляемой сотнями терабайт, в силу вступают законы больших чисел — если что-то нежелательное, например, сбой при операции чтения, случается раз в год на нескольких дисках, то в системе с тысячами дисков такое может происходить несколько раз в день. Какие изменения происходят в системе после неплановой отключения питания? На маленьких установках,

Вычислительная система TERA-10 работает под управлением ОС Linux и построена на базе 4352 двухъядерных процессоров Intel Itanium 2 (Montecito), а 544 сервера NovaScale 6160 компании Bull, каждый из которых содержит по восемь двухъядерных процессоров, предназначены для обработки данных. Объем памяти TERA-10 составляет 27 Тбайт, а дискового пространства — 1 Пбайт.

В Японии был представлен суперкомпьютерный центр производительностью 59 TFLOPS, состоящий из компьютера Hitachi мощностью 2,15 TFLOPS и компьютера IBM Blue Gene Solution мощностью 57,3 TFLOPS. Суперкомпьютер используется для научных исследований в области ядерной физики и физики частиц.

В Институте машиноведения им. А.А. Благонравова РАН (ИМАШ РАН) был установлен векторный суперкомпьютер NEC SX, который будет использоваться для моделирования и расчетов в виброакустических исследованиях. Суперкомпьютер состоит из одного узла, в котором установлено 4 векторных процессора с пиковой производительностью 8 GFLOPS. Сегодня наблюдается возрождение интереса к векторной технологии, которая особенно хорошо подходит для решения задач прогноза погоды и моделирования климатических изменения, доказавшей свою эффективность в метеорологических службах Франции, Великобритании, ФРГ, Дании, Австралии и Сингапуре. Кстати, суперкомпьютер Earth Simulator (симулятор Земли), состоящий из 640 узлов и 5120 процессоров, имел производительность в 35,9 TFLOPS, что позволило ему оставаться одной из самых мощных системой в мире с 2002 по 2004 гг. в списке TOP500.

В начале 2007 года корпорация IBM и американский Национальный центр исследования атмосферы (National Center for Atmospheric Research, NCAR) заявили о завершении монтажа первой очереди суперкомпьютера BlueICE, являющегося составной частью научной сети ICESS (Integrated Computing Environment for Scientific Simulation). BlueICE построен на основе SMP модулей IBM P5 575 с процессорами Power5+ 1,9 ГГц. Система имеет 4 Тбайт оперативной памяти и хранилище данных емкостью 150 Тбайт. Расчетная пиковая производительность суперкомпьютера составляет 12 TFLOPS. Эта система применяется для моделирования таких сложных явлений как турбулентность, состояние океанических вод, региональное изменение климата, что, наконец, позволит приблизиться к созданию точных кратко-среднесрочных прогнозов погоды.

В очередную редакцию TOP500 в 2007 году попали четыре российских системы и самый мощный из них — «СКИФ Cyberia» с пиковой производительностью 12 TFLOPS, работающий в Томском государственном университете. Комплексный экологический мониторинг атмосферы и гидросферы, контроль за разливом рек, распространением пожаров и эпидемий, рациональное использование лесных и минеральных ресурсов, новые конкурентоспособные методы разведки нефтегазовых месторождений, восстановление загрязненных почв, проектирование ракетно-космической техники и безопасного шахтного оборудования, создание новых видов ракетного топлива и сверхтвердых покрытий с помощью нанотехнологий — часть задач, которые ученые ТГУ решают с помощью «СКИФ Cyberia».

Ведет разработки и планирует внедрение собственных кластеров и SMP-систем ряд государственных университетов из Нижнего Новгорода, Санкт-Петербурга и других вузовских центров России. Традиционно сильные позиции суперкомпьютеров в российских научных учреждениях подкрепляются проникновением высокопроизводительных кластеров в производственную сферу. Так, суперкомпьютер мощностью 0,9 TFLOPS установлен в НПО «Сатурн», разработчике и производителе авиадвигателей, причем планируется на порядок увеличить производительность вычислительного центра.

Александр Семенов

после аварии происходит простая проверка целостности локальной файловой системы. Если попытаться запустить такую проверку в большой кластерной файловой системе, то она может занять несколько дней. Именно по этим причинам горячий интерес вызывает явление «silent corruption» («тихое искажение»). Проблема в том, что ряд искажений данных в больших файловых системах остается неизвестным до того времени, пока они не приводят к каким-то неприятностям. В подтверждение этого имеются результаты серии экспериментов с записью файлов на диск, чтением только что записанных файлов и сравнением прочитанных данных с теми, что были записаны. Запись, считывание и сравнение производились фоновым процессом, который запускался на 3500 машинах. Оказалось, что прочитанная с диска информация не совпадала с ранее записанным оригиналом примерно в тысяче случаев при общем объеме передачи из оперативной памяти на диск 41 Пбайт. Ошибки (несовпадения) были встречены на 170 машинах, а средний темп несовпадений составил 2-5 раз в день. В эксперименте не удалось найти строгую корреляцию с каким-то типом программного обеспечения и в большинстве случаев не просматривается корреляции с типом оборудования.

Источники таких ошибок интуитивно очевидны — в технических данных на многие дисководы указано, что вероятность ошибок составляет 10^{14} (или одна ошибка на 10^{14} бит) при выполнении операции ввода/вывода. Таким образом, возникает примерно одна ошибка (один бит неверный) при считывании около 11,4 Тбайт. Если учесть, что информация с поверхности диска считывается примерно со скоростью 900 Мбит/с, то в среднем можно ожидать одну ошибку за 30,9 часов непрерывного чтения. В реальных дисковых системах при использовании многих сотен дисководов это время может сократиться до минут. Кроме того, ошибка в считывании управляющей информации, описывающей расположение данных на диске, может привести к считыванию полностью искаженного фрагмента данных, объем которого кратен размеру стрипа в RAID (64 или 128 Кбайт). Естественно, имеются некоторые методы предотвращения искажения информации, в частности, отмечается устойчивость по отношению к ошибкам ввода/вывода файловой системы Sun ZFS, по сравнению, например, с системой XFS. Несколько кластеров уже используют систему ZFS, а ряд организаций планируют это в ближайшее время.

Наравне с надёжностью хранения данных такую же важность имеют задачи оценки производительности процессоров.

ИЗМЕРЕНИЯ БЫСТРОДЕЙСТВИЯ

Оценка быстродействия процессоров является нетривиальной задачей по многим причинам. Для компании производителя этот параметр — прямая реклама. Для потребителя может оказаться неприятным открытием, что новая машина, которая на каких-то тестах оценивается как более быстрая в сравнении со старой, тратит больше времени на выполнение конкретной задачи. Простая истина состоит в том, что лучший тест для любой машины — это конкретная задача.

Среди особенностей вычислений в физике высоких энергий следует отметить, что обработка больших пакетов данных обычно больше напоминает выполнение программ, в которых преобладают операции над числами в формате с фиксированной точкой. Например, наличие сложных и дорогих процессоров для работы с вещественной арифметикой не дадут здесь никакого эффекта. С другой стороны, нет никакой возможности установить свою типовую программу на все типы процессоров, поэтому одной из главных задач выполнения измерений является поиск подходящего теста, который даёт хорошую корреляцию с типовыми задачами. Если типовая задача выполняется на двух машинах с разницей в 10% производительности, то хотелось бы подобрать известный тест, который показывал бы близкое значение.

Ряд авторов отмечают, что существующие распространенные методы оценки (например, SPECint2000) не отражают изменение производительности на конкретном классе задач физики высоких энергий. Предлагается больше внимания обратить на тест SPECint2006, более соответствующий современной архитектуре микропроцессоров (многоядерность, большой кэш второго уровня на каждом ядре и т.п.).

Среди параметров, характеризующих производительность, используется еще один важный показатель — производительность машины на ватт потребляемой мощности. Это неудивительно, поскольку экономия хотя бы 1 ватт на машину при парке из 500 машин может вылиться в экономию более 2 КВт.

SCIENTIFIC LINUX

Состояние дел с дистрибутивом Scientific Linux (www.scientificlinux.org, www.linux-ink.ru) обсуждается весьма оживленно, поскольку этот дистрибутив Linux является фактическим стандартом для институтов физики высоких энергий. Дистрибутив в значительной степени является производной от RedHat, и сегодня доступна для тестов версия Scientific Linux 5.x. Практически все организации, вовлечённые в эксперименты по физике высоких энергий, планируют переход в ближайшее время на стабильную Scientific Linux 4.x. Дистрибутив Scientific Linux, естественно, отличается от других дистрибутивов ОС Linux. Среди основных требований к нему можно назвать следующие:

- высокая надежность работы базовых программных средств (6 месяцев и более без перезагрузки и без проблем с десятками-сотнями заданий каждый день является нормой);
- доступность и надежное функционирование популярных кластерных файловых систем и распределенных файловых систем (в частности, AFS);
- независимость от быстро меняющихся коммерческих интересов, что реализуется в виде финансирования разработчиков правительством США;
- приемлемый уровень поддержки, который, в основном, реализуется в виде дискуссионных списков.

ТЕНДЕНЦИИ

Хотелось бы обратить внимание на ряд важных тенденций в области вычислительных кластеров.

Аппаратные компоненты и коммуникации. На всех кластерах для приложений физики высоких энергий планируется увеличение объемов памяти и вычислительной мощности примерно на 30% в ближайшем году. Общее для всех вычислительных кластеров — быстрое обновление оборудования. В среднем за год устанавливается дополнительно примерно треть новых машин. Большая часть вновь приобретаемых базируется на Intel Xeon 5160 (Woodcrest, 3 ГГц) или близком по параметрам Opteron, что позволяет поддерживать возраст важных компонентов кластеров не более трех лет.

Закончилась гонка по частоте процессоров и началось соревнование по количеству ядер процессоров. Отмечается, что обычная многопоточность процессора не является эффективной для параллельных вычислений с MPI или OpenMP. Для получения серьезного ускорения необходимо использовать мультиядерные процессоры. Рост количества ядер в процессоре стимулировал рост популярности использования программного RAID, что предопределило, в свою очередь, преимущества реализации дисковой памяти в виде «дискового ящика» (например, Sun Fire X4500 — 48 x 500 Гбайт на шасси 4U) с использованием программного RAID 5(6).

Непросто установить большое количество машин (несколько сотен и более) — проблемы с охлаждением стоек, потребляемой мощностью, подготовкой помещений. В связи с этим предложено интересное решение размещения всего оборудования вычислительного кластера в стандартном транспортном контейнере. Если требуется большой кластер, то можно поставить еще контейнеры, начиненные компьютерным оборудованием.

Внутренние сети вокруг кластеров имеют коммуникации на основе 10G Ethernet. Внешние каналы, как правило, имеют емкость 10 Гбит и более, причем емкость внешних каналов связи наращивается быстро — по ряду оценок к 2013 году ожидается достижение емкости около 1 Тбит/с для каналов между кластерами на разных континентах.

Общесистемное программное обеспечение. На большинстве кластеров физики переходят к кластерным файловым системам и прекращают использование NFS в качестве основной файловой системы. В качестве кластерных файловых систем используются, как правило, GPFS или Lustre. Несколько кластеров используют PANASAS, обсуждается переход от этой файловой системы к другим. Почти все кластеры работают под управлением ОС Scientific Linux. Многие организации используют версию 3.x, но все планируют переход к Scientific Linux 4.x. Пакетные системы выполнения вычислительных заданий различны: используются PBS и его производные, а также LSF, Condor, SGE. ■

Андрей Е. Шевель (Andrey.Shevel@pnpi.spb.ru) — заведующий отделом вычислительных систем Института ядерной физики (Санкт-Петербург).