Динамические виртуализованные вычислительные кластеры для ФВЭ в ИЯФ СО РАН

А. М. Сухарев

Институт ядерной физики им. Г. И. Будкера Сибирского отделения Российской академии наук,
Россия, 630090, г. Новосибирск, проспект Академика Лаврентьева, д. 11

Новосибирский государственный университет,
Россия, 630090, г. Новосибирск, ул. Пирогова, д. 2.

E-mail: A.M.Suharev@inp.nsk.su

Несколько экспериментальных групп Института ядерной физики им. Будкера СО РАН принимают участие в локальных и международных проектах физики высоких энергий (ФВЭ). Их требования к вычислительным ресурсам и вычислительной среде могут сильно отличаться, часто являясь просто несовместимыми.

С другой стороны, в непосредственной близости от института имеются вычислительные центры, предоставляющие учёным доступ к суперкомпьютерам, и условия работы в них тоже различны.

Концепция динамических виртуализованных вычислительных кластеров позволила интегрировать эти внешние ресурсы в удобном и привычном для каждой пользовательской группы виде.

Динамический виртуализованный вычислительный кластер представляет собой набор виртуальных машин, выполняющихся как обычные пользовательские задания в удалённых вычислительных центрах и обеспечивающих нужную вычислительную среду для пользователей локальных групп ИЯФ. Виртуальные машины запускаются и останавливаются автоматически по мере появления и завершения пользовательских заданий.

Динамические виртуализованные вычислительные кластеры успешно применяются в ИЯФ СО РАН с 2011 года. Мы готовы как к созданию новых пользовательских групп, так и к расширению доступных вычислительных ресурсов.

Ключевые слова: вычисления для физики высоких энергий, виртуализация

Введение

Институт ядерной физики Сибирского отделения Российской академии наук ведёт эксперименты с детекторами КМД-3 и СНД на e^+e^--коллайдере ВЭПП-2000 и с детектором КЕДР на e^+e^--коллайдере ВЭПП-4М. Группы физиков ИЯФ участвуют в международных коллаборациях ATLAS, CMS и LHCB на Большом адронном коллайдере (ЦЕРН), Belle II в KEK (Япония) и других. В институте разрабатывается проект нового e^+e^--коллайдера с высокой светимостью и детектора для него.

Все эти группы требуют больших объёмов вычислительных ресурсов для обработки экспериментальных данных и моделирования изучаемых физических процессов.

В непосредственной близости от ИЯФ СО РАН в новосибирском Академгородке располагаются Информационно-вычислительный центр Новосибирского государственного университета (ИВТ НГУ, NUSC) [веб-сайт ИВЦ НГУ] и Сибирский суперкомпьютерный центр Института вычислительной математики и математической геофизики СО РАН (ССКЦ СО РАН, SSCC) [веб-сайт ССКЦ]. Развёрнутая в Академгородке суперкомпьютерная сеть Новосибирского научного центра (NSC/SCN) соединяет их с ИЯФ. Сеть построена по топологии «звезда» с центром в Институте вычислительных технологий СО РАН (ИВТ СО РАН) [веб-сайт ИВТ], имеет пропускную способность 10 Гбит/с, и изолирована от сетей общего назначения. Вычислительная установка общего назначения ИЯФ (BINP/GCF) является шлюзом в NSC/SCN, а также имеет собственный вычислительный кластер и предоставляет тем группам, которым это требуется, виртуальные машины (ВМ) для интерактивной работы пользователей.

Общая схема сети, вычислительных ресурсов и пользовательских групп ИЯФ показана на рис. 1.

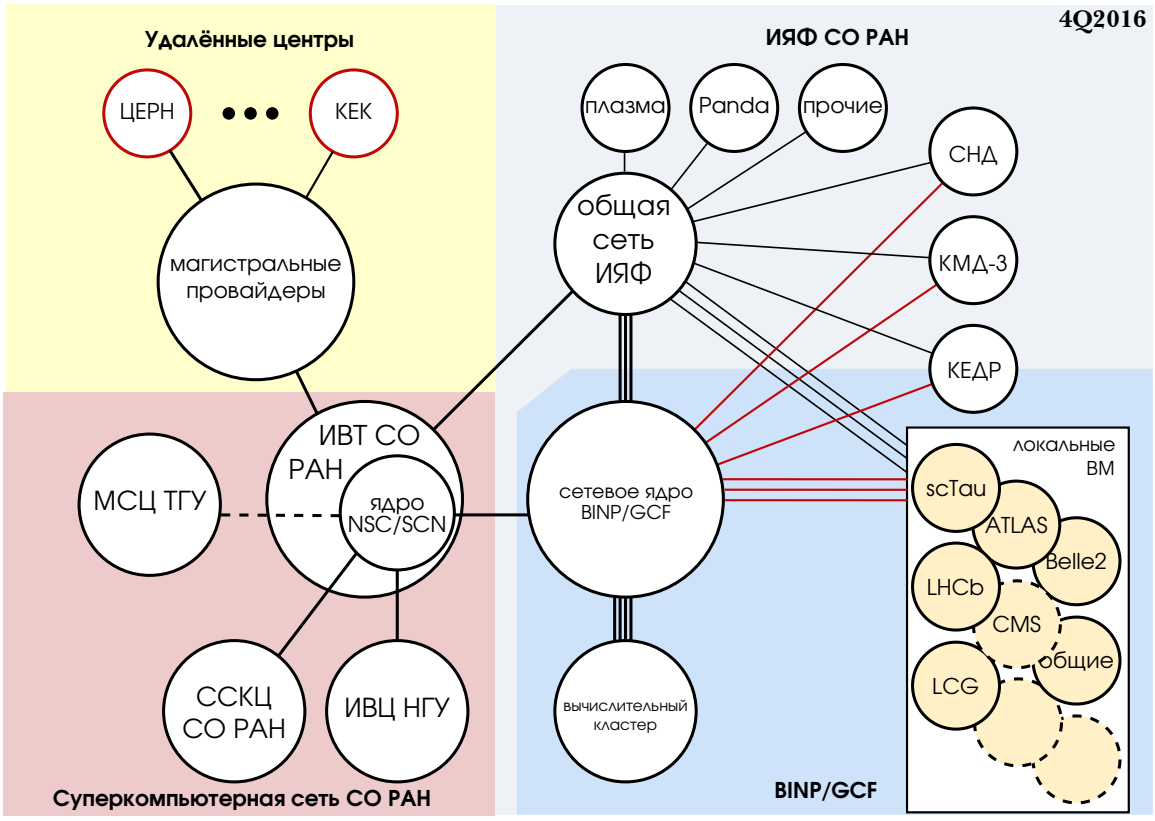

Рис. 1. Пользовательские группы ИЯФ СО РАН и доступные им вычислительные ресурсы.

Вычислительные задачи ФВЭ и их запуск в суперкомпьютерных центрах

Программное обеспечение для обработки данных и моделирования современных экспериментов физики высоких энергий, как правило, представляет собой однопоточные программы, которые могут выполняться параллельно на независимых наборах данных; в этом смысле специальной параллелизации программ не требуется.

Каждая экспериментальная группа создаёт своё программное обеспечение самостоятельно, с учётом особенностей своей установки и доступных человеческих ресурсов. Это приводит к большому разнообразию версий поддерживаемых операционных систем, версий стандартных библиотек подпрограмм для ФВЭ, используемых систем управления заданиями — то есть вычислительных сред в целом.

С другой стороны, вычислительные центры (ВЦ) общего назначения имеют свои собственные вычислительные среды, которые, к тому же, часто привязаны к конкретному поставщику оборудования.

Очень привлекательной возможностью для работы в этих условиях выглядит виртуализация. Запуская виртуальные машины с вычислительной средой экспериментальной группы внутри обычного пользовательского задания в удалённом ВЦ, можно сохранить традиционный сценарий работы членов каждой группы, изолировать друг от друга вычислительные среды групп и ВЦ, уменьшить трудозатраты системных администраторов.

Развёртывание такой виртуализованной инфраструктуры включает в себя

- создание и локальное тестирование базового образа виртуальных машин, которые будут предоставлять конкретной экспериментальной группе её стандартную вычислительную среду,
- установку образа и набора скриптов запуска и интеграции систем управления заданиями в удалённом вычислительном центре,
- настройку сети для обеспечения доступа ВМ к данным,
- запуск локального сервиса интеграции систем управления заданиями.

Таким образом формируется динамический виртуализованный вычислительный кластер (ДВВК), в котором ВМ автоматически запускаются в удалённых ВЦ при появлении локальных пользовательских заданий и останавливаются по их выполнении, освобождая вычислительные ресурсы.

В качестве системы виртуализации выбрана KVM [веб-сайт KVM]. Она присутствует по умолчанию в современных дистрибутивах ОС Linux, не требует использования специально модифицированного ядра ОС и демонстрирует хорошую стабильность работы.

Образы дисков ВМ располагаются в файловой системе удалённого ВЦ и используются в режиме «snapshot», делая возможным запуск любого количества ВМ с одного неизменного мастер-образа. Входные и выходные данные размещаются на серверах в ИЯФ и доступны через файловую систему NFS. Если ВМ требуется связь с сетью Интернет, она идёт через BINP/GCF. Для старта ВМ в очередь системы управления заданиями удалённого ВЦ ставится скрипт, который, попадая на конкретный вычислительный узел, запускает требуемое количество процессов ВМ с заданными параметрами, контролирует их загрузку, при необходимости отправляет им сигнал завершения.

Сервис интеграции систем управления заданиями выполняется на стороне локальной системы, следит за появлением новых пользовательских заданий, связывается с удалёнными ВЦ по

протоколу ssh, чтобы сделать запрос на запуск новых ВМ, и отправляет команду на выключение ВМ, когда все задания выполнены.

Примеры использования ДВВК

КЕДР. [Anashin, 2013] Вычислительная среда эксперимента КЕДР около 10 лет назад была зафиксирована на ОС Scientific Linux CERN 3 с архитектурой i386 и системе управления заданиями Sun Grid Engine 6.2. Она стала пилотной для развёртывания первого динамического виртуализованного вычислительного кластера, и с 2011 года подавляющее большинство заданий физического анализа и моделирования выполняется именно на нём.

ATLAS. ATLAS [Aad, 2008] — большая коллаборация, активно использующая для своих вычислительных потребностей в первую очередь GRID-ресурсы. Тем не менее, локальная система управления заданиями и «собственный» вычислительный кластер иногда оказываются более удобными для физиков. С 2012 года группа ATLAS в ИЯФ пользуется собственным ДВВК для физического анализа.

Belle II. Группа ИЯФ участвует в коллаборации Belle II, в том числе и вычислениях для неё [Krokovny, 2015]. В 2014 году создан ДВВК, принимающий задания, централизованно распространяемые через систему DIRAC [Tsaregorodtsev, 2014]. Суммарный вклад кластера в вычислительные ресурсы Belle II составил около 3%.

LHCB. В 2016 году развёрнут ДВВК для группы LHCB [Augusto Alves, 2008], работающий по схеме, аналогичной Belle II. С учётом имевшегося опыта время развёртывания кластера было минимальным. Суммарный вклад в вычислительные ресурсы LHCB составил около 0.3%.

Заключение

Разработанное в ИЯФ СО РАН программное обеспечение позволяет формировать динамические виртуализованные вычислительные кластеры на основе географически распределённых ресурсов. Эти кластеры успешно применяются для решения задач локальных экспериментальных групп ИЯФ СО РАН и для обработки заданий международных коллабораций ФВЭ, предоставляя каждой группе пользователей необходимую ей вычислительную среду и изолируя её от вычислительных сред удалённых ВЦ.

ДВВК могут легко развёртываться для новых групп пользователей и расширяться при подключении к новым удалённым вычислительным центрам. В наших ближайших планах — подключение к суперкомпьютеру СКИФ Cyberia в Межрегиональном супервычислительном центре Томского государственного университета (МСЦ ТГУ).

Работа выполнена при участии специалистов ИВЦ НГУ, ССКЦ СО РАН, ИВТ СО РАН и с использованием вычислительных ресурсов ИВЦ НГУ и ССКЦ СО РАН.

Список литературы

Aad G. et al. // The ATLAS Experiment at the CERN Large Hadron Collider. — Journal of Instrumentation. — 2008. — Vol. 3 No. 08.

Anashin V. V. et al. // The KEDR detector. — Physics of particles and nuclei. — 2013. — Vol. 44 No. 4. — P. 657–702.

Augusto Alves A., Jr et al.// The LHCb Detector at the LHC. — Journal of Instrumentation. — 2008. — Vol. 3 No. 08.

Krokovny P. // Belle II distributing computing. — Journal of Physics: Conference Series. — 2015. — Vol. 608 No. 1.

Tsaregorodtsev A. and the Dirac Project // DIRAC Distributed Computing Services. — Journal of Physics: Conference Series. — 2014. — Vol. 513 No. 3.
ICT web site [Electronic resource]: http://www.ict.nsc.ru
NUSC web site [Electronic resource]: http://nusc.nsu.ru
SSCC web site [Electronic resource]: http://www2.sscc.ru
KVM project web site [Electronic resource]: http://www.linux-kvm.org

Dynamical virtualized computing clusters for HEP at Budker INP

A. M. Sukharev

Budker Institute of Nuclear Physics of Siberian Branch Russian Academy of Sciences,
11, Lavrentiev av., Novosibirsk, 630090 Russia

Novosibirsk State University,
2, Pirogov st., Novosibirsk, 630090 Russia

E-mail: A.M.Suharev@inp.nsk.su

There are several experimental groups at Budker Institute of Nuclear Physics participating in local and international high energy physics (HEP) projects. Their requirements on computing resources and environments vary widely, often being incompatible.

On the other hand, there are several computing sites nearby the institute, providing supercomputer resources for academic users, each site having its own specific setups.

The dynamical virtualized computing cluster concept allowed to integrate these remote resources for Budker INP users in a convenient for each users group manner.

Such a cluster is basically a set of virtual machines running like conventional user jobs at remote computing sites, virtual machines providing local Budker INP user group with its specific computing environment. Virtual machines start and stop when user jobs arrive and terminate.

The dynamical virtualized computing clusters are successfully used in the Budker INP since 2011, and we look forward to add more user groups and computing resources.

Keywords: computing for high energy physics, virtualization