ГРИД-САЙТ МОРСКОГО ГИДРОФИЗИЧЕСКОГО ИНСТИТУТА НАН УКРАИНЫ

Д.В. Бородин, В.В. Фомин, В.А. Иванов

Морской гидрофизический институт
НАН Украины
г. Севастополь, ул. Капитанская, 2
E-mail: borodin112@gmail.com

Дано описание структуры грид-сайта Морского гидрофизического института НАН Украины. Приведены результаты тестирования производительности вычислительной системы. Описан общий подходы использования грид-технологий при решении задач численного моделирования.

Введение. Вычислительный кластер (ВК) Морского гидрофизического института НАН Украины предоставляет вычислительные ресурсы для решения задач численного моделирования атмосферных процессов [1], ветровых волн [2, 3] и течений [4] в Азово-Черноморском регионе. Нередко возникают условия, при которых данные задачи требуют больших вычислительных мощностей, и, в виду ограниченности ресурсов ВК, простаивают в очереди планировщика задач. Также нередки и обратные ситуации, когда ресурсы ВК простаивают из-за отсутствия задач. Такая обстановка характерна для вычислительных кластеров средней производительности, количество которых среди имеющихся в распоряжении институтов Национальной академии наук Украины является подавляющим. Еще пару лет назад это было серьезной проблемой. Однако, благодаря тому, что в Украине, вслед за остальным миром, начали широко использоваться распределённые информационные системы, базирующиеся на грид-технологиях, данная проблема начала решаться.

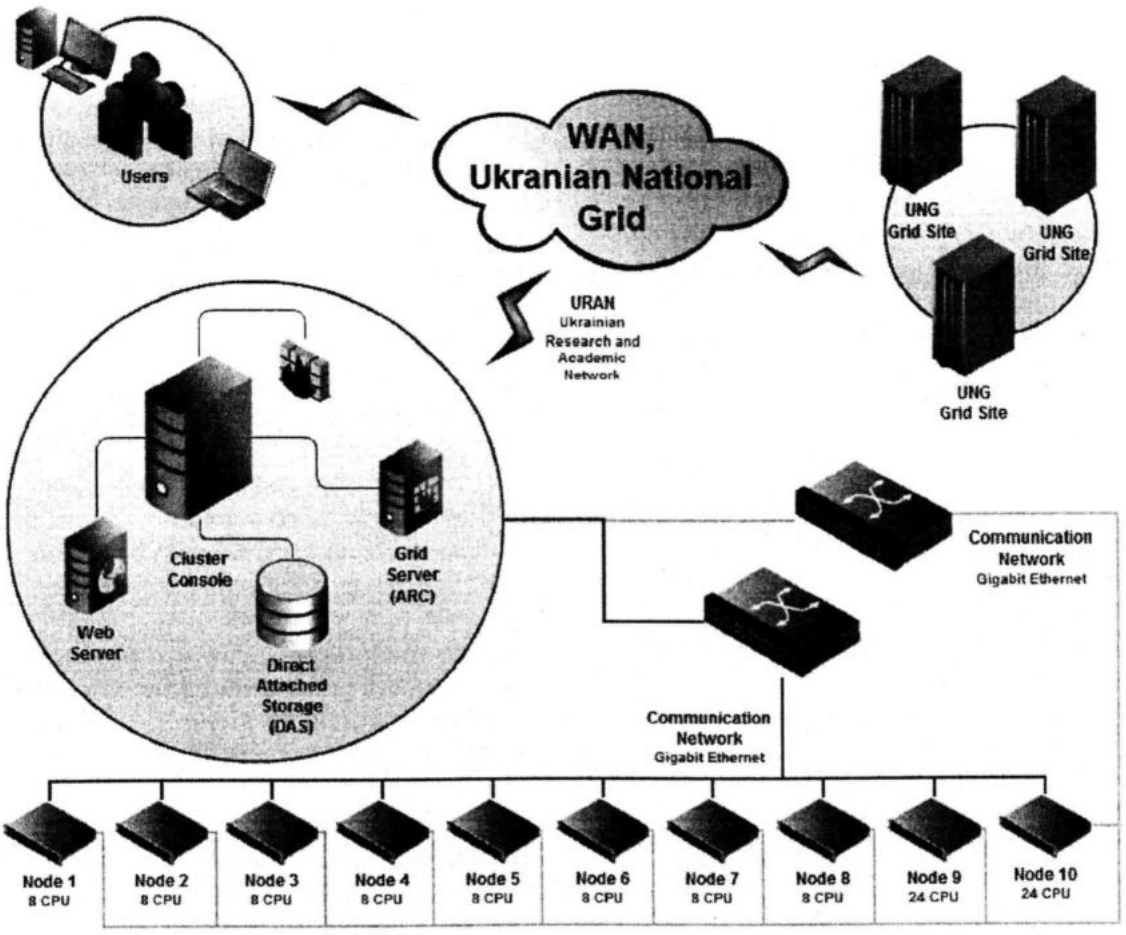
Грид (англ. grid – сеть) – это географически распределенная, согласованная, открытая и стандартизованная среда разделения вычислительных и информационных ресурсов [5]. Подобно тому, как планировщик задач на локальном кластере «ищет» и предоставляет свободные вычислительные ресурсы для вновь сформированных задач, грид «ищет» и предоставляет эти ресурсы по распределенным кластерам, находящимся в грид-среде. Развитие грид-технологий определило ближайшие перспективы многих прикладных направлений, в том числе науки о Земле, охватывающие широкий спектр тем, связанных с земной корой, атмосферой, океаном и их взаимодействием. Данные технологии предоставляют доступ к большим вычислительным ресурсам и хранилищам данных и, как следствие, большим вычислительным мощностям по сравнению с локальными кластерами.

В связи с этим, в рамках государственной программы развития и использования грид-технологий [6] в МГИ НАН Украины на основе имеющегося ВК был создан грид-сайт (ГС), представляющий собой совокупность ресурсов и сервисов грид-инфраструктуры, предоставляемую учреждением для коллективного использования. Ресурсы грид-сайта, аналогично ВК, включают вычислительные ресурсы, программное обеспечение, ресурсы хранения данных и сетевую инфраструктуру.

В данной статье представлена архитектура созданного грид-сайта, результаты тестирования его производительности, а также дано описание общего подхода использования грид-инфраструктуры на примере решения задач ретроспективного анализа морских течений и численного моделирования ветрового волнения.

Архитектура грид-сайта. Архитектура ГС (рис. 1) состоит из трех концептуальных, взаимодействующих составляющих: аппаратной, программной и грид-инфраструктурой. Аппаратная и программная составляющие, в свою очередь, в совокупности являются структурой кластерных систем типа Beowulf [7].

Структура Beowulf (вычислительный кластер) – вычислительная система, состоящая из широко распространённого аппаратного обеспечения, работающая под управлением операционной системы, распространяемой с исходными кодами, к коим относятся дистрибутивы семейства GNU/Linux и BSD. Главной особенностью вычислительных кластеров является их масштабируемость – возможность увеличения количества узлов системы с пропорциональным увеличением ее производительности.



Р и с. 1. Архитектура грид-сайта МГИ НАН Украины

Преимущества использования структуры Beowulf:

- стоимость построения существенно ниже стоимости суперкомпьютера;
- возможность увеличения производительности системы за счет масштабируемости;
- широкая распространённость и, как следствие, доступность аппаратного и программного обеспечения.

Структура Beowulf включает как минимум один управляющий узел, содержащий необходимые для функционирования системы сервисы. Остальные узлы являются рабочими (вычислительными) и используются непосредственно для проведения вычислений.

Аппаратная составляющая ГС состоит из 11-ти серверных станций с процессорами узлов линейки Intel Xeon серии «E», дискового массива VessRAID 1740s, коммуникационной и транспортной сетей, 2-х гигабитных управляемых коммутаторов D-Link DGS-1210-24 и Planet GSW-2404SF и 5-ти источников бесперебойного питания APC Smart.

Серверные станции являются узлами ГС, из которых один используется в качестве управляющего, а остальные 10-ть – вычислительные. Процессоры на узлах распределены следующим образом:

- на восьми вычислительных узлах – по два 4-ядерных процессора E5440 (с частотой 2.83 GHz) при 16-ти GB оперативной памяти на каждом;
- на двух вычислительных узлах – по два 6-ядерных процессора E5645 (с частотой 2.4 GHz), при 32-х GB оперативной памяти на каждом (данный процессор поддерживает технологию одновременной мультипоточности – «hyper-threading», за счет чего каждое его физическое ядро определяется операционной системой как два логических);
- и на управляющем узле – по два 4-ядерных процессора E5607 (с частотой 2.27 GHz) при 12-ти GB оперативной памяти.

Используемый дисковый массив – VessRAID 1740s является прямо подключенным хранилищем (DAS) и обеспечивает хранение результатов расчетов, резервных копий системы и прочих данных. Общая физическая емкость составляет 11 TB. Хранилище подключается к управляющему узлу посредством интерфейса SAS 4X, обеспечивающего скорость передачи данных до 12 GB/s. Управление хранилищем осуществляется как при помощи встроенного веб-интерфейса, так и посредством внешней LCD-панели.

Сети ГС базируются на полнодуплексной технологии Gigabit Ethernet, обеспечивающей скорость обмена данными по сети в 1 GB/s. Транспортная сеть используется для обмена данными различных клиент-серверных приложений и сервисов, а коммуникационная – для обмена данными при распределенных вычислениях.

Таким образом, аппаратно ГС характеризуется общим количеством ядер (CPU) – 120 единиц, общим количеством оперативной памяти (RAM) – 0.2 TB и хранилищем данных (DAS) – 11 TB.

Программная составляющая ГС представляет собой непосредственно операционную систему (ОС), а также необходимые для полноценного функционирования ГС сервисы и библиотеки.

В качестве операционной системы ГС использует дистрибутив GNU/Linux – Debian 6.0 для 64-хбитных систем [8], поскольку его отличительными чертами является жесткая политика по отношению к пакетам программного обеспечения и высокое качество выпускаемых версий этого обеспечения.

Под необходимыми сервисами понимаются: сервер удаленного доступа (ssh), сервер имен (dns), сервер сетевой файловой системы (nfs), сервер точного времени (ntp), сервер доступа к каталогам (ldap), система локального управления ресурсами (СЛУР), система мониторинга ресурсов, средства виртуализации, межсетевой экран, а также средства ограничения ресурсов. Функции, выполняемые данными сервисами:

- ssh-сервер – первостепенно необходим для обеспечения работы параллельной среды MPI [9], а именно – беспарольной авторизации на вычислительных узлах и обмена данными во время вычислений. Помимо этого, наличие данного сервиса на управляющем узле обеспечивает удаленный доступ к ГС с автономных рабочих станций;
- dns-сервер – обеспечивает работу внутреннего домена ГС, что позволяет прочим сервисам работать не с ip-адресами узлов, а с их доменными именами, что в свою очередь повышает масштабируемость системы в целом;
- nfs-сервер – обеспечивает общее файловое пространство узлам ГС, устраняя тем самым необходимость среды MPI дублировать исполняемые файлы на каждый вычислительный узел отдельно;
- ntp-сервер – обеспечивает синхронизацию локального времени ГС с серверами точного времени, что повышает корректность результата расчетов параллельных моделей, критичных к идентичности локального времени на вычислительных узлах;
- ldap-сервер – обеспечивает аутентификацию пользователей, использующих MPI, на вычислительных узлах. Данный подход устраняет необходимость создавать пользователей и группы на вычислительных узлах ГС, так как они считываются из базы данных на управляющем узле, что также увеличивает масштабируемость системы;
- СЛУР – обеспечивает распределение задач среди доступных вычислительных ресурсов, а также постановку их в очередь в случае отсутствия ресурсов. В качестве такой системы ГС использует некоммерческий дистрибутив семейства PBS – Torque [10]. Система включает сам сервер, планировщик задач и клиенты для вычислительных узлов;
- Система мониторинга ресурсов – позволяет наблюдать за статистикой использования ресурсов и историей вычислений ГС в реальном времени. В качестве такой системы ГС использует Ganglia – масштабируемую распределенную систему мониторинга кластеров;
- Средства виртуализации – обеспечивают функционирование на ГС виртуальных серверов, используемых для различных дистрибутивов промежуточного программного обеспечения грид-инфраструктуры;

• Межсетевой экран и средства ограничения ресурсов – обеспечивают сетевую и внутреннюю безопасность системы, соответственно. Первый осуществляет контроль и фильтрацию проходящих через ГС сетевых пакетов в соответствии с разработанными правилами безопасности, вторые – используют механизмы ограничений подключаемых модулей аутентификации PAM и системы квотирования доступного количества физической памяти Quota для снижения вероятности нанесения вреда системе пользователями ГС.

Под необходимыми библиотеками понимаются инструменты разработки, включающие компиляторы языков высокого уровня Си, Си++ и Фортрана, а также библиотеки MPI для параллельных вычислений. На ГС присутствуют как GNU компиляторы (gcc, g++, gfortran), обеспечивающие достаточную точность вещественных типов данных, так и некоммерческие оптимизирующие Intel компиляторы (icc, icpc, ifort), обеспечивающие повышенную их точность. Параллельная среда представлена дистрибутивами – Mpich2 и OpenMPI, собранными как с поддержкой GNU, так и Intel компиляторов.

Наличие данных библиотек позволяет пользователям ГС работать в готовой параллельной среде, а также дает возможность провести тестирование (сл. раздел) процессорной производительности полученной среды в рамках созданной структуры Beowulf.

Грид-инфраструктура представляет собой промежуточное программное обеспечение (ППО, grid middleware). На текущем этапе развития грид-технологий в Украине в качестве ППО используются gLite [11] и Nordugrid ARC [12], причем последнее – в подавляющем большинстве. Поэтому в данной статье грид-инфраструктура ГС рассмотрена с точки зрения промежуточного программного обеспечения ARC.

ARC представляет собой программное решение, позволяющие совместно использовать географически распределенные вычислительные ресурсы и ресурсы хранения данных, и позволяет этим ресурсам быть доступными через стандартные интерфейсы.

Базовая структура ARC следуют общепринятым подходам построения грид-среды [13]. ARC использует единую информационную систему для оптимизации доступа. Клиентское программное обеспечение может сделать запрос этой информационной системе, чтобы узнать, какие ресурсы имеются в наличии, чтобы потом запустить грид-задачи пользователя на наилучших имеющихся ресурсах. Для пользователей, все эти сложности скрыты: они просто формулируют свои задачи в виде сценария на специальном языке xRSL [14] и отправляют их в грид-среду, не подозревая, какие ресурсы будут использованы. Тексты сценариев, как правило, содержат: название задания; расположение исполняемой программы и ее аргументы; списки входных и выходных файлов; название файлов стандартного потока вывода и ошибок.

Сервисы ARC работают на виртуальном сервере управляющего узла ВК, среди них:

• A-REX – грид-менеджер, осуществляет запуск заданий локально, контролирует процесс выполнения;

• GridFTP Server – осуществляет прием заданий, создает для каждого задания отдельную директорию на время его выполнения;

• ARIS – информационный агент, собирающий и передающий данные о состоянии ресурсов в грид-среду.

Общая схема запуска задания в среде ARC с точки зрения клиента такова: сценарий задания в качестве аргумента передается команде пользовательского интерфейса – ngsub; заданию присваивается ссылка (JobID), по которой отслеживается его текущий статус командой пользовательского интерфейса – ngstat; по завершению задания скачивается результат командой пользовательского интерфейса – ngget.

Производительность грид-сайта. Тестирование ГС проводилось набором тестов – HPL Benchmark, который применяется при формировании списка 500 наиболее производительных вычислительных систем по всему миру [15].

В результате проведения серии тестовых расчетов с большим межпроцессорным взаимодействием (исключая

управляющий узел) была достигнута максимальная производительность:

$R_{max} \approx 0.6$ TFlops.

Это соответствует 0.6×10^{12} операций с плавающей точкой в секунду. Пиковая (теоретически максимальная) же производительность рассчитывается по формуле:

$R_{peak} = R_{peak\ узла} \times N = f \times n \times k \times N$,

где N – количество вычислительных узлов; f – тактовая частота процессора; n – количество операций с плавающей запятой, выполняемых за один такт; k – общее количество процессорных ядер в вычислительном узле.

Исходя из представленной аппаратной конфигурации (исключая управляющий узел), а также, учитывая, что для линейки Intel Xeon серии «E» $n = 4$ Flop/s, имеем пиковую производительность:

$R_{peak} = 0.95$ TFlops.

Таким образом, коэффициента полезного действия системы, рассчитываемый по формуле:

$КПД = R_{max} \times 100 / R_{peak}$,

составляет 63 %. КПД показывает, насколько эффективна вычислительная система [16] и в значительной степени определяется (при одинаковых характеристиках вычислительных узлов) коммуникационной средой и объемом оперативной памяти этой системы. Так как аппаратная структура вычислительных узлов гетерогенна, а коммуникационная среда (Gigabit Ethernet) обеспечивает недостаточно высокую скорость обмена данных, можно считать, что полученное значение (63 %) соответствует действительности, но, несмотря на это – попадает в интервал полезного КПД для суперкомпьютеров и вычислительных кластеров (60 – 83 %).

Использования грида в задачах численного моделирования. Стоит сразу отметить, что грид не является технологией параллельных вычислений. Формально, в грид-среде осуществляется запуск определенного количества отдельных задач на территориально распределенные ресурсы. Следовательно, эффективно решаться в грид-среде могут те задачи, которые, в свою очередь, могут быть разбиты на группу подзадач, не обменивающихся между собой данными (иными словами, вычисления для каждой такой подзадачи выполняются независимо).

Рассмотрим некоторые задачи, из тех, что локально обслуживаются вычислительным кластером МГИ:

1. Многовариантные расчеты полей ветровых волн. Простейший пример многовариантных расчетов ветровых волн – расчеты стационарных полей волнения для большого количества градаций скорости и направления постоянного по времени и однородного по пространству приводного ветра. Для определенных значений скорости и направления ветра задачи полностью независимы, решать их можно в любом порядке;

2. Ретроспективный анализ ветрового волнения. Ретроспективный анализ ветрового волнения применяется для получения статистических характеристик и климатических тенденций полей волнения за достаточно большой период времени. В этом случае есть возможность разбить исходный временной интервал на ряд вспомогательных перекрывающихся интервалов [17], и выполнять расчеты независимо на каждом из них.

3. Ансамблевый метод прогноза ветрового волнения. Прогноз, составленный на основе усреднения по ансамблю, обеспечивает в среднем более высокое качество оценки, чем прогноз, рассчитанный при одном детерминированном поле ветра. Расчеты для каждого участника ансамбля полностью независимы.

Задачи 1 и 3 – можно разбить на независимые подзадачи, каждая со своими входными параметрами, но общим временным интервалом. Задачу 2 можно разбить на независимые подзадачи, осуществив декомпозицию временного интервала.

Вообще говоря, эти задачи, с точки зрения технической реализации, ничем не отличаются. Для любой из них необходимо будет составить набор xRSL-сценариев грид-заданий, соответствующих ее подзадачам. Основа этих сценариев будет одна: перечисление входных файлов (исполняемый файл численной модели, файлы инициализации модели, содержащие необходимые значения параметров и пр.), выходных файлов (файлов с результатов расчетов, файлов ошибок и

пр.), запрашиваемых ресурсов и т.д. Отличаться будет только содержание файлов инициализации модели, а именно – значения параметров.

Таким образом, общий подход использования грид-инфраструктуры для решения задач численного моделирования, допускающих декомпозицию по параметрам инициализации используемой модели следующий:

1. составление набора файлов инициализации модели;
2. генерация для каждого такого файла сценария на языке xRSL;
3. отправка сгенерированных сценариев в грид-очередь (ngsub);
4. мониторинг выполнения заданий (ngstat) и получение результата по завершению (ngget);
5. обработка результатов моделирования, если в ней есть необходимость.

Данные этапы просто автоматизировать, для чего достаточно использовать возможности командного интерпретатора и встроенные утилиты unix-систем (такие, как SED, AWK и пр.). В [17] рассмотрено программное решение, дающее возможность на основе численной модели морской циркуляции решать задачу ретроспективного анализа морских течений в грид-среде Nordugrid ARC, представляющее собой сценарий командного интерпретатора bash.

Заключение. Модернизация вычислительного кластера МГИ НАН Украины до грид-сайта в составе грид-сегмента Украинского Национально Грида расширила потенциальные возможности проведения численного моделирования благодаря задействованию дополнительных вычислительных ресурсов. Сделан очередной шаг в развитии информационных и грид-технологий в институте.

СПИСОК ЛИТЕРАТУРЫ

1. *http://www.wrf-model.org.*
2. *http://swan-model.sourceforge.net.*
3. *Полонский А.Б., Фомин В.В., Гармашов А.В.* Характеристики ветрового волнения Черного моря // Доповіді Національної Академії наук України. – 2011. – № 8. – С. 108 – 112.
4. *http://imedea.uib-csic.es/users/toni/sbpom.*
5. *Foster I., Kesselman C., Tuecke S.* The Anatomy of the Grid: Enabling Scalable Virtual Organizations // International Journal of High Performance Computing Applications – 2001. – 15(3). – P. 200 – 222.
6. *Распоряжение* Кабинета Министров Украины № 1421-р от 5 ноября 2008 года «об одобрении Концепции Государственной целевой научно-технической программы внедрения и использования грид-технологий на 2009 – 2013 г.»
7. *Sterling T.* Beowulf cluster computing with Linux // Massachusetts Institute of Technology. – 2002.
8. *http://www.debian.org.*
9. *Антонов А.С.* Параллельное программирование с использованием технологии MPI. - М.: Издательство Московского университета, 2004. – 72 с.
10. *http://www.adaptivecomputing.com/products/open-source/torque.*
11. *http://glite.cern.ch.*
12. *http://www.nordugrid.org/arc.*
13. *P. Eerola.* The NorduGrid architecture and tools // Computing in High Energy and Nuclear Physics, La Jolla, California 24 – 28 March 2003.
14. *http://nordugrid.org/documents/xrsl.pdf.*
15. *http://www.top500.org.*
16. *Шнитман В.* Современные высокопроизводительные компьютеры. // Информационно-аналитические материалы центра информационных технологий – http://www.citforum.ru/hardware/svk/contents.shtml.
17. *Бородин Д.В., Фомин В.В., Иванов В.А.* Об использовании грид-технологий в задачах реанализа морских течений // Системы контроля окружающей среды. – Севастополь: МГИ НАН Украины, 2011. – С. 128 – 131.