ПОСТРОЕНИЕ ВЫЧИСЛИТЕЛЬНОГО КЛАСТЕРА НА ОСНОВЕ PC В ИССЛЕДОВАТЕЛЬСКОМ ИНСТИТУТЕ

А.Е. Шевель [1]

Петербургский Институт Ядерной Физики
188350, Гатчина, Ленинградской обл.,
Россия

Аннотация. В статье описывается построение вычислительного кластера на основе PC в исследовательском институте [1]. Освещаются ряд вопросов связанных с построением такого типа кластеров.

ВВЕДЕНИЕ

В 1997 в Петербургском Институте Ядерной Физики была осознана необходимость очередной смены поколений централизованных вычислительных ресурсов. Предыдущая смена поколений централизованных вычислителей (от старых мэйнфреймов к UNIX-серверам) в институте имела место в период 1990-1994 годы.

Очевидными мотивами последней смены поколений явилось значительное моральное устаревание еще работающего вычислителя SGI Power Series, который имеет 6 процессоров MIPS R3000, 224 MB памяти. Машина работает под управлением операционной системы IRIX. Число пользователей - около 100. Примерно половина занимается расчетами, остальные - обработкой текстов, поиск информации в Интернет, обмен информацией, прочее.

1) e-mail:*shevel@pnpi.spb.ru*

СОВРЕМЕННЫЕ ВЫЧИСЛИТЕЛЬНЫЕ УСТАНОВКИ

Поскольку имевшиеся средства на модернизацию были весьма ограничены, то был предпринят интенсивный анализ современных решений по максимальному снижению удельной стоимости вычислительной мощности. Во многих конкретных научных проектах для получения необходимой вычислительной мощности используются недорогие компьютеры на базе Intel, PowerPC, AMD и других. Понятно, что недорогим может быть лишь то, что производится массовым тиражом. Отсюда проистекает интерес к таким платформам как Intel. Поскольку каждый компьютер весьма недорог, то можно купить много таких компьютеров и использовать их параллельно.

Среди интересных решений по построению вычислителей на основе PC можно отметить самую мощную вычислительную установку не сегодняшний день - это Sandia National Labs Albuquerque [2] суммарной производительностью 2.1 TFLOP. Общее число процессоров составляет 9472 (в 1997 году это были PentiumPro200). Интересные реализации вычислителей на основе PC были предприняты и в других местах: наиболее известная разработка в NASA под названием Beowulf [3], а также ряд менее крупных систем: кластер в университете Duke [5], кластер Avalon в LANL [6] - последний на основе 140 процессоров Alpha 533MHz 21164A под управлением ОС Linux, кластер Калифорнийского университета [4], атакже кластер NERSC [8].

Наибольшую известность получил экспериментальный кластер Beowulf, построенный в NASA [3]. Особенностью систем типа Beowulf является использование массовых оборудования, которое продается почти в любом канцелярском магазине плюс использование кластеров как параллельных вычислителей. В качестве операционной системы везде использовалась ОС Linux. В тех случаях когда использовались мультипроцессорные платы, то использовалось ядро SMP.

ВЫЧИСЛИТЕЛЬНЫЙ КЛАСТЕР В ФИЗИКЕ ВЫСОКИХ ЭНЕРГИЙ

В профессиональной среде под кластером часто имеют в виду довольно разные архитектуры. В данной статье под термином кластер имеется в виду вычислительная установка со следующими важными признаками:

- кластер состоит из нескольких вычислительных машин;
- кластер находится под единым менеджментом и на нем проводится единая политика использования ресурсов кластера;
- кластер представляется пользователю как единый ресурс; пользователь не обязан ничего знать о структуре кластера и может воспринимать его как одну вычислительнуюмашину.

В физике высоких энергий довольно давно (с середины 90-х) стали использоваться в качестве вычислителей небольшие персональные станции на основе PC. С того же времени операционная система Linux для таких станций стала фактическим стандартом. По мнению автора наибольшее влияние на молчаливое принятие Linux в качестве фактического стандарта явилась деятельность Виктора Балашова по переводу и поддержке CERN Program Library на платформе Linux.

Процессоры

Возможная производительность вычислительной системы можно оценить сравнивая данные процессоров. Мы пытаемся сравнивать одни из самых популярных и дешёвых процессоров Intel с наиболее быстродействующими процессорами DEC Alpha. Процессоры сравнивают разными способами. Сотрудники лаборатории NERSC предлагают свои оценки [22]. Особенность этих оценок состоит в независимости от фирм производителей процессоров. В таблице 1 приведена некоторая выжимка из представленных данных. Совершенно не обязательно, что на конкретных задачах конкретные процессоры демонстрируют именно такую производительность, как показано в таблице 1. Главное, что произведены сравнения в примерно одинаковых условиях, тем самым мы получаем некую основу для обсуждения.

Table 1. Некоторые сравнения процессоров

На основе процессора	Производительность в MFLOPs
Alpha-600	94.2
PII-Xeon/450	64.2
Alpha-433	64.1
PIII/450	59.9
T3E-900	58.1
PII/400	56.4
PII/400/(Fujitsu)	53.6
Alpha-333	51.4
Alpha-533	50.7
Celeron/400	49.5
Celeron-300A	40.8
PII/333	33.6
PII/266/(SDRAM)	33.4
PII/266/(DRAM)	28.9
PPro-200	26.6

Из таблицы 1 следует, что, например, Pentium II/266 отличается от одного из самых быстрых по общему мнению процессоров всего в три раза. Значит мы могли бы рассчитывать, что три Pentium II/266 могут выполнять примерно ту же работу, что один Alpha 600. Аналогичные выводы напрашиваются, если посмотреть данные корпораций выпускающие микропроцессоры, хотя они дают несколько иные цифры производительности в MFLOPs.

ПАРАЛЛЕЛИЗМ

Уместно заметить о каком виде параллелизма идёт речь в данной статье, поскольку здесь обсуждается централизованная вычислительная установка, предназначенная в основном для обслуживания задач в области физики высоких энергий. К основным особенностям таких задач относятся:

- невысокая степень алгоритмического параллелизма (заметную часть программного тела составляют операторы сравнения и условного ветвления), т.е. важны как операции с числами в плавающем формате, так и целочисленные вычисления;
- в большинстве случаев имеет место обработка или генерация статистически независимых объектов данных, именуемых событиями, которые можно обрабатывать независимо;
- заметный объём готовых к исполнению программ (десятки мегабайтов);
- относительно большой ввод/вывод (во время выполнения программ могут потребоваться сканировать файлы размером в несколько гигабайтов и более).

В связи с этим обсуждаются главным образом независимые задания по обработке данных, которые могут выполняться как на одном процессоре, так и на совершенно разных машинах. Иными словами требуется некоторый диспетчер, который сможет планировать выполнение независимых заданий на машинах вычислителя в соответствии с набором критериев. Задания могут быть как интерактивными так и не интерактивными. Обычным набором критериев являются уровень загрузки процессора, процент использования области свопирования, интенсивность ввода/вывода, наличие на конкретной машине конкретных программ и другие условия.

ДИСПЕТЧЕРИЗАЦИЯ ЗАДАНИЙ НА ВЫЧИСЛИТЕЛЬНОМ КЛАСТЕРЕ

Диспетчеризация заданий или пакетная обработка заданий на ЭВМ имеет более чем 30-летнюю историю. Большая часть проблем по построению очередей заданий, максимизации производительности вычислительной установки в целом были поставлены и, в основном, решены в 60-70 годах на машинах типа мейнфрейм. С переходом от культуры мейнфреймов к рабочим станциям, интерес к пакетной обработке был на некоторое время утерян. Он появился вновь, когда в конце 80-х многие стали обращать внимание на эффективность финансовых вложений в компьютерную инфраструктуру государственных организаций и коммерческих компаний.

Есть и особенности, которые отличают современные системы пакетной обработки от прежних систем. В прошлом, как правило, рассматривались одномашинные системы, даже если они были многопроцессорными. Сейчас, пакетные системы обработки заданий, которые часто называют системами балансировки загрузки машин, чаще обсуждаются в многомашинной гетерогенной среде. Не накладывается никаких условий на расположение машин в сети: машины могут располагаться как в одном сегменте сети и одном помещении, так и в разных сегментах сети и разных зданиях или на разных континентах.

В настоящее время около 10 подобных систем в среде UNIX имеют широкое распространение. Неплохая подборка ссылок на разные виды пакетных диспетчеров находится на странице [10]. Описания отдельных систем могут быть найдены в [9]. Часть систем имеют родственные связи друг с другом. На рисунке 1 показано примерное родовое дерево систем пакетной обработки.

Почти все упомянутые системы, если они имеются на платформе Linux, могут рассматриваться в качестве кандидатов системы балансировки нагрузки в прототипе вычислительного кластера. В связи с этим можно принять во внимание степень распространения и другие частные особенности, например, доступность системы пакетной обработки для конкретной организации. Могут рассматриваться несколько коммерческих систем, которые вполне устраивают по своим функциональным возможностям: CODINE [16], LSF [18]. Имеются и некоммерческие системы, например,

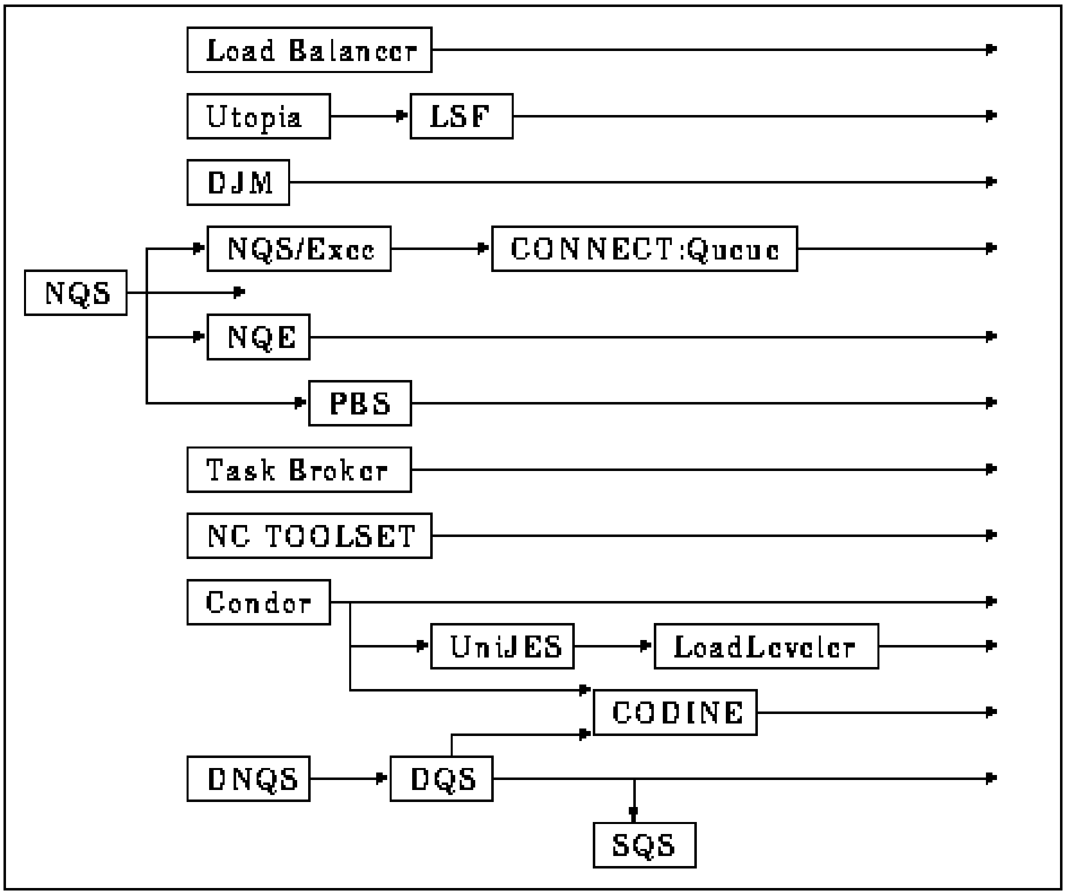

Рис. 1. Родовое дерево систем пакетной обработки.

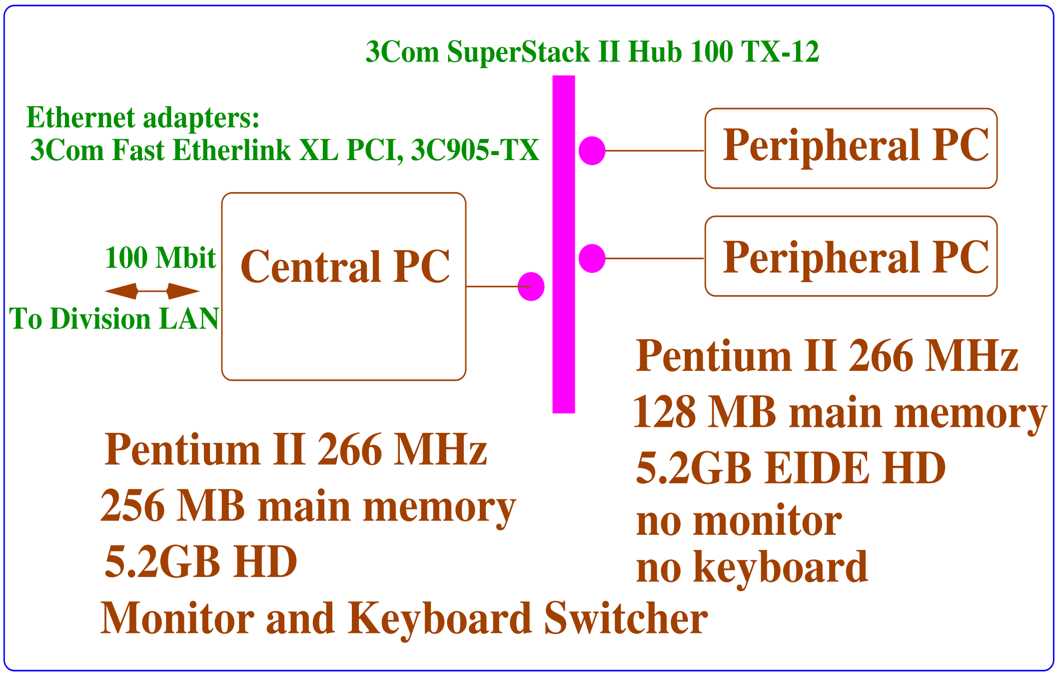

Рис. 2. Структура прототипа вычислительного кластера.

Condor [12], NQS [11], BPS [13] и другие.

Кроме систем пакетной обработки заданий, полезно отметить несколько подсистем смешанного назначения, которые могут быть использованы как для балансировки распределения потока заданий между различными узлами вычислительного кластера, так и для распараллеливания отдельных сложных программ: PVM [17], MPI [20,21].

ТЕХНИЧЕСКАЯ ОРГАНИЗАЦИЯ ПРОТОТИПА

Структура вычислительного кластера на основе PC приведена на рисунке 2.

При выборе компонентов исходили из минимума затрат. Однако минимум получали не за счёт наиболее дешёвых компонентов, а благодаря соответствующим архитектурным решениям. Так, в кластере на несколько компьютеров используется единственный монитор монитор и одна клавиатура.

Вычислительный кластер или ферма состоит в настоящее время из трёх компьютеров [7] на основе Pentium II/266 MHz. Один из компьютеров является основным или центральным, остальные – сателлиты. Сателлиты имеют лишь EIDE интерфейс к дискам по 5.2 GB. Только центральный узел имеет SCSI интерфейс. Как показано на рисунке 2 все компьютеры кластера объединены в локальную сеть кластера посредством концентратора 3COM 100 Mbit.

Как правило пользователи логируются на центральный компьютер (как правило с X-сессией). В обычных условиях не требуется логироваться на сателлитные компьютеры или даже знать об их существовании. Основной компьютер имеет 256 MB оперативной памяти, остальные по 128 MB. Все общие библиотеки программ, находятся на центральном узле кластера, на остальных машинах доступ к программам реализован посредством NFS.

Центральный компьютер играет роль рутера для сателлитов при связи с локальной сетью института.

Чтобы использовать один монитор и одну клавиатуру мы использовали механический переключатель, который позволяет переключать 4 компьютера, что достаточно для данного прототипа.

Пакетные задания (неинтерактивные и интерактивные) запускаются средствами CODINE. Основным критерием для распределения заданий по машинам в данном случае является степень загрузки процессора.

В настоящее время общее количество пользователей кластера превышает 50.

Общая стоимость использованного оборудования (в январе 1998) для прототипа кластера составила около $10K.

ТЕХНОЛОГИЯ ДАЛЬНЕЙШЕЙ МОДЕРНИЗАЦИИ И СТЕПЕНЬ МАСШТАБИРОВАНИЯ

Если ничего не менять в оснастке имеющегося вычислительного кластера, то его можно расширить до 4-х компьютеров (переключатель имеет четыре положения). Увеличив число переключателей и включив их каскадно можно легко убрать это ограничение. Так, использовав два переключателя, вы сможете обслужить 7 машин, а с тремя переключателями можно подключить 10 машин и т.д. Таким образом, можно относительно безболезненно наращивать число машин до 16. Использовав более дорогие многопозиционные переключатели клавиатуры и терминала можно довести число машин до ста и более.

Если использовать многопроцессорные конфигурации, то общее количество процессоров в такой системе можно сделать много больше числа машин. Реальный максимум на сегодняшний день составляет 4 процессора Pentium III/500 MHz. Иными словами, общее число процессоров относительно легко доводится до 400 штук. Следует отметить, что кластер с таким числом процессоров не будет дешёвым, однако вычислительный кластер из 200 процессоров Pentium III/500MHz можно построить из оборудования, которое продаётся "на каждом углу". Такие системы сейчас предлагают многие компании, например, /citextreme. Там же приведён пример 32-процессорной конфигурации (128 MB на процессор, двухпроцессорные материнские карты, по 11 GB дисководы) общей стоимостью около $60K.

Как всем хорошо известно, удельная стоимость процессоров падает довольно регулярно (примерно вдвое за 15 месяцев). Падают цены и на сетевое оборудование. Таким образом следует быть готовым к относительно частым модернизациям централизованных вычислительных ресурсов. Относительно редкие обновления вычислительного оборудования (скажем раз в пять лет) могут оказаться недешёвой затеей по двум причинам:

- поддержка морально и физически устаревшего оборудования как правило оказывается дорогостоящей;
- массовая смена всего имеющегося оборудования по прошествии нескольких лет может оказаться заметно дороже, чем вы можете себе позволить.

В случае с рассматриваемым кластером можно заметить, что модернизацию можно проводить очень постепенно (по одному компьютеру), по мере появления возможностей у организации, а не заменять сразу всё. Высвободившиеся в процессе модернизации машины можно использовать как настольные машины ещё ряд лет.

Таким образом, такая схема модернизации оборудования позволяет весьма эффективно использовать инвестиции в компьютерную инфраструктуру организации или предприятия.

СПИСОК ССЫЛОК

1. *Петербургский Институт Ядерной Физики* `http://www.pnpi.spb.ru/`.
2. *Кластер в лаборатории Sandia* `http://www.sandia.gov/ASCI/Red.htm`
3. *Кластер Beowulf* `http://www.beowulf-underground.org/underground/clemsonFAQ.html`
4. *Кластер Калифорнийского университета* `http://aeneas.ps.uci.edu/aeneas/`
5. *Кластер в университете Duke* `http://www.phy.duke.edu/brahma/`
6. *Кластер в Лос Аламос* `http://cnls.lanl.gov/avalon/`
7. *Проект BCFpc* `http://www.pnpi.spb.ru/pcfarm`.
8. *The NERSC PC Cluster Project* `http://www.nersc.gov/research/FTG/pcp/`
9. *Пакетный сервер* `http://www.halcyon.com/camellia/bjsfaq.htm`
10. *Ссылки на различные системы диспетчеризации пакетов* `http://www.cmpharm.ucsf.edu/~srp/batch/systems.html`
11. *Система NQS* `http://www.shef.ac.uk/uni/projects/nqs/`
12. *Система балансировки загрузки рабочих станций Condor* `http://www.cs.wisc.edu/condor/`
13. *Batch Portable System* `http://pbs.mrj.com/`
14. *Обзор пакетных диспетчеров (1995)* `http://www.npac.syr.edu/techreports/html/0700/abs-0748.html`
15. *Ещё один обзор пакетных диспетчеров (1996)* `http://www.crpc.rice.edu/NHSEreview/CMS/index.html`

16. *COmputing in DIstributed Network Environments - CODINE*
`http://www.genias.de/products/codine/description.html`
17. *Параллельные виртуальные машины - pvm*
`http://www.epm.ornl.gov/pvm/pvm_home.html`
18. *Load Sharing Facility – LSF*
`http://www.platform.com/platform/platform.nsf/webpage/LSF?OpenDocument`
19. *Каталог программного обеспечения для параллельных вычислений*
`http://www.nhse.org/rib/repositories/ptlib/catalog/`
20. *Message Passing Interface – MPI – Интерфейс Передачи Сообщений*
`http://mpi-softtech.com/company/what-is-mpi.html`
21. *Message Passing Interface – MPI*
`http://www.vcpc.univie.ac.at/information/language_standards/MPI/`
22. *Сравнение процессоров по производительности*
`http://www.nersc.gov/research/FTG/pcp/performance.html`
23. *Экстремальные машины*
`http://www.xtreme-machines.com/`
24. *32-процессорная конфигурация*
`http://www.xtreme-machines.com/x-configure.html`