Ордена Ленина

ИНСТИТУТ ПРИКЛАДНОЙ МАТЕМАТИКИ

имени М. В. Келдыша РАН

А. О. Лацис

МВС-900: вариант МВС-1000 на базе локальной сети Windows-машин.

Москва
2003

УДК 681.3.06

А. О. Лацис
МВС-900: вариант МВС-1000 на базе локальной сети Windows – машин.

АННОТАЦИЯ

Описывается способ реализации архитектуры многопроцессорного суперкомпьютера МВС-1000 на базе локальной сети персональных ЭВМ под управлением Microsoft Windows. Рассматриваются области применения описанного программного комплекса как в учебных, так и в производственных целях, а также в рамках технологий метакомпьютинга.

A. O. Lacis
MVS-900: an MVS-1000 Version Based on Windows Workstation LAN.

ABSTRACT

An MVS-1000 supercomputer architecture implementation based on an MS Windows workstation LAN is described. Possible applications of the proposed software, including training, production computations and metacomputing technologies, are discussed.

ОГЛАВЛЕНИЕ

1. Эволюция МВС-1000: недостающее звено.

Первые варианты МВС-1000 строились из специализированных, только для этой цели изготовленных, аппаратных компонент [1]. Технологический прогресс в области производства готовых компонент общего назначения (персональных компьютеров, оборудования локальных сетей) привел к повсеместному переходу разработчиков суперкомпьютеров на технологию ***кластеров выделенных рабочих станций***. По этой технологии, параллельный суперкомпьютер строится исключительно из готовых компонент общего назначения, то есть представляет собой обычную локальную сеть, только расположенную компактно, и специально **выделенную** для использования в качестве суперкомпьютера.

Современные варианты МВС-1000 [2] строятся именно по этой технологии: МВС-1000 на базе специализированного оборудования больше не выпускаются (хотя несколько выпущенных ранее установок по-прежнему используются).

С переходом на кластерную технологию понятие «МВС-1000» перестало ассоциироваться преимущественно с аппаратурой. Сегодня МВС-1000 – это, прежде всего, **архитектура операционной системы**, то есть конкретный облик машины с точки зрения пользователя, программиста и администратора. Отметим, что архитектура ОС коллективного пользования МВС-1000 сложилась, в основном, еще на до – кластерном этапе, и сама по себе мало зависит от используемой аппаратуры. С переходом на использование в качестве базы исключительно Linux – кластеров, она лишь несколько уточнилась и конкретизировалась. Именно этот, сегодняшний вариант архитектуры МВС-1000 мы и будем далее иметь в виду.

Парк используемых сегодня кластерных МВС-1000 включает МВС-1000М, расположенную в Межведомственном суперкомпьютерном центре РФ (768 процессоров), и примерно десяток установок по 16 – 32 процессора, расположенных в различных городах и организациях. Эти серийно выпускаемые установки называются МВС-1000/16 и МВС-1000/32,

соответственно. Все кластерные МВС-1000 оснащаются одной и той же ОС, ориентированной на применение в режиме коллективного пользования.

Архитектурное единство машин создает естественную иерархию применения: пользователи, имеющие доступ к небольшим установкам в составе региональных или корпоративных ВЦ, учатся, отлаживают программы и считают небольшие задачи на местных мощностях. Когда их потребности перерастают возможности местной МВС-1000/16, они обращаются к услугам Межведомственного суперкомпьютерного центра. Переучиваться или переписывать программы уже не приходится – местная и «центральная» машины архитектурно идентичны.

Такое двухуровневое построение парка МВС-1000, безусловно, является большим шагом вперед по сравнению с до – кластерным периодом, когда возможности доступа к суперкомпьютеру, в том числе для обучения и отладки, ограничивались одной – двумя «центральными» машинами. Однако, на современном этапе и двух уровней уже не достаточно.

Предлагается дополнить уже имеющуюся двухуровневую «линейку» моделей МВС-1000 младшей моделью: реализацией архитектуры МВС-1000 на базе локальной сети персональных ЭВМ под управлением MS Windows. Такой вариант МВС-1000 получил название МВС-900.

2. Основные черты архитектуры МВС-900.

2.1. Замечания о порядке дальнейшего изложения.

Предполагается, что читатель, в основном, знаком с архитектурой МВС-1000 как таковой. При необходимости сведения на эту тему можно получить, например, на сайте Межведомственного суперкомпьютерного центра РФ (www.jscc.ru), а также на сайте Испытательной лаборатории проекта МВС, ИПМ им. М. В. Келдыша РАН (www.kiam.ru). Мы же ограничимся рассмотрением реализации этой архитектуры на локальной сети Windows – машин (МВС-900), а также некоторых соображений о том, чем

именно МВС-900 могла бы быть полезна с точки зрения стратегии применения МВС-1000.

2.2. Реализация узла Linux – кластера на Windows – машине.

Отображение архитектуры МВС-1000 на локальную сеть персональных ЭВМ под управлением MS Windows осуществляется в два этапа:

- на Windows – машине реализуется узел, то есть запускается Linux,
- из полученных таким образом Linux – машин строится Linux – кластер.

В свою очередь, способов запуска Linux на машине под управлением Windows существует несколько (точнее, довольно много). Мы ограничимся описанием двух из них, выбранных по следующим принципам: простота установки, минимальность требований к базовой ОС (MS Windows), отсутствие необходимости переразмечать диск и переустанавливать базовую ОС, минимальность вероятности распространения ошибок администратора МВС-900 в базовую ОС.

Первый способ – наиболее простой, но и наименее удобный в использовании: вариантная начальная загрузка. Машина загружается либо под управлением MS Windows, либо под управлением Linux, и, соответственно, работает либо как универсальный компьютер, либо в составе МВС-900, но никогда – одновременно. Для удовлетворения перечисленным выше требованиям максимальной простоты и минимальной «травматичности» такой установки Linux, следует выбрать дистрибутив, который не требует отдельного, специально «под Linux» размеченного, раздела диска, а обеспечивает старт Linux непосредственно из Windows. При этом специфичная для Linux файловая система целиком размещается внутри специального файла (а не раздела диска), в то время как файловая система Windows монтируется из Linux и остается доступной. От Windows при таком способе установки Linux требуются лишь две вещи:

- наличие достаточного запаса свободного места на диске,

- умение загружаться «в режиме командной строки» (или, как его еще называют, «режиме DOS»). Последнее необходимо, поскольку начальные загрузчики Linux обычно требуют для своей работы именно этого режима.

В качестве дистрибутива Linux, специально собранного в расчете на такой режим использования, был выбран DragonLinux (http://sourceforge.net/projects/dragonlinux/) - очень удобный, небольшой дистрибутив на базе Slackware, предельно легко устанавливаемый и, что не менее важно, убираемый с машины путем простого удаления директории, безо всякого вмешательства в Windows. В МВС-900 этот дистрибутив принят за базу по причине минимальности и простоты обращения с ним. Отметим, что в «железных» версиях МВС-1000 используется RedHat, что создает некоторые различия в администрировании.

Мы уже отмечали выше, что вариантная загрузка – довольно неудобный способ реализации узла, и мы рекомендуем его нашим пользователям лишь в том случае, если на машине, по каким-либо причинам, установлена версия Windows на основе Windows-95, а не NT – например, Windows-95, Windows-98 или Windows ME. В этом случае нет проблем с режимом командной строки, а вариантная загрузка (либо Windows, либо Linux) альтернативы не имеет. Более штатной следует признать ситуацию, когда установлена версия Windows на базе технологии NT: Windows NT 4, Windows-2000 или Windows XP. В этом случае, мы рекомендуем гораздо более удобный, считающийся в МВС-900 основным,

Второй способ. Этот способ основан на применении технологии ***виртуальных машин***, а именно – на мониторе виртуальных машин фирмы VMware (www.vmware.com). При этом способе, монитор виртуальных машин устанавливается в Windows, и для выполнения Linux организуется виртуальная машина. Дистрибутив Linux используется тот же – DragonLinux. Преимуществом этого способа является одновременная работа Windows и Linux на одной машине, а также возможность конфигурировать по своему

усмотрению виртуальную машину. Например, на управляющей машине МВС-900 желательно иметь три разных сетевых интерфейса (два – строго обязательны). Виртуальная машина легко может быть сконфигурирована с тремя виртуальными сетевыми картами, в то время как по первому способу пришлось бы действительно вставлять в компьютер две или три сетевых карты.

При использовании виртуальной машины в качестве узла кластера, естественно, возникает вопрос об эффективности, прежде всего – о быстродействии процессора. Следует отметить, что технология виртуальных машин, прекрасно зарекомендовавшая себя еще в 80-е годы на компьютерах ряда ЕС, вовсе не заключается в программной эмуляции кода одной операционной системы под управлением другой. Виртуальная машина выполняется под управлением MS Windows на физической машине в качестве процесса, и, как и в любом другом процессе, не привилегированные команды (арифметические, переходов и им подобные) выполняются физическим процессором напрямую. Эмулируются лишь привилегированные команды, преимущественно связанные с вводом – выводом. Это позволяет предположить, что вычислительно – емкие приложения могут выполняться на виртуальной машине практически так же быстро, как на физической, если бы на нее непосредственно была установлена та же операционная система. Измерения это подтверждают – даже на весьма слабой (300 мегагерц) машине тест по умножению матриц замедлился, при переходе от физической машины к виртуальной, лишь на 7%. На более мощной машине относительное замедление будет еще меньше, поскольку «виртуализация» каждой привилегированной команды «стоит», грубо говоря, фиксированного числа команд физических, независимо от того, сколько времени эти команды выполняются. Замедление работы виртуального жесткого диска по сравнению с физическим также мало существенно, причем тем менее существенно, чем мощнее физическая машина.

Набор виртуального оборудования, которым оснащена та или иная виртуальная машина, задается при ее конфигурировании, средствами диспетчера виртуальных машин. Мы можем оснащать виртуальную машину эмулируемыми дисками (в действительности им соответствуют просто файлы), эмулируемыми сетевыми картами и пр.

Наконец, поскольку виртуальная машина совпадает по системе команд с физической, мы можем запустить на ней операционную систему – например, DragonLinux, получив так необходимый нам узел МВС-900. В отличие от описанного выше способа вариантной загрузки, DragonLinux выполняется в Windows на правах процесса. При этом, конечно, виртуальная машина конкурирует за вычислительную мощность с другими процессами, но общая работоспособность Windows сохраняется в полной мере. В те отрезки времени, когда в Windows отсутствуют другие вычислительно – емкие приложения, практически вся мощность физической машины автоматически оказывается в распоряжении машины виртуальной.

Например, если проводится занятие студенческой группы по параллельным вычислениям, физические машины, на которых реализованы узлы МВС-900, можно использовать одновременно как терминальные машины для доступа студентов к виртуальному вычислителю.

2.3. Реализация сетевой инфраструктуры МВС-900 на Windows – сети.

Сеть – главное «слабое место» МВС-900. Причин тому две.

Первая и главная – в том, что физическое сетевое оборудование, обычно используемое, скажем, в недорогом ВУЗовском компьютерном классе, плохо подходит для кластера. Чтобы получить приемлемое для хоть сколько-нибудь широкого круга задач сетевое быстродействие, даже в наиболее дешевых «железных» кластерах применяется коммутируемый Fast Ethernet, а в аппаратно реализованных вариантах МВС-1000 минимальная сетевая конфигурация – это две параллельных сети Fast Ethernet, коммутируемых независимо. Если в сети, на которой Вы устанавливаете

МВС-900, часть узлов подключена по 10-мегабитному кабелю, или предполагаемые для использования в качестве узлов 4 компьютера подключены через hub совместно с еще 20 машинами – толку от такой сети будет мало. Заведомо работоспособная конфигурация получится, если снабдить узлы – физически – дополнительными адаптерами Fast Ethernet, и соединить эти дополнительные адаптеры в отдельный сегмент локальной сети коммутатором. При нынешних ценах на сетевое оборудование, вряд ли затраты на такую доработку будут непосильны даже для самых стесненных в средствах пользователей. Впрочем, формальную работоспособность системы обеспечит, в принципе, даже модемное соединение – вопрос только в широте класса задач, для успешного распараллеливания которых такой производительности сети будет достаточно.

Вторая причина слабости сети – в том, что накладные расходы на ее виртуализацию (если узел реализован по второму варианту) довольно высоки. Пропускная способность «точка – точка» для TCP – соединения между виртуальными Linux – машинами, при отображении виртуальных сетевых адаптеров на отдельный физический сегмент сети, составляет всего 6 - 7.5 мегабайт в секунду на машинах с тактовой частотой 300-400 мегагерц (для физических машин эта величина достигает 11.5), и лишь при частоте процессора в 1 гигагерц и выше поднимается до 10 мегабайт в секунду. Латентность, в свою очередь, может достигать 400 микросекунд. Впрочем, опытный образец МВС-1000, изготовленной по до – кластерной технологии, имел примерно такую же пропускную способность канала, при гораздо большей латентности, и, тем не менее, успешно применяется в производственных расчетах до сих пор.

3. МВС-900 и проблемы стратегии применения МВС-1000.

В предыдущем разделе мы обсудили конкретный способ превращения локальной сети Windows – машин в виртуальный суперкомпьютер. Способ

это далеко не единственный, и, в некоторых отношениях, не самый эффективный. Почему же предлагается именно он? Какие именно проблемы в применении МВС-1000 (всех моделей и вариантов) можно надеяться решить путем использования именно МВС-900?

Проблемы обучения новых пользователей технологиям параллельных вычислений, с одной стороны, и доступа к централизованно предоставляемым, удаленным вычислительным установкам, с другой – общеизвестны. Состоят эти проблемы в том, что параллельная техника дорога, мало доступна и требует от пользователя знаний и навыков, отличных от тех, которыми он уже овладел на своем настольном компьютере. Общеизвестны и технологии преодоления этих проблем.

Если речь идет об **обучении**, можно воспользоваться имитатором параллельной машины на настольном компьютере, причем степень внешнего сходства (функционального) такого имитатора с настоящей машиной может быть очень высокой. При этом, конечно, ни о каком скоростном выигрыше от параллельного выполнения программ и речи быть не может, то есть предлагается средство исключительно для обучения и отладки.

Если же речь идет об упрощении и удешевлении доступа к **реальной** многопроцессорной вычислительной мощности, для небольших, вспомогательных и методических расчетов, то для отбора такой мощности из локальной сети Windows – машин существует масса программных технологий. Прежде всего приходят в голову реализации MPI и PVM, специально предназначенные для таких локальных сетей, то есть простейшие программные инструменты для кластеров невыделенных рабочих станций. Существуют и более современные программные системы, позволяющие использовать локальную сеть как суперкомпьютер исключительно в фоновом режиме, с разной степенью прозрачности для пользователей, работающих непосредственно на компьютерах такой сети.

Все эти технологии характеризуются, во-первых, существенным отличием порядка работы пользователя от «железных», аппаратно

реализованных, суперкомпьютеров, и во-вторых – заметной дополнительной нагрузкой на службу администрирования и поддержки используемой таким образом локальной сети.

Таким образом, в зависимости от того, хотим мы учиться, или же проводить вспомогательные и методические расчеты, мы можем воспользоваться либо реалистичным, но не настоящим макетом суперкомпьютера, либо суперкомпьютером почти настоящим, но неудобным, трудоемким в использовании и совсем не таким, в смысле порядка работы, как настоящие машины.

В действительности, опыт показывает, что эти две задачи – обучения параллельным вычислительным технологиям и их вспомогательного использования – разделять нельзя. Существует всего лишь одна задача – освоение всего технологического цикла применения суперкомпьютерной техники, от продумывания особенностей параллельной реализации алгоритма до привыкания к конкретному порядку нажатия на клавиши при запуске счета, включая осознание пользователем, на собственном опыте, того, действительно ли ему нужен для его приложений параллельный суперкомпьютер. Задача эта может быть решена только в комплексе. С этой точки зрения способность виртуального суперкомпьютера «понимать» те же команды, что и настоящая машина, не отделима от его способности так же, как и настоящая машина, давать скоростной выигрыш, зависящий от того, сколько сил и времени программист потратил на разработку параллельной программы.

Легко видеть, что МВС-900 – это техническое решение, ориентированное именно на такую технологическую комплексность. Мы не делим пользователей на тех, кто «еще учится», и тех, кто «уже считает», а программные инструменты – на отдельные учебные и отдельные производственные. Предлагается единый инструмент, и инструмент этот – настоящий суперкомпьютер, архитектурно идентичный МВС-1000, а не тренажер, лишь создающий иллюзию параллельных вычислений (и иллюзию

обучения им). Одновременно обеспечивается возможность плавного перехода с МВС-900 на малые, а с них – на средние и большие конфигурации «железных» МВС-1000.

Очень важно, с этой точки зрения, еще и то, что МВС-900 обладает очень высокой степенью замкнутости в смысле администрирования и сопровождения. Она не разделяет с локальной сетью, на которой она реализована, ни сетевых адресов, ни дисциплины и прав доступа. Фактически, МВС-900 администрируется совершенно отдельно, причем очень похоже на то, как администрируется настоящая МВС-1000, например, МВС-1000/16. Это очень важно, поскольку организации, планирующие приобретение малых конфигураций МВС-1000, получают возможность заранее подготовить не только пользователей, но и администратора, и на практике оценить сложность и трудоемкость сопровождения установки.

4. МВС-900 и метакомпьютинг.

Еще одна важная область применения МВС-900, на которой хотелось бы остановиться – это бурно развивающиеся в последнее время технологии метакомпьютинга. В идеале развитие этих технологий само по себе должно привести к решению тех проблем, которые мы рассматривали в предыдущем разделе. В наиболее общей постановке, метакомпьютинг предполагает динамическое и прозрачное для пользователя формирование виртуального параллельного компьютера из всевозможных доступных в Сети вычислительных ресурсов. Практически, на сегодняшний день, системы метакомпьютинга распадаются на несколько классов, решающих эту задачу разными способами. На одном полюсе находятся системы объединения небольшого числа очень мощных машин в единый, еще более мощный, виртуальный компьютер (Globus [3]), на другом – системы прозрачного отбора вычислительной мощности в сравнительно больших наборах маломощных компьютеров (Condor [4] и т. п.).

Системы первого класса ориентированы на выполнение классических параллельных программ, системы второго класса – в основном на массовый вариантный счет, причем на задачи с небольшим объемом обрабатываемых данных, но очень большим объемом самой обработки.

Промежуточное место занимают системы прозрачного отбора вычислительной мощности невыделенных локальных сетей, ориентированные на выполнение параллельных программ [5]. Теоретически такие системы могли бы быть полезны именно для решения поставленной нами задачи удешевления и упрощения доступа пользователей к параллельным вычислительным ресурсам. Однако, технологии этого класса пока не очень широко применяются, и тому есть довольно очевидное объяснение. Метакомпьютер, по определению, весьма сложен и трудоемок в поддержании и администрировании. Реальная совокупная себестоимость поддержания инфраструктуры метакомпьютера на мощностях, интенсивно используемых, в основном, для других целей, зачастую выше, чем себестоимость покупки кластера выделенных рабочих станций.

Однако, при использовании в качестве «материала» для создания метакомпьютера не отдельных рабочих станций, а виртуальных кластеров МВС-900, значительная часть проблем со сложностью и трудоемкостью администрирования и сопровождения решается автоматически.

Как мы уже говорили выше, администрирование МВС-900 практически полностью отделено от администрирования локальной сети, на которой она реализована. Все сложные трудоемкие действия по обеспечению безопасности, авторизации, настройке сетевых маршрутов, правил доступа и т. п. выполняются для физических и виртуальных машин независимо. Как следствие, установка «нижнего этажа» метакомпьютера не создает дополнительной нагрузки на администрирование и сопровождение физической вычислительной мощности. Виртуальный кластер, конечно, нуждается в администрировании, но администрировать кластер выделенных рабочих станций гораздо легче, чем метакомпьютер. И уж, конечно,

построение метакомпьютера из небольшого числа централизованно управляемых виртуальных кластеров – гораздо более обозримая задача, чем построение системы отбора вычислительной мощности из сети нескольких десятков независимых Windows – станций.

Немаловажно и то, что как пользователь виртуального кластера, так и разработчик «верхнего этажа» - собственно метакомпьютерного программного обеспечения – получают в свое распоряжение машины, не отличимые при сетевом доступе от «настоящих, железных». Так необходимый сегодня экспериментальный материал для обкатки технологий метакомпьютинга (да и параллельных вычислений вообще) оказывается дешевым в двух смыслах:

- во-первых, не требующим выделенной аппаратуры,
- во-вторых, что гораздо важнее, не требующим, в отличие от классических систем метакомпьютинга, дорогостоящих и трудоемких экспериментов с сетевыми настройками и системами авторизации на «живых», эксплуатируемых вычислительных мощностях.

На задаче комплексной «обкатки» технических решений в области метакомпьютинга хотелось бы остановиться чуть подробнее. Представляется, что проблема имеет, как минимум, два аспекта.

Первый, достаточно очевидный, состоит в том, что сами технологии метакомпьютинга нуждаются в широкомасштабных испытаниях с целью выбора наиболее подходящих. Как уже отмечалось, испытания эти на «живых» системах, как правило, неподъемно трудоемки.

Второй, ничуть не менее важный, аспект состоит в том, что успех в применении технологий метакомпьютинга подразумевает коренной пересмотр подхода к методикам распараллеливания. Для метакомпьютера характерны неоднородность коммуникационной среды и «узкие места» в коммуникациях, от которых современные пользователи, «избалованные» высокоскоростными сетями «железных» суперкомпьютеров, уже успели

отвыкнуть. Потребуется вернуться к применявшимся на заре параллельных вычислений методикам «щадящего» использования сети, как минимум – достичь четкого понимания того, насколько и при каких условиях те или иные параллельные алгоритмы являются коммуникационно чувствительными. Виртуальный кластер в этом отношении сам по себе является идеальным полигоном, поскольку при виртуализации основные накладные расходы ложатся именно на сеть, а не на быстродействие процессора.

Литература.

1. А. В. Забродин, В. К. Левин. Опыт разработки параллельных вычислительных технологий. Создание и развитие семейства МВС. Труды Всероссийской научной конференции «Высокопроизводительные системы и их приложения», Черноголовка, 2000.
2. Г. И. Савин, Б. М. Шабанов, А. В. Забродин, В. К. Левин, В. В. Каратанов, В. В. Корнеев, Г. С. Елизаров. Структура многопроцессорной вычислительной системы МВС-1000М. Труды Всероссийской научной конференции «Высокопроизводительные системы и их приложения», Черноголовка, 2000.
3. http://www.globus.org Сетевой ресурс.
4. http://www.cs.wisc.edu/condor Сетевой ресурс.
5. IBM grid computing - Grid Environments. http://www-1.ibm.com/grid/grid_environments.shtml Сетевой ресурс.