Использование технологии *VDI* для работы с большими сеточными моделями

П.И. Каледа, В.Д. Никоноров (АО "НИКИЭТ")

Для работы с большими сеточными моделями необходимо, чтобы рабочая станция пользователя обладала большим объемом оперативной памяти **(ОП)** и высокопроизводительной графической картой. Кроме того, мы выдвинули требование непосредственного подключения к файловой системе суперкомпьютера, чтобы избежать копирования данных между рабочей станцией для сеточных моделей и суперкомпьютером. При помощи технологии виртуализации рабочих мест (*Virtual Desktop Infrastructure* – ***VDI***) можно обеспечить удаленную работу с виртуальными рабочими местами, удовлетворяющими этим требованиям. В предлагаемой статье представлены описание стенда для тестирования данного подхода и результаты испытаний.

1. Введение

При выполнении задач численного моделирования процессов качество и размер расчетной сеточной модели непосредственно влияют на точность получаемых результатов. Размеры сеточных моделей постоянно увеличиваются в связи с повышающимися требованиями к точности. Современные размеры сеточных моделей таковы, что при использовании сеточных пре-/постпроцессоров, работающих с общей памятью (*ANSYS Meshing / ANSYS Icem CFD* и аналогов), объема оперативной памяти и графических ресурсов, доступных обычно на персональных компьютерах, становится недостаточно. По этой причине применяются дорогостоящие специализированные рабочие станции с высокопроизводительными графическими картами. Однако стоимость таких станций делает нецелесообразной их установку каждому пользователю, работающему с большими сеточными моделями.

С другой стороны, при использовании в работе с пре-/постпроцессором данных, полученных в результате расчетов на суперкомпьютере, необходимо иметь непосредственное подключение к системе хранения данных суперкомпьютера (в нашем случае – файловая система *Lustre* [1]), чтобы избежать копирования данных между суперкомпьютером и машиной, на которой происходит работа с сеточными моделями.

В совокупности эти два обстоятельства – высокая стоимость рабочей станции с достаточным количеством ресурсов и необходимость подключения к системе хранения данных суперкомпьютера – приводят к идее централизации ресурсов, требующихся для организации рабочих мест. Такой подход предусматривает наличие рядом с суперкомпьютером нескольких рабочих станций, подключенных к системе хранения данных суперкомпьютера, на которых пользователи могут осуществлять решение своих задач удаленно. При этом рабочие станции не закрепляются за конкретными пользователями, а являют собой некий общедоступный разделяемый ресурс.

2. Описание проблемы

При организации удаленного доступа к приложениям с повышенными требованиями к графической подсистеме основной проблемой становится доставка рабочего стола, на котором работает приложение, до пользователя. В случае, когда удаленное приложение использует *3D*-графику, оно, как правило, не может полноценно задействовать функционал *GPU*. Например, обработка *3D*-команд происходит на центральном процессоре сервера (центральный процессор может выполнять эту задачу, но делает это не так эффективно, как *GPU*), или приложение передает *3D*-команды на клиентское устройство [2] (объем информации, который приложению необходимо передавать для отображения картинки на экран, становится настолько внушительным, что отзывчивость этого удаленного рабочего стола будет низкой или очень низкой, какой бы протокол для удаленного доступа к рабочему столу ни использовался).

3. Подходы к решению проблемы

Для решения проблемы низкой отзывчивости необходимо исключить обработку *3D*-команд на центральном процессоре и передачу *3D*-команд от приложения через локальную сеть к *GPU* пользователя. Делается это путем прямой обработки *3D*-команд локальным *GPU* сервера и последующей отсылки растрового изображения, уже сформированного на *GPU*, на клиентское устройство.

В ходе исследования рынка авторами были рассмотрены открытая технология, реализующая этот функционал (связка перехватчика *3D*-команд *VirtualGL* и инструмента для удаленного доступа *TurboVNC*), и проприетарные решения: *VMware Horizon View* и *Citrix XenDesktop*. Два последних решения не просто предоставляют средства доставки пользователю удаленного рабочего стола, а сочетают такие средства с виртуализацией рабочих мест (*VDI*).

Решения с виртуализацией рабочих мест представляются более гибкими и эффективными в аспекте распределения ресурсов между пользователями, интеграции с существующей информационной инфраструктурой предприятия, организации управления доступом пользователей к рабочим

местам и выбора ПО, используемого на удаленных рабочих станциях. В отличие от них, связка *VirtualGL+TurboVNC* не имеет встроенных средств разделения ресурсов между несколькими пользователями, равно как и готовых средств интеграции с системой централизованной авторизации, используемой в ОАО НИКИЭТ (доменная сеть под управлением *Microsoft Active Directory*). Решения, совмещающие доставку приложений или рабочих столов пользователю с виртуализацией рабочих мест, как правило, не используют собственную подсистему аутентификации, а интегрируются в инфраструктуру предприятия – например, *Microsoft Active Directory*.

4. Выбор подхода

Решения по виртуализации рабочих мест (*VDI*) сегодня предлагают многие компании: *Vmware* [3], *Citrix* [4], *Microsoft* [5], *RedHat* [6] и другие, но классическая виртуализация рабочих мест, использующаяся уже достаточно давно, обычно не предполагает интенсивную нагрузку на графическую подсистему. Так, решение от *RedHat* вообще не поддерживает *3D*-графику, а решение от *Microsoft* плохо подходит для приложений, которые используют графический интерфейс *OpenGL*, так как ориентировано на интерфейс *DirectX*. Решение *Citrix XenServer* официально не поддерживает работу с сетью стандарта *InfiniBand* (которая необходима, в том числе, для доступа к данным), и, следовательно, не способно обеспечить многопользовательский высокоскоростной доступ к файловому хранилищу суперкомпьютера.

Итак, остается решение на базе ***VMware vSphere*** с доставкой при помощи *VMware Horizon View* или *Citrix XenDesktop*. В связи со сложностями, возникшими при настройке взаимодействия между *XenDesktop* и гипервизором *vSphere*, наш выбор был сделан в пользу ***Horizon View***. Виртуализация рабочих мест на платформе *vSphere* позволяет получить максимум от графического процессора *NVIDIA* и в то же время обеспечить быстрый доступ к файловому хранилищу суперкомпьютера. Также отметим, что данное решение интегрируется с *Atcive Directory,* что существенно упрощает задачу управления пользовательскими учетными записями.

5. Описание эксперимента

5.1. Аппаратное обеспечение для работы с графикой

С точки зрения плотности размещения виртуальных рабочих мест, специализированные решения для виртуализации *NVIDIA GRID K1* и *K2* [7] имеют преимущество перед классическими картами линейки *Quadro* для рабочих станций, поскольку карты *GRID* имеют на борту несколько (четыре для *GRID K1* и два для *GRID K2*) графических процессоров, управление каждым из которых можно передать отдельной виртуальной машине.

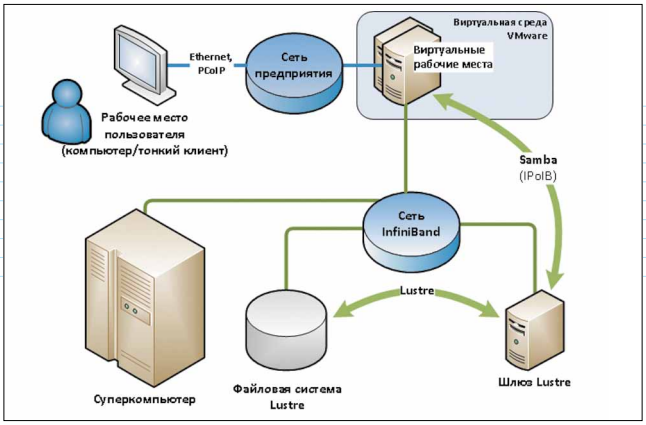

Рис. 1. Принципиальная схема решения

5.2. Общая схема протестированного решения

Принципиальная схема протестированного решения изображена на рис. 1. Виртуальное рабочее место доставляется пользователю по сети предприятия. Графический процессор *NVIDIA* предоставляется в монопольное распоряжение виртуальной машины (технология *VMDirectPath*), чтобы *3D*-приложение могло использовать его напрямую.

5.3. Подключение виртуальных рабочих мест к файловому хранилищу суперкомпьютера

Следующий вопрос – подключение к системе хранения данных суперкомпьютера, которая построена на основе параллельной файловой системы ***Lustre***. Важно, что и *VMware Horizon View*, и *Citrix XenDesktop* поддерживают доставку лишь рабочих мест под управлением *Windows*. В связи со сложностями, возникающими при настройке доступа *windows*-пользователей к *linux*-объектам (пользователям рабочих станций нужно иметь доступ к своим папкам в ФС *Lustre*), наиболее простым подходом представляется использование “шлюза”.

Под шлюзом в данном случае подразумевается промежуточный *Linux*-сервер, который подключается к *Lustre*-хранилищу и реэкспортирует ресурс, к которому он получает доступ, виртуальной машине по протоколу *CIFS*. Подключение шлюза к системе хранения данных суперкомпьютера организуется посредством программного *Lustre*-клиента – аналогично узлам кластера.

Возможны различные варианты организации шлюза:

1. Виртуальный шлюз *Lustre* – в эту виртуальную машину средствами гипервизора передается *InfiniBand*-адаптер, и гостевая операционная система работает с ним напрямую. Виртуальным рабочим местам данные предоставляются по протоколу *CIFS* через внутреннюю сеть гипервизора.
2. Физический шлюз *Lustre* и доступ к *CIFS*-ресурсам по каналу *InfiniBand*. Гипервизор

Табл. 1. Сервер *HP ProLiant DL 380p Gen8*

Процессор	2 процессора *Intel Xeon E5-2630 v2* (6 ядер, 2.60 *GHz*, 15 *Mb Cache*)
Оперативная память	320 *Gb* (20×16 *Gb*) *DDR3-1600 MHz*
Графическая карта	1 карта *NVIDIA GRID K2* (два *GPU*, аналогичных *Quadro K5000*)
Диски	4 диска по 300 *Gb*, 15 000 *rpm*
Адаптер *InfiniBand*	*HP Infiniband QDR/Ethernet 10Gb 2-port 544FLR-QSFP Adapter*

VMware ESXi может предоставлять ресурсы *InfiniBand*-адаптера виртуальным машинам посредством виртуального сетевого адаптера, который гостевая операционная система видит как десятигигабитный *Ethernet*-адаптер.

Физический шлюз *Lustre* и доступ к *CIFS*-ресурсам по каналу *Ethernet* (10 *Gbit/s*). Этот вариант имеет смысл в том случае, когда на предприятии уже имеется сетевая инфраструктура, построенная на основе стандарта *10G Ethernet*.

В нашем случае такой инфраструктуры не было, но присутствовала *InfiniBand*-инфраструктура, поэтому выбор был сделан в пользу варианта №2.

5.4. Описание аппаратной части

Экспериментальный стенд был построен на серверной платформе *HP ProLiant DL 380p Gen8*, его конфигурация указана в табл. 1.

5.5. Описание виртуальных машин, развернутых на стенде

На стенде были развернуты следующие машины:

✓ Два виртуальных рабочих места с ОС *Windows 7 64-bit* (именно они доставлялись пользователям, и на них осуществлялось построение сеточных моделей):

- ***vDesktop-1***: шесть виртуальных процессоров, графический процессор *NVIDIA GRID K2*; ~192 *Gb* оперативной памяти (это максимум ОП, который поддерживается *Windows 7*; для *Windows 8* ограничение составляет 512 *Gb*, для *Windows Server 2012* – 4 *Tb* [8]);
- ***vDesktop-2***: шесть виртуальных процессоров, ~100 *Gb* оперативной памяти, графический процессор *NVIDIA GRID K2*;

✓ ***VMware vCenter Server*** – сервер управления виртуальной средой *VMware vSphere*;

✓ ***VMware Horizon View Connection Server*** – диспетчер подключений к виртуальным рабочим местам;

✓ **контроллер домена** – необходим для авторизации пользователей и хранения учетных записей виртуальных машин;

✓ **маршрутизатор** – виртуальная машина, необходимая для маршрутизации пакетов между различными сетями стенда.

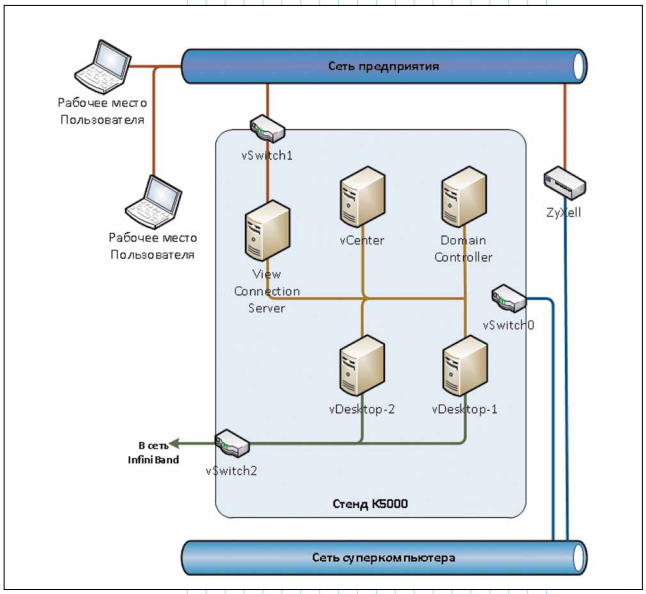

Рис. 2. Схема сети Ethernet

5.6. Подключение виртуальных машин к сети предприятия и высокоскоростной сети суперкомпьютера

На стенде были созданы виртуальные сетевые коммутаторы для следующих *TCP/IP*-сетей:

- внутренняя сеть *VMware Horizon View* (*vSwitch0*) – к ней были подключены виртуальные рабочие места и все инфраструктурные серверы;
- сеть предприятия (*vSwitch1*) – к ней был подключен *VMware Horizon View Connection Server*, к которому пользователи обращались для подключения к виртуальным рабочим местам;
- сеть *Lustre* (*vSwitch2*) использовалась для подключения виртуальных рабочих мест к шлюзу *Lustre*;
- внутренняя сеть суперкомпьютера (*vSwitch3*), через которую осуществлялось управление физическим сервером.

Упрощенная схема сети *TCP/IP* представлена на рис. 2.

Схема сети *InfiniBand* (высокоскоростной коммуникационной сети суперкомпьютера) приведена на рис. 3. Сеть включает в себя

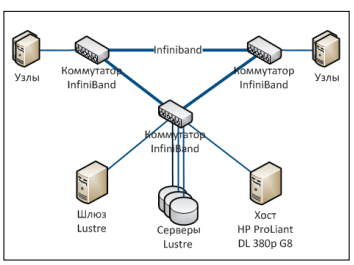

Рис. 3. Схема сети InfiniBand

три коммутатора, к каждому из них подключены 20 узлов кластера, в том числе серверы *Lustre*. Между собой каждые два коммутатора соединены избыточными связями (эти соединения обозначены на рисунке более жирными линиями). Шлюз подключен к коммутатору, обслуживающему серверы *Lustre*.

6. Результаты тестирования

В качестве оконечного оборудования на клиентской стороне можно использовать как программные клиенты, так и тонкие/нулевые клиенты. В нашем тестировании использовались программные клиенты *VMware Horizon View Client* и тонкие клиенты *HP t610 Plus* (*E4T96AA*). В тестировании принимали участие представители различных подразделений АО НИКИЭТ.

6.1. Общие тесты производительности

Была проведена серия тестов виртуального рабочего места *vDesktop*-1. Для тестирования использовалось ПО *SPECwpc*; результаты приведены в табл. 2.

Из таблицы можно сделать следующие выводы:

- При решении реальных инженерных задач виртуальная машина сопоставима по производительности с рабочей станцией уровня *HP Z420*;
- В некоторых задачах с повышенными требованиями к *CPU* (таких, как решение уравнений Пуассона) производительность виртуальной машины оказалась существенно выше, чем у *Z420*. Причина в том, что кэш процессора *E5-2630 v2* в сервере *VDI* много больше кэша *E3-1280 v3* и *E5-1660 v2*, установленных на рабочих станциях *Z230* и *Z420* соответственно.
- Производительность графики соизмерима с производительностью графики на аппаратных решениях; падения производительности за счет виртуализации нет или оно незначительно (~10%).
- Производительность *CPU* на реальных задачах также достаточна. Влияние системы виртуализации незаметно (*E5-1660v2* быстрее *E5-2630v2* на 30%, а на результатах расчетов тестовых задач такой большой разницы не видно).
- Наблюдается некоторая задержка в передаче видеопотока от виртуальной машины к пользователю. Возможно, “обычного” тонкого клиента для передачи изображения высокого разрешения недостаточно. Необходимо протестировать, как поведут себя клиенты с аппаратной поддержкой протокола *PCoIP*.

6.2. Работа с сеточными моделями

Представителем расчетного подразделения были произведены тесты по использованию виртуального рабочего места *vDesktop*-2 для разработки сеточных *CAE*-моделей *ANSYS CFX*. В качестве расчетных моделей использовались условия экспериментов по проливкам тепловыделяющей сборки из 19-ти стержней в ИТ СО РАН и по продувкам сборки из 37-ми стержней в МГТУ на кафедре Э7 (рис. 4). Размеры сеточных моделей составляли 10…100 млн. элементов, объем занимаемой оперативной памяти – 1 млн. элементов ≈ 1 *Gb*.

На физической рабочей станции *HP Z620* (два процессора I*ntel Xeon E5 2640 v2*, 48 *Gb* оперативной памяти) комфортно работать, когда занимаемый сеточной моделью объем оперативной памяти не превышает 40 *Gb*. Напротив, работа на виртуальной машине позволяла весьма эффективно снять пики загрузки операций с сеточными моделями, требующими до 80 *Gb* оперативной памяти. В результате расчетов на

Табл. 2. Ускорение в сравнении с рабочими станциями

Тест	***HP Z230 Intel Xeon E3-1280 v3* 3.6 *GHz*, 16 *Gb* ОЗУ, графическая карта *Quadro K4000***	***HP Z420 Intel Xeon E5-1660 v2* 3.7*GHz*, 32 *Gb* ОЗУ, графическая карта *Quadro K5000***
Entertainment (Maya, HD Video, Tessellation):	1.1×	0.8×
Financial (использование финансовых программ: стат. обработка больших баз данных)	1.2×	0.9×
CAD (*SolidWorks, NX, CATIA, Showcase*)	1.5×	1.0×
CAE (расчеты *CFD*)	1.7×	1.15×
Energy (в основном, расчеты на *CPU*)	2.1×	1.3×
LifeSciences (расчеты на *CPU* сложных математических задач, постобработка больших баз данных)	1.5×	1.0×
General (*7zip*, компилирование и т.д.)	2.3×	1.9×
Уравнения Пуассона	4.0×	2.7×

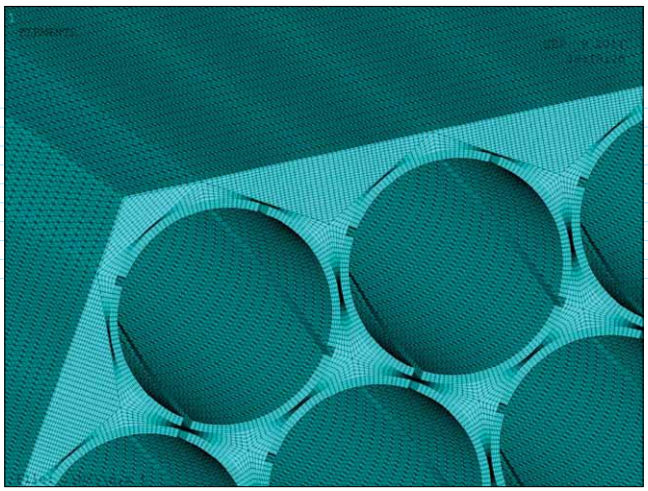

Рис. 4. Пример сеточной модели тепловыделяющей сборки из 37-ми стержней – 80 млн. ячеек

моделях в десятки миллионов ячеек были получены сведения о механизмах межъячеистого обмена массой теплоносителя и их количественные характеристики. Полученные результаты хорошо согласуются с результатами экспериментов.

6.3. Задачи рендеринга

В процессе тестирования виртуального десктопа *vDesktop-2* для задач рендеринга в программных продуктах *Adobe After Effects* и *Adobe 3DS Max* были получены следующие результаты:

- Рендеринг *3D*-модели из *Adobe 3DS Max* на виртуальном десктопе *VDI* был выполнен за 1 час 12 мин. Для просчета аналогичной модели на рабочей станции *HP Z230* (*Intel Core i7-4770*, 32 *Gb* ОП) было затрачено 43 мин.
- Рендеринг видеоролика из программного продукта *Adobe After Effects* на *VDI* был выполнен за 10 часов 32 мин. Для рендеринга аналогичного видеоролика на рабочей станции *HP* Z620 (два процессора *Intel Xeon E5-2640*, 48 *Gb* ОП) было затрачено 17 часов 08 мин.

Исходя из вышеизложенного, можно констатировать, что время рендеринга видеороликов на виртуальных машинах *VDI* является приемлемым, и данное решение вполне применимо для этих задач.

Заключение

Описанный в данной статье стенд обеспечивает двум пользователям одновременную работу с большими сеточными моделями. Размер моделей ограничивается размером оперативной памяти: для 192 *Gb* оперативной памяти (машина *vDesktop-1*) потолок составляет порядка 200 млн. контрольных объемов; для 100 *Gb* оперативной памяти (машина *vDesktop-2*) потолок вдвое меньше – порядка 100 млн. контрольных объемов (эти оценки справедливы для *ANSYS Mechanical APDL*).

Важно, что виртуализация позволяет легко перераспределить оперативную память между виртуальными рабочими местами. Это крайне полезно, когда на разных этапах расчетов используются сеточные модели разных размеров с разными требованиями к оперативной памяти.

В завершение нужно сказать также несколько слов о масштабировании. Для данного подхода главными характеристиками оборудования являются объем ОП и количество графических процессоров. Оба эти параметра можно увеличить: в настоящий момент на рынке представлены серверы с объемом оперативной памяти до 1.5 *Tb*, а также решения, допускающие установку нескольких карт *NVIDIA GRID*. Оборудование такого уровня позволило бы увеличить объем оперативной памяти, доступной виртуальному рабочему месту, и/или количество виртуальных рабочих мест, которые можно разместить на одном физическом сервере. Кроме того, описываемое решение допускает горизонтальное масштабирование за счет увеличения числа физических серверов.

Авторы благодарят *Hewlett-Packard Russia* за оборудование, предоставленное для тестирования, компании *ANSYS, Inc* и КАДФЕМ Си-Ай-Эс за предоставление временных лицензий на программное обеспечение *ANSYS*, а также сотрудников АО НИКИЭТ – Огнерубова Д.А., Мариничева Д.В. и Куликова А.А. за помощь в тестировании решения.

Литература

1. Сайт проекта *Lustre* // wiki.lustre.org/index.php/Main_Page
2. Сайт проекта *VirtualGL*, документация // svn.code.sf.net/p/virtualgl/code/trunk/doc/index.html#hd003
3. Сайт компании *Vmware*, страница о *Horizon* // www.vmware.com/ru/products/horizon-view
4. Сайт компании *Citrix*, страница о *XenDesktop* // www.citrix.ru/products/xendesktop/overview.html
5. Сайт компании *Microsoft*, страница о *VDI* // www.microsoft.com/ru-ru/windows/enterprise/products-and-technologies/virtualization/vdi.aspx
6. Сайт компании *RedHat*, страница о виртуализации // www.redhat.com/products/cloud-computing/virtualization
7. Сайт корпорации *NVIDIA*, страница о технологии *GRID* // www.nvidia.com/object/nvidia-grid.html
8. Сайт компании *Microsoft*, информация о максимальном размере ОЗУ // msdn.microsoft.com/en-us/library/windows/desktop/aa366778%28v=vs.85%29.aspx