ИССЛЕДОВАНИЕ ПРОИЗВОДИТЕЛЬНОСТИ КЛАСТЕРНЫХ СИСТЕМ ХРАНЕНИЯ ДАННЫХ В ЗАДАЧАХ ОБРАБОТКИ ДАННЫХ СЕЙСМОРАЗВЕДКИ

Е.А. Курин, Е.Л. Музыченко

ВВЕДЕНИЕ

Сейсмическая разведка была и остаётся основным методом поисков и разведки месторождений нефти и газа. Сейсмические данные используются на всех этапах и стадиях нефтегазопоисковых работ [1]. В последние годы основной объём сейсморазведочных работ выполняется в районах со сложными сейсмогеологическими условиями (глубоководные районы, солянокупольная тектоника, вечная мерзлота, переходная зона, значительные колебания рельефа поверхности наблюдений и пр.). Всё это предъявляет повышенные требования к качеству полевых наблюдений и к совершенствованию графа обработки данных. Как правило, в таких условиях необходимо применять самые совершенные обрабатывающие алгоритмы, которые требуют использования значительных вычислительных ресурсов вплоть до суперкомпьютеров [2].

Несмотря на непрерывный рост производительности компьютеров, необходимость в значительных вычислительных ресурсах с каждым годом увеличивается из-за внедрения более плотных систем сейсмических наблюдений и новых ресурсоёмких алгоритмов обработки данных. Объём получаемых в процессе наблюдений данных в настоящее время может составлять десятки, а в отдельных случаях, и сотни терабайт на одну исследуемую площадь. Таким образом, системы хранения и передачи данных являются критически важными элементами вычислительных систем для обработки сейсмических данных. В настоящей работе на примере кластерных файловых систем российских суперкомпьютеров «Чебышев» и «Ломоносов» изучаются вопросы производительности доступа к данным в приложении к решению некоторых задач обработки результатов сейсмических наблюдений.

ВЫЧИСЛИТЕЛЬНЫЕ СИСТЕМЫ

Суперкомпьютер «Чебышев», построенный в 2008 году и установленный в НИВЦ МГУ имени М.В. Ломоносова, состоит из 625 двухпроцессорных вычислительных узлов на базе процессоров Intel Xeon E5472, связанных между собой коммуникационной сетью Infiniband DDR. Большинство узлов — бездисковые. В качестве кластерной файловой системы используется устройство Panasas ActiveStor 5000 общей ёмкостью 60 терабайт. В качестве сети передачи данных от файловой системы к вычислительным узлам используется Gigabit Ethernet.

Вторая исследуемая система, «Ломоносов», построена в 2009 году и установлена в Суперкомпьютерном центре МГУ имени М.В.Ломоносова. В общей сложности, она имеет 4446 вычислительных узлов, большинство из которых — на базе процессоров Intel Xeon X5570. Узлы связаны между собой коммуникационной сетью Infiniband QDR. Большая часть узлов не имеет собственных жёстких дисков. Кластерная файловая система построена из базе программного обеспечения с открытым исходным кодом Lustre. Суммарный доступный размер файловой системы превышает 300 терабайт. Доступ вычислительных узлов к файловой системе осуществляется при помощи общей коммуникационной сети Infiniband QDR.

СТАНДАРТНАЯ ОБРАБОТКА СЕЙСМИЧЕСКИХ ДАННЫХ

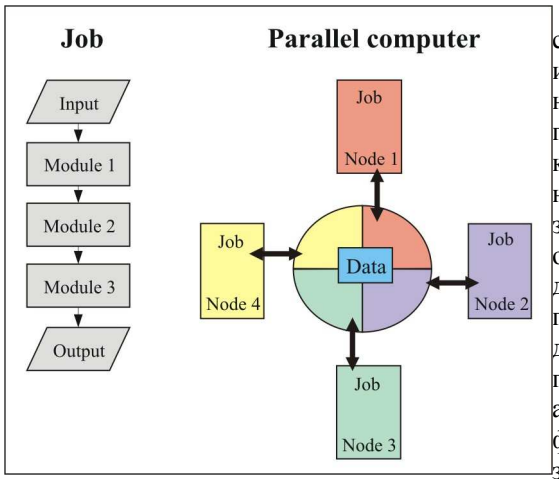

Рис. 1. Схема доступа к данным при стандартной обработке.

Как правило, полученные в результате наблюдений сейсмические данные осложнены различного рода регулярными и нерегулярными помехами, что препятствует их непосредственной интерпретации. Поэтому необходимо провести так называемую стандартную обработку, в ходе которой производится подавление помех и выделение сигнала, несущего информацию о структуре исследуемого участка земной коры. Как правило, граф стандартной обработки организован в виде конвейера, в котором некоторая порция данных последовательно проходит через ряд обрабатывающих процедур. При этом чтение, запись и обработка каждой порции данных производится независимо от других порций данных, как показано на Рис.1. Здесь «job» обозначает задание на обработку, а «module» - отдельную обрабатывающую процедуру, например, фильтр. Как правило, операции доступа к сейсмическим записям (трассам) осуществляются в последовательном режиме, кроме операции сортировки данных, при которой необходим произвольный доступ к большой части файла.

В качестве теста, адекватно воспроизводящего чтение и запись данных при стандартной обработке, была разработана простая MPI-программа [3] со следующим алгоритмом работы:

- Каждый MPI-процесс последовательно записывает свой файл на кластерной файловой системе, при этом размер файла существенно превышает размер оперативной памяти вычислительного узла;
- После очистки буферов файловой системы в оперативной памяти MPI-процесс последовательно считывает ранее созданный файл;
- После очистки буферов файловой системы MPI-процесс считывает файл в так называемом «режиме сортировки», когда блоки данных (трассы) прочитываются не последовательно, а с некоторым интервалом, на два-три порядка превышающим размер блока.

На каждом этапе работы программы производится измерение скорости доступа к файлу отдельного процесса, а также суммарной скорости чтения-записи всеми процессами.

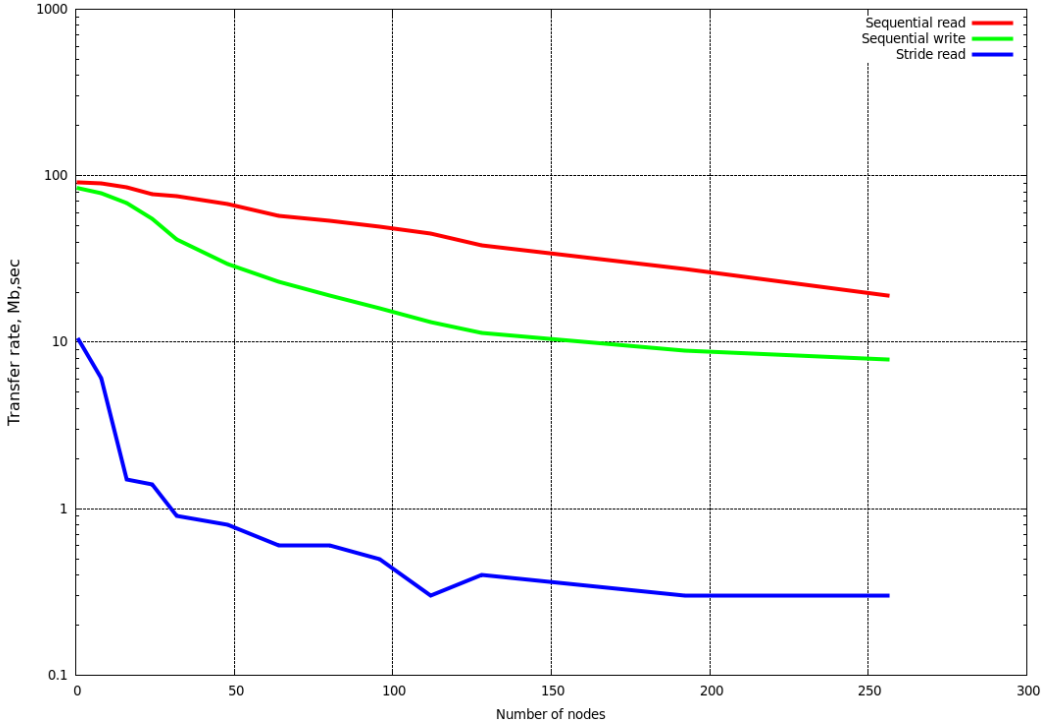

Рис. 2(а). Результаты тестов чтения-записи данных для одного узла суперкомпьютера «Чебышев».

На Рис.2(а) приведены результаты работы программы на суперкомпьютере «Чебышев» для одного вычислительного узла. По горизонтальной оси отложено число одновременно работающих вычислительных узлов, а по вертикальной — скорость доступа в мегабайтах в секунду. Для удобства анализа графиков вертикальная шкала представлена в логарифмическом масштабе. Красным цветом показана кривая зависимости скорости последовательного чтения от количества «читающих» узлов, зелёным - скорости последовательной записи, синим — скорости чтения «в режиме сортировки». Во всех тестах размер блока данных составлял 8 килобайт, а размер файла в полтора раза превышал размер оперативной памяти, установленной на вычислительном узле. Графики показывают достаточно стабильное и предсказуемое поведение СХД при последовательном доступе к файлам. Можно отметить, что для небольшого (до десяти) количества узлов скорость чтения «в режиме сортировки» достаточно высока, то есть сопоставима с показателями, получаемыми при использовании локальных жёстких дисков. Однако, для большего количества узлов она быстро уменьшается.

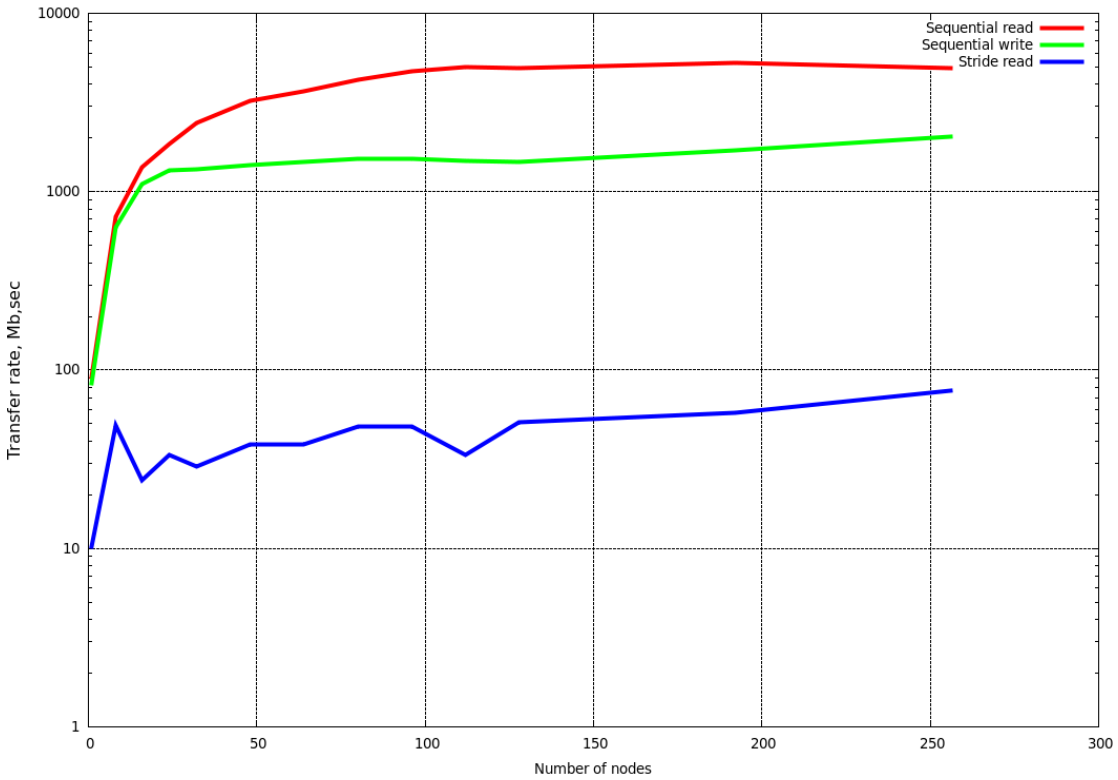

Рис. 2(б). Суммарные результаты тестов чтения-записи данных суперкомпьютера «Чебышев».

На Рис.2(б) представлены показатели суммарной производительности, полученные в том же численном эксперименте. Максимальная суммарная скорость чтения достигается при использовании более 100 вычислительных узлов. Другими словами, чтобы в полной мере использовать возможности данной системы, приложение должно состоять не менее чем из 100 процессов. Что касается суммарной скорости чтения «в режиме сортировки», то она не превышает 100 мегабайт в секунду при любом количестве узлов-клиентов.

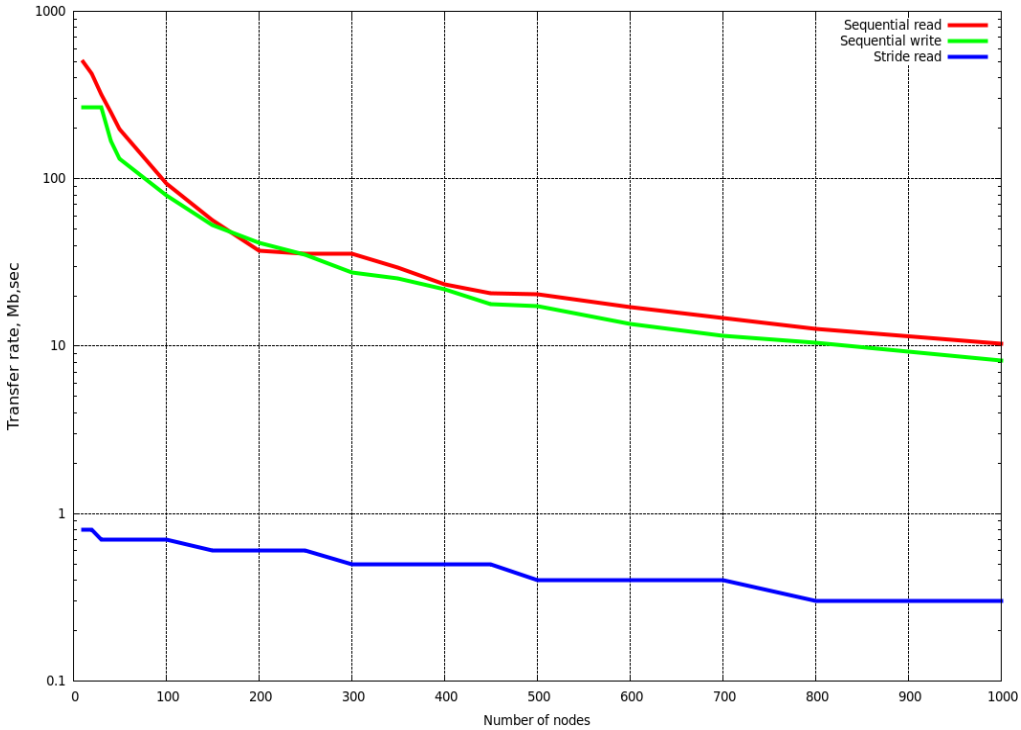

Рис. 3(а). Результаты тестов чтения-записи данных для одного узла суперкомпьютера «Ломоносов».

Перейдём к анализу результатов работы тестовой программы, полученных на суперкомпьютере «Ломоносов». На Рис.3(а) показаны показатели скорости доступа к данным для одного узла в зависимости от количества одновременно работающих узлов, а на Рис 3(б) — суммарные показатели для всех узлов. Видно, что

скорость чтения «в режиме сортировки» мала даже для небольшого количества узлов. Что касается суммарной скорости чтения-записи в последовательном режиме, то благодаря использованию сети Infiniband она достигается уже при использовании 25-30 узлов.

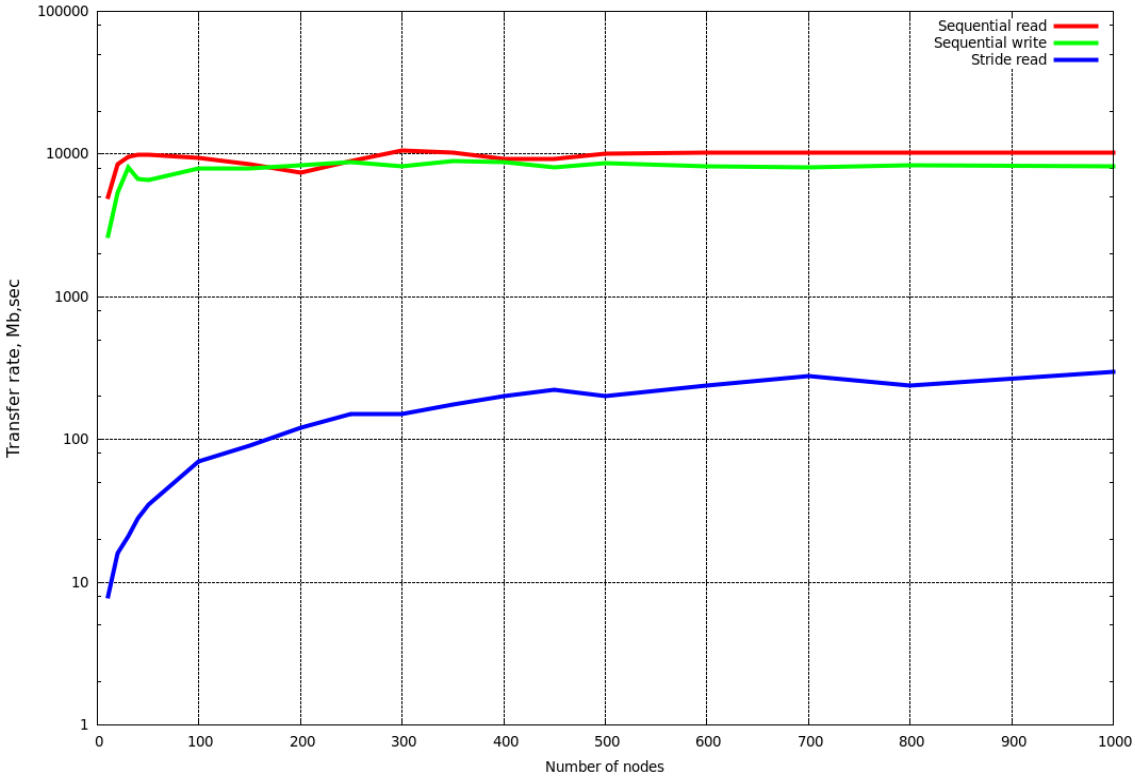

Рис. 3(б). Суммарные результаты тестов чтения-записи данных суперкомпьютера «Ломоносов».

ПРОЦЕДУРЫ ОБРАБОТКИ, ТРЕБУЮЩИЕ ДОСТУПА КО ВСЕМ ДАННЫМ

К сожалению, не для всех обрабатывающих процедур можно организовать подобную простую схему доступа к данным. Некоторые процедуры обработки и построения глубинных сейсмических изображений для получения одной выходной трассы требуют доступа к значительной части входных данных при том, что оперативной памяти одного вычислительного узла недостаточно для хранения необходимой части данных. В качестве примера подобной процедуры мы рассмотрим алгоритм SRME подавления кратных волн-помех, связанных с дневной поверхностью, при обработке данных площадных (3D) наблюдений.

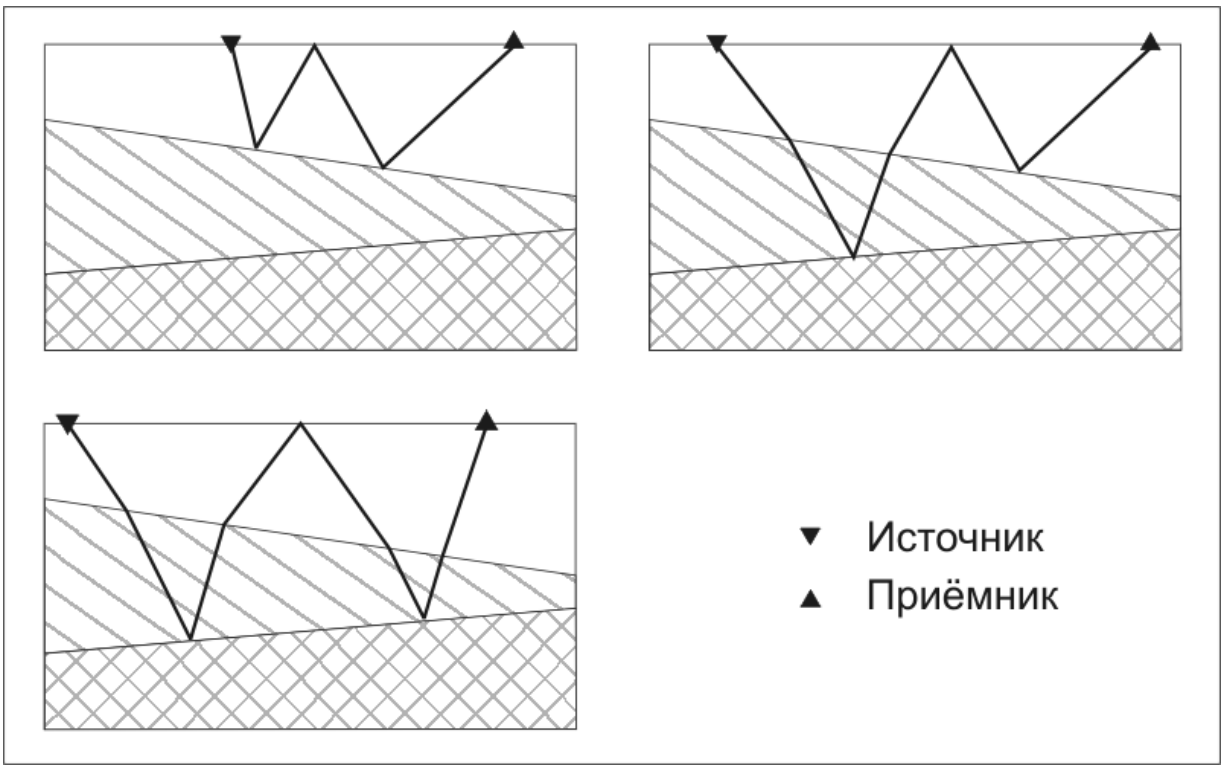

Рис. 4. Лучевая схема образования кратных волн, связанных с дневной поверхностью.

При обработке данных, особенно, полученных на акваториях, необходимо произвести ослабление регулярных помех, вызванных многократными отражениями от жёстких глубинных границ и дневной поверхности — кратных волн. На Рис. 4 приведена лучевая схема образования кратных волн, связанных с дневной поверхностью. К применявшимся ранее методам, основанным на различии каких-либо свойств однократных и кратных отражений, в настоящее время добавились двухшаговые методы, где на первом этапе производится прогнозирование поля помех, а на втором — его адаптивное вычитание из исходного поля. Одним из самых эффективных способов является метод SRME (Surface Related Multiple Elimination), где для прогнозирования поля волн-помех используются только сами данные [4]. Математическое выражение, на основе которого осуществляется прогнозирование трассы $M(S,R,t)$ модели кратных волн, может быть записано в виде

$$M(S,R,t)=f(t)*\sum_{Z} D(S,Z,t)*\tilde{D}(Z,R,t) \quad ,$$

где S и R - соответственно координаты положения источника и приемника, $D(S,Z,t)$ - трассы исходного волнового поля, $\tilde{D}(Z,R,t)$ - трассы исходного поля после предварительной обработки с целью учета косинуса угла выхода луча отраженной волны, $f(t)$ - компенсирующий фильтр. Как следует из формулы, трассы массива $D(S,Z,t)$ сгруппированы по координате общего пункта взрыва (ОПВ(S)) , а массива $\tilde{D}(Z,R,t)$ - по координате общего пункта приема (ОПП(R)). При этом вычисления сводятся к суммированию в пределах некоторой апертуры взаимных сверток трасс сейсмограмм ОПВ(S) и ОПП(R).

Особенностью всех схем 3D (площадных) наблюдений является то, что, как правило, имеются достаточно «плотные» сейсмограммы ОПВ, но не удается осуществить подборку трасс ОПП. Для решения этой проблемы предлагались различные методы “восстановления” отсутствующих трасс, например, [5], [6]. В настоящей работе применяется алгоритм, предложенный в работах [7], [8]. Суть метода состоит в поиске трассы, источник и приёмник которой находятся на наименьшем (в смысле наименьших квадратов) расстоянии от “источника” и “приёмника” отсутствующей трассы, а затем в применении разностных кинематических поправок с учётом изменения направления “источник-приёмник” (азимута).

В работе рассматриваемого алгоритма можно выделить следующие элементарные блоки (в порядке убывания используемого процессорного времени):

- Поиск ближайшей трассы к отсутствующей для её замещения. Обычно эта проблема сводится к поиску нужного значения в нескольких связанных по значению ассоциативных массивах большого размера (1-10 гигабайт).
- Нелинейное масштабирование временного ряда (ввод разностных кинематических поправок).
- Дискретная свёртка и дискретное преобразование Фурье.

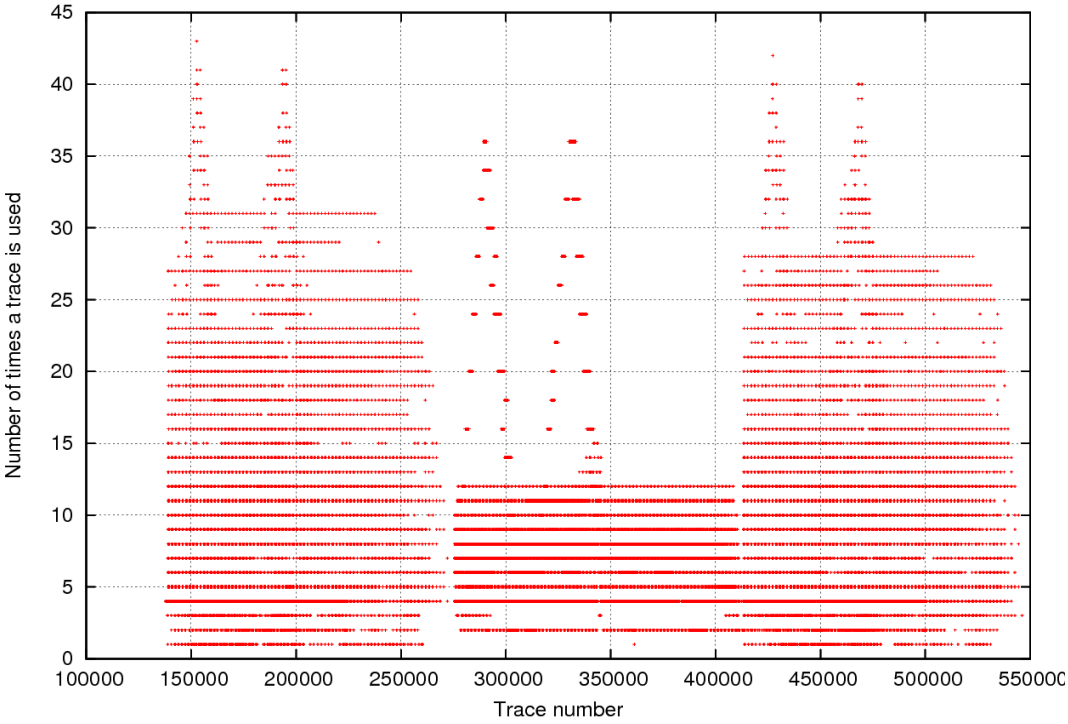

Рис. 5. Гистограмма доступа к трассам при прогнозировании поля кратных волн для одной сейсмограммы ОПВ.

Основной проблемой для эффективной реализации данного алгоритма на суперкомпьютерах с распределённой памятью является необходимость произвольного доступа к данным, размер которых зачастую превышает размер доступной оперативной памяти всех вычислительных узлов. Так, при непосредственной программной реализации описанного алгоритма и использовании локальных дисков вычислительного узла до 50 процентов процессорного времени приходится на ожидание операций ввода-вывода. Как показывают результаты тестов с квазипроизвольным доступом («режим сортировки»), рассмотренные выше, подобная реализация для кластеров с доступом к данным, расположенным на кластерной СХД, обладает крайне ограниченной масштабируемостью. Более того, очевидно, что при таком доминировании времени доступа к данным в общем времени расчётов всякая оптимизация кода лишена практического смысла.

Чтобы исследовать проблему и найти пути её решения, мы построили гистограмму использования трасс входного набора данных для получения одной сейсмограммы ОПВ модели кратных волн (Рис.5). Анализ шаблона доступа к трассам на диске показывает, что одна и та же трасса может быть использована несколько, иногда десятки, раз. Для оптимизации скорости доступа к данным предлагается метод, состоящий в предварительном определении наборов {источник, приёмник, промежуточная точка}, в которые вносит вклад та или иная трасса. Затем производится сортировка полученных таблиц с ключом – номером трассы. После чего на каждом вычислительном узле организуется квазипоследовательный доступ к трассам и соответствующие вычисления, то есть применяется *потоковая модель доступа* к исходным данным.

Принимая во внимание, что объединённые сейсмограммы ОПВ (super-shot gathers) при морских широкоазимутальных наблюдениях, имеют, как правило, существенно разный размер, в программе была реализована динамическая балансировка загрузки вычислительных узлов на основе популярной для MPI-приложений модели “master-slaves”.

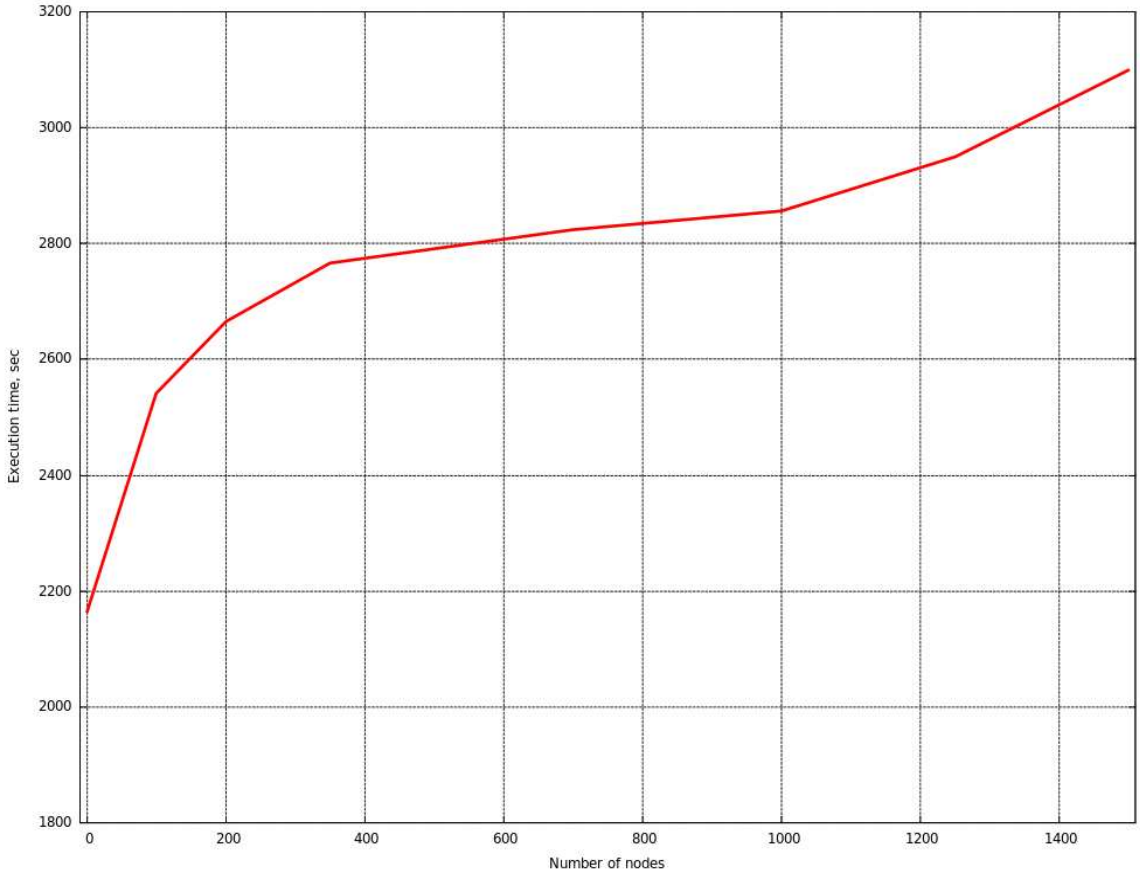

Рис. 6. Зависимость времени прогнозирования поля кратных волн для одной сейсмограммы ОПВ от количества одновременно работающих вычислительных узлов.

Тестирование проводилось на суперкомпьютере “Ломоносов”, установленном в МГУ имени М.В.Ломоносова. На Рис. 6 приведена зависимость времени вычислений для одной сейсмограммы общего пункта взрыва (ОПВ) на одном узле в зависимости от количества используемых узлов. Видно, что предложенный метод позволил достичь хорошей масштабируемости вычислений для значительного количества вычислительных узлов.

Следует отметить, что в данном примере все вычисления на отдельных узлах были распараллелены при помощи директив OpenMP, и программа была собрана при помощи компилятора C++ Intel версии 12. На Рис.7 приведён анализ использования процессорного времени на каждом вычислительном узле при одновременной работе 1000 узлов. По горизонтальной оси отложено время с момента начала расчётов, а по вертикальной — процент используемых в данный момент ресурсов процессоров (ядер). Красным цветом показана часть времени, используемая процессорами для, собственно, вычислений (user), а зелёным - для обслуживания системных вызовов (sys) и ожидания ввода-вывода (wait). Видно, что проблема ввода-вывода была успешна решена, и примерно 90 процентов процессорного времени занимают вычисления. Подобная организация потоков данных позволила произвести дальнейшую оптимизацию работы алгоритма, в том числе, для гибридных вычислительных систем на основе графических процессоров.

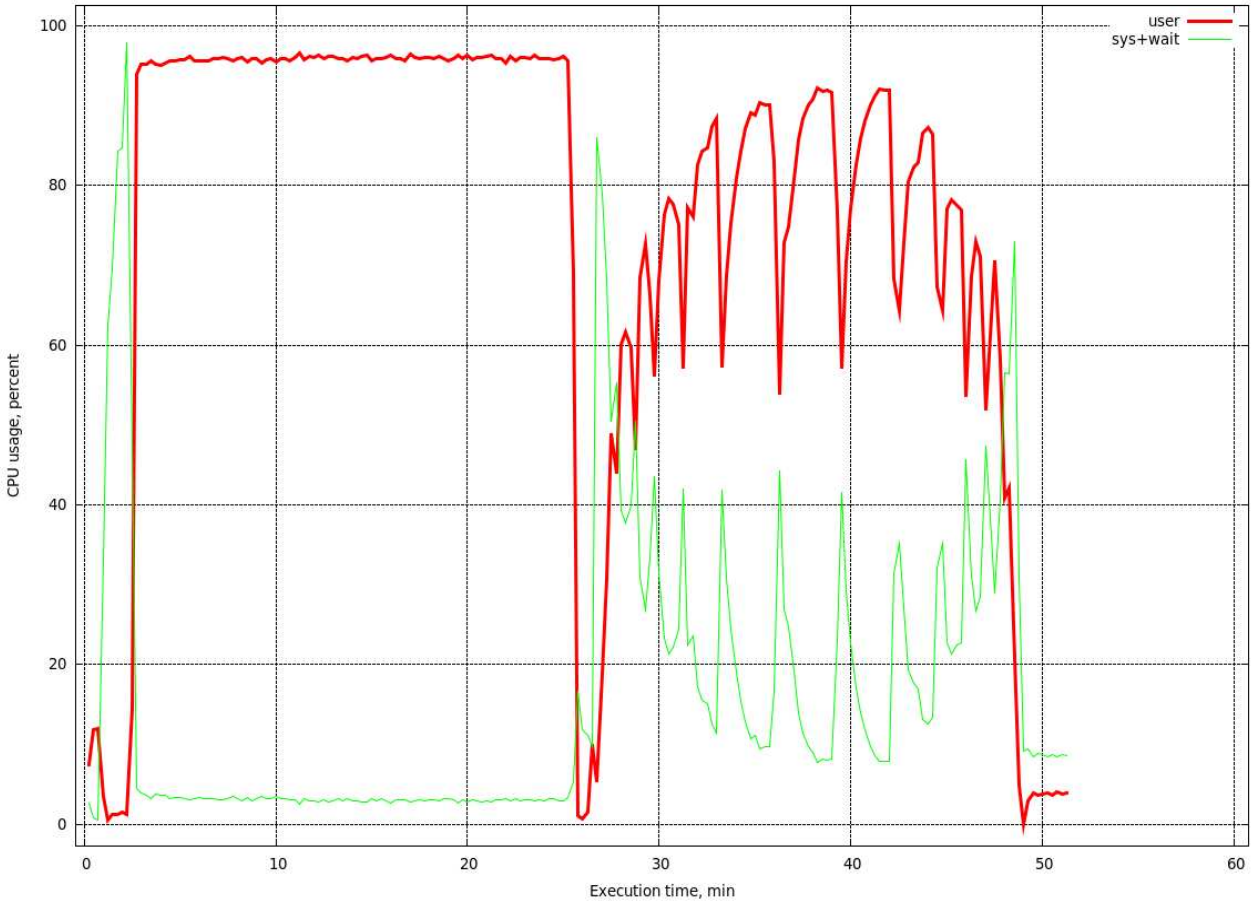

Рис. 7. Использование процессорных ресурсов одного вычислительного узла при прогнозировании поля кратных волн для одной сейсмограммы ОПВ.

ВЫВОДЫ

Система хранения данных (СХД) является критически важным элементом оборудования вычислительного центра для обработки больших объёмов сейсмических данных. Тестирование СХД российских суперкомпьютеров «Чебышев» и «Ломоносов» как на задачах стандартной обработки данных сейсморазведки, так и на особо требовательной к вычислениям программе подавления кратных волн-помех методом 3D SRME, позволяет сделать следующие выводы:

- Более быстрая сеть для соединения СХД и вычислительных узлов— существенное преимущество, позволяющее более эффективно использовать ресурсы суперкомпьютера.
- При программной реализации алгоритмов обработки следует учитывать, что все имеющиеся на сегодняшний момент кластерные СХД могут обеспечить приемлемую производительность только при последовательном доступе.
- Тщательный учёт возможностей СХД позволяет эффективно реализовывать сложные процедуры обработки данных сейсморазведки для различных вычислительных систем, в том числе, с использованием бездисковых вычислительных узлов.
- Существующие кластерные СХД способны обеспечить эффективную работу программ на суперкомпьютерах «петафлопсного» уровня производительности, однако, для бесперебойного обеспечения данными подобных программ на суперкомпьютерах «экзафлопсного» уровня потребуется существенное, на два-три порядка, увеличение скорости доступа.

Авторы признательны М.С.Денисову («ГЕОЛАБ») за полезные советы при разработке рассматриваемого в статье алгоритма подавления кратных волн-помех и за критические замечания при подготовке настоящего текста, А.А.Наумову («Т-Платформы») и С.А.Жуматию (НИВЦ МГУ имени М.В.Ломоносова) за всестороннюю помощь при проведении тестирования файловых систем суперкомпьютеров «Чебышев» и «Ломоносов», Д.Е.Локштанову (Statoil) за предоставление интересных и сложных данных и постановку геофизических задач.

ЛИТЕРАТУРА:

1. Ампилов Ю.П., 2008, От сейсмической интерпретации к моделированию и оценке месторождений

нефти и газа, М., Спектр
2. Курин Е.А., 2010, Сейсморазведка и суперкомпьютеры // Труды Международной суперкомпьютерной конференции «Научный сервис в сети Интернет: суперкомпьютерные центры и задачи»
3. Курин Е.А., 2007, О производительности компьютеров для обработки данных сейсморазведки // Технологии сейсморазведки, №4, 69-76
4. Berkhout A.J., Verschuur D.J., 1997, Estimation of multiple scattering by iterative inversion, Parts 1 and Part 2, Geophysics, 62, 1586-1595, 1596-1611.
5. Pica A., Poulain G., David B., Magesan M., Baldock S., Weisser T., Hugonnet P., and Hermann Ph., 2005, 3D Surface-Related Multiple Modeling, 67th EAGE Mtg., Exp. Abstracts.
6. Levin S.A., 2002, Prestack Poststack 3D Multiple Prediction, 72nd SEG Mtg., Exp. Abstracts.
7. Kurin E., Denisov M.S., Lokshtanov D., 2006, A method for 3D surface-related multiples prediction in case of coarse sampling // SEG-EAGE Conference Saint Petersburg 2006.
8. Денисов М.С., Курин Е.А, 2007, Способы прогнозирования кратных волн по данным площадных морских наблюдений // Технологии сейсморазведки, 2, 73–78.