# О некоторых подходах к оптимизации работы суперкомпьютеров

**Д.В.Андрюшин, В.С.Горбунов, Л.К.Эйсымонт**

Суперкомпьютеры предназначены для решения задач, которые не могут быть решены на других средствах из-за присущей им вычислительной сложности и большого объёма обрабатываемых данных. Особую важность имеют суперкомпьютеры высшего диапазона производительности, каждый проект создания такого суперкомпьютера по-своему уникален, отличается не только достижением высоких показателей пиковых характеристик, но и тщательно выполненной оптимизацией программно-аппаратных средств, продуманными технологиями оптимального использования.

При любой оптимизации центральный вопрос - эффективность взаимодействия с памятью, это главный фактор, влияющий на развиваемую на задачах реальную производительность суперкомпьютеров. Проблема обеспечения высокого уровня реальной производительности актуальна в последние два десятка лет, поскольку разрыв пиковой и реальной производительности непрерывно увеличивался и в настоящее время, в зависимости от пространственно-временной локализации работы с памятью задач, может достигать от 10 до 1000 раз.

Определенных успехов в решении проблемы роста разрыва реальной и пиковой производительности достичь в результате выполнения некоторых зарубежных и отечественных проектов последних десяти лет. Наиболее известна программа DARPA HPCS создания высокопродуктивных перспективных суперкомпьютеров с глобально адресуемой памятью, ориентированных на достижение реальной петафлопсной производительности на широком классе задач. Программа выполнялась с июня 2002 по декабрь 2010 года. Аналогичные программы были немедленно запущены в Китае и Японии, близкие проекты выполнялись и в России (например, проекта “Ангара”). Главные образцы суперкомпьютеров, использующих результаты этих программ, предполагалось создать после 2015 года.

Часть результатов программ предназначались для открытого рынка. Судя по появившимся на рынке и в рейтингах новым суперкомпьютерам, это и происходит: американские IBM Power 775 (проект IBM PERCS программы DARPA HPCS) и Cray XC30 (проект Cray Cascade программы DARPA HPCS); китайские Tianhe-1A и Tianhe-2 (Национальный институт оборонных технологий (NUDT), программа №863); японский K-компьютер

(Институт физических и химических исследований (RIKEN), Fujitsu, Проект суперкомпьютеров следующего поколения (NGSP)).

Одновременно с возникновением новых аппаратно-программных средств появились и новые технологи оптимизации работы суперкомпьютеров. Эти технологии важны не только для решения текущих проблем, но и в проектах разработки будущих суперкомпьютеров высшего диапазона производительности экзафлопсного уровня ($10^{18}$ операций над вещественными числами в секунду или 1 Экзафлопс, Эфлопс), одновременно обладающих и высокой энергетической эффективностью вычислений в 50 Гфлопс/Вт, что в десятки раз выше современного уровня. Сложность таких работ, что повышает важность технологий оптимизации, еще и в том, что они выполняются на фоне исчерпания возможностей КМОП-технологий и появления технологий так называемой пост-Муровской эры, технологий, которые будут применяться после окончания действия закона Мура.

Новые технологии оптимизации работы суперкомпьютеров предполагают дальнейшее совершенствование вычислительных методов и алгоритмов, а также и аппаратно-программных средств.

**Системное представление о технологиях оптимизации.**

Общую картину условий применения и содержания технологий оптимизации работы суперкомпьютеров, что включает как оптимизацию приложений, так и оптимизацию самих суперкомпьютеров, сформулируем в форме следующих пяти тезисов, комментарии к которым будут далее.

**Тезис 1.** Микропроцессоры, основные компоненты современных суперкомпьютерных комплексов, чрезвычайно чувствительны к пространственно-временной локализации обращений к памяти, чем хуже пространственно-временная локализация обращений, тем ниже реальная производительность. Тенденция современных приложений к ухудшению пространственно-временной локализации, отразилась на характере целевых тестов, задаваемых в технических заданиях на разработку перспективных суперкомпьютеров.

**Тезис 2.** Плохая пространственно-временная локализация и деградация реальной производительности может происходить вследствие использования достаточно простых и естественных шаблонов работы с памятью, поэтому для создания эффективных приложений важны знания о профиле их работы с памятью, знания о возможностях оборудования по

эффективности выполнения обращений к памяти и знания о возможных методах оптимизации программ.

**Тезис 3.** Методы оптимизации программ могут быть разделены на два направления: первое - повышение эффективности за счет улучшения пространственно-временной локализации обращений к памяти в приложении; второе - обеспечение нечувствительности (толерантности) приложения к задержкам выполнения операций работы с памятью за счет использования иерархических массово-мультитредовых моделей организации программ над общей виртуальной сегментно-страничной памятью, а также резкого снижения количества обращений к памяти за счет применения управляемых данными потоковых моделей вычислений.

**Тезис 4**. Методы оптимизации суперкомпьютеров - обеспечение толерантности и поддержка потоковых моделей вычислений, связаны с использованием массово-мультитредовых моделей организации программ над общей виртуальной сегментно-страничной памятью с управляемым отображением на физическую память, а также применением потоковых моделей вычислений. Для наиболее эффективной поддержки таких моделей требуется особая организация процессоров, сетей и модулей памяти, что решалось в программах DARPA HPCS и будет решаться в экзафлопсных программах текущего десятилетия. Эти компоненты должны быть спроектированы с целью обеспечения высокой пропускной способности выполнения операций с памятью. Между тем, новейшие исследования показали, что требуемые модели вычислений и памяти могут быть реализованы и на стандартном оборудовании через схемы эмуляции, базирующиеся на сверхлегких тредах и высокой многоядерности оборудования.

**Тезис 5.** Важнейшее направление оптимизации работы суперкомпьютеров – специализация, как аппаратных, так и программных средств, создание аппаратно-программных комплексов. Это приобретает особую важность на фоне развернутых работ по оптимизации применения КМОП-технологий, а также внедрения новых технологий пост-Муровской эры, часть которых, как уже известно, будет представлена в виде специализированных блоков аналогового типа.

В представляемой автором организации ФГУП ”НИИ ”Квант” перечисленные тезисы воплощены в виде уже практически применяемых технологий, а также технологий, создаваемых в настоящее время с прицелом

на ближнюю и дальнюю перспективу. Обобщая существующие и будущие технологии можно сказать, что они нацелены на решение таких задач:

- исследование процессов выполнения приложений на суперкомпьютерах и возможностей собственно суперкомпьютеров на тестах с синтезируемой нагрузкой и тестах приложений разного уровня и специализации;

- предсказание производительности приложений исходя из полученных знаний об их выполнении и выявленных возможностей аппаратных средств, оценка возможностей оптимизации и выбор подходящих методов;

- проведение оптимизации программных средств и собственно суперкомпьютеров.

В работах по новым технологиям используются наработки и опыт использования созданной многоуровневой методики оценочного тестирования оборудования, ее описание и результаты ее использования уже были опубликованы а ряде работ, а краткое описание приведено в [1]. Результаты работ по методам оптимизации работы суперкомпьютеров будут применены не только в текущих исследованиях и разработках, но и в цикле работ по моделированию вариантов экзафлопсных систем, выполнение которых в настоящее время инициировано в РАН на базе мультипроцессорной гибридной вычислительной системы МГВС [2].

**Комментарии к тезисам и используемые технологии**

Чувствительность оборудования к пространственно-временной локализации обращений к памяти можно продемонстрировать, например, на результатах оценочного тестирования одного процессорного ядра микропроцессора Intel E5-2660 Sandy Bridge (пиковая производительность ядра составляет 17,6 Гфлоп/с при тактовой частоте 2.2 ГГц) на тестах Euroben (www.hpcresearch.nl/euroben).

На рис.1 демонстрируется связанная с ухудшением временной локализации обращений к памяти деградация реальной производительности на тесте поэлементного умножения векторов, если усложнять доступ к его элементам. Лучшая реальная производительность была ~ 3 Гфлоп/с, это около 17% от пиковой. При увеличении длины вектора и усложнении доступа к элементам происходит 10-кратная деградация реальной

производительности до 1.7% от пиковой, отчетливо заметна ступенчатость деградации, что связано с выходами за пределы кэш-памятей L1, L2 и L3.

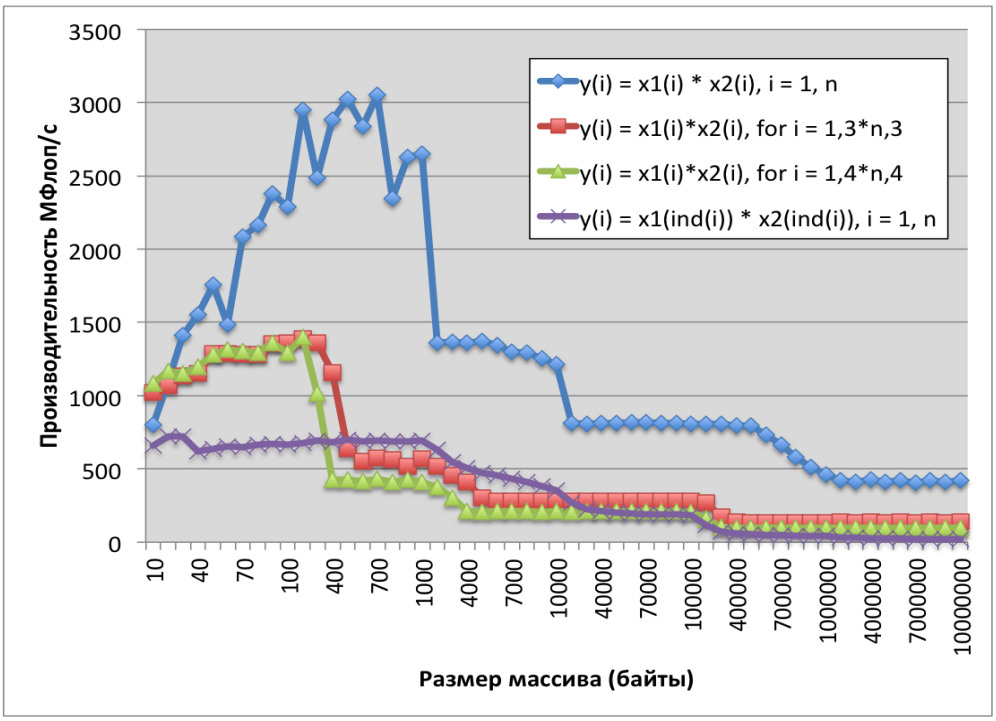



**Рис.1 Деградация реальной производительности одного ядра Intel Xeon E5-2660 (Sandy Bridge) при усложнении доступа к памяти.**

На рис.2 показано влияние увеличения временной локализации тестов на повышение реальной производительности. Временная локализация выше, чем чаще используются одни и те же операнды. Видно, что с увеличением количества операций, приходящихся на одно обращение к памяти, реальная производительность значительно увеличивается, но она, все-таки, далека от пиковой. Лучший результат получен на тесте вычисления полинома 9-й степени по схеме Горнера, в котором на одно обращение к памяти приходится 18 вычислительных операций. Лучшая реальная производительность в этом случае составляет 51 %, а при увеличении вектора деградирует не так сильно, до 42 %.

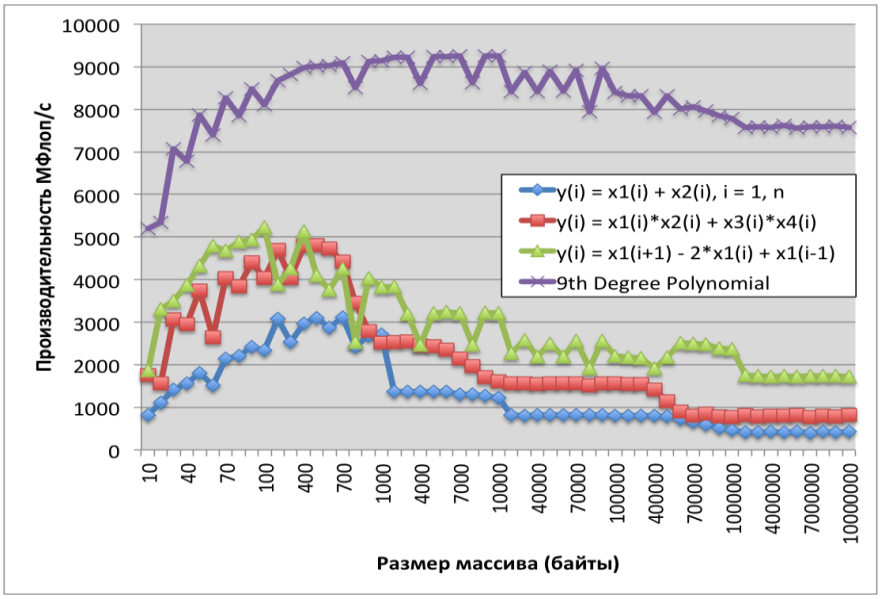



**Рис.2 Рост реальной производительности одного ядра Intel Xeon E5-2660 (Sandy Bridge) при увеличении количества вычислительных операций, приходящихся на одно обращение к памяти**

Понятие пространственно-временной локализации введено и формализовано в середине прошлого десятилетия. Количественная оценка может быть получена по профилю работы памяти с программой, формируемой в процессе ее выполнения в пространстве “адреса выдаваемых обращений - время выполнения программы”. Примеры таких профилей имеются на рис.3 и 5, каждое обращение к памяти на таком профиле – это точка, цвет отражает тип обращения - на считывание, запись или атомарная операция. Вид таких профилей и метрики пространственно-временной локализации позволяют получить исходные данные о программе, что может быть использовано при ее оптимизации. Сотрудниками ФГУП ”НИИ ”Квант” и СПбГПУ была разработана методика снятия таких характеристик по работающим реальным приложениям, которая в ближайшее время будет расширена по линии повышения управляемости и избирательности, качества визуализации и анализа.

Вычисляемая метрика пространственно-временной локализации приложения может быть использована при оптимизации программы в сочетании со знанием характеристик оборудования при выполнении обращений к памяти с заданной пространственно-временной локализацией. Такая характеристика получается на тестах с синтезируемой нагрузкой по обращениям к памяти с регулярно изменяемой пространственно-временной

локализацией. Пример такого теста – тест APEX-map, который строит некоторую APEX-поверхность (см.рис.3), представляющую собой среднее количество тактов, приходящихся на одно обращение к памяти на считывание в зависимости от пространственной и временной локализации таких обращений, которую задает тест. Построение APEX-поверхностей входит в базовую методику оценочного тестирования, применяемую на предприятии [1].

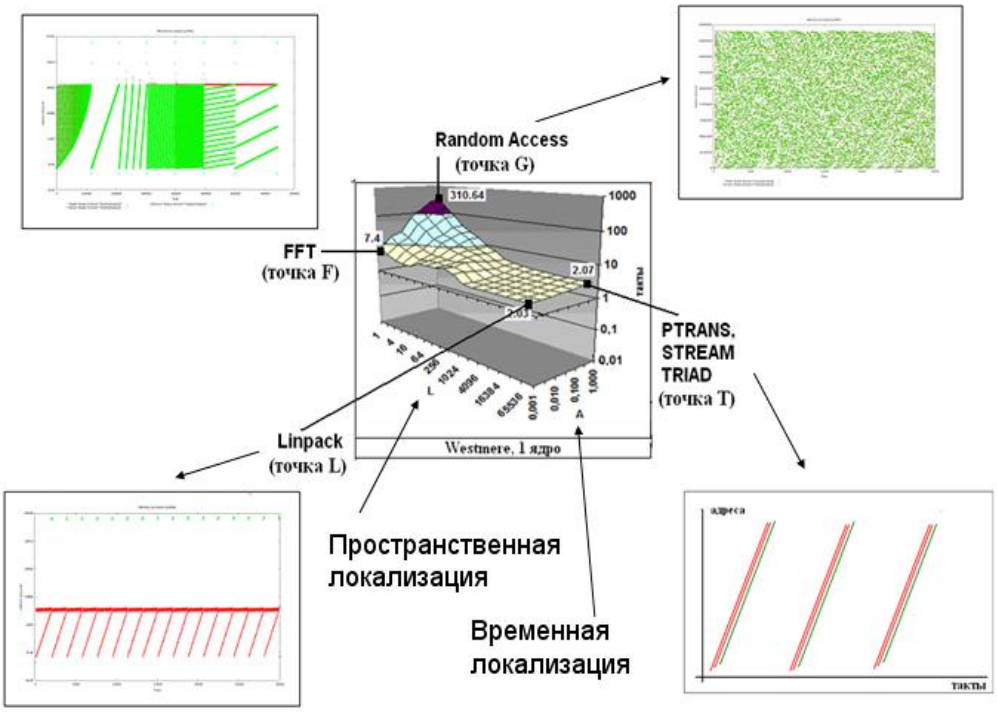



**Рис.3 - Характеристики подсистемы памяти для обращений с синтезируемой пространственно - временной локализацией на тесте APEX-map и реальные профили адресов обращений некоторых тестов в граничных точках.**

Существует методика перехода от вычисляемой пространственно-временной локализации к соответствующей ей синтезируемой пространственно-временной локализации теста APEX-map. Собственно говоря, по этой методике граничным точкам L, F, T и G были поставлены в соответствие граничные тесты производительности, обозначенные на рис. 3. Также для любой прикладной задачи можно найти соответствующую ей точку на APEX-поверхности, это один из ключевых моментов в технологии оптимизации работы суперкомпьютеров, развиваемой на предприятии. Эта методика также освоена и развивается на нашем предприятии. Пример такого

отображения представлен на рис.4. Поясним, как этим можно воспользоваться а процессе оптимизации работы суперкомпьютера.

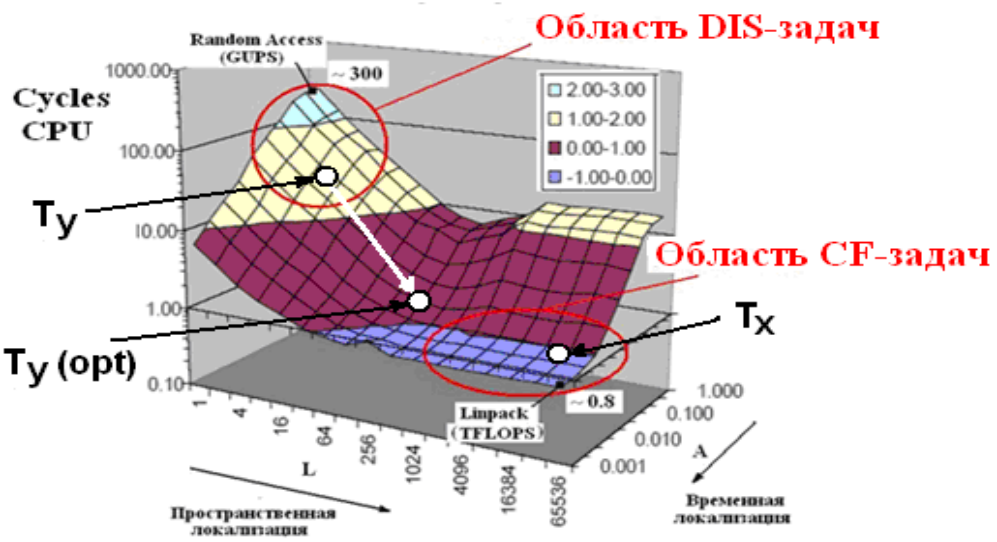



**Рис.4 Отображение прикладных задач на APEX-поверхность и варианты их оптимизации.**

Пусть подсистема памяти некоторого процессора характеризуется APEX-поверхностью, представленной на рис. 4. Пусть для некоторой задачи $T_X$ был снят профиль взаимодействия с памятью, по этому профилю была вычислена пространственно-временная локализация этой задачи, далее на APEX-поверхности найдена точка, которая по синтезируемой пространственно-временной локализации соответствует задаче $T_X$ . Эта точка на рис. 4 так и обозначена. Видно, что задача $T_X$ использует подсистему памяти в режиме хорошей пространственно-временной локализации (область CF-задач, дружественных к применению кэш-памяти), так что оптимизация этой задачи за счет оптимизации работы с памятью вряд ли возможна, поскольку память работает и так в хорошем режиме локализации.

Предположим теперь, что есть другая задача $T_Y$. Для этой задачи также был снят профиль взаимодействия с памятью, найдена точка на APEX-поверхности, соответствующая по синтезируемой локализации пространственно-временной локализации взаимодействия с памятью задачи $T_Y$ . Эта точка на рис.4 так и обозначена. Видно, что в данном случае подсистема памяти используется задачей $T_Y$ в режиме плохой пространственно-временной локализации (область DIS-задач, задач с интенсивной нерегулярной работой с памятью). Для повышения эффективности этой задачи можно попытаться с применением методов

прямой оптимизации улучшить режим ее пространственно-временной локализации и перейти в лучший режим, соответствующий точке $T_Y(opt)$. Часть таких методов применялась в организации при работе с появившимися на рынке многосокетными серверными платами. По линии перспективных работ предприятия поставлена цель систематизации этих и других методов такой прямой оптимизации, экспериментального их опробования.

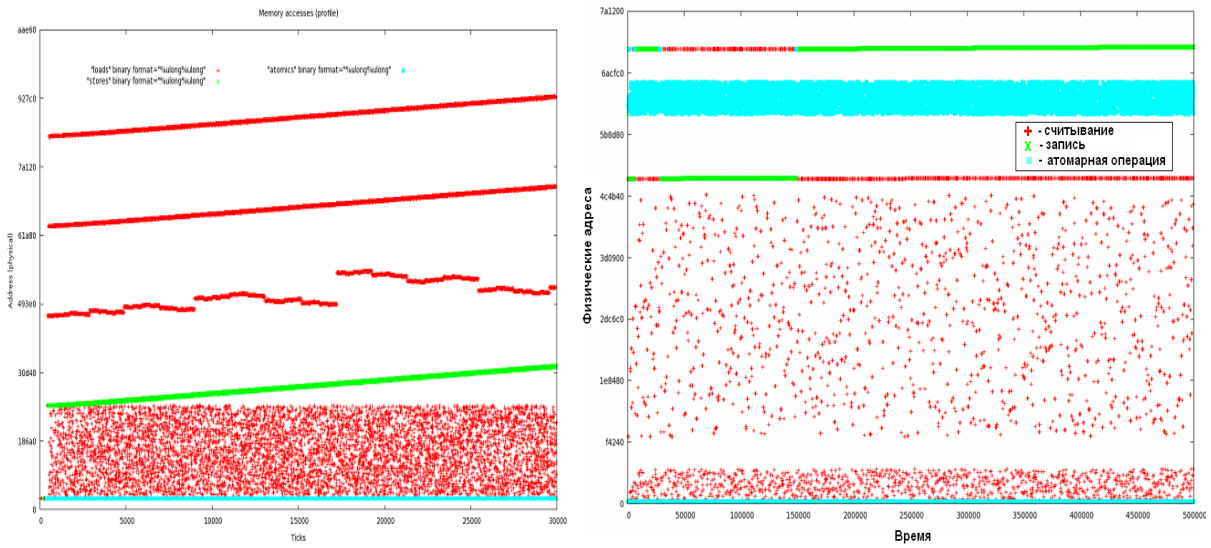
Методы прямой оптимизации за счет повышения пространственно-временной локализации не всегда применимы, т.е. если обратиться к рис. 4, то не всегда можно переместиться в точку $T_Y(opt)$ для задачи $T_Y$. Для оптимизации взаимодействия с памятью в этом случае применяется другая модель вычислений, которая обеспечивает сокращение среднего времени выполнения обращений к памяти (видимого пользователем) не за счет использования быстродействующих кэш-памятей, а за счет высокого темпа выполнения этих обращений. Такой прием называется обеспечением толерантности работы с памятью. Суть этого метода в том, чтобы поддерживать одновременное выполнение большого количества обращений к памяти. Обозначим это количество N. Тогда если выдача обращений и прием результатов их выполнения будет происходить с большим темпом $\Delta T$, то даже при больших временах обращений T, но таких больших N, что обеспечивается соотношение $T = \Delta T \times N$ (Правило Литтла), для приложения будет создаваться иллюзия работы с памятью со временем выполнения обращения $\Delta T$ .

Методы прямой оптимизации и обеспечения толерантности поддерживаются в рамках предложенной специалистами организации модели организации параллельных программ HPGAS, которую в определенном смысле можно считать производной от принципов работы массово-мультитредового суперкомпьютера с глобально адресуемой память. В настоящее время в виде перспективного направления работ предприятия намечена реализация модели HPGAS через схемы программной эмуляции на кластерных суперкомпьютерах. Принципиальная возможность такой схемы была недавно показана в работах специалистов Вашингтонского университета и экспериментально проверена специалистами предприятия. В процессе исследований предполагается также проверка возможностей использование при этом принципов, заложенных в организации процессоров и иерархической памяти российского проекта “Ангара”, а также результаты работ по мультитредовой платформе на базе легких q-тредов специалистов Ливерморской национальной лаборатории Лоуренса, Национальной

лаборатории Сандиа и Международной ассоциации университетов RENCI . Практическую проработку реализации предложенной модели HPGAS предполагается осуществлять на кластерах с сетью МВС-экспресс (www.kiam.ru/mvs/resourses/k100.html).

Проблема работы с памятью может решаться и кардинально, посредством принципиального резкого сокращения таких обращений в программах. Это достигается применением управляемых данными потоковых моделей вычислений, а также применением специализированных блоков для решения заданных задач. Этими направлениями на предприятии занимаются достаточно активно в плане разработки и использования ускорительных блоков на ПЛИС.

В заключение приводимых комментариев вернемся к вопросу оправданности постановки проблемы работы с памятью для современных и перспективных суперкомпьютеров. В приведенном тезисе 1 было декларировано, что целевые тесты, являющиеся частью технических заданий новых суперкомпьютеров, стали отличаться плохой пространственно-временной локализацией. Это напрямую относится к граничным тестам APEX-поверхности в точках L, F, T и G, представленной на рис.2. Эти тесты использовались в программе DARPA HPCS для задания контрольных показателей, они являются также тестами в профессиональном рейтинге HPC Challenge Class 1 Award. Отметим, что ранее в большей степени принимался во внимание тест Linpack, которому соответствует точка L с наилучшей пространственно-временной локализацией. Сейчас это не так, о чем говорит даже недавняя работа Дж.Донгара [3], признавшего неадекватность этого теста и соответствующего рейтинга Top500 современным приложениям, что вводит в заблуждение пользователей и разработчиков суперкомпьютеров. Вместо этого теста в [3] был предложен тест HPCG умножения разреженной матрицы на вектор, обычное обозначение SpMV. Это тест с плохой пространственно-временной локализации, развиваемый на нем уровень реальной производительности от порядка трех процентов и до десятых долей от пиковой. Это уже второй случай критики адекватности теста Linpack, ранее он критиковался в связи с введением рейтинга Graph500 на тесте BFS поиска вширь в графе. Профили этих новых тестов представлены на рис. 5.




**Тест SPMV** **Тест BFS**

**Рис.5 Профили обращений к памяти теста SPMV и теста BFS рейтинга Graph500.**

Сложность этих и современных целевых тестов с плохой пространственно-временной локализацией в том, что на них кэш-память процессоров работает неэффективно. Задержки выполнения одного обращения в память составляют, в целом, несколько сотен тактов процессора, а это значит, что процессор все это время будет простаивать, что и приводит к низкому уровню производительности в сравнении с пиковой – от единиц до десятых долей процента. Предлагаемые многошаговые технологии оптимизации работы суперкомпьютеров, которые уже применяются и развиваются на предприятии, нацелены именно на преодоление этой проблемы.

[1] Горбунов В., Эйсымонт Л. Комплексная методика тестирования производительности суперкомпьютеров, профессиональный подход. Журнал “Вычислительные методы и программирование”, 2013 (в печати)

[2] Горбунов В., Елизаров Г. Эйсымонт Л. Экзафлопсные суперкомпьютеры: достижения и перспективы. «Открытые системы», №7, 2013.

[3] Dongara J., Heroux M.A. Toward a New Metric for Ranking High Performance Computing Systems. Sandia Report SAND2013-4744, June 2013, 18 pp