КОМПЛЕКСНАЯ МЕТОДИКА ТЕСТИРОВАНИЯ ПРОИЗВОДИТЕЛЬНОСТИ СУПЕРКОМПЬЮТЕРОВ, ПРОФЕССИОНАЛЬНЫЙ ПОДХОД

В.С. Горбунов, Л.К. Эйсымонт

В работе [1] отмечалась важность оценочного тестирования производительности суперкомпьютеров, как общедоступного, рейтингового (Top500, Graph500, Green500, HPC Challenge Class 1 Awards), так и профессионального, разнопланового и углубленного (методики HPC Challenge, HPEC Challenge и др.). Были упомянуты появившиеся в начале 2000-х годов федеральные центры оценочного тестирования США и Китая, а также существовавший ранее при Минпромнауки РФ Центр независимого межведомственного тестирования суперкомпьютерных систем (ЦТСС), организованный в 2000-м году. В состав ЦТСС входили ведущие специалисты в данной области из ИПМ им. М. В. Келдыша РАН (инициатор создания ЦТСС), ФГУП “НИИ ”Квант”, ОАО ”НИЦЭВТ”, ИММ РАН, НИВЦ МГУ, ФГУП «РФЯЦ-ВНИИЭФ» (г.Саров), Межведомственного суперкомпьютерного центра РАН.

Предлагалось воссоздание ЦТСС в обновленном виде, сначала как рабочей группы (далее – Центр) ФГУП ”НИИ”Квант” и Санкт-Петербургского Государственного Политехнического Университета (СПбГПУ). Данная работа связана с темой оценочного тестирования, отражает важные современные тенденции и освещает некоторые работы Центра за последнее время.

1. О чем говорит динамика рейтинговых списков

Возникновение рейтингового теста – говорит об актуальности соответствующего класса приложений. Темп улучшения значений характеристик на тесте (рейтинговая динамика) говорит об интенсивности и направленности работ.

В последние годы стали популярны DIS-задачи (Data Intensive Systems), отличающиеся плохой пространственно-временной локализацией и непредсказуемостью адресов обращений к памяти, высокой интенсивностью операций с памятью в сравнении с вычислительными. На этих задачах неэффективно работает кэш-память, поэтому DIS-задачи еще называют как “недружественные” к ней. Класс DIS-задач представляют тесты: BFS (рейтинг Graph500); G-RandomAccess, G-FFT, EP-STREAM-Triad (три из четырех тестов рейтинга HPC Challenge Class 1 Awards). Противоположность DIS-задач - CF-задачи (Cache-Friendly, “дружественные” к кэш-памяти), имеющие хорошую пространственно-временную локализацию и предсказуемость адресов обращений к памяти, они представляются тестами: Linpack (рейтинг Top500), G-HPL (рейтинг HPC Challenge Class 1 Awards).

Таблица 1. Динамика первых пяти мест списка Top500

Top500		Место в рейтинге (Tflops)				
год	месяц	1	2	3	4	5
2010	июнь	1759.0	1271.0	1042.0	831.7	825.5
		XT5	Nebulae	Roadrunner	XT5	BG/P
	ноябрь	2566.0	1759.0	1271.0	1192.0	1054.0
		Tianhe-1A	XT5	Nebulae	Tsubame	XE6
2011	июнь	8162.0	2566.0	1759.0	1271.0	1192.0
		K-comp	Tianhe-1A	XT5	Nebulae	Tsubame
	ноябрь	10510.0	2566.0	1759.0	1271.0	1192.0
		K-comp	Tianhe-1A	XT5	Nebulae	Tsubame
2012	июнь	16324.8	10510.0	8162.0	2897.0	2566.0
		BG/Q	K-comp	BG/Q	IBM-cluster	Tianhe-1A
	ноябрь	17590.0	16324.8	10510.0	8162.4	4141.2
		XK7	BG/Q	K-comp	BG/Q	BG/Q
Динамика 2012/2005 (Nov, разы)		62.69	178.8	165.78	133.81	108.13
Динамика 2012/2010 (Nov, разы)		6,86	9.28	8.27	6.85	3.93

В таблице 1 дана динамика первых пяти мест списка Top500 с июня 2005 по ноябрь 2012 года. Видно, что показатели для первого места увеличились почти в 63 раза, следующие четыре места подтягиваются к лидеру, темп их улучшения более сотни. В сравнении с ноябрем 2010 ускорение первого места лишь 6.4 раза, для следующих четырех мест - не более десяти. В таблице 2 приведена динамика первых пяти мест списка Graph500. Если сравнивать показатели ноября 2012 года с ноябрем 2010 года, то по первому месту разница в 2330 раз (на Top500 – всего в 6.4 раза), для следующих четырех мест ускорение от 2000 до 4800 раз (в Top500 – 4-9 раз).

Таблица 2. Динамика первых пяти мест списка Graph500

Graph500		Место в рейтинге (GTEPS(scale))				
год	месяц	1	2	3	4	5
2010	ноябрь	**6.6** (36)	**5.22** (32)	**1.22** (29)	**1.17** (29)	**0.533** (29)
		BG/P	XT4	XMT	XMT	X5670+IB
2011	июнь	**43.47** (37)	**25.08** (37)	**19.96** (36)	**18.51** (38)	**18.42** (38)
		Lomonosov	XE6	XT4	BG/P	BG/P
	ноябрь	**253.4** (32)	**112.7** (37)	**103.1** (37)	**99.86** (36)	**92.34** (36)
		BG/Q	XE6	Lomonosov	Tsubame	BG/P
2012	июнь	**3541** (38)	**3541** (38)	**508.1** (35)	**358.1** (38)	**317.09** (35)
		BG/Q	BG/Q	p7-775	K-comp	HP-cluster
	ноябрь	**15383** (40)	**10461** (39)	**5848** (38)	**5524** (40)	**2567** (37)
		BG/Q	BG/Q	BG/Q	K-comp	BG/Q
Динамика 2012/2010 (разы)		**2330**	**2004**	**4793**	**4721**	**4816**

Такое отличие динамики списков Top500 и Graph500 является впечатляющим количественным подтверждением выводов аналитиков, что оптимизация суперкомпьютеров для выполнения DIS-задач в

настоящее время наиболее актуальна. Такие же выводы можно сделать, рассматривая характеристики на тестах рейтинга HPC Challenge Class 1 Award.

Таблица 3. Динамика первых двух мест на списка HPC Challenge Class 1 Award

HPCC Awards	G-HPL (TFlops)		G-RandomAccess (GUPS)		G-FFT (TFlops)		EP-Stream-Triad (Tbyte/s)	
(Nov)	1	2	1	2	1	2	1	2
2005	259	-	35	-	2.3	-	160	-
	BG/L	-	BG/L	-	BG/L	-	BG/L	-
2006	259	67	35	17	2.311	1.122	160	55
	BG/L	BG/L	BG/L	BG/L	BG/L	XT3	BG/L	p5-575
2007	259	94	35.5	33.6	2.870	2.311	160	77
	BG/L	XT3	BG/L	XT3	XT3	BG/L	BG/L	XT3
2008	902	259	103	35.5	5.080	2.870	330	160
	XT5	BG/L	BG/P	BG/L	BG/P	XT3	XT5	BG/L
2009	1533	736	117	103	11.0	8.0	398	267
	XT5	XT5	BG/P	BG/P	BG/P	BG/P	XT5	BG/P
2010	1533	736	117	103	11.88	11.7	398	267
	XT5	XT5	BG/P	BG/P	SX-9	XT5	XT5	BG/P
2011	2118	1533	121	117	34.7	11.88	812	398
	K-comp	XT5	K-comp	BG/P	K-comp	SX-9	K-comp	XT5
2012	9796	1533	2021	472	205.9	132.7	3857	525
	K-comp	XT5	p7-775	K-comp	K-comp	p7-775	K-comp	p7-775
Динамика 2012/2006 (разы)	37.8	22.88	57.75	7.62	89.52	118.27	24.01	4.27
Динамика 2012/2010 (разы)	6.4	2.08	17.27	4.58	17.33	11.34	9.69	1.97

В таблице 3 дана динамика первых двух мест списка HPC Challenge Class 1 Award. Наиболее сильная динамика на тестах G-FFT и G-RandomAccess. Однако она ниже, чем была на тесте BFS списка Graph500 (таблица 2), что объясняется более простой работой с памятью на тесте BFS.

Динамика трех рейтинговых списков демонстрирует актуальность характеристик, связанных с выполнением операций с памятью. Отметим, что такие характеристики были заложены в основу базовой методики оценочного тестирования, применяемой в Центре, которая будет рассмотрена далее. Однако сначала остановимся на разных схемах использования методик оценочного тестирования.

2. Схемы оценочного тестирования

Пусть A = {A1, A2 … AN} - это набор приложений, в эффективном выполнении которых на суперкомпьютере S заинтересован заказчик. Будем считать, что суперкомпьютер S имеет следующие компоненты: Hw – аппаратная часть; SwOS - операционная система; SwRT/LIB - системы поддержки выполнения программ (run-time системы) и библиотеки; SwC - компиляторы. Показателями эффективности выполнения приложений будет достижение некоторых значений характеристик Q={Q1, Q2, ... QM}.

Пусть также имеется набор тестов {TI, J}, где I – уровень теста, J – номер теста на уровне, который входит *в методику оценочного тестирования*, включающую дополнительно: порядок использования тестов, подходы к трактовке результатов на тестах, инструментальные средства обработки и накопления результатов, средства анализа, средства профилирования приложений. Рассмотрим часто применяемые схемы оценочного тестирования производительности.

Схема 1. Пропускается набор приложений A на суперкомпьютере S и фиксируются значения характеристик Q. На этом оценочное тестирование завершается. Если значения характеристик Q устраивают, то суперкомпьютер S оценивается положительно, иначе - отрицательно. Это самая простая схема.

Схема 2. Пропускается набор приложений A на суперкомпьютере S и фиксируются значения характеристик Q, но на этом работа не заканчивается, поскольку есть необходимость Q улучшить.

Обычно изначально пробуются разные варианты SwC, SwRT/LIB и SwOS. При этом удобнее использовать подмножества тестового набора {TI, J}, которые поставлены в соответствие имеющимся приложениям. Обозначим такие подмножества в виде AK ~ {TI, J (k)}. Эта схема демонстрирует “запоздалое” применение профессионального оценочного тестирования. К сожалению, бывает, что к нему прибегают, когда суперкомпьютер S уже приобретен.

Схема 3. Имеется суперкомпьютер S, но набора приложений A либо нет, либо они по какой-то причине малодоступны, подмножества AK ~ {TI, J (k)} есть или нет. Это типичная ситуация, когда суперкомпьютер делается по заказу и лишь по общей договоренности по его спецификациям.

В этом случае оценочное тестирование может вестись на тестовом наборе {TI, J}, это уже профессиональный уровень, результаты обсуждаются и согласуются с заказчиком, по ним проводится оптимизация S. Примеры такого тестирования и оптимизации даны в работе [7].

Схема 4. Суперкомпьютер S надо еще разработать, нет или мало приложений A, но общее видение их имеется, есть тестовый набор {TI, J}, на котором необходимо достичь определенных значений характеристик, которые задаются относительно достижимых на уже существующем эталонном суперкомпьютере. Подмножества AK ~ {TI, J (k)} для предполагаемых приложений A также создаются и согласуются. В этом случае набор {TI, J} и подмножества {TI, J (k)} выступают как часть технического задания на разработку S. Обычно в процессе такой работы {TI, J} и {TI, J (k)} дорабатываются. Например, сейчас есть необходимость доработки тестов применяемой в Центре методики с учетом ставших актуальными задач на динамических графах. В частности, представляет интерес использование формата STINGER хранения таких графов в глобально адресуемой памяти [12].

Такая схема применялась при разработке петафлопсных суперкомпьютеров программы DARPA HPCS и проекта Ангара [11], а сейчас - экзамасштабных суперкомпьютеров [13]. В этом случае работа ведется на имитационных моделях и макетах. Интересна обнаруженная недавно возможность применения при макетировании и даже создании опытных образцов кластерных суперкомпьютеров, поскольку огромное количество имеющихся в них процессорных ядер неожиданно эффективно и относительно дешево позволяет эмулировать перспективные архитектуры. Например, доказана возможность эффективной эмуляции массово-мультитредовых архитектур с глобально адресуемой памятью [10], а также возможность отработки перспективных операционных систем и систем поддержки выполнения программ [14].

При любой схеме главным результатом оценочного тестирования являются знания о работе оборудования и приложений. Известна крылатая фраза – “не можешь измерить – не можешь улучшить”.

3. Многоуровневая методика оценочного тестирования

Работа ЦТСС Минпромнауки РФ началась в свое время с создания собственной методики оценочного тестирования [15]. Уже тогда применялся способ оценки границ реальной производительности и ее сравнения с декларируемой пиковой: верхняя на задаче Linpack, нижняя – на задачах аэрогидродинамики тестового пакета NASA NPB. Это отражало разную работу с памятью в приложениях.

Далее методика была обобщена в ОАО ”НИЦЭВТ”[16, 17] и получила развитие в ФГУП ”НИИ”Квант”[1]. В основу положен принцип оценки “снизу вверх”, от основных элементов до вычислительной системы в целом. Считается, что главным элементом, влияющим на эффективность, является подсистема памяти. Методика отображает и степень детализации, с которой требуется получить знания о тестируемой системе — на нижних уровнях находятся тесты компонентов системы, а на верхних - пользовательские программы. Уровни тестирования можно выбирать в зависимости от применяемой схемы тестирования. Детализация тестовой нагрузки в методике делается целенаправленно, со знанием ожидаемых особенностей работы оборудования.

Первый уровень – оценка подсистемы памяти на тесте APEX-MAP, он искусственно меняет пространственно-временную локализацию адресов обращений и определяет среднее количество тактов на одно обращение. По результатам строится APEX-поверхность – зависимость тактов процессора на обращение от временной и пространственной локализации.

Второй уровень – граничные тесты, соответствуют четырем предельным значениям пространственно-временной локализации: Linpack (хорошая пространственная и временная локализация), Random Access (одновременно плохая пространственная и временная), PTRANS и TRIAD (плохая временная, хорошая пространственная), FFT (хорошая временная, плохая пространственная локализация).

Третий уровень – специально подобранные тесты для детального исследования оборудования, а именно: процессорных функциональных устройств выполнения арифметико-логических операций и операций с памятью, внутренней и внешней сети вычислительных узлов, системы ввода-вывода.

Подгруппа 3.1. Эффективность устройств выполнения операций с памятью при разных по сложности схемах доступа к памяти.

Подгруппа 3.2. Эффективность арифметико-логических устройств процессора на вычислениях разной интенсивности с разными операциями и операндами на фоне простейшей, наилучшей по пространственно-временной локализации схемы работы с памятью.

Подгруппа 3.3. Эффективность арифметико-логических устройств процессора на вычислениях разной интенсивности с разными операциями и операндами на фоне сложных схем доступа к памяти с плохой пространственно-временной локализацией.

Подгруппа 3.4. Тесты внутриузловой и межузловой коммуникационной сети.

Подгруппа 3.5. Комплексные тесты вычислительного узла.

Подгруппа 3.6. Тесты подсистемы ввода-вывода.

Четвертый уровень - общие и специальные базовые алгоритмы прикладных программ: стандартные математические функции, векторные операции, векторно-матричные операции (включая с разреженными матрицами), матричные операции, операции с битовыми матрицами, операции специальных преобразований при обработке сигналов, операции целочисленной арифметики многократной точности. Обычно из тестов этого уровня составляются подмножества {TI, J (k)}, которые сопоставляются с приложениями.

Пятый уровень - ядра разных приложений, а именно: научные расчеты (линейная алгебра и аэрогидродинамика); сжатие текстов и изображений; защита информации; обработка текстов и изображений; задачи оптимизации и поиска; задачи на высокорегулярных структурах (сеточные методы и клеточные алгоритмы); задачи на деревьях и графах; тесты разных моделей организации параллельных программ.

Шестой уровень – тесты модельных приложений, а также операционной системы и системы планирования прохождения заданий.

Седьмой уровень – тесты в виде прикладных программ, немного упрощенных для проведения тестирования.

Имеется автоматизированная система обработки файлов с результатами оценочного тестирования, а также база данных результатов с графическим интерфейсом для анализа полученных результатов, информационно-аналитическая база по тематике суперкомпьютеров и их оценочному тестированию.

4. Некоторые результаты и варианты их применения

Рассмотренная методика тестирования активно применялась в последние несколько лет, что отражено в достаточно большом количестве публикаций и выступлений. Объектами тестирования были:

- суперскалярные микропроцессоры от Intel и AMD;
- 64-ядерный микропроцессор фирмы Tilera;
- графические сопроцессоры фирм Nvidia и ATI;
- коммуникационные средства фирм Mellanox, Qlogic и Arista;
- высокореактивная сеть МВС-express собственной разработки;
- 2-х, 4-х и 8-ми сокетные серверные платы;
- фрагменты суперкомпьютеров с однорейловыми и многорейловыми коммуникационными сетями;
- фрагменты облачных суперкомпьютеров.

Среди недавних работ Центра можно выделить сравнительное оценочное тестирование Xeon Phi и Xeon Sandy Bridge [2] и работы [3, 4], которые рассмотрим подробнее, поскольку они демонстрируют возможность достаточно необычного использования получаемых при оценочном тестировании знаний для оптимизации вычислений на имеющемся оборудовании.

В работе [3] реализована методика измерения пространственной (SL) и временной (TL) локализации при выполнении программ на некотором процессоре. Для этого автоматически локализовались команды работы с памятью в битовом коде приложения и проводилась трассировка адресов обращений к памяти в этих командах. По таким трассам адресов в соответствии с методикой [5] вычислялась измеряемая локализация <SL,TL> как всего приложения, так и задаваемых отдельных команд, реализована возможность выборочной выдачи профилей обращений к памяти в координатах “адреса-время”.

В соответствии с [5] был еще реализован способ перехода от измеряемой <SL,TL> к точке <SL*,TL*> APEX-поверхности [6] подсистемы памяти используемого приложением процессора. Переход к <SL*,TL*> позволяет увидеть, в каком режиме локализации используется подсистема памяти на этом приложении. Такое знание о приложении можно использовать для его оптимизации посредством улучшения локализации, если это требуется и возможно, но имеется еще один вариант, он связан с результатами второй работы [4], суть его в следующем.

В работе [4] приведены результаты оценки эффективности одновременного выполнения множества операций с памятью с плохой пространственно-временной локализацией в многоядерных суперскалярных процессорах, а также эффективности способов программной эмуляции массово-мультитредовых архитектур посредством корутин на суперскалярных процессорах. Цель такого исследования — оценка возможности разработки в будущем для кластерных суперкомпьютеров с многоядерными суперскалярными микропроцессорами некоторой системы поддержки выполнения программ (SwRT/LIB), в которой за счет резкого увеличения одновременно выполняемых операций с памятью (как и в массово-мультитредовых архитектурах) обеспечивается толерантность к задержкам выполнения операций с памятью. В работе [10] экспериментально показывается, что такой подход не сильно проигрывает при работе с локальной памятью вычислительного узла и окажется более эффективным при реализации операций с глобально адресуемой памятью, отображаемой на физические памяти множества узлов. В работе [9] идея использования программной эмуляции была обобщена, а в работе [4] началась экспериментальная проверка результатов работы [10], что открывает путь к реализации планов работы [9].

Если такая реализация окажется успешной, то в битовом коде оптимизируемого приложения можно будет найти обращения к памяти с плохой пространственно-временной локализацией и заменить их на обращения к памяти, реализуемые через систему эмуляции массово-мультитредовой архитектуры. Это должно повысить толерантность приложения к задержкам выполнения операций с памятью, предоставив одновременно

возможность комфортной работы с огромной глобально адресуемой памятью, отображаемой на физические памяти множества узлов. Ставится цель повышения за счет этого эффективности DIS-приложений на кластерных суперкомпьютерах в 5-10 раз. Таким образом, можно было бы решить давнюю проблему эффективного выполнения OpenMP программ на множестве серверных плат, а не на одной, как это возможно сейчас [7] или на специализированном суперкомпьютре [8].

Заключение

Оценочное тестирование и применение ее результатов – это огромная экспериментальная работа по изучению сложных объектов и освоению полученных знаний. В силу этого в ней целесообразно участие многих специалистов и организаций, она должна быть открытым проектом федерального уровня. Это мнение будет предложено для обсуждения на дискуссии “Нужна ли национальная система оценки суперкомпьютеров?”, она должна состояться в рамках Четвертого Московского Суперкомпьютерного Форума этого года (МСКФ-2013).

ЛИТЕРАТУРА:

1. В.С.Горбунов, Л.К.Эйсымонт, А.В.Речинский, В.С.Заборовский, П.В.Забеднов. Суперкомьютеры для промышленности – вопросы тестирования, анализа и разработки. Материалы 2-й Всесоюзной конференции “Суперкомпьютерные технологии” (СКТ-2012), стр.360-364.
2. А.В.Мищенков, С.А.Фёдоров, Д.В.Андрюшин. Оценочное тестирование Intel Xeon Phi как массово-многоядерной SMP-системы. Научно-технический семинар «Высокопроизводительные и встраиваемые вычисления» (в рамках молодежной научной конференции «Студенты и молодые ученые — инновационной России», 23-24 мая 2013 года), СПбГПУ, презентация.
3. В.И.Максимов, С.А.Фёдоров, Д.В.Андрюшин. Реализация методики измерения пространственно-временной локализации приложения и её отображение на синтезируемую APEX-поверхность используемого оборудования. Презентация, см. [2].
4. В.А.Макаров, С.А.Фёдоров, Л.К.Эйсымонт. Эффективность многофазных операций считывания/записи для глобально адресуемых систем памяти. Презентация, см.[2].
5. J.Weinberg et al. Quantifying Locality In The Memory Access Patterns of HPC Applications. SC’05, 12 pp.
6. E.Strohmaier H.Shan. Apex-Map: A Global Data Access Benchmark to Analyze HPC Systems and Parallel Programming Paradigms. SC’05, 14pp.
7. А.Речинский, В.Горбунов, Л.Эйсымонт, Суперкластер с глобально адресуемой памятью//Открытые системы, №7, 2011.
8. J.Multby. Graph Oriented Computing and XMT Architecture. Cray Inc. 26 slides. Climate Knowledge Discovery Workshop, March 2011, Hamburg.
9. В.С.Горбунов, Л.К.Эйсымонт, А.А.Соколов, А.В.Зайцев, В.С.Заборовский, В.Б.Семеновский. Экзамасштабные технологии: инкапсуляция иерархической структуры суперкомпьютеров в модели HPGAS. см. [1], с. 29-34.
10. J.Nelson et al. Crunching Large Graphs with Commodity Processors. USENIX HotPar 2011, 6 pp.
11. А.А.Семенов, А.А.Соколов., Л.К.Эйсымонт. Архитектура глобально адресуемой памяти мультитредово-потокового суперкомпьютера. Журнал «Электроника: НТБ», №1, 2009 г., с. 50-56.
12. Ediger D., McColl R.M., Reidy J., Bader D.A. STINGER: High Performance Data Structure for Streaming Graphs. 2012, 5 pp.
13. В.Горбунов, Л.Эйсымонт. Экзафлопсный барьер: проблемы и решения //Открытые системы, №6, 2010, с. 12-15.
14. US Department of Energy Exascale Operating System and Runtime Software Report, December 28, 2012, 52 pp.
15. Л.К.Эйсымонт. Оценочное тестирование высокопроизводительных систем: цели, методы, результаты и выводы. "Суперкомпьютерные вычислительно-информационные технологии в физических и химических исследованиях" (30 октября - 2 ноября), Сборник лекций, Черноголовка, 2000, с.33-42.
16. М.Кудрявцев, Эйсымонт Л., Мошкин Д., Полунин М. Суперкластеры — между прошлым и будущим // Открытые системы, №8, 2008.
17. М.В.Кудрявцев, Д.В.Мошкин, М.А.Полунин, Л.К.Эйсымонт. Оценочное тестирование кластеров на базе процессоров AMD Barcelona и Shanghai с сетями Infiniband DDR и QDR. Журнал «Вычислительные методы и программирование»,том10,№1, 2009, с. 215-223.