Копия текста публикации со страницы
http://www.pcweek.ru/themes/detail.php?ID=112308

PC Week/RE №27-28 (633-634) 22 июля — 11 августа 2008

Рейтинг Top 500. Соревнование с гандикапом

Автор: Денис Воейков
17.07.2008

Если верить статистическим данным, то отечественная суперкомпьютерная отрасль сейчас находится на небывалом подъеме. Подъем этот можно отслеживать по разным показателям, однако общепринятой практикой является ориентация на международный рейтинг Top 500, где российские (установленные на территории нашей страны) машины в последнее время занимают все больше строчек, причем отнюдь не в самом хвосте.

Казалось бы, ничто не могло омрачить эту оптимистичную картину. Однако в силу того, что Россия в рейтинге представлена сразу несколькими производителями, конкурирующими между собой и соревнующимися за более высокие позиции, с некоторых пор весьма актуальной стала проблема разночтений в трактовке правил ведения Top 500. Как выяснилось, у российских участников рынка нет единого их понимания, что вызывает жаркие споры и взаимные упреки.

Попытка редакции разобраться в том, что же именно неоднозначно интерпретируется в правилах и какие претензии конкурентов друг к другу обоснованны, а какие нет, привела к совершенно неожиданным результатам.

Где тестировать?

Рейтинг пятисот мощнейших машин мира традиционно составляется два раза в год (в ноябре и июне) ведущими экспертами США из Государственного научно-исследовательского вычислительного центра министерства энергетики, а также из университетов Мангейма и Теннеси. Строчки в Top 500 распределяются в соответствии с тем, какую производительность продемонстрировали системы при прохождении специального стандартного теста Linpack.

Одним из первых неоднозначных нюансов, с которыми сталкиваешься при изучении порядка подачи заявок в Top 500, является тот факт, что в правилах в явном виде не оговорено, где именно должен тестироваться суперкомпьютер. Дело в том, что далеко не все машины сразу собираются на месте их будущей работы. Весьма широко распространена практика сборки и настройки вычислительной системы на заводе производителя, после чего суперкомпьютер разбирается и отсылается заказчику, который заново его монтирует и настраивает. При такой схеме вполне естественно, что обе стороны склонны прогонять Linpack-тест прямо на заводе, чтобы заранее определить производительность. И вот тут возникает резонный вопрос: нужно ли заново проводить тест на площадке заказчика, или для рейтинга достаточно первоначальных “заводских” данных?

Как показало общение с игроками рынка и непосредственными владельцами машин, мнения на этот счет сильно расходятся. Для всех очевидно, что повторная настройка суперкомпьютера отнюдь не гарантирует тех же самых результатов теста. Процесс этот многоступенчатый, и к моменту официальной публикации Top 500 производительность машины по Linpack может ощутимо измениться как в большую, так и в меньшую сторону. За примерами далеко ходить не надо. Из переписки с представителем Межведомственного суперкомпьютерного центра (МСЦ) РАН в Москве, где расположилась система МВС-100К (56-я строка списка Top 500), выяснилось, что к моменту выхода предыдущей (осенней) редакции рейтинга их машина демонстрировала “чуть худшие” результаты, чем те, что были поданы в Top 500 производителем (компанией Hewlett-Packard) на основании заводских испытаний.

С одной стороны, сложно не прислушаться к тем, кто, апеллируя к здравому смыслу, настаивает, что суперкомпьютеры создаются отнюдь не для прогонки тестов Linpack. По мнению представителей Hewlett-Packard, абсолютно естественно, что настройки системы могут различаться на этапах нагрузочного тестирования (во время опытной эксплуатации), тестирования с целью получения максимальных результатов Linpack и в период промышленной эксплуатации. Как отмечают в компании, тестирование Linpack предполагает долгий итерационный процесс тюнинга параметров системы и её конфигурации. Что характерно, речь идет не о повышении эффективности системы как таковой, а всего лишь о подборе оптимальных параметров, чтобы получить максимальные результаты при исполнении конкретной программы. Машина, показавшая при аналогичной архитектуре более высокую эффективность на Linpack, вовсе не обязательно лучше. Просто на специальную настройку было затрачено больше времени.

Кроме того, по свидетельству специалистов HP, для достижения максимальных Linpack-результатов систему часто приводят в состояние, в котором ее эффективность в целом существенно снижается. Очевидный пример — работающие на узлах средства мониторинга. В погоне за Linpack-показателями их, как правило, отключают, хотя реальная эксплуатация без них не бывает: никому не нужна производительность без должного уровня мониторинга и управляемости системы.

Однако нельзя, наверное, сбрасывать со счетов и мнение тех, кто склонен трактовать правила Top 500 иначе и ратует за предельную адекватность рейтинга реальности. Если формальный по своей природе рейтинг оперирует параметром работоспособности, достижение предельного значения которого слишком обременительно для владельца машины, то, может, стоит просто отказаться от гонки за строчками в Top 500? В общем, существует категория специалистов, считающая, что в тот промежуток времени, когда система уже установлена у заказчика, но свежая редакция Top 500 еще не опубликована, суперкомпьютер должен находиться в полной готовности и быть способным повторить тест, продемонстрировав заявленные ранее результаты.

Чтобы разрешить этот спор, мы обратились непосредственно к одному из главных кураторов рейтинга Джеку Донгарре. Ему были заданы два конкретных вопроса: правда ли, что достаточно провести тест Linpack на заводе, не повторяя его на месте окончательной установки компьютера, и подразумевается ли при этом, что к моменту выхода Top 500 заявленный компьютер обладает теми же настройками, какие были у него на заводе? В ответ заокеанский эксперт сообщил, что это правда, но только в том случае, если система будет размещена на официальной площадке в пределах двух месяцев (по всей видимости — после проведения заводского теста).

Остается лишь гадать, почему речь идет именно о двух месяцах и по какой причине это правило не является общественным достоянием (не опубликовано в открытом доступе на сайте Top 500). Тем не менее факт остается фактом: если при монтаже системы заказчик проявит расторопность, то комитет не настаивает на перепроверке теста.

Перенос данных

Следующий нюанс составления рейтинга, заслуживший неоднозначную оценку участников рынка, — это легитимность произвольного переноса результатов теста от одной схожей машины на другую. С точки зрения сотрудников компании IBM, в соответствии с принятой практикой при подаче заявок в Top 500 позволено использовать ранее полученные данные для кластерных систем, построенных на той же архитектуре, если стандартный тест производительности Linpack к тому времени еще не был выполнен на заявляемой системе и при условии, что эта последняя является заведомо более мощной (быстродействующей) без модификаций и интерполяций. В дальнейшем консервативная оценка производительности Rmax уточняется в последующих редакциях списка Top 500.

Активным противником законности такого подхода выступает директор Института программных систем РАН, научный руководитель суперкомпьютерной программы "СКИФ-ГРИД" Сергей Абрамов. Он отмечает, что IBM в данном случае ссылается на некий подпункт в разделе правил администрирования Top 500, но трактует его неверно. С его точки зрения, в упомянутом параграфе речь идет не о том, как нужно (и можно) подавать данные для рейтинга, а о том, как модераторам приходится интерпретировать неполные заявки. Если комитет по какой-либо причине не получает сведений о производительности машины, то в редких случаях ("in a few cases") этот параметр переносится от такой же, но менее мощной модели.

Реальность такова, что комитет Top 500, по всей видимости, спокойно относится к переносам, практикуемым IBM. По подсчетам г-на Абрамова, в осенней редакции таковых было 19, и все они касались машин именно этой компании — остальных производителей (Sun Microsystems, NEC, SGI, Hitachi, Linux Networx, Dell, Cray, Fujitsu, Dawning и пр.) это не касается. Однако такая лояльность, как мы сейчас увидим, в известной степени вредит репутации рейтинга по той причине, что отдельные его фрагменты выглядят просто абсурдно и провоцируют общественность на обсуждение различного рода домыслов.

Так, несколько месяцев назад Сергей Абрамов подробно проанализировал фрагмент ноябрьского (последнего на тот момент) рейтинга, приведенный нами в таблице. В таблице мы убрали колонки "континент" и "регион", а также характеристики, одинаковые для всех машин: производитель — IBM, семейство процессоров — Intel EM64T, процессор — Intel EM64T Xeon 53xx (Clovertown), тактовая частота процессора — 2333 МГц, семейство суперЭВМ — IBM Cluster, ОС — Linux, архитектура — Cluster.

Во всех семи строчках последовательно расположились машины IBM одного семейства с одинаковыми процессорами равной тактовой частоты, с идентичными архитектурой и ОС. Различен их так называемый интерконнект и платформы, однако, как указывает г-н Абрамов, для результатов Linpack-теста это несущественно. Общеизвестно, что если компьютеры нормально отлажены, то они должны иметь приблизительно одинаковые соотношения Linpack- и пиковой производительностей (КПД), а сами эти параметры — линейно зависеть от числа процессоров. (Погрешность настройки, по утверждению г-на Абрамова, не может превышать 3%.)

Что же мы видим в нашей таблице (колонки с 9-й по 11-ю)? Число процессоров у всех машин разное (различие между старшей и младшей составляет 67%), а вот Linpack-производительность почему-то везде одинаковая — 9287 Гфлопс.

Если представить себе, что эти данные объективны, то придется признать, что за исключением Тихоокеанской северо-западной национальной лаборатории (США) все остальные организации приобрели десятки и сотни лишних процессоров.

Как мы уже упоминали выше, подобные аномалии объясняются тем, что компания-производитель может воспользоваться поправкой к правилам подачи заявок в Top 500. Но Сергей Абрамов считает такую практику порочной, полагая что при злоупотреблениях переносами (без прогонки реального теста) какие-то позиции рейтинга обязательно окажутся "мертвыми душами" — часть машин просто не будет собрана и настроена на местах к положенному сроку.

С учетом того, что две из семи упомянутых машин располагаются на территории нашей страны, редакция посчитала небезынтересным уточнить реальные сроки монтажа оборудования в российских вузах и сравнить их с описаниями "дедлайнов" в правилах Top 500.

Сроки инсталяции

На официальном сайте рейтинга в соответствующем разделе четко говорится: для того чтобы претендовать на место в осенней редакции списка, необходимо подать заявку с протоколом теста Linpack до 1 октября и не позднее 1 ноября запустить машину на постоянном месте будущей работы ("All system reported have to be installed by November 1st"). С учетом этих положений мы попросили представителей Уфимского государственного авиационного технического университета (УГАТУ) и Сибирского федерального университета (СФУ) ответить на несколько вопросов: когда в каждом конкретном случае была установлена и настроена машина (до 1 ноября 2007 г. или после)? проводился ли на ней Linpack-тест? (какие дал результаты?) и кто подавал заявку в Top 500?

Как сообщил проректор по информатизации УГАТУ Руслан Хисамутдинов, его вуз заявку на рейтинг не подавал, так как к началу ноября подавать было еще просто нечего. Тесты Linpack начали проводиться только в середине января и продемонстрировали 15,8 Тфлопс. О вхождении в Top 500 администрация университета узнала в день выхода рейтинга, что стало для нее полной неожиданностью.

Из Сибири ответил научный руководитель Института космических и информационных технологий СФУ, член-корреспондент РАН Владимир Шайдуров. С его слов, суперкомпьютер в их учебно-лабораторный корпус был доставлен и установлен к началу декабря. О том, кем была подана заявка в Top 500, г-н Шайдуров не сообщил. Попытка получить комментарии у IBM закончилась неудачей – на конкретно поставленные нами вопросы представители компании не ответили.

В сложившейся ситуации для нас стала очевидной необходимость разузнать судьбу всех прочих суперЭВМ из ноябрьского (последнего на тот момент) рейтинга, которые были установлены в российских организациях (благо таковых осталось всего пять).

Первой на очереди оказалась машина HP, расположившаяся в Межведомственном суперкомпьютерном центре (МСЦ) РАН в Москве, — МВС-100К, 33-я строка осеннего

списка Top 500 за 2007 г. Один из ведущих сотрудников МСЦ, пожелавший впоследствии остаться неназванным, о конкретных сроках установки суперкомпьютера умолчал, но при этом поставил под сомнение “дедлайн” 1 ноября и отослал интервьюера читать правила Top 500.

Тут надо отметить, что в отношении сроков сборки этой машины “недоброжелатели” уже давно указывали на весьма существенную странность. Дело в том, что в отличие от общепринятой практики хвастаться достижениями мирового уровня в данном случае первых журналистов допустили к суперкомпьютеру только 27 февраля 2008 г.(!).

Второй суперкомпьютер (HP, № 242 осеннего 2007-го Top 500) расположился в Вятском государственном университете (ВятГУ) в Кирове. В телефонном разговоре не представившаяся дама со ссылкой на начальника отдела телекоммуникаций и ИТ университета Илью Карпасова заверила нас, что к 1 ноября система была работоспособна, тест Linpack проводился именно на ней (причем как на заводе, так и в стенах вуза) и заявку подал сам ВятГУ. Причину проведения официального открытия только 25 января 2008 г. она объяснила тем, что презентацию решили приурочить к выездному заседанию местного правительства.

Из общения с пресс-службой Hewlett-Packard также стало ясно, что по официальным данным компании обе рейтинговые машины к 1 ноября стояли на своих местах.

Относительно системы “СКИФ Cyberia” (№ 199 осеннего 2007-го Top 500), находящейся в Томском государственном университете (ТГУ), комментарии были даны с косвенной ссылкой на руководителя центра кластерных технологий компании “Т-Платформы” Андрея Слепухина, который и подавал заявку в Top 500. С этой машиной все просто. Во-первых, она вошла в рейтинг еще в июне прошлого года (дата установки — 10 февраля, дата теста Linpack — 15 февраля) и в ноябрьскую редакцию перекочевала автоматически. А во-вторых, к поставке этого суперкомпьютера имеет непосредственное отношение Сергей Абрамов, который, как мы знаем, является ярым сторонником формального соблюдения правил Top 500.

А вот следующая машина заслуживает отдельного разговора. Речь еще об одной установке МСЦ РАН — MVS-15000BM (IBM, № 408 осеннего 2007-го Top 500), попавшей в рейтинг аж в 2006 г. Дело в том, что по данным из открытых источников этот кластер недавно был разделен на пять частей, которые по отдельности разошлись по учреждениям Москвы, Санкт-Петербурга, Казани, Владивостока и Черноголовки. Возникает резонный вопрос, когда же именно случилось дробление, означающее, что единая система, способная фигурировать в рейтинге, прекратила своё существование. Обращаемся к архивам новостных лент и выясняем, что, по всей видимости, как минимум 84 из 1148 процессоров начали свою работу в Институте автоматики и процессов управления Дальневосточного отделения РАН уже 25 октября 2007 г.(!) Комментарии представителей IBM и МСЦ РАН на этот счет получить не удалось.

На 430-й строчке прошлого рейтинга присутствовала еще одна российская система производства IBM, однако каких-либо открытых данных о ней нет. Принадлежит машина некой коммерческой “индустриальной компании”, очевидно, пожелавшей остаться в тени.

Мы посчитали нужным обратиться за комментариями все к тому же Джеку Донгарре. И, как выяснилось, не зря. Теперь остается только гадать, как бы отреагировал г-н Донгарра на наше письмо, приди оно на месяц раньше. Но, к сожалению, к тому моменту, когда все сведения о российских суперЭВМ из ноябрьского Top 500 были собраны, уже наступил

июнь и вышла новая редакция рейтинга. Это и дало г-ну Донгарре повод ответить в том смысле, что если ошибки и были, они все устранены в новом рейтинге.

Ну что ж, июньский, так июньский.

На верхней строчке в нем расположился вычислительный комплекс IBM Roadrunner, созданный для Лос-Аламосской национальной лаборатории министерства энергетики США. Это первая машина, сумевшая перешагнуть петафлопсный рубеж (квадриллион операций с плавающей запятой в секунду), поэтому ее выхода компьютерное сообщество ожидало с огромным нетерпением, и завершение теста Linpack широко освещалось в мировой прессе. Судя по западным публикациям, Roadrunner разогнался до петафлопса в первых числах июня, а его поставка в лабораторию была намечена на август. Заглядываем в правила Top 500 и видим там условия вхождения в июньский рейтинг: необходимо подать заявку с протоколом теста до 15 апреля и не позднее 15 мая запустить машину на постоянном месте работы.

За всеми позициями пятисотстрочного рейтинга комитет Top 500, возможно, уследить не в состоянии, но в то, что заокеанские эксперты не знакомы со всеми аспектами создания Roadrunner’a, поверить решительно невозможно. Вопрос редакции Джеку Донгарре выглядел следующим образом: судя по публикациям в СМИ, суперкомпьютер Roadrunner явно нарушает требование по срокам установки у заказчика. Верно ли, что этот случай является единственным исключением из указанного правила, сделанным из-за важности события — взятия барьера в 1 петафлопс?

Внимание, ответ! “Нет, правило двух месяцев применялось и в прошлом”. Итак, вхождение разогнанного в июне суперкомпьютера в рейтинг, заявка на участие в котором должна была быть подана до 15 апреля, с точки зрения комитета Top 500 также оправдывается загадочным правилом двух месяцев.

Очевидно, о петафлопсном суперкомпьютере однозначно знают и западные конкуренты IBM. Однако раз рейтинг до сих пор не изменился под их гневным давлением, значит, и их всё вполне устраивает. То есть правило "двух месяцев" (чем бы оно ни было) фактически признается и ими. Остается, однако, совершенно непонятно, зачем комитету Top 500 понадобилось формулировать весьма жесткие условия подачи заявок если на деле они подменяются мало кому известным правилом "двух месяцев".

Фрагмент мирового рейтинга Top 500 (ноябрь 2007 г.)

Место	Страна	Организация	Сегмент	Применение	Компьютер	Interconnect	Число процессоров	Пиковая производительность	Linpack-производительность	КПД (Linpack / пиковая)	Размер задачи (Nmax)
1	2	3	4	5	6	7	8	9	10	11	12
180	Italy	ENEA	Research	Geophysics	BladeCenter HS21 Cluster, Xeon quad core 2,33 GHz, Infiniband	Infiniband	2560	23889,9	9287	38,87%	0
181	Russia	Ufa State Aviation Technical University	Academic	Aerospace	BladeCenter HS21 Cluster, Xeon quad core 2,33 GHz, Infiniband	Infiniband	2128	19858,5	9287	46,77%	0
182	United Kingdom	Finance L	Industry	Finance	BladeCenter HS21 Cluster, Xeon quad core 2,33 GHz, GigEthernet	Gigabit Ethernet	1904	17768,1	9287	52,27%	0
183	Russia	Siberian National University	Academic	Not Specified	BladeCenter HS21 Cluster, Xeon quad core 2,33 GHz, Infiniband	Infiniband	1808	16872,3	9287	55,04%	0
184	France	Financial Institution (P)	Industry	Finance	BladeCenter HS21 Cluster, Xeon quad core 2,33 GHz, GigEthernet	Gigabit Ethernet	1680	15677,8	9287	59,24%	0
185	United States	Research	Research	Research	xSeries x3550 Cluster Xeon quad core, 2,333 GHz, GigEthernet	Gigabit Ethernet	1560	14557,9	9287	63,79%	0
186	United States	Pacific Northwest National Laboratory	Research	Not Specified	spray cooled xSeries x3550 Cluster Xeon quad core, 2,333 GHz, GigEthernet	Gigabit Ethernet	1536	14539	9287	63,88%	616000