На пути к вычислениям экзафлопсного класса

В. Б. Бетелин,
д. ф.-м. н., профессор,
академик РАН, директор ФНЦ НИИСИ
betelin@niisi.msk.ru

1. Суперкомпьютеры тера- и петафлопсного класса на основе массовых коммерческих технологий

До начала 1990-х гг. суперкомпьютер представлял собой уникальное, дорогостоящее, штучное изделие, создаваемое на основе последних технологических достижений в области микроэлектроники, электроники, межсоединений элементов и т. д., с целью максимального увеличения производительности дорогого единичного центрального процессора обработки данных. В начале 1990-х гг. в США была сформулирована идея увеличения производительности компьютера не за счет увеличения производительности единичного дорогого, уникального процессора, а за счет использования большого количества параллельно работающих дешевых коммерческих микропроцессоров, объединенных коммуникационной сетью. Эта идея была успешно апробирована при создании в 1993 г. компьютера Intel Paragon и далее была положена министерством энергетики США в основу концепции программы ASCI — программы сохранения боеспособности ядерных арсеналов США с помощью суперкомпьютерного моделирования. В рамках этой программы ведущими компаниями США (Intel, IBM, HP, Sun и т. д.) на основе государственного финансирования и были разработаны массовые коммерческие технологии создания суперкомпьютеров, представляющих собой масштабируемую сеть из десятков и сотен тысяч параллельно работающих коммерческих микропроцессоров, соединенных с помощью коммерческих коммуникационных СБИС. Одновременно национальными лабораториями США была созданы технологии создания масштабируемых приложений для таких суперкомпьютеров. На основе этих технологий в течение 12 лет были созданы суперЭВМ с производительностью от единиц терафлопс (1997 г.) до тысяч терафлопс (2009 г.). Эти суперкомпьютеры обеспечили решение предусмотренных программой ASCI конкретных задач по поддержанию и развитию ядерных боеприпасов на основе суперкомпьютерных технологий. Существенно важно, что в технических заданиях на создание в рамках ASCI всего ряда суперкомпьютеров ключевым требованием являлась их производительность на семи тестовых задачах. Так, например, суперкомпьютер ASCI BLUE должен был в течение одного часа поддерживать производительность один терафлопс на тесте SPPM (моделирования нестационарного гидродинамического процесса).

Стратегически важным для экономики США коммерческим результатом программы ASCI является формирование за последние 15 лет новой ниши мирового ИТ-рынка — высокопроизводительных систем обработки данных (ВСОД, англ. HPC — High Performance Computing) и их аппаратных и программных компонент (вычислительные и коммуникационные узлы, конструктивы, соединители и т. д.), т. е. не массовых, дорогостоящих (по сравнению с персональными компьютерами) систем обработки данных на основе ключевых массовых коммерческих технологий. Безусловными технологическими лидерами и мирового ИТ-рынка, и этой новой его ниши, являются компании США (Intel, AMD, IBM, HP, SGI, CRAY, Sun и т. д.), которые создают и развивают ключевые для этого рынка (и его новой ниши) массовые технологии проектирования и производства коммерческих микропроцессоров и коммуникационных СБИС, чипов оперативной памяти, межсоединений элементов и т. д., а также обеспечивают их разработку и массовый выпуск. В результате этой деятельности компаний-лидеров с 1997 г. по настоящее время

стоимость одного терафлопса для суперкомпьютеров тера- и петафлопсного класса на основе массовых коммерческих вычислительных и коммуникационных узлов, снизилась в тысячи раз — с $60 млн в 1997 г. до $10-20 тыс. в 2014 г.

Однако уже сейчас есть целый спектр невоенных, коммерческих задач, которые не могут быть решены на основе использования суперкомпьютеров тера- и петафлопсного класса и которые, предположительно, могут быть решены в полном объеме только на суперкомпьютерах экзафлопсного класса. К числу таких задач, по мнению Департамента энергетики США, относится, например, совершенствование методов разведки и добычи углеводородов. В связи с этим необходимо отметить, что в 2013 г. французский нефтяной гигант Тотал приобрел Пангеа — самый мощный суперкомпьютер в мире, принадлежащий частной компании. Его пиковая производительность составляет 2,3 Пфлопс, а стоимость — $77,8 млн с рассрочкой на четыре года [1, 2]. Производитель Пангеа — компания SGI. Основное назначение этого суперкомпьютера — поиски нефти и газа. Агентство Рейтер сообщает, что компания Бритиш Петролеум разворачивает суперкомпьютер производительностью 2 Пфлопс с теми же целями.

Еще одним примером коммерческой задачи, для решения которой требуется суперкомпьютер экзафлопсного класса, являются финансовые спекуляции на валютных рынках, производимые в реальном масштабе времени с временем ответа порядка секунд и долей секунд. В 2014 г. бизнесмен Джон Фитцпатрик организовал компанию, цель которой — создание экзафлопсного компьютера для решения этих задач на основе микропроцессоров компании Intel. Мощность и стоимость такого суперкомпьютера составят, по прикидкам, 1000 МВт и $50 млрд соответственно [3].

2. Вычисления экзафлопсного класса на основе гибридной архитектуры суперкомпьютера

Стимулирующее влияние все эти годы на компьютерную индустрию высокопроизводительных вычислений оказывала метрика, положенная в основу формирования списка топ-500, которая учитывала только скорость выполнения плавающих операций. Однако в последние годы эта метрика, как стимул достижения производительности 1 эксафлопс, т. е. 10^{18} плавающих операций в секунду, стала оказывать скорее негативное, чем позитивное влияние на развитие индустрии высокопроизводительных вычислений. Дело в том, что по мере развития микроэлектроники стоимость реализации арифметических операций снижалась существенно быстрее стоимости операций доступа к памяти. Как следствие, сегодня время и энергозатраты на решение многих реальных задач определяются количеством требуемых «дорогих» обращений к памяти, а не количеством требуемых «дешевых» арифметических операций над данными, уже извлеченными из памяти [4].

Поэтому в последнее время стали говорить о суперкомпьютерах, обеспечивающих вычисления экзафлопсного класса, т. е. о вычислениях, требующих выполнения порядка 10^{18} операций в секунду (арифметических или доступа к памяти), памяти для рабочего множества с числом элементов порядка 10^{18} и вдесятеро большей оперативной памяти для хранения исходных данных и результатов их обработки.

Ключевыми проблемами создания такого суперкомпьютера экзафлопсного класса являются обеспечение экономически и технически приемлемых энергопотребления и надежности его функционирования, а также стоимости создания и эксплуатации. Так, например, гипотетическое создание Джоном Фитцпатриком экзафлопсного суперкомпьютера стоимостью $50 млрд и энергопотреблением 1000 МВт возможно и экономически приемлемо для игр на валютном рынке объемом $5 трлн и планируемым ежемесячным оборотом $200 млрд. Однако даже для нефтегазовой отрасли, не говоря уже о машиностроении и науке, такое решение очевидно неприемлемо ни технически, ни экономически.

Приемлемый уровень энергопотребления суперкомпьютера экзафлопсного класса должен, по данным DARPA, составлять не более 20 МВт, а энергоэффективность не менее 50 Гфлопс/Вт, что в 50 раз превосходит соответствующие показатели гипотетического суперкомпьютера Джона Фитцпатрика. Наиболее реальный путь достижения такой энергоэффективности — использование гибридных архитектур на основе графических микропроцессоров/ускорителей типа NVidia, Intel MIC и многоядерных микропроцессоров с «легкими ядрами» типа архитектуры PowerPC A2 разработки IBM (18 ядер, SIMD-акселератор, 16 вычислительных ядер, одно — управляющее, одно — резервное). Действительно, из первых 10 систем в списке топ-500 (осень 2014 г.): пять — первого типа, включая TIANHE-2 (55 Пфлопс, 17,8 МВт, энергоэффективность — 3 Гфлопс/Вт), четыре — второго типа, включая SEQUOIA на основе BLUE GENE/Q (20 Пфлопс, 7,9 МВт, энергоэффективность — 2,5 Гфлопс/Вт) и только один К COMPUTER (11,2 Пфлопс, 12,7 МВт) на основе многоядерных микропроцессоров SPARC 64 с «тяжелыми ядрами», имеющий наихудший показатель энергоэффективности 0,88 Гфлопс/Вт. Это подтверждает вывод, который в мае 2013 г. сделал в своем докладе заместитель директора Berkley National Laboratory Horst Simon о том, что невыполнение компаний IBM контракта по созданию суперкомпьютера BLUE WATERS на основе микропроцессора POWER 7 с «тяжелыми ядрами» — завершающее событие в развитии этого типа микропроцессоров и что будущее за суперкомпьютерами с гибридной архитектурой на основе графических микропроцессоров /ускорителей и микропроцессоров с «легкими ядрами» [5].

3. Энергоэффективность и надежность — ключевые проблемы на пути к экзафлопсным вычислениям

3.1. Энергоэффективность

Энергоэффективность суперкомпьютера экзафлопсного класса, с энергопотреблением не более 20 МВт, должна составлять не менее 50 Гфлопс/Вт.

Суперкомпьютер TITAN, установленный в Oak Ridge National Lab состоит из 18688 вычислительных узлов, каждый из которых содержит один NVidia Tesla K20X GPU. Производительность этой системы на тесте LINPACK составляет 17,59 Пфлопс, а энергопотребление 8,2 МВт, т. е. энергоэффективность системы TITAN равна 2,14 Гфлопс/Вт, и на пути движения к экзафлопсному суперкомпьютеру этот показатель необходимо увеличить более чем в 20 раз. Такое радикальное улучшение энергоэффективности потребует разработки целого комплекса новых технических и технологических решений, помимо уменьшения технологических норм проектирования микропроцессоров. Дело в том, что объем энергозатрат, необходимых для выполнения арифметических операций, хорошо масштабируется с уменьшением проектных норм, в то время как объем энергозатрат, необходимых для передачи сигналов по проводникам, масштабируется гораздо хуже.

Действительно, по данным компании NVidia, энергетическая стоимость операции 64-битного умножения с накоплением (чтение трех аргументов, запись результата) для микропроцессора, изготовленного по технологическим нормам 28 нм составляет 20 пикоджоулей (пДж). Следовательно, энергетическая стоимость 10^{18} таких операций составит 20 МВт, т. е. экзафлопсный суперкомпьютер будет тратить в этом случае только на выполнение арифметических операций 20 МВт, весь лимит мощности, установленной требованиями DARPA. К этому, однако, следует добавить энергозатраты:

- на выборку инструкций, которые составляют для современного супер скалярного out-of-order микропроцессора 2000 пДж/инструкция, то есть эквивалентные выполнению 100 арифметических операций;
- запись результата в LI-КЭШ – 50 пДж, что эквивалентно выполнению 2,5 арифметических операций;
- запись результата в DRAM – 16000 пДж, что эквивалентно выполнению 800 арифметических операций.

Энергетическая стоимость передачи 256 бит на расстояние 10 мм составляет внутри 28 нм чипа 256 пДж, а для аналогичного 10 нм чипа, – 200 пДж, т. е. улучшается всего в 1,3 раза. Для операции умножения с накоплением переход на технологию 10 нм улучшает ее энергоэффективность в 2,3 раза с 20 до 8,7 пДж.

Тем не менее, ожидается, что переход от 28 нм технологического процесса к 7 нм обеспечит улучшение энергоэффективности примерно в 4,3 раза. Дополнительное улучшение в 1,9 раза планируется получить за счет уменьшения напряжение питания с 0,9 В (28 нм) на более низкое напряжение 0,7 В (7 нм). Дальнейшие надежды на улучшение энергоэффективности в 2,5 раза связывается с локальными архитектурами и схемными решениями, а также новыми решениями на уровне проектирования системы в целом, обеспечивающими повышение энергоэффективности коммуникационных узлов, сопоставимое с аналогичным показателем вычислительных узлов.

Одним из таких новых технологических решений, которые планирует реализовать компания NVidia, является интеграция в 3D-сборке внутри одного чипа ускорителя и универсального ядра процессора, а также DDR памяти. Как ожидается, это позволит снизить энергозатраты на один пин с 15 пДж до 9, т. е. почти в 1,7 раза [6].

В 2010 г. компанией NVidia были начаты работы по проекту ECHELON, цель которого – создание гибридного микропроцессора производительностью 10 Тфлопс/чип по проектным нормам 10 нм для экзафлопсного компьютера 2018 г. Микропроцессор будет состоять из 8 универсальных и 256 вычислительных ядер, которые через L2-кэш размером 128 Мбайт будут выходить на общую память с производительностью обменов до 4 Тб/с [7].

В ревизии этого проекта от 2014 г. планируется создать микропроцессор с производительностью 16 Тфлопс/чип с энергопотреблением 230 Вт по проектным нормам 7 нм, и на его основе экзафлопсный суперкомпьютер с энергоэффективностью лучше 50 Гфлопс/Вт.

Существенно важно, что для проектируемого экзафлопсного суперкомпьютера компанией NVidia приводится достаточно детальный анализ энергопотребления и производительности на десятке задач трехмерного моделирования. Однако пока не проведены и не опубликованы какие-либо оценки надежности проектируемого суперкомпьютера при решении этих задач.

3.2. Надежность

Имеющийся опыт эксплуатации суперкомпьютеров петафлопсного класса свидетельствует о том, что теоретические оценки надежности их функционирования, вообще говоря, недостоверны. Так, например, для суперкомпьютера TITAN с пиковой производительностью 27 Пфлопс, что составляет 2,7% от 1 Экзафлопс, прогнозное время между отказами, требующими вмешательства оператора, должно быть не более 12 часов. Однако по факту 18-месячной эксплуатации суперкомпьютера среднее время составило 48 часов. При этом был отмечен необъяснимый разброс по частоте отказов между различными вычислительными модулями. Для сравнения, фактическое время между отказами, требующими вмешательства оператора, для суперкомпьютера BLUE WATERS (13 Пфлопс в пике, 1,3% от Экзафлопс) составило всего 4,2 часа [8].

Одна из наиболее вероятных причин этих проблем – недостаточный объем аппаратных средств контроля и коррекции ошибок в массовых коммерческих графических картах, поскольку реализация таких средств влечет за собой, например, увеличение энергопотребления на 20-25% и, конечно, увеличение стоимости карты.

3.3. Ошибка – ординарное, а не исключительное событие

Суперкомпьютер экзафлопсного класса, который компания NVidia планирует создать в рамках проекта ECHELON, будет содержать 629145600 вы-

числительных и 614900 универсальных ядер, 4915200 контроллеров доступа к памяти и сотни тысяч коммуникационных контроллеров на всех уровнях (узел, модуль, стойка) [7]. Принципиальная проблема обеспечения надежности функционирования такой системы с миллиардом активных аппаратных элементов заключается в том, что в этих условиях вероятность возникновения ошибки на небольшом временном интервале близка к единице, то есть ошибка для суперкомпьютера экзафлопсного класса — это ординарное, а не исключительное (как для тера- и петафлопсных систем) событие.

Поэтому построение экзафлопсной системы на основе глобальных механизмов синхронизации и контрольных точек слишком дорого и неэффективно, если вообще возможно. Альтернативой является реализация такой системы на основе локальных асинхронных механизмов парирования и аппаратных и программных ошибок на всех основных уровнях:

- математические модели и алгоритмы;
- прикладное и базовое программное обеспечение;
- аппаратура суперкомпьютера;
- элементная база (прежде всего микропроцессоры и коммуникационные контроллеры).

То есть, обеспечение необходимого уровня надежности должно, в частности, основываться на избыточности реализации в аппаратуре и программном обеспечении дополнительных асинхронных средств самоконтроля их функционирования и парирования различного рода ошибок.

Примером такого подхода на уровне разработки математических методов и алгоритмов может служить метод асинхронной хаотичной итерации для решения блочной системы линейных уравнений (СЛАУ), приведенный в работе [11]:

$H : E \to E; x = (x_1, \dots, x_m) \to ((Hx)_1, \dots, (Hx)_m);$

```
Async_Parallel.for any i {
        until convergence_i do {
                read x from common memory;
                compute xnew_i = H (x)_i;
                overwrite x_i in common memory
                with xnew_i;
        }
        done_i;
}
```

4. Проект CORAL — предэкзафлопсный уровень вычислений

С января 2014 г. Департамент энергетики США ускоренными темпами реализует проект CORAL по разработке и изготовлению для трех национальных лабораторий Департамента семейства суперкомпьютеров, обеспечивающих предэкзафлопсный уровень вычислений. Сокращение C_O_R_A_L расшифровывается как «The **C**ollaboration of **O**ak **R**idge, **A**rgonne and **L**ivermore national Labs». Цели проекта CORAL — обеспечение научного, экономического и военного лидерства США в мире и поддержание, на этой основе, боеспособности ядерных арсеналов США [9].

В рамках проекта CORAL в 2017 г. в три национальные лаборатории (OAK RIDGE, ARGONNE и LIVERMORE) должны быть поставлены три суперкомпьютера двух различных архитектур, производительностью 100-200 Пфлопс каждый. Разработка и поставка CORAL-систем ведется не в рамках модели продавец–покупатель, а в рамках постоянного сотрудничества разработчика-изготовителя и потребителя на всех стадиях разработки–изготовления–эксплуатации. Единый бюджет проекта предусматривает как затраты на проведение исследований, так и индивидуальные затраты на изготовление и четырехлетнее сопровождение каждого из трех суперкомпьютеров проекта. Основные технические требования к претендентам на разработку/изготовление CORAL-систем, выдвинутые заказчиком в начале 2014 г., включали:

- предоставление в составе конкурсной документации полного описания архитектуры предлагаемой CORAL-системы, всех ее компонент и путь ее развития к экзафлопсу;
- обязательное использование двух разных архитектур от двух разных изготовителей, обеспечение пиковой производительности не менее 100 Пфлопс;
- не менее чем четырехкратное ускорение на четырех научных тестах (рекордный счет) и шестикратное ускорение на смеси технических задач с умеренными требованиями к производительности (производственный счет);
- энергопотребление не должно превышать 20 МВт, а интервал между отказами, требующими вмешательства оператора, должен составлять не менее 144 часов;
- предоставление в составе конкурсной документации результатов измерения и/или предсказания производительности на специальных тестах с изложением методик измерений/предсказаний;
- анализ и описание возможных конфигураций CORAL-системы (память, коммуникации, подсистема ввода/вывода, минимальная конфигурация), а также варианты модернизации в середине срока эксплуатации;
- обеспечение совместимости снизу вверх с первыми экзафлопсными компьютерами, на основе технологий программирования MPI, Open MP, OpenACC и CUDA;
- использование существующих приложений в CORAL-системах должно обеспечиваться без радикального изменения программной модели. Типичные приложения — «рабочие лошади» сегодняшнего дня должны работать на закупаемых системах без «капитального ремонта».

По результатам конкурса в ноябре 2014 г. консорциум IBM, NVidia, MELLANOX выиграл контракт, предусматривающий затраты $100 млн на исследования и разработки, и $325 млн на изготовление, введение в эксплуатацию к 2017 г. и пятилетнее сопровождение двух суперкомпьютеров проекта CORAL — производительностью не менее 150 Пфлопс каждый — для Oak Ridge and Livermore National Labs. Оба суперкомпьютера будут использовать процессоры семейства Power фирмы IBM, графические процессоры-ускорители фирмы NVidia и коммуникационную систему семей-

ства Infiniband фирмы MELLANOX, разработанные с использованием полупроводниковых технологий с топологическим размером не больше 14 нм.

В 2015 г. Департамент энергетики США объявил о выделении еще $200 млн на разработку и ввод в эксплуатацию третьего суперкомпьютера «АВРОРА» для Argonne National Labs.

5. Закон Мура после 10 нм

Проблему продления действия закона Мура после преодоления рубежа 10 нм обсуждали на конференции ISSCC 2015 г. представители крупнейших микроэлектронных компаний Intel, Samsung, TSMC, IBM и др. [10]. По-видимому, 10 нм технологический процесс будет последним, построенным Intel по чисто кремниевой технологии. Наиболее вероятно, что в 7 нм технологическом процессе кремний заменят полупроводники группы III-V, такие как индий арсенида галлия (In Ga As). Основная научная и технологическая проблема, которую необходимо при этом решить — обеспечение приемлемого уровня пространственных дефектов, порожденных разницей типов объединяемых кристаллических решеток. Компании IMEC и IBM уже добились заметных успехов в решении этой проблемы, создав образцы высокоэффективных транзисторов на основе полупроводников группы III-V. Тем самым созданы реальные предпосылки к продлению закона Мура по крайней мере еще на 5-7 лет, и, что существенно важно, предотвращению опасности «ширпотребизации» (commoditization) рынка электронных компонент и даже рынка аппаратуры обработки и передачи данных на их основе. То есть, предотвращение слияния брендовых продуктов в общую массу функционально идентичных GENERIC продуктов, конкурирующих между собой только по цене. Производство ширпотреба не требует проведения затратных НИОКР и характеризуется низкой нормой прибыли.

Необходимо отметить, что с точки зрения показателя коммерческой эффективности технических процессов — стоимости миллиона транзисторов, наиболее эффективным является 28 нм технологический процесс, для которого этот показатель равен 2,7 цента/миллион транзисторов. Дальнейшее уменьшение проектных норм влечет ухудшение этого показателя: для 20 нм —2,8 цента, для 16/14нм — 2,9 цента (Handel Jones, IBS Inc.).

Список использованных источников

1. Mark Hachman, March 25, 2013. World's Most Powerful Private Supercomputer Will Hunt Oil, Gas. http://insights.dice.com/2013/03/25/worlds-most-powerful-private-supercomputer-will-hunt-oil-gas-2.
2. Компания Total выбрала SGI для модернизации суперкомпьютера Pangea. http://www.hpcwire.com/off-the-wire/sgi-selected-by-total-to-upgrade-pangea-supercomputer.
3. John Fitzpatrick is forming a company to make an Exaflop computer this year using 1000 megawatts of power, costing $50 billion and using Intel processors. http://nextbigfuture.com/2014/02/john-fitzpatrick-is-forming-company-to.html.
4. GPU and the Future of Parallel Computing, 2011. http://www.cs.nyu.edu/courses/spring12/CSCI-GA.3033-012/ieee-micro-echelon.pdf.
5. No Exascale for You!, An Interview with Berkeley Lab's Horst Simon, May 15, 2013. http://www.hpcwire.com/2013/05/15/no_exascale_for_you_an_interview_with_nersc_s_horst_simon.
6. http://gpu.cs.uct.ac.za/Slides/Echelon.pdf.
7. Scaling the Power Wall: A Path to Exascale, SC14, November 16-21, 2014, New Orlean. https://www.cs.utexas.edu/users/skeckler/pubs/SC_2014_Exascale.pdf.
8. Lessons Learned From the Analysis of System Failures at Petascale: The case of BlueWaters. https://courses.engr.illinois.edu/ece542/sp2014/finalexam/papers/bluewaters.pdf.
9. Department of Energy Awards $425 Million for Next Generation Supercomputing Technologies, November 14, 2014. http://energy.gov/articles/department-energy-awards-425-million-next-generation-supercomputing-technologies.
10. Intel forges ahead to 10nm, will move away from silicon at 7nm, Feb 23, 2015. http://arstechnica.com/gadgets/2015/02/intel-forges-ahead-to-10nm-will-move-away-from-silicon-at-7nm.
11. D. Chazan, W. Miranker. Chaotic relaxation, LinearAlgebra Appl. 2 (1969) 199–222.

On the way to ekzaflops class computing

V. B. Betelin, Director FNTS NIISI, professor, academician RAS.

Фонд «Сколково» создал и запустил программу по микрогрантам

Отныне участникам «Сколково» доступен широкий набор инструментов грантовой поддержки, включающий большие гранты размером до 30, 150 и 300 млн рублей (в зависимости от стадии проекта), предоставляемые для комплексной реализации проекта, минигранты в сумме до 5 млн рублей, которые имеют более узкую направленность, и микрогранты до 1,5 млн руб., предоставляемые на решение отдельных небольших задач в рамках реализации проекта.

Подробнее о том, как подать заявку на микрогрант «Сколково», можно узнать:

https://sk.ru/foundation/grants-experts/p/grants-experts_microgrants.aspx.