УДК 004.272.3

ЭКСАФЛОПНЫЕ СУПЕР-ЭВМ: КОНТУРЫ АРХИТЕКТУРЫ[1]

С.А. Степаненко, В.В. Южаков

Исследованы архитектурные аспекты вычислительных систем эксафлопной производительности. Получены оценки параметров вычислительной и коммуникационной сред. Показаны необходимость и возможности применения архитектурных средств масштабирования эффективности.

Ключевые слова: гибридная архитектура, средства масштабирования, эффективность, реконфигурируемая структура, минимизация, длительность обменов, топологическое резервирование.

ВВЕДЕНИЕ

Задача эффективного применения супер-ЭВМ актуальна в течение всей истории развития вычислительной техники. Это обусловлено как наличием сложнейших задач, для решения которых собственно и разрабатываются супер-ЭВМ, так и большими ресурсами, требуемыми для создания последних.

Достижение эффективности требует учета свойств архитектуры вычислительных систем в прикладных программах и реализации в архитектуре средств, позволяющих ускорить выполнение вычислений. На различных этапах эволюции вычислительной техники применялись различные архитектурные средства — от введения кэш-памяти до создания специализированных вычислителей, аппаратно реализующих алгоритмы [1].

В настоящей работе исследуются архитектурные аспекты, которые с большой вероятностью будут присущи супер-ЭВМ эксафлопной производительности, необходимость которой и возможности создания показаны, например, в публикациях [2, 3].

Эти аспекты обусловлены объективными факторами — энергопотреблением системы эксафлопной производительности и числом задействованных в ней процессорных ядер, определяющим степень параллелизма. Цель работы заключается в:

— обосновании необходимости применения гибридных архитектур для достижения эксафлопной производительности;

— качественной оценке параметров вычислительной и коммуникационной сред;

— описании архитектурных средств масштабирования эффективности, позволяющих на различных уровнях параллелизма учитывать особенности исполняемых процессов, что при прочих равных условиях позволяет уменьшить длительность вычислений и достигнуть практически приемлемых значений производительности и эффективности.

1. ЭТАПЫ ЭВОЛЮЦИИ АРХИТЕКТУРЫ ВЫЧИСЛИТЕЛЬНЫХ СИСТЕМ

Этапы эволюции вычислительных систем согласно, например, работе [1] можно охарактеризовать применяемыми дисциплинами вычислений и архитектурами, реализующими эти дисциплины.

Для достижения производительности $1—10^9$ оп/с оказалось достаточно SISD-дисциплины (Single Instruction Single Data) и однопроцессорной архитектуры.

Достижение $10^{12}—10^{15}$ оп/с потребовало MIMD-дисциплины (Multiple Instructions Multiple Data) и мультипроцессорной архитектуры с разделенной памятью.

Достижение 10^{18} оп/с — эксафлопс — предполагает применение MIMD- и SIMD-дисциплин (Single Instruction Multiple Data) вычислений, реализуемых гибридными архитектурами. Процессорные элементы в них содержат универсальные про-

[1] Статья написана по материалам доклада на шестой международной конференции «Параллельные вычисления и задачи управления» (PACO'2012), Москва, 24 — 26 октября 2012 г.

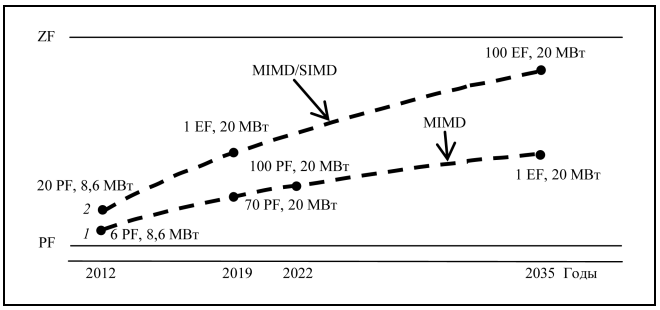

Рис. 1. Оценки производительности и потребляемой мощности: *1* — CPU-only «Titan»; *2* — CPU-Accelerated «Titan»; PF, EF и ZF — пета-, экса- и зеттафлопс соответственно

цессоры — MIMD-компонент и арифметические ускорители — SIMD-компонент.

Применение SIMD-компонентов позволяет гибридной системе достигнуть при определенных условиях производительности 10^{18} оп/с, потребляя 10—20 МВт (в тех же условиях для MIMD системы потребуется не менее 100 МВт); количество MIMD-ядер универсальных процессоров и SIMD-ядер ускорителей составит в системе соответственно ~10^7 и 10^8 шт.

Значения производительности и потребляемой мощности, требуемые для систем, реализующих MIMD-дисциплину и MIMD/SIMD-дисциплину, показаны на рис. 1.

Эффективное задействование гибридных архитектур требует разработки соответствующих вычислительных процессов и анализа их особенностей, в частности, выделения фрагментов, «быстро» исполняемых универсальными процессорами (MIMD-компонентом), и фрагментов, «быстро» исполняемых арифметическими ускорителями (SIMD-компонентом). В свою очередь это влечет необходимость применения нового прикладного и системного программного обеспечения.

Масштабность и трудоемкость создания качественно новых аппаратных и программных средств породили многочисленные исследовательские проекты, выполняемые в различных странах и направленные на освоение гибридных архитектур [2, 5].

Из результатов исследований, выполняемых в мире, следует, что эксафлопная производительность может быть достигнута в результате комплекса взаимозависимых работ, которые включают в себя разработку и создание:

— оптимальной архитектуры, позволяющей обеспечить эффективное исполнение приложений вычислительной системой, состоящей из ~10^8 ядер;

— аппаратных компонентов, удовлетворяющих конструктивным ограничениям и требованиям надежности;

— прикладного и системного программного обеспечения, реализующего управление ресурсами и надежное исполнение приложений на разных уровнях параллелизма;

— экспериментальных систем, позволяющих верифицировать проектные решения.

Удовлетворительным результатом этих работ, приемлемым для практики, будет создание вычислительной машины, имеющей пиковую производительность не менее 1 эксафлопс и соответствующую пропускную способность средств обмена информацией, энергопотребление 10—20 МВт, занимающую 100—200 стоек, оснащенную системным программным обеспечением, позволяющим эффективно распараллеливать приложения на ~10^8 процессов, а также соответствующим прикладным программным обеспечением, допускающим эффективное исполнение с указанным параллелизмом.

Оценим параметры компонентов и некоторые архитектурные средства, требуемые для достижения такого результата.

2. ПАРАМЕТРЫ АППАРАТНЫХ КОМПОНЕНТОВ

Ключевыми аппаратными компонентами служат:

— процессоры для научных расчетов, в качестве которых в ближайшей перспективе рассматриваются MIMD/SIMD-процессоры (MIMD — универсальная часть, SIMD — арифметические ускорители), называемые также гибридными; в более отдаленной — MIMD/SIMD/FPGA;

— система межпроцессорного обмена, включая средства реализации коммуникационной среды.

Оценим параметры вычислительной среды и коммуникационной среды, необходимые для достижения эксафлопной производительности.

2.1. Параметры и состав вычислительной среды

Вычислительный компонент эксафлопной машины (включающий в себя не только процессоры, но и память) должен обеспечить достижение эксафлопной производительности при «разумном» значении энергопотребления — 10—20 МВт и технологической надежности.

Первое может быть достигнуто совместным применением MIMD- и SIMD-компонентов. Вследствие сравнительно простой структуры, энергопотребление, конструктивные размеры и стоимость, приходящиеся на единицу производительности SIMD-компонентов, примерно в 10 раз меньше по сравнению с MIMD-компонентами.

Из приведенных в работах [6—8] данных следуют представленные в табл. 1 значения удельной производительности для MIMD- и SIMD-компо-

Таблица 1

Значения удельной производительности *q*, Гфлопс/Вт

Год	MIMD	SIMD
2014	2—4	24
2016	4—8	50
2018	10—15	100

нентов. В соответствии с этими значениями возможна разработка MIMD/SIMD-процессоров производительностью (значения до и после символа «/» означают соответственно производительность MIMD- и SIMD-компонентов):

— 500...1000 Гфлопс/4000...8000 Гфлопс в 2014 г. проектные нормы 22 нм;

— 1000...2000 Гфлопс/10 000...16 000 Гфлопс в 2017 г. проектные нормы 17 нм (заметим, что в планах ведущих производителей микросхем — 8 нм в 2017 г. [6]).

Потребляемая мощность процессора постоянна: 300—500 Вт.

Можно показать, что вычислительная среда пиковой производительностью 1000 Пфлопс, из которых 100 и 900 Пфлопс составляют производительность MIMD-компонента и SIMD-компонента соответственно, при указанных условиях будет потреблять около 19 МВт, из них 10 МВт приходится на MIMD-компонент и 9 МВт на SIMD-компонент.

В составе такой вычислительной среды понадобится задействовать $50 \cdot 10^3$—$90 \cdot 10^3$ MIMD/SIMD-процессоров пиковой производительностью (1000—2000)/(10 000—16 000) Гфлопс каждый.

Полагаем, что MIMD/SIMD-процессор содержит (100...200) MIMD-ядер и (1000...2000) SIMD-ядер. Общее количество MIMD/SIMD-ядер в системе составит $\sim 10^7/10^8$ шт.

2.2. Коммуникационная среда

Оценим параметры коммуникационной среды, требуемые для объединения указанного количества процессоров в единую систему определенной выше производительности.

2.2.1. Уровни параллелизма и структура соединений

Будем различать четыре уровня параллелизма: процессор, вычислительный блок, стойка и система. Их иерархия показана на рис. 2 (см. третью полосу обложки).

Полагаем, что в процессоре связь между MIMD- и SIMD-компонентами и образующими их ядрами осуществляется внутрипроцессорными средствами.

Структура гибридного процессорного элемента показана на рис. 3. Он содержит несколько MIMD/SIMD-процессоров и коммутатор *K*, через который осуществляется его взаимодействие с другими элементами.

В процессорном элементе задействованы каналы I—III уровней, реализующие соответственно связи между процессорными элементами в вычислительном блоке, в стойке и в системе.

Укажем на идентичность рассматриваемой структуры связей и структуры, примененной, например, в *K* компьютере [9] и в Cray XC [10].

2.2.2. Оценки параметров коммуникационной среды

Функционирование современных процессоров требует ~1500 внешних выводов на его корпусе. Полагаем, что это количество, определяемое механическими параметрами, не изменится. Чтобы уменьшить количество связей, реализуемых проводными соединениями, MIMD/SIMD-процессоры, задействованные в процессорном элементе, объединяют на общей «подложке» или в виде трехмерной сборки. Это позволяет микроэлектронными технологиями реализовать связи между процессорами, а также внешний интерфейс, через который осуществляется связь с системой межпроцессорного обмена. Примером внешнего интерфейса служит совокупность одновременно задействованных разъемов интерфейса PCI Express, Hypertransport или QPI. Возможны другие конструктивные элементы.

Для определенности в расчетах будем использовать характеристики процессорного элемента, производительность MIMD/SIMD-компонентов которого составляет 8 Тфлопс/64 Тфлопс; такая производительность в 2017 г. может быть достигнута объединением на одной подложке восьми MIMD/SIMD-процессоров производительностью 1 Тфлопс/8 Тфлопс.

В качестве каналов связи будем использовать каналы IB 12xHDR (480 · 480) Гбит/с, параметры которых указаны в сообщении [11]; пропускная спо-

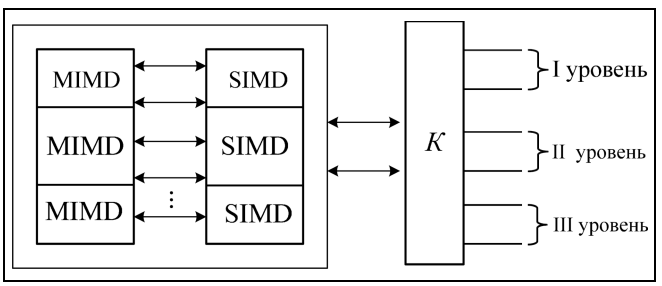

Рис. 3. Процессорный элемент

собность одного линка составляет (40 + 40) Гбит/с = = (5 + 5) Гбайт/с.

Каждый канал содержит 12 линков, его пропускная способность составляет $v_к = (0{,}005 \cdot 12) = 0{,}06$ Тбайт/с.

Для достижения производительности 1 Тфлопс потребуется 2^{14} процессорных элементов.

Полагаем, что вся система содержит 128 стоек, в каждой стойке 8 блоков, в каждом блоке 16 процессорных элементов.

Каналы первого уровня применяются для объединения 16-ти процессорных элементов в вычислительный блок, достаточно длины $l_1 = 20$ см. Каналы второго уровня — для объединения вычислительных блоков в стойке, достаточно длины $l = 2$ м. Каналы третьего уровня объединяют стойки, достаточно длины $l = 20$ м.

Реализация каналов первого и второго уровней возможна путем применения многослойных печатных плат. Реализация каналов третьего уровня, по-видимому, невозможна без применения многомодовых оптических средств связи.

Оценим три варианта топологии среды: 3D тор, гиперкуб (H) и dragonfly (DF).

При расчете значений параметров среды полагаем, что выполняются условия:

— размерности среды 3D тор: 32 × 32 × 16; причем, по координате z объединены 32 элемента (два блока) в топологию 1D тор; четыре 1D тора (восемь блоков) в стойке объединяются по координате x, восемь стоек образуют ряд по координате x; 16 рядов по координате y, по 8 стоек в каждом, образуют систему; каждый процессорный элемент содержит 6 каналов;

— размерности среды H: $n_1 = 4$ (2^4 элементов в блоке), $n_2 = 3$ (2^3 блоков в стойке), $n_3 = 7$ (2^7 стоек в системе); каждый процессорный элемент содержит 14 каналов;

— размерности среды DF: 16 элементов (blades) объединены полносвязным графом в блок (chassis), в каждом элементе 15 каналов первого уровня; 8 блоков объединяются полносвязным графом в стойку (group), в каждом элементе 7 каналов второго уровня; стойки объединены полносвязным графом в систему, каждая пара стоек соединена восемью каналами; у каждого элемента 8 каналов третьего уровня; в скобках указаны термины, используемые для этой среды в работе [10]; размерности рассматриваемой здесь среды и среды, описанной в той же работе [10], также близки; каждый процессорный элемент содержит 30 каналов.

Для каждой из рассматриваемых топологий — 3D тор, H и DF, в табл. 2 указаны значения C_i — количество связей среды i-уровня и L_i — суммарная длина этих связей. Символом D обозначено значение диаметра — наибольшего расстояния между процессорными элементами; γ — отношение суммарной пропускной способности каналов связи процессорного элемента к производительности его SIMD-компонента.

Заметим, что значения γ в табл. 2 существенно меньше $\gamma = 0{,}12$ для системы Cray XE6 [12] и $\gamma = 0{,}2$ для объявленного IBM проекта системы «Blue Waters» [13].

Приведенные в табл. 2 данные демонстрируют реальность создания системы эксафлопной производительности. Они иллюстрируют достоинства и недостатки рассмотренных топологий, влияние которых понадобится оценивать на этапе создания систем, исходя из достигнутого технологического уровня.

3. АРХИТЕКТУРНЫЕ СРЕДСТВА МАСШТАБИРОВАНИЯ ЭФФЕКТИВНОСТИ

Рассмотренные варианты вычислительной системы характеризуются:

— гибридной (неоднородной) структурой процессорных элементов;

— сложностью коммуникационной среды и, как следствие, сравнительно малым значением отношения пропускной способности каналов связи процессорного элемента к его производительности.

В этих условиях необходимы инструментальные средства, которые позволяют учитывать архитектурные особенности вычислительной системы и обеспечивают масштабирование эффективности на различных уровнях параллелизма.

Таблица 2

Значения параметров коммуникационных сред

Топология	Уровень I	Уровень II	Уровень III	D	γ
	C_1, шт./L_1, км	C_2, шт./L_2, км	C_3, шт./L_3, км		
3D	$0{,}18 \cdot 10^6/36$	$0{,}16 \cdot 10^6/319$	$0{,}2 \cdot 10^6/3{,}9 \cdot 10^3$	40	0,005
H	$0{,}39 \cdot 10^6/78$	$0{,}29 \cdot 10^6/600$	$0{,}69 \cdot 10^6/14 \cdot 10^3$	14	0,013
DF	$1{,}5 \cdot 10^6/294$	$0{,}69 \cdot 10^6/1376$	$0{,}78 \cdot 10^6/15 \cdot 10^3$	5	0,028

Рассмотрим средства архитектурного масштабирования эффективности:

— реконфигурацию структуры гибридных процессорных элементов, посредством вариации количества MIMD-ядер и задействованных с ними SIMD-ядер, в соответствии с особенностями выполняемого вычислительного процесса для достижения наибольшей в заданных условиях производительности и эффективности;

— применение бесконфликтных множеств источников и приемников и/или декомпозиции вычислительного процесса на подпроцессы и размещение их по элементам среды в соответствии с особенностями элементов и топологией связей среды в целях минимизации длительностей обменов информацией между процессорными элементами;

— топологическое резервирование процессорных элементов, позволяющее при отказах элементов сохранять неизменными топологию среды и ее производительность на данном процессе, тем самым сохранять значение эффективности, достигнутое перечисленными средствами реконфигурации структуры и минимизации длительностей обменов.

3.1. Гибридные реконфигурируемые структуры

Значения ускорения вычислений гибридными системами и их эффективность зависят от особенностей решаемой задачи и параметров вычислительной среды.

К особенностям задачи, точнее, алгоритма ее решения, относятся длительности нераспараллеливаемых фрагментов, количество и тип операций обмена информацией, синхронность вычислительных процессов и т. п.

Для гибридных архитектур (в отличие от однородных) характерно, что вычислительный процесс распределяется в начале между MIMD- и SIMD-компонентами и лишь затем между процессорами, образующими эти компоненты.

Результирующее ускорение зависит от ускорений достигаемых на MIMD- и SIMD-компонентах и от размера «долей» вычислительного процесса, приходящихся на эти компоненты.

Варьируя производительностями MIMD- и SIMD-компонентов — в частности, количеством задействованных в них ядер, можно изменять длительности выполнения вычислительного процесса.

В работе [14] получены оценки длительности вычислений гибридными системами (процессорными элементами) в зависимости от соотношений между фрагментами вычислительного процесса и производительностью MIMD- и SIMD-компонентов, выполняющих эти фрагменты. Предложены критерии динамической реконфигурации структуры процессора, предусматривающие разделение ядер MIMD- и SIMD-компонентов на определенные взаимодействующие подмножества, состав и производительность которых определяются в соответствии с параметрами исполняемого процесса.

Варьирование составом и производительностью MIMD- и SIMD-компонентов позволяет, исходя из определенных первичных свойств процесса, получить максимальное для заданных условий ускорение вычислений.

В частности, коэффициент ускорения вычислений гибридной системой, содержащей q ядер и один ускоритель, по сравнению с системой, содержащей одно ядро универсального процессора, имеет вид:

$$K_{q,1} = \frac{q}{\varphi + (1-\varphi)q/\rho},$$

если задействованы одно ядро и q ускорителей, то $K_{1,q} = \dfrac{q}{\varphi q + (1-\varphi)/\rho}$, где $0 \leqslant \varphi \leqslant 1$ — доля вычислительного процесса, выполняемого универсальным процессором (доля MIMD-фрагмента), $\rho > 1$ — коэффициент ускорения по сравнению с универсальным ядром процессора, достигаемый применением ускорителя на SIMD-фрагменте.

В качестве иллюстрации приведем, согласно работе [15], пример вычислений значений потенциала Морзе по программе молекулярной динамики гибридной системой, содержащей четырехядерный процессор Intel Core i7920 и арифметические ускорители Nvidia Tesla C2050.

Длительность вычислений одним ядром значений потенциала Морзе для задачи размером $55 \times 55 \times 55$ периодов кристаллической решетки составила $T_1 = 22{,}96$ с. Этот вычислительный процесс можно разделить на два фрагмента: MIMD-фрагмент, выполняемый ядром в течение $T_M = 7{,}07$ с (следовательно, $\varphi = 7{,}7/22{,}96 \approx 0{,}31$) и SIMD-фрагмента, выполняемый ускорителем в течение $T_s = 2{,}8$ с (имеем $\rho = (1-\varphi)\, T_1/T_s = 5{,}67$). Длительность выполнения этого процесса в режиме умножения (weak scaling) гибридной системой, содержащей одно ядро универсального процессора и четыре ускорителя, составляет $T_{1,4} = 30{,}3$ с, а длительность выполнения системой, содержащей четыре ядра и один ускоритель $T_{4,1} = 18{,}3$ с, т. е. система из четырех ядер и одного ускорителя на этом процессе в 1,65 раз быстрее системы из одного ядра и четырех ускорителей.

На рис. 4 (см. третью полосу обложки) для рассматриваемого вычислительного процесса по программе молекулярной динамики приведены графики: $K_{q,1}$ — значения ускорения вычислений по сравнению с одним ядром, достигаемые гибридной системой, содержащей q ядер и один ускори-

тель; $K_{1,q}$ — значения ускорения вычислений по сравнению с одним ядром, достигаемые гибридной системой, содержащей одно ядро и q ускорителей.

Из значений $K_{q,1}$ и $K_{1,q}$ следует, что этот вычислительный процесс быстрее выполняется системой с большим количеством ядер.

Другие подробности применения гибридных систем изложены в работах [14, 16]. Изложенный метод может быть применен на первом уровне параллелизма.

3.2. Минимизация длительностей обменов

Минимизация длительностей обменов достигается взаимной адаптацией вычислительного процесса и структуры связей между процессорными элементами с целью исключения конфликтов при выполнении обменов и уменьшения расстояний обменов. Возможности адаптации зависят как от топологии вычислительной среды, так и от свойств вычислительного процесса (явные схемы, регулярные связи и т. д.) С увеличением сложности вычислительной системы актуальность (и результативность) этих средств возрастает.

3.2.1. Бесконфликтные множества

Различные топологии мультипроцессорных сред накладывают различные принципиальные ограничения на количество свободных непересекающихся маршрутов, исключающих возникновение конфликтов в процессе передачи информации.

Пусть S и R соответственно множества источников и приемников такие, что для любой пары $a \in S$ и $b \in R$ существует свободный маршрут длины, не превышающей $l_{\max}$ при условии, что все остальные источники и приемники из множеств S и R также выполняют парные обмены ($l_{\max}$ — диаметр среды, расстояние достаточное для соединения любой пары «источник — приемник» из данной среды). Другими словами, S и R — такие множества источников и приемников, находящиеся на максимальном для данной среды расстоянии, при задействовании которых для любой пары «источник — приемник» существует и заранее известен свободный маршрут.

Таблица 3

Мощности и численность бесконфликтных множеств

Среда	3D тор	H^m	DF
C	$\omega^{2/3}$	$\omega/2 = 2^{m-1}$	$\omega/2$
G	$\omega^{2/3} 2^{\omega^{2/3}}$	$\binom{\log_2\omega}{\log_2\log_2\omega} \log_2\omega \cdot 2^{\frac{\omega}{\log_2\omega}-1}$	$2^{\omega-1}$
$l_{\max}$	$(3/2)\omega^{1/3}$	$m = 2^k - 1,\ k = 0,1,2$	2

Пусть $C = \|S\| = \|R\|$ — мощности этих множеств, G — количество таких множеств для сред с различной топологией. Количество процессорных элементов в среде обозначим ω. Оценки C и G для сред с различными топологиями приведены в работе [17].

Значения C, G и $l_{\max}$ для сред с топологиями 3D тор, H^m и DF указаны в табл. 3 (для среды DF оценки приведены лишь для двух первых уровней (chasis и group) [10]). Из них следует, что сравнительно хорошими коммуникационными возможностями характеризуются среды H^m и DF.

По определению бесконфликтных множеств, выполнение обменов между принадлежащими им источниками и приемниками свободно от конфликтов; потребуются затраты времени лишь на построение маршрута длины $l_{\max}$ и передачи информации по этому маршруту. Это позволяет в определенных ранее условиях сохранять практически неизменной эффективность среды при наращивании ее сложности.

Недостаток рассматриваемого метода заключается в сравнительно «малой» численности бесконфликтных множеств, по сравнению с общей численностью $2^{\omega-1}$ множеств источников и приемников, содержащих по $\omega/2$ элементов; исключение составляют лишь среда N — полный матричный коммутатор и первые два уровня среды DF; однако их аппаратная реализация очень сложна.

3.2.2. Декомпозиция и размещение процессов

Другим средством масштабирования эффективности минимизацией длительностей обменов служит размещение источников и приемников на минимальном расстоянии с целью достижения наименьшего значения длительности обменов.

Рассмотрим возможности, предоставляемые различными топологиями для реализации одного класса вычислительных процессов, относящихся к наиболее применимым (рис. 5, *а* — см. третью полосу обложки). Потребуем, чтобы «массовые» обмены выполнялись только между подпроцессами, расположенными на соседних процессорных элементах — непосредственно соединенных каналом. Тем самым попытаемся создать условия, при которых с увеличением числа процессорных элементов для обменов задействуются лишь соседние коммутаторы.

Средства минимизации длительностей обменов включают в себя:

— декомпозицию вычислительного процесса в соответствии с особенностями процессорных элементов;

— размещение полученных подпроцессов по элементам в соответствии с направлениями обменов между ними.

Возможности декомпозиции и размещения вычислительных процессов, предоставляемые топологией 3D тор, изложены, например, в работах [18, 19].

Покажем возможность применения других топологий, в частности, H^n, из которой можно получить 1D, 2D и 3D торы разных размерностей.

В гиперкуб размерности n, обозначаемый как H^n, помещаются (вкладываются) с сохранением физического соседства: 1D тор из 2^n процессов; 2D тор из $2^{n1}\times 2^{n2}$ процессов, где $n_1 + n_2 = n$; 3D тор из $2^{n1}\times 2^{n2}\times 2^{n3}$ процессов, где $n_1 + n_2 + n_3 = n$.

В 3D тор можно помещать 3D, 2D и 1D торы меньших размерностей; в 2D тор можно помещать 2D и 1D торы меньших размерностей.

Потребуем, чтобы и для трехмерного, и для двумерного процесса обмены с соседями по каждому измерению обеспечивались одинаковыми связями. Тогда, в качестве процессорного элемента целесообразно использовать элемент, содержащий 2^m процессоров, где m — число, кратное 3 и 2.

Процессорный элемент, содержащий $2^6 = 64$ процессора, позволяет размещать на нем «квадраты» размерностью $2^3\times 2^3$ и «кубики» $2^2\times 2^2\times 2^2$.

Для вычисления требуемого отображения процессов и исполняющих их элементов может применяться прикладная программа, результатом выполнения которой служит таблица соответствия. Эта таблица передается системным средствам, которые загружают процессы на соответствующие процессоры.

Описанные средства легко распространяются на топологию DF и позволяют обеспечить физическое соседство компонентов, образующих вычислительный процесс. Эти средства применимы в условиях современных аппаратных платформ — вычислительных модулей из нескольких, в частности, гибридных многоядерных процессоров на общей памяти.

В табл. 4 приведены значения производительности [20], достигнутые на тесте NPB LU [21]. В столбцах 1 — значения производительности, оп/с, для варианта размещения процессов в соответствии со структурой связей вычислительной системы, приведенной на рис. 5, *б* (для вычисления соответствия использовался код Грэя), в столбцах 2 — значения производительности для последовательного размещения процессов, обычно реализуемого системным планировщиком. В частности, на тесте NPB LU система (класс C) из 512 процессорных ядер при оптимальном размещении процессов показала производительность в 1,72 раза большую по сравнению с достигаемой при «обычном» последовательном размещении.

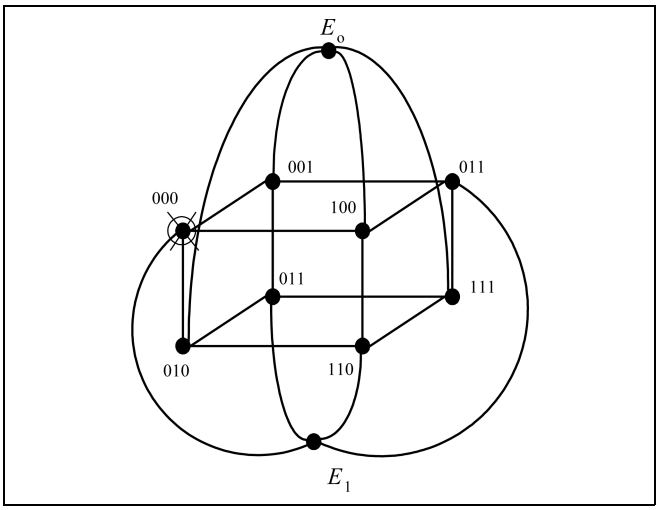

Рис. 6. Топологическое резервирование H^3 избыточными элементами

Представленные в табл. 4 данные показывают, что эффект от применения декомпозиции и размещения подпроцессов возрастает с увеличением числа процессорных элементов (ПЭ) (ядер), задействованных в процессе вычисления. Этот эффект иллюстрируется рис. 6.

3.3. Средства отказоустойчивого масштабирования эффективности

Сбои и отказы отдельных элементов обусловлены как аппаратными, так и программными эффектами, характер и источник которых «некогда» выяснять в процессе счета, их надо исключать и изолировать.

Таблица 4

Значения производительности на тесте NPB LU

Число процессорных ядер, шт.	Класс B		Класс C		Класс D	
	1	2	1	2	1	2
128	88 950	73 478	97 072	95 398	83 421	83 008
256	164 537	108 826	177 953	152 613	223 547	221 760
512	—	—	283 926	164 283	409 697	406 182

Архитектурные средства обеспечения надежности (дополняющие технологические и схемотехнические достижения) должны не только устранять источники сбоев и отказов, но и сохранять эффекты масштабирования эффективности, достигаемые в результате применения средств, описанных ранее.

Масштабирование эффективности может быть достигнуто применением методов топологического резервирования [22, 23], позволяющих обеспечить в случае отказов и сбоев элементов неизменность топологии среды и ее производительности.

Могут быть применены два метода топологического резервирования. Их отличительными особенности:

— сохранение топологии вычислительной среды, выполняющей вычислительный процесс (деградации в случае отказа не происходит);

— идентичность резервных и резервируемых элементов.

Первый метод [22] основан на введении избыточных процессорных элементов. Он иллюстрируется на рис. 6, где E_0 и E_1 — резервные элементы. Если, например, откажет элемент (000), он и его каналы связи заменяются элементом E_0 и его каналами.

Пусть $p = p(t)$ — вероятность безотказной работы процессорного элемента на интервале t. $P = p^{\omega}$ — вероятность безотказной работы среды из ω элементов на интервале t. Можно показать, что для d-кратного резервирования, когда на $\log\omega = n$ элементов вводится d резервных, вероятность безотказной работы среды на интервале t составит:

$$P = p^{\omega} + \left(\begin{array}{c} \omega \\ 1 \end{array} \right) p^{\omega - 1}(1 - p)^1 + \ldots + \left(\begin{array}{c} \omega \\ d \end{array} \right) p^{\omega - d}(1 - p)^d.$$

Второй метод [23] основан на избирательном резервировании части вычислительной среды. Процессорные элементы, занятые выполнением одного вычислительного процесса, резервируются в случае необходимости другими процессорными элементами этой же среды. Эти элементы изначально могут использоваться для выполнения других, менее «важных» процессов, которые при необходимости резервирования удаляются. Множество процессорных элементов, предоставляемое резервируемому процессу, имеет ту же топологию и мощность, что и исходное.

Второй метод иллюстрируется на рис. 7, где процесс, исполняемый элементами (0000, 0001, 0010, 0011), может быть перенесен на другую плоскость. Например, в случае отказа элемента (0000) процесс можно перенести на элементы (1000, 1001, 1010, 1011).

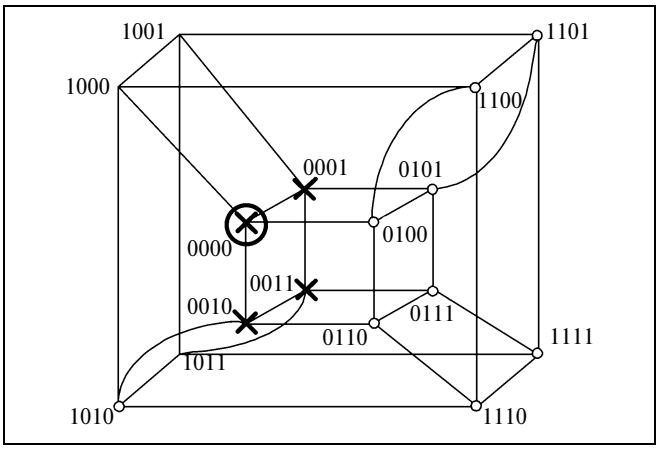

Рис. 7. Топологическое избирательное резервирование H^4 выделенными элементами

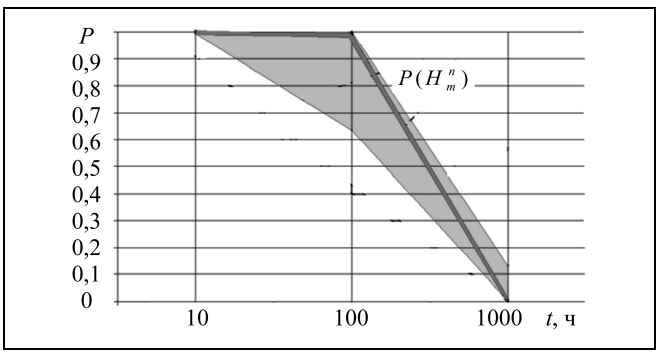

Рис. 8. Топологическое избирательное резервирование: вероятность выполнения процесса

В общем виде вероятность $P(H_m^n)$ выполнения средой с топологией H^n вычислительного процесса, занимающего в ней подмножество процессорных элементов H^m, где $m < n$, запишется как

$$\sum_{i=0}^{d} \left(\begin{array}{c} \omega \\ i \end{array} \right) p^{\omega - i}(1 - p)^i \leqslant P(H_m^n) \leqslant \sum_{i=0}^{\mu} \left(\begin{array}{c} \omega \\ i \end{array} \right) p^{\omega - i}(1 - p)^i,$$

где $d = 2^{n-m}$, $\mu = \left(\begin{array}{c} n \\ 0 \end{array} \right) + \left(\begin{array}{c} n \\ 1 \end{array} \right) + \ldots + \left(\begin{array}{c} n \\ m \end{array} \right)$.

На рис. 8 показаны значения вероятностей безотказного выполнения средой из 128 элементов процесса длительностью t, требующего половину элементов среды.

Процессорный элемент имеет длительность наработки на отказ (MTBF) 10^4 ч. Вероятность безотказного выполнения процесса длительностью $t = 100$ ч без применения топологического избирательного резервирования составляет $0{,}28 \leqslant P \leqslant 0{,}53$, с применением резервирования $0{,}64 \leqslant P \leqslant 0{,}99$, наиболее вероятное значение близко к 0,99 (показано утолщенной линией).

Оба метода могут быть применены для сред с различными топологиями. Разумеется, различные топологии обеспечивают различные оценки вероятностей выполнения вычислительного процесса.

Эффективность вычислительной системы зависит также от длительности записи и чтения контрольных точек [24, 25].

Избыточные элементы могут использоваться для хранения контрольных точек, для этого они располагаются аналогично резервным на минимальном расстоянии от элементов, непосредственно выполняющих вычисления.

Неизменность топологии среды и ее производительность исключают необходимость каких бы то ни было изменений исполняемых программ и процессов в случае отказов.

Реализация средств топологического резервирования на различных уровнях параллелизма применительно к процессору (резервирование ядер MIMD- и SIMD-компонентов), вычислительному блоку (резервирование процессорных элементов), стойке (резервирование блоков) и т. п. позволяет создавать среды с наперед заданными значениями вероятностей исполнения вычислительного процесса определенной длительности, занимающего заданное число элементов.

4. УРОВНИ ПАРАЛЛЕЛИЗМА И КОНТУРЫ АДАПТИРУЕМОЙ АРХИТЕКТУРЫ

Из изложенного следует, что перспективный вариант достижения эксафлопной производительности заключается в применении гибридных архитектур, реализующих различные дисциплины вычислений и допускающих реконфигурацию компонентов в соответствии с особенностями исполняемого процесса.

Эффективное применение гибридных архитектур предусматривает создание системного и прикладного программного обеспечения, позволяющего как можно «сильнее» задействовать возможности аппаратных средств, в частности, возможности их адаптации к особенностям исполняемой программы.

Реализуемая реконфигурация — средство создания архитектуры, динамически адаптируемой к особенностям исполняемого процесса на первом уровне параллелизма — на уровне MIMD/SIMD-компонентов. Эти компоненты более общие, по сравнению с узкофункциональными арифметическими и логическими устройствами, обычно рассматриваемыми при построении реконфигурируемых систем. Их задействование в соответствии с особенностями исполняемого вычислительного процесса позволяет ускорить вычисления.

Задействование бесконфликтных множеств процессорных элементов, маршрутизация в соответствии с топологией среды и размещение процессов по процессорным элементам, минимизирующее длительности обменов, обеспечивают эффективное задействование коммуникационной среды на втором и третьем уровнях параллелизма в соответствии с особенностями исполняемой программы.

Топологическое резервирование на различных уровнях позволяет реализовывать отказоустойчивое масштабирование эффективности среды.

В свою очередь, в создаваемых программах должны быть учтены особенности и параметры вычислительной системы — наличие и состав MIMD- и SIMD-компонентов, структура и характер связей между элементами, модулями и стойками.

Представляется сомнительным, что без реализации перечисленных средств гибридную систему эксафлопной производительности удастся эффективно использовать на содержательных задачах.

Изложенное означает принципиально новый уровень взаимозависимости аппаратных и программных средств (именуемый в литературе «co-design»), реализуемый через специальный инструментарий, позволяющий максимально задействовать возможности аппаратуры и использовать алгоритмические особенности прикладных программ.

ЗАКЛЮЧЕНИЕ

Исследованы архитектурные особенности вычислительных систем, необходимые для достижения эксафлопной производительности.

Оценены параметры процессорной среды и коммуникационной среды.

Показана целесообразность применения архитектурных средств масштабирования эффективности, включающих в себя:

— реконфигурацию структуры гибридных процессорных элементов посредством вариации количества MIMD-ядер и задействованных с ними SIMD-ядер в соответствии с особенностями выполняемого вычислительного процесса в целях достижения наибольшей в заданных условиях производительности и эффективности;

— применение бесконфликтных множеств источников и приемников и/или декомпозиции вычислительного процесса на подпроцессы и раз-

мещение их по элементам среды в соответствии с особенностями элементов и топологией связей среды; этим достигается минимизация длительностей обменов информацией между процессорными элементами;

— топологическое резервирование элементов среды, позволяющее при отказах элементов сохранять неизменными топологию среды и ее производительность на данном процессе. Тем самым сохраняется значение эффективности, достигнутое перечисленными средствами реконфигурации структуры и средствами минимизации длительностей обменов, т. е. обеспечивается отказоустойчивое масштабирование эффективности.

В результате достигается динамическая адаптируемость архитектуры к особенностям исполняемой программы (при условии, что в самой программе учтены возможности архитектуры), что в свою очередь должно обеспечить эффективность применения эксафлопных супер-ЭВМ.

ЛИТЕРАТУРА

1. *Цилькер Б.Я., Орлов С.А.* Организация ЭВМ и систем. — СПб.: Питер, 2004.
2. *Концепция* по развитию технологии высокопроизводительных вычислений на базе супер-ЭВМ эксафлопного класса на 2012—2020 гг. — URL: http://www.rosatom.ru/wps/wcm/connect/rosatom/rosatomsite/aboutcorporation/nauka/ (дата обращения 13.11. 2013).
3. *SC11* Keynote by Nvidia CEO Jen-Hsun Huang. — URL: http://blogs.nvidia.com/2011/11/exascale-an-innovator%E2%80 %99s-dilemma/ (дата обращения 13.11.2013).
4. *Stevens R. and White A.* A DOE Laboratory plan for providing exascale applications and technologies for critical DOE mission needs. — URL: http://computing.ornl.gov/workshops/scidac2010/presentations/r_stevens.pdf (дата обращения 16.12.2013).
5. *International* Exascale Software Project. — URL: www.exascale.org (дата обращения 13.11. 2013).
6. *URL:* http://www.ecmwf.int/newsevents/meetings/workshops/2010/high_performance_computing_14th/presentations/barkai.pdf (дата обращения 13.11.2013).
7. *SC'09* Exascale Panel. Steve Scott. Cray Chief Technology Officer. Exhibitor Forum, SC'09. — URL: http://www.exascale.org/mediawiki/images/c/c1/SC09_Exascale_panel_Scott.pdf (дата обращения 13.11.2013).
8. *The Future* of GPU Computing. — URL: http://www.nvidia.com/content/GTC/documents/SC09_Dally.pdf (дата обращения 13.11.2013).
9. *Tomohiro Inoue*. Fujutsu Limited. The 6D Mesh/Torus Interconnect of K Computer. — URL: http://www.fujitsu.com/downloads/TC/sc10/interconnect-of-k-computer.pdf (дата обращения 13.11.2013).
10. *Alverson B., Froese E., Kaplan L. and Roweth D.* Cray Inc. Cray XC Series Network. — URL: http://www.cray.com/Assets/PDF/products/xc/CrayXC30Networking.pdf (дата обращения 13.11.2013).
11. *Infiniband* Roadma. — URL: http://www.infinibandta.org/content/pages.php?pg=technology_overview (дата обращения 13.11.2013).
12. *IBM* Blue aters. — URL: http://www.ncsa.illinois.edu/Blue Waters (дата обращения 13.11.2013).
13. *Cray* Titan. — URL: http://www.knoxnews.com/news/2011/mar/07/oak-ridge-lab-to-add-titanic-supercomputer/ (дата обращения 13.11.2013).
14. С*тепаненко С.А.* Оценки ускорения вычислений гибридными системами // Пятая междунар. конф. «Параллельные вычисления и задачи управления' PACO 2010»: плен. докл., 26—28 октября 2010 г., Москва / ИПУ РАН. — М., 2010. — С. 61—71.
15. *Крючков И.А., Степаненко С.А., Рыбкин А.С.* Реализация статической маршрутизации и оптимального размещения вычислительных процессов в мультипроцессорных средах // Молодежь в науке: Сб. докл. Шестой науч.-техн. конф. — Саров, 2008. — С. 172—176.
16. *Степаненко С.А.* Топологическое резервирование мультипроцессорных сред выделенными элементами // Тр. РФЯЦ—ВНИИЭФ. — 2005. — № 10. — С. 50—60.
17. *Применение* арифметических ускорителей для расчета задач молекулярной динамики по программному комплексу МД / Б.Л. Воронин, А.М. Ерофеев, С.В. Копкин и др. // Вопросы атомной науки и техники. Сер. Математическое моделирование физических процессов. — 2009. — Вып. 2.
18. *Степаненко С.А.* Топологическое резервирование мультипроцессорных сред // Вопросы атомной науки и техники. Сер. Математическое моделирование физических процессов. — 2002. — Вып. 4. — С. 55—60.
19. *NASA*, NAS Parallel Benchmars. — URL: http://www.nas.nasa.gov/Resources/Software/npb.html (дата обращения 13.11.2013).
20. *Barrett B., Barrett R., Brandt J.,* et al. Report of Experiments and Evidence for ASC L2 Milestone 4467 — Demonstration of a Legacy Application's Path to Exascale; Sandia Report, SAND2012-1750, Printed March 2012.
21. *Daly J.T.* A higher order estimate of the optimum checkpoint interval for restart dumps. Los Alamos National Laboratory. M/S, Los Alamos, NM 87545, USA. 28 December 2004. — URL: http://www.sciencedirect.com (дата обращения 13.11.2013).
22. *Hao Yu, I.-Hsin Chung, Jose Moreira*. Topology Mapping for Blue Gene/L Supercomputer. SC2006 November 2006. — URL: http://www.ibm.com (дата обращения 13.11.2013).
23. *Network* Resiliency for Cray XETM Systems. — URL: http://fs.hlrs.de/projects/craydoc/docs/books/S-0032-3101/html-S-0032-3101/index.html (дата обращения 13.11.2013).
24. *Степаненко С.А.* Коммуникационные параметры мультипроцессорных сред // Сб. докл. IX Междунар. семинара по супервычислениям и математическому моделированию. Саров, 3—7 октября 2006 г. — Саров, 2006. — С. 96.
25. *Stepanenko S.A.* Estimated speedups of hybrid reconfigurable systems // XIV International conference «Supercomputing and Mathematical Modeling» / RFNC — VNIIEF, Sarov, October 1—5, 2012. — Sarov, 2012.

Статья представлена к публикации членом редколлегии В.Г. Лебедевым.

Степаненко Сергей Александрович — д-р физ.-мат. наук, нач. отдела, ☎ (83130) 4-53-54, ✉ ssa@vniief.ru,

Южаков Василий Васильевич — зам. нач. отдела, ☎ (83130) 2-79-66, ✉ v.v.yuzhakov@vniief.ru,

Российский федеральный ядерный центр — Всероссийский научно-исследовательский институт экспериментальной физики, г. Саров.