ISSN 2079-3316 ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ № 4(18), 2013, с.61—90

УДК 004.272.3

С. А. Степаненко, В. В. Южаков

Эксафлопсные суперЭВМ. Контуры архитектуры[1]

Аннотация. Исследованы архитектурные аспекты вычислительных систем эксафлопной производительности. Оценены параметры вычислительной и коммуникационной сред. Показано, что для достижения эксафлопной производительности необходимы гибридные системы. Процессорные элементы этих систем содержат ядра универсальных процессоров и арифметические ускорители. Они реализуют MIMD и SIMD дисциплины вычислений соответственно.

Эффективное задействование эксафлопных гибридных систем требует принципиально нового программного обеспечения и средств архитектурного масштабирования эффективности.

Применение перечисленных средств иллюстрируется на примерах тестовых программ молекулярной динамики и NPB LU.

В результате достигается динамическая адаптируемость архитектуры к особенностям исполняемой программы, что в свою очередь обеспечивает эффективность применения эксафлопных суперЭВМ.

Ключевые слова и фразы: Гибридные архитектуры, архитектурные средства масштабирования эффективности, гибридные реконфигурируемые структуры, минимизация длительности обменов, топологическое резервирование.

Введение

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

[1] Статья рекомендована к публикации Программным комитетом НСКФ-2012

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 использовались различные архитектурные средства — от введения КЭШ памяти до создания специализированных вычислителей, аппаратно реализующих алгоритмы [1].

Ниже исследуются архитектурные аспекты, которые с большой вероятностью будут присущи суперЭВМ эксафлопсной производительности, необходимость которой и возможности создания показаны, например, в[2].

Эти аспекты обусловлены объективными факторами – энергопотреблением системы эксафлопсной производительности и количеством задействованных в ней процессорных ядер, определяющим степень параллелизма.

В этой работе:

- дано обоснование необходимости применения гибридных архитектур для достижения эксафлопной производительности;
- приведены качественные оценки параметров вычислительной среды и коммуникационной среды; для последней оценены три варианта топологии;
- изложены архитектурные средства масштабирования эффективности, позволяющие 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, [1] можно охарактеризовать применяемыми дисциплинами вычислений и архитектурами, реализующими эти дисциплины.

Для достижения производительности 10^0–10^9 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 процессоры — MIMD компонента и арифметические ускорители — SIMD компонента.

Применение SIMD компонент 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 штук.

Рис. 1 из [3] иллюстрирует значения производительности и потребляемой мощности, требуемые для систем, реализующих MIMD дисциплину и MIMD/SIMD дисциплину.

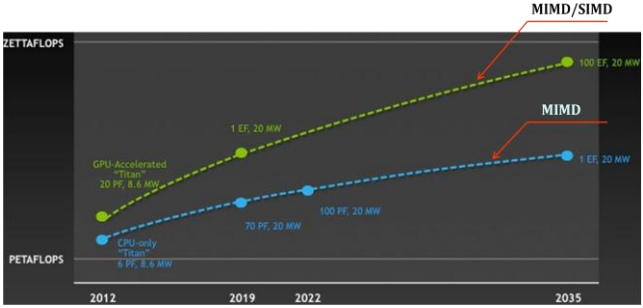

Рис. 1. Оценка производительности и потребляемой мощности

Эффективное задействование гибридных архитектур требует разработки соответствующих вычислительных процессов и анализа их особенностей, в частности выделения фрагментов «быстро» ис-

полняемых универсальными процессорами (MIMD компонентой) и фрагментов «быстро» исполняемых арифметическими ускорителями (SIMD компонентой). 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 [4].

Из результатов исследований, выполняемых в мире, следует, что эксафлопсная производительность может быть достигнута в результате комплекса взаимозависимых работ, которые включают следующее:

- разработку оптимальной архитектуры, позволяющей обеспечить эффективное исполнение приложений вычислительной системой из ~10^8 ядер;
- создание аппаратных компонентов, удовлетворяющих конструктивным ограничениям и требованиям надёжности;
- разработку прикладного и системного программного обеспечения, реализующего управление ресурсами и надежное исполнение приложений на разных уровнях параллелизма;
- создание экспериментальных систем, позволяющих верифицировать проектные решения.

Удовлетворительным результатом этих работ, приемлемым для практики, будет создание вычислительной машины со следующими свойствами:

- пиковая производительность не менее 1 Эксафлопс и соответствующая пропускная способность средств обмена информацией,
- энергопотребление 10–20 МВт,
- размер 100–200 стоек,
- системное программное обеспечение позволяет эффективно распараллеливать приложения на ~10^8 процессов,
- прикладное программное обеспечение допускает эффективное исполнение с указанным параллелизмом.

Оценим параметры компонентов и некоторые архитектурные средства, требуемые для достижения указанной цели.

2. Параметры аппаратных компонентов

Ключевыми аппаратными компонентами являются:

- процессоры для научных расчётов, 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 не только процессоры, но и память) должен обеспечить достижение эксафлопной производительности при «разумном» значении энергопотребления — 10–20 МВт и технологической надёжности.

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 приведённых в [5]–[7] данных следуют представленные в Таблица 1 значения q Гфлопс/Вт — удельные производительности для MIMD и SIMD компонентов.

Таблица 1. Значения удельной производительности

	MIMD Гфлопс/Вт	**SIMD Гфлопс/Вт**
2014	2–4	24
2016	4–8	50
2018	10–15	100

В соответствии с указанными в Таблица 1 значениями возможна разработка MIMD/SIMD процессоров производительностью (значения до и после символа / означают соответственно производительность MIMD и SIMD компонент):

- 500–1000 Гфлопс / 4000–8000 Гфлопс в 2014 г., проектные нормы 22 нм;
- 1000–2000 Гфлопс / 10000–16000 Гфлопс в 2017 г., проектные нормы 17 нм.

Заметим, что в планах ведущих производителей микросхем — 8 нм в 2017 г. ([5]).

Потребляемая мощность процессора постоянна — 300–500 Вт.

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

В составе этой вычислительной среды понадобится задействовать $50 \cdot 10^3 - 90 \cdot 10^3$ MIMD/SIMD процессоров пиковой производительностью $(1000-2000)/(10000-16000)$ 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 следующие уровни параллелизма: процессор, вычислительный блок, стойка и система. Их иерархия показана на Рис. 2.

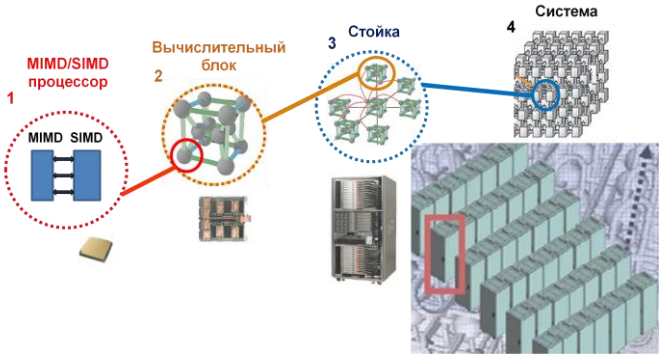

Рис. 2. Уровни параллелизма

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, через который осуществляется его взаимодействие с другими элементами.

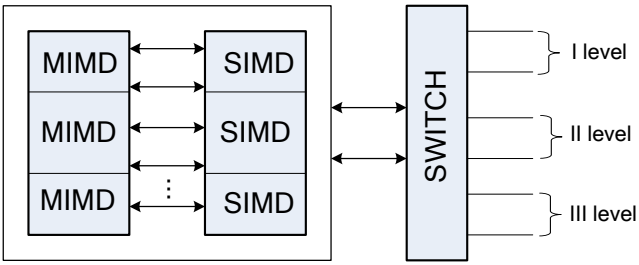

Рис. 3. Структура гибридного процессорного элемента

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

Укажем идентичность рассматриваемой структуры связей, примененной, например, в K компьютере [8] и в Cray XC [9].

2.2.2. Оценки параметров коммуникационной среды

Функционирование современных процессоров требует примерно 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 является совокупность одновременно задействуемых разъемов интерфейса PCI Express, Hypertransport или QPI. Возможны другие конструктивные элементы.

Для определенности в расчетах будем использовать процессорный элемент, производительность MIMD/SIMD компонент которого составляет 8 Тфлопс / 64 Тфлопс; эта 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 [10]; пропускная способность одного линка составляет $(40 + 40)$Гбит/с $= (5 + 5)$ Гбайт/с.

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

Для достижения производительности 1 Eflops 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 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 применением 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 следующие условия:

- размерности среды 3D тор равны: 32x32x16; 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 ка-

налов 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 [9]; размерности рассматриваемой здесь среды и среды из [9] также близки; каждый процессорный элемент содержит 30 каналов.

Для каждой из рассматриваемых топологий — 3D тор, H и DF, в Таблица 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; символом γ — отношение суммарной пропускной способности каналов связи процессорного элемента к производительности его SIMD компоненты.

Таблица 2. Значения параметров коммуникационных сред

	Уровень 1	**Уровень 2**	**Уровень 3**	**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

Заметим, что значения γ в Таблица 2 существенно меньше γ=0.12 для системы Cray XE6 [13] и γ= 0.2 для объявленного IBM проекта системы Blue Waters [11].

Приведенные в Таблица 2 данные демонстрируют реальность создания системы эксафлопной производительности. Они иллюстрируют достоинства и недостатки рассмотренных топологий, влия-

ние которых понадобится оценивать на этапе создания систем, исходя из достигнутого технологического уровня.

3. Архитектурные средства масштабирования эффективности

Рассмотренные варианты вычислительной системы характеризуются следующими факторами:

- гибридная (неоднородная) структура процессорных элементов;
- сложность коммуникационной среды, и как следствие, сравнительно малое значение 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.

Рассмотрим следующие пути архитектурного масштабирования эффективности:

- реконфигурация структуры гибридных процессорных элементов посредством вариации количества MIMD и задействованных с ними SIMD ядра, 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 с целью 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 указанными выше 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 то, 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 компонент, в частности, количеством задействованных в них ядер, можно изменять длительности выполнения вычислительного процесса.

В [15] получены оценки длительности вычислений гибридными системами (процессорными элементами) в зависимости от соотношений между фрагментами вычислительного процесса и производительностью MIMD и SIMD компонент, выполняющих эти фрагменты. Предложены критерии динамической реконфигурации структуры процессора, предусматривающие разделение ядер MIMD и SIMD компонент на определенные взаимодействующие подмножества, состав и производительность которых определяются в соответствии с параметрами исполняемого процесса.

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

В частности, коэффициенты ускорения вычислений гибридной системой, содержащей q ядер и 1 ускоритель, по сравнению с одним ядром универсального процессора, имеют вид

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

Если задействованы 1 ядро и q ускорителей, то

$$K_{1,q} = \frac{q}{\varphi\cdot q+(1-\varphi)\frac{1}{\rho}}, \quad (2)$$

где $0 \le \varphi \le 1$ 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 сказанного приведем согласно [18] 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 55x55x55 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, $\varphi = \frac{7,07}{22,96} \approx 0,31$) и SIMD фрагмента, выполняемый ускорителем в течение T_s=2,8 с (имеем ρ=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 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 ускоритель;

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

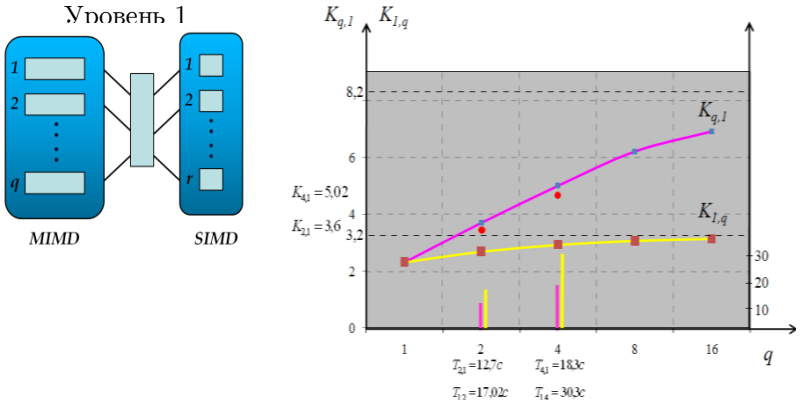

Рис. 4. Значения коэффициентов ускорения гибридными элементами различных конфигураций

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 [15],[26].

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, причем 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 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.

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 [25].

Значения C, G и l_{max} для сред с топологиями 3D тор, H^m и DF указаны в Таблица 3 (для среды DF оценки приведены лишь для двух первых уровней (chasis и group) [9]).

ТАБЛИЦА 3. МОЩНОСТИ И КОЛИЧЕСТВА БЕСКОНФЛИКТНЫХ МНОЖЕСТВ

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

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 выше условиях сохранять практически неизменной эффективность среды при наращивании её сложности.

Недостатком рассматриваемого метода является сравнительно «малое» количество бесконфликтных множеств, по сравнению с общим количеством $2^{\omega-1}$ 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 является размещение источников и приемников на минимальном расстоянии с целью достижения наименьшего значения длительности обменов.

Рассмотрим возможности, предоставляемые различными топологиями, для реализации одного класса вычислительных процес-

сов, относящихся к наиболее применимым (см. Рис. 5а). 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 количества процессорных элементов для обменов задействуются лишь соседние коммутаторы.

Средства минимизации длительностей обменов включают:

- 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 [23], [24].

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 H^n, помещаются (вкладываются) с сохранением физического соседства:

- 1D тор из 2^n процессов;

- 2D тор из $2^{n1}x2^{n2}$ процессов, где $n_1 + n_2 = n$;

- 3D тор из $2^{n1}x2^{n2}x2^{n3}$ 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 2^3x2^3 и «кубики» $2^2x2^2x2^2$.

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

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

В Таблица 4 приведены согласно [16] значения производительности, достигнутые на тесте NPB LU [20]. В столбце 1 — значения производительности для варианта размещения процессов в соответствии со структурой связей вычислительной системы, приведенной на Рис. 5б (для вычисления соответствия использовался код Грэя (Gray)), в столбце 2 — значения производительности для последовательного размещения процессов, обычно реализуемого системным планировщиком.

В частности, на тесте NPB LU класс C система из 512 процессорных ядер при оптимальном размещении процессов показала производительность в 1,72 раза большую, по сравнению с достигаемой при «обычном» последовательном размещении.

Таблица 4. Значения производительности на тесте NPB LU

Количество процессорных ядер, шт.	Класс B		Класс C		Класс D	
	1 оп/с	2 оп/с	1 оп/с	2 оп/с	1 оп/с	2 оп/с
128	88950	73478	97072	95398	83421	83008
256	164537	108826	177953	152613	223547	221760
512			283926	164283	409697	406182

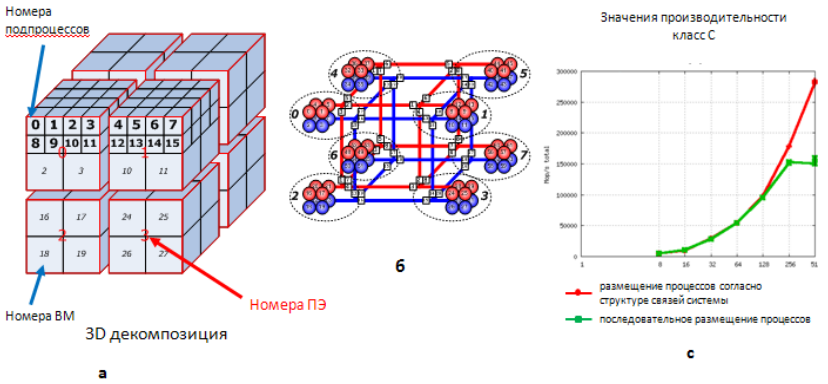

Рис. 5. Структура системы и значения производительности

Представленные в Таблица 4 данные показывают, что эффект от применения декомпозиции и размещения подпроцессов возрастает с увеличением количества процессорных элементов (ядер), задействованных в процессе вычисления. Этот эффект иллюстрируется Рис. 5c.

3.3. Средства отказоустойчивого масштабирования эффективности

Сбои и отказы отдельных элементов обусловлены как аппаратными, так и программными эффектами, характер и источник ко-

торых «некогда» выяснять в процессе счета, их надо исключать и изолировать.

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, указанных в предыдущих разделах.

Масштабирование эффективности может быть достигнуто применением методов топологического резервирования [19],[17], позволяющих обеспечить в случае отказов и сбоев элементов неизменность топологии среды и ее производительности.

Могут быть применены два метода топологического резервирования. Отличительными особенностями обоих являются:

- сохранение топологии вычислительной среды, выполняющей вычислительный процесс (деградации в случае отказа не происходит);
- идентичность резервных и резервируемых элементов.

Первый метод [19] основан на введении избыточных процессорных элементов. Он иллюстрируется 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).

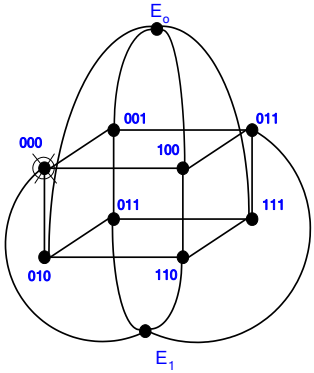

Рис. 6. Топологическое резервирование H^3 избыточными элементами

Пусть $p{=}p(t)$ — вероятность безотказной работы процессорного элемента на интервале t. Обозначим $P{=}p^{\omega}$ — 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

$$P = p^{\omega} + \binom{\omega}{1} p^{\omega-1}(1-p)^1 + \cdots + \binom{\omega}{d} p^{\omega-d}(1-p)^d.$$

Второй метод [17] 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).

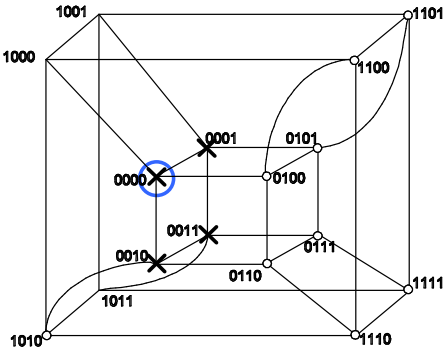

Рис. 7. Топологическое избирательное резервирование H^4 выделенными элементами

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 имеет вид

$$(3) \qquad \sum_{i=0}^{d} \binom{\omega}{i} p^{\omega-i}(1-p)^i \le P(H_m^n) \le \sum_{i=0}^{\mu} \binom{\omega}{i} p^{\omega-i}(1-p)^i,$$

где

$$d = 2^{n-m},$$

$$\mu = \binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{m}.$$

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 часов. Вероятность безотказного выполнения процесса длительностью $t = 100$ час без применения топологического избирательного резервирования составляет $0{,}28 \le \rho \le 0{,}53$, с применением резервирования $0{,}64 \le \rho \le 0{,}99$, наиболее вероятное значение близко к 0.99 (показано красной линией).

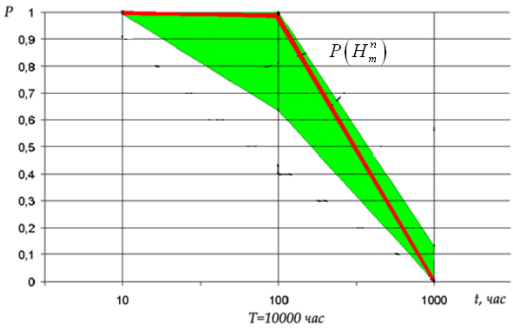

Рис. 8. Топологическое избирательное резервирование. Вероятность выполнения процесса

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 [21],[22].

Избыточные элементы могут использоваться для хранения контрольных точек; 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), к вычислительному блоку (резервирование процессорных элементов), к стойке (резервирование блоков) и т.п. позволяет создавать среды с наперед заданными значениями вероятностей исполнения вычислительного процесса определенной длительности, занимающего заданное количество элементов.

4. Уровни параллелизма и контуры адаптируемой архитектуры

Из приведенного выше следует, что перспективным вариантом достижения эксафлопсной производительности является применение гибридных архитектур, реализующих различные дисциплины вычислений и допускающих реконфигурацию компонент 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 — есть средство создания архитектуры, динамически адаптируемой к особенностям исполняемого процесса на первом уровне параллелизма — на уровне MIMD/SIMD компонент. 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, 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, задействуемое 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 компонент, структура и характер связей между элементами, модулями и стойками.

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

Вышеизложенное 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

В работе исследованы архитектурные особенности вычислительных систем, необходимые для достижения эксафлопсной производительности.

Оценены параметры процессорной среды и коммуникационной среды.

Показана целесообразность применения архитектурных средств масштабирования эффективности, включающих:

- реконфигурацию структуры гибридных процессорных элементов посредством вариации количества MIMD и задействованных с ними SIMD ядер; в соответствии с особенностями выполняемого вычислительного процесса с целью 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; 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 указанными выше 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 суперЭВМ.

Список литературы

[1] Цилькер Б. Я., Орлов С. А. *Организация ЭВМ и систем.* С.-Пб., 2004 г.

[2] Концепция по развитию технологии высокопроизводительных вычислений на базе суперЭВМ эксафлопного класса на 2012-2020 гг. [Электронный ресурс]. Режим доступа: http://www.rosatom.ru/wps/wcm/connect/rosatom/rosatomsite/aboutcorporation/nauka/

[3] SC11 Keynote by Nvidia CEO Jen-Hsun Huang [Электронный ресурс]. Режим доступа: http://blogs.nvidia.com/2011/11/exascale-an-innovator%E2%80%99s-dilemma/

[4] Rick Stevens and Andy White. *A DOE Laboratory plan for providing exascale applications and technologies for critical DOE mission needs* [Электронный ресурс]. Режим доступа: http://computing.ornl.gov/workshops/SCIDAC2010/r_stevens.pdf

[5] International Exascale Software Project URL: www.exascale.org

[6] [Электронный ресурс]. Режим доступа:: http://www.ecmwf.int/newsevents/meetings/workshops/2010/high_performance_computing_14th/presentations/barkai.pdf

[7] SC'09 Exascale Panel. Steve Scott. Cray Cheef Technology Officer. Exhibitor Forum, SC'09

[8] The Future of GPU Computing [Электронный ресурс]. Режим доступа: http://www.nvidia.com/content/GTC/documents/SC09_Dally.pdf

[9] Tomohiro Inoue. Fujutsu Limited. *The 6D Mesh/Torus Interconnect of K Computer.* [Электронный ресурс]. Режим доступа: http://www.fujitsu.com/downloads/TC/sc10/interconnect-of-k-computer.pdf

[10] Bob Alverson, Edwin Froese, Larry Kaplan and Duncan Roweth. Cray Inc. Cray XC Series Network. [Электронный ресурс]. Режим доступа:: http://www.cray.com

[11] Infiniband Roadmap

[Электронный ресурс]. Режим доступа: http://www.infinibandta.org/content/pages.php?pg=technology_overview

[12] IBM Blue Waters [Электронный ресурс]. Режим доступа: http://www.ncsa.illinois.edu/BlueWaters

[13] Cray Titan [Электронный ресурс]. Режим доступа: http://www.knoxnews.com/news/2011/mar/07/oak-ridge-lab-to-add-titanic-supercomputer/

[14] Liu N., Carothers C., Cope J. Ross R. Model and Simulation of Exascale Communication Network. [Электронный ресурс]. Режим доступа: http://www.mcs.anl.gov/uploads/cels/papers/P1937-0911.pdf

[15] Степаненко С. А. *Оценки ускорения вычислений гибридными системами.* // Пленарные доклады Пятой международной конференции "Параллельные вычисления и задачи управления" PACO'2010 Москва, 26–28 октября 2010 г. М.: Учреждение Российской академии наук. Институт проблем управления им. В. А. Трапезникова РАН с. 61–71, ISBN 978-5-91450-062-4.

[16] Крючков И. А., Степаненко С. А., Рыбкин А. С. *Реализация статической маршрутизации и оптимального размещения вычислительных процессов в мультипроцессорных средах.* // «Молодежь в науке». Сборник докладов шестой научно-технической конференции. Саров, 2008 г. с. 172–176.

[17] Степаненко С. А. *Топологическое резервирование мультипроцессорных сред выделенными элементами.* // Труды РФЯЦ-ВНИИЭФ №10, 2005 г. с. 50–60.

[18] Воронин Б. Л. Ерофеев А. М., Копкин С. В., Крючков И. А., Рыбкин А. С., Степаненко С. А., Южаков В. В. *Применение арифметических ускорителей для расчета задач молекулярной динамики по программному комплексу МД.* // «Вопросы атомной науки и техники». Сер. Математическое моделирование физических процессов. 2009 г., вып. 2.

[19] Степаненко С. А. *Топологическое резервирование мультипроцессорных сред* // Вопросы атомной науки и техники. Сер. Математическое моделирование физических процессов. 2002 г. вып. 4, с. 55–60.

[20] NASA, «NAS Parallel Benchmars» [Электронный ресурс]. Режим доступа: http://www.nas.nasa.gov/Resources/Software/npb.html

[21] Barrett B., Barrett R., Brandt J. and others. *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.

[22] J. T. Daly. *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. [Электронный ресурс]. Режим доступа: http://www.sciencedirect.com.

[23] Hao Yu, I-Hsin Chung, Jose Moreira. *Topology Mapping for Blue Gene/L Supercomputer.* SC2006 November 2006. [Электронный ресурс]. Режим доступа: http://www.ibm.com

[24] Network Resiliency for Cray XETM Systems. [Электронный ресурс]. Режим доступа: http://fs.hlrs.de/projects/craydoc/docs/books/S-0032-3101/html-S-0032-3101/index.html

[25] Степаненко С. А. *Коммуникационные параметры мультипроцессорных сред.* // Сборник докладов IX Международного семинара по супервычислениям и математическому моделированию. Саров, 3–7 октября 2006 г., с. 96.

[26] Stepanenko, S. A. 2012. Estimated speedups of hybrid reconfigurable systems. XIV International conference "Supercomputing and Mathematical Modeling". RFNC-VNIIEF, Sarov, October 1–5, 2012, p. 120 [in Russian].

Рекомендовал к публикации *чл.-корр. РАН С. М. Абрамов*

Об авторах:

Сергей Александрович Степаненко

Доктор физико-математических наук, лауреат Государственной премии Российской Федерации, главный научный сотрудник РФЯЦ-ВНИИЭФ, автор более 50 публикаций.

e-mail: ssa@vniief.ru

Василий Васильевич Южаков

Начальник группы научно-исследовательского отдела РФЯЦ-ВНИИЭФ.

e-mail: V.V.Yuzhakov@vniief.ru

Образец ссылки на публикацию:

С. А. Степаненко, В. В. Южаков. Эксафлопсные суперЭВМ. Контуры архитектуры // Программные системы: теория и приложения: электрон. научн. журн. 2013. Т. 4, № 4(18), с. 61–90.
URL: http://psta.psiras.ru/read/psta2013_4_61-90.pdf

S. A. Stepanenko, V. V. Yuzhakov. Exascale supercomputers. Architectural outlines.

Abstract. Architectural aspects of exascale supercomputers are explored. Parameters of the computing environment and interconnect are evaluated. It is shown that reaching exascale performances requires hybrid systems. Processor elements of such systems comprise CPU cores and arithmetic accelerators, implementing the MIMD and SIMD computing disciplines, respectively.

Efficient exascale hybrid systems require fundamentally new applications and architectural efficiency scaling solutions, including:

- process-aware structural reconfiguring of hybrid processor elements by varying the number of MIMD cores and SIMD cores communicating with them to attain as high performance and efficiency as possible under given conditions;
- application of conflict-free sets of sources and receivers and/or decomposition of the computation to subprocesses and their allocation to environment

elements in accordance with their features and communication topology to minimize communication time;

- application of topological redundancy methods to preserve the topology and overall performance achieved by the above communication time minimization solutions in case of element failure thus maintaining the efficiency reached by the above reconfiguring and communication minimization solutions, i.e. to provide fault-tolerant efficiency scaling.

Application of these solutions is illustrated by running molecular dynamics tests and the NPB LU benchmark.

The resulting architecture displays dynamic adaptability to program features, which in turn ensures the efficiency of using exascale supercomputers. (*in Russian*)

Key Words and Phrases: Hybrid architectures, architectural efficiency scaling solutions, hybrid reconfigurable structures, minimization of communication time, topological redundancy.