© 2012 г. С.А. СТЕПАНЕНКО, д-р физ.-мат. наук,
В.В. ЮЖАКОВ
(Российский федеральный ядерный центр Всероссийский научно-исследовательский институт экспериментальной физики, Саров)

ЭКСАФЛОПНЫЕ СУПЕРЭВМ КОНТУРЫ АРХИТЕКТУРЫ

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

EXASCALE SUPERCOMPUTERS. ARCHITECTURAL OUTLINES / S.A. Stepanenko, V.V. Yuzhakov (Russian Federal Nuclear Center – All-Russian Research Institute of Experimetnal Physics. Institute of theoretical and mathematical physics, 37 Mira Avenue, Sarov, Nizhny Novgorod Region, Russia, 607188, E-mail: ssa@vniief.ru)
Architectural aspects of exascale computing systems are explored.
Assessments of the performance of computing systems and communication environment are made. Necessity and potential of architectural efficiency scaling solutions are demonstrated.

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.

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 эксафлопной производительности, необходимость которой и возможности создания показаны в [2].

Эти аспекты обусловлены объективными факторами – энергопотреблением системы эксафлопной 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, предназначенные для эффективного задействования аппаратных компонент. Эти средства учитывают определенные особенности вычислительных процессов, что при прочих равных условиях позволяет уменьшить длительность вычислений;
– отмечены возможности адаптации (специализации) архитектуры эксафлопной системы к особенностям исполняемых процессов, предоставляемые рассматриваемыми средствами и основанные на взаимном влиянии свойств аппаратуры и программного обеспечения.

1 Этапы эволюции архитектуры вычислительных систем

Этапы эволюции вычислительных систем согласно [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. При этом количество MIMD ядер универсальных процессоров и SIMD ядер ускорителей составит в системе соответственно $\sim 10^7$ и 10^8 штук.

На рисунке 1 приведены согласно [3] значения производительности и потребляемой мощности для систем, реализующих MIMD дисциплину и MIMD/SIMD дисциплину.

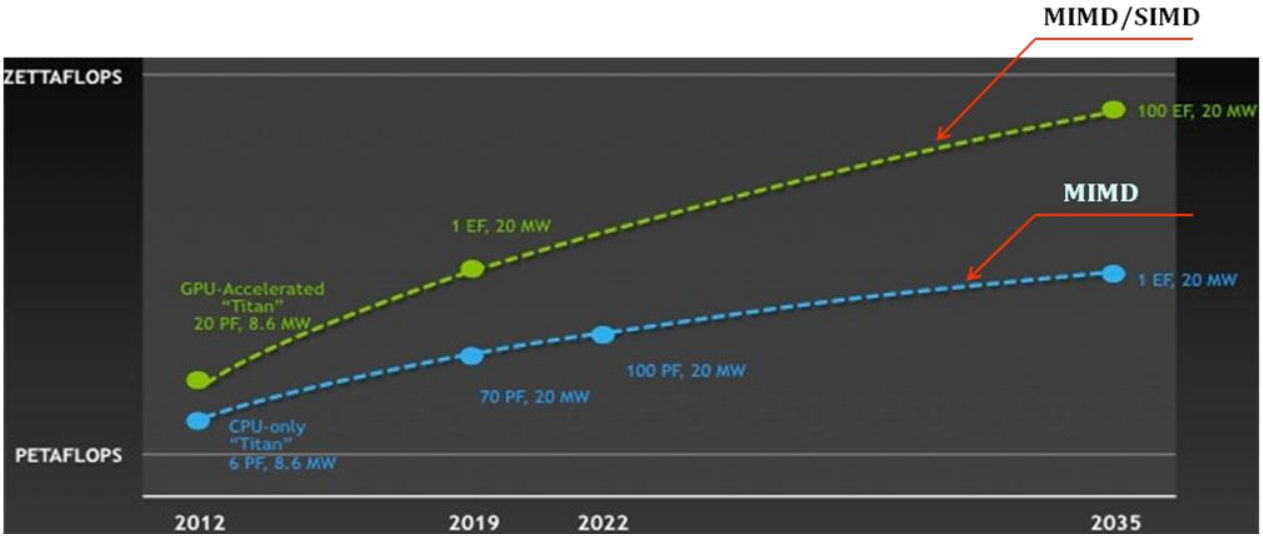


Рисунок 1 – Оценка производительности и потребляемой мощности

Эффективное задействование гибридных архитектур требует разработки соответствующих вычислительных процессов и анализа их особенностей, в частности, выделения фрагментов «хорошо» исполняемых универсальными процессорами (MIMD компонентой) и фрагментов для арифметических ускорителей (SIMD компонента). В свою очередь это влечет необходимость создания 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, 5].

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

- разработку оптимальной архитектуры, позволяющей обеспечить эффективное исполнение приложений 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 машины, имеющей пиковую производительность не менее 1 Эксафлопс и соответствующую пропускную способность средств обмена информацией, энергопотребление 10-20 МВт, занимающую 100-200 стоек, оснащённую 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 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 (СМПО), включая средства реализации коммуникационной среды.

Оценим параметры вычислительной 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 [6-8] данных следуют представленные в таблице 1 значения q Гфлопс/Вт – удельные производительности для MIMD и SIMD компонентов.

Таблица 1 – Значения удельной производительности

	MIMD Гфлопс/Вт	**SIMD Гфлопс/Вт**
2010	0,5-1,0	2
2012	1-2	8
2014	2-4	24
2016	4-8	50
2018	10-15	100

В соответствии с указанными в таблице 1 значениями возможна разработка ряда MIMD/SIMD процессоров производительностью:

- 256-512 Гфлопс / 1500-2000 Гфлопс в 2012г. проектные нормы 30/30 нм;
- 500-1000 Гфлопс / 4000-8000 Гфлопс в 2014г. проектные нормы 22/22 нм;
- 1000-2000 Гфлопс / 10000-16000 Гфлопс в 2017г. проектные нормы 17/17 нм.

Заметим, что в планах Intel 15 нм в 2013г. и 8 нм в 2017г. [6].

Потребляемая мощность процессора постоянна – 300-500 Вт.

Можно показать, что вычислительная среда пиковой производительностью 1 000 Пфлопс, из которых 100 Пфлопс и 900 Пфлопс составляют производительность MIMD компоненты и SIMD компоненты соответственно, при указанных условиях будет потреблять 19 000 КВт, из них 10 000 КВт приходится на MIMD компоненту и 9 000 КВт на SIMD компоненту.

В составе этой вычислительной среды понадобится задействовать $50 \cdot 10^3 - 90 \cdot 10^3$ MIMD/SIMD процессоров пиковой производительностью (1 000 - 2 000)/(10 000 – 16 000) Гфлопс каждый.

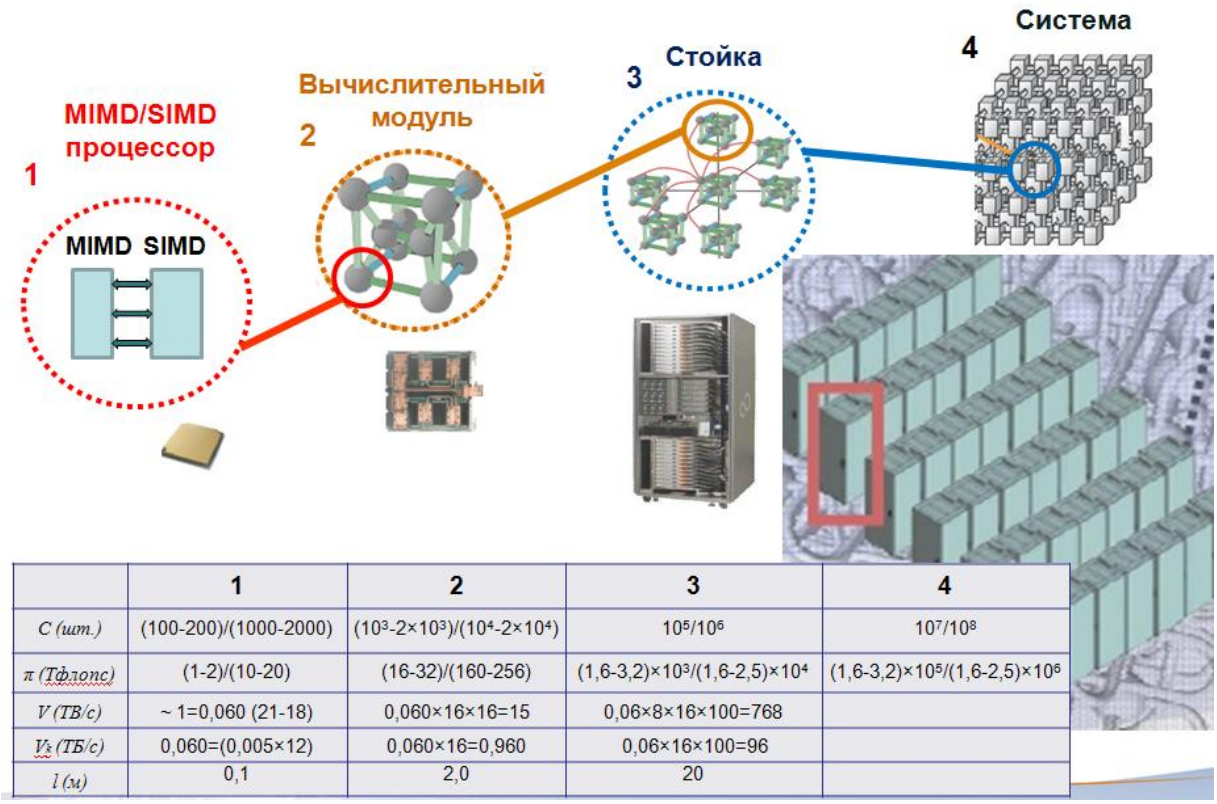
Полагаем, что MIMD/SIMD процессор содержит (100 - 200) MIMD ядер и (1 000 - 2 000) SIMD ядер. Общее количество MIMD/SIMD ядер в системе составит $\sim 10^7/10^8$ шт.

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 модуль, стойка и система. Их иерархия показана на рисунке 2.



	1	2	3	4
C (шт.)	(100-200)/(1000-2000)	$(10^3\text{-}2\times10^3)/(10^4\text{-}2\times10^4)$	$10^5/10^6$	$10^7/10^8$
π (Тфлопс)	(1-2)/(10-20)	(16-32)/(160-256)	$(1{,}6\text{-}3{,}2)\times10^3/(1{,}6\text{-}2{,}5)\times10^4$	$(1{,}6\text{-}3{,}2)\times10^5/(1{,}6\text{-}2{,}5)\times10^5$
V (ТВ/с)	~ 1=0,060 (21-18)	0,060×16×16=15	0,06×8×16×100=768	
$V_Б$ (ТБ/с)	0,060=(0,005×12)	0,060×16=0,960	0,06×16×100=96	
l (м)	0,1	2,0	20	

Рисунок 2 – Уровни параллелизма

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

MIMD/SIMD процессор и коммутатор, через который осуществляется его взаимодействие с другими процессорами, образуют гибридный процессорный элемент, показанный на рисунке 3.

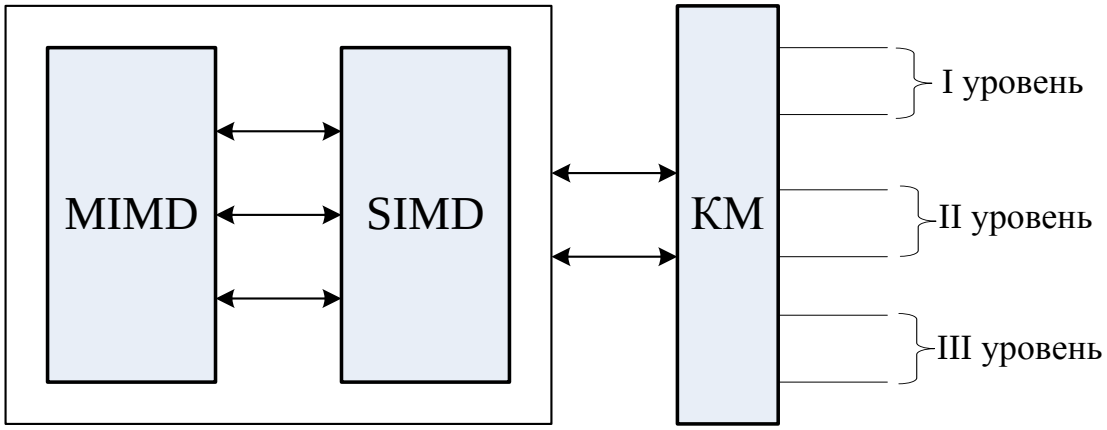


Рисунок 3 – Процессорный элемент

В процессорном элементе задействованы каналы I, II и III уровней, реализующие соответственно связи между процессорными элементами внутри ВМ, стойки и системы.

Укажем идентичность рассматриваемой структуры связей, примененной японцами в К компьютере [9], опубликованной в эксафлопном проекте Nvidia Echelon [10]; эти структуры воспроизведены на рисунках 4а и 4б соответственно.

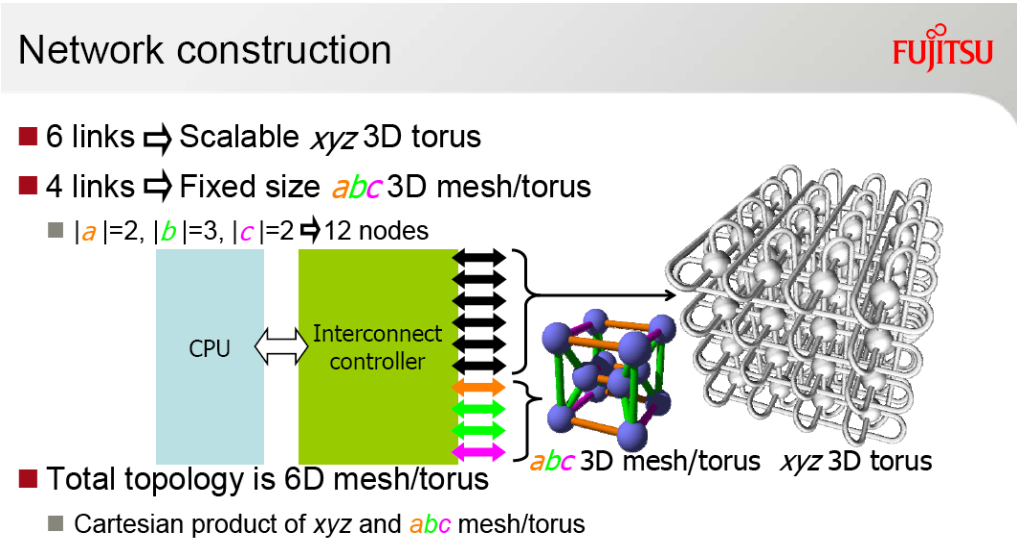


Рисунок 4а - K Computer

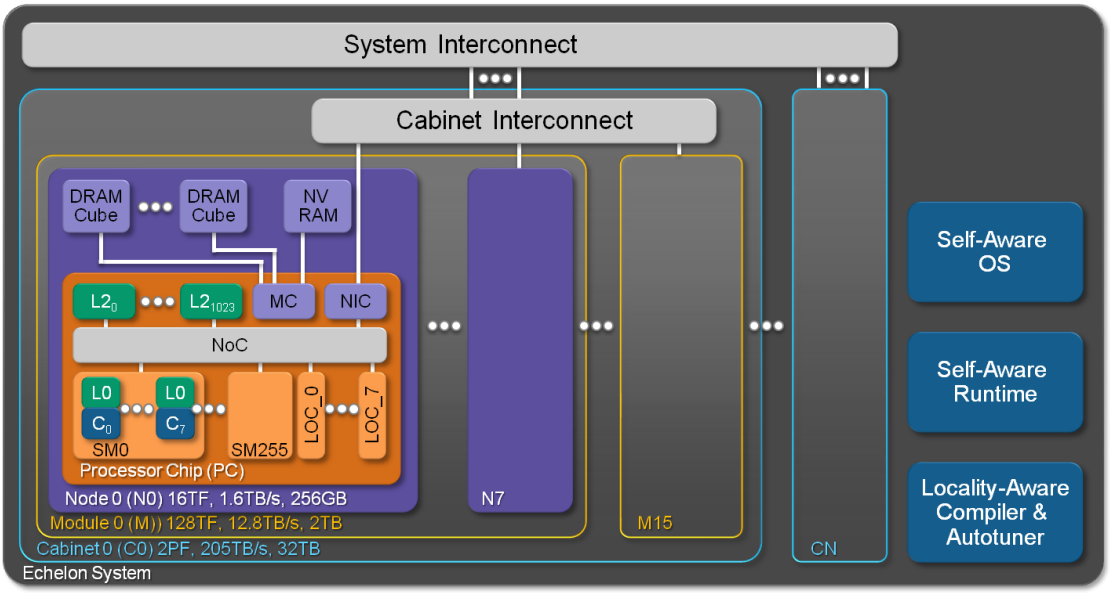


Рисунок 4б - Обобщенная схема проекта NVIDIA Echelon

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, применяют объединение процессоров в вычислительный модуль, реализуемое, например, 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 или Hypertransport. Возможны другие конструктивные элементы.

Для определенности в расчетах полагаем, что в результате разработки вычислительного модуля должно быть реализовано конструктивное объединение:

- 4 процессоров суммарной производительностью 1 Тфлопс / 8 Тфлопс в 2014г.;
- 8 процессоров суммарной производительностью 4 Тфлопс / 64 Тфлопс в 2017г.;
- 16 процессоров суммарной производительностью 16 Тфлопс / 256 Тфлопс в 2019г.

Дальнейший анализ выполним по отношению к последнему варианту – 16 MIMD/SIMD процессоров производительностью 16 Тфлопс / 256 Тфлопс – предназначенному для достижения 1 Ефлопс; другие варианты оцениваются аналогично.

В качестве каналов связи будем рассматривать каналы IB 12xHDR (480*480) Гбит/с, планируемые к выпуску в 2014г. [11]; пропускная способность одного линка составляет (40 + 40) Гбит/с = (5 + 5) Гбайт/с.

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

В качестве топологии рассмотрим для простоты и определенности гиперкуб, 3D тор и k-арный d-тор.

Полагаем, что вся система содержит $2^7 = 128$ стоек, в каждой стойке $2^7 = 128$ модулей, в каждом модуле $2^4 = 16$ процессоров.

Выбранные значения размерностей позволят варьировать параметрами в случае технологических трудностей на том или ином уровне.

Для реализации соединений внутри модуля достаточно линий связи длиной $l = 10$см. Чтобы объединить модули внутри стойки достаточно длины $l = 2$м. Для объединения стоек достаточно $l = 20$м.

Оценим количество линий связи, требуемое для соединения указанных процессоров при выполнении условия $\frac{\nu}{\pi} = 0{,}1$ Байт/флопс, где ν - суммарная пропускная способность средств обмена информацией между процессорным элементом и внешней средой, π - пиковая производительность процессорного элемента.

Именно такое соотношение выполняется в известных системах: для проекта IBM системы Blue Waters [12] π=1 Тфлопс, ν=192 Гбайт/с, $\frac{\nu}{\pi} = 0{,}2$; для системы Cray XE6 [13] π=210 Гфлопс, ν=25,6 Гбайт/с, $\frac{\nu}{\pi} = 0{,}12$; в эксафлопном проекте Nvidia [10] π=13 Тфлопс, ν=1 Тбайт/с, $\frac{\nu}{\pi} = 0{,}1$.

Коммуникационная среда первого уровня применяется для объединения 16 процессоров в вычислительный модуль. Коммуникационная среда второго уровня – для объединения вычислительных модулей в стойке. Коммуникационная среда третьего уровня объединяет стойки.

Для каждой из рассматриваемых топологий – Г-гиперкуб, 3D тор и 16-арный 5d тор [14], в таблице 2 указаны значения C_i – количество связей среды i-уровня и L_i – суммарная длина этих связей. Реализация среды второго уровня возможна применением

многослойных печатных кроссплат. Реализация среды третьего уровня, по-видимому, невозможна без применения многомодовых оптических средств связи.

В таблице 2 сведены полученные выше значения параметров коммуникационных сред. Символом D обозначено значение диаметра – наибольшего расстояния между процессорными элементами.

Таблица 2 – Значения параметров коммуникационных сред

	Уровень 1	Уровень 2	Уровень 3	D	$\frac{\nu}{\pi}$
	C_1, шт./ L_1, км	C_2, шт./ L_2, км	C_3, шт./ L_3, км		
Г	$6\cdot10^6$ / 629	$11\cdot10^6$ / $22\cdot10^3$	$11\cdot10^6$ / $220\cdot10^3$	16	$0{,}1 \div 0{,}05$
3D	$4\cdot10^6$ / 432	$1{,}5\cdot10^6$ / $3\cdot10^3$	$3\cdot10^6$ / $62\cdot10^3$	(8+16+64)=88	$0{,}036 \div 0{,}018$
k=16,d=5	$3\cdot10^6$ / 314	$3{,}1\cdot10^6$ / $6{,}3\cdot10^3$	$6\cdot10^6$ / $125\cdot10^3$	$5\cdot15=75$	$0{,}03 \div 0{,}015$

Приведенные в таблице 2 данные иллюстрируют достоинства и недостатки рассмотренных топологий, влияние которых понадобится оценивать на этапе создания систем, исходя из достигнутого технологического уровня.

3 Архитектурные средства масштабирования эффективности

Рассмотренная выше вычислительная система характеризуется следующими факторами:

— гибридная (неоднородная) структура процессорных элементов;

— сложность коммуникационной среды, выражающаяся в больших значениях диаметра (и соответственно больших задержках) и возможно большого разброса пропускной способности каналов, обусловленного большим диапазоном расстояний.

В этих условиях необходимы инструментальные средства, позволяющие учитывать в программном обеспечении архитектурные особенности вычислительной системы на различных уровнях параллелизма и обеспечивающие масштабирование эффективности.

Рассмотрим следующие средства архитектурного масштабирования эффективности:

— реконфигурация структуры процессорных элементов, состоящая в вариации количества MIMD и задействованных с ними SIMD ядер [15];

— минимизации длительностей обменов посредством декомпозиции процессов на подпроцессы в соответствии с особенностями вычислителей и размещение подпроцессов с учетом направлений обменов информацией между ними [16];

— топологическое резервирование, обеспечивающее выполнение заданного вычислительного процесса с заданной вероятностью на заданном количестве процессоров [17].

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 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 производительностью 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 в зависимости от соотношений между фрагментами вычислительного процесса и производительностью MIMD и SIMD компонент, выполняющих эти фрагменты. Предложены средства динамической реконфигурации структуры процессора, обеспечивающие разделение ядер MIMD и SIMD компонент на определенные, соответствующие друг другу подмножества (соединенные между собой), состав и производительность которых определяются в соответствии с параметрами исполняемого процесса.

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

В качестве иллюстрации сказанного приведем пример вычислений по программе молекулярной динамики [18].

Для процесса, выполняемого одним ядром за $T = 22{,}96\,\text{с}$ и состоящего из MIMD фрагмента, выполняемого за $T_M = 7{,}07\ \text{с}$, и SIMD фрагмента, выполняемого ускорителем за $T_S = 2{,}8\ \text{с}$, на рисунке 6 показано значение ускорения:

- $\tilde{K}_{q,1}$, достигаемое увеличением q - количества ядер и одним ускорителем;
- $K_{1,r}$, достигаемое увеличением r - количества ускорителей и одним ядром.

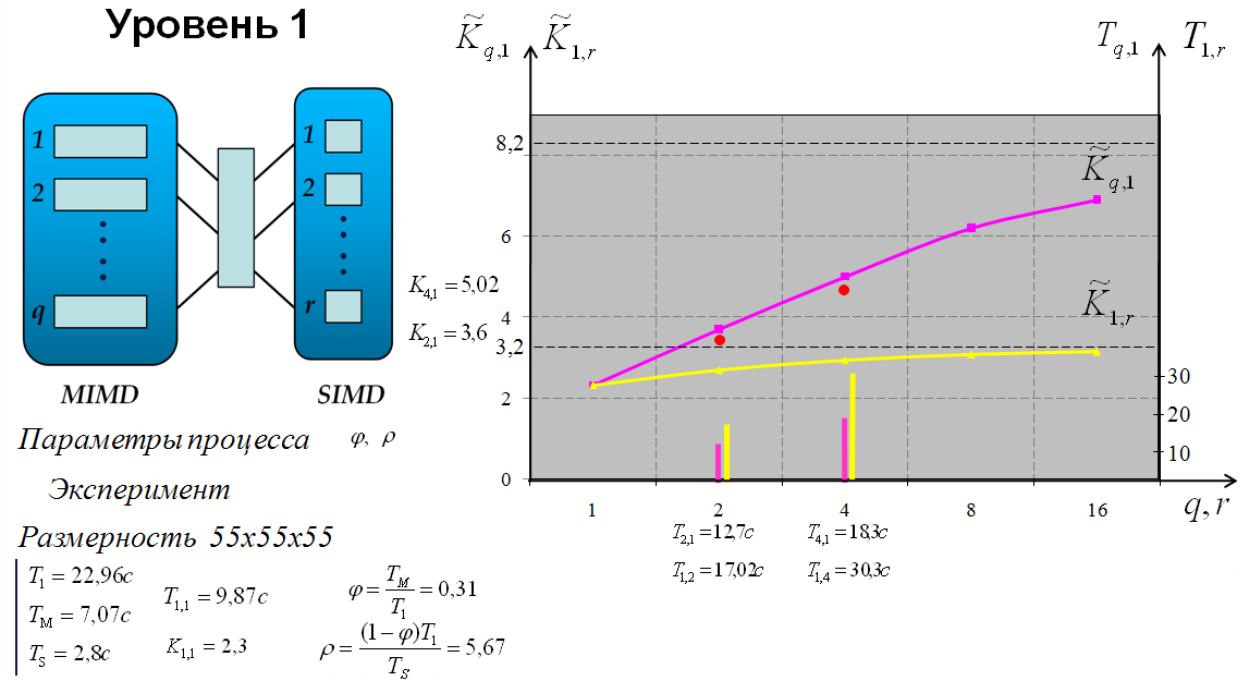


Рисунок 6 – Значения коэффициентов ускорения гибридными элементами различных конфигураций

Для данного процесса вычислитель, содержащий 16 ядер и один ускоритель, примерно в ~ 2,5 раза «быстрее» вычислителя, содержащего 1 ядро и 16 ускорителей.

3.2 Минимизация длительностей обменов

Эти средства предполагают, прежде всего, взаимную адаптацию вычислительного процесса и структуры связей между процессорными элементами с целью минимизации расстояний обменов и исключения конфликтов при выполнении обменов. Возможности адаптации зависят как от топологии вычислительной среды (2D, 3D, Γ^n, T^n), так и от свойств вычислительного процесса (явные схемы, регулярные связи и т.д.) С увеличением сложности машины актуальность и результативность этих средств возрастает.

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.

В общем случае в гиперкуб размерности n, обозначаемый Γ^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 согласно заданным параметрам исходного процесса и мультипроцессорной среды может применяться прикладная программа, результатом выполнения которой является таблица соответствия. В качестве исходных данных она передается системным средствам, которые загружают процессы на соответствующие процессоры.

Изложенные средства позволяют обеспечить требуемое исходному вычислительному процессу физическое соседство образующих его компонент. Они применимы в условиях современных аппаратных платформ – вычислительных модулей из нескольких, в частности, многоядерных процессоров на общей памяти.

В таблице 3 приведены согласно [16] значения производительности, достигнутые на тесте NASA LU. В столбце 1 – значения для варианта размещения процессов в соответствии со структурой связей вычислительной системы, приведенной на рисунке 7, в столбце 2 – значения производительности для последовательного размещения процессов, обычно реализуемого системным планировщиком. В частности, на тесте NASA 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 3 – Значения производительности на тесте NASA 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	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

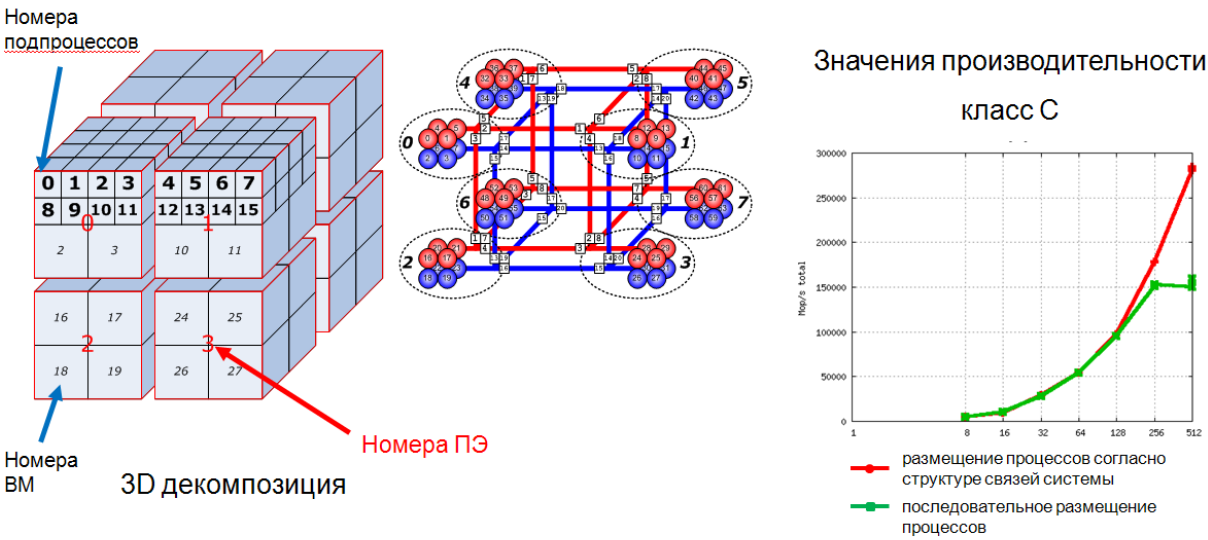


Рисунок 7 – Структура системы и значения производительности

Представленные в таблице 3 данные показывают, что эффект от применения декомпозиции и размещения подпроцессов возрастает с увеличением количества процессорных элементов (ядер), задействованных в процессе вычисления.

3.3 Средства топологического резервирования

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.

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.

Это может быть достигнуто применением методов топологического резервирования [16], позволяющих обеспечить в случае отказов и сбоев неизменность топологии среды и ее производительности; в результате полностью исключается 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 модулю (резервирование процессорных элементов, резервирование модулей) и т.п. позволяет создавать среды с наперед заданными значениями вероятностей исполнения вычислительного процесса определенной длительности.

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 основан на введении избыточных процессорных элементов.

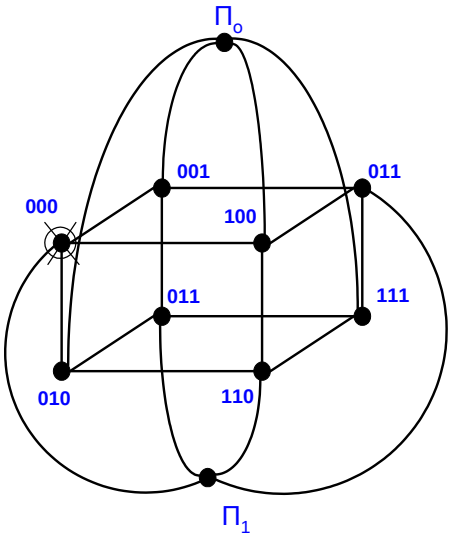


Рисунок 8 – Резервирование Γ^3 избыточными элементами

Второй метод обеспечивает выполнение определенного вычислительного процесса, задействуя в случае необходимости в качестве резервных элементы, выполняющие другие менее «важные» процессы; последние в случае резервирования удаляются.

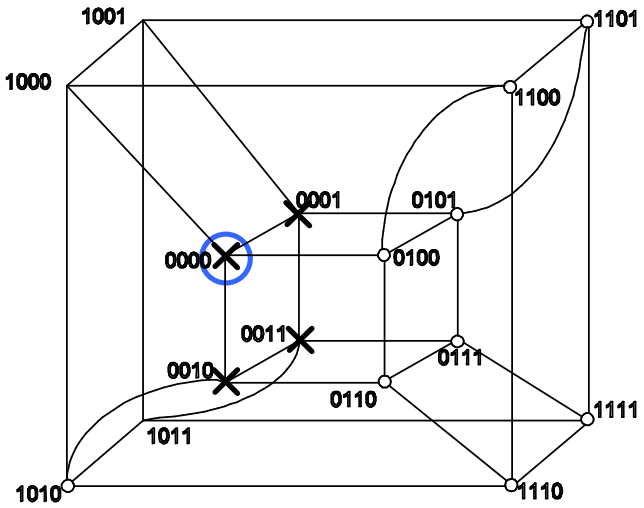


Рисунок 9 – Топологическое резервирование Γ^4 выделенными элементами

Оба метода могут быть применены для сред с различными топологиями, допускающими описание элементов кодами Грэя. К этим топологиям, в частности, относятся 1D тор, 2D тор, 3D тор, k-арный d-тор и другие. Разумеется, различные топологии обеспечивают различные оценки вероятностей выполнения вычислительного процесса.

4 Уровни параллелизма и контуры адаптируемой архитектуры

Из приведенного выше следует, что весьма перспективным вариантом достижения эксафлопной 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 адаптации к особенностям исполняемой программы.

Реализуемая реконфигурация – есть средство создания архитектуры, 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. Однако их задействование в соответствии с особенностями исполняемого вычислительного процесса позволяет ускорить вычисления.

Распределение ядер в процессорных элементах согласно декомпозиции процессов, маршрутизацию в соответствии с топологией среды и оптимальное размещение процессов по процессорным элементам, учитывающее направления обменов, можно рассматривать как средство адаптации структуры вычислительной системы к исполняемой программе на втором и третьем уровнях параллелизма.

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 эксафлопной 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.

По-видимому, ни эти инструментальные средства, ни тем более результаты их применения в виде адаптированных под заданные программы архитектур (структур), не будут поставляться открыто, например, в силу закрытости программ и неизвестности их параметров.

Создание вычислительных систем эксафлопного класса в силу их сложности в течение длительного времени будет носить единичный характер и требует интегрированной разработки новых архитектур, адаптируемых в соответствии с особенностями вычислительных процессов.

Заключение

В работе 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:

- реконфигурацию структуры гибридных процессорных элементов в соответствии с особенностями исполняемого процесса;

- декомпозицию вычислительных процессов на подпроцессы и оптимизацию размещения последних с целью уменьшения длительностей обменов;

- средства топологического резервирования, позволяющие обеспечить с заданной вероятностью безотказное выполнение вычислительного процесса определенной длительности, занимающего заданное количество процессорных элементов; это достигается либо резервированием избыточными элементами (при необходимости – неоднократным), либо применением в качестве резервных тех элементов среды, которые не задействованы данным процессом.

В результате достигается 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., 2004г.

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 [Электронный ресурс]. Режим доступа: www.exascale.org

6 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 An Nvidia Exascale Machine in 2017.

9 Tomohiro Inoue. Fujutsu Limited. The 6D Mesh/Torus Interconnect of K Computer.

10 Bill Dally Chief Scientist, NVIDIA Bell Professor of Engineering, Stanford University. From Here to ExaScale Challenges and Potential Solutions

11 Infiniband Roadmap 072611

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 2010г. М.: Учреждение Российской академии наук. Институт проблем управления им. В.А.Трапезникова РАН стр.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. 2009г., вып.2.