УДК 519.6

ПРОГНОЗ ПАРАМЕТРОВ ПОДСИСТЕМ ВЫЧИСЛИТЕЛЬНОЙ СИСТЕМЫ ЭКСАФЛОПСНОГО КЛАССА

Ю. Г. Бартенев, Ю. А. Бондаренко, В. Ф. Спиридонов
(РФЯЦ-ВНИИЭФ)

Приводятся оценки возрастаний пропускной способности коммуникационной сети в целом, объема памяти всех уровней и производительности графической подсистемы в зависимости от увеличения числа вычислительных узлов (серверов) и роста их производительности. В качестве модели вычислительного алгоритма использованы явные разностные схемы для трехмерных задач газовой динамики. Предполагается, что при переходе от существующей ЭВМ к проектируемой вычислительной системе эксафлопсного класса время счета одной задачи на всем вычислительном поле не меняется.

Ключевые слова: суперЭВМ, параметры вычислительной системы, масштабирование вычислительной системы, пропускная способность системы межпроцессорных обменов, быстродействие и объем оперативной и внешней памяти, быстродействие системы графической обработки, трехмерные нестационарные задачи, явные разностные схемы.

Введение

Определению параметров подсистем (коммуникационной, оперативной и внешней памяти, графической обработки) ЭВМ при росте ее производительности всегда уделялось достаточное внимание [1—4]. Предлагаемый в данной работе прогноз строится на основе формул для задач трехмерного моделирования, отражающих возрастание количества вычислений на расчет газодинамического процесса по явной схеме при повышении подробности сетки. Отметим, что подход к оценке характеристик перспективной ЭВМ, основанный на модели вычислений, аналогичен подходу И. Д. Софронова [1].

Цель данной работы — дать разработчикам вычислительной системы (суперЭВМ) эксафлопсного класса ориентиры на длительную перспективу по необходимым параметрам: пропускной способности коммуникационной подсистемы, быстродействию подсистем памяти, объему подсистем памяти и других подсистем в зависимости от быстродействия суперЭВМ и ее вычислительных компонентов. Ожидается, что использование таких ориентиров в виде математических соотношений позволит при проектировании мощной суперЭВМ более точно определить задачи по ее разработке.

Обычно считается, что указанные характеристики зависят линейно от производительности вычислителя. Но это оправданно при небольшом возрастаниии производительности. Практика показывает, что такая простая зависимость тем больше завышает достаточный для эффективного использования суперЭВМ уровень требований, чем более мощная суперЭВМ создается.

При решении задачи в одной и той же геометрической области в одном и том же интервале физического времени исходная математическая модель (модель явной газовой динамики на структурированных сетках, например, построенная по методике ЛЭГАК [5]) дает сверхлинейные (нелинейные в степени $k > 1$) зависимости количества вычислений от объема обрабатываемых данных с $k = 4/3$ (кубическая сетка), $k = 5/3$ (цилиндрическая сетка) и $k = 6/3 = 2$ (сферическая сетка) при линейной зависимости шага по времени от минимального диаметра ячейки сетки (он уменьшается линейно, квадратично, кубично при сгущении указанных сеток во всех направлениях). Модель отражает характерную для сеточных методов декомпозицию трехмерной счетной сетки на подобласти (параобласти), преимущественную схему обмена данными меж-

ду соседними параобластями, при которой объем передаваемой на каждом временном шаге информации пропорционален числу ячеек (и количеству вычислений) параобласти в степени 2/3.

По-видимому, подобные (выше линейной) зависимости количества вычислений от объема обрабатываемой информации присущи не только методикам явной газовой динамики, но и связанным с ними методикам (модели теплопроводности, переноса частиц), использующим аппроксимацию исходных дифференциальных уравнений на сетках. Одновременное с увеличением области увеличение числа ячеек в газодинамической задаче, увеличение числа частиц при молекулярно-динамическом моделировании ослабляют указанные зависимости, вплоть до линейной зависимости шага по времени от шага по сетке (или от числа частиц). Ряд других факторов, наоборот, усиливают степень нелинейности. Например, использование неявных схем, приводящих к решению сеточных уравнений, даст дополнительный множитель нелинейности, по-видимому, не меньше, чем логарифм от отношения числа ячеек двух сеток.

Все эти случаи, включая ориентацию суперЭВМ на специфику превалирующих классов задач, расчетные технологии и их изменение в перспективе, учесть затруднительно. Поэтому грубые оценки параметров суперЭВМ должны уточняться на этапах средне- и краткосрочных заданий на проектирование.

Предлагаемые ниже оценки представлены в виде формул, которые дают диапазон выбора зависимости между количеством вычислений (в флопах), и объемом данных задачи (в байтах). Их можно назвать соотношениями подобия при переходе от одной ЭВМ к другой, сохраняющем подсистемы второй ЭВМ настолько же сбалансированными, насколько они сбалансированы в первой ЭВМ. В этом диапазоне из определенных соображений указываются наиболее надежные с точки зрения авторов соотношения, которые далее применяются для оценки параметров суперЭВМ с повышенной производительностью.

Неформально, *сбалансированными* характеристиками подсистем (вычислительной, коммуникационной, памяти и др.) суперЭВМ будем называть такие характеристики, которые позволяют считать в многозадачном режиме все целевые классы задач (и обрабатывать результаты расчетов), в том числе основные классы задач с использованием не менее половины вычислительного ресурса, за некоторое допустимое время с достаточной для потребителя эффективностью распараллеливания. Для большей определенности можно сказать, что на представительных с точки зрения потребителя тестовых задачах характерного размера должна достигаться производительность не ниже заданной.

Более подходящим показателем является *относительная сбалансированность*, учитывающая опыт счета на некоторой существующей ЭВМ. Если ее характеристики принять за приемлемый эталон эффективности счета задач[1], то новую ЭВМ будем считать сбалансированной, если эффективность счета задач, соответствующая увеличенной вычислительной емкости, будет такой же, как на существующей (*старой*) ЭВМ.

Применяя предлагаемые формулы, можно проектировать новую ЭВМ *сбалансированным образом*, используя параметры старой ЭВМ, если параметры ее подсистем сбалансированы. Для проверки применимости формул для эксафлопсной ЭВМ прогнозируемые параметры сравниваются с другими существующими прогнозами.

Математическая модель

Математическая модель базируется на следующих положениях:

- при увеличении объема данных задачи на новой ЭВМ относительно старой ЭВМ в D раз возрастание количества вычислений A определяется формулой $A = D^k$ (рассматриваются случаи $k = 1,\ 4/3,\ 5/3,\ 2$);
- суперЭВМ состоит из множества связанных коммуникационной сетью (без учета структуры) вычислительных узлов (без учета их структуры), на каждом из которых размещаются данные одной параобласти задачи;
- объем *обменной* информации одного занимаемого задачей вычислительного узла суперЭВМ (данных граничных ячеек параобласти, которыми он обменивается с другими узлами) увеличивается, как объем обрабатываемых этим узлом данных в степени 2/3;

[1] Вообще говоря, усложняя и удорожая машину, формально можно получать более высокую эффективность счета, однако затраты на ее создание могут не окупить прирост реальной производительности, который ограничен, в том числе и алгоритмической эффективностью распараллеливания реальных задач. Поэтому всегда приходится останавливаться на каком-то приемлемом уровне эффективности распараллеливания.

– физическая модель не меняется, количество вычислений в одной ячейке на одном шаге по времени не зависит от числа ячеек сетки; соответственно количество вычислений (в флопах) на одном шаге по времени возрастает пропорционально D (увеличению числа ячеек сетки), а общее количество вычислений в задаче возрастает за счет D и увеличения числа временных шагов (для рассматриваемой модели в D^{k-1} раз);
– размер одной записи в файловую систему новой ЭВМ увеличивается в D раз; задача на новой ЭВМ делает такое же число записей (считаем, что это число определяется характерными точками эволюции моделируемого процесса и не зависит от подробности сетки) и за то же время счета, что на старой ЭВМ;
– графическая система новой ЭВМ обрабатывает в D раз увеличенный объем данных за то же время, что на старой ЭВМ;
– одна соответственно увеличенная большая задача решается на всем вычислительном поле новой ЭВМ за то же время, что на старой ЭВМ (считаем, что это относится и к случаю одновременного счета множества задач — каждая соответственно увеличенного объема).

Пусть

T — возрастание времени счета на новой ЭВМ относительно старой;

n — увеличение числа вычислительных узлов (серверов) вычислительного поля на новой ЭВМ относительно старой;

p — возрастание производительности вычислительного узла;

$P = np$ — возрастание производительности новой ЭВМ относительно старой;

s — необходимое возрастание быстродействия памяти вычислительного узла (сервера);

b — необходимое возрастание пропускной способности внешнего коммуникационного интерфейса вычислительного узла (сервера);

$B = nb$ — возрастание пропускной способности коммуникационной сети в целом[2];

V — необходимое увеличение объема памяти всех уровней: оперативной, внешней (файловой подсистемы);

S — необходимое возрастание быстродействия всех уровней файловой подсистемы;

G — необходимое возрастание производительности графической подсистемы.

Проведя выкладки (подробно они даны в Приложении), получим соотношения, отраженные в табл. 1.

Естественное предположение о неизменности времени счета задачи ($T = 1$, что принимается далее) при возрастаниях A и P соответствует тому, что более мощная суперЭВМ нужна для решения более вычислительноемких задач, которые невозможно решить за разумные сроки на старой ЭВМ. Кроме того, указанные в табл. 1 соотношения обеспечивают сохранение достигнутой эффективности счета.

Каждым из указанных соотношений можно пользоваться в зависимости от того, какой класс задач будет превалировать на суперЭВМ. Случаи $k = 1$ и $k = 2$ дают крайние значения в этих

Таблица 1

Параметры ЭВМ в зависимости от увеличения производительности и числа вычислительных узлов

Параметр	$k = 1$	$k = 4/3$	$k = 5/3$	$k = 2$
$b = \frac{p^{1-1/(3k)} n^{1/3-1/(3k)}}{T^{1/(3k)}}$	$\frac{p^{2/3}}{T^{1/3}}$	$\frac{p^{3/4} n^{1/12}}{T^{1/4}}$	$\frac{p^{4/5} n^{2/15}}{T^{1/5}}$	$\frac{p^{5/6} n^{1/6}}{T^{1/6}}$
$B = \frac{p^{1-1/(3k)} n^{4/3-1/(3k)}}{T^{1/(3k)}}$	$\frac{p^{2/3} n}{T^{1/3}}$	$\frac{p^{3/4} n^{5/4}}{T^{1/4}}$	$\frac{p^{4/5} n^{17/15}}{T^{1/5}}$	$\frac{p^{5/6} n^{7/6}}{T^{1/6}}$
$S = (TP)^{1/k}$	TP	$(TP)^{3/4}$	$(TP)^{3/5}$	$(TP)^{1/2}$
$V = (TP)^{1/k}$	TP	$(TP)^{3/4}$	$(TP)^{3/5}$	$(TP)^{1/2}$
$G = (TP)^{1/k}$	TP	$(TP)^{3/4}$	$(TP)^{3/5}$	$(TP)^{1/2}$

[2]Предполагаем, что коммуникационная система новой ЭВМ обеспечивает такое же отношение реальной пропускной способности к пиковой, как на старой ЭВМ.

неравенствах. Тем не менее ряд задач может удовлетворять этим условиям. Например, $k = 1$ при увеличении числа ячеек одновременно с увеличением размеров расчетной области; $k = 2$ при кубической зависимости шага по времени от шага регулярной сетки. Эти случаи отбрасываем, не считая характерными для общей массы задач.

Тогда следует придерживаться следующего "коридора" значений характеристик при переходе от старой ЭВМ к новой:

$n^{1/12}p^{3/4} < b < n^{2/15}p^{4/5}$ и наиболее надежные $b = n^{2/15}p^{4/5}$ и $B = n^{1/3}P^{4/5}$;

$P^{3/5} < V < P^{3/4}$ и наиболее надежное $V = P^{3/4}$;

$P^{3/5} < G < P^{3/4}$ и наиболее надежное $G = P^{3/4}$;

$P^{3/5} < S < P^{3/4}$ и наиболее надежное $S = P^{3/4}$;

$s \geq p$ (скорость выполнения ряда алгоритмов ограничивается быстродействием памяти).

Прокомментируем соотношения, которые представляются авторам наиболее надежными. Формально они соответствуют наиболее сильному условию (максимизирующему параметр) в указанном диапазоне. Равенство $b = n^{2/15}p^{4/5}$ формально соответствует соотношению $A = D^{5/3}$, которое (не говоря уже о большей степени $A = D^2$) из-за "дороговизны" вряд ли можно ожидать от методов и прикладных программ, предназначенных для суперЭВМ эксафлопсного класса. Можно было бы придерживаться соотношения $A = D^{4/3}$, к которому приводят некоторые методики счета. Однако здесь не учтен ряд факторов, несколько усиливающих коммуникационный поток, как то: коллективные операции с небольшим объемом данных, обмен данными по ребрам и вершинам ячеек параобластей, многократный обмен данными граничных ячеек на одном шаге по времени в итерационных решателях СЛАУ и обмен данными (не только граничными) параобластей при построении некоторых предобусловливателей, обмен данными при вычислительной балансировке и др. Поэтому для надежности прогноза поднимаем коммуникационные требования до уровня $A = D^{5/3}$.

С другой стороны, для требуемого быстродействия и объема периферийных систем не следует ориентироваться на снижающее параметр соотношение $A = D^{5/3}$ ввиду "аддитивных" добавок. Например, распределенная по вычислительным узлам оперативная память должна содержать в каждом узле данные системы управления и буфера обмена информации, объем которых не уменьшится на новой ЭВМ; наряду с большими задачами на новой суперЭВМ будут считаться задачи класса старой ЭВМ с таким же объемом памяти, число которых, как правило, не уменьшается; с увеличением числа вычислительных узлов снижается надежность, вызывающая повышение нагрузки на файловую систему для записи контрольных точек; также она повышается из-за увеличения количества записей для более детального исследования процесса в характерных точках. Поэтому для надежности прогноза поднимаем требования к объемам памяти до уровня $A = D^{4/3}$.

Сделаем некоторые пояснения и выводы.

1. Отношения возрастаний объема памяти, производительности файловой и графической подсистем к возрастанию производительности ЭВМ с увеличением производительности ЭВМ уменьшаются (при $P > 1$ получаем $S/P < 1$, $V/P < 1$, $G/P < 1$).

2. Отношение B/P возрастания пропускной способности коммуникационной системы ЭВМ к увеличению производительности ЭВМ, а также отношение b/p возрастания пропускной способности коммуникационного интерфейса вычислительного узла к увеличению производительности узла возрастают при увеличении числа узлов без повышения их производительности (при $p = 1$, $n > 1$ и $k = 5/3$ получаем $B/P = b/p = n^{2/15} > 1$).

При возрастании производительности ЭВМ за счет возрастания производительности узлов и неизменном их количестве эти отношения уменьшаются (при $p > 1$, $n = 1$ и $k = 5/3$ получаем $B/P = b/p = p^{-1/5} < 1$).

В общем случае

$B/P = b/p = p^{-1/(3k)}n^{1/3-1/(3k)}$,

и $B/P = b/p = 1$ при $p = n^{k-1}$.

Для $k = 5/3$ получаем

$B/P = b/p = (n/p^{3/2})^{2/15}$,

и $B/P = b/p = 1$ при $n = p^{3/2}$.

Если алгоритм распараллеливания требует обмена данными всей параобласти (что не свойственно геометрической декомпозиции), а не только относящимися к границе, то нужно брать $b = p$. Не менее сильное соотношение требуется в случае ускорения счета в P раз задач класса старой ЭВМ, основанных на геометрической декомпозиции ($b = pn^{1/3}$). Является ли актуальным счет таких задач на всей суперЭВМ эксафлопсного класса и в каком объеме — вопрос открытый. Можем только сказать, что рассмотренные более слабые зависимости пригодны для прогнозирования параметров ЭВМ, ориентиро-

ванной на массовый счет широкого класса актуальных задач, требующих эксафлопсной производительности.

Предлагаемые зависимости b и B на самом деле не такие уж слабые, и их достижение требует больших усилий.

Во-первых, для массы задач достаточна более слабая зависимость.

Во-вторых, если коммуникационные требования обеспечивают эффективный счет задачи на всем вычислительном поле ЭВМ, то они обеспечат не менее эффективный счет более мелких задач на части ЭВМ. Например, проведение одного и того же числа расчетов задач класса старой ЭВМ на новой ЭВМ, имеющей в p раз меньше узлов, будет более эффективным, чем на старой ЭВМ[3]. Вычислительно менее сложная задача (на более грубой сетке) на пропорционально меньшем ресурсе посчитается быстрее[4].

В-третьих, десятилетнее движение от терафлопса к петафлопсу (увеличение производительности суперЭВМ в 1000 раз) дало ускорение двухпроцессорных узлов приблизительно в 100 раз и рост пропускной способности их коммуникационного интерфейса примерно в 40—50 раз. Согласно формулам предстоящее ускорение в 100 раз вычислительных узлов и в 1000 раз всей ЭВМ требует не меньшего ускорения коммуникации ($b = 50$).

Отметим, что рассматривается именно внешний интерфейс узлов. Если узел — *составной*, т. е. состоит из m связанных внутренней сетью узлов с суммарным возрастанием их производительности $p_0 = pm$ и суммарным возрастанием пропускной способности их коммуникационного интерфейса $b_m = bm$, и как целое имеет внешний интерфейс с другими такими же составными узлами, то возрастание пропускной способности внешнего интерфейса b_0 составного узла связано с b_m, как

$$b_0 = b_m^{2/3},$$

если данные составных параобластей располагать строго компактно на составных узлах. Отметим, что это не зависит от k в формуле $A = D^k$, что следует из $P =$const в формуле

$$b = p^{2/3} P^{1-1/3k}, \qquad (1)$$

которая получается при исключении n из формулы для b в табл. 1.

Это дает основание для разработки иерархической структуры коммуникационной системы суперЭВМ, где более крупные вычислители имеют меньшую удельную (относительно производительности) пропускную способность, чем более мелкие вычислители.

Конечно, лучше, если коммуникационная сеть будет иметь большую пропускную способность, чем по формулам. Но ввиду неполного распараллеливания вычислений, дисбаланса вычислений в реальных задачах повышение скорости сверх достаточной не всегда оправданно (дороже, сложнее без должной отдачи).

Сравнение с другими прогнозами

Сравним прогноз на основе рассмотренных соотношений с прогнозами характеристик эксафлопсной ЭВМ, сделанными Ж. Донгарра [2] и А. Гейстом [3], путем рассмотрения в качестве старой суперЭВМ вычислительной системы Jaguar, установленной в ANL США.

Как видно из табл. 2, предлагаемые формулы в основном дают значения, попадающие в коридор значений авторов работ [2, 3]. "Выбивается" из него объем внешней памяти, потребность в которой оба автора оценивают ниже, чем авторы данной статьи. Возможно, внешняя память Jaguar реализована "с запасом". Прогноз, предлагаемый в данной статье, наиболее близок прогнозу Донгарра.

Предложение параметров эксафлопсной ЭВМ

Прогноз на 2020 г. — довольно реалистичный срок возможного создания эксафлопсной ЭВМ (рассматриваем вариант 1,3 Эфлопс, чтобы получить 1 Эфлопс на тесте Linpack) — в России вряд ли должен опираться на более чем двукратное увеличение производительности вычислительного узла каждые 2 года. Это приводит в 2020 г. к 15—30 Тфлопс с двойной точностью для *легкого* узла, если с указанным темпом масштабировать производительность самого производительного — гибридного варианта узла.

[3] Объемы данных параобластей возрастут в p раз, их границ — в $p^{2/3}$, а скорость обмена — более чем в $p^{4/5}$ раз при $n > 1$.

[4] При уменьшении в 8 раз сетки задачи (в 16 раз — вычислений) и в 16 раз — числа вычислительных узлов время вычислений на временном шаге возрастает в 2 раза, а время обмена — только в $2^{2/3}$раза.

Таблица 2

Сравнение предлагаемого прогноза с прогнозами Донгарра и Гейста

Параметр	Jaguar	Прогноз		
		Донгарра	Гейста	предлагаемый
Пиковая производительность ЭВМ	2,34 Пфлопс	1 Эфлопс	1 Эфлопс	1,33 Эфлопс
Объем оперативной памяти ЭВМ	0,3 Пбайт	32—64 Пбайт	10 Пбайт	35 Пбайт
Производительность узла	125 Гфлопс	1,2—15 Тфлопс	1—10 Тфлопс	15 Тфлопс
Быстродействие памяти узла	25 Гбайт/с	2—4 Тбайт/с	0,2—0,4 Тбайт/с	$\geqslant$ 3 Тбайт/с
Пропускная способность коммуникационного интерфейса узла	3,5 Гбайт/с	200—400 Гбайт/с	50 Гбайт/с*	198 Гбайт/с
Число узлов в ЭВМ	18 700	$\sim 10^5$—10^6	$\sim 10^6$	88 887
Объем внешней памяти ЭВМ	15 Пбайт	500—1 000 Пбайт	300 Пбайт	1 750 Пбайт
Быстродействие ввода-вывода ЭВМ	0,2 Тбайт/с	60 Тбайт/с	20 Тбайт/с	24 Тбайт/с

*Возможно, Гейст имел в виду быстродействие одного из шести каналов коммутатора узла, так как указал аналогичный параметр в Jaguar 2009 г. равным 1,5 Гбайт/с.

В настоящее время легкий гибридный узел ($\sim$0,6—1,1 Тфлопс) — это один универсальный многоядерный микропроцессор и один сопроцессор-ускоритель; легкий универсальный узел ($\sim$0,15 Тфлопс) — это два универсальных многоядерных микропроцессора.

Надежнее рассчитывать в 2020 г. на максимальную производительность вычислительного узла, достигаемую в 2018 г., — 15 Тфлопс[5]. Это приводит приблизительно к 90 000 узлам, что сопоставимо с самыми большими по числу узлов суперЭВМ настоящего времени — Jaguar ($\sim$ 19 000) и Blue Gene ($\sim$ 70 000), и значит, конструкция такой сложности вполне реализуема.

На основании этих выводов, выведенных формул, опыта разработки и эксплуатации ЭВМ, проведения расчетов и с учетом мнения ряда зарубежных специалистов можно дать следующие ориентиры для параметров подсистем отечественной эксафлопсной ЭВМ:

- быстродействие оперативной памяти узла $\sim$ 2—4 Тбайт/с;
- пропускная способность коммуникационного интерфейса узла $\sim$100—400 Гбайт/с[6];
- объем оперативной памяти ЭВМ $\sim$ 30 Пбайт, узла $\sim$350—700 Гбайт;
- объем параллельной файловой системы $\sim$ 600 Пбайт;
- объем системы долговременного хранения не более 5 000 Пбайт;
- производительность графической подсистемы $\sim$2 Пфлопс.

Указанные значения представляются доступными технически и достаточными для эффективного счета больших задач — примерно с 10^{12}—10^{13} ячеек сетки и несколькими килобайтами данных для каждой ячейки. По-видимому, число ячеек порядка $O\left(10^6\right)$ на одном узле будет являться нижним пределом для эффективного распараллеливания на всем множестве узлов.

Создание эксафлопсной ЭВМ из *тяжелых* вычислительных узлов, например составных узлов, содержащих 16 универсальных микропроцессоров и их сопроцессоров-ускорителей, объединенных внутренней сетью, позволяет поднять производительность узла до $\sim$250 Тфлопс. Это приводит к следующим изменениям:

- число узлов $\sim$ 5 000;
- быстродействие оперативной памяти $\sim$ 40 Тбайт/с;
- пропускная способность коммуникационного интерфейса узла $\sim$ 0,75—1—3 Тбайт/с;
- объем памяти узла 6 Тбайт.

Полезное свойство *укрупнения* узлов связано со снижением пропускной способности коммуникационного интерфейса узла относительно еди-

[5]По прогнозам [4, 5] производительность вычислительного узла в 2018 г. составит 10 Тфлопс.

[6]400 Гбайт/с, вероятно, "покроет" весь мыслимый набор приложений эксакласса 2020 г.

ницы производительности вычислительного узла и снижением объема оборудования коммуникационной сети. Например, при повышении производительности узла в 16 раз пропускную способность интерфейса достаточно поднять в 6,34 раза (согласно формуле (1)).

Однако вариант создания суперЭВМ из тяжелых узлов приводит не к решению проблемы, а к ее переформулировке, и как ее решать проще — сейчас окончательно не ясно. По-видимому, более целесообразно снижать требования на элементы суперЭВМ за счет повышения ее сложности, чем наоборот.

Кроме того, такой вариант суперЭВМ до сих пор приводил к их удорожанию, после 2005 г. такие суперЭВМ с рекордной производительностью не строились. Одна из последних суперЭВМ на тяжелых узлах ASCI White (100 Тфлопс) стоила $\sim$200 млн долларов. Это обусловлено повышением стоимости разработки таких вычислительных узлов (серверов) и меньшей востребованностью их на рынке (штучный товар).

Заключение

Рассмотренные ориентировочные оценки для параметров суперЭВМ эксафлопсного класса не претендуют на завершенность, и хотелось бы надеяться на отклик в виде их анализа на базе иных моделей вычислений эксафлопсного класса. Это способно привести к уточнению параметров суперЭВМ или их разнообразию благодаря ориентации на разные классы алгоритмов.

Приложение. Вывод формул подобия

Поясним подробнее вывод формул, приведенных в табл. 1.

В основе приводимых здесь рассуждений лежит предположение, что на узле (вычислительном сервере с большим числом процессоров на общей памяти) рассчитывается связный кусок задачи, так что все межузловые обмены информацией можно рассматривать, не вникая в особенности алгоритмов распараллеливания (на общей памяти) внутри узла. Узел рассматривается как независимый вычислитель. Коммуникационная сеть взаимодействует с узлом только как с неделимой единицей и тем самым обеспечивает обмен информацией между узлами.

Рассматривается явный алгоритм, типичный для нестационарных задач трехмерной газовой динамики, в предположении, что используется метод геометрической декомпозиции с нарезкой на подобласти (параобласти) с топологией соседства типа трехмерной решетки (если пренебрегать взаимодействием по ребрам и вершинам ячеек — потоки информации в диагональных направлениях через ребра и вершины считаем пренебрежимо малыми по сравнению с потоками информации через грани ячеек).

Введем обозначения, описывающие параметры трехмерной задачи:

N — полное число точек сетки в задаче;

N_1, N_2 — полные числа точек в трехмерной задаче для старой и новой ЭВМ соответственно;

N_{par} — полное число точек сетки в параобласти, рассчитываемой на одном узле (сервере);

n_1, n_2 — полные числа узлов для старой и новой ЭВМ соответственно;

$n = n_2/n_1$ — увеличение числа узлов (серверов);

p — возрастание производительности одного узла (сервера);

$P = np$ — возрастание полной производительности ЭВМ;

$C = N^k$ ($k = 4/3,\ 5/3,\ 2$) — условная *стоимость* трехмерного расчета (число арифметических операций с точностью до коэффициента пропорциональности равно произведению числа точек сетки N на число шагов по времени; число шагов по времени прямо пропорционально $N^{1/3}$ для кубической сетки, $N^{2/3}$ для полистовой сетки с квадратной сеткой в листах и $N^{3/3} = N$ для сетки сферического типа)[7];

$T = \dfrac{C_2/P_2}{C_1/P_1} = \dfrac{(N_2/N_1)^k}{P_2/P_1} = \dfrac{(N_2/N_1)^k}{P}$ — возрастание астрономического времени на расчет.

Отсюда получаем:

$D = N_2/N_1 = (TP)^{1/k} = (Tnp)^{1/k}$ — увеличение полного числа точек сетки в трехмерной задаче, или объема данных;

$A = C_2/C_1 = (N_2/N_1)^k = D^k$ — возрастание количества вычислений на расчет;

$\dfrac{N_{par(2)}}{N_{par(1)}} = \dfrac{N_2/n_2}{N_1/n_1}$ — увеличение числа точек сетки трехмерной задачи на одном узле.

Отсюда

[7] Такие зависимости связаны с тем, что в явных схемах обычно шаг по времени прямо пропорционален минимальному геометрическому размеру ячейки сетки и в случае полистовой сетки и сферической сетки вблизи *центра* и оси симметрии зависит дополнительно от угла между листами и между сферическими столбцами соответственно.

$\frac{N_{par(2)}}{N_{par(1)}} = \frac{N_2/n_2}{N_1/n_1} = \frac{N_2/N_1}{n} = \frac{(Tnp)^{1/k}}{n} = \frac{(Tp)^{1/k}}{n^{1-1/k}}$ — увеличение полного числа точек сетки трехмерной задачи в одной параобласти (на одном узле);

$\left(\frac{N_{par(2)}}{N_{par(1)}}\right)^{2/3} = \left[\frac{(Tp)^{1/k}}{n^{1-1/k}}\right]^{2/3} = \frac{(Tp)^{2/(3k)}}{n^{2/3-2/(3k)}}$ — увеличение площади поверхности параобласти, т. е. количества информации, пересылаемой на одном шаге по времени от одного узла к другому;

$N_t = C/N = N^{k-1}$ — полное число шагов по времени в одном трехмерном расчете;

$t = N_{t2}/N_{t1} = (N_2/N_1)^{k-1} = \left[(Tnp)^{1/k}\right]^{k-1} = (Tnp)^{1-1/k}$ — увеличение полного числа шагов по времени в одном трехмерном расчете;

$\left(\frac{N_{par(2)}}{N_{par(1)}}\right)^{2/3} t = \frac{(Tp)^{2/(3k)}}{n^{2/3-2/(3k)}}(Tp)^{1-1/k} n^{1-1/k} = (Tp)^{1-1/(3k)} n^{1/3-1/(3k)}$ — увеличение количества информации, пересылаемой от одного узла к другому за все время проведения одного трехмерного расчета;

b_1, b_2 — пропускные способности межузловых каналов связи для старой и новой ЭВМ соответственно;

$b = b_2/b_1$ — возрастание пропускной способности каналов связи между узлами (внешнего коммуникационного интерфейса узлов).

Исходим из предположения, что при переходе к новой ЭВМ количество узлов, с которыми обменивается информацией данный узел, не меняется (это некоторая постоянная, равная числу геометрических соседей). Поэтому потери времени на межузловые обмены за все время счета одной трехмерной задачи возрастают в следующее число раз:

$$\frac{(N_{par(2)}/N_{par(1)})^{2/3} t}{b} = \frac{(Tp)^{1-1/(3k)} n^{1/3-1/(3k)}}{b}.$$

Потребуем, чтобы возрастание полного времени на межузловые обмены было равно возрастанию астрономического времени на полный счет задачи, т. е. T. Это требование можно назвать гипотезой о том, что новая ЭВМ так же самосогласованна, как и старая. Из него получаем уравнение

$$\frac{(N_{par(2)}/N_{par(1)})^{2/3} t}{b} = \frac{(Tp)^{1-1/(3k)} n^{1/3-1/(3k)}}{b} = T,$$

откуда следует оценка

$$(Tp)^{1-1/(3k)} n^{1/3-1/(3k)} = Tb,$$

или

$$b = \frac{p^{1-1/(3k)} n^{1/3-1/(3k)}}{T^{1/(3k)}}.$$

Пусть B_1, B_2 — пропускные способности коммуникационных сетей старой и новой ЭВМ соответственно; L_1, L_2 — полное число одновременно действующих каналов связи коммуникационных сетей старой и новой ЭВМ соответственно. Полагаем, что полная пропускная способность коммутационной сети равна произведению пропускной способности одного канала связи на полное число одновременно действующих каналов, т. е.

$$B_j = b_j L_j, \qquad j = 1, 2.$$

Поэтому $B = B_2/B_1 = (b_2/b_1)(L_2/L_1) = b(L_2/L_1)$ — возрастание пропускной способности коммутационной сети в целом.

Для трехмерной задачи газовой динамики с геометрической декомпозицией в виде трехмерной решетки одновременно на одном шаге по времени задействованы каналы связи, соответствующие трехмерной решетке[8]. Количество таких каналов равно (с точностью до множителя, одинакового для старой и новой ЭВМ)

$$L_j = n_j, \qquad j = 1, 2.$$

Поэтому

$$B = \frac{B_2}{B_1} = b\left(\frac{L_2}{L_1}\right) = b\left(\frac{n_2}{n_1}\right) = bn = \frac{p^{1-1/(3k)} n^{4/3-1/(3k)}}{T^{1/(3k)}}.$$

Пусть V — увеличение памяти всех уровней. Расходуемая память прямо пропорциональна полному числу точек сетки в задаче, независимо от увеличения числа шагов по времени. Поэтому

$$V = \frac{N_2}{N_1} = (Tnp)^{1/k} = (TP)^{1/k}.$$

Список литературы

1. *Софронов И. Д.* Оценка параметров вычислительной машины, предназначенной для решения задач механики сплошной среды // Числ. методы мех. спл. среды. 1975. Т. 6, № 3. С. 98—147.

[8] Это не надо понимать как топологию коммутационной сети, эта топология может быть любой, хоть *каждый с каждым*, но реально во время счета шага по времени используются только каналы, обеспечивающие трехмерную решетку или ее усложнение с учетом диагоналей.

2. *Dongarra J.* Architecture-aware algorithms for scalable performance and resilience on heterogeneous architectures // ASCAC Meeting. American Geophysical Union (AGU). Washington. August 24—25, 2010. http: // www.sc.doe.gov / ascr / ASCAC / Meetings/Aug10/Dongarra.pdf.

3. *Geist A.* Paving the roadmap to exascale // SciDAC Review. Special Issue. 2010. P. 52—59. http://www.scidacreview.org/1001/index.htm.

4. *Мастер Н. С.* Intel — путь к эксафлопсу // ПАВТ-2011. Москва, МГУ им. М. В. Ломоносова. Март 2011 г. http://agora.parallel.ru/pavt2011.

5. *Бахрах С. М., Величко С. В., Спиридонов В. Ф. и др.* Методика ЛЭГАК-3D расчета трехмерных нестационарных течений многокомпонентной сплошной среды и принципы ее реализации на многопроцессорных ЭВМ с распределенной памятью // Вопросы атомной науки и техники. Сер. Математическое моделирование физических процессов. 2004. Вып. 4. С. 41—50.

Статья поступила в редакцию 20.02.12.