МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФГБНУ «Научно-исследовательский институт – Республиканский исследовательский научно-консультационный центр экспертизы»

Приоритетное направление развития науки, технологий и техники «Информационно-телекоммуникационные системы»

РАЗВИТИЕ СУПЕРКОМПЬЮТЕРНЫХ ТЕХНОЛОГИЙ – ОСНОВНОЙ КОМПОНЕНТ РАЗВИТИЯ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ

АНАЛИТИЧЕСКИЙ ОБЗОР

Подготовлен при финансовой поддержке Минобрнауки России.
Использованы материалы, предоставленные экспертами Федерального реестра экспертов научно-технической сферы Минобрнауки России

МОСКВА 2015

Введение

Почему суперкомпьютеры и суперкомпьютерные технологии? Почему в настоящее время им придается такое большое значение, которое проявляется в том, что ряд стран (США, Япония, Китай, страны Европы) приняли специальные программы на федеральном уровне, проводят масштабные организационные мероприятия по консолидации усилий различных организаций и выделяют большие деньги под развитие суперкомпьютерных технологий?

На это существует ряд причин. Во-первых, сама область информационных технологий (ИТ) стремительно меняется в последние годы. Эти изменения характеризуются, прежде всего высокой динамикой, ростом получаемых и накапливаемых данных и появлением новых, революционных технологий. Все четыре основных тренда, определяющих современное состояние ИТ: мобильность, Большие данные, облачные вычисления, робототехника и искусственный интеллект не могли бы быть сформированы и получить развитие без прогресса в области разработки высокопроизводительных вычислительных устройств. Во-вторых, как оказалось, не только потребности в ИТ сфере, но и проблемы системного характера, обозначившиеся в последние годы в самой области вычислительных устройств и грозящие нарушить поступательный прогресс во всей сфере ИТ, привлекают внимание к развитию суперкомпьютерных технологий, да так, что этим озабочены на уровне правительств стран.

В данном докладе рассмотрены проблемы, существующие в данной области, проанализированы пути их решения на примере отдельных стран и рассмотрены возможные варианты их решения для нашей страны.

1 Анализ суперкомпьютерной отрасли

Рассмотрим более детально содержательное наполнение обозначенных во введении вызовов.

Представление об объеме и скорости накопления данных в мире дают цифры, приведенные на рисунке 1 [1]. Выделим значение 25 PB/yr ($25x10^{15}$ петабайт/год), которое относится к данным, получаемым в результате выполнения экспериментов в европейском ЦЕРНе (LHC).

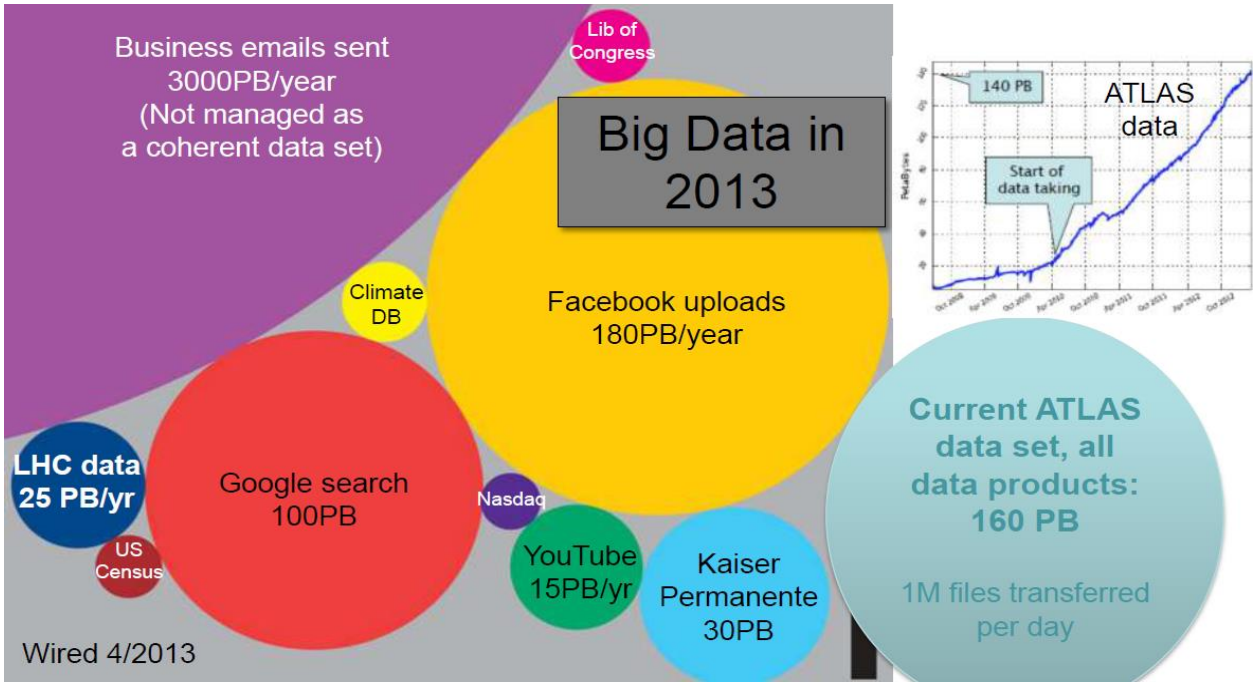

Рисунок 1 – Объем накопленной информации в мире (PB – петабайт, 10^{15} байт)

Как видно из рисунка 1, это далеко не самый большой информационный ресурс. Тем не менее, даже при таком, относительно «небольшом» объеме данных, ЦЕРН вынужден задействовать грид-систему с вовлечением в обработку данных вычислительные устройства, включая суперкомпьютеры, расположенные в различных странах по всему миру (рисунки 2 и 3).

На рисунке 2 приведены цифры, поясняющие происхождение столь большого объема данных [1]. В процессе работы коллайдера непрерывно происходит ~109 взаимодействий «протон-протон» в секунду с образованием 1600 частиц в каждом соударении. Это приводит к тому, что скорость накопления «сырой» информации происходит лавинообразно и составляет 1

PB/s (Raw data rate from LHC detector). И только благодаря тому, что после предварительной обработки объем исходных данных уменьшается на шесть порядков (с 10^{15} до 6 10^{9} байт), появляется возможность выполнять их обработку в приемлемые сроки. В ином случае, это занимало бы годы. На рисунке 3 показана сеть компьютеров, вовлеченных в вычисления, которая охватывает практически все страны мира.

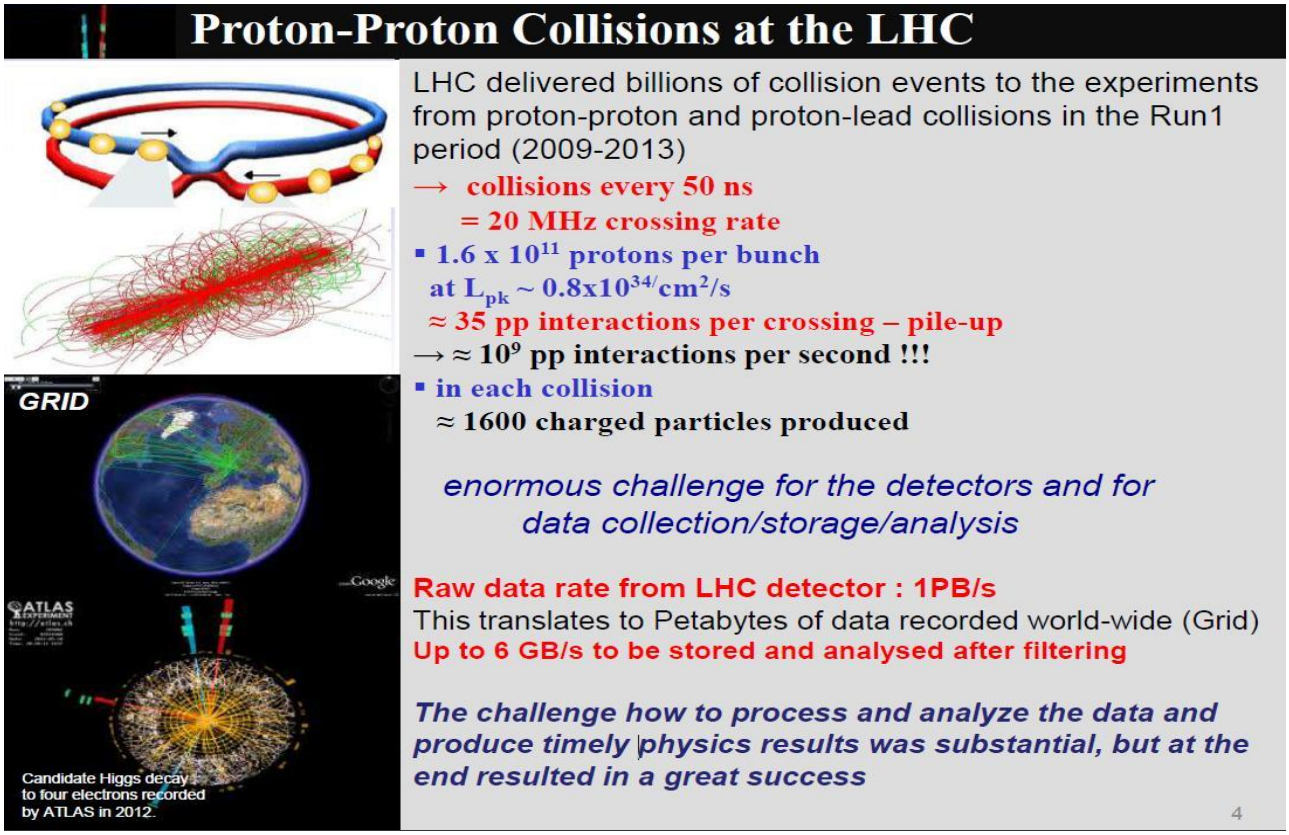

Рисунок 2 – Процесс формирования и трансформации данных в ЦЕРНе [1]

Многие изменения в области ИТ связывают с наступлением эпохи «Интернета вещей» или «Интернета всего», когда огромное количество материальных, но не живых объектов, число которых прогнозируется на уровне 50 млрд к 2020 году, будут наделены возможностью обмениваться информацией между собой. При этом весь поток данных будет необходимо обрабатывать в реальном времени. Предполагается, что это потребует массового применения высокопроизводительных вычислительных устройств (рисунок 3) [1].

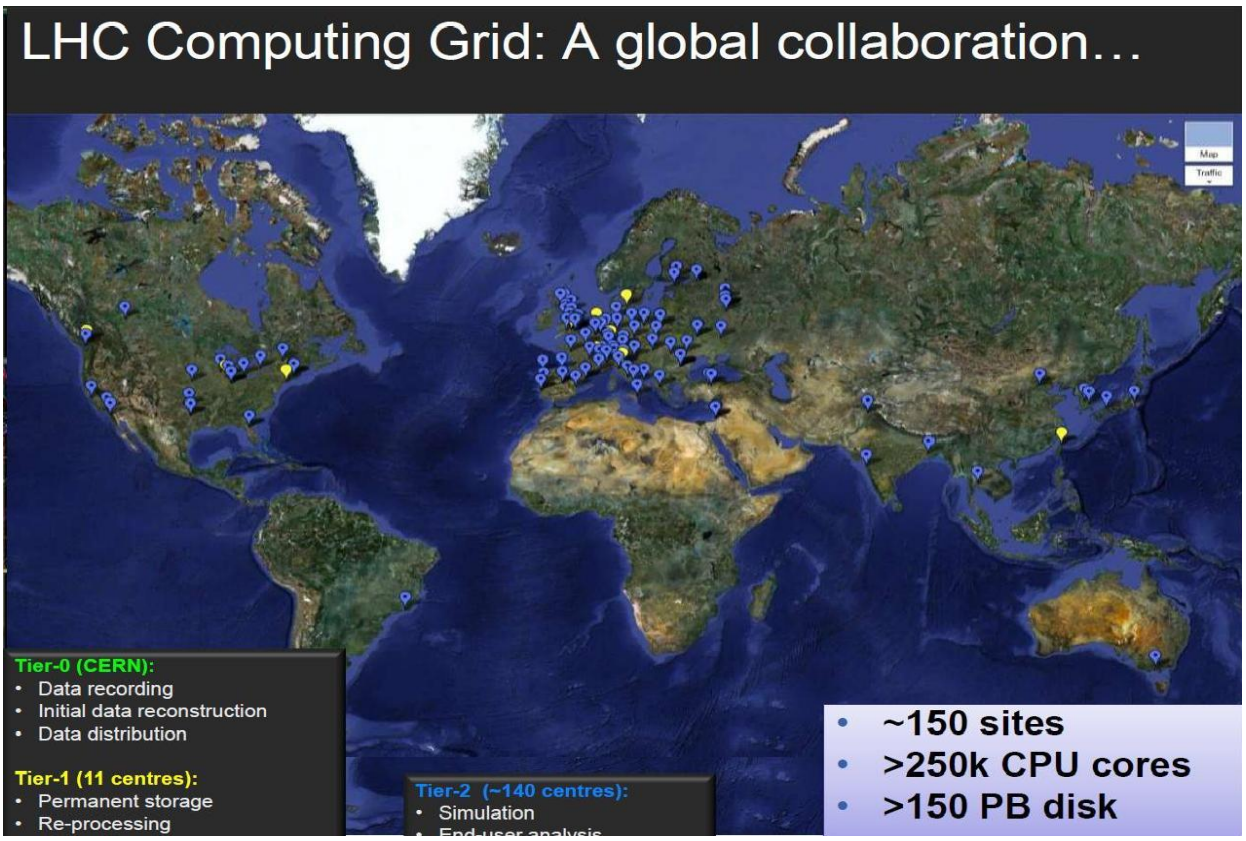

Рисунок 3 – Грид-система ЦЕРНа, включающая более 250 000 вычислительных узлов и 150 петабайтный объем хранения данных

Помимо общих трендов, отдельные специальные задачи (ядерное моделирование, моделирование климата, моделирование биологических систем, разведка, др.) сегодня уже требуют экзафлопсной производительности (10^{18} флопс, то есть, 10^{18} операций с числом с плавающей запятой, выполняемое компьютером за одну секунду). Ставятся принципиально новые задачи (например, сбор и обработка разнородной информации, распространяемой в социальных сетях, то есть, в глобальном масштабе, моделирование потока материальных частиц при конструировании оптимальных форм транспортных устройств и т. д.), решение которых связывают с достигнутыми, а то и ожидаемыми пиковыми возможностями производительности компьютеров.

Эти и другие факторы и определяют ключевую потребность в области ИТ – потребность в высокопроизводительных вычислительных устройствах, поскольку без них уже невозможно ставить и решать определенные задачи,

как и невозможно принципиально говорить о прогрессе в этой области. Здесь исследуются и отрабатываются самые передовые технологии, как это происходит, например, в Формуле 1 в автомобильной промышленности. Разница лишь в том, что скорость изменений в ИТ сфере настолько высока, что требуется опережающий темп роста скорости вычислений. Так, что суперкомпьютерные технологии здесь определяют облик ИТ уже не завтрашнего, а фактически и сегодняшнего дня.

Что же происходит в области суперкомпьютеров (СК) или суперкомпьютерных технологиях (СКТ) и почему такое внимание к существующим там проблемам?

Наиболее интегральной оценкой, характеризующей достижения в области СКТ, является величина пиковой производительности, демонстрируемой на специальных тестах, и которая имеет сегодня величину десятков петафлопс. В планах достичь к 2018 - 2020 годам экзафлопсного уровня (10^{18} флопс). Эта задача, ввиду ее значимости, получила специальное название «ExaScale» - «ЭкзаМасштаб». И на этапе «штурма» экзафлопсного уровня начинают проявляться физические ограничения КМОП – технологий. То есть, эволюционный путь развития компьютерной техники, базирующийся на данной технологии, начинает подходить к своему пределу. Данная технология, как прогнозируется, должна исчерпать свой запас к 2020 годам. Чтобы идти дальше, нужны инновационные решения.

Дадим краткое пояснение начинающим проявляться ограничениям КМОП-технологий, чтобы осознать глубину стоящих в области суперкомпьютерных технологий проблем.

1. Наиболее принципиальное ограничение связано с существующей зависимостью вычислений и выделяемого тепла, т. е. связи между информацией и термодинамикой (принцип Неймана-Лэндауэра), известное как «ограничение Лэндауэра» [2, 3, 4].

Ограничение Лэндауэра утверждает, что при необратимом процессе обработки информации, когда возможна потеря битов информации, затраты

на обработку одного бита (или выделение тепла при потере одного бита) не могут быть меньше величины

$$kT \ln 2 = 2.85 \times 10^{-21} = 2.85\ zJ\ (зептоджоулей),$$

где: k – константа Больцмана, а T – температура по Кельвину, при расчете берется комнатная температура в 25^0C или 298.15 K.

На рисунке 4 приведен график, построенный Лэндауэром [3], показывающий динамику снижения затрат на обработку одного бита в зависимости от улучшения технологий, привязанного к временной оси. Уровень kT (пунктирная линия) соответствует ограничению Лэндауэра.

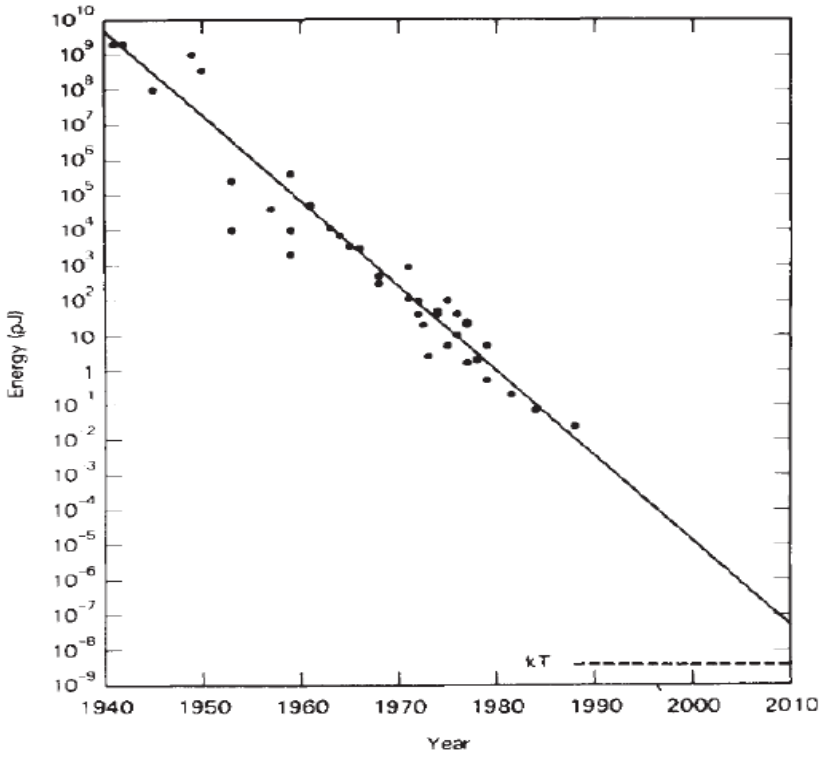

Рисунок 4 – Динамика снижения затрат на обработку одного бита

Сегодня современным технологиям 22 нм и современным компьютерам, которые являются принципиально необратимыми (нереверсивными), соответствует уровень энергозатрат на один бит информации составляет не менее 100000 - 1000000 kT, что в пересчете на энергозатраты всей вычислительной системы, при ее требуемой производительности на уровне экзафлопс, даст потребление энергии на уровне сотен и даже тысяч мегават. Это является неприемлемым.

Справка. Под необратимой (или нереверсивной) логикой работы компьютера понимается то, что тепловые процессы, сопровождающие вычисления в таком компьютере, протекают только в одном направлении. В отличии от этого, ***обратимыми (реверсивными)*** *являются компьютеры, над созданием которых в настоящее время ведутся активные исследования, процессы перехода системы из одного равновесного состояния в другое возможно провести в обратном направлении (через ту же последовательность промежуточных равновесных состояний), то есть, когда сама система и окружающие тела возвращаются к исходному состоянию.*

То есть, традиционные технологии, в силу невозможности с их помощью приближения к уровню энергозатрат на обработку одного бита порядка kT Ln2, не способны принципиально преодолевать экзафлопсный уровень производительности по энергетическим соображениям. Нужны новые, инновационные решения.

2. Из соотношений неопределенности Гейзенберга следует, что при достижении по энергозатратам на обработку одного бита информации значения kT Ln2 («ограничение Лэндауэра»), минимальный размер стороны объема вещества (независимо от используемых технологий), который может быть установлен в устойчивое состояние 0 или 1, независимо от теплового шума электронов в этом объеме, равен 1.5 нм [5]. При этом достижение меньшего размера, даже если технологии это будут позволять, будет бессмысленным, так как на этом уровне начнут сказываться квантовые эффекты, т. е. процесс будет неконтролируемым.

Это дает примерную оценку предельному значению технологической нормы в 4 … 5 нм, переход на который ожидается в районе 2020 года, что определяет технологический предел в процессе миниатюризации при производстве микросхем. То есть, тот прогресс, который всегда связывался с переходом на все меньшие технологические нормы, будет «исчерпан» после достижения величин в 4…5 нм. Заметим, что компания Intel уже работает на уровне 14 нм, другие брэнды в производстве микроэлектронных плат – на уровне 22…28 нм.

3. Проблемы экономической эффективности. Достигнутая плотность размещения транзисторов на плате приближается к своему пределу и логически ожидаемое снижение стоимости чипа в пересчете на стоимость одного транзистора (Cost per wafer reduction) при переходе на технологическую норму (Technology node) 22/20 нм приобретает отрицательную динамику (рисунок 5) [3].

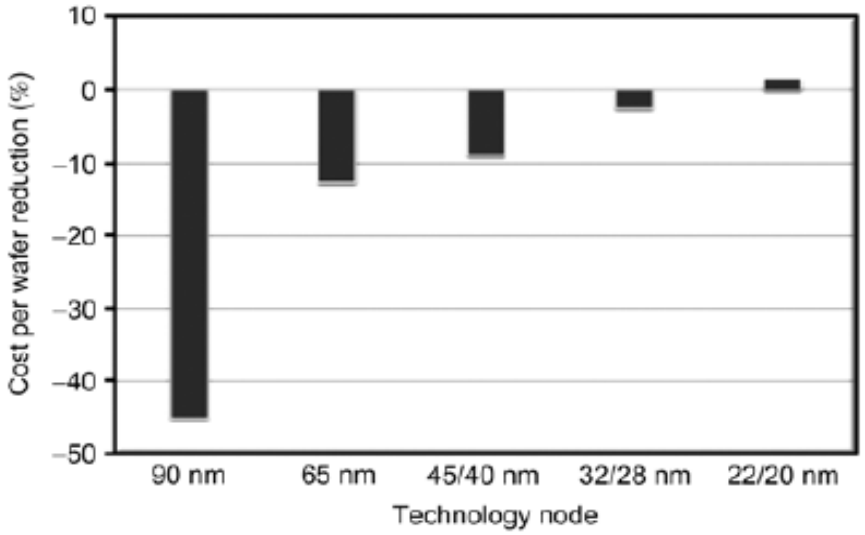

Рисунок 5 – Негативная динамика уменьшения стоимости одного транзистора с улучшением технологического процесса, начинающая проявляться с технологической нормы 22/20 нм

То есть, переход на меньшие технологические нормы в общем экономически выгоден, оценивая его по совокупности всех параметров, но уже не так очевиден, как это происходило ранее. Можно говорить о замедлении или достижении некоторого порога в получении экономического эффекта от дальнейшего процесса микроминиатюризации.

4. С середины первого десятилетия текущего века перестало выполняться следствие теории Деннара (появившейся в 70-е годы прошлого века и успешно подтверждавшееся все предыдущие десятилетия). В соответствии с теорией масштабируемое уменьшение размеров транзистора при небольшом уменьшении прикладываемого напряжения не только значительно увеличивает плотность транзисторов на кристалле, но и производительность (частоту, на которой он работает). То же касается

известного закона Мура, который, в смысле удвоения количества транзисторов на кристалле каждые 18 месяцев, продолжает выполняться. Однако с 2002 - 2003 гг. нет роста таких макроуровневых характеристик процессорного ядра, как тактовая частота микропроцессора (Clock Speed, MHz), потребляемая мощность (Power, W), количество команд, выдаваемых на выполнение за один такт (Perf/Clock, ILP), что демонстрируется на рисунке 6 [3]. Поэтому нет заметного роста и производительности процессорного ядра.

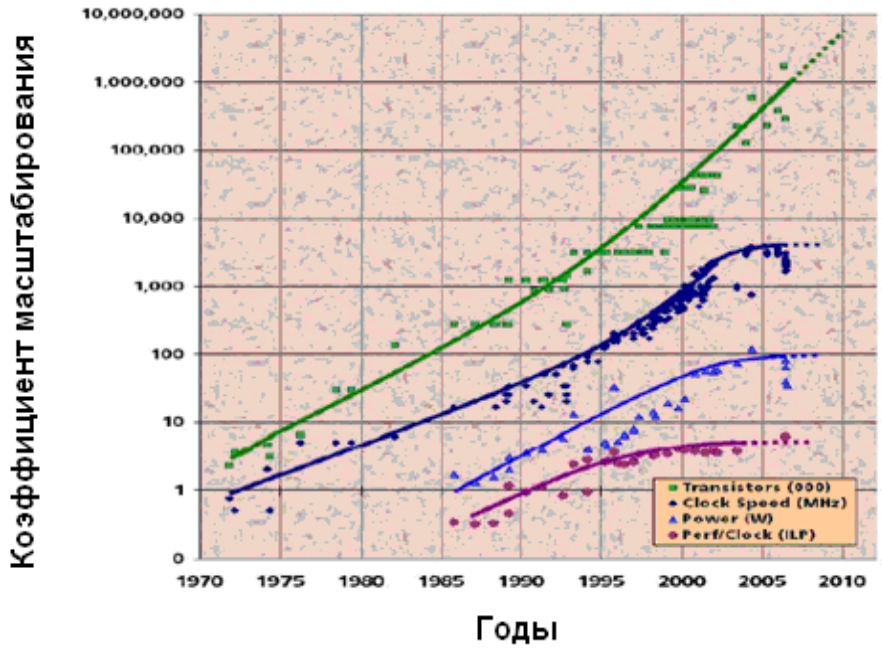

Рисунок 6 – Выполнение закона Мура и ограничение его влияния на производительность

5. Проблема энергоэффективности. Известна оценка энергопотребления на одно обращение к памяти, которое считается наиболее затратным из всех операций, выполняемых в процессе вычислений (примеры других операций – операция непосредственного вычисления, обращение к КЭШ-памяти, выборка команд, дешифрация, распределение синхросигналов и т.д.). На сегодня, это величина в размере 16 000 пДж (на одно обращение) или 62,5 пДж/бит (на обработку одного бита информации, рисунок 7) [6]. Это означает, что при производительности в 10^{18} флопс, которой стремятся достичь, при обращении к памяти один раз на 10 операций и выполнении

реальных счетных задач энергозатраты на вычисления на устройствах, собранных по традиционным КМОП-технологиям, составят ~ 4 ГВт. То есть, такое вычислительное устройство будет потреблять энергию, вырабатываемую целой электростанцией (для представления значения полученной величины, можно сопоставить ее с мощностью Братской ГЭС, которая составляет 4,52 ГВт). Очевидно, что такие энергозатраты неприемлемы и их следует уменьшать, как минимум на два порядка.

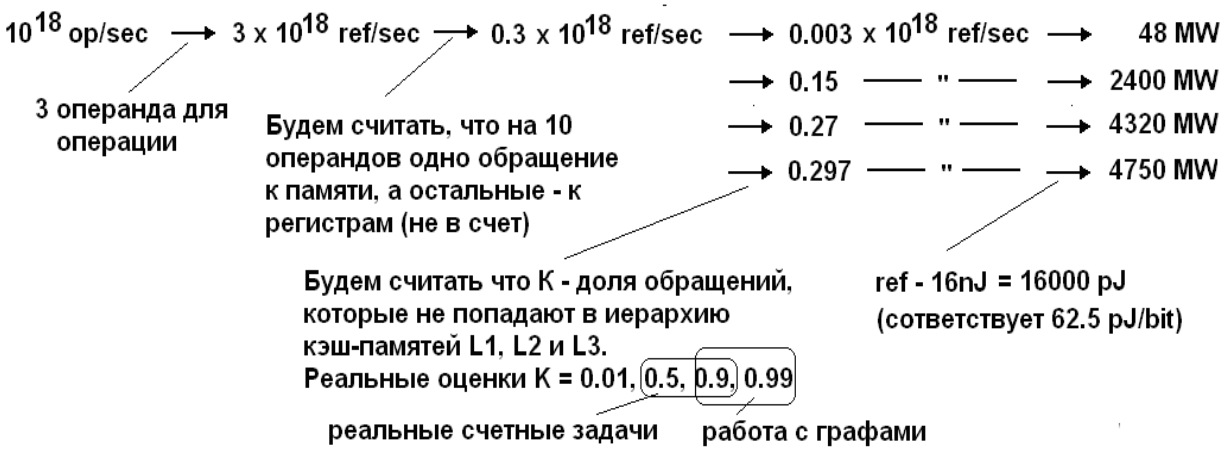

Рисунок 7 – Оценка энергозатрат на вычисления при использовании традиционных КМОП-технологий и традиционных архитектурных решений

Решение этой задачи является далеко непростым. На рисунке 8 [7] показаны прогнозные величины энергозатрат, которые возможно достичь за счет оптимизационных решений, оставаясь принципиально на традиционных технологических решениях.

Здесь: Conventional Design - обычный подход к архитектуре, что дает среднюю величину энергопотребления на одну операцию в 2500 пДж, Minimize Overhead - при оптимизированной микроархитектуре, Minimize DRAM eng. - при оптимизированной микроархитектуре и памяти.

Но, традиционные решения позволяют снизить энергозатраты только на порядок, до сотен мегаватт, в то время как необходимо их уменьшение до десятков мегаватт. Широко цитируется цифра в 20 МВт, обозначенная в качестве ориентира и которую планируется не превысить при построении экзафлопсного компьютера.

Варианты реализации	Направления энергозатрат (pJ/FLOP)					
	FPU	Local	Global	Off-chip	Overhead	DRAM
Conventional Design	39.0	26.0	72.0	65.0	1818.0	480.0
	Всего: 2500 pJ/FLOP					
Minimize Overhead	15.0	16.0	45.0	65.0	10.0	480.0
	Всего: 631 pJ/FLOP					
Minimize DRAM eng.	5.1	4.0	2.3	2.6	10.0	3.8
	Всего: 28 pJ/FLOP					

FPU – выполнение собственно операции в функциональном устройстве;
Local – чтение операндов из регистров и L1D кэш-памяти ядра;
Global – работа с глобальными кэш-памятями микропроцессора уровня L2|L3;
Off-chip – внекристальные взаимодействия, в основном, интерфейс с памятью;
Overhead – накладные расходы, связанные с подкачкой команд в кэш команд, выборкой команд, дешифрацией, управлением их выполнения, распределением синхросигналов, потерями из-за токов утечки (+ невычислительные операции - ?);
DRAM - выполнение операций непосредственно с модулями DRAM-памяти.

Рисунок 8 – Прогноз энергопотребления в вычислительном узле для технологии 22 нм

Существуют и иные проблемы, в том числе, большое количество затрачиваемых операций на обращение к памяти («стена памяти»), проблема отвода тепла от микропроцессорных ядер, сложность архитектурных решений при распараллеливании работы процессоров и связанное с этим существенное усложнение процесса программирования и т. д.

Перечисленные проблемы, связанные, как принято говорить, с эволюционным путем развития вычислительных устройств, определяют потенциал данного направления, который, если оценивать его одним параметром - производительностью вычислений, оценивается в пределах 30…130 экзафлопс, что не мало, но это величина пиковой производительности, а это значит, что, как показывает практика, на реальных задачах производительность будет на 2 порядка меньше, этого в ближайшем будущем уже недостаточно. Очевидно, что этот путь развития компьютерной отрасли будет вырабатываться до конца, в предположении, что это позволит удовлетворять потребности ИТ сферы на ближайшие годы, вплоть до 2020 года. Однако, приходит ясное осознание того, что дальнейший прогресс связан исключительно с инновационным путем развития данной сферы.

2 Анализ программ инновационного развития СКТ

Понимание важности роли инновационных технологий для решения проблем в суперкомпьютерной области на уровне «проблемы государственного масштаба» в мире впервые было зафиксировано американским Сенатом в Законе “О возрождении суперкомпьютеров в США” в конце ноября 2004. Хотя практически работы в этом направлении на федеральном уровне начались еще с 2002 года с запуска программы DARPA HPCS (где HPCS – High Productivity Computing Systems, - «высокопроизводительные компьютерные системы»). Аналогичные программы были практически немедленно запущены в Китае и Японии. Заметим, что в России, к сожалению, в это время не было сформировано аналогичной программы, принимаемые программы были нацелены на развитие лишь эволюционных технологий и то, в узком кругу ведомств и организаций.

После успешного завершения программы DARPA HPCS в 2010 году, в США в 2013 году была запущена следующая программа - DARPA STARnet (где STARnet – Semiconductor Technology Advanced Research Network, - «исследования в области перспективных полупроводниковых технологий»), которая отличается высоким уровнем мобилизации исследовательских ресурсов страны и их беспрецедентной координацией.

Ниже приведены более подробные сведения по этой программе для понимания того, насколько серьезно отношение к данному вопросу в США. По примеру США аналогичные программы принимались в Китае, Японии и странах Европы, но до сих пор нет в России, где исследования ведутся относительно разрозненно и без координации на федеральном уровне.

В рамках проекта была создана сеть из шести мощных исследовательских центров в области полупроводников, которые были открыты при следующих университетах:

– University of Illinois (at Urbana-Champaign);

– University of Michigan;

– University of Minnesota;

– University of Notre Dame;

– University of California at Los Angeles;

– University of California at Berkeley.

В целом, в проекте задействовано 39 университетов, порядка 150 специалистов и 400 аспирантов. Кроме DARPA, SRC и университетов, в проекте также участвуют Исследовательская лаборатория ВВС США (U.S. Air Force Research Laboratory, AFRL) и Ассоциация полупроводниковой промышленности (Semiconductor Industry Association, SIA), а также шесть партнеров из промышленности: Applied Materials, GLOBALFOUNDRIES, IBM, Intel Corporation, Micron Technology, Raytheon, Texas Instruments and United Technologies.

Такое внимание со стороны промышленности логично, так как осознание ограниченности перспектив современных КМОП-технологий хотя и может позволить «продержаться» 144 миллиардному рынку микроэлектроники США до конца десятилетия, но физические ограничения неумолимы, поэтому нужны долговременные фундаментальные исследования, соответственно, без их поддержки этот рынок можно «потерять».

Был организован ряд центров с определенной функциональной специализацией.

Центр исследований по будущим архитектурам (The Center for Future Architectures Research (C-FAR)). Образован при Мичиганском университете (University of Michigan) и ориентирован на разработку архитектур периода 2020-2030 гг. Это направление работ основано на предположении, что специализированные на конкретные приложения архитектуры усилят возможности промышленных технологий и таким образом “продлят жизнь” КМОП-технологий.

Центр исследований по материалам, интерфейсам и новым архитектурам спинтроники (Spintronic Materials, Interfaces and Novel Architec-

tures (C-SPIN)). Образован при университете Миннесоты (University of Minnesota) и ориентирован на рассмотрение технологий памяти и вычислений на базе использования спинов электронов с целью выяснения их потенциала преодолеть проблемы современных КМОП-технологий.

Центр функционально ускоренного проектирования наноматериалов (The Center for Function Accelerated nano Material Engineering (FAME)). Образован при Калифорнийском университете Лос-Анжелеса (University of California, Los Angeles) и ориентирован на изучение необычных материалов, включая наноструктуры со свойствами квантового уровня. Исследования нацелены на поддержку создания аналоговых логических элементов и памяти для реализации вычислений "за пределами двоичных вычислений".

Центр технологии систем с низким потреблением энергии (The Center for Low Energy Systems Technology (LEAST)). Образован при университете Нотр Дамм (University of Notre Dame) и ориентирован на выполнение исследований и разработок в области материалов и устройств с чрезвычайно низким потреблением энергии.

Центр систем на базе информационных технологий наноуровня (The Center for Systems on Nanoscale Information Fabrics (SONIC)). Образован при Иллинойском университете (University of Illinois at Urbana-Champaign) и ориентирован на исследования преимуществ перехода от детерминированных моделей вычислений к статистическим моделям.

Центр исследований по системам в виде групп агентов терра-уровневого масштаба (Terra Swarm Research Center (TerraSwarm)). Образован при Калифорнийском университете в Беркли (University of California, Berkeley) и ориентируется на исследования в области приложений распределенных самоорганизующихся систем масштаба города.

Помимо данной программы, известно о нацеленности создания в совместном проекте Японии и Америки заказных специализированных суперкомпьютеров зетта-уровня (~ 20^{21} флопс) и йотта-уровня (~ 20^{24} флопс) на технологиях квантовых клеточных автоматов и реверсивной логике. Эти су-

перкомпьютеры предназначаются для реализации систем воздушно-космической обороны тихоокеанского региона и западного побережья США. В Китае к 2020 г. планируется создание суперкомпьютера зета-уровня также на квантовых клеточных автоматах, k-значной и реверсивной логике для решения задач криптоанализа. Головная организация проекта – NUDT, в которой велся проект «Удар грома» и который контролируется военной разведкой Китая.

Краткое резюме из вышесказанного.

В развитых странах мира суперкомпьютерной тематике придается беспрецедентное внимание, основанное на осознании:

– запросов сегодняшнего, но что более важно, завтрашнего дня, а точнее, следующего десятилетия;

– ограниченности существующих технологий;

– принципиальной важности поиска новых, инновационных решений, которые должны обеспечить конкурентные преимущества и укрепить безопасность стран в будущем десятилетии.

Это внимание выражено, прежде всего, в масштабности организационных мероприятий, глубине подхода к проблеме и скоординированности усилий научного сообщества и коммерческих структур, инициированных на государственном уровне.

3 Ситуация в суперкомпьютерной отрасли в России

В открытых программах развития отрасли СКТ следует выделить два направления:

– целевые программы Союзного Государства (координатор этих программ - Минобрнауки России);

– общие ФЦП развития высокотехнологичных отраслей, где часть мероприятий посвящены тематике развития СКТ, но подбор даже этих мероприятий выглядит достаточно хаотично, а координаторами этих ФЦП выступают самые разные Министерства.

Целевые программы Союзного Государства (все завершенные):

– Программа «СКИФ» (2000-2004). «Разработка и освоение в серийном производстве семейства моделей высокопроизводительных вычислительных систем с параллельной архитектурой (суперкомпьютеров) и создания прикладных программно-аппаратных комплексов на их основе»;

– Программа «ТРИАДА» (2005-2008). «Развитие и внедрение в государствах-участниках Союзного государства наукоемких компьютерных технологий на базе мультипроцессорных вычислительных систем»;

– Программа «СКИФ-ГРИД» (2007-2010). «Разработка и использование программно-аппаратных средств ГРИД-технологий и перспективных высокопроизводительных (суперкомпьютерных) вычислительных систем семейства «СКИФ».

ФЦП:

– ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2007-2011 г.». Координатор – Минобрнауки России. Государственные заказчики – Минобрнауки России, МГУ им. М.В. Ломоносова. Общий объем финансирования - 21.6 млрд руб., в том числе на НИОКР – около 19.7 млрд руб.;

– ФЦП «Исследования и разработки по приоритетным направлениям развития научно-технологического комплекса России на 2014-2020 г.». Координатор – Минобрнауки России. Государственные заказчики – Минобрнауки России, МГУ им. М.В. Ломоносова. Общий объем финансирования – 239.023 млрд. рублей, в том числе на НИОКР – около 131.2 млрд руб.;

– ФЦП «Развитие электронной компонентной базы и радиоэлектроники на 2008 – 2015 г.». Координатор – Минпромторг России. Государственные заказчики: Минпромторг России, Минобрнауки России, Роскосмос, ГК "Росатом". Общий объем финансирования – 13.0 млрд руб., в том числе на НИОКР – 9.63 млрд руб.;

– ФЦП «Национальная технологическая база на 2007 – 2011 г.». Координатор – Минпромторг России. Государственные заказчики: Минпромторг

России, Минобрнауки России, Роскосмос, Российская академия наук, Сибирское отделение Российской академии наук, ГК "Росатом". Общий объем финансирования – 9.55 млрд руб., в том числе на НИОКР – 7.19 млрд руб.;

– Перечень проектов по направлению «Развитие суперкомпьютеров и грид-технологий», утвержден Комиссией при Президенте РФ по модернизации и технологическому развитию экономики России. Государственный заказчик – ГК «Росатом», головной исполнитель – ФГУП «РФЯЦ-ВНИИЭФ». Стоимость работ по проекту в 2010 году составила 3.395 млрд. рублей из федерального бюджета и 908 млн. рублей из внебюджетных источников соответственно;

– ФЦП «Информационное общество, на 2012-2020 г.». Координатор – Минкомсвязь России. Общий объем финансирования – 88.0 млрд, руб. в том числе: 2011 г. – 3.1 млрд руб.; 2012 г. – 3.1 млрд руб.; 2013 г. – 3.1 млрд руб.; 2014 – 2020 г. – 78.7 млрд руб.

Анализ перечисленных программ показывает, что области суперкомпьютерных технологий в каждой ФЦП уделено внимание, поставлены цели и задачи, намечены плановые показатели, которых следовало/следует достичь. Но, это внимание выглядит фрагментарным и поверхностным, предполагающим более детальное их развитие в специальных документах, распределяющих задачи между ведомствами. Эти ведомства, однако, зачастую преследуют свои интересы, не обязательно совпадающие со стратегическими целями.

Лишь в мае 2011 г. была впервые предложена относительно целостная концепция по экзафлопсным технологиям «Развитие технологии высокопроизводительных вычислений на базе суперЭВМ эксафлопсного класса (2012-2020 гг.)» – ГК «Росатом». Общий объем запрашиваемого финансирования – около 49 млрд рублей, но оно до сих пор не открыто, хотя ведущее предприятие Росатома (Саровский центр (ВНИИЭФ)) получило часть финансирования на создание нового суперкомпьютера с пиковой производительностью в несколько петафлопс. Насколько это будет способствовать общей стратегии развития отрасли СКТ в России, покажет время.

Что в итоге мы видим по достигнутым результатам в России и ведущих зарубежных странах?

Результаты реализованных зарубежных программ в настоящее время нашли практическое воплощение:

– в виде коммерчески доступных на специальных внутренних рынках суперкомпьютеров нового типа (США – суперкомпьютеры IBM Power 775 и Cray XC30, Китай – суперкомпьютеры Tianhe-1 и Tianhe-2, Япония – K - компьютер);

– в виде специальных массово-мультитредовых суперкомпьютеров с глобально адресуемой памятью для военных и разведслужб (в США, Японии и Китае), они уже находятся в опытной эксплуатации, на боевом дежурстве ожидаются после 2017 года.

Достижение таких результатов обеспечило лидирующее положение перечисленных стран, как в научно-технической, так и военной областях, особенно в приложениях, связанных с обработкой больших данных (BigData). Приложения класса BigData важны, например, для решения разведывательных задач и задач воздушно-космической обороны, создания национальных центров управления войсками, решения социально-экономических задач, а также задач класса ведения кибервойн, включая манипулирование сознанием целых народов.

Количественные оценки сложившегося соотношения качества отечественных и зарубежных суперкомпьютеров на основе информации из мировых рейтингов:

– в рейтинге Top500 (задачи так называемого CF-класса, для которых характерны хорошая пространственная и временная локализация обращений к памяти, это достаточно редкий случай для современных практических задач) – отставание не менее 10 раз;

– в рейтинге Graph500 и HPC Challenge (задачи DIS-класса, т. е. задачи с плохой пространственной и/или временной локализацией обращений к памяти, подавляющий в настоящее время класс задач) – отставание в 100-1000

раз. Если сравнивать со специальными массово-мультитредовыми суперкомпьютерами США, Японии и Китая, то этот показатель увеличивается еще не менее, чем на два-три порядка, то есть, до 10^4 – 10^6 раз. Если при этом учитывать и ставший эффективно доступным ресурс глобально адресуемой оперативной памяти в десятки петабайт, то по объему такой памяти отставание в 10^4 раз.

Обратим внимание, что приведенные показатели относятся к отечественным суперкомпьютерам, построенным с применением зарубежных эволюционных СКТ. Если из-за санкций придется перейти на отечественную элементно-компонентную базу, то по результатам уже проведенных экспериментов отставание увеличится еще на 1-2 порядка, а по энергетике – не менее, чем на порядок.

Почему так произошло?

Одна из основных причин, но не единственная, - это системный подход в постановке работ в ведущих зарубежных странах, особенно в США. Изначально в постановке работ участвуют значительные группы квалифицированных экспертов, как раз формирующих системное видение проблемы. Это определяет, в дальнейшем, затрагиваемые области науки, соответственно, коллективы и организации, работающие в данных областях, после чего на государственном уровне в рамках общей задачи для них определяются подзадачи, финансирование и осуществляется координация работ.

При важности всех этапов работы над проектом, обратим внимание на начальную стадию процесса – экспертную работу. Например, в США постановка целей и задач, выбор основных вариантов решений задач по созданию перспективных суперкомпьютеров петафлопсной производительности был в прошлом десятилетии выполнен сначала специальными группами экспертов DARPA и NSA (Агентство Национальной Безопасности) с привлечением специалистов из академической среды для формирования ведомственных программ. Затем, по инициативе аппарата Президента США, на федеральном уровне, для формирования уже федеральной программы по восстановлению

в США отрасли СКТ была составлена экспертная группа HPC Task Force (оперативная группа экспертов в области СКТ) из компетентных, непосредственно участвующих в исследованиях и разработках специалистов.

Эта группа появилась в начале прошлого десятилетия. Достойный подражания образец выработанных ею решений по возрождению отрасли СКТ в США и созданию петафлопсных суперкомпьютеров – это работа Workshop on “The Roadmap for Revitalization of High-End Conputing” [8]. Примеры результатов более поздних работ подобных групп, но уже по тематике экзафлопсных систем, – работы *Dongarra J. et al DARPA’s HPCS Program: History, Models, Tools, Languages* и *Amarasinghe S. et al. ExaScale Software Study: Software Challenges in Extreme Scale Systems*. [9, 10].

Эксперты и даже целые рабочие группы экспертов есть и в России. В связи с этим появляется ряд вопросов, а именно:

– насколько зависит высокая эффективность зарубежных суперкомпьютерных проектов от работы экспертов?

– чем отличается работа упомянутых зарубежных экспертных групп и отечественных экспертов?

– не являются ли недостатки работы на экспертном уровне и отношения к результатам этой работы одной из причин сложившегося серьезного отставания в области СКТ, так как средства у нас вкладываются в данную отрасль достаточно большие, сопоставимые с зарубежными?

Прежде всего, ответим на вопрос, насколько вообще важна экспертная работа.

Мировая практика показывает, что исправление заложенных на начальной стадии проекта ошибок обходятся, на стадии НИР, на один порядок дороже потраченных средств, на стадии ОКР – на 2-3 порядка, на стадии серийного производства – потерей стратегии и конкуренции. Эти цифры, усредненные по результатам накопившегося «печального» опыта, наглядно демонстрируют важность экспертной работы, особенно, на начальных этапах работ, начиная просто с формирования объективной информационной среды

для выявления тенденций и проблем, и продолжая выбором тематики работ, определения задач и вариантов их решения. Именно поэтому, для привлеченных зарубежных экспертов, участвовавших в постановке упоминавшихся и ставших знаковыми проектами, было выделено более полутора лет только на то, чтобы они могли взвешенно и обоснованно определить цели и задачи в соответствии с существующими вызовами. Результаты такого подхода не могли не сказаться [11].

Что у нас?

Создается впечатление, что наличие экспертного сообщества научно-технической сферы только формально обеспечивает выполнение приведенных выше экспертных функций. На уровне принятия решений, по всей видимости, экспертное сообщество работает в пассивном режиме, то есть, когда необходима оценка уже подготовленных проектов или результатов их реализации. Попытка активизации экспертного сообщества была предпринята в рамках Федерального реестра экспертов научно-технической сферы Минобрнауки России. Из экспертов в области ИТ, была выделена группа экспертов в области высокопроизводительных вычислительных устройств. Группа экспертов включает более 40 ученых и специалистов в области СКТ, не учитывая ученых в достаточно близких смежных областях (например, элементная база и др.). Экспертами Федерального реестра за последние 3 года было разработано более 100 предложений и 30 аналитических докладов по перспективным направлениям и тематике исследований и разработок в области суперкомпьютерных технологий. Эта работа проведена, что особенно важно, в рамках независимого сообщества экспертов.

В ведомствах экспертная деятельность, по всей видимости, так же ведется, но играет, далеко не приоритетную роль в выборе тематики исследований и разработок. По сложившейся практике, решения принимаются в узком кругу чиновников и ученых высокого ранга, директоров предприятий, где значительное влияние на принимаемые темы проектов оказывает экономическая заинтересованность и “благоразумие” потенциальных исполните-

лей, не желающих браться за хотя и рисковые, но реально необходимые темы. Плюс ко всему, сам механизм учета предложений (экспертов федерального или ведомственного уровня) не отработан и не гарантирует оптимальности принимаемых решений в рамках формирования ФЦП.

Все это позволяет в целом говорить о пассивном характере экспертной деятельности в нашей стране. Иначе ситуация в области СКТ была бы иной. Главный вывод, который напрашивается по результатам анализа состояния СКТ в России, состоит в том, что у нас нет системного подхода к постановке задачи в решении проблем в области СКТ. Это было сделано в США, Китае и Японии, а следом в европейских странах, где такие задачи формулировались, прежде всего, на основе активной и независимой экспертно-аналитической работы и прогнозирования, при обеспечении соответствующего уровня ее влияния на принимаемые решения по планируемым в стране исследованиям и разработкам, и организации соответствующих работ.

4 Возможный вариант начала решения проблемы

Очевидно, что подъем суперкомпьютерной отрасли не сводиться только к вопросу создания экспертно-аналитической службы, как органа, «ключевого и снимающего основные проблемы». И очевидно, что это целый комплекс мероприятий: подготовка кадров, исследования по будущим архитектурам и новым материалам, разработка систем с низким потреблением энергии, работы в области разработки новых алгоритмов вычислений, и т. д. и т. п. Но активная экспертная работа, в том контексте, который был определен выше, способна придать системный характер всем тем работам, которые велись и ведутся в России, заметим, при достаточно большом финансировании, но которые не приводят к результатам стратегического характера, что уже начинает сказываться на отставании страны в различных рейтинговых списках, а в 2020-х годах может привести к отсутствию у нас вообще такой отрасли, как суперкомпьютеры, так как все придется заимствовать.

Для этого, целесообразно принятие инфраструктурных решений на уровне государственного управления, в рамках которых сформировать экспертно-аналитическую службу и наделить ее соответствующим статусом, которая, будучи составной частью межведомственной рабочей группы по суперкомпьютерам, вела бы упреждающую прогнозную работу, принимала участие в координации проводимых работ и, самое главное, ее рекомендации учитывались.

Собственно говоря, так организована экспертная информационно-аналитическая служба в Китае (головная группа при NUDT, университете оборонных технологий Китая), Японии (головная группа при генеральном штабе сил самообороны) и США в рамках программы NITRD (главная федеральная программа по информационным технологиям), где есть группа - Task Force, которая была создана в начале 2000-х годов в процессе возрождения отрасли СКТ в этой стране. Она действует и сейчас, как один из проектов NITRD. Основные направлениями изучения экспертной группы TaskForce были с учетом современной терминологии следующими:

– перспективные суперкомпьютерные технологии в области элементной базы, системотехники, конструктивов, моделей вычислений и программного обеспечения (инновационные СКТ);

– системы на базе коммерчески доступных компонентов (эволюционные СКТ);

– заказные суперкомпьютеры высшего диапазона производительности (по нашей терминологии – суперкомпьютеры стратегического назначения (СКСН)), они особенны тем, что во все времена должны не менее, чем на один-два порядка превосходить по реальной производительности суперкомпьютеры, создаваемые кем-либо по доступным на рынке технологиям (это суперкомпьютеры на базе инновационных СКТ);

– операционные системы, системы обеспечения отказоустойчивости, системы поддержки выполнения программ;

– выработка критериев оценки суперкомпьютеров, оценочное тестирование, извлечение конструктивных инженерных знаний из результатов оценочного тестирования для оптимизации суперкомпьютеров и программ для них, оценки продуктивности программирования;

– отслеживание приложений для суперкомпьютеров и предъявляемых ими требований;

– инфраструктура сетей суперкомпьютерных центров и самих центров, вопросы закупки, внедрения и эксплуатации суперкомпьютеров.

Все перечисленные направления актуальны и сейчас и могут быть положены в основу работы отечественной экспертно-аналитической службы.

Создание экспертно-аналитической службы может служить начальным толчком к переходу на системный уровень в организации работ по суперкомпьютерной тематике, с консолидацией их в отдельной федеральной целевой программе (ФЦП СКТ, далее - Программа).

Работа экспертно-аналитической службы может быть показана на примере координации работ в рамках названной Программы (рисунок 9).

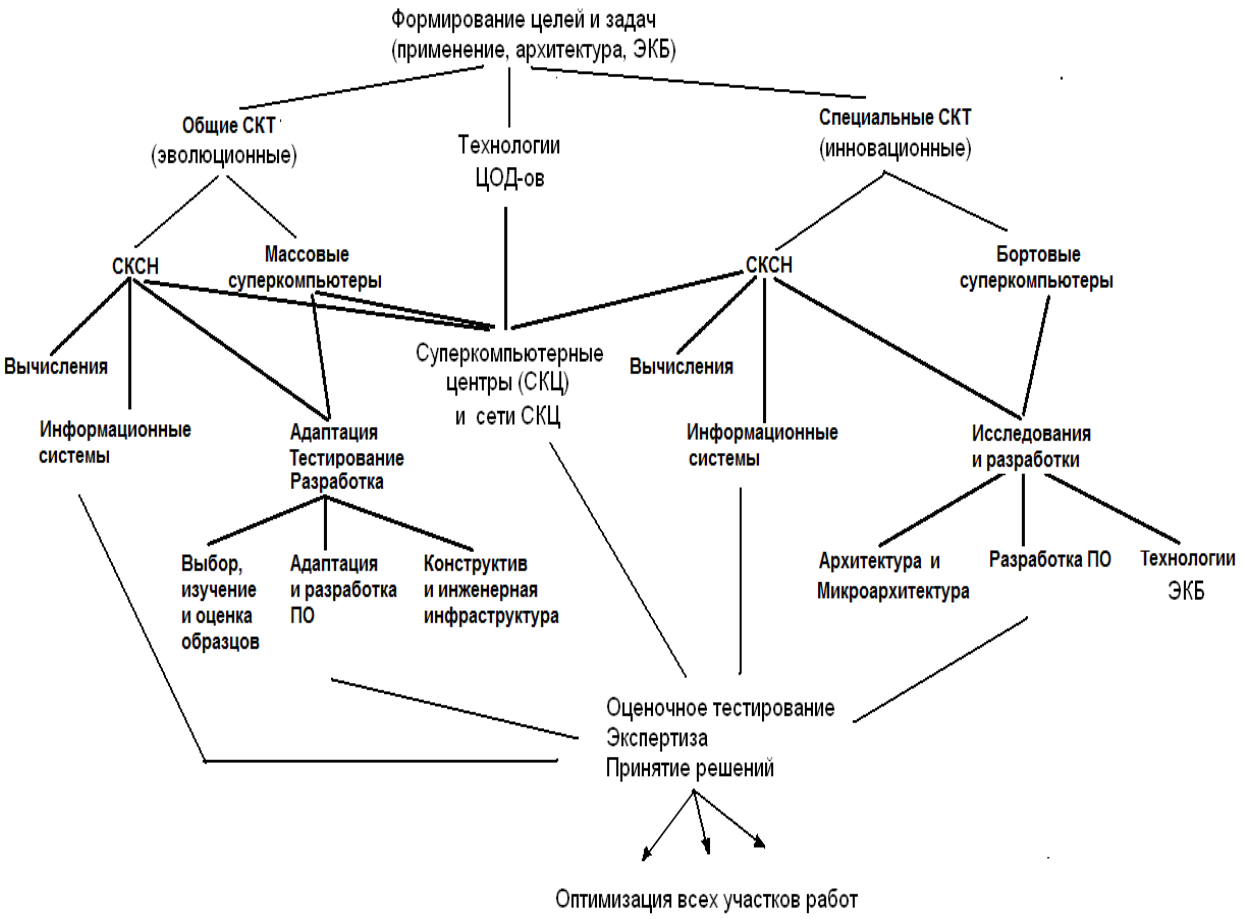

Рисунок 9 – Круг задач и участие экспертов в рамках работ по СК тематике

Такая модель предполагает наличие двух, тесно взаимодействующих между собой, органов: Научно-технического координационного совета (далее - Совет) и Экспертно-аналитической группы.

Совет может состоять из ведущих ученых и специалистов страны, представителей государственных заказчиков Программы, а также организаций промышленности и ведомств, использующих разрабатываемые в рамках Программы изделия. Совет вырабатывает рекомендации по выполнению научно-исследовательских и опытно-конструкторских работ, пользуясь рекомендациями Экспертно-аналитической группы, как на этапе подготовки обсуждений, так и в их процессе.

При Совете создается группа высокопрофессиональных экспертов-аналитиков (экспертно-аналитическая группа), способных контролировать выбор и выполнение проектов Программы, проводить и/или руководить ин-

формационно-аналитической работой по выполняемым в мире работам данной области. Совет и Экспертно-аналитическая группа периодически устраивают рабочие группы-конференции по актуальным вопросам выполнения Программы. Производится публикация в периодических научно-технических изданиях по тематике суперкомпьютеров и суперкомпьютерных вычислений. По некоторым направлениям работ, представляющим особую государственную важность, Совет и Экспертно-аналитическая группа участвуют в организации тендеров, решая вопрос по выбору исполнителя.

Уже сейчас, основываясь на мнении независимых экспертов, хотя и не организованных в составе Экспертно-аналитической службы, возможно сформулировать некоторые оценки, рекомендации и прогнозы в СКТ тематике.

Эволюционное направление развития СК

Это направление основано на комбинировании малоядерных суперскалярных микропроцессоров (предпочтительно, фирм Intel и AMD) с сопроцессорами-ускорителями в виде графических микропроцессоров, других специфических микропроцессоров такого типа, а также ПЛИС и, возможно, китайских микропроцессоров и элементов коммуникационных сетей.

Предварительные оценки показывают, что применение гибридных архитектур позволяет создавать вычислительные системы производительностью:

– 10 Пфлопс в 2015 г. с энергопотреблением 2 000 – 3 000 кВт, состоящую из $10^5/5•10^6$ MIMD/SIMD ядер;

– 100 Пфлопс в 2017 г. с энергопотреблением 4 500 – 5 000 кВт, состоящую из $10^6/5•10^7$ MIMD/SIMD ядер;

– 1000 Пфлопс в 2020 г. с энергопотреблением 20 000 кВт, состоящую из $10^7/5•10^8$ MIMD/SIMD ядер.

Работы этого направления могут быть организованы на базе следующих организаций:

– для массового рынка – Т-платформы, РСК-Технологии, Открытые технологии, ОАО «НИЦЭВТ»;

– для отдельных областей приложений и заказчиков с особенными требованиями по производительности и функциональности – ВНИИЭФ (г. Саров), НИИ МВС (г. Таганрог), ИТМ и ВТ, ФГУП НИИ «Квант», ИСП РАН, ИПС РАН, ИПМ им. М. В. Келдыша РАН;

– для областей, где требуется применение компонентной базы отечественного производства – МЦСТ/ОАО «ИНЭУМ», НИИСИ РАН, Росэлектроника (г. Фрязино), ОАО “НИЦЭВТ”.

Работу заказчиков и поставщиков в условиях быстрой смены аппаратных компонентов на рынке эволюционных технологий может облегчить совместное углубленное оценочное тестирование и анализ нового оборудования и программного обеспечения, для чего в рамках ФЦП СКТ может быть создан Центр оценочного тестирования. Ранее такой центр был при Минобрнауки РФ в НИИ РИНКЦЭ. Целесообразно его восстановить.

Инновационное направление развития СК

Для решения этих задач необходимо проведение исследований по следующим основным компонентам и проблемным направлениям:

– элементная база и схемотехника микропроцессоров – низкотемпературные КМОП-схемы, электрооптические схемы с оптическими внутрикристальными и внекристальными соединениями, технология схем одноквантовой логики (RSFQ-технологии), реверсивная логика, многозначная логика;

– элементная база и схемотехника памяти – оптимизация ячеек памяти, разработка 3-х мерных кристаллов памяти, интеллектуализация устройств управления модулями памяти и обеспечение отказоустойчивости;

– разработка перспективных микропроцессоров, реализующих новые архитектурные принципы (мультитредовость, потоковость, встроенная па-

мять, реконфигурируемые функциональные устройства и поля функциональных устройств, адаптивные), а также новые системотехнические решения (распространение синхросигналов по оптическим внутрикристальным каналам, асинхронная организация работы блоков);

– разработка новых алгоритмов работы операционных систем и систем обеспечения отказоустойчивости, включая методы мониторинга и самонастройки;

– разработка параллельных алгоритмов решения важнейших прикладных задач с использованием модели распределенной общей памяти и алгоритмического обеспечения отказоустойчивости вычислений;

– разработка новых методов организации вычислений и работы с памятью - реверсивные вычисления, транзакционная память.

– разработка компиляторов языков параллельного программирования нового поколения для высокопродуктивного программирования типа Chapel, X10, Fortress;

– выработка критериев объективной оценки реальной производительности суперкомпьютеров, подготовка тестов оценочного тестирования, разработка методик и инструментальных средств извлечения конструктивных инженерных знаний из результатов оценочного тестирования с целью оптимизации суперкомпьютеров и программ для них, разработка методик и средств оценки продуктивности программирования для суперкомпьютеров;

– изучение требований перспективных приложений для суперкомпьютеров, создание методик и систем предсказания реальной производительности, подготовка предложений по алгоритмам и архитектуре суперкомпьютеров для решения перспективных трудно решаемых задач;

– создание инфраструктуры сетей суперкомпьютерных центров и самих центров, решение вопросов закупки, внедрения и эксплуатации суперкомпьютеров, постановки прикладных исследований с использованием суперкомпьютеров в интересах обеспечения национальной безопасности и решения важнейших научно-технических и социально-экономических задач.

Так могут быть сформулированы как модель, так и уже сформированные рекомендации по развитию отрасли СКТ, которые могут положить начало действительно системному подходу к решению проблемы.

Заключение

Придание федерального статуса теме СКТ в России является важной и насущной задачей, как в плане поддержания и развития конкурентных технологий, так и в вопросах безопасности. При этом, формальная сторона вопроса не так важна, как содержательная. Если, как минимум, статус экспертного мнения будет сохранен в пассивном варианте, то ситуация не измениться. Так же в других проблемных областях СКТ, которые были рассмотрены в докладе, без содержательных инфраструктурных преобразований общую проблему не решить. Акцент на важности решения проблем в экспертной области был сделан с точки зрения ее исключительной важности на начальном, определяющем вектор направления работ и, соответственно, целеполагающем этапе реализации столь масштабного проекта.

Список литературы

1. Материалы Пятого Московского Суперкомпьютерного Форума, 21. 10.2014 г., http://www.ospcon.ru/node/107252
2. Landauer R. Irreversibility and Heat Generation in the Computing. IBM Journal, July. 1961, pp.183-191.
3. Bennet C.H., Landauer R. The Fundamental Phisical Limits of Computation. Scientifi American, July 1985, 38-46 (повторная публикация – Scientific American, June, 2011)
4. Frank M.P. Physical Limits of Computing. IEEE Computing in Science & Engineering, May/June, 2002, 22 pp.
5. Zhirnov V.V. et al. Limits to Binary Logic Switch Scaling – A Gedanken Model. Proceedings of the IEEE.Vol.91, No 11, 2003, pp.1934-1939.
6. Ming L., Review of advanced CMOS technology for post-Moore era. Institute of Microelectronics, Peking University, SCIENCE CHINA, Physics, Mechanics&Astronomy, December 2012, Vol.55, N12, 2316-2325
7. Keckler S.W., Dally W.J. *et al*. GPUs and the Future of Parallel Computing. IEEE MICRO, September/Octouber 2011, pp.7-17.
8. Workshop on "The Roadmap for Revitalization of High-End Conputing", edt.D.A.Reed, June 16-18, 2003, 110 pp. http://cra.org/uploads/documents/resources/rissues/supercomputing.web_.pdf
9. Dongarra J. et al DARPA's HPCS Program: History, Models, Tools, Languages. Technical Report, Sun Diego Supercomputer Center, 2008, 94 pp.
10. Amarasinghe S. et al. ExaScale Software Study: Software Challenges in Extreme Scale Systems. DARPA IPTO, US Air Force Research Laboratory, September 14, 2009, 153 pp.
11. Kogge P. et al. ExaScale Computing Study: Technology Challenges in Achieving Exascale Systems. DARPA IPTO, US Air Force Research Laboratory, September 28, 2008, 278 pp.