4.5. СВЕРХВЫСОКОПРОИЗВОДИТЕЛЬНЫЕ ВЫЧИСЛИТЕЛЬНЫЕ СИСТЕМЫ СЕМЕЙСТВА CRAY X

Семейство сверхвысокопроизводительных ВС Cray X разрабатывается Cray Incorporation – глобальным лидером в области суперкомпьютеров. В конце 2002 года анонсированы две модели: Cray X1 и Cray X2 соответственно с производительностью порядка $5 \cdot 10^{13}$ опер./с и $8 \cdot 10^{14}$ опер./с.

Система Cray X1 является самым высокопроизводительным средством обработки информации первого десятилетия 21 века. Она расценивается как промежуточный этап в решении стратегической проблемы США (и, в частности, корпорации Cray) достичь к 2010 году скорости вычислений 1 PetaFLOPS, т.е. одного квадриллиона или 10^{15} операций с плавающей запятой в секунду. Данная проблема была поставлена в 1999 г. в докладе Президентского консультационного комитета по информационным технологиям (President's Information Technology Advisory Committee). В США считается, что создание высокопроизводительных ВС осуществляется исключительно в интересах национальной безопасности (!).

Вычислительная система Cray X1 предназначается как для академических, так и прикладных исследований, для решения сверхтрудоемких (high-end) задач науки, техники, экономики и военной сферы. Разработка ВС Cray X1 получила поддержку от нескольких агентств правительства США, включая Агентство национальной безопасности (NSA – National Security Agency).

4.5.1. Особенности архитектуры Cray X1

Максимальная конфигурация ВС Cray X1 состоит из 4096 элементарных процессоров (ЭП), имеет производительность 52,4 TeraFLOPS и обладает памятью емкостью 34–64 Терабайта. Вес такой конфигурации ВС составляет примерно 230 т (при воздушном охлаждении) или 170 т (при жидком хладагенте). Цена 16-процессорной ВС (1,64 TFLOPS) – 16,4 млн. долларов.

Система Cray X1 была официально анонсирована в ноябре 2002 г. Первые поставки Cray X1 (в упрощенных конфигурациях, но допускающих модернизацию) произведены в конце 2002 г. и первом квартале 2003 г. К числу первых организаций, которые приобрели конфигурации Cray X1, относятся: Научно-исследовательский центр высокопроизводительных вычислений Армии США (AHPCRC – U.S. Army High Performance Computing Research Center), Испанский национальный институт метеорологии (Spain's National Institute of Meteorology), Оук – Риджская национальная лаборатория (ORNL – Oak Ridge National Laboratory) Отдела энергетики США (U.S. Department of Energy).

Cray X1 – это MIMD-система с общей распределенной памятью, ее архитектура впитала в себя достижения как PVP-, так и MPP-систем. Данная ВС основывается на тороидальной топологии и имеет широкую полосу пропускания и низкую латентность (малые задержки при передаче информации между ресурсами). Cray X1 характеризуется высокой надежностью и живучестью, а также масштабируемостью. Диапазоны возможных конфигураций, производительности и емкости памяти Cray X1 соответственно равны: 8 – 4096 процессоров, 102,4 GFLOPS – 52,4 TFLOPS и 32 Г байт – 64 Т байт.

В систему Cray X1 вложен новейший набор команд, активные исследования по которому велись в корпорации Cray в течение 10 лет. Считается, что архитектура ВС с этих набором команд будет отвечать достижениям в интегральной технологии по крайней мере в течение десятилетия. Набор команд Cray X1 весьма прост, в нем нет сложных и избыточных инструкций. Набор рассчитан на использование очень больших регистровых файлов, поддерживает 64- и 32-разрядные вычисления, реализует новый механизм синхронизации, обеспечивающий масштабируемость ВС и др. В результате Cray X1 обладает рядом преимуществ по сравнению с другими архитектурами суперкомпьютеров:

– высоким вычислительным параллелизмом (при низкой пропускной способности инструкций);

– незначительной сложностью управления;

– небольшим энергопотреблением, соотнесенным к одной операции в секунду;

– низкой латентностью.

Итак, архитектура ВС Cray X1 позволяет формировать конфигурации, адекватные областям применения, параметрам решаемых суперсложных задач.

Система Cray X1 – это композиция множества мультипроцессорных узлов, коммуникационной сети между узлами и средств ввода-вывода данных. Среда программирования Cray X1 поддерживается специальным сервером.

4.5.2. Вычислительный узел Cray X1

Система Cray X1 может иметь в своем составе от 2 до 1024 однородных вычислительных узлов (ВУ). Каждый ВУ имеет в совеем составе 4 элементарных процессора (ЭП) и распределенную общедоступную оперативную память (рис. 4.8). Взаимодействие между процессорами и оперативной памятью в узле осуществляется при помощи коммутатора ВУ (Crossbar).

Каждый ЭП – это специально спроектированный конвейерный (или векторный) процессор, обладающий производительность 12,8 GigaFLOPS (при обработке 64-разрядных операндов). Процессор поддерживает также арифметику над 32-разрядными данными.

Элементарный процессор относится к типу мультипотоковых процессоров (MSP – Multi-Streaming Processors), если придерживаться терминологии Cray Inc. Вообще, такой процессор по сути является конвейерным (или векторным), но с той особенностью, что он состоит из множества небольших конвейеров (pipes), работающих параллельно. В Cray X1 функциональная структура MSP усовершенствована, в процессоре дополнительно имеются схемы синхронизации и кэш-память. Следует отметить, что читатель без труда обнаружит сходство функциональных структур мультипотокового процессора Cray и конвейерной ВС STAR-100 (см. 4.2.1 и рис. 4.3).

Элементарный процессор Cray X1 (рис. 4.9) состоит из 4 секций обработки информации (СОИ), 4 блоков кэш-памяти и коммутатора ЭП. Каждая из секций обработки включает в себя скалярный блок с кэш-памятью для данных (СБ & КЭШ) и пару векторных конвейеров (ВК). Скалярный блок имеет тактовую частоту 400 МГц и способен выполнять 2 операции за такт. Быстродействие 4 скалярных блоков ЭП составляет 3,2 GIPS ($3{,}2 \cdot 10^9$ операций с фиксированной запятой в секунду).

Векторные конвейеры ЭП работают параллельно, синхронно и с тактовой частотой 800 МГц. Их суммарная производительность оценивается величинами 12,8 GFLOPS или 25,6 GFLOPS при обработке 64- или 32-разрядных данных. (В самом деле, на каждый из конвейеров поступает два вектора-операнда, следовательно, за один такт 8 конвейеров способны обработать 16 элементов векторов).

Кэш-память ВУ обеспечивает когерентность между быстродействием при обработке информации и скоростью ввода данных, т.е. она играет роль сверхоперативной буферной памяти между секциями обработки информации и оперативной памятью (см. 4.2.1). Кэш-память состоит из 4 блоков, ее суммарная емкость 2 М байт.

Коммутатор ЭП (Crossbar) обеспечивает доступ каждой секции обработки информации к любому блоку кэш-памяти.

Полоса пропускания в направлении от кэш-памяти к секциям обработки информации равна 102,4 Г байт/с, а наоборот – 51,2 Г байт/с. Четыре канала (рис. 4.9) между кэш-блоками и оперативной памятью обеспечивают обмен информацией со скоростью 76,8 Г байт/с.

В пределах вычислительного узла имеется оперативная память, доступная каждому ЭП. Память ВУ формируется из Rambus DRAM-микросхем, производимых Samsung Electronics Co. Ltd. Rambus-чипы характеризуются значительными емкостью и пропускной способностью. Прогресс в наращивании емкости памяти в Cray X1 отражает табл. 4.2.

Т а б л и ц а 4.2

Годы	Емкость памяти Cray X1		
	Rambus-чип	Вычислительный узел	Cray X1 с 4096 ЭП
2002	256 М бит	16 Г байт	16 Т байт
2003	512 М бит	32 Г байт	32 Т байт
2004	1 Г бит	64 Г байт	64 Т байт

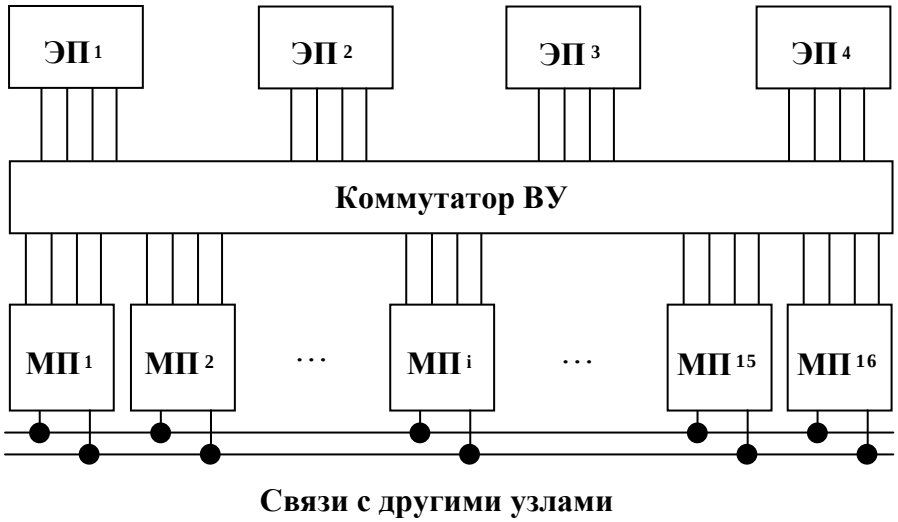

Рис. 4.8. Вычислительный узел Cray X1

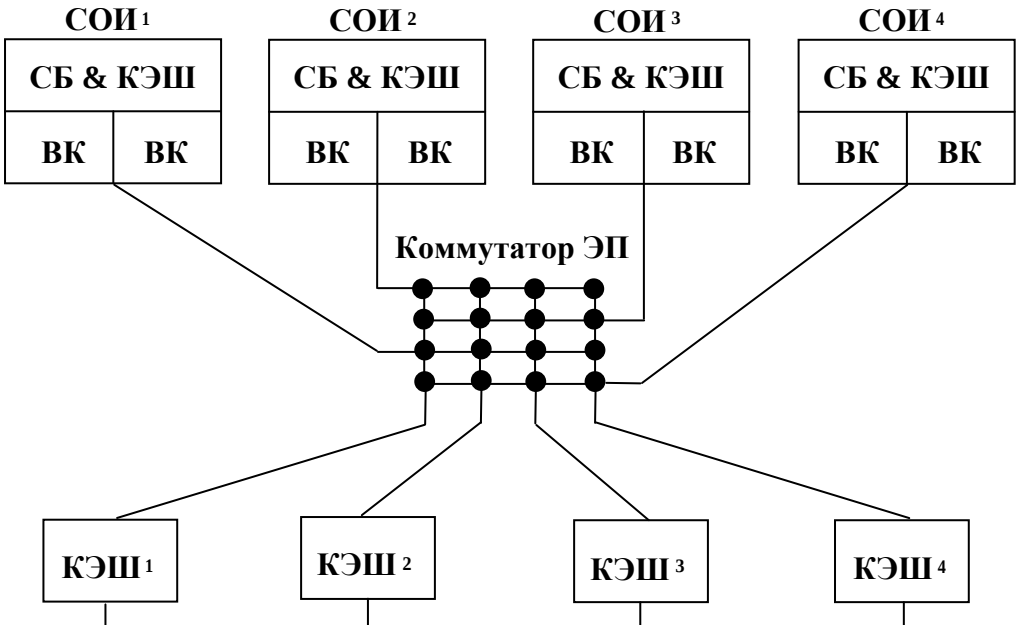

Рис. 4.9. Элементарный процессор Cray X1

Память любого узла (рис. 4.8) представляется множеством из 16 четырехканальных модулей (МП); для максимизации ее пропускной способности используется 16 контроллеров.

Каждый элементарный процессор (при помощи своих 4 кэш-блоков) имеет доступ (через коммутатор ВУ) к каждому модулю оперативной памяти узла. При этом любой кэш-блок ЭП связан только со своей группой из 4 модулей памяти. Поскольку любая из четырех секций обработки информации (рис. 4.9) связана через коммутатор со всеми блоками $\{КЭШ_i\}$, $i = \overline{1,4}$, то в пределах узла любая $СОИ_i$ имеет доступ к любому модулю памяти $МП_j$, $j = \overline{1,16}$.

Пропускная способность "канала" между элементарным процессором и оперативной памятью в вычислительном узле составляет 34,1 Г байт/с.

Оперативная память вычислительного узла доступна для других ВУ системы (рис. 4.8); этот доступ реализуется при помощи специальных маршрутизаторов.

Итак, все модули памяти в системе Cray X1 физически распределены по вычислительным узлам (и, следовательно, по элементарным процессорам), но логически они доступны каждому ЭП. Заключаем: оперативная память ВС Cray X1 является и распределенной, и общей.

Следует отметить, что элементарный процессор является основным функциональным элементом Cray X1. Он конструктивно выполнен в виде многокристального модуля.

Конструкция вычислительного узла Cray X1 оформлена в виде платы, содержащей 4 конструктивных модуля – процессора, схемы памяти и коммутатора.

4.5.3. Коммуникационная сеть Cray X1

Взаимодействие между вычислительными ресурсами (узлами и, следовательно, элементарными процессорами и памятью) в системе Cray X1 осуществляется через коммуникационную сеть. Архитектурные решения, заложенные в коммуникационную сеть, позволили достичь в сверхвысокопроизводительной системе Cray X1 высокой надежности и живучести, масштабируемости, большой пропускной способности и незначительной латентности (задержки) при передаче информации между ресурсами. Так, например, пропускная способность сети в 64-процессорной конфигурации Cray X1 (819,2 GFLOPS, 256 Г байт) с жидкостным охлаждением равна 400 Гигабайт/с.

В системе Cray X1 для реализации коммуникационной сети применен модифицированный двумерный тор (Modified 2D Torus). В чем состоит суть модификации 2D-тора и что является его вершиной?

Ранее (см. 4.5.2) было отмечено, что вычислительный узел (рис. 4.8) обладает двумя маршрутами для связи с другими ВУ в пределах системы Cray X1. (Если быть более

точным, то в узле имеется 16 пар отдельных маршрутов по одному на каждый из 16 модулей памяти, что гарантирует живучесть и необходимую полосу пропускания связей между ВУ). Один из этих маршрутов используется для того, чтобы организовать в системе Cray X1 множество связных пар: “ВУ с нечетным номером – смежный ВУ с четным номером”. Другой маршрут служит для подключения ВУ к маршрутизатору (Router).

В качестве вершины (Vertex) двумерного тора используется композиция из четырех пар вычислительных узлов и двух маршрутизаторов ($М_1$, $М_2$), работающих на 4 внешних связи (рис. 4.10). Индекс в обозначении $ВУ_k$, $k=\overline{1,8}$, не является физическим номером вычислительного узла, он дает лишь информацию о четности или нечетности номера. Маршрутизаторы $М_1$ и $М_2$ обеспечивают два параллельных канала связи данной вершины с соседними вершинами в 2D-торе.

Очевидно, что структура (граф) вершины в системе Cray X1 обладает диаметром, равным 2. (Диаметр графа – максимальное расстояние, определяемое на множестве кратчайших путей между парами вершин). Это следует из того, что маршрутизатор не имеет задержки, сравнимой со временем обмена информацией между памятями различных ВУ. Значит, в вершине обмен информацией между любыми ВУ осуществляется с использованием максимум одного транзитного узла, или, говоря иначе, он производится посредством двух пересылок (“hops”): из данного ВУ – в транзитный, а затем в ВУ – приемник.

Двумерный тор Cray X1 – это “бублик”, на поверхности которого размещена двумерная структура, вершины (рис. 4.10) которой связаны в двух направлениях: по окружности “бублика” и по окружности его сечения. Примерами таких структур могут служить трех- и четырехмерные гиперкубы (с числами вершин и ребер, равными 8 и 3 или 16 и 4, см. рис. 3.2). Не требуется особого воображения увидеть в этих гиперкубах двумерные торы.

Последний (4-мерный) гиперкуб использован в конфигурации Cray X1, состоящей из 128 вычислительных узлов (512 элементарных процессоров). Ясно, что в этой конфигурации ВС сама вершина имеет, в свою очередь, свою структуру из 8 ВУ и двух маршрутизаторов, а ребра в гиперкубе отражают двойные каналы межвершинных связей (рис. 4.10).

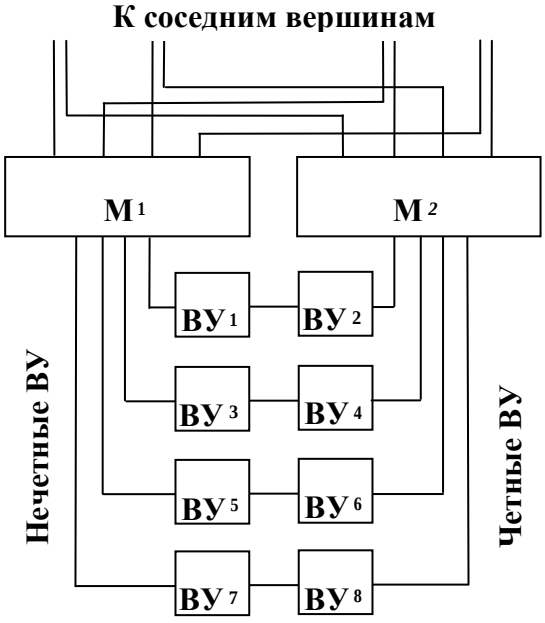

Рис. 4.10. Вычислительная вершина двумерной тороидальной системы Cray X1

Четырехмерный гиперкуб (рис.3.2) – это структура из трехмерного куба внутри такого же куба и с ребрами между соответствующими вершинами этих кубов.

Четырехмерный куб Cray X1 характеризуется тем, что вершины (точней их маршрутизаторы) каждого трехмерного куба входят в двойные циклы. Следовательно, в четырехмерном кубе один из концентрических циклов представляет вычислительные узлы (рис. 4.10) с нечетными номерами, а другой – ВУ с четными номерами. Далее, в четырехмерном гиперкубе Cray X1 имеют место также связи между соответствующими циклами с нечетными ВУ двух трехмерных кубов, а также между циклами с четными узлами этих же кубов. Заметим, что связи между циклами с нечетными ВУ и с четными узлами организуются в пределах вершин и так как это показано на рис. 4.10.

Оценим задержки, которые существуют при передачи информации между вычислительными узлами в четырехмерном гиперкубе Cray X1. Ясно, что максимальное расстояние между любыми двумя нечетными ВУ (или четными узлами) равно 4, т.е. для обмена информацией между этими узлами потребуется максимум 4 пересылки (“hops”). Максимальное расстояние между любыми нечетным и четным ВУ увеличивается на единицу. Здесь используется не 3, а 4 транзитных узла (необходима пересылка между узлами внутри вершины).

Рассмотрим как решается проблема масштабирования структуры ВС Cray X1. Корпорация Cray в проекте Cray X1 не использует структуру в виде гиперкуба при наращивании количества вычислительных узлов за пределами 128. Гиперкубы больших размерностей, чем 4, потребовали бы включения в состав вершин дополнительных маршрутизаторов (см. рис. 4.10), что породило бы набор неоднородных вершин. Вместо этого Cray Inc. “растягивает” четырехмерный куб, превращая его в 2D-тор с большей “окружностью”, и тем самым увеличивает число вершин.

Итак, система Cray X1 имеет иерархическую структуру сети связей между вычислительными ресурсами. На каждом структурном уровне используется свой тип графа межресурсных связей и свой тип вычислительных ресурсов – элементов обработки информации (см. табл. 4.3). Такое структурное решение в системе Cray X1 позволило достичь оптимума по эффективности в условиях технических и технологических ограничений на рубеже между 20 и 21 столетиями.

Т а б л и ц а 4.3

Структурные уровни Cray X1	Вычислительные элементы структуры
Макроуровень – двумерный тор (рис. 3.2).	Вершина – четырехполюсник, отражающий композицию из 8 вычислительных узлов (ВУ) и 2 маршрутизаторов ($М_1$, $М_2$). Элементы $М_1$ и $М_2$ формируют два двумерных канала.
Структура вершины – граф, состоящий из 4 пар связных ВУ с четными и нечетными номерами и 2 маршрутизаторов, каждый из которых соединен ребрами либо с четными, либо с нечетными узлами (рис. 4.10).	Вычислительный узел – двухполюсник, представляющий композицию из 4 элементарных процессоров (ЭП), 16 модулей памяти (МП) и коммутатора ВУ. Маршрутизатор – четырехполюсник по внешним связям, позволяет формировать двумерные структуры.
Структура вычислительно узла – граф, дающий связность каждого из четырех ЭП с каждым из 16 модулей памяти (рис. 4.8).	Элементарный процессор – четырехполюсник, соответствующий композиции из 4 секций обработки информации (СОИ), 4 блоков кэш-памяти и коммутатора ЭП. Коммутатор ВУ обеспечивает связность между процессорами $ЭП_1$ – $ЭП_4$ и модулями памяти ($МП_1$ – $МП_{16}$).

Структура элементарного процессора – граф, создающий связность каждой из четырех СОИ с каждым из 4 блоков кэш-памяти (рис. 4.9).	Секция обработки информации – композиция из скалярного блока с кэш-памятью (СБ & КЭШ) и двух векторных конвейеров (ВК). Коммутатор ЭП дает связность между $СОИ_1$ – $СОИ_4$ и блоками $КЭШ_1$ – $КЭШ_4$.

4.5.4. Средства ввода-вывода Cray X1

Средства ввода-вывода информации ВС Cray X1 распределены по ее вычислительным узлам. Каждый ВУ располагает 4 каналами ввода-вывода (I/O System Port Channels). Пиковая полоса пропускания одного канала ввода-вывода составляет 1,2 Гигабайт/с.

Каналы ввода-вывода Cray X1 служат для подключения дисков и других периферийных устройств. Предусматривается возможность использования волоконно-оптических линий связи.

Поддержка различных сетевых протоколов (в частности для гигабитной Ethernet) осуществляется специальным сервером CNS (Cray Network Server).

4.5.5. Конструкция системы Cray X1

Для формирования Cray X1 используются корпуса двух вариантов, с воздушным и водяным охлаждением. В корпусе первого варианта размещается 4 вычислительных узла (16 элементарных процессоров), а второго варианта – 16 узлов (64 ЭП).

В табл. 4.4. приведены физические характеристики конструктивов для системы Cray X1.

Таблица 4.4.

Тип корпуса ВС	Размер площадки, $м^2$	Вес, кг
Основной с воздушным охлаждением	$0{,}9 \times 1{,}5$	895
Основной с жидкостным охлаждением	$1{,}3 \times 2{,}6$	2610
Для средств ввода-вывода	$0{,}75 \times 1{,}1$	512

4.5.6. Программное обеспечение Cray X1

Архитектура сверхвысокопроизводительной ВС Cray X1 по сути является объединением архитектур PVP и MPP. Поэтому ее операционная система впитала в себя все лучшее из PVP UNICOS и MPP UNICOS/mk.

Среди средств программирования Cray X1 имеются языки высокого уровня (FORTRAN, C), интерфейсы передачи сообщений (MPI) и др.

В Cray X1 среда программирования поддерживается специальным сервером – CPES (Cray Programming Environment Server). В частности, компиляторы работают не на самой системе Cray X1, а на CPES.

4.5.7. Области применения системы Cray X1

Cray X1 – универсальная сверхвысокопроизводительная масштабируемая вычислительная система. Архитектура Cray X1 позволяет формировать конфигурации ВС, в которых достигается оптимум между быстродействием, емкостью памяти, надежностью

и ценой и которые адекватны областям применения. Всё разнообразие видов деятельности человека, связанных с трудоемкими вычислениями, и составляет прикладные области для Cray X1. Но главными областями для данной ВС все же являются наука, техники и экономика (как гражданской, так и военных сфер).

Области применения Cray X1:

- фундаментальные научные исследования, вычислительная математика, физика, химия, астрономия (включая астрофизику), биология, науки о Земле;

- биотехнические исследования (изучение геномов организмов и строения белка), биоинформатика и моделирование биологических процессов;

- медицина и фармакология; виртуальная хирургия; создание лекарств; предсказание естественной пандемии и локальных эпидемий;

- экология, окружающая среда, климат, погода; моделирование процессов ионосферной плазменной физики; моделирование климата и атмосферы, долгосрочное, краткосрочное и очень краткосрочное прогнозирование погоды; изучение взаимодействия между океанической, воздушной и земной средами; моделирование цунами и воздушных течений, предсказание сильных штормов (бурь);

- аэрокосмические исследования и индустрия; вычислительная механика; моделирование летательных аппаратов; механика сплошных сред; аэро- и гидродинамика; проектирование в авиации и космонавтике, анализ эффективности горючего;

- экономика, моделирование национальной экономики и инвестиций в отрасли народного хозяйства (энергетику, транспорт и др.);

- оборона и военные приложения; моделирование и планирование обороны, наступления и боя; управление сложными техническими системами; предсказание климата после военных действий и погоды после боя; борьба с биотерроризмом и эпидемиями;

- промышленность, машиностроение, энергетика, металлургия; создание новых материалов; нанотехнологии и наноматериалы; анализ, проектирование и обеспечение безопасности в автомобилестроении; нефтехимический сейсмический анализ и др.