4.4. МАССОВО-ПАРАЛЛЕЛЬНЫЕ ВЫЧИСЛИТЕЛЬНЫЕ СИСТЕМЫ CRAY

В конце 80-х годов прошлого столетия ряд компаний (Thinking Machines, Kendal Square, NCube, MasPar и Mieko) успешно проводили исследования и разработки новых архитектур суперВС. В таких суперВС (Massively Parallel Processing Systems) высокая эффективность достигалась за счет применения большого количества элементарных (простых) процессоров.

Системы с массовым параллелизмом (MPP-системы) стали альтернативой для векторно-параллельных ВС (PVP-систем). Это создало угрозу лидерству Cray Research Inc. в области суперкомпьютеров. В 1991 г. была анонсирована программа, предусматривающая 5-летний срок для приобретения фирмой Cray Research лидерства по MPP–системам. Программа была поддержана агентством ARPA (Advanced Research Projects Agency – агентство перспективных исследовательских проектов), сумма финансирования с 1991 г. по 1996 г. составила 37,7 млн. долларов. В рамках данной программы было разработано семейство массово-параллельных ВС, включающее возможные конфигурации моделей: Cray T3D и Cray T3E. Эти ВС обеспечивают обработку информации с производительностью от десятков GigaFLOPS до TeraFLOPS и предоставляют память емкостью от Гигабайт до Терабайт.

4.4.1. Вычислительная система Cray T3D

Система Cray T3D – первая MPP-система корпорации Cray Research, ее разработка была завершена в 1993 г. Она позволила Cray Research Inc. быстро захватить лидерство на рынке MPP-систем. Допустимые количества элементарных процессоров в конфигурациях системы Cray T3D – 32 – 2048, а диапазоны производительности и емкости памяти соответственно равны 5 – 300 GFLOPS и 512 Мегабайт – 128 Гигабайт. Система в максимальной конфигурации никогда не строилась; обычная конфигурация Cray T3D – 64-процессорная, она обеспечивала быстродействие, равное 10 GFLOPS.

Архитектура системы Cray T3D – это MIMD, а сама ВС принадлежит к виду распределенных. В системе достаточно полно воплощены принципы модели коллектива вычислителей (см. 3.1.1). Последнее позволило, в частности, достичь в Cray T3D высокой надежности и живучести, а также масштабируемости (варьируемости числа процессоров: 32, 64, 128, 256, 512, 1024 или 2048). Следовательно, архитектура Cray T3D приспособлена к формированию конфигураций с заданной производительностью и/или стоимостью.

Система Cray T3D работает под управлением хост-системы (Host System – управляющая ВС). Среди функций хост-системы – производительная подготовка программ (включающая компиляцию) и ввод-вывод данных для Cray T3D. В качестве хост-системы могут быть использованы, в частности, конфигурации ВС Cray Y-MP и Cray C90. Между хост-ВС и системой Cray T3D имеется высокоскоростной канал связи (200 Мбайт/с).

Вычислительная система Cray T3D – это композиция множества вычислительных узлов, коммуникационной сети (или сети межузловых связей), каналов ввода-вывода информации и средств синхронизации.

1. Вычислительный узел Cray T3D

Все вычислительные узлы (ВУ), составляющие ВС Cray T3D, – однородные. Узел (Processing Element Node) системы включает в себя (рис. 4.6) два одинаковых элементарных процессора (ЭП) и локальный коммутатор (ЛК).

Элементарный процессор (или PE, Processing Element – процессорный элемент) представляется композицией из микропроцессора, локальной памяти (ЛП) и устройства управления памятью (УУП). Микропроцессор – это DEC 21064 Alpha chip (или просто

DEC Alpha), т.е. RISC-процессор типа Alpha фирмы DEC. (Reduced Instruction Set Computer – компьютер с сокращенным набором команд). Микропроцессор имеет кэш-память для команд и кэш-память для данных. Набор команд предусматривает и логические, и арифметические операции целочисленной и вещественной арифметики. Для архитектуры DEC Alpha характерно следующая спецификация:

- разрядность – 64;
- тактовая частота – 150 МегаГц;
- производительность: 150 MFLOPS, 300 MIPS (3 инструкции за цикл);
- емкость кэш-памяти для команд и данных: 8 К байт и 8 К байт;
- виртуальное адресное пространство – 43 бита;
- технология – КМОП (комплементарная “металл – окисел – проводник”);
- геометрические нормы – 0,75 мкм.

Ясно, что адресное пространство DEC Alpha позволяет организовать оперативную память емкостью в несколько Терабайт (2^{43} байт).

Локальная память ЭП – это DRAM-память емкостью 16–64 М байт (Dynamic Random Access Memory – динамическая память с произвольной выборкой). Каналы связи микропроцессора с локальной памятью ЭП характеризуются малой задержкой и высокой пропускной способностью. Устройство управления памятью ЭП (Support Circuitry) осуществляет поддержку обмена данными между элементарными процессорами.

Локальный коммутатор обеспечивает непосредственную связь вычислительного узла с соседними узлами и представляет собой шестиполюсник. В состав ЛК входят: сетевой маршрутизатор, сетевой интерфейс и контроллер прямого доступа к памяти.

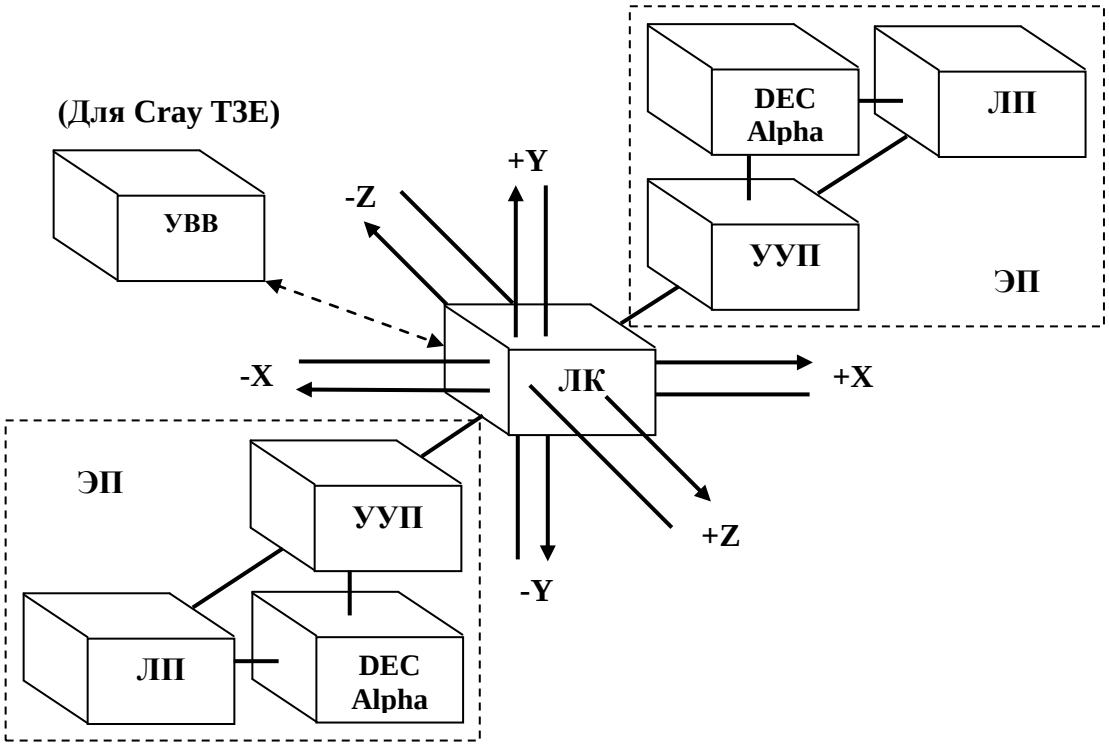

Рис. 4.6. Вычислительный узел системы Cray T3D (Cray T3E)

Сетевой маршрутизатор (Network Router) ВУ – основной элемент формирования коммуникационной сети Cray T3D. Он способен работать с 3 парами двунаправленных межузловых связей (Links), следовательно, позволяет создавать трехмерные структуры ВС. Маршрутизатор каждого ВУ определяет путь перемещения каждого пакета данных и может осуществлять параллельный транзит данных по всем трем межузловым связям.

Сетевой интерфейс (Network Interface) вычислительного узла специальным образом кодирует информацию перед ее пересылкой по коммуникационной сети другому ВУ или в канал ввода-вывода. Сетевой интерфейс служит также для приема данных от других ВУ или из канала ввода-вывода и распределяет их между элементарными процессорами данного вычислительного узла.

Контроллер прямого доступа к памяти (Block Transfer Engine) осуществляет асинхронное перераспределение данных в пределах всей распределенной памяти ВС Cray T3D , т.е. перераспределение информации, находящейся в локальной памяти разных ЭП системы, без прерывания работы самих элементарных процессоров.

Следует подчеркнуть, что в системе Cray T3D память – распределенная, точнее: она физически распределенная, но логически общая. Каждый ЭП имеет непосредственный доступ к своей локальной памяти, но он может обратиться и к локальной памяти другого ЭП, не прерывая его работы. Обращение к памяти другого ЭП осуществляется лишь в 6 раз медленнее, чем обращение к своей собственной локальной памяти. Такие возможности памяти поддерживаются аппаратурой вычислительных узлов и коммуникационной сетью.

2. Коммуникационная сеть Cray T3D

Коммуникационная сеть (Interconnect Network) системы Cray T3D предназначается для реализации обменов информацией между вычислительными узлами, а также между ВУ и каналами ввода-вывода. Она образуется из связей (Communication Links) и сетевых маршрутизаторов (Network Routers) как вычислительных узлов, так и каналов ввода-вывода информации (рис. 4.7).

Ориентация ВС на решение трехмерных сложных задач предопределила ее структуру, именно, трехмерную структуру коммуникационной сети. В системе Cray T3D каждый вычислительный узел связан с соседними по трем направлениям X, Y и Z, причем по каждому направлению вершины образуют замкнутое кольцо. Говоря точнее, структура коммуникационной сети Cray T3D является циркулянтным графом с тремя образующими (см. 3.1.2) или *трехмерным (3D) тором* (рис. 4.7). Каждая связь между двумя соседними узлами – это два однонаправленных канала передачи данных, такая связь допускает одновременный обмен информацией в противоположных направлениях.

Быстродействие коммуникационной сети Cray T3D по каждому из двух направлений передачи информации – 140 Мегабайт/с.

Двунаправленный трехмерный тор имеет преимущества перед “незамкнутыми” трехмерными топологиями:

– возможность быстрой связи граничных узлов и небольшая латентность (задержка) при передаче информации между вершинами (максимальное расстояние между вершинами для конфигурации из 128 ЭП равно 6, а для 2048 ЭП – 12);

– повышенная живучесть структуры – возможность выбора маршрутов для обхода поврежденных узлов и связей.

Следует подчеркнуть, что в трехмерном торе каждый вычислительный узел непосредственно связан с 6 соседними узлами, и он входит в три кольца ВУ, соответствующих направлениям X, Y, Z. Например, в 3D-торе 64 ($4^*4^*4^*$) имеется 64 узла, причем каждый узел входит в три кольца из 4 ВУ.

Адресация (нумерация) вычислительных узлов Cray T3D подразделяется на физическую, логическую и виртуальную. Каждому ВУ присвоен свой *физический адрес*, определяющий его абсолютное расположение в системе; этот адрес используется непосредственно аппаратурой. Вычислительному узлу может быть присвоен также *логический адрес*, определяющий его расположение в логической конфигурации системы, которая уже и будет представлять собой трехмерный тор. Например, 512-процессорная конфигурация системы Cray T3D реально содержит 260 физических ВУ, четыре из которых составляют резерв. Следовательно, логические конфигурации ВС суть

физические системы с повышенной надежностью. Виртуальная адресация узлов введена для того, чтобы пользователю предоставлять дополнительный сервис: он не должен учитывать при программировании физические и логические адреса ВУ, а может вводить свои (виртуальные) адреса вычислительных узлов. При этом каждой программе пользователя из трехмерного тора будет выделен вполне определенный прямоугольный параллелепипед, на котором и будет исполняться данная программа (не учитывая средств операционной системы).

Рис. 4.7. Фрагмент структуры коммуникационной сети Cray T3D (Cray T3E)
– вычислительный узел; – узел ввода; – узел вывода

Любой из адресов ВУ представляется трехкомпонентным вектором: (x, y, z), который однозначно определяет расположение узла в трех измерениях коммуникационной сети Cray T3D. Физический адрес ВУ – это абсолютный неизменяемый номер узла в сети, а логический и виртуальный адреса ВУ являются относительными адресами (относительно абсолютного).

Относительные адреса получают при помощи операции "смещения" исходного адреса. Поясним смысл относительной адресации на примере 3D-тора вида 64 ($4^*4^*4^*$):

$x,$	$y,$	z	
$2,$	$0,$	3	– исходный адрес,
$3,$	$2,$	0	– относительный адрес,
$+1,$	$+2,$	-3	–"смещения".

Легко заметить, что требуемый относительный адрес в рассматриваемом торе может быть получен и при помощи смещения $(-3, -2, +1)$. Приведенный пример демонстрирует простоту механизма преобразования физического адреса ВУ в его логический адрес.

В системе Cray T3D отображение логических адресов на физические адреса вычислительных узлов обеспечивается таблицей маршрутизации, загружаемой в сетевые маршрутизаторы. Гибкость средств отображения адресов и маршрутизации позволяет логически изолировать неисправные вычислительные узлы в Cray T3D.

3. Каналы ввода-вывода Cray T3D

Каналы ввода-вывода (Input/Output Gateways – порты ввода-вывода) предназначаются для обмена информацией между Cray T3D и управляющей системой (Host System) или кластером ввода-вывода (Input/Output Cluster).

Канал ввода-вывода Cray T3D представляется композицией из узлов ввода и вывода и низкоскоростного устройства передачи запросов и ответов.

Функциональные структуры узлов ввода и вывода предельно близки к структуре вычислительного узла, в состав каждого из первых двух узлов входят элементарный процессор, локальный коммутатор и высокоскоростные схемы ввода или вывода соответственно. Локальный коммутатор любого из узлов ввода или вывода включает в свой состав сетевой маршрутизатор (Network Router), сетевой интерфейс (Network Interface) и контроллер асинхронного прямого доступа к памяти (Block Transfer Engine). Однако, в отличие от вычислительного узла, здесь сетевой маршрутизатор рассчитан на работу со связями только по направлениям X и Z.

Узлы ввода и вывода соединяются друг с другом не только посредством своих маршрутизаторов, но и через устройство передачи запросов и ответов.

Каналы ввода-вывода включаются в коммуникационную сеть Cray T3D только в “кольцо” направлений X и Z (см. рис. 4.7). Хост-ВС подсоединяется к каналам через высокоскоростные (200 М байт/с) схемы ввода и вывода, а также через низкоскоростные устройства передачи запросов и ответов. В комплексе “хост-ВС – Cray T3D” запросы и ответы используются для управления потоком передачи данных через высокоскоростные схемы ввода и вывода.

4. Средства синхронизации Cray T3D

В системе Cray T3D весь коллектив вычислительных узлов и каналы ввода-вывода работают синхронно. Это достигается при помощи генератора тактовых импульсов (Clock), который посылает импульсы одновременно и в вычислительные узлы, и в узлы ввода и вывода. Генератор работает на частоте 150 МГц.

Для синхронизации параллельных вычислительных процессов (реализуемых в различных ЭП) в системе Cray T3D имеются специальные аппаратурные средства. Эти средства – распределенные, т.е. они организуются из специальных локальных схем поддержки синхронизации, расположенных в элементарных процессорах.

В системе Cray T3D осуществлена аппаратурная реализация механизмов синхронизации “барьер” и “эврика” (Barrier/Eureka). Для реализации механизма “барьер” в каждой ветви параллельной программы задается точка синхронизации, при достижении которой каждый элементарный процессор должен ждать до тех пор, пока остальные ЭП не дойдут до своих точек, и лишь после этого все процессоры могут продолжать работу дальше. Ясно, что такая синхронизация требуется перед осуществлением коллективных обменов информацией между ветвями параллельной программы (см. 3.3.5). Механизм синхронизации “эврика” запускает продолжение всех параллельных процессов, если из них даже только один достиг точки синхронизации. Механизмы синхронизации необходимы для реализации программирования, характерного и для SIMD-, и для MIMD-архитектур.

5. Эффективность системы Cray T3D

Не претендуя на полноту исследования, здесь мы оценим эффективность ВС Cray T3D по Амдалу. Закон Амдала обычно применяют при оценке ускорения векторных ВС (с конвейерной организацией вычислений). Система Cray T3D относится к ВС с массовым параллелизмом, вместе с тем она принадлежит к числу изделий фирмы Cray Research Inc., поэтому по Амдалу рассчитывают и ее эффективность. В табл. 4.1 приведены максимальные значения для коэффициента ускорения по Амдалу:

$$\chi^* \le \frac{1}{f + (1-f)/n},$$

где f – доля последовательных вычислений при реализации программы, n – число элементарных процессоров.

Т а б л и ц а 4.1.

Число ЭП	Доля последовательных вычислений				
	50 %	25%	10%	5%	2%
32	1,94	3,66	7,80	12,55	19,75
512	1,99	3,97	9,83	19,28	45,63
2048	2,00	3,99	9,96	19,82	48,83

Из табл. 4.1 видно, что ускорение тем выше, чем меньше доля последовательных вычислений. Избежать последовательных участков в параллельной программе нельзя; в самом деле, в программе всегда присутствуют сугубо последовательные действия, например, инициализация и операции ввода-вывода. Но табл. 4.1 заставляет задуматься и над основным: как достичь при работе на ВС с массовым параллелизмом линейной зависимости ускорения от числа элементарных процессоров? Резюмируя опыт работы пользователей Cray T3D, а главное, опираясь на наши отечественные результаты по параллельному программированию, можно заключить, что кардинальный путь повышения эффективности ВС с массовым параллелизмом связан с методикой крупноблочного распараллеливания сложных задач (см. 3.3.6).

4.4.2. Вычислительная система Cray T3E

Система с массовым параллелизмом Cray T3E – преемник опыта, полученного при создании и эксплуатации ВС Cray T3D. Вычислительная система Cray T3E была построена в 1995 г. Количество элементарных процессоров в ВС – 16–2048, диапазон производительности – 14,4 GigaFLOPS – 2,76 TeraFLOPS, емкость памяти – 1 Гигабайт – 1 Терабайт. Цена 128-процессорной конфигурации Cray T3E – 3–4 млн. долларов.

Производительность ВС определяется как количеством процессоров, так и возможностями базового микропроцессора. Выделяется несколько модификаций ВС: Cray T3E, Cray T3E-900, Cray T3E-1200, Cray T3E-1200E, Cray T3E-1350, с тактовыми частотами от 300 до 675 МегаГц. Барьер производительности 1 TeraFLOPS, т.е. 10^{12} операций с плавающей запятой в секунду над 64-разрядными данными, был впервые преодолен на системе Cray T3E-1200 в 1998 году

Архитектура ВС Cray T3E относится к классу MIMD, но она более развитая по сравнению с совместимой архитектурой Cray T3D. Отметим две архитектурные особенности Cray T3E:

1) мультипрограммирование – возможность одновременной реализации нескольких параллельных программ на различных подсистемах;
2) масштабируемость – варьируемость количества элементарных процессоров с квантом 4 или 8 (производятся модулями с 4 или 8 ЭП в зависимости от вида охлаждения ВС, воздушного или жидкостного).

Следует заметить, что в Cray T3D был реализован только монопрограммный режим. Следовательно, если для решения какой-либо задачи не требовались все ресурсы ВС, то имели место простои неиспользованных ЭП. В Cray T3E мультипрограммирование позволяет избежать простоев элементарных процессоров. Далее, в системе Cray T3D

допускались конфигурации только с числом ЭП, кратным 32. В системе Cray T3E существенно шире возможности по формированию конфигураций, адекватных сферам применения.

Функциональные структуры систем Cray T3D и Cray T3E на макроуровне полностью идентичны. Каждая из них представляется композицией множества вычислительных узлов, коммуникационной сетью в виде трехмерного тора, каналов ввода-вывода и средств синхронизации. Однако при технической реализации Cray T3E нашли место новшества. Ниже будет продолжен анализ архитектурных возможностей ВС Cray T3E.

1. Вычислительный узел Cray T3E

Ядром элементарного процессора вычислительно узла (рис. 4.6) в любой модификации Cray T3E служит микропроцессор семейства DEC 21164 Alpha. Остановимся на спецификации микропроцессора для модификации ВС Cray T3E-1350, т.е. DEC 21164 Alpha (EV5.6):

- разрядность операндов: 32 или 64;
- тактовая частота – 675 МегаГц;
- производительность: 1350 MFLOPS (2 операции с плавающей запятой за такт), 2700 MIPS (4 инструкции за такт);
- быстродействие канала к памяти – 1200 Мегабайт/с.

Элементарный процессор располагает своей локальной памятью, емкость которой варьируется в пределах от 64 до 512 Мегабайт (в зависимости от модификации ВС, в частности). В Cray T3E-1350 локальная память ЭП имеет емкость в пределах: 250 – 512 М байт и формируется из 64-мегабайтных DRAM-схем.

В вычислительном узле ВС Cray T3E в отличие от ВУ системы Cray T3D, предусмотрена специальная связь (Link) для непосредственного подключения устройств ввода-вывода (УВВ) информации (рис. 4.6). Эта связь предоставляет потенциальную возможность подключения к любому вычислительному узлу внешних средств. Однако далеко не все ВУ должны оснащаться устройствами ввода-вывода. Если же такое подключение имело место, то устройство ввода-вывода становилось общим ресурсом для четырех вычислительных узлов.

2. Коммуникационная сеть Cray T3E

Сеть межузловых связей Cray T3E – трехмерный тор с двунаправленными каналами (рис. 4.7). Она имеет крайне малое время задержки при пересылке сообщений (обладает низкой латентностью, Latency) и характеризуется значительной шириной полосы пропускания. Так, например, модификация Cray T3E-1350 имеет быстродействие 650 Мегабайт/с в каждом из двух направлений передачи информации. Данная сеть в 3,4 раза превосходит по быстродействию аналогичную сеть Cray T3E.

3. Каналы ввода-вывода Cray T3E

В системе Cray T3E реализована возможность осуществлять обмен информацией с внешней средой через множество каналов ввода-вывода (портов).

Каналы ввода-вывода Cray T3E интегрированы в трехмерную коммуникационную сеть так, что их количество всегда пропорционально числу элементарных процессоров в любой конфигурации системы. Таким образом, при масштабировании ВС происходит и адекватное масштабирование пропускной способности каналов ввода-вывода.

Ясно, что все каналы ввода-вывода (все их узлы ввода и вывода) Cray T3E закоммутированы в два гигакольца (GigaRings), данные по которым перемещаются в противоположных направлениях. Суммарная пропускная способность этих гигаколец равна величине 1 Гигабайт/с; максимальная полоса пропускания любого интерфейса гигакольца составляет 500 Мегабайт/с.

4. Конструктивные особенности системы Cray T3E

Вычислительная система Cray T3E изготовляется в двух вариантах корпусов: с воздушным и жидкостным охлаждением. В первом варианте конструктивный модуль для компоновки (масштабирования) ВС представляется платой из 4 элементарных процессоров, а при применении жидкостного охлаждения подобный модуль имеет 2 платы (8 ЭП). Следовательно, в системах с воздушным или жидкостным охлаждением масштабирование осуществляется на величину, кратную 4 или 8 ЭП, соответственно. При этом каждая 4-процессорная плата имеет только один вывод на разъем корпуса для ее включения в гигакольцо ввода-вывода.

В корпусе с жидкостным охлаждением для системы (например, Cray T3E-1350) размещается 272 элементарных процессора, из которых 256 ЭП являются основными, а остальные 16 ЭП составляют избыточность (резерв). Следовательно, на каждые 16 основных ЭП предусматривается один избыточный процессор. Максимальная конфигурация ВС Cray T3E размещается в 8 корпусах и насчитывает 2176 элементарных процессоров, из которых число основных ЭП равно 2048.

Ясно, что в любой конфигурации Cray T3E избыточность оценивается 6,25%, и она используется компонентами операционной системы и обеспечивает высокий уровень надежности ВС в целом.

5. Программное обеспечение Cray T3E

Архитектурные особенности MPP-систем потребовали от Cray Research Inc. разработки нового программного обеспечения, учитывающего мировой опыт и традиции в параллельном программировании, а также в программировании PVP-систем Cray.

Операционная система UNICOS/mk, разработанная для ВС Cray T3E, по сути является распределенной и масштабируемой версией UNICOS (последняя использовалась в PVP-системах: Cray-1, Cray X-MP, Cray-2; UNICOS – в свою очередь производная от системы UNIX).

Масштабируемая ОС (Scalable Operating System) UNICOS/mk разделена на программы – серверы, распределенные по элементарным процессорам ВС Cray T3E. Локальные серверы ОС обрабатывают запросы, специфичные для каждого ЭП ВС. Глобальные серверы обеспечивают общесистемные возможности такие, как управление процессами и файловые операции. Последние серверы размещаются в специальных системных ЭП и не дублируются в пределах ВС.

Система UNICOS/mk поддерживает масштабируемую архитектуру ввода-вывода Cray T3E. Она использует стандартные утилиты и команды ОС UNIX, следовательно она обеспечивает знакомую операционную среду для пользователей и администраторов.

Средства программирования Cray T3E

– Языки программирования и компиляторы: FORTRAN 90, C и C++, они используются для написания программ и их преобразования в эквивалентные объектные программы (на машинном языке).

– Пакет поддержки параллельного программирования MPT (Message Passing Toolkit) реализует взаимодействия между ветвями параллельной программы. Пакет включает широко применяемые интерфейсы передачи сообщений: MPI, MPI-2 и PVM.

– Отладчик (Cray Total View Debugger) используется для отладки прикладных параллельных программ на уровне исходного текста. Он позволяет пользователям отображать и анализировать информацию о параллельных процессах.

– Интерактивная среда (Cray Program Browser) применяется для отображения и редактирования файлов и прикладных программ.

– Обучающая система (MPP Apprentice) дает рекомендации по повышению производительности MPP-системы, отображает данные по производительности и интерпретирует их.

– Библиотеки оптимизированных параллельных прикладных программ.

6. Выводы

В результате многолетней работы фирмы Cray Research Inc. по развитию архитектуры средств обработки информации пройден путь от канонической конвейерной ВС до вычислительных систем с массовым параллелизмом. Последние системы с достаточной полнотой основываются на модели коллектива вычислителей (см. 3.1). Они по своей архитектуре вплотную подошли к распределенным ВС с программируемой структурой, архитектурно гибкие образцы которых были разработаны и построены еще в середине 70-х годов 20 века Отделом вычислительных систем Сибирского отделения АН СССР совместно с промышленными организациями (см. гл.7).