Лекция 10
Массово-параллельные вычислительные системы (MPP systems)

Ефимов Александр Владимирович
E-mail: alexandr.v.efimov@sibguti.ru

Курс «Архитектура вычислительных систем»
СибГУТИ, 2019

Массово-параллельные вычислительные системы

✓***Массово-параллельная ВС*** (Massively Parallel Processing System) – распределённая ВС, состоящая из большого количества вычислительных узлов.

✓Каждый узел является независимым вычислителем, содержащим как минимум один элементарный процессор, память и локальный коммутатор для связи с другими узлами системы

✓Термин «массово» подчёркивает наличие тысяч и даже миллионов элементарных процессоров в системе

Развитие ВС Cray

Сеймур Крей (Seymour Cray) возле Cray-1, 1974

April 6, 1972
Открытие Cray Research, Inc. (CRI) в Chippewa Falls, WI

1975
Ввод в эксплуатацию первой суперВС Cray-1™

1976
Установка Cray-1 (Los Alamos National Laboratory)
Заказ на Cray-1 (National Center for Atmospheric Research)

1977
Выпуск первого коммерчески доступного компилятора с автоматической векторизацией

1979
Объявление о выходе Cray-1S™

1981
Выход Cray-2™ (жидкостная система охлаждения)

Развитие ВС Cray

Сеймур Крей (Seymour Cray) возле Cray-1, 1974

1982
Cray X-MP™ (1-ый многопроцессорный суперкомпьютер)
Cray-1M™
1988
Cray Y-MP™ (1 GFLOPS)
1989
Cray Y-MP 2E™ (первая система с воздушным охлаждением)
1990
Cray XMS™ и Cray EL™
1991
Cray Y-MP 8™, Cray Y-MP EL™ и Cray Y-MP 4E™
Cray C90™ (1 GFLOPS процессор)
1992
Cray Y-MP M90™ и Cray S-MP™
Выпуск первого компилятора языка Fortran 90

Развитие ВС Cray

Сеймур Крей (Seymour Cray) возле Cray-1, 1974

1993

Cray EL92™ и Cray EL98™

Cray T3D™ (MPP System)

1994

Cray T90™ и Cray J90™

1995

Cray T3E™ (MPP, 1 TFLOPS)

Cray T94™

Заключен контракт с 1-ым российским клиентом – Росгидромет (Cray YMP-8E и сервер приложений Cray EL98)

1996

Silicon Graphics покупает Cray Research

2000

Cray SV1ex™

2001

Cray SX-6™

Alpha Linux (суперкластерная система)

Развитие ВС Cray

Сеймур Крей (Seymour Cray) возле Cray-1, 1974

2002

Cray X1™ (51 TFLOPS)

2004

Cray XD1™, Cray XT3™, Cray X1E™

2006

Cray XT4™, Cray XMT™

2007

Cray XT5™ (MPP system)

Cray XT5h™ (гибридный суперкомпьютер)

2008

Cray XT™ (переход барьера в 1 PFLOPS)

2009

Cray XT5, Cray XT6™

“Anyone can build a fast CPU. The trick is to build a fast system.”

Seymour Cray

Массово-параллельные вычислительные системы Cray

Сеймур Крей (Seymour Cray) возле Cray-1, 1974

2010

Cray XE6™ (следующее поколение MPP)

2012

Cray XK7™ (560 640 cores, 17.59 PFLOPS)

7 место top500 (июнь 2018)

2017

Cray XC50™ (361 760 cores, 19.59 PFLOPS)

6 место top500 (июнь 2019)

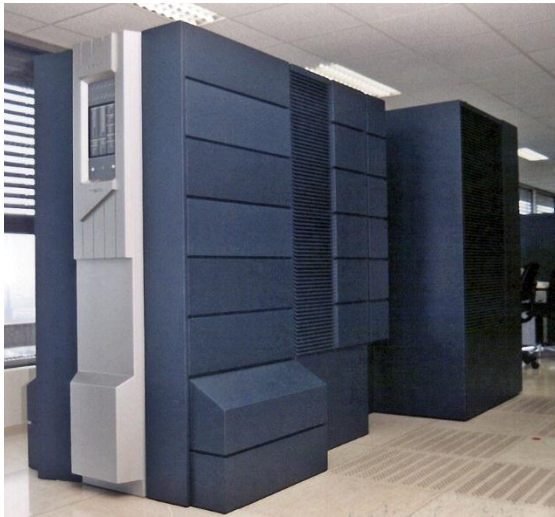
Вычислительная система Cray T3D

- Cray T3D – первая MPP-система корпорации Cray Research
- Допустимые количества ЭП в конфигурациях Cray T3D – 32 – 2048
- Диапазоны производительности и емкости памяти соответственно равны 5 – 300 GFLOPS и 512 Мбайт – 128 Гбайт
- Обычная конфигурация Cray T3D – 64-процессорная, она обеспечивала быстродействие, равное 10 GFLOPS
- Архитектура системы Cray T3D – MIMD,

  а сама ВС – распределенная
- Система Cray T3D работает под управлением

  хост-системы (Cray Y-MP и Cray C90)
- Монопрограммный режим

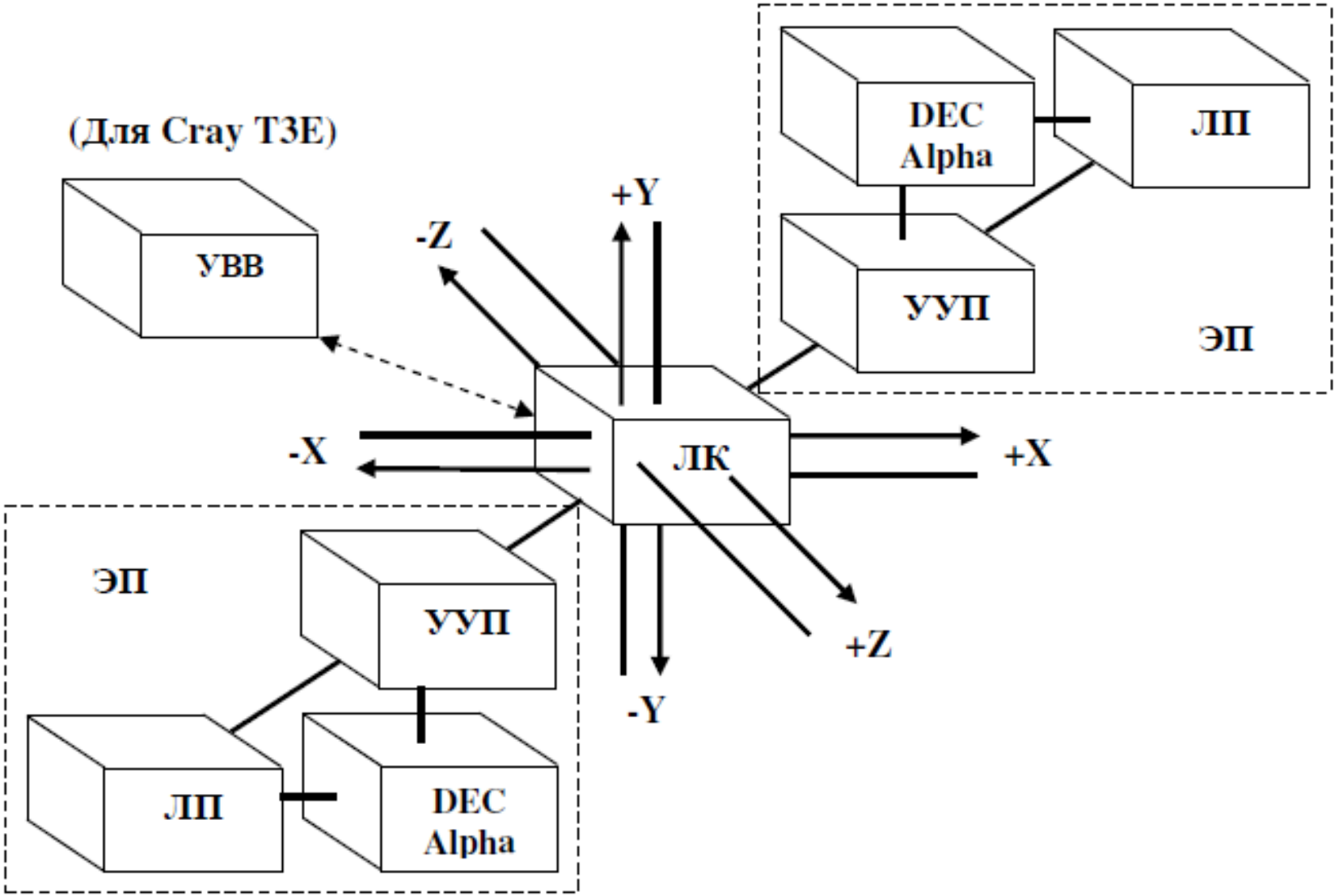
  функционирования



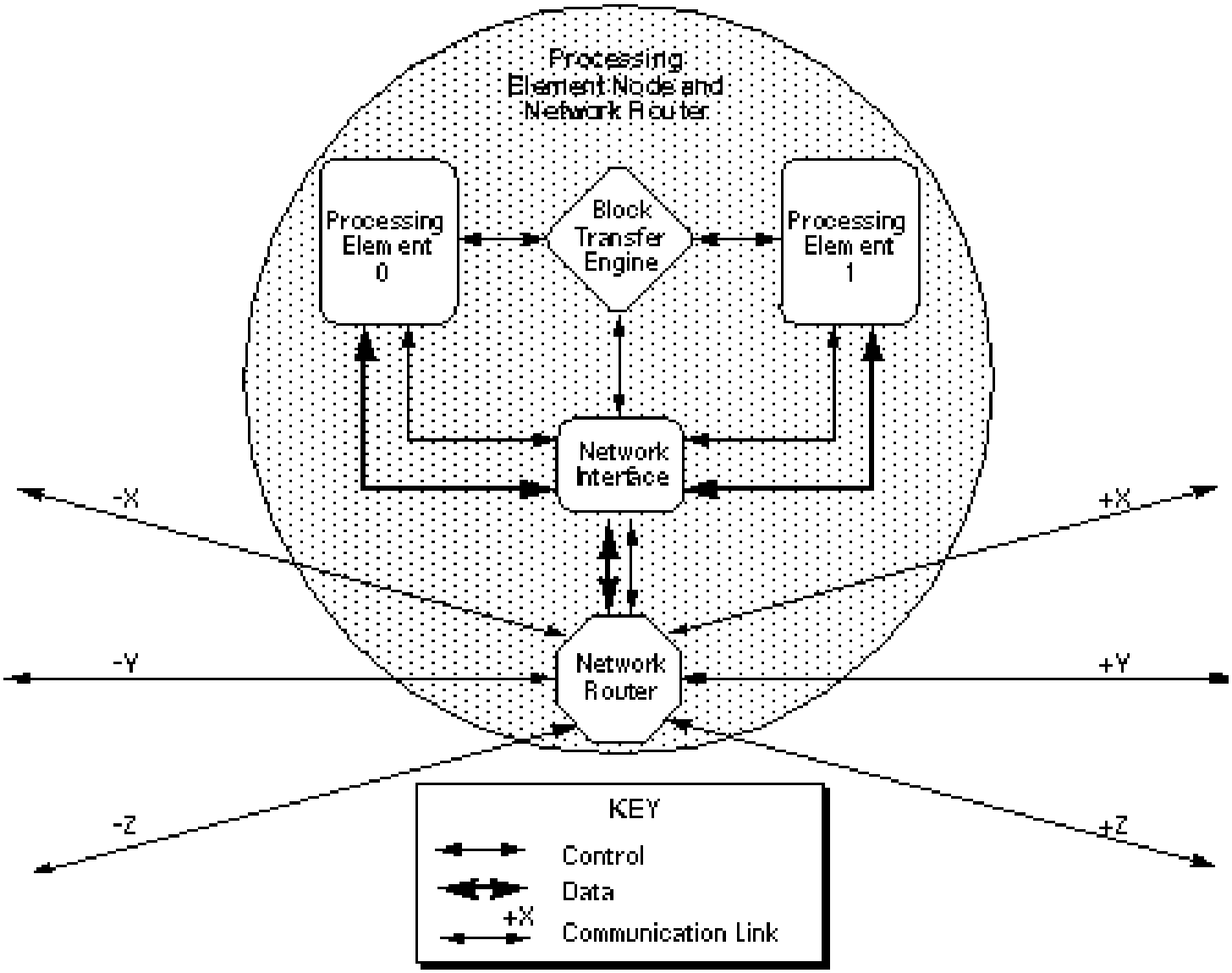
Вычислительный узел Cray T3D

- Все вычислительные узлы (ВУ) – однородные
- ВУ (Processing Element Node) системы включает в себя два одинаковых элементарных процессора (ЭП) и локальный коммутатор (ЛК)
- ЭП (Processing Element) представляется композицией из микропроцессора, локальной памяти (ЛП) и устройства управления памятью (УУП)
- Микропроцессор – это DEC 21064 Alpha chip (RISC-процессор типа Alpha фирмы DEC)
- ЛП – DRAM-память емкостью 16 – 64 Мбайт
- ЛК обеспечивает непосредственную связь ВУ с соседними узлами и представляет собой шестиполюсник. В состав ЛК входят: сетевой маршрутизатор, сетевой интерфейс и контроллер прямого доступа к памяти

Вычислительный узел Cray T3D



Вычислительный узел Cray T3D



Вычислительный узел Cray T3D

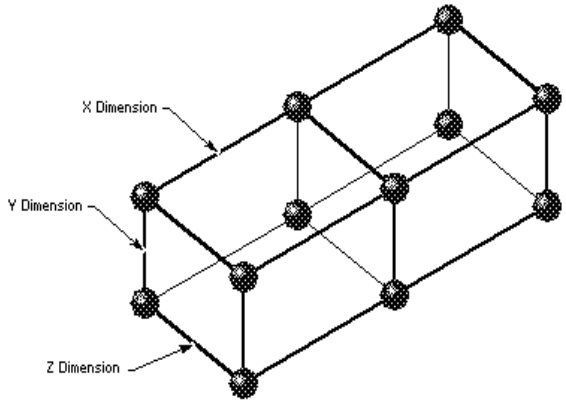
- *Сетевой маршрутизатор* (Network Router) ВУ – основной элемент формирования коммуникационной сети Cray T3D. Он способен работать с 3 парами двунаправленных межузловых связей, следовательно, позволяет создавать трехмерные структуры ВС

- *Сетевой интерфейс* (Network Interface) ВУ специальным образом кодирует информацию перед ее пересылкой по коммуникационной сети другому ВУ или в канал ввода-вывода

- *Контроллер прямого доступа к памяти* (Block Transfer Engine) осуществляет асинхронное перераспределение данных в пределах всей распределенной памяти ВС Cray T3D

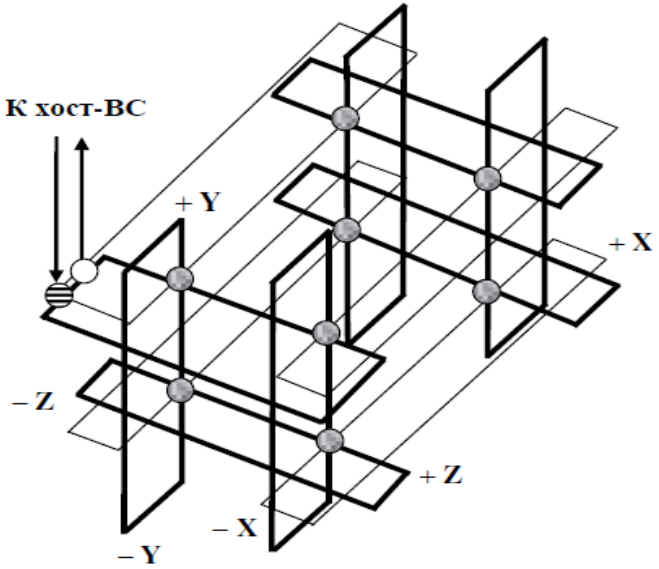
Коммуникационная сеть Cray T3D

- *Коммуникационная сеть* (Interconnect Network) системы Cray T3D предназначается для реализации обменов информацией между вычислительными узлами, а также между ВУ и каналами ввода-вывода



- Каждый ВУ связан с соседними по трем направлениям *X, Y и Z ,* причем по каждому направлению вершины образуют замкнутое кольцо

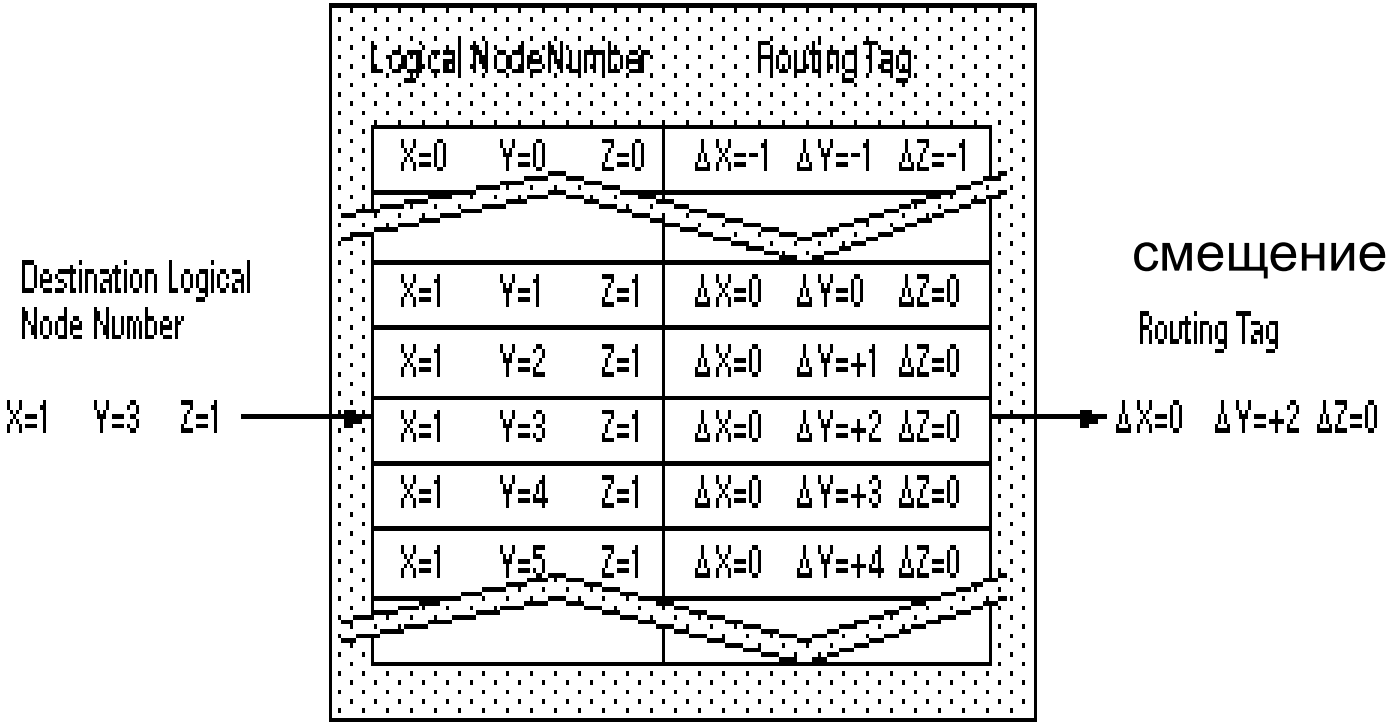
- Структура коммуникационной сети Cray T3D является циркулянтным графом с тремя образующими или *трехмерным (3D) тором*



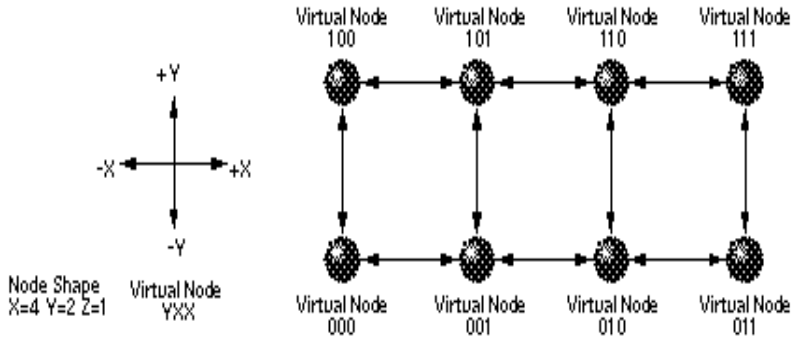
- Любой из адресов ВУ представляется трехкомпонентным вектором: (*x, y, z)*

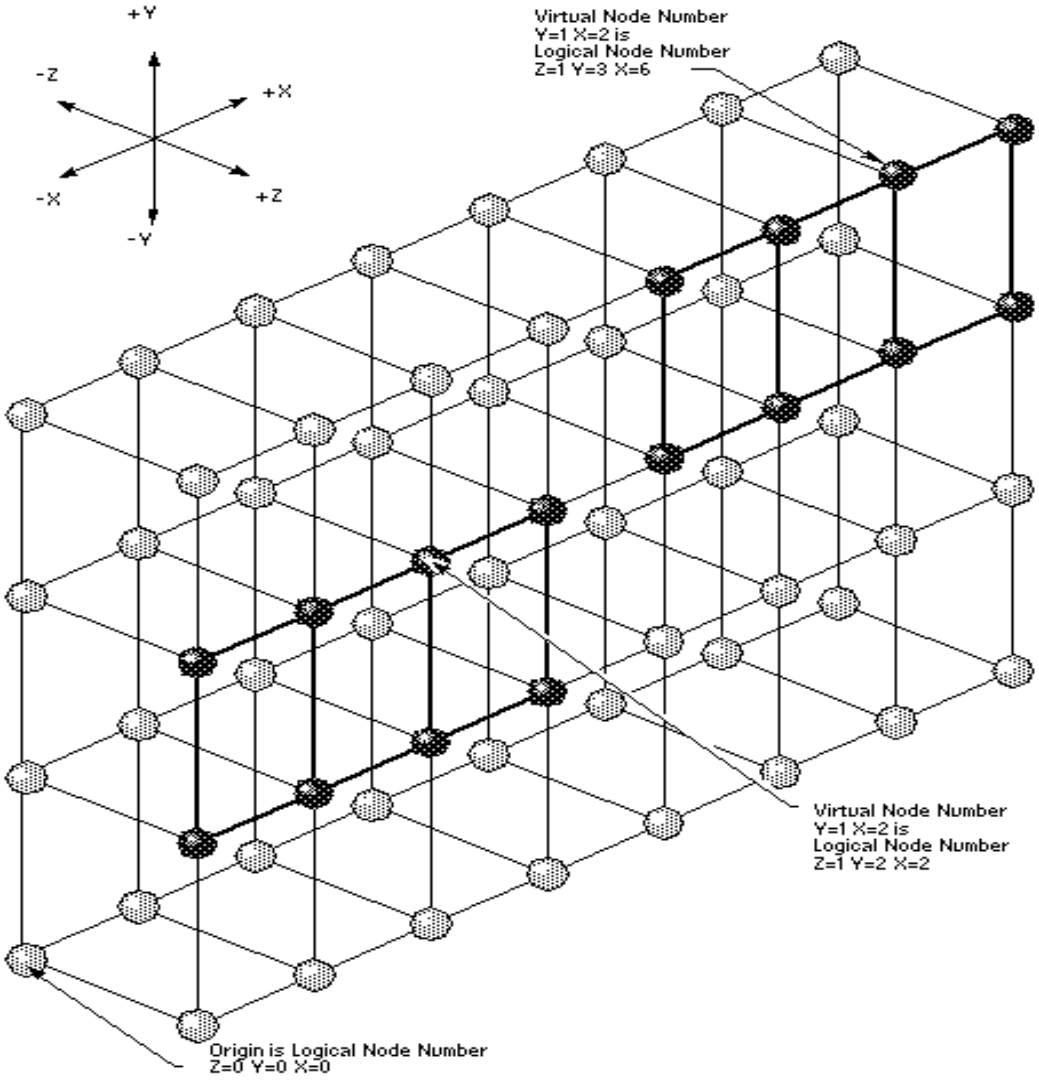
Коммуникационная сеть Cray T3D

- *Физический адрес* ВУ – это абсолютный неизменяемый номер узла в сети
- *Логический* и *виртуальный адреса* ВУ являются относительными адресами (относительно абсолютного)



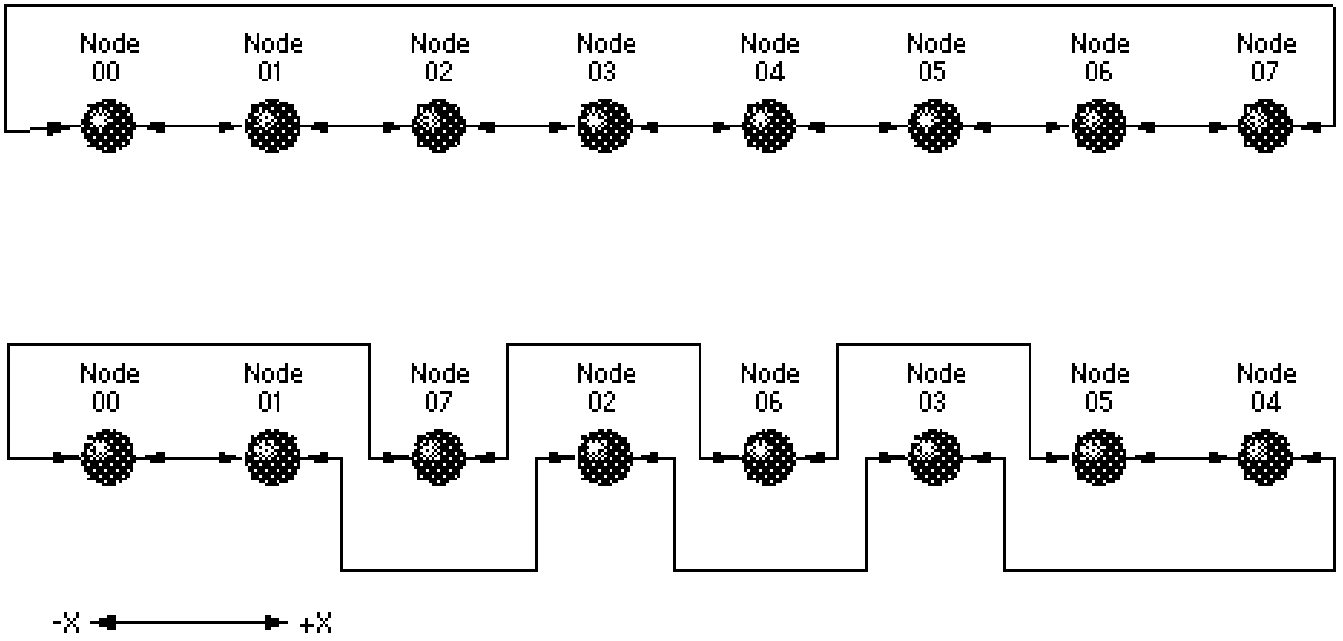
Коммуникационная сеть Cray T3D





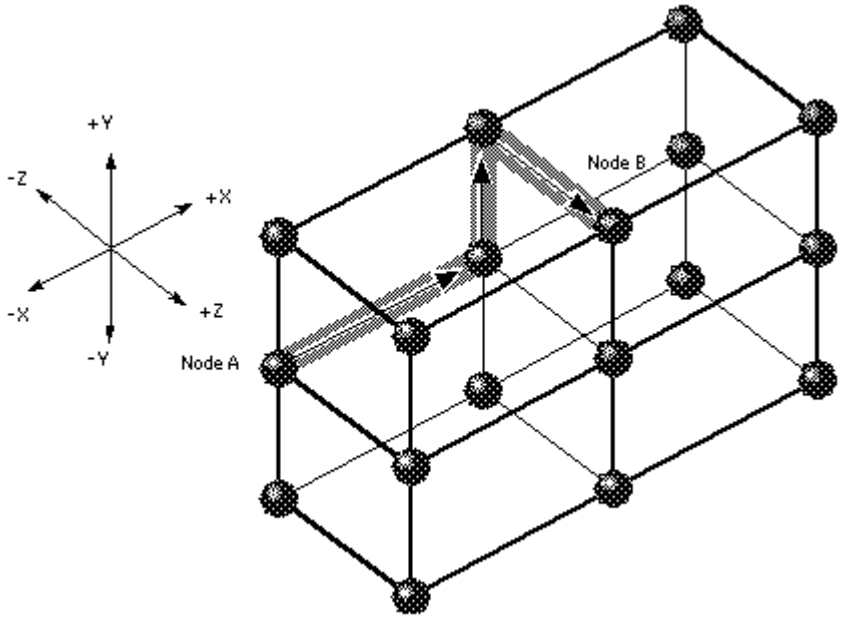
Коммуникационная сеть Cray T3D

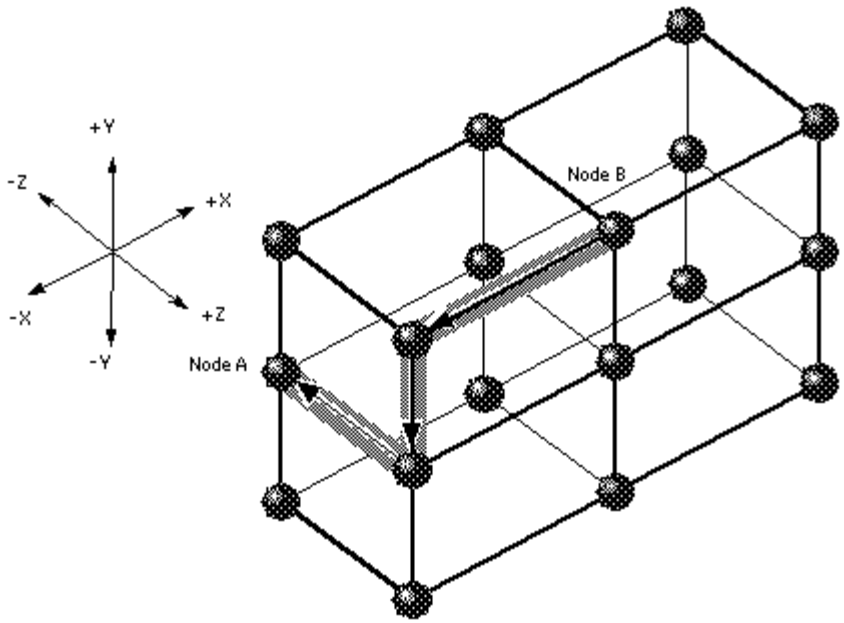
- *Чередование* – физическое размещение ВУ так, чтобы максимальное расстояние между узлами было минимальным

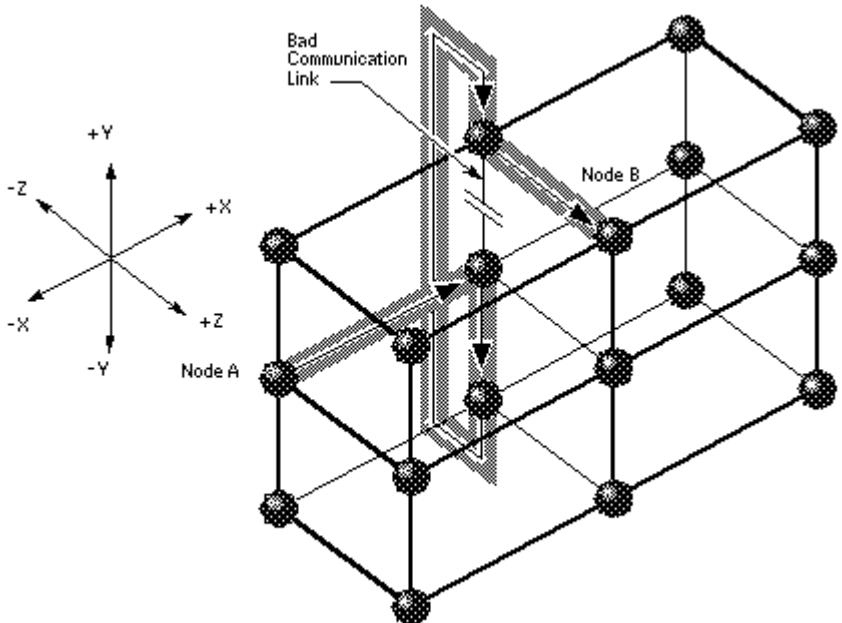


- Чередование организовано по осям X и Z. Это минимизирует длину физических коммуникационных связей в Cray T3D.

Коммуникационная сеть Cray T3D



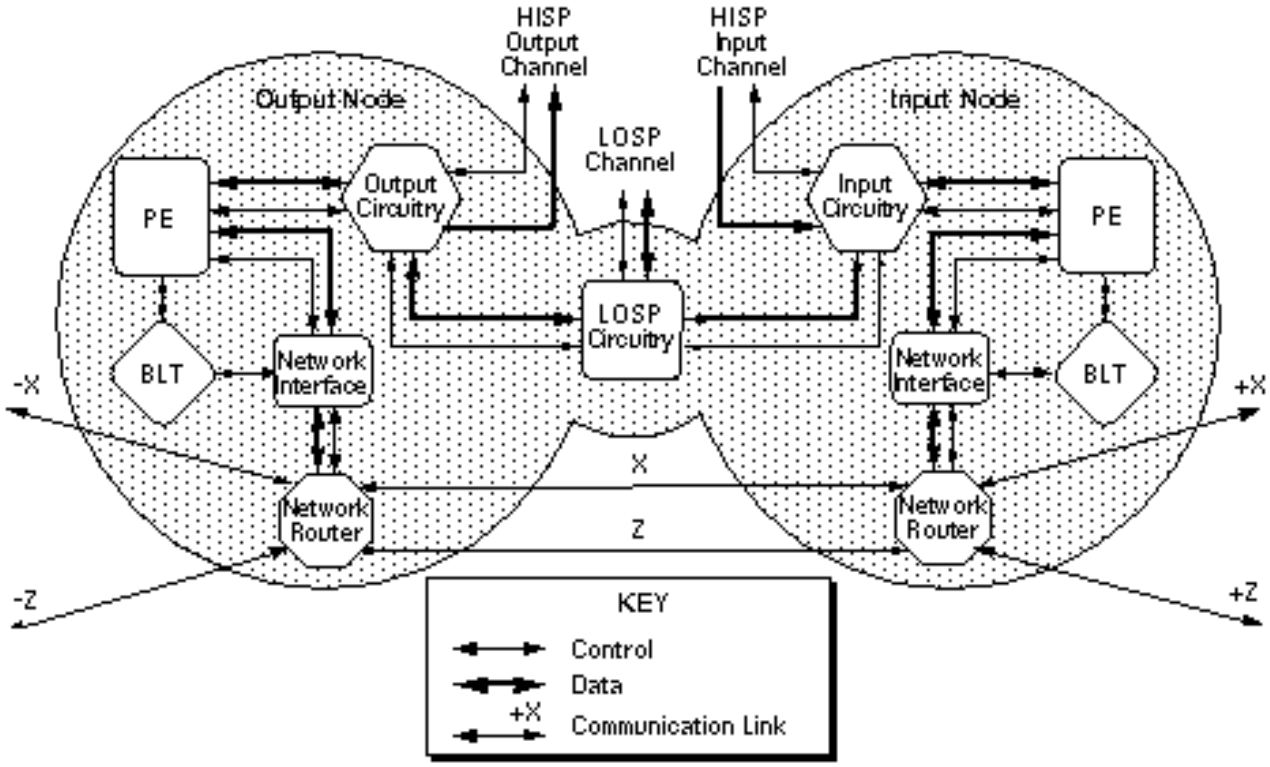




Каналы ввода-вывода Cray T3D

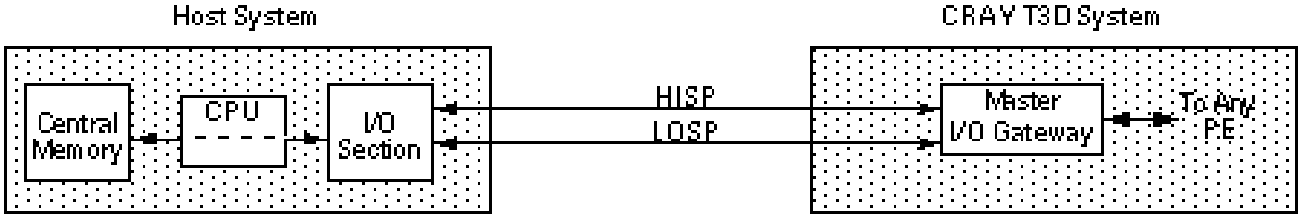
- *Каналы ввода-вывода* (Input/Output Gateways) предназначаются для обмена информацией между Cray T3D и управляющей системой (Host System) или кластером ввода-вывода Input/Output Cluster)
- Канал ввода-вывода Cray T3D представляется композицией из *узлов ввода* (Input node) и *вывода* (Output node) и *низкоскоростного устройства передачи запросов и ответов* (LOSP circuitry).

low-speed (LOSP, 6 Mbytes/s) and high-speed (HISP, 200 Mbytes/s) channel connectors

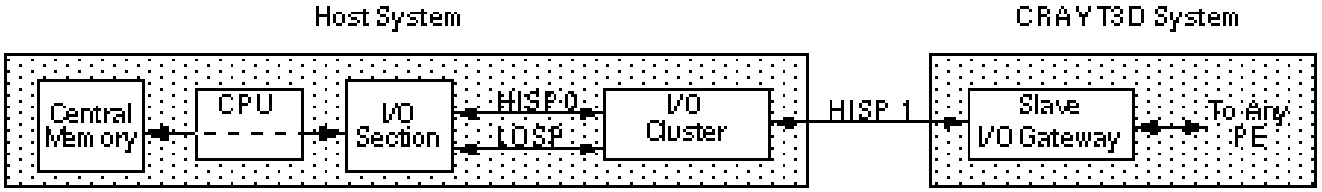


Каналы ввода-вывода Cray T3D

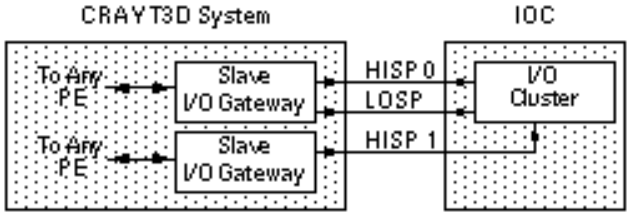
- Конфигурация 1 соединяет главный (master) канал ввода-вывода с хост-системой через HISP-канал



- Конфигурация 2 соединяет подчинённый (slave) канал ввода-вывода с кластером ввода-вывода, который в свою очередь соединен с ЦП хост-системы



- Конфигурация 3 соединяет подчинённый (slave) канал ввода-вывода с кластером ввода-вывода



Средства синхронизации Cray T3D

- Генератор тактовых импульсов (Clock), который посылает импульсы одновременно и в вычислительные узлы, и в узлы ввода и вывода
- Генератор работает на частоте 150 МГц
- Аппаратурная реализация механизмов синхронизации «барьер» и «эврика» (Barrier/Eureka)

Вычислительная система Cray T3E

- Количество элементарных процессоров в ВС – 16 – 2048
- Диапазон производительности – 14,4 GFLOPS – 2,76 TFLOPS
- Ёмкость памяти – 1 Гбайт – 1 Тбайт
- Цена 128-процессорной конфигурации Cray T3E – 3 – 4 млн. долларов
- Несколько модификаций ВС: Cray T3E, Cray T3E-900, Cray T3E-1200, Cray T3E-1200E, Cray T3E-1350, с тактовыми частотами от 300 до 675 МГц
- Барьер производительности 1 TFLOPS, 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

Вычислительная система Cray T3E

- Cray T3E относится к классу MIMD
- Две архитектурные особенности Cray T3E:

  мультипрограммирование – возможность одновременной реализации нескольких параллельных программ на различных подсистемах;

  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).
- Функциональные структуры систем Cray T3D и Cray T3E на макроуровне полностью идентичны.

Вычислительный узел Cray T3E

- Микропроцессор семейства DEC 21164 Alpha

  Спецификация микропроцессора для модификации ВС Cray T3E-1350:

  – разрядность операндов: 32 или 64;

  – тактовая частота – 675 МГц;

  – производительность: 1350 MFLOPS (2 операции с плавающей запятой за такт), 2700 MIPS (4 инструкции за такт);

  – быстродействие канала к памяти – 1200 Мбайт/с.
- ЭП располагает своей локальной памятью, ёмкость которой варьируется в пределах от 64 до 512 Мбайт (в зависимости от модификации ВС)

  В Cray T3E-1350 локальная память ЭП имеет емкость в пределах: 250 – 512 Мбайт и формируется из 64-мегабайтных DRAM-схем.
- В вычислительном узле ВС Cray T3E в отличие от ВУ системы Cray T3D, предусмотрена специальная связь для непосредственного подключения устройств ввода-вывода (УВВ) информации

Коммуникационная сеть Cray T3E

- Сеть межузловых связей Cray T3E – трехмерный тор с двунаправленными каналами
- Крайне малое время задержки при пересылке сообщений (обладает низкой латентностью, Latency) и характеризуется значительной шириной полосы пропускания
- Модификация Cray T3E-1350 имеет быстродействие 650 Мбайт/с в каждом из двух направлений передачи информации

Каналы ввода-вывода Cray T3E

- Реализована возможность осуществлять обмен информацией с внешней средой через множество каналов ввода-вывода
- Каналы ввода-вывода Cray T3E интегрированы в трехмерную коммуникационную сеть так, что их количество всегда пропорционально числу элементарных процессоров в любой конфигурации системы
- Все каналы ввода-вывода (и 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 Гбайт/с; максимальная полоса пропускания любого интерфейса гигакольца составляет 500 Мбайт/с

Конструктивные особенности системы Cray T3E

- Два варианта корпусов: с воздушным и жидкостным охлаждением
- В системах с воздушным или жидкостным охлаждением масштабирование осуществляется на величину, кратную 4 или 8 ЭП, соответственно
- 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
- Максимальная конфигурация ВС Cray T3E размещается в 8 корпусах и насчитывает 2176 элементарных процессоров, из которых число основных ЭП равно 2048

Программное обеспечение Cray T3E

- *Операционная система* UNICOS/mk, разработанная для ВС Cray T3E, по сути является распределенной и масштабируемой версией UNICOS
- *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

Сверхвысокопроизводительные вычислительные системы семейства Cray X

Особенности архитектуры системы Cray X1

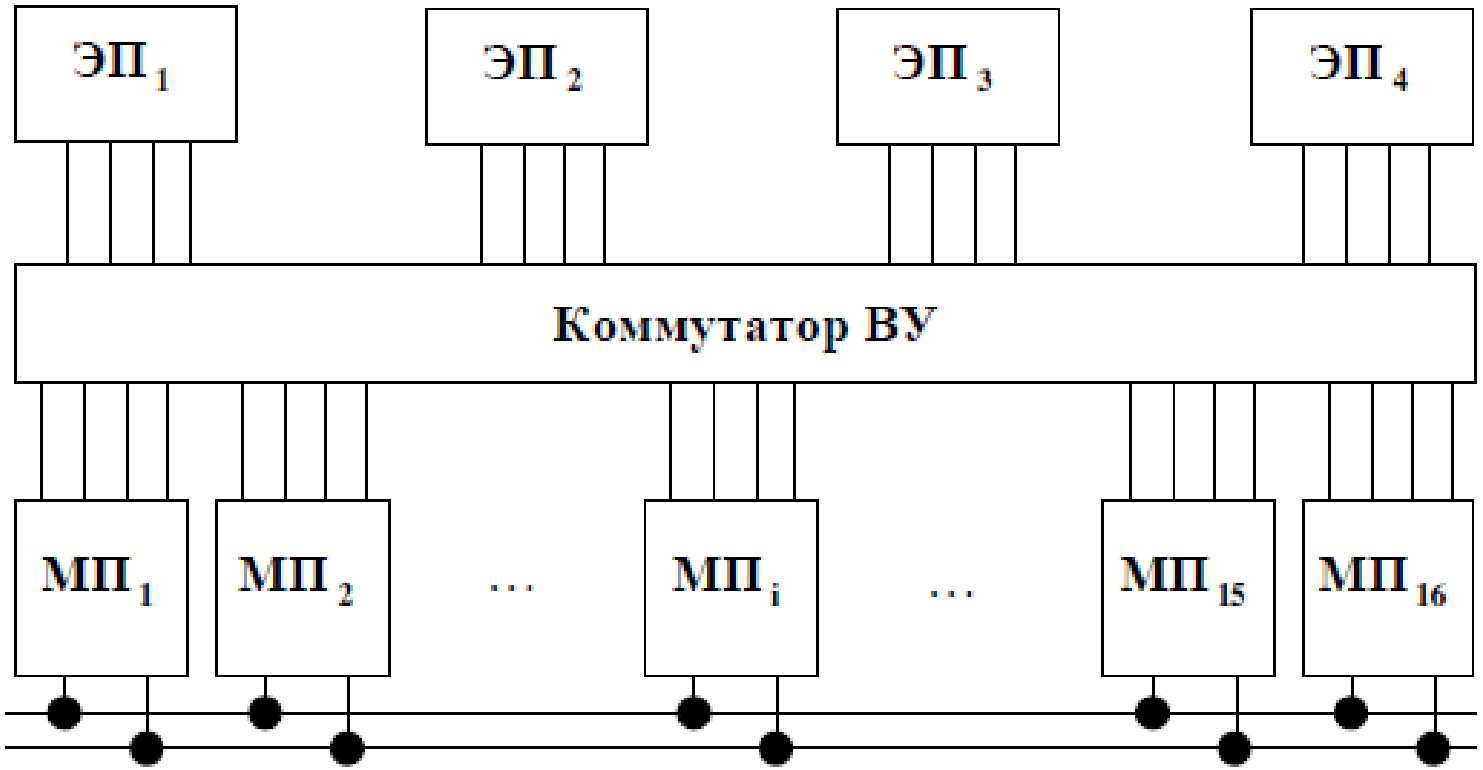
- Максимальная конфигурация ВС Cray X1 состоит из 4096 ЭП (или 49152 вычислителей, среди которых 32768 векторных конвейеров и 16384 скалярных блоков)
- Производительность – 52,4 TFLOPS
- Память – 16 – 64 Тбайта
- Вес такой конфигурации ВС составляет примерно 230 т (при воздушном охлаждении) или 170 т (при жидком хладагенте)
- Система Cray X1 была официально анонсирована в ноябре 2002 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).

Особенности архитектуры системы Cray X1

- Cray X1 – это MIMD-система с общей распределенной памятью (Distributed Shared Memory)
- ВС основывается на тороидальной топологии и имеет широкую полосу пропускания и низкую латентность
- 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 Гбайт – 64 Тбайт
- В систему Cray X1 вложен новейший набор команд, активные исследования по которому велись в корпорации Cray в течение 10 лет
- Незначительная сложность управления
- Небольшое энергопотребление, соотнесенное к одной операции в секунду

Вычислительный узел Cray X1

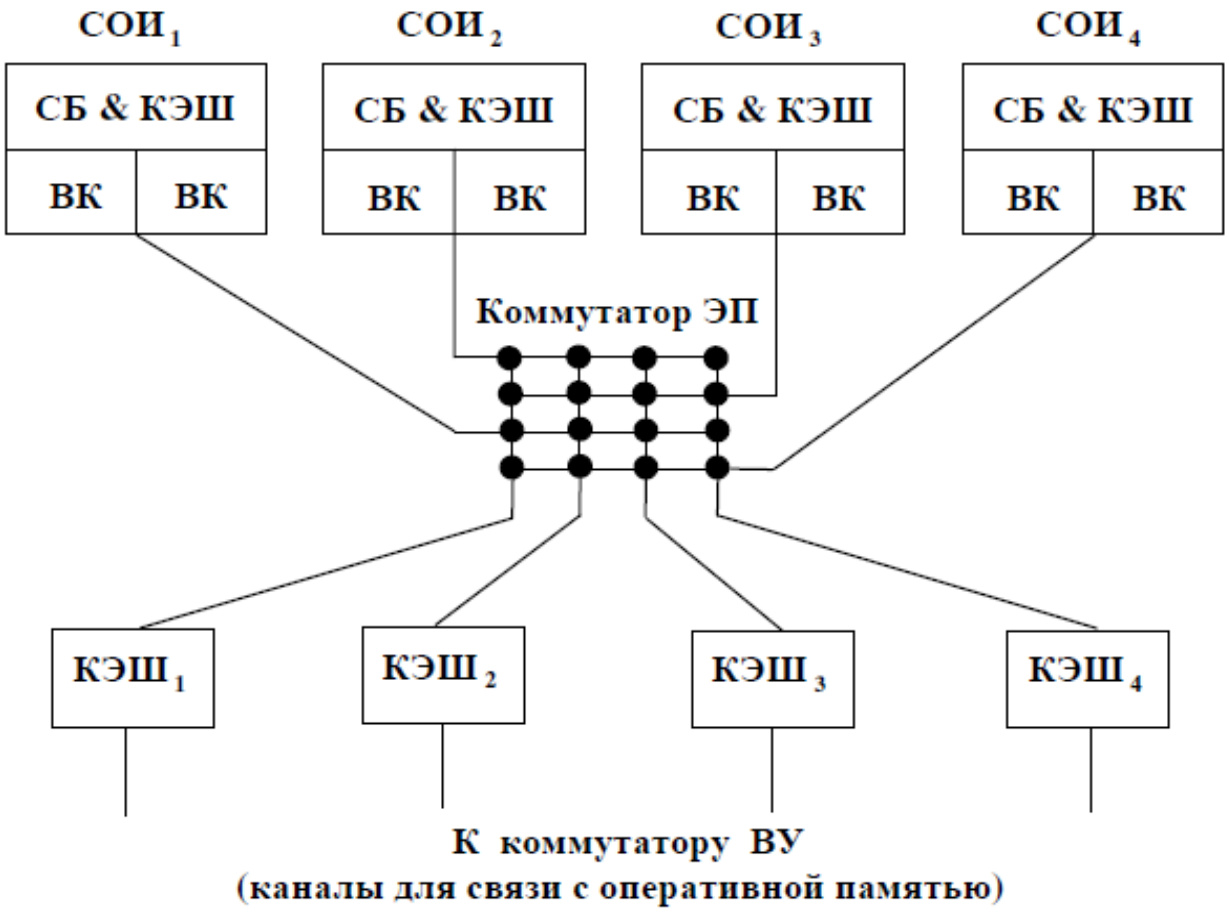
- Система Cray X1 может иметь в своем составе от 2 до 1024 однородных вычислительных узлов
- Каждый ВУ содержит 4 ЭП и распределенную общедоступную оперативную память
- Взаимодействие между процессорами и оперативной памятью в узле осуществляется при помощи коммутатора



Вычислительный узел Cray X1

- Каждый ЭП – это специально спроектированный конвейерный (или векторный) процессор, обладающий производительность 12,8 GFLOPS (при обработке 64-разрядных операндов)
- ЭП относится к типу мультипотоковых процессоров (MSP – Multi-Streaming Processors), если придерживаться терминологии компании Cray

СОИ – секция обработки информации
ВК – векторный конвейер
СБ – скалярный блок



Вычислительный узел Cray X1

- 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.

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

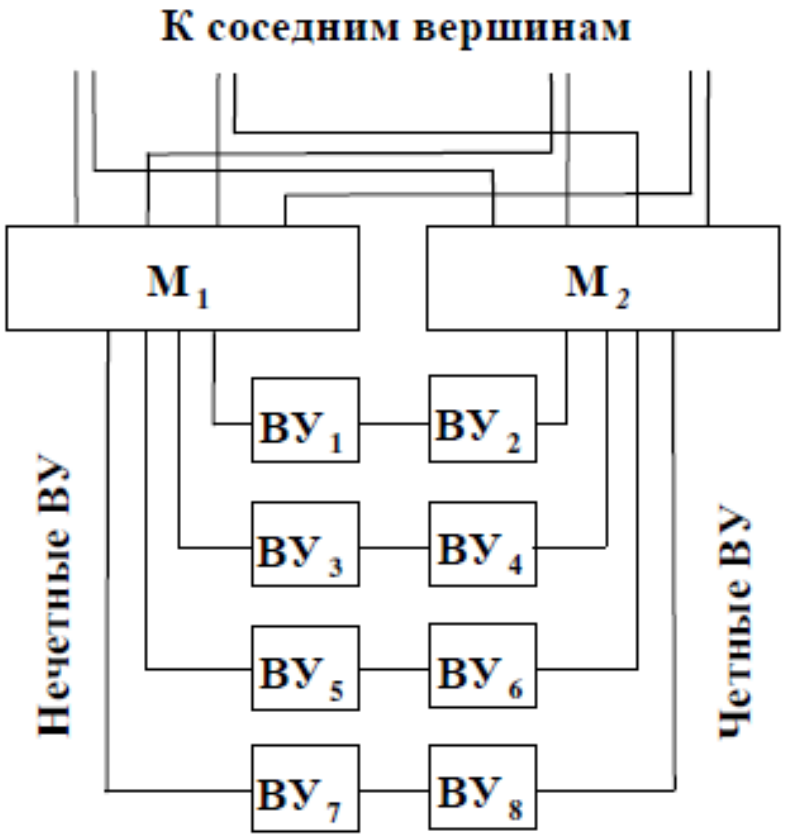
Вычислительный узел Cray X1

- Память любого узла 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
- Пропускная способность канала между ЭП и оперативной памятью в вычислительном узле составляет 34,1 Гбайт/с
- Оперативная память вычислительного узла доступна для других ВУ системы; этот доступ реализуется при помощи специальных маршрутизаторов
- Все модули памяти в системе Cray X1 физически распределены по вычислительным узлам (и, следовательно, по элементарным процессорам), но логически они доступны каждому ЭП.

Оперативная память ВС Cray X1 является и распределенной, и общей.

Коммуникационная сеть Cray X1

- Модифицированный двумерный тор
- В качестве вершины двумерного тора используется композиция из четырех пар вычислительных узлов и двух маршрутизаторов (M_1, M_2), работающих на 4 внешних связи
- Индекс в обозначении ВУ не является физическим номером вычислительного узла в пределах ВС, он дает лишь информацию о четности или нечетности номера
- Маршрутизаторы M_1 и M_2 обеспечивают два параллельных канала связи данной вершины с соседними вершинами в 2D-торе



Коммуникационная сеть Cray X1

Структурные уровни Cray X1	Вычислительные элементы структуры
Макроуровень – двумерный тор	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 Диаметр сети межузловых связей равен 2.	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	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	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 $КЭШ_1$ – $КЭШ_4$.

Средства ввода-вывода Cray X1

- Средства ввода-вывода информации ВС Cray X1 распределены по ее вычислительным узлам
- Каждый ВУ располагает 4 каналами ввода-вывода
- Пиковая полоса пропускания одного канала ввода-вывода составляет 1,2 Гбайт/с
- Каналы ввода-вывода Cray X1 служат для подключения дисков и других периферийных устройств

Конструкция системы Cray X1

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

Программное обеспечение системы Cray X1

- Операционная система UNICOS/mp
- Компиляторы с языков Фортран и Си, включающие возможности автоматической векторизации и распараллеливания
- Специальные оптимизированные библиотеки
- Интерактивный отладчик
- Средства для анализа производительности
- Среда программирования поддерживается специальным сервером – CPES (Cray Programming Environment Server)

Области применения системы Cray X1

- Фундаментальные научные исследования, вычислительная математика, физика, химия, астрономия (включая астрофизику), биология, науки о Земле
- Биотехнические исследования, биоинформатика и моделирование биологических процессов
- Медицина и фармакология; виртуальная хирургия; создание лекарств
- Предсказание естественной пандемии и локальных эпидемий
- Экология, окружающая среда, климат, погода
- Аэрокосмические исследования и индустрия
- Экономика, моделирование национальной экономики и инвестиций
- Оборона и военные приложения
- Промышленность, машиностроение, энергетика, металлургия

Анализ архитектуры вычислительной системы Cray X1

- Параллельность выполнения операций
- Программируемость структуры
- Конструктивная однородность
- Технико-экономические соображения и необходимость достижения высокой производительности и архитектурного совершенства с неизбежностью заставили главного разработчика конвейерных ВС – Cray стать на платформу распределенных средств обработки информации, полностью основанных на модели коллектива вычислителей
- Архитектурным совершенством начала 21 века являются сверхвысокопроизводительные вычислительные системы семейства Cray X

Архитектура Cray XC50

Piz Daint - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100

Site:	Swiss National Supercomputing Centre (CSCS)
System URL:	http://www.cscs.ch/computers/piz_daint_piz_dora/index.html
Manufacturer:	Cray Inc.
Cores:	387,872
Memory:	365,056 GB
Processor:	Xeon E5-2690v3 12C 2.6GHz
Interconnect:	Aries interconnect
Performance	
Linpack Performance (Rmax)	21,230 TFlop/s
Theoretical Peak (Rpeak)	27,154.3 TFlop/s
Nmax	3,743,232
HPCG [TFlop/s]	496.978
Power Consumption	
Power:	2,384.24 kW (Submitted)
Power Measurement Level:	3
Measured Cores:	387,872
Software	
Operating System:	Cray Linux Environment

Архитектура Cray XC50

Figure 2. A hybrid Cray XC50 module.

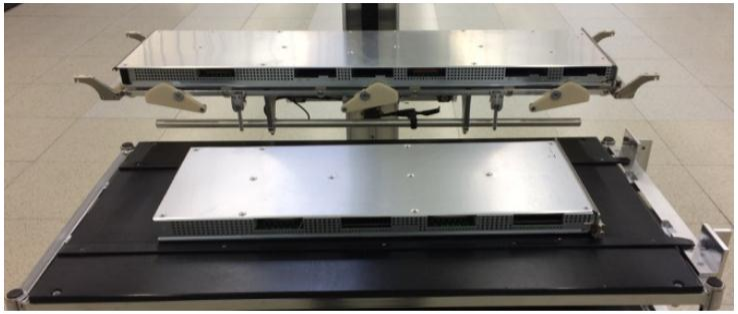
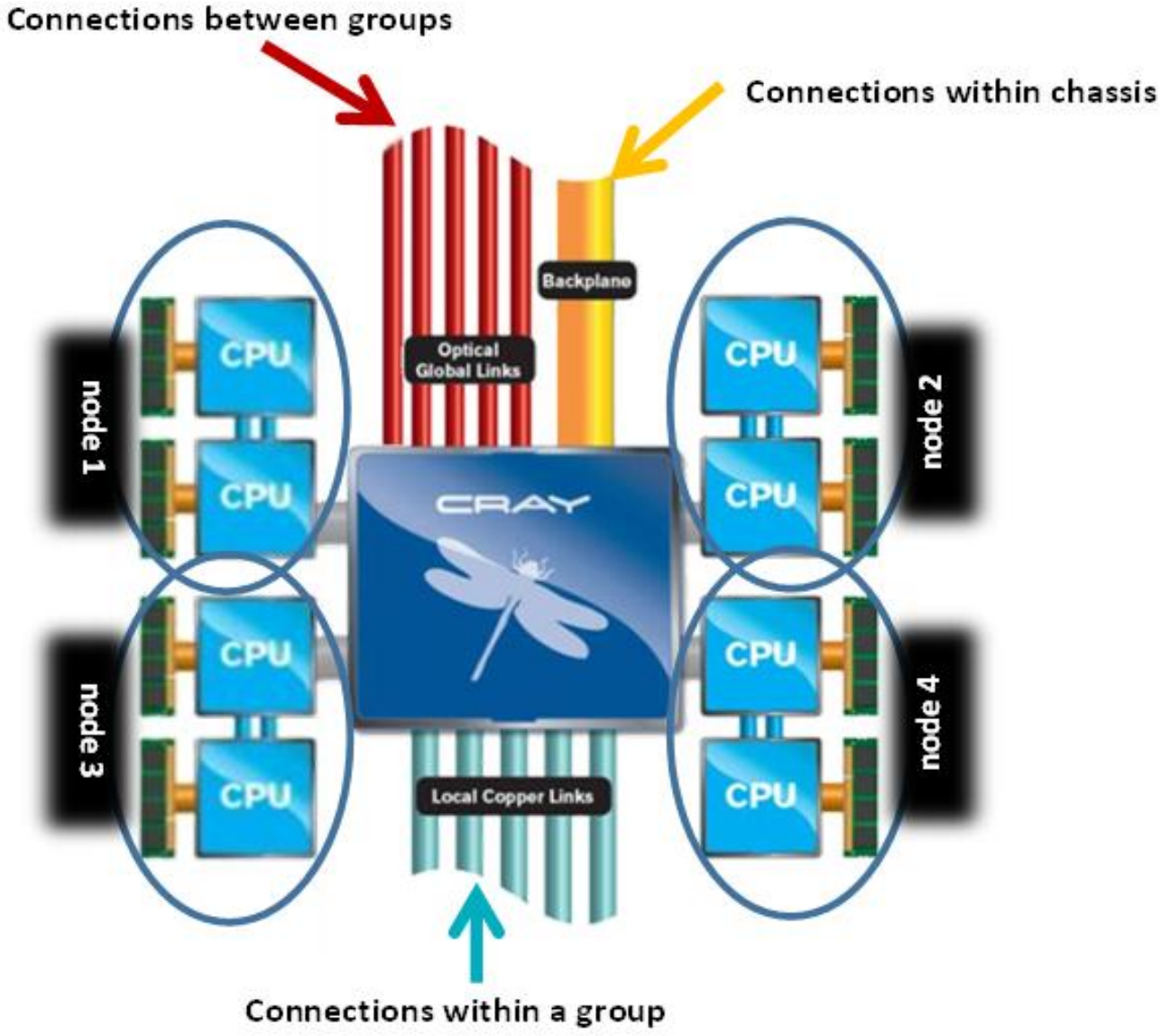


Figure 1. Comparison of Cray XC40 (front) and XC50 (rear) compute modules.

Figure 3. The front of a Cray XC50 service module.

Архитектура вычислительного модуля Cray XC50



Интерконнект Cray XC50

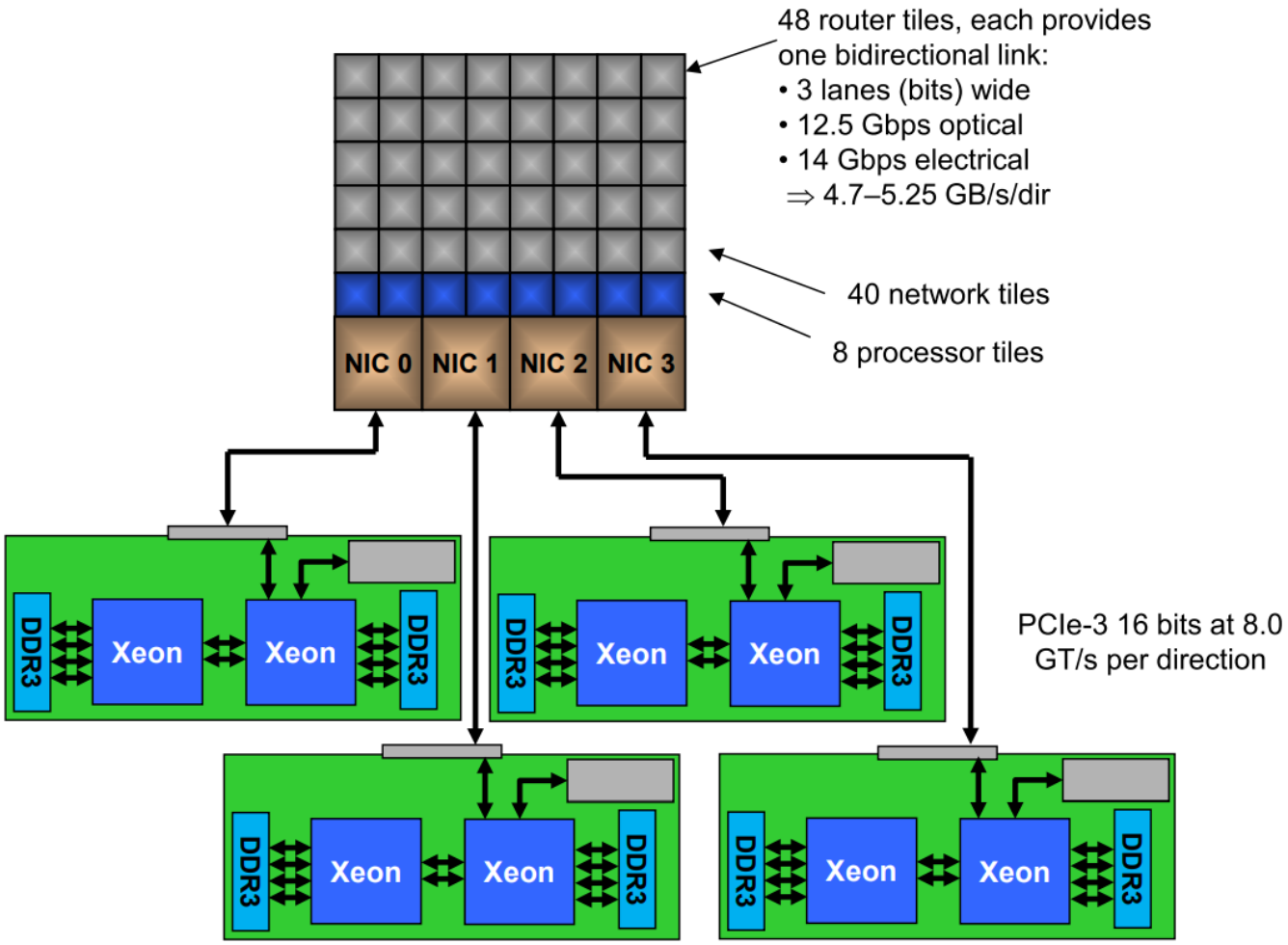


Figure 3: *A single Aries system-on-a-chip device provides network connectivity for the four nodes on a Cray XC blade.*

Интерконнект Cray XC50

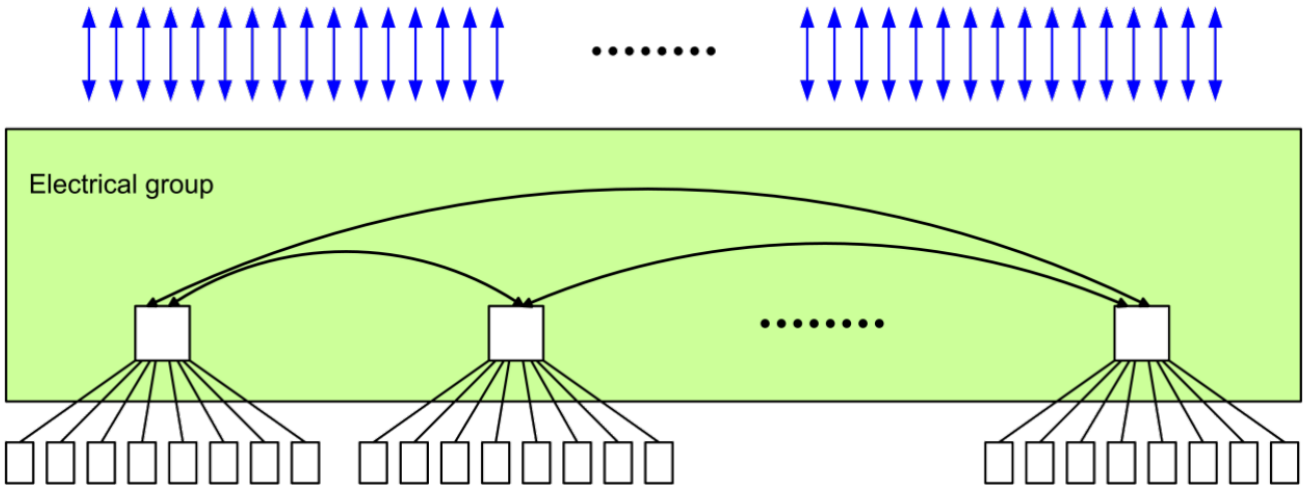


Figure 1: Short electrical links connect both the NICs to the routers and the routers in each group of a Dragonfly network. The routers in a group pool their global links (shown in blue).

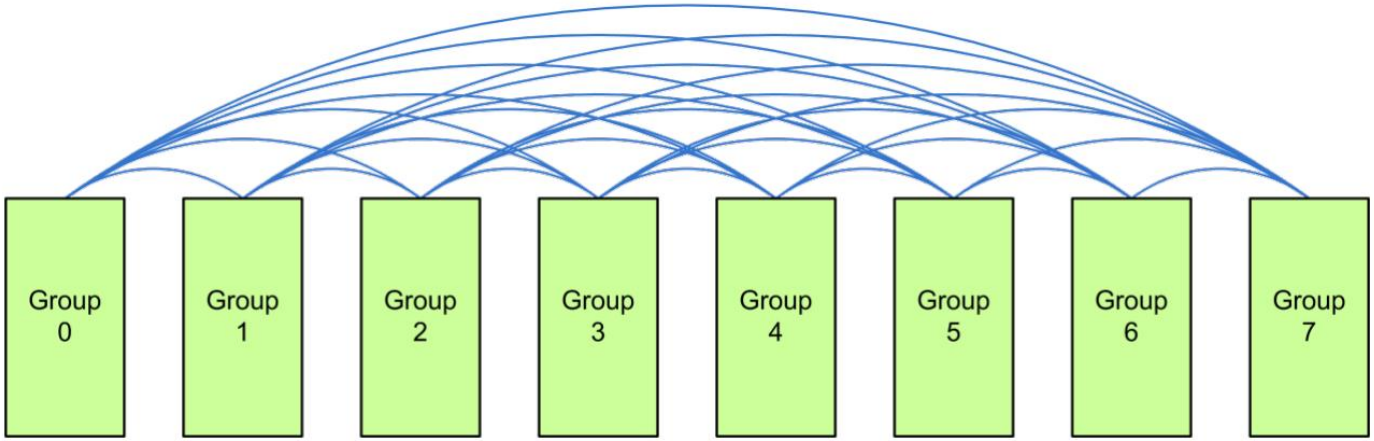
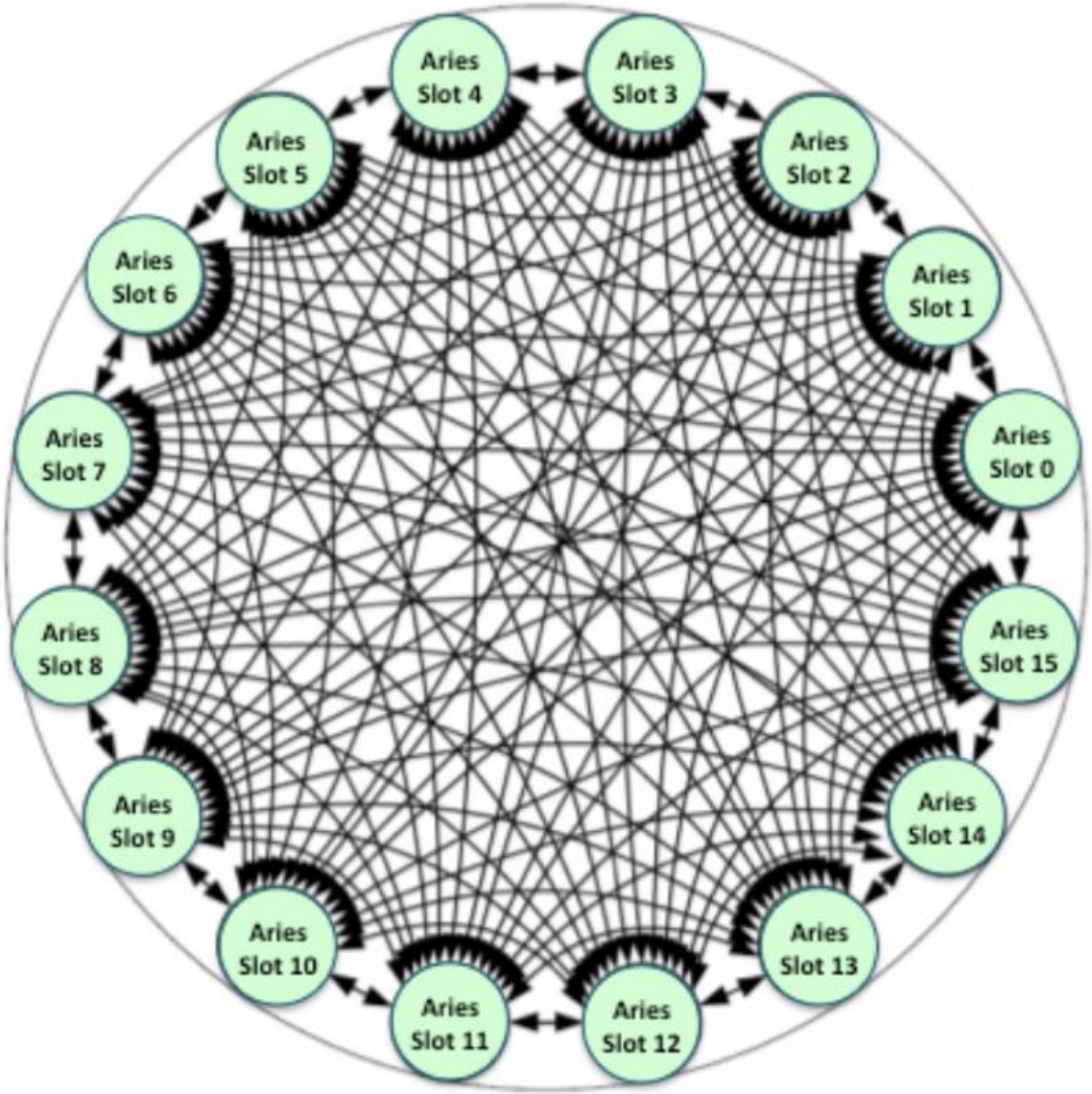


Figure 2: Global links connect Dragonfly groups together — these links are optical in a large system.

Топология макроструктуры Dragonfly Cray XC50



Литература

Хорошевский В.Г. Архитектура вычислительных систем. Учебное пособие. – М.: МГТУ им. Н.Э. Баумана, 2005; 2-е издание, 2008.

Хорошевский В.Г. Инженерные анализ функционирования вычислительных машин и систем. – М.: “Радио и связь”, 1987.