Суперкомпьютерные технологии.

Лекционно-практический курс
для студентов 5 курса факультета ВМиК МГУ
октябрь – декабрь 2011 г.

Лекция 1

7 октября 2011 г.

Н.Н. Попова

доцент кафедры АСВК

popova@cs.msu.su

План лекции

- Учебный план курса
- Содержание и требования к выполнению практических заданий
- Обзор технологий ПП, необходимых для выполнения заданий
- Архитектура ВС **pSeries 690 Регатта**
- Архитектура и программное обеспечение ВС BlueGene/P

Учебный план

Лекции, семинары, выполнение практических заданий

- Итоговая оценка: зачет
- Форма отчетности: отчет в электронном виде.
- Лекции - Обзор суперкомпьютерных систем и программных технологий.
  с 7 октября по 28 октября
- Семинарские занятия
  с 10 октября 2011 г.
- Выполнение практических заданий на высокопроизводительных системах:
  - IBM pSeries 690 Регатта (www.regatta.cs.msu.su)
  - IBM BlueGene/P (hpc.cmc.msu.ru)
  - «Ломоносов» (parallel.ru/cluster)

Список TOP500 (Июнь, 2011 г.)

12	Forschungszentrum Juelich (FZJ) Germany	JUGENE - Blue Gene/P Solution / 2009 IBM	294912	825.50	1002.70	2268.00
13	Moscow State University - Research Computing Center Russia	Lomonosov - T-Platforms T-Blade2/1.1, Xeon X5570/X5670 2.93 GHz, Nvidia 2070 GPU, Infiniband QDR / 2011 T-Platforms	33072	674.11	1373.06	

Содержание курса

- Лекции - **Обзор суперкомпьютерных систем и технологий параллельного программирования.**

Тема 1. **7 октября** : «Архитектура и ПО ВС Регатта и BlueGene/P» Н.Н.Попова

Тема 2. **14 октября**: «Архитектура и программное обеспечение суперкомпьютера «Ломоносов»» А.Корж

Тема 3. **21 октября**: «Технология OpenMP, гибридное MPI/OpenMP программированине» В.А.Бахтин

28 октября: «Технология OpenMP, гибридное MPI/OpenMP программированине» Продолжение. В.А.Бахтин

- Семинарские занятия – **обсуждение и сдача практических заданий**
  - с 10 октября 2011 г.
  - см. расписание компьютерных классов

Примерный план семинарских занятий

- 10 - 22 октября – выдача заданий, логинов, Регатта
- 24 октября – 2 ноября – выполнение заданий на BlueGene/P
- 2 ноября – 30 ноября – «Ломоносов»
- 1 – 20 декабря– оформление и сдача отчетов

Практические задания.

Базовое задание.

- **Тема**: Исследование эффективности решения систем линейных уравнений Ax = b на параллельных ВС.
- **Дано**: матрица A, правая часть – вектор b
- **Требуется:**

  на основе представленной MPI-реализации параллельной программы провести оптимизацию параллельного алгоритма и исследовать его эффективность
- **Варианты методов**: метод сопряженных градиентов, метод сопряженных градиентов с предобусловливателем Якоби, метод Якоби, метод Гаусса-Зейделя, метод Гаусса
- Для каждого из вариантов базовая параллельная программа задается при выдаче задания.
- **Требование к зачету**: реализация задания нв вычислительных системах IBM pSeries690 Регатта, BlueGene/P и «Ломоносов»
- **Форма отчетности** по курсу – **отчет**, представленный в электронном виде, с контрольной сдачей отчета преподавателям. Вместе с отчетом должны быть представлены тексты параллельных программ

Информационные ресурсы по заданию

- Методические материалы, инструкции:

  http://angel.cs.msu.su/~popova
- Исходные тексты, методические материалы:

  regatta.cs.msu.su/~popova/SuperComp2011

  assignment1.tar.gz, assignment2.tar.gz, …

Рекомендации по выполнению заданий

- Оптимизация программ с использование оптимизирующих возможностей компиляторов (анализ и настройка соответствующих опций компиляторов).
- Оптимизация с использованием расширенного набора MPI-функций (использование асинхронных передач, коллективных операций обмена, совмещенных операций и др.)
- Гибридное MPI/OpenMP программирование с использованием директив OpenMP в MPI-программах.

Также:

- Модификация параллельного алгоритма
- Использование библиотечных функций (BLAS, ESSL, PESSL,MKL, Lapack)
- Разработка параллельной программы

Исследование эффективности

- Ускорение S
- Эффективность η
- Пусть: T_1 - время выполнения наилучшего последовательного алгоритма,

  T_n – время выполнения параллельной программы на n процессорах

$$S = \frac{T_1}{T_n}$$

$$\eta = \frac{T_1}{nT_n} = \frac{S}{n}$$

Графическое представление ускорения

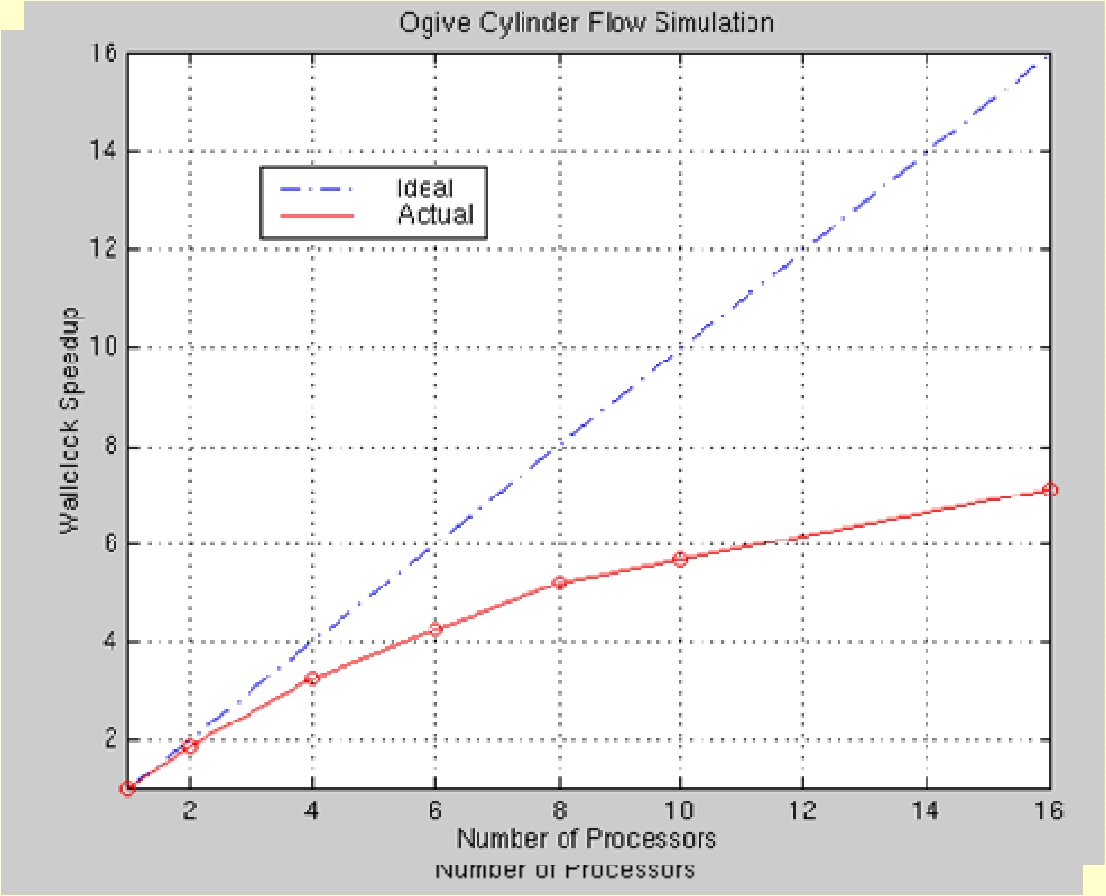
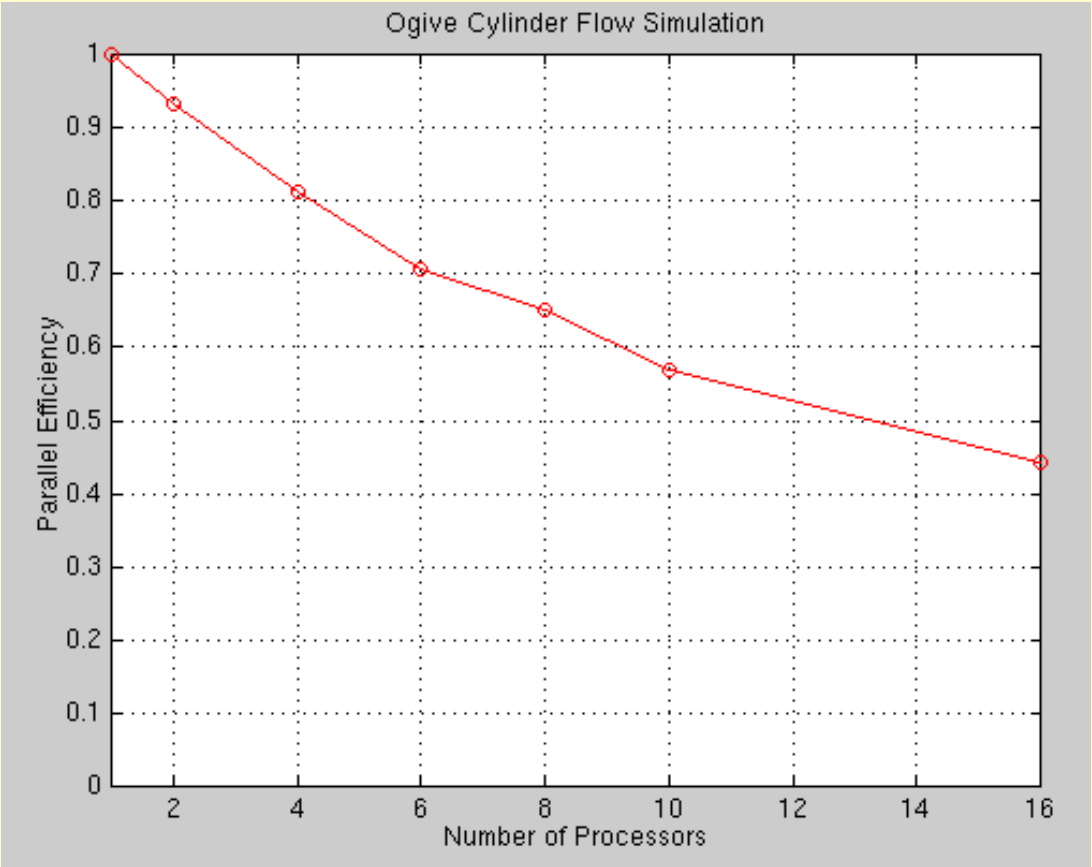


График эффективности



Обзор технологий параллельного программирования, необходимых для выполнения заданий

ПО для выполнения задания (1)

- Утилиты ОС для выхода удаленные машигны

  ssh (UNIX-подобные системы)

  putty, WinSCP – Windows

- Компиляция программ:

  xlc (IBM), icc (Intel), gcc

  Скрипты для компиляции MPI- программ

  mpicc – **Регатта**

  mpixlc – BlueGene/P (MPI)

  mpixlc_r – BlueGene/P (MPI+OpenMP)

  mpicc, mpiicc – «Ломоносов»

ПО IBM для выполнения задания (2)

- Системы управления заданиями:

  LoadLeveler – IBM

- Постановка задания в очередь на выполнение

  mpisubmit – Регатта

  mpisubmit.bg – BlueGene/P

  llsubmit – BlueGene/P

- Контроль за прохожлением задания:

  llq – LoadLeveler

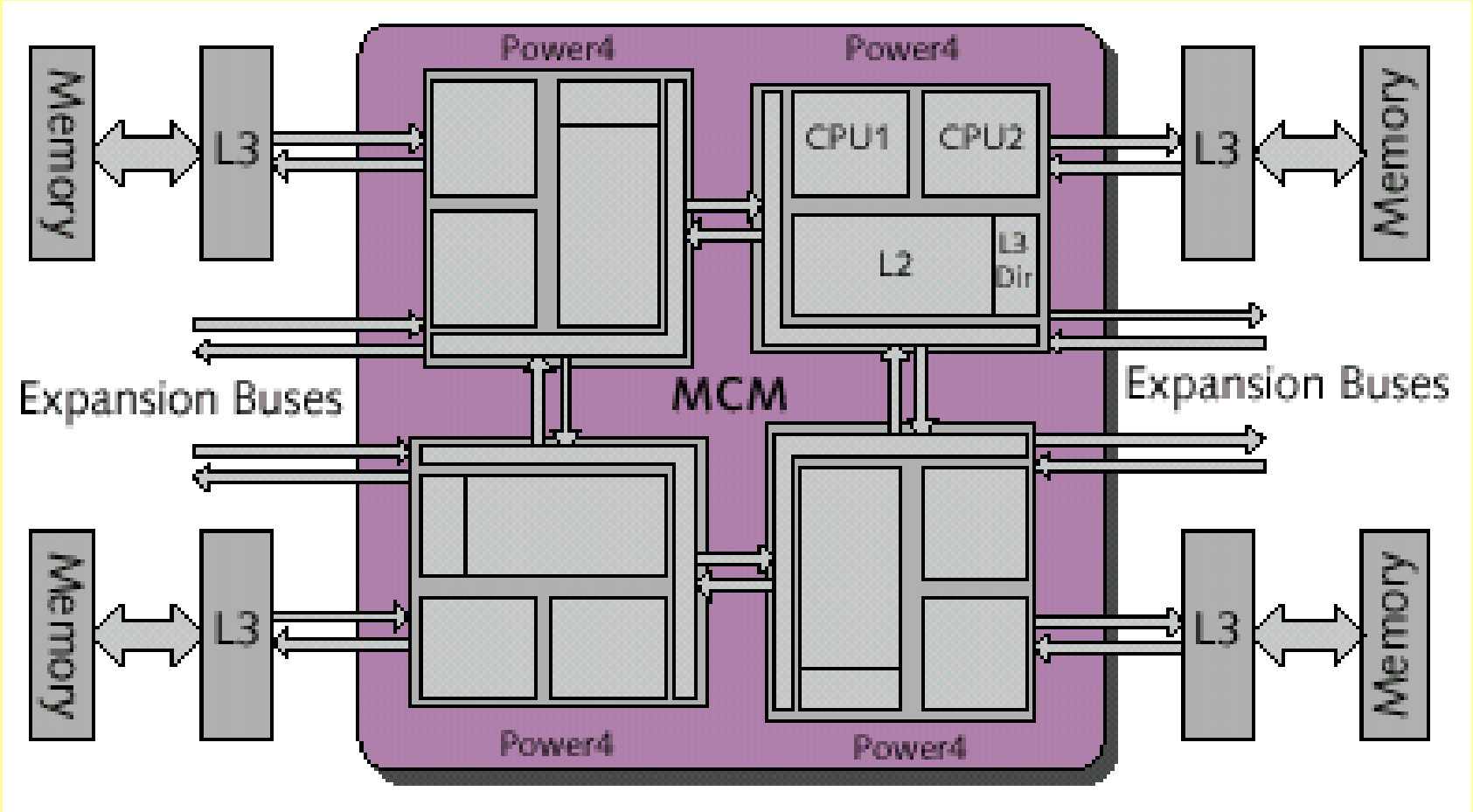
Вычислительные системы

p-series 690 Регатта

- 16-процессорная SMP система
- IBM Power4 процессоры
- 1.3 GH - тактовая частота
- 83 GFlops -максимальная производительность
- 64 Gbytes ОЗУ
- 32 KB L1 cache на процессор
- 1.41 MB L2 cache (общий для 2-ух процессор.)
- 128 MB L3 cache (общий для 8 процессоров)

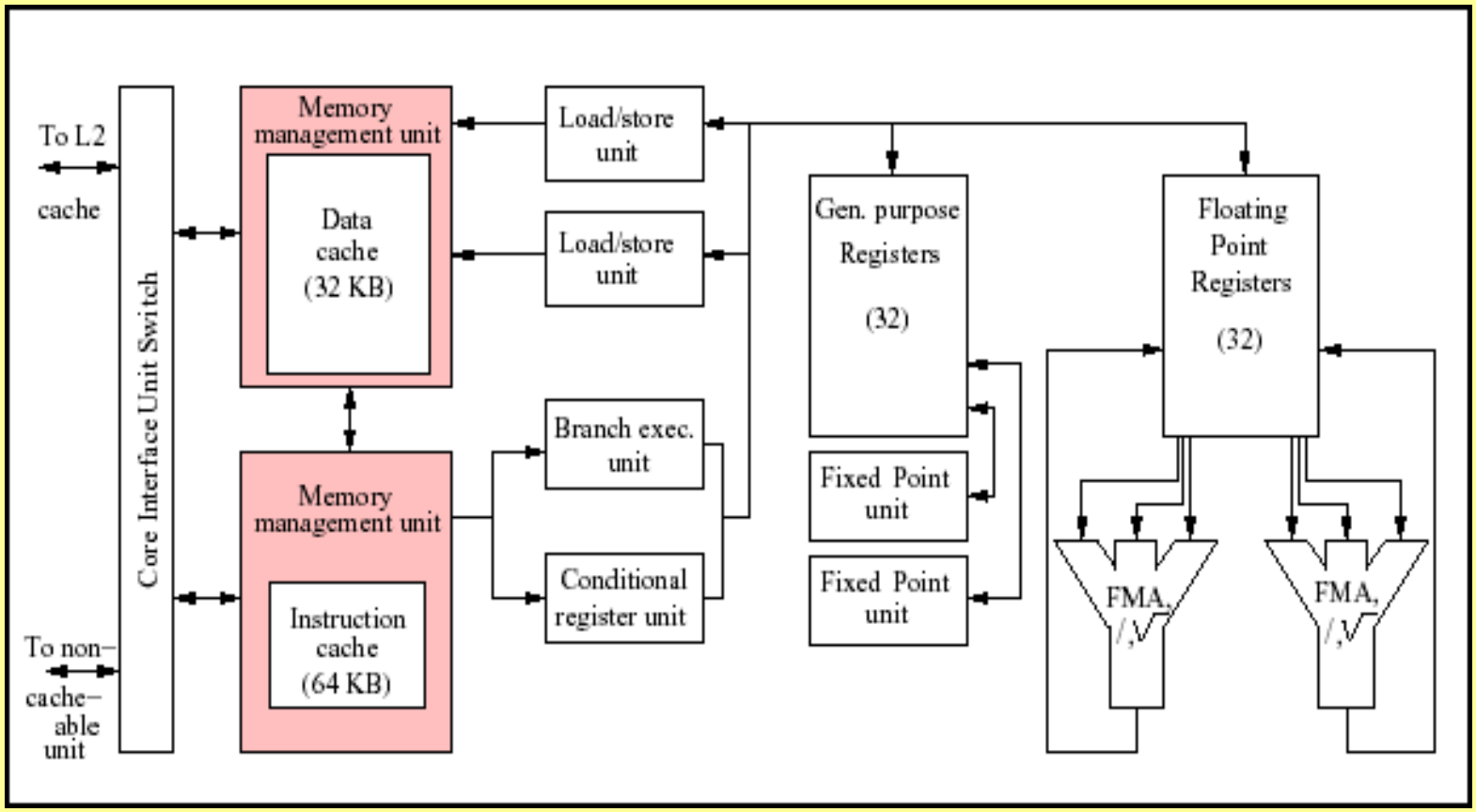
Архитектура IBM pSeries690 Regatta

/



Архитектура IBM pSeries690 Regatta.

Процессор POWER4.



IBM компиляторы

	Посл.	MPI	OpenMP	Mixed
Fortran 77	xlf	mpxlf	xlf_r	mpxlf_r
Fortran 90	xlf90	mpxlf90	xlf90_r	mpxlf90_r
Fortran 95	xlf95	mpxlf95	xlf95_r	mpxlf95_r
C	cc	mpcc	cc_r	mpcc_r
	xlc	mpxlc	xlc_r	mpxlc_r
C++	xlC	mpCC	xlC_r	mpCC_r

Опции компиляторов

- Уровни оптимизации
  - **-O** базовая оптимизация
  - **-O2** то же, что -O
  - **-O3** более агрессивная оптимизация
  - **-O4** еще более агрессивная оптимизация: межпроцедурный анализ (IPA) , оптимизация с учетом особенностей архитектуры
  - **-O5** агрессивный IPA

IBM компиляторы (1)

- -q64
  - если надо больше, чем 2GB
  - **Просто увеличивает адресное пространство**
  - -O5 -qarch=pwr4 -qtune=pwr4 и
  - -O3 (-qhot) -qarch=pwr4 -qtune=pwr4

IBM компиляторы (2)

- При использовании O3 или ниже рекомендуется оптимизация под архитектуру (выполняется автоматически для -O4 и -O5):

  -qarch=auto оптимизация под архитектуру

  -qtune=auto оптимизация под процессор

  -qcache=auto оптимизация под кэш

Общие рекомендации

- Профилировка для определения hot spots (gprof)
- Для ключевых функций – возможность замены на библиотечные вызовы (ESSL)
- Настройка опций компиляторов
- Использование MASS & MASSV библиотек
- Inlining часто используемых небольших функций
- Ручная настройка программы:

  - оптимизация доступа к памяти
  - оптимизация использования сопроцессоров и функциональных устройств
  - оптимизация I/O

Опции компилятора для OpenMP

- _r суффикс для имени компиляторов например, xlc_r
- –qsmp=omp флаг
  - указание компилятору интерпретировать OpenMP директивы
- Автоматическое распараллеливание

  -qsmp
  - имеет смысл использовать

OpenMP

- Автоматическое распараллеливание

  -qreport=smplist

  - Исходный текст
  - mycode.lst
  - Информация о распараллеленных циклах

- Размер стека
  - По умолчанию 4 MB
  - Может быть увеличена

    setenv XLSMPOPTS $XLSMPOPTS\:stack=*size*

    где *size* – размер в байтах

Математическая библиотека (1)

- MASS library
  - **M**athematical **A**cceleration **S**ub**S**ystem
- sqrt, rsqrt, exp, log, sin, cos, tan, atan, atan2, sinh, cosh, tanh, dnint, x**y
- Fortran и C

Математическая библиотека (2)

- Опции:

  Fortran: -lmass

  C: -lmass -lm

- Ускорение:

exp	2.4
log	1.6
sin	2.2
complex atan	4.7

Математическая библиотека (3)

- Использование векторизованных функций
  - Требует незначительных изменений в программе
  - не переносимы
  - Дают хороший результат
- Линковка с использованием –lmassv
- Вызов функций:
  - Префикс к именам функций
    - vs для 4-byte вещественных чисел (single precision)
    - v для 8-byte вещественных чисел (double precision)

Математическая библиотека (4)

- пример: одинарная точность

  call vsexp(y,x,n)

  – x вектор длины n

  – Y вектор длины n

- ускорение

	4-byte	8-byte
exp	9.7	6.7
log	12.3	10.4
sin	10.0	9.8
complex atan	16.7	16.5

LoadLeveler

- Система управления заданиями на многопользовательских системах, состоящих из нескольких вычислительных узлов
- Оптимизирует использование имеющихся вычислительных ресурсов
  - Учет приоритета задач и пользователей
  - Динамическое распределение ресурсов
  - Допускается использование разнородных вычислительных узлов
  - Используется для запуска как последовательных, так и параллельных задач
- Пользователь формулирует задания в виде командных файлов
- Поддерживает очередь заданий

Основные команды LoadLeveler (1)

llsubmit – постановка задачи в очередь

Пример:

```
popova@regatta:~/SuperComp2011/assignment5$ llsubmit job.cmd
llsubmit: The job "regatta.hpc.47739" has been submitted.
```

mpisubmit – скрипт постановки в очередь MPI-программ

llq – просмотр текущего статуса очереди

llcancel – удаление задачи из очереди

Основные команды LoadLeveler (2)

popova@regatta:...C_2011/assignment5

```
popova@regatta:~/SC_2011/assignment5> mpisubmit -n 4 gauss mat_1024.mat
llsubmit: Stdin job command file written to "/tmp/loadlx_stdin.8100.Rikkd7".
llsubmit: The job "regatta.24944" has been submitted.
popova@regatta:~/SC_2011/assignment5> llq
Id                       Owner      Submitted   ST PRI Class        Running On
------------------------ ---------- ----------- -- --- ------------ -----------
regatta.24944.0          popova     10/7  00:08 R  50  test_class   regatta2
regatta.24939.0          tiger      10/6  17:01 HS 50  test_class

2 job step(s) in queue, 0 waiting, 0 pending, 1 running, 1 held, 0 preempted
popova@regatta:~/SC_2011/assignment5> llq
Id                       Owner      Submitted   ST PRI Class        Running On
------------------------ ---------- ----------- -- --- ------------ -----------
regatta.24944.0          popova     10/7  00:08 R  50  test_class   regatta2
regatta.24939.0          tiger      10/6  17:01 HS 50  test_class

2 job step(s) in queue, 0 waiting, 0 pending, 1 running, 1 held, 0 preempted
popova@regatta:~/SC_2011/assignment5> llq
Id                       Owner      Submitted   ST PRI Class        Running On
------------------------ ---------- ----------- -- --- ------------ -----------
regatta.24939.0          tiger      10/6  17:01 HS 50  test_class

1 job step(s) in queue, 0 waiting, 0 pending, 0 running, 1 held, 0 preempted
popova@regatta:~/SC_2011/assignment5> ls
assignment.pdf  gauss.24943.out  gauss_elimination.c  generator_relax.cpp  mat_1024.mat
gauss*          gauss.24944.out  generator*           Instructions.pdf     Regatta_user_guide.pdf
popova@regatta:~/SC_2011/assignment5>
```

Дополнительные команды (скрипты) LoadLeveler

Автоматически составляют командные файлы для MPI и OpenMP задач

```
ompsubmit –n 4 –w 00:10:00 myprog
mpisubmit –n 4 –w 00:10:00 myprog
```

Схема выполнения заданий на системе Regatta

- **Выход на удаленную систему.**

  ssh –X ivanov@regatta.cs.msu.su

- **Копирование файлов с заданием в текущий каталог**

  cp ~popova/SC_2011/assignment_1.tar.gz .

- **Распаковка архива**

  gzip –d assignment_1.tar.gz

  tar –xvf assignment_1.tar

- **Компиляция программы**

  mpicc –o ass1 cg.c

- **Постановка в очередь на выполнение**

  mpisubmit –w 10:00 –n 8 ass1

- **Просмотр состояния очереди**

  llq

- **Удаление задания из очереди в случае необходимости**

  llcancel <id>

- **Копирование результатов на локальную машину.**

Материалы лекции

- http://angel.cs.msu.su/~popova

- www.ibm.com/software/awdtools/xlcpp/library/
- IBM POWER4, IBM Journal of Research and Development(Vol. 46, No. 1, Jan. 2002), www.research.ibm.com/journal/rd46-1.html
- www.redbooks.ibm.com

  The POWER4 Processor.Introduction and Tuning Guide,SG24-7041-00, Nov. 2001

  http://www.mhpcc.edu/training/workshop/loadleveler/MAIN.html

  http://www.redbooks.ibm.com/abstracts/sg246038.html

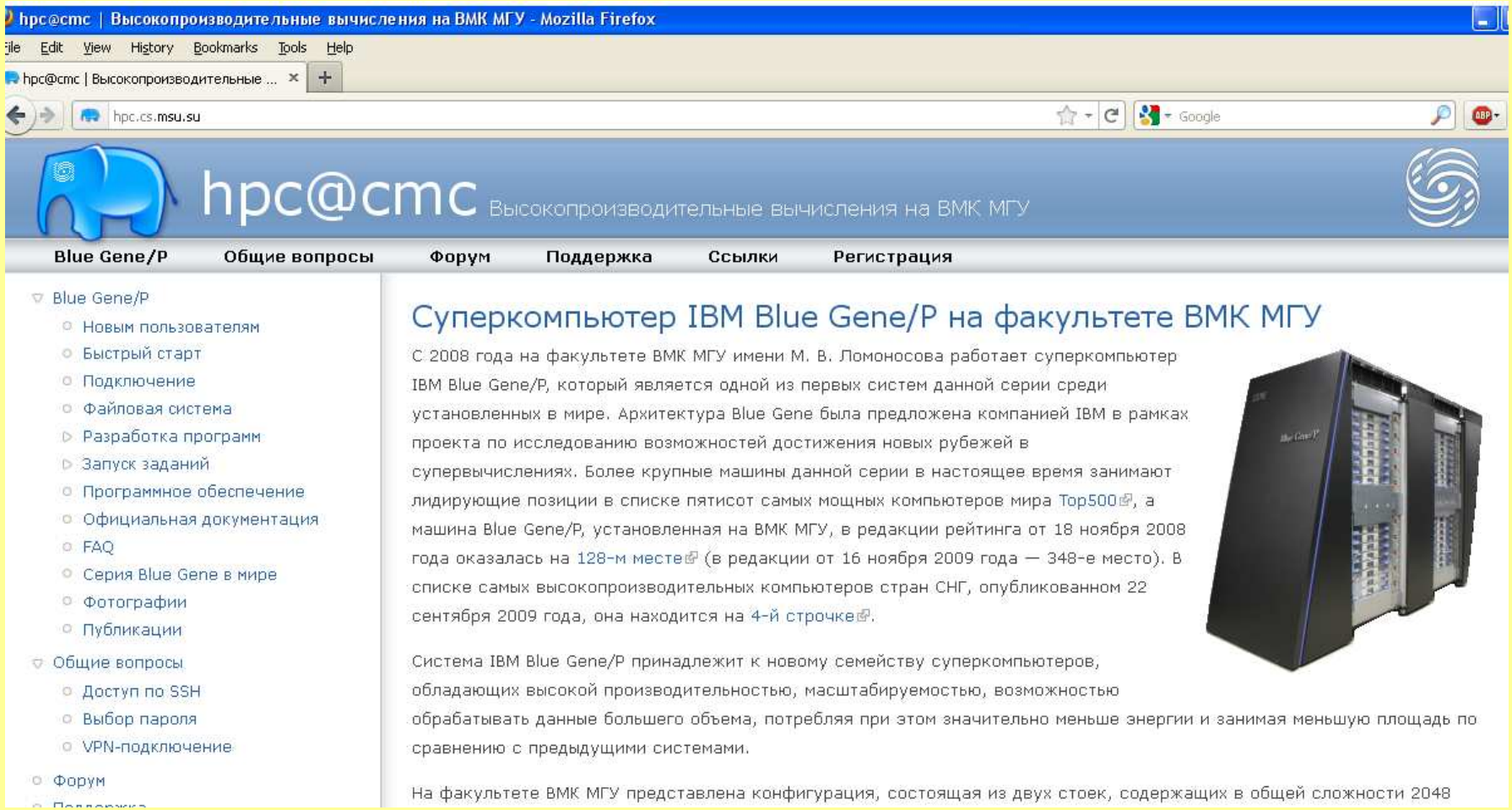
  http://www.hlrn.de/doc/quickstart/qs_loadl.html

Архитектура и программное обеспечение массивно-параллельной вычислительной системы

IBM Blue Gene / P

http://hpc.cs.msu.su

Дополнительная информация на сайте



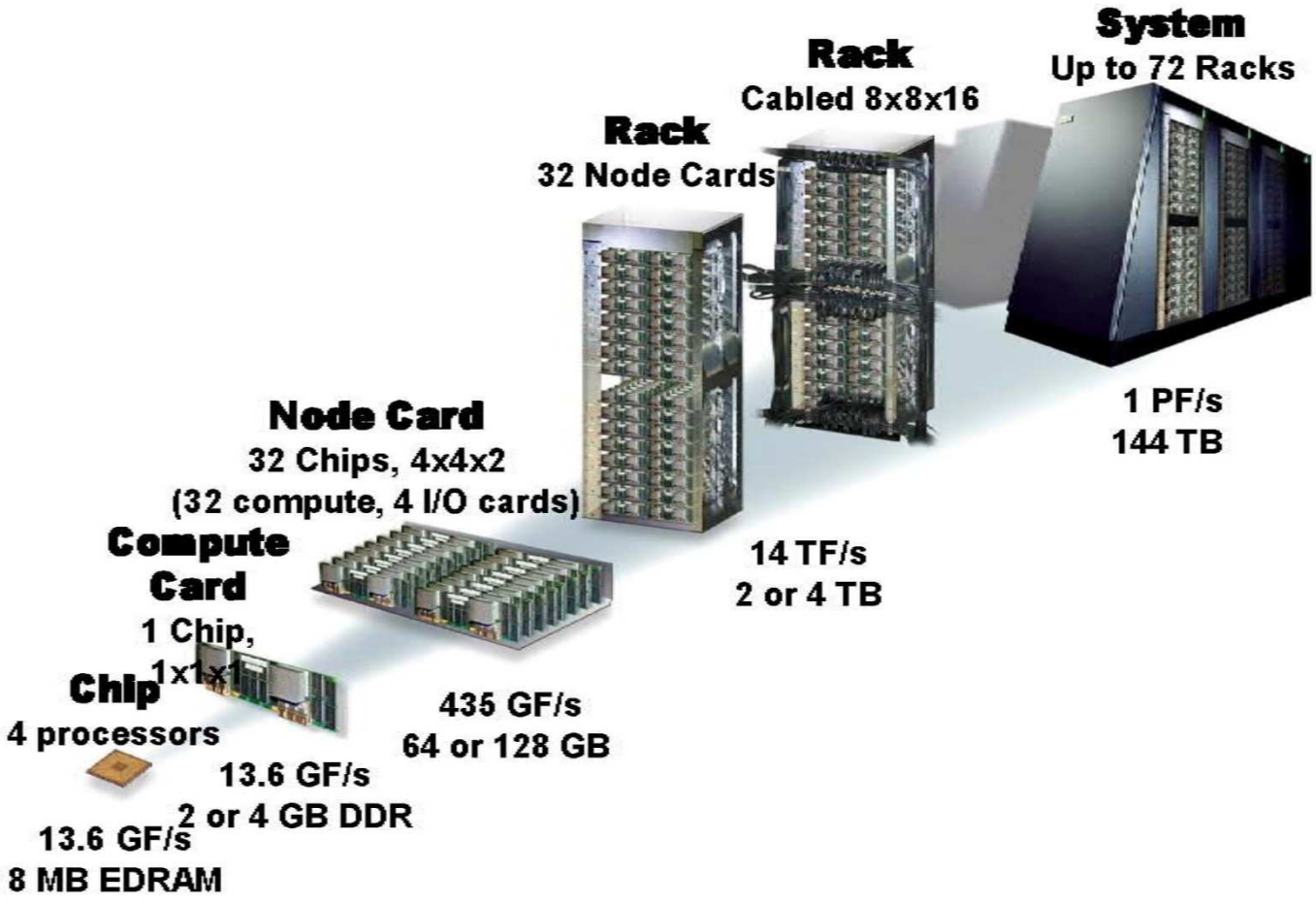
История проекта Blue Gene

- Blue Gene/L
  - Начинался как массивно-параллельный компьютер для изучения фолдинга белков
  - Первый прототип был собран в 2004 г.
    - занял первую строчку в Top 500 с производительностью в 70.72 Тфлоп/с
  - 2-х ядерный чип
- Blue Gene/P
  - Продолжение линейки Blue Gene
  - Увеличена частота процессора и объем памяти
  - 4-х ядерный чип
  - Самая большая система на основе Blue Gene/P установлена в Германии (JUGENE)
    - 1 Пфлоп/с пиковая, 825 Тфлоп/с реальная
- Blue Gene/Q
  - Ожидается в 2011 году, производительность ~20 Пфлоп/с
  - 8-ядерный чип

Общая характеристика систем Blue Gene

- Массивно-параллельные системы с распределенной памятью
- Высокая плотность упаковки
  - процессоры с низким энергопотреблением (40 W ~ лампочка)
- Высокопроизводительный интерконект
  - несколько комутационных подсистем для различных целей
- Ультра легкая ОС
  - выполнение вычислений и ничего лишнего
- Стандартное ПО Standard software
  - Fortran/C/C++ и MPI

Blue Gene/P Hardware



Blue Gene/P

1 стойка

- 1024 четырехъядерных вычислительных узлов
- 13.6 GF/s - производительность одного вычислительного узла
- 13.9 Tflops - производительность 1 стойки
- 2 GB - оперативная память одного узла
- 2 TB - суммарная оперативная память
- 8 узлов ввода/вывода
- Размеры - 1.22 x 0.96 x 1.96
- занимаемая площадь 1.17 кв.м.
- энергопотребление (1 стойка) - 40 kW (max)

BlueGene/P факультета ВМиК

- пиковая производительность 27.8 Tflop/s
- 2 стойки
- 2048 4-ех ядерных узлов
- общий объем ОЗУ 4 ТВ
- http://hpc.cs.msu.ru

Компоненты Blue Gene/P

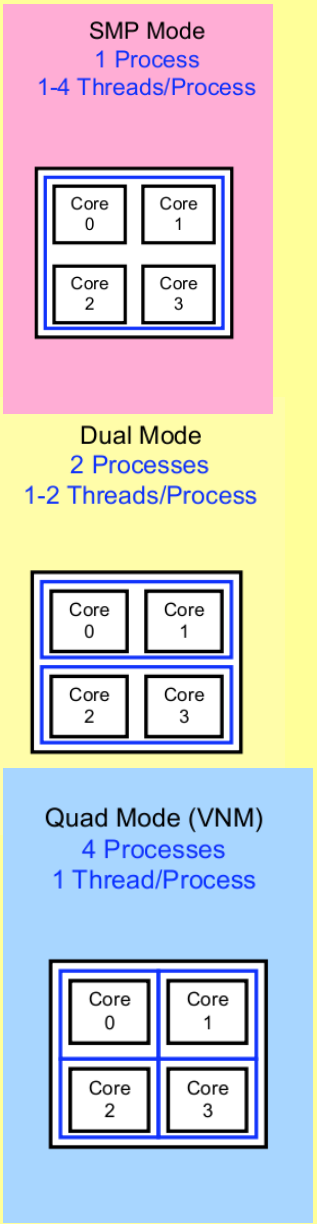
- Основная единица – четырехядерный вычислительный узел, ядро – PowerPC 450 850Mhz + память (2GB)
- Плата = 32 вычислительных узла + до 2х узлов ввода-вывода
- Стойка – 32 платы
- Итоговое число ядер на стойку - 4096

Характеристики вычислительного узла

- 4 ядерный 32-битный процессор PowerPC 850 Мгц
  - Двойное устройство для работы с вещественными числами с плавающей точкой (double precision)
  - 2 Гб памяти
  - Работает под управлением облегченной ОС
    - Создание процессов и управление ими
    - Управление памятью
    - Отладка процессов
    - Ввод-вывод
  - Объем виртуальной памяти равен объему физической

Характеристики вычислительного узла

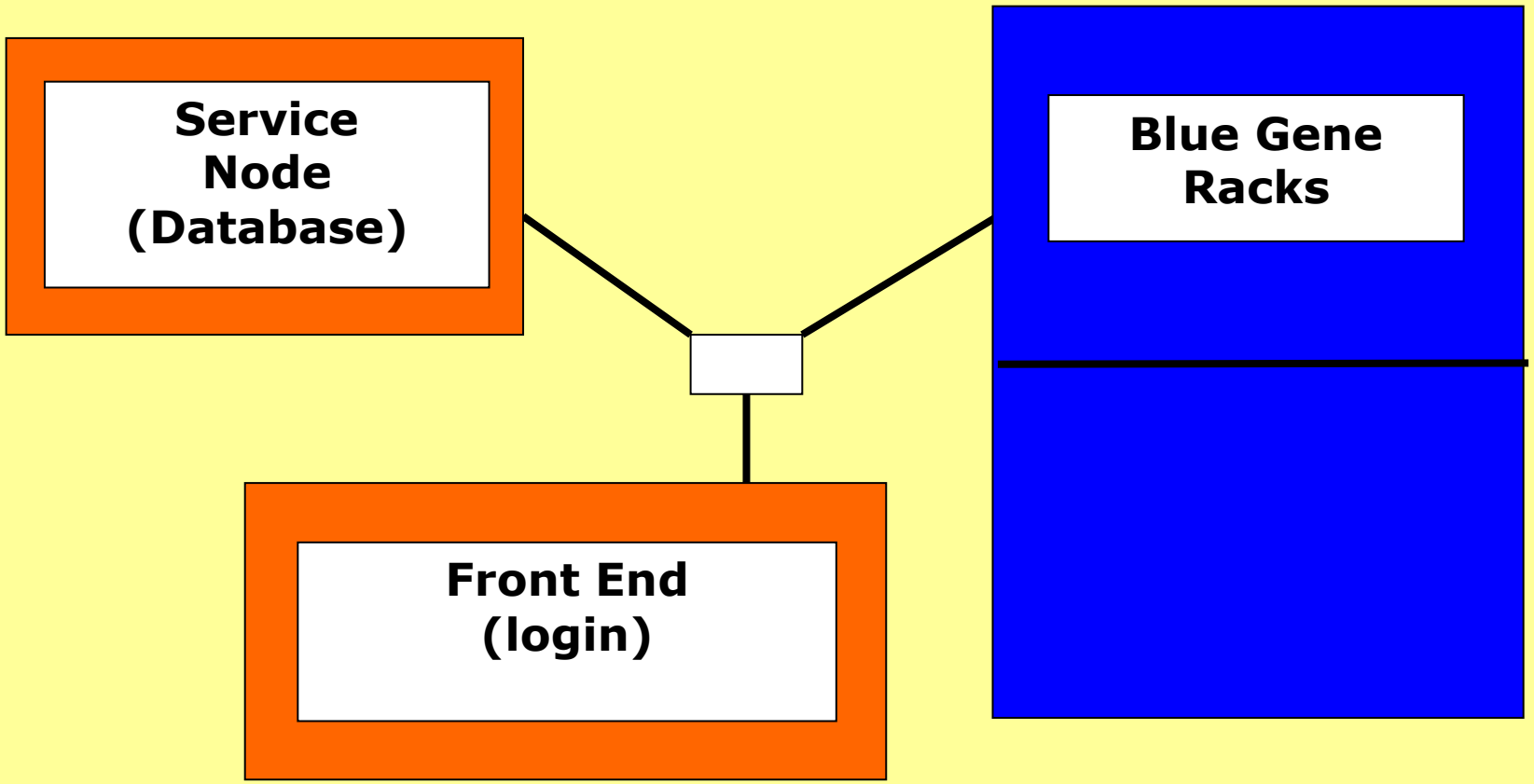
- 3 режима использования ядер
  - **SMP**: 1 MPI процесс из 4 SMP нитей, 2 Гб памяти
  - **DUAL**: 2 MPI процесса по 2 SMP нити, 1 Гб памяти на MPI процесс
  - **VNM**: 4 MPI процесс



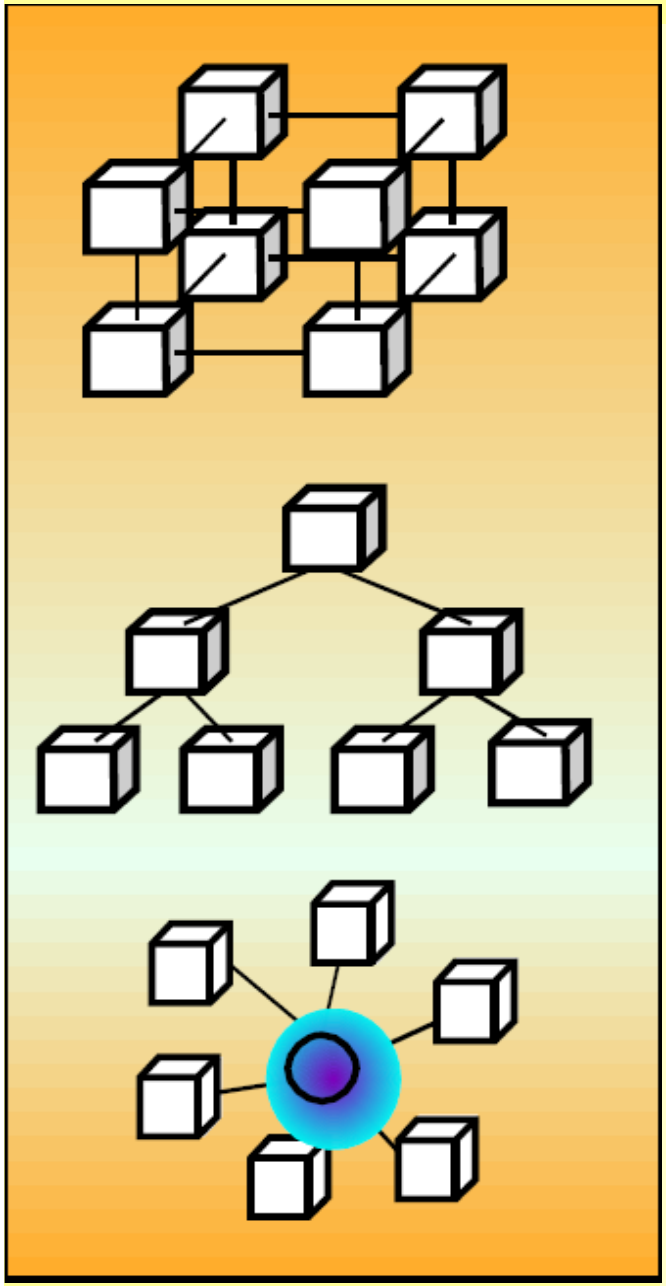
Компоненты Blue Gene/P

- Помимо вычислительных узлов, в состав системы также входят:
  - узлы ввода-вывода
  - узел управления системой
  - не менее одного узла front end (через них осуществляется доступ пользователей к системе)
  - сеть, связывающая компоненты системы
  - специализированная сеть для сообщения между сервисным узлом и узлами ввода-вывода

BlueGene/P



Коммуникационные сети

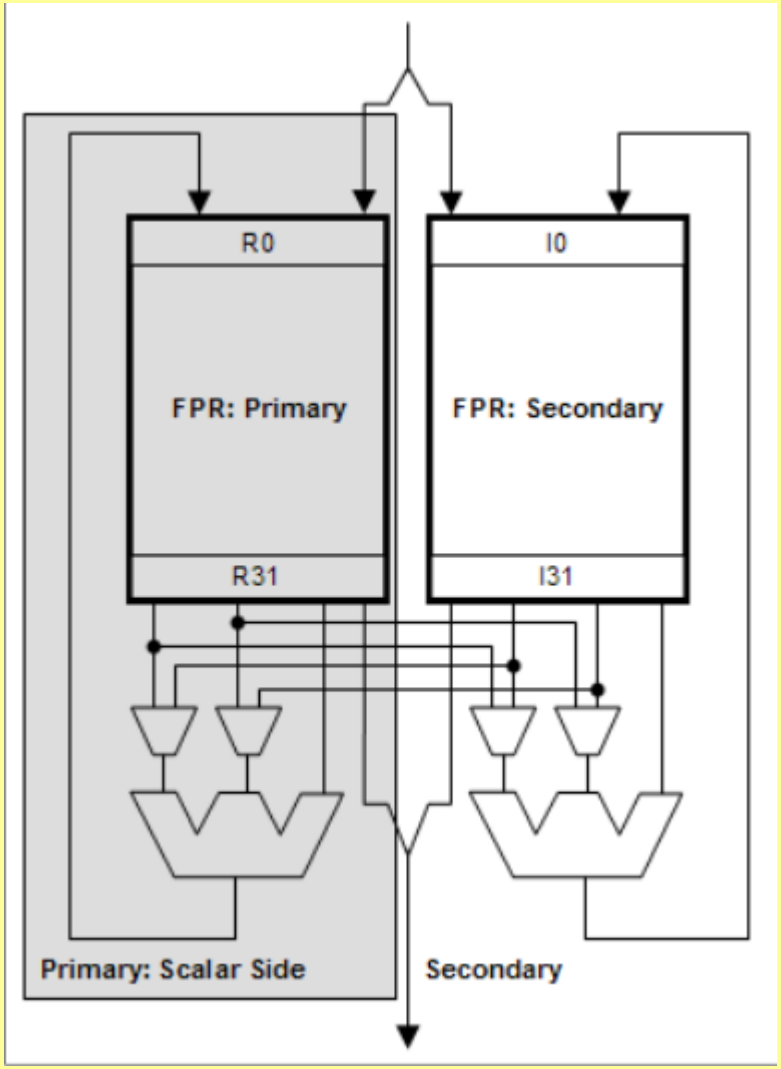


- **3-мерный тор**
  - Виртуальная аппаратная маршрутизация без буферизации
  - 3.4 Гбит/с на всех 12 портах (5.1 ГБ/с на узле)
  - Аппаратные задержки: 0.5 мс между соседними узлами, 5 мс между самыми далекими
  - Основная коммуникационная сеть
    - Используется в том числе для многих коллективных операций
- **Коллективная сеть – дерево**
  - Для глобальной коммуникации один-ко-всем (broadcast, reduction)
  - 6.8 ГБ/с на порт
  - Соединяет все вычислительные узлы и узлы ввода-вывода
  - Используется для коллективных операций и коммуникатора MPI_COMM_WORLD
- **Высокоскоростная сеть для глобальных прерываний**
  - Для MPI_Barrier

Память

- Оперативная память – до 2GB на вычислительный узел, пропускная способность 13.6GBps
- Трёхуровневый кэш:
  - L1 – отдельный для каждого ядра, размер 32Kb
  - L2 – отдельный для каждого ядра, используется для предварительной выборки информации из кэша L1. Считывает\записывает по 16b за одно обращение.
  - L3 – разделен на две части по 4MB, доступ к ним имеют все четыре ядра, для каждого есть канал чтения и канал записи. Связан с 10-гигабитной сетью (в том случае, если на карте имеется узел ввода-вывода)

Double Hammer FPU



- SIMD инструкции могут выполняться одновременно на двух FPU
- Параллельные операции load/store
- Данные **должны** быть выровнены по 16-байтовой границе
  - Иначе производительность будет значительно снижена
  - Даже хуже, чем при использовании только одного FPU
- Компилятор сможет сгенерировать SIMD инструкции, только если данные в памяти расположены подряд (stride-one access)
  - Хотя при более высоких (-O4, -O5) уровнях оптимизации компилятор попытается сгенерировать SIMD инструкции и для данных, расположенных не подряд
  - -O3 –qarch=450d –qtune=450

BlueGene/P ПО (1)

- Linux® на узлах ввода\вывода
- MPI (MPICH2) и OpenMP (2.5)
- Стандартное семейство компиляторов IBM XL: XLC/C++, XLF
- Компиляторы GNU
- Система управления заданиями LoadLeveler
- Файловая система GPFS
- Инженерная и научная библиотека подпрограмм (ESSL)

BlueGene/P ПО (2)

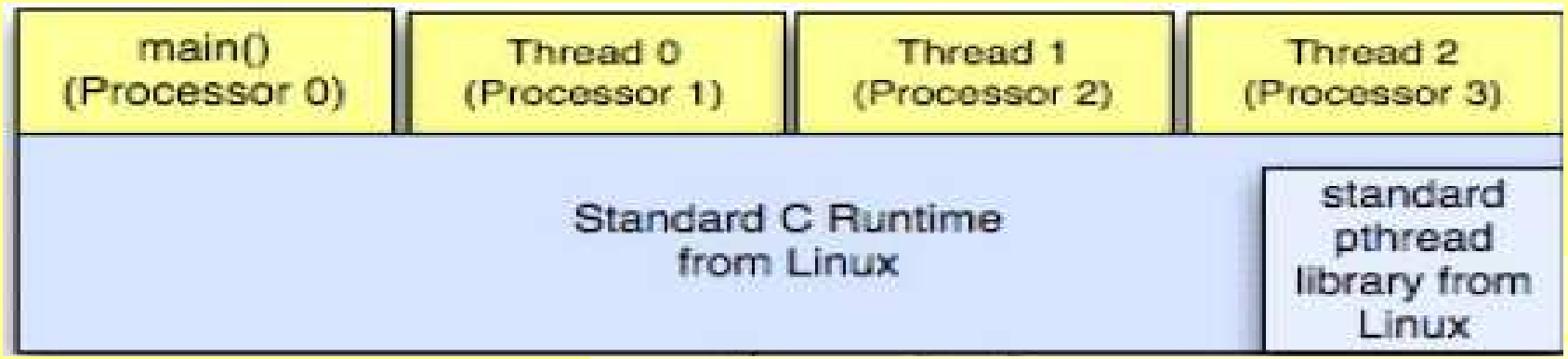
- Compute Node Kernel (CNK)
  - Минимальное ядро – обработка сигналов, передача системных вызовов к узлам ввода-вывода, старт-завершение задач, поддержка нитей
  - “linux-подобная” ОС
    - Нет некоторых системных вызово (fork() в основном)
      - Ограниченная поддержка mmap(), execve()
    - Однако, большинство приложений, которые работают под Linux, портируются на BG/P

Компиляторы Blue Gene

- IBM XL компиляторы (xlc, xlf77, xlf90)
- работают на front end узлах
  - Fortran: **mpixlf**, **mpixlf90**, **mpixlf95**
  - C: **mpixlc**
  - C++: **mpixlcxx**
- обычно являются скриптами
- GNU компиляторы существуют, но малоэффективны: **mpicc**

Режимы выполнения процессов в системе Blue Gene/P

- **Symmetrical Multiprocessing (SMP) Node Mode.**
- Физический узел выполняет **1 MPI-процесс**, внутри которого выполняются максимум **4 нити**.
- mpirun ... -mode smp ...

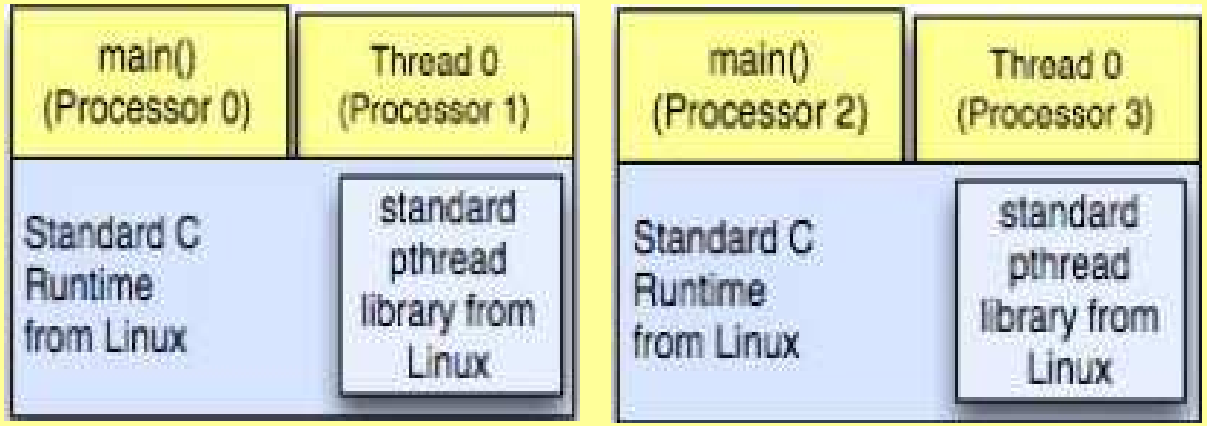


Режимы выполнения процессов в системе Blue Gene/P

- **Virtual Node Mode (VN).**
- На каждом физическом узле выполняются **4 MPI-процесса**.
- Устанавливается по умолчанию в mpirun

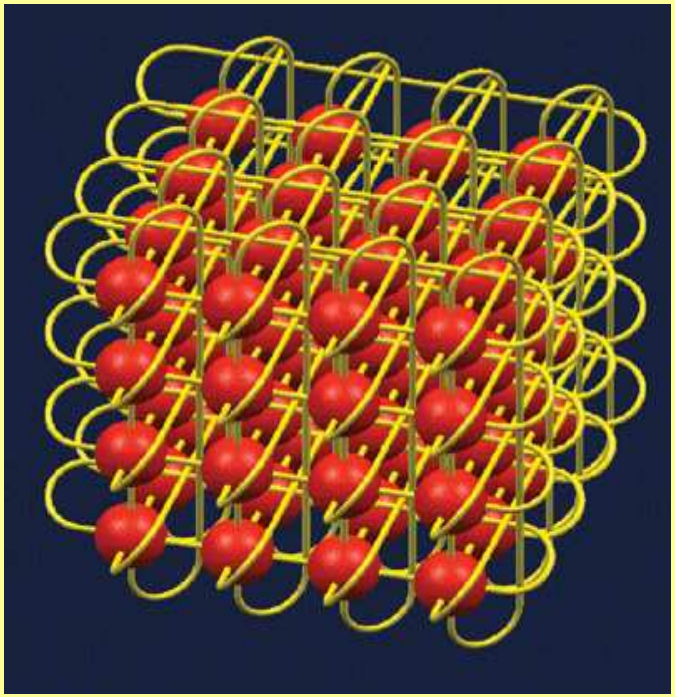
Режимы выполнения процессов в системе Blue Gene/P

- **Dual Node Mode (DUAL).**
- На одном физическом узле выполняются **2 MPI-процесса**, у каждого из которых максимум по **2 нити** (всего на одном узле выполняется не более 4 нитей).
- mpirun ... -mode DUAL ...



Mapping

(распределение процессов по процессорам)



По умолчанию распределение MPI-процессов в системе Blue Gene/P происходит в порядке XYZT,

где <X, Y, Z> - координаты процессора в торе,

T – номер ядра внутри процессора.

Сначала увеличивается X-координата, затем Y, затем Z, затем T.

Mapping

mpirun в системе Blue Gene/P позволяет распределять процессы двумя способами:

- с помощью аргумента командной строки
  –mapfile TXYZ (задаем порядок TXYZ или другие перестановки XYZT).

- создать свой map файл, указать в командной строке
  –mapfile my.map, где my.map – имя файла.

Синтаксис файла распределения очень прост – четыре целых числа в каждой строке задают координаты для каждого MPI-процесса (первая строка задает координаты для процесса с номером 0, вторая строка – для процесса с номером 1 и т.д.).

Очень важно, чтобы этот файл задавал корректное распределение, с однозначным соответствием между номером процесса и координатами <X, Y, Z, T>.

http://www.hpcwire.com/hpcwire/2011-10-06/russia_seeks_rocket_simulation_system.htmlNews

- October 06, 2011
- **Russia Seeks Rocket Simulation System**
- The business of testing rockets isn't a cheap one, and Russian scientists are looking for less expensive, quicker ways to analyze new designs as they race toward space exploration goals. Modeling and simulation, which is used to model everything from car crashes to more streamlined beer cans, is on the agenda as Russia looks to speed time to rocket development.

  Roscocosmos, the Russian state space organization, has published a tender for development of "manufacturing technology of a cluster compute system with hybrid architecture for imitational modeling of rocket and launchers' real flight conditions," reports CNews. According to the proposal, Russia is prepared to set aside around $1.74 million for the rocket testing cluster.

  Russian space officials claim they require a system to be capable of providing peak performance of up to 10 teraflops, hold 20 GB RAM and offer 4000 GB of disk space.

  The tender goes on to note that the agency is looking for a contractor that can not only deliver this "manufacturing technology" but that can also provide a sample of such compute system (with CPU, GPU architecture), which will be installed at other sites in the space agency's network of research and development centers.

Ссылки

- http://www.ibm.com/servers/deepcomputing/bluegene.html
- http://www3.ibm.com/systems/deepcomputing/bluegene/
- IBM System Blue Gene Solution: Blue Gene/P Application Development, SG24-7287-00
  http://www.research.ibm.com/journal/rd/521/tocpdf.html