

33. Программно-аппаратная архитектура суперкомпьютеров Ломоносов и Blue Gene/P.

IBM Blue Gene – название архитектуры семейства суперкомпьютеров. Проект начался в 1999 году, изначально использовался для фолдинга белков и расшифровки генома. Всего было 3 поколения: BG/L, BG/P и BG/Q. Это массивно-параллельная (а не кластерная) система, то есть все конструктивные элементы (процессоры, коммуникация) объединены одним аппаратным решением энергоэффективность, быстрая коммуникация. Долго занимал первые места в топ-500.

Одна стойка BG/P состоит из 1024 4-х ядерных вычислительных узлов с производительностью 13.6 GF/s и 2 Gb оперативной памяти. Таким образом, стойка имеет производительность 13.6 TF/s память 1 Tb.

Конфигурация на ВМК включает в себя 2 стойки: 2048 4-х ядерных узлов с 4 Tb памяти и пиковой производительностью 27.2 TF/s. Реальная производительность по тесту Linpack – 23.5 TF/s или 85% от пиковой.

Основной единицей является 4-х ядерный 32-х битный процессор PowerPC 450 850 Mhz. Узлы объединяются в узловую карту, которая содержит 32 вычислительных узла и до 2-х узлов ввода-вывода для обращения к внешней памяти. 32 узловые карты составляют одну стойку.

PowerPC 450 850 Mhz производит операции над числами с плавающей точкой, имеет 2 Gb оперативной памяти, причем физическая память равна виртуальной, и управляется облегченной ОС. ОС умеет создавать, управлять и отлаживать процессы, управлять памятью и производить операции ввода-вывода. Память имеет трехуровневый кэш. Первый и второй уровни – свои для каждого ядра объемом 32 kb и 16 b соответственно. Второй используется для выборки информации из первого. Третий уровень общий для всех ядер, разделен на две части по 4 Mb.

Кроме того, в процессорах используется Double Hammer FPU – сопроцессор, ускоряющий производство операций с вещественными числами.

Процессы могут выполняться в 3-х режимах:

1. SMP-режим: 1 MPI процесс на узел, разделенный на 4 потока, используется 2 Гб памяти на процесс
2. Dual-режим: 2 MPI процесса на узел, разделенный по 2 потока, используется 1 Гб памяти на процесс
3. Quad-режим: 4 MPI процесса на узел, по 1 потоку на каждый, используется 512 Кб памяти на процесс

Для коммуникации между узлами используется несколько коммуникационных сетей:

1. 3-х мерный тор – основная сеть, связывает процессорные элементы и реализует все передачи для MPI. Каждый процессор имеет связь с ближайшими соседями в 3-х мерной решетке. У каждого узла есть 6 соседей. Получается, процессоры образуют “куб”, боковые грани которого замыкаются между собой, то есть тор.
2. Коллективная сеть-дерево – соединяет все вычислительные узлы с узлами ввода-вывода. Имеет большую пропускную способность, но поддерживаются только для прямоугольных решеток.
3. Высокоскоростная сеть глобальных прерываний – для быстрой барьерной синхронизации.

В общем, архитектура BlueGene/P состоит из: вычислительных узлов, многопроцессорный front-end узел (с полноценной ОС, для подготовки программ к выполнению на вычислительных узлах) и служебный узел (мониторинг состояния системы).

Состав ПО: Linux (узлы в-в), компиляторы IBM XL (C++, Fortran) и GNU, система управления заданиями LoadLeveler (нужна для запуска заданий), файловая система GPFS и библиотеки подпрограмм (ESSL и MASS).

ОС узла - Compute Node Kernel. Это урезанная Linux, без некоторых системных вызовов (в основном, динамического создания процессов).

Суперкомпьютер Ломоносов – гибридный суперкомпьютер (помимо ЦП традиционной архитектуры используют для вычислений специализированные процессоры е.г. графические). В нем используется 3 вида вычислительных узлов и процессоры с различной архитектурой. Основные узлы (90% производительности) – blade-платформа T-blade.

Пиковая производительность – 1.7 Пфлопс.

Типы вычислительных узлов:

1. Тип T-Blade2(link is external)(УВ1). 12 Гб оперативной памяти, 2 процессора Intel Xeon 5570 по 4 ядра, всего процессоров 8320, всего узлов 4160.
2. Тип T-Blade1(УВ2). 24 Гб оперативной памяти, 2 процессора Intel Xeon 5570 по 4 ядра, всего процессоров 520, всего узлов 260.
3. Тип T-Blade2(link is external)(УВ1). 24 Гб оперативной памяти, 2 процессора Intel Xeon 5670 по 6 ядер, всего процессоров 1280, всего узлов 640.
4. Тип T-Blade1(УВ2). 48 Гб оперативной памяти, 2 процессора Intel Xeon 5670 по 6 ядер, всего процессоров 80, всего узлов 40.
5. Тип Узлы на базе IBM® Cell (УВ3). 16 Гб оперативной памяти, процессор PowerXCell 8i с 8 ядрами, всего процессоров 60, всего узлов 30.

Состав ПО: Языки программирования: C/C++, Fortran 77/90/95; Наборы компиляторов: Intel 12, GNU 4.4, Pathscale, PGI; Среды исполнения: OpenMPI 1.4, MVAPICH 1.1, IntelMPI 4;

Все узлы связаны тремя независимыми сетями:

1. Системная сеть - QDR InfiniBand, 40 Гбит/сек
2. Сервисная сеть - Ethernet, 10 Гбит/сек, 1 Гбит/сек и 100 Мбит/сек
3. Управляющая сеть - Ethernet, 10 Гбит/сек и 1 Гбит/сек
4. Сеть барьерной синхронизации и сеть глобальных прерываний, Т-Платформы