ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ

ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«СИБИРСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ТЕЛЕКОММУНИКАЦИЙ И ИНФОРМАТИКИ»

КАФЕДРА ВС

РАСЧЕТНО-ГРАФИЧЕСКОЕ ЗАДАНИЕ

по дисциплине

«Архитектура вычислительных систем»

Вариант №4

Выполнил:

студент группы ИВ-823

Шиндель Э. Д.

Проверил:

кандидат технических наук

Ефимов А. В.

Новосибирск 2020

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Задание: проанализировать мультиархитектуру суперВС Sunway TaihuLight (№4 в списке Top500) [1].

Sunway TaihuLight - китайский суперкомпьютер, который с июня 2016 по июнь 2018 года являлся самым производительным суперкомпьютером в мире с производительностью 93 петафлопса, согласно тестам LINPACK [2].

1. АНАЛИЗ АРХИТЕКТУРЫ СУПЕРВС

Суперкомпьютер Sunway TaihuLight включает шесть систем: вычислительную, сетевую, внешней памяти, обслуживания и диагностики, электропитания и охлаждения, программного обеспечения (рис. 1) и представляет собой иерархическую систему.

Рис. 1 – Общая архитектура Sunway TaihuLight.

Конфигурация этого суперкомпьютера включает сорок вычислительных стоек с 40 960 узлами с одним процессором в каждом, то есть в общей сложности

получается 10 649 600 процессорных ядер. Пиковая производительность Sunway TaihuLight достигает немыслимого значения в 125 PFLOPS, а в тесте Linpack несколько ниже – 93 PFLOPS. Работает он под управлением собственной операционной системы Sunway Raise OS 2.0.5 на базе Linux.

Система внешней памяти включает сеть внешней памяти, подсистему управления и дисковый массив с общей памятью 20 Пбайт, имеющий системную консоль, управляющий сервер и сеть для управления всей системой. Донгарра в докладе в министерстве энергетики США указывал на наличие в Sunway TaihuLight 288 SSD-дисков емкостью по 800 Гбайт и общую пропускную способность ввода-вывода на уровне 288 Гбайт/с.

Система обслуживания и диагностики отвечает за интерактивное управление узлами, суперузлами и вычислительной системой в целом — мониторинг состояния Sunway TaihuLight, обнаружение сбоев, ведение протоколов и т. п.

Охлаждение вычислительной и сетевой систем основано на непрямом водяном охлаждении, а в системе внешней памяти применяется смешанное воздушно-водяное охлаждение.

1.1. АНАЛИЗ ВЫЧИСЛИТЕЛЬНОГО СУПЕРУЗЛА

Суперкомпьютер Sunway TaihuLigh состоит на базе однопроцессорных вычислительных узлов, имеющих память по 32 Гбайт (по 8 Гбайт на CG). Два вычислительных узла образуют карту, четыре карты — плату, а из 32 плат собирается суперузел на 256 вычислительных узлов. Из четырех суперузлов состоит кабинет, а весь Sunway TaihuLight состоит из 40 кабинетов.

Суперузел имеет полносвязанный скрещиваемый коммутатор для поддержки вычислительно интенсивных работ. Чтобы соединить все кабинеты создали собственную сеть соединений PCIe 3.0, которую они назвали «Sunway Network». Сеть соединяет коммутаторы, оборудование для совместного использования ресурсов и все суперузлы с помощью 7-дюймовых кабелей, которые передают данные со скоростью 70 терабайт в секунду.

Сетевая система суперкомпьютера трехуровневая: на верхнем находится центральная коммутаторная сеть, через которую работают все суперузлы; на среднем — суперузловая сеть; на нижнем — сеть совместного использования ресурсов, связывающая ресурсы с суперузлами и обеспечивающая службы коммуникаций ввода-вывода и отказоустойчивости вычислительных узлов.

1.2. АНАЛИЗ АРХИТЕКТУРЫ ПРОЦЕССОРА SW26010

Данный суперкомпьютер использует многоядерные 64-битные RISC-процессоры SW26010, базирующиеся на архитектуре ShenWei [2].

Процессор SW26010 (рис. 2) [4, 5] имеет 260 ядер и включает четыре группы ядер CG (Core Group). Каждая группа содержит кластер из 64 вычислительных элементов (Computing Processing Element, CPE), которые и образуют основу вычислительной мощности процессора. Кроме CPE, связанных в кластере решеткой-массивом 8×8, каждая группа CG имеет одно свое ядро общего назначения — процессорный элемент управления (Management Processing Element, MPE). Процессор SW26010 имеет тактовую частоту 1,45 ГГц, однако по какой технологии он изготовлен, точных сведений нет.

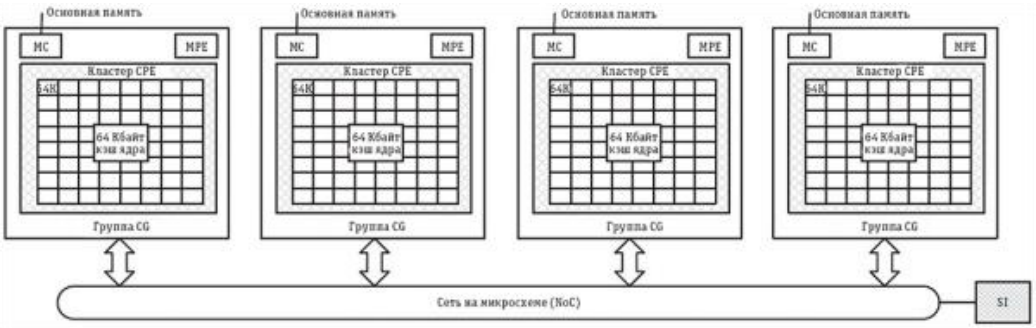

Рис. 2 – Архитектура SW26010.

Элементы CPE и MPE имеют 64-разрядную RISC-архитектуру с поддержкой 256-разрядных векторов (по четыре числа двойной точности, вдвое короче, чем у KNL) и команд «умножить и сложить». Каждый MPE содержит два векторных конвейера, соответственно, все 256 CPE дают 256×8 DP-результатов за такт и еще четыре MPE дают 4×8×2 результатов. И если умножить это на 1,45 ГГц, получается, что суммарная пиковая производительность SW26010 составляет 3062 GFLOPS.

Кроме процессорных ядер CPE и MPE, каждая CG содержит свой работающий с DDR3 контроллер памяти (Memory Controller, MC), а объединяются CG через общую для SW26010 «сеть на микросхеме» (Network on Chip, NoC). Процессор поддерживает связь с другими устройствами через системный интерфейс SI (рис. 1). Разработчики говорят об «общей глубоко объединенной многоядерной архитектуре» (Deeply Fused Many-Core, DFMC). Число CPE и топология их соединения в других DFMC-процессорах могут отличаться, а топология NoC — это решетка, кольцо или перекрестный коммутатор (crossbar), и она может подбираться в зависимости от количества компонентов в DFMC (число CG, MPE и MC может меняться [5]). Поэтому DFMC можно охарактеризовать как высокомасштабируемую гетерогенную микроархитектуру, ориентированную на высокопроизводительные вычисления.

Каждая группа CG имеет собственное адресное пространство памяти, связывающей ядро MPE и кластер CPE через контроллер MC. Ядра CPE и MPE имеют совместимую 64-разрядную RISC-архитектуру, напоминающую DEC

Alpha [5], образованную из 212 команд, работающих с 8-, 16-, 32- и 64-разрядными операндами. Кроме того, CPE поддерживают коммуникации на уровне регистров и передачи потоков данных, а также синхронизацию в их кластере. Суперскалярный уровень в MPE поддерживает декодирование сразу четырех команд и возможность параллельного выполнения семи команд, а в CPE декодируются и выполняются параллельно по две команды [6].

В DFMC память разделяется между MPE и CPE, а когерентность кэша достигается только между MPE. В MPE имеется двухуровневый кэш: кэш первого уровня команд (I) и данных (D) емкостью по 32 Кбайт и общий кэш второго уровня для команд и данных (256 Кбайт). В этих кэшах используются строки размером 128 байт. Кэши первого уровня в MPE являются четырехканальными наборно-ассоциативными, а кэш второго уровня — восьмиканальным. В CPE устроено по-другому: здесь имеются кэш команд первого уровня емкостью 16 Кбайт и сверхоперативная память SPM (scratch pad memory) емкостью 64 Кбайт с NUMA-задержками. Кроме того, в кластере CPE предусмотрен еще общий кэш второго уровня для команд [5]. В отличие от кэша, сверхоперативная память не содержит копию данных из основной памяти и может быть сконфигурирована как быстрый буфер, напрямую управляемый пользователем, или как программно-эмулируемый кэш, который дает автоматическое кэширование данных, что, естественно, сильно уменьшает производительность. Здесь усматривается некая аналогия со сверхоперативной памятью MCDRAM в KNL. Выбор SPM по сравнению с D-кэшем позволяет упростить аппаратную реализацию.

Слабое место SW26010 — иерархия памяти. Поэтому важнейшей при оптимизации программ для SW26010 задачей является уменьшение нагрузки на память, особенно в CPE-кластере.

Полнофункциональное ядро MPE может работать в системном или пользовательском режимах, поддерживая функции прерывания и управления памятью и реализуя внеочередное спекулятивное выполнение команд. На MPE эффективно выполняются последовательные части программ.

Ядра CPE работают только в пользовательском режиме, имеют ограниченные функции и не поддерживают прерывания. В них реализовано статическое предсказание переходов и поддерживаются команды с плавающей запятой типа «умножить и сложить» и деление / извлечение квадратного корня. CPE и MPE могут работать независимо, а SPM оснащен собственным адресным пространством, поэтому в CPE имеются команды для обмена данными между SPM и основной памятью. В CG есть общая память, но архитектура DFMC ориентируется на «кооперативную» технику, и CPE аппаратно поддерживают синхронизацию и межъядерные коммуникации, в частности передачу потоков данных и коммуникации на уровне регистров. При доступе в основную память, включая передачу потоков данных, CPE могут получить копию из кэша MPE.

Для взаимодействия ядер CPE внутри CG имеется аппаратура сети CPE_NET, которая и обеспечивает, например, коммуникации регистров и интерфейс с NoC.

Ядра MPE, занимая небольшую часть площади процессора и потребляя малую долю электроэнергии, предназначены для повышения общей производительности, а на CPE построена упрощенная микроархитектура, позволяющая получить высокую производительность для распараллеленных приложений с плавающей запятой. Для программиста независимость MPE от CPE означает, что можно сделать программу для выполнения на MPE или на кластере CPE. В целом, скорее всего, для оптимизации программ на SW26010 требуется больше ресурсов на программирование, чем в случае KNL.

Процессор SW26010 использует технологию систем на кристалле: кроме контроллеров MC, там интегрированы интерфейсы PCIe 3.0, Gigabit Ethernet и стандарта JTAG для тестирования и отладки микросхем.

1.3. ДОСТУПНЫЕ ТЕХНОЛОГИИ ПРОГРАММИРОВАНИЯ

Система программного обеспечения Sunway TaihuLight базируется на Linux-подобной операционной системе Raise OS 2.0.5 с компиляторами для языков Фортран, Си/C++ с расширениями Sunway OpenACC (применяется OpenACC 2.0 со специальными дополнениями для Sunway TaihuLight; OpenACC 2.0 поддерживается в компиляторе GCC 6). Для распараллеливания внутри CG применяется Sunway OpenACC, для четырех групп CG внутри процессора SW26010 можно применять распараллеливание на OpenMP или MPI, а между узлами — MPI. Обеспечиваются и базовые математические библиотеки, для чего применима библиотека xMath, сопоставимая с коммерческими библиотеками типа MKL и др.

1.4. ОБЛАСТЬ ЭФФЕКТИВНОГО ПРИМЕНЕНИЯ

В [4] указано, что Sunway TaihuLight ориентирован не только на традиционные HPC-вычисления, но и на обработку больших массивов данных. В Национальном суперкомпьютерном центре в Уси, где установлен Sunway TaihuLight, выполняется множество различных приложений, включая расчеты атмосферных моделей и приложения вычислительной гидродинамики, молекулярной динамики и др.

2. ПАРАМЕТРЫ СУПЕРВС

2.1. БИСЕКЦИОННАЯ ПРОПУСКНАЯ СПОСОБНОСТЬ

Бисекционная пропускная способность – это пропускная способность соединений, удаленных при разделении вычислительной сети на примерно равные части. В Sunway TaihuLight она составляет 70 Тбайт/с.

2.2. ПРОПУСКНАЯ СПОСОБОНОСТЬ СЕТИ

Пропускная способность сети – это максимально допустимая скорость обработки трафика, которая определяется стандартами сети. Она показывает, какой максимальный объем может быть передан в единицу времени. Эта величина не зависит от загруженности сети, так как отражает именно максимально возможную скорость. Фактически, пропускная способность показывает, с какой скоростью выполняются внутренние сетевые операции, такие, как передача пакетов с данными по сети через все коммуникационные узлы. Пропускная способность зависит от качеств и характеристик физической среды, то есть от наличия медного кабеля или оптического волокна. В Sunway TaihuLight пропускная способность канала сети равна 16 Гбайт/с.

2.3. ДИАМЕТР СЕТИ

Диаметр сети определяет минимальный путь, по которому проходит сообщение между двумя наиболее удаленными друг от друга узлами сети. В Sunway TaihuLight диаметр сети равен 7.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. TOP500 Supercomputer Sites [Электронный ресурс]. - Электрон. текстовые данные. - URL: https://www.top500.org/system/178764/ (Дата обращения: 20.12.2020).
2. Sunway TaihuLight [Электронный ресурс]. - Электрон. Текстовые данные. - URL: https://ru.wikipedia.org/wiki/Sunway_TaihuLight (Дата обращения: 20.12.2020).
3. Китайский процессорно-суперкомпьютерный путь [Электронный ресурс]. - Электрон. Текстовые данные. - URL: https://www.osp.ru/os/2017/01/13051592 (Дата обращения: 20.12.2020).
4. Fu H. et al. The Sunway TaihuLight supercomputer: system and applications // Science China Information Sciences. — 2016. — Т. 59. — № 7. — С. 072001.
5. Fu H. et al. The Sunway TaihuLight supercomputer: system and applications // Science China Information Sciences. — 2016. — Т. 59. — № 7. — С. 072001.
6. Meng D. et al. Hybrid Implementation and Optimization of OpenFOAM on the SW26010 Many-core Processor. 2016. [Электронный ресурс]. – Электрон. Текстовые данные. - URL: http://hpc.sjtu.edu.cn/hpcchina16_openfoam.pdf (дата обращения: 24.12.2020).