ОСНОВНЫЕ ОСОБЕННОСТИ CRAY-СИСТЕМ

Перевод: Мусенко Е.А.

Оригинал: Workload Management and Application Placement for the Cray Linux Environment, S-2496-4001, Cray Inc.

Суперкомпьютеры Cray XE и Cray XK являются системами массовой параллельной обработки (massively parallel processing - MPP). Cray объединяет как коммерческие продукты так и компоненты с открытым кодом (open source) со специально разработанными аппаратными средствами и программным обеспечением для создания системы, которая может работать в больших масштабах.

MPP-системы Cray базируются на основе Redstorm-технологий и были разработаны совместно с Cray Inc. и американского министерства энергетики Sandia National Laboratories. Системы Cray предназначены для работы с приложениями, которые требуют обработки в крупных масштабах, высокой пропускной способности сетей и сложных связей. Стандартными приложениями являются такие приложения, которые создают подробное моделирование со сложной геометрией во времени и пространстве, включающие в себя множество различных материальных компонентов. Такие длительные и ресурсоемкие приложения нуждаются в системе, которая является программируемой, масштабируемой, надежной и управляемой.

Основными особенностями Cray-систем являются:

- ✓ высокая производительность;
- ✓ масштабируемость;
- ✓ отказоустойчивость.

Масштабируемость. Cray-системы предназначены для масштабирования более 1 миллиона процессоров. Возможность масштабирования до таких размеров связано со структурой компонент системы:

- ✓ базовым компонентом является узел. Существует два типа узлов. Сервисные узлы обеспечивают поддержку функций, таких как управление пользовательской средой, обработка ввода вывода, а также запуск системы.

  Вычислительные узлы запускают пользовательские приложения. Поскольку процессоры вставляются в стандартные сокеты и заказчик будет иметь возможность обновления узлов, как только более быстрые процессоры станут доступны. Двойной разъем вычислительных узлов состоит из подмножества узлов с неоднородным доступом к памяти (non-uniform memory access - NUMA), которые определяют границы смерти процессора. Каждый NUMA-узел состоит из набора вычислительных ядер и памяти. Операции узла Inter-NUMA вне определенного вычислительного узла будут обрабатываться медленнее, чем узловые операции Intra-NUMA – это «неравномерность» доступа к памяти внутри и между узлами NUMA. На вычислительной пластине Cray XE6, которая состоит из процессоров AMD Opteron 6100 и 6200 серий, имеются два кристалла в одном контейнере. Таким образом, на одном вычислительном узле находятся четыре NUMA-узла.
- ✓ Cray-системы используют простую модель памяти. Каждый экземпляр распределенного приложения имеет свой процессор и локальную память. Удаленная память – это память на других узлах, которые запускают связанные экземпляры приложения. Тем не менее системная поддержка Cray односторонних моделей, таких как PGAS(Parallel Global Address Space) языки и DMAPP(Distributed Memory Application) API, которые позволяют программам обрабатывать пространства памяти приложения как распределенную глобальную память.
- ✓ Система взаимосвязанной сети связывает расчет и сервисные узлы. Это ресурс маршрутизации данных, который Cray-системы используют для поддержки высокой степени связи при увеличении числа узлов. Cray-системы используют полную 3D топологию сети вида тор. Но Cray XE6 использует двумерную топологию.

Отказоустойчивость. Особенности функции отказоустойчивости систем Cray:

- ✓ Узел проверки жизнеспособности (NHC – Node Health Checker) выполняет проверку вычислительных узлов на жизнеспособность, достаточную для поддержки работы конкретного приложения. Если не достаточно, то HNC удаляет все узлы, неспособные поддерживать выполнение приложения из пула ресурсов
- ✓ Инструменты, помогающие администраторам восстанавливать систему или ошибку узла, включает в себя backup-утилиту, резервное копирование, аварийного переключения узла загрузки и оперативную перезагрузку.
- ✓ Технология кода коррекции загрузки (ECC – error correction code), которая обнаруживает и исправляет ошибки хранения многоразрядных данных и их передачи.
- ✓ Lustre отказоустойчивая файловая система. В случае установки Lustre в автоматическое состояние отказоустойчивости, Lustre переключается в режим сервиса ожидания, если основной узел выходит из строя или Lustre временно находится на техническом обслуживании.
- ✓ Системные процессорные платы Cray XE имеют резервные модули регулировки напряжения (VRMs) или VRMs с резервными схемами.
- ✓ Имеется несколько резервных контроллеров, которые обеспечивают автоматическое переключение функций, и несколько портов Fibre Channel и InfiniBand-коннекторов для доступа к данным диска.

Основными программными компонентами Cray-систем являются:

- ✓ Набор инструментов для приложений, которые включают в себя:
  - Cray Application Development Environment (CADE)
    - Message Passing Toolkit (MPI, SHMEM)
    - Математические и научные библиотеки (LibSci, PETSc, ACML, FFTW, Fast_mv)

  - Инструменты для моделирования и управления данными (NetCDF, HDF5)
  - Отладчик GNU (lgdb)
  - GCC, C++ и Fortran компиляторы
  - Java (для разработки программ для сервисных узлов)
- Набор инструментов для размещения приложений:
  - Application Level Placement Scheduler (ALPS) утилиты планирования и запуска приложений
  - Режим совместимости кластера (Cluster Compatibility Mode) позволяет пользователям запускать приложения с индивидуальным системным обеспечением на Cray-системах
  - контрольные точки/перезапуск
- Дополнительно:
  - C, C++ и Fortran 95 компиляторы от PGI и PathScale
  - glibc библиотеки (подмножество вычислительных узлов)
  - Разделенное глобальное адресное пространство (PGAS) программной модели, включающей Fortran 2008 с co-массивами(co-arrays), Unified Parallel C (UPC) и Cray’s Chapel
  - Управление рабочей нагрузкой систем (PBS Professional, Moab TORQUE, платформа LSF)
  - Отладчик TotalView
  - Отладчик DDT
  - Набор инструментов для анализа производительности (CrayPAT)
  - Inter Compiler Support
  - Среда компиляции Cray (CEE)
    - Cray C и компиляторы
    - Cray C++ компиляторы
    - Fortran 2003 компилятор

    - Cray C компилятор поддерживает Unified Parallel C и Cray Fortran компилятор поддерживает со-массивы (co-arrays). Все компиляторы CCE поддерживают OpenMP
    - Инструменты для CUDA
- Cray Application Development Supplement (CADES) для Linux-платформ
- Сервисы операционной системы. Операционная система, Cray Linux Environment (CLE), адаптирована к требованиям служб и вычислительных узлов.
- Поддержка параллельных файловых систем. Cray поддерживает Lustre – параллельную файловую систему. CLE позволяет системе Cray также использовать файловые системы, такие как NFS, проецируя их на вычислительные узлы с помощью DVS (Cray Data Virtualization Services)
- Инструменты системного управления и администрирования:
    - System Management Workstation (SMW), единая точка контроля для системного администрирования
    - Hardware Supervisory System (HSS), следит за системой и описателями связанных компонент. HSS не зависит от вычислительных и сервисных аппаратно связанных компонент, она имеет свою собственную сеть
    - Comprehensive System Accounting (CSA), программный пакет, который выполняет стандартную обработку логирования. CSA имеет открытый код, включая в себя изменения в ядро Linux, так что есть возможность собрать несколько видов системных ресурсов используемых данных, чем при стандартном Fourth Berkeley Software Distribution (BSD) логировании и использует LDAP (Lightweight Directory Access Protocol)