Сетевая коммуникация в кластерах

Доклад на семинаре «высокопроизводительные вычисления»

Чуканов Вячеслав 6057/2

14.11.2011

Содержание

Кластер

- Кластер
  - Совокупность процессоров, объединенных компьютерной сетью и предназначенных для решения одной задачи, как правило, большой вычислительной сложности
- 2 класса кластеров
  - Кластеры специальной разработки с быстродействием ~Tflops
  - Кластеры, строящиеся на базе имеющихся локальных сетей из ПК
- MPI – Message Passing Interface
  - Наиболее распространенный интерфейс параллельного программирования

Компьютерные сети

- Сеть
  - Сложный комплекс взаимосвязанных и согласованно функционирующих программных и аппаратных компонентов
- Компоненты сети
  - Компьютеры
  - Коммуникационное оборудование
  - Операционные системы
  - Сетевые приложения
- Адресное пространство
  - Линейное и иерархическое

Модель OSI

- Open System Interconnection
- 7 уровней взаимодействия
  - Прикладной
  - Представительный
  - Сеансовый
  - Транспортный
  - Сетевой
  - Канальный
  - Физический

Коммуникационные протоколы

Модель OSI 1	IBM/Microsoft 2	 3	TCP/IP 4	Novell 5
Прикладной	SMB, MPICH	MPICH	Telnet, FTP, MPICH и др.	NSP, SAP
Представительный				
Сеансовый	NetBIOS	NBT		
Транспортный		TCP	TCP	SPX
Сетевой	NBF	IP	IP, RIP, OSPF	IPX, RIP, NLSP
Канальный	Протоколы Ethernet, Fast Ethernet, Token Ring, ATM, X25 и др.			
Физический	Коаксиал, витая пара, оптоволокно			

Типы коммуникаций

- SCI
- Ethernet
- Myrinet
- QsNet
- Infiniband

…

SCI

- Scalable Coherent Interface
- Производитель:
  - Dolphin Interconnect Solutions, Норвегия
  - ОАО «НИЦЭВТ», Россия
- Топология:
  - кольцо, 2-х/3-х мерный тор
  - коммутируемые кольца
- Пропускная способность
  - физическая скорость передачи — 667 Мб/сек
  - на уровне MPI — от 200 до 325 Мб/сек
- Программное обеспечение:
  - драйверы для Linux, Windows NT, Solaris
  - ScaMPI – коммерческое ПО от Scali Computer
  - SISCI API - интерфейс нижнего уровня от Dolphin

Gigabit Ethernet

- Gigabit Ethernet
  - **1000BASE-SX**, IEEE 802.3z
    - Многомодовое волокно
    - Дальность прохождения сигнала без повторителя до 550 метров
  - **1000BASE-LX**, IEEE 802.3z
    - Одномодовое волокно.
    - Дальность прохождения сигнала без повторителя до 5 километров
- 10-гигабитный Ethernet
- 100-гигабитный Ethernet
- Терабитный Ethernet

Myrinet

- Myrinet (ANSI/VITA 26-1998)
  - 28% кластерных установок на 2005 год
  - 2% кластерных установок на 2009 год
- Производитель:
  - компания Myricom
- Топология:
  - коммутируемая (матрица 8x8)
  - коммутаторы поддерживают до 128 портов
  - Fat Tree
- Программное обеспечение:
  - низкоуровневый API GM, MPICH/GM, PVM/GM, стек TCP/IP
  - коммерческие продукты — MPIPro от Scali

QsNet

- Производитель:
  - Quadrics Ltd.
- Топология:
  - Fat Tree до 1024 узлов (QsNet I)
  - Fat Tree до 4096 узлов (QsNet II)
- Программное обеспечение:
  - под Linux распространяется с исходными текстами по лицензии GNU GPL
  - поддерживает MPI (специализированную версию MPICH) и TCP/IP.
- Пропускная способность на уровне MPI около 900 МБ/сек
- Время задержки 3 мкс

Infiniband

- InfiniBand Trade Association (~40 компаний, включая IBM, Intel, Sun)
- Infiniband использует двунаправленную последовательную шину (как PCI-e)
- 3 уровня производительности
  - 10 GB/s
  - 20 GB/s
  - 40 GB/s
- Свободная топология
- Независимые виртуальные полосы
  - До 16 полос на соединение
  - Отдельный контроль пропускной способности для каждой полосы
- Используется многими протоколами и API
  - RDMA (Remote Direct Memory Access)
  - IPoIB (IP over Infiniband)

Производительность

Сеть	Скорость передачи данных	Время задержки, мкс
SCI	физ. скорость – 667 Мб/с на уровне MPI – 200-325 Мб/с	2-3 4
QsNet	на уровне MPI – 900 Мб/с	3
Myrinet	на уровне MPI – 250 Мб/с	10
Ethernet	физ. скорость до 1000 Мб/с на уровне MPI – 500 Мб/с	50
Infiniband	на уровне MPI – 800 Мб/с	5-7

Литература

- Интернет-энциклопедия www.wikipedia.ru
- Г. И. Шпаковский, А. Е. Верхотуров, Н. В. Серикова “Руководство по работе на вычислительном кластере”
- Официальный сайт InfiniBand Trade Association http://www.infinibandta.org/