Технология InfiniBand: архитектура, текущее состояние и перспективы развития

Андрей Слепухин
Т-Платформы

andrey.slepuhin@t-platforms.ru *http://www.t-platforms.ru*

Немного истории...

- 1999, август – слияние FIO и NGIO
- 2000, октябрь – версия 1.0 спецификации
- 2002 – первые демонстрации решений
- 2003 – широкий спектр решений в массовой продаже
- Конец 2003 – 12x12 4x кроссбар на одном чипе, поддержка PCI-Express
- 2004, сентябрь, версия 1.2 спецификации

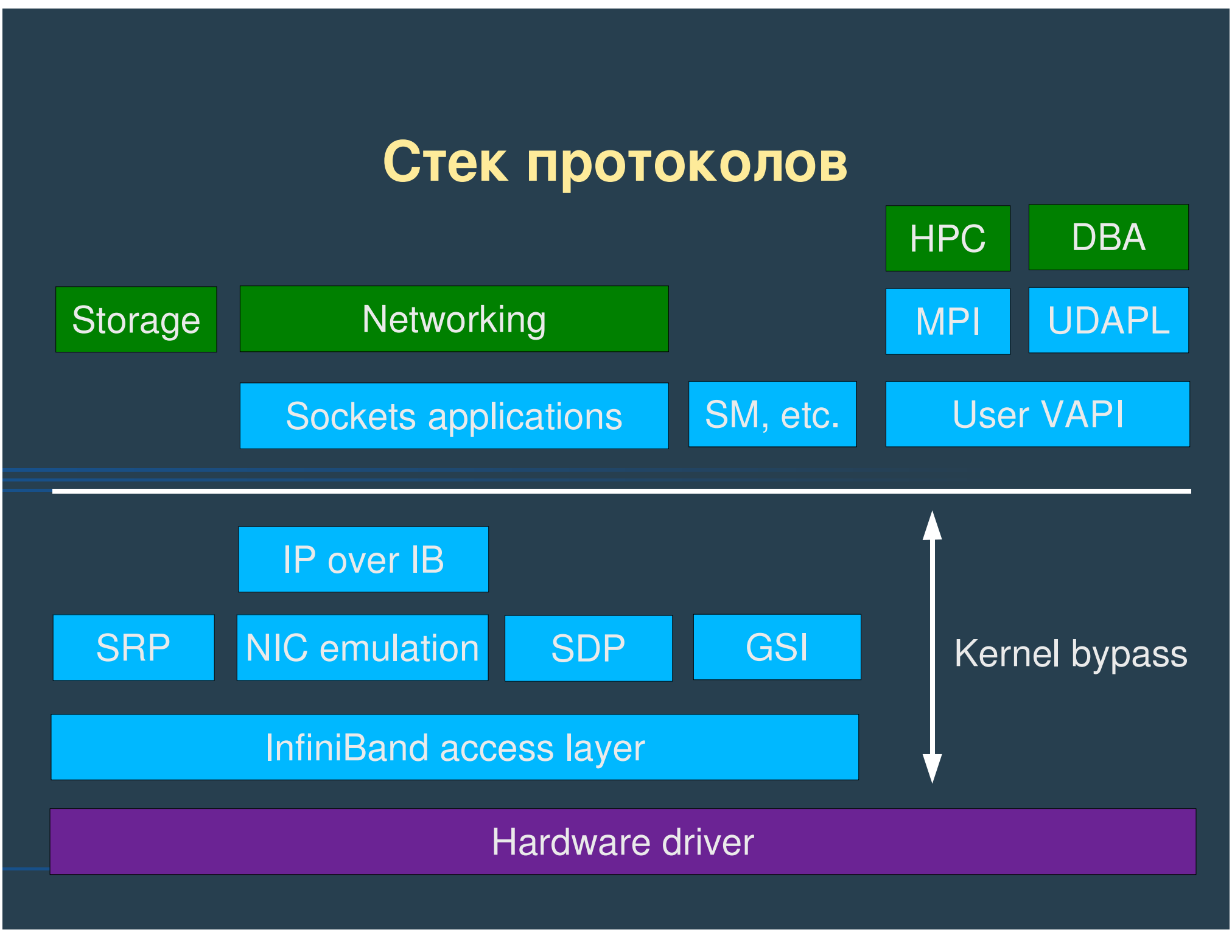
Архитектура

- QP (queue pair)-based
  - SQ
  - RQ
  - CQ
- Поддерживаемые операции:
  - Send/Receive
  - RDMA Read/Write
  - Atomic
  - Multicast

Архитектура (2)

- Поддерживаемые виды сервиса:
  - Reliable Connection/Datagram
  - Unreliable Connection/Datagram
  - Raw Datagram
- Virtual Lanes
  - 15 Vlanes + Management Lane
- Service Level
- Credit-based flow control
- Congestion avoidance

Стек протоколов



Реализации MPI

- MPICH
  - OSU-MVAPICH (университет Огайо)
  - MPICH-VMI (NCSA)
- LAM
- Scali MPI
- LA-MPI (Лос-Аламос)
- MPI/Pro (Verari)

Производительность

- Пропускная способность:
  - PCI-X – ~800Mb/sec
  - PCI-Express – 971Mb/sec (unidirectional), 1878Mb/sec (bidirectional)
  - PCI-Express (dual ports) – 1495Mb/sec (unidirectional), 2720Mb/sec (bidirectional)
- Задержка:
  - PCI-X – ~5.2usec
  - PCI-Express – ~4.0usec

Производители чипов

- Mellanox
  - InfiniBridge
  - InfiniScale
  - InfiniHost
  - InfiniScale III
  - InfiniHost III Ex
- Intel
- Agilent
- Fujitsu
- IBM – ???

Производители оборудования

- Mellanox – PCI-X, PCI-Express адаптеры, коммутаторы до 144 портов
- InfiniCon – PCI-X адаптеры, коммутаторы до 288 портов, IBtoFC, IBtoGbE
- Voltaire – PCI-X адаптеры, коммутаторы до 288 портов, IBtoFC, IBtoGbE
- Topspin – PCI-X, PCI-Express адаптеры, коммутаторы до 96 портов, IBtoFC, IBtoGbE

Кластер СКИФ К-1000

- 8 стоек, 288 узлов,
  36 коммутаторов
- 576 процессоров
  AMD Opteron 2.2GHz
- Суммарный объем памяти
  1152Gb
- Производительность
  - Пиковая: 2534.4 Gflops
  - Linpack: 2032 Gflops
    (80.1% от пиковой)
- Потребляемая мощность 89kW

Перспективы

- DDR и QDR соединения (5 и 10 Gbit/sec на одну пару), пропускная способность до 120 Gbit/sec
- Уменьшение задержки до ~1usec
- Интеграция с существующими серверными решениями
- InfiniBand direct-attached storage