МЦСТ ЭЛЬБРУС

Организация сетевого взаимодействия в вычислительных кластерах семейства «Эльбрус»

Авторы: *Белянин Игорь Валерьевич*
Петраков Павел Юрьевич

Докладчик: Белянин Игорь Валерьевич

АО «МЦСТ»

9 февраля 2016

О компании МЦСТ

- Разработка ключевых компонентов компьютера
  - Процессоры
  - Контроллеры
  - BIOS
  - Ядро ОС
- Собственные средства разработки
- Доверенная платформа
  - Собственная архитектура процессоров и юж. моста
  - Исходные коды на всё используемое ПО
  - Verilog на всю используемую аппаратуру

МП Эльбрус-4С

Высокопроизводительный МП

- **Ядро:** 4 ядра улучшенной архитектуры Эльбрус
  - ✓ Поддержка 64-битной многопоточной двоичной трансляции
- **Тактовая частота:** 800 МГц
- **Производительность:** 50 GFlops
- **Кэш-память:** L2 4 * 2 МБ
- **Встроенные интерфейсы:**
- ✓ Память: 3 * DDR3-1600, до 96 ГБ
- ✓ Ввод-вывод: 2 x IOLink, 2 x 1 ГБ/с в одну сторону
- ✓ 3 когерентных межпроцессорных линка 12ГБ/с
- **Рассеиваемая мощность:** 45 Вт
- **Количество транзисторов:** 960 млн
- **Технология:** 65 нм, 9 слоев металла
- **Площадь кристалла:** 380 мм^2

МЦСТ ЭЛЬБРУС

Стоечный сервер

Типовой сервер высотой 1U/2U

ИНЭУМ

- **Процессоры:** 4 * Эльбрус-4С
- **ОЗУ:** 48 Гбайт (до 384 ГБ)
- **Видеоконтроллер:** SM 718
- **Интерфейсы:**
  - ✓ PCI Express 1.0 x8 (2 канала)
  - ✓ PCI 32 бит, 33 МГц (2 канала)
  - ✓ Gigabit Ethernet (2 канала)
  - ✓ SATA2 (8 каналов)
  - ✓ IOLink (2 канала)
  - ✓ USB 2.0 (4 канала)
  - ✓ Audio, RS-232
  - ✓ Разъём для модуля IPMI
- **Охлаждение:** воздушное
- **Конструктив:** Стоечный 19”
- **Группа исполнения:** 1.1

МЦСТ ЭЛЬБРУС

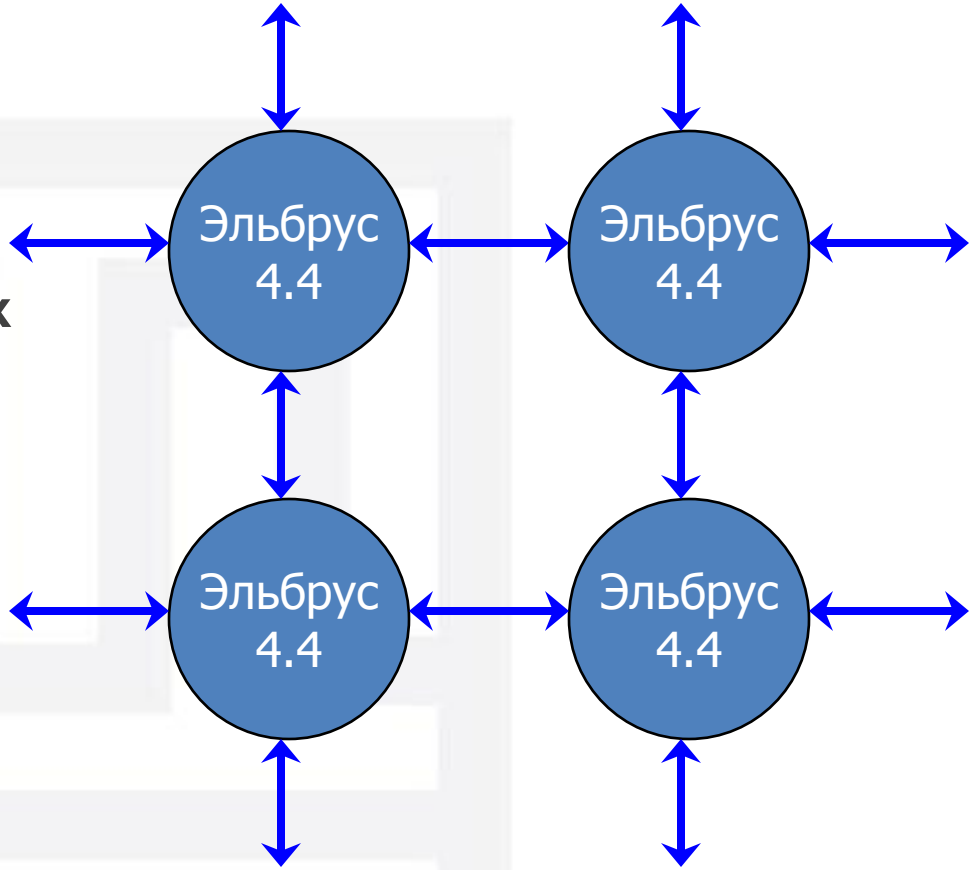
Вычислительный кластер

ИНЭУМ

- **Процессоры:** Эльбрус-4С
- Серверы: Эльбрус-4.4 (1U)
- Евромеханический шкаф 47 U – 2 шт;
- Количество серверов – до 64
- Количество процессоров – до 256
- Объем дисковой памяти – до 1.5 ПБайт
- Система охлаждения - воздушная
- Потребляемая мощность – 20 кВт
- Производительность – 13,8 Тфлопс
- Контроллер межмашинного обмена (разработан в ИНЭУМ)

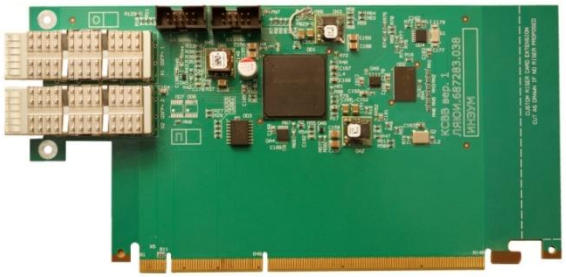
Коммуникационная сеть

- **Количество узлов:** до 256
- **Топология: 2D/3D тор**
- **Отсутствие дополнительных устройств для организации сети(маршрутизаторов)**
- **Отсутствие импортных ограничений на компоненты**
- **Соединения:** активный оптический кабель
- **Цена**



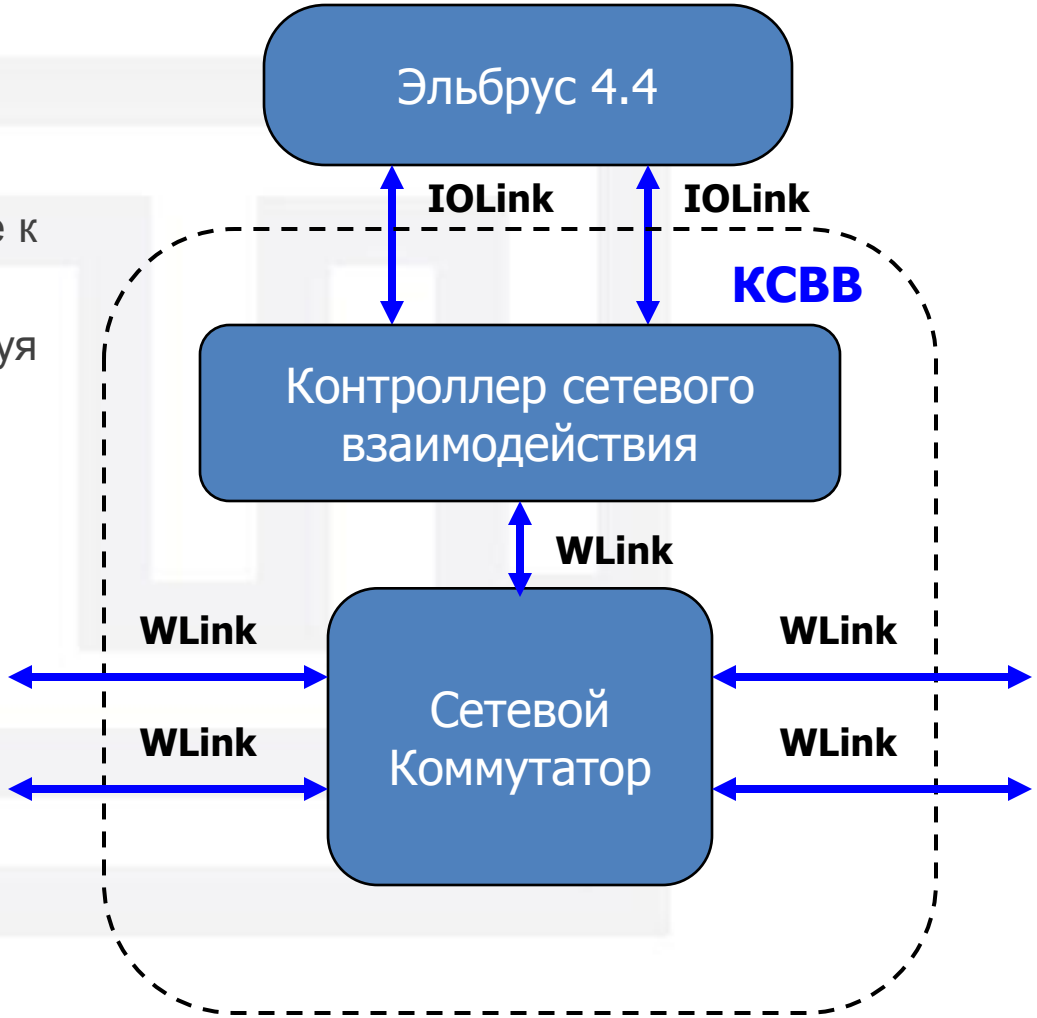
Контроллер скоростного ввода/вывода

- **Подключение:** напрямую к процессорам
- **Встроенный сетевой коммутатор**
- **Системный интерфейс:** до 3 x IOLink (2 в текущей реализации)
- **Пропускная способность системных интерфейсов:** до 1 ГБ/с каждый
- **Количество внешних линков:** до 8 (4 в текущей реализации)
- **Разъем:** QSFP+(«гидра»)
- **Пропускная способность внешних линков:** до 5 Гб/с
- **Реализация на ПЛИС:** Altera Cyclone V



Общая схема

- **Особенности:**
  - ✓ Одновременное подключение к двум процессорам (ccNUMA)
  - ✓ Подключение напрямую (минуя южный мост)
  - ✓ Используются встроенные в ПЛИС трансиверы
  - ✓ Собственный протокол
  - ✓ Собственные контроллеры и интерфейсы линков

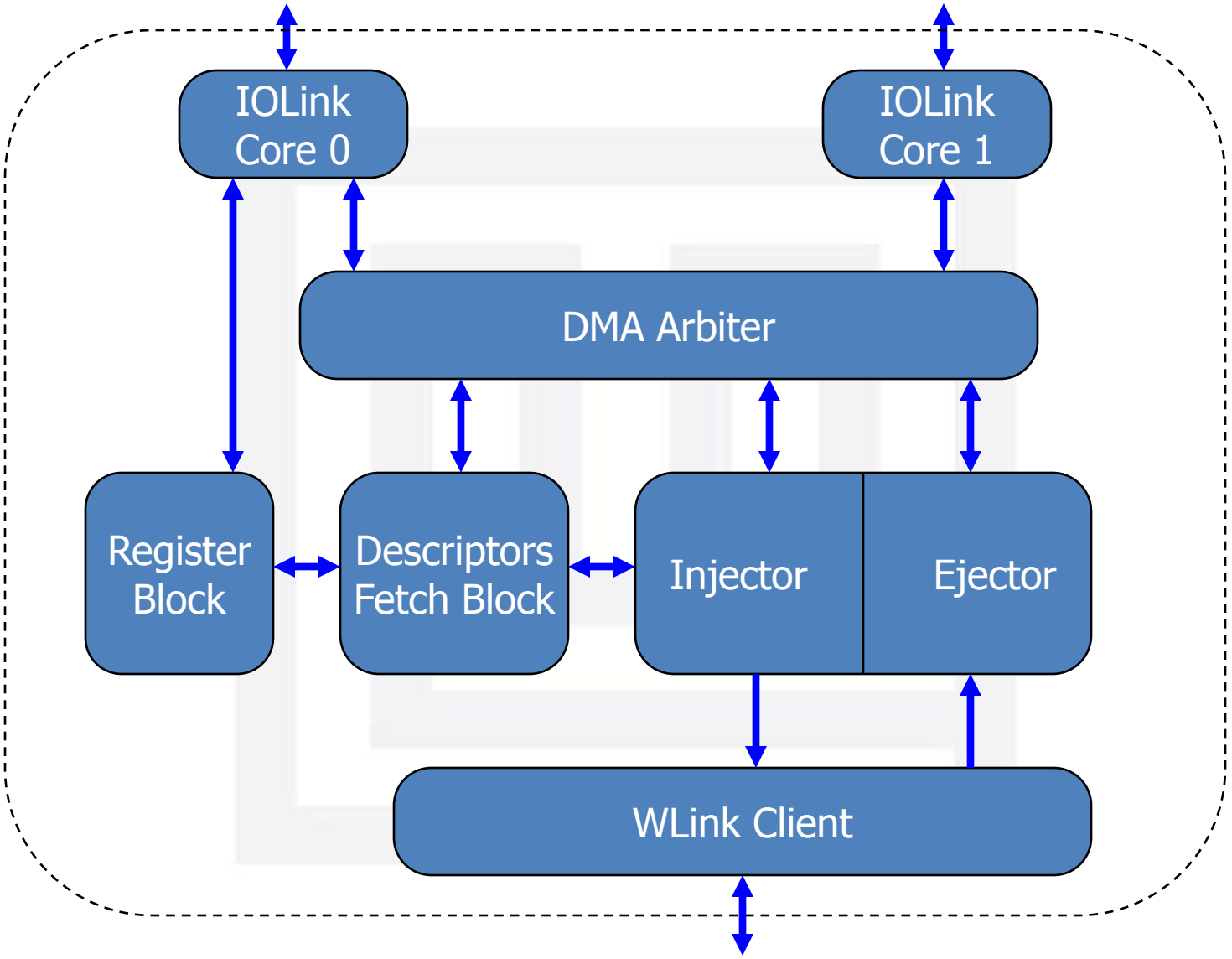


Контроллер сетевого взаимодействия

- **Программная модель:** PCI - устройство
- **До 32-х независимых блока программных ресурсов**
- **Блок программных ресурсов:**
  - ✓ Кольцевая очередь на передачу
  - ✓ Кольцевые очередь на приём
  - ✓ Блок для передачи без использования DMA
  - ✓ Копии регистров очередей в памяти
  - ✓ Блок управления прерываниями
- **Блок «общих» сетевых регистров**
- **Поддерживаемые типы обменов:**
  - ✓ Put/Get – запись/чтение из памяти удаленного узла
  - ✓ Msg – сообщение удаленному узлу
  - ✓ Doorbell – короткое сообщение удаленному узлу (удаленное прерывание)
  - ✓ Maintenance – операции доступа в сетевые регистры узлов и сетевых коммутаторов

МЦСТ ЭЛЬБРУС

Контроллер сетевого взаимодействия

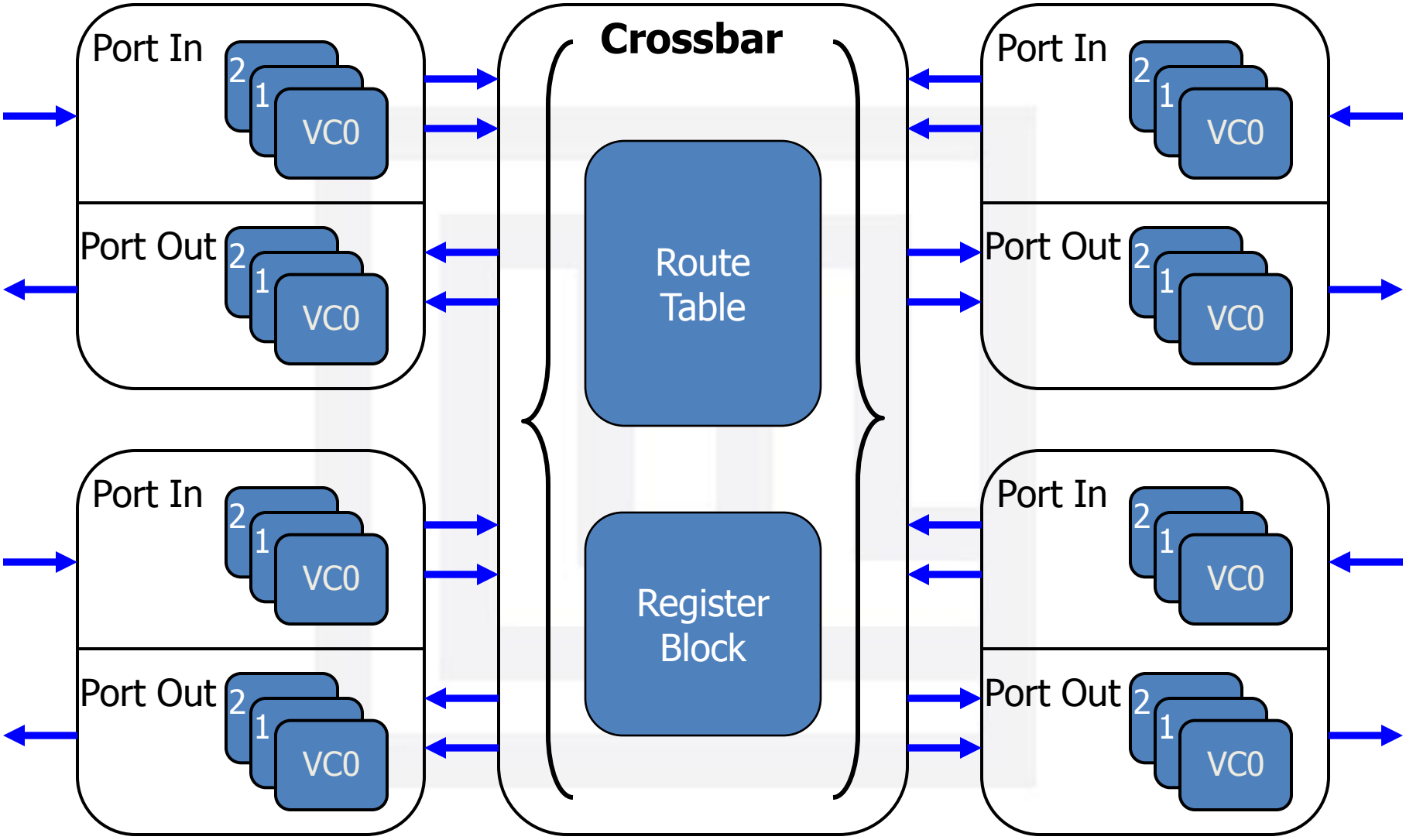


МЦСТ ЭЛЬБРУС

Сетевой коммутатор

- **Маршрутизация сетевых обменов**
- **Инициализация сети (построение таблиц маршрутизации):**
  - ✓ Ручная
  - ✓ Автоматическая
- **Типы маршрутизации:**
  - ✓ Детерминированная
  - ✓ Адаптивная
- **Буферы «Store-and-Forward»**
- **Поддержка виртуальных каналов**
- **Масштабируемость:**
  - ✓ Количество портов
  - ✓ Количество виртуальных каналов
  - ✓ Ширина внутренних шин

Сетевой коммутатор



Программное обеспечение

- **Драйвер программного интерфейса**
- **Сетевой драйвер (TCP/IP)**
- **Поддержка MPI (MPICH 3.1.4)**
- **Особенности реализации MPI:**
  - ✓ OS-bypass
  - ✓ RDMA операции
  - ✓ Совмещение вычислений и передачи сообщений

Заключение

- **Планы работ:**
  - ✓ Оценка эффективности сети на реальных задачах
  - ✓ Увеличение пропускной способности сети
  - ✓ Изменение топологии: 3D tor
- **Оптимизации:**
  - ✓ Уменьшение latency
  - ✓ Уменьшение количества стадий хранения сетевого обмена (уход от использования «Store-and-Forward»)
  - ✓ Оптимизация ПО с учетом особенностей архитектуры «Эльбрус»

МЦСТ ЭЛЬБРУС

Спасибо за внимание!

Вопросы?

Докладчик: Белянин Игорь Валерьевич

09 февраля 2015

Контакты:

Тел: +7-916-024-96-17, e-mail: igor.v.belyanin@mcst.ru