

«Как собрать кластер самому?»

mini HOWTO

Зачем нужен кластер?

- Что считать?
 - Определить перечень приложений, класс задач, алгоритмов, методов
- Как считать?
Классифицировать потребности
 - По процессорам (сколько, каких)
 - По памяти (сколько, какая)
 - По спецвычислителям (нужны ли, какие)

«Устаревший» способ построения кластера

- Проект Beowulf (www.beowulf.org). 1993 год!
- Построение кластера из доступных на рынке компонент.
- Компоненты:
 - Набор системных блоков (рабочих станций)
 - Стандартные коммутаторы
 - Базовое ПО:
 - Linux kernel (+ channel-bonding драйвер)
 - MPI (LAM, MPICH)



1998 год. Суперкомпьютер Avalon (Los Alamos National Lab)

313 000\$ (~50 GFLOPS; CPU: 140 x DEC/Alpha 533MHz)

112 место в TOP500

Рядом в TOP500 коммерческая система: IBM SP2 (Scalable Power System 2)

(152 процессора POWER)



Технология лезвий

- Патент датируется 2002 годом (US patent 6411506, Christopher Hipp & David Kirkeby, "High density web server chassis system and method", published 2002-06-25, issued 2002-06-25, assigned to RLX Technologies)
- Бывшие сотрудники компании Compaq, позже RLX Technologies была куплена Hewlett Packard.



Технология лезвий

- **Блейд-серверы, лезвия (англ. blade)** — компьютерные серверы с компонентами, вынесенными и обобщёнными в корзину для уменьшения занимаемого пространства.
Корзина (англ. enclosure) — шасси для блейд-серверов, предоставляющая им доступ к общим компонентам, например, блокам питания и сетевым контроллерам.
- Для вычислений компьютеру требуются как минимум следующие части (машина Тьюринга):
 - память, содержащая исходные данные,
 - процессор, выполняющий команды,
 - память для записи результатов.
- Остальные компоненты, типичные для компьютера, выполняют вспомогательные для вычислений функции, такие как ввод и вывод, обеспечение питания.

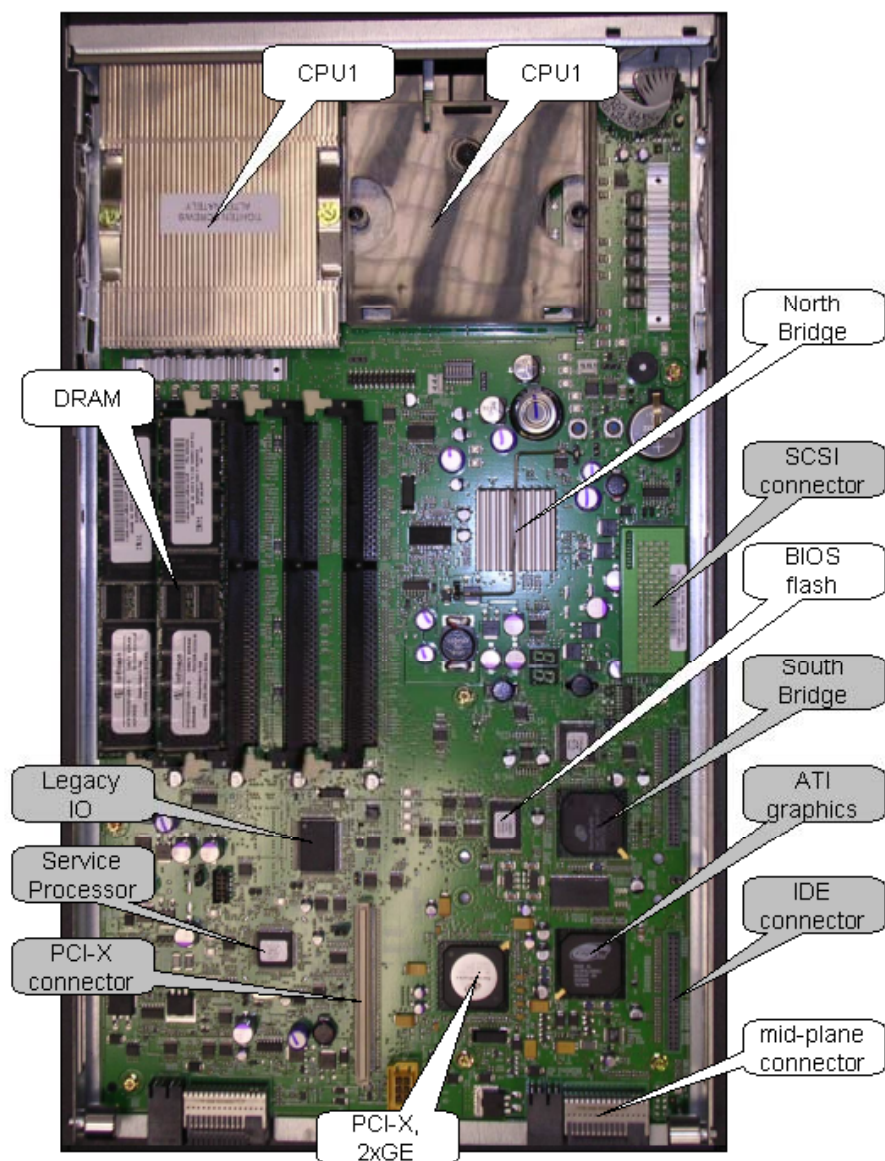
Технология лезвий

- **Преимущества:**
 - Компактность (2-ух — 3-ех кратное)
 - Гибкость (заменяемые компоненты)
 - Централизация (управление)
- **Недостатки («с натяжкой»!):**
 - Охлаждение
 - Уникальность
 - Цена заметно выше традиционных серверов
 - ... ?? (вероятно больше нет недостатков)

IBM JS20 Blade Server: 2 * PowerPC 970 2.2GHz
(пиковая = 17,8 GFLOPS)

Всего: 84 узла (пиковая = 1478,4 GFLOPS)

Размер: стойка 42U

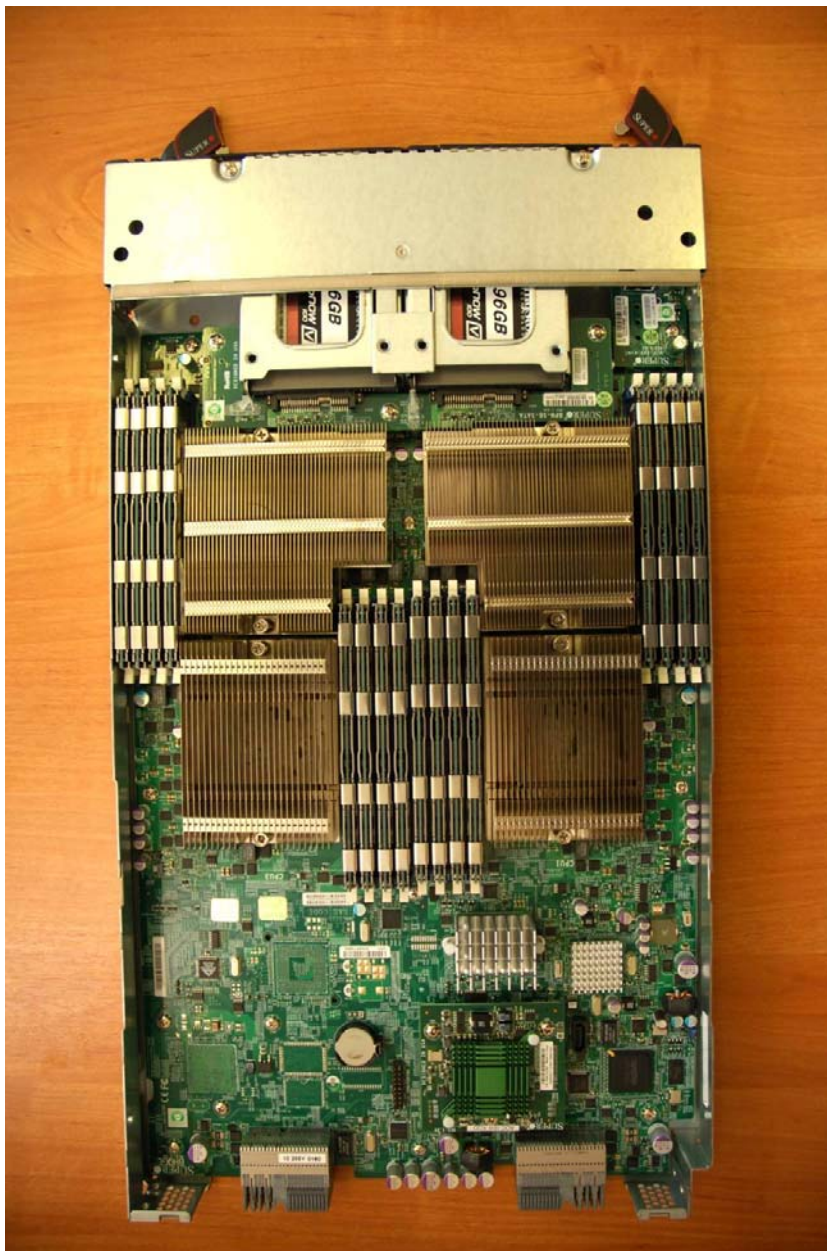


Лезвия кластера MBC15K



Лезвия SMH11

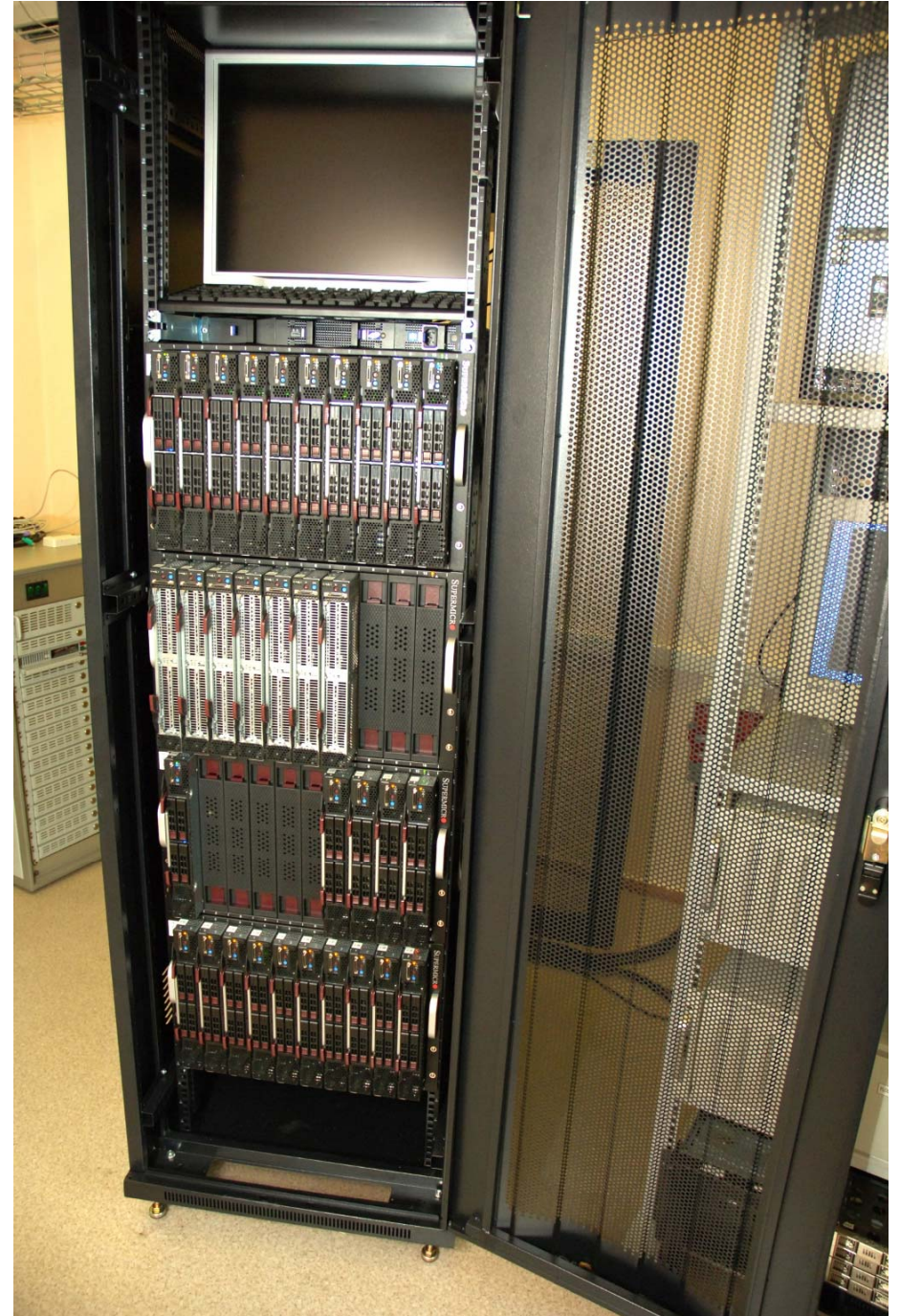
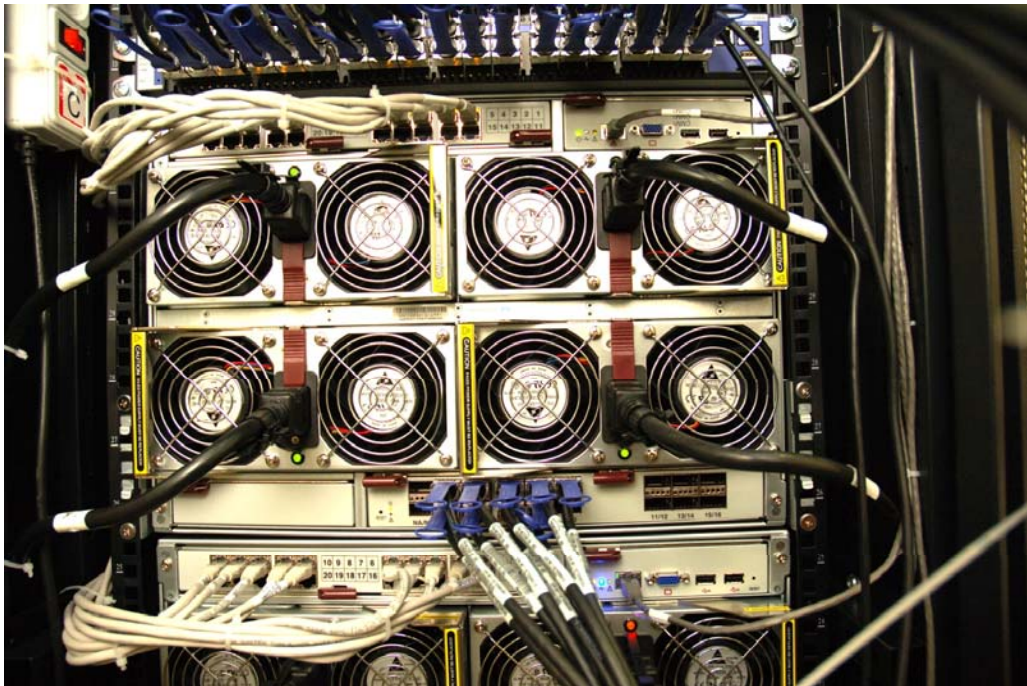
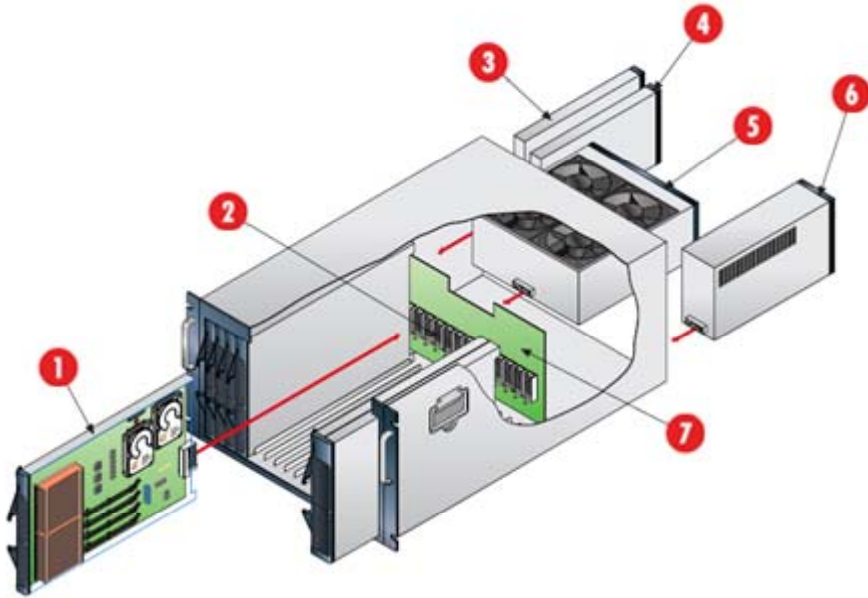
4 x Opteron (64/128, 100/400)



2 x Xeon (32, 32, 2 * Tesla S2050)



Шасси (enclosure) SMH11



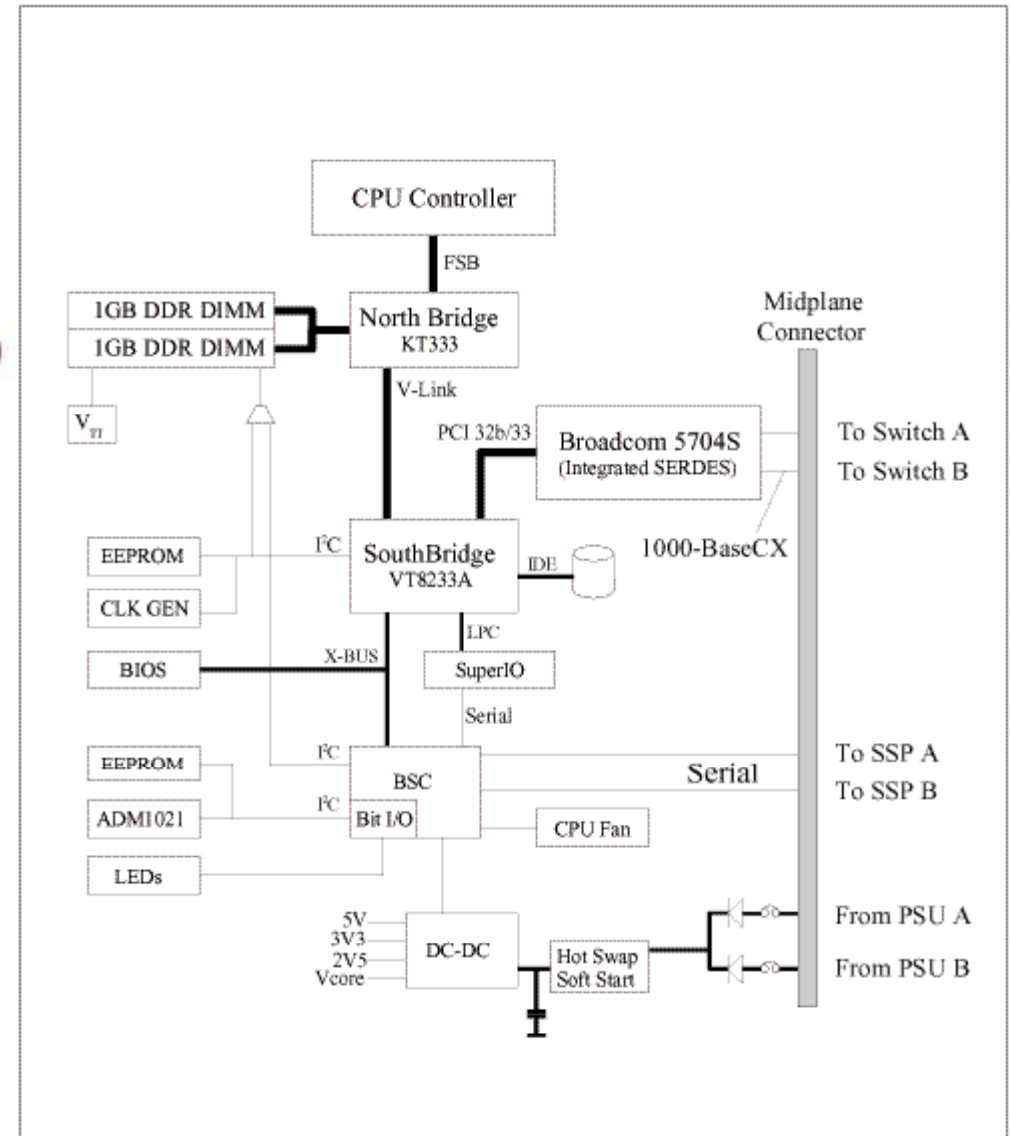
Компоненты блейд-системы

- Питание

- MBC15K (IBM): 4 * 2000W, резерв.: **N + N**
- SMH11 (Supermicro): 4 * 2500W, резерв.: **N + 1**



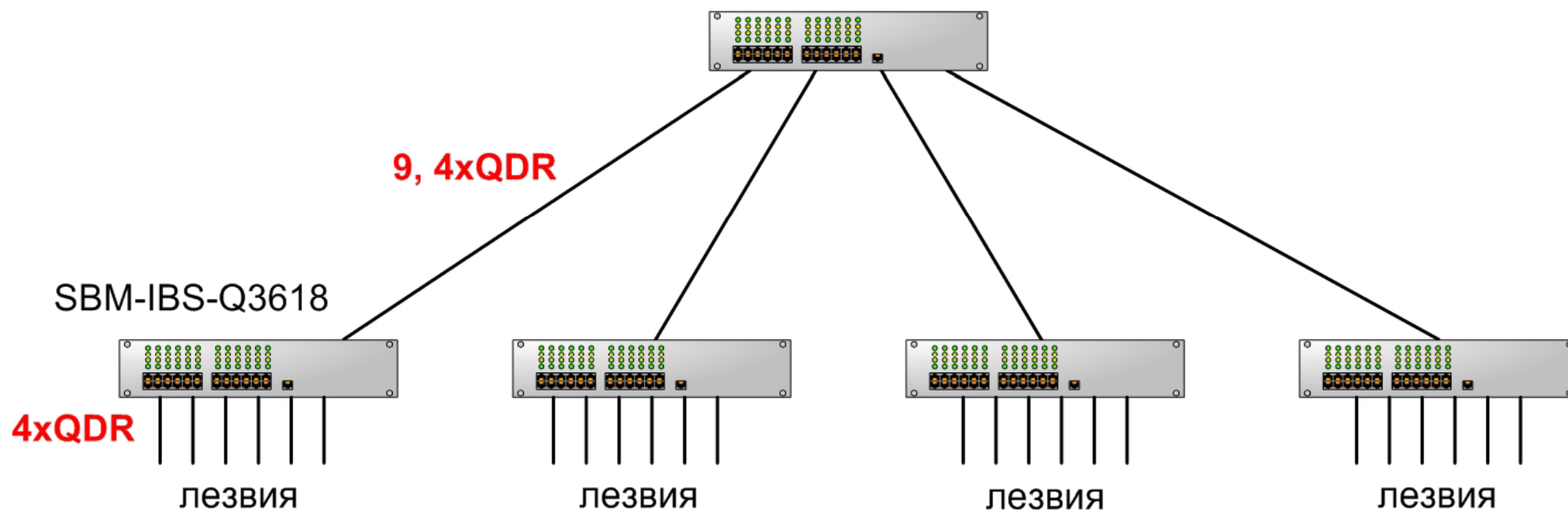
- Сеть
 - path-through (проброс порта)
 - switch (коммутатор)



Infiniband Fabric SMH11

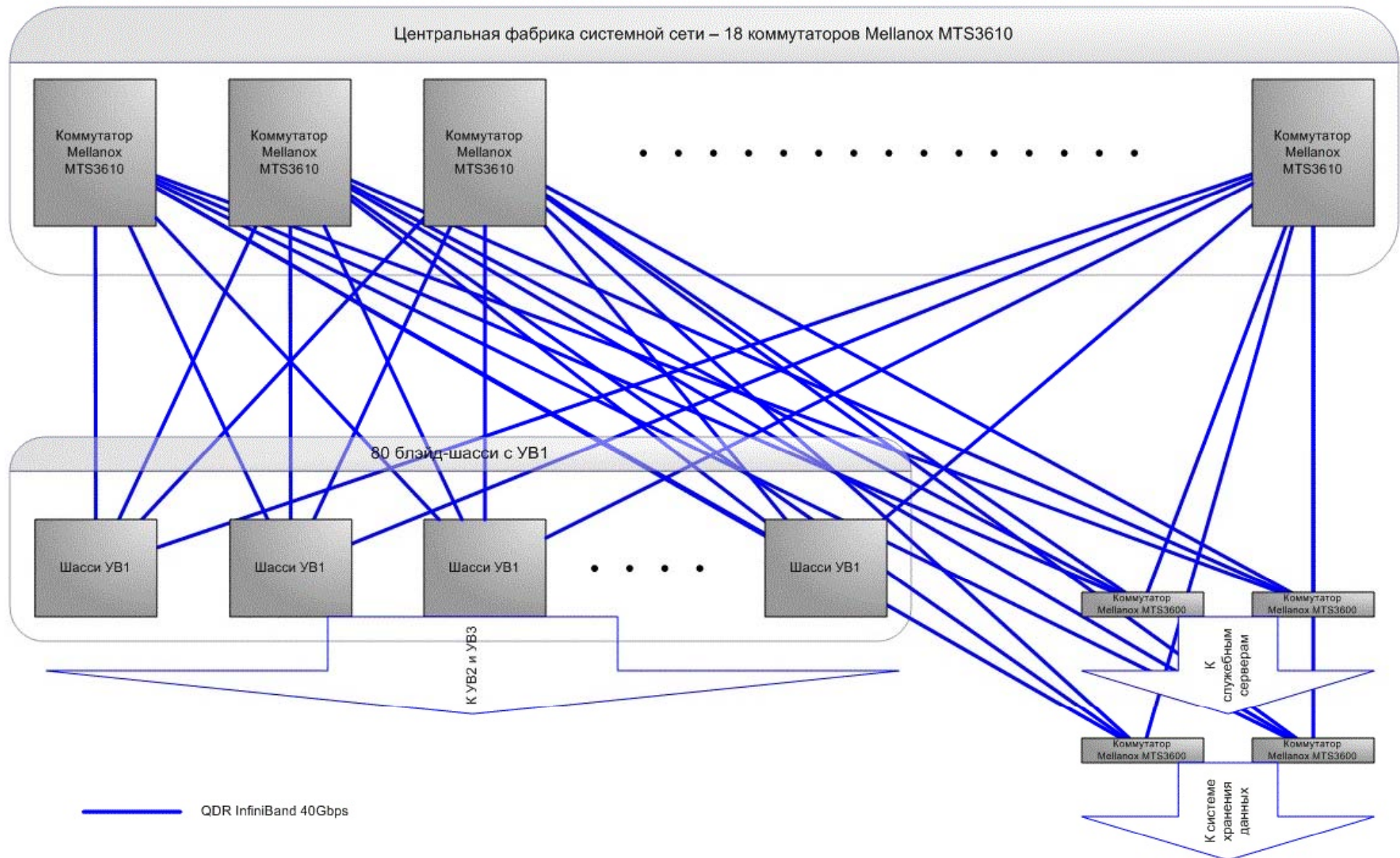


Switch Mellanox 3618
36-port 4x QDR

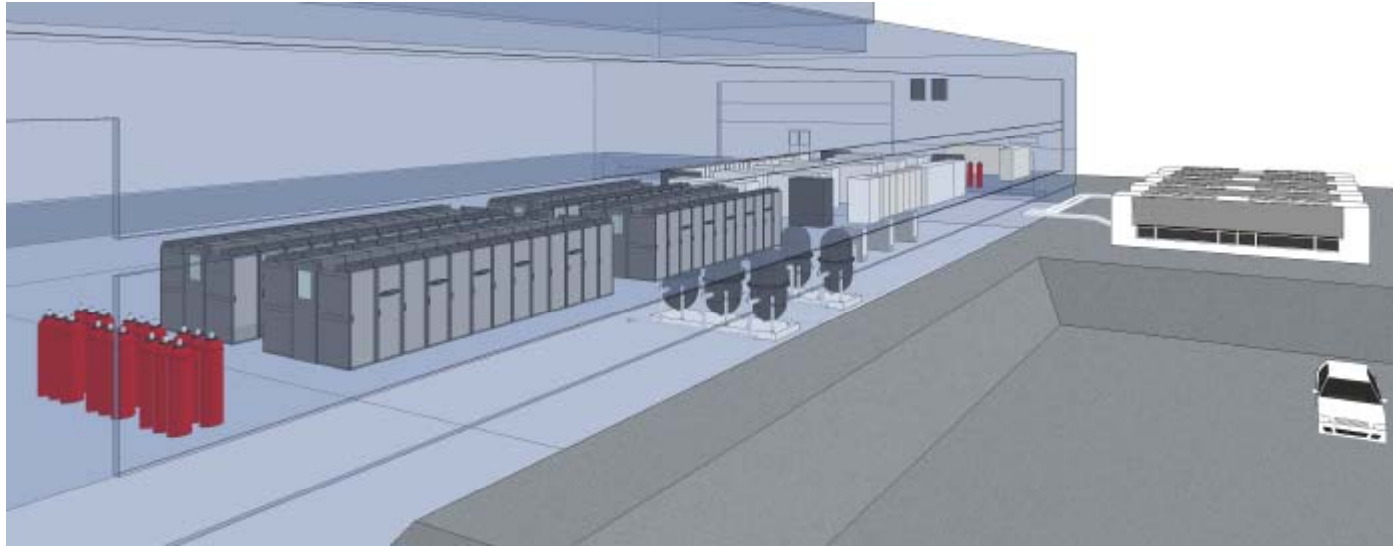


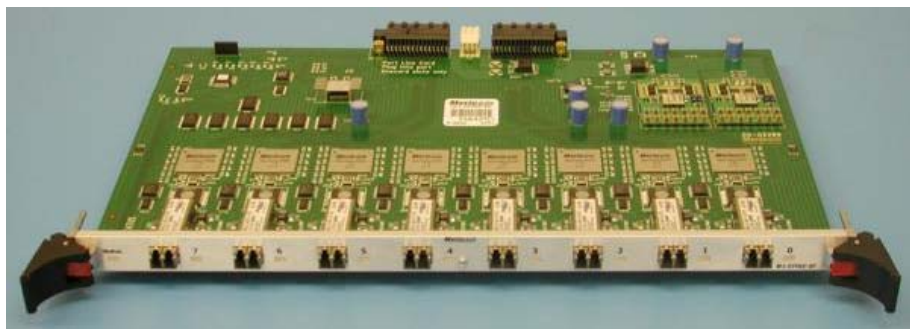
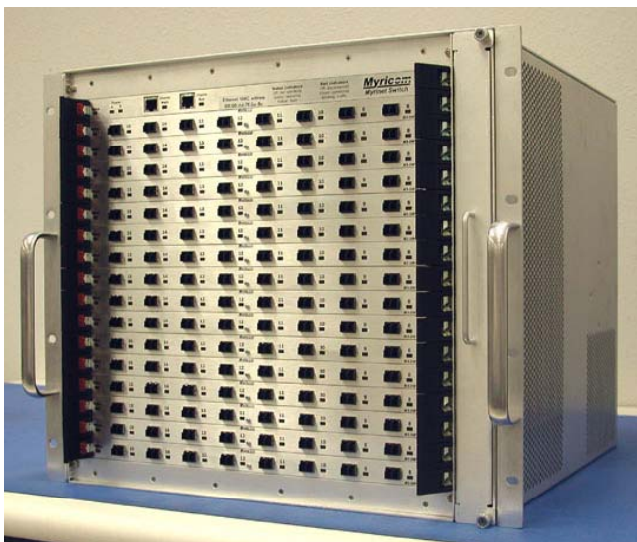
Infiniband Fabric МГУ «Ломоносов»

Топология системной сети

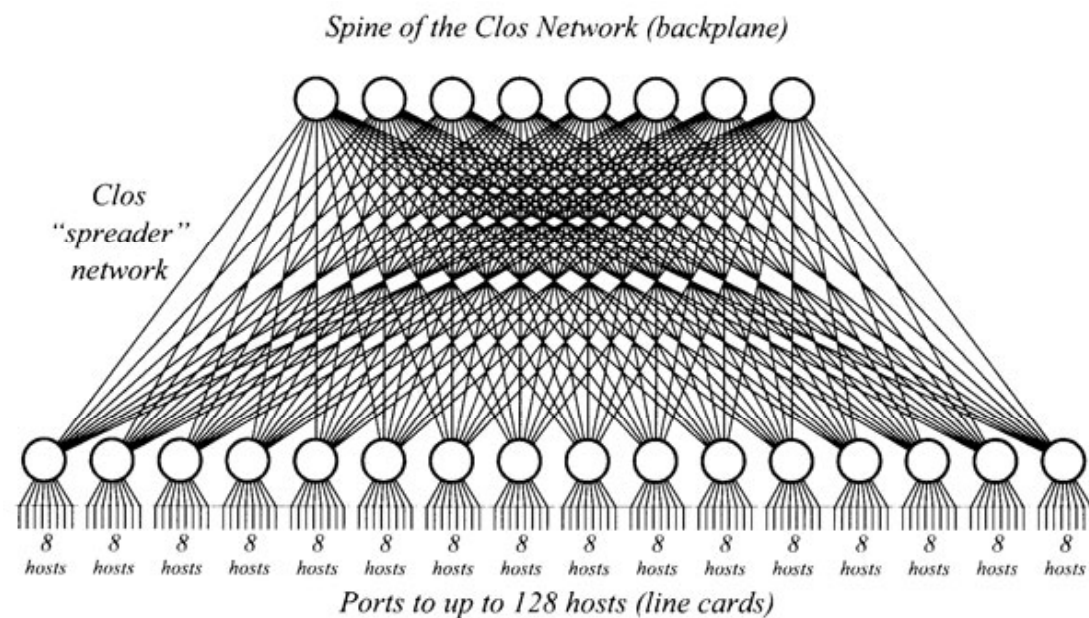
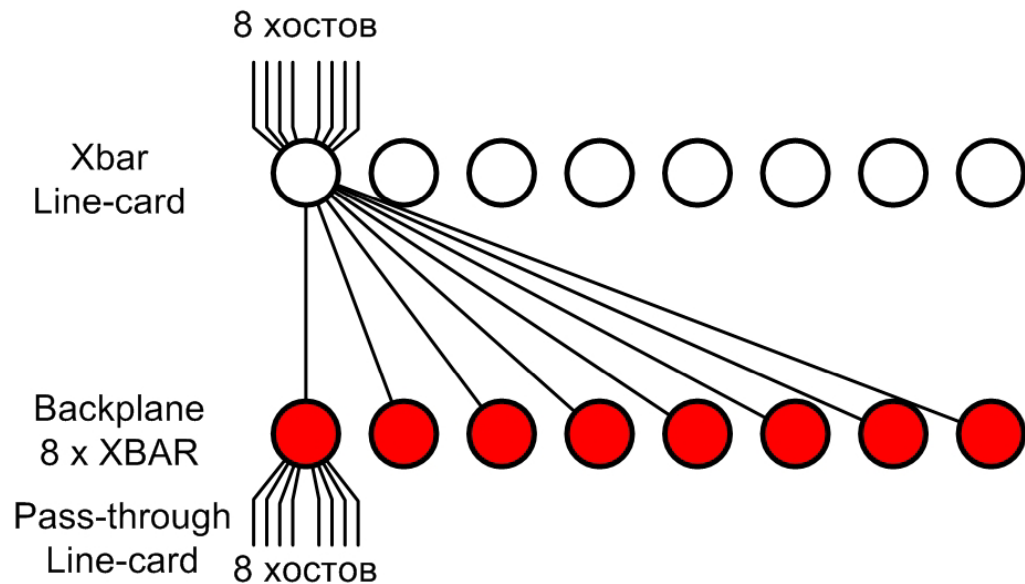


Кластер МГУ «Ломоносов»





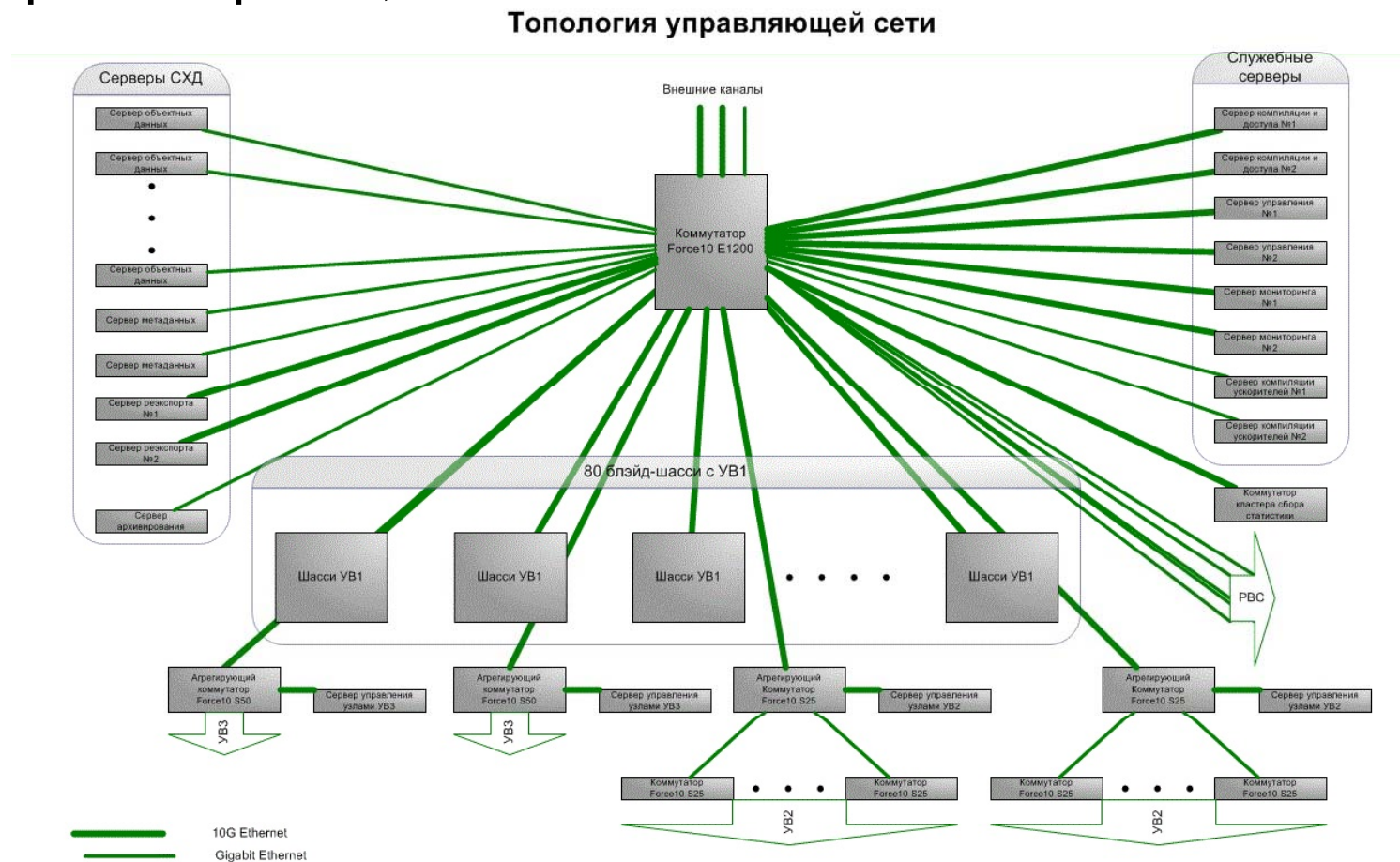
Myrinet CLOS MBC15K



(Gigabit, 10G) Ethernet

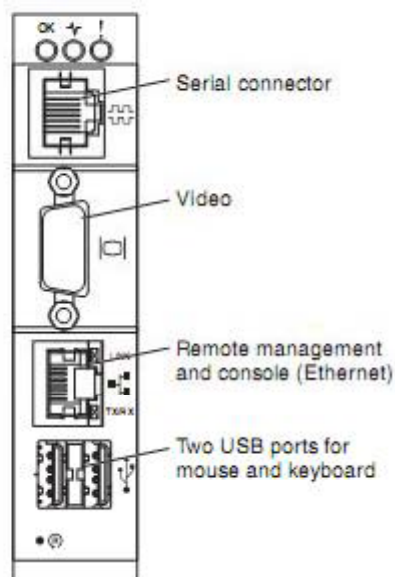
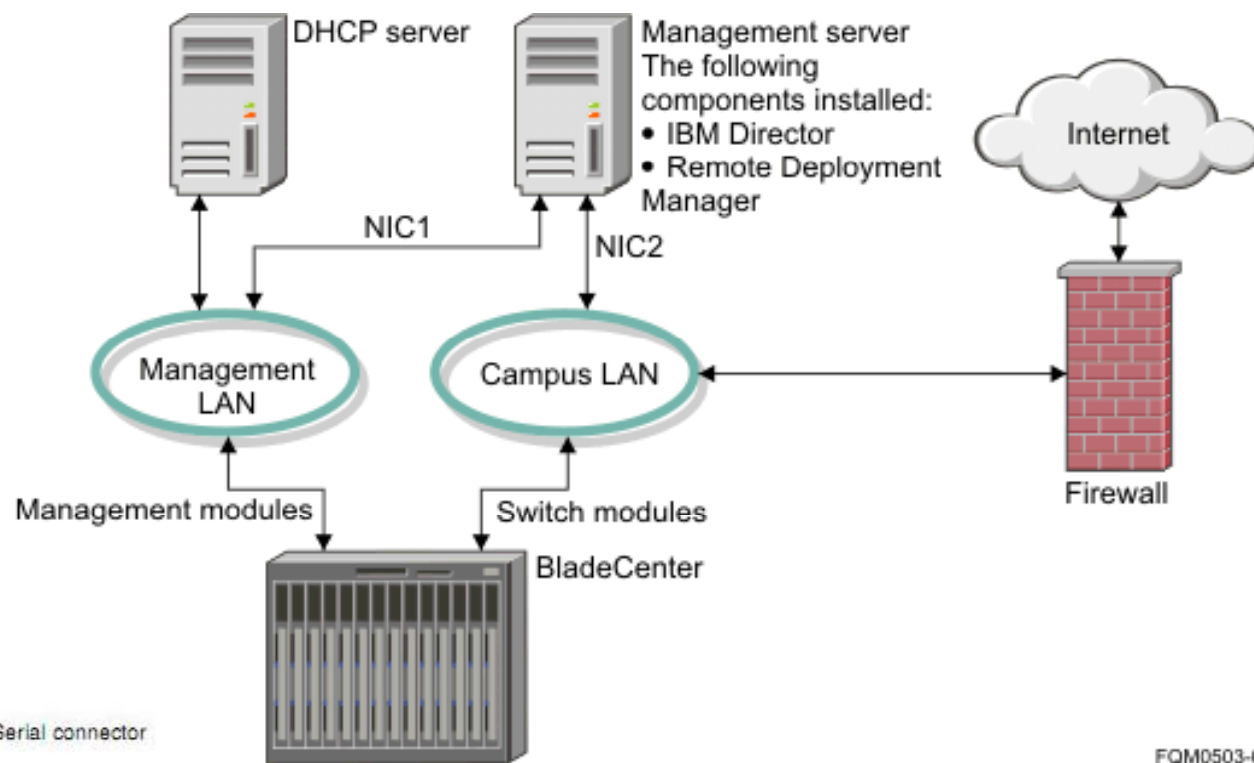
- Дерево коммутаторов:

- Корень: магистральный коммутатор или стек магистральных коммутаторов;
- Ветки: агрегирующие коммутаторы;
- Листья: лезвия, соединенные с коммутатором на шасси через Midplane;

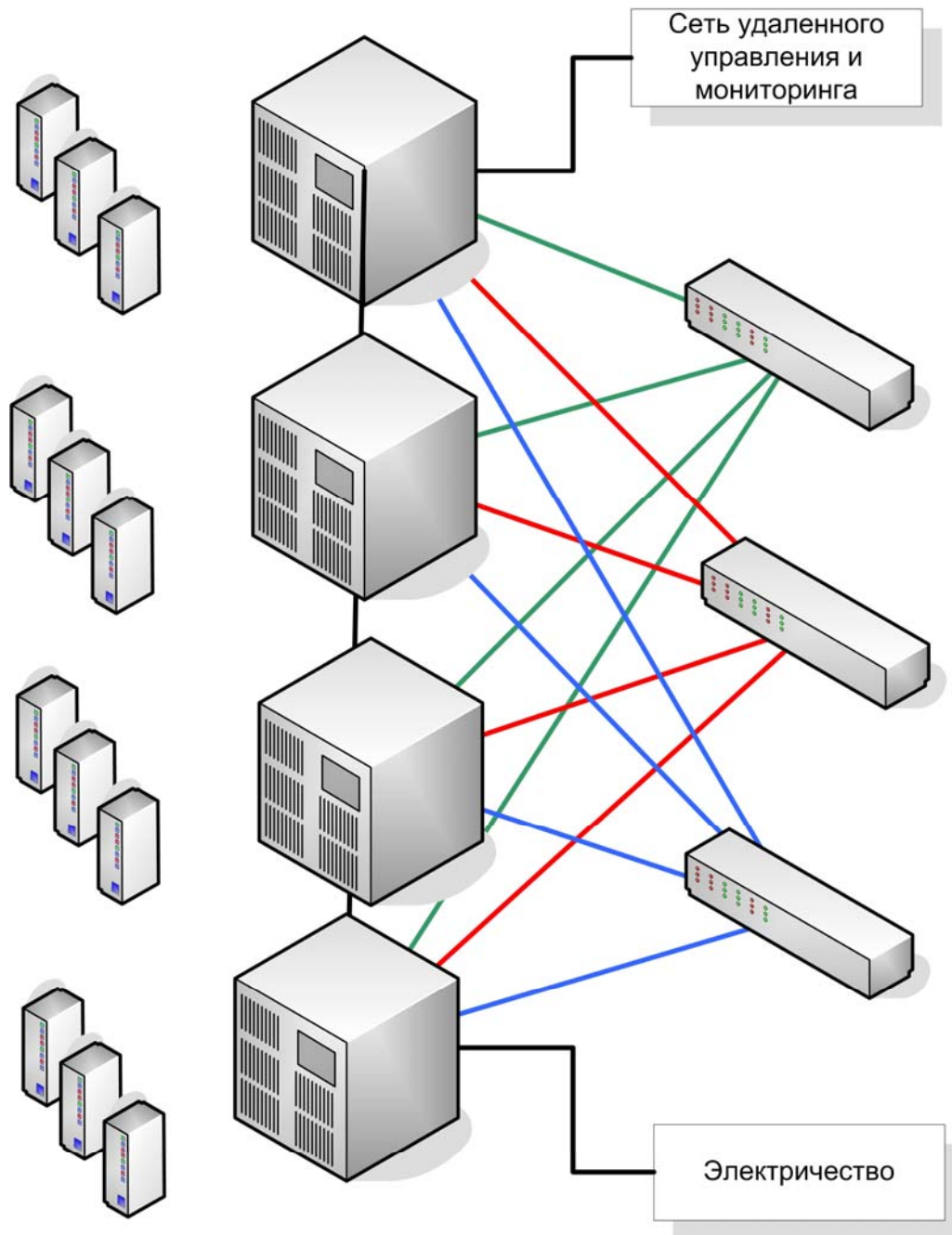


Управляющий модуль

- Средство удаленного управления и мониторинга:
 - Web-interface
 - IMPI
 - SNMP



Общая схема SMH11



Узлы:

- Узлы типа S (17 шт.):
 - 4 x Opteron 6164 (64 Гб)
- Узлы типа T (8 шт.):
 - 2 x Intel Xeon + 2 x GPU
- Узлы типа X (10 шт.):
 - 4 x Opteron 6174 (128Гб)

Сети:

- Gigabit Ethernet для запуска задач
- Gigabit Ethernet для передачи файлов (+ локальная дисковая память)
- Infiniband для взаимодействия процессов
- Дополнительно:
 - Управляющая сеть (мониторинг)
 - Электрическая сеть.

Инфраструктура

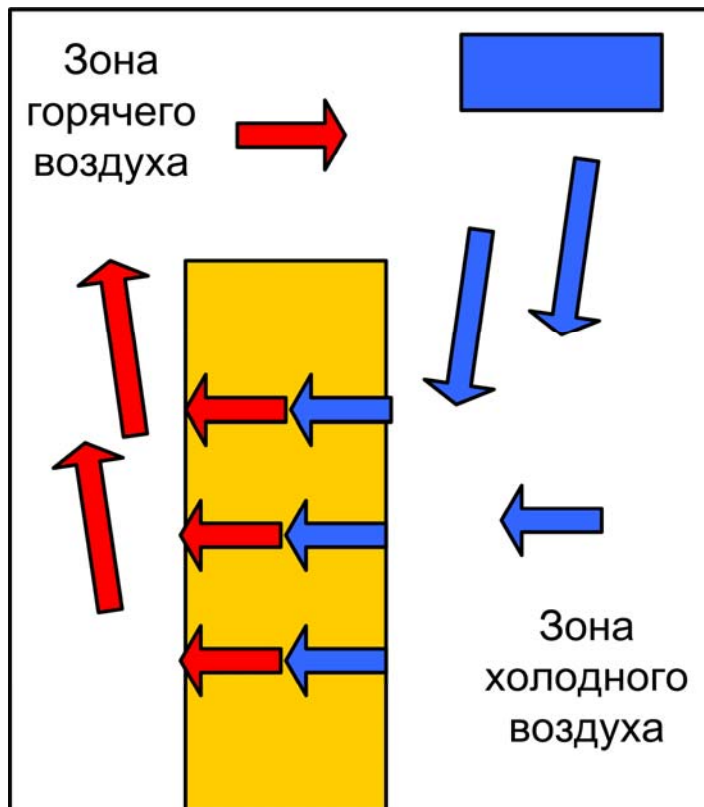
- Система бесперебойного питания:
 - Расчет мощности (~30-35 кВт)
 - Количество батарей



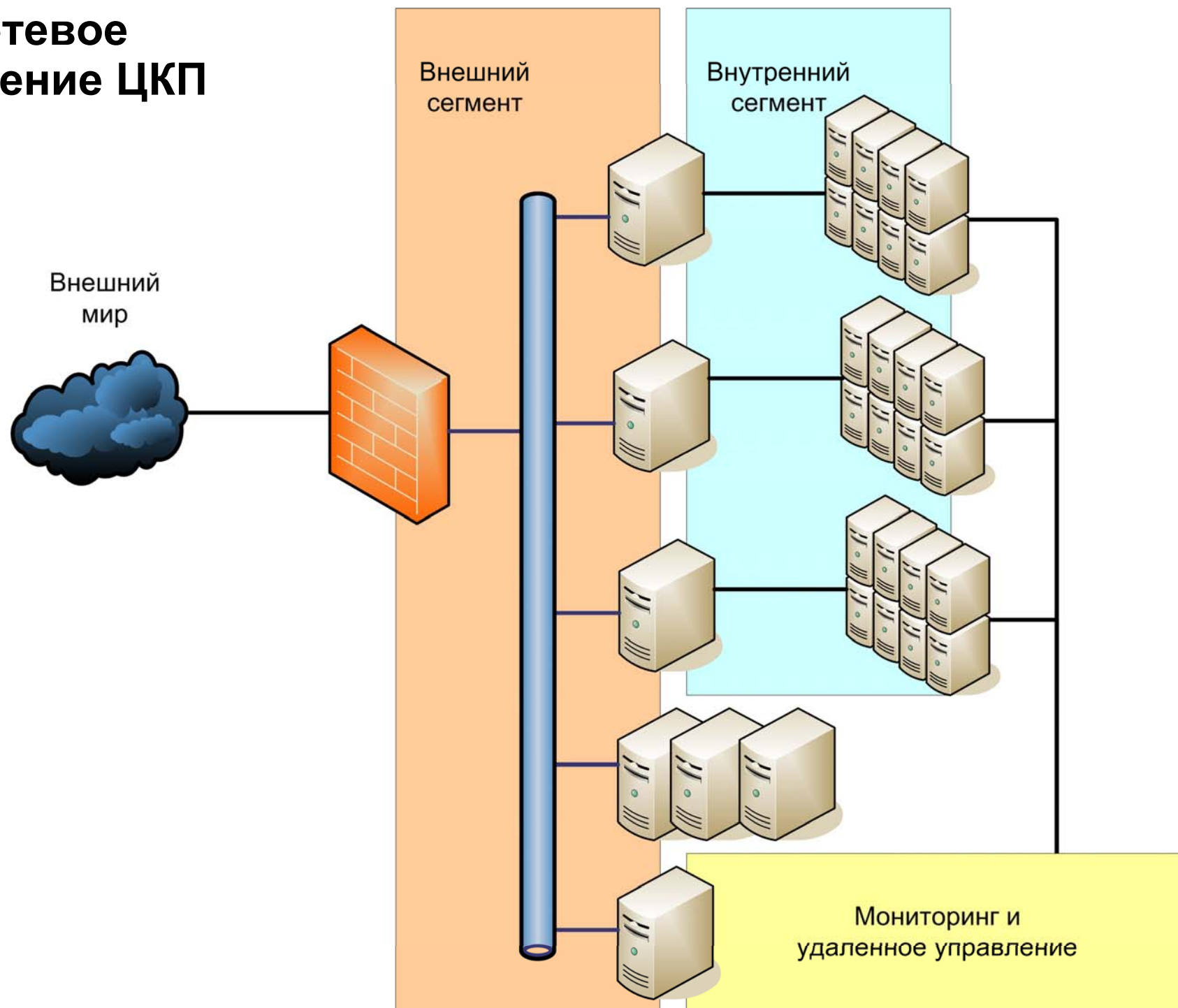
Тип батареи (Ач)	F (17)													
Конфигурация	Нагрузка (ВА)													
	2000	4000	6000	8000	10000	15000	19000	22500	25000	27000	30000	35000	40000	45000
BAT ONL-33	3,1 ч	1,2 ч	41,8 мин	26,9 мин	21,7 мин	11,3 мин	7,9 мин	<7 мин	<6 мин	<5 мин	<5 мин	<5 мин	<5 мин	<5 мин
2 BAT ONL-33	7,1 ч	3,1 ч	2,1 ч	1,2 ч	56,5 мин	28,2 мин	23 мин	18,5 мин	15,3 мин	12,8 мин	11,3 мин	9,1 мин	7 мин	<7 мин
3 BAT ONL-33	10,2 ч	4,9 ч	3,1 ч	2,4 ч	1,8 ч	56,5 мин	36,9 мин	28,2 мин	26 мин	24,3 мин	21,7 мин	17,5 мин	13,2 мин	11,3 мин
4 BAT ONL-33	14,4 ч	7,1 ч	4,2 ч	3,1 ч	2,6 ч	1,4 ч	1 ч	47,3 мин	38,2 мин	30,8 мин	28,2 мин	25 мин	21,7 мин	18,5 мин

Инфраструктура

- Система охлаждения (обеспечить отвод тепла)
 - Расчет мощности охл. (= потребл. мощность)
 - 2 конд. по 14 кВт
 - Циркуляция (порьба с «застойными» зонами горячего воздуха)



Сетевое окружение ЦКП



IP адресное пространство

- Внешняя сеть:
 - 94.198.16.224 / 28
 - 94.198.20.160 / 28
- Внутренняя сеть
 - 10.200.0.0 / 16
 - 10.200.11.0 / 24 – сеть управления
 - 10.200.111.0 / 24 – сеть коммуникационная
 - 10.200.211.0 / 24 – сеть файловая
 - ... аналогично другие кластера
- Сеть мониторинга
 - 192.168.200.0 / 24

DNS сервер: пространство имен

- **cc.dvo.ru**
 - smh11.cc.dvo.ru – управляющая машина
 - smh11-1.cc.dvo.ru – smh11-4.cc.dvo.ru - шасси
 - smh11-bmc-s1.cc.dvo.ru ... - управление узлами
 - smh11-s1.cc.dvo.ru, smh11-s1x.cc.dvo.ru – узлы
 - smh11-t1.cc.dvo.ru ...
 - smh11-x1.cc.dvo.ru
 - mvs15k.cc.dvo.ru
 - ...
- **intra** – файловая сеть
- **myri.intra** – коммуникационная сеть myrinet
- **iband.intra** – коммуникационная сеть infiniband

DHCP сервер: MAC-IP соответствие

```
subnet 10.200.11.0 netmask 255.255.255.0 {
    option routers 10.200.11.1;
    option domain-name-servers 94.198.16.226, 94.198.20.162, 94.198.16.3;
    option domain-name "cc.dvo.ru";
    filename "pxelinux.0";
        #pool {
        #    allow members of "pxeclient";
        #    range 10.200.11.2 10.200.11.170;
        #}
    host smh11-s1.cc.dvo.ru {
        hardware ethernet 00:30:48:ff:97:c2;
            #eth0    -    00:30:48:ff:97:c2
            #eth1    -    00:30:48:ff:97:c3
        option host-name smh11-s1;
        fixed-address 10.200.11.1;
    }
    host smh11-s1x.cc.dvo.ru {
        hardware ethernet 00:30:48:ff:97:c3;
            #eth0    -    00:30:48:ff:97:c2
            #eth1    -    00:30:48:ff:97:c3
        fixed-address 10.200.11.101;
    }
}
```

Обслуживающие сервера

- DNS & DHCP сервера
- Firewall
- Сервер доступа и авторизации (LDAP)
- Интернет-сервер (www & ftp; wiki)
- Сервера СУБД (MySQL, PostgreSQL)

Firewall

- Политика

- Доступ по ssh разрешен только из доверенных сетей

```
TRUSTED_NETS=(  
    "10.0.0.0/8"           # private IPv4 address space (class A)  
    "172.16.0.0/12"        # private IPv4 address space (16 class B)  
    "192.168.0.0/16"       # private IPv4 address space (256 class C)  
    "94.198.16.0/21"       # from RIPE: IACP FEBRAS  
    "83.149.192.0/18"      # from RIPE: CNTK RAN  
    "62.76.0.0/16"        # from RIPE: RB-NET  
    "77.34.0.0/15"        # from RIPE: PRIMORYE NET  
    "109.126.0.0/18"      # from RIPE: VLADLINK NET  
    "46.37.128.0/19"      # from RIPE: VLADLINK NET
```

- пакет **fail2ban**: отслеживание журналов входа, блокировка при обнаружении нелегальных действий по подбору пароля

Интернет-сервер



ЦЕНТР КОЛЛЕКТИВНОГО ПОЛЬЗОВАНИЯ
"ДАЛЬНЕВОСТОЧНЫЙ"
ВЫЧИСЛИТЕЛЬНЫЙ РЕСУРС

[ИНФОРМАЦИЯ](#) 
[контакты](#)
[суперкомпьютеры](#)
[текущее использование](#)
[кластеров](#)
[мониторинг](#)

[ДОКУМЕНТАЦИЯ](#) 
[описание кластера](#)
[руководство](#)
[пользователя](#)
[руководство](#)
[программиста](#)
[система очередей](#)
[учебная литература](#)
[софт](#)

[ПУБЛИКАЦИИ](#)
[НАШИ ПРОЕКТЫ](#)
[ПЕРСОНАЛИИ](#)
[ССЫЛКИ](#)

[на главную](#)

[НОВОСТИ](#)
[АРХИВ](#)
[РЕГИСТРАЦИЯ](#)
[ДОСТУП НА КЛАСТЕР](#)

Описание кластеров

Действующие (по состоянию на 01.10.2011)



- Название: SMH11
- Ввод в эксплуатацию: 2011 год
- Пиковая производительность: 18 TFLOPS
- [Характеристики и описание](#)



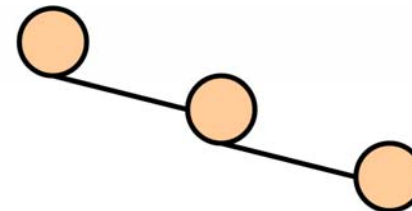
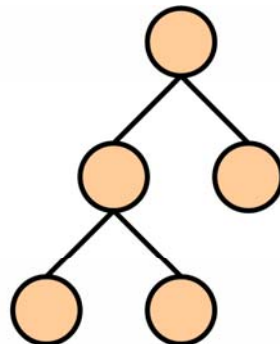
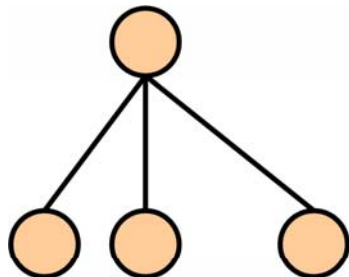
- Название: MBC1000/17
- Модернизирован: 2011 год
- Ввод в эксплуатацию: 2001 год
- Пиковая производительность: 1.26 TFLOPS
- [Характеристики и описание](#)

Программное окружение. Системная часть

- Выбор ОС – выбора нет: Linux!
 - **RHEL** – Red Hat Enterprise Linux
 - **SLES** – SuSE Linux Enterprise Server
 - **CentOS** – Community Enterprise Linux
 - **Scientific Linux** – оптимизация пакетов для научных приложений
 - **Debian**
 - **Gentoo** – наш выбор
- Windows Compute Cluster Server

Установка. Стратегии

- Полностью «ручная» установка
- «Ручная» установка + клонирование
 - Бездисковый кластер
 - Схемы клонирования
- Автоматизированная установка по сети:
 - FAI (Fully Automatic Installation)
 - AutoYAST
 - ...



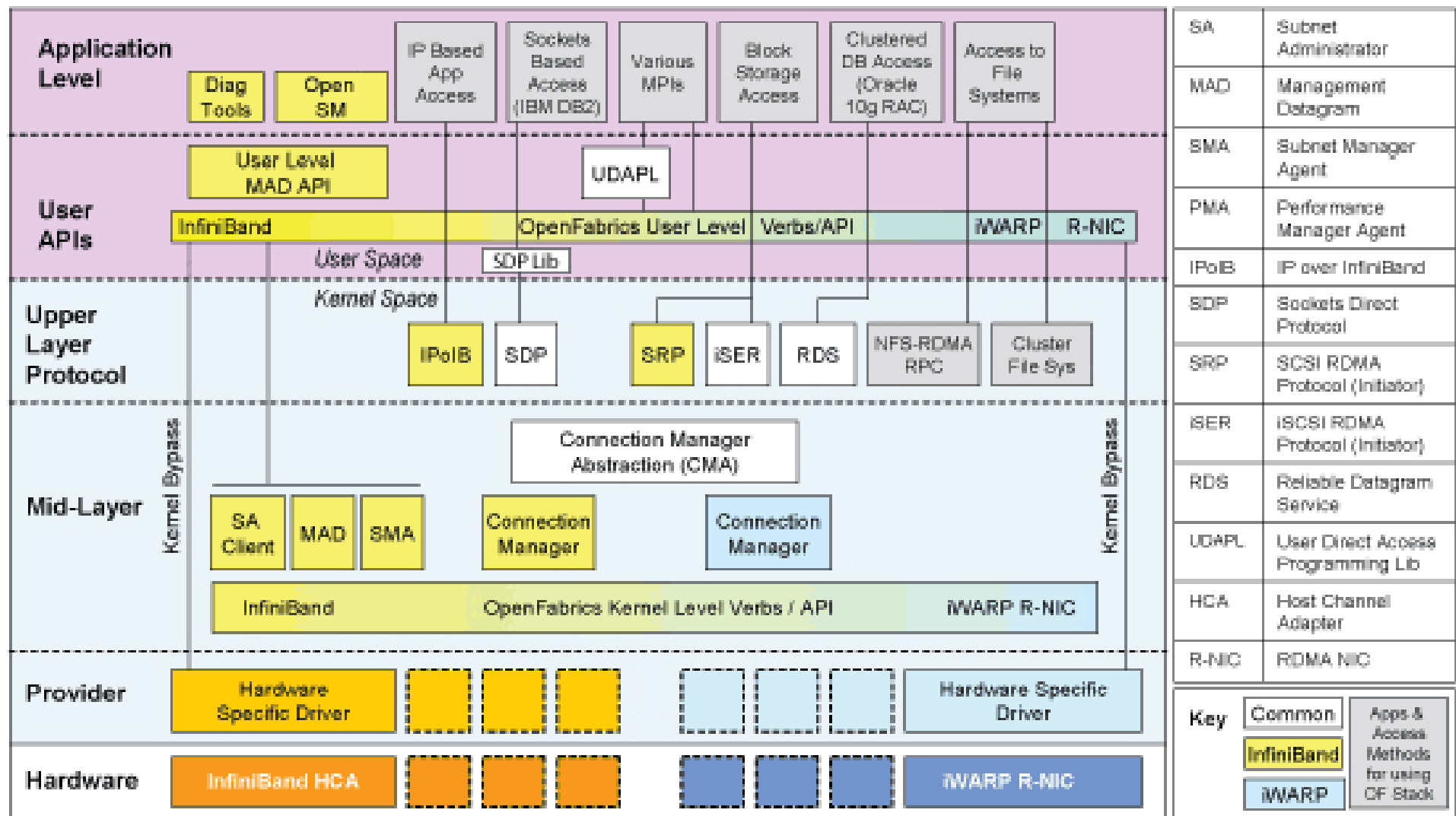
«Ручная» установка Gentoo

- Почему Gentoo?
 - Открыто все (можно разобраться, почему не работает, вплоть до исходников)
 - Оптимизировать сборку всей системы (в том числе оптимизировать сборку компилятора, которым далее компилируются остальные пакеты системы)
- Установка MBS15K:
 - Установили один узел.
 - Написали скрипт клонирования
 - Запустили скрипт, сделали стартовую перезагрузку
- Установка SMH11 (загрузка образа системы по сети):
 - Установили один узел. Сделали образ системы (~10 Гбайт)
 - На базе загрузочного диска Gentoo minimal сделали собственную сборку, которая копирует образ системы из сетевой директории в RAM-диск и заливает его на локальный диск, потом перезагружается.
 - Инициировать загрузку каждого узла
- Установка MBS1000/17 (7 узлов)
 - Установили один узел.
 - Клонировали диски по простой схеме.

Настройка сети

Infiniband: OFED

The OpenFabrics Enterprise Distribution (OFED™) is open-source software for RDMA and kernel bypass applications. OFED is used in business, research and scientific environments that require highly efficient networks, storage connectivity and parallel computing.



Пакет OFED

- Ориентирован на установку на основные дистрибутивы в HPC: SLES, RHEL, ...
- Было необходимо установить на Gentoo:
 - Разобрали по кусочкам OFED installer
 - Определили общий список пакетов (более 40)
 - Определили список пакетов, попадающий в разделы установки:
 - hpc_kernel_packages
 - hpc_user_packages
 - Определили порядок установки
 - Провели установку из исходников
 - Определили порядок загрузки драйверов
 - Тестирование

hpc_kernel_packages

01. kernel-ib - MUST BE! (здесь драйвер карточки и infiniband kernel core)

hpc_user_packages

01. libibverbs – базовые понятия Infiniband

02. libmlx4 - userspace driver for MLX4

03. librdmact - RDMA функциональность

04. dapl - DAPL: Direct Access Programming Library

05. libibumad – Management Datagram

06. libibmad - Management Datagram

07. opensm – Subnet manager

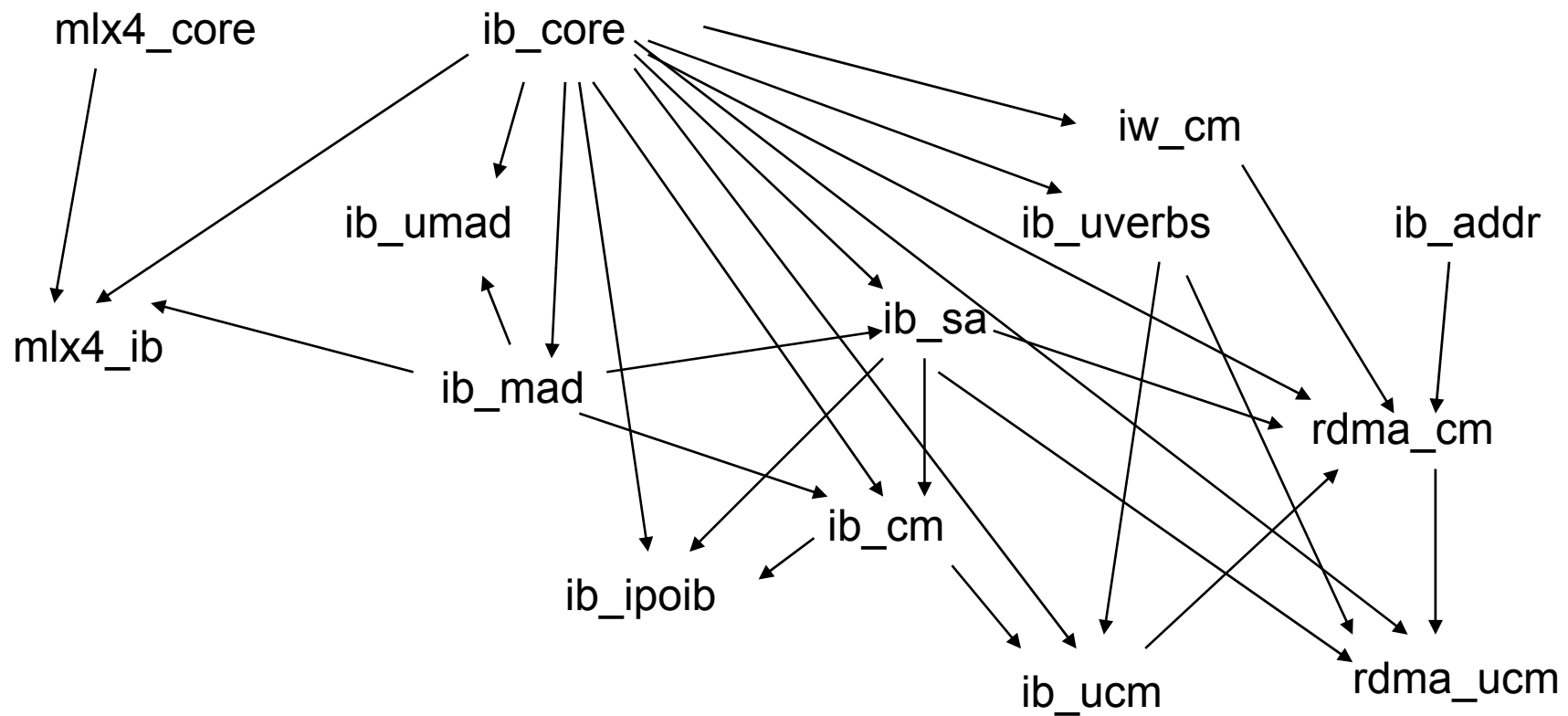
08. infiniband-diags - диагностика

09. ibutils - диагностика

10. qperf – тесты

11. perftest - тесты

Дерево модулей Infiniband



Удаленное управление

- SSH Hostbased authentication:
 - Прозрачная авторизация пользователей на основе доверия имени хоста, с которого он приходит
 - На управляющую машину попадаем с паролем smh11.cc.dvo.ru
 - Теперь с smh11.cc.dvo.ru можем двигаться дальше по пространству имен
 - Имени smh11.cc.dvo.ru все доверяют, поэтому можно без пароля зайти на любой узел
 - Узлы доверяют друг другу по своим именам, поэтому можно зайти с любого узла на любой
 - Реализация на основе RSA

Сетевая файловая система

- Определить точки монтирования NFS:
 - Каталог /common – используется для хранения общей для всех пользователей информации
 - Чаще всего это сборка open-source программ
 - Информация для систем управления пользователями и задачами
 - Каталог /system – системная информация
 - Образы систем
 - Конфигурация системы
 - Место для сборки системных и пользовательских пакетов
 - Для других нужд администратора кластера

Автоматизация конфигурирования

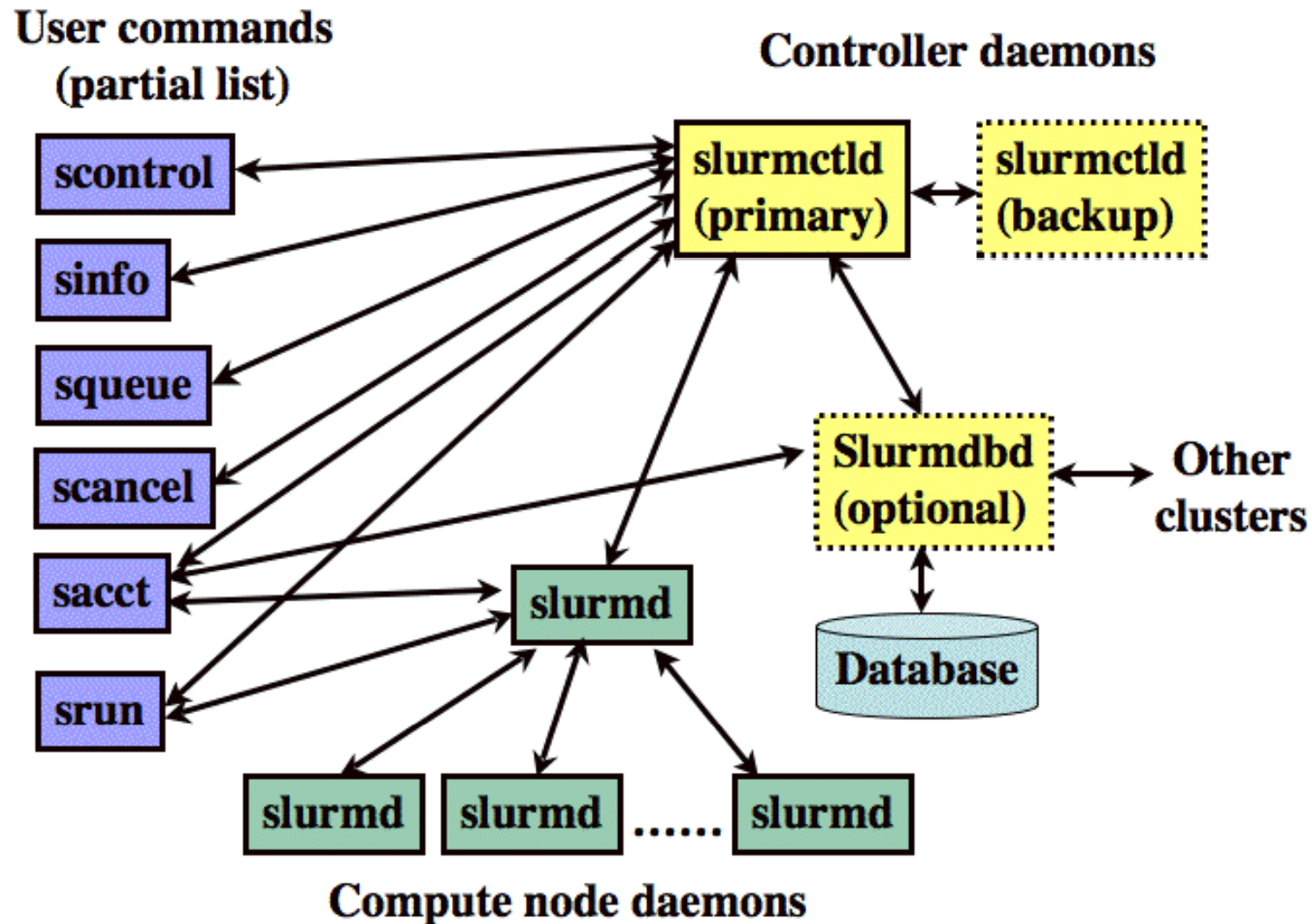
- В процессе эксплуатации возникает необходимость «на лету» менять конфигурацию узлов (незначительно!)
 - Например добавить или удалить точку монтирования сетевой файловой системы
- Системы управления конфигурациями:
 - Например puppet – IT leader solution
 - ...
- Сделали скрипт синхронизации
 - config.all
 - config.master

Система управления заданиями

- Torque
 - На управляющую машину ставится сервер системы (pbs_server)
 - На узлы ставится мини-сервер, обеспечивающий запуск и управление данным заданием на узле (pbs_mom)
 - Определяется список узлов (nodes)
 - Определяется список очередей, их характеристики (queues)
 - Запуск системы: mom -> server

Система управления заданиями

- SLURM



Установка планировщика ресурсов

- MAUI Cluster Scheduler
 - Устанавливается из коробки
 - Определить диспетчер ресурсов, с которым будет работать MAUI
 - Определить классы/очереди/разделы – группы (списки) узлов
 - Определить параметры планирования:
 - Приоритеты
 - Веса
 -
- Запустить систему: кластер почти готов!

Средства разработки

- Языки программирования
 - традиционно: C/C++, Fortran 77/90/95
- GNU GCC – **Gentoo: предустановлен**
- Intel C/C++ Compiler
- XLC C/C++ Compiler & XLF Compiler (IBM)
- Portland Group Compilers
- CUDA/OpenCL
-

Средства разработки

- Средства отладки:
 - gdb/ddd – традиционно
 - TotalView
 - Intel Cluster Studio
 - ...
- К сожалению многие средства разработки платные – закладывать в смету расходов на кластер!

Средств профилирования/анализа

- `strace` - трассировка системных вызовов
- `gperf` – профилирование скорости работы функций программы (требуется сборки с флагом `–pg` – `profiling generator`)
- `valgrind` – «рассказывает» все о работе приложения с памятью; определяет утечки, порчу динамической кучи; не требует специальной сборки со стороны приложения
- ...

Реализация MPI: OpenMPI

```
#!/bin/bash

export CC=gcc
export CFLAGS=-I/opt/ofed/include
export LDFLAGS=-L/opt/ofed/lib
export CXX=g++
export CXXFLAGS=${CFLAGS}
export F77=gfortran
export FFLAGS="-ffixed-form -fopenmp"
export FC=gfortran
export FCFLAGS="-ffree-form -fopenmp"

../configure \
    --prefix=/opt/openmpi \
    --enable-mpi-f77 \
    --enable-mpi-f90 \
    --enable-mpi-profile \
    --enable-mpi-cxx \
    --enable-mpi-cxx-seek \
    --enable-heterogeneous \
    --enable-mpi-threads \
    --enable-openib-rdmacm \
    --with-openib=/opt/ofed \
    --with-openib-libdir=/opt/ofed/lib
```


Реализация MPI: OpenMPI

- `mpirun` имеет параметр `mca`
 - MCA: Modular Component Architecture – немнолитная архитектура
 - Состоит из:
 - Frameworks – определенная функциональность (например, групповые операции MPI)
 - Components – интерфейсы функциональности
 - Modules – реализация интерфейсов
- MCA можно «по кусочкам» собрать динамически, можно собрать статически
- Проигрывает в производительности (по нашим тестам)

Реализация MPI: MVAPICH2

- Ориентирован на Infiniband и системы с большим количеством узлов
 - Поддерживает различные схемы запуска процессов
 - Простой (fork)
 - Дерево (proxy)
 - Поддерживает RDMA CM (прямое чтение/запись в память другого процесса на другом узле коммуникационной сети)
 - Поддерживает стандарты MPI1.1 и MPI2.0

Реализация MPI: MVAPICH2

```
#!/bin/bash
```

```
export CC=gcc
export CFLAGS=-I/opt/ofed/include
export LDFLAGS=-L/opt/ofed/lib
export F77=gfortran
export FFLAGS="-ffixed-form -fopenmp"
export F90=gfortran
export F90FLAGS="-ffree-form -fopenmp"
export CXXFLAGS=${CFLAGS}
```

```
./configure \
  --prefix=/opt/mvapich2 \
  --enable-error-checking=all \
  --enable-error-messages=all \
  --enable-timing=all \
  --enable-g=log \
  --enable-f77 \
  --enable-f90 \
  --enable-cxx \
  --enable-romio \
  --enable-debuginfo \
  --enable-mpcoll \
  --enable-async-progress \
  --enable-mpe \
  --enable-rdma-cm \
  --enable-xrc \
  --with-device=ch3 \
  --with-mpe \
  --with-cluster-size=medium \
  --with-dapl-include=/opt/ofed/include \
  --with-dapl-libpath=/opt/ofed/lib \
  --with-ib-include=/opt/ofed/include \
  --with-ib-libpath=/opt/ofed/lib \
  --with-rdma=gen2
```

Выбор реализации MPI

- MPI-selector
 - На уровне системы активирует определенные MPI-переменные среды окружения пользователя
 - MPIRUN_HOME
 - LD_LIBRARY_PATH
 - PATH
 - ...
 - Неудобство: требует перезагрузки сессии после переключения на другую реализацию MPI

Пользовательское ПО

- Прикладные базовые библиотеки
 - BLAS (AMCL, IML, ...)
 - LAPACK
 - ScaLAPACK
 - fftw
 - HDF5
 - ...
- Для этого ПО есть /common, которая доступна всем

Пользовательское ПО

- ПО со свободной лицензией:
 - octave – численная математика (аналог matlab)
 - NAND – молекулярная динамика
 - GAMESS – квантовая химия
 - GROMACS – молекулярная динамика
 - abinit – квантовая химия
 - OpenFOAM - гидродинамика
 -
- gnuplot – система визуализации научных данных
- mplayer – ala K-Lite Codec Pack для Linux
- /common

Пользовательское ПО

- Лицензионный софт:
 - ANSYS – система инженерных расчетов
 - VASP – квантовая механика
 - MOE: Molecular Operating Environment – молекулярная динамика
- В /common нельзя – не позволяет лицензия
- Только в директорию пользователя
- Для нескольких пользователей: можно определить специальную группу с правами запуска