Центры обработки данных

Суперкомпьютеры

Облака

Лекция 3

29 октября 2018

Дата центры
Центры обработки данных (ЦОД)

ЦОД

Структура ЦОД

- Информационная инфраструктура
  - серверное оборудование (обработка данных);
  - сетевые хранилища (хранение данных).

- Телекоммуникационная инфраструктура
  - взаимосвязь элементов дата-центра;
  - передача данных между дата-центром и пользователями.

- Инженерная инфраструктура, обеспечивающая функционирование основных систем дата-центра
  - кондиционирование для поддержания температуры и уровня влажности;
  - бесперебойное электроснабжение для автономной работы дата-центра;
  - охранно-пожарная сигнализация;
  - система пожаротушения;
  - система управления питанием;
  - система удаленного IP-контроля;
  - система контроля доступа (безопасность).

Типы ЦОД (по размеру)

- **Крупные дата-центры** имеют своё здание, специально сконструированное для обеспечения наилучших условий размещения. Обычно они имеют свои каналы связи, к которым подключают серверы.
- **Средние дата-центры** обычно арендуют площадку определённого размера и каналы определённой пропускной способности (обычно измеряется в Мбит/с).
- **Малые дата-центры** размещаются в малоприспособленных помещениях. Часто ими используется оборудование плохого качества, а также предоставляется самый минимум услуг.
- **Контейнерные дата-центры**. Стойки с оборудованием размещаются в стандартных ISO контейнерах размером 20 и 40 футов. Имеют преимущества т.к. могут перевозиться автомобильным и железнодорожным транспортом

Типы ЦОД (по надежности)

По стандарту **TIA/EIA-942**:

- **Tier 1 (N)** — отказы оборудования или проведение ремонтных работ приводят к остановке работы всего дата-центра; в дата-центре отсутствуют фальшполы, резервные источники электроснабжения и источники бесперебойного питания; инженерная инфраструктура не зарезервирована;
- **Tier 2 (N+1)** — имеется небольшой уровень резервирования; в дата-центре имеются фальшполы и резервные источники электроснабжения, однако проведение ремонтных работ также вызывает остановку работы дата-центра;
- **Tier 3 (N+1)** — имеется возможность проведения ремонтных работ (включая замену компонентов системы, добавление и удаление вышедшего из строя оборудования) без остановки работы дата-центра; инженерные системы однократно зарезервированы, имеется несколько каналов распределения электропитания и охлаждения, однако постоянно активен только один из них;
- **Tier 4 (2(N+1))** — имеется возможность проведения любых работ без остановки работы дата-центра; инженерные системы двукратно зарезервированы, то есть продублированы как основная, так и дополнительная системы (например, бесперебойное питание представлено двумя ИБП, работающими по схеме N+1).

Уровни надежности ЦОД

Выделено 4 уровня надежности:

- Tier I: время простоя 28,8 часов в год, коэффициент отказоустойчивости 99,671%;
- Tier II: 22,0 часа в год, 99,749%;
- Tier III: 1,6 часа в год, 99,982%;
- Tier IV: 0,4 часа в год, 99,995%.

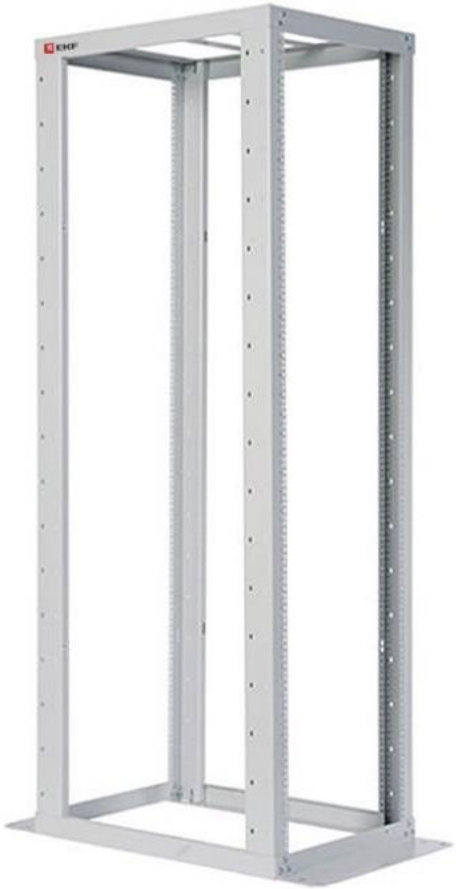
ЦОД. Стойки с серверами

ЦОД. Стойки с серверами

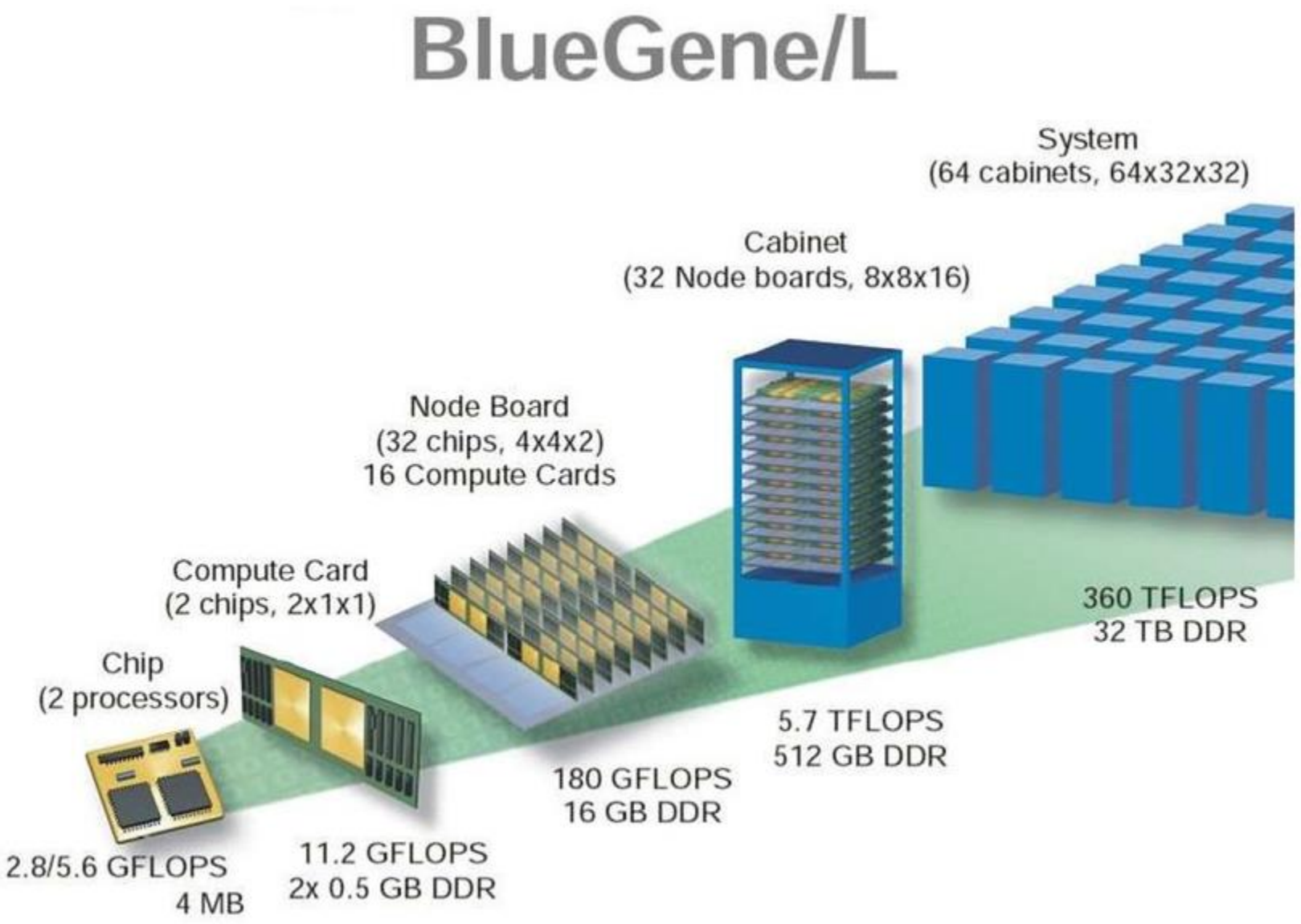
Сервер

Блейд сервер

Серверные шкафы и стойки

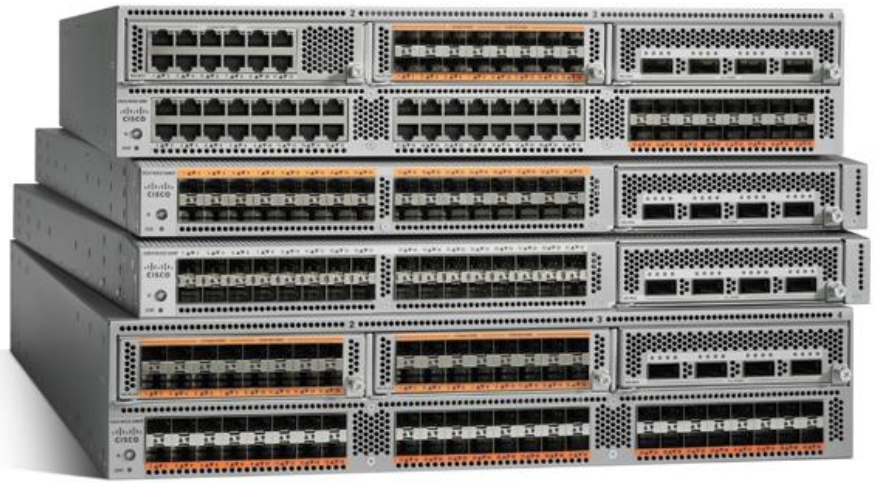


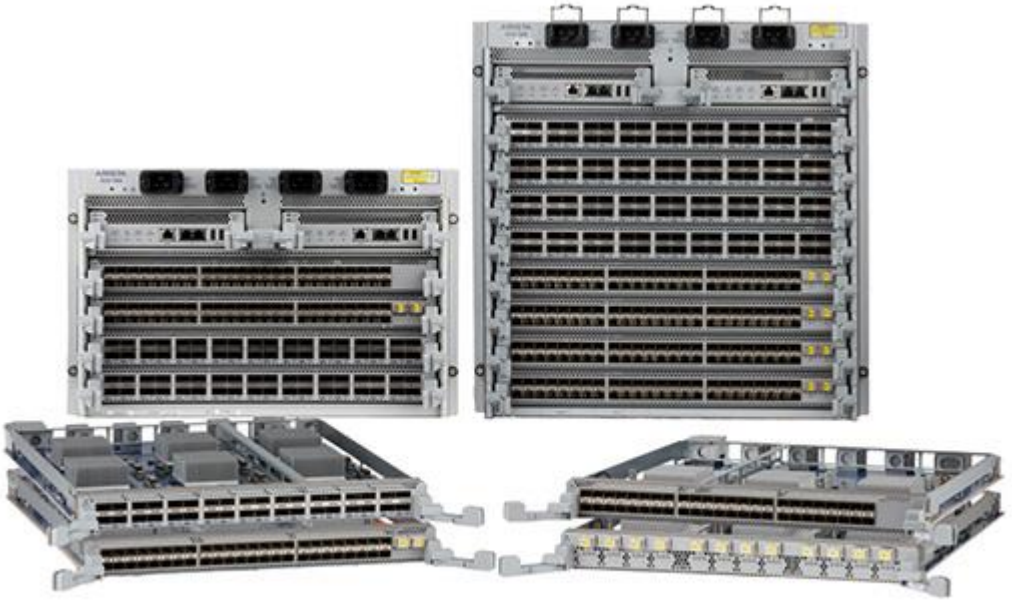
Уровни иерархии ВУ в ЦОД



Телекоммуникационная инфраструктура ЦОД

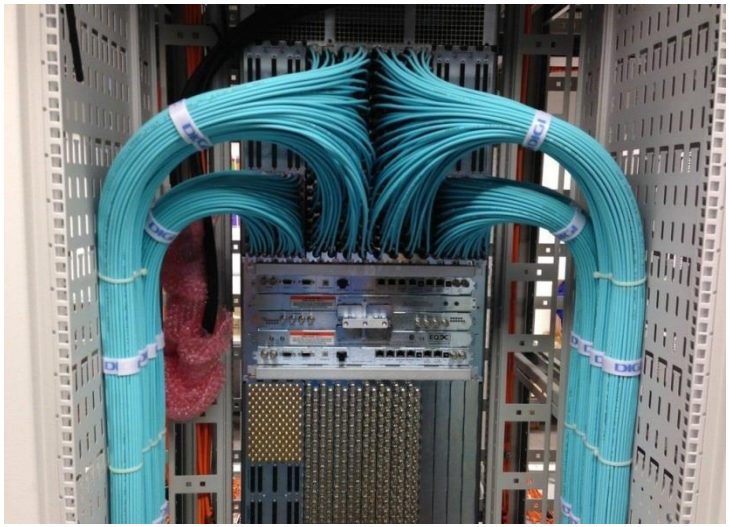
- Телекоммуникационная инфраструктура
  - взаимосвязь элементов дата-центра;
  - передача данных между дата-центром и пользователями.

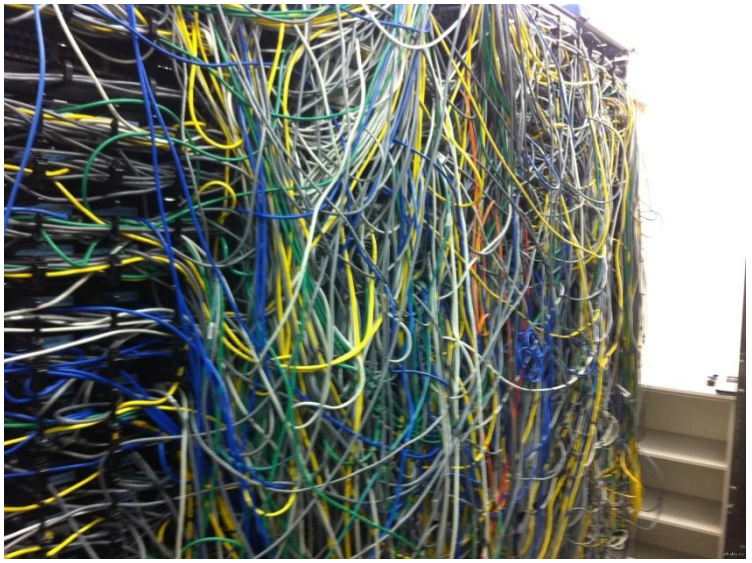


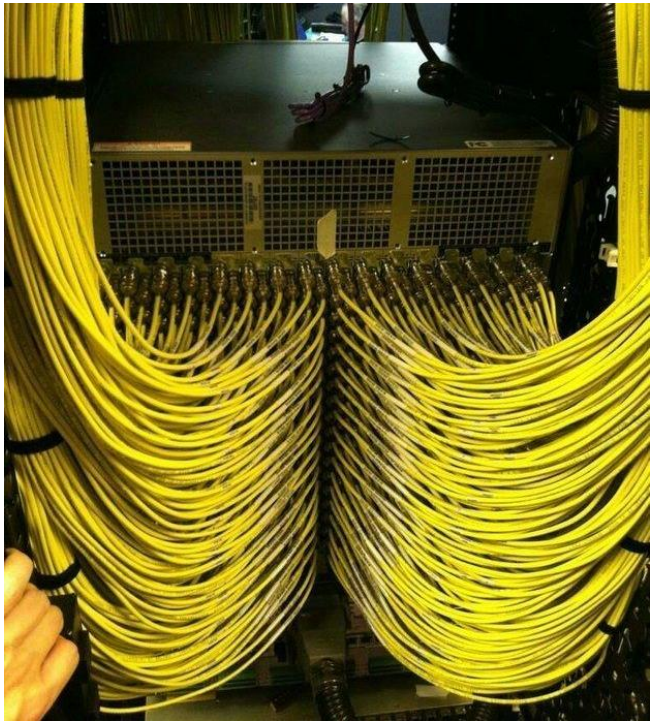


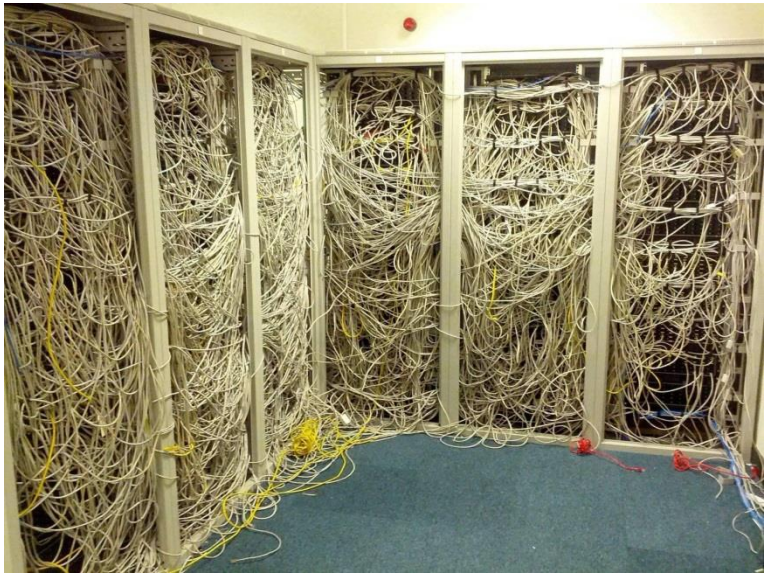
ЦОД. Кабельные системы

ЦОД. Кабельные системы







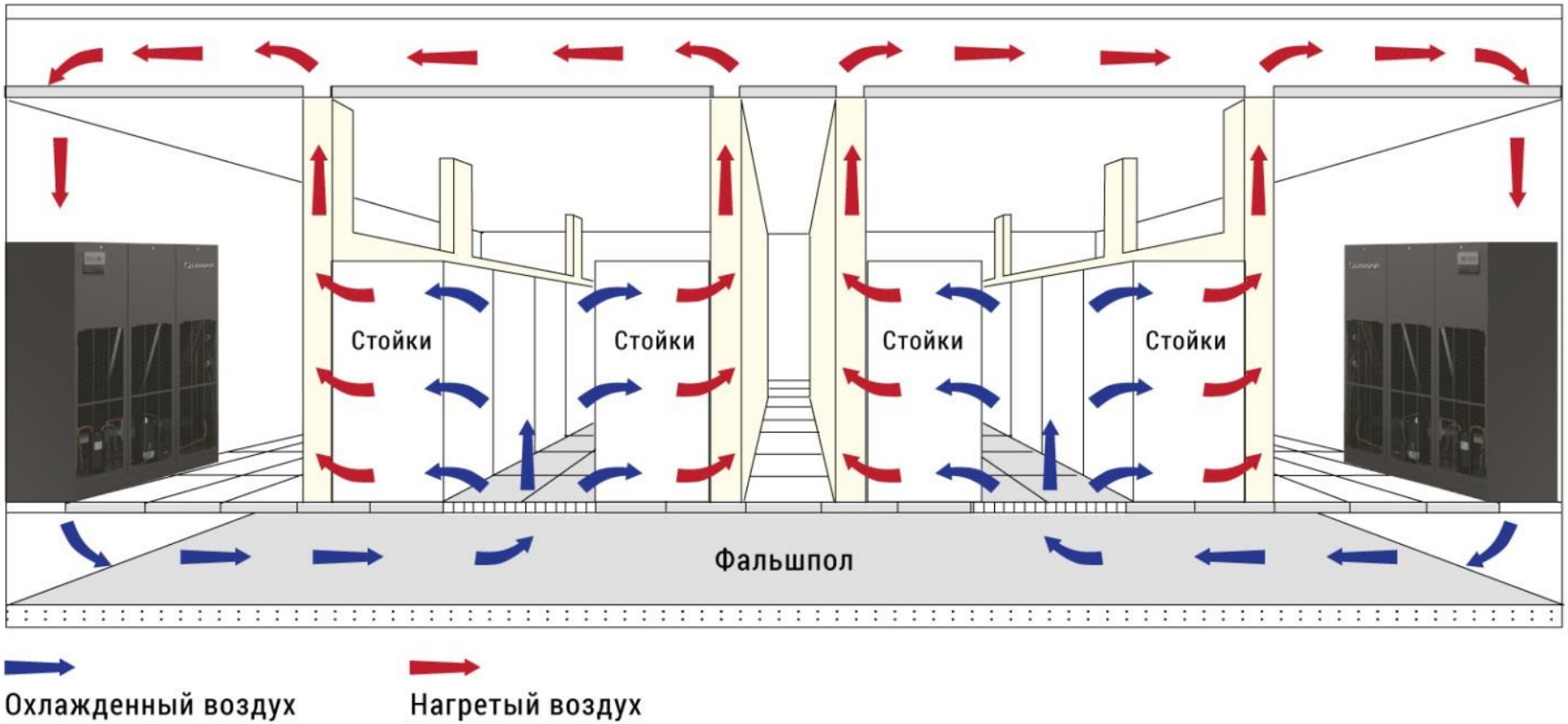


Инженерная инфраструктура ЦОД

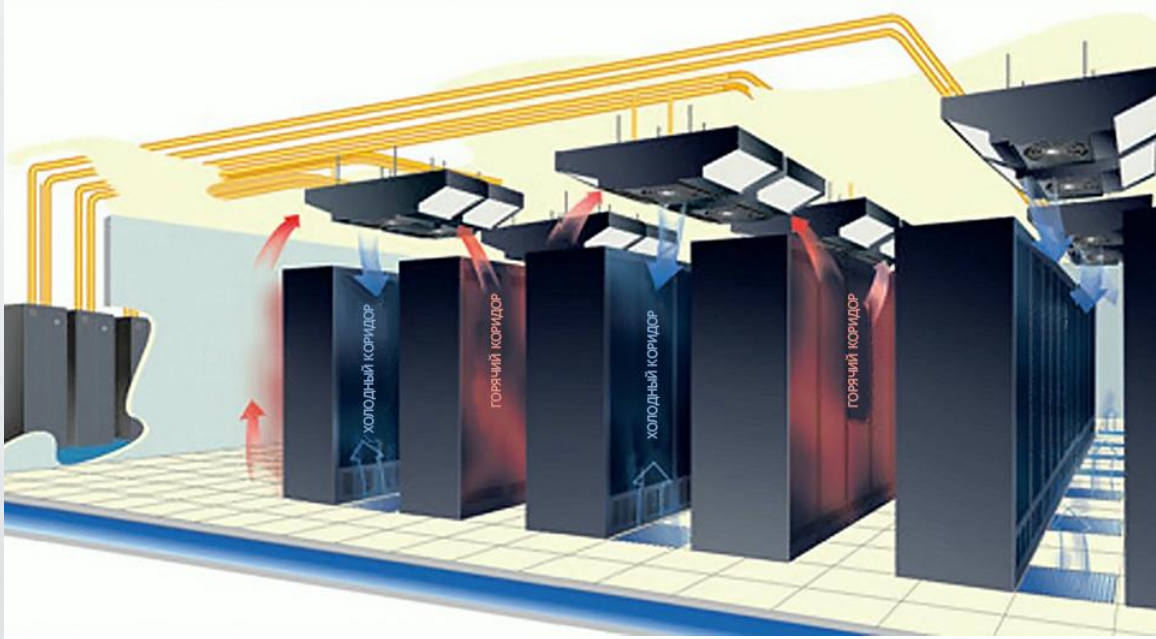
Инженерная инфраструктура обеспечивает функционирование основных систем дата-центра

- кондиционирование для поддержания температуры и уровня влажности;
- бесперебойное электроснабжение для автономной работы дата-центра;
- охранно-пожарная сигнализация;
- система пожаротушения;
- система управления питанием;
- система удаленного IP-контроля;
- система контроля доступа (безопасность).

Воздушное охлаждение

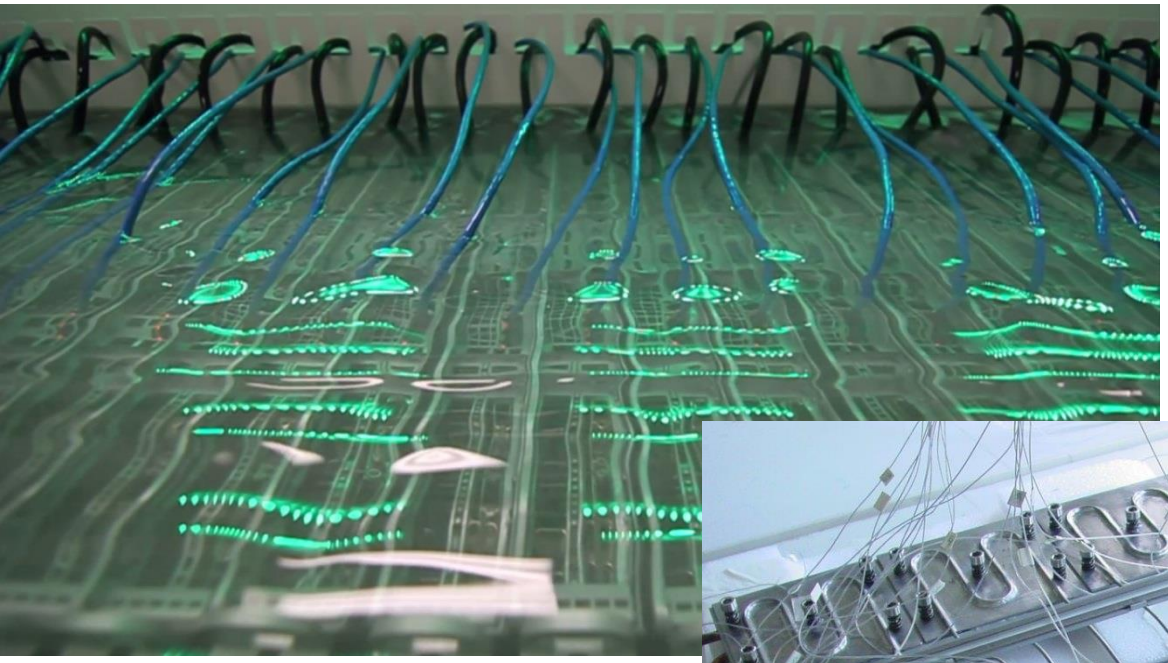


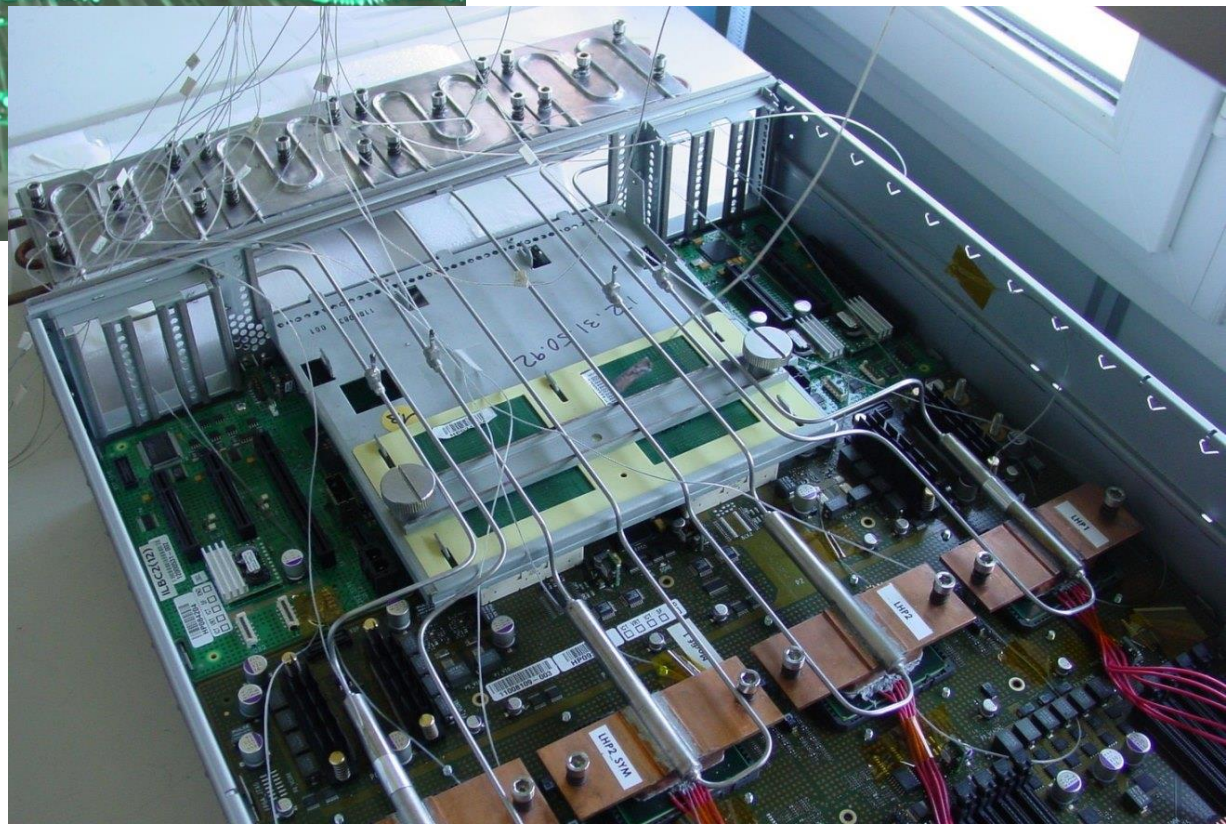
Воздушное охлаждение



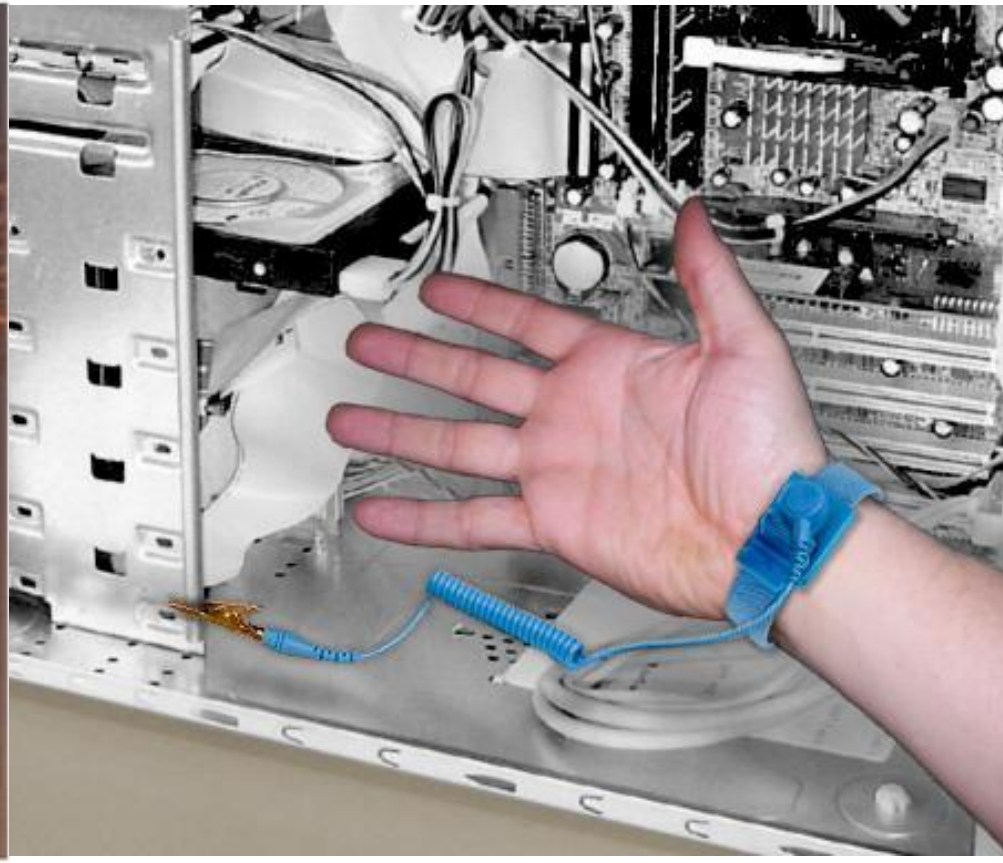
Пассивное воздушное охлаждение

Жидкостное охлаждение





Статическое электричество и контроль влажности в ЦОД



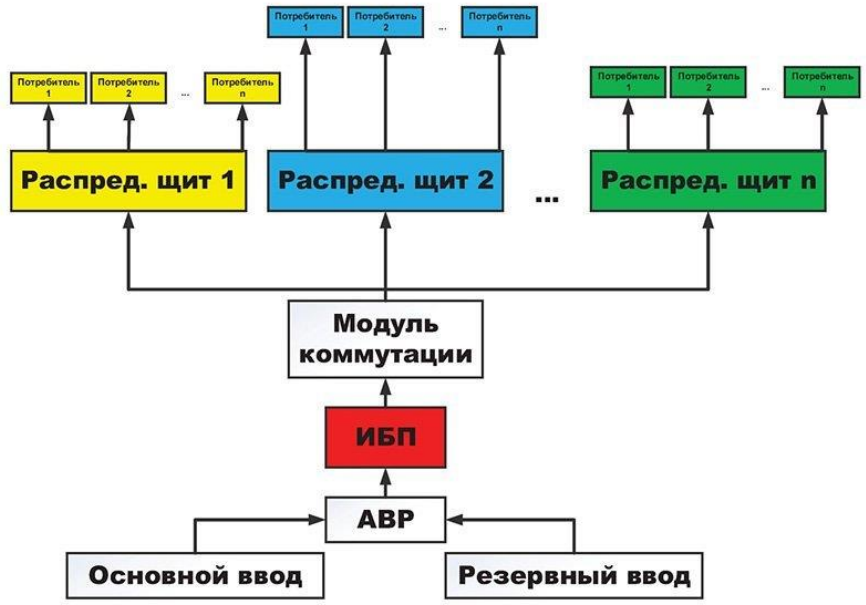
Бесперебойное электропитание ЦОД

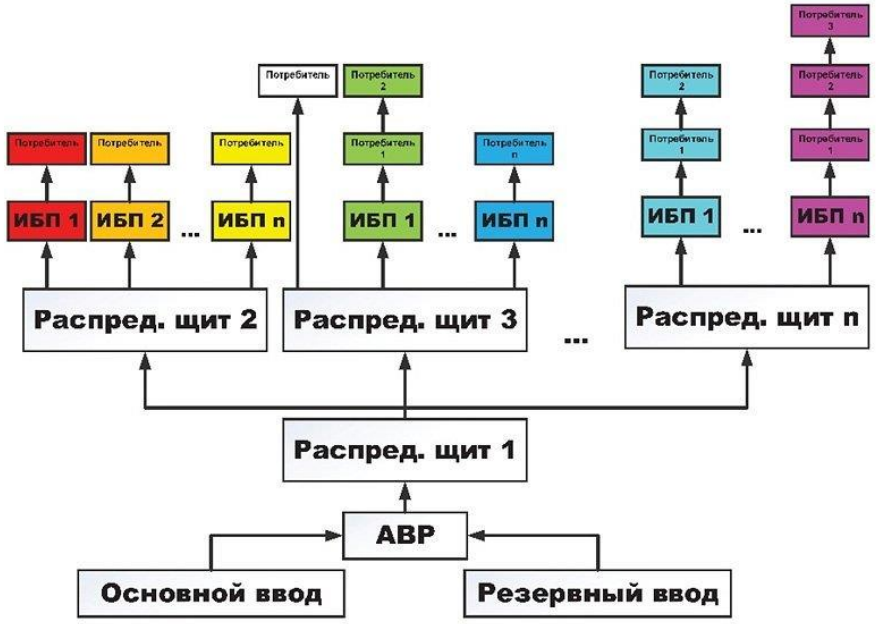
Бесперебойное электропитание ЦОД

Построение электропитания в ЦОД происходит по одной из 3 схем:

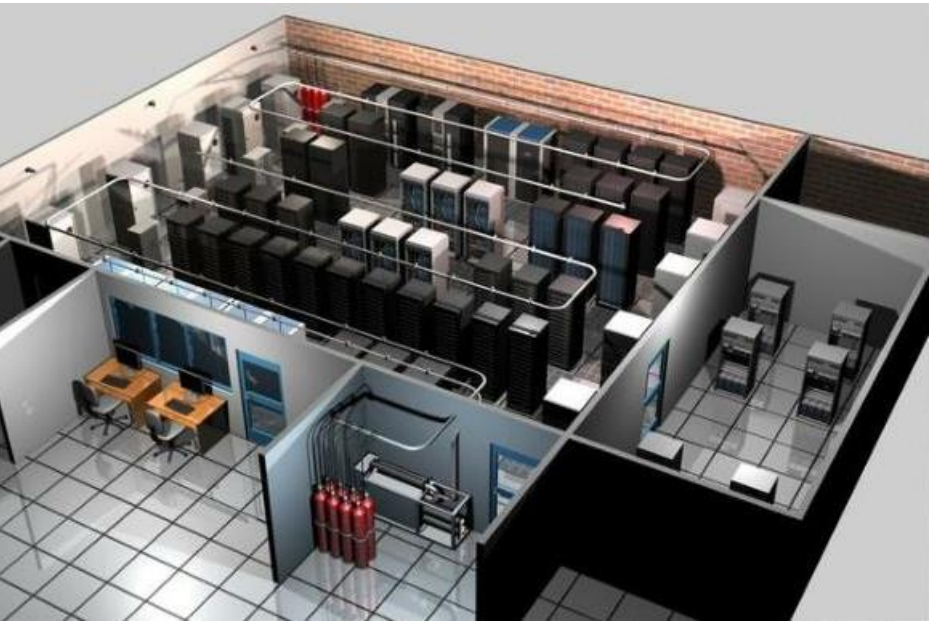
- централизованной;
- распределённой;
- комбинированной.

В каждой из них источник бесперебойного питания занимает главное место и является не только средством защиты, но и связующим звеном между различными элементами системы электроснабжения.





Система пожаротушения в ЦОД



гипоксический метод (постоянное снижение содержания кислорода в помещении до уровня ниже 14% путем введения азота);
изоляция (ввод инертного газа в помещение при возникновении возгорания, чтобы снизить уровень кислорода ниже 14%) с помощью азота, аргона, аргонита или инергена;
ингибирование (впрыскивание в помещение галогенированного газа, снижающего содержание кислорода и препятствующего процессу горения);
охлаждение (распыление мелкодисперсного водяного тумана на область горения, что приводит к снижению уровня кислорода на местном уровне и охлаждает зону возникновения огня);
порошок/аэрозоль (выброс порошковой химии и распыление продуктов горения аэрозоля).

Инфраструктура ЦОД

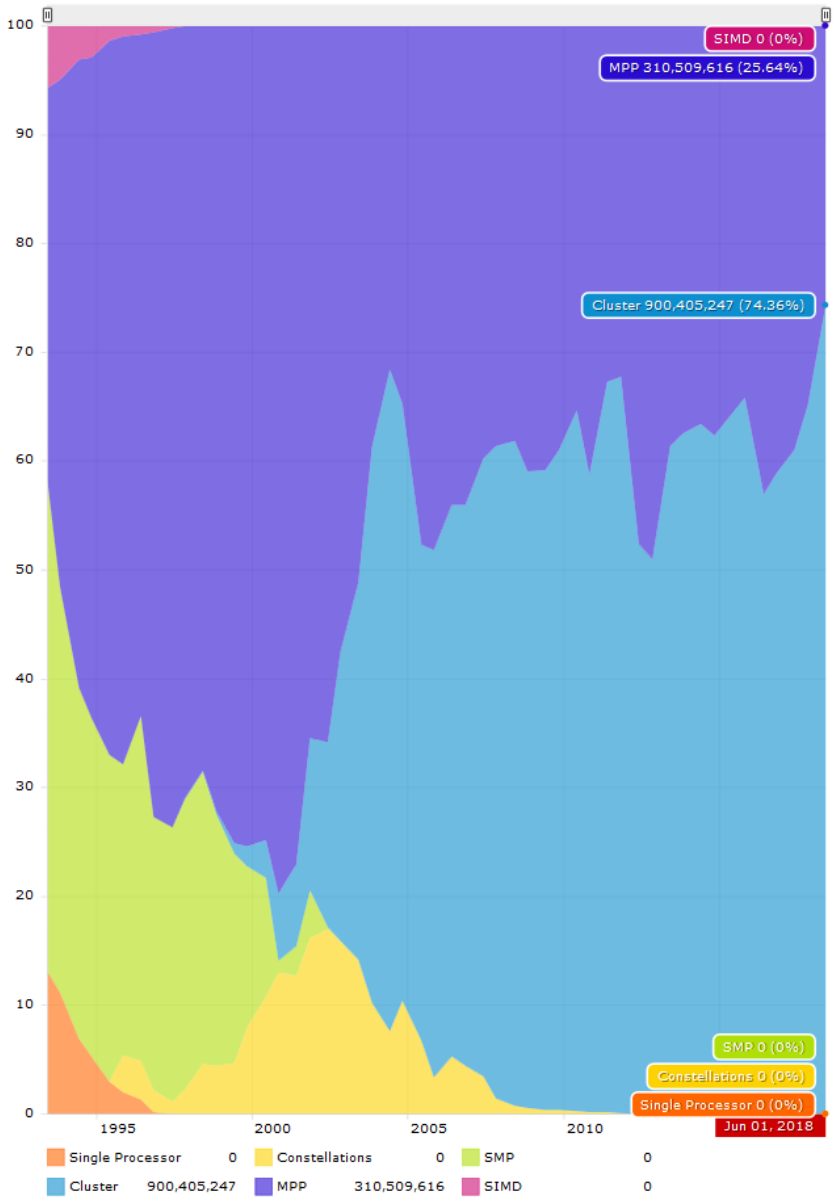
Безопасность

ЦОД. Управление и контроль

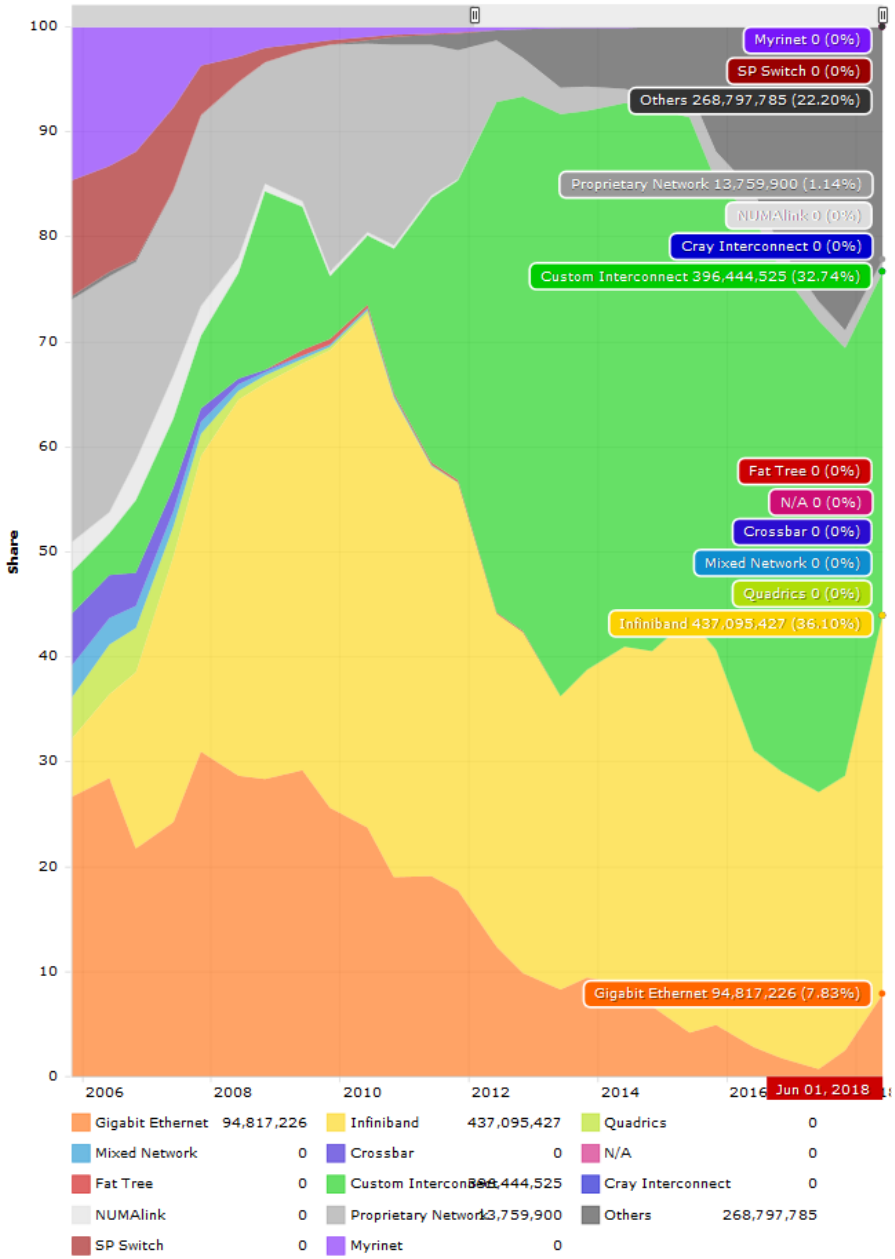
Суперкомпьютеры

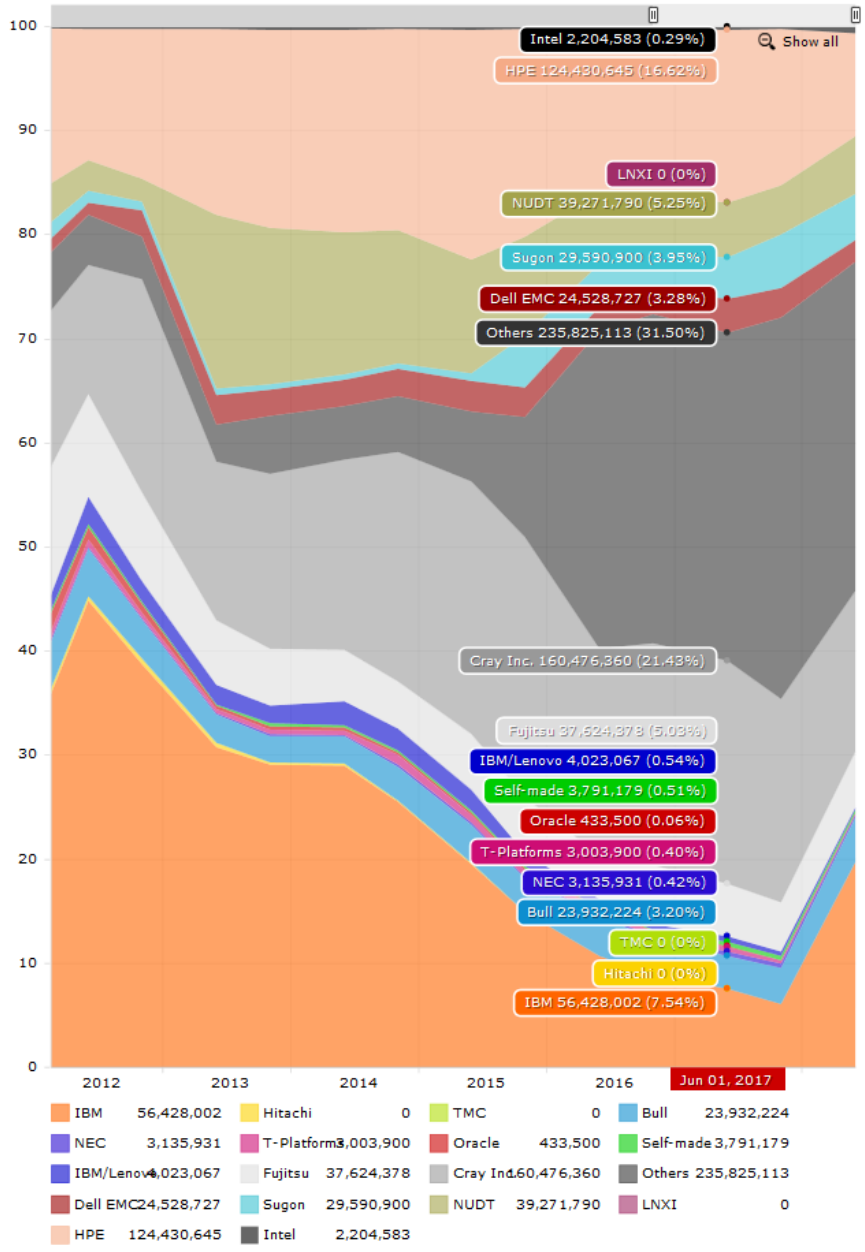
TOP500

Rank	System	Cores	Rmax (TFlop/s)	Rpeak (TFlop/s)	Power (kW)
1	**Summit** - IBM Power System AC922, IBM POWER9 22C 3.07GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband , IBM DOE/SC/Oak Ridge National Laboratory United States	2,282,544	122,300.0	187,659.3	8,806
2	**Sunway TaihuLight** - Sunway MPP, Sunway SW26010 260C 1.45GHz, Sunway , NRCPC National Supercomputing Center in Wuxi China	10,649,600	93,014.6	125,435.9	15,371
3	**Sierra** - IBM Power System S922LC, IBM POWER9 22C 3.1GHz, NVIDIA Volta GV100, Dual-rail Mellanox EDR Infiniband , IBM DOE/NNSA/LLNL United States	1,572,480	71,610.0	119,193.6	
4	**Tianhe-2A** - TH-IVB-FEP Cluster, Intel Xeon E5-2692v2 12C 2.2GHz, TH Express-2, Matrix-2000 , NUDT National Super Computer Center in Guangzhou China	4,981,760	61,444.5	100,678.7	18,482
5	**AI Bridging Cloud Infrastructure (ABCI)** - PRIMERGY CX2550 M4, Xeon Gold 6148 20C 2.4GHz, NVIDIA Tesla V100 SXM2, Infiniband EDR , Fujitsu National Institute of Advanced Industrial Science and Technology (AIST) Japan	391,680	19,880.0	32,576.6	1,649
6	**Piz Daint** - Cray XC50, Xeon E5-2690v3 12C 2.6GHz, Aries interconnect , NVIDIA Tesla P100 , Cray Inc. Swiss National Supercomputing Centre (CSCS) Switzerland	361,760	19,590.0	25,326.3	2,272
7	**Titan** - Cray XK7, Opteron 6274 16C 2.200GHz, Cray Gemini interconnect, NVIDIA K20x , Cray Inc. DOE/SC/Oak Ridge National Laboratory United States	560,640	17,590.0	27,112.5	8,209
8	**Sequoia** - BlueGene/Q, Power BQC 16C 1.60 GHz, Custom , IBM DOE/NNSA/LLNL United States	1,572,864	17,173.2	20,132.7	7,890
9	**Trinity** - Cray XC40, Intel Xeon Phi 7250 68C 1.4GHz, Aries interconnect , Cray Inc. DOE/NNSA/LANL/SNL United States	979,968	14,137.3	43,902.6	3,844
10	**Cori** - Cray XC40, Intel Xeon Phi 7250 68C 1.4GHz, Aries interconnect , Cray Inc. DOE/SC/LBNL/NERSC United States	622,336	14,014.7	27,880.7	3,939

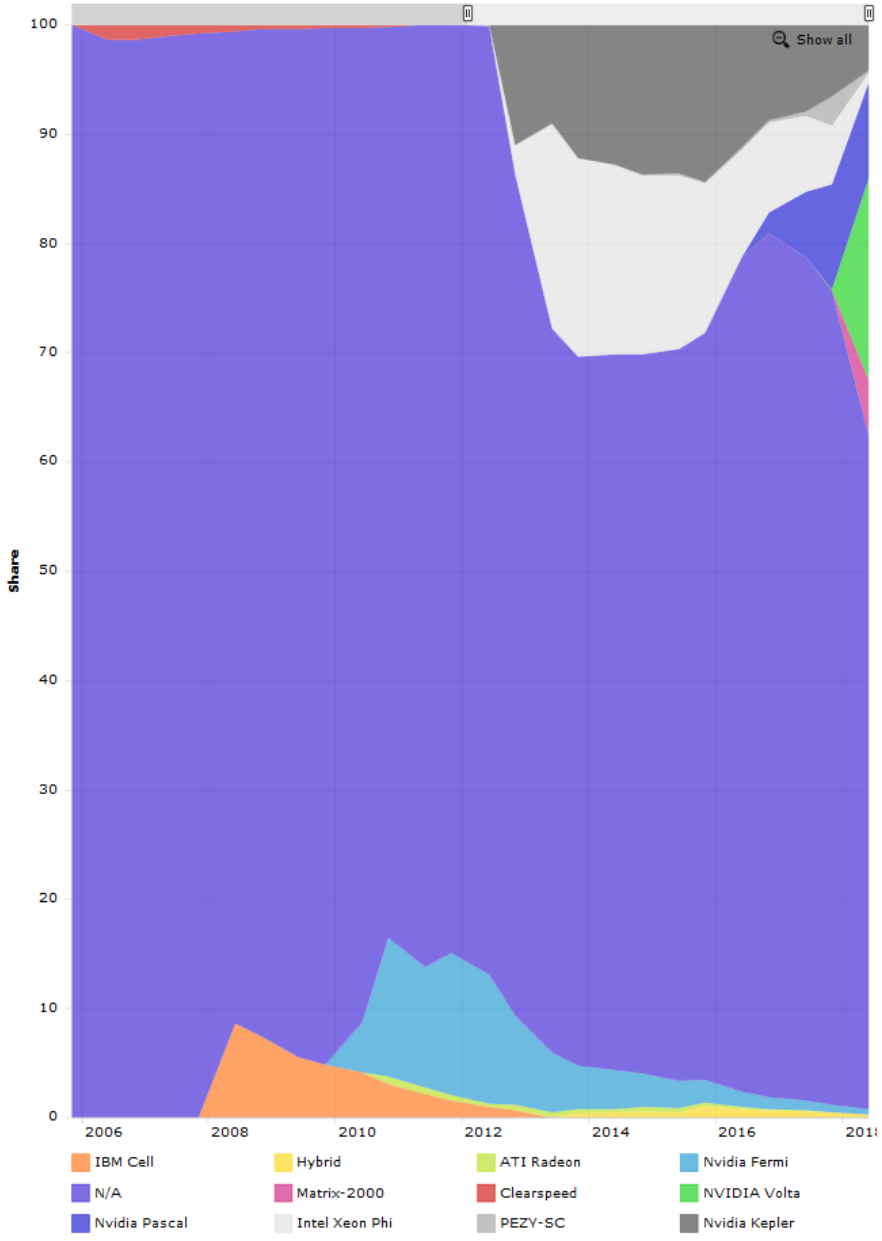


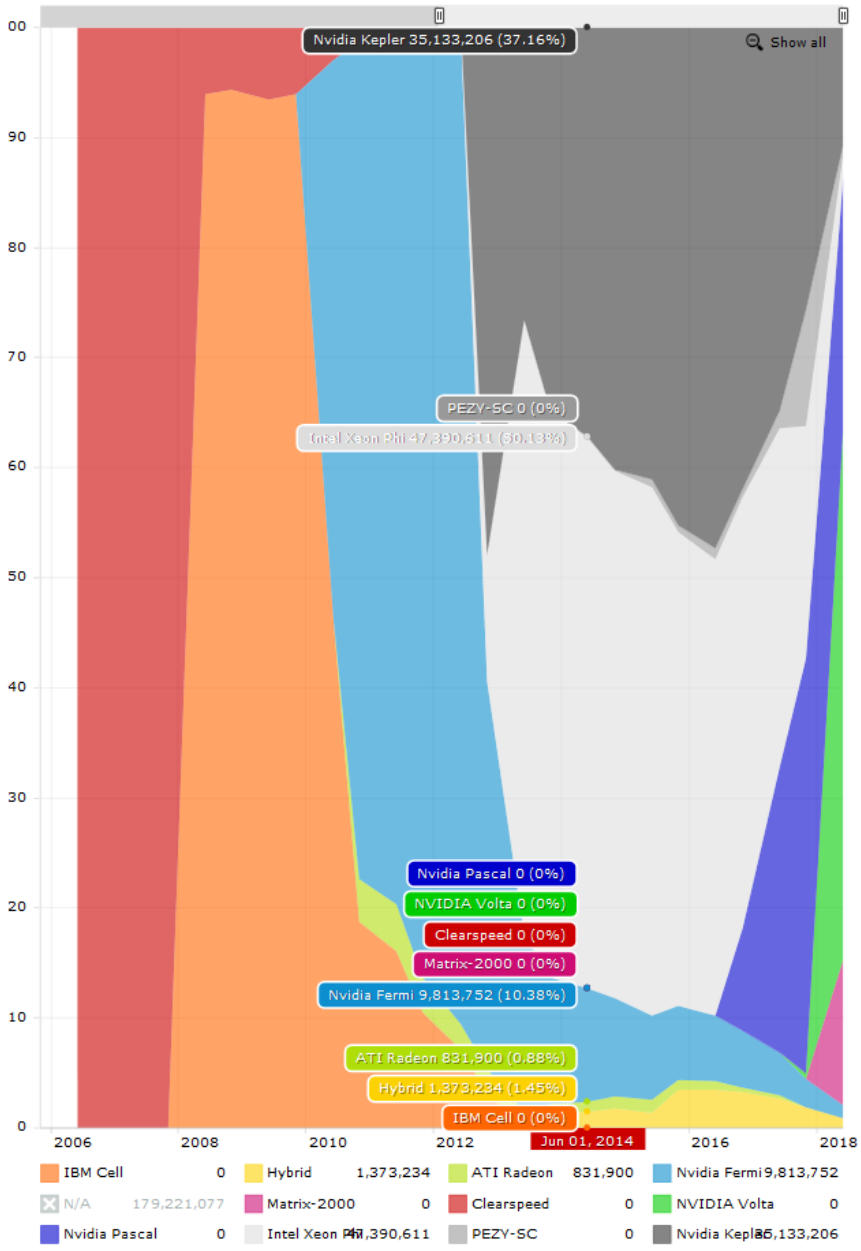
Статистика по TOP500 (сети, производители)





Статистика по ускорителям





Тяньхэ-2

- 29 мая 2013г. НОАК
- Производительность: 33,86 Пфлопс(Linpack)
- Теоретическая

пиковая: 54,9 Пфлопс.

- 720 м^2
- 200—300 миллионов

долларов.

Тяньхэ-2. Compute node

16 тысяч узлов:

- 3 120 000 вычислительных ядер и 1,4 петабайта оперативной памяти
- 2 Intel Ivy Bridge CPU+3 Intel Xeon Phi
- 64 ГБ DDR3 ECC памяти (16 модулей)
- Производительность одного узла достигает 3.432 тфлопс, из них 0.422 тфлопс - за счет процессоров Ivy Bridge.

Тяньхэ-2. Процессоры

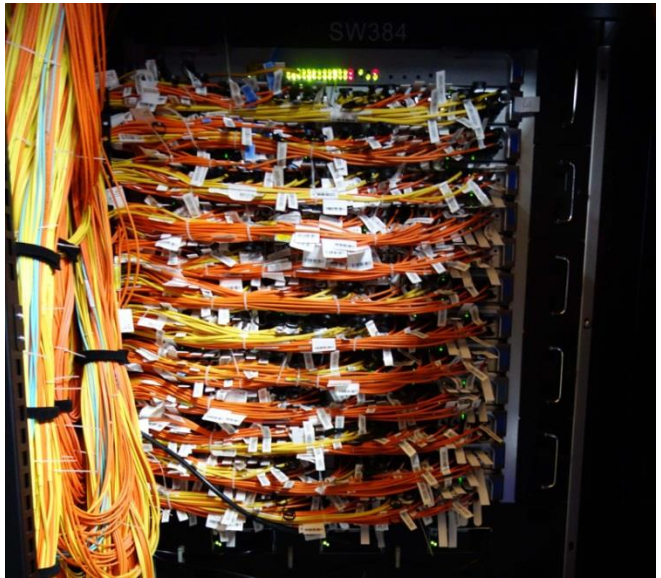
4096 процессоров Galaxy FT-1500:

- SparcV9, 16 ядер, 4 SIMD
- 40 нм, 1,8 ГГц
- Производительность: 144 Гфлопс

Тяньхэ-2. Сеть

Сеть TH-Express 2:

- топология Fat tree, использующая 13 576-портовых коммутаторов (ASIC)
- оптоэлектронные гибридные транспортные технологии.



Тяньхэ-2. Применение

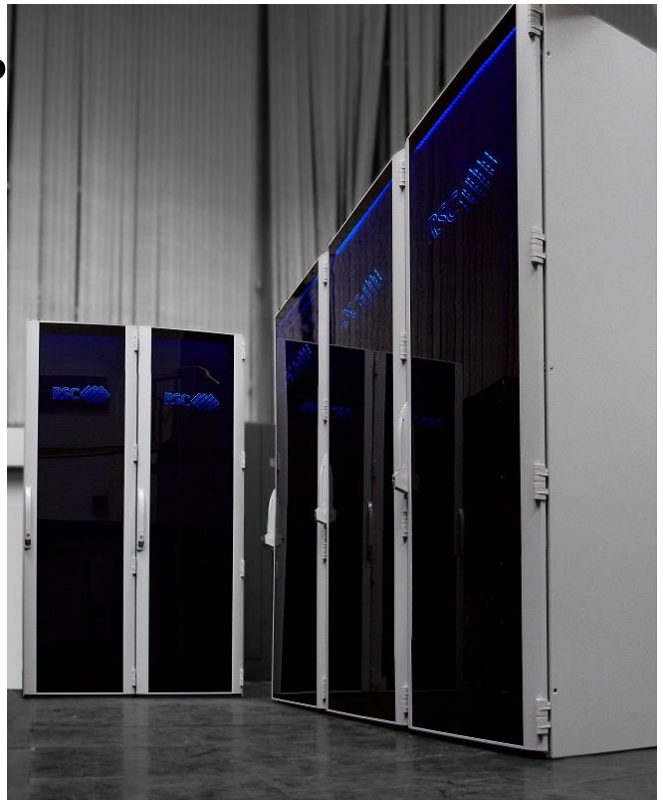
Программа 863

- Крупные проекты по освоению космического пространства
- Сейсмическое моделирование и др. исследования (газовая и нефтяная промышленность)

МВС-10П (84 место)

- 5 апреля 2013г. РСК+МСЦ РАН
- Производительность на тесте LINPACK - 375.7 ТФЛОПС
- Пиковая производительность

  523.8 ТФЛОПС.
- 20 кв. м.



МВС-10П. Вычислительный модуль

Гибридная архитектура суперЭВМ(СК)

- 2 процессора: Xeon E5-2690 (Xeon) (8 ядер каждый)
- 2 сопроцессора: Intel Xeon Phi (MIC) (61 ядро)
- Общее количество ядер составляет - 3,3 тысячи.

МВС-10П. Вычислительные узлы

207 вычислительных узлов.

Сервер очередей различает 2 варианта ВУ:

- host-узел с 16-ю процессорами (cpus) Xeon (2 процессора по 8 ядер, 32 логических процессора cpus) (nodeX)
- MIC-узел с 4-мя процессорами (cpus) MIC (1 процессор MIC с 61 ядром) (nodeX-micY).

МВС-10П. Архитектура «РСК Торнадо»

Прямое жидкостное охлаждение для Intel Xeon Phi.

- *Энергоэффективность* (4,8 раза улучшить рекорд по энергоэффективности для России и стран СНГ, достигнув значения более 1680 мегафлопс/Вт).
- *Высочайшая вычислительная плотность (*181 TFLOPS на одну стойку).

МВС-10П. Применение

- задача обработки сейсмических данных при разведке нефти (метод Wave-Equation Migration)
- приложения, построенные на реализации так называемых методов Monte-Carlo
- приложение для расчетов прогноза погоды WRF

СК «Ломоносов» версия 1

«Ломоносов»: общая информация

- Самый мощный суперкомпьютер в России и Восточной Европе
- Установлен в 2009 г. в МГУ
- В 2011 г. пиковая производительность увеличена в >2 раза
- Разработчик: компания «Т-Платформы»

«Ломоносов»: характеристики

- Пиковая производительность: 1373 Тфлопс
- Реальная производительность: 674 Тфлопс
- Оперативная память: 83 ТБ
- Энергопотребление: 2,8 МВт
- Занимаемая площадь: 252 кв. м.

«Ломоносов»: архитектура

- 4 вида вычислительных узлов:
  - TB2-XN (Intel Xeon)
    - 4160 узлов (2 Intel Xeon 5570 (4 ядра), 12 ГБ)
    - 640 узлов (2 Intel Xeon 5670 (6 ядер), 24 ГБ)
  - TB1.1 (Intel Xeon)
    - 260 узлов (2 Intel Xeon 5570 (4 ядра), 24 ГБ)
    - 40 узлов (2 Intel Xeon 5670 (6 ядер), 48 ГБ)
  - PeakCell S
    - 30 узлов (2 PowerXCEll 8i (8 ядер), 16 ГБ)
  - TB2-TL (Intel Xeon, NVIDIA Tesla)
    - 777 узлов (2 NVIDIA Tesla X2070, 2 Intel Xeon 5630)

«Ломоносов»: архитектура

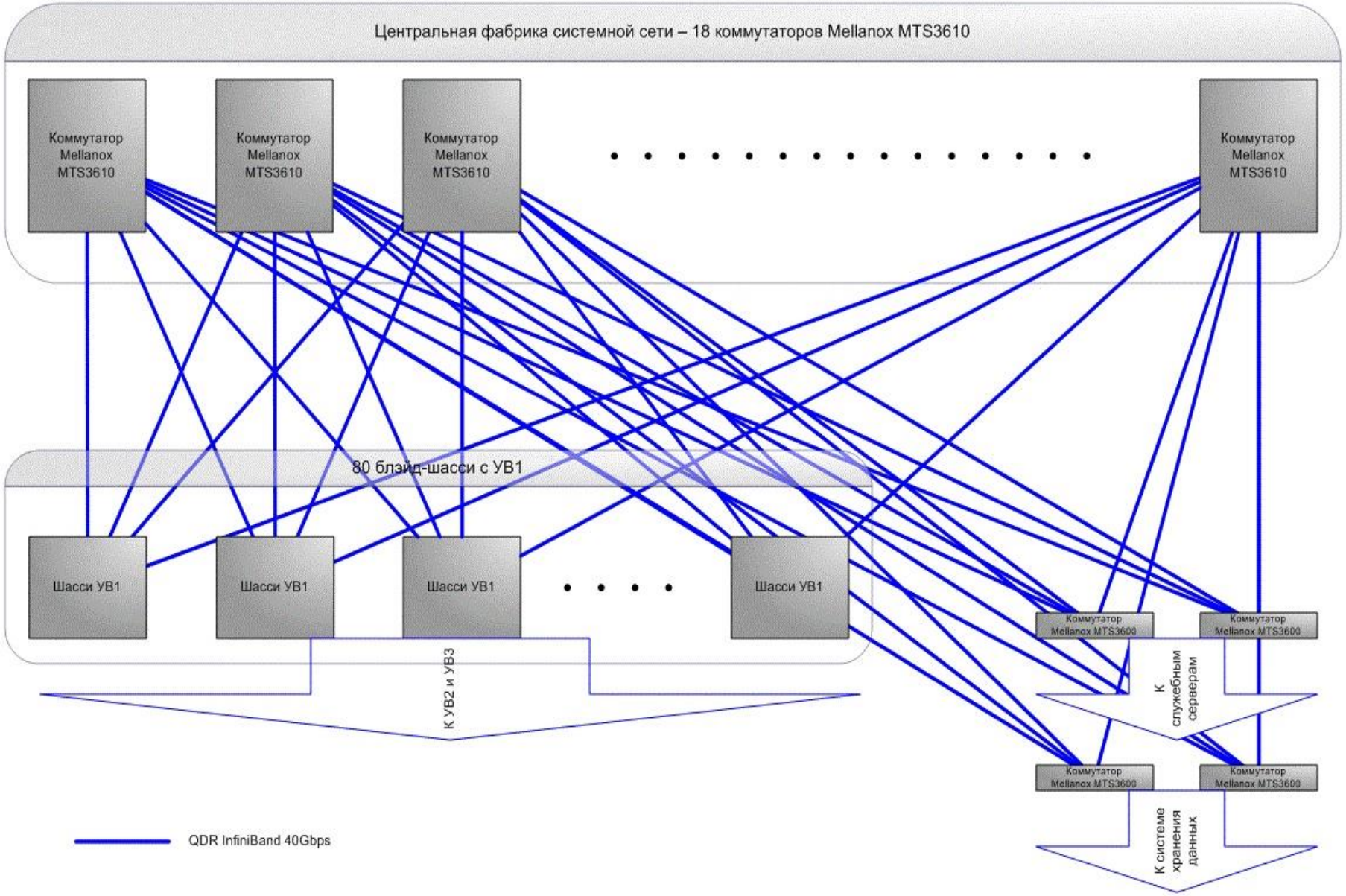
- 3-уровневая система хранения данных:
  - высокоскоростное дисковое хранилище (500 ТБ, параллельная файловая система Lustre)
  - хранилище повышенной надежности (300 ТБ NAS)
  - роботизированная ленточная библиотека для резервного копирования данных (1000 ТБ)

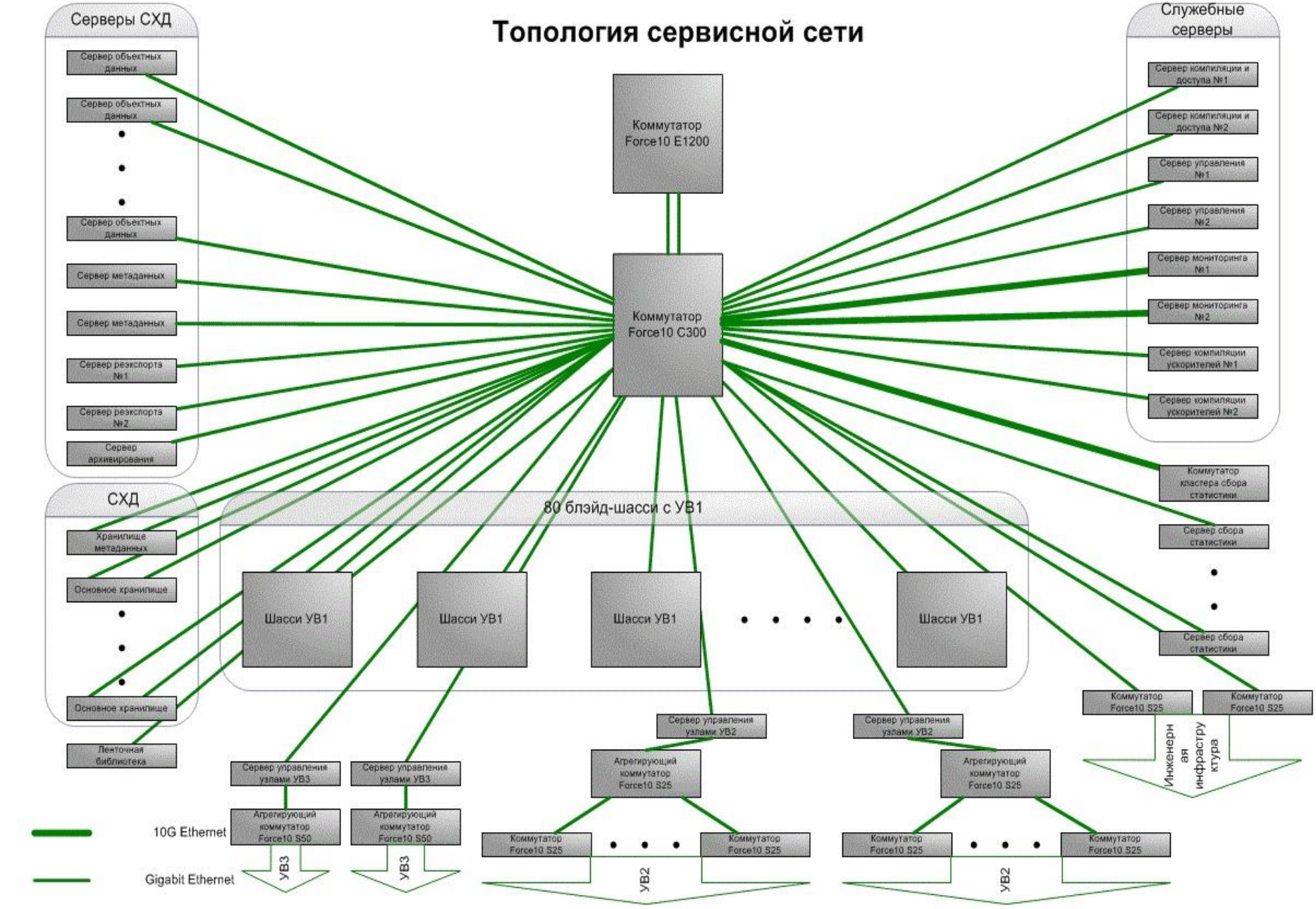
«Ломоносов»: архитектура

Узлы связаны 3 независимыми сетями:

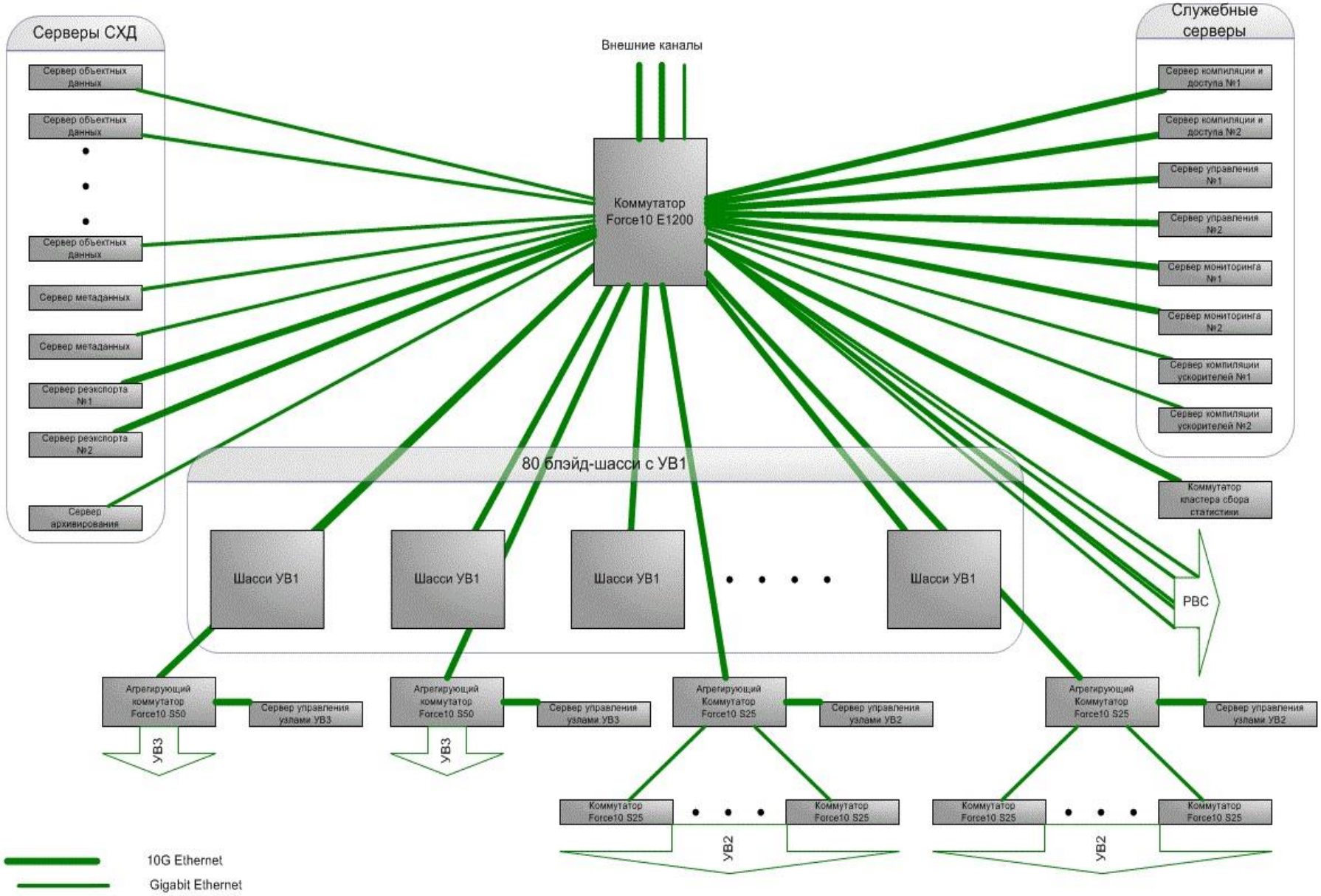
- Системная сеть - QDR InfiniBand, 40 Гбит/сек
- Сервисная сеть - Ethernet, 10 Гбит/сек, 1 Гбит/сек и 100 Мбит/сек
- Управляющая сеть - Ethernet, 10 Гбит/сек и 1 Гбит/сек

Топология системной сети




Топология сервисной сети
Серверы СХД
Сервер объектных данных
Сервер объектных данных
Сервер объектных данных
Сервер метаданных
Сервер метаданных
Сервер реэкспорта №1
Сервер реэкспорта №2
Сервер архивирования
СХД
Хранилище метаданных
Основное хранилище
Основное хранилище
Ленточная библиотека
Коммутатор Force10 E1200
Коммутатор Force10 C300
80 блэйд-шасси с УВ1
Шасси УВ1
Шасси УВ1
Шасси УВ1
Шасси УВ1
Служебные серверы
Сервер компиляции и доступа №1
Сервер компиляции и доступа №2
Сервер управления №1
Сервер управления №2
Сервер мониторинга №1
Сервер мониторинга №2
Сервер компиляции ускорителей №1
Сервер компиляции ускорителей №2
Коммутатор кластера сбора статистики
Сервер сбора статистики
Сервер сбора статистики
Коммутатор Force10 S25
Коммутатор Force10 S25
Инженерная инфраструктура
Сервер управления узлами УВ3
Сервер управления узлами УВ3
Агрегирующий коммутатор Force10 S50
Агрегирующий коммутатор Force10 S50
УВ3
УВ3
Сервер управления узлами УВ2
Сервер управления узлами УВ2
Агрегирующий коммутатор Force10 S25
Агрегирующий коммутатор Force10 S25
Коммутатор Force10 S25
Коммутатор Force10 S25
Коммутатор Force10 S25
Коммутатор Force10 S25
УВ2
УВ2
10G Ethernet
Gigabit Ethernet

Топология управляющей сети



«Ломоносов»: применение

- Изменения климата
- Динамика атмосферы
- Моделирование при разработке нанотехнологий
- Поиск молекул при разработке лекарств

«Ломоносов»: особенности

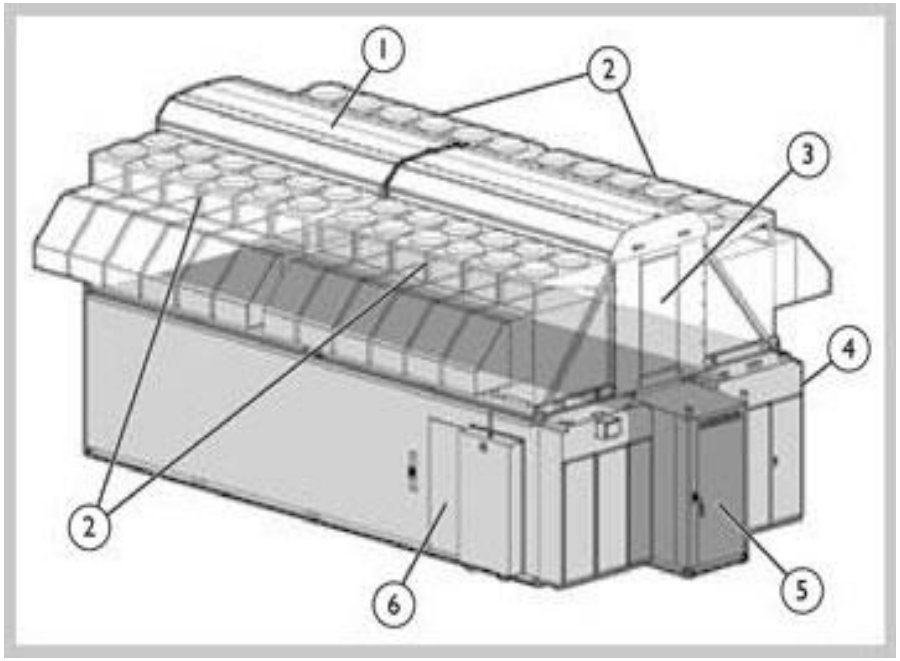
- Высокая вычислительная плотность
- Специализированная ОС Clustrx
- Система теплоотвода и режим охлаждения “free-cooling”

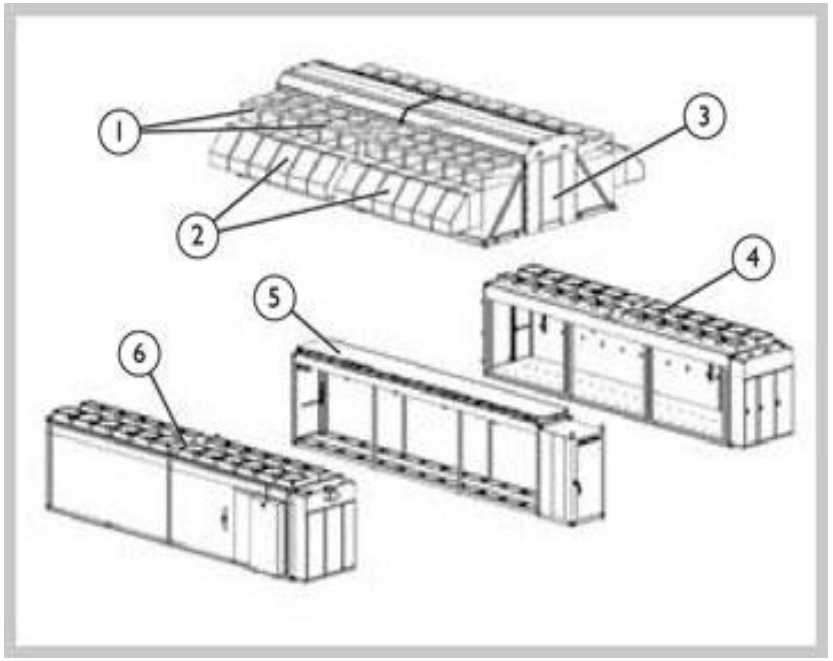
Airbus «EcoPOD»

«EcoPOD»: характеристики

- Пиковая производительность: 296 Тфлопс
- Реальная производительность: 244 Тфлопс
- Число ядер: 24192
- Энергопотребление: 2*1.3 МВт
- Занимаемая площадь: 2*89 кв. м.
- Расположение: г. Тулуза, г. Гамбург

«EcoPOD»: схема





- 1, 2, 3, 5 – система охлаждения
- 4, 6 – IT-модули

«EcoPOD»: преимущества

- Энергоэффективность
- Высокая вычислительная плотность
- Быстрота запуска
- Вариативность

Облако

Модель обеспечения сетевого доступа по требованию к общему пулу конфигурируемых вычислительных ресурсов

Обязательные характеристики облака

- Универсальный доступ по сети
- Объединение ресурсов *(resource pooling)*
- Эластичность
- Самообслуживание по требованию *(self service on demand)*
- Учёт потребления ресурсов

Модели обслуживания

- **SaaS** *(Software-as-a-Service)*
- **PaaS** *(Platform-as-a-Service)*
- **IaaS** *(Infrastructure-as-a-Service)*

Характеристики облаков

- ***Самообслуживание по требованию*** (*self service on demand*) — потребитель самостоятельно определяет и изменяет вычислительные потребности, такие как серверное время, скорости доступа и обработки данных, объём хранимых данных без взаимодействия с представителем поставщика услуг;
- ***Универсальный доступ по сети*** — услуги доступны потребителям по сети передачи данных вне зависимости от используемого терминального устройства;
- ***Объединение ресурсов*** (*resource pooling*) — поставщик услуг объединяет ресурсы для обслуживания большого числа потребителей в единый *пул* для динамического перераспределения мощностей между потребителями в условиях постоянного изменения спроса на мощности; при этом потребители контролируют только основные параметры услуги (например, объём данных, скорость доступа), но фактическое распределение ресурсов, предоставляемых потребителю, осуществляет поставщик (в некоторых случаях потребители всё-таки могут управлять некоторыми физическими параметрами перераспределения, например, указывать желаемый центр обработки данных из соображений географической близости);
- ***Эластичность*** — услуги могут быть предоставлены, расширены, сужены в любой момент времени, без дополнительных издержек на взаимодействие с поставщиком, как правило, в автоматическом режиме;
- ***Учёт потребления*** — поставщик услуг автоматически исчисляет потреблённые ресурсы на определённом уровне абстракции (например, объём хранимых данных, пропускная способность, количество пользователей, количество транзакций), и на основе этих данных оценивает объём предоставленных потребителям услуг.

Виды облаков

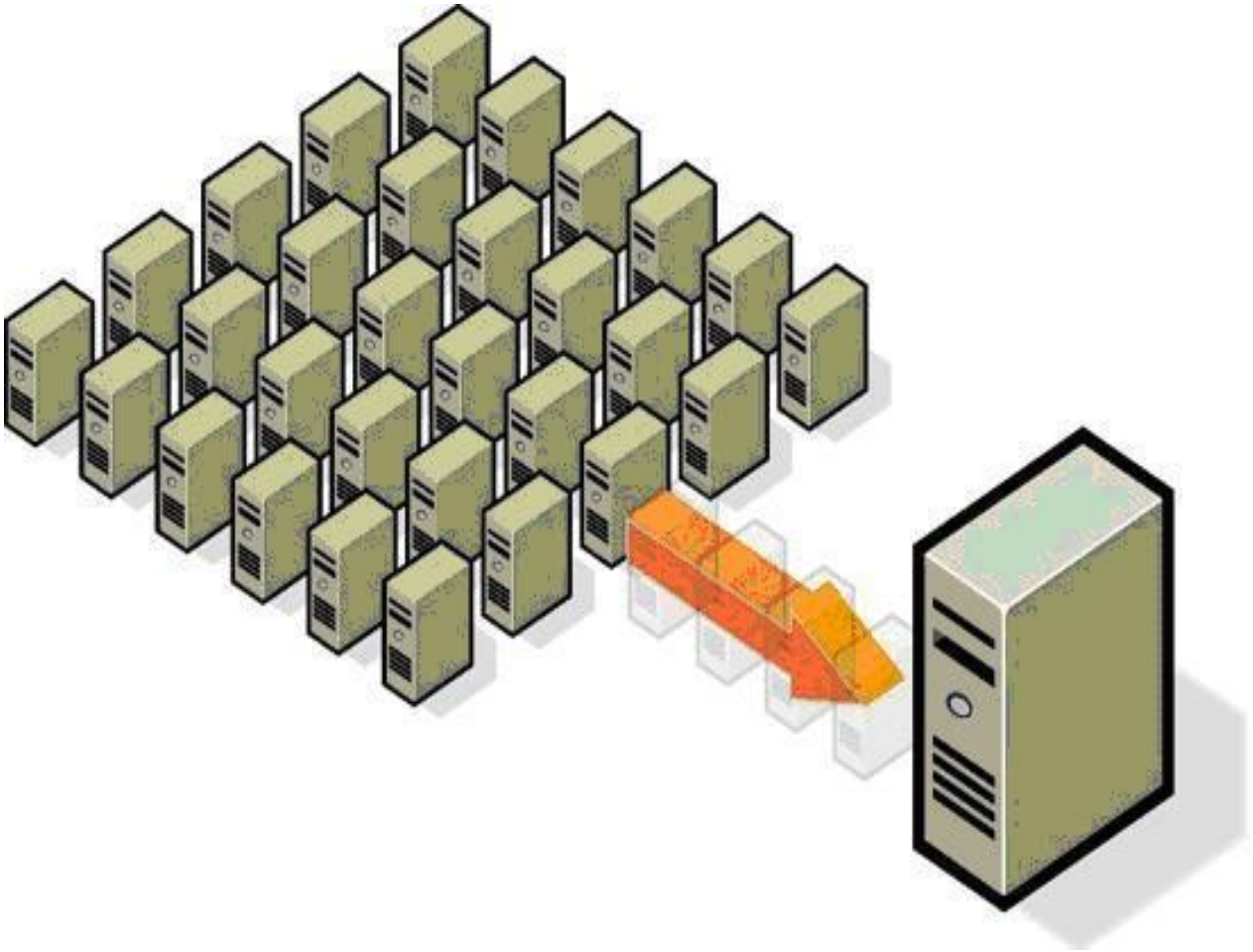
- **Частное облако** (*private cloud*) — инфраструктура, предназначенная для использования одной организацией, включающей несколько потребителей (например, подразделений одной организации), возможно также клиентами и подрядчиками данной организации. Частное облако может находиться в собственности, управлении и эксплуатации как самой организации, так и третьей стороны (или какой-либо их комбинации), и оно может физически существовать как внутри, так и вне юрисдикции владельца.

- **Публичное облако** (*public cloud*) — инфраструктура, предназначенная для свободного использования широкой публикой. Публичное облако может находиться в собственности, управлении и эксплуатации коммерческих, научных и правительственных организаций (или какой-либо их комбинации). Публичное облако физически существует в юрисдикции владельца — поставщика услуг.

- **Общественное облако** (*community cloud*) — вид инфраструктуры, предназначенный для использования конкретным сообществом потребителей из организаций, имеющих общие задачи (например, миссии, требований безопасности, политики, и соответствия различным требованиям). Общественное облако может находиться в кооперативной (совместной) собственности, управлении и эксплуатации одной или более из организаций сообщества или третьей стороны (или какой-либо их комбинации), и оно может физически существовать как внутри, так и вне юрисдикции владельца.

- **Гибридное облако** (*hybrid cloud*) — это комбинация из двух или более различных облачных инфраструктур (частных, публичных или общественных), остающихся уникальными объектами, но связанных между собой стандартизованными или частными технологиями передачи данных и приложений (например, кратковременное использование ресурсов публичных облаков для балансировки нагрузки между облаками).

Виртуализация



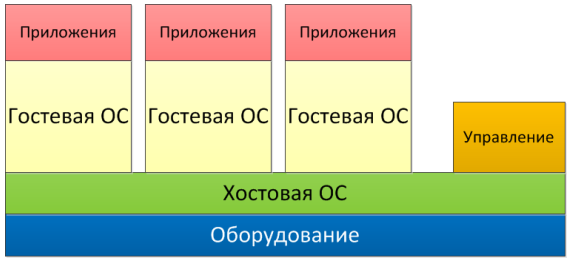
Преимущества виртуализации

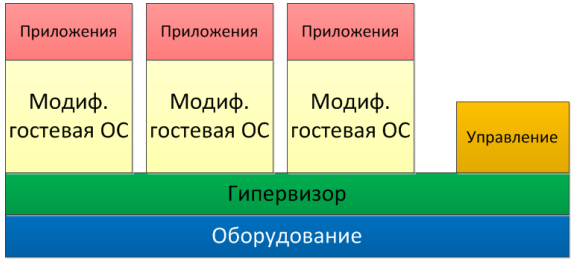
- Эффективное использование вычислительных ресурсов;
- Сокращение расходов на инфраструктуру;
- Снижение затрат на программное обеспечение;
- Повышение гибкости и скорости реагирования системы;
- Несовместимые приложения могут работать на одном компьютере;
- Повышение доступности приложений и обеспечение непрерывности работы предприятия;
- Возможности легкой архивации;
- Повышение управляемости инфраструктуры.

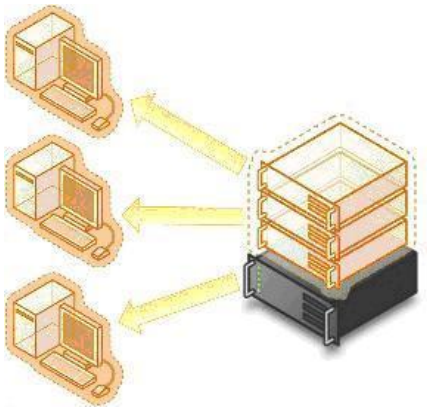
Виртуальной машиной *будем называть программную или аппаратную среду, которая скрывает настоящую реализацию какого-либо процесса или объекта от его видимого представления*.

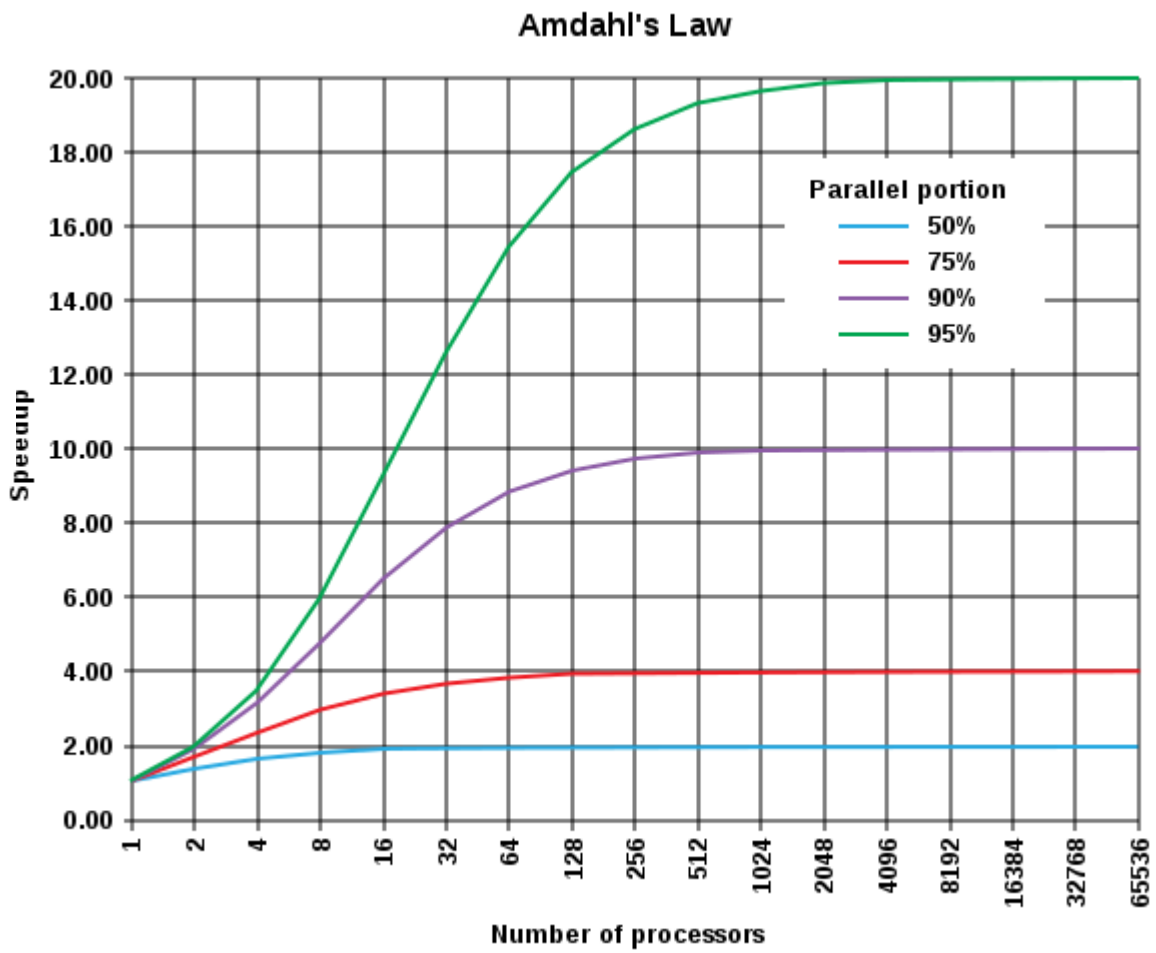
Виды виртуализации

1. виртуализация серверов
   – полная виртуализация
   – паравиртуализация
2. виртуализация на уровне операционных систем;
3. виртуализация приложений;
4. виртуализация представлений.







Закон Амдала

$$S_p = \frac{1}{\alpha + \frac{1-\alpha}{p}}$$