Аппаратные решения для построения вычислительных систем. От «персонального кластера» до 100-терафлопного суперкомпьютера

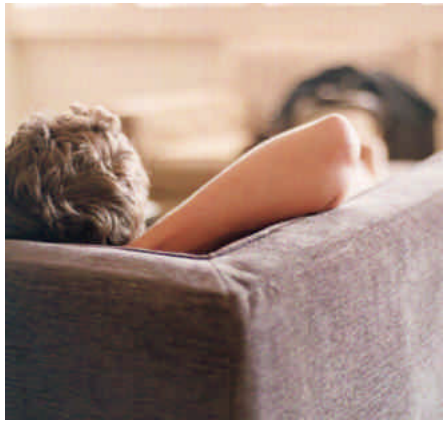
Форум программы «Университетский кластер»,
январь 2009

Евгений Лагунцов
Системный архитектор, HP Ambassador,
Руководитель направления
«Масштабируемые вычислительные комплексы»,
HP Россия

Поговорим о…

- Проблемы
- Тенденции
- Платформы для вычислительных комплексов
- Аппаратное обеспечение «Университетского кластера»
- А что вокруг?

Основные проблемы, стоящие перед администраторами, пользователями, разработчиками вычислительных комплексов и отраслью в целом

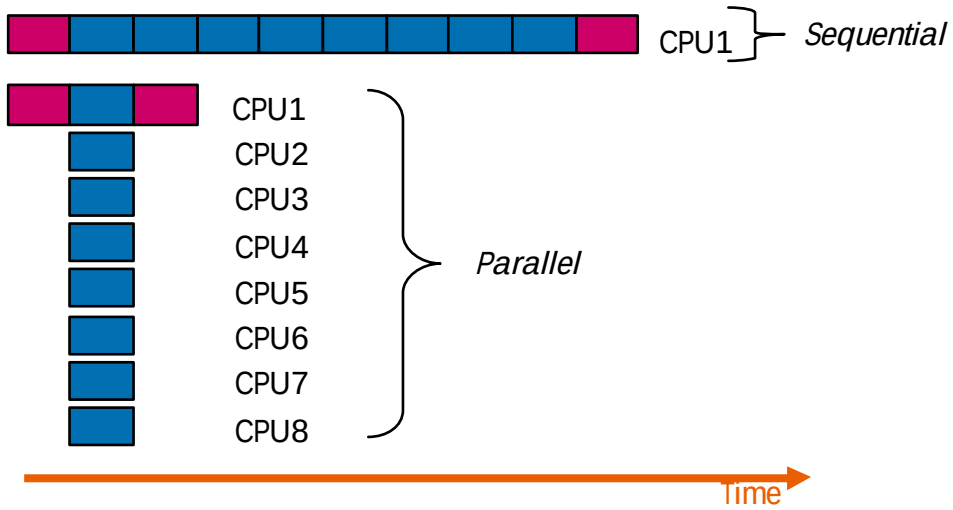


Первое и самое главное – прикладная часть

- Система должна быть востребована
- Вычислительный комплекс сам по себе никому не нужен – нужно решение конкретных задач
- Те, перед кем стоят эти задачи, зачастую не знают о том, что суперкомпьютер в принципе может им помочь
- Те, кто об этом догадываются, зачастую не знают как именно
- Чтобы решать конкретные задачи, должен быть написан соответствующий программный код
- Этот код должен эффективно (и быстро) работать на данном аппаратном обеспечении
- Разработка эффективного кода для современных вычислительных систем – огромная проблема

Как получить высокую производительность?

- Единственный способ ускорить выполнение расчетов – выполнять их параллельно на большом количестве процессоров
- Просто ли это сделать? Нет. Не все алгоритмы допускают эффективное распараллеливание
- В идеале: 1000 процессоров выполняют программу в 1000 раз быстрее одного
- В реальной жизни (закон Амдала):
  - В любой программе есть фазы которые можно распараллелить, и фазы, которые должны выполняться последовательно
  - Между параллельными ветвями всегда есть взаимодействие и конкуренция
  - Программа, параллельная на 90%, никогда не будет выполняться быстрее более чем в 10 раз, даже на 1000 процессоров



Не все параллельные процессы одинаково параллельны…

- Мало просто написать параллельный код, необходимо принимать во внимание специфику аппаратуры, прежде всего иерархию памяти и ее удаленность от процессора:
  - кэш ядра,
  - разделяемый кэш процессора,
  - локальная память,
  - удаленная память в том же узле,
  - другой узел в том же «сегменте» коммуникационной сети,
  - другой узел в удаленном «сегменте» сети.
- Два общающихся между собой процесса могут быть в одном процессоре, а могут – в разных узлах
- Только за счет изменения привязки процессов к физическим процессорам возможно обеспечить повышение реальной производительности на десятки процентов, а иногда и в разы – благодаря более эффективному обмену между процессорами и снижению конкуренции за доступ к памяти

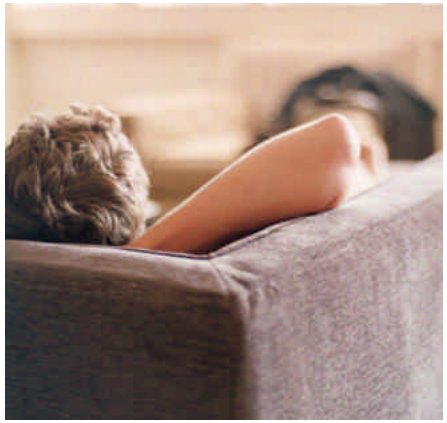
Современные системы не очень сбалансированы для нагрузок HPC

- Один современный процессор теоретически может сделать 4cores*4flops/tick*3GHz=48GFlops, 48 МИЛЛИАРДОВ операций с плавающей точкой двойной точности в секунду
- Один такт процессора – доли наносекунды; а задержка при обращении к оперативной памяти – десятки наносекунд, т.е. десятки или даже сотни тактов
- Главная проблема : обеспечить должную пропускную способность от процессора к памяти – процессору нужно «скормить» те данные, которые он теоретически может обработать
- Зная архитектуру и используя средства профилирования, только за счет минимальных изменений в коде (например, переставления операторов) возможно получить повышение производительности на десятки процентов, а иногда и в разы – благодаря более активному использованию кэша и более эффективной работе с памятью

Второе, но не менее главное – технические трудности

- Наличие места
- Электропитание
- Теплоотвод
- Кабели-кабели-кабели
- Шум
- Эффективный мониторинг и управление
- Простота в обслуживании
- Возможности развития, апгрейда
- Вопросы совместимости и переносимости

Тенденции в отрасли. Куда идут высокопроизводительные вычисления?



1. Стандартизация

- Процессорная архитектура x86 – Intel Xeon или AMD Opteron
- Стандартные микросхемы адаптеров: Broadcom, Mellanox, QLogic, Emulex…
- Стандартные микросхемы коммутаторов: BNT, Cisco, Mellanox…
- ОС Linux или Microsoft Windows

- Как следствие – полная совместимость со сложившейся экосистемой, стандартные драйверы, стандартное ПО включая коммерческое и open-source

Microsoft Windows Compute Cluster Server 2003

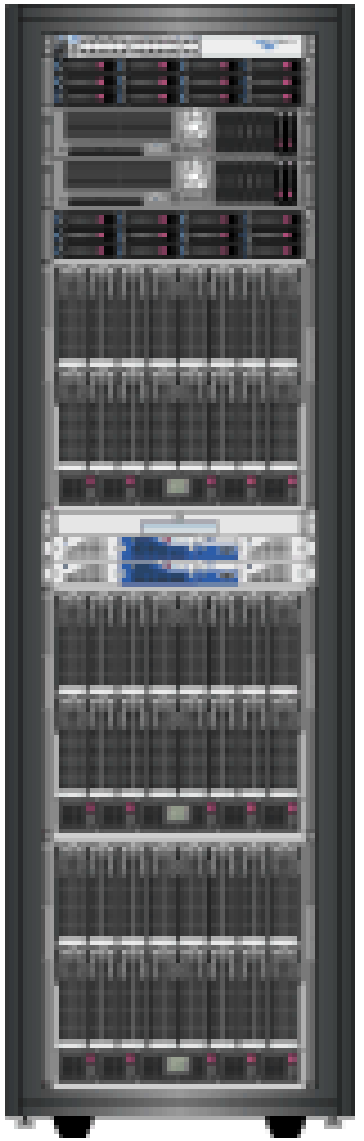
Mellanox TECHNOLOGIES

intel Xeon inside

hp invent

2. Блейдизация

- Максимальная компакность (до 12ТФлоп, 1024 ядер, 8ТБ памяти на шкаф)
- Интегрированные коммуникационные системы GigE и Infiniband
- Средства оптимизации систем электропитания и теплоотвода
- Расширенные средства управления и мониторинга
- Линейная масштабируемость от единиц до тысяч вычислительных узлов
- Подавляющее большинство систем в top500 – на базе блейд-решений, из них на базе HP BladeSystem – 201 (более 40%)



3. Персонализация

Развитие «персональных суперкомпьютеров»:

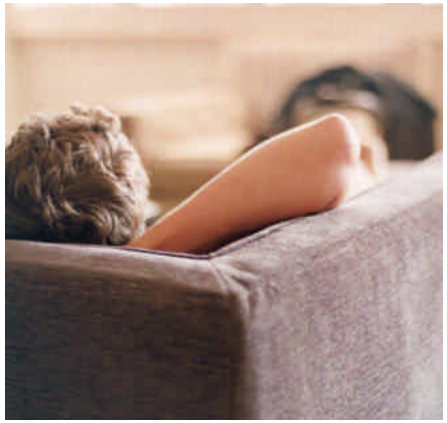
- Простота управления
- Компактность
- Производительность порядка 1ТФлоп
- Простота подключения
- Легкость обслуживания
- Отсутствие серьезных требований к инфраструктуре
- Отсутствие кабелей
- Те же технологии, что используются для построения больших систем, но в уменьшенном варианте

4. Диверсификация

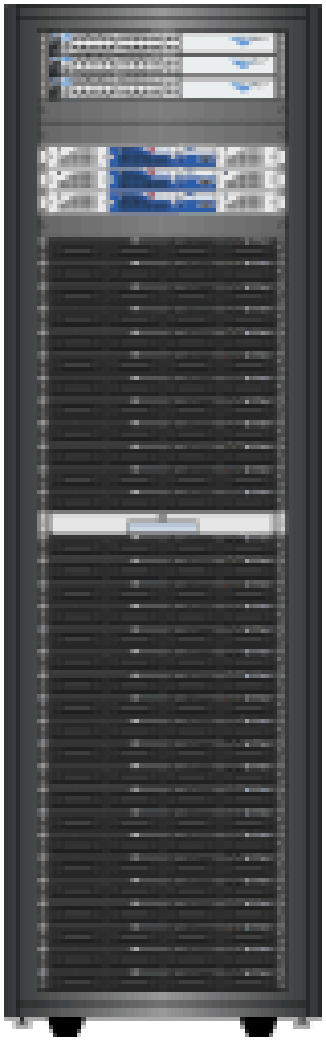
5. Акселерация

- Перенос вычислений с центральных процессоров на специфические устройства – GPU, FPGA
- Разнообразные способы подключения:
  - Слот PCI
  - Процессорный разъем
  - Слот HTX
  - Даже слоты DIMM
- Фантастическая теоретическая производительность – до 100 раз быстрее обычного x86
- НО не на всех операциях
- Требуют специфического подхода к программированию
- Многие алгоритмы не реализуются в принципе

Платформы для вычислительных комплексов. Что такое HP BladeSystem

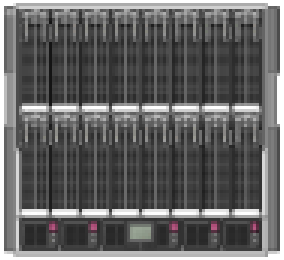


BladeSystem против традиционных серверов. Цифры.



Вариант 1.

- 32 сервера (2x Xeon 5450, 16GB, InfiniBand DDR, 2xGE, 250GB HDD)
- 3 коммутатора InfiniBand DDR
- 2 коммутатора 1/10GigE
- 1 консольный коммутатор
- Монтажный шкаф 42U
- Энергопотребление: **14.7кВт**
- 192 кабеля внутри (2x32 UTP, 32 IB, 2x32 10A-Power, 1x32 Console)
- 319,668 очень условных единиц



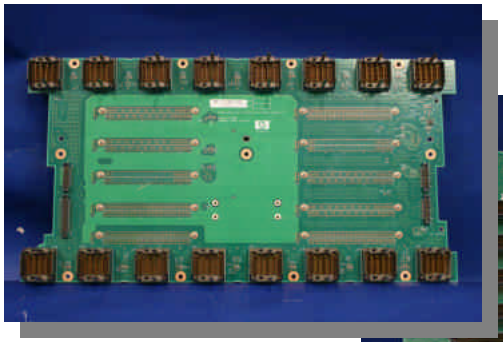
Вариант 2.

- 32 сервера (2x Xeon 5450, 16GB, 2xGE, InfiniBand DDR, 250GB HDD)
- 4 интегрированных коммутатора 1/10Gb
- 2 интегрированных коммутатора InfiniBand 16:16
- Серверная полка 10U (19”)
- Энергопотребление: **8.4кВт**
- 0 кабелей внутри
- 299,682 очень условных единиц

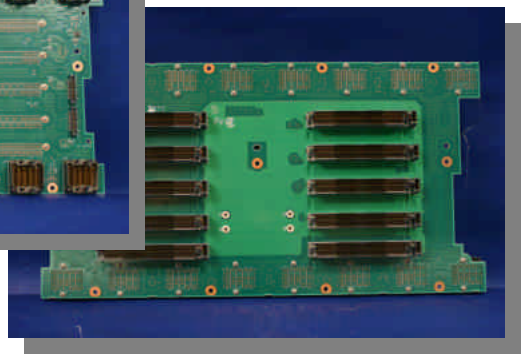
Блейд-инфраструктура HP

Серверная полка

Вентилятор



Сигнальная панель и панель питания



Блок питания

Интегрированный коммутатор

Адаптер ввода-вывода

Блейд-сервер

Модуль управления

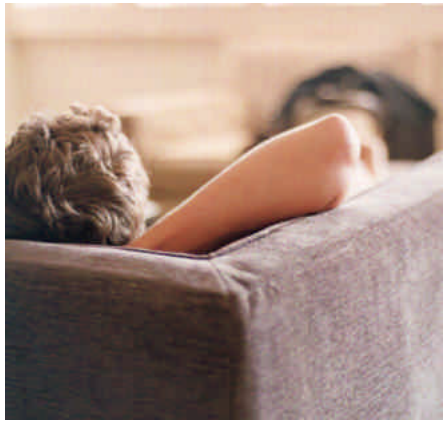
Блейд-серверы

Двухпроцессорные серверы
до 8 ядер, до 128GB RAM

Четырехпроцессорные серверы
до 24 ядер, до 192GB RAM

Модули расширения,
блейд-системы хранения,
блейд-рабочие станции

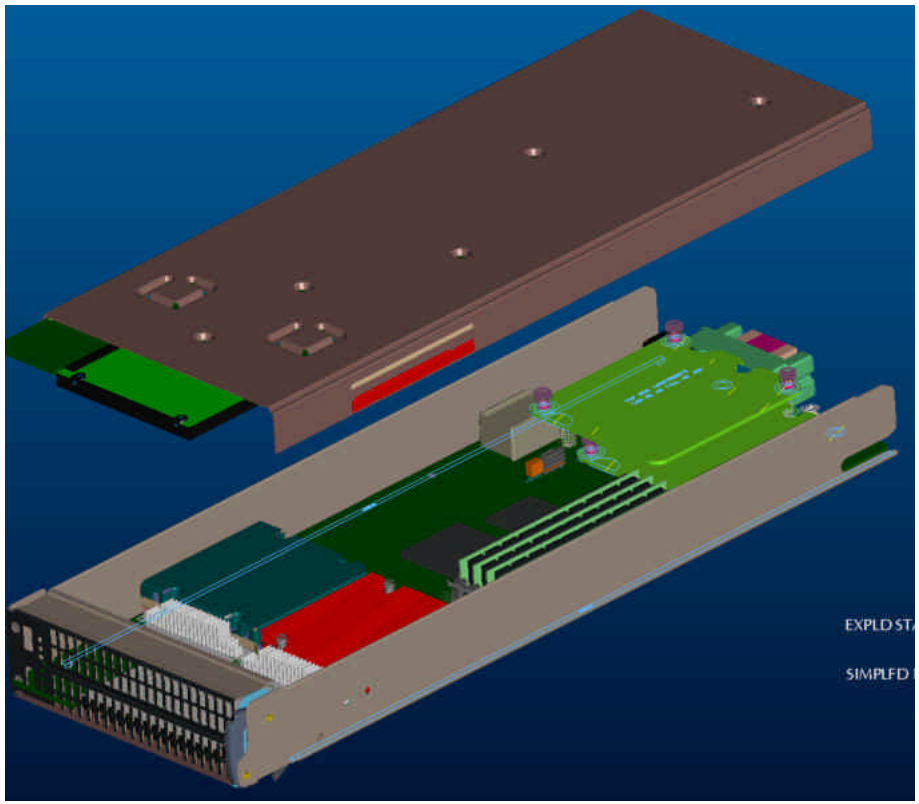
Платформа для «Университетского кластера»



HP BladeSystem c3000 – решение для построения компактных высокопроизводительных систем

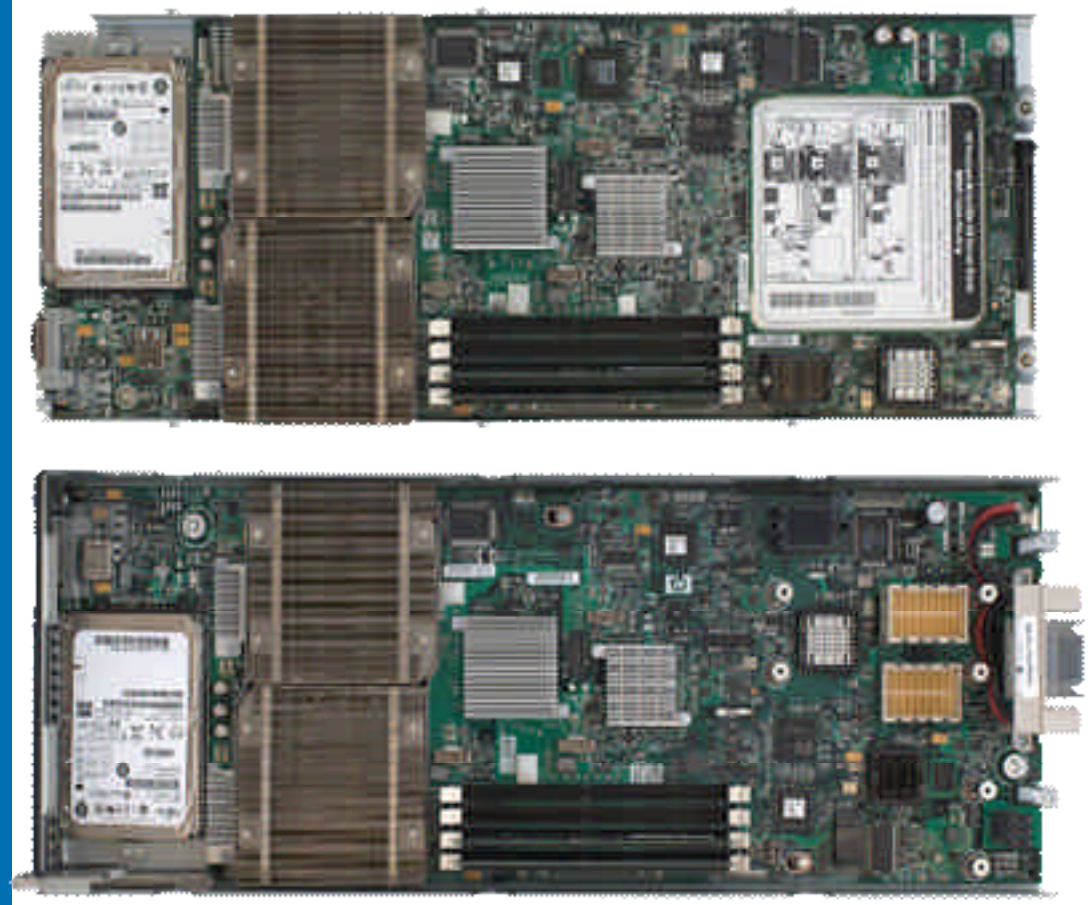
- Полка HP BladeSystem c3000 – 6U в шкафу или менее 0,5 кв.м. «под столом»
- Включается в стандартные розетки 220В
- До 16 вычислительных узлов – до 1,5ТФлоп
- До 1024ГБ оперативной памяти
- Интегрированные средства сетевых коммуникаций:
  - Gigabit Ethernet,
  - 10GigE (опционально),
  - Infiniband DDR (ConnectX, InfiniScale IV),
  - FC 8Gbps (опционально).
- Возможность интеграции систем хранения данных – до 2ТВ на узел

HP ProLiant BL2x220c



- 2 полнофункциональных сервера в одном конструктиве => 32 сервера на серверную полку 10U, **128 серверов** и **1024 ядра** на монтажный шкаф 42U
- Специализированный чипсет, память DDR2 => уникальное соотношение **производительность/Вт**
- Никакого comminucation sharing – каждый сервер работает со своими **выделенными линиями** сетевых коммуникаций
- Очевидный минус: серверы можно вынимать из полки только парами. Для большинства scale-out систем это непринципиально

HP ProLiant BL2x220c



Каждый узел (их два в одном конструктиве) содержит:

- 2 процессора Dual-, Quad-core Intel Xeon 5200 и 5400
- 4 слота под оперативную память (до 32GB)
- диск SATA 250GB или SSD 64GB
- 2 интегрированных адаптера GigE
- 1 дополнительный слот для установки плат расширения (GigE, 10GigE, FC, Infiniband DDR)
- Контроллер управления iLO

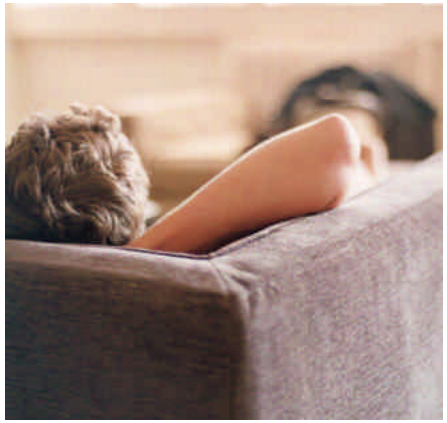
Конфигурация «Учебного кластера»

- Полка HP BladeSystem c3000 в напольном исполнении
- Интегрированный коммутатор GigE
- Управляющий узел BL260cG5: Xeon 5410 2,33GHz, 4GB RAM, 2x120GB, 2xGE
- 4 вычислительных узла на базе BL2x220cG5: 2xXeon 5410 2,33GHz, 8GB RAM, 120GB HDD, 2xGE
- Пиковая производительность – 335GFlops (еще год назад эта система могла бы попасть в top50)
- В эту же полку можно поставить еще 10 вычислительных узлов, добавить InfiniBand – получим 1.1TFlop и место в текущем рейтинге top50 (сентябрь 2008)

Есть куда расти. Например, 50TFlops

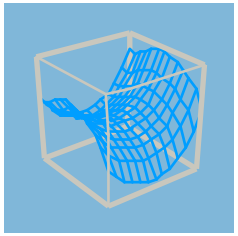
- 49,1TFlops пиковой производительности;
- 512 вычислительных узлов;
- до 16ТБ оперативной памяти;
- Параллельная система хранения на 144ТБ «сырого» пространства;
- Коммуникационная сеть Infiniband DDR 1:1;
- Энергопотребление – порядка 130кВт;
- 6 (!!!) стандартных монтажных шкафов 42U.

Вычислительный комплекс – это не только вычислительные узлы. Что вокруг?

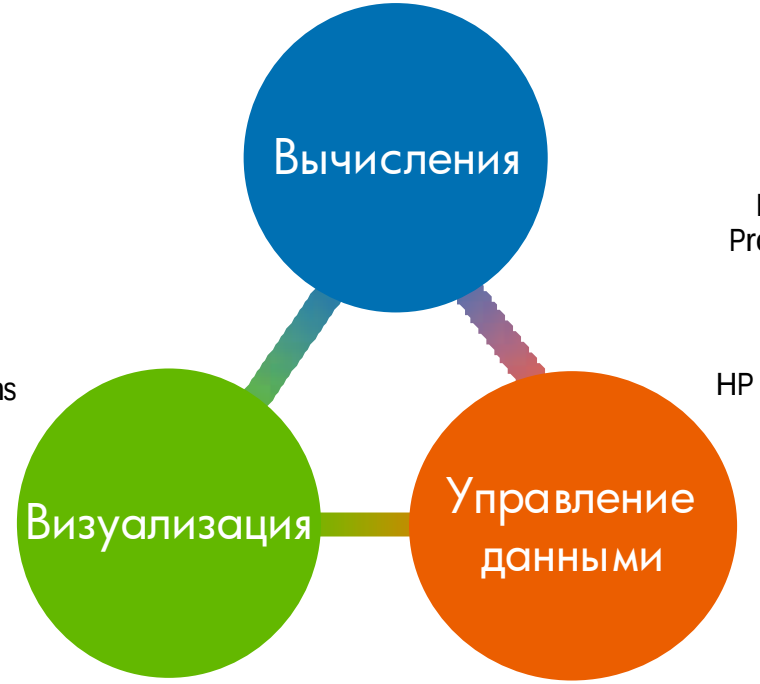


Высокопроизводительные вычисления
Важно не просто посчитать

HP Cluster Platforms

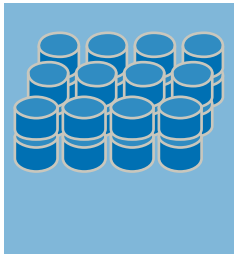


HP Workstations Scalable Visualization



HP Integrity & ProLiant Servers

HP StorageWorks Scalable File Share



Необходимые компоненты вычислительного комплекса:

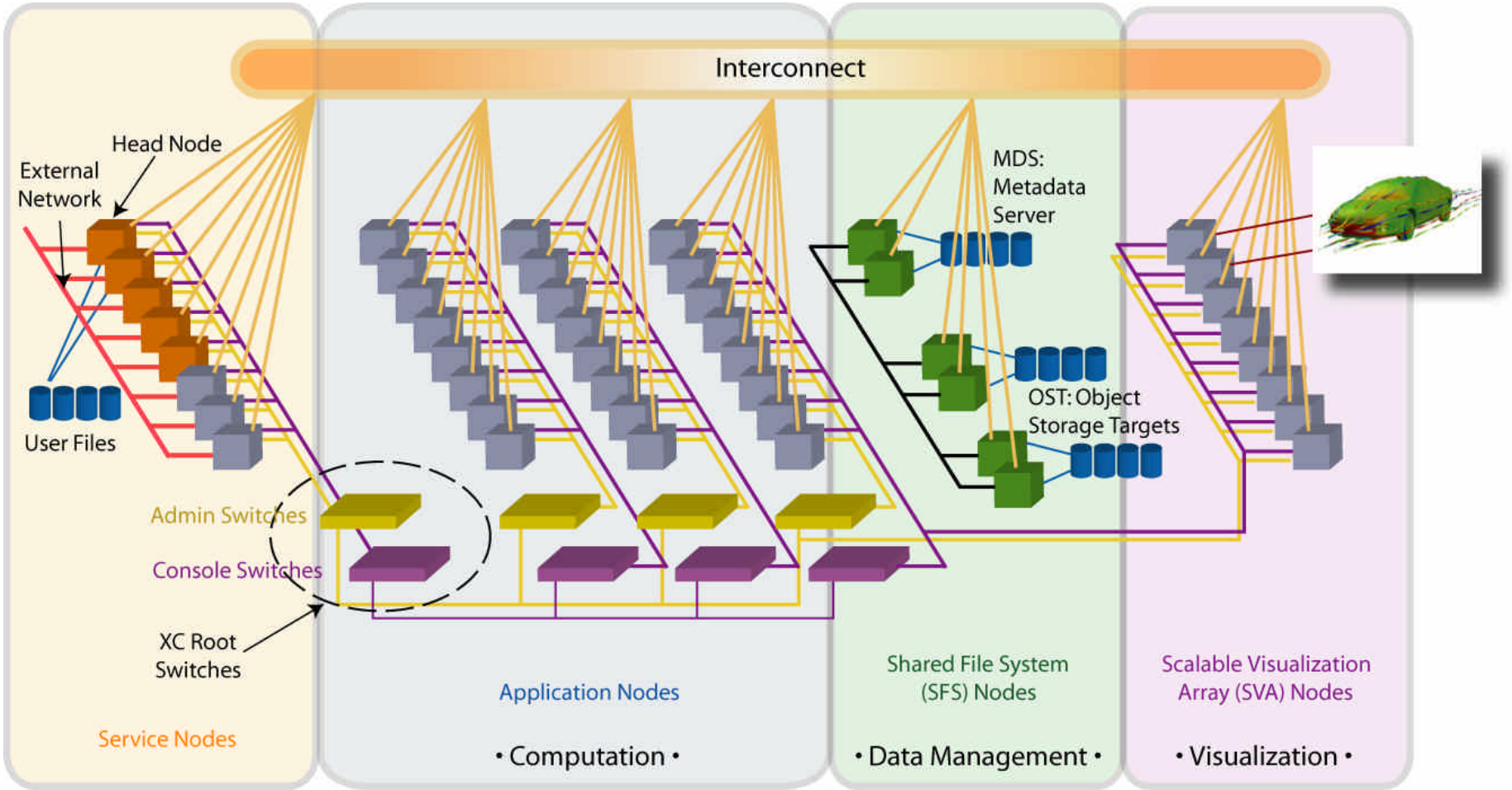
- Вычислительная платформа
- Система хранения данных, обеспечивающая хранения массивов данных и максимально производительный доступ к ним
- Система визуализации результатов

объединенные

- высокопроизводительными сетевыми соединениями
- системным и управляющим программным обеспечением

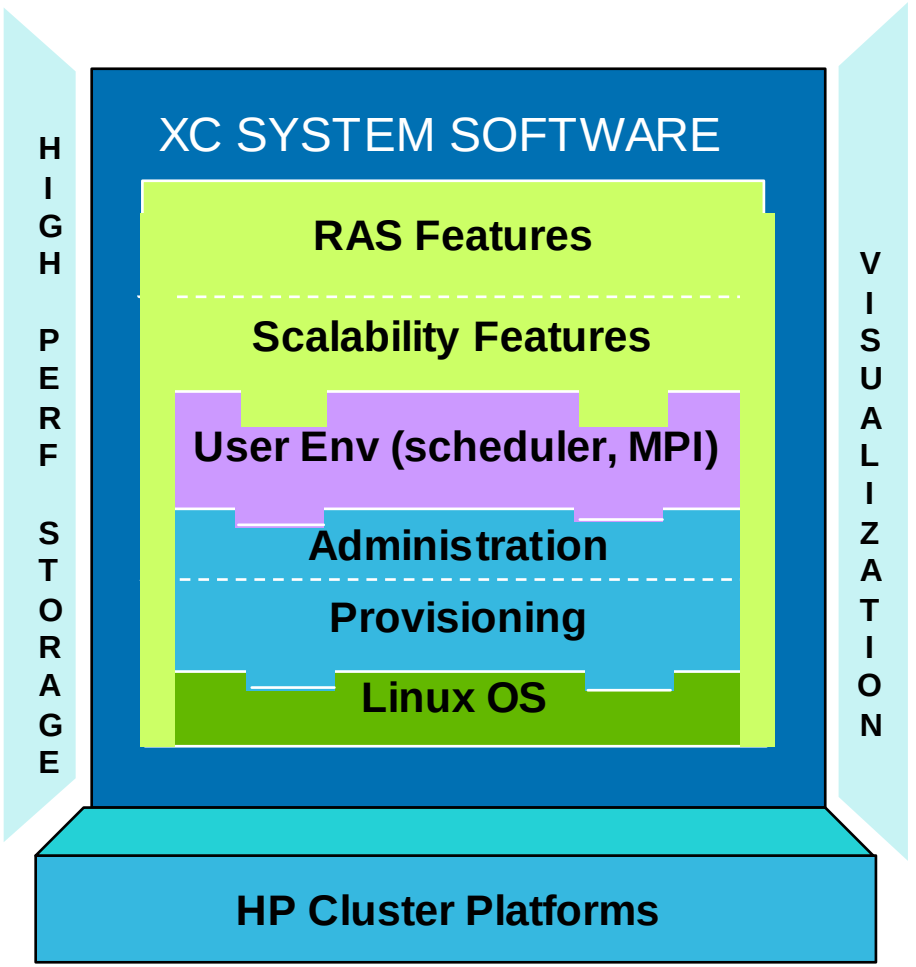
Общая архитектура HP Cluster Platform

Интегрированное решение: XC, SFS и SVA

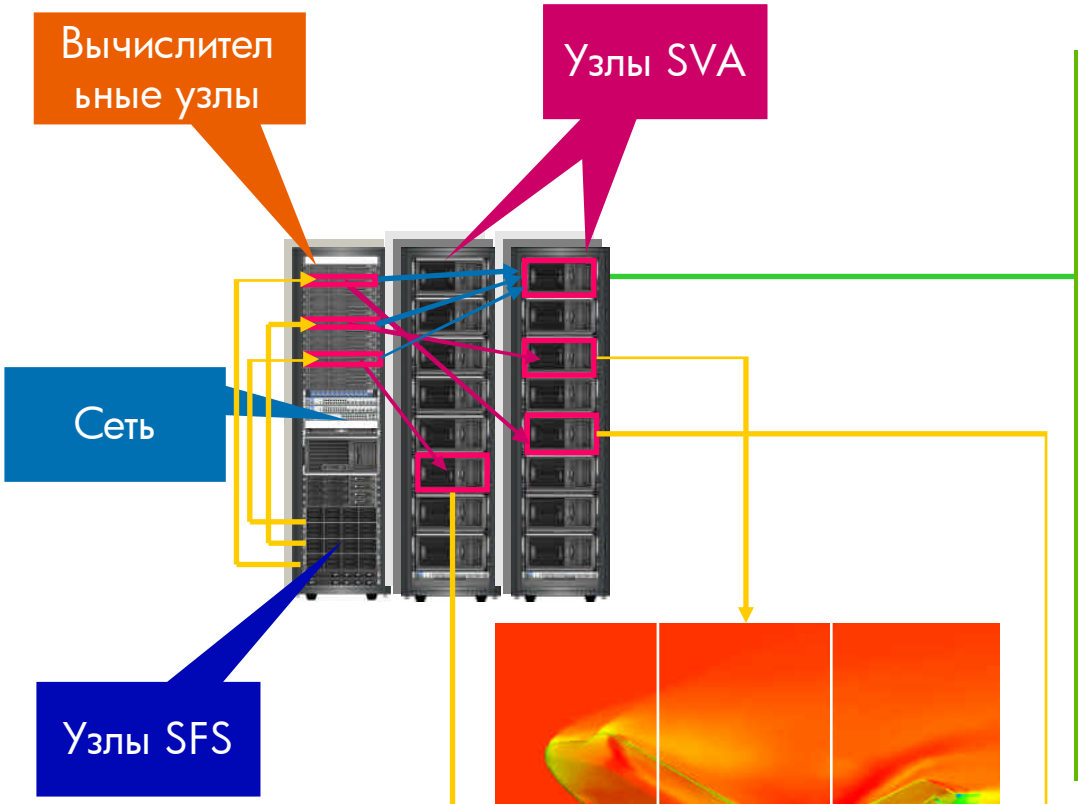


HP XC – интегрированная система

- Полностью законченное интегрированное решение
- Включает в себя все необходимые системные средства:
  - Операционная система,
  - Средства развертывания, мониторинга и управления,
  - Средства управления заданиями,
  - Библиотека MPI.
- Простота ввода в эксплуатацию (5 команд для установки всей системы)
- Масштабируемость до тысяч узлов
- Интеграция с системами визуализации и системами хранения данных
- Единая точка входа по всем вопросам

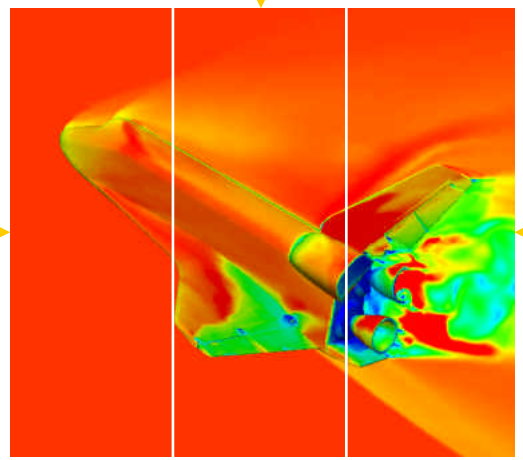


Scalable Visualization Array (SVA)



- ✓ Стандартные рабочие станции
- ✓ Профессиональные графические карты
- ✓ Параллельный рендеринг и композирование
- ✓ Одно или мульти-панельные устройства вывода
- ✓ Удаленная визуализация

Usage example:
CFD post-processing



Масштабирование
Разделяемость
Доступность

HP StorageWorks Scalable File Share

Масштабируемая производительная система хранения для кластеров Linux

Большая пропускная способность: решение проблемы I/O Bottleneck для Linux кластеров

Кластер серверов данных, формирует единый виртуальный масштабируемый файл-сервер

- **Масштабируемая производительность**
  - 200 MB/s to 35 GB/s (more by request)
- **Масштабируемая емкость**
  - 2 TB to 512 TB (more by request)
- **Масштабируемая цена/производительность**
  - 2X to 5X bandwidth/$ price/performance advantage
- **Масштабируемые подключения**
  - От десятков до тысяч клиентов
- **Масштабируемая надежность**
  - Различные схемы отказоустойчивости
- **Масштабируемая простота**
  - Интегрированная система
  - Разработана, интегрирована и поддерживается HP

HP StorageWorks 9100 Extreme Data Storage System

Интегированное аппаратно-программное решение

Новинка!

1. Блейд-системы
   - Стандартная полка c7000
   - До 16 серверов в кластере с балансировкой нагрузки
2. Системы хранения
   - До 12TB/U
   - От 246TB, расширение до 820TB
   - Масштабируется блоками по 82TB
3. Программное обеспечение
   - Интегрированные средства управления
   - Кластерная симметричная файловая система
   - Поддержка различных протоколов
     - NFS, HTTP, DirectIO


invent