



T-BLADE 2 ТЕХНИЧЕСКОЕ РУКОВОДСТВО

Заявление об отсутствии гарантий:

Данный документ предназначен только для информационных целей и может содержать неточности. Чтобы получить самую последнюю информацию, свяжитесь с представителем компании "Т-Платформы".

Оглавление

1. Общий обзор	4
2. Вычислительный узел	5
2.1. Плата блейд-модуля.....	5
2.1.1. Геометрия и размеры платы.....	5
2.1.2. Снабор средств вычислительного узла	6
2.1.3. Специальные средства вычислительного узла.....	7
2.2. Питание и средства подключения блейд-модуля.....	7
2.3. Радиатор блейд-модуля	8
3. Модули коммутации InfiniBand	9
4. Управляющий модуль (Management and Switch Module, MSM).....	10
4.1. Блок процессора управления	10
4.2. Блок коммутаторов Gigabit Ethernet	11
4.3. Блок специальных сетей (FPGA)	11
4.4. Блок глобального распределения синхросигналов	11
4.5. Порты управляющего модуля	12
5. Объединительная плата	13
6. Шасси	14
6.3. Подсистема охлаждения	15
7. Другие средства.....	16
7.1. Процедура включения питания	16
7.2. Процедура экстренного завершения работы системы	16
8. Управление кластером и мониторинг	16
9. Базовые требования к инфраструктуре	17
9.1. Электропитание.....	17
9.2. Охлаждение	17
9.3. Инфраструктура шкафа.....	17
9.4. Полы и размещение оборудования.....	17
10. Совместимость с операционной и файловой системами.....	17
11. Спецификация системы	18
Шасси T-Blade 2.....	18
Вычислительный узел T-Blade 2	18
Внешние порты и сети T-Blade 2.....	18
Приложение 1. Диаграмма блейд-модуля	19
Приложение 2. Диаграмма коммуникаций QDR IB внутри шасси.....	20
Приложение 3. Диаграмма модуля коммутации и управления	21

1. Общий обзор

Познакомьтесь с T-Blade 2 - решением с лидирующей вычислительной плотностью, разработанным компанией "Т-Платформы" специально для крупнейших в мире суперкомпьютерных инсталляций на платформе x86.

Представленная в 2009 г. система T-Blade 2 позволяет создавать не имеющие аналогов решения для заказчиков с самыми высокими требованиями и превосходно дополняет другие элементы портфеля вычислительных систем от компании "Т-Платформы", включая вычислительные узлы Cell BE 1RU на базе PowerXCell 8 и систему T-Blade 1 в конструктиве 5RU из 10 узлов.

Основу конструкции T-Blade 2 составляет использование стандартных вычислительных компонентов, интерконнекта и подсистемы управления, объединенных в единое решение и оптимизированных для высокопроизводительных вычислительных задач, чтобы обеспечить правильное сочетание производительности, плотности, избыточности и управляемости.

T-Blade 2 - платформа, разработанная для достижения высокой плотности, производительности и надежности:

- Шасси высотой 7U для установки в стандартную 19" стойку.
- 16 вычислительных модулей, подключаемых в "горячем" режиме, с 64 процессорами серии Intel Xeon 56xx; каждый вычислительный модуль содержит на одной плате 2 двухпроцессорных вычислительных узла.
- 2 интегрированных 36-портовых коммутатора QDR InfiniBand.
- Интегрированный управляющий модуль с коммутаторами GbE и специальными средствами НРС.
- Воздушное охлаждение, использующее заменяемые в "горячем" режиме высокопроизводительные вентиляторы.
- Блоки питания мощностью 11 кВт, резервируемые по схеме N+1.



Хотя в T-Blade 2 применяются стандартные технологии, компания "Т-Платформы" использует свою техническую экспертизу и опыт разработки архитектуры систем для преодоления "узких мест" в масштабируемости, столь характерных для кластеров из массовых серверов стандартной архитектуры. В основном причиной этих "узких мест" является низкая эффективность коллективных коммуникаций, для которых необходима специальная поддержка на аппаратном уровне: лишь с ее помощью удается поддерживать превосходную производительность приложений в системах с большим числом узлов.

Платформа T-Blade 2 оснащается специализированными глобальными сетями барьерной синхронизации и прерываний, обеспечивающих эффективную масштабируемость приложений даже в системах, содержащих несколько тысяч узлов. Глобальная сеть барьерной синхронизации поддерживает быструю синхронизацию между задачами в крупномасштабных приложениях, а глобальная сеть прерываний значительно снижает негативное влияние джиттера ОС, синхронизируя процесс планирования в масштабе всей системы. В результате процессоры эффективнее взаимодействуют друг с другом, обеспечивая высокую степень масштабируемости даже для самых требовательных параллельных приложений.

Этими двумя сетями управляет встроенная микросхема FPGA в составе интегрированного в шасси управляющего модуля (см. Приложение). Модуль осуществляет мониторинг всех подсистем и компонентов, обеспечивая удаленное управление системой, экстренное завершение работы и т.д.

Для быстрой и эффективной передачи данных и снижения нагрузки на сеть в системах с большим числом узлов в системе T-Blade 2 интегрированы коммутаторы InfiniBand, с общей пропускной способностью в 1,6 Тбит/с.

Компания "Т-Платформы" разработала T-Blade 2 как надежное и высокопроизводительное решение. В корпусе системы нет ни жестких дисков, ни кабелей, что значительно снижает вероятность отказов из-за механических неисправностей внутри вычислительного узла. Надежность еще более увеличивается за счет наличия в каждом корпусе заменяемых в "горячем" режиме блоков питания и вентиляторов охлаждения, резервируемых по схеме N+1.

Для эффективной поддержки масштабируемости приложений до уровня петафлопсных вычислений суперкомпьютерные вычисления требуют глубоких изменений в системном программном обеспечении. В качестве опции с T-Blade 2 поставляется полный стек программного обеспечения по типу "все-в-одном", включающий в себя оптимизированное ядро ОС Linux и системные библиотеки, а также все необходимые программные компоненты для управления и мониторинга. Этот интегрированный стек программного обеспечения превращает систему в готовое к использованию решение (out-of-the-box), сокращает время ее инсталляции и затраты на администрирование.

Поставляемое опционально ядро Linux предусматривает специализированную поддержку глобальной сети барьерной синхронизации и глобальной сети прерываний, оптимизацию, значительно увеличивающую скорость межузловых коммуникаций по сравнению с более традиционной архитектурой. Системное управляющее ПО (часть пакета Clustrx OS) обеспечивает эффективную масштабируемость мониторинга до 12000 вычислительных узлов на один сервер управления с возможностями мониторинга в масштабе времени, близком к реальному. Кроме того, платформа T-Blade 2 оснащается прогрессивной технологией снижения энергопотребления и поддерживает выделение ресурсов на основе топологии, что улучшает эффективность использования памяти и увеличивает производительность реальных приложений.

Вычислительная платформа НРС T-Blade 2 прошла эксплуатационную проверку: это основной "строительный элемент" кластера "Ломоносов", с пиковой производительностью в 420 ТФлопс, развернутого в Московском Государственном Университете. Эта система находится на 13-ом месте в опубликованном в июне 2010 года списке TOP500 самых быстрых компьютеров мира и считается самым крупным суперкомпьютером в Восточной Европе.

2. Вычислительный узел

Вычислительный узел T-Blade 2 состоит из одной печатной платы (PCB) с двумя отдельными двухпроцессорными узлами на ней, смонтированными на специально спроектированном радиаторе для обеспечивающем необходимое охлаждение всего модуля. Каждая плата блейд-модуля поставляется с предустановленными модулями памяти (Рисунок 1) в упаковке, включающей 16-подобных узлов.

2.1. Плата блейд-модуля

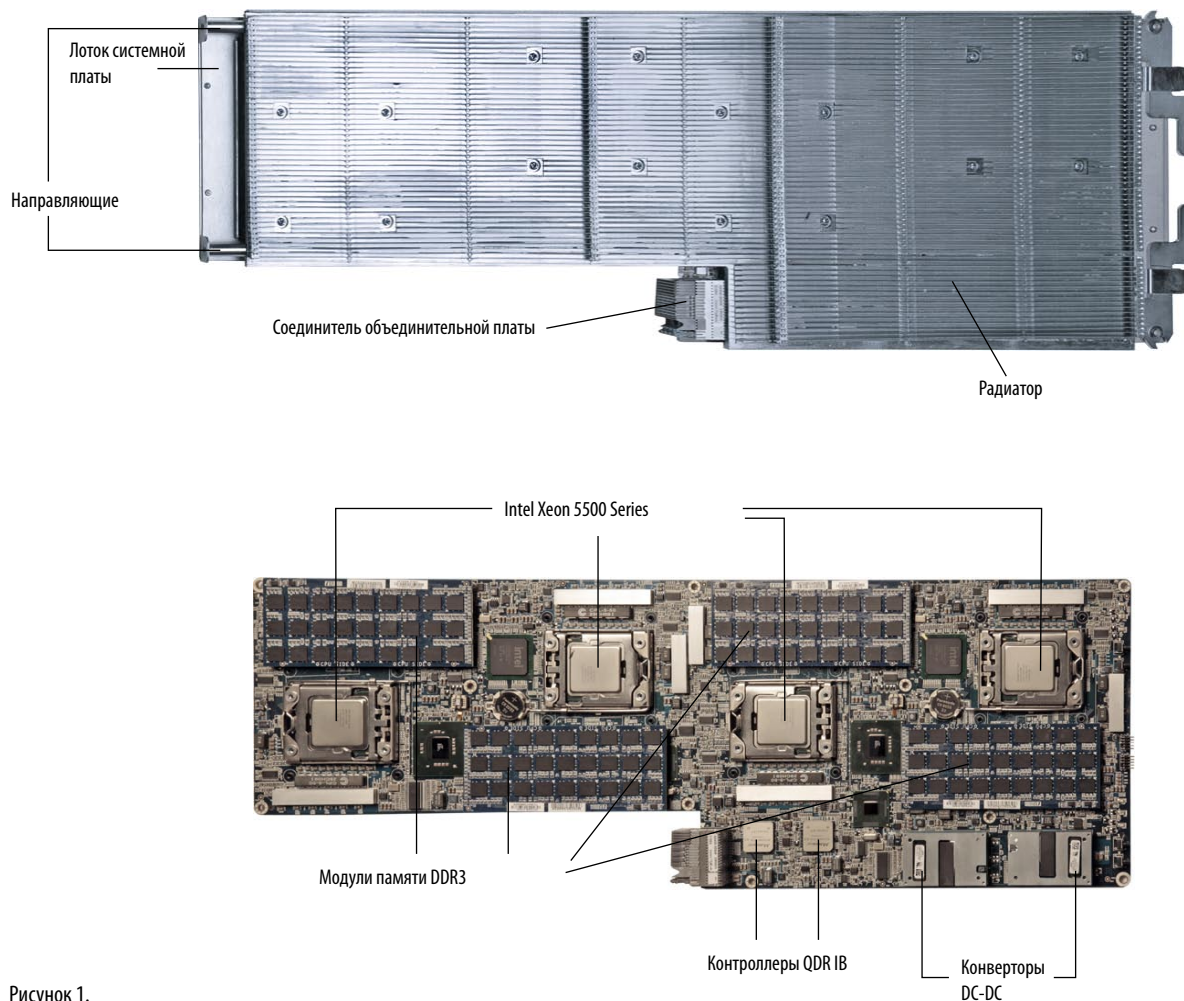


Рисунок 1.

2.1.1. Геометрия и размеры платы

Геометрия блейд-модуля PCB представлена на Рисунке 1. Эта плата, оптимизированная для воздушного охлаждения, для надежности соединяется с радиатором и с лотком платы с помощью монтажных отверстий. Каждая плата подключается к объединительной платы (backplane) с помощью разъема питания и сигнального разъема, смонтированного на PCB, поддерживается механическими направляющими, которыми снабжается каждый лоток.

2.1.2. Сбор средств вычислительного узла

Каждый из двух вычислительных узлов на печатной плате оснащается:

- Двумя процессорами Intel Xeon серии 56xx (Westmere) с тепловым конвертом (TDP) 95 Вт каждый и тактовой частотой до 2,93 ГГц.
- 4 модулями памяти (Рисунок 2) в специальном исполнении мезанинной платы с 27 микросхемами памяти (9 микросхем на каждый из трех каналов), обеспечивающими 6 или 12 гигабайтов памяти на каждый сокет ЦП. Используются трехканальная небуферизованная память ECC DDR3-1333 RAM.
- Набором микросхем Intel Tylersburg 24D+ ICH10 с соединениями QPI, обеспечивающими скорость коммуникаций между ЦП 6,4 тыс. передач/с (3,2 ГГц).
- Однопортовой микросхемой Mellanox ConnectX QDR InfiniBand, соединенной с набором микросхем Tylersburg через канал PCIe 2.0 x8.
- Однопортовым контроллером GbE, подключенным к ICH10.
- Контроллером USB для флэш-накопителей miniSD или microSD и коннектором для флэш-карты.
- Управляющим контроллером (Baseboard management controller, BMC) и отдельным однопортовым контроллером GbE для подключения BMC (см. раздел “Управление вычислительным узлом и мониторинг”, где обсуждается BMC, а также набор средств управления и мониторинга).
- Одним последовательным портом, соединенным с BMC.
- Для реализации специальных кластерных функций (описанных в разделе “Вычислительный узел и специальные средства”) пять сигнальных линий GPIO на управляющем модуле проходят от микросхемы ICH10 через объединительную плату к микросхеме FPGA.
- Централизованной архитектурой хранения без интегрированного на печатной плате контроллера SAS/SATA; каналы интегрированного контроллера SATA не задействованы, а соответствующая часть BIOS отключена.
- Двумя микрокнопками “Питание” и “Сброс” на заднем крае блейд-модуля.

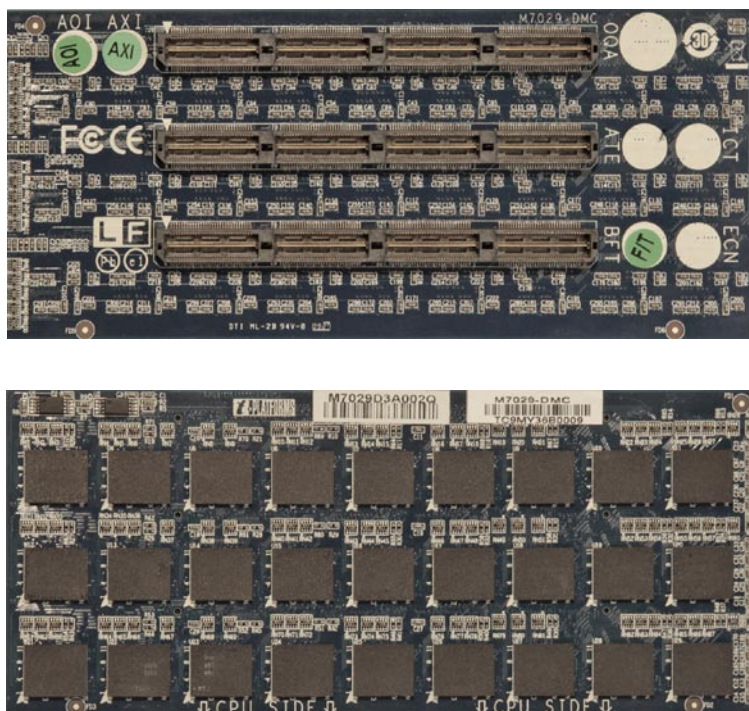


Рисунок 2. Модуль памяти DDR3.

2.1.3. Специальные средства вычислительного узла

- Для синхронизации в системе можно использовать внутренние или внешние тактовые генераторы. Их выбор осуществляется с помощью специальной микросхемы IDT5T9GL02, и вывод выбора режима синхронизации этой микросхемы соединяется с переключателем DIP на задней нижней части печатной платы, что позволяет выбирать режим синхронизации узлов вручную. Кроме того, светодиодный индикатор (LED) на задней части каждого блейд-модуля показывает, какой тактовый генератор часы использует система - внутренний или внешний.
- Два вывода микросхемы ICH10 через объединительную панель соединяются с микросхемой FPGA модуля управления.
- 5 выводов GPIO микросхемы ICH10 каждого вычислительного узла через объединительную плату соединяются с микросхемой FPGA модуля управления.

Управление и мониторинг каждого вычислительного узла реализованы на основе стандартного контроллера управления материнской платы (base-board management controller, BMC). BMC доступен через модуль управления по выделенному интерфейсу GbE, исключающим возможное влияние на производительность приложений. BMC предлагает следующие средства мониторинга оборудования:

- Контроль температуры каждого ядра ЦП.
- Температура памяти - один датчик температуры на блок памяти, выделенный каждому ЦП.
- Два дополнительных температурных датчика на модуле регулятора напряжения ЦП.
- Напряжение на ядре каждого ЦП: 3,3, 5, 12 и 5В дежурного напряжения.
- Состояние ошибок памяти и PCI; в случае ошибок памяти для помощи в устранении ошибки указывается канал памяти.
- Счетчики InfiniBand и GbE (опционально).

Поддерживаются следующие возможности управления:

- Удаленное включение/отключение/сброс питания.
- Удаленный KVM-over-IP и полный доступ к консоли SOL, начиная с процесса загрузки и включая процедуру POST и настройки BIOS.
- Обновление BIOS вычислительного узла и сохранение/восстановление NVRAM без использования зависящих от ОС утилит.
- Выбор приоритета загрузки.
- Удаленное управление с помощью микропрограммного обеспечения BMC.
- Все средства BMC доступны удаленно через Telnet и SSH с использованием ключа аутентификации для дополнительной защиты.

2.2. Питание и средства подключения блейд-модуля

Все питание и соединения блейд-модуля осуществляются через объединительную плату и сигнальные разъемы. На входе разъема питания - -48 В постоянного тока (DC) и ток минимум 18 А на каждый канал -48 В.

Через сигнальный разъем на объединительную плату разводятся следующие сигналы:

- 2 канала QDR InfiniBand 4x (1 на каждый вычислительный узел).
- 4 канала GbE (1 GbE + 1 BMC GbE на каждый вычислительный узел).
- 2 канала внешней синхронизации (1 на блейд-модуль); с помощью расширителя синхронизации можно использовать на каждом вычислительном модуле один внешний источник синхронизации (тактовый генератор).
- 4 удаленных канала прерывания (2 на блейд-модуль).
- 10 каналов глобальной сети барьерной синхронизации (5 каналов GPIO на блейд-модуль).

2.3. Радиатор блейд-модуля

Плата блейд-модуля соединяется с лотком и специально спроектированным радиатором для отвода тепла, закрывающим всю плату. Этот радиатор используется также в качестве направляющей для платы блейд-модуля, что помогает надежно вставлять ее в отсеки корпуса. Геометрия радиатора представлена на Рисунке 3.

Базовая пластина радиатора имеет переменную толщину в соответствии с высотой компонентов платы блейд-модуля. Мастика с низким термическим сопротивлением обеспечивает хороший тепловой контакт между компонентами и радиатором.

В основе конструкции радиатора и выбора материалов - обширный термический анализ, выполненный компанией "Т-Сервисы" для достижения оптимального дизайна вычислительного узла. Эти инженерные результаты стали основой окончательной конструкции, которая не только энергоэффективнее, но и легче первоначальных вариантов, что снижает общий вес системы T-Blade 2 до 153 кг.

Предусматривается также механизм фиксации, смонтированный на задней кромке радиатора для установки и извлечения блейд-модуля (Рисунок 3).

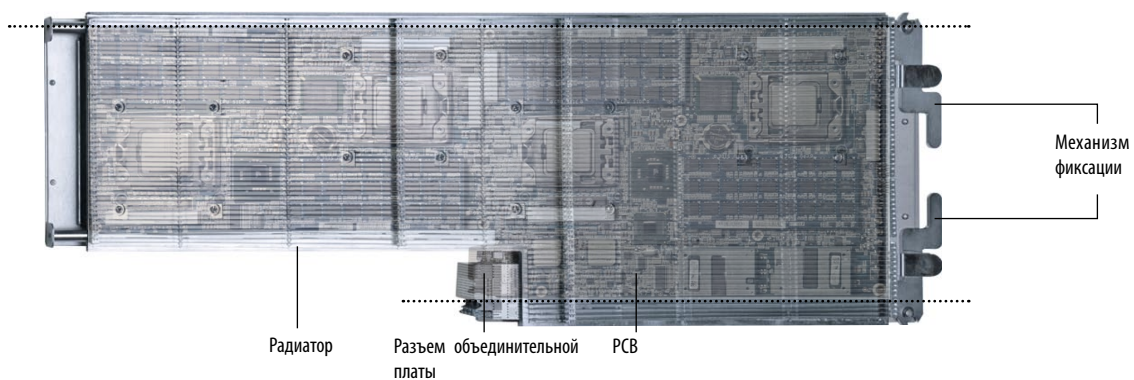


Рисунок 3.

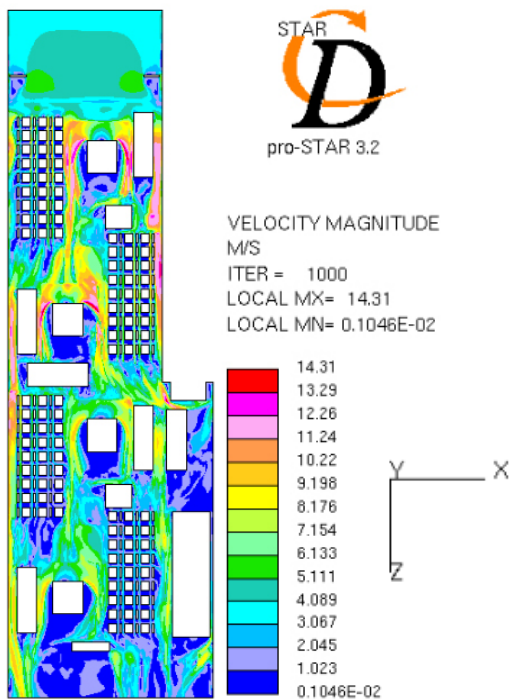


Рисунок 4.

Скорость и распределение воздушных потоков между платой и радиатором для варианта конструкции с вентилятором 60CFM.

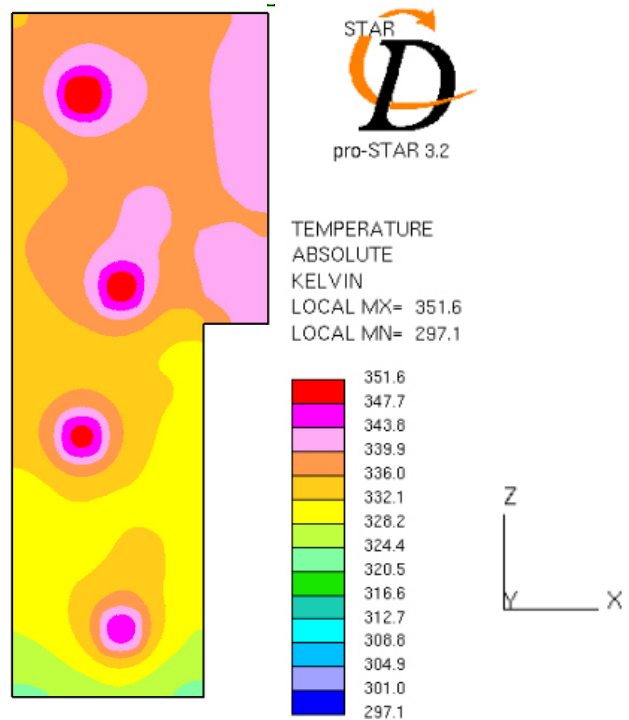


Figure 5.

Распределение температуры в основании радиатора в варианте на основе алюминия.

3. Модули коммутации InfiniBand

Система T-Blade 2 содержит два резервируемых модуля коммутации InfiniBand, каждый из которых расположен в задней части модульного шасси. В модулях коммутаторов InfiniBand применяется логика Mellanox InfiniScale IV QDR InfiniBand. Всего предусмотрено 36 портов QDR InfiniBand 4x: 16 портов QDR разводятся через объединительную плату на вычислительные модули, а 20 портов QDR выводятся на заднюю панель, что обеспечивает дополнительную гибкость при подключении централизованной системы хранения или разнородных узлов (см. Приложение 2). Каждый порт блейд-модуля подключается последовательно, при этом порты 1-16 подключены к одному коммутатору, а порты 17-36 - к другому. В топологии внутри шасси используется два независимых коммутатора, а в топологии между шасси допускается любой вариант, поддерживаемый InfiniBand Subnet Manager.



Рисунок 6. Модули QDR InfiniBand

- Порты внешней панели реализованы на базе разъемов QSFP.
- Удаленное управление и мониторинг модуля коммутации InfiniBand реализованы с помощью канала I²C, который через объединительную плату подключается к модулю управления.
- Модуль коммутации InfiniBand питается через разъем питания объединительной платы напряжением -48В постоянного тока.

Задняя панель модуля коммутации IB содержит:

- 20 разъемов QSFP портов InfiniBand.
- Светодиодные индикаторы инициализации канала (Link initialized) и активности канала (Link active) для каждого порта.
- Два монтажных крепления (механизм фиксации, Рисунок 1).
- Вентиляционные отверстия.

4. Управляющий модуль (Management and Switch Module, MSM)

MSM предусматривает следующие функции:

- Управление и мониторинг вычислительных модулей и шасси.
- Коммутация Gigabit Ethernet.
- Поддержка глобальной сети барьерной синхронизации и глобальной сети прерываний
- Глобальное распределение синхронизирующих импульсов.

MSM представляет собой устройство высотой 1U, расположенное в задней части корпуса. Он состоит из четырех функциональных блоков:

- Блок процессора управления (4.1)
- Блок коммутаторов Gigabit Ethernet (4.2)
- Блок специальных сетей (4.3)
- Глобальный блок распределение синхросигналов (4.4)



Рисунок 7. Управляющий модуль

Для крупномасштабных инсталляций модуль управления соединяется со специализированным внешним коммутатором распределение внешних синхросигналов для уменьшения джиттера ОС, повышения предсказуемости и производительности приложений. Этот модуль управления постоянно собирает информацию о программных и аппаратных событиях, консолидирует их для передачи узлу управления кластером и специальному защищенному устройству - "черному ящику". В сочетании с сервисами управления Clustrx данный двухэтапный процесс обеспечивает возможность мониторинга 12 тыс. событий в секунду в близком к реальному режиму времени с использованием лишь одного двухпроцессорного модуля управления (подробнее об этом рассказывается в Разделе 8).

4.1. Блок процессора управления

Блок процессора управления содержит следующие компоненты:

- Низковольтный ЦП класса Intel Yonah.
- Intel 3100 MICH.
- Один слот памяти DDR2 с установленным модулем ECC 2 Гбайта.
- Двухпортовый контроллер GbE на шине PCI-E (Intel 82576).
- BMC (AST2050).
- Микросхема SIO.
- Диск SSD емкостью 40 Гбайт.

Предусмотрены следующие интерфейсы:

- 1 интерфейс GbE с Intel 82576 на коммутатор вспомогательной сети.
- 1 интерфейс GbE с Intel 82576 на коммутатор управления сетью.
- 1 интерфейс 10/100 Ethernet от BMC, использующий один из двух портов RJ45 на задней панели. Это позволяет обращаться к BMC и процессору модуля управления удаленно в случае проблем с коммутаторами GbE.
- 1 интерфейс PCI-E к блоку специальных сетей (FPGA).
- 1 видеоинтерфейс от BMC к порту D-SUB на задней панели.
- 2 интерфейса USB к портам на задней панели.
- 1 интерфейс USB к модулю ЖК-панели (через разъем объединительной платы).
- 1 интерфейс RS-232 к порту RJ45 на задней панели.
- 1 интерфейс RS-232 к процессору управления коммутаторами GbE (Marvel 88F5181).
- Интерфейс шины I²C, который идет к объединительной плате (для вентиляторов, PSU и управления коммутатором InfiniBand). Шина I²C совместно используется BMC и ЦП модуля управления, позволяет BMC управлять включением и выключением отдельных блоков питания.

Подробности см. в Приложении 3.

4.2. Блок коммутаторов Gigabit Ethernet

Блок коммутаторов Gigabit Ethernet состоит из двух отдельных коммутаторов:

- Коммутатора вспомогательной сети, используемого для общего доступа к узлу (SSH, управление заданиями, и т.д.).
- Коммутатора управляющей сети, используемого для доступа к BMC.

Каждый из двух коммутаторов содержит две состыкованные микросхемы Marvell DX270. Оба коммутатора управляются процессором Marvell 88F5181, ОС Linux и специальной микропрограммой.

Коммутаторы вспомогательной сети имеют следующие соединения:

- 32 канала GbE, которые идут через объединительную плату к вычислительным узлам.
- 1 канал GbE к сетевому контроллеру управляющего ЦП (Intel 82576).
- 1 канал GbE, который идет к Marvell 88F5181.
- 2 восходящих канала 10G (uplink) к портам XFP на задней панели модуля управления.

Коммутатор управляющей сети имеет следующие соединения:

- 32 канала GbE, которые идут через объединительную плату к BMC вычислительных узлов.
- 1 канал GbE к сетевому контроллеру управляющего ЦП (Intel 82576).
- 1 канал GbE к Marvell 88F5181.
- 2 восходящих канала GbE к портам RJ45 на задней панели управляющего модуля.

Один из двух портов RJ45 на задней панели также может использоваться для удаленного доступа к BMC модуля управления. Процессор Marvell 88F5181 предусматривает два интерфейса RS-232 к ЦП модуля управления.

Более подробную информацию можно найти в Приложении 3.

4.3. Блок специальных сетей (FPGA)

Блок специальных сетей используется для реализации глобальной сети барьерной синхронизации и сети глобальных прерываний. В нем используется микросхема FPGA Xilinx XC5VLX50T FPGA (XC5VLX50T-3FFG665C), а также флэш-накопитель для загрузки FPGA.

Предусмотрены следующие интерфейсы:

- Канал PCI-E от FPGA к микросхеме Intel 3100 MICH.
- 32x5 (всего 160) одножильных каналов от FPGA к выводам GPIO микросхем ICH10 вычислительных узлов.
- 2 одножильных канала от FPGA к выводам IRQ микросхем ICH10 вычислительных узлов (они разделяются на 32x2 канала на объединительной плате).
- 5 дифференциальных каналов от FPGA к портам RJ45 на задней панели для глобальной сети барьерной синхронизации.
- 2 дифференциальных канала от FPGA к портам RJ45 на задней панели для глобальной сети прерываний.
- 4 дифференциальных каналов от MGT на FPGA к порту RJ45 на задней панели (зарезервировано для будущего использования).

Более подробную информацию можно найти в Приложении 3.

4.4. Блок глобального распределения синхросигналов

Глобальное распределение синхросигналов используется для настройки частоты ЦП и памяти всех вычислительных узлов. Распределение синхросигналов может осуществляться через тактовый генератор модуля управления или внешний тактовый генератор. Для лучшего качества сигнала при внешней синхронизации следует использовать передачу дифференциальными сигналами; ввод дифференциальных синхросигналов преобразуется на модуле управления в одножильный. Выбор внешнего или встроенного источника тактовой генерации осуществляется микросхемой выбора синхронизации: режим выбирается переключателем на задней панели управляющего модуля. Выбранный синхросигнал поступает на объединительную плату, а затем разделяется по всем вычислительным узлам.

Более подробную информацию можно найти в Приложении 3.

4.5. Порты управляющего модуля

На модуле управления присутствуют следующие порты:

- Сигнальный порт объединительной платы предусматривает следующие сигналы:
 - 32 канала GbE от вычислительных узлов к коммутатору вспомогательной сети.
 - 32 канала GbE от BMC вычислительных узлов к коммутатору сети управления.
 - Канал I²C к микросхеме Intel 3100 MICH у BMC узла управления.
 - 32x5 каналов GPIO от FPGA.
 - 2 канала прерывания от FPGA.
 - Сингал Rlock на вычислительные узлы.
 - Канал USB к модулю ЖК-панели.
 - Разъем питания объединительной платы предусматривает основное питание 48В и вход питания 5В дежурного напряжения для BMC.

Порты на задней панели:

- 2 порта XFP 10GbE для восходящих соединений (uplink) коммутаторов вспомогательной сети (ETH1 и ETH2).
- 2 порта RJ45 для восходящих соединений коммутаторов управляющей сети (ETH3 и ETH4).
 - Один порт является портом GbE.
 - Другой - это порт 100 Мбит/с, разделяемых между несколькими устройствами.
- 1 порт RJ45 100 Мбит/с для глобальной сети барьерной синхронизации - каналы 1-4 (блок FPGA) (SPN1).
- 1 порт RJ45 100 Мбит/с для глобальной сети барьерной синхронизации - канал 5, глобальная сеть прерываний - каналы 1 и 2, ввод внешней синхронизации (блок FPGA) (SPN2)
- 1 порт RJ45 100 Мбит/с для 4 дифференциальных сигналов, зарезервирован для будущего использования (блок FPGA) (RIO).
- 1 порт D-SUB для видео (VGA).
- 2 порта USB.
- 1 порт RJ45 для порта RS-22, последовательная консоль (Serial).

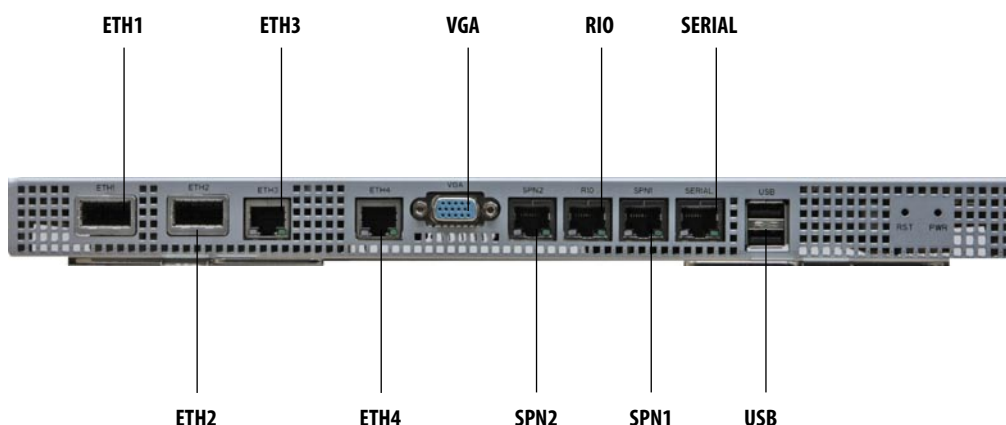


Рисунок 8. Порты задней панели управляющего модуля

5. Объединительная плата

Объединительная плата (backplane) используется для подачи питания на вычислительные модули, модули коммутации InfiniBand, управляющего модуля (MSM), а также для обеспечения коммуникаций между этими компонентами. Кроме того, для вычислительных модулей поддерживается функция "горячего" подключения. Объединительная плата доставляет электропитание мощностью до 13 кВт и реализует стандартную систему передачи сигналов с использованием высокочастотных каналов InfiniBand QDR (32 канала @ 40 ГТ/с - миллиардов пересылок в сек).

Объединительная плата содержит лишь небольшое число активных компонентов; они отвечают только за логику глобального распределения синхросигналов. Она представляет собой низкопрофильную плату, оптимизированную для эффективного пропускания воздуха и основанную на 24-слойной печатной плате. На задней части объединительной платы находятся следующие разъемы:

- 16 комбинированных разъемов питания/передачи сигналов к вычислительным модулям.
- 2 комбинированных разъемов питания/передачи сигналов к модулям коммутации InfiniBand.
- 2 разъемов питания/передачи сигналов к MSM.

Передняя часть объединительной платы служит для ввода питания. На передней стороне размещаются следующие компоненты:

- I²C разъемов к PSU.
- I²C разъемов к модулям вентиляторов.
- Порт USB для ЖК-панели.

Общая геометрия объединительной платы представлена на Рисунке 9.

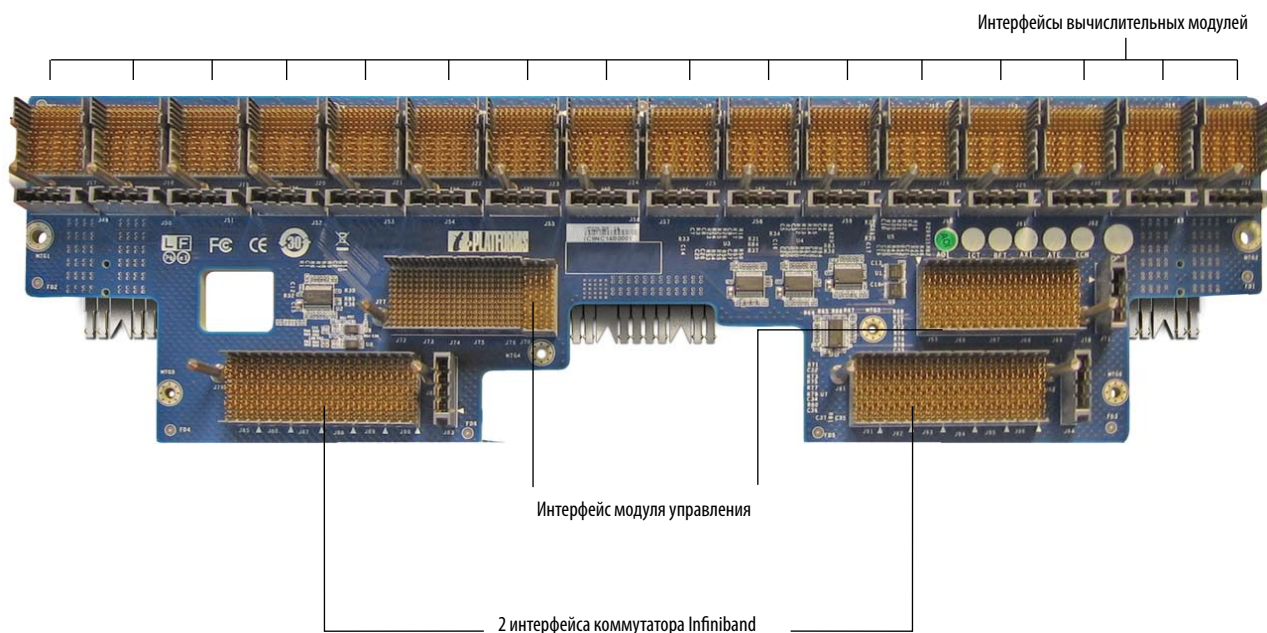


Рисунок 9. Объединительная плата системы, вид сзади.

6. Шасси

В шасси T-Blade 2 устанавливаются следующие заменяемые в эксплуатационных условиях компоненты (field-replaceable unit, FRU):

- 16 вычислительных модулей “горячего” подключения.
- 2 модуля коммутации InfiniBand.
- 1 управляющий.
- 10+2 заменяемых в “горячем” режиме 80-мм модулей вентиляторов.
- 6 заменяемых в “горячем” режиме блоков питания.

Кроме того, имеется 4-канальная ЖК-панель с кнопочным управлением, подключаемая через объединительную плату к порту USB на управляющем модуле. Основной выключатель питания на передней панели корпуса отсутствует: узлы запитываются удаленно или с помощью микрокнопок включения/выключения/сброса отдельных вычислительных модулей.

Размеры шасси:

- Высота: 310 мм (7U)
- Ширина (без направляющих и монтажных скоб): 430 мм, помещается в стандартную 19” стойку.
- Глубина: 860 мм

Шасси можно монтировать в стандартную 19” стойку с помощью фиксированных направляющих (без возможности выдвигания), и на передней стороне шасси имеются монтажные скобы, которые можно прикручивать к вертикальным направляющим стойки. Компания “Т-Платформы” рекомендует заказчикам устанавливаться в стойку пустое шасси, и уже затем наполнять его модулями. Полностью укомплектованное шасси T-Blade 2 весит примерно 153 кг, поэтому обращаться с ним нужно осторожно.

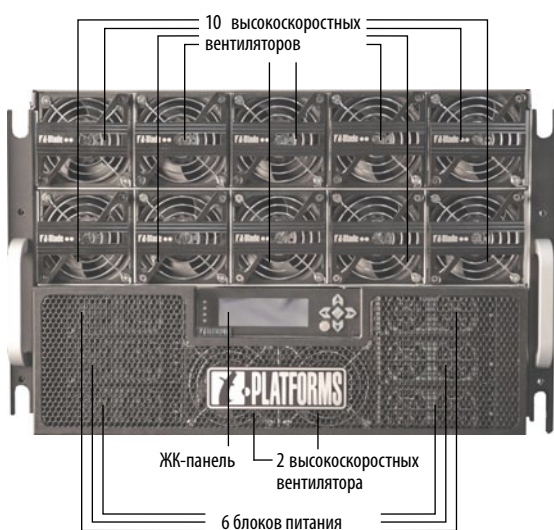


Рисунок 10. Вид передней части.



Рисунок 11. Вид задней части.



Рисунок 12. Извлечение блейд-модуля.



Рисунок 13. Извлечение модуля коммутации IB.

6.2. Подсистема питания

Подсистема питания T-Blade 2 состоит из шести блоков питания. Они соединены через шесть плат сопряжения с T-образной платой распределения питания (Power Distribution Board), которая подключает заменяемые в “горячем” режиме блоки питания (и вентиляторы системы) к объединительной плате системы. Блоки питания и вентиляторы - единственные компоненты сторонних производителей в системе T-Blade 2; подробнее о сборке системы T-Blade 2 рассказывается в документе “Схема и компоненты системы T-Blade 2”.

Модуль питания 2725W Lineage (модель CP2725AC54Z) имеет компактную конструкцию 1RU (Рисунок 14). Являясь составной частью системы распределения питания T-Blade 2, он оснащается воздушным фильтром (воздух забирается спереди), интерфейсом RS485, резервируемыми каналами I²C, PFC, защитой от падения и повышения напряжения в сети и термической защитой.

Более подробную информацию о PSU можно найти по ссылке:

<http://www.lineagepower.com/BinaryGet.aspx?ID=f75abb6f-abc4835-b811-f3b6dabd98b5>



Рисунок 14. Системные блоки питания.

6.3. Подсистема охлаждения

T-Blade 2 использует воздушное охлаждение с продувкой спереди назад, а корпус T-Blade 2 разделен на верхнюю и нижнюю зоны воздушных потоков. Проектирование охлаждения T-Blade 2 было связано с обширными техническими разработками, термическим моделированием и физическим тестированием, чтобы охлаждение системы было достаточным даже при устойчивых пиковых нагрузках.

Охлаждение блейд-модуля обеспечивается десятью заменяемыми в “горячем” режиме 80-мм вентиляторами (Рисунок 15). Вентиляторы питаются постоянным током, позволяют контролировать скорость вращения и управлять ею через систему пульсово-волновой модуляции (PWM). Каждый вентилятор помещен в отдельный заменяемый в “горячем” режиме модуль, и специальные заслонки в шасси предотвращают обратный воздушный поток в случае отказа или удаления вентилятора. Модули вентиляторов оборудованы небольшой платой с контроллером вентилятора, соединенным через интерфейс I²C с управляющим модулем.

Под ЖК-панелью находятся два дополнительных 80-мм вентилятора, обеспечивающих охлаждение модулей коммутаторов InfiniBand и управляющего модуля. Каждая из двух групп вентиляторов резервируется по схеме N+1 и использует унифицированный дизайн высокопроизводительных вентиляторов со скоростью вращения 14000 об/мин. Для увеличения срока службы они располагаются впереди корпуса и работают при температуре, близкой к температуре окружающей среды.



Рисунок 15. Вентиляторы системы.

7. Другие средства

7.1. Процедура включения питания

После подключения T-Blade 2 к источнику питания блоки питания переходят в состояние ожидания (режим Stand By), подавая на управляющий модуль только 5В дежурного напряжения (Vsb). Затем блок питания включается с помощью удаленного соединения с BMC или кнопкой питания. В этот момент модуль управления, модули коммутации InfiniBand и BMC на вычислительных модулях начинают процедуры инициализации. После завершения инициализации вычислительные модули включаются с помощью удаленного соединения со своих BMC.

7.2. Процедура экстренного завершения работы системы

Система T-Blade 2 имеет функцию экстренного завершения работы (Emergency Shutdown). Она реализована в управляющем модуле (MSM). В зависимости от версии микропрограммного обеспечения в случае критического события управления можно завершать работу и выключать отдельный вычислительный узел или весь корпус. Процедура Emergency Shutdown помогает избежать физического повреждения электроники системы.

Не забывайте, что каждый вычислительный узел можно также включить, выключить или перезапустить систему вручную с помощью двух микрокнопок на задней части PCB управляющего модуля.

8. Управление кластером и мониторинг

Усовершенствованные возможности мониторинга и управления T-Blade 2 реализованы на уровне кластера. ПО Clustrx OS TP Edition и ее пакет мониторинга Clustrx Watch тесно интегрированы с системой T-Blade 2 и ее управляющим модулем, а также предусматривают утилиты режима командной строки для сложных скриптов. Clustrx - операционная система для высокопроизводительных вычислительных приложений петафлопного уровня производительности, разработанная компанией T-Massive Computing, входящей в группу компаний "Т-Платформы". Для подсистемы управления ОС Clustrx требуется консольная система и один управляющий узел. Она реализует следующие функции:

- Автоматизирует развертывание кластера с помощью одной консоли и одного установочного диска DVD.
- Доступ через CLI и GUI.
- Оптимизирована таким образом, что установка кластера и его настройка требует заурядных навыков системного администрирования.
- GUI установки сокращает время базовой установки и настройки кластера: начиная от установки системы до готовности к тесту Linpack проходит всего два часа (для систем со стандартной топологией без учета времени установки системы хранения данных).
- Система мониторинга отслеживает вычислительную и инфраструктурную подсистемы кластера (на момент подготовки этого документа - кроме коммутаторов Ethernet и InfiniBand).
- Конфигурацию системы мониторинга можно настроить на доставку уведомлений системным администраторам.
- Подсистема управления автоматически реагирует и точно корректирует операции в случае отклонения от нормальной работы, включая автоматическое отключение питания оборудования.
- Задержка мониторинга сведена к минимуму, что обеспечивает своевременность корректирующих мер. Для каждого вычислительного узла собирается до 150 событий в секунду (в ближайшем будущем это число будет увеличено до 300-500 событий).
- В сочетании с Clustrx CNL (Compute Node Linux) пакет мониторинга и управления Clustrx готов к применению в приложениях экзафлопного уровня производительности, а операционная среда охватывает кластеры уровня L3, включающие до 12000 вычислительных узлов (планируется поддержка кластеров L4, содержащих до 210000 вычислительных узлов).
- На вычислительных узлах поддерживаются различные операционные системы; на текущий момент проверка прошла Clustrx OS (поддерживается или будут поддерживаться RHEL, SUSE, Windows и др. ОС).
- Поддерживаются различные параллельные файловые системы, включая Lustre и Panasas.



С системами T-Blade 2 можно также использовать некоторые другие популярные пакеты управления, если они совместимы с IPMI. Так называемые коннекторы устройств можно разработать по запросу.

9. Базовые требования к инфраструктуре

Для обеспечения правильной и бесперебойной работы системе T-Blade 2 необходима профессиональная инсталляция. Такая инсталляция обычно входит в услуги развертывания системы высокопроизводительных вычислений, предоставляемые специалистами по инсталляции компании "Т-Платформы". Проектирование серверной и поддерживающей инфраструктуры - составная часть установки каждой системы HPC на базе T-Blade 2, и, как правило, компания "Т-Платформы" тесно работает с заказчиками над проектированием и развертыванием инсталляции всей системы HPC "под ключ".

В следующих разделах описываются основные предпосылки для развертывания высокопроизводительного кластера на базе T-Blade 2.

9.1. Электропитание

- 80-415 В AC, 5-проводная трехфазная система распределения электропитания.
- Каждая система T-Blade 2 должна поставляться как минимум с одним ИБП мощностью 11 кВт.
- Каждая система T-Blade 2 должна снабжаться выделенным рубильником или автоматическим выключателем на 32 А.

9.2. Охлаждение

- Для охлаждения шасси каждой системы T-Blade 2 требуется 600 кубических футов воздуха в минуту (CFM) с обдувом спереди назад (1019 кубометров в час).
- Система охлаждения площадки должна быть способна работать при окружающей температуре до 55°C, смешивая достаточное количество холодного воздуха для поддержки номинального режима работы систем.
- Требуемая производительность система охлаждения - 11 кВт для каждой системы T-Blade 2.
- Температура входящего воздушного потока должна быть в диапазоне 10°C - 30°C; в соответствии с правилами построения энергоэффективной архитектуры температура входящего воздуха не должна быть менее 20°C, а оптимальная температура - 25°C

9.3. Инфраструктура шкафа

- Шкаф, соответствующий стандарту EIA 310-D (или более поздней версии).
- Глубина шкафа - не менее 900 мм; рекомендуются более новые шкафы с глубиной 1000 мм.
- В зависимости от соединительных кабелей, общий вес корпуса с подсоединенной кабельной инфраструктурой может варьироваться от 130 до 200 кг

9.4. Полы и размещение оборудования

- Рекомендуются антистатические фальш-полы, выдерживающие шкаф с полной нагрузкой.
- Рекомендуемая ширина переднего коридора - не менее 1,0 м.
- Рекомендуемая ширина заднего коридора 0,9 м.
- Итоговая ширина коридора может зависеть также от индивидуальных требований охлаждения.

10. Совместимость с операционной и файловой системами

Являясь высокопроизводительной вычислительной системой, основанной на стандартах, T-Blade 2 поддерживает большое число дистрибутивов Linux, включая RHEL и SUSE. Она позволяет также использовать с основными дистрибутивами Linux пакет Clustrx Watch, реализующий функции детального мониторинга. Для заказчиков, заинтересованных в инсталляции на базе Windows, на системе T-Blade 2 может работать ОС Windows HPC Server 2008.

Примечание: The Clustrx T-Platforms Edition OS - единственный дистрибутив, который в настоящее время поддерживает на платформе T-Blade 2 функциональность глобальной сети барьерной синхронизации и глобальной сети прерываний.

Поддерживаемые в настоящее время файловые системы включают в себя Lustre и Panasas (но не ограничиваются ими).

11. Спецификация системы

Шасси T-Blade 2

Форм-фактор	<ul style="list-style-type: none">16 подключаемых в “горячем” режиме вычислительных модуля (32 двухпроцессорных вычислительных узла в корпусе 7U)Два модуля с 36 портами QDR InfiniBandВыделенный модуль управления
Пиковая производительность на корпус	4.5 ТФлопс
Плотность	384 четырехъядерных процессора (2304 ядер) на стандартную 19” стойку 42U
Пиковая производительность на стойку	27 ТФлопс
ОЗУ	<ul style="list-style-type: none">До 384 Гбайт (1,5 Гбайт на ядро)До 768 Гбайт (3 Гбайт на ядро)
Потребляемая мощность на корпус (максимальная конфигурация)	11 кВт
Производительность/потребляемая мощность	0,4 гигафлопс/Вт
Архитектура охлаждения	12 заменяемых в горячем режиме вентиляторов охлаждения в передней части шасси
Рабочая температура	10-30°C
Размеры (ВхШхГ), мм	310x430x860 мм
Вес системы (в полной конфигурации)	152,6 кг

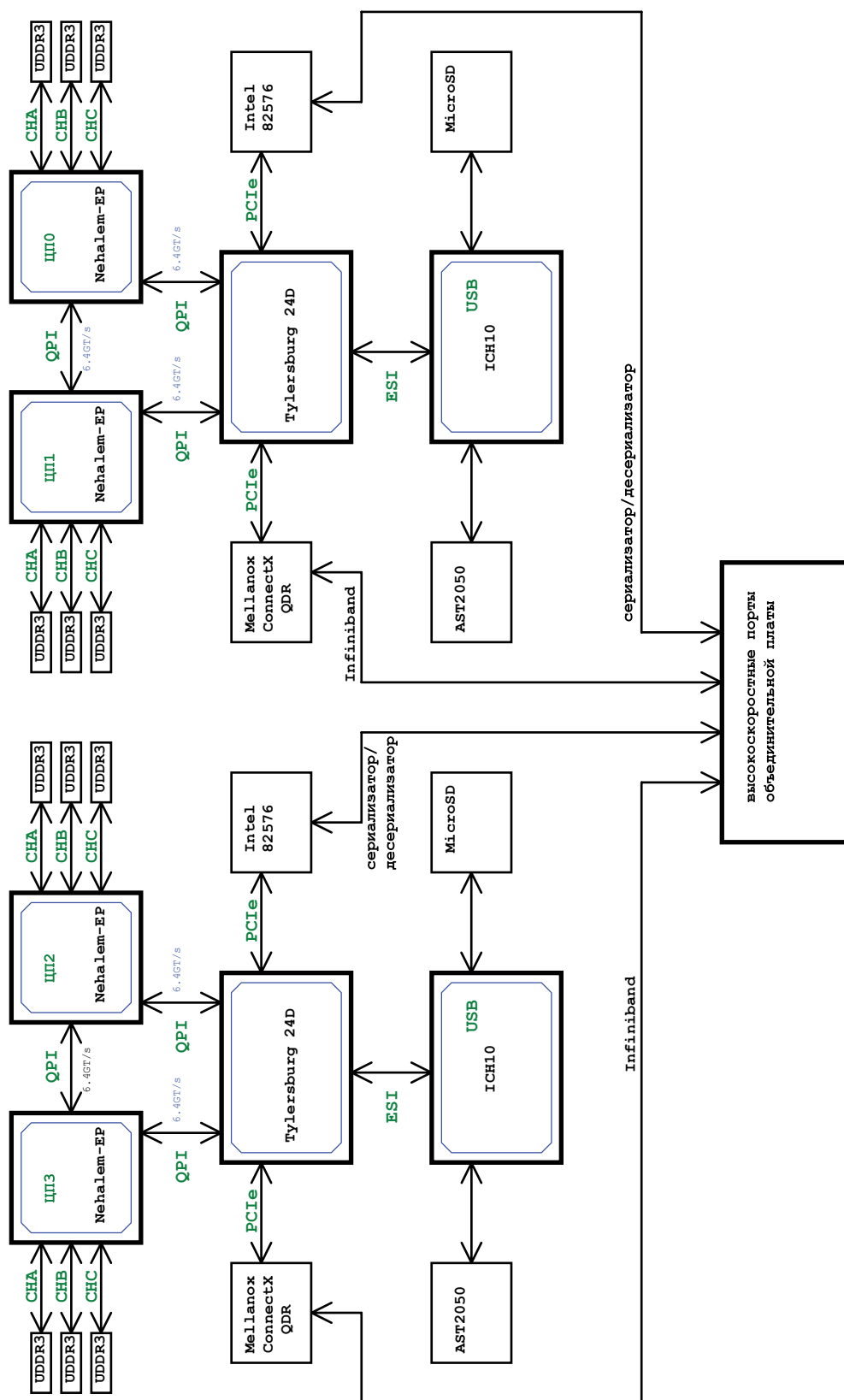
Вычислительный узел T-Blade 2

Мощность/тип процессора	2 шестиядерных процессора Intel Xeon E5600, до 2,93 ГГц
Набор микросхем	Intel 5520+ICH10
ОЗУ	До 24 Гбайт памяти DDR3-1333/1066/800
Встроенная система хранения	Слот MicroSD
Слоты расширения	Нет
Интерфейс Ethernet	1 порт GbE
Поддерживаемый встроенный интерконнект	QDR InfiniBand
Пропускная способность сетевого интерфейса	40 Гбайт/с
Светодиодный индикатор	Питание, ID системы
Управление	Интегрированный сервисный процессор с поддержкой KVM over IP
Размеры (ВхШхГ)	26x225x612 мм

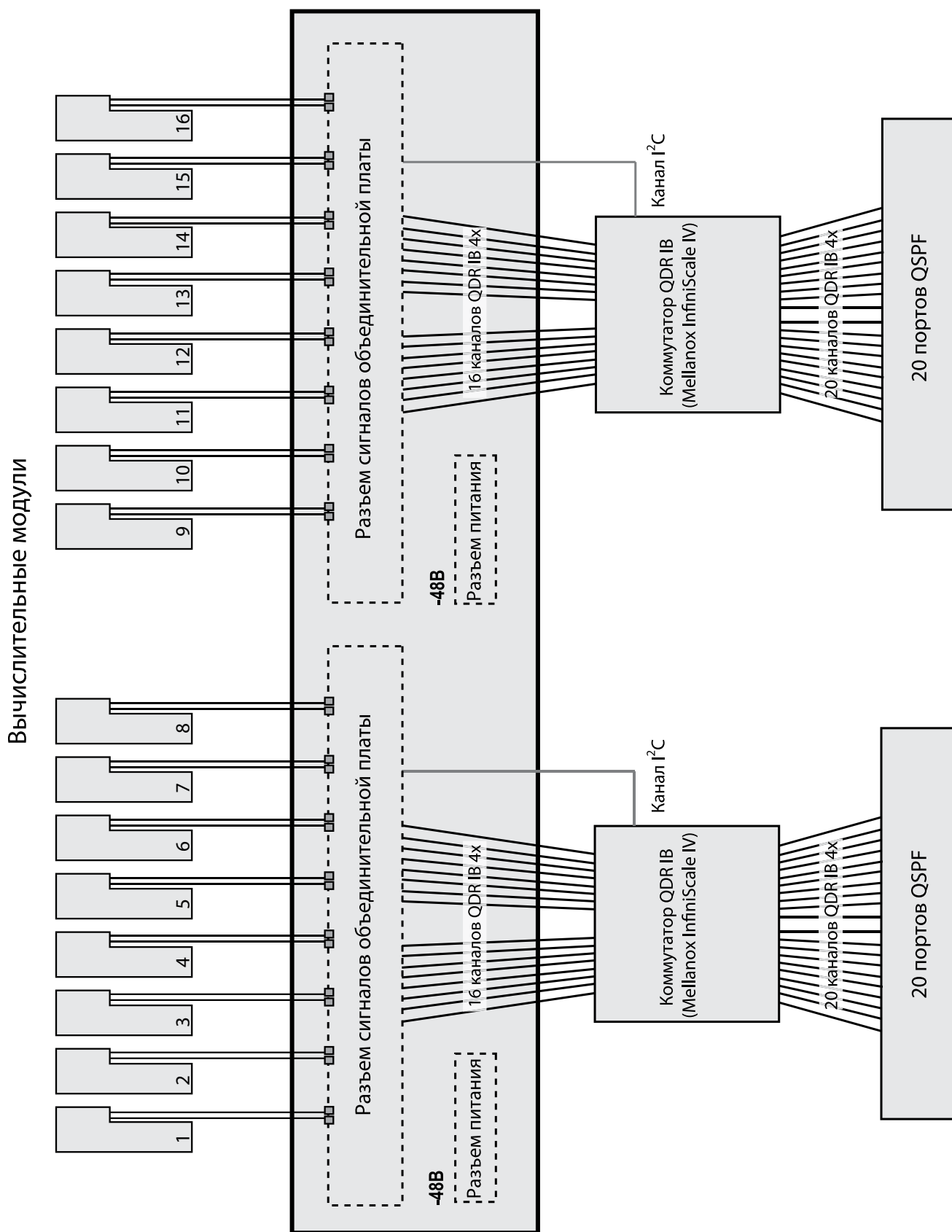
Внешние порты и сети T-Blade 2

Системная сеть	QDR InfiniBand 40 Гбит/с 40 внешних портов на стойку
Управляющая (вспомогательная) сеть	10G Ethernet, 2 внешних порта на корпус
Сервисная сеть	Один внешний порт GbE и один внешний порт 100 Мбит/с на корпус
Глобальная сеть барьерной синхронизации	1 восходящий порт для поддержки топологии больших систем
Глобальная сеть прерываний	1 восходящий порт для поддержки топологии больших систем

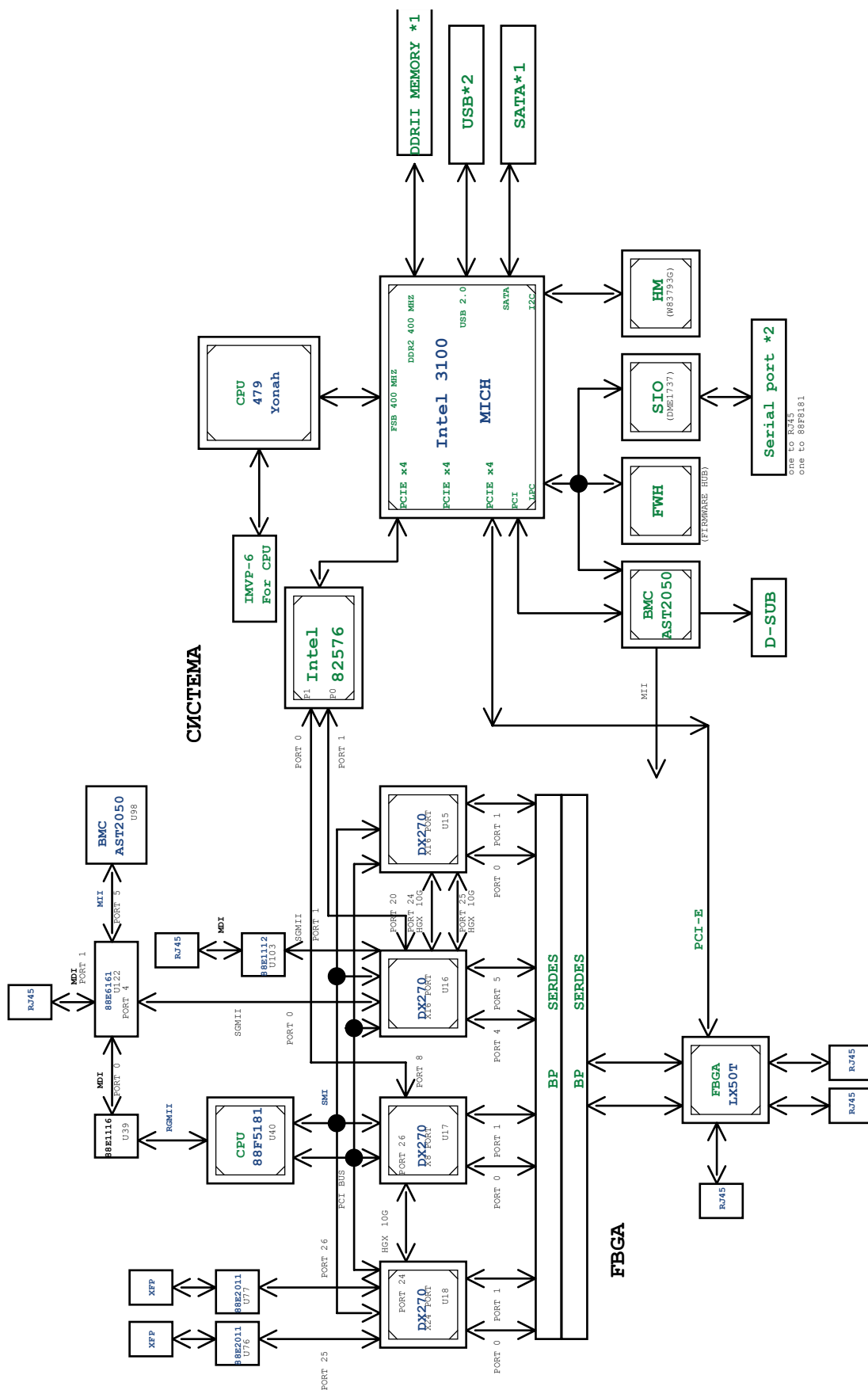
Приложение 1. Диаграмма блейд-модуля



Приложение 2. Диаграмма коммуникаций QDR IB внутри шасси



Приложение 3. Диаграмма модуля коммутации и управления



Заявление об отсутствии гарантий:

Данный документ предназначен только для информационных целей и может содержать неточности. Чтобы получить самую последнюю информацию, свяжитесь с представителем компании "Т-Платформы".

О компании “Т-Платформы”

Компания “Т-Платформы”, основанная в 2002 году, предлагает комплексные системы, программные продукты и услуги для высокопроизводительных вычислений. Системы, которые она поставляет заказчикам, постоянно находятся в рейтинге самых мощных в мире суперкомпьютеров TOP500. Система “Ломоносов”, установленная компанией “Т-Платформы” в Московском государственном университете, получила широкое признание как суперкомпьютер №1 в Восточной Европе и суперкомпьютер №13 в мире.

“Т-Платформы” является поставщиком полного цикла для тех заказчиков, которые хотят воспользоваться выгодами технологий НРС, но не располагают ресурсами, необходимыми для полноценного освоения и внедрения суперкомпьютерной техники. Спектр услуг, предлагаемых компанией “Т-Платформы”, начинается с раннего этапа анализа и документирования требований заказчика и продолжается до проектирования суперкомпьютерного центра “под ключ”. В семействе НРС-систем T-Blade компании применяется надежная операционная система Clustrx®, созданная специально для высокопроизводительных вычислений, которая гарантирует масштабируемость, необходимую для перехода от петамасштаба вычислений к экзамасштабу.

“Т-Платформы” предоставляют также уникальные дополнительные услуги, обеспечивая комплексное моделирование, имитационное моделирование и анализ, и обладают глубокими техническими знаниями и опытом в таких областях, как вычислительная гидродинамика, структурный анализ и другие экстремальные вычислительные дисциплины, оказывая поддержку такого уровня, который не в состоянии обеспечить большинство других поставщиков платформ НРС.

Компания “Т-Платформы” входит в состав группы компаний “Т-Платформы”, которая состоит из компаний “Т-Платформы”, “Т-Сервисы”, “Т-Massive Computing” и “Т-Проектирование” и имеет офисы в Ганновере, Москве, Киеве и Тайбэе.

Дополнительная информация: www.t-platforms.ru.



Ряд стоек T-Blade 1 установки “Чебышев” производительностью 60 ТФлопс в МГУ

“Т-Платформы”
Россия, Москва
Ленинский проспект, 113/1, офис Е-307
Телефон: +7 (495) 956 54 90
Факс: +7 (495) 956 54 15

tPlatforms GmbH
Woehlerstrasse 42, D-30163,
Hannover, Germany
Tel.: +49 (511) 203 885 40
Fax.: +49 (511) 203 885 41

info@t-platforms.ru
<http://www.t-platforms.ru>

© Т-Платформы, 2010

“Т-Платформы”, T-Platforms, логотип T-Platforms, T-Blade, Clustrx TPEdition – товарные знаки или зарегистрированные товарные знаки ОАО “Т-Платформы”. Другие торговые марки и товарные знаки являются собственностью соответствующих владельцев.

Настоящий документ предназначен исключительно для информационных целей. Компания “Т-Платформы” оставляет за собой право вносить изменения без дополнительного уведомления в любые упоминаемые в настоящем документе изделия.