T-ПЛАТФОРМЫ

T-PLATFORMS A-CLASS

СУПЕРКОМПЬЮТЕРЫ С ВОДЯНЫМ ОХЛАЖДЕНИЕМ

T-Platforms A-Class

Новая система T-Platforms A-Class — это энергоэффективное суперкомпьютерное решение с максимальной масштабируемостью и вычислительной плотностью, предназначенное для создания вычислительных систем мультипетафлопсного диапазона. Новая система является нашим наиболее совершенным и комплексным продуктом, вобравшим в себя пожелания заказчиков, многолетний опыт и экспертизу компании.

- Теплоизолированное стоечное шасси 52U с охлаждением компонентов горячей водой.
- 256 узлов с пиковой производительностью 535 Тфлопс.
- Масштабируемость до 192 систем (102,8 Пфлопс).
- Пиковая производительность в расчёте на ватт потребляемой энергии — 3570 Мфлопс/Вт.
- Встроенная двухуровневая коммутация двух независимых сетей Gigabit/10Gigabit Ethernet.
- Встроенная коммутация двух независимых сетей FDR InfiniBand.
- Поддержка топологий 3D и 4D torus, flattened butterfly, hypercube.
- Низкий уровень шума за счёт применения водяного охлаждения.

Модульная архитектура

A-Class является системой уровня стойки с интегрированной сигнально-силовой объединительной платой и подсистемами электропитания и охлаждения. Уникальное 52-юнитовое шасси объединяет в единый вычислительный ресурс 2 головных узла, 256 вычислительных узлов и 60 коммутаторов InfiniBand и Ethernet.

Два независимых модуля управления находятся в верхней части шасси. Основное пространство в системе занимают 8 независимых вычислительных секций, содержащих по 32 вычислительных узла и необходимые сетевые коммутаторы.

Подсистема электропитания также имеет модульный характер, и поделена на 8 независимых групп блоков питания.

Архитектура шасси поддерживает различные конфигурации перспективных вычислительных модулей на базе однопроцессорных и двухпроцессорных узлов или узлов с несколькими ускорителями.

Таким образом, модульный характер системы позволяет заказчикам постепенно наращивать вычислительные и сетевые ресурсы A-Class, устанавливая внутри шасси однотипные или разные вычислительные модули, по мере их выхода на рынок.

Производительность и масштабируемость системы

Пиковая производительность одного шасси в максимальной конфигурации из 256 узлов с ускорителем составляет 535,6 терафлопс. Производительность суперкомпьютера наращивается до 102,8 петафлопс за счёт объединения 192 шасси в единый комплекс.

Первоначальная конфигурация системы A-Class основана на вычислительных узлах с одним процессором Intel® Xeon® E5-2600 v3 и опциональным ускорителем NVIDIA Tesla™ K40. Данная конфигурация обеспечивает общую сбалансированность системы для решения вычислительных задач разного класса. Схема «1+1» позволяет достичь наибольшей эффективности вычислений за счёт полной поддержки технологии GPU Direct и уменьшения накладных расходов, связанных с обеспечением кэш-когерентности двухпроцессорного узла.

Производительность процессоров и ускорителей сбалансирована с доступной пропускной способностью интерфейсов на уровне 3,34 ГБ/с на Тфлопс для FDR InfiniBand или 5,97 ГБ/с на Тфлопс для перспективного EDR InfiniBand.

Производительность системы будет расти за счет поддержки перспективных микропроцессорных архитектур, возможных изменений конфигураций узлов и сетевой среды.

Система охлаждения

Управляющие, вычислительные и коммутационные модули A-Class охлаждаются горячей водой на основе принципа разницы температур. Электронные платы перечисленных модулей крепятся непосредственно на специальный радиатор, плотно прилегающий к компонентам плат для отвода выделяемого тепла.

Блоки питания охлаждаются циркулирующим внутри шасси воздухом, отводящим излучаемое тепло через внутренний водоохлаждаемый теплообменник. Для минимизации излучения тепла в окружающее пространство конструкция шасси теплоизолирована. Технология прямого охлаждения горячей водой позволяет достичь пиковой энергоэффективности A-Class в 3570 Мфлопс/Вт и существенно снизить операционный шум в вычислительном зале.

Применение горячей воды в качестве теплоносителя позволяет внедрить круглогодичный режим «свободного охлаждения» системы, без затрат на закупку компрессоров, холодильных машин и их эксплуатацию. В зимнее время заказчики могут повторно использовать уже нагретую воду для обогрева помещений.

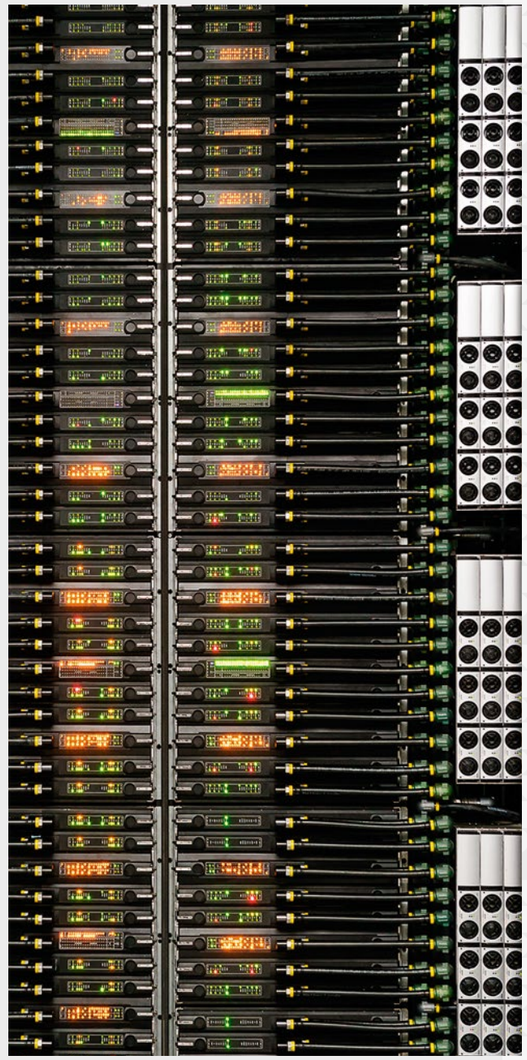
Отказоустойчивость

Система реализует как аппаратные, так и перспективные программные средства повышения отказоустойчивости суперкомпьютера. Два независимых модуля управления с выделенными сетевыми фабриками Ethernet поддерживают горячую замену, обеспечивая функции отказоустойчивого управления и мониторинга системы. Каждый модуль состоит из головного сервера управления и Ethernet- и InfiniBand-коммутаторов верхнего уровня. Независимые серверы управления A-Class позволяют отслеживать состояние компонентов системы и управлять нагрузкой и конфигурациями установленного ПО.

Отказоустойчивая архитектура электроснабжения системы реализована на основе 8 независимых групп высокоэффективных блоков питания, поддерживающих как горячую замену индивидуальных блоков питания, так режим резервирования N+1 в рамках каждой группы, позволяющий продолжить работу всех вычислительных узлов даже в случае потери одного блока питания в каждой группе.

Дополнительная надёжность системы A-Class обеспечивается мониторингом на уровнях шасси, секции и узла. Отслеживаются значения температуры, энергопотребления и других контрольных показателей аппаратных средств. Мониторинг системы жидкостного охлаждения осуществляется с помощью набора датчиков протечек, влажности и давления жидкости. В случае аварийной ситуации система управления автоматически отключает подачу воды и электропитания к шасси.

Для безопасной эксплуатации и упрощения обслуживания A-Class в режиме «онлайн», система охлаждения и электрическая и сетевая инфраструктуры разнесены по разным зонам стоечного шасси.



Сетевая архитектура

Система A-Class обладает сбалансированной интегрированной сетевой инфраструктурой. Вычислительные и управляющие узлы системы подключены через высокоскоростные интерфейсы к двум независимым сетевым фабрикам FDR InfiniBand и двум независимым фабрикам Ethernet-сетей.

Одна из сетей InfiniBand используется для MPI-трафика, вторая служит для подключения к СХД. Фабрика для обмена MPI-трафиком поддерживает различные топологии коммутации вычислительных узлов без использования выделенных корневых коммутаторов, и предоставляет возможность создания произвольной однородной топологии соединения коммутаторов, входящих в состав системы. Для создания максимально эффективной конфигурации системы доступны топологии трехмерного и четырехмерного тора, «плоской бабочки» и гиперкуба.

Ethernet-сети системы A-Class построены на основе двухуровневой топологии. Коммутаторы нижнего уровня располагаются в вычислительных секциях. Коммутатор верхнего уровня каждой из Ethernet-сетей включён в состав отдельного управляющего модуля. Данная организация сети обеспечивает отказоустойчивость и управляемость системы в случае отказа любого компонента управляющих модулей или Ethernet-сетей. По такому же принципу устроена и сеть InfiniBand для подключения к СХД.

ПО управления и мониторинга

Для управления и мониторинга системы A-Class Т-Платформы предлагают собственный программный комплекс ClustrX HPC Pack, который включает в себя следующие компоненты:

- система управления кластером;
- модуль управления пользователями;
- различные менеджеры ресурсов и системы мониторинга на выбор клиента;
- средства управления оборудованием;
- ClustrX Safe — система автоматического отключения оборудования в чрезвычайных ситуациях (САОО).

ClustrX HPC Pack обладает широким функционалом:

- поддержка распределённых сервисных узлов;
- поддержка виртуальных машин;
- возможность работать с разными ОС в рамках одного кластера;
- поддержка локальной и бездисковой загрузки узлов с помощью подключений Ethernet, InfiniBand и iSCSI;
- поддержка различных файловых систем и баз данных;
- гибко настраиваемый виджет-ориентированный интерфейс оператора позволяет представлять нужную информацию о работе системы в удобном и понятном виде.

Также возможно применение других программных комплексов для управления и мониторинга суперкомпьютеров на базе системы A-Class.

Основные характеристики шасси

Конструктив	Специализированное стоечное шасси с воздушно-жидкостным охлаждением, интегрированным высокоскоростным сигнальным и силовым бэкплейном
Габаритные размеры	Ширина 1500 мм × глубина 800 мм × высота 2400 мм, (~52U)
Система управления	2 независимых управляющих модуля, каждый с сервером и коммутаторами управляющих сетей InfiniBand и Ethernet верхнего уровня
Организация вычислительных ресурсов	8 вычислительных секций, в каждой расположены коммутаторы сетей FDR InfiniBand и Ethernet, а также 32 вычислительных узла, поддерживающих установку ускорителей
Сети InfiniBand	Встроенная коммутация двух независимых сетей FDR InfiniBand (двухуровневая для сети доступа к СХД; различные однородные топологии фабрики обмена MPI-трафиком — torus, flattened butterfly, hypercube)
Сети Ethernet	Встроенная двухуровневая коммутация двух независимых сетей Gigabit/10Gigabit Ethernet
Организация системы электропитания	8 × 12 отсеков для высокоэффективных блоков питания (3 кВт, эффективность до 97,2%) с поддержкой горячей замены, электропитание каждой секции обеспечивается с резервированием уровня N+1
Шина питания	48 В постоянного тока с функцией измерения потребления
Потребляемая мощность	Система электропитания обеспечивает совокупную мощность до 256 кВт с учётом резервирования
Параметры электропитания	Входящее напряжение — 380 В трёхфазного переменного тока
Шумность	Низкий уровень операционного шума

Основные характеристики системы охлаждения

Тип	Жидкостный
Охлаждение компонентов	Управляющие модули, вычислительные узлы, встроенные коммутаторы — прямое охлаждение горячей водой. Блоки питания — через жидкостные теплообменники в изолированном пространстве шасси.
Температурные показатели	Температура воды на входе: до 45 °C. Температура воды на выходе: более 50 °C при температуре воды на входе 45 °C.
Прокачиваемый объём воды	До 10,5 л/с
Особенности	Круглогодичное свободное охлаждение при температуре окружающей среды до 35 °C. На территории России рекомендовано применение двухконтурной системы охлаждения.

Основные характеристики системы управления

Схема управления	2 независимых модуля управления системой с поддержкой горячей замены
Состав модуля управления	• Однопроцессорный сервер управления; • Коммутатор верхнего уровня сети Ethernet • Коммутатор верхнего уровня сети FDR InfiniBand, предназначенной для подключения к СХД
Охлаждение модуля управления	Охлаждение горячей водой
Конфигурация сервера управления	1 × Intel E5-2600 v3, TDP 145 Вт; до 32 ГБ DDR4-2133 Reg. ECC, 4 модуля модуля объёмом 8 ГБ 2 × HDD 2.5" 2 × 10GbE SFP+ 2 × FDR InfiniBand (QSFP и бэкплейн)

Основные характеристики вычислительного модуля

Описание	Вычислительный модуль с поддержкой горячей замены (включает 4 вычислительных узла)

Конфигурация вычислительного узла

Процессор	1 × Intel E5-2600 v3, TDP до 145 Вт
Память	до 32 ГБ DDR4-2133 Reg. ECC, 4 модуля объёмом 8 ГБ
Локальные накопители	SSD 256 ГБ (опция)
Сети	2 фабрики GbE
Интерконнект	2 фабрики InfiniBand FDR 56 Гб/с
Ускоритель	1 × NVIDIA Tesla K40 (SXM), TDP 235 Вт

Основные характеристики сетевой инфраструктуры

Ethernet	• 2 независимых сети Gigabit/10Gigabit Ethernet с двухуровневой коммутацией. • Два коммутатора Ethernet нижнего уровня в каждой из вычислительных секций, по 1 коммутатору на сеть. • Один коммутатор Ethernet верхнего уровня в каждом из управляющих модулей. • Каждый коммутатор нижнего уровня подключён к обоим коммутаторам верхнего уровня. • Каждый коммутатор верхнего уровня имеет два соединения 10Gigabit Ethernet для подключения к внешней сети Ethernet. • Общее количество коммутаторов каждой из двух сетей Ethernet — 1 коммутатор верхнего уровня, 8 коммутаторов нижнего уровня. • Общее число внешних портов 10Gigabit Ethernet — 4 шт.
Сеть подключения к СХД	• Сеть с двухуровневой топологией. • Каждая вычислительная секция содержит 1 коммутатор InfiniBand нижнего уровня. • Один коммутатор InfiniBand верхнего уровня в каждом из управляющих модулей. • Каждый коммутатор нижнего уровня подключён к обоим коммутаторам верхнего уровня. • Каждый коммутатор верхнего уровня имеет 18 соединений FDR InfiniBand для подключения к внешней сети InfiniBand. • Общее количество коммутаторов сети подключения к СХД — 2 коммутатора верхнего уровня, 8 коммутаторов нижнего уровня. • Общее число внешних портов FDR InfiniBand сети подключения к СХД — 36 шт.
Сеть обмена MPI-трафиком	• Сеть с вариативной топологией. • Каждая вычислительная секция содержит 4 коммутатора InfiniBand. • Каждый коммутатор имеет 28 соединений FDR InfiniBand для подключения к внешней сети InfiniBand или объединения узлов системы A-Class в самодостаточную сеть обмена MPI-трафиком. • Общее количество коммутаторов сети обмена MPI-трафиком — 32 шт. • Общее число внешних портов FDR InfiniBand сети обмена MPI-трафиком — 896 шт.

Программное обеспечение

ОС	CentOS версий 6.4 и выше, Linux с ядром версии не ниже 2.6.32
Комплекс управления и мониторинга	ClustrX HPC Pack (включает систему аварийного отключения оборудования ClustrX Safe)
Прочее	MPI, другие библиотеки и прикладное ПО — по выбору заказчика

ОАО «Т-Платформы»

Россия, Москва,
Ленинский проспект, д. 113/1, офис В-705.
тел: +7(495) 956 54 90
факс: +7(495) 956 54 15

www.t-platforms.ru

tPlatforms GmbH

Wöhlerstraße 42, D-30163, Hannover, Germany
tel: +49 (511) 203 885 40
fax: +49 (511) 203 885 41

www.t-platforms.com