УТВЕРЖДАЮ

Директор Института программных систем РАН

_____ А.К.Айламазян

_____ 1999 г.

УТВЕРЖДАЮ

Генеральный директор

"НИО "Кибернетика" НАН Беларуси

_____ В.С. Танаев

_____ 1999 г.

КОНЦЕПЦИЯ

СОЗДАНИЯ МОДЕЛЕЙ СЕМЕЙСТВА

СУПЕРКОМПЬЮТЕРОВ

(Редакция 1 от 15.12.1999 г.)

СОГЛАСОВАНО:

От исполнителя Совместной Программы

Директор ИЦМС ИПС РАН

_____ С.М. Абрамов

СОГЛАСОВАНО:

От головного исполнителя Совместной Программы

Заместитель генерального директора

НИО "Кибернетика" НАН Беларуси

_____ В.В. Анищенко

Исполнительный директор Совместной Программы

_____ Н.Н. Парамонов

От предприятия "Суперкомпьютерные системы"

Генеральный директор

_____ Ю.В. Татур

От ГП "НИИ ЭВМ"

Заместитель директора по научной работе

_____ Д.Б. Жаворонков

От АО "НИЦЭВТ"

Генеральный директор

_____ В.В. Митрофанов

От ГП "МНИЭВТ"

Главный инженер

_____ К.Ю. Карлинович

От НИКТП "Белмикросистема"

Заместитель директора по научной работе

_____ А.К. Белоус

Содержание

Введение

Концепция создания моделей семейства суперкомпьютеров (далее по тексту **Концепция**) разработана на основе научно-технического задела, созданного российскими и белорусскими специалистами в области высокопроизводительных вычислений, с учетом основных положений совместной белорусско-российской программы "Разработка и освоение в серийном производстве семейства моделей высокопроизводительных вычислительных систем с параллельной архитектурой (суперкомпьютеров) и создание прикладных программно-аппаратных комплексов на их основе" (далее по тексту Совместная Программа)

Целью концепции является выработка единого сбалансированного стратегического подхода для всех многочисленных исполнителей Совместной Программы на всех этапах создания моделей семейства суперкомпьютеров.

Концепция отражает основополагающие принципы создаваемых по Совместной Программе суперкомпьютерных систем:

- базовые архитектурные решения;
- основные принципы и решения в части базового (общесистемного) программного обеспечения;
- основные принципы и решения в части аппаратных средств;
- идеологию создания моделей семейства суперкомпьютеров;
- основные принципы разработки конструкторской документации, проведения испытаний суперкомпьютерных систем и организации их серийного производства;
- общую схему реализации прикладных суперкомпьютерных конфигураций.

Основные принципы изложены с различной степенью детализации, что отражает различия в глубине их проработки, естественные для **этапа разработки** Совместной Программы. На соответствующих **этапах реализации** Совместной Программы отдельные принципы и решения будут развиваться и уточняться (детализироваться).

1. Двухуровневая открытая масштабируемая архитектура

1.1 Кластерный и потоковый уровни суперкомпьютеров

Концепция создания моделей семейства суперкомпьютеров базируется на двухуровневой масштабируемой архитектуре, обеспечивающей возможность совместного (в рамках одной вычислительной системы) использования двух различных архитектурных аппаратных решений (уровней):

1-й уровень - классический ***кластер (тесносвязанная сеть) из вычислительных узлов*** (Рис. 1, позиция 2), реализованных с использованием компонент широкого применения (стандартных микропроцессоров, модулей памяти, жестких дисков и материнских плат, в том числе материнских плат с поддержкой SMP);

2-й уровень - модули ***однородной вычислительной среды*** (ОВС—Рис. 1, позиция 4)—вычислительная среда с топологией плоской решетки из большого числа микропроцессоров, объединенных потоковыми линиями передачи данных.

Этим двум архитектурным аппаратным решениям соответствуют различные подходы к организации параллельных вычислений:

- ***для кластеров*** разработан целый ряд моделей организации параллельного счета, поддерживаемых соответствующими программными средствами—Т-система, MPI, PVM, Norma, DVM и др.
- ***в архитектуре ОВС*** для организации параллельного исполнения задачи наиболее адекватна модель потоковых вычислений (data-flow).

При дальнейшем изложении Концепции для обозначения первой и второго архитектурных уровней двухуровневой архитектуры используются соответственно термины - кластерный уровень и потоковый уровень или уровень ОВС.

Кластерный уровень (Рис. 1, позиция 2)—тесно связанная сеть (кластер) вычислительных узлов, работающих под управлением ОС Linux—одного из клонов широко используемой многопользовательской универсальной операционной системы UNIX. Для организации параллельного выполнения прикладных задач на данном уровне используется:

- оригинальная система поддержки параллельных вычислений—Т-система, реализующая автоматическое динамическое распараллеливание программ;
- классические системы поддержки параллельных вычислений, обеспечивающие эффективное распараллеливание прикладных задач различных классов (как правило—задач с явным параллелизмом): MPI, PVM, Norma, DVM и др.

Потоковый уровень (Рис. 1, позиция 4)—сеть вычислительных узлов, каждый из которых—это модуль на базе однородной вычислительной среды (Рис. 1, позиция 7). Модуль состоит из большого числа (десятки тысяч) простых однобитовых процессоров. Потоковый уровень позволяет эффективно реализовывать те задачи (фрагменты прикладных проблем), которые неэффективно решаются на кластерном уровне суперкомпьютера.

Аппаратные и программные ***средства поддержки взаимодействия кластерного и потокового уровня*** (Рис. 1, позиция 5 и 6) суперкомпьютера позволяет эффективно запускать из одного уровня суперкомпьютера фрагмент задачи для решения на другом уровне.

Предпосылкой объединения двух—кластерного и потокового,—аппаратных решений и соответствующих им программных средств для организации параллельной обработки в рамках одной вычислительной системы, является то, что эти два подхода своими сильными сторонами компенсируют недостатки друг друга. Тем самым, в общем случае, каждая прикладная проблема может быть разбита на:

- фрагменты со сложной логикой вычисления, с крупноблочным (явным статическим или скрытым динамическим) параллелизмом—такие фрагменты эффективно реализовывать на кластерном уровне с использованием Т-системы и других систем поддержки параллельных вычислений (MPI, PVM, Norma, DVM и др.);
- фрагменты с простой логикой вычисления, с конвейерным или мелкозернистым явным параллелизмом, с большими потоками информации, требующими обработки в реальном режиме времени—такие фрагменты эффективно реализовывать в ОВС.

Пропорции такого деления прикладной проблемы определяют:

- пропорции объемов программного обеспечения для нее в части кластерного и потокового уровней;
- эффективный состав вычислительной системы для данной прикладной задачи—количество вычислительных узлов кластерного и потокового уровня, набор и характеристики необходимого коммутационного оборудования и т.д.

В рамках единой идеологии на базе двухуровневой архитектуры предусматривается:

- создание моделей семейства суперкомпьютеров только на основе программно-аппаратных средств кластерного уровня - модели суперкомпьютеров с кластерной архитектурой;
- создание моделей семейства суперкомпьютеров только на основе программно-аппаратных средств потокового уровня - модели суперкомпьютеров с потоковой архитектурой или с архитектурой ОВС;
- создание моделей семейства суперкомпьютеров на основе программно-аппаратных средств обоих архитектурных уровней (кластерного и потокового) и средств поддержки взаимодействия этих уровней - модели суперкомпьютеров со смешанной (двухуровневой) архитектурой;
- объединение моделей суперкомпьютеров с разными или одинаковыми архитектурными решениями (кластерная, потоковая, смешанная) в единую вычислительную систему.

Архитектура является открытой и масштабируемой, т.е. не накладывает жестких ограничений к программно-аппаратной платформе узлов кластера, топологии вычислительной сети, конфигурации и диапазону производительности суперкомпьютеров. Такой подход позволяет создавать прикладные суперкомпьютерные системы в широком диапазоне

производительности—от изделий класса высокопроизводительных серверов и мультипроцессорных рабочих станций (миллиарды операций в секунду) до вычислительных систем с массовым параллелизмом сверхвысокой производительностью (триллионы операций в секунду).

1.2 Сетевые средства поддержки взаимодействия вычислительных узлов суперкомпьютера

Для организации взаимодействия вычислительных узлов суперкомпьютера в его составе используются различные сетевые (аппаратные и программные) средства в совокупности образующие четыре системы передачи данных:

Системная сеть кластера (СС-SAN) объединяет узлы кластерного уровня в кластер. Данная сеть поддерживает масштабируемость кластерного уровня суперкомпьютера, а также пересылку и когерентность данных во всех вычислительных узлах кластерного уровня суперкомпьютера. Системная сеть кластера строится на основе специализированных технологий класса SCI, cLan или Myrinet, предназначенных для эффективной поддержки кластерных вычислений и соответствующей программной поддержки на уровне ОС Linux и систем организации параллельных вычислений (Т-система, MPI, PVM и др.).

Вспомогательная сеть суперкомпьютера (ВС-LAN с протоколом TCP/IP) объединяет узлы кластерного уровня в обычную (TCP/IP) локальную сеть. Данная сеть может быть реализована на основе широко используемых сетевых технологий класса Fast Ethernet, Gigabit Ethernet или ATM. Сеть ВС-LAN предназначена для управления системой, для подключения рабочих мест пользователей, интеграции суперкомпьютера в локальную сеть предприятия и/или в глобальные сети. Кроме того, данный уровень может быть использован и системой организации параллельных кластерных вычислений (Т-система, MPI, PVM и др.) для вспомогательных целей (основные потоки информации, возникающие при организации параллельных кластерных вычислений, передаются через системную сеть кластера SAN).

Примечание: В некоторых случаях аппаратура системной сети кластера (SAN) позволяет без ущерба для реализации кластерных вычислений поддержать на этой же аппаратуре реализацию сети TCP/IP. В этих случаях, в некоторых моделях суперкомпьютера аппаратные части обоих сетей (SAN и LAN TCP/IP) могут быть совмещены.

Потоковая сеть ОВС (ПС-ОВС)—совокупность каналов и средств управления ими, объединяющая необходимое количество модулей ОВС (МОВС) для выполнения конкретной программы. Данная сеть обеспечивает пересылку данных между модулями ОВС и масштабируемость потокового уровня суперкомпьютера.

Средства взаимодействия двух уровней суперкомпьютера (СВКУиОВС)—обеспечивают возможность взаимодействия между кластерным и потоковым уровнем суперкомпьютера и реализуются в рамках сетей СС-SAN или ВС-LAN (Рис. 1, позиция 6). То есть, при реализации в модулях ОВС соответствующих сетевых интерфейсов, модули ОВС в принципе могут выступать в качестве устройств:

- системной сети кластера (СС-SAN) и/или
- вспомогательной сети суперкомпьютера (ВС-LAN с протоколом TCP/IP).

Примечание: Взаимодействие двух уровней суперкомпьютера в принципе может осуществляться и средствами аппаратной и программной реализации передачи данных (например по протоколу SCSI) между некоторыми (возможно всеми) узлами кластерного уровня и некоторыми (возможно всеми) модулями ОВС вне рамок сетей СС-SAN или ВС-LAN (Рис. 1, позиция 5).

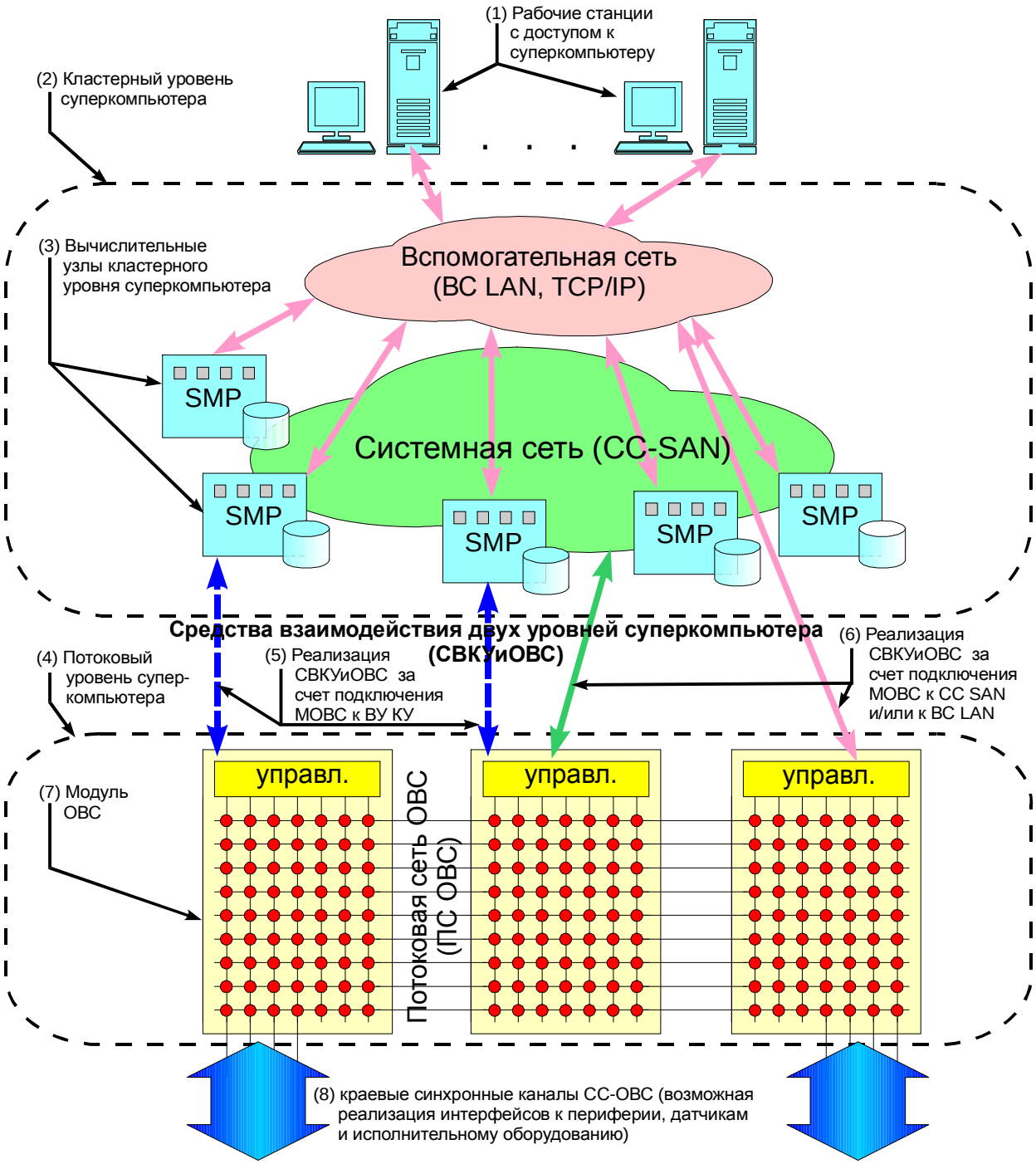

Рис.1 Двухуровневая архитектура суперкомпьютера

1.3 Средства поддержки передачи данных между суперкомпьютером и вспомогательной периферией

Основная периферия (в первую очередь дисковые накопители) может подключаться (непосредственно через системную шину вычислительных устройств и соответствующие аппаратные адаптеры) к вычислительным узлам кластерного уровня.

Кроме того, в отдельных моделях семейства суперкомпьютеров могут быть использованы и другие, дополнительные средства обмена (например на базе устройств семейства Fiber Channel) с периферийными устройствами. Такие средства предназначаются для обеспечения высокоскоростного обмена данными с внешними устройствами—дисковыми накопителями высокой емкости и т.п., а также для решения задачи объединения в единую высокопроизводительную параллельную вычислительную систему нескольких суперкомпьютеров.

Отдельно надо отметить, что БМ ОВС могут быть эффективно использованы для реализации (Рис. 1, позиция 8) высокоскоростного потокового обмена:

- со стандартной компьютерной периферией и/или
- с нестандартными устройствами—датчиками и/или исполнительными механизмами той или иной прикладной системы, создаваемой с использованием суперкомпьютера (подключение систем управления оружием, систем съема информации об окружающей обстановке и т.п.).

1.4 Анализ особенностей и преимуществ (по сравнению с аналогичными разработками) предложенной схемы реализации суперкомпьютеров

Предложенная двухуровневая схема реализации аппаратных и программных средств суперкомпьютера обладает рядом преимуществ (по сравнению с аналогичными разработками), позволяющими достичь и превысить мировой уровень сегодняшних результатов в суперкомпьютерной отрасли. К таким особенностям технических решений, положенным в основу Совместной Программы относятся:

- **в части Т-системы**: обеспечивается автоматическое динамическое распараллеливание программ и, таким образом, достигается освобождение программиста от большинства аспектов разработки параллельных программ, свойственных различным системам ручного статического распараллеливания:

– обнаружение готовых к выполнению фрагментов задачи (процессов);
– их распределение по процессорам;
– их синхронизацию по данным.

Все эти (и другие) операции выполняются в Т-системе автоматически и в динамике (во время выполнения задачи). Тем самым достигается **более низкие затраты на разработку параллельных программ и более высокая их надежность.**

По сравнению с использованием распараллеливающих компиляторов, Т-система обеспечивает более глубокий уровень параллелизма во время выполнения программы и более полное использование вычислительных ресурсов мультипроцессоров. Это связано с принципиальными алгоритмическими трудностями (алгоритмически неразрешимыми проблемами), не позволяющими во время компиляции (в статике) выполнить полный точный анализ и предсказать последующее поведение программы во время счета.

Кроме указанных выше принципиальных преимуществ Т-системы перед известными сегодня методами организации параллельного счета, в реализации Т-системы имеется ряд технологических находок, не имеющих аналогов в мире, в частности:

– **реализация понятия "неготовое значение"** и поддержка корректного выполнения некоторых операций над неготовыми значениями. Тем самым поддерживается возможность выполнение счета в некотором процессе-потребителе в условиях, когда часть из обрабатываемых им значений еще не готова—не вычислена в соответствующем процессе-поставщике. Данное техническое решение обеспечивает обнаружение более глубокого параллелизма в программе;
– **оригинальный алгоритм динамического автоматического распределения процессов по процессорам.** Данный алгоритм учитывает особенности неоднородных распределенных вычислительных сетей. По сравнению с известными алгоритмами динамического автоматического распределения процессов по процессорам (например, с диффузионным алгоритмом и его модификациями), алгоритм Т-системы имеет существенно более низкий трафик межпроцессорных передач. Тем самым, Т-система обеспечивает снижение накладных расходов на организацию параллельного счета и предъявляет менее жесткие требования к пропускной способности аппаратуры объединения процессорных элементов в кластер.

- **в части ОВС:** архитектура ОВС позволяет использовать естественный параллелизм решаемой задачи вплоть до битового уровня, то есть уровня структуры обрабатываемых данных, а также позволяет строить конвейеры произвольной глубины. ОВС предоставляет возможность одновременной обработки множества независимых некогерентных потоков, а также обеспечивает удобную организацию ввода/вывода данных на вычислительную матрицу ОВС.

Фактически, при решении конкретной функции или самостоятельной задачи, на ОВС путем ввода новой программы организуется **спецпроцессор**, реализующий решаемую

функцию или задачу с наибольшей эффективностью. На матрице ОВС одновременно могут решаться несколько независимых задач и функций, причем механизм перезагрузки сегментов ОВС позволяет перезагружать часть матрицы без остановки выполнения еще незавершенных задач. В качестве западного аналога ОВС могут выступать систолические структуры. Но, реализуя все возможности систолических структур, ОВС обладает значительно большей гибкостью и перестраиваемостью. В частности, ОВС обладает полной аппаратной и программной масштабируемостью, что позволяет строить на базе матрицы ОВС вычислительные системы с большим быстродействием. Производительность ОВС, теоретически, растет линейно с увеличением рабочей частоты поля и площади вычислительной матрицы.

ОВС позволяет создавать системы с высоким уровнем надежности и отказоустойчивости, эффективно реализовывать нейросетевые алгоритмы.

За счет оригинальных технических решений в части ОВС улучшаются важнейшие характеристики изделий:

- потребляемая мощность на единицу объема;
- габариты и вес;
- соотношение стоимость/производительность;
- производительность на единицу объема.

- **в части суперкомпьютерной вычислительной системы в целом** (с учетом обоих уровней - кластерного и потокового) вычислительная система, построенная на базе двух вычислительных моделей, предлагает пользователям уникальные возможности для оптимального решения прикладных задач.

Высокопроизводительные вычислительные системы с параллельной архитектурой (***суперкомпьютеры***), реализуемые на базе двухуровневой архитектуры позволяют реализовать любые виды параллелизма. Архитектура является открытой и масштабируемой, то есть не накладывает жестких ограничений к программно-аппаратной платформе узлов кластера, топологии вычислительной сети, конфигурации и диапазону производительности суперкомпьютеров. Вычислительные системы, создаваемые на базе основополагающего концептуального архитектурного принципа могут оптимально решать как классические вычислительные задачи математической физики и линейной алгебры, так и специализированные задачи обработки сигналов, моделирования виртуальной реальности, задачи управления сложными системами в реальном времени.

2. Базовое (общесистемное) программное обеспечение суперкомпьютера

2.1 Программное обеспечение кластерного уровня суперкомпьютера

2.1.1 Операционная система вычислительных узлов кластерного уровня

В качестве ОС в универсальном кластерном суперкомпьютере используется операционная система Linux, являющаяся клоном ОС Unix. Операционная система Linux является одной из самых надежных, эффективных и перспективных операционных систем, которую сегодня многие коммерческие и государственные организации выбирают в качестве базовой для приложений и перспективных разработок в области параллельных вычислений.

Основания для выбора ОС Linux:

- ОС Linux распространяется свободно, то есть бесплатно и с исходными текстами. Это дает возможность модифицировать тексты ОС Linux, если это потребуется для реализации прикладной задачи.
- Вследствие того, что ОС Linux распространяется в виде исходных текстов наличие в коде операционной системы различных «закладок» и «троянских коней» полностью исключено.
- Функциональные возможности ОС Linux и ее утилит развиваются огромной армией добровольных программистов-разработчиков (сегодня 7-10 миллионов установок в мире),

что обеспечивает непрерывность ее тестирования и корректировки ошибок в исходных текстах.
- Распространение ОС Linux не подвержено каким-либо ограничениям каких-либо стран или фирм.
- ОС Linux является открытой, то есть она реализована не только для платформ класса IBM PC, но и для многих других аппаратных платформ.

Выбор одной из самых развитых на сегодняшний день операционных систем—ОС Linux—обеспечивает наличие поддержки на кластерном уровне суперкомпьютера всех существующих (и возможных в будущем) требований в части:
- периферийных устройств (включая поддержку подключения к LAN/WAN/Internet);
- различных режимов функционирования (многопользовательский, удаленный с криптографическими протоколами доступа, оконный интерфейс и т.п.);
- сервисных программных средств.

2.1.2 Системы поддержки параллельных вычислений на кластерном уровне суперкомпьютера

Для управления работой суперкомпьютера и организации процесса параллельных вычислений на кластерном архитектурном уровне наряду с ОС Linux используются ***системы поддержки параллельных вычислений***: Т-система, MPI, PVM, DVM, Norma и т.д.:
- ***Классические системы поддержки параллельных вычислений***—MPI, PVM, DVM, Norma и др.—являются эффективным средством реализации параллельного исполнения на кластерном уровне суперкомпьютера тех прикладных задач (фрагментов задач), для которых параллелизм является явным—то есть он может быть вскрыт и описан в статике (до запуска задачи) с использованием конструкций и понятий, принятых в данных системах.
- ***Т-система*** позволяет организовать параллельное исполнение прикладных задач (фрагментов задач) как с явным, так и со скрытым (динамическим) параллелизмом.

Т-система - система поддержки параллельных вычислений на кластерном уровне суперкомпьютера, включающая в себя операционную среду поддержки исполнения параллельных приложений и соответствующие средства программирования (систему программирования).

Использование Т-системы является одним из базовых концептуальных принципов, обеспечивающих достижение современного мирового уровня в области параллельных вычислений. Т-система—оригинальное программное обеспечение для мультипроцессоров, реализующее концепцию автоматического динамического распараллеливания программ.

Задачей Т-системы является обеспечение такой организации процесса параллельных вычислений, которая позволила бы эффективно использовать вычислительную мощность всех кластерных узлов вычислительной системы в целом. В рамках Т-системы процесс вычислений представляется в виде вычислительной сети, включающей в себя данные и Т-процессы—элементарные подзадачи (гранулы параллелизма). Используя соответствующие алгоритмы внешнего планирования—алгоритмы назначения Т-процессов к вычислению в том или ином вычислительном узле кластерного уровня,—Т-система обеспечивает высокий уровень использования вычислительной мощности каждого из выделенных данной прикладной задаче вычислительных узлов, а тем самым—и высокий коэффициент утилизации вычислительной мощности всей вычислительной системы в целом.

Особенно эффективно использование Т-системы в задачах с неявным параллелизмом, т.е. задачах, параллелизм которых в статике (в момент написания программы или ее компиляции) трудно выявить. Тем самым, применение иных подходов (MPI, PVM и др.) к данным задачам затруднен, а в некоторых случаях—невозможен. К такому классу можно отнести некоторые задачи, искусственного интеллекта, компьютерной алгебры, сложные алгоритмы обработки нечисловой информации и др.

Программы, разработанные для Т-системы без переработки и без перекомпиляции могут исполняться на мультипроцессоре с любым числом процессоров и с любой коммуникационной архитектурой. Это позволяет:

- Без переработки системного и прикладного программного обеспечения расширять конфигурации систем (добавлять число процессоров в систему) и тем самым достигать ускорения счета прикладных программ.
- Без переработки системного и прикладного программного обеспечения исполнять задачи на частично неисправной системе (с вышедшими из строя компонентами). В принципе, возможно достижение более сильной устойчивости к отказам аппаратуры—достижение эффекта продолжения счета без перезапуска задачи после выхода из строя части оборудования. При этом выбираются те компоненты системы, которые позволяют наиболее эффективно реализовать процесс.

В качестве языка программирования Т-системы используется расширение языка Си новыми Т-конструкциями.

Т-система позволяет использовать вычислительные узлы, обладающие различным уровнем вычислительной мощности. Благодаря этому свойству, кластерный архитектурный уровень создаваемых суперкомпьютеров может строиться из разных по мощности вычислительных узлов.

2.2 Программное обеспечение потокового архитектурного уровня.

Так как управление БМ ОВС реализуется на основе стандартного процессора, то очевидно он должен использовать ту же операционную систему, что и кластерный уровень-Linux. Для выполнения задач на ОВС не требуется (по крайней мере на первом этапе реализации Совместной Программы) наличия ядра Т-системы в контроллере.

Для успешного функционирования потокового уровня необходима разработка таких программных средств как:

- диспетчер задач
- компоновщик программ для матрицы ОВС
- драйверы связи БМ ОВС
- драйверы поддержки соответствующих внешних устройств, подключаемых к матрице ОВС.

В состав программного обеспечения потокового архитектурного уровня входят также необходимые инструментальные средства разработчика программ для ОВС:

- Трансляторы для языков программирования различного уровня абстракции.
- Библиотеки стандартных программ для ОВС.
- Интегрированные средства подготовки и отладки программ.

2.3 Программные средства сопряжения кластерного и потокового архитектурных уровней.

Программные средства сопряжения *в части кластерного уровня* должны быть поддержаны:

- набором драйверов устройств, обеспечивающих сопряжение кластерного и потокового уровня;
- базовой библиотекой стандартных примитивов обмена информацией и управления ОВС;
- библиотекой прикладных задач и подпрограмм, реализуемых с использованием ОВС;
- структурами данных и программными механизмами, обеспечивающими:
  - передачу Т-процесса, из которого осуществляется взаимодействие с ОВС, в один из вычислительных узлов кластерного уровня, имеющих физический интерфейс с ОВС;
  - осуществление удаленного вызова функции/прикладной задачи из вычислительного узла кластерного уровня, не имеющего интерфейса с ОВС с использованием механизмов, предназначенных для распределенной работы с файлами.

В части ОВС программные средства сопряжения должны включать в себя реализованный в виде специализированной библиотеки набор предназначенных для загрузки из кластерной компоненты в ОВС фрагментов программного кода, каждый из которых непосредственно реализует в ОВС ту или иную прикладную задачу или фрагмент вычислений, в частности:

- получает из кластерного уровня наборы входных данных;
- организует и осуществляет выполнение в ОВС вычислений в соответствии с алгоритмом решения соответствующей прикладной задачи;

- передает из ОВС в кластерный уровень наборы данных, содержащие результаты вычислений.

Описанный набор программных средств, структур данных и механизмов поддерживает возможности:

- из Т-программы эффективно передать на вычисление в ОВС фрагмент решаемой задачи;
- из выполняемого в ОВС кода передать на выполнение в кластерную компоненту фрагмент решаемой задачи.

3. Аппаратная реализация суперкомпьютера

3.1 Аппаратная реализация кластерного уровня суперкомпьютера.

Основным конструктивом для аппаратной реализации кластерного уровня является ***базовый вычислительный модуль кластерного уровня (БВМ КУ)***. Аппаратная реализация БВМ КУ основывается на использовании готовых изделий массового применения, широко представленных на сегодняшнем компьютерном рынке.

В состав аппаратных средств БВМ КУ входят:

- стандартные (монопроцессорные, 2-х и 4-х процессорные SMP) Intel-совместимые материнские (системные) платы;
- микропроцессоры класса Intel Pentium II/III и выше (или совместимых процессоров других фирм, например AMD Athlon) с тактовыми частотами микропроцессоров 400MHz и выше;
- стандартные модули памяти (объем ОЗУ вычислительного узла порядка 0.5–1GB и выше);
- стандартный жесткий диск (объемом 6–10GB и выше);
- адаптеры для подключения БВМ КУ к:

– системной сети кластера (СС-SAN) и ;

– вспомогательной сети суперкомпьютера (ВС-LAN с протоколом TCP/IP).

Кроме того, в БВМ КУ предусматривается наличие стандартных средств (адаптеров, интерфейсов) для ***возможного*** подключения дополнительного набора периферийных устройств к ВУ (дополнительных НЖМД, монитора, клавиатуры и т.п.).

Широкое представительство перечисленных выше комплектующих для БВМ КУ на компьютерном рынке предопределяет разумные цены на эти изделия и стабильность поставок.

Примечание: В некоторых моделях суперкомпьютера возможна реализация БВМ КУ на базе иной (по сравнению с Intel) платформе, поддержанной в ОС Linux. Например, возможно использование платформ на базе процессора Alpha и др.

Аппаратные средства системной сети кластера (СС-SAN) реализуются на основе специализированных технологий класса SCI, cLan или Myrinet, предназначенных для эффективной поддержки кластерных вычислений.

Аппаратные средства вспомогательной сети суперкомпьютера (ВС-LAN с протоколом TCP/IP) реализуются на основе сетевых технологий класса Fast Ethernet, Gigabit Ethernet или ATM.

Примечание: В некоторых случаях аппаратура системной сети кластера (СС-SAN) позволяет без ущерба для реализации кластерных вычислений поддержать на этой же аппаратуре реализацию сети TCP/IP. В этих случаях в некоторых моделях суперкомпьютера аппаратные части обоих сетей (СС-SAN и ВС-LAN TCP/IP) могут быть совмещены.

Конструктивное исполнение БВМ КУ может быть в виде отдельной стойки с соответствующим оборудованием (системные платы, периферийные адаптеры, сетевое оборудование, блоки питания и т.п.) или в виде модуля, встраиваемого в конструктив более высокого уровня.

Конструктив стойки может быть оригинальной разработки или базироваться на стандартных монтажных стойках различных типоразмеров, снабженных средствами вентиляции и электропитания.

Таким образом, отдельный вычислительный узел кластера представляет собой полноценную самостоятельную монопроцессорную или мультипроцессорную (SMP)

вычислительную систему. Количество типоразмеров БВМ КУ определяется исходя из требуемого диапазона производительности семейства суперкомпьютеров и требований оптимизации создания прикладных суперкомпьютерных систем (массогабаритные параметры, стоимость, конструктивная реализация, область применения и др.)

3.2 Аппаратная реализация потокового архитектурного уровня.

Потоковый уровень (уровень ОВС) суперкомпьютера представляет собой совокупность большого числа (10^4-10^7) однобитовых процессоров с набором команд, аналогичным RISC-процессору, реализованных на СБИС.

Основным конструктивом для аппаратной реализации уровня ОВС является базовый вычислительный модуль ОВС (БВМ ОВС).

БВМ ОВС является самостоятельной конструктивной и функциональной единицей, которая может функционировать как самостоятельно, так и в составе сложного вычислительного комплекса.

Каждый базовый вычислительный модуль представляет собой открытый по производительности, объемам запоминающих устройств, типам и количествам интерфейсов аппаратно-программный вычислительный комплекс и состоит из двух основных частей:

- поля однородной вычислительно-запоминающей среды (ОВЗС);
- контроллера модуля.

Поле ОВС представляет собой вычислительную структуру, состоящую из однотипных процессоров, образующих прямоугольную решетку.

После программирования поле превращается в проблемно-ориентированный вычислитель, реализующий заданную потоковую программу. Каждый процессор поля выполняет одну из команд потоковой программы на протяжении всего времени работы, вплоть до смены программы. Распределение команд по процессорам производится как компилятором в процессе трансляции, так и контроллером модуля в процессе загрузки исполняемой программы.

Управление обработкой производится как потоками данных, проходящими через поле, так и контроллером модуля, обеспечивающим смену программ.

Возможны два режима работы поля ОВС: статический и динамический. При статическом режиме поле не перепрограммируется до тех пор, пока данные от источника в ходе выполнения данной задачи не будут получены, обработаны и переданы потребителю. Динамический режим подразумевает перепрограммирование поля или отдельных его частей, то есть смену потоковой программы в ходе обработки данных одной задачи. Промежуточные данные при этом сохраняются в памяти БВМ ОВС.

Контроллер модуля предназначен для решения задач управления и обмена информацией на уровне БВМ ОВС. Контроллер должен также обеспечивать возможность подключения устройств из номенклатуры периферийных устройств, подключаемых через стандартные интерфейсы. Одной из функций контроллера является обеспечение сопряжения БВМ ОВС с БВМ КУ.

Поле ОВЗС реализуется на основе оригинальных СБИС. В зависимости от достигнутого уровня технологии предусматривается использование корпусного варианта СБИС и технологии многокристальных модулей (МКМ) бескорпусной сборки. Используемый тип СБИС оказывает непосредственное влияние на конструктив поля ОВЗС и БВМ ОВС в целом.

Основой для реализации контроллера модуля является конструктив системного блока стандартной ПЭВМ.

Периферийные устройства могут подключаться к контроллеру БВМ, используя адаптеры, подключаемые к системной шине контроллера.

Предусматривается также возможность подключения периферийных устройств непосредственно к полю ОВЗС.

Основные функциональные блоки БВМ ОВС комплектуются в стойке, оснащенной блоками питания и средствами вентиляции.

Конструктив стойки может быть оригинальной разработки или базироваться на стандартных монтажных стойках различных типоразмеров.

Предусматривается также реализация БВМ ОВС в виде модуля, встраиваемого, например, в БВМ КУ или в конструктив более высокого уровня. Количество типоразмеров БВМ ОВС определяется исходя из требуемого диапазона производительности семейства суперкомпьютеров и требований оптимизации создания прикладных суперкомпьютерных систем (массогабаритные параметры, стоимость, конструктивная реализация, область применения и др.).

3.3 Базовые конфигурации суперкомпьютерных систем

Базовые конфигурации суперкомпьютерных систем (БКСС) отображают весь спектр возможных прикладных суперкомпьютерных реализаций на базе основополагающего концептуального принципа создания семейства суперкомпьютеров—двухуровневой открытой, масштабируемой архитектуры.

Конфигурация вычислительной системы (по определению)—это совокупность функциональных частей вычислительной системы и связей между ними, обусловленная основными техническими характеристиками этих частей, а также характеристиками решаемых задач обработки данных.

В состав базовых конфигураций входят базовые вычислительные модули кластерного и потокового уровней (БВМ КУ и БВМ ОВС), сетевые средства поддержки взаимодействия вычислительных узлов (СС-SAN, ВС-LAN-TCP/IP, ПС ОВС, СВ КУ и ОВС) и соответствующее базовое (общесистемное) и прикладное программное обеспечение.

Количество типов БКСС определяется с учетом следующих основных моментов:

- диапазон производительности создаваемых на базе БКСС прикладных суперкомпьютерных систем и характерные области их применения;
- принятый порядок разработки суперкомпьютерных реализаций, включая разработку конструкторской документации и проведение соответствующих испытаний;
- освоение суперкомпьютерных конфигураций в производстве, включая подготовку производства, создание необходимого производственного задела, проведение различных типов испытаний, потребительский спрос, поставку конкретных прикладных суперкомпьютерных конфигураций;
- техническое обслуживание у потребителей суперкомпьютерных изделий.

3.4 Аппаратная реализация суперкомпьютерных систем разного назначения

На основе базовых конфигураций суперкомпьютерных систем, БВМКУ и БВМОВС могут быть созданы три типа вычислительных систем соответствующих идеологии двухуровневой архитектуры:

- **кластерные** - системы, состоящие из БВМ КУ и средств объединения вычислительных узлов в кластер (кластерный архитектурный уровень);
- **потоковые** - системы, состоящие из БВМ ОВС и средств их объединения (потоковый архитектурный уровень);
- **смешанные** - универсальные системы, включающие БВМ КУ, БВМ ОВС, средства объединения базовых вычислительных модулей, а также средства поддержки взаимодействия кластерного и потокового уровней—смешанная (двухуровневая) архитектура.

4. Модели семейства суперкомпьютеров

Основные концептуальные принципы создания семейства суперкомпьютеров (открытая масштабируемая двухуровневая архитектура, набор базовых вычислительных модулей и конфигураций и др.) позволяют оптимальным способом создавать для каждой конкретной прикладной проблемы адекватную суперкомпьютерную конфигурацию. В связи с этим фактически отождествляются понятия модель и прикладная конфигурация.

Модели (прикладные конфигурации) идентифицируются в соответствии с ***классификатором семейства суперкомпьютеров***.

Идентификатор модели суперкомпьютера, включающий идентифицирующие признаки, определяет:

- диапазон производительности;
- класс решаемых задач;
- тип архитектуры—потоковая, кластерная, смешанная (двухуровневая);
- типы используемых вычислительных узлов;
- тип базовой конфигурации суперкомпьютерных систем (БКСС);
- тип средств сопряжения вычислительных узлов и т.п.

Учитывая наличие естественного диапазона влияния идентифицирующих признаков, практически каждая конкретная модель суперкомпьютера может включать несколько конкретных прикладных суперкомпьютерных конфигураций (модификации моделей).

Специфика моделей и их модификаций отражается в эксплуатационной документации.

5. Конструкторская документация

Конструкторская документация разрабатывается на базовые модули, имеющие самостоятельную поставку, и на базовые конфигурации суперкомпьютерных систем (БКСС).

Конструкторская документация выполняется в едином для всех исполнении БВМ или БКСС групповом варианте в соответствии с действующими стандартами.

В соответствии с групповым принципом документация содержит постоянные и переменные данные. Постоянные данные—это технические параметры и характеристики, являющиеся общими для всех исполнений каждого типа БВМ или БКСС. Переменные данные отражают отличительные характеристики конкретных исполнений БВМ или конкретных конфигураций, реализуемых на основе данного типа БКСС (то есть отражают специфику моделей суперкомпьютеров и их модификаций).

Конструкторские документы (спецификации, формуляры, паспорта, комплекты монтажных частей, программного обеспечения и т.п.) являются открытыми, то есть содержат разделы, заполняемые для конкретных реализаций.

Технические условия также оформляются по групповому принципу—в них указываются диапазоны изменения параметров и технических характеристик, допускаемые конфигурации, комплектность и т.д. для БВМ и БКСС соответствующего типа, а также предусматривается возможность для указания конкретных значений этих характеристик для конкретных реализаций.

Групповое построение конструкторской документации адекватно отражает возможности архитектурной идеологии (открытость, масштабируемость), позволяя оптимальным способом организовать серийное производство широкой номенклатуры моделей суперкомпьютеров, наиболее полно удовлетворяющих предъявляемым пользовательским требованиям.

6. Проведение испытаний

Суперкомпьютерные конфигурации на различных этапах разработки и производства должны подвергаться ряду проверок и испытаний:

- проверка качества разработки (предварительные, приемочные испытания);
- проверка качества подготовки и освоения серийного производства (квалифицированные, сертификационные испытания);
- проверка стабильности технологического процесса (периодические испытания);
- проверка качества каждой изготовленной прикладной реализации (приемосдаточные испытания).

Предварительные, приемочные, квалификационные, сертификационные и периодические испытания проводятся для каждого типа БВМ, а также для каждого типа базовых конфигураций суперкомпьютерных систем.

На этих испытаниях проверяются параметры базовых исполнений, а также возможность реализации базовых характеристик в диапазонах, указанных в соответствующей КД.

На приемосдаточных испытаниях проверяются параметры конкретной прикладной реализации на соответствие требованиям конкретного пользователя.

Для обеспечения проведения всех типов испытаний должно быть предусмотрено, наряду с базовым (общесистемным) программным обеспечением, соответствующее тестовое программное обеспечение.

7. Организация серийного производства

Объем работ по подготовке серийного производства моделей семейства суперкомпьютеров определяется исходя из принципов настоящей концепции и принятых в процессе реализации Совместной Программы технических решений—количество типов БВМ ОВС, БВМ КУ и БКСС, базовые конструктивы, технология изготовления печатных плат, реализация СБИС для ОВС, соотношение оригинальных и покупных изделий и т.п.

Для обеспечения оперативности исполнения заказа на конкретные суперкомпьютерные изделия предприятие-изготовитель обеспечивает необходимый задел соответствующих конструктивов для изготовления базовых вычислительных модулей и конкретных конфигураций на их основе, а также заключает соответствующие долгосрочные соглашения (рамочные договора) с поставщиками покупных изделий (СБИС, процессорных модулей, модулей сетевого оборудования, периферийных устройств и других необходимых компонентов). Объем необходимого задела элементов собственного изготовления определяется с учетом маркетинговых исследований, соглашений с потенциальными пользователями о намерениях, экономической ситуации и ряда других факторов.

8. Концептуальная технологическая схема реализации прикладных суперкомпьютерных конфигураций

Реализация прикладных суперкомпьютерных конфигураций базируется на следующей концептуальной технологической схеме:

- Поставку конкретной прикладной суперкомпьютерной конфигурации (модели суперкомпьютера) конечному пользователю осуществляет предприятие-поставщик комплексных решений или **системный интегратор**. В конкретных случаях системными интеграторами могут быть предприятия-разработчик или изготовитель суперкомпьютерных конфигураций.
- Системный интегратор с учетом основополагающих концептуальных архитектурных принципов создания семейства суперкомпьютеров анализирует структуру конкретной прикладной проблемы и оценивает удельный вес фрагментов (задач) с крупноблочным динамическим параллелизмом и с мелкозернистым (явным) параллелизмом.
- На основании структурного анализа прикладной проблемы с учетом требуемого диапазона производительности и других, специфических для прикладной проблемы факторов, системный интегратор определяет необходимую конфигурацию прикладного комплекса (количество и типы базовых вычислительных модулей, тип базовой конфигурации суперкомпьютерной системы, периферийное оборудование и т.п.), а также структуру базового (общесистемного) программного обеспечения и необходимые для решения проблемы приложения.
- Завод-изготовитель суперкомпьютерных систем на основании заказа системного интегратора изготавливает в соответствии с предоставленной спецификацией требуемую прикладную суперкомпьютерную конфигурацию (модель суперкомпьютера), проводит ее приемосдаточные испытания и поставляет системному интегратору с комплектом эксплуатационной документации, отражающим специфику модели.

- Системный интегратор разрабатывает прикладное программное обеспечение (своими силами или силами контрагентов), обеспечивает выполнение прикладных задач на модели, разрабатывает необходимую эксплуатационную документацию и сдает суперкомпьютерную систему конечному пользователю в виде комплексного решения ("под ключ").
- Системный интегратор обеспечивает гарантийное и послегарантийное техническое обслуживание суперкомпьютерной системы в процессе ее эксплуатации у пользователя, например, привлекая завод-изготовитель или другие специализированные предприятия.

9. Обобщенная система требований по созданию моделей семейства суперкомпьютеров

Система требований по созданию моделей семейства суперкомпьютеров базируется на следующих основных регламентирующих документах:

- Совместная белорусско-российская программа "Разработка и освоение в серийном производстве семейства моделей высокопроизводительных вычислительных систем с параллельной архитектурой (суперкомпьютеров) и создание прикладных программно-аппаратных комплексов на их основе".
- Концепция создания моделей семейства суперкомпьютеров.
- Общее техническое задание (ОТЗ) на создание моделей семейства суперкомпьютеров.
- Общие технические требования (ОТТ) на создание моделей семейства суперкомпьютеров.
- Технические задания (ТЗ) на реализацию конкретных программных мероприятий (проектов) Совместной Программы.

Совместная Программа определяет:

- этапность выполнения работ;
- основных исполнителей программных мероприятий (в том числе, координаторов Совместной Программы, головного исполнителя и исполнителя Совместной Программы от РБ и РФ);
- перечень и наименование проектов Совместной Программы;
- ориентировочные сроки выполнения и ориентировочные объемы финансирования проектов.

Точные сроки выполнения и объемы финансирования определяются соответствующими договорными соглашениями по конкретным проектам.

Концепция определяет общие для всего семейства принципы создания суперкомпьютерных систем на всех этапах технологического процесса—от базовых концептуальных архитектурных принципов до процессов разработки, освоения в серийном производстве и поставки конечному пользователю конкретных прикладных суперкомпьютерных конфигураций.

В общем техническом задании на создание моделей семейства суперкомпьютеров в сжатой форме сформулированы общие технические принципы создания семейства суперкомпьютеров, базирующиеся на положениях Концепции.

ОТЗ разрабатывается для каждого этапа реализации Совместной Программы—ОТЗ на создание моделей первого ряда семейства суперкомпьютеров (I этап реализации Совместной Программы) и ОТЗ на создание моделей второго ряда семейства суперкомпьютеров (II этап реализации Совместной Программы).

В ОТТ сформулированы общие технические требования к моделям семейства суперкомпьютеров, базирующиеся на положениях ОТЗ.

ОТТ также разрабатываются для каждого этапа реализации Совместной Программы—ОТТ к моделям первого ряда семейства суперкомпьютеров и ОТТ к моделям второго ряда семейства суперкомпьютеров.

Разделение общих технических принципов и требований на два документа (ОТЗ и ОТТ) отражает объективную целесообразность разработки ОТТ после согласования ОТЗ из-за разных (по очевидным причинам) сроков отработки концептуальных принципов и технических требований.

Технические задания на проведение работ по реализации конкретных проектов разрабатываются с учетом основных положений Совместной Программы, Концепции, ОТЗ и ОТТ и отражают специфику конкретных проектов.

Заключение

Реализация базовых принципов Концепции позволяет создать перспективные суперкомпьютерные системы, соответствующие, а по некоторым позициям (в частности, по архитектурным принципам и способам их реализации, стоимостным и другим показателям), превосходящие современный мировой уровень суперкомпьютерной отрасли.

Создаваемые на базе Концепции модели семейства суперкомпьютеров позволяют перекрыть широкий диапазон производительность и областей применения - от класса высокопроизводительных серверов и мультипроцессорных рабочих станций (миллиарды операций в секунду) до вычислительных систем с массовым параллелизмом сверхвысокой производительности (триллионы операций в секунду).

Концептуальные принципы создания моделей семейства суперкомпьютеров, в частности, возможность создания суперкомпьютерных систем на основе только кластерной или потоковой архитектурных компонент, позволяют уже на первом этапе реализации Совместной Программы освоить выпуск моделей суперкомпьютеров среднего класса (10-100 ГФлопс).

На базе существующего научно-технического задела реально освоение в производстве суперкомпьютеров с кластерной архитектурой с пиковыми производительностями, например, 10, 25, 50 и 100 ГФлопс, что соответствует по вычислительной мощности суперкомпьютерам SUN 4000, SGI Origin 2000/195, HP SPP-2000, Cray T3 E-900; Cray T3 E-1200 с числом процессоров от 8 до 256.

Выпуск таких моделей позволит закрыть широкие потребности в моделях серверов и суперкомпьютеров среднего класса и подготовить базу для создания суперкомпьютерных систем сверхвысокой производительности (100-10,000 Гфлопс и выше).

На основе типовых БВМ ОВС могут быть созданы встраиваемые и/или бортовые вычислительные системы, предназначенные для управления сложными системами, что значительно расширяет сферу применения данных изделий.

Основополагающие принципы Концепции позволяют создавать на их основе прикладные системы, оптимально соответствующие требованиям конкретного заказчика, и оптимально использовать производственные мощности предприятия-изготовителя с учетом специфики рынка сбыта высокопроизводительных вычислительных систем.