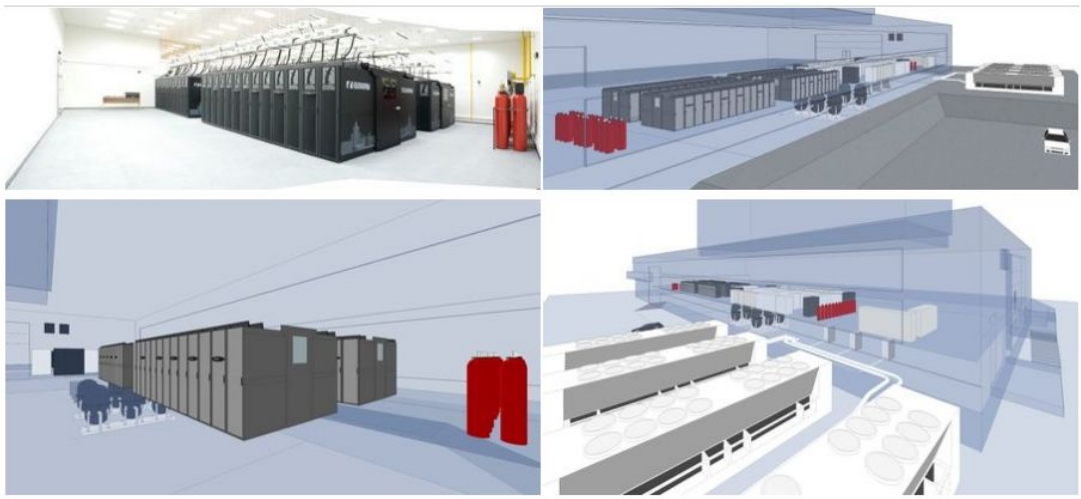
Программно-аппаратная архитектура Ломоносов и BlueGene\P

Суперкомпьютер «Ломоносов», 2015 год (из лекций)

Пиковая производительность	1700.21 TFlop/s
Производительность (Linpack)	901.90 TFlop/s
Эффективность	53%
Вычислительных узлов (Intel)	5 104
Вычислительных узлов (ГПУ)	1 065
Процессоры Intel Xeon 5570, 5670	12 346
NVIDIA Tesla X2070	2 130
Число процессорных ядер (x86)	52 168
Число процессорных ядер (ГПУ)	954 240
Оперативная память	92 ТБайт
Коммуникационная сеть	QDR Infiniband / 10 GE
Система хранения данных	1.75 ПБайт, Lustre, NFS, ...
Операционная система	Clusrtx T-Platforms Edition
Занимаемая площадь (вычислитель)	252 м2
Энергопотребление (вычислитель)	2.8 МВт

Суперкомпьютерный комплекс "Ломоносов"

Общий вид комплекса:



Суперкомпьютер "ЛОМОНОСОВ"

Суперкомпьютер «Ломоносов» — первый **гибридный суперкомпьютер** такого масштаба в России и Восточной Европе. В нём используется 3 вида вычислительных узлов и процессоры с различной архитектурой. В качестве основных узлов, обеспечивающих свыше 90 % производительности системы, используется blade-платформаT-Blade2. Предполагается использовать суперкомпьютер для решения ресурсоёмких вычислительных задач в рамках фундаментальных научных исследований, а также для проведения научной работы в области разработки алгоритмов и программного обеспечения для мощных вычислительных систем.

Последние несколько лет рост интереса к суперкомпьютерной тематике связан с выходом на рынок так называемых **гибридных суперкомпьютеров. То есть суперкомпьютеров, которые наряду с центральным процессором традиционной архитектуры используют для вычислений специализированные процессоры, в частности, графические.**

Общая характеристика

Основные технические характеристики суперкомпьютера "Ломоносов"	
Пиковая производительность	1,7 Пфлопс
Производительность на тесте Linpack	901.9 Тфлопс
Число вычислительных узлов x86	5 104
Число графических вычислительных узлов	1 065
Число вычислительных узлов PowerXCell	30
Число процессоров/ядер x86	12 346 / 52 168
Число графических ядер	954 240
Оперативная память	92 ТБ
Общий объем дисковой памяти вычислителя	1,75 ПБ
Основной тип процессора	Intel Xeon X5570/Intel Xeon 5670, Nvidia X2070
Число типов вычислительных узлов	8
Основной тип вычислительных узлов	ТВ2-XN
System/Servise/Management Network	QDR Infiniband 4x/10G Ethernet/Gigabit Ethernet

Система хранения данных	Параллельная файловая система Lustre, файловая система NFS, иерархическая файловая система StorNext, система резервного копирования и архивирования данных
Операционная система	Clustrx T-Platforms Edition
Занимаемая площадь	252 м^2
Потребление энергии	2,6 МВт
Вес всех составляющих	Более 75 тонн
Производитель	Т-Платформы(link is external)

Площади помещений:

- Вычислитель: 252 кв. м
- СБЭ (система бесперебойного электропитания): 246 кв.м.
- ГРЩ (главный распределительный щит): 85 кв. м.
- Климатическая система: 216 кв. м.

Энергопотребление:

- Пиковая мощность вычислителя (1,7 Tflops): 2,6 МВт
- Средняя мощность инфраструктуры: 740 КВт.
- Пиковая мощность инфраструктуры при внешней температуре 35 цельсия: 1,2 МВт
- Средняя суммарная мощность комплекса: 2,57 МВт
- Пиковая суммарная мощность комплекса (при 35 цельсия): 3,05 МВт.

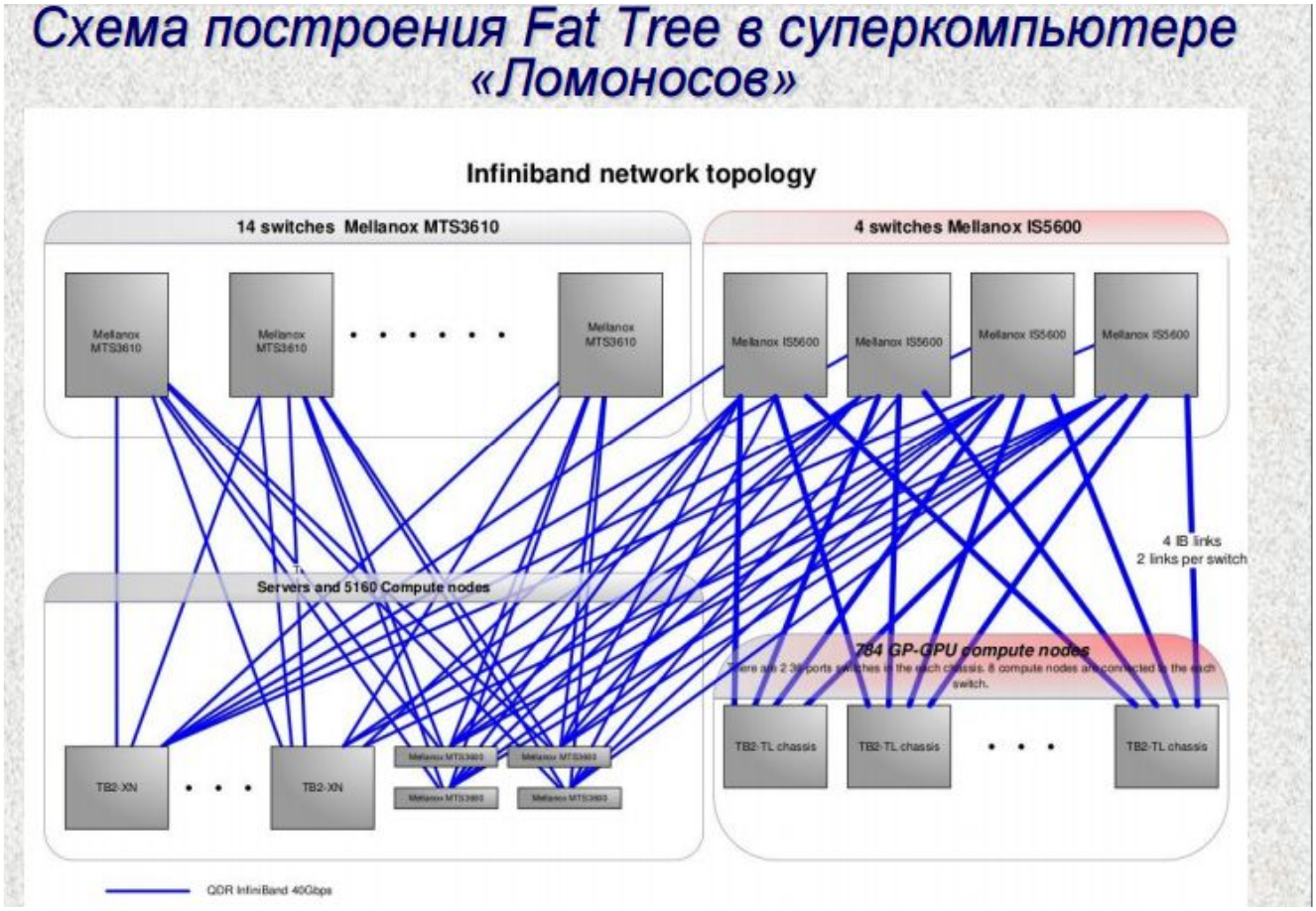
Вычислительные узлы и сети Группы вычислительных узлов:

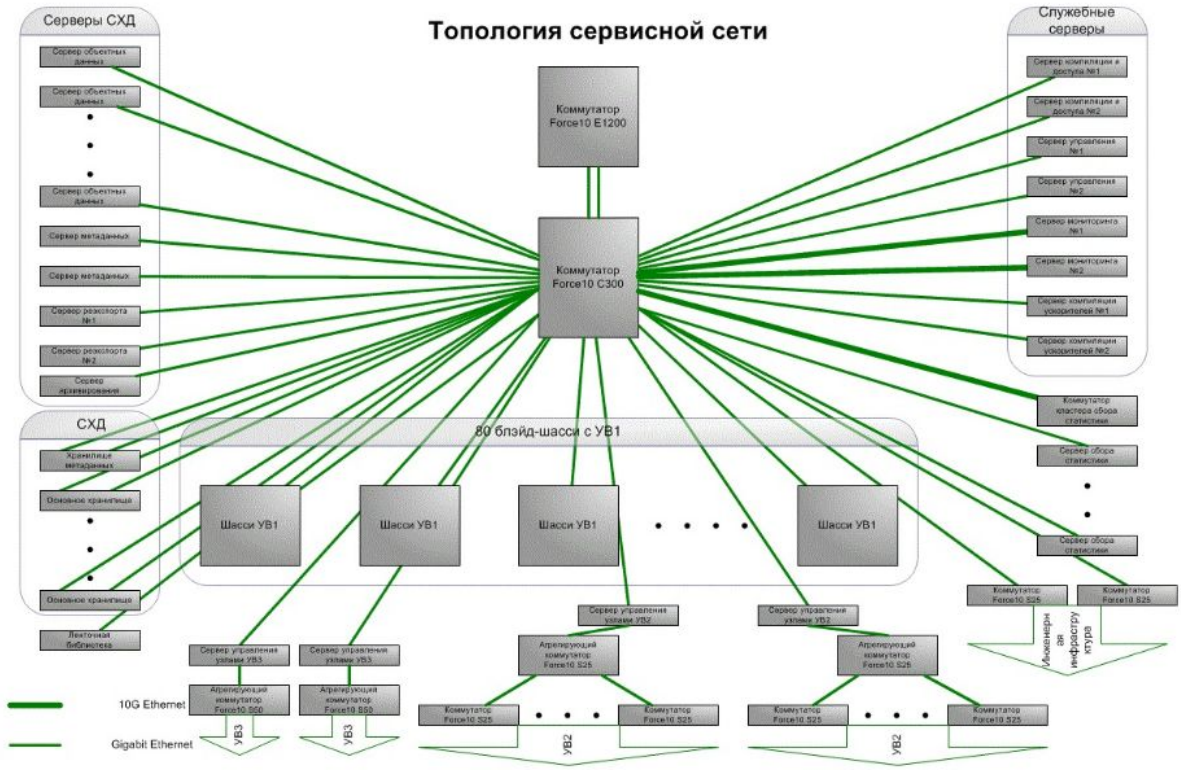
Тип	**Процессоры**	**Кол-во ядер**	**Опер. память, ГБ**	**Сумм. кол-во процес.**	**Сумм. кол-во ядер**	**Кол-во узлов**
T-Blade2(link is external)(УВ1)	2 x Intel® Xeon 5570 Nehalem	2 x 4	12	8 320	33 280	4 160
T-Blade1(УВ2)	2 x Intel® Xeon 5570 Nehalem	2 x 4	24	520	2 080	260
T-Blade2(link is external)(УВ1)	2 x Intel® Xeon 5670 Westmere	2 x 6	24	1 280	7 680	640
T-Blade1(УВ2)	2 x Intel® Xeon 5670 Westmere	2 x 6	48	80	480	40
Узлы на базе IBM® Cell (УВ3)	PowerXCell 8i	8	16	60	480	30

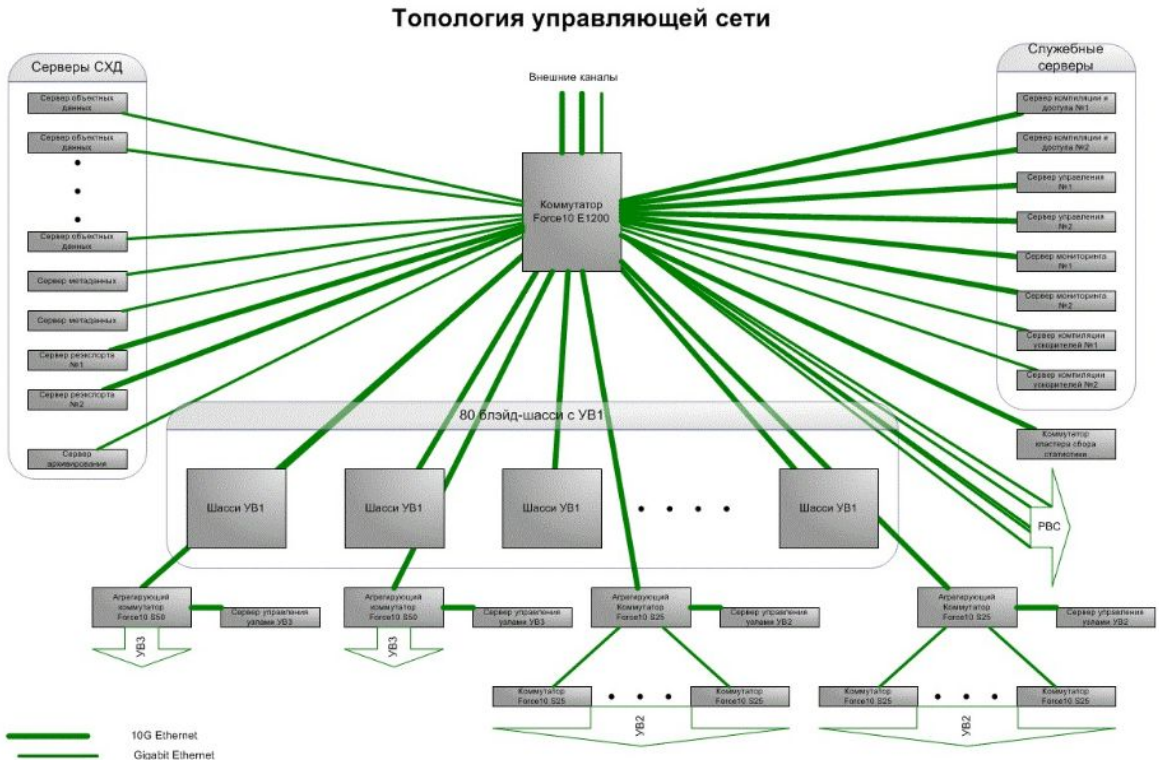
Все узлы связаны тремя независимыми сетями:

- Системная сеть - QDR InfiniBand, 40 Гбит/сек (схема)
- Сервисная сеть - Ethernet, 10 Гбит/сек, 1 Гбит/сек и 100 Мбит/сек (схема)
- Управляющая сеть - Ethernet, 10 Гбит/сек и 1 Гбит/сек (схема)
- Сеть барьерной синхронизации и сеть глобальных прерываний, Т-Платформы

Сеть **fat tree** (рус. *утолщенное дерево*) — топология компьютерной сети, изобретённая Чарльзом Лейзерсоном из MIT, является дешевой и эффективной для суперкомпьютеров[1]. В отличие от классической топологии дерево, в которой все связи между узлами одинаковы, связи в утолщенном дереве становятся более широкими (толстыми, производительными по пропускной способности) с каждым уровнем по мере приближения к корню дерева. Часто используют удвоение пропускной способности на каждом уровне.







Программное обеспечение

- Средства архивации данных: bacula 3 (Т-Платформы), StorNext (Quantum), NetBackup (Symantec)
- Передача файлов: SCP, SFTP
- Управление заданиями и ресурсами: SLURM 2.0
- Среды исполнения: OpenMPI 1.4, MVAPICH 1.1, IntelMPI 4

- Языки программирования: C/C++, Fortran 77/90/95
- Наборы компиляторов: Intel 12, GNU 4.4, Pathscale, PGI
- Средства отладки и анализа производительности: Intel® ITAC 12, grpof 4, Intel® vTune 4, Intel® Thread Checker, Acumem ThreadSpotter, IDB, Allinea DDT
- Системы контроля версий: SVN, GIT
- Языки сценариев: Perl, Python

HPC-приложения MOLPRO, версия 2010.1 (установлено в /opt/molpro2010.1/) - доступен ТОЛЬКО сотрудникам МГУ.

Система хранения данных

Что снижает производительность компьютеров с распределенной памятью?

1. Закон Амдала
2. Латентность передачи по сети
3. Пропускная способность каналов передачи данных
4. Особенности использования SMP-узлов
5. Балансировка вычислительной нагрузки
6. Возможность асинхронного счета и передачи данных
7. Особенности топологии коммуникационной сети
8. Производительность отдельных процессоров
9. ...

http://users.parallel.ru/wiki/pages/22-config

Конфигурация суперкомпьютеров

Ломоносов-2

Раздел	узлов/x86 ядер (x86 ядер на узел)	GPU-карт на узле	память	диски	max ядро-часов	max/default время (часов на задачу)	max задач	max запущенных задач	max ядер на задачу, не более
compute .	1024 / 14336 (14)	1	64 ГБ (4,5 ГБ/ядро)	нет	нет	72/24	3	3	нет

- Объём памяти на GPU: 11.56 GB
- Модель GPU: Tesla K40s
- Модель CPU: Intel Xeon E5-2697 v3 2.60GHz

Ломоносов

Раздел	узлов/x86 ядер (x86 ядер на узел)	GPU-карт на узле	память	диски	max ядро-часов	max/default время (часов на задачу)	max задач	max запущенных задач	max ядер на задачу, не более
regular4 .	4096 / 32768 (8)	0	12 ГБ (1,5 ГБ/ядро)	нет	нет	72/24	3	3	1024
regular6 .	596 / 7152 (12)	0	12 ГБ (1 ГБ/ядро)	нет	нет	72/24	3	3	512
hdd4	260 / 2080 (8)	0	12 ГБ (1,5 ГБ/ядро)	есть	нет	72/24	2	2	256
hdd6	32 / 384 (12)	0	48 ГБ (4 ГБ/ядро)	есть	нет	72/24	2	2	128
gpu	830 / 6640 (8)	2	24 ГБ (3 ГБ/ядро)	нет	нет	72/24	3	3	256
smp	1 / 128 (128)	0	2ТБ (16 ГБ/ядро)	есть	нет	72/24	3	2	128
test	64 / 512 (8)	0	12 ГБ (1,5 ГБ/ядро)	нет	нет	0,25 (15 минут)/0,25	3	1	128
gputest	16 / 128 (8)	2	24 ГБ (3 ГБ/ядро)	нет	64	0,25 (15 минут)/0,25	3	1	64

Очередь по умолчанию - **regular4** “max задач” включает в себя и задачи на счёте и в очереди.

- Объём памяти на GPU: 5.25 GB
- Модель GPU: Tesla X2070
- Модель CPU:
  - regular4, hdd4: Intel Xeon X5570 2.93GHz
  - gpu: Intel Xeon E5630 2.53GHz
  - regular6, hdd6: Intel Xeon X5670 2.93GHz

Описание вычислительного комплекса IBM Blue Gene/P

IBM Blue Gene/P — массивно-параллельная вычислительная система, которая состоит из двух стоек, включающих **8192 процессорных ядер** (2 x 1024 четырехъядерных вычислительных узлов), **с пиковой производительностью 27,9 терафлопс** (27,8528 триллионов операций с плавающей точкой в секунду).

Массивно-параллельные системы с распределенной памятью
• Высокая плотность упаковки
– процессоры с низким энергопотреблением (40 W ~лампочка)
• Высокопроизводительный интерконект
– несколько комутационных подсистем для различных целей
• Ультра легкая ОС
– выполнение вычислений и ничего лишнего
• Стандартное ПО Standard software
– Fortran/C/C++ и MPI

- 2048 4-ех ядерных узлов
- пиковая производительность 27.2 Tflop/s
- Реальная производительность по тесту Linpack: 23.2 Тфлоп/с
- 85% от пиковой
- общий объем ОЗУ 4 ТВ

Характеристики системы:

- две стойки с вычислительными узлами и узлами ввода-вывода
- 1024 четырехъядерных вычислительных узла в каждой из стоек
- 16 узлов ввода-вывода в стойке (в текущей конфигурации активны 8, т.е. одна I/O-карта на 128 вычислительных узлов)
- выделенные коммуникационные сети для межпроцессорных обменов и глобальных операций
- программирование с использованием MPI, OpenMP/pthreads, POSIX I/O
- высокая энергоэффективность: **~ 372 MFlops/W** (см. список Green500)
- система воздушного охлаждения

1 стойка (rack, cabinet) состоит из двух midplane'ов.
В midplane входит 16 node-карт (compute node card),
на каждой из которых установлено 32 вычислительных узла (compute card).
Midplane, 8 x 8 x 8 = 512 вычислительных узлов,
— минимальный раздел, на котором становится доступна топология трехмерного тора;
для разделов меньших размеров используется топология трехмерной решетки.
Node-карта может содержать до двух узлов ввода-вывода (I/O card).
Вычислительный узел включает в себя четырехъядерный процессор, 2 ГБ общей памяти и сетевые интерфейсы.

- 1024 четырехъядерных вычислительных узлов
- производительность одного вычислительного узла – 13.6 GF/s
- производительность 1 стойки– 13.9 Tflops
- оперативная память одного узла – 2 GB
- суммарная оперативная память в стойке– 2 TB
- узлов ввода/вывода 8 – 64
- Размеры - 1.22 x 0.96 x 1.96
- занимаемая площадь 1.17 кв.м.
- энергопотребление (1 стойка) - 40 kW (max)

Микропроцессорное ядро:

- модель: PowerPC 450
- рабочая частота: 850 MHz
- адресация: 32-битная
- кэш инструкций 1-го уровня (L1 instruction): 32 KB
- кэш данных 1-го уровня (L1 data): 32 KB
- кэш предвыборки (L2 prefetch): 14 потоков предварительной выборки (stream prefetching): 14 x 256 байтов
- два блока 64-битной арифметики с плавающей точкой (Floating Point Unit, FPU), каждый из которых может выдавать за один такт результат совмещенной операции умножения-сложения (Fused Multiply-Add, FMA)
- пиковая производительность: 2 FPU x 2 FMA x 850 MHz = 3,4 GFlop/sec per core

Вычислительные узлы и I/O-карты в аппаратном смысле неразличимы и являются взаимозаменяемыми, разница между ними состоит лишь в способе их использования. У них нет локальной файловой системы, поэтому все операции ввода-вывода перенаправляются внешним устройствам.

Вычислительной узел:

- четыре микропроцессорных ядра PowerPC 450 (4-way SMP)
- пиковая производительность: 4 cores x 3,4 GFlop/sec per core = 13,6 GFlop/sec
- пропускная способность памяти: 13,6 GB/sec
- 2 ГБ общей памяти
- 2 x 4 МБ кэш-памяти 2-го уровня (в документации по BG/P носит название L3)
- легковесное ядро (compute node kernel, CNK), представляющее собой Linux-подобную операционную систему, поддерживающую значительное подмножество Linux-совместимых системных вызовов
    - Создание процессов и управление ими
    - Управление памятью
    - Отладка процессов
    - Ввод-вывод
- асинхронные операции межпроцессорных обменов (выполняются параллельно с вычислениями)
- операции ввода-вывода перенаправляются I/O-картам через сеть коллективных операций
- Двойное устройство для работы с вещественными числами с плавающей точкой (double precision)
- Объем виртуальной памяти равен объему физической

Узел ввода-вывода:

- не учитывается при расчете пиковой производительности
- использует сеть коллективных операций для коммуникаций с вычислительными узлами
- подключен к внешним устройствам через Ethernet-порт посредством 10-гигабитный функциональной сети
- операционная система на основе Linux (Mini-Control Program, MCP) с минимальным набором пакетов, необходимых для поддержки клиента сетевой файловой системы и Ethernet-подключений

Коммуникационные сети:

- **трехмерный тор (three-dimensional torus)**
  - сеть общего назначения, объединяющие все вычислительные узлы; предназначена для операций типа «точка-точка»
  - вычислительный узел имеет двунаправленные связи с шестью соседями
  - пропускная способность каждого соединения — 425 MB/s (5,1 GB/s для всех 12 каналов)
  - латентность (ближайший сосед):
    - 32-байтный пакет: 0,1 μs
    - 256-байтный пакет: 0,8 μs
- **глобальные коллективные операции (global collective)**
  - коммуникации типа «один-ко-многим» (broadcast-операции и редукция)
  - используется вычислительными узлами для обменов с I/O-картами
  - каждый вычислительный узел и I/O-карта имеют три двунаправленные связи
  - пропускная способность каждого соединения — 850 MB/s (1,7 GB/s для двух каналов)
  - латентность (полный обход): 3,0 μs
- **глобальные прерывания (global interrupt)**
  - операции барьеров и прерываний (глобальные AND- и OR-операции)
- **функциональная сеть**
  - соединяет узлы ввода-вывода с внешним окружением
  - 10-гигабитная оптическая Ethernet-сеть
- **сервисная сеть (service/control)**
  - загрузка, мониторинг, диагностика, отладка, доступ к счетчикам производительности
  - гигабитная Ethernet-сеть (4 соединения на стойку)

Чтобы разгрузить процессорное ядро от операций, связанных с передачей сообщений по сети трехмерного тора, используется устройство прямого доступа к памяти (direct memory access, DMA). Кроме уменьшения нагрузки на ядро, этот механизм уменьшает вероятность взаимной блокировки процессов, обменивающихся сообщениями, которая может возникнуть вследствие ошибок программиста.

Окружение Blue Gene/P включает

- **фронтэнд (front end node)** — система, открытая для доступа по протоколу SSH; служит для доступа пользователей на вычислительный комплекс; вся связь с комплексом осуществляется только через эту машину; предназначена для разработки пользователями программ, компилирования проектов и постановки задач в очередь; работа с ней осуществляется в интерактивном режиме
- **сервисный узел (service node)** — обеспечивает контроль над системой Blue Gene/P; к этой машине доступа по SSH нет

- **систему управления высокопроизводительной параллельной файловой системой IBM General Parallel File System (GPFS)**

Назначение	**Модель**	**Процессоры**	**Количество**
фронтэнд	IBM pSeries 55A	POWER5+	2
сервисный узел			2
GPFS-серверы	IBM x3650	Intel Xeon	16

Для коммутации оптических линий служит высокопроизводительный свитч IBM Ethernet Switch B08R со 112 портами 10 Gb Ethernet: 64 порта используются для подключения узлов ввода-вывода вычислительной системы Blue Gene/P, 32 порта служат для подключения GPFS-серверов, управляющих параллельной файловой системой, к четырем портам подключены фронтэнды и сервисные узлы, остальные 12 портов используются для инфраструктурных нужд, либо зарезервированы для будущего использования. Гигабитный Ethernet скоммутирован на четыре 48-портовых свитча Cisco.

Для организации файлового хранилища используется 16 дисковых систем DS3512, каждая из которых включает по две дополнительных дисковых полки EXP3512. В основу сети хранения данных положены два 80-портовых коммутатора IBM System Storage SAN80B-4 Fibre Channel.

ОС вычислительного узла BlueGene P

• Compute Node Kernel (CNK)

– "linux-подобная" ОС

– Нет некоторых системных вызово (fork() в основном). Ограниченная поддержка mmap(), execve().

– Минимальное ядро – обработка сигналов, передача системных вызовов к узлам ввода-вывода, старт-завершение задач, поддержка нитей

– Большинство приложений, которые работают под Linux, портируются на BG/P

Компиляторы Blue Gene

• IBM XL компиляторы (xlc, xlf77, xlf90)

• работают на front end узлах

– Fortran: mpixlf, mpixlf90, mpixlf95

– C: mpixlc

– C++: mpixlcxx

• обычно являются скриптами

• GNU компиляторы существуют, но малоэффективны: mpicc