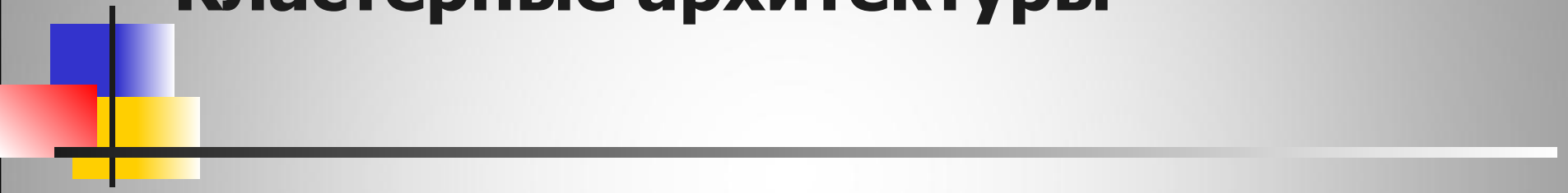
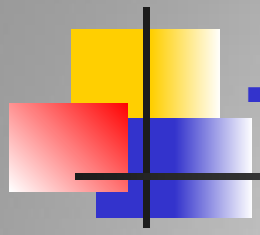
Кластерные архитектуры



**концепция, особенности организации,
примеры разработок**

Введение



- Кластер - это группа вычислительных машин, которые связаны между собою и функционируют как один узел обработки информации.

- Самым важным при построении кластерной системы является то ,что для пользователя или прикладной задачи вся совокупность вычислительной техники выглядит как один компьютер.

- Узлы кластера - серверы, рабочие станции и даже обычные персональные компьютеры.

- Преимущество кластеризации для повышения работоспособности становится очевидным в случае сбоя какого-либо узла: при этом другой узел кластера может взять на себя нагрузку неисправного узла, и пользователи не заметят прерывания в доступе.

- Впервые в классификации вычислительных систем термин "кластер” определила DEC (Digital Equipment Corporation) 1983 г.
- До середины 80-х годов под технологиями суперкомпьютеров понимали исключительно создание особо мощных процессоров.
- Первые кластеры компании Digital были построены на машинах VAX. Эти машины уже не производятся, но все еще работают на площадках, где были установлены много лет назад.
- Общие принципы, заложенные при их проектировании, остаются основой при построении кластерных систем и сегодня.
- В 1998 году в Лос-Аламосской национальной лаборатории астрофизик Майкл Уоррен с коллегами построили суперкомпьютер Avalon, который представлял собой Linux-кластер на базе процессоров Alpha 21164A с тактовой частотой 533 МГц.
- Первоначально Avalon состоял из 68 процессоров, затем был расширен до 140. В каждом узле установлено по 256 Мбайт оперативной памяти, жесткий диск на 3 Гбайт и сетевой адаптер Fast Ethernet. Общая стоимость проекта Avalon составила 313 тыс. долл.
- Создание кластеров на основе дешёвых персональных компьютеров, объединённых сетью передачи данных, продолжилось в 1993 г. силами Американского аэрокосмического агентства (NASA), затем в 1995 г. получили развитие кластеры Beowulf, специально разработанные на основе этого принципа. Успехи таких систем подтолкнули развитие grid-сетей.

- **Высокая готовность**

- В случае сбоя программного обеспечения на одном узле приложение продолжает функционировать (либо автоматически перезапускается) на других узлах кластера;

- Сбой или отказ узла (или узлов) кластера по любой причине (включая ошибки персонала) не означает выхода из строя кластера в целом;

- **Масштабирование**

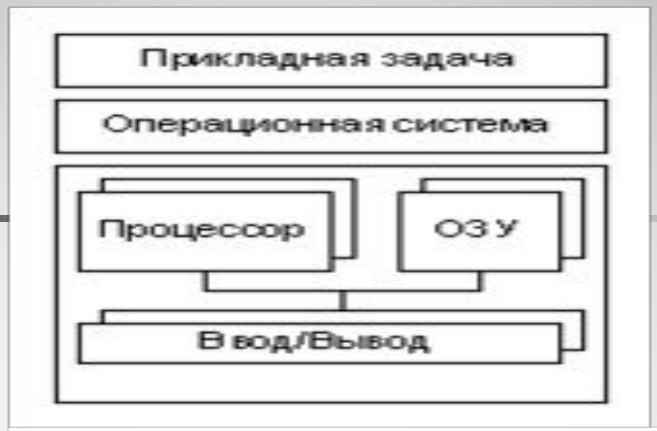
- Вертикальное (увеличение числа процессоров в многопроцессорных системах или добавление новых адаптеров или дисков)

- Горизонтальное (предоставляет возможность добавлять в систему дополнительные компьютеры и распределять работу между ними)

- **Высокое быстродействие**
- **Общий доступ к ресурсам**
- **Удобство обслуживания**

аппаратных ресурсов

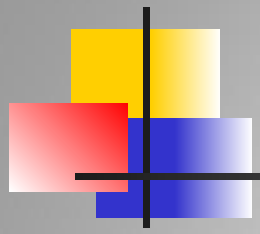
Тесно связанная мультипроцессорная система



Умеренно связанная мультипроцессорная система

Слабо связанная мультипроцессорная система

Кластеры



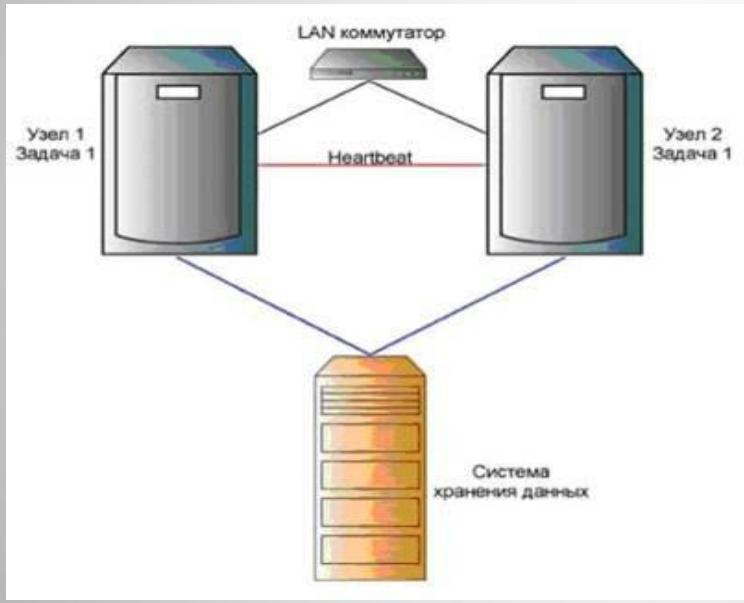
- Высокоскоростные (High Performance, HP)
- Ситемы высокой готовности(High Availability, HA)
- Смешанные системы

- Высокоскоростные кластеры используются для задач, которые требуют значительной вычислительной мощности.
- Классическими областями, в которых используются подобные системы, являются:
  - обработка изображений
  - научные исследования
  - промышленность
- Системы высокой готовности используются везде, где стоимость возможного простоя превышает стоимость затрат, необходимых для построения кластерной системы, например:
  - биллинговые системы
  - банковские операции
  - электронная коммерция
  - управление предприятием

- Все типы систем высокой готовности имеют общую цель - минимизацию времени простоя.
- Стоимость систем высокой готовности намного превышает стоимость обычных систем.

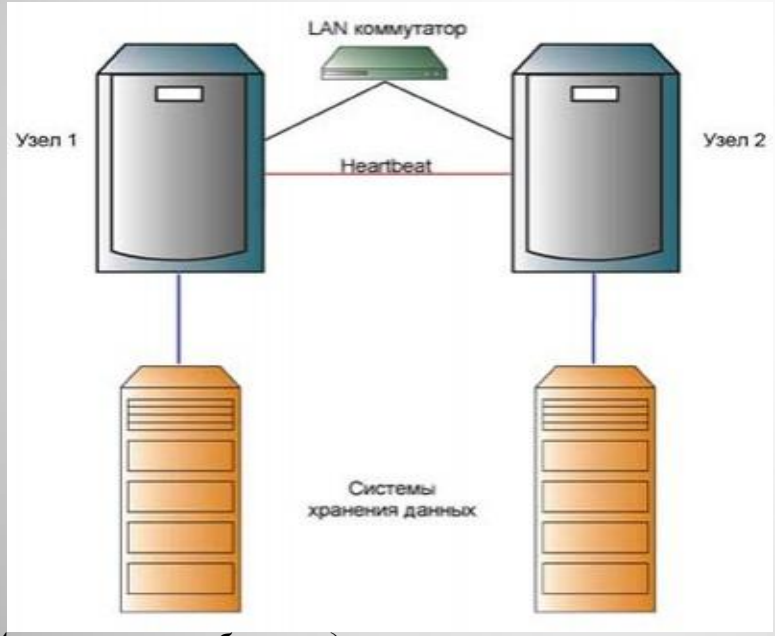
- **Shared Disk Architecture**
- Архитектура с общими дисками ,классически используется для построения кластерных систем высокой готовности, ориентированных на обработку больших объемов данных.
- **Особенности :**
- На компьютерах, входящих в состав кластера, запускается одно приложение, хотя это условие не является обязательным .
- Сервер приложений должен координировать доступ к дискам. Для этого в локальной сети между серверами должны быть установлены постоянные соединения
- **Недостатки :**
- Низкая степень масштабируемости
- Невозможность совместного использования ресурсов, расположенных на больших расстояниях.
- Усложняется управление удаленными данными.

- Система с общими дисками может оказаться эффективным решением задачи, если в задаче удается логически разделить данные для того, чтобы запрос из некого подмножества можно было бы обработать с использованиям части данных .
- Производительность может увеличиваться как путем наращивания числа процессоров и объемов оперативной памяти в каждом узле кластера, так и посредством увеличения количества самих узлов.



Архитектура с общими дисками

- **Shared Nothing Architecture**
- Архитектура без разделения ресурсов, в которой каждый узел системы имеет собственную оперативную память и собственные диски, которые не разделяются между отдельными узлами системы. В таких системах разделяется только общий коммуникационный канал между узлами системы.
- Производительность систем может увеличиваться путем добавления процессоров, объемов оперативной и внешней (дисковой) памяти в каждом узле, а также путем наращивания количества таких узлов.



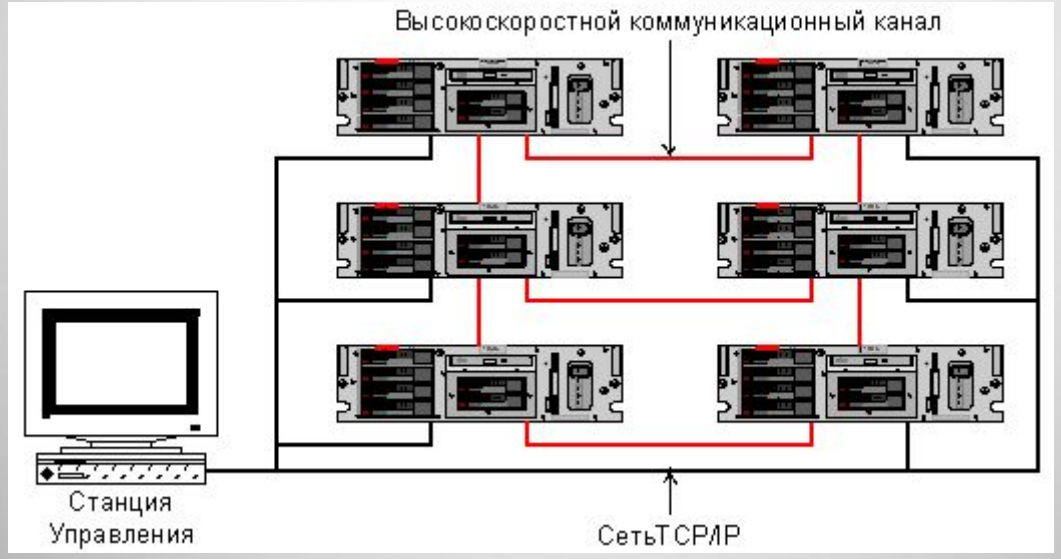
Архитектура без разделения ресурсов

Кластеры для высокопроизводительных вычислений предназначены для параллельных расчётов.

Быстродействие High Performance кластерной системы определяется быстродействием узлов и связей между ними.

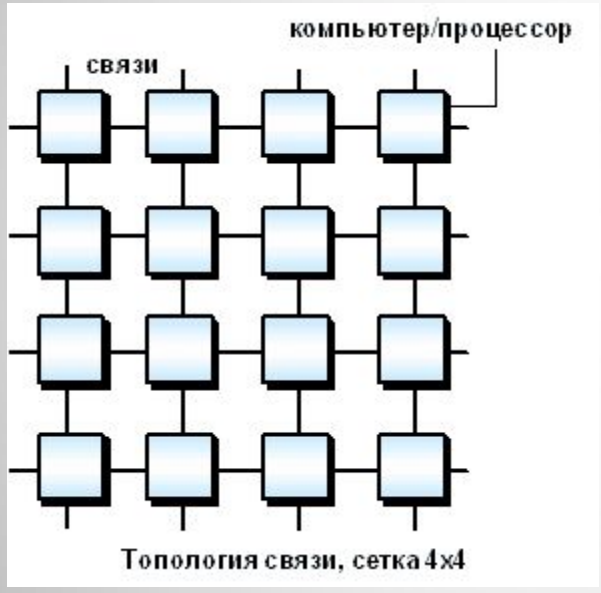
Основная проблема высокоскоростного кластера - невозможность избегать необходимости передавать данные от одной подзадачи другой в ориентированной на параллельное вычисление задаче .

При проектировании высокоскоростных кластерных систем и расчета их быстродействия, следует учитывать потери быстродействия, связанные с обработкой и передачей данных в узлах кластера, т.к. быстродействие передачи данных между центральным процессором и оперативной памятью узла значительно превышает скоростные характеристики систем межкомпьютерного взаимодействия.



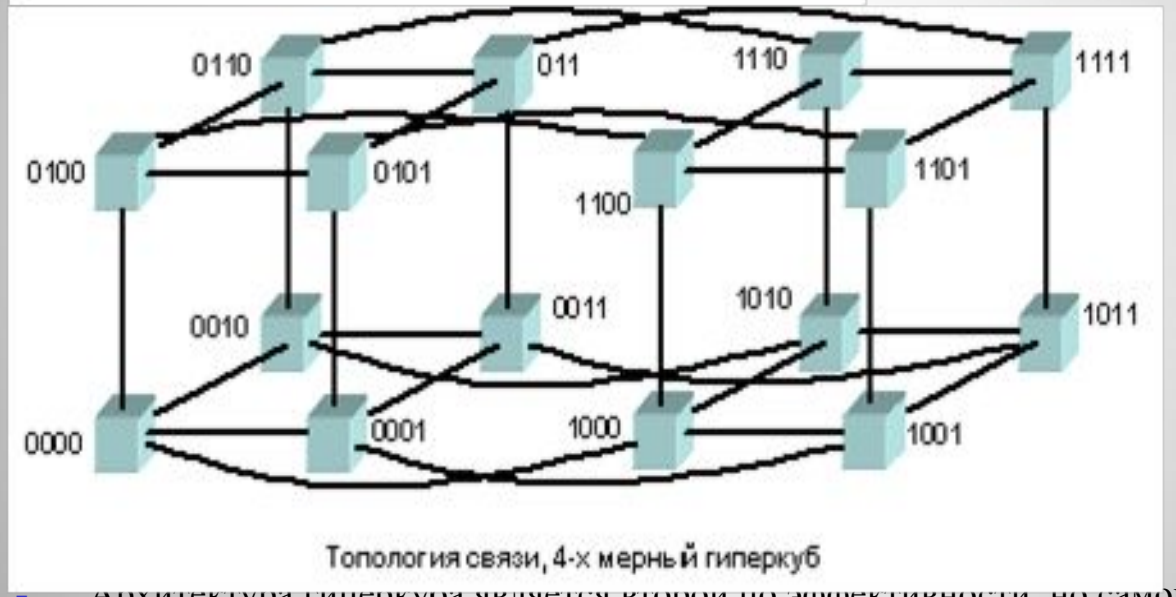
- Архитектура кластерной системы (способ соединения процессоров друг с другом) определяет ее производительность, чем тип используемых в ней процессоров. Критическим параметром, влияющим на величину производительности такой системы, является расстояние между процессорами.

- Схема соединения процессоров в виде плоской решетки



- Примеры гиперкубов

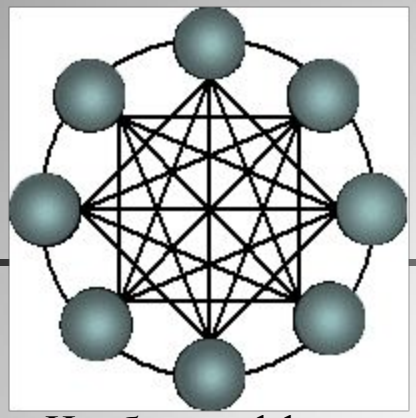
Топология связи, 3-х мерный гиперкуб



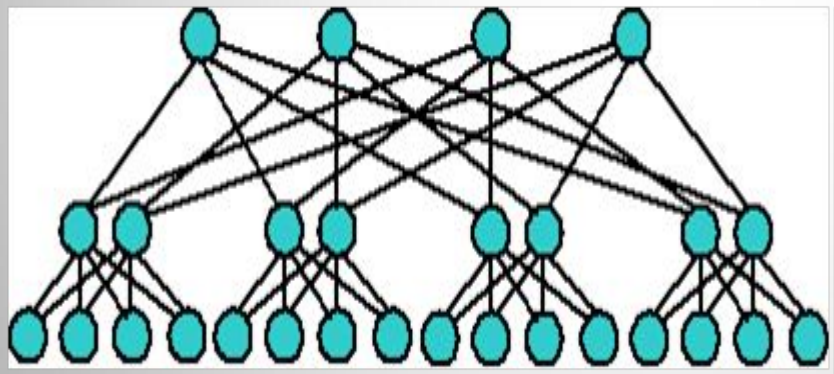
Топология связи, 4-х мерный гиперкуб

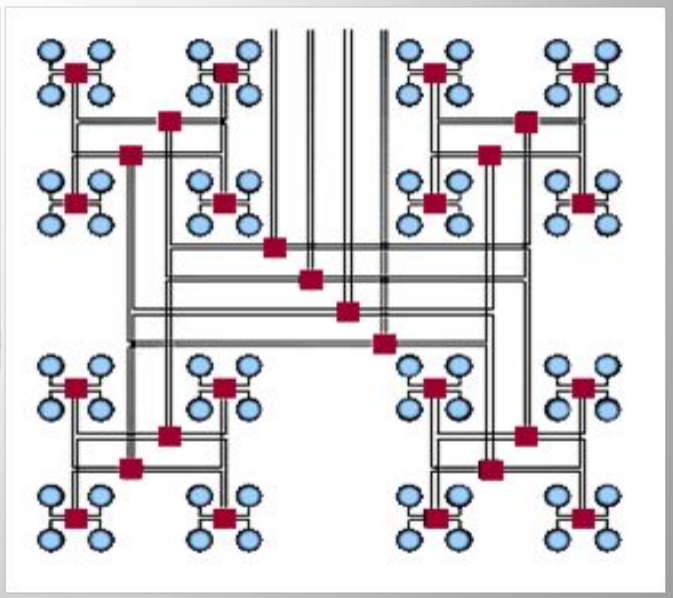
- Архитектура гиперкуба является второй по эффективности, но самой наглядной. Используются и другие топологии сетей связи: трехмерный тор, "кольцо", "звезда" и другие.

- Архитектура кольца с полной связью по хордам (Chordal Ring)



- Наиболее эффективной является архитектура с топологией "толстого дерева" (fat-tree).
- Кластерная архитектура "Fat-tree"





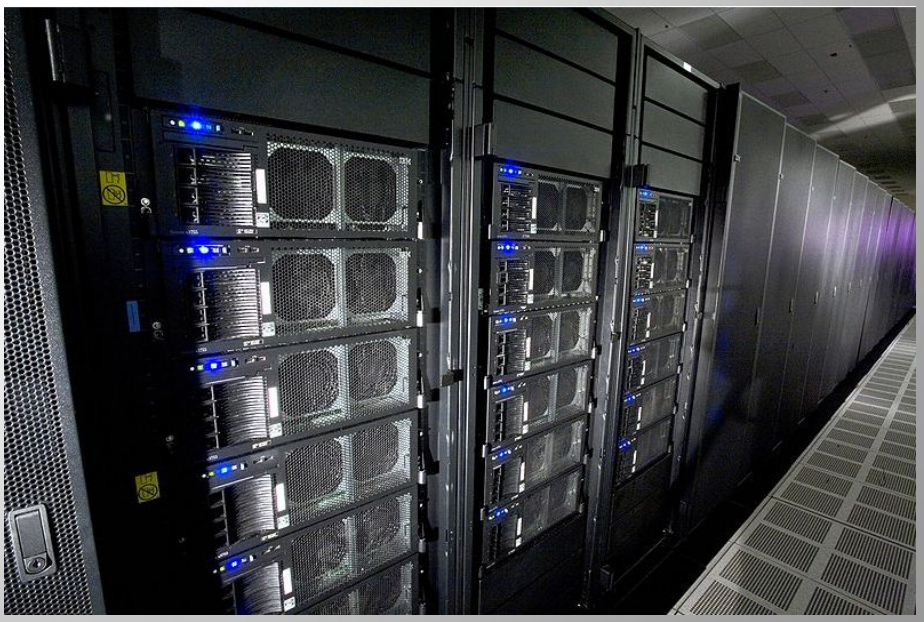
Зарубежные разработки

- **TOP500(июнь 2009) :**

Фирма, тип системы, микропроцессора и межпроцессорной сети; страна	Ввод в дейст. гг.	Колич. процес-сорных ядер	Производит. [Tflops]		Энер-потр. кВт	Top 500 06.09
			пиков.	тест Linp-k		
IBM Roadrunner; Cell 8i; 3,2 ГГц; 2Opteron; 1,8 ГГц; Infiniband	2008	129600	1456,7	**1105,0**	2483	**1**
Cray XT5 Jaguar; 4Opteron; 2,3 ГГц	2008	150152	1381,4	**1059,0**	6951	**2**
IBM BG/P; 2Pow.PC450; 0,85 ГГц; Герм.	2007-9	294912	1002,7	**825,5**	2268	**3**
SGI Altix; 4Xeon; 3/2,7 ГГц; Inf-b-d	2008	51200	608,8	**487,0**	2090	**4**
IBM Blue Gene/L; PowerPC440; 0,7 ГГц	2005-7	212992	596,4	**478,2**	2330	**5**
Sun Runger; 4Opteron; 2 ГГц; Inf-band	2008	62976	579,4	**433,2**	2000	**8**
Bull; 4Xeon; 2,9 ГГц; Inf-band; Герм.	2009	26304	308,3	**274,8**	1549	**10**
Cray XT4 Jaguar; 4Opteron; 2,1 ГГц	2006-8	30976	260,2	**205,0**	1581	**12**
IBM Blue Gene/P; Саудов. Аравия	2009	65536	222,8	**185,2**	504	**14**
Dawning;4Opteron;1,9 ГГц; Inf-b-d; КНР	2007-8	30720	233,5	**180,6**		**15**
HP Индия; 4Xeon; 3 ГГц; Infiniband	2007-8	14384	172,6	**132,8**	786	**18**
NEC; SX9/E, Earth Simul.; 3,2 ГГц; Яп.	2009	1280	131,1	**122,4**		**22**
IBM Power575; Pow.6; 4,7 ГГц; I-d; В-бр.	2008	8320	156,4	**115,9**	1330	**23**
Lenovo; 4Xeon; 3 ГГц; Inf-band; КНР	2008	12216	146,0	**102,8**		**31**
Hitachi 4Opt.; 2,3 ГГц; Myrinet 10G; Яп.	2008	12288	113,1	**83,0**	639	**42**
МСЦ РАН, HP; 4Xeon; 3ГГц; Infinib-d	2007-8	7920	95,0	**71,3**	327	**54**
NEC; 4Xeon; 2,8 ГГц; Infiniband	2009	5376	60,2	**50,8**	186	**77**
НИВЦ МГУ, Т-Пл.; 4Xeon; 3 ГГц; In-b.	2008	5000	60,0	**47,2**	265	**82**
IBM; 2Xeon; 2,66 ГГц; GigEth-net	2008	3528	37,6	**17,1**	329	**500**

- Местоположение- Лос-Аламосская национальная лаборатория в Нью-Мексико, США
- 1 место TOP500 на июнь 2009 г.
- Пиковая производительность - 1,456 PFlops.

- Площадь ~ 12 000 кв.футов (1100 м²)
- Вес - 226 тонн.
- Стоимость IBM Roadrunner - 133 миллиона долларов.
- **Кластер Roadrunner**
- Суммарная информация:
- 6480 двухъядерных Opteron с 51,8 Тб ОЗУ
- 12 960 Cell процессоров с 51,8 Тб ОЗУ
- 216 узлов ввода-вывода System x3655
- 26288-портовых маршрутизаторов
- ISR2012 Infiniband 4x DDR
- 296 корпусов



- Местоположение- Национальный центр компьютерных исследований в Окридже, штат Теннеси
- 2 место TOP500 на июнь 2009 г.
- Пиковая производительность - 1,38 PFlops.
- Суперкомпьютер имеет массово-параллельную архитектуру, то есть состоит из множества автономных ячеек .
- **Каждая вычислительная ячейка содержит :**

1) 2 четырехъядерных процессоров AMD Opteron 2356 (Barcelona) с частотой 2.3 ГГц
- 2)16 ГБ памяти DDR2-800,
- 3) роутер SeeStar 2+
- Всего содержится 149’504 вычислительных ядер, более 300 ТБ памяти, более 6 ПБ дискового пространства.

В период с июля по ноябрь 2009 г. происходит апгрейд ячеек: процессоры Opteron заменяются с 4-ядерных на 6-ядерные. Апгрейд планируется произвести в 5 этапов, каждый раз отсоединяя и модернизируя часть ячеек.

- **TOP500(июнь 2009) :**

Дислокация	Количество	Σ[Tf/s]	Верхние места
1. США	**291**	13721	1–2
2. Великобрит.	**44**	1248	25
3. Германия	**30**	2208	3
4. Франция	**23**	1005	20
5. Китай	**21**	788	15
9. Индия	**6**	247	18
10. Италия	**6**	186	46
15. Россия	**4**	167	54
Прочие 15 стран	**26**	739	14
В с е г о		22608	

- **СНГ — Top50(июнь 2009):**
- 1. МВС-100К (МСЦ РАН,НР) – 54 позиция ТОП500
- 2. СКИФ МГУ - 82 позиция ТОП500

- Пиковая производительность - 95,04 TFlops.

Технические средства "МВС-100К":

990 вычислительных модулей (7920 процессорных ядер)

- **Вычислительный модуль**

4-х процессорный сервер HP Proliant :

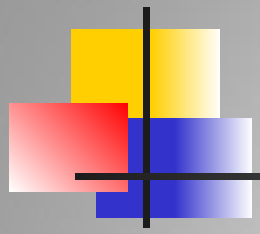
1. 2 4-х микропроцессора Intel Xeon, работающих на частоте 3 ГГц;
2. оперативную память DDR2 > 4 ГБайт;
3. жёсткий диск объёмом > 36 ГБайт;
4. интерфейсная плата HP Mezzanine Infiniband DDR;
5. два интегрированных контроллера Gigabit Ethernet.

- управляющая станция и узел доступа на
- базе двух процессоров Intel Xeon;
- коммуникационная сеть Infiniband DDR,
- построенная с использованием коммутаторов
- Voltaire и Cisco;
- транспортная сеть Gigabit Ethernet;
- управляющая сеть Gigabit Ethernet;
- системная консоль;

- июнь 2008 -1-ое место ТОП 50
- июнь 2009 -2-ое место
- совместная разработка МГУ, Института программных систем (ИПС) РАН и компании «Т-Платформы» .
- является 7-м по мощности среди всех суперкомпьютеров, использующихся в мировой системе образования.
- Общая стоимость проекта - 231 млн. руб.

- **Технические характеристики суперкомпьютера «СКИФ МГУ»:**

Число вычислительных узлов/процессоров	625/1250
Конструктив узла	blade
Количество монтажных шкафов вычислительного кластера	14
Тип процессора	четырехъядерный Intel® Xeon® E5472, 3,0 ГГц
Пиковая производительность	60 TFlops
Производительность на тесте Linpack	47.17 TFlops (78% от пиковой)
Тип системной сети	DDR InfiniBand (Mellanox ConnectX)
Скорость передачи сообщений между узлами	не менее 1450 Мб/сек
Задержка при передаче пакетов данных	не более 2.2 мкс
Тип управляющей (вспомогательной) сети	Gigabit Ethernet
Тип сервисной сети	IPMI + СКИФ-ServNet
Оперативная память	5.5 ТБ
Дисковая память узлов	15ТБ
Тип системы хранения данных	T-Platforms ReadyStorage ActiveScale Cluster
Объем системы хранения данных	60 ТБ
Занимаемая площадь	96 м2
Потребляемая мощность вычислительного кластера	330 кВт
Потребляемая мощность установки в целом	520 кВт (в пике возможно до 720кВт)
Суммарная длина кабельных соединений	более 2 км



На данный момент в России 42 установки с производительностью > 1 TFLOPS.

Для попадания в список Top50 теперь требуется производительность в тесте Linpack не менее 432 GFLOPS.

Систем в списке ТОП50 :

- Intel - 38
- AMD -6
- IBM-5
- HP-1