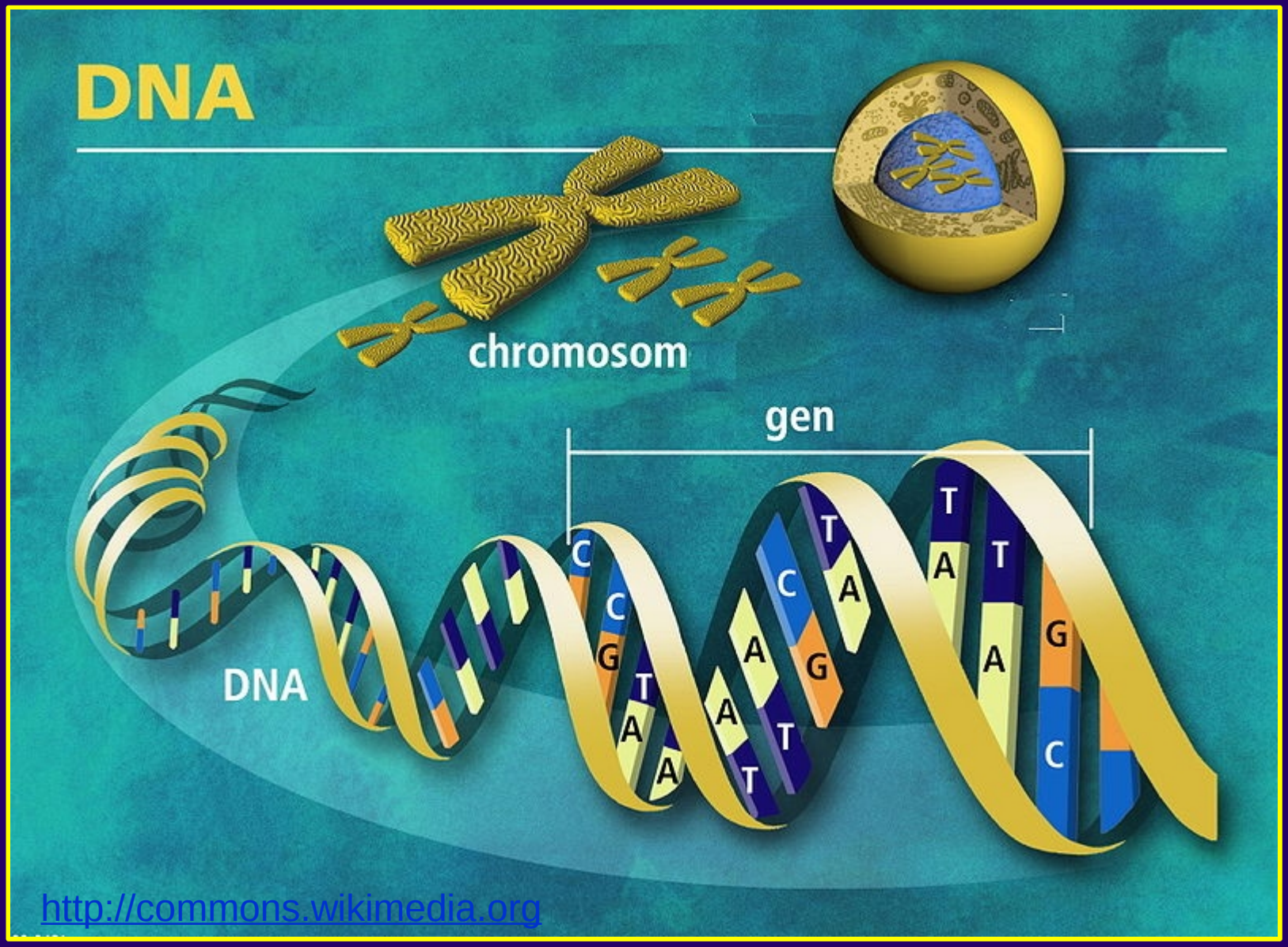
Ubuntu 13.10 Install Fest
МИЭМ НИУ ВШЭ 2.11.2013

Linux для обработки геномов

Сергей Науменко
Ренат Арифулов
Нина Попова

Лаборатория эволюционной геномики ФББ МГУ
Институт проблем передачи информации РАН
РХТУ им. Д.И. Менделеева

Секвенирование



Секвенирование – прочтение последовательности ДНК или РНК

Проект “геном человека”

- 3,5 миллиарда нуклеотидов
- 13 лет
- 3 миллиарда долларов

Цели проекта:

- Все гены
- Полный геном
- Базы данных
- Новые инструменты
- Передача технологий
- Разработка этических , правовых и социальных аспектов

http://web.ornl.gov/sci/techresources/Human_Genome/index.shtml

Высокопроизводительное секвенирование

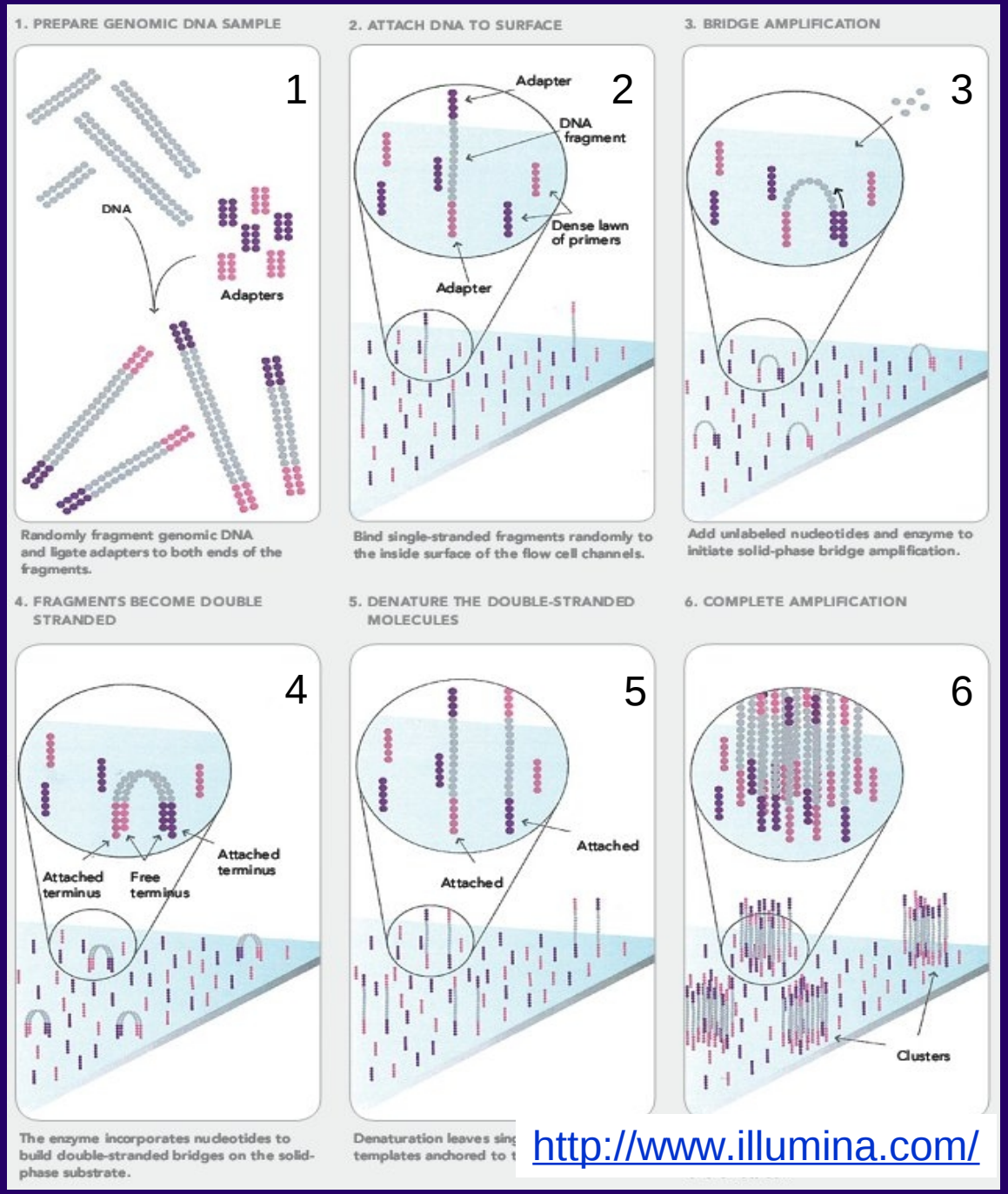
1)Фрагментация
ДНК, пришивание адаптеров

2) Связывание фрагментов с подложкой

3-4) Добавление нуклеотидов, образование двойных цепочек

5)Денатурация
(разрыв двойных цепочек, отрыв одного конца от подложки)

6) Амплификация - многократное копирование фрагментов ,
образование кластеров



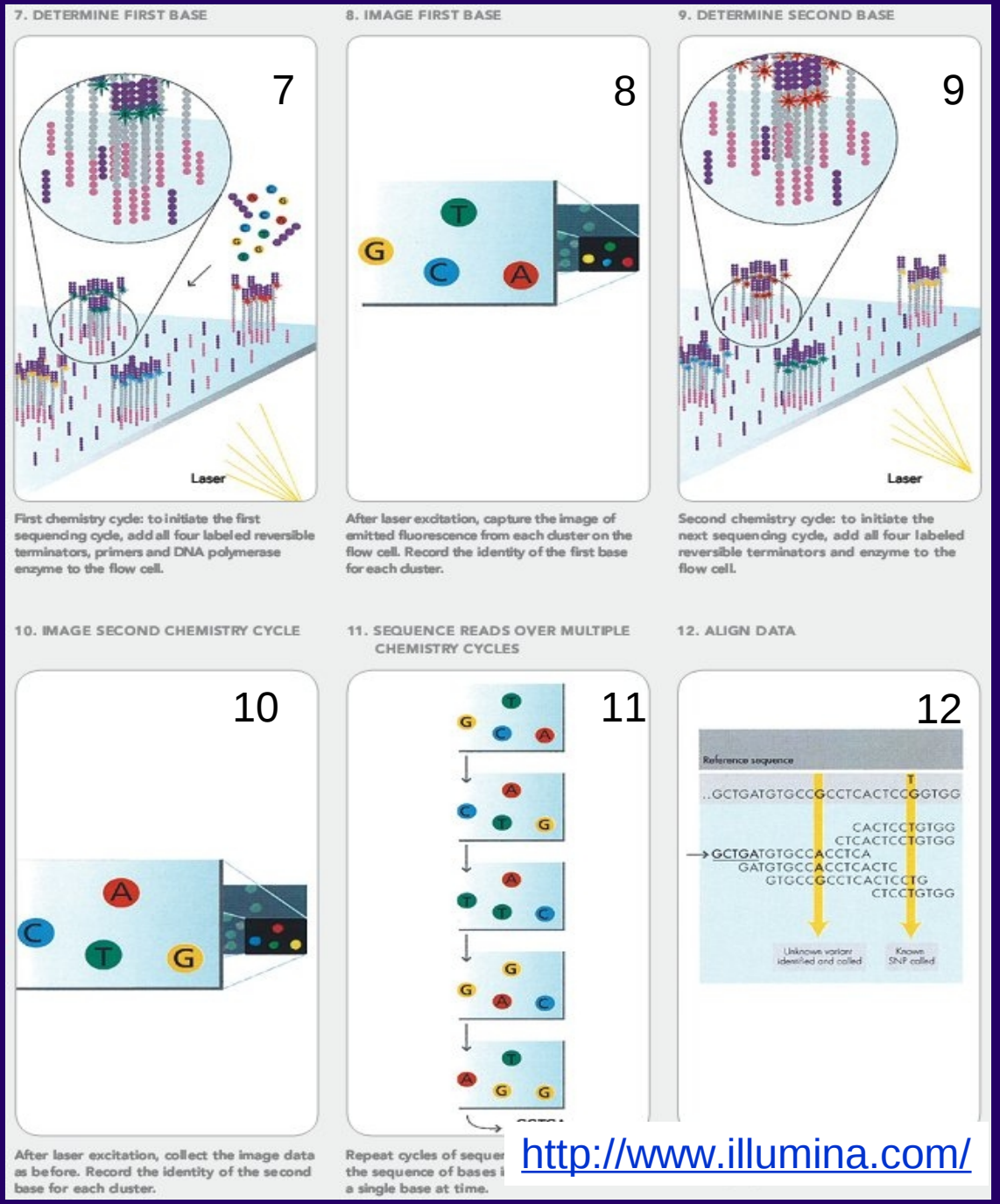
http://www.illumina.com/

Высокопроизводительное секвенирование

7-8) Добавление меченых нуклеотидов, определение первого нуклеотида при помощи лазера

9-11) Чтение остальных нуклеотидов

12) Анализ данных



Illumina HiSeq 2000:

- Две недели
- 600 миллиардов нуклеотидов
- 17 геномов человека с покрытием 10x
- 25 тысяч долларов США

http://www.illumina.com/

Высокопроизводительное секвенирование

в мире

Пекинский институт геномики (128 секвенаторов)

华大基因 BGI

http://www.genomics.cn/en/index

wellcome trust sanger institute

Институт Сэнгера

http://www.sanger.ac.uk/

BROAD INSTITUTE

http://www.broadinstitute.org/

JGI DOE JOINT GENOME INSTITUTE

http://www.jgi.doe.gov/

в России

Лаборатория эволюционной геномики ФББ МГУ

http://evolgenomics.fbb.msu.ru/

ЦНИИ эпидемиологии Роспотребнадзора

ЦНИИ МЗ РФ

http://www.pcr.ru/

Центр геномный исследований НОЦ СФУ

http://genome.sfu-kras.ru/

ИОГЕН РАН, лаборатория эволюционной геномики

Институт общей генетики им. Н.И. Вавилова РАН

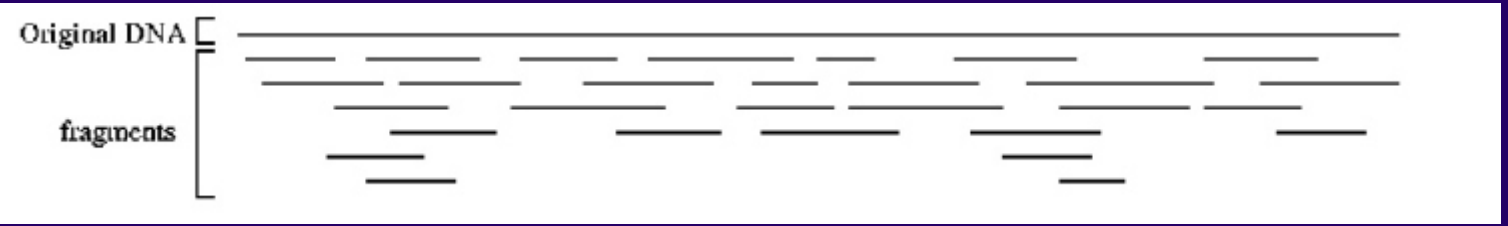
http://vigg.ru/

Основные задачи обработки данных

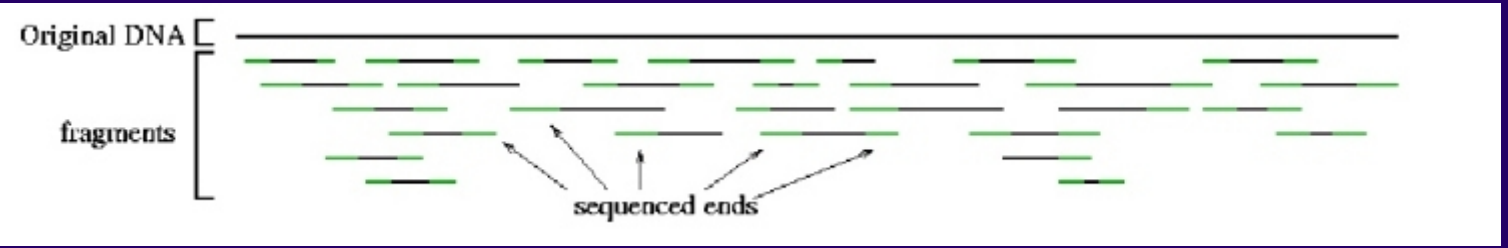
- Приём данных с секвенаторов
- Первичная обработка данных
- Сборка геномов и транскриптомов
- Картирование
- Аннотация
- Выравнивание
- Филогенетика
- Сравнительная геномика
- Популяционная генетика

Задача сборки генома

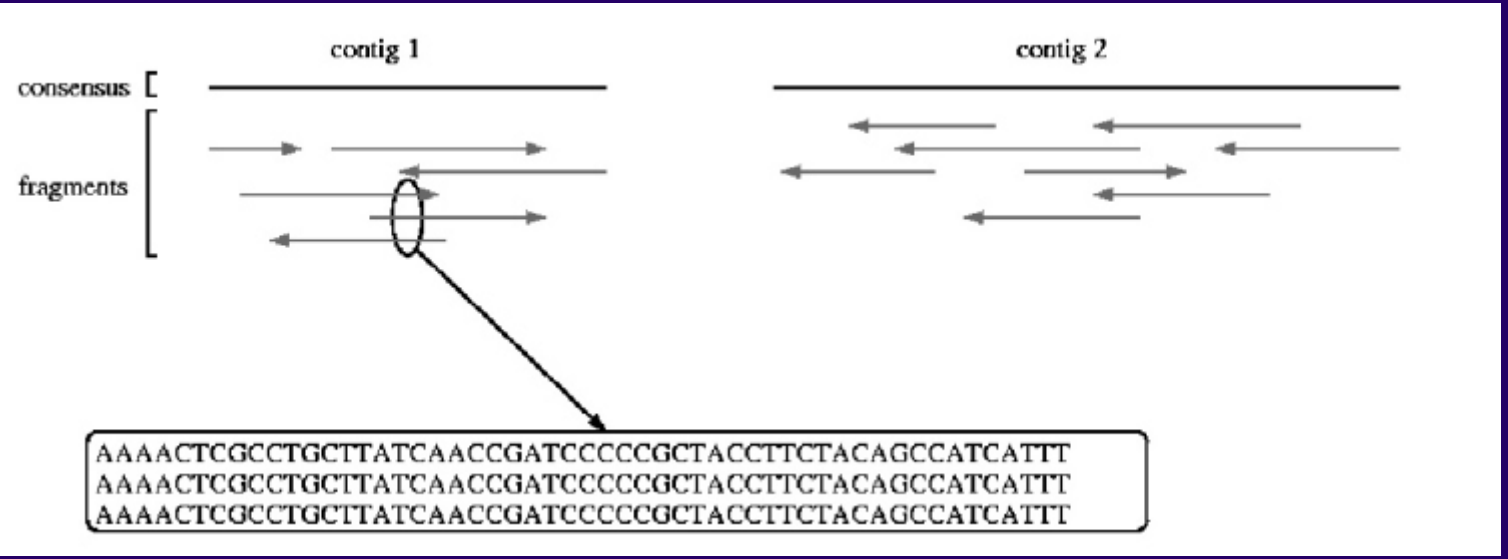
Исходная ДНК разрезается на фрагменты.



Фрагменты прочитываются с двух концов.



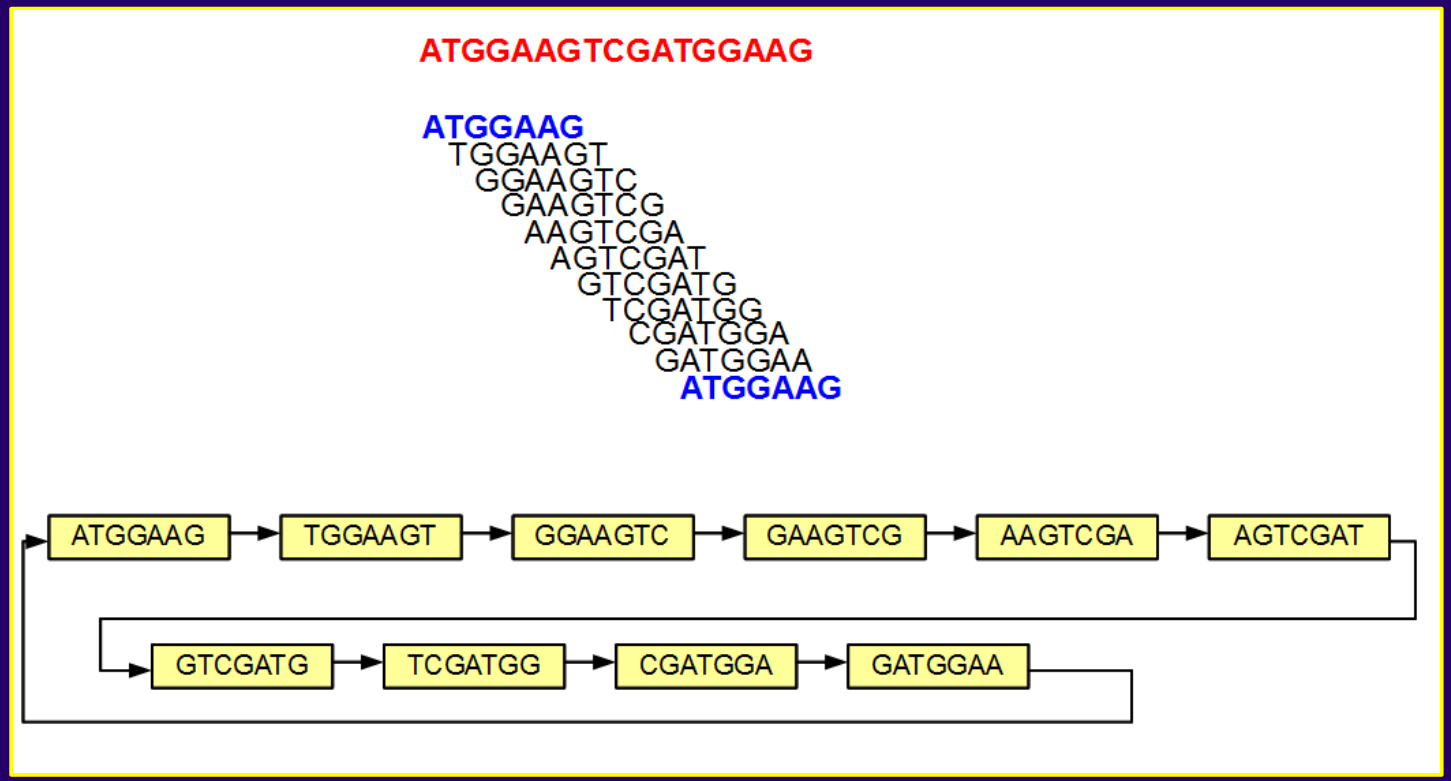
Схожие фрагменты объединяются в длинные строки.

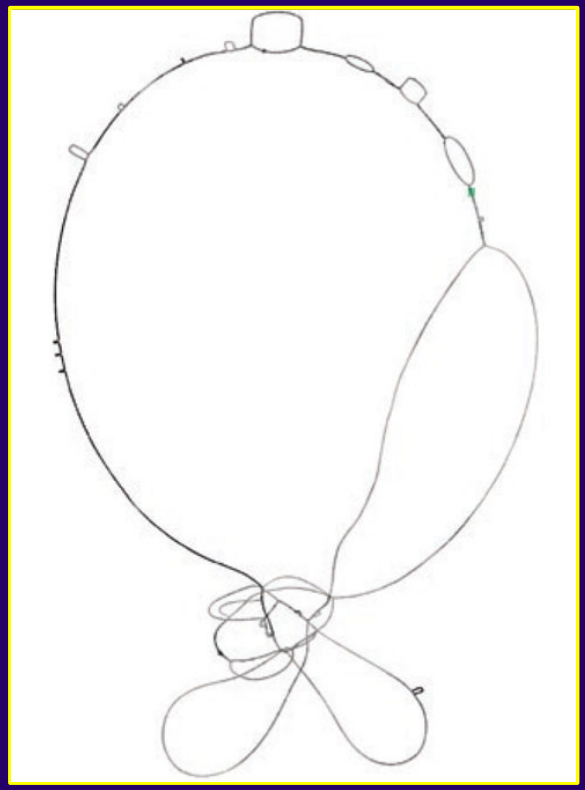


http://www.cbcb.umd.edu/research/assembly_primer.shtml

Алгоритм сборки генома

- Фрагменты разбиваются на слова (k-mer)
- Строится строковый граф
- Удаляются дублирующие пути
- На графе находится максимальный путь

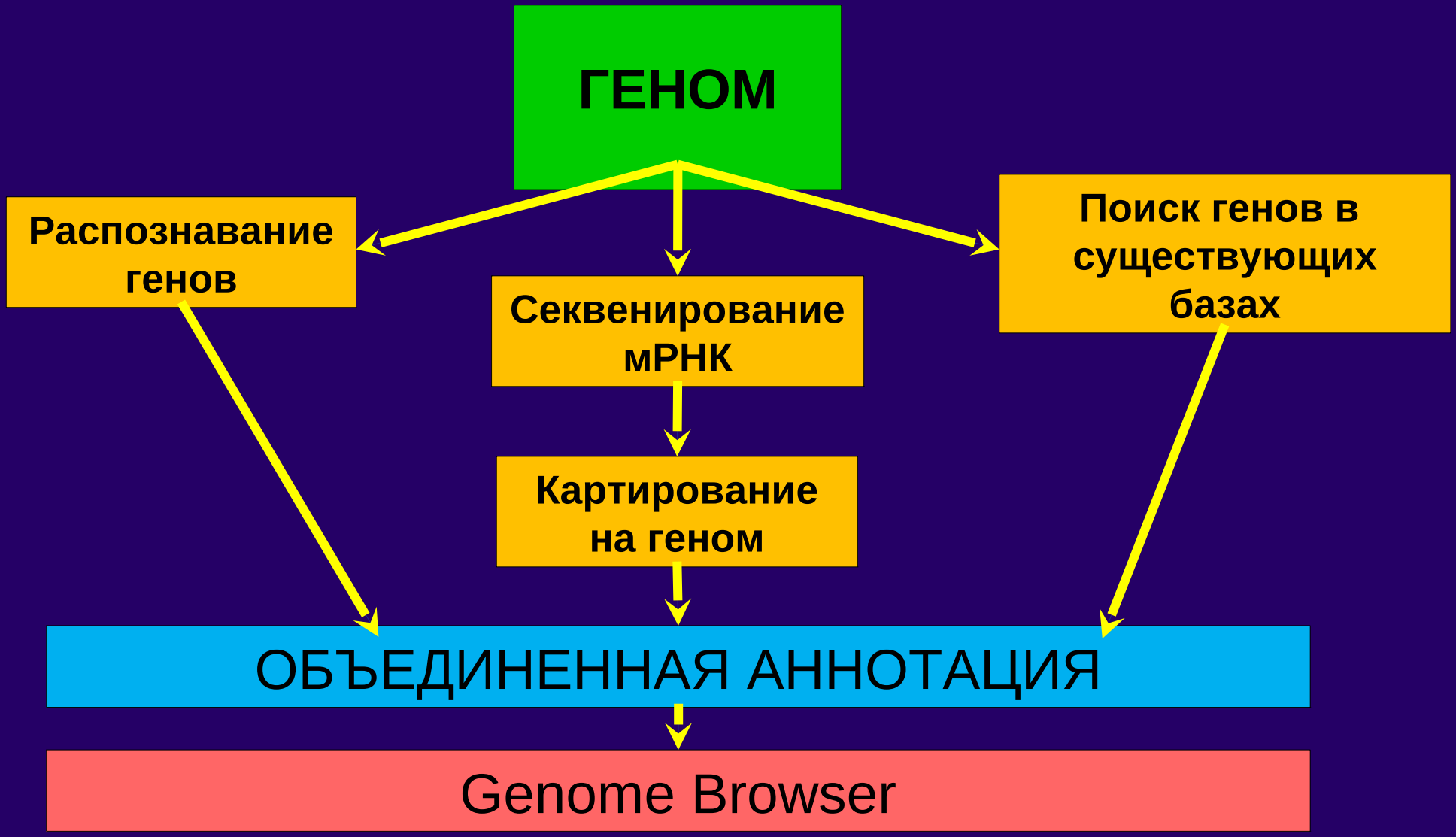




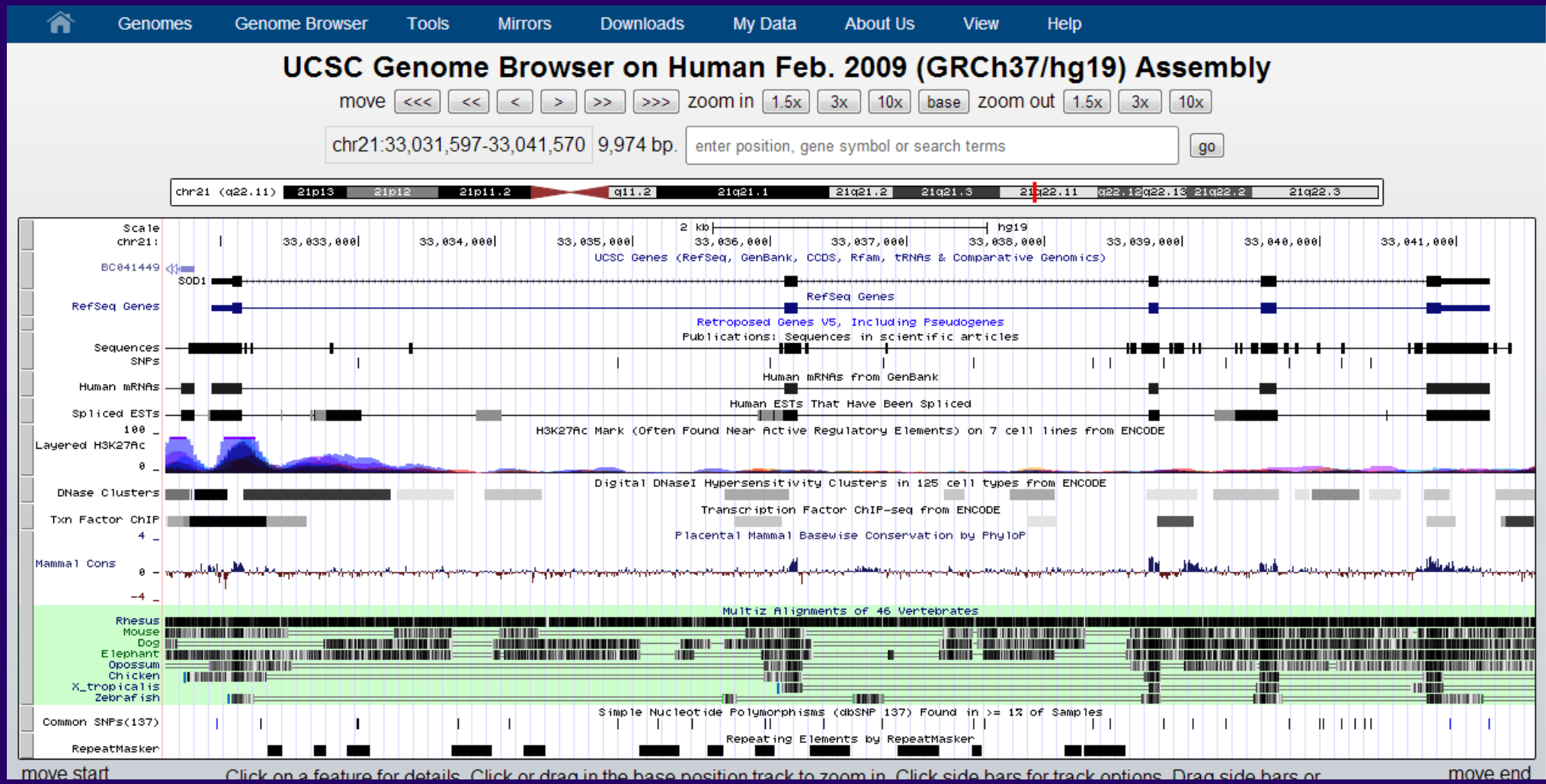
Paul Flicek & Ewan Birney. *Nature Methods* 6

Аннотация генома – «разметка» генома

выделение последовательностей белок-кодирующих генов и РНК



Собранный и аннотированный геном

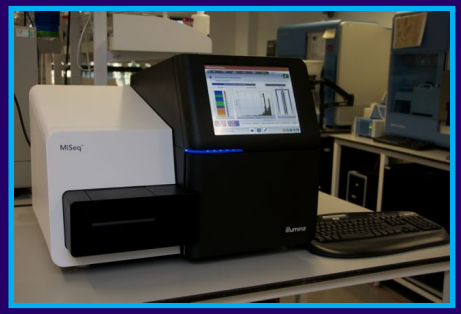


Потоки данных

Пользователи:
(более 50)

Illumina HiSeq2000

Illumina MiSeq



Центр обработки данных

Лаборатория эволюционной геномики ФББ МГУ

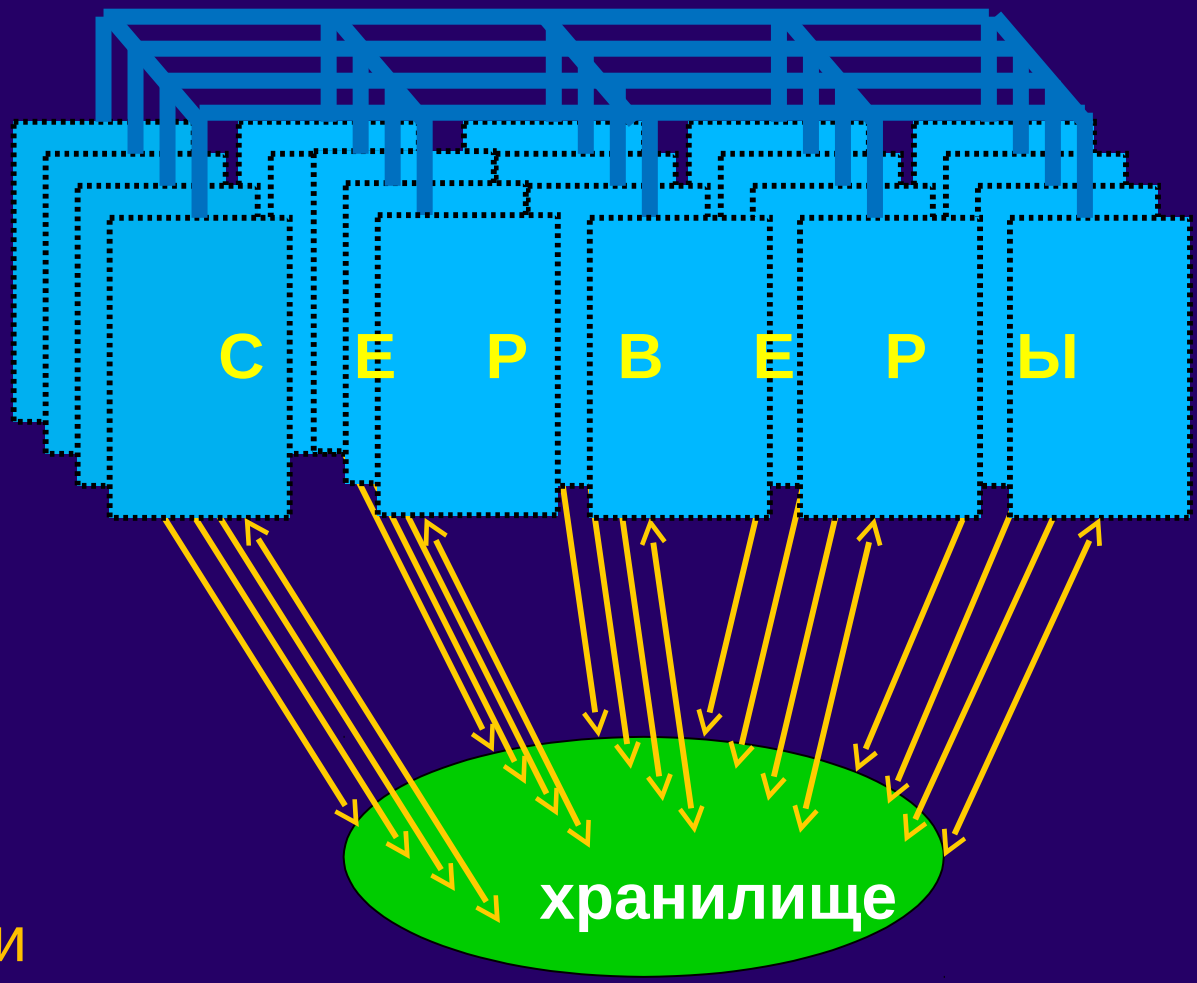
Лаборатория биоинформатики ФББ МГУ

УНЦ Биоинформатика

Вычислительный кластер

Предназначен для решения задач:

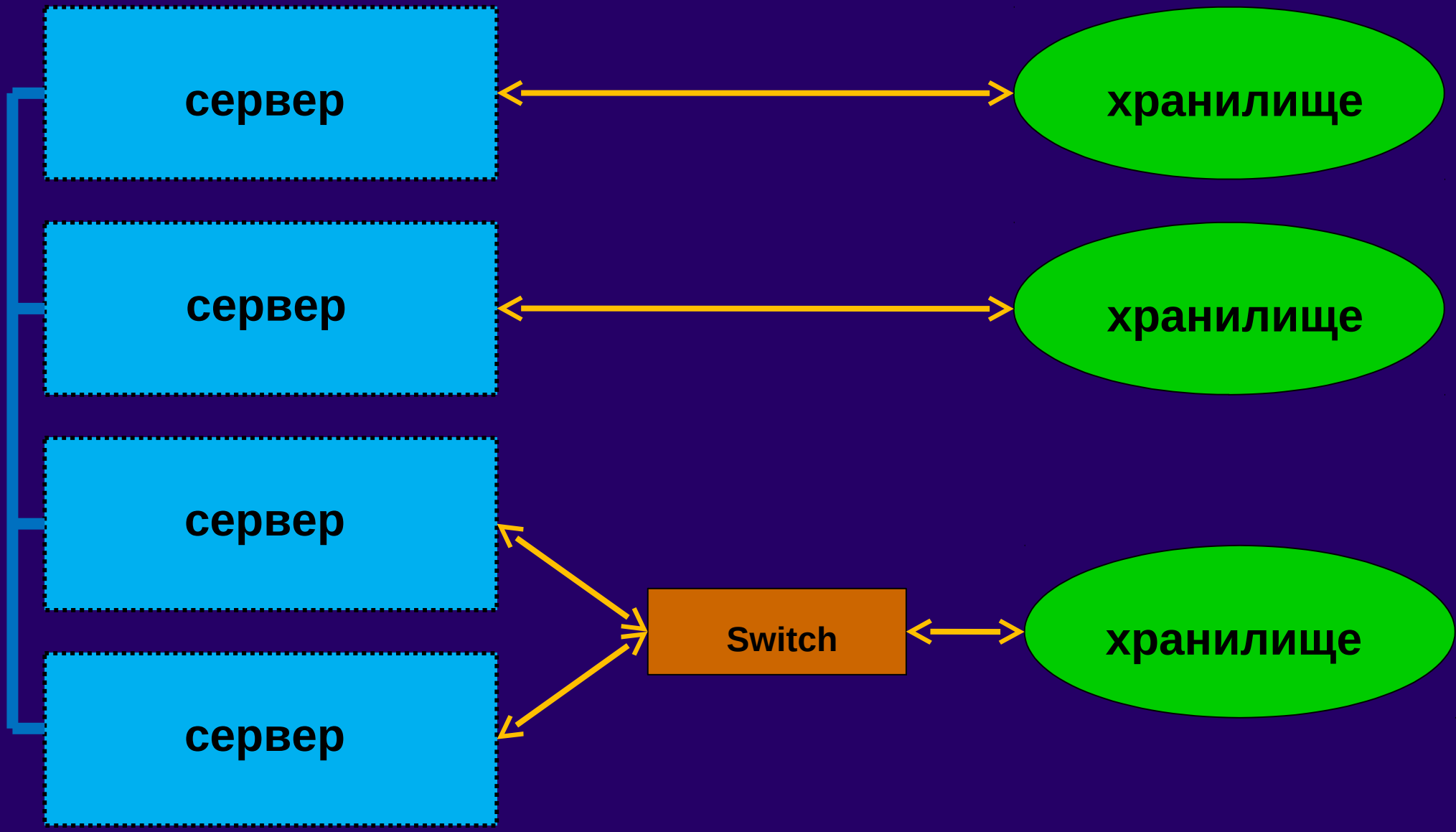
- Вычислительной гидродинамики
- Квантовой химии
- Физики высоких энергий



Особенности:

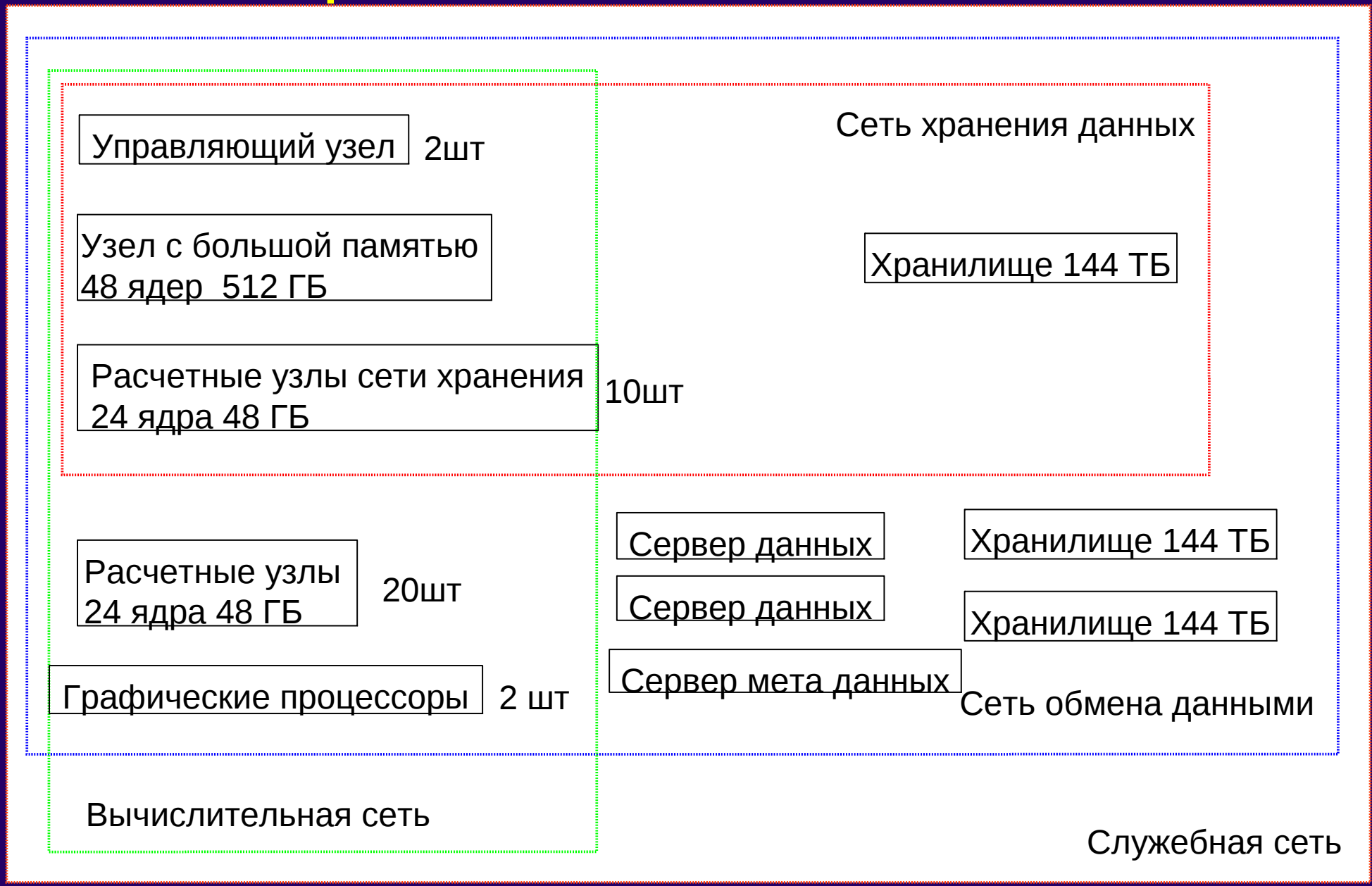
- Много ядер
- Мало оперативной памяти
- Небольшое хранилище

Центр обработки данных



Количество серверов сопоставимо с количеством хранилищ

Схема ЦОД лаборатории эволюционной геномики ФББ МГУ



Программное обеспечение

- Управление ресурсами (очередь задач) – torque
- Мониторинг – nagios
- Управление конфигурациями – puppet
- Файловые системы (XFS, lustre)
- OS – Enterprise Linux (Scientific Linux)
- Биоинформатические пакеты

Биоинформатические пакеты

- Velvet
- Soapdenovo
- Platanus
- GATK
- MUMmer
- Clustal
- R
- Biopython
- BioPerl
- PHYLYP
- SHRiMP
- STAR
- Agalma
- Bambus
- Bamtools
- Blast
- Blat
- RAxML

- MCScanX
- Python
- Pal2nal
-

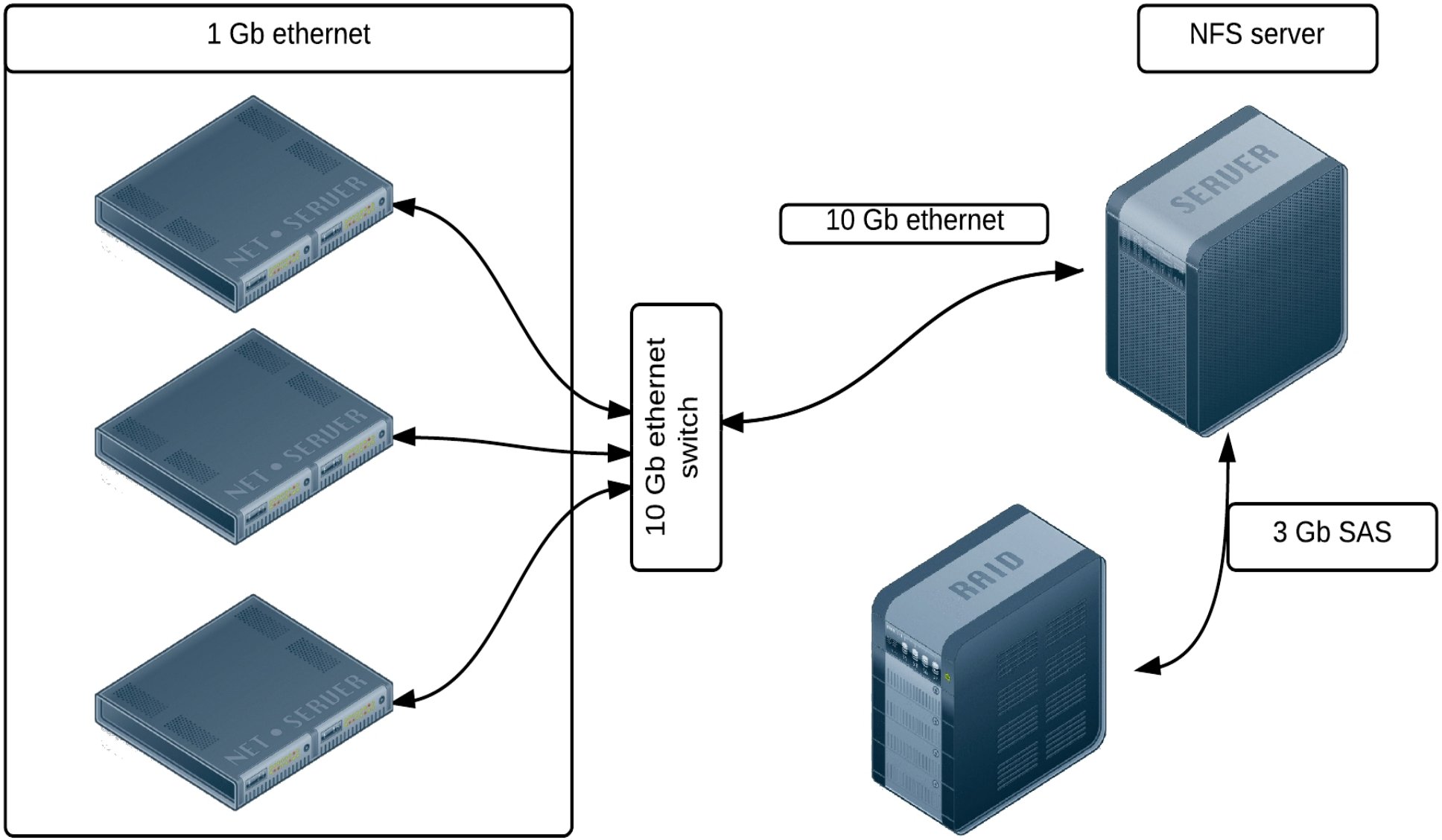
AdapterRemoval

- MCScanX
- HaploMerger
- Paml
- Mrbayes
- geneid
- bowtie
- HaploMerger
- Megan

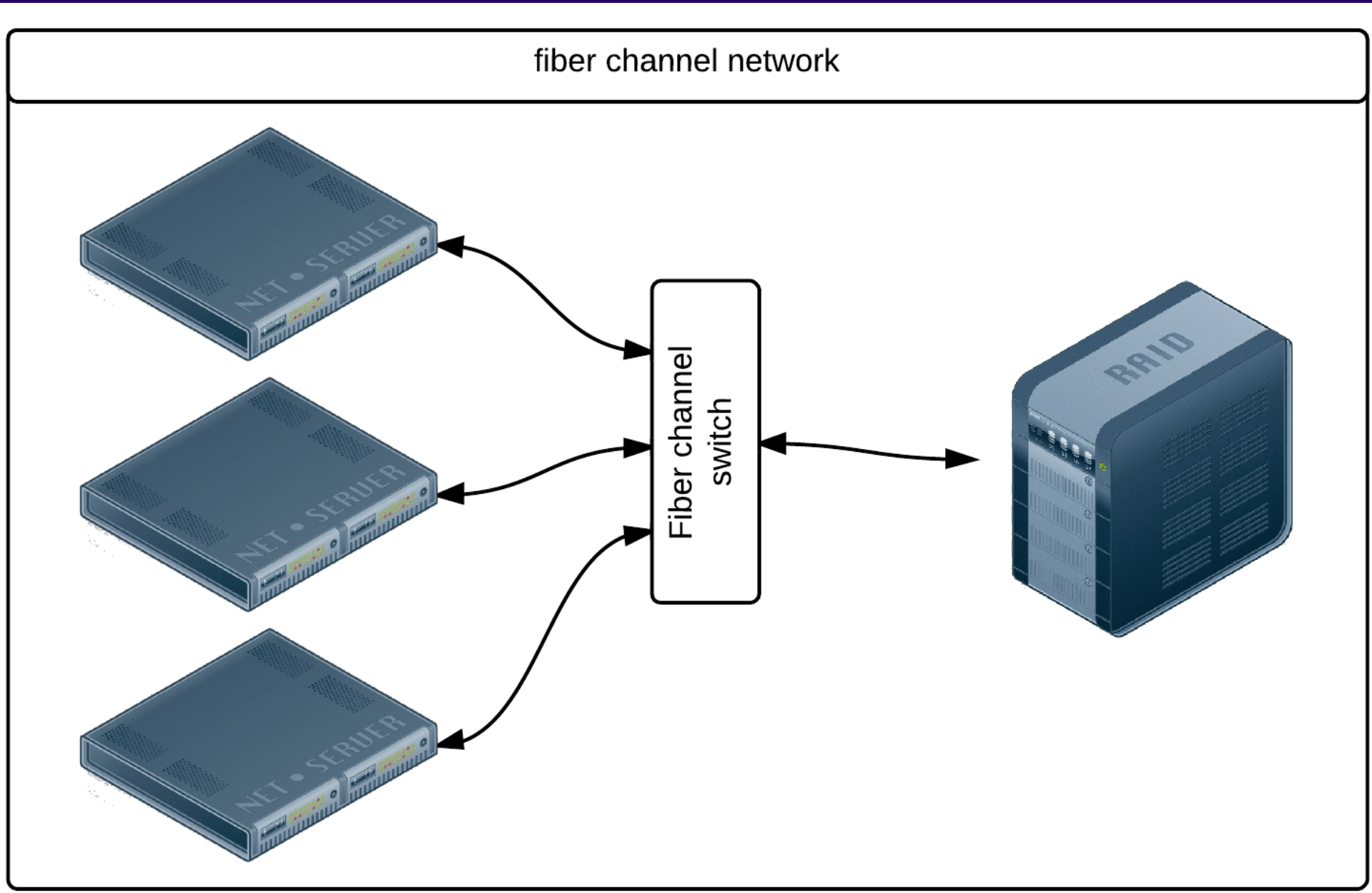
- Cuda
- Orthomcl
- Jellyfish
- Cegma
- Annovar
- Abyss
- AdapterRemoval
- Statistics-Descriptive
- beagle-lib

- libsequence
- BaseSpaceSHRE
- wise
- RepeatMasker
- И др

NFS Сервер


1 Gb ethernet
NFS server
10 Gb ethernet
10 Gb ethernet switch
3 Gb SAS
NET • SERVER
SERVER
RAID

Сеть хранения данных



Результаты тестов чтения/записи для файловых систем NFS, OCFS2, GFS2

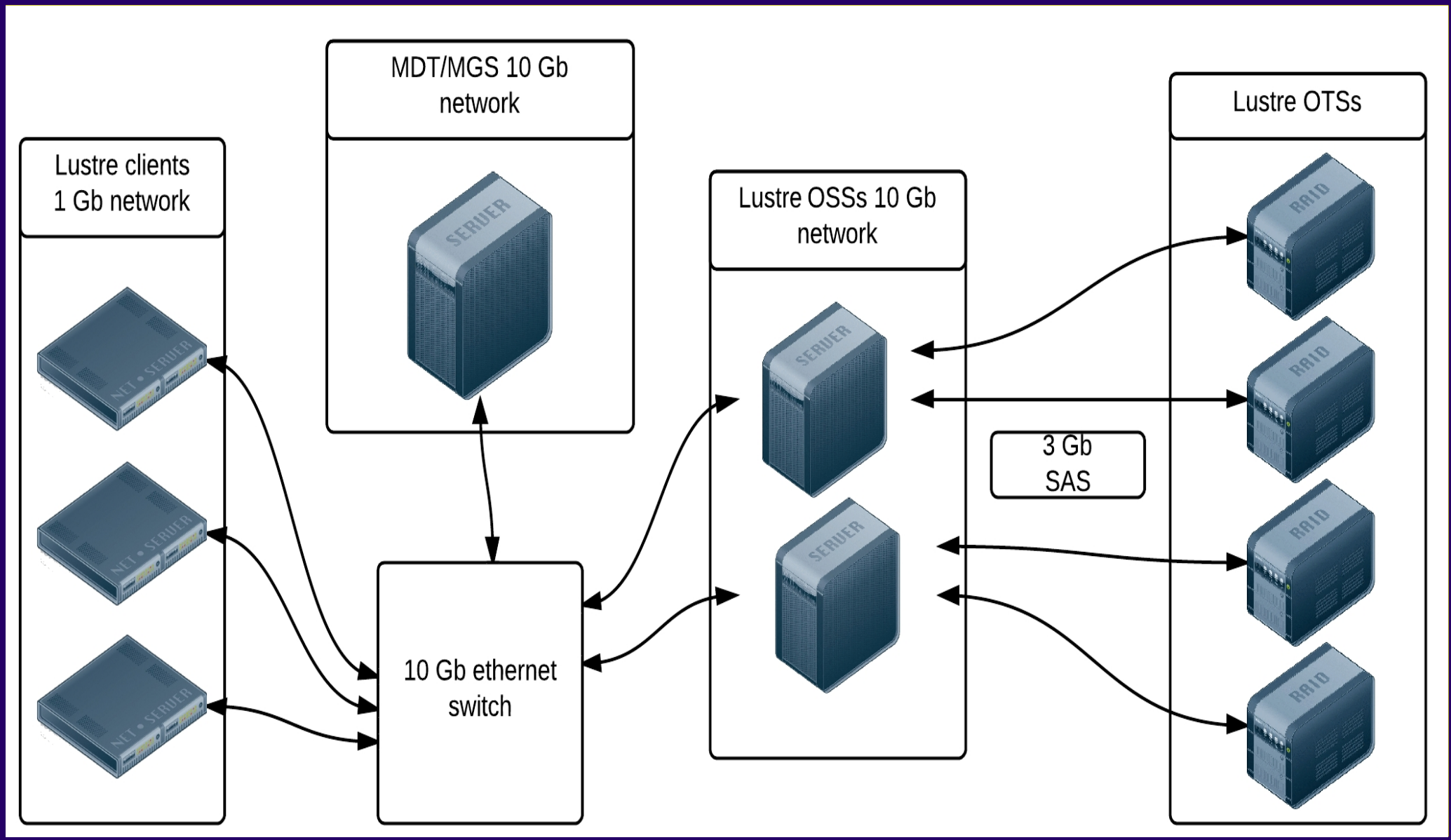
dd (write), oflag=direct

Nodes	OCFS2	GFS2	NFS4
1	120 MB/s	59 MB/s	57 MB/s
2	78 MB/s	54 MB/s	53 MB/s
3	64 MB/s	45 MB/s	51 MB/s

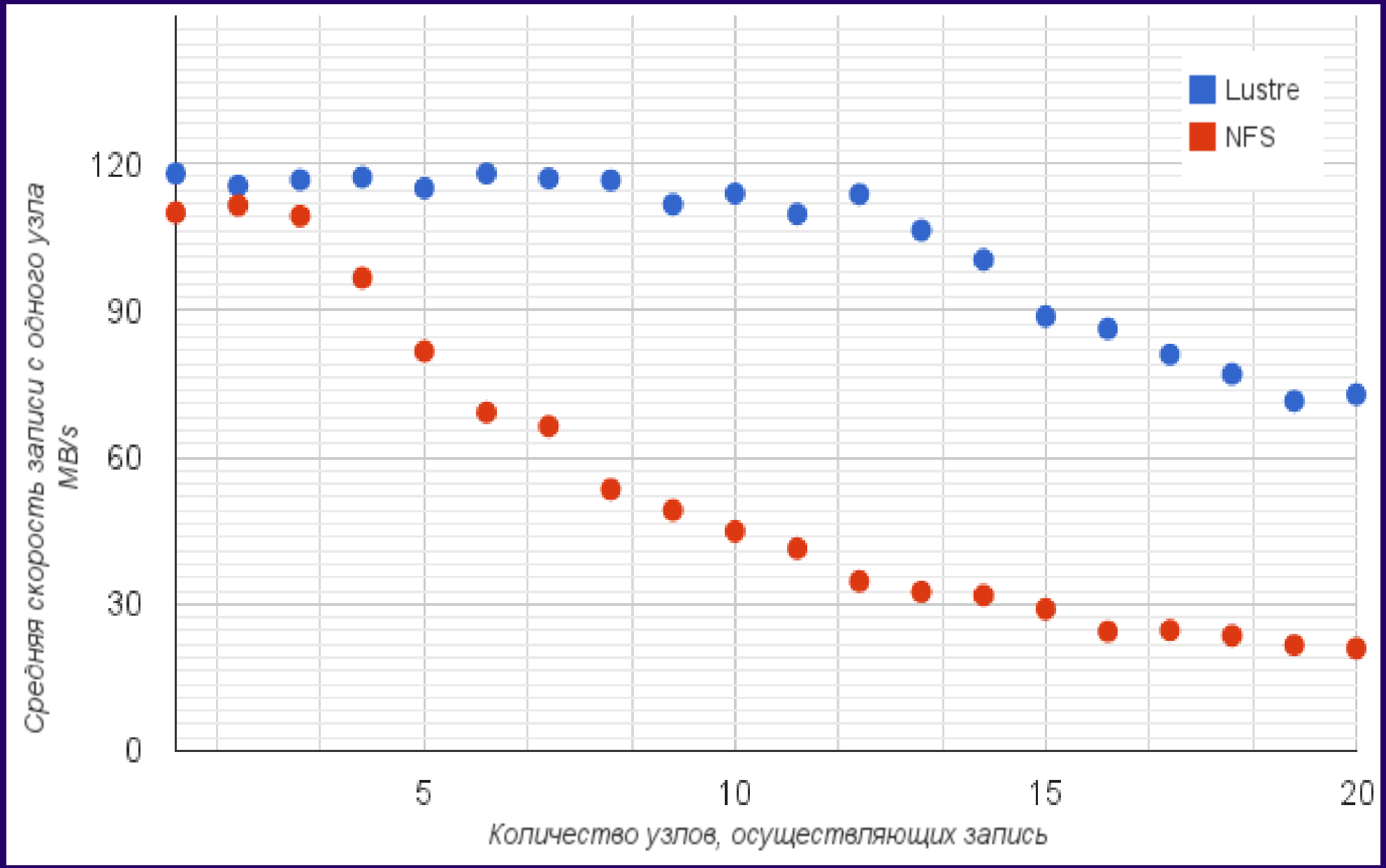
dd (read), iflag=direct

Nodes	OCFS2	GFS2	NFS4
1	274 MB/s	174 MB/s	29 MB/s
2	178 MB/s	90 MB/s	29 MB/s
3	124 MB/s	73 MB/s	29 MB/s

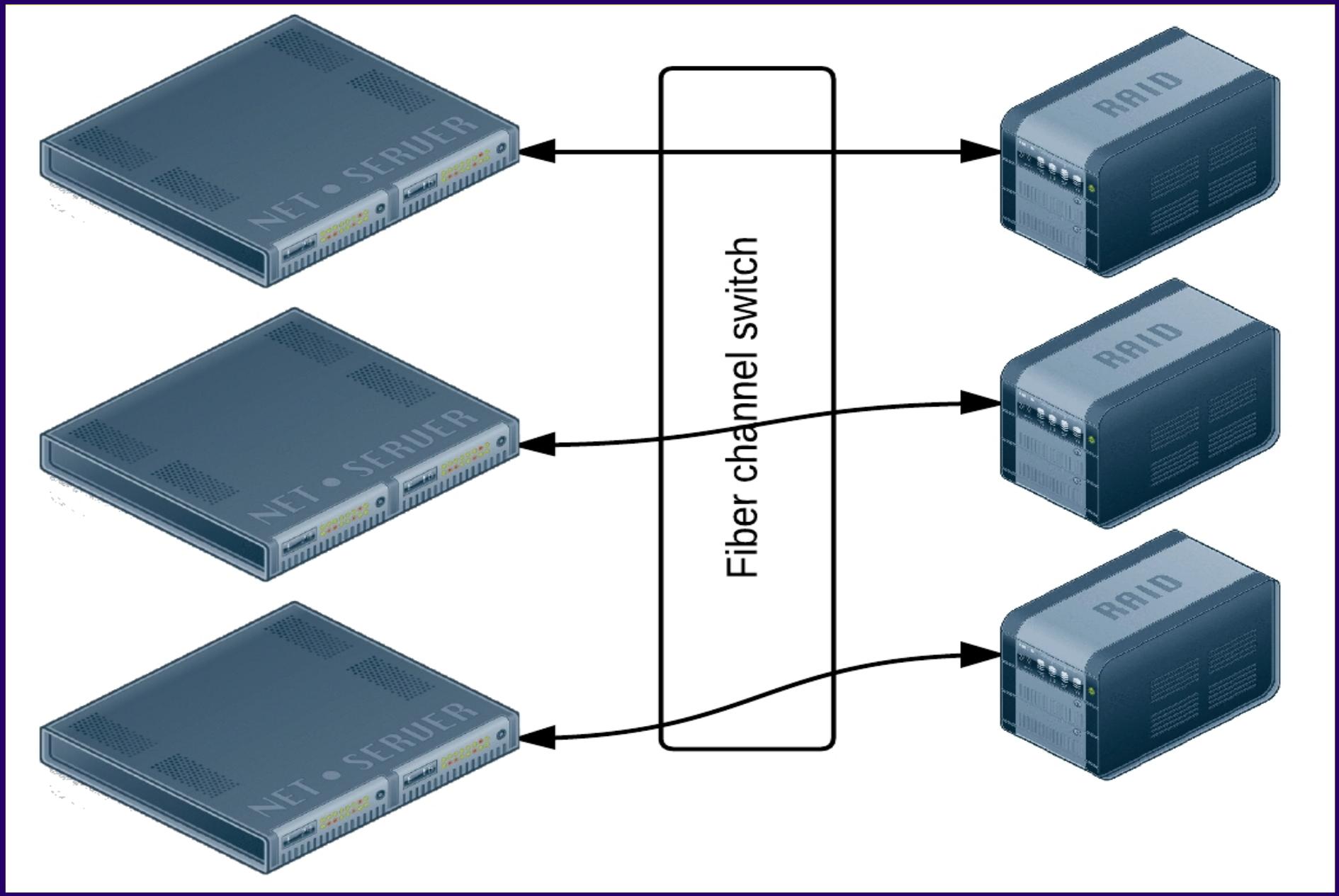
Распределенная файловая система lustre®



Сравнение производительности файловых систем Lustre и NFS



Распределение дисковых ресурсов по проектам



Выводы

- Высокопроизводительное секвенирование – ключевая технология для современной биологии и медицины
- Для обработки геномных данных необходимы соответствующие вычислительные мощности
- Оптимизация потоков данных критична для работы ЦОД
- Сочетание распределенной файловой системы Lustre и инфраструктуры Fiber Channel является оптимальным решением для ЦОД