

Санкт-Петербургский государственный университет  
информационных технологий, механики и оптики  
кафедра вычислительной техники

# Разработка архитектуры и методов организации слабосвязанных архивных систем для автоматизации проектирования

Лукьянов Н.М.

Руководитель:

к.т.н., доц. Тимченко Б.Д.

2011

# Тенденции развития CAD систем

- Растущая производительность рабочих станций
- Переход от суперкомпьютерных архитектур к распределенным высокопроизводительным системам
- Уменьшение цикла разработки изделия
  - По данным Boeing 60 месяцев -> 12 месяцев
- Снижение стоимости разработок
  - По данным Boeing 7 млрд. -> 1 млрд. \$
- Интеграция с PLM и ERP системами

- Увеличение объема хранения периодически используемых данных
- Переход к распределенному хранению данных
- Необходимость снижения совокупной стоимости владения инфраструктурой



# Задачи работы

- Разработать архитектуру распределенной системы архивного хранения и доставки файловых данных, построенную на слабосвязанных гетерогенных узлах с узкими каналами связи
- Предложить подход к разделению программной и аппаратной составляющих системы
- Разработать метод адаптации системы в части управления, как размещением, так и доставкой контента.

# Информационное обеспечение САПР

Архивируемые данные:

- *выполненные проекты*
- *фотографии объектов, материалов, помещений*
- *видеоотчеты пусконаладочных работ и т.д.*
- *3D модели*
- *подписанная КД*

Объем данных на 1 организацию:

Средний объем проекта – 500 Гб

Проектов в год – 30

Глубина работы с архивом – 5 лет

Общий объем хранения архивной системы (на организацию) – 75 Тб

## Классические решения



Центролизированные NAS, SAN системы:

*высокая стартовая стоимость и высокая TCO*



Ленточные (LTO) и DVD-RAM библиотеки:

*Низкая скорость доставки короткая жизнь носителей*

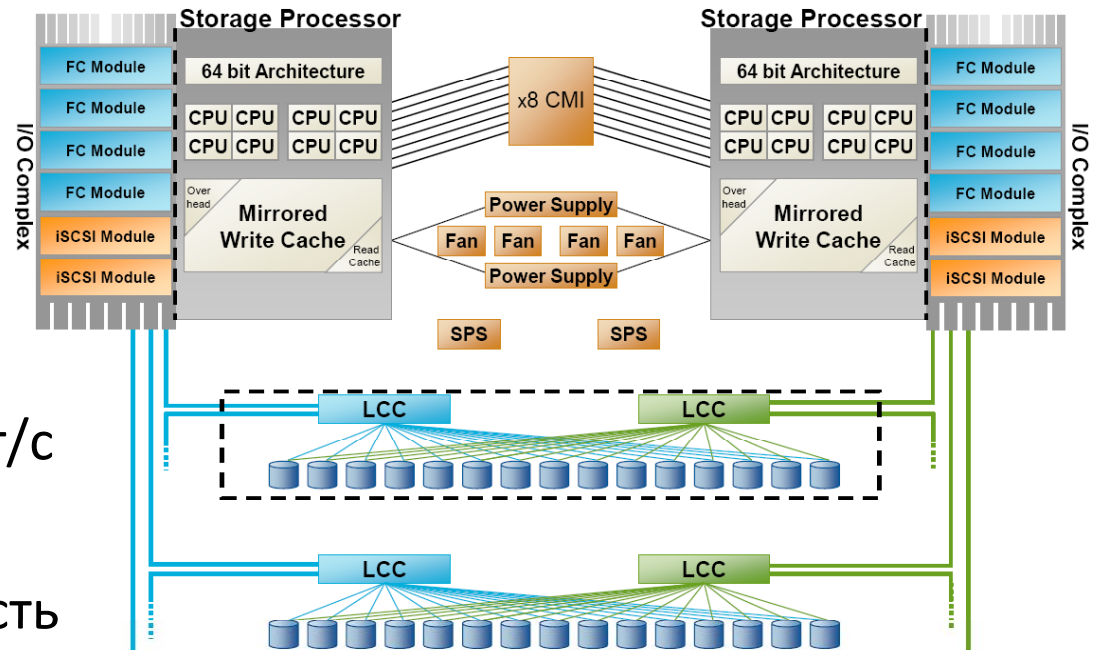
# СХД EMC “Celerra”

## EMC Celerra CX4-120

- Средний ценовой диапазон
- TCO 500\$ / Тб / год
- SAN сеть 8 Гбит/с
- NAS NFS, CIFS 10 Гбит/с

## Ограничения

- высокая гранулярность масштабирования
- практически локальная система
- распределенность обеспечивается только дополнительными СХД



## Нет поддержки

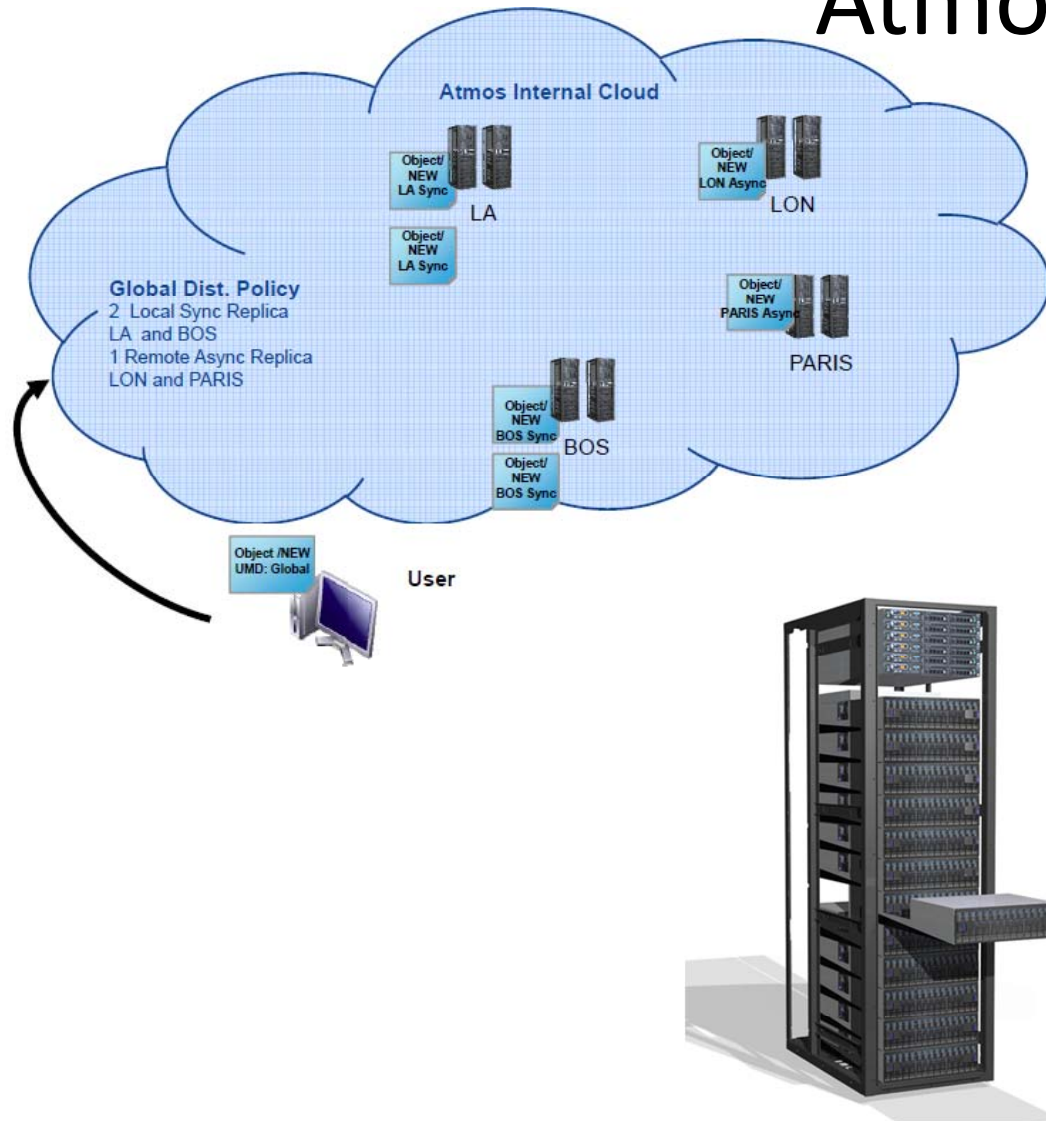
- FTP и ACL
- HTTP
- WebDAV

# Слабосвязанные системы

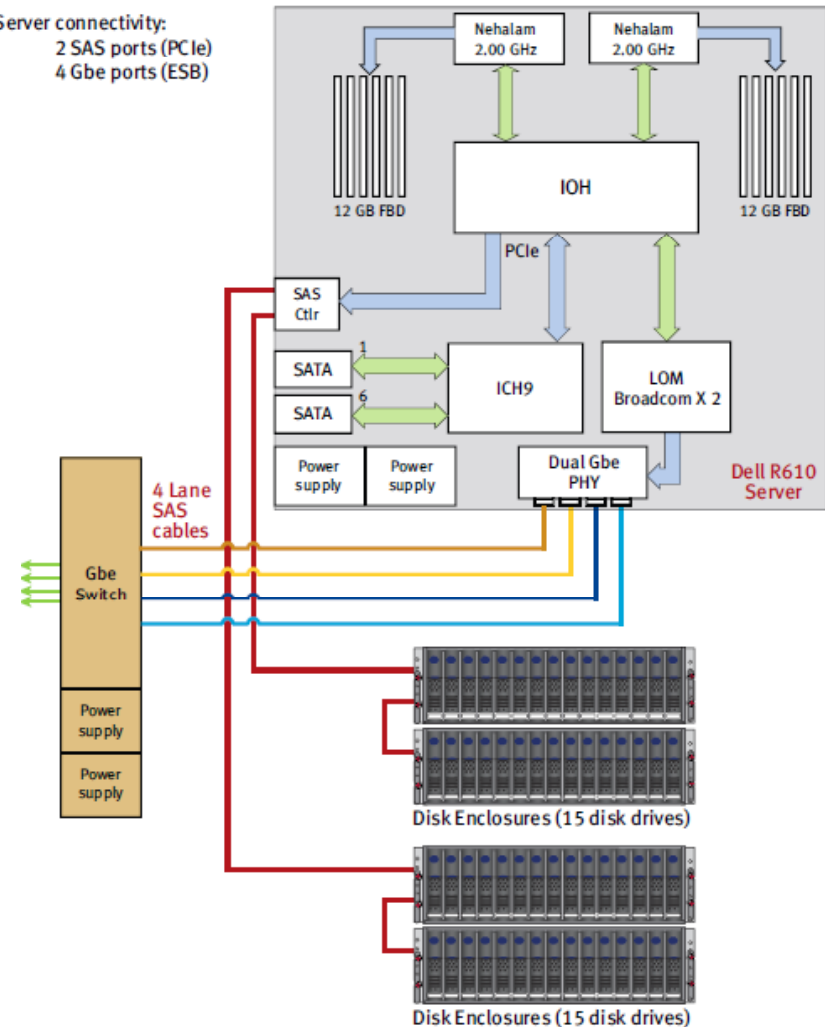
- ***слабая связанность***
  - передача данных по публичным или нестабильным каналам с ограниченной пропускной способностью
  - работа за пределами ЦОД
- ***гетерогенность***
  - унаследованные аппаратные компоненты
  - различное программное окружение
- ***хранение неструктурированных данных***
  - файлы, а не данные в РБД
- ***эластичное развитие (масштабирование)***
  - улучшение характеристик без принципиального изменения архитектуры
  - плавное увеличение объемов дискового пространства
- ***открытость для развития***
  - универсальный (общедоступный) протокол доступа к системе
  - возможность интеграции в существующие сервисы



# Слабосвязанная система EMC “Atmos”

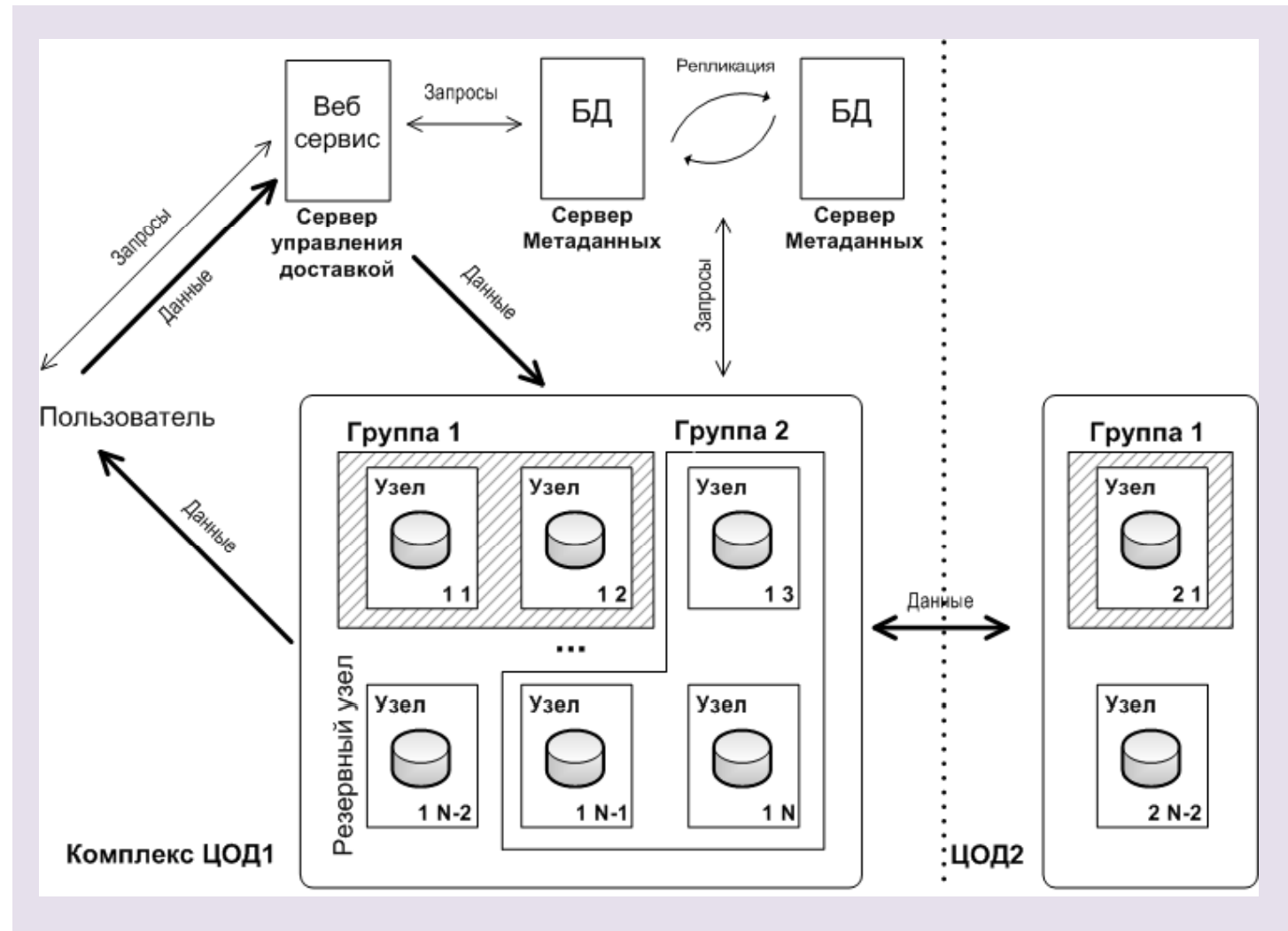


Server connectivity:  
2 SAS ports (PCIe)  
4 Gbe ports (ESB)



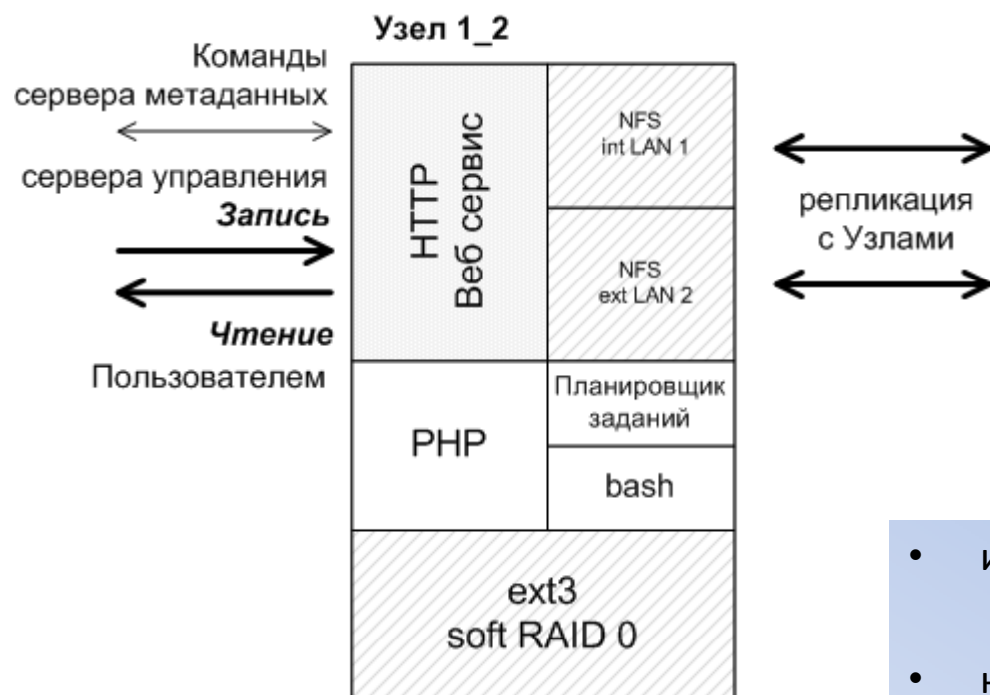
# Архитектура системы

- Открытые протоколы передачи данных:
  - для всей системы **HTTP**
  - между узлами **NFS**
- Ввод/ вывод файлов поверх **HTTP** с помощью программных компонент
- Открытая БД **PostgreSQL**





# Узел хранения



## Загрузка файла, запрос 1/3:

```
PUT /newfolder1/adk45x_r4324.jpg HTTP/1.1
accept: */*
date: Mon, 02 Aug 2010 18:02:34 GMT+3
content-type: application/octet-stream
host: 172.16.34.51
content-length: 18471
x-access-key: _N45$0KL4sdMpDj_hI8ZRnE=
content-md5: tv6qMXw6eTuumbd40yndvJ==
```

## Загрузка файла, ответ узла:

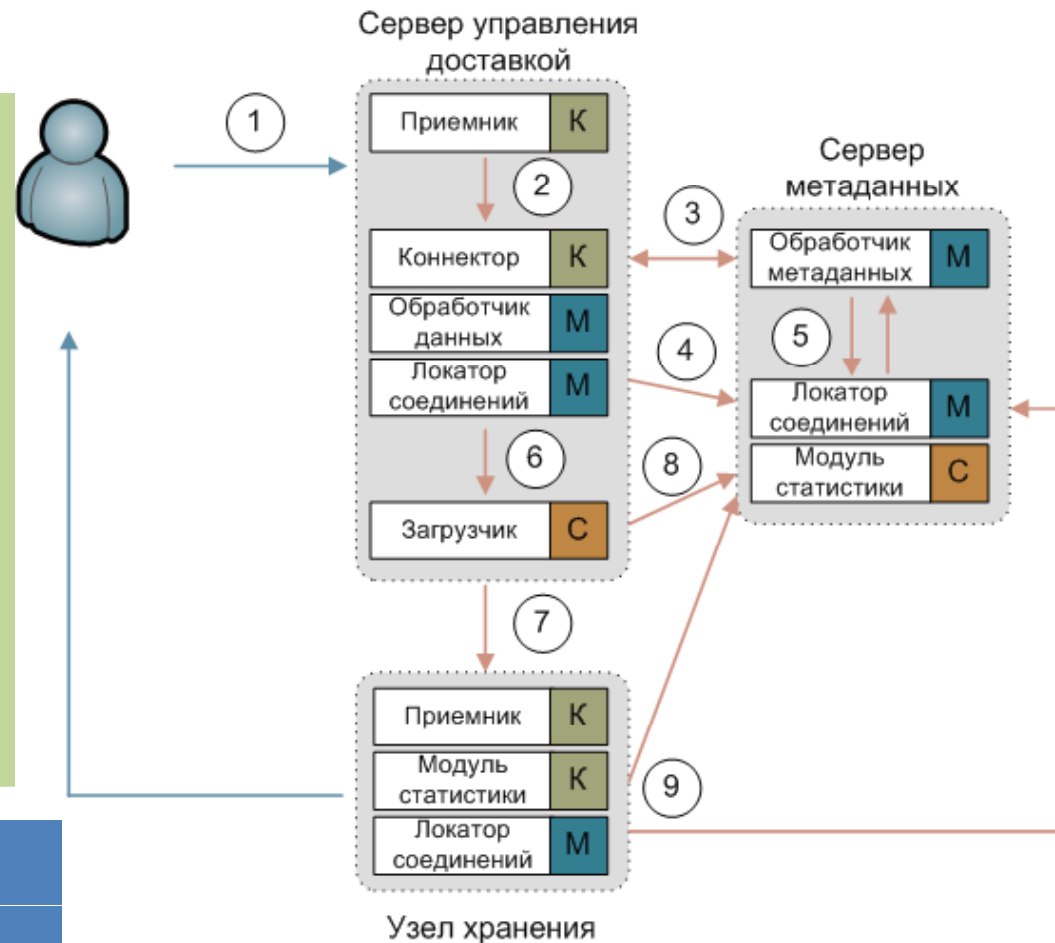
```
HTTP/1.1 200 OK
Date: Mon, 02 Aug 2010 18:02:34 GMT+3
Server: Apache
Content-Length: 0
Connection: close
Content-Type: text/plain;
charset=UTF-8
```

- именование:
  - **node(ЦОД)\_(номер).domain.name**
- нет резервирования данных внутри узла:
  - резервные узлы внутри комплекса
- 2 сетевых интерфейса:
  - внутренний  $\geq 1\text{Gb/s}$
  - внешний 100 Mbit/s
- передача данных:
  - **чтение** пользователем напрямую
  - **запись** сервером управления

# Сервер метаданных

## Основные функции:

- Взаимодействие с серверами управления доставкой
- Хранение ссылок на файлы
- Составление таблицы узлов данных по рейтингу на основе загрузки узла
- Динамическая выборка ссылок



| Параметры узла      | Весовой коэф-т |
|---------------------|----------------|
| Загрузка процессора | 0,15           |
| Загрузка канала     | 0,75           |
| Заполнение диска    | 0,1            |

- 10 мегабайт метаданных – 1 Гб полезных данных
- 4 запроса в БД метаданных для выборки файла

# Управление количеством реплик

| Параметр                        | Балл<br>$W_k$ | Вес.<br>коэф<br>$w_k$ |
|---------------------------------|---------------|-----------------------|
| Популярность файла $P$          | 4             | 0,45                  |
| Степень важности $I$            | 2             | 0,22                  |
| Тип файла $T$                   | 2             | 0,22                  |
| Свободное место<br>носителя $F$ | 1             | 0,11                  |

$W_k$  - баллы экспертной оценки  
тонкой настройки системы

$$w_k = \frac{W_k}{\sum_k W_k}$$

расчета весового  
коэффициента  $w_k$   
для  $k$ -ого параметра

$N_{add}$  – наибольшее число реплик,  
разрешенное в системе

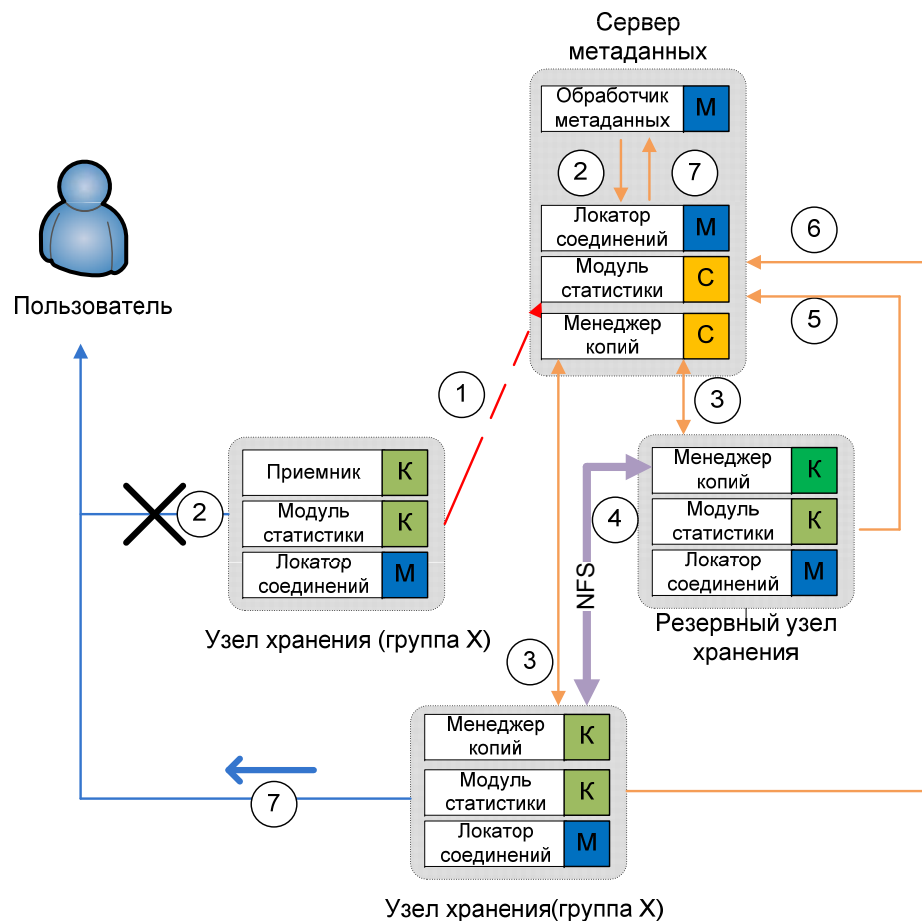
$P, I, F, T_{max}$  – максимальные  
значения показателей

Расчет рекомендуемого количества реплик для заданного файла:

$$N_i = N_{min} + N_{add} * (w_{pop} \frac{P_i}{P_{max}} + w_{imp} \frac{I_i}{I_{max}} + w_{free} \frac{F_i}{F_{max}} + w_{type} \frac{T_i}{T_{max}})$$

$N_{min}$  – минимальное число реплик (три)

# Восстановление данных



## Особенности алгоритма:

- прямое копирование данных через внутреннее высокоскоростное соединение **int LAN 1** по протоколу **NFS**
- минимальное взаимодействие с сервером метаданных
  - блокировка поврежденного узла (**3 запроса**)
  - изменение таблицы предпочтительности узлов (**2 запроса**)
  - ввод в эксплуатацию узла (**3 запроса**)
- Корректировка только записей об изменении IP адреса узла и его имени.
- Отсутствие необходимости аппаратного RAID контроллера на узлах

### Группа X

А RW node1\_2.domain.name  
А RW node1\_8.domain.name  
А RW node1\_6.domain.name

Исключение узла  
Восстановление данных

### Группа X

А RW node1\_2.domain.name  
А RW node1\_8.domain.name  
**node1\_7.domain.name**

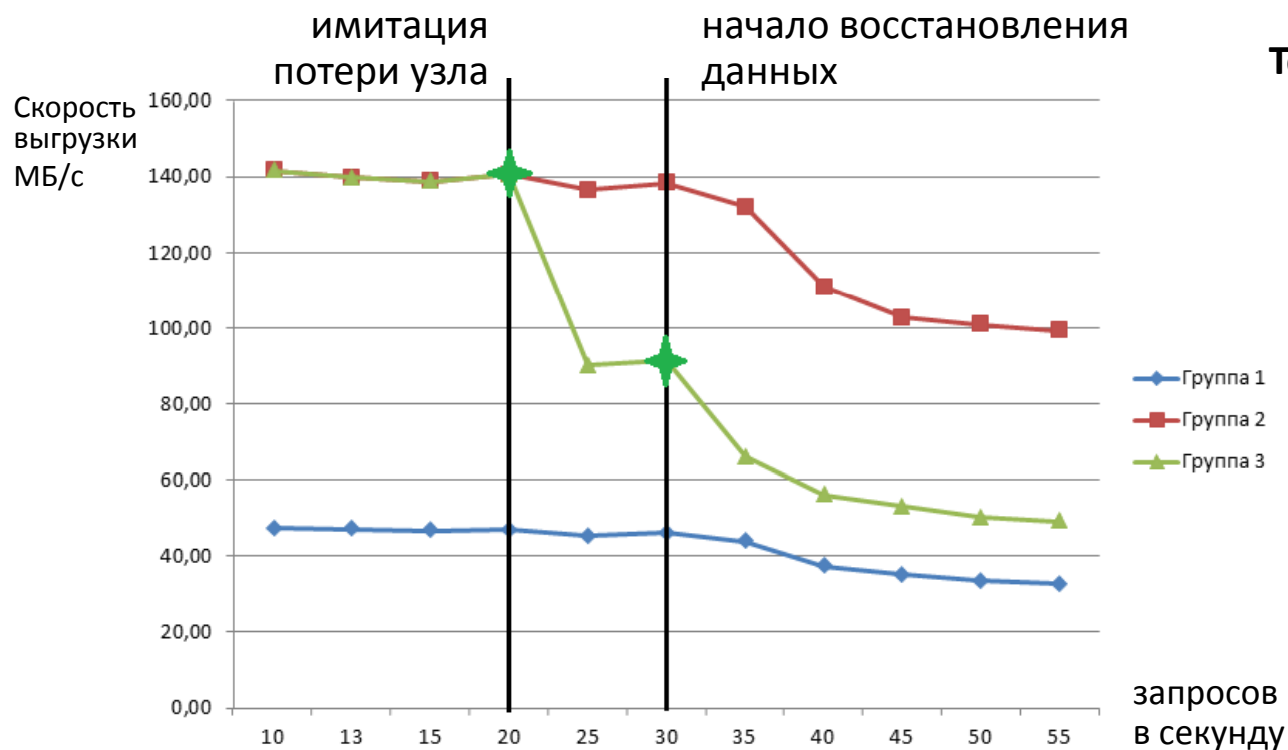
Изменение ключей node1\_6 на node1\_7

### Группа X

А RW node1\_2.domain.name  
А RW node1\_8.domain.name  
А RW node1\_7.domain.name

# Восстановление данных

| Тип           | Объем дисков, МБ | Кол-во дисков, шт. | Уровень RAID | Пропускная способность NAS, МБ/с | Время восстановления, ч | Нормированное время, ч |
|---------------|------------------|--------------------|--------------|----------------------------------|-------------------------|------------------------|
| SAS           | 300              | 4                  | 5            | 109                              | 4,7                     | 7,8                    |
| SATA          | 1500             | 4                  | 5            | 94                               | 31,5                    | 10,5                   |
| Узел хранения | 1500             | 1                  | нет          | 81                               | 20,8                    | 20,8                   |
|               | 1500             | 2                  | 0            | 97                               | 18,2                    | 18,2                   |



## Тестовые группы:

- Группа 1 – 1 узел
- Группа 2 – 3 узла
- Группа 3 – 3 узла / восстановление данных

## Экспериментальный запрос:

синхронный  
15 файлов x 100 кб

# Результаты работы

- Разработана архитектура распределенной системы архивного хранения данных, обеспечивающая, в отличие от централизованных систем:
  - практически неограниченное, плавное масштабирование
  - использование гетерогенных, экономически доступных аппаратных компонент
- Разработан программный метод адаптации системы в части размещения и доставки контента:
  - повышение скорости доставки
  - снижение затрат архивной памяти
- Предложен авторский алгоритм управления количеством реплик, обеспечивающий достижение требуемой доступности хранимых данных
- Разработан программный метод восстановления данных с отказавших узлов системы, исключающий безвозвратную потерю данных
- Реализована и опытно эксплуатируется распределенная система хранения файлового контента в области архивного хранения и представления видеоданных:
  - плавное расширение и соответствующее снижение единовременных капитальных затрат
  - сохранение работоспособности системы при отказах узлов



**СПАСИБО ЗА ВНИМАНИЕ!**