Предпосылки для проекта Data Lake на российских ресурсах

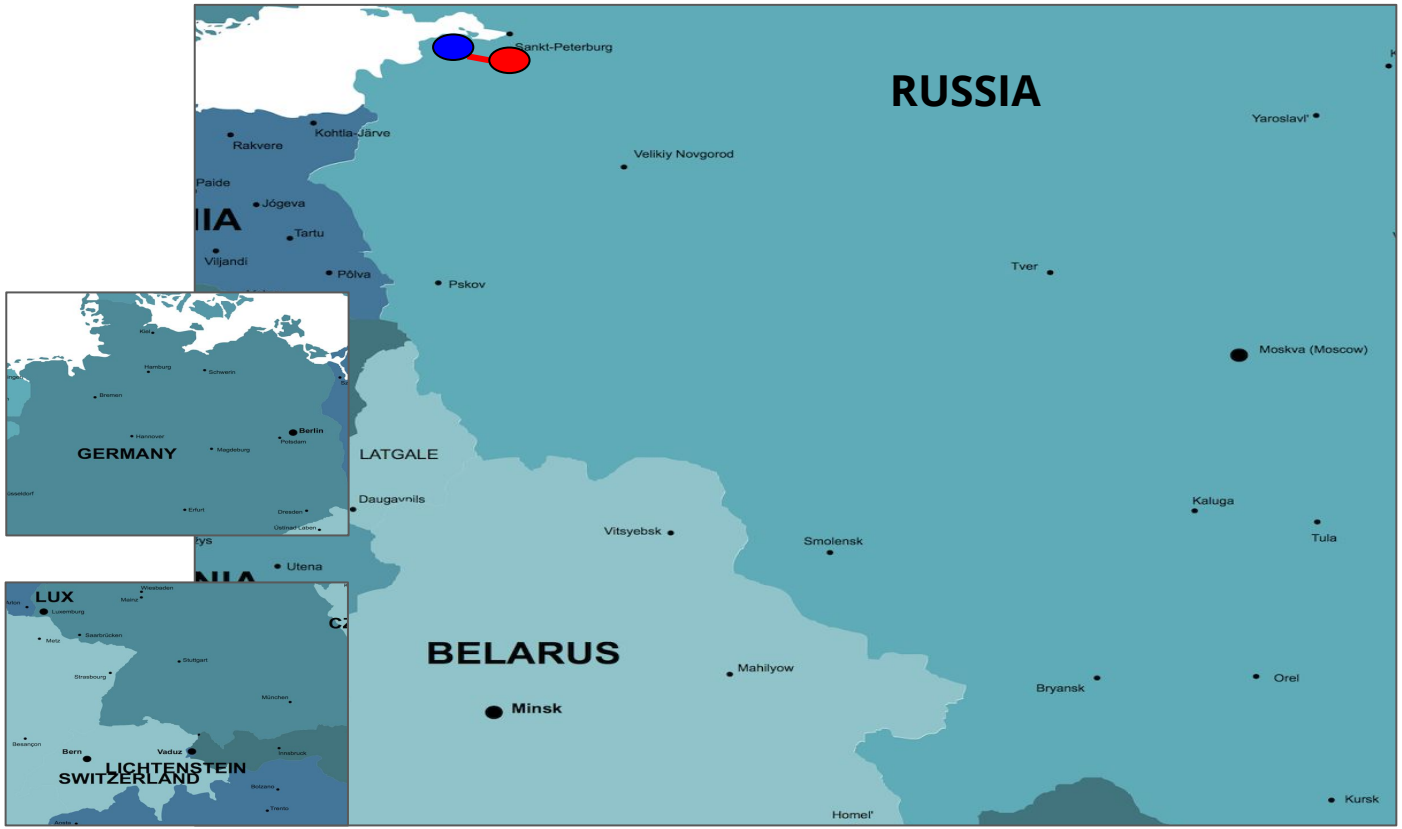
История развития проекта FedStor в России

История участников

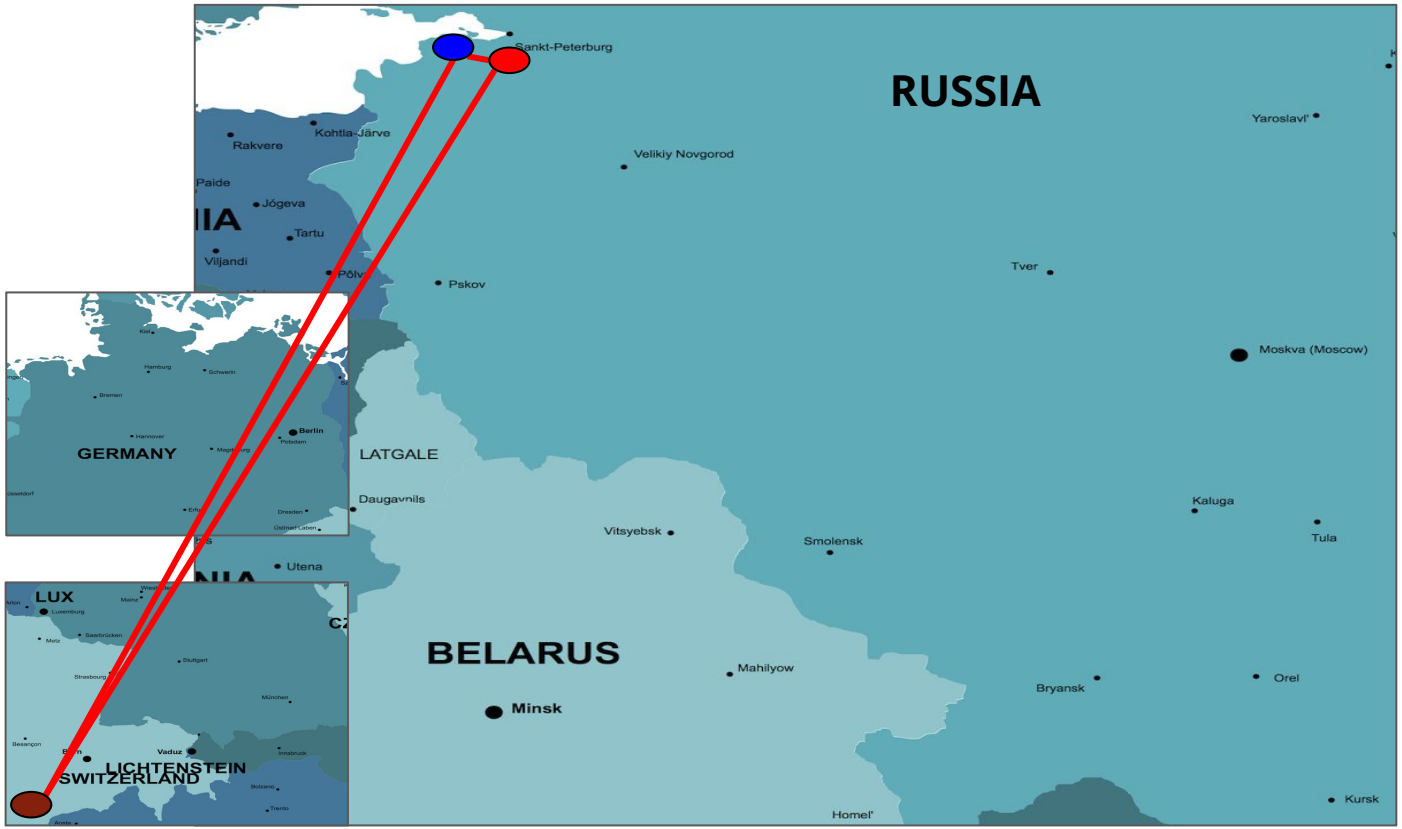
История структуры

История тестов

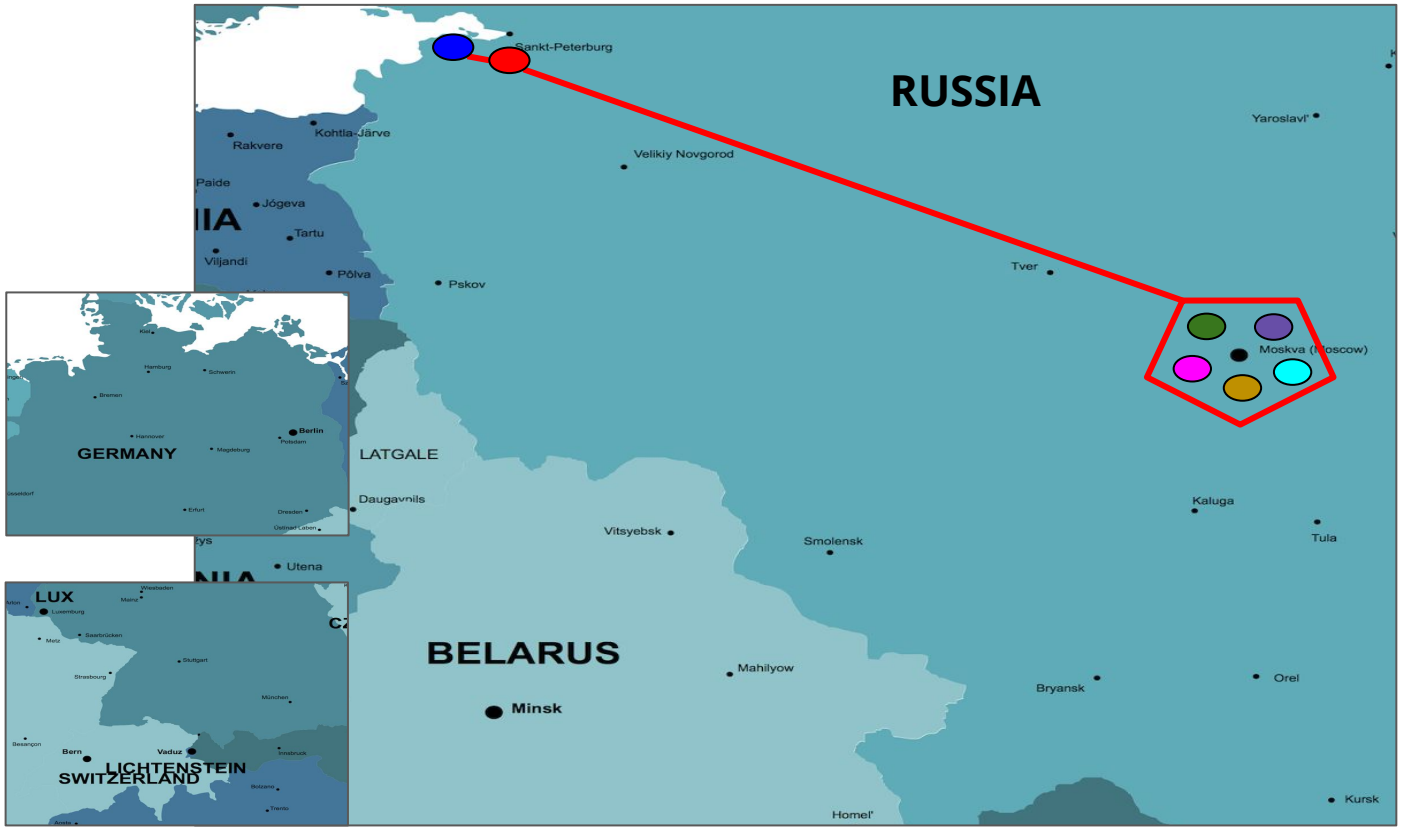
Развитие в картинках (конец 2015 г.)



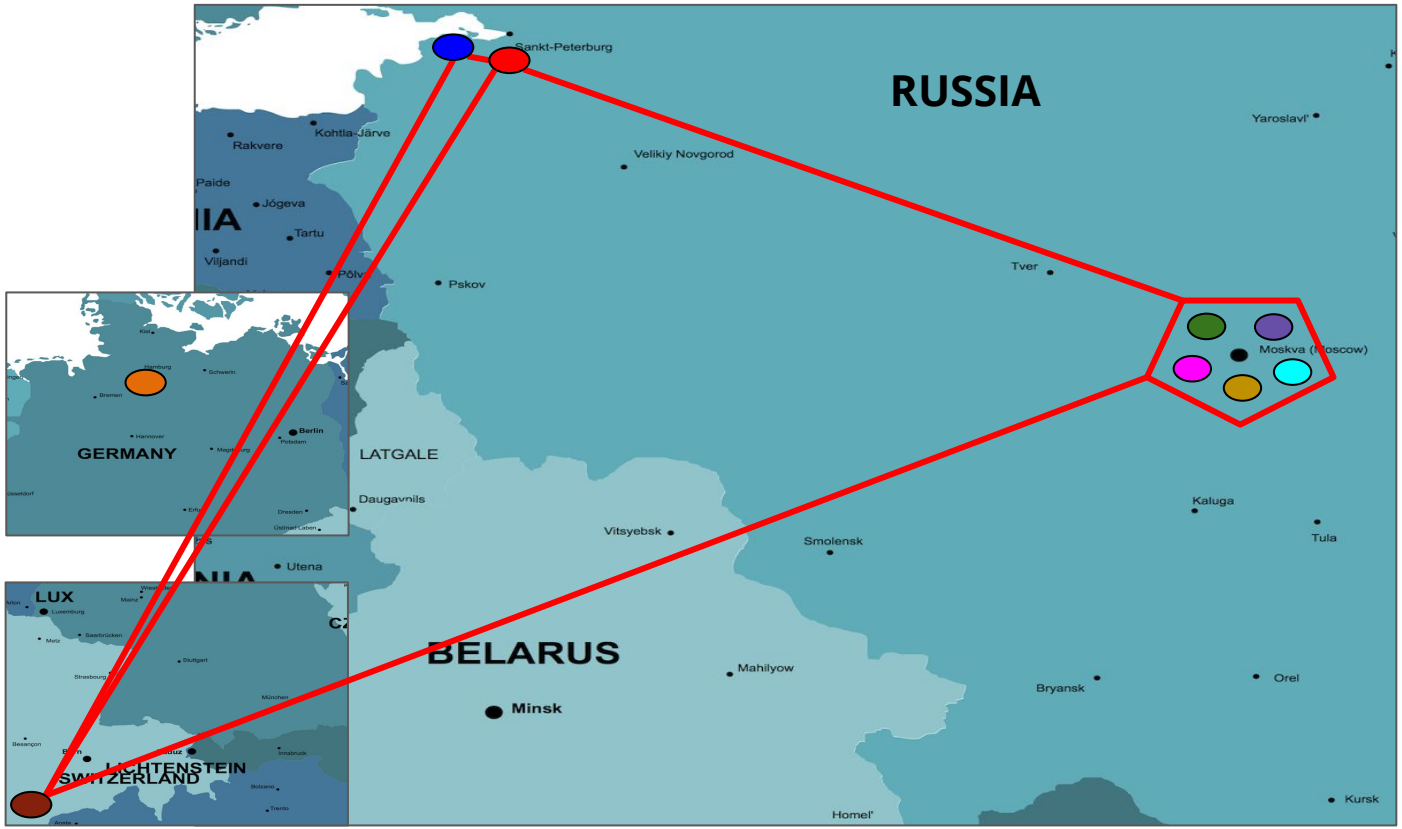
Развитие в картинках (начало 2016 г.)



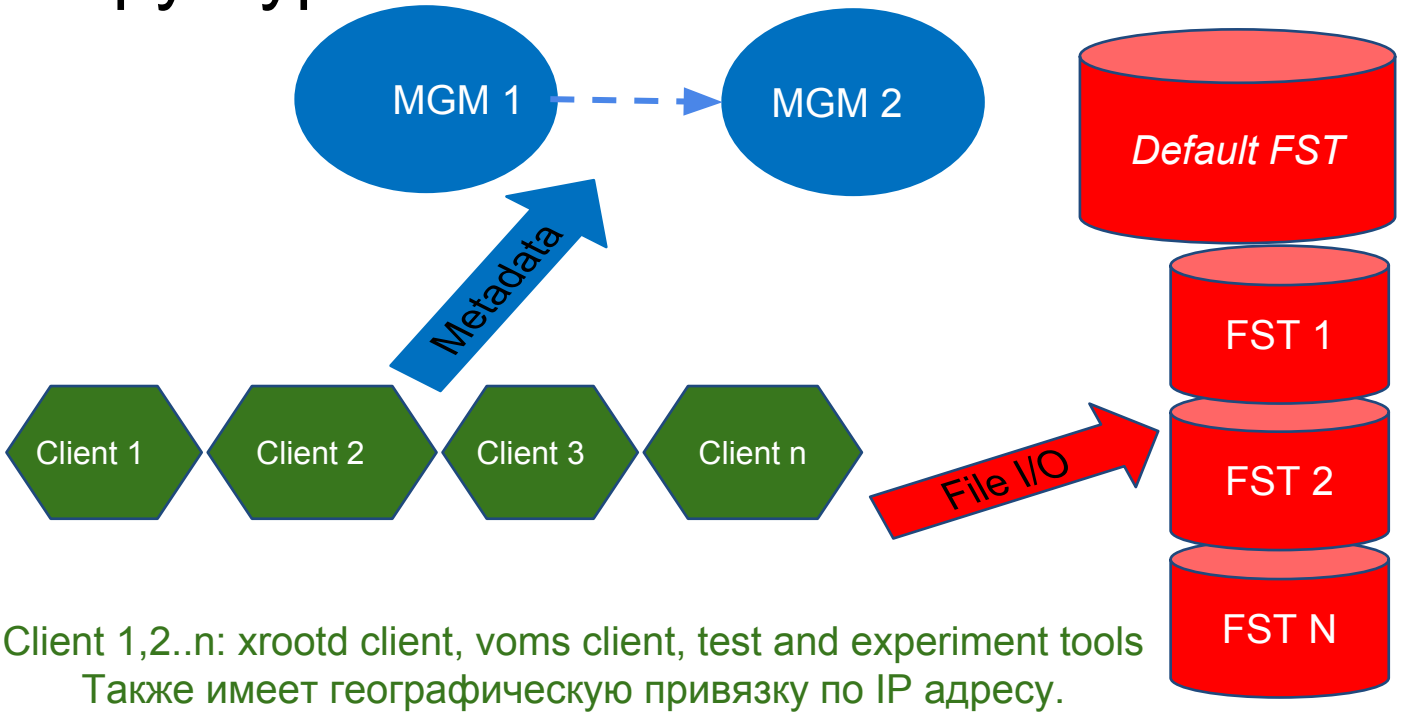
Развитие в картинках (2016 – 2017 гг.)



Развитие в картинках (2017 г.)



Структура



Client 1,2..n: xrootd client, voms client, test and experiment tools
Также имеет географическую привязку по IP адресу.

Созданы сценарии разворачивания каждого элемента

MGM 1: Точка доступа, пространство имён, метаданные

MGM 2: Резервная копия MGM1 только для чтения

FST 1,2..N: файловый сервер, с географической привязкой

Default FST: Файловый сервер по умолчанию. В одном из сценариев сохраняет реплики всех данных.

Тесты

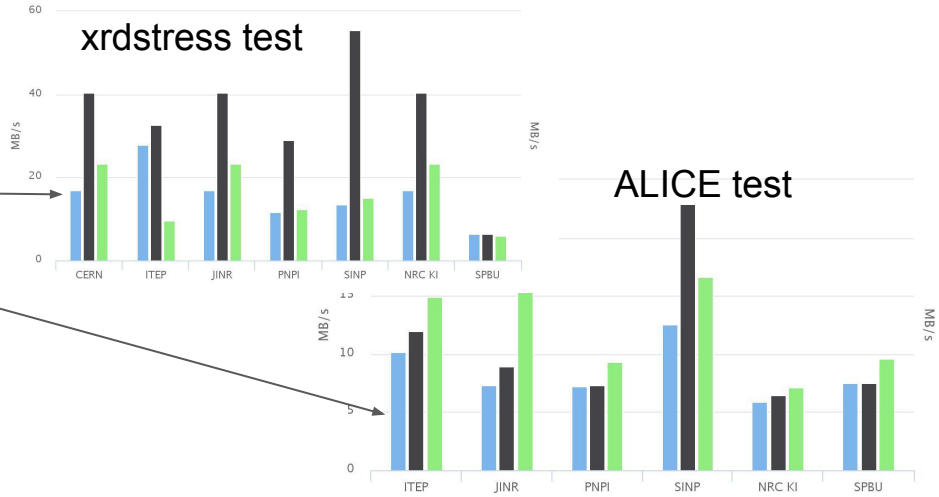
Синтетические тесты

- Bonnie++: file and metadata I/O test for mounted file systems (FUSE)
- xrdstress: EOS-bundled file I/O stress test via xrootd protocol

Тесты на реальных данных:

- ATLAS test: standard ATLAS TRT reconstruction workflow with Athena
- ALICE test: sequential ROOT event processing (thanks to Peter Hristov)

- Все тесты был автоматизированы на этапе выполнения и, частично, на этапе визуализации.
- xrdstress был сильно доработан в процессе исследований.
- тесты экспериментов были специально подготовлены для оценок производительности распределённого хранилища.
- Bonnie++ был интересен изначально, дабы показать зависимость метаданных от положения MGM сервера.



Сценарии

Желаемая политика:

- При чтении обращаться к ближайшей реплике;
- При записи создавать две реплики: на ближайшем T2 и на указанном T1.

Реализация в EOS:

- *Гибридная* политика позволяет располагать первую реплику на ближайшем узле, а вторую на случайном;

В настоящий момент протестировано три политики репликации:

1. Одна реплика без геотэгов: размещение на случайном узле;
2. Одна реплика с геотэгами: размещение на ближайшем узле, совпадающем по геотэгу с клиентом; если совпадений не найдено, размещение на узле по умолчанию;
3. Две реплики с геотэгами: первая на ближайшем узле, вторая на случайном узле.

Мониторинг: Perfsonar + MadDash + Ganglia

Throughput

> 100Mb/s

10Mb/s ÷ 100Mb/s

< 10Mb/s

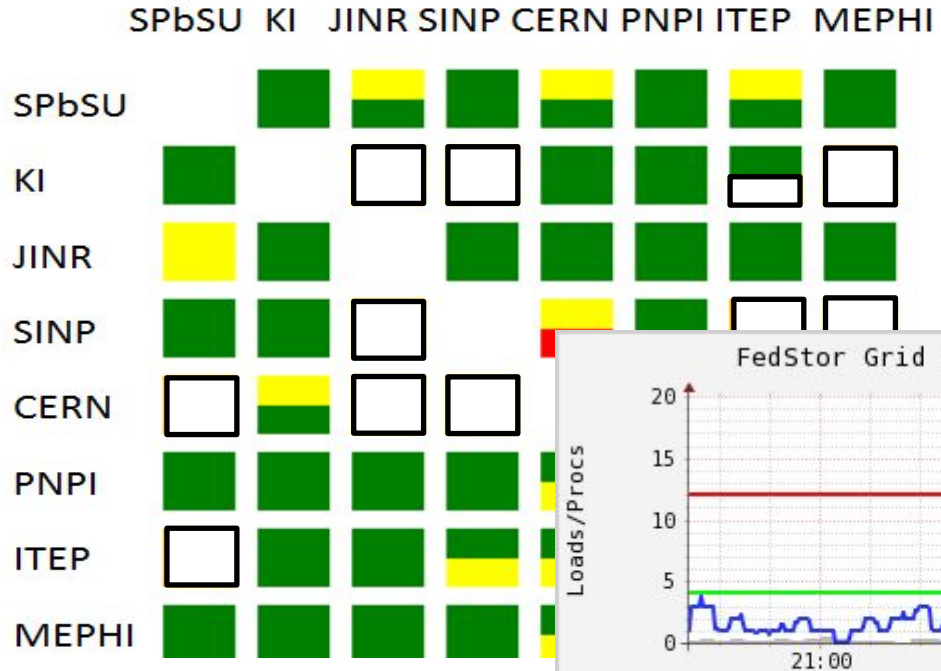
No data

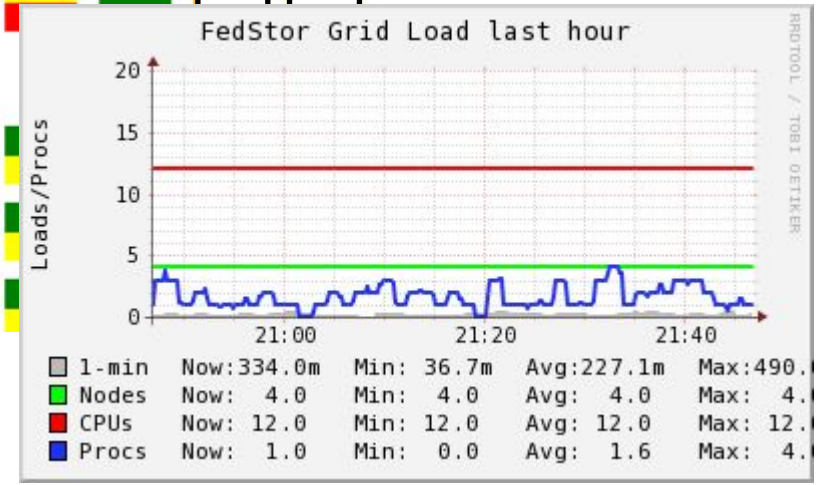
X

Y

Throughput from Site X to Site Y

Throughput from Site Y to Site X





Предпосылки к следующему этапу

1. Накоплен большой опыт работы в EOS и dCache в распределённых конфигурациях.
2. Существуют отработанные схемы развертывания с выбором политик чтения-записи и репликации данных.
3. Существует развёрнутая система мониторинга.
4. Существуют отработанные системы автоматизированных тестов.