Могут ли данные управлять аппаратной конфигурацией дата-центра?

Антон Катенев

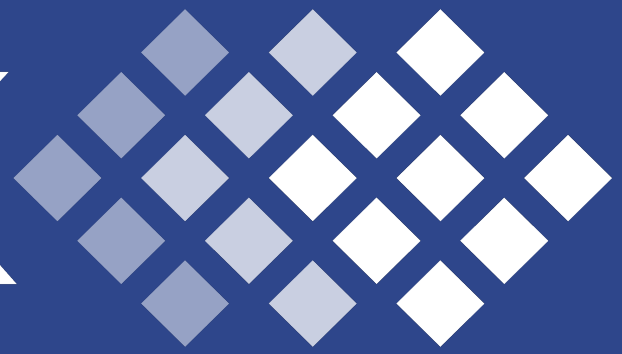
Павел Лавренко

Дмитрий Подгайный

ОИЯИ

РСК

Группа компаний



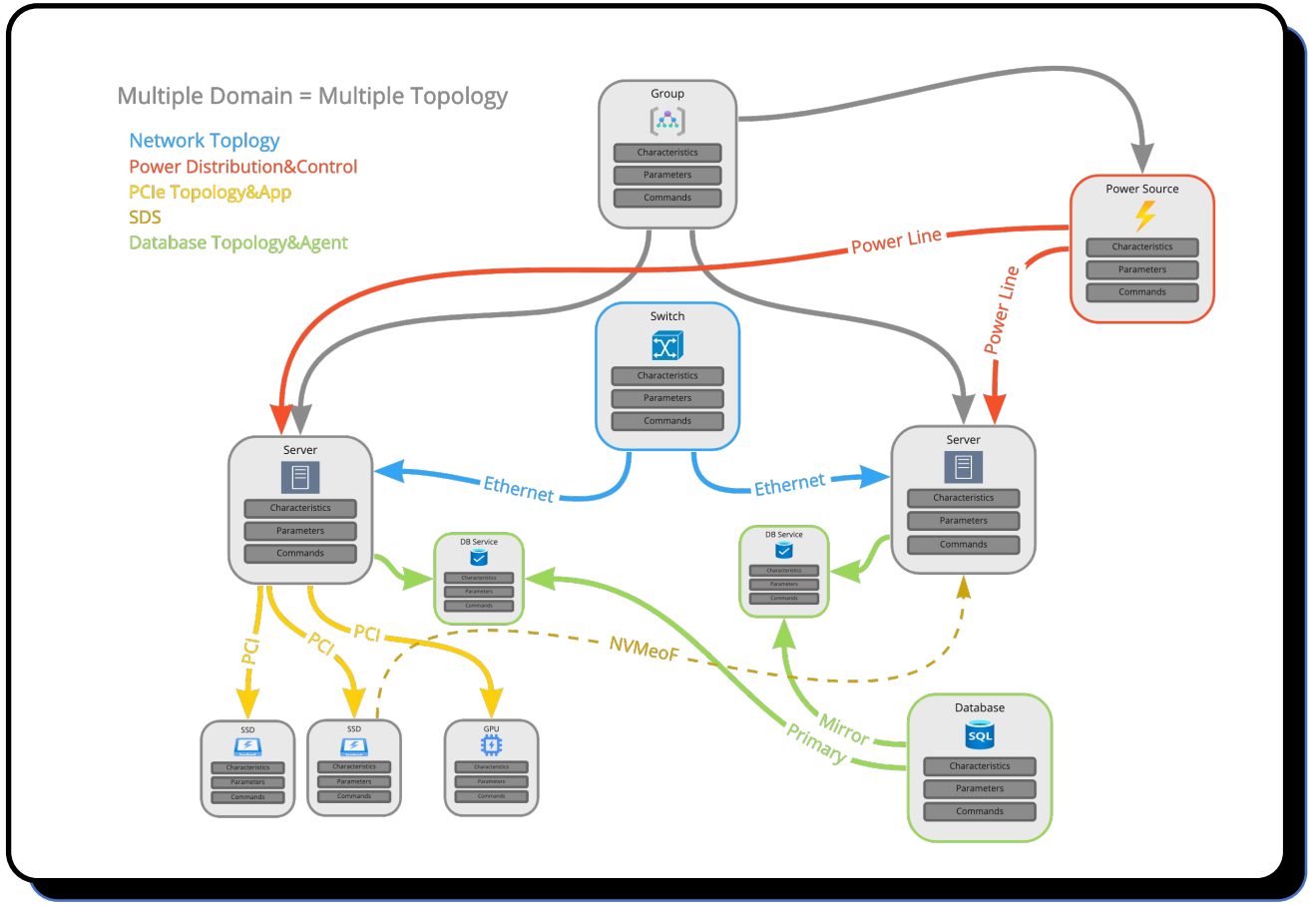
Проблемы в проектировании инфраструктуры суперкомпьютера

- Оптимальную архитектуру сложно спроектировать
- Интегрировать несколько суперкомпьютеров между собой из-за потери контекста при интеграции
- В эксплуатации всё превращается в

Мультидоменная платформа приложений управления

FOLIAGE

- Знает обо всех объектах и связях дата-центра
- Выполняет приложения управления
- Поддерживает жизненный цикл приложений управления
- Предоставляет SDK для разработчика



**Суперкомпьютер «Говорун»
в Объединенном институте ядерных
исследований в г. Дубна**

Большие данные в физике высоких энергий

9.8 Gbps

Сырой поток данных

38 Pb

Сырых данных в год

8 Pb

После системы отбора

1400 Kb

Размер события

8 месяцев в году

Длительность эксперимента

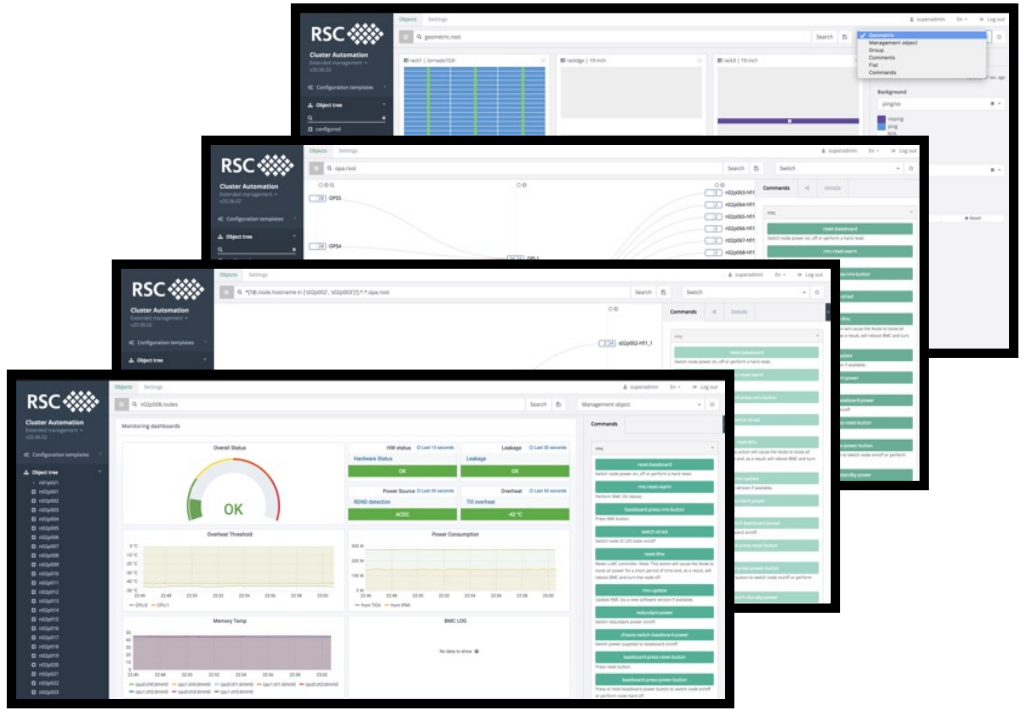
1:5–1:30

Степень сжатия

Какими же аспектами можно управлять?

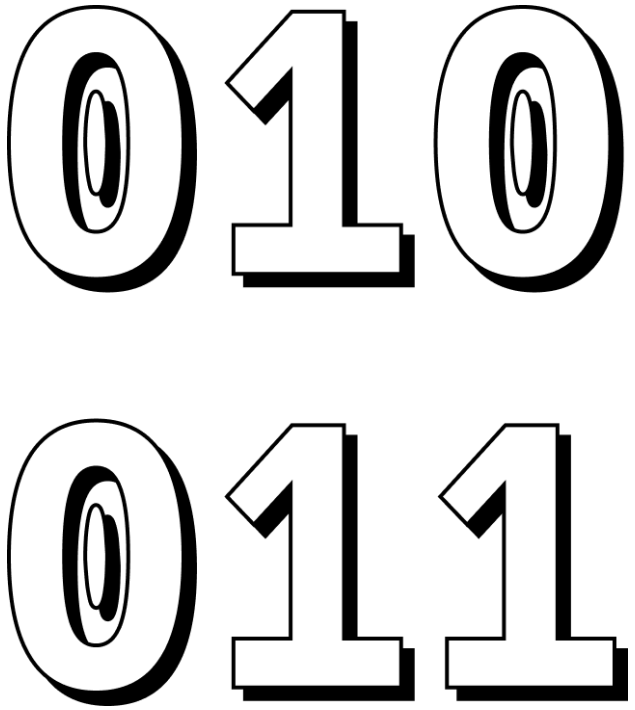
Сегодня мы затронем только управление данными

- Системами электропитания и охлаждения оборудования
- Инвентаризацией оборудования
- Сетями
- **Системами хранения данных**
- Серверами
- Операционными системами
- Конфигурациями ОС
- Облачными или контейнерными средами
- **Данными**
- Прикладными приложениями
- Шаблонами задач
- Вычислительными задачами
- Результатами расчетов



Требования к данным на жизненном цикле

- [x] Храниться в правильном месте
- [x] Использовать правильную СХД
- [x] Не занимать место зря
- [x] Храниться достаточно надежно
- [x] Быть доступными



Классическая схема схема управления

Главный архитектор

Сетевой инженер

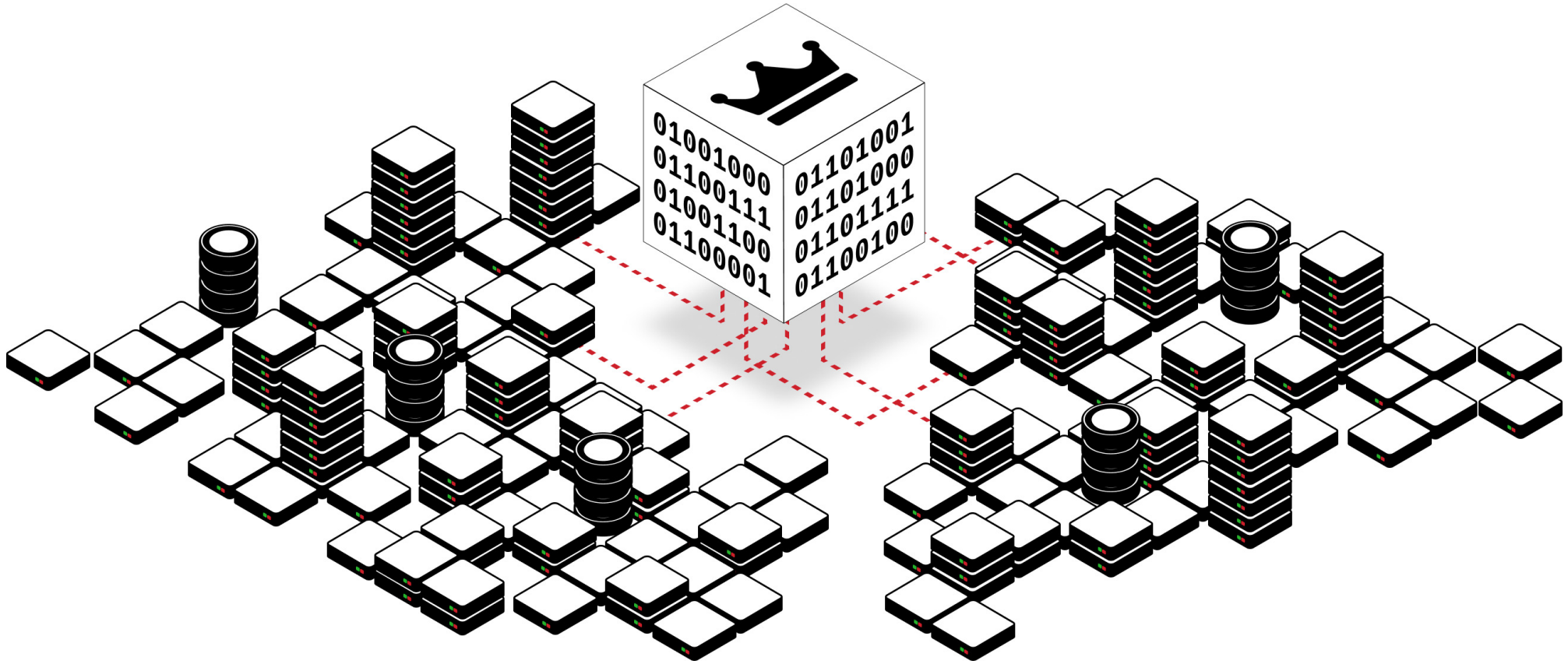
Администратор Баз данных

Администратор Серверов

Администратор СХД

DevOps

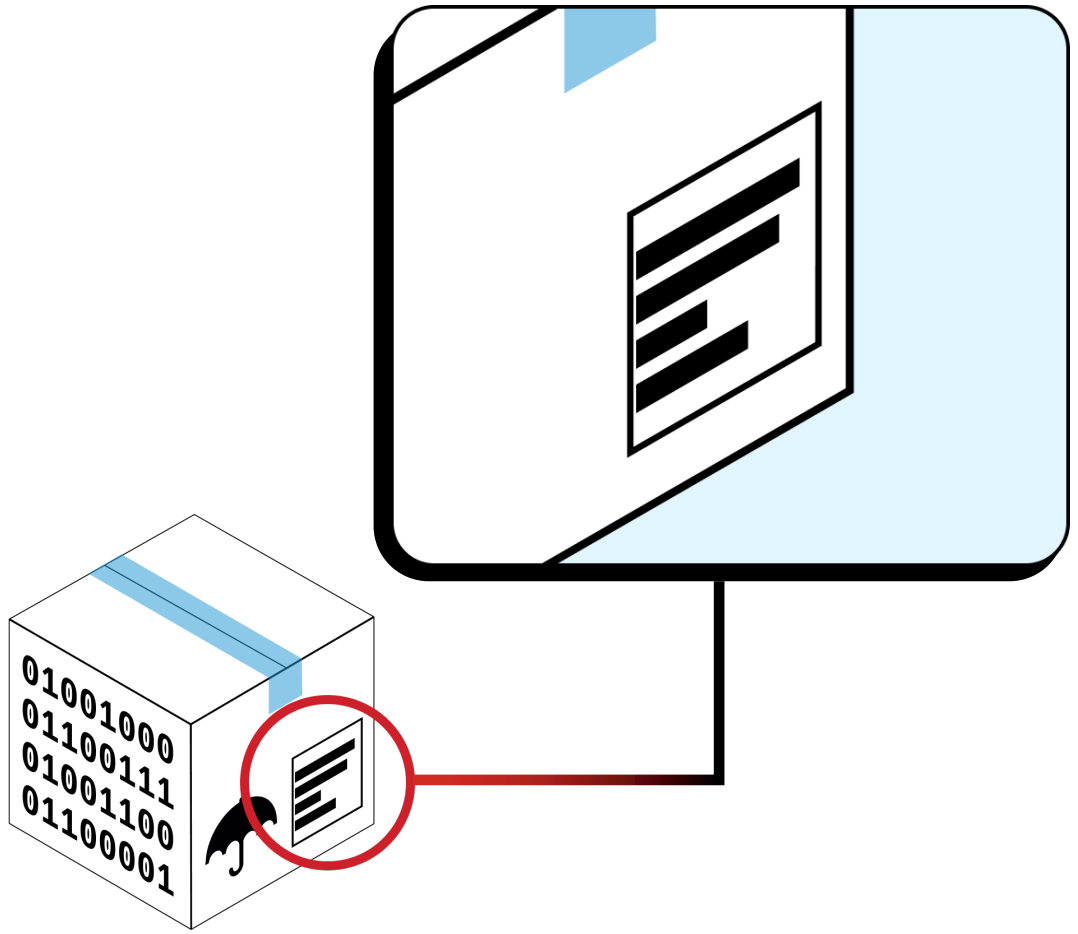
Данные могут управлять инфраструктурой



Данные в HPC пакетные

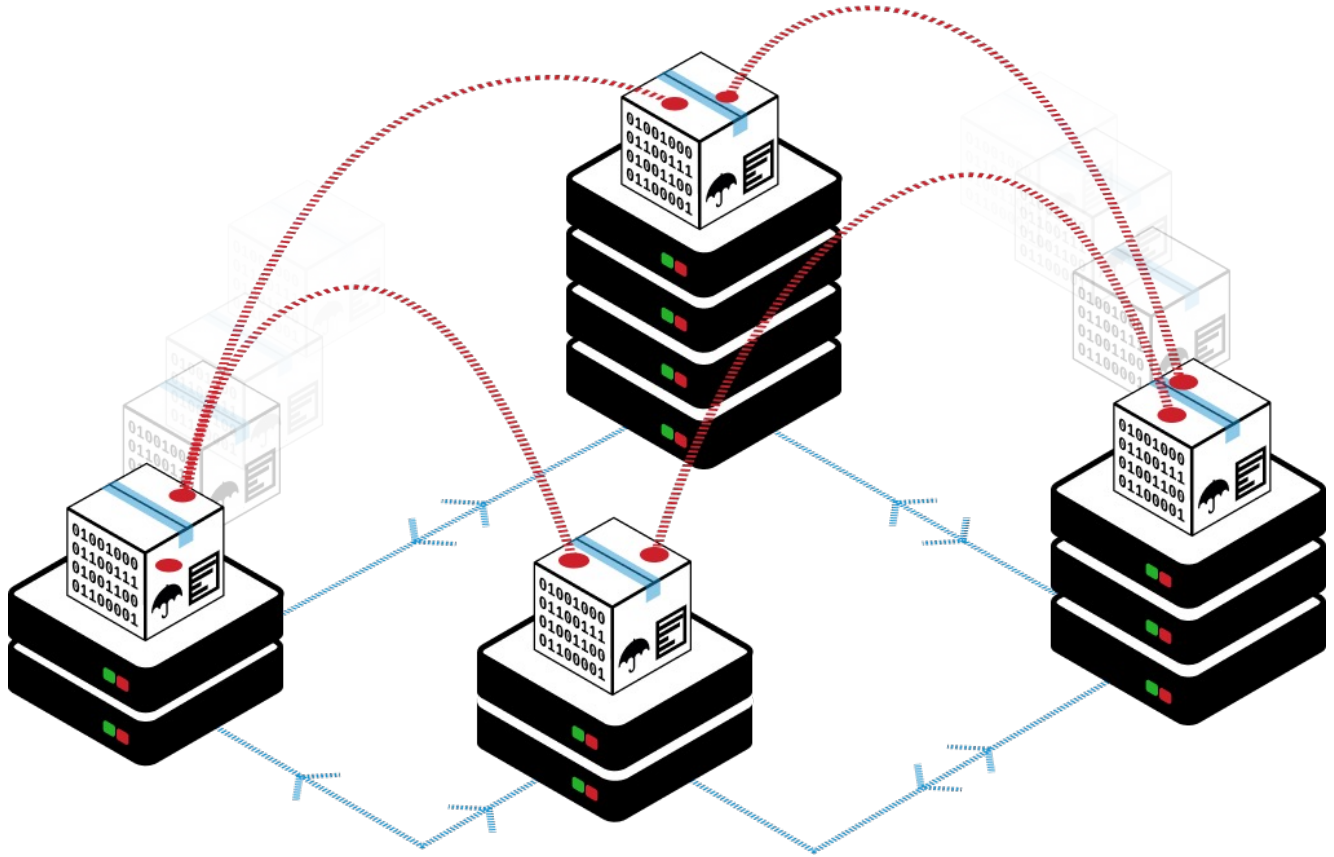
«Датасет» имеет:

- свой жизненный цикл
- свою интенсивность и частоту использования
- свои требования к СХД и надёжности и длительности хранения



Данные не остаются на месте

- Хранятся
- Перемещаются
  - Двигаются к месту обработки
  - Возвращаются «домой»
- Переносятся на более быстрое или медленное хранилище



Управление жизненным циклом данных

1 **Система Data Management**

Контейнеры для данных,
где они хранятся

2 **Система Data Moving**

Механизмы движения

3 **Система декларативного управления ”Smart Data”**

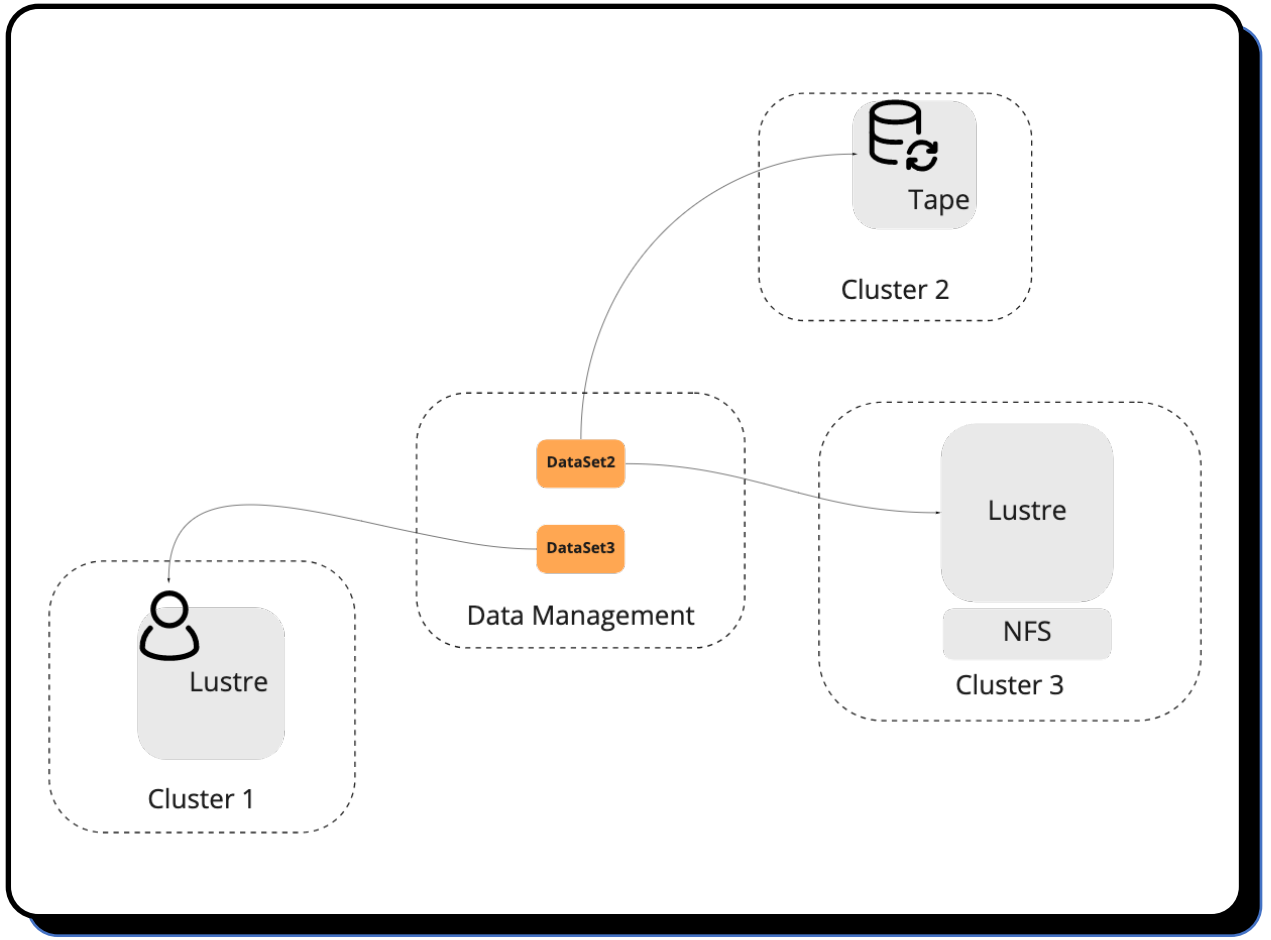
Правила автоматического движения

4 **Система хранения «По запросу»**

Динамически создаваемые пункты назначения для данных

Подсистема Data Management

- Двигать к пользователю
- Двигать к месту обработки
- Двигать к месту хранения
- Уничтожить

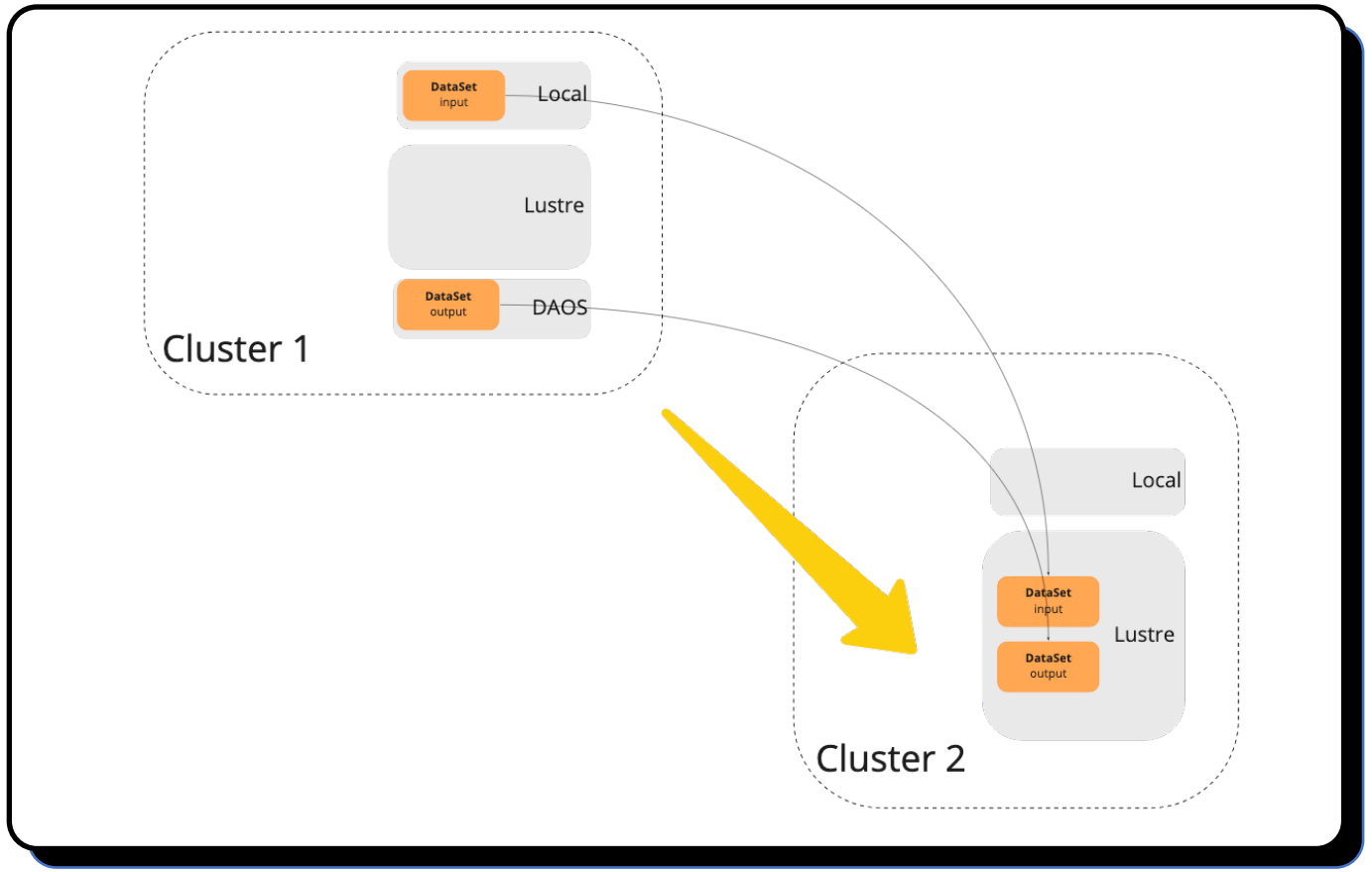


Очень часто данные «в гостях»

Где хранить?

Как получить доступ?

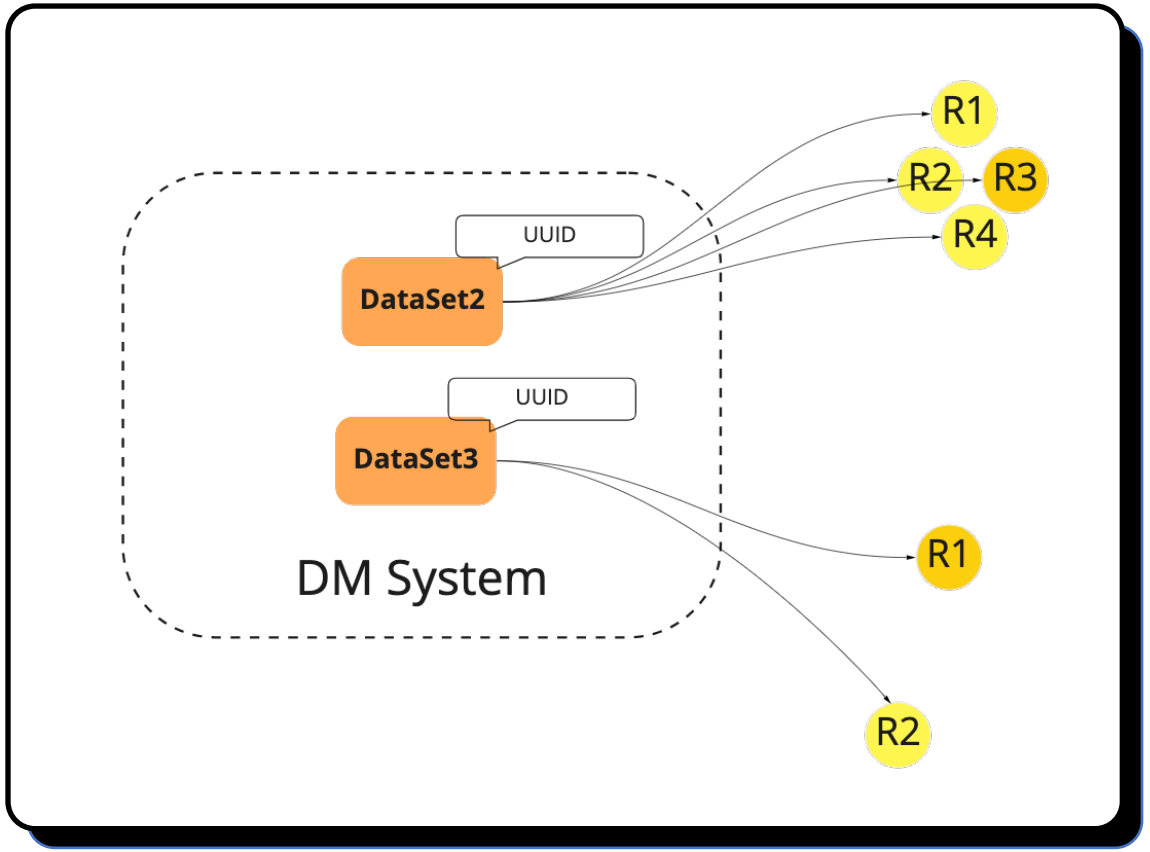
Как долго можно хранить?



Датасет

- Уникальный идентификатор
- Владелец
- Метаданные

А данные хранятся отдельно и могут иметь множество копий (реплик)

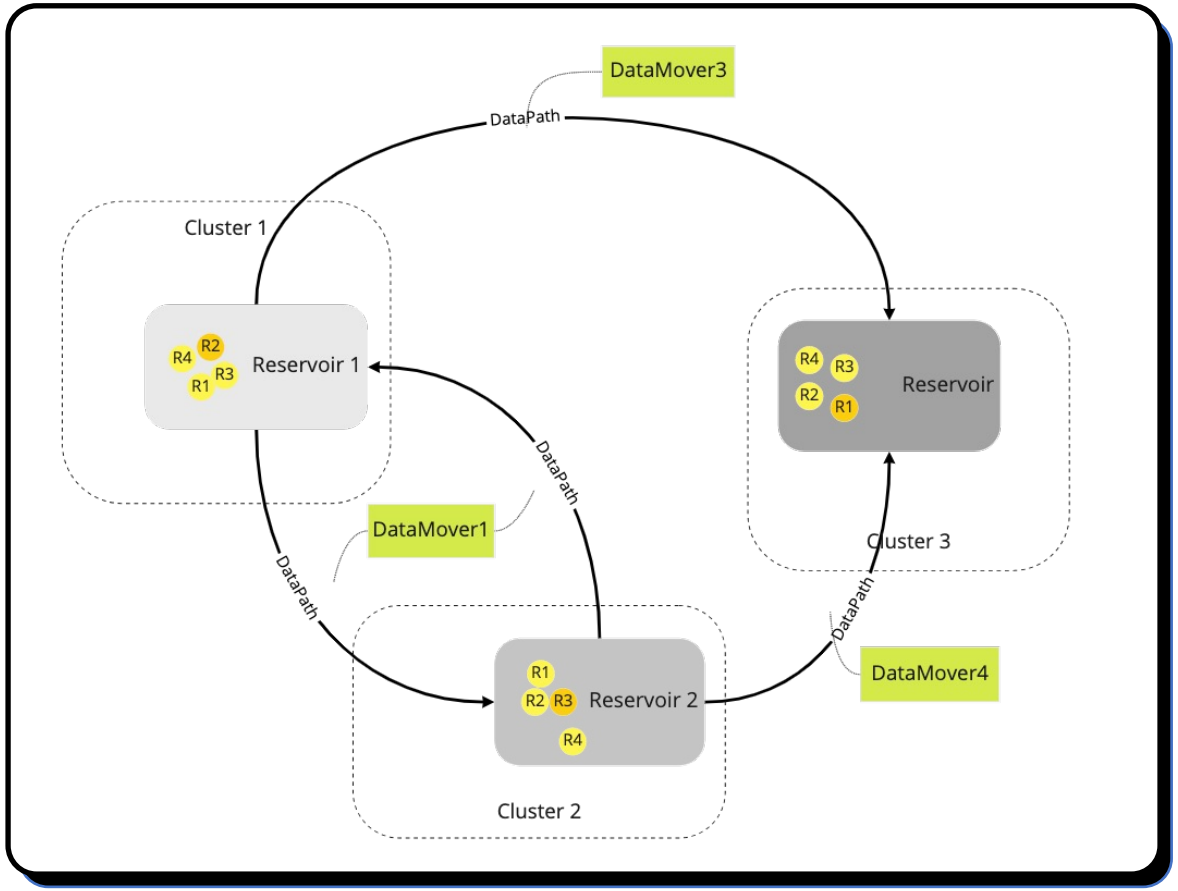


Жизненный цикл датасета

Реплики данных хранятся
в резервуарах:

- Директории
- Базы данных
- Объектные контейнеры

Данные могут реплицироваться между резервуарами разными способами с помощью дата-муверов

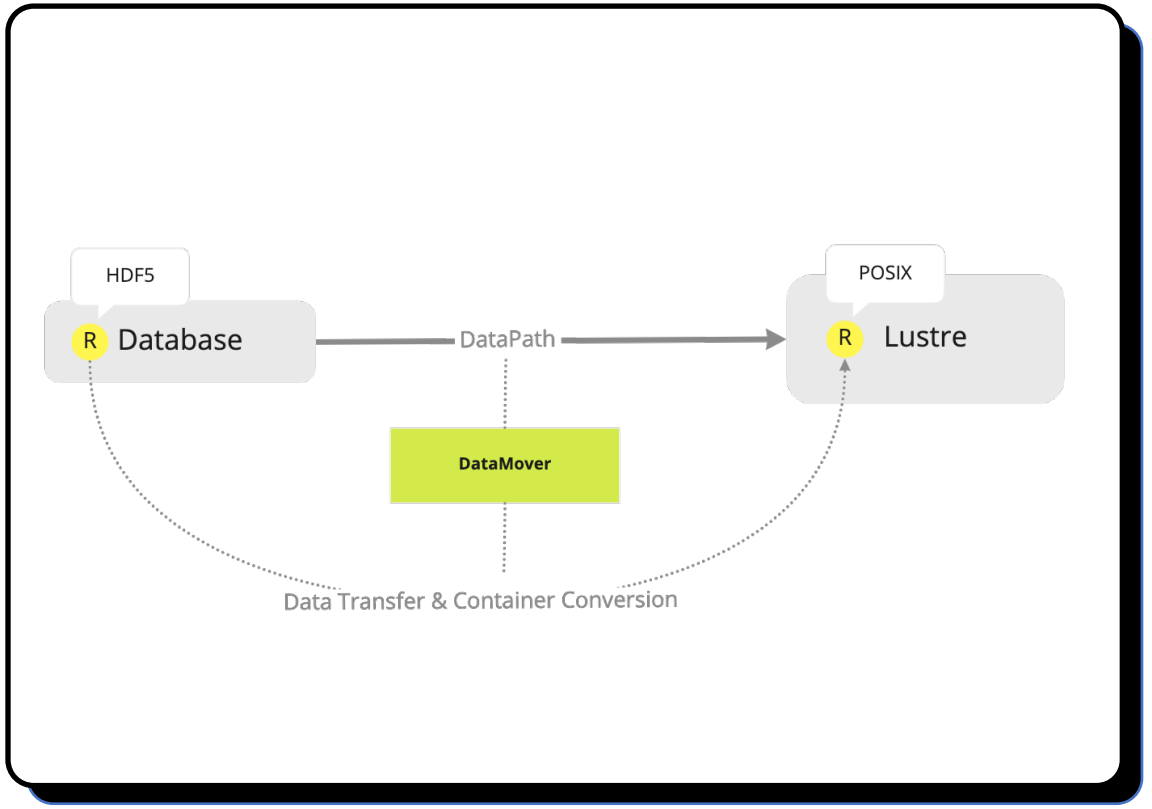


Дата-муверы

Произвольные способы перемещения

- `rsync` или `copy`
- массовое параллельное копирования
- пользовательский метод

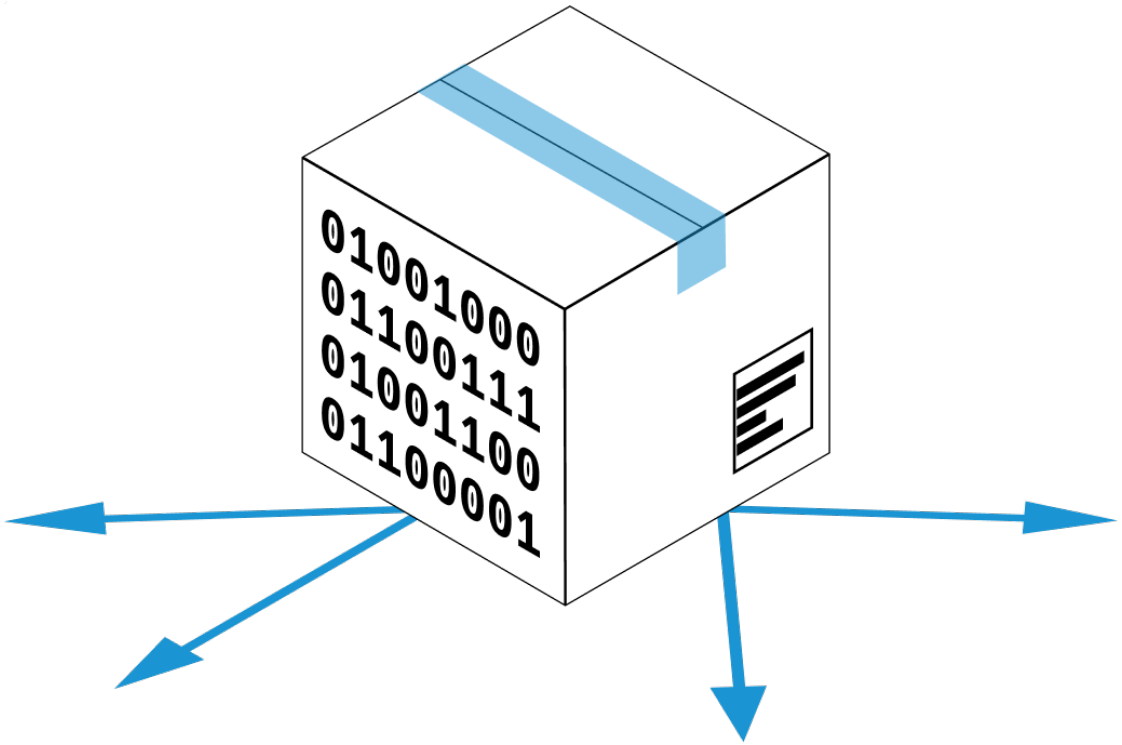
По дороге данные могут быть обработаны или вообще преобразованы



Система правил движения

Система управления пытается «выполнить» цель наилучшим образом

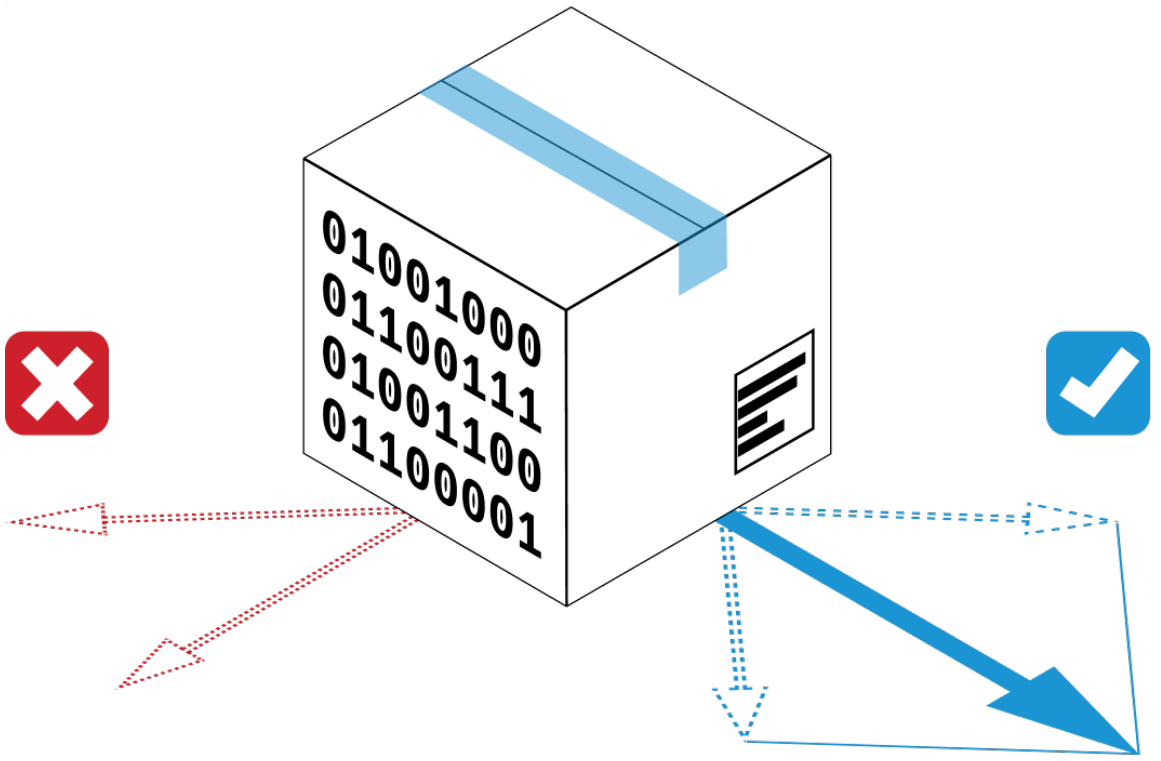
Декларативная система на основании «целей»



Система правил движения

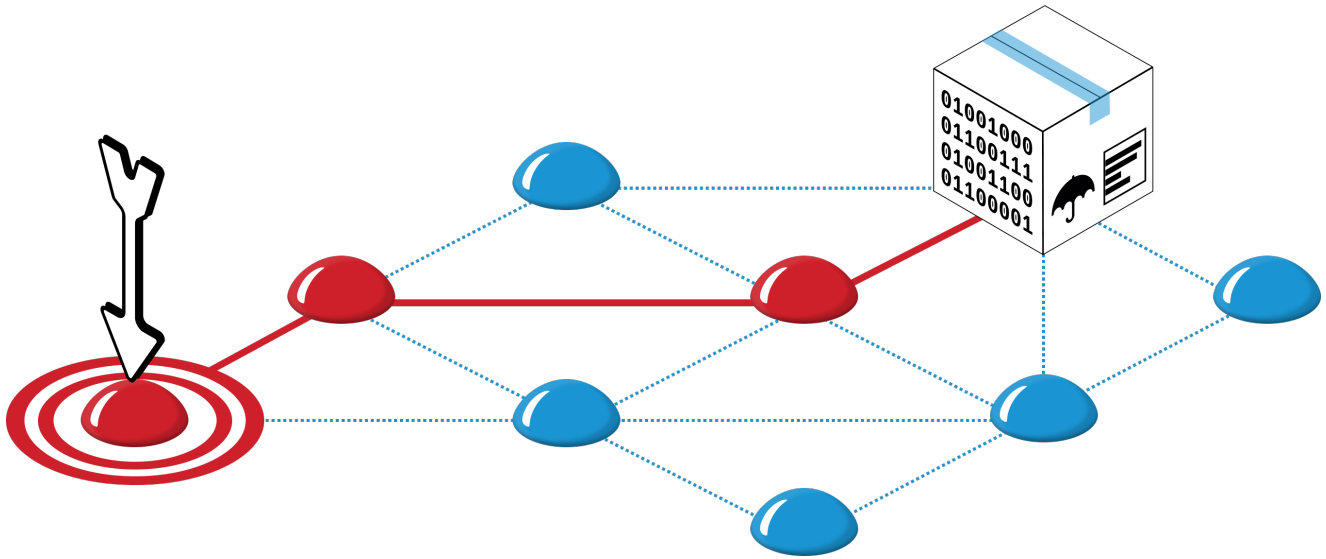
Система управления пытается «выполнить» цель наилучшим образом

Декларативная система на основании «целей»



Сейчас жизненный цикл данных выглядит вот так

1. Установка целей
2. Анализ графа потенциальных мест хранения
3. Вычисление вектора движения
4. Перемещение данных по системе



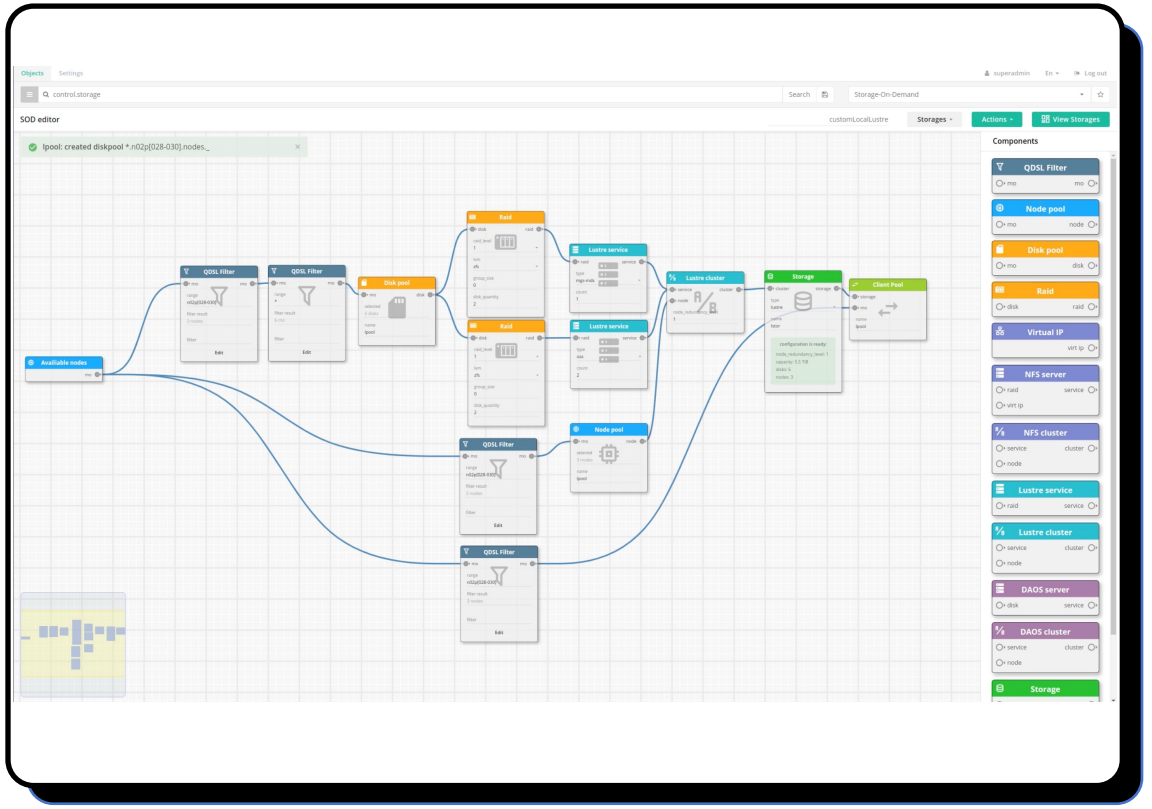
Связь с архитектурой датацентра

Система адаптируется
к существующим условиям
дата-центра

Снижение цены ошибки
при изначальном планировании
архитектуры датацентра

Данные сами формируют для себя среду

- Адаптации недостаточно
- Компонуемые среды (Composable Disaggregated Infrastructure) позволяют собирать СХД «на лету»:
  - Из доступных компонентов
  - В зависимости от требований
  - На время жизни



Требования к СХД «по–запросу» могут сильно различаться

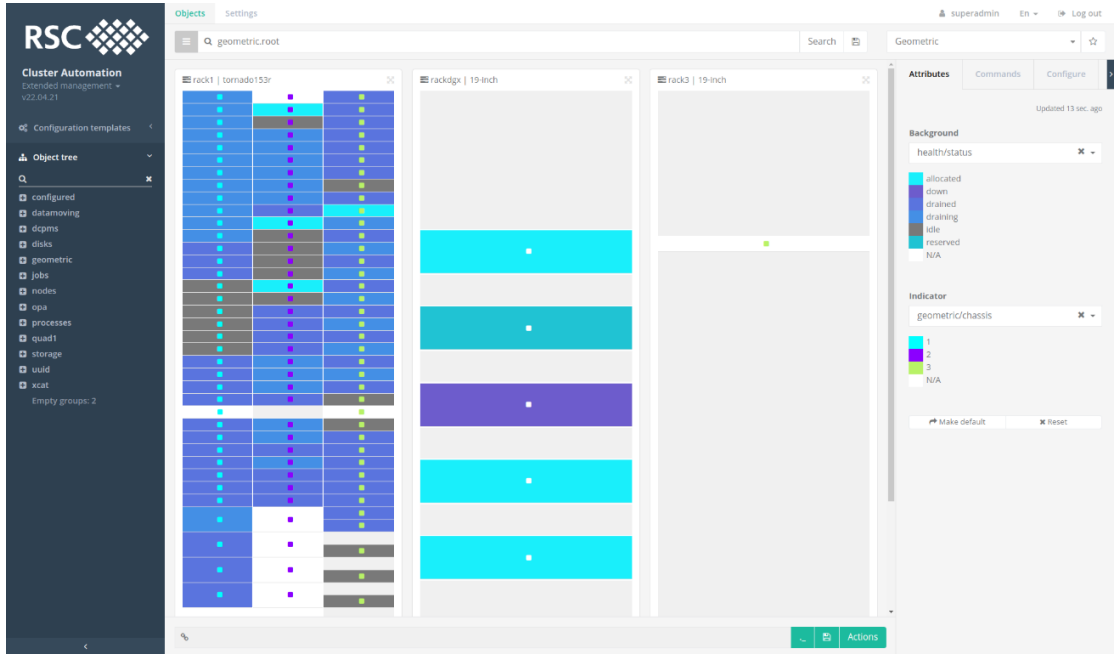
Для создания контрольной точки (Check Point) вычислительной задачи

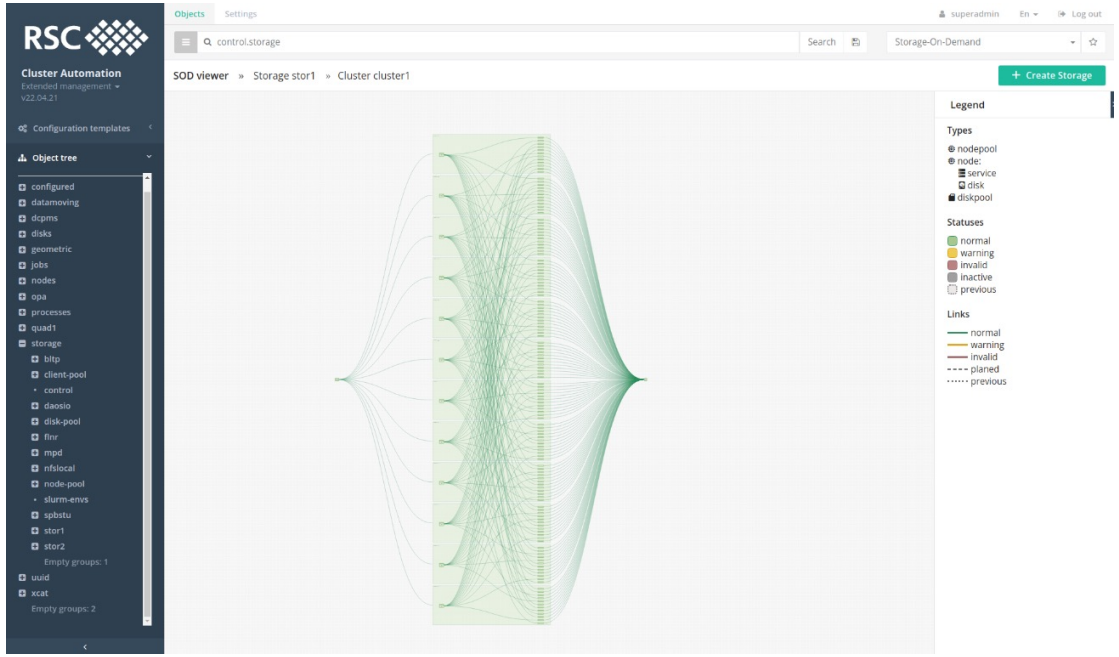
Для совместной работы с данными очень быстрое кластерное хранилище (Lustre, DAOS)

Для резервного копирования БД

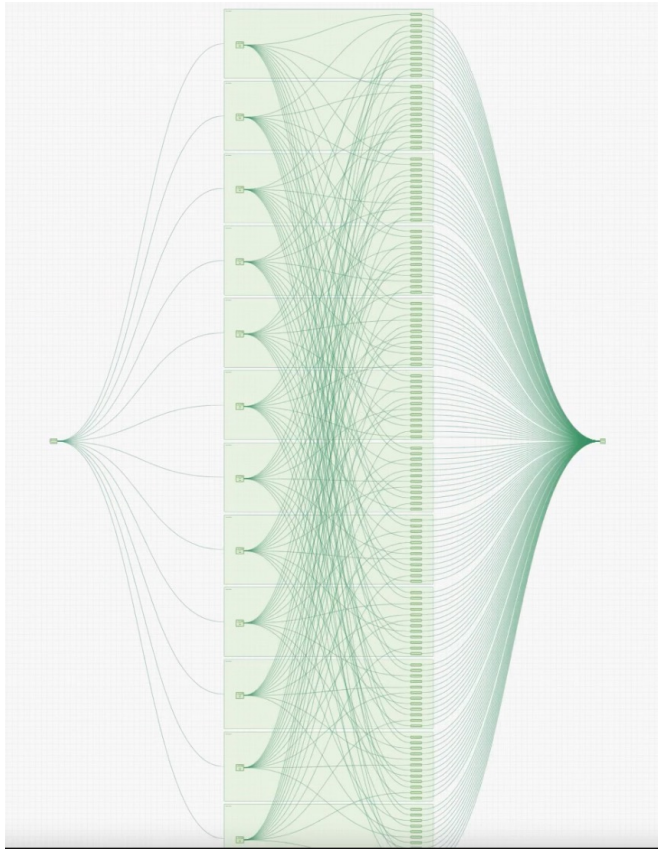
Система управления знает о ресурсах все

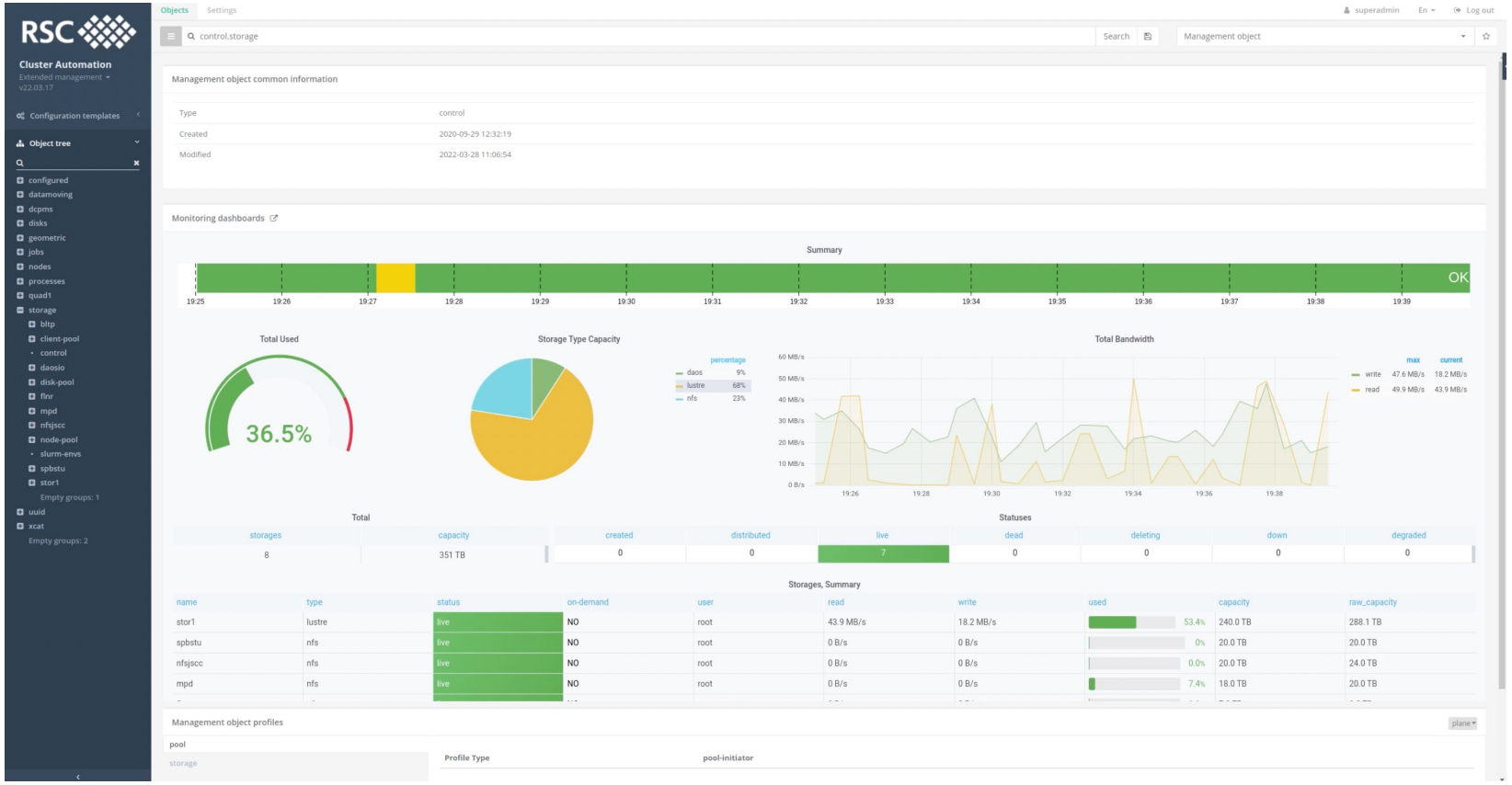
Сервера | Сети | Диски | PCIe-линии | А также топологии питания и охлаждения!





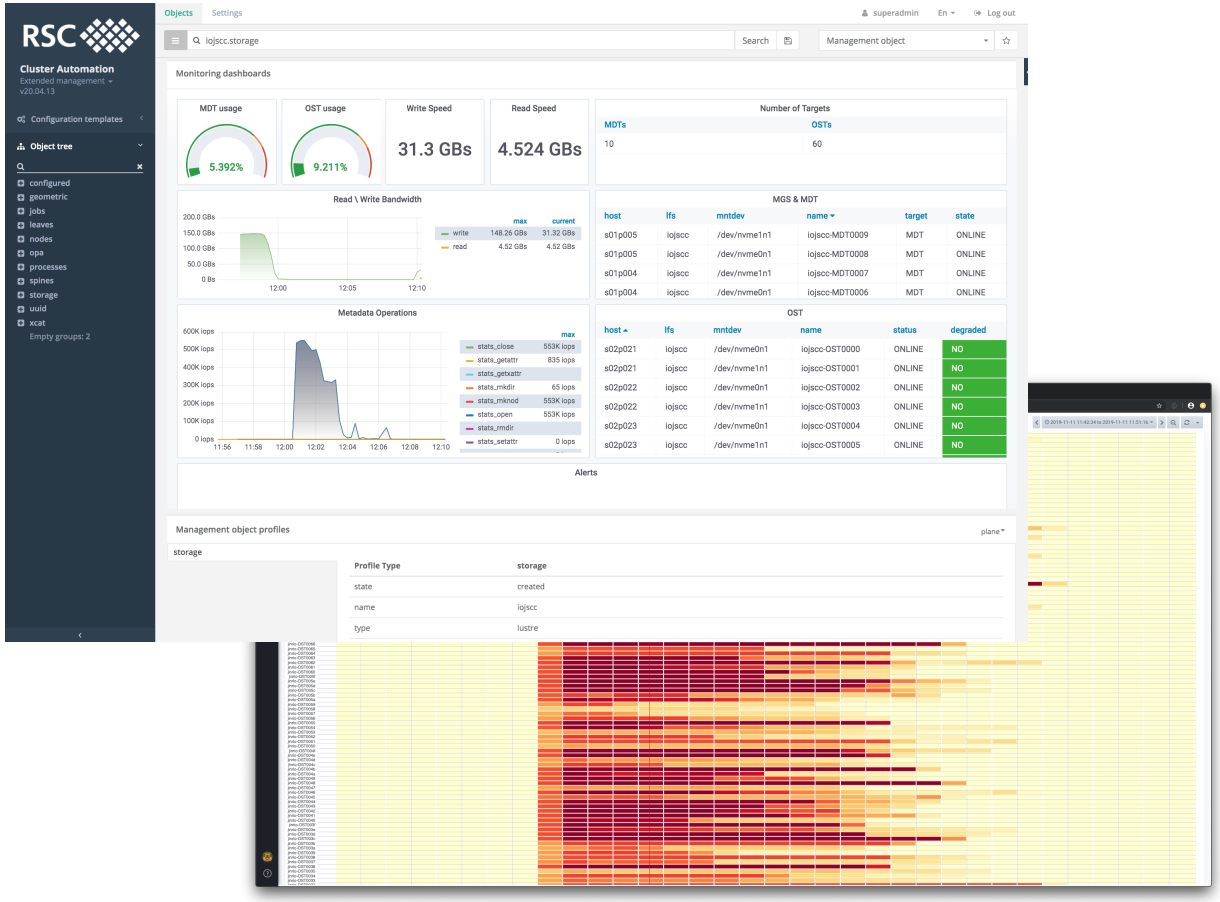
Работающая система хранения





СХД «по запросу»

- Короткоживущие
- Постоянные



Три системы в рейтинге

IO500

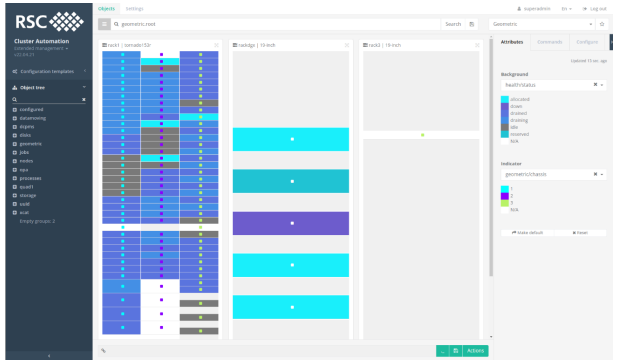
Объединенный институт ядерных исследований

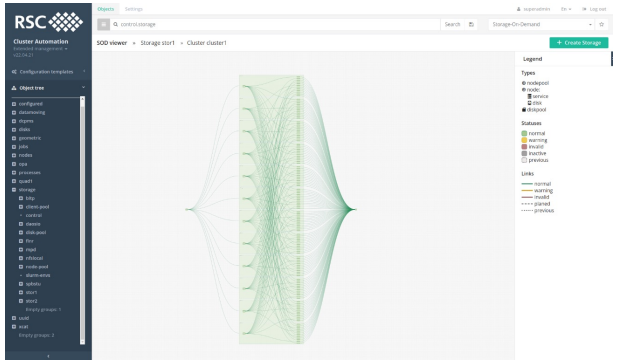
МЦ

ПОЛИТЕХ
Санкт-Петербургский политехнический университет Петра Великого

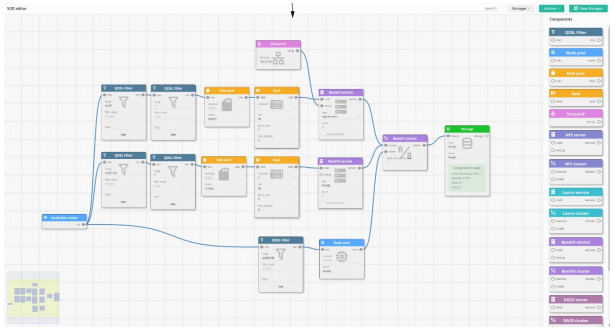
Вот как всё работает в совокупности

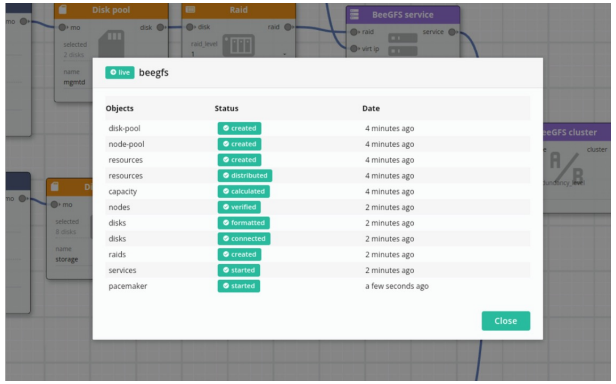
1. Компонуемая платформа



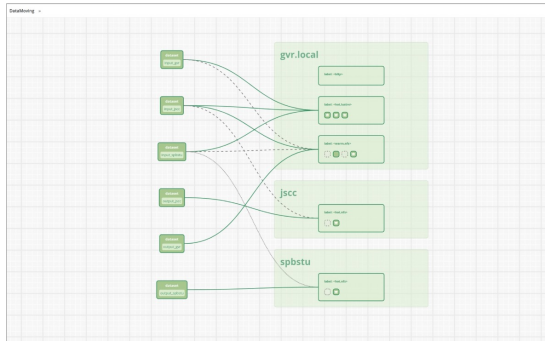


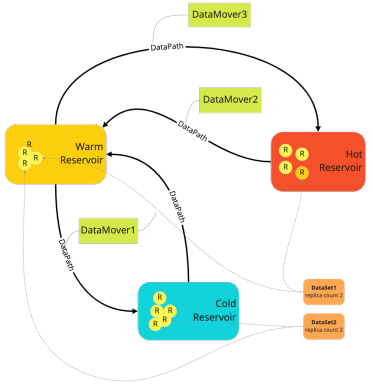
2. Система хранения «по запросу»





3. Система Data Management





Результаты внедрения

Мы сократили сроки обработки

с 25 до 10 дней

Подход Data Management в компонуемых аппаратных средах ускорил виртуальный эксперимент MPD (Multi-Purpose Detector)

Мультидоменная платформа приложений управления

FOLIAGE

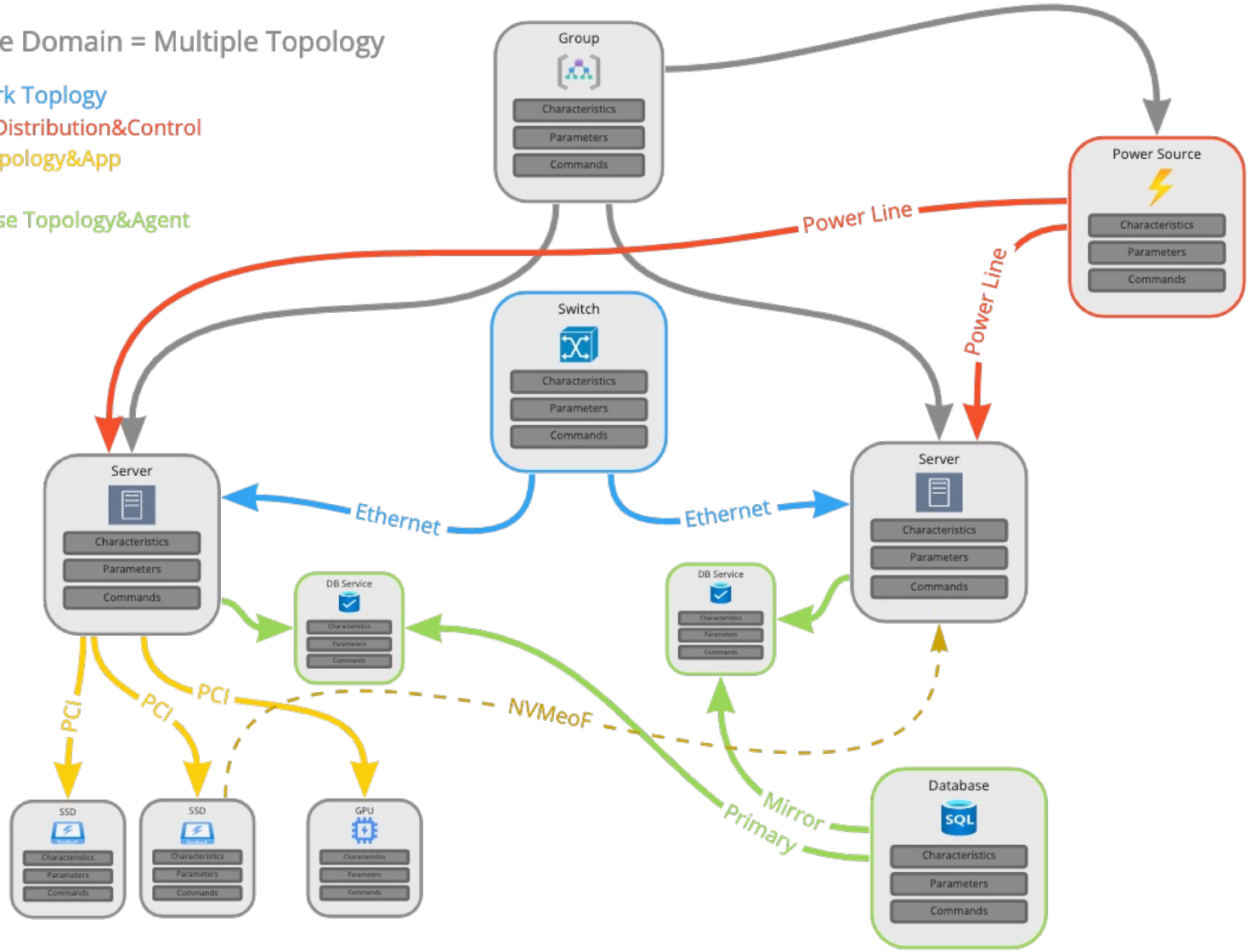
- Знает обо всех объектах и связях дата-центра
- Выполняет приложения управления
- Поддерживает жизненный цикл приложений управления
- Предоставляет SDK для разработчика

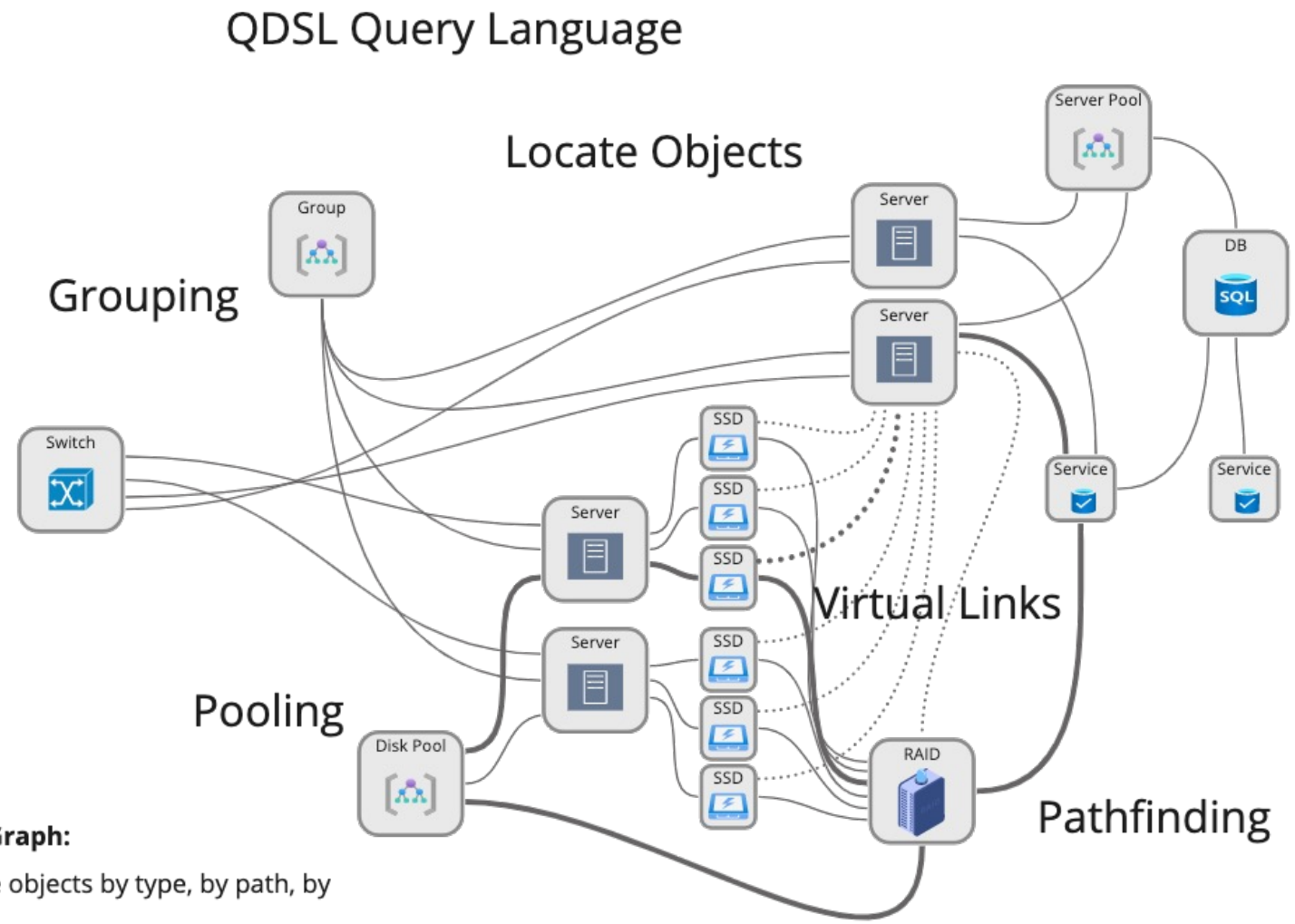
Разработчик теперь:

1. Разрабатывает только полезные функции
2. Размещает функции в графе исполнения

Multiple Domain = Multiple Topology

Network Toplogy
Power Distribution&Control
PCIe Topology&App
SDS
Database Topology&Agent





Object and Links Graph:

- Find and locate objects by type, by path, by attributes, etc.
- Explore and construct useful paths
- Create virtual links

Any object profile can be used as shared context and information exchange point, between functions, and even between applications

DNS-like: slot1.chassis2.rack3.geometric, n01p001.nodes.root, switch2.access.networks

Wildcard:

..rack1.geometric, *.chassis1.*.geometric

Ranges: [n01p001, n02p002, n03p003].nodes, sky[08-10], n01p[001-002, 010-011]

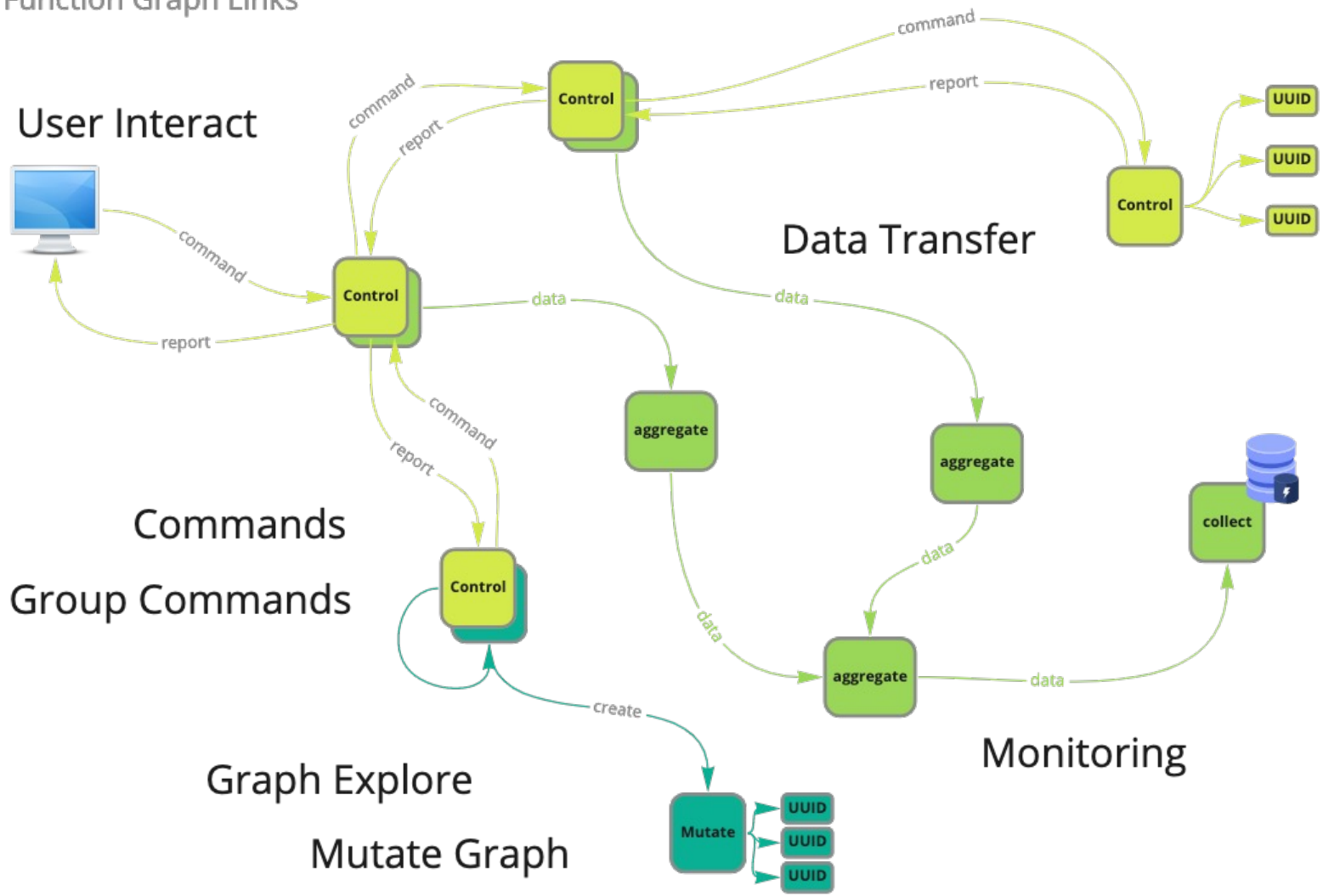
Filters: *[?@.metrics.temp > 70?]

Examples:

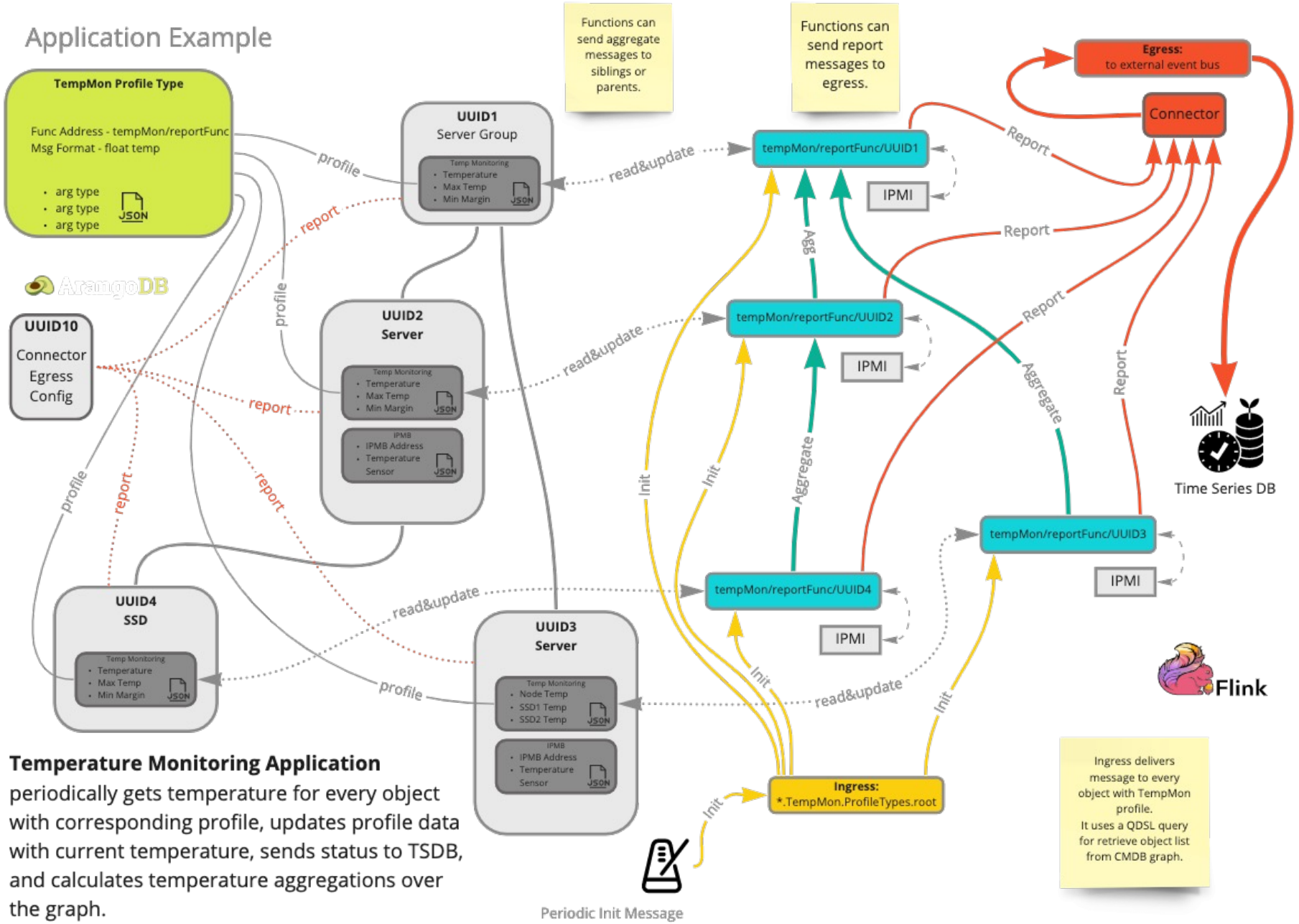
<[?@.power.avg_1m > 300?].geometric

All nodes with 1 minute power consumption average is greater then 300W

Function Graph Links

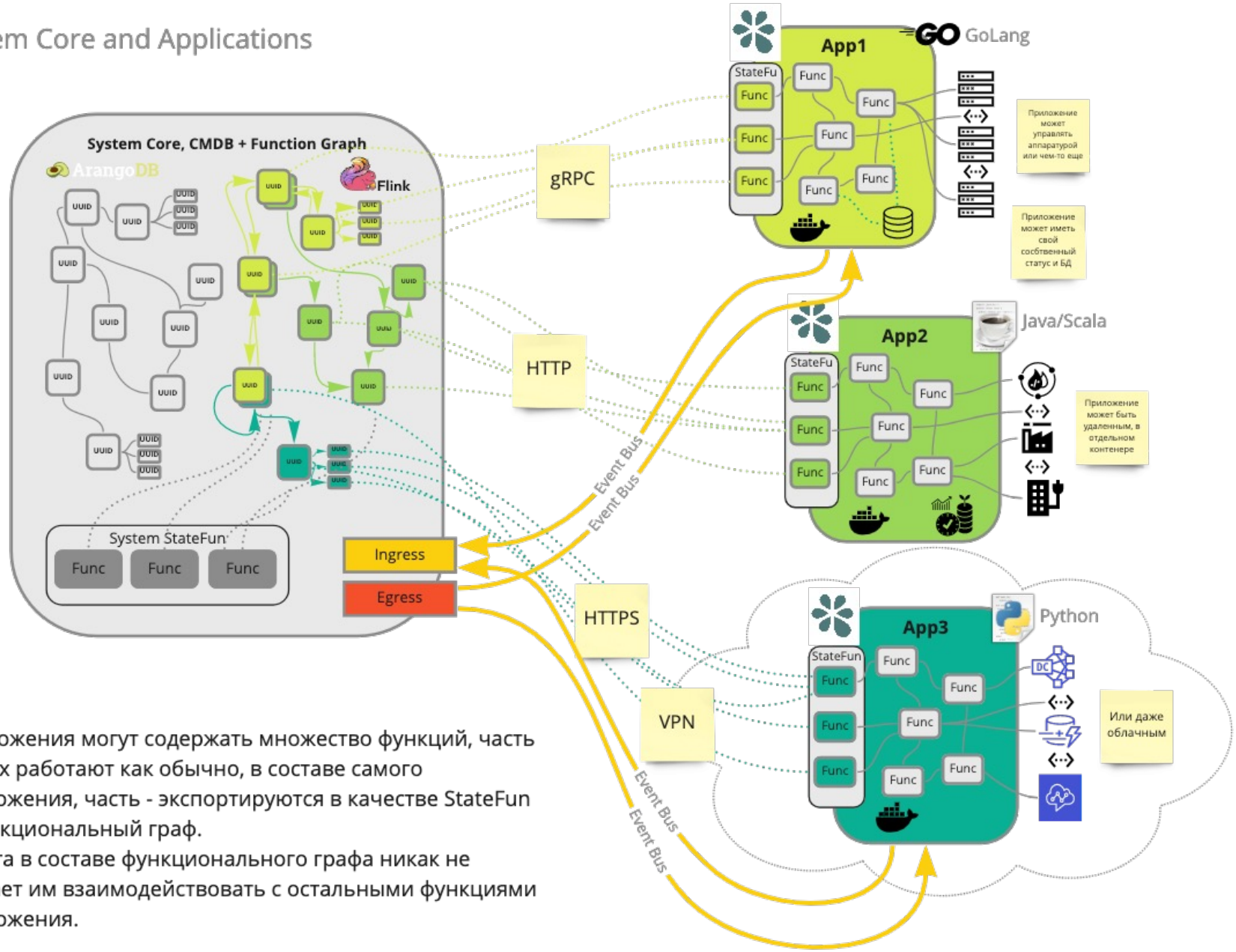


Application Example



Temperature Monitoring Application
periodically gets temperature for every object with corresponding profile, updates profile data with current temperature, sends status to TSDB, and calculates temperature aggregations over the graph.

System Core and Applications



Приложения могут содержать множество функций, часть из них работают как обычно, в составе самого приложения, часть - экспортируются в качестве StateFun в функциональный граф.
Работа в составе функционального графа никак не мешает им взаимодействовать с остальными функциями приложения.

Спасибо!

#RSC #ОИЯИ

#Суперкомпьютеры

#Компонуемость

#Системы управления

#СХД

#Кластера

#Datamanagement