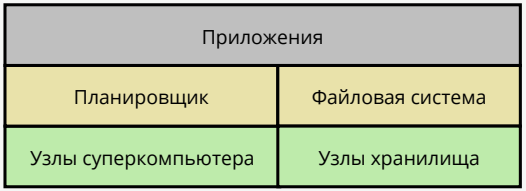
Параллельные файловые системы и планировщики пакетных задач

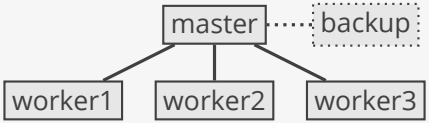
2022

План

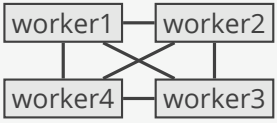


- Параллельные файловые системы.
- Хранилища ключ-значение.
- Планировщики задач.

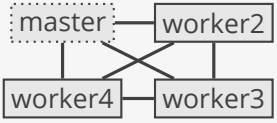
Архитектура



- CouchDB v1
- SLURM



- CouchDB v2
- Gluster



- CouchBase

Gluster

Gluster — высокопроизводительная, отказоустойчивая, параллельная файловая система.

- Нет сервера метаданных.
- Отказоустойчивость.
- Совместимость с Hadoop.

Пример создания тома

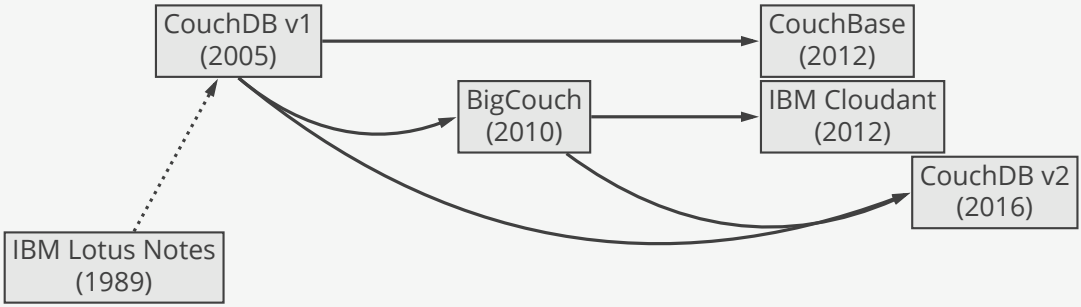
```
$ gluster volume create my-volume \
    disperse-data 4 redundancy 2 transport tcp \
    m1:/var/lib/gdata \
    m2:/var/lib/gdata \
    m3:/var/lib/gdata \
    m4:/var/lib/gdata \
    m5:/var/lib/gdata \
    m6:/var/lib/gdata
Creation of my-volume has been successful
Please start the volume to access data.

$ gluster volume start my-volume
Starting my-volume has been successful
```

Способы распределения данных

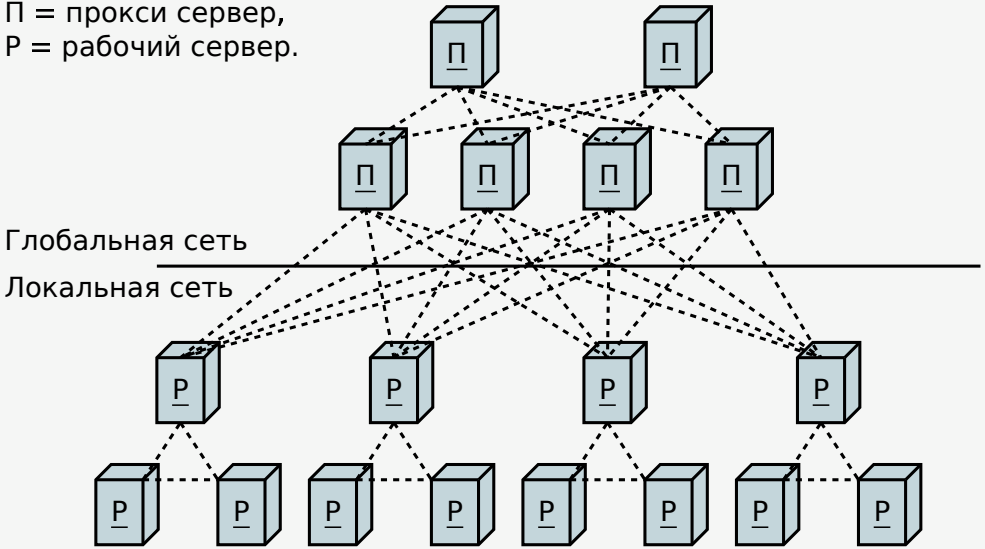
- Репликация (replicated volume).
- Фрагментация (striped volume).
- Репликация + фрагментация.
- Циклические коды (dispersed volume).

CouchDB

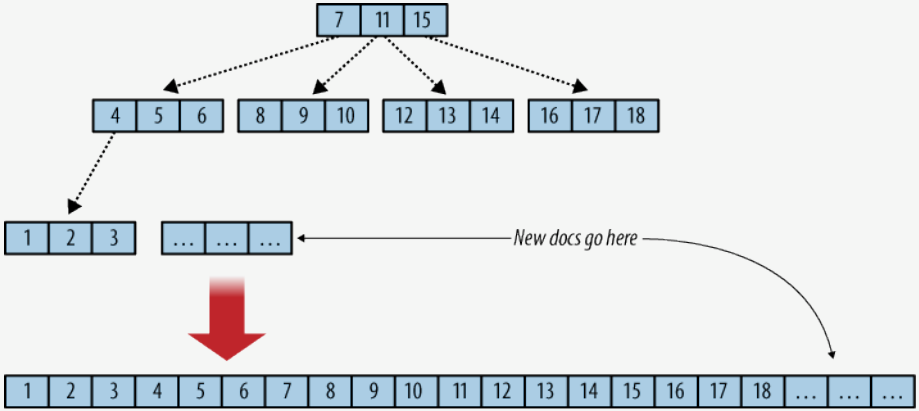
CouchDB (Cluster of Unreliable Commodity Hardware) — документ-ориентированная база данных, доступная как веб сервис.



Архитектура CouchDB



Толстые деревья

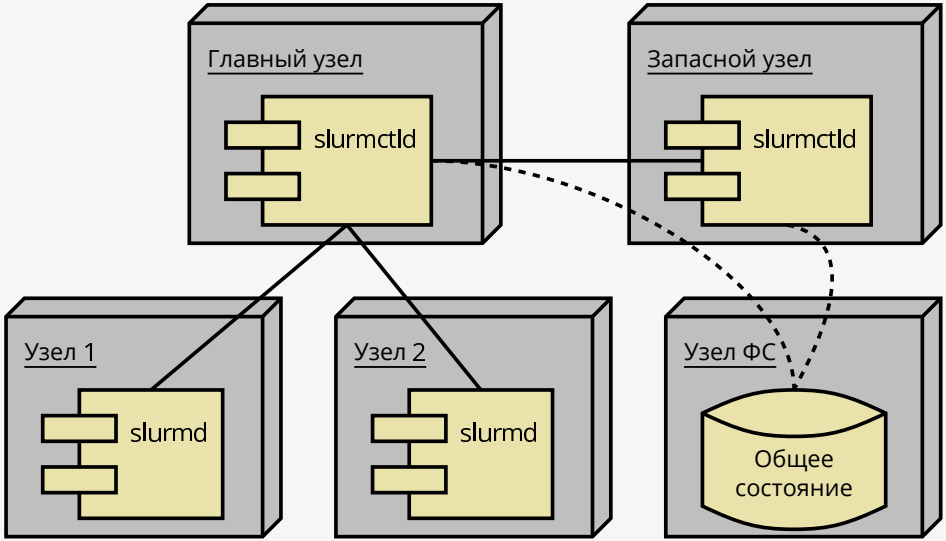


SLURM

SLURM (Simple Linux Utility for Resource Management) — планировщик пакетных задач для суперкомпьютеров.

- Ресурсы выделяются в эксклюзивное пользование приложению.
- Используется на многих суперкомпьютерах из списка TOP500.

Архитектура SLURM



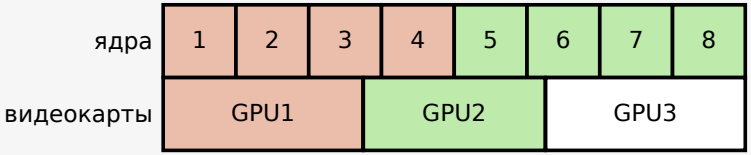
Резервирование ресурсов

Хочу 1 узел, 4 ядра и 1 видеокарту:

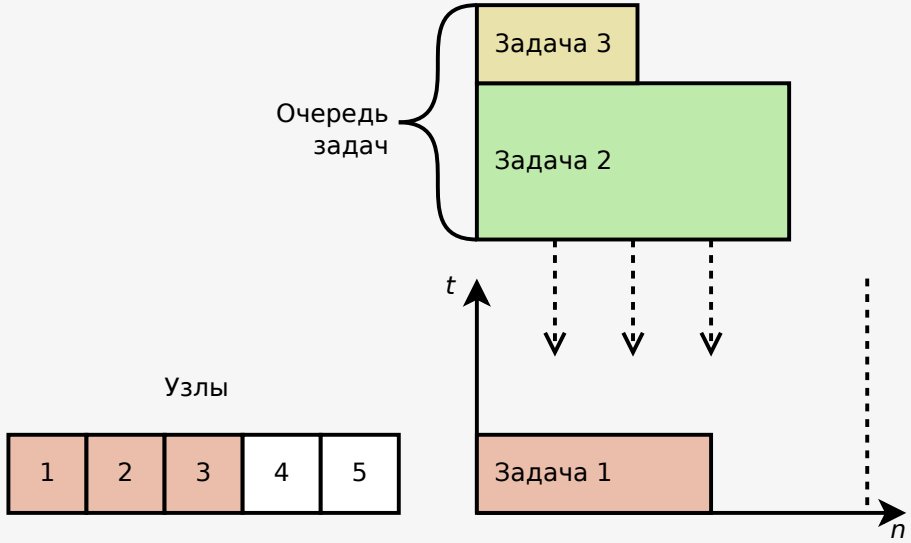
```
sbatch \
    --nodes=1 \
    --tasks=4 \
    --gres=gpu:1
```

Ресурсы:

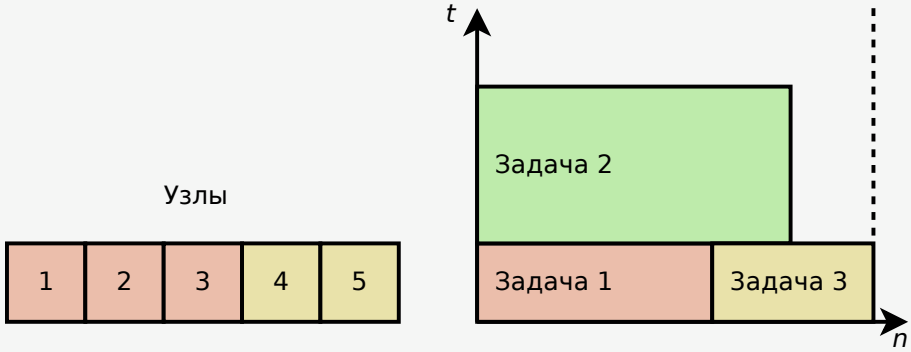
- ядра
- оперативная память
- дисковое пространство
- видеокарты
- лицензии
- образы операционной системы



Алгоритм планирования



Алгоритм планирования



Контрольные точки восстановления

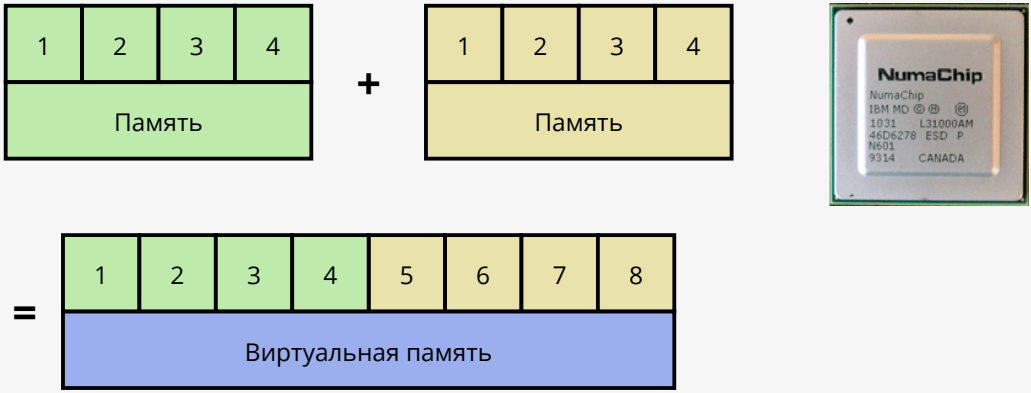
CRIU:

```
$ criu dump -t PID
$ criu restore ...
```

BLCR:

```
$ sbatch --checkpoint 30 ...
$ scontrol checkpoint create JOBID
$ scontrol checkpoint restart JOBID
```

Виртуальная общая память



Узлы + хранилище + планировщик = суперкомпьютер