ПРИМЕНЕНИЕ СОВРЕМЕННЫХ ПАРАЛЛЕЛЬНЫХ ФАЙЛОВЫХ СИСТЕМ В СУПЕРКОМПЬЮТЕРНОМ ЦЕНТРЕ

Межведомственный суперкомпьютерный центр РАН
– филиал ФГУ ФНЦ НИИСИ РАН

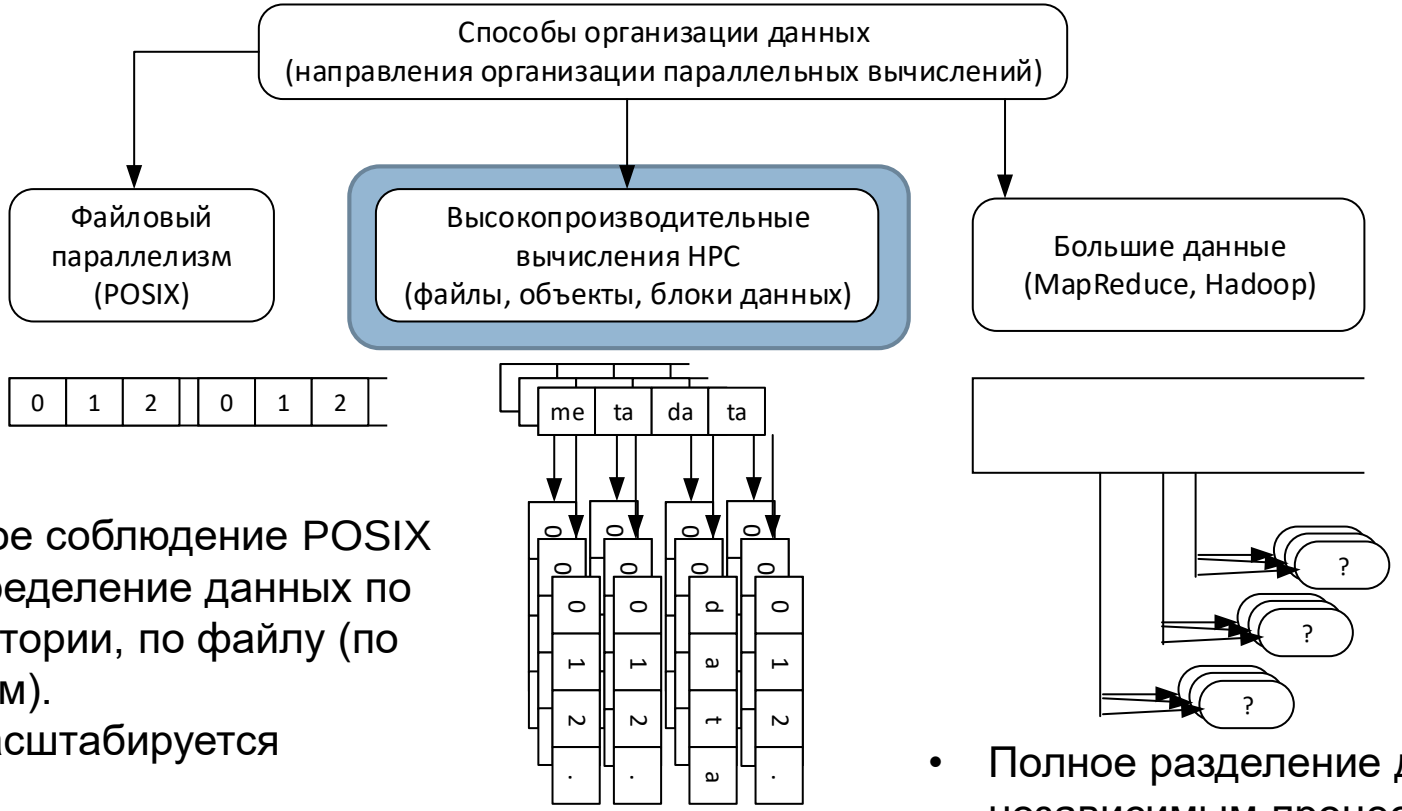
Аладышев Олег Сергеевич (к.т.н.)

Москва 2023

Суперкомпьютерные дни в России

Содержание

Направления параллельных вычислений

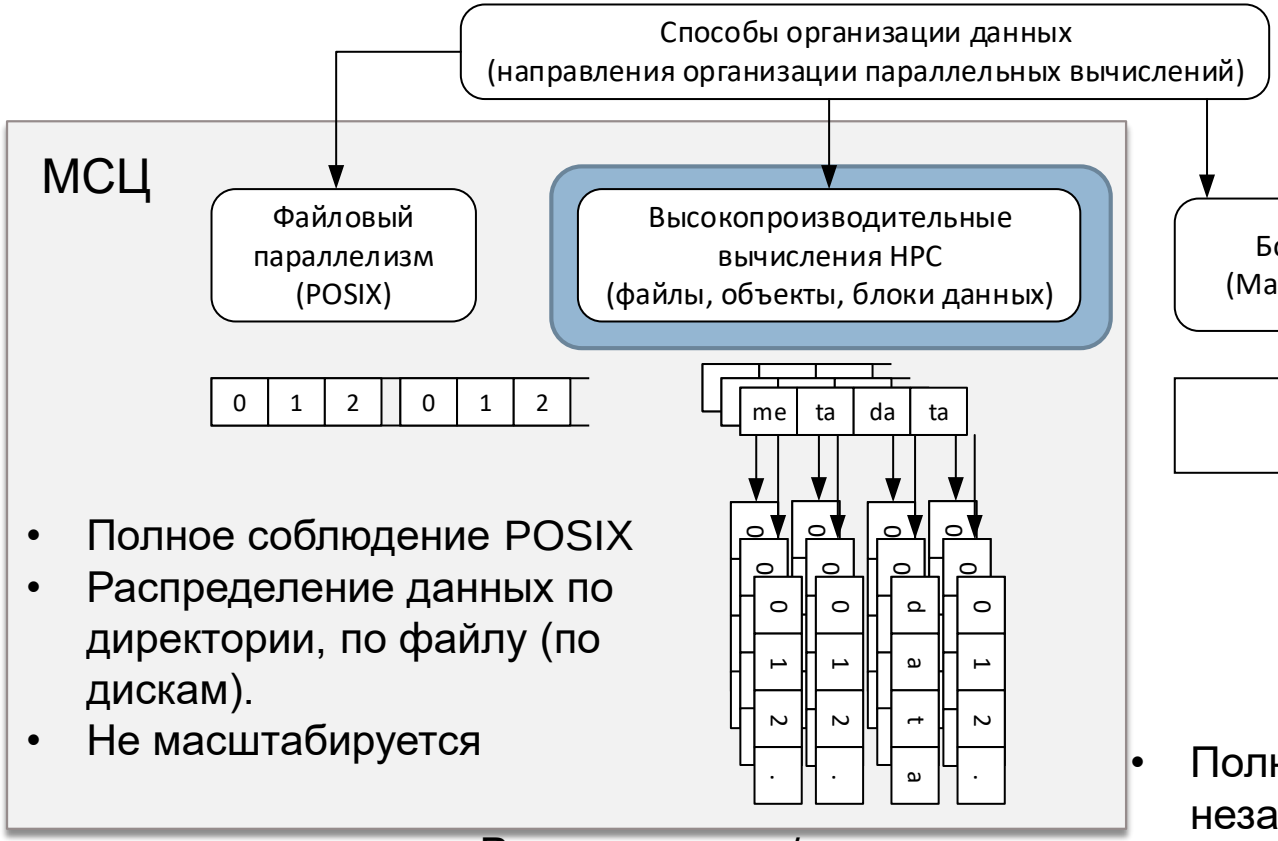


- Полное соблюдение POSIX
- Распределение данных по директории, по файлу (по дискам).
- Не масштабируется

- Распределение/разделение данных объектам (файлам, блокам)
- Частичное несоблюдение POSIX
- Массовый параллельный ввод-вывод, данные к процессу

- Полное разделение данных по независимым процессам
- Идеальное масштабирование
- Перемещение процесса к данным.
- POSIX не нужен

Направления параллельных вычислений



- Полное соблюдение POSIX
- Распределение данных по директории, по файлу (по дискам).
- Не масштабируется

- Распределение/разделение данных по объектам (файлам, блокам)
- Частичное несоблюдение POSIX
- Массовый параллельный ввод-вывод, данные к процессу

- Полное разделение данных по независимым процессам
- Идеальное масштабирование
- Процесс к данным.
- POSIX не нужен

Где нужен ввод-вывод (IO)

1. IO нужен самим параллельным приложениям HPC
   a) Декомпозиция данных. **Ввод**, распределение по решающему полю
   b) ~~Счёт и обмен данными между параллельными процессами~~.
   c) Генерация, обработка данных
   d) Вывод временных и/или конечных результатов.

2. Обеспечение работы СК
   a) Профайлы пользователей, системное и пользовательское ПО
   b) Обеспечение надёжности СК, гарантированности получения вычислительного результата - **контрольная точка** (КТ)

Проблема «контрольной точки»

КТ – «контрольная точка»

1. Системная КТ
2. Пользовательская КТ

Общие свойства КТ:

- запись на внешнюю систему хранения, операция последовательной записи;
- в массовом порядке;
- записываются только те данные, которые могут быть востребованы для последующего счета;
- размер блока операции выбирается достаточно большой.

*КТ – основная форма использования систем хранения HPC

Файловые системы выч. кластера

Цели:
Программная организация доступа к внешним данным

Задачи:
Разделение доступа к данным между вычислительными узлами

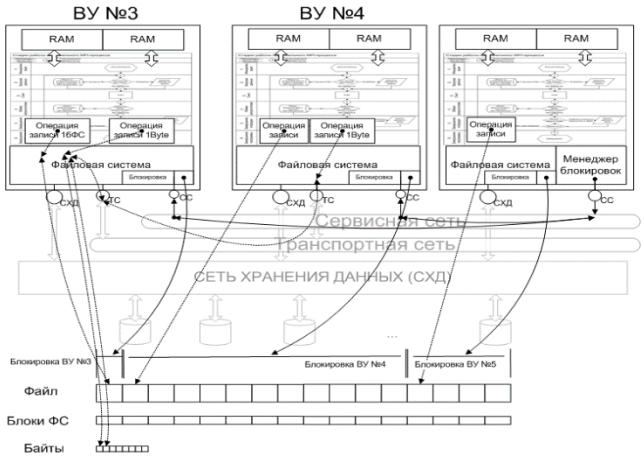
Общие требования к ФС:
•POSIX (полное удовлетворение только на одном узле)
•Масштабируемость (наращивание объёмов и производительности)
•Обеспечение независимости процессов разных ВУ по данным (механизм блокировок). Параллельный доступ к данным и метаданным.
•Репликация, целостность данных, восстанавливаемость после сбоев

Структура:
Данные (последовательность байт)
Метаданные (информация о файле, объекте или блоке, права доступа, карта (layout) расположения данных, директории, …)

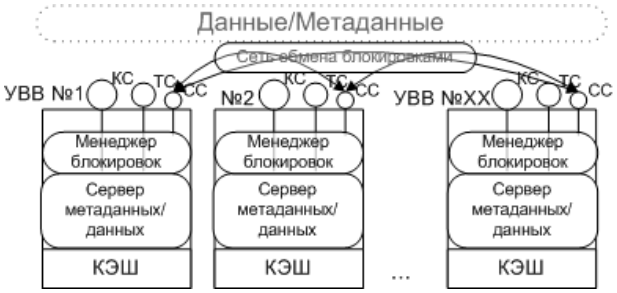
Параллельная файловая система

Разделение данных по дискам, упреждающее чтение, кэширование - отложенная запись.

Поддержка больших директорий
Регистрация изменений и восстановление метаданных
Механизм блокировок данных и метаданных

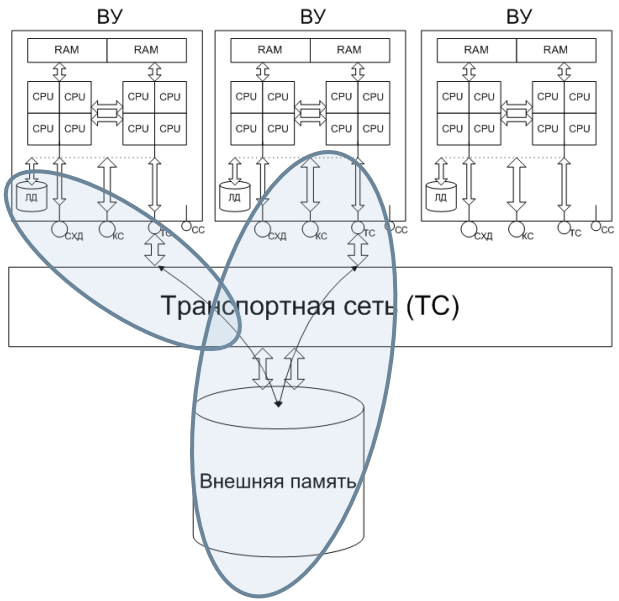


Маленькие файлы, хвосты больших файлов
Управление блокировками



Таблицы размещения, разделение прав на добавление
Глобальные метаданные ФС
Отказоустойчивость, определение момента сбоя (fail IO узла, сбой канала или карты, сбой дисков)

Внешняя память кластера HPC



Если кроме вычислительного узла (ВУ) локальную память никто не использует, то называем её **внутренней или локальной**.

ЛД – локальный диск узла, может быть использован как для локальной системы хранения, так и для внешней.

Признак **внешней** системы хранения данных (ВСХД).
Носители данных логически вынесены за пределы ВУ, за решающее поле:

- диски через сеть
- локальный диск через сеть
- ввод-вывод (IO) на внешнюю память или на локальный диск другого узла

Организации внешней системы хранения данных (ВСХД)

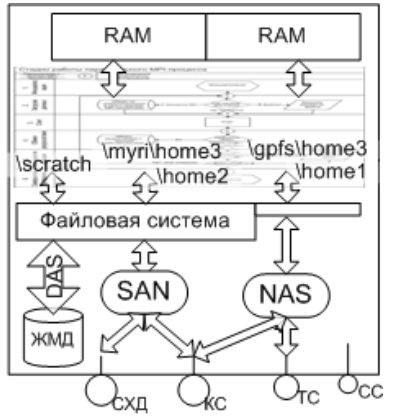
Монтирование системы хранения на вычислительном узле кластера (ВУ):

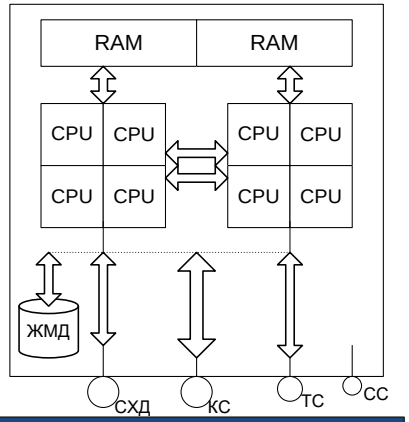
- монтирование с локального диска
- монтирование через сеть хранения данных
- монтирование через сеть tcp/ip

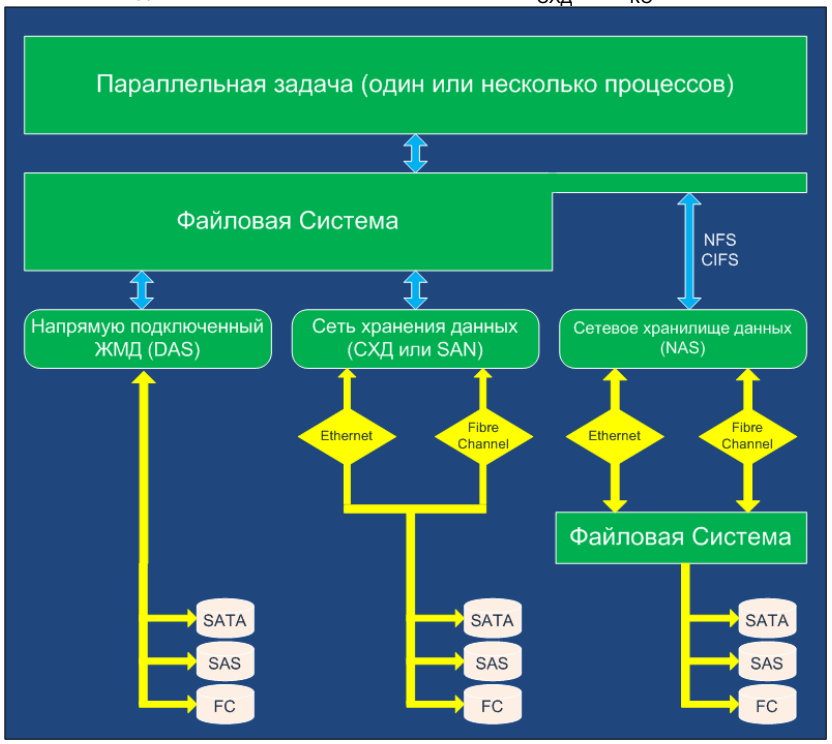
2 момента:

- на ВУ есть серверная часть ФС требуется выделение оперативной памяти и процессорной мощности

- на ВУ нет серверной части ФС не требуется выделение оперативной памяти для ФС (только для процессов на ВУ – клиент ФС)







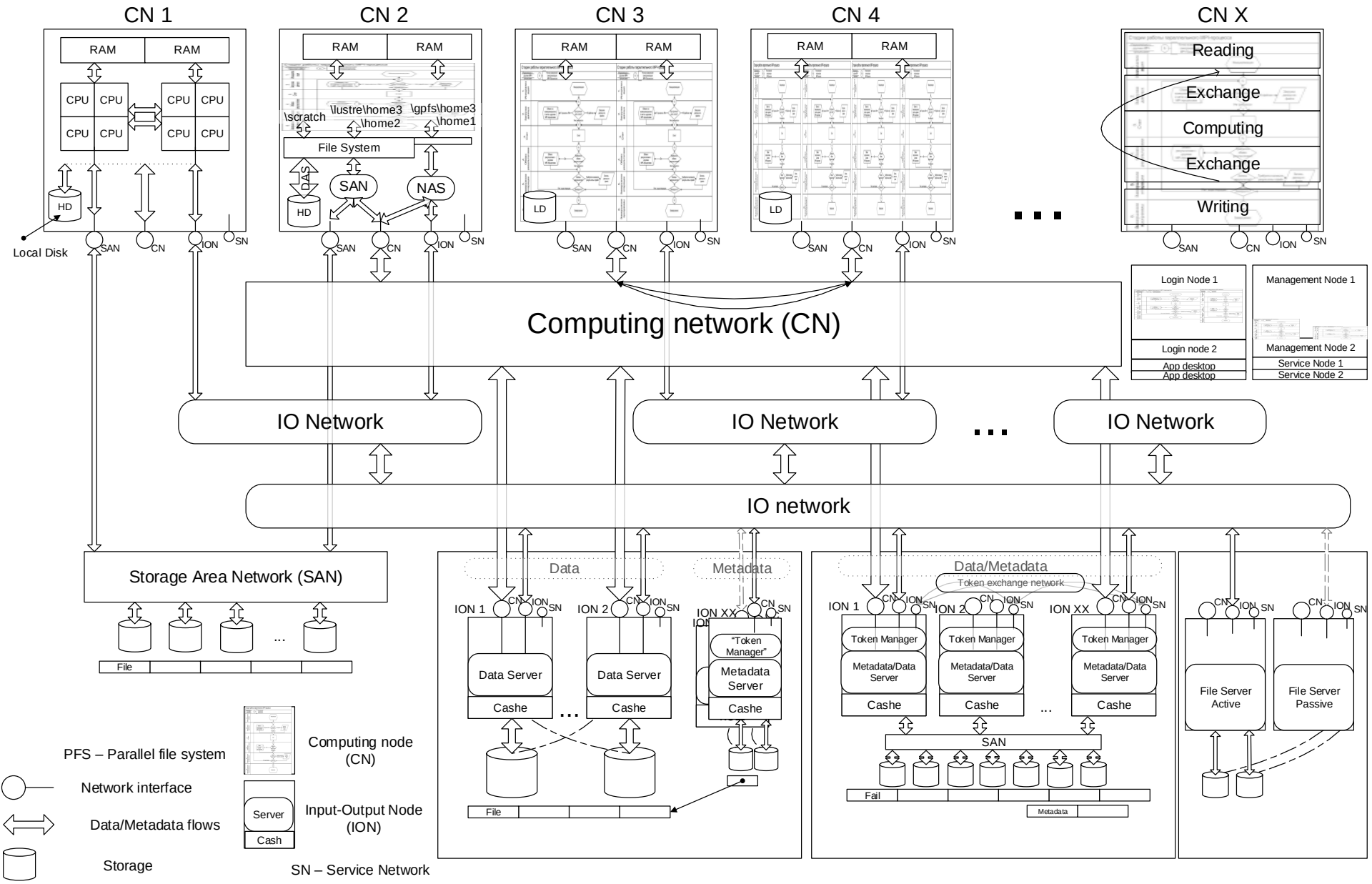
ВСХД суперкомпьютера

Структура вычислительного кластера:

Решающее поле – набор вычислительных узлов (ВУ)
Сети различных типов и назначений:
- управляющая сеть кластера
- коммуникационная сеть для низколатентной связи ВУ
- сеть ввода-вывода (файлы, объекты, БД, …)
- сеть хранения данных (сеть дисков)
- наложенные сети (например NVMe over OMP)

Узлы ввода-вывода (серверы)
- NFS серверы
- кластеры ввода-вывода (серверы данных и/или метаданных)
- дисковые массивы

High performance computing cluster



Системы хранения МСЦ РАН

Высокодоступные (HA) NFS

- Универсальный /home1 (запуск задач запрещён)

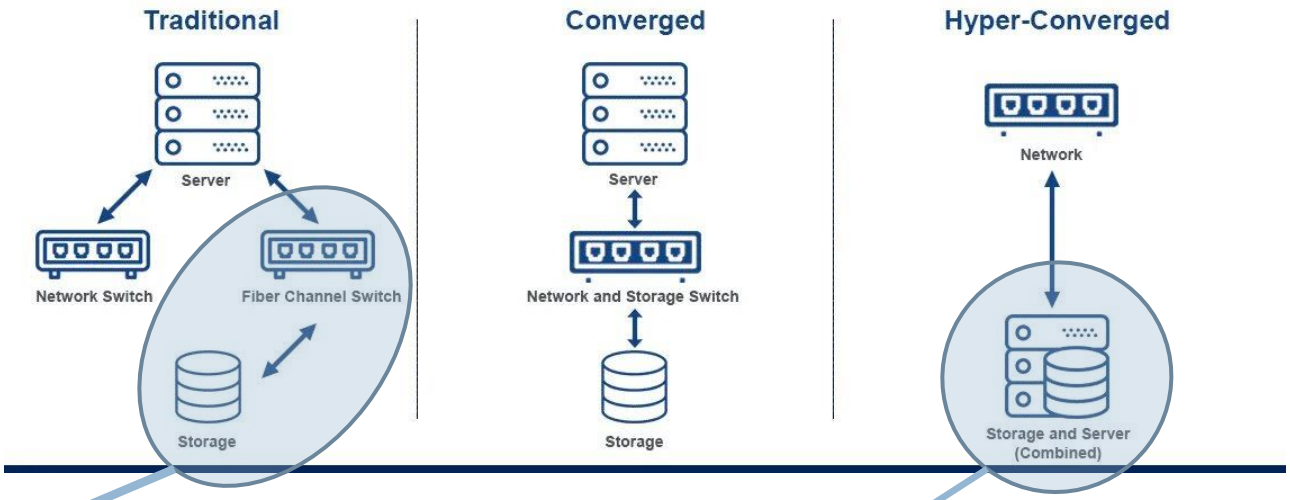
Надёжные NFS для проектов и ПО

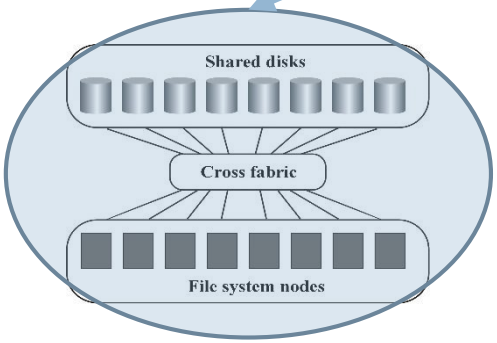
- /opt/software /common (системное ПО кластера)
- /opt/cluster_software (прикладное ПО кластера)
- /home2-6 (проекты пользователей)
- -

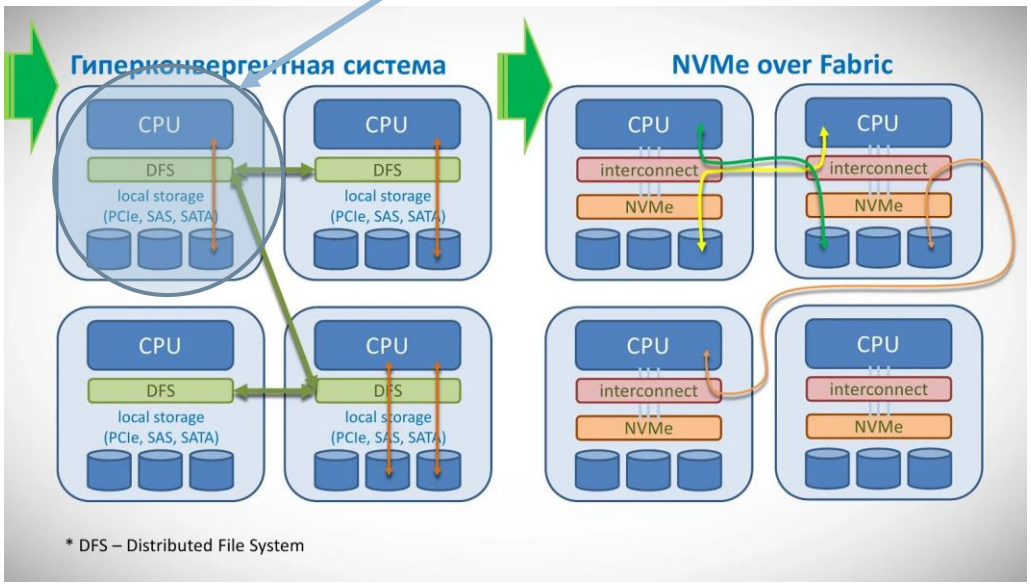
Высокоскоростные системы хранения для сопровождения расчётов и КТ

- Внешняя /lustre/lstor (конвергентная с NVMe over fabric)
- ЛД /scratch (не используется более)
- ЛД /tmpfs

Разновидности систем хранения







Традиционные - NFS
Конвергентные – узлы IO
Гиперконвергентные – ВУ+IO

DAOS РСК

Серверный процесс DAOS server instance может быть запущен как на физическом Linux-сервере, так и внутри виртуального Linux-сервера или контейнера. Его подпроцессы DAOS Engine осуществляют доступ к локальным PMEM-устройствам и NVMe-дискам, расположенным на данном узле хранения данных.

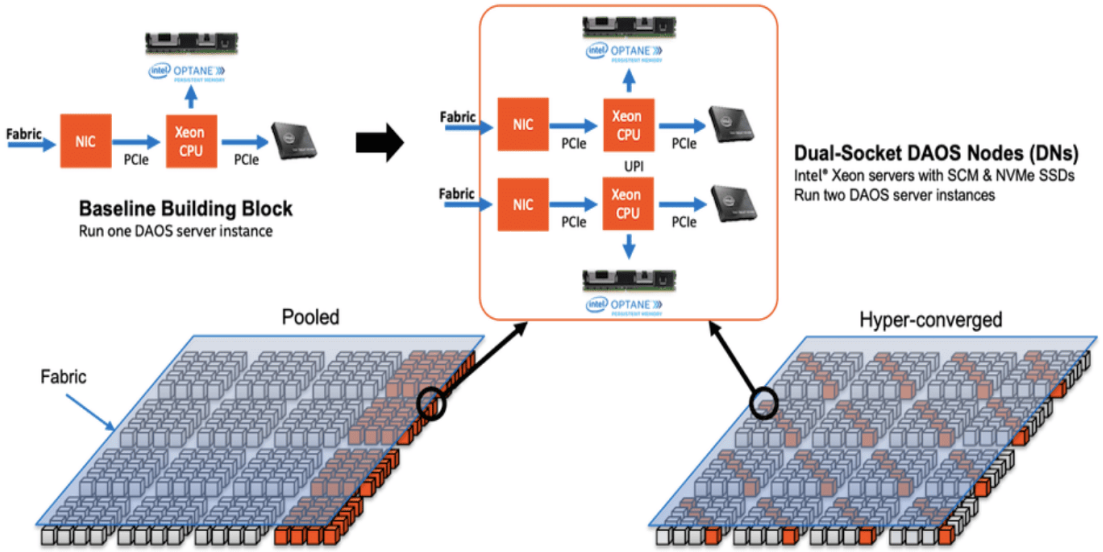


Figure 2-1. DAOS Storage

Современный дата-центр может состоять из тысяч вычислительных узлов, на которых могут быть запущены десятки тысяч **виртуальных вычислительных машин**, *compute instances*, обмен данными между которыми осуществляется посредством коммуникационной сети с высокой пропускной способностью. В гиперконвергентной вычислительной системе все или часть вычислительных узлов может также выступать в качестве узлов хранения данных, *storage nodes*.

*https://sod.rsc-tech.ru/operator-guide/daos/basic-info/

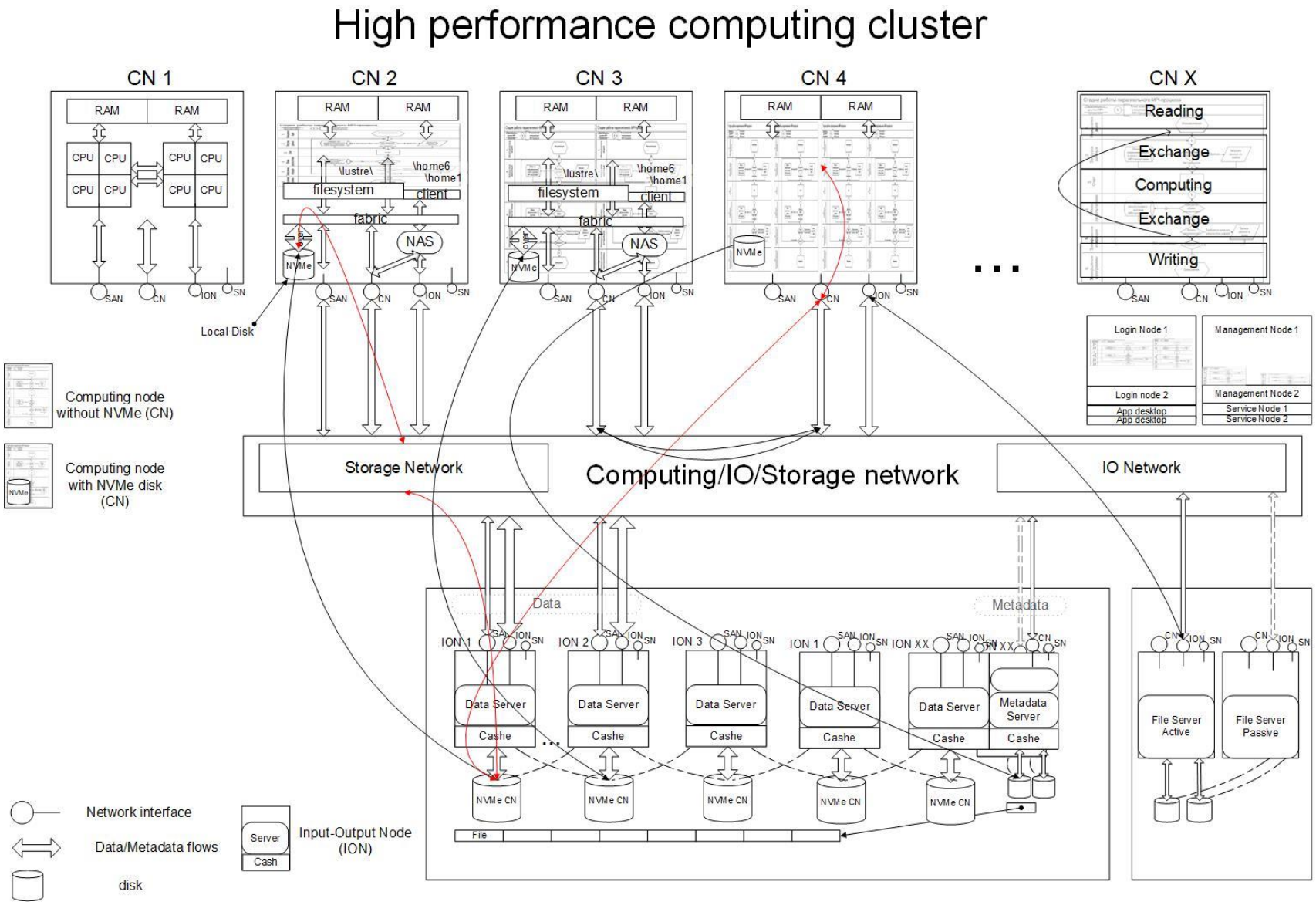
Порядок работы высокоскоростной системы хранения

Схематично порядок работы сетевых систем хранения, кластеров систем хранения и конвергентных систем можно описать следующим образом:

1. ВУ проводит операции ввода-вывода с данными ВСХД
2. Запросы проходят по IO-сети до серверов IO
3а. Если данные в локальной памяти сервера IO, ответы поступают ВУ через IO-сеть.
3б. Если данные не в локальной памяти IO-сервера, сервер берёт данные с внешних носителей по сети хранения данных и отвечают по IO-сети.

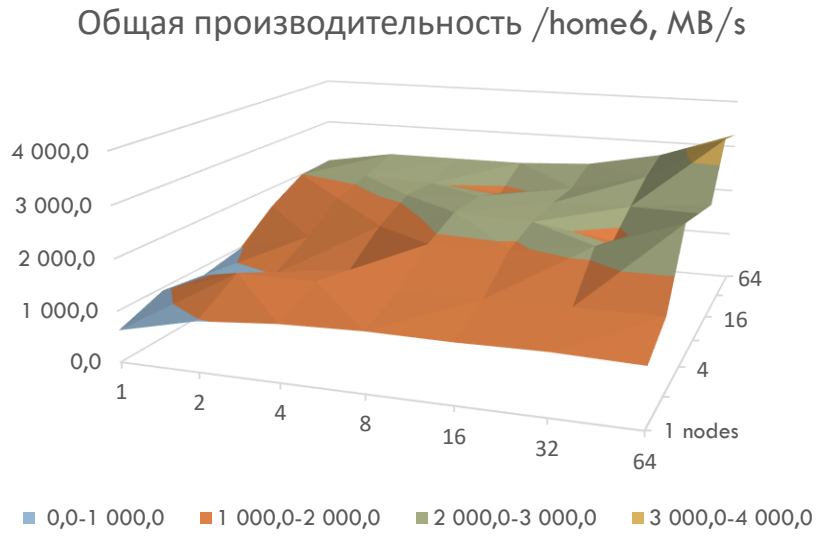
См. схему на следующем слайде, где в качестве всех сетей: вычислительной; ввода-вывода и хранения данных, выступает одна среда OmniPath

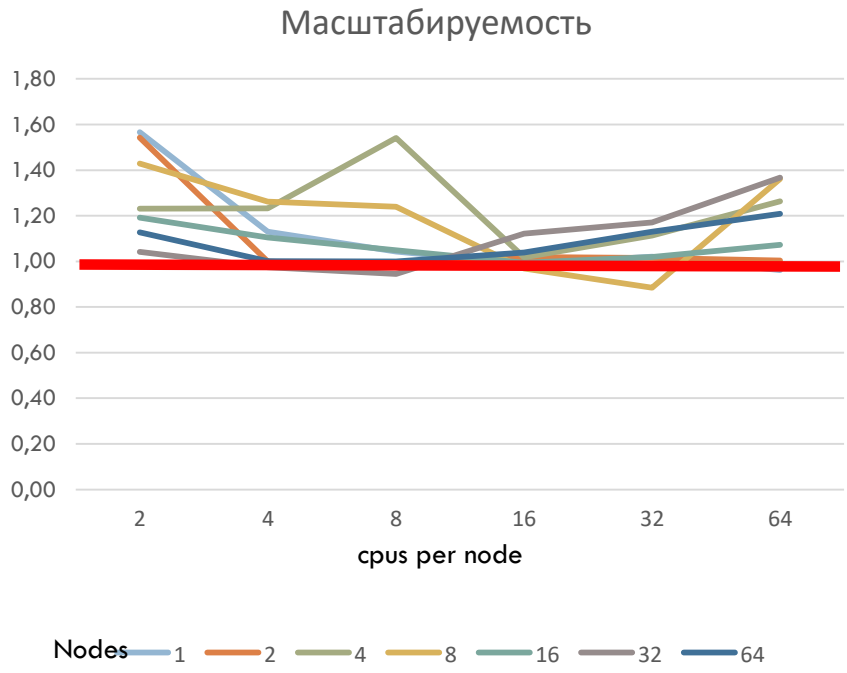
Высокоскоростная система хранения



Особенности систем хранения (NFS)

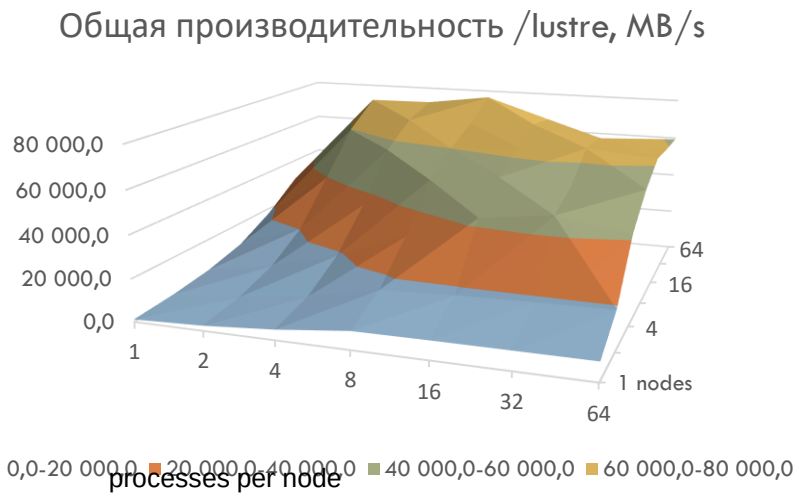
/home6
2 сервера (активный + пассивный) в режиме отказоустойчивости (pcs), zfs, 4 пула raidz2 по 14+2 HDD (SAS)
OmniPath
~ 650 ТБ для общей очереди

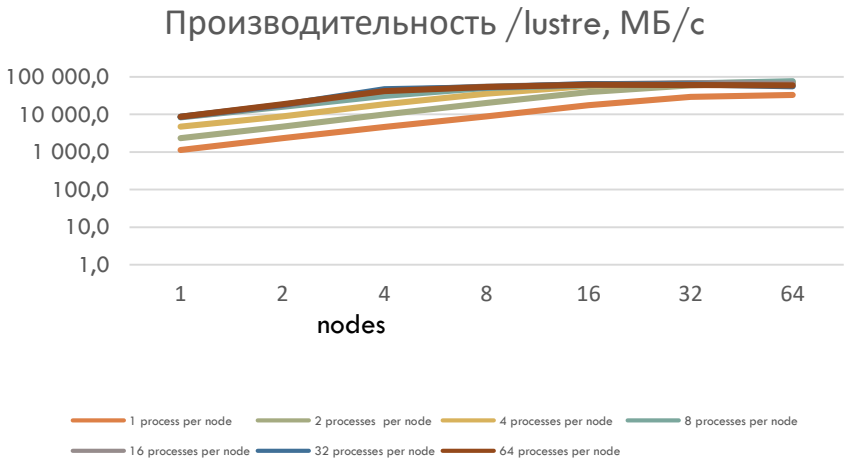


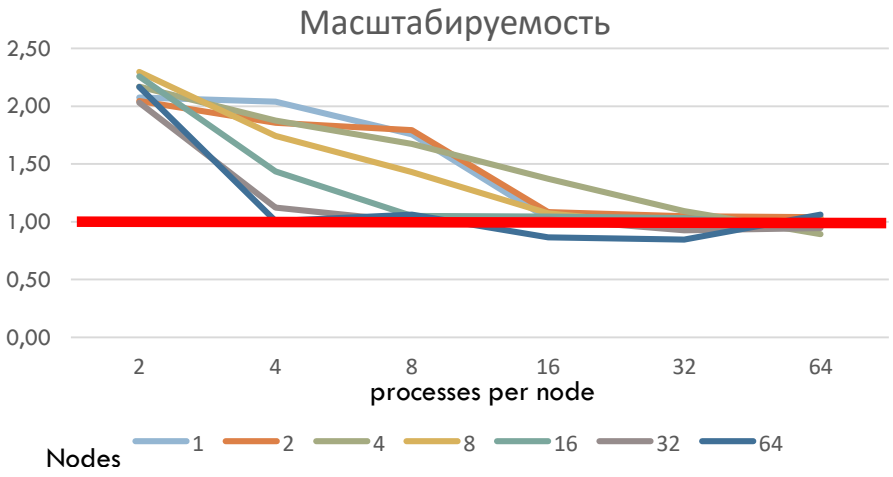


Особенности систем хранения (NVMe over OMP)

Nodes/ cpus_per	1	2	4	8	16	32	64
1	1 132,0	2 348,0	4 789,0	8 414,8	8 626,0	8 681,0	8 718,0
2	2 330,0	4 764,0	8 852,0	15 868,0	17 199,0	18 030,0	18 740,0
4	4 615,3	10 017,0	18 781,0	31 449,0	43 155,0	47 158,0	42 119,0
8	8 883,7	20 408,0	35 597,0	50 961,0	54 749,0	52 591,0	54 128,0
16	17 579,2	39 719,0	57 043,0	69 514,0	62 737,0	64 635,0	61 864,0
32	29 444,6	59 798,0	67 200,0	67 264,0	68 750,0	63 679,0	60 262,0
64	33 045,3	71 588,0	72 007,0	76 507,0	66 113,0	55 877,0	59 318,0







Соотношение в производительности

Универсальные (высокая доступность, сохранность)

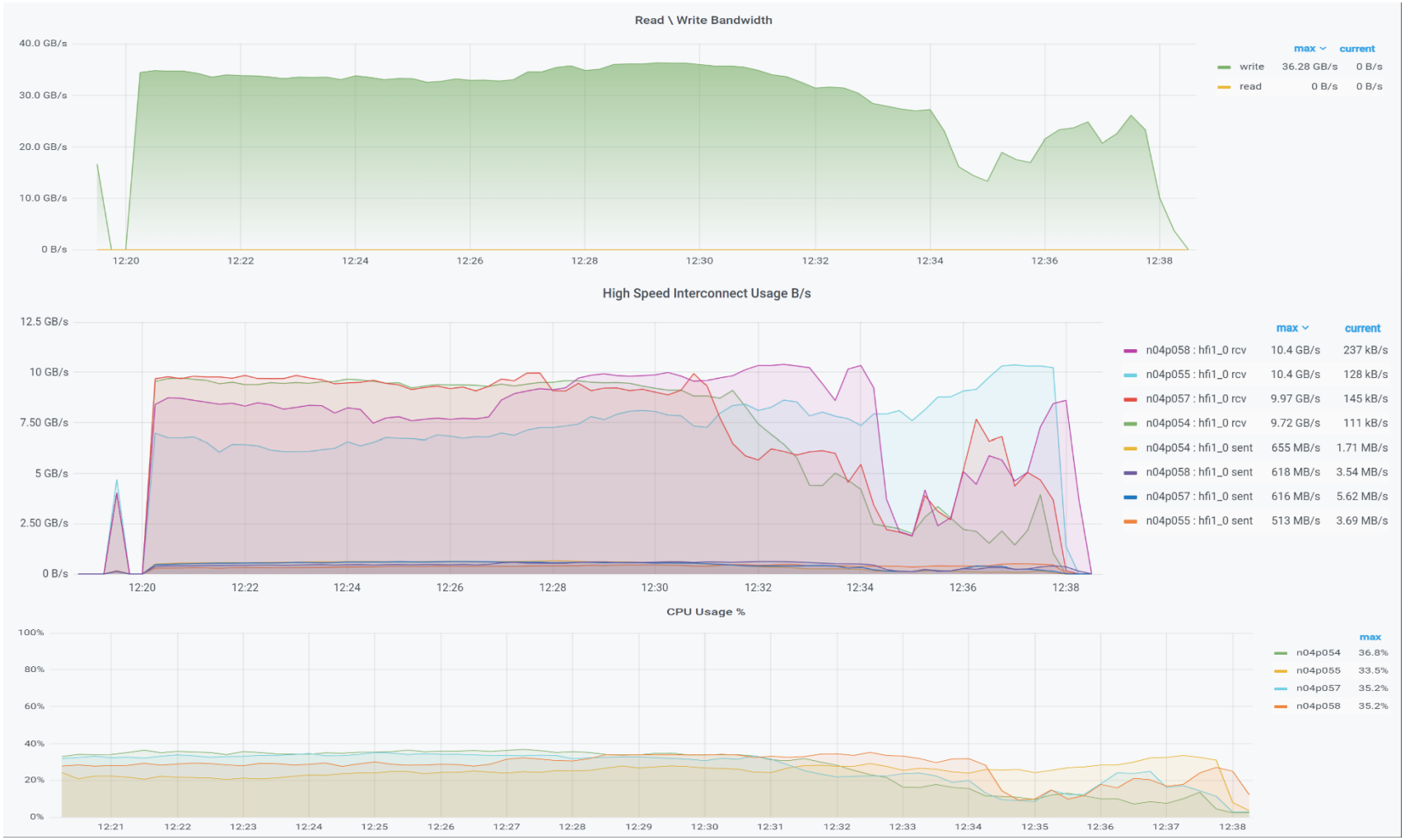
Проектные (доступность, сохранность)

Высокоскоростные (производительность)

Загрузка

OMP 100Gb/s ~12,5 GB/s
5 OSS = 5*12,5 = max 62,5 GB/s (на графике 4)

64 nodes x 64 MPI per node (48 cpus per node)

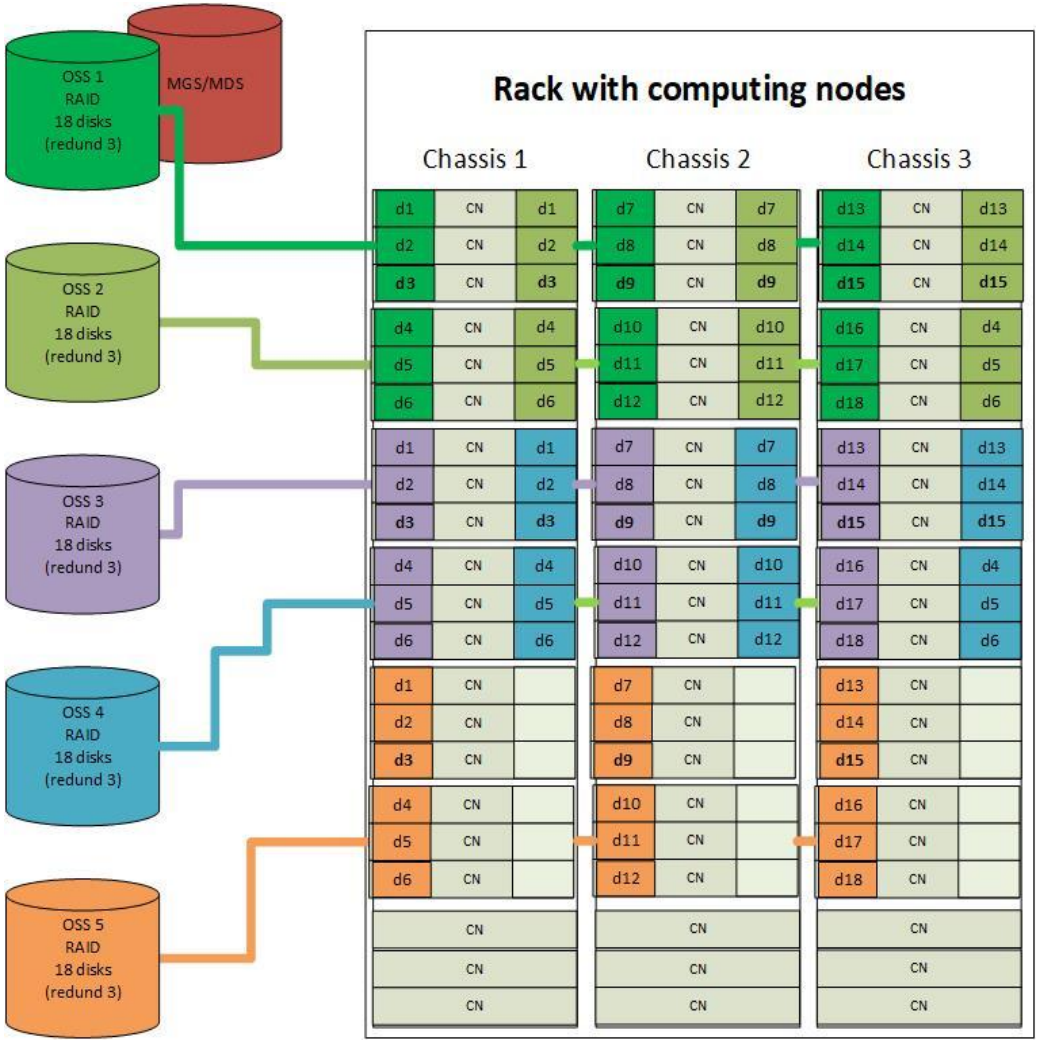


Надежность, избыточность

Полная отказоустойчивость в случае размещения дисков на вычислительных узлах в системах коллективного пользования не имеет смысла.

Путём распределения устройств по независимым зонам добиваемся максимальной надёжности в рамках имеющихся возможностей:
- уменьшение вероятности отказа – распределение по зонам,
- контроль времени ремонта – резерв узлов с дисками на период ремонта,
- долговечности – удаление старых данных,
- сохранности нет – доп. пространство для обслуживания

Постоянный контроль за состоянием



Заключение

- Нюансы гипер конвергентных архитектур в **большом** вычислительном кластере коллективного пользования:
  - Серверная нагрузка ввода-вывода мешает исполняющимся заданиям
  - Трудно обеспечить стабильную работу системы хранения
  - Необходимо распределять RAID диски по зонам надёжности
- IO сервера выделяются из однородного решающего поля, т.к. их нагрузка может достигать 40% от производительности
- Применяются надёжные системы хранения, надёжные/производительные, производительные
- Архитектура системы хранения + архитектура файловой системы + архитектура сети определяют наилучшую конфигурацию потоков ввода-вывода относительно количества вычислительных узлов, занятых задачей
- Улучшать надёжность системы можно и нужно, всегда учитывая компромисс надёжно-производительность