Отдел суперкомпьютерного моделирования

Международная научная конференция «Суперкомпьютерные дни в России»

Москва
26.09.2022

Сравнение производительности параллельной СХД суперкомпьютера с разными версиями файловой системы Lustre

Чулкевич Р.А., Козырев В.И., Шамсутдинов А.Б., Костенецкий П.С.

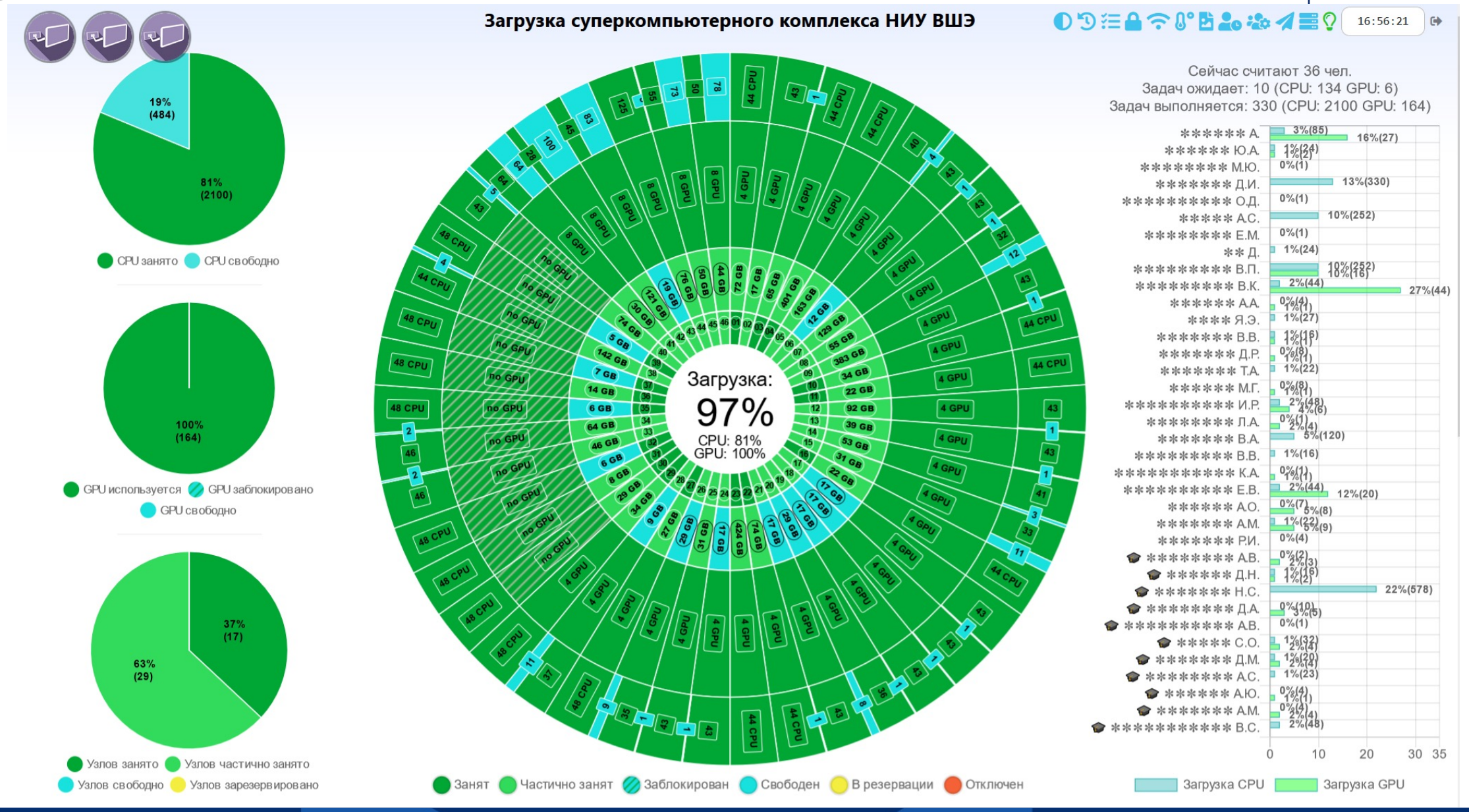
Характеристики суперкомпьютера *cHARISMa* (*Computer of HSE for Artificial Intelligence and Supercomputer Modelling*)

- **10 место в ТОП 50 СНГ**
- Пиковая производительность: **2 Петафлопс** (2 квадриллиона операций в секунду над числами с двойной точностью)
- LINPACK-производительность: **927.4 Терафлопс**
- **46** вычислительных узлов
  - **6** узлов **с 1 ТБ** ОЗУ**, 8 GPU A100 80 ГБ SXM**
  - **10** узлов с **1,5 ТБ** ОЗУ, 4 GPU V100 32 ГБ
  - **19** узлов с **768 ГБ** ОЗУ, 4 GPU V100 32 ГБ
  - **11** узлов с 384 ГБ ОЗУ для расчётов на CPU
- **2** управляющих узла
- **148 GPU NVIDIA Tesla A100 80 ГБ**
- **16 GPU NVIDIA Tesla V100 32 ГБ**
- **2584** ядер центральных процессоров
- Оперативная память: **40,3 ТБ RAM + 7.5 ТБ GPU Memory**
- Дисковая память: **1,15 ПБ**
  - параллельная СХД на базе Lustre **840 ТБ**
  - локальные диски **128 ТБ**
  - сервер резервного копирования **182 ТБ**
- Коммуникационная сеть: **2 x InfiniBand EDR** (**2x100 Гбит/с**, топология **FatTree**)

Суперкомпьютер "cHARISMa" в цифрах

- **379** активных пользователей (всего 974)
- **64** подразделений НИУ ВШЭ активно используют суперкомпьютер
- **822 000** задач выполнено пользователями
- **91** научный проект сейчас в работе
- **179** проектов выполнено за 2020-2021 г.
- **106+** научных статей опубликовано пользователями за 3 года
- **25** научных статей в журналах из I квартиля SCOPUS
- **1130** выполнена техподдержка по заявкам пользователей
- **24/7** обеспечена бесперебойная работа суперкомпьютера

Загрузка суперкомпьютера



HPC TaskMaster - система для обнаружения неэффективных и некорректно запущенных вычислительных задач

1. Kostenetskiy, P. S., Shamsutdinov, A. B., Chulkevich, R. A., Kozyrev, V. I., Antonov D. A. (2022) HPC TaskMaster – Task Efficiency Monitoring System for the Supercomputer Center. Parallel computational technologies (PCT) 2022

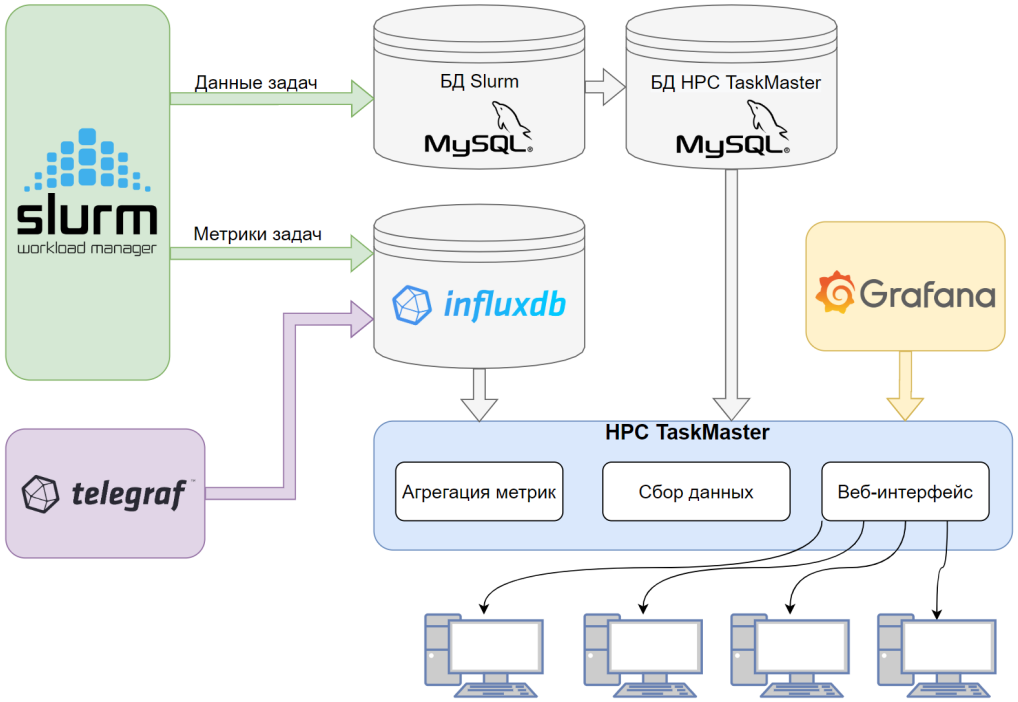
2. Voevodin, V. V., Chulkevich, R. A., Kostenetskiy, P. S., Kozyrev, V. I., Maliutin, A. K., Nikitenko, D. A., Rykovanov, S. G., Shamsutdinov, A. B., Shkandybin, Y. N., & Zhumatiy, S. A. (2021). Administration, Monitoring and Analysis of Supercomputers in Russia: a Survey of 10 HPC Centers. *Supercomputing Frontiers and Innovations*, *8*(3), 82–103

3. *Костенецкий П.С., Шамсутдинов А.Б.* Разработка системы мониторинга эффективности задач на суперкомпьютере cHARISMa // Параллельные вычислительные технологии ПаВТ'2021, г. Волгоград

4. *Костенецкий П.С., Шамсутдинов А.Б., Чулкевич Р.А., Козырев В.И.* HPC TaskMaster – система мониторинга эффективности задач суперкомпьютера // Суперкомпьютерные дни в России: труды международной конференции (27-28 сентября 2021 г., г. Москва). Москва: Издательство МГУ, 2021.

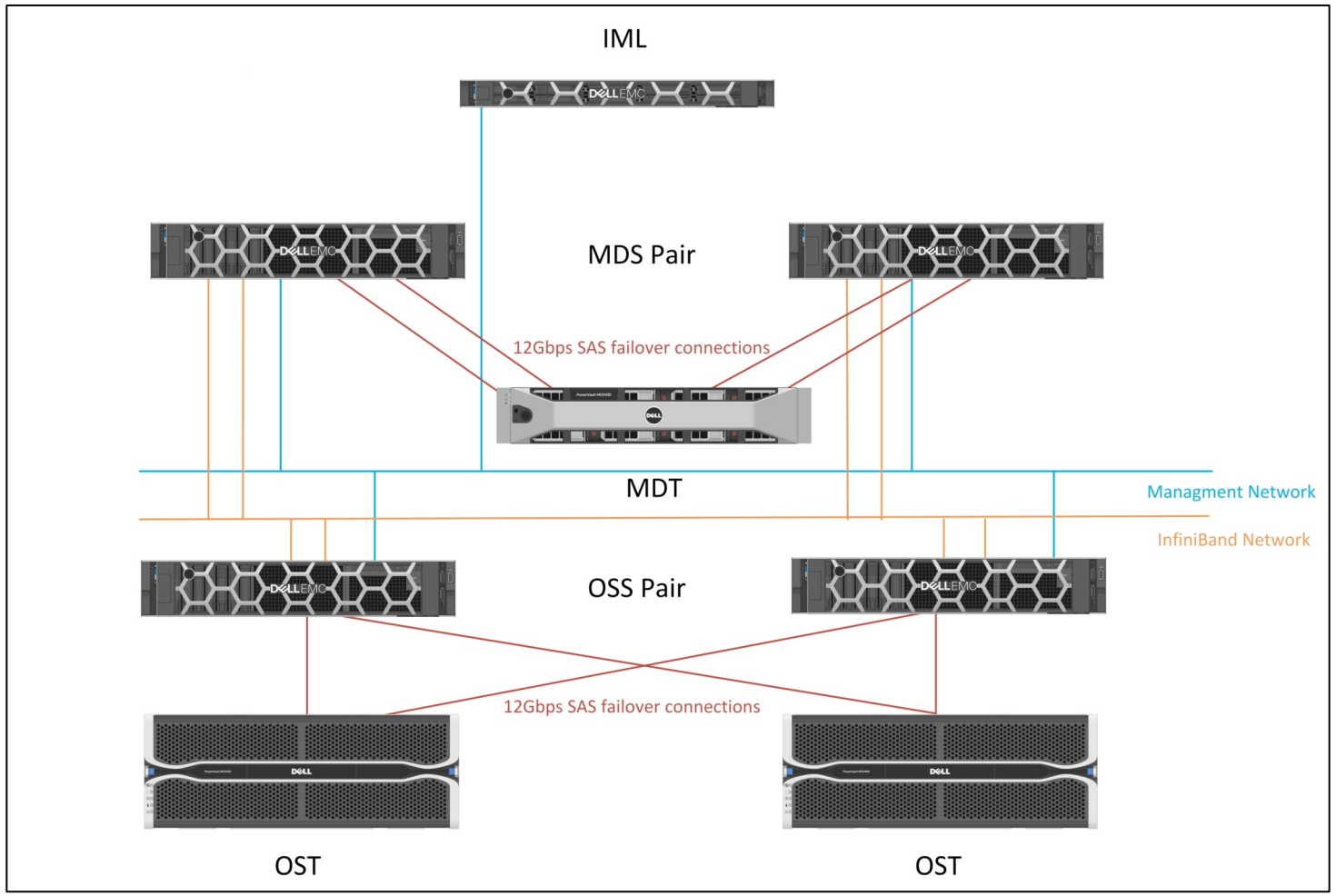
5. Open Source / HPC TaskMaster · GitLab. URL: https://git.hpc.hse.ru/open-source/hpc-taskmaster

Личный кабинет пользователя суперкомпьютера: https://lk.hpc.hse.ru



- Открытый исходный код
- Система собирает информацию о задачах, а не об узлах кластера
- Метрики задач автоматически анализируются на наличие проблем
- Для каждой задачи формируется вывод
- Система интегрирована в личный кабинет пользователя суперкомпьютера

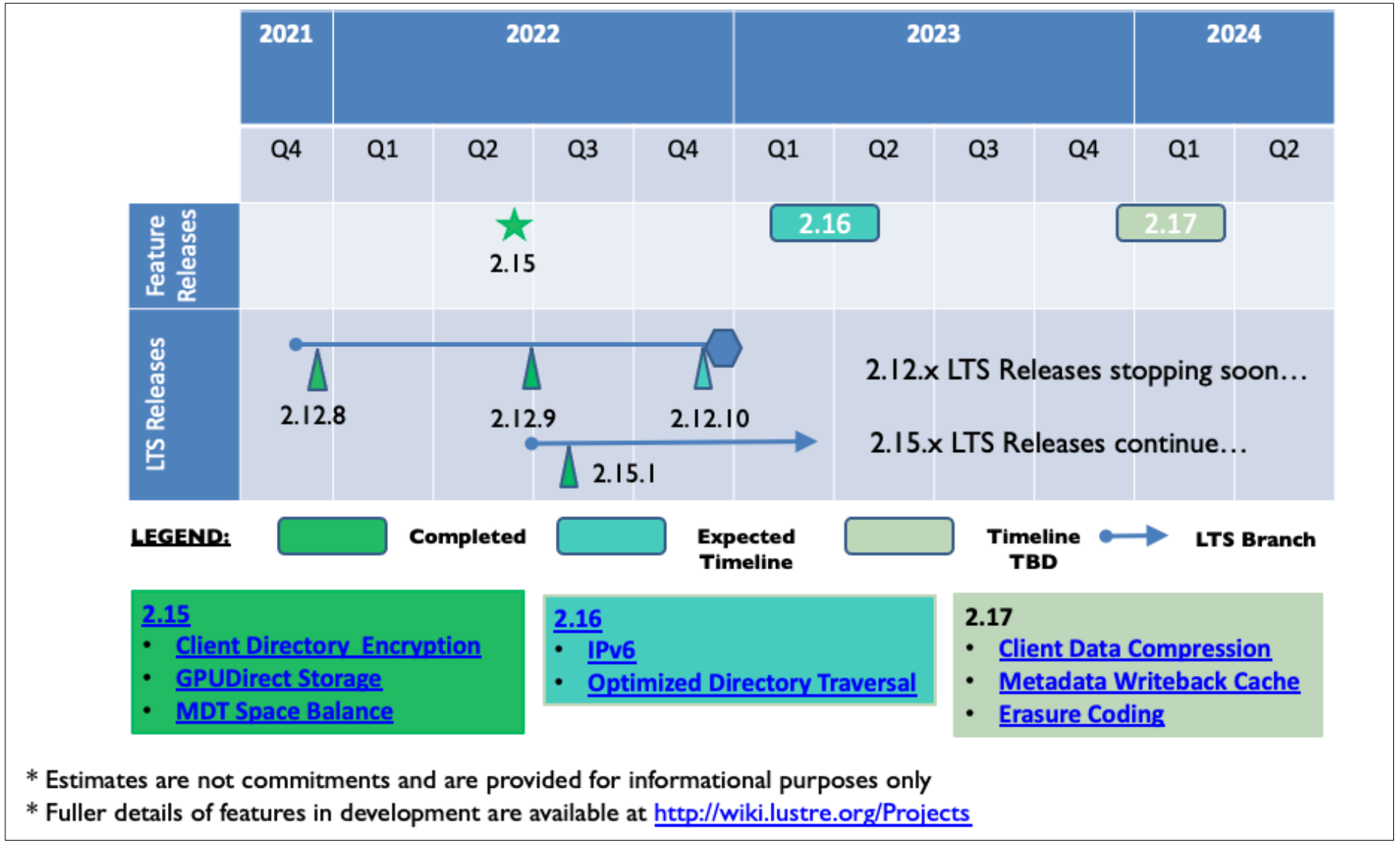
Файловая система суперкомпьютера - Lustre



Lustre – параллельная и распределённая файловая система, широко используемая в суперкомпьютерах. Lustre строится на базе нескольких серверов и объектных хранилищ:

- MDS (Meta Data Server): сервер, управляющий доступом к файловым метаданным – имени, размеру, владельцу, расположению и т.д.
- MDT (Meta Data Target): хранилище для файловых метаданных.
- OSS (Object Storage Server): сервер, управляющий доступом к содержимому файлов.
- OST (Object Storage Target): хранилище для содержимого файлов.

Lustre Roadmap: обновление LTS



Версии
до обновления:

- Клиент Lustre – 2.11.0
- Сервер Lustre – 2.10.6
- IML – 4.0.9

В конце II квартала 2022 вышла LTS-версия 2.15, несшая в себе : GPUDirectStorage – технология, позволяющая ускорить процессы обучения нейросетей и прочего IO-взаимодействия GPU с файловой системой.

https://www.lustre.org/roadmap

Сборка Lustre для Centos 7

Матрица совместимости с портала wiki.lustre.org

Lustre Version	2.12*	2.13	2.14	2.15
Release Date	2018-12-21	2019-12-5	2021-02-19	2022-06-16
Server	RHEL 7.6	RHEL 7.7	RHEL 8.3	RHEL 8.5
Client	RHEL 7.6, SLES12 SP3, Ubuntu 18.04	RHEL 7.7, SLES12 SP4, Ubuntu 18.04	RHEL 8.3, SLES15 SP2, Ubuntu 20.04	RHEL 8.5, SLES15 SP3, Ubuntu 20.04

Коммит из git проекта

git://git.whamcloud.com / fs / lustre-release.git / commit

summary | shortlog | log | commit | commitdiff | tree
(parent: 3e5dc84) | patch

LU-15875 kernel: kernel update RHEL7.9 [3.10.0-1160.66.1.el7] 97/47397/2

```
author     Jian Yu <yujian@whamcloud.com>
           Thu, 19 May 2022 21:20:41 +0300 (11:20 -0700)
committer  Oleg Drokin <green@whamcloud.com>
           Sat, 11 Jun 2022 09:02:45 +0300 (06:02 +0000)
commit     8bb18332e3cf5fb84a0a74c212dbbcacca6ca33d
tree       0e73e9367bddf70f443ff47bcf4d2e02acdb1264        tree | snapshot
parent     3e5dc84be447e16a84d02dc5b4400bc547c52459        commit | diff
```

```
LU-15875 kernel: kernel update RHEL7.9 [3.10.0-1160.66.1.el7]

Update RHEL7.9 kernel to 3.10.0-1160.66.1.el7.

Test-Parameters: trivial clientdistro=el7.9 serverdistro=el7.9

Change-Id: I9e8ab33edd6cacbbf895399962027827a1befd5b
Signed-off-by: Jian Yu <yujian@whamcloud.com>
Reviewed-on: https://review.whamcloud.com/47397
Tested-by: jenkins <devops@whamcloud.com>
Tested-by: Maloo <maloo@whamcloud.com>
Reviewed-by: Yang Sheng <ys@whamcloud.com>
Reviewed-by: Minh Diep <mdiep@whamcloud.com>
Reviewed-by: Oleg Drokin <green@whamcloud.com>
```

```
lustre/ChangeLog                                  diff | blob | history
lustre/kernel_patches/targets/3.10-rhel7.9.target.in  diff | blob | history
lustre/kernel_patches/which_patch                 diff | blob | history
```

Lustre primary development and releases

Сложности обновления

Проблемы зависимостей:

- модуль ядра Lustre зависит от модуля ядра MLNX (InfiniBand)
- модуль ядра MLNX может зависеть от версии ядра (или релиза ОС)

Проблемы совместимости:

- старые версии Lustre (<2.12) несовместимы с новыми версиями MLNX (>5.0)
- новые версии Lustre (>2.14) несовместимы со старыми версиями MLNX (<5.0)
- клиент Lustre не требует модифицированного ядра для работы
- сервер Lustre требует модифицированного ядра для работы

Для обновления Lustre (и клиента, и сервера) нужно:

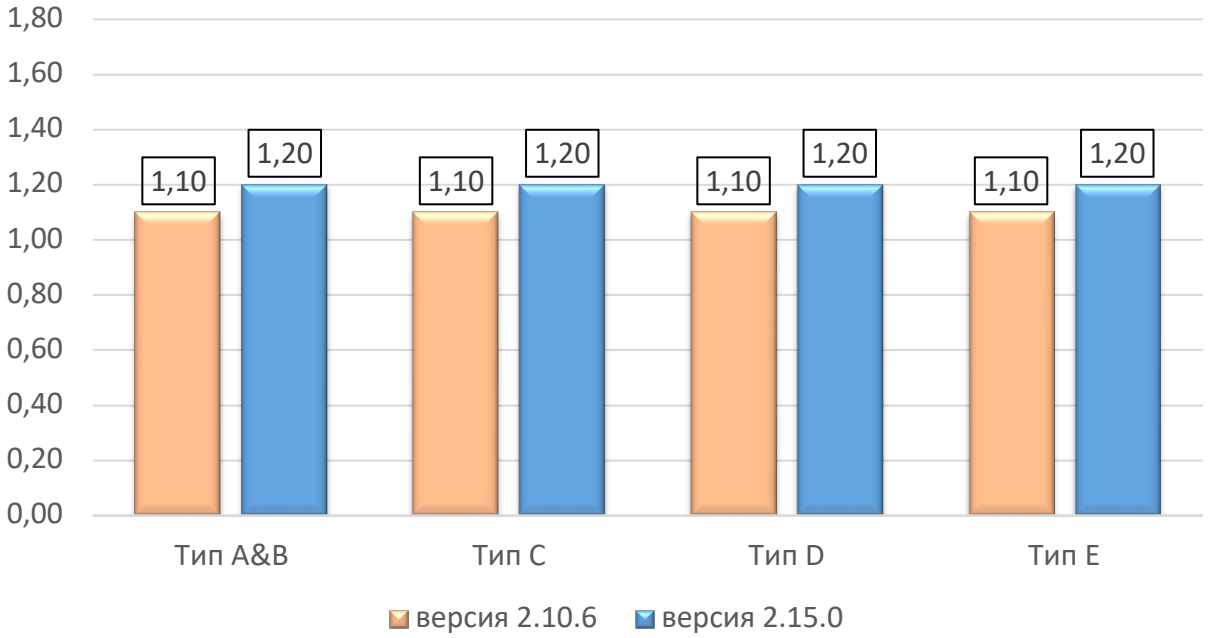
- обновить ядро,
- обновить модули ядра (в частности для InfiniBand) до совместимости,
- выполнить патчинг и сборку ядра для серверов Lustre,
- выполнить сборку клиент-части (на обычном ядре) и сервер-части Lustre (на патченном ядре).

Версии системного программного обеспечения на суперкомпьютере

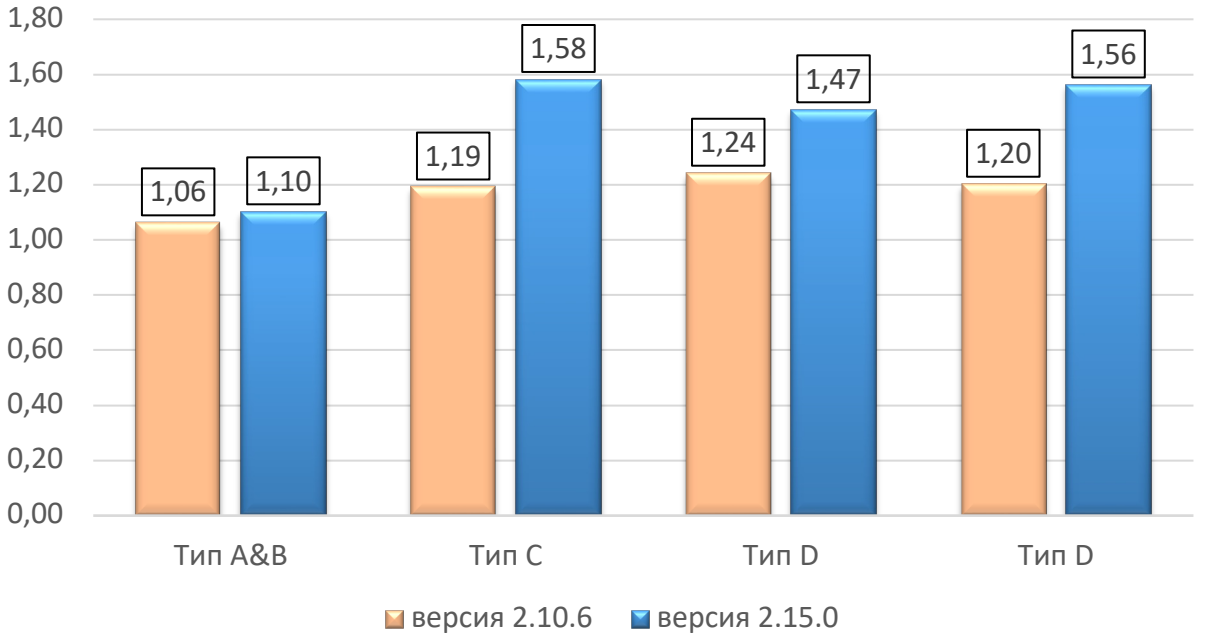
Программное обеспечение	До обновления	После обновления
Клиент Lustre	2.11.0	2.15.0
Сервер Lustre	2.10.6	2.15.0
Ядро на вычислительных узлах	3.10.0-957.5	3.10.0-1160.59
Ядро на серверах Lustre	3.10.0-957.5	3.10.0-1160.49
Драйвера Mellanox (InfiniBand)	4.5-1.0.1.0	5.6-1.0.3.3

Результаты сравнения производительности

СКОРОСТЬ ЧТЕНИЯ (AVG), GB/s



СКОРОСТЬ записи (AVG), GB/s



Тип узла	Конфигурация	Ускорение чтения, %	Ускорение записи, %
A,B	Dell C4140K, Xeon Gold 6152, 4xV100	9	3.8
C	Dell C4140M, Xeon Gold 6240R, 4xV100	9	32.8
D	Dell R640, Xeon Gold 6248R	9	18.5
E	HPE XL675dG10, EPYC 7702, 8xA100	9	30