Сравнение производительности параллельной СХД суперкомпьютера с разными версиями файловой системы Lustre

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

Национальный исследовательский университет "Высшая школа экономики"

Суперкомпьютер "cHARISMa" [1] активно используется 64 подразделениями НИУ ВШЭ для проведения научных исследований и учебной работы. Суперкомпьютер представляет собой высокопроизводительный вычислительный кластер с 46 вычислительными узлами, и параллельной СХД. Шесть вычислительных узлов кластера оснащены восьмью GPU NVIDIA A100 80ГБ SXM в каждом, 29 узлов с большим объемом оперативной памяти 768-1536 ГБ оснащены четырьмя графическими ускорителями NVIDIA Tesla V100 32 ГБ SXM в каждом, а для задч не требующих GPU в составе кластера есть 11 вычислительных узлов без с более мощными центральными процессорами. Система хранения данных (СХД) суперкомпьютера построена на базе параллельной сетевой файловой системы Lustre [3]. СХД состоит из двух Object Storage Server (OSS), двух Lustre Metadata Service (MDS), Lustre Metadata Target (MDT), Integrated Manager for Lustre (IML).

В июне 2022 года отделом суперкомпьютерного моделирования НИУ ВШЭ выполнен большой комплекс работ по обновлению системного программного обеспечения суперкомпьютера "cHARISMa". Обновление проходило в несколько этапов. На всех вычислительных узлах кластера было обновлено ядро и релиз ОС, стек программного обеспечения вычислительной сети InfiniBand и микропрограммное обеспечение. Далее были обновлены клиенты Lustre на вычислительных узлах кластера. На заключительном этапе были обновлены ОС, драйвера и управляющее ПО серверов СХД Lustre. Версии используемого программного обеспечения приведены в таблице 1. Файловая система Lustre была обновлена до актуального LTS-релиза 2.15.0. Особенностью данной версии стала поддержка NVIDIA GPUDirect Storage [2], благодаря которой можно ускорить обучение искусственных нейронных сетей. Обновление Lustre позволило значительно повысить скорость работы с файлами и обеспечить совместимость с новейшими версиями прикладного и системного программного обеспечения.

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

Таблица 1. Версии системного программного обеспечения на суперкомпьютере

При обновлении отслеживались зависимости между новыми и старыми версиями программного, микропрограммного и аппаратного обеспечения. Чтобы сократить период обновления и не беспокоить пользователей, предварительно была создана виртуальная копия суперкомпьютера со всем установленным ПО и на ней отработаны сценарии обновления. Подготовительные работы длились три месяца. Далее сценарий обновления был выпол-

нен «на чистовик» за 48 часов. В результате работ сохранены все данные пользователей и обеспечена обратная совместимость, позволяющая запускать ранее подготовленные пользователями научные задачи без перекомпиляции.

Для сравнения производительности старой и новой версиий программного обеспечения СХД проводилось тестирование скорости чтения/записи с использованием утилиты *dd*. На каждом вычислительном узле 25 раз выполнялась работа с файлами размером 1 ГБ. Полученные результаты на тестах записи для четырех типов вычислительных узлов с различными характеристиками [1] приведены в таблице 2.

Вычислительные узлы	**До обновления**	**После обновления**	**Ускорение,%**
	AVG, GB/s	AVG, GB/s	
01-26 Dell C4140K, Xeon Gold 6152	1.06	1.1	3.77
26-29 Dell C4140M,Xeon Gold 6240R	1.19	1.58	32.77
30-40 Dell R640, Xeon Gold 6248R	1.24	1.47	18.54
41-46 HPE XL675dG10+,EPYC7702	1.2	1.56	30

Таблица 2. Сравнение скорости записи до и после обновления

В результате обновления ОС и файловой системы Lustre существенно выросла средняя скорость записи. При относительно похожих конфигурациях, наибольшее ускорение от обновления получили вычислительные узлы на базе CPU Intel Xeon Gold 6240R и AMD EPYC 7702 (32,8% и 30% соответственно), а вычислительные узлы с более старыми процессорами Intel Xeon Gold 6152 дали меньшее ускорение - 3,8%. Предположительно, разница в ускорении вызвана оптимизацией новой ОС и Lustre под новейшее программное и аппаратное обеспечение.

Также после обновления значительно повысилась эффективность системы кэширования файловой системы. При повторном обращении к файлу на СХД скорость его чтения увеличивается, т.к. содержимое файла размещается в кэше файловой системы. Ускорение достигает четырех раз при доступе с вычислительного узла, выполнявшего первое чтение, и до трех раз – с других узлов.

Благодаря произведенному комплексу работ, выполнение программ, осуществляющих ввод/вывод в файлы, заметно ускорилось, что повысило общую производительность суперкомпьютера НИУ ВШЭ. Авторы рекомендуют обновление ОС и файловой системы Lustre до версии 2.15.0 на всех современных вычислительных кластерах.

Литература

1. Kostenetskiy P.S., Chulkevich R.A., Kozyrev V.I. HPC Resources of the Higher School of Economics // Journal of Physics: Conference Series. 2021. Т. 1740, № 1.

2. NVIDIA GPUDirect Storage Benchmarking and Configuration Guide:: NVIDIA GPUDirect Storage Documentation. (n.d.). Retrieved May 10, 2022, from https://docs.nvidia.com/gpudirect-storage/configuration-guide/index.html

3. Dai, D., Gatla, O. R., Zheng, M. (2019). A Performance Study of Lustre File System Checker: Bottlenecks and Potentials. IEEE Symposium on Mass Storage Systems and Technologies, 2019-May, 7–13. https://doi.org/10.1109/MSST.2019.00-20