Статистический анализ загрузки кластера «Уран»*

А.С. Игумнов

ИММ УрО РАН

В 2012 году на кластере «Уран» была запущена система тотального сбора показателей загрузки вычислительных узлов. С периодичностью в одну минуту с каждого из узлов собирается информация о загрузке CPU и GPU, выделении памяти, использовании сетей Ethernet и Infiniband. Главной целью внедрения этой системы является создание предпосылок для объективной оценки соответствия конфигурации кластера потоку вычислительных задач.

В основу архитектуры системы сбора статистики были положены два базовых принципа:

1. непрерывный сбор информации о нагрузке со всех узлов;
2. независимость от системы управления заданиями.

Сбор статистики на узлах выполняется с помощью модифицированной утилиты collectl [2]. Из собираемых данных удалены параметры, слабо связанные с отладкой кластера, и добавлены специфические параметры, например, загрузка сети Infiniband. При анализе статистики данные о нагрузке узлов объединяются с информацией о выполнявшихся на этих узлах прикладных задачах, извлекаемой из лог-файлов системы управления потоком задач. В настоящий момент написаны модули, обрабатывающие лог-файлы систем запуска СУППЗ и Slurm.

В процессе работы над системой сбора статистики были опробованы различные хранилища данных, в том числе, базы данных SQL. Был сделан вывод о том, что на текущем этапе развития проекта для хранения статистики вполне подходят плоские текстовые файлы с разбиением по месяцам и вычислительным узлам. Разбиение на отдельные файлы позволяет производить быструю выборку данных, относящихся к одной задаче, которая характеризуется группой выделенных ей вычислительных узлов и временем выполнения. При статистическом анализе загрузки кластера в целом во многих случаях достаточно иметь возможность прочитать все накопленные данные без необходимости делать какие-либо выборки или сортировки.

Примерная оценка объемов данных такова: 20 собираемых параметров, представленных в среднем 4-х разрядными десятичными числами, при сборе с периодичностью в одну минуту дают файл размером примерно 20x5x60x24 = 140 КБ в сутки на процессор, или 140 МБ в сутки для системы с 1000 вычислительных ядер. Такие объёмы без труда помещаются в оперативную память даже в тех случаях, когда необходимо проанализировать поведение задачи, использовавшей тысячу ядер в течение нескольких суток.

Отдельно рассматривались вопросы о возможности изменения в будущем набора измеряемых параметров и об обеспечении совместимости накопленных файлов статистики различных версий. Для обеспечения необходимой гибкости было принято решение использовать формат файла с разделителями и фиксированной семантикой отдельных полей.

В 2013 году на основе накопленных данных планируется провести поиск наиболее значимых для анализа поведения кластера наборов параметров и развитие соответствующей системы визуализации.

* Работа выполнена в рамках программы Президиума РАН № 18 "Алгоритмы и математическое обеспечение для вычислительных систем сверхвысокой производительности" при поддержке УрО РАН (проект 12-П-1-1034).