Управление нагрузкой в кластерном комплексе «Инпарком»

Ющенко Руслан Андреевич[1]

[1] *Институт кибернетики им. В.М. Глушкова НАН Украины, пр. Ак. Глушкова, 40, Киев, Украина*

yruslan@ukr.net

Аннотация. *Современные операционные системы не содержат собственных средств по управлению ресурсами и координации заданий вычислительного кластера. Поэтому разработчиками кластерных комплексов семейства «Инпарком» создали систему управления кластером «ACMS». Рассмотрены предпосылки, архитектура и функциональность этой системы, а также особенности реализации, позволившие защитить комплекс от теплового удара летом, от переохлаждения зимой и значительно сэкономить электроэнергию.*

Ключевые слова

Кластеры, системы управления кластером, мониторинг, температура, электроэнергия, администрирование

1 Введение

Хорошо спроектированная система должна иметь максимально легкий (но необязательно простой) для использования интерфейс, сколь бы ни была сложной ее внутренняя структура. В частности, кластерным комплексом справедливо можно назвать любое множество компьютеров, соединенных в локальную сеть и использующихся для решения общей (shared, а не general) задачи. Если речь идет о кластерах для высокопродуктивных вычислений, обычно предполагают однородность среды (аппаратуры, ОС, библиотек, и прочего), наличие инструментов распараллеливания в системах с распределенной памятью (MPI здесь вне конкуренции) и системы управления ресурсами (SLURM, TORQUE, OpenPBS, и т.д.). И хотя этого достаточно для выполнения параллельных вычислений, в процессе работы очень быстро возникают дополнительные требования.

Концептуально, кластерные комплексы семейства «Инпарком», которые Институт кибернетики им. В.М. Глушкова НАНУ разрабатывает совместно с ГНПП «Электронмаш», являются готовым инструментом, который с одной стороны предоставляет инфраструктуру для выполнения параллельных приложений и разработки параллельных программ, а с другой стороны содержат собственные параллельные библиотеки для решения задач [1]. За 6 лет работы линейка комплексов «Инпарком» включает диапазон от 8 до 512 ядер, но даже на начальных этапах исследований стало очевидным, что для эффективной работы комплекса необходимо разработать собственную систему управления. В результате была создана система «ACMS», состоящая из набора взаимодействующих служб. Круг ее обязанностей рос по мере разработки комплексов. Архитектура системы спроектирована для возможности добавления новых компонент, гибкой настройки и возможности быстро наладить ее функционирование на новом оборудовании и системном программном обеспечении [2]. Это позволило ее поддерживать одновременно для всех комплексов линейки, которые отличаются типами процессоров, типом сети, графическими процессорами и др.

Система управления позволила подойти централизованно к обеспечению выполнения правил и политик комплексов, что позволило решить такие административные задачи, которые обычными средствами практически не решаются. В частности, темой данного доклада является описание подсистемы управления нагрузкой комплекса. Согласно правилам на комплексах «Инпарком» узлы выключаются автоматически при превышении температурного порога, а включаются при наличии заданий в очереди. Как заставить эти правила выполняться автоматически, если его подсистемы функционируют независимо, т.к. разработаны совершенно разными группами людей? Для управления температурой, например, можно использовать Ganglia или Zabbix, а

для управления заданиями TORQUE или SLURM. Но как сделать так, чтобы они работали сообща? В системе управления кластером «ACMS» это достигается за счет *службы мониторинга оборудованием*, которую можно настроить получать температурные данные, например, в Ganglia, и *службы мониторинга заданий*, которую можно настроить работать как с TORQUE, так и SLURM. Настройка на конкретную подсистему потребует программирования, поскольку без стандартов невозможно заранее рассчитывать на определенный интерфейс подобных программ, но в целом такой подход обеспечивает гибкость. Если добавить к требованиям необходимость вести общий журнал, выводить графики нагрузки, учитывать процессорное время в разрезе пользователей, оповещать об аварийных ситуациях, выдавать отчеты о работе комплекса за период, и прочее – без использования централизованной системы управления кластером не обойтись. В данной работе бегло рассмотрены подходы, которые коллектив разработчиков комплексов «Инпарком» применил к проектированию такой системы, и которые впоследствии оказались удачными.

2 Система управления кластером

На рис. 1. проиллюстрирована роль системы управления кластером «ACMS» в комплексах «Инпарком».

Рис. 2. Роль системы управления кластером.

Основная идея – поставить между пользовательским интерфейсом и системным ПО комплекса промежуточный слой, задачей которого является координация, обеспечение безопасности, мониторинг и учет. Причем этот промежуточный слой является совершенно прозрачным для пользователя. Например, когда пользователь на любом из комплексов «Инпарком» заходит в терминал и запускает команду «mpirun» для запуска параллельной программы, он видит все так, как если бы он работал непосредственно в среде MPI. Но на самом деле команда «mpirun» является лишь оберткой, а «за кулисами» выполняется выделение необходимых узлов, постановка задачи в очередь, сохранение в базу данных графиков нагрузки и учетной информации. Причем аналогичные действия выполняются, если пользователь запускает параллельную программу из веб-интерфейса, либо если использует так называемое «интеллектуальное ПО», подробнее про которое можно прочитать в работе [3].

Приведенный подход помогает контролировать задания, которые пришли из вне, например, через грид-инфраструктуру. Обычно, задание приходит, ему предоставляются ресурсы исключительно на основе доверия к виртуальной организации, представляющей пользователя. И больше об этом задании мало что удается узнать. Используя централизованную систему управления кластером, можно не только учесть потраченное процессорное время, но и посмотреть, какие ресурсы использовало задание, посмотреть журнал выполнения (все, что программа писала в stdout/stderr). Кроме того, информация о выполнении заданий помещается в реляционную базу данных, поэтому легко получить агрегированную выборку за период для составления всякого рода отчетов.

3 Мониторинг и управление нагрузкой

Мониторинг и управление – пассивная и активная сторона одного и того же процесса. *Мониторинг* подразумевает сбор данных о текущих заданиях (выполняемых либо ожидающих) и о состоянии оборудования. *Управление* подразумевает выполнение некоторых действий в случае необходимости. Собранные данные доступны администратору комплекса в графическом виде, как показано на рис. 2:

Рис. 2. Мониторинг нагрузки комплекса.

Данные об оборудовании собираются в основном из датчиков, доступных по сети по интерфейсу IPMI [4]. Этот интерфейс позволяет получить доступ к аппаратным функциям узлов кластера по сети. IPMI позволяет удаленно включать, выключать, сбрасывать компьютеры, интерактивно управлять компьютерами через текстовую и графическую KVM-консоль. На каждом узле запущена служба, которая периодически получает данные из датчиков узла и через определенные интервалы времени отправляет их на управляющий сервер, который аккумулирует их и сохраняет в БД. Такой подход хорошо масштабируется на большóе число компьютеров, т.к. службе мониторинга на управляющем сервере нужно только обрабатывать входящие данные от узлов и соответствующим образом корректировать свое состояние.

Кластер в процессе работы генерирует тепло, пропорциональное потраченной электроэнергии. Например, мощность полностью включенного комплекса «Инпарком-256» составляет 15-20 КВт в зависимости от нагрузки, плюс 50-100% для системы охлаждения. Такое количество энергии способно быстро нагреть комнату выше допустимой для оборудования температуры, что, в свою очередь, может привести к необратимым неполадкам. Для охлаждения помещения используют кондиционеры, но ведь они могут выйти из строя. В самом худшем случае персонал не успеет отреагировать на стремительно повышающуюся температуру в комнате. Система «ACMS» следит за температурой в помещении и при превышении соответствующих порогов: а) не пускает новые задачи на выполнении, б) прерывает выполняющиеся задачи, в) аварийно выключает узлы комплекса. Позже этот подход применили для того, чтобы включать комплекс зимой, если возникают неполадки с системой обогрева помещения, см. Рис. 3:

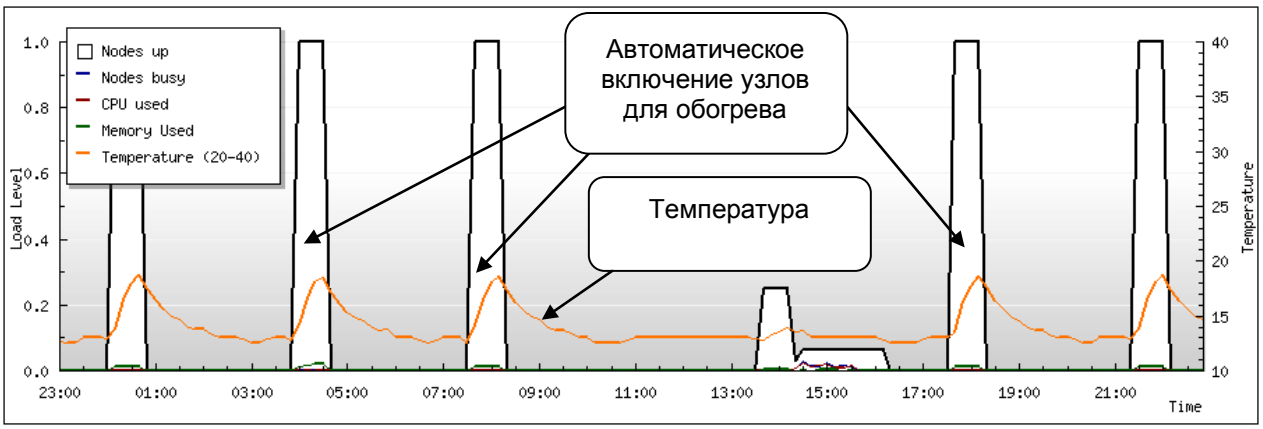

Рис. 3. Автоматическое включение зимой для самообогрева

«Зачем все узлы кластера включены, когда на нем в данный момент ничего не считается?» – это вопрос, которым задались исследователи после начала экономического кризиса в 2008 году [5]. Конечно, кластерный комплекс проектируется с расчетом на максимальную нагрузку, иначе он сам по себе не окупается. Но ведь нагрузка не равномерна, и бывают ситуации, когда комплекс простаивает со всеми включенными узлами. Служба управления нагрузкой использует данные о текущем состоянии кластера для определения того, когда его нужно автоматически включить или выключить. Если узел простаивает больше получаса – он выключается, если пришло новое задание и узлов не хватает – включается. Механизм включения – IPMI, но можно использовать и «WakeOnLAN». Время включения узлов до полной готовности – примерно 3 минуты, что совершенно не задерживает выполнение заданий, см. рис. 4:

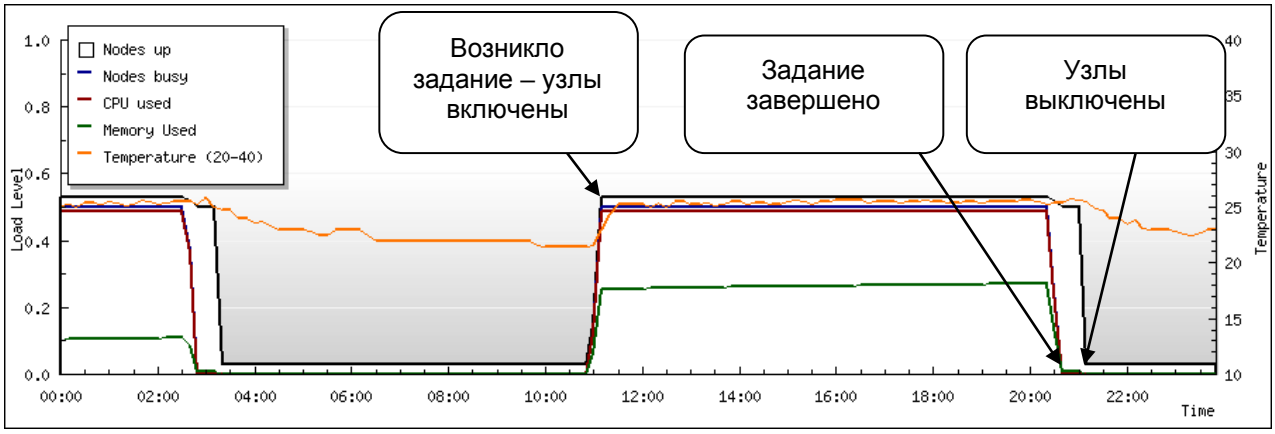

Рис. 4. Автоматическое включение и выключение на основе текущей очереди заданий

Хочется отметить, что аварийное включение/выключение компьютеров при понижении/повышении температуры и автоматическое включение/выключение на основе наличия заданий в очереди – потенциально конфликтные механизмы управления и поэтому должны быть реализованы в одной службе. Легко можно представить ситуацию, когда одна служба выключает компьютеры из-за превышения температуры, а другая – включает, т.к. в очереди есть задания. Имеем ситуацию, когда в одной и той же среде существуют два агента с противоположными целями. В этой связи надежнее всего использовать одну службу (одного агента), обладающую доступом к максимально полным и оперативным данным о состоянии комплекса, и принимающую решения о необходимости изменения этого состояния и выполняющую соответствующие действия.

4 Выводы

Сейчас становится очевидным, что кластерный комплекс как система нуждается в централизованном механизме мониторинга и управления. Подход, использующий систему управления кластером как набор взаимодействующих служб, примененный для контроля нагрузки в комплексе «Инпарком», показал свою эффективность, позволив тратить электроэнергию только тогда, когда на нем действительно выполняются вычисления, что позволило в разы снизить затраты.

Ссылки

[1] *Молчанов И.Н., Перевозчикова О.Л., Химич А.Н.* Опыт разработки семейства кластерных комплексов Инпарком // Кибернетика и системный аналіз. – 2009.– №6. – С. 88-96.
[2] *Ющенко Р.А.* Система управления кластером для комплексов семейства «Инпарком» // Проблеми програмування. – 2010.– №2-3. с. 155-161.
[3] *Молчанов И.Н., Мова В.И., Стрюченко В.А.* Интеллектуальные MIMD-компьютеры Инпарком – база для организации численных экспериментов в инженерии и науке //Технологические системы. – №40. – 2007. – С. 12-22.
[4] http://www.intel.com/design/servers/ipmi/
[5] http://www.rce-cast.com/components/com_podcast/media/10RCE-slurm.mp3