Разработка и моделирование системы управления энергопотреблением кластерных вычислительных систем

Д.В. Гордиенко

Новосибирский государственный университет
dimgordi@gmail.com

В настоящей работе исследуется возможность уменьшения затрат на энергоресурсы за счет отключения питания простаивающих вычислительных узлов ЦОД и повышения эффективности работы систем охлаждения. В силу того, что быстрые циклы включения/выключения могут напротив, повысить расход энергии, износ оборудования и время ожидания задач в очереди, отключение не всегда целесообразно. Таким образом, необходимо тщательное исследование работы кластера и разработка эффективного алгоритма отключения вычислительных узлов. Разработана технология сбора и обработки информации о текущей нагрузке на элементы системы электропитания, температуре отдельных компонент вычислительной системы и загрузке процессоров. Сформулированы критерии, характеризующие эффективность политики управления электропитанием, и подходы к моделированию системы.

Методика опробована на оборудовании информационно-вычислительного центра НГУ.

1. Введение

В процессе работы вычислительного комплекса существуют временные интервалы, когда пиковая загрузка кластера крайне высока, но когда она достигнет своего максимального значения, через какое-то время начинается спад. То есть загрузка кластера очередью задач начинает снижаться, по переменно то освобождая, то снова загружая вычислительные ресурсы. Данный факт может быть связан с целым множеством факторов, таких как отток пользователей к более новым вычислительным комплексам, окончание проектов крупных заказчиков, временные финансовые трудности клиентов итд.

Нами предлагается комплексный подход к мониторингу и анализу высвобождаемых вычислительных ресурсов кластера совместно с методом полного отключения питания неиспользуемых узлов, а также к контролю по работе охладительных систем в период низкой загрузки ЦОД.

Цель нашей работы состоит в повышении энерго-эффективности кластерных вычислительных систем. Энерго-эффективность, это отношение количества CPU-часов использованных для выполнения расчетов к объему потребленной системой электроэнергии. Для достижения этой цели мы разрабатываем программный комплекс (IDC - “Input Data Center”), позволяющий сократить энергопотребление за счет отключения простаивающих узлов кластера. IDC включает в себя компоненты мониторинга и управления. Существует ряд вычислительных комплексов, где реализованы аналогичные идеи, предназначенные для снижения энергопотребления кластера.

В настоящей статье изложены принципы, принятые за основу при проектировании собственного комплекса и оценки получаемой экономии для одной вычислительной системные для снижения энергопотребления кластера.

2. Описание IDC (Input Data Center)

Кластерная вычислительная система порождает последовательность задач. Каждая задача в свою очередь определяет набор параметров, необходимых для ее запуска, таких, например, как гарантированное необходимое время на выполнение задачи *(w)*, фактическое время выполнения *(t)* а также необходимое количество выделяемых вы-

числительных узлов *(N)*. Таким образом, последовательность задач может быть представлена последовательностью троек

$$\ldots (t_i, N_i, w_i), (t_{i+1}, N_{i+1}, w_{i+1}), \ldots.$$

Основным объектом разрабатываемого программного комплекса является Input Data Center («IDC»). Данный модуль выполняет функцию сбора информации со всего кластера, анализа полученных данных, принятия решения об отключении/включении узлов.

В процессе нашей работы мы реализовали сбор следующих данных:

1. *Температура.* В нашем распоряжении оказались датчики APC (расположенные внутри шкафа) и сенсоры в «лезвиях» - доступ к которым осуществляется с использованием интерфейса «iLO» (Integrated Lights-Out), полученная информация позволяет контролировать данный параметр внутри каждого узла, а также внутри каждого из четырех шасси. (Рис.1-b, c)
2. *Уровень влажности.* Датчик APC расположен внутри шкафа.
3. *Потребление электроэнергии.* Данные APC.
4. *Загрузка узла.* Данные, доступные в логах CMU (Cluster Management Utility): загрузка процессоров и оперативной памяти узлов. А также, что немаловажно, общее время работы системы после последней операции выключения или перезагрузки.
5. *Постановка задачи в очередь.* Информация планировщика задач (в нашем случае это продукт Altair – PBS Professional v.9.2).

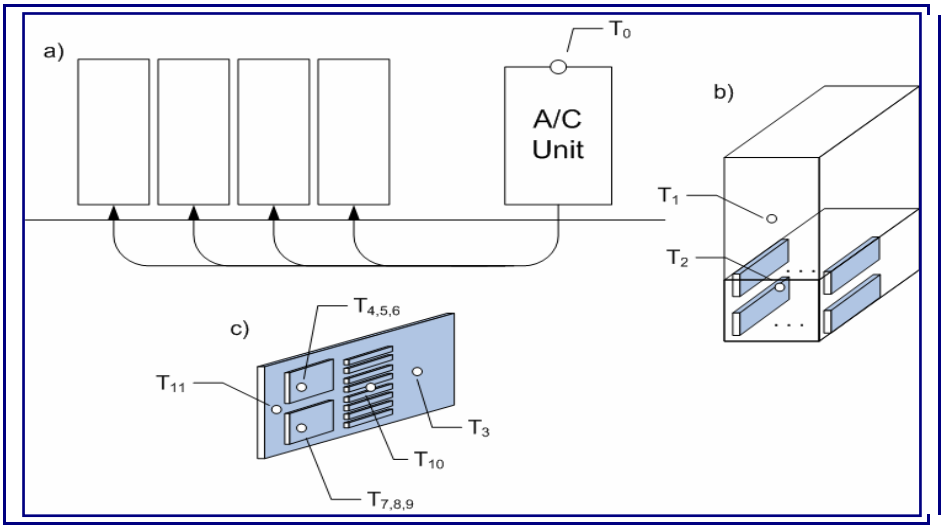

Рисунок 1.

Первостепенная задача IDC состоит в том, чтобы регулярно собирать доступные статистические данные, т.е. осуществлять мониторинг системы.
Предполагается, что связь со всеми компонентами будет осуществляться через коммуникационную среду кластера с использованием безопасного сетевого протокола Secure Shell (ssh) (Рис.2).

На следующем этапе, после получения информации, IDC приступает к ее анализу. Заключительный этап цикла состоит в том, чтобы зависимости от полученных результатов отправить команду на отключение/включение вычислительного узла или какой-либо компоненты системы охлаждения. Здесь важно выбрать наиболее адекватный алгоритм, который предполагает определение важнейших параметров системы в зависимости от особенностей функционирования ЦОД.

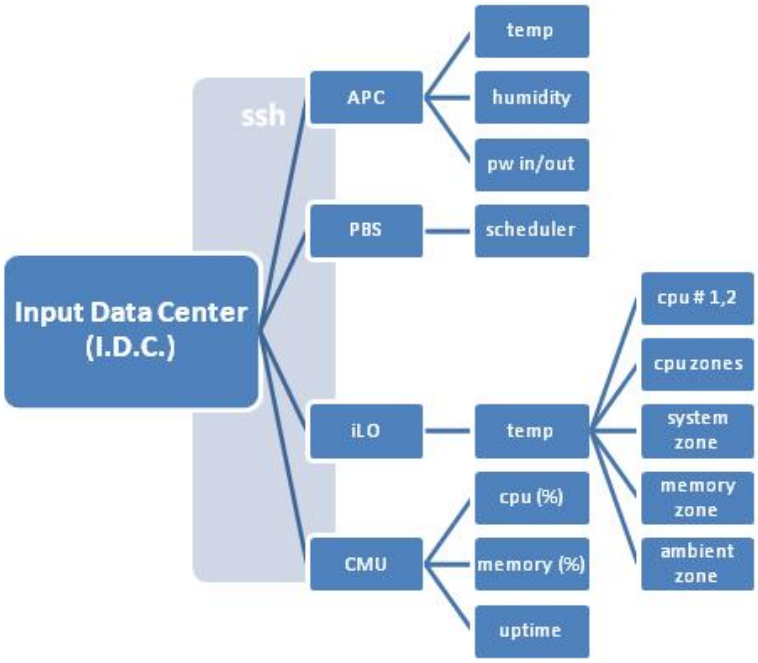

Рисунок 2.

После сбора полной статистики пора получить данные о свободных –простаивающих узлах и принять решение об их отключении. Возможностей определить, что узел свободен множество. Начнем с того, что мы всегда можем сконфигурировать планировщик задач таким образом, чтобы каждый раз, после окончания выполнения задачи, выполнялся какой-либо скрипт, который будет оповещать IDC о том, что в системе появился свободный узел. Однако, есть смысл в том, чтобы скрипт оповещал IDC не сразу, а через какой-то небольшой промежуток времени, т.е. дать возможность планировщику, если это необходимо, занять узел
новой задачей, если такова имеется.

Следующий этап - обращение к планировщику для анализа очереди задач. Алгоритм отключения должен избегать часто повторяющегося цикла включения/выключения, с целью снижения негативного влияния на долговечность оборудования кластера. Вместе с тем, необходимо не допустить значительного увеличения времени ожидания для «быстрых» задач. Для этого мы вводим параметр α —это число постоянно включенных «дежурных» узлов. α должно быть минимальным, но при этом удовлетворять особенностям функционирования кластера. Предполагается, что данный параметр определяется следующими методами:

1. Статистически, т.е. анализируем историю задач, которые требовали для расчета минимальное количество времени, выясняем, какое в среднем количество узлов они запрашивали.
2. Исходя из специфики задач, для которых предназначен кластер.

3. Теория массового обслуживания как подход для построения математической модели

Нами рассматривается возможность представления вычислительных узлов и планировщика как единой системы массового обслуживания (СМО). Для этого рассмотрим, чем характеризуется СМО, и что из ниже перечисленного представлено в нашей системе:

1.Поток поступающих сообщений - в нашем случае поток сообщений представлен

очередью задач, поступающих в планировщик.
2.Система обслуживания – вычислительный кластер, включая коммуникационную среду и обслуживающий сервер.
3.Характеристики качества:
◦ Среднее время задержки задачи (время ожидания).
◦ Средняя длинна очереди.
◦ Энергопотребление
◦ Температура
4.Дисциплина обслуживания
• Алгоритм отключения простаивающих узлов.
• Построение имитационной модели в виде временной сети Петри высокого уровня.

4. Выводы и результаты

Разработанная схема минимально привязана к аппаратным и коммуникационным особенностям кластера. Ставка делается на использование стандартных средств, имеющихся в инструментарии большинства вычислительных кластерах. Имея информацию о распределении тепла по шасси (узлам) мы можем проанализировать, на сколько оптимально распределены задачи по узлам с точки зрения минимизации концентрации тепла внутри одного шасси.

Для определения алгоритма включения/выключения узлов, сформулированы следующие критерии:
1. Сокращение времени ожидания задач.
2. Сокращение объема потребляемой энергии.
3. Сокращение числа включений/выключений узлов.
4. Избежание циклов включения/выключения оборудования, следующих в короткой последовательности, для обеспечения безопасной эксплуатации оборудования. Определяющим фактором является разница температур в рабочем и выключенном состоянии. Если разница температур составляет 100 градусов, то контакты начинают сыпаться после 1000 циклов. С другой стороны, при условиях, близких к машинному залу, среднее время жизни медных контактов на полиамидной основе составляет более миллиона циклов. С предварительным прогревом, ожидаемое число циклов до отказа -- 10E15 [2].

На основании исследования последовательности задач, поступающей на реальный вычислительный кластер ИВЦ НГУ, мы можем сделать вывод о возможности сокращения энергопотребления примерно на 10%. Данная оценка является грубой, однако в будущем, именно построение адекватной математической модели должно дать более строгий ответ.

5. Основные перспективы развития и усовершенствования

В данной работе не была рассмотрена подробно технология управления системой кондиционирования. Упор делался на управление питанием вычислительных узлов, однако, очевидным является факт возможности реализации следующей схемы: имея информацию о распределении тепла по корзинам (шасси) мы можем проанализировать, на сколько оптимально распределены задачи по узлам с точки зрения минимизации концентрации тепла внутри корзины. В будущем, это даст возможность распределять задачи по лезвиям таким образом, чтобы тепловыделение лезвий было наименьшим, как следствие понижение температуры внутри монтажного шкафа, снижение нагрузки на систему кондиционирования, продление срока эксплуатации компонент кластера и в итоге повышение надежности всего центра обработки данных.

6. Краткий обзор исследований в данной области

Вопросом экономии электроэнергии для высокопроизводительных вычислительных систем некоторые специалисты задались достаточно давно. Первые значимые работы, где данная проблема рассматривалась в рамках вычислительных кластеров и центров обработки данных, начали появляться начиная с 2004-2005 годов. Некоторые, наиболее близкие и интересные работы бы хотелось выделить отдельно.

Во-первых, это совместное исследование корейских и австралийских ученых в области возможностей понижения энергопотребления ЦОД с применением технологии DVS (Dynamic Voltage Scaling) [1]. Во-вторых, работа Sandeep Gupta из Темпэ (Аризона, США), в которой подробно рассматривается вопрос тепловыделение кластера с целью оптимизации работы планировщика задач [3]. В-третьих, известная работа японских ученых из университета Киото, которая представляет собой результат почти 4 летнего исследования в области поиска возможных путей экономии электроэнергии для суперкомпьютера ВЦ университета Киото [4].

Однако все эти работы имеют свои особенности и были начаты в то время, когда некоторые предоставляемые сегодня функции для управления кластером были недоступны. Более того, ни в одной из работ не была произведена попытка разработки модели автоматизированной программной системы, позволяющей использовать все преимущества современных коммуникационных сред кластера, с целью выполнения одной из наиболее простых операции для ЭВМ — выключение и включение по мере необходимости. Вместе с этим, те результаты, которые были получены в выше указанных работах применимы в основном для слишком специфической конфигурации высокопроизводительного кластера.

Литература

1. Kyoung Hoon Kim, Rajkumar Buyya, Jong Kim - CCGrid 2007 – “Power Aware Scheduling of Bag-of-Tasks Application with Deadline Constraints on DVS-Enabled Data Centers”.
2. J. M. Kallis and M. D. Norris. Effect of steady-state operating temperature on power cycling durability of electronic assemblies. Technical report, Hughes Aircraft Company, 1996.
3. Sandeep K.S. Gupta – “Thermal Management of Data Centers Through Thermal-Aware Job Scheduling”, 2008.
4. Junichi Hikita, Akio Hirano, Hiroshi Nakashima - IPDPS 2008 – “Saving 200kWand $200 K/year by power-aware job/machine scheduling.”