XXIX Неделя науки СПбГТУ. Материалы межвузовской научной конференции. Ч.V: С.45-47, 2001.

УДК 681.324.06

А. П. Лубанец (5 курс, каф. АиВТ), М. В. Хлудова, к.т.н., доц.

ПОСТРОЕНИЕ ВЫСОКОПРОИЗВОДИТЕЛЬНЫХ РАСПРЕДЕЛЕННЫХ ВЫЧИСЛИТЕЛЬНЫХ СИСТЕМ НА ОСНОВЕ КЛАСТЕРОВ. ТЕХНОЛОГИЯ MOSIX

Основные направления совершенствования вычислительных систем это повышение их производительности, увеличение надежности и уменьшение показателя цена/производительность. Существует несколько подходов к решению этой проблемы. Первый из них – применение суперкомпьютеров. Характеризуется самой высокой производительностью при аналогичной стоимости и сравнительно небольшой надежностью, ввиду сложности архитектуры системы и трудоемкости обслуживания. Второй подход – применение высокопроизводительных серверов с резервированием самых критичных частей системы – памяти всех уровней иерархии и процессоров. Данные системы обладают небольшой производительностью, средней надежностью и более низким отношением цена/производительность. И, наконец, третий подход, самый современный и интенсивно развивающийся, – построение распределенных вычислительных систем – кластеров, которые обладают самой высокой надежностью, низкой стоимостью и производительностью, приближающейся к суперкомпьютерной.

Основные цели, ставящиеся перед кластерными системами, можно разбить на четыре основных группы: отказоустойчивые системы, высокопроизводительные системы, системы распределения нагрузки, т.е. системы с высоким коэффициентом готовности, и системы, объединяющие в себе вышеперечисленные качества в том или ином соотношении. Задачи, решаемые при проектировании конкретного кластера, заключаются в точном определении его параметров и структуры, в зависимости от которых, можно получить или вычислительный инструмент, или отказоустойчивый сервер, или сервер с высоким коэффициентом готовности, или универсальную вычислительную систему с возможностями мейнфрейма.

Объектом исследования представленной работы является технология MOSIX (Multicomputer OS for UNIX), разрабатываемая в Hebrew University (Иерусалим, Израиль) с 1982 года. Помимо изучения всех аспектов и перспектив самой технологии, на её основе был построен действующий трехузловой кластер и произведено тестирование его производительности в “идеальном” и в реальном вариантах при различных механизмах IPC (именованные, неименованные каналы и очереди сообщений) на основе написанного программного обеспечения.

MOSIX – системное программное обеспечение и одноименная технология для реализации масштабируемых вычислительных кластеров. Распространяется по лицензии GPL GNU. Существует для нескольких платформ, в том числе и для ОС Linux. Достоинство MOSIX заключается в том, что ПО распространяется бесплатно. Данная технология характеризуется использованием уровня ядра для встраивания своих модулей и системного API, адаптивным алгоритмом распределения глобальных ресурсов кластера, простотой масштабирования при минимальных аппаратных затратах. Цель MOSIX-технологии – объединение большого количества компьютеров (до 65536) на основе сетей TCP/IP, вне зависимости от топологии, для совместной работы в качестве единой системы.

Алгоритмы функционирования MOSIX ориентированы на стохастический контроль изменения использования ресурсов узлов кластера, поддержку равномерной нагрузки на всех узлах кластера и миграцию процессов с более загруженного узла на простаивающий при дисбалансе системы хотя бы на одном узле. Основные алгоритмы MOSIX – алгоритм

балансировки нагрузки (load-balancing) и предотвращение истощения памяти (memory ushering). Они прозрачны для приложений и пользователя и препятствуют перегруженности процессоров и уменьшению количества свободной оперативной памяти каждого из узлов кластера. MOSIX – легко масштабируемая технология с динамическим формированием структуры, подразумевающей прозрачный вход/выход узлов в конфигурацию. Технология является попыткой улучшения общей производительности кластера при динамическом распределении и автоматическом перераспределении рабочей нагрузки, а также ресурсов между узлами вычислительного кластера любого размера. MOSIX поддерживает многопользовательский режим распределения вычислительного времени для выполнения как последовательных многопроцессных, так и параллельных задач. Технология подразумевает использование в качестве узлов кластера только вычислительные системы с одной и той же ОС. Сами узлы могут быть как однопроцессорными, так и SMP-системами. Если поступает запрос на изменение ресурсов кластера, процесс может автоматически под управлением ядра MOSIX мигрировать на другой узел сколь угодно раз, пока не будет достигнуто оптимальное распределение ресурсов (либо по истечению MTTL - Migration Time To Live). Управление миграцией процессов осуществляется как на уровне системного администратора, так и на уровне пользователя. Это особенно важно при запуске задач, привязанных к конкретному узлу, и задач с непредсказуемым использованием ресурсов и временем выполнения. Технологией не подразумевается понятия “главный узел”, т.к. все узлы кластера равноправны.

Описанная технология была детально изучена в процессе работы над реализацией кластера на ПО MOSIX версии 0.97.4 для ОС Linux (kernel 2.2.15). Тестирование построенного трехузлового кластера производилось в два этапа.

На первом этапе тестировалась модель так называемого “идеального” кластера, т.е. проверялась корректная работа алгоритма балансировки нагрузки “идеальным” программным обеспечением. При тестировании использовались программы с алгоритмом 100% несвязанных параллельных потоков решения гипотетической математической задачи. Априорно было показано, что производительность кластера должна быть 300% (по числу узлов). Эксперимент подтвердил предположение. Получены зависимости производительности кластера от числа узлов, зависимость времени выполнения от типа узла и количества узлов в кластере и рост производительности кластера в зависимости от используемого основного узла.

На втором этапе проводилось реальное тестирование работы кластера при выполнении многопроцессных гипотетических задач, взаимодействующих при помощи доступных для технологии MOSIX способов IPC (именованные и неименованные каналы, очереди сообщений). Для тестирования были написаны три класса программ, использующих перечисленные IPC. Каждый класс подзадач подразумевал под собой проведение трех экспериментов с разной интенсивностью обмена. Были получены диаграммы и численные значения для всех трех классов задач. Наиболее оптимальные результаты были достигнуты при использовании очередей сообщений, при этом пиковая производительность кластера была около 185%. Результатом данного тестирования явилось полное подтверждение концепции параллельных программ и программ с совпадающими участками кода.

На основании проведенных опытов были сделаны выводы о следующем:

- высокая производительность на полностью параллельных задачах;
- приемлемая производительность в задачах со средней интенсивностью IPC;
- практически нулевой выигрыш производительности на задачах с максимальной интенсивностью IPC;
- наиболее приемлемый способ IPC при использовании MOSIX – очереди сообщений;
- зависимость производительности от скорости процессора и объема памяти на стартовом узле;

- для получения большей производительности необходим переход к технологиям коммуникации Fast Ethernet, Myrinet, Gigabit Ethernet.

По результатам исследования, реализации и тестирования был сделан вывод о том, что MOSIX представляет собой новое поколение кластерных технологий, соответствует концепции Distributed OS (распределенная ОС) и является весьма перспективной для применения технологией получения универсальных вычислительных кластеров.