Анализ статистических данных отказов кластерных систем национальной лаборатории Лос-Аламоса

Е.С. Пименов, А.Ю. Поляков

Сибирский государственный университет телекоммуникаций и информатики, Новосибирск
Институт физики полупроводников СО РАН, Новосибирск

В данной работе описан созданный расширяемый инструментарий анализа статистических данных Лос-Аламоса, позволяющий определить такие параметры, как среднее время наработки на отказ, среднее время восстановления узла и т.п.

Введение

Распределенные вычислительные системы (ВС) являются основным инструментом высокопроизводительной обработки информации. Современные ВС большемасштабны, они формируются из $10^5 – 10^6$ вычислительных ядер, а их производительность достигает десятков PFLOPS [1]. При этом ВС – сложные технические комплексы, для организации эффективного отказоустойчивого функционирования которых требуется анализ и постоянное совершенствование архитектурных и алгоритмических решений.

Основной задачей распределенных ВС является решение поступающих задач, при этом в процессе их функционирования возможно возникновение отказов. Согласно статистическим исследованиям [2], основными причинами отказов ВС являются аппаратурные ошибки вычислительных узлов. В 2005 году национальной лабораторией Лос-Аламоса (США) были опубликованы наборы статистических данных, включающие в себя информацию о функционировании 23 вычислительных кластеров различных архитектур и конфигураций, собранную в течение 10 лет. Данная информация представляет большой интерес с точки зрения исследования распределенных ВС и может быть использована при построении их имитационных моделей, а также при принятии решений по их конфигурированию, например для определения оптимального интервала между формированием контрольных точек.

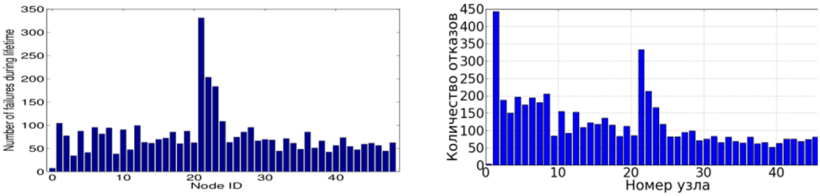
Отказы были классифицированы по следующим категориям:

1) аппаратурные сбои;
2) программные ошибки;

3) сетевые ошибки;
4) человеческий фактор;
5) ошибки в окружении;
6) неопределенные.

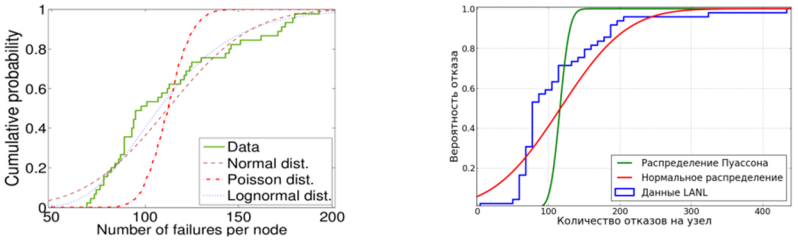
Процесс сбора статистики состоял из нескольких этапов. Ошибка обнаруживалась автоматической системой мониторинга. Оператор получал уведомление и добавлял в базу запись об отказе. Затем инженер устранял проблему, и оператор возвращал узел в систему, дополняя созданную запись информацией о времени устранения неисправности и ее причине. Поскольку в процессе присутствовал человеческий фактор, предоставленная статистика может быть неполной или некорректной. Следовательно, для ее анализа требуется фильтрация входных данных.

Наиболее полные исследования статистики были приведены в статье [2]. Авторами представлены различные показатели функционирования ВС. В частности, приведен график распределения количества отказов по вычислительным узлам системы №2 (рис. 1,а). Также была построена функция распределения количества отказов на узел и выполнено ее приближение тремя различными распределениями (рис. 2,а). Основной вывод авторов заключается в том, что нормальное и логнормальное распределение являются лучшим приближением по сравнению с распределением Пуассона.



а) б)

Рис. 1. Распределение количества отказов по узлам системы №2:
а) результаты из статьи [2], б) результаты, полученные разработанным инструментарием HPCStatTool



а) б)

Рис. 2. Функция распределения отказов по вычислительным узлам системы №2: а) результаты из статьи [2]; б) результаты, полученные разработанным инструментарием HPCStatTool

Инструментарий HPCStatTool

Авторами [2] выполнен анализ лишь части данных, доступных в статистике Лос-Аламоса. В частности, рассмотрено всего несколько систем. Инструментарий, с помощью которого были получены результаты, не был размещен в свободном доступе.

Для более подробного анализа доступных данных потребовалась разработка собственного инструментария, который был назван HPCStatTools. Созданный программный пакет позволяет:

1) гибко определять набор входных данных;
2) задавать методы их анализа;
3) выполнять визуализацию результатов.

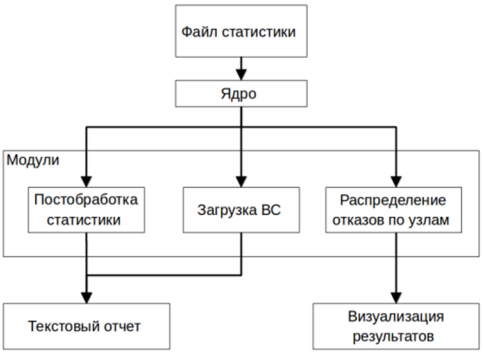


Рис. 3. Архитектура инструментария HPCStatTools

Архитектура разработанного инструментария представлена на рис. 3. Он состоит из ядра и набора расширений, обеспечивающих анализ различных аспектов статистических данных. На уровне ядра производится построчное чтение и проверка корректности входных данных из журналов статистики. Прошедшие проверку данные передаются независимо функционирующим модулям обработки статистики. Эти модули должны реализовывать интерфейс для обработки порции данных и формирования отчета.

Анализ отказов Лос-Аламосской национальной лаборатории

В данной работе приведены результаты исследований для систем из таблицы.

Таблица. Исследуемые системы Лос-Аламосской национальной лаборатории

Номер системы	Количество узлов	Количество ядер	Объем ОЗУ на одном узле, Гб
2	49	6152	128
3	128	512	4
16	16	2048	32
18	1024	4096	16
20	512	2048	16

С помощью разработанного инструментария было получено распределение отказов по вычислительным узлам системы №2 (рис. 1,б). Полученный график отличается от исходного, поскольку в оригинальной статье не был обозначен метод фильтрации исходных данных. Несмотря на это, результаты сопоставимы, а в большей части графика удалось получить точное совпадение результатов. Большой интерес представляет аналогичный график для системы №20, представленный на рис. 4. Очевиден неравномерный характер количества отказов в системе: узлы 1–256 отказывали значительно чаще. Нами было выдвинуто предположение о неравномерной загрузке системы, которое было подтверждено с применением имитационного моделирования, детали будут описаны далее. Система была загружена лишь наполовину, а в качестве алгоритма планирования использовался планировщик типа FIFO, использующий при распределении задач первые свободные узлы системы.

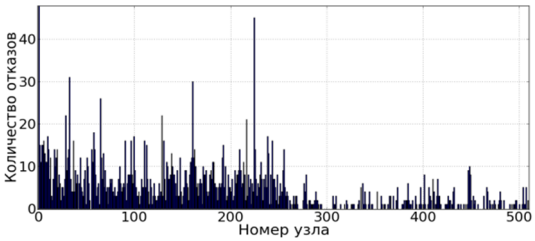


Рис. 4. Распределение отказов по вычислительным узлам системы №20

Также в работе было выполнено построение графиков функций распределения количества отказов на узел. Выводы в работе [2] справедливы не для всех случаев: согласно нашим наблюдениям нормальное распределение является лучшим приближением для SMP-систем, в то время как для кластерных ВС распределение Пуассона является достаточно качественным (рис. 5).

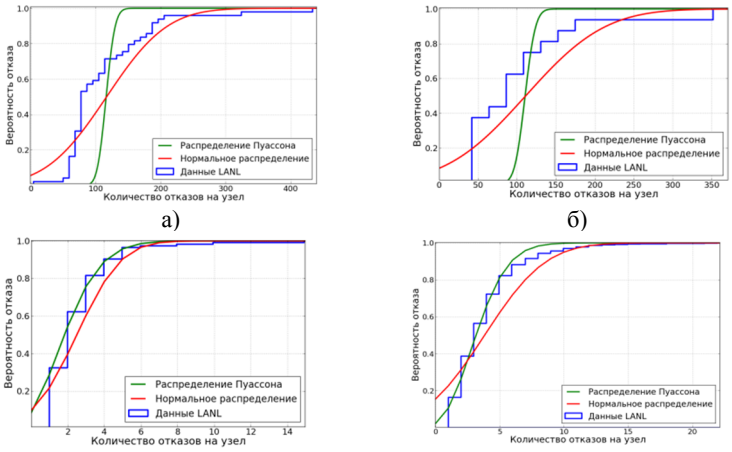


Рис. 5. Функции распределения отказов по узлам и их приближения для систем: а) 2, б) 16, в) 3, г) 18

Анализ загрузки ВС Лос-Аламоса

Помимо информации об отказах статистика Лос-Аламоса содержит данные загрузки систем. Разработанный в данной работе

инструментарий позволяет оценить характеристики потоков задач из представленной статистики.

На рис. 6 показано распределение времени решения задач, поступавших на систему №20. Значительное количество задач требовало 1, 5 и 5,5 часов для решения.

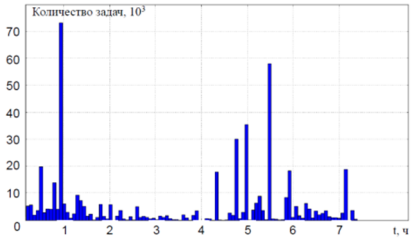


Рис. 6. Распределение времени решения задач на ВС №20

Разработанный инструментарий позволяет обработать статистику потока задач и привести ее к виду, удобному для использования в качестве входных данных имитационной модели. Также предусмотрена возможность получить восстановленную загрузку без проведения моделирования для проверки адекватности модели.

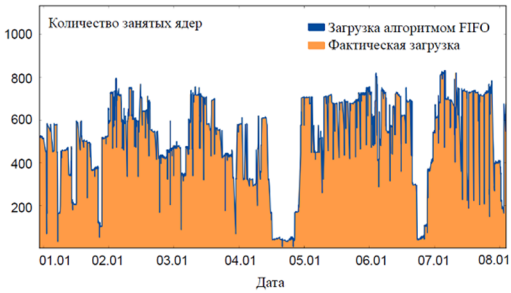


Рис. 7. Восстановленная и моделируемая загрузка системы №20

По извлеченной статистике было проведено имитационное моделирование системы, функционирующей под управлением FIFO-планировщика. На рис. 7 видно, что моделируемая загрузка полностью совпадает с восстановленной, что подтверждает корректность модели. Вместе с этим заметим, что пиковая загрузка

системы не превышает половину узлов, что подтверждает высказанное ранее предположение о причине неравномерного распределения отказов по узлам данной ВС.

Заключение

Разработан программный инструментарий HPCStatTools, который был адаптирован для обработки статистических данных об эксплуатации ВС лаборатории Лос-Аламоса. Выполнен анализ статистики отказов, который согласуется с результатами, опубликованными в работах других исследователей. Сделаны выводы по поводу причин дисбаланса распределения отказов в системе №20. Проведено исследование алгоритма распределения ресурсов на системе №20. Показано, что использовался простой планировщик FIFO. Выполнена визуальная оценка отказов в SMP–кластерах и кластерах, построенных на базе серийного оборудования. Полученные результаты свидетельствуют о корректности созданного инструментария.

Литература

1. *TOP500* [Электронный ресурс] – Режим доступа: http://www.top500.org/
2. *Schroeder B.* A large-scale study of failures in high-performance computing systems / Bianca Schroeder, Garth A. Gibson // Proceedings of the International Conference on Dependable Systems and Networks. – IEEE Computer Society Washington, DC, USA, 2006. P. 249–258.