ВЕРОЯТНОСТНО-СТОИМОСТНОЙ ПОДХОД К ОПТИМИЗАЦИИ РАСПРЕДЕЛЕННЫХ СИСТЕМ ХРАНЕНИЯ ДАННЫХ ФИЗИЧЕСКИХ ЭКСПЕРИМЕНТОВ

В.В. Трофимов, А.В. Нечаевский, Г.А. Ососков, Д.И. Пряхина [a]

Объединенный институт ядерных исследований,Россия, 141980, Московская обл., г. Дубна, ул. Жолио-Кюри, д. 6

E-mail: [a] pryahinad@jinr.ru

В рамках работ по созданию системы хранения и обработки данных установок *BM@N* и *MPD*, входящих в состав комплекса *NICA*, возникает проблема выбора оптимальной конфигурации необходимого компьютерного и сетевого оборудования. Для решения этой проблемы предложена и реализована схема макро-моделирования, рассматривающая процесс перемещения данных, как поток байтов, имеющий статистическую природу, без анализа отдельных частей этого потока. Для оценки различных конфигураций оборудования используется вероятностный подход, при котором определяются вероятности потерь информации, поступающей с детекторов для каждой из этих конфигураций, и выбирается с учетом экономических факторов та, для которой эта вероятность не превышает заданный предел, а цена минимальна. Описаны структура программы моделирования, предложенные статистические критерии качества работы системы и алгоритмы моделирования. Приведены результаты предварительных вычислений для выбора оптимальной системы приёма и хранения данных эксперимента BM@N.

Ключевые слова: физический эксперимент, моделирование, оптимизация распределенных систем хранения данных

1. Введение

В рамках работ по созданию системы хранения и обработки данных установок *BM@N* (*Baryonic Matter at Nuclotron*) и *MPD* (*Multi-Purpose Detector*), входящих в состав комплекса NICA (*Nuclotron based Ion Collider facility*) [1], возникает проблема выбора оптимальной конфигурации необходимого компьютерного и сетевого оборудования. Для решения этой проблемы требовалось разработать и исследовать модель перемещения данных внутри системы. Предыдущий опыт моделирования авторов [2-4] показал, что описанные в литературе подходы моделирования процессов обработки потока заданий в распределенных и облачных системах [5, 6] основаны на детализации потока данных до уровня пакета или файла, что неизбежно приводит к сложной организации программ и большим вычислительным затратам. В качестве примера можно указать, что такое детальное моделирование передачи данных с установки BM@N в течение только одной недели программой, описанной в [4], потребовало 4 млрд. расчетных циклов. Кроме того, следует признать, что подход с детальным описанием параметров компьютерных и сетевых систем в базах данных моделирующих программ для решения прогнозных задач малоперспективен в силу того, что эти параметры быстро устаревают, а излишняя детализация приводит к затруднениям в принятии оптимального решения по выбору системной конфигурации. К тому же, вышеуказанные программы не ориентированы на учет ценовых показателей моделируемых систем.

В то же время известны и другие подходы к прогнозному моделированию, относящиеся к сложным системам иной природы [7-9], опыт которых может быть учтен и в рассматриваемом случае. Речь идет о макроэкономическом вероятностно-стоимостном подходе, в котором упор сделан на такие наиболее существенные факторы, как критерии функционирования системы, например, потери или их отсутствие и экономические затраты (или прибыль) и их правдоподобие, т.е. вероятность осуществления тех или иных их значений.

Авторами статьи предложена и реализована схема моделирования, рассматривающая процесс перемещения данных, как поток байтов, имеющий статистическую природу, без анализа отдельных частей этого потока. Для оценки различных конфигураций оборудования используется вероятностно-статистический подход, при котором определяются вероятности потерь информации, поступающей с детекторов для каждой из рассматриваемых конфигураций. Оптимальной конфигурацией считается та, что имеет минимальную стоимость при заданном допустимом уровне потерь.

2. Подход к моделированию

При разработке *Technical Design Report* (*TDR*) [10] на систему сбора и хранения данных возникает проблема поиска конфигурации оборудования, которое обеспечит функционирование системы с заданным качеством при минимальной цене. Рассмотрим подход к поиску модели такой конфигурации, который основан на определении статистических зависимостей качества работы установки от ценовых вложений в оборудование хранения и передачи данных. Привлечение аппарата статистики и случайных чисел необходимо, поскольку некоторые параметры модели случайны (например, интенсивность потока данных), и для их оценки необходимо привлечение статистических методов, а для моделирования требуются генераторы случайных величин с заданными распределениями.

Специфика работы физической установки на ускорителе заключается в том, что набор данных с экспериментального триггера происходит не постоянно во время сеанса, а в течение активных периодов, которые прерываются для настройки аппаратуры, калибровки детектора, профилактики оборудования и т.д. Поток данных с детектора тоже не является в общем случае стационарной величиной. Если система включает хранилища нескольких уровней, то во время таких остановок новые данные поступать не будут, но уже поступившие могут быть переданы на следующие уровни для долговременного хранения. Количество информации, которую система может сохранить, определяется набором параметров, задающих размеры дисковых массивов, ширину каналов, типы ленточных накопителей, топологию и иерархию системы хранения.

Критерием качества работы системы была выбрана полнота сохранения информации за период моделирования, которая определяется как отношение объема сохраненной в системе к информации, произведенной детектором. Удобнее оперировать величиной, отражающей уровень потерь информации в системе, т.е. величиной, получаемой, как единица минус значение полноты. Такой критерий представляется логичным, поскольку абсолютная величина потерь будет зависеть от интенсивности событий в триггере детектора. Событие, при котором происходит потеря, случается, когда триггер системы поставляет информацию, но ее некуда разместить, так как сам триггер не располагает буфером.

Уровень потерь также будет величиной случайной, поскольку зависит от факторов, имеющих случайный характер. Поэтому в качестве численного выражения качества работы системы устанавливается предельное значение вероятности того, что потери не превысят заданного порога. Все конфигурации оборудования должны характеризоваться ценой. Под ценой здесь понимается метрика, включающая в себя различные факторы: стоимость оборудования, стоимость эксплуатации, требования к персоналу и т.д.

Суммируя вышесказанное, задачу поиска оптимальной конфигурации можно сформулировать так: среди множества конфигураций, для которых вероятность относительных потерь не превысят заданного порога, найти конфигурацию с минимальной ценой. Для упрощения будем считать, что, если в результате моделирования относительные потери не превышают заданный уровень, система данной конфигурации работоспособна.

Для оценки вероятности работоспособности системы проводят N вычислительных экспериментов с фиксированными параметрами конфигурации. В каждом эксперименте определяют относительные потери и подсчитывают количество экспериментов M, в которых относительные потери не превысили заданного значения. Отношение M к N будем считать оценкой вероятности того, что выбранная конфигурация работоспособна. Если ввести минимальный порог вероятности работоспособности системы, то оптимальной можно считать конфигурацию, в которой цена минимальная, а вероятность работоспособности системы выше заданного порога.

3. Поиск оптимальной конфигурации для системы сбора и хранения данных эксперимента *BM@N*

Рассмотрим процедуру вычисления оптимальной конфигурации в системе, обслуживающей конкретный детектор. В качестве источника информации выступает детектор событий эксперимента *BM@N*, последовательно соединенный с буфером приема данных (*DAQ*), фермой предварительной обработки с дисковым буфером для промежуточного хранения данных, фермой долговременного хранения с дисковым буфером и роботизированной библиотекой. (рисунок 1).

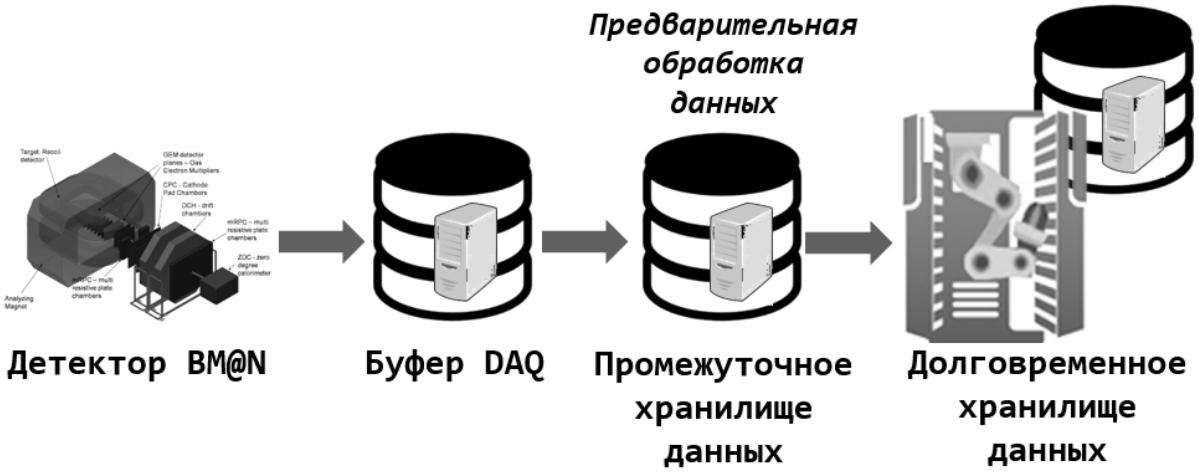

Рисунок 1. Схема системы сбора и хранения данных эксперимента *BM@N*

Поток данных возникает в детекторе, который производит буферы заданного объема и с определенной частотой. Все компоненты системы соединены между собой каналами передачи данных. В простейшем случае поток данных идет в одном направлении, от детектора к ферме долговременного хранения. В работе ускорителя и детектора возникают паузы, обусловленные, как запланированными событиями (калибровка, обслуживание аппаратуры), так и случайными отказами оборудования. Значения вероятностей таких отказов и длительности восстановления можно предположить, исходя из опыта эксплуатации аналогичных установок. Из-за перерывов в работе, обусловленных всеми перечисленными факторами, количество данных, поступающих в систему, будет меньше максимального значения, равного произведению размера буфера на частоту событий и на астрономическое время сеанса. Если по какой-то причине промежуточное звено не может принять данные (сбой, переполнение буфера), то объект, находящийся в потоке данных «выше» по иерархии не может записать их с последующим стиранием, и вынужден будет хранить у себя. Если этот объект – детектор, то данные будут потеряны, поскольку детектор не имеет собственного буфера.

В терминах модели детектор будет рассматриваться, как источник данных, фермы с дисковыми массивами, как участники, драйвы магнитофонов – также как участники, но со специфическим свойством восстановления свободного объема путем смены лент, а такое оборудование передачи данных, как каналы, процессы передачи данных с их характеристиками, рассматриваются как связи. Плановые события моделируются в простейшем случае последовательностью «стробирующих импульсов». Все дисковые массивы модели имеют ограниченный объем. При исчерпании свободного места запись в участника прекращается. Исключением является запись на ленту. Если заканчивается место на ленте, запись информации на этот драйв прекращается на время, необходимое для смены ленты, затем запись возобновляется.

В этой постановке рассмотрим общие вопросы применения вероятностно-стоимостного метода, а также процедуру поиска оптимальной конфигурации. Для того, чтобы определить уровень потерь при работе системы определенной конфигурации, надо рассчитать потери для всех возможных конфигураций установки. Разделим задачу перебора конфигураций и моделирования передачи данных на ее реализацию в двух программных модулях: модуле построения заданий (конфигураций) и модуле, реализующем саму модель. При построении заданий необходимо задать количество вычислительных экспериментов, общее для всех заданий, параметры конфигурации, если они фиксированные, или их диапазон и шаг изменения. Сложность задачи растет, как произведение количества вариантов по каждому параметру. Для систем с 10-20 параметрами количество заданий может достигать десятков тысяч. В общем случае модуль построения заданий должен иметь логический фильтр, отсеивающий задания с конфигурациями параметров, которые не могут существовать.

Модуль построения заданий формирует все возможные конфигурации системы и записывает их в базу данных в виде заданий на моделирование. После этого запускается диспетчер, который вызывает процесс моделирования для одной конфигурации (одного задания) из базы данных. Поскольку задания независимы друг от друга, порядок из запуска не имеет значения. Единственным требованием для запуска процесса оценки полных результатов моделирования является завершение всех заданий пакета. Диспетчер заданий выбирает задание из базы, отмечает факт выбора и вызывает модуль моделирования. Диспетчер может быть в единственном экземпляре, но, если количество заданий большое, имеет смысл использовать несколько диспетчеров на разных вычислителях, каждому из которых разрешен доступ к базе заданий.

Модуль, выполняющий задания, является, собственно, моделью перемещения информации в системе. Работа модели начинается с чтения параметров вычислительного эксперимента из базы данных. Затем заданное количество раз (в зависимости от продолжительности моделируемого эксперимента) выполняется программный цикл, который состоит из следующих шагов: (1) розыгрыш случайных величин; (2) моделирование процесса передачи данных от триггера до фермы хранения; (3) определение процента суммарных потерь за цикл и сравнение с заданным порогом.

Превышение порога считается отказом, после которого вычисления прекращаются, а объект данной конфигурации при разыгранных случайных величинах считается

неработоспособным. Факт неработоспособности объекта в одном из вычислительных экспериментов не означает, что такая конфигурация параметров не может быть применена, вывод о качестве работы можно сделать только на основании расчета вероятности по всем экспериментам. Такая последовательность вычислений повторяется заданное пользователем число раз. После этого вычисляется вероятность отказа для заданной конфигурации оборудования. После завершения работы всего пакета конфигураций, среди них выбирается та, или те, для которой вероятность не превышает заданный предел, а цена минимальна.

Предложенная схема допускает параллельный расчет вариантов, что позволяет анализировать значительное количество вариантов (десятки тысяч). Как перспектива развития этого подхода рассматривается создание классов, которые позволят гибко менять топологию системы хранения данных. В существующем варианте программа моделирования ориентирована только на анализ потерь при работе установок *BM@N* и *MPD*.

4. Результаты моделирования для эксперимента *BM@N*

В качестве примера рассмотрим задачу определения зависимости вероятности работоспособности системы сбора и хранения данных установки *BM@N* от ценовых вложений в систему дисковой памяти промежуточного хранилища данных. Для этого будем изменять количество одновременных потоков чтения данных с буфера *DAQ*, при этом для добавления одного дополнительного потока чтения шириной 150 МБ/сек необходимо 100 условных единиц стоимости. Детектор производит буферы объемом 0.5 МБ с частотой 3 кГц. Вероятность сбоя оборудования в любой момент времени равна 0.005, а время, необходимое на восстановление — 100 сек. Период моделирования составляет 48 часов. После предварительных расчетов были проанализированы вероятности работоспособности системы сбора с разным количеством потоков чтения данных при изменяющихся порогах возможных потерь (рисунок 2). Предположим, что изначально в системе 33 потока чтения. По графику видно, что для достижения вероятности работоспособности системы более 0.5% при 1% пороге допустимых потерь, необходимо добавить 3 потока чтения, т.е. вложить 300 условных ценовых единиц, но в случае, если допустимо 2% потерь, достаточно добавить 2 потока чтения, т.е. вложить 200 условных ценовых единиц.

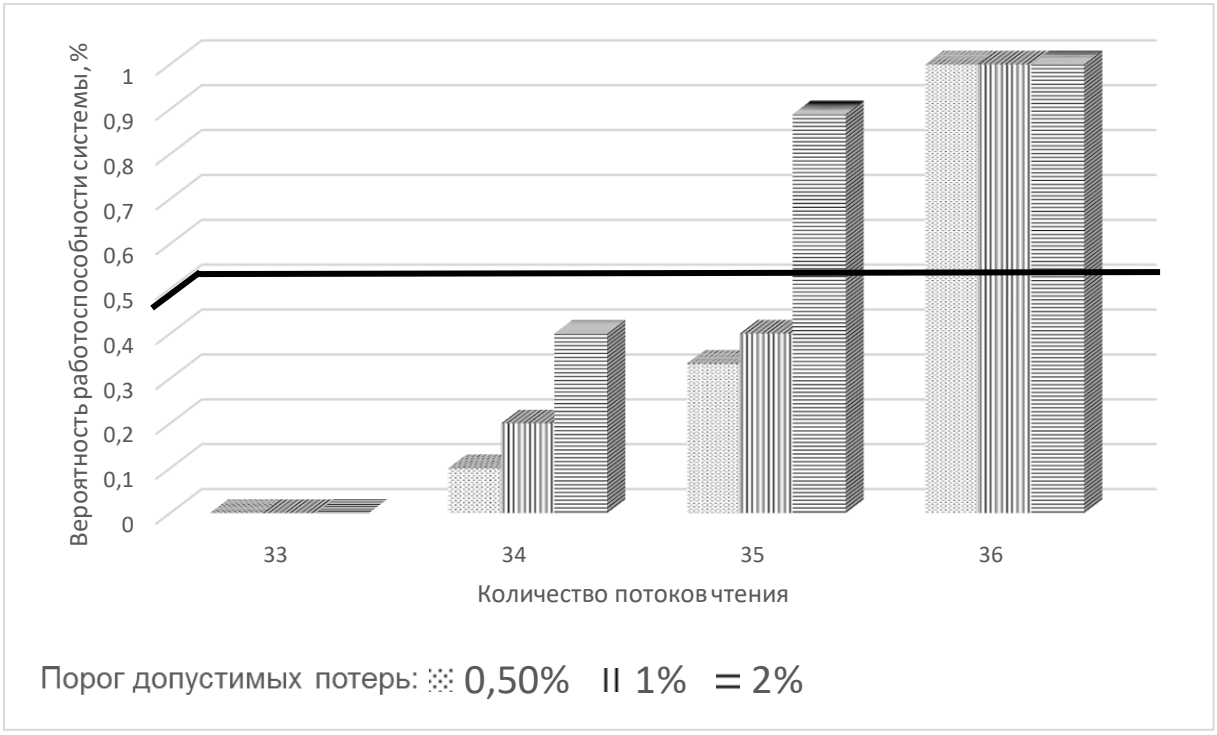

Рисунок 2. Вероятность работоспособности системы сбора с разным количеством потоков чтения данных при изменяющихся порогах возможных потерь

5. Заключение

Разработана, реализована и протестирована новая схема макро-моделирования, основанная на вероятностно-ценовом подходе, для оценки различных конфигураций оборудования. Полученные к настоящему времени результаты моделирования позволили вести с проектировщиками систем *DAQ* и триггеров содержательные и аргументированные дискуссии по поводу параметров потоков данных, способствующие принятию мотивированных решений. Приведенный пример показывает, как оптимизировать систему приема и хранения данных установки *BM@N* (сократить вероятность потерь) путем добавления 2 или 3 дополнительных потоков чтения шириной 150 МБ/сек, что обойдется в 200 или 300 условных ценовых единиц вложений соответственно. В настоящее время программа ориентирована только на анализ потерь при работе установок *B@MN* и *MPD*, но в дальнейшем она будет доработана путем создания классов, которые позволят гибко менять топологию системы хранения данных. Также планируется апробирование параллельного расчета различных вариантов конфигураций оборудования, что позволит анализировать значительное количество вариантов.

Список литературы

[1] NICA - Nuclotron-based Ion Collider fAcility. Available at: http://nica.jinr.ru/complex.php (accessed 23.09.2017)

[2] V. Korenkov, A. Nechaevskiy, G. Ososkov, D. Pryahina, V. Trofimov, A. Uzhinskiy and N. Voytishin. The JINR Tier1 Site Simulation for Research and Development Purposes // European Physical Journal (EPJ) Web of Conferences February 2016: Vol.108, article number - DOI: 10.1051/epjconf/201610802033

[3] V.V. Korenkov, A.V. Nechaevskiy, G.A. Ososkov, D.I. Pryahina, V.V. Trofomov, A.V. Uzhinskiy. Simulation concept of NICA-MPD-SPD Tier0-Tier1 computing facilities // Particles and Nuclei Letters. 2016. V. 13, no. 5, pp. 1074-1083

[4] В.В. Кореньков, А.В. Нечаевский, Г.А. Ососков, Д.И. Пряхина, В.В. Трофимов, Ю.К. Потребеников, А.В. Ужинский. Моделирование системы распределенной обработки данных эксперимента BM@N в составе комплекса T0-T1 NICA // CEUR Workshop Proceedings 2016: Vol.1787, с.307-311 - ISSN 1613-0073

[5] Rajkumar Buyya. GridSim: A Grid Simulation Toolkit for Resource Modelling and Application Scheduling for Parallel and Distributed Computing. Available at: http://www.buyya.com/gridsim/ (accessed 23.06.2015)

[6] Rodrigo N. Calheiros. CloudSim: A Framework for Modeling and Simulation of Cloud Computing Infrastructures and Services. Available at: http://www.cloudbus.org/cloudsim/ (accessed 23.06.2015)

[7] Хрусталев Л.Н. Численные методы решения задачи промерзания протаивания грунта // Известия Сибирского отделения АН СССР.: Серия технических наук. 1966. Вып. 2, № 6, с. 148-154

[8] Хрусталев Л.Н., Пустовойт Г.П. Вероятностно-статистические расчеты оснований зданий в криолитозоне // Новосибирск. Издательство «Наука». 1988. 253 с.

[9] Л.Н. Хрусталев, И.В. Давыдова. Прогноз потепления климата и его учет при оценке надежности оснований зданий на вечномерзлых грунтах // Криосфера Земли. 2007. Т. XI, №2, с. 68–75

[10] A. Dolbilov, Yu. Minaev, V. Mitsyn, A. Nechaevskiy, Yu. Potrebenikov, D. Pryahina, O. Rogachevsky, B. Shchinov, T. Strizh, V. Trofimov. Network and computing infrastructure for the NICA complex at JINR. Technical Design. Version 1.02. Dubna. May 15, 2018. Available at: http://mpd.jinr.ru/wp-content/uploads/2018/05/NICA_computing_TDR.pdf (accessed 23.06.2018)

PROBABILITY-COST APPROACH TO OPTIMIZING OF DISTRIBUTED DATA STORAGE SYSTEMS FOR PHYSICAL EXPERIMENTS

V.V. Trofimov, A.V. Nechaevskiy, G.A. Ososkov, D.I. Priakhina [a]

Joint Institute for Nuclear Research,

Russia, 141980, Moscow region, Dubna, Joliot-Curie, 6

E-mail: [a] pryahinad@jinr.ru

There is a problem of choosing the optimal configuration of the necessary computer and network equipment for the storage and processing system of BM@N and MPD experiments that the NICA complex part. A macro-modeling scheme is proposed and implemented to solve this problem. The process of data movements is considered as a byte stream that has a statistical nature, without analyzing individual parts of this stream. A probabilistic approach is used to assess the different equipment configurations: the losses probability of information coming from the detectors for each of these configurations are defined, and one configuration is selected for which this probability does not exceed a predetermined limit and the price is minimal. The structure of the simulation program, the proposed statistical criteria for the system quality and simulation algorithms are described. The results of preliminary calculations for the selection of the optimal data storage system for the BM@N experiment are presented.

Keywords: physical experiment, simulation, optimization of distributed data storage systems