ОБОЗРЕНИЕ
ПРИКЛАДНОЙ И ПРОМЫШЛЕННОЙ
Том 27 МАТЕМАТИКИ Выпуск 1
2020

М. Г. Коновалов, **Р. В. Разумчик** (Москва, ИПИ ФИЦ ИУ РАН). **Новый способ построения стратегий распределения нагрузки в почти ненаблюдаемых системах с параллельной обработкой информации.**

УДК 519.87+519.24

Резюме: Рассматривается проблема эффективного распределения единственного потока однородных заданий в почти ненаблюдаемых системах с параллельной обработкой информации. Предлагается новая методика для порождения легко масштабируемых стратегий диспетчеризации, которые при любом количестве серверов зависят от единственного неизвестного параметра. Оптимальные значения параметра могут быть оценены по имитируемой траектории системы.

Ключевые слова: Системы с параллельным обслуживанием, диспетчеризация, стратегии размещения заданий, управление при неполном наблюдении.

По мере наращивания производительности вычислительных систем за счет создания и внедрения более мощных технических средств все острее встает проблема эффективного использования ресурсов. Это связано с тем, что, несмотря на потенциально большие возможности по обработке информации, вследствие имеющих различную (в том числе и стохастическую) природу эффектов, системы оказываются либо слабо загруженными, либо, наоборот, перегруженными, а требования к качеству функционирования при этом не выполняются. Доклад посвящен изложению нового приема повышения производительности систем за счет оптимизации потоков нагрузки. В основе методики — недавно полученные авторами экспериментальные результаты [1]. Для примера предлагается постановка задачи, возникшая из определенного круга практических приложений, связанных с распределенными компьютерными системами передачи, обработки и хранения данных, в которых по тем или иным причинам динамическая информация о системе отсутствует.

Система состоит из конечного (возможно весьма большого) числа параллельно и независимо друг от друга работающих гетерогенных обслуживающих ресурсов (серверов), которые выполняют однородные задания, направляемые на них диспетчером, в режиме разделения процессора [2]. При этом диспетчер, осуществляя выбор сервера для выполнения очередного задания, не имеет возможности отложить решение и может руководствоваться только априорной информацией о системе и всей предысторией принятых решений. Динамическая информация о состоянии системы (например, о числе заданий в серверах и пр.), полностью отсутствует. Ставится задача нахождения стратегии распределения заданий между ресурсами, которая будет наилучшей с точки зрения стационарного среднего времени отклика.

Математическая формализация оптимизационной задачи не представляет принципиальной трудности. Однако ее аналитическое решение невозможно, за исключением частных случаев (например, при простейшем потоке заданий с экспоненциально распределенными размерами). Недоступность для выбора управлений текущих состояний системы очень сужает множество допустимых стратегий: применимыми в общем случае являются рандомизированная стратегия [3, Раздел 3] и программная стратегия [4]. В докладе описывается принципиально новый способ порождения стратегий

диспетчеризации, использующих всю доступную диспетчеру информацию. Несмотря на простоту реализации, они остаются слишком сложными для теоретического анализа. В основе методики — идея использования для порождения действий виртуальных вспомогательных процессов, зависящих от единственного неизвестного параметра и синхронизованных по моментам поступления заданий с основной системой. Ввиду того, что априорная информация дает возможность осуществлять имитацию траектории системы, значение неизвестного параметра может быть подобрано и оптимизировано. Численные эксперименты показывают, что новый подход позволяет создавать стратегии, сопоставимые со всеми известными в мировой литературе и часто превосходящими их по простоте реализации и критерию стационарного среднего отклика.

СПИСОК ЛИТЕРАТУРЫ

1. *Konovalov M.*, *Razumchik R.* A simple dispatching policy for minimizing mean response time in non-observable queues with SRPT policy operating in parallel. — Communications of the ECMS, 2020, v. 34, № 1, p. 398–402.
2. *Яшков С. Ф.* Анализ очередей в ЭВМ. М.: Радио и связь, 1989.
3. *Коновалов М. Г.*, *Разумчик Р. В.* Обзор моделей и алгоритмов размещения заданий в системах с параллельным обслуживанием. — Информ. и ее примен., 2015, т. 9, в. 4, с. 56–67.
4. *Hordijk A.*, *van der Laan D.* Periodic routing to parallel queues and billiard sequences. — Mathematical Methods of Operations Research, 2004. v. 59, № 2, p. 173–192.

УДК 519.87+519.24

Konovalov M. G., ***Razumchik R. V.*** (Moscow, Institute of Informatics Problems, FRC CSC RAS)**. New approach for the development of dispatching policies in non-observable systems with parallel service.**

Abstract: The problem of efficient dispatching (with respect to the long-run mean response time) of single flow of homogeneous jobs in non-observable systems with parallel service is revisited. New approach is proposed, which allows one to develop easy scalable dispatching policies. Irrespective of the system size new policies depend on a single unknown parameter, which optimal values can be estimated from a simulated trajectory of the system.

Keywords: Parallel service, dispatching, job allocation, decision making under non-observability.