

ПРОЯВЛЕНИЕ ЗАКОНА АМДАЛЯ-ГУСТАВСОНА НА ПРИМЕРЕ РЕАЛИЗАЦИИ АЛГОРИТМА К-СРЕДНИХ



А.И. Демидчук
Заведующий учебно-научной «Лаборатории высокопроизводительных вычислений», Республика Беларусь



Д.Ю. Перцев
Ассистент кафедры электронных вычислительных машин БГУИР, магистр технических наук, Республика Беларусь

Д.В. Кришталь
Студент БГУИР, Республика Беларусь



Д.И. Самаль
Заведующий кафедрой электронных вычислительных машин БГУИР, кандидат технических наук, доцент, Республика Беларусь



М.М. Татур
Профессор кафедры электронных вычислительных машин БГУИР, доктор технических наук, Республика Беларусь

Белорусский государственный университет информатики и радиоэлектроники, tatur@bsuir.by.

K-means algorithm used to verify Amdahl's law for cluster architecture based on GPU. Some practical experiments to analyze the performance and theoretical assumption verification are presented. Also we will show that there's a theoretical limit of cluster nodes after which the increase in nodes count doesn't lead to time reduction.

Известно, что производительность параллельной машины при решении прикладной задачи не может возрасти линейно с увеличением числа процессорных элементов. Так, общее время вычислений складывается из времени работы параллельного сопроцессора и времени «накладных расходов», связанных с загрузкой либо выгрузкой данных сопроцессора, а также реализации участков программ, не поддающихся распараллеливанию, подготовительных операций, сервисных функций и т. п. Влияние объема обрабатываемых данных и числа процессорных элементов на время вычислений характеризуют эффективность программно-аппаратной реализации в целом. Графики качественных зависимостей такого влияния для различных аппаратных платформ должны иметь форму, представленную на рисунке 1[1]. Получение численных зависимостей необходимо для проек-

тирования параллельных вычислительных систем на конкретных аппаратных платформах под заданные технические характеристики. С этой целью был проведен следующий эксперимент с использованием вычислительного кластера БГУИР.

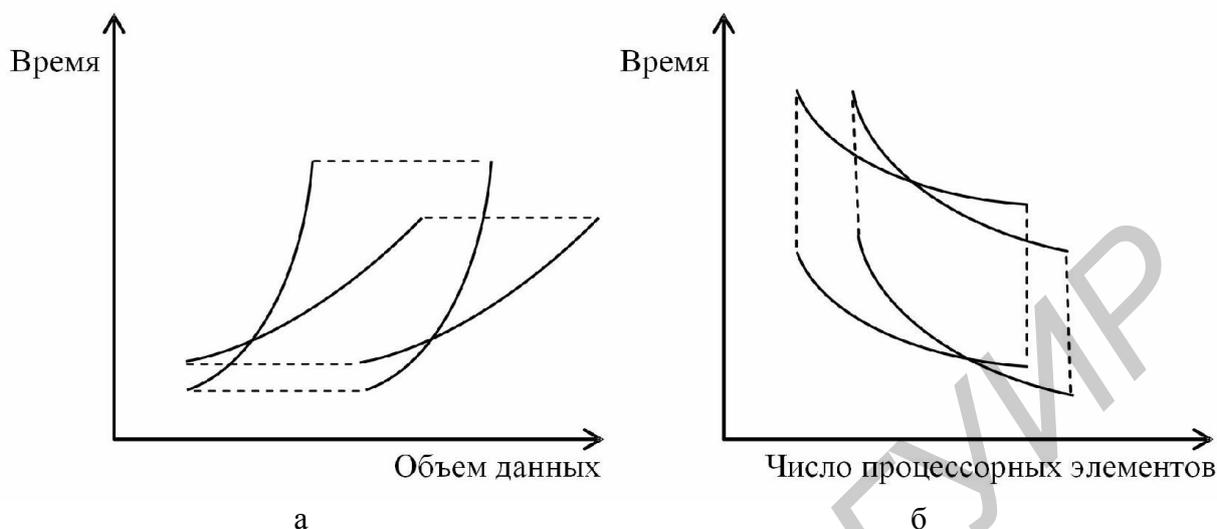


Рис. 1. Качественные зависимости времени вычислений тестовой задачи: а) от объема обрабатываемых данных; б) от числа процессорных элементов

В качестве аппаратной реализации использовался вычислительный кластер, состоящий из 7 узлов и включающий 2 альтернативные вычислительные конфигурации (в контексте главного вычислительного модуля):

- 2 CPU Intel© Xeon E5-2650 и 32 Gb RAM стандарта DDR3 на каждом узле;
- 2 GPU NVIDIA© Tesla M2075 (число ядер CUDA – 448, частота работы ядра – 1,15 ГГц, 6Gb памяти) на каждом узле.

Вычислительные узлы кластера соединены по сети InfiniBand с пропускной способностью 10Gb/s. Для обмена сообщениями между параллельными процессами использовалась реализация программного интерфейса OpenMPI.

В качестве тестовой задачи использовался алгоритм кластеризации k-средних [2].

Для проведения эксперимента был сгенерирован тестовый набор с 10^7 образов на 100 признаков (координат) и заданным числом кластеров. Из данного набора создавались подмножества с меньшим числом объектов, кратных 10^6 , для исследования зависимости времени от объема обрабатываемых данных.

Корректность реализации алгоритма k-средних проверялась с помощью программы Weka 3[3].

На рисунке 2 представлена полученная зависимость среднего времени обработки от числа узлов кластера, на рисунке 3 – зависимость от числа объектов (образов) для кластеризации (число вычислительных узлов – 7).

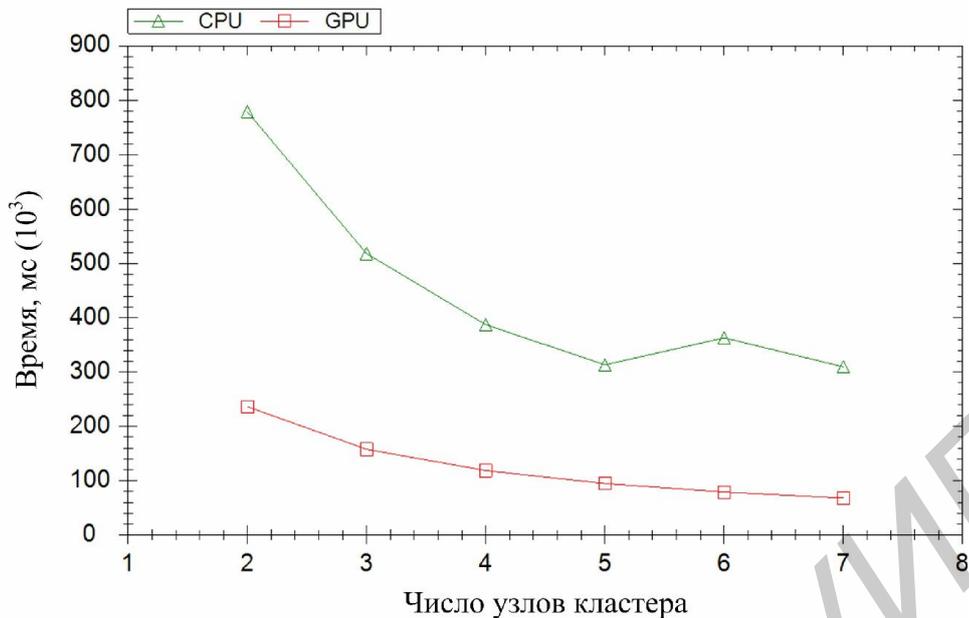


Рис. 2. Зависимость времени обработки от числа узлов кластера

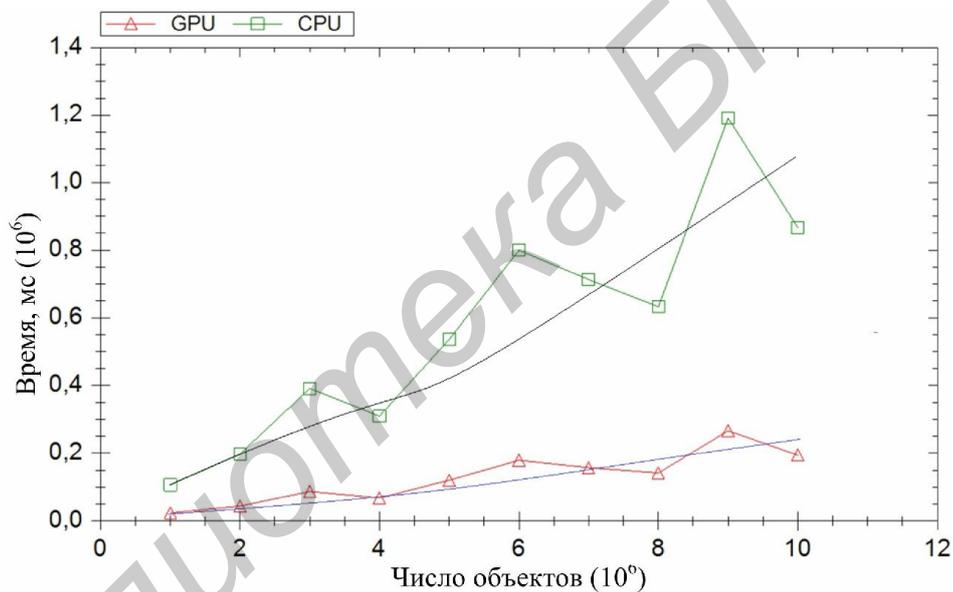


Рис. 3. Зависимость времени обработки от числа объектов (число узлов – 7)

Скорость выборки входных данных из хранилища (RAID массив жестких дисков) так же существенно влияет на общую производительность системы. На рисунке 4 представлена средняя скорость чтения в зависимости от числа узлов кластера обращающихся к хранилищу.

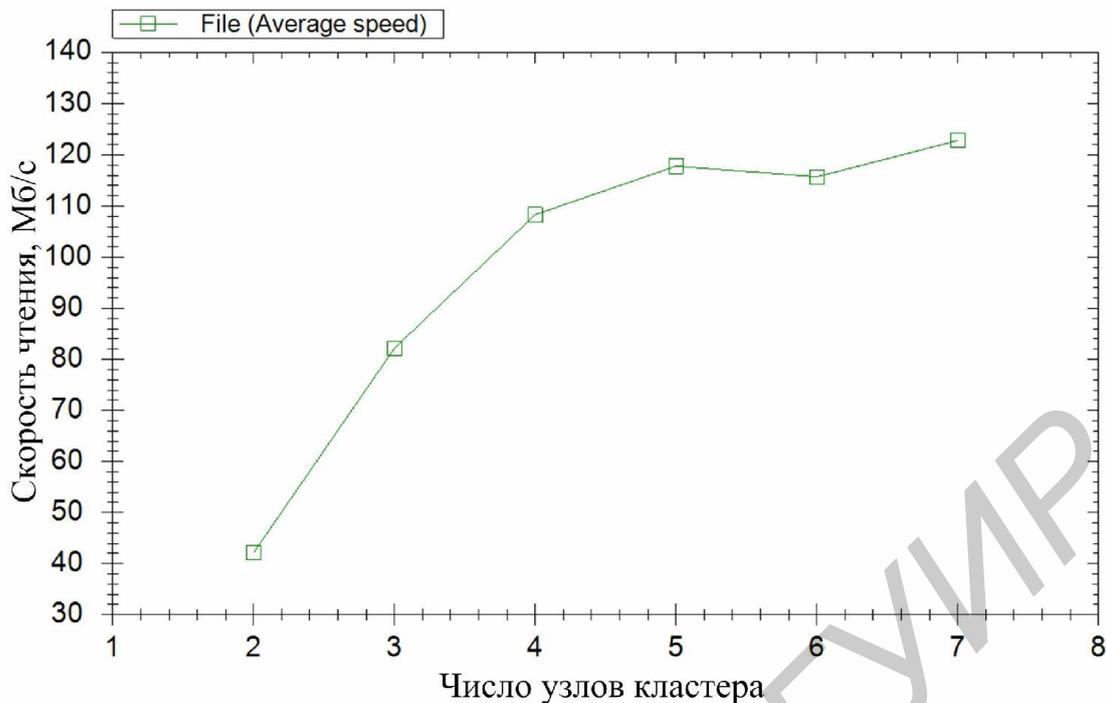


Рис. 4. Зависимость скорости чтения от числа узлов

Исходя из полученных графиков, можно сделать следующие выводы:

1. Увеличение числа узлов кластера приводит к уменьшению времени обработки;
2. Существует предел числа узлов кластера, после которого время перестает уменьшаться;
3. Время увеличивается экспоненциально с увеличением вычислительной нагрузки (в примере за счет увеличения числа объектов для кластеризации);
4. Увеличение скорости чтения ограничено пропускной способностью среды передачи данных и задержками доступа к жестким дискам.

Литература

1. Татур М.М. Особенности построения вычислителей интеллектуальной обработки данных. – М.: Информатика, 2015. – №1(45). – с.39-44.
- 2/ Мандель И.Д. Кластерный анализ. — М.: Финансы и статистика, 1988. – 176 с.
- 3/ Weka 3: Data Mining Software in Java [Electronic resource]. – Mode of access: <http://www.cs.waikato.ac.nz/ml/weka>. – Date of access: 12.05.2015.
- 4/ Intel® Xeon® Processor E5-2650 [Electronic resource]. – Mode of access: http://ark.intel.com/products/81705/Intel-Xeon-Processor-E5-2650-v3-25M-Cache-2_30-GHz. – Date of access: 12.05.2015.
- 5/ NVIDIA© Tesla M2075 [Electronic resource]. – Mode of access: http://www.nvidia.com/docs/IO/43395/BD-05837-001_v01.pdf. – Date of access: 12.05.2015.