Proceedings of the 2016 International Conference on Industrial Engineering and Operations Management
Kuala Lumpur, Malaysia, March 8-10, 2016

Cost Minimization in Data Batch Processing

Alia Al Sadawi
Abdulrahim Shamayleh
Malick Ndiaye
Industrial Engineering Department
American University of Sharjah
Sharjah, UAE
alia.h.sadawi@gmail.com
ashamayleh@aus.edu
mndiaye@aus.edu

***Abstract*—In today's world, data communication and execution- whether online or offline- became essential in all business fields. Considering the case of offline execution, data files are collected in groups called input batches and processed using available resources to get the required aggregated output in the form of "output batch", this process is called Data Batching. In most applications of batch processing, the target is to execute required tasks within specified time frame while fulfilling all predecessors' requirements and constraints set by the client using minimum resources. The data batch process terms and conditions are stated in the contract between service provider and the client - Known as Service Level Agreement (SLA). In this work we present an algorithm to minimize the total cost of data batch processing.**

Keywords— Data batch process, scheduling, cost minimization, parallel processors, process time.

I. INTRODUCTION

Batch processing has been associated with mainframe computers since the earliest days of electronic computing in the 1950s. It is used in many industries and in many application such as the banking industry billing function and issuing reports, etc.. Its main feature is the ability to handle huge amount of data files and that made it attractive to many organizations in multiple business fields[1] [2].

In data batch process jobs are scheduled in the form of input batches, assigned to available servers while satisfying all constraints and priorities then processed within specified time frame to obtain the output batch. Our objective was including all types of associated costs. The developed algorithm can be utilized to analyze and realize the optimal resources allocation to minimize the total operation cost.

A. *Data Batch Process:*

Data batch process is the execution of collections of programs ("jobs") on a computer that run at a scheduled time or on as needed basis without interaction with users and with no or minimal interaction with a computer operator [1] [2].

A major part of the workload on mainframe computers consists of batch processing. A large mainframe often will run several thousand batch jobs every day [1]. This "network" of jobs represents a business workflow with complex interrelations requiring careful scheduling and prioritizing to ensure that all batch jobs run in the correct order and meet strict deadlines [1]. In general, Jobs are gathered in a queue and will run when the user places a request. Job scheduler will schedule them as per pre-determined policy and the server will start the batch process when " MOM" –mother of executing jobs- gives the order [2]. "Fig. 1", illustrates a typical data batching system [2].

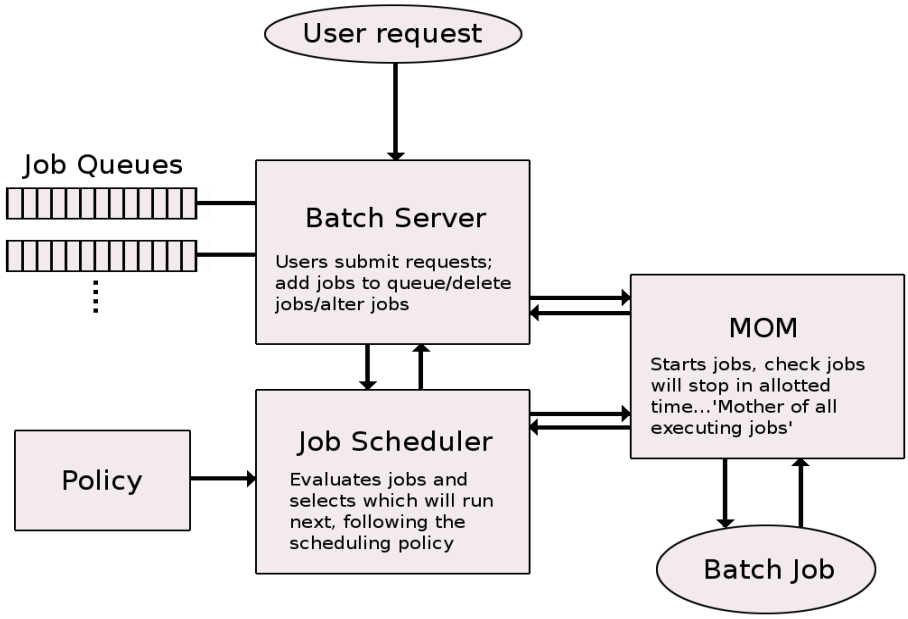

Fig. 1. Data batching process

B. *When and where to use data batching?*

Data batching is used when non-continuous (non-real time) processing of data, instructions, or materials is accepted. Batch processing is used in data transmission for very large files or where a fast response time is not critical.

Most mature mainframe systems rely on batch jobs to perform significant portions of the total application logic. The types of tasks undertaken include [1] [2]:

- Merging the day's transactions into master files.
- Sorting data files for optimal processing the following day.
- Providing daily, weekly, monthly and annual reports, bills , payroll and statements.

C. *Batch window:*

A batch window is "a period of less-intensive online activity", when the computer system is able to run batch jobs without interference from online systems [1].

II. PROBLEM STATEMENT

Batch processing plays a critical role in daily operations carried out in most organizations in different business fields. Data batching process handles data files through allocating them as batches to the available servers for execution and obtaining the output files also as a batch while satisfying data files priorities and predecessors' requirements set by the client within specified time frame. It is clear from the above demonstration that a major component of the data batch process is missing and that is cost. The objective of this work is to address the problem of minimizing the cost of resources (processors and software), needed to process a group of tasks according to a certain set of priorities while satisfying all the predecessors and constraints and maintaining the lowest possible cost.

III. Literature Review

Modern computer network and distribution systems are known for their massive complexity which makes traditional approaches to resource allocation sort of impractical and hard to optimize in addition to making it extremely costly. Several papers in the literature dealt with different aspects of resource allocation such as methods, effectiveness, optimization and process time. Only few considered cost as a way of optimizing the resource allocation issue whether it is utilized in online or batch systems.

A. Page, T. Keane and T. Naughton [3] considered a combination of algorithms and strategy to schedule tasks to processors dynamically in a heterogeneous distributed system. C. Mendez, J. Cerda, I. Grossmann [4] presented different state of the art optimization methods for short term batch scheduling. S. Lim and S. Cho. [5], illustrated a method of process scheduling which adapts to users' preferences and aims to arrange CPU time to multiple processes for providing users with more efficient

throughput using Fuzzy inference systems (FIS). F. Xhafa and A. Abraham [6] dealt with Grid technologies where heuristic and meta-heuristic methods were illustrated and found to be appropriate for Grid scheduling. In K. Aida, [7] researchers investigated the effect of job size characteristics on job scheduling performance in a parallel computer system. Multiple job scheduling algorithms performance (such as FCFS, LJF, SJF, First-Fit and Backfilling) were evaluated.

Papers addressing batch process optimization were applied also in chemical engineering field. B. Srinivasan, S. Palanki and D. Bonvin [8] studied achieving batch process optimization by reducing production cost, improving production quality, meeting safety requirements and environmental regulations. In the first paper, researchers assumed the existence of accurate model and they discussed the characterization of the optimal solution . B. Srinivasan, S. Palanki, E. visser and D. Bonvin [9] included uncertainty in their optimality study since accurate models of industrial processes are rarely available. Optimality was achieved by tracking the necessary conditions using measurement as a way to optimize uncertain batch process. In R. Zhou, L. Li, W. Xiao and H. Dong [10] a cost –optimal schedule scheme and water allocation network were combined and designed simultaneously which incorporated into a single mathematical programming model addressing the interaction between the two systems.

D. Ferguson, C. Nikolaou, J. Sairamesh and Y. Yemini [11] adopted the human economic model and implemented it on resource allocation in computer network system. K. Kuwabara, Y. Nishibe, T. Ishida and T. Suda [12] presented market-based approach, where resources are allocated to activities through buying and selling of resources between agents and resource allocation in a multi-agent system .Chun and Culler [13] presented performance analysis of market based batch schedulers for clusters of workstations using user-centric performance metrics. J. Sairamesh, D. Ferguson and Y. Yemini [14] Proposed a new methodology based on economic models to provide Quality of Service (QoS) guarantees to competing traffic classes in packet networks. C. Yeo and R. Buyya [15] outlined a taxonomy that describes how market-based resource management systems can support utility-driven cluster computing.

It is concluded from literature review that only few research studies take operations cost into consideration, papers were mostly dealing with scheduling methods or minimizing the total execution time and that is the gap this research is trying to fill.

IV. BANK DATA BATCHING MODEL

In the proposed work, a bank will be used as an example of application. The bank has a huge amount of data that need to be processed. Due to the quantity and diversity of data and lack of resources (i.e. processors), the bank outsourced the job. A private company take charge of arranging and processing the data, and aggregating the output. In other words, the proposed model assumes the presence of a third party who provide the service of processing the bank data. The third party company is leasing servers and software from the only source like IBM and performing the batch process for the bank under a service level agreement (SLA). The (SLA) governs the relationship between the client (i.e. bank) and the service provider (i.e. private company). The scheduling problem was tackled from the service provider point of view rather than from IBM and the customer side.

First step in the batching process is scheduling the batches taking into consideration job priorities, predecessors and other constraints. This means that jobs won't be processed simultaneously but rather as per their schedule. In our example,6 job files were assumed and scheduled as per the following illustrative "Fig. 2", [16]:

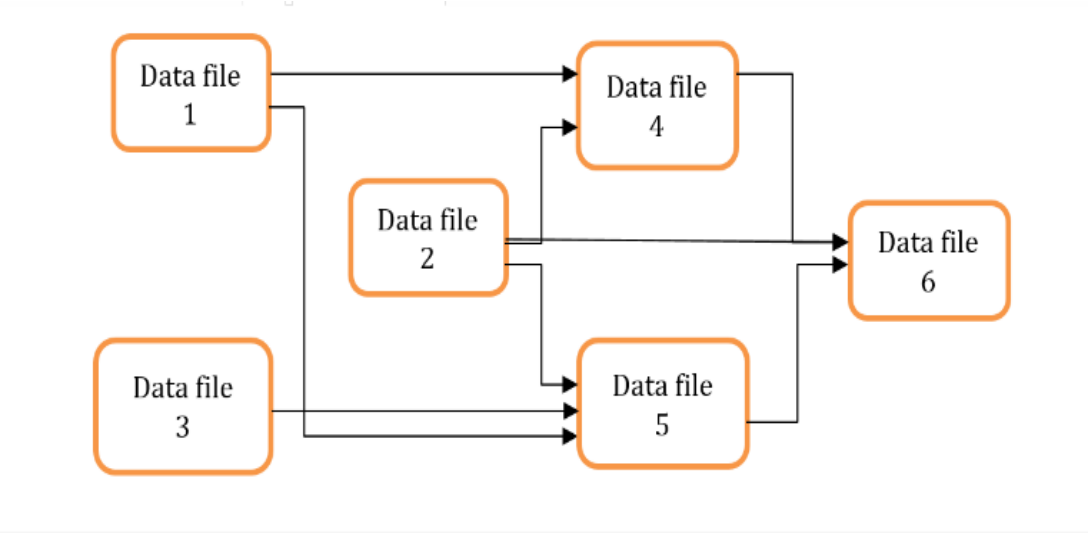

Fig. 2. Job scheduling example

Different costs involved in the process will be considered which are:

First: Basic leasing cost as per the agreement between servers host like IBM and the service provider.

Second: Software cost, which is the cost of renting the necessary software from IBM .

Third: rental cost, which is simply the extra cost endorsed by service provider due to renting additional servers from IBM. This happens in the case of job overload from the bank side. Obviously rental cost will be much higher than leasing cost since it was not negotiated in the first place .

Fourth: penalty cost. Clearly this is the fine that the service provider has to pay in case of failing to execute the data batch process as per the SLA.

Fifth: the opportunity cost representing the cost of idling a resource for any period of time due to inefficient task allocation.

Assuming multiprocessing method will be adopted meaning that several processors will be running simultaneously. It is assumed that the company has two servers, however server reservation is not allowed. Also, there are two types of instructions: CI (computing instruction) and I/O reading. It is worth mentioning that I/O reading operation is longer than CI operation [16].

No pre-assignment to processors is allowed. In addition, each file may need certain computing instruction (CI) processes and many input/output (I/O) reading. With multiple processing, any task can be divided among more than one processor for execution. It is worth mentioning here that a certain processor can execute one job at a time. [16].

A dynamic approach is presented and called: Batch Data Processes Scheduling Cost (BDPSC) algorithm which is an extension to BDPS developed by Osman et. al. [16], It is a dynamic iterative frame work used for assigning required tasks to available resources taking into consideration the predecessors, constrains and priority of each job. The dynamic framework is characterized by constant change in task allocation to available resources to achieve maximum utilization and prevent idling any resource during any period of time.

The following are the steps of BDPSC, starting from step two (BDPSC) enters a repetitive loop:

1) Preparatory stage: initial data is set up & BDPSC algorithm begins.

2) Increment & update iteration clock by 1 time unit and consider allocating available data files to several processors for CI processing.

3) Setting up data files subset, include data files not preceded in the data file subset and set data file weight based on precedence / dependency matrix.

4) Solve an integer network optimization model to determine the allocation of data files to different available processors.

Using the above mentioned assumptions, the objective function can be written as follows:

Min (Z_T) =

$$\sum_{i=1}^{I'}\sum_{k=1}^{K}\left(\frac{Csv}{\beta i\,\alpha\, i}\right)X_{ik} \;+\; \sum_{k=1}^{K}P_k\left(\frac{Csf+Ch}{\beta i\,\alpha\, i}\right)\left(1-\sum_{i=1}^{I'}X_{ik}\right) \;+\; \sum_{i=1}^{I'}\sum_{r=1}^{V}\left(\frac{Cesv+Ceh}{\beta i\,\alpha\, i}\right)Y_{ir} \qquad (1)$$

The objective function (1) is to minimize data file allocation cost while taking into consideration priority and weight of each file included in each subset at any time T. Notations are shown in the below table:

TABLE I. NOTATION

	Description
I	is the set of all data files at any time T
I'	is a subset of data file that are available for CI processing at any time T
K	is the set of all processors
R	Is the set of extra processors available in case of overload
V	Is the subset of extra processors needed in case of overload at any discrete time T
f_i^T	is equal to 1 if data file i is available for CI processing at discrete time T, and 0 otherwise
C_{sf}	is the software fixed basic leasing cost per unit time (= Basic cost /total unit times).
C_{sv}	is the software variable basic leasing cost per unit time (= Basic cost /total unit times).
C_h	is the hardware basic leasing cost per unit time (= Basic cost /total unit times)
C_{esv}	is the extra variable software rental cost per unit time (= Basic cost /total unit times)
C_{eh}	is the extra hardware rental cost per unit time (= Basic cost /total unit times)
q_i	is equal to the number of times data file i been CI processed; is incremented by 1 every time data file i being CI processed
n_i	is equal to the number IC processing tasks required for data file i
P_k	is equal 1 if processor/server k is available to receive data file, and 0 otherwise.
W_r	is equal 1 if processor/server R is available to receive data file, and 0 otherwise.
T	Clock discrete time
x_{ik}	is 1 if data file i allocated to processor k, and 0 otherwise
x_{ir}	is 1 if data file i allocated to extra processor R, and 0 otherwise
α_i	The data file weight based precedence/dependency matrix
β_i	The data file scheduling priority
e_i	The multiprocessing of data file i

Subject to:

$$\sum_{k=1}^{K} X_{ik} + \sum_{r=1}^{V} Y_{ir} \le M_i \; f_i^T \qquad \forall \; i \in I' \qquad \text{Where} \quad M_i = min\{ e_i, n_i - q_i^T, K+V \} \tag{2}$$

Constraint (2) ensures that the total file allocation for any file i at a certain time T doesn't exceed either the multiprocessing ei of that file , required number of processing nor the total number of available basic and extra processors .

$$\sum_{i=1}^{I'} X_{ik} \le P_k^T \qquad \forall \; k \in K \tag{3}$$

$$\sum_{i=1}^{I'} Y_{ir} \le W_r^T \qquad \forall \; r \in R \tag{4}$$

Constraint (3) ensures that exactly one data file is allocated to a single basic processor.

Constraint (4) ensures that exactly one data file is allocated to a single extra processor.

$$X_{ik} = 0 \quad or \quad 1 \qquad \forall \; i \in I \; and \; k \in K \tag{5}$$

$$Y_{ir} = 0 \quad or \quad 1 \qquad \forall \; i \in I \; and \; r \in R \tag{6}$$

Constraint (5) declares that the decision variable X_{ik} is binary, meaning that a file i either be assigned to basic processor or not.

Constraint (6) declares that the decision variable Y_{ir} is binary, meaning that a file i either be assigned to extra processor or not.

5) Update the availability of files and processors.

6) Termination condition is checked.

Steps 2-6 of the (BDPSC) will be repeated until all tasks are allocated to available resources.

V. RESULTS AND CONCLUSION

The following result was obtained using (BDPSC) algorithm, all files were allocated to servers within 3 time units and no extra servers were needed , also all basic servers were utilized except for the last time unit were one file was left for allocation. The below table demonstrates the result:

TABLE II. RESULT

T	**Files**					
	1	***2***	***3***	***4***	***5***	***6***
0	k:1	k:2				
1		k:1	k:2			
2				k:1	k:2	
3						k:1

In this paper, a dynamic scheduling algorithm for batch data processing (BDPSC) was proposed to solve the problem of allocating (scheduling) jobs in the form of input batches to available servers satisfying all constraints, predecessors and priorities within specified time frame while including all types of associated software and hardware costs. This work represents a model that can be used by companies to analyze and realize their optimal resource allocation which minimizes the total operation cost.

This research would contribute to the literature by introducing a cost approach of optimizing hardware and software resources management, look at the scheduling problem from the service provider point of view rather than from IBM and the customer side. We will be rrunning the final developed algorithm on real life examples to enhance the quality of schedules and minimize overall cost.

VI. ACKNOWLEDGMENT

The authors would like to acknowledge the support provided by The American University of Sharjah (AUS).

REFERENCES

[1] "Batch Applications—The Hidden Asset" Mainframe Migration Alliance, 2006.

[2] High Performance Computing Center North (HPC2N) " What is a batch system " [online] available at: https://www.hpc2n.umu.se/support/beginners_guide
[Accessed 1 March 2015].

[3] A. Page, T. Keane and T. Naughton. "Multi-heuristic dynamic task allocation using genetic algorithms in a heterogeneous distributed system." Journal of Parallel and Distributed Computing 70:758-766, 2010.

[4] C. Mendez, J. Cerda, I. Grossmann. "State-of-the-art review of optimization methods for short-term scheduling of batch processes." Computers and Chemical Engineering. 30:913-946, 2006.

[5] S. Lim and S. Cho. "Intelligent os process scheduling using fuzzy inference with user models." Lecture Notes in Computer Science 4570:725-734, 2007.

[6] F. Xhafa and A. Abraham. "Computational models and heuristic methods for Grid scheduling problems." Future Generation Computer Systems 26:608-621, 2010.

[7] K. Aida. "Effect of job size characteristics on Job scheduling performance." Lecture Notes in Computer Science Volume, 1911, pp 1- 17, 2000.

[8] B. Srinivasan, S. Palanki and D. Bonvin. "Dynamic optimization of batch processes, I. Characterization of the nominal solution", Computers and Chemical Engineering, 27 1 _/26, 2002.

[9] B. Srinivasan, S. Palanki, E. visser and D. Bonvin. "Dynamic optimization of batch processes, II. Role of measurements in handling uncertainty", Computers and Chemical Engineering, 27_/44, 2002.

[10] R. Zhou, L. Li, W. Xiao and H. Dong. "Simultaneous Optimization of batch process schedules and water – allocation network", Computers and Chemical Engineering, 33 (2009) 1153–1168, 2008.

[11] D. Ferguson, C. Nikolaou, J. Sairamesh and Y. Yemini. "Economic Models for Allocating Resources in Computer Systems." Market- based control: a paradigm for distributed resource allocation, P. 156-183, 1996.

[12] K. Kuwabara, Y. Nishibe, T. Ishida and T. Suda. "An quilibratory Market- Based Approach for Distributed Resource Allocation and Its Applications to Communication Network Control", Market-based control: a paradigm for distributed resource allocation, P. 53-73,1996.

[13] Chun and Culler. "User-centric Performance Analysis of Market-based Cluster Batch Schedulers." ,CCGRID '02 Proceedings of the 2nd IEEE/ACM International Symposium on Cluster Computing and the Grid , P. 30, 2002

[14] J. Sairamesh, D. Ferguson and Y. Yemini . "An Approach to Pricing, Optimal Allocation and Quality of Service Provisioning In High- Speed Packet Networks.", proceedings of the INFOCOM, 1995.

[15] C. Yeo and R. Buyya . "A taxonomy of market-based resource management systems for utility-driven cluster computing", Journal Software—Practice & Experience, Volume 36 Issue 13, P.1381-1419, 2006.

[16] M. Osman, M. Ndiaye and A. Shamayleh. "Dynamic scheduling for batch data processing in parallel systems", 2014.

BIOGRAPHY

Alyaa Abdulameer is currently a GTA and a master student in the engineering college-Engineering Systems Management program- at the American University of Sharjah (AUS). Eng. Abdulameer holds a Bachelor of Science degree in Electronics Engineering. She worked as Electronics and Sales engineer for Honeywell Middle East, Johnson Controls International and Nahas Intertard. Also occupied the position of Communication Division Director for Unified Technologies.

Abdulrahim Shamayleh is an Assistant professor in the Department of Industrial Engineering at American University of Sharjah.

Malick Ndiaye is an Associate Professor in the Department of Industrial Engineering at American University of Sharjah.