On the Throughput Optimization in Large-Scale Batch-Processing Systems

Sounak Kar
TU Darmstadt, Germany

Robin Rehrmann
TU Dresden, Germany

Arpan Mukhopadhyay
University of Warwick, UK

Bastian Alt
TU Darmstadt, Germany

Florin Ciucu
University of Warwick, UK

Heinz Koeppl
TU Darmstadt, Germany

Carsten Binnig
TU Darmstadt, Germany

Amr Rizk
Universität Ulm, Germany

ABSTRACT

We analyze a data-processing system with n clients producing jobs which are processed in *batches* by m parallel servers; the system throughput critically depends on the batch size and a corresponding sub-additive speedup function that arises due to overhead amortization. In practice, throughput optimization relies on numerical searches for the optimal batch size which is computationally cumbersome. In this paper, we model this system in terms of a closed queueing network assuming certain forms of service speedup; a standard Markovian analysis yields the optimal throughput in $\omega\left(n^4\right)$ time. Our main contribution is a mean-field model that has a unique, globally attractive stationary point, derivable in closed form. This point characterizes the asymptotic throughput as a function of the batch size that can be calculated in $O(1)$ time. Numerical settings from a large commercial system reveal that this asymptotic optimum is accurate in practical finite regimes.

1. INTRODUCTION

We consider a closed system where n clients generate jobs to be processed by m parallel servers. Each client alternates between being in an *active* or an *inactive* state; in the former it produces a job and in the latter it awaits the response. Note that each client can have at most one job in the system as it produces a new job no sooner than its previous one finished execution. The servers process jobs in batches of size k; once k clients produce k jobs, these are sent for batch processing and may have to wait in a central queue if all servers are busy. All times are assumed to be exponentially distributed[1]. In the active state, a client produces a *job* with rate λ, the batching step has a rate M, and a *batch* is served by one server with rate $\mu(k)$; see Fig. 1. This model is representative for some real-world data-processing systems such as databases employing Multi Query Optimization [7, 6, 5]. In addition to the single job type case, we consider a generalized model with multiple job types under priority constraints such as *read* and *write* jobs in a database with essentially different average processing times; see [3] for details.

[1]We will show that this technically convenient assumption is valid by fitting the parameters to real-world system traces.

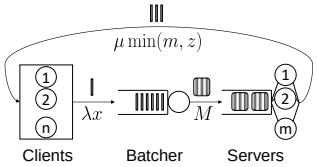


Figure 1: A closed queueing system with n clients and m servers. Clients are either active or inactive and produce jobs at rate λx when x of them are *active*. The batcher produces batches of size k at rate $M\lfloor y/k \rfloor$ when y jobs are available. The service station consists of a single queue and m parallel servers, each having a service rate μ; with z batches, the overall *batch* service rate is $\mu \min(m, z)$.

Classical literature of closed queueing systems has been focused on proving the product form property of the steady state distribution and characterization of such systems [1]. Closest to our setup from Fig. 1 is [2], where the existence of product form was investigated under service batching. However, this work does not readily apply to our problem as the conditional routing probabilities of jobs/batches in our case is state-dependent due the FCFS nature of the service. Even by approximating FCFS using random service order, we cannot *directly compute* the system throughput due to lack of a method to derive the corresponding normalizing constant.

The main contribution of this paper is the throughput optimization in a closed batching system which requires finding the optimal batch size. An exact analysis by solving the balance equations in a corresponding Markov model requires at least $\omega\left(n^4\right)$ computational time. Hence, we provide a mean-field model that yields an exact result in $O(1)$ time in the asymptotic regime where both n and m are proportionally scaled. Due to the deterministic nature of the limiting process and its globally attractive stationary point, this leads to a simple optimization problem that yields the asymptotically optimal batch size either in closed form or numerically.

We demonstrate the practical relevance of our results by analyzing a large commercial database system where a job refers to a query, e.g., an SQL string, which can execute *read* or *write* operations. Here, batching involves merging queries into a new SQL string, whose execution time depends on many factors such as the operation type. Moreover, the shared overhead amongst the individual queries leads to a certain speedup in the batch execution time [5] that is generally a function of the number of batched jobs k and the job types.

2. MAIN RESULT

In real-world data-processing systems, the number of clients served is usually large, making the Markovian approach computationally infeasible. Hence we adopt a mean-field approach where the number of servers m scales with the number of clients n. We assume that the batching step is instantaneous, i.e., the number of jobs in the batching station jumps from $(k-1)$ to 0 upon the arrival of a new job. Apart from tractability, this assumption is also motivated by the fact that the batching step is about 50 times faster than service in the database used for experiments.

Let $X^{(n)}(t)$ denote the number of active clients in the system at time $t \geq 0$, implying the number of queries in the system is $n - X^{(n)}(t)$. Then, $(X^{(n)}(t), t \geq 0)$ is a Markov process on $\{0, 1, \ldots, n\}$ and is ergodic by irreducibility and finiteness of the state space. However, it is difficult to obtain a closed form solution of the stationary distribution $\pi^{(n)}$ by solving $\pi^{(n)} Q^{(n)} = 0$ because of the non-linear state dependent rates [3]. An alternative and immediate approach is to obtain a bound on the system throughput as follows. Under the stationary distribution the following equality must hold:

$$\lambda \mathbb{E}\left[X^{(n)}\right] = k\mu(k)\mathbb{E}\left[\min\left(m, \left\lfloor \frac{n - X^{(n)}}{k} \right\rfloor\right)\right] , \quad (1)$$

$$\text{implying } \lambda \mathbb{E}\left[X^{(n)}\right] \leq k\mu(k) \min\left(m, \frac{n - \mathbb{E}\left[X^{(n)}\right]}{k}\right) ,$$

by Jensen's inequality. Now, the expected throughput of the system $\mathbb{E}\left[\Theta^{(n)}\right]$ for a given batch size k is given by the RHS (and hence the LHS) of (1), which gives

$$\mathbb{E}\left[\Theta^{(n)}\right] \leq \min\left(k\mu(k)m, \frac{n\lambda\mu(k)}{\lambda + \mu(k)}\right) . \quad (2)$$

Note that we dropped the dependency on k in $\Theta^{(n)}$ for brevity. Next we show that the bound in (2) is asymptotically tight as $n, m \to \infty$ with $m = \alpha n$ for some $\alpha > 0$. To this end, we consider the process $(w^{(n)}(t) := X^{(n)}(t)/n, t \geq 0)$ which denotes the fraction of active clients and we show that it converges to a deterministic limit with a unique fixed point. Note that $(w^{(n)}(t), t \geq 0)$ is a *density dependent jump Markov process* [4] with rates

$$q^{(n)}(w \to w - 1/n) = n\lambda w$$

$$q^{(n)}(w \to w + k/n) = n\mu(k) \min\left(\alpha, \frac{1}{n}\left\lfloor \frac{n - nw}{k} \right\rfloor\right) .$$

The next theorem contains our main result:

Theorem 1. *(i) If $w^{(n)}(0) \to w_0 \in [0,1]$ as $n \to \infty$ in probability, then we have*

$$\sup_{0 \leq t \leq T} \|w^{(n)}(t) - w(t)\| \to 0$$

in probability as $n \to \infty$, where $(w(t), t \geq 0)$ is the unique solution of the following ODE:

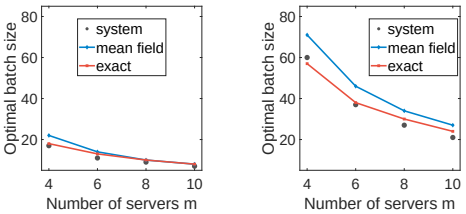
$$\dot{w}(t) = f(w(t)), \quad w(0) = w_0,$$

$$\text{with} \quad f(w) = k\mu(k) \min\left(\alpha, \frac{1-w}{k}\right) - \lambda w, \quad w \in [0,1] .$$

(ii) For any $w_0 \in [0,1]$, we have $w(t) \to w^*$ as $t \to \infty$, where w^* is the unique solution of $f(w^*) = 0$, i.e.,

$$w^* = \min\left(\frac{\mu(k)}{\lambda + \mu(k)}, \frac{\alpha k \mu(k)}{\lambda}\right)$$

(iii) The sequence of stationary measures $\pi_w^{(n)}$ of the process $(w^{(n)}(t), t \geq 0)$ converges weakly to δ_{w^*} as $n \to \infty$.



(a) $n = 100$ clients (b) $n = 300$ clients

Figure 2: Experimental evaluation: The observed optimal batch sizes k^* vs. the model estimates with increasing number of servers m. The system receives only *read* jobs and the comparison is for different number of clients n. *Exact* represents standard CTMC analysis. The optimal batch size decreases with higher number of servers due to server idling.

Details of the proof can be found in [3] where a similar result for the case with multiple job types has also been derived. The above theorem implies the weaker result that

$$\lim_{n\to\infty} \lim_{t\to\infty} \mathbb{E}\left[w^{(n)}(t)\right] = \lim_{t\to\infty} \lim_{n\to\infty} \mathbb{E}\left[w^{(n)}(t)\right] = w^*.$$

Equivalently, we have the convergence of the normalized throughput: $\Theta^{(n)}/n \to \lambda w^*$ as $n \to \infty$. The optimal asymptotic throughput follows by maximizing the fraction of active clients w^* over the batch size k as

$$k^* = \max_k \min\left(\frac{\mu(k)}{\lambda + \mu(k)}, \frac{\alpha k \mu(k)}{\lambda}\right) . \quad (3)$$

Assuming that $\mu(\cdot)$ takes a subadditive form, we estimate it by probing batch service times for a fraction of possible batch sizes. Thus k^* can be found by solving (3) with a calculation time that is independent of the system size n. Fig. 2 shows the validity of the models for *read* jobs and for different system sizes. Comprehensive evaluation results can be found in [3].

3. REFERENCES

[1] K. M. Chandy and A. J. Martin. A characterization of product-form queuing networks. *Journal of the ACM*, 30(2):286–299, 1983.

[2] W. Henderson and P. G. Taylor. Product form in networks of queues with batch arrivals and batch services. *Queueing Syst.*, 6(1):71–87, 1990.

[3] S. Kar, R. Rehrmann, A. Mukhopadhyay, B. Alt, F. Ciucu, H. Koeppl, C. Binnig, and A. Rizk. On the throughput optimization in large-scale batch-processing systems. *Performance Evaluation*, 2020.

[4] T. G. Kurtz. Solutions of ordinary differential equations as limits of pure jump markov processes. *Journal of Applied Probability*, 7(1):49–58, 1970.

[5] R. Rehrmann, C. Binnig, A. Böhm, K. Kim, W. Lehner, and A. Rizk. OLTPshare: The case for sharing in OLTP workloads. *Proc. VLDB Endow.*, 11(12):1769–1780, 2018.

[6] T. K. Sellis. Multiple-query optimization. *ACM Trans. Database Syst.*, 13(1):23–52, Mar. 1988.

[7] A. Thomson, T. Diamond, S.-C. Weng, K. Ren, P. Shao, and D. J. Abadi. Calvin: Fast distributed transactions for partitioned database systems. In *Proceedings of the ACM SIGMOD International Conference on Management of Data*, pages 1–12, 2012.