Asymptotic Analysis of Probabilistic Scheduling for Erasure-Coded Heterogeneous Systems

Rooji Jinan[1], Gaurav Gautam[1], Parimal Parag[1], and Vaneet Aggarwal[2]
[1]Indian Institute of Science, Bangalore 560012, India, email: {roojijinan,gauravgautam,parimal}@iisc.ac.in.
[2] Purdue University, West Lafayette IN 47906, USA, email: vaneet@purdue.edu.

ABSTRACT

We consider (k, k) fork-join scheduling on a large number (say, N) of parallel servers with two sets of heterogeneous rates. An incoming task is split into k sub-tasks and dispatched to k servers according to a probabilistic selection policy, with parameter p_s being the selection probability of slower servers. Mean task completion time admits an integral form, and thus it is analytically intractable to compute p_s that minimizes it. In this work, we provide an upper bound on the mean task completion time, and determine p_s that minimizes this upper bound. Numerically, this choice has been shown to be near-optimal.

1. INTRODUCTION

With increasing shift towards horizontal scaling of resources, distributed computing has become very popular. In distributed computing, a task is divided into smaller sub-tasks and distributed to multiple servers. The task completion time is limited by the slowest server. In practice, servers are heterogeneous, i.e., some servers are fast and some are slow. If these servers are treated equally, some servers will be congested and others will be under utilized. This leads to an increase in mean task completion time, and potential revenue loss for the service provider.

In this work, we consider two classes of servers that we call slow and fast servers. An incoming task is forked to k servers and, on completion of all k sub-tasks, the task leaves the system. We call assignment of these k sub-tasks to servers as scheduling. We are interested in finding a scheduling policy that minimizes mean completion time of incoming tasks. Given the heterogeneous nature of servers, it is difficult to identify the set of k servers for each incoming task that minimizes the mean completion time. For optimal performance, the server assignment depends on four parameters namely arrival rate, number of sub-tasks k, ratio of number of slow servers to number of fast servers, and ratio of server speeds.

We propose a probabilistic policy, where a sub-task is sent to a slow server with probability p_s and to a fast server with probability $1 - p_s$. We choose server uniformly at random without replacement within each class. In the proposed policy, we use selection probability p_s to distribute the sub-tasks among the servers. As a result, finding optimal policy is akin to a problem of finding the optimal selection probability that minimizes the mean task completion time.

Related Works: Common load balancing strategies designed to reduce the mean task completion time in a distributed computing system include the join shortest queue [14], the join smallest work [1, 7], the water filling policy [11], etc. The "power-of -d" variants [1, 10, 13] of these policies are also popular. Other efficient dispatching policies for parallel server systems include the size interval task assignment policy [6], Redundant-to-Idle queue [3], load balancing with timed replicas [9] etc. However, these policies are designed for a system of parallel homogeneous servers.

Comparison of various load balancing algorithms for heterogeneous systems can be found in [2]. More recent load balancing strategies for a system of heterogeneous parallel servers can be found in [4, 5, 12]. In [4], a "power-of-d" type load balancing policy for a system with heterogeneous servers with good response time and stability characteristics is studied. An algorithmic solution to the load balancing policy in a heterogeneous system posed as a stochastic optimization problem is given in [5]. A general "power-of-d" framework for heterogeneous servers is considered in [8]. An algorithm that yields a product form stationary distribution is studied in [12]. In all these works for heterogeneous servers, task is not sub-divided into multiple sub-tasks, which is the focus of our work.

Our contributions: We analytically compute the mean task completion time, under the proposed probabilistic policy, when the number of servers is arbitrarily large. This is achieved using the asymptotic independence of workload distribution. Finding the optimal probability selection parameter p_s that minimizes the mean response time is analytically intractable. Thus, we find a tight upper bound on the mean response time, and the probability selection parameter p_s that minimizes this bound. This probability serves as an approximation for the optimal selection probability, and we numerically verify that this approximation is tight.

2. SYSTEM MODEL

We consider a system of N heterogeneous servers with the set of slow and fast servers denoted by E_s and $E_f = [N] \setminus E_s$ respectively. We denote the number of slow and fast servers by $N_s \triangleq |E_s|$ and $N_f \triangleq N - N_s$ respectively. The fraction of slow servers is denoted by $f_s \triangleq \frac{N_s}{N}$. For this system, we assume a Poisson arrival of tasks with homogeneous rate $N\lambda$. Each arriving task is subdivided into k sub-tasks, and dispatched to k distinct servers selected out of N. We assume that the number of subtasks $k \leq \min(N_s, N_f)$. The task is assumed to be completed when all k sub-tasks are completed, and it leaves the system. The sub-task completion time at server i for task n is denoted by a random variable X_i^n. We assume that $(X_i^n : i \in [N], n \in \mathbb{N})$ is independent for servers $[N]$ and across tasks $n \in \mathbb{N}$. The sub-task completion time distribution at server i is denoted by G_{X_i}, and we assume that this distribution is identical for servers with same rate. The completion time distribution at slow and fast servers is denoted by G_s and G_f respectively. The service rates of slow and fast servers are denoted by μ_s and μ_f respectively, where $\mu_s < \mu_f$. That is, $\mathbb{E}X_i^n = \frac{1}{\mu_s}\mathbb{1}_{\{i \in E_s\}} + \frac{1}{\mu_f}\mathbb{1}_{\{i \in E_f\}}$.

We consider a probabilistic selection of k servers out of N. Servers are selected sequentially, and are chosen to be either slow or fast with probabilities $(p_s, \bar{p}_s)$ respectively. If the server is selected to be slow or fast, then it is chosen to be one of the slow or fast servers uniformly at random. For task n, let E^n be the k-set of probabilistically selected servers, then we denote the random set of selected slow and fast servers by $I_s^n \triangleq E^n \cap E_s$ and $I_f^n \triangleq E^n \cap E_f$ respectively, and denote the random number of slow and fast servers as $K_s^n \triangleq |I_s^n|$ and $K_f^n \triangleq k - K_s^n$ respectively. For task n, we can write the probability of selecting k_s slow servers as

$$q(k_s) \triangleq P\{K_s^n = k_s\} = \binom{k}{k_s} p_s^{k_s}(1-p_s)^{k-k_s}. \qquad (1)$$

Consequently, we can compute the probability that a slow server $i \in E_s$ is selected by the dispatcher for an incoming task, as $\sum_{k_s=1}^{k} q(k_s)\frac{\binom{N_s-1}{k_s-1}}{\binom{N_s}{k_s}} = \frac{1}{N_s}\sum_{k_s=1}^{k} k_s q(k_s) = \frac{kp_s}{N_s}$. This probability is independent of the incoming task, and hence the arrival at each slow server is a thinned Poisson process with arrival rate $\lambda_s \triangleq \lambda N \frac{kp_s}{N_s} = \frac{\lambda k p_s}{f_s}$. Analogously, we can compute the probability that a server $i \in E_f$ is selected by the dispatcher for an incoming task as $\frac{k\bar{p}_s}{N_f}$ independent of the task. Consequently, the arrival process at each fast server is thinned Poisson process with arrival rate $\lambda_f \triangleq (\lambda k \bar{p}_s)/\bar{f}_s$.

3. MEAN TASK COMPLETION TIME

We denote the marginal workload at server i seen by nth incoming task by W_i^n, and its limiting distribution by F_{W_i} such that $F_{W_i}(x) \triangleq \lim_{n\to\infty} P\{W_i^n \leqslant x\}$. If one of the k sub-tasks for the nth task is dispatched to a server $i \in E^n$, then the sub-task completion time at this server is denoted by $T_i^n \triangleq W_i^n + X_i^n$. Since the sub-task completion times are *i.i.d.* , W_i^n and X_i^n are independent and for any $x \in \mathbb{R}_+$, $F_{T_i^n}(x) \triangleq P\{W_i^n + X_i^n \leqslant x\} = \int_{\mathbb{R}_+} P\{W_i^n \leqslant x - y\} dG_{X_i}(y)$. Due to symmetry in the system, the marginal workload distribution is identical at all slow servers and at all fast servers. The limiting distribution for marginal workload at a slow and a fast server is denoted by F_s and F_f respectively. We denote the limiting distribution for sub-task completion time at any server i as $J_{T_i} : \mathbb{R}_+ \to [0,1]$, which can be written for any $x \in \mathbb{R}_+$, as $J_{T_i}(x) = \lim_{n\to\infty} P\{T_i^n \leqslant x\} = \int_{y\in\mathbb{R}_+} F_{W_i}(x-y)dG_{X_i}(y)$. It follows that limiting distribution of sub-task completion times are identical up to the parameters for slow and fast servers and we denote them by J_s and J_f respectively.

The completion time for task n is denoted by T^n, and is the maximum of the sub-task completion times at the selected E^n servers, and written as $T^n \triangleq \max_{i\in E^n} T_i^n$. The limiting distribution of task completion times is denoted by $H : \mathbb{R}_+ \to [0,1]$, and defined as $H(x) \triangleq \lim_{n\to\infty} P\{T^n \leqslant x\}$ for all $x \in \mathbb{R}_+$. In the following, we present our technical results. In the following results, it is assumed that the workloads at individual queues are independent of each other as the number of servers is asymptotically large. Proofs are omitted due to space constraints.

Theorem 1. *The limiting distribution of mean task completion time is given by* $H(x) = \sum_{k_s=0}^{k} q(k_s)J_s(x)^{k_s}J_f(x)^{k-k_s}$.

Corollary 1. *The mean task completion time for the heterogeneous system under consideration is given by*

$$\lim_{n\to\infty} \mathbb{E}[T^n] = \sum_{k_s=0}^{k} q(k_s)\left[\int_{w\in\mathbb{R}_+} [1 - J_s(w)^{k_s}J_f(w)^{k-k_s}\right] dw.$$

Memoryless sub-task completions: We note that when the service time is exponentially distributed, each queue observed in isolation is an $M/M/1$ queue. That is, when sub-task completion times at slow and fast servers are distributed exponentially with rates μ_s and μ_f respectively, and the respective loads are defined as $\rho_s \triangleq \frac{\lambda_s}{\mu_s}$ and $\rho_f \triangleq \frac{\lambda_f}{\mu_f}$, the limiting marginal workload distribution at slow and fast servers are $F_s(w) = 1 - \rho_s e^{-(\mu_s-\lambda_s)w}$ and $F_f(w) = 1 - \rho_f e^{-(\mu_f-\lambda_f)w}$. Furthermore, the limiting sub-task completion times for slow and fast servers are $J_s(x) = 1 - e^{-(\mu_s-\lambda_s)x}$, $J_f(x) = 1 - e^{-(\mu_f-\lambda_f)x}$. Thus, the resulting limiting mean of task completion time is $\int_{x\in\mathbb{R}_+} dx[1 - (1 - p_s e^{-(\mu_s-\lambda_s)x} - \bar{p}_s e^{-(\mu_f-\lambda_f)x})^k]$.

Theorem 2. *The optimal selection probability p_s for the slow servers that minimizes the limiting mean of task completion time for $k = 1$ is the solution of the following equation* $\mu_s\left(1 - \frac{\lambda p_s}{f_s \mu_s}\right)^2 = \mu_f\left(1 - \frac{\lambda \bar{p}_s}{\bar{f}_s \mu_f}\right)^2$.

Remark 1. *The optimal selection probability p_s is the positive root of a quadratic equation. We can verify that* $p_s \in [0,1]$.

Remark 2. *For exponentially distributed sub-task completion times, we can write the limiting mean of response time as*
$\sum_{x=1}^{k}(-1)^{x-1}\binom{k}{x}\sum_{i=0}^{x}\binom{x}{i}\frac{p_s^i(1-p_s)^{x-i}}{i(\mu_s-\lambda_s)+(x-i)(\mu_f-\lambda_f)}$.
Here, the analytical computation of the optimal selection probability p_s seems intractable for $k > 1$. However, the optimal selection probability can be numerically evaluated.

Upper bound:

Remark 3. *For exponentially distributed sub-task completion times, the limiting mean of sub-task completion time at server i is* $\mathbb{E}T_i^n = \frac{1}{\mu_s-\lambda_s}\mathbb{1}_{\{i\in E_s\}} + \frac{1}{\mu_f-\lambda_f}\mathbb{1}_{\{i\notin E_s\}}$.

Theorem 3. *The mean task completion time for the heterogeneous system under consideration for exponentially distributed sub-task completion times, is upper bounded as*

$$\lim_{n\to\infty} \mathbb{E}T^n \leqslant \frac{kp_s f_s}{\mu_s f_s - \lambda k p_s} + \frac{k\bar{p}_s\bar{f}_s}{\mu_s\bar{f}_s - \lambda k\bar{p}_s}.$$

Remark 4. *The upper bound on the limiting mean of task completion time for exponentially distributed sub-task completion times is minimized by the selection probability p_s for slow servers, that solves the following equation for a constant k.*

$$\mu_s\Big(1-\frac{k\lambda p_s}{f_s\mu_s}\Big)^2=\mu_f\Big(1-\frac{k\lambda\bar{p}_s}{\bar{f}_s\mu_f}\Big)^2. \qquad (2)$$

4. NUMERICAL RESULTS

We have carried out the analysis under the regime of asymptotically large number of servers N, which yields asymptotic independence of marginal workload distribution at individual servers. We first demonstrate that this assumption is robust, by empirically computing the limiting mean of task completion time for a system with finite number of servers $N \in \{20, 40\}$, for different values of selection probability $p_s \in [0, 1]$. This empirical curve and the theoretically obtained expression under asymptotic independence assumption is plotted in Figure 1. Observe that the asymptotic independence assumption remains robust even for finite N, and gets better as N increases. In addition, the mean task completion time is a convex function of selection probability p_s and thus has a unique minimum.

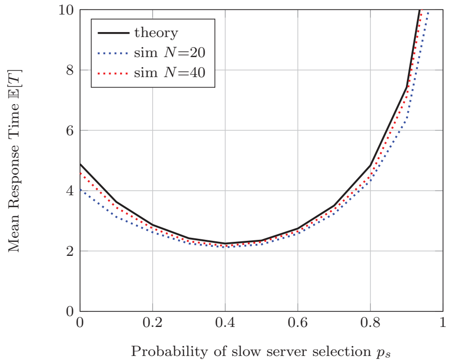

Figure 1: Comparison of mean completion time obtained theoretically and empirically as a function of selection probability p_s for the choice of parameters $k = 10, \lambda = 0.09, \mu_s = 2, \mu_f = 2.4, f_s = 0.5$.

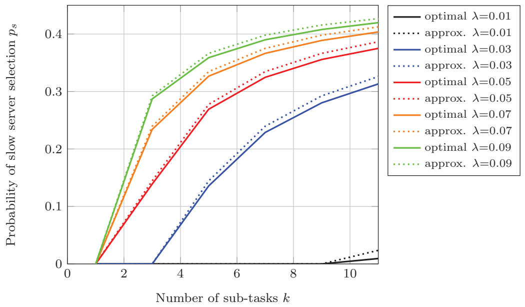

Figure 2: Comparison of optimal selection probability p_s and its approximation as a function of number of sub-tasks k for the choice of system parameters $N = 40, \mu_s = 2, \mu_f = 2.4, f_s = 0.5$.

Next, we plot the numerically evaluated optimal selection probability p_s as mentioned in Remark 2 as a function of number of tasks $k \in \{1, \dots, 10\}$ for a system with $N = 40$ servers in Figure 2, and for different arrival rates $\lambda \in \{0.01, \dots, 0.09\}$. We also plot the value of selection probability p_s as a function of k, that minimizes the upper bound on the limiting mean of task completion time, given in Remark 4. We observe that this sub-optimal selection probability is close to numerically evaluated optimal selection probability for different arrival rates and number of sub-tasks. We observe that the difference between two probabilities increases with increase in the number of sub-tasks. In addition, we note that the optimal selection probability of slow servers increases with arrival rate λ and number of sub-tasks k.

5. REFERENCES

[1] U. Ayesta, T. Bodas, and I. M. Verloop. On redundancy-d with cancel-on-start aka join-shortest-work (d). *ACM SIGMETRICS Performance Evaluation Review*, 46(2):24–26, Jan. 2019.

[2] S. Banawan and N. Zeidat. A comparative study of load sharing in heterogeneous multicomputer systems. In *Proceedings. 25th Annual Simulation Symposium*, pages 22–31, 1992.

[3] K. Gardner, M. Harchol-Balter, A. Scheller-Wolf, M. Velednitsky, and S. Zbarsky. Redundancy-d: The power of d choices for redundancy. *Operations Research*, Aug. 2017.

[4] K. Gardner, J. A. Jaleel, A. Wickeham, and S. Doroudi. Scalable load balancing in the presence of heterogeneous servers. *Performance Evaluation*, 145:102151, Jan. 2021.

[5] G. Goren, S. Vargaftik, and Y. Moses. Stochastic coordination in heterogeneous load balancing systems. *arXiv preprint arXiv:2105.09389*, May 2021.

[6] M. Harchol-Balter. *Performance Modeling and Design of Computer Systems: Queueing Theory in Action*. Cambridge University Press, New York, NY, USA, 2013.

[7] T. Hellemans and B. V. Houdt. On the power-of-d-choices with least loaded server selection. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 2(2):27, June 2018.

[8] J. A. Jaleel, S. Doroudi, K. Gardner, and A. Wickeham. A general "power-of-d" dispatching framework for heterogeneous systems, 2021.

[9] R. Jinan, A. Badita, T. Bodas, and P. Parag. Load balancing policies with server-side cancellation of replicas. *arXiv preprint arXiv:2010.13575*, 2020.

[10] M. Mitzenmacher. The power of two choices in randomized load balancing. 12(10):1094–1104, Oct. 2001.

[11] S. Shneer and A. Stolyar. Large-scale parallel server system with multi-component jobs. *arXiv preprint arXiv:2006.11256*, 2020.

[12] M. van der Boor and C. Comte. Load balancing in heterogeneous server clusters: Insights from a product-form queueing model. In *2021 IEEE/ACM 29th International Symposium on Quality of Service (IWQOS)*, pages 1–10. IEEE, June 2021.

[13] N. D. Vvedenskaya, R. L. Dobrushin, and F. I. Karpelevich. Queueing system with selection of the shortest of two queues: An asymptotic approach. *Problemy Peredachi Informatsii*, 32(1):20–34, 1996.

[14] W. Winston. Optimality of the shortest line discipline. *Journal of Applied Probability*, pages 181–189, Mar. 1977.