Dynamic Job Replication for Balancing Fault Tolerance, Latency, and Economic Efficiency: Work in Progress

Vladimir Marbukh
National Institute of Standards and Technology
100 Bureau Drive, Stop 8920, Gaithersburg,
MD 20899-8920, USA
marbukh@nist.gov

ABSTRACT
Recent research has demonstrated benefits of replication of requests with canceling, which initiates multiple concurrent replicas of a request and uses the first successful result immediately removing the remaining replicas of the completed request from the system. This paper suggests that benefits of replication may come at the risk of abrupt system transition to an undesirable highly congested equilibrium. To expose, evaluate, and ultimately manage these risk/benefit trade-offs, we generalize replication strategy by: (a) accounting for possible inefficiency of "remote" service, (b) allowing replication only when static routing fails to identify idle "local" server, and (c) requiring one or more replicas of the same request to be completed to improve fault tolerance using majority rule decision. Due to intractability of the Markov performance model, our analysis is based on mean-field and fluid approximations. Future research should evaluate accuracy of assertions based on these approximations, and ultimately develop practical solutions for optimization of various performance trade-offs in distributed systems with replication.

Categories and Subject Descriptors
C.4 [**Performance of Systems**]: – *modeling techniques, performance attributes, reliability, availability, and serviceability.*

General Terms
Algorithms, Management, Performance, Design, Theory.

Keywords
Dynamic job replication, fault tolerance, latency, economic efficiency, risk/benefit trade-offs.

1. INTRODUCTION
Recent research [1]-[4] has demonstrated potential performance and reliability benefits of request replication with canceling, which initiates multiple concurrent replicas of a request and uses the first successful result immediately removing the remaining replicas of the completed request from the system. These benefits are due to avoiding or reducing idle time for servers, and resiliency to inherent local demand/capacity imbalances resulted from exogenous demand variability and limited reliability of the system components. However, analysis [1]-[4] has not taken into account replication overhead, e.g., due to "remote" service being less efficient than "local" service in distributed systems, which may create risks of performance loss as level of replication increases. Evaluation and optimization of these risk/benefit tradeoffs in large-scale distributed systems is a challenging problem due to intractability of the corresponding Markov performance model.

This short paper reports on work in progress on these tradeoffs exposure, evaluation, and management. To this end we generalize replication strategies [1]-[4] by (a) accounting for "inefficient" remote service, (b) allowing replication only when static routing fails to identify idle server, and (c) requiring one or more replicas of the same request to be completed to improve fault tolerance using majority rule decision. Due to intractability of the Markov performance model, we propose mean-field and fluid approximations for performance of the dynamic replication. Our analysis under mean-field and fluid approximations reveals that benefits of dynamic replication may come at the risk of abrupt system transition to an undesirable highly congested equilibrium. Future research should verify this claim and develop practically viable techniques for the tradeoff optimization in large-scale, distributed systems.

This paper is organized as follows. Section 2 introduces the dynamic job replication strategy. Section 3 proposes mean-field and fluid approximate performance models for this strategy. These approximations require solving system of non-linear fixed-point equations of dimension equal to the number of service groups. Section 4 specifies these approximations to a case of homogeneous system, when the fixed-point systems simplify to single fixed-point equation. Section 5 discusses some performance implications of dynamic replication. Finally, section 6 concludes and outlines directions of future research.

2. DYNAMIC JOB REPLICATION
Consider a system with I classes of jobs and J classes of servers, where class $j=1,..,J$ includes N_j servers. Jobs of class $i=1,..,I$ arrive following a Poisson process of rate Λ_i, and have exponentially distributed service time with average $1/\mu_{ij}$ on a class j server. Static load balancing strategy $S(q_{ij})$, determined by matrix $Q=(q_{ij})$, routes portion q_{ij} of arriving class i requests

to server group j, where $q_{ij} \geq 0$, $\sum_{j=1}^{J} q_{ij} \leq 1$, and $q_{i0} := 1 - \sum_{j=1}^{J} q_{ij}$ is rejection probability. Strategy $S(q_{ij})$ guarantees low latency at least in a case of large service groups $N_j >> 1$ under condition that accepted load can be sustained on average:

$$(1/N_j)\sum_{i=1}^{I} \Lambda_i q_{ij} / \mu_{ij} < 1, \; j = 1,..,J, \qquad (1)$$

However, due to variability of the exogenous demand or limited reliability of system elements, condition (1) may be occasionally violated, which results in unacceptably long delays. Dynamic load reallocation is typically considered as a viable alternative to overprovisioning which may be too expensive for commercial systems. However, efficient implementation of dynamic load reallocation requires real-time dynamic information on the server availability and service rates μ_{ij} affected by breakdowns or slowdowns. In this short paper, we assume tight latency requirements, which motivates us to consider the following two-stage job routing strategy.

At the first stage, an arriving job of class $i = 1,..,I$ is rejected with probability q_{i0}, and occupies an available server from group $j = 1,..,J$ selected with probability q_{ij}. A situation when all group j servers being already occupied indicates a demand/capacity mismatch, which is mitigated with job replication strategy $D(d_i, \mathbf{J}_i)$. This strategy immediately replicates an accepted job of class $i = 1,..,I$, which failed to find available server at the first stage, and one replica immediately occupies an available server in each server group $j \in \mathbf{J}_i \subseteq \{1,..,J\}$. Service time for a replica of class i job on a group j server is distributed exponentially with average $1/\mu_{ij}$. As soon as $d_i \geq 1$ replicas are serviced, the original job is considered serviced and all replicas leave the system.

Note that since $q_{ij} > 0$ indicates efficient service for a class i job on server group j, $q_{ij} > 0$ may imply $j \in \mathbf{J}_i$, despite we do not require this. Parameter d_i determines fault tolerance for class i jobs since $d_i \geq 3$ completed replicas can be used for majority rule decision. Note that strategy $S(q_{ij}) \& D(d_i, \mathbf{J}_i)$ interpolates between pure static strategy $S(q_{ij})$ if $\mathbf{J}_i = \varnothing$ and pure replication strategy $D(d_i, \mathbf{J}_i)$.

Assuming that a serviced job of class i brings revenue $w_i = w_i(d_i)$, the system performance is characterized by the revenue rate $W = \sum_i w_i \Lambda_i$, where w_i is an increasing function of fault tolerance d_i: $w_i = w_i(d_i)$. Due to intractability of Markov performance model of a queuing systems with replication, we model queuing system with tight latency requirements by the corresponding loss system with tight loss requirements for class i jobs: $\pi_i \leq \varepsilon$ for some $0 < \varepsilon << 1$. This loss system differs from the original queuing system only in that jobs are lost instead of being queued. Since loss rates π_i depend on the fault tolerances $d = (d_1,..,d_I)$ and replication sets $\mathbf{J} = (\mathbf{J}_1,..,\mathbf{J}_I)$: $\pi = \pi(d, \mathbf{J})$, the system optimization problem becomes as follows:

$$\max_{d,\mathbf{J}} \sum_i w_i(d_i)\Lambda_i \qquad (2)$$

subject to

$$\pi_i(d, \mathbf{J}) \leq \varepsilon \qquad (3)$$

3. PERFORMANCE MODELS

Let $\delta_j = 1$ if all servers of class j are occupied, and $\delta_j = 0$ otherwise. It is known [5] that for static resource allocation without replication, due to reversibility of the underlying Markov process, steady-state random variables δ_j are jointly statistically independent for $j = 1,..,J$:

$$P(\delta_1,..,\delta_J) = \prod_{j=1}^{J} [p_j^{\delta_j} (1 - p_j)^{1-\delta_j}], \qquad (4)$$

where $p_j = P(\delta_j = 1)$. In a general case of job replication, (4) does not hold exactly, and we propose a mean-field approximation [6], which assumes that (4) holds approximately. This approximation can be used when $I, J, |\mathbf{J}_i| >> 1$, and thus the aggregate demand on each server group is approximately Poisson.

Under this approximation, rejection probabilities $\pi_i \approx \tilde{\pi}_i$, where

$$\tilde{\pi}_i = \sum_j q_{ij} \tilde{p}_j \sum_{\mathbf{J}_i^0 : |\mathbf{J}_i^0| \leq d_i - 1} \left[\left(\prod_{j \in \mathbf{J}_i \setminus \mathbf{J}_i^0} \tilde{p}_j \right) \left(\prod_{j \in \mathbf{J}_i^0} (1 - \tilde{p}_j) \right) \right], \qquad (5)$$

approximate overflow probabilities are:

$$\tilde{p}_j = \frac{1}{\tilde{Z}_j} \sum_{n_{1j} + .. + n_{Ij} = N_j} \prod_{i=1}^{I} \frac{1}{n_{ij}!} \left(q_{ij} \frac{\Lambda_i}{N_j \mu_{ij}} + \tilde{\rho}_{ij} \right)^{n_{ij}}, \qquad (6)$$

and $\widetilde{Z}_j$ is the normalization constant. In (6), "effective" utilization

$$\widetilde{\rho}_{ij} = \begin{cases} \dfrac{\Lambda_i}{N_j \widetilde{\mu}_{ij}} \sum\limits_{k \in \mathbf{J}_i \setminus j} q_{ik} \widetilde{p}_{ik} & if \quad j \in \mathbf{J}_i \\ 0 & otherwise \end{cases}, \qquad (7)$$

"effective" service rate for a replica of a class i job on a class $j \in \mathbf{J}_i$ server is a sum of the service rates on this server and service rates for other replicas of this job:

$$\widetilde{\mu}_{ij} = \sum_{\delta_k} \left(\sum_{k \in \mathbf{J}_i \setminus j} (1-\delta_k) \mu_{ik} \right) \prod_{k \in \mathbf{J}_i \setminus j} \widetilde{p}_k^{\delta_k} (1-\widetilde{p}_k)^{1-\delta_k}, \quad (8)$$

and summing is over $\delta_k = 0,1 \colon \sum_k \delta_k \le |\mathbf{J}_i| - d_i + 1$.

Substituting (7)-(8) into right-hand side of (6), we obtain a closed system of J non-linear algebraic fixed-point equations for approximate overflow probabilities $\widetilde{p}_j$, $j = 1,..,J$. After solving this system, one can approximate rejection probabilities by (5) which enter constraints (3).

Fluid performance model describes case of large service groups: $N_j \to \infty$, $j = 1,..,J$

$$\widetilde{p}_j = \left[1 - 1 \Big/ \sum_i \left(q_{ij} \frac{\Lambda_i}{N_j \mu_{ij}} + \widetilde{\rho}_{ij} \right) \right]^+, \qquad (9)$$

where $[x]^+ := \max(0, x)$. Substituting (7)-(8) into right-hand side of (9) we obtain a closed system of fixed point equations describing system under fluid approximation.

4. HOMOGENEOUS SYSTEM

In distributed systems, each job class often has a "native/local" service group: $I = J$, which are more "efficient" then "non-native/remote" services: $\mu_{ij} < \mu_{ii}$, $i, j = 1,..,I$; $i \ne j$. In such systems, natural static routing strategy attempts to access local service: $q_{ii} = 1$, and it is natural to determine level of replication for class i jobs by the maximum allowed loss in service efficiency $\chi_i \colon \mathbf{J}_i = \mathbf{J}_i(\chi_i) = \{ j : \mu_{ij} \le (1+\chi_i)\mu_{ii}, j = 1,..,I \}$.

Consider a homogeneous system with native services, where $\Lambda_i = \Lambda$, $\mu_{ii} = \mu$, $\chi_i = \chi$, $N_i = N$, $d_i = d$, $\dim \mathbf{J}_i(\chi) = K(\chi)$ are the same for all job classes $i = 1,..,I$, and mean-field system (6)-(8) has homogeneous solution $\widetilde{p}_j = \widetilde{p}$, $i = 1,..,I$, which can be found from the corresponding single fixed-point equation. One may think of such a system as comprised of server groups located on a regular grid, where efficiency of non-native/remote service decreases with increase in the "distance" from the native/local service group.

It can be shown that the corresponding homogeneous mean-field equation has the following form:

$$\widetilde{p} = \frac{1}{N!} \left(\frac{\Lambda}{\mu_{eff}} \right)^N \Bigg/ \sum_{k=0}^{N} \frac{1}{k!} \left(\frac{\Lambda}{\mu_{eff}} \right)^k, \qquad (10)$$

where effective service rate is an increasing function of overflow probability $\widetilde{p}$ and replication inefficiency χ:

$$\mu_{eff} = \mu \varphi(\widetilde{p}, \chi), \qquad (11)$$

and form of function φ in (11) can be derived from (6)-(8).

Solution to fixed-point system (10)-(11) is shown in Figure 1 for small service groups N and low inefficiency of non-native services χ, and in Figure 2 for large N, χ.

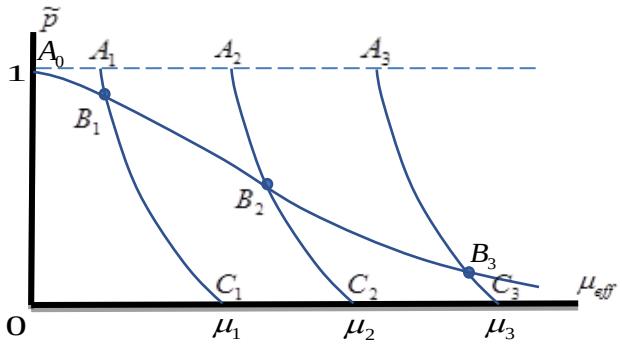

Figure 1. Solution to (10)-(11): small N, χ.

Curve $A_0B_1B_2B_3$ in Figure 1 and curve $A_0B_1B_2^1B_2^2B_2^3B_3$ in Figure 2 describe function (10). Curves A_kC_k in Figures 1 and 2 describe function (11) for $\mu = \mu_k$, where $\mu_1 < \mu_2 < \mu_3$.

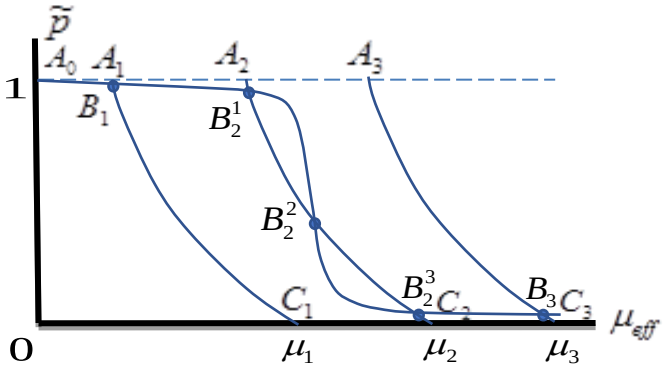

Figure 2. Solution to (10)-(11): large N, χ.

5. PERFORMANCE IMPLICATIONS

Figure 3 plots service group overflow probability $\tilde{p}$, given by (10)-(11), as a function of the exogenous load $\rho = \Lambda/(N\mu)$.

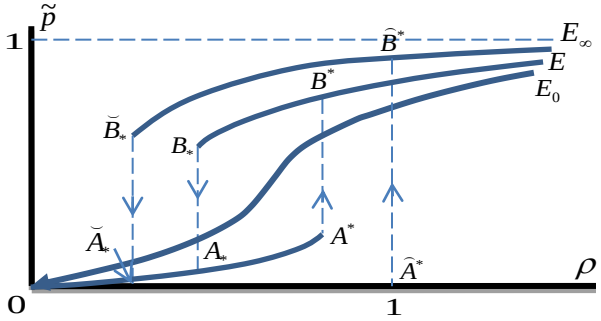

Figure 3. Server group overflow vs. exogenous load.

Curve $0E_0$ corresponds to situation shown in Figure 1, when system (10)-(11) has unique solution for any $\rho > 0$. Curves with hysteresis loops correspond to situation shown in Figure 2, when system (10)-(11) has two stable equilibria separated by unstable equilibrium. The stable equilibrium with lower overflow probability $\tilde{p}$ describes desirable operational system regime, and another stable equilibrium describes undesirable system regime. These two regimes coexist as metastable. Economics drives system to the stability boundary of the operational region creating risk of abrupt/discontinuous transition to undesirable regime.

Figure 4 plots phase diagram of system (10)-(11) in parameters (χ, ρ), where χ characterizes level of replication directly related to the allowed level of service inefficiency, and ρ characterizes exogenous load.

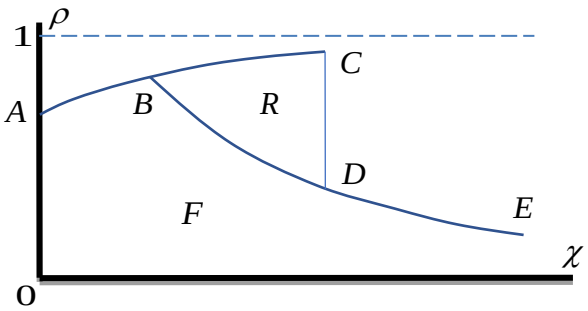

Figure 4. Phase diagram of system (10)-(11).

In region F system (10)-(11) has unique stable equilibrium, corresponding to the operational regime. In region R this equilibrium is locally stable and coexists with another locally stable equilibrium describing undesirable overloaded regime. Selection of the design parameter χ intended to maximize the sustainable exogenous utilization ρ yields point C in Figure 4. However, this solution on the boundary of the operational region $F \bigcup R$, is unstable with respect to unavoidable disturbances due to exogenous demand variability or limited reliability of the system components. These disturbances will result in abrupt/discontinuous system transition to undesirable operating point D. Risk of such a transition can be reduced by moving operating point along curve CB towards point B, where this risk completely diappears, at the cost of certain loss in sustained exogenous load. Note that in terminology [7], curve CB represents a "risk-aware" optimal selection of level of replication characterized by allowed level of service inefficiency χ.

6. CONCLUSION

This paper argues that benefits of request replication with cancelling [1]-[4] are inherently associated with certain risks. To evaluate the corresponding risk/benefit tradeoffs, we introduced a generalized replication strategy which accounts for relative inefficiency of "remote" service and has controlled level of replication and fault tolerance. Due to intractability of the Markov performance model, our analysis under mean-field and fluid approximations indicates that request replication is beneficial or harmful depending if the system is "light" or "heavy" respectively.

Future research should evaluate accuracy of the proposed approximations and the corresponding performance assertions through simulations. Observations made in this paper suggest that system performance optimization involves dynamic request replication, which reduces level of request replication with increase in the systemic congestion. Currently we are investigating such adaptive strategies.

7. REFERENCES

[1] G. Ananthanarayanan, A. Ghodsi, S. Shenker, and I. Stoica. Why let resources idle? aggressive cloning of jobs with Dolly. Memory, 2012.

[2] K. Gardner, M. Harchol-Balter, and A. Scheller-Wolf, A better model for job redundancy: Decoupling server slowdown and job size," in Proceedings of IEEE MASCOTS, London, UK, 2016.

[3] Z. Qiu, J. F. Pérez, and P. G. Harrison, Tackling latency via replication in distributed systems, 7th ACM/SPEC, Delft, Netherlands, 2016.

[4] G. Joshi, Boosting the throughput of a multiserver system via adaptive task replication, ACM Sigmetrics, MAMA, Illinois, USA, 2017.

[5] F.P. Kelly, Reversibility and Stochastic Networks, Cambridge University Press, 2011.

[6] N. Antunes, C. Fricker, P. Robert, and D. Tibi. Stochastic networks with multiple stable points. ArXiv:math.PR/0601296, 2006.

[7] V. Marbukh, "Fragility risks of low latency dynamic queuing in large-scale clouds: complex system perspective," IFIP Networking, 2017.