Thus, the probability that any backend is busy at any given point of time be denoted as:

\[ P_{\text{busy}} = \frac{\lambda}{n \cdot \mu}. \]  (2)

Similarly, let the probability that any backend is idle at any given point of time be denoted as:

\[ P_{\text{idle}} = 1 - P_{\text{busy}} = 1 - \frac{\lambda}{n \cdot \mu}. \]  (3)

Let \( P_r \) denote the probability that a request finds an idle backend after \( r \) retries, i.e., \( P_r \) denotes the probability that a request is assigned to busy backends during the first \( r \) attempts, and only in the \( r + 1^{st} \) retry, the request is assigned to an idle backend. Thus,

\[ P_r = (P_{\text{busy}})^r \cdot P_{\text{idle}}. \]  (4)

Let \( W_r \) denote the waiting time of the request that finds an idle backend after \( r \) retries. \( W_r \) consists of \( r \) round-trip latencies for the first \( r \) attempts, \( r \) delays of time \( \Delta \) each enforced by the frontend, and a single frontend-to-backend communication latency for the last successful attempt. Thus,

\[ W_r = r \cdot (d_1 + d_2) + r \cdot \Delta + d_1 = r \cdot (d_1 + d_2 + \Delta) + d_1. \]  (5)

\section*{APPENDIX}

\textbf{Assumption 1.} The average time to send a message from any frontend to any backend is \( d_1 \), and the average time to send a message from any backend to any frontend is \( d_2 \). Requests arrive at an average rate of \( \lambda \) and that the average time to service a request is \( 1/\mu \).

\textbf{Definition 1. RAND LB Policy:} Upon receipt of a request from the broker, the frontend immediately forwards it to a random backend. If the backend is busy, it sends back a busy response to the frontend. Upon receipt of a busy response, the frontend forwards the request to another random backend after \( \Delta \) time units.

\textbf{Theorem 1 (Waiting Time Distribution).} If there are \( n \) backends, and requests are assigned to backends as per the RAND LB Policy, then the expected \( p^{th} \) percentile waiting time of the requests can be approximated as:

\[ \omega_p = d_1 + \left( \frac{\ln (1 - \frac{p}{100})}{\ln \left( \frac{\lambda}{n \cdot \mu} \right)} - 1 \right) \cdot (d_1 + d_2 + \Delta). \]  (1)

\textbf{Proof.} Assume that \( d_1, d_2, \Delta \ll 1/\mu \). As a result, the duration for which the system remains non-work-conserving is negligible. Thus, the probability that any backend is busy at any given point of time is given by the overall system utilization:

\[ P_{\text{busy}} = \frac{\lambda}{n \cdot \mu}. \]  (2)

For \( P_{\text{busy}} \), we have:

\[ P_{\text{idle}} = 1 - P_{\text{busy}} = 1 - \frac{\lambda}{n \cdot \mu}. \]  (3)

Let \( P_r \) denote the probability that a request finds an idle backend after \( r \) retries, i.e., \( P_r \) denotes the probability that a request is assigned to busy backends during the first \( r \) attempts, and only in the \( r + 1^{st} \) retry, the request is assigned to an idle backend. Thus,

\[ P_r = (P_{\text{busy}})^r \cdot P_{\text{idle}}. \]  (4)

Let \( W_r \) denote the waiting time of the request that finds an idle backend after \( r \) retries. \( W_r \) consists of \( r \) round-trip latencies for the first \( r \) attempts, \( r \) delays of time \( \Delta \) each enforced by the frontend, and a single frontend-to-backend communication latency for the last successful attempt. Thus,
Hence, Eq. 1 is proved.

**Theorem 2 (Response Time Distribution).** If there are \( n \) backends, if the requests are assigned to backends as per the **RAND LB Policy**, and the CDF of the request execution times is denoted by \( \text{CDF}_{\text{exe}}(x) \), then the probability that the response time of a request is less than or equal to \( RT \) is given by

\[
\sum_{r=0}^{r_{\text{max}}} \left( \frac{\lambda}{n \cdot \mu} \right)^r \cdot \left( 1 - \frac{\lambda}{n \cdot \mu} \right) \cdot \text{CDF}_{\text{exe}}(RT - WT_r),
\]

where \( r_{\text{max}} = \left\lfloor \frac{(RT - d_1)}{(d_1 + d_2 + \Delta)} \right\rfloor \) and \( WT_r = r \cdot (d_1 + d_2 + \Delta) + d_1 \).

**Proof.** Let \( Y \) and \( Z \) denote the random variables corresponding to request execution times and waiting times, respectively. Let \( R = Y + Z \) denote the random variable corresponding to the request response time. Thus, by the general formula for the distribution of the sum of two independent discrete variables:

\[
P(R \leq RT) = \sum_{z=\infty}^{\infty} P(Z = z) \cdot P(Y \leq RT - z)
\]

(since \( Z \) can only take discrete values based on the number of retries, i.e., for \( r \) retries, \( Z = WT_r = r \cdot (d_1 + d_2 + \Delta) + d_1 \), and since \( r \) varies from 0 to \( \infty \))

\[
P(R \leq RT) = \sum_{r=0}^{\infty} P(Z = WT_r) \cdot P(Y \leq RT - WT_r)
\]

(since \( R, Y, Z \) are non-negative, \( R \leq RT \) implies that \( Z = WT_r \leq RT \), which in turn implies that \( r \leq r_{\text{max}} = \left\lfloor \frac{(RT - d_1)}{(d_1 + d_2 + \Delta)} \right\rfloor \))

\[
P(R \leq RT) = \sum_{r=0}^{r_{\text{max}}} P(Z = WT_r) \cdot P(Y \leq RT - WT_r)
\]

(since \( P(Z = WT_r) \) is equivalent to \( P_r = (P_{\text{busy}})^r \cdot P_{\text{idle}} \) in Eq. 4, and since \( P_{\text{busy}} = \left( \frac{\lambda}{n \cdot \mu} \right) \) and \( P_{\text{idle}} = 1 - \left( \frac{\lambda}{n \cdot \mu} \right) \) for the **RAND LB Policy**)

\[
P(R \leq RT) = \sum_{r=0}^{r_{\text{max}}} \left( \frac{\lambda}{n \cdot \mu} \right)^r \cdot \left( 1 - \frac{\lambda}{n \cdot \mu} \right) \cdot \text{CDF}_{\text{exe}}(RT - WT_r)
\]

(since \( P(Y \leq RT - WT_r) \) is equivalent to the CDF of the execution time distribution)

\[
P(R \leq RT) = \sum_{r=0}^{r_{\text{max}}} \left( \frac{\lambda}{n \cdot \mu} \right)^r \cdot \left( 1 - \frac{\lambda}{n \cdot \mu} \right) \cdot \text{CDF}_{\text{exe}}(RT - WT_r).
\]

Hence, Eq. 7 is proved. \( \square \)