Queueing Modeling

The arrival of transactions to the system can be regarded as a Poisson process with parameter λ. The transactions are divided into two parts, intra-shard transactions and cross-shard transactions. For a balanced sharding scheme, the probability of a transaction involving n accounts being an intra-shard transaction can be calculated as

$$\mathrm{P}_{\text{intra}} = \frac{1}{k^{n-1}}.$$

Thus, the arrival rate of intra-shard transactions λ_{int} and cross-shard transactions λ_{cr} can be determined by

$$\begin{aligned} \lambda_{\text{int}} &= \lambda\, \mathrm{P}_{\text{intra}} = \frac{\lambda}{k^{n-1}}; \\ \lambda_{\text{cr}} &= \lambda\,(1 - \mathrm{P}_{\text{intra}}) = \lambda \left(1 - \frac{1}{k^{n-1}}\right). \end{aligned}$$

Suppose that on average, the number of the input accounts is x. Then the rate of generating the first-phase subtransactions is

$$\lambda_{\text{sub}_1} = x\lambda_{\text{cr}} = x\lambda \left(1 - \frac{1}{k^{n-1}}\right).$$

Now consider an arbitrary shard $i \in \{1, 2, \cdots, k\}$. The mining process of shard i is modelled as an $M/M/1$ queue. Due to the homogeneity assumption of those shards, one can denote the service rate of each shard as μ and the shards share the same arrival rate, denoted by λ_0.

When the system attains its statistical equilibrium, as shown in [1], pp. 42, the departure process of a $M/M/1$ queue is the same Poisson as the arrival process, from which we can deduct that the rate of generating the second-phase subtransactions λ_{sub_2} satisfies that

$$\lambda_{\text{sub}_2} = (n-x)\lambda_{\text{sub}_1} = (n-x)x\lambda \left(1 - \frac{1}{k^{n-1}}\right).$$

Notice that, λ_{sub_2} also serves as an input sequence. So for shard i, along with the homogeneous assumption, we have

$$\begin{aligned} \lambda_0(k) &= \frac{\lambda_{\text{int}} + \lambda_{\text{cr}} + \lambda_{\text{sub}_2}}{k} \\ &= \frac{1}{k^n} + x(1+n-x)\left(\frac{1}{k} - \frac{1}{k^n}\right). \end{aligned}$$

Consider the growth of λ_0, we have

$$\begin{aligned} &\lambda_0(k+1) - \lambda_0(k) \\ &= \frac{-B_{n,x}(k+1)^{n-1}k^{n-1} + (B_{n,x}-1)\left[(k+1)^n - k^n\right]}{(k+1)^n k^n} < 0 \end{aligned}$$

as $k \to \infty$, where $B_{n,x} = x(1+n-x) > 0$. This indicates that when k is sufficiently large, λ_0 will decrease as k gets bigger.

Now we can calculate the expectation of the waiting time W_q as [1], pp. 37,

$$W_q = \frac{1}{\mu - \lambda_0} - \frac{1}{\mu}$$

If we consider the variation of W_q with λ_0,

$$\frac{\mathrm{d}W_q}{\mathrm{d}\lambda_0} = \frac{1}{(\mu - \lambda_0)^2} > 0$$

So minimizing W_q is the same to minimize λ_0. And since we have known that when k is sufficiently large, λ_0 decrease as k tends to infinity, one can conclude that if we want to shorten the waiting time W_q, we only need to increase the number of shards, i.e. integer k.

References

[1] U. N. Bhat, *An Introduction to Queueing Theory*, ser. Statistics for Industry and Technology. Boston: Birkhäuser Boston, 2008.