15-829 **Lec 10: More Queueing Tools: Redundancy & Fork-Join**, Chpt 3, 11, 13

Goal for Today

Several of you are doing research related to one of these models:

1. **Redundancy Model** ("min" model)
   - Same request sent to multiple servers.
   - Use first result to complete.

2. **Limited Fork-Join Model** ("max" model)
   - Break request into many pieces, each sent to different server.
   - Request done when all pieces are done.

Before we can do this, we need to cover the math on MINs and MAXs.

Understanding MINs: General Distribution

Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim X$. Assume we know X.

Want to understand:

$$W = \min(X_1, \ldots, X_n)$$

Our Goal: Express $f_W(t)$ in terms of $f_X(t)$

Question: What is $\overline{F}_W(t) = \mathbf{P}\left\{W > t\right\}$?

$$\begin{aligned}
\mathbf{P}\left\{W > t\right\} &= \mathbf{P}\left\{\min(X_1, \ldots, X_n) > t\right\} \\
&= \mathbf{P}\left\{X_1 > t \ \& \ X_2 > t \ \& \ \ldots \ \& \ X_n > t\right\} \\
&= \mathbf{P}\left\{X_1 > t\right\} \cdot \mathbf{P}\left\{X_2 > t\right\} \cdots \mathbf{P}\left\{X_n > t\right\} \\
&= \left(\mathbf{P}\left\{X > t\right\}\right)^n
\end{aligned}$$

Question: What is $F_W(t)$?

$$F_W(t) = 1 - \left(\mathbf{P}\left\{X > t\right\}\right)^n$$

Question: What is $f_W(t)$?

$$\begin{aligned}
F_W(t) &= \int_0^\infty f_W(t) dt \\
f_W(t) &= \frac{d}{dt} F_W(t) = \frac{d}{dt} \int_0^\infty f_W(t) dt \\
&= -n \cdot \left(\mathbf{P}\left\{X > t\right\}\right)^{n-1} \cdot \left(-f_X(t)\right) \\
&= n f_X(t) \left(\mathbf{P}\left\{X > t\right\}\right)^{n-1}
\end{aligned}$$

Point is that we know everything about X, so we can get density of W.

Understanding MINs: Exponential Distribution

Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim X$. Assume we know $X \sim \text{Exp}(\mu)$.

Want to understand: $W = \min(X_1, \ldots, X_n)$

Question: What is $\overline{F}_W(t)$? What is $f_W(t)$?

$$\begin{aligned} \overline{F}_W(t) &= \left(\mathbf{P}\left\{X > t\right\}\right)^n \\ &= \left(e^{-\mu t}\right)^n \\ &= e^{-n\mu t} \end{aligned}$$

$$\begin{aligned} F_W(t) &= 1 - e^{-n\mu t} \\ f_W(t) &= -1 \cdot e^{-n\mu t} \cdot (-n\mu) \\ &= (n\mu)e^{-n\mu t} \end{aligned}$$

Question: What does this tell us about the distribution of W?

$$W \sim \text{Exp}(n\mu) .$$

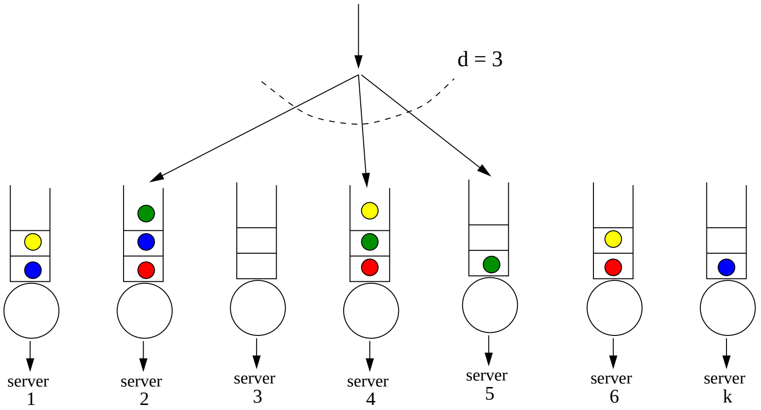
Question: What is $\mathbf{E}\left[X\right]$? What is $\mathbf{E}\left[W\right]$?

$$\begin{aligned} \mathbf{E}\left[X\right] &= \frac{1}{\mu} \\ \mathbf{E}\left[W\right] &= \frac{1}{n\mu} \end{aligned}$$

Question: How would we get $\mathbf{E}\left[W\right]$ if we didn't have $X \sim \text{Exp}(\mu)$?

$$\mathbf{E}\left[W\right] = \int f_W(t) \cdot t dt .$$

Redundancy-d Model



- Each job creates d copies of itself.
- The d copies go to different random servers.
- Job is complete as soon as any ONE copy is done. Remaining copies are cancelled.

Question: Let T_i represent the response time at queue i.
What is T, the overall response time?

$$T = \min(T_1, T_2, \ldots, T_d) \ .$$

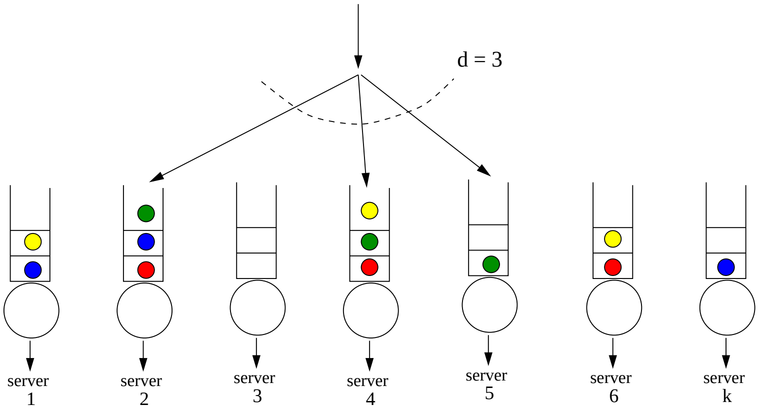
Question: How do we compute this?

Problem is that the T_i's are *NOT* independent. If one queue is very long, then it got most of the jobs, so other queues got fewer jobs. Thus the math gets very complicated (see the exact result in papers below). However, if k is high, and d is low, it's close enough to pretend that the T_i's are independent. So

$$\overline{F}_T(t) = \mathbf{P}\left\{T > t\right\} = \left(\mathbf{P}\left\{T_i > t\right\}\right)^d \ .$$

Relevant papers:
– Gardner et al. "Redundancy-d: The Power of d Choices for Redundancy" Operations Research, vol. 65, no. 4, 2017.
– Gardner et al. "Reducing Latency via Redundant Requests: Exact Analysis." SIGMETRICS 2015.
– Gardner et al. "A Better Model for Job Redundancy: Decoupling Server Slowdown and Job Size." IEEE MASCOTS 2016.

Redundancy-d Model



- Each job creates d copies of itself.
- The d copies go to different random servers.
- Job is complete as soon as any one copy is done.
  Remaining copies are cancelled.

$$T = \min(T_1, T_2, \ldots, T_d) \ .$$

Question: What is the distribution of T_i?

Generally, you need to obtain this through measurement. However in the case where the job size is Exponentially distributed and the arrival process into queue i is a Poisson process, then it turns out [Perf Modeling & Design, Chpt 13] that T_i is Exponentially-distributed.

Question: If T_i is Exponentially-distributed with mean $\mathbf{E}[T_i]$, what does this say about $\mathbf{E}[T]$?

T is Exponentially distributed with mean $\frac{\mathbf{E}[T_i]}{d}$.

NOTE: For the unique case of Exponential job sizes, where copies have *independent* sizes, redundancy doesn't add load to the system, because if $x \leq d$ copies are running simultaneously, they each take $1/x$ of the time of a single jobs, and the other copies don't contribute to the load since they never run. So load is exactly the same in d copy system as if there were only 1 copy per job. IN REALITY, however, the copies are NOT independently sizes, since all copies of a job are the same. Here load is in fact added. See the MASCOTS 2016 paper above!

Understanding MAXs: General Distribution

Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim X$. Assume we know X.

Want to understand:

$$Z = \max(X_1, \ldots, X_n)$$

Question: What is $F_Z(t) = \mathbf{P}\{Z \le t\}$?

$$\begin{aligned}
\mathbf{P}\{Z \le t\} &= \mathbf{P}\{\max(X_1, \ldots, X_n) \le t\} \\
&= \mathbf{P}\{X_1 \le t \ \& \ X_2 \le t \ \& \ \ldots \ \& \ X_n \le t\} \\
&= \mathbf{P}\{X_1 \le t\} \cdot \mathbf{P}\{X_2 \le t\} \cdots \mathbf{P}\{X_n \le t\} \\
&= (\mathbf{P}\{X \le t\})^n
\end{aligned}$$

Question: What is $f_Z(t)$?

$$\begin{aligned}
f_W(t) &= \frac{d}{dt} F_Z(t) \\
&= n \cdot (\mathbf{P}\{X > t\})^{n-1} \cdot f_X(t)
\end{aligned}$$

Understanding MAXs: Exponential Distribution

Let $X_1, X_2, \ldots, X_n$ be i.i.d. $\sim X$. Suppose $X \sim \text{Exp}(\mu)$.

Want to understand:

$$Z = \max(X_1, \ldots, X_n)$$

Question: What is $F_Z(t) = \mathbf{P}\left\{Z \leq t\right\}$?

$$\begin{aligned} \mathbf{P}\left\{Z \leq t\right\} &= \left(\mathbf{P}\left\{X \leq t\right\}\right)^n \\ &= (1 - e^{-\mu t})^n \end{aligned}$$

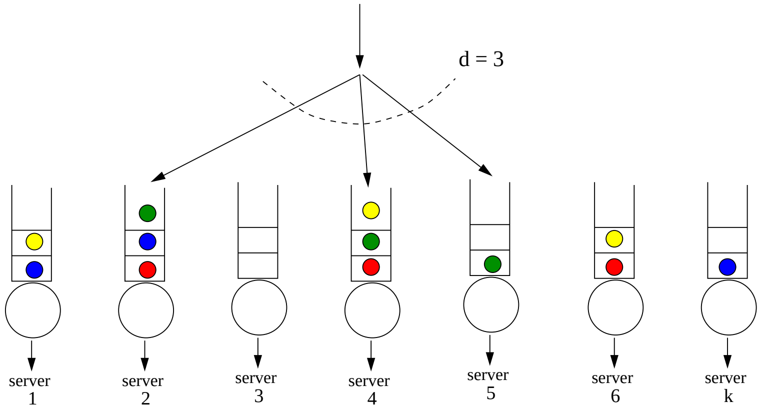
Question: What is $\mathbf{E}\left[Z\right]$?

One could of course derive $f_Z(t)$ and then integrate. However there's an easier way to see this [see our class textbook]:

$$\mathbf{E}\left[Z\right] = \frac{1}{n\mu} + \frac{1}{(n-1)\mu} + \cdots + \frac{1}{\mu}\ .$$

Limited Fork-Join Model

Model is analyzed here: Wang et al.“Delay Asymptotics and Bounds for Multi-Task Parallel Jobs.” Queueing Systems , vol. 91, no. 3-4, March 2019, pp. 207–239.



- Each job split into d parts.
- The d parts go to different random servers.
- Job is complete only when ALL parts are done.

Question: Let T_i represent the reponse time at queue i.
What is T, the overall response time?

$$T = \max(T_1, T_2, \ldots, T_d) \ .$$

Question: How do we compute this?

The problem is that the T_i's are NOT independent. However, if k is high, and d is low, it is close enough to pretend that the T_i's are independent. Thus

$$\overline{F}_T(t) = \mathbf{P}\left\{T \leq t\right\} = \left(\mathbf{P}\left\{T_i \leq t\right\}\right)^d \ .$$

Question: Generally, we need to obtain the distribution of T_i through measurement. However, if job “parts” are Exponentially-distributed in size, and thus T_i is Exponentially distributed with mean $\mathbf{E}[T_i]$, what does this say about $\mathbf{E}[T]$?

$$\mathbf{E}[T] = \frac{\mathbf{E}[T_1]}{d} + \frac{\mathbf{E}[T_1]}{d-1} + \frac{\mathbf{E}[T_1]}{d-2} + \cdots + \mathbf{E}[T_1] \ .$$