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 study these, we need to cover 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 $\overline{F}_W(t) = \mathbf{P}\{W > t\}$?

Question: What is $F_W(t)$?

Question: What is $f_W(t)$?

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

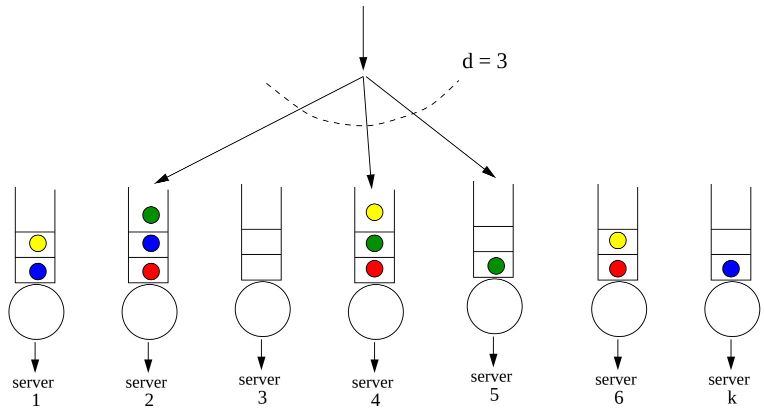
Question: What is $\overline{F}_W(t)$? What is $f_W(t)$?

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

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

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

Redundancy-d Model



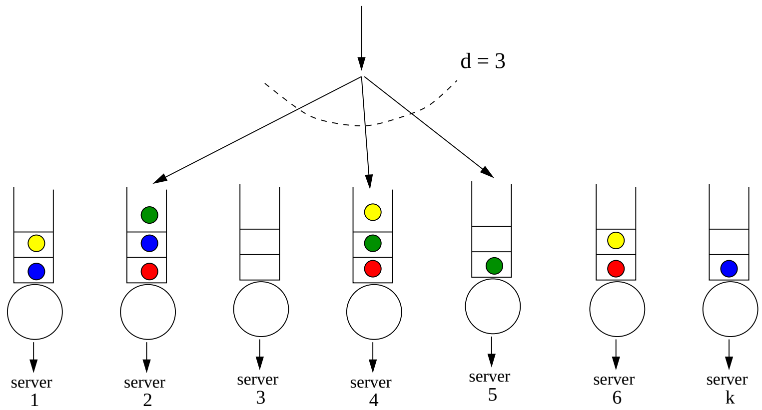
- Each job creates d copies of itself.
- The d copies go to different random servers.
- Job is complete as soon as any ONE copy is done.
  Remaining copies are cancelled.

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

Question: How do we compute this?

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?

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

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 $F_Z(t) = \mathbf{P}\{Z \leq t\}$?

Question: What is $f_Z(t)$?

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:

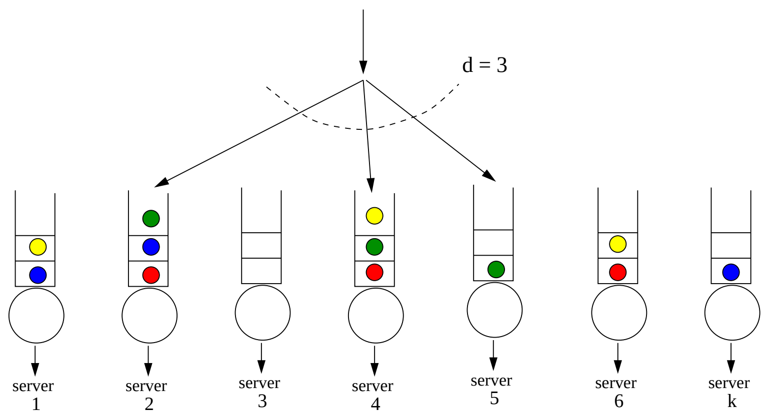
$$Z = \max(X_1, \ldots, X_n)$$

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

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

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?

Question: How do we compute this?

Question: Generally, we obtain the distribution of T_i through measurement. However, if job "parts" have Exponentially-distributed sizes, and thus T_i is Exponentially distributed with mean $\mathbf{E}[T_i]$, what can we say about $\mathbf{E}[T]$?