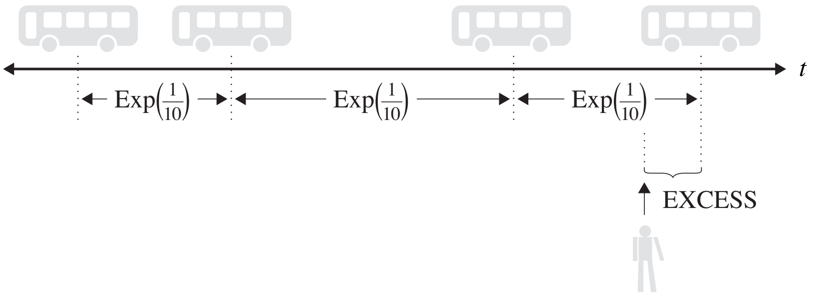
Finishing last lecture ... back to waiting for the bus



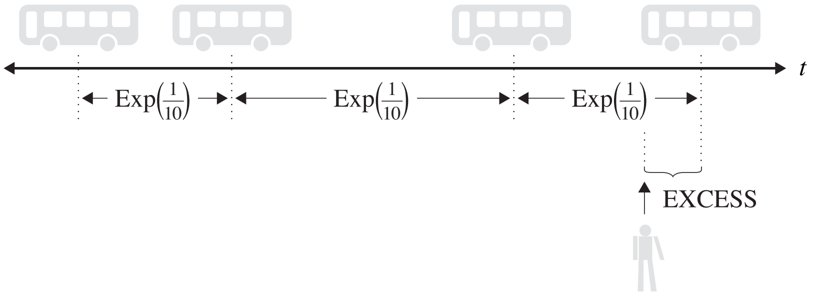
Assume times between buses are i.i.d. and are denoted by r.v. A.

Question: How is A_e defined?

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

Question: How high can $\mathbf{E}[A_e]$ be for general A? ________________

Finishing last lecture ... back to waiting for the bus



Question: What is the expected time between buses **as observed by person arriving at random time?**

Question: What is the Inspection Paradox? Why does it happen?

Back to P-K formula for the M/G/1?

Question: If S denote the job size distribution, how is S_e defined?

Question: What is the P-K formula for $\mathbf{E}\left[T_Q\right]$?

Question: How does load ρ influence delay?

Question: How does C_S^2 influence delay?

Question: How can we have high delay but low load?

From M/G/1 to G/G/1

Question: Express $\mathbf{E}\left[T_Q\right]^{M/G/1}$ in terms of $\mathbf{E}\left[S_e\right]$.

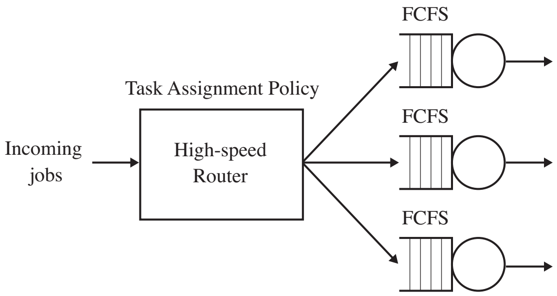
Question: Express $\mathbf{E}\left[T_Q\right]^{M/G/1}$ in terms of C_S^2.

Question: Express $\mathbf{E}\left[T_Q\right]^{M/G/1}$ in terms of $\mathbf{E}\left[T_Q\right]^{M/M/1}$.

Question: Consider the G/G/1 queue. Hard to analyze because don't have PASTA. [Kingman 1961] has an approximation for $\mathbf{E}\left[T_Q\right]^{G/G/1}$. Take a guess ...

Today: Task Assignment in Server Farms

This topic is full of OPEN PROBLEMS. The job of the Dispatcher (router) is to route each arriving job (task) to one of the servers.



<u>Assumptions</u>:

1. k identical servers.
2. Job sizes denoted by r.v. S. Typically assume S has high variability.
3. Job size may or may not be known a priori. Assume known to start.
4. Jobs not preemptible. Processed FCFS and run to completion.

<u>Goal</u>: Find a task assignment policy which minimizes $\mathbf{E}\left[T_Q\right]$.

Question: What are some ideas for policies?

Random & Round-Robin

Question: Draw a picture of Random. What is ρ?

Question: What is $\mathbf{E}[T_Q]$ under the Random policy?

Question: How do Random and Round-Robin compare?

JSQ

Question: What's the advantage of JSQ over Random and Round-Robin?

Question: What are the similarities/differences between JSQ and LWL?

JSQ vs. LWL

Question: Why is LWL likely to be better than JSQ?

Question: When is LWL likely to *not* be much better than JSQ?

Question: What are some weaker versions of JSQ?

See Lu, Xie, Kliot, Geller, Larus, Greenberg. "Join-Idle-Queue: A novel load balancing algorithm for dynamically scalable web servers." Performance Evaluation, Vol 68, no. 11, 2011.

JSQ and LWL

Question: Can we analyze JSQ?

Question: Can we analyze LWL?

LWL vs. M/G/k

Question: How do LWL and Central-Queue M/G/k compare with respect to performance?

Question: What are the advantages of M/G/k over LWL?

Question: What analysis is known for M/G/k?

See Gupta, Harchol-Balter, Dai, Zwart. "On the Inapproximability of M/G/k: Why Two Moments of Job Size Distribution are Not Enough." QUESTA, 2010.

Commercial Break: Announcements

1. Zhouzi's office hours today! GHC 6003.

2. There have been a lot of illnesses. Come find me if you're missing lecture notes. I am also happy to email lecture notes to you!

3. I'm teaching PnC next semester (15-259). Always looking for TAs who enjoy teaching!

SITA

Question: What are the advantages of SITA?

Question: How do we analyze SITA?

Question: How do we choose the size cutoffs for SITA?

See Harchol-Balter and Rein Vesilo. “To Balance or Unbalance Load in Size-Interval Task Allocation.” Probability in the Engineering and Informational Sciences, 2010.

SITA vs. LWL

Question: What are pros/cons of SITA vs LWL?

Question: What is typical behavior of SITA vs. LWL under high C_S^2?

SITA vs. LWL, cont.

Question: What might cause LWL to look much better than SITA?

See papers:

– Harchol-Balter et al. “Surprising Results on Task Assignment,” SIGMETRICS 2009.

– Harchol-Balter et al. “Why Segregating Short Jobs from Long Jobs is Not Always a Win,” Allerton 2009.

– Scheller-Wolf et al. “Delay moments for FIFO GI/GI/s queues.” Queueing Systems 1997.

– Scheller-Wolf et al. “New bounds for expected delay in FIFO GI/GI/s queues.” Queueing Systems 1997.

– Scheller-Wolf et al. “Structural Interpretation and Derivation of Necessary and Sufficient Conditions for delay moments in FIFO Multiserver Queues” Queueing Systems 2006.

What if we don't know the job size?

Question: Which of our task assignment policies work if we don't know the job size a priori?

Question: Can we do anything like SITA without knowing the job size?

See Harchol-Balter "Task Assignment by Guessing Size," Journal of the ACM 2002.