Response Time Modeling I

- service time T_S time for the disk/array to perform the operation. Determined by seek latencies, rotational latencies, and transfer times.

- response time T_R time spent in the disk/array; waiting (queueing) time + T_S.

- N is the number waiting within the queue.

- λ is the operation arrival rate. This is the workload for the disk/array.

- Little's Law $N = T_R \lambda$

Response Time Modeling II

- basic queueing theory.

  λ is the operation arrival rate

  μ is the service rate $= \dfrac{1}{T_S}$.

  $\rho = \lambda / \mu$

- result $N = \dfrac{\rho}{1-\rho}$.

Using Little's law, we have

$$T_R = N / \lambda = \frac{1/\mu}{1-\rho} = \frac{1}{\mu-\lambda}.$$