UCB CS 162 - Operating Systems and Systems Programming "Lecture 19 - Filesystems 1- Performance, Queueing Theory, Filesystem Design"

Lecture 19 - Filesystems 1: Performance, Queueing Theory, Filesystem Design

When we measure the performance of an I/O system, we have to consider queueing time, because the data can pass through multiple queues before our I/O request is satisfied. These queues are present in the OS itself (the ready queue, wait queue, etc. that the process has to go through), and the I/O device itself.

Each queue has a latency L and a bandwidth B. In a simple system, the (maximum) bandwidth is the inverse of the latency.

$$B = \frac{1}{L}$$

If we consider a pipelined system, if we have a pipeline of K stages, our bandwidth increases K times, as expected.

$$B = \frac{K}{L}$$

Another important parameter could be the number of things/operations inside the system at any time. If the latency is L, and the current bandwidth is B, then there are $ B*L $ things/operations inside the system at this time (roughly, assuming the bandwidth isn't constantly changing). This is known as Little's Law.