**CS 153**

**Design of Operating Systems**

**Ch. 8 Scheduling**

**8.1 Performance concepts and terminology**

*utilization* – fraction of that interval spent servicing requests

*throughput* – number of requests serviced during that interval divided by the length of the interval…rate of requests serviced

*waiting time* – time spent waiting in queues for service to begin

*latency* – time from start of service to first output

*processing time* – time required for the server to complete that request

*service time* = latency + processing time

*response time* = waiting time + latency

*turnaround time* = waiting time + latency + processing time

*transfer rate (bandwidth)* – quantity of data divided by transmission time

*backlog* – set of requests waiting for a service at an instant

*queue length* – number of requests in the backlog

*length of a schedule* – total service time required to complete the schedule; the sum of the service times for all requests in the backlog under a schedule

*schedule-sensitive* – if there is a backlog for which different schedules have different lengths for that backlog

**Little’s law.**

number of concurrent jobs = throughput \* average length of a job

N = throughput \* turnaround time (average)

Number waiting = throughput \* waiting time (average)

N = utilization \* (interval / number of jobs)

= utilization / throughput