Queueing Networks

Raj Jain
Washington University in Saint Louis
Saint Louis, MO 63130
Jain@cse.wustl.edu

Audio/Video recordings of this lecture are available at:

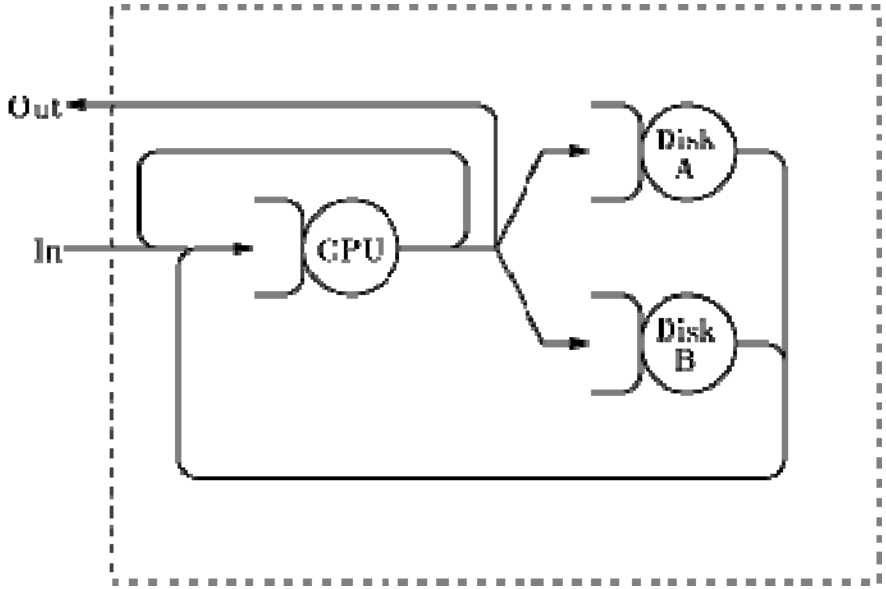
http://www.cse.wustl.edu/~jain/cse567-11/

Overview

1. Open and Closed Queueing Networks
2. Product Form Networks
3. Queueing Network Models of Computer Systems

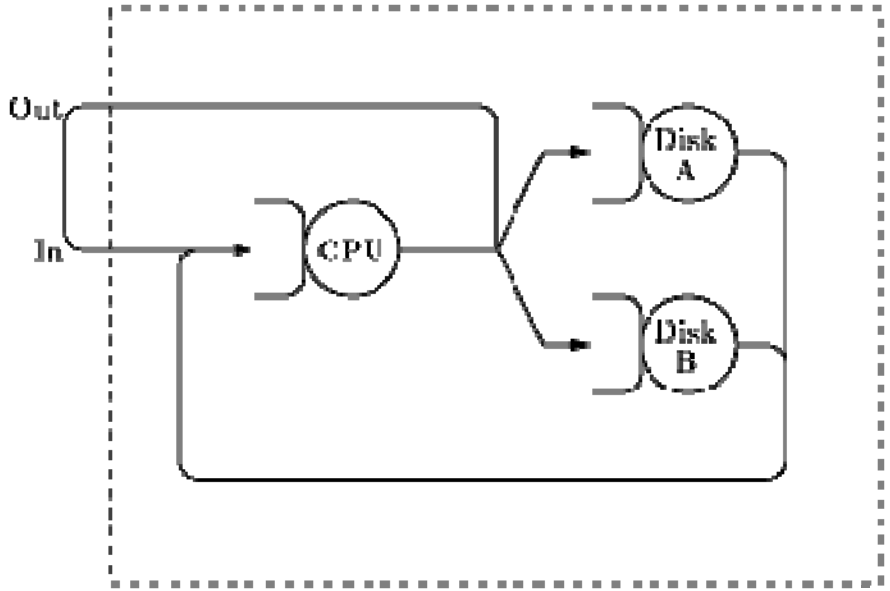
Open Queueing Networks

- **Queueing Network**: model in which jobs departing from one queue arrive at another queue (or possibly the same queue)
- **Open queueing network**: external arrivals and departures
  - Number of jobs in the system varies with time.
  - Throughput = arrival rate
  - Goal: To characterize the distribution of number of jobs in the system.



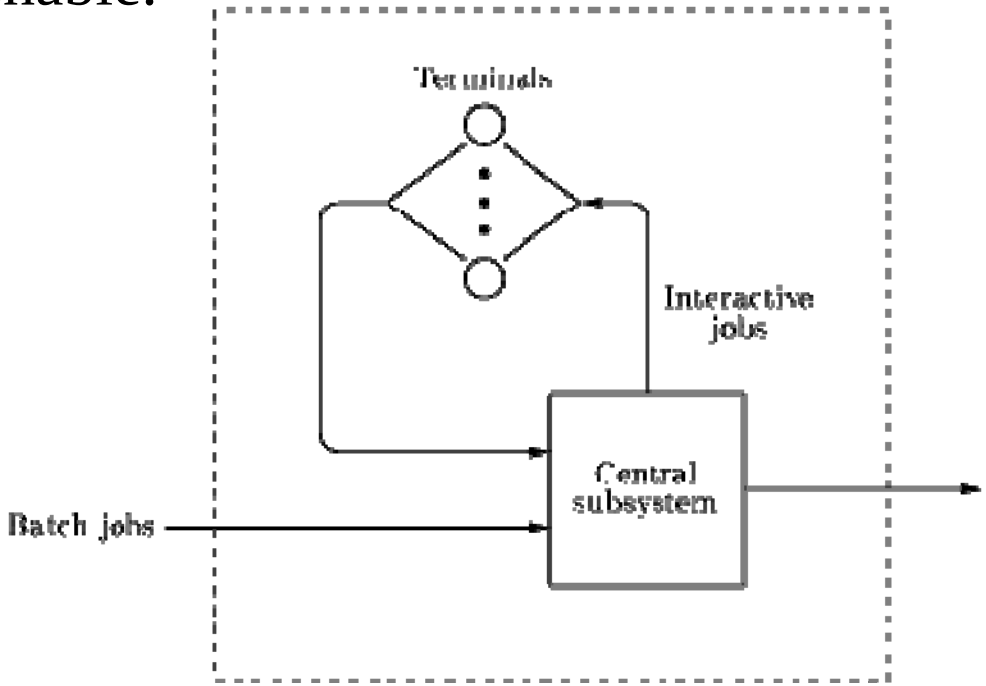
Closed Queueing Networks

- Closed queueing network: No external arrivals or departures
  - Total number of jobs in the system is constant
  - `OUT' is connected back to `IN.'
  - Throughput = flow of jobs in the OUT-to-IN link
  - Number of jobs is given, determine the throughput

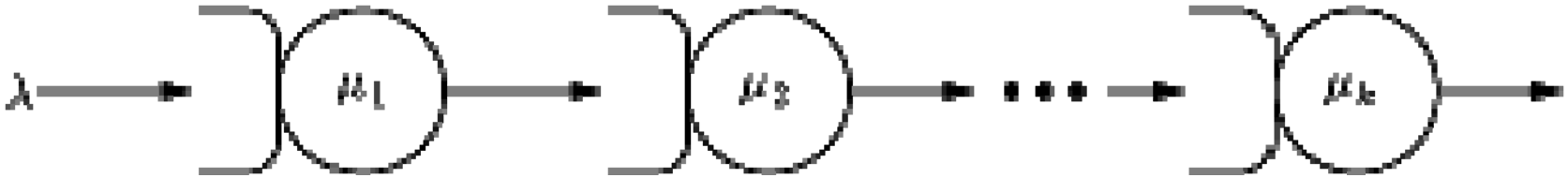


Mixed Queueing Networks

- **Mixed queueing networks**: Open for some workloads and closed for others ⇒ Two classes of jobs. **Class** = types of jobs. All jobs of a single class have the same service demands and transition probabilities. Within each class, the jobs are indistinguishable.



Series Networks



- *k M/M/1* queues in series
- Each individual queue can be analyzed independently of other queues
- Arrival rate= λ. If μ_i is the service rate for i^{th} server:

$$\text{Utilization of } i^{th} \text{ server} \rho_i = \lambda/\mu_i$$

$$\text{Probability of } n_i \text{ jobs in the } i^{\text{th}} \text{ queue } = (1-\rho_i)\rho_i^{n_i}$$

Series Networks (Cont)

- Joint probability of queue lengths:

$$P(n_1, n_2, n_3, \ldots, n_M)$$
$$= (1-\rho_1)\rho_1^{n_1}(1-\rho_2)\rho_2^{n_2}(1-\rho_3)\rho_3^{n_3}\cdots(1-\rho_M)\rho_M^{n_M}$$
$$= p_1(n_1)p_2(n_2)p_3(n_3)\cdots p_M(n_M)$$

$\Rightarrow$ product form network

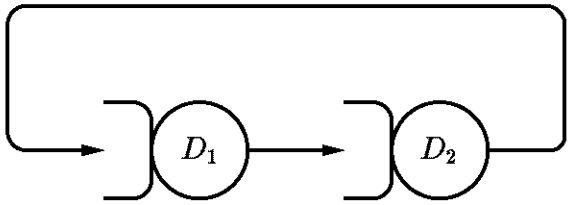
Product-Form Network

- Any queueing network in which:

$$P(n_1, n_2, \ldots, n_M) = \frac{1}{G(N)} \prod_{i=1}^{M} f_i(n_i)$$

- When $f_i(n_i)$ is some function of the number of jobs at the ith facility, $G(N)$ is a normalizing constant and is a function of the total number of jobs in the system.

Example 32.1

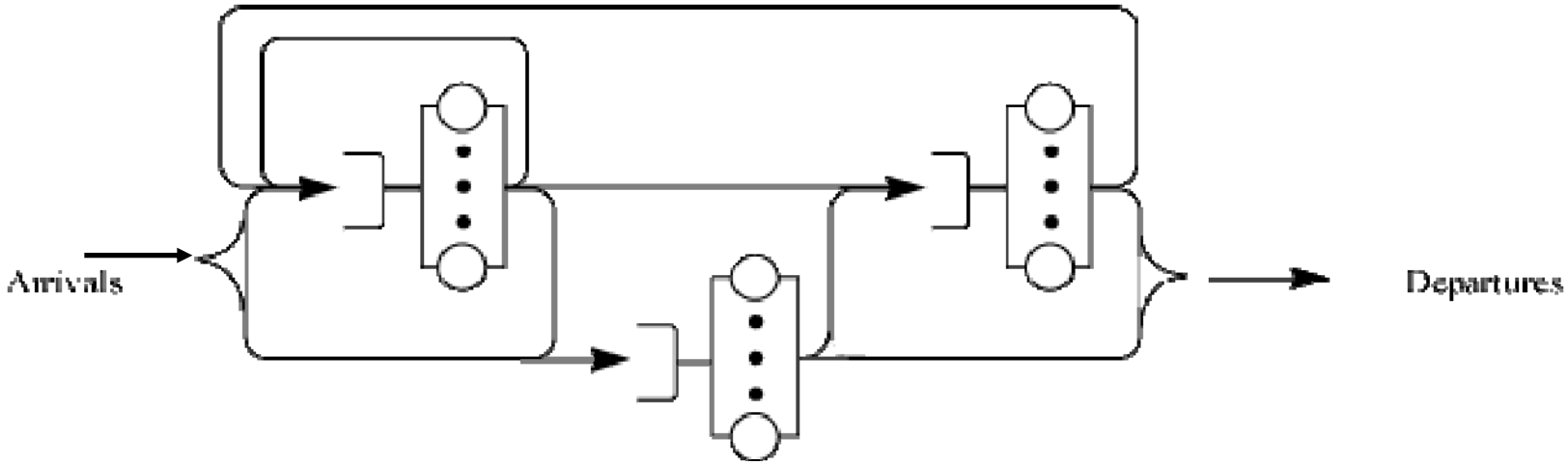


- Consider a closed system with two queues and N jobs circulating among the queues:
- Both servers have an exponentially distributed service time. The mean service times are 2 and 3, respectively. The probability of having n_1 jobs in the first queue and $n_2=N-n_1$ jobs in the second queue can be shown to be:

$$P(n_1, n_2) = \frac{1}{3^{N+1} - 2^{N+1}} (2^{n_1} \times 3^{n_2})$$

- In this case, the normalizing constant $G(N)$ is $3^{N+1}-2^{N+1}$.
- The state probabilities are products of functions of the number of jobs in the queues. Thus, this is a ***product form network***.

General Open Network of Queues



- Product form networks are easier to analyze
- Jackson (1963) showed that any arbitrary open network of m-server queues with exponentially distributed service times has a product form

General Open Network of Queues (Cont)

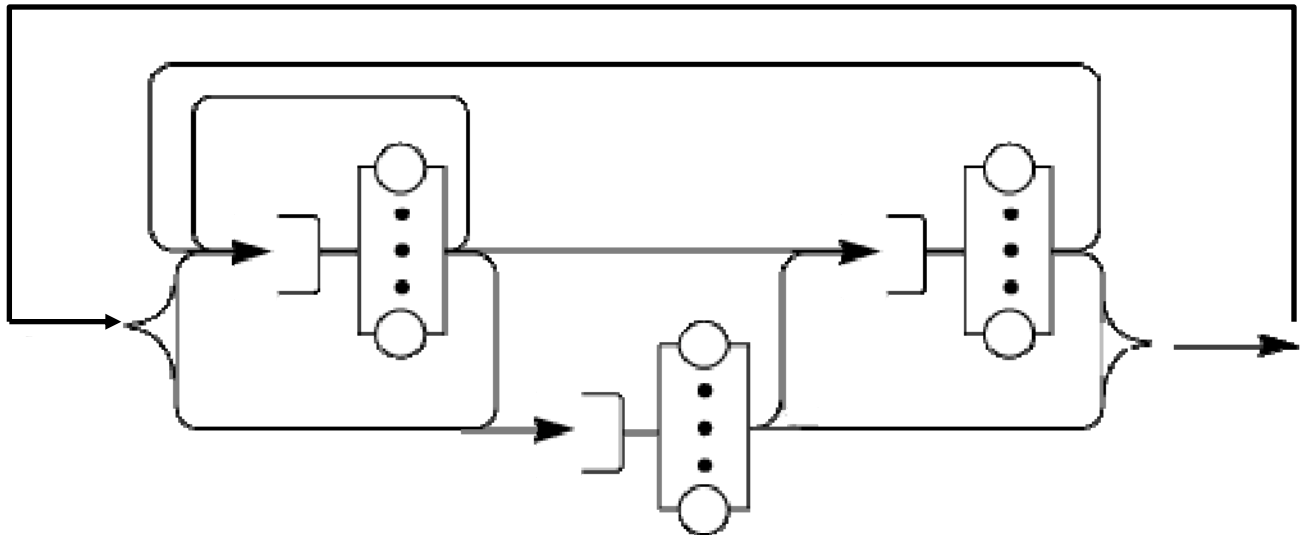
- If all queues are single-server queues, the queue length distribution is:

$$\begin{aligned} &P(n_1, n_2, n_3, \ldots, n_M) \\ &= (1-\rho_1)\rho_1^{n_1}(1-\rho_2)\rho_2^{n_2}(1-\rho_3)\rho_3^{n_3}\cdots(1-\rho_M)\rho_M^{n_M} \\ &= p_1(n_1)p_2(n_2)p_3(n_3)\cdots p_M(n_M) \end{aligned}$$

- Note: Queues are not independent *M/M/1* queues with a Poisson arrival process.
- In general, the internal flow in such networks is not Poisson. Particularly, if there is any feedback in the network, so that jobs can return to previously visited service centers, the internal flows are not Poisson.

Closed Product-Form Networks

- Gordon and Newell (1967) showed that any arbitrary closed networks of m-server queues with exponentially distributed service times also have a product form solution.
- Baskett, Chandy, Muntz, and Palacios (1975) showed that product form solutions exist for an even broader class of networks.



BCMP Networks

1. **Service Disciplines**:
   - First-come-first-served (FCFS),
   - Processor sharing (PS),
   - Infinite servers (IS or delay centers), and
   - Last-come-first-served-preemptive-resume (LCFS-PR).
2. **Job Classes**: The jobs belong to a single class while awaiting or receiving service at a service center, but may change classes and service centers according to fixed probabilities at the completion of a service request.
3. **Service Time Distributions**:
   - At FCFS service centers, the service time distributions must be identical and exponential for all classes of jobs.

BCMP Networks(Cont)

- At other service centers, where the service times should have probability distributions with rational Laplace transforms;
- Different classes of jobs may have different distributions.

4. State Dependent Service:

- The service time at a FCFS service center can depend only on the total queue length of the center.
- The service time for a class at PS, LCFS-PR, and IS center can also depend on the queue length for that class, but not on the queue length of other classes.
- Moreover, the overall service rate of a subnetwork can depend on the total number of jobs in the subnetwork.

BCMP Networks(Cont)

5. Arrival Processes:

- In open networks, the time between successive arrivals of a class should be exponentially distributed.
- No bulk arrivals are permitted.
- The arrival rates may be state dependent.
- A network may be open with respect to some classes of jobs and closed with respect to other classes of jobs.

Non-Markovian Product Form Networks

By Denning and Buzen (1978)

1. **Job Flow Balance**: For each class, the number of arrivals to a device must equal the number of departures from the device.
2. **One Step Behavior**: A state change can result only from single jobs either entering the system, or moving between pairs of devices in the system, or exiting from the system. This assumption asserts that simultaneous job-moves will not be observed.
3. **Device Homogeneity**: A device's service rate for a particular class does not depend on the state of the system in any way except for the total device queue length and the designated class's queue length. This assumption implies the following:

Non-Markovian PFNs (Cont)

a. Single Resource Possession: A job may not be present (waiting for service or receiving service) at two or more devices at the same time.

b. No Blocking: A device renders service whenever jobs are present; its ability to render service is not controlled by any other device.

c. Independent Job Behavior: Interaction among jobs is limited to queueing for physical devices, for example, there should not be any synchronization requirements.

d. Local Information: A device's service rate depends only on local queue length and not on the state of the rest of the system.

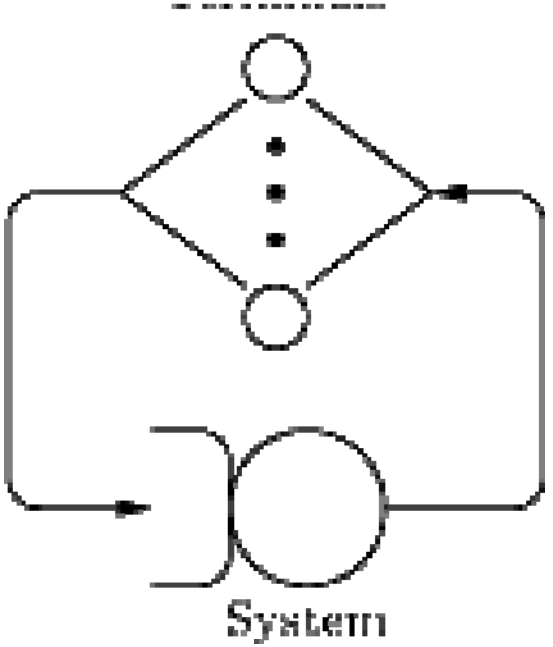
Non-Markovian PFNs (Cont)

e. Fair Service: If service rates differ by class, the service rate for a class depends only on the queue length of that class at the device and not on the queue lengths of other classes. This means that the servers do not discriminate against jobs in a class depending on the queue lengths of other classes.

4. Routing Homogeneity: The job routing should be state independent.

The routing homogeneity condition implies that the probability of a job going from one device to another device does not depend upon the number of jobs at various devices.

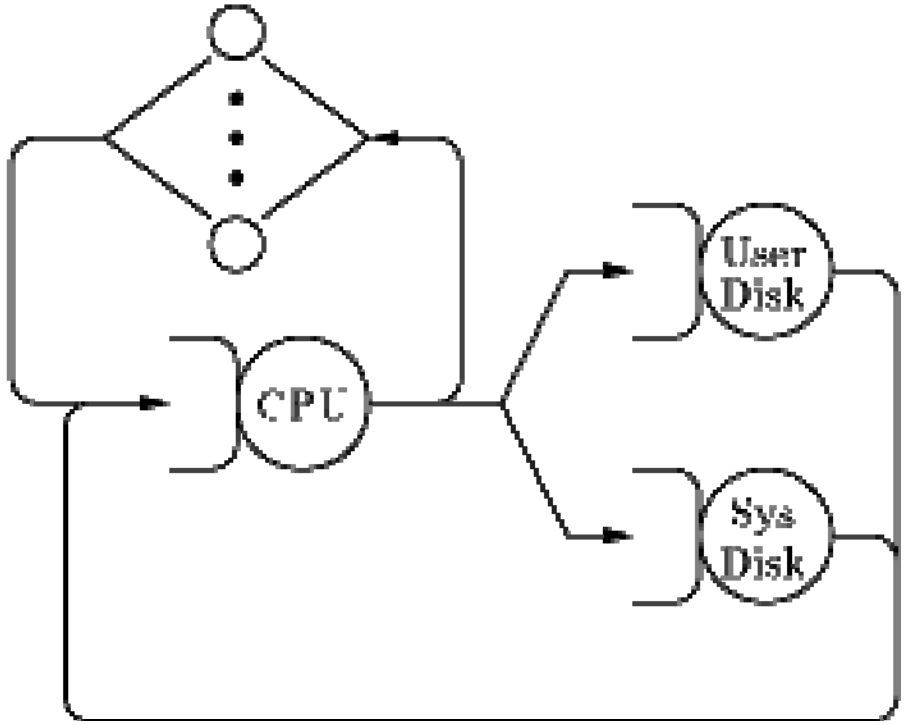
Machine Repairman Model



- Originally for machine repair shops
- A number of working machines with a repair facility with one or more servers (repairmen).
- Whenever a machine breaks down, it is put in the queue for repair and serviced as soon as a repairman is available
- Scherr (1967) used this model to represent a timesharing system with *n* terminals.
- Users sitting at the terminals generate requests (jobs) that are serviced by the system which serves as a repairman.
- After a job is done, it waits at the user-terminal for a random ``think-time'' interval before cycling again.

Central Server Model

- Introduced by Buzen (1973)
- The CPU is the ``central server'' that schedules visits to other devices
- After service at the I/O devices the jobs return to the CPU



Types of Service Centers

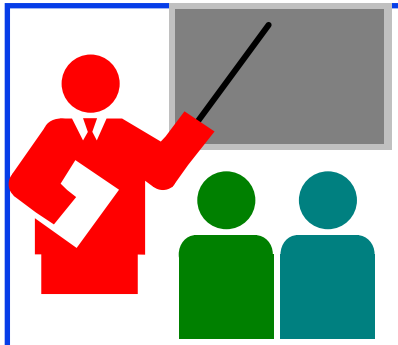
Three kinds of devices

1. **Fixed-capacity service centers**: Service time does not depend upon the number of jobs in the device

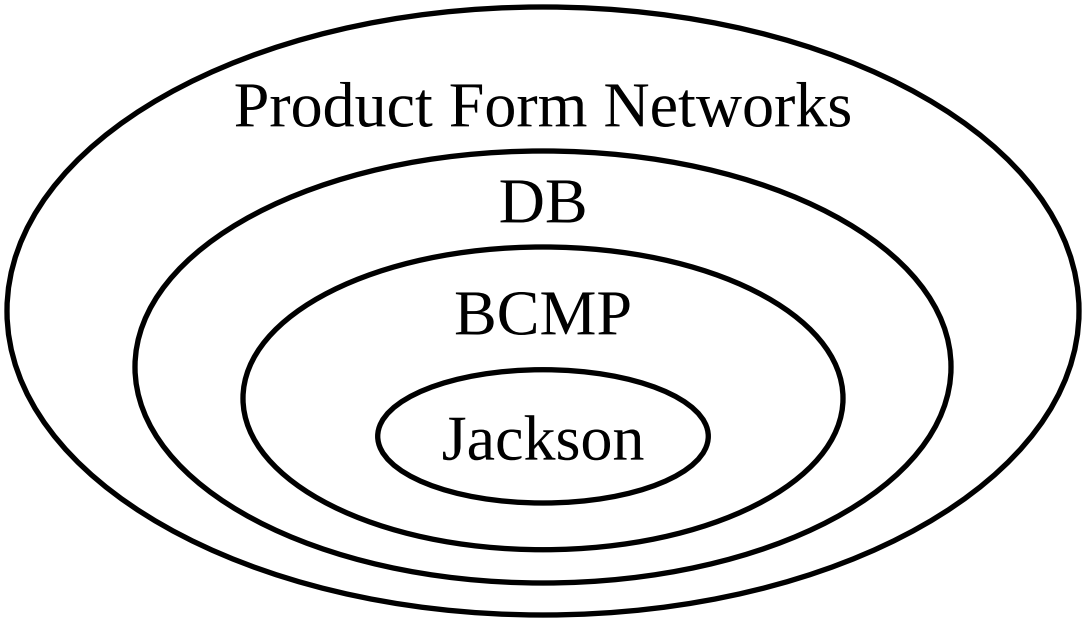
For example, the CPU in a system may be modeled as a fixed-capacity service center.

2. **Delay centers or infinite server**: No queueing. Jobs spend the same amount of time in the device regardless of the number of jobs in it. A group of dedicated terminals is usually modeled as a delay center.
3. **Load-dependent service centers**: Service rates may depend upon the load or the number of jobs in the device., e.g., *M/M/m* queue (with $m \geq 2$)

A group of parallel links between two nodes in a computer network is another example



Summary



- Product form networks: Any network in which the system state probability is a product of device state probabilities
- Jackson: Network of M/M/m queues
- BCMP: More general conditions
- Denning and Buzen: Even more general conditions

Homework 32

- In a series network of three routers, the packets arrive at the rate of 100 packets/second. The service rate of the three routers is 250 packets/s, 150 packets/s, and 200 packets/s.
- Write an expression for the state probability of the system.
- Calculate the probability of having 2 packets at each of the three routers.