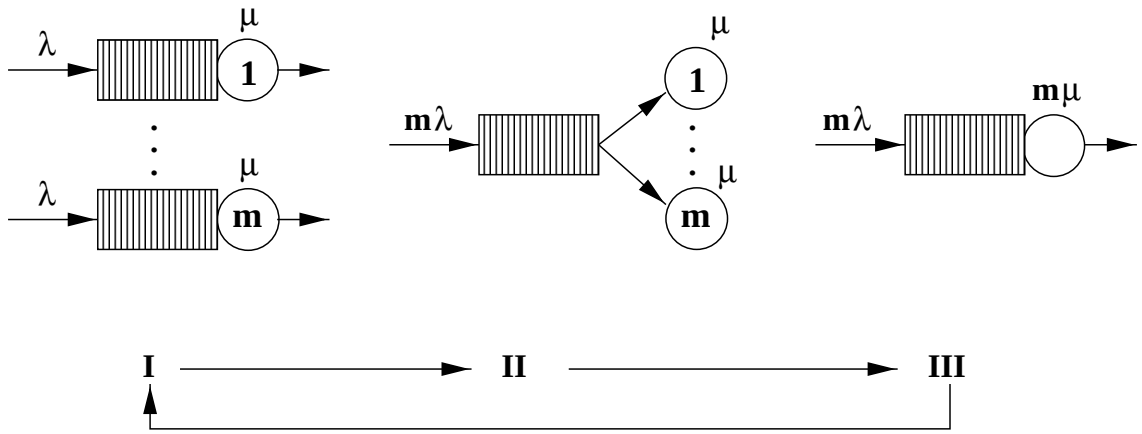
Service Engineering July 7, 1997
Revised June 1998, 19/12/1999

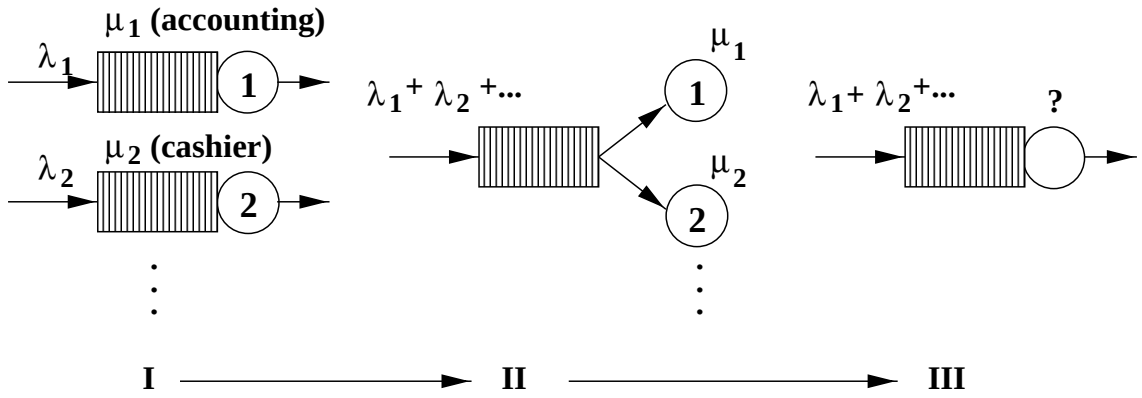
Leading to Q-Nets *(Motivating via Flexibility)*

1. In Kleinrock (Vol 2, pg 279), the following 3 models are compared:



Homogeneous (statistically identical) servers: suffices to analyze *EOS (Economies of Scale)* and some effects of *pooling*.

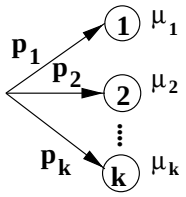
2. The above is not natural for the analysis of *specialization*: each server performs something else.



Model II is not very relevant, when different tasks are considered.
(It is appropriate when servers perform the same task, but have varying service rates.)

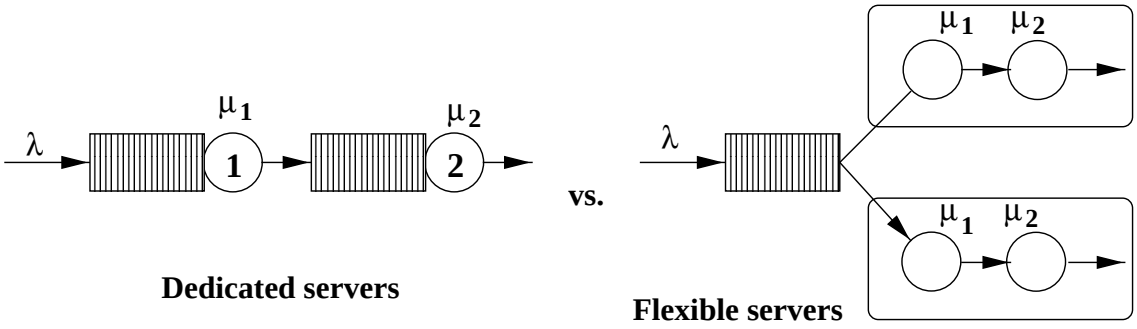
Model III is a single flexible server, capable of doing everything.
How to model the service time of a flexible server?

In this case, hyper-exponential or, more generally, phase-type distribution, seems reasonable.

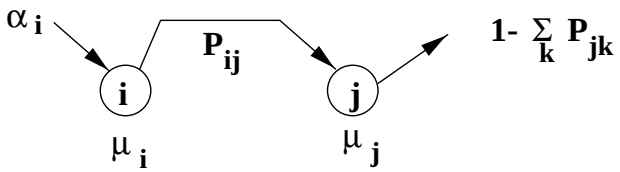


$$p_i = \frac{\lambda_i}{\lambda_1 + \lambda_2 + \ldots}$$

For example, compare



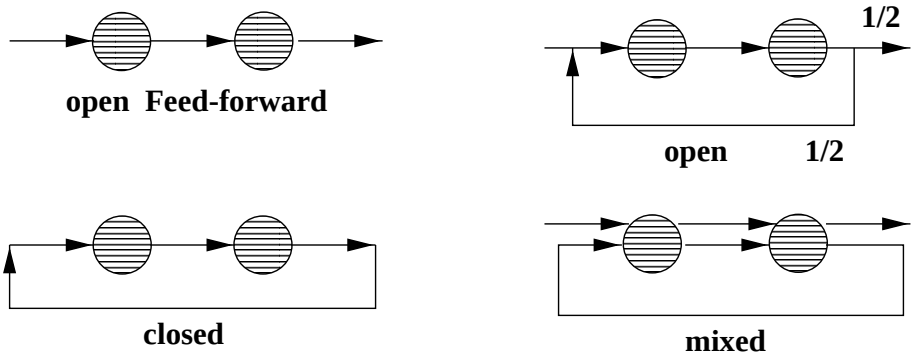
Dedicated servers: 2 single-server queues in series.
Flexible servers: 2-server queues with phase-type service distribution.
Generally,



Reference: Mandelbaum, A. & Reiman, M.I., "On Pooling in Queueing Networks", *Management Science*, 44, 971-981, 1998.

Jackson Networks (Open)

Examples.



Closed model seems to be less important for services (at least for now).

Model

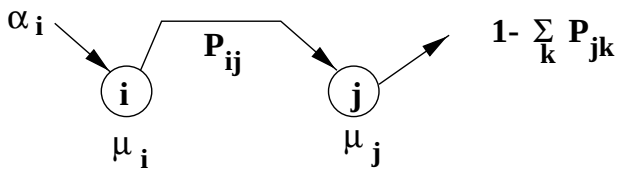
Services - K service stations, indexed by $i, j, k = 1, \ldots, K$.
Each station j has a single server that provides service with duration $\sim \exp(\mu_j)$.
Services are independent.

Arrivals - External arrivals: Poisson (α_j) to station j
Alternatively: a single Poisson (α) arrival;
splits to j with probability P_{0j}, thus $\alpha_j = \alpha P_{0j}$

Routing (Markovian) - After service at station j, move on to station k with probability P_{jk}, join queue at k (hence, move out from j with probability $1 - \sum_{k=1}^{K} P_{jk}$), etc.

Independence - Mechanisms of arrivals, services, switches are independent of each other.

Data $\alpha = (\alpha_j) \geq 0,\ \mu = (\mu_j) \geq 0,\ P = [P_{jk}]$ substochastic.



(1) **Assumptions** $\alpha \neq 0,\ \mu > 0,\ \exists\, [I - P]^{-1}$.
(Equivalent to finitely many visits to all states: recall Phase-type.)
For example, P irreducible and $\exists j \ni \sum_k P_{jk} < 1$.

Restrictions Single server (generalizable, even with exact analysis).
Exponentiality (generalizable only approximately).

We consider a "Dumb" model with *Homogeneous* customers.
(Extensions exist, but, in general, they are very complicated).

The **Fluid-View** (Traffic Equations)

$$\lambda_j = \alpha_j + \sum_i \delta_i P_{ij}, \qquad j = 1, \ldots, K$$

inflow rate to j ↗ ↖ outflow rate, out of i

$$\underline{\delta_j = \lambda_j \wedge \mu_j \ .}$$

Vector form: $\boxed{(*) \quad \lambda = \alpha + (\lambda \wedge \mu)P}$ (or $\lambda = \alpha + P^T(\lambda \wedge \mu)$) .

Fact: Under assumptions (1), $(*)$ has a unique solution (proof via Fixed Point results).

3 possibilities for every station:

$\lambda_j > \mu_j$ overloaded (supercritical);
$\lambda_j = \mu_j$ critically loaded;
$\lambda_j < \mu_j$ properly loaded (nonstandard terminology).

Bottleneck (strict) $\lambda_j \geq \mu_j$ $(\lambda_j > \mu_j)$

What is to be expected of strict bottlenecks: explode at rate $(\lambda_j - \mu_j)$.
What is to be expected of other bottlenecks: null-recurrence.

(2) Assume: $\lambda_j < \mu_j,\ \forall\, j$; equivalently $\rho_j = \frac{\lambda_j}{\mu_j} < 1$. **(traffic intensity)**
Then $(*)$ reduces to $\lambda = \alpha + \lambda P$

$$\boxed{\lambda = \alpha[I - P]^{-1}} \qquad (= \alpha R, \text{ where } R = [I - P]^{-1} \text{ – fundamental matrix})$$

Traffic equations for a network without bottlenecks:

$$\begin{aligned} \lambda_j &= \text{arrival rate to } j \qquad \text{(external + internal)} \\ &= \text{departure rate.} \end{aligned}$$

Solve $\lambda = \alpha + \lambda P$ by $\lambda = \alpha[I - P]^{-1} = \alpha R$, that is

$$\lambda_j = \sum_i \alpha_i R_{ij} = \text{ load on } j, \text{ measured in average number of visits per time-unit.}$$

R_{ij} = average number of visits to station j, of a customer that enters at station i.

MJP Representation $X(t) = (X_1(t), \ldots, X_K(t)), \quad t \geq 0$

$\uparrow$

in station 1

State $n = (n_1, \ldots, n_K), \quad n_k = 0, 1, 2, \ldots$

Transitions:

- external arrival to j: $X_j(t) \to X_j(t) + 1$, at rate α_j.
- service completion at j, followed by a transfer to k:

$$X_j(t) \to X_j(t) - 1, \; X_k(t) \to X_k(t) + 1, \text{ at rate } \mu_j P_{jk} \; .$$

- service completion at j, followed by a departure from network

$$X_j(t) \to X_j(t) - 1, \text{ at rate } \mu_j \left(1 - \sum_k P_{jk}\right) \; .$$

We can write, in principle, steady-state equations, and solve them.

Jackson's Theorem (1957, 1963; seminal contribution to the Theory of Q-Nets)

Under assumptions (1) and (2), the network is ergodic with stationary distribution

$$\pi(n) = \textstyle\prod_{j=1}^{K} \pi_j(n_j) \; , \qquad n = (n_1, \ldots, n_K) \tag{3}$$

where

$$\pi_j(x) = \rho_j^x (1 - \rho_j) \; , \qquad x = 0, 1, 2, \ldots .$$

Description of Proof(s):

If, for some j, $\lambda_j = 0$, then station j is empty in steady state ($\pi_j(0) = 1$), and it can be excluded from consideration. Thus, assume $\lambda_j > 0, \; j = 1, \ldots, K$, without loss of generality. This implies irreducibility of X (e.g., state $(0, 0, \ldots, 0)$ communicates with all other states).

- One can now verify directly that (3) solves the steady-state equations (Asmussen, pg. 68).
- An alternative indirect way exploits the notion of reversibility (e.g., this is the proof in Bertsekas & Gallager, pp. 223–225).
- There is also a heuristic explanation/proof, due to J. Walrand, also described in Bertsekas & Gallager, pp. 227–228.

– There are additional directions, each trying to add insight into this wonderful surprising result.

Why wonderful? So simple! Every station, in isolation, is **M/M/1-like** with parameters λ_j, μ_j; *and* stations are *independent* in equilibrium, when viewed at a snapshot (but not at different times).

Why surprising! Internal flows are typically *not* Poisson and one expects a "lot of dependence". (Incidentally, outflows *are* Poisson; internal flows are Poisson only when there are no loops.)

Network Performance

Think in terms of a single exogenous arrival stream Poisson (α), which splits to station j with probability α_j/α.

Jackson's Theorem: $X_j \sim$ Geometric $(p = 1 - \rho_j)$, independent.

$$E(X_j) = \frac{\rho_j}{1-\rho_j} \quad .$$

L = total # in system: $E(L) = \sum_{j=1}^{K} \frac{\rho_j}{1-\rho_j}$, $\quad \mathrm{Var}(L) = \sum_{j=1}^{K} \frac{\rho_j}{(1-\rho_j)^2}$.

Time in system by Little:

$$E(W) = \frac{1}{\alpha} E(L) = \frac{1}{\alpha} \sum_{j=1}^{K} \frac{\rho_j}{1-\rho_j} \qquad \left(\alpha = \sum_{j=1}^{K} \alpha_j\right)$$

$$= \sum_j \frac{\lambda_j}{\alpha} \frac{1}{\mu_j} \frac{1}{1-\rho_j} = \sum_j \frac{\lambda_j}{\alpha} E(W_j).$$

↑ steady-state delay at j

Recall: $\lambda_j = \sum_{i=1}^{K} \alpha_i R_{ij}$, where $R = [R_{ij}] = [I - P]^{-1}$.

↑

mean # of visits to j by a customer entering at i

Hence,

$$E(W) = \sum_i \frac{\alpha_i}{\alpha} \underbrace{\sum_j R_{ij} E(W_j)}$$

Prob. to start at station i

total sojourn time for a customer that enters at i