Markovian queueing networks (cont’)
Kelly network

- K single-server FIFO nodes with infinite waiting room
- R classes of customers. Class-k defined by deterministic route $c_r = (c_r(1), \ldots, c_r(n_r))$
  
  $c_r(i) =$ identity node visited by class-r at $i$-th visit
  
  $n_r =$ number of nodes visited by class-r customer
  
  Same node can be visited several times, i.e.
  
  $c_r(i_1) = \ldots = c_r(i_j)$

- Class-r arrivals are Poisson with rate $\lambda_r > 0$

- Service times at node k are exponentially distributed with rate $\mu_k$

- All above rvs are mutually independent
Kelly network (cont')

A job shop model
Kelly network (con’t)

\( \hat{\lambda}_{k,r} \) = rate of class-\( r \) customers entering node \( k \)

\[
\hat{\lambda}_{k,r} = \lambda_r \sum_{i=1}^{n_r} 1(c_r(i) = k)
\]

If class-\( r \) customers enter node \( k \) only once then

\( \hat{\lambda}_{k,r} = \lambda_r \)

\( \hat{\lambda}_k \) = total arrival rate in node \( k \)

\[
\hat{\lambda}_k = \sum_{r=1}^{R} \hat{\lambda}_{k,r}
\]
\( M(t) = [M_{k,r}(t)]_{1 \leq k \leq K, 1 \leq r \leq R} \) state of network at time \( t \)

\( M_{k,r}(t) = \# \text{ class-}r \text{ customers in node } k \text{ at time } t \)

\[ \{M(t), t \geq 0\} \text{ is not a CTMC} \]

To see this, assume that system is in state \( M=[m_{k,r}]_{k,r} \), where \( M \) is an arbitrary \( K \)-by-\( R \) matrix with nonnegative integer-valued entries (i.e. \( m_{k,r} = 0,1,... \) for each \( k \) and \( r \))

Since we do not know the class of the customer in a server, we cannot find transition rates of the type

\[ M \rightarrow M - E(k,r) \text{ if } m_{k,r} > 0 \]

where \( E(k,r) \) is a \( K \)-by-\( R \) matrix whose all entries are zero except entry \( (k,r) \) which is one
We obtain a CMTC if one knows the class of the customer along with its position in its route (e.g. position i if it visits the ith node on its route) at each position of each node (recall that nodes are FIFO)

Position in route is needed in case customer visits a node more than once as one needs to know where to route it next

[See related analysis done in Lecture 12 for multiclass LFCS queue in isolation]

One can write balance equations and solve them

By aggregating states, we get the following result:
Result (Limiting distribution of number of jobs of each class in Kelly network)

If \( \hat{\lambda}_k < \mu_k \) for \( k=1,...,K \) (stability) then

\[
\pi(M) = \lim_{t \to \infty} P(M(t)=M) \text{ exists for each K-by-R matrix } M = [m_k,r] \text{ regardless of the initial state, and}
\]

\[
\pi(M) = \prod_{k=1}^{K} \left[ \left(1 - \frac{\hat{\lambda}_k}{\mu_k}\right) \left(\sum_{r=1}^{R} m_{k,r}\right)! \prod_{r=1}^{R} \frac{1}{m_{k,r}!} \frac{\hat{\lambda}_{k,r}}{\mu_k} \right]
\]

Performance measures

- \( N_{k,r} = \# \text{ class-}r \text{ customers in node } k \)

\[
N_{k,r} = \sum_{m_{l,s} \geq 0} \sum_{l=1,\ldots,K} m_{k,r} \prod_{l=1}^{K} \left( 1 - \frac{\hat{\lambda}_l}{\mu_l} \right) \left( \sum_{s=1}^{R} m_{l,s} \right) ! \prod_{s=1}^{R} \frac{1}{m_{l,s} !} \frac{\hat{\lambda}_{l,s}}{\mu_l}
\]

- \( N_k = \# \text{ customers in node } k = \sum_{1 \leq r \leq R} N_{k,r} \)

\[
N_k = \frac{\hat{\lambda}_k}{\mu_k - \hat{\lambda}_k}
\]
Performance measures (con’t)

- $T_r = \text{expected sojourn time class-}r\ \text{customers}$

Recall:
- $c_r = (c_r(1), ..., c_r(n_r))$
- Expected # class-r in node k: $N_{k,r} = \frac{\hat{\lambda}_{k,r}}{\mu_k - \hat{\lambda}_k}$

By Little $T_r = \frac{1}{\lambda_r} \sum_{i=1}^{n_r} N_{c_r(i),r}$ so that

- $T = \text{expected sojourn time of an arbitrary cust.}$

$T = \frac{1}{\sum_{r=1}^{R} \lambda_r} \sum_{k=1}^{N} N_k \ (\text{Little})$ so that

$T = \frac{1}{\sum_{r=1}^{R} \lambda_r} \sum_{k=1}^{N} \frac{\hat{\lambda}_k}{\mu_k - \hat{\lambda}_k}$
Web servers

Model as Kelly network: 4 classes

- $c_1 = (A)$
  - $\lambda_1 = 0.5\lambda_A$
- $c_2 = (AB, B, BA)$
  - $\lambda_2 = 0.5\lambda_A$
- $c_3 = (B)$
  - $\lambda_3 = 0.8\lambda_B$
- $c_4 = (BA, A, AB)$
  - $\lambda_4 = 0.2\lambda_B$

$\lambda_A = 2000\text{req/sec}$

$\lambda_B = 500\text{req/sec}$
Each server processes 2000 req/sec.
Bandwidth of links A->B and B->A is 32Mb/sec.
Average size documents & requests 1Kbytes

\[ \mu_A = \mu_B = 2000 \text{ req/sec} \quad \mu_{AB} = \mu_{BA} = 4000 \text{ req/sec} \]
 Arrival rates = ?  
\[c_1 = (A) \quad c_2 = (AB, B, BA) \quad c_3 = (B) \quad c_4 = (BA,A, AB)\]
\[
\begin{align*}
\lambda_1 & = 0.5\lambda_A \\
\lambda_2 & = 0.5\lambda_A \\
\lambda_3 & = 0.8\lambda_B \\
\lambda_4 & = 0.2\lambda_B
\end{align*}
\]

\[\hat{\lambda}_{k,r} = \text{arrival rate class-}r \text{ in node } k\]
\[\hat{\lambda}_k = \text{arrival rate node } k = \sum_{r=1}^{R} \hat{\lambda}_{k,r}\]

\[
\begin{align*}
\hat{\lambda}_{A,1} & = \lambda_1 \\
\hat{\lambda}_{AB,1} & = 0 \\
\hat{\lambda}_{BA,1} & = 0 \\
\hat{\lambda}_{B,1} & = 0 \\
\hat{\lambda}_{A,2} & = 0 \\
\hat{\lambda}_{AB,2} & = \lambda_2 \\
\hat{\lambda}_{BA,2} & = \lambda_2 \\
\hat{\lambda}_{B,2} & = \lambda_2 \\
\hat{\lambda}_{A,3} & = 0 \\
\hat{\lambda}_{AB,3} & = 0 \\
\hat{\lambda}_{BA,3} & = 0 \\
\hat{\lambda}_{B,3} & = \lambda_3 \\
\hat{\lambda}_{A,4} & = \lambda_4 \\
\hat{\lambda}_{AB,4} & = \lambda_4 \\
\hat{\lambda}_{BA,4} & = \lambda_4 \\
\hat{\lambda}_{B,4} & = 0
\end{align*}
\]

\[
\begin{align*}
\hat{\lambda}_A & = \lambda_1 + \lambda_4 \\
\hat{\lambda}_{AB} & = \lambda_2 + \lambda_4 \\
\hat{\lambda}_{BA} & = \lambda_2 + \lambda_4 \\
\hat{\lambda}_B & = \lambda_2 + \lambda_3
\end{align*}
\]
\[ \hat{\lambda}_A = \lambda_1 + \lambda_4 = 0.5\lambda_A + 0.2\lambda_B < \mu_A = 2000 \]
\[ \hat{\lambda}_{AB} = \lambda_2 + \lambda_4 = 0.5\lambda_A + 0.2\lambda_B < \mu_{AB} = 4000 \]
\[ \hat{\lambda}_{BA} = \lambda_2 + \lambda_4 = 0.5\lambda_A + 0.2\lambda_B < \mu_{BA} = 4000 \]
\[ \hat{\lambda}_B = \lambda_2 + \lambda_3 = 0.5\lambda_A + 0.8\lambda_B < \mu_B = 2000 \]

- Stability condition: \(0.5\lambda_A + 0.8\lambda_B < 2000\)
- \(T_A\) = expected response time in zone A?
\[ T_A = (1/\lambda_A)E[\# \text{ customers which have entered the system from zone A}] \]
\[ = (1/\lambda_A) \sum_{k=A,AB,BA,B} \sum_{r=1,2} N_{k,r} \quad \text{with} \]
\[ N_{k,r} = \frac{\hat{\lambda}_{k,r}}{\mu_k - \hat{\lambda}_k} \]
\( T_A \) = expected response time in zone A?

\[
T_A = \frac{1}{\lambda_A} \left( \frac{\lambda_1}{\mu_A - \hat{\lambda}_A} + \frac{\lambda_2}{\mu_{AB} - \hat{\lambda}_{AB}} + \frac{\lambda_2}{\mu_{BA} - \hat{\lambda}_{BA}} + \frac{\lambda_2}{\mu_A - \hat{\lambda}_A} \right)
\]

\[
T_A = \frac{0.5}{2000 - 0.5\lambda_A - 0.2\lambda_B} + \frac{1}{4000 - 0.5\lambda_A - 0.2\lambda_B} + \frac{0.5}{2000 - 0.5\lambda_A - 0.8\lambda_B}
\]

as \( \lambda_1 = \lambda_2 = 0.5\lambda_A, \mu_A = \mu_B = 2000, \mu_{AB} = \mu_{BA} = 4000 \)

Similarly

\[
T_B = \frac{0.2}{2000 - 0.5\lambda_A - 0.2\lambda_B} + \frac{0.4}{4000 - 0.5\lambda_A - 0.2\lambda_B} + \frac{0.8}{2000 - 0.5\lambda_A - 0.8\lambda_B}
\]
\[ T = \text{expected response time arbitrary request?} \]

\[
T = \frac{1}{\sum_{r=A,AB,BA,B} \lambda_r} \sum_{r=A,AB,BA,B} \frac{\hat{\lambda}_k}{\mu_k - \hat{\lambda}_k}
\]

\[
T = \frac{1}{\lambda_A + \lambda_B} \left( \frac{0.5\lambda_A + 0.2\lambda_B}{2000 - 0.5\lambda_A - 0.2\lambda_B} + \frac{\lambda_A + 0.4\lambda_B}{4000 - 0.5\lambda_A - 0.2\lambda_B} + \frac{0.5\lambda_A + 0.8\lambda_B}{2000 - 0.5\lambda_A - 0.8\lambda_B} \right)
\]

Also obtained as

\[
T = \frac{\lambda_A}{\lambda_A + \lambda_B} T_A + \frac{\lambda_B}{\lambda_A + \lambda_B} T_B
\]
Approximating a constant $D$ by an Erlang rv

$X_N = \text{Erlang}(\mu, N)$

$E[X_N] = N/\mu$, $E[(X_N)^2] = N(N+1)/\mu^2$, $\text{var}(X_N) = N/\mu^2$

We want $E[X_N] = N/\mu = D$ so that $\mu = N/D$ yielding $\text{var}(X_N) = D/N$

We can select $N$ to make $D/N$ arbitrary small

Example: $D=1$, $P(|X_N-D|< 0.1) > 0.95$ for $N=400$

$> 0.98$ for $N=600$
Miscellaneous result (cont')

- Expected waiting time in M/D/1 FIFO queue:

\[
E[W_{M/D/1}] = \frac{\lambda D^2}{2(1-\rho)} = \frac{\rho D}{2(1-\rho)}
\]

- Expected waiting time in M/Erlang(N/D,N)/1 FIFO queue:

\[
E[W_{M/Erl/1}] = \frac{\lambda \sigma^{(2)}}{2(1-\rho)} = \frac{\lambda D^2(N+1)/N}{2(1-\rho)} = \frac{\rho D}{2(1-\rho)} \left(1 + \frac{1}{N}\right)
\]

Relative error:

\[
0 \leq \frac{E[W_{M/Erl/1}] - E[W_{M/D/1}]}{E[W_{M/D/1}]} = \frac{1}{N}
\]
Call admission

- Infinite waiting room FCFS server queue fed by \(N\) independent Poisson processes with rate \(\lambda_i, i=1,\ldots,N\)

- Service times required by class-\(i\) customers are iid rvs with cdf \(G_i(x)\), mean \(1/\mu_i\) and 2\(^{nd}\) order moment \(\sigma_i^{(2)}\)

- All above rvs are independent

\(\Rightarrow\) Overall arrival process is Poisson with rate \(\lambda := \sum_{1 \leq i \leq N} \lambda_i\)

Q1: Service time distribution of an arbitrary customer?
**Q1:** Service time distribution of an arbitrary customer?

**A:** \[ G(x) := P(\text{service time new customer} < x) = \sum_{1 \leq i \leq N} P(\text{service time new customer} < x \mid \text{class i}) \frac{\lambda_i}{\lambda} = \frac{1}{\lambda} \sum_{1 \leq i \leq N} \lambda_i G_i(x) \]

Mean and 2\textsuperscript{nd} order moment of arbitrary customer given by

\[
\frac{1}{\mu} := \frac{1}{\lambda} \sum_{i=1}^{N} \rho_i \quad \text{and} \quad \frac{\sigma^{(2)}}{\sigma} := \frac{1}{\lambda} \sum_{i=1}^{N} \lambda_i \bar{\sigma}_i \quad \text{resp., with} \quad \rho_i := \frac{\lambda_i}{\mu_i}
\]

This is M/G/1 queue with arrival rate \( \lambda \), mean service time \( 1/\mu \)

Stable if \( \lambda/\mu < 1 \iff \sum_{i=1}^{N} \rho_i < 1 \)
Call admission (con’t)

Optimization problem: Assume there are $n_1$ sessions of type 1, ..., $n_N$ sessions of type N when a session of type i arrives (=new flow of packets seeks admission)

Q1: Should one accept it if one wants $E[W] \leq \alpha$?

For a stable $M/G/1$ queue

$E[W] = \frac{\lambda \sigma^{(2)}}{2(1 - \rho)}$ \hspace{1cm} (\(\rho<1\))

Here, $E[W] = \frac{\sum_{i=1}^{N} n_i \lambda_i \overline{\sigma}_i^{(2)}}{2 \left(1 - \sum_{i=1}^{N} n_i \rho_i \right)}$, provided $\sum_{1 \leq i \leq N} n_i \rho_i < 1$
Call admission: Effective bandwidth (e.b)

**Q1:** Should we accept it if one wants $E[W] \leq \alpha$?

$$E[W] = \frac{\sum_{i=1}^{N} n_i \lambda_i \sigma_i^{(2)}}{2 \left(1 - \sum_{i=1}^{N} n_i \rho_i\right)}$$

provided $\sum_{1 \leq i \leq N} n_i \rho_i < 1$

$$E[W] \leq \alpha \iff \sum_{i=1}^{N} n_i \lambda_i \sigma_i^{(2)} \leq 2\alpha \left(1 - \sum_{i=1}^{N} n_i \rho_i\right)$$

or

$$\sum_{i=1}^{N} n_i \alpha_i \leq 1 \quad \text{with} \quad \alpha_i := \rho_i + \frac{\lambda_i \sigma_i^{(2)}}{2\alpha} \quad \alpha_i = \text{e.b. of session } i \quad (\alpha_i > \rho_i)$$

- If $\alpha = \infty$ then $\alpha_i = \rho_i$ and constraint becomes

  stability condition $\sum_{i=1}^{N} n_i \rho_i < 1$
Q1: Should one accept new session $i$ if one wants $E[W] \leq \alpha$?

A: Yes, if

$$\alpha_i + \sum_{i=1}^{N} n_i \alpha_i \leq 1$$

with

$$\alpha_i := \rho_i + \lambda_i \frac{\sigma_i^{(2)}}{2\alpha}$$

Nice “decoupling” result
Q2: Should one accept a new session $i$ if one wants $P(W \geq b) \leq q$?

$P(W > b)$ not available for $M/G/1$ queue (unless $G = M$)
Kingman’s bound

- GI/GI/1 queue:
  - interarrival times $\{\tau_n\}_n$ iid rvs
  - service times $\{\sigma_n\}_n$ iid rvs
All these rvs are independent

Kingman’s bound for GI/GI/1: $P(W \geq x) \leq \exp(-\theta x)$
where $\theta > 0$ such that $\phi(\theta) := E[\exp(\theta(\sigma_n - \tau_n))] \leq 1$

See (simple) proof pp. 53-54 of Lecture Notes
Kingman’s bound (cont’)

Kingman’s bound for GI/GI/1: \( P(W \geq x) \leq \exp(-\theta x) \)

where \( \theta > 0 \) such that \( \phi(\theta) := E[\exp(\theta(\sigma_n - \tau_n))] = 1 \)

Such \( \theta > 0 \) exists when queue stable as
\( \phi(0) = 1 \) and \( \phi'(0) = E[\sigma_n] - E[\tau_n] = 1/\mu - 1/\lambda < 0 \)

as long as \( \phi(\theta) \) exists* for some \( \theta > 0 \)

(can be shown that \( \lambda < \mu \) stability condition)

* Not the case for heavy-tailed distributions, like Pareto
\( P(X > x) \sim x^{-\alpha}, \alpha > 0 \) (\( E[X] = \infty \) when \( 0 < \alpha < 1 \))
Kingman’s bound (cont’)

E.g. M/Erlang(\(\mu/N,N\))/1 queue with \(\lambda < \mu\)

\[ \phi(\theta) = \left(\frac{\mu}{\mu-N\theta}\right)^N \left(\frac{\lambda}{\lambda+\theta}\right) \]

Take \(\lambda = 2\), \(\mu = 3\), \(N=5\)

\[ P(W \geq x) \leq \exp(-\theta^*x) \]
Q2: Should one accept a new session if one wants $P(W \geq b) \leq q$?

If $\phi(-\log(q)/b) \leq 1$ then $P(W \geq b) \leq q$ as $P(W \geq b) \leq \exp(-(-\log(q)/b)b)$ by Kingman's bound

= q

Call admission: Effective bandwidth (con't)
Q2: Should one accept a new session \(i\) if one wants \(P(W \geq b) \leq q\)?

With \(n_1\) sessions of type 1, \(\ldots\), \(n_N\) sessions of type \(N\), we (easily) find

\[
\phi(\theta) = \sum_{i=1}^{N} \frac{\lambda_i n_i}{\lambda + \theta} \int_{0}^{\infty} e^{\theta x} dG_i(x)
\]

with \(\lambda = \sum_{1 \leq i \leq N} \lambda_i n_i\), \(G_i(x) = \text{cdf of service times of class-} \ n\)

See pp. 56-57 of Lecture Notes

Hence

\[
\phi(-\log(q)/b) \leq 1 \iff \sum_{1 \leq i \leq N} n_i \beta_i \leq 1
\]

with \(\beta_i := \frac{\lambda_i b (1 - \phi_i(-\log(q)/b))}{\log(q)}\)

e.b. session \(i\) and

\[
\phi_i(\theta) := \int_{0}^{\infty} e^{\theta x} dG_i(x)
\]
Q2: Should one accept new session i if one wants $P(W \geq b) \leq q$?

A: Yes, if $\beta_i + \sum_{i=1}^{N} n_i \beta_i \leq 1$ with

$$\beta_i = \frac{\lambda_i b (1 - \phi_i (-\log(q) / b))}{\log(q)}$$

Another nice “decoupling” result

Call admission

Extensions exist for **bursty traffic** (Poisson traffic does not model well traffic in Internet - to be seen later in this class)

See

and references therein