*Este texto é uma tradução livre e parcial do Capítulo 12 do livro "*Digital Telephony*" de J. C. Bellamy.

Análise de Tráfego

- **O projeto de uma rede de comunicações é elaborado assumindo-se que nem todos os usuários da rede solicitarão serviço ao mesmo tempo.**
- A quantidade exata de equipamentos a serem empregados não pode ser determinada, devido à natureza aleatória das solicitações de serviço.
- Redes de comunicações podem ser projetadas para atender instantaneamente toda a demanda, exceto quando há a ocorrência de picos muito raros ou inesperados.
- No entanto, esta solução não é econômica, porque uma grande parte do equipamento comum permanece inutilizada durante períodos normais de carga da rede.
- **O objetivo básico da Análise de Tráfego é prover métodos para determinar a relação custo-benefício de diferentes configurações e tamanhos de redes de comunicações.**

- **Tráfego em uma rede de comunicações se refere à agregação de todas as solicitações de serviços de usuários que são atendidas pela rede.**
- **As solicitações de serviço são consideradas pelas redes como aleatórias e de duração (tempo) de serviço usualmente imprevisíveis.**
- **Por esta razão, a análise de tráfego consiste primeiramente em caracterizar as solicitações de serviço e a duração do serviço solicitado sob uma abordagem probabilística.**

- **O desempenho de uma rede pode ser avaliado em termos de quanto tráfego ela transporta sob cargas normais ou médias e com que freqüência o volume de tráfego excede a capacidade da rede.**

*Este texto é uma tradução livre e parcial do Capítulo 12 do livro "*Digital Telephony*" de J. C. Bellamy.

Técnicas de Análise de Tráfego podem ser divididas em duas categorias:

1. **Sistemas que admitem Perdas (*Loss Systems*)**

2. **Sistemas que admitem Atrasos (*Delay Systems*)**

A categoria apropriada para análise de um particular sistema depende do tratamento que o sistema dá ao tráfego excedente:

- Em um *Loss System* o tráfego excedente é rejeitado sem ser servido.
- Em um *Delay System* o tráfego excedente é colocado em uma fila até que condições de serviço se tornem disponíveis.

- A medida básica de desempenho para um **sistema com perdas** é a probabilidade de rejeição, ou probabilidade de bloqueio de chamadas.
- A medida de desempenho de um **sistema que opera atrasos** é tomada em termos do atraso no serviço. Muitas vezes é considerada uma medida do atraso médio, outras vezes, uma medida de atraso que exceda a um parâmetro que especifique um determinado valor de atraso.

- Sistemas com **chaveamento por circuito** operam como sistemas com perdas, pois o excesso de tráfego é bloqueado e não servido, sem a possibilidade de nova tentativa por parte do usuário.
- Chamadas perdidas (*lost calls*) realmente representam perda de receita para os provedores de serviço.
- Redes com chaveamento por circuito incorporam certas características de atraso em adição à característica de perdas de chamadas. Por exemplo, acessos a receptores digitais, processadores de chamadas, etc, são controlados por filas.

- Sistemas do tipo ***store-and-forward*** ou de **chaveamento por pacotes** possuem as características básicas de um sistema que opera atrasos.
- Muitas vezes, no entanto, uma operação de chaveamento por pacotes pode também conter certos aspectos de um sistema com perdas: tamanho de filas limitados e circuitos virtuais podem implicar em perdas, sob circunstâncias de sobrecarga de tráfego.

*Este texto é uma tradução livre e parcial do Capítulo 12 do livro "*Digital Telephony*" de J. C. Bellamy.

4.1 Caracterização de Tráfego

* Devido à natureza aleatória do tráfego em redes, a análise de tráfego envolve fundamentos de teoria de probabilidades e processos estocásticos.
* Neste contexto, a análise de probabilidades de bloqueio é tratada em teoria de congestionamento e a análise de probabilidades de atraso é tratada em teoria de filas.
* Ainda neste contexto, estes tópicos são abordados sob o ponto de vista da análise de fluxo de tráfego.

* Em **redes comutadas por circuito**, o fluxo de dados não é tão importante quanto a duração da conexão, ou seja, o tempo que os equipamentos permanecem efetivamente ocupados.
* Uma rede comutada por circuitos estabelece um circuito *end-to-end* envolvendo vários recursos da rede (*links* de transmissão e estágios de chaveamento) que são mantidos ocupados durante a duração da chamada.
* Do ponto de vista da rede, é o tempo de ocupação destes recursos que é importante.
* No entanto, o chaveamento por circuitos envolve certos aspectos de fluxo de tráfego no processo de estabelecimento de uma conexão.
* O estabelecimento de uma conexão requer o fluxo entre fonte e destino, adquirindo, mantendo e liberando certos recursos no processo.
* O controle do fluxo durante o estabelecimento de uma conexão durante períodos de sobrecarga na rede é uma função vital do gerenciamento de redes.

* Em **redes que operam via chaveamento por mensagens ou pacotes** o interesse principal está justamente no fluxo de informação, pois nestes sistemas o tráfego nos *links* de transmissão é diretamente relacionado à atividade das fontes.

A natureza imprevisível do tráfego em comunicações decorre do resultado **de dois processos subjacentes aleatórios**:

1. a ocorrência de chamadas (*call arrivals*) e
2. o tempo de duração destas chamadas (*holding time*).

Call Arrivals

Uma chamada gerada por qualquer particular usuário é assumida como se ocorresse puramente ao acaso e é totalmente independente das chamadas de outros usuários.

Por esta razão, o nº de chamadas durante qualquer particular intervalo de tempo é indeterminado.

Holding Time

Na maior parte dos casos, os tempos de duração das chamadas também são distribuídos aleatoriamente.

Em algumas aplicações, no entanto, este elemento de aleatoriedade pode ser removido assumindo duração de chamadas constantes (por exemplo, no caso de redes de comunicações que utilizam pacotes de tamanho fixo).

A carga de tráfego apresentada a uma rede é fundamentalmente dependente tanto da **freqüência de chamadas** quanto do **tempo médio de duração de cada particular chamada**.

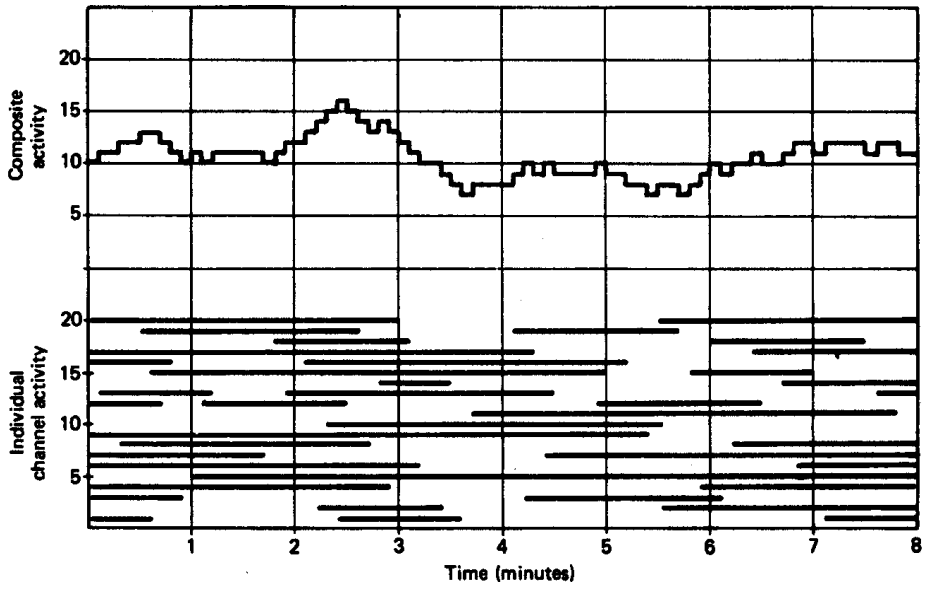


Figura 4.1: Perfil de Atividade do Tráfego da Rede
(todas chamadas sendo realizadas com sucesso).

- A Figura 4.1 descreve uma situação na qual tanto as chamadas, quanto os tempos de duração das chamadas de 20 diferentes fontes são imprevisíveis.
- A parte inferior da figura descreve a atividade de cada fonte individual, enquanto que a parte superior apresenta o total instantâneo de toda a atividade.
- Se assumirmos que as 20 fontes são conectadas a um grupo tronco, a curva de atividade apresenta o nº de circuitos em uso, a qualquer particular tempo.
- Note que o nº máximo de circuitos em uso, a qualquer instante de tempo, é 16 e a utilização média é de ≈ 11 circuitos.
- Em termos gerais, os troncos são chamados de servidores e um grupo tronco é chamado de um grupo servidor.

*Este texto é uma tradução livre e parcial do Capítulo 12 do livro "*Digital Telephony*" de J. C. Bellamy.

Medidas de Tráfego

- Uma medida de capacidade de uma rede de comunicações pode ser tomada pela determinação do volume de tráfego transportado sobre um dado intervalo de tempo.

- O volume de tráfego é essencialmente a soma de todos os tempos de duração das chamadas que são transportadas durante o intervalo.

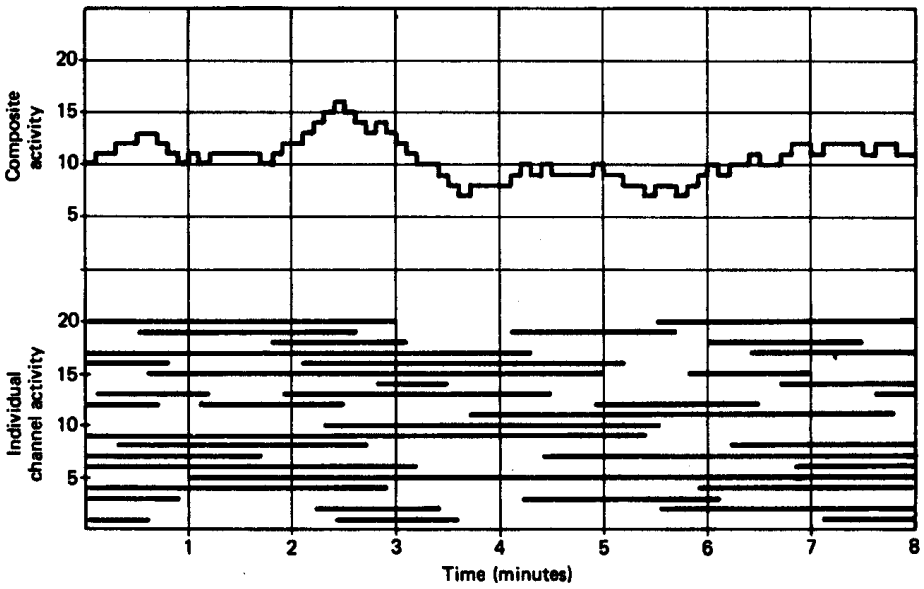


Figura 4.1: Perfil de Atividade do Tráfego da Rede (todas chamadas sendo realizadas com sucesso).

- O volume de tráfego representado na Figura 4.1, que é repetida acima, pode ser calculado através da área abaixo da curva de atividade, resultando em aproximadamente 84 minutos (de chamadas).

- Uma medida mais útil de tráfego é a intensidade de tráfego (também chamada fluxo de tráfego).
- A intensidade de tráfego é obtida dividindo o volume de tráfego pelo intervalo de tempo durante o qual ele é medido.

- Na Figura 4.1, o valor da intensidade de tráfego é 84min/8min=10.5, representando a atividade média durante o intervalo de tempo de observação.

- Embora a intensidade de tráfego seja fundamentalmente adimensional (tempo dividido por tempo) é usualmente expressa em unidades de Erlangs.
- A intensidade de tráfego também pode ser expressa em termos de centenas de segundos (de duração de chamada) por hora, ou seja, *century call seconds* (CCS) por hora.
- A relação entre Erlangs e unidades CCS pode ser derivada observando que há 3600s em 1 hora. Assim, 1 Erlang = 36 CCS, ou (3600s/100s) = 36 centenas de s.
- A capacidade máxima de um único servidor (canal) é 1 Erlang, o que equivale a dizer que o servidor está sempre ocupado.
- Desta forma, pode-se dizer que a capacidade máxima, em Erlangs, de um grupo de servidores é simplesmente igual ao nº de servidores.
- Devido ao fato de que o tráfego em um sistema com perdas sofre probabilidades de bloqueio infinitas, quando a intensidade de tráfego é igual ao nº de servidores, a atividade média é necessariamente menor do que o nº de servidores.
- Similarmente, sistemas que operam com atraso, operam a uma capacidade menor do que a capacidade total, na média, devido ao fato de que infinitos atrasos ocorrem quando a carga média se aproxima do nº de servidores.

* Dois parâmetros importantes que são usados para caracterizar o tráfego são:
  - a taxa média de chamadas, λ e
  - o tempo de duração de chamada médio (*holding time*), t_m.

* A intensidade de tráfego *A* é expressa em Erlangs, então

$$A = \lambda . t_m \qquad (4.1)$$

  onde λ e t_m são expressos em unidades de tempo (por exemplo, chamadas por segundo e segundos por chamada, respectivamente).

* Note que a intensidade de tráfego *A* é somente uma medida da utilização média durante um intervalo de tempo e não reflete o relacionamento entre chamadas e duração de chamadas.

* Ou seja, muitas chamadas curtas podem produzir a mesma intensidade de tráfego do que poucas chamadas longas.

* Na maior parte dos estudos de análise de tráfego os resultados são dependentes apenas da intensidade de tráfego.

* Em alguns casos, no entanto, os resultados são também dependentes dos padrões de chamadas individuais e das distribuições do tempo de duração das chamadas.

* Redes públicas de telefonia são tipicamente analisadas em termos da atividade média durante a hora mais ocupada do dia.

* O uso de medidas de tráfego que consideram a hora mais ocupada do dia para projetar e analisar redes telefônicas implica em conciliar duas situações bastante distintas no projeto:

  1ª: a utilização média global, a qual inclui virtualmente horas noturnas, em que não há utilização de serviço e

  2ª: os picos de curta duração, que podem ocorrer ao acaso ou como decorrência, por exemplo, de intervalos comerciais na programação de televisão, ou quando rádios promovem concursos que demandam ligações telefônicas para a emissora de rádio, etc.

* Medidas de tráfego tomadas na hora mais ocupada indicam que um telefone residencial individual está tipicamente em uso entre 5 a 10% da hora ocupada.

* Desta forma, cada telefone representa uma carga de tráfego que está entre 0.05 e 0.1 Erlangs.

* O tempo médio de duração da chamada está entre 3 e 4 min, indicando que um telefone típico está envolvido em 1 ou 2 chamadas durante a hora mais ocupada do dia.

* Telefones comerciais usualmente produzem padrões de carga diferentes dos padrões de telefones residenciais.

* Um telefone comercial possui geralmente uma carga de utilização maior.

* As horas mais ocupadas de tráfego comercial são freqüentemente diferentes da hora mais ocupada com relação ao tráfego residencial.

* A Figura 4.2 mostra uma representação gráfica da dependência típica do volume de tráfego em função da hora do dia, para ambas as fontes de tráfego (residencial e comercial).

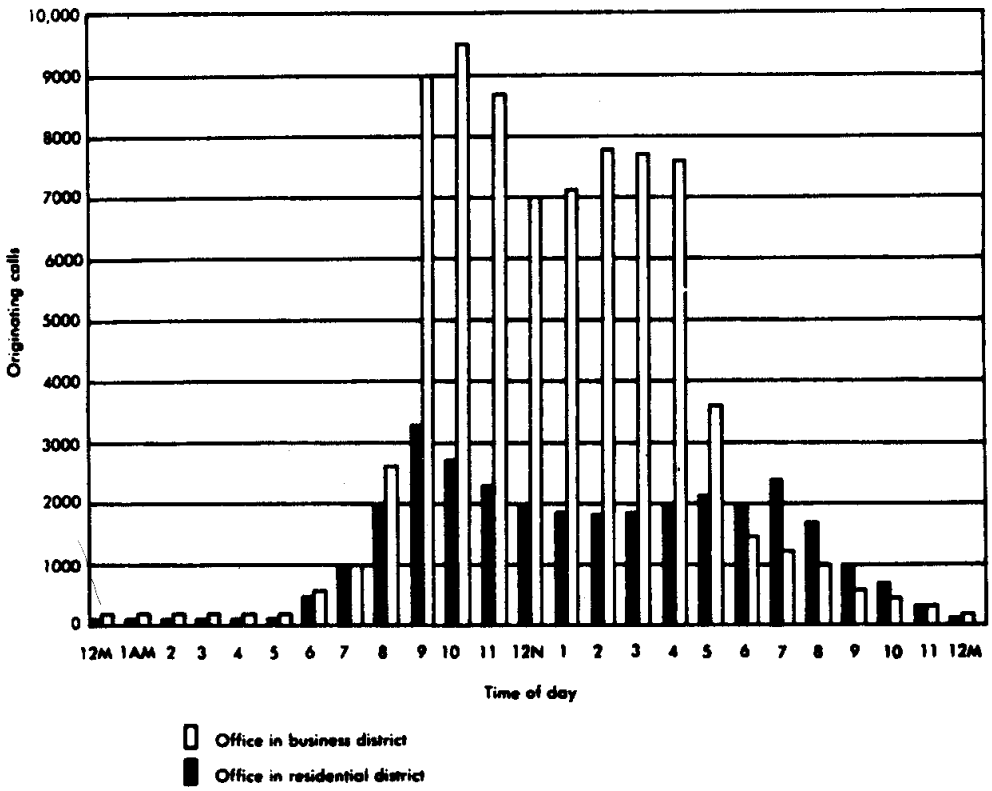


Fig. 4.2: Dependência típica do volume de tráfego em função da hora do dia.

* Esta característica diferenciada de tráfego é utilizada no projeto de troncos de uma rede telefônica, visando tirar vantagem da variação de padrões de chamadas entre diferentes centrais telefônicas.

* Troncos que conectam centrais *toll* de áreas residenciais são freqüentemente mais ocupados durante as horas do entardecer, enquanto que troncos de áreas comercias são obviamente mais ocupados durante o meio da manhã ou o meio da tarde.

* Conclui-se, então, que a engenharia de tráfego depende não apenas do volume de tráfego global, mas também dos padrões de volume de tráfego relativos a determinados intervalos de tempo dentro da rede.

Tempos de *Setup* e *Release*

- Um certo cuidado precisa ser tomado quando se está determinando a carga de tráfego total de um sistema, a partir da carga de linhas individuais ou circuitos troncos.
- Em alguns casos pode ser importante incluir tempos referentes a estabelecimento e liberação de chamadas nos tempos médios de duração de chamadas.
- Um tempo de estabelecimento de conexão de 10 s não é particularmente significativo para uma chamada de voz de duração de 4 min, mas pode dominar o tempo de duração quando se trata de uma chamada breve (curtas mensagens de dados, por exemplo).
- Tempos de *setup* se tornam mais significativos quando se está determinando a carga de tráfego total de um sistema na presença de sobrecarga de tráfego.
- Neste caso, uma porcentagem significativa da carga global vai ser representada pelas tentativas de chamada, porque o volume de tentativas é crescentemente maior do que o efetivo estabelecimento das chamadas.

Capacidade de Tráfego x Tráfego Transportado

- Uma distinção importante a ser feita, quando se está discutindo tráfego em uma rede de comunicações, é a diferença entre a capacidade de tráfego de uma rede e o tráfego que é de fato transportado (que de fato flui através da rede, decorrente do estabelecimento de chamadas, fluxo de dados, ...).

- A capacidade de tráfego de uma rede é o tráfego total que pode ser transportado pela rede, em uma condição em que a rede seja capaz de atender a todas as chamadas, à medida que elas acontecem.

- Fatores econômicos geralmente impedem que uma rede possa suportar a quantidade total de tráfego solicitada, implicando em que uma pequena porcentagem do tráfego solicitado tipicamente seja bloqueada, ou sofra atraso.

- Quando as chamadas bloqueadas são rejeitadas pela rede, o sistema é chamado *Blocked Calls Cleared* ou *Lost Calls Cleared* (LCC).

- Quando as chamadas bloqueadas são atrasadas, o sistema é chamado *Lost Calls Delayed* (LCD).

- Em sistemas ***Lost Calls Cleared***, em essência, as chamadas bloqueadas são perdidas (desaparecem e nunca retornam).
- Esta consideração é especialmente apropriada quando a rede é formada por grupos tronco, que possibilitam rotas alternativas (neste caso, uma chamada bloqueada é normalmente atendida por um outro grupo tronco e, na verdade, não retorna para o segmento de rede que a rejeitou).
- O tráfego transportado em um sistema que trabalha com perdas é sempre menor do que o tráfego solicitado da rede.

- Um sistema ***Lost Calls Delayed***, por outro lado, não rejeita chamadas bloqueadas, mas retém as chamadas até que os meios para estabelecer a comunicação estejam disponíveis.
- Considerando que a média de longo termo do tráfego solicitado é menor do que a capacidade da rede, um sistema que opera com atrasos transporta, em tese, todo o tráfego solicitado.
- Se o número de chamadas que podem ficar esperando por serviço é limitado, entretanto, um sistema que opera com atraso também apresenta algumas das propriedades de um sistema que opera com perdas.
- Por exemplo, se a fila para reter chamadas que foram bloqueadas for finita, as novas solicitações de chamadas que chegam quando a fila está cheia são descartadas.

4.1.1 Modelos de Distribuição de Chamadas (*Call Arrivals*)

Um dos dois processos subjacentes aleatórios responsáveis pela imprevisibilidade do tráfego é a natureza da ocorrência de chamadas λ (*call arrivals*).

Por simplificação, usaremos o termo "chamada" para definir: tráfego solicitado da rede; tráfego que chega à rede e solicitação de chamada.

- A consideração mais fundamental da análise de tráfego clássica é que **as solicitações de chamadas são independentes**.
- Ou seja, uma chamada gerada por uma fonte não é relacionada a uma chamada gerada por outra fonte qualquer.
- Mesmo sendo inválida em algumas situações, esta consideração é útil para a maior parte das aplicações.
- Nos casos inválidos (onde as chamadas que chegam tendem a ser correlacionadas) resultados úteis podem ainda ser obtidos a partir de modificações dos métodos que consideram as solicitações de chamadas como aleatórias.
- **O fato de assumir chamadas aleatórias provê uma formulação matemática que pode ser ajustada para produzir soluções aproximadas a problemas que seriam, de outra forma, matematicamente intratáveis.**

Intervalos entre Chamadas Modelados por Uma Distribuição Exponencial Negativa

Seja λ a taxa média de chamadas que são geradas por um grande grupo de fontes independentes (linhas de assinantes).

Assuma que:

1. Somente uma chamada pode ocorrer em qualquer intervalo de tempo suficientemente pequeno.
2. A probabilidade de ocorrer uma chamada em qualquer intervalo de tempo suficientemente pequeno é diretamente proporcional ao tamanho do intervalo (a probabilidade de uma chamada é $\lambda\Delta t$, onde Δt é o tamanho do intervalo).
3. A probabilidade de uma chegada ocorrer em qualquer particular intervalo é independente do que ocorreu em outros intervalos (de chamadas que possam ter ocorrido em outros intervalos).

Pode-se mostrar que a distribuição de probabilidade dos intervalos entre chamadas é

$$P_0(\lambda t) = e^{-\lambda t}. \qquad (4.2)$$

A Equação 4.2 define a probabilidade de que nenhuma chamada ocorra em um intervalo de tempo *t* aleatoriamente selecionado.

O que equivale à probabilidade de que *t* segundos transcorram entre uma chamada e a outra.

Exemplo 4.1:

Assumindo que cada uma das 10000 linhas de assinantes origine uma chamada por hora, com qual freqüência duas chamadas chegarão com menos do que 0.01 segundos entre elas?

Solução:

A taxa média de chamadas é

$$\lambda = 10000/3600 = 2.78 \quad \text{chamadas/s}.$$

A partir da Equação 4.2, a probabilidade de que não haja chamadas em um intervalo de 0.01 segundos é dada por

$$P_0(\lambda t = 2.78 \times 0.01 = 0.0278) = e^{-\lambda t} = e^{-0.0278} = 0.973.$$

Então, 1 - 0.973 = 0.027, ou 2.7% das chamadas ocorrem em um intervalo de tempo menor do que 0.01 segundos a partir da chamada anterior.

Como a taxa de chamadas é 2.78 chamadas/segundo, a taxa de ocorrência de intervalos entre chamadas menores do que 0.01 segundos é

$$2.78 \times 0.027 = 0.075 \text{ vezes/segundo}.$$

Para derivar a distribuição de intervalos entre chamadas baseados na exponencial negativa foram feitas 3 considerações:

1. Somente uma chamada pode ocorrer em qualquer intervalo de tempo suficientemente pequeno.

2. A probabilidade de ocorrer uma chamada em qualquer intervalo de tempo suficientemente pequeno é diretamente proporcional ao tamanho do intervalo (a probabilidade de uma chamada é $\lambda\Delta t$, onde Δt é o tamanho do intervalo).

3. A probabilidade de uma chegada em qualquer particular intervalo é independente do que ocorreu em outros intervalos.

As primeiras duas considerações podem ser intuitivamente justificadas, para a maior parte das aplicações.

A **terceira**, no entanto, implica em certos aspectos do comportamento das fontes que nem sempre podem ser considerados como certos:

- Certos eventos, tais como intervalos comerciais em programação de televisão, podem estimular as fontes a fazerem suas ligações ao mesmo tempo.

  Neste caso, a distribuição baseada na exponencial negativa ainda pode ser adequada (no entanto, para uma taxa de chamadas muito superior, durante o intervalo comercial).

- Uma outra implicação de se assumir chamadas independentes envolve o nº de fontes e não apenas os seus padrões de chamadas:

  - Se o nº de fontes disponíveis para gerar solicitações de chamadas for **finito** e constante e
  - a probabilidade de uma chamada em qualquer intervalo de tempo pequeno for independente de outras chamadas, o que é assumido em (3),
  - poderá ocorrer que algumas das fontes tenham se tornado ocupadas em um intervalo de tempo imediatamente anterior ao considerado e
  - não possam gerar chamadas no intervalo de tempo em questão,
  - o que reduzirá a taxa média de solicitação de chamadas.
  - A redução da taxa média de solicitação de chamadas fará com que os intervalos de tempo entre chamadas sejam sempre maiores do que os intervalos de tempo expressos na Equação 4.2 ($P_0(\lambda t) = e^{-\lambda t}$).

- A única vez que a taxa de chamadas é realmente independente da atividade da fonte é quando existe um nº **infinito** de fontes.

- Se o nº de fontes é grande e sua atividade média é relativamente baixa, fontes ocupadas não reduzem de forma apreciável a taxa de chamadas.

- Por exemplo, considere uma central local que serve a 10000 assinantes com 0.1 Erlang de atividade cada um.

- Normalmente há 1000 *links* ativos e 9000 assinantes disponíveis para gerar novas chamadas.

- Se o nº de assinantes ativos aumenta por um fator improvável de 50%, passando a 1500 linhas ativas, o nº de assinantes inativos é reduzido para 8500, uma mudança de apenas 5.6%.

- O que demonstra que a taxa de chamadas é relativamente constante sobre uma ampla gama de diferentes atividades de fonte. Caso em que considerar o nº de fontes como infinito é justificado.

Distribuição de Chamadas pelo Modelo de Poisson

A equação $P_0(\lambda t) = e^{-\lambda t}$ provê uma forma de determinar a distribuição de tempos entre chamadas, mas não permite determinar quantas chamadas podem ocorrer em algum intervalo de tempo arbitrário.

Usando as mesmas considerações apresentadas, entretanto, **a probabilidade de *j* chamadas ocorrerem em um intervalo de tempo *t*** pode ser determinada através da lei de probabilidades de Poisson, expressa na Equação 4.3.

$$P_j(\lambda t) = \frac{(\lambda t)^j}{j!} e^{-\lambda t} \quad . \tag{4.3}$$

Quando $j = 0$, a probabilidade de não haverem chamadas em um intervalo de tempo *t* é $P_0(t)$, conforme Equação 4.2.

* A Equação 4.3 assume chamadas que são independentes e ocorrem a uma dada taxa média λ.

* A distribuição de probabilidades de Poisson é utilizada para chamadas provenientes de um grande número de fontes independentes.

* A Equação 4.3 define a probabilidade de acontecerem exatamente *j* chamadas em *t* segundos.

A probabilidade de ocorrerem *j* **ou mais** chamadas em *t* segundos é dada pela Equação 4.4:

$$P_{\geq j}(\lambda t) = \sum_{i=j}^{\infty} P_i(\lambda t) = 1 - \sum_{i=0}^{j-1} P_i(\lambda t) = 1 - P_{<j}(\lambda t) \tag{4.4}$$

onde $P_i(\lambda t)$ é definida pela Equação 4.3.

Exemplo 4.2:

Em um *link* que opera sob chaveamento por mensagem ocorrem 4 chamadas por minuto. Qual é a probabilidade de que 8 ou mais chamadas ocorram em um intervalo arbitrário de 30 segundos?

Solução:

O número médio de chamadas em um intervalo de 30 segundos é dado por

$$\lambda t = 4 \times \frac{30}{60} = 2 .$$

A probabilidade de 8 ou mais chamadas (quando a média é 2) é:

$$P_{\geq 8}(2) = \sum_{i=8}^{\infty} P_i(2) = 1 - \sum_{i=0}^{7} P_i(2) =$$

$$= 1 - e^{-2}\left(1 + \frac{2^1}{1!} + \frac{2^2}{2!} + \frac{2^3}{3!} + \ldots + \frac{2^7}{7!}\right) = 0.0011$$

.

Exemplo 4.3:

Qual é a probabilidade de que um bloco de dados formado por 1000 bits sofra exatamente 4 erros enquanto ele está sendo transmitido sobre um *link* de transmissão com uma taxa de erros de bits (BER) de 10^{-5} ?

Solução:

Assumindo erros independentes (esta consideração é questionável em muitos *links* de transmissão), pode-se obter a probabilidade de que ocorram exatamente 4 erros, a partir da distribuição de Poisson.

O nº médio de erros (de chamadas) é dado por $\lambda t = 10^3 \times 10^{-5} = 0.01$. Assim,

$$\text{prob}\,(4\ \text{erros}) = P_4(0.01) = \frac{(0.01)^4}{4!} e^{-0.01} = 4.125 \times 10^{-10}\ .$$

4.1.2 Modelos de Distribuição de Tempos de Duração de Chamada (*Holding Time*)

O segundo processo aleatório subjacente responsável pela imprevisibilidade do tráfego é o tempo médio de duração de chamada t_m (*holding time*).

* Para determinar as probabilidades de bloqueio em um sistema que opera com perdas, ou atrasos em um sistema que opera com atrasos, em alguns casos é suficiente conhecer o tempo médio de duração de chamadas.

* Em outros casos, é necessário conhecer a distribuição de probabilidade dos tempos de duração de chamadas para obter os resultados desejados.

* As duas mais comuns distribuições de probabilidade assumidas para o tempo de duração de chamadas são:

  (1) a distribuição conhecida por **tempo de duração de chamadas constante** e

  (2) a distribuição conhecida por **tempo de duração de chamadas exponencial**.

Tempo de Duração de Chamadas Constante

- Embora os tempos de duração de chamadas não possam ser assumidos constantes para conversações de voz convencionais, eles constituem uma consideração razoável para atividades como: requerimentos de processamento de chamadas, sinalização de endereçamento entre centrais e *playback* de mensagens gravadas.
- Além disso, tempos de duração de chamada constantes são obviamente válidos para tempos de transmissão de redes que trabalham com pacotes de tamanho fixo.

Quando mensagens com tempos de duração de chamadas constantes ocorrem, pode-se utilizar a Equação $P_j(\lambda t) = \frac{(\lambda t)^j}{j!} e^{-\lambda t}$ para determinar a distribuição de probabilidade de canais ativos.

Assumindo que todas as solicitações sejam atendidas,

⇨ a probabilidade de *j* canais estarem ocupados em qualquer particular instante de tempo é =

⇨ à probabilidade de que *j* chamadas tenham ocorrido no intervalo de tempo de tamanho t_m imediatamente precedente ao instante de tempo em questão.

Como o nº médio de circuitos ativos sobre todo o intervalo de tempo é dado pela intensidade de tráfego $A = \lambda t_m$, a probabilidade de *j* circuitos estarem ocupados é dependente somente da intensidade de tráfego. Assim,

$$P_j(\lambda t_m) = P_j(A) = \frac{A^j}{j!} e^{-A} \qquad (4.5)$$

onde:

λ = taxa de chamadas,

t_m = tempo de duração de chamadas constante e

A = intensidade de tráfego em Erlangs.

Tempo de Duração de Chamadas Exponencial

O tipo mais comum de distribuição de duração de chamadas utilizado para conversações telefônicas convencionais é a distribuição de duração de chamadas exponencial.

$$P(>t) = e^{-t/t_m}, \tag{4.6}$$

onde t_m é o tempo médio de duração de chamadas.

* A Equação 4.6 especifica a probabilidade de que uma duração de chamada exceda um valor *t.*
* Esta relação é derivada a partir da natureza do processo de terminação de uma chamada.
* As conversações reais de voz exibem uma correspondência muito próxima a uma distribuição exponencial.
* A distribuição exponencial possui a propriedade de que a probabilidade de uma terminação é independente de o quão longa uma chamada em andamento tenha sido.
* **Ou seja, não interessa o quão longa foi uma chamada, a probabilidade de que ela dure outros *t* segundos é definida pela Equação 4.6.**
* Tempos de duração de chamadas exponenciais representam o processo mais aleatório possível.
* Nem sempre o conhecimento de o quão longa foi a duração de uma chamada em andamento provê qualquer informação sobre quando uma chamada irá terminar.

- Combinar um processo de chamadas de Poisson com um processo de duração de chamadas exponencial para obter a probabilidade da distribuição de circuitos ativos é mais complicado do que no caso em que se considerava a distribuição de tempo de duração de chamadas constante, porque as chamadas podem durar indefinidamente.
- O resultado final, entretanto, prova ser dependente apenas do tempo de duração de chamada médio.
- Assim, a Equação 4.5 é válida tanto para tempos de duração de chamada exponenciais, quanto para tempos de duração de chamadas constantes (ou qualquer distribuição de tempos de duração de chamadas).
- Desta forma:
- A **probabilidade de *j* circuitos estarem ocupados em qualquer particular instante de tempo** (assumindo um processo de chamadas de Poisson e que todas as solicitações de chamadas são atendidas imediatamente) é dada por

$$P_j(A) = \frac{A^j}{j!} e^{-A} \qquad (4.7)$$

onde *A* é a intensidade de tráfego em Erlangs.

- Este resultado é verdadeiro para qualquer distribuição de tempos de duração de chamadas.

Exemplo 4.4:

Assuma que um grupo tronco tenha um número de canais suficiente para imediatamente transportar todo o tráfego oferecido por um processo de Poisson, com uma taxa de chamadas de 1 chamada por minuto.

Assuma que o tempo médio de duração de chamada é igual a 2 minutos.

Qual a porcentagem do tráfego total que é transportado pelos primeiros 5 circuitos, e quanto do tráfego é transportado pelos circuitos remanescentes?

(Assuma que o tráfego é sempre transportado sobre os circuitos de numeração inferior.)

Solução:

A intensidade de tráfego do sistema é *A*=1x2=2 Erlangs.

A intensidade de tráfego transportada pelos *i* circuitos ativos é exatamente *i* Erlangs.

A probabilidade de *i* circuitos estarem ocupados em qualquer particular instante de tempo é dada por $P_i(A) = \frac{A^i}{i!} e^{-A}$

Assim, o tráfego transportado pelos 5 primeiros circuitos pode ser determinado como segue:

$$A_5 = 1P_1(2) + 2P_2(2) + 3P_3(2) + 4P_4(2) + 5P_5(2) =$$

$$= e^{-2}\left(2 + \frac{2\times 2^2}{2!} + \frac{3\times 2^3}{3!} + \frac{4\times 2^4}{4!} + \frac{5\times 2^5}{5!}\right) = 1.89 \text{ Erlangs}$$

De forma que todos os circuitos remanescentes transportam $(2 - 1.89) = 0.11$ Erlangs.

Os primeiros 5 circuitos transportam 94.5% do tráfego, enquanto todos os circuitos remanescentes transportam apenas 5.5% do tráfego.

Se há 100 fontes, 95 circuitos extras são necessários para transportar os 5.5%.

Exemplo extraído da dissertação: "Sistematização de Procedimentos para Projeto e Implementação de Sistemas VoIP em Redes LAN/WAN Utilizando a Recomendação ITU-T H.323", de Emerson Baumgarten de Oliveira.

7.1.2 ***Passo 2*** *- Definir a demanda telefônica e o interesse de tráfego (Erlang).*

Modificar toda uma estrutura de uma empresa no que tange ao uso de um sistema VoIP, além de envolver conceitos de redes de dados, deve também levar em consideração a teoria definida pelo matemático dinamarquês A. K. Erlang.

Porém, é um tanto difícil isolar os fatores que determinem o comportamento do usuário de um sistema de telefonia, seja ele convencional ou VoIP. Ao longo do dia é possível que um determinado sistema passe por períodos de ociosidade e de congestionamento, logo a saída é usar uma abordagem probabilística para viabilizar os estudos e simulações necessárias.

Inicialmente pode parecer simples o dimensionamento da quantidade de canais que irá interligar dois pontos. Contudo, como os recursos devem sempre ser otimizados ao máximo, fica difícil alocar um número de canais capaz de atender completamente a demanda, ainda mais quando a tendência é essa demanda crescer no tempo e superar essa capacidade. Além disso, conforme observado no início, existem inúmeros fatores imprevisíveis que podem provocar um pico inesperado de demanda a ponto de ultrapassar a atual capacidade de atendimento, criando um estado de congestionamento.
Logo, um projeto de telefonia deve definir um número de canais que garanta que a probabilidade de haver um excesso de demanda, ou congestionamento, não seja maior do que um valor considerado razoável.

Consequentemente, o projeto de um sistema telefônico envolverá três variáveis:

- número de canais que estarão à disposição dos usuários;
- a demanda do sistema, ou seja, o volume de tempo do total das ligações solicitadas em uma hora;
- congestionamento provável da central, ou seja, o provável percentual de chamadas que encontrarão o sistema ocupado.

Para compreender melhor a unidade de medida de tráfego telefônico, será feito um exemplo:

Um sistema telefônico com 10 linhas, onde cada linha recebe, em média, 2 chamadas/hora e essas tem duração média de 3 minutos. Qual a sua demanda?

Como chegam ao sistema 10 × 2 = 20 chamadas por hora, que ocupam 20 × 3 = 60 minutos = 1 hora, logo a demanda é de 1 hora por hora, ou seja, 1 Erlang.

Voltando aos conceitos de Erlang, este constatou que as chamadas poderiam ser aproximadas por uma distribuição de probabilidades do tipo de Poisson. A partir desse resultado, ele conseguiu relacionar três variáveis básicas: L, c e d, podendo ser resumido pela seguinte fórmula:

$$c(L,d) = \frac{\frac{d^L}{L!}}{1 + \frac{d}{1!} + \frac{d^2}{2!} + \frac{d^3}{3!} + \frac{d^4}{4!} + \ldots + \frac{d^L}{L!}}$$

onde:

L = número de canais

d = demanda do sistema

c = congestionamento provável

Assim o congestionamento do sistema é função do número de canais e da demanda do sistema. Aplicando a fórmula acima, no exemplo abaixo, obtêm-se:

Um sistema com L = 15 canais e demanda d = 10 Erlangs, terá um congestionamento c = c (15, 10) = 0.036, ou seja, 3,6% das chamadas receberão o sinal de ocupado.

Posto isto, o projeto de um sistema telefônico, e consequentemente, de um sistema VoIP também, estarão resolvidos se for expresso o número L de canais em termos da demanda d a ser atendida e do congestionamento provável c que se esteja disposto a aceitar. Logo, o problema básico da telefonia é achar a função L = L (c, d).

Como a fórmula de Erlang dá uma relação do tipo c = c (L, d), enquanto que o problema do projeto consiste em achar o número L de canais em função do grau de congestionamento aceitável e da demanda envolvida, existem duas alternativas:

- Tratar c = c (L, d) como uma equação na incógnita L;
- Utilizar uma tabela de valores L para uma grande quantidade de possibilidades de d e c.

Como exemplo, será utilizado o caso a seguir: Calcular o número de canais L necessários para um sistema telefônico capaz de atender uma demanda de 55 Erlangs com congestionamento de 1%.

Usando a primeira alternativa, a equação c (L, 55) = 0,01 terá de ser resolvida por processo iterativo. Para isso é utilizado, de acordo com cálculos via a fórmula de Erlang, para d < 75 Erlangs e com c = 0.01 os L são dados aproximadamente por 6 + d/0.85. Consequentemente, para d = 55, tem-se que L ≈ 70. A partir daí, obtém-se uma seqüência de intervalos envolvendo o valor de L e que vão diminuindo sucessivamente:

c (70, 55) = 0.0074

c (65, 55) = 0.0228

c (67, 55) = 0.0151

c (69, 55) = 0.0095

c (68, 55) = 0.0128

Como L deve ser um número inteiro, o número de canais será 69.

Usando a segunda alternativa, a obtenção da resposta fica mais fácil, pois, baseado no valor de x% de congestionamento e no valor da demanda em Erlangs, utilizam-se tabelas que auxiliam neste cálculo, como é demonstrado na Tabela 15 abaixo [SI75].

Canais	1%	1,2%	1,5%	2%	3%	5%	7%	10%	15%	20%	30%	40%	50%
58	45,1	45,8	46,6	47,8	49,6	52,6	55	58,2	63,3	68,4	79,8	94,3	114,1
59	46	46,7	47,5	48,7	50,6	53,6	56	59,3	64,5	69,7	81,3	96	116,1
60	46,9	47,6	48,4	49,6	51,6	54,6	57,1	60,4	65,6	70,9	82,7	97,6	118,1
61	47,9	48,5	49,4	50,6	52,5	55,6	58,1	61,5	66,8	72,1	84,1	99,3	120,1
62	48,8	49,4	50,3	51,5	53,5	56,6	59,1	62,6	68	73,4	85,5	101	122,1
63	49,7	50,4	51,2	52,5	54,5	57,6	60,2	63,7	69,1	74,6	87	102,6	124,1
64	50,6	51,3	52,2	53,4	55,4	58,6	61,2	64,8	70,3	75,9	88,4	104,3	126,1
65	51,5	52,2	53,1	54,4	56,4	59,6	62,3	65,8	71,4	77,1	89,8	106	128,1
66	52,4	53,1	54	55,3	57,4	60,6	63,3	66,9	72,6	78,3	91,2	107,6	130,1
67	53,4	54,1	55	56,3	58,4	61,6	64,4	68	73,8	79,6	92,7	109,3	132,1
68	54,3	55	55,9	57,2	59,3	62,6	65,4	69,1	74,9	80,8	94,1	111	134,1
69	**55,2**	55,9	56,9	58,2	60,3	63,7	66,4	70,2	76,1	82,1	95,5	112,6	136,1
70	56,1	56,8	57,8	59,1	61,3	64,7	67,5	71,3	77,3	83,3	96,9	114,3	138,1
71	57	57,8	58,7	60,1	62,3	65,7	68,5	72,4	78,4	84,6	98,4	115,9	140,1
72	58	58,7	59,7	61	63,2	66,7	69,6	73,5	79,6	85,8	99,8	117,6	142,1
73	58,9	59,6	60,6	62	64,2	67,7	70,6	74,6	80,8	87	101,2	119,3	144,1
74	59,8	60,6	61,6	62,9	65,2	68,7	71,7	75,6	81,9	88,3	102,7	120,9	146,1
75	60,7	61,5	62,5	63,9	66,2	69,7	72,7	76,7	83,1	89,5	104,1	122,6	148
76	61,7	62,4	63,4	64,9	67,2	70,8	73,8	77,8	84,2	90,8	105,5	124,3	150
77	62,6	63,4	64,4	65,8	68,1	71,8	74,8	78,9	85,4	92	106,9	125,9	152
78	63,5	64,3	65,3	66,8	69,1	72,8	75,9	80	86,6	93,3	108,4	127,6	154
79	64,4	65,2	66,3	67,7	70,1	73,8	76,9	81,1	87,7	94,5	109,8	129,3	156
80	65,4	66,2	67,2	68,7	71,1	74,8	78	82,2	88,9	95,7	111,2	130,9	158

Tabela 15 - Valores para cálculo de tráfego telefônico.

4.2 Loss Systems

* O Exemplo 4.4 provê uma indicação das probabilidades de bloqueio que surgem quando o nº de servidores (circuitos) é menor do que a carga máxima de tráfego possível (nº de fontes).
* No exemplo, 94.5% do tráfego é transportado por apenas 5 circuitos. A implicação deste fato é que a probabilidade de bloqueio, se apenas 5 circuitos estiverem disponíveis, será de 5.5%.
* Há uma pequena, mas importante, diferença entre a probabilidade de 6 ou mais circuitos estarem ocupados (como pode se obtido pela Equação 4.7) e a probabilidade de bloqueio que surge quando existem apenas cinco circuitos.
* A diferença básica para a discrepância é indicada na Figura 4.3, a qual descreve o mesmo padrão de tráfego gerado por 20 fontes, mostrado na Figura 4.1.

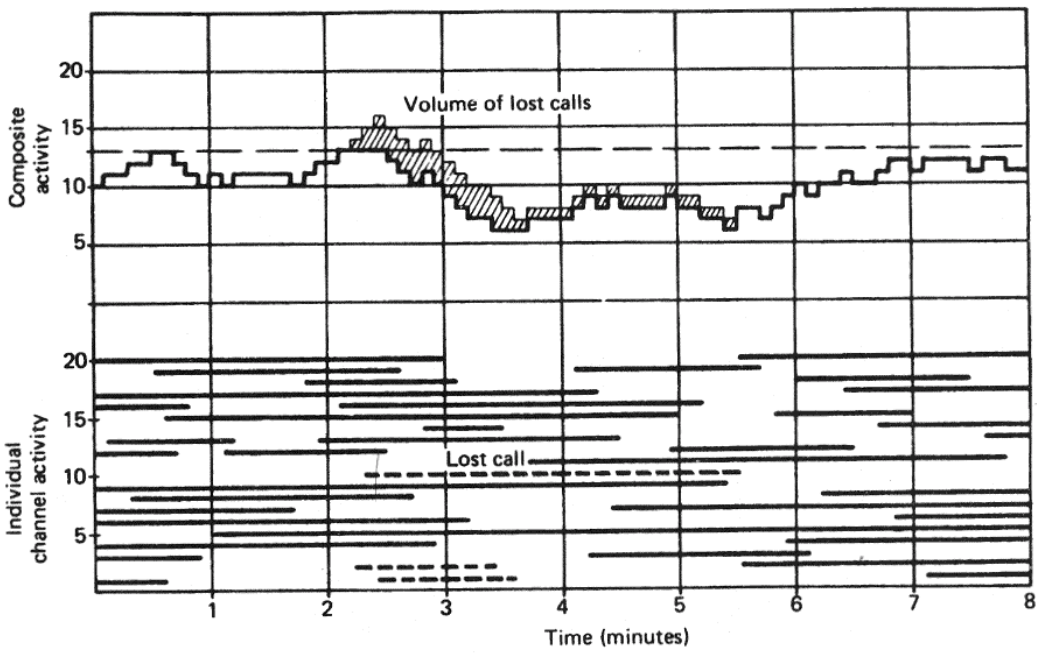


Figura 4.3: Perfil de Atividade *blocked calls cleared* (13 canais).

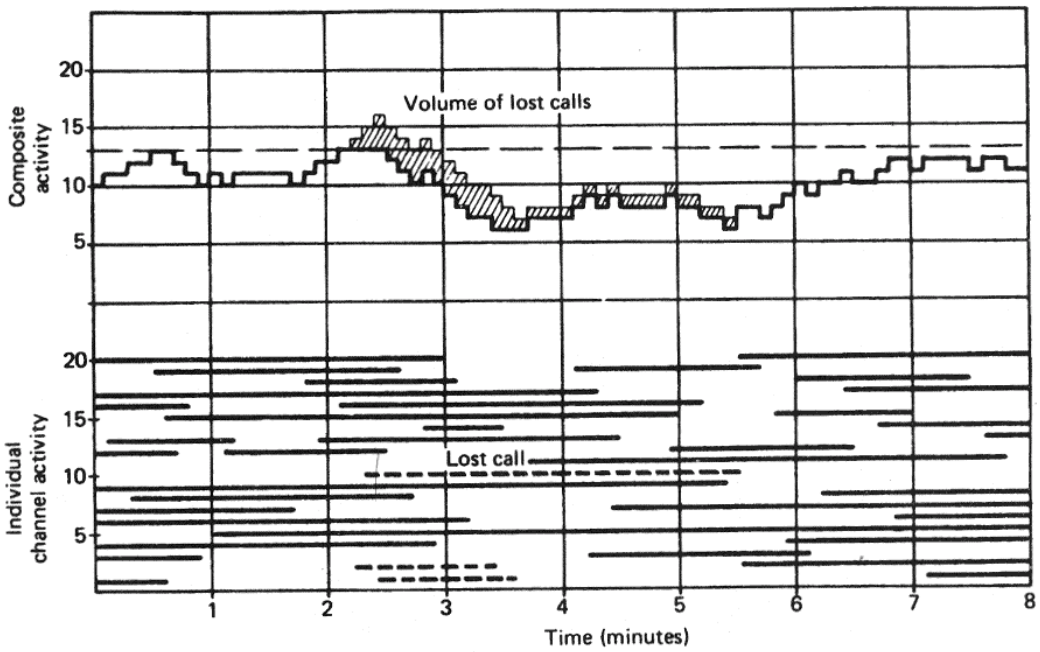


Figura 4.3: Perfil de Atividade *blocked calls cleared* (13 canais).

* Na Figura 4.3, no entanto, é assumido que apenas 13 circuitos estão disponíveis para transportar o tráfego. Assim, as três chamadas, em $t = 2.2,\ 2.3 \text{ and } 2.4 \min$ são bloqueadas e é assumido que tenham deixado o sistema.
* A quantidade total de volume de tráfego perdido é indicada pela área sombreada, a qual é a diferença entre todo o tráfego que é atendido à medida que é solicitado e o tráfego transportado por um sistema que opera sob perda (*blocked calls cleared system*), com 13 circuitos.
* A mais importante característica a ser notada na Figura 4.3 é que a chamada que chega em $t = 2.8$ não é bloqueada, mesmo quando o perfil original indica que ela chega quando todos os 13 circuitos estão ocupados.
* A razão pela qual ela não é bloqueada é que as chamadas previamente bloqueadas deixaram o sistema e portanto reduziram o congestionamento para chamadas subseqüentes.
* Assim, a porcentagem de tempo que o perfil do tráfego original é igual ou superior a 13 não é equivalente à probabilidade de bloqueio quando apenas 13 circuitos estão disponíveis.

Chamadas Bloqueadas Descartadas
Lost Calls Cleared (LCC)

* Este tipo de sistema não oferece a possibilidade de espera em uma fila para as chamadas solicitadas que não são completadas.

* Ou seja, para cada usuário que solicita um serviço, é assumido que não há um tempo de *setup* e que o usuário recebe imediatamente o acesso ao canal (desde que haja, pelo menos, um canal disponível).

* Se não há canais disponíveis, o usuário solicitante é bloqueado e fica sem acesso ao sistema, tendo que tentar novamente mais tarde.

* Este tipo de modelo foi considerado pela primeira vez em 1917, pelo matemático dinamarquês Erlang e é denominado Chamadas Bloqueadas Descartadas (*Blocked Calls Cleared*).

* A formulação da probabilidade de bloqueio para um sistema que descarta as chamadas não atendidas assume que as chamadas obedecem a uma distribuição de Poisson.

* Um processo de Poisson é um processo em que as ocorrências são variáveis aleatórias independentes. Este tipo de processo implica na consideração de fontes infinitas.

* Um sistema *Lost Calls Cleared* conduz à derivação da fórmula de Erlang de primeiro tipo,

* ou fórmula Erlang B,

* (também conhecida como a fórmula *Lost Calls Cleared* ou *Blocked Calls Cleared*).

* A fórmula pode ser expressa por $E_{1,N}(A)$, conforme

$$B = E_{1,N}(A) = \frac{A^N}{N!\sum_{i=0}^{N}\left(\frac{A^i}{i!}\right)} \qquad (4.8)$$

onde *N* é o número de canais (servidores) e *A* é a intensidade de tráfego total oferecida, dada por λt_m (Erlangs).

* Um aspecto fundamental da formulação de Erlang, que é considerado chave para a teoria moderna de processos estocásticos é o conceito de equilíbrio estatístico.

* Basicamente, o equilíbrio estatístico implica em que a probabilidade de que um sistema esteja em um estado particular (nº de circuitos ocupados em um grupo tronco) é independente do instante de tempo em que o sistema é examinado.

* Para que um sistema esteja em equilíbrio estatístico, um longo tempo deve passar (vários tempos médios de duração de chamadas) entre o instante de tempo em que o sistema está em um estado conhecido, até que seja novamente examinado.

* Por exemplo, quando é iniciada a transmissão em um grupo tronco, não há circuitos ocupados.
* Por um curto intervalo de tempo o sistema terá, provavelmente, apenas poucos circuitos ocupados.
* À medida que o tempo passa, entretanto, o sistema atinge o equilíbrio.
* Neste ponto, o estado mais provável do sistema é que ele tenha $A = \lambda t_m$ circuitos ocupados.

* Estando no estado de equilíbrio, é tão provável para o sistema receber uma nova solicitação de chamada quanto uma terminação de chamada. Assim,
* Se o nº de circuitos ativos > média $A \rightarrow$ uma terminação é mais provável do que uma solicitação de chamada.
* Se o nº de circuitos ativos < média $A \rightarrow$ uma solicitação de chamada é mais provável do que uma terminação.

➔ A Equação 4.8 especifica a probabilidade de bloqueio para um sistema:

– com solicitações de chamadas aleatórias originadas por um número infinito de fontes e

– com distribuições arbitrárias de tempo de duração de chamadas.

$$B = E_{1,N}(A) = \frac{A^N}{N!\sum_{i=0}^{N}\left(\frac{A^i}{i!}\right)} \qquad (4.8)$$

N = nº de canais (servidores)

A = intensidade de tráfego total, dada por λt_m (Erlangs).

* A probabilidade de bloqueio resultante da Equação 4.8 é plotada na Figura 4.4 como função da intensidade de tráfego solicitada para vários números de canais.

*Este texto é uma tradução livre e parcial do Capítulo 12 do livro "*Digital Telephony*" de J. C. Bellamy.

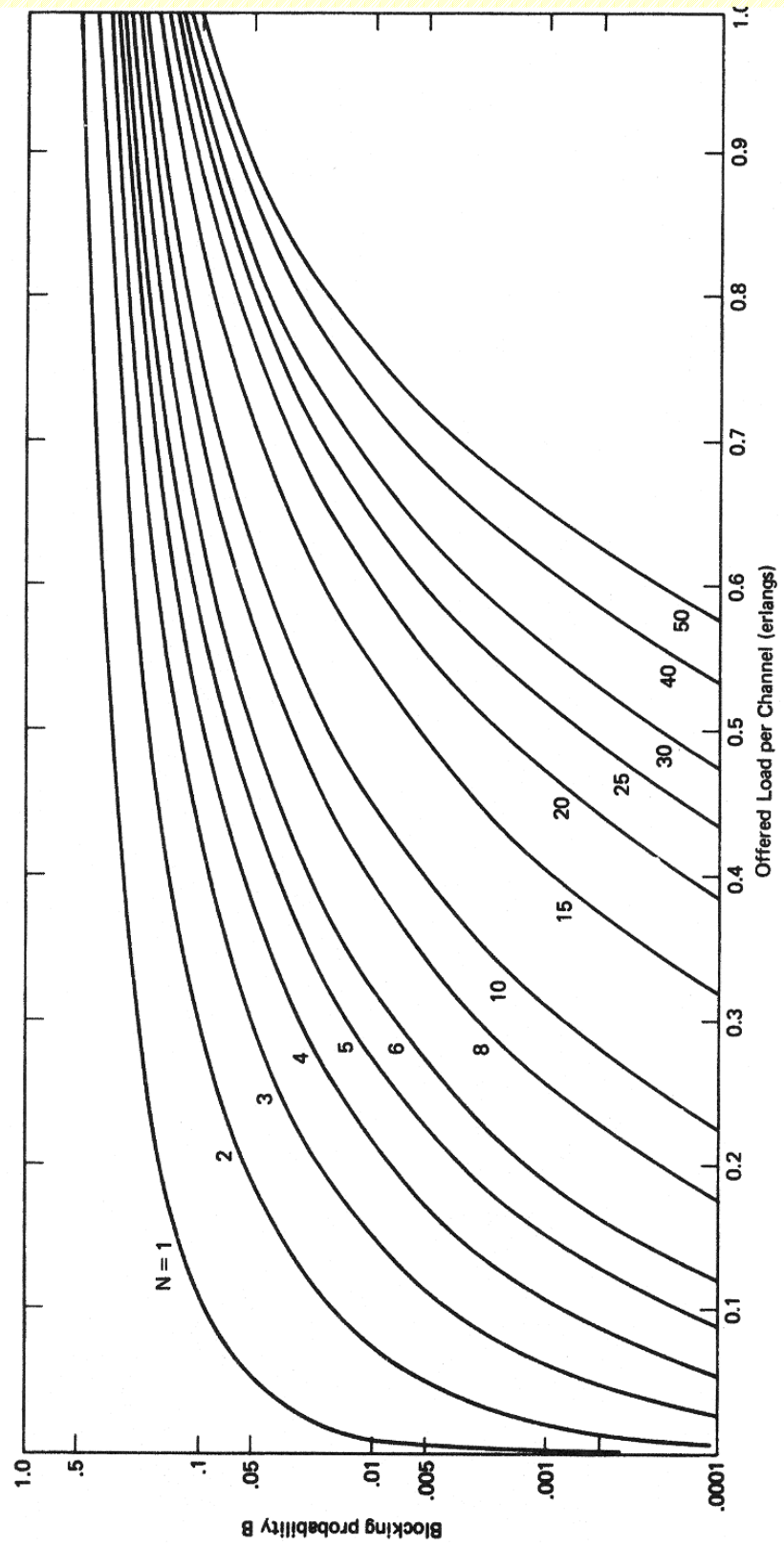


Figura 4.4: Probabilidade de bloqueio de sistemas LCC.

➔ Uma representação gráfica mais útil do resultado de Erlang é dada na Figura 4.5, a qual apresenta a utilização de canal para várias probabilidades de bloqueio e número de servidores (canais).

* A utilização de canal ρ representa o tráfego transportado por cada circuito.

$$\rho = \frac{(1-B)A}{N}, \quad (4.9)$$

onde: A = intensidade de tráfego total solicitada,

N = nº de canais (servidores),

B = probabilidade de bloqueio e

$(1-B)A$ = tráfego transportado.

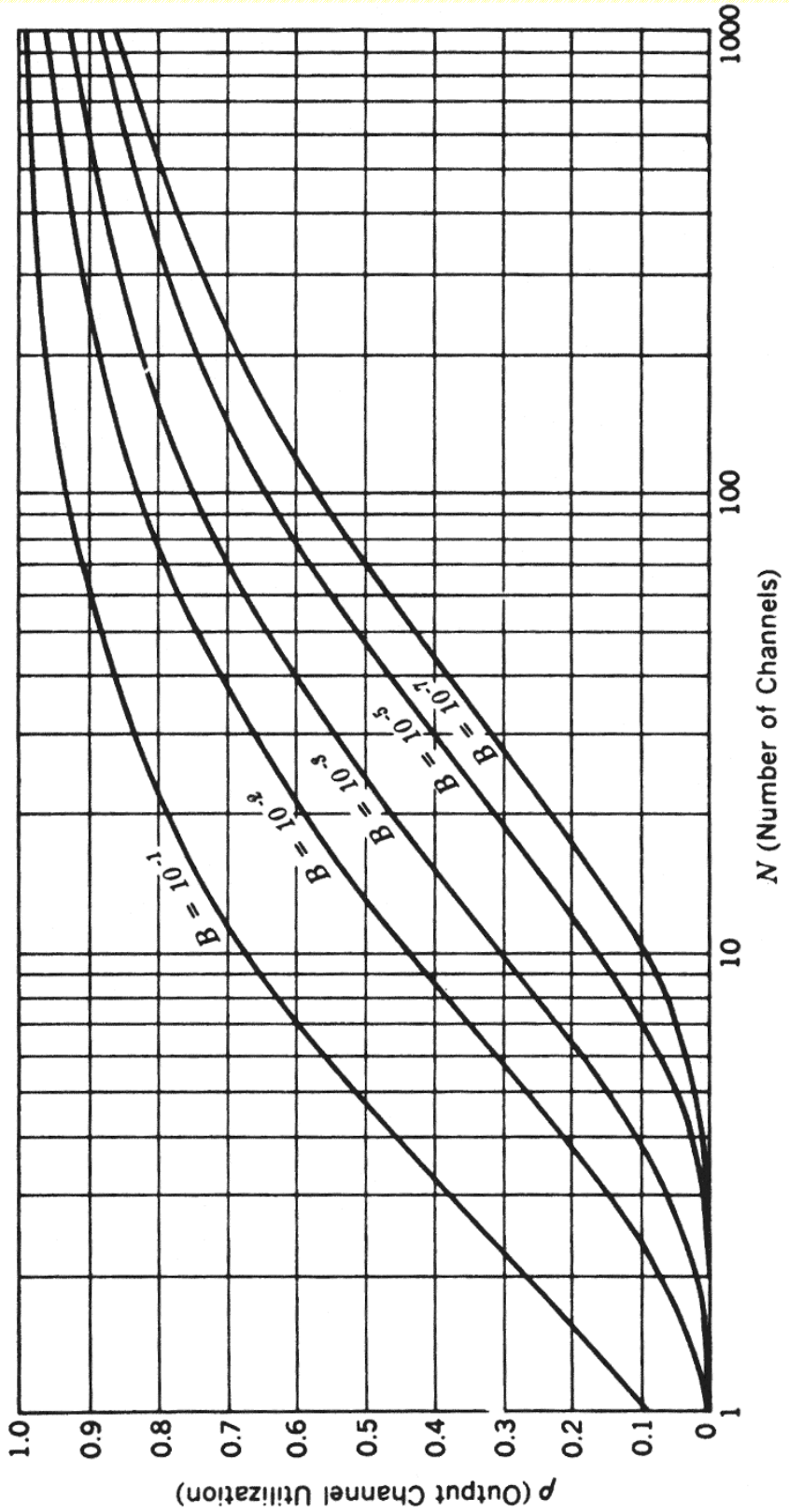


Figura 4.5: Utilização de canais para sistemas LCC.

Exemplo 4.5:

Uma linha T1 será usada como grupo tronco entre dois sistemas PBX.

Quanto tráfego o grupo tronco pode transportar para uma probabilidade de bloqueio de 0.1?

Qual é a intensidade de tráfego solicitada?

Solução:

- A partir do gráfico mostrado na Figura 4.5, a utilização do circuito de saída para $B = 0.1$ e $N = 24$ (T1) é $\rho = 0.8$.

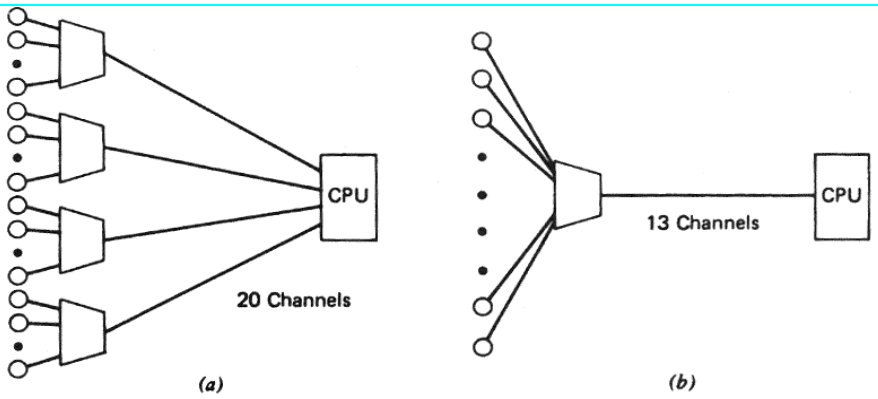
- Como $\rho = \dfrac{(1-B)A}{N} \Rightarrow (1-B)A = \rho\, N$ e, assim, a intensidade de tráfego que o grupo pode transportar $(1-B)A$ será igual a $\rho\, N = (0.8 \times 24) = 19.2$ Erlangs.

- Sendo a probabilidade de bloqueio $B = 0.1$, o nível máximo de tráfego solicitado será

$$A = \frac{\rho\, N}{(1-B)} = \frac{0.8 \times 24}{(1-0.1)} = 21.3 \text{ Erlangs.}$$

Exemplo 4.6:

Quatro *clusters* de terminais de dados são conectados a um computador por meio de circuitos contratados (Figura 4.6).



Na Figura 4.6 (a) o tráfego proveniente dos *clusters* usa grupos separados de circuitos compartilhados.	Na Figura 4.6 (b) o tráfego proveniente de todos os *clusters* é concentrado em um grupo comum de circuitos.

Determine o nº total de circuitos requeridos em ambos os casos, quando o valor máximo suportado para *B* é 5%.

Assuma que há 22 terminais em cada *cluster* e que cada terminal está ativo 10% do tempo.

Utilize a tabela abaixo, onde são listados valores máximos de intensidade de tráfego (*A*) para várias probabilidades de bloqueio (*B*) e nºs de servidores (*N*).

N/*B*	0.01	0.05	0.1	0.5	1.0	2	5	10	15	20	30	40
1	.0001	.0005	.001	.005	.010	.020	.053	.111	.176	.250	.429	.667
2	.014	.032	.046	.105	.153	.223	.381	.595	.796	1.00	1.45	2.00
3	.087	.152	.194	.340	.455	.602	.899	1.27	1.60	1.93	2.63	3.48
4	.235	.362	.439	.701	.869	1.09	1.62	2.05	2.50	2.95	3.89	5.02
5	.452	.649	.762	1.13	1.36	1.66	2.22	2.88	3.45	4.01	5.10	6.60
6	.728	.996	1.15	1.62	1.91	2.28	2.96	3.76	4.44	5.11	6.51	8.19
7	1.05	1.39	1.58	2.16	2.50	2.94	3.74	4.67	5.46	6.23	7.86	9.80
8	1.42	1.83	2.05	2.73	3.13	3.63	4.54	5.60	6.50	7.37	9.21	11.4
9	1.83	2.30	2.56	3.33	3.78	4.34	5.37	6.55	7.55	8.52	10.6	13.0
10	2.26	2.80	3.09	3.96	4.46	5.08	6.22	7.51	8.62	9.68	12.0	14.7
11	2.72	3.33	3.65	4.61	5.16	5.84	7.08	8.49	9.69	10.9	13.3	16.3
12	3.21	3.88	4.23	5.28	5.88	6.61	7.95	9.47	10.8	12.0	14.7	18.0
13	3.71	4.45	4.83	5.96	6.61	7.40	8.83	10.5	11.9	13.2	16.1	19.6
14	4.24	5.03	5.45	6.66	7.35	8.20	9.73	11.5	13.0	14.4	17.5	21.2
15	4.78	5.63	6.08	7.38	8.11	9.01	10.6	12.5	14.1	15.6	18.9	22.9
16	5.34	6.25	6.72	8.10	8.88	9.83	11.5	13.5	15.2	16.8	20.3	24.5
17	5.91	6.88	7.38	8.83	9.65	10.7	12.5	14.5	16.3	18.0	21.7	26.2
18	6.50	7.52	8.05	9.58	10.4	11.5	13.4	15.5	17.4	19.2	23.1	27.8
19	7.09	8.17	8.72	10.3	11.2	12.3	14.3	16.6	18.5	20.4	24.5	29.5
20	7.70	8.83	9.41	11.1	12.0	13.2	15.2	17.6	19.6	21.6	25.9	31.2

Solução:

- O tráfego solicitado por cada *cluster* é $(22 \times 0.1) = 2.2$ Erlangs.
- Como o nº médio de circuitos ativos é muito menor do que o nº de fontes, uma análise que considera nº infinito de fontes pode ser usada.
- A partir da Tabela, o nº de circuitos requerido para $B = 5\%$, a uma intensidade de tráfego de 2.2 Erlangs é 5.
- Assim, a configuração mostrada na **Figura 4.6 (a)** requer um total de 20 circuitos.

- O tráfego total solicitado ao concentrador da configuração da **Figura 4.6 (b)** é $(4 \times 2.2) = 8.8$ Erlangs.
- A partir da Tabela, o nº de circuitos requerido para $B = 5\%$, a uma intensidade de tráfego de 8.8 Erlangs é 13.

N/B	0.01	0.05	0.1	0.5	1.0	2	5	10	15	20	30	40
1	.0001	.0005	.001	.005	.010	.020	.053	.111	.176	.250	.429	.667
2	.014	.032	.046	.105	.153	.223	.381	.595	.796	1.00	1.45	2.00
3	.087	.152	.194	.340	.455	.602	.899	1.27	1.60	1.93	2.63	3.48
4	.235	.362	.439	.701	.869	1.09	1.62	2.05	2.50	2.95	3.89	5.02
5	.452	.649	.762	1.13	1.36	1.66	2.22	2.88	3.45	4.01	5.10	6.60
6	.728	.996	1.15	1.62	1.91	2.28	2.96	3.76	4.44	5.11	6.51	8.19
7	1.05	1.39	1.58	2.16	2.50	2.94	3.74	4.67	5.46	6.23	7.86	9.80
8	1.42	1.83	2.05	2.73	3.13	3.63	4.54	5.60	6.50	7.37	9.21	11.4
9	1.83	2.30	2.56	3.33	3.78	4.34	5.37	6.55	7.55	8.52	10.6	13.0
10	2.26	2.80	3.09	3.96	4.46	5.08	6.22	7.51	8.62	9.68	12.0	14.7
11	2.72	3.33	3.65	4.61	5.16	5.84	7.08	8.49	9.69	10.9	13.3	16.3
12	3.21	3.88	4.23	5.28	5.88	6.61	7.95	9.47	10.8	12.0	14.7	18.0
13	3.71	4.45	4.83	5.96	6.61	7.40	8.83	10.5	11.9	13.2	16.1	19.6
14	4.24	5.03	5.45	6.66	7.35	8.20	9.73	11.5	13.0	14.4	17.5	21.2
15	4.78	5.63	6.08	7.38	8.11	9.01	10.6	12.5	14.1	15.6	18.9	22.9
16	5.34	6.25	6.72	8.10	8.88	9.83	11.5	13.5	15.2	16.8	20.3	24.5
17	5.91	6.88	7.38	8.83	9.65	10.7	12.5	14.5	16.3	18.0	21.7	26.2
18	6.50	7.52	8.05	9.58	10.4	11.5	13.4	15.5	17.4	19.2	23.1	27.8
19	7.09	8.17	8.72	10.3	11.2	12.3	14.3	16.6	18.5	20.4	24.5	29.5
20	7.70	8.83	9.41	11.1	12.0	13.2	15.2	17.6	19.6	21.6	25.9	31.2

* O Exemplo 4.6 demonstra que a consolidação de grupos com pequeno tráfego em um grupo de tráfego maior pode prover ganhos significativos no nº total de circuitos.
* Grupos grandes são mais eficientes do que múltiplos pequenos grupos porque é improvável que os pequenos grupos venham a se tornar sobrecarregados ao mesmo tempo (assumindo chamadas independentes).
* Na verdade, o tráfego excedente em um grupo pode ser acomodado em outros grupos que estejam desocupados.
* Assim, aqueles circuitos que são necessários para acomodar picos de tráfego, mas estão normalmente desocupados, são utilizados com mais eficiência quando o tráfego é combinado em um grupo.
* Esta característica é uma das responsáveis pela integração de tráfego de voz e dados em uma rede comum.
* A economia total em termos de custo de transmissão é mais significativa quando as intensidades de tráfego individuais são menores.
* Desta forma, pode-se dizer que a área periférica de uma rede é a mais beneficiada pela concentração do tráfego.

Exemplo 4.7: O que acontece às probabilidades de bloqueio discutidas no Exemplo 4.6, Figuras 4.6 (a) e (b) quando a intensidade de tráfego é aumentada por 50%?

Solução:

- Na configuração da **Figura 4.6 (a)**, o nº de circuitos requerido para $B = 5\%$, a uma intensidade de tráfego de 2.2 Erlangs é 5.
- A partir da Tabela, para 5 circuitos e uma intensidade de tráfego aumentada por 50% (3.3 Erlangs), a probabilidade de bloqueio ficará em torno de 14%.
- Assim, se a intensidade de tráfego de cada grupo aumenta de 2.2 p/ 3.3 Erlangs, a probabilidade de bloqueio da configuração da Figura 4.6 (a) aumenta de 5% para quase 14%.
- O tráfego total solicitado ao concentrador da configuração da **Figura 4.6 (b)** será $(4 \times 3.3) = 13.2$ Erlangs.
- A partir da Tabela, para os 13 de circuitos requeridos no Exemplo 4.6, a uma intensidade de tráfego de 13.2 Erlangs, a probabilidade de bloqueio será 20%.
- Assim, na configuração da Figura 4.6 (b), um aumento de 50% na intensidade de tráfego causa um aumento de 400% na probabilidade de bloqueio, que passa de 5 para 20%.

N/B	0.01	0.05	0.1	0.5	1.0	2	5	10	15	20	30	40
1	.0001	.0005	.001	.005	.010	.020	.053	.111	.176	.250	.429	.667
2	.014	.032	.046	.105	.153	.223	.381	.595	.796	1.00	1.45	2.00
3	.087	.152	.194	.340	.455	.602	.899	1.27	1.60	1.93	2.63	3.48
4	.235	.362	.439	.701	.869	1.09	1.62	2.05	2.50	2.95	3.89	5.02
5	.452	.649	.762	1.13	1.36	1.66	2.22	2.88	3.45	4.01	5.10	6.60
6	.728	.996	1.15	1.62	1.91	2.28	2.96	3.76	4.44	5.11	6.51	8.19
7	1.05	1.39	1.58	2.16	2.50	2.94	3.74	4.67	5.46	6.23	7.86	9.80
8	1.42	1.83	2.05	2.73	3.13	3.63	4.54	5.60	6.50	7.37	9.21	11.4
9	1.83	2.30	2.56	3.33	3.78	4.34	5.37	6.55	7.55	8.52	10.6	13.0
10	2.26	2.80	3.09	3.96	4.46	5.08	6.22	7.51	8.62	9.68	12.0	14.7
11	2.72	3.33	3.65	4.61	5.16	5.84	7.08	8.49	9.69	10.9	13.3	16.3
12	3.21	3.88	4.23	5.28	5.88	6.61	7.95	9.47	10.8	12.0	14.7	18.0
13	3.71	4.45	4.83	5.96	6.61	7.40	8.83	10.5	11.9	13.2	16.1	19.6
14	4.24	5.03	5.45	6.66	7.35	8.20	9.73	11.5	13.0	14.4	17.5	21.2
15	4.78	5.63	6.08	7.38	8.11	9.01	10.6	12.5	14.1	15.6	18.9	22.9
16	5.34	6.25	6.72	8.10	8.88	9.83	11.5	13.5	15.2	16.8	20.3	24.5
17	5.91	6.88	7.38	8.83	9.65	10.7	12.5	14.5	16.3	18.0	21.7	26.2
18	6.50	7.52	8.05	9.58	10.4	11.5	13.4	15.5	17.4	19.2	23.1	27.8
19	7.09	8.17	8.72	10.3	11.2	12.3	14.3	16.6	18.5	20.4	24.5	29.5
20	7.70	8.83	9.41	11.1	12.0	13.2	15.2	17.6	19.6	21.6	25.9	31.2

O Exemplo 4.7 ilustra uma importante consideração no projeto de redes: as probabilidades de bloqueio são muito sensíveis a aumentos na intensidade de tráfego, particularmente quando os canais são fortemente utilizados.

* Como grupos tronco grandes utilizam os canais de forma mais eficiente, são mais vulneráveis a aumentos na intensidade de tráfego do que um conjunto de grupos menores projetados para prover o mesmo grau de serviço.
* Ainda, grupos tronco grandes são mais afetados por falhas na transmissão do que vários grupos menores.
* Em ambos os casos, a vulnerabilidade dos grandes grupos ocorre porque operam com menor reserva de capacidade do que múltiplos grupos pequenos.

Um segundo aspecto demonstrado da análise de bloqueio realizada no Exemplo 4.7 é que os resultados são altamente dependentes de quão acurada for a estimação da intensidade de tráfego.

* Valores precisos de intensidade de tráfego em geral não são disponíveis e, mesmo quando são, não contêm informação sobre o crescimento esperado do tráfego.
* A análise de probabilidades de bloqueio é útil porque, apesar das limitações, permite uma comparação objetiva entre diferentes tamanhos e configurações de redes .
* O projeto cuja relação custo/benefício for a mais adequada para um dado grau de serviço deverá ser escolhido, mesmo que as intensidades de tráfego sejam hipotéticas.

4.3 Probabilidades de Bloqueio em Redes

* Em redes de comunicações (com mais do que um roteador) as probabilidades de bloqueio precisam ser calculadas de uma forma global (*end-to-end*). Assim, precisam ser considerados:
  - As interações do tráfego em várias rotas da rede.
  - O efeito da sobrecarga de algumas rotas.

4.3.1 Probabilidades de Bloqueio *End-to-End*

* Em geral, em uma rede de comunicações, as conexões envolvem uma série de *links* de transmissão, cada um deles selecionado a partir de um conjunto de alternativas.
* O procedimento mais simples para determinar a probabilidade de bloqueio *end-to-end* é semelhante à análise de probabilidades de bloqueio para as chaves comutadoras, que estudamos anteriormente.
* A Figura 4.7 apresenta um grafo de probabilidades para a análise de probabilidades *end-to-end*.

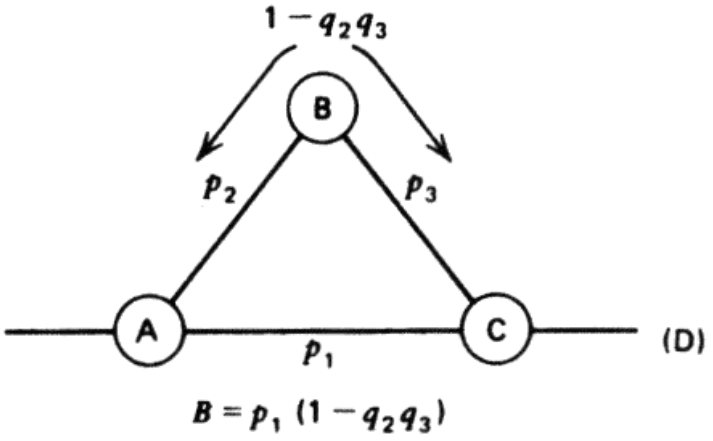


Figura 4.7: Grafo para análise de probabilidades de bloqueio *end-to-end*.

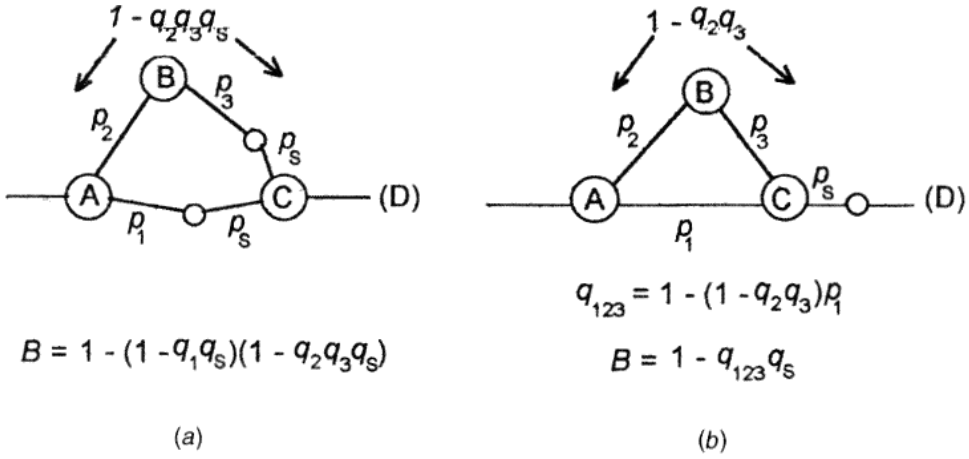
* Para a determinação da equação presente na Figura 4.7, foram feitas várias simplificações:

1. A probabilidade de bloqueio das chaves não é incluída. (válido para chaves digitais modernas).

Quando for necessário incluir probabilidades de bloqueio inerentes a chaves, considera-se a presença de uma fonte de bloqueio em série com os grupos troncos.

Quando mais do que um *link* passa por uma mesma chave, conforme mostra o nó C da Figura 4.7, torna-se necessário ainda considerar a correlação existente entre as probabilidades de bloqueio associadas a cada *link*.

A Figura 4.8 descreve duas abordagens adequadas para considerar a correlação entre as probabilidades dos *links* que passam por um mesmo nó (o *link* comum é o *link* que une os nós C e D), quando se inclui a probabilidade de bloqueio inerente às chaves.



(a) Abordagem otimista:

as probabilidades de bloqueio são independentes. Neste caso, as probabilidades de bloqueio estarão em série com os *links* individuais.

(b) Abordagem conservadora:

as probabilidades de bloqueio são completamente correlacionadas. Neste caso, as probabilidades de bloqueio estarão em série com o *link* comum.

Figura 4.8: Correlação existente entre as probabilidades de bloqueio de cada *link* incorporada na determinação da probabilidade de bloqueio *end-to-end*.

2. A segunda simplificação feita para derivar a equação presente na Figura 4.7 envolve assumir independência para as probabilidades de bloqueio de grupos tronco.

Neste caso, a composição das probabilidades de bloqueio de *n* rotas paralelas é simplesmente o produto das probabilidades respectivas $B = p^n$.

Similarmente, a independência implica em que a probabilidade de bloqueio de *n* caminhos em série seja $B = 1 - q^n$, onde *q* é a probabilidade de que o *link* esteja desbloqueado $(q = 1 - p)$.

- Na realidade, probabilidades de bloqueio nunca serão completamente independentes.

- Especialmente quando uma grande quantidade de tráfego sobre uma rota é resultante da ocorrência de sobrecarga em outras rotas.

- Sempre que a primeira rota estiver ocupada, é provável que uma grande parte do tráfego seja desviada para a segunda rota (fenômenos correlacionados).

Desta forma, é mais provável que uma rota alternativa esteja ocupada quando uma rota primária estiver ocupada.

*Este texto é uma tradução livre e parcial do Capítulo 12 do livro "*Digital Telephony*" de J. C. Bellamy.

4.4 *Delay Systems*

* A segunda categoria de análise de tráfego diz respeito a sistemas que **atrasam** as solicitações de chamadas que não são atendidas até que os recursos para que a chamada seja completada estejam disponíveis.

* Estes sistemas são conhecidos por *delay systems*, *waiting-call systems* e *queuing systems*.

* Uma fila (*queue*) pode consistir de dispositivos com capacidade de armazenamento, como também pode apenas consistir de uma lista de fontes aguardando por serviço, em que o armazenamento de mensagens é responsabilidade das próprias fontes.

* Os fundamentos da teoria de filas são devidos aos pesquisadores de tráfego em telecomunicações.

* Erlang apresentou a primeira solução para o sistema de atraso do tipo mais básico.

* **Exemplos de sistemas que operam atrasos em telecomunicações são:** chaveamento por mensagem, chaveamento por pacote, multiplexação estatística por divisão no tempo, comunicações de dados multi-ponto, distribuição automática de chamadas, processamento de camadas, etc. É importante que sejam citados, ainda, alguns sistemas PBX que operavam como *loss systems* e passaram a operar como *delay systems*.

* Em geral, sistemas que operam admitindo atrasos permitem uma maior utilização dos servidores do que sistemas que operam admitindo perdas.

* A melhora na utilização dos servidores é obtida porque os picos de chamadas são suavizados pelas filas.

* Mesmo quando as chamadas são aleatórias, os servidores "enxergam" um padrão de chamadas regular, devido à espera em filas.

- A Figura 4.9 apresenta o mesmo padrão de tráfego mostrado nas Figuras 4.1 e 4.3.
- Neste caso, entretanto, a sobrecarga de tráfego é atrasada até que as terminações de chamadas liberem canais ocupados.

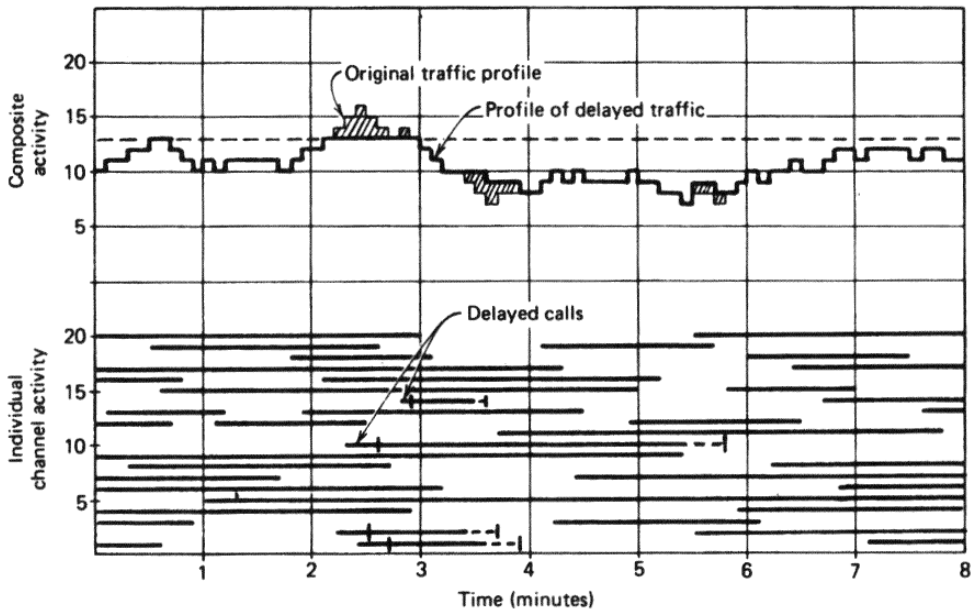


Figura 4.9: Perfil de Atividade *blocked calls delayed* (13 canais).

- **Na maior parte da análise assumiremos que todo o tráfego solicitado do sistema é atendido.**
- Uma implicação desta consideração é que a intensidade de tráfego solicitada *A* deve ser menor do que o número de servidores *N*.
- Mesmo nesta situação, há dois casos em que o tráfego transportado pode ser menor do que o tráfego solicitado:

  (1) Algumas fontes podem cansar de esperar em uma longa fila e abandonar a solicitação.

  (2) A capacidade de armazenamento de solicitações pode ser finita e as solicitações podem ocasionalmente ser rejeitadas pelo sistema.

- **Uma segunda consideração na análise que segue é que o número de fontes é considerado infinito.**
- Em um sistema que admite atrasos pode haver um nº finito de fontes, em um sentido físico, mas infinito no sentido operacional, porque cada fonte pode ter um nº arbitrário de solicitações pendentes.
- Quando se considera o atendimento a todas as chamadas e nº de fontes infinito, surge a implicação adicional de serem consideradas filas de capacidade infinita.
- Mesmo quando a intensidade de tráfego solicitada é menor do que o nº de servidores, não existe limite estatístico para o nº de chamadas que ocorrem em um curto período de tempo.
- Assim, a fila para um sistema sem perdas deve ser arbitrariamente longa.
- Em um sentido prático, apenas filas finitas podem ser efetivamente realizadas.
- Assim, sempre poderá ocorrer uma chance estatística de bloqueio.

- Quando são tratados sistemas que admitem atrasos, é conveniente separar o tempo total que uma solicitação permanece no sistema em:

(1) Tempo de espera e

(2) Tempo de duração de chamada (ou tempo de serviço).

- Em contraste com *loss systems*, o desempenho dos *delay systems* é geralmente dependente da distribuição de tempos de serviço e não apenas do valor médio t_m .

- Duas distribuições de tempo de serviço são consideradas:

(1) Tempo de serviço constante.

Que representa os serviços mais determinísticos.

(2) Tempo de serviço exponencial.

Que representa os serviços mais aleatórios existentes.

Assim, um sistema que opere com alguma outra distribuição de tempo de serviço terá o desempenho situado entre o desempenho produzido por estas duas distribuições.

- **O propósito básico desta análise é determinar a distribuição de probabilidades dos tempos de espera.**
- A partir da determinação da distribuição de probabilidades, o tempo médio de espera é facilmente determinado.
- Muitas vezes apenas se deseja determinar **o tempo médio de espera.**
- A maior parte das vezes, no entanto, se está interessado na **probabilidade de que o tempo de espera exceda a algum valor especificado**.
- Em ambos os casos, os tempos de espera são dependentes dos seguintes fatores:

(1) Intensidade e natureza probabilística do tráfego solicitado.

(2) Distribuição dos tempos de serviço.

(3) Número de servidores.

(4) Número de fontes.

(5) Heurística adotada para operação da fila.

- Uma heurística de operação de filas envolve um certo número de fatores:

1. **O 1º deles diz respeito à maneira pela qual as chamadas em espera são selecionadas.**

   Normalmente, as chamadas são selecionadas da seguinte maneira: a 1ª solicitação que chega é a 1ª a ser atendida (*first-come*, *first-served basis* – FCFS *basis*), a qual costuma também ser chamada *first-in*, *first-out* – FIFO.

   Algumas vezes, no entanto, o sistema servidor não mantém uma fila, mas atende as fontes fazendo um rodízio entre as fontes em espera.

   Outras vezes, as chamadas em espera podem ser selecionadas aleatoriamente.

   Além disso, variações adicionais de serviço surgem se qualquer destes esquemas for acrescido de uma heurística para estabelecer prioridades, que permita a algumas chamadas serem atendidas antes de outras, na fila.

2. **Um 2º aspecto da heurística de serviço que deve ser considerado é o comprimento da fila.**

   Se o tamanho máximo da fila é menor do que o nº efetivo de fontes, pode ocorrer bloqueio (da mesma forma que ocorre em sistemas *lost calls cleared*). Resultando que 2 características do grau de serviço devem ser consideradas: a probablidade de atraso e a probabilidade de bloqueio.

   Um exemplo de um sistema com as características de atraso e bloqueio é um distribuidor de chamadas automático com mais circuitos de acesso do que atendentes (operadores).

   Neste sistema, as solicitações de chamadas são colocadas em uma fila.

   Quando o sistema está fortemente carregado, entretanto, o bloqueio ocorre antes que o distribuidor de chamadas automático tenha sido alcançado.

★ Para simplificar a caracterização dos particulares sistemas, a teoria de filas adota uma notação concisa para a classificação de vários tipos de sistemas que operam atrasos. Esta notação é mostrada na Figura 4.10.

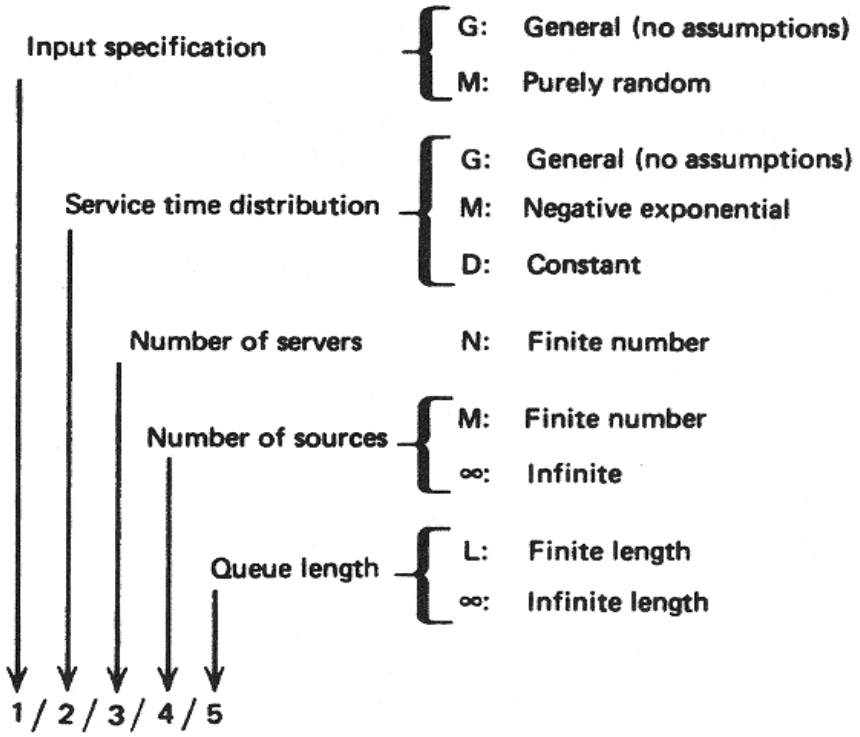


Figura 4.10: Notação de sistema baseado em filas.

★ O formato apresentado na Figura representa uma extensão do formato mais utilizado, que é abreviado, eliminando a última ou as 2 últimas entradas.

★ Entradas eliminadas são assumidas infinitas.

★ Por exemplo:

- Um sistema com um único servidor (*number of servers*), com entradas aleatórias (*input specification*) e tempos de serviço modelados por uma exponencial negativa (*service time distribution*) é usualmente especificado como M/M/1.
- Tanto o nº de fontes (*number of soucers*) quanto o tamanho de fila permitido (*queue length*) são assumidos infinitos.

4.4.1 Modelamento Exponencial de Tempos de Serviço

O *delay system* de análise mais simples é um sistema com chamadas aleatórias, tempos de serviço modelados por uma exponencial negativa e nº de servidores finito: M/M/N.

Uma distribuição de chamadas aleatória é modelada por uma distribuição exponencial negativa de intervalos de tempo entre chamadas.

Assim, na notação curta, a letra M sempre irá se referir a distribuições modeladas por exponenciais negativas (um M é usado porque uma distribuição puramente aleatória é sem memória (***M****emoryless*)).

Em um sistema M/M/1 e todos os outros sistemas que iremos considerar, é assumido que chamadas são atendidas na ordem em que chegam.

A análise que segue também assume que a probabilidade de uma chamada é independente do nº de solicitações já presentes na fila (fontes infinitas).

A partir destas considerações, a probabilidade de que uma chamada experimente congestionamento e seja, portanto, atrasada, foi derivada por Erlang como:

$$\text{prob}\left(\text{delay}\right) = p\left(>0\right) = \frac{NB}{N - A\left(1 - B\right)} \tag{4.10}$$

onde:

N = nº de servidores
A = tráfego solicitado (Erlangs)
B = probabilidade de bloqueio para um sistema *lost calls cleared* (Equação 4.8).

$$B = E_{1,N}\left(A\right) = \frac{A^N}{N!\sum_{i=0}^{N}\left(\frac{A^i}{i!}\right)} \tag{4.8}$$

A probabilidade de atraso $p(>0)$, descrita na Equação 4.10, é referida de várias formas na literatura:

* segunda fórmula de Erlang,
* fórmula Erlang C,
* fórmula *Lost Calls Delayed* – LCD e
* $E_{2,N}(A)$.

★ Para sistemas de um único servidor $(N=1)$, a probabilidade de atraso se reduz a ρ, que é simplesmente a utilização ou tráfego transportado pelo servidor.

★ Assim, a probabilidade de atraso para um sistema de um único servidor é também igual à carga solicitada λt_m (assumindo $\lambda t_m < 1$).

★ A **distribuição de tempos de espera** para chamadas aleatórias, tempos de serviço aleatórios e heurística de serviço FIFO é expressa por

$$p(>t) = p(>0)\,e^{-(N-A)t/t_m} \qquad (4.11)$$

onde:

$p(>0)$ = probabilidade de atrasos dada na Equação 4.10

t_m = tempo de serviço médio da distribuição de tempo de serviço modelada pela exponencial negativa

★ A Equação 4.11 define a probabilidade de que uma chamada que chegue a um instante de tempo aleatoriamente escolhido seja atrasada por mais do que t/t_m tempos de serviço.

★ **A Figura 4.11 representa a relação expressa pela Equação 4.11, apresentando as capacidades de tráfego de vários números de servidores, como uma função dos tempos aceitáveis de atraso.**

★ Dado um tempo "alvo" (tolerado) de atraso t/t_m, a Figura 4.11(a) apresenta a máxima intensidade de tráfego se o atraso "alvo" é excedido por apenas 10% das solicitações de chamadas.

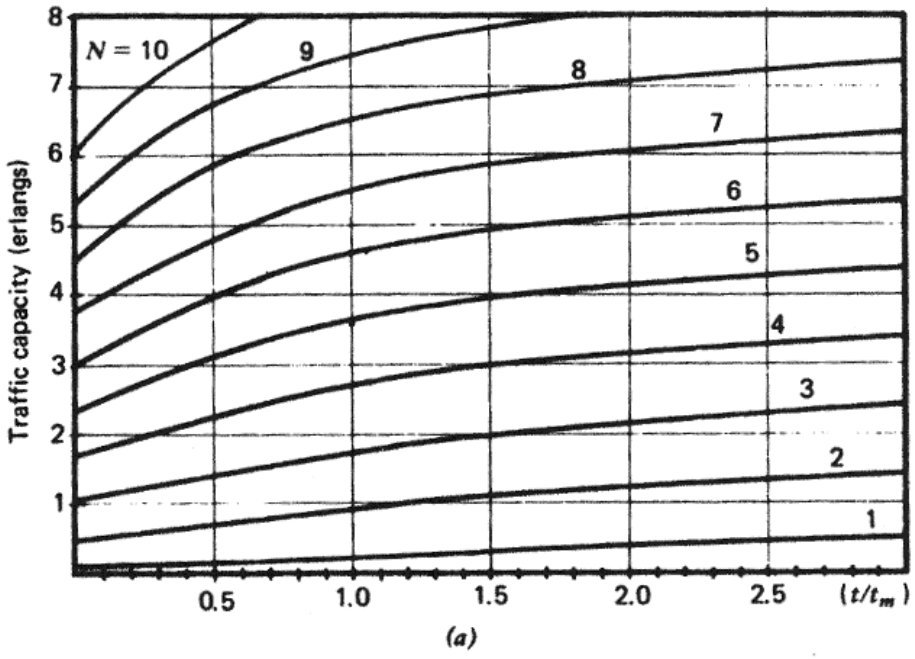


Figura 4.11(a): Capacidade de tráfego de sistemas multi-canais que admitem atrasos, com tempos de serviço exponenciais: Probabilidade de exceder *t*, $p(>t)=10\%$.

$$p(>t) = p(>0)e^{-(N-A)t/t_m} \qquad (4.11)$$

De forma similar, a Figura 4.11(b) apresenta a máxima intensidade de tráfego se o atraso "alvo" é excedido por apenas 1% das solicitações de chamada.

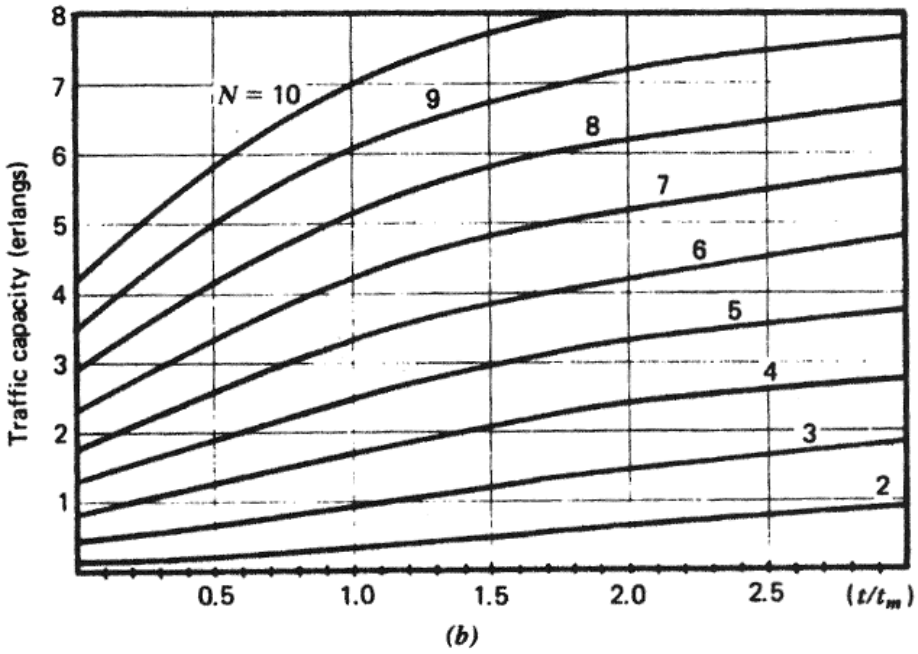


Figura 4.11(b): Capacidade de tráfego de sistemas multi-canais que operam atrasos com tempos de serviço exponenciais: Probabilidade de exceder *t*, $p(>t)=1\%$. $\left(p(>t)=p(>0)e^{-(N-A)t/t_m} \quad (4.11) \right)$

★ Integrando a Equação 4.11 sobre todo o tempo, o tempo de espera médio para todas as solicitações de chamada resulta em

$$\bar{t} = \frac{p(>0)t_m}{N-A} \qquad (4.12)$$

★ Note que $\bar{t}$ é o atraso esperado para todas as solicitações de chamada.

★ O atraso médio daquelas chamadas que efetivamente sofreram atraso é denotado por

$$t_w = \frac{t_m}{N-A} \qquad (4.13)$$

Exemplo 4.8: Uma rede comutada por mensagem deve ser projetada para 95% de utilização de seus *links.*

Assumindo comprimentos de mensagem exponencialmente distribuídos e uma taxa de 10 mensagens por minuto, qual é o tempo médio de espera, e qual é a probabilidade que o tempo de espera exceda 5 minutos?

Assuma que a rede comutada por mensagem utiliza um único canal entre cada par de nós, de forma que haja um único servidor e uma única fila para cada *link* de transmissão.

Solução:

São dados: $\rho = 0.95$ e $\lambda = 10$ mensagens por minuto.

Para sistemas de um único servidor $(N = 1)$, a probabilidade de atraso $p(>0)$ se reduz a ρ, que é simplesmente a utilização ou tráfego transportado pelo servidor.

Assim, a probabilidade de atraso para um sistema de um único servidor é também igual à intensidade de tráfego solicitado $A = \lambda t_m$ (desde que $\lambda t_m < 1$).

Como *N*=1 $\Rightarrow$ $\rho = p(>0) = \lambda t_m$ e o tempo de serviço médio pode ser determinado por

$$t_m = \frac{\rho}{\lambda} = \frac{0.95}{10} = 0.095 \,\text{min}\,.$$

O tempo de espera médio (não incluindo o tempo de serviço) é determinado por

$$\bar{t} = \frac{p(>0)\,t_m}{N - A} \Rightarrow \bar{t} = \frac{0.95 \times 0.095}{1 - 0.95} = 1.805 \,\text{min}$$

Através da Equação 4.11 $\left(p(>t) = p(>0)\,e^{-(N-A)t/t_m} \right)$, pode-se determinar a probabilidade do tempo de espera exceder 5 min, conforme

$$p(>5) = 0.95\, e^{-(1-0.95)5/0.095} = 0.068$$

Desta forma, 6.8% das mensagens experimentam atrasos de fila de mais do que 5 minutos.

4.4.2 Modelamento Constante de Tempos de Serviço

O tipo de *delay system* que será abordado é um sistema com chamadas aleatórias, tempos de serviço constantes e um único servidor (M/D/1).

São assumidas fontes infinitas e uma heurística de serviço FIFO.

O tempo de espera médio para um único servidor com tempo de serviço constante é determinado por

$$\bar{t} = \frac{\rho \ t_m}{2(1-\rho)} \tag{4.14}$$

onde: $\rho = A$ é a utilização do servidor.

★ A Equação 4.14 produz um tempo de espera médio que é exatamente a metade daquele de um sistema com um único servidor com tempos de serviço exponenciais.

★ Tempos de serviço exponenciais causam maiores atrasos médios porque há dois processos aleatórios envolvidos na geração do atraso:

★ Em ambos os tipos de sistemas, os atrasos ocorrem quando um surto de chamadas excede a capacidade dos servidores.

★ Com tempos de serviço exponenciais, entretanto, longos atrasos também surgem devido aos excessivos tempos de serviço introduzidos por umas poucas chamadas.

Se o perfil de atividade de um sistema de tempo de serviço constante (M/D/1) é comparado com o perfil de atividade de um sistema de tempo de serviço exponencial (M/M/1):

⇨ o sistema M/D/1 estará ativo por períodos de tempo mais freqüentes e mais curtos. Ou seja,

⇨ o sistema M/M/1 tem uma variância maior na duração de seus períodos ocupados.

A atividade média de ambos os sistemas é igual à utilização do servidor ρ.

Então, a probabilidade de atraso para um sistema de um único servidor com tempo de serviço constante é idêntico àquele para tempos de serviço exponenciais:

$$p(>0) = \lambda t_m .$$

A probabilidade de congestionamento para maiores valores de *N* é relativamente próxima à probabilidade de congestionamento para tempos de serviço exponenciais.

Assim, a Equação 4.11

$$p(>t) = p(>0)e^{-(N-A)t/t_m}$$

pode ser usada como uma boa aproximação para $p(>0)$ para sistemas de múltiplos servidores com distribuições de tempo de serviço arbitrárias.

Para sistemas de um único servidor, com tempos de duração de chamadas constantes, a probabilidade de ocorrer atraso maior do que um valor arbitrário *t* é

$$p(>t) = p[>(k+r)t_m] =$$

$$= 1-(1-\rho)\sum_{i=0}^{k}\frac{\rho^i (i-t/t_m)^i e^{-\rho(i-t/t_m)}}{i!} = \qquad (4.15)$$

$$= 1-(1-\rho)e^{\lambda t}\sum_{i=0}^{k}\frac{(i\rho-\lambda t)^i e^{-i\rho}}{i!}$$

onde:

t_m = tempo de serviço médio

k = maior quociente inteiro de t/t_m

r = resto de t/t_m

ρ = utilização do servidor = λt_m

A Figura 4.12 apresenta comparações de distribuições de tempo de espera para sistemas de um único servidor com tempos de serviço exponenciais e constantes.

Para cada par de curvas, a curva superior é para tempos de serviço exponenciais e a curva inferior para tempos de serviço constantes.

Como todas as outras distribuições de tempo de serviço produzem probabilidades de atraso entre estes dois extremos, a Figura 4.12 provê uma indicação direta do intervalo de variação dos atrasos possíveis.

*Este texto é uma tradução livre e parcial do Capítulo 12 do livro "*Digital Telephony*" de J. C. Bellamy.

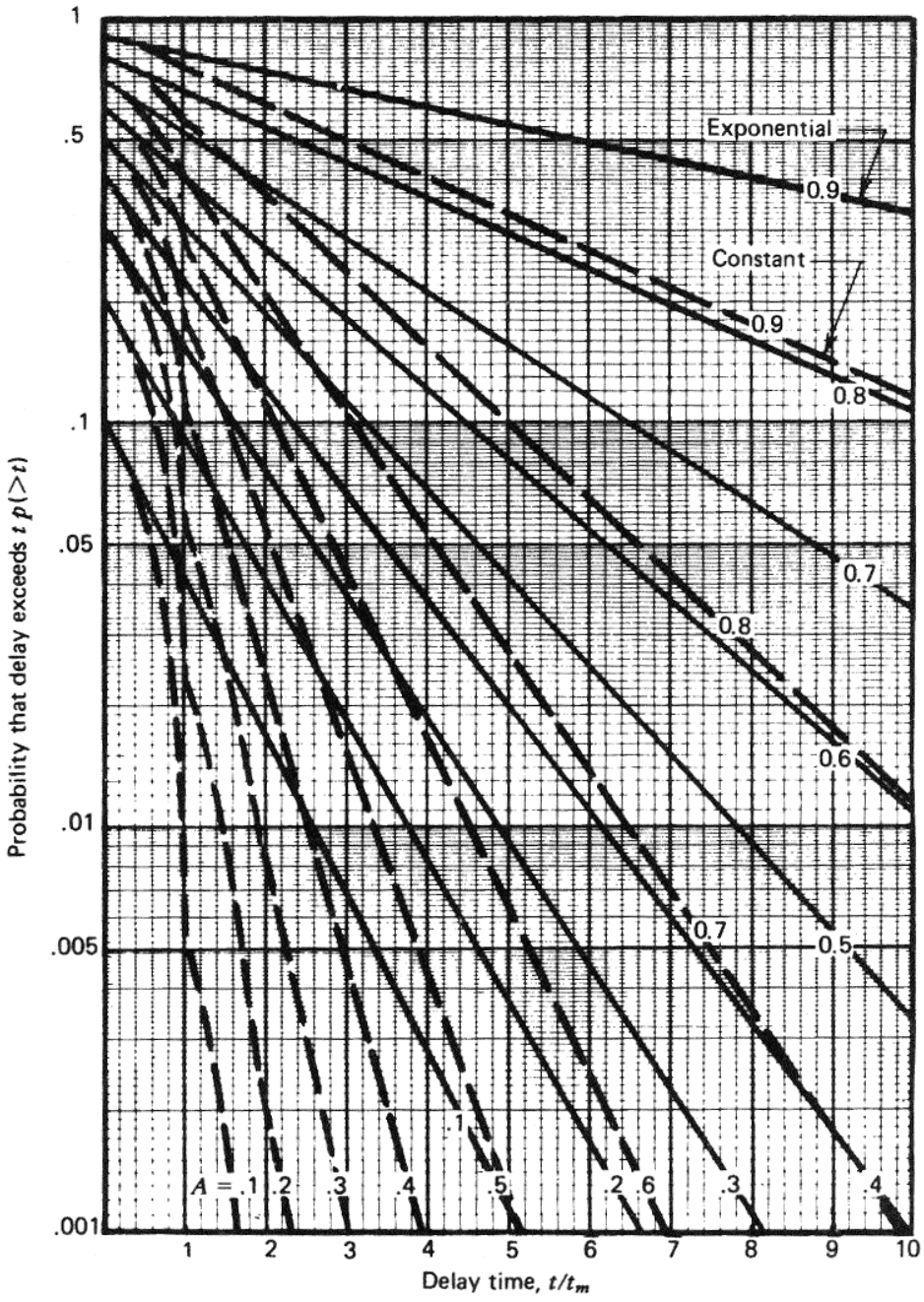


Figura 4.11: Probabilidades de atraso de sistemas de um único servidor (tempos de serviço exponenciais e constantes).

Exemplo 4.9:

Uma rede opera por chaveamento de pacotes de 300 bits sobre linhas de 9600 bps.

Se a utilização de um *link* for 90%, qual é o atraso médio?

Qual a porcentagem de pacotes que sofrem mais do que 0.35s de atraso?

Qual é o atraso médio se a carga solicitada aumentar 10%?

Solução:

Comprimentos de mensagem de 300 bits e uma taxa de dados de 9600 bps implicam que o tempo de serviço (de tamanho fixo) seja

$$300/9600 = 0.031\text{s}\,.$$

A partir da Equação 4.14 $\left(\bar{t} = \dfrac{\rho\ t_m}{2(1-\rho)}\right)$, o tempo de espera médio é

$$\bar{t} = \frac{0.9 \times 0.031}{2(1-0.9)} = 0.14\text{s}$$

O atraso médio total no nó é obtido adicionando o tempo de espera médio ao tempo de serviço, assim:

Atraso médio = 0.140 + 0.031 = 0.171s

Como o tempo de serviço é 0.031s, 0.35s de atraso ocorrem quando o tempo de espera é

0.35-0.031=0.319.

Isto corresponde a

0.319/0.031=10 tempos de serviço.

A partir da Figura 4.11, a probabilidade de atraso para t/t_m =10 é aproximadamente 0.12.

Assim, 12% dos pacotes experimentam atrasos maiores do que 0.35s.

Um aumento de 10% na intensidade de tráfego implica que a nova carga oferecida é 0.99 Erlang.

A partir da Equação 4.14, o tempo de espera médio se torna

$$\bar{t} = \frac{0.99 \times 0.031}{2(1-0.99)} = 0.53\text{s}$$

Então, quando a carga solicitada aumenta por apenas 10%, o atraso médio no nó aumenta aproximadamente 9 vezes, para um valor de

1.53+0.031=1.56s!

Assim como nos sistemas com perdas, o desempenho de sistemas com atrasos é muito sensível a aumentos na intensidade de tráfego, especialmente quando os circuitos são fortemente utilizados.

Desta forma, o fluxo de controle é um aspecto crítico em uma operação comutada por pacotes, particularmente quando se deseja operar em tempo real.

4.5 Modelamento com Nº Finito de Fontes

⇨ É possível modelar sistemas LCC ou LCD, considerando-se um nº finito de usuários, no entanto, as expressões resultantes acabam por ser muito mais complicadas do que as expressões para os modelos Erlang.

⇨ Além disso, o modelamento se torna inadequado para casos em que o nº de usuários é muitas ordens de magnitude maior do que o de canais disponíveis.

⇨ Sendo assim, as expressões de Erlang conduzem a uma estimativa conservadora do grau de serviço pois, como o nº real de usuários é finito, os resultados sempre predizem uma probabilidade de bloqueio um pouco maior do que aquela que pode, de fato, ocorrer.

⇨ Valores para as expressões de Erlang são tabelados, de forma que a análise das combinações desejadas de grau de serviço, tráfego e nº de canais necessários, se torna mais prática.