2.
Teoria das Filas

Segundo Fogliatti (2007), a teoria das filas consiste na modelagem analítica de processos ou sistemas que resultam em espera e tem como objetivo determinar e avaliar quantidades, denominadas medidas de desempenho, que expressam a produtividade/operacionalidade desses produtos, e de posse destas informações buscar meios para minimizar os impactos negativos das esperas nos processos.

Teixeira e Pinheiro (2010) apresentam uma visão mais específica, quando afirmam que a teoria das filas é um ramo da probabilidade que estuda a formação de filas, através de análises matemáticas precisas e propriedades mensuráveis das filas. Ela provê modelos para demonstrar previamente o comportamento de um sistema que ofereça serviços cuja demanda cresce aleatoriamente, tornando possível dimensioná-lo de forma a satisfazer os clientes e ser viável economicamente para o provedor do serviço, evitando desperdícios e gargalos.

2.1.
Características estruturais dos sistemas de fila

Existem muitos tipos diferentes de filas. Em Moreira (2007), as filas são estruturadas de acordo com a Figura 5, em quatro partes principais: a fonte de clientes; a chegada de clientes; o processo de seleção; e o posto de atendimento. Os clientes são indivíduos de uma população que chegam ao local da prestação do serviço de acordo com determinado comportamento estatístico, para serem atendidos de acordo com um critério de seleção preestabelecido e serão atendidos de acordo com características próprias.

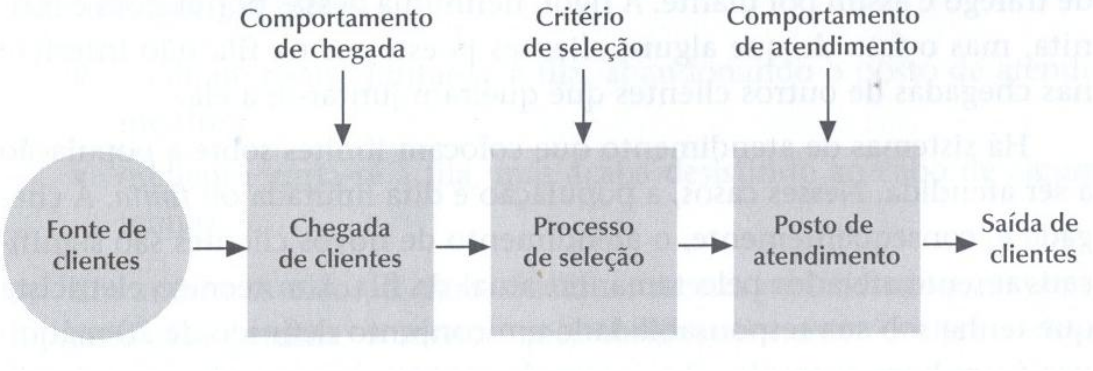

Figura 5: Estrutura de uma fila
Fonte: Moreira 2007

2.2. Fonte de clientes

Os clientes pertencem a uma população maior, onde todos são clientes potenciais. As fontes podem ser finitas ou infinitas, sendo infinitas aquelas em que a probabilidade de chegada não é afetada de forma significativa pelo fato de que alguns clientes já estão aguardando na fila, já na chegada finita ocorre o contrário.

2.3. Modelos de chegadas

Segundo Moreira (2007), o processo de chegadas dos usuários é especificado pelo comportamento do fluxo de chegadas dos mesmos ao sistema. Se forem conhecidos o número de chegadas e os instantes de tempo em que elas acontecem, esse processo é denominado determinístico; caso contrário, tem-se um comportamento aleatório constituindo um processo estocástico caracterizado por uma distribuição de probabilidade. Para essa distribuição, é necessária a especificação de um parâmetro denominado taxa de chegadas, que representa o número médio de usuários que chegam ao sistema por unidade de tempo.

Usualmente as taxas de chegada são representadas por λ e há duas formas tradicionais de se falar sobre a chegada de clientes para o atendimento: número de clientes que chegam em um dado intervalo de tempo e o tempo decorrido entre

duas chegadas consecutivas. É muito comum nos estudos de teoria das filas se utilizar da distribuição de Poisson para configurar a taxa de chegada à fila e atendimento.

A Tabela 2 apresenta as distribuições de probabilidade utilizadas em cada uma das formas tradicionais de chegada de clientes.

Tabela 2: Distribuições de probabilidade utilizadas nas taxas de chegada de clientes

Grandezas	Distribuição de chegada	Médias
Número de chegadas na unidade de tempo (taxa de chegada)	Poisson	λ
Tempo decorrido entre duas chegadas consecutivas	Exponencial	$1/\lambda$

Fonte: Moreira (2007)

2.4. Modelos de atendimento

Segundo Moreira (2007), o processo de atendimento é especializado pelo comportamento do fluxo de usuários atendidos e a sua caracterização é análoga à do processo de chegadas.

Usualmente nos modelos de atendimento usa-se o parâmetro μ e, assim como nos modelos de chegadas, existem duas nomenclaturas comumente utilizadas, associadas a este parâmetro: número de atendimentos na unidade de tempo e tempo decorrido entre dois atendimentos consecutivos. Cada uma dessas

grandezas apresenta uma distribuição de probabilidade, como é ilustrado na Tabela 3.

Tabela 3: Taxa de atendimento de clientes no sistema de filas

Grandezas	Distribuição de atendimento	Médias
Número de atendimentos na unidade de tempo(taxa de atendimento	Poisson	μ
Tempo decorrido entre dois atendimentos consecutivos	Exponencial	$1/\mu$

Fonte: Moreira (2007)

Os modelos de atendimento podem apresentar diversas configurações: canal único, canal múltiplo, atendimento único, atendimento múltiplo. O canal único se configura por ter apenas uma instalação de atendimento, podendo ter um ou mais postos de atendimento, porém em série. O canal múltiplo apresenta mais de um canal de atendimento em paralelo, atuando de forma independente. O atendimento múltiplo é realizado por mais de uma instalação de atendimento em série, dependente uma da outra. Já o atendimento único consiste na realização do atendimento feita integralmente em um posto, independente de qualquer outro posto. As figuras Figura 6, Figura 7, Figura 8, Figura 9, exemplificam a combinação destes conceitos.

Figura 6: Canal único, atendimento único.
Fonte: Moreira 2007

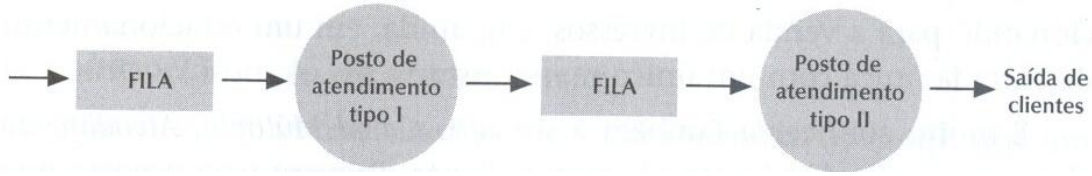

Figura 7: Canal único, atendimento múltiplo.
Fonte: Moreira 2007

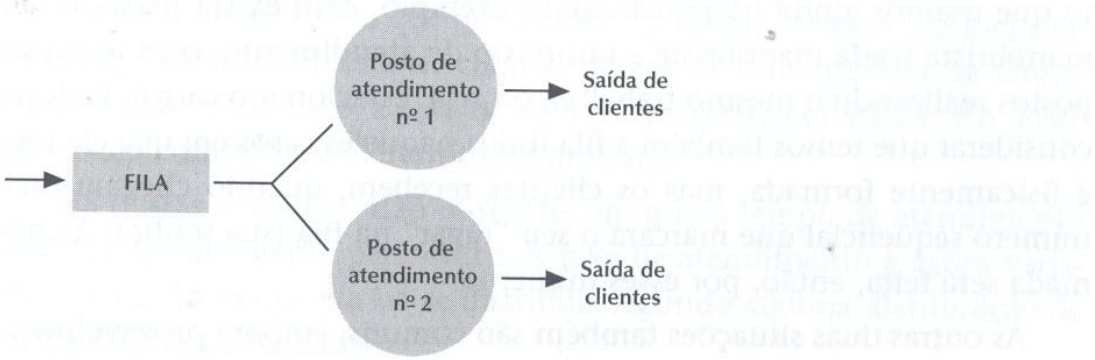

Figura 8: Canal múltiplo, atendimento único.
Fonte: Moreira 2007

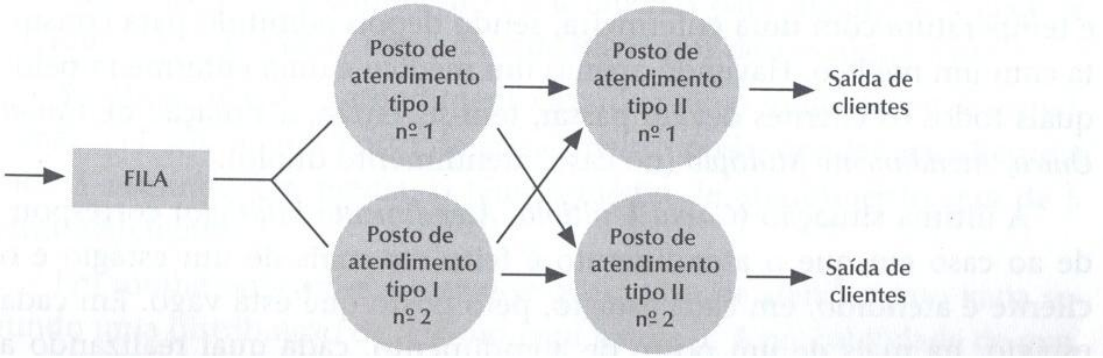

Figura 9: Canal múltiplo, atendimento múltiplo.
Fonte: Moreira 2007

Os canais heterogêneos modelam canais múltiplos com diferentes distribuições de tempo de atendimento. Porém por mais que este tipo de abordagem consiga modelar casos reais com menor distorção da realidade que o canal único, este tema não é amplamente abordado na literatura de Teoria das Filas. Estes modelos são matematicamente complexos na medida em que há mais canais de atendimentos e suas características se distanciam da abordagem clássica de canal único.

Faria (1991) aponta dois modelos, Gumbel e Krishnamoorth, para canais múltiplos. Grumbel propõe fila única com mais de um posto de atendimento onde a probabilidade de chegada é igual para todos os postos. As chegadas são representadas por uma função exponencial negativa, com uma fonte de clientes infinita já que a média de clientes na fila independe do comprimento da mesma. A taxa de atendimento é uma função exponencial, independente também do comprimento da fila. O modelo de Krishnamoorthi estabelece canais heterogêneos de serviço seguindo duas disciplinas distintas. As chegadas e atendimentos são regidas por distribuições exponenciais negativas. Porém o autor destaca que a probabilidade de chegada para cada um dos postos é diferente diante do conhecimento dos clientes sobre as taxas de atendimento, ou seja, o cliente tenderá a escolher o posto de maior taxa. Ainda, o processo de escolha pode se distinguir da forma FIFO, uma vez que é possível que o cliente sendo atendido pelo posto de menor taxa pode acabar de ser atendido após o cliente que estava imediatamente atrás e foi atendido pelo posto de menor taxa.

2.5. Disciplinas das filas

As disciplinas de filas se referem às regras que o servidor vai empregar para decidir qual será o próximo cliente da fila a ser atendido. As disciplinas mais comuns são demonstradas na Tabela 4.

Tabela 4: Regras de definição das disciplinas de filas

FIFO	Primeiro a chegar é o primeiro a sair
LIFO	Último a chegar é o primeiro a sair
SIRO	Atendimento aleatório
PRI	Atendimento por Prioridade
GD	Outra Ordem

Fonte Borba (2007)

2.6. Notação de sistema de filas

Para a descrição de um sistema com fila, será utilizada neste artigo a notação proposta por Kendall (1953), que é da forma A/B/C/D/E, onde as siglas são:

- A: Representa a distribuição do tempo entre chegadas sucessivas;
- B: Representa a distribuição do tempo de atendimento;
- C: Representa o número de postos de atendimento em paralelo;
- D: Representa a capacidade física do sistema;
- E: Representa a disciplina de atendimento.

Usualmente para definir as distribuições do tempo de chegada e do tempo de atendimento, são utilizadas as seguintes siglas para as distribuições mais comuns:

- D: Representa uma distribuição determinística ou degenerada
- M: Representa uma distribuição exponencial
- Ek: Representa uma distribuição de Erlang do tipo k
- G: Representa uma distribuição geral (não específica)

Alguns autores no caso das siglas D e E, às vezes, simplificam a notação de Kendall, com isto, admite-se um sistema de filas com capacidade ilimitada

(infinita) e com disciplina de atendimento FIFO (primeiro a chegar é o primeiro a sair).

2.7. Medidas de desempenho de sistemas de filas

De acordo com Moreira (2007), as características operacionais de uma fila são números ou indicadores de desempenho calculados para um dado modelo de filas adotado.

Dentre as medidas de desempenho citadas por diversos autores, as seguintes variáveis foram escolhidas na definição e cálculo de desempenho de um sistema de filas:

- r: abreviação de $\frac{\lambda}{\mu}$
- ρ: taxa de utilização do servidor; é uma medida de congestionamento do servidor;
- P(0): probabilidade de que o sistema esteja ocioso;
- P(n): probabilidade de que haja n clientes esperando ou sendo atendidos no sistema;
- P(N>k): probabilidade de que haja mais de "k" clientes na fila;
- Lq : número médio de clientes na fila;
- L: número médio de clientes no sistema.
- Wq: tempo médio de clientes em espera na fila;
- W: tempo médio de clientes em espera no sistema;

2.8. Modelos básicos de filas

Segundo Ferrari (2008), a maior parte dos modelos elementares de filas de espera baseia-se no processo de nascimento e morte (markoviano). No contexto das filas de espera, um nascimento corresponde à chegada de um novo cliente e uma morte corresponde à partida de um cliente.

A seguir, serão apresentados alguns modelos básicos que compõe um sistema de filas, com ênfase nos processos probabilísticos descritos em Fogliatti (2007). Nestes modelos é comum que os tempos entre chegadas e os tempos de atendimento sigam distribuições exponenciais e as nuances descritas a seguir.

- Modelo M/M/1/∞/FIFO: Existe um único posto de atendimento, não existe limitação de capacidade no espaço reservado para a fila de espera, sendo que a ordem de acesso de usuários ao serviço segue a ordem de chegada dos mesmos ao sistema (FIFO).

- Modelo M/M/1/K/FIFO: Apresenta um único posto de atendimento, porém existe uma limitação de capacidade no espaço reservado para a fila de espera, sendo que a ordem de acesso de usuários ao serviço segue a ordem de chegada dos mesmos ao sistema (FIFO). A taxa de ingresso ao sistema (λ) se difere da taxa de chegada λ para $n \geq K$, pela existência da limitação de capacidade no sistema (K).
- Modelo M/M/C/∞/FIFO: Existem "C" postos de atendimento, não existe limitação de capacidade no espaço reservado para a fila de espera, sendo que a ordem de acesso de usuários ao serviço segue a ordem de chegada dos mesmos ao sistema (FIFO).

- Modelo M/M/C/K/FIFO: Existem "C" postos de atendimento, porém existe uma limitação de capacidade no espaço reservado para a fila de espera, sendo que a ordem de acesso de usuários ao serviço segue a ordem de chegada dos mesmos ao sistema (FIFO). A taxa de ingresso ao sistema (λ) se difere da taxa de chegada λ para $n \geq K$, pela existência da limitação de capacidade no sistema (K).

As equações das medidas de desempenho dos modelos de filas apresentados acima são apresentadas na Tabela 5.

Tabela 5: Medidas de desempenho de uma fila

	M/M/1/∞/FIFO	*M/M/1/K/FIFO*	*M/M/C/∞/FIFO*	*M/M/C/K/FIFO*
r	-		$\frac{\lambda}{\mu}$	$\frac{\lambda}{\mu}$
ρ	$\frac{\lambda}{\mu}$	$\frac{\lambda}{\mu}$	$\frac{\lambda}{c\mu}$	$\frac{\lambda}{c\mu}$
P(0)	$1-\rho$	$\begin{cases}\frac{1}{K+1} \Leftrightarrow \rho = 1 \\ \frac{1-\rho}{1-\rho^{K+1}} \Leftrightarrow \rho \neq 1\end{cases}$	$(\sum_{n=0}^{c-1} r^n/n! + \frac{r^c}{c!(1-\rho)})^{-1} \quad (r/c = \rho < 1)$	$(\sum_{n=0}^{c-1} (r^n/n!) + (r^c/c!)\frac{1-\rho^{k-c+1}}{1-\rho})^{-1} \quad (\rho \neq 1),$ $(\sum_{n=0}^{c-1} (r^n/n!) + (r^c/c!)(K-c+1))^{-1} \quad (\rho = 1).$
P(n)	$\rho^n(1-\rho) \forall n \geq 0$	$\begin{cases}\frac{1}{K+1} \Leftrightarrow \rho = 1 \\ \frac{(1-\rho)\rho^n}{1-\rho^{K+1}} \Leftrightarrow \rho \neq 1\end{cases}$	$\begin{cases}P(0)\frac{r^n}{n!} \Leftrightarrow 1 \leq n \leq c \\ P(0)\frac{r^n}{c^{n-c}} \Leftrightarrow n \geq c\end{cases}$	$\begin{cases}\frac{r^n}{n!}P_0 \Leftrightarrow 1 \leq n \leq c-1 \\ \frac{r^n}{c!c^{n-c}}P_0 \Leftrightarrow c \leq n \leq k\end{cases}$
P(c)	$\left(\frac{\lambda}{\mu}\right)^k$	$\begin{cases}\frac{(K+1-k)}{K+1} \Leftrightarrow \rho = 1 \\ \frac{\rho^k(1-\rho^{K-k+1})}{(1-\rho^{K+1})} \Leftrightarrow \rho \neq 1\end{cases}$	$1-\sum_{n=0}^{k-1} P(n)$	$1-\sum_{n=0}^{k-1} P(n)$
	$\frac{\rho^2}{(1-\rho)}$	$L-1+P_0$	$\frac{P(0)cr^{c+1}}{c!(c-r)^2}$	$\frac{s^s\rho^{s+1}}{s!(1-\rho)^2}[1-\rho^{k-s}-(1-\rho)(k-s)\rho^{k-s}]P(0)$
	$\frac{\rho}{1-\rho}$	$\begin{cases}\frac{K}{2} \Leftrightarrow \rho = 1 \\ \frac{\rho[1+K\rho^{K+1}-\rho^K(K+1)]}{(1-\rho)(1-\rho^{K+1})} \Leftrightarrow \rho \neq 1\end{cases}$	$r+\left[\frac{r^{c+1}c}{c!(c-r)^2}\right]P(0)$	$L_q + \lambda(1-p_k)/\mu = L_q + r(1-p_k).$
W_q	$\frac{\rho}{\mu-\lambda}$	$\frac{L_q}{\lambda(1-P_K)}$	$\left[\frac{r^c\mu}{(c-1)!(c\mu-\lambda)^2}\right]P(0)$	$\frac{L_q}{\lambda(1-P_K)}$
W	$\frac{1}{\mu-\lambda}$	$\frac{L}{\lambda(1-P_K)}$	$\frac{1}{\mu}+\left[\frac{r^c\mu}{(c-1)!(c\mu-\lambda)^2}\right]P(0)$	$\frac{L}{\lambda(1-P_K)}$

Fonte: Fogliatti