Avaliação e Desempenho
Aula 18

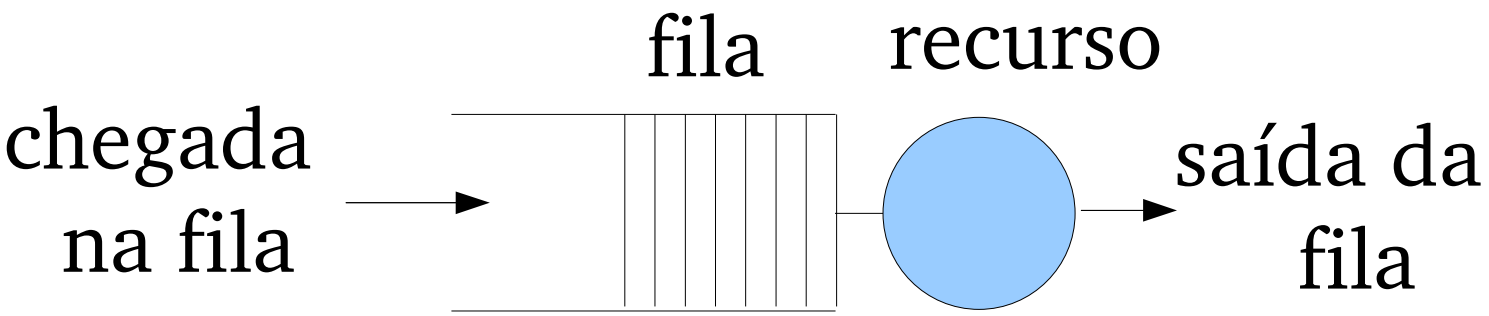
Aula passada

- Fila com buffer finito
- Fila com buffer infinito
- Medidas de interesse: vazão, número médio de clientes na fila, taxa de perda.

Aula de hoje

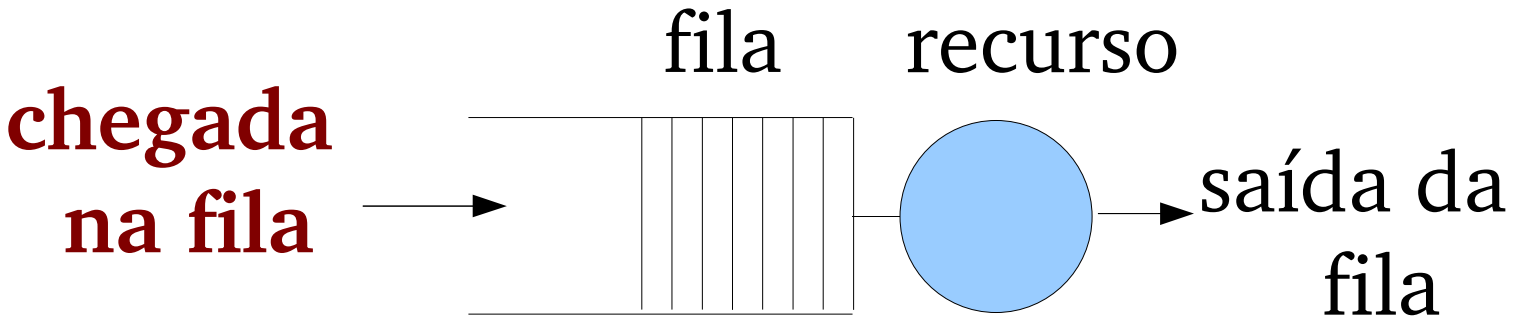
- Parâmetros de uma fila
- Medidas de desempenho
- Cálculo do tempo de espera
- Resultado de Little

Parâmetros da Fila



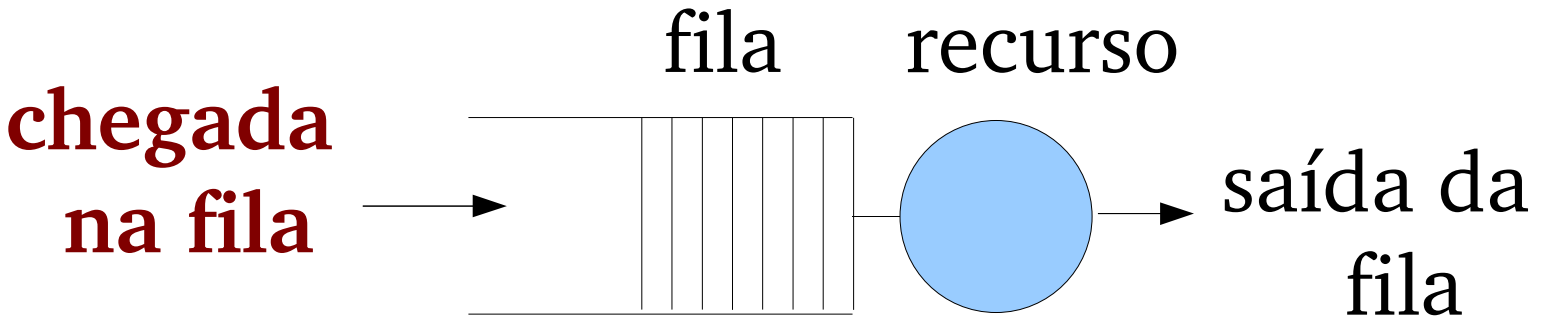
- Processo de chegada (como os pedidos chegam ao recurso)
- Processo de serviço (como os pedidos são servidos)
- Capacidade de armazenamento da fila
- Número de recursos (estações de serviço)
- Política de atendimento (como escolher o próximo da fila)

Fila – Processo de Chegada



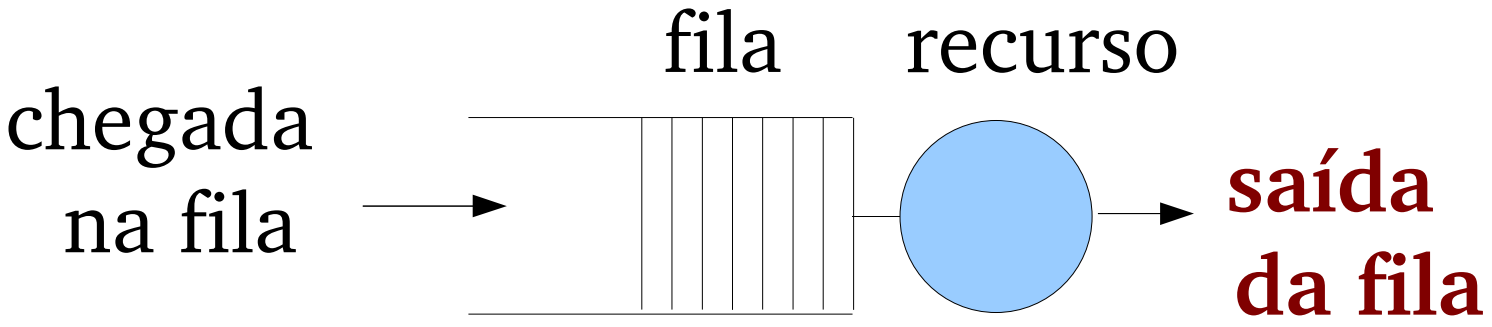
- Como definir *processo de chegada (demanda)*?
- No banco, como pessoas chegam ao banco
- Na rede, como pacotes chegam ao roteador
- No disco, como pedidos de leitura chegam
- **Abstração matemática**

Fila – Processo de Chegada



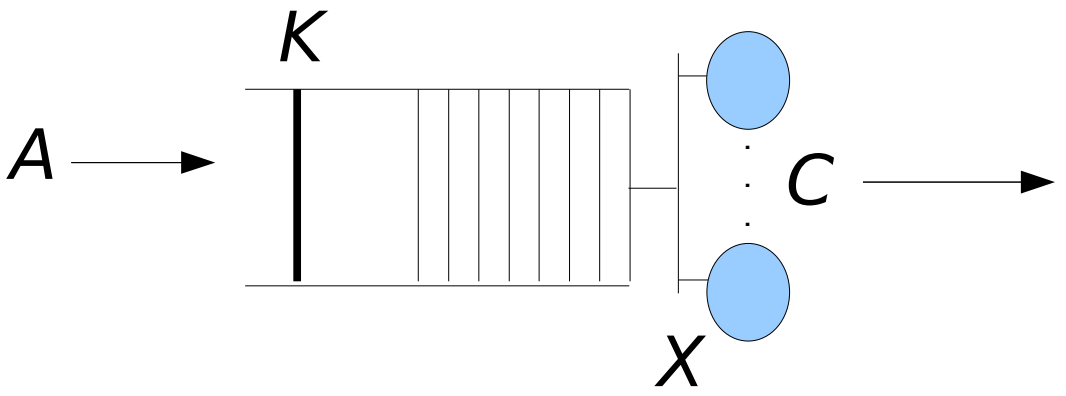
- ***Modelo matemático*** para abstrair processo de chegada
- Chegada não é determinística
- Necessidade de representação aleatória
- Ex. processo de Poisson

Fila – Processo de Serviço



- Como definir *processo de serviço*?
- Quanto tempo leva para atender um pedido?
- Depende do sistema, geralmente aleatório
- *Modelo matemático* para abstrair o processo de serviço (distribuição exponencial, lognormal, etc.)

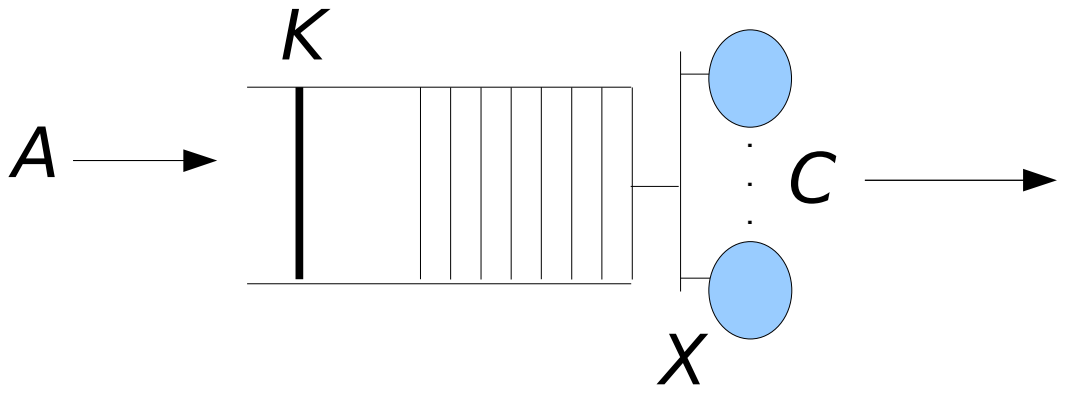
Parâmetros de uma Fila



- A : intervalo de tempo entre chegadas
- X : tempo de serviço (de 1 pedido)
- K : capacidade de armazenamento da fila (infinita?)
- C : número de estações de serviço
- P : política de atendimento dos elementos em fila

A, X geralmente são variáveis aleatórias

Notação de Filas

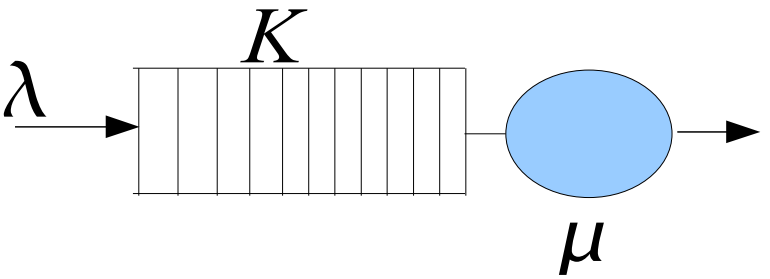


$$A / X / C / K - P$$

- distribuição do tempo entre chegadas
- distribuição do tempo de serviço
  - M = exponencial (memoryless)
  - D = determinístico
  - Er = Erlang
- número de estações de serviço
- capacidade de armazenamento
  - *default*: infinito
- política de atendimento
  - *default*: FIFO

- Exemplo: M/M/1 ← Qual é a fila?

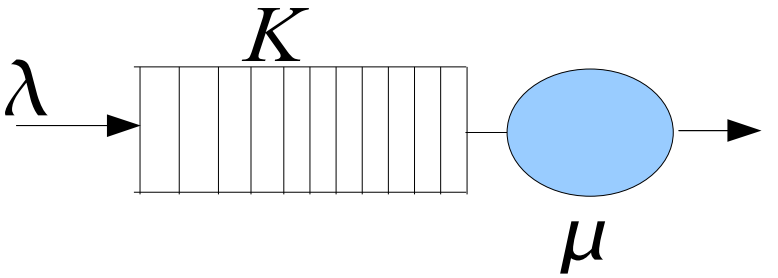
Medidas de Desempenho em Filas

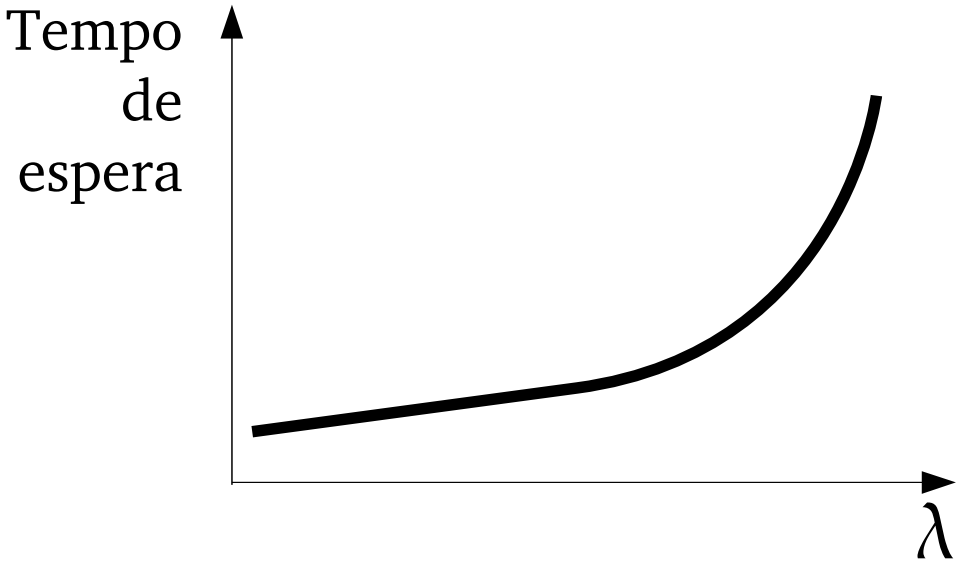


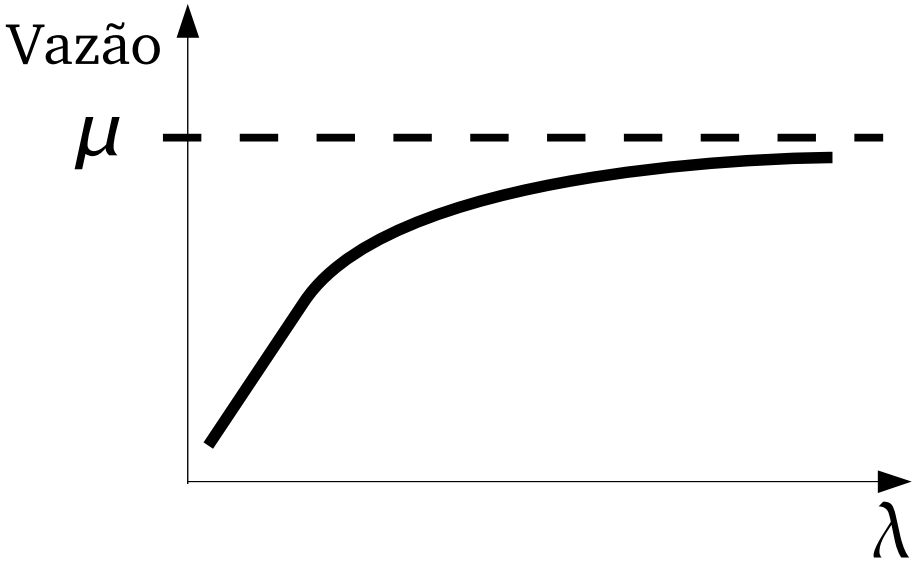
- **Utilização:** fração de tempo que o servidor está ocupado
- **Tempo de espera:** quantidade de tempo que cada elemento espera na fila
- **Vazão (*Throughput*):** quantidade de elementos servidos pela fila por unidade de tempo
- **Fração de descarte:** fração de elementos descartados por falta de espaço na fila

Medidas de Desempenho em Filas

- Comportamento típico



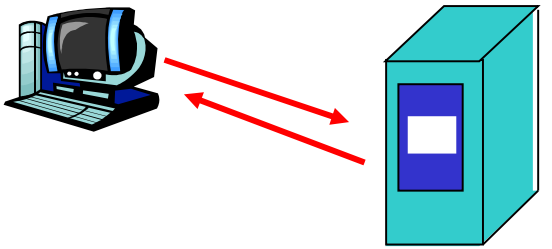




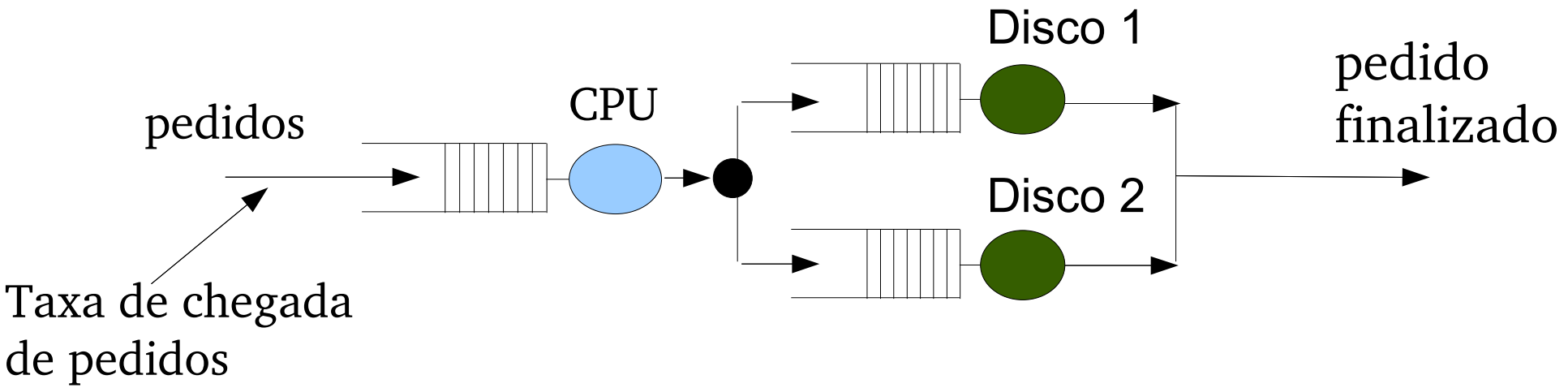
- ***Tradeoff*** entre tempo de resposta e vazão

Modelagem do Servidor Web

- Pedidos de objeto chegam ao servidor
- Processar pedidos
  - CPU, discos

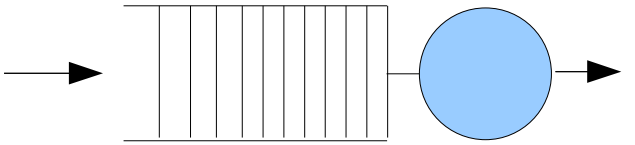


Servidor Web



- Modelagem através de filas
- Qual é o tempo médio de espera de um pedido?

Avaliando o Desempenho de uma Fila



Problema

- Dado uma fila, qual seu desempenho?

Como assim?

- Parâmetros da fila (demanda, capacidade, etc.)

Como assim?

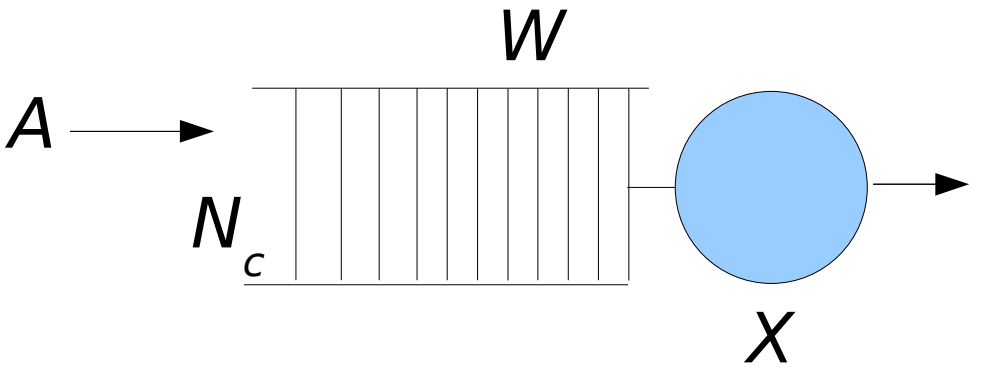
- Medida: Tempo médio de espera no sistema

Vamos calcular isto!

Tempo de Espera

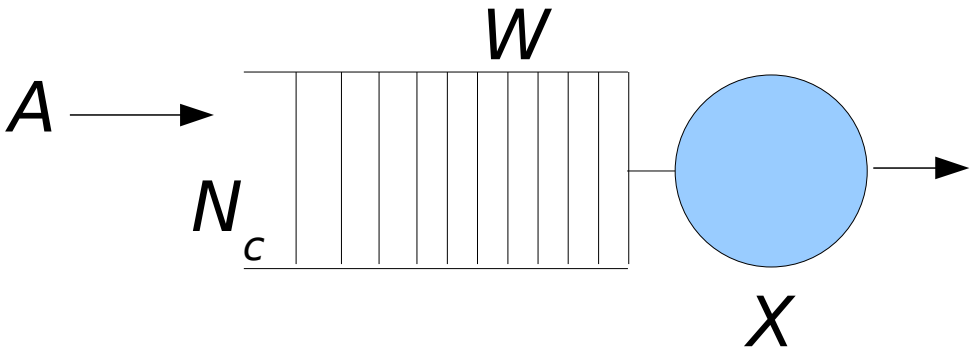
- Como calcular o tempo de espera em uma fila?
  - tempo decorrido desde o instante de chegada até o instante de saída
- **IDÉIA: Decompor o tempo de espera em tempos menores**
- Tempo para finalizar o pedido sendo atendido no instante de chegada
- Tempo para atender cada um dos pedidos da fila
- Tempo para atender o pedido que acabou de chegar

Variáveis Aleatórias de uma Fila



- N_c : número de pedidos na fila quando nosso pedido chega (incluindo o que está sendo atendido)
- R : tempo residual (tempo para finalizar atendimento do pedido sendo atendido)
- X_i : tempo necessário para atender o i-ésimo pedido da fila ($i=1, 2, ..., N_c - 1$)
- W : tempo de espera de um pedido (do instante de chegada até o instante de saída)

Tempo de Espera

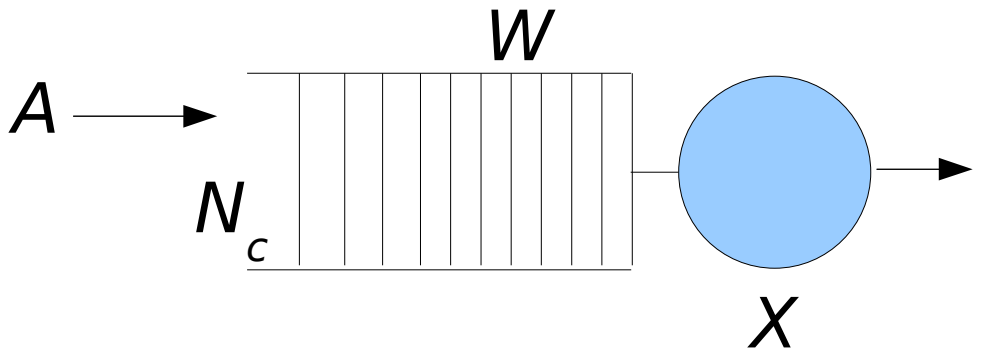


- W: tempo de espera do pedido que chegou
  - N_c pedidos no sistema
  - 1 em atendimento + $N_c - 1$ na fila
- Quanto vale W?

$$W = R + X_1 + X_2 + \ldots + X_{N_c-1} + X$$

- R: tempo residual do elemento em atendimento
- X_i: tempo de serviço do i-ésimo elemento da fila
- X: tempo de serviço do elemento que acabou de chegar

Tempo Médio de Espera



- Aplicar *valor esperado* – E[.]
- Tempo **médio** de espera: **E[W]**

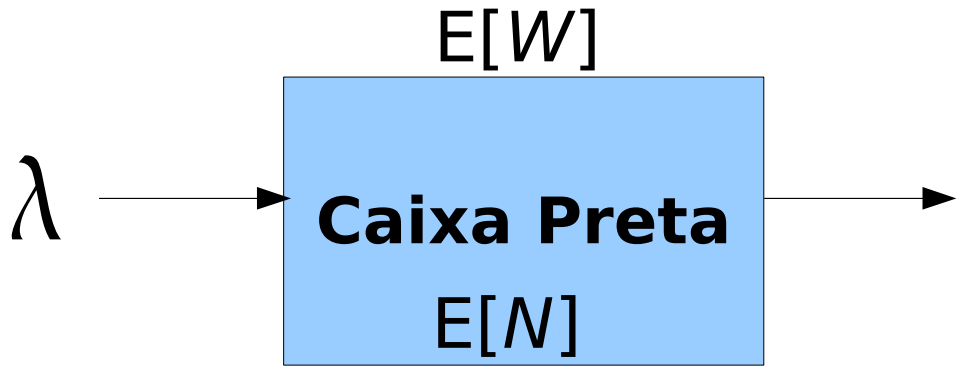
$$E[W] = E[R + X_1 + X_2 + \ldots + X_{N_c-1} + X]$$

$$E[W] = E[R] + E[X_1] + \ldots + E[X_{N_c-1}] + E[X]$$

$$E[W] = E[R] + E[N_c]E[X]$$

X_i são idênticamente distribuídas

Resultado de Little

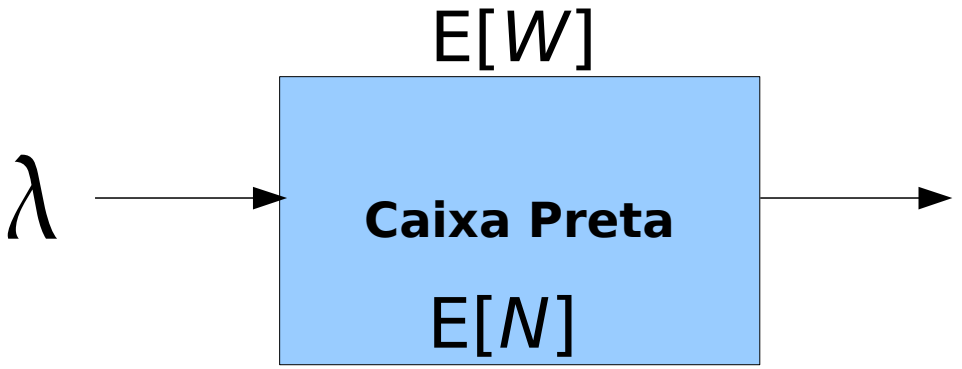


- λ : taxa média de chegada
- E[W] : tempo médio que cada pedido permanece dentro da caixa preta
- E[N] : número médio de pedidos dentro da caixa (depois de um tempo muito grande)
- Qual é a relação entre eles?

$$E[N] = \lambda E[W]$$

Vamos provar isto hoje!

Resultado de Little



- Número médio dentro do sistema é igual ao produto da taxa média de chegada pelo tempo médio de permanência no sistema
- Intuitivo, mas poderoso (não depende de nenhuma distribuição)
- Exemplo:

E[W] = 3.5 minutos

$\lambda = 2$ clientes por minuto

E[N] = ?

Resolvendo o Tempo Média de Espera

- Duas equações:

$$E[W] = E[R] + E[N_c]E[X]$$

$$E[N] = \lambda \; E[W]$$

- Suposição (I) : $E[N] = E[N_c]$
- Suposição (II) : $E[R] = E[X]$
- Substituindo nas equações acima, temos

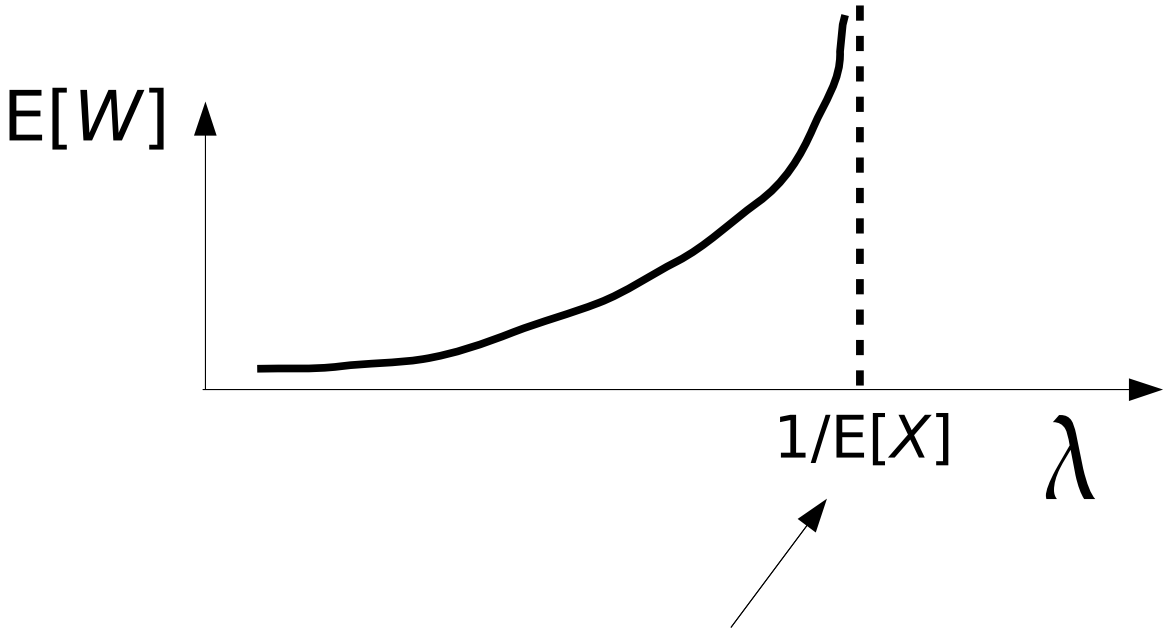
$$E[W] = E[X] + \lambda E[W]E[X]$$

$$E[W] = \frac{E[X]}{1 - \lambda E[X]}$$

equação em função dos parâmetros da fila

Gráfico do Tempo Médio de Espera

$$E[W] = \frac{E[X]}{1 - \lambda E[X]}$$



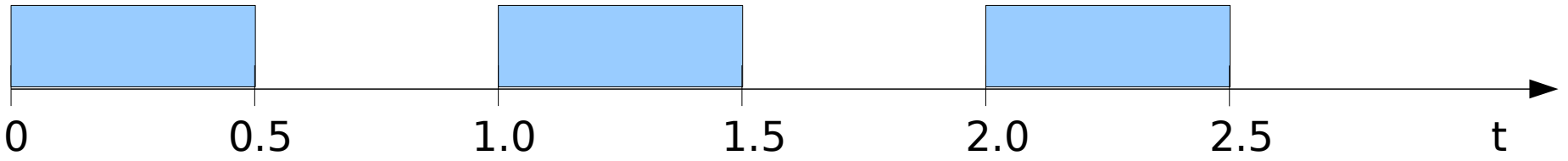
Suposição I: $E[N] = E[N_c]$

- Duas variáveis (medidas)
  - Número de elementos no sistema
  - N – em um instante de tempo qualquer
  - N_c – no instante de uma chegada
- Em geral, $E[N]$ é diferente de $E[N_c]$

Exemplo?

Exemplo

- Suponha sistema de fila determinístico
  - tempo entre chegadas 1s, tempo de serviço 0.5s
- Evolução do sistema no tempo?



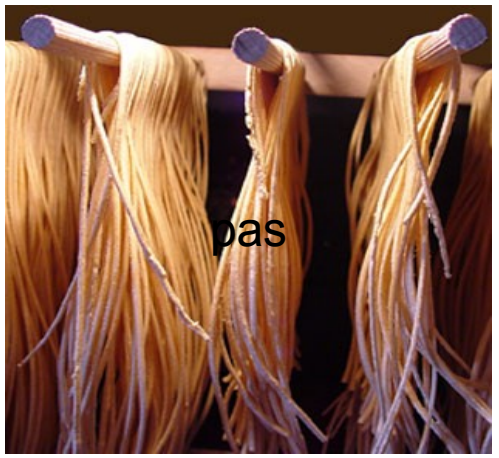
- Quanto vale E[N] e $E[N_c]$?
  - E[N] = 0.5
  - $E[N_c] = 0$

→ **Diferentes!**

Suposição I

- $E[N] = E[N_c]$
- Quando isto é verdade?

PASTA → **P**oisson **A**rrivals **S**ee **T**ime **A**verages



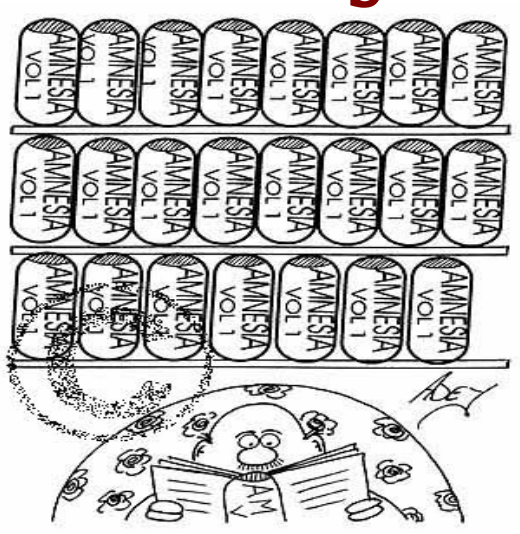
• Cliente que chega encontra, em média, a mesma situação que um observador em um tempo aleatório.
• Intuição: chegadas de Poisson ocorrem aleatoriamente no tempo.

- Quando o processo de chegada é *Poisson*
  - $E[N] = E[N_c]$

Suposição II

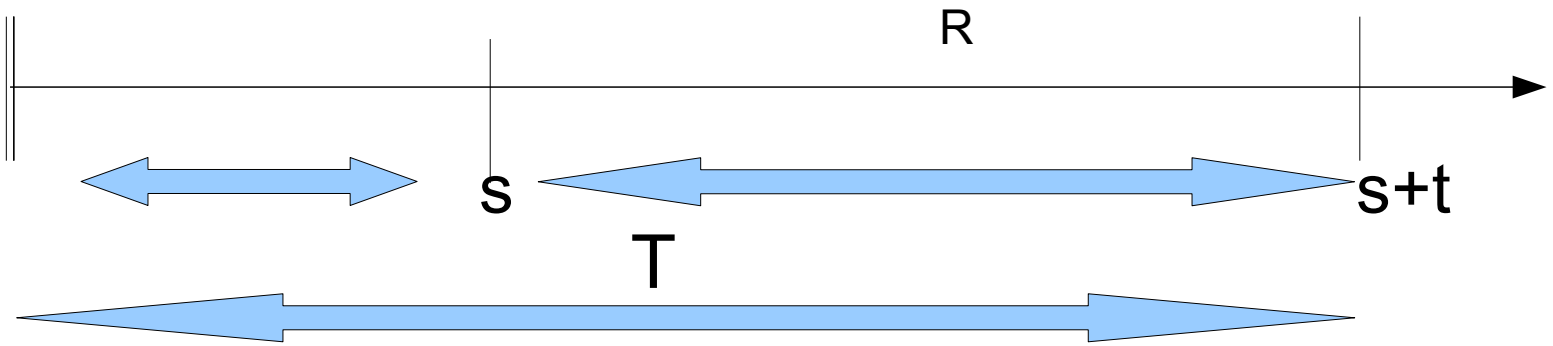
- E[R] = E[X]
- Quando isto é verdade?

Memoryless → Distribuição não tem memória



- Quando o tempo de serviço é *exponencial*

Propriedade Memoryless



$$P[R > t | T > s] = ???$$

$$s, t > 0$$

$$P[T - s > t | T > s]$$

$$P[T > t + s | T > s]$$

Propriedade Memoryless

- Propriedade de ***memoryless***
  - Distribuição da *probabilidade condicional* é igual a distribuição da probabilidade original (para o restante do tempo)

$$P[T > s+t \mid T > s] = P[T > t] \qquad s, t > 0$$

- Exemplo

$$P[T > 4\,0 \mid T > 3\,0] = P[T > 1\,0]$$ ← **Correto**

$$P[T > 4\,0 \mid T > 3\,0] = P[T > 4\,0]$$ ← **Errado!**

- A chance de um evento não ocorrer nos próximos 10 segundos é igual a dos primeiros 10 segundos!

Provando a Propriedade Memoryless para Exponencial

- Distribuição exponencial (parâmetro λ)

$$P[T>t]=e^{-\lambda t}$$

- Propriedade

$$P[T>s+t\,|\,T>s]=P[T>t]$$

- Prova (aplicar definições)

$$P[T>s+t\,|\,T>s]=\frac{P[(T>s+t)\cap(T>s)]}{P[T>s]}=\frac{P[T>s+t]}{P[T>s]}=\frac{e^{-(s+t)\lambda}}{e^{-s\lambda}}=$$

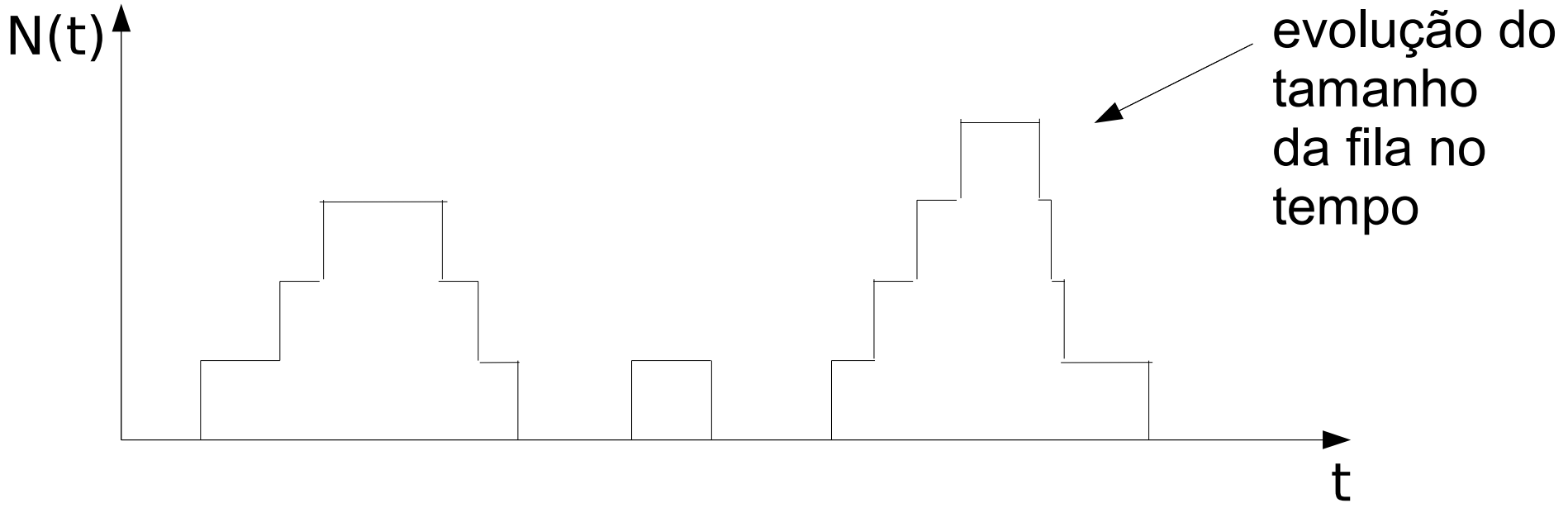
$$=e^{-t\lambda}=P[T>t]$$

Voltando a Suposição II

- E[R] = E[X] ?
- Quando o tempo de serviço é memoryless
  - exponencial (e somente exponencial)

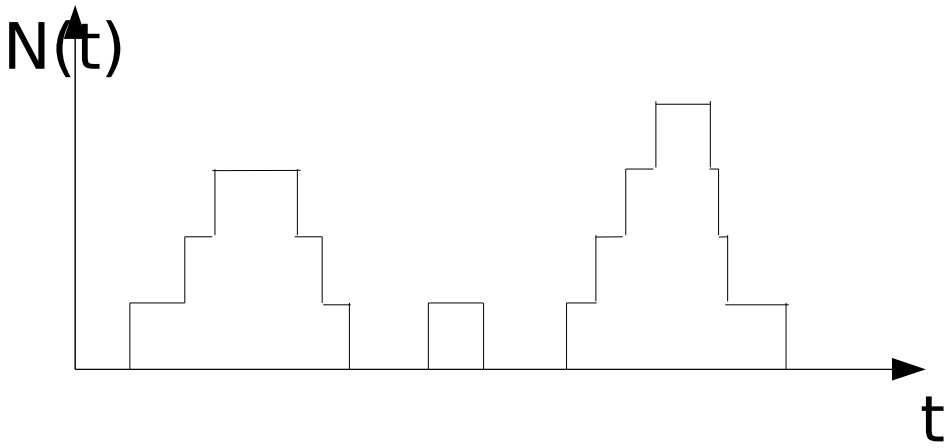
- "pedido em atendimento começa a ser atendido no momento da chegada"

Calculando as Medidas de Interesse de uma Fila



- Quanto vale E[N]?
  - tamanho médio da fila (no tempo)
- Quanto vale E[W]?
  - tempo médio de espera (de um pedido)

Calculando E[N]



$$E[N(T)] = \frac{\sum_{i=0}^{\infty} i * T_i(T)}{T}$$

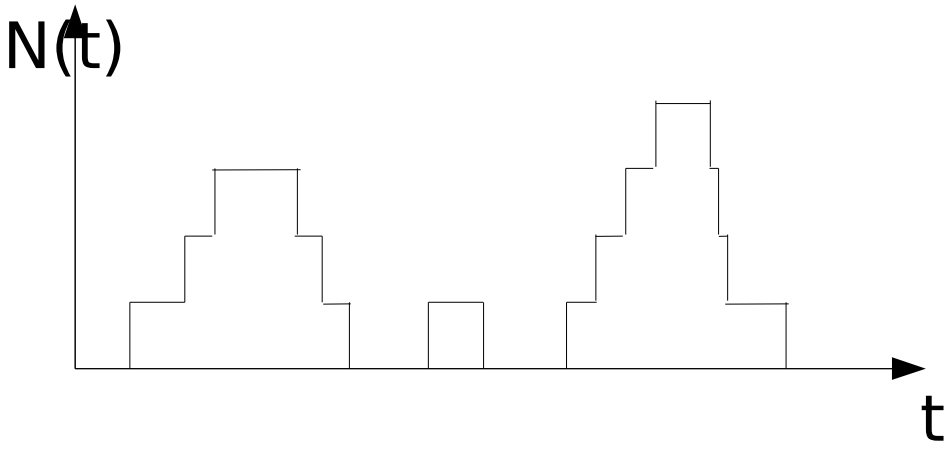
- onde T é um tempo qualquer e $T_i(T)$ é o tempo que a fila permanece com i elementos
- Observação: $\sum_{i=0}^{\infty} T_i(T) = T$

$$\sum_{i=0}^{\infty} i * T_i(T) = \int_{t=0}^{t=T} N(t)\,dt$$

- Logo:

$$E[N(T)] = \frac{\int_{t=0}^{t=T} N(t)\,dt}{T}$$

Calculando E[W]

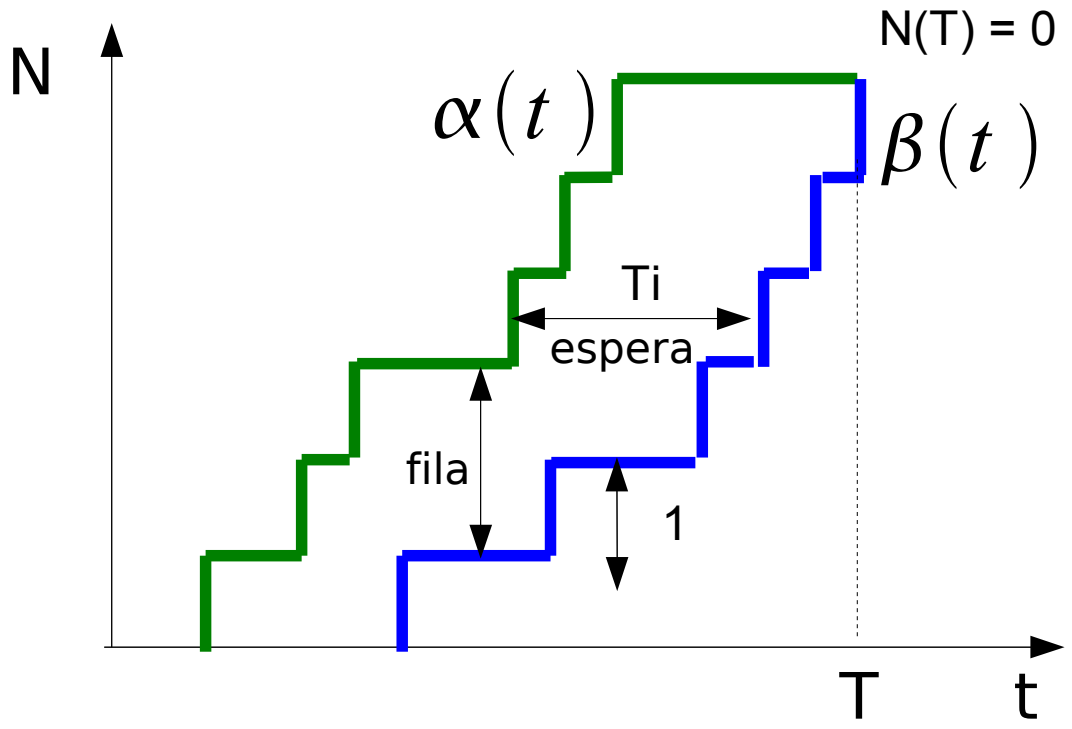


$$E[W] = \frac{\sum_{i=1}^{n} W_i}{n}$$

- onde n é um número de saídas do sistema qualquer e W_i é o tempo de espera da i-ésima saída
- Observação: $W_i = S_i - C_i$
  - onde C_i e S_i são os instantes de chegada e saída do i-ésimo elemento
- Logo:

$$E[W] = \frac{\sum_{i=1}^{n} S_i - C_i}{n}$$

Outra Evolução da Fila

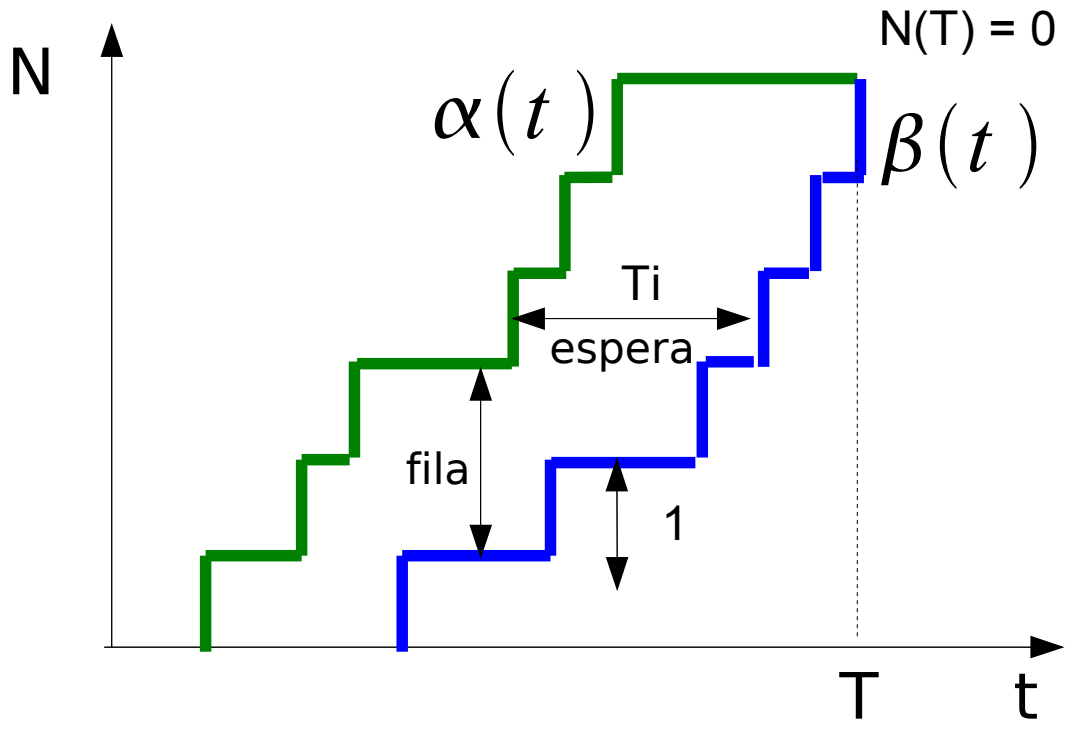


$\alpha(t)$: número total de chegadas até t

$\beta(t)$: número total de saídas até t

- Quanto vale N(t)? → $\alpha(t) - \beta(t)$
- Área do gráfico: → $\int_T \alpha(t) - \beta(t)\, dt = \int_T N(t)\, dt$
- A mesma área pode ser expressa por: → $\sum_{i=1}^{\alpha(T)} T_i$

Provando Resultado de Little



$\alpha(t)$: número total de chegadas até t

$\beta(t)$: número total de saídas até t

- Igualando as duas áreas:

$$\frac{\int_T N(t)dt}{T} = \frac{\alpha(T)}{T}\frac{\sum_{i=1}^{\alpha(T)} T_i}{\alpha(t)}$$

$$E[N] = \lambda E[W] \longleftarrow \text{Resultado de Little!}$$