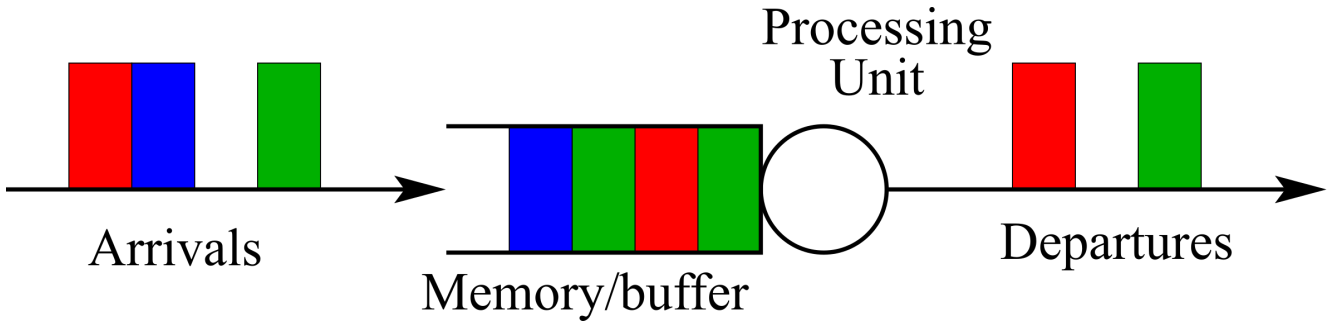
Classification of single queues: many features



Specifications:

- Deterministic vs stochastic models
- Fluid ($\mathbb{R}_+$) vs discrete ($\mathbb{N}$) models over time and data
- Rhythm of arrivals (periodic, Poisson, unknown, ...)
- Rhythm of departures / services
- Service/scheduling policy (FIFO, priorities, Round-robin ...)
- Sizes of buffer / of packets ...

Classification of single queues: Kendall's notation (1953)

Kendall's notation for a single queue: $A/B/c/K/N/P$ where

- A denotes the distribution of inter-arrival times
- B denotes the distribution of service times
- c is the number of servers (possibly ∞)
- K is the size of the buffer (possibly ∞)
- N is the size of the population served (possibly ∞)
- P is the scheduling policy (e.g. FIFO)

Some usual notations:

- M (Markov): i.i.d. random times of exponential law
- D (Deterministic): constant time (one fixed value)
- G (General): i.i.d. random times of unknown law

By default: $K = N = \infty$, $P =$ FIFO, e.g. $\mathrm{M}/\mathrm{M}/c = \mathrm{M}/\mathrm{M}/c/\infty/\infty/\mathrm{FIFO}$

Classical questions about queues

Some classical questions:

- What is the average waiting time for a client which arrives ?
- What fraction of time is the server busy ?
- What is the distribution of the queue length ?
- What is the shape of the output traffic ?

⚠ **Beware of the definition of average**, e.g. it may mean

- empirical average of a parameter over one or several trajectories, observed over finite or infinite interval of time (*time average over finite or infinite horizon*)
- average of a parameter with regard to its probability distribution at some fixed time, the system is sometimes assumed to be at a *stationary state* (if it exists)

Classical questions: M/M/1 queues

Examples: M(λ)/M(μ)/1 queue at the stationary state/mode.
Math model: Continuous Time Markov Chain (requires $\lambda < \mu$ to have an invariant distribution)
Some classical questions:

- Distribution of the queue length (buffer+server) = geometric invariant distrib
- Average number of clients (buffer+server) = $\lambda/(\mu - \lambda)$
- Average waiting time (including service time) = $1/(\mu - \lambda)$
- Shape of the output traffic = Poisson of intensity λ
- Fraction of time the server is busy = λ/μ

Classical questions: G/G/1 queues

Examples: G/G/1 queue

Math model: not markovian in general (requires results about renewal processes)

Definition (Renewal process)

A renewal process $(T_n)_{n\in\mathbb{N}}$ *of rate* μ *is defined by* $T_0 = 0$ *and for* $n \geq 1$, $T_n = I_1 + \cdots + I_n$ *where* $(I_n)_{n\geq 1}$ *i.i.d. r.v. with values in* $\mathbb{R}_+$ *such that* $\mathbb{E}(I_0) = 1/\mu$. *It is associated with its counting process* $N_t = \max\{n \mid T_n \leq t\}$ *for* $t \in \mathbb{R}_+$.

Theorem (LLN for renewal processes)

Let (T_n) *be a renewal process of rate* μ *and let* (N_t) *be its counting process. Then*

$$\frac{N_t}{t} \xrightarrow[t\to\infty]{a.s.} \mu \text{ and } \frac{\mathbb{E}(N_t)}{t} \xrightarrow[t\to\infty]{} \mu$$

Classical questions: G/G/1 queues

Proposition (Stability of G/G/1 queues)

In a queue with finite initial size, where arrivals follows a renewal process of rate λ and service times are i.i.d. with average $1/\mu$, if $\lambda < \mu$, then the queue size will reach 0 in finite time a.s.

Proof: Let T_n arrival time of new client n. LLN ensures $\frac{T_n}{n} \xrightarrow{a.s.} \frac{1}{\lambda}$. Let Z_0 time to serve all clients initially in the queue, let s_i time to serve new client i, $\mathbb{E}(s_i) = 1/\mu$. Suppose that server always remains busy, then new client n will leave at time $Z_0 + S_n$ with $S_n = s_1 + \cdots + s_n$. LLN ensures $\frac{Z_0+S_n}{n} \xrightarrow{a.s.} \frac{1}{\mu}$. Since $1/\mu < 1/\lambda$, it would mean that for large n, client n leaves the queue before its arrival. Impossible.

Little's law: general statement

Little's law for a single queue (in short)

$$\text{average nb of clients } \overline{N} \text{ in a queue}$$
$$=$$
$$\text{average arrival rate } \lambda \times \text{average sojourn time } \overline{T}$$

Use: intuitive relation avoiding sometimes complex calculations

Warning: ⚠ *average* yet to define ...

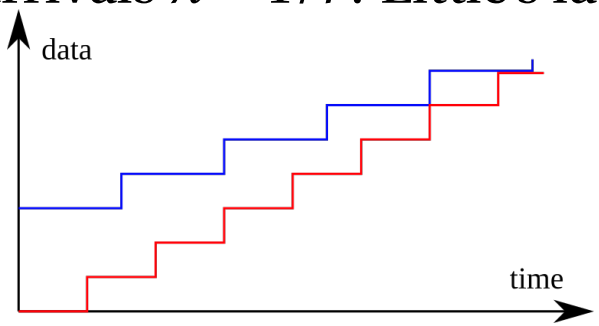
Examples: estimating the capacity of a router

Router model: 1 server + 1 buffer
Experiment: expose the router to an input traffic with throughput rate = 100 packets/sec and measure the number of packets in the whole router and in the sole buffer. Each packet of size 900 bytes.
Measures: no packet lost, average nb of packets in the system = 5.5, in the buffer = 4.8
Analysis:

- Average nb of packets in the server = $5.5 - 4.8 = 0.7$ packets
- Average arrival rate at the server = 100 packets/sec (no loss)
- Average sojourn time = $0.7/100 = 7 \times 10^{-3}$ sec (with Little's law)
- Router capacity = $1/7.10^{-3}$ packets/sec $\approx 1Mb/s$

Examples: D/D/1 and M/M/1 queues

Little's law for D(r)/D(s)/1: constant inter-arrival time $r \in \mathbb{R}_+$ and service time $s \in \mathbb{R}_+$. Queue size tends to ∞ iff $r < s$. If $r \geq s$ and even if initially the queue is not empty, then asympotically at most 1 client in the queue, with waiting time = service time s and during time $r - s$ empty queue before next client arrives. Thus $\overline{N} = s/r$, $\overline{T} = s$ and the rate of arrivals $\lambda = 1/r$. Little's law works.



Little's law for M(λ)/M(μ)/1: at stationary state when $\lambda < \mu$

$$\overline{N} = \frac{\lambda}{\mu - \lambda} = \lambda \times \frac{1}{\mu - \lambda} = \lambda \times \overline{T}$$

Deterministic general case: finite horizon

Pseudo-inverse: $f^{(-1)}(n) = \inf\{t | f(t) \geq n\}$, useful for staircase functions A et B $\nearrow$ and right continuous.

Definitions & notations:

- $A(t)$: nb $n \in \mathbb{N}$ of clients *arrived* between time 0 and $t \in \mathbb{R}_+$.
- $B(t)$: nb $n \in \mathbb{N}$ of clients *left* between time 0 and $t \in \mathbb{R}_+$.
- $N(t) \stackrel{\text{def}}{=} A(t) - B(t)$: nb of clients in the queue at time t.
- $T(n) \stackrel{\text{def}}{=} B^{(-1)}(n) - A^{(-1)}(n)$: delay for client n° n (if FIFO).

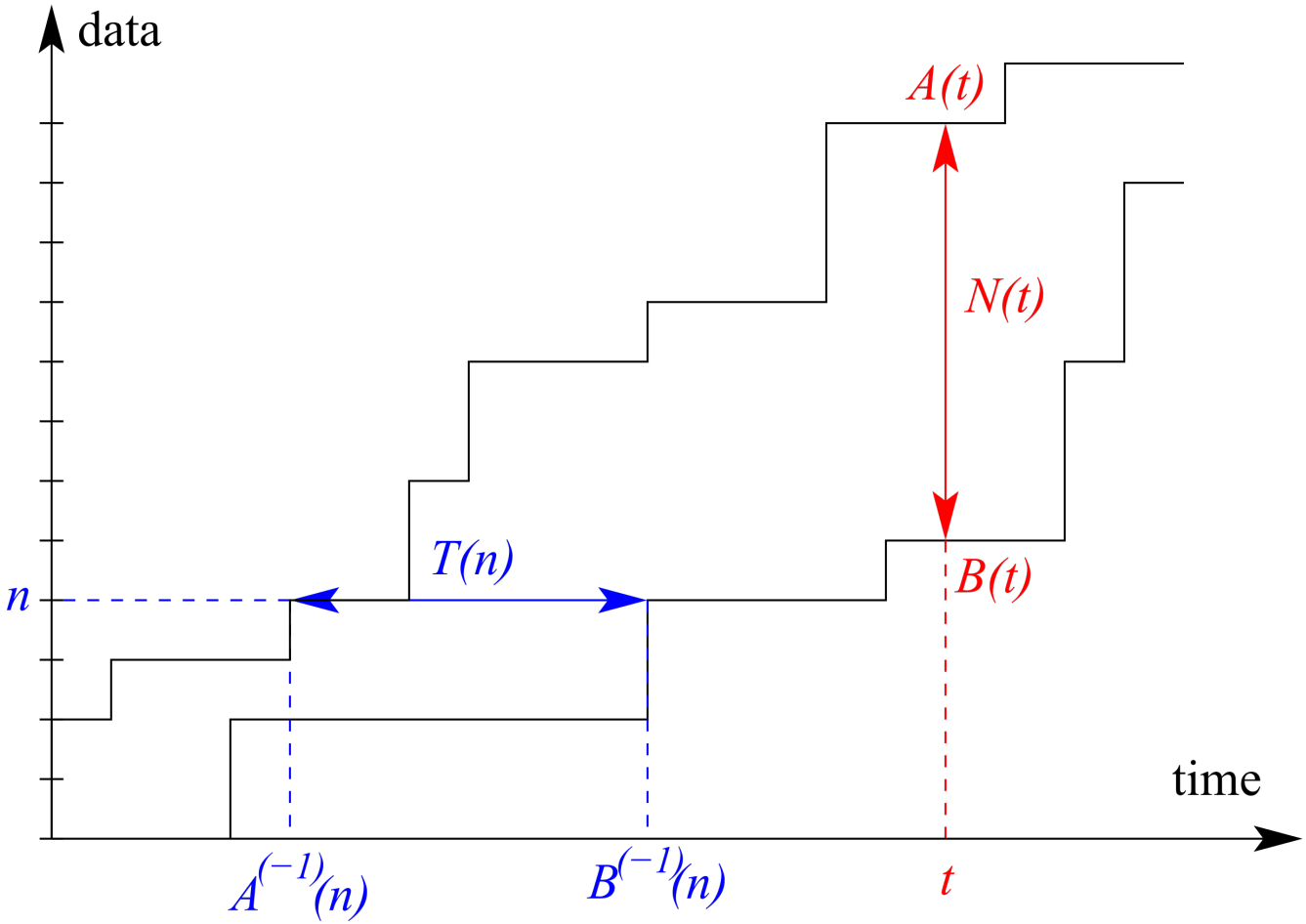
Averages: "finite horizon" = over $[0, t]$,

- $\lambda \stackrel{\text{def}}{=} A(t)/t$ (arrival rate)
- $\overline{T} \stackrel{\text{def}}{=} \frac{1}{A(t)} \sum_{i=1}^{A(t)} T(i)$ (average delay)
- $\overline{N} \stackrel{\text{def}}{=} \frac{1}{t} \int_{x=0}^{t} N(x)dx$ (average load)

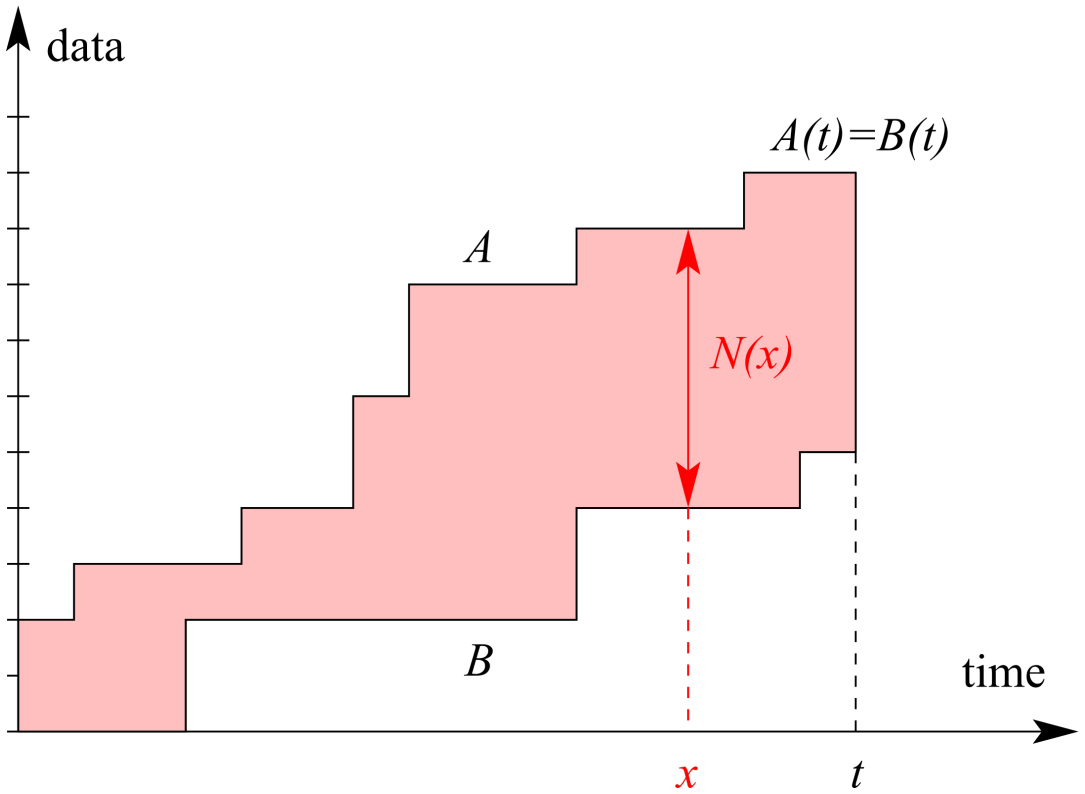
Theorem (Little 1961)

If $A(t) = B(t)$, *then* $\overline{N} = \lambda \overline{T}$.

Deterministic general case: notations

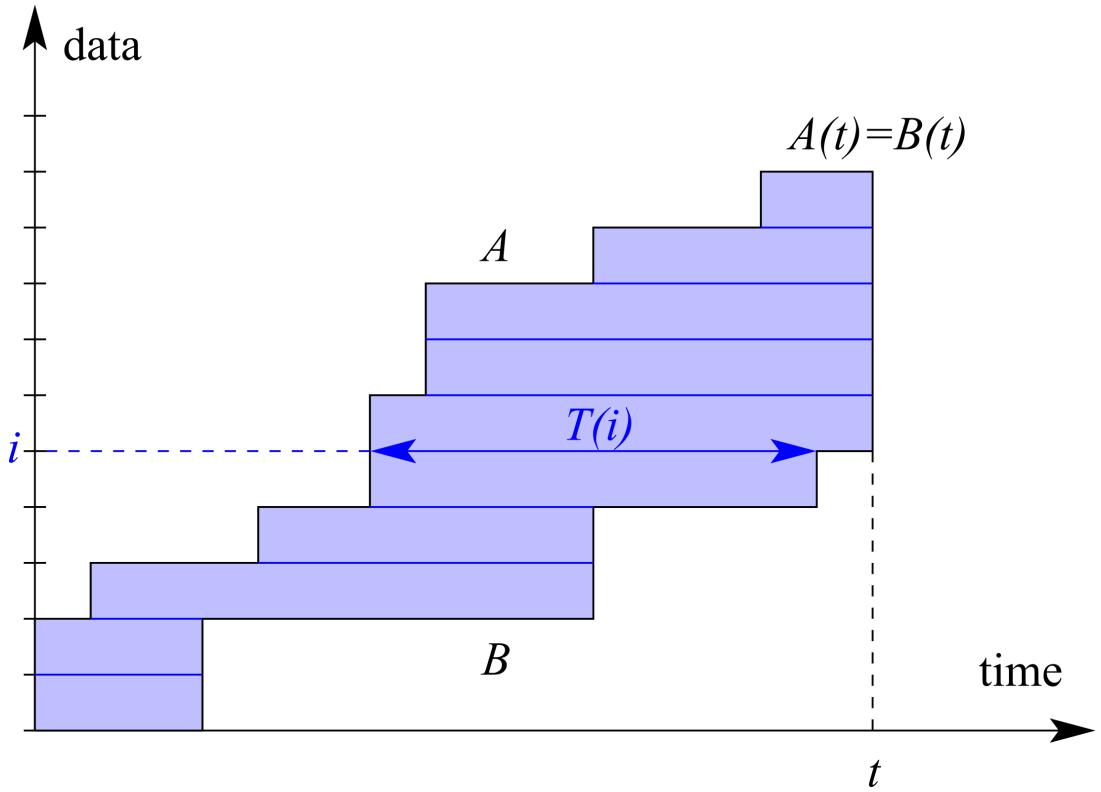


Deterministic general case: finite horizon



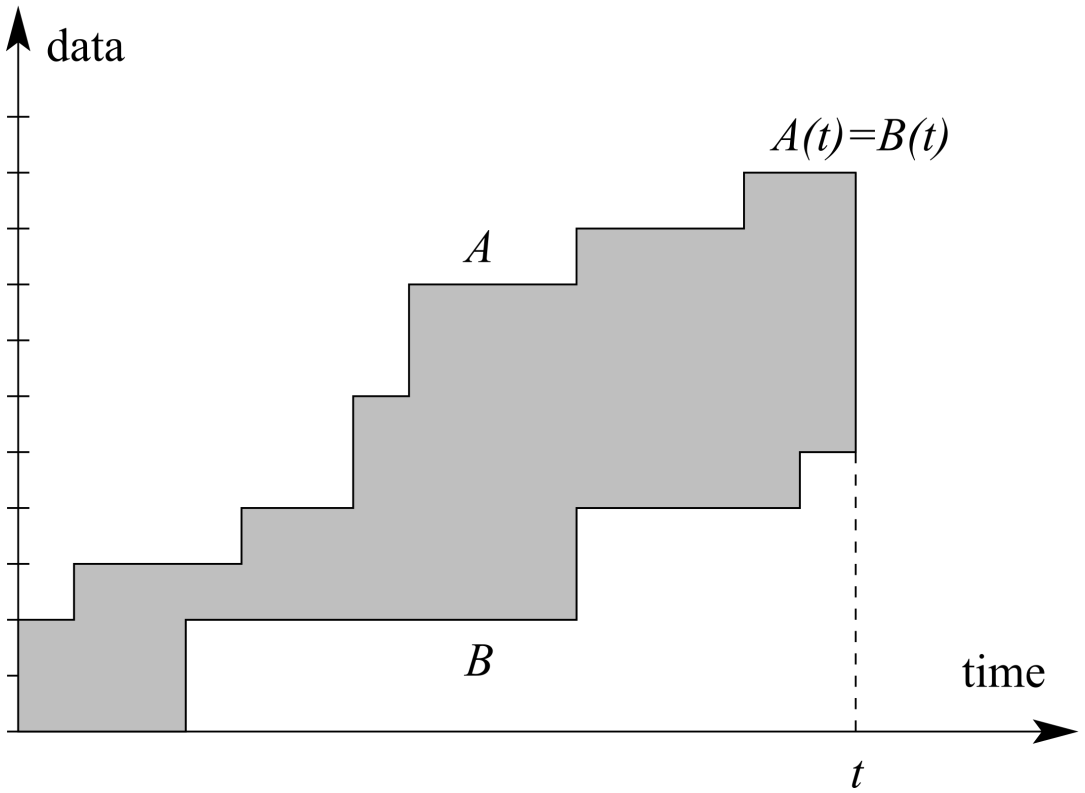
$$\int_{x=0}^{t} N(x)\,dx =$$

Deterministic general case: finite horizon



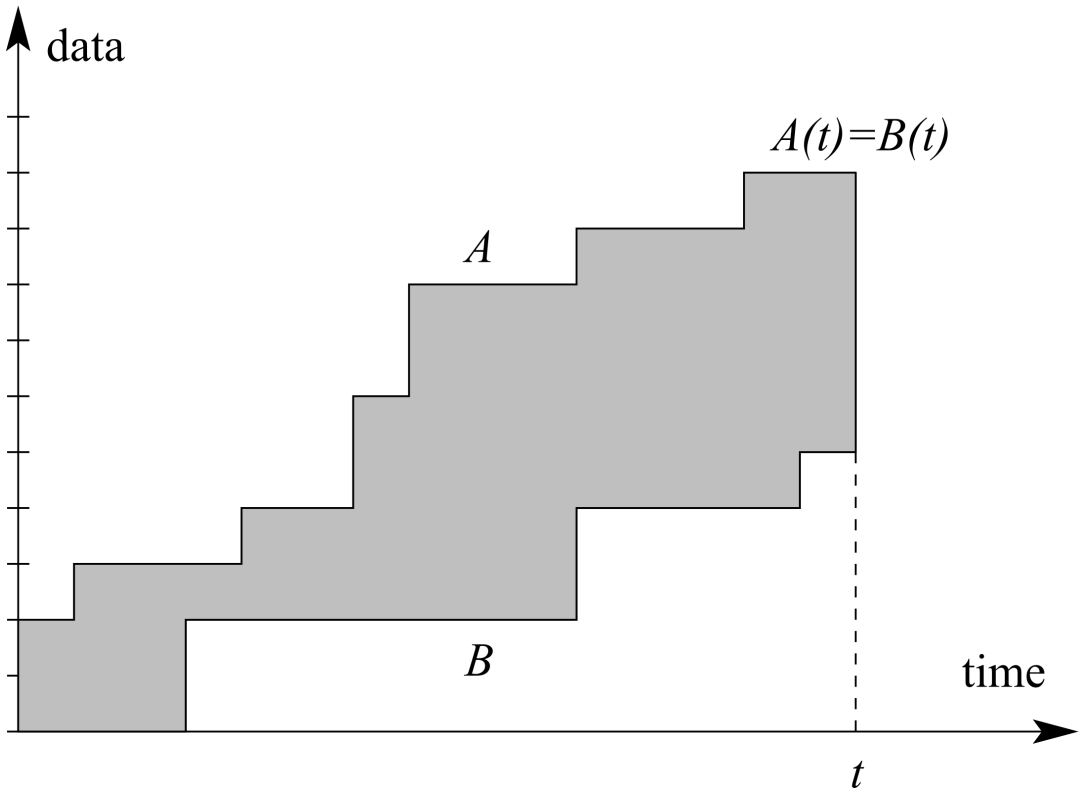
$$\int_{x=0}^{t} N(x)\,dx = \sum_{i=1}^{A(t)} T(i)$$

Deterministic general case: finite horizon



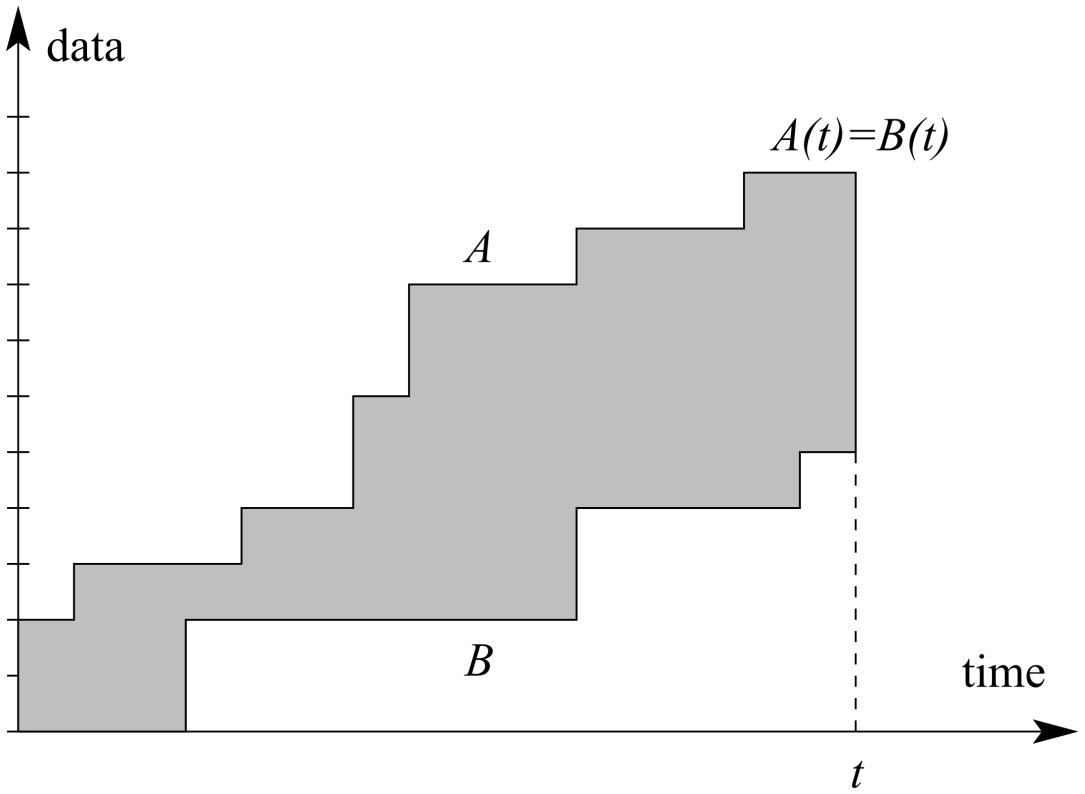
$$\frac{1}{t}\int_{x=0}^{t} N(x)\,dx = \frac{1}{t}\sum_{i=1}^{A(t)} T(i)$$

Deterministic general case: finite horizon



$$\frac{1}{t}\int_{x=0}^{t} N(x)\,dx = \frac{1}{t}\sum_{i=1}^{A(t)} T(i) = \frac{A(t)}{t}\,\frac{1}{A(t)}\sum_{i=1}^{A(t)} T(i)$$

Deterministic general case: finite horizon



$$\overline{N} = \frac{1}{t}\int_{x=0}^{t} N(x)\,dx = \frac{1}{t}\sum_{i=1}^{A(t)} T(i) = \frac{A(t)}{t}\,\frac{1}{A(t)}\sum_{i=1}^{A(t)} T(i) = \lambda \overline{T}$$

Deterministic general case: ∞ horizon

Pseudo-inverse: $f^{(-1)}(n) = \inf\{t | f(t) \geq n\}$, useful for staircase functions A et B $\nearrow$, right continuous, with limit $+\infty$.

Definitions & notations:

- $A(t)$: nb $n \in \mathbb{N}$ of clients *arrived* between time 0 and $t \in \mathbb{R}_+$.
- $B(t)$: nb $n \in \mathbb{N}$ of clients *left* between time 0 and $t \in \mathbb{R}_+$.
- $N(t) \stackrel{\text{def}}{=} A(t) - B(t)$: nb of clients in the queue at time t.
- $T(n) \stackrel{\text{def}}{=} B^{(-1)}(n) - A^{(-1)}(n)$: delay for client n° n (if FIFO).

Asymptotic averages: provided such limits exist,

- $\lambda \stackrel{\text{def}}{=} \lim_{t \to +\infty} \frac{A(t)}{t}$ (arrival rate)
- $\overline{T} \stackrel{\text{def}}{=} \lim_{n \to +\infty} \frac{1}{n} \sum_{i=1}^{n} T(i)$ (average delay)
- $\overline{N} \stackrel{\text{def}}{=} \lim_{t \to +\infty} \frac{1}{t} \int_{x=0}^{t} N(x)\,dx$ (average load)

Theorem (Brumelle/Stidham 1971/1974)

If λ and $\overline{T}$ exist and are finite, then $\overline{N}$ exists and $\overline{N} = \lambda \overline{T}$.

Deterministic general case: ∞ horizon

Hypotheses: $A, B : \mathbb{R}_+ \to \mathbb{N}$, $\nearrow$, right continuous, with limit $+\infty$.

Lemma

$$\lim_{t\to+\infty} \frac{A(t)}{t} = \lambda \iff \lim_{n\to+\infty} \frac{A^{(-1)}(n)}{n} = \frac{1}{\lambda} \text{ (true for } \lambda \in \mathbb{R}_+ \cup \{+\infty\})$$

Lemma

$$\text{If } \overline{T} \text{ finite, then } \lim_{n\to+\infty} \frac{T(n)}{n} = 0.$$

Lemma

$$\text{If } \overline{T} \text{ and } \lambda \text{ finite, then } \lim_{t\to+\infty} \frac{A(t)}{t} = \lambda \implies \lim_{t\to+\infty} \frac{B(t)}{t} = \lambda.$$

Deterministic general case: ∞ horizon

Hypotheses: $A, B : \mathbb{R}_+ \to \mathbb{N}$, $\nearrow$, right continuous, with limit $+\infty$.

Lemma

$\lim_{t\to+\infty} \frac{A(t)}{t} = \lambda \iff \lim_{n\to+\infty} \frac{A^{(-1)}(n)}{n} = \frac{1}{\lambda}$ *(true for* $\lambda \in \mathbb{R}_+ \cup \{+\infty\}$*)*

Lemma

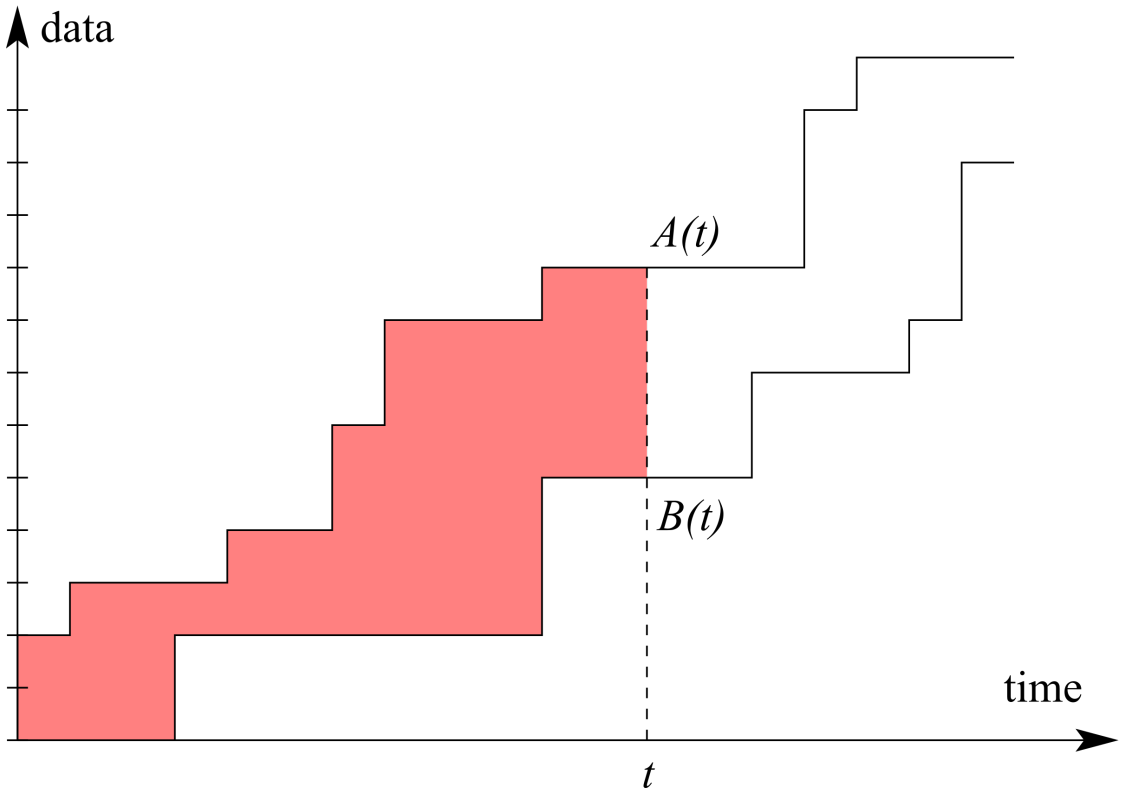
If $\overline{T}$ *finite, then* $\lim_{n\to+\infty} \frac{T(n)}{n} = 0$.

Lemma

If $\overline{T}$ *and* λ *finite, then* $\lim_{t\to+\infty} \frac{A(t)}{t} = \lambda \Longrightarrow \lim_{t\to+\infty} \frac{B(t)}{t} = \lambda$.

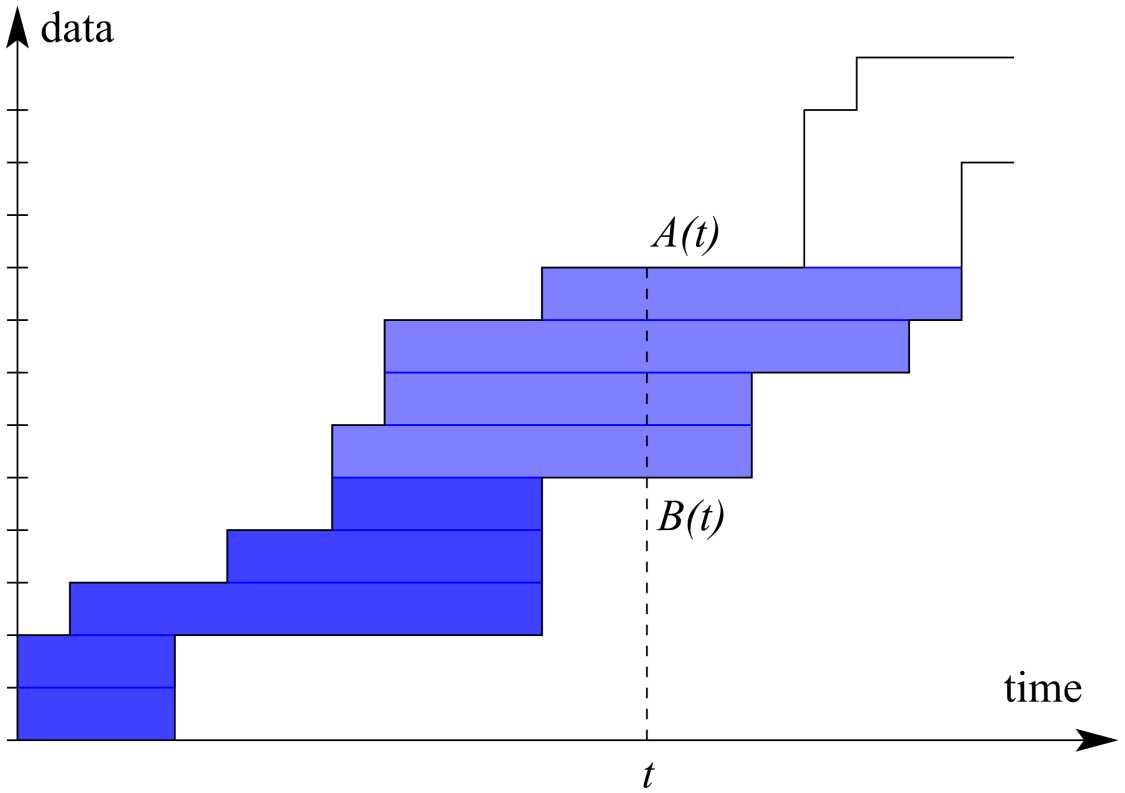
- **Lemma 1** Suppose $\lim_{t\to+\infty} \frac{A(t)}{t} = \lambda$. By def of $A^{(-1)}$ and right continuity of A,
  $A(A^{(-1)}(n)) \geq n \geq A(A^{(-1)}(n)-1)$. Thus,
  $$\frac{A(A^{(-1)}(n))}{A^{(-1)}(n)} \geq \frac{n}{A^{(-1)}(n)} \geq \frac{A(A^{(-1)}(n)-1)}{A^{(-1)}(n)} \underset{n\to+\infty}{\Longrightarrow} \lambda \geq \lim_{n\to+\infty} \frac{n}{A^{(-1)}(n)} \geq \lambda \text{ (since } A^{(-1)}(n) \to +\infty)$$
  Suppose $\lim_{n\to+\infty} \frac{A^{(-1)}(n)}{n} = \frac{1}{\lambda}$. By def of $A^{(-1)}$, $A^{(-1)}(A(t)) \leq t < A^{(-1)}(A(t)+1)$. Thus,
  $$\frac{A^{(-1)}(A(t))}{A(t)} \leq \frac{t}{A(t)} < \frac{A^{(-1)}(A(t)+1)}{A(t)} \underset{t\to+\infty}{\Longrightarrow} \frac{1}{\lambda} \leq \lim_{t\to+\infty} \frac{t}{A(t)} \leq \frac{1}{\lambda} \text{ (since } A(t) \to +\infty)$$
- **Lemma 2** If $\overline{T} < +\infty$, then if $n \to +\infty$, $\frac{T(n)}{n} = \frac{1}{n}\sum_{i=1}^{n} T(i) - \frac{n-1}{n}\frac{1}{n-1}\sum_{i=1}^{n-1} T(i) \to \overline{T} - 1 \times \overline{T} = 0$.
- **Lemma 3** If $\frac{A(t)}{t} \to \lambda$, then $\frac{A^{(-1)}(n)}{n} \to \frac{1}{\lambda}$ (Lemma 1).
  By def of T, $B^{(-1)}(n) = A^{(-1)}(n) + T(n) \underset{\text{Lemma 2}}{\Longrightarrow} \frac{B^{(-1)}(n)}{n} = \frac{A^{(-1)}(n)}{n} + \frac{T(n)}{n} \to \frac{1}{\lambda} + 0 = \frac{1}{\lambda}$.
  Thus $\frac{B(t)}{t} \to \lambda$ (Lemma 1).
  Lemma 1 enables to switch from vertical view $(A(t)/t)$ to horizontal view to exploit the hypothesis about delays $T = B^{(-1)} - A^{(-1)}$, then to come back to the vertical view $(B(t)/t)$.

Deterministic general case: ∞ horizon



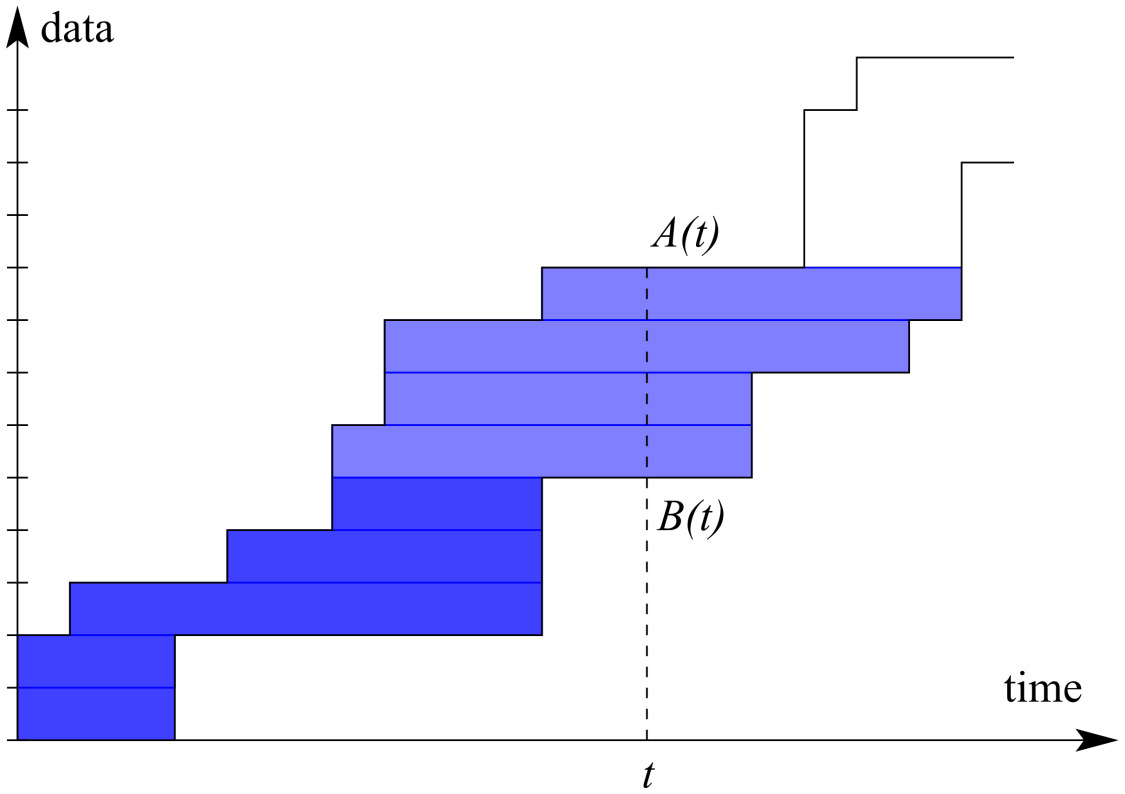
$$\int_{x=0}^{t} N(x)\,dx$$

Deterministic general case: ∞ horizon



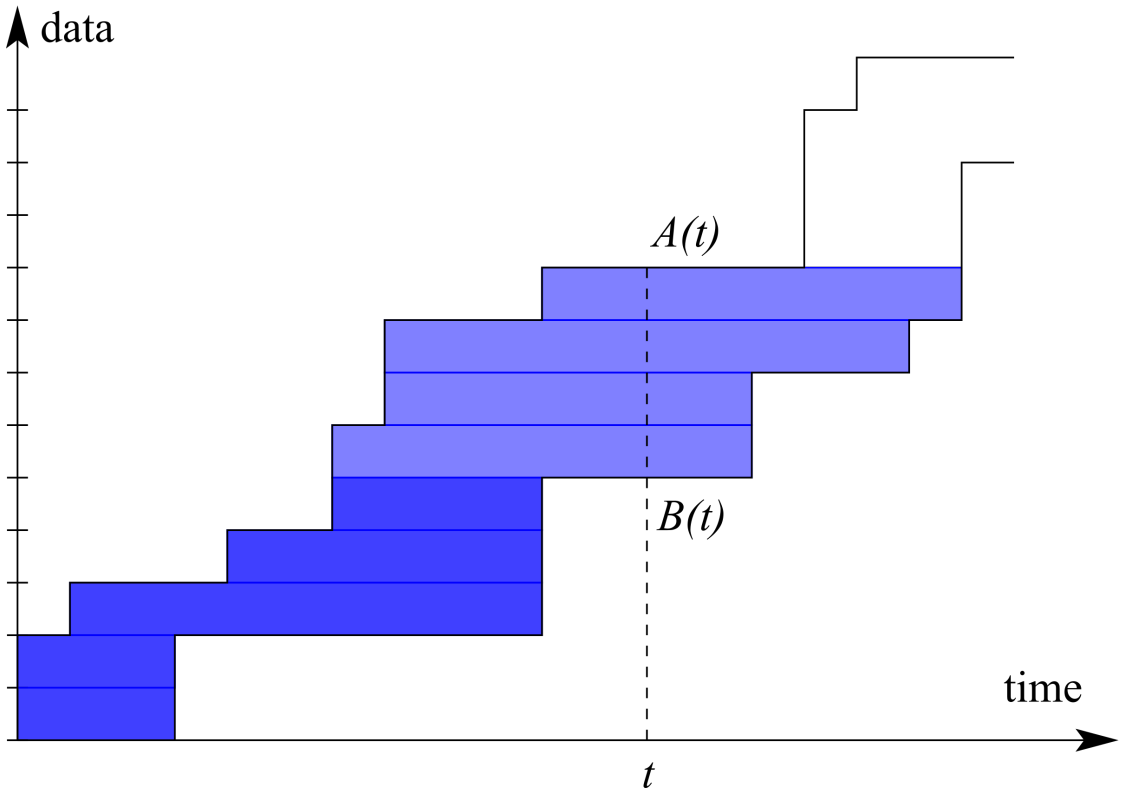
$$\sum_{i=1}^{B(t)} T(i) \leq \int_{x=0}^{t} N(x)\,dx \leq \sum_{i=1}^{A(t)} T(i)$$

Deterministic general case: ∞ horizon

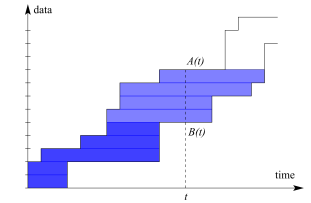


$$\frac{1}{t}\sum_{i=1}^{B(t)} T(i) \leq \frac{1}{t}\int_{x=0}^{t} N(x)\,dx \leq \frac{1}{t}\sum_{i=1}^{A(t)} T(i)$$

Deterministic general case: ∞ horizon



$$\frac{B(t)}{t}\frac{1}{B(t)}\sum_{i=1}^{B(t)} T(i) \leq \frac{1}{t}\int_{x=0}^{t} N(x)\,dx \leq \frac{A(t)}{t}\frac{1}{A(t)}\sum_{i=1}^{A(t)} T(i)$$



$$\frac{B(t)}{t}\frac{1}{B(t)}\sum_{i=1}^{B(t)}T(i) \leq \frac{1}{t}\int_{x=0}^{t}N(x)dx \leq \frac{A(t)}{t}\frac{1}{A(t)}\sum_{i=1}^{A(t)}T(i)$$

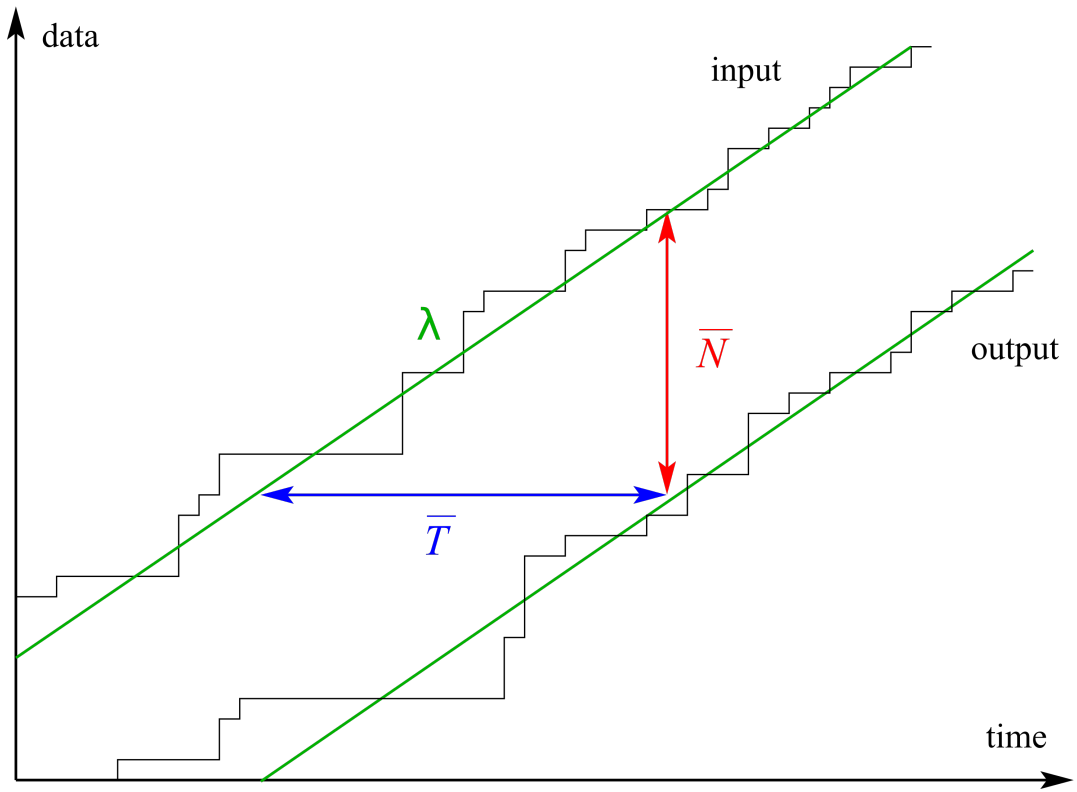
- During the last step, when one makes $t \to +\infty$, the hypotheses $\overline{T}$ and λ finite are used, to avoid an undefinite product. Note that it is the first time that we use $\lambda < +\infty$ (we already used $\overline{T} < +\infty$ in preliminary lemmas).

Little's law: applications

Practice: short calculations without precise modeling (assuming that Little's law is robust enough to work for the studied system).

Theory: if a model is ergodic, asympotic average of a parameter over one single trajectory = average of the parameter for the invariant distribution. Thus Little's law enables to derive some formulas about averages for the invariant distribution (e.g. average sojourn time from average buffer size in M/M/1 queues).

Little's law: a reminder figure



PASTA property: definition

Vocabulary: PASTA = Poisson Arrivals See Time Averages

Framework: probabilistic model of a system, at stationary state (if it exists), with Poissonian input traffic (each arrival induces state transition in the system). For each state of the system, focus on two probabilities:

1. Probability of the state seen by a random outside observer
   π_i = probability that system in state i at a random instant
2. Probability of the state seen by an arriving client
   π_i^* = probability that system in state i just before (a randomly chosen) arrival

Remark: in general $\pi_i \neq \pi_i^*$ but the Poisson effect yields an equality

PASTA property: counter-examples

Example 1: accessing your own laptop (one client, one server)

$$\begin{cases} \text{state } 0 = \text{ laptop free} \\ \text{state } 1 = \text{ laptop occupied} \end{cases}$$

$$\begin{cases} \pi_0^* = 1 \quad \text{(your own laptop is always free when needed)} \\ \pi_1^* = 0 \end{cases}$$

$$\begin{cases} \pi_0 = \text{ proportion of time the laptop is free } (< 1) \\ \pi_1 = \text{ proportion of time the laptop is occupied } (> 0) \end{cases}$$

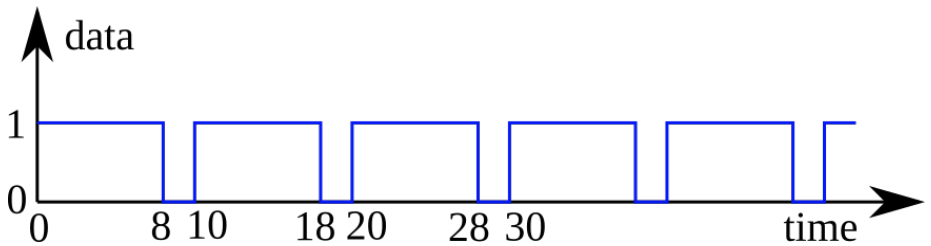
Remark: here arrival process is not Poisson, when an arrival has occurred (i.e. you have started to work with your laptop) for a while it is unlikely that another arrival occurs (i.e. you have stopped the previous session and started a new one). Thus arrivals at different times are not independent.

PASTA property: counter-examples

Example 2: deterministic periodic behavior (one client, one server), e.g. one new request every 10 sec and service time = 8 sec.
Framework a bit different here since not a probabilistic model, but one can study the PASTA property by setting a large time interval $[0, t]$ where the random observer chooses observation time uniformly (state 0 = server free, state 1 = server occupied).

$$\begin{cases} \pi_0^* = 1 & \text{(server always free before next arrival)} \\ \pi_1^* = 0 \end{cases}$$

$$\begin{cases} \pi_0 \approx 0.2 \\ \pi_1 \approx 0.8 \end{cases}$$



PASTA property: Poisson processes

Theorem (PASTA property for Poisson processes)

With the previous notations, if the input traffic is a Poisson process, then for any state i of the system, $\pi_i^* = \pi_i$.

Proof: Arrival history before the instant of consideration, irrespective whether we are considering a random instant or an arrival instant, are stochastically the same: a sequence of arrivals with exponentially distributed interarrival times. Moreover remaining time to next arrival has the same exponential distribution irrespective of the time that has already elapsed since the previous arrival.
Both come from memoryless property of exponential distribution (the same view also holds in reversed time, i.e. looking backwards).

PASTA property: Poisson processes

Theorem (PASTA property for Poisson processes)

With the previous notations, if the input traffic is a Poisson process, then for any state i of the system, $\pi_i^* = \pi_i$.

Proof: Since the stochastic characterization of the arrival process before the instant of consideration is the same, irrespective how the instant has been chosen) the state distributions of the system (induced by the past arrivals processes) at the instant of consideration must be the same in both cases.