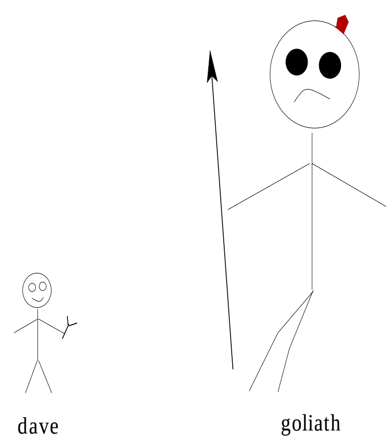
Before proceeding to bad news:

⟶ connection between heavy-tailedness and `google`?

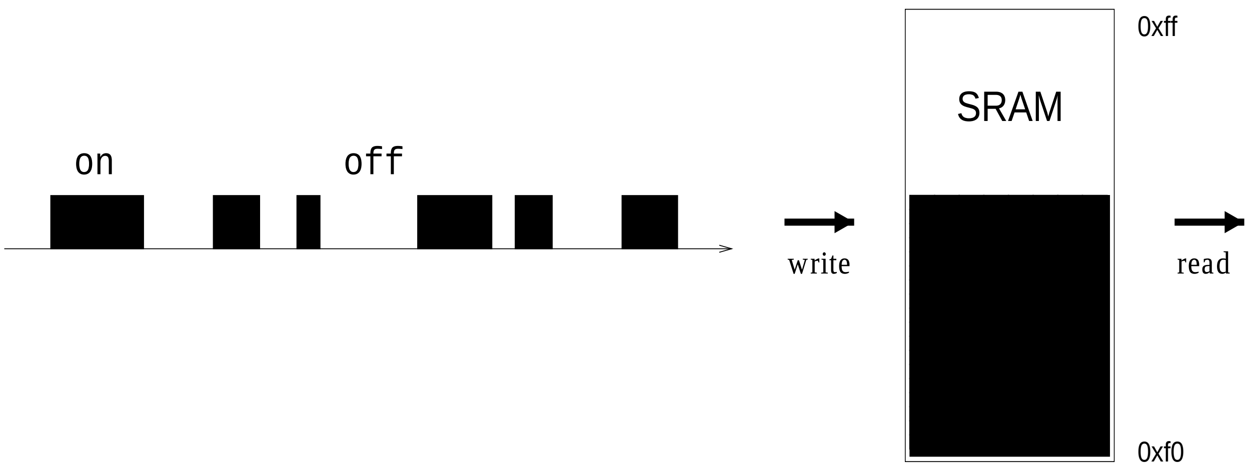
⟶ saga of two lucky kids (aka "grad students")

⟶ lesson to be drawn?



Now, to the bad news!

Bad news #1: queueing



- influx rate (**write**) < outflux rate (**read**)

  → else buffer will grow out of bound

- during on-time: if **write** rate < **read** rate

  → then what?

  → economy dictates opposite (suppose 1/2)

  → hence: during on-time buffer grows (McDonald's)

Since on/off input is random, so is the buffer/memory occupancy

$\longrightarrow$ at time t, could be 10 KB, 120 KB, etc.

$\longrightarrow$ i.e., $\Pr\{Q(t) = 10000\}$ = some value, ...

Want to know: in the long-run ($t \rightarrow \infty$) what is $Q(t)$?

$\longrightarrow$ write as $Q(\infty)$

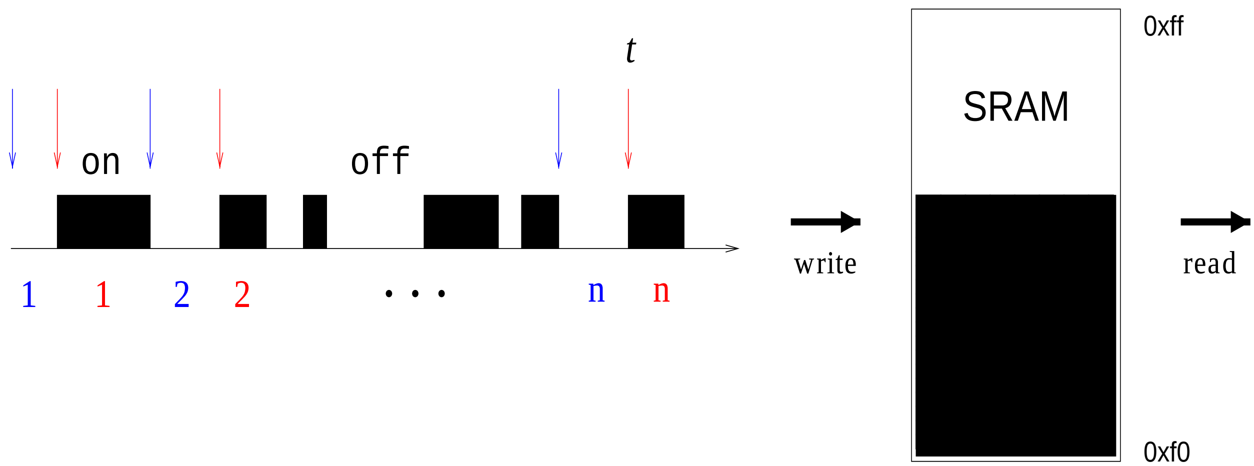
$\longrightarrow$ practical interest: $\Pr\{Q(\infty) > x\}$?

$\longrightarrow$ corresponds to excessive delay, buffer loss, etc.

Case I: what shape does $\Pr\{Q(\infty) > x\}$ take when both on and off periods are exponential?

$\longrightarrow$ assume i.i.d. (with be^{-bt})

$\longrightarrow$ first, switch from time unit to count unit



- alternating on/off periods: mutually independent
- how many on and off periods (n) at time t?

  $\rightarrow n \approx t/(E[\text{on}] + E[\text{off}]) = t/(\frac{1}{b} + \frac{1}{b}) = bt/2$

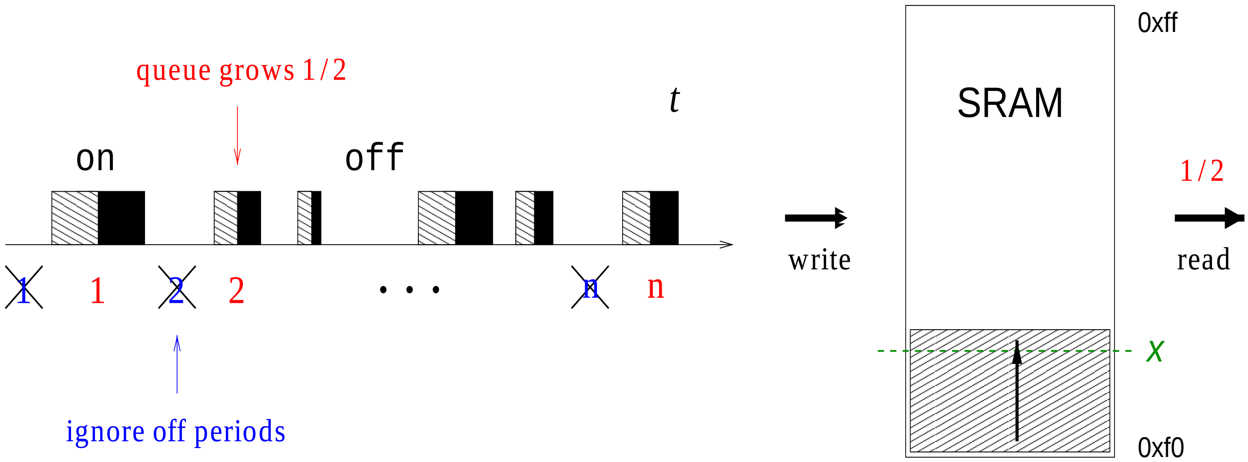
  $\rightarrow$ for large t

- sum of n on-periods: S_n

$\longrightarrow$ let's upper-bound $\Pr\{Q(t) > x\}$

$\longrightarrow$ i.e., $\Pr\{Q(t) > x\} < \boxed{?}$

Upper-bounding idea:



- worst-case viewpoint

  → ignore the beneficial effect of off periods

  → McDonald's: groups of people arrive without pause

- thus, to have $Q(t) > x$:

  → $S_n/2 > x$

  → i.e., at the very least

  → hence: $\Pr\{Q(t) > x\} < \Pr\{S_n > 2x\}$

- need to upper bound $\Pr\{S_n > 2x\}$

  $\longrightarrow$ for large x (i.e., $2x > nE[\text{on}]$)

$$\begin{aligned}\Pr\{S_n > 2x\} &= \Pr\{S_n - nE[\text{on}] > 2x - nE[\text{on}]\} \\ &= \Pr\{S_n/n - E[\text{on}] > 2x/n - E[\text{on}]\}\end{aligned}$$

  $\longrightarrow$ by LLN S_n is concentrated around its mean!

- we can apply large deviation bound

  $\longrightarrow$ $\Pr\{|\frac{S_n(t)}{n} - p| > \varepsilon\} < e^{-an}$

  $\longrightarrow$ here: $\varepsilon = 2x/n - E[\text{on}]$

  $\longrightarrow$ recall: a depends on ε

- facts: shape of $a(\varepsilon)$

  $\longrightarrow$ binary case: $a = \varepsilon \log \frac{\varepsilon}{p} + (1 - \varepsilon) \log \frac{1-\varepsilon}{1-p}$

  $\longrightarrow$ exponential case: $a = b\varepsilon - 1 - \log b\varepsilon$

  $\longrightarrow$ for large ε (same as large x): $a \approx b\varepsilon$

- apply large deviation bound to S_n

$$\begin{aligned}
\Pr\{S_n > 2x\} &= \Pr\{S_n/n - E[\text{on}] > 2x/n - E[\text{on}]\} \\
&< e^{-an} \\
&\approx e^{-b\varepsilon n} \\
&= e^{-b(2x/n - E[\text{on}])n} \\
&= e^{-2bx + bE[\text{on}]n} \\
&= e^{-2bx+n} \\
&< e^{-bx}
\end{aligned}$$

$\rightarrow$ for sufficiently large x (used several times)

Thus: $\Pr\{Q(t) > x\} < e^{-bx}$ for large x and t

$\longrightarrow$ $\Pr\{Q(\infty) > x\} < e^{-bx}$ for large x

$\longrightarrow$ prob. of queue growing large: exponentially small

$\longrightarrow$ for exponential traffic: buffering is effective

$\longrightarrow$ extra buffer/memory y buys a lot:

$$\Pr\{Q(\infty) > x + y\} < e^{-b(x+y)} = e^{-bx}e^{-by}$$

Remarks: analysis method

⟶ con: only holds for large x

⟶ pro: very general/powerful

⟶ for exponential case: “excessive force”

⟶ somewhat like “catching fly with a cannon”

⟶ can use more elementary methods

⟶ a course in queueing theory (Markovian input)

⟶ problem: doesn’t extend to heavy-tailed input

⟶ but the Internet is heavy-tailed!

Case II: shape of $\Pr\{Q(\infty) > x\}$ when off-period is exponential but on-period is heavy-tailed?

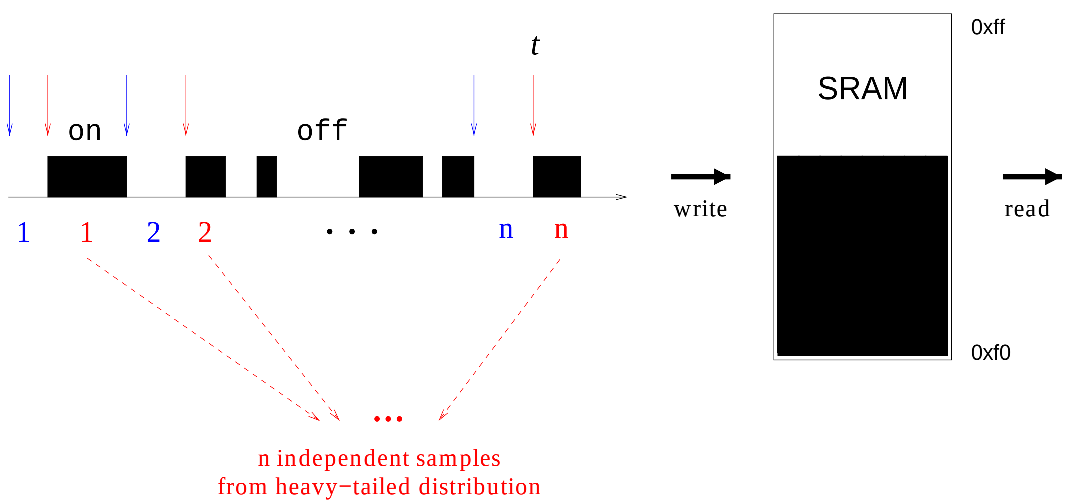
$\longrightarrow$ want to show: $\Pr\{Q(\infty) > x\}$ is heavier as well

$\longrightarrow$ want to contrast with exponential case

$\longrightarrow$ let's lower-bound: $\Pr\{Q(t) > x\} > \boxed{?}$

$\longrightarrow$ why upper-bounding not enough?

Lower-bounding idea:



$\longrightarrow$ sampling viewpoint: ok since i.i.d.

$\longrightarrow$ wait till first long ($> 2x$) on-period

- how long must one wait?

  $\longrightarrow$ on the order of $1/\Pr\{Z > x\} = 1/cx^{-\alpha} \propto x^{\alpha}$

  $\longrightarrow$ so, time scale of interest: $t = O(x^{\alpha})$

- number on and off periods at time t

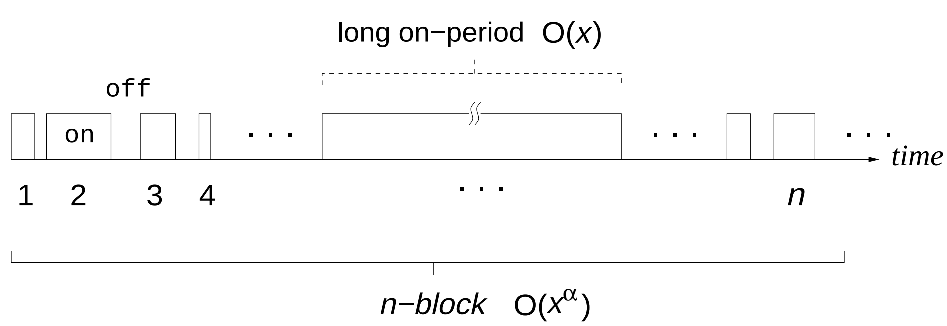
  $\longrightarrow n \approx t/(E[\text{on}] + E[\text{off}]) = \delta t = O(x^{\alpha})$

  $\longrightarrow$ for large t (hence large x)

- now: $\Pr\{Q(t) > x\} \approx$ fraction of time during $O(x^{\alpha})$ where queue is bigger than x

  $\longrightarrow O(x/x^{\alpha}) = O(x^{1-\alpha})$

  $\longrightarrow$ where did we apply similar reasoning?



- note: we ignored the contribution of other on periods

  $\longrightarrow$ hence: lower-bound

- thus: for large x and t

  $\longrightarrow$ $\Pr\{Q(t) > x\} > O(x^{1-\alpha})$

  $\longrightarrow$ tail $\Pr\{Q(\infty) > x\}$ is at least polynomially heavy

  $\longrightarrow$ can also show polynomially upper-bounded

  $\longrightarrow$ much more likely to overcrowd

  $\longrightarrow$ buffering is not as effective: marginal gain small

  $\longrightarrow$ modern view: bandwidth-centric resource provisioning

Remarks:

- heavy-tailed on-times and resultant heavy-tailed queueing was a big surprise

  $\longrightarrow$ grabbed CS, EE, statistics/probability, OR, some physicists, etc. by surprise!

  $\longrightarrow$ huge scientific impact

- one technical aside: for heavy-tailed i.i.d. variables

  $\Pr\{Z_1 + \cdots + Z_n > x\} = \Pr\{\max\{Z_1, \ldots, Z_n\} > x\}$

  $\longrightarrow$ for large x

  $\longrightarrow$ when the sum is large, one guy is to blame!

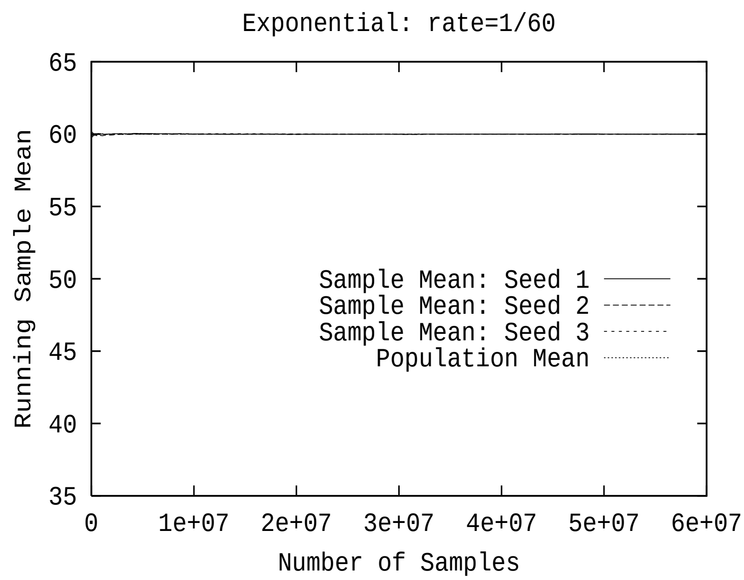
  $\longrightarrow$ single long on-period picture: accurate

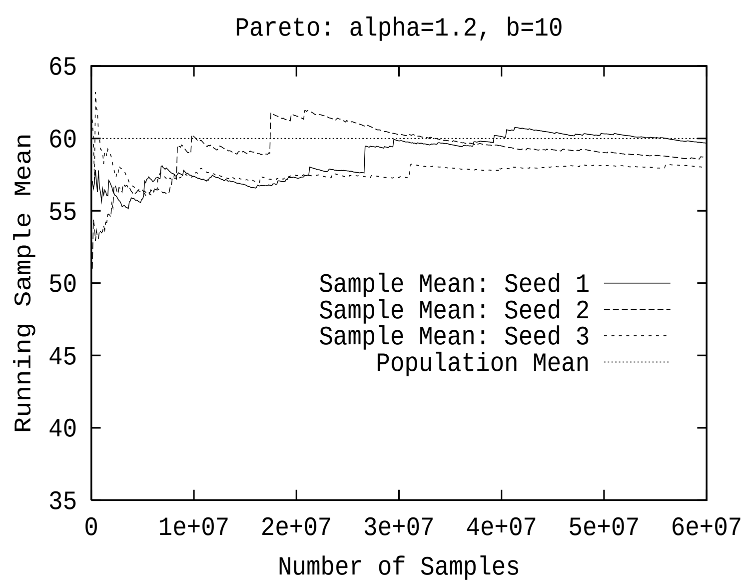
  $\longrightarrow$ yields upper bound

  $\longrightarrow$ starkly different from exponential: equal blame

  $\longrightarrow$ implication to sampling and simulation: slow convergence

Sample mean convergenge rate: exponential vs. Pareto

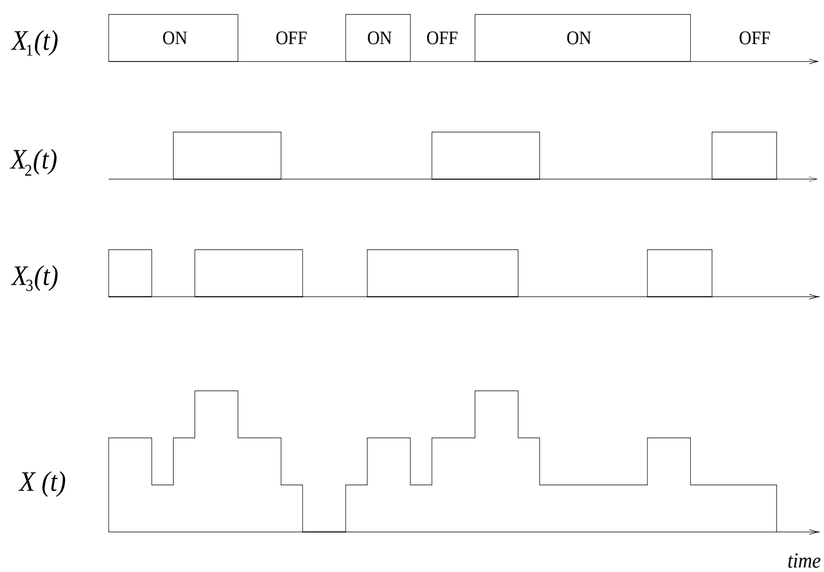




Lastly: characteristic of aggregate traffic

⟶ multiple on/off sources

Recall:

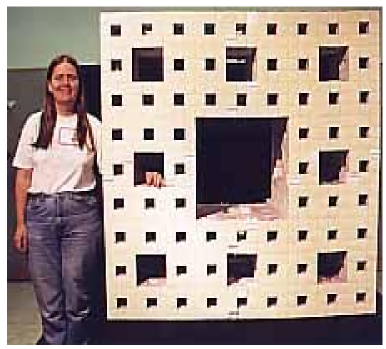


⟶ with many on/off sessions, what does $X(t)$ look like?

⟶ it's fractal, i.e., self-similar!

Some fractal objects:

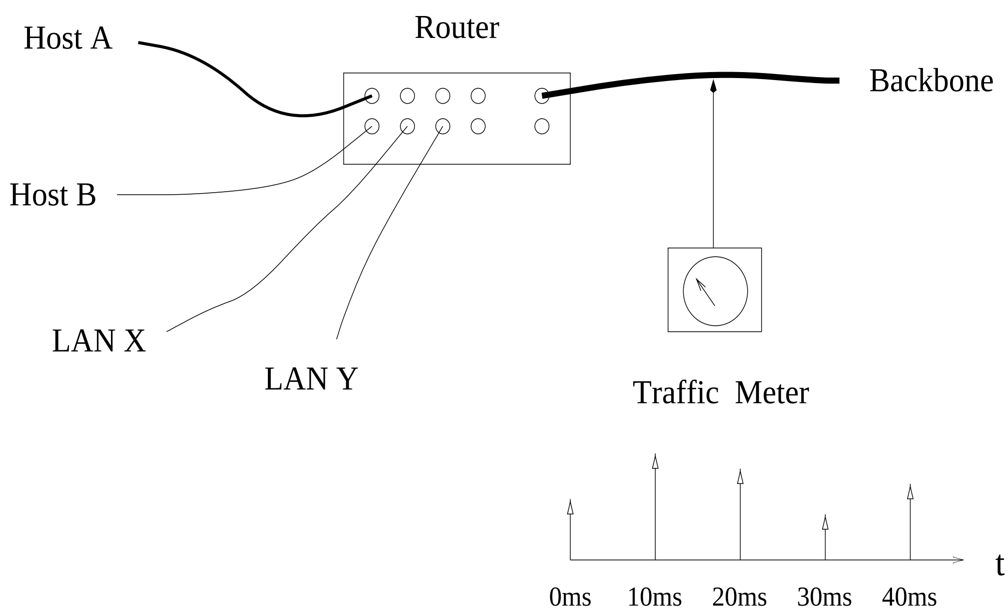
Menger sponge (picture from `www.ics.uci.edu/∼eppstein`):



Fractal fern:

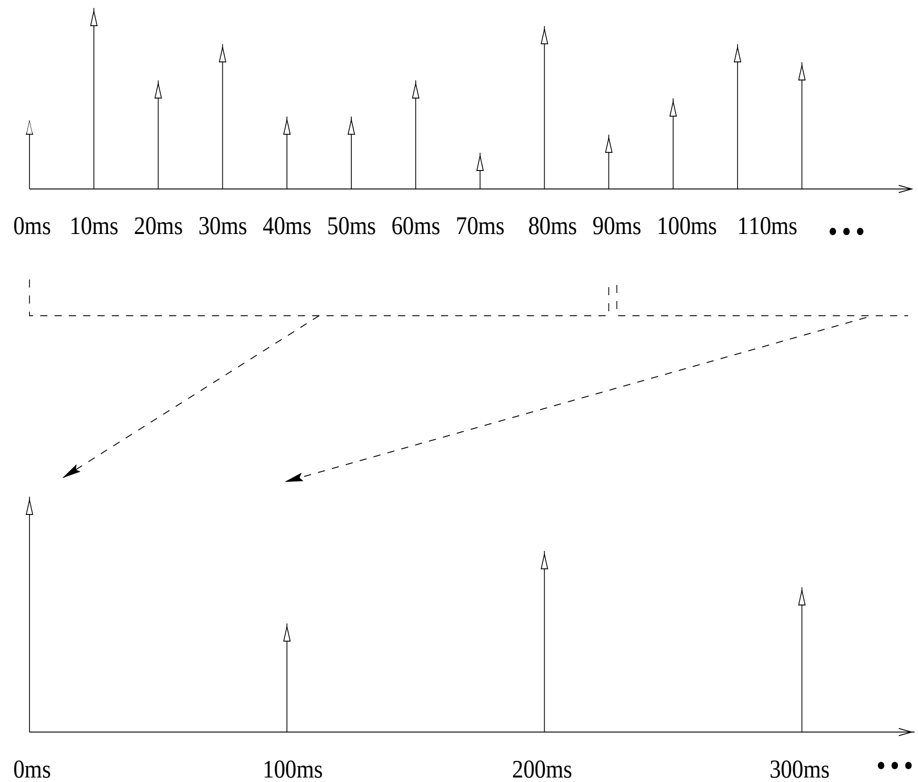
$\longrightarrow$ are fractal objects random?

Internet traffic: measurement



$\longrightarrow$ traffic time series (at 10ms granularity)

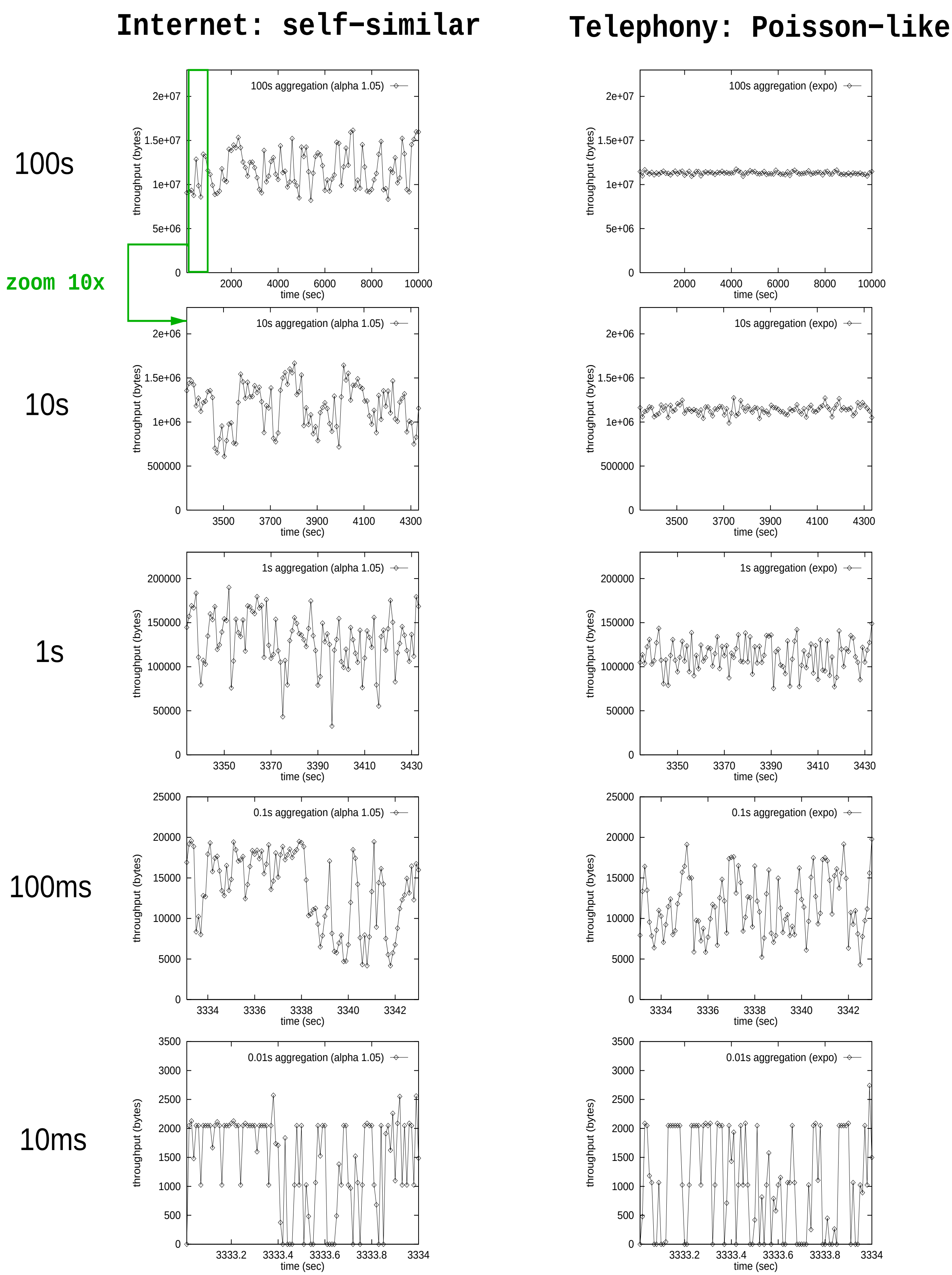
Aggregation (time):



⟶ analogous to computing sample mean

⟶ aggregation over multiple time scales

⟶ what to expect?


Internet: self-similar
Telephony: Poisson-like
100s
10s
1s
100ms
10ms
zoom 10x
100s aggregation (alpha 1.05)
100s aggregation (expo)
10s aggregation (alpha 1.05)
10s aggregation (expo)
1s aggregation (alpha 1.05)
1s aggregation (expo)
0.1s aggregation (alpha 1.05)
0.1s aggregation (expo)
0.01s aggregation (alpha 1.05)
0.01s aggregation (expo)
throughput (bytes)
time (sec)

We observe:

- for Internet traffic burstiness preserved across time scales five orders of magnitude apart

  → Poisson traffic: smoothes out quickly

- if traffic were uncorrelated in time, by LLN should smooth out

  → how fast should it smooth out?

Self-similar burstiness viewpoint:

Time aggregation of $X(t)$ at level m means

$$X^{(m)}(i) = \frac{1}{m} \sum_{t=m(i-1)+1}^{mi} X(t).$$

Since $X(t)$ are random variables, $X^{(m)}(i)$ in time series is analogous to computing the sample mean.

The visual phenomenon of "burstiness preservation" corresponds to

$$\mathrm{var}(X^{(m)}(i)) \approx \mathrm{var}(X(t))$$

for a range of time scales m.

If the $X(t)$'s were independent, then

$$\mathrm{var}(X^{(m)}(i)) = \sigma^2 m^{-1}$$

where σ^2 is the variance of the $X(t)$'s.

$\longrightarrow$ elementary fact

Consider rewriting expression with parameter H as

$$\sigma^2 m^{-2(1-H)}$$

where $1/2 \leq H < 1$.

If $H = 1/2$, then we have previous expression σ^2/m.

$\longrightarrow$ σ^2 decays at rate m^{-1}

If $1/2 < H < 1$, then rate of decay is slower.

$\longrightarrow$ $m^{-\beta}$ where $0 < \beta < 1$

Thus, if $H \approx 1$, then can expect

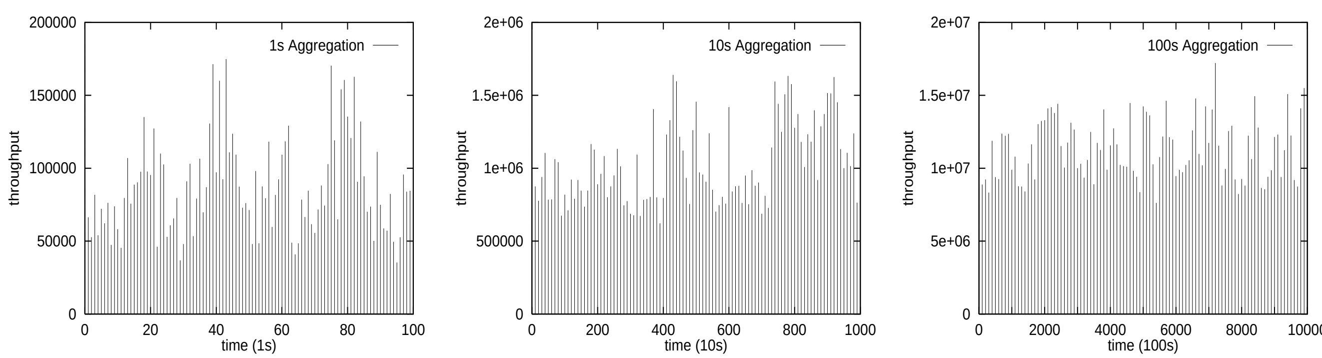
$$\mathrm{var}(X^{(m)}(i)) \approx \mathrm{var}(X(t))$$

$\longrightarrow$ burstiness dies out very slowly w.r.t. scaling

$\longrightarrow$ empirically: H is 0.8–0.9 range

$\longrightarrow$ $X(t)$ must be strongly correlated in time

$\longrightarrow$ what causes it?



The principal cause: heavy-tailed file sizes!

$\longrightarrow$ present impacts distant future

$\longrightarrow$ recall predictability discussion

$$\Pr\{Z > x + y \,|\, Z > y\} = \left(\frac{y}{y+x}\right)^{\alpha}$$

$\longrightarrow$ predictability also leads to long-term correlation

$\longrightarrow$ consequences: heavy-tailed queueing

$\longrightarrow$ periods of over- and under-utilization

$\longrightarrow$ bad for resource provisioning

Internet: self-similar **Telephony: Poisson-like**