Heavy Tails: Performance Models and Scheduling Disciplines

Part I – Introduction and Methodology

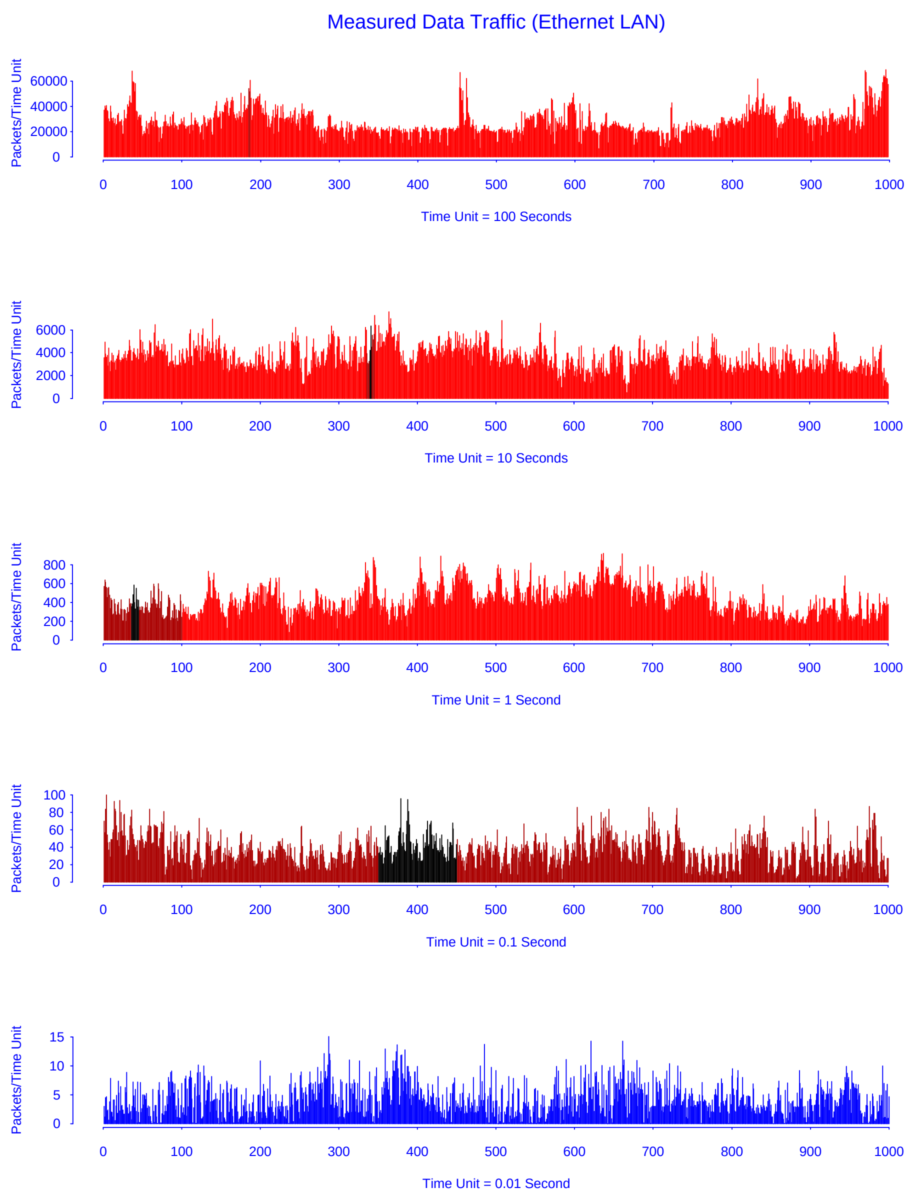
Onno Boxma

Heavy Tails:
Performance Models and Scheduling Disciplines
Part I – Introduction and Methodology

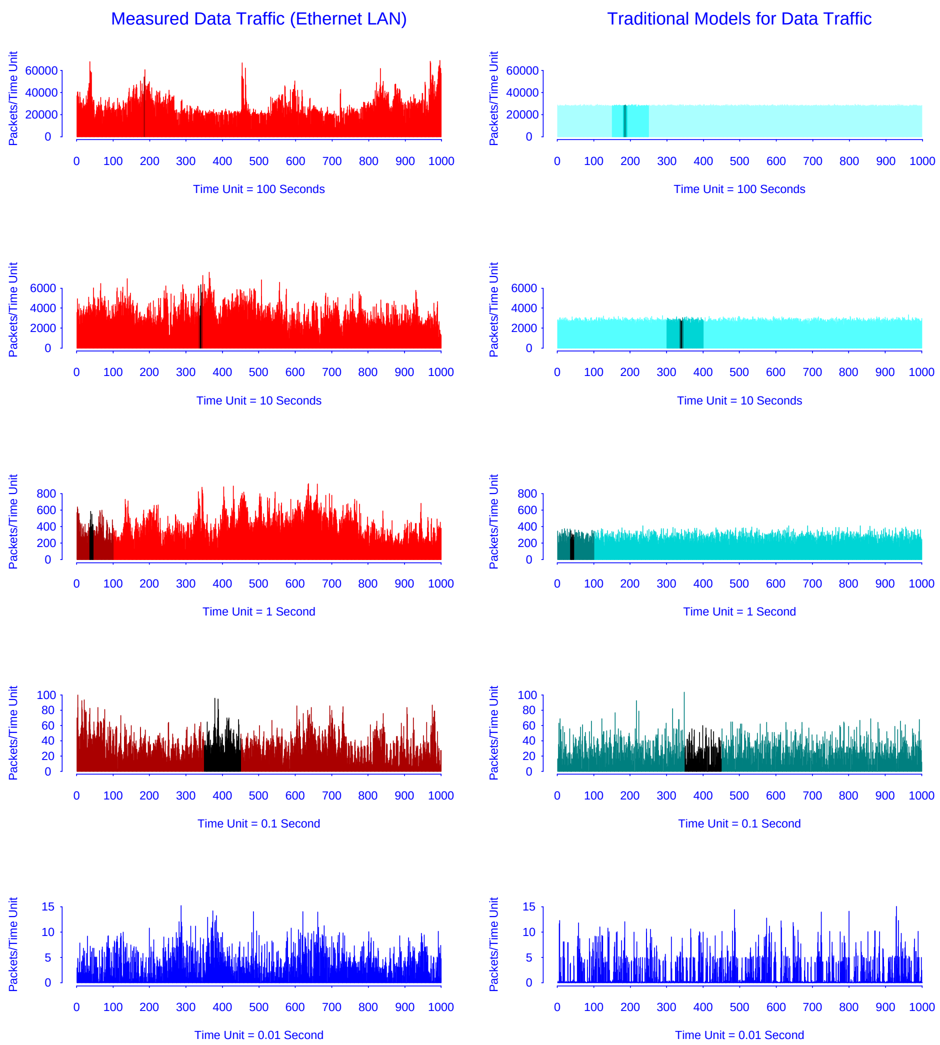
Tales to tell:

- traffic measurements and statistical analysis
- traffic modeling
- heavy-tailed input: performance analysis
- heavy-tailed input: damage control

traffic measurements and statistical analysis



traffic measurements and statistical analysis



The Bellcore data (courtesy: W. Willinger)

Heavy Tails:
Performance Models and Scheduling Disciplines
Part I – Introduction and Methodology

Tales to tell:

- traffic measurements and statistical analysis
- traffic modeling
- heavy-tailed input: performance analysis
- heavy-tailed input: damage control

traffic modeling

W **is** **heavy-tailed** **if** $\mathrm{e}^{\epsilon x}\mathrm{P}(W > x) \to \infty$**,** $\forall \epsilon > 0$**.**
Examples:

Pareto: $\mathrm{P}(B > x) = (\frac{\theta}{\theta + x})^{\nu}, \quad x \geq 0$**.**

Weibull: $\mathrm{P}(B > x) = \mathrm{e}^{-(x/b)^c}, \quad 0 < c < 1, \quad x \geq 0$**.**

lognormal: $\mathrm{P}(B > x) = \int_x^\infty \frac{1}{y\sigma\sqrt{2\pi}} \mathrm{exp}[\frac{-[\log(y/m)]^2}{2\sigma^2} \mathrm{d}y], \quad x \geq 0$**.**

Heavy Tails: Performance Models and Scheduling Disciplines Part I – Introduction and Methodology

Tales to tell:

- traffic measurements and statistical analysis
- traffic modeling
- heavy-tailed input: performance analysis
- heavy-tailed input: damage control

heavy-tailed input: performance analysis

Key questions:

- **tail behavior of key performance measures**
  $M/G/1$
  if

$$\mathsf{P}(B > x) \sim x^{-\nu}, \quad x \to \infty,$$

then

$$\mathsf{P}(W_{FCFS} > x) \sim ??, \quad x \to \infty.$$

heavy-tailed input: performance analysis

Key questions:

- **tail behavior of key performance measures**
  $M/G/1$
  if

$$\mathsf{P}(B > x) \sim x^{-\nu}, \quad x \to \infty,$$

then

$$\mathsf{P}(W_{FCFS} > x) \sim Cx^{1-\nu}, \quad x \to \infty.$$

Tales to tell:

- traffic measurements and statistical analysis
- traffic modeling
- heavy-tailed input: performance analysis
- heavy-tailed input: damage control

heavy-tailed input: performance analysis/damage control

Key questions:

- **tail behavior of key performance measures**
- **multiclass systems: effect of one class on another class**

heavy-tailed input: performance analysis/damage control

Key questions:

- tail behavior of key performance measures
- multiclass systems: effect of one class on another class
- influence of the service discipline

heavy-tailed input: performance analysis/damage control

$M/G/1$

if

$$\mathsf{P}(B > x) \sim x^{-\nu}, \quad x \to \infty,$$

then

$$\mathsf{P}(W_{FCFS} > x) \sim Cx^{1-\nu}, \quad x \to \infty.$$

but perhaps

$$\mathsf{P}(W_{smart} > x) \sim Dx^{-\nu}, \quad x \to \infty?$$

Ref.: V. Anantharam, QUESTA 33, 1999

heavy-tailed input: performance analysis/damage control

Key questions:

- tail behavior of key performance measures
- multiclass systems: effect of one class on another class
- influence of the service discipline
- fairness

Organization of the tutorial:

- **Part I. Tail behavior of key performance measures: Methodology**

- **Part II. Workload asymptotics of GPS systems**
  (Sem Borst)

- **Part III. Delay asymptotics for various scheduling strategies; PS in integrated services environments**
  (Sindo Nunez)

- **Part IV. Scheduling in practice; fairness**
  (Mor Harchol-Balter)

Organization of Part I:

1. introduction

2. regular variation

3. $M/G/1$ FCFS: workload and waiting time

4. $M/G/1$: busy period

5. $M/G/1$ LCFS preemptive resume: sojourn time

6. epilogue

2. Regular variation

M/G/1 queue: $B(x) = \mathsf{P}(B < x)$,
LST $\mathsf{E}[\mathrm{e}^{-sB}]$, **mean** $\mathsf{E}B < \infty$.

arrival rate λ, **load** $\rho := \lambda \mathsf{E}B < 1$.

Regularly varying service time distribution:

$$\mathsf{P}(B > x) = x^{-\nu} L(x).$$

2. Regular variation

M/G/1 queue: $B(x) = \mathsf{P}(B < x)$,
LST $\mathsf{E}[\mathrm{e}^{-sB}]$, **mean** $\mathsf{E}B < \infty$.

arrival rate λ, **load** $\rho := \lambda \mathsf{E}B < 1$.

Regularly varying service time distribution:

$$\mathsf{P}(B > x) = x^{-\nu} L(x).$$

$L(\cdot)$ **is slowly varying:** $\lim_{t \to \infty} \frac{L(tx)}{L(t)} = 1, \ \forall x > 0$.

(we mainly consider $1 < \nu < 2$: $\mathrm{Var}(B) = \infty$.**)**

Example: Pareto. $\mathsf{P}(B > x) = (\frac{\theta}{\theta + x})^{\nu}, \quad x \geq 0$.

Key lemma for regularly varying tails (Bingham & Doney):

let $n < \nu < n+1$.

Equivalent are:

$$\mathsf{P}(Y > x) \sim \mathbf{x}^{-\nu} L(x), \quad x \to \infty,$$

and

$$\mathsf{E}[\mathsf{e}^{-sY}] - \sum_{j=0}^{n} \mathsf{E}Y^j \frac{(-s)^j}{j!} \sim -\Gamma(1-\nu)\mathrm{s}^{\nu} L(\frac{1}{s}), \quad s \downarrow 0.$$

3. $M/G/1$ FCFS: waiting time (and workload)

$G/G/1$ FCFS

Cohen (JAP, 1973):

$$\mathsf{P}(B > x) \sim \mathbf{x}^{-\nu} L(x), \quad x \to \infty, \ \nu > 1 \Longleftrightarrow$$

$$\mathsf{P}(W > x) \sim \frac{\rho}{1-\rho} \frac{1}{\nu - 1} \frac{1}{\mathsf{E}B} \mathbf{x}^{1-\nu} L(x), \quad x \to \infty.$$

3. $M/G/1$ FCFS: waiting time (and workload)

$G/G/1$ FCFS

Cohen (JAP, 1973):

$$\mathsf{P}(B > x) \sim \mathbf{x}^{-\nu} L(x), \quad x \to \infty, \ \nu > 1 \Longleftrightarrow$$

$$\mathsf{P}(W > x) \sim \frac{\rho}{1-\rho} \frac{1}{\nu - 1} \frac{1}{\mathsf{E}B} \mathbf{x}^{1-\nu} L(x), \quad x \to \infty.$$

In fact (Pakes, JAP 1975, for subexponential B^{res}):

$$\mathsf{P}(W > x) \sim \frac{\rho}{1-\rho} \mathsf{P}(B^{res} > x), \quad x \to \infty.$$

Note that $\mathsf{P}(B^{res} > x) = \int_x^\infty \frac{\mathsf{P}(B>u)}{\mathsf{E}B} \mathrm{d}u$
and $\mathsf{E}B^{res} = \frac{\mathsf{E}B^2}{2\mathsf{E}B}$.

Four proofs for $M/G/1$:

1. **Direct proof**

2. **Proof via LST and Bingham-Doney**

3. **Proof via sample-path argument**

4. **Proof via conditional moments method (see Part III)**

$$\mathsf{P}(W > x) \sim \frac{\rho}{1-\rho}\mathsf{P}(B^{res} > x), \quad x \to \infty.$$

1. Direct proof. See next slide:

$$W =^d B_1^{res} + \ldots + B_K^{res},$$

with $K \sim \mathsf{geom}(\rho)$**.**

Hence

$$\begin{aligned} \mathsf{P}(W > x) &= (1-\rho)\sum_{n=0}^{\infty} \rho^n \mathsf{P}(B_1^{res} + \ldots + B_n^{res} > x) \\ &\sim (1-\rho)\sum_{n=0}^{\infty} \rho^n n \mathsf{P}(B^{res} > x) \\ &= \frac{\rho}{1-\rho}\mathsf{P}(B^{res} > x), \quad x \to \infty. \end{aligned}$$

Indeed (Pollaczek-Khintchine formula for $M/G/1$):

$$\mathrm{E}[\mathrm{e}^{-sW}] = \frac{(1-\rho)s}{s-\lambda+\lambda\mathrm{E}[\mathrm{e}^{-sB}]}$$

Indeed (Pollaczek-Khintchine formula for $M/G/1$):

$$\mathsf{E}[\mathrm{e}^{-sW}] = \frac{(1-\rho)s}{s-\lambda+\lambda\mathsf{E}[\mathrm{e}^{-sB}]} = \frac{1-\rho}{1-\lambda\frac{1-\mathsf{E}[\mathrm{e}^{-sB}]}{s}}$$

Indeed (Pollaczek-Khintchine formula for $M/G/1$):

$$\begin{aligned} \mathrm{E}[\mathrm{e}^{-sW}] &= \frac{(1-\rho)s}{s-\lambda+\lambda \mathrm{E}[\mathrm{e}^{-sB}]} \\ &= \frac{1-\rho}{1-\lambda\frac{1-\mathrm{E}[\mathrm{e}^{-sB}]}{s}} \\ &= \frac{1-\rho}{1-\rho\frac{1-\mathrm{E}[\mathrm{e}^{-sB}]}{s\mathrm{E}B}} \end{aligned}$$

Indeed (Pollaczek-Khintchine formula for $M/G/1$):

$$\begin{aligned} \mathsf{E}[\mathsf{e}^{-sW}] &= \frac{(1-\rho)s}{s-\lambda+\lambda\mathsf{E}[\mathsf{e}^{-sB}]} \\ &= \frac{1-\rho}{1-\lambda\frac{1-\mathsf{E}[\mathsf{e}^{-sB}]}{s}} \\ &= \frac{1-\rho}{1-\rho\frac{1-\mathsf{E}[\mathsf{e}^{-sB}]}{s\mathsf{E}B}} \\ &= \frac{1-\rho}{1-\rho\mathsf{E}[\mathsf{e}^{-sB^{res}}]} \end{aligned}$$

Indeed (Pollaczek-Khintchine formula for $M/G/1$):

$$\begin{aligned}
\mathsf{E}[\mathrm{e}^{-sW}] &= \frac{(1-\rho)s}{s-\lambda+\lambda\mathsf{E}[\mathrm{e}^{-sB}]} \\
&= \frac{1-\rho}{1-\lambda\frac{1-\mathsf{E}[\mathrm{e}^{-sB}]}{s}} \\
&= \frac{1-\rho}{1-\rho\frac{1-\mathsf{E}[\mathrm{e}^{-sB}]}{s\mathsf{E}B}} \\
&= \frac{1-\rho}{1-\rho\mathsf{E}[\mathrm{e}^{-sB^{res}}]} \\
&= (1-\rho)\sum_{n=0}^{\infty}\rho^n(\mathsf{E}[\mathrm{e}^{-sB^{res}}])^n,
\end{aligned}$$

Indeed (Pollaczek-Khintchine formula for $M/G/1$):

$$\begin{aligned}
\mathsf{E}[e^{-sW}] &= \frac{(1-\rho)s}{s-\lambda+\lambda\mathsf{E}[e^{-sB}]} \\
&= \frac{1-\rho}{1-\lambda\frac{1-\mathsf{E}[e^{-sB}]}{s}} \\
&= \frac{1-\rho}{1-\rho\frac{1-\mathsf{E}[e^{-sB}]}{s\mathsf{E}B}} \\
&= \frac{1-\rho}{1-\rho\mathsf{E}[e^{-sB^{res}}]} \\
&= (1-\rho)\sum_{n=0}^{\infty}\rho^n(\mathsf{E}[e^{-sB^{res}}])^n,
\end{aligned}$$

so

$$W =^d B_1^{res} + \ldots + B_K^{res}.$$

Remark

$$\mathsf{P}(B_1^{res} + \ldots + B_n^{res} > x) \sim n\mathsf{P}(B^{res} > x), \quad x \to \infty,$$

holds for subexponential (e.g., regular varying) distributions.

Crucial property: if sum is large, it is most likely due to one big term

(Catastrophy principle)

2. Proof via LST and Bingham-Doney

If $\mathsf{P}(B > x) \sim x^{-\nu} L(x), \quad x \to \infty$, $1 < \nu < 2$,

then (Bingham-Doney):

$$
\begin{aligned}
1 - \mathsf{E}[\mathrm{e}^{-sB^{res}}] &= 1 - \frac{1 - \mathsf{E}[\mathrm{e}^{-sB}]}{s\mathsf{E}B} \\
&\sim -\frac{\Gamma(1-\nu)}{\mathsf{E}B} s^{\nu-1} L(1/s), \quad s \downarrow 0,
\end{aligned}
$$

because

$$
\mathsf{E}[\mathrm{e}^{-sB}] - 1 + s\mathsf{E}B \sim -\Gamma(1-\nu) s^{\nu} L(\frac{1}{s}), \ s \downarrow 0.
$$

2. Proof via LST and Bingham-Doney

If $\mathsf{P}(B > x) \sim x^{-\nu} L(x), \quad x \to \infty$**,**

then (Bingham-Doney):

$$\begin{aligned} 1 - \mathsf{E}[e^{-sB^{res}}] &= 1 - \frac{1 - \mathsf{E}[e^{-sB}]}{s\mathsf{E}B} \\ &\sim -\frac{\Gamma(1-\nu)}{\mathsf{E}B} s^{\nu-1} L(1/s), \quad s \downarrow 0. \end{aligned}$$

Combine with

$$\mathsf{E}[e^{-sW}] = \frac{1-\rho}{1-\rho\mathsf{E}[e^{-sB^{res}}]} = \frac{1-\rho}{1-\rho+\rho(1-\mathsf{E}[e^{-sB^{res}}])},$$

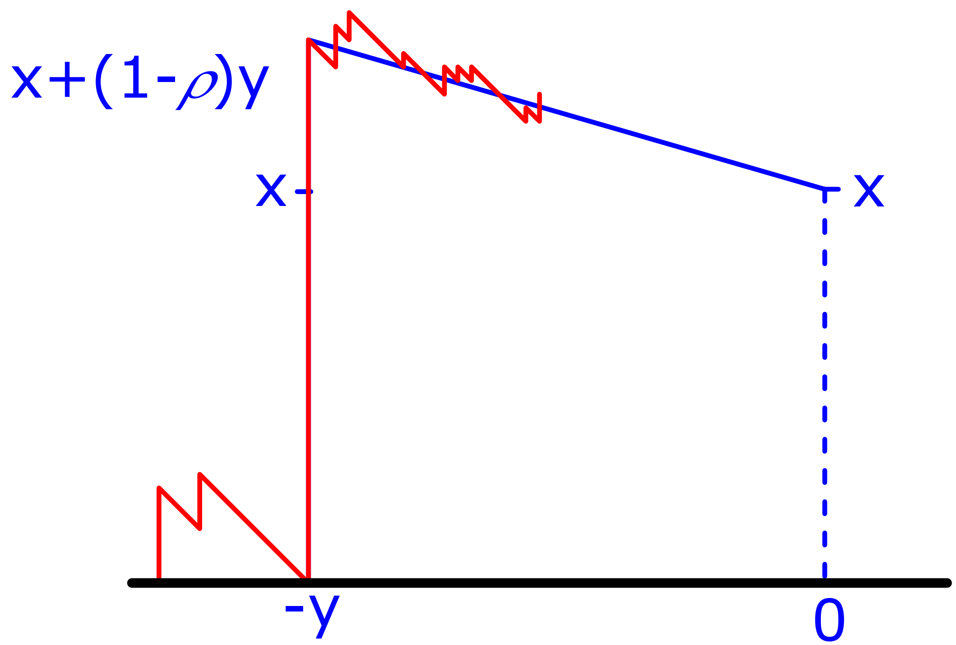
to get, for $s \downarrow 0$**:**

$$1 - \mathsf{E}[e^{-sW}] \sim -\frac{\rho}{1-\rho} \frac{\Gamma(1-\nu)}{\mathsf{E}B} s^{\nu-1} L(1/s).$$

Now again apply Bingham-Doney.

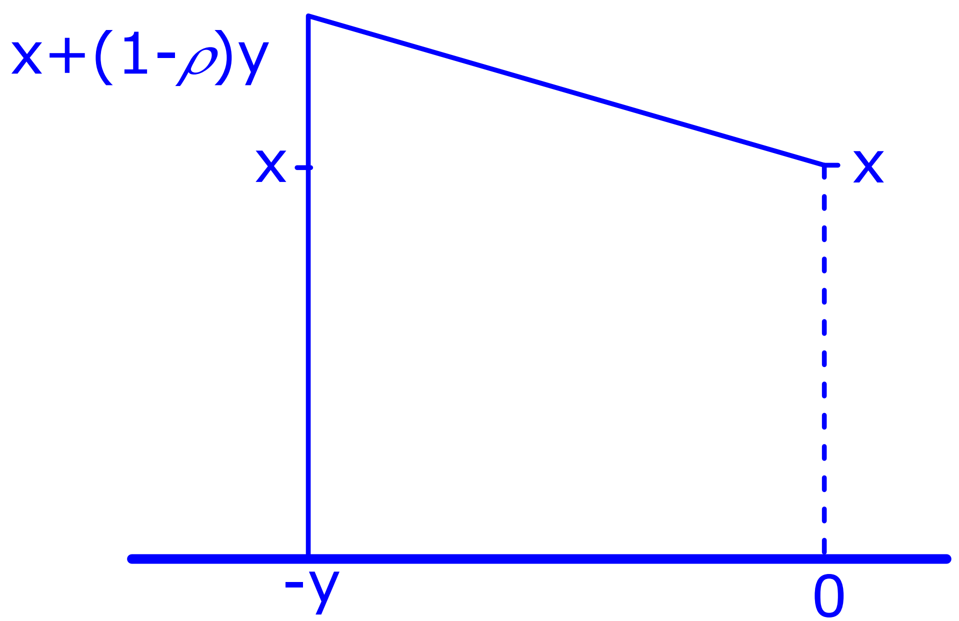
3. Proof via sample-path argument

$\mathsf{P}(W > x) = \mathsf{P}(V > x), \quad x > 0$ **(PASTA)**



3. Proof via sample-path argument

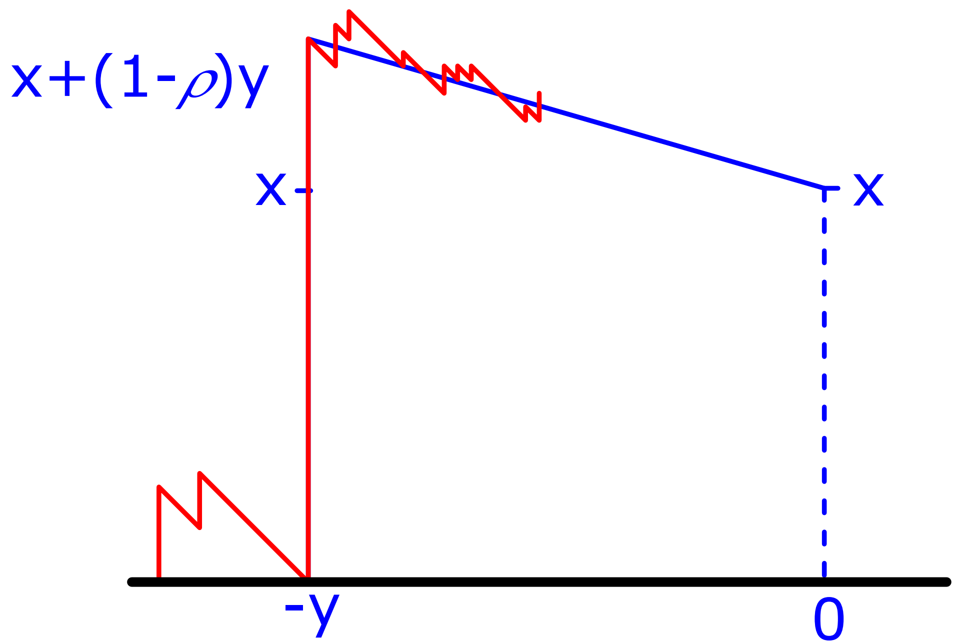
$\mathsf{P}(W > x) = \mathsf{P}(V > x), \quad x > 0$ **(PASTA)**



$$\mathsf{P}(V > x) \sim \int_{y=0}^{\infty} \mathsf{P}(B > x + (1-\rho)y)\lambda \mathsf{d}y$$

3. Proof via sample-path argument

$\mathsf{P}(W > x) = \mathsf{P}(V > x), \quad x > 0$ **(PASTA)**



$$\mathsf{P}(V > x) \sim \int_{y=0}^{\infty} \mathsf{P}(B > x + (1-\rho)y)\lambda \mathrm{d}y$$

$$= \int_{z=x}^{\infty} \mathsf{P}(B > z)\lambda \frac{\mathrm{d}z}{1-\rho} = \frac{\rho}{1-\rho}\int_{z=x}^{\infty} \frac{\mathsf{P}(B > z)}{\mathsf{E}B}\mathrm{d}z.$$

Rigorous proof: provide a lower and upper bound that asymptotically coincide.

Lower bound: "easy".

$$\mathsf{P}(V > x) \geq \frac{\rho}{1-\rho+\delta} \int_{x(1+\epsilon)}^{\infty} \frac{\mathsf{P}(B > z)}{\mathsf{E}B} \mathrm{d}z.$$

Use the Law of Large Numbers to show that one big jump gives this lower bound.

Rigorous proof: provide a lower and upper bound that asymptotically coincide.

Lower bound: “easy”.

$$\mathsf{P}(V > x) \geq \frac{\rho}{1 - \rho + \delta} \int_{x(1+\epsilon)}^{\infty} \frac{\mathsf{P}(B > z)}{\mathsf{E}B} \mathsf{d}z.$$

Use the Law of Large Numbers to show that one big jump gives this lower bound: $A(-y, 0) > (\rho - \delta)y$ **a.s.**

$$\mathsf{P}(V > x) \geq \int_{y=0}^{\infty} \mathsf{P}(B + A(-y, 0) - y > x)\lambda \mathsf{d}y$$

Rigorous proof: provide a lower and upper bound that asymptotically coincide.

Lower bound: "easy".

$$\mathsf{P}(V > x) \geq \frac{\rho}{1-\rho+\delta} \int_{x(1+\epsilon)}^{\infty} \frac{\mathsf{P}(B > z)}{\mathsf{E}B} \mathrm{d}z.$$

Use the Law of Large Numbers to show that one big jump gives this lower bound: $A(-y,0) > (\rho - \delta)y$ **a.s.**

$$\begin{aligned}
\mathsf{P}(V > x) &\geq \int_{y=0}^{\infty} \mathsf{P}(B + A(-y,0) - y > x)\lambda \mathrm{d}y \\
&\geq \int_{y=0}^{\infty} \mathsf{P}(B > x(1+\epsilon) + y(1-\rho+\delta)) \\
&\quad \mathsf{P}(A(-y,0) - \rho y \geq -\epsilon x - \delta y)\lambda \mathrm{d}y
\end{aligned}$$

Rigorous proof: provide a lower and upper bound that asymptotically coincide.

Lower bound: “easy”.

$$\mathsf{P}(V > x) \geq \frac{\rho}{1-\rho+\delta}\int_{x(1+\epsilon)}^{\infty}\frac{\mathsf{P}(B > z)}{\mathsf{E}B}\mathrm{d}z.$$

Use the Law of Large Numbers to show that one big jump gives this lower bound: $A(-y,0) > (\rho-\delta)y$ **a.s.**

$$\begin{aligned}
\mathsf{P}(V > x) &\geq \int_{y=0}^{\infty}\mathsf{P}(B + A(-y,0) - y > x)\lambda\mathrm{d}y \\
&\geq \int_{y=0}^{\infty}\mathsf{P}(B > x(1+\epsilon) + y(1-\rho+\delta)) \\
&\quad\ \mathsf{P}(A(-y,0) - \rho y \geq -\epsilon x - \delta y)\lambda\mathrm{d}y \\
&\geq (1-\tilde{\epsilon})\int_{y=0}^{\infty}\mathsf{P}(B > x(1+\epsilon) + y(1-\rho+\delta))\lambda\mathrm{d}y.
\end{aligned}$$

Rigorous proof: provide a lower and upper bound that asymptotically coincide.

Lower bound: "easy".

$$\mathsf{P}(V > x) \geq \frac{\rho}{1-\rho+\delta}\int_{x(1+\epsilon)}^{\infty} \frac{\mathsf{P}(B > z)}{\mathsf{E}B}\mathrm{d}z.$$

Upper bound: "hard".

$$\mathsf{P}(V > x) \leq \frac{\rho}{1-\rho-\delta}\int_{x(1-\epsilon)}^{\infty} \frac{\mathsf{P}(B > z)}{\mathsf{E}B}\mathrm{d}z + \mathrm{o}(x^{1-\nu}).$$

Include all other scenario's (like two big jumps) and show that they can (asymptotically) be neglected.

De Meyer and Teugels (1980):

$$\mathsf{P}(P > x) \sim \frac{1}{1-\rho}\mathsf{P}(B > (1-\rho)x), \quad x \to \infty.$$

Proof 2: apply Bingham-Doney to $\mathsf{E}[\mathrm{e}^{-sP}]$**.**
$\mathsf{E}[\mathrm{e}^{-sP}]$ **satisfies**

$$x = \mathsf{E}[\mathrm{e}^{-(s+\lambda(1-x))B}].$$

De Meyer and Teugels (1980):

$$\mathsf{P}(P > x) \sim \frac{1}{1-\rho}\mathsf{P}(B > (1-\rho)x), \quad x \to \infty.$$

Proof 2: apply Bingham-Doney to $\mathsf{E}[\mathrm{e}^{-sP}]$**.**
$\mathsf{E}[\mathrm{e}^{-sP}]$ **satisfies**

$$x = \mathsf{E}[\mathrm{e}^{-(s+\lambda(1-x))B}].$$

This implies ($1 < \nu < 2$**):**

$$\mathsf{E}[\mathrm{e}^{-sP}] - 1 + s\frac{\mathsf{E}B}{1-\rho} \sim -\Gamma(1-\nu)(1-\rho)^{-(\nu+1)}s^{\nu}L(\frac{1}{s}), \quad s \downarrow 0.$$

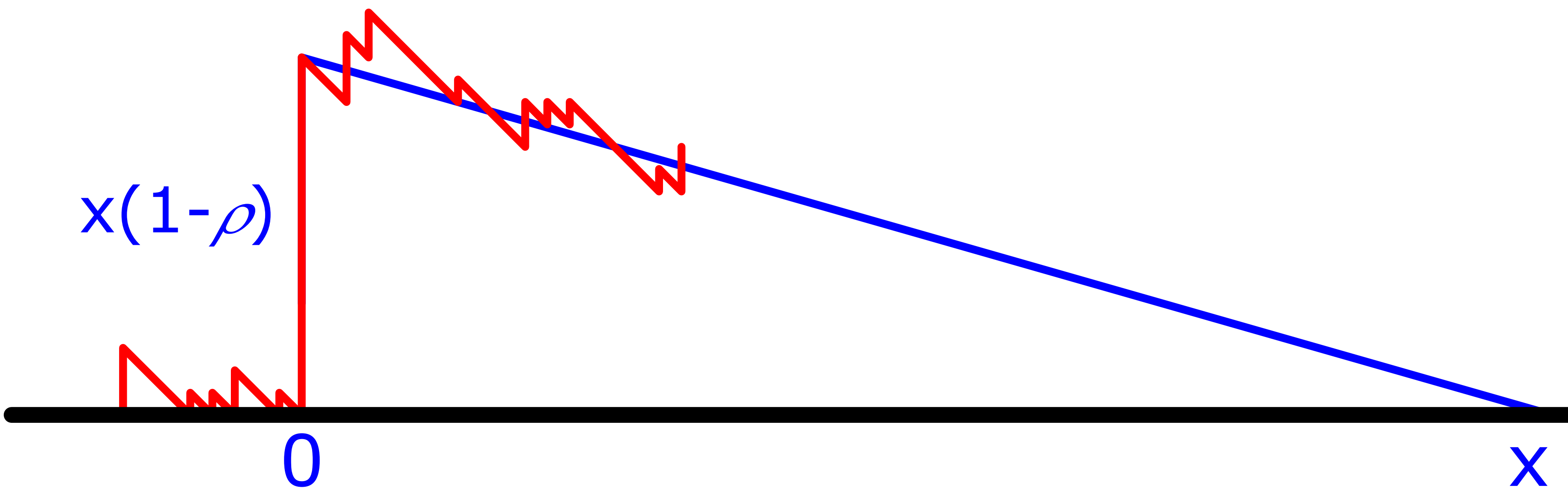
Hence

$$\mathsf{P}(P > x) \sim \frac{1}{1-\rho}((1-\rho)x)^{-\nu}L(x), \quad x \to \infty.$$

De Meyer and Teugels (1980):

$$\mathsf{P}(P > x) \sim \frac{1}{1-\rho}\mathsf{P}(B > (1-\rho)x), \qquad x \to \infty.$$

Proof 3: sample-path argument

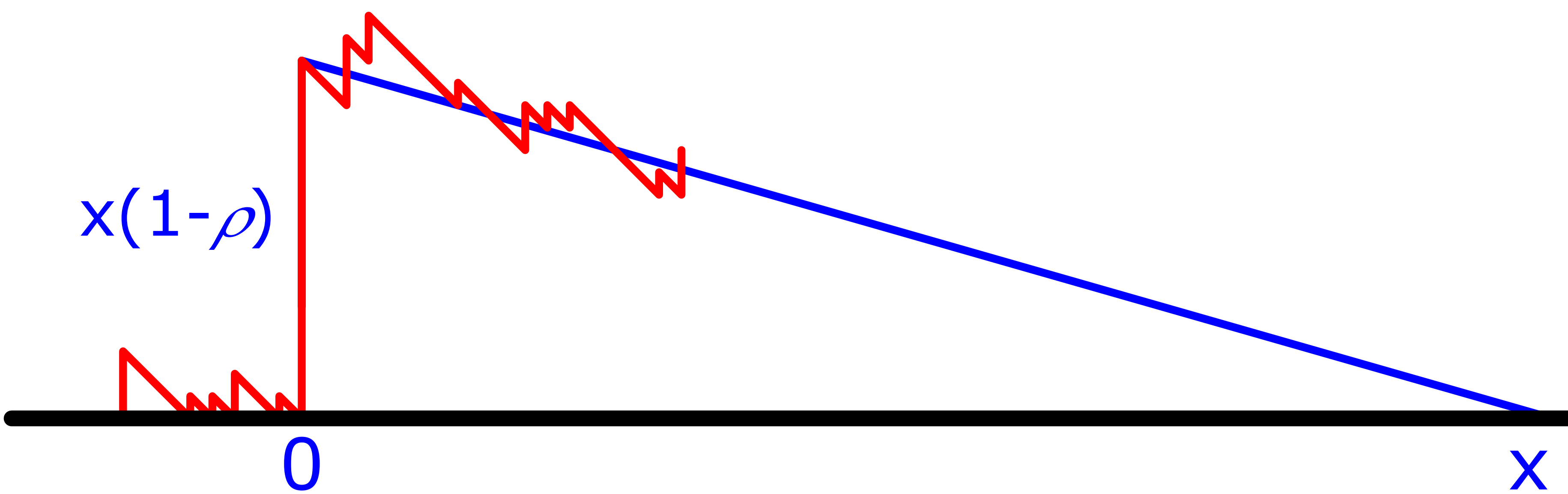


$M/G/1$: busy period

De Meyer and Teugels (1980):

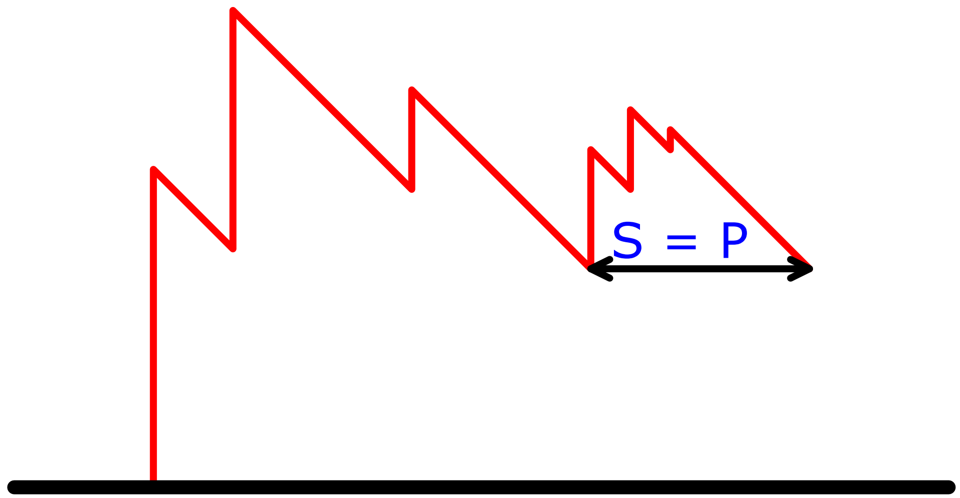
$$\mathsf{P}(P > x) \sim \frac{1}{1-\rho}\mathsf{P}(B > (1-\rho)x), \quad x \to \infty.$$

Proof 3: sample-path argument

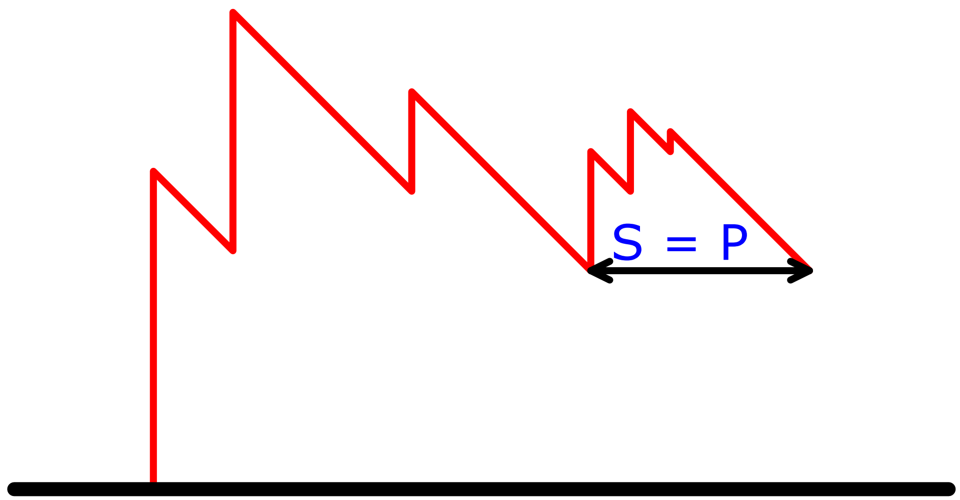


Remember that $\frac{1}{1-\rho} = \mathsf{E}$[**number of customers in busy period**].

5. $M/G/1$ LCFS preemptive resume: sojourn time



5. $M/G/1$ LCFS preemptive resume: sojourn time



Sojourn time $S =^d P$ busy period

Hence

$$\mathsf{P}(S_{LCFS-PR} > x) \sim \frac{1}{1-\rho}\mathsf{P}(B > (1-\rho)x), \quad x \to \infty.$$

Note: $O(x^{-\nu})$ instead of $O(x^{1-\nu})$!

6. Epilogue

Warning: be careful with using results like

$$P(W > x) \sim \frac{\rho}{1-\rho} P(B^{res} > x) \text{ or } Cx^{1-\nu}$$

as approximation.
Such an approximation is typically bad
(Abate, Choudhury & Whitt, QUESTA 16, 1994).

Taking a few more terms does not always help
(Willekens & Teugels, QUESTA 10, 1992).

Abate et al. suggest: Use their numerical algorithm for LST inversion, if such LST is known.

Reference:

The impact of the service discipline on delay asymptotics
Sem Borst, Onno Boxma, Sindo Núñez, Bert Zwart
to appear in Performance Evaluation, 2003