Service Engineering

April 30, 1997
Last Revised 11/11/2003

A <u>Deterministic</u> Model of a Service Station (Fluid View)

Primitives

$Z(0)$	initial content
$\alpha(t)$	input rate
$\mu(t)$	*potential* service rate

Model: (Think cumulants)

$$\text{Inflow:} \quad A(t) = \int_0^t \alpha(u)du, \quad t \geq 0\,;$$

$$\text{Potential Outflow:} \quad M(t) = \int_0^t \mu(u)du, \quad t \geq 0\,.$$

- We could start with primitives A, M, in which case they need not be continuous; for example, they could be counting processes.

Netflow: $X(t) = Z(0) + A(t) - M(t), \quad t \geq 0.$

Introduce $Y(t)$ = cumulative potential lost during $[0, t]$.

$\Rightarrow$ Outflow: $D = M - Y$ (**A** arrivals; **D** departures)

$\Rightarrow$ Balance:

$$\begin{aligned} Z(t) &= Z(0) + A(t) - D(t) \\ &= Z(0) + A(t) - [M(t) - Y(t)] \\ &= X(t) + Y(t), \qquad t \geq 0\,. \end{aligned}$$

Model $Z = X + Y$

Feasible $Z \geq 0,\ Y \uparrow 0 \quad (Y(0) = 0)$;

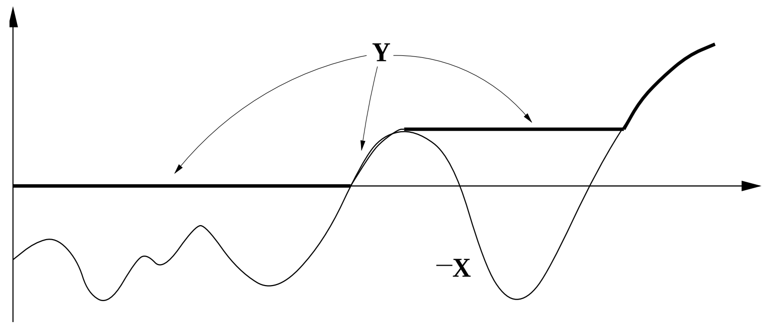
Efficient Y least (hence, Y unique);

Existence: $Y = \overline{(-X)^+} \quad (Y = -\underline{X}$, when $Z(0) = 0)$;

$\underline{X}(t) = \inf_{0 \leq u \leq t} X(u)$, which is called the **lower envelope** of X.

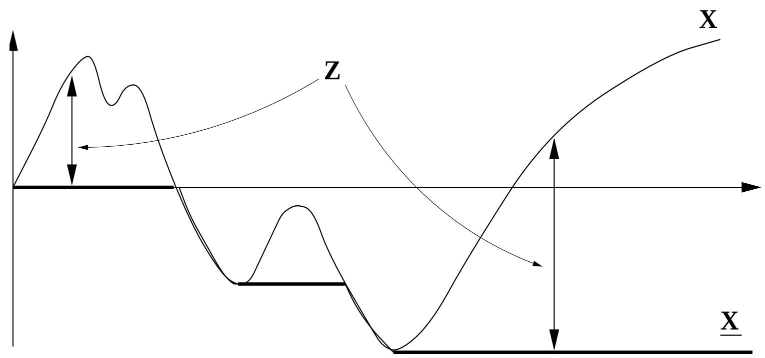
"Proof"

Least $Y \uparrow 0$
s.t. $Y \geq -X$



When $Z(0) = 0$:

$Z = X - \underline{X}$,
$\underline{X}$ = lower envelope.



Equivalent characterization via complementarity: (LCP/DCP)

$$Y \text{ least } <=> ZdY = 0, \text{ i.e. } \begin{array}{l} Y \text{ increases at } t \\ \text{only when } Z(t) = 0\,. \end{array}$$

In words: potential lost due to idleness.

Claim (Skorohod) Given $X \in$ RCLL (**R**ight **C**ontinuous **L**eft **L**imit),

there exists a unique (Y, Z) such that

$$\begin{aligned} Z &= X + Y, \\ Z &\geq 0, \quad Y \uparrow 0, \\ ZdY &= 0\,. \end{aligned}$$

Proof Existence by checking $Y = \overline{(-X)^+}$ $(= -\underline{X \wedge 0})$.

Uniqueness by Lyapunov-function argument:

(Note: if minimality is established, then uniqueness is automatic.)

If (Y_i, Z_i), $i = 1, 2$, are two solutions, then consider

$$\eta = \frac{1}{2}(Y_1 - Y_2)^2.$$

Assume, for simplicity, continuous Y_i's, in which case differentiate:

$$d\eta = (Y_1 - Y_2)(dY_1 - dY_2) = (Z_1 - Z_2)(dY_1 - dY_2) = -Z_1 dY_2 - Z_2 dY_1 \leq 0 \ .$$

Deduce that η decreases, but also

$$\eta(0) = 0 \Rightarrow \eta \equiv 0 \Rightarrow Y_1 \equiv Y_2.$$

Outflow $\quad D(t) = M(t) - Y(t) = \int_0^t \delta(u)du, \quad$ where $\delta(u) =$ outflow rate,

$$\Rightarrow \quad Y(t) = \int_0^t [\mu(u) - \delta(u)]du \ .$$

In terms of rates: $dY \geq 0$ implies $\delta \leq \mu$.

Now, either

$\delta = \mu$ or

$\delta < \mu \Leftrightarrow dY > 0,$

$\Rightarrow Z = 0$ (since $ZdY = 0$),

$\Rightarrow d(X + Y) = 0$ (consider a neighbourhood and differentiate),

$\Rightarrow (\alpha - \mu) + (\mu - \delta) = \alpha - \delta = 0.$

Thus (Hall, pg. 190, Def. 6.6),

$$\delta(t) = \begin{cases} \mu(t) & \text{when } Z(t) > 0, \\ \alpha(t) & \text{when } Z(t) = 0 \,. \end{cases}$$

Note that the above is *not* a direct definition of δ, since it uses Z, which is defined in terms of δ.

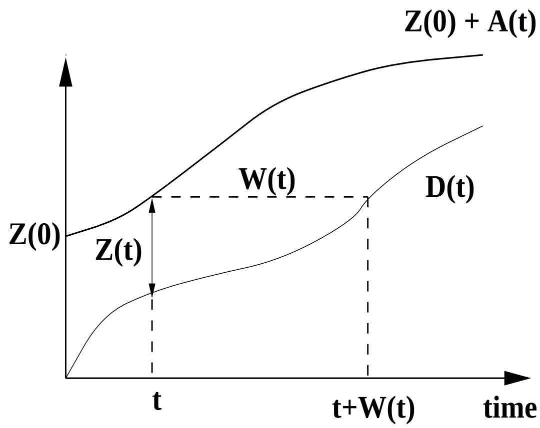
How to calculate **Delay**?

Define

$$W(t) = \text{work-load at time } t$$
$$(= \text{time to process all that is present at time } t)$$
$$= \text{under FCFS, virtual waiting time.}$$

Defining relation for W:

$$D(t + W(t)) = Z(0) + A(t)$$



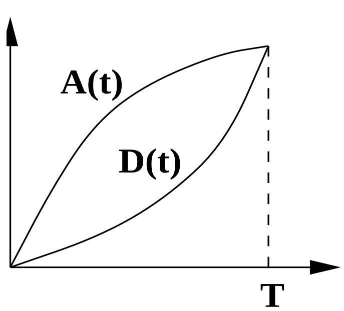
Hence, $Z(t + W(t)) = Z(0) + A(t + W(t)) - A(t)$.

MOP's over a finite horizon T:

Averages **Inflow**: $\bar{\alpha} = \frac{1}{T}\int_0^T \alpha(t)dt$;

Outflow: $\bar{\delta} = \frac{1}{T}\int_0^T \delta(t)dt$;

Throughput: λ, defined when $\bar{\alpha} = \bar{\delta}$ as their common value.



eg. $\lambda = \frac{1}{T}A(T) = \frac{1}{T}D(T)$.

Queue length (Inventory): $\bar{Z} = \frac{1}{T}\int_0^T Z(t)dt = \frac{1}{T}\times$ Area.

Delay: $\bar{W} = \frac{1}{A(T)}\int_0^T W(t)dA(t) \quad \left(= \frac{\int_0^T W(t)\alpha(t)dt}{\int_0^T \alpha(t)dt} \right).$

↑
Rieman-Stiltjes

Intuition:

- Discrete arrivals $\Rightarrow \bar{W} = \frac{1}{A(T)} \sum_{n=1}^{A(T)} W_n$ (as in Hall, Chap. 2);
- Absolutely continuous: $\alpha(t)dt$ arrivals during $(t, t+dt)$, each suffering a delay of $W(t)$.

Little's Conservation Law: $\bar{Z} = \lambda \cdot \bar{W}$.

Cumulative lost potential $Y(T)$.

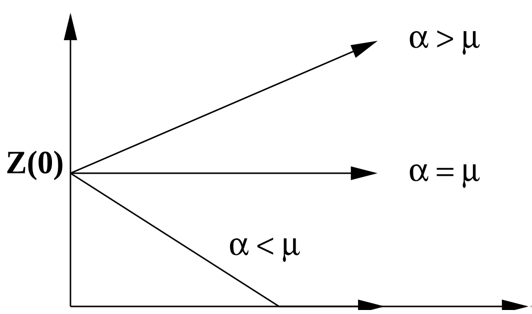
Efficiency $\varepsilon(T) = 1 - \frac{Y(T)}{M(T)} =$

actual ↘

$$= \frac{D(T)}{M(T)} \quad \left(= \frac{\int_0^T \delta(t)dt}{\int_0^T \mu(t)dt}, \text{ when applicable} \right).$$

potential ↗

Example *constant rates* $\alpha(t) \equiv \alpha$, $\mu(t) \equiv \mu$.
(linear model)



overloaded (supercritical) $(\rho > 1)$

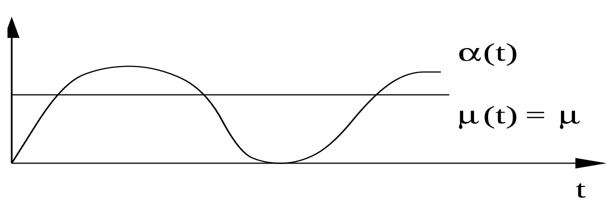
critically loaded (critical) $(\rho = 1)$

underloaded (subcritical) $(\rho < 1)$

Definition: $\rho = \alpha/\mu$ **traffic (flow) intensity**.

Natural *extension*: piecewise constant rates, as in National Cranberry (HBS case).

Example *periodic rates* e.g.



(If α has a period $T_\alpha = 8$, μ has a period $T_\mu = 3$, take period $T = T_\alpha \cdot T_\mu = 24$.)

Long-run: $\bar{\alpha} = \frac{1}{T}\int_0^T \alpha(t)dt; \quad \bar{\mu} = \frac{1}{T}\int_0^T \mu(t)dt;$

$\rho = \bar{\alpha}/\bar{\mu}$ (Heyman-Whitt).

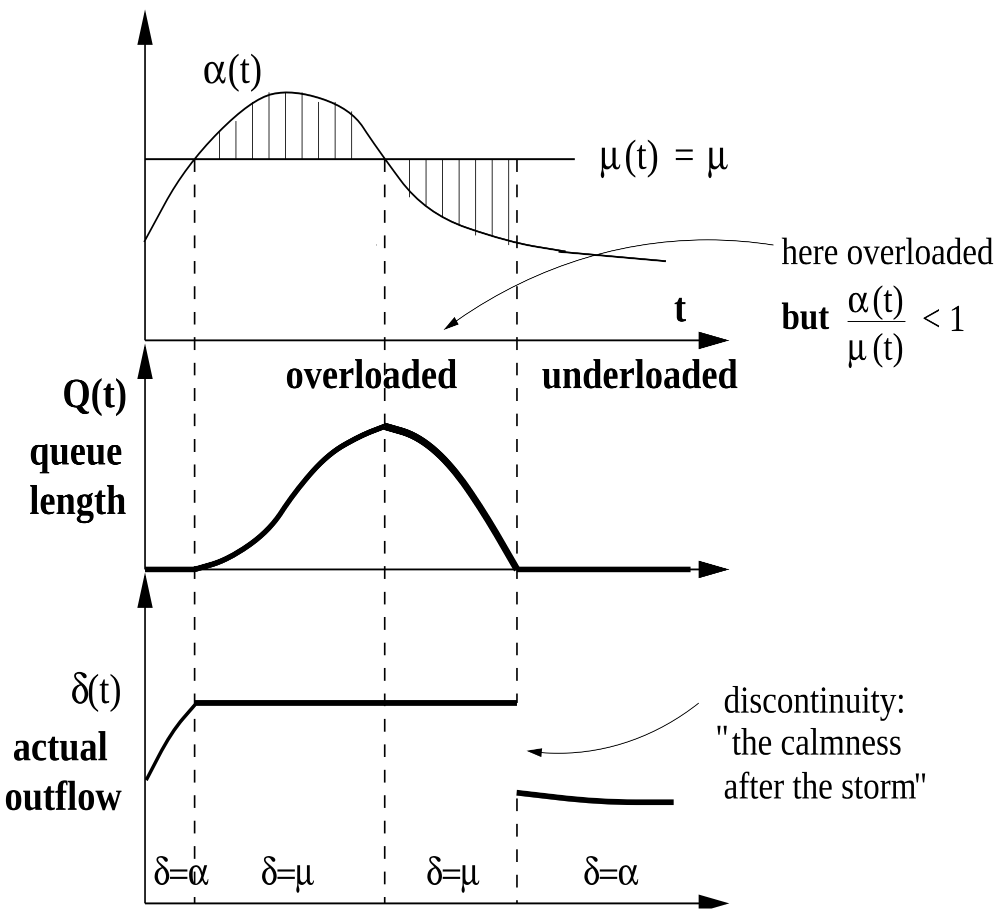
Short-run: Phase-transitions (different from Hall, pg. 189–190, that has stagnant → growth → decline → stagnant).

Short-Run Phase Transitions

Overloaded at t : $Z(t) > 0$;

Underloaded : $Z(t) = 0$ and $\delta(t) < \mu(t)$ (excess capacity, $dY(t) > 0$);

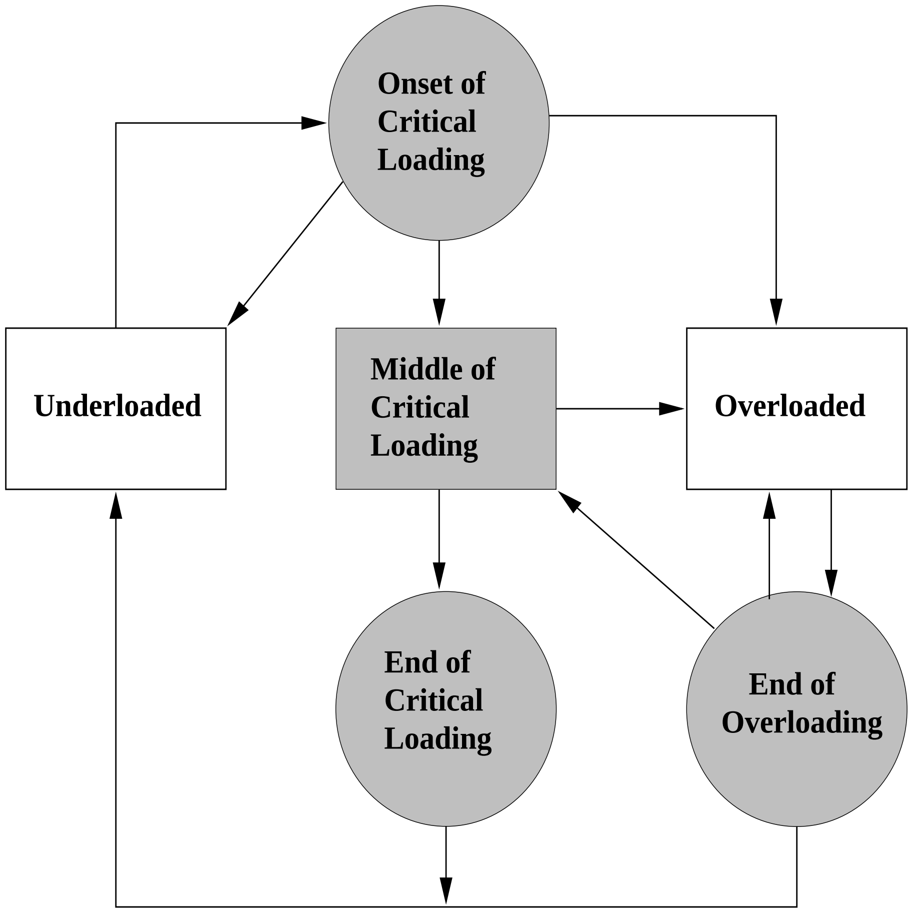
Critically loaded : $Z(t) = 0$ and $\delta(t) = \mu(t)$ (balanced capacity, $dY(t) = 0$).



The analogue of ρ, traffic intensity, is here (assume $Z(0) = 0$):

$$\rho(t) = \sup_{0 \le s \le t} \frac{\int_s^t \alpha(u)du}{\int_s^t \mu(u)du} \qquad \begin{cases} > 1 & \text{overloaded} \\ = 1 & \text{critically loaded} \\ < 1 & \text{underloaded} \end{cases}$$

For finer approximations, we must acknowledge more phases, as depicted in the following figure.



Phase transition diagram for the asymptotic regions.
(Massey & Mandelbaum.)

References:

– Hall, R.W., "*Queueing Methods for Service and Manufacturing*", Prentice Hall, 1991.

– Harrison, J.M., "*Brownian Motion and Stochastic Flow Systems*", Wiley, 1985.

– Mandelbaum, A. and Massey, William, A., "Strong approximations for time-dependent queues", *Math. of Operations Research,* 20, 33-64, 1995.

Mathematical Framework

Reflection Mapping (Regulator) $\quad X \to X - \underline{X \wedge 0}$

$$(X \to X - \underline{X}\,, \quad \text{when } X(0) = 0)\,.$$

Fundamental:

- Flow analysis (Fluid Models);
- Economics;
- Stochastic Processes;
  - Skorohod (needed cumulant Y!);
  - Queueing Models (later);
- Approximations.

Idea of Approximations: $\quad Z = f(X), \quad f$ continuous (Lipshitz).

Hence, $\quad X \approx \tilde{X}$ implies $Z \approx \tilde{Z} = f(\tilde{X})$

$$X \approx \bar{X} \quad \text{fluid} \quad \Rightarrow \bar{Z} = f(\bar{X}) \quad \text{fluid approximations.}$$

$$X \approx \bar{X} + \hat{X} \text{ diffusion} \quad \Rightarrow \hat{Z} = f(\bar{X} + \hat{X}) \quad \text{diffusion refinements.}$$

Reference: Harrison, Chapter 2 (which covers also finite buffers, and two-node networks).