Economy of Scale
in Multiserver Service Systems:
A Retrospective

Ward Whitt

IEOR Department

Columbia University

Ancient Relics

- A. K. Erlang (1924) On the rational determination of the number of circuits. In The Life and Works of A. K. Erlang, E. Brockmeyer, H. L. Halstrom and A. Jensen (editors), Danish Academy of Technical Sciences, Copenhagen, 1948.

- S. Halfin and W. Whitt (1981) Heavy-traffic limits for queues with many exponential servers. Operations Research 29, 567-588.

- D. R. Smith and W. Whitt (1981) Resource sharing for efficiency in traffic systems. The Bell System Technical Journal 60, 39-55.

- T. Levitt (1972) Production-line approach to service. Harvard Business Review, September-October, 41-52.

Customer Contact Centers

- Telephone Call Centers

- e-Contact

N. Gans, G. Koole and A. Mandelbaum (2002) Telephone call centers: a tutorial and literature review. Manufacturing and Service Operations Management (M&SOM), to appear.

Economy of Scale

- **Qualitative**

Bigger is Better

- **Quantitative**

Bigger is How Much Better?

The $\mathrm{M/M/s/0}$ and $\mathrm{M/M/s/\infty}$ Models

Parameters

$$
\begin{aligned}
\lambda &= \text{arrival rate} \\
\mu &= \text{service rate} \\
s &= \text{number of servers} \\
a &= \lambda/\mu = \text{offered load} \\
\rho &= a/s = \text{traffic intensity}
\end{aligned}
$$

Erlang Loss Formula

steady-state blocking probability

$$B(s, \lambda, \mu) = B(s, a) = \frac{a^s/s!}{\sum_{k=0}^{k=s} a^k/k!}$$

truncated Poisson distribution

$$B(s, a) = \frac{aB(s, a)}{s + aB(s-1, a)},$$

where $B(0, a) = 1$ **and** $a = \lambda/\mu$**.**

recursion for efficient computation

Erlang Delay Formula

steady-state delay probability
i.e., probability an arrival must wait before beginning service

$$C(s, \lambda, \mu) = C(s, a) = \frac{B(s, a)}{1 - \rho + \rho B(s, a)}$$

expected steady-state waiting time:

$$EW(s, \lambda, \mu) = C(s, a) \frac{1}{\mu(1 - \rho)}$$

(Can start from $B(s, a)$).

Economy of Scale

- **Qualitative**

Bigger is Better

- **Quantitative**

Bigger is How Much Better?

Qualitative Economy of Scale

$$B(s_1+s_2,\lambda_1+\lambda_2,\mu) \le \frac{\lambda_1}{\lambda_1+\lambda_2}B(s_1,\lambda_1,\mu) + \frac{\lambda_2}{\lambda_1+\lambda_2}B(s_2,\lambda_2,\mu)$$

$$C(s_1+s_2,\lambda_1+\lambda_2,\mu) \le \frac{\lambda_1}{\lambda_1+\lambda_2}C(s_1,\lambda_1,\mu) + \frac{\lambda_2}{\lambda_1+\lambda_2}C(s_2,\lambda_2,\mu)$$

$$EW(s_1+s_2,\lambda_1+\lambda_2,\mu) \le \frac{\lambda_1}{\lambda_1+\lambda_2}EW(s_1,\lambda_1,\mu) + \frac{\lambda_2}{\lambda_1+\lambda_2}EW(s_2,\lambda_2,\mu)$$

D. R. Smith and W. Whitt (1981)

Qualitative Economy of Scale: Short Proofs

The same argument works for all three.

One-dimensional monotonicity:

$$B(ts, ta) \quad \textbf{is decreasing in} \quad t\ .$$

Convexity in s:

$$B\left(\frac{\lambda_1 s_1 + \lambda_2 s_2}{\lambda_1 + \lambda_2}, a\right) \leq \frac{\lambda_1}{\lambda_1 + \lambda_2} B(s_1, a) + \frac{\lambda_2}{\lambda_1 + \lambda_2} B(s_2, a)$$

for $0 < \lambda_i < \infty$, $i = 1, 2$.

Jagers and van Doorn (1986, 1991)

Qualitative Economy of Scale: Short Proofs

Need to extend $B(s, a)$ to all real positive s:

$$\frac{1}{B(s, a)} = a \int_0^\infty (1+t)^s e^{-at} \, dt$$

Jagerman (1974)

Use to prove:

- **one-dimensional monotonicity**

- **convexity in real s.**

Qualitative Economy of Scale: Short Proofs

The same argument works for all three (B, C and EW).

$$
\begin{aligned}
B(s_1+s_2,\lambda_1+\lambda_2,\mu) &\leq \frac{\lambda_1}{\lambda_1+\lambda_2} B\left(\frac{\lambda_1+\lambda_2}{\lambda_1} s_1, \lambda_1+\lambda_2, \mu\right) \\
&\quad + \frac{\lambda_2}{\lambda_1+\lambda_2} B\left(\frac{\lambda_1+\lambda_2}{\lambda_2} s_2, \lambda_1+\lambda_2, \mu\right) \\
&\leq \frac{\lambda_1}{\lambda_1+\lambda_2} B\left(s_1, \lambda_1+\lambda_2, \mu\right) \\
&\quad + \frac{\lambda_2}{\lambda_1+\lambda_2} B\left(s_2, \lambda_1+\lambda_2, \mu\right) \quad .
\end{aligned}
$$

Use convexity in step 1.

Use one-dimensional monotonicity in step 2.

Stochastic-Comparison Proof: Loss Model

Little's Law: $L = \lambda W$

$$EQ(s, \lambda, \mu) = \lambda(1 - B(s, \lambda, \mu))\mu^{-1}$$

or, equivalently,

$$\lambda B(s, \lambda, \mu) = \lambda - \mu EQ(s, \lambda, \mu) \ ,$$

so that

$$(\lambda_1 + \lambda_2)B(s_1 + s_2, \lambda_1 + \lambda_2, \mu) \\ \leq \lambda_1 B(s_1, \lambda_1, \mu) + \lambda_2 B(s_2, \lambda_2, \mu)$$

if and only if

$$EQ(s_1 + s_2, \lambda_1 + \lambda_2, \mu) \\ \geq EQ(s_1, \lambda_1, \mu) + EQ(s_2, \lambda_2, \mu)$$

Suffices to show: $EQ_{1+2} \geq EQ_1 + EQ_2$

Stochastic-Comparison Proof: Delay Model

Little's Law: $L = \lambda W$

$$EQ(s, \lambda, \mu) = \lambda(EW(s, \lambda, \mu) + \mu^{-1})$$

or, equivalently,

$$\lambda EW(s, \lambda, \mu) = EQ(s, \lambda, \mu) - \lambda\mu^{-1} \ ,$$

so that

$$(\lambda_1 + \lambda_2)EW(s_1 + s_2, \lambda_1 + \lambda_2, \mu) \\ \le \lambda_1 EW(s_1, \lambda_1, \mu) + \lambda_2 EW(s_2, \lambda_2, \mu)$$

if and only if

$$EQ(s_1 + s_2, \lambda_1 + \lambda_2, \mu) \\ \le EQ(s_1, \lambda_1, \mu) + EQ(s_2, \lambda_2, \mu) \ .$$

Suffices to show: $EQ_{1+2} \le EQ_1 + EQ_2$

Stochastic-Comparison Proofs: Summary

By Little's Law, $L = \lambda W$, we need to show:

For Loss Model, $EQ_{1+2} \geq EQ_1 + EQ_2$

For Delay Model, $EQ_{1+2} \leq EQ_1 + EQ_2$

We establish stronger results:

For Loss Model, $Q_{1+2} \geq_{st} Q_1 + Q_2$

For Delay Model, $Q_{1+2} \leq_{st} Q_1 + Q_2$

stochastic order: $X \leq_{st} Y$ **if** $Ef(X) \leq Ef(Y)$ **for all nondecreasing real-valued functions** f**.**

Stronger Stochastic Comparisons

We show:

For Loss Model, $Q_{1+2} \geq_{st} Q_1 + Q_2$

For Delay Model, $Q_{1+2} \leq_{st} Q_1 + Q_2$

by establishing two stronger comparisons:

- **likelihood-ratio ordering,** $\leq_r$ $(\mathbf{M/M/s}/*)$

$X \leq_r Y$ **if** $\dfrac{P(X{=}k{+}1)}{P(X{=}k)} \leq \dfrac{P(Y{=}k{+}1)}{P(Y{=}k)}$ **for all** k.

- **sample-path stochastic order,** $\leq_{st}$ $(\mathbf{A/GI/s}/*)$

$\{X(t) : t \geq 0\} \leq_{st} \{Y(t) : t \geq 0\}$ **if**
$E[f(\{X(t) : t \geq 0\})] \leq E[f(\{Y(t) : t \geq 0\})]$
for all nondecreasing real-valued functions f**.**

Economy of Scale

- **Qualitative**

Bigger is Better

- **Quantitative**

Bigger is How Much Better?

Quantitative Economy of Scale

Perform asymptotics as $a \to \infty$ and $s \to \infty$.

Heavy Traffic: $\rho = a/s \to 1$

with $(1-\rho)\sqrt{s} \to \beta$ **for** $-\infty < \beta < \infty.$

Square-Root Safety Factor

We should have $s \approx a + c\sqrt{a}$ **for** $c = c(\beta)$**.**

known by Erlang (1924)

Quantitative Economy of Scale: Loss Model

As $a \longrightarrow \infty$**,**

$$B(a + c\sqrt{a}, a) = \left(\frac{1}{\sqrt{a}}\right)\left(\frac{\phi(c)}{\Phi(c)}\right) + O\left(\frac{1}{a}\right)$$

Erlang (1924), Jagerman (1974)

$\Phi(x) = P(N(0,1) \leq x)$, **normal cdf**

ϕ **normal density,** $\Phi(x) = \int_{-\infty}^{x} \phi(u)\, du$

Quantitative Economy of Scale: Delay Model

As $a \to \infty$ (or $\lambda \to \infty$ with μ fixed),

$$C(a + c\sqrt{a}, a) = \left[1 + c\frac{(1 - \Phi(c))}{\phi(c)}\right]^{-1} + \quad O\left(\frac{1}{\sqrt{a}}\right)$$

Erlang (1924), Halfin and Whitt (1981)

$$EW(\lambda + c\sqrt{\lambda}, \lambda, 1) = \left(\frac{c}{\sqrt{a}}\right)\left[1 + c\frac{(1 - \Phi(c))}{\phi(c)}\right]^{-1} + \quad O\left(\frac{1}{a}\right)$$

Erlang (1924), Halfin and Whitt (1981)

$\Phi(x) = P(N(0,1) \leq x)$, **normal cdf**

ϕ **normal density**, $\Phi(x) = \int_{-\infty}^{x} \phi(u)\, du$

Stochastic-Process Limit

As $a \rightarrow \infty$ **and** $s \rightarrow \infty$

with $(1-\rho)\sqrt{s} \rightarrow \beta$ **for** $-\infty < \beta < \infty$**,**

$$\frac{Q_a(t) - a}{\sqrt{a}} \Rightarrow L(t),$$

where L **is a diffusion process.**

(convergence for stochastic processes)

Recent Work

Halfin and Whitt did stochastic-process limit for $\mathrm{GI}/\mathrm{M}/\mathrm{s}/\infty$.

Now extend from $\mathrm{GI}/\mathrm{M}/\mathrm{s}/\infty$ to $\mathrm{G}/\mathrm{H}^*/\mathrm{s}/\infty$

http://www.research.att.com/~wow