QUEUEING THEORY REVISITED -- SOME NEW RESULTS

W Blunden
Emeritus Professor

U Vandebona
Lecturer
University of
New South Wales
Sydney, New South Wales

ABSTRACT

Queueing Theory provides an unassailable rationale for most discrete flow problems -- cars on the road, customers at cash-register channels, travellers at baggage counters, trucks at loading docks, ships in port. Most importantly it demonstrates that traffic congestion is not an evil, per se, but rather a feedback mechanism that spreads the peak, changes the route or mode and eventually brings the land-use and transport resources into balance. It is also a most useful tool for the operational analysis and functional design of transport facilities and systems. It is surprising therefore that it is not more generally applied in the on-the-job sense.

One reason perhaps is that, with the exception of the most simple situations, the analytical results are clumsy and awkward to evaluate. In this paper the authors discuss the results of some recent research on the development of simple and widely applicable formulae that give accurate results for both single and multiple channel queueing systems and take account of a wide range of traffic handling characteristics. The intractable problem of time dependent demands and computer simulation techniques are also discussed.

INIRODUCIION

Queueing is without doubt the most widely observed phenomen in transport processes - vehicles at intersections, ships moving through a port, trucks using loading/unloading docks, passengers at baggage counters, customers at telling windows and cash register channels, motorists at toll barriers, aircraft approaching runways and numerous other transport operations, even relatively static ones such as parking. All these traffic elements have a unique common characteristic - their capacity i.e. the maximum rate at which they can pass traffic. Queueing occurs when trucks, ships, cars, people arrive in a more or less random manner to use the facilities described above. Ihe manner of specifying their manner of arrival must be time specific i.e. as a rate of arrival or demand. Much philosophical and analytical confusion arises from a failure to distinguish clearly between the rate of demand and the total demand.

Queueing systems fall into two main categories - single channel and multiple channel. Traffic flow facilities - intersections, roads, train tracks are included in the former; and terminal facilities - docks, service counters, parking lots the latter.

IHE RLVISIIAIION

The theory of queues and waiting lines is rigorously based on probability theory and its application at the turn of the century to traffic was due to the pioneering work of A. K. Erlang, a Danish scientist and telephone engineer. It is not surprising therefore that its initial application was to telephone traffic problems. It was embraced world wide and has been supported by an enormous literature of high quality. What is surprising however is the limited use made of it in the field of transport. The first two editions of the Highway Capacity Manual make no use of it whatsoever and the current massive Third Edition accords it one passing reference. During the second half of the century a good deal of repioneering has been carried out by staff and students of the School of Traffic Engineering of the University of New South Wales (1,2,3,4,5,6) and by some notable workers elsewhere (7, 8,9,10). Even so it does not seem to have been embraced with much enthusiasm by the professional traffic engineer and planner. Hence the motive for this revisitation.

In a conceptual and philosophical sense queueing theory provides the analyst and engineer with the Hooks Law or Ohms Law of traffic. It establishes the basic characteristic curve that relates the delay, travel time or cost of passage along or through any transport facility to the time intensity of the traffic demand. This characteristic is univerally applicable to the wide range of transport devices and facilities mentioned above and is of the form shown in Figure 1. It has very important general properties -

(a) All the curves in the queueing family are highly non-linear - at low values of the traffic intensity the extra delay (i.e. the queueing delay), which represents in a general way the inefficiency of the flow or handling process will increase slowly at first but as the traffic intensity approaches unity (i.e. saturation) the increase in delay is dramatic and at the onset of saturation becomes infinite even though the facility itself may still be passing traffic quite satisfactorily.

(b) The non-linearity of these curves is accentuated as they move to the right. This results from successive curves representing flow situations in which the quality of the flow process is "better" - either the demand stream is more regular and/or the

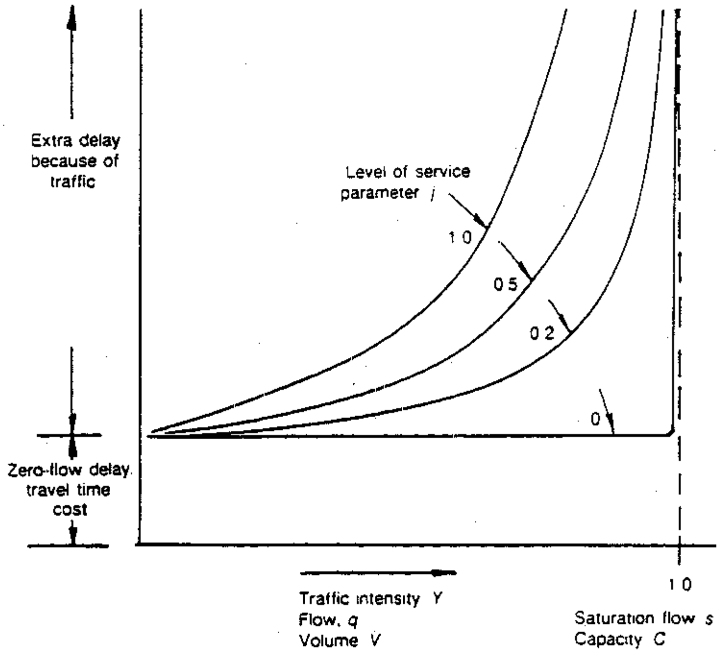

Fig. 1 Ihe Queueing Characteristic

service operation more uniform. If both were "perfect" there would be no queueing delay over the whole range of flow (no load to full load); but as soon as saturation is reached the queue would build-up indefinitely along the capacity asymptote

These fundamental properties require that the flow process is statistically stationary i.e. the mean demand rate remains steady over all time. If this be so then the non-linearity results in a powerful feedback influence that guarantees the stability of the flow operation by limiting the demand on the particular facility by shifting some of it to alternative routes or different modes or other destinations. In such circumstances there would be no need to hiss out the term traffic congestion as if it were a sin but instead consider the merits (or otherwise) of the adjusted situation.

But in practice traffic demands do not persist indefinitely. However as will be shown later even quite short overloads cause very rapid increases in delay and the feedback effects are strong enough to maintain stability by first causing the demand to spread-out in time so as to maintain its instantaneous rate at just around capacity. The implications of these important considerations are intuitively understood by professional transport operator and the experienced road user who as a result resort to better time scheduling of their transport activities. However at the political and bureaucratic planning level the high degree of dependence of the voting public on the motor car and the paranoia that congestion engenders all to oiten influences the decision maker to opt for expensive and ineffectual solutions - the Sydney Harbour Tunnel Project is a good example of the need to fully understand the implications of congestion feedback.

But quite apart from the basic conceptual significance of the queueing theory rationale in overall transport system planning the detailed analytical assessment and economic evaluation of nearly all transport schemes and projects calls for meaningful calculations of operating costs. These depend on the analysts ability to make accurate assessments of travel times and delays under normal operating conditions i.e. at traffic intensities below saturation. This return visit to the queueing theory mansion has provided the authors with the opportunity to look more closely at some of its treasures and to discover some new ones.

THE STATE OF THE ART

The starting point for all queueing analysis is knowledge of the form of the distributions of the arrivals and the service times (time to buy a ticket, pay a toll, start-up at an intersection). These distributions are many and varied but those for particulat operations and processes are surprisingly "immutable". Many such distributions have been obtained experimentally by staff and students at the University of New South Wales (5,6) The distribution function is generally of the Gamma type and may be conveniently approximated by the Erlang family shown in Figure 2. These distributions are particularly apposite for queueing applications as they are very easy to generate for computer simulation studies.

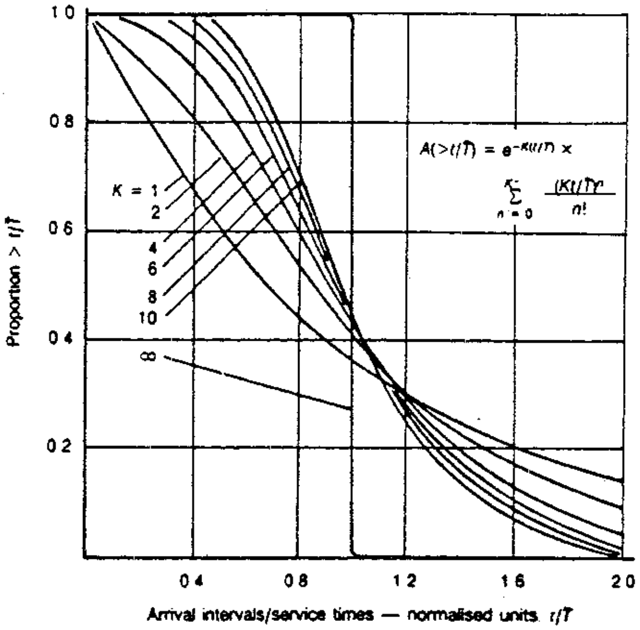

Fig. 2 Family of Erlang arrival/service distributions

The describing parameters of the above distributions are the -

- the mean arrival interval/service time, TBAR
- the Erlang K Number which ranges from 1 for an exponential distribution to infinity for a uniform one.

Leaving aside for the present peak and other highly time dependent demands it is generally safe to assume that the arrival distributions are exponentially distributed. Service or handling distributions are more varied but in the absence of specific data they may be conveniently divided into two main categories -

- practiced and repetitous tasks such as starting-up at intersections, paying a toll, boarding a bus have high Erlang Numbers and for initial calculations may be assumed to be constant (i.e. K = infinity);

- tasks which are influenced by many "independent" factors - checking a trolly of groceries, length of stay in a parking space, baggage check-in tend to randomness (i.e. K = 1).

The former are generally associated with traffic flow processes and single channel operation and the latter with terminal facilities with many channels.

Io facilitate comparisons of the performance characteristics the independent variable is usually represented as the traffic intensity which is the ratio of the mean arrival rate to the total capacity rate (all channels and is designated here as Y. The output (i.e. the delay) is expressed in mean service time units - IBAR. The determining parameters of the whole range of queueing devices then reduce to just two - the Erlang Number, K and the Number of channels, M. The principal measures of performance are the average delay in the queue - WBAR and the probability that the system is full, often known as the Erlang Loss Probability - PL. Both are extremely useful for assess-the economic merit of additional economic investment - the former especially relevant to flow facilities and the latter in the evaluation of terminals.

There are three results of great significance to the analyst - philosophically, practically and computationally -

(a) The Erlang Loss Formula

$$PL = ((A**M)/(1 + A + (A**2)/(2!) \ldots\ldots\ldots\ldots (A**M)/(M!))$$

where $A = M*Y$: A is known as the Traffic Load.

This result is not only of great significance from a design point of view but is also most useful mathematically in simplifying the many computationally awkward formulae for multiple channel systems. It is tabulated for a representative range of values of A and M in Table 1.

Iable 1 Probability of Loss for Various Iraffic Loads and Channels

(a)

Traffic load	Number of channels			
A	*M* = 1	2	3	4
0 1	9	0.5	0	0
0 2	17	2	0 1	0
0 3	23	3	0 3	0
0 4	29	5	0 7	0
0 5	33	8	1	0.2
0.6	38	10	2	0.3
0 7	41	13	3	0.5
0 8	44	15	4	0.8
0 9	47	18	5	1
1 0	50	20	6	2
2 0	67	40	21	10
3 0	75	53	34	21
4.0	80	62	45	31

(b)

Traffic load *A*	Number of channels							
	M = 1	5	10	20	30	40	50	100
1	50	0 3	0	0	0	0	0	0
2	67	4	0	0	0	0	0	0
3	75	11	0	0	0	0	0	0
4	80	20	0.5	0	0	0	0	0
5	83	28	2	0	0	0	0	0
10	91	56	21	0 2	0	0	0	0
20	95	76	54	16	0.8	0	0	0
30	97	84	68	38	13	1	0	0
40	98	87	76	52	30	12	2	0
50	98	90	80	61	42	25	10	0
100	99	95	90	80	70	61	51	8

(b) Average Delav for all systems with random arrivals and exponential service or handling times

DBAR = PD/(M - A)

where DBAR is the average delay in the queue in mean service time units, SBAR;

PD is the probability of delay - a useful performance measure in its own right; it is related to PL, viz -

PD = M*PL/((M - A*(1 - PL))

(c) Ihe Average Delay for a single channel system with exponential arrivals and any service distribution - the Pollaczek-Khinchine Formula

WBAR = DBAR*(1 + C**2)/2

where WBAR is the mean delay in the queue for any service distribution

C is the coefficient of variation of that distribution; if the distribution is approximated by an Erlang K curve then C**2 = 1/K and WBAR may be written -

WBAR = DBAR*(1 + K)/(2*K)

IHE NEW RESULTS

Ihe above remarkable results allow a very wide range of queueing problems to be analysed. However there are three questions that are not satisfactorily answered in the literature which militate against the full universality of their application, viz -

(a) Does the Erlang Loss Formula hold for non-exponential service distributions?

(b) Is there a simple correction to the general delay formula that would hold for any service distribution when the number of channels is greater than one?

(c) Can the arbitrary service time distribution in the Pollaczek-Khinchine Formula be taken to include a deterministically programmed sequence?

These questions have been addressed in a heuristic manner and affirmative answers have been validated by extensive direct simulation.

Ihe Erlang Loss Formula

Io test Proposition (a) a direct simulation of the rejection probabilities of arrivals ata single channel facility with constant service was carried out and the results compared with the standard PL Formula which is generally considered rigorously valid for exponential service. The results are summarised in Iable 2.

Table 2   Probability of Loss Comparison

Traffic Intensity	Erlang Loss Formula	Simulation of CSQ
0.5	0.3333	0.3243
0.9	0.4707	0.4663

The General Delay Formula

Figure 3 shows the results of a simulation study on the multiple queue system with constant service and demonstrates a surprisingly simple result. It is that the effect of regularising the service is the same for many channels as for one. The importance of this is further accentuated when it is noted that the queueing delay is dominated overwhelmingly by the number of channels in the system - falling to a negligible amount when the number exceeds 10. In view of these results it is clear that the general formula given in (b) above would be closely approximated for all queueing situations with exponential arrivals and any Erlang service distribution by introducing the (1 + K)/(2*K) factor of the Pollaczek-Khinchine Formula, viz -

DBAR = (PD/(M-A))*((1+K)/(2*K))

Programmed Service Times

The question as to whweher a programmed sequence of service times constitutes an "arbitrary" distribution is a subtle one. It is neverthe less a challenging one particularly as it opens the way to a rational solution of the intersection delay problem. The many attempts to solve it range from the rationally based approximations of Webster, Miller, Newell, Blunden to the ingenious curve fitting effort presented in the new Highway Capacity Manual. Here we investigate the direct application of the Pollaczek-Khinchine result by regarding the sequence of short constant start-up headways during the green interval followed by a long departure service time equal to the red time to be an arbitrary service time distribution. On the basis that such a sequence is deterministic it seems reasonable to assume that its variance is zero i.e. K = infinity. This assumption has been tested for a representative range of service time sequences. The mean start-up headway or mean "service "time is readily calculated, viz -

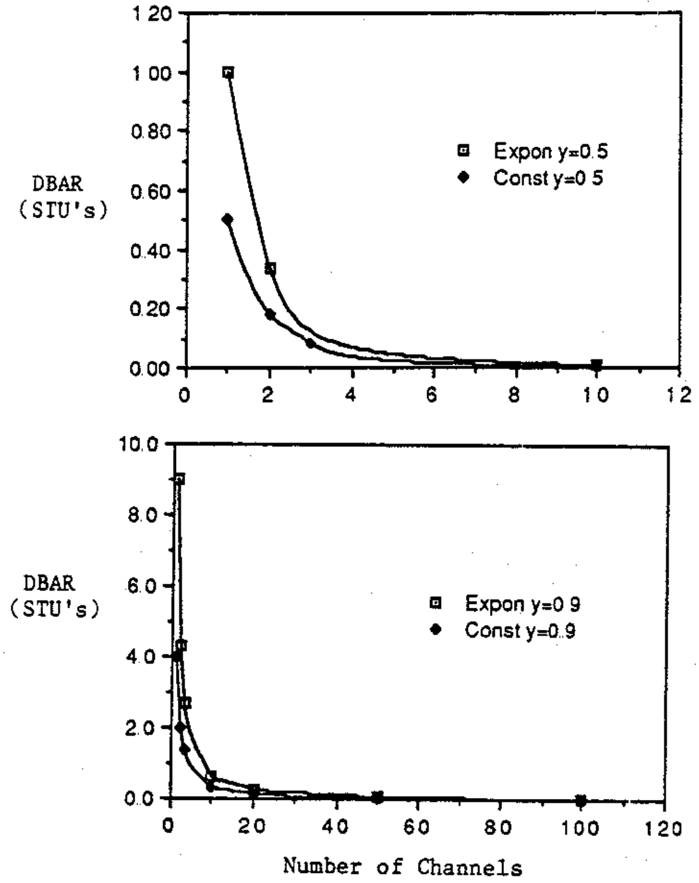

Fig. 3 MULTICHAN Simulation Results

SBAR = (INT(s*g*c/3600)*3600/s) +(1-g)*c)/(INI(s*g*c/3600) + 1)

where s is the start-up rate of the approach lane (vehicles/hour)

c is the cycle time (seconds)

g is the effective green time ratio.

The delay per vehicle, DBAR is then given by -

DBAR = TBAR *(1 + Y/(2*(1 - Y)) for values of Y < 1

This result together with some other well known intersection delay results and the new Highway Capacity Manual Formula are compared with those of an extensive simulation study in Table 3. Ihe simulation was carried out on an intersection approach with the following operational characteristics -

start-up capacity - 1500 vehicles/hour/lane

cycle time - 60 seconds

effective green time ratio 0.5

flow range - 75 to 675 vehicles/hour

simulation - 20 runs of an hours actual duration for each flow.

Iable 3 Intersection Simulation Results

Average Flow	Y	Sim'l'd Delay	Std. Dev	P-K Formula	HCM	Webster
75	0.1	9.86	2.36	4.77	6.00	7.28
150	0.2	10.28	2.39	5.09	6.34	8.11
225	0.3	10.47	2.32	5.49	6.74	8.89
300	0.4	11.28	2.48	6.01	7.22	9.86
375	0.5	12.05	2.67	6.78	7.83	11.21
450	0.6	13.30	2.92	7.92	8.64	12.91
525	0.7	14.63	3.23	9.87	9.80	15.39
600	0.8	17.85	4.76	13.57	11.75	19.91
675	0.9	24.84	9.15	24.88	15.97	31.51

The Overloaded Intersection

On this problem there is much confusion. It has been made clear earlier in this paper that overloading for an extended time period is virtually impossible. Short term overloads of some 10 - 15 minutes do seem common enough in urban traffic. Even so it is very difficult to decide whether these peaks are true overloads or in the context of say, a two hour busy period they are little more than random fluctuations anyway. However without debating this difficult issue in more detail here it is pertinent to note that if serious overloading does in fact take place deterministic queueing models provide a sound rationale for their analysis. For an intersection approach or any other facility operating at a Y value in excess of 1 for a period of time T seconds the average delay per vehicle may be written -

$$\text{DBAR} = (1 - g)*c/2 + (Y - 1)*T/2$$

This is a surprisingly simple result when one considers the the curve-fitting ingenuity used in obtaining the new Highway Capacity Manual Formula and that produced by the Australian Road Research Board for local application. Once again the validity of this result was tested by a simulation study and the results are summarised below in Table 4.

Table 4 Simulation Study of an Overloaded Intersection

Time (mins)	Overload Ratio	1.1	1.2	1.3
15	SIM'L'N	53.0	79.3	125.6
	DBAR	60.0	105.0	150.0
	HCM	64.5	121.5	198.2
30	SIM'L'N	92.5	169.4	250.1
	DBAR	105.0	195.0	285.0
	HCM	*	*	*
60	SIM'L'N	133.2	277.9	470.9
	DBAR	195.0	375.0	555.0
	HCM	*	*	*
120	SIM'L'N	245.2	561.9	875.7
	DBAR	375.0	735.0	1095.0
	HCM	*	*	*

* The Highway Capacity Manual Formula is calibrated for an assumed 15 minute overload period only.

If the simulation results can be considered to establist the delay benchmark then the DBAR formula estimates, even though high, are of the right order. Ihe HCM figures are significantly higher again for the 15 minute overload situation, As the formula does not include time explicitly the results for the longer periods have no real meaning, except that they are what the Formula yields. However the important conclusion that emerges is that despite the discrepancies the delay consequences of persistent overloading are dramatic enough to reinforce the earlier remarks concerning the stabilising influence of the queueing mechanism.

CONCLUDING REMARKS

Queueing Theory is an esoteric topic. Its importance in transport and land-use planning demands that it be reviewed and represented from time to time. In doing this it is hoped that the new results and the powerful role of computer simulation will aid its better understanding by the operator and analyst. Ihe "theory" tag is perhaps one reason for the diffidence of the practitioner. However it should be remembered that whilst tables, graphs, even simulation summarise data there is no substitute for rationally based formulae for they summarise knowledge and provide a framework for innovative application.

In particular the investigations reported-on here show that within the limits of accuracy of the simulations which are high the following results emerge -

(a) the Erlang Loss Formula is valid for any Erlang form of the service time distribution:

(b) the Average Delay for any number of channels and any Erlang service distribution can be closely approximated by an extension of the Pollaczek-Khinchine Formulae:

(c) a deterministic service time sequence can be very reasonably represented as a distribution with zero variance:

(d) for situations such as traffic signals subject to appreciable short-term input overloads a simple deterministic delay formula provides the most realistic analytical model for delay calculations.

REFERENCES

1. Blunden, W R and Black, J A (1984) Ihe Land-use Iransport System, 2nd Ed, Pergamon Press, Sydney.

2. Jones, J H and Blunden, W R (1968) Ship Iurn-around Iime at the Port of Bangkok, Journal of Waterways and Harbors Division, Proceedings of the American Society of Civil Engineers, vol 94 WW2, pp135-149.

3. Pretty, R L and Blunden, W R (1964) On the Computer Simulation of a Single Channel Queueing Facility for a Wide Range of Arrival and Departure Distributions, Proceedings of the Australian Road Research Board, vol 5, part 1, pp248-260.

4. Alfa, A S, Black, J A and Blunden, W R (1985) On the Iemporal Distribution of Peak Traffic Demaands: A model and its Calibration, Forum Papers, 10th Australian Transport Research Forum, vol 2 pp1-18.

5. Blunden, W R (1966) On the Iraffic Effects of Frontage Land-uses on Urban Main Roads and Arterials, Proceedings of the Australian Road Research Board, vol3, part 1, pp158-182.

6. Miller, A J (1968) Ihe Capacity of Signalised Intersections in Australia, Australian Road Research Board Bulletin No 3.

7. Brockmeyer, E, Halstrom, H L and Jensen, A (1960) Ihe Life and Works of A. K. Erlang, Applied Mathematics and Computing Machinery Series, No 6 (Copenhagen: Acta Polytechnica).

8. Lee, A M (1966) Applied Queueing Iheory (Iondon:Macmillan).

9. Webster, F V 1958) Settings for Fixed-cycle Iraffic Signals, Road Research Laboratory, Technical Paper No 39, Iondon.

10. Newell, G F (1960) Queues for Fixed-cycle Iraffic Lights, Annals of Mathematical Statistics, vol 31, no 3, pp589-604.