An Erlang formula for the Internet

Jim Roberts
james.roberts@inria.fr
14 February 2011
The failure of Poisson modeling?

• "Wide area traffic: the failure of Poisson modeling"
  - V. Paxson and S. Floyd, Sigcomm 1994
  - cited in 3898 papers since
• packet and TCP connection arrival processes are self-similar
  - no simple models! no simple formulas!
• but ... session arrivals in the busy period are Poisson

from Leland et al., IEEE ToN, 1994
Erlang and the Internet?
The Erlang formula
The Erlang loss formula

• assuming
  - Poisson call arrivals at rate \( \lambda \)
  - independent holding times of mean \( h \)
  - full accessibility to \( N \) trunks
  - lost calls cleared
• the probability of blocking is

\[
B = \frac{A^N / N!}{\sum_{i=0}^{N} A^i / i!}
\]

  - where \( A = \lambda \times h \) is the offered traffic
• an example of the capacity-demand-performance relation, essential for all engineering
Lessons from the Erlang formula

• performance is insensitive to the holding time distribution
• the essential traffic characteristic is $A (= \lambda \times h)$, the offered traffic or expected demand
• the network realizes scale economies
  - admissible load for given blocking increases with capacity

![Graph showing admissible load vs. capacity for different values of B]
The Erlang formula is not all we need...

- a simple utilization limit (e.g., $A/N < 80\%$) suffices for large $N$
- use the Engset formula for limited traffic sources
- use approximations for
  - limited availability
  - overflow and dynamic routing
  - delay systems
  - repeat attempts
  - etc...
- ... but it underpins our understanding of what is essential for network engineering and what traffic control objectives are reasonable and feasible
Useful generalizations
Poisson session arrivals

- sessions consist of a random number of calls interspersed by silent intervals
- assuming:
  - Poisson session arrivals
  - full accessibility to N trunks
  - blocked calls are cleared, the session proceeds with a new silence
- the probability of blocking is

\[
B = \frac{A^N / N!}{\sum_{i=0}^{N} A^i / i!}
\]
Poisson session arrivals

- Sessions consist of a random number of calls interspersed by silent intervals.
- Assuming:
  - Poisson session arrivals
  - Full accessibility to N trunks
  - Blocked calls are cleared, the session proceeds with a new silence.
- The probability of blocking is

\[
B = \frac{A^N / N!}{\sum_{i=0}^{N} A^i / i!}
\]
Poisson session arrivals

- sessions consist of a random number of calls interspersed by silent intervals
- assuming:
  - Poisson session arrivals
  - full accessibility to N trunks
  - blocked calls are cleared, the session proceeds with a new silence
- the probability of blocking is

\[
B = \frac{A^N / N!}{\sum_{i=0}^{N} A^i / i!}
\]

- for general, possibly correlated holding times and silences (cf. Bonald, 2006)
The multirate generalization

- the multirate loss system
  - connections with different constant bit rates \( \{c_i\} \)
  - Poisson traffics \( \{a_i\} \) where \( a_i = \lambda_i \times h_i \)
  - link of capacity \( C \)
- a product form steady state distribution \( \pi(n) \) and a recurrence relation for overall occupancy \( f(x) \) \( (f(x) = \sum_{\{n: nc=x\}} \pi(n)) \)
- insensitive to connection holding time distribution and true for Poisson session traffic

\[
\pi(n) = \prod_i (a_i^{n_i} / n_i!) \pi(0)
\]

\[
xf(x) = \sum_i a_i d_i f(x - c_i) \text{ for } x \leq C
\]
Internet traffic
Internet traffic: packets, flows, sessions

- A flow is a succession of packets, local in time and place, corresponding to an instance of some application.
- Packet and flow arrivals are self-similar, but session arrivals are Poisson!
- If flows were constant rate we could use the multirate Erlang recursion to size links for "negligible congestion"
Internet traffic: packets, flows, sessions

- A flow is a succession of packets, local in time and place, corresponding to an instance of some application.
- Packet and flow arrivals are self-similar, but session arrivals are Poisson!
- If flows were relatively low rate, we could use the multirate Erlang recursion to size links for "negligible congestion".
- But Internet flows use TCP and have an "elastic" rate.
Traffic theory for elastic traffic

• assuming
  - Poisson flow arrivals at rate $\lambda$
  - independent flow sizes of mean $\sigma$
  - perfect fair sharing of link rate $C$
• the probability $n$ flows are in progress is
  \[ \pi(n) = \rho^n (1 - \rho) \]
  - where $\rho = \frac{\lambda \times \sigma}{C}$
• the expected flow rate is
  \[ \gamma = C (1 - \rho) \]
• these results are true for a general flow size distribution
• they are also true for the general Poisson session traffic model
• but... most Internet flows cannot attain rate $C$
Accounting for limited flow rates

- assuming
  - Poisson session arrivals
  - flow rate \( \leq c \)
  - perfect fair sharing of link rate \( C \) when \( nc > C \)
- we readily derive state probabilities and mean flow rate
  - in particular, "probability of congestion" = \( \Pr[\text{flow rate} < c] \) is given by the Erlang delay formula

![Graphs showing loss and delay vs. capacity for different values of B and D.](image)
Accounting for a mix of limited flow rates

- assuming
  - Poisson session arrivals
  - class i flow rate \( \leq c_i \)
  - balanced fair sharing of link rate \( C \) when \( \Sigma n_i c_i > C \)
- we derive recurrence relations for relevant performance measures
  - extending the recurrence relations for the multirate Erlang model (cf. Bonald and Virtamo, 2005)
  - e.g., "probability of congestion" = \( \Pr[\text{class i flow rate} < c_i] = f^+/f^- + f^+ \)

where \( f^- = \Sigma_{n \leq C} f(x), \ f^+ = \Sigma_{n > C} f(x) \) and

\[
xf(x) = \sum_i a_i d_i f(x - c_i) \text{ for } x \leq C
\]
\[
Cf(x) = \sum_i a_i d_i f(x - c_i) \text{ for } x > C
\]
Discrepancies

• sharing is not balanced fair but max-min fair or proportional fair or...
  - but simulations show that performance does not depend critically on the fairness objective
• not all flows are elastic (e.g., streaming and conversational flows)
  - but high rate flows need to be adaptive ("TCP friendly"); assuming they are, it is conservative to suppose they are elastic (cf. Bonald & Proutière, 2004)
  - and the rate of low rate flows is hardly impacted by imposing fairness (especially max-min fairness)
• flow throughput depends on the network path
  - but $E[\text{throughput}] \approx \min_i \{c_i, C_i(1 - \rho_i)\}$ for links $l$ in flow path
Candidate Erlang formulas for the Internet

- the probability of congestion = probability network link imposes rate reduction
  - by the multirate recurrence or an approximation based on the Erlang delay formula
- or some other balanced fair formula (cf. Bonald, 2010)?
- expected throughput $\approx \min_i \{c_i, C(1 - \rho_i)\}$
- or $Pr[\text{max-min fair rate} > p] < \varepsilon$ for given flow peak rate $p$ and some tolerance $\varepsilon$
  - ensures all flows of rate $< p$ suffer negligible delay
- what matters is that we have capacity-demand-performance relations
  - as long as the network realizes some kind of per-flow fairness
Conclusions
Lessons from the Erlang formula for the Internet

• the **success** of Poisson modelling: performance depends essentially only on mean load and flow rate limits
• "fairness is good for you"!
  - impose (max-min) fairness in router queues
• performance deteriorates in overload
  - apply overload controls
• scale economies imply very limited scope for service differentiation
  - typically, all is good or all is bad!
Capacity-demand-performance

• understanding the capacity-demand-performance relation is paramount for successful engineering
• we have this relation for the Internet, assuming fair sharing between flows
  - the "Erlang formula(s) for the Internet"
• these relations demonstrate the limits and possibilities for meeting "service level agreements"
  - if flows can be reliably identified, sizing the network to meet performance targets is easy!

bit transport is a commodity
QoS is a scam!
References

• Bonald & Proutière, 2004
  - *On performance bounds for balanced fairness*
    Performance Evaluation 2004

• Bonald and Virtamo, 2005
  - *A recursive formula for multirate systems with elastic traffic*
    IEEE Communications Letters 2005

• Bonald, 2006
  - *The Erlang model with non-Poisson call arrivals*
    Proc. of SIGMETRICS / Performance 2006

• Bonald, 2010
  - *Habilitation à Diriger des Recherches*