How long will that movie take to download over the Internet?

A Cascade of Queues

TADEUSZ CZACHÓRSKI
Institute of Theoretical and Applied Informatics,
Polish Academy of Sciences
tadek@iitis.gliwice.pl
Prof. Tadeusz Czachórski, director of the PAS Institute of Theoretical and Applied Informatics, works on modeling and performance evaluation of computer systems and networks, and studying next-generation Internet issues.

Vast volumes of data are constantly flowing across telecommunications networks, including the Internet. Every time a user connects to a webpage, a bitstream flows across the network to his or her computer

Text, images, music, films, telephone calls - all this content is transmitted digitally. Networks are also used for distributed calculations that require communication between computers. All such transmitted bits are organized into larger units - packets of data - containing their destination address. Their transport is supervised by communications protocols striving to detect potential transmission biases and find a better route between the sender and the recipient, usually via intermediate nodes.

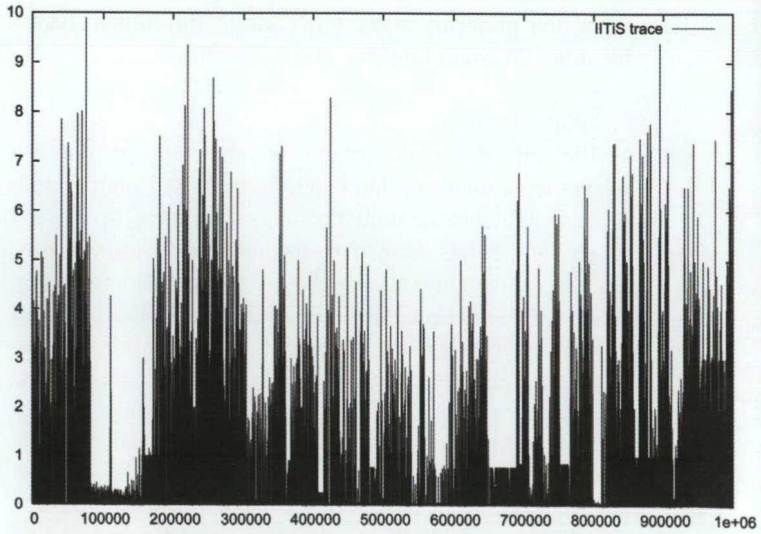
Fast and reliable

The speed of network operation means that there are many bits and packets en route between different nodes. Time of transmission between a given pair of nodes can be calculated easily when the distance between them and the speed of transmission along the connection are known; however, the waiting time at each is unknown.

The movement of packets across a network can be compared to road traffic: streams of vehicles traveling along routes of varying capacity. The flow is highly irregular; the intensity of traffic across a typical connection varies with time. Since each node can be reached by many routes in the network, and acts as a switch directing the arriving packets to their output paths, the packet stream in a given direction frequently exceeds a node's output capacity. In this event, incoming packets get queued up: they are saved in buffers and sent according to network availability. This may result in backlogs, which again have an analog in road traffic: packets/vehicles get held up in such a queue before they can start moving again. Estimating the waiting time is extremely important, since users need the transmission time to be as short and repeatable as possible.

Reliability of transmission is equally important; when a packet queue buffer is full, further incoming packets will not be saved. The Internet transmission control protocol (TCP) tries to deal with this to ensure the reliability of transmission: the recipient confirms when packets are received, and if such confirmation is not received, the packet is sent out again. However, this in itself may introduce further delays, and during "live" transmissions it is not always needed; packets which arrive late have already caused a certain irreparable disruption to the transmitted sounds or images, and are likely no longer useful. As such, it is important to maintain network conditions that minimize the likelihood of packet loss.

Network nodes classify packets by higher and lower priority, and prioritize sending them accordingly to provide a higher quality of service. It is also possible to reserve a certain capacity along the entire route between the sender and the receiver, although this is not a particularly effective solution across the entire network. Network operators must seek a compromise: on one hand, it is important that a network be used to the best of its capacity; on the other, the more a network is being used, the more the quality of service may drop (with node



Changes in intensity of packet movement over time. Measurements taken at the PAS Institute of Theoretical and Applied Informatics; horizontal axis: packet number; vertical axis: time interval in seconds between packets

queues potentially building up, increasing the likelihood of overload), resulting in poor client satisfaction.

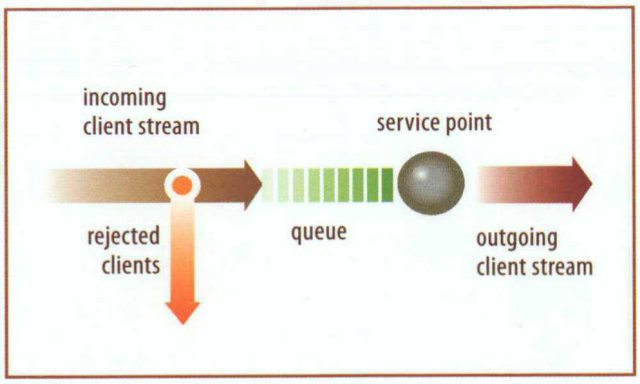
Queue modeling

Network design and development can be assisted by modeling techniques. Queuing theory has proven useful for describing traffic intensity across computer networks. The basic model examined here is as follows: clients approach a service point at time intervals which are a known random variable (described by a certain probability distribution) and are queued up waiting to be served; service time is also random. It is necessary to define the distribution of queue length, that is the probability that the queue is empty or that it contains 1, 2, or n clients, the probability distribution for waiting time, and the probability of a queue with a finite capacity becoming overfilled. Service points can be interconnected to make a network by defining client routes between them.

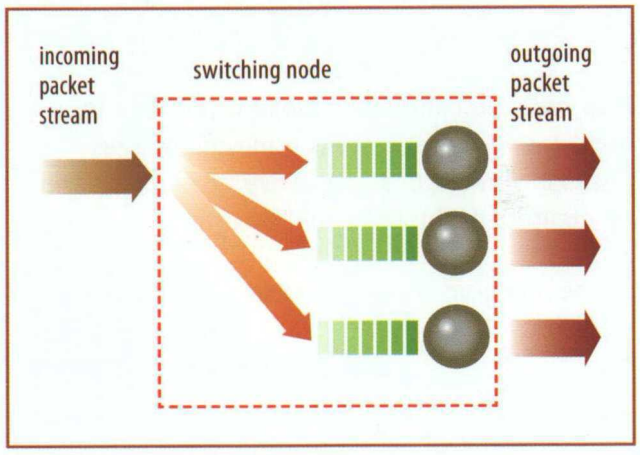
Queuing models have been constructed for various applications. Apart from the obvious – where clients are people and service points are real-life locations such as post office counters or supermarket checkouts – there are many other interpretations; the "clients" might be ships docking at ports (here, "service time" is the unloading time), parcels sent to warehouses (storage time), or vehicles arriving at junctions (the time taken to cross). In telecommunications, queuing models have been used for around a century, since Erlang and Molina first used them to describe the operation of telephone exchanges: such an exchange is able to serve a given number of connections, where clients are the people making the connections, and service time is the duration of each conversation. In this instance, it is necessary to define the probability that all channels will be busy and a given client might not be served.

Today, queuing models in communications require increasing computational resources; we are interested in quantitative calculation results referring to specific networks and their parameters. The community working on issues of Internet traffic is huge, including organizations measuring traffic flow as well as national and internationals institutions and programs aiming to improve the principles of transmission in communication protocols. The Polish Future Internet Engineering project, for instance, encompasses nine universities, PAS institutes, and other scientific institutions.

The PAS Institute of Theoretical and Applied Informatics has in fact been studying queuing models for computer systems and networks for the last three decades. The most frequently used methods are Markov chains, diffusion approximation, and continuous approximation. A Markov chain is a random process, simple in mathematical terms; however, the problem here is the number of unknown variables and the equations to be solved. Diffusion approximation, in turn, uses analogies



Model of a network service point and queue



Model of a packet switching node (router)

between the diffusion process and queue length: differential equations describing the position of a particle in diffuse motion are here used to describe the probability distribution of queue length. Continuous approximation is a simplified version of this method – it only uses mean values of packet stream flow, its size and the resulting queue length – and as such it cannot be used to define probability distributions. However, the differential equations it uses are simpler, and the calculations can be completed in a reasonable time. Apart from these analytical models, use is also made of computer simulations of protocol operation and queue behavior.

These techniques and others are making today's networks not only faster but also increasingly reliable – so that ultimately we can all spend less time wondering just how long it will take that film to download. ■

Further reading:

Czachórski T. (1999). *Kolejkowe modele w ocenie efektywności sieci i systemów komputerowych.* [Queuing Models for Performance Evaluation of Computer Systems and Computer Networks]. Gliwice: Pracownia Komputerowa Jacka Skalmierskiego.

Hassan M., Jain R. (2004). *High Performance TCP/IP Networking: Concepts, Issues, and Solutions.* Prentice-Hall.

Kobayashi H., Mark B.L. (2009). *System Modeling and Analysis: Foundations of System Performance Evaluation.* Prentice Hall.