RESOURCE PROVISIONING AND NETWORK TRAFFIC

Network engineering:

• Feedback traffic control
  → closed-loop control ("adaptive")
  → small time scale: msec
  → mainly by end systems
  → e.g., congestion control

• Resource provisioning
  → open-loop control ("in advance")
  → large time scale: seconds, minutes, and higher
  → mainly by service providers

Question: what do ISPs do to keep customers happy and make money (or lose less money)?
Resource provisioning: two main resources

- Bandwidth allocation
  → primary

- Buffer allocation
  → resource dimensioning: long-term
  → also called network planning: months, years
  → on-demand resource allocation: short-term
  → i.e., second, minutes, hours

Turns out:
  → same principles apply to both
  → take ISP’s viewpoint
  → granularity: user-, session-, and packet-level
User- and Session-Level Resource Provisioning

Basic set-up:

- aggregate demand at access switch
- \( n \) users or CPE (customer premises equipment)

Set-up applies to:
- Telephone switch: TDM slot per session/user
- Dial-up modem pool: e.g., AOL Internet access
- Broadband access service: e.g., IP address pool
Basic building block: access switch

→ function: aggregation

→ performance benefit?

→ old banking trick: keep fraction of total deposit

→ observation: not all customers withdraw at once (?)

Networking: not all customers access network at once

→ affords efficiency & economy

• can keep fewer T1 lines

• can keep smaller modem pool

• can keep fewer IP addresses

• can keep less bandwidth

Note: a calculated risk

→ sometimes very many users connect at once

→ access denied: blocking
In what other major sector is “old banking trick” employed?

What makes old banking trick possible?

\[ \rightarrow \quad \text{one of the few “laws of engineering”} \]

Law of large numbers (LLN): the sum of many independent random variables concentrates around the mean

\[ \rightarrow \quad \text{i.e., very few outliers} \]

\[ \rightarrow \quad \text{also, typically, mean} \ll \text{maximum} \]

Ex.: Suppose there are \( n \) users subscribing to Verizon in West Lafayette.

\[ \rightarrow \quad \text{how many users will make a call at time} \ t? \]
Assuming:

- $X_i(t) = 0$ if no call by user $i$ at $t$, 1 if call
- $\Pr\{X_i(t) = 1\} = p$
- users make calling decisions independent of each other
  → i.e., $X_1(t), X_2(t), \ldots, X_n(t)$ are i.i.d.
  → note: same as coin tossing
- total calls at time $t$
  → $S_n(t) = X_1(t) + X_2(t) + \cdots + X_n(t)$
- average number of calls
  → $E[S_n(t)] = E[X_1(t)] + \cdots + E[X_n(t)] = np$
  → hence, $E[S_n(t)/n] = p$
  → independence needed?
- LLN: $\Pr\{|\frac{S_n(t)}{n} - p| > \varepsilon\} \to 0$ as $n \to \infty$ for any $\varepsilon > 0$
  → weak LLN
  → strong LLN?
Thus, for sufficiently large $n$ deviation from the mean is rare.

\[ \rightarrow \text{Verizon can expect } np \text{ calls at time } t \]
\[ \rightarrow \text{with large } n, \text{ very close to } np \text{ calls} \]
\[ \rightarrow \text{but, how large is “large”?} \]
\[ \rightarrow \text{does WL have sufficiently many customers?} \]

To be useful for engineering, we need to know more

\[ \rightarrow \text{rate of convergence} \]

Large deviation bound:

\[ \rightarrow \Pr\{|\frac{S_n(t)}{n} - p| > \varepsilon\} < e^{-an} \]
\[ \rightarrow \text{constant } a \text{ depends on } \varepsilon \]
\[ \rightarrow \text{exponential decrease in } n \]
\[ \rightarrow \text{also, holds for all } n \]
\[ \rightarrow \text{engineering: blocking probability} \]
Large deviation bound gives simple prescription for resource provisioning:

- measure $p$ (historical data); ISP knows $n$
- determine acceptable blocking probability $\delta$
  - e.g., $\delta = 0.00001$
  - i.e., one in 10000 calls gets blocked
- find $\varepsilon$ such that
  \[
  \Pr\{|\frac{S_n(t)}{n} - p| > \varepsilon\} = \Pr\{|S_n(t) - np| > n\varepsilon\} < e^{-an} = \delta
  \]
  - note: $\varepsilon$ determines excess capacity allocated
  - recall: $a$ depends on $\varepsilon$ (called rate function)
  - one of the main tools used by ISPs/telcos

From ISP’s perspective, is this enough for making resource provisioning decision?

---- what crucial element may be missing?
Connection lifetime or duration

→ also called call holding time
→ for how long a resource (e.g., modem) is busy
→ if fixed, then previous formula holds
→ user session property
→ in general: connection lifetime is variable
→ e.g., average telephone call: 7 minutes

Let \( L \) denote connection lifetime (assuming i.i.d. across all users)

→ by measurement, ISP knows its distribution
→ consider average lifetime \( E[L] \)
→ consider two time instances \( t \) and \( t + E[L] \)
→ what to do?
View system in terms of time granularity $E[L]$:

- use large deviation formula to estimate connection arrivals during time window $[t, t + E[L])$  
  \[ \rightarrow \text{excess capacity } n\varepsilon \text{ above and beyond mean } np \]

- use distribution of $L$ to estimate $\Pr\{L > E[L]\}$  
  \[ \rightarrow \text{may refine } \varepsilon \text{ to } \varepsilon' \ (\varepsilon < \varepsilon') \]
  \[ \rightarrow \text{for } E[L] \text{ not-too-small may not be needed (why?)} \]
Remarks:

- LLN: principal engineering tool used by large transit providers and large access providers
  → “largeness” is key
  → even though components are random, system is well-behaved and predictable
  → apply at ingress/egress and backbone links
  → measurement-based tool: traffic matrix
• sometimes can apply central limit theorem (CLT): aggregate has Gaussian (normal) distribution
  \( \rightarrow \) in practice: not very useful
  \( \rightarrow \) e.g., tail of Gaussian: not very accurate
  \( \rightarrow \) deviation estimate valid only for moderate \( \varepsilon \)
  \( \rightarrow \) may not even look Gaussian!
  \( \rightarrow \) needs very large \( n \)
  \( \rightarrow \) large deviation bound: holds for all \( n \)

• aggregation over time window \( [t, t + E[L]] \)
  \( \rightarrow \) a single user can have 2 or more sessions
  \( \rightarrow \) may violate independence assumption (across users)
  \( \rightarrow \) independence over time: separate matter
we assumed discrete number of resources

→ e.g., 10000 modems, 50000 IP addresses, 1000 T1 lines, etc.

→ valid viewpoint at user/session granularity

→ also applies to packet granularity

→ as long as independence over time holds

How does session arrival for a single user over time look like?

→ aggregation over time

→ resource provisioning: buffering

→ vs. aggregation over users (bandwidth)
Session arrivals over time:

- apply coin tossing idea over time
- before: one coin per user
- now: one user has multiple coins
  → coins are assumed to be i.i.d. with probability $p$
  → apply LLN over time!
Over discrete time window $[1, m]$, same bounds apply; for user $i$:

$$
\rightarrow S_i(m) = X_i(1) + X_i(2) + \cdots + X_i(m)
$$

$$
\rightarrow Pr\{|\frac{S_i(m)}{m} - p| > \varepsilon\} < e^{-am}
$$

Thus: if we have $mp + m\varepsilon$ resources (e.g., buffer), then can buffer user $i$’s service requests (could be even packets) over time $[1, m]$ without “loss”

$$
\rightarrow \text{loss probability} < e^{-am}
$$

$$
\rightarrow \text{before: blocking probability}
$$

$$
\rightarrow \text{in practice: } m \text{ can’t be too high}
$$

$$
\rightarrow \text{buffering } \Rightarrow \text{ delay penalty}
$$

$$
\rightarrow \text{some applications require quick response time}
$$
One refinement: what does the time spacing between successive arrivals look like?

\[ \rightarrow \text{prob. session will arrive after } k \text{ steps: } (1 - p)^k p \]

\[ \rightarrow \text{called geometric distribution (where did we see it?)} \]

\[ \rightarrow \text{most important: exponentially decreasing in } k \]

Corresponding distribution in continuous time:

\[ \rightarrow be^{-bt} (t \text{ in place of } k) \]

\[ \rightarrow \text{exponential distribution} \]

\[ \rightarrow \text{essentially equivalent to geometric distribution} \]

\[ \rightarrow \text{important property: memoryless} \]