This document describes a user-based performance metric called Equivalent Circuit Rate that offers some advantages in evaluating MBWA systems. We present an analytical model, simulation results, and issues involved in using this for MBWA systems.
Equivalent Circuit Rate: A User-based Performance Metric for Shared Packet Access Networks

IEEE 802.20 Standards Meeting
July 20-25, 2003

N. K. Shankaranarayanan
shankar@research.att.com
AT&T Labs-Research
Middletown, NJ

Vincent Park
park@flarion.com
Flarion Technologies
Bedminster, NJ
Talk Outline

• ECR (Equivalent Circuit Rate) measure
• Analytical model for interactive traffic
• Shared-TDMA downlink simulation
• ECR used for MBWA
• We want to characterize performance & capacity
  • Quantify user experience
  • Determine number of users
• We will take a user-centric view
Shared v. Dedicated Access

Assume backbone & server are ideal except for finite latency / propagation delay

User

Shared Packet Access

Gateway Router

Internet Backbone

Interactive Data (Web) Server

User

User

Dedicated Access (circuit wireless, dial-up modem..)

User

User

User
Equivalent Circuit Rate (ECR) of a shared packet access network:

- Dedicated rate required to get same user performance
- Useful and intuitive measure for users, planners, marketing
- Equivalence based on statistics of application (e.g. mean or percentile of web page delay)
ECR References

   Full paper: complete analytical model and wireless simulations

   Mixed web & dataphone traffic workloads

   Cable modem scenario
ECR Model: Clarifications

• Shared v. dedicated equivalence - very general
• Equivalence can be based on:
  – Mean delay
  – 90th percentile delay
  – etc

• Analytical model
  – Interactive traffic model
  – Mean delay

• Simulation work (shared TDMA wireless system):
  – Downlink simulation
  – Web browsing traffic model
  – Mean page delay, 90th percentile page delay, etc.
Talk Outline

- ECR (Equivalent Circuit Rate) measure
- Analytical model for interactive traffic
- Shared-TDMA downlink simulation
- ECR used for MBWA
Analytical Model for Interactive Traffic

- Closed queueing network with finite population of ON/OFF sources
- OFF state independent of ON & Delay state => interactive traffic model
- Model each page (response) as one job for the server
- Per-user workload decreases as number of users increases

Closed Queueing Model

Interactive user workload model

\[ Delay = \text{Wait Time} + \text{Service Time} + \text{Latency} \]

- Average OFF (think) time = \( \frac{1}{\lambda} \)
- Average service time = \( \frac{1}{\mu} \)
- Average latency = \( E[D] \)

Per-user offered load

\[ a = \frac{\frac{1}{\mu}}{\frac{1}{\lambda} + \frac{1}{\mu} + E[D]} \approx \frac{\text{average user rate}}{\text{shared network rate}} \]
Page delay v. users

- Average delay - normalized to minimum (single-user case)
- Saturation phenomenon: Linear rise in delay beyond $M^* = 1/a$
- Network sizing $\sim 0.7M^* = 0.7/a$ for wide range of $a$

**Diagram: Page delay v. users**

- Normalized average delay vs. number of users, $M$

  - Slope = 1
  - $a = 0.03$
  - $a = 0.01$
  - $M^* = 100$
  - $M^* = 33$

**Diagram: Page delay v. $M/M^*$**

- Normalized average delay vs. degree of saturation, $M/M^*$

  - $a = 0.001$
  - $a = 0.003$
  - $a = 0.01$
  - $a = 0.03$

  - $aM = 1$ (saturation)

July 2003 - 802.20 Meeting
Equivalent Circuit Rate (ECR) v. users

- Assume network rate of 4.0 Mbps shared by M users
- ECR = rate required for same average page delay
- ECR $\geq$ 4.0/M Mbps $\Rightarrow$ multiplexing gain

ECR v. users

Normalized ECR v. M/M*

$M = M^* = 1/a$ (saturation)
Normalized ECR v. Link utilization

- Normalized ECR ~ \(1 - \text{Utilization}\) for wide range of \(a\)!
- ECR ~ \(\text{Shared Rate} \times (1 - \text{Utilization})\)
- Universal relationship – quite general
- Intuitive on hindsight

July 2003 - 802.20 Meeting
Talk Outline

- ECR (Equivalent Circuit Rate) measure
- Analytical model for interactive traffic
- Shared-TDMA downlink simulation
- ECR used for MBWA
Simulation Model

- Web workload
- FreeBSD TCP code

**TCP/IP packets**

**User queues**

**Mobile Stations**

- 1
- 2
- N

**BASE STATION**

**Link Level**
8 slots in 20 ms frame

**Shared channel rate:** 176 to 440 kbps
( constant rate per simulation )

**Block error rate:** 0 to 40%

**Backbone network**
round-trip latency:
2 ms to 200 ms

July 2003 - 802.20 Meeting
Page delay: Analysis and Simulation

- Different network rates
- 200 ms latency

- Different latencies
- 440 kbps network rate
ECR: Mean & percentile page delay

- ECR result is same when calculated using two methods to capture user experience: average & 90th percentile page delay
- Good match with analytical results (ECR based on average page delay)

- 200 ms latency
- No errors
- 9-13 kbps user rate

Network can be sized based on ECR target

July 2003 - 802.20 Meeting
ECR and other end-to-end rate measures

- ECR uses end-to-end delay to characterize only the access portion
- ECR reflects effective access network bandwidth available to user
- ECR insensitive to delay, but other measures are not
- Differences decrease at high loads
Different radio channel conditions

- 176 to 440 kbps with different error rates
- Includes link-level retransmission
- Linear ECR v. utilization curves
Markov channel model
- Time-varying
- Degradation with load (interference)
- ECR decrease with load
  - channel degradation
  - sharing
ECR for different user modem rate limits

![Graph showing ECR for different user modem rate limits]

- 55 kbps/slot, 8-slot terminal
- 55 kbps/slot, 3-slot terminal
- 55 kbps/slot, 1-slot terminal

July 2003 - 802.20 Meeting
Mixed channel conditions

- Users in Group 1:
  - Farther from base
  - Low rate
- Users in Group 2:
  - Closer to base
  - High rate

Uniform Case Results:
Only low rate:
\[ a_1 = 0.067, \quad M_1^* = \frac{1}{a_1} = 15 \]
Only high rate:
\[ a_2 = 0.028, \quad M_2^* = \frac{1}{a_2} = 35 \]

Non-uniform Case:
???
Delay v. utilization for mix of channel conditions: insensitive to fraction of mix
Number of users for mixed channel rates

Simulation & theory for various fractions
(3-slot terminals, round-robin sharing)

Theory:

\[
M_{\text{tot}} = \frac{1}{E(a)} = \sum_{k} \frac{1}{p_k a_k}
\]

\[
M_k = p_k M_{\text{tot}}
\]
4 groups: hi/lo channel rates & hi/lo modem rate limits

4 equal groups:
55 kbps, 3-slot
55 kbps, 1-slot
22 kbps, 3-slot
22 kbps, 1-slot

July 2003 - 802.20 Meeting
Mix of traffic loads: Web and dataphone (WAP-like) users

- **Wireless channel:** Time-varying TDMA channel with mean rate of 220 kbps
- **Web user:** average rate ~ 9 kb/s
- **Dataphone:** average rate ~ 0.65 kbps
- Small size of dataphone (WAP-like) payload resulted in inefficient use of TDMA slots
Talk Outline

- ECR (Equivalent Circuit Rate) measure
- Analytical model for interactive traffic
- Shared-TDMA downlink simulation
- ECR used for MBWA
Using ECR for MBWA Dimensioning: E.g., Number of Active Users

• Based on modeling Web traffic (dominant)
• Web traffic model details…
  – Mean “Web-request” size: 53 Kbytes
  – Mean “Viewing” time: 40 sec
• Additional system model details…
  – Shared Link Rate: 1.5 Mbps
  – E[D]: 450 ms (accounts for RTT and TCP slow start)
• Corresponding User Workload: ~10.8 kbps
Saturation Point suggests number of active users that should be concurrently supported

![Graph showing Equivalent Circuit Rate and Mean Response Time](image)

July 2003 - 802.20 Meeting
Effect of changing shared link rate

Saturation Point ($M^*$) and Recommended Maximum Number of Users

Number of Users

Shared Link Rate (Mbps)

$M^*$

Max users
Improving Fidelity of Model for MBWA Dimensioning

• Selecting the Shared Link Rate
  – Function of user distribution in cell/sector?
  – Function of interference?
  – System might not offer full aggregate rate to one user
• Modeling mix of individual user rate limits
  – Due to channel conditions
  – Due to administrative policy, e.g., policing
• Modeling a mix of users with different workloads
  – Need traffic models for other interactive applications (e.g., instant messaging?)
  – Need to defined mix of user traffic models
• Inclusion of non-interactive traffic models
  – Aggregate rate of CBR traffic flows (e.g., VoIP) can be subtracted from Shared Link Rate prior to modeling/analysis
ECR in MBWA Simulations

- Separate computation of ECR reference (e.g.
  - Pre-compute equivalence metric (e.g. mean delay) v. circuit rate for a single user with dedicated circuit connection for same conditions of {traffic model, protocol model, backbone network etc.} - this is independent of MBWA system
  - Map MBWA user performance to ECR using lookup
- Runtime computation of Effective Serving Rate ~ ECR
  - \[ \text{ECR} = \text{ESR} \sim \frac{\sum \text{user bytes transferred}}{\sum \text{“serving time”}} \]
  - “serving time” = time when user queue is not empty
- ECR may be useful as comparison tool
- Easy to track ECR - useful as input for research, e.g. analytical model can be extended
Value of using ECR for MBWA

• Technical
  – Forces one to think about, and quantify various rates
    • aggregate rate, shared rate, user rate
  – ECR v. utilization relationship may reveal “capacity” behavior under load
  – Insights into system behavior

• Planning, Operations, and Marketing
  – Simple model, simple rules of thumb
  – End users & planners can easily relate to ECR
  – ECR highlights multiplexing advantage of shared systems for bursty traffic
Conclusion

- Useful and robust new measure of user-perceived performance: ECR (Equivalent Circuit Rate)
- Simple analytical model for interactive data in a shared packet network
- ECR & network sizing depend on simple parameters: (shared channel rate, average user rate)
- Results validated by TCP-based simulations of typical Web workloads in shared TDMA wireless system
Backup Slides
Focus on Web-based applications since they dominate Internet usage

Users alternate between ON (busy) state and OFF (idle) state

**ON state duration (Web page delay)** depends on:
- Page size statistics - no/size of files in page - typically heavy-tailed
- Server and network resources
- HTTP and TCP protocol dynamics

**OFF state duration (Think time)**
- Determined by human behavior - typically heavy-tailed
- Independent of ON state \(\Rightarrow\) feedback, upper limit on workload

\[\text{TCP connection establishment & HTTP request for web file(s)}\]
\[\text{Addl. HTTP requests to same server (same TCP connection in HTTP v1.1)}\]
Web Workload Model

HTTP v1.0

Embedded files per page

ON ON ON ON
OFF OFF OFF

time

= HTTP request and TCP connection establishment

HTTP v1.1

ON ON ON ON
OFF OFF OFF

time

= HTTP request and TCP connection establishment followed by period during which the web page is received by the user

July 2003 - 802.20 Meeting
Web Browsing Workload Model

OFF (think) time statistics*
- Pareto distribution
- 4 sec median, 12 sec mean
- 2 sec min, 10 mins. max

Web page statistics*
- Lognormal distribution
- 10 KB median, 20 KB mean
- 100 B min, 100 KB max

Lower channel bit rate OR more users...
- Longer ON times (Web page delay)
- Same Web pages, Same OFF (Think) times

* Web traffic statistics based on Boston University SURGE model (1998)

July 2003 - 802.20 Meeting
We can assume exponential distributions..

- Formulas can be easily derived by assuming **exponential distributions** for **service time and think time**.
- For processor-sharing (round-robin) service policy, the formulas involving mean statistics are also valid:
  - for any **rational service time** and **think time distributions**
  - for any **general service time distribution** and exponential think time distribution
Table I. Simulation parameters for network and traffic model.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Access network bit rate</td>
<td>176 to 440 kbps</td>
</tr>
<tr>
<td>Round-trip backbone propagation delay</td>
<td>2 to 400 ms</td>
</tr>
<tr>
<td>Queueing policy</td>
<td>Round robin</td>
</tr>
<tr>
<td>TCP window size</td>
<td>64 KB</td>
</tr>
<tr>
<td>TCP MSS</td>
<td>512 bytes</td>
</tr>
<tr>
<td>Web page size distribution</td>
<td>Log-normal ($\mu = 9.5$, $\sigma = 1.8$) in bytes</td>
</tr>
<tr>
<td></td>
<td>$f(x) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}$</td>
</tr>
<tr>
<td></td>
<td>100 B min, 100 kB max</td>
</tr>
<tr>
<td>Median/Mean Web page size</td>
<td>10 kB / 20 kB</td>
</tr>
<tr>
<td>OFF time distribution</td>
<td>Pareto ($k = 2$ s, $\alpha = 2$)</td>
</tr>
<tr>
<td></td>
<td>$f(x) = \alpha k^\alpha / x^{\alpha+1} \ , x \geq k$</td>
</tr>
<tr>
<td></td>
<td>2 s min, 10 min max</td>
</tr>
<tr>
<td>Median/ Mean OFF time</td>
<td>4 s / 12 s</td>
</tr>
<tr>
<td>Workload for single user</td>
<td>9 to 13 kbps</td>
</tr>
</tbody>
</table>
Web page delay behavior

- 440 kbps rate
- No errors
- 200 ms latency
- 9 to 13 kbps user rate