ETLA ELINKEINOELÄMÄN TUTKIMUSLAITOS
THE RESEARCH INSTITUTE OF THE FINNISH ECONOMY

Keskusteluaiheita
Discussion Papers
21.3.2012

No 1268

Queueing Systems with Fractional Number of Servers

Valeriy Naumov* – Olli Martikainen**

* ETLA – The Research Institute of the Finnish Economy, valeriy.naumov@etla.fi
** ETLA – The Research Institute of the Finnish Economy, olli.martikainen@etla.fi

ISSN 0781–6847

Contents

Abstract

In this paper we introduce multi-server queueing systems that can be considered as extensions of conventional M/M/*s* queue to fractional number of servers. We show that the extended Erlang's delay function can be used to calculate delay probabilities for such systems. This approach enables the delay analysis in networks with fractional number of servers in the nodes using classical methods.

Key words: Erlang's delay function, multi-server system, state dependent service times

JEL: C61, C62, C68

Tiivistelmä

Tässä artikkelissa esittelemme monen palvelimen jonojärjestelmän, jossa traditionaalinen M/M/s jonokäsite laajennetaan murtolukumäärälle palvelimia. Näytämme, että yleistettyä Erlangin viivefunktiota voidaan käyttää tällaisen järjestelmän viivetodennäköisyyksien laskemiseen. Esitetty menetelmä mahdollistaa viiveiden laskennan klassisia menetelmiä käyttäen yleistetyissä jonojärjestelmissä, joiden solmuissa on murtolukumäärä palvelimia.

1 Introduction

We introduce multi-server queueing systems that can be considered as extensions of conventional M/M/*s* queue to fractional number of servers. We show that the extended Erlang's delay function can be used to calculate delay probabilities for such systems. This approach enables the delay analysis in networks with fractional number of servers in the nodes using classical methods.

Erlang's loss function and Erlang's delay function have been widely used for the analysis of multi-server systems since their appearance in (Erlang, 1917). These functions give the blocking probability for systems without waiting places and delay probability for systems without losses. Analytical continuation extends these functions to non-integral value of the number of servers (Bretschneider, 1973), (Jagers and van Doorn, 1991). Extended Erlang's loss function is used for the analysis of complex systems with losses by teletraffic engineering methods like equivalent random traffic and Hayward's approximation (Iversen, 2011). Equivalent random traffic for queues method as well as optimal server allocation requires calculation with extended Erlang's delay function (Nightingale, 1976), (Down and Karakostas, 2008), (Naumov and Martikainen, 2011). But, to the author's knowledge, there are no descriptions of systems which can be interpreted as multi-server system with fractional number of servers. In this paper we introduce a queueing system, in which the delay probability is given by extended Erlang's delay function, and which for integer value of parameter *s* operates as conventional multi-server queueing system.

2 M/M/*s* queue

Consider queueing system with *s* servers and infinite number of waiting places. Assume that arrival process is Poisson with arrival rate λ, service times are exponentially distributed with service rate μ and $\rho = \lambda/\mu < s$. Then the mean residence time in the system is given by the following well known formula

$$T = \frac{1}{\mu} + \frac{C_s(\rho)}{s\mu - \lambda}. \tag{1}$$

Here Erlang's delay function $C_s(\rho)$ gives the probability that a customer has to wait in the queue:

$$C_s(\rho) = \frac{\frac{\rho^s}{s!}\left(\frac{s}{s-\rho}\right)}{\sum_{k=0}^{s-1}\frac{\rho^k}{k!} + \frac{\rho^s}{s!}\left(\frac{s}{s-\rho}\right)}. \quad (2)$$

Erlang's delay function $C_s(\rho)$ can be expressed through Erlang's loss function $B_S(\rho)$ as

$$C_s(\rho) = \frac{sB_s(\rho)}{s-\rho(1-B_s(\rho))}, \quad (3)$$

where

$$B_S(\rho) = \frac{\rho^s}{s!}\left(\sum_{k=0}^{s}\frac{\rho^k}{k!}\right)^{-1}. \quad (4)$$

For traffic engineering purposes, Bretschneider has proposed in (Bretschneider, 1973) an analytic continuation of Erlang's loss function to non-integral values of *s* by

$$B_s(\rho) = \frac{\rho^s e^{-\rho}}{\int_\rho^\infty t^s e^{-t}dt}, \quad (5)$$

which can be rewritten as (Jagerman, 1974):

$$B_s(\rho) = \left(\rho\int_0^\infty e^{-\rho t}(1+t)^s\,dt\right)^{-1}. \quad (6)$$

For all integer *s* this function coincides with (4). Similar analytic continuations of Erlang's delay function, which for all integer *s* coincides with (2), have been proposed in (Jagers and van Doorn, 1991) by

$$C_s(\rho) = \left(\rho\int_0^\infty e^{-\rho t}(1+t)^{s-1}tdt\right)^{-1}. \quad (7)$$

It is easy to prove that formula (3) also remains valid for functions (6) and (7) (see, e.g., Karsten et al. 2011).

For calculation of extended Erlang's delay function $C_s(\rho)$ we use formula (3) in which extended Erlang's loss function $B_s(\rho)$ is computed in Matlab environment (Martinez and Martinez, 2002). We rewrite (5) as

$$B_s(\rho) = \frac{f_{s+1}(\rho)}{1-F_{s+1}(\rho)}, \quad (8)$$

where $f_\gamma(x)$ and $F_\gamma(x)$ are the probability density function and the cumulative distribution function of the standard gamma distribution respectively,

$$f_\gamma(x)=\frac{1}{\Gamma(a)}x^{\gamma-1}e^{-x}, \qquad F_\gamma(x)=\frac{1}{\Gamma(a)}\int_0^x t^{\gamma-1}e^{-t}dt\,.$$

In Matlab environment these functions can be computed as $f_\gamma(x)=\text{gampdf}(x,\gamma,1)$ and $F_\gamma(x)=\text{gamcdf}(x,\gamma,1)$.

3 Extended M/M/*s* queue

In this section we describe a family of multi-server systems with Poisson arrival processes and state dependent service rates, which can be interpreted as extensions of the queueing system M/M/s to non-integral value of s. Let $R(x)$ be continuous function, which is positive for $x>0$ and satisfies:

$$\lfloor x\rfloor \le R(x) < \lfloor x\rfloor +1,\ x>0\,, \tag{9}$$

where $\lfloor x\rfloor$ is the integer part of x. Let also s be a positive number, $n=\lfloor s\rfloor$, and $r=R(s)$. Consider a queueing system with Poisson arrival process, $n+1$ servers, and infinite number of waiting places (see Figure 1). Servers 1,2,…, n have same service rate μ, while service rate of the server n+1 depends of the customer presence in the queue. If the queue is empty, server n+1 has service rate $(r-n)\mu$, otherwise it has service rate $(s-n)\mu$. We call servers 1,2,…, n as fast servers and the server n+1 as slow server, since its service rate is always less that μ. Note, that for $0<s<1$, the system has only one slow server.

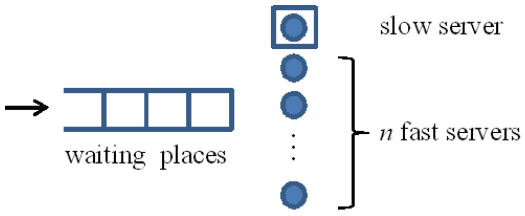

Figure 1. Queueing system with fractional number of servers.

Arriving customer first checks whether one of fast servers is free. If all fast servers are busy it checks whether the slow server is free. The first available server starts service. If no servers are available the customer waits in the queue. If fast server finishes service at a time when the slow server is busy, the customer processed at the slow server is immediately moved to the vacant fast server and its service is resumed from where it was left. Therefore slow server

may be busy only if there are no free fast servers. As soon as a server becomes free and the queue is not empty a customer is selected from the queue and dispatched to the free server.

If s is close to n+1 then service rate at the slow server is close to μ. It processes customers as fast server and the system behaves as M/M/n+1 queue. If $s>1$ is close to n then the slow server unlikely has enough time to finish service: when a fast server finishes its service the work processed at the slow server is interrupted and continued at the fast server. In this case the slow server is used as additional waiting place and the system behaves as M/M/n queue.

Let λ be arrival rate, and $\rho=\lambda/\mu<s$. It is easy to see that the number of customers in the system is the birth-and-death process with birth rates $\lambda_k=\lambda$ and death rates given by

$$\mu_k=\begin{cases} k\mu, & 1\le k\le n,\\ r\mu, & k=n+1,\\ s\mu, & k\ge n+2. \end{cases} \tag{10}$$

The stationary distribution of the number of customers in the system is given by

$$p(0)=\left(\sum_{k=0}^{n}\frac{\rho^k}{k!}+\frac{s\rho^{n+1}}{(s-\rho)rn!}\right)^{-1}, \tag{11}$$

$$p(k)=\begin{cases} p(0)\dfrac{\rho^k}{k!}, & 0\le k\le n,\\ p(0)\dfrac{\rho^n}{n!}\left(\dfrac{\rho}{r}\right)\left(\dfrac{\rho}{s}\right)^{k-n-1}, & k\ge n+1. \end{cases} \tag{12}$$

Therefore for the mean residence time in the system we have (Lazowska, 1984):

$$T=\sum_{k=1}^{\infty}\frac{k}{\mu_k}p(k-1)=\frac{1}{\mu}+\frac{C_{r,s}(\rho)}{s\mu-\lambda}, \tag{13}$$

where

$$C_{r,s}(\rho)=\frac{p(n)}{r(s-\rho)}\left((1-\delta)s^2+\delta\rho s+(s-r)(s-\rho)^2\right). \tag{14}$$

Note that the probability $p(n)$ in (14) can be expressed through Erlang's loss function as

$$p(n)=\frac{r(s-\rho)E_n(\rho)}{r(s-\rho)+s\rho E_n(\rho)}.$$

From (9) it follows that for integer s, function $R(s)$ satisfies $R(s)=n$, and it is easy to see that the mean residence time given by (13) coincides with the mean residence time in the M/M/s queue. Therefore function $C_{R(s),s}(\rho)$ in (13) can be considered as extension of Erlang's delay function $C_s(\rho)$. We consider three options for the function $R(s)$:

$$1)\quad R_1(s) = s\frac{B_n(\rho)s + B_n(\rho)(n-\rho)(s-\rho+\rho B_S(\rho))}{B_S(\rho)s + B_n(\rho)(s-\rho)(s-\rho+\rho B_S(\rho))} = s\frac{n-\rho+\dfrac{C_S(\rho)}{B_S(\rho)}}{s-\rho+\dfrac{C_S(\rho)}{B_n(\rho)}}; \tag{15}$$

$$2)\quad R_2(s) = s; \tag{16}$$

$$3)\quad R_3(s) = n + \frac{\delta\sqrt{c}}{\delta\sqrt{c}+(1-\delta)\sqrt{\rho}},\ \delta = s-n. \tag{17}$$

Case 1. If we equalise right sides of (1) and (13), for the function $R(s)$ we get formulas (15). In this case for any positive s the mean response time in the described system is given by formula (1), and for the delay probability we have $C_{R_1(s),s}(\rho) = C_S(\rho)$. We may say that the system with $R(s) = R_1(s)$ is a truly extension of the system M/M/s to non-integral value of s.

Case 2. In the simplest case with $R(s) = s$ the service rate of the slow server does not depend of the number of customers in the queue. The number of customers in the system is the birth-and-death process with birth rates $\lambda_k = \lambda$ and death rates given by

$$\mu_k = \begin{cases} k\mu, & 1 \le k \le n, \\ s\mu, & k \ge n+1, \end{cases}$$

and for the mean residence time in the system we have

$$C_{s,s}(\rho) = p(n)\left(\frac{s}{s-\rho} - \delta\right).$$

Case 3. This case does not require calculation of gamma distribution but $C_{R_3(s),s}(\rho)$ gives better approximation to the delay probability $C_S(\rho)$ then $C_{s,s}(\rho)$. Figures 2 and 3 show delay probability $C_{R_j(s),s}(\rho)$ for functions $R_1(s)$, $R_2(s)$, $R_3(s)$, and ρ/*s*=0.1, 0.5, 0.9.

4 Conclusion

In this paper we introduced a family of multi-server queueing systems that can be considered as extensions of conventional M/M/*s* queue to fractional number of servers *s*. One member of the family, specified by the function (15), has the mean residence time given by formula (1) with extended Erlang's delay function. Other family members can be used for approximate

calculation of delays in multi-server queues with fractional number of servers. Each system has Poisson arrival process and state-dependent service rates. Therefore our approach enables the analysis of open and closed networks with nodes having fractional number of servers using well known methods (Lazowska, 1984).

References

A.K. Erlang. Løsning af nogle Problemer fra Sandsynlighedsregningen af Betydning for de automatiske Telefoncentraler, Elektroteknikeren, 13, 5–13,1917.

G. Bretschneider. Extension of the equivalent random method to smooth traffics, Extension of the Equivalent Random Method to Smooth Traffics. *Proc. of 7th International Teletraffic Congress*, Stockholm, 411/1–9, 1973.

D.L. Jagerman. Some properties of the Erlang loss function. *Bell System Technical J.*, 53(3), 525–551, 1974.

L. Kleinrock. *Queueing Systems, Volume 1*. Wiley, New York, 1975.

D.T. Nightingale. Computations with smooth traffics and Wormald chart, *Proc. of 8th International Teletraffic Congress*, Sydney, 145/1–7, 1976.

M. Raiser. Mean value analysis and convolution method for queue-dependent servers in closed queueing networks, *Performance Evaluation*, 1(1), 7–18, 1981.

E.D. Lazowska et.al. *Quantitative system performance computer system analysis using queueing network models,* Prentice-Hall, London, 1984.

A.A. Jagers and E.A. van Doorn. Convexity of functions which are generalizations of the Erlang loss function and the Erlang delay function. *SIAM Review*, 33(2), 281–282, 1991.

W.L. Martinez and A. R. Martinez. *Computational Statistics Handbook with MATLAB,* Chapman & Hall/CRC Press, New York, 2002.

D.G. Down and G. Karakostas. Maximizing throughput in queueing networks with limited flexibility, *European J. of Operational Research*, 187(1), 98–112, 2008.

V. Naumov and O. Martikainen. Method for throughput maximization of multiclass network with flexible servers. *ETLA Discussion paper 1261*, The Research Institute of the Finnish Economy, Helsinki, 1–27, 2011.

V.B. Iversen. Teletraffic Engineering and Network Planning, DTU Course 34340, Technical University of Denmark, Lyngby, 2011.

F. Karsten, M. Slikker, and G.-J. van Houtum. Resource pooling and cost allocation among independent service providers, *Working Paper 352*, Beta Research School for Operations Management and Logistics, Eindhoven, 2011.

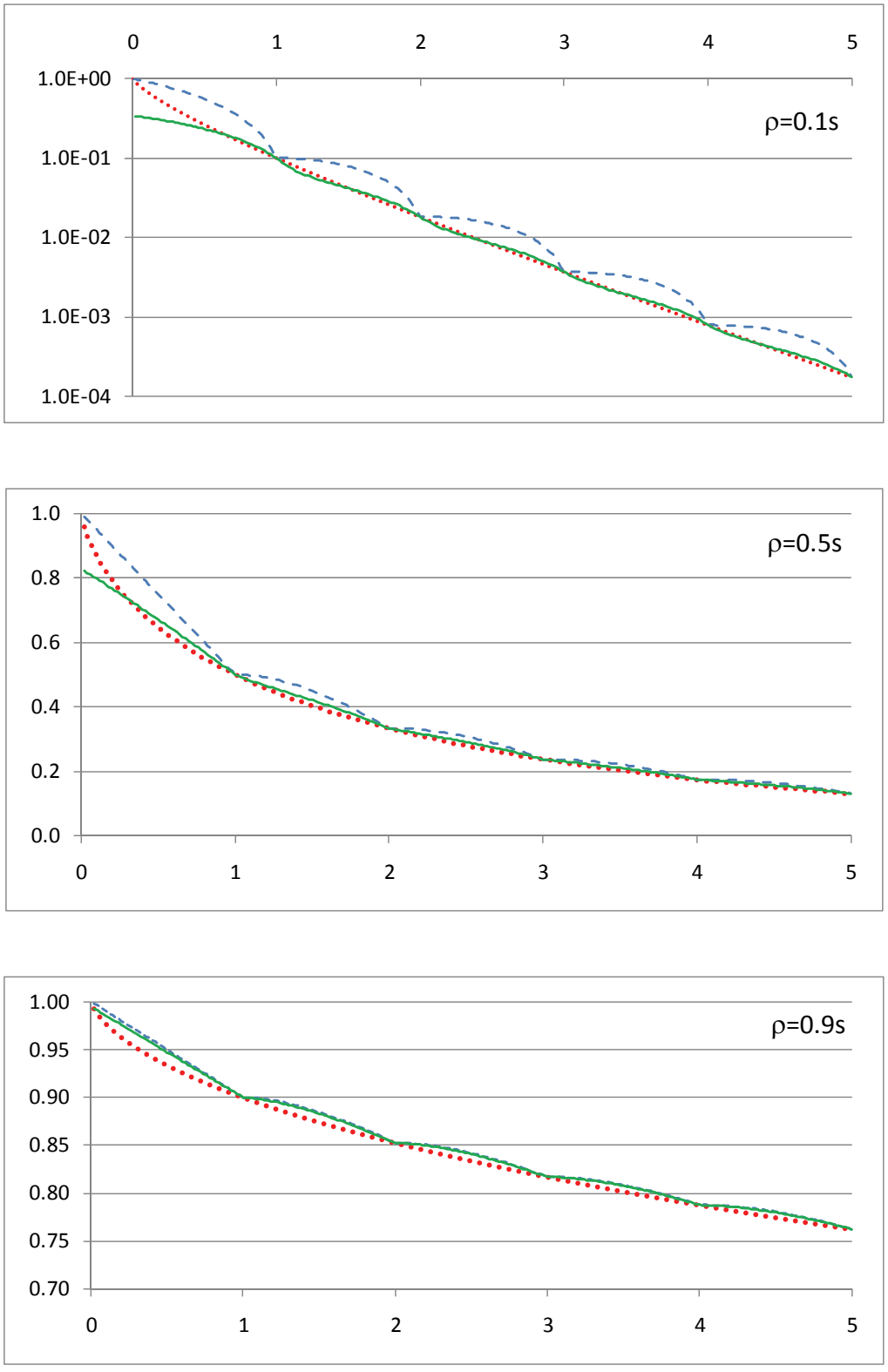

Figure 2. Probability of delay $C_{R(s),s}(\rho)$ for $0 < s \leq 5$,
R$_1$, R$_2$, R$_3$.

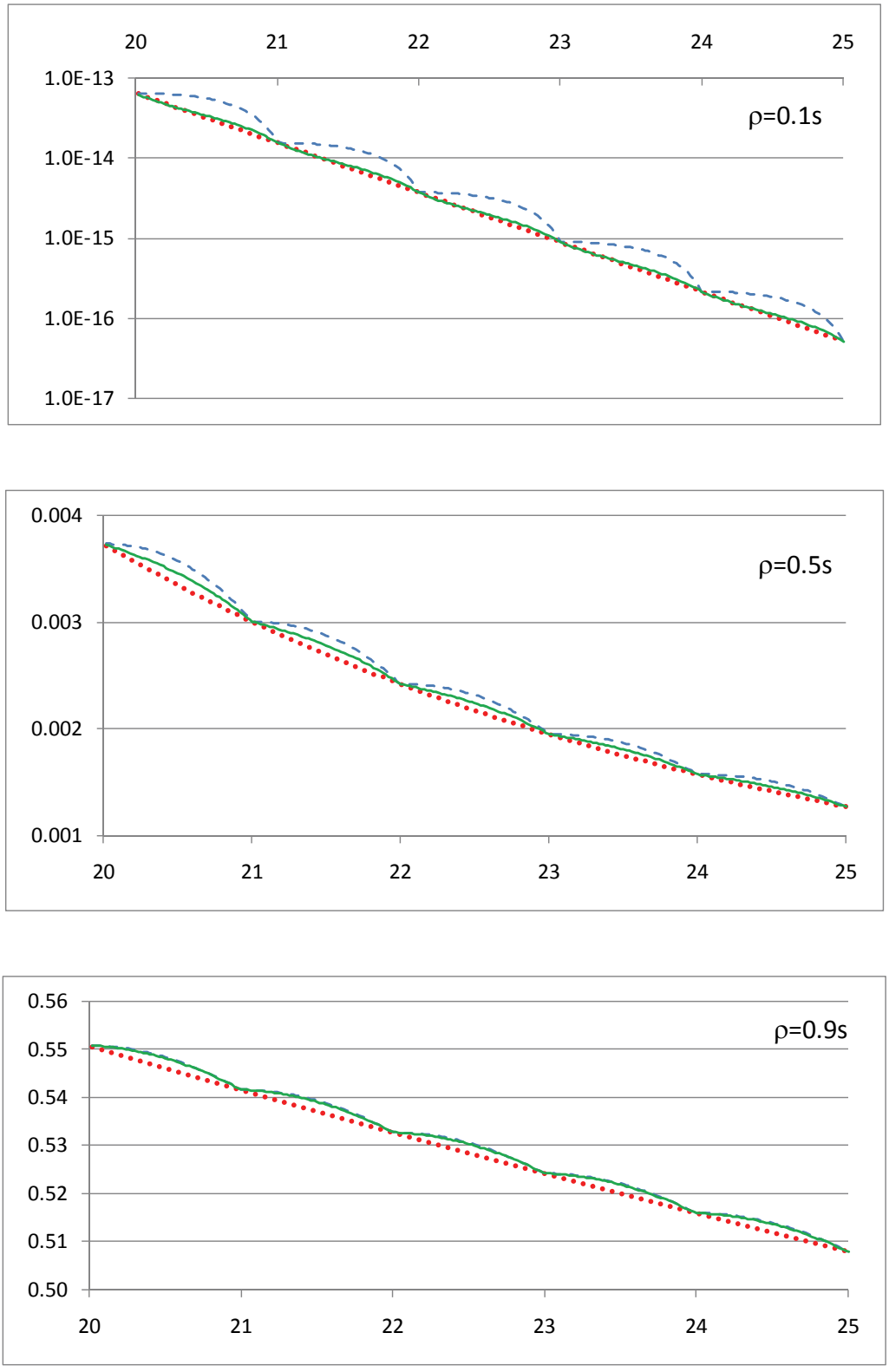

Figure 3. Probability of delay $C_{R(s),s}(\rho)$ for $20 \le s \le 25$,
R$_1$, R$_2$, R$_3$.

Aikaisemmin ilmestynyt ETLAn Keskusteluaiheita-sarjassa

Previously published in the ETLA Discussion Papers Series

No 1253 *Ari Hyytinen – Mika Maliranta*, Firm Lifecycles and External Restructuring. 17.06.2011. 34 p.

No 1254 *Timo Seppälä – Olli Martikainen*, Europe Lagging Behind in ICT Evolution: Patenting Trends of Leading ICT Companies. 22.06.2011. 18 p.

No 1255 *Paavo Suni – Pekka Ylä-Anttila*, Kilpailukyky ja globaalin toimintaympäristön muutos. Suomen koneteollisuus maailmantaloudessa. 19.08.2011. 39 s.

No 1256 *Jari Hyvärinen*, Innovaatiotoiminta: Näkemyksiä hyvinvointialaan ja työelämän kehittämiseen. 31.8.2011. 28 s.

No 1257 *Terttu Luukkonen – Matthias Deschryvere – Fabio Bertoni – Tuomo Nikulainen*, Importance of the Non-financial Value Added of Government and Independent Venture Capitalists. 2.9.2011. 28 p.

No 1258 *Ari Hyytinen – Mika Pajarinen – Pekka Ylä-Anttila*, Finpron vaikuttavuus – Finpron palveluiden käytön vaikutukset yritysten kansainvälistymiseen ja menestymiseen. 15.9.2011. 32 s.

No 1259 *Kari E.O. Alho*, How to Restore Sustainability of the Euro? 19.9.2011. 27 p.

No 1260 *Heli Koski*, Does Marginal Cost Pricing of Public Sector Information Spur Firm Growth? 28.9.2011. 15 p.

No 1261 *Valeriy Naumov – Olli Martikainen*, Method for Throughput Maximization of Multiclass Networks with Flexible Servers. 13.12.2011. 19 p.

No 1262 *Valeriy Naumov – Olli Martikainen*, Optimal Resource Allocation in Multiclass Networks. 14.12.2011. 17 p.

No 1263 *Jari Hyvärinen*, Innovaatiotoiminta: Suomi globaalitaloudessa. 30.12.2011. 49 s.

No 1264 *Jari Hyvärinen*, Productivity: An International Comparison. 30.12.2011. 20 p.

No 1265 *Jukka Lassila – Tarmo Valkonen – Juha M. Alho*, Fiscal Sustainability and Policy Rules under Changing Demographic Forecasts. 21.12.2011. 32 p.

No 1266 *Reijo Mankinen – Olavi Rantala*, Ulkomaanliikenteen palveluiden arvonlisäverotuksen käyttöönoton vaikutukset laiva- ja lentoliikenteeseen. 11.1.2012. 29 s.

No 1267 *Ville Kaitila – Pekka Ylä-Anttila*, Investoinnit Suomessa. Kehitys ja kansainvälinen vertailu. 30.1.2012. 34 s.

Elinkeinoelämän Tutkimuslaitoksen julkaisemat "Keskusteluaiheita" ovat raportteja alustavista tutkimustuloksista ja väliraportteja tekeillä olevista tutkimuksista. Tässä sarjassa julkaistuja monisteita on mahdollista ostaa Taloustieto Oy:stä kopiointi- ja toimituskuluja vastaavaan hintaan.

Papers in this series are reports on preliminary research results and on studies in progress. They are sold by Taloustieto Oy for a nominal fee covering copying and postage costs.

Julkaisut ovat ladattavissa pdf-muodossa osoitteessa: www.etla.fi/julkaisuhaku.php

Publications in pdf can be downloaded at www.etla.fi/eng/julkaisuhaku.php

ETLA
Elinkeinoelämän Tutkimuslaitos
The Research Institute of the Finnish Economy
Lönnrotinkatu 4 B
00120 Helsinki

Puh. 09-609 900
Fax 09-601 753
www.etla.fi
etunimi.sukunimi@etla.fi

ISSN 0781–6847