

HOE GROOT MOET EEN WACHTRUIMTE ZIJN?

N. M. van Dijk*

Dit artikel heeft een tweeledige doel: concreet en illustratief.

- (i) *Konkreet wordt aan de hand van een klassiek wachttijdmodel onderzocht welke invloed de grootte van een wachtruimte heeft op de prestatie van een bedieningssysteem, zoals een bank, een informatienummer of een ziekenhuiswachtkamer, waarbij de vraag de capaciteit overschrijdt. Zowel theoretisch als numeriek wordt onderbouwd dat grote wachtruimtes daarbij vermeden moeten worden.*
- (ii) *Illustratief wordt met dit resultaat ondersteund hoe een betrekkelijk eenvoudige kwantitatieve aanpak tot praktisch zinvolle conclusies kan leiden.*

1. INLEIDING

Algemene achtergrond en motivatie

Wachten is een universeel probleem waar elk van ons maar al te vaak mee geconfronteerd wordt: bij de bushalte, in de bank, op het postkantoor, in de supermarkt, bij de dokter, in het ziekenhuis, bij algemene informatienummers, in files op de snelweg, bij een stoplicht, bij het inchecken op een vliegveld, of in restaurants, om maar eens een greep te doen.

Hoogst verwonderlijk echter en in strijd met hedendaagse logistieke modewoorden als snelheid en efficiëntie, lijken wij ons min of meer bij dit wachten neer te leggen. "Overmacht" en een irritant maar onvermijdelijk "bijverschijnsel" zijn zoal geluiden waarmee het dikwijls wordt afgedaan. Desondanks zullen slechts weinigen zich niet van tijd tot tijd ergeren aan het verschijnsel wachten. Zo werd in een artikel in de "Boston Globe" door een anonieme schrijver (Mr. Thoughtful) gesteld; "Waiting is the worst in the world", (zie [2]).

Theoretische modelvorming en analyse van wachttijdproblematiek staat bekend onder de naam *wachttijdtheorie* oftewel *queueing*. Dit vakgebied is begin 20-er jaren ontstaan t.b.v. het zo efficiënt mogelijk ontwikkelen en besturen van telefooncentrales. Tot op de dag van vandaag vervult dit vakgebied nog steeds een belangrijke rol in de telecommunicatie, zoals voor het evalueren en ontwikkelen van gecombineerde spraak en gegevens (voice-data) transmissies over digitale kanalen (ISDN-netwerken). (Zie bijv. Schwartz 1987 en Van Dijk et al 1990).

Een andere belangrijke toepassing van wachttijdtheorie werd begin 60-er jaren gevonden in de opkomst van computer netwerk systemen, hetgeen eveneens tot op de dag van vandaag een intensief onderzoeksgebied vormt en bekend staat onder de naam: "Computer Prestatie Analyse" (zie bijv. Lavenberg 1982, King 1990 en Harrison en Patel 1992).

* Universiteit van Amsterdam
Fac. der Economische Wetenschappen
en Econometrie
Roetersstraat 11
1018 WB AMSTERDAM
Tel: *31 20-525.4215

Helaas echter blijkt zelfs tot op nu, begin 90-er jaren de toepassing van wachttijdtheorie voornamelijk beperkt gebleven te zijn tot deze twee gebieden. Relatief zeer *weinig kennis* in de wachttijdtheorie inmiddels opgedaan, lijkt *gebruikt te worden* voor andere en wellicht belangrijker alledaagse wachttijdsituaties, zoals bijvoorbeeld in postkantoren, in supermarkten, bij de burgerlijke stand, in ziekenhuissituaties, bij recreatieparken etc.

Enkele fraaie uitzonderingen hierop vindt men bijvoorbeeld vermeld in Halpern 1979 (Brandbestrijding), Sehlinger en Finley (Walt Disney World) en de Telegraaf 1992 (Supermarkten).

Voor deze discrepantie zijn verschillende gedeeltelijke verklaringen te bedenken. Zo bezitten verscheidene service instanties min of meer een monopolie positie of zijn uniek zodat de klant simpelweg ongewenste wachtsituaties voor lief moet nemen. Een maar al te bekend voorbeeld hiervan is de hoog oplopende wachttijd in ziekenhuizen om de artsen te vrijwaren van te moeten wachten op een patiënt (zie bijv. Vissers 1992). Een andere belangrijke verklaring is terug te vinden in **psychologische** factoren die vaak een dominerende rol kunnen vervullen ten opzichte van kwantitatieve argumenten. (zie bijv. Palm 1953, Mann 1969, Martin 1983, Morrow 1984, Maister 1985, Larson 1987 and Pruyn 1992). In het bijzonder blijkt het verschil tussen objectieve en gepercipiëerde wachttijden soms groot te zijn en afhankelijk van allerlei bijfactoren (zie Pruyn 1992).

Ten slotte kunnen ook **andere niet direct zichtbare** en kwantificeerbare factoren een alles dominerende rol vervullen, zoals bijv. geïllustreerd door Martin 1983, aan de hand van een vliegveld-case, waar de klachten van passagiers bij bagage banden niet verdwenen ondanks reductie van de wachttijd tot slechts 8 minuten.

Dit alles echter neemt niet weg, dat de **kwantitatieve** bestudering van wachttijdsituaties toch in beginsel een eerste natuurlijk middel zou moeten zijn om wachtsituaties te **evalueren en te verbeteren**. Zo werd onlangs door de PTT besloten in vele postkantoren over te gaan op een systeem van één lange wachtrij i.p.v. afzonderlijke wachtrijen voor afzonderlijke loketten met als hoofdbepalend (psychologisch) argument, dat dit eerlijker is omdat men nooit in de verkeerde rij staat. Een kwantitatieve analysering van dit probleem echter zou tot een genuanceerdere aanbeveling leiden (zie bijv. Volkskrant 1992 en Rothkopf and Rech 1987), waarbij differentiatie naar verschillende handelingen een belangrijke rol kan spelen. Een soort gelijke opmerking geldt bijvoorbeeld ook voor ziekenhuizen.

Doel

Dit artikel beoogt in de eerste plaats te illustreren dat kwantitatieve wachanalyses van duidelijk nut kunnen zijn ook voor vrij universeel optredende situaties. Hiertoe wordt als speciaal verschijnsel het veelvuldig voorkomen van te grote wachtruimtes aan de kaak gesteld, hetgeen ontleend is aan een specifiek praktisch onderzoeksproject.

Specifiek wachttijdprobleem

Het afwijzen van klanten ziet men in het algemeen wel als het allerlaatste dat mag plaatsvinden. Als gevolg daarvan zal men veelal streven naar een zo veel mogelijk accepteren van service verzoeken, ook wanneer een systeem reeds zeer vol zo niet overvol is. Dit wordt tot bijna elke prijs doorgevoerd, in het bijzonder:

de prijs van het wachten

De grootte van wachtfaciliteiten lijkt dan ook in de praktijk eerder door technische of opslag

beperkingen bepaald dan door bewuste afkapping. Deze beperkingen staan veelal vrij grote wachtfaciliteiten toe.

Met name onder hoge werkbelasting kan dit tot zeer extreme wachttijden leiden. Aangezien systeem capaciteiten normaliter worden afgestemd op *gemiddelden* en niet op de "hoogste belastingsgraad", komen dit soort situaties helaas vrij veel voor. Speciale aandacht daarvoor is derhalve geboden.

In een recent artikel van Biaocchi 1992 is de invloed van de grootte van de wachtruimte gedetailleerd bestudeerd als de wachtruimte naar een oneindige grootte groeit. De hierin verkregen resultaten echter richten zich uitsluitend op asymptotische resultaten en convergentiesnelheden en niet op resultaten voor een specifieke (ook relatief kleine) grootte.

Resultaat

In het kader van bovenstaande beoogt dit verhaal twee punten naar voren te brengen die beide primair gericht zijn op praktische waarde:

- (i) Dat **grote** wachtruimtes onder hoge bezettingsgraad **vermeden** moeten worden.
- (ii) Dat **een simpele**, zowel experimentele als theoretische **kwantitatieve analyse** tot deze conclusie leidde.

Het eerste punt is van mogelijke praktische waarde. Het toont aan dat wachttijden aanzienlijk teruggebracht kunnen worden tegen een slechts zeer kleine verhoging van afwijzingskansen. Het tweede punt is algemeen bedoeld om naar de "buitenwereld" over te brengen dat **kwantitatieve** en soms zelfs zeer eenvoudige modellen tot **praktisch nuttige** en soms verbazingwekkende conclusies kunnen leiden. Meer **erkenning van een kwantitatieve aanpak** zou derhalve op zijn plaats zijn.

Presentatie

Om te illustreren dat de vereiste kwantitatieve analyse soms betrekkelijk eenvoudig kan blijven, is gekozen voor een presentatie die **ook voor de niet met wachttijdtheorie of operationele research bekende lezer** te volgen moet zijn, slechts gebruik makend van simpele calculus. Voor hen die wel hiermee bekend zijn enerzijds met verontschuldiging voor de eenvoud, maar anderzijds met de hoop dat ook zij hun kennis meer open stellen voor alledaagse praktische problematiek.

2. MODEL BESCHRIJVING

We beschouwen een bedieningssysteem dat in de wachttijdtheorie wordt aangeduid met $M | M | c | c+m$. Dit systeem heeft c bedienden en m additionele wachtplaatsen. Klanten arriveren volgens een Poisson proces met parameter λ en vereisen een exponentiële hoeveelheid bediening met parameter μ . M.a.w. klanten komen willekeurig aan met intensiteit λ per tijdseenheid en elke bediende kan μ klanten per tijdseenheid afhandelen. Preciezer, gedurende een willekeurig klein tijdsinterval ter lengte Δt waarbij k bedienden werkzaam zijn zal met kans

$$\begin{cases} \lambda \Delta t + o(\Delta t): & \text{een klant arriveren} \\ k \mu \Delta t + o(\Delta t): & \text{een bediening beëindigen} \end{cases}$$

met $o(\Delta t)$ van orde $(\Delta t)^2$. Indien alle wachtplaatsen en dus ook bedienden bij binnenkomst van een klant bezet zijn wordt deze klant afgewezen en gaat verloren.

We zijn geïnteresseerd in het zogenaamde "*steady state*" gedrag van dit systeem oftewel, ruwweg gezegd, prestatie maten bezien of uit gemiddeld over lange tijdsperioden dan wel bij observatie op een willekeurig moment. In het bijzonder gelden als meest belangrijke prestatie indicatoren de *verlieskansen B* dat een binnenkomende klant wordt afgewezen en de (verwachte) *wachttijd W* die een geaccepteerde klant zal ondergaan. Hierbij beperken we ons tot de juist zo praktisch belangrijke situatie waarbij de werkbelasting groter is dan de capaciteit van het systeem aangegeven met:

$$\rho = \frac{\text{aangeboden werklast per tijdseenheid}}{\text{maximale bedieningscapaciteit per tijdseenheid}} = \frac{\lambda}{c \mu} > 1$$

STEADY STATE

We bepalen hiertoe eerst de evenwichtskansen $\pi(k)$ dat zich op een willekeurig moment k klanten in het systeem bevinden. Deze kansen kunnen direct bepaald worden uit de welbekende *geboortesterfte* oftewel *uit=instroom* vergelijkingen, gegeven door:

$$\begin{cases} \pi(k)k\mu = \pi(k-1)\lambda & (k=1, \dots, c) \\ \pi(k)c\mu = \pi(k-1)\lambda & (k=c+1, \dots, c+m) \end{cases}$$

en welke geïnterpreteerd kunnen worden als het gelijkstellen van de kans om uit een toestand te verdwijnen en om in die toestand terecht te komen (bezien gedurende een kleine tijd en per tijdseenheid), zodat als netto resultaat de kans op die toestand ongewijzigd blijft. Als oplossing vindt men:

$$\begin{cases} \pi(k) = \pi(0) \frac{1}{k!} \left[\frac{\lambda}{\mu} \right]^k & (k = 0, \dots, c) \\ \pi(k) = \pi(0) \frac{1}{c!} \left[\frac{\lambda}{\mu} \right]^c \rho^{k-c} & (k = c+1, \dots, c+m) \\ \pi(0) = \left[\sum_{k=0}^c \frac{1}{k!} \left[\frac{\lambda}{\mu} \right]^k + \frac{1}{c!} \left[\frac{\lambda}{\mu} \right]^c \sum_{s=1}^m \rho^s \right]^{-1} \end{cases}$$

De grootheden B en W kunnen hieruit direct berekend worden als:

$$\begin{cases} B = \pi(c+m) & : \text{Verlieskansen} \\ W = \sum_{k=0}^{c+m-1} \pi(c+k) \frac{(k+1)}{c\mu} & : \text{Verwachte wachttijd} \end{cases}$$

De grootheid B voor het speciale geval $m=0$ staat bekend als *Erlang's verliesformule* (zie [1]) en zal hierna aangeduid worden met het symbool E .

3. ASYMPTOTIEK

We beogen nu het gedrag van B en W te onderzoeken als functie van de wachtruimte omvang m . Intuïtief mag men verwachten dat de verlieskans B minstens groter is dan:

$$[1 - \text{fraktie van de aangeboden lading die maximaal afgehandeld kan worden}] = [1 - \frac{1}{\rho}] = [(\rho - 1)/\rho]$$

Het zal blijken dat deze waarde inderdaad een ondergrens is en van bovenaf benaderd wordt als m heel groot wordt. Maar hoe snel is deze benadering en wat is daarbij de prijs die we moeten betalen in termen van wachttijd als we de wachtruimte laten groeien? Deze vragen zullen we hieronder nader onderzoeken.

Verlieskans B

Vanwege de normalisatie constante $\pi(0)$ is het handiger het asymptotisch gedrag van $1/B$ i.p.v. B te onderzoeken. Dit leidt namelijk tot:

$$\begin{aligned} \frac{1}{B} &= \left[\sum_{k=0}^c \frac{1}{k!} \left(\frac{\lambda}{\mu} \right)^k + \frac{c^c}{c!} \rho^c \sum_{s=1}^m \rho^s \right] / \left[\frac{c^c}{c!} \rho^{c+m} \right] \\ &= \frac{1}{\rho^m} \left\{ \frac{c!}{\rho^c c^c} \left[\sum_{k=0}^c \frac{1}{k!} \left(\frac{\lambda}{\mu} \right)^k \right] + \sum_{s=1}^m \rho^s \right\} \\ &= \frac{1}{\rho^m} \frac{1}{E} + \frac{1}{\rho^m} \rho \left[\frac{\rho^{m-1}}{\rho-1} \right] \\ &= \frac{\rho}{\rho-1} + \frac{1}{\rho^m} \left[\frac{1}{E} - \frac{\rho}{\rho-1} \right] \end{aligned}$$

Het omgekeerde van de verlieskans B convergeert dus bij toenemende grootte van de wachtruimte naar:

$$\boxed{\frac{\rho}{\rho-1} = \left[\frac{\rho-1}{\rho} \right]^{-1} = \left[1 - \frac{1}{\rho} \right]^{-1}}$$

met een snelheid van orde

$$\boxed{\left(\frac{1}{\rho} \right)^m \text{ waarbij } \rho > 1}$$

Door tevens op te merken dat $E > 1 - \frac{1}{\rho} = \frac{\rho-1}{\rho}$, wordt deze limiet van beneden benaderd met

constante:

$$\left[\frac{1}{E} - \frac{\rho}{\rho - 1} \right]$$

Deze constante hangt alleen af van de intensiteiten λ en μ en het aantal bedienden c ; maar **niet** van de **grootte van de wachtruimte: m !** Als gevolg van de convergentie $(1/\rho)^m$ met $\rho > 1$ en het feit dat deze constante ook vrij snel klein wordt voor *redelijk kleine c waarde*, zal dus ook de verlies- kans B voor realistische waarden als $\rho \sim 1.2 - 1.5$ de limietwaarde van bovenaf vrij snel dicht benaderen, zeg voor $m \geq 10$ en soms zelfs kleiner. Dit wordt geïllustreerd in figuur 1.

Wachttijd W

De afhankelijkheid van de wachttijd als functie van de grootte van de wachtruimte kan direct uitgewerkt worden volgens:

$$\begin{aligned} W &= \sum_{k=0}^{m-1} \pi(c+k) \left[\frac{k+1}{c\mu} \right] \\ &= \pi(0) \frac{1}{c!} \left[\frac{\lambda}{\mu} \right]^c \left[\frac{1}{c\mu} \right] \sum_{k=0}^{m-1} (k+1) \rho^k \\ &= \pi(0) \frac{1}{c!} \left[\frac{\lambda}{\mu} \right]^c \left[\frac{1}{c\mu} \right] \frac{d}{d\rho} \left[\sum_{k=0}^{m-1} \rho^{k+1} \right] \\ &= \pi(0) \frac{1}{c!} \left[\frac{\lambda}{\mu} \right]^c \left[\frac{1}{c\mu} \right] \left[\frac{(m+1)(\rho-1)\rho^m - \rho^{m+1} + 1}{(\rho-1)^2} \right] \\ &= \pi(0) \frac{1}{c!} \left[\frac{\lambda}{\mu} \right]^c \frac{\rho^m}{c\mu} \left[\frac{(m+1)(\rho-1) - \rho}{(\rho-1)^2} \right] + \pi(0) \frac{1}{c!} \left[\frac{\lambda}{\mu} \right]^c \frac{1}{c\mu} \frac{1}{(\rho-1)^2} \end{aligned}$$

$$= [m+1] \left[\frac{1}{c\mu} \right] \left[\frac{B}{\rho-1} \right] - \frac{B}{c\mu} \frac{\rho}{(\rho-1)^2} + \pi(c) \frac{1}{c\mu} \frac{1}{(\rho-1)^2}$$

Met B en analoog $\pi(c)$ snel naderend naar een constante als hierboven aangetoond met toenemende m , neemt de afhankelijkheid al snel een lineaire gedaante aan, als eveneens in figuur 1 zichtbaar gemaakt. Ruwweg leidt elke extra wachtplaats daarbij tot een toename van de wachttijd ter grootte van:

$$\left[\frac{1}{c\mu} \frac{B}{\rho-1} \cong \frac{\rho}{c\mu} \right]$$

Conclusie

Voor bedieningssystemen met hoge werkbelasting moet men uiterst voorzichtig zijn met het toelaten van grote wachtruimten. In zulke situaties zullen grote i.p.v. kleinere ruimten nauwelijks een verbetering van verliezen opleveren maar wel snel tot onacceptabele wachttijden leiden.

EVALUATIE

Aan de hand van een simpel wachttijdmodel en eenvoudige calculus werd het effect bestudeerd van het vergroten van wachtruimten voor eindige bedieningssystemen onder hoge werkbelasting. Zowel theoretisch als numeriek onderbouwd kon zo getoond worden dat **grote wachtruimten**, in het bijzonder in dit soort situaties, **vermeden** moeten worden. Het bewust maken van dit type resultaat kan van praktisch nut zijn voor prestatieverbetering.

Naast de resultaten op zich werd hiermee tevens geïllustreerd dat **kwantitatieve technieken** en onderbouwing, ook zelfs berustend op meest **simple modellen**, van belang kunnen zijn voor **alledaagse praktische situaties**.

Dankwoord

Met dank aan drs. E van der Sluis voor de grafische ondersteuning en aan de referent voor het kritisch lezen en commentariëren van eerdere versies.

Referenties

1. Baiocchi A. (1992): "Asymptotic Behaviour of the Loss Probability of the M/G/1/K and G/M/1/K Queues", Queueing Systems 10, 235-248.
2. Boston Globe, 1984. Confidential Chat: "Playing the Waiting Game", (November 17).
3. Brockmeyer, Hållstrom and Jensen (1948): "The Life and Works of A.K. Erlang", Trans 2, Copenhagen: OATS, 1-277.
4. Cooper, R.J. (1987): "An Introduction to Queueing", Wiley and Sons.
5. Halpern, J. (1979): "Fire Loss Reduction: Fire Detectors vs. Fire Stations". Mgmt. Sci. 1082-1092.
6. Harrison, P.G. and Patel, N.M. (1992): "Performance Modelling of Communication Networks and Computer Architectures", Addison Wesley.
7. Larson, R.C. (1987): "Perspectives on Queues: Social Justice and the Psychology of Queueing". Operations Research 35, 895-905.
8. King, P.J. (1990): "Computer and Communication Systems Performance Modelling", Prentice Hall.
9. Lavenberg, S.S. (1982): "Computer Performance Modelling Handbook", Academic Press.
10. Mann, L. (1969): "Queue Culture: The Waiting Line as a Social System". Am J. Sociol. 75, 340-354.

11. Maister, D.H.(1986): " The Psychology of Waiting Lines", Chap. 8 in "The Service Encounter", Czepiel, J.A., Solomon, M.R. and Surprenant, D.F. (eds.). D.C. Heath. Lexington, Mass.
12. Martin, A. (1983): "Perception and Value Management", Think Proactive 8, 95-101.
13. Morrow, L. (1984): "Waiting as a Way of Life", Time, July 23 (1984), p. 65.
14. Palm, C. (1953): "Methods of judging the Annoyance Caused by Congestion", TELE (English ed.) no. 2, 1-20.
15. Pruyn, Th. (1992): "Psychologische Beleving van Rijen". Presentatie EUROFORUM Studiedag: "Het Verkorten van Rijen en Wachttijden", 4 december 1992.
16. Rothkopf, M.H. and Rech P. (1987): "Perspectives on Queues: Combining Queues is not Always Beneficial", Opns. Res. 35, 906-909.
17. Schwartz, M. (1987): "Telecommunication Networks", Addison Wesley
18. Sehlinger, R. and Finley, J. (1985): "The Unofficial Guide to Walt Disney World", Menasha Ridge Press, Hillsborough N.C.
19. Telegraaf d.d. 19 november 1992: "Nooit meer in de Rij", (Interview met N.M. van Dijk n.a.v. Supermarktketen Hoogvliet).
20. Van Dijk, N.M., Gerrand, P. Henderson, W.T., Warfield, B. (1990): "Traffic Theories for New Telecommunication Services".
21. Vissers, J.M.H. (1992): "Terugdringen van Wachtijden in Poliklinieken", Presentatie EUROFORUM studiedag: Het Verkorten van Rijen en Wachtijden, d.d. 4 december 1992.
22. Volkskrant, d.d. 24 april 1992: "Waarom Wachten Altijd te Lang Duurt", (Interview met N.M. van Dijk).

ontvangen	4 - 12-1992
geaccepteerd	4 - 6 -1993

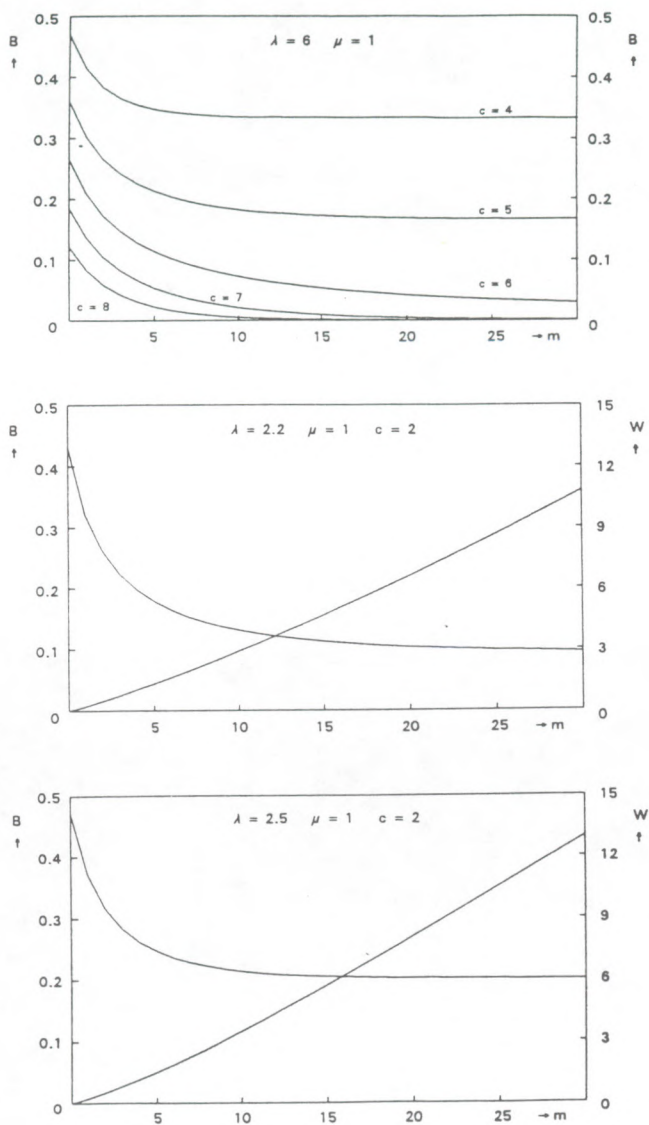


Figure 1 Loss probability and expected waiting time in different systems.

