Norwegian University of Science and Technology
Department of Telematics

Contact during exam [Faglig kontakt under eksamen]:
Bjarne E. Helvik (92667)

EXAM IN COURSE [EKSAMEN I EMNE]
TTM4110 Dependability and Performance with Discrete event Simulation [Pålitelighet og ytelse med simulering]

Thursday [Torsdag] 2007-12-06
09:00 – 13:00

The English version starts on page 2.

Den norske bokmålsutgaven starter på side 8.

Hjelpemidler:
C - Graham Birtwistle: DEMOS - A system for Discrete Event Modelling on Simula. Collection of formulas for TTM4110 Dependability and Performance with Discrete event Simulation. NB! the collection of formulas are attached. Predefined simple calculator.
[Graham Birtwistle: DEMOS - A system for Discrete Event Modelling on Simula. Formelsamling i fag TTM4110 Pålitelighet og ytelse med simulering. NB! Formelsamlingen er vedlagt. Forhåndsbestemt enkel kalulator.]

Sensur 2008-01-10

English version[1]

In this exam we will regard an enterprise operating a call centre[2] and the ICT infrastructure of this enterprise. See Figure 1 for an illustration. Denote this enterprise Quality Response and for short QR. The enterprise receives VoIP calls via the Internet requesting information, assistance and similar. A call centre host manages the interaction and must be operative for the calls to be handled. A call is transferred by a router and a switch to a *call agent*. In fault free operation, there is no capacity limitation in the technical part of the system and one of the routers may handle the entire traffic load.

The calls stem from an infinite number of callers (sources) and the aggregated call intensity is λ. The duration of a call is negatively exponentially distributed with expected value $\mu^{-1} = 180\,s$. The duration of the calls are i.i.d. QR has n call agents. The workstation of each agent is connected to one switch which is connected to two routers. There are k switches. (For the sake of simplicity, it is assumed that n/k is an integer.) In addition, the workstations may reach the routers via a wireless access point. The system components fail independently with a constant intensities. The failure intensities are: λ_r for the router, λ_h for the call centre host, λ_s for the switch, λ_a for the call agent workstation and λ_w for the wireless access point. The interconnection between these system components as well as the Internet is for the moment assumed to be fault free. If a system component fails, all calls using this component are lost.

QR markets its service with the following statements: "During busy hours, at least 99.5 % of the calls to QR reach a call centre agent. This requirement may be relaxed if equipment at QR fails. The probability that an established call is interrupted due to failure is less than $5 \cdot 10^{-4}$. QR guarantees that no more than one hour accumulated time per year shall the call centre have less than 75% its busy hour capacity."

a) Define the QoS statements of QR in terms of common technical dependability and performance parameters.

b) Assume that calls to QR that do not find a vacant call agent, receive a busy tone and are cleared. What kind of system is this; use Kendall's notation? Give an expression for the probability that a call does not find a vacant call agent. What is the fraction of the time that all agents are busy? Justify the answers.

To increase the utilization of the call agents and the relative number calls served, QR let all callers that do not find a vacant agent queue for one. Entering the queue, the caller gets the voice message: "You are number i in the queue. The expected waiting time is τ_i." Empirical experience from other call centres is that the caller then immediately closes the call (hangs up the phone) and leaves the system with a probability q_i. For the sake of simplicity, it is assumed that those who do not immediately leave, stay in the system until they are served. The call centre host ensures that calls are served in the order of arrival.

c) Draw a diagram depicting a continuous time discrete state Markov model of the number of callers in the system. Denote the probability that there is j callers in the system by p_j, and establish a set of equations that may be used to determine p_j.

[1] In case of divergence between the English and the Norwegian version, the English version prevails.

[2] A call centre is a centralised office used for handling a large volume of calls.

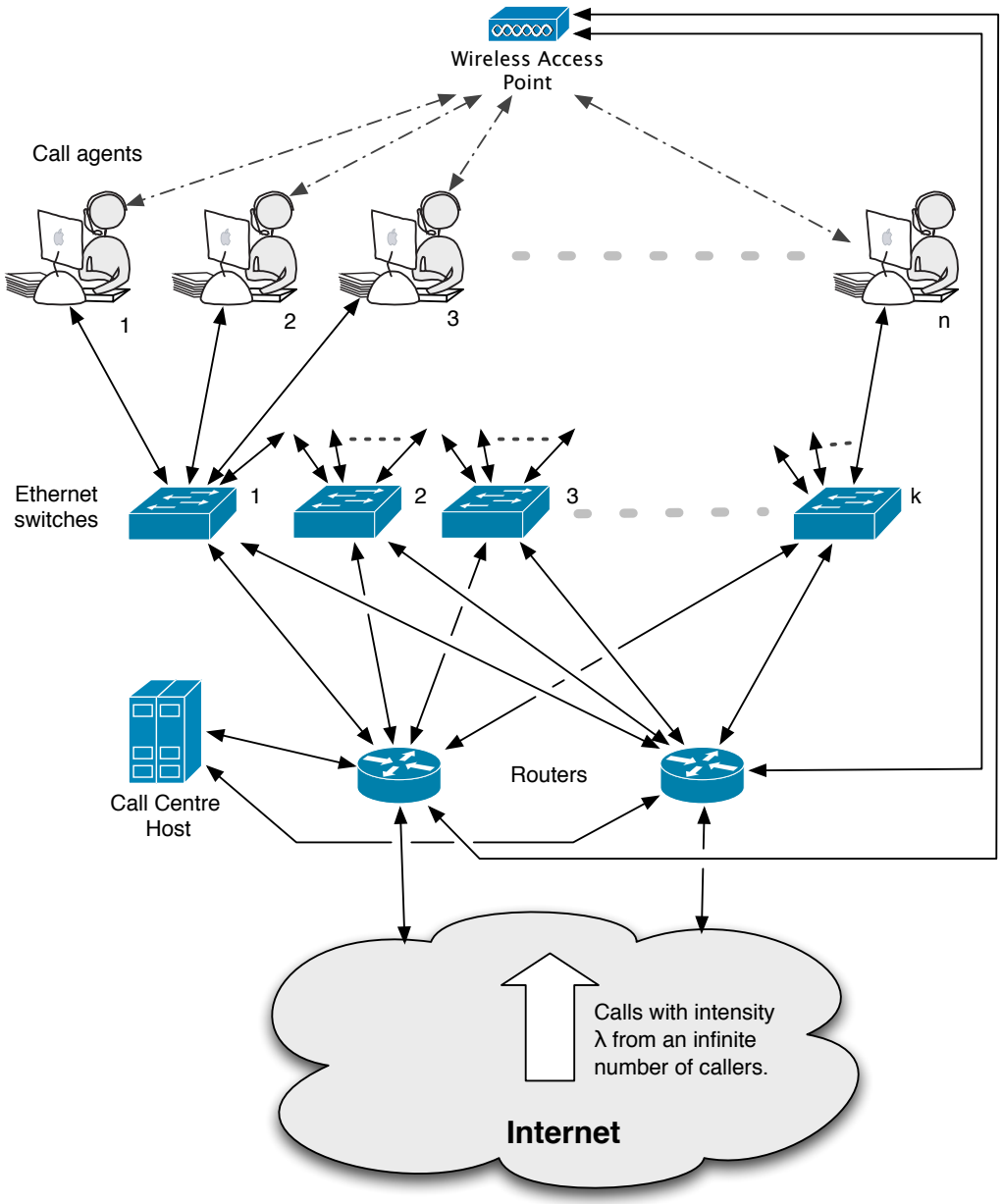


Figure 1: Illustration of the ICT infrastructure of QR, including the call agents.

d) Assume that the probabilities p_j are found. Establish expressions for the carried traffic, the lost traffic, the probability the a caller gets a voice message ("You are number $\cdots$.") and the expected waiting time for those who wait.

In the next questions, do not regard the wireless access point as a part of the system.

e) For the system, it is required that the probability for an established call to be interrupted due to failure is less than $5 \cdot 10^{-4}$. This translates into a requirement for the interruptions intensity ξ of ongoing calls, $\xi \leq 0.01\text{hour}^{-1}$. Find an expression for ξ.

QR's ICT infrastructure has a system failure when the call centre has less than 75% of its full capacity in terms of call agents that handle calls. To predict the system availability with respect to this requirement, two simplifying assumptions are made: 1) All system components are repaired independently and the mean down time of all components is assumed to be d. 2) The workstations are assumed not to fail, i.e. $\lambda_a = 0$.

f) How will the simplifying assumptions bias the prediction of the system availability? Justify the answer. Draw a dependability model of the system using the above simplifying assumptions and find an expression for the system availability A_S.

Assume that the above analysis shows that the availability requirement is not met. Instead of the costly operation of introducing more switches and re-cable the workstation access, etc., QR introduces a wireless access point reachable from all workstations as shown in Figure 1. The access point can handle at most k_w workstations simultaneously.

g) If $k_w = n/k$, $\lambda_w = \lambda_s$ and the simplifying assumptions may be used, how may the model found in question f be modified to include the access switch?

The above approximation do not yield a sufficient accuracy. Regard the subsystem consisting of the switches and the wireless access. Use that $\lambda_w \neq \lambda_s$, that switches and the wireless access point have different repair times, i.e. $d_w \neq d_s$, and that at most one component is repaired at a time. Repair times may be assumed to be negatively exponentially distributed and independent of each other.

h) Make a state diagram (Markov model) of the switches and the wireless access point, when $k = 3$, which may be used to find the availability of this subsystem with respect to the system failure requirement. Define clearly each state and which states that yields a working and failed system.

Assumptions made are still too rough. The maximum number of workstations handled by the wireless access point is different from n/k. Links between the workstation and the switch do also fail with an intensity $\lambda_l > 0$. To rotate workload among the agents, each workstation has to release its connection to the access point after eight calls. Assume that the time from a call agent becomes accessible for calls and until he/she receives a call is a randomly distributed time `timeToCall`.

i) Assume a specific number $m > k_w$ of workstations using the wireless access. Also assume that the wireless access point cannot fail. Make a simulation model to study the time a workstation is not connected to the wireless access point and the call agent is thus unreachable. Draw an activity diagram for the model. (Note that in this question, only the use of the wireless access point is studied.)

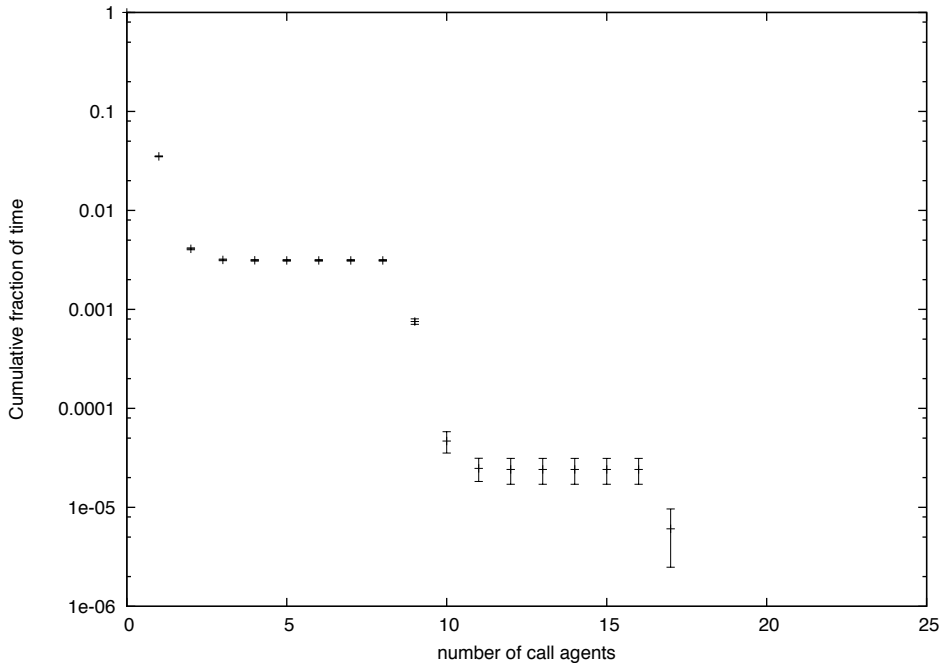


Figure 2: The cumulative fraction of time at least n_w call agent's workstations need to use the wireless access point.

The complete code of a simulator of the call agents using switches and the wireless access point is implemented in the Simula/DEMOS given in Listing 1starting on Page 6. The switches and the links between the workstations and the switches may fail. The workstation themselves and the wireless access point do not fail. Workstations are indexed i, j, where $i = 1, \ldots, k$ represents the switch number and $j = 1, \ldots, n/k$ represents workstation number on a specific switch. The simulation time unit is [hour].

j) What are the entities defined in the simulator? What is the purpose of the `Z` resource? What does the `state` attribute of the `CallAgent` and the `X` and `Y` resources represent? What is defined as the global state, i.e. variable `globalState`, of the system? (Note that it is *not* required that an activity diagram of the simulator is made.)

The graph in Figure 2 plots the cumulative fraction of time at least n_w call agent's workstations need to use the wireless access point given with 95%-confidence intervals. The failure intensities, repair times, the number of call agents, the number of switches and the simulation time are those used in the Simula/DEMOS code above, i.e. $\lambda_s = 2\,\mathrm{year}^{-1}$, $\lambda_l = 1\,\mathrm{year}^{-1}$, $d_s = 3\,\mathrm{hours}$, $d_l = 48\,\mathrm{hours}$, $n = 24$ and $k = 3$. The confidence intervals are obtained by running ten replications of the simulation.

k) Give an interpretation of the shape of the graph related to the structure of the system. What is the minimal number of workstations k_w the wireless access point must be able to handle in order to meet the availability guarantee of "no more than one hour accumulated time per year shall the call centre have less than 75% its busy hour capacity[3]" with the parameters used for the simulation? Is the the number of replications sufficient to conclude with a negligible risk (e.g. less then a chance of 2.5% of making a wrong decision)? Justify the answer.

[3]The busy hour capacity is in this context $n = 24$ call agents on line.

Listing 1: Switch and wireless access simulator

```
begin
  external class demos="demos.atr";
  demos
  begin

    integer n= 24;
    integer k= 3;
    integer i,j,globalState,cumulativeObs;
    real totalSimTime,remainingSimTime,samplePeriod;
    ref(Res) Z;
    ref(Bin) array X(1:k,1:n/k), Y(1:k,1:n/k);
    ref(RDist) switchFailure,switchRepair,linkFailure,linkRepair;
    ref(CallAgent) array agent(1:n);
    real array obsIn(0:n),cumulativeRelTime(0:n);

    Entity class EthSwitch(i,X,Y);integer i;ref(Bin) array X,Y;
    begin
      integer j;
      loop:
      hold(switchFailure.sample);
      for j:=1 step 1 until n/k do X(i,j).give(1);
      Z.acquire(1);
      hold(switchRepair.sample);
      for j:=1 step 1 until n/k do Y(i,j).give(1);
      Z.release(1);
      repeat;
    end;

    Entity class Link(X,Y);ref(Bin) X,Y;
    begin
      loop:
      hold(linkFailure.sample);
      X.give(1);
      Z.acquire(1);
      hold(linkRepair.sample);
      Y.give(1);
      Z.release(1);
      repeat;
    end;

    Entity class CallAgent(X,Y);ref(Bin) X,Y;
    begin
      integer state;
      loop:
      state:=0;
      X.take(1);
      state:=1;
      Y.take(1);
      repeat;
    end;

    Z:-new Res("Z",1);
    switchFailure:-new NegExp("switchFail",2/(365.25*24));
    switchRepair:-new NegExp("switchRepair",1/3);
    linkFailure:-new NegExp("linkFail",1/(365.25*24*4));
    linkRepair:-new NegExp("linkRepair",1/48);
```

```
    for i:=1 step 1 until k do
    begin
      for j:=1 step 1 until n/k do
      begin
        X(i,j):-new Bin("Xbin",0);
        Y(i,j):-new Bin("Ybin",0);
        new Link("link",X(i,j),Y(i,j)).schedule(0.0);
        agent((i-1)*n/k+j):-new CallAgent("callAgent",X(i,j),Y(i,j));
        agent((i-1)*n/k+j).schedule(0.0);
      end;
      new EthSwitch("switch",i,X,Y).schedule(0.0);
    end;
    samplePeriod:=1;
    totalSimTime:=1000*365.25*24;
    remainingSimTime:=totalSimTime;
    while remainingSimTime>0 do
    begin
      hold(samplePeriod);
      globalState:=0;
      for i:=1 step 1 until n do globalState:=globalState+agent(i).state;
      obsIn(globalState):=obsIn(globalState)+1;
      remainingSimTime:=remainingSimTime-samplePeriod;
    end;

    for i:=0 step 1 until n do
    begin
      cumulativeObs:=cumulativeObs+obsIn(n-i);
      cumulativeRelTime(n-i):=cumulativeObs*samplePeriod/totalSimTime;
    end;

    for i:=0 step 1 until n do
    begin
      outInt(i,2);outFix(cumulativeRelTime(i),8,12);outImage;
    end;
  end;
end;
```

Norsk bokmål utgave[4]

I denne eksamensoppgaven vil vi studere et firma som driver et såkalt "call-centre" og IKT infrastrukturen til dette firmaet. Denne er illustrert i figur 1 på side 3. Firmaet betegnes QR. Det mottar VoIP samtaler fra Internett, hvor det etterspørres informasjon, assistanse og liknende. En call-centre vertsmaskin [host] håndterer disse interaksjonene, og denne må være operativ for at anropene skal bli betjent. En samtale blir etablert mellom innringeren og en såkalt "call agent" via en ruter og en svitsj. Under feilfri drift setter den tekniske delen av systemet ingen kapasitetsbegrensninger. En ruter kan føre hele trafikklasten.

Anropene kan antas å komme fra en uendelig stor mengde potensielle trafikk-kilder. Den aggregerte anropsintensiten er λ. Varigheten av samtalene er negativt eksponensialfordelt med forventning $\mu^{-1} = 180\,s$. Varighetene er uavhengige og identisk fordelte. QR har n call-agenter. Arbeidsstasjonen til en agent er knyttet til én svitsj, som igjen er knyttet til to rutere. Det er k svitsjer i systemet. (For å unngå unødig komplikasjon, antas at n/k er et heltall.) I tillegg kan arbeidsstasjonene nå en ruter via et trådløst aksesspunkt. Systemkomponentene feiler uavhengig av hverandre og feilintensitetene er konstante. Feilintensitetene er: λ_r for en ruter, λ_h for call-centre-verten, λ_s for en svitsj, λ_a for call-agentens arbeidstasjon og λ_w for det trådløse aksesspunktet. Forbindelsen mellom de ulike systemkomponentene og mellom system og Internett antas inntil videre å være feilfrie. Hvis en systemkomponent feiler, så vil alle samtalene som benyttet denne systemkomponenten gå tapt.

QR markedsfører tjenestene sine med følgende utsagn: "I travel time skal minst 99,5 % av anropene til QR nå en call-agent. Dette kravet modereres når det inntreffer utstyrsfeil. Sannsynligheten for at en etablert interaksjon mellom en innringer og en call-agent (samtale) skal bli avbrudt, skal være mindre enn $5 \cdot 10^{-4}$. QR garanterer at samlet i løpet av ett år, skal ikke call-centret ha mindre enn 75% av sin travel time kapasitet i mer en én time."

a) Definer de ovenstående QoS utsagnene fra QR som tekniske pålitelighets- og ytelsesparametre.

b) Anta at anrop til QR som ikke finner en ledig call-agent, mottar opptatt tone og kobles ned. Hva slags system er dette; bruk Kendalls notasjon? Gi et uttrykk for sannsynligheten for at et anrop ikke skal finne en ledig call-agent. Hva er den relative andelen av tiden som alle agentene er opptatt? Grunngi svarene.

For å øke utnyttelsen av call-agentene og det relative antall anrop som blir betjent, endres systemet slik at alle anrop som ikke finner en ledig call-agent forsøkes satt i kø. Idet anropet blir satt i kø får anroper talemeldingen "Du er nummer i i køen. Den antatte ventetiden er τ_i." Erfaringsdata fra andre call-centre viser at innringeren da avslutter (legger på røret) og anropet forlater systemet med en sannsynlighet q_i. For enkelhets skyld antas det at de som ikke forlater systemet med en gang blir inntil de blir betjent. Call-centre vertsmaskinen sørger for at ankomst og betjeningsrekkefølge er den samme.

c) Tegn et diagram som viser en kontinuerlig tid, diskret rom Markov modell av antall anrop/samtaler i systemet. Kall sannsynligheten for at det er j anrop/samtaler i systemet for p_j, og sett opp et sett av likninger som kan benyttes for å bestemme p_j.

[4] I tilfelle uoverensstemmelse mellom den engelske og norske utgaven, er det den engelske som er gjeldende. Engelske betegnelser anvendes hvor ingen norsk oversettelse ble funnet.

d) Anta at p_j er funnet. Sett opp uttrykk for ført trafikk, tapt trafikk, sannsynligheten for at en anroper får talemeldingen ("Du er nummer ...") og den forventede ventetid for de som må vente.

I de neste spørmålene skal ikke det trådløse aksesspunktet betraktes som en del av systemet.

e) For systemet er det et krav om at sannsynligheten for at en etablert samtale skal bli brudt pga. feil skal være mindre enn $5{\cdot}10^{-4}$. Dette kan omsettes til et krav til avbruddsintensiten ξ av pågående samtaler, $\xi \leq 0.01\text{hour}^{-1}$. Finn et uttrykk for denne intensiteten.

QRs IKT infrastruktur har en systemfeil når call-centret har mindre enn 75% av sin fulle kapasitet regnet av antall call-agenter som kan håndtere anrop. For å prediktere systemtilgjengeligheten med hensyn på dette kriteriet gjøres to forenklende antakelser: 1) Alle systemkomponentene repareres uavhengig av hverandre og midlere nedetid er antatt å være d for samtlige komponenter. 2) Det antas at arbeidsstasjonene ikke feiler, dvs. $\lambda_a = 0$.

f) Hvordan vil de forenklende antakelsene påvirke (eng: bias) prediksjonen av systemtilgjengeligheten? Begrunn svaret. Tegn en pålitelighetsmodell av systemet hvor de ovennevnte forenklende antagelsene er benyttet og finn et uttrykk for systemtilgjengeligheten A_s.

Anta at den ovenstående analysen viser at tilgjengelighetskravet ikke blir møtt. Istedet for å gjennomføre kostbare tiltak med anskaffelse av flere svitsjer og re-kabling av aksessen til arbeidsstasjonene osv., introduserer QR et trådløst aksesspunkt som kan nås fra alle arbeidsstasjoner, som vist i figur 1 på side 3. Aksesspunktet kan håndtere maksimalt k_w arbeidsstasjoner samtidig.

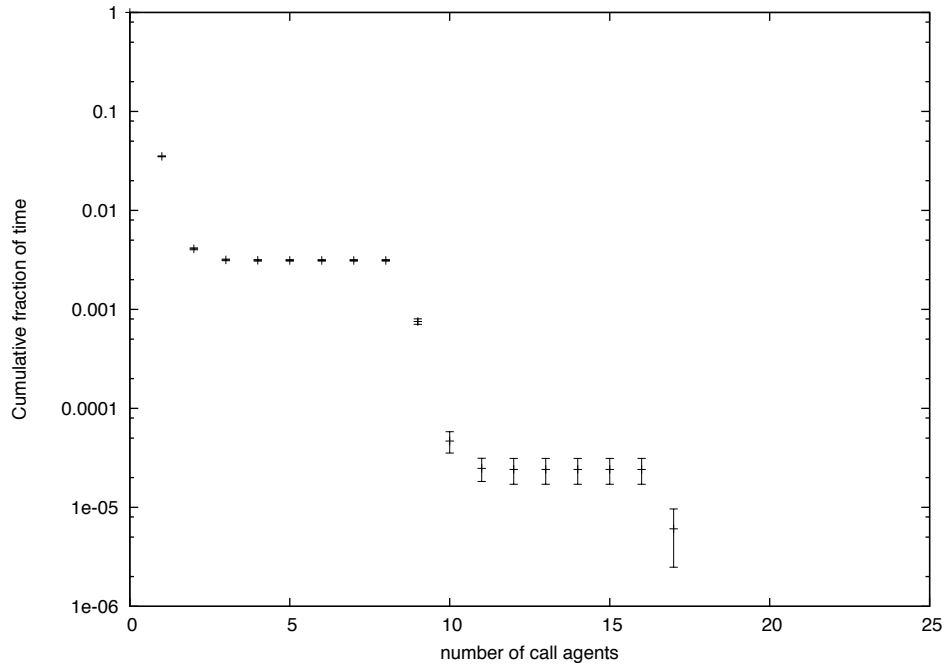
g) Hvis $k_w = n/k$, $\lambda_w = \lambda_s$ og de forenklende antakelsene ovenfor kan benyttes, hvordan kan modellen som ble funnet i spørsmål f) modifiseres til å inkludere det trådløse aksesspunktet?

Den ovenstående tilnærmingen gir ikke tilstrekkelig nøyaktighet. Betrakt subsystemet bestående av svitsjene og det trådløse aksesspunktet. Anvend at $\lambda_w \neq \lambda_s$, at svitsjene og det trådløse aksesspunktet har ulike reparasjonstider, i.e. $d_w \neq d_s$, og at maksimalt en systemkomponent repareres ad gangen. Reparasjonstidene kan antas å være negativt eksponensialfordelte.

h) Lag et tilstandsdiagram (Markov modell) av svitsjene og det trådløse aksesspunktet når $k = 3$, som kan benyttes til å finne tilgjengeligheten til dette subsystemet med hensyn på systemfeilkravet. Definer tydelig hver tilstand og angi hvilke tilstander som representerer et hhv. arbeidende og feilet system.

Antakelsene som må gjøres er forsatt for grove. Det maksimale antallet arbeidsstasjoner som kan håndteres av det trådløse aksesspunktet er forskjellig fra n/k. Lenkene mellom arbeidsstasjonene og svitsjene vil også feile med en intensitet $\lambda_l > 0$. For å fordele arbeidsbelastningen mellom call-agentene, vil en arbeidsstasjon frigjøre sin tilknytning til det trådløse aksesspunktet etter at åtte anrop er håndtert. Anta at tiden fra en call-agent blir ledig for anrop til han/hun får et anrop er en tilfeldig fordelt tid `timeToCall`.

i) Anta at et spesifikt antall $m > k_w$ arbeidsstasjoner bruker det trådløse aksesspunktet. Anta også at det trådløse aksesspunktet ikke feiler. Lag en simuleringsmodell for å undersøke den tiden en arbeidsstasjon *ikke* er tilknyttet det trådløse nettet og som en konsekvens av dette, call-agenten ikke kan nås. Tegn et aktivitetsdiagram for modellen. (Merk at i denne deloppgaven skal kun bruken av det trådløse aksesspunktet underesøkes.)



Figur 3: Den kumulative andelen av tiden minst n_w call-agenters arbeidsstasjoner trenger å bruke det trådløse aksesspunktet.

Den fullstendige koden for call-agentene som bruker svitsjene og det trådløse aksesspunktet er implementert i Simula/DEMOS og vist i listing 1 som starter på side 6. Svitsjene, og lenkene mellom arbeidsstasjonene og svitsjene kan feile. Selve arbeidsstasjonene og det trådløse aksesspunktet antas ikke å feile. Arbeidsstasjonene er indeksert i, j, hvor $i = 1, \ldots, k$ representerer svitsjnummeret og $j = 1, \ldots, n/k$ representerer arbeidsstasjonens nummer på en spesifikk svitsj. Enheten for simuleringstid er en time [hour].

j) Hva er entititene som er definert i simulatoren? Hva er hensikten med `Z` ressursen? Hva representerer `state` attributten til `CallAgent`en og hva representerer `X` og `Y` ressursene? Hva er den globale tilstanden i systemet definert til å være, dvs. hva representerer variabelen `globalState`? (Merk at det *ikke kreves* at det lages et aktivitetsdiagram for simulatoren.)

Grafen i figur 3 viser den kumulative andelen av tiden minst n_w call-agenters arbeidsstasjoner trenger å bruke det trådløse aksesspunktet gitt med 95%-konfidens intervall. Feilintensitetene, reparasjonstidene og antall call-agenter er de samme som ble brukt i Simula/DEMOS koden ovenfor, dvs. $\lambda_s = 2\,\text{aar}^{-1}$, $\lambda_l = 1\,aar^{-1}$, $d_s = 3\,\text{timer}$, $d_l = 48\,\text{timer}$, $n = 24$ og $k = 3$. Konfidensintervallene er funnet ved å kjøre ti replikeringer av simuleringen.

k) Gi en fortolkning av formen på grafen relatert til strukturen til systemet. Hva er det minimale antall arbeidsstasjoner k_w det trådløse aksesspunktet må være i stand til å håndtere for å imøtekomme tilgjengelighetsgarantien på "QR garanterer at samlet i løpet av ett år, skal ikke call-centret ha mindre enn 75% av sin travel time kapasitet i mer en én time[5]" med de parametrene som ble brukt i simuleringen? Er antall replikasjoner tilstrekkelig til å kunne konkludere med en neglisjerbar risiko (f.eks. mindre enne en sjanse på 2.5% for å trekke en gal konklusjon)? Begrunn svaret.

[5]Travel time kapasitet er i denne sammenhengen $n = 24$ call-agenter som kan håndtere anrop.