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 12.

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.

Dependability:

Working criteria is that at least $\lceil 0.75 \cdot n \rceil$ of the call agents shall be accessible.

The availability is: $A \geq 1 - \frac{1}{365.24 \cdot 24} = 0,99989$.

After the removal of the request of finding the spesific call interuption rate ξ is removed, a sufficient answer is: For an eteblishe connection/call we hav that $R(T_{\text{call}}) > 1 - 5 \cdot 10^{-4}$ where T_{call} is the duration of a call; or better $\int_0^\infty R(t)\mu e^{-\mu t} dt > 1 - 5 \cdot 10^{-4}$ which caluclated yields the original answer idetenical to the one below

The service reliability is given by a call completion probability of $1 - 5 \cdot 10^{-4}$. Since service/call completion and interruptions are competing Poisson processes, the completion probability is $1 - 5 \cdot 10^{-4} \leq \frac{\mu}{\mu + \xi}$ and hence the interruption intensity is $\xi \leq \frac{\mu}{1999} = 1/(359820\,s) \approx 1/(100\,h)$.

Performance:

The call congestion $B \leq 1 - 0.995 = 5\,10^{-3}$.

Missing QoS requirements/statements include:

[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.

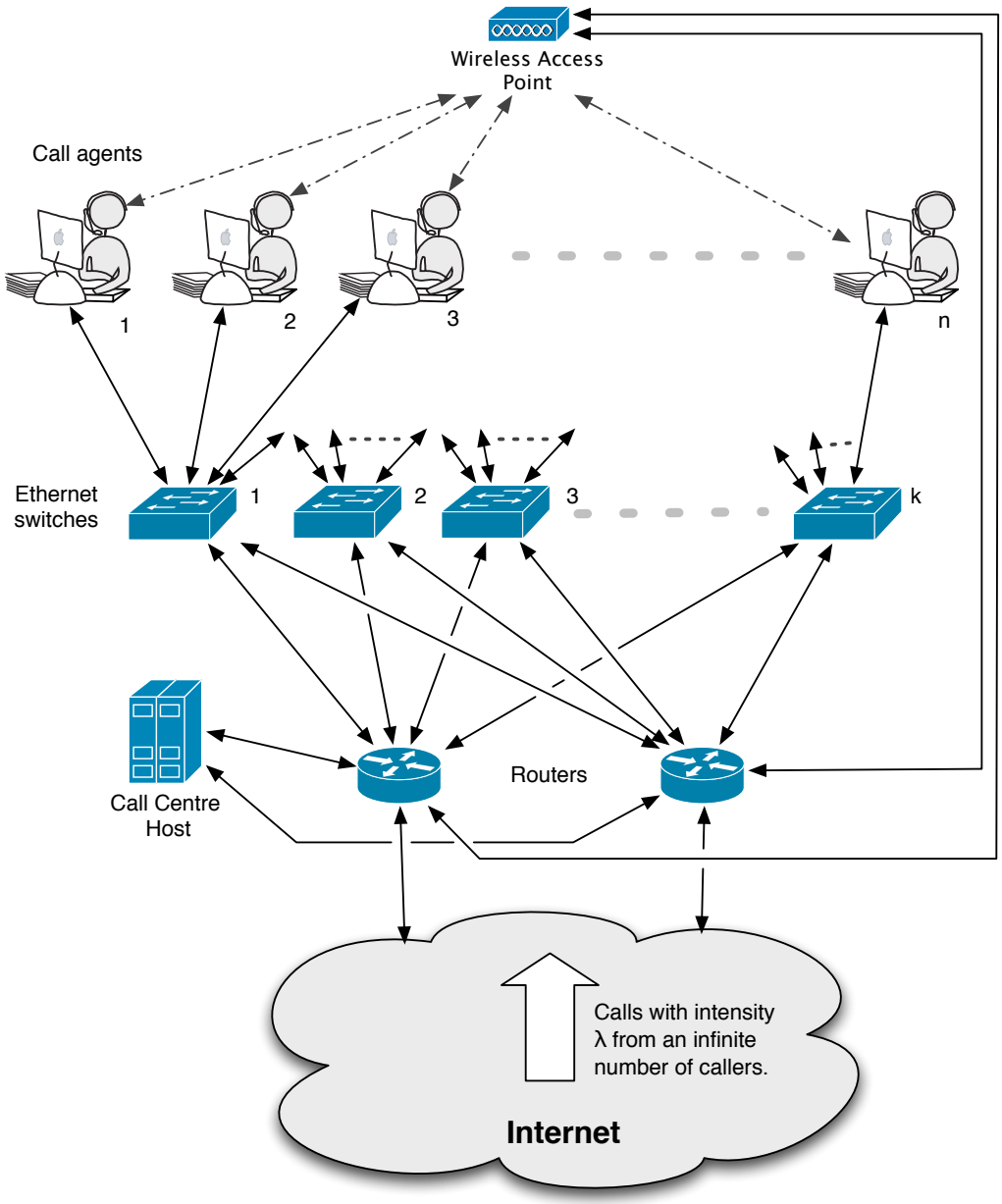


Figure 1: Illustration of the ICT infrastructure of QR, including the call agents.

- The delay before a call centre agent responds

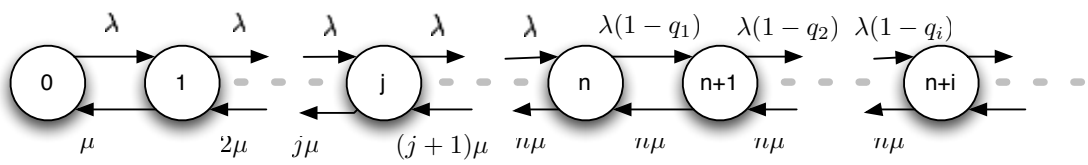
- The voice quality or parameters that influence the voice quality like delay, jitter and packet loss [This may also depend on the service provided by QR's ISP].

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.

Due to the infinite number of potential callers the arrival process is Poisson. The server process is given to be Poisson. There are n servers and no queue, hence we have a M/M/n/n system, i.e. an Erlang's loss system. The blocking probability is given by Erlang's B-formula $E_n(A) = \frac{A^n/n!}{\sum_{\nu=0}^{n} A^\nu/\nu!}$, where $A = \lambda/\mu$. For an Erlang's loss system, call congestion and time congestion are equal, and all agents are busy $E_n(A)$ of the time.

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.



$p_{j-1}\lambda = p_j j\mu,\ j = 1, \ldots, n,\ p_j\lambda(1 - q_{j-n+1}) = p_{j+1}n\mu,\ j = n, \ldots, \infty$ and $\sum_{j=0}^{\infty} p_j = 1$.

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.

Carried traffic: $A' = \sum_{j=0}^{\infty} \min(j, n)p_j$.

Lost traffic: $A'' = \lambda/\mu - A'$.

Probability the a caller gets a voice message: $P(\text{all agents busy}) = \mathrm{P_W} = \sum_{j=n}^{\infty} p_j$

Expected waiting time for those who wait: $P_W^{-1} \sum_{j=n}^{\infty} p_j(j-n+1)/(n\mu)$ (Probability of finding j in the system on arrival given that a caller has to queue $P_W^{-1}p_j$. The number of completions necessary before service may start $j - n + 1$. The expected time between completions $(n\mu)^{-1}$.)

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.

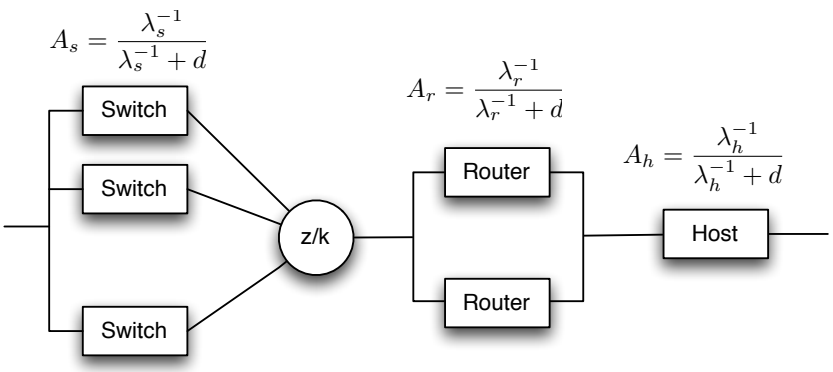
All the used system components for a reliability series structure since established "connections" are lost if affected by failure. Hence: $\xi = \lambda_r + \lambda_h + \lambda_s + \lambda_a$.

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.

Disregarding potential failures gives a too optimistic prediction of the availability. Assuming independent repair increases the repair resources in multiple failure situations and yields also too optimistic results.

For the system to be non-failed, at least 75% of the agents must be accessible. Hence, the number of switches that must be working is z, where $z\frac{n}{k} \geq 0.75\,n$, i.e. $z = \lceil 0.75\,k \rceil$.



$A_S = A_h(1 - (1 - A_r)^2) \sum_{i=z}^{k} A_s^i (1 - A_s)^{k-i}$

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.

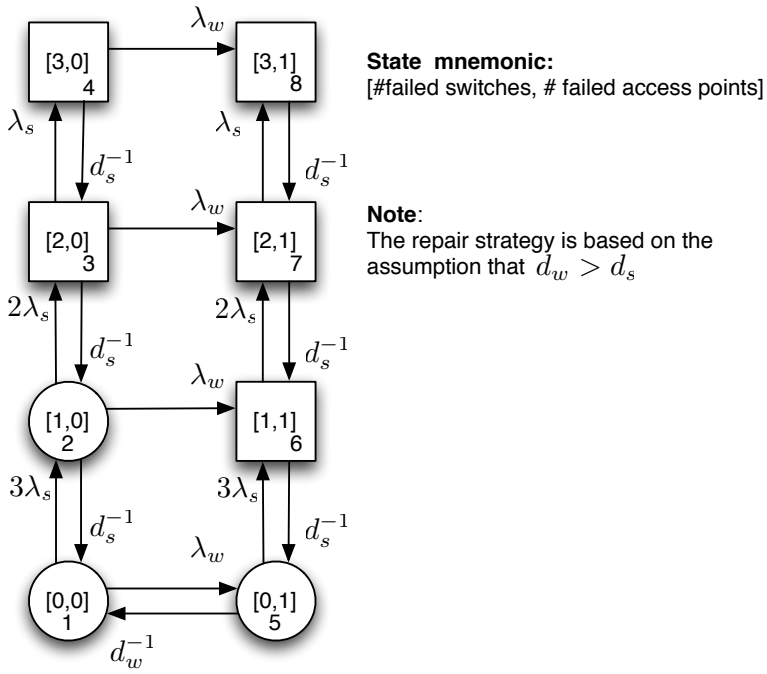
g) If $k_w = n/k$, $\lambda_w = \lambda_s$ and the simplifying assumptions may be used, how may the model found in question f be modified to include the wireless access point?

We add one more switch to the system and one more may fail before the system fails. Hence the "z-out-of k" structure should be replaced by a "z out-of (k+1)" structure.

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.

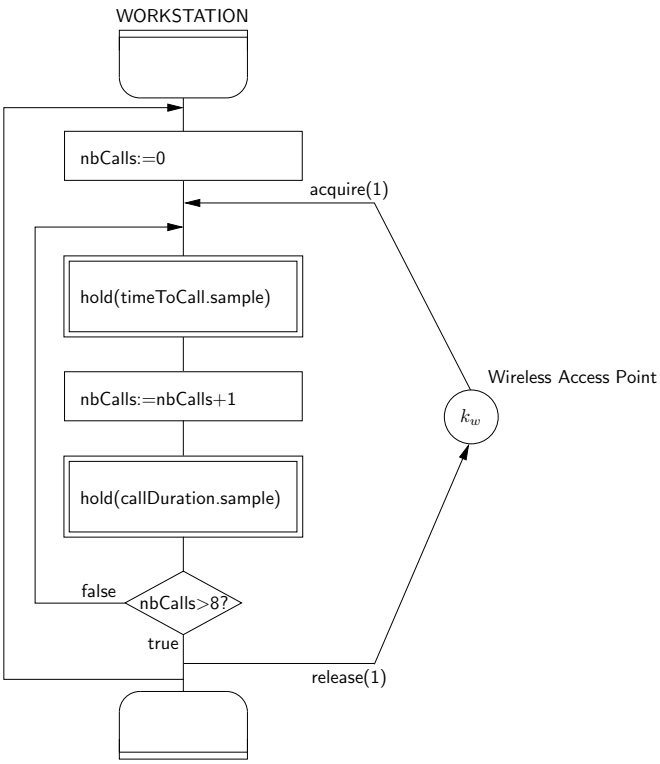
Note that at most one repair can take place at a time.



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.)

We model the workstations as entities and the wireless access point as k_w resources (type res, mutual exclusion). The workstation is unreachable while waiting for a resource.



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 10. 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 switches, links and call agents are entities

Z resource type: res. Purpose: mutual exclusion. Model the repairman and ensures that no more than one system component is repaired at a time. state = 0 when a workstation is in normal operation using the switch and awaiting a failure. The resource X or X(i,j) represents the link and switch failures affecting the access of the workstation. In state = 1 the workstation tries to use the wireless access point awaiting the repair of the failure(s) affecting the workstation [3]. Y represents a completed repair.

The variable`globalState` represents the number of workstations that tries to use the wireless access.

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. $\lambda_s = 2\,\text{year}^{-1}$, $\lambda_l = 1\,\text{year}^{-1}$,

[3] Note the we do not include in the model whether the workstation gets a wireless access or not. The model of question i may be merged with the current to achieve that.

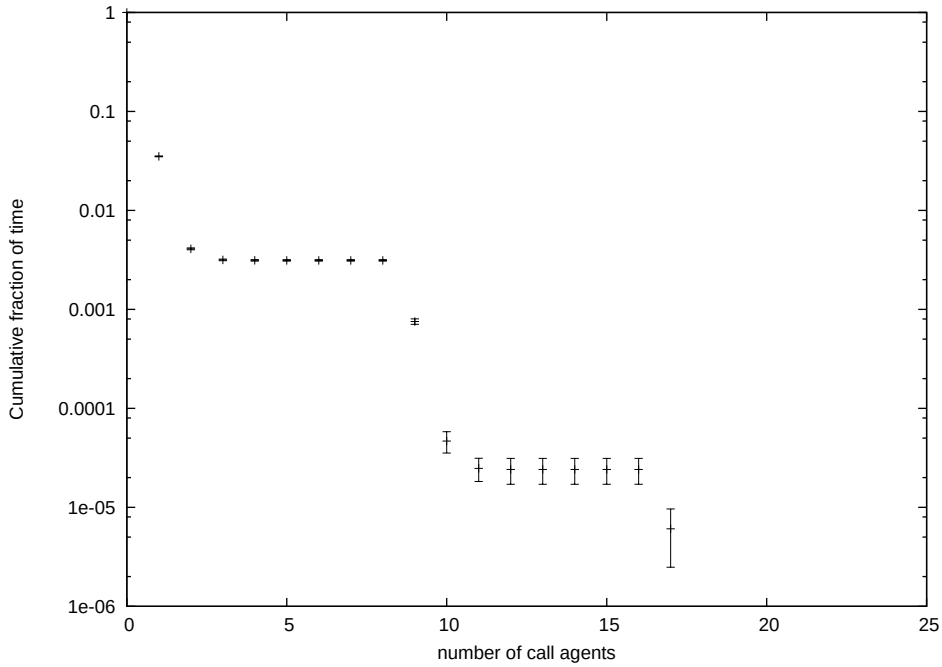


Figure 2: The cumulative fraction of time at least n_w call agent's workstations need to use the wireless access point.

$d_s = 3$ hours, $d_l = 48$ 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[4]" 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.

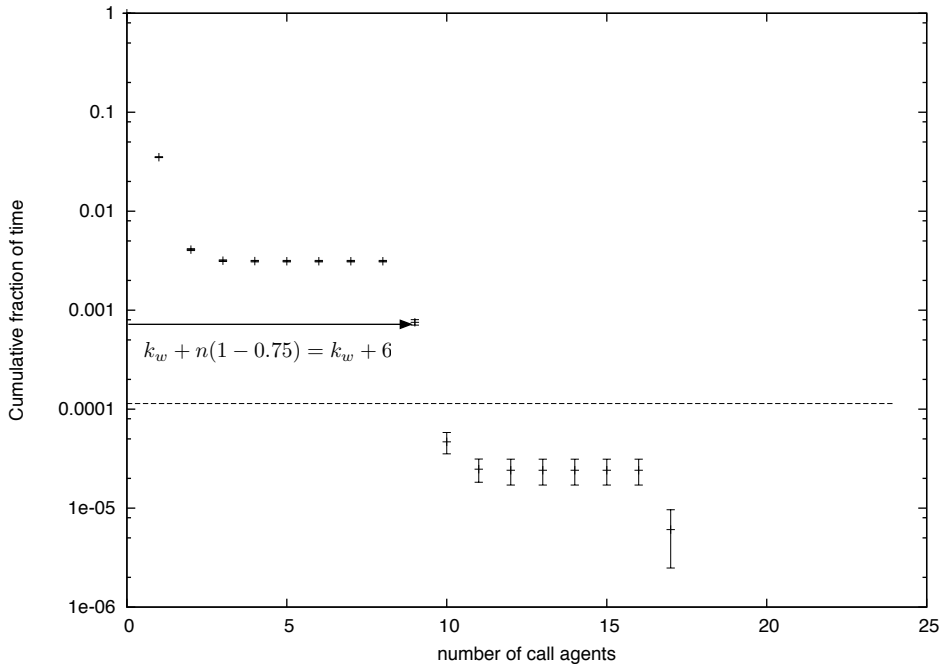
The probability that more than one of the system's component are failed simultaneously is small but not negligible. When a link fails, the workstation connected to a switch via that link needs to use the wireless access point. The left part of the curve shows primarily the contribution from link failures of one an two links. When a switch fails, all the workstations connected to the switch, (here 8 workstations) need to use the wireless access point. This yield the flat part of the curve up to 8. The next point represent one switch failure and on link failure, etc. Next we see the effect of two simultaneous switch failures, etc.

The availability guarantee is that no more than one hour per year the call centre have less than 75% of the capacity available, that is when $n = 24$ less than 18 call agents are available or no more than 6 call agents are unavailable. The minimal number of workstations the wireless access point must be able to handle, k_w, so that the availability guarantee is achieved, is the minimal number such that the relative time k_w+7 or more call agents are unavailable is less than or equal to one hour per year is: $(365.25 \cdot 24)^{-1} = 1.14 \cdot 10^{-4}$; and correspondingly$k_w$+6 are above this requirement. Reading the graph , we

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

find: $k_w = 10 - 7 = 3$. *Note, $k_w = 3$ is sufficient as a correct answer.*

The "decision line" $(365.25 \cdot 24)^{-1} = 1.14 \cdot 10^{-4}$ is not covered by any 95% confidence intervals. Hence, it is a low risk involved in concluding the number of channels needed.



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;
```