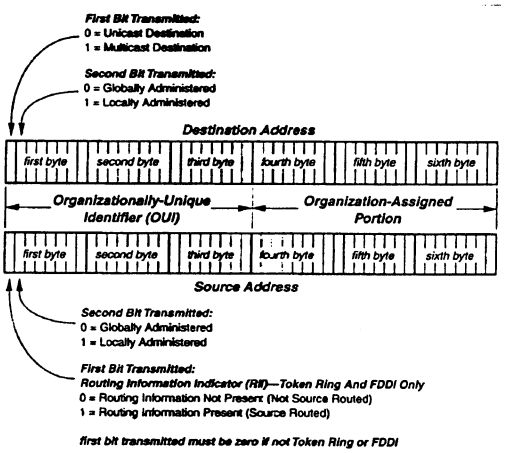
Q1 answer

a) The key to MAC layer protocols is a uniform addressing structure so that LAN hardware can easily identify stations on the network and transmit packets between them. The address can identify both the transmitting station (*source address*) and the receiving station (*destination address*) and are usually referred to as the *MAC address* of the station on the LAN.



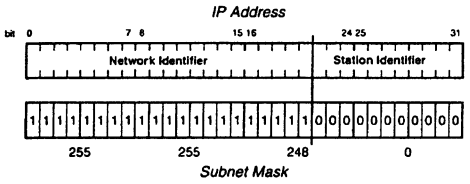
The 48bitMAC address space is split into two halves:

- A *unicast address* identifies a single device or network interface. Often referred to as *individual, physical* or *hardware addresses.*
- A *multicast address* identifies a group of logically related devices. Often referred to as *group* or *logical addresses.*

The first bit of the destination address defines if it is a unicast (0) or a multicast (1) address. Source addresses are always unicast so the first bit is zero except in token ring or FDDI, where is describes how the packet is routed. The second bit of the address defines whether the address is globally unique (0) or locally unique (1) to the LAN. Globally unique addresses are assigned by the manufacturer. A 48-bit address allows for approx 281 million million addresses.

The 48-bit MAC address is divided into two parts. The first 24 bits constitute the organisationally unique identifier (OUI), which indicates which organisation (typically the manufacturer) is responsible for assigning the remaining 24 bits of the address. A MAC address is usually expressed in canonical byte form *aa-bb-cc-dd-ee-ff* where the first 3 pairs of bytes are the OUI and the last 3 pairs of bytes are set by the manufacturer.

IP addresses are 32bit long, fixed-length fields which comprise two portions:



- ✓ The network identifier, which indicates the network on which the addressed station resides.
- ✓ The station identifier, which denotes the individual station within the network to which the address refers. IP station identifiers are locally unique, only being meaningful in the context of the identified network.

Each IP address has an associated subnet mask of the same length (32 bits). The network identifier portion of the address is defined by the portion of the subnet mask set to 1's. The rest is the station identifier. The convention is that the network bits and the station bits are set in a contiguous fashion. This is a CONVENTION which is broken at the risk of being rapidly ostracised by your internetworking peers. The contiguous subnet can be stored as a 5bit word, which greatly simplifies the router look up process.

b) When a bridge receives a MAC address is must locate the relevant port mapping from the bridge address table. In order for a bridge to locate a port to which a destination station is attached it must search for it in the bridge address table. This should be as efficient as possible to minimise potential delays in the catenet. One of the most efficient ways in which an address location can be found is to use the 48 bit address itself as a pointer to the location of the required port allocation information. However, a 48bit address implies at least 2^{47} required memory locations, which is prohibitively large.

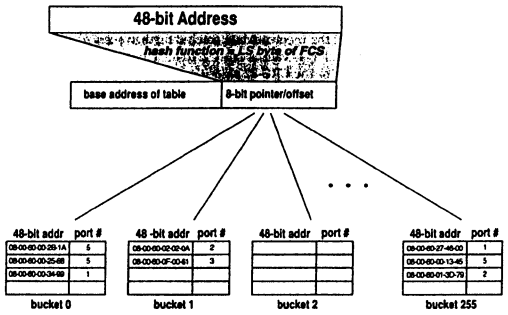
It should be noted that the routing look up process is considerably more complex than that for a bridge, as there is a complex address structure to be broken down into a pair of variable fields. Hence the search algorithms are considerably more 'intelligent'. Unlike in a bridge, where the search is for a fixed length address field, the IP network identifier has a variable length from 0 (usually specifying a default route) to 32 bits (for host-specific routes). A given destination address may yield multiple matches with entries in the router table corresponding to different network identifier lengths. The IP routing rules

specify that the routing result returned should be for the one for the largest number of matching bits. Hence the look up process implies a search for the longest match against a variable length field.

There are appropriate algorithms for such searches including many based on a compressed binary tree (such as a radix or PATRICIA tree) data structure. In addition, a binary tree can be incorporated with the ARP cache mapping process. In some structures, the entire layer 2 addressing table can also be included. Router look ups are often implemented as hardware state machines using complex data structures and content addressable memory.

c) There are two main techniques: The hash table and content addressable memory (CAM).
A way to simplify the search process is to map the address onto a small pointer space, using what is known as a hash function. Hashing is taking the 48bit address and producing a shorter field, which will then point at a subset of locations in memory. The hash function must:

- ✓ Produce a consistent value for the same address (be repeatable)
- ✓ Produce a relatively uniform distribution of pointers for a given set of inputs.



There are many different hash functions in the literature, however there is an automatic generation algorithm, which will perform the hash function on the data in the form of the FCS generator and checking circuitry. As the frame passes through the FCS circuit, there is a point when the 48bit destination address has passed through it, hence we only need to sample the output from the FCS at this point in time to generate a suitable hashed address. This is usually done with a linear feedback shift register (LFSR). The address table is organised as hash buckets, with the hash function (the bits from the LFSR) pointing to the appropriate bucket.

The original idea of using the address as a pointer to the table entry is used in CAM structures, which offer very fast look up speeds, however they are very limited in size due to the space needed to layout the require circuitry and they are 4-5 times more expensive that conventional memory structures. CAMs can store 32000 addresses up to 64 bits wide and have access times of nsec.

The hash function does not help simplify an IP address look-up as it is a variable field length search, however CAMs can be used to form the basic structure in an IP look-up state machine

d) Mapping the destination to a local data-link address (ARP mapping) – The structure of the station identifier section of the IP address does not provide a simple mapping onto 48 bit data link addresses. It is not possible to determine the 48 bit data link address solely from the station identifier. Hence for packets destined for a locally connect network (such as the last hop port) must undergo a second look up process to find the destination station. In some instances this could be a separate look up operation in the ARP cache or continuation of the router look up process. Either way it will comprise one of three classes:

1. The packet is destined for the router itself. ie the destination IP address corresponds to one of the IP addresses of the router. This packet will be passed the higher layer entities in the router.
2. The ARP mapping for the indicated station is unknown. In this case, the router must initiate a discovery procedure (ARP request). This may take some time (it is a form of layer 3 flooding) and could result in the packet being dropped. This is not part of the fast path. ARP requests are not normally part of the steady state operation of a router.
3. The station is destined for a known station on the directly attached network. In this, very common case, the router successfully determines the mapping from the ARP cache and continues the routing process.

e) subnet 255.255.242.0 has 22 contiguous ones and 10 zero. No of stations = 2^{10} = 1024
This is a very limited number of stations for a modern LANs and catenets today. There are two possible solutions.

1. Go to IP version 6 which has a larger address space (64bits) giving more freedom. This is difficult to implement in reality due to the legacy of IPv4 and the complexity of IPv6

2. Make the ARP resolution process discover the station each time. When a packet is received, the router floods the LAN with packets an waits for the station with the matching IP address (and therefore MAC address to respond to the flood. This is Ok as the flood is only within the LAN and does not threaten congestion. Once the response is received, it then uses the attached MAC address to send the packet to the final destination station. This mapping will be remembered for a short time in case other packets in the sequence arrive, but will eventually time out and have to be learnt again.

Q2 answer

a) A bridge is a piece of OSI layer 2 hardware, which allows frames to be passed between LANs that have different geographical locations and even different LAN protocols (MACs). A bridge is a data-link layer device and any interconnection of LANs via bridges is often referred to as a catenet. Some basic bridge principles:

- Each station has a globally unique 48 bit unicast address.
- There is a table within the bridge which maps station addresses to bridge ports.
- The bridge acts in promiscuous mode, it receives (or attempts to) every frame on every port.

When a frame is received on any port, the bridge extracts the destination address from the frame, looks it up in the table and determines the port to which the address maps. If the looked up port from the table is the same as the one on which the frame arrived then the frame is discarded, as it assumes that the station on that port will have already received the intended frame (filtering). If the frame received by the bridge is mapped to a different port than the one it arrived on, it then forwards the frame to the port and onto the appropriate LAN.

A bridge tries to make a catenet appear transparent to end stations, as if it were a single LAN. Hence higher layer services will expect a LAN-like performance from the catenet below it. A LAN data-link must exhibit certain properties. Hard Invariants, Non duplication of frames, Sequential delivery of frames. These invariants and the complexity of the address table limit the scaling on the catenet, hence a bridge is not a suitable device to maintain a global network such as the internet.

As with the technological evolution of the bridge into the switch, the advance of technology has allowed the use of network layer routing or layer 3 switching to be implemented as part of a LAN. Using modern silicon integration and application specific ICs (ASICs) it is now possible to build routers for similar costs as layer 2 devices. Wire speed devices now marketed as 'layer 3 switches' and the difference between layer 2 and 3 switching depends on the needs of the station and administration. There are a whole host of layer 3 protocols to choose from including IP, IPX, DECnet, Phase IV, AppleTalk and the OSI CLNP, however the majority of networks and network vendors have migrated towards the internet protocol (IP) as the preferred protocol of choice. Hence most routers are based on IP, with a few offering limited IPX functionality. The global acceptance of the layer 3 IP address has meant that it is an ideal method for making a global network structure.

The router is not hampered by the restrictions of hard invariants, hence the layer 3 functionality allows it to learn network structures and routing paths as well as optimise its operation through a whole host of criterion such as latency, number of hops or error rate. The cost of this is in operational complexity.

b) The address table is built automatically by considering the source address of frames received by the bridge. The bridge will look up the destination address in order to assign a port, and at the same time it will look up the source address to see if it has ever heard from that particular station before. In an entry is not found for the source address in the table, then a new entry is created. If there is already an entry, then it is updated to the port from which the frame was originally received. Over time the bridge will learn a port mapping for all active stations on the LANs.

If a bridge only ever learned address to port mappings, then two problems would occur: If old entries are not removed from the table, then its size will increase and will eventually take too long to search through. If a station moves from one port to another, then frames will be sent to the incorrect port until the moved station decides to transmit itself. This could take forever with a poorly designed upper layer structure.

The simple solution to both these issues is to age entries on the address table until they become stale entries, which have expired and are removed from the table. The definition of activity is based on appearance of the source address only. In a typical bridge address table, there will be a series of bits stored, which indicate the age, and status of a table entry. The commonest way is to use 3 bits, the valid bit (V) and the hit bit (H) dictate how old the entry is. The V bit indicates if an entry is currently valid in the table, and the H bit indicates that a source address has appeared within the last ageing cycle. The typical length of the ageing cycle is around 300 seconds (5 minutes). Some tables use a third bit as a static bit to indicate an address entry which cannot be aged or changed.

A router undergoes a similar process to prevent stagnation of its table. It will either periodically download routes from other sites or time out in active entries from its table.

c) The spanning tree protocol uses accelerated table ageing to prevent instability in its structure, especially when events such a link failure lead to changes in the STO topology. If the topology changes it is usually due to component or link failure, management intervention or network evolution. STP treats these as boundary events and can be slow to react to changes. Hence a smoothing procedure is employed to prevent possible shocks.

- Provides for explicit topology change and notification.
- Provides for acknowledgement of the changes.
- Uses timers to:

1. Prevent rapid transition between blocking and forwarding states. Hence minimise transient loops
2. Allow bridges to participate in the election of new designated bridges or ports. This prevents recursive changes

Before transitioning to the forwarding state as a result of a topology change, a port will wait for the topology change information to propagate through the catenet. Weird things can happen when a topology changes, hence when the topology bit is set in a BPDU the ageing timers for address tables are set to a much shorter duration, which means they will purge old values much quicker. This is important because, when topologies change, huge chunks of addresses will appear to shift from port to port.

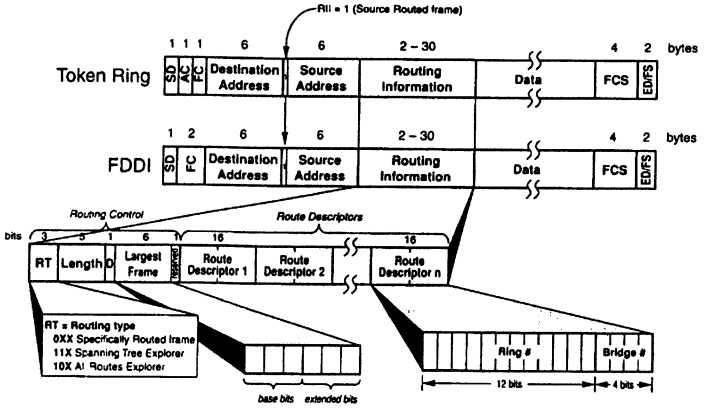
d) The roles of the station and bridge are completely reversed in source routing when compared to the operation of transparent bridges. The end stations are totally responsible for the transportation of traffic across the catenet and the bridges are totally unaware. Essentially source routing is a connection oriented process, with a call setup procedure followed by data transmission. The chosen path is then used for the entire transmission process as if it were connected via a virtual circuit. Once the data transmission has been completed, the path is closed and no record of the route is kept by either end station. For some classes of source routed traffic, there is a need to be able to send frames without first setting up a virtual circuit and establishing a path through the network.

- Multicast frame – It is not feasible to set up a virtual channel with a single source and multiple destinations. Multicast frames are generally connectionless.
- Route discovery – We need to send frames without duplication in order to set up the source route.

In order to do this we set up a loop free path through the topology which includes all of the rings in the catenet. Effectively a spanning tree. In source routing, a frame contains a list of all the rings and bridges that it must traverse to reach its destination. In order to do this each ring and bridge must have some form of identification. Each ring has a unique 12bit number (giving 4096 rings in a single catenet). Bridges are give a 4bit number from 1 to 15 (0 is reserved).

Each end station in the catenet connection must learn all about the available paths in the catenet through the process of route discovery. There may be multiple paths through a catenet and the stations must select the route to use. Normally both stations will use the same route forward and in reverse, with the route being stored in cache until the connection is over. During route discovery, the station may learn that the end station is on the same ring in which case normal token procedure is invoked and source routing not used. During the discovery process the end stations will also learn the maximum transmission unit (MTU) which can be handled across the path, which will then set the MTU used in data transmission.

Source routing inserts extra fields into the MAC frame, hence it is necessary to indicate presence of a source routed frame. This is done by using the first bit of the 48bit source address in the MAC frame, which normally defines unicast or multicast address format. This bit is normally wasted as multicast source address is meaningless, hence it is used to indicate (set to 1) a source routed frame.

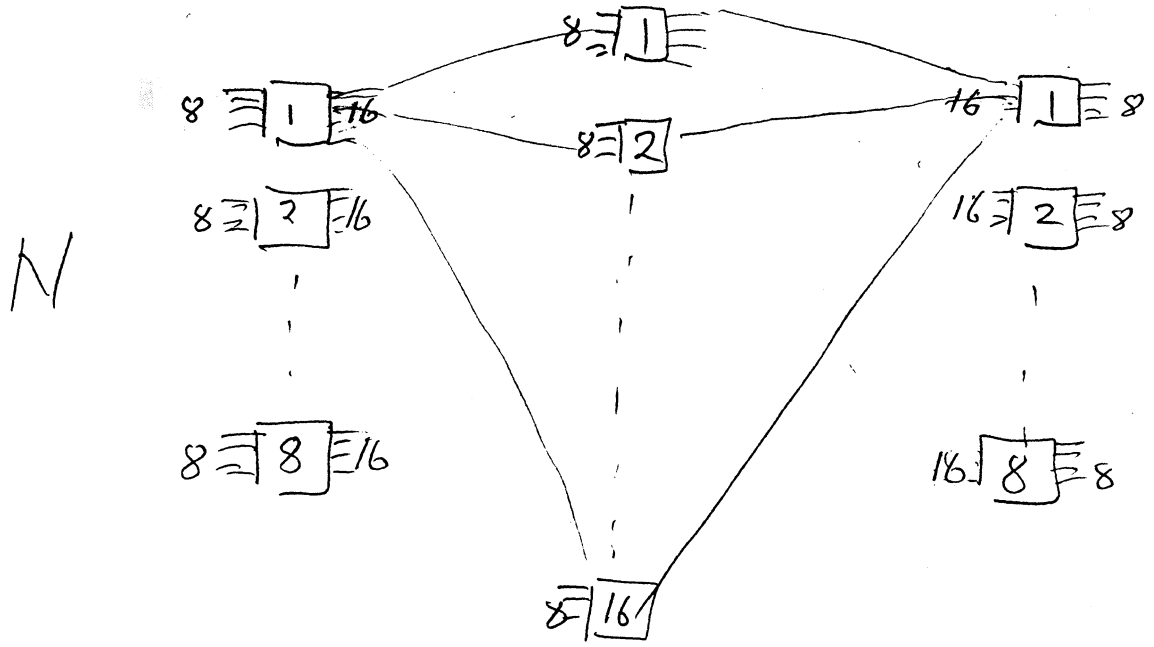


e) The use of a address table in source routing is quite different as there is no need to have a look-up table in the strict sense, just a mapping of the nearest links which lead to other networks. Hence in the pure form, ageing would not be helpful in a source routed sytem as it effectively ages itself when each route is discovered.

If the source routing system uses a spanning tree, then an ageing process could help maintain the tree when topology changes occur. This assumes that an automatic from of STP is used in the source routing system through the broadcast of BPDUs. This would work as long as topology changes were relatively rare. This is the case with source routing as it was originally based on ring type MAC LAN structures.

4815 ~~8~~)

Q3



Input stage : r crossbars of dimensions n x m r=8, n=8, m=16

Output stage : r crossbars of dimensions m x n r=8 n=8 m=16

Centre stage : m crossbars of dimensions r x r.

For strictly non-blocking $m > 2n-1$ i.e. $16 > 2\times8-1$

Crosspoint count

Input stage = 8 x 8 x 16 = 1024

Output stage = 8 x 8 x 16 = 1024

Centre stage = 8 x 8 x 16 = 1024

3072

A full crossbar would have 64 x 64 crosspoints

= 4096

ie Clos save 1024 crosspoints

A Clos multistage network lowers the number of cross points required.

A disadvantage is that it increases the complexity of

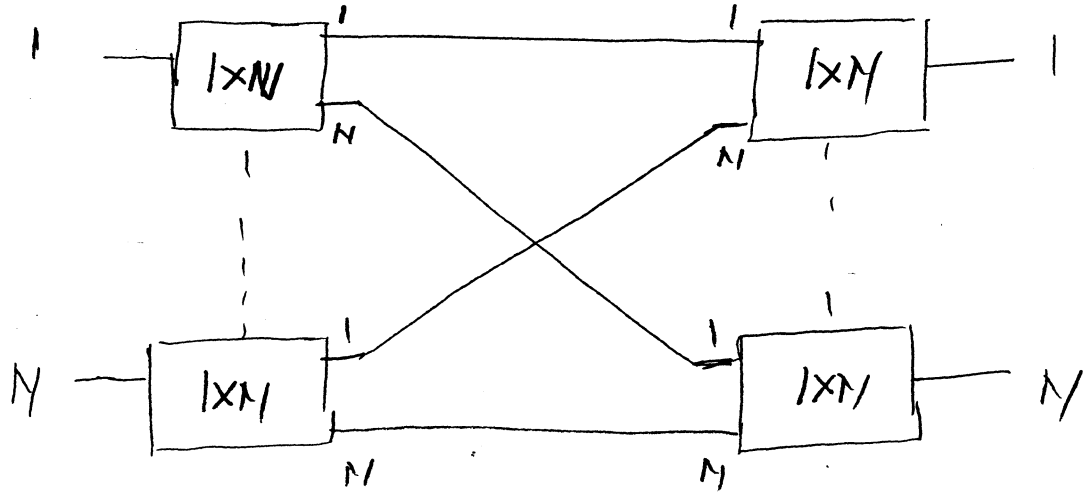
control and it may increase the latency.

b) In the synchronous digital hierarchy (SDH) add-drop multiplexers and cross-connects are the key components of the transport layer. They are often connected in a ring topology with the cross-connects linking the rings.

The synchronous cross-connects manage and reconfigure the network by setting up semi-permanent interconnections at a virtual container (VC) level.

Add-drop multiplexers allow channels to be added or dropped at nodes or stations. They are designed to allow a wide variety of tributaries; synchronous, plesiochronous, from LANs or from ATM networks. They may be duplicated either by traditional 1+1 protection or "east-west" mode for ring loops

c) The 'router-selector' architecture has been proposed for making scalable optical switches operating between single mode fibres.



2N beam deflectors are required for an N×N switch

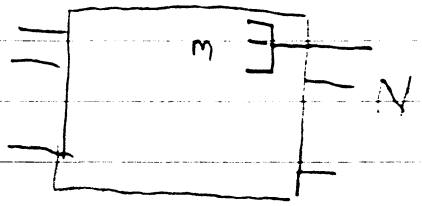
ie if N = 64

128 beam deflectors give 64×64 connectivity

Question 4

a) In circuit switching the benchmark of full connectivity (strict sense non-blocking) is crossbar connectivity in which the number of crosspoints (and therefore connection paths) scales as N^2, where N is the number of transmitters and receivers.
The connections change at the message rate.
'Arbitration' is carried out to prevent multiple inputs from connecting to a single output.
In packet switching no such arbitration is possible, and multiple inputs may send to a single output port. As a consequence packet switches would have to scale as N^3 if there was to be no packet loss. This is not practical, so some statistical packet loss must be accepted, (perhaps 1 in 10^{10}).
Packet switch connectivity must change at the packet rate, i.e it must change much faster than circuit switches

b) Consider a fully loaded, fully available $N \times N$ circuit switch:



Let all input packets or cells be copied to all output ports. However, in given time slot only m ($m \ll N$) are accepted. All other cells are lost.
Assume that the probability that K packets arrive destined for a particular output (for large N) is given by the Poisson distribution, i.e. the packets have an independent, uniform address distribution.
This probability therefore is $a(K)_{N\to\infty} = \frac{\rho^K}{K!} e^{-\rho}$

where ρ = the probability that a packet arrives in a given time slot.

The mean fraction of cells that make it to the output is given by

$$p_{av}(N\to\infty) = \left[\frac{pe^{-p} + \frac{p^2e^{-p}}{2!} + \frac{p^3e^{-p}}{3!} + \dots}{p}\right] = \frac{1-e^{-p}}{p}$$

(obtained by summing the probabilities that 1, 2, 3 --- cells will arrive)

Assuming a fully loaded, fully non-blocking switch, $p=1$
$\therefore$ 63.2% of input traffic gets through
36.8% is lost

ie a very unacceptable packet loss.

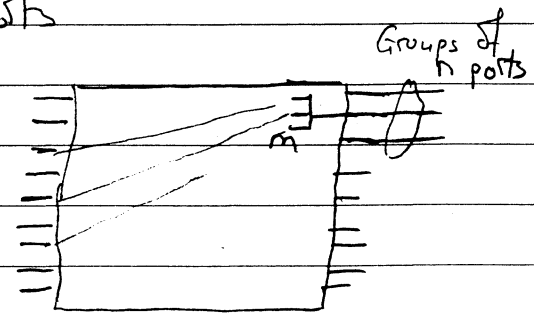
c) Data traffic tends to have very different characteristics to "human related" processes and so does not follow Poisson statistics. Packet traces tend to "bursty" or self-similar. The arrival statistics are sometimes said to be "fractal-like". Aggregation of this kind of traffic does not smooth out the statistics (as might reasonably be expected) but intensifies the self-similarity. Despite this — to a first approximation — the results obtained by assuming Poisson statistics are a useful guide.

d) Increasing the number of paths through the switch decreases the rate of packet loss below that stated in b). This is usually referred to as the speed-up, m.
On the basis of Poisson statistics setting $m \simeq 12$ for a 90% loaded switch ($p = 0.9$) gives a cell loss probability of 10^{-10}. This applies over a

wide range of N values.

So far we have described an output queued switch (since there was no queuing on the input or in the switch)

To facilitate the output queuing we can group the N output ports into groups of n ports



This means that further switching must be carried out in the output stages (not shown) to direct packets to the correct output within the group of n. However this grouping allows buffers to be shared and has a dramatic effect on the speed up required

For a group size of 16 the speed up is down to $m = 2$ even if the number of ports (N) is 1000. (this would apply for a cell loss probability of 10^{-10})