DISTRIBUTED SYSTEMS PRINCIPLES AND PARADIGMS

PROBLEM SOLUTIONS

ANDREW S. TANENBAUM

MAARTEN VAN STEEN

Vrije Universiteit
Amsterdam, The Netherlands

PRENTICE HALL

UPPER SADDLE RIVER, NJ 07458

SOLUTIONS TO CHAPTER 1 PROBLEMS

1. **Q:** What is the role of middleware in a distributed system?

   A: To enhance the distribution transparency that is missing in network operating systems. In other words, middleware aims at improving the single-system view that a distributed system should have.

2. **Q:** Explain what is meant by (distribution) transparency, and give examples of different types of transparency.

   A: Distribution transparency is the phenomenon by which distribution aspects in a system are hidden from users and applications. Examples include access transparency, location transparency, migration transparency, relocation transparency, replication transparency, concurrency transparency, failure transparency, and persistence transparency.

3. **Q:** Why is it sometimes so hard to hide the occurrence and recovery from failures in a distributed system?

   A: It is generally impossible to detect whether a server is actually down, or that it is simply slow in responding. Consequently, a system may have to report that a service is not available, although, in fact, the server is just slow.

4. **Q:** Why is it not always a good idea to aim at implementing the highest degree of transparency possible?

   A: Aiming at the highest degree of transparency may lead to a considerable loss of performance that users are not willing to accept.

5. **Q:** What is an open distributed system and what benefits does openness provide?

   A: An open distributed system offers services according to clearly defined rules. An open system is capable of easily interoperating with other open systems but also allows applications to be easily ported between different implementations of the same system.

6. **Q:** Describe precisely what is meant by a scalable system.

   A: A system is scalable with respect to either its number of components, geographical size, or number and size of administrative domains, if it can grow in one or more of these dimensions without an unacceptable loss of performance.

7. **Q:** Scalability can be achieved by applying different techniques. What are these techniques?

   A: Scaling can be achieved through distribution, replication, and caching.

8. **Q:** What is the difference between a multiprocessor and a multicomputer?

A: In a multiprocessor, the CPUs have access to a shared main memory. There is no shared memory in multicomputer systems. In a multicomputer system, the CPUs can communicate only through message passing.

9. **Q:** A multicomputer with 256 CPUs is organized as a 16×16 grid. What is the worst-case delay (in hops) that a message might have to take?

A: Assuming that routing is optimal, the longest optimal route is from one corner of the grid to the opposite corner. The length of this route is 30 hops. If the end processors in a single row or column are connected to each other, the length becomes 15.

10. **Q:** Now consider a 256-CPU hypercube. What is the worst-case delay here, again in hops?

A: With a 256-CPU hypercube, each node has a binary address, from 00000000 to 11111111. A hop from one machine to another always involves changing a single bit in the address. Thus from 00000000 to 00000001 is one hop. From there to 00000011 is another hop. In all, eight hops are needed.

11. **Q:** What is the difference between a distributed operating system and a network operating system?

A: A distributed operating system manages multiprocessors and homogeneous multicomputers. A network operating system connects different independent computers that each have their own operating system so that users can easily use the services available on each computer.

12. **Q:** Explain how microkernels can be used to organize an operating system in a client-server fashion.

A: A microkernel can separate client applications from operating system services by enforcing each request to pass through the kernel. As a consequence, operating system services can be implemented by (perhaps different) user-level servers that run as ordinary processes. If the microkernel has networking capabilities, there is also no principal objection in placing those servers on remote machines (which run the same microkernel).

13. **Q:** Explain the principal operation of a page-based distributed shared memory system.

A: Page-based DSM makes use of the virtual memory capabilities of an operating system. Whenever an application addresses a memory location that is currently not mapped into the current physical memory, a page fault occurs, giving the operating system control. The operating system can then locate the referred page, transfer its content over the network, and map it to physical memory. At that point, the application can continue.

14. **Q:** What is the reason for developing distributed shared memory systems? What do you see as the main problem hindering efficient implementations?

A: The main reason is that writing parallel and distributed programs based on message-passing primitives is much harder than being able to use shared memory for communication. Efficiency of DSM systems is hindered by the fact, no matter what you do, page transfers across the network need to take place. If pages are shared by different processors, it is quite easy to get into a state similar to thrashing in virtual memory systems. In the end, DSM systems can never be faster than message-passing solutions, and will generally be slower due to the overhead incurred by keeping track of where pages are.

15. **Q:** Explain what false sharing is in distributed shared memory systems. What possible solutions do you see?

A: False sharing happens when data belonging to two different and independent processes (possibly on different machines) are mapped onto the same logical page. The effect is that the page is swapped between the two processes, leading to an implicit and unnecessary dependency. Solutions include making pages smaller or prohibiting independent processes to share a page.

16. **Q:** An experimental file server is up 3/4 of the time and down 1/4 of the time, due to bugs. How many times does this file server have to be replicated to give an availability of at least 99%?

A: With k being the number of servers, we have that $(1/4)^k < 0.01$, expressing that the worst situation, when all servers are down, should happen at most 1/100 of the time. This gives us $k=4$.

17. **Q:** What is a three-tiered client-server architecture?

A: A three-tiered client-server architecture consists of three logical layers, where each layer is, in principle, implemented at a separate machine. The highest layer consists of a client user interface, the middle layer contains the actual application, and the lowest layer implements the data that are being used.

18. **Q:** What is the difference between a vertical distribution and a horizontal distribution?

A: Vertical distribution refers to the distribution of the different layers in a multitiered architectures across multiple machines. In principle, each layer is implemented on a different machine. Horizontal distribution deals with the distribution of a single layer across multiple machines, such as distributing a single database.

19. **Q:** Consider a chain of processes $P_1, P_2, \ldots, P_n$ implementing a multitiered client-server architecture. Process P_i is client of process P_{i+1}, and P_i will return a reply to P_{i-1} only after receiving a reply from P_{i+1}. What are the

main problems with this organization when taking a look at the request-reply performance at process P_1?

A: Performance can be expected to be bad for large n. The problem is that each communication between two successive layers is, in principle, between two different machines. Consequently, the performance between P_1 and P_2 may also be determined by $n-2$ request-reply interactions between the other layers. Another problem is that if one machine in the chain performs badly or is even temporarily unreachable, then this will immediately degrade the performance at the highest level.

SOLUTIONS TO CHAPTER 2 PROBLEMS

1. **Q:** In many layered protocols, each layer has its own header. Surely it would be more efficient to have a single header at the front of each message with all the control in it than all these separate headers. Why is this not done?

   A: Each layer must be independent of the other ones. The data passed from layer $k+1$ down to layer k contains both header and data, but layer k cannot tell which is which. Having a single big header that all the layers could read and write would destroy this transparency and make changes in the protocol of one layer visible to other layers. This is undesirable.

2. **Q:** Why are transport-level communication services often inappropriate for building distributed applications?

   A: They hardly offer distribution transparency meaning that application developers are required to pay significant attention to implementing communication, often leading to proprietary solutions. The effect is that distributed applications, for example, built directly on top of sockets are difficult to port and to interoperate with other applications.

3. **Q:** A reliable multicast service allows a sender to reliably pass messages to a collection of receivers. Does such a service belong to a middleware layer, or should it be part of a lower-level layer?

   A: In principle, a reliable multicast service could easily be part of the transport layer, or even the network layer. As an example, the unreliable IP multicasting service is implemented in the network layer. However, because such services are currently not readily available, they are generally implemented using transport-level services, which automatically places them in the middleware. However, when taking scalability into account, it turns out that reliability can be guaranteed only if application requirements are considered. This is a strong argument for implementing such services at higher, less general layers.

4. **Q:** Consider a procedure *incr* with two integer parameters. The procedure adds one to each parameter. Now suppose that it is called with the same variable twice, for example, as *incr*(i, i). If i is initially 0, what value will it have afterward if call-by-reference is used? How about if copy/restore is used?

   A: If call by reference is used, a pointer to i is passed to *incr*. It will be incremented two times, so the final result will be two. However, with copy/restore, i will be passed by value twice, each value initially 0. Both will be incremented, so both will now be 1. Now both will be copied back, with the second copy overwriting the first one. The final value will be 1, not 2.

5. **Q:** C has a construction called a union, in which a field of a record (called a struct in C) can hold any one of several alternatives. At run time, there is no sure-fire way to tell which one is in there. Does this feature of C have any implications for remote procedure call? Explain your answer.

   A: If the runtime system cannot tell what type value is in the field, it cannot marshal it correctly. Thus unions cannot be tolerated in an RPC system unless there is a tag field that unambiguously tells what the variant field holds. The tag field must not be under user control.

6. **Q:** One way to handle parameter conversion in RPC systems is to have each machine send parameters in its native representation, with the other one doing the translation, if need be. The native system could be indicated by a code in the first byte. However, since locating the first byte in the first word is precisely the problem, can this actually work?

   A: First of all, when one computer sends byte 0, it always arrives in byte 0. Thus the destination computer can simply access byte 0 (using a byte instruction) and the code will be in it. It does not matter whether this is the low-order byte or the high-order byte. An alternative scheme is to put the code in all the bytes of the first word. Then no matter which byte is examined, the code will be there.

7. **Q:** Assume a client calls an asynchronous RPC to a server, and subsequently waits until the server returns a result using another asynchronous RPC. Is this approach the same as letting the client execute a normal RPC? What if we replace the asynchronous RPCs with asynchronous RPCs?

   A: No, this is not the same. An asynchronous RPC returns an acknowledgement to the caller, meaning that after the first call by the client, an additional message is sent across the network. Likewise, the server is acknowledged that its response has been delivered to the client. Two asynchronous RPCs may be the same, provided reliable communication is guaranteed. This is generally not the case.

8. **Q:** Instead of letting a server register itself with a daemon as is done in DCE, we could also choose to always assign it the same endpoint. That endpoint can then be used in references to objects in the server's address space. What is the main drawback of this scheme?

   A: The main drawback is that it becomes much harder to dynamically allocate objects to servers. In addition, many endpoints need to be fixed, instead of just one (i.e., the one for the daemon). For machines possibly having a large number of servers, static assignment of endpoints is not a good idea.

9. **Q:** Give an example implementation of an object reference that allows a client to bind to a transient remote object.

A: Using Java, we can express such an implementation as the following class:

```
public class Object_reference {
    InetAddress server_address;     // network address of object's server
    int         server_endpoint;    // endpoint to which server is listening
    int         object_identifier;  // identifier for this object
    URL         client_code;        // (remote) file containing client-side stub
    byte[]      init_data;          // possible additional initialization data
}
```

The object reference should at least contain the transport-level address of the server where the object resides. We also need an object identifier as the server may contain several objects. In our implementation, we use a URL to refer to a (remote) file containing all the necessary client-side code. A generic array of bytes is used to contain further initialization data for that code. An alternative implementation would have been to directly put the client-code into the reference instead of a URL. This approach is followed, for example, in Java RMI where proxies are passed as reference.

10. **Q:** Java and other languages support exceptions, which are raised when an error occurs. How would you implement exceptions in RPCs and RMIs?

    A: Because exceptions are initially raised at the server side, the server stub can do nothing else but catch the exception and marshal it as a special error response back to the client. The client stub, on the other hand, will have to unmarshal the message and raise the same exception if it wants to keep access to the server transparent. Consequently, exceptions now also need to be described in an interface definition language.

11. **Q:** Would it be useful to also make a distinction between static and dynamic RPCs?

    A: Yes, for the same reason it is useful with remote object invocations: it simply introduces more flexibility. The drawback, however, is that much of the distribution transparency is lost for which RPCs were introduced in the first place.

12. **Q:** Some implementations of distributed-object middleware systems are entirely based on dynamic method invocations. Even static invocations are compiled to dynamic ones. What is the benefit of this approach?

    A: Realizing that an implementation of dynamic invocations can handle *all* invocations, static ones become just a special case. The advantage is that only a single mechanism needs to be implemented. A possible disadvantage is that performance is not always as optimal as it could be had we analyzed the static invocation.

13. Q: Describe how connectionless communication between a client and a server proceeds when using sockets.

A: Both the client and the server create a socket, but only the server binds the socket to a local endpoint. The server can then subsequently do a blocking read call in which it waits for incoming data from any client. Likewise, after creating the socket, the client simply does a blocking call to write data to the server. There is no need to close a connection.

14. Q: Explain the difference between the primitives mpi_bsend and mpi_isend in MPI.

A: The primitive mpi_bsend uses buffered communication by which the caller passes an entire buffer containing the messages to be sent, to the local MPI runtime system. When the call completes, the messages have either been transferred, or copied to a local buffer. In contrast, with mpi_isend, the caller passes only a pointer to the message to the local MPI runtime system after which it immediately continues. The caller is responsible for not overwriting the message that is pointed to until it has been copied or transferred.

15. Q: Suppose you could make use of only transient asynchronous communication primitives, including only an asynchronous receive primitive. How would you implement primitives for transient *synchronous* communication?

A: Consider a synchronous send primitive. A simple implementation is to send a message to the server using asynchronous communication, and subsequently let the caller continuously poll for an incoming acknowledgement or response from the server. If we assume that the local operating system stores incoming messages into a local buffer, then an alternative implementation is to block the caller until it receives a signal from the operating system that a message has arrived, after which the caller does an asynchronous receive.

16. Q: Now suppose you could make use of only transient synchronous communication primitives. How would you implement primitives for transient *asynchronous* communication?

A: This situation is actually simpler. An asynchronous send is implemented by having a caller append its message to a buffer that is shared with a process that handles the actual message transfer. Each time a client appends a message to the buffer, it wakes up the send process, which subsequently removes the message from the buffer and sends it its destination using a blocking call to the original send primitive. The receiver is implemented similarly by offering a buffer that can be checked for incoming messages by an application.

17. Q: Does it make sense to implement persistent asynchronous communication by means of RPCs?

A: Yes, but only on a hop-to-hop basis in which a process managing a queue passes a message to a next queue manager by means of an RPC. Effectively, the service offered by a queue manager to another is the storage of a message. The calling queue manager is offered a proxy implementation of the interface to the remote queue, possibly receiving a status indicating the success or failure of each operation. In this way, even queue managers see only queues and no further communication.

18. Q: In the text we stated that in order to automatically start a process to fetch messages from an input queue, a daemon is often used that monitors the input queue. Give an alternative implementation that does not make use of a daemon.

A: A simple scheme is to let a process on the receiver side check for any incoming messages each time that process puts a message in its own queue.

19. Q: Routing tables in IBM MQSeries, and in many other message-queuing systems, are configured manually. Describe a simple way to do this automatically.

A: The simplest implementation is to have a centralized component in which the topology of the queuing network is maintained. That component simply calculates all best routes between pairs of queue managers using a known routing algorithm, and subsequently generates routing tables for each queue manager. These tables can be downloaded by each manager separately. This approach works in queuing networks where there are only relatively few, but possibly widely dispersed, queue managers.

A more sophisticated approach is to decentralize the routing algorithm, by having each queue manager discover the network topology, and calculate its own best routes to other managers. Such solutions are widely applied in computer networks. There is no principle objection for applying them to message-queuing networks.

20. Q: How would you incorporate persistent asynchronous communication into a model of communication based on RMIs to remote objects?

A: An RMI should be asynchronous, that is, no immediate results are expected at invocation time. Moreover, an RMI should be stored at a special server that will forward it to the object as soon as the latter is up and running in an object server.

21. Q: With persistent communication, a receiver generally has its own local buffer where messages can be stored when the receiver is not executing. To create such a buffer, we may need to specify its size. Give an argument why this is preferable, as well as one against specification of the size.

A: Having the user specify the size makes its implementation easier. The system creates a buffer of the specified size and is done. Buffer management becomes easy. However, if the buffer fills up, messages may be lost. The alternative is to have the communication system manage buffer size, starting

with some default size, but then growing (or shrinking) buffers as need be. This method reduces the chance of having to discard messages for lack of room, but requires much more work of the system.

22. **Q:** Explain why transient synchronous communication has inherent scalability problems, and how these could be solved.

A: The problem is the limited geographical scalability. Because synchronous communication requires that the caller is blocked until its message is received, it may take a long time before a caller can continue when the receiver is far away. The only way to solve this problem is to design the calling application so that it has other useful work to do while communication takes place, effectively establishing a form of asynchronous communication.

23. **Q:** Give an example where multicasting is also useful for discrete data streams.

A: Passing a large file to many users as is the case, for example, when updating mirror sites for Web services or software distributions.

24. **Q:** How could you guarantee a maximum end-to-end delay when a collection of computers is organized in a (logical or physical) ring?

A: We let a token circulate the ring. Each computer is permitted to send data across the ring (in the same direction as the token) only when holding the token. Moreover, no computer is allowed to hold the token for more than T seconds. Effectively, if we assume that communication between two adjacent computers is bounded, then the token will have a maximum circulation time, which corresponds to a maximum end-to-end delay for each packet sent.

25. **Q:** How could you guarantee a minimum end-to-end delay when a collection of computers is organized in a (logical or physical) ring?

A: Strangely enough, this is much harder than guaranteeing a maximum delay. The problem is that the receiving computer should, in principle, not receive data before some elapsed time. The only solution is to buffer packets as long as necessary. Buffering can take place either at the sender, the receiver, or somewhere in between, for example, at intermediate stations. The best place to temporarily buffer data is at the receiver, because at that point there are no more unforeseen obstacles that may delay data delivery. The receiver need merely remove data from its buffer and pass it to the application using a simple timing mechanism. The drawback is that enough buffering capacity needs to be provided.

26. **Q:** Imagine we have a token bucket specification where the maximum data unit size is 1000 bytes, the token bucket rate is 10 million bytes/sec, the token bucket size is 1 million bytes, and the maximum transmission rate is 50 million bytes/sec. How long can a burst of maximum speed last?

A: Call the length of the maximum burst interval Δt. In an extreme case, the bucket is full at the start of the interval (1 million bytes) and another $10\Delta t$ comes in during that interval. The output during the transmission burst consists of $50\Delta t$ million bytes, which should be equal to $(1+10\Delta t)$. Consequently, Δt is equal to 25 msec.

SOLUTIONS TO CHAPTER 3 PROBLEMS

1. **Q:** In this problem you are to compare reading a file using a single-threaded file server and a multithreaded server. It takes 15 msec to get a request for work, dispatch it, and do the rest of the necessary processing, assuming that the data needed are in a cache in main memory. If a disk operation is needed, as is the case one-third of the time, an additional 75 msec is required, during which time the thread sleeps. How many requests/sec can the server handle if it is single threaded? If it is multithreaded?

   A: In the single-threaded case, the cache hits take 15 msec and cache misses take 90 msec. The weighted average is $2/3 \times 15 + 1/3 \times 90$. Thus the mean request takes 40 msec and the server can do 25 per second. For a multithreaded server, all the waiting for the disk is overlapped, so every request takes 15 msec, and the server can handle 66 2/3 requests per second.

2. **Q:** Would it make sense to limit the number of threads in a server process?

   A: Yes, for two reasons. First, threads require memory for setting up their own private stack. Consequently, having many threads may consume too much memory for the server to work properly. Another, more serious reason, is that, to an operating system, independent threads tend to operate in a chaotic manner. In a virtual memory system it may be difficult to build a relatively stable working set, resulting in many page faults and thus I/O. Having many threads may thus lead to a performance degradation resulting from page thrashing.

3. **Q:** In the text, we described a multithreaded file server, showing why it is better than a single-threaded server and a finite-state machine server. Are there any circumstances in which a single-threaded server might be better? Give an example.

   A: Yes. If the server is entirely CPU bound, there is no need to have multiple threads. It may just add unnecessary complexity. As an example, consider a telephone directory assistance number (like 555-1212) for an area with 1 million people. If each (name, telephone number) record is, say, 64 characters, the entire database takes 64 megabytes, and can easily be kept in the server's memory to provide fast lookup.

4. **Q:** Statically associating only a single thread with a lightweight process is not such a good idea. Why not?

   A: Such an association effectively reduces to having only kernel-level threads, implying that much of the performance gain of having threads in the first place, is lost.

5. **Q:** Having only a single lightweight process per process is also not such a good idea. Why not?

   A: In this scheme, we effectively have only user-level threads, meaning that any blocking system call will block the entire process.

6. **Q:** Describe a simple scheme in which there are as many lightweight processes as there are runnable threads.

   A: Start with only a single LWP and let it select a runnable thread. When a runnable thread has been found, the LWP creates another LWP to look for a next thread to execute. If no runnable thread is found, the LWP destroys itself.

7. **Q:** Proxies can support replication transparency by invoking each replica, as explained in the text. Can (the server side of) an object be subject to a replicated invocation?

   A: Yes: consider a replicated object A invoking another (nonreplicated) object B. If A consists of k replicas, an invocation of B will be done by each replica. However, B should normally be invoked only once. Special measures are needed to handle such replicated invocations.

8. **Q:** Constructing a concurrent server by spawning a process has some advantages and disadvantages compared to multithreaded servers. Mention a few.

   A: An important advantage is that separate processes are protected against each other, which may prove to be necessary as in the case of a superserver handling completely independent services. On the other hand, process spawning is a relatively costly operation that can be saved when using multithreaded servers. Also, if processes do need to communicate, then using threads is much cheaper as in many cases we can avoid having the kernel implement the communication.

9. **Q:** Sketch the design of a multithreaded server that supports multiple protocols using sockets as its transport-level interface to the underlying operating system.

   A: A relatively simple design is to have a single thread T waiting for incoming transport messages (TPDUs). If we assume the header of each TPDU contains a number identifying the higher-level protocol, the tread can take the payload and pass it to the module for that protocol. Each such module has a separate thread waiting for this payload, which it treats as an incoming request. After handling the request, a response message is passed to T, which, in turn, wraps it in a transport-level message and sends it to tthe proper destination.

10. **Q:** How can we prevent an application from circumventing a window manager, and thus being able to completely mess up a screen?

    A: Use a microkernel approach by which the windowing system including the window manager are run in such a way that all window operations are

required to go through the kernel. In effect, this is the essence of transferring the client-server model to a single computer as we also explained in Chap. 1.

11. Q: Explain what an object adapter is.

A: An object adapter is a generic program capable of accepting incoming invocation requests and passing these on to the server-side stub of an object. An adapter is primarily responsible for implementing an invocation policy, which determines if, how, and how many multiple threads are used to invoke an object.

12. Q: Mention some design issues for an object adapter that is to support persistent objects.

A: The most important issue is perhaps generating an object reference that can be used independently of the current server and adapter. Such a reference should be able to be passed on to a new server and to perhaps indicate a specific activation policy for the referred object. Other issues include exactly when and how a persistent object is written to disk, and to what extent the object's state in main memory may differ from the state kept on disk.

13. Q: Change the procedure thread_per_object in the example of the object adapters, such that all objects under control of the adapter are handled by a single thread.

A: The code stays almost the same, except that we need to create only a single thread that runs a modified version of thread_per_object. This thread is referred to as adapter_thread (its creation is not shown). When the demultiplexer calls the adapter, it puts a message in the buffer of adapter_thread, after which the latter picks it up as before. The adapter thread invokes the appropriate stub, and handles the response message.

```
#include <header.h>
#include <thread.h>
#define MAX_OBJECTS     100
#define NULL            0
#define ANY             -1

METHOD_CALL invoke[MAX_OBJECTS];    /* array of pointers to stubs    */
THREAD *root;                       /* demultiplexer thread          */
THREAD *adapter_thread              /* thread that runs single_thread*/

void single_thread(long object_id) {
    message     *req, *res;                 /* request/response message*/
    unsigned    size;                   /* size of messages          */
    char        **results;                  /* array with all results*/

    while(TRUE) {
        get_msg(&size, (char*) &req);       /* block for invocation request */

        /* Pass request to the appropriate stub. The stub is assumed to      */
        /* allocate memory for storing the results.                          */
        (invoke[req->object_id]*)(req->size, req->data, &size, results);

        res = malloc(sizeof(message)+size); /* create response message   */
        res->object_id = object_id;         /* identify object           */
        res->method_id = req.method_id;     /* identify method           */
        res->size = size;                   /* set size of invocation results*/
        memcpy(res->data, results, size);   /* copy results into response */
        put_msg(root, sizeof(res), res);    /* append response to buffer  */
        free(req);                          /* free memory of request     */
        free(*results);                             /* free memory of results*/
    }
}

void invoke_adapter(long oid, message *request) {
    put_msg(adapter_thread, sizeof(request), request);
}
```


14. Q: Is a server that maintains a TCP/IP connection to a client stateful or stateless?

A: Assuming the server maintains no other information on that client, one could justifiably argue that the server is stateless. The issue is that not the server, but the transport layer at the server maintains state on the client. What the local operating systems keep track of is, in principle, of no concern to the server.

15. Q: Imagine a Web server that maintains a table in which client IP addresses are mapped to the most recently accessed Web pages. When a client connects to the server, the server looks up the client in its table, and if found, returns the registered page. Is this server stateful or stateless?

A: It can be strongly argued that this is a stateless server. The important issue with stateless designs is not if *any* information is maintained by the server on its clients, but rather how *accurate* that information has to be. In this example, if the table is lost for what ever reason, the client and server can still properly interact as if nothing happened. In a stateful design, such an interaction would be possible only after the server had recovered from a possible fault.

16. Q: To what extent does Java RMI rely on code migration?

A: Considering that object references are actually portable proxies, each time an object reference is passed, we are actually migrating code across the

network. Fortunately, proxies have no execution state, so that support for simple weak mobility is all that is needed.

17. Q: Strong mobility in UNIX systems could be supported by allowing a process to fork a child on a remote machine. Explain how this would work.

A: Forking in UNIX means that a complete image of the parent is copied to the child, meaning that the child continues just after the call to fork. A similar approach could be used for remote cloning, provided the target platform is exactly the same as where the parent is executing. The first step is to have the target operating system reserve resources and create the appropriate process and memory map for the new child process. After this is done, the parent's image (in memory) can be copied, and the child can be activated. (It should be clear that we are ignoring several important details here.)

18. Q: In Fig. 3-13 it is suggested that strong mobility cannot be combined with executing migrated code in a target process. Give a counterexample.

A: If strong mobility takes place through thread migration, it should be possible to have a migrated thread be executed in the context of the target process.

19. Q: Consider a process P that requires access to file F which is locally available on the machine where P is currently running. When P moves to another machine, it still requires access to F. If the file-to-machine binding is fixed, how could the systemwide reference to F be implemented?

A: A simple solution is to create a separate process Q that handles remote requests for F. Process P is offered the same interface to F as before, for example in the form of a proxy. Effectively, process Q operates as a file server.

20. Q: Each agent in D'Agents is implemented by a separate process. Agents can communicate primarily through shared files and by means of message passing. Files cannot be transferred across machine boundaries. In terms of the mobility framework given in Sec. 3.4, which parts of an agent's state, as given in Fig. 3-19, comprise the resource segment?

A: The resource segment contains all references to local and global resources. As such, it consists of those variables that refer to other agents, local files, and so on. In D'Agents, these variables are primarily contained in the part consisting of global program variables. What makes matters simple, is that virtually all resources in D'Agents are nontransferrable. Only agents can move betweeen machines. Because agents are already named by global references, namely an *(address, local-id)* pair, transforming references to resources in the presence of migration is relatively simple in D'Agents.

21. Q: Compare the architecture of D'Agents with that of an agent platform in the FIPA model.

A: The main distinction between the two is that D'Agents does not really have a separate directory service. Instead, it offers only a low-level naming service by which agents can be globally referenced. The management component in the FIPA architecture corresponds to the server in D'Agents, whereas the ACC is implemented by the communication layer. The FIPA model provides no further details on the architecture of an agent, in contrast to D'Agents.

22. Q: Where do agent communication languages (ACLs) fit into the OSI model?

A: Such languages are part of the application layer.

23. Q: Where does an agent communication language fit into the OSI model, when it is implemented on top of a system for handling e-mail, such as in D'Agents? What is the benefit of such an approach?

A: It would still be part of the application layer. An important reason for implementing ACLs on top of e-mail, is simplicity. A complete, worldwide communicaton infrastructure is available for handling asynchronous message-passing between agents. Essentially, such an approach comes close to message-queuing systems discussed in Chap. 2.

24. Q: Why is it often necessary to specify the ontology in an ACL message?

A: In this context, an ontology can best be interpreted as a reference to a standard interpretation of the actual data contained in an ACL message. Usually, data in message-passing systems is assumed to be correctly interpreted by the sender and receiver of a message. Agents are often considered to be highly independent from each other. Therefore, it can not always be assumed that the receiving side will interpret the transferred data correctly. Of course, it is necessary that there is common agreement on the interpretation of the ontology field.

SOLUTIONS TO CHAPTER 4 PROBLEMS

1. **Q:** Give an example of where an address of an entity *E* needs to be further resolved into another address to actually access *E*.

   A: IP addresses in the Internet are used to address hosts. However, to access a host, its IP address needs to be resolved to, for example, an Ethernet address.

2. **Q:** Would you consider a URL such as *http://www.acme.org/index.html* to be location independent? What about *http://www.acme.nl/index.html*?

   A: Both names can be location independent, although the first one gives fewer hints on the location of the named entity. Location independent means that the name of the entity is independent of its address. By just considering a name, nothing can be said about the address of the associated entity.

3. **Q:** Give some examples of true identifiers.

   A: Examples are ISBN numbers for books, identification numbers for software and hardware products, employee numbers within a single organization, and Ethernet addresses (although some addresses are used to identify a machine instead of just the Ethernet board).

4. **Q:** How is a mounting point looked up in most UNIX systems?

   A: By means of a mount table that contains an entry pointing to the mount point. This means that when a mounting point is to be looked up, we need to go through the mount table to see which entry matches a given mount point.

5. **Q:** Jade is a distributed file system that uses per-user name spaces (see In other words, each user has his own, private name space. Can names from such name spaces be used to share resources between two different users?

   A: Yes, provided names in the per-user name spaces can be resolved to names in a shared, global name space. For example, two identical names in different name spaces are, in principle, completely independent and may refer to different entities. To share entities, it is necessary to refer to them by names from a shared name space. For example, Jade relies on DNS names and IP addresses that can be used to refer to shared entities such as FTP sites.

6. **Q:** Consider DNS. To refer to a node *N* in a subdomain implemented as a different zone than the current domain, a name server for that zone needs to be specified. Is it always necessary to include a resource record for that server's address, or is it sometimes sufficient to provide only its domain name?

   A: When the name server is represented by a node *NS* in a domain other than the one in which *N* is contained, it is enough to give only its domain name. In that case, the name can be looked up by a separate DNS query. This is not possible when *NS* lies in the same subdomain as *N*, for in that case, you would need to contact the name server to find out its address.

7. **Q:** Is an identifier allowed to contain information on the entity it refers to?

   A: Yes, but that information is not allowed to change, because that would imply changing the identifier. The old identifier should remain valid, so that changing it would imply that an entity has two identifiers, violating the second property of identifiers.

8. **Q:** Outline an efficient implementation of globally unique identifiers.

   A: Such identifiers can be generated locally in the following way. Take the network address of the machine where the identifier is generated, append the local time to that address, along with a generated pseudo-random number. Although, in theory, it is possible that another machine in the world can generate the same number, chances that this happens are negligible.

9. **Q:** Give an example of how the closure mechanism for a URL could work.

   A: Assuming a process knows it is dealing with a URL, it first extracts the scheme identifier from the URL, such as the string *ftp:*. It can subsequently look up this string in a table to find an interface to a local implementation of the FTP protocol. The next part of the closure mechanism consists of extracting the host name from the URL, such as *www.cs.vu.nl*, and passing that to the local DNS name server. Knowing where and how to contact the DNS server is an important part of the closure mechanism. It is often hard-coded into the URL name resolver that the process is executing. Finally, the last part of the URL, which refers to a file that is to be looked up, is passed to the identified host. The latter uses its own local closure mechanism to start the name resolution of the file name.

10. **Q:** Explain the difference between a hard link and a soft link in UNIX systems.

    A: A hard link is a named entry in a directory file pointing to the same file descriptor as another named entry (in possibly a different directory). A symbolic link is a file containing the (character string) name of another file.

11. **Q:** High-level name servers in DNS, that is, name servers implementing nodes in the DNS name space that are close to the root, generally do not support recursive name resolution. Can we expect much performance improvement if they did?

    A: Probably not: because the high-level name servers constitute the global layer of the DNS name space, it can be expected that changes to that part of the name space do not occur often. Consequently, caching will be highly effective, and much long-haul communication will be avoided anyway. Note that recursive name resolution for low-level name servers is important, because in that case, name resolution can be kept local at the lower-level domain in which the resolution is taking place.

12. **Q:** Explain how DNS can be used to implement a home-based approach to locating mobile hosts.

A: The DNS name of a mobile host would be used as (rather poor) identifier for that host. Each time the name is resolved, it should return the current IP address of the host. This implies that the DNS server responsible for providing that IP address will act as the host's name server. Each time the host moves, it contacts this home server and provides it with its current address. Note that a mechanism should be available to avoid caching of the address. In other words, other name servers should be told *not* to cache the address found.

13. **Q:** A special form of locating an entity is called anycasting, by which a service is identified by means of an IP address (see, for example, Sending a request to an anycast address, returns a response from a server implementing the service identified by that anycast address. Outline the implementation of an anycast service based on the hierarchical location service described in Sec. 4.2.4.

A: Each service has a unique identifier associated with it. Any server implementing that service, inserts its network-level address into the location service at the directory node of the leaf domain in which the server resides. Lookup requests use the identifier of the service, and will automatically return the nearest server implementing that service.

14. **Q:** Considering that a two-tiered home-based approach is a specialization of a hierarchical location service, where is the root?

A: The root is formed jointly by all home locations, but is partitioned in such a way that each mobile entity has its own root server.

15. **Q:** Suppose it is known that a specific mobile entity will almost never move outside domain D, and if it does, it can be expected to return soon. How can this information be used to speed up the lookup operation in a hierarchical location service?

A: Simply encode the domain D in the identifier for the entity that is used for the lookup operation. The operation can then be immediately forwarded to the directory node *dir(D)*, from where the search continues.

16. **Q:** In a hierarchical location service with a depth of k, how many location records need to be updated at most when a mobile entity changes its location?

A: Changing location can be described as the combination of an insert and a delete operation. An insert operation requires that at worst $k+1$ location records are to be changed. Likewise, a delete operation also requires changing $k+1$ records, where the record in the root is shared between the two operations. This leads to a total of $2k+1$ records.

17. **Q:** Consider an entity moving from location A to B, while passing several intermediate locations where it will reside for only a relatively short time. When arriving at B, it settles down for a while. Changing an address in a hierarchical location service may still take a relatively long time to complete, and should therefore be avoided when visiting an intermediate location. How can the entity be located at an intermediate location?

A: Combine the hierarchical location service with forwarding pointers. When the entity starts to move, it leaves behind a forwarding pointer at A to its next (intermediate) location. Each time it moves again, a forwarding pointer is left behind. Upon arrival in B, the entity inserts its new address into the hierarchical location service. The chain of pointers is subsequently cleaned up, and the address in A is deleted.

18. **Q:** When passing a remote reference from process P_1 to P_2 in distributed reference counting, would it help to let P_1 increment the counter, instead of P_2 ?

A: No, because in that case, process P_2 may decide to immediately remove its reference before P_1 has a chance to increment the counter at the object. Consequently, when P_2 sends its decrement message, the counter may drop to zero and the object may be incorrectly removed.

19. **Q:** Make clear that weighted reference counting is more efficient than simple reference counting. Assume communication is reliable.

A: Creation or removal of reference can be done with only a single message, as is also the case with simple reference counting. Passing a reference is much cheaper, as it can be done with only one message containing the copied reference with its partial weight set to half of the weight of the original reference. There is thus no need to contact the object, in contrast to simple reference counting.

20. **Q:** Is it possible in generation reference counting, that an object is collected as garbage while there are still references, but which belong to a generation the object does not know of?

A: No. Whenever a reference of generation k is removed (so that possibly the associated entry in $G[k]$ becomes zero), the object is always informed about any outstanding references of generation $k+1$. If that generation was not yet known to the object, it will be at the time a reference of generation k is removed.

21. **Q:** Is it possible in generation reference counting, that an entry $G[i]$ becomes smaller than zero?

A: Yes. Suppose a reference of generation k is copied, leading to a reference of generation $k+1$. If the latter is removed before the reference from which it

was copied, $G[k+1]$ will be decremented by one. If no reference of generation k had yet been removed, $G[k+1]$ will then drop below zero.

22. **Q:** In reference listing, if no response is received after sending a ping message to process P, the process is removed from the object's reference list. Is it always correct to remove the process?

A: No. It may be possible that the process is temporarily unreachable due to a network partition, for example, caused by a failing router or gateway. In that case, the reference is lost and the skeleton may falsely decide that the object can be removed if the list becomes empty.

23. **Q:** Describe a very simple way to decide that the stabilization step in the tracing-based garbage collector of Lang et al. has been reached.

A: Assume that each process can decide whether its local garbage collector has changed any of the markings on proxies or skeletons during the local propagation step. A very simple, nonscalable solution, is to report whether or not any changes have been made to a central coordinator. As soon as that coordinator has recorded that no more changes have occurred, it tells each process to start reclaiming garbage.

SOLUTIONS TO CHAPTER 5 PROBLEMS

1. **Q:** Name at least three sources of delay that can be introduced between WWV broadcasting the time and the processors in a distributed system setting their internal clocks.

A: First we have signal propagation delay in the atmosphere. Second we might have collision delay while the machines with the WWV receivers fight to get on the Ethernet. Third, there is packet propagation delay on the LAN. Fourth, there is delay in each processor after the packet arrives, due to interrupt processing and internal queueing delays.

2. **Q:** Consider the behavior of two machines in a distributed system. Both have clocks that are supposed to tick 1000 times per millisecond. One of them actually does, but the other ticks only 990 times per millisecond. If UTC updates come in once a minute, what is the maximum clock skew that will occur?

A: The second clock ticks 990,000 times per second, giving an error of 10 msec per second. In a minute this error has grown to 600 msec. Another way of looking at it is that the second clock is one percent slow, so after a minute it is off by 0.01×60 sec, or 600 msec.

3. **Q:** Add a new message to Fig. 5-7 that is concurrent with message A, that is, it neither happens before A nor happens after A.

A: The solution cannot involve 0 or it would be ordered. Thus it must be a message from 1 to 2 or from 2 to 1. If it departs or arrives from 1 after 16, it will be ordered with respect to A, so it must depart or arrive before 16. The possibilities are a message leaving process 2 at 0 and arriving at process 1 at 8, or a message leaving process 1 at 0 and arriving at process 2 at 10. Both of these are concurrent with A.

4. **Q:** To achieve totally-ordered multicasting with Lamport timestamps, is it strictly necessary that each message is acknowledged?

A: No, it is sufficient to multicast any other type of message, as long as that message has a timestamp larger than the received message. The condition for delivering a message m to the application, is that another message has been received from each other process with a large timestamp. This guarantees that there are no more messages underway with a lower timestamp.

5. **Q:** Consider a communication layer in which messages are delivered only in the order that they were sent. Give an example in which even this ordering is unnecessarily restrictive.

A: Imagine the transfer of a large image which, to that end, has been divided into consecutive blocks. Each block is identified by its position in the original image, and possibly also its width and height. In that case, FIFO ordering is

not necessary, as the receiver can simply paste each incoming block into the correct position.

6. **Q:** Suppose that two processes detect the demise of the coordinator simultaneously and both decide to hold an election using the bully algorithm. What happens?

   A: Each of the higher-numbered processes will get two *ELECTION* messages, but will ignore the second one. The election will proceed as usual.

7. **Q:** In Fig. 5-12 we have two *ELECTION* messages circulating simultaneously. While it does no harm to have two of them, it would be more elegant if one could be killed off. Devise an algorithm for doing this without affecting the operation of the basic election algorithm.

   A: When a process receives an *ELECTION* message, it checks to see who started it. If it itself started it (i.e., its number is at the head of the list), it turns the message into a *COORDINATOR* message as described in the text. If it did not start any *ELECTION* message, it adds its process number and forwards it along the ring. However, if it did send its own *ELECTION* message earlier and it has just discovered a competitor, it compares the originator's process number with its own. If the other process has a lower number, it discards the message instead of passing it on. If the competitor is higher, the message is forwarded in the usual way. In this way, if multiple *ELECTION* messages are started, the one whose first entry is highest survives. The rest are killed off along the route.

8. **Q:** Many distributed algorithms require the use of a coordinating process. To what extent can such algorithms actually be considered distributed? Discuss.

   A: In a centralized algorithm, there is often one, fixed process that acts as coordinator. Distribution comes from the fact that the other processes run on different machines. In distributed algorithms with a nonfixed coordinator, the coordinator is chosen (in a distributed fashion) among the processes that form part of the algorithm. The fact that there is a coordinator does not make the algorithm less distributed.

9. **Q:** In the centralized approach to mutual exclusion (Fig. 5-13), upon receiving a message from a process releasing its exclusive access to the critical region it was using, the coordinator normally grants permission to the first process on the queue. Give another possible algorithm for the coordinator.

   A: Requests could be associated with priority levels, depending on their importance. The coordinator could then grant the highest priority request first.

10. **Q:** Consider Fig. 5-13 again. Suppose that the coordinator crashes. Does this always bring the system down? If not, under what circumstances does this happen? Is there any way to avoid the problem and make the system able to tolerate coordinator crashes?

   A: Suppose that the algorithm is that every request is answered immediately, either with permission or with denial. If there are no processes in critical regions and no processes queued, then a crash is not fatal. The next process to request permission will fail to get any reply at all, and can initiate the election of a new coordinator. The system can be made even more robust by having the coordinator store every incoming request on disk *before* sending back a reply. In this way, in the event of a crash, the new coordinator can reconstruct the list of active critical regions and the queue by reading file from the disk.

11. **Q:** Ricart and Agrawala's algorithm has the problem that if a process has crashed and does not reply to a request from another process to enter a critical region, the lack of response will be interpreted as denial of permission. We suggested that all requests be answered immediately, to make it easy to detect crashed processes. Are there any circumstances where even this method is insufficient? Discuss.

    A: Suppose a process denies permission and then crashes. The requesting process thinks that it is alive, but permission will never come. One way out is to have the requester not actually block, but rather go to sleep for a fixed period of time, after which it polls all processes that have denied permission to see if they are still running.

12. **Q:** How do the entries in Fig. 5-16 change if we assume that the algorithms can be implemented on a LAN that supports hardware broadcasts?

    A: Only the entries for the distributed case change. Because sending a point-to-point message is as expensive as doing a broadcast, we need only send one broadcast message to all processes requesting the entry to the critical section. Likewise, only one exit broadcast message is needed. The delay becomes $1+(n-1)$: one delay coming from the broadcast request, and an additional $n-1$ as we still need to receive a message from each other process before being allowed to enter the critical section.

13. **Q:** A distributed system may have multiple, independent critical regions. Imagine that process 0 wants to enter critical region *A* and process 1 wants to enter critical region *B*. Can Ricart and Agrawala's algorithm lead to deadlocks? Explain your answer.

    A: It depends on the ground rules. If processes enter critical regions strictly sequentially, that is, a process in a critical region may not attempt to enter another one, then there is no way that it can block while holding a resource (i.e., a critical section) that some other process wants. The system is then deadlock free. On the other hand, if process 0 may enter critical region *A* and then try to enter critical region *B*, a deadlock can occur if some other process tries to acquire them in the reverse order. The Ricart and Agrawala algorithm itself does not contribute to deadlock since each critical region is handled independently of all the others.

14. **Q:** In Fig. 5-17 we saw a way to update an inventory list atomically using magnetic tape. Since a tape can easily be simulated on disk (as a file), why do you think this method is not used any more?

A: The primary reason is probably that people have gotten greedier and want to do more than they used to. If the users were content to simply maintain the inventory by making a run once a day, one could do it on disk. The problem is that now everyone is demanding instantaneous online access to the data base, which makes it impossible to store the inventory as a tape image.

15. **Q:** In Fig. 5-25(d) three schedules are shown, two legal and one illegal. For the same transactions, give a complete list of all values that x might have at the end, and state which are legal and which are illegal.

A: The legal values are 1, 2, and 3. The illegal values are 4, 5, and 6. The 4 is achieved by first running the middle transaction, then incorrectly interleaving the other two. The 5 is achieved in schedule 3. The 6 is achieved by first setting x to 0 three times, then incrementing it three times.

16. **Q:** When a private workspace is used to implement transactions on files, it may happen that a large number of file indices must be copied back to the parent's workspace. How can this be done without introducing race conditions?

A: One way is by setting a lock at the top of the system to prevent any activity at all until all the indices have been overwritten. It would probably be wise to create an intentions list before starting to guard against crashes.

17. **Q:** Give the full algorithm for whether an attempt to lock a file should succeed or fail. Consider both read and write locks, and the possibility that the file was unlocked, read locked, or write locked.

A: The algorithm is as follows. If the file is unlocked, the attempt always succeeds. If the file is read locked and the attempt is for another read lock, it also succeeds. In all other cases, it fails (i.e., write locking a locked file fails, as does read locking a file already locked for writing).

18. **Q:** Systems that use locking for concurrency control usually distinguish read locks from write locks. What should happen if a process has already acquired a read lock and now wants to change it into a write lock? What about changing a write lock into a read lock?

A: A read lock can only be converted to a write lock if there are no other readers. Doing this conversion is effectively releasing the lock, then immediately trying to reacquiring it as a write lock. This operation fails if there are other readers. If there are other processes waiting to acquire a write lock, it is a matter of policy whether it should succeed; there is no technical objection. Downgrading a lock from write to read is always legal and should always work. (Doing so, however, implicitly gives priority to waiting readers over waiting writers, but this is a legal choice.)

19. **Q:** With timestamp ordering in distributed transactions, suppose a write operation *write(T_1,x)* can be passed to the data manager, because the only, possibly conflicting operation *write(T_2,x)* had a lower timestamp. Why would it make sense to let the scheduler postpone passing *write(T_1,x)* until transaction T_2 finishes?

A: By waiting until transaction T_2 finishes, the scheduler may avoid a cascaded abort in the case T_2 aborts.

20. **Q:** Is optimistic concurrency control more or less restrictive than using timestamps? Why?

A: It is more restrictive because by using the optimistic scheme, if a transaction tries to commit and discovers that another transaction has modified a file it has used, it always aborts. With timestamps, the transaction can sometimes commit, namely, if the other transaction has a lower timestamp.

21. **Q:** Does using timestamping for concurrency control ensure serializability? Discuss.

A: Yes. Stronger yet, it either completely serializes the transactions in their timestamp order, or it aborts some of them.

22. **Q:** We have repeatedly said that when a transaction is aborted, the world is restored to its previous state, as though the transaction had never happened. We lied. Give an example where resetting the world is impossible.

A: Any situation in which physical I/O has occurred cannot be reset. For example, if the process has printed some output, the ink cannot be removed from the paper. Also, in a system that controls any kind of industrial process, it is usually impossible to undo work that has been done.

SOLUTIONS TO CHAPTER 6 PROBLEMS

1. **Q:** Access to shared Java objects can be serialized by declaring its methods as being synchronized. Is this enough to guarantee serialization when such an object is replicated?

   A: No. The problem is that access to each replica is serialized. However, different operations at different replicas may be executed at the same time, leaving the replicated instance variables in an inconsistent state.

2. **Q:** Consider a monitor as discussed in Chap. 1. If threads are allowed to block in a replicated monitor, what do we need to guarantee when signaling a condition variable?

   A: That the same thread is awakened in each replica, so that exactly the same flow of control takes place in all replicas. In practice, this means that the runtime system should be in full control of thread scheduling at each replica.

3. **Q:** Explain in your own words what the main reason is for actually considering weak consistency models.

   A: Weak consistency models come from the need to replicate for performance. However, efficient replication can be done only if we can avoid global synchronizations, which, in turn, can be achieved by loosening constistency constraints.

4. **Q:** Explain how replication in DNS takes place, and why it actually works so well.

   A: The basic idea is that name servers cache previously looked up results. These results can be kept in a cache for a long time, because DNS makes the assumption that name-to-address mappings do not change often.

5. **Q:** During the discussion of consistency models, we often referred to the contract between the software and data store. Why is such a contract needed?

   A: If a program expects a sequentially consistent data store and cannot live with anything less, the store must provide sequential consistency. However, to improve performance, some systems provide a weaker model. It is then essential that the software agrees to abide by the rules imposed by this model. Generally, it means that programs obeying the rules will perceive what looks like a sequentially consistent data store.

6. **Q:** Linearizability assumes the existence of a global clock. However, with strict consistency we showed that such an assumption is not realistic for most distributed systems. Can linearizability be implemented for physically distributed data stores?

   A: Yes. Linearizability assumes loosely synchronized clocks, that is, it assumes that several events may happen within the same time slot. Those events need to be ranked adhering to sequential consistency.

7. **Q:** A multiprocessor has a single bus. Is it possible to implement strictly consistent memory?

   A: Yes. The bus serializes requests so they appear at the memory in absolute time order.

8. **Q:** Why is $W_1(x)a \;\; R_2(x)NIL \;\; R_3(x)a$ not legal for Fig. 6-7(b)?

   A: It violates data coherence.

9. **Q:** In Fig. 6-0, is 000000 a legal output for a distributed shared memory that is only FIFO consistent? Explain your answer.

   A: Yes. Suppose (a) runs first. It prints 00. Now (b) runs. If the store into (a) has not arrived yet, it also prints 00. Now (c) runs. If neither store has arrived yet, it too prints 00.

10. **Q:** In Fig. 6-8, is 001110 a legal output for a sequentially consistent memory? Explain your answer.

    A: Yes. If the processes run in the order (a), (c), (b), this result is obtained.

11. **Q:** At the end of Sec. 6.2.2, we discussed a formal model that said every set of operations on a sequentially consistent data store can be modeled by a string, H, from which all the individual process sequences can be derived. For processes P_1 and P_2 in Fig. 6-9, give all the possible values of H. Ignore processes P_3 and P_4 and do not include their operations in H.

    A: It is clear that P_2 cannot read 1 before it is written, so all values of H will have the write of 1 before the read of 1. Program order is respected, so the write of 2 must come after the read of 1. The only possibilities that remain are these:
    $H = W(x)a \;\; R(x)a \;\; W(x)b \;\; W(x)c$
    $H = W(x)a \;\; R(x)a \;\; W(x)c \;\; W(x)b$

12. **Q:** In Fig. 6-13, a sequentially consistent memory allows six possible statement interleavings. List them all.

    A: The six statement interleavings are as follows:
    (1) a=1; if (b==0); b=1; if (a==0);
    (2) a=1; b=1; if (a==0); if (b==0);
    (3) a=1; b=1; if (b==0); if (a==0);
    (4) b=1; if (a==0); a=1; if (b==0)
    (5) b=1; a=1; if (b==0); if (a==0);
    (6) b=1; a=1; if (a==0); if (b==0);

13. **Q:** It is often argued that weak consistency models impose an extra burden for programmers. To what extent is this statement actually true?

    A: It really depends. Many programmers are used to protect their shared data through synchronization mechanisms such as locks or transactions. The main

idea is that they require a coarser grain of concurrency than one offered at the level of only read and write operations. However, programmers do expect that operations on synchronization variables adhere to sequential consistency.

14. Q: In most implementations of (eager) release consistency in distributed shared memory systems, shared variables are synchronized on a release, but not on an acquire. Why is acquire needed at all then?

A: Acquire is needed to delay a process trying to access shared variables while another process is doing so.

15. Q: Does Orca offer sequential consistency or entry consistency? Explain your answer.

A: Formally, Orca provides only entry consistency, as concurrent operations on the same object are properly serialized. However, when using totally-ordered broadcasting as the means to implement ordering on operations, *all* operations are globally ordered, independent of the object on which they are performed. In that case, it offers sequential consistency.

16. Q: Does totally-ordered multicasting by means of a sequencer and for the sake of consistency in active replication, violate the end-to-end argument in system design?

A: Yes. The end-to-end argument states that problems should be solved at the same level in which they occur. In this case, we are dealing with the problem of totally-ordered multicasting for achieving consistency in active replication. In primary-based protocols, consistency is achieved by first forwarding all operations to the primary. Using a sequencer, we are actually doing the same but at a lower level of abstraction. In this case, it may have been better to use a primary-based protocol in which updates are propagated by sending operations.

17. Q: What kind of consistency would you use to implement an electronic stock market? Explain your answer.

A: Causal consistency is probably enough. The issue is that reactions to changes in stock values should be consistent. Changes in stocks that are independent can be seen in different orders.

18. Q: Consider a personal mailbox for a mobile user, implemented as part of a wide-area distributed database. What kind of client-centric consistency would be most appropriate?

A: All of them, actually. What it boils down to is that the owner should always see the same mailbox, no matter whether he is reading or updating it. In fact, the simplest implementation for such a mailbox may well be that of a primary-based local-write protocol, where the primary is always located on the user's mobile computer.

19. Q: Describe a simple implementation of read-your-writes consistency for displaying Web pages that have just been updated.

A: The simplest implementation is to let the browser always check whether it is displaying the most recent version of a page. This requires sending a request to the Web server. This scheme is simple as it is already implemented by many systems.

20. Q: Give an example where client-centric consistency can easily lead to write-write conflicts.

A: In our description of client-centric consistency models, we assumed that each data item had a single owner, which had exclusive write access. Dropping this assumption is enough to generate write-write conflicts. For example, if two independent users of the same data eventually bind to the same replica, their respective updates may need to be propagated to that replica. At that point, update conflicts will become apparent.

21. Q: When using a lease, is it necessary that the clocks of a client and the server, respectively, are tightly synchronized?

A: No. If the client takes a pessimistic view concerning the level at which its clock is synchronized with that of the server, it will attempt to obtain a new lease far before the current one expires.

22. Q: Consider a nonblocking primary-backup protocol used to guarantee sequential consistency in a distributed data store. Does such a data store always provide read-your-writes consistency?

A: No. As soon as the updating process receives an acknowledgement that its update is being processed, it may disconnect from the data store and reconnect to another replica. No guarantees are given that the update has already reached that replica. In contrast, with a blocking protocol, the updating process can disconnect only after its update has been fully propagated to the other replicas as well.

23. Q: For active replication to work in general, it is necessary that all operations be carried out in the same order at each replica. Is this ordering always necessary?

A: No. Consider read operations that operate on nonmodified data or commutative write operations. In principle, such situations allow ordering to be different at different replicas. However, it can be hard or impossible to detect whether, for example, two write operations are commutative.

24. Q: To implement totally-ordered multicasting by means of a sequencer, one approach is to first forward an operation to the sequencer, which then assigns it a unique number and subsequently multicasts the operation. Mention two alternative approaches, and compare the three solutions.

A: Another approach is to multicast the operation, but defer delivery until the sequencer has subsequently multicast a sequence number for it. The latter happens after the operation has been received by the sequencer. A third approach is to first get a sequence number from the sequencer, and then multicast the operation.

The first approach (send operation to sequencer), involves sending one point-to-point message with the operation, and a multicast message. The second approach requires two multicast messages: one containing the operation, and one containing a sequence number. The third approach, finally, costs one point-to-point message with the sequence number, followed by a multicast message containing the operation.

25. Q: A file is replicated on 10 servers. List all the combinations of read quorum and write quorum that are permitted by the voting algorithm.

A: The following possibilities of (read quorum, write quorum) are legal. (1, 10), (2, 9), (3, 8), (4, 7), (5, 6), (6, 5), (7, 4), (8, 3), (9, 2), and (10, 1).

26. Q: In the text, we described a sender-based scheme for preventing replicated invocations. In a receiver-based scheme, a receiving replica recognizes copies of incoming messages that belong to the same invocation. Describe how replicated invocations can be prevented in a receiver-based scheme.

A: Assume object A invokes object B. As before, each invocation is assigned the same, unique identifier by all replicas of A. Each replica subsequently multicasts its invocation request to the replicas of B. When a replica of B receives an invocation request, it checks whether it had already received that request from one of A's replicas. If not, the invocation is carried out at the local replica, and a reply message is multicast to all replicas of A. If the invocation had already been done, the incoming request is further ignored. Likewise, a replica of A passes only the first incoming reply for an outstanding invocation request to its local copy of the object, while all subsequent replies to that same invocation request are simply ignored.

27. Q: Consider causally-consistent lazy replication. When exactly can an operation be removed from a write queue?

A: When it is known that the operation has been performed everywhere. This can be detected by keeping track of the last update (from, say L_i) carried out in each copy L_j. If each L_j has performed an operation W from L_i with timestamp $ts(W)$, L_i can remove all operations W' with a smaller timestamp (i.e., with each $W'[k] \leq W[k]$). Therefore, each L_j sends an acknowledgement to L_i containing the timestamp of the most recently performed write operation.

28. Q: For this exercise, you are to implement a simple system that supports multicast RPC. We assume that there are multiple, replicated servers and that each client communicates with a server through an RPC. However, when dealing with replication, a client will need to send an RPC request to each replica. Program the client such that to the application it appears as if a single RPC is sent. Assume you are replicating for performance, but that servers are susceptible to failures.

SOLUTIONS TO CHAPTER 7 PROBLEMS

1. **Q:** Dependable systems are often required to provide a high degree of security. Why?

   A: If, for example, the responses given by servers cannot be trusted to be correct because some malicious party has tampered with them, it hardly makes sense to talk about a dependable system. Likewise, servers should be able to trust their clients.

2. **Q:** What makes the fail-stop model in the case of crash failures so difficult to implement?

   A: The fact that, in practice, servers simply stop producing output. Detecting that they have actually stopped is difficult. As far as another process can see, the server may just be slow, or communication may (temporarily) be failing.

3. **Q:** Consider a Web browser that returns an outdated cached page instead of a more recent one that had been updated at the server. Is this a failure, and if so, what kind of failure?

   A: Whether or not it is a failure depends on the consistency that was promised to the user. If the browser promises to provide pages that are at most T time units old, it may exhibit performance failures. However, a browser can never live up to such a promise in the Internet. A weaker form of consistency is to provide one of the client-centric models discussed in Chap. 7. In that case, simply returning a page from the cache without checking its consistency may lead to a response failure.

4. **Q:** Can the model of triple modular redundancy described in the text handle Byzantine failures?

   A: Absolutely. The whole discussion assumed that failing elements put out random results, which are the same as Byzantine failures.

5. **Q:** How many failed elements (devices plus voters) can Fig. 7-2 handle? Give an example of the worst case that can be masked.

   A: In each row of circles, at most one element can fail and be masked. Furthermore, one voter in each group can also fail provided it is feeding a faulty element in the next stage. For example, if all six elements at the top of their respective columns all fail, two of the three final outputs will be correct, so we can survive six failures.

6. **Q:** Does TMR generalize to five elements per group instead of three? If so, what properties does it have?

   A: Yes, any odd number can be used. With five elements and five voters, up to two faults per group of devices can be masked.

7. **Q:** For each of the following applications, do you think at-least-once semantics or at most once semantics is best? Discuss.

   (a) Reading and writing files from a file server.
   (b) Compiling a program.
   (c) Remote banking.

   A: For (a) and (b), at least once is best. There is no harm trying over and over. For (c), it is best to give it only one try. If that fails, the user will have to intervene to clean up the mess.

8. **Q:** With asynchronous RPCs, a client is blocked until its request has been *accepted* by the server (see Fig. 2-12). To what extent do failures affect the semantics of asynchronous RPCs?

   A: The semantics are generally affected in the same way as ordinary RPCs. A difference lies in the fact that the server will not be processing the request while the client is blocked, which introduces problems when the client crashes in the meantime. Instead, the server simply does its work, and attempts to contact the client later on, if necessary.

9. **Q:** Give an example in which group communication requires no message ordering at all.

   A: Multicasting images in small fragments, where each fragment contains the (x,y) coordinate as part of its data. Likewise, sending the pages of a book, with each page being numbered.

10. **Q:** In reliable multicasting, is it always necessary that the communication layer keeps a copy of a message for retransmission purposes?

    A: No. In many cases, such as when transferring files, it is necessary only that the data is still available at the application level. There is no need that the communication layer maintains its own copy.

11. **Q:** To what extent is scalability of atomic multicasting important?

    A: It really depends on how many processes are contained in a group. The important thing to note is, that if processes are replicated for fault tolerance, having only a few replicas may be enough. In that case, scalability is hardly an issue. When groups of different processes are formed, scalability may become an issue. When replicating for performance, atomic multicasting itself may be overdone.

12. **Q:** In the text, we suggest that atomic multicasting can save the day when it comes to performing updates on an agreed set of processes. To what extent can we guarantee that each update is actually performed?

    A: We cannot give such guarantees, similar to guaranteeing that a server has actually performed an operation after having sent an acknowledgement to the

client. However, the degree of fault tolerance is improved by using atomic multicasting schemes, and makes developing fault-tolerant systems easier.

13. Q: Virtual synchrony is analogous to weak consistency in distributed data stores, with group view changes acting as synchronization points. In this context, what would be the analogon of strong consistency?

A: The synchronization resulting from individual multicasts, be they totally, causally, or FIFO ordered. Note that view changes take place as special multicast messages, which are required to be properly ordered as well.

14. Q: What are the permissable delivery orderings for the combination of FIFO and total-ordered multicasting in Fig. 7-14?

A: There are six orderings possible:

Order 1	Order 2	Order 3	Order 4	Order 5	Order 6
m1	m1	m1	m3	m3	m3
m2	m3	m3	m1	m1	m4
m3	m2	m4	m2	m4	m1
m4	m4	m2	m4	m2	m2

15. Q: Adapt the protocol for installing a next view G_{i+1} in the case of virtual synchrony so that it can tolerate process failures.

A: When a process P receives G_{i+k}, it first forwards a copy of any unstable message it has, regardless to which previous view it belonged, to every process in G_{i+k}, followed by a flush message for G_{i+k}. It can then safely mark the message as being stable. If a process Q receives a message m that was sent in G_j (with $j < i + k$), it discards it when Q was never in G_j. If the most recently installed view at Q is G_l with $l > j$, message m is also discarded (it is a duplicate). If $l = j$ and m had not yet been received, process Q delivers m taking any additional message ordering constraints into account. Finally, if $l < j$, message m is simply stored in the communication layer until G_j has been installed.

16. Q: In the two-phase commit protocol, why can blocking never be completely eliminated, even when the participants elect a new coordinator?

A: After the election, the new coordinator may crash as well. In this case, the remaining participants can also not reach a final decision, because this requires the vote from the newly elected coordinator, just as before.

17. Q: In our explanation of three-phase commit, it appears that committing a transaction is based on majority voting. Is this true?

A: Absolutely not. The point to note is that a recovering process that could not take part in the final decision as taken by the other process, will recover to a state that is consistent with the final choice made by the others.

18. Q: In a piecewise deterministic execution model, is it sufficient to log only messages, or do we need to log other events as well?

A: More logging is generally needed: it concerns *all* nondeterministic events, including local I/O and, in particular, system calls.

19. Q: Explain how the write-ahead log in distributed transactions can be used to recover from failures.

A: The log contains a record for each read and write operation that took place within the transaction. When a failure occurs, the log can be replayed to the last recorded operation. Replaying the log is effectively the opposite from rolling back, which happens when the transaction needed to be aborted.

20. Q: Does a stateless server need to take checkpoints?

A: It depends on what the server does. For example, a database server that has been handed a complete transaction will maintain a log to be able to redo its operations when recovering. However, there is no need to take checkpoints for the sake of the state of the distributed system. Checkpointing is done only for local recovery.

21. Q: Receiver-based message logging is generally considered better than sender-based logging. Why?

A: The main reason is that recovery is entirely local. In sender-based logging, a recovering process will have to contact its senders to retransmit their messages.

SOLUTIONS TO CHAPTER 8 PROBLEMS

1. **Q:** Which mechanisms could a distributed system provide as security services to application developers that believe only in the end-to-end argument in system's design, as discussed in Chap. 5?

   A: None. Applying the end-to-end argument to security services means that developers will not trust anything that is not provided by their own applications. In effect, the distributed system as a whole is considered to be untrusted.

2. **Q:** In the RISSC approach, can all security be concentrated on secure servers or not?

   A: No, we still need to make sure that the local operating systems and communication between clients and servers are secure.

3. **Q:** Suppose you were asked to develop a distributed application that would allow teachers to set up exams. Give at least three statements that would be part of the security policy for such an application.

   A: Obvious requirements would include that students should not be able to access exams before a specific time. Also, any teacher accessing an exam before the actual examination date should be authenticated. Also, there may be a restricted group of people that should be given read access to any exam in preparation, whereas only the responsible teacher should be given full access.

4. **Q:** Would it be safe to join message 3 and message 4 in the authentication protocol shown in Fig. 8-15, into $K_{A,B}(R_B,R_A)$?

   A: Yes, there is no reason why the challenge sent by Alice for Bob cannot be sent in the same message.

5. **Q:** Why is it not necessary in Fig. 8-12 for the KDC to know for sure it was talking to Alice when it receives a request for a secret key that Alice can share with Bob?

   A: Suppose that Chuck had sent the message "I'm Alice and I want to talk to Bob." The KDC would just return $K_{A,KDC}(K_{A,B})$ which can be decrypted only by Alice because she is the only other entity holding the secret key $K_{A,KDC}$.

6. **Q:** What is wrong in implementing a nonce as a timestamp?

   A: Although a timestamp is used only once, it is far from being random. Implementations of security protocols exist that use timestamps as nonces, and which have been succusfully attacked by exploiting the nonrandomness of the nonces.

7. **Q:** In message 2 of the Needham-Schroeder authentication protocol, the ticket is encrypted with the secret key shared between Alice and the KDC. Is this encryption necessary?

   A: No. Because Bob is the only one who can decrypt the ticket, it might as well have been sent as plaintext.

8. **Q:** Can we safely adapt the authentication protocol shown in Fig. 8-19 such that message 3 consists only of R_B?

   A: In principle, if R_B is never used again, then returning it unencrypted should be enough. However, such randomness is seldom found. Therefore, by encrypting R_B, it becomes much more difficult for Chuck to break in and forge message 3.

9. **Q:** Devise a simple authentication protocol using signatures in a public-key cryptosystem.

   A: If Alice wants to authenticate Bob, she sends Bob a challenge R. Bob will be requested to return $K_B^-(R)$, that is, place his signature under R. If Alice is confident that she has Bob's public key, decrypting the response back to R should be enough for her to know she is indeed talking to Bob.

10. **Q:** Assume Alice wants to send a message m to Bob. Instead of encrypting m with Bob's public key K_B^+, she generates a session key $K_{A,B}$ and then sends $[K_{A,B}(m),K_B^+(K_{A,B})]$. Why is this scheme generally better? (*Hint:* consider performance issues.)

    A: The session key has a short, fixed length. In contrast, the message m may be of arbitrary length. Consequently, the combination of using a session key and applying public-key cryptography to a short message will generally provide much better performance than using only a public key on a large message.

11. **Q:** How can role changes be expressed in an access control matrix?

    A: Roles, or protection domains in general, can be viewed as objects with basically a single operation: enter. Whether or not this operation can be called depends on the role from which the request is issued. More sophisticated approaches are also possible, for example, by allowing changes back to previous roles.

12. **Q:** How are ACLs implemented in a UNIX file system?

    A: Each file has three associated entries: one for the owner, one for a group that is associated with the file, and one for everyone else. For each entry, the access rights can essentially be specified as *read*, *write*, *execute*.

13. **Q:** How can an organization enforce the use of a Web proxy gateway and prevent its users to directly access external Web servers?

A: One way is to use a packet-filtering gateway that discards all outgoing traffic except that directed to a few, well-known hosts. Web traffic is accepted provided it is targeted to the company's Web proxy.

14. Q: Referring to Fig. 8-29, to what extent does the use of Java object references as capabilities actually depend on the Java language?

A: It is independent of the Java language: references to secured objects still need to be handed out during runtime and cannot be simply constructed. Java helps by catching the construction of such references during compile time.

15. Q: Name three problems that will be encountered when developers of interfaces to local resources are required to insert calls to enable and disable privileges to protect against unauthorized access by mobile programs as explained in the text.

A: An important one is that no thread switching may occur when a local resource is called. A thread switch could transfer the enabled privileges to another thread that is not authorized to access the resource. Another problem occurs when another local resource needs to be called before the current invocation is finished. In effect, the privileges are carried to the second resource, while it may happen that the caller is actually not trusted to access that second resource. A third problem is that explicitly inserting calls to enable and disable privileges is suspect to programming errors, rendering the mechanism useless.

16. Q: Name a few advantages and disadvantages of using centralized servers for key management.

A: An obvious advantage is simplicity. For example, by having N clients share a key with only a centralized server, we need to maintain only N keys. Pairwise sharing of keys would add up to $N(N-1)/2$ keys. Also, using a centralized server allows efficient storage and maintenance facilities at a single site. Potential disadvantages include the server becoming a bottleneck with respect to performance as well as availability. Also, if the server is compromised, new keys will need to be established.

17. Q: The Diffie-Hellman key-exchange protocol can also be used to establish a shared secret key between three parties. Explain how.

A: Suppose Alice, Bob, and Chuck want to set up a shared secret key based on the two publicly known large primes n and g. Alice has her own secret large number x, Bob has y, and Chuck has z. Alice sends $g^x \bmod n$ to Bob; Bob sends $g^y \bmod n$ to Chuck; and Chuck sends $g^z \bmod n$ to Alice. Alice can now compute $g^{xz} \bmod n$, which she sends to Bob. Bob, in turn, can then compute $g^{xyz} \bmod n$. Likewise, after receiving $g^x \bmod n$ from Alice, Bob can compute $g^{xy} \bmod n$, which he sends to Chuck. Chuck can then compute

$g^{xyz} \bmod n$. Similarly, after Chuck receives $g^y \bmod n$ from Bob, he computes $g^{yz} \bmod n$ and sends that to Alice so that she can compute $g^{xyz} \bmod n$.

18. Q: There is no authentication in the Diffie-Hellman key-exchange protocol. By exploiting this property, a malicious third party, Chuck, can easily break into the key exchange taking place between Alice and Bob, and subsequently ruin the security. Explain how this would work.

A: Assume Alice and Bob are using the publicly known values n and g. When Alice sends $g^x \bmod n$ to Bob, Chuck need simply intercept that message, return his own message $g^z \bmod n$, and thus make Alice believe she is talking to Bob. After intercepting Alice's message, he sends $g^z \bmod n$ to Bob, from which he can expect $g^y \bmod n$ as reply. Chuck is now in the middle.

19. Q: Give a straightforward way how capabilities in Amoeba can be revoked.

A: The object's owner simply requests that the server discard all registered *(rights, check)*-pairs for that object. A disadvantage is that *all* capabilities are revoked. It is difficult to revoke a capability handed out to a specific process.

20. Q: Does it make sense to restrict the lifetime of a session key? If so, give an example how that could be established.

A: Session keys should always have a restricted lifetime as they are easier to break than other types of cryptographic keys. The way to restrict their lifetime is to send along the expiration time when the key is generated and distributed. This approach is followed, for example, in SESAME.

21. Q: What is the role of the timestamp in message 6 in Fig. 8-38, and why does it need to be encrypted?

A: The timestamp is used to protect against replays. By encrypting it, it becomes impossible to replay message 6 with a later timestamp. This example illustrates a general application of timestamps in cryptographic protocols.

22. Q: Complete Fig. 8-38 by adding the communication for authentication between Alice and Bob.

A: Alice sends to Bob the message $M=[K_{B,AS}(A,K_{A,B}),K_{A,B}(t)]$, where $K_{B,AS}$ is the secret key shared between Bob and the AS. At that point, Bob knows he is talking to Alice. By responding with $K_{A,B}(t+1)$, Bob proves to Alice that he is indeed Bob.

23. Q: Consider the communication between Alice and an authentication service AS as in SESAME. What is the difference, if any, between a message $m_1 = K_{AS}^{+}(K_{A,AS}(data))$ and $m_2 = K_{A,AS}(K_{AS}^{+}(data))$?

A: Message m_1 must be decrypted by the authentication service, but results in a message that both the authentication service and Alice can understand.

With message m_2, the authentication service or Alice must first decrypt the message, but only the authentication service can understand its content. In this respect, the two messages are quite different.

24. Q: Define at least two different levels of atomicity for transactions in electronic payment systems.

A: Money atomicity is what we defined in the text: the actual payment should take place entirely, or should fail altogether. A coarser grain of atomicity is goods atomicity. In that case, the whole transaction succeeds if and only if the purchased goods are delivered and paid for.

25. Q: A customer in an e-cash system should preferably wait a random time before using coins it has withdrawn from the bank. Why?

A: Otherwise, the bank may start noticing a relation between deposited coins and the fact that withdrawals from that specific user are taking place. This would violate anonymity.

26. Q: Anonymity of the merchant in any payment system is often forbidden. Why?

A: It is often legally required to know the payee to warrant against the purchase of stolen goods.

27. Q: Consider an electronic payment system in which a customer sends cash to a (remote) merchant. Give a table like the ones used in Figs. 8-44 and 8-0 expressing the hiding of information.

A:

		Information				
		Merchant	**Customer**	**Date**	**Amount**	**Item**
Party	**Merchant**	Full	Partial	Full	Full	Full
	Customer	Full	Full	Full	Full	Full
	Bank	None	None	None	None	None
	Observer	Full	Full	Full	None	None

SOLUTIONS TO CHAPTER 9 PROBLEMS

1. Q: Why is it useful to define the interfaces of an object in an Interface Definition Language?

A: There are several reasons. First, from a software-engineering point of view, having precise and unambiguous interface definitions is important for understanding and maintaining objects. Furthermore, IDL-based definitions come in handy for generating stubs. Finally, and related to the latter, if an IDL definition has been parsed and stored, supporting dynamic invocations becomes easier, as the client-side proxy can be automatically constructed at runtime from an interface definition.

2. Q: Give an example in which CORBA's dynamic invocation mechanisms come in handy.

A: When a process implements a gateway between two different ORBs, it can dynamically construct an invocation for an object at the target ORB without knowing in advance what that object actually looks like. Instead, it can find out dynamically.

3. Q: Which of the six forms of communication discussed in Sec. 2.4 are supported by CORBA's invocation model?

A: Transient asynchronous communication (by means of CORBA one-way requests) and response-based transient synchronous communication (by means of CORBA synchronous requests). Deferred synchronous request can be viewed as a combination of two one-way requests, possibly with a blocking client waiting for the response.

4. Q: Outline a simple protocol that implements at-most-once semantics for an object invocation.

A: A simple solution is to let a proxy retransmit an invocation request, but telling the server explicitly that it is a retransmission. In that case, the server may decide to ignore the request and return an error to the client. In a more sophisticated solution, the server can cache results of previous invocations and check whether it still has those results when receiving a retransmitted request.

5. Q: Should the client and server-side CORBA objects for asynchronous method invocation be persistent?

A: In general, they should be persistent, allowing the client or server to shut down and later restart and fetch the objects from disk. However, in theory, there is no hard reason to demand that these objects should be persistent.

6. Q: In a GIOP *Request* message, the object reference and method name are part of the message header. Why can they not be contained in the message body only?

A: The server will need to demultiplex incoming messages to the appropriate object adapter. It can only look at message headers as it should know nothing about the actual invocation. Demultiplexing can be done on an object reference alone, but further demultiplexing may be needed by also looking at the method that is to be invoked.

7. **Q:** Does CORBA support the thread-per-object invocation policy that we explained in Chap. 3?

   A: Yes, although it can be done only by associating a single object per POA, which in turn should provide the single-threading policy. In other cases, the thread-per-object policy depends on what the underlying ORB provides.

8. **Q:** Consider two CORBA systems, each with their own naming service. Outline how the two naming services could be integrated into a single, federated naming service.

   A: Integration can be straightforward. Because naming contexts are regular objects, a naming context in a foreign name space can be registered under a specific name in the naming graph. Resolving that name will return an IOR referring to the foreign naming context. To a client, a regular object reference is returned for which it can invoke the *resolve* method as usual.

9. **Q:** In the text, we state that when binding to a CORBA object, additional security services may be selected by the client ORB based on the object reference. How does the client ORB know about these services?

   A: In principle, they are encoded as part of an IOR, for example, in the *Components* field of a tagged profile. Such information can be made available by the object itself when its associated object server (or rather, POA) generates its IOR.

10. **Q:** If a CORBA ORB uses several interceptors that are not related to security, to what extent is the order in which interceptors are called important?

    A: It really depends, but in general, if security is needed, most interceptors (at the client side) will be called after calling the access control interceptor, but certainly before the secure invocation interceptor. The latter is strictly required because the secure invocation interceptor produces encrypted messages in which only the destination is visible. Any modification to such a message will lead to a fault when checked for integrity.

11. **Q:** Illustrate how a transactional queue can be part of a distributed transaction as mentioned in the text.

    A: In this case, an application may need to send and receive messages from components running on other machines, and which possibly need to access local databases. By turning the operations on a transactional queue into a nested transaction, and likewise using nested transactions for the other groups of operations, it becomes possible to construct one big distributed transaction.

12. **Q:** In comparing DCOM to CORBA, does CORBA provide standard marshaling or custom marshaling? Discuss.

    A: CORBA provides standard marshaling, but allows customization of proxies and skeletons by means of interceptors. In this sense, custom marshaling in DCOM can be seen as a work-around for the lack of general interception techniques. However, there is no reason why a CORBA developer should not write his own custom proxies, as long as they adhere to language-dependent mapping of the IDL specifications of the object's interfaces. The drawback of this approach, however, is that it requires these custom proxies to be stored in an interface repository from where they can be retrieved by clients. Unlike DCOM, CORBA provides no standard support for storing and retrieving customized proxies. In DCOM, such matters are handled by the registry.

13. **Q:** In DCOM, consider a method m that is contained in interface X, and which accepts a pointer to interface Y as input parameter. Explain what happens with respect to marshaling when a client holding an interface pointer to a proxy implementation of X, invokes m.

    A: When m is invoked, the proxy implementing X will have to marshal the proxy implementing Y. If Y is based on standard marshaling, then, in principle, only the interface identifier of Y needs to be marshaled into the message comprising the marshaled invocation of m. If Y is based on custom marshaling, X will have to marshal Y using the class object associated with Y that handles the marshaling of Y. The CLSID of that class object is prepended to the marshaled version of Y in order to allow the receiver to load the appropriate code to unmarshal Y again.

14. **Q:** Does custom marshaling in DCOM require that the process receiving a marshaled proxy runs on the same type of machine as the sending process?

    A: No, provided that a marshaled proxy can be represented in a machine-independent way. A developer of object-specific proxies will have to take this design issue into account. For example, assume the sending process runs on an Intel-based Windows machine whereas the receiver is running as a Solaris process on a SPARC station. When the sender marshals its proxy, it provides a CLSID identifying the class object that is needed to unmarshal the proxy by the receiver. When the Solaris implementation of DCOM receives this CLSID, it will have to be able to load its specific implementation of the identified class object. That class object is then used to instantiate an object to unmarshal the proxy and return an interface pointer to the receiving process. If the proxy can be unmarshaled to something that can be readily understood by the Solaris implementation of DCOM, then no problems need to occur.

15. **Q:** Outline an algorithm for migrating an object in DCOM to another server.

    A: The main issues that need to be dealt with is migrating the state to the target object server, and having each client rebind to the migrated object. One

possible approach is the following. First, a copy of the object is created on the target server using the object's CLSID. Then, the state is marshaled and transferred to the new server, where the newly created object is instructed to unmarshal that state. The old copy is replaced by an object implementing a forwarding pointer that offers the same interface as the moved object. Each time an invocation request is forwarded, the invoking client will be notified that it should rebind to the object at its new location.

16. **Q:** To what extent does JIT activation violate or comply with the notion of objects?

A: JIT activation essentially assumes that objects have no state. In other words, an object is reduced to a collection of (stateless) methods. In general, such collections are hardly considered to resemble objects.

17. **Q:** Explain what happens when two clients on different machines use the same file moniker to bind to a single object. Will the object be instantiated once or twice?

A: It really depends. If no special measures are taken, two copies of the object will be constructed and initialized with the data from the same file. However, with a special DCOM object server and possibly adapted moniker, it should be possible to create only a single, shared instance.

18. **Q:** What would be the obvious way to implement events in Globe without violating the principle that objects are passive?

A: The best way would be to implement a simple layer on top of Globe objects in which a thread would occasionally poll the underlying object to see if any state changes have occurred. Such a thread could then invoke a callback interface as provided by a client application. In effect, such an implementation changes a pull-style model into a push-style model.

19. **Q:** Give an example in which the (inadvertent) use of callback mechanisms can easily lead to an unwanted situation.

A: If a callback leads to another invocation on the same object, a deadlock may arise if locks are needed to protect shared resources. Situations as these are very hard to control in a general way.

20. **Q:** In CORBA, nodes in naming graph, such as directories, are also considered as objects. Would it make sense to also model directories as distributed shared objects in Globe?

A: Yes. This is best explained by considering a root directory in a worldwide naming service. In most cases, the root directory hardly changes. Consequently, its state, which consists of a collection of object references (in the form of object handles), can be widely replicated. A similar reasoning will show that for many lower-level nodes, a different approach to replication is

needed as the update-to-read ratio is very different. These differences can be dealt with by means of distributed shared objects.

21. **Q:** Assume a Globe object server has just installed a persistent local object. Also, assume that this object offers a contact address. Should that address be preserved as well when the server shuts down?

A: If the address is not preserved, the server should remove it from the Globe location service before shutting down. When the object is loaded again, the server should offer a new address and insert that address into the location service. For efficiency reasons, Globe also supports persistent contact addresses that remain the same even when a server shuts down.

22. **Q:** In our example of using Kerberos for security in Globe, would it be useful to encrypt the ticket with a key shared between the object and the ticket granting service, instead of the key $K_{B,TGT}$?

A: It depends. Using a single object key would allow secure and efficient multicasting of a shared session key to all replicas. However, because the object key needs to be shared between all processes that have a replica, we need to trust that all processes will indeed keep that key a secret.

SOLUTIONS TO CHAPTER 10 PROBLEMS

1. **Q:** Is a file server implementing NFS version 3 required to be stateless?

   A: No, but the NFS protocols are such that it is possible to provide an implementation using stateless servers.

2. **Q:** NFS does not provide a global, shared name space. Is there a way to mimic such a name space?

   A: A global name space can easily be mimiced using a local name space for each client that is partially standardized, and letting the automounter mount the necessary directories into that name space.

3. **Q:** Give a simple extension to the NFS lookup operation that would allow iterative name lookup in combination with a server exporting directories that it mounted from another server.

   A: If a lookup operation always returns an identifier for the server from which a directory was mounted, transparent iterative name lookups across multiple servers would be easy. Whenever a server looks up a mount point on which it mounted a remote file system, it simply returns the server's ID for that file system. The client can then automatically mount the directory as well, and contact its associated server to continue name resolution.

4. **Q:** In UNIX-based operating systems, opening a file using a file handle can be done only in the kernel. Give a possible implementation of an NFS file handle for a user-level NFS server for a UNIX system.

   A: The problem to be solved is to return a file handle that will allow the server to open a file using the existing name-based file system interface. One approach is to encode the file name into the file handle. The obvious drawback is that as soon as the file name changes, its file handles become invalid.

5. **Q:** Using an automounter that installs symbolic links as described in the text makes it harder to hide the fact that mounting is transparent. Why?

   A: After the symbolic link has been followed, the user will not be in the expected directory, but in a subdirectory used by the automounter. In other words, the local home directory for Alice will be */tmp_mnt/home/alice* instead of what she thinks it is, namely */home/alice*. Special support from the operating system or shell is needed to hide this aspect.

6. **Q:** Suppose the current denial state of a file in NFS is *WRITE*. Is it possible that another client can first successfully open that file and then request a write lock?

   A: Yes. If the second client requires read/write access (i.e., value *BOTH*) but no denial (i.e., value *NONE*), it will have been granted access to the file. However, although a write lock may actually be granted, performing an actual write operation will fail. Remember that share reservation is completely independent from locking.

7. **Q:** Taking into account cache coherence as discussed in Chap. 6, which kind of cache-coherence protocol does NFS implement?

   A: Because multiple write operations can take place before cached data is flushed to the server, NFS clients implement a write-back cache.

8. **Q:** Does NFS implement entry consistency? What about release consistency?

   A: Yes. Because share reservations and file locking are associated with specific files, and because a client is forced to revalidate a file when opening it and flush it back to the server when closing it, it can be argued that NFS implements entry consistency. Note that this scheme also comes close to eager release consistency, except that not all files need to be flushed, as would be the case with release consistency, which, by definition, works on all shared data.

9. **Q:** We stated that NFS implements the remote access model to file handling. It can be argued that it also supports the upload/download model. Explain why.

   A: Because a server can delegate a file to a client, it can effectively support whole-file caching and putting that client in charge of further handling of the file. This model comes close to uploading a file to a client and downloading it later when when the client is finished.

10. **Q:** In NFS, attributes are cached using a write-through cache coherence policy. Is it necessary to forward all attributes changes immediately?

    A: No. For example, when appending data to a file, the server does not really need to be informed immediately. Such information may possibly be passed on when the client flushes its cache to the server.

11. **Q:** To what extent will the duplicate-request cache as described in the text actually succeed in implementing at-most-once semantics?

    A: Although the scheme will generally work, there are still numerous problems. For example, the duplicate-request cache is often implemented in volatile memory. Consequently, when the server crashes, the cache is lost, which may eventually lead to processing duplicate requests when the server later recovers. Another problem is that cache entries, in general, will eventually need to be removed to make space for new requests. To do this safely, the client should acknowledge previous replies, possibly by piggybacking ACKs on next requests.

12. **Q:** In NFS, what kind of security measures can be taken to prevent a malicious client from reclaiming a lock it never was granted when a server enters its grace period?

A: One simple solution is to pass an attribute certificate to a client when it first claims a lock. That certificate should be shown again when reclaiming the lock later.

13. Q: Fig. 10-21 suggests that clients have complete transparent access to Vice. To what extent is this indeed true?

A: All issues concerning file access and caching are handled by Venus, so in that sense, user processes can indeed remain unaware of the organization of servers and the placement of volumes. However, when disconnected, it may be impossible to achieve failure transparency if it turns out that a file is not in the cache, or that a conflict needs to be manually resolved.

14. Q: Using RPC2's side effects is convenient for continuous data streams. Give another example in which it makes sense to use an application-specific protocol next to RPC.

A: File transfer. Instead of using a pure RPC mechanism, it may be more efficient to transfer very large files using a protocol such as FTP.

15. Q: If a physical volume in Coda is moved from server *A* to server *B*, which changes are needed in the volume replication database and volume location database?

A: The volume location database needs to be updated with the new location. No other changes are needed.

16. Q: What calling semantics does RPC2 provide in the presence of failures?

A: Considering that the client will be reported an error when an invocation does not complete, RPC2 provides at-most-once semantics.

17. Q: Explain how Coda solves read-write conflicts on a file that is shared between multiple readers and only a single writer.

A: The problem is solved by "defining it away." The semantics of transactions underlying file sharing in Coda, permit treating all readers as accessing the shared file before the writer opened it. Note that read-write conflicts within a specific time interval cannot be solved in this way.

18. Q: In Coda, is it necessary to encrypt the clear token that Alice receives from the AS when she logs in? If so, where does encryption take place?

A: Yes, otherwise the session key in the clear token can be seen by any other process. Encryption takes place as part of the secure RPC. As is also shown in Fig. 10-32, after mutual authentication between Alice and the AS has taken place, the AS generates a session key as part of secure RPC that is used to encrypt the clear token when returning it to Alice.

19. Q: If a file on a Vice server needs its own specific access control list, how can this be achieved?

A: A solution is to create a separate directory containing only that file, and to associate the required access control list with that directory. If necessary, a symbolic link to the file can be created in the directory where that file logically belongs.

20. Q: Does it make sense for a Vice server to offer only *WRITE* permissions on a file? (Hint: consider what happens when a client opens a file.)

A: No. The point is that in Coda, whenever a client opens a file, it receives a copy of that file to cache. This effectively already gives it *READ* permission. Only a Venus process may possibly enforce write-only permission, but in that case, Vice servers will need to trust clients.

21. Q: Can union directories in Plan 9 replace the *PATH* variable in UNIX systems?

A: Yes. By mounting various file systems in a specific order at the same mount point, and subsequently looking only for executables at that mount point, there is no need to look in any other subdirectory.

22. Q: Can Plan 9 be efficiently used as wide-area distributed system?

A: Without modifications, probably not. An important obstacle is keeping files at a server and using a write-through caching policy. In effect, all update operations need to be passed between a client and the server, which introduces a considerable latency in the case of wide-area networks.

23. Q: What is the main advantage of using stripe groups in xFS compared to the approach in which a segment is fragmented across all storage servers?

A: The main advantage is scalability. In a system with many servers, writing a segment requires that all servers are contacted. Using stripe groups, it becomes possible to limit the size of a group and in this way also limits the amount of network traffic and servers that need to be contacted.

24. Q: Using self-certifying path names, is a client always ensured it is communicating with a nonmalicious server?

A: No. SFS does not solve naming problems. In essence, a client will have to trust that the server named in the path can actually be trusted. In other words, a client will have to put its trust in the name and name resolution process. It may very well be the case that a malicious SFS server is spoofing another server by using its IP address and passing the other server's public key.

SOLUTIONS TO CHAPTER 11 PROBLEMS

1. **Q:** To what extent is e-mail part of the Web's document model?

   A: E-mail is not part of the document model underlying the Web, but rather a separate system that has been integrated with hypertext documents by means of client-side software only. For example, most Web browsers provide additional support for handling e-mail, but the actual e-mail messages are not related to hypertext documents in any way.

2. **Q:** The Web uses a file-based approach to documents by which a client first fetches a file before it is opened and displayed. What is the consequence of this approach for multimedia files?

   A: One of the problems is that in many cases such files are first fetched and stored locally before that can be opened. What is needed, however, is to keep the file at the server and set up a data stream to the client. This approach is supported in the Web, but requires special solutions at both the client and the server.

3. **Q:** Why do persistent connections generally improve performance compared to nonpersistent connections?

   A: The real gain is in avoiding connection setup, which requires a 3-way handshake in the case of TCP.

4. **Q:** Explain the difference between a plug-in, an applet, a servlet, and a CGI program.

   A: A plug-in is a piece of code that can be dynamically loaded from a browser's *local* file system. In contrast, an applet is dynamically downloaded from the server to the browser, for which reason security is generally more important. Note that many plug-ins can be dynamically downloaded as well, but generally not without manual interference by the user. A servlet is comparable to a plug-in, but is entirely handled at the server side. A CGI program is a piece of code that runs in a separate process on the same machine as the server.

5. **Q:** Outline the design of a general-purpose URN-to-URL naming service, that is, a service that resolves a URN to the URL of a copy of a document named by the URN.

   A: Such a service has already been discussed in Chap. 4, namely a worldwide hierarchical location service. Instead of using object identifiers as input, the service accepts a URN and looks up the URL of the "nearest" copy of the named document.

6. **Q:** In WebDAV, is it sufficient for a client to show only the lock token to the server in order to obtain write permissions?

   A: No, the client should also identify itself as the rightful owner of the token. For this reason, WebDAV not only hands over a token to a client, but also registers which client has the token.

7. **Q:** Instead of letting a Web proxy compute an expiration time for a document, a server could do this instead. What would be the benefit of such an approach?

   A: An important benefit would be that the expiration time is consistent in a hierarchy of caches. For example, if a browser cache computes a longer expiration time than its associated Web proxy, this would mean that one user would get to see a possibly stale document, while other users that access the same proxy are returned a fresher version.

8. **Q:** Does the Akamai CDN follow a pull-based or push-based distribution protocol?

   A: Because replicas are fetched on demand after a client has been redirected to a CDN server, Akamai is seen to follow a pull-based protocol.

9. **Q:** Outline a simple scheme by which an Akamai CDN server can find out that a cached embedded document is stale without checking the document's validity at the original server.

   A: A simple approach followed by Akamai is to include a hash value of the embedded document in its modified URL. It is important to realize that any client will always retrieve all modified URLs when fetching the main document. Subsequent lookups will eventually lead to a CDN server, which can then conclude that a cached document is no longer valid by comparing hash values in the modified URL of the requested and the cached document, respectively.

10. **Q:** Would it make sense to associate a replication strategy with each Web document separately, as opposed to using one or only a few global strategies?

    A: Probably, considering the wide variety of usage patterns for Web documents. Many documents such as personal home pages, are hardly ever updated, and if so, they are updated by only a single person, making these documents suitable for caching and replication. Dynamically generated documents containing timely information require a different strategy, especially if they need to be replicated for performance. Note that the current Web can hardly differentiate such documents.

11. **Q:** URLs in Notes contain an identifier of a document. To what extent is this an improvement compared to the file names used in the Web?

    A: The main improvement is that even if a document changes its (local) name, this change will not make its URL invalid as would be the case in the Web. The only requirement is that the database name remains the same. Note in this respect, that most URLs in the Web become invalid not because a

document is moved to another server, but only because their local name is changed.

12. **Q:** Does the URL scheme in Notes imply that operations are uniquely named worldwide?

    A: In fact, yes. Because operations are accessed only at a particular database, which is accessed through a globally unique identifier, one could argue that operations are also uniquely named.

13. **Q:** Consider a cluster of Domino servers and suppose a Notes client contacts a busy Domino server in this cluster for the first time. Explain what happens.

    A: Because the client contacts the server for the first time, we can assume it knows nothing about the cluster the server is part of. Consequently, it cannot contact another server and the client's request will have to be deferred until a later time.

14. **Q:** In the text, we described a conflict between two modified Notes documents. It is also possible that a conflict arises between a deleted copy and a modified one. Explain how this conflict arises and how it can be resolved.

    A: The conflict is easily explained by considering deletion as a special modify operation, leaving only a deletion stub behind. If the other copy is subsequently modified more than once, one can argue that this "undoes" the deletion. However, if there is exactly one modification opposed to the deletion, it is harder to decide which operation should prevail. In this case, Notes follows the policy of removing the document.

15. **Q:** To what extent can write-write conflicts occur in the case of cluster replication in Notes?

    A: Cluster replication differs only from the general replication scheme in the frequency by which updates are propagated. Conflict detection and resolution is exactly the same. However, because the time during which replicas are not synchronized is much smaller, there will generally be much fewer conflicts to resolve compared to general replication.

16. **Q:** Give an alternative solution to using cross-certificates in Notes.

    A: Instead of using cross-certificates, two organizations could use a trusted third party that hands out certificates. In other words, they could make use of a certification authority.

17. **Q:** Executing downloaded code in Notes is often based on trust. Is this safe?

    A: In principle, no. However, Notes provides strong authentication and access control to shared databases. In other words, it is not as open as, for example, the Web. In this sense, it is generally safer provided trust can be put into proper system administration.

SOLUTIONS TO CHAPTER 12 PROBLEMS

1. **Q:** What type of coordination model would you classify the message-queuing systems discussed in Chap. 2?

   A: Considering that in message-queuing systems processes are temporally uncoupled, but will otherwise have to agree on message format and destination queues, they classify as mailbox coordination models.

2. **Q:** Referential uncoupling in TIB/Rendezvous is supported by subject-based addressing. What else contributes to referential uncoupling in this system?

   A: The fact that messages are self describing. In that way, there is no need for processes to agree in advance to the format of the messages they exchange. Instead, a receiving process can dynamically determine what an incoming message actually contains. Of course, semantical issues are still left to the application.

3. **Q:** Outline an implementation of a publish/subscribe system based on a message-queuing system like that of IBM MQSeries.

   A: Such an implementation can be accomplished by using a message broker. All publish/subscribe messages are published to the broker. Subscribers pass their subscriptions to the broker as well, which will then take care to forward published messages to the appropriate subscribers. Note that the broker makes use of the underlying message-queuing network.

4. **Q:** To what is a subject name in TIB/Rendezvous actually resolved, and how does name resolution take place?

   A: A name is resolved to the current group of subscribers. Name resolution takes place by properly routing message to those subscribers. In TIB/Rendezvous, routing takes place through a combination of multicasting and filtering messages at rendezvous and router daemons.

5. **Q:** Outline a simple implementation for totally-ordered message delivery in a TIB/Rendezvous system.

   A: Use a sequencer to which all messages are sent. Subscribers pass subscriptions to this sequencer. The FIFO-ordered message delivery of the TIB/Rendezvous system will then guarantee that all messages multicast by the sequencer are delivered to every subscriber in the same order.

6. **Q:** When a message is delivered to a process P as part of a TIB/Rendezvous transaction, P confirms that delivery. Is it possible for the transaction layer at P to automatically send a confirmation to the transaction daemon?

   A: Yes, it could. When a message is delivered to P as part of a transaction, what happens is that the message is removed from an event queue local to P.

Removal of a message can be interpreted by the transaction layer as delivery, so that it can automatically send a confirmation to the transaction daemon.

7. **Q:** Assume a process is replicated in a TIB/Rendezvous system. Give two solutions to avoid so that messages from this replicated process are not published more than once.

   A: A first solution is to attach a message identifier to each published message, and to let subscribers discard duplicates by taking a look at the identifiers. The main drawback of this approach is the waste of network bandwidth. Another solution is to assign a coordinator among the replicas, and let only the coordinator actually publish messages. This solution is similar to assigning a coordinator in the case of invocations with replicated distributed objects.

8. **Q:** To what extent do we need totally-ordered multicasting when processes are replicated in a TIB/Rendezvous system?

   A: Assuming that duplicate messages are either detected by subscribers, or avoided altogether, total ordering is not an issue at all. In this case, the FIFO-ordering delivery semantics are sufficient to let subscribers process the messages in the order they were published by the replicated process.

9. **Q:** Describe a simple scheme for PGM that will allow receivers to detect missing messages, even the last one in a series.

   A: A simple scheme is to use sequence numbers. Whenever a sender has no more data to send, it should announce by multicasting a special control message. If that control message is lost, a receiver will start complaining in the usual way. The sender can then simply retransmit the control message. In essence, this is also the solution adopted in PGM.

10. **Q:** Can ledgers in TIB/Rendezvous that are implemented as files guarantee that messages are never lost, even in the presence of crashing processes?

    A: No. If a ledger is not flushed to disk before the sending process crashes, messages may be lost that later require retransmission by a receiver. The idea of certified message delivery is to increase the reliability of message delivery. By no means is it a solution to 100 percent guaranteed delivery.

11. **Q:** How could a coordination model based on generative communication be implemented in TIB/Rendezvous?

    A: This is actually not very difficult. What makes TIB/Rendezvous different from, for example, the JavaSpaces in Jini, is that processes are still temporally coupled. If published messages are temporarily stored, it would be possible for subscribers to read them even when the publisher of messages no longer exists. What is needed is for each subscriber to record a published message that it has already read, so that receiving duplicates can be avoided.

12. **Q:** Apart from the temporal uncoupling in Jini, in your opinion, what is the most important difference between Jini and TIB/Rendezvous when considering coordination models?

    A: An important difference is the fact that TIB/Rendezvous uses character-string naming to match publishers and subscribers, whereas JavaSpaces uses marshaled object references. Another important difference is that JavaSpaces allows tuple instances to be removed upon reading, thus providing a means to synchronize processes. A comparable facility is not available in TIB/Rendezvous.

13. **Q:** A lease period in Jini is always specified as a duration and not as an absolute time at which the lease expires. Why is this done?

    A: Especially in wide-area systems, it may happen that clocks on different machines give very different times. In that case, specifying the expiration of a lease as an absolute time is simply too inaccurate as the holder of the lease may have a very different idea when the lease expires than the processes that handed out the lease. With durations, this difference becomes less an issue, provided some guarantees can be given that the transmission time of a lease is relatively low.

14. **Q:** What are the most important scalability problems in Jini?

    A: One important problem is related to the fact the Jini uses a multicast protocol to locate lookup services. In a wide-area system, this protocol will have to be replaced by something different if Jini is to scale to large numbers of users and processes. A second problem is related to matching templates to tuples in JavaSpaces. Again, special measures will be needed to search a JavaSpace that may be potentially distributed across a large-scale network. No efficient solutions to these problems are yet known.

15. **Q:** Consider a distributed implementation of a JavaSpace in which tuples are replicated across several machines. Give a protocol to delete a tuple such that race conditions are avoided when two processes try to delete the same tuple.

    A: Use a two-phase commit protocol. In phase one, the process doing the delete sends a message to all the JavaSpace servers holding the tuple asking them to lock the tuple. When all of them are locked, the delete is sent. If a second delete happens simultaneously, it can happen that some servers have locked one tuple and some have locked the other. If a server cannot grant a request because the tuple is already locked, it sends back a negative acknowledgement, which causes the initiator to abort the delete, unlock all the tuples it has locked, wait a random time, and try again later.

16. **Q:** Suppose a transaction T in Jini requires a lock on an object that is currently locked by another transaction T'. Explain what happens.

A: Transaction T will continue to block until the lock is either released or until the lease on the transaction expires. In the latter case, transaction T is simply aborted.

17. Q: Suppose a Jini client caches the tuple it obtained from a JavaSpace so that it can avoid having to go to the JavaSpace the next time. Does this caching make any sense?

A: Caching is senseless because the tuple will have been removed from the JavaSpace when it was returned; it is ready for the client to keep. The main idea behind caching is to keep data local to avoid another server access. In the case of Jini, a JavaSpace is often used to explicitly synchronize processes. Caching does not play a role when process synchronization is explicitly needed.

18. Q: Answer the previous question, but now for the case that a client caches the results returned by a lookup service.

A: This is a completely different situation. The lookup service stores information on the whereabouts of services. In this case, it may indeed make sense for a client to cache previously returned results and try to contact the returned services before going to the lookup service again.

19. Q: Outline a simple implementation of a fault-tolerant JavaSpace.

A: The simplest approach is to implement a JavaSpace on a single server with stable storage. Write operations succeed only if the tuple has been safely written to storage. In a more advanced setting, a distributed JavaSpace can be used in which a server group is used to mask process failures. In that case, the servers may need to follow a two-phase commit protocol for each write operation.