

RESOURCE AVAILABILITY IN INTELLIGENT TELECOMMUNICATIONS NETWORKS

The present invention relates to telecommunications networks and in particular to such networks using distributed computers.

5 Modern telecommunications networks are called upon to provide many diverse services apart from the simple interconnection of telephony. In order to fully utilise the network operators offer many services across their networks to potential customers.

Services such as Freephone, calling cards, premium rate services
10 "televoting" and data transfer services and the like are often provided using service specific dedicated computers. As such services as televoting are transient in nature in so far as they take place over a limited period of time having a dedicated computer for use in that service is an expensive waste of resource. Further, using a dedicated computer may result in network congestion if a substantial number of
15 calls occur over a short period to the location of the dedicated service features and data. Additionally, in order to protect services multiple service specific dedicated units may be required. This, in common with customers having multiple service requirements, results in data duplication, for example, and potential data maintenance problems.

20 It has been proposed to utilise a network of computers each capable of being programmed to provide a number of different services. Although this does not eliminate the need for dedicated computer units for certain services, particularly when new service offerings are under test, it results in a more effective usage of expensive computing equipment. Additionally, the distributed computing
25 arrangement enables the network operator to provide a particular service at differing points in the network thereby enabling a more equal distribution of traffic in the network.

However, the network operator needs to ensure that conflicting applications are not running in the same computing resource and that where a
30 number of services are required to operate contemporaneously sufficient computing resource is available.

According to the present invention there is provided a telecommunications network comprising a service creation facility and a computational platform having

a multiplicity of programmable resources each capable of providing at least one function in a service package comprising a plurality of functions defined by the service creation facility, and control means responsive to data from the service creation facility defining such a service package to allocate the functions of the package within the multiplicity of programmable resources and to direct calls on the telecommunications network to the service package in accordance with the allocations characterised in that when the service creation facility defines an additional service package for future implementation data is forwarded to the control means by the service creation facility such data defining the functions and resources required by the additional package and the time at which such functions and resources are required, and the control means compares available functions and resources for the specified period with the required functions and resources to determine whether the additional service package can be provided.

A distributed computing environment in accordance with the invention will now be described by way of example only with reference to the accompanying drawings of which:

Figure 1 is an architecture diagram of a telecommunications network using the distributed computing environment;

Figure 2 is an architecture diagram showing an intelligence layer of Figure 1 in more detail;

Figure 3 is an architecture diagram showing the computing environment layer of Figure 2 in greater detail;

Figures 4A & 4B show data interchanges which occur within the network of Figure 1 during a particular kind of service call; and

Figure 5 is a block schematic diagram of a service creation environment as used in a telecommunications network in accordance with the invention.

Referring first to figure 1, the invention is primarily for use in a telecommunications network 1 comprising an intelligence domain 2, a transport domain 3, and customer premises equipment domain 4. This "intelligent" network is managed by an operational support system 5 (OSS).

The transport domain 3 comprises one or more telecommunications networks either of switched or non-switched kind. For the purposes of the present invention it is only necessary to mention that the type of network in the transport

area 3 is not significant and may be a data or TDM/PCM voice network, mobile network, data network or other managed broadband system.

The OSS 5 contains functionality which manages customer services and network capability and interfaces to the intelligence domain 2 to provide a high level of integrated management across the whole intelligent network 1. The customer premises equipment domain 4 contains the interface between customer premises equipment and the various transport networks and services within the network 1.

Again, for the purposes of this invention the actual customer premises equipment is not significant to the implementation of the invention merely being an interface between customer and network services provided.

Considering now figure 2, the intelligence domain 2, on which service control functions for call services and features are created, can be divided into the overall applications 21 and the computing environment 22. Interconnection to the transport network 3 is shown purely for completeness.

Now referring to figure 3, the computing environment 2 may be considered as comprising an encapsulation layer 221, distributed computing 222, a support infrastructure (or "computational platform") 223 and internal networking arrangements 224.

The networking arrangement 224 provides data communications (31) within the computing environment 22 and an interface with the external transport network 3 (not shown in figure 3). This allows the computing environment to be independent of the communications technology in use externally.

In addition, network control functions 33 and an operational support system 32 for the computing environment 22 are provided by the networking layer 224.

The computational platform 223 comprises a number of general purpose computers 34 (only one of which is shown) with their respective associated operating systems 35. It should be noted that the computational platform may be physically widely distributed, there being no requirement for co-location of the computers 34.

This distributed computing arrangement provides the functionality to support the real time applications and associated management as required by the

applications layer 21. As may be appreciated not all of the computers 34 will be of the same type. Processing power, memory availability and connected peripheral equipment will vary from unit to unit particularly where certain services require specialised hardware. The encapsulation layer 221 maintains a record of resources
5 available within the distributed computer system. Such resources may include specialised transport and voice peripherals, for example, for which the encapsulation layer has suitable application programming interfaces (API). The encapsulation layer 221 includes an OSS server 36 provided to manage the resources through respective APIs.

10 Considering now the distributed computing environment layer 222, various functions are provided to link the applications required at "run time" to customers with the transport network (3 of figure 2). The customers have no need to know where the particular application resides. Thus throughout the intelligent network, the distributed computing environment 22 contains coherent service applications
15 which may be present in, say, a single computer 34 or may be present in a number of differing physical locations. Certain service applications may be replicated throughout the system.

 The location function 40 provides location transparency to map logical names of addressable resources, services and functions to physical locations,
20 providing the required service to the customer. Embedded within the location function 40 (for access from a service creation environment) is a trading function which matches queries for a particular peripheral or service type to register (or cancel) service offerings. For example, if a peripheral is required capable of providing a voice announcement for, say, a televoting service and to collect and
25 store multi-frequency tone responses then the trading function in association with the resource function will identify which (if any) physical locations can satisfy the service. If more than one location is capable of providing the service then the resource control function 41 considers factors such as load balancing and security to determine which service is selected.

30 Having selected the logical service in response to service creation request and resource control 41 decisions, the location function 40 uses its directory to map the logical service to its physical address(es).

Policy service 42 and security service 43 functions control access to applications and computer resources. The policy service 42 will hold user specific policy in respect of specific failures. Thus while the OSS (5 of figure 1) dictates what occurs in the event of a machine failure, link failure or process failure the distributed computer environment 22 may refer to policy service 42 if other conditions (such as network congestion for example) occur.

Quality of Service (QoS) requirements are also included in the policy service function 42 such that if a component of the system fails to satisfy the requirements, remedial action can be taken by the system by reallocating computational resource (223) or applying protective controls to the network access limiting calls to the pre-allocated resource.

Resource services 41 provides management of the platform resources 223 in an efficient and coherent manner. Functions included in the resource service 41 include monitoring, migration, control, configuration, prioritisation and load balancing.

Each of these functions is now considered in turn in respect of its specific activity.

The monitoring function provides a check on the behaviour of the distributed computer environment 22. Included within the monitoring function is the ability to filter information gathered in respect of loading or faults and determining when significant events have occurred. The monitoring function may arrange to take appropriate action in consultation with the policy service 42. For example, if a particular named function fails to meet its QoS (Quality of Service) requirement, then the monitor function will examine the associated policy statement from the policy server 42. If this specifies that the service should be moved to another area, then the monitor function refers to a factory service 44 to initiate a further instance of the service required.

The migration function is responsible for moving applications to another location. This service may be on another machine or may be a further instance of the same process on the same machine within a computational platform 223.

The control function co-ordinates starting and stopping of services provided. This provides for the starting and stopping of new services in such a way that no calls are lost.

The configuration function maintains a configuration map of the current topology of the computing platform 22 in terms of machines, components currently active, communication links, data stores etc.

The prioritisation function schedules applications in accordance with the
5 priority allocated to the particular service. The function does not decide the priority of an application but provides a priority mechanism which allows changes in application priorities.

Load balancing function maps service requests on to particular instances of a service depending on the current and expected loading of that service. In
10 particular, where the trading function (hereinbefore described) has responded to the service creation function for a particular type of service when more than one instance of that service matches the request the load balancing function determines the most suitable service to be used.

The factory service 44 is responsive to service creation environment and
15 resource function 41 to set up or cease particular instances of a service application. For instance if the resource control service 41 determines that a further instance of a particular service is required within the computational platform 223 in order to meet an increase in the level of demand for that service, it will instruct the factory service to start the appropriate processes in an appropriate
20 machine 34. Similarly, where there is no longer a need for a particular service or where the management system determines that a particular process is deadlocked or no longer functional, then the factory service is responsible for stopping the particular instance of the service from operation.

Finally, within the distributed computing environment 222 the data
25 distribution function 45 provides the functionality of a distributed database management system to present a single logical view of all the data distributed and replicated across a variety of databases within the computational platform or within the computing environment 22. Within the data distribution services function a data migration function ensures that data is located most closely to
30 where it is needed. Any data distribution fragmentation or replication within the distributed computing environment 22 is controlled by this function. The data distribution function ensures that there is transparency of the location of the data required by applications running in the computational platform 223.

Within the distributed computing environment 222 there is also an event notification service responsive to pre-determined events to notify the occurrence of such an event to any resource previously registered as requiring such notification. For Example, if a particular service engine holds cached data in respect of a particular customer any update of data held by the data service causes notification of the update to that service engine.

To assist understanding of the operation of the system of figure 3, consider a service such as the proprietors "0800" or "Freefone" service.

Referring to figures 4A & 4B each potential interaction is numbered in sequence and operation may occur as follows:

51 the transport server has terminated a call that requires a freefone number translation. Although the transport server knows that this call must be directed to a service engine that is capable of processing that call, it does not know the physical address of such a service engine. An implementation may of course take short cuts and could "cache" such information. The location service is asked for the physical address of a service engine. The trader function is invoked which returns a logical service engine that is capable of supporting the required freefone number translation. This logical name is translated into a physical address by the directory service. Again an implementation may store the physical address in the trader which would not require the use of a directory service (in this case). However, if there is more than one suitable service engine then which one should be selected. The trader does not have any decision making capability and the load balancing function within the resource control service must be used. This in turn may have to use the policy function within the policy service.

52 the load balancing function is given a list of service engines from the trader, it decides which one is to be used. However, it may have to consult the policy function within the policy service.

53 the policy function is asked to supply a policy relating to a particular performance condition that the load balancing function has encountered.

54 the policy is returned.

55 given that a logical service engine has been identified its physical address is required. The directory function within the location service is used.

55A It is noted that if at step 51 the trading function had identified only a single logical service engine to provide the required service then steps 52 to 55 above would not be necessary.

56 the physical address of the service engine is returned to the
5 transport server.

57 the service engine is invoked. Although this has the ability to process the service provider's service logic it does not have the service logic for that particular freefone call. For instance 0800 645743 has been dialled, the service engine needs the profile for service provider "645743" which will contain
10 the routing instructions and associated numbers. For instance there could be time of day routing, day of week routing, week of year, automatic call distribution, call barring, each of these would contain the appropriate destination number and instructions.

58 the service engine requires the service provider's profile. It needs to
15 know the address of the data distribution service to ask it for the profile. An implementation may of course keep the interfaces to components at "well-known" addresses and not have to involve the trader. This may prove acceptable in some cases given the limitations of such an approach are known and understood.

59 as for 52 above, although this time the load balancing function is
20 looking for a data distribution service.

60 as for 53 above, although this time the policy function is looking for a data distribution service.

61 policy for data distribution service is returned.

62 as for 55, although this time it is looking for the physical address of
25 a data distribution service.

62A again if the trading function had identified only a single appropriate data distribution service steps 59 to 62 would be unnecessary and the trader would interact directly with the directory function.

63 the directory function returns the physical address of a data
30 distribution service to the service engine.

64 the service engine asks the data distribution function for the required service provider's profile.

65 the data distribution service consults the data service for the profile (again the trader may be involved but of course in an implementation the data distribution service may hold the addresses of data services and take other short cuts).

5 66 the required data is returned to the data distribution service.

67 the required data is returned and processed by the service engine to produce routing instructions (including destination number).

68 the appropriate routing instructions (including destination number) are returned to the transport server.

10 69 the call is routed to the specified destination.

The 0800 or Freephone service noted above is only one type of service which is provided. Other services using the intelligent network 1 will include such applications as telephone voting, telemetering, information services, home control,
15 on line entertainment and so on.

Thus, as each new type of service is identified it is only necessary to provide appropriate peripherals and software somewhere within the intelligence layer 2 of figure 1.

To facilitate understanding of the invention it is necessary to understand
20 the manner in which the service creation environment currently operates. The service creation environment is the point at which customer requirements as specified by the individual customer are translated into network operational requirement and software provision.

The service creation environment is shown schematically in Figure 5 to
25 which reference is now made and comprises four levels. At the network level 104 (hardware capability) the physical network platforms and peripherals are provided. The second level 103 provides software for the network platforms to provide features. the third level 102 builds services from features whilst the fourth 101 uses services packages to provide services in response to customer requirements.

30 Thus at level 101 an operator at a workstation in the service creation environment accepts customer information as to the service required by the customer. Using available service packages the operator seeks to provide the service defined by the customer, for example by modifying an existing service

package such as that previously described in respect of the Freefone service. This procedure is carried out using icons and linking them on a computer screen to enable a tool to assemble appropriate software which results in the new service being deployed if practical 105 by use of the factory resource 44 as hereinbefore
5 described.

If the customer's requirements cannot be met at the first level, then the customer's requirements are referred to a product definition team who are responsible for providing new services. In the product definition level 102, existing software and hardware combinations are considered to determine whether the
10 available service features can be combined to provide the required service package. If this is the case then a new feature description 106 is created and deployed as an available service package 107.

If the existing service features on the network are inadequate to provide the required new service package then the requirement may be referred to a
15 software development team.

In the software development phase 103 the software team will consider all of the resources available within the intelligent network (2 of Figures 3) to determine whether software provision of the required feature is practical. If this is the case then new software 108 may be written to provide the required new
20 feature at 109.

Should the software development level be unable to provide the required service then it may be necessary for the network designer to add hardware capability to the network to enable a new service to be provided.

Whilst the above description is reactive in nature, that is to say is
25 responsive to customer requests at level 101, the service provider may go through the phases of providing relevant features and network capability in anticipation of a demand for a service at any time.

A detailed description of the service creation process may be found in the BT Technology Journal vol 13, no. 2, April 1995, page 80 at seq in an article by G
30 D Turner entitled "Service Creation".

Now, such service creation in response to specific customer requests has previously resulted in a specific platform or service engine being allocated to a particular customer. No consideration of existing availability and network

capability in terms of quantity of resource or quality of service to be provided to the customer has been taken into account. This often results in over provision of network platforms and hardware or a waste of resources which sit idle for a considerable period of time. During creation of a service type, when the network
5 service is specified the operational requirements of that service will also be generated. Such requirements include (but are not necessarily limited to) data defining priority, reliability (both machine and call time application), incompatible call time application identities, concurrency information, deadlines, call time application, life time, operating requirements and quality of service requirements.

10 The priority data is used to enable the selection of call time applications to be selected or rejected in the event of congestion in the network, failure of equipment or deadlock in the computing system.

Machine reliability specifies the characteristics of machines in the computational platform which are capable of execution of the call time application.
15 The data stored here includes information on processor availability, memory availability and the communications availability to the particular machine.

In an alternative to the machine reliability specification being created a call time application reliability may be used. This specifies that the distributed computer environment itself manages the call time application. In view of the
20 capability of the distributed computer environment 222 to reallocate resources in event of failure this may be a preferred option.

Incompatibility will define call time applications which should not run concurrently on the same machine. A list of other call time applications which should not be allowed to execute concurrently can be provided. Applications
25 included in this list may be second instances of the same call time application if this is likely to result in deadlock or conflicting use of the same peripheral.

Concurrency data specifies whether a number of sub-programmes within the application require to operate concurrently. For example, a call time application may spread "threads" to interact with security service or an audit service whilst
30 the call is in progress. This data may specify the number of threads required to enable it to operate in the manner required to meet deadlines.

Deadlines define a time period during which the call time application has to complete its execution. Call time application depends upon timely information

being transferred across the network. This information may be used for example within the resource function 41 of figure 3.

Call time application life time provides the start and finish date and time for a particular application. The call time application will be determined by the service provider who may choose to offer a network service on, say, one day a week every week or on a particular time pattern on a daily basis. Knowing when a call application is going to be present enables better utilisation of the overall resources within the computational platform and the network.

The operating requirements define the computer resources which will be required by the call time application. Thus, call processor unit requirement and input/output (i/o) and data server activities need to be identified. Thus a call time application may interact with a number of sources including for example data servers, OSS servers, input/output service etc. Requirements where a local data store is required by a call time application must also be defined together with the power requirement, for example the number of MIPs or "standard" transactions required.

The operating requirement may also identify data which should be cached. This would improve the performance of the call time application, for example, by avoiding further applications to data distribution services.

Quality of service (QoS) information really defines the service which the customer expects. For example, QoS may define a limit of 500 milliseconds to process calls within the intelligent network 2 but would have a longer boundary since if the call process takes 600 milliseconds on occasion, the end customer will probably not notice. However, it is necessary to know when calls within the network 2 have missed the QoS deadline. This enables long-term resourcing adjustments to take place.

It should be noted that call time applications are not always machine independent and whilst the network 2 provides a general purpose distributed computing environment it may include a number of non-standard machines and, referring again to figure 3, the resource function 41 will know which specialised peripherals and machines are available and where. The data stored above in respect of each application by the service creation environment is used to ensure that the distributed computing environment 22 is used effectively. Thus, if a

customer requests a service creation environment to provide a particular service at a particular time, the service creation environment will forward information to the distributed computing environment 222. The data, including the quality of service and the deadline data together with the application live time is used by the distributed computing environment to consult a call time application diary and a computational platform diary. The computational platform diary contains the schedule in relation to the computational platform 223 showing the availability of the varying machines within the platform and scheduled machine downtime for servicing or other purposes, upgrading and introduction of new machines and the like. The platform diary also includes details of available power in respect of machine processing capabilities and which machines can be grouped to meet operational requirements received from the service creation function. Comparing the information received from the service creation environment with call time application diary and computational platform diaries enables the distributed computing environment 222 to accept the call time application, reject the application or to offer an alternative, lower grade, quality of service.

The call time application diary will also be checked for incompatible call time applications and, assuming that the application is accepted, will be updated to note the requirement for revised call time application.

When a call time application is to be run the factory function 4 in the distributed computing environment causes the application to be started in the computational platform as dictated by the resource function 41. The location of the service is noted such that when calls arrive from the network 3 they are routed by the location broker to the physical address of a machine capable of handling the call. Where a number of computers 34 are running the same application in respect of a particular service, then the location broker will effect distribution of calls through the machine groups specified in accordance with a scheduling algorithm. Call time applications may be added or removed as instructed by the service creation function in the first instance. In the event of failure of one or more machines being detected, then in accordance with the priority data and QoS requirements previously downloaded the resource function 421 will seek to establish the application in another compatible machine.

While, as hereinbefore described, reference is made to the deployment of call time application programs, the programs may be readily available at all times in the environment 222, the service creation facility populating data related to the call time applications for example by providing service data which maps dialled
5 numbers to other numbers of the network. In this case tagging information (ie QOS, computer power requirements etc) may be tagged also to the data. Thus, use of the call time application diary may map the requirements of the service data populated to the availability of computation platform resource.

The call time application diary calculation may, where a multiplicity of
10 computers each have capability of running applications for particular service use a tag indicative of the number MIPS required to run the service into the number of CTU cycles required. This will enable mapping of services onto the run time environment in a way which may be more efficient than complete machine allocation within the service creation environment by allocating computer resources
15 up to the total number of computer cycles available in a particular machine..

CLAIMS

1. A telecommunications network comprising a service creation facility and a computational platform having a multiplicity of programmable resources each
5 capable of providing at least one function in a service package comprising a plurality of functions defined by the service creation facility, and control means responsive to data from the service creation facility defining such a service package to allocate the functions of the package within the multiplicity of programmable resources and to direct calls on the telecommunications network to
10 the service package in accordance with the allocations characterised in that when the service creation facility defines an additional service package for future implementation data is forwarded to the control means by the service creation facility such data defining the functions and resources required by the additional package and the time at which such functions and resources are required, and the
15 control means compares available functions and resources for the specified period with the required functions and resources to determine whether the additional service package can be provided.
2. A telecommunications network as claimed in claim 1, further characterised
20 in that the data forwarded also includes quality of service data defining a maximum time period allowed for handling a call to the service package and the control means also determines whether the functions and resources available can handle calls within said maximum time period.
- 25 3. A telecommunications network as claimed in claim 2, further characterised in that if the control means determines that calls cannot be handled within the maximum time period the control means calculates the time period which would be taken to handle such calls and returns data defining that period to the service creation facility.
30
4. A telecommunications network as claimed in claim 1, further characterised in that if the control means determines that the package cannot be met with defined commencement and/or termination times the control means determines



whether the required functions and resources would be available with a different commencement and/or termination time and forwards data defining said different commencement and/or termination time to the services creation facility.

5 5. A telecommunications network as claimed in any preceding claim, further characterised in that data defining the resources which make up the computational platform and the times at which such resources are available is accessible by the control means and from the service creation facility such that any planned change in resource availability is used in the determination of whether the additional
10 service package can be provided.

6 A telecommunications network as claimed in any preceding claim, further characterised in that the data forwarded also specifies comparative priority of the additional service package whereby, if in operation one or more of the resources in
15 the computational platform becomes unavailable, the control means may re-allocate resources from service packages having a lower comparative priority to service packages having a higher comparative priority.

7. A telecommunications network as claimed in any preceding claim,
20 characterised in that the control means determines from the number of MIPS or computer cycles required to implement a service package whether the service package can be provided.

