



HAL
open science

Distributed Systems—Towards a Formal Approach

Gerard Le Lann

► **To cite this version:**

Gerard Le Lann. Distributed Systems—Towards a Formal Approach. 7th IFIP Congress 1977, Aug 1977, Toronto, Canada. pp.155-160. hal-03504338

HAL Id: hal-03504338

<https://inria.hal.science/hal-03504338>

Submitted on 29 Dec 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

DISTRIBUTED SYSTEMS—TOWARDS A FORMAL APPROACH

GÉRARD LE LANN
IRISA—Université de Rennes—BP 25 A
35 031 Rennes Cedex, France

Packet-switching computer communication networks are examples of distributed systems. With the large scale emergence of mini and micro-computers, it is now possible to design special or general purpose distributed systems. However, as new problems have to be solved, new techniques and algorithms must be devised to operate such distributed systems in a satisfactory manner. In this paper, basic characteristics of distributed systems are analysed and fundamental principles and definitions are given. It is shown that distributed systems are not just simple extensions of monolithic systems. Distributed control techniques used in some planned or existing systems are presented. Finally, a formal approach to these problems is illustrated by the study of a mutual exclusion scheme intended for a distributed environment.

1. INTRODUCTION

Computer communication networks using packet-switching technology provide for the interconnection of data-processing equipments of any kind. Such systems, sometimes simply referred to as computer networks, may be viewed as multi-macroprocessors whenever the goals of resource-sharing are achieved. With the large-scale emergence of mini and microcomputers, it is now possible to envision building general or special purpose multimini and multimicrocomputers to be operated in a non-centralized manner. The need for automatic resource-sharing arises here as in a similar way it does for multimacroprocessor systems.

Two kinds of resources must be considered :
- system resources, multi-accessed by users and for which multiplexing is required (hidden sharing)
- user resources, which users agree to share according to some protocol of their own (explicit sharing).

This paper discusses the problems of system resource-sharing in a distributed environment. An example of a user-sharing problem is distributed data-base sharing.

2. DISTRIBUTED SYSTEMS-ELEMENTS FOR A FORMAL APPROACH

Experimental and public packet-switching computer communication networks have been built and operated since 1968 ; examples are Arpanet, [7], Cyclades, [13], EIN, [1], Telenet, [17] and Datapac, [4]. The communication subnet of these networks is an example of a distributed system : all nodes have equal rights and responsibilities and no central "controller" is needed for the subnet to switch packets. Other examples are multicomputers like DCS [6] and Pluribus [8]. Finally, some multimicroprocessors currently planned by manufacturers will include "distributed features", in particular, those designed to have high availability characteristics.

A definition of what is meant by distributed system is needed. Then, analysis of the fundamental properties of computer systems makes it possible to tell whether or not a given system has distributed features.

2.1 Definitions

In a computer system, system resources are not accessed as such by users but through a set of services usually referred to as an operating system. Examples of services which we call operating functions are : communication, user scheduling, resource allocation, hardware resource handling, system data management, ... These functions are run through pro-

cesses called logical entities.

Let $F = \{f_i, i \in I\}$ be the set of the operating functions and $E_i = \{e_j^i, j \in J(i)\}$ be the set of entities participating in function f_i . At any instant t , it is possible to define $s_t(e_j^i)$ as the instantaneous state of entity e_j^i . It is therefore theoretically possible to define the global state of E_i at instant t as the vector $S_t(E_i) = \{\dots, s_t(e_j^i), \dots \text{ for all } j \in J(i)\}$.

A system will be said to be f_i -centralized if there exists $k \in J(i)$ such that $S_t(E_i)$ is known to e_k^i .

An example is a system in which $\dim(E_i) = 1$; another example is a system in which entities run the operating function f_i by using a common "system table".

A system will be said to be totally centralized if it is f_i -centralized for any $i \in I$.

In contrast, a system will be said to be f_i -distributed if there does not exist $k \in J(i)$ such that $S_t(E_i)$ is known to e_k^i .

A system will be said to be totally distributed if it is f_i -distributed for any $i \in I$.

Typical cases of distributed systems are systems in which cooperating entities do not share the same physical space and/or do not have a common time reference. In such systems, an entity may get either a partial and coherent view of the system or a complete but incoherent view of the system, coherence meaning that observations are made at the same moment in the system (absolute time). This absence of uniqueness both in time and space has very important consequences.

2.2 Time and space

It is well known in Physics that the sentence "event E occurred at time t " is meaningless if there is no indication about which time reference is used. Similarly, the statement "I am in location l " has no meaning as long as a space reference and a time reference have not been defined. Then, to the purpose of describing the behaviour of an entity, we will use a precise time-space reference where internal states, time lengths, names and instants may be defined.

We define the absolute Reference as the ideal reference such that speed of communication between this reference and any other time-space reference is infinite and where every space location may be given a unique name.

(i) System properties

An entity may be viewed as a finite-state automaton; a decision to switch to a new state is possible through the observation of a specific sequence of events received during a time period $(t-a, t)$, measured in the local time reference.

property Q : a is a finite value

property M : there are several time-space references for entities in the system

All systems studied here are assumed to have properties Q and M.

The propagation delay between two entities is the time needed, as measured in the absolute Reference, to transmit an elementary signal from one entity to the other.

property P₁ : for any pair of entities, propagation delays are fixed, finite and known with absolute accuracy ; they may be different for each pair

property P₂ : propagation delays are variable, finite and their values are not known with absolute accuracy

property P₃ : propagation delays are variable, finite and known a posteriori with absolute accuracy

property P₄ : propagation delays are variable but their values are upper bounded.

We should mention that these properties are common to all systems that are spacially distributed with finite propagation delays, including, for example, conventional logic design. Formalization of these properties was felt necessary so as to infer from them, basic principles which should be useful to distributed system designers.

(ii) Classification

- Systems with properties Q, M and P₁ will be called Perfect Multireference Systems (PMS)
- Systems with properties Q, M and P₃ will be called Quasi-Perfect Multireference Systems (QPMS)
- Systems with properties Q, M and P₂ will be called Multireference Systems (MS).

Let V be a sequence of events occurring within a system ; let E be the set of entities observing V. Events may be observed in two ways : referenced within a time-space reference R_i, i ∈ I, with I being the set of the time-space references of E or referenced within the absolute Reference. It may be interesting to consider the following problems :

- (a) is it possible to build in R_i, for any i ∈ I, the absolute chronological ordering of events (as observed in the absolute Reference) ?
- (b) do all the entities of E observe identically the set of events V ?

Answers to these questions are given in table 1.

Table 1

	PMS	QPMS	MS
(a)	YES	YES	NO
(b)	YES*	NO	NO

* if the value a (property Q) is the same for all entities in the system.

A graphical representation of properties P₂, P₃ and P₄ may help to answer questions (a) and (b) ; an example is given in fig. 1.

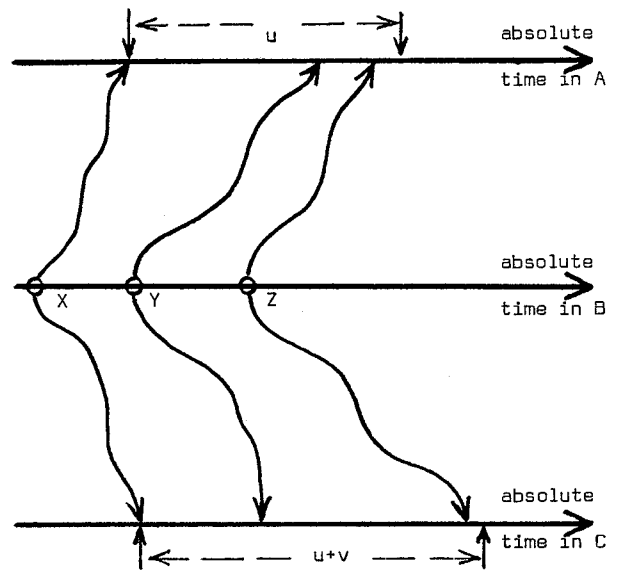


Fig. 1

In this example, three time-space references are represented. Three events X, Y, Z originated in B must be reported to A and C. It is easy to see that there may exist cases for which A, B and C will not be in agreement ; for instance, C may miss the observation of event Z, whatever the finite value of $u+v$. Thus, if time is the only dimension used to achieve cooperation between entities, systems having property P₄ are identical to Multireference Systems.

(iii) Principles for non-perfect systems (QPMS and MS)

(a) Principle of time nondeterminancy

For any given sequence of events, it is impossible to prove that two different entities will observe identically this sequence of events. An example of an event is a change of internal state ; as these changes are observable from the outside only through communication, one may state another principle ;

(b) Principle of relativistic observation

In a non-perfect system, it is impossible to prove that any two entities will have the same global view of a given subset of the system.

As a consequence, the environment of any entity cannot perceive the pair $(t, e(t))$ with an absolute certainty, t being the entity local time and $e(t)$ the current internal state of the entity. It will be possible either to know $e(t)$ and to state that the entity will reach that state in a predefined time interval Δt with some probability or to pick up an instant t and not being able to associate with certainty a given state $e(t)$. Then we have the following principle :

(c) Principle of state nondeterminancy

The instantaneous state of an entity may only be expressed in terms of possible values associated with some probabilities.

This means that the global state of a non-perfect system does not exist. As a consequence, according to the definition, these systems are totally distri-

vector $V = \{1, 1\}$ with their neighbours, 1 being the minimum value of the loads most recently received by the entity and i being the identity of the corresponding processor. Then, at any moment, upon receiving an external request, an entity is able to route it immediately to the less loaded processor. Stability conditions must be computed according to hardware performances and processing requirements.

(ii) Circulating vector technique

A successor is defined for each entity, such that all entities are located on a virtual ring. On this ring, one or several load vectors circulate, the dimension of the vectors being equal to the number of entities. The individual algorithm simply consists of each entity updating its own component with the current load value upon receiving a vector, recording a copy of it and transmitting this vector on to the ring. Notice that loss of a vector or creation of several vectors is not catastrophic to the system.

The efficiency of these techniques has been evaluated by means of a simulation model. Some results may be found in [10]. These mechanisms may be used as they stand to distribute the load evenly in a system or they may be used in connection with some other mechanisms when it is necessary to take locality constraints into account.

3.3 Distributed allocation of resources

The problem is the following one : U-entities (users) must be allocated R-entities (resources) ; specific entities are in charge of multiplexing several U-entities and performing the resource allocation (in a system described somewhere else [9], these entities are called controllers). A communication subsystem is used by the controllers to send their requests directly to R-entities ; how should deadlocks be avoided ? From several techniques, the circulating control token scheme is now presented.

For every controller, a successor is defined such that controllers are located on a virtual ring. One representation of a n -controller ring may be $\{i \rightarrow i+1, \text{modulo } n, i \in \{0, n-1\}\}$, each integer i being the identity of a controller. Asynchronous and natural time division is achieved by the means of a unique control token circulating on this virtual ring ; a controller is allowed to send allocation requests only when it owns the control token ; R-entities are provided with waiting files in which requests are recorded, up to a pre-defined limit (congestion control). When all requests have been answered, the control token is transmitted to the successor on the ring ; later, the U-entity will receive a message from each requested R-entity indicating that it has now moved to the first position in the file and that utilization of the resource is allowed. The average number of R-entities to be requested at a time is not independent of the circulating speed of the control token and it influences directly the system-wide job throughput. Extensive simulation described in [12] has been performed to evaluate the performances of this technique.

This is an example of a technique which must be shown to survive failures ; of major concern is the guarantee that the control token (CT) is never lost and that there is only one token on the ring. A protocol fulfilling this requirement exists and is now presented.

4. A SECURE PROTOCOL TO ACHIEVE MUTUAL EXCLUSION IN A DISTRIBUTED SYSTEM

We will discuss problems related mainly to controller failures such as what we should do when the controller which owns the CT goes down and thus removes the CT from the ring ?

4.1 Ring failures

First, assume that an error control mechanism based on "life messages", [7, 9], is provided at the hardware level which allows for the virtual ring reconfiguration. Then, controller i may be temporarily excluded from the ring, the successor of controller $i-1$ being controller $i+1$.

Second, let us briefly discuss problems related to failures of the interconnection structure (unibus, multibus, digital loop, multidrop telephone line, radio/satellite communication channel, matrix switch, store-and-forward networks,...). Transmission errors are easily recovered by using a simple mechanism like the Window technique. If the structure does not provide for more than one physical path between any pair of controllers, then it is of no help to design a failure tolerant distributed protocol ; if the structure does so, then failure of a structure subset can be controlled and recovered by using well known techniques like adaptive routing or alternate fixed routing.

4.2 The protocol

The protocol consists of a precedence rule and an election phase algorithm.

(i) Hypothesis

- controllers identities are integer values : $\{0, n-1\}$ for n controllers (H1)
- controllers may access the header of messages circulating on the ring
- the CT and other tokens are empty messages (only the header)
- each controller owns a timer ; this timer is reset at each CT occurrence
- the system is asynchronous ; timer values are not necessarily identical ; if they are, no consequence can be drawn from this (principle of time nondeterminancy) ; each controller is provided with its own time clock (H2)
- the sequence of messages on the ring is unchanged, i.e. messages received by a controller are retransmitted FIFO
- creation of a CT is an instantaneous action
- when timer awakes, the controller generates instantaneously a token carrying its own identity ; this token is candidate for being the new CT

(ii) Precedence rule

A controller which has generated a token and which receives the CT before its own token has completed a period must remove this token from the ring.

Indeed, "early" generation of a token may be due to large round trip times for the CT, short timer values and so forth. In any case, as a CT is circulating on the ring, there is no need for local action.

(iii) Election phase algorithm

The problem is to design an algorithm such that one can prove that, when the control token (CT) is lost, there is a unique controller to be elected as responsible for regenerating a new CT (constraint A), in a finite time delay (constraint B).

Let I be the set of controllers participating in the election i.e. controllers for which timers awake between the loss of CT and regeneration of a new CT ; let $S(i)$, $i \in I$, be the set of tokens identities as recorded by controller i after a complete rotation of token i ; obviously, one of these identities will be i itself.

Uniqueness in the choice for several controllers is guaranteed if :

condition (a) : the algorithm is unique for all controllers
 condition (b) : value of $S(i)$, $i \in I$ is the same for all controllers.

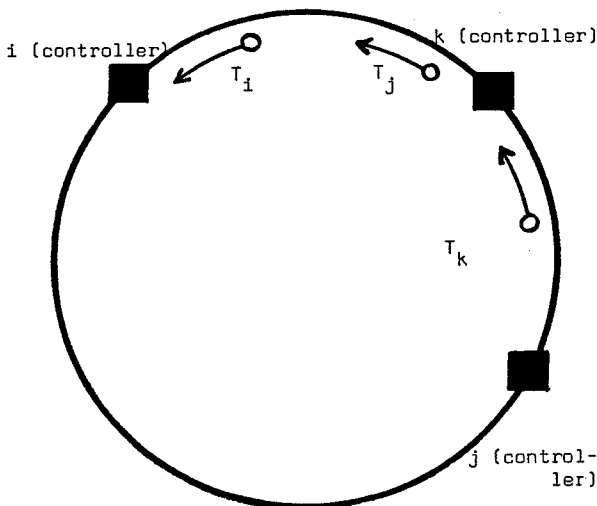
Unfortunately, there are cases for which condition (b) is not true, see fig. 2 ; T_j being generated after T_i crossed controller j and before T_k crossed the same controller, one is left with a situation where $S(i) = \{i,k\}$ and $S(k) = \{i,j,k\}$.

One may be tempted to solve the problem by using one of these solutions :

Solution 1 : for each token crossing a controller, the local timer is reset to its initial value ; a new token is generated locally only when the timer awakes.

Solution 2 : solution 1 plus : each controller must remove from the ring any token circulating after the first token so that only that one will be left on the ring.

It should be noted that in this case, all controllers, even those not participating in the election (which means permanently), are required to record crossing of any token. It is not difficult to conclude that these solutions are not acceptable, because of H2 and condition (b).



T_x : token of controller x
 T_i and T_k have completed a revolution

Fig. 2

As the set of controllers is strictly ordered (H1), a simple algorithm may be proposed :

Ω : if $i = \min S(i)$, then entity i immediately generates the new CT

In order to demonstrate that Ω satisfies constraint A, let us describe first the state-transition table of entity i on the ring (table 2).

Table 2

external events state (i)	0	1	2	3	4
α	β	α	α	α	α
β	β	α	γ	β	α^*
γ	β	α	γ	γ	α

Comments :

- 0 : awaking of the timer
- 1 : reception of the control token
- 2 : reception of a candidate token, the identity of which is smaller than i
- 3 : reception of a candidate token, the identity of which is larger than i
- 4 : reception of the candidate token i (after one complete revolution)
- α : idle, control token timer is set
- β : candidate token timer is set and i is prepared to regenerate a new control token
- γ : candidate token timer is set and i is not responsible for the control token regeneration
- α^* : generation of the new control token and immediate switching to state α

We will use the following notation :

- $I(CT,x)$ = instant of control token reception by entity x
- $I(t(x),y)$ = instant of reception of the candidate token x by entity y
- $I(x,o)$ = instant of generation of a candidate token by entity x
- $I(x,x)$ = occurrence of event 4.

Let us imagine that two entities x and y generate "simultaneously" (during the same revolution on the ring) a control token, thus violating constraint A and we will show that this situation is impossible. Let us assume, for example, that $identity(x) < identity(y)$.

Entity y will generate a new token if and only if state (y) at time $I(y,y)$ is β ; this implies : $\bar{1}$ and $\bar{2}$ between $I(y,o)$ and $I(y,y)$ where \bar{n} means non occurrence of event n .

Identically, assuming that x will generate a new token implies $(x < y) : \bar{1}$ between $I(x,o)$ and $I(x,x)$.

It is easy to show that a subset of these conditions leads to a contradiction.

$\bar{2}$ between $I(y,o)$ and $I(y,y)$ $I(t(x),y) > I(y,y)$ for entity y ; $\bar{1}$ between $I(x,o)$ and $I(x,x)$ $I(CT,x) > I(x,x)$ with the CT received by x being the token generated by y . This constraint and the FIFO hypothesis imply that for entity $y : I(t(x),y) < I(y,y)$.

It has thus been demonstrated that the CT cannot be generated by two different entities during one revolution on the ring.

Constraint B is obviously fulfilled by Ω and it is possible to compute bound values for the time T needed to regenerate a new CT. If x is the identity of the first controller to initiate the election phase after loss of the CT and if θ is the maximum value of the time required for a controller to process a token and to hand it down to its neighbour on the ring, then we have :

$$R \leq T \leq x(R - \theta) + \theta$$

T being counted from the instant of timeout for controller x .

(iv) Failures during the election phase

When considering the failure of a controller participating in the election phase, two problems must be tackled. Failure of the controller which is precisely the one being elected by the other controllers as responsible for generating the new CT is not catastrophic ; protection against infinite waiting is provided by timers ; the election phase will only be longer than for a failure-free situation.

The other problem is what to do with tokens generated by controllers which have failed before tokens have completed a period. One protocol may be that only controller i is allowed to remove T_i from the ring.

This is acceptable provided that failures are either exceptional or of short duration. Otherwise, other controllers must be allowed to destroy tokens pertaining to dead controllers ; for instance, the last elected controller may be responsible for this.

Another solution is based on mutual help and mutual suspicion principles being applied in the whole system.

When a controller failure is detected, its neighbours help to "clean" the situation. One of the actions to be taken could be precisely to withdraw from the ring all tokens and messages generated by this failing controller. These actions are then dependant upon an error control mechanism situated at another logical level in the system.

Problems related to ring reconfiguration after a failure, reintegration of a controller on the ring as well as a second protocol achieving mutual exclusion are analysed in detail in [11].

We should mention the practical utility of the election protocol. Controllers are autonomous not only while the system is running but also at the initialization phase. No external action is needed as controllers will undertake spontaneously an election when the system is turned on. In this sense, this approach is different from the one described in [5] as here there is no need for a stabilization to be achieved after initialization.

5. CONCLUSION

Because of hardware technology trends, distributed systems are receiving more and more attention. Importantly, this kind of system seems to fulfill user needs more satisfactorily and more easily than conventional and centralized systems : as processors may migrate, people do not have to, fully modular systems are easier to maintain, to expand and so forth. In this paper, an attempt has been made to clarify the concept of distributed system ; the nature of such systems has been analysed, definitions and design principles have been given and specific techniques have been presented and discussed.

ACKNOWLEDGMENTS

Thanks are due to the referees for their comments and to Sessori, Ministère de l'Industrie et de la Recherche, for its financial support.

REFERENCES

- [1] D.L.A. Barber, A European Informatics Network : achievement and prospects, Third ICC, Toronto, Aug. 1976, 44-50.
- [2] V.G. Cerf and R. Kahn, A protocol for packet network intercommunication, IEEE Trans. on Communications, vol. COM-22, issue 5, May 1974, 637-648.
- [3] J.F. Chambon et al., Spécifications fonctionnelles des stations de transport du Réseau Cyclades, IRIA Technical Report SCH 502.3, May 1973, 97 p. (French).
- [4] W.W. Clipsham et al., Datapac network overview, Third ICC, Toronto, Aug. 1976, 131-136.
- [5] E.W. Dijkstra, Self-stabilizing systems in spite of distributed control, Com. of the ACM, vol. 17, issue 11, Nov. 1974, 643-644.
- [6] D.J. Farber et al., The Distributed Computing System, Seventh Annual IEEE Computer Society Int. Conf., San Francisco, Feb. 1973, 31-34.
- [7] F.E. Heart et al., The interface message processor for the Arpa computer network, AFIPS, SJCC 1970, vol. 36, 551-576.
- [8] F.E. Heart et al., A new minicomputer/multi-processor for the Arpa network, NCC 1973, vol. 42, 529-537.
- [9] G. Le Lann and R. Negaret, Operating principles for a distributed multimicroprocessor, First European Symp. on Microarchitecture of Computer Systems, North-Holland, Nice, June 1975, 219-222.
- [10] G. Le Lann et al., Distribution of access and data in large data bases, Int. Symp. on Technology for Selective Dissemination of Information, IEEE, San-Marino, Sept. 1976, 94-98.
- [11] G. Le Lann, Introduction à l'analyse des systèmes multiréférentiels, Thèse de Doctorat d'Etat, University of Rennes, May 1977, 180 p. (French).
- [12] R. Negaret, Etude de l'allocation de ressources dans les systèmes informatiques répartis, Thèse de Doctorat de 3è cycle, University of Rennes, Dec. 1976, 190 p. (French).
- [13] L. Pouzin, Presentation and major design aspects of the Cyclades computer network, Third ACM/IEEE Data Communication Symp., Tampa, Nov. 1973, 80-85.
- [14] L. Pouzin, Distributed congestion control in a packet network : the channel load limiter, INWG Protocol Note 36, June 1976, 6 p.
- [15] L. Pouzin, Flow control in data networks-methods and tools, Third ICC, Toronto, Aug. 1976, 467-474.
- [16] W.L. Price, Simulation of packet-switching networks controlled on isarithmic principles, Third ACM/IEEE Data Communication Symp., Tampa, Nov. 1973, 44-49.
- [17] L.G. Roberts, Telenet principles and practice, On Line Symp. on Communications Networks, London, Sept. 1975, 315-329.
- [18] D.L. Russell and T.H. Bredt, Error resynchronization in producer-consumer systems, Fifth ACM Symp. on Operating Systems Principles, Austin, Nov. 1975, 106-113.
- [19] H. Zimmermann, The Cyclades end-to-end protocol, Fourth ACM/IEEE Data Communication Symp., Quebec City, Oct. 1975, 7.21-7.26.