Exclusion mutuelle distribuée, algorithmes arborescents et protocoles Ethernels

Gérard Le Lann
INRIA – IMARA
Gerard.Le_Lann@inria.fr

Je n'aime pas les gens qui parlent pendant que je les interromps …

Exclusion mutuelle distribuée non bloquante : hypothèses

Soit un jeu à n participants, noms uniques ∈ [1, n], n > 1, qui doivent utiliser un objet partagé C. Un joueur dispose de deux actions sur C :

❖ A, activée pour tenter d'acquérir C (la durée d'utilisation de C est finie bornée)

❖ L, pour libérer C après utilisation.

C peut prendre 3 états, 0 (libre), 1 (utilisé), ∇(conflit), notés e(C).

Une action A exécutée par j à t produit l'une des réponses suivantes :

❒ **Gagné** (C est alloué à j) :

• e(C) = 0 à t, j est le seul demandeur || **e(C) : 0 ⇒ 1**

❒ **Perdu** (C n'est pas alloué à j) :

• e(C) = 1 à t || **e(C) reste à 1** (jusqu'à L par k ≠ j, et alors e(C) ⇐ 0)

❒ **Perdu** (C n'est pas alloué à j) :

• e(C) = 0 à t, j n'est pas le seul demandeur à t || **e(C) : 0 ⇒ ∇** || e(C) = ∇ pendant une durée σ, après quoi e(C) ⇐ 0

• e(C) = ∇ à t || **e(C) reste à ∇**

Les joueurs observent les transitions d'état et les états de C, mais ne peuvent se concerter. A activable librement. Tout joueur peut s'arrêter à tout moment.

Exclusion mutuelle distribuée non bloquante : propriétés

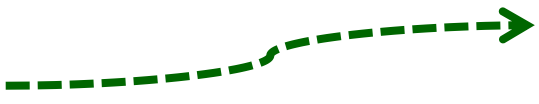
- Sûreté logique : à tout instant, au plus 1 joueur utilise C
- Vivacité : tout joueur qui active A se voit allouer C en un temps fini*
- Non oisiveté : e(C) ≠ 0 si au moins 1 joueur a activé A et C ne lui est pas encore alloué.

*Un joueur qui utilise C peut ne pas exécuter L en un temps fini (arrêt accidentel).

Problème classique (Dijkstra, …, 60's) en version centralisée : exclusion mutuelle pour entrer en section critique (programmes concurrents).
Solution (*sucre syntaxique*) : variable *mutex*, virtualisation de C, e(*mutex*) = {0, 1}

Depuis les 70's, nombreuses déclinaisons du problème (et des solutions) en version distribuée :

- Sémaphores
- Estampilles
- Jeton circulant
- Tris arborescents



© IFIP, NORTH-HOLLAND PUBLISHING COMPANY (1977)
https://www.rocq.inria.fr/novaltis/.../IFIP%20Congress%201977.pdf
Format de fichier: PDF/Adobe Acrobat - Afficher
de G LE LANN - 1977 - Cité 428 fois - Autres articles
GERARD *LE LANN*. IRISA—Université dc ... problems is illustrated by the study of a *mutual exclusion* scheme intended for a *distributed* envi- ronment. 1 .

Tris arborescents // Tree searches

Buts : en cas de conflit, trier les joueurs permet de leur affecter des instants d'activation à nouveau de A qui sont différents (élimination de la concomitance)
Principe : isoler des sous-arbres, chacun d'entre eux contenant au plus 1 joueur actif

Disciplines : théorie de l'information, structures de données, algorithmique distribuée, modélisation/calcul analytique, analyse combinatoire, …

- Algorithmes probabilistes

Les joueurs choisissent leurs parcours selon une même loi aléatoire ➔ les parcours peuvent ne pas différer (probabilité calculable) ➔ probabilité non nulle de violation de la propriété de vivacité (a fortiori, pas de borne de terminaison)

- Algorithmes déterministes

Les joueurs choisissent leurs parcours selon une même loi déterministe ➔ les parcours finissent sûrement par différer ➔ terminaison certaine en temps fini borné

- Analyses comportementales // performances (débits, délais)

Stochastiques : moments de distribution (espérances, écart-types, …), comportements asymptotiques
Pires cas : bornes inf (débits), bornes sup (délais)

```
FUNCTION GetQuorum (Tree: NetworkHierarchy): QuorumSet;
  VAR left, right : QuorumSet;
  BEGIN
  IF Empty (Tree) THEN
       RETURN ({});
  ELSE IF GrantsPermission(Tree↑.Node) THEN
                RETURN ((Tree↑.Node) ∪ GetQuorum (Tree↑.LeftChild));
                OR
                RETURN ((Tree↑.Node) ∪ GetQuorum (Tree↑.RightChild));(*line 9*)
          ELSE
                left←GetQuorum(Tree↑.left);
                right←GetQuorum(Tree↑.right);
                IF (left = ∅ ∨ right = ∅) THEN
                     (* Unsuccessful in establishing a quorum *)
                      EXIT(-1);
                ELSE
                     RETURN (left ∪ right);
                END; (* IF *)
          END; (* IF *)
  END; (* IF *)
END GetQuorum
```

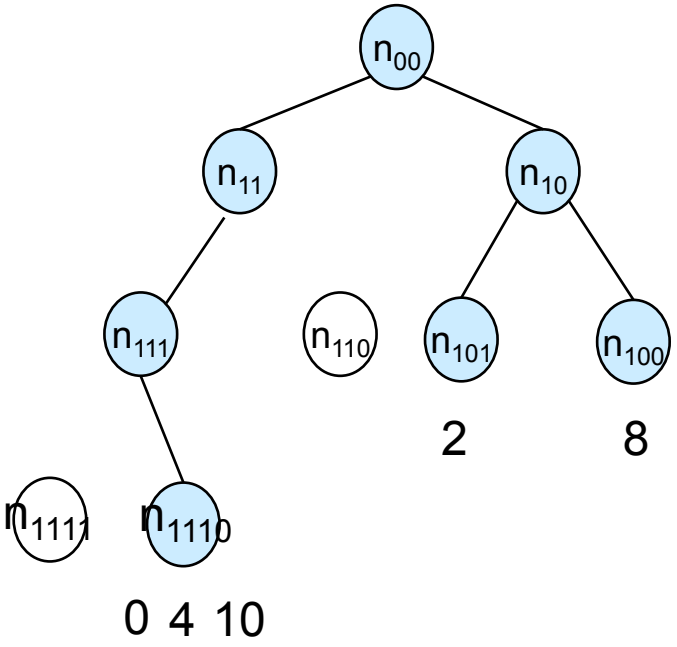
Figure: Algorithm for constructing a tree-structured quorum.

Preuves de propriétés (cf. # 3) requises pour toute solution/algorithme

(« home work »)

Exemple du tri binaire – Algorithme probabiliste

A chaque réponse « perdu », tirage aléatoire sur {0, 1}
0 ➔ tenter A à nouveau à t = transition e(C) ⇐ 0
1 ➔ tenter A à nouveau à t + σ et attendre transition e(C) ⇐ 0



1, n_{00} : e(C) = 0 à t. les joueurs **0, 2, 4, 8, 10** font A à t || e(C) ⇐ ∇ pendant σ ; les joueurs **2, 8** tirent 0, les autres tirent 1

2, n_{10} : quand e(C) ⇐ 0, les joueurs **2, 8** font A || e(C) ⇐ ∇ pendant σ ; le joueur **8** tire 0, le joueur **2** tire 1

3, n_{11} : quand e(C) ⇐ 0, les joueurs **0, 4, 10** font A || e(C) ⇐ ∇ pendant σ ; ils tirent 1

4, n_{100} : quand e(C) ⇐ 0, le joueur **8** fait A || e(C) ⇐ 1, **8** acquiert C, puis fait L, e(C) ⇐ 0

5, n_{101} : quand e(C) ⇐ 0, le joueur **2** fait A || e(C) ⇐ 1, **2** acquiert C, puis fait L, e(C) ⇐ 0

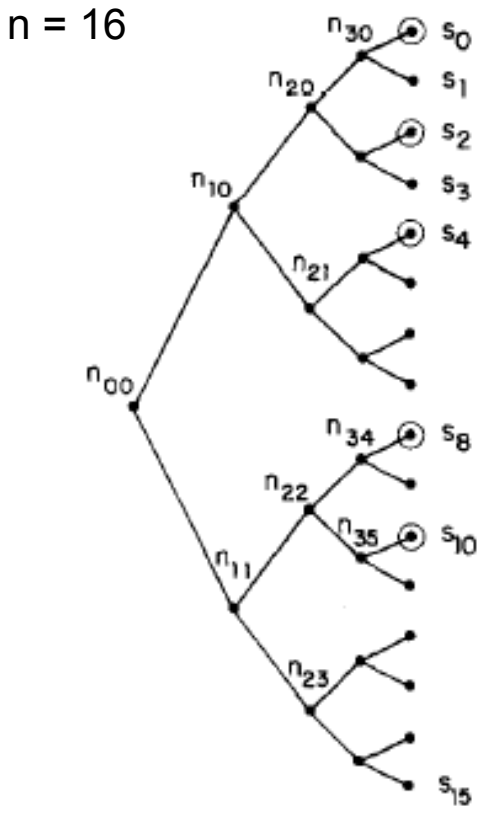
6, n_{110} : e(C) reste à 0

7, n_{111} : σ après transition e(C) 1 ⇒ 0, les joueurs **0, 4, 10** font A || e(C) ⇐ ∇ pendant σ ; ils tirent 0

…/…

Exemple du tri binaire – Algorithme déterministe

A chaque réponse « perdu », ensemble des noms dichotomisé (itératif) :
j ∈ moitié basse ➔ tenter A à nouveau à t = transition e(C) ⇐ 0
j ∈ moitié haute ➔ tenter A à nouveau à t + σ et attendre transition e(C) ⇐ 0



1, n_{00} : e(C) = 0 à t, les joueurs **0, 2, 4, 8, 10** font A à t || e(C) ⇐ ∇ pendant σ

2, n_{10} : quand e(C) ⇐ 0, les joueurs **0, 2, 4** font A || e(C) ⇐ ∇ pendant σ

3, n_{11} : quand e(C) ⇐ 0, les joueurs **8, 10** font A || e(C) ⇐ ∇ pendant σ

4, n_{20} : quand e(C) ⇐ 0, les joueurs **0, 2** font A || e(C) ⇐ ∇ pendant σ

5, n_{21} : quand e(C) ⇐ 0, le joueur **4** fait A || e(C) ⇐ 1, **4** acquiert C, puis fait L, e(C) ⇐ 0

6, n_{30} : quand e(C) ⇐ 0, le joueur **0** fait A || e(C) ⇐ 1, **0** acquiert C, puis fait L, e(C) ⇐ 0

…/…

Tri binaire déterministe – Analyse markovienne des délais

δ: delay for completing a tree search (acquiring C) as a function of offered load λ (activation rate of A)

$$E\{\delta\} \leqslant \frac{6.05\lambda(c-2.88\lambda)}{1-(2.88\lambda)^2} + \frac{1.05(c-2.88\lambda)^2}{(1-(2.88\lambda)^2\,\overline{I}_l(\lambda)} + 0.321$$

$$E\{\delta\} \geqslant \max\left[\frac{0.72\lambda}{1-(2.88\lambda)^2} + \frac{1}{2}\,\overline{I}_l(\lambda),\ \overline{I}_l(\lambda)\right].$$

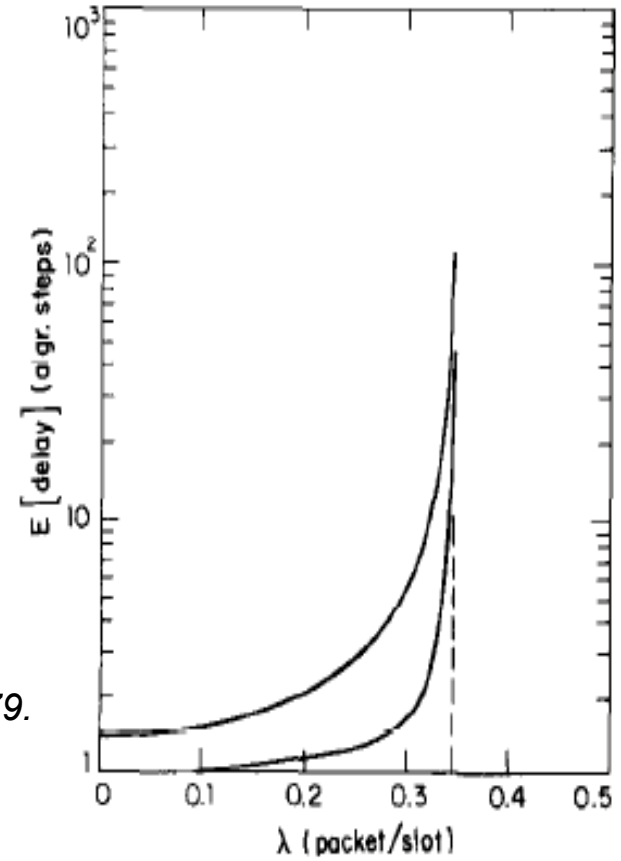


Fig. 5. Upper and lower bounds to average delay versus average arrival rate for binary tree/Poisson source system. Note: An algorithmic step equals one round-trip delay plus two slots.

J. I. Capetanakis, "*Tree algorithms for packet broadcast channels*," *IEEE Trans. Inform. Theory, vol. IT-25, 505-515, 1979.*

B. S. Tsybakov and V. A. Mikhailov, *"Free synchronous packet access in a broadcast channel with feedback*," *Prubl. Inform. Transmission, vol.14*, pp. 259-280, 1979.

Tri binaire déterministe – Analyse pires cas des délais

(analyse combinatoire)

$$\xi_k^t = \begin{cases} \frac{m^{\left\lceil log_m\left(m\left\lfloor\frac{k}{2}\right\rfloor\right)\right\rceil}-1}{m-1} + m\left\lfloor\frac{k}{2}\right\rfloor\left\lfloor log_m\left(\frac{t}{m\left\lfloor\frac{k}{2}\right\rfloor}\right)\right\rfloor - \left(k - m\left\lfloor\frac{k}{2}\right\rfloor\right) & \text{si } k \in \{2,\ldots,t\}, \\ 0 & \text{si } k = 1, \\ 1 & \text{si } k = 0. \end{cases}$$

$$t = m^n,\ m \in \mathbb{N}^* \setminus \{1\},\ n \in \mathbb{N}^*.$$

$\sigma\ \xi^t_k$: worst-case delay for isolating k leaves (completing a tree search // acquiring C) in a t-leaf balanced m-ary tree

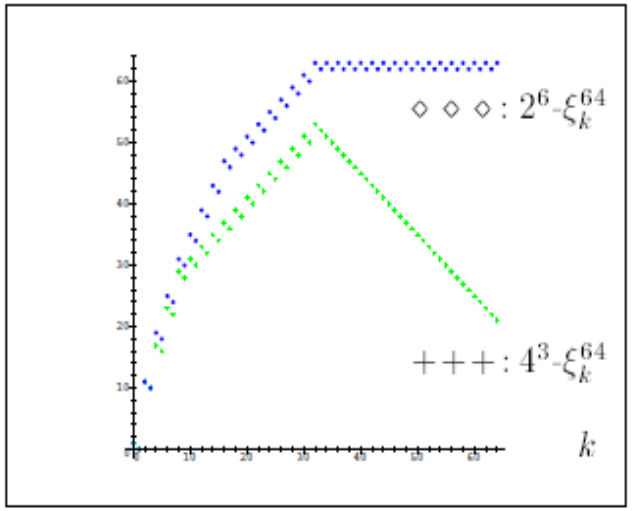


Fig. 2: Worst-case search times for 64-leaf balanced binary and quaternary trees

J.-F. Hermant, G. Le Lann, "A protocol and Correctness Proofs for Real-Time High-Performance Broadcast Networks", 18th IEEE International Conference on Distributed Computing Systems (ICDCS 98), Amsterdam, The Netherlands, 26-29 May 1998, pp. 360-369.

Exclusion mutuelle distribuée et canaux de communication en accès multiple

La grande famille des protocoles CSMA

70's Palo Alto

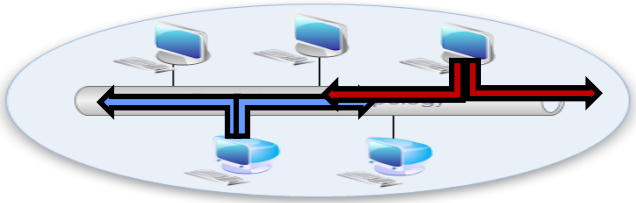
CSMA/CD et la naissance de Ethernet (probabliste)

Messages transmis sur câbles bidirectionnels

CSMA: Carrier Sense Multiple Access
CA: collision avoidance
CD: collision detection
CR: collision resolution

Action A : carrier sense (CS)*
si e(C) ≠ 0 alors attendre e(C) ⇐ 0
si e(C) = 0 alors commencer émission
et poursuivre CS
tantque e(C) ≠ ∇ poursuivre/terminer émission
sinon arrêter et faire BEB(d_a)

*test de porteuse



$n \leq 1024$

BEB: binary exponential backoff
= tri binaire probabiliste

CD correspond à e(C) = ∇. BEB (d_a) ? Tirage uniforme d'entiers sur l'intervalle $[0, 2^a-1]$, a = nb de tentatives A infructueuses ➔ entier b ➔ $d_a = b\,\sigma$
➔ choix aléatoire de l'une parmi 2^a feuilles d'un arbre binaire

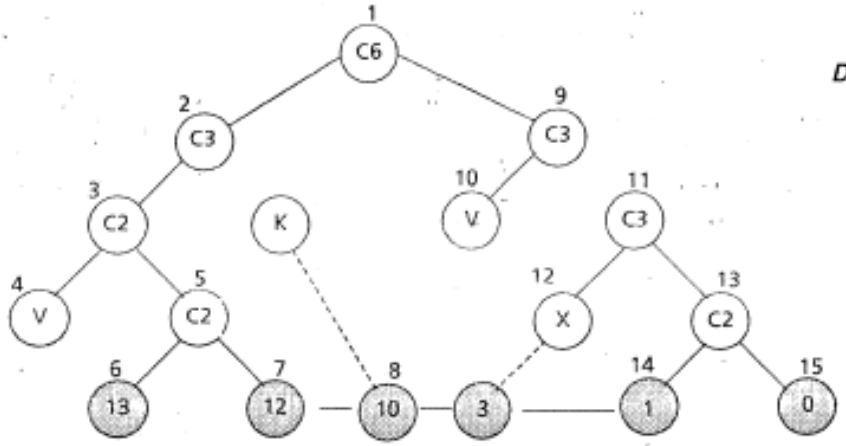
Exclusion mutuelle distribuée et canaux de communication en accès multiple

La grande famille des protocoles CSMA

80's Rocquencourt

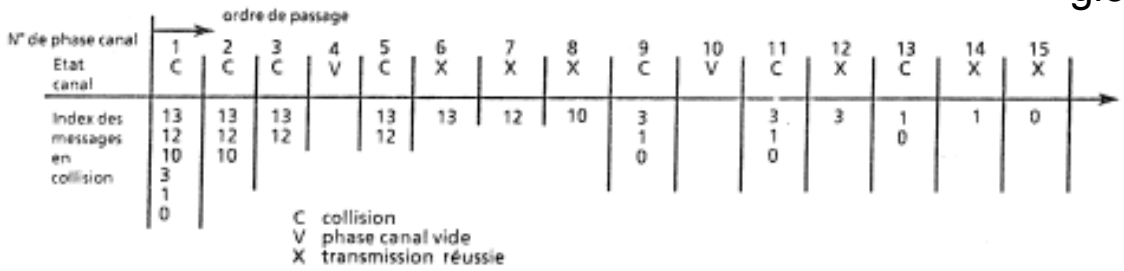
CSMA/CD/DCR et la naissance de Ethernet déterministe

CSMA: Carrier Sense Multiple Access
CA: collision avoidance
CD: collision detection
CR: collision resolution



les feuilles contiennent les valeurs d'index associés aux messages transmis

Figure 1
Diagramme temporel de la résolution de collision (mode général)



DCR: deterministic collision resolution = tri binaire déterministe

Exemple numérique

σ: ch_slot time = 40 µs

durée utilisation C : durée message = 460 µs

➔ taux utilisation = 0,92 avec taux global d'arrivées = 2 messages/ms

Les protocoles CSMA à tris arborescents sont non bloquants

On vérifie aisément qu'un joueur k défaillant (qui s'arrête) ne peut bloquer le jeu.

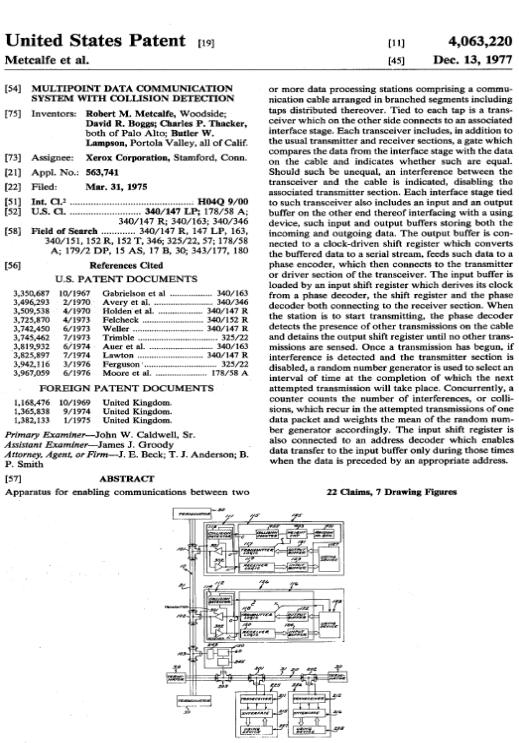
- Soit k est inactif ➔ « too bad » for k
- Soit k est actif, mais n'a pas encore acquis C ➔ « too bad » for k
- Soit k est actif et possède C

  ➔ sur défaillance de k, transition d'état de(C) : 1 $\Rightarrow$ 0 (arrêt de transmission)

  ➔ les autres joueurs détectent la libération de C, et poursuivent le jeu (comme si k avait fait L)

Cette propriété n'est pas obtenue aussi aisément dans le cas de ressource C passive.

Les algorithmes à verrouillage/sémaphores un peu trop « simples » sont bloquants.



United States Patent [19]
Metcalfe et al.

[11] 4,063,220
[45] Dec. 13, 1977

[54] MULTIPOINT DATA COMMUNICATION SYSTEM WITH COLLISION DETECTION

[75] Inventors: Robert M. Metcalfe, Woodside; David R. Boggs; Charles P. Thacker, both of Palo Alto; Butler W. Lampson, Portola Valley, all of Calif.

[73] Assignee: Xerox Corporation, Stamford, Conn.

[21] Appl. No.: 563,741

[22] Filed: Mar. 31, 1975

[51] Int. Cl.² H04Q 9/00

[52] U.S. Cl. ..... 340/147 LP; 178/58 A; 340/147 R; 340/163; 340/346

[58] Field of Search ..... 340/147 R, 147 LP, 163, 340/151, 152 R, 152 T, 346; 325/22, 57; 178/58 A; 179/2 DP, 15 AS, 17 B, 30; 343/177, 180

[56] References Cited

U.S. PATENT DOCUMENTS

3,350,687	10/1967	Gabrielson et al.	340/163
3,496,293	2/1970	Avery et al.	340/346
3,509,538	4/1970	Holden et al.	340/147 R
3,725,870	4/1973	Felcheck	340/152 R
3,742,450	6/1973	Weller	340/147 R
3,745,462	7/1973	Trimble	325/22
3,819,932	6/1974	Auer et al.	340/163
3,825,897	7/1974	Lawton	340/147 R
3,942,116	3/1976	Ferguson	325/22
3,967,059	6/1976	Moore et al.	178/58 A

FOREIGN PATENT DOCUMENTS

1,168,476	10/1969	United Kingdom.
1,365,838	9/1974	United Kingdom.
1,382,133	1/1975	United Kingdom.

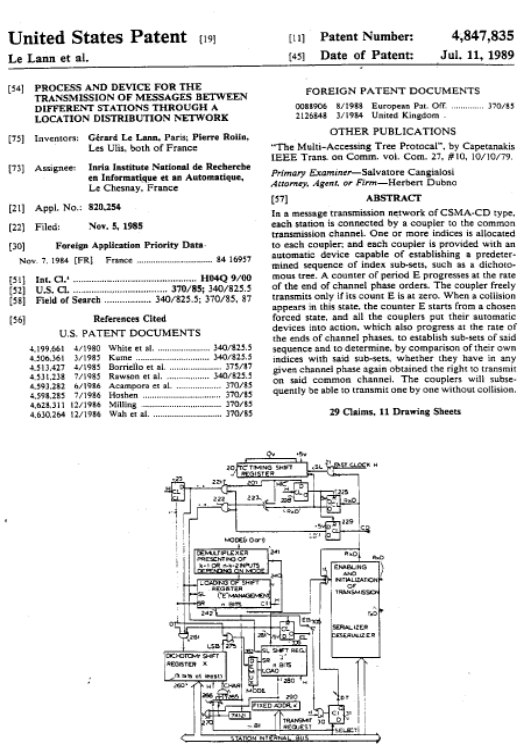
Primary Examiner—John W. Caldwell, Sr.
Assistant Examiner—James J. Groody
Attorney, Agent, or Firm—J. E. Beck; T. J. Anderson; B. P. Smith

[57] ABSTRACT

Apparatus for enabling communications between two or more data processing stations comprising a communication cable arranged in branched segments including taps distributed thereover. Tied to each tap is a transceiver which on the other side connects to an associated interface stage. Each transceiver includes, in addition to the usual transmitter and receiver sections, a gate which compares the data from the interface stage with the data on the cable and indicates whether such are equal. Should such be unequal, an interference between the transceiver and the cable is indicated, disabling the associated transmitter section. Each interface stage tied to such transceiver also includes an input and an output buffer on the other end thereof interfacing with a using device, such input and output buffers storing both the incoming and outgoing data. The output buffer is connected to a clock-driven shift register which converts the buffered data to a serial stream, feeds such data to a phase encoder, which then connects to the transmitter or driver section of the transceiver. The input buffer is loaded by an input shift register which derives its clock from a phase decoder, the shift register and the phase decoder both connecting to the receiver section. When the station is to start transmitting, the phase decoder detects the presence of other transmissions on the cable and detains the output shift register until no other transmissions are sensed. Once a transmission has begun, if interference is detected and the transmitter section is disabled, a random number generator is used to select an interval of time at the completion of which the next attempted transmission will take place. Concurrently, a counter counts the number of interferences, or collisions, which recur in the attempted transmissions of one data packet and weights the mean of the random number generator accordingly. The input shift register is also connected to an address decoder which enables data transfer to the input buffer only during those times when the data is preceded by an appropriate address.

22 Claims, 7 Drawing Figures

.../...



United States Patent [19]
Le Lann et al.

[11] Patent Number: 4,847,835
[45] Date of Patent: Jul. 11, 1989

[54] PROCESS AND DEVICE FOR THE TRANSMISSION OF MESSAGES BETWEEN DIFFERENT STATIONS THROUGH A LOCATION DISTRIBUTION NETWORK

[75] Inventors: Gérard Le Lann, Paris; Pierre Rolin, Les Ulis, both of France

[73] Assignee: Inria Institute National de Recherche en Informatique et an Automatique, Le Chesnay, France

[21] Appl. No.: 820,254

[22] Filed: Nov. 5, 1985

[30] Foreign Application Priority Data

Nov. 7, 1984 [FR] France 84 16957

[51] Int. Cl.⁴ H04Q 9/00

[52] U.S. Cl. 370/85; 340/825.5

[58] Field of Search 340/825.5; 370/85, 87

[56] References Cited

U.S. PATENT DOCUMENTS

4,199,661	4/1980	White et al.	340/825.5
4,506,361	3/1985	Kume	340/825.5
4,513,427	4/1985	Borriello et al.	375/87
4,531,238	7/1985	Rawsen et al.	340/825.5
4,593,282	6/1986	Acampora et al.	370/85
4,598,285	7/1986	Hoshen	370/85
4,628,311	12/1986	Milling	370/85
4,630,264	12/1986	Wah et al.	370/85

FOREIGN PATENT DOCUMENTS

0088906	8/1988	European Pat. Off. 370/85
2126848	3/1984	United Kingdom .

OTHER PUBLICATIONS

"The Multi-Accessing Tree Protocal", by Capetanakis IEEE Trans. on Comm. vol. Com. 27, #10, 10/10/79.

Primary Examiner—Salvatore Cangialosi
Attorney, Agent, or Firm—Herbert Dubno

[57] ABSTRACT

In a message transmission network of CSMA-CD type, each station is connected by a coupler to the common transmission channel. One or more indices is allocated to each coupler; and each coupler is provided with an automatic device capable of establishing a predetermined sequence of index sub-sets, such as a dichotomous tree. A counter of period E progresses at the rate of the end of channel phase orders. The coupler freely transmits only if its count E is at zero. When a collision appears in this state, the counter E starts from a chosen forced state, and all the couplers put their automatic devices into action, which also progress at the rate of the ends of channel phases, to establish sub-sets of said sequence and to determine, by comparison of their own indices with said sub-sets, whether they have in any given channel phase again obtained the right to transmit on said common channel. The couplers will subsequently be able to transmit one by one without collision.

29 Claims, 11 Drawing Sheets

Patented by INRIA at the request of the French Navy. Fielded in various places:

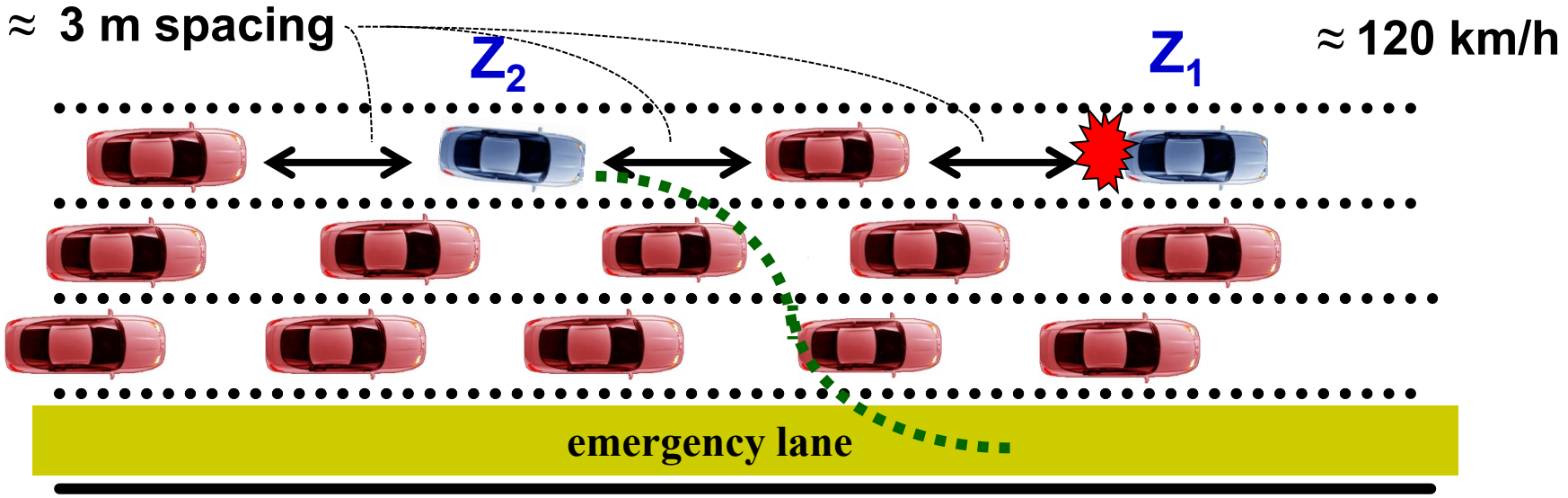
- *Nuclear aircraft carrier Charles-de-Gaulle*
- *Frigates and submarines*
- *Industrial sites (Saint-Gobain, …)*
- *European spatial base (Ariane launcher) in Kourou (French Guyana)*

Mais noms et n sont connus ➔ systèmes distribués « fermés »
Domaine « fermé » : les algorithmes distribués « classiques »

Si ensemble des joueurs/processus est ouvert
➔ noms ≡ inconnus, n ≡ inconnu

➔ Problèmes ouverts (domaine quasi-vierge) : algorithmes pour systèmes cyberphysiques !

Open problems: safety-critical inter-vehicular communications (ad hoc networks of driverless/automated cars)



shared contention-prone resource C: a multiaccess radio channel

How to avoid (collective) collisions whenever risk-prone maneuvers must be undertaken. Examples:

- Z_1 decelerates abruptly
- Z_2 needs to reach the emergency lane asap

radio → câbles → radio
La grande famille des protocoles CSMA

60's **Hawaï Alohanet**
Static / radio || CSMA

70's **Ethernet**
Static / cables
|| CSMA-CD

TCP/IP et la naissance de Internet
Vint Cerf
2012 : VP Google

Robert Metcalfe
2012 : Polaris Venture Partners et
Prof. of Innovations UTexas at Austin

…/…

Cyber Nets

80's **Sensor Nets**
Static / radio || CSMA-CA

Internet des Objets

90's **MANETs**
Mobile ad hoc nets
Mobile / radio || CSMA-CA

…/…

00's **VANETs**
Vehicular ad hoc nets
Mobile / radio || CSMA-CA

CyberPhysical Nets

Défis scientifiques (retombées concrètes majeures) : renommage des joueurs/processus ???, CSMA-??? déterministe, etc.