# Distributed algorithms The Leader Election Protocol (IEEE 1394)

J.R. Abrial and Dominique Cansell and Dominique Méry

26th IFIP WG 6.1 International Conference on Formal Methods for Networked and Distributed Systems September 26-29 2006, Paris, France

- Background :-)
- An informal presentation of the protocol :-)
- Step by step formal design :-
- Short Conclusion. :-)

#### **IEEE 1394 High Performance Serial Bus (FireWire)**

- It is an international standard
- There exists a widespread commercial interest in its correctness
- Sun, Apple, Philips, Microsoft, Sony, etc involved in its development
- Made of three layers (physical, link, transaction)
- The protocol under study is the Tree Identify Protocol
- Situated in the Bus Reset phase of the physical layer

## The Problem (1)

- The bus is used to transport digitized video and audio signals
- It is "hot-pluggable"
- Devices and peripherals can be added and removed at any time
- Such changes are followed by a bus reset
- The leader election takes place after a bus reset in the network
- A leader needs to be chosen to act as the manager of the bus

- After a bus reset: all nodes in the network have equal status
- A node only knows to which nodes it is directly connected
- The network is connected
- The network is acyclic

#### BASIC

- IEEE. *IEEE Standard for a High Performance Serial Bus. Std 1394-1995*. 1995

- IEEE. IEEE Standard for a High Performance Serial Bus (supplement). Std 1394a-2000. 2000

#### GENERAL

- N. Lynch. *Distributed Algorithms*. Morgan Kaufmann. 1996

- R. G. Gallager et al. *A Distributed Algorithm for Minimum Weight Spanning Trees*. IEEE Trans. on Prog. Lang. and Systems. 1983.

#### MODEL CHECKING

- D.P.L. Simons et al. *Mechanical Verification of the IEE 1394a Root Contention Protocol using Uppaal2* Springer International Journal of Software Tools for Technology Transfer. 2001
- H. Toetenel et al. Parametric verification of the IEEE 1394a Root
  Contention Protocol using LPMC Proceedings of the 7th International
  Conference on Real-time Computing Systems and Applications. IEEE
  Computer Society Press. 2000

## THEOREM PROVING

 M. Devillers et al. Verification of the Leader Election: Formal Method Applied to IEEE 1394. Formal Methods in System Design. 2000

- J.R. Abrial et al. *A Mechanically Proved and Incremental Development of IEEE 1394*. To be published 2002

- We are given a connected and acyclic network of nodes
- Nodes are linked by bidirectional channels
- We want to have one node being elected the leader in a finite time
- This is to be done in a distributed and non-deterministic way
- Next are two distinct abstract animations of the protocol





































- Formal definition and properties of the network
- A one-shot abstract model of the protocol
- Presenting a (still abstract) loop-like centralized solution
- Introducing message passing between the nodes (delays)
- Modifying the data structure in order to distribute the protocol



## Let ND be a set of nodes (with at least 2 nodes)



## Let gr be a graph built and defined on ND



## gr is a symmetric and irreflexive graph

#### gr is a graph built on ND $gr \subseteq ND \times ND$

gr is a graph built on ND

gr is defined on ND

 $gr \subseteq ND \times ND$ 

 $\operatorname{dom}\left(gr\right) = ND$ 

gr is a graph built on ND

gr is defined on ND

 $gr \subseteq ND \times ND$ 

$$\operatorname{dom}\left(gr\right)=ND$$

$$gr = gr^{-1}$$

gr is symmetric

gr is a graph built on ND

gr is defined on ND

gr is symmetric

gr is irreflexive

 $gr \subseteq ND \times ND$ 

 $\mathrm{dom}\left(gr\right)=ND$ 

 $gr = gr^{-1}$ 

 $\mathsf{id}(ND) \cap gr = \emptyset$


# gr is connected and acyclic

- A tree is a special graph

- A tree has a root

- A tree has a, so-called, father function

- A tree is acyclic

- A tree is connected from the root



### A tree t built on a set of nodes



## t is a function defined on ND except at the root



# Avoidind cycles



- Given
  - a set ND
  - a subset  $p \ {\rm of} \ ND$
  - a binary relation t built on ND
- The inverse image of p under t is denoted by  $t^{-1}[p]$

$$t^{-1}[p] \cong \{ x \mid x \in ND \land \exists y \cdot (y \in p \land (x,y) \in t) \}$$

- When t is a partial function, this reduces to

$$\{x \mid x \in \mathsf{dom}(t) \land t(x) \in p\}$$

- If p is included in its inverse image, we have then:

 $\forall x \cdot (x \in p \Rightarrow x \in \text{dom}(t) \land t(x) \in p)$ 

- Notice that the empty set enjoys this property

 $\emptyset \subseteq t^{-1}[\emptyset]$ 

- The property of having no cycle is thus equivalent to:

The only subset p of ND s.t.  $p \subseteq t^{-1}[p]$  is EMPTY

$$\forall p \cdot \begin{pmatrix} p \subseteq ND \land \\ p \subseteq t^{-1}[p] \\ \Rightarrow \\ p = \emptyset \end{pmatrix}$$

 $r \text{ is a member of } ND \quad r \in ND$ 

- r is a member of ND  $r \in ND$
- *t* is a function  $t \in ND \{r\} \rightarrow ND$

- $r \text{ is a member of } ND \quad r \in ND$
- *t* is a function  $t \in ND \{r\} \rightarrow ND$

t is acyclic

$$\forall p \cdot \begin{pmatrix} p \subseteq ND \land \\ p \subseteq t^{-1}[p] \\ \Rightarrow \\ p = \emptyset \end{pmatrix}$$

### *t* is acyclic: equivalent formulations

$$\forall p \cdot \begin{pmatrix} p \subseteq ND \land \\ p \subseteq t^{-1}[p] \\ \Rightarrow \\ p = \emptyset \end{pmatrix} \Leftrightarrow \forall q \cdot \begin{pmatrix} q \subseteq ND \land \\ r \in q \land \\ t^{-1}[q] \subseteq q \\ \Rightarrow \\ ND \subseteq q \end{pmatrix}$$

# This gives an Induction Rule

$$\forall q \cdot \begin{pmatrix} q \subseteq ND \land \\ r \in q \land \\ \forall x \cdot (x \in ND - \{r\} \land t(x) \in q \Rightarrow x \in q) \\ \Rightarrow \\ ND \subseteq q \end{pmatrix}$$

#### $r \text{ is a member of } ND \quad r \in ND$

*t* is a function  $t \in ND - \{r\} \to ND$ 

t is acyclic

$$\forall q \cdot \begin{pmatrix} q \subseteq ND \land \\ r \in q \land \\ t^{-1}[q] \subseteq q \\ \Rightarrow \\ ND \subseteq q \end{pmatrix}$$



# A spanning tree t of the graph gr

### The predicate spanning (r, t, gr)

r, t is a tree tree (r, t)

t is included in gr  $t \subseteq gr$ 

- Defining a relation fn linking a node to the possible spanning trees of gr having that node as a root:

$$fn \subseteq ND \times (ND \leftrightarrow ND)$$
  

$$\forall (r,t) \cdot \begin{pmatrix} r \in ND \land \\ t \in ND \leftrightarrow ND \\ \Rightarrow \\ (r,t) \in fn \Leftrightarrow \text{ spanning } (r,t,gr) \end{pmatrix}$$

### Totality of relation $fn \Rightarrow$ Connectivity of gr

Functionality of relation  $fn \Rightarrow$  Acyclicity of gr

$$gr \subseteq ND \times ND$$
  
dom  $(gr) = ND$   
 $gr = gr^{-1}$   
id  $(ND) \cap gr = \emptyset$ 

$$fn \in ND \rightarrow (ND \leftrightarrow ND)$$

$$\forall (r,t) \cdot \begin{pmatrix} r \in ND \land \\ t \in ND \leftrightarrow ND \\ \Rightarrow \\ t = fn(r) \Leftrightarrow \text{ spanning } (r,t,gr) \end{pmatrix}$$

- Variables rt and ts

 $rt \in ND$  $ts \in ND \leftrightarrow ND$ 



- Introducing a new variable, tr, corresponding to the "tree" in construction
- Introducing a new event: the progression event
- Defining the invariant
- Back to the animation : Observe the construction of the tree



















- The green arrows correspond to the tr function

- The blue nodes are the domain of tr

- The function tr is a forest (multi-tree) on nodes

- The red nodes are the roots of these trees

#### The predicate invariant (tr)

#### $tr \in ND \nrightarrow ND$

#### The predicate invariant (tr)

 $tr \in ND \twoheadrightarrow ND$ 

$$\forall p \cdot \begin{pmatrix} p \subseteq ND & \land \\ ND - \text{dom}(tr) \subseteq p & \land \\ tr^{-1}[p] \subseteq p \\ \Rightarrow \\ ND \subseteq p \end{pmatrix}$$
## The predicate invariant (tr)

 $tr \in ND \twoheadrightarrow ND$ 

$$\forall p \cdot \begin{pmatrix} p \subseteq ND & \land \\ ND - \operatorname{dom}(tr) \subseteq p & \land \\ tr^{-1}[p] \subseteq p \\ \Rightarrow \\ ND \subseteq p \end{pmatrix}$$

 $\operatorname{dom}(tr) \triangleleft (tr \cup tr^{-1}) = \operatorname{dom}(tr) \triangleleft gr$ 



- Introducing the new event "progress"
- Refining the abstract event "elect"

- Back to the animation : Observe the "guard" of progress





When a red node x is connected to AT MOST one other red node y then event "progress" can take place

```
progress \widehat{=}
   ANY x, y where
      x, y \in gr \land
      x \notin \text{dom}(tr) \land
      y \notin \operatorname{dom}(tr) \land
      qr[\{x\}] = tr^{-1}[\{x\}] \cup \{y\}
   THEN
      tr := tr \cup \{x \mapsto y\}
   END
```

## To be proved

$$invariant(tr) \land$$
  

$$x, y \in gr \land$$
  

$$x \notin tr \land$$
  

$$y \notin tr \land$$
  

$$gr[\{x\}] = tr^{-1}[\{x\}] \cup \{y\}$$
  

$$\Rightarrow$$
  

$$invariant(tr \cup \{x \mapsto y\})$$





When a red node x is ONLY connected to blue nodes then event "elect" can take place

> elect  $\widehat{=}$ ANY x WHERE  $x \in ND \land$   $gr[\{x\}] = tr^{-1}[\{x\}]$ THEN rt, ts := x, trEND

```
elect \widehat{=}
BEGIN
rt, ts : spanning (rt, ts, gr)
END
```

```
elect \widehat{=}

ANY x WHERE

x \in ND \land

gr[\{x\}] = tr^{-1}[\{x\}]

THEN

rt, ts := x, tr

END
```

## To be proved

$$invariant(tr) \land$$
  

$$x \in ND \land$$
  

$$gr[\{x\}] = tr^{-1}[\{x\}]$$
  

$$ts = tr$$
  

$$\Rightarrow$$
  

$$spanning(x, ts, gr)$$

- 15 proofs

- Among which 9 were interactive (one is a bit difficult !)

- 12 proofs

- Among which 5 were interactive (one is a bit difficult !)

- Animation of the current model





































- Nodes are communicating with their neighbors
- This is done by means of messages
- Messages are acknowledged
- Acknowledgements are confirmed
- Next is a local animation














Invariant (1)

 $msg \in ND \twoheadrightarrow ND$ 

 $ack \in ND \nrightarrow ND$ 

 $tr \subseteq ack \subseteq msg \subseteq gr$ 

#### Node x sends a message to node y

send\_msg  $\cong$ ANY x, y where  $x, y \in gr \land$  $x \notin \operatorname{dom}(tr) \land$  $y, x \notin tr \land$  $gr[\{x\}] = tr^{-1}[\{x\}] \cup \{y\} \land$  $y, x \notin ack \land$  $x \notin \text{dom}(msg)$ THEN  $msg := msg \cup \{x \mapsto y\}$ END

Node y sends an acknowledgement to node x



#### Node x sends a confirmation to node y



# Invariant (2)

$$\forall (x, y) \cdot \begin{pmatrix} x, y \in msg - ack \\ \Rightarrow \\ x, y \in gr \land \\ x \notin \operatorname{dom}(tr) \land y \notin \operatorname{dom}(tr) \land \\ gr[\{x\}] = tr^{-1}[\{x\}] \cup \{y\} \end{pmatrix}$$

$$\forall (x, y) \cdot \begin{pmatrix} x, y \in ack \land \\ x \notin \operatorname{dom}(tr) \\ \Rightarrow \\ x, y \in gr \land \\ y \notin \operatorname{dom}(tr) \land \\ gr[\{x\}] = tr^{-1}[\{x\}] \cup \{y\} \end{pmatrix}$$

- Explaining the problem
- Proposing a partial solution

- Towards a better treatment

- Back to the local animation





















## **Discovering Contention**













## **Discovering Contention**













## **Discovering Contention**











- Node y discovers the contention with node x because:
  - It has sent a message to node  $\boldsymbol{x}$
  - It has not yet received a response from node  $\boldsymbol{x}$
  - It receives instead a message from node  $\boldsymbol{x}$

- Node  $\boldsymbol{x}$  also discovers the contention with node  $\boldsymbol{y}$ 

- Assumption: The time between both discoveries IS SUPPOSED TO BE BOUNDED BY  $\tau$  ms

- The time  $\tau$  is the maximum transmission time between 2 connected nodes
- Each node waits for  $\tau$  ms after its own discovery

 After this, each node thus knows that the other has also discovered the contention

- Each node then retries immediately
- **PROBLEM**: This may continue for ever

- Each node waits for  $\tau$  ms after its own discovery
- Each node then choses with equal probability:
  - either to wait for a short delay
  - or to wait for a large delay
- Each node then retries

- Question: Does this solves the problem ?

- Are we sure to eventually have one node winning ?

- Answer: Listen carefully to Caroll Morgan's lectures

Node y discovers a contention with node x

send\_ack 
$$\widehat{=}$$
  
ANY  $x, y$  WHERE  
 $x, y \in msg-ack \land$   
 $y \notin dom(msg)$   
THEN  
 $ack := ack \cup \{x \mapsto y\}$   
END

contention  $\widehat{=}$ ANY x, y WHERE  $x, y \in msg-ack \land$   $y \in dom (msg)$ THEN  $cnt := cnt \cup \{x \mapsto y\}$ END

- Introducing a dummy contention channel: cnt

 $cnt \in ND \Rightarrow ND$ 

 $cnt \subseteq msg$ 

 $ack \cap cnt = \emptyset$ 

#### Solving the contention (simulating the $\tau$ delay)



- 73 proofs

- Among which 34 were interactive

- The representation of the graph gr is modified

- The representation of the tree tr is modified

- Other data structures are localized

## The graph gr and the tree tr are now localized

 $nb \in ND \rightarrow \mathbb{P}(ND)$ 

$$\forall x \cdot (x \in ND \Rightarrow nb(x) = gr[\{x\}])$$

 $sn \in ND \to \mathbb{P}(ND)$ 

 $\forall x \cdot (x \in ND \Rightarrow sn(x) \subseteq tr^{-1}[\{x\}])$ 

# **Localization (2)**

 $bm \subseteq ND$ 

bm = dom (msg)

 $bt \subseteq ND$ 

bt = dom(tr)

 $ba \in ND \rightarrow \mathbb{P}(ND)$ 

 $\forall x \cdot (x \in ND \Rightarrow ba(x) = ack^{-1}[\{x\}])$ 

- Node x is elected the leader

elect 
$$\widehat{=}$$
  
ANY  $x$  WHERE  
 $x \in ND \land$   
 $nb(x) = sn(x)$   
THEN  
 $rt := x$   
END

- Node x sends a message to node y (y is unique)

send\_msg 
$$\cong$$
  
ANY  $x, y$  WHERE  
 $x \in ND-bm \land$   
 $y \in ND-(ba(x)\cup sn(x)) \land$   
 $nb(x) = sn(x)\cup \{y\}$   
THEN  
 $msg := msg \cup \{x \mapsto y\} \parallel$   
 $bm := bm \cup \{x\}$   
END

- Node y sends an acknowledgement to node x

```
send_ack \widehat{=}
   ANY x, y where
      x, y \in msg \land
      x \notin ba(y) \land
      y \notin bm
   THEN
      ack := ack \cup \{x \mapsto y\} \quad \parallel
      ba(y) := ba(y) \cup \{x\}
   END
```

- Node x sends a confirmation to node y

```
progress \widehat{=}
   ANY x, y where
      x, y \in ack \land
      x \notin bt
   THEN
      tr := tr \cup \{x \mapsto y\} \quad \|
      bt := bt \cup \{x\}
   END
```

- Node y receives confirmation from node x







- 29 proofs

- Among which 19 were interactive

- 119 proofs

- Among which 63 were interactive

- Establishing the mathematical framework

- Establishing the mathematical framework
- Resolving the mathematical problem in one shot

- Establishing the mathematical framework
- Resolving the mathematical problem in one shot

- Resolving the same problem on a step by step basis

- Establishing the mathematical framework
- Resolving the mathematical problem in one shot
- Resolving the same problem on a step by step basis
- Involving communication by means of messages

- Establishing the mathematical framework
- Resolving the mathematical problem in one shot
- Resolving the same problem on a step by step basis
- Involving communication by means of messages
- Towards the localization of data structures