# Distributed algorithms <br> 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

## This Session

- 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


## The Problem (2)

- 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


## References (1)

## 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


## References (2)

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.


## References (3)

## 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


## References (4)

## 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

## Informal Absract Properties of the Protocol

- 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






O














## Summary of Development Process

- 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


$$
g r \subseteq N D \times N D
$$

$g r$ is a graph built on $N D$
$g r$ is defined on $N D$
$g r \subseteq N D \times N D$
$\operatorname{dom}(g r)=N D$
$g r$ is a graph built on $N D$
$g r$ is defined on $N D$
$g r$ is symmetric
$g r \subseteq N D \times N D$
$\operatorname{dom}(g r)=N D$
$g r=g r^{-1}$
$g r$ is a graph built on $N D$
$g r$ is defined on $N D$
$g r$ is symmetric
$g r$ is irreflexive

$$
g r \subseteq N D \times N D
$$

$$
\operatorname{dom}(g r)=N D
$$

$$
g r=g r^{-1}
$$

$i d(N D) \cap g r=\emptyset$


## A Little Detour Through Trees

- 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



- Given
- a set $N D$
- a subset $p$ of $N D$
- a binary relation $t$ built on $N D$
- The inverse image of $p$ under $t$ is denoted by $t^{-1}[p]$
$t^{-1}[p] \cong\{x \mid x \in N D \wedge \exists y \cdot(y \in p \wedge(x, y) \in t)\}$
- When $t$ is a partial function, this reduces to

$$
\{x \mid x \in \operatorname{dom}(t) \wedge t(x) \in p\}
$$

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

$$
\forall x \cdot(x \in p \Rightarrow x \in \operatorname{dom}(t) \wedge t(x) \in p)
$$

- 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 $N D$ s.t. $p \subseteq t^{-1}[p]$ is EMPTY

$$
\forall p \cdot\left(\begin{array}{c}
p \subseteq N D \wedge \\
p \subseteq t^{-1}[p] \\
\Rightarrow \\
p=\emptyset
\end{array}\right)
$$

The predicate tree $(r, t)$

## The predicate tree $(r, t)$

$r$ is a member of $N D \quad r \in N D$

## The predicate tree ( $r, t$ )

$r$ is a member of $N D \quad r \in N D$
$t$ is a function

$$
t \in N D-\{r\} \rightarrow N D
$$

## The predicate tree $(r, t)$

$r$ is a member of $N D \quad r \in N D$
$t$ is a function

$$
t \in N D-\{r\} \rightarrow N D
$$

$t$ is acyclic

$$
\forall p \cdot\left(\begin{array}{l}
p \subseteq N D \wedge \\
p \subseteq t^{-1}[p] \\
\Rightarrow \\
p=\emptyset
\end{array}\right)
$$

## $t$ is acyclic: equivalent formulations

$$
\forall p \cdot\left(\begin{array}{l}
p \subseteq N D \wedge \\
p \subseteq t^{-1}[p] \\
\Rightarrow \\
p=\emptyset
\end{array}\right) \quad \Leftrightarrow \quad \forall q \cdot\left(\begin{array}{l}
q \subseteq N D \wedge \\
r \in q \wedge \\
t^{-1}[q] \subseteq q \\
\Rightarrow \\
N D \subseteq q
\end{array}\right)
$$

## This gives an Induction Rule

$$
\forall q \cdot\left(\begin{array}{l}
q \subseteq N D \wedge \\
r \in q \wedge \\
\forall x \cdot(x \in N D-\{r\} \wedge t(x) \in q \Rightarrow x \in q) \\
\Rightarrow \\
N D \subseteq q
\end{array}\right)
$$

## The predicate tree $(r, t)$

$r$ is a member of $N D \quad r \in N D$
$t$ is a function

$$
t \in N D-\{r\} \rightarrow N D
$$

$t$ is acyclic

$$
\forall q \cdot\left(\begin{array}{c}
q \subseteq N D \wedge \\
r \in q \wedge \\
t^{-1}[q] \subseteq q \\
\Rightarrow \\
N D \subseteq q
\end{array}\right)
$$



A spanning tree $t$ of the graph gr

The predicate spanning $(r, t, g r)$
$r, t$ is a tree
$t$ is included in $g r$
tree $(r, t)$
$t \subseteq g r$

## The graph $g r$ is connected and acyclic (1)

- Defining a relation $f n$ linking a node to the possible spanning trees of $g r$ having that node as a root:

$$
\begin{aligned}
& f n \subseteq N D \times(N D \leftrightarrow N D) \\
& \forall(r, t) \cdot\left(\begin{array}{l}
r \in N D \wedge \\
t \in N D \leftrightarrow N D \\
\Rightarrow \\
(r, t) \in f n \Leftrightarrow \operatorname{spanning}(r, t, g r)
\end{array}\right)
\end{aligned}
$$

## The graph $g r$ is connected and acyclic (2)

Totality of relation $f n \Rightarrow$ Connectivity of $g r$

Functionality of relation $f n \Rightarrow$ Acyclicity of $g r$

## Summary of constants $g r$ and $f n$

$$
\begin{aligned}
& g r \subseteq N D \times N D \\
& \operatorname{dom}(g r)=N D \\
& g r=g r^{-1} \\
& \text { id }(N D) \cap g r=\emptyset \\
& f n \in N D \rightarrow(N D \leftrightarrow N D) \\
& \forall(r, t) \cdot\left(\begin{array}{l}
r \in N D \wedge \\
t \in N D \leftrightarrow N D \\
\Rightarrow \\
t=f n(r) \Leftrightarrow \text { spanning }(r, t, g r)
\end{array}\right.
\end{aligned}
$$

## Election in One Shot: Building a Spanning Tree

- Variables $r t$ and $t s$

$$
\begin{aligned}
& r t \in N D \\
& t s \in N D \leftrightarrow N D \\
& \begin{array}{l}
\text { elect } \cong \\
\text { BEGIN } \\
r t, t s: \text { spanning }(r t, t s, g r) \\
\text { END }
\end{array}
\end{aligned}
$$

## First Refinement (1)

- 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






O





- The green arrows correspond to the $t r$ function
- The blue nodes are the domain of $t r$
- The function $t r$ is a forest (multi-tree) on nodes
- The red nodes are the roots of these trees


## The predicate invariant (tr)

$$
t r \in N D \leftrightarrow N D
$$

## The predicate invariant (tr)

$$
\begin{aligned}
& \operatorname{tr} \in N D \mapsto N D \\
& \forall p \cdot\left(\begin{array}{l}
p \subseteq N D \quad \wedge \\
N D-\operatorname{dom}(t r) \subseteq p \quad \wedge \\
t r^{-1}[p] \subseteq p \\
\Rightarrow \\
N D \subseteq p
\end{array}\right)
\end{aligned}
$$

## The predicate invariant (tr)

$$
\begin{aligned}
& \operatorname{tr} \in N D \rightarrow N D \\
& \forall p \cdot\left(\begin{array}{l}
p \subseteq N D \quad \wedge \\
N D-\operatorname{dom}(t r) \subseteq p \quad \wedge \\
t r^{-1}[p] \subseteq p \\
\Rightarrow \\
N D \subseteq p
\end{array}\right)
\end{aligned}
$$

$$
\operatorname{dom}(t r) \triangleleft\left(t r \cup t r^{-1}\right)=\operatorname{dom}(t r) \triangleleft g r
$$



## First Refinement (2)

- 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

$$
\begin{aligned}
& x, y \in g r \wedge \\
& x \notin \operatorname{dom}(\operatorname{tr}) \wedge \\
& y \notin \operatorname{dom}(\operatorname{tr}) \wedge \\
& g r[\{x\}]=\operatorname{tr}^{-1}[\{x\}] \cup\{y\}
\end{aligned}
$$

## THEN

$$
\operatorname{tr}:=\operatorname{tr} \cup\{x \mapsto y\}
$$

END

## To be proved

$$
\begin{aligned}
& \text { invariant }(t r) \wedge \\
& x, y \in g r \wedge \\
& x \notin \operatorname{tr} \wedge \\
& y \notin t r \wedge \\
& g r[\{x\}]=\operatorname{tr}^{-1}[\{x\}] \cup\{y\} \\
& \Rightarrow \\
& \text { invariant }(\operatorname{tr} \cup\{x \mapsto y\})
\end{aligned}
$$




When a red node $x$ is ONLY connected to blue nodes then event "elect" can take place
elect $\xlongequal{\widehat{ }}$
ANY $x$ WHERE

$$
\begin{aligned}
& x \in N D \wedge \\
& \operatorname{gr}[\{x\}]=\operatorname{tr}^{-1}[\{x\}]
\end{aligned}
$$

THEN

$$
r t, t s:=x, t r
$$

END

## elect $\widehat{=}$

## BEGIN

$r t, t s$ : spanning ( $r t, t s, g r$ )
END
elect $\widehat{=}$
ANY $x$ WHERE

$$
\begin{aligned}
& x \in N D \wedge \\
& \operatorname{gr}[\{x\}]=\operatorname{tr}^{-1}[\{x\}]
\end{aligned}
$$

## THEN

$$
r t, t s:=x, t r
$$

END

## To be proved



## Summary of First Refinement

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


## Current state of the model

- 12 proofs
- Among which 5 were interactive (one is a bit difficult !)
- Animation of the current model






O














## Second Refinement

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





# Sending a message 



## Receiving a message

## Sending Acknowledgement



## Receiving Acknowledgement

 Sending Confirmation


## Invariant (1)

$$
\begin{aligned}
& m s g \in N D \mapsto N D \\
& a c k \in N D \mapsto N D \\
& t r \subseteq a c k \subseteq m s g \subseteq g r
\end{aligned}
$$

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

## send_msg $\widehat{=}$

ANY $x, y$ WHERE

$$
\begin{aligned}
& x, y \in g r \wedge \\
& x \notin \operatorname{dom}(t r) \wedge \\
& y, x \notin \operatorname{tr} \wedge \\
& g r[\{x\}]=t r^{-1}[\{x\}] \cup\{y\} \wedge \\
& y, x \notin a c k \wedge \\
& x \notin \operatorname{dom}(m s g)
\end{aligned}
$$

THEN

$$
m s g:=m s g \cup\{x \mapsto y\}
$$

END

Node $y$ sends an acknowledgement to node $x$

## send ack $\widehat{=}$

ANY $x, y$ WHERE
$x, y \in m s g-a c k \wedge$
$y \notin \operatorname{dom}(m s g)$

## THEN

$$
\text { ack }:=\operatorname{ack} \cup\{x \mapsto y\}
$$

END

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

## progress

ANY $x, y$ WHERE
$x, y \in a c k \wedge$
$x \notin \operatorname{dom}(t r)$

## THEN

$\operatorname{tr}:=\operatorname{tr} \cup\{x \mapsto y\}$
END

## Invariant (2)

$$
\begin{aligned}
& \forall(x, y) \cdot\left(\begin{array}{l}
x, y \in m s g-a c k \\
\Rightarrow \\
x, y \in g r \wedge \\
x \notin \operatorname{dom}(t r) \wedge y \notin \operatorname{dom}(t r) \wedge \\
g r[\{x\}]=t r^{-1}[\{x\}] \cup\{y\}
\end{array}\right) \\
& \forall(x, y) \cdot\left(\begin{array}{l}
x, y \in a c k \wedge \\
x \notin \operatorname{dom}(t r) \\
\Rightarrow \\
x, y \in g r \wedge \\
y \notin \operatorname{dom}(t r) \wedge \\
g r[\{x\}]=t r^{-1}[\{x\}] \cup\{y\}
\end{array}\right)
\end{aligned}
$$

## Second Refinement: The problem of contention

- Explaining the problem
- Proposing a partial solution
- Towards a better treatment
- Back to the local animation







# Sending a message 



## Sending another message



## Discovering Contention





# Sending a message 



## Sending another message







## Sending another message



## Discovering Contention





# Sending a message 



## Receiving a message

## Sending Acknowledgement



## Receiving Acknowledgement

 Sending Confirmation


## Discovering the Contention (1)

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


## Discovering the Contention (2)

- Node $x$ also discovers the contention with node $y$
- 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


## A Partial Solution

- 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


## A Better Solution (1)

- 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


## A Better Solution (2)

- 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 $x$

$$
\begin{aligned}
& \text { send ack } \widehat{=} \\
& \text { ANY } x, y \text { WHERE } \\
& x, y \in m s g-a c k \wedge \\
& y \notin \operatorname{dom}(m s g) \\
& \text { THEN } \\
& \quad a c k:=a c k \cup\{x \mapsto y\} \\
& \text { END }
\end{aligned}
$$

contention $\widehat{=}$
ANY $x, y$ WHERE

$$
x, y \in m s g-a c k \wedge
$$

$y \in \operatorname{dom}(m s g)$
THEN
$c n t:=c n t \cup\{x \mapsto y\}$
END

- Introducing a dummy contention channel: cnt

$$
\begin{aligned}
& c n t \in N D \rightarrow N D \\
& c n t \subseteq m s g \\
& \text { ack } \cap c n t=\emptyset
\end{aligned}
$$



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

solve_contention $\widehat{=}$
ANY $x, y$ WHERE

$$
x, y \in c n t \cup c n t^{-1}
$$

## THEN

$$
\begin{aligned}
& m s g:=m s g-c n t \quad \| \\
& c n t:=\emptyset
\end{aligned}
$$

## END

## Summary of Second Refinement

- 73 proofs
- Among which 34 were interactive


## Third Refinement: Localization

- The representation of the graph $g r$ is modified
- The representation of the tree $t r$ is modified
- Other data structures are localized

The graph $g r$ and the tree $t r$ are now localized

$$
\begin{aligned}
& n b \in N D \rightarrow \mathbb{P}(N D) \\
& \forall x \cdot(x \in N D \Rightarrow n b(x)=\operatorname{gr}[\{x\}]) \\
& s n \in N D \rightarrow \mathbb{P}(N D) \\
& \forall x \cdot\left(x \in N D \Rightarrow \operatorname{sn}(x) \subseteq \operatorname{tr}^{-1}[\{x\}]\right)
\end{aligned}
$$

## Localization (2)

$$
\begin{aligned}
& b m \subseteq N D \\
& b m=\operatorname{dom}(m s g) \\
& b t \subseteq N D \\
& b t=\operatorname{dom}(t r) \\
& b a \in N D \rightarrow \mathbb{P}(N D) \\
& \forall x \cdot\left(x \in N D \Rightarrow b a(x)=a c k^{-1}[\{x\}]\right)
\end{aligned}
$$

## - Node $x$ is elected the leader

## elect $\widehat{=}$

ANY $x$ WHERE

$$
\begin{aligned}
& x \in N D \wedge \\
& n b(x)=\operatorname{sn}(x)
\end{aligned}
$$

THEN

$$
r t:=x
$$

END

- Node $x$ sends a message to node $y$ ( $y$ is unique)
send msg $\widehat{=}$
ANY $x, y$ WHERE

$$
\begin{aligned}
& x \in N D-b m \wedge \\
& y \in N D-(b a(x) \cup \operatorname{sn}(x)) \wedge \\
& n b(x)=\operatorname{sn}(x) \cup\{y\}
\end{aligned}
$$

THEN

$$
\begin{aligned}
& m s g:=m s g \cup\{x \mapsto y\} \\
& b m:=b m \cup\{x\}
\end{aligned}
$$

## END

- Node $y$ sends an acknowledgement to node $x$


## send ack $\widehat{=}$

 ANY $x, y$ WHERE$x, y \in m s g \wedge$
$x \notin b a(y) \wedge$
$y \notin b m$
THEN

$$
\begin{aligned}
& a c k:=a c k \cup\{x \mapsto y\} \\
& b a(y):=b a(y) \cup\{x\}
\end{aligned}
$$

END

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

progress $\cong$
ANY $x, y$ WHERE

$$
\begin{aligned}
& x, y \in a c k \wedge \\
& x \notin b t
\end{aligned}
$$

THEN

$$
\begin{aligned}
\operatorname{tr} & :=\operatorname{tr} \cup\{x \mapsto y\} \\
b t & :=b t \cup\{x\}
\end{aligned}
$$

## END

- Node $y$ receives confirmation from node $x$


## rev_cnf $\widehat{ }$

ANY $x, y$ WHERE
$x, y \in \operatorname{tr} \wedge$
$x \notin \operatorname{sn}(y)$
THEN

$$
\operatorname{sn}(y):=\operatorname{sn}(y) \cup\{x\}
$$

END

## contention $\widehat{=}$

ANY $x, y$ WHERE

$$
\begin{aligned}
& x, y \in c n t \cup c n t^{-1} \wedge \\
& x \notin b a(y) \wedge \\
& y \in b m
\end{aligned}
$$

THEN
$c n t:=c n t \cup\{x \mapsto y\}$
END

## solve contention $\widehat{ }$

ANY $x, y$ WHERE
$x, y \in c n t \cup c n t^{-1}$

## THEN

$m s g:=m s g-c n t \quad \|$
$b m:=b m-\operatorname{dom}(c n t)$
$c n t:=\emptyset$
END

## Summary of Third Refinement

- 29 proofs
- Among which 19 were interactive


## Main Summary

- 119 proofs
- Among which 63 were interactive


## Conclusion: a Systematic Approach to Distribution

- Establishing the mathematical framework



## Conclusion: a Systematic Approach to Distribution

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


## Conclusion: a Systematic Approach to Distribution

- Establishing the mathematical framework
- Resolving the mathematical problem in one shot
- Resolving the same problem on a step by step basis


## Conclusion: a Systematic Approach to Distribution

- 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


## Conclusion: a Systematic Approach to Distribution

- 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

