Link Aggregation Control Protocol - Update

Presentation to the Link Aggregation Task Force, September 1998

Tony Jeffree

Overview

- Uses the best bits of the presentations by Finn/Wakerley/Fine, & Jeffree
- Comments taken on board from July meeting
  - Keys
  - Crowd machine removed - FFS
  - Work done on initial behaviour of Automatic mode
- Further work done on the protocol description & operation

Basic assumptions/objectives

- If aggregation is possible, it will happen automatically
- If not possible, links operate normally
- Determinism
- Rapid convergence
- Low risk of misconfiguration
- Low risk of duplication or misordering

Specific Objectives - 1

- Ability to configure “speak if spoken to” Ports (= Automatic mode) and “speak anyway” Ports (= Desirable mode)
- Ability to configure “Relaxed” operation for Ports that can hardware detect link failure, or “Nervous” operation for Ports that cannot

Specific Objectives - 2

- Fast detection of presence/absence of partners on initialisation
- Accommodation of hardware that can control transmit/receive independently, and of hardware that cannot
- Accommodation of hardware that may take significant time (> protocol re-transmission time) to change state

Specific Objectives - 3

- Fast detection of cases where aggregation cannot occur => activate as individual link
- Ability to determine which physical Ports can/cannot aggregate with which Aggregate Ports
- Very low probability of misdelivery
- Low probability of loss
- Low probability of reporting good link with only partial connectivity

Identifying link characteristics

- Many characteristics that contribute
  - Standardised in .3: Link speed, duplex/non-duplex…etc
  - Other characteristics…e.g., administrative, non-standardised
- A Link is allocated a single **Key** value
- All Links in a system that share the same Key can potentially aggregate;
- Links that are not capable of aggregation are allocated unique Keys

Identifying Links that can Aggregate

- System ID plus Key provides a global identifier
- The set of links between 2 systems that can aggregate are identified by concatenating the System ID and Key at each end of the links
- Hence, for two systems S and T that use C and D respectively as Key values for some links, then all links with {SC, TD} (interchangeably, {TD, SC}) can aggregate together

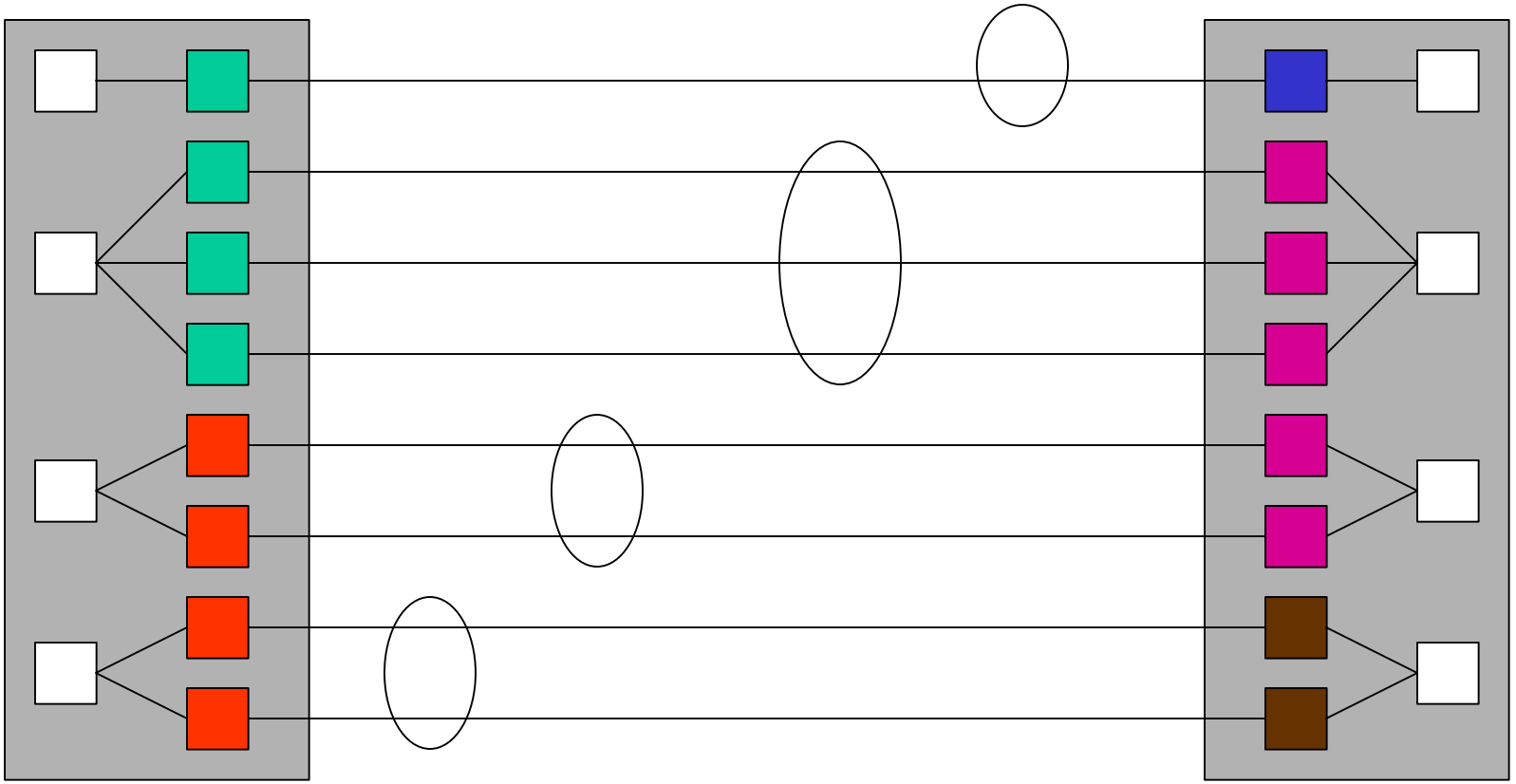
Detecting Aggregation possibility

- Aggregation possibility can be detected simply by exchanging System Ids and Keys across a link; each system can then see whether any other Links exist with the same {SC,TD} value.
- If other links in a system exist with the same {SC, TD} then they can all be added to the same Aggregate
- Simplifying assumption: no limit on aggregation size - allocate more capabilities if it is necessary to impose such a limit.

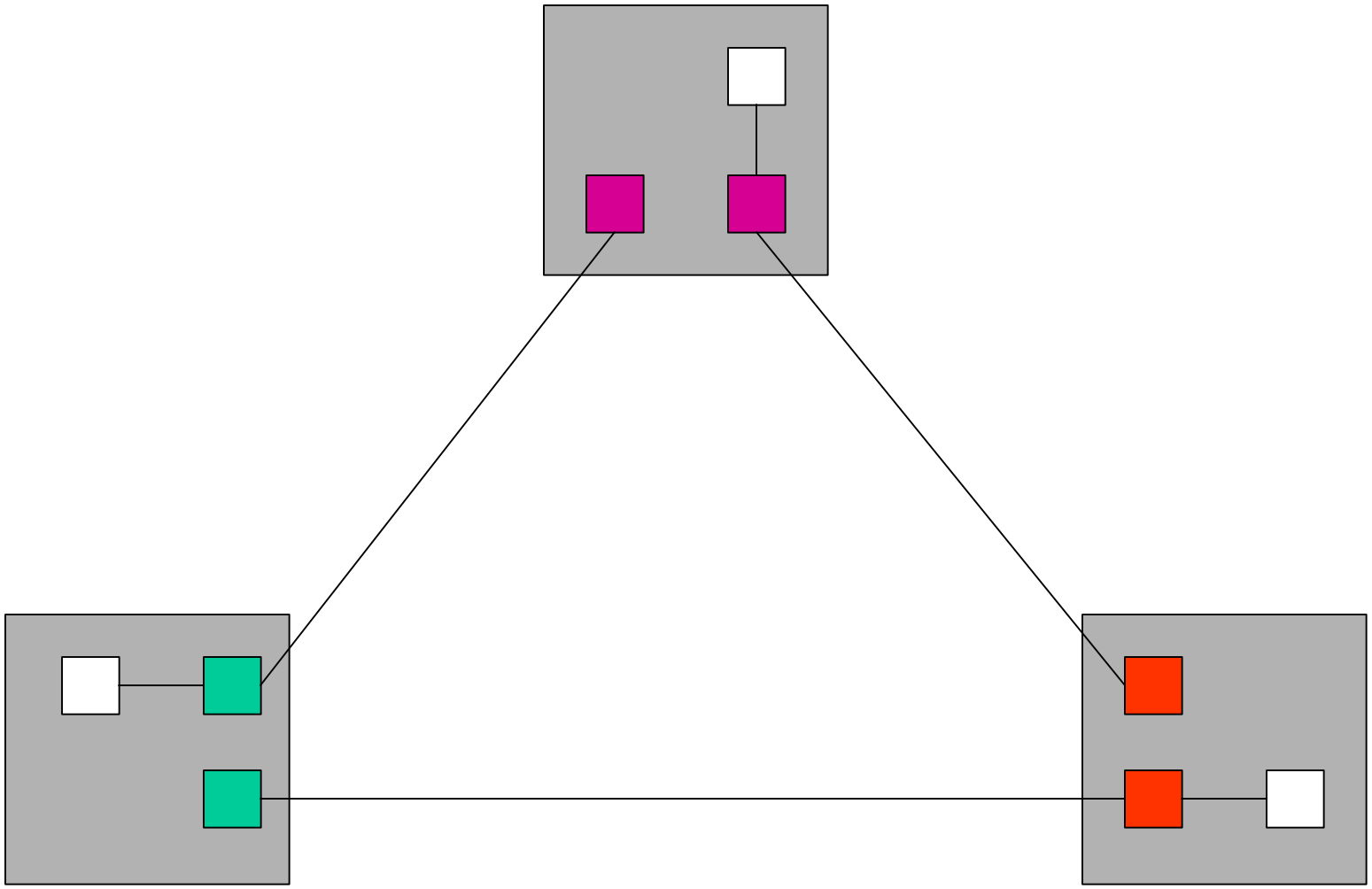
Effect of Keys - Example

System ID = A

System ID = B



Consequence of too few Agports



Prevention of Duplication/Reordering

- *Collect* once you are in the right aggregation
- Don’t *Distribute* until you know that the other end is Collecting
- Stop Distribution/Collection on a Link prior to moving it to a new aggregation
- BUT also need to accommodate equipment which cannot switch collector/distributor independently
- Need to “flush” other links if Conversations are re-allocated as a result of adding/removing links

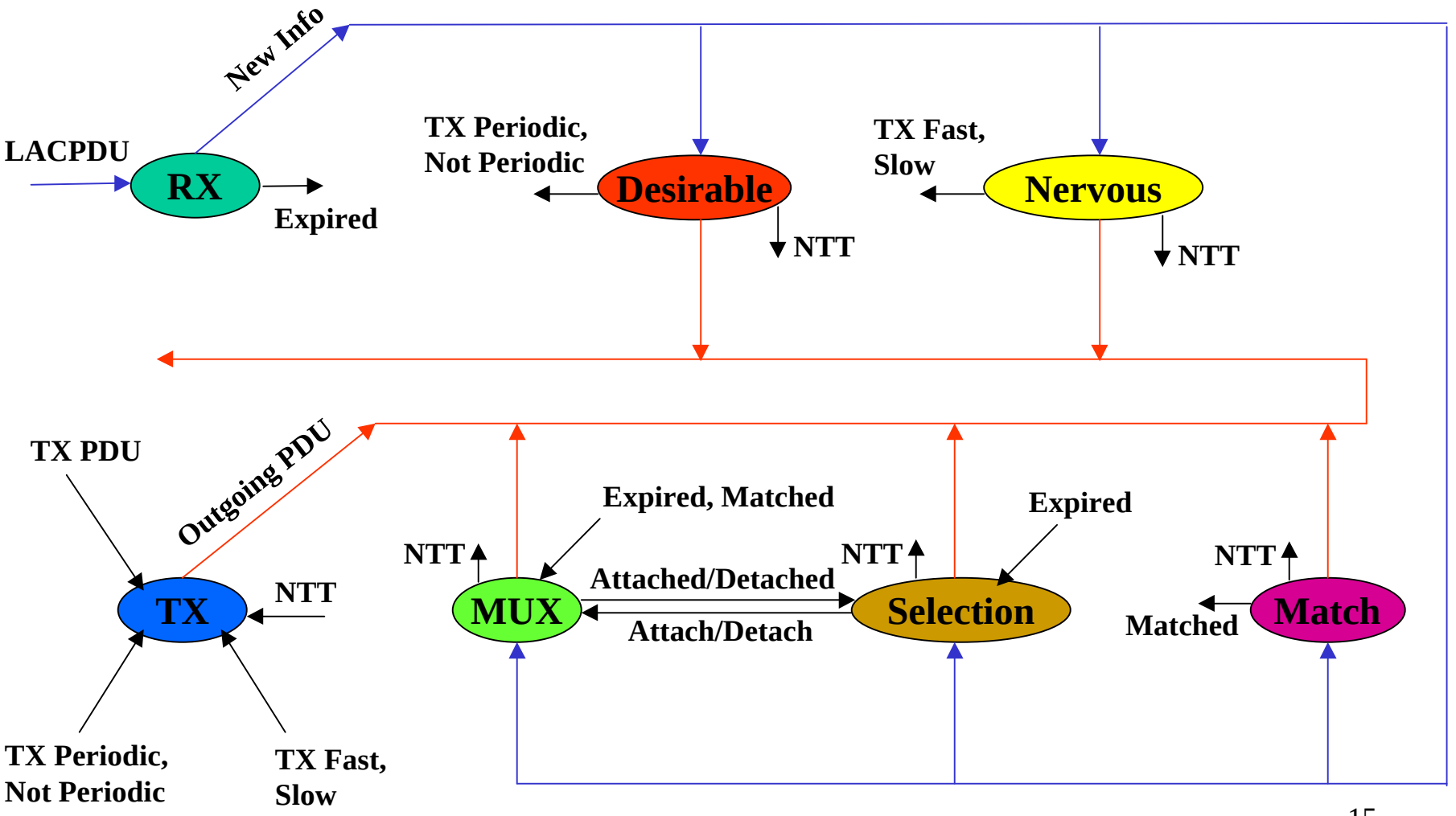
Protocol basics

- If the other guy doesn’t get it, say it again
- Assumption that packet loss is very low
- Communicate *state*, not *commands*
- *Need to Tell* if local state has changed, if information is old, or if the other guy does not get it
- Tell the other party what you know. When you are both agreed - aggregate

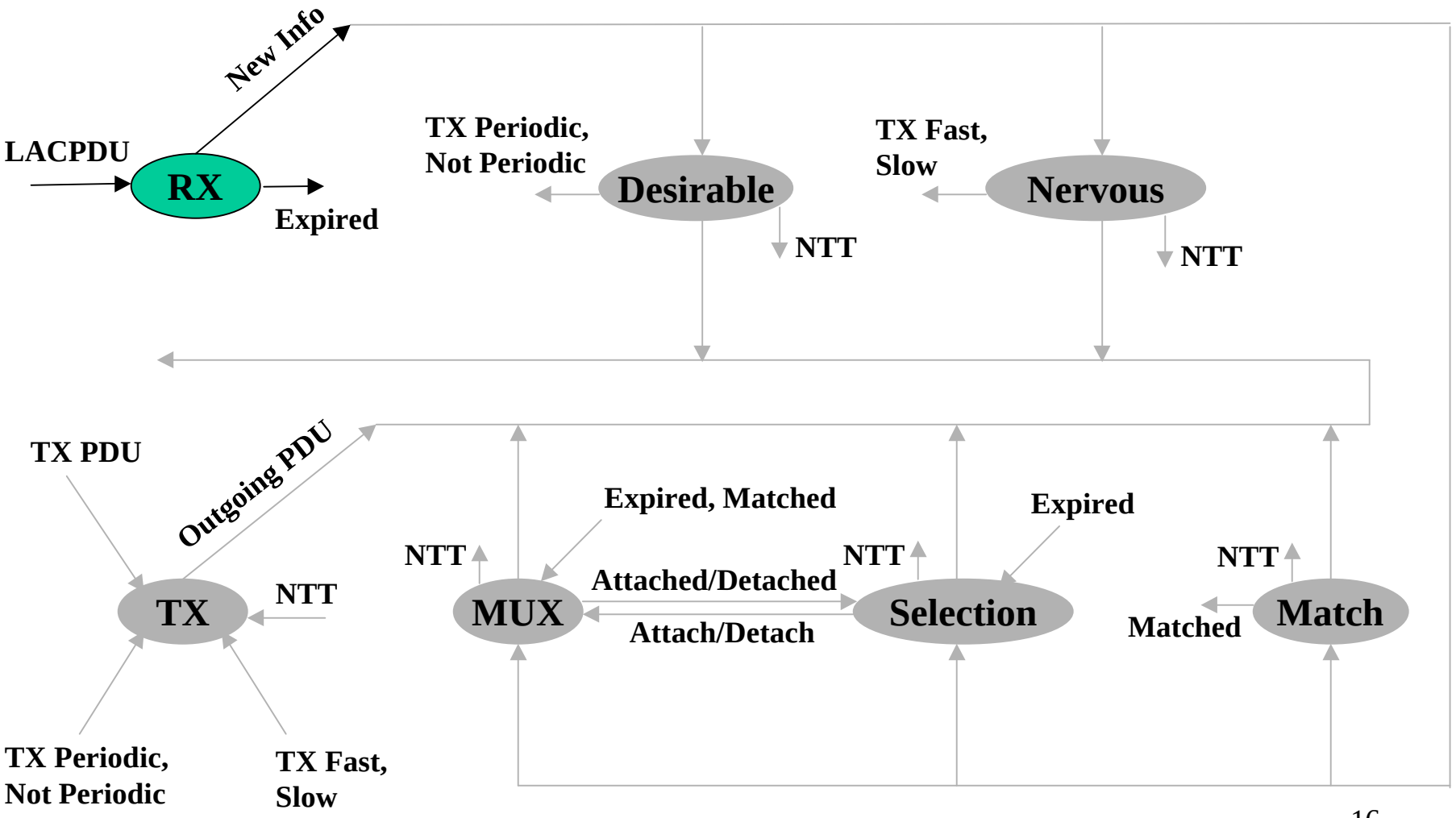
Flush protocol operation something like...

- Flush ID sent (along with normal message content). Sender chooses ID value.
- Recipient's NTT is asserted by receipt of Flush ID; Flush ID saved by recipient & sent in subsequent messages till message received with no Flush ID.
- Note: Does not fix the case of a link failing.

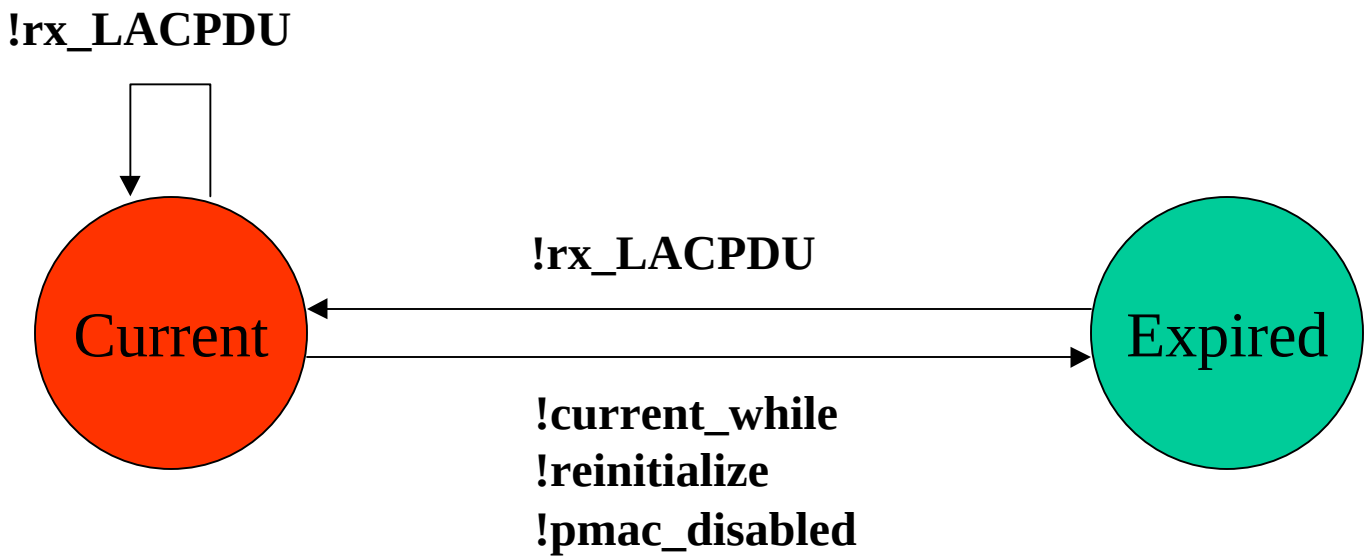
The Big Picture



RX


New Info
LACPDU
RX
Expired
TX Periodic,
Not Periodic
Desirable
NTT
TX Fast,
Slow
Nervous
NTT
TX PDU
Outgoing PDU
TX
NTT
TX Periodic,
Not Periodic
TX Fast,
Slow
Expired, Matched
NTT
MUX
Attached/Detached
Attach/Detach
NTT
Selection
Expired
NTT
Match
Matched

RX State Machine



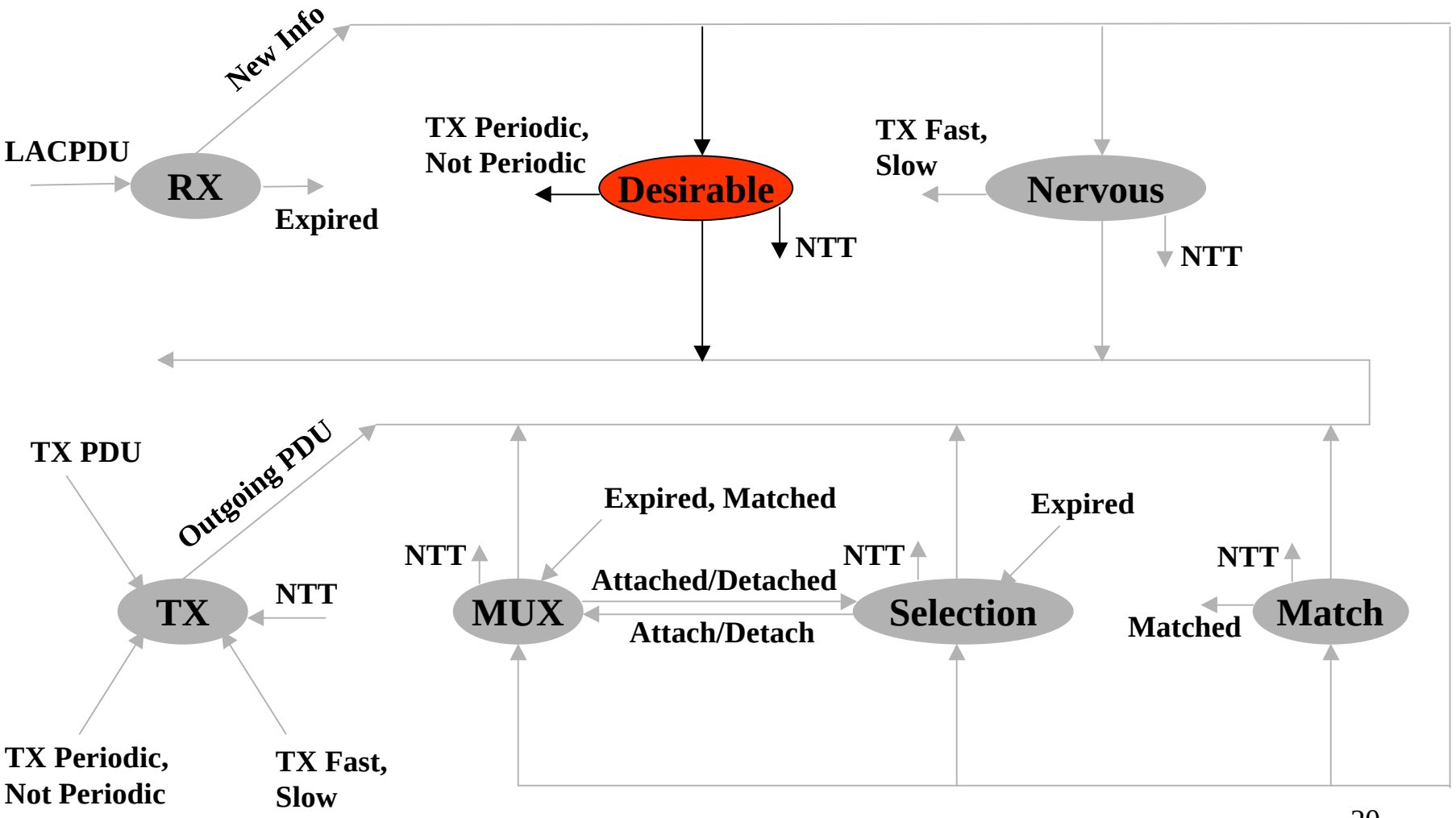
Information communicated

- My Port
- My System
- My Key
- My State:
  - Desirable/Auto
  - Nervous/Cool
  - Aggregate/Individual
  - In Sync/Out of Sync
  - Collector On/Off
  - Distributor On/Off
- Partner System
- Partner Key
- Partner State:
  - Desirable/Auto
  - Nervous/Cool
  - Aggregate/Individual
  - In Sync/Out of Sync
  - Collector On/Off
  - Distributor On/Off

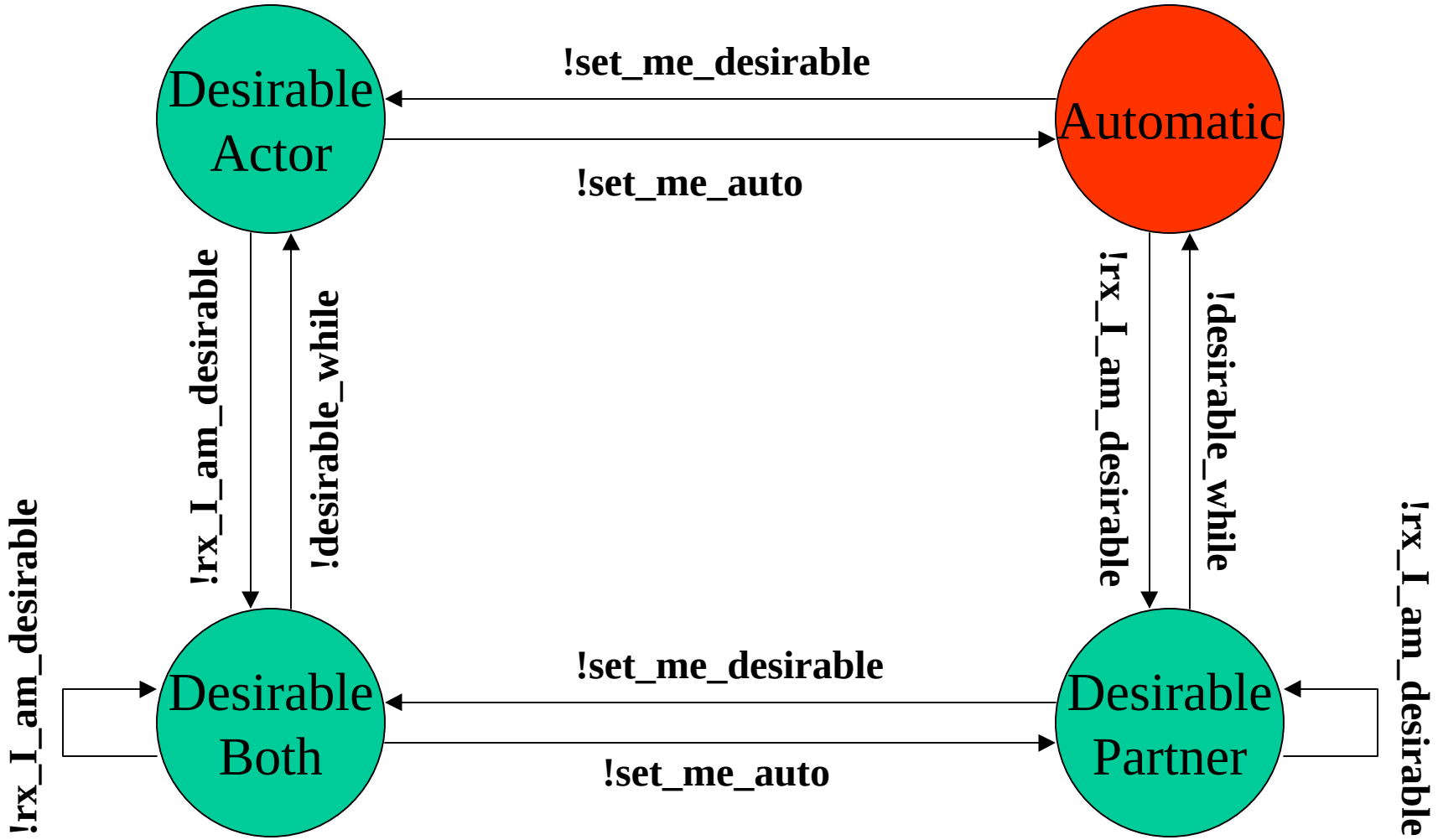
RX functionality recap

- Receives & unpacks incoming LACPDUs
- Signals availability of new information to other state machines
- Maintains knowledge of whether partner information is current or expired
- Expiry signalled to Selection and Mux machines

Desirable


New Info
LACPDU
RX
Expired
TX Periodic,
Not Periodic
Desirable
NTT
TX Fast,
Slow
Nervous
NTT
TX PDU
Outgoing PDU
TX
NTT
TX Periodic,
Not Periodic
TX Fast,
Slow
NTT
Expired, Matched
MUX
Attached/Detached
Attach/Detach
NTT
Expired
Selection
NTT
Matched
Match

Desirable - State Machine



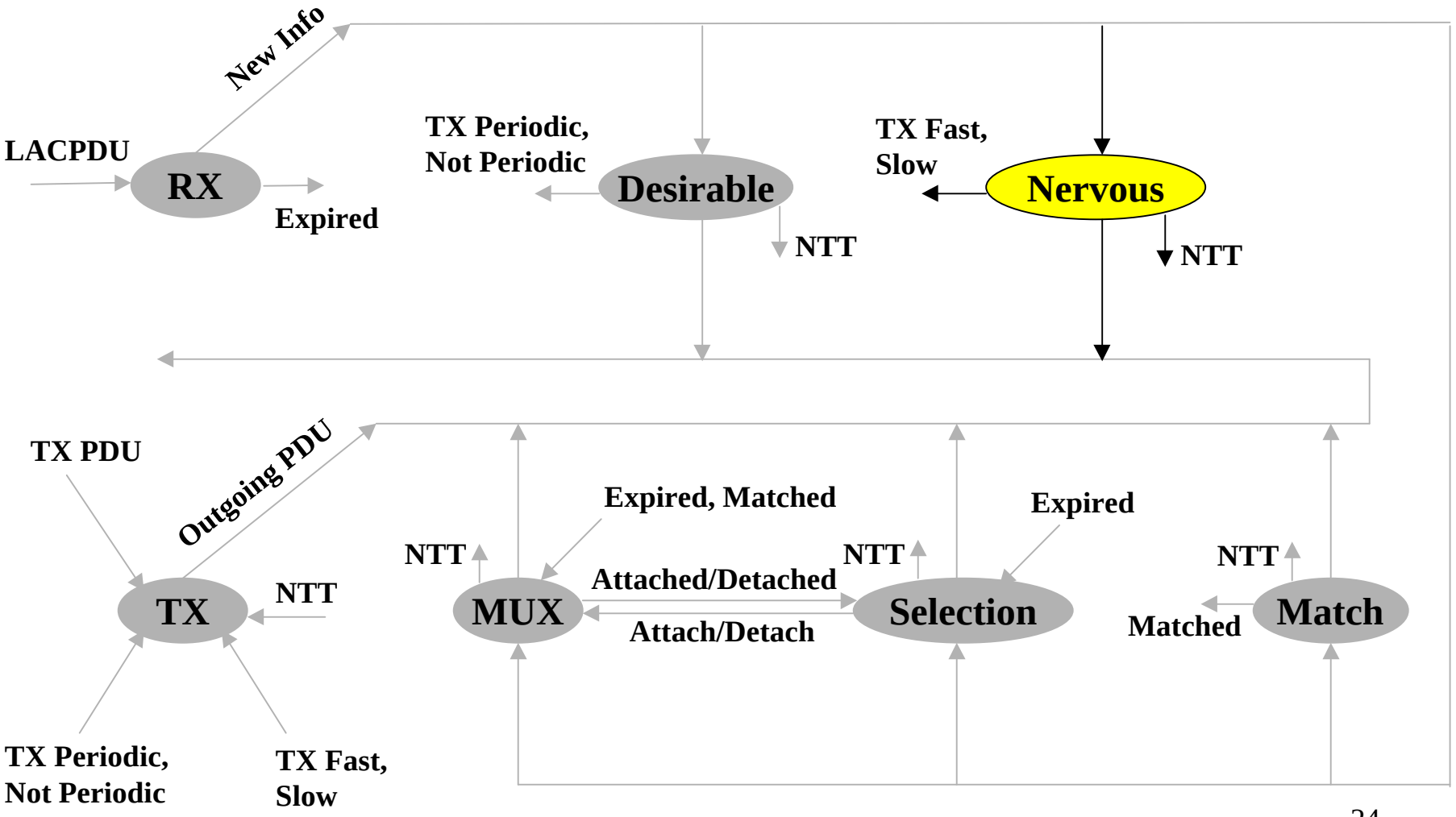
Desirable - Functionality Recap

- Determines whether or not this Port will generate periodic LACPDU transmissions
- *TX Periodic* if the actor or its partner are (or are believed to be) *desirable*
- *TX Not periodic* if the actor and its partner are (or are believed to be) *automatic*
- If *No periodic transmissions* this must be an individual link
- NTT if partner doesn’t know my state

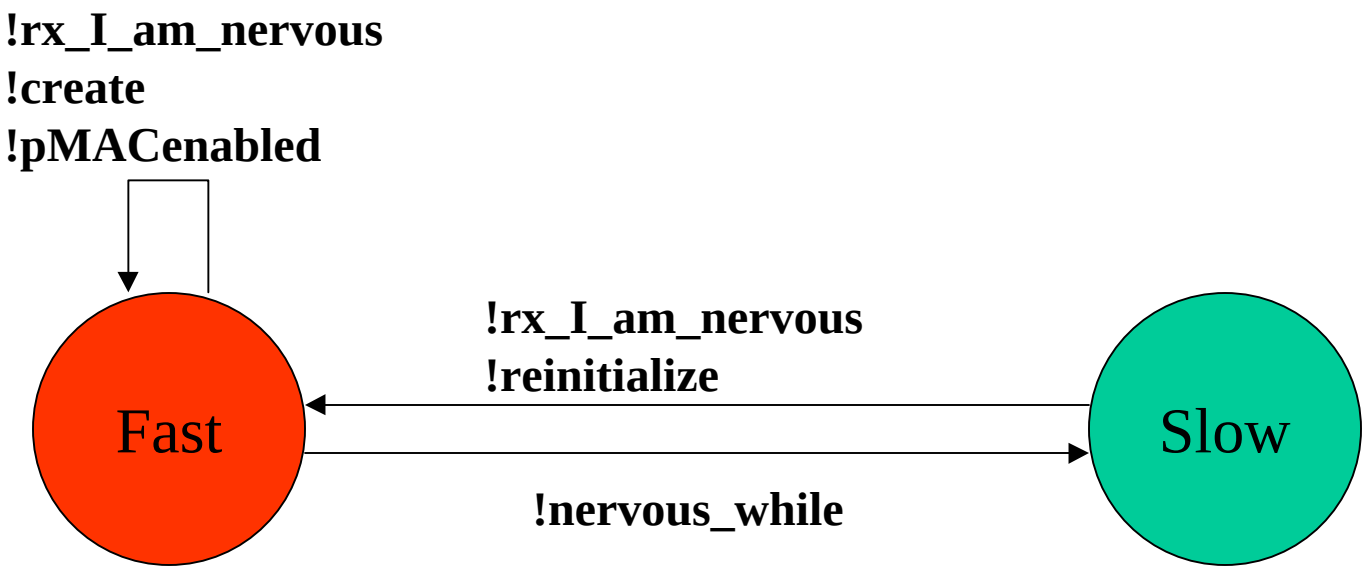
Desirable - Initial state

- Depends on whether
  1) Actor and partner both see h/w status changes
  2) Acceptable to wait before enabling a link
  3) Acceptable to immediately enable a link, then take it down/re-enable on seeing a protocol partner
- Automatic or Desirable Partner if answer to 1 or 3 is yes but 2 is no

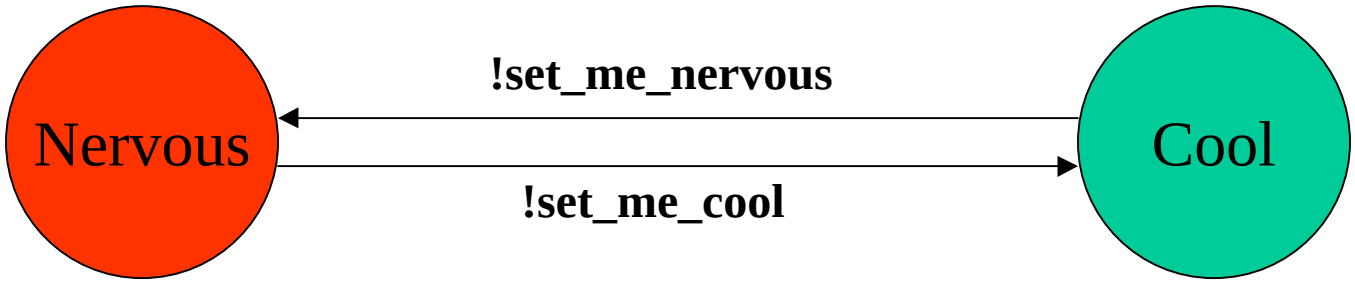
Nervous


New Info
LACPDU
RX
Expired
TX Periodic,
Not Periodic
Desirable
NTT
TX Fast,
Slow
Nervous
NTT
TX PDU
Outgoing PDU
TX
NTT
TX Periodic,
Not Periodic
TX Fast,
Slow
Expired, Matched
NTT
MUX
Attached/Detached
Attach/Detach
NTT
Selection
Expired
NTT
Match
Matched

Nervous - Partner's Anxiety State Machine



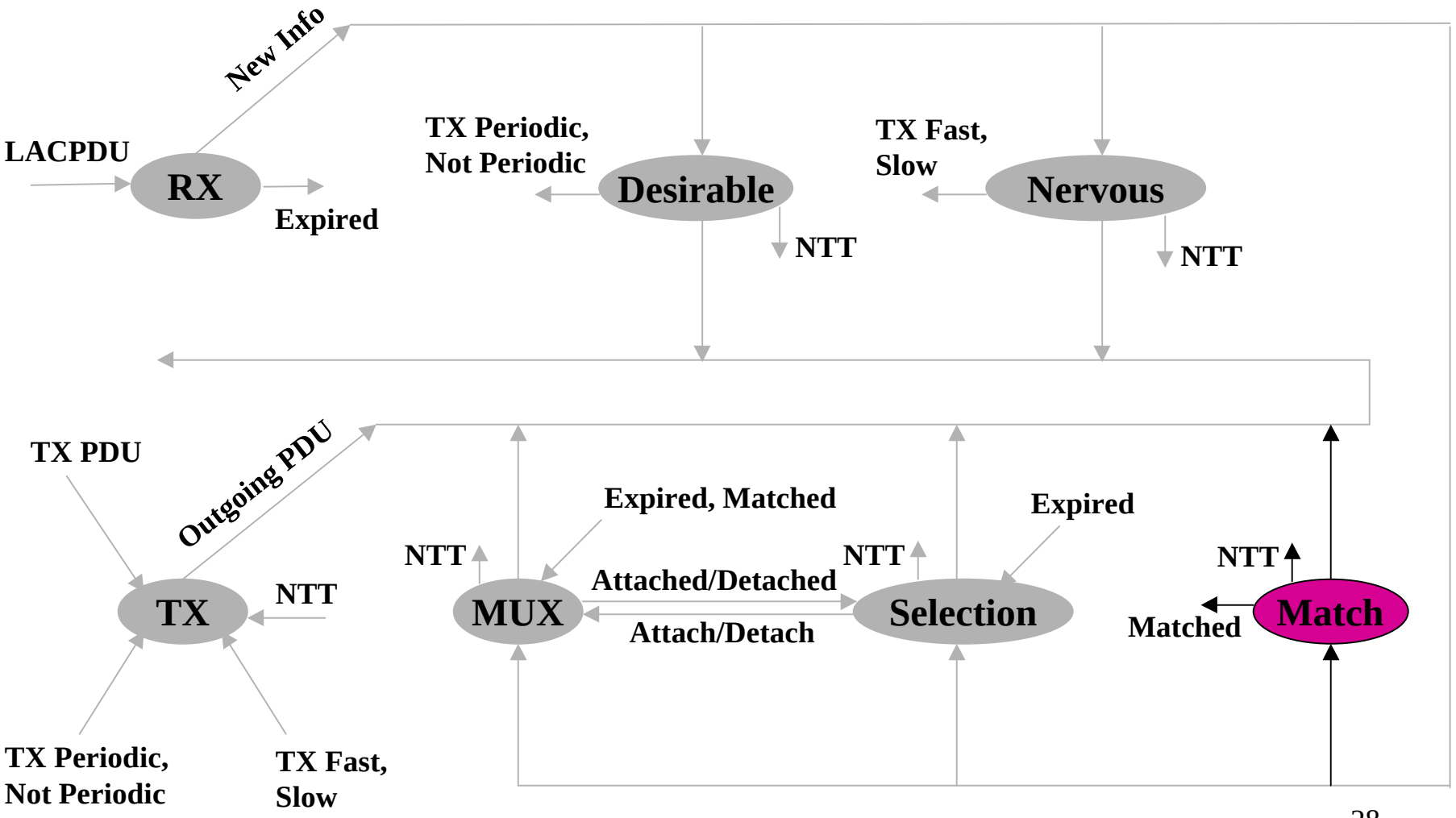
Nervous -
My Anxiety State Machine



Nervous - Functionality Recap

- Controls whether periodic LACPDU transmission is *fast* or *slow*
- Speed depends upon the nervous condition of the partner, not the actor
- Initial state: Partner is nervous

Match


New Info
LACPDU
RX
Expired
TX Periodic,
Not Periodic
Desirable
NTT
TX Fast,
Slow
Nervous
NTT
TX PDU
Outgoing PDU
TX
NTT
TX Periodic,
Not Periodic
TX Fast,
Slow
NTT
Expired, Matched
MUX
Attached/Detached
Attach/Detach
NTT
Expired
Selection
NTT
Matched
Match

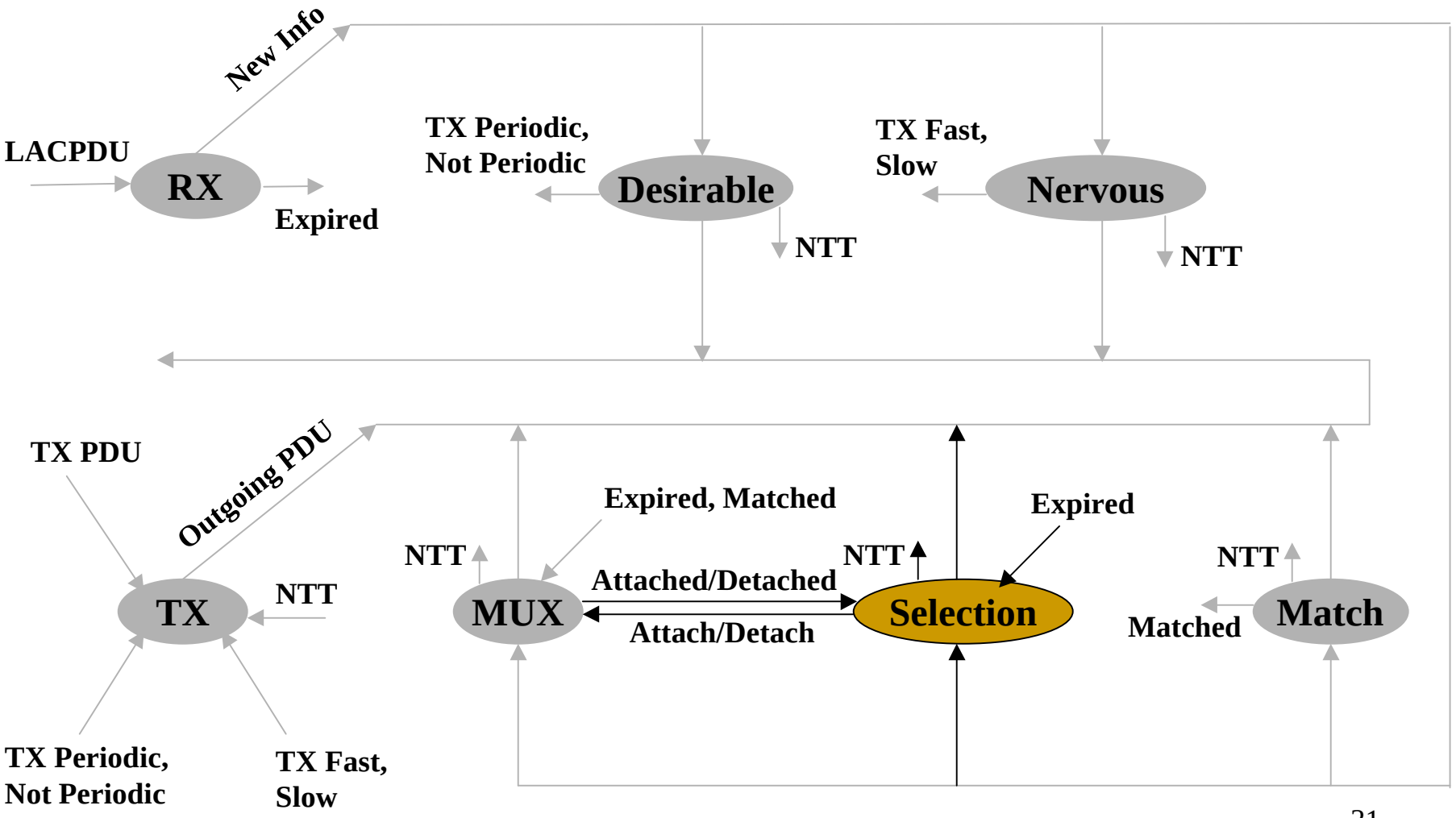
Match Logic

- Matched if:
  - No Partner
  - Matched Individual (Partner believes this link is Individual, or Actor believes this link is Individual & Partner's view agrees)
  - Matched Aggregate

Match - Functionality Recap

- Determines whether participants have both agreed on the protocol information exchanged to the extent that the PhyPort can safely be used in an aggregate
- State of match feeds into Mux state machine
- Initial state: No match

Selection


New Info
LACPDU
RX
Expired
TX Periodic,
Not Periodic
Desirable
NTT
TX Fast,
Slow
Nervous
NTT
TX PDU
Outgoing PDU
TX
NTT
TX Periodic,
Not Periodic
TX Fast,
Slow
NTT
Expired, Matched
MUX
Attached/Detached
Attach/Detach
NTT
Expired
Selection
NTT
Matched
Match

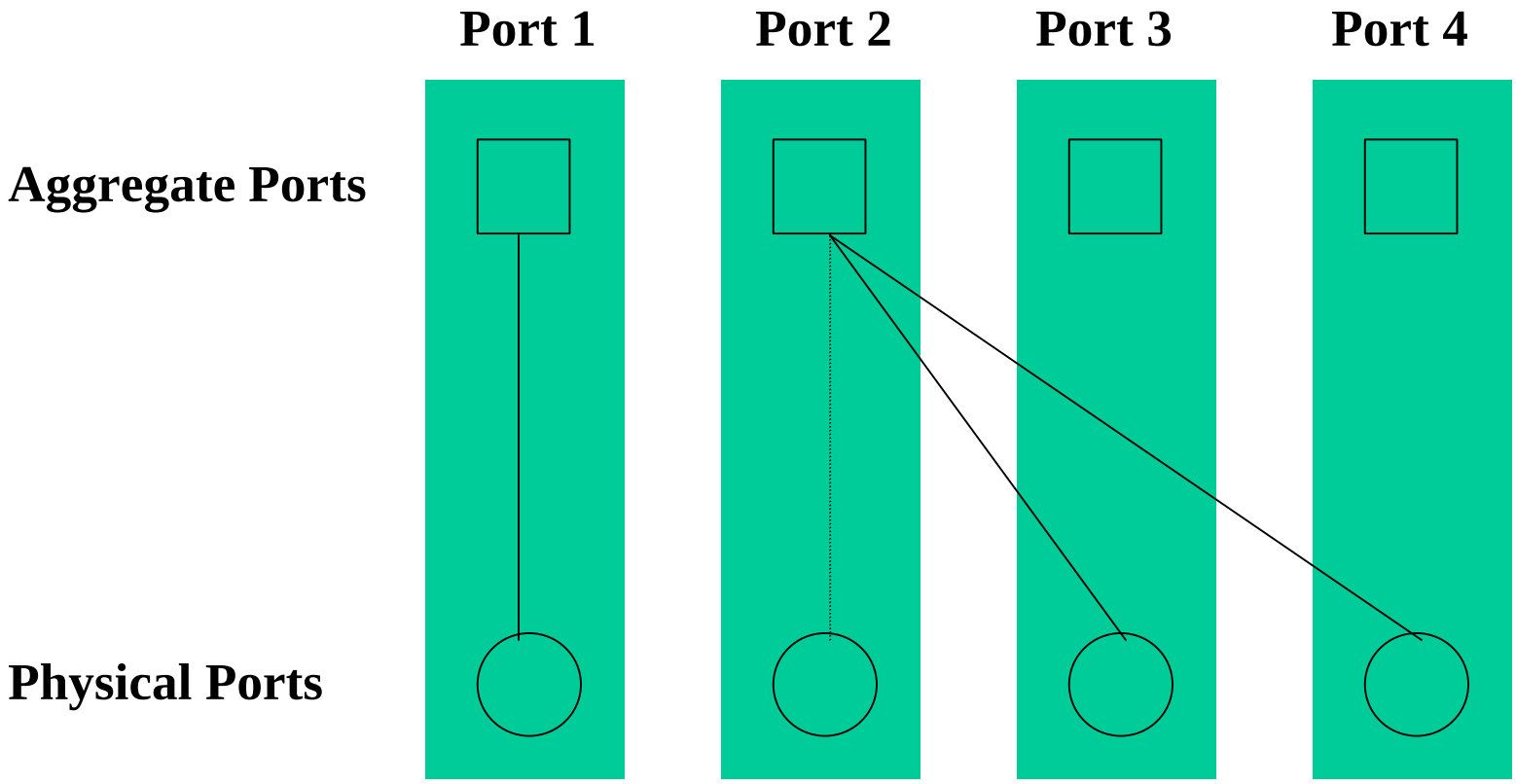
Selection - Assumptions

- No additional MAC addresses (over those allocated per physical MAC) required
- Determinism in allocation of PhyPorts to AgPorts
- Result is intuitive to the user
- Compatible with alternative views

Selection - Rules

- Each MAC has a PhyPort and an AgPort
- Aggregation = attachment of a PhyPort to an AgPort (its own, or someone else's)
- Each PhyPort is always attached to one AgPort
- The PhyPort of an individual link always attaches to its own AgPort
- The lowest numbered AgPort is always used for an aggregate, even if its PhyPort is disabled

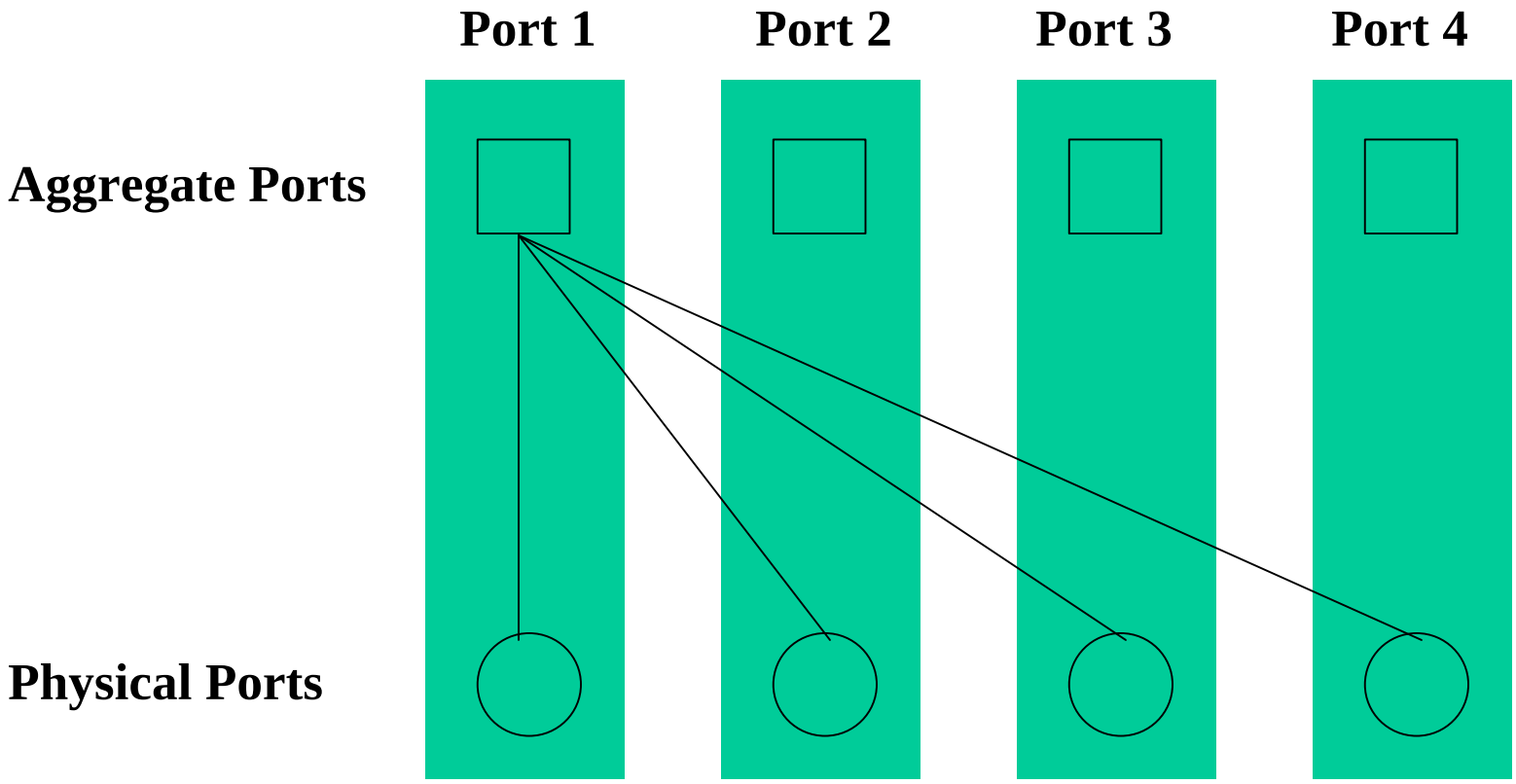
Selection - Legal Example 1



Selected

Selected & Attached

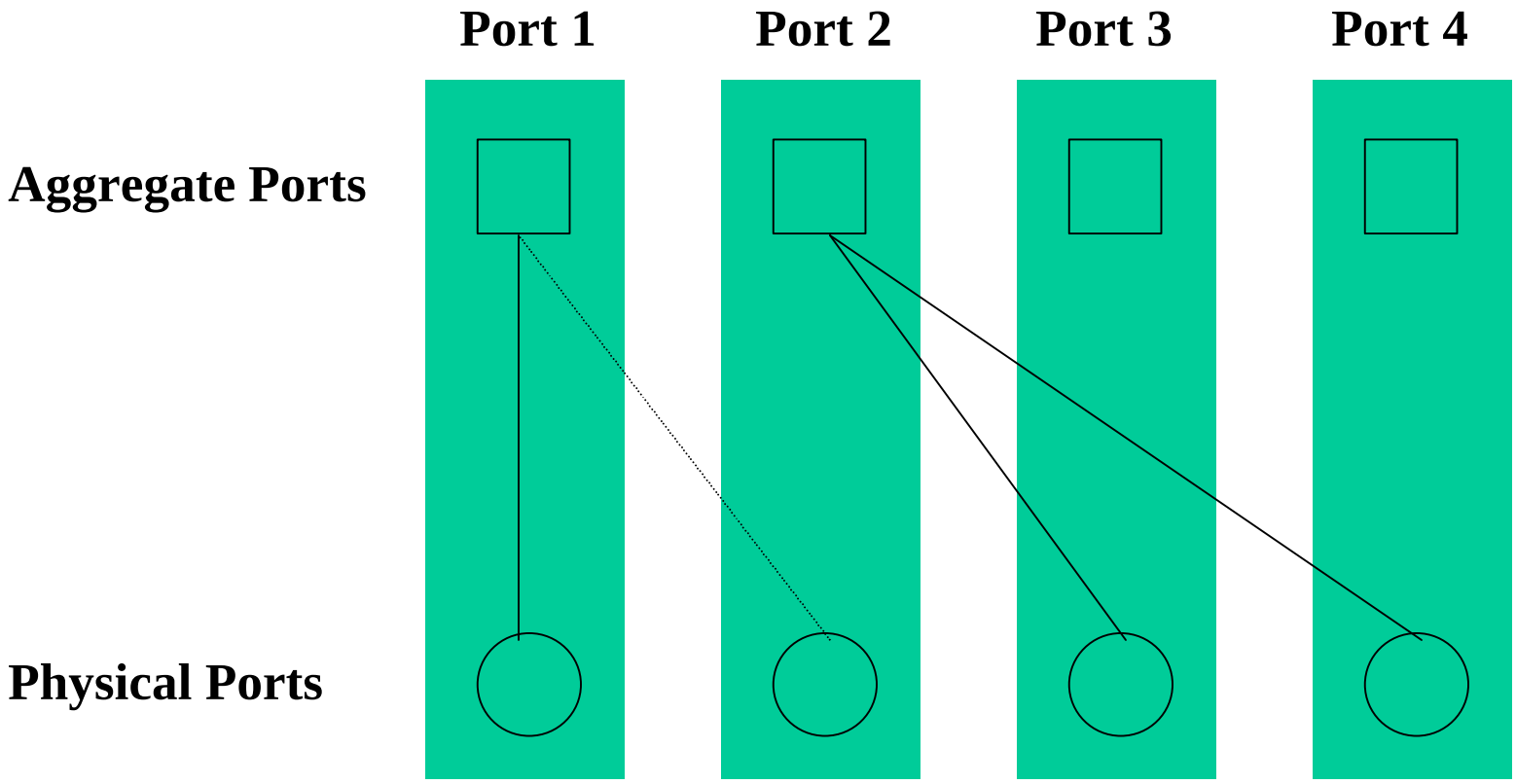
Selection - Legal Example 2



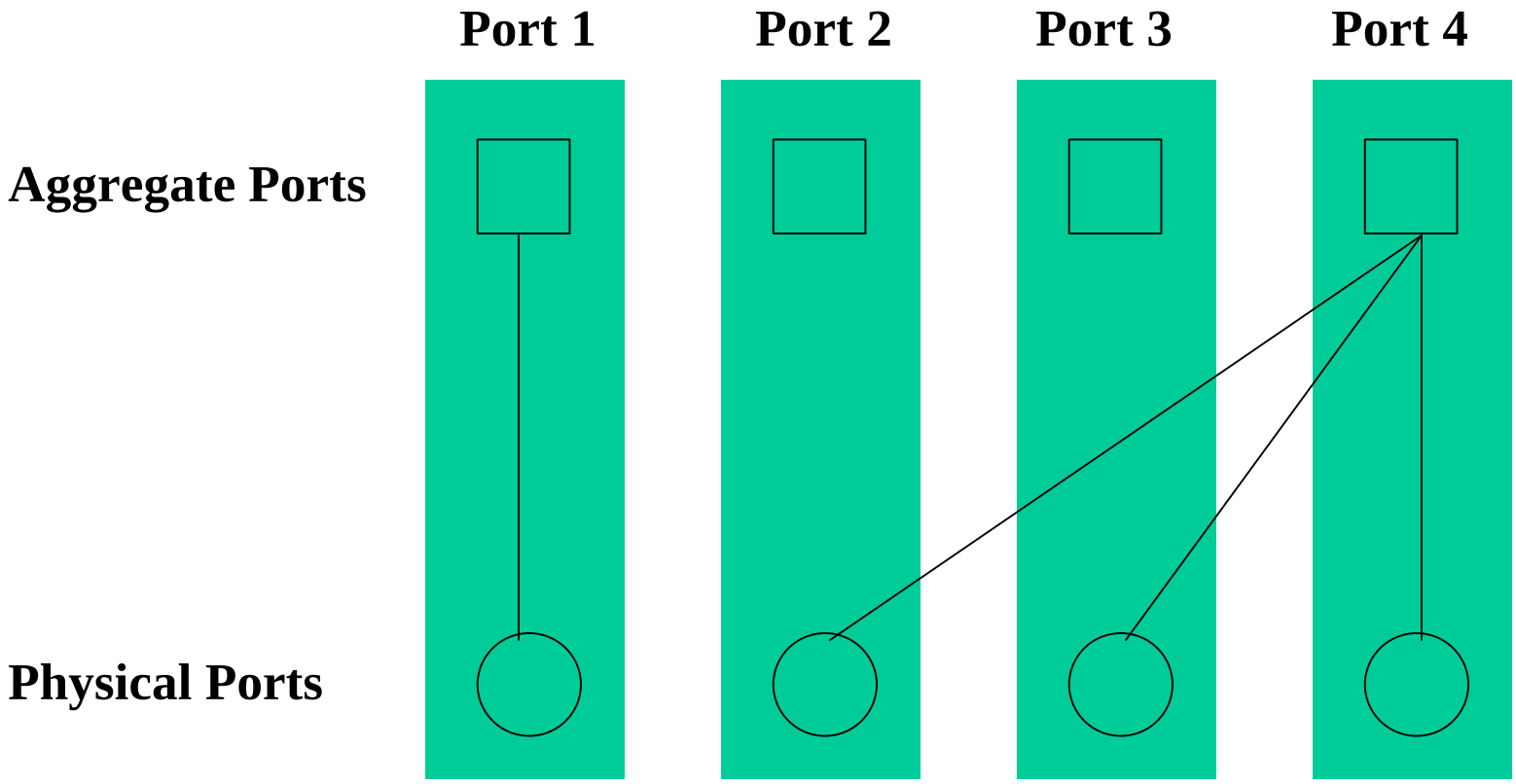
Selected

Selected & Attached

Selection - Illegal Example 1



Selection - Illegal Example 2



Selected

Selected & Attached

Selection Logic (1)

- Determines
  - Partner's System ID and Key
  - Whether this link is an individual link
  - Whether the partner has changed (ID or Key)
- Updated on
  - New information received
  - Selection Wait time expiry
  - Management changes to my parameters

Selection Logic (2)

- Individual if
  - RX machine is expired
  - Actor believes the link to be individual
  - partner believes the link to be individual
- If Individual, AgPort selected is own AgPort
- If not Individual, AgPort selected is lowest numbered AgPort with same local/remote system ID & Key

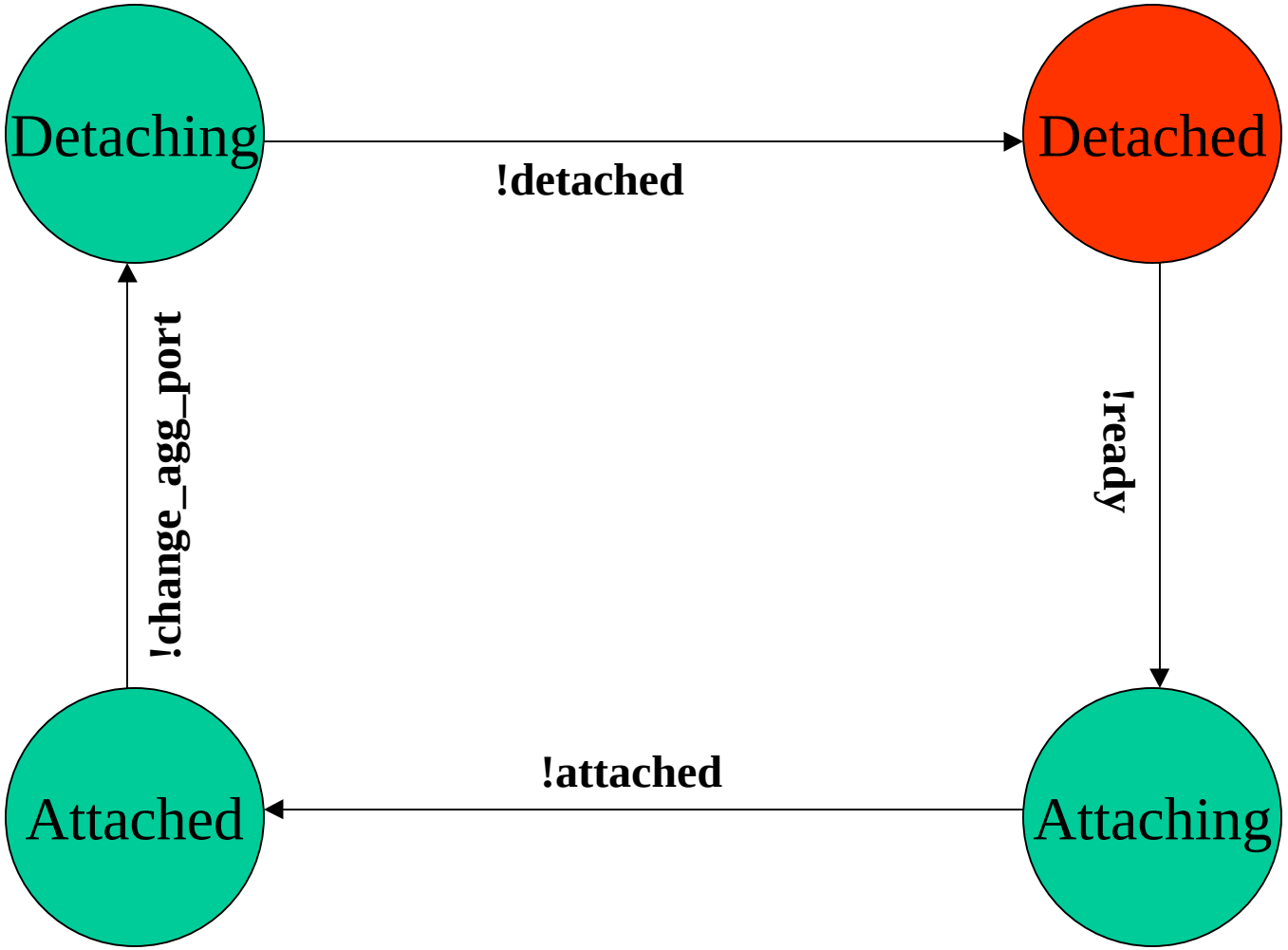
Selection Machine

- Attaches PhyPort to selected AgPort
- On change of selection
  - Detaches PhyPort from old AgPort
  - Waits for dust to settle
  - Attaches to new AgPort
  - May involve evicting other PhyPorts from their current AgPort

Selection States

- Detached, Attaching, Attached, Detaching
- Equivalent to:
  - attach (Administrative state - signalled to Mux)
  - attached (Operational state - signalled from Mux)

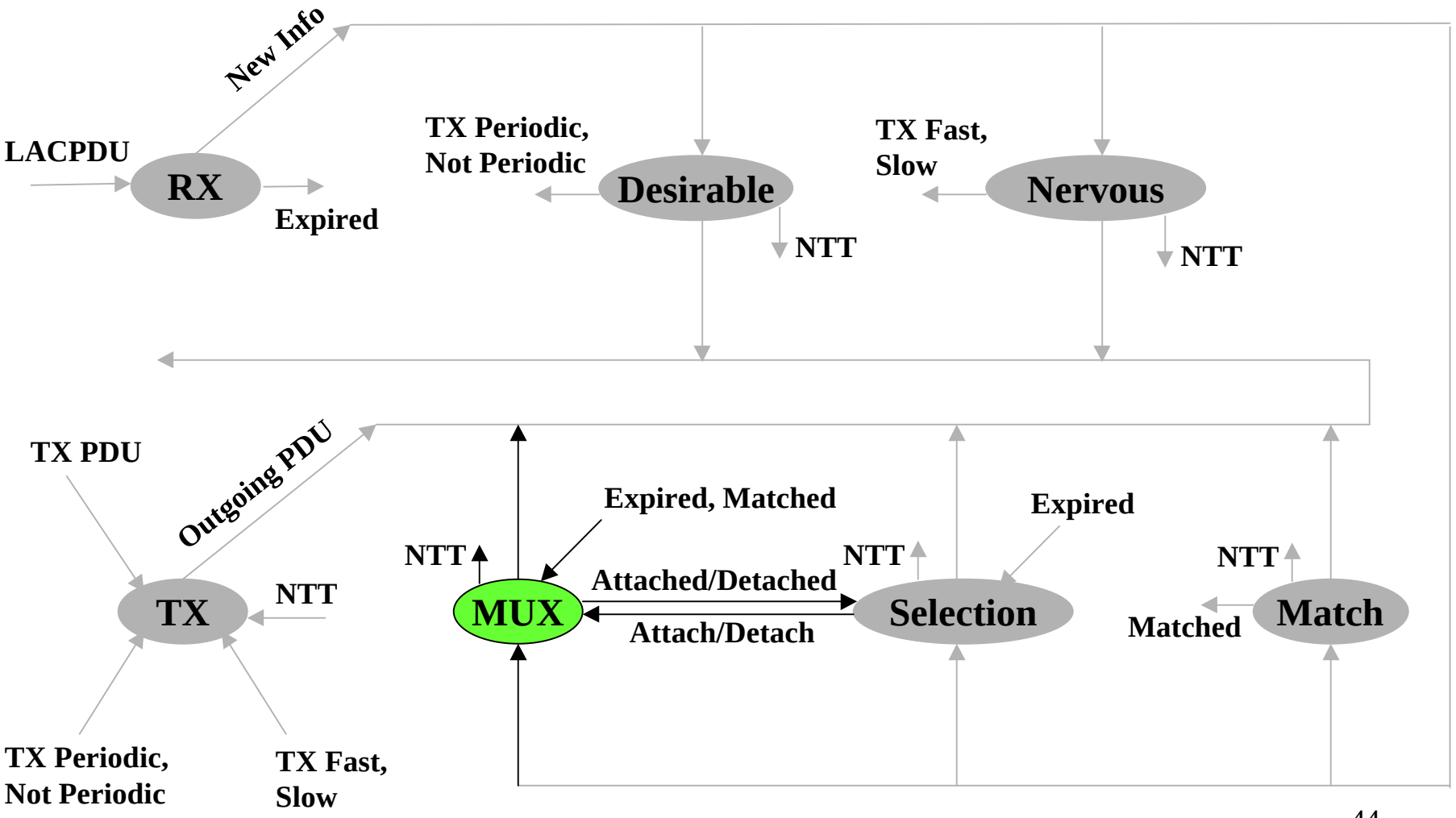
Selection - State Machine



Selection - Functionality Recap

- Determines whether the link is in the right aggregate or not
- If not in the right one, removes it
- If not in an aggregate, finds the right one for it to be in and adds it
- Takes account of the need to wait for other links to select the same aggregate

MUX


New Info
LACPDU
RX
Expired
TX Periodic,
Not Periodic
Desirable
NTT
TX Fast,
Slow
Nervous
NTT
TX PDU
Outgoing PDU
TX
NTT
TX Periodic,
Not Periodic
TX Fast,
Slow
Expired, Matched
NTT
MUX
Attached/Detached
Attach/Detach
NTT
Selection
Expired
NTT
Match
Matched

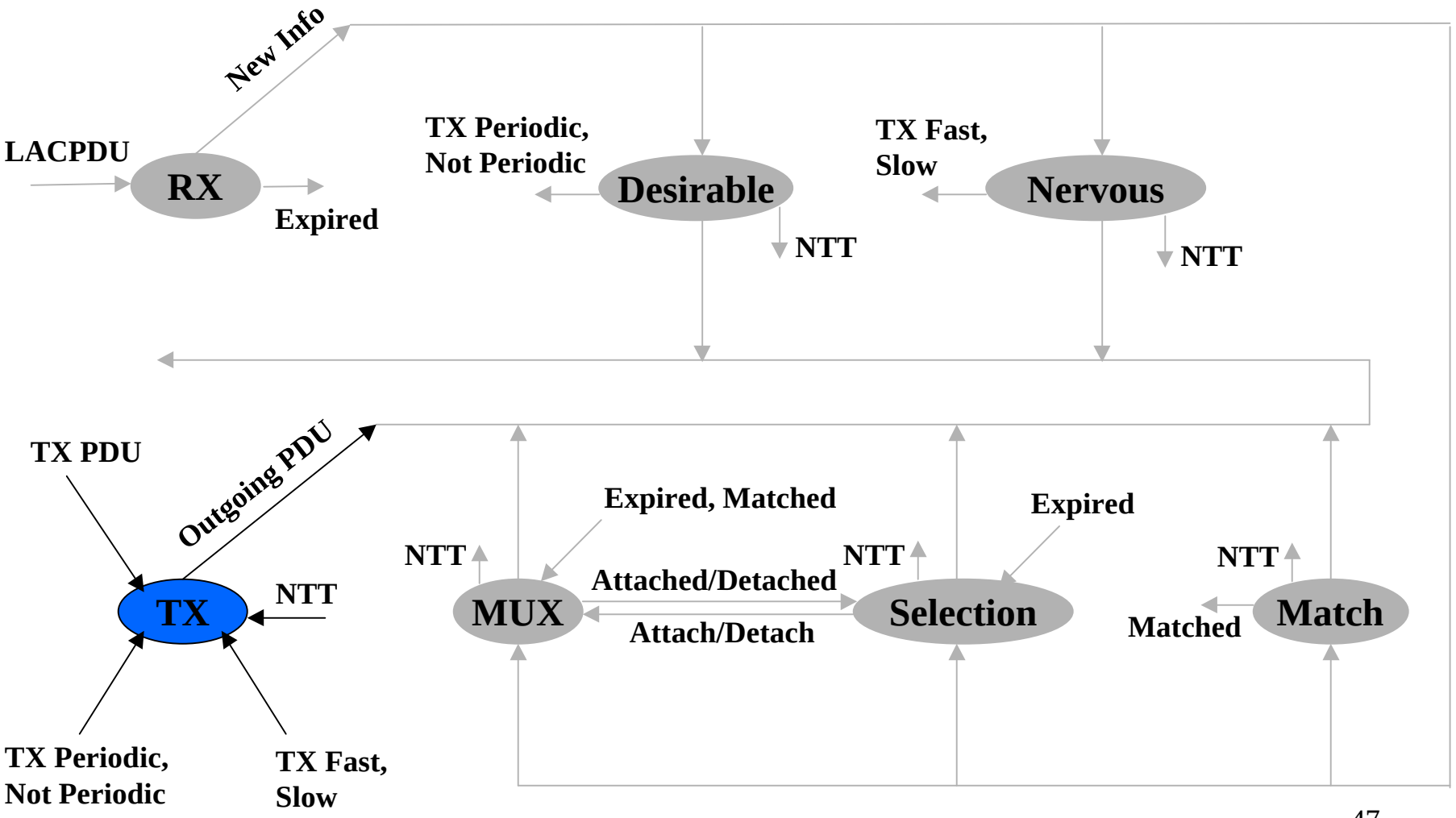
Mux - States and Goals

- States: In Sync, Out of Sync
- Goals
  - Partner or Actor Out of Sync: turn off collector & distributor
  - Actor and Partner In Sync: turn on collector
  - Actor and Partner In Sync, Partner's Collector is on: turn on distributor
  - Above rules also apply to **coupled** mux h/w
  - If mux h/w is **independent**, and if Partner's collector is turned off, then turn off distributor

Mux - Functionality Recap

- When *in synch*, takes the necessary steps to turn on collector and distributor
- When *out of synch*, takes the necessary steps to turn off collector and distributor
- Signals *attached*, *detached* when its done
- Initial state: off

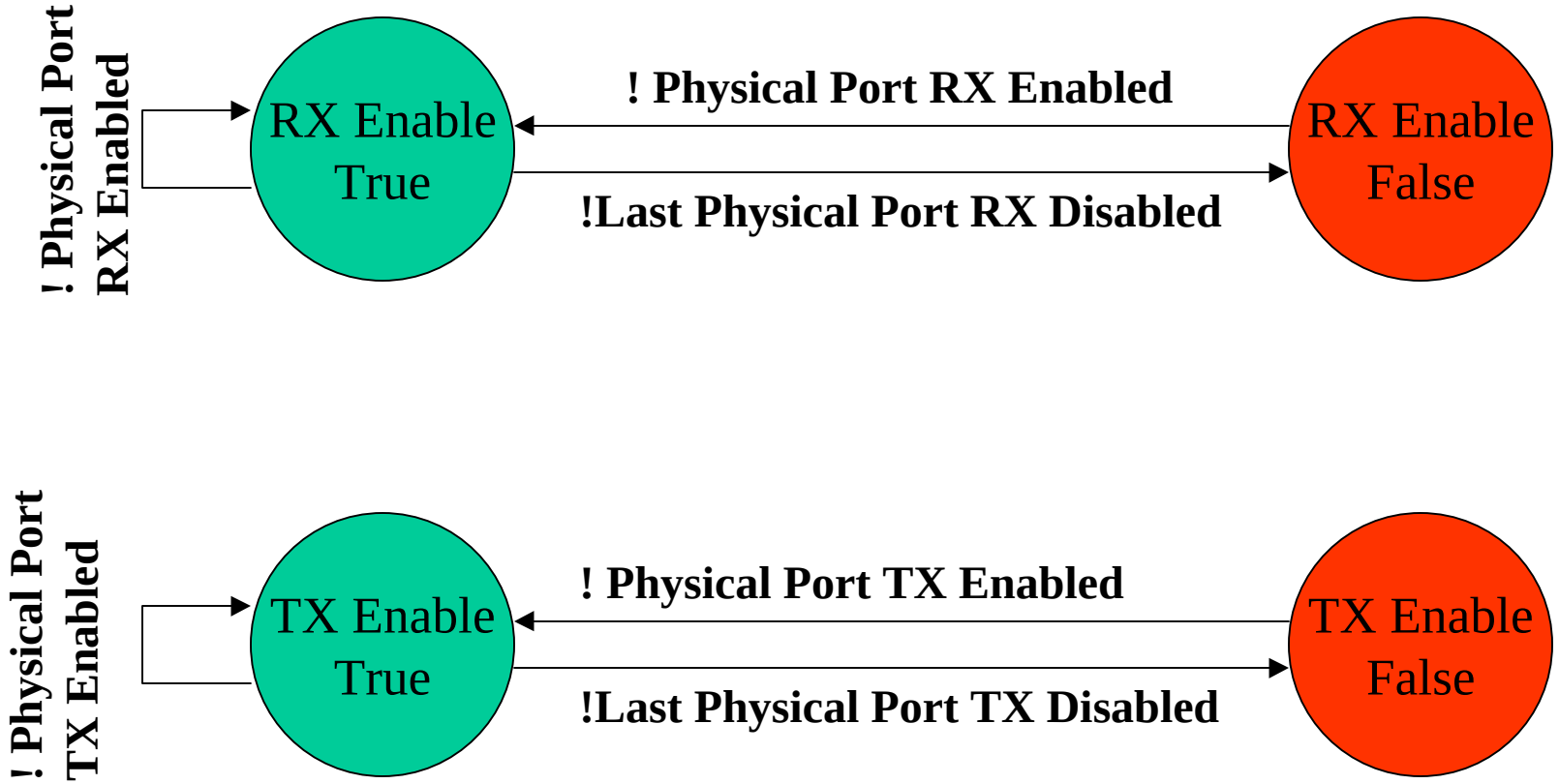
TX


New Info
LACPDU
RX
Expired
TX Periodic,
Not Periodic
Desirable
NTT
TX Fast,
Slow
Nervous
NTT
TX PDU
Outgoing PDU
TX
NTT
TX Periodic,
Not Periodic
TX Fast,
Slow
Expired, Matched
NTT
MUX
Attached/Detached
Attach/Detach
NTT
Selection
Expired
NTT
Match
Matched

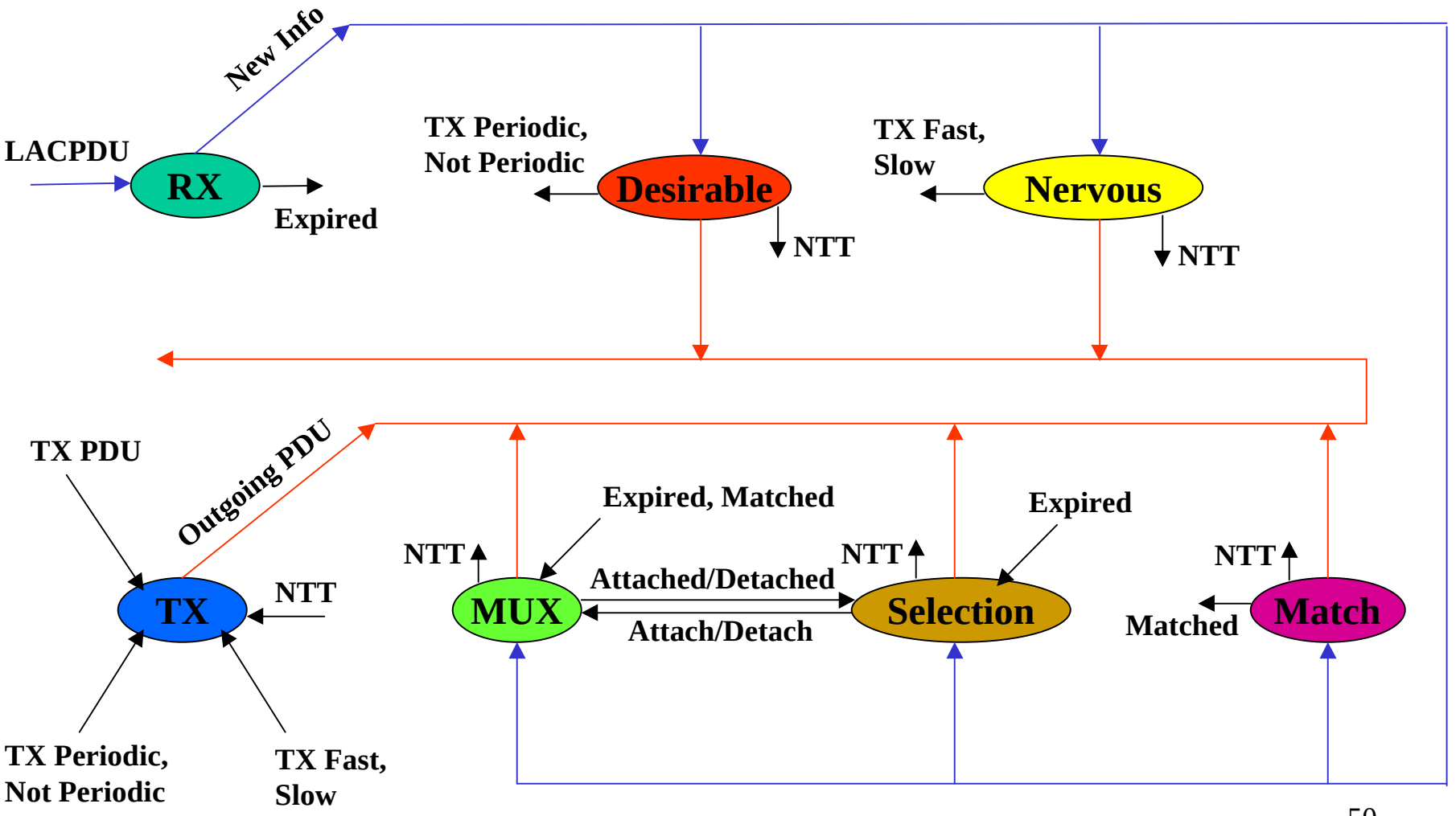
TX - Functionality Recap

- Causes LACPDUs to be generated if:
- NTT
- Desirable
  - Frequency depends on *fast* or *slow*signal from Nervous state machine

Aggregate Port - State Machines



The Big Picture



Summary

- Covers (majority of) functionality described by Finn/Wakerly/Fine & Jeffree
- Fully describes the process of reaching agreement & the actions taken to join & leave aggregations
- Separate state machines improve clarity
- Flush protocol yet to be included