Link Aggregation Protocol Techniques

802.3 Trunking Study Group
IEEE Interim, April 1998

Norman Finn, Michael Fine, John Wakerly
Cisco Systems, Inc.

Outline

- Requirements
- Scope of this contribution
- Overall scheme:
  - Per-link state machine
  - Aggregation Monitor
- Expressing aggregation possibilities
- Choosing the links to aggregate
- Related issues

Requirements

- Initial bring-up of aggregated link may take on the order of 1 second.
- Creating or adding to an aggregation must favor reliability over speed where those traits conflict.
- Support required for incremental addition/deletion of links to/from an aggregation.
- Loss of one member of a link must be detected as quickly as possible to prevent black-holing packets.

Scope of this Contribution

- Describes key elements of a protocol for creating or augmenting an aggregation.
- Does not directly address removal of a link from an aggregation.

Overall Scheme

- Each device runs an aggregation monitor to decide which links aggregate together.
- Each link runs a state machine which:
  - ensures that the link is bi-directional; and
  - exchanges information required by the aggregation monitor.
- Monitor reprograms hardware and initiates synchronization when links are in appropriate state(s), and when partner is ready, enables the aggregated link.

Per-Link State Machine (1)

- Link transmits LinkAg protocol packet regularly every time T. Packet includes:
  - my device ID and transmitting port ID
  - partner's device ID and transmitting port ID
  - aggregation possibilities
  - synchronization flag
- Both ends reach the same conclusion as to which links to aggregate, because they use the same algorithm.

Per-Link State Machine (2)

S1-> S2: Link comes up

-> S3: Data frame received

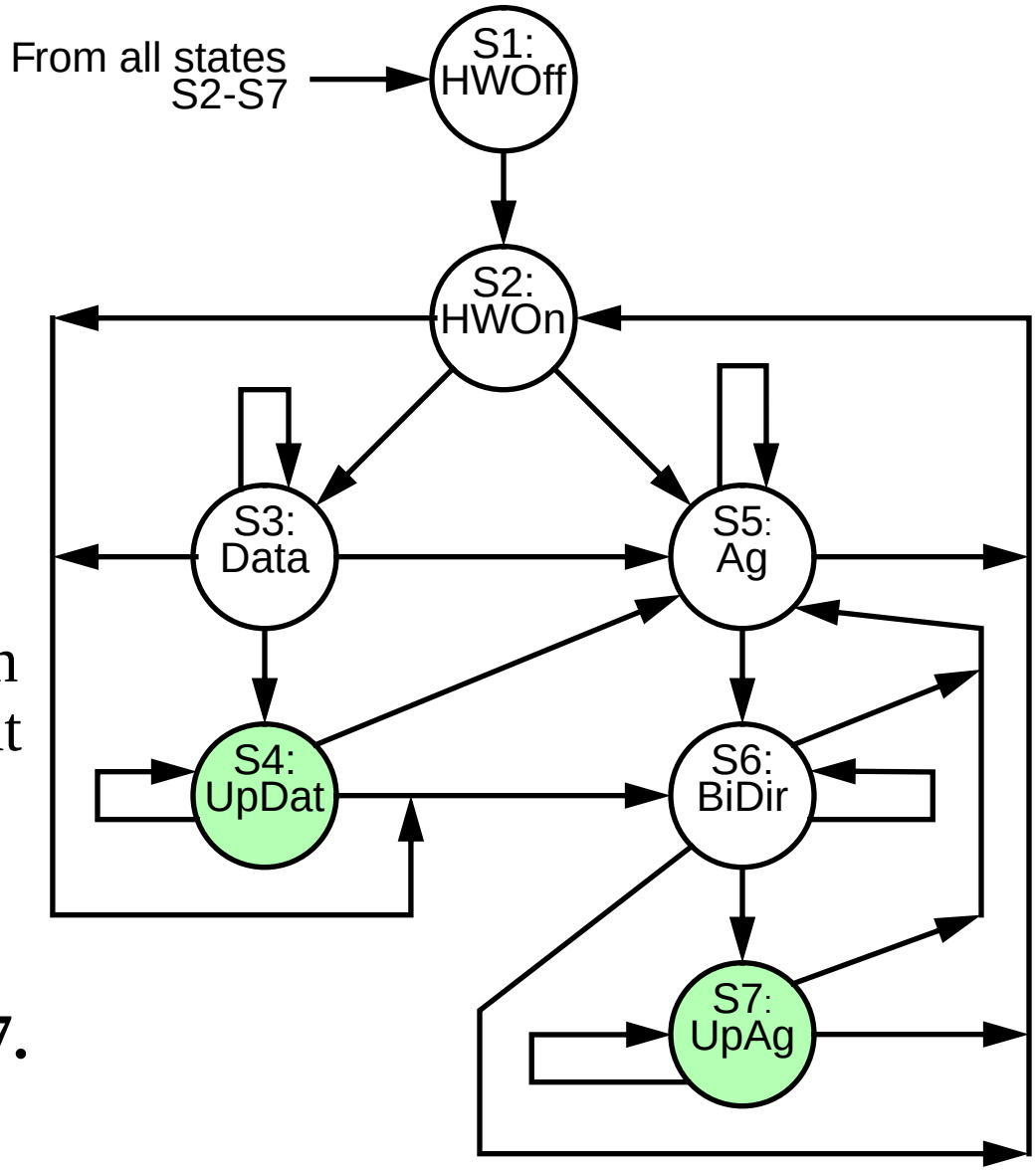
-> S4: Time-out: data, but no LAg packet received for some time

-> S5: LAg packet received not mentioning me

-> S6: Bi-directional link known (LAg packet received that mentions me)

-> S7: Second bi-directional LAg packet received

Link is usable only in S4 or S7.



Aggregation Monitor

- When any link reaches S4/UpDat, link is “aggregated” as a single link. (This device is connected to a non-LinkAg device.)
- When one link reaches S7/UpAg, link is aggregated with all other links attached to same device and aggregable together which have reached either S6/BiDir or S7/UpAg.
  - Delays aggregation one cycle (time T) in order to come up all at once, instead of one link at a time
- Enable aggregation when partner is ready.

Expressing aggregation possibilities (1)

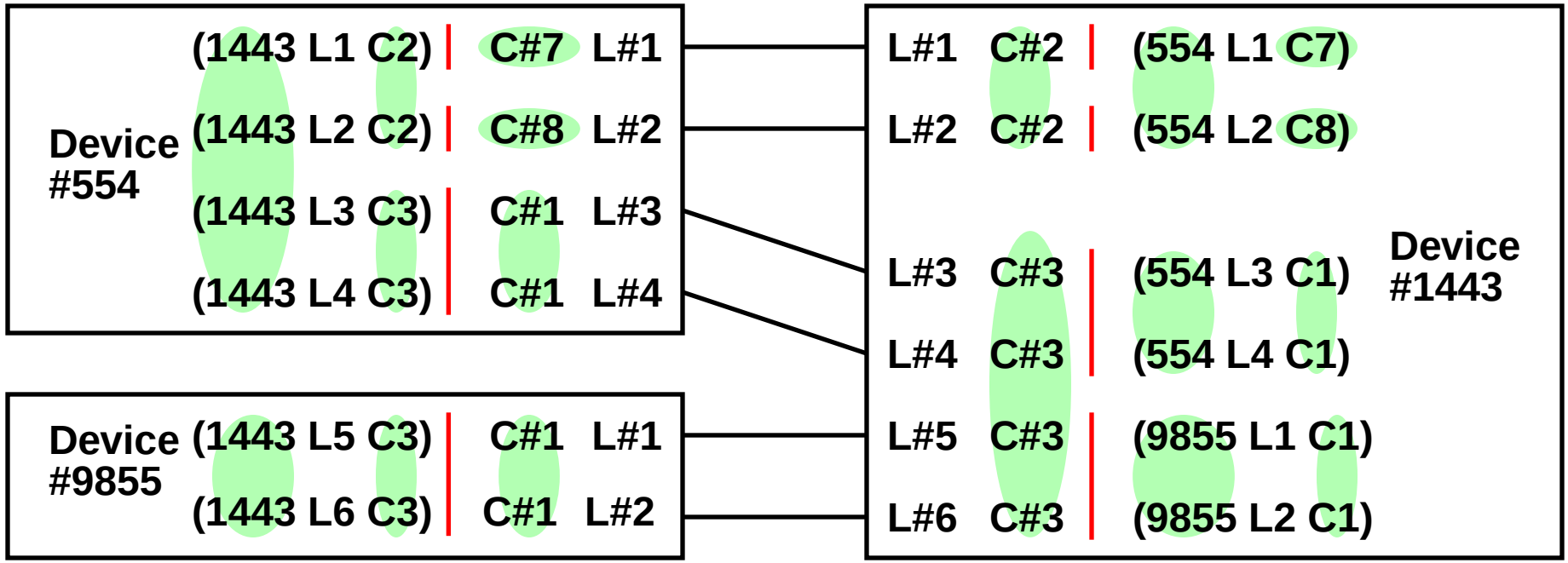
- Many hardware- and software-imposed limitations on aggregation, e.g.:
  - Aggregate only by pair or 4-tuple
  - Aggregate only on same line card
  - Aggregate any links among 16 up to maximum of 4
  - Aggregate any combo, but one link must be chosen from set of “master” links
- Enumeration among o(n!) possibilities is sufficient, but not practical.
- Must provide sysadmin control.

Expressing aggregation possibilities (2)

- Each link assigned a “category” number.
  - Category assigned by transmitting device
  - Category number has meaning only among links on same device, and only for comparison for equality with other links’ category numbers.
  - Category number is constant, modulo changes in HW parameters such as link speed or full-duplex.
- Category numbers may be overridden by sysadmin to further limit aggregation.
- Not perfectly expressive, but captures most capabilities.

Choosing the links to aggregate

- Two links may be aggregated together if and only if they have the same <device number, category number> at both ends of the link.
- Monitors form largest possible aggregation.



Related Issues

- Where Category numbers are inadequate, initialize number to collect too much
  - Can alter numbers dynamically towards less aggregation.
  - Category numbers cannot change towards greater aggregation without hardware change (e.g. disconnect) or operator intervention.
- Need more controls on transmission of Link Ag packets to minimize “hello” traffic.
- Details/controls can be added: allow “silent partners”, manage multi-point connections...