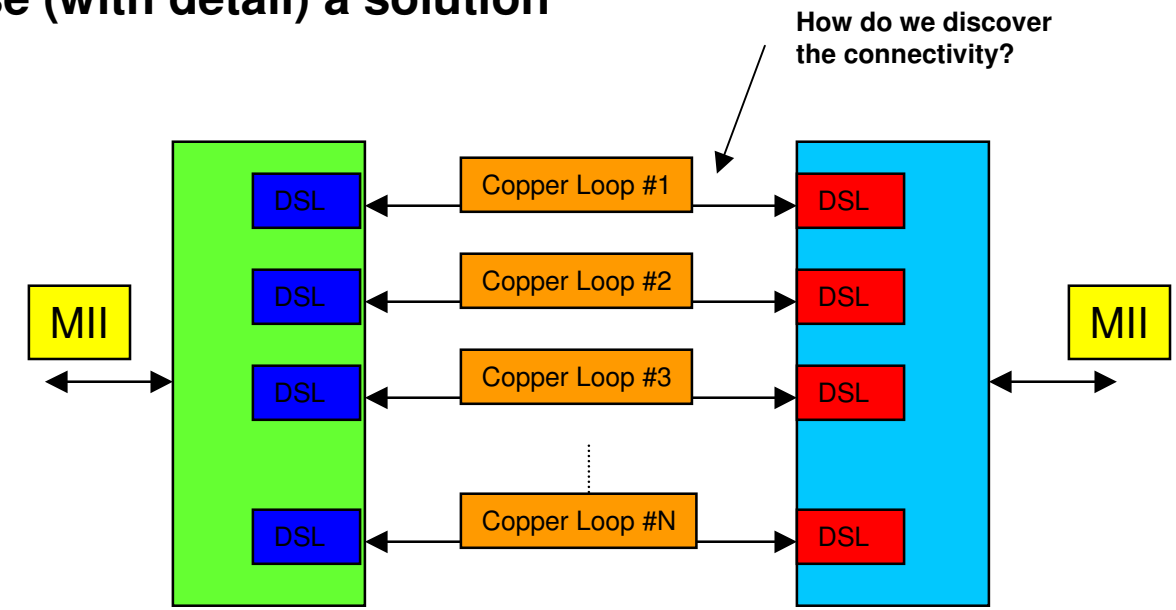
Loop Aggregation Discovery

Hugh Barrass

(supplemental to baseline proposed by Klaus Fosmark)

Loop Aggregation Discovery

- **Loop aggregation as described by Klaus Fosmark (fosmark_1_0302.pdf) requires a simple and resilient discovery mechanism**
- **Consider options**
- **Propose (with detail) a solution**



Definitions and assumptions

- **Define "aggregateable PHY"**

  Physical layer interface for multiple PMI's which supports aggregation across those PMI's. This may be one or more devices. There may be multiple separate "aggregateable PHYs" within one high density device.

- **Define "dumb CPE"**

  A CPE device which does not have any capability (or intention) to control any part of the physical layer connection. Such a device may not have a processor capable of communicating across the MDIO interface. It may not be programmed with a MAC address or be capable of responding to MAC control frames.

- **Assume aggregation control as described by Klaus**

  At least within each "aggregateable device"

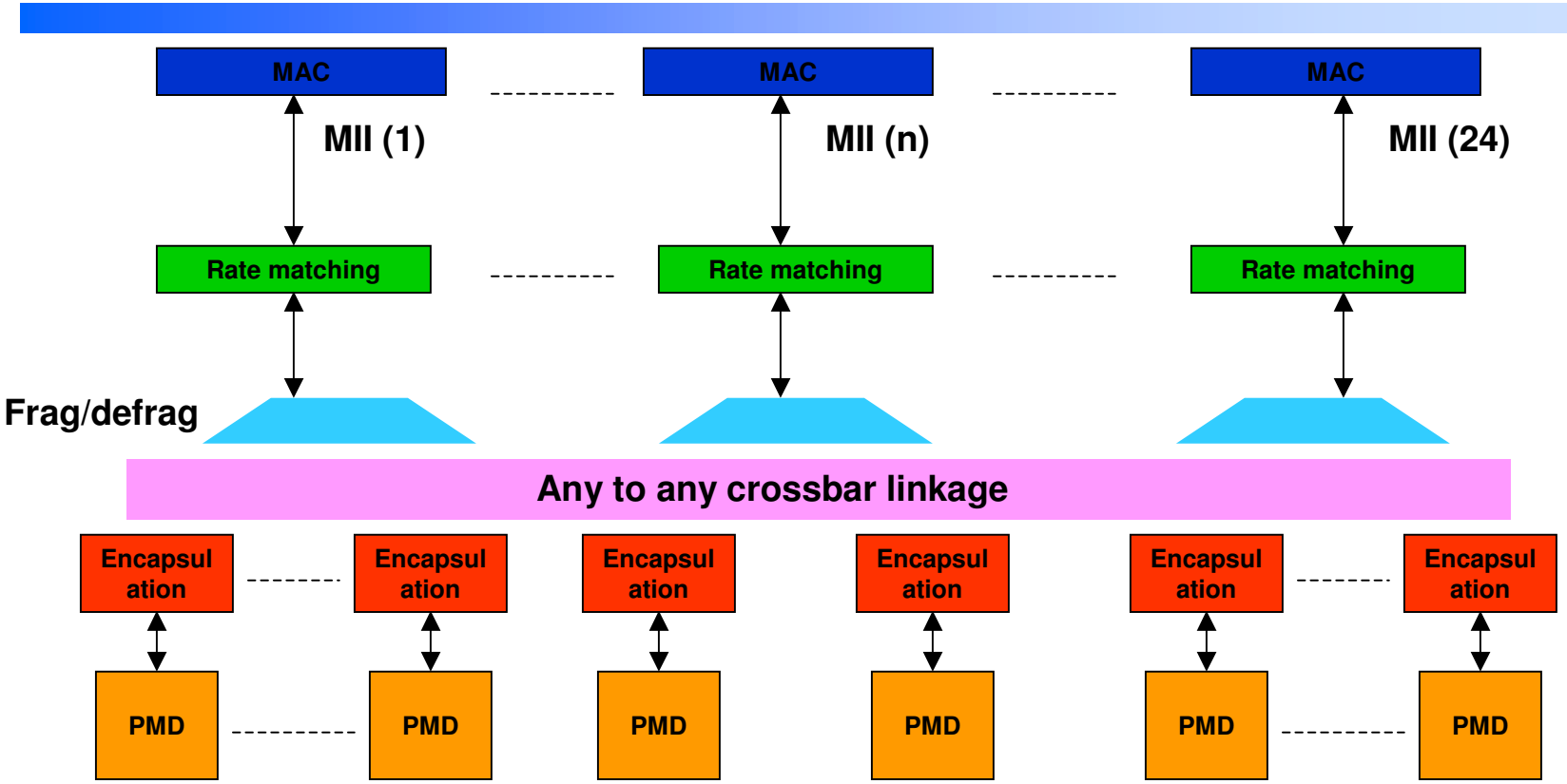
- **Assume host (LT) is smart, CPE may be dumb**

  Mechanism must work in the presence of this asymmetry

- **Assume only default connectivity with remote PHY devices**

  At least one PMI will be connected to an MII, no more can be assumed

Reminder of “dream system”



- Any connection between MII’s and PMI’s
- Implementation could use arbitrary number of MII’s vs PMI’s
- System vendor could choose optimal ratio

Discovery problems

- **Default settings at either end may not permit passage of frames across all loops**
  - **e.g. # PMI > # MII ➔ some PMI's not used until loop aggregation initialized**
- **"Loop confusion" means that no assumptions can be made about connectivity**
  - **Can't assume that PMI-2 at LT connects to PMI-2 at NT (even if PMI-1 connects)**
- **Possibility that alien devices are performing similar function in same plant**
  - **More unbundling difficulties…**
- **Host may address PCS (& thus MII) directly, or may address PMA/PMD (thus PMI) directly**
  - **Mechanism for explicit access to EOC/VOC for a specific loop**

Discovery options

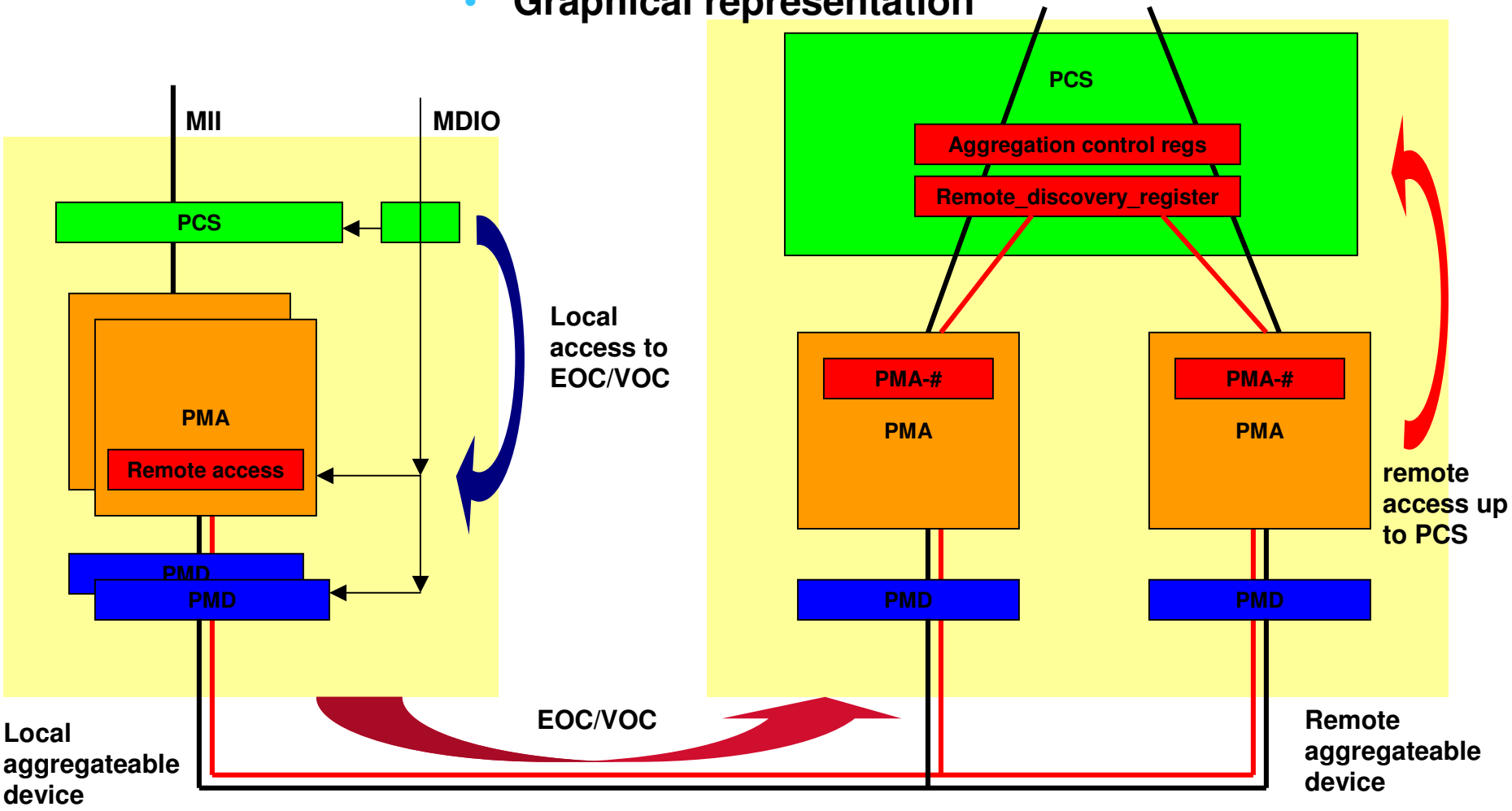
- **Extend 802.3ad - LACP**
  - **Differences to link aggregation too great**
  - **Danger of overloading existing function (& breaking it!)**
- **New frame based discovery protocol**
  - **Frames not passed across all loops by default**
  - **Doesn't cater for dumb CPEs**
  - **No way to tell MAC control which PMI received the frame**
- **Use new PHY layer mechanism**
  - **Low level, limited flexibility**
  - **Keeps "architectural purity" of layers**
- **==> Propose PHY layer mechanism**

Required PHY functions

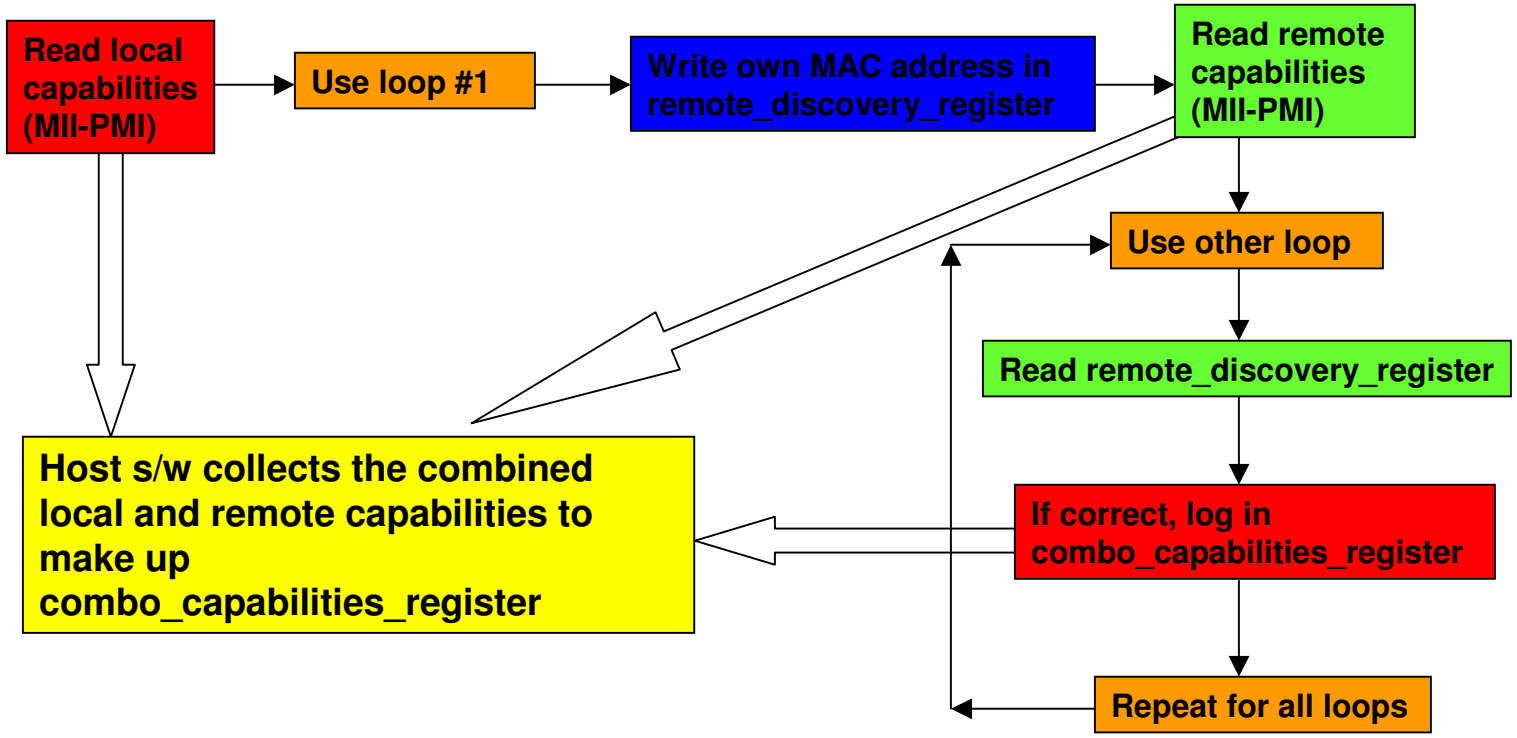
- **Mechanism for MDIO access to local & remote PMA/PMD, PCS**
  - Already a requirement for OAM, PMD and loop aggregation control
- **PMA/PMD address remotely readable**
  - Physical (relative to device) or logical (relative to system)?
- **Remotely writeable and readable register in PCS**
  - Ideally >48 bit, may require multiple operations to read/write
  - Only needed in PCS #1 in any aggregateable PHY device
  - Referred to as “remote_discovery_register”
- **Host access to remote register**
  - Uses MDIO addressing to access local PMA/PMD (1-32)
  - This gives access to remote PMA/PMD across loop
  - Read/write “remote_discovery_register”

Required PHY functions (2)

- Graphical representation



Discovery process



NB: could use other unique code instead of MAC. Also could add pseudo-random extension to MAC to increase security

- Note that this process is implemented in the host system
  - It is assumed that the host system will optimize serial vs parallel implementation
- This process must be repeated for each "aggregation domain"

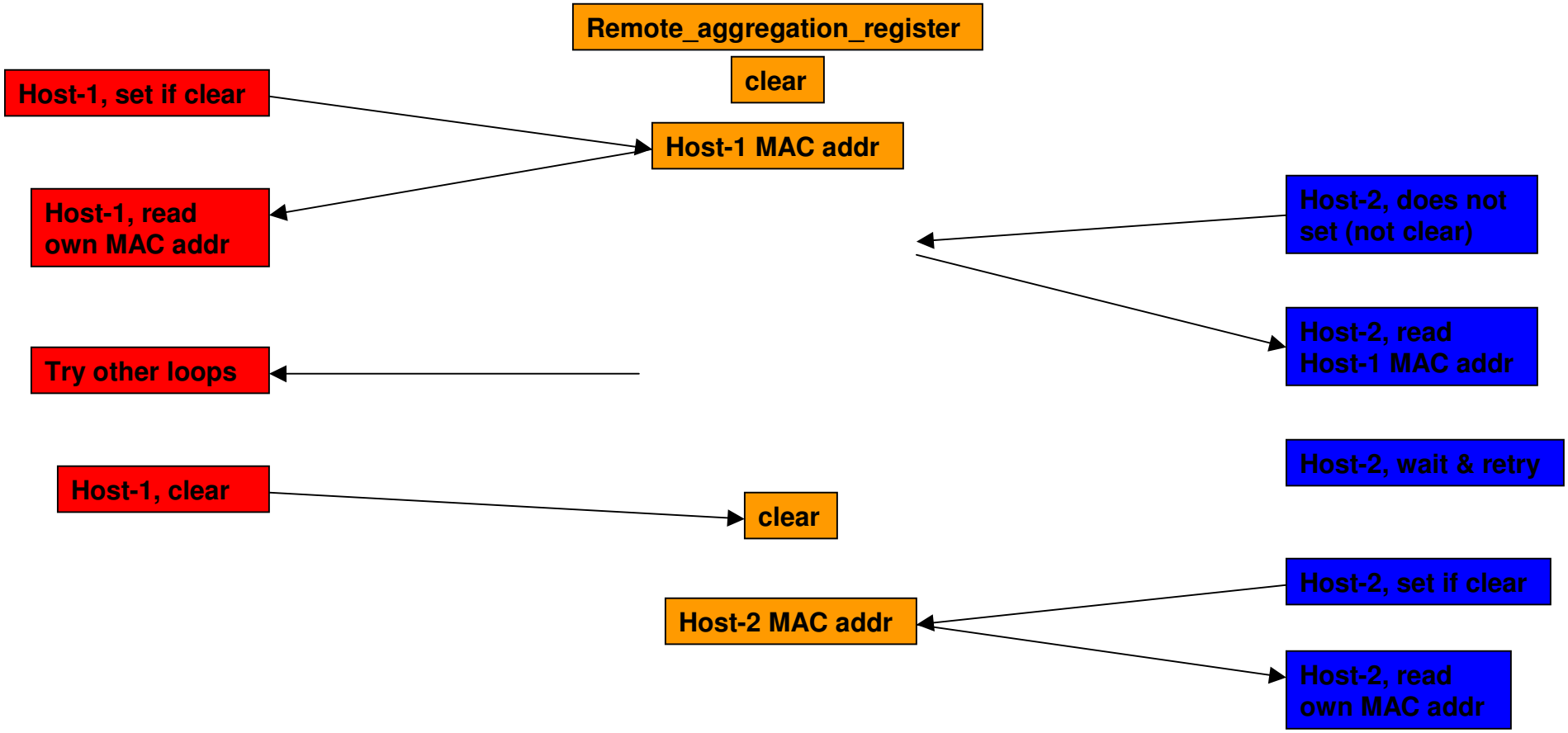
After discovery

- Host has a map of all possible aggregation possibilities
- 802.3ah does not specify how to use aggregation
  - Manual, automatic etc.
- Control of loop aggregation is master-slave
  - Similar approach to PMD control
  - Program remote aggregation control registers, then local

Problem!

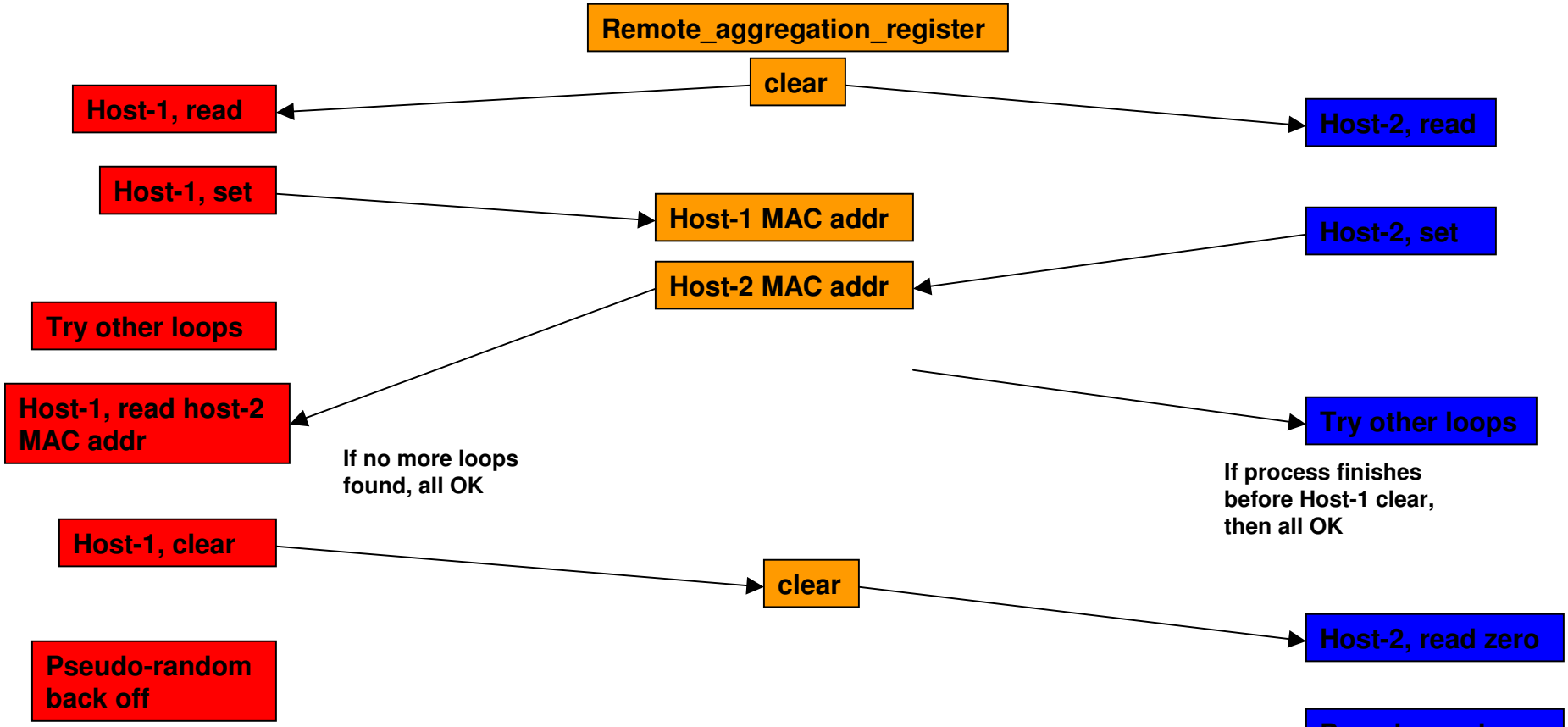
- What if two hosts are trying to discover in the same binder?
- Needs a mechanism for collision detection
  - … and then back-off or avoidance
- Write host MAC address into remote_aggregation_register
  - Gives means of collision detection
  - Needs atomic operation for guaranteed collision avoidance
- Explore collision scenarios with and without atomic operation
  - “read register and write xxx if zero”
  - Similar to operations used in parallel computing

Collision detection/avoidance



- Requires atomic "set if clear" operation
- CPE register should self-clear after timeout period

Collision detection/back off



- No atomic operation
- In most cases, read then set is not interrupted by other host => same as for atomic operation

Summary

- Method for Loop Aggregation discovery
- Compatible with Klaus' baseline
  - fosmark_1_0302.pdf
- Uses PHY layer communication for control of PHY layer aggregation
  - Architectural purity!
- Minimal overhead
  - R/w pathway through remote PMA/PMD to remote PCS
  - 48 bit register (maybe use 16 bit with multiple writes)
  - Larger register will ease parallel operation for large concentrator
- Special requirements
  - Atomic operation
  - Auto reset timeout

Proposal

- Add remote_aggregation_register
  - 48 bit (3 x 16 bit access?) or 64 bit
- One per PCS (aggregateable PHY)
  - Accessible through EOC/VOC
- Atomic operation
  - “set if clear”
  - Or else pseudo random back off
- Only register is implemented in PHY, all else in host s/w
  - Algorithm for reference