LL-FC: Problem(s) Identification

Mitch Gusat and Cyriel Minkenberg
IBM ZRL Jan 2007

Goal: Prepare for Monterey Interim

1. Set objectives of an ideal LL-FC

2. Identify the issues re. PAUSE (deviations from 1)
    - agree on terminology

3. Select those problems that we can solve in '07

4. Later: Solution candidates

5. Set the LL-FC agenda for Monterey

Generic Objectives of LL-FC

- I. Correct by design:
  - Doesn't encourage 'simple' and faulty implementations.
  - Doesn't lead to inter-operability faults thru standard misinterpretation by vendor.

- Correctness defines the semantics of
  - (a) lossy-lossless operation,
  - (b) priority level / QoS support,
  - (c) deadlock freedom.

- II. Efficiency options: Enables high performance, Bw- and power-efficient operation. Allows vendor differentiation thru options for
  - ✓ (a) e.g. load balancing/AR, reliable delivery;
  - ✓ (b) options against, e.g., HOL-blocking, hogging, transient HS congestion, persistent HS congestion.
-
- III. 3-way Compatibility:
  - Backwards w/ legacy Ethernet; w/ established IETF protocols (IP, TCP, etc.).
  - Upwards w/ storage and cluster transports RDMA, MPI, iSCSI etc.
  - Forwards w/ new apps that may directly use the native capabilities of LL-FC.

Improved PAUSE: A Hollow Strawman

- 10GigE is a discontinuity in the Ethernet evolution
  - opportunity to address new needs and markets
  - however, improvements are needed
- Requirements of next-generation PAUSE
  1. Correct by design, not implementation
     1. Deadlock-free
     2. No HOL_1- and, possibly reduced HOL_2-blocking

        Note: Do not try to address high-order HOL-blocking at link layer
  2. Configurable for both lossy and lossless operation
  3. QoS / 802.1p support
  4. Enables virtualization / 802.1q
  5. Beneficial or neutral to CM schemes (BCN, TCP, ...)
  6. Legacy PAUSE-compatible
  7. Simple to understand and implement by designers
     1. Min. no. of flow control domains: h/w queues and IDs in Ether-frame
  8. Compelling to use => always enabled...!

Boilerplate: Principles of LL-FC. Orthogonality and 5D Control

- LL-FC shall orthogonalize the following 4 dimensions
  - Correctness
    1. C1: lossy - lossless operation
    2. C2: deadlock prevention and recovery
    3. C3: priority classes / QoS service levels
  - Performance
    4. P1: low-order HOL-blocking, resource hogging, transient congestion
    5. [P2: high-order HOL-blocking, persistent congestion (realm of CM).]

- If the solutions to the above 5 concerns are mutually exclusive (can not coexist in implementation or are not simultaneously operational), any such limitation/constrain will be explicitly stated, including the consequences thereof.

Deadlock Taxonomy:
Deadlocks Possible in a Datacenter Interconnects

- The following types of dlocks may affect an Ethernet system:
  - 1. (DLK1) circular dependency
    - ✓ a) memory-2-memory circular dependency (inter-switch, CD = 1st order deadlock)
    - ✓ b) Load/Store (Rq/Reply) circular dependecy (transaction-induced deadlock)
  - 2. (DLK2) priority blocking (PB=2nd order, improperly called deadlocks);
  - 3. (DLK3) routing loop (RL=3rd order)
- Specific deadlock cases will be illustrated in Monterey

PAUSE Issues

- PAUSE-related issues interfere with BCN simulations
- Correctness
  - ➢ Deadlocks (some of them...)
    - ✓ cycles in the routing graph (if multipath adaptivity is enabled)
      - multiple solutions exist
    - ✓ circular dependencies (in bidir fabrics)
  - ➢ BCN can't help this => Solutions required

- Performance (to be elaborated in a future report)
  - ➢ low-order HOL-blocking and memory hogging
    - ✓ Non-selective PAUSE causes hogging, i.e., monopolization of common resources: e.g. shared memory may be monopolized by frames for a congested port (as shown here)
    - ✓ Consequences
      - best: reduced throughput
      - worst: unfairness, starvation, saturation tree, collapse
    - ✓ properly tuned, BCN can address this problem