LIGHT-WEIGHT PROTOCOLS FOR WIRE-SPEED ORDERING

Hans Eberle, Larry Dennison, SC 2018

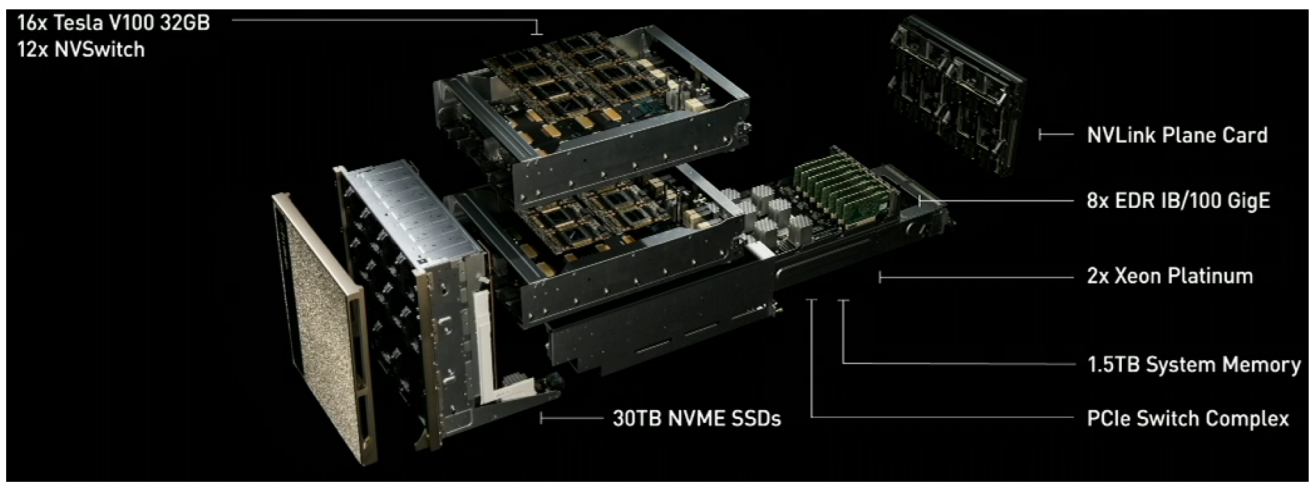
INTRODUCTION

Accelerated Systems are Becoming Pervasive

Summit Supercomputer: 9,216 POWER9 CPUs + 27,648 Tesla V100 GPUs

- Provide selective ordering for out-of-order networks
- Make accelerators first-class networking nodes

NVIDIA DGX-2: 2 Xeon CPUs + 16 Tesla GPUs

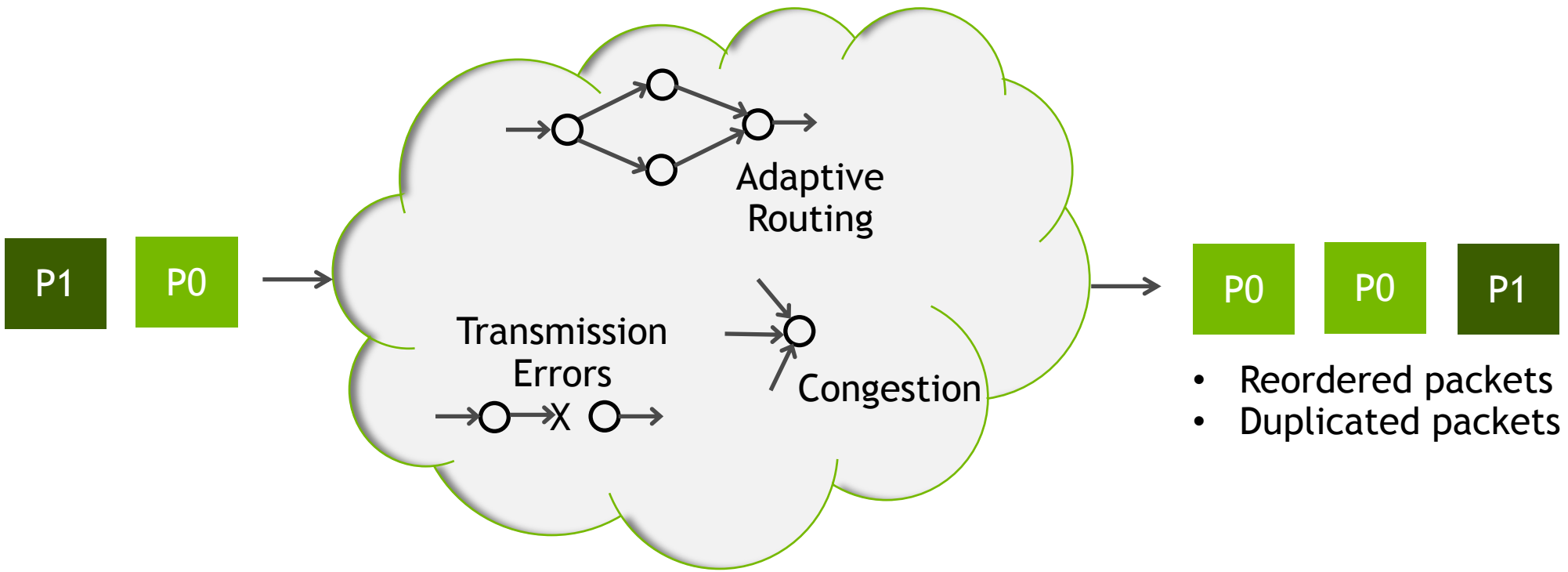


MEMORY INTERCONNECT FOR ACCELERATED SYSTEMS

Properties

- Load/Store semantics
  - Seamless on-/off-chip communication
  - Accelerators have limited NIC resources
- Relaxed memory model
  - No ordering, strict ordering, relaxed ordering
- Advanced networking features
  - Adaptive routing, multipathing
  - Congestion management
  - End-to-end error recovery

OUT-OF-ORDER INTERCONNECTS



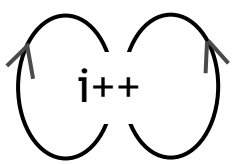
ORDERING SOLUTIONS

Related Work

- Deterministic routing for same-address accesses (CRAY Aries)
  - Doesn’t exploit available path diversity
- Sliding window protocols (TCP)
  - Not optimized for multipathing
  - Long control loops
- One-at-a-time stateful delivery for non-idempotent operations (Gen-Z)
  - Limited throughput

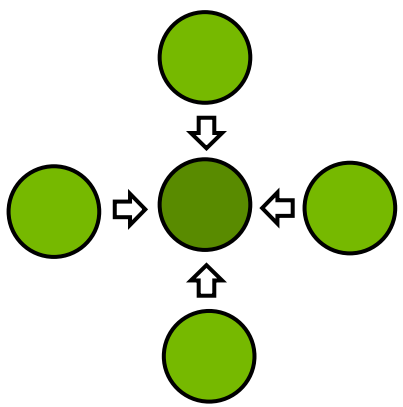
USE CASES

SC-LOC
"Hot Atomics"

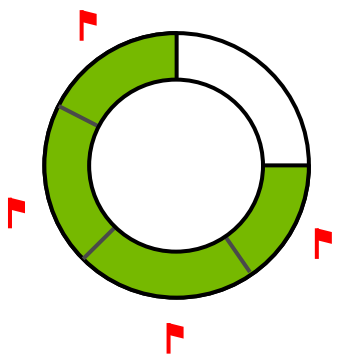


I/O Device Transfer

HALO Exchange

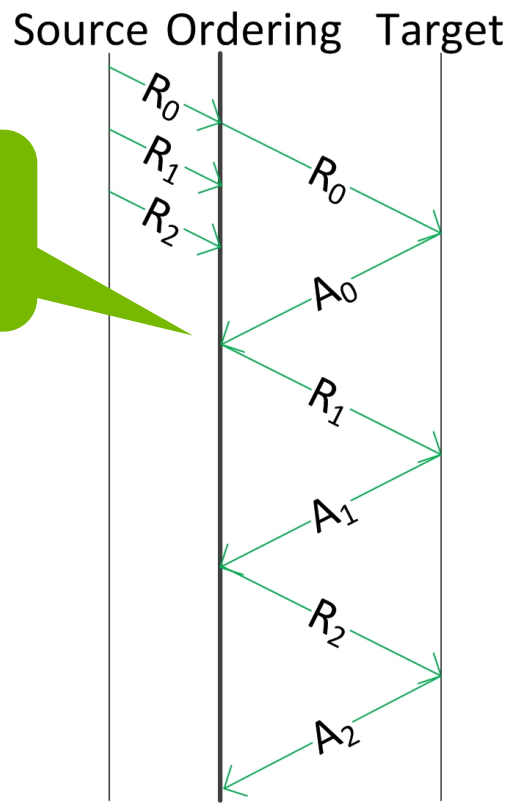


Producer/Consumer Queue



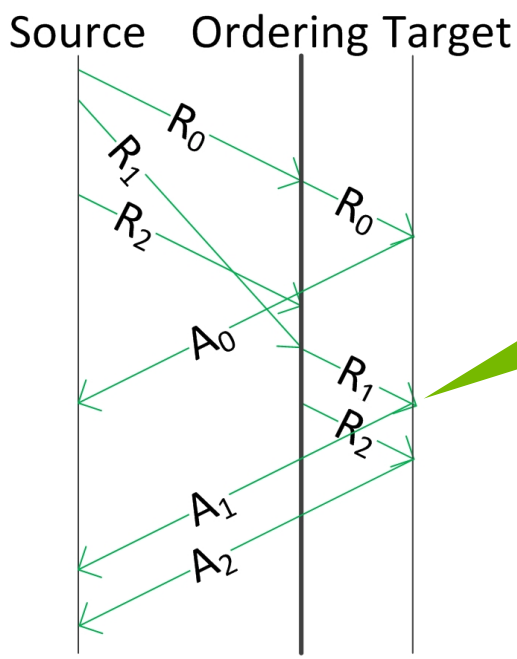
ACCELERATION TECHNIQUES

Source-Side Ordering



REQ/ACK transactions are serialized

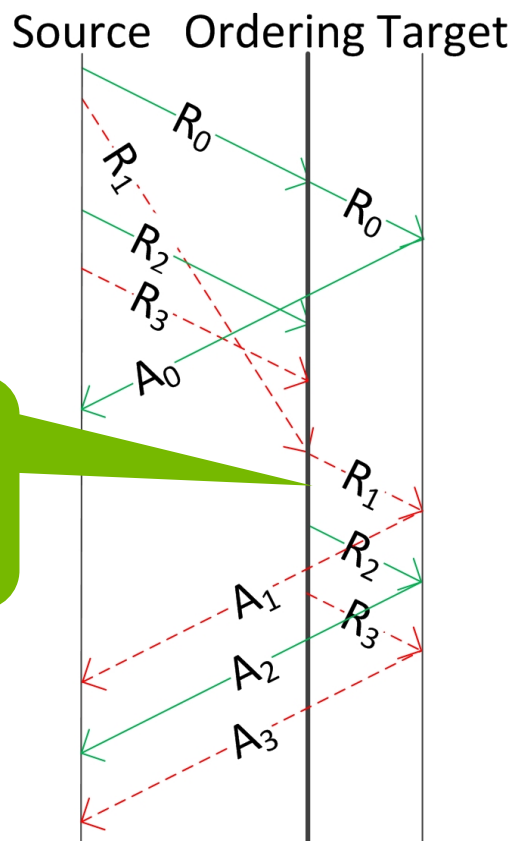
Target-Side Ordering



REQ/ACK transactions are overlapped

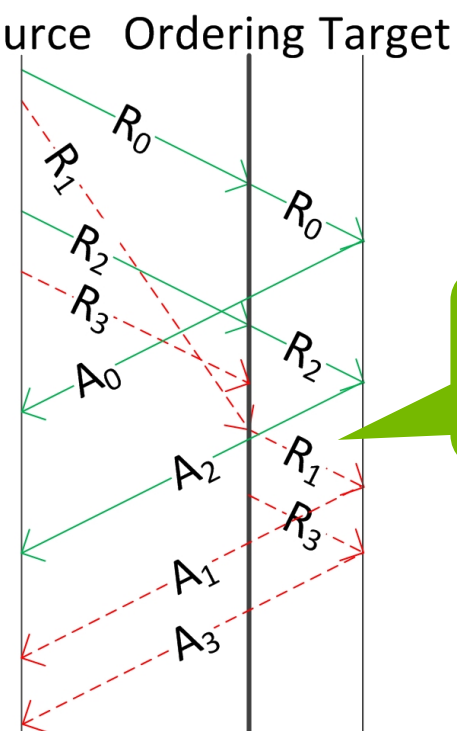
ACCELERATION TECHNIQUES

Coarse-Grained Ordering



Single flow creates unnecessary dependencies

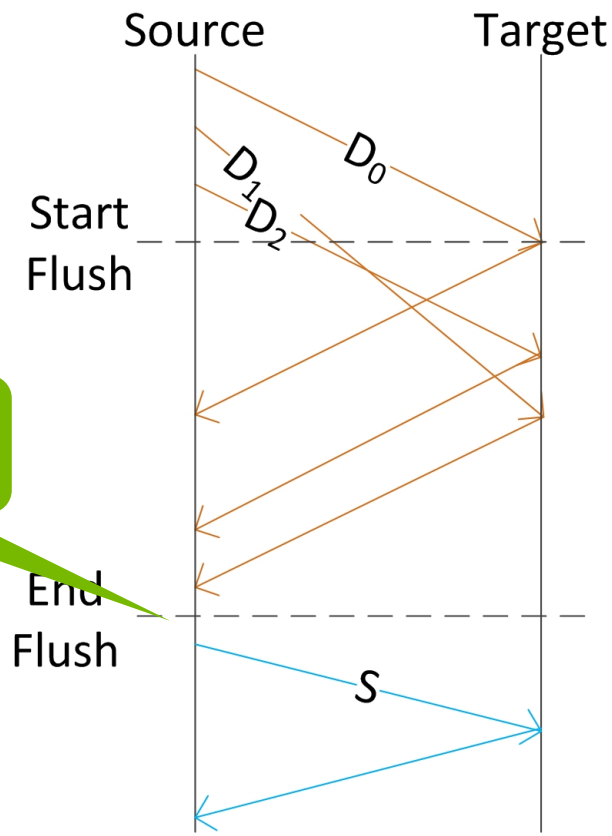
Fine-Grained Ordering



Separate flows decouple dependencies

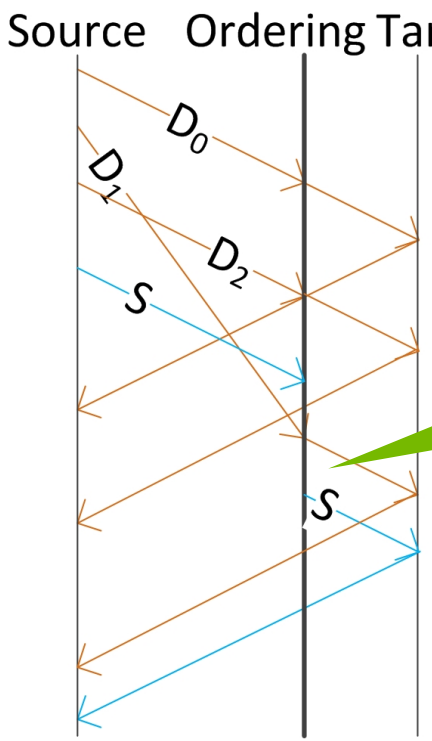
ACCELERATION TECHNIQUES

Source-Side Synchronization



Flush adds RTT

Target-Side Synchronization



S immediately released after last D

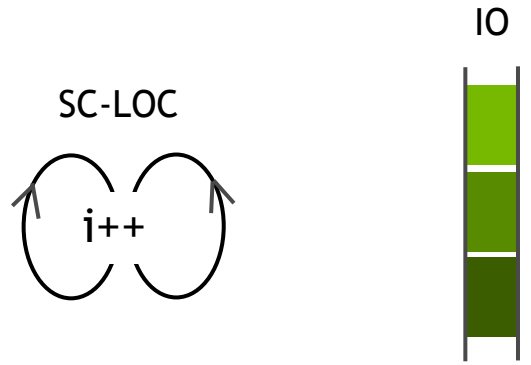
PROTOCOL HIGHLIGHTS

- Slow mode (one-at-a-time transfers) and fast mode (overlapped transfers)
- Light-weight connections with speculative connection setup
- Small reorder buffer (∝ network skew) and small replay buffer (∝ network RTT)
- No timeouts on regular protocol path
- Exactly-once delivery option
- Unreliable transport layer

TWO ORDERING PROTOCOLS

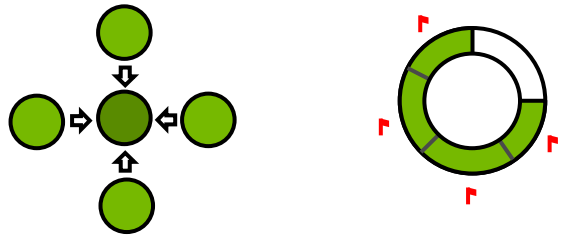
Use Cases Have Different Needs

1. Ordered Transfers
   - Strict sequential ordered packet delivery
   - Exactly-once delivery option

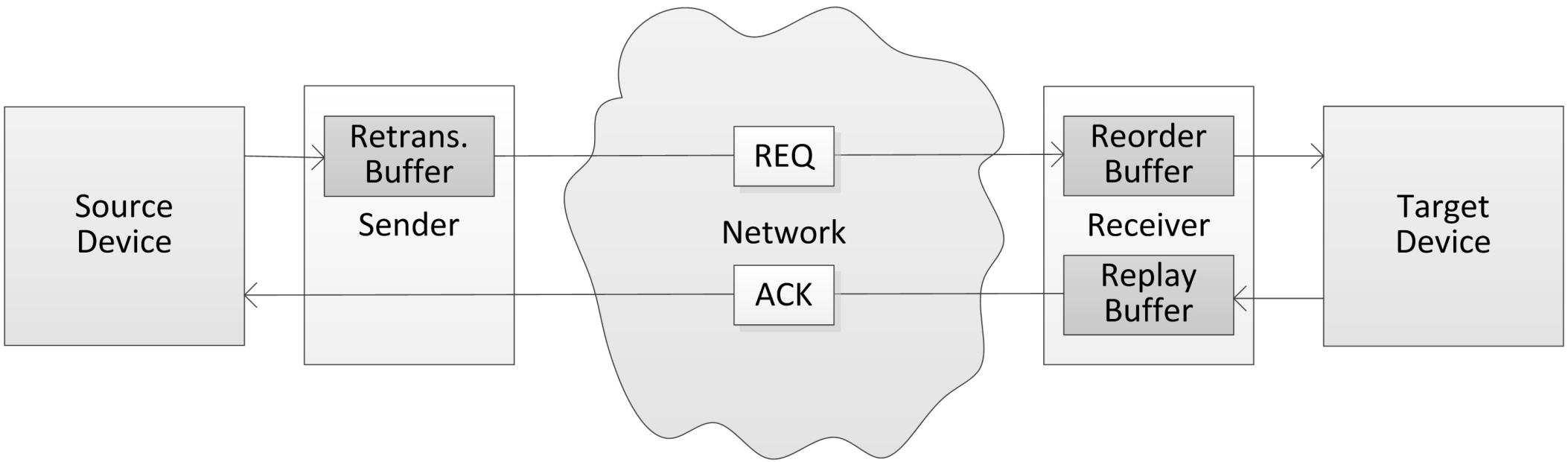


2. Synchronized Transfers
   - Unordered bulk transfer followed by synchronization operation

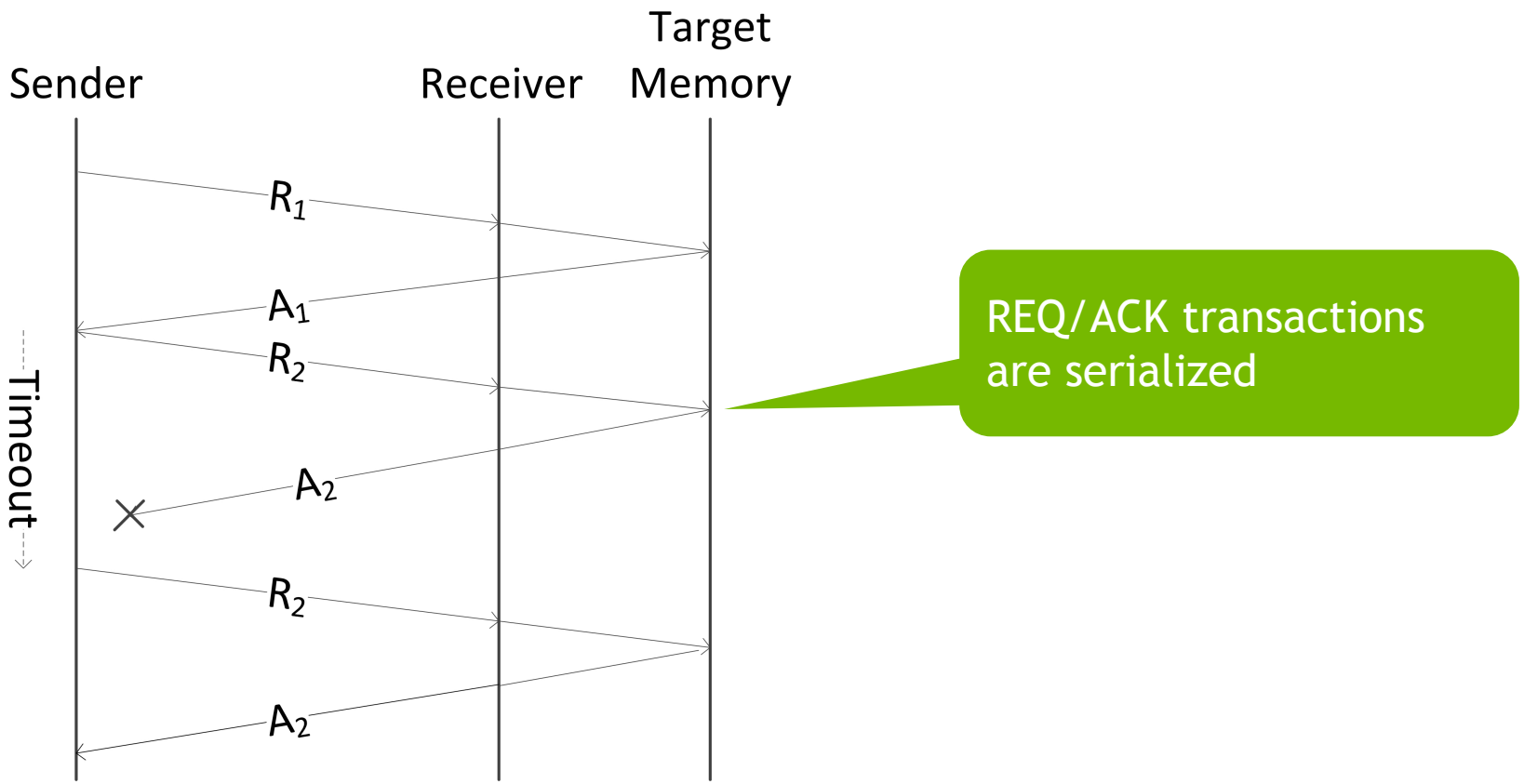
Producer/Consumer Communication



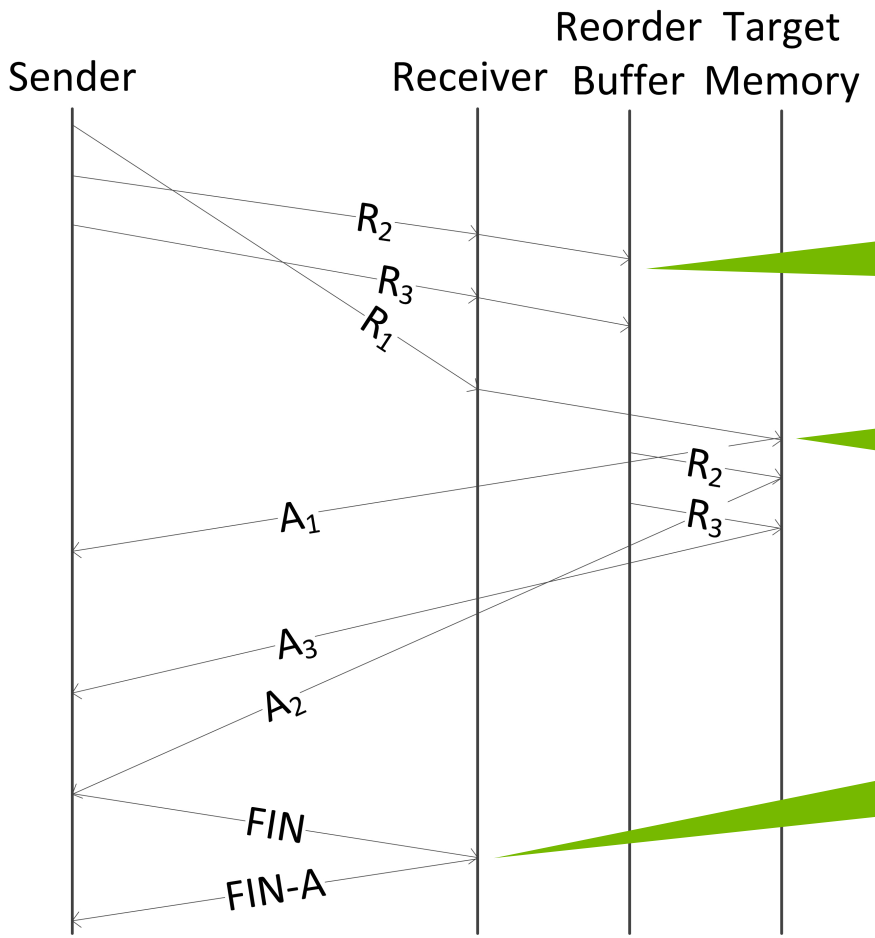
OVERVIEW


Source Device
Retrans. Buffer
Sender
REQ
Network
ACK
Reorder Buffer
Receiver
Replay Buffer
Target Device

ORDERED TRANSFER, SLOW MODE



ORDERED TRANSFER, FAST MODE

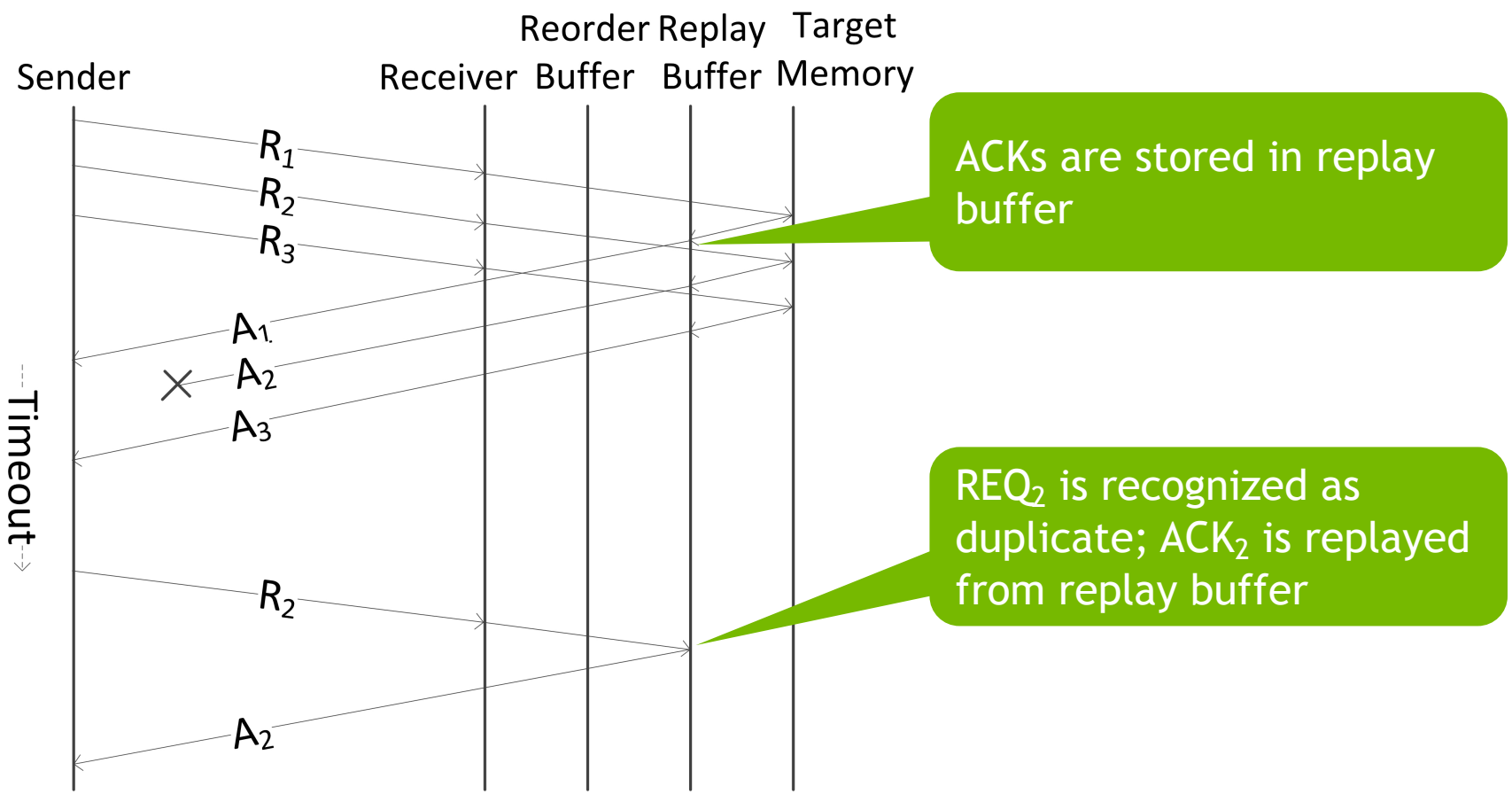


REQ_2 and REQ_3 arrive out-of-order and stored in reorder buffer

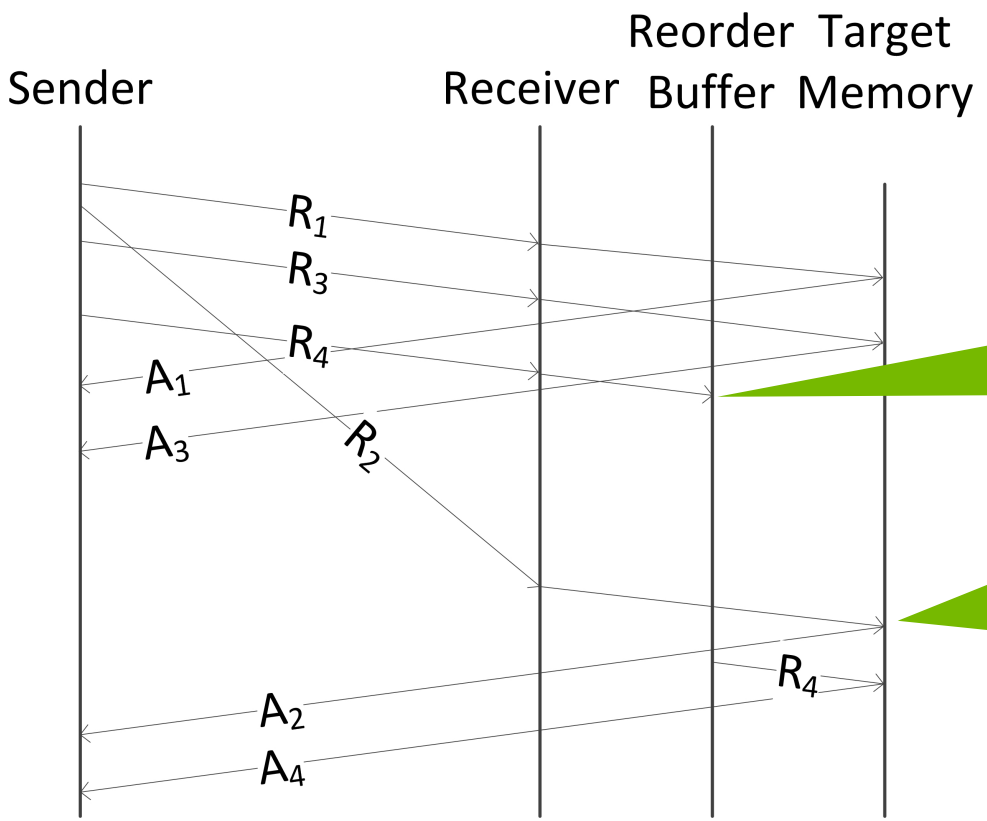
REQ_1 arrives in-order and releases REQ_2 and REQ_3 from reorder buffer

Connection is closed once all ACKs are received

EXACTLY-ONCE DELIVERY



SYNCHRONIZED TRANSFER



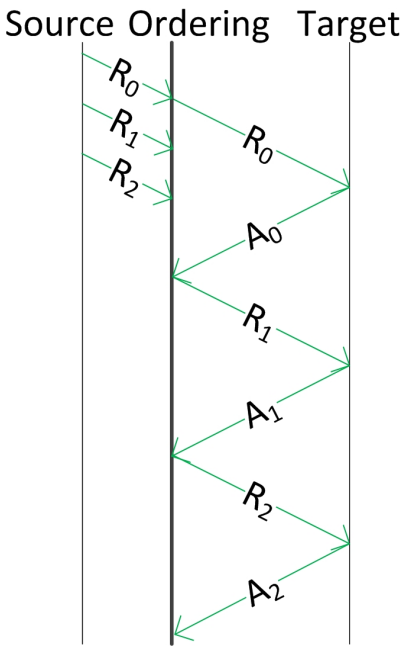
Synchronization operation is held back until all data packets are delivered

Data packet delivery is unordered

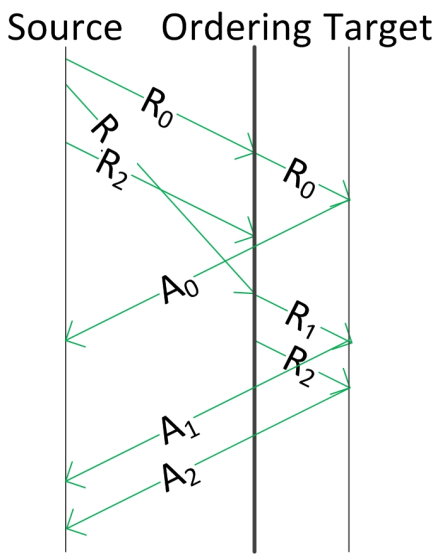
EVALUATION

- Simulated with Bksim cycle-accurate network simulator
- Dragonfly topology
  - 1,056 nodes
  - 33 groups, 8 15-port routers per group
  - 40 ns local channel latency, 500 ns global channel latency
  - 1 flit/ns channel bandwidth
  - PAR6/2 adaptive routing algorithm

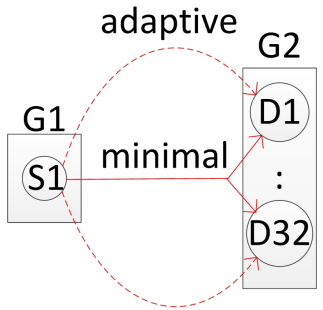
SOURCE-SIDE VS. TARGET-SIDE ORDERING



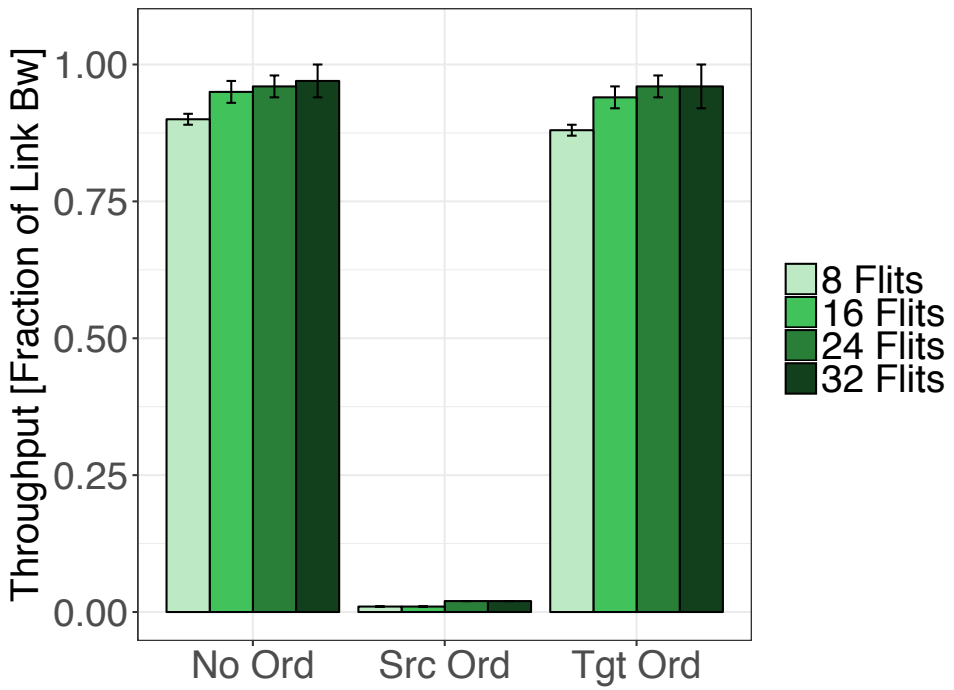
Source-Side Ordering
(Ordered Transfer,
Slow Mode)



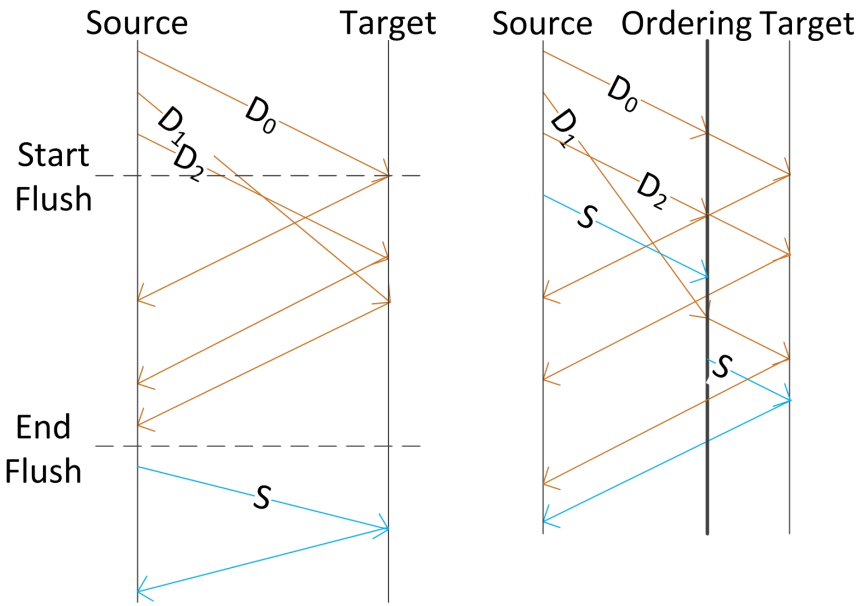
Target-Side Ordering
(Ordered Transfer,
Fast Mode)



10 packets/stream
8..32 flits/packet
1/3 of packets arrive ooo

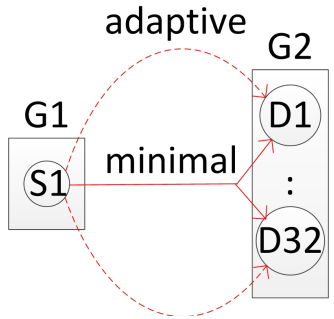


SOURCE-SIDE VS. TARGET-SIDE SYNCHRONIZATION

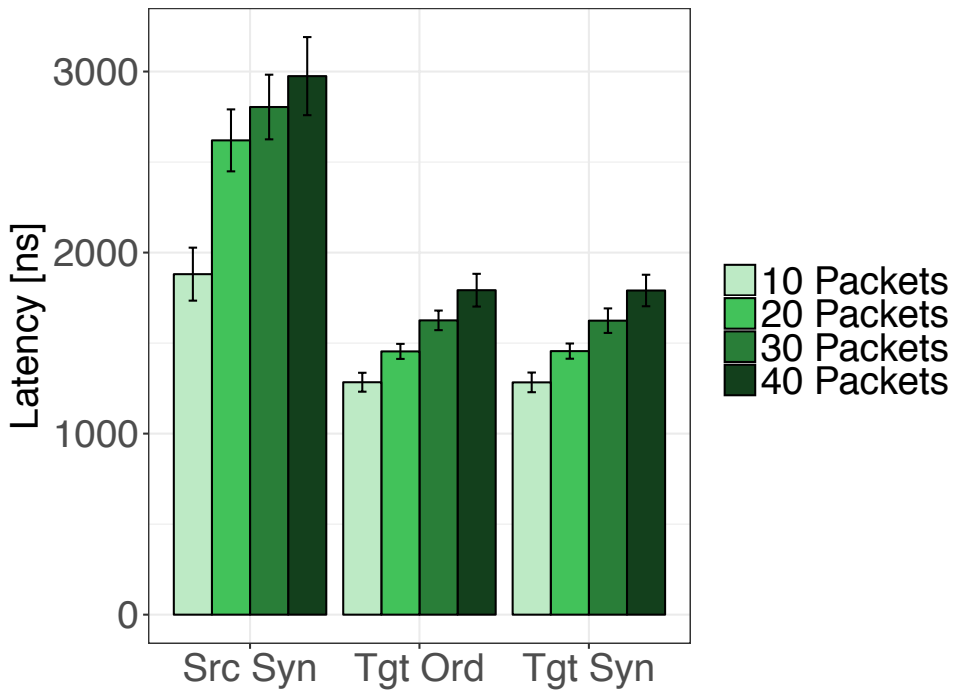


Source-Side Synchr.
(Unordered Transfers
+ Flush + EOD)

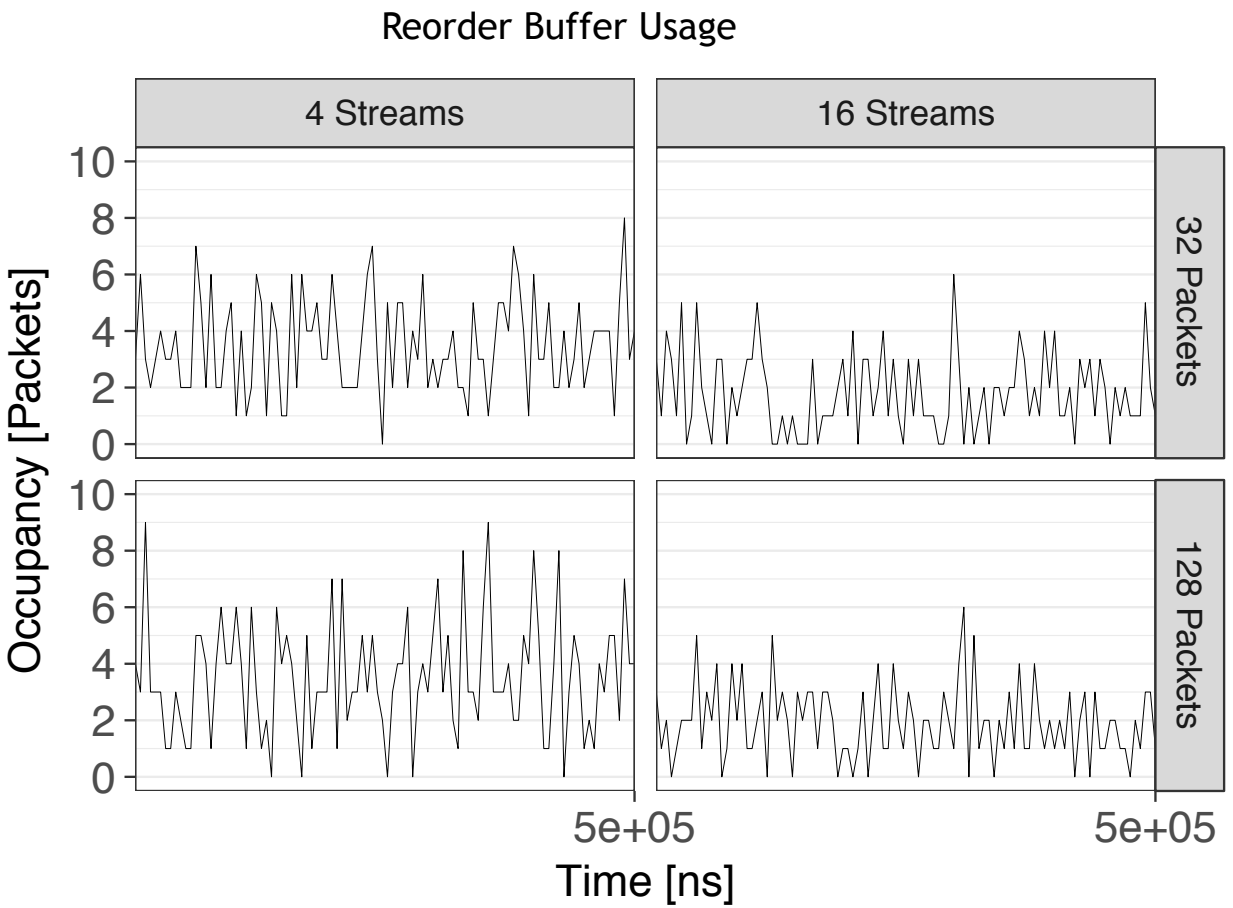
Target-Side Synchr.
(Synchronized
Transfer)

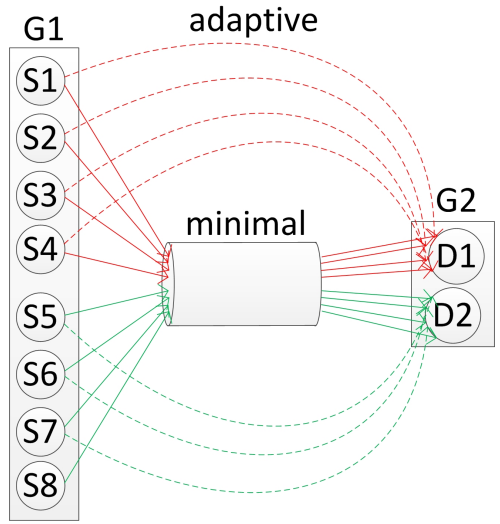


10..40 packets/stream
16 flits/packet
1/3 of packets arrive ooo



REORDER BUFFER USAGE

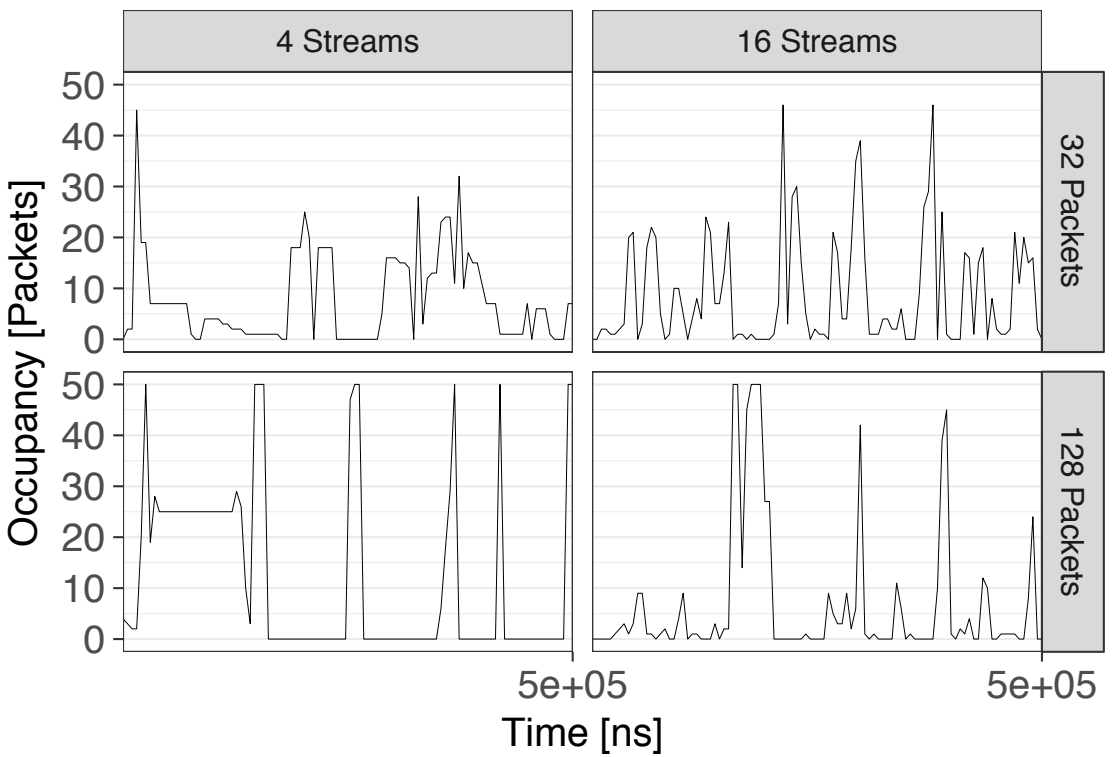




2 x 4/16 sources, 2 targets
Combined load is 0.7 link capacity
Packet distribution is uniform random
No limit on number of connections
32/128 packets/stream
16 flits/packet
50 reorder buffers
1/4 to 1/3 packets arrive ooo

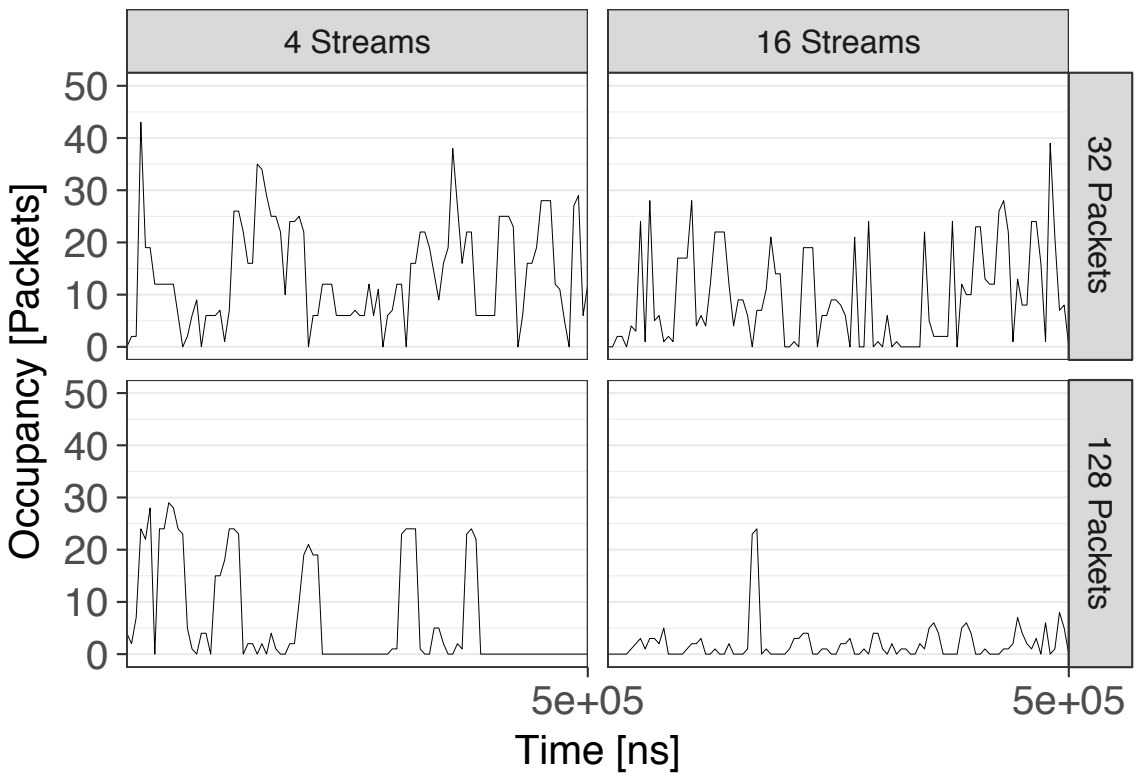
RESOURCE EXHAUSTION

Reorder Buffer Usage w/o Throttling



2 receiver connections

Reorder Buffer Usage with Throttling



Max. 25 outstanding packets per stream

SUMMARY

Light-weight Ordering Protocols for OoO Memory Interconnect

- Protocols extend relaxed memory model of manycore accelerators
- *Ordered transfer protocol* for strict ordering
- *Synchronized transfer protocol* for relaxed ordering
- *Exactly-once delivery* option for non-idempotent operations
- Wire-speed operation
- Modest resource requirements


NVIDIA