R2D2: RAPID AND RELIABLE DATA DELIVERY IN DATA CENTERS

Berk Atikoglu, Mohammad Alizadeh, Jia Shuo Yue, Balaji Prabhakar, Mendel Rosenblum
Department of Electrical Engineering, Stanford University

Introduction

- Two main sources of unreliability on the network
  - **Corruption.** Increasing line rates coupled with constant BERs mean that errors happen more frequently.
  - **Congestion.** Buffer overruns at switches cause large amounts of packet drops; see incast, figure 2.
  - Cause timeouts, increased transfer time
- Existing approaches
  - **High resolution timer (HRT) [1]**
    - Reduce RTO to ~100 µs
    - Expensive to implement in software
  - **Switches with large buffers**
    - Cache all packets, reduce incast occurrence
    - Complex, expensive implementation

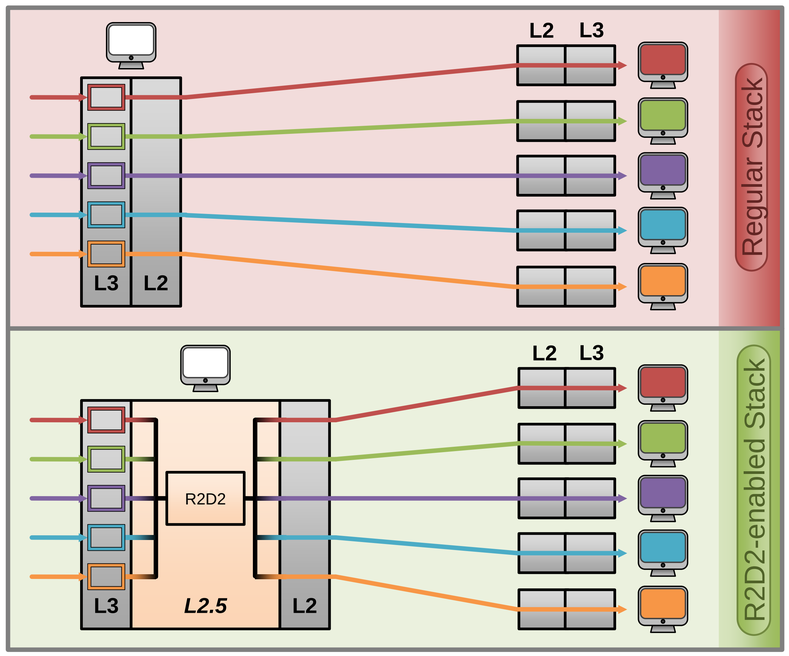

Figure 1. R2D2 and L2.5 in the network stack.

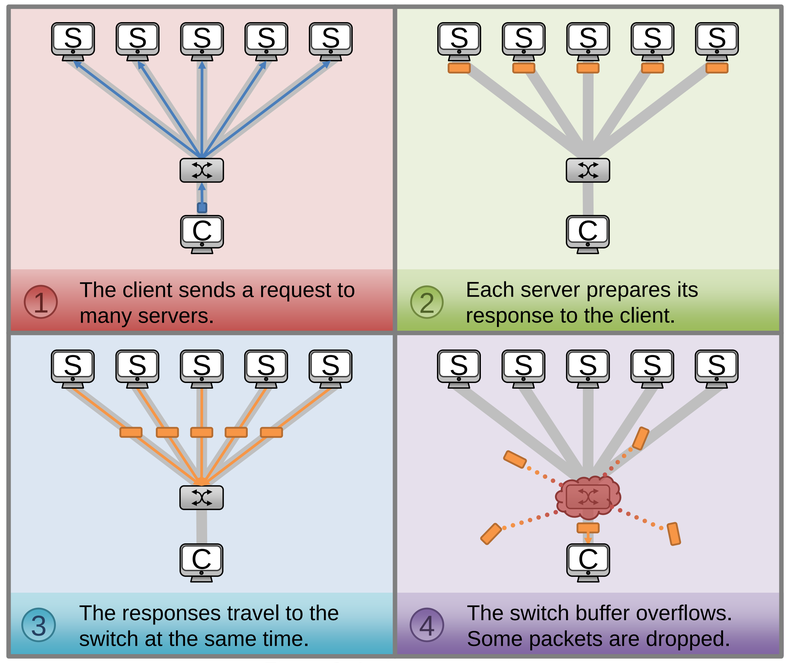

Figure 2. An incast episode.

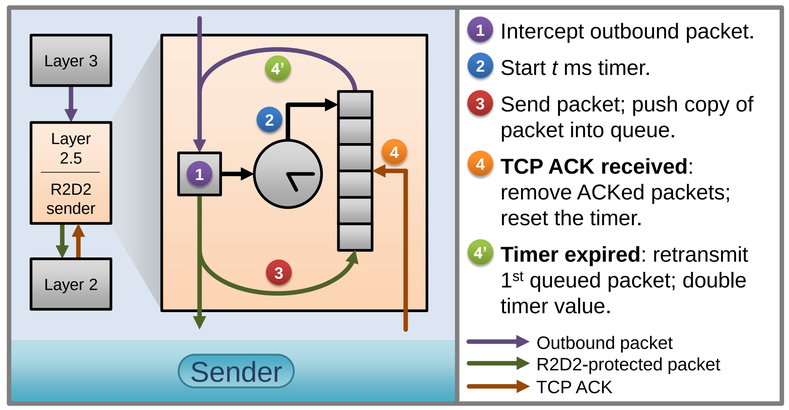

Figure 3. The operation of R2D2.

Testing

- Linux kernel module for versions 2.6.18 up; support for SMP
- Packets are captured and processed with Netfilter hooks
- Requires disabling TCP timestamps and segmentation offloading

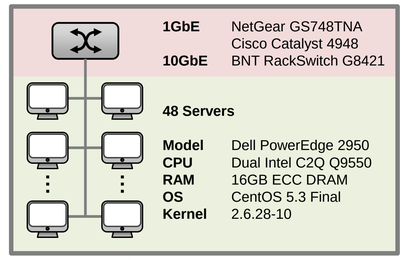

Figure 4. Test setup.

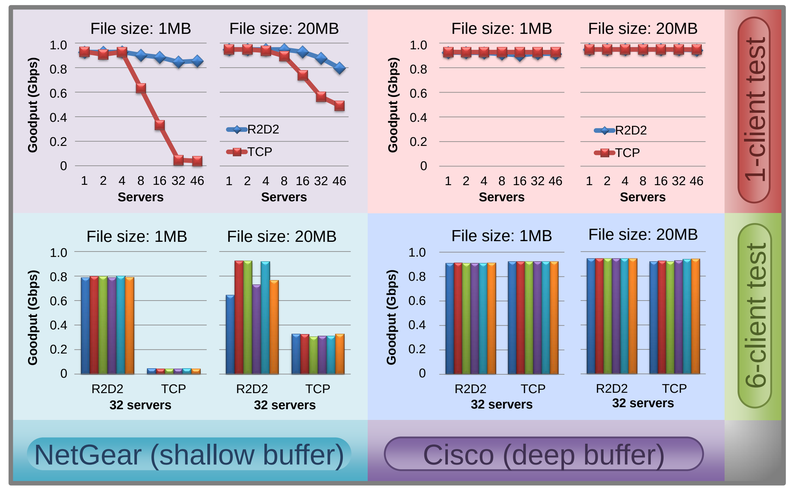

Figure 5. GbE switch goodput test.

- Client requests *b* bytes from *s* servers
- each server responds with *b*/*s* bytes
- Retransmission ratio (RR):

$$\frac{\text{Retransmitted packets}}{\text{Total packets sent by TCP}}$$

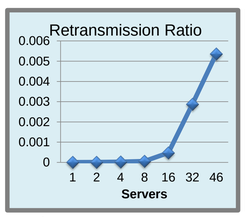

Figure 6. NetGear 1 Client 20MB test (RR).

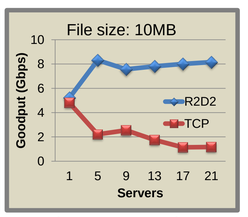

Figure 7. 10GbE switch goodput test.

R2D2: leveraging homogeneity to manage flows

- **Homogeneity in data centers**
  - few hops between hosts (3-5 hops)
  - small, uniform RTTs (100 – 400 µs)
  - high link speeds (1GbE, 10GbE)
- **Features of R2D2**
  - Aggregate all flows into “meta-flow”
  - Single retransmission timer
  - Little per-flow state
  - Reliable (but not guaranteed) delivery
- **Little per-flow state**
  - All flows use the same base timer
  - Each flow maintains one value: number of base timer expirations before retransmission
- **No encapsulation**
  - Does not modify outbound packets
  - Protected packets may safely cross L3 domain boundaries
- **Reliable (but not guaranteed) delivery**
  - Packet is transmitted a maximum number of times, then dropped
  - Increasing timer value also limits the number of packets put onto wire
- **Incremental deployment**
  - White-list approach: can protect individual flows based on IP subnets / TCP port numbers

Conclusion

- R2D2 provides reliable delivery
  - Simple, efficient, cost-effective
  - Kernel module is easily deployed
  - Improves reliability in data center networks
- Next steps
  - test on real data center workloads
  - Deployment in real data centers

[1] V. Vasudevan, A. Phanishayee, H. Shah, E. Krevat, D. G. Andersen, G. Ganger, G. A. Gibson, and B. Mueller, “Safe and Effective Fine-grained TCP Retransmissions for Datacenter Communication,” In Proc. ACM SIGCOMM, 2009.