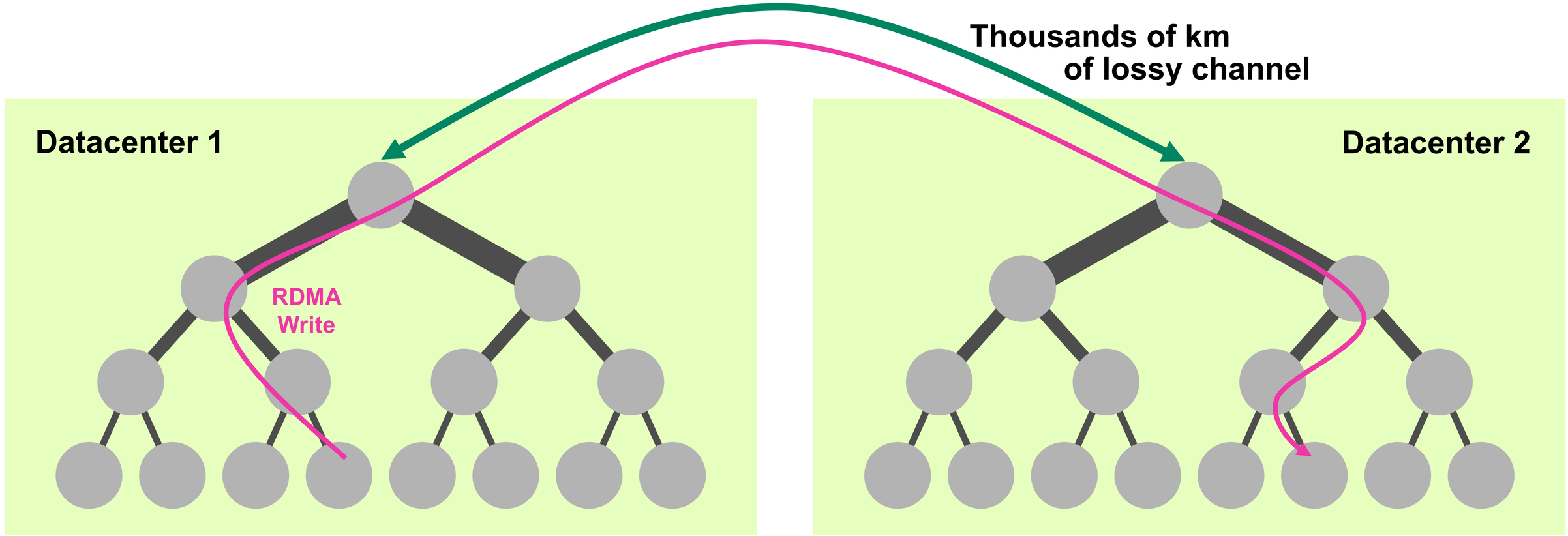
SDR-RDMA: Software-Defined Reliability Architecture for Planetary Scale RDMA Communication

Mikhail Khalilov, Siyuan Shen, Marcin Chrapek, Tiancheng Chen, Kenji Nakano, Peter-Jan Gootzen, Salvatore Di Girolamo, Rami Nudelman, Gil Bloch, Jithin Jose, Abdul Kabbani, Sreevatsa Anantharamu, Jie Zhang, Konstantin Taranov, Zhuolong Yu, Scott Moe, Mahmoud Elhaddad, Nicola Mazzoletti, Torsten Hoefler

Use-case for long-distance RDMA

"*The **biggest issue** we are now having **is not a compute glut, but it's power**…* " — Satya Nadella

"*We had to scale to more compute, and that compute was not available as part of one cluster. **We had to go to multi-cluster training**...*" — from OpenAI's "Pre-Training GPT-4.5"

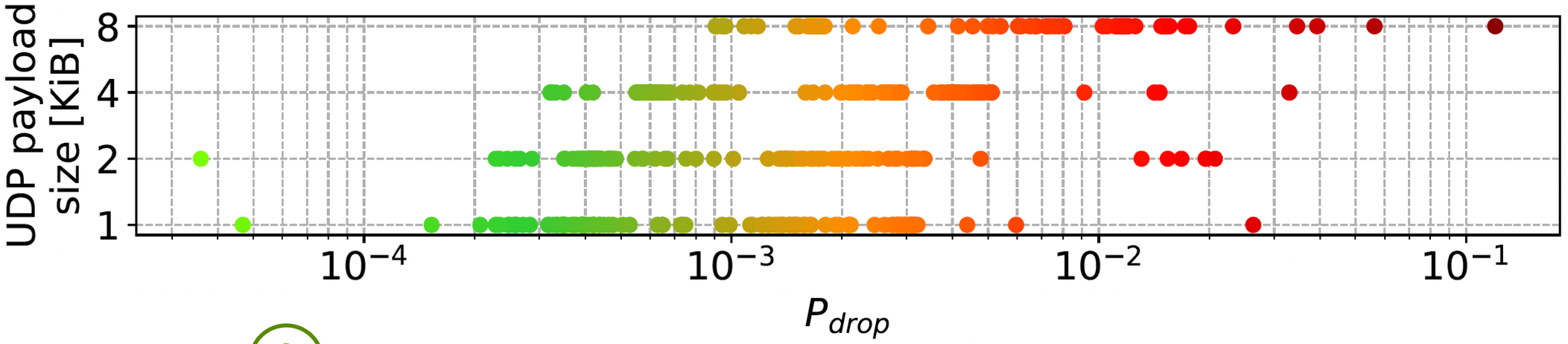


The diversity of long-distance RDMA links

Thousands of kilometers = **Milliseconds** of latency

Large latency variability
Livermore to Oak Ridge **26ms**
Lugano to Jülich **5ms**

Drop probabilities vary considerably based on the distance, used cable, and other traffic.



Private deployments with optimized cables can reach 10^{-8} [1]

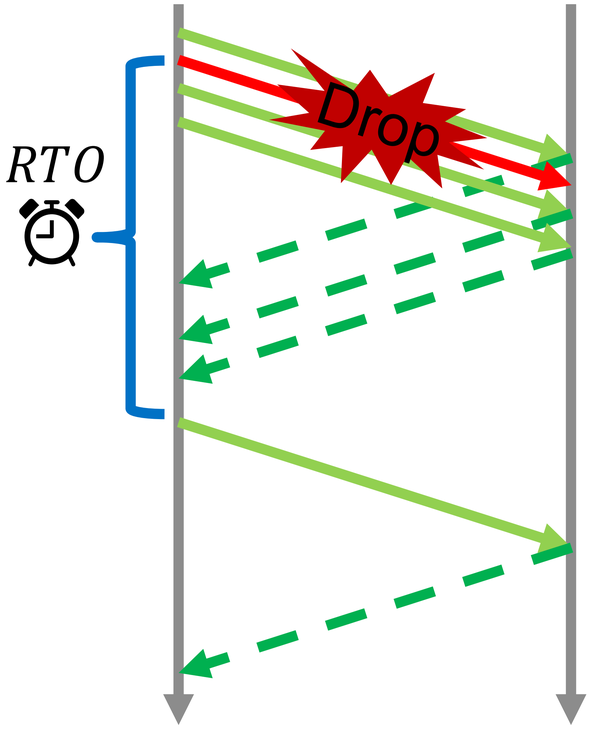
[1] Understanding and mitigating packet corruption in data center networks. Danyang Zhuo, Monia Ghobadi, Ratul Mahajan, Klaus-Tycho Förster, Arvind Krishnamurthy, and Thomas Anderson. 2017 SIGCOMM

Reliability challenge

Selective Repeat (SR)

Default in modern NICs

Sender **Receiver**

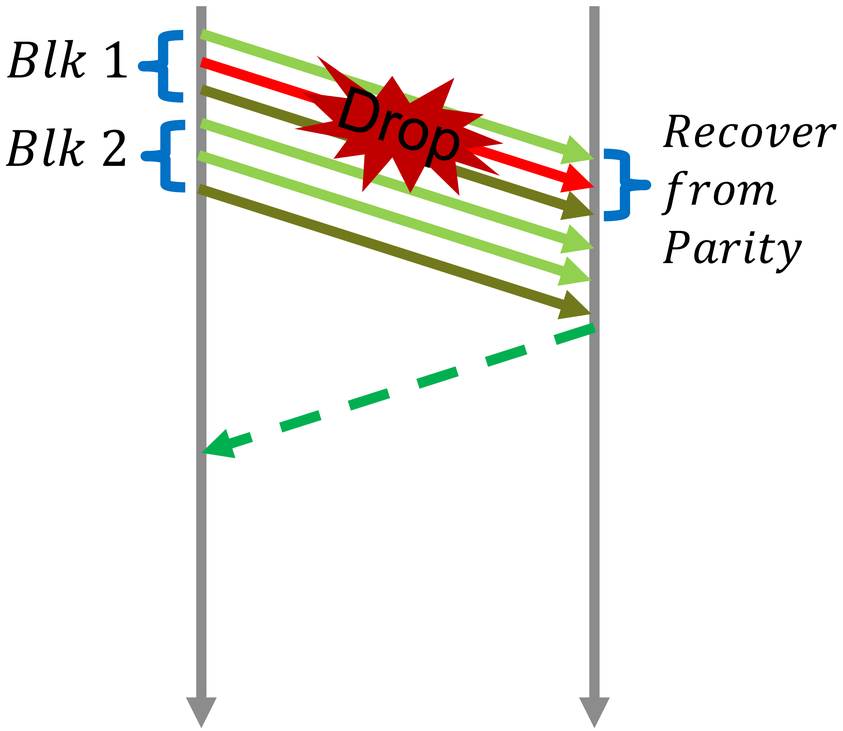


≥1 RTT to selectively repeat
Millisecond RTTs!

Erasure Coding (EC)

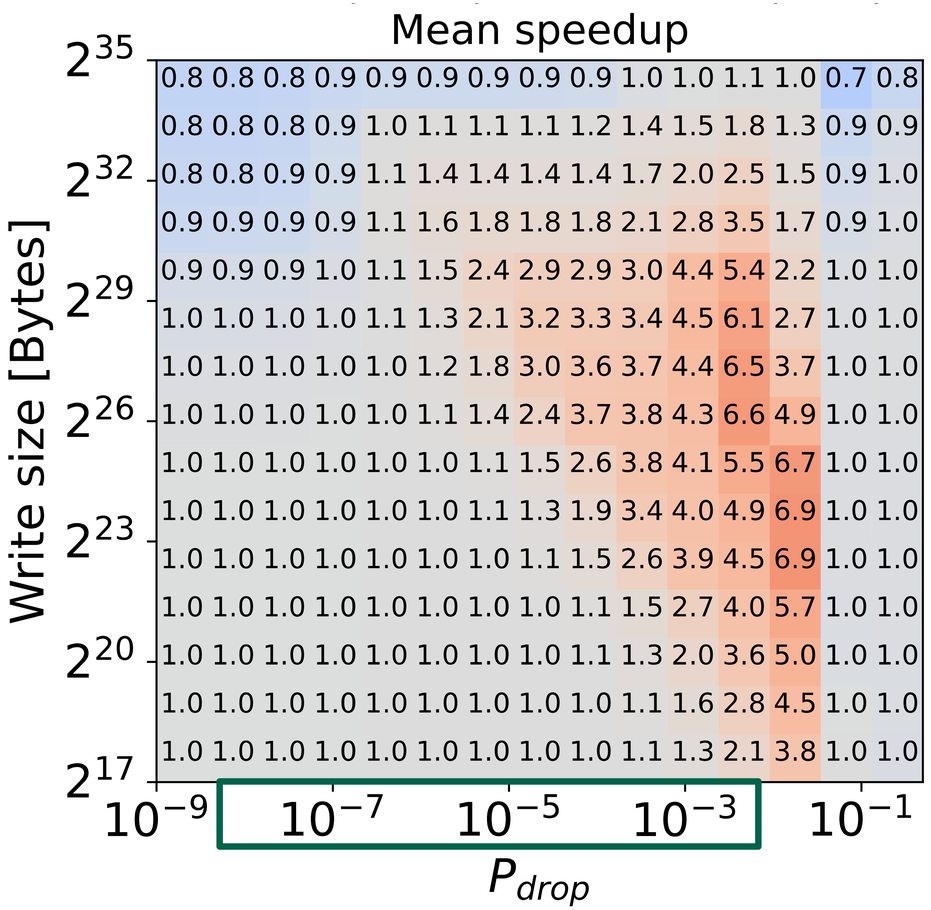
Not available in hardware

Sender **Receiver**



Avoids timeouts

! Speedup of EC over SR up to 6.9

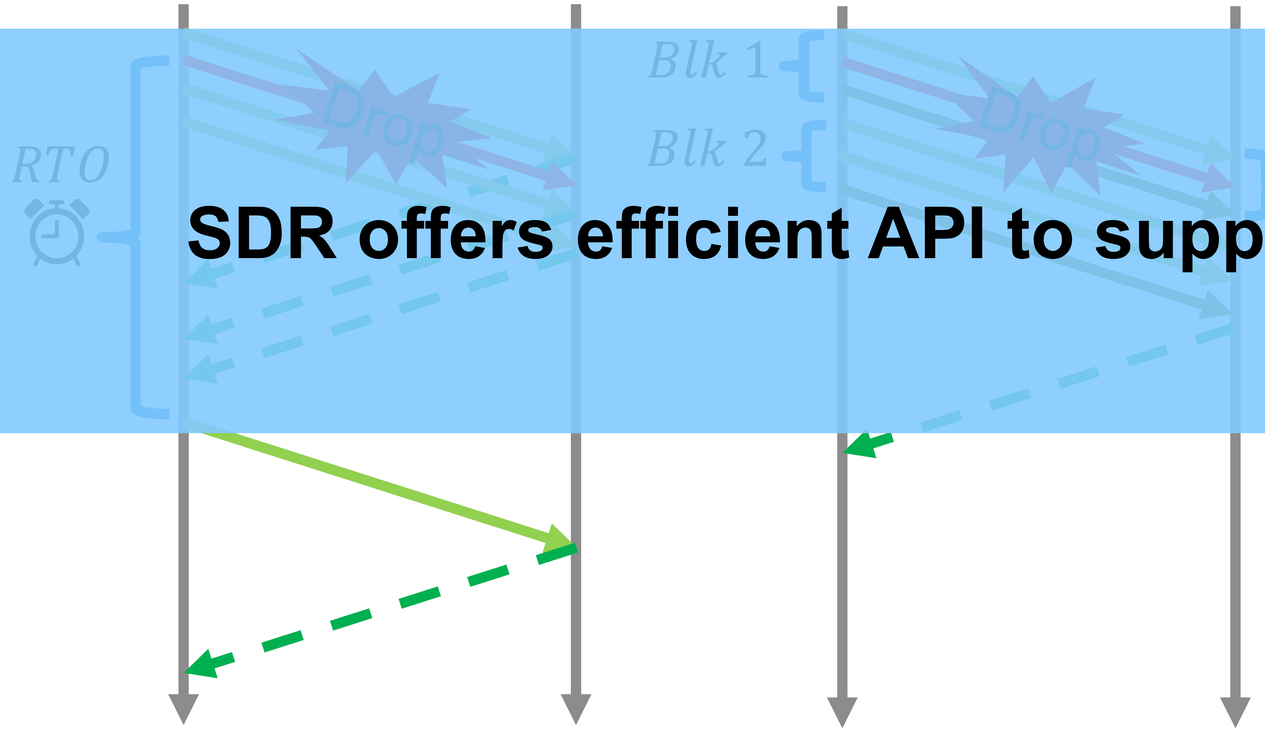


Reliability challenge

Selective Repeat (SR)

Default in modern NICs

Sender **Receiver**



≥1 RTT to selectively repeat
Millisecond RTTs!

Erasure Coding (EC)

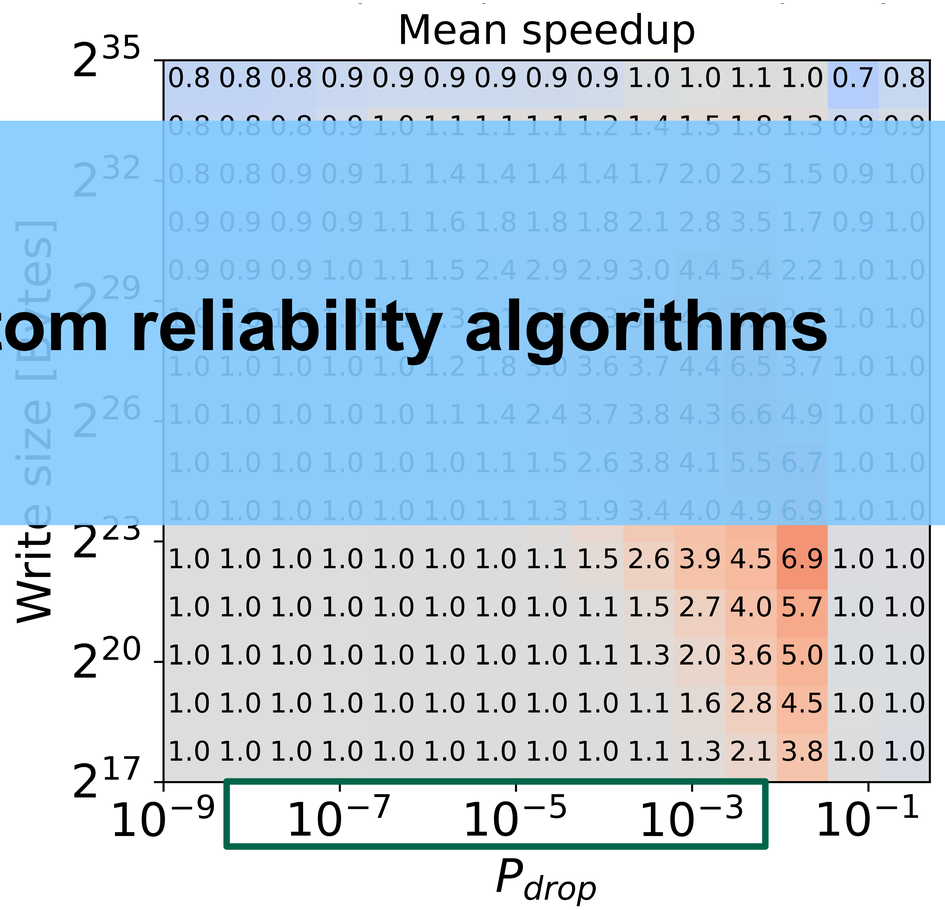
Not available in hardware

Sender **Receiver**

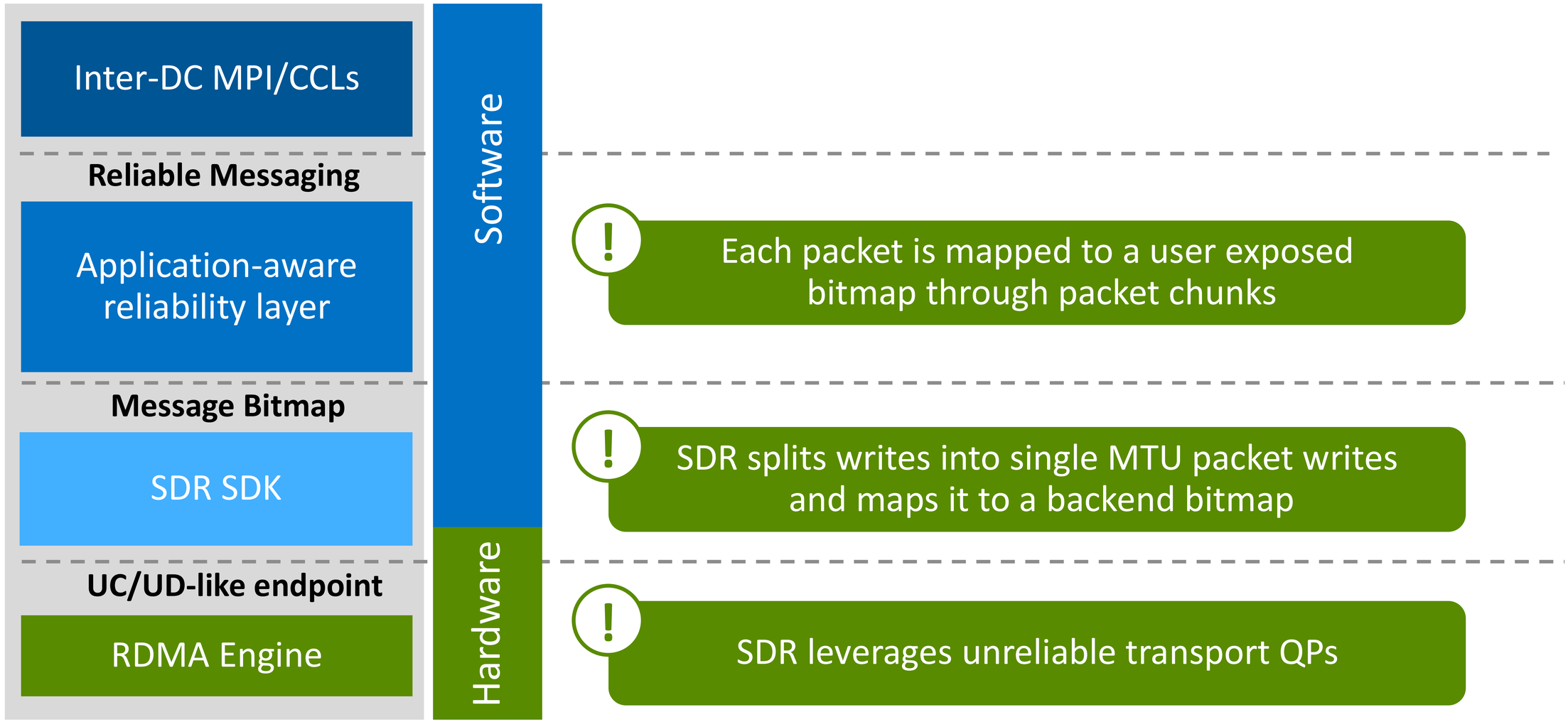
Avoids timeouts

SDR offers efficient API to support custom reliability algorithms

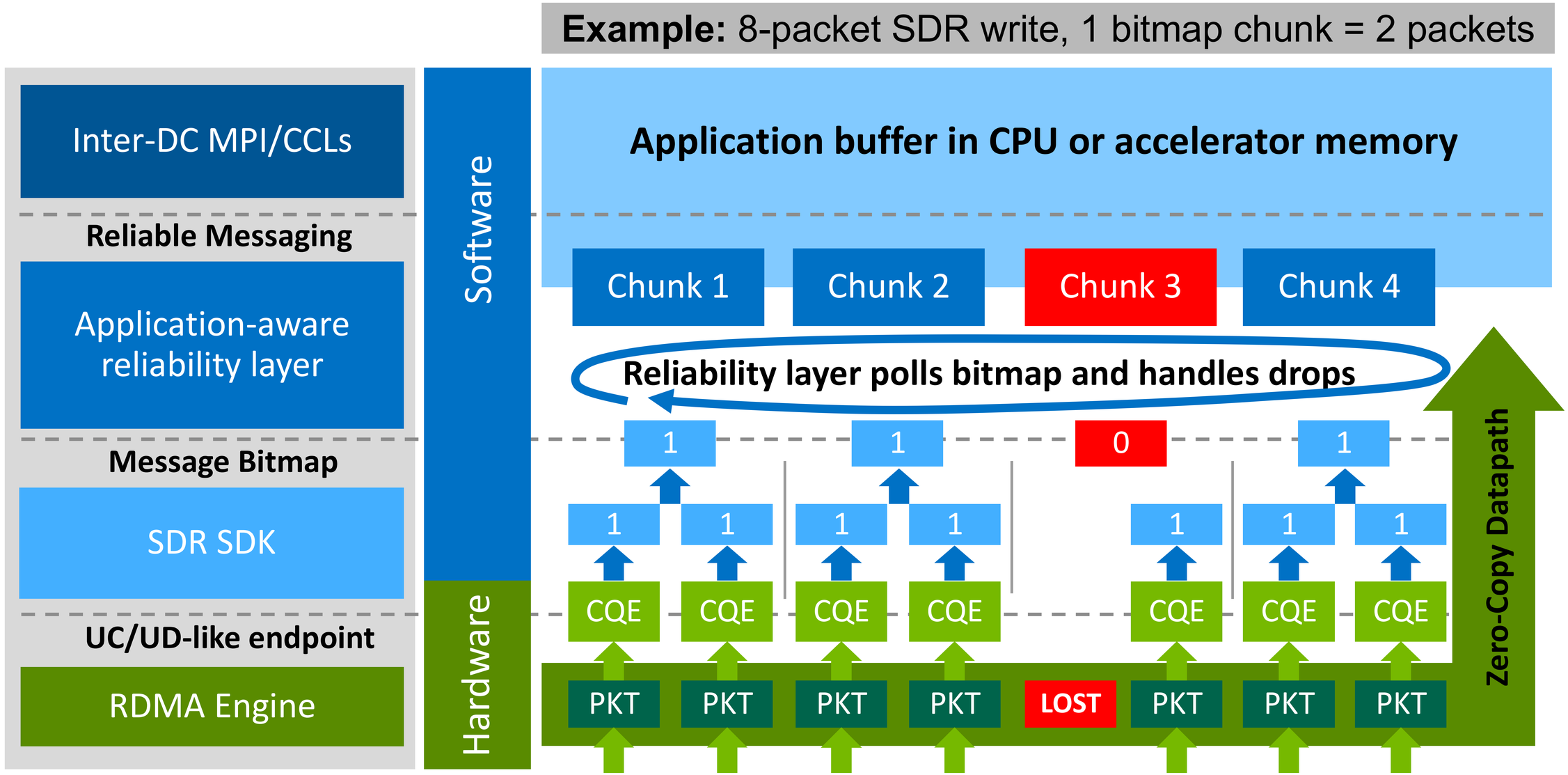
! Speedup of EC over SR up to 6.9



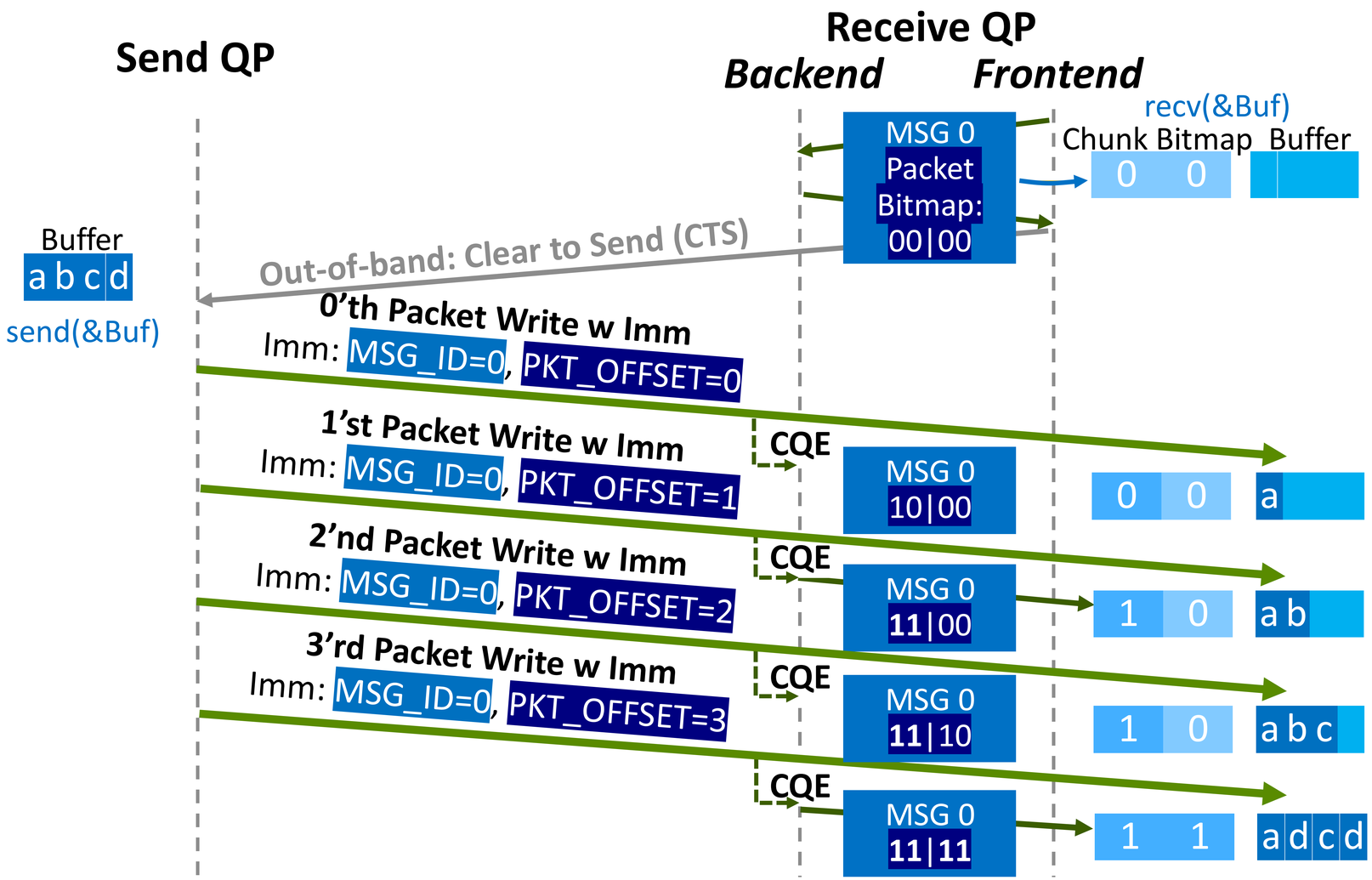
Software-Defined Reliability (SDR)

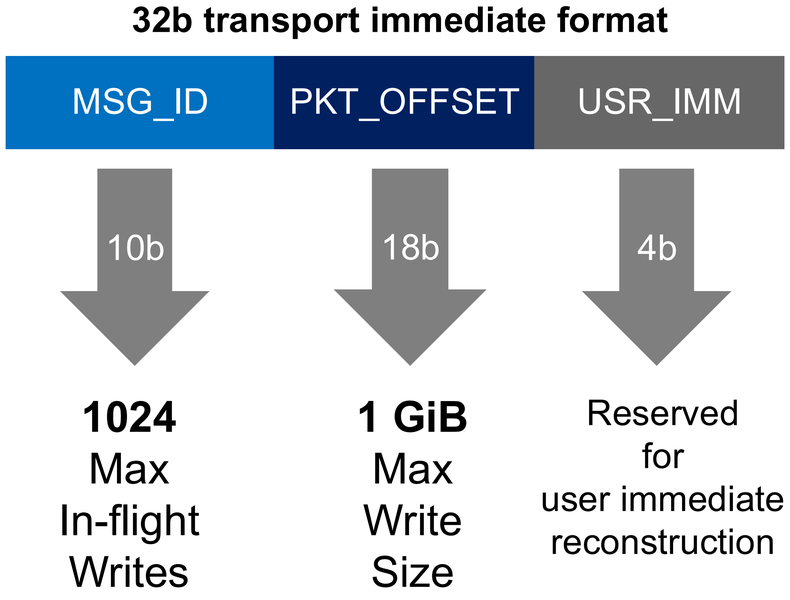


Software-Defined Reliability



SDR protocol design

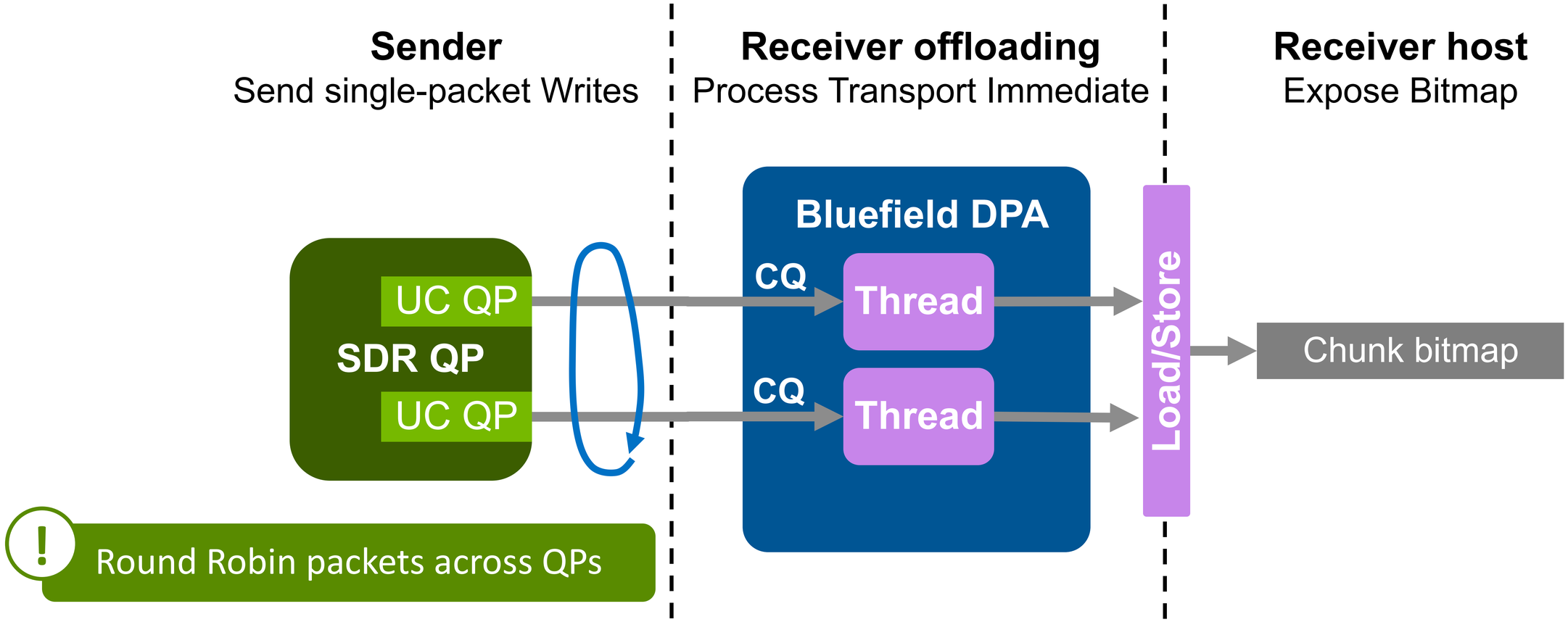




Offloaded implementation

Splitting messages into MTU writes and bitmap processing are expensive on the CPU

We leverage NVIDIA BF3 Datapath Accelerator for offloading

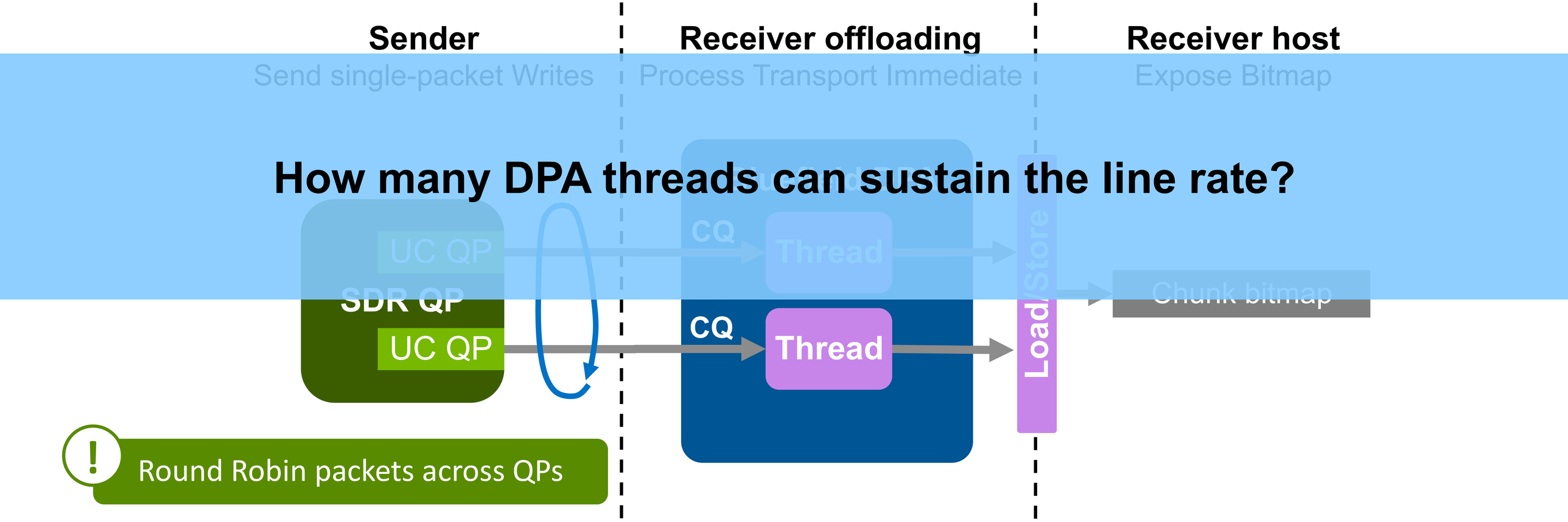


Round Robin packets across QPs

Offloaded implementation

! Splitting messages into MTU writes and bitmap processing are expensive on the CPU

! We leverage NVIDIA BF3 Datapath Accelerator for offloading

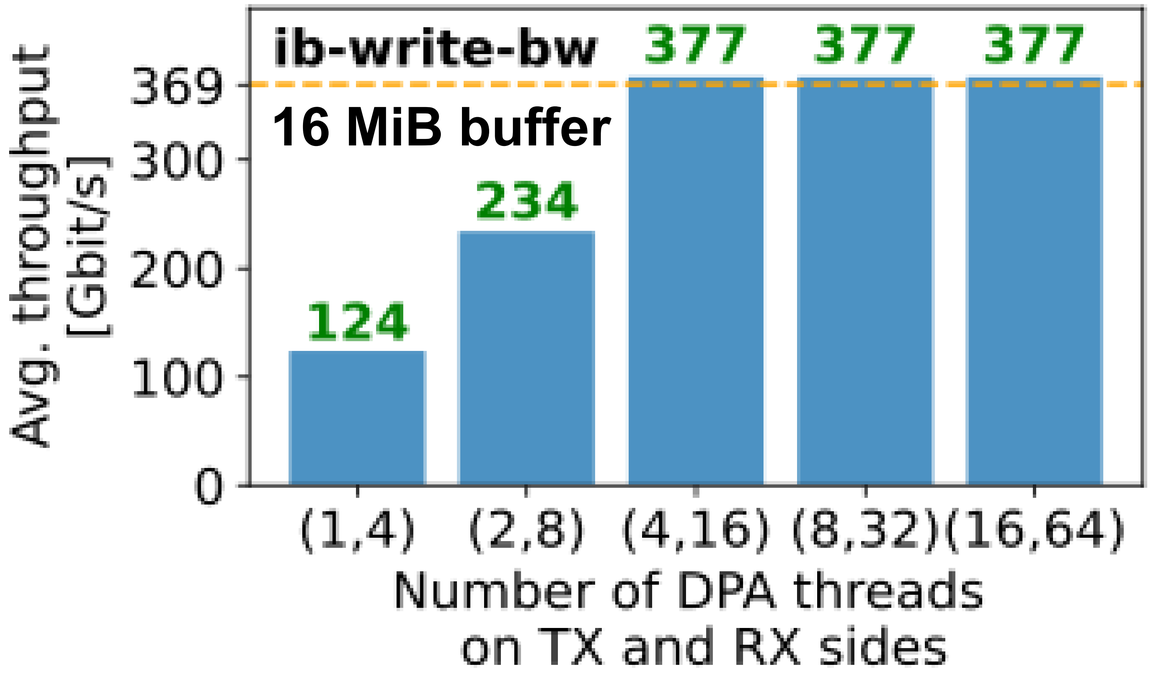


! Round Robin packets across QPs

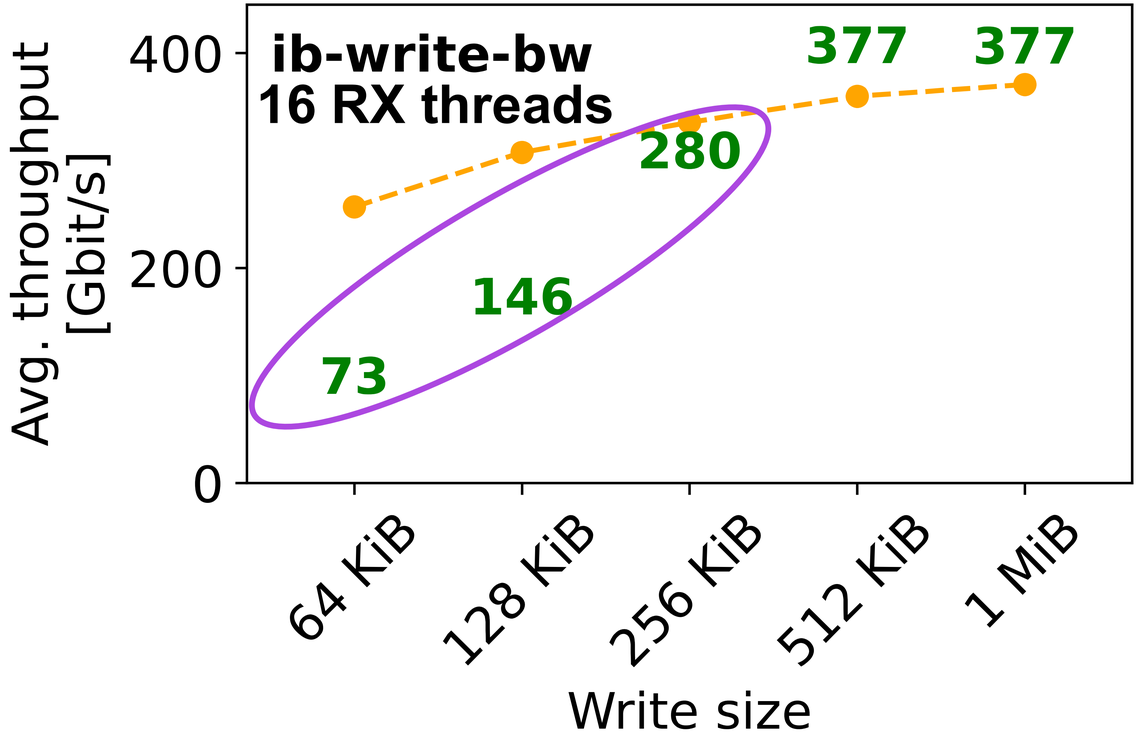
Offloading performance

Israel-1 supercomputer with 400Gbit BF3 NICs

! Line rate achieved with just 20/256 threads



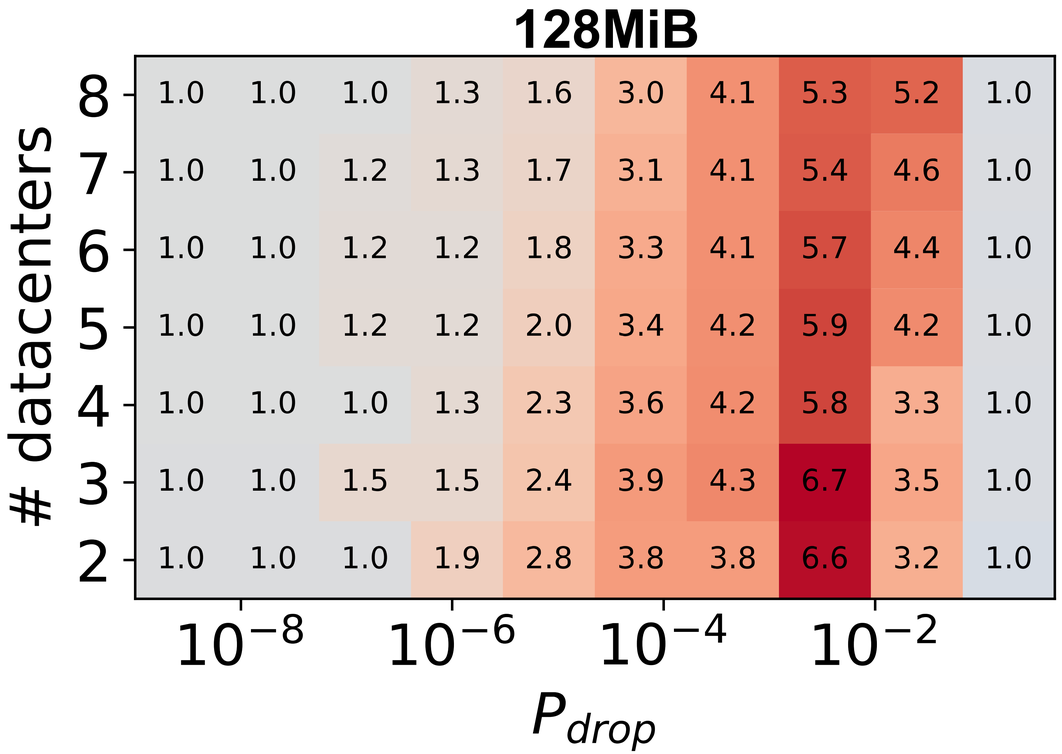
! Receive repost overhead



Impact of reliability on ring Allreduce

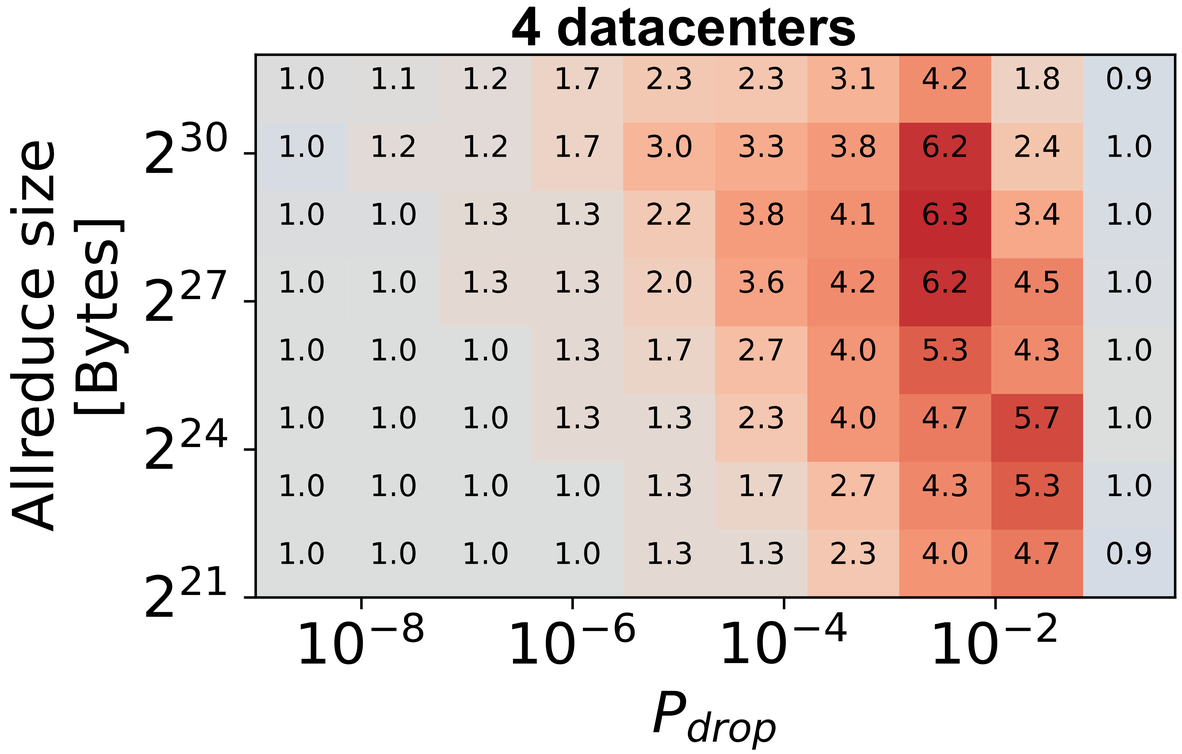
! Theoretical model to quantify SDR-based EC

! Inefficiencies stack over the stages



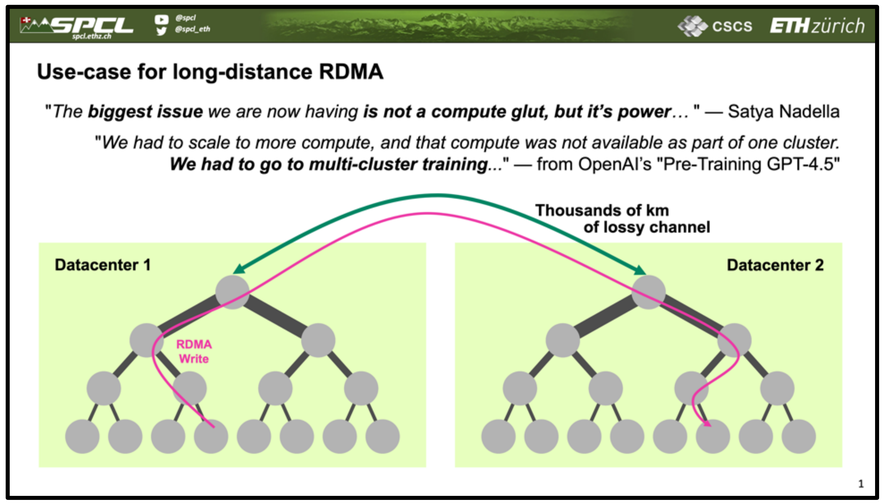
+ Reliability is lower bounded by per-stage cost times the number of stages.

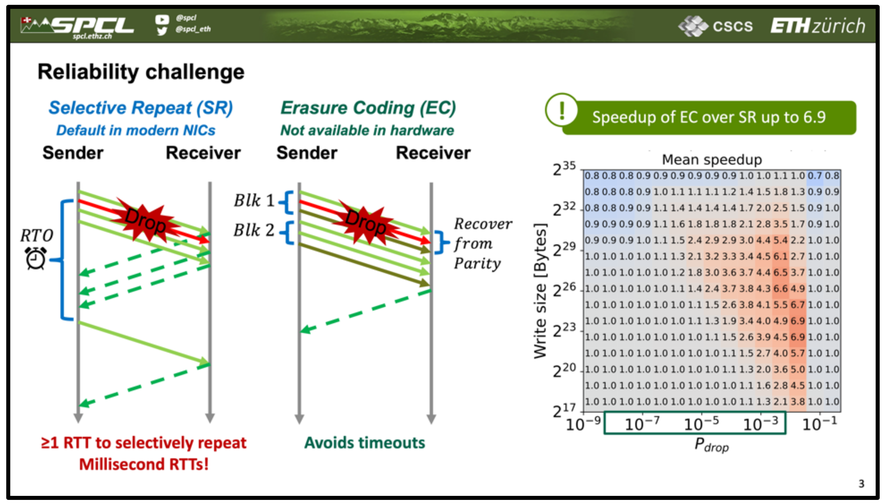
! Overheads like in the P2P case

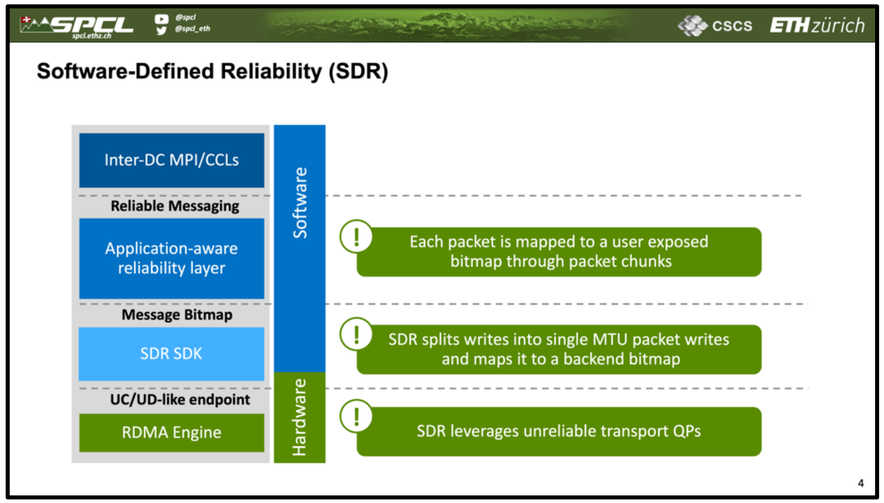


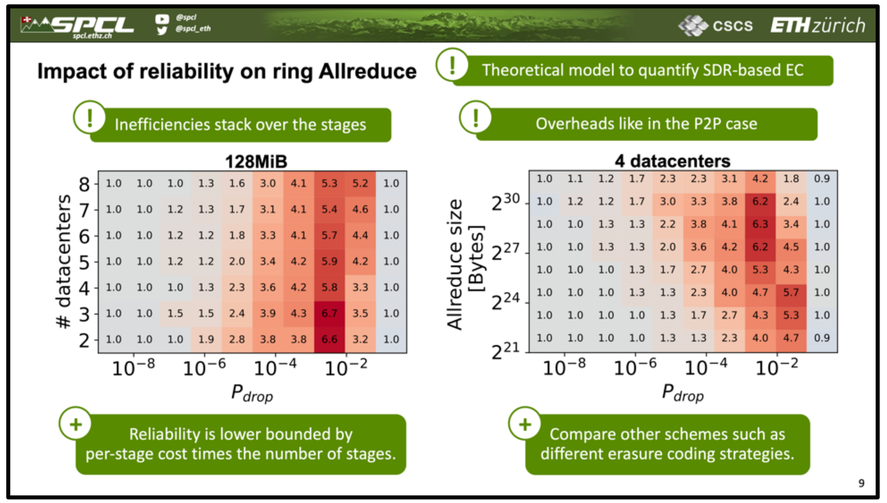
+ Compare other schemes such as different erasure coding strategies.

Conclusions









More of SPCL's research:

youtube.com/@spcl — 180+ Talks

twitter.com/spcl_eth — 1.4K+ Followers

github.com/spcl — 3.8K+ Stars

… or spcl.ethz.ch

Artifacts Available · Artifacts Evaluated Functional · Results Reproduced

github.com/spcl/sdr-rdma-artifact