SAND2023-07608C

Sandia National Laboratories

Exceptional service in the national interest

MODELING AND BENCHMARKING THE POTENTIAL BENEFIT OF EARLY-BIRD TRANSMISSION IN FINE-GRAINED COMMUNICATION

Whit Schonbein, Scott Levy, Matthew G. F. Dosanjh, W. Pepper Marts, Elizabeth Reid, Ryan E. Grant

2023 August, ICPP, Salt Lake City, Utah, United States

Sandia National Laboratories

Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

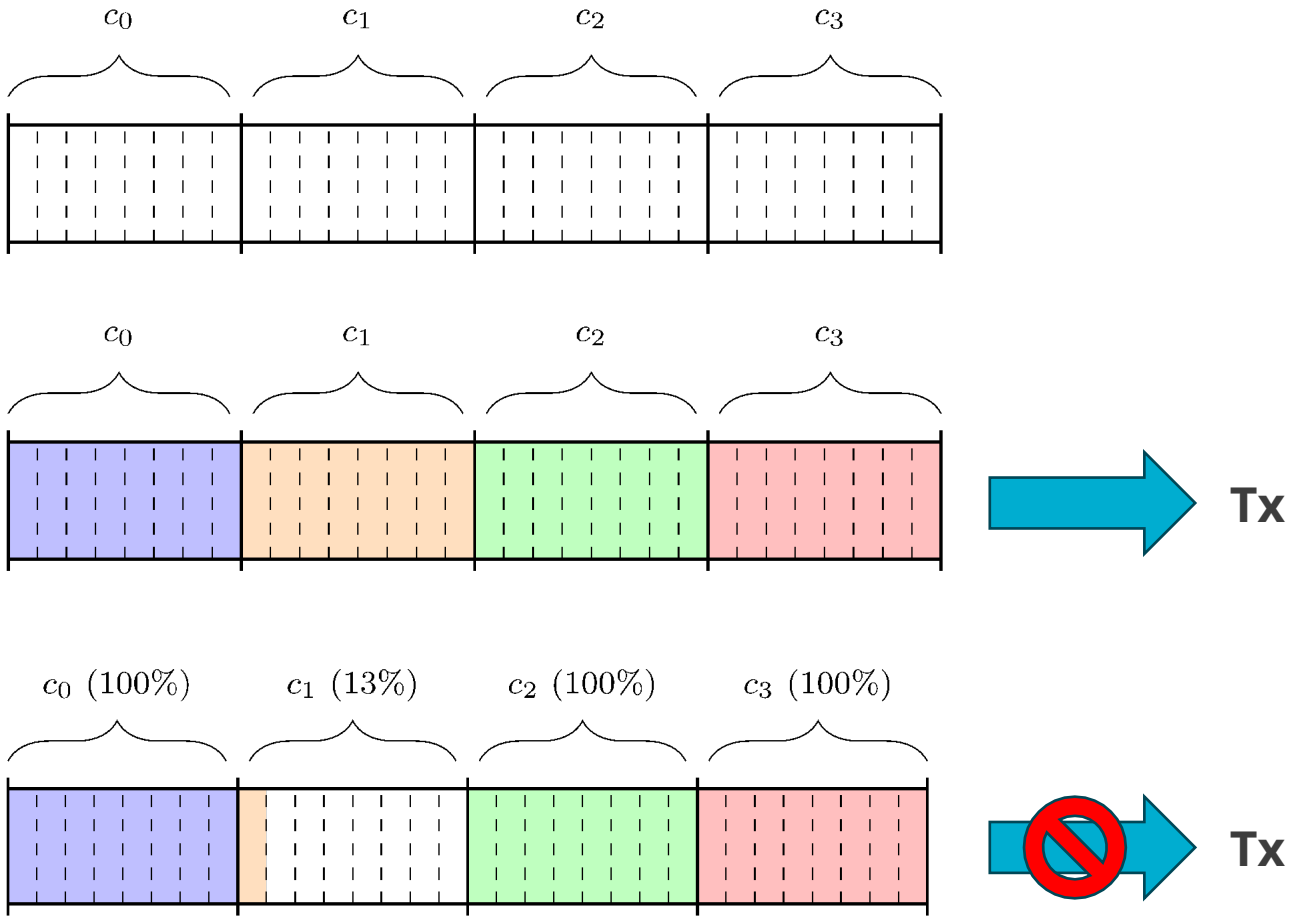
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia LLC, a wholly artment of Energy's National Nuclear Security Administration under contract DE-NA0003525.

INTRODUCTION

In HPC applications, buffers to be communicated often have multiple contributors.

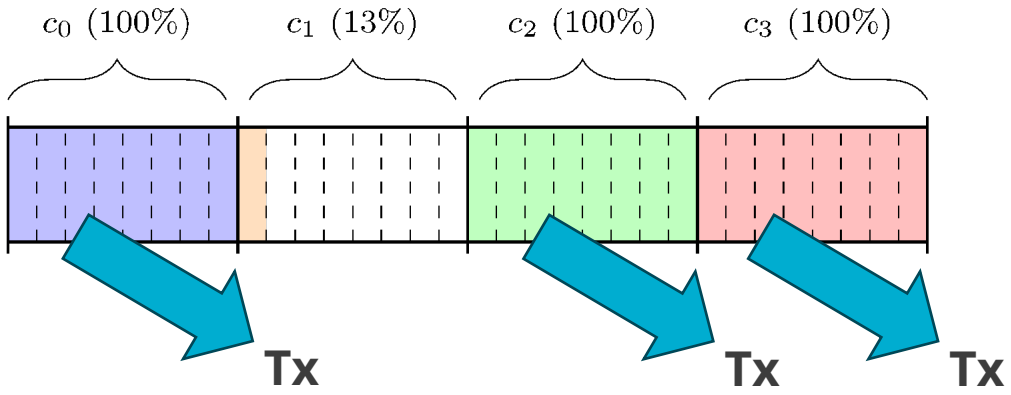
In **traditional communication**, the buffer is sent only when all of the contributors have finished

A delay in one contributor delays the sending of *all* the data



INTRODUCTION

In **fine-grained communication**, transmission can be initiated on completed *partitions* of buffers



Potential benefits versus traditional communication:

- **Early work**: receiver has data to work with earlier than otherwise
- **Early delivery**: entire buffer contents may be received earlier than otherwise

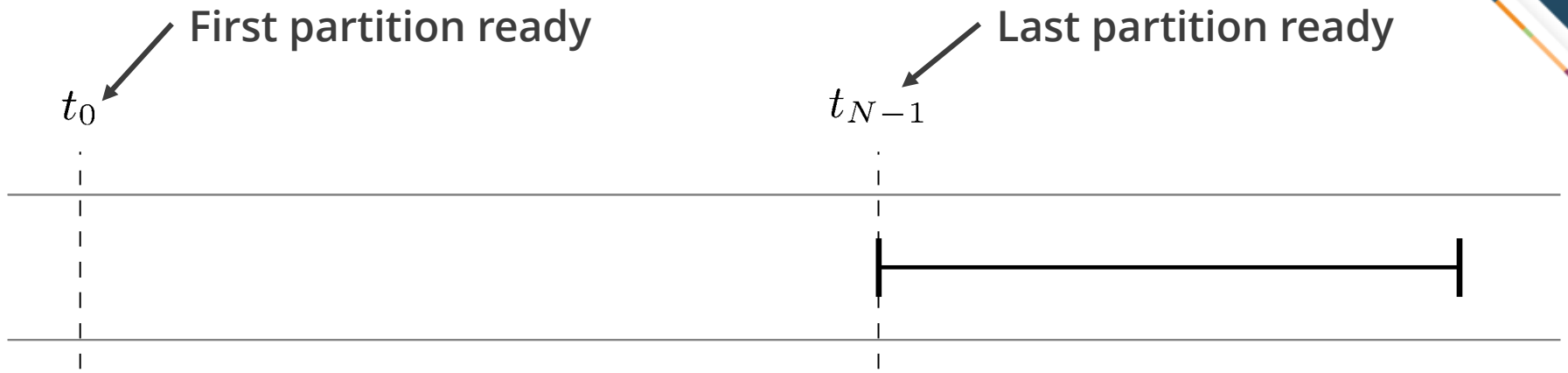
See also:
MPI 4.0 Partitioned Communication API (Chapter 4)

INTRODUCTION

- Questions
  - What are the impacts of communication overheads?
  - How many partitions are most effective for a given buffer size?
  - What are the consequences of the relative completion times of contributors?

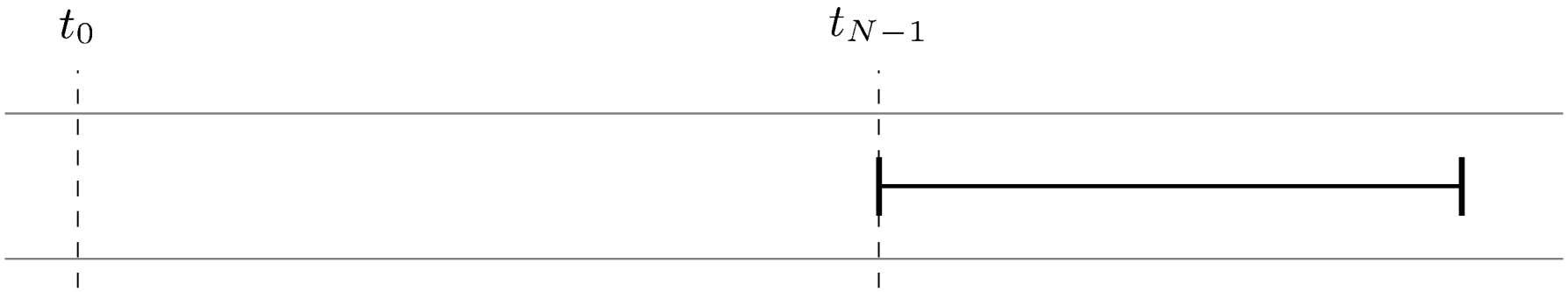
INTRODUCTION

Traditional

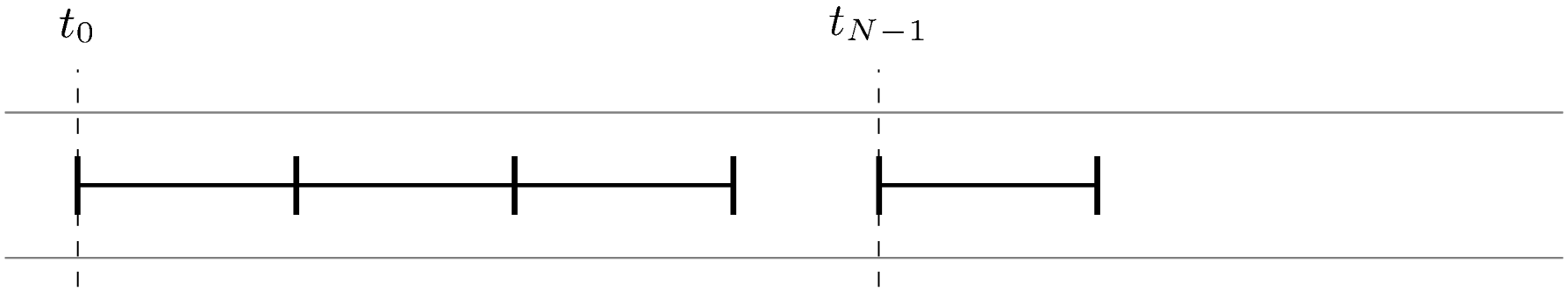


INTRODUCTION

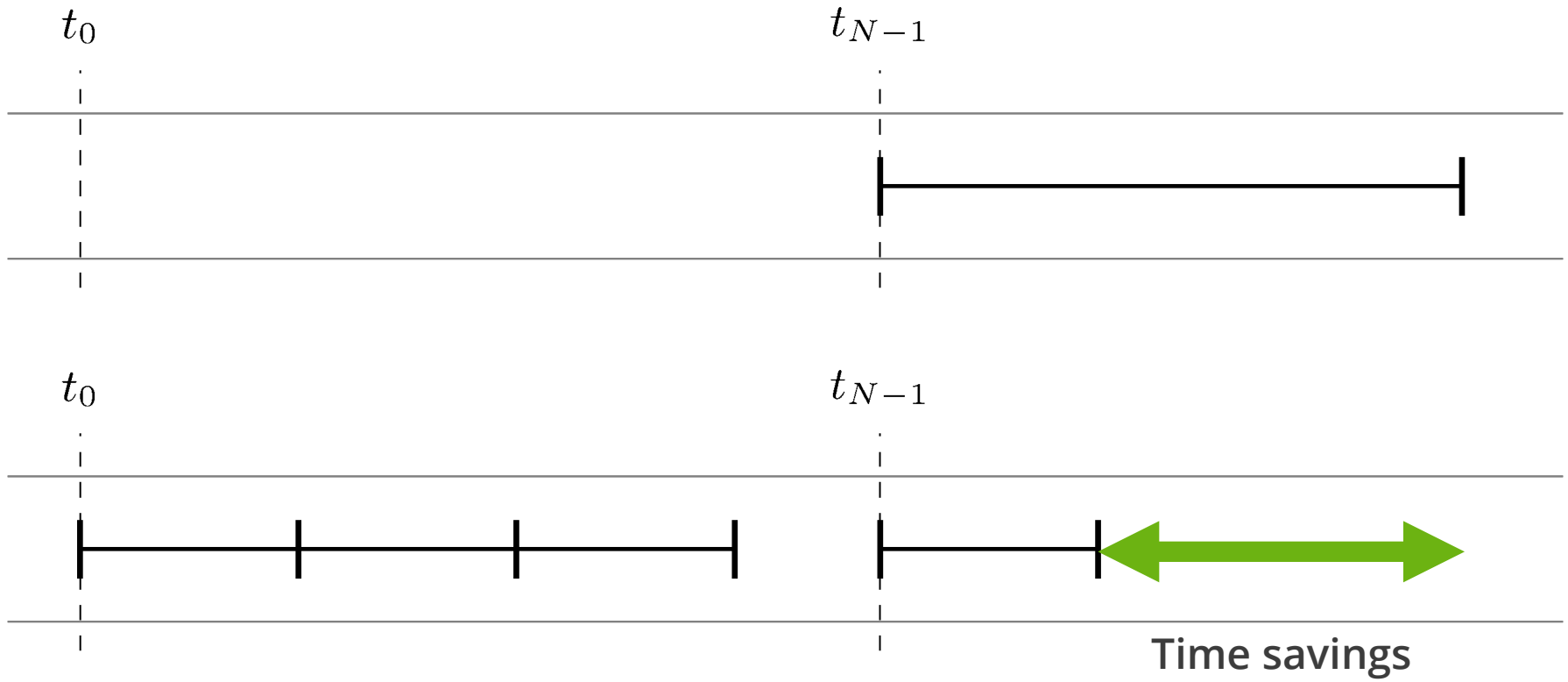
Traditional



Many-before-one

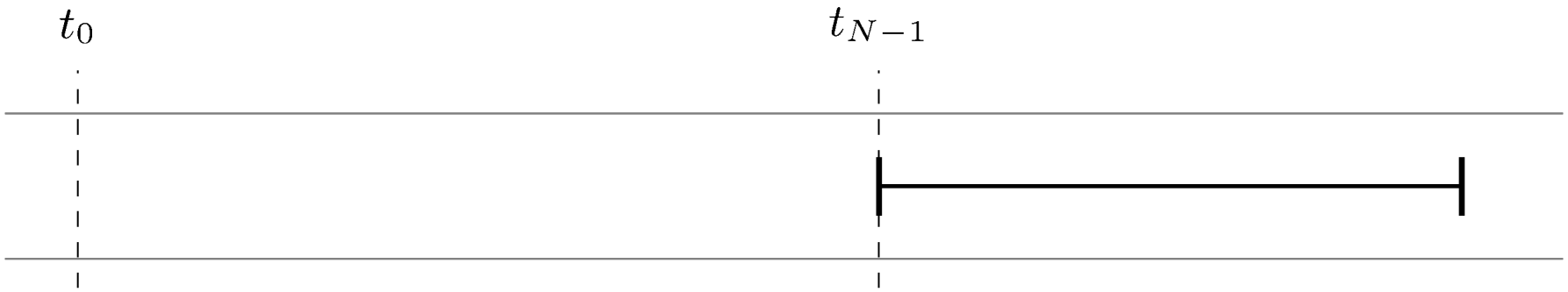


INTRODUCTION

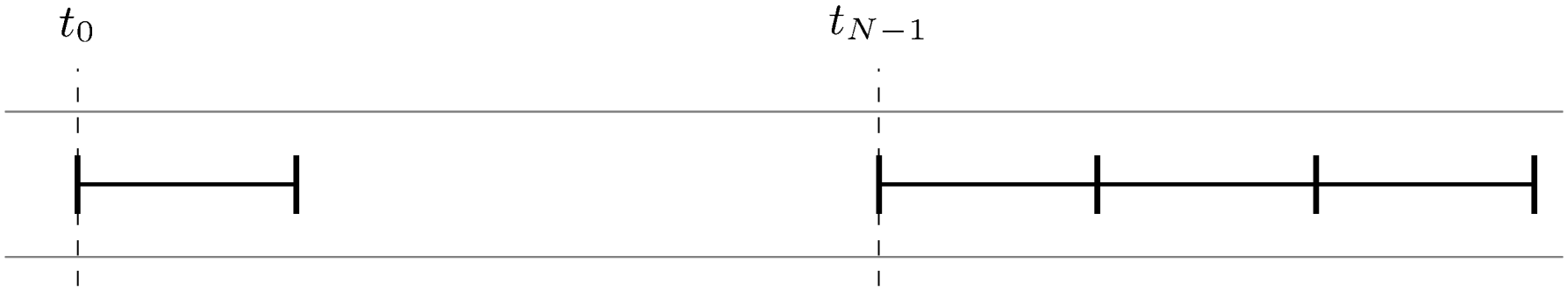


INTRODUCTION

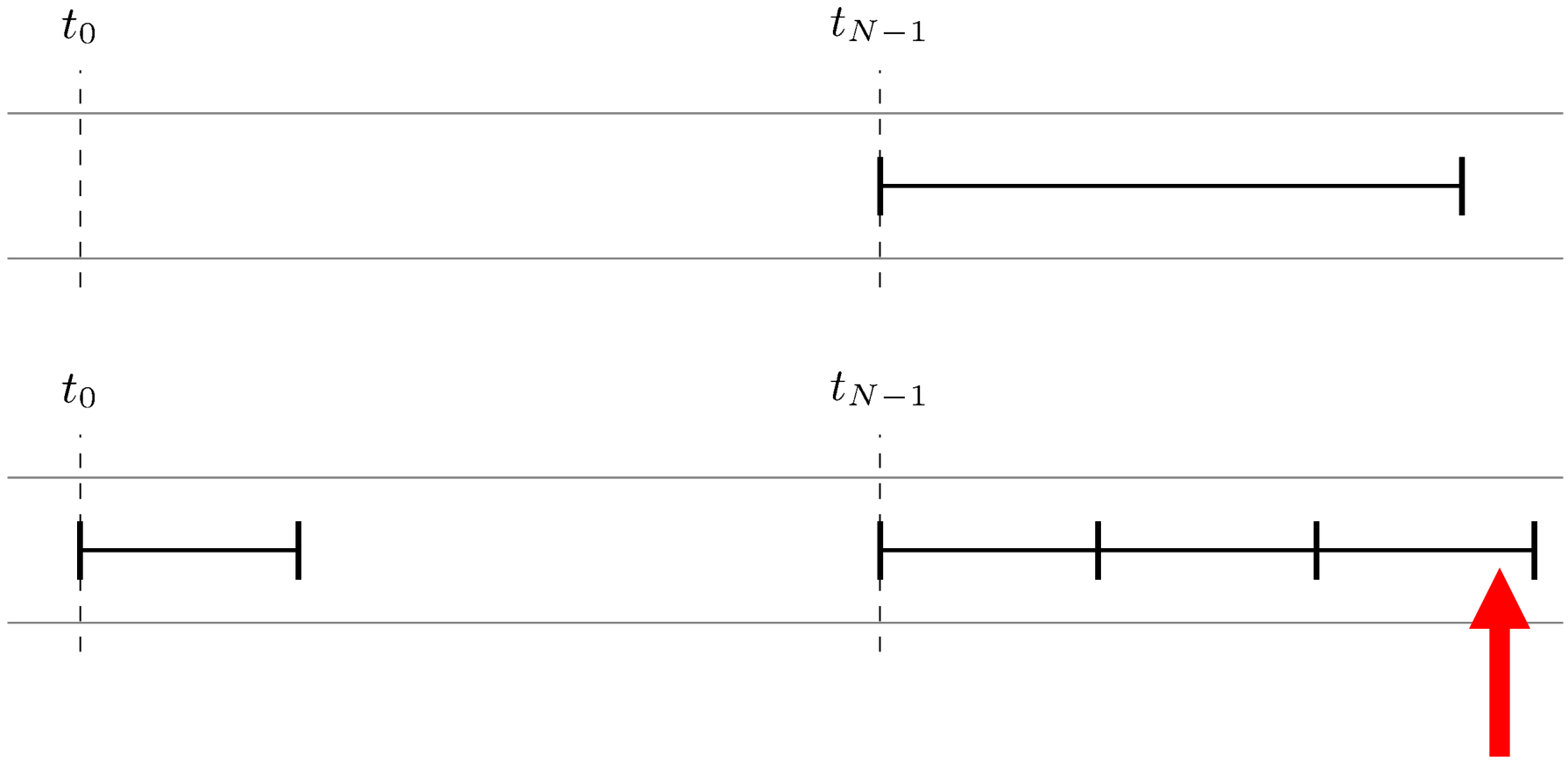
Traditional



One-before-many



INTRODUCTION



INTRODUCTION

- Questions
  - What are the impacts of communication overheads?
  - How many partitions are most effective for a given buffer size?
  - What are the consequences of the relative completion times of contributors?
  - What happens on real-world networks?

INTRODUCTION

Strategy: Consider these questions from the perspective of **modeling** plus **benchmarks**

Model: Identify trends in fine-grained communication behavior

Benchmarks: Consider how those trends manifest in real-world networks (or not)

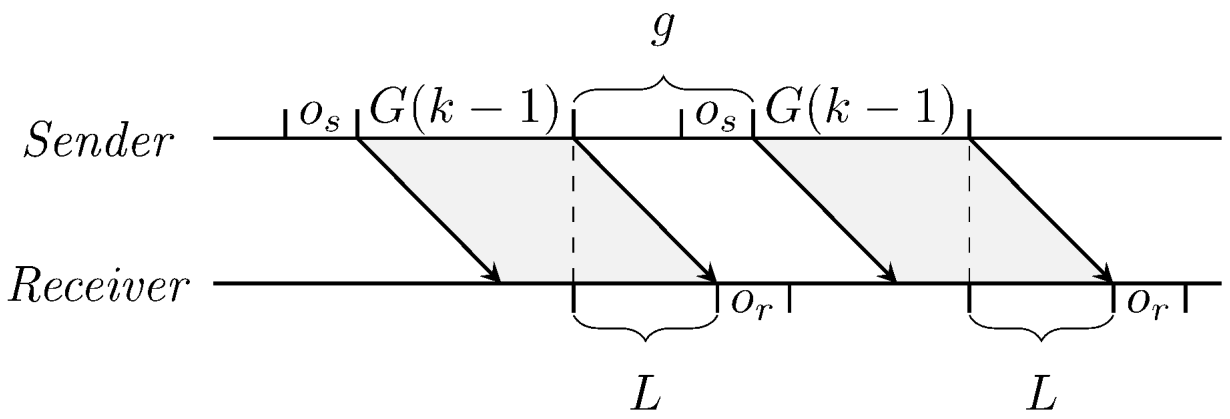
MODELING FINE-GRAINED COMMUNICATION

MODELING FINE-GRAINED COMMUNICATION

- LogP family of models
  - Long history of providing guidance on the design of parallel algorithms
  - Model communication time as linear function of overheads, inter-message gaps, etc.
  - For this study we extend the LogGP variant

MODELING FINE-GRAINED COMMUNICATION

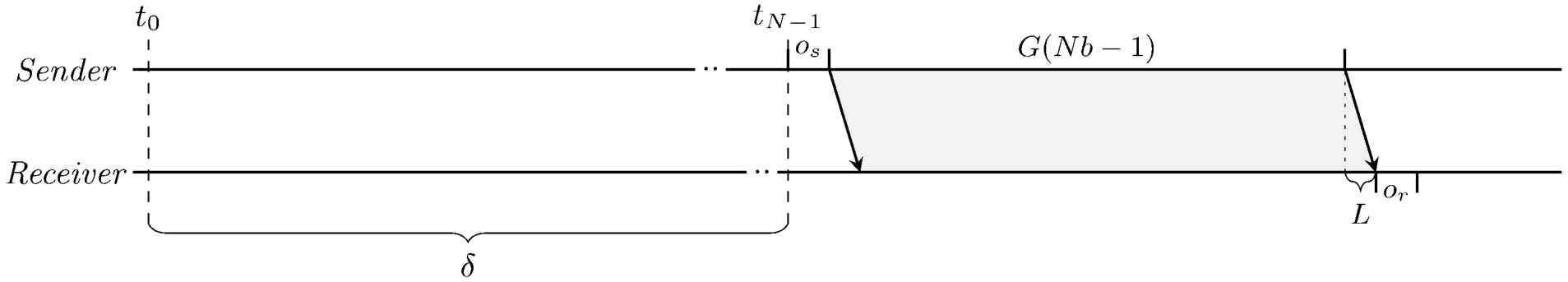
LogGP Model	
k	Number of bytes to send
L	Latency of sending a message between processes
o_s, o_r	Sender and receiver processor time required
g	Minimum time between consecutive messages
G	Inter-byte gap



$$o_s + 2G(k-1) + g + L + o_r$$

MODELING FINE-GRAINED COMMUNICATION

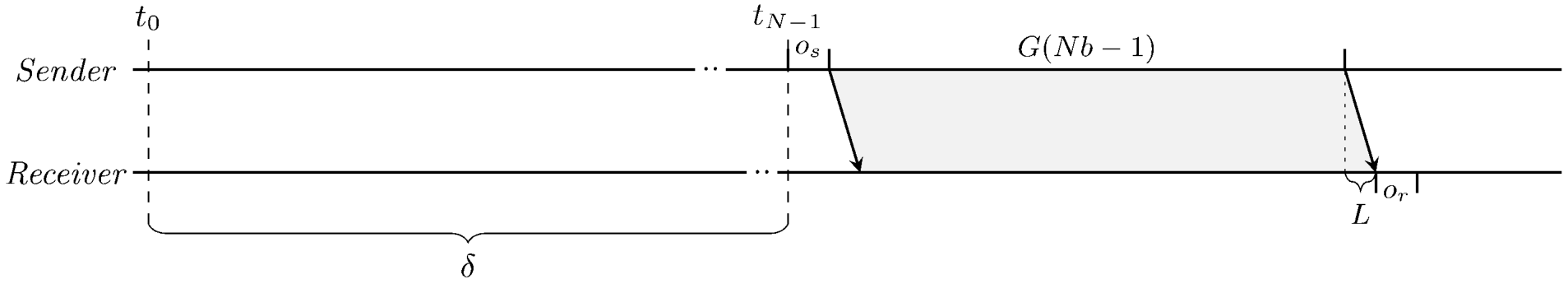
Traditional



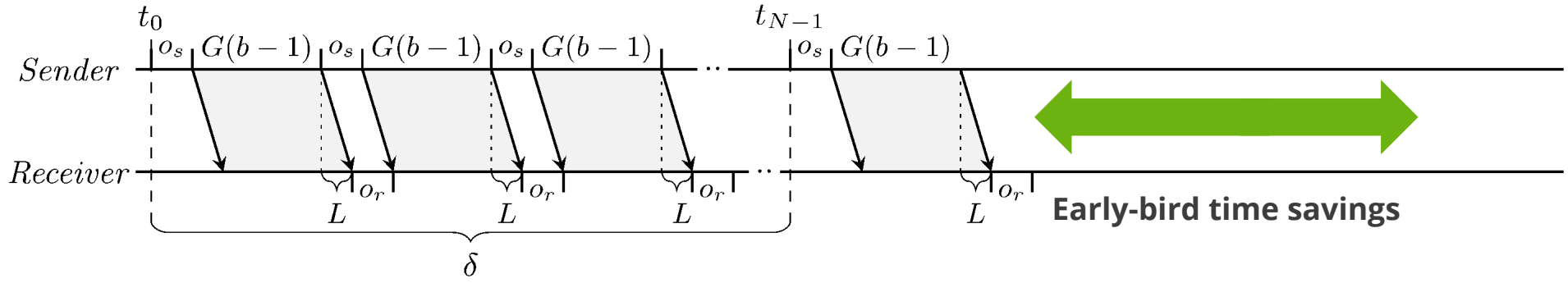
Extended LogGP Model	
N	Number of partitions
b	Bytes per part (= *k*/*N*)
t_i	Time when part *i* is ready (0 <= *i* < N)
delta	`delay' between part 0 and part *N*-1

MODELING FINE-GRAINED COMMUNICATION

Traditional

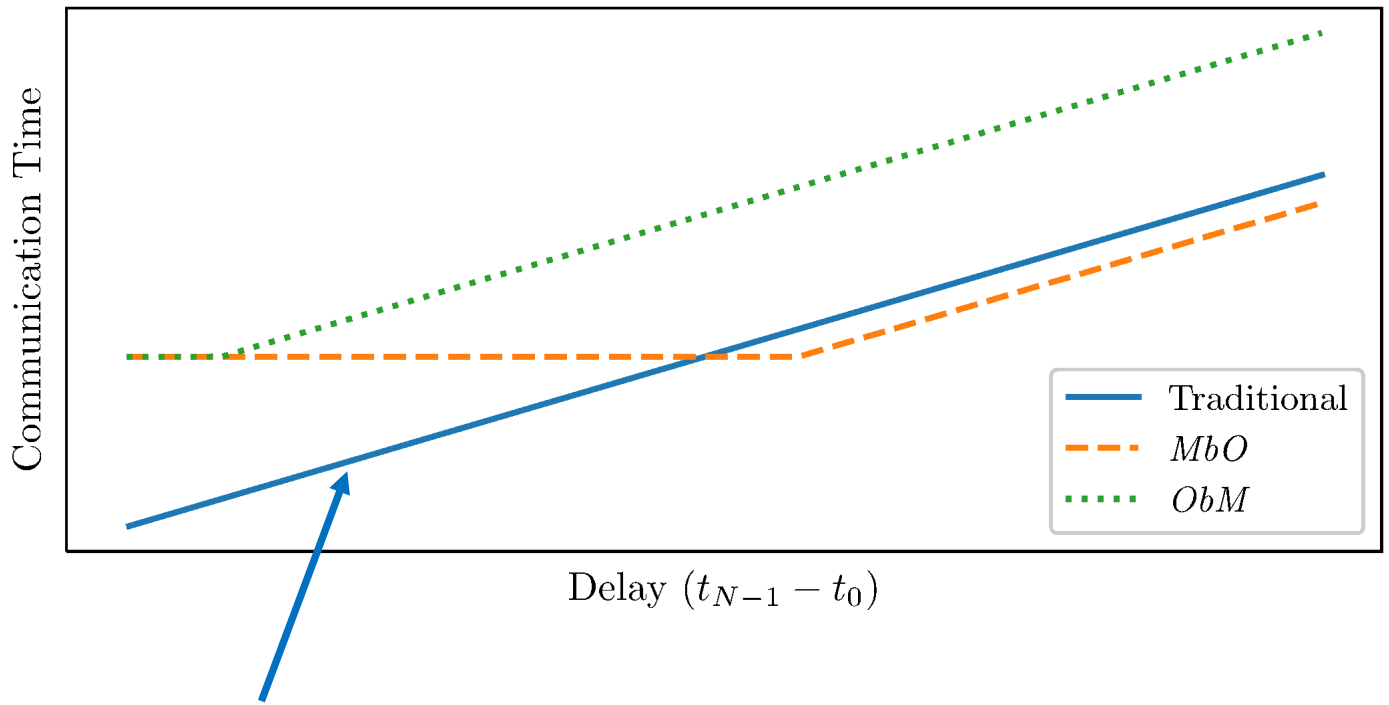


Many before one



MODELING FINE-GRAINED COMMUNICATION

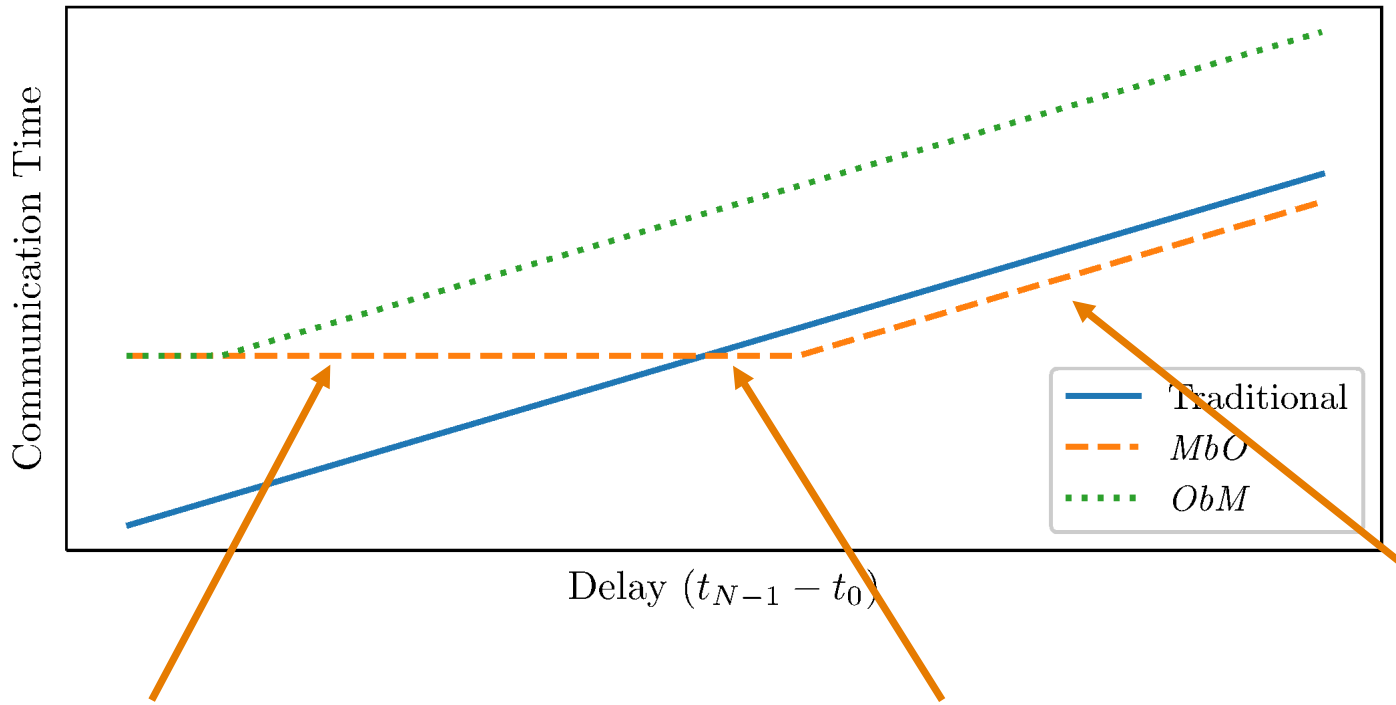
Model Result: Impact of completion times on minimum delay required for benefit



Traditional: As delay increases, communication time increases

MODELING FINE-GRAINED COMMUNICATION

Model Result: Impact of completion times on minimum delay required for benefit



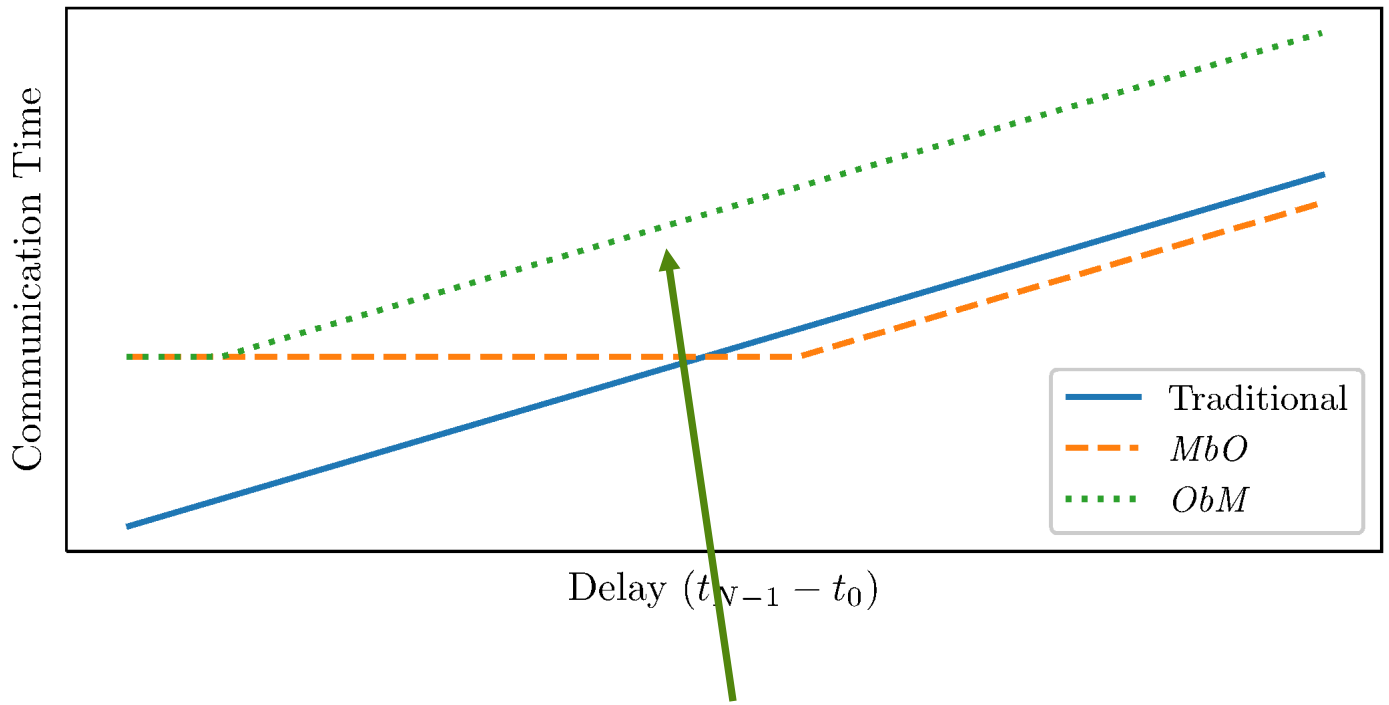
Many-before-One: For small delays, partitions stack; performs worse than traditional

Eventually, delay is sufficiently large to cover a portion of the communication of N-1 partitions; fine-grained performs better than traditional

Further increases merely delay the sending of the last partition; maximum benefit is reached

MODELING FINE-GRAINED COMMUNICATION

Model Result: Impact of completion times on minimum delay required for benefit

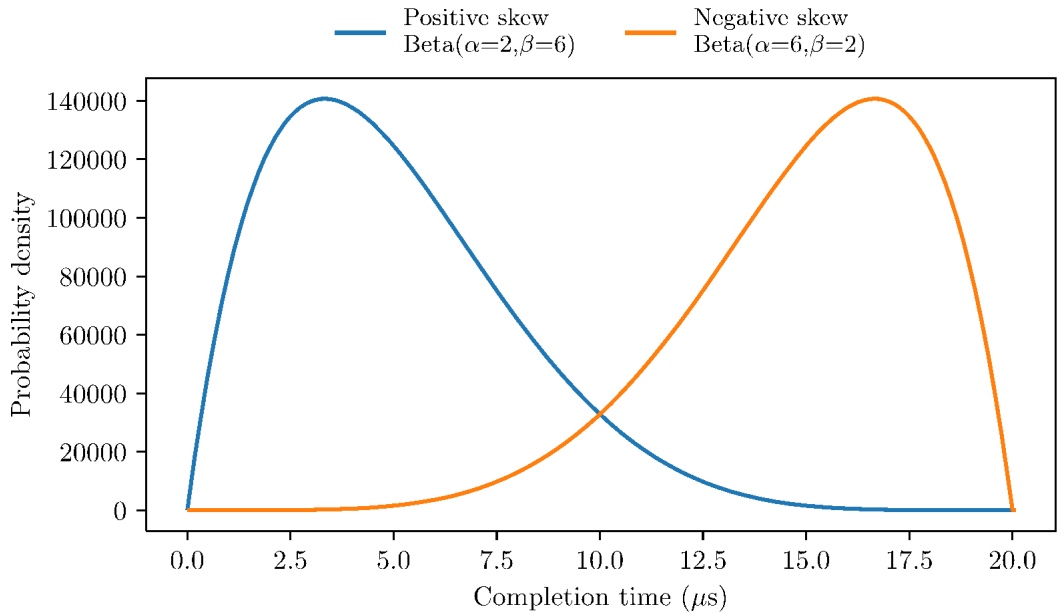


One-before-Many: No amount of delay can overcome the cost of the additional overheads; always performs worse than traditional

BENCHMARKING FINE-GRAINED COMMUNICATION

BENCHMARKING FINE-GRAINED COMMUNICATION

- Benchmarks
  - Standard latency (ping-pong) benchmark
  - Uses MPI non-blocking sends and receives
  - Extended to allow control of when sends are issued (the “completion schedule”)
    - Positive skewed beta distribution (analogous to Many-then-One scenario)



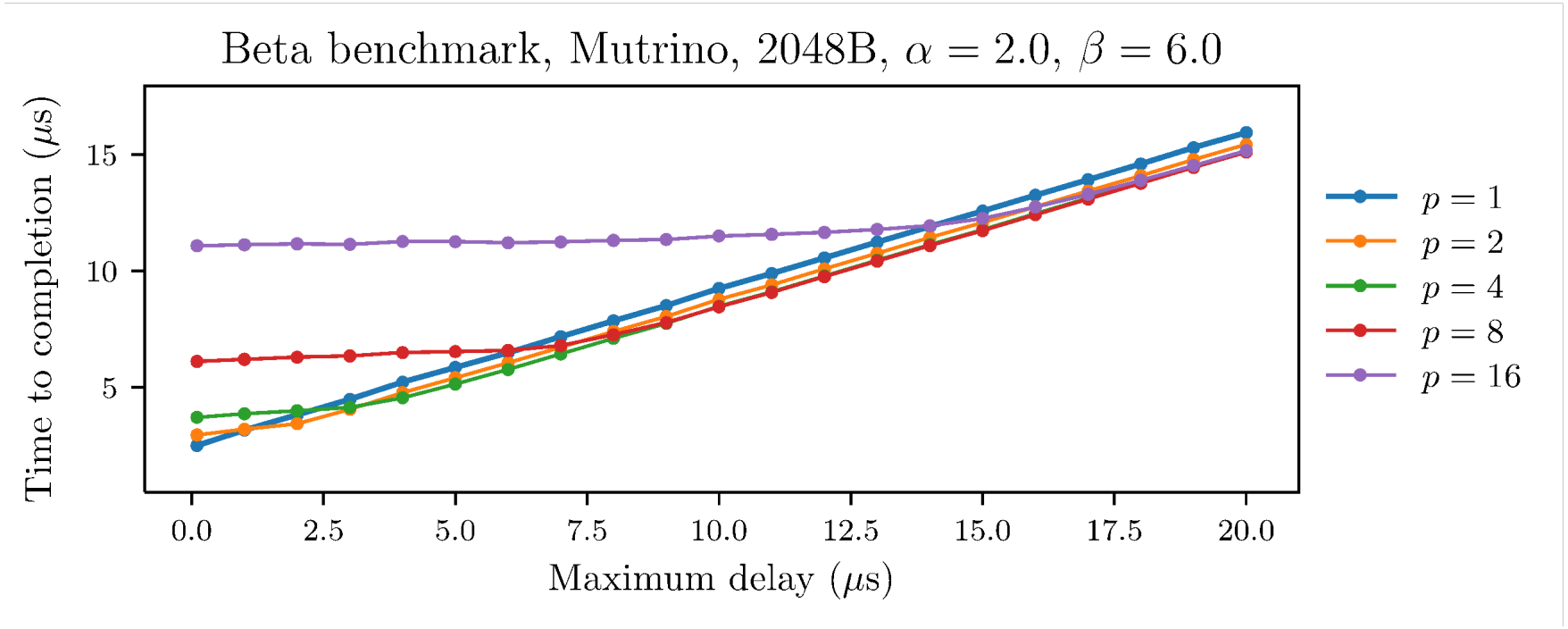
Beta distributions

BENCHMARKING FINE-GRAINED COMMUNICATION

System	Compute	Network
Manzano	Intel Cascade Lake 8268 2.9 GHz	Intel Omni-Path 100 Series (fat tree)
Mutrino	Intel Xeon ES-2698 2.5 GHz (Cray XC40)	Cray Aries (dragonfly)
Stria	Arm Cavium Thunder-X2 2.0 GHz	Mellanox ConnectX-5 Infiniband EDR (fat tree)

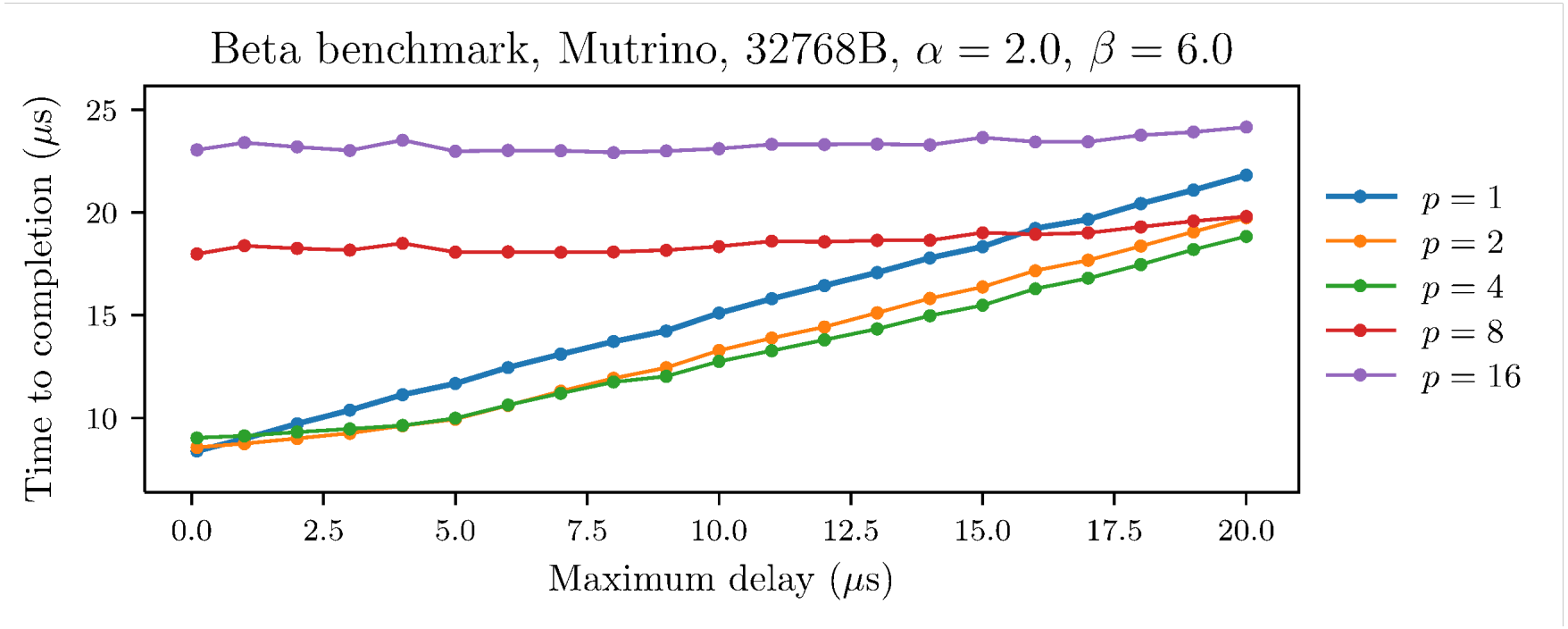
- Vary total buffer size from 1KiB to 2MiB, number of partitions from 1 (traditional) to 128.
- 25000 trials per configuration across 50 runs.
- Vary maximum possible beta distribution delay from 0 usecs to 20 usecs

BENCHMARKING FINE-GRAINED COMMUNICATION

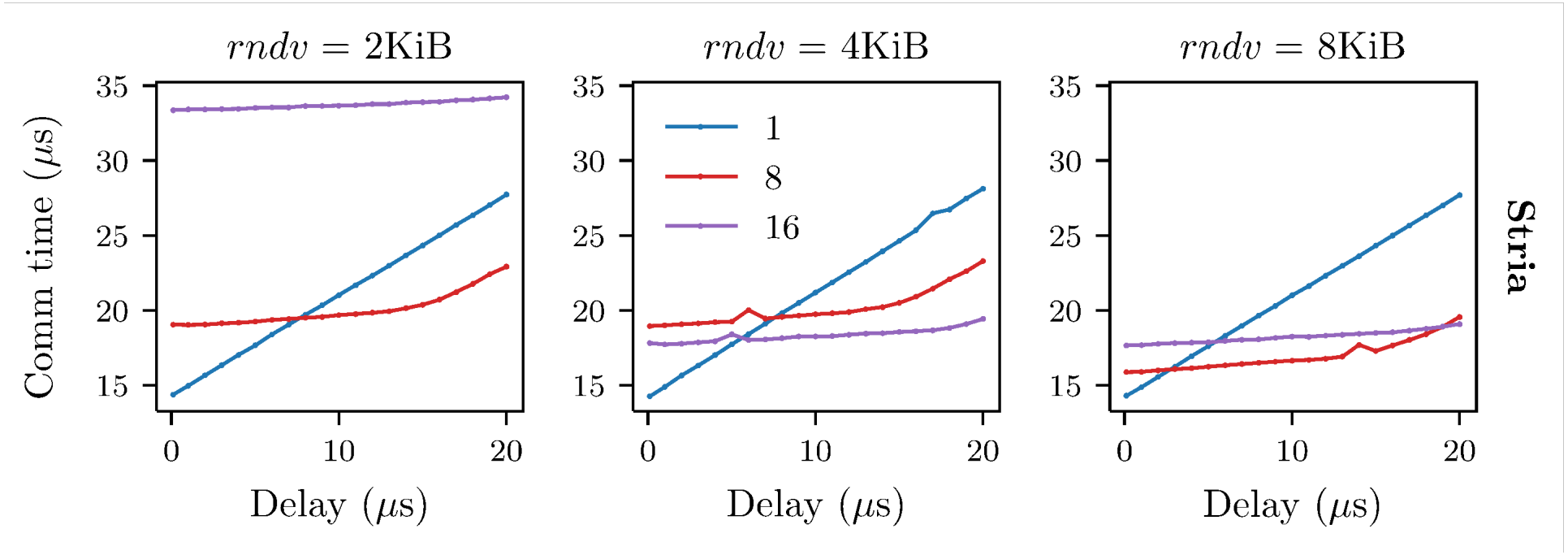


Beta positive skew (analogous to many-before-one)

BENCHMARKING FINE-GRAINED COMMUNICATION



BENCHMARKING FINE-GRAINED COMMUNICATION



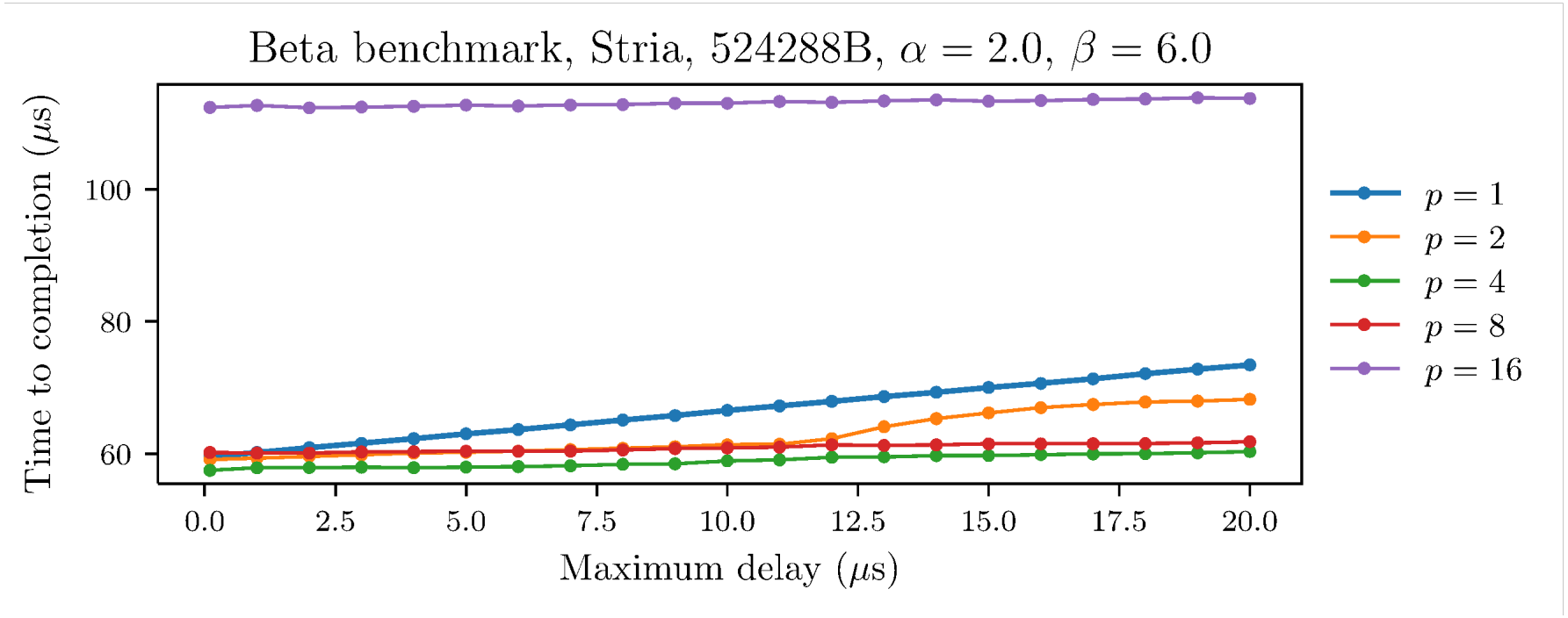
Buffer size = 32KiB

All communication is rendezvous: p = 8 outperforms p = 16

p = 16 is below threshold: Outperforms p = 8

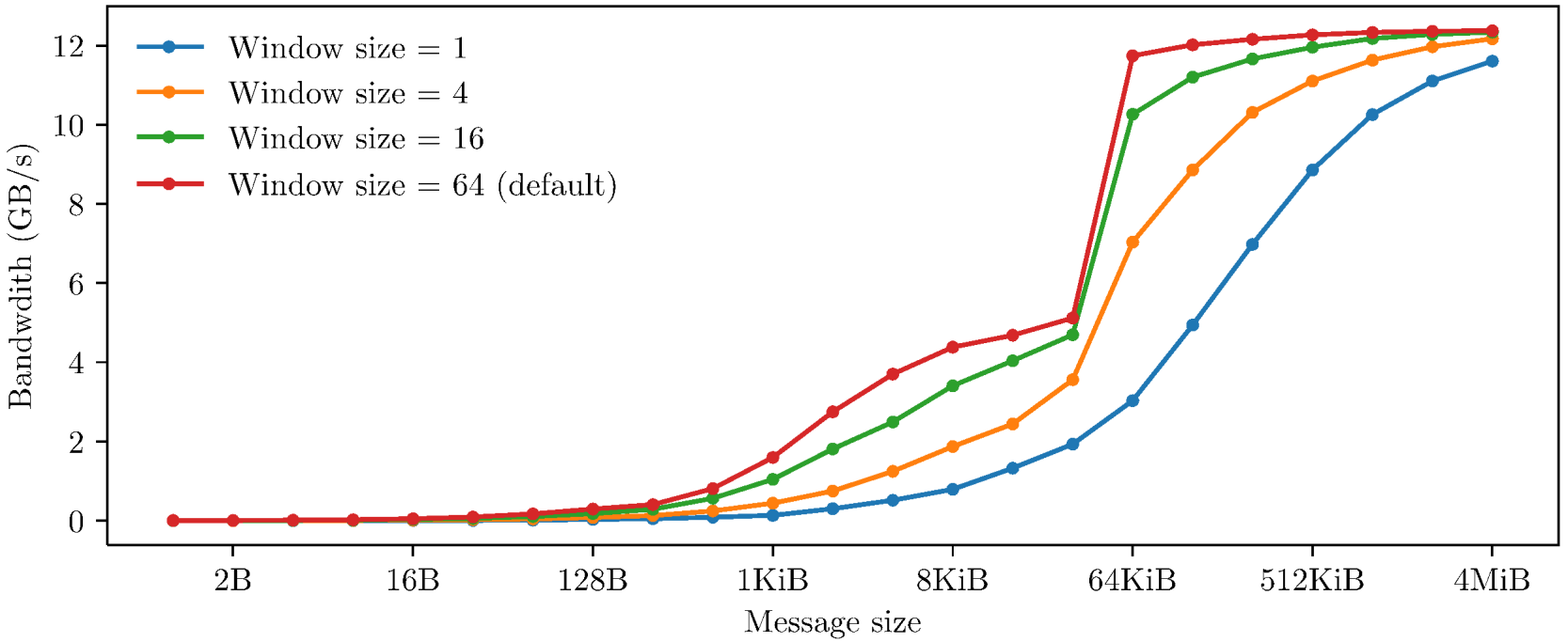
p = 8 and p = 16 are below threshold: p = 8 once again outperforms p = 16

BENCHMARKING FINE-GRAINED COMMUNICATION



BENCHMARKING FINE-GRAINED COMMUNICATION

- OSU bandwidth benchmark (`osu_bw`)
- Window size: number of communication operations posted before `MPI_Waitall` is called.
- Bandwidth jump for larger window sizes between 32KiB and 64KiB
- So, when 512KiB is split into 16 partitions, each partition is 32KiB, below the bandwidth jump



CONCLUSION

SUMMARY

- Model can provide guidance regarding fine-grained communication behavior, especially for small or mid-size messages
  - Temucin et al., “A Dynamic Network-Native MPI Partitioned Aggregation Over InfiniBand Verbs” Cluster 2023
- Understanding application characteristics (e.g. completion schedules) is critical to securing a benefit from fine-grained communication.
  - Marts et al., “Measuring Thread Timing to Assess the Feasibility of Early-bird Message Delivery”, P2S2 2023
- Understanding network details is equally important to securing benefits of fine-grained communication. The model provides a point of reference for isolating network-specific behaviors.

Whit Schonbein - wwschon@sandia.gov

Q&A