TECHNISCHE UNIVERSITÄT DRESDEN

The Hebrew University of Jerusalem

Faculty of Computer Science Institute of Systems Architecture, Operating Systems Group

OVERHEAD OF A DECENTRALIZED GOSSIP ALGORITHM ON THE PERFORMANCE OF HPC APPLICATIONS

ELY LEVY, AMNON BARAK, AMNON SHILOH,
MATTHIAS LIEBER, CARSTEN WEINHOLD, HERMANN HÄRTIG

MOTIVATION

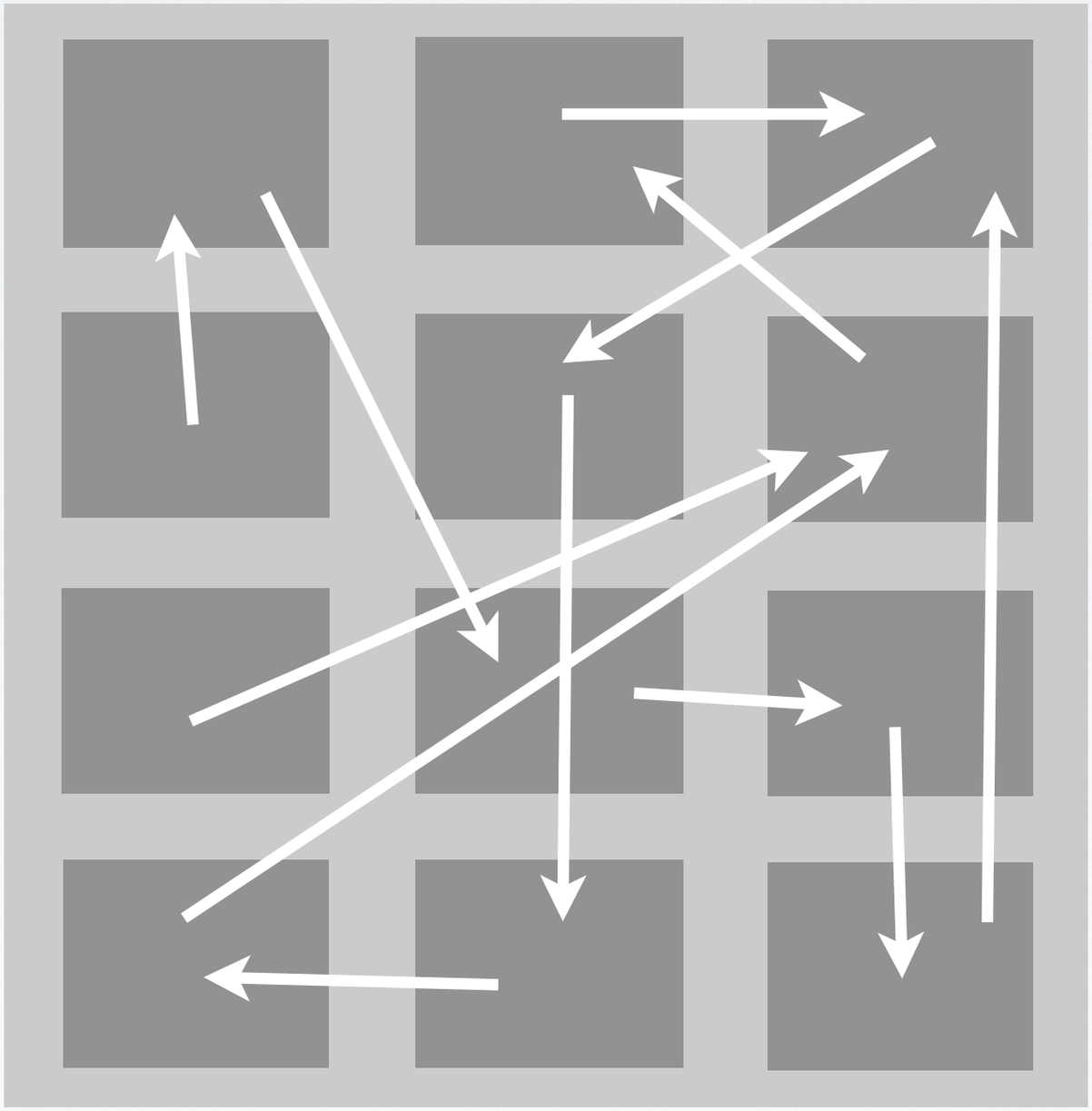
- Management tasks in supercomputers:
  - Process placement
  - Load management
  - System monitoring
- Up-to-date information required to make informed decisions

REQUIREMENTS

- Low overhead on application performance
- Scalability:
  - Decentralized information dissemination
  - Decentralized decision making
- Fault tolerance

TECHNISCHE UNIVERSITÄT DRESDEN

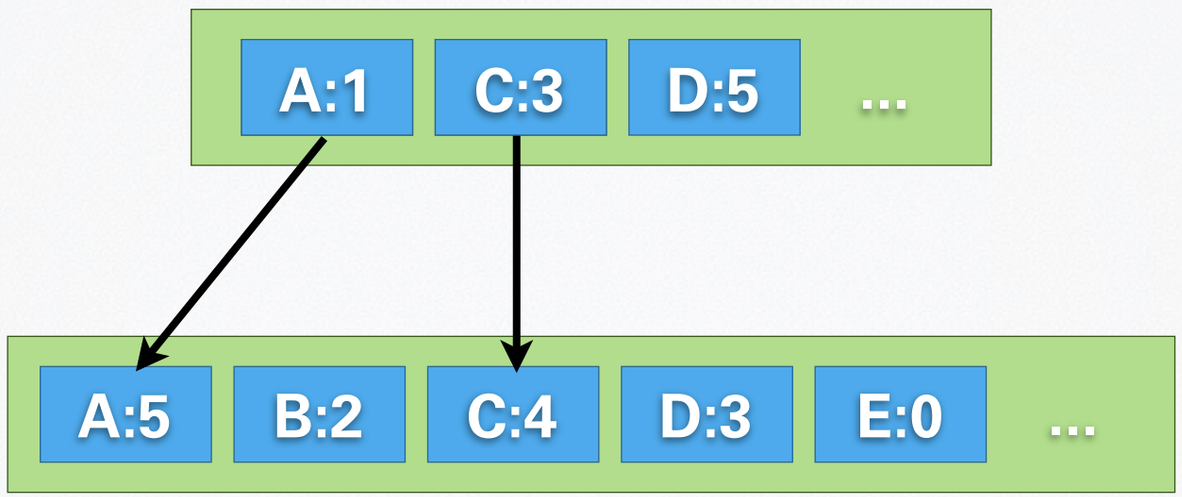
RANDOMIZED GOSSIP



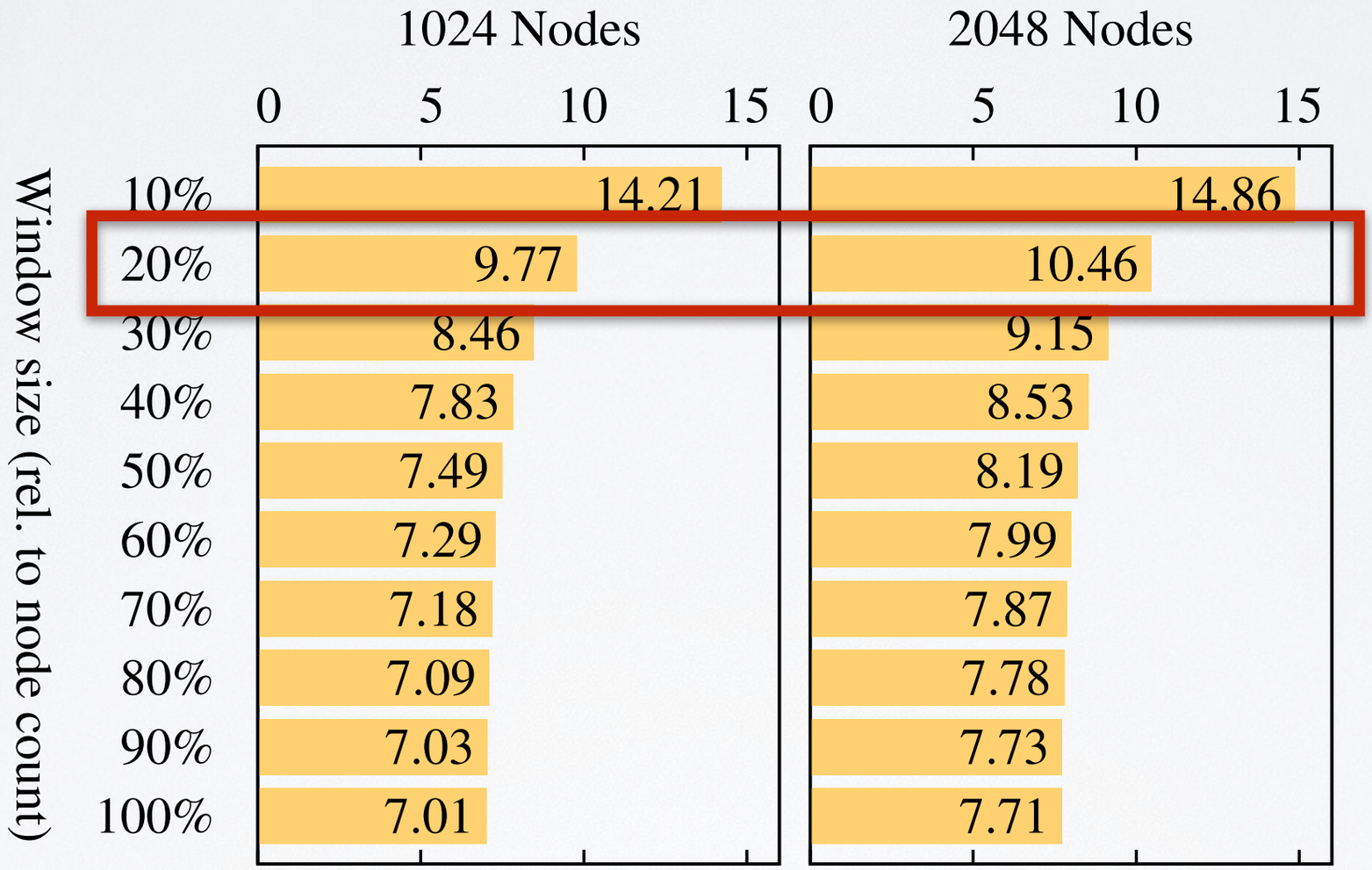
MERGING WINDOWS

Node A

Node E



WINDOW SIZE



How much data to send?

- Small window sizes already yield good average age
- Diminishing return for larger window sizes
- Example: 20% of 1024 nodes w/ 1KiB per node → 200 KiB

HARDWARE

- BlueGene/Q at Jülich (JUQUEEN)
- 28.672 nodes total (used 1024 – 8192)
- 16 cores per node (PowerPC A2 @ 1.6 GHz)
- 5D torus network (10 links per node)
- 2 GB/s per link send + receive
- Total bandwidth per node: 40 GB/s
- 2.6 µs worst-case latency

GOSSIP ALGORITHM

- MPI-based implementation (`MPI_Bsend`)
- Gossip algorithm runs on 1 core
- Application uses remaining 15 cores

- How to run two programs on BG/Q?
  - Gossip algorithm + application linked together
  - MPI communicators configured to hide every 16th core from application
  - Wrapped all uses of `MPI_COMM_WORLD`

BENCHMARKS

Heavy network usage

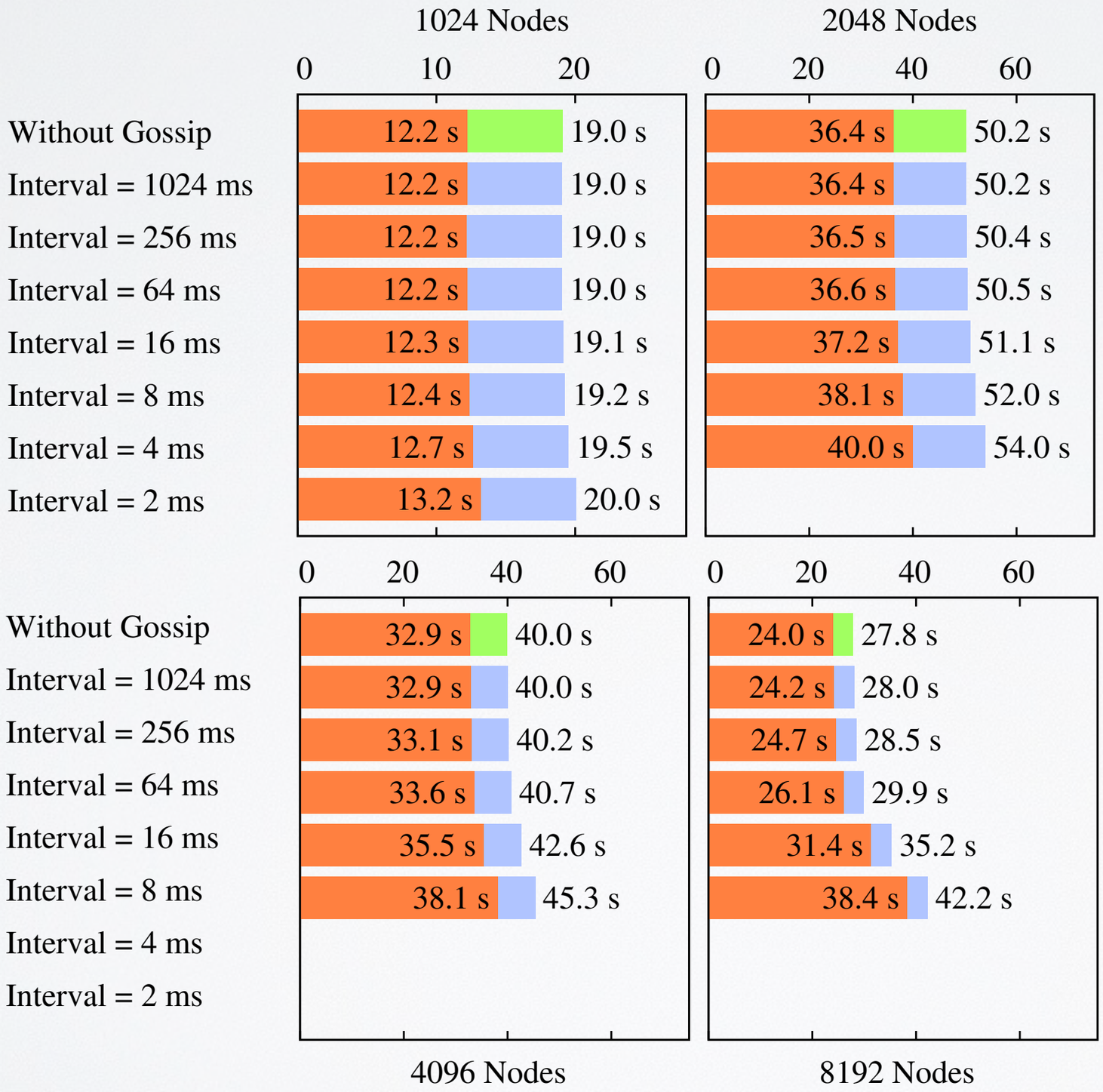
- HPCC suite: **MPI-FFT**
- HPCC suite: **PTRANS**
- Application: **COSMO-SPECS+FD4**

Moderate network usage

BENCHMARKS

- HPCC suite: **MPI-FFT**
- HPCC suite: **PTRANS**
- Application: **COSMO-SPECS+FD4**

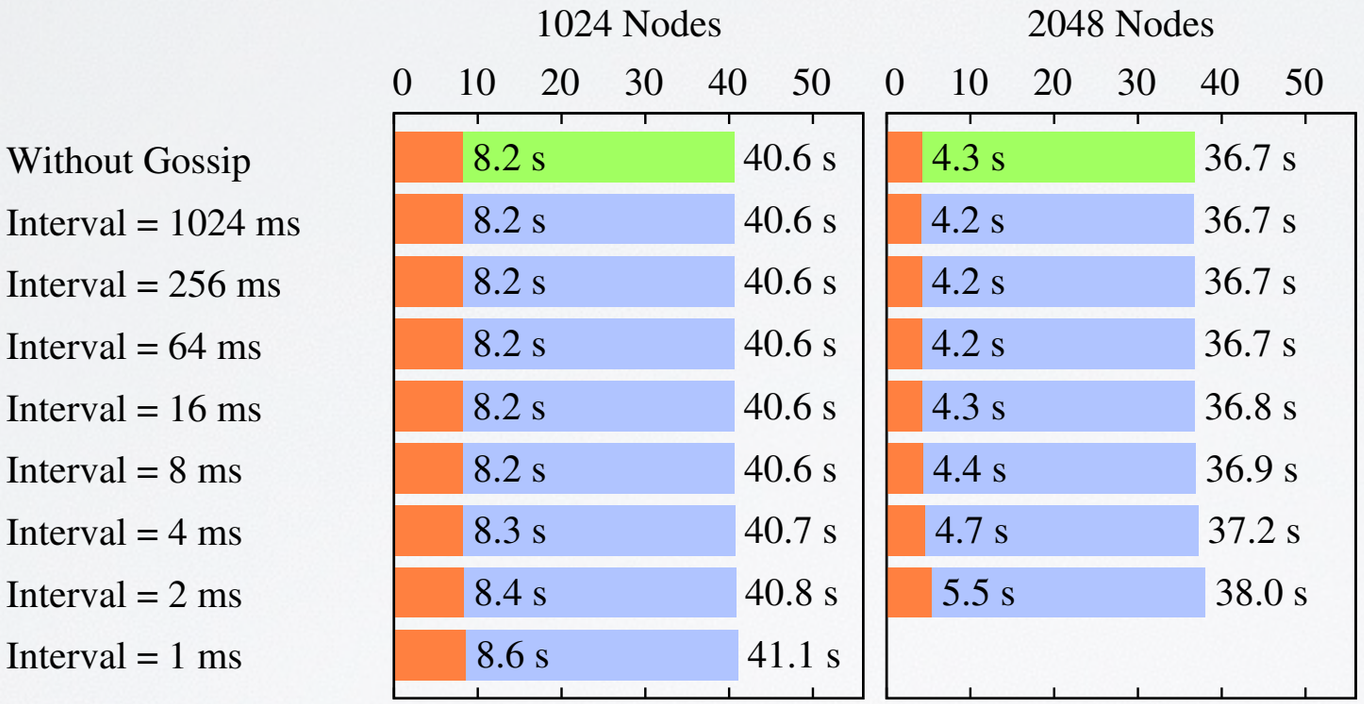
HPCC: MPI-FFT



Benchmark: Fast Fourier Transform

- All-to-all communication pattern
- Stresses bisection bandwidth
- 1024 nodes: 136 million vector elements (2025 GiB)
- 2048+ nodes: 544 million vector elements (8100 GiB)

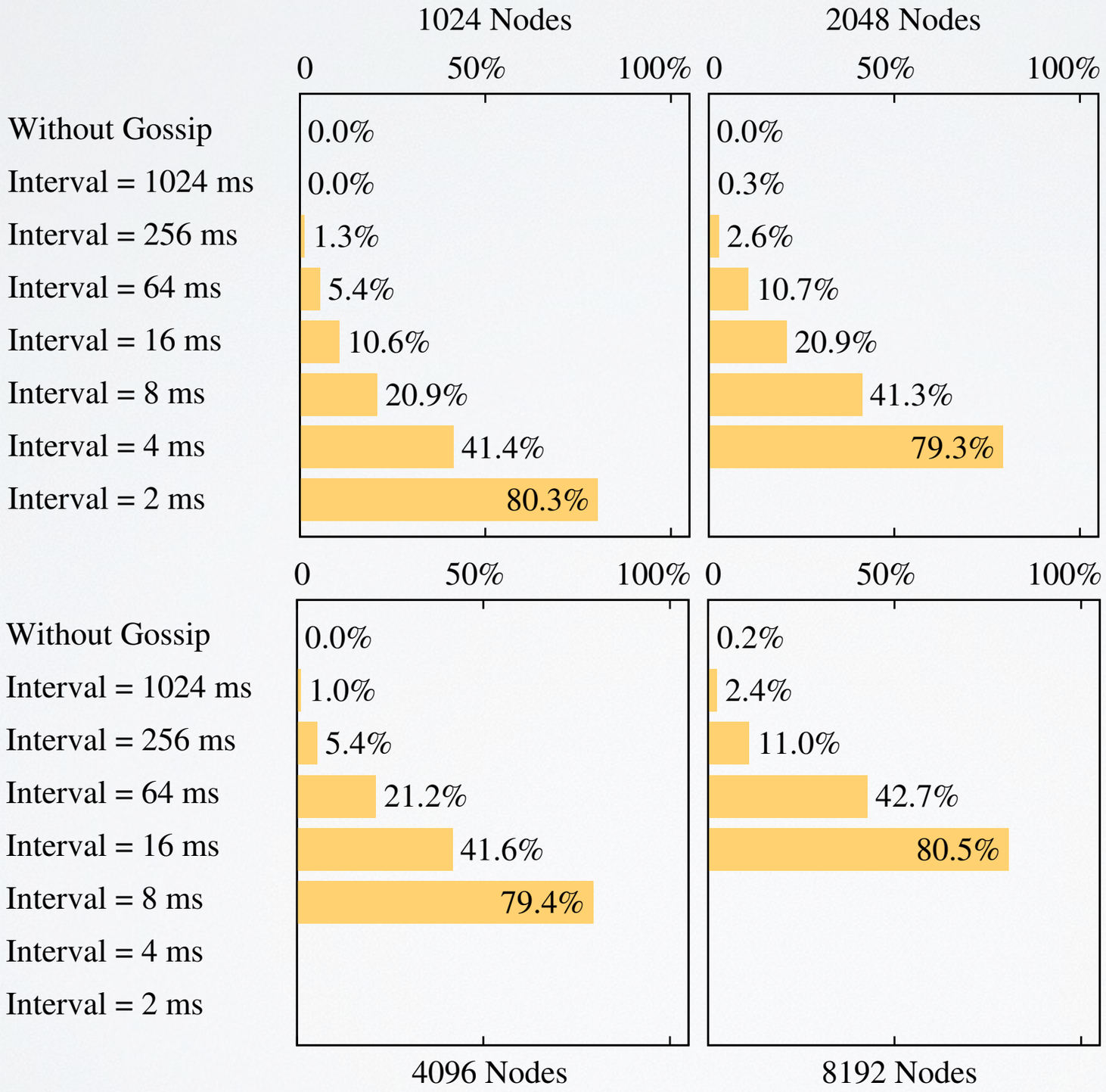
COSMO-SPECS+FD4



Benchmark: Atmospheric Simulation

- COSMO: static regular communication
- SPECS: dynamic, irregular communication
- Model coupling: dynamic, irregular, small volume
- Partitioning: collectives
- Migration: highly local, mostly between neighbors

GOSSIP SCALABILITY



Computational complexity:
O(n·log(n))

RATE VS WIN SIZE

	Window Size			
Gossip Interval	10 %	20 %	40 %	80 %
1 ms	3.8%			
2 ms	1.8%	4.7%		
4 ms	1.1%	2.6%	6.3%	17.2%
8 ms	0.7%	1.4%	3.2%	8.7%
16 ms	0.3%	0.7%	1.8%	4.5%
32 ms	0.3%	0.4%	0.9%	2.6%
64 ms	0.2%	0.3%	0.6%	1.5%

More Data

More Data

TECHNISCHE UNIVERSITÄT DRESDEN

REACTION TIME?

- Gossip intervals + average vector age:
  - 256 ms → **2-3 s**
  - 1024 ms → **10 s**

- Applicability for system services:
  - Global load information (allocation, …)
  - Local load balancing (MOSIX-like, …)
  - System monitoring (node health, …)

FUTURE WORK

- Other types of network (Infiniband, Cray, …)
- Fault tolerance, loss of messages

- Must adapt for exascale systems:
- Incomplete knowledge at each node
  - Groups of gossip nodes
  - Smaller vectors
- Hierarchical gossip for global view

TECHNISCHE UNIVERSITÄT DRESDEN

CONCLUSIONS

- Gossip algorithm scales to thousands of nodes
- Increasing window size causes more overhead than decreasing gossip interval
- Collective MPI communication most sensitive
- Gossip intervals of **256–1024 ms** with **no noticeable overhead** (in most cases)
- **Average age** of information at nodes in the order of **2–3 s** with gossip interval of 256 ms

FFMK.tudos.org

SPPEXA German Priority Programme 1648 **Software for Exascale Computing**