Rethinking Design Paradigm of Graph Processing System with a CXL-like Memory Semantic Fabric

Xu Zhang, Yisong Chang, Tianyue Lu, Ke Zhang, Mingyu Chen

State Key Lab of Processors, Institute of Computing Technology, Chinese Academy of Sciences

University of Chinese Academy of Sciences

Executive Summary

- **GraCXL is a distributed shared memory system**
  - Graph processing
  - Connected with CXL over fabric
  - Novel design paradigm
  - 1.33x-8.92x and 2.48x-5.01x improvement on CPU and FPGA

Custom FPGA Board with a Xilinx Zynq UltraScale+ MPSoC (ZynqMP) Chip

Outline

Background

Motivation

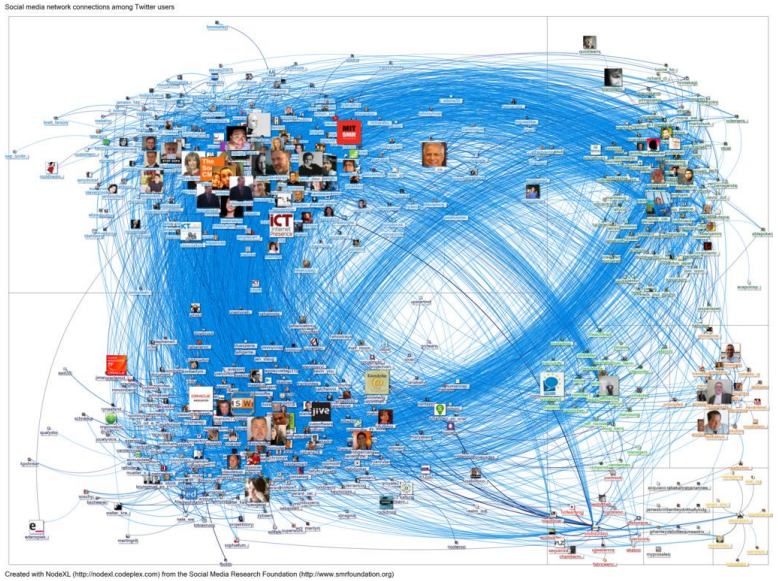
GraCXL

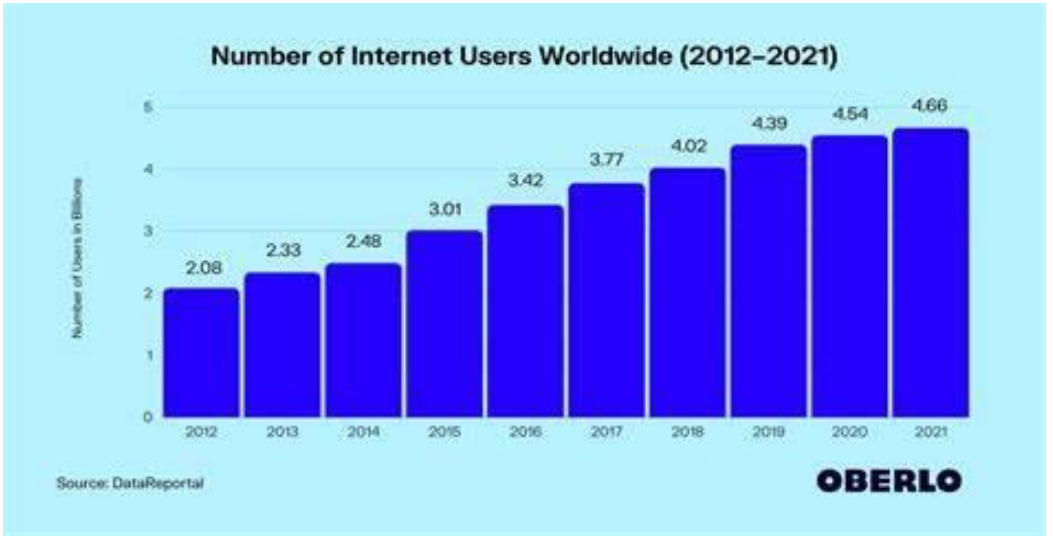
Evaluation

Conclusion

Graph Processing

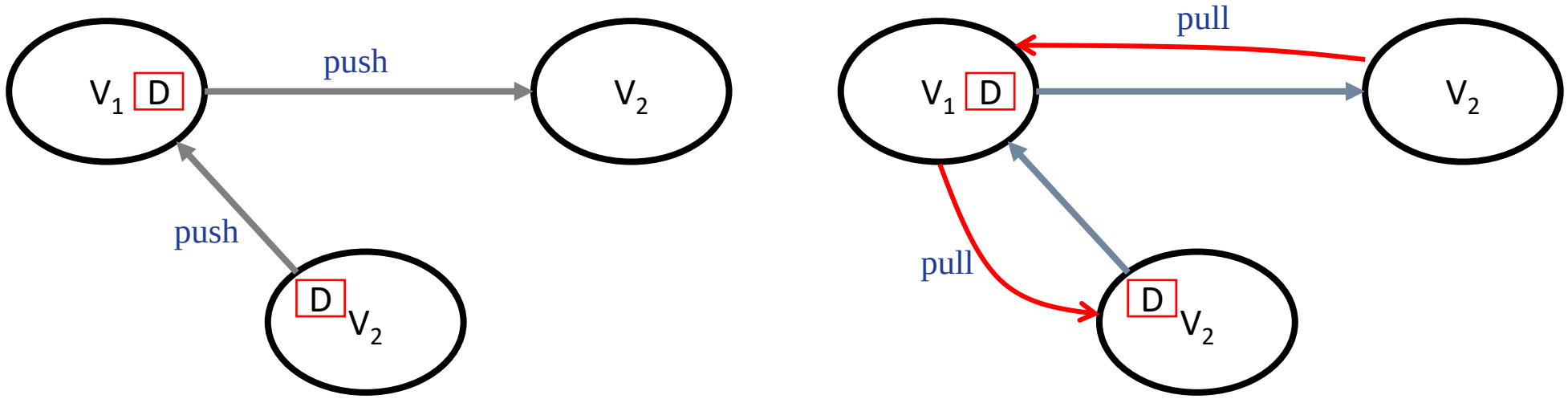
- Large scale graphs run out of single machine memory





Graph Processing Framework for Single Machine

- **Ligra is one of shared memory execution model for multicore system**
  - supersteps
  - pull or push

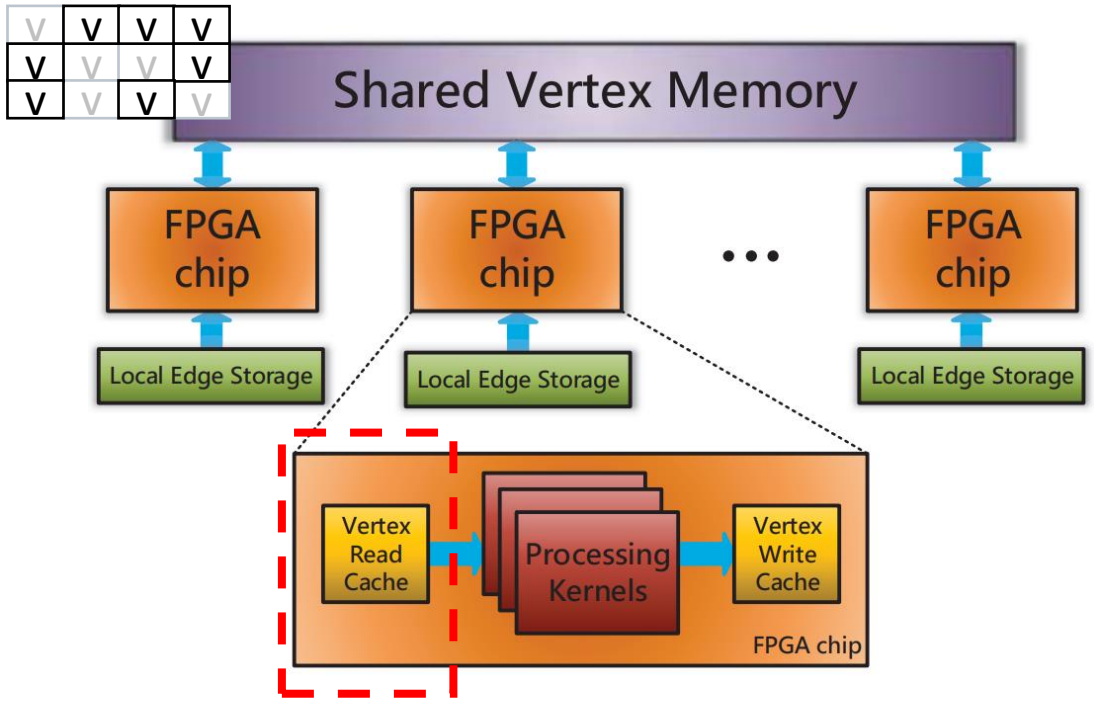


Julian Shun and Guy E. Blelloch. Ligra: A Lightweight Graph Processing Framework for Shared Memory. Proceedings of the ACM SIGPLAN Symposium on Principles and Practice of Parallel Programming (PPoPP), pp. 135-146, 2013.

Existing Distributed Shared Memory System (1/3)

FPGP

- A global shared vertex memory
- Each node maintains a vertex cache
- Redundant data movements

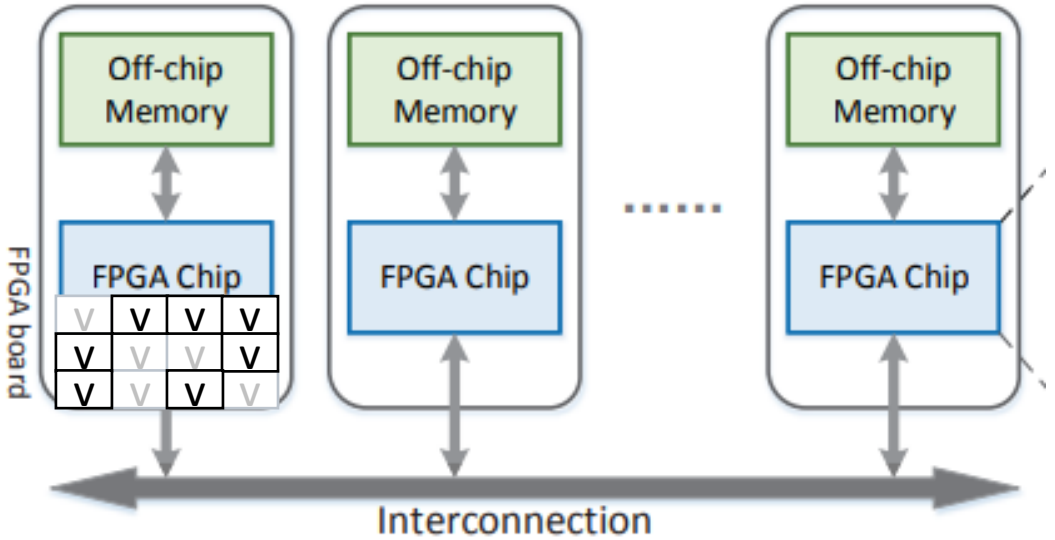


Guohao Dai, Yuze Chi, Yu Wang, and Huazhong Yang. 2016. FPGP: Graph Processing Framework on FPGA A Case Study of Breadth-First Search. The 2016 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA '16).

Existing Distributed Shared Memory System (2/3)

❖ ForeGraph

- Distributed shared memory
- Redundant data movements

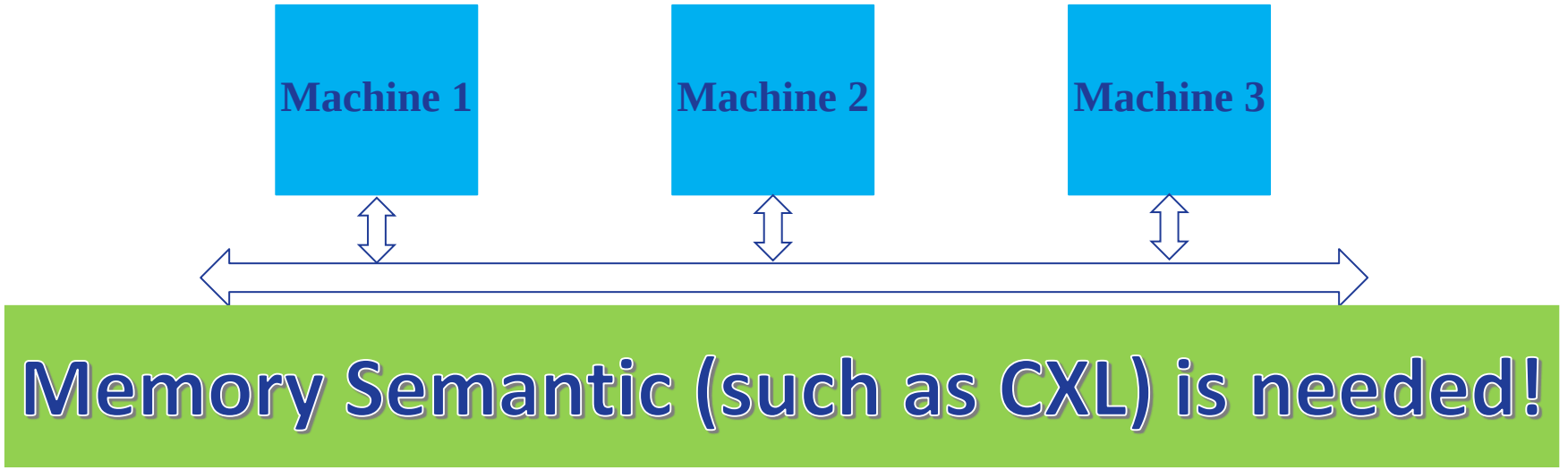


Guohao Dai, Tianhao Huang, Yuze Chi, Ningyi Xu, Yu Wang, and Huazhong Yang. 2017. ForeGraph: Exploring Large-scale Graph Processing on Multi-FPGA Architecture. The 2017 ACM/SIGDA International Symposium on Field-Programmable Gate Arrays (FPGA '17).

Existing Distributed Shared Memory System (3/3)

❖ SHMEMGraph

- Partitioned global address space
- Provide API to explicit transfer data



H. Fu, M. Gorentla Venkata, S. Salman, N. Imam and W. Yu, "SHMEMGraph: Efficient and Balanced Graph Processing Using One-Sided Communication," 2018 18th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID)

Pros & Challenges using CXL over fabric

- **Extending CXL over fabric to build multi-machine systems**
- **Pros**
  - Explicit movement: byte-addressable load-store (ld-st) semantics
  - Redundant movement: fine-granularity
- **Challenges**
  - Longer latency: 1-10us
  - Limited semantic
  - Unavailable hardware and platforms

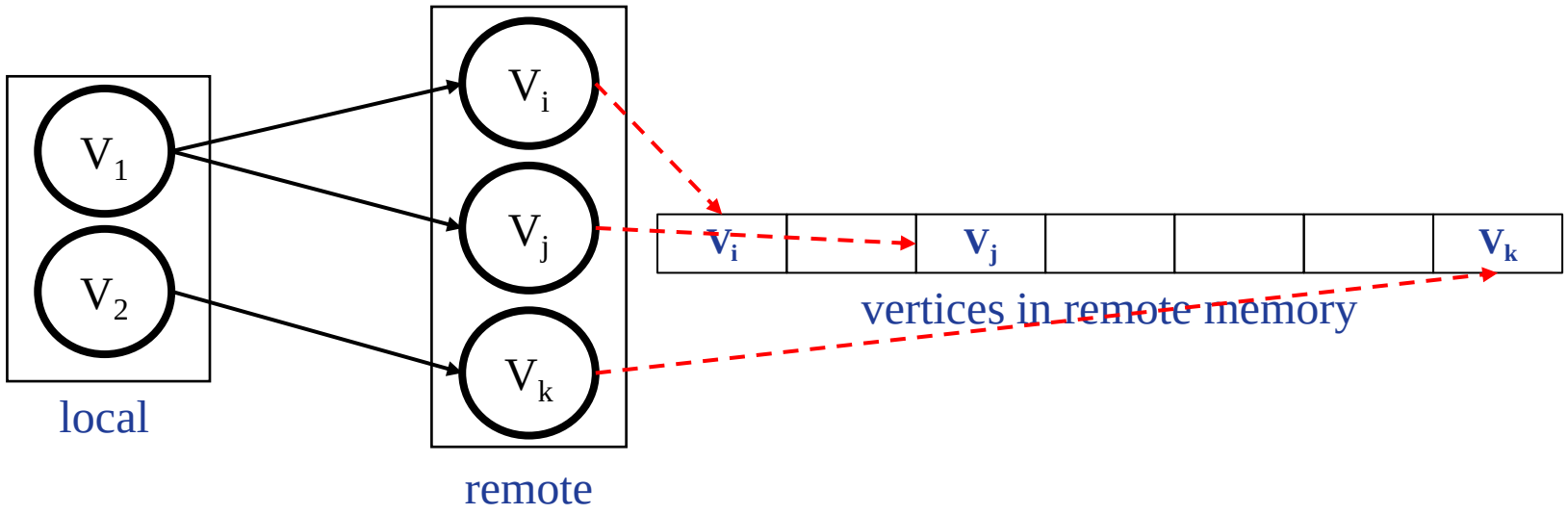
Outline

- Background
- Motivation
- GraCXL
- Evaluation
- Conclusion

Extending Ligra to DSM: What causes slow down

❖ Insights for pull mode

- Cache unfriendly: randomly traverse remote vertices
  - Waste cache capacity and fabric bandwidth ☹

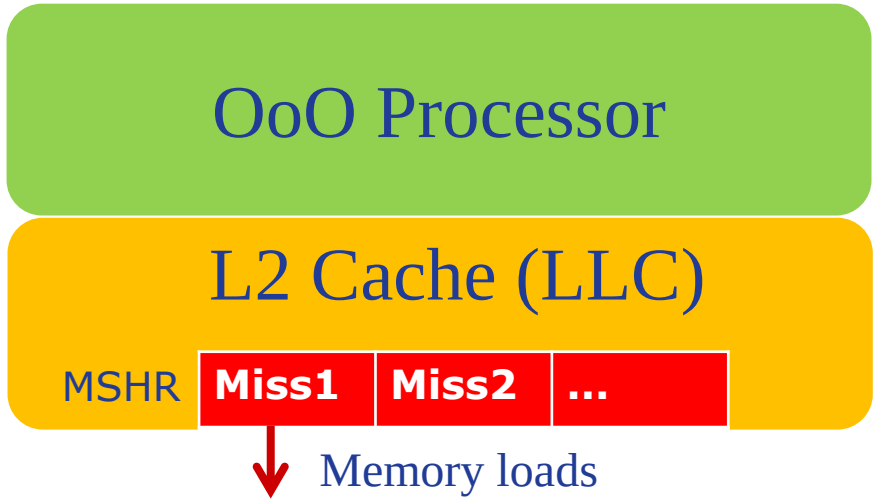


Extending Ligra to DSM: What causes slow down

❖ Insights for pull mode

- Limited outstanding loads constrained by MSHRs
- Computing stalled ☹

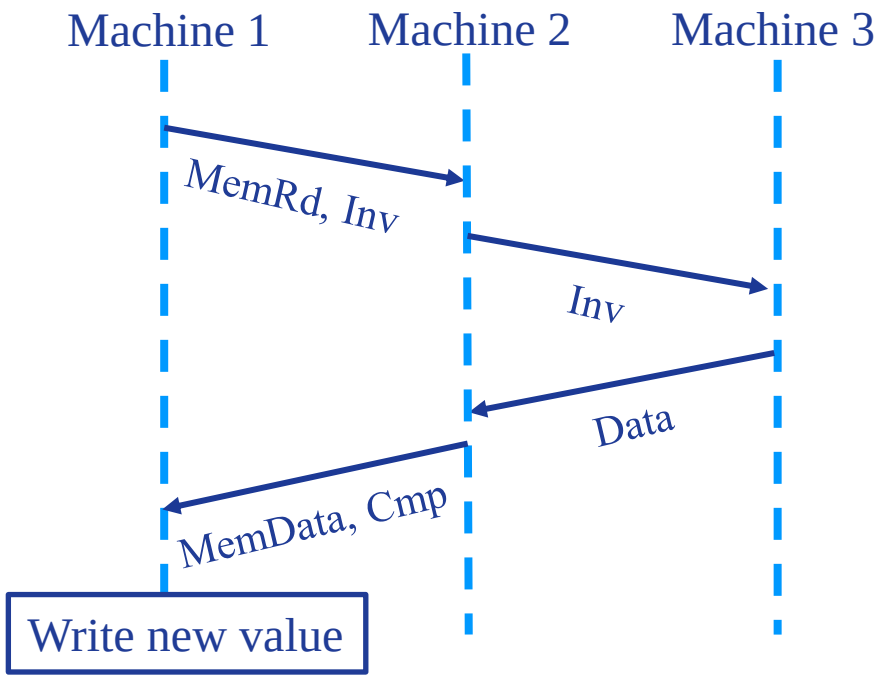
MSHR: only **12** in single Intel Skylake!



Extending Ligra to DSM: What causes slow down

❖ Insights for push mode

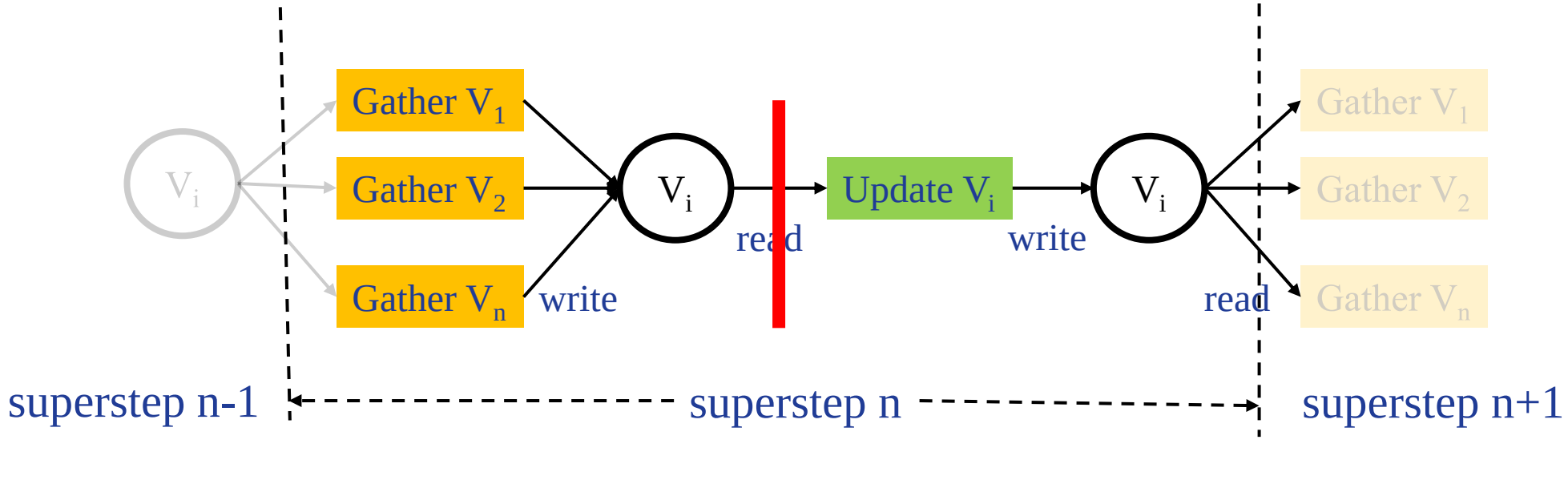
- Large amount of cache coherent related operations
  - Unbearable latency overhead ☹



Extending Ligra to DSM: What causes slow down

❖ Insights for push mode

- The serialized computing phases
  - Caused by data dependency
  - Unable to hide latency with computing ☹

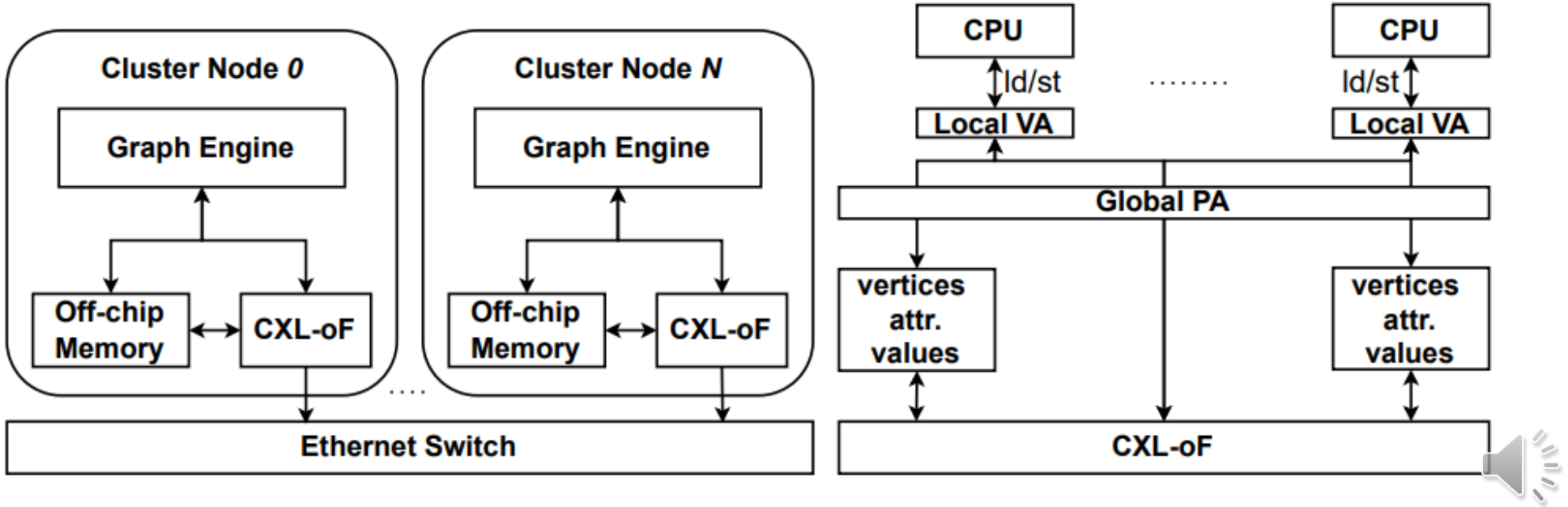


Outline

- Background
- Motivation
- GraCXL
- Evaluation
- Conclusion

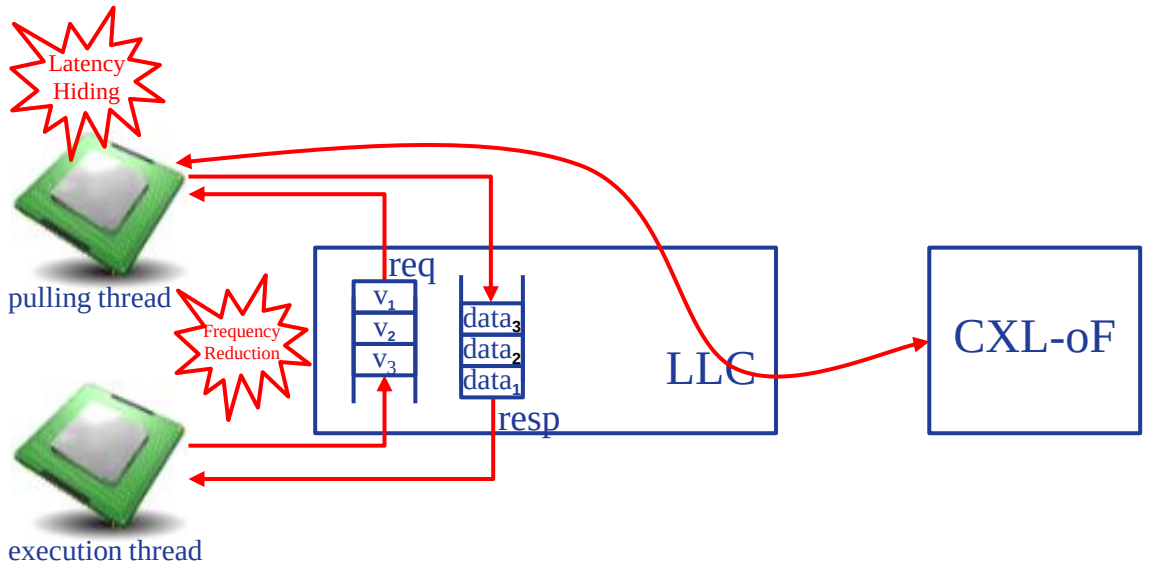
GraCXL Overview

- Each node exposes a portion of local memory to global address space
- Graph Engine (CPU or FPGA) accesses remote memory via CXL-oF
- CXL-oF accesses local memory bypassing Graph Engine



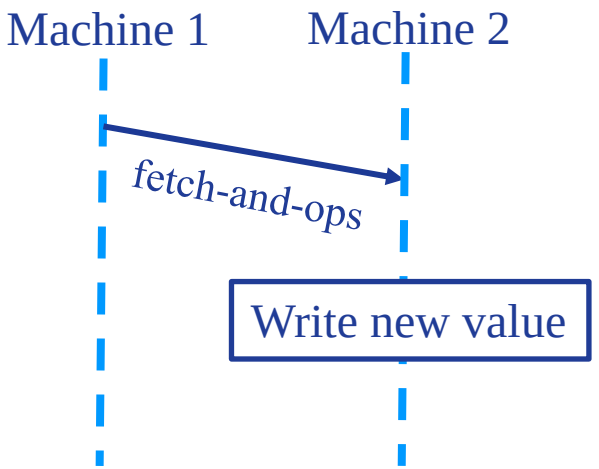
GraCXL: CPU platform

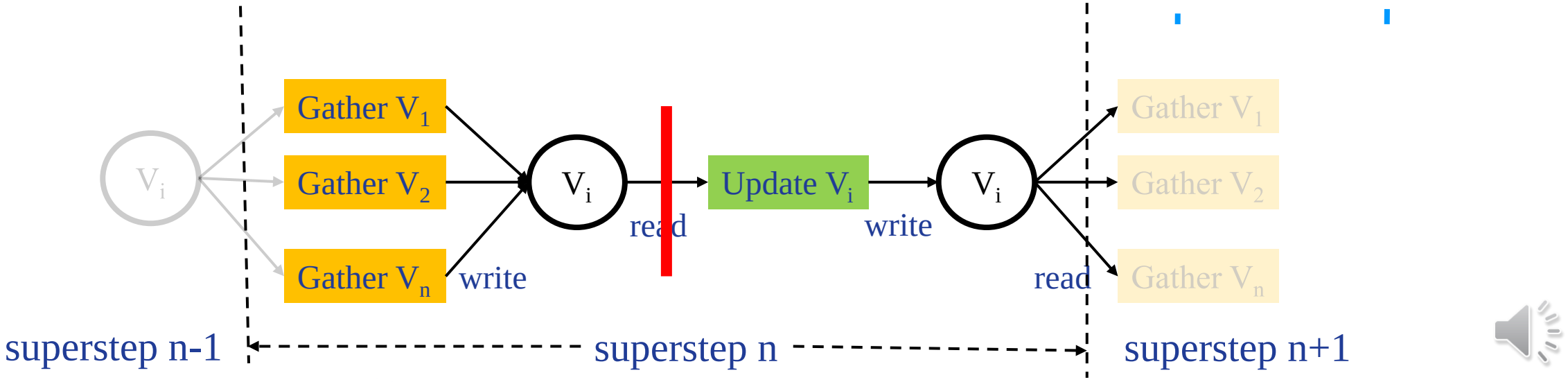
- Frequency Reduction
  - Iteratively traversing remote vertices
- Latency Hiding
  - Execution thread compute results
  - Pulling thread loads remote data



GraCXL: FPGA platform

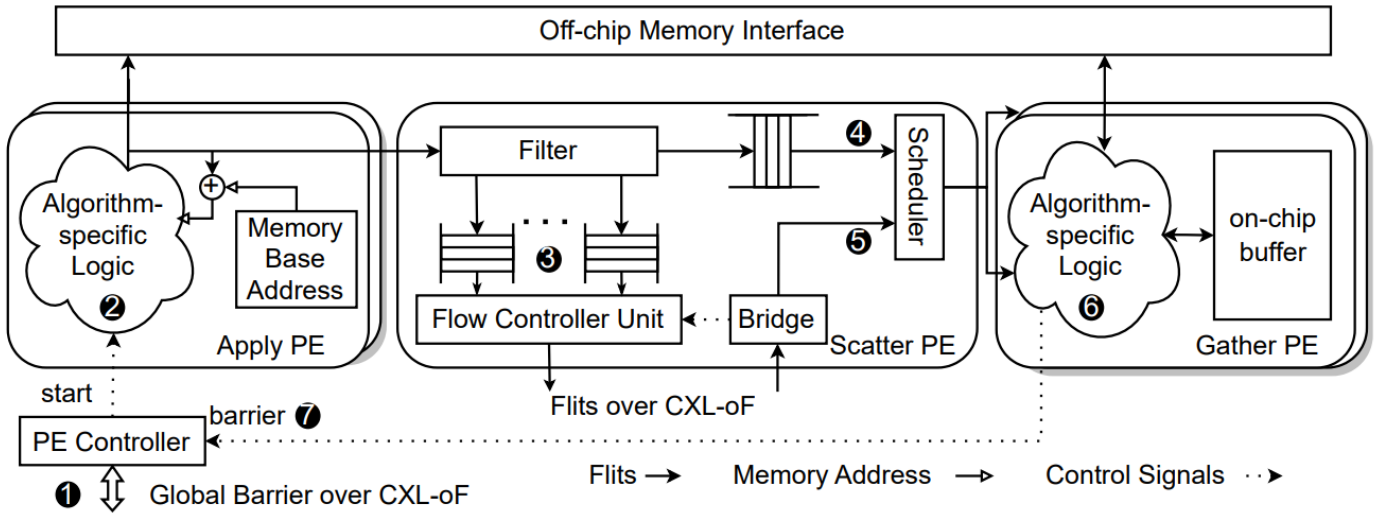
- Frequency Reduction
  - Extend CXL protocol with *fetch-and-ops*
- Latency Hiding
  - Adjusting sequence of execution phases

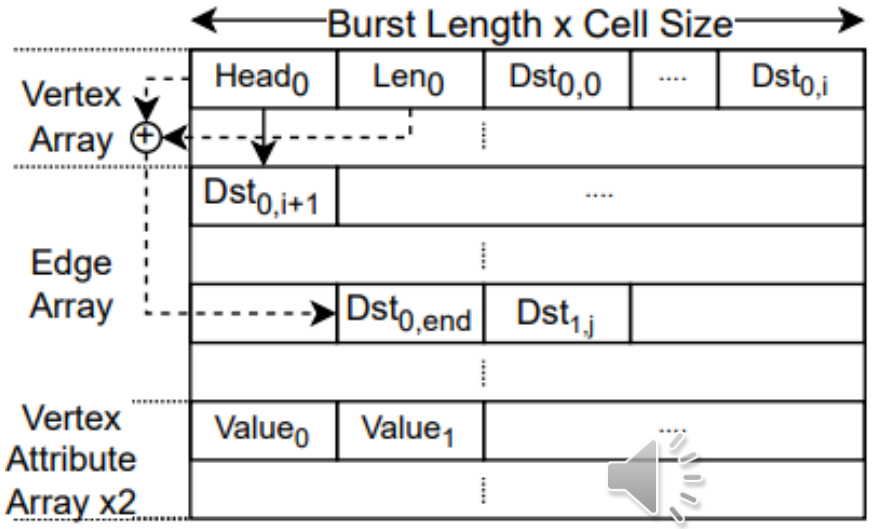




GraCXL: Accelerator Architecture

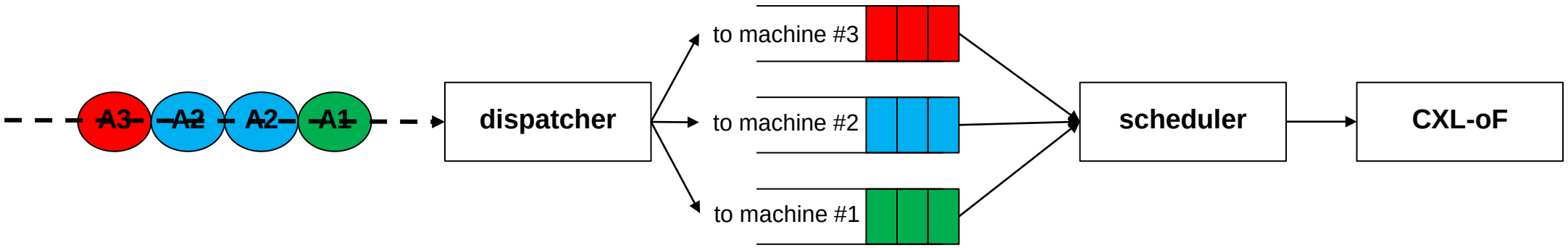
- Apply PEs (APEs): read attribute values and execute the apply kernel ❷
- Scatter PE (SPE): schedules the flits to remote ❸ or local ❹❺
- Gather PEs (GPEs): conduct the atomic operations ❻
- Graph representation: optimized CSR structure





GraCXL: Flow Control Unit Architecture

- Dispatch queue: equally share egress bandwidth of CXL-oF
- Scheduler: insert dynamic interval



Outline

- Background
- Motivation
- GraCXL
- Evaluation
- Conclusion

Implementation

❖ Prototype

- Four custom FPGA nodes based on Xilinx ZynqMP chips.
- Each node is attached to an 100Gbps Ethernet Switch

Custom FPGA Board with a Xilinx
Zynq UltraScale+ MPSoC (ZynqMP) Chip

Performance Evaluations

❖ Workloads

- Three real-world graphs
- Two scale-free synthetic graphs

❖ Metrics:

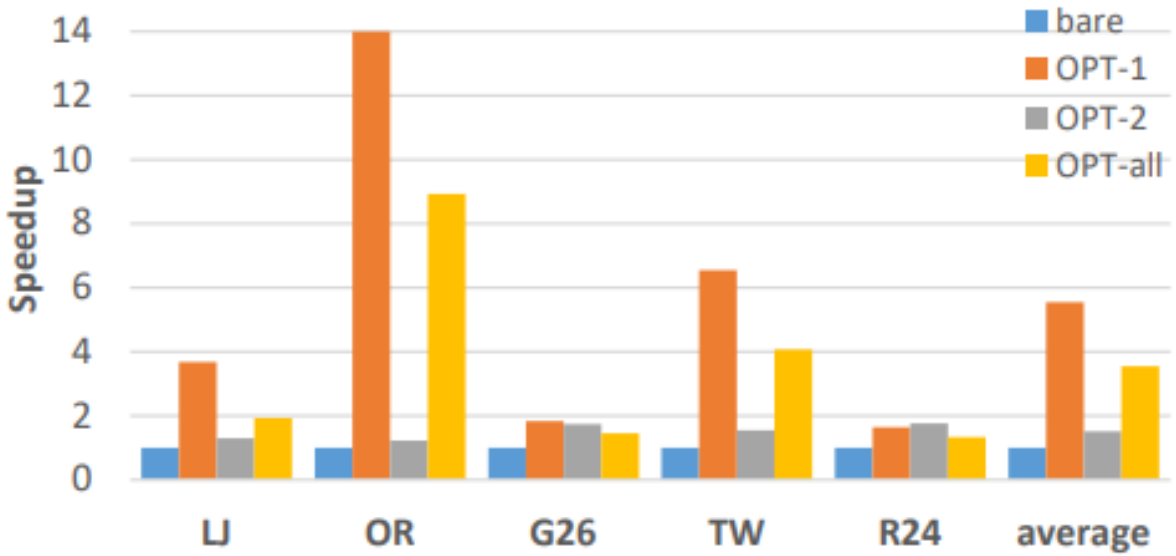
- Throughput: traversed edges per second (TEPS)

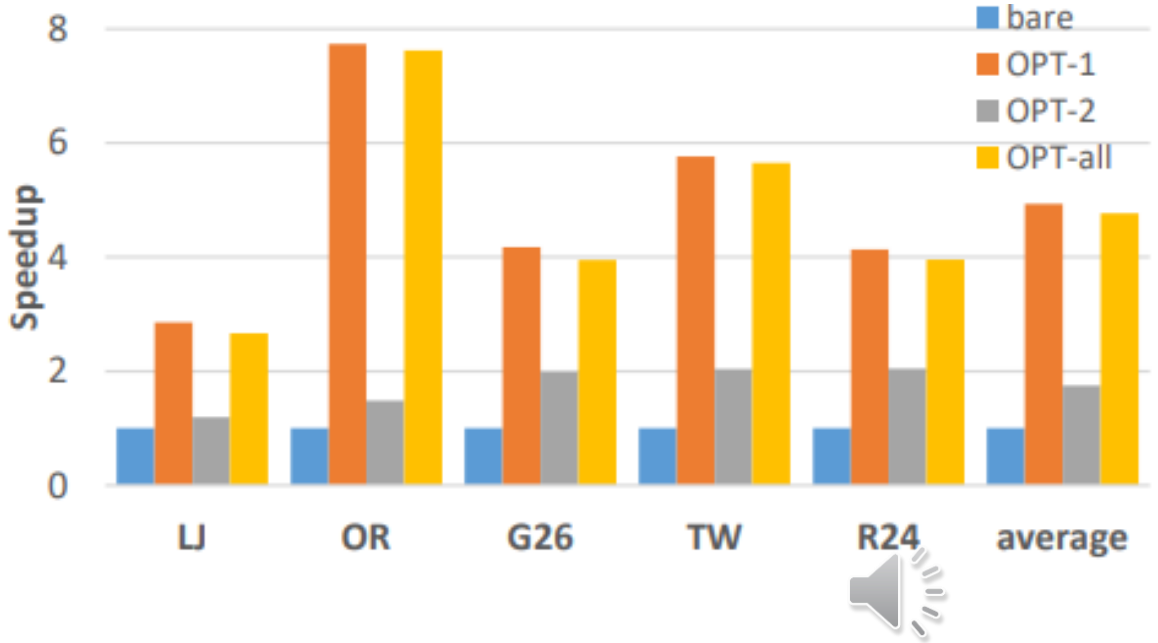
Workload	# Vertices	# Edges	Average Degree
Real-world Graphs [28]			
com-orkut (OR)	3.07 M	234 M	76.2
soc-LiveJournal1 (LJ)	4.84 M	69 M	14.3
twitter-rv (TW)	41.65 M	1468.37 M	35.3
Synthetic Graphs			
graph500-26 (G26)	64 M	1024 M	16
RMAT-24 (R24)	16 M	512 M	32

J. Leskovec and A. Krevl, "SNAP Datasets: Stanford large network dataset collection," http://snap.stanford.edu/data, Jun. 2014.
D. Chakrabarti and C. Faloutsos, "The (recursive matrix) graph generator," in Graph Mining. Springer, 2012, pp. 8

Performance Evaluations (CPU)

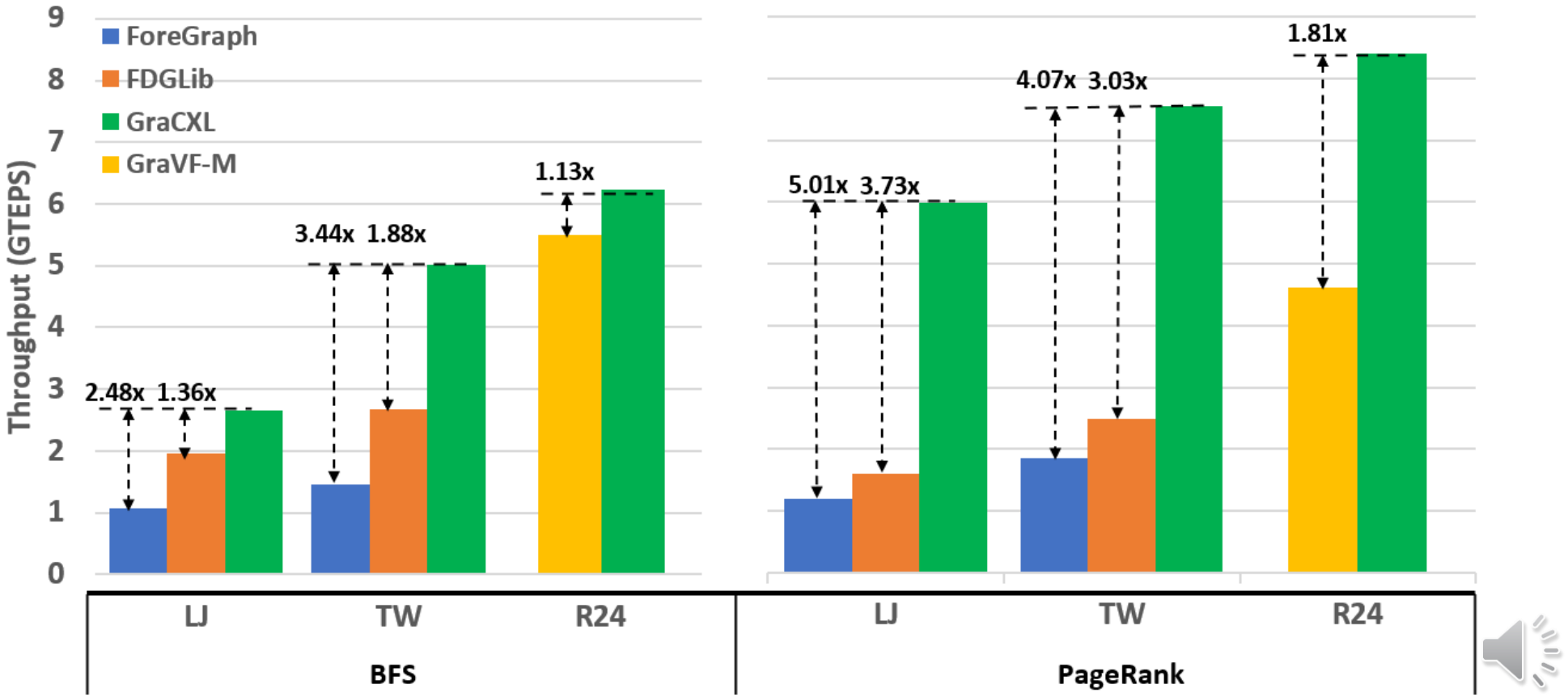
- Frequency reduction (OPT-1) provides 1.83x-14x throughput improvement
- Latency hiding (OPT-2) provides 1.19x-1.98x throughput improvement
  - Inter threads communication overhead ☹





Performance Evaluations (FPGA)

- GraCXL obtains 2.48x-5.01x progress in throughput against the state-of-the-art DSM system



Conclusion

- A series of new design paradigm are needed
- We design and implement DSM system atop CXL-oF
- GraCXL achieve throughput improvement on CPU and FPGA
  - latency hiding and frequency reduction
  - Other techniques?

Rethinking Design Paradigm of Graph Processing System with a CXL-like Memory Semantic Fabric

Xu Zhang, Yisong Chang, Tianyue Lu, Ke Zhang, Mingyu Chen

Thanks for Listening!

Contact me: zhangxu19s@ict.ac.cn

About me: https://zxhero.github.io/cv/