Intra-node MPI Collectives and the XHC framework for improved performance

George Katevenis
Foundation for Research and Technology – Hellas (FORTH)

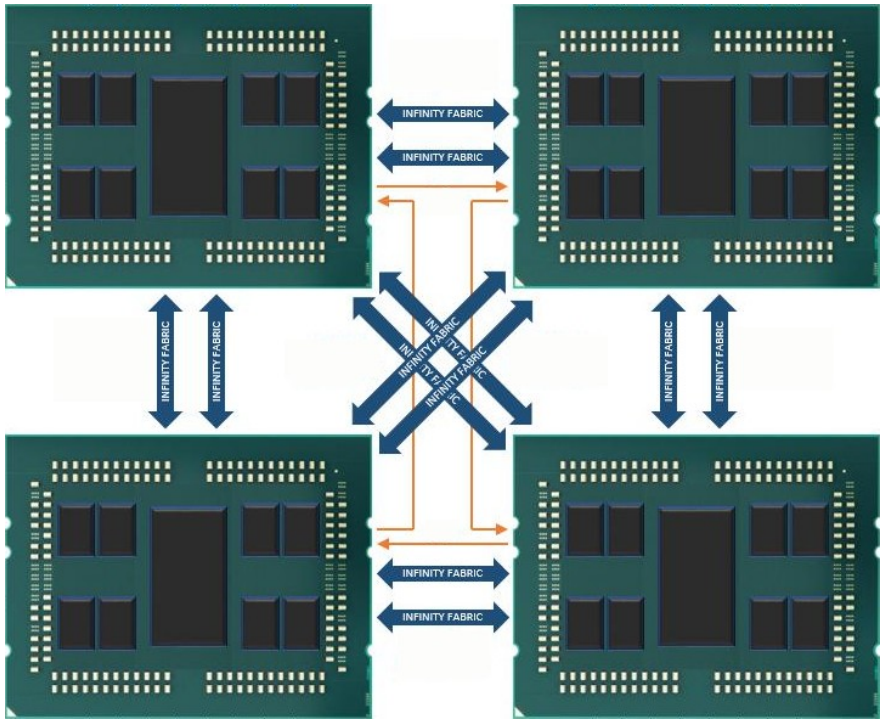
EUPEX

EuroHPC Joint Undertaking

This project has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 101033975. The JU receives support from the European Union's Horizon 2020 research and innovation programme and France, Germany, Italy, Greece, United Kingdom, Czech Republic, Croatia.

Problem statement

- Modern HPC nodes are complex
  - High core counts, elaborate topologies
  - Architectures, cache coherence schemes, interconnects, …, vary across systems
- Goal: Optimal MPI collective communication, across the board, for such nodes



AMD Epyc Rome system, 4 sockets with 64 cores across 8 chiplets each

Message Passing Interface (MPI)

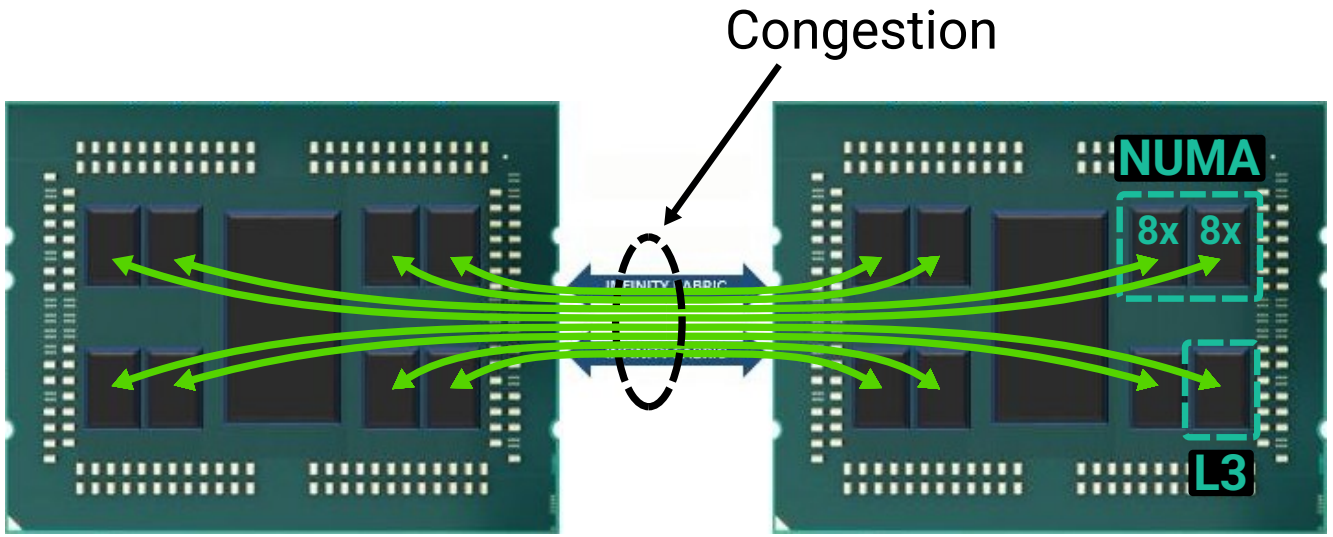
- Communication standard for parallel & distributed computing
  - Many implementations (OpenMPI, MPICH, …)
  - Widely relied upon in HPC
- **Collective communication** – involves all processes
  - **One-to-all** – Broadcast
  - **All-to-one** – Reduce, Gather
  - **All-to-all** – Allreduce, Allgather

XPMEM Hierarchical Collectives (XHC)

- Optimized intra-node collectives in OpenMPI
  - Part of upstream OpenMPI v6 (upcoming)
  - Specialization in node-level performance
- Hierarchical & topology aware algorithms
- Data copies with XPMEM (Cross-Partition Memory)
  - For large messages – very low overhead IPC
- Combines existing techniques, efficiently, with new features, insights, optimizations

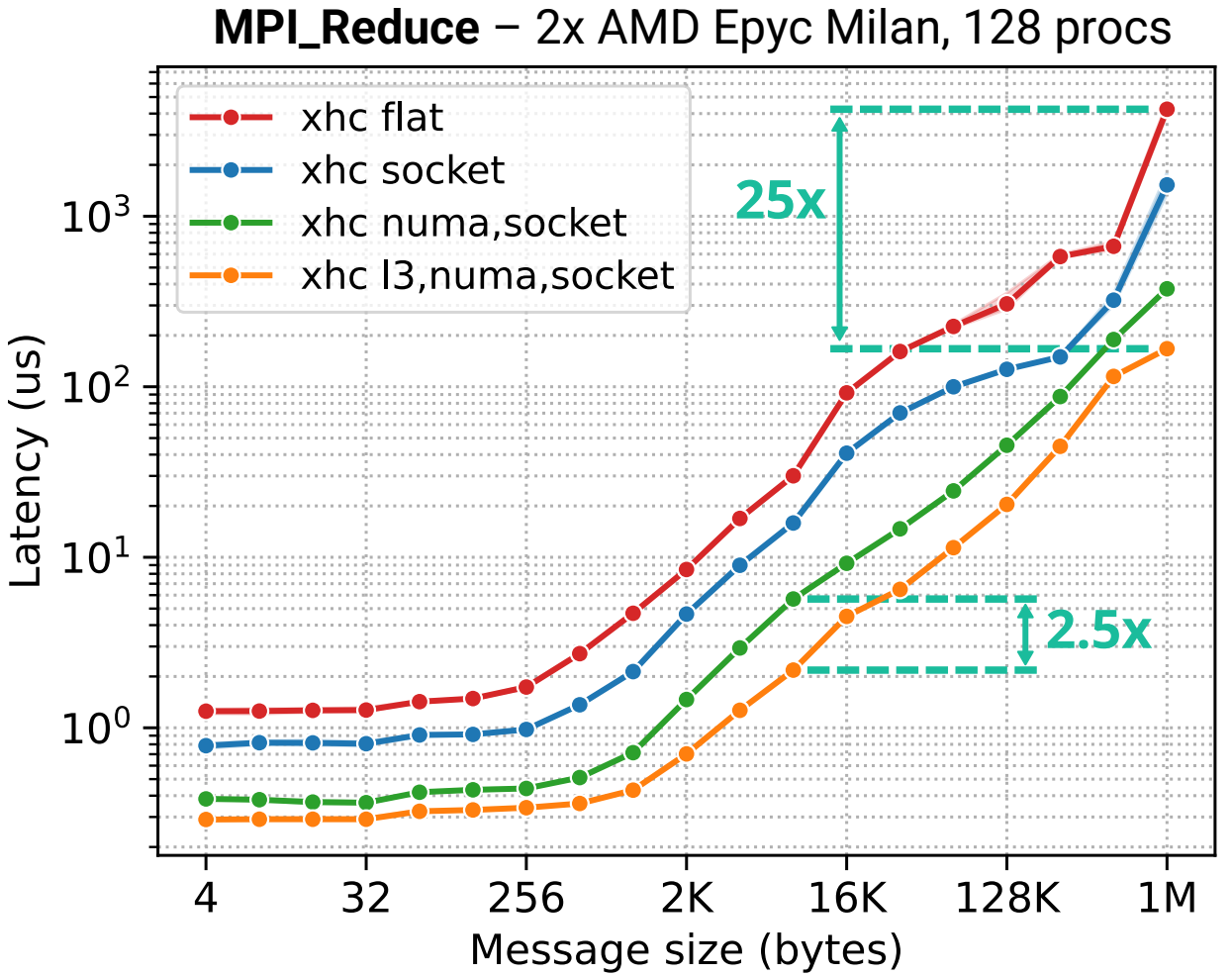
Hierarchy

- Concurrent communication across *distance* results in congestion
- **Partition** into smaller groups, perform part of operation locally
- Topology discovery through Hwloc
  - CPU socket, NUMA node, caches
- User-configurable sensitivity to topological features, *n*-level



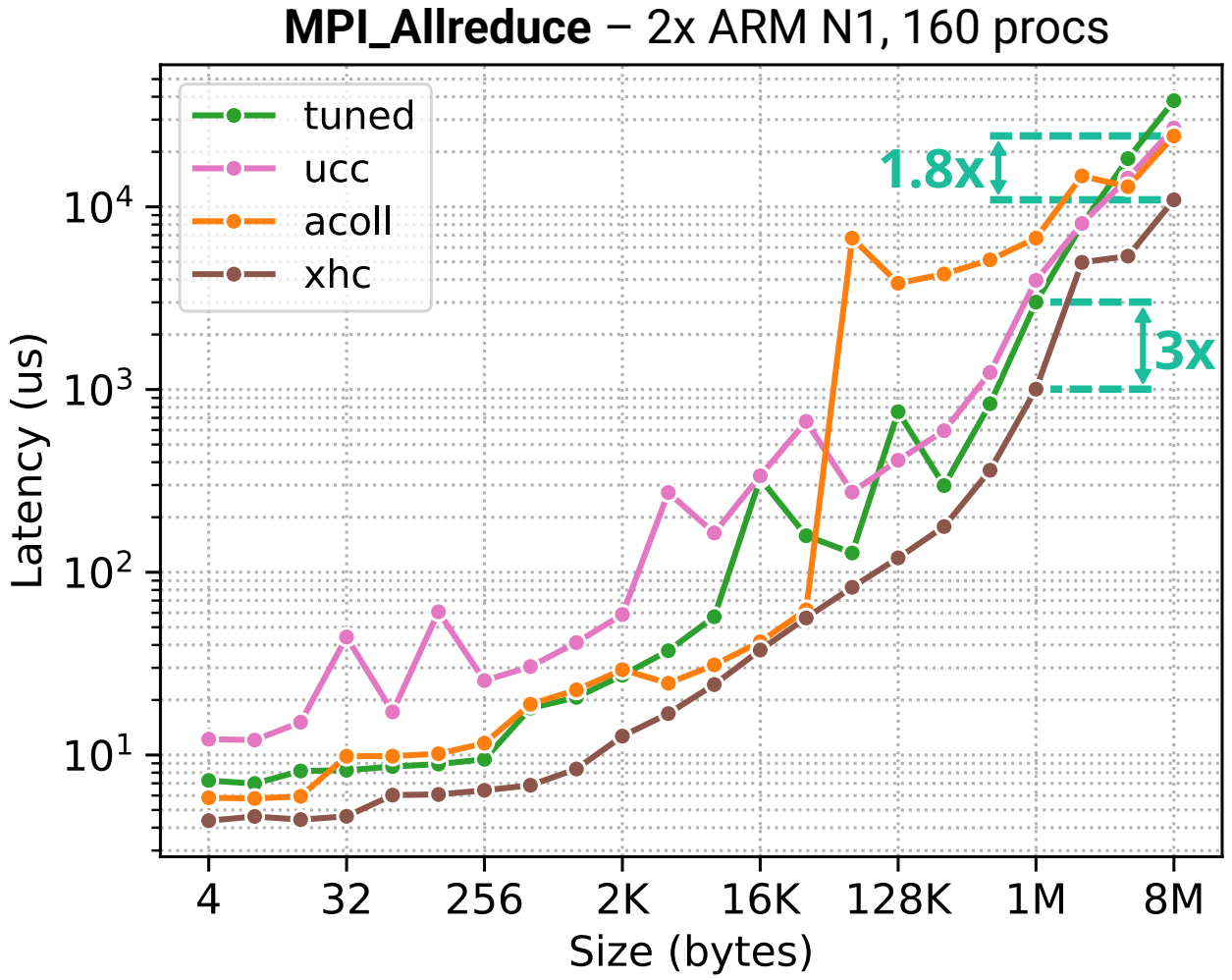
Hierarchy – Performance

- Hierarchical algorithms very useful
  - Especially on ‘fat’ nodes & elaborate collectives
- Beneficial to consider multiple/all topological features
  - *Though not always!...*



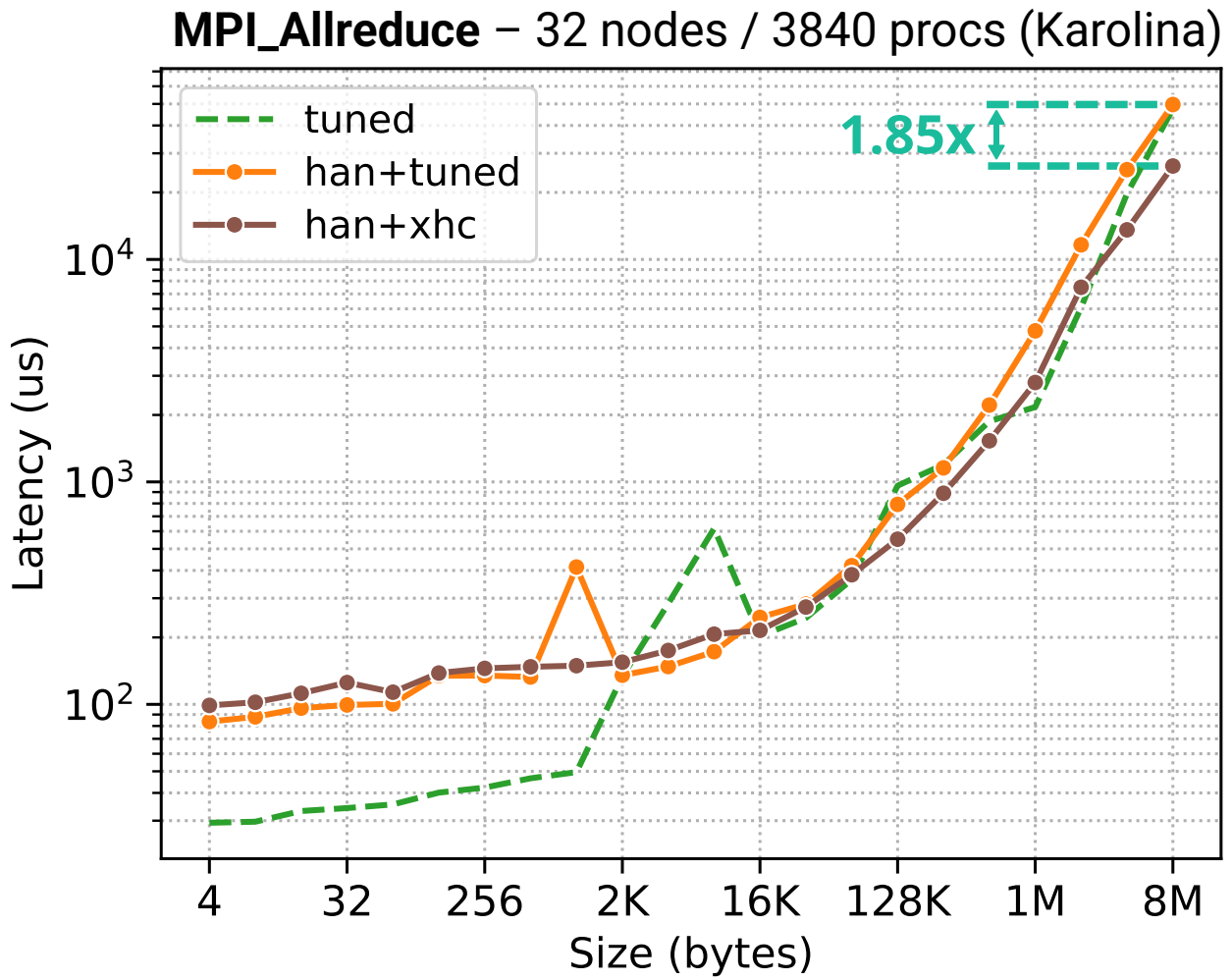
Intra-node performance

- Lowest latency across all sizes
- Consistently good performance



Inter-node: HAN+XHC

- Multi-node via HAN (Hierarchical AutotuNed)
  - Default since OpenMPI v5
  - XHC in intra-node phase
- Preliminary results →
- Work on HAN(+XHC) warranted/planned



EUPEX

Thank you!

Publications

1. A framework for **hierarchical** single-copy MPI **collectives** on **multicore** nodes
   CLUSTER ‘22
   G. Katevenis, M. Ploumidis, M. Marazakis
   DOI 10.1109/CLUSTER51413.2022.00024

2. Impact of **Cache Coherence** on the Performance of **Shared-Memory** based MPI Primitives: A Case Study for **Broadcast** on Intel **Xeon Scalable** Processors
   ICPP ‘23
   G. Katevenis, M. Ploumidis, M. Marazakis
   DOI 10.1145/3605573.3605616