Performance Analysis Report: MPI Latency and Bandwidth Regression Tests

Submitted by :

Paul Timileyin Aromolaran – paul.aromolaran.001@student.uni.lu

Jabin Tasnim Khan Urmy – jabin.urmy.001@student.uni.lu

Ngu Frerashyno Ndah – frerashyno.ngu.001@student.uni.lu

Abstract & Introduction

Abstract

- This report details a ReFrame-based regression testing suite for MPI intranode and internode communication performance on ULHPC clusters (Aion, Iris).
- Utilizes OSU MicroBenchmarks (osu_latency, osu_bw) for point-to-point measurements.
- Evaluated four process placement scenarios and three binary sourcing methods (Source, EasyBuild, EESSI).
- Presents latest performance results, including several instances where performance **deviated from expectations**, and critically examines the baseline reference values used in ReFrame for future regression testing.

Introduction

- **Goal:** Develop a robust regression testing suite to monitor MPI latency and bandwidth, detecting performance variations due to software/hardware changes or configuration drift.
- **Importance:** Consistent MPI performance is critical for HPC application scalability and efficiency.

Benchmarks & Core Aim

Benchmarks Used

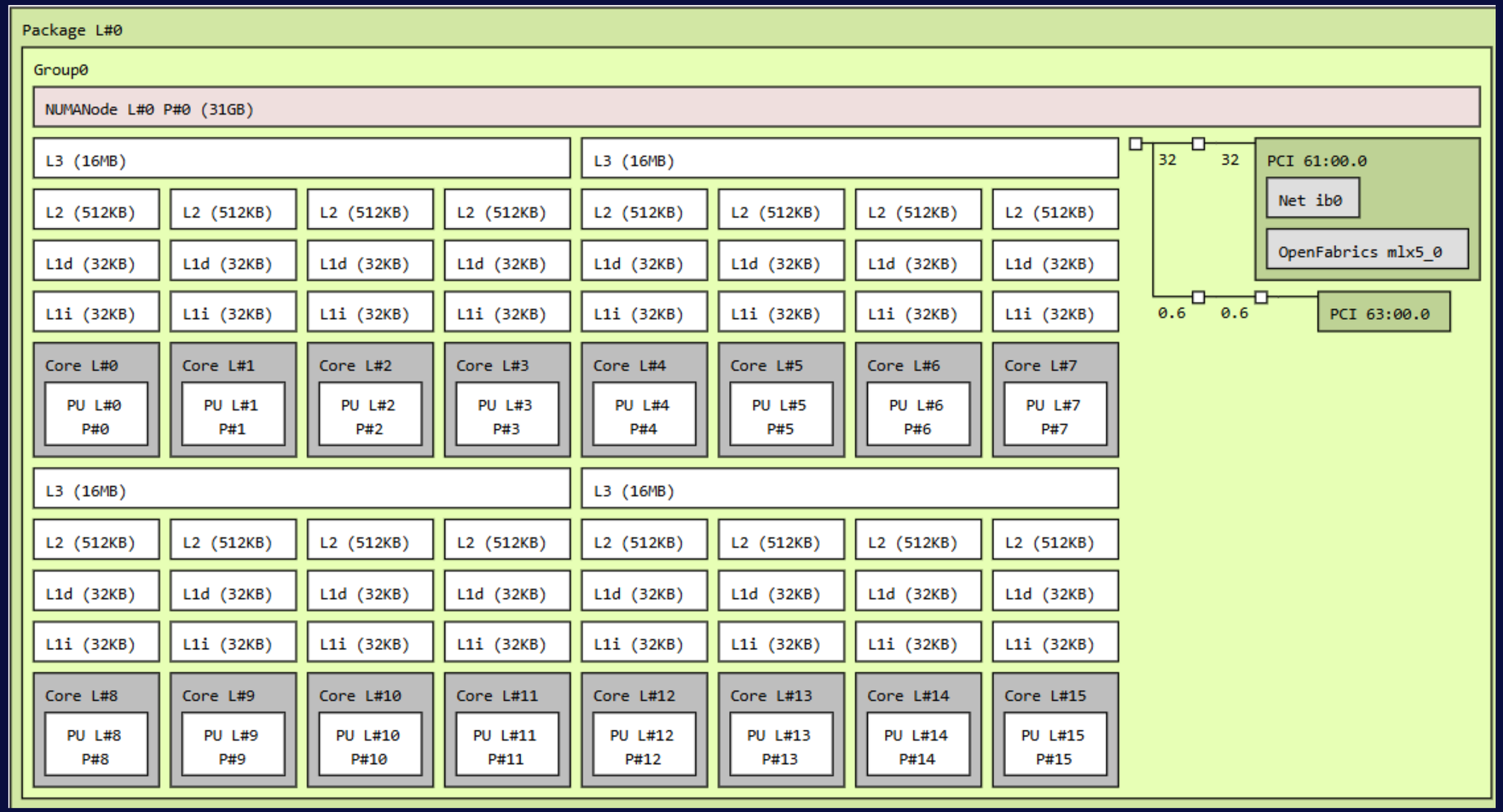
- osu_latency: Measures small message latency
- osu_bw: Measures large message bandwidth

Core Aim

Compare current data to references to spot regressions or improvements.

ReFrame scripts contain reference values; measured performance is compared, **determining whether performance aligned with expectations or warranted further review.** The results.txt file, reflecting latest runs, now indicates **various performance outcomes**.

System Architecture (AION)



Methodology - Test Design

Target Systems

- **Aion:** AMD EPYC based cluster.
- **Iris:** Intel Skylake based cluster (regular CPU nodes).
  - *Note: SameSocketDifferentNuma scenario is not applicable/tested on Iris.*

Benchmarks & Messages

- **osu_latency:**
  - Message Size: 8192 bytes (typical small message latency).
  - Iterations: 100 warm-up, 1000 measurement.
- **osu_bw:**
  - Message Size: 1,048,576 bytes (1MB) (peak bandwidth scenario).
  - Iterations: 100 warm-up, 1000 measurement.

Process Placement

Process Placement Scenarios (2 MPI processes):

1. **Same NUMA Node (SameNumaNode):** Processes on cores within the same NUMA node, memory local. (Best-case intra-node)
2. **Same Socket, Different NUMA Nodes (SameSocketDifferentNuma):** (Aion-specific) Processes on same socket, different NUMA nodes.
3. **Different Sockets (DifferentSockets):** Processes on same compute node, different physical sockets.
4. **Different Nodes (DifferentNodes):** Processes on different physical compute nodes. (Inter-node)

Methodology - Compilation & ReFrame Implementation

OSU MicroBenchmarks (Version 7.2)
Sourcing:

1. Source Compilation (SOURCE): Compiled directly from source using foss/2023b toolchain.
2. EasyBuild Compilation (EASYBUILD): Compiled using an EasyBuild recipe with foss/2023b toolchain.
3. EESSI Binaries (EESSI): Pre-compiled binaries from EESSI software stack (OSU-Micro-Benchmarks/7.2-gompi-2023b).

ReFrame Implementation:

- Regression tests implemented using the ReFrame framework.
- Specific SLURM options and OpenMPI MCA parameters enforced process placement.
- Performance extracted by parsing benchmark output.
- **Crucially:** ReFrame scripts contain reference values. Measured performance is compared against these, **determining if the measurement is within the expected range or if it signals a deviation.**
    - The results.txt file, reflecting the latest runs, shows a **variety of performance outcomes relative to these references**, forming the basis of this analysis.

Results Overview: NUMA Performance Benchmarks

Results - Overview & General Trends

General Performance Trends Observed:

Bandwidth (Higher is Better):

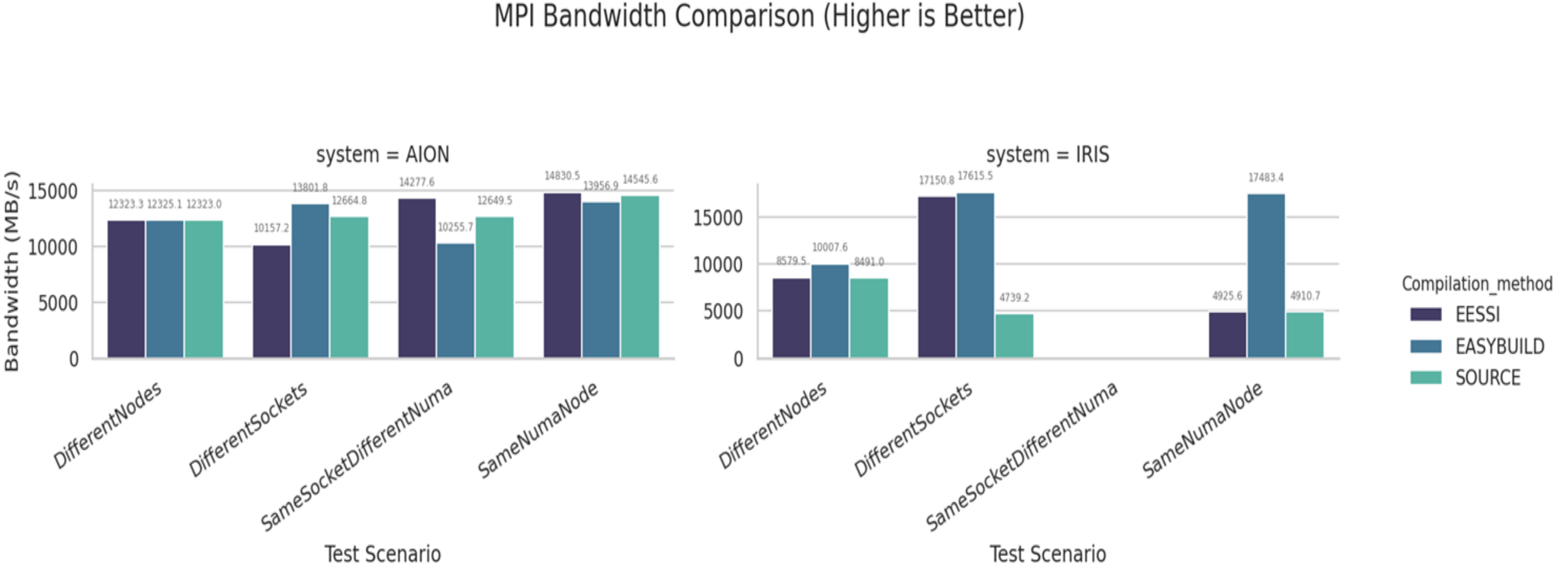
- Typically highest for SameNumaNode (with exceptions on Iris).
- Decreases for DifferentSockets, SameSocketDifferentNuma (Aion), and lowest for DifferentNodes.

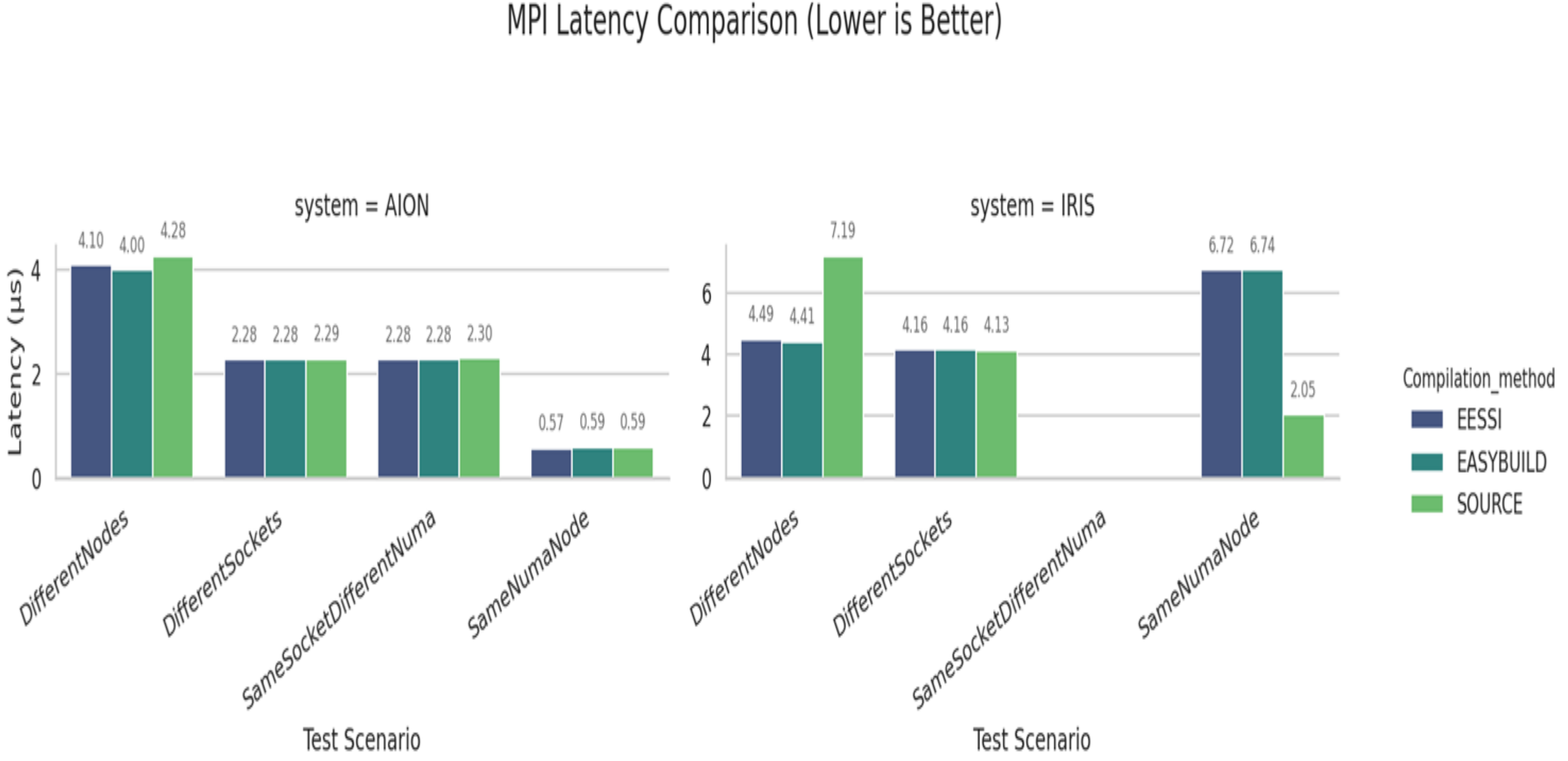
Latency (Lower is Better):

- Generally lowest for SameNumaNode.
  - *Note: These often **indicated performance better than very strict reference values, thus highlighting them for review.***
- Increases as communication crosses NUMA, socket, and inter-node boundaries.

Latest results.txt shows a variety of performance outcomes relative to reference values.

- This presentation will analyze observed values and critically discuss the appropriateness of the underlying ReFrame reference values that led to these outcomes.


MPI Bandwidth Comparison (Higher is Better)
system = AION
system = IRIS
Bandwidth (MB/s)
15000
10000
5000
0
12323.3
12325.1
12323.0
10157.2
13801.8
12664.8
14277.6
10255.7
12649.5
14830.5
13956.9
14545.6
8579.5
10007.6
8491.0
17150.8
17615.5
4739.2
4925.6
17483.4
4910.7
DifferentNodes
DifferentSockets
SameSocketDifferentNuma
SameNumaNode
Test Scenario
Compilation_method
EESSI
EASYBUILD
SOURCE

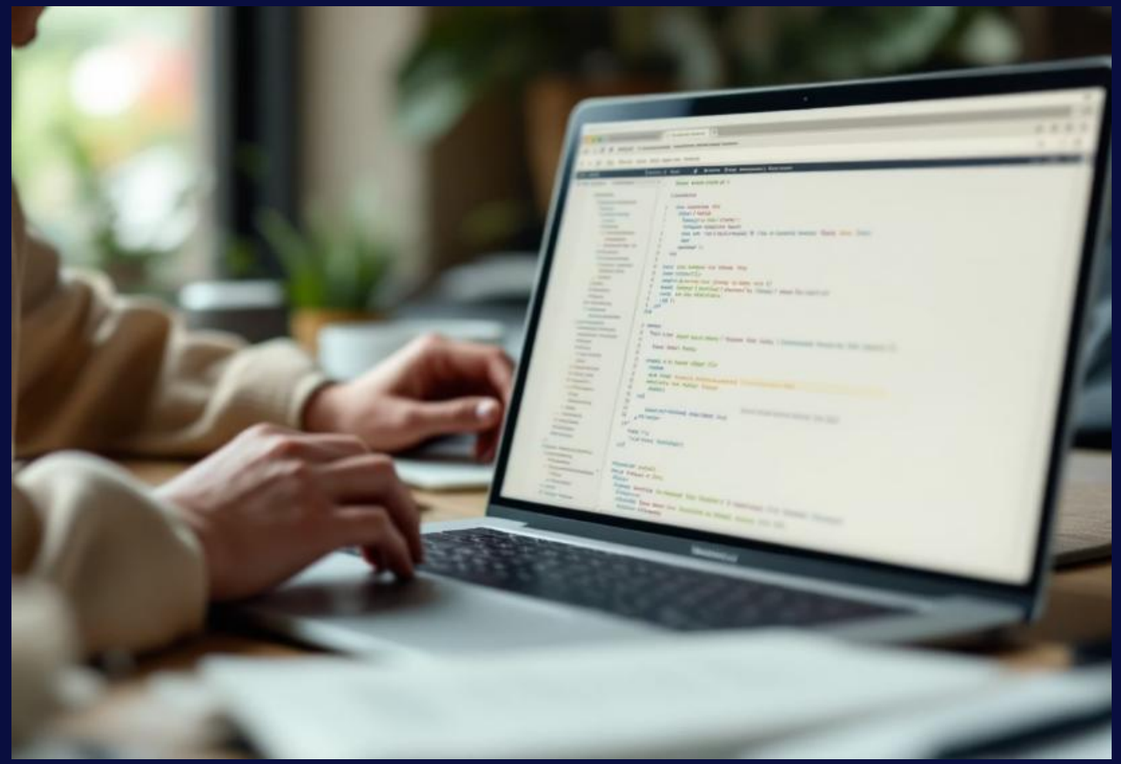

MPI Latency Comparison (Lower is Better)
system = AION
system = IRIS
Latency (µs)
0
2
4
6
4.10
4.00
4.28
2.28
2.28
2.29
2.28
2.28
2.30
0.57
0.59
0.59
4.49
4.41
7.19
4.16
4.16
4.13
6.72
6.74
2.05
DifferentNodes
DifferentSockets
SameSocketDifferentNuma
SameNumaNode
Test Scenario
Compilation_method
EESSI
EASYBUILD
SOURCE

Reproducibility of Regression Testing

All ReFrame tests, configuration files, EasyBuild recipe, and report are stored in a Git repository:

https://github.com/PaulAroo/Regression-testing

Instructions for cloning, loading modules, and running tests are detailed in the README.md.



```
#!/bin/bash
#SBATCH --job-name="rfm_OsuDifferentNodes_bf0b4b03"
#SBATCH --ntasks=2
#SBATCH --ntasks-per-node=1
#SBATCH --cpus-per-task=1
#SBATCH --output=rfm_job.out
#SBATCH --error=rfm_job.err
#SBATCH --exclusive
#SBATCH --nodes=2
#SBATCH --partition=batch
#SBATCH --qos=normal
#SBATCH --time=0-00:10:00
module load env/testing/2023b
module load toolchain/foss/2023b
module load tools/EasyBuild
export OMPI_MCA_hwloc_base_report_bindings=1
srun --cpus-per-task=1 --nodes=2 /mnt/aiongpfs/users/paromolaran/project_playground/stage/
    aion/batch/foss-2023b/OsuBuildSource_0de34a51/osu-micro-benchmarks-7.2/c/mpi/pt2pt/
    standard/osu_bw -m 1048576:1048576 -x 100 -i 1000
```

```
cpu-bind=MASK - aion-0079, task··0··0 [94304]: mask 0×1 set
cpu-bind=MASK - aion-0079, task··1··1 [94305]: mask 0×2 set
[aion-0079:94304] MCW rank 0 bound to socket 0[core 0[hwt 0]]:
    [B/./.]
    [./.]
[aion-0079:94304] MCW rank 0 bound to socket 0[core 0[hwt 0]]:
    [B/./.]
    [./.]
[aion-0079:94305] MCW rank 1 bound to socket 0[core 1[hwt 0]]: [./
    B/./.]
    [./.]
[aion-0079:94305] MCW rank 1 bound to socket 0[core 1[hwt 0]]: [./
    B/./.]
    [./.]
```

THANK YOU