# Communication Performance Assessment of Sapphire Rapids Architecture

Jackson Wesley; Mentor: David DeBonis | HPC-ENV


## Abstract

The Sapphire Rapids architecture from Intel is an approach to microarchitecture design using chiplets, discrete and modular chips assembled into a package. Because communication is the dominating cost in parallel applications, it is valuable to understand the on node and off node bandwidth of these microarchitectures. In this study we present the bandwidth trends over a range of message sizes between chiplets on Rocinante. By observing differences in bandwidth performance at certain packet sizes as well as comparing between different types of message communication, we hope to allow for communication improvements in HPC codes of interest.


## Specs

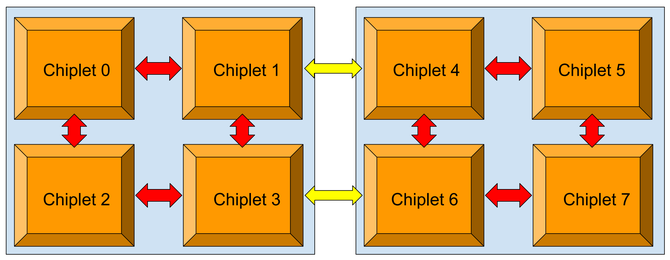
- Rocinante has two partitions:
  - Standard – Intel Xeon Platinum 8479 with 256 GB DDR memory per node.
  - HBM – Intel Xeon CPU Max 9480 with 128 GB HBM memory per node.
- Uses Cray Slingshot 11
- 56 cores per socket across the 4 chiplets. 2 sockets per node

## Bandwidth Tests

The tests used were from the Fabtests library, a testing library that utilizes Libfabric to implement various tests. These benchmarks are structured to be similar to OSU MPI benchmarks.

The following bandwidth benchmarks are used in this project:

- fi_msg_bw: Message transfers, connected endpoints
- fi_rma_bw: RMA read and write
- fi_rdm_tagged_bw: Tagged message, reliable-datagram

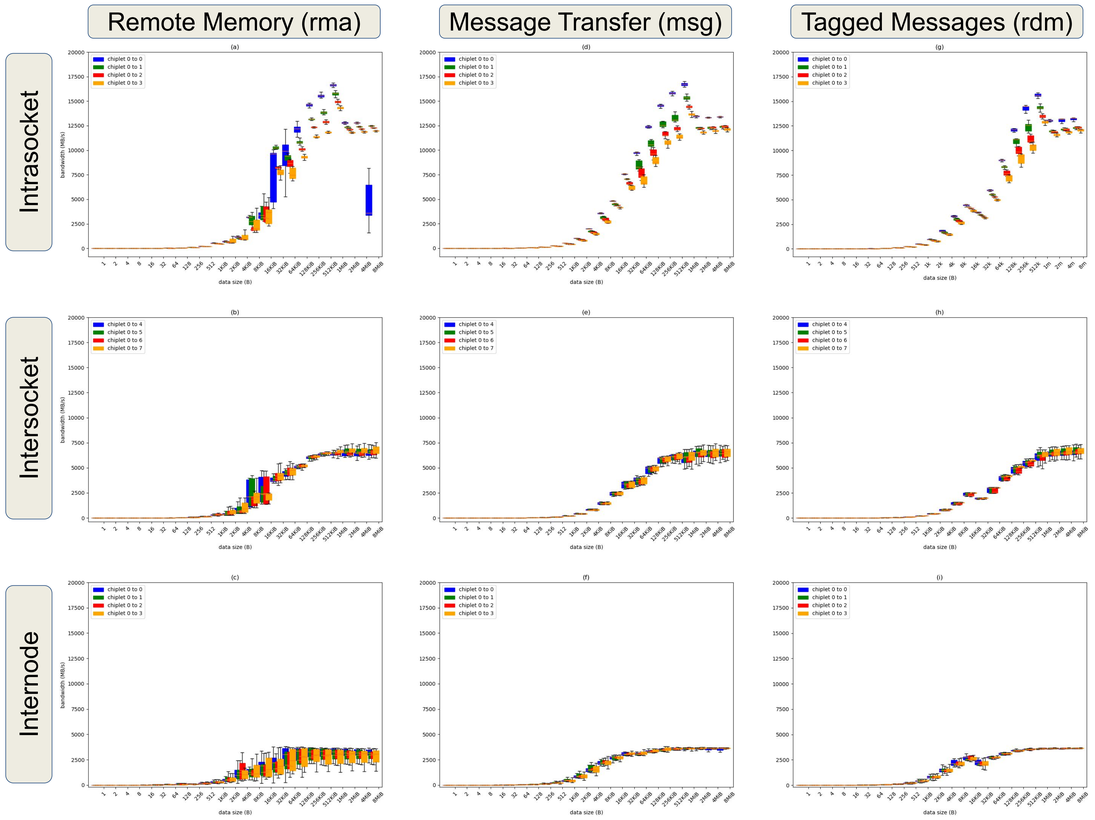



## Test Methodology

- Tests on Rocinante were performed over both DDR and HBM partitions with notable comparisons here being made between tests performed on the DDR partition of Rocinante.
- Bandwidth measurements are averaged over 10 runs per node after an initial warmup of 1024 iterations.
- Tests are performed on 5 nodes and trends are averaged.
- Chiplet communication is specified by pinning each end to a cpu.
  - The first CPU of each chiplet is used, IE chiplet 0 uses CPU 00, except for when communicating to itself, which chiplet 0 pins CPU 01 as the other end of communication.

## Figures

- For this set of tests, we display the communication bandwidth from host chiplet 0 to client chiplets.
- Boxplot shows the average (lines within box), quartiles (boxes), and min & max (whiskers).




## Observations

- The tail of bandwidth tests observed display that after data sizes of 1MiB, single core bandwidth lowers for intrasocket tests (figures a, d, g). This can be explained by some limit in the page size or buffer size when transferring data is reached.
- Intrasocket bandwidth is significantly higher than intersocket (figures b, e, h), which is an expected result. Transmitting data between chiplets on the same socket is a closer and more direct pathway than sending off of the socket.
- Internode bandwidth (figures c, f, i) is lower than intersocket bandwidth, but with less variation.
- RDM tagged messages have lower bandwidth at 16KiB than before or after that data size, then resuming similar trends to other test variations. Possibly due to an eager to rendezvous logic in libfabric communication.
- Bandwidth appears less predictable when measuring single core bandwidth over rma intrasocket. This is specifically when chiplet 0 tries to transmit over rma on chiplet 0 and to chiplet 1.

## Outside Observations

- Single core HBM bandwidth tests (not shown) were overall lower than DDR partition.
- MPI bandwidth tests have been run and data collected using the OSU Micro Benchmarks. Though not compared here, there are similarities in the trends when compared to Fabtests, as well as interesting differences. MPI bandwidth has shown to be significantly higher, which led to some questions of settings used in Fabtests.
- This study is occurring in parallel with another ongoing study at Sandia National Laboratories in which the same tests are being conducted on similar architectures.

## Future work

- Measuring the performance of message sending off-socket through the network against through memory.
- Performance modeling to find overhead costs incurred by Cray MPICH that relies on Libfabric to work.
- Investigate the effects of overlapping communication and its influence on performance.
- When stable, utilization of these benchmarks as tests within Pavilion system testing would be useful for finding possible bandwidth or latency outliers.

Los Alamos NATIONAL LABORATORY
NNSA National Nuclear Security Administration
Managed by Triad National Security, LLC, for the U.S. Department of Energy's NNSA.
LA-UR-24-28301