NEKRS: OVERVIEW

- nekRS solves the governing equations using a spectral element discretization
- nekRS employs high-order spectral elements in which the solution, data, and test functions are represented as locally structured Nth-order tensor product Lagrange polynomials on a set of E globally unstructured curvilinear hexahedral brick elements.
- Within each element, the Gauss–Lobatto–Legendre (GLL) quadrature points are set as the nodes of the Lagrange basis polynomials.
- For polynomial order N, there are $(N+1)^3$ GLL points in an element, leading to $E(N+1)^3$ total GLL points in the mesh.
- nekRS, like many other CFD codes, uses domain decomposition to solve very large meshes on distributed systems with MPI (each sub-domain is assigned to a distinct MPI rank)
- The size of the mesh, E, the polynomial order, N, and the number of MPI ranks, M, determine the computational load on each MPI rank (approx. $E(N+1)^3 /M$)

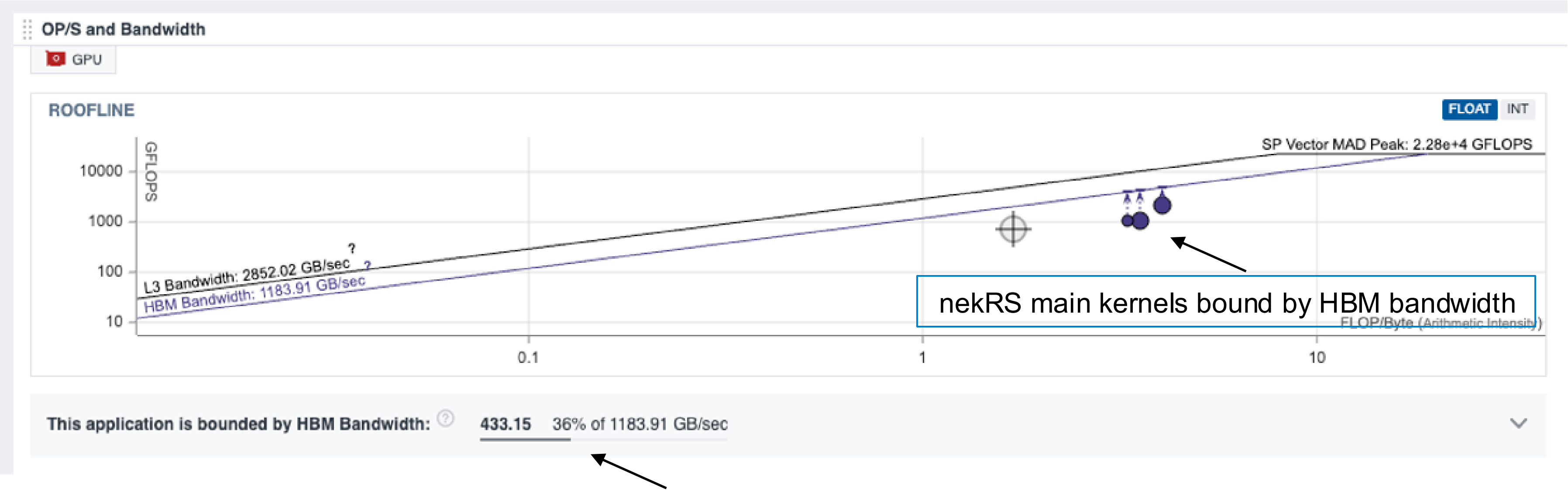
Detailed reference: https://arxiv.org/pdf/2104.05829

NEKRS: OVERVIEW

- Common polynomial order N for science runs:
  - —7 is most common (512 GLL points per element, a power of 2)
  - —5, 9 also used
- Common loading for good scaling and GPU utilization:
  - —4-5x10^6 GLL points per MPI rank for both Frontier and Aurora
  - —Approx. 8-10x10^3 mesh elements per MPI rank

ROOFLINE (4M GLL POINTS PER RANK, N=7)

- nekRS roofline collected with Intel Advisor on single Aurora PVC tile
  - nekRS main kernels are **HMB bandwidth bound**
  - nekRS application as a whole is also **HBM bandwidth bound**


OP/S and Bandwidth
GPU
ROOFLINE
FLOAT
INT
GFLOPS
10000
1000
100
10
0.1
1
10
SP Vector MAD Peak: 2.28e+4 GFLOPS
L3 Bandwidth: 2852.02 GB/sec
HBM Bandwidth: 1183.91 GB/sec
FLOP/Byte (Arithmetic Intensity)

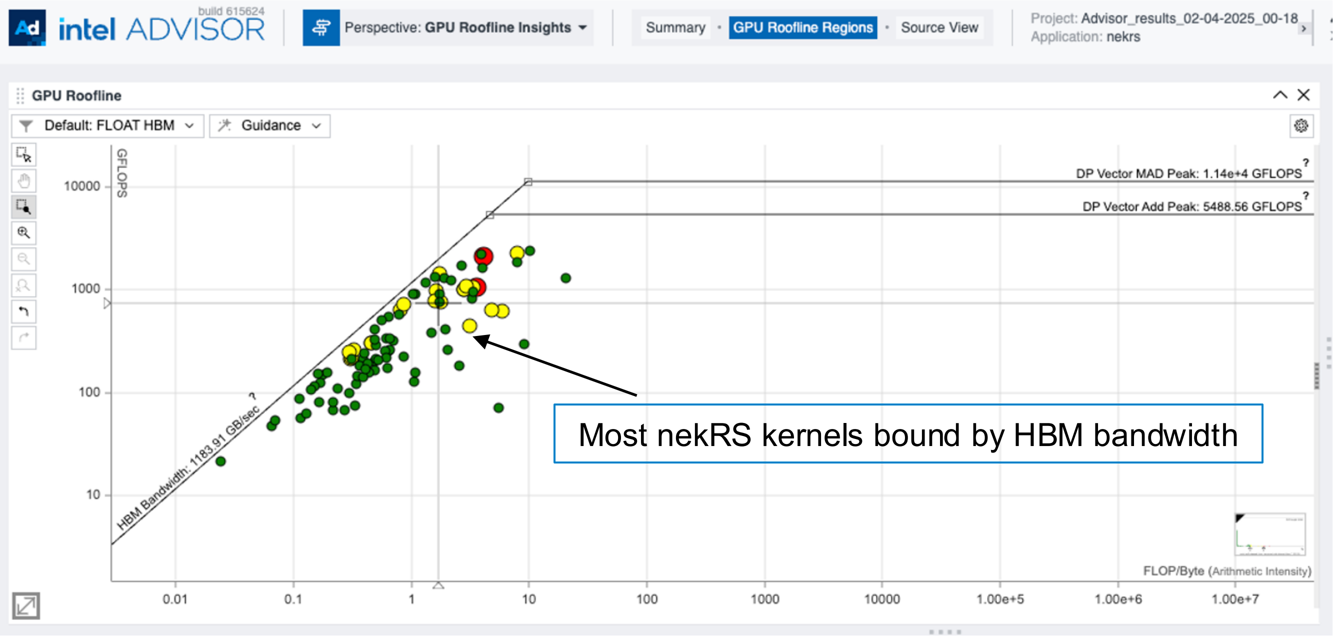
This application is bounded by HBM Bandwidth: 433.15 36% of 1183.91 GB/sec

nekRS main kernels bound by HBM bandwidth

nekRS application bound by HBM bandwidth

ROOFLINE (4M GLL POINTS PER RANK, N=7)

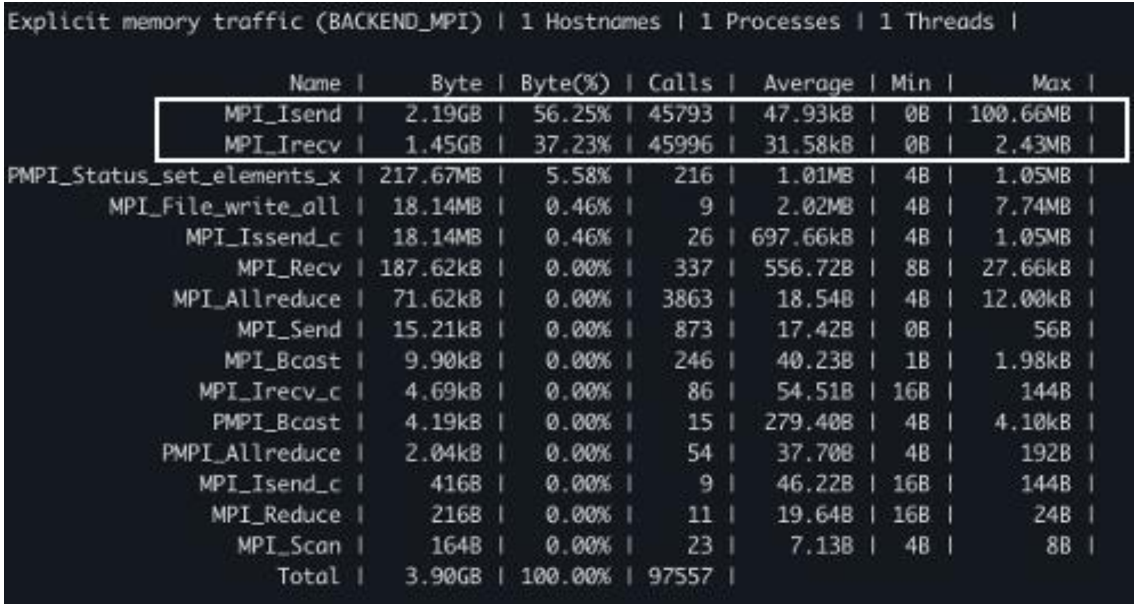
- nekRS roofline collected with Intel Advisor on single Aurora PVC tile
  - Most nekRS kernels are **HBM bandwidth bound**
  - Dependency on loading (GLL points per GPU) and p-order to be explored further


Ad intel ADVISOR
build 615624
Perspective: GPU Roofline Insights
Summary
GPU Roofline Regions
Source View
Project: Advisor_results_02-04-2025_00-18
Application: nekrs
GPU Roofline
Default: FLOAT HBM
Guidance
GFLOPS
10000
1000
100
10
DP Vector MAD Peak: 1.14e+4 GFLOPS
DP Vector Add Peak: 5488.56 GFLOPS
HBM Bandwidth: 1183.91 GB/sec
0.01
0.1
1
10
100
1000
10000
1.00e+5
1.00e+6
1.00e+7
FLOP/Byte (Arithmetic Intensity)

Most nekRS kernels bound by HBM bandwidth

MPI PROFILES

- nekRS high-speed network usage
  - Collected MPI memory traffic usage with iprof on Aurora
  - **>50% of memory traffic done by Isend/Irecv** (point-to-point communication dominates time step loop)



```
Explicit memory traffic (BACKEND_MPI) | 1 Hostnames | 1 Processes | 1 Threads |

                     Name |     Byte | Byte(%) | Calls |   Average | Min |      Max |
                MPI_Isend |   2.19GB |  56.25% | 45793 |   47.93kB |  0B | 100.66MB |
                MPI_Irecv |   1.45GB |  37.23% | 45996 |   31.58kB |  0B |   2.43MB |
PMPI_Status_set_elements_x | 217.67MB |   5.58% |   216 |    1.01MB |  4B |   1.05MB |
        MPI_File_write_all |  18.14MB |   0.46% |     9 |    2.02MB |  4B |   7.74MB |
              MPI_Issend_c |  18.14MB |   0.46% |    26 |  697.66kB |  4B |   1.05MB |
                  MPI_Recv | 187.62kB |   0.00% |   337 |   556.72B |  8B |  27.66kB |
             MPI_Allreduce |  71.62kB |   0.00% |  3863 |    18.54B |  4B |  12.00kB |
                  MPI_Send |  15.21kB |   0.00% |   873 |    17.42B |  0B |      56B |
                 MPI_Bcast |   9.90kB |   0.00% |   246 |    40.23B |  1B |   1.98kB |
               MPI_Irecv_c |   4.69kB |   0.00% |    86 |    54.51B | 16B |     144B |
                PMPI_Bcast |   4.19kB |   0.00% |    15 |   279.40B |  4B |   4.10kB |
            PMPI_Allreduce |   2.04kB |   0.00% |    54 |    37.70B |  4B |     192B |
               MPI_Isend_c |     416B |   0.00% |     9 |    46.22B | 16B |     144B |
                MPI_Reduce |     216B |   0.00% |    11 |    19.64B | 16B |      24B |
                  MPI_Scan |     164B |   0.00% |    23 |     7.13B |  4B |       8B |
                     Total |   3.90GB | 100.00% | 97557 |
```