HPC·AI
ADVISORY COUNCIL
NETWORK OF EXPERTISE

CASTEP
Performance Benchmarking and Profiling

Feb 2019

Note

- **The following research was performed under the HPC Advisory Council activities**
  - Compute resource - HPC Advisory Council Cluster Center

- **The following was done to provide best practices**
  - CASTEP performance overview
  - CASTEP profiling information

- **More info on CASTEP**
  - http://www.castep.org/

CASTEP

- **CASTEP is a full-featured materials modelling code**
- **Based on a first-principles quantum mechanical description of electrons**
- **CASTEP uses robust methods of a plane-wave basis set and pseudo-potentials**
- **Using density functional theory, it can simulate a wide range of properties of materials proprieties including energetics, structure at the atomic level and vibrational properties**
- **CASTEP has a wide range of spectroscopic features that link directly to experiment**
  - Such as infra-red and Raman spectroscopies, NMR, and core level spectra
- **The code is developed by the Castep Developers Group (CDG) who are all UK based academics**

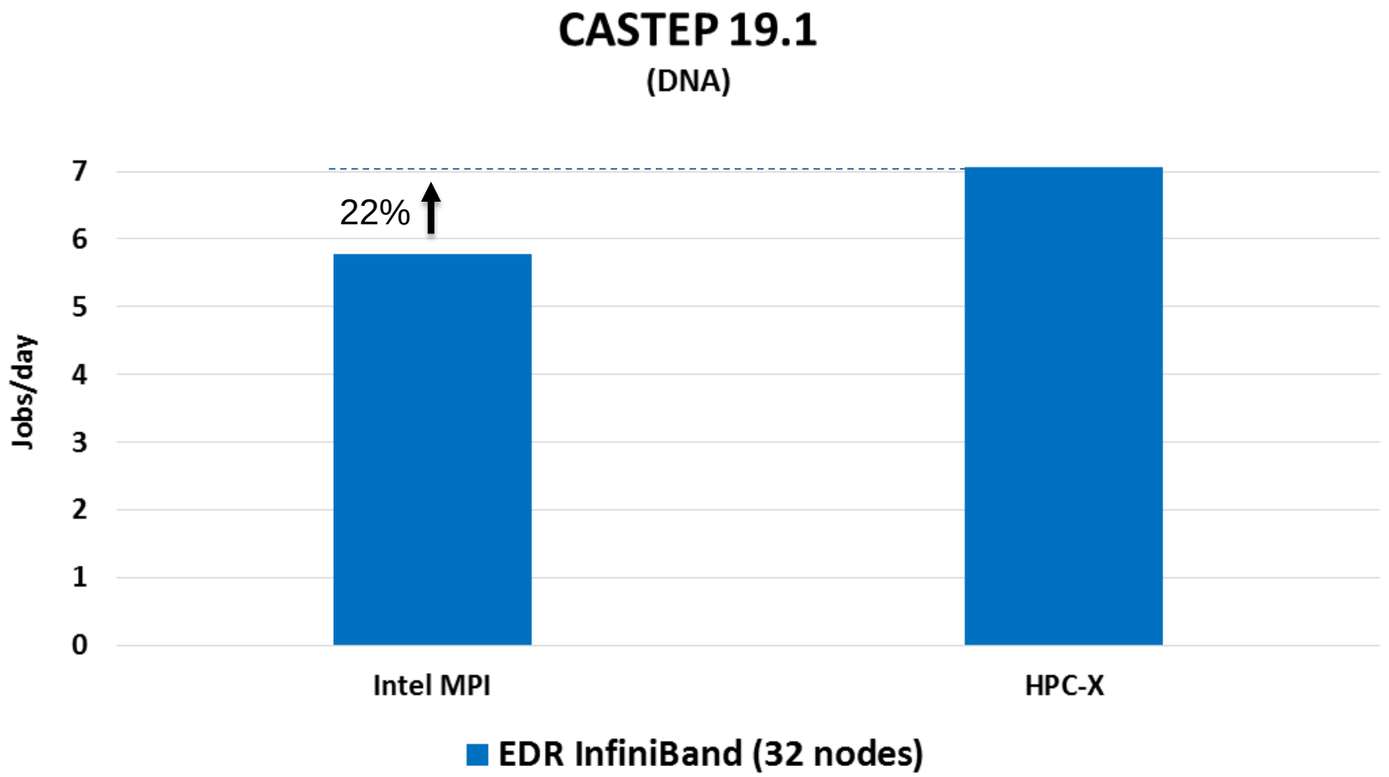
Cluster Configuration

- **Helios Cluster**
  - Supermicro SYS-6029U-TR4 / Foxconn Groot 1A42USF00-600-G 32-node cluster
  - Dual Socket Intel(R) Xeon(R) Gold 6138 CPU @ 2.00GHz
  - Mellanox ConnectX-5 EDR 100Gb/s InfiniBand/VPI adapters
  - Mellanox Switch-IB 2 SB7800 36-Port 100Gb/s EDR InfiniBand switch
  - Memory: 192GB DDR4 2677MHz RDIMMs per node
  - 1TB 7.2K RPM SSD 2.5" hard drive per node

- **Software**
  - OS: RHEL 7.5, MLNX_OFED 4.4
  - MPI: HPC-X 2.2
  - CASTEP 19.1

CASTEP Performance – MPI Comparision

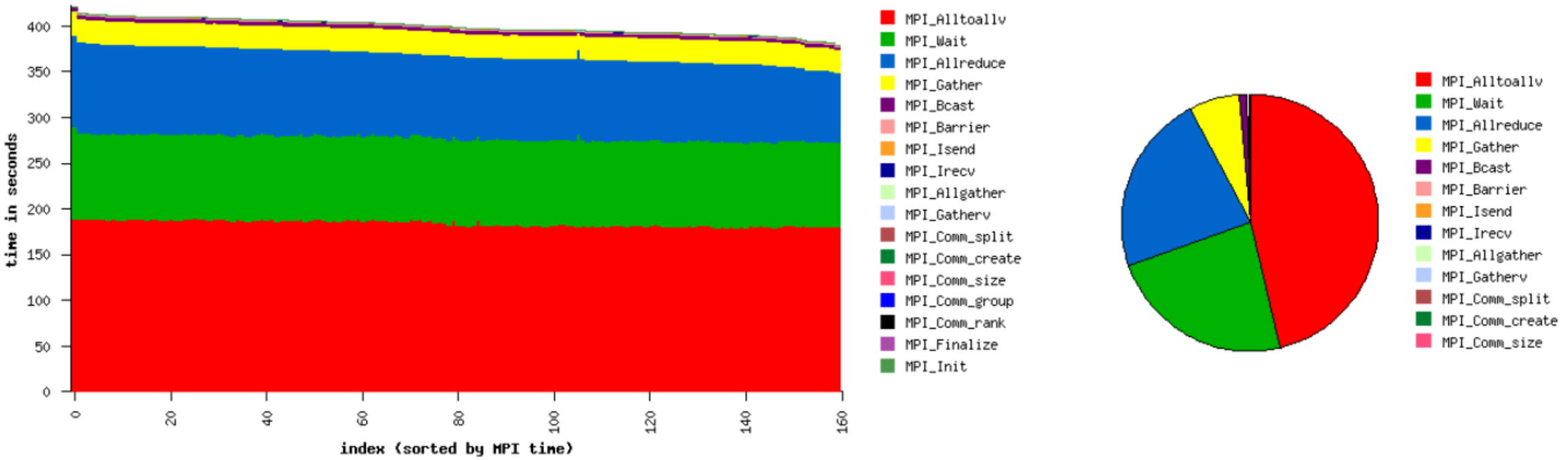
- **DNA Benchmark (large Input)**



Higher is better

CASTEP Profiling

- Benchmark: “ham8_1” (4 nodes, 160 cores)
- 35% of the time is spent in MPI



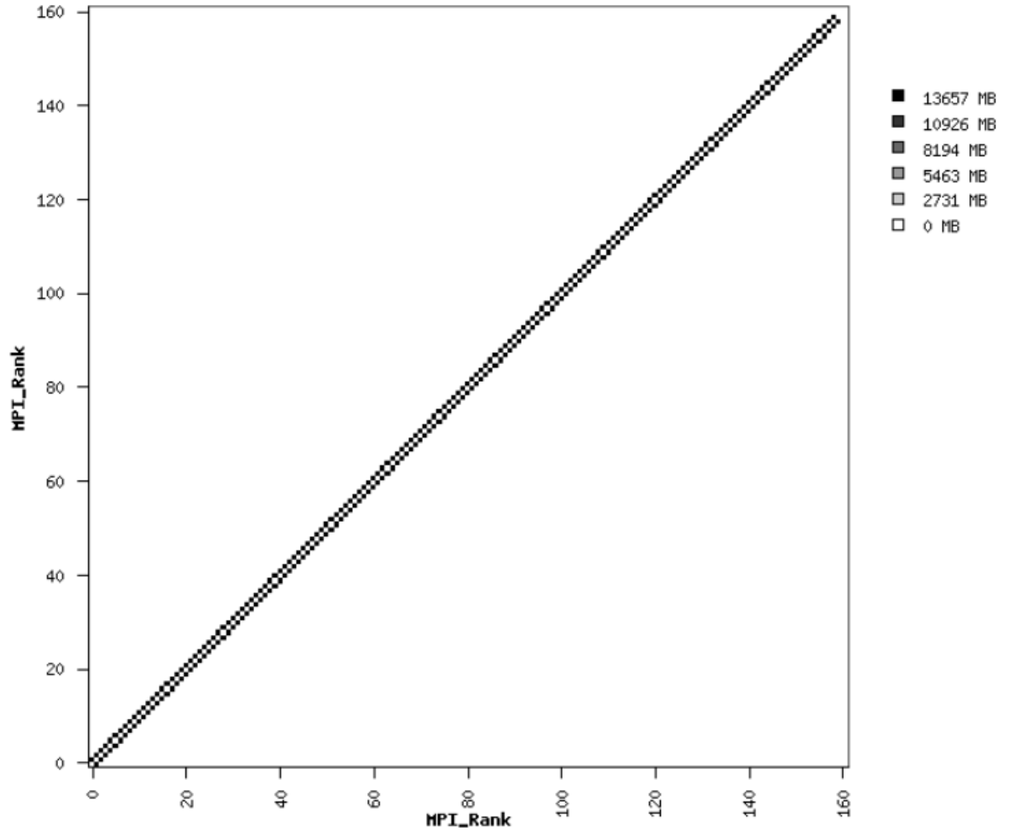
CASTEP Profiling

- **Communication Statistics**
- **Benchmark: “ham8_1” (4 nodes, 160 cores)**

Communication Event Statistics (% detail, --- error)									
	Comm Size	Buffer Size	Ncalls	Total Time	Avg Time	Min Time	Max Time	%MPI	%Wall
MPI_Alltoallv	160	640	28427520	1.554733e+04	5.469113e-04	3.440400e-04	8.391100e-03	24.33	8.68
MPI_Wait	0	0	111180432	1.493563e+04	1.343369e-04	0.000000e+00	1.106000e-02	23.38	8.33
MPI_Alltoallv	160	10240	21378840	1.049319e+04	4.908215e-04	2.410400e-04	6.462100e-03	16.42	5.85
MPI_Gather	160	8192	595520	4.004433e+03	6.724263e-03	5.240400e-04	3.830000e-02	6.27	2.23
MPI_Allreduce	160	655360	2148320	3.573959e+03	1.663607e-03	9.620200e-04	1.508200e-02	5.59	1.99
MPI_Alltoallv	160	12288	7126280	3.407276e+03	4.781283e-04	2.408000e-04	6.469000e-03	5.33	1.90
MPI_Allreduce	160	4	4953280	2.528469e+03	5.104636e-04	1.907300e-06	8.155700e+00	3.96	1.41
MPI_Allreduce	160	3145728	331200	1.908094e+03	5.761154e-03	4.545900e-03	1.258700e-02	2.99	1.06
MPI_Allreduce	160	2097152	413120	1.786817e+03	4.325176e-03	3.027000e-03	1.023100e-01	2.80	1.00
MPI_Allreduce	160	4194304	201600	1.606039e+03	7.966461e-03	4.558100e-03	1.972200e-01	2.51	0.90
MPI_Allreduce	160	2621440	140160	6.194720e+02	4.419749e-03	3.809000e-03	1.104100e-02	0.97	0.35
MPI_Allreduce	1	4194304	262400	6.189354e+02	2.358748e-03	4.940000e-04	1.405100e-02	0.97	0.35
MPI_Allreduce	160	640	163680	3.584327e+02	2.189838e-03	1.406700e-05	2.107000e-02	0.56	0.20
MPI_Allreduce	160	393216	297440	3.293487e+02	1.107278e-03	5.979500e-04	6.487100e-03	0.52	0.18
MPI_Allreduce	160	327680	209920	2.238997e+02	1.066596e-03	4.971000e-04	6.552000e-03	0.35	0.12
MPI_Bcast	160	4	691200	1.888959e+02	2.732869e-04	0.000000e+00	4.951200e-01	0.30	0.11
MPI_Barrier	160	0	960	1.767835e+02	1.841495e-01	5.960500e-06	8.727200e-01	0.28	0.10
MPI_Allreduce	160	8	9139840	1.629248e+02	1.782578e-05	9.536700e-07	1.180000e-02	0.26	0.09
MPI_Bcast	160	14336	12160	1.339389e+02	1.101472e-02	4.053100e-06	4.280300e-02	0.21	0.07
MPI_Alltoallv	160	1024	76145	1.003772e+02	1.318238e-03	6.291900e-04	5.705400e-02	0.16	0.06
MPI_Allreduce	160	1310720	9280	9.003176e+01	9.701698e-03	1.877100e-03	5.079200e-02	0.14	0.05
MPI_Allreduce	160	524288	60480	8.807522e+01	1.456270e-03	8.060900e-04	1.101400e-02	0.14	0.05
MPI_Bcast	160	25165824	4960	7.994003e+01	1.611694e-02	1.531100e-02	2.475600e-02	0.13	0.04
MPI_Bcast	160	655360	119520	7.832396e+01	6.553209e-04	1.399500e-04	6.282100e-03	0.12	0.04
MPI_Allreduce	160	114688	280640	7.760110e+01	2.765148e-04	1.378100e-04	5.524200e-03	0.12	0.04
MPI_Bcast	160	1048576	62880	6.381318e+01	1.014841e-03	3.061300e-04	9.431800e-03	0.10	0.04
MPI_Allreduce	160	458752	44640	5.918353e+01	1.325796e-03	7.009500e-04	6.158800e-03	0.09	0.03

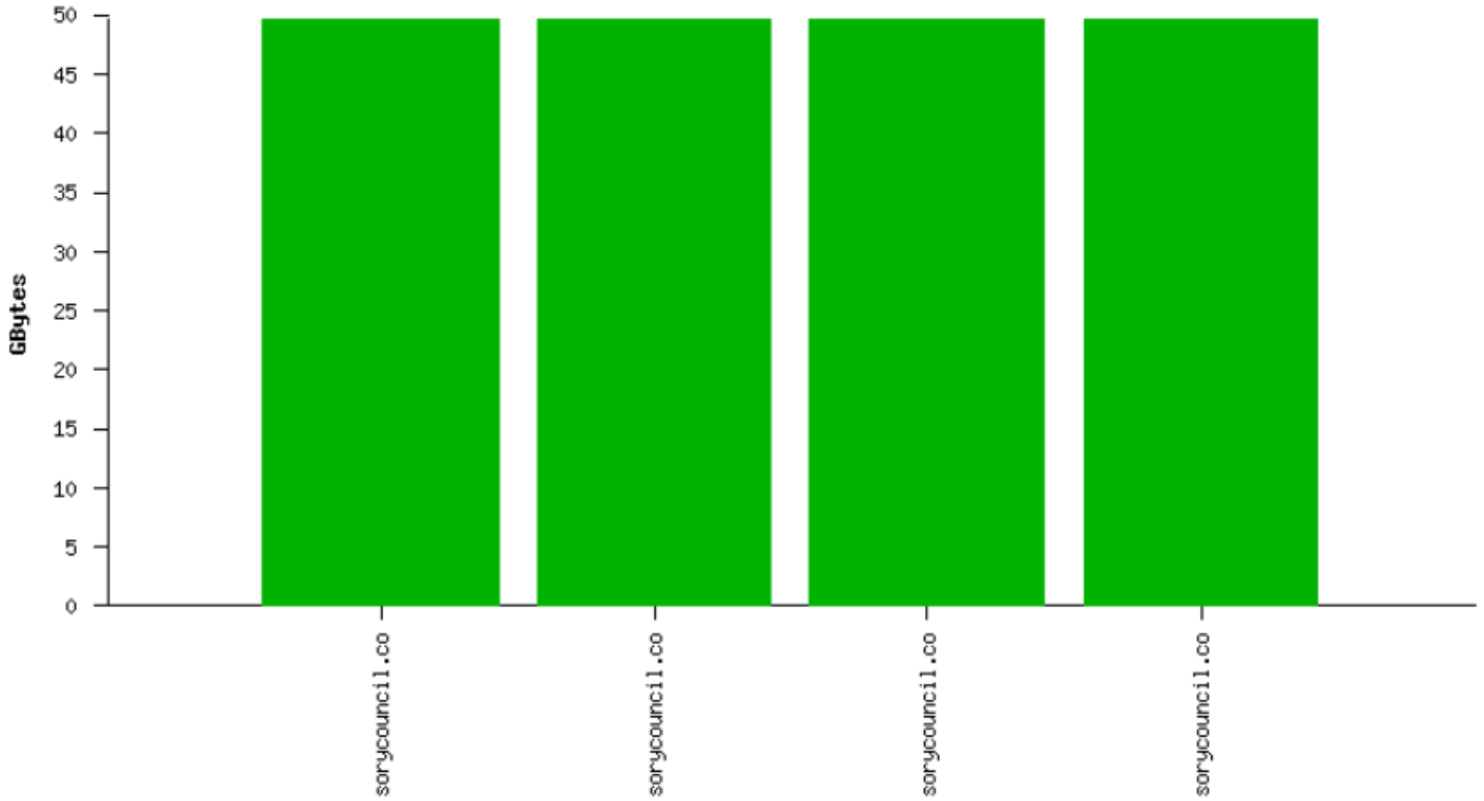
CASTEP Profiling

- Near core communication for point-to-point
- Benchmark: “ham8_1” (4 nodes, 160 cores)



CASTEP Profiing

- Memory usage of ~50GB per node
- Benchmark: “ham8_1” (4 nodes, 160 cores)



Summary

- **CASTEP performance testing**
  - HPC-X Enables 22% higher performance versus Intel MPI using 32 nodes and the “DNA “ input set

- **CASTEP profiling on “ham8_1”**
  - MPI communication accounts for 35% of overall wall clock time
  - MPI_Alltoallv is 46% of MPI, MPI_Wait is 29% of MPI and MPI_Allreduce is 28% of MPI
  - Most point to point communications are between near ranks

- **CASTEP mpirun command**

mpirun -np $nranks --map-by node --bind-to core -report-bindings --display-map -mca coll_hcoll_enable 1 -mca coll_hcoll_np 0 -x HCOLL_IB_IF_INCLUDE=mlx5_0:1 -x HCOLL_MAIN_IB=mlx5_0:1 -x HCOLL_SBGP=basesmsocket,basesmuma,p2p -x HCOLL_BCOL=basesmuma,basesmuma,ucx_p2p -x HCOLL_ENABLE_MCAST_ALL=1 -x HCOLL_MCAST_NP=0 -x HCOLL_ML_HYBRID_ALLTOALLV_RADIX=0 -x UCX_RC_MLX5_TM_ENABLE=n -x UCX_DC_MLX5_TM_ENABLE=n -mca pml ucx -x UCX_NET_DEVICES=mlx5_0:1 ./castep.mpi polyA20-no-wat

HPC·AI ADVISORY COUNCIL
NETWORK OF EXPERTISE

Thank You