

hpcscan version 1.1

Performance benchmarks on Shaheen II (KAUST)

December 2020

- 1 Introduction
- 2 Shaheen II (KAUST)
- 3 Test Case Memory
- 4 Test Case Grid
- 5 Test Case Comm
- 6 Test Case FD_D2
- 7 Test Case Propa
- 8 Summary
- 9 Acknowledgements

- 1 Introduction
- 2 Shaheen II (KAUST)
- 3 Test Case Memory
- 4 Test Case Grid
- 5 Test Case Comm
- 6 Test Case FD_D2
- 7 Test Case Propa
- 8 Summary
- 9 Acknowledgements

Introduction

This document presents a characterization of the computing nodes and interconnect of the supercomputer Shaheen II at KAUST. The full set of test cases embedded in `hpcscan` is used in various configurations.

List of test cases in this study

Test Case	Objectives	Remark
Memory	Assess memory bandwidth	Scalability analysis on a single node
Grid	Assess bandwidth of grid operations	Analyse effect of the grid size
Comm	Assess inter-node communication bandwidth	Analyse effect of subdomain decomposition
FD_D2	Assess FD spatial derivative computation bandwidth	Analyse effect of FD stencil order
Propa	Find optimal configuration for the wave propagator	Explore range of parameters
Propa	Scalability analysis of wave propagator on multiple nodes	Analyse effect of the FD stencil order

General settings

- All tests are performed in single precision
- Best performance is reported over 10 tries for each case (unless stated otherwise)
- Grids dimensions are 3D
- Grids sizes range from 500 MB up to 4 GB per node

- 1 Introduction
- 2 **Shaheen II (KAUST)**
- 3 Test Case Memory
- 4 Test Case Grid
- 5 Test Case Comm
- 6 Test Case FD_D2
- 7 Test Case Propa
- 8 Summary
- 9 Acknowledgements

Machine Shaheen II / Cray XC40

- Computing nodes Intel Haswell 2.3 Ghz dual socket (16 cores / socket)
- RAM 128 GB with Peak memory BW 136.5 GB/s
- Peak performance Single Prec. 2.36 TFLOP/s / Double Prec. 1.18 TFLOP/s
- Interconnect Cray Aries with Dragonfly topology
 - 60 GB/s optical links between groups
 - 8.5 GB/s copper links between chassis
 - 3.5 GB/s backplane within a chassis
 - 5 GB/s PCIe from node to Aries router



- 1 Introduction
- 2 Shaheen II (KAUST)
- 3 Test Case Memory**
- 4 Test Case Grid
- 5 Test Case Comm
- 6 Test Case FD_D2
- 7 Test Case Propa
- 8 Summary
- 9 Acknowledgements

Test Case Memory - Description

Benchmark objective

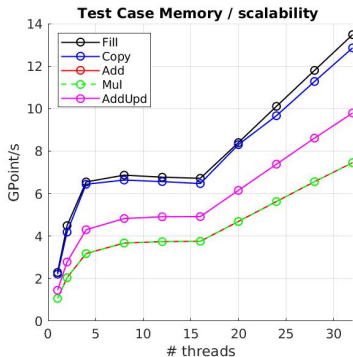
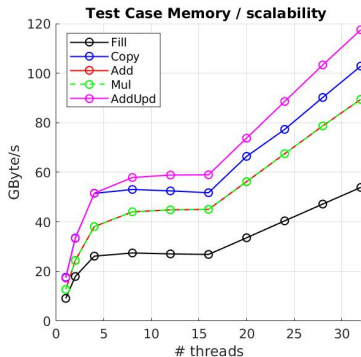
Assess memory bandwidth

- Measure GByte/s and GPoint/s for simple operations on memory arrays
- Scalability analysis on a single node
- Get a reference to compare with for the following tests

Benchmark configuration

- Scalability on 1 node with 1 to 32 threads
- Baseline kernel
- Array size 4 GB
- Reproduce results with `./script/testCase_Memory/hpcscanMemory.sh`
- Total 10 configurations, Elapsed time about 4 minutes

Test Case Memory - Results ¹



¹Updated Dec 22, 2020

- 1 Introduction
- 2 Shaheen II (KAUST)
- 3 Test Case Memory
- 4 Test Case Grid**
- 5 Test Case Comm
- 6 Test Case FD_D2
- 7 Test Case Propa
- 8 Summary
- 9 Acknowledgements

Benchmark objective

Assess bandwidth of grid operations

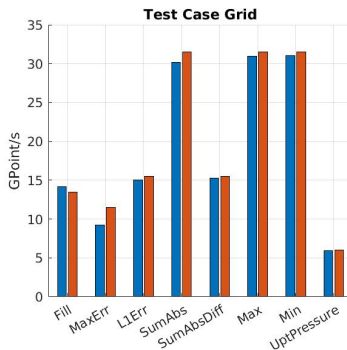
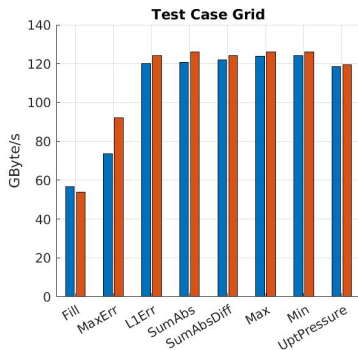
- Measure GByte/s and GPoint/s for simple and complex operations on 3D grids
- Analyse effect of the grid size

Benchmark configuration

- 1 node with 32 threads
- Baseline kernel
- 2 grid sizes
 - Small size 500 MB ($500 \times 500 \times 500$ points)
 - Medium size 4 GB ($1000 \times 1000 \times 1000$ points)
- Reproduce results with `./script/testCase_Grid/hpcscanGrid.sh`
- Total 2 configurations, Elapsed time less than 1 minute

Test Case Grid - Results ²

Blue=small grid, Red=medium grid



ApplyBoundaryCondition performs at 713/846 GBytes (89/105 Gpoint/s)

²Updated Dec 23, 2020

- 1 Introduction
- 2 Shaheen II (KAUST)
- 3 Test Case Memory
- 4 Test Case Grid
- 5 Test Case Comm**
- 6 Test Case FD_D2
- 7 Test Case Propa
- 8 Summary
- 9 Acknowledgements

Benchmark objective

Assess inter-node communication bandwidth

Point to point communication

- Half-duplex BW with MPI_Send from proc X to proc 0
- Full-duplex BW with MPI_Sendrecv between proc X and proc 0

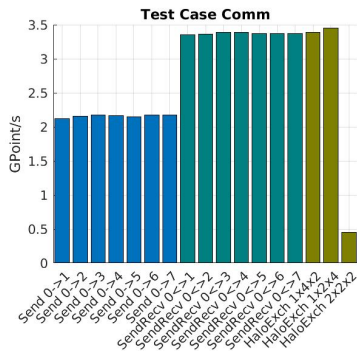
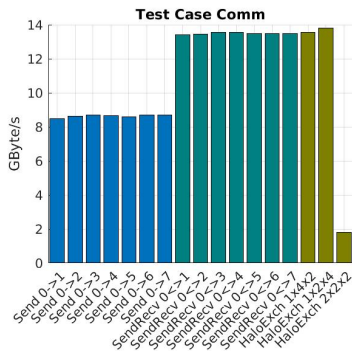
Collective communication

- Grid halos exchange (used in FD kernels) with MPI_Sendrecv
- Analyse effect of subdomain decomposition geometry

Benchmark configuration

- 8 nodes with 1 MPI/node & 32 threads/node
- Baseline kernel
- Grid size 4 GB (1000 × 1000 × 1000 points)
- FD order O8
- Subdomain decomposition: 1x4x2, 1x2x4 & 2x2x2
- Reproduce results with `./script/testCase_Comm/hpcscanComm.sh`
- Total 3 configurations: Elapsed time less than 1 minute

Test Case Comm - Results ³



³Updated Dec 26, 2020

- 1 Introduction
- 2 Shaheen II (KAUST)
- 3 Test Case Memory
- 4 Test Case Grid
- 5 Test Case Comm
- 6 Test Case FD_D2**
- 7 Test Case Propa
- 8 Summary
- 9 Acknowledgements

Benchmark objective

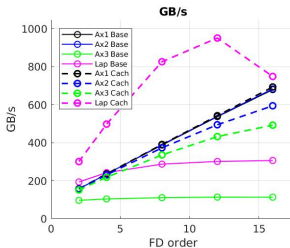
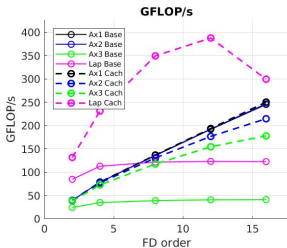
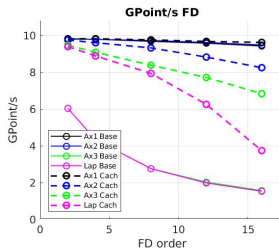
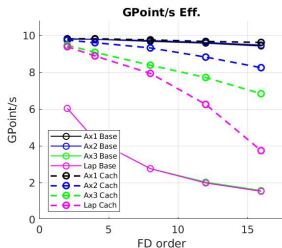
Assess FD spatial derivative computation bandwidth

- Directional derivatives
 - Axis 1, $W = \partial_{x1}^2(U)$
 - Axis 2, $W = \partial_{x2}^2(U)$
 - Axis 3, $W = \partial_{x3}^2(U)$
- Laplacian $W = \Delta(U)$
- Analyse effect of FD stencil order
- Try different implementations of FD computation

Benchmark configuration

- 1 node with 32 threads
- 2 test modes: Baseline & CacheBlk
- Grid size 4 GB (1000 × 1000 × 1000 points)
- FD orders 2, 4, 8, 12 & 16
- Reproduce results with `./script/testCase_FD_D2/hpcscanFD_D2.sh`
- Total 10 configurations: Elapsed time about 2 minutes

Test Case FD_D2 - Results ⁴



- 1 Introduction
- 2 Shaheen II (KAUST)
- 3 Test Case Memory
- 4 Test Case Grid
- 5 Test Case Comm
- 6 Test Case FD_D2
- 7 Test Case Propa**
- 8 Summary
- 9 Acknowledgements

Test Case Propa - Description

Benchmark objective

Find optimal configuration for the wave propagator regarding accuracy/cost

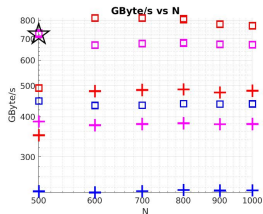
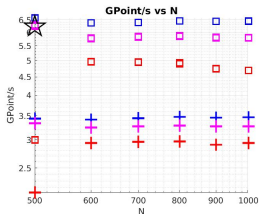
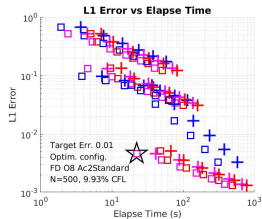
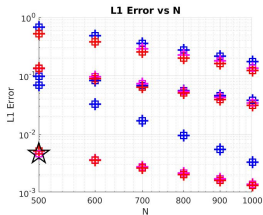
- Explore range of grid sampling
- Explore range of time step
- Explore range of FD order
- Try different implementations of the propagator

Benchmark configuration

- 1 node with 32 threads
- Test mode CachBlk
- 2 propagator implementations: Ac2Standard and Ac2SplitComp
- FD orders 4, 8 & 12
- Time step 100, 50 and 10% of stability time step
- Grid size from 500x500x500 (500 MB) to 1000x1000x1000 (4 GB)
- nt from 101 to 2311 (depending of the configuration) & ntry = 4
- Reproduce results with
`./script/testCase_Propa/paramAnalysis/hpcscanPropaParamAnalysis.sh`
- Total 108 configurations: Elapsed time about 12 hours

Test Case Propa - Results ⁵

Blue=FD O4, Pink=FD O8, Red=FD O12 / Square=Ac2Standard, Cross=Ac2SplitComp



⁵ Updated Dec 28, 2020

Test Case Propa - Description

Benchmark objective

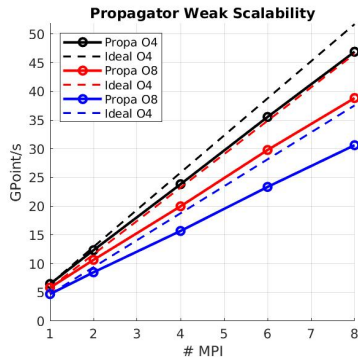
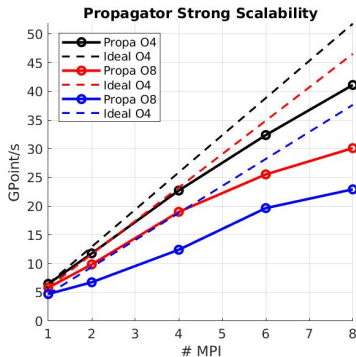
Scalability analysis of the wave propagator on multiple nodes

- Strong and weak scalability
- Analyse effect of the FD stencil order

Benchmark configuration

- From 1 node to 8 nodes with 32 threads/node
- Test mode CachBlk
- Propagator implementation Ac2Standard
- FD orders 4, 8 & 12
- Strong scalability: Grid size 1000x1000x1000 (4 GB)
- Weak scalability: Grid size from 1000x1000x1000 (4 GB) to 1000x4000x2000 (32 GB)
- $nt = 100$
- Reproduce results with
./script/testCase_strongWeakScalability/hpcscanPropaStrongWeakScalability.sh
- Total 30 configurations: Elapsed time about 1h 15min

Test Case Propa - Results ⁶



⁶ Updated Dec 28, 2020

- 1 Introduction
- 2 Shaheen II (KAUST)
- 3 Test Case Memory
- 4 Test Case Grid
- 5 Test Case Comm
- 6 Test Case FD_D2
- 7 Test Case Propa
- 8 Summary**
- 9 Acknowledgements

Test Case Memory

- Highest memory BW for Add+Update 118 GB/s [9.8 GPoint/s] (86 % of peak BW)
- Lowest memory BW for Fill 54 GB/s [13.5 GPoint/s] (40 % of peak BW)
- All cores (32 threads) are required to reach best perf. on the node

Test Case Grid

- Highest memory BW for L1Err, SumAbs, SumAbsDiff, GetMin & GetMax 125 GB/s (92 % of peak BW)
- In terms of GPoint/s: 15 for L1Err & SumAbsDiff and 30 for SumAbs, GetMin & GetMax
- Lowest memory BW for Fill 54 GB/s [13.5 GPoint/s] (40 % of peak BW)
- Minor effect of the grid size on the performance
- Results are consistent with the test case Memory

Test Case Comm

- Half-duplex BW about 8.6 GByte/s [2.1 GPoint/s]
- Full-duplex BW about 13.4 GByte/s [3.3 GPoint/s]
- Communication BW is similar between 8 nodes (no unbalance observed)
- Grid halos exchange: large effect of subdomain decomposition geometry
 - Best BW: 1x4x2 or 1x2x4 with 13.6 GByte/s [3.4 GPoint/s]
 - Worst BW: 2x2x2 with 1.8 GByte/s [0.5 GPoint/s]

Test Case FD_D2

- Large benefit of cache blocking
- Significant effect of grid dimension and index (very bad performance for n3 without cache blocking)
- Min BW 50 GFLOP/s (∂_{x3}^2 O2) = 2 % peak BW [apparent Mem. BW 150 GB/s]
- Max BW 370 GFLOP/s (Δ O8) = 16 % peak BW [apparent Mem. BW 900 GB/s]
- Apparent Mem. BW 150-900 GB/s (110-660 % Peak Mem. BW) = shows data in-cache effect
- Typical stencils of interest for geophysical applications
 - Δ O4 BW = 8-10 GPoint/s
 - Δ O8 BW = 7-9 GPoint/s
 - Δ O12 BW = 3-5 GPoint/s
- Parallel efficiency with 8 nodes 55 to 86 % (depends on workload on Shaheen)

Test Case Propa

- TO DO

Content

- 1 Introduction
- 2 Shaheen II (KAUST)
- 3 Test Case Memory
- 4 Test Case Grid
- 5 Test Case Comm
- 6 Test Case FD_D2
- 7 Test Case Propa
- 8 Summary
- 9 Acknowledgements**

Acknowledgements

- KAUST ECRC and KSL for access and support on Shaheen II