

# 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 KAUST supercomputer Shaheen II. 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 are 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

## Benchmark objective

### Assess memory bandwidth

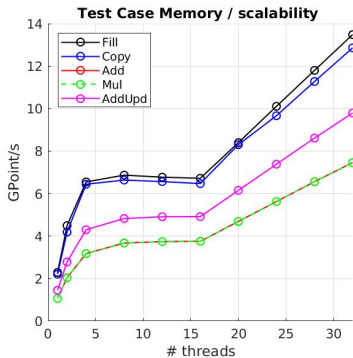
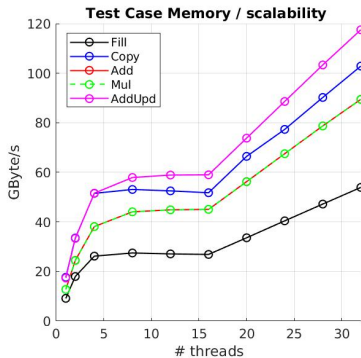
- Measure GByte/s and GPoint/s for simple operations on memory arrays
- Scalability analysis on a single node
- Get a reference to compare with for the following tests

## Benchmark configuration

- Scalability on 1 node with 1 to 32 threads
- Baseline kernel
- Grid size 4 GB (1000 × 1000 × 1000 points)
- Reproduce results with `./script/testCase_Memory/hpcscanMemory.sh`
- Total 10 configurations, Elapsed time about 4 minutes



# Test Case Memory - Results <sup>1</sup>



<sup>1</sup>Updated 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

# Test Case Grid - Description

## Benchmark objective

### Assess bandwidth of grid operations

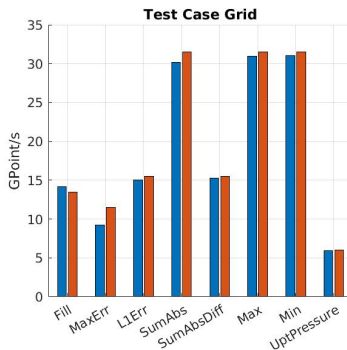
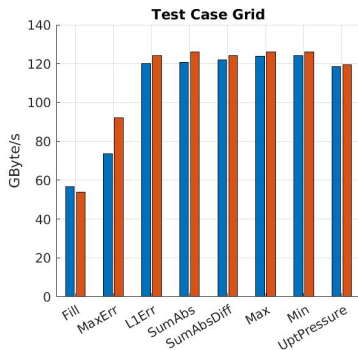
- Measure GByte/s and GPoint/s for simple and complex operations on 3D grids
- Analyse effect of the grid size

## 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 <sup>2</sup>

Blue=small grid, Red=medium grid



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

<sup>2</sup>Updated 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

MPI point to point communication

- Send with MPI\_Send from proc X to proc 0 (Half-duplex BW)
- Send and receive with MPI\_Sendrecv between proc X and proc 0 (Full-duplex BW)

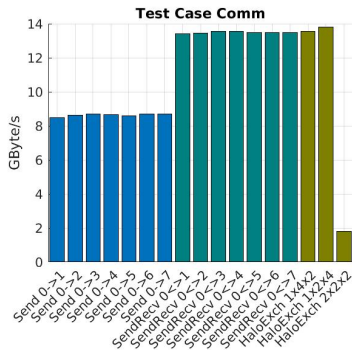
MPI collective communication

- Exchange of halos 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 <sup>3</sup>



<sup>3</sup>Updated 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

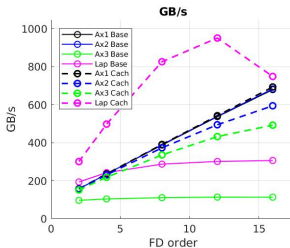
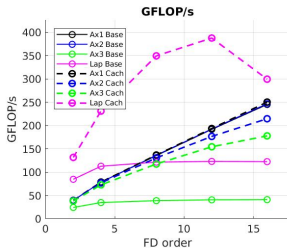
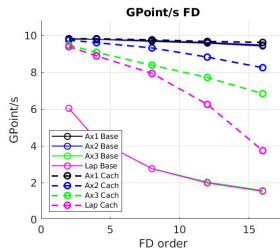
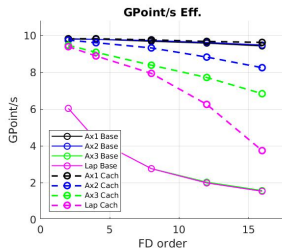
### Assess FD spatial derivative computation bandwidth

- 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 <sup>4</sup>



- 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**

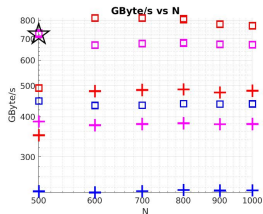
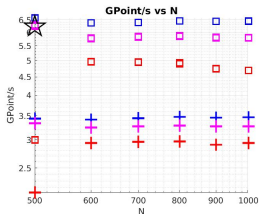
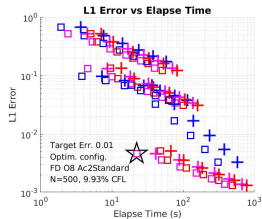
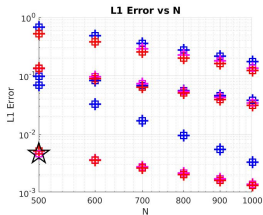
- Explore range of grid sampling
- Explore range of time step
- Explore range of FD order
- Try different implementations of the propagator

## 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 <sup>5</sup>

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



<sup>5</sup> Updated Dec 28, 2020

# Test Case Propa - Description

## Benchmark objective

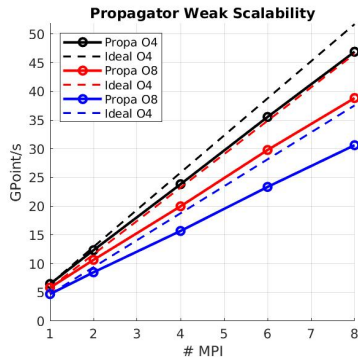
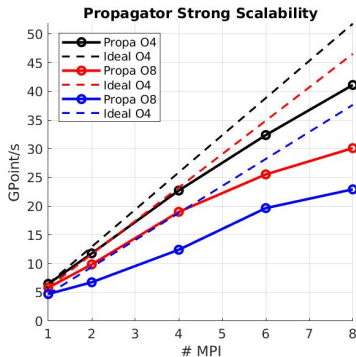
### Scalability analysis of the wave propagator on multiple nodes

- 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 <sup>6</sup>



<sup>6</sup> 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

- Measured memory BW between 91 to 122 GB/s (67-90 % of peak BW)
- Low BW 59 GB/s for Fill (43 % of peak BW)
- Multiply (= imaging condition) performs at 7.6 Gpoint/s

## Test Case Grid

- L1 Err., Get Min & Max: 125 GB/s close to peak BW (92 % Peak Mem. BW)
- Low perf for Fill: 54-58 GB/s (40-43 % Peak Mem. BW)
- Max Err. 72-91 GB/s (53-67 % Peak Mem. BW)
- Pressure update 6 GPoint/s (120 GB/s, 88 % Peak Mem. BW)

## Test Case Comm

- TO DO

## 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 ( $\partial_{x3}^2$  O2) = 2 % peak BW [apparent Mem. BW 150 GB/s]
- Max BW 370 GFLOP/s ( $\Delta$  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
  - $\Delta$  O4 BW = 8-10 GPoint/s
  - $\Delta$  O8 BW = 7-9 GPoint/s
  - $\Delta$  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