SMART-PETSc: High-Performance MPI Library to Boost Performance of the PETSc Library

Drs. Donglai Dai (PI), **Sreevatsa (Sreev) Anantharamu** (Lead Developer), Hari Subramoni, D. K. Panda

s.anantharamu@x-scalesolutions.com

X-ScaleSolutions

http://x-scalesolutions.com

SMART-PETSc

Multi-core Processors

PETSc

“PETSc, the Portable, Extensible Toolkit for Scientific Computation, includes a large suite of scalable parallel linear and nonlinear equation solvers, ODE integrators, and optimization algorithms for application codes written in C, C++, Fortran, and Python. In addition, PETSc includes support for managing parallel PDE discretizations including parallel matrix and vector assembly routines.” (Used by more than 30 scientific toolkits/libraries)

⊕

MVAPICH

“The MVAPICH2 software, based on MPI 3.1 standard, delivers the best performance, scalability and fault tolerance for high-end computing systems and servers using InfiniBand, Omni-Path, Ethernet/iWARP, RoCE(v1/v2), Cray Slingshot 10 and 11, and Rockport Networks networking technologies. This software is being used by more than 3,325 organizations in 90 countries worldwide to extract the potential of these emerging networking technologies for modern systems.”

GPU accelerators
high compute density, high performance/watt
>9.7 TFlop DP on a chip

High Performance Interconnects – InfiniBand (DPU), Slingshot, Omnipath (IPU)
<1usec latency, 200-400Gbps Bandwidth>

- SMART-PETSc is a co-designed PETSc + MVAPICH2 middleware
- Goal: Deliver best performance for PETSc end-applications via co-design to take full advantage of modern HPC architecture features
- Challenge: How?
- Thanks to support from DOE SBIR Phase-II and -I

Modern exascale machines (Frontier, El Capitan, Aurora) and cloud HPC (AWS, Azure)

Team

- Drs. Donglai Dai, Sreevatsa (Sreev) Anantharamu, Hari Subramoni, D. K. Panda (Led by X-ScaleSolutions, LLC)
- Drs. Richard Tran Mills, Junchao Zhang (Argonne National Lab)
- Dr. Victor Eijkhout (Texas Advanced Computing Center)
- Drs. Sameer Shende, Allen Malony (ParaTools, Inc)

Presentation Outline

- MPI communication patterns in PETSc
- Optimizations
  - ➢ Matrix-vector multiplication kernel
  - ➢ Finite Difference PETSc application
  - ➢ Finite Element PETSc application
  - ➢ Intra-node bandwidth on modern CPUs
  - ➢ MVAPICH2-DPU with PETSc
  - ➢ Co-designed rendezvous protocols
- Profiling
  - ➢ TAU + PETSc via perfstubs
- Conclusions

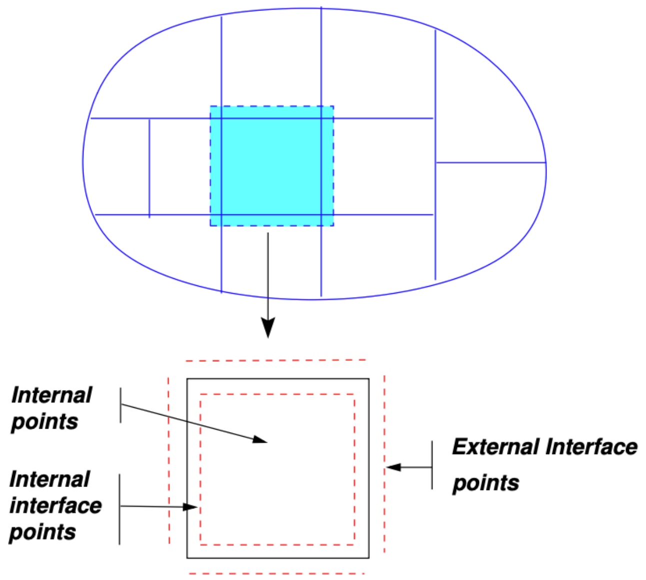
MPI communication patterns in PETSc

Point-to-point

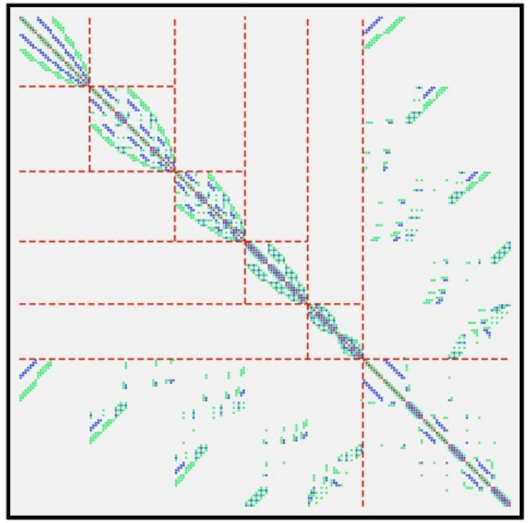
- Near-neighbor
- Parallel matrix-vector multiplication, assembly
- Solutions transfers (prolongators/restrictors) in multi-grid/-level preconditioners
- Krylov Solvers, Preconditioners
- libCEED exascale library (matrix-free back-end of PETSc) needs two rounds of point-to-point communication for each matrix-vector multiplication (compared to just one round for the usual assembled matrices)

Collectives

- Inner-products in Krylov solvers
- Coarse level solution in multigrid preconditioner



Partitioning PDE problems



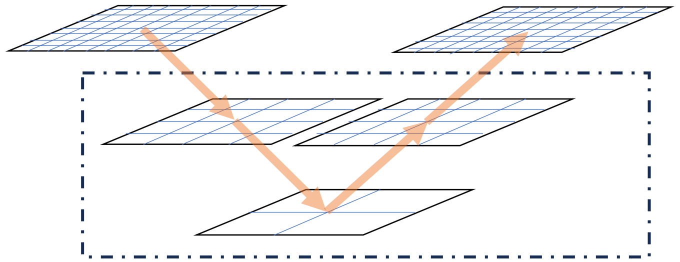
Sparsity pattern of a parallel sparse matrix (Cover page of Prof. Saad's book)

Compute $r_0 := b - Ax_0$, $p_0 := r_0$.
For $j = 0, 1, \ldots$, *until convergence Do:*

$$\alpha_j := (r_j, r_j)/(Ap_j, p_j)$$
$$x_{j+1} := x_j + \alpha_j p_j$$
$$r_{j+1} := r_j - \alpha_j A p_j$$
$$\beta_j := (r_{j+1}, r_{j+1})/(r_j, r_j)$$
$$p_{j+1} := r_{j+1} + \beta_j p_j$$

EndDo

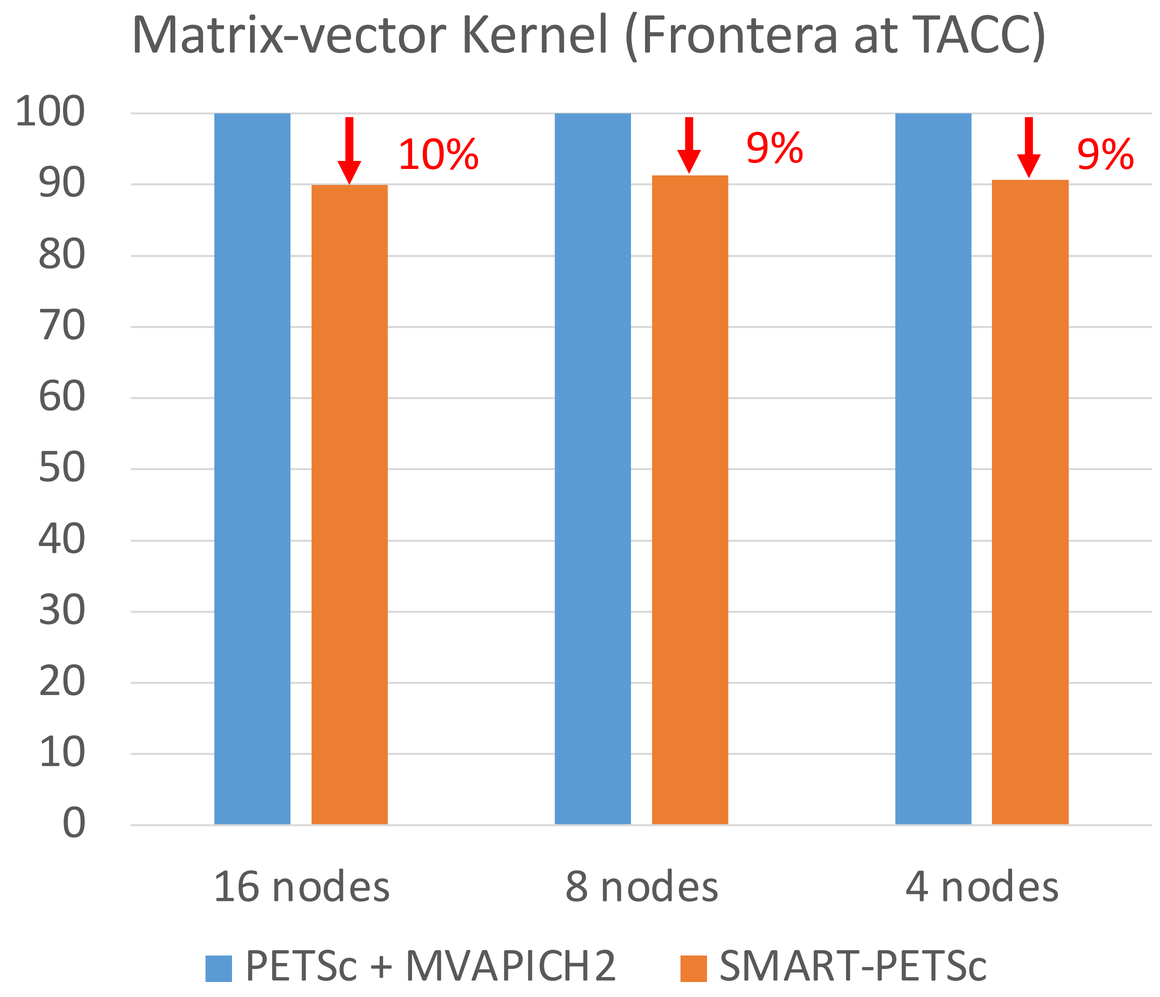
Classical CG, (,) denotes an inner product

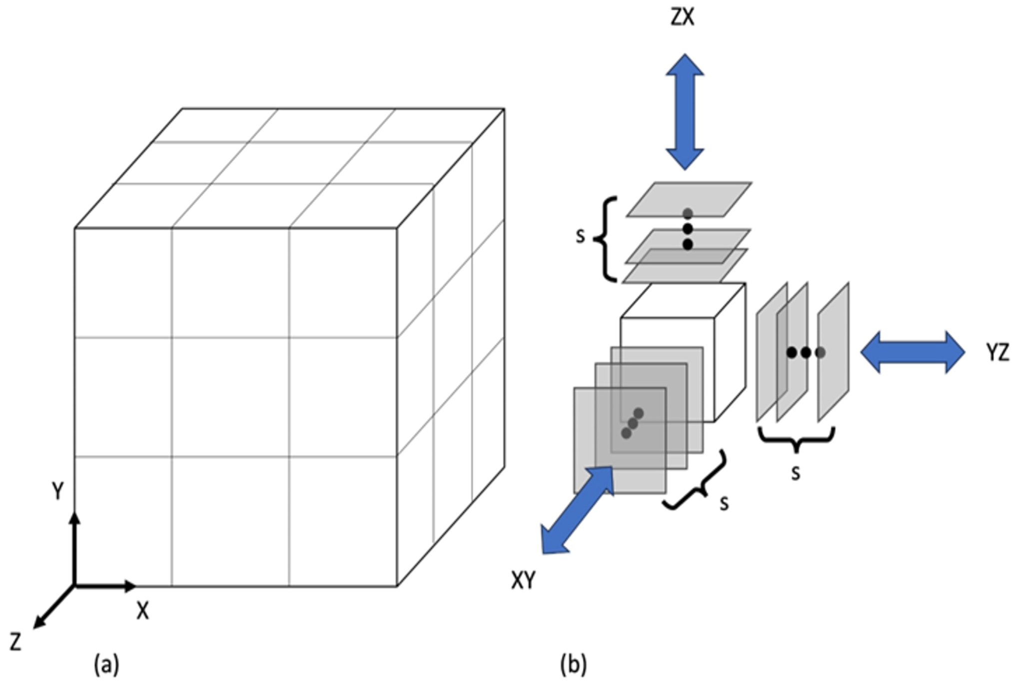


Multigrid

Optimizations: Matrix-vector kernel

- High-order finite-difference Laplacian stencil (stencil width=5)
- Up to 10% benefit
- Environment variable MV2_SMART_PETSC_MATVEC_OPT=1



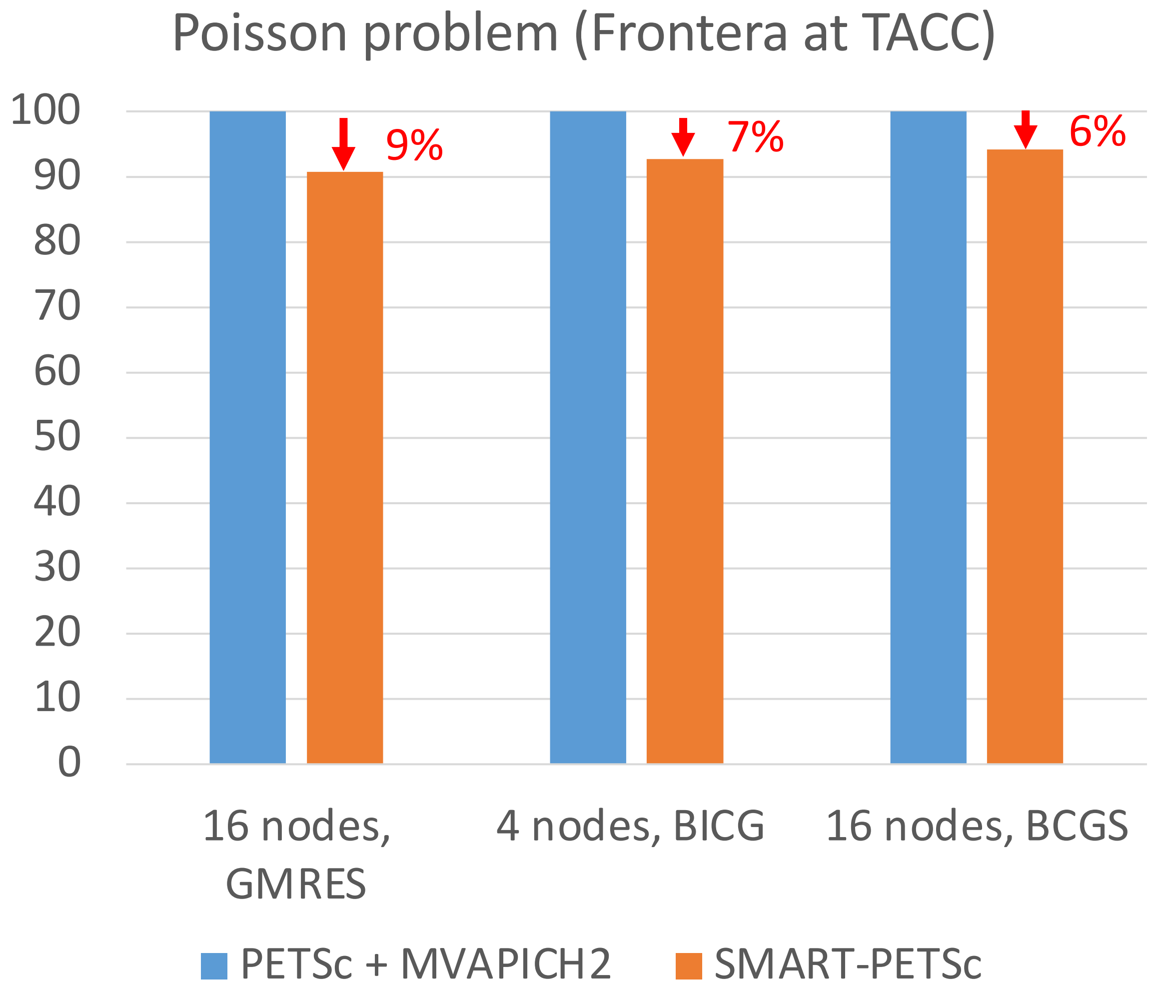


Compute grid and communication pattern

Processors	Intel 8280 “Cascade Lake”
Cores/Node	56 (28 per socket)
Memory/Node	192GB DDR-4
Network	Mellanox Infiniband, HDR-100

Frontera system specification

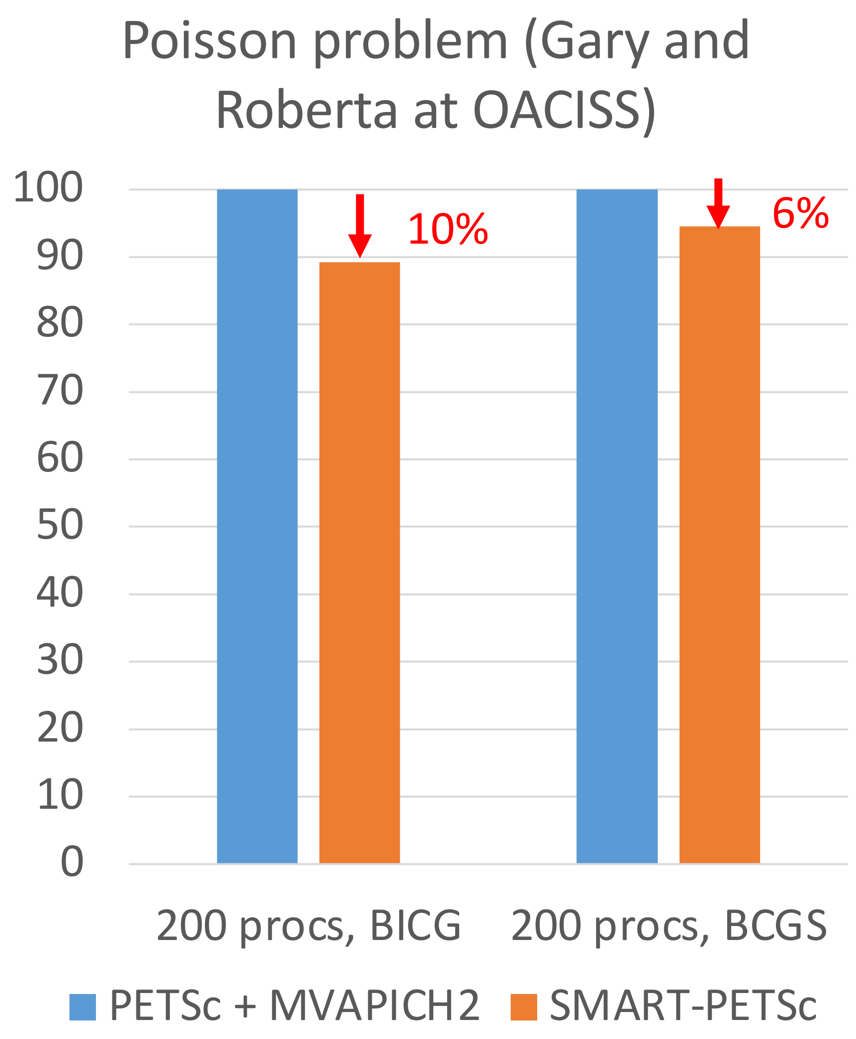
Optimizations: Finite Difference PETSc application

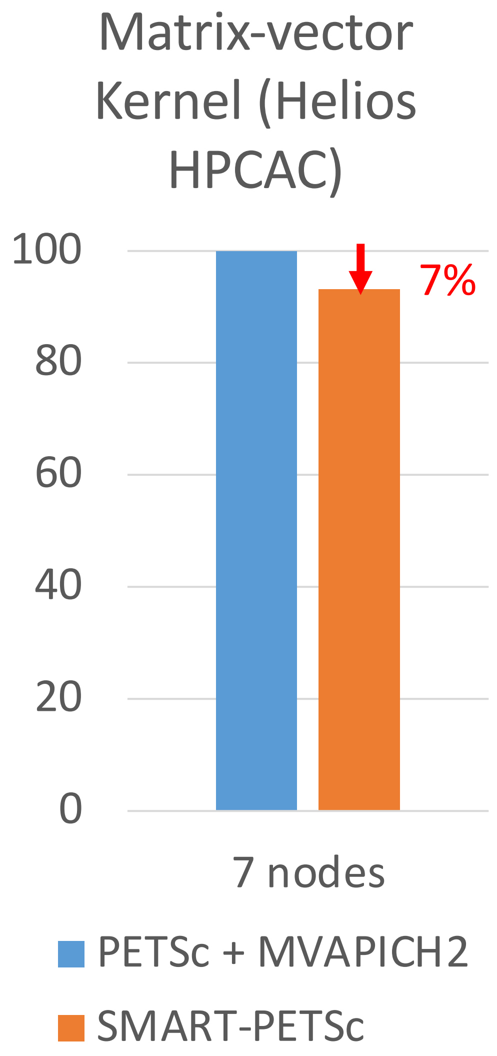


- Poisson problem (with non-periodic boundary condition)
- 10th order finite-difference spatial discretization (non-symmetric due to non-periodic boundaries)
- Encountered in fluid, solid, and heat transfer applications
- Up to 9% application-level benefit on TACC
- Different Krylov subspace solvers
- GMRES – Generalized Minimal Residual
- BICG – Biconjugate Gradient
- BCGS – Biconjugate Gradient-Stabilized

Optimizations: Finite Difference PETSc application

- Other CPUs
- Gary and Roberta – AMD EPYC, Helios – Intel Gold





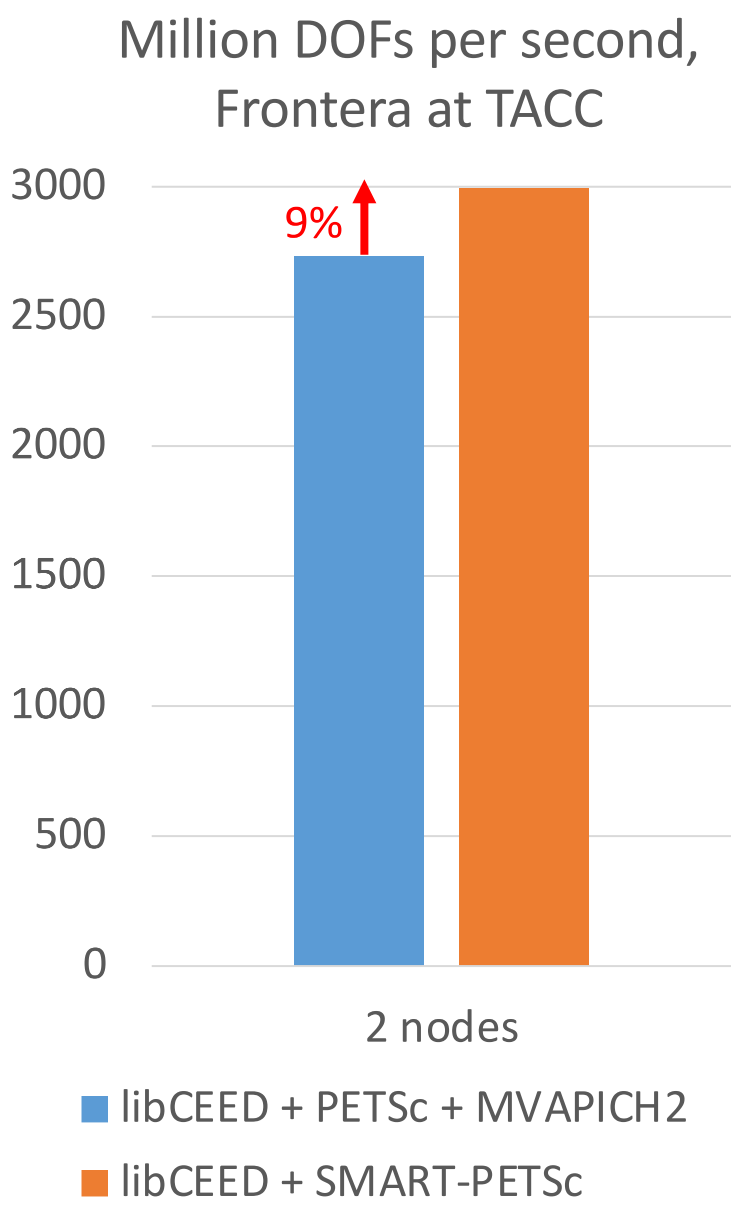
Processors	AMD EPYC
Cores/Node	16, 192
Memory/Node	384GB DDR-4
Network	Mellanox Infiniband, NDR-400

Gary and Roberta at OACISS

Processors	Intel Gold
Cores/Node	40
Memory/Node	192GB DDR-4
Network	Mellanox Infiniband, HDR-200

Helios at HPCAC

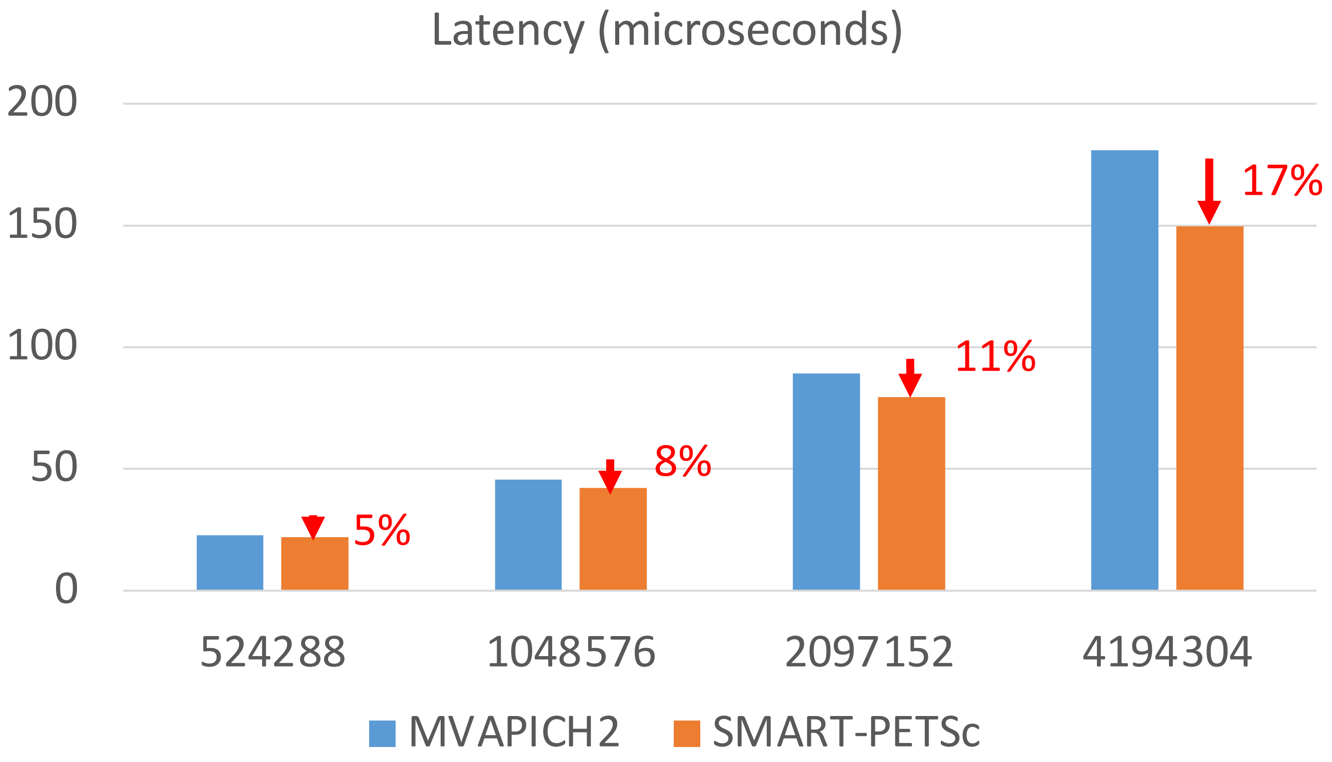
Optimizations: Finite Element PETSc application

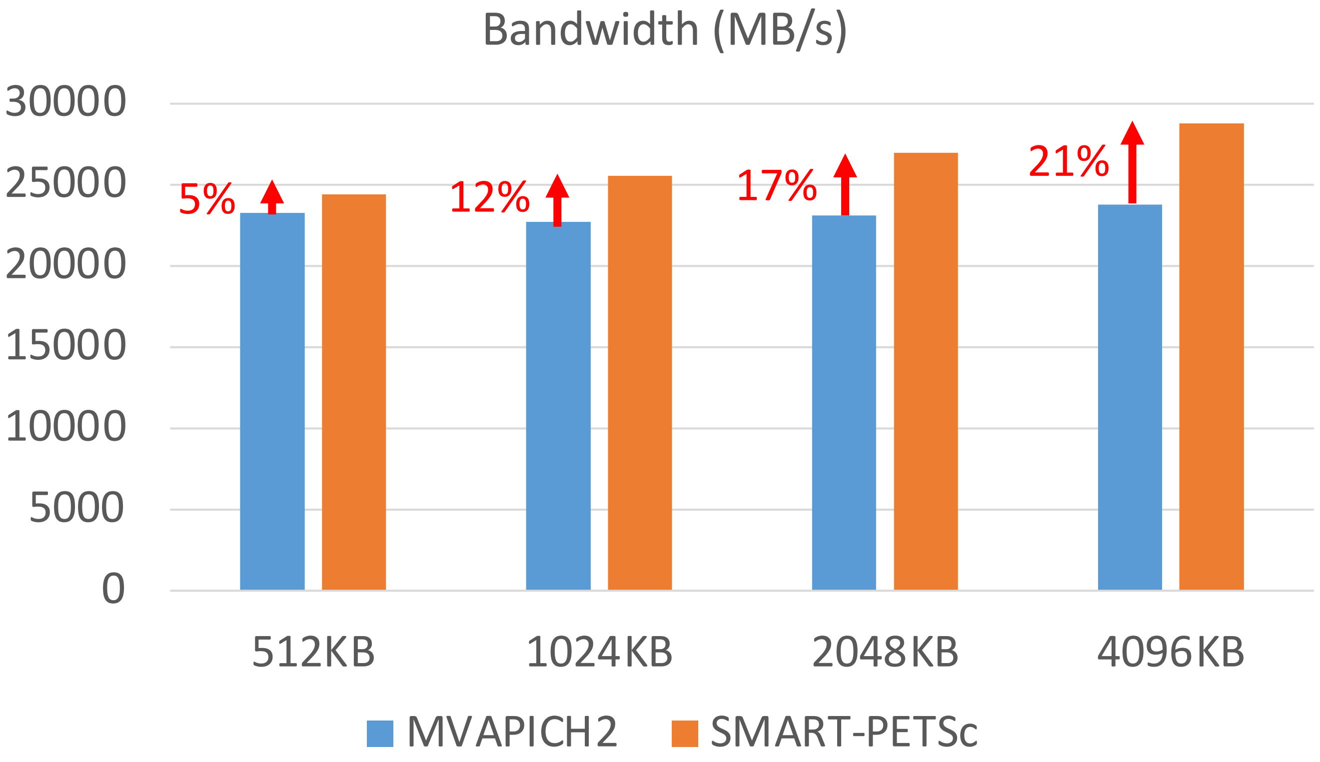


- libCEED
- Matrix-free back-end of many exascale high-order finite element libraries
- Sum-factorization, high throughput
- Uses PETSc tools to setup and perform MPI communication
- CPU, libxsmm, optimized blocked AVX512 instructions
- Bakeoff problem #2, conjugate gradient with mass matrix on a three-dimensional vector

Optimizations: Intra-node bandwidth on modern CPUs

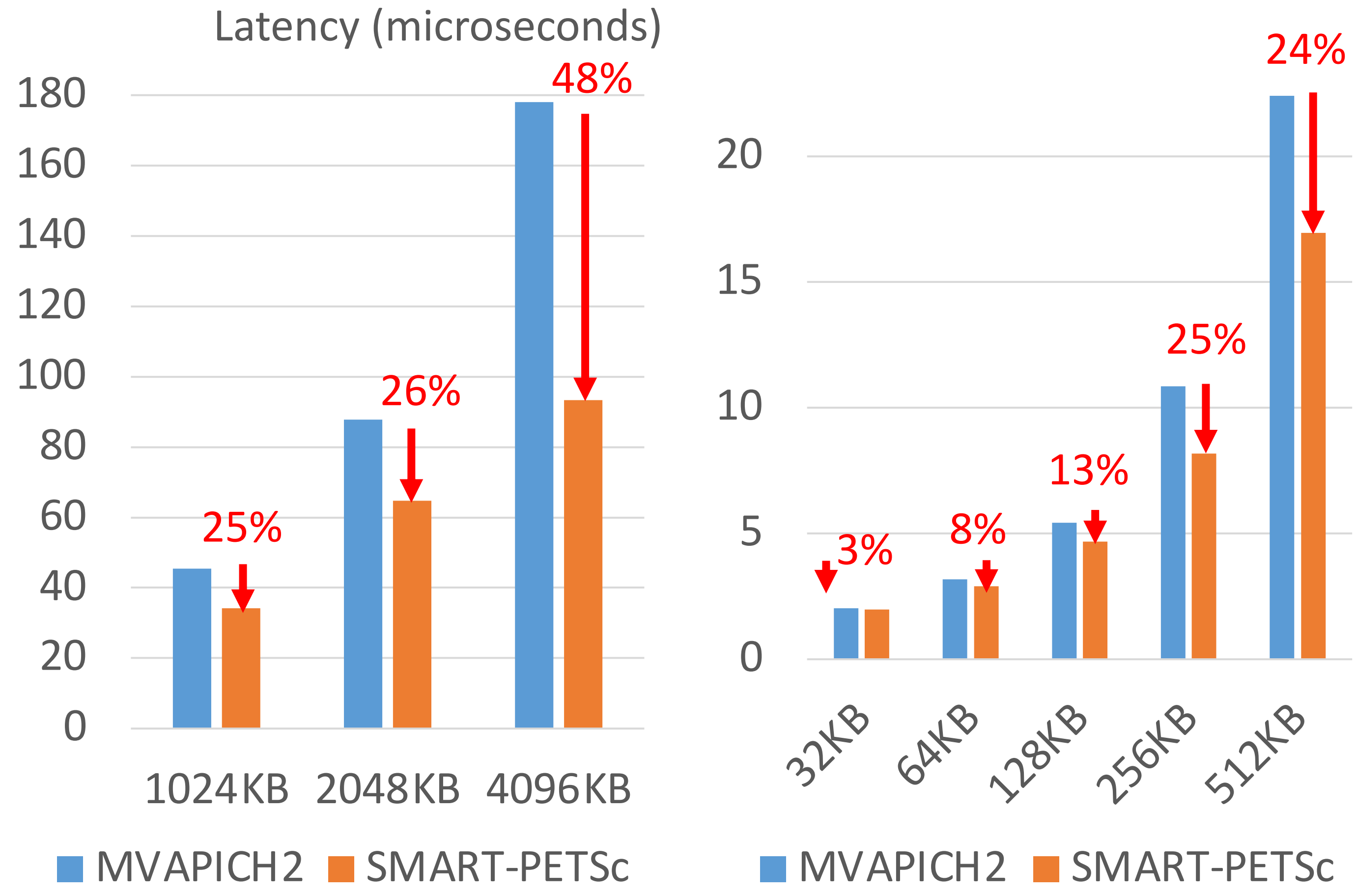
- AMD EPYC, MPI-only
- Up to 17% latency reduction and 21% bandwidth improvement for large message sizes
- osu_latency and osu_bw microbenchmark
- Environment variable MV2_SMART_PETSC_OPT=1 to turn on enhancement

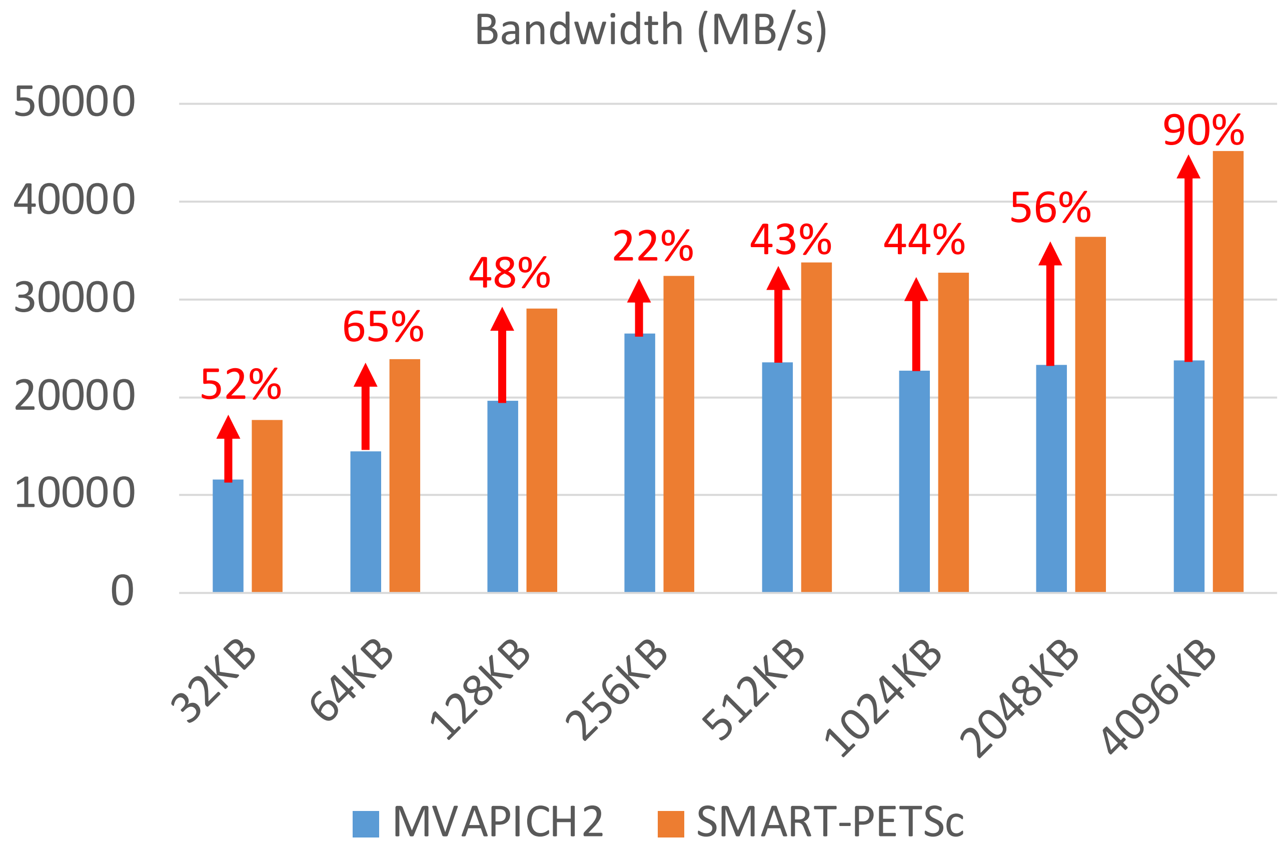




Optimizations: Intra-node bandwidth on modern CPUs

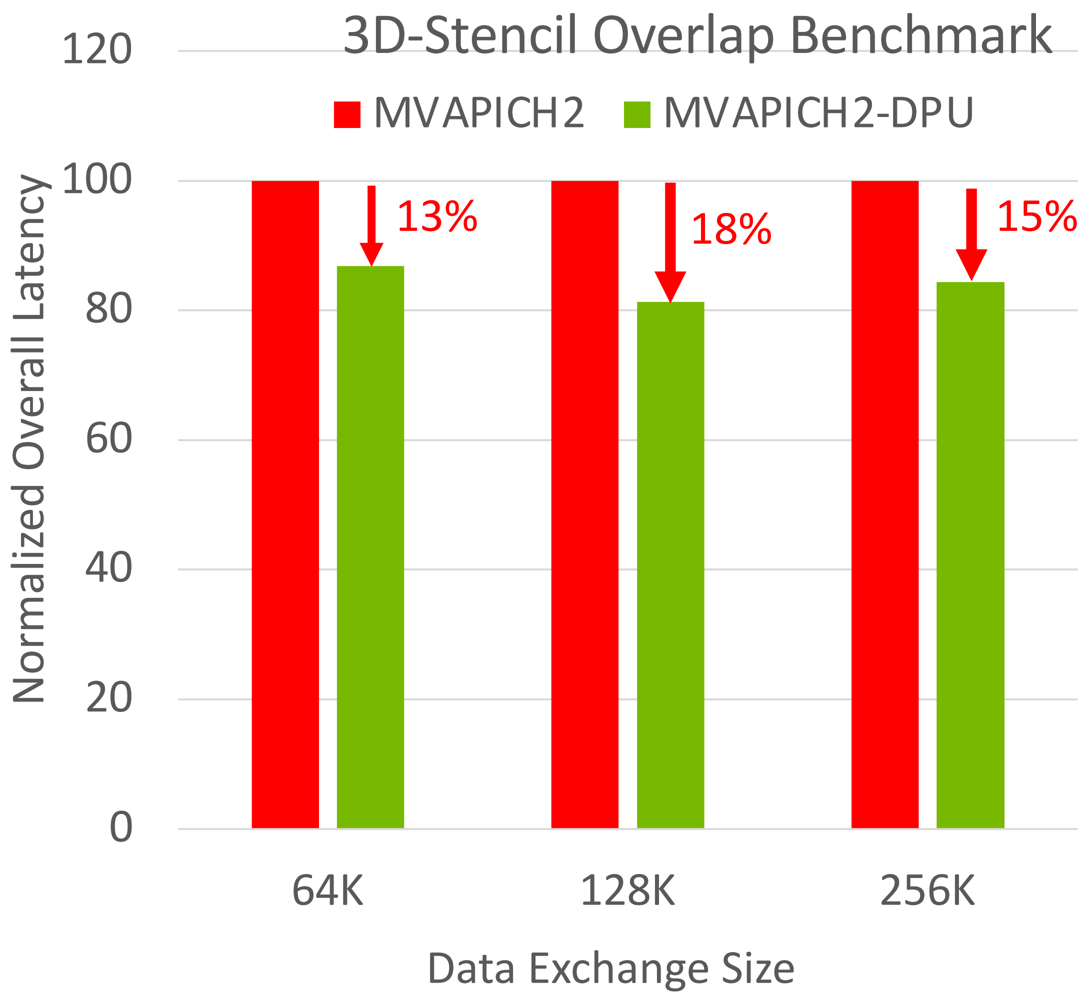
- MPI+OpenMP with AMD EPYC
- Up to 48% latency reduction and 90% bandwidth increase
- osu_bw microbenchmark
- Environment variable MV2_SMART_PETSC_OPT=1 to turn on enhancement



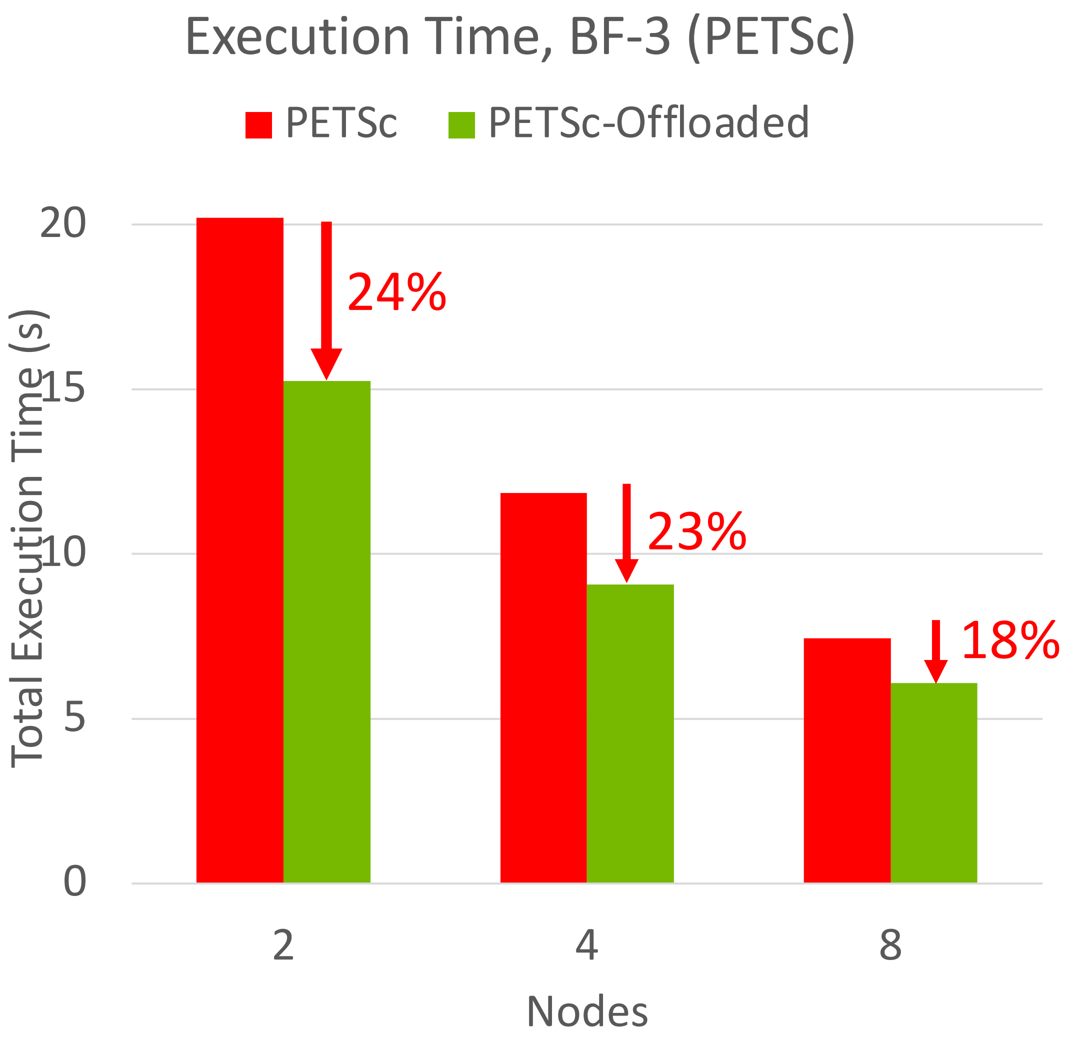


Optimizations: MVAPICH2-DPU with PETSc

- Using Bluefield DPU/SmartNICs to offload point-to-point, reduction and some computation



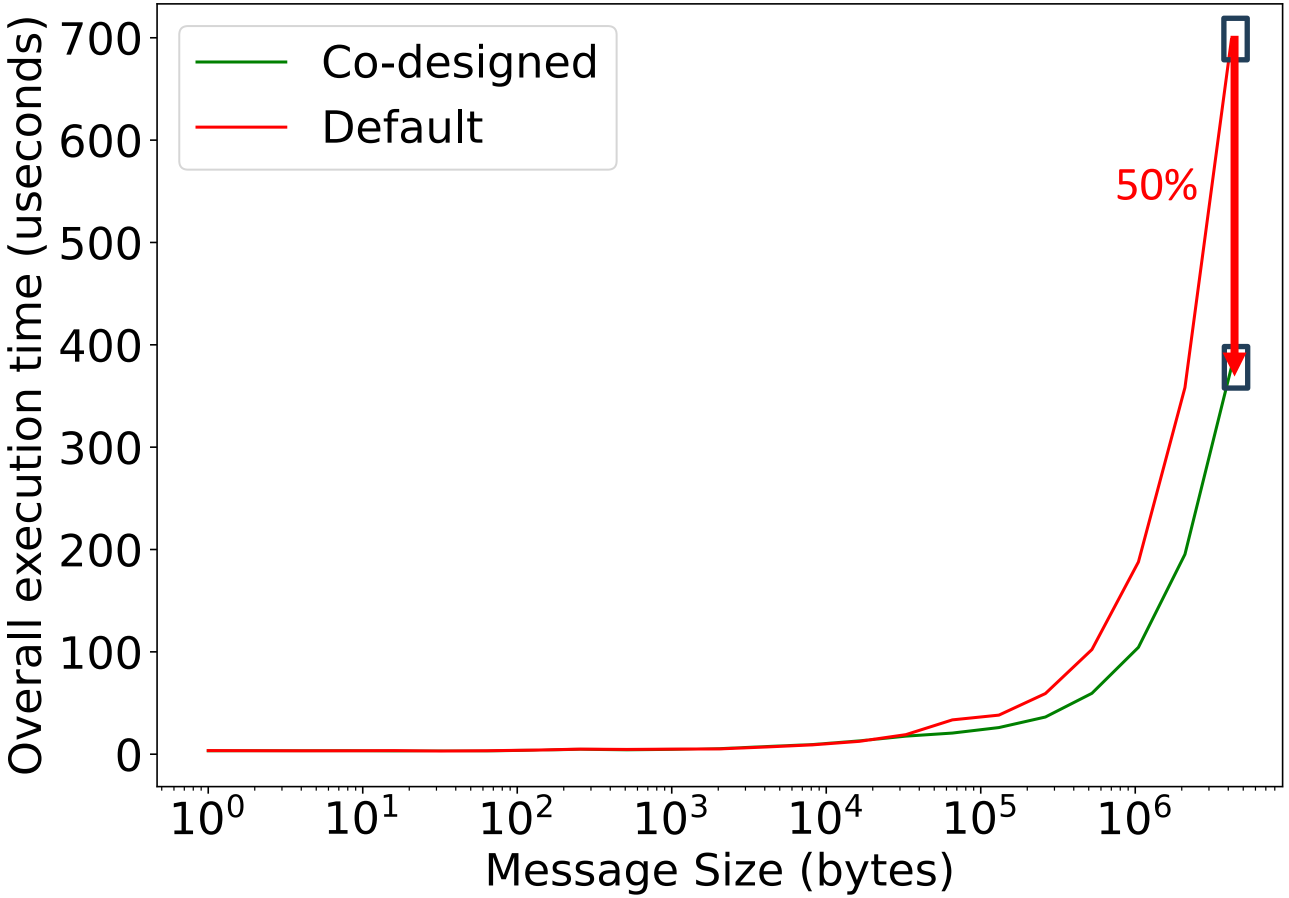
Communication pattern similar to PETSc DMDA



27-point Laplacian stencil PETSc application

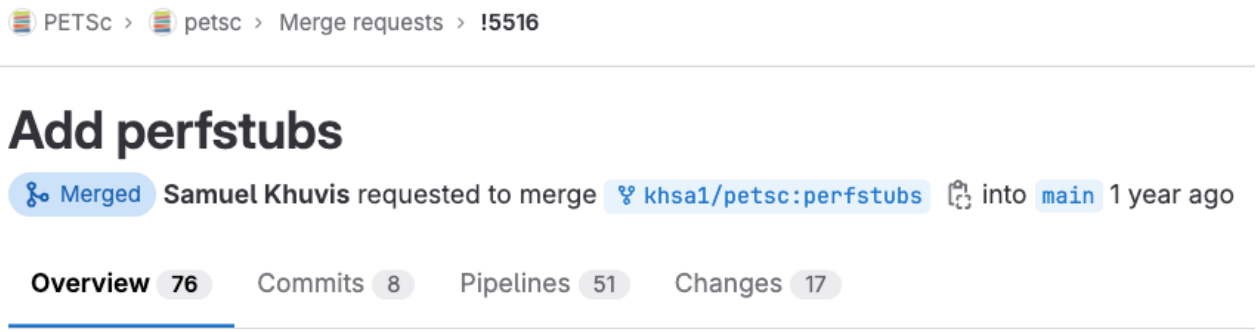
Optimizations: Co-designed rendezvous protocols

- Co-designed rendezvous protocols
- 3D-stencil benchmark (communication pattern similar to PETSc DMDA)
- Up to 50% performance benefit
- Demonstrates the maximum potential of such co-design enhancements
- Currently, porting it to PETSc



Profiling: TAU + PETSc via perfstubs

- Added interface for easy profiling of PETSc with TAU
- Contributed back to PETSc public repo
- Is now the de facto standard procedure to profile PETSc with TAU



Github merge request

```
Reading Profile files in profile.*

NODE 0;CONTEXT 0;THREAD 0:
---------------------------------------------------------------------------------------
%Time    Exclusive    Inclusive       #Call      #Subrs  Inclusive Name
              msec   total msec                          usec/call
---------------------------------------------------------------------------------------
100.0           26        1,838           1       41322    1838424 .TAU application
 73.2            1        1,345           2         168     672950 SNESSolve
 62.2            3        1,142           2        1282     571442 SNESJacobianEval
 62.0        1,136        1,138           2          76     569494 DMPlexJacobianFE
 60.1        0.046        1,105           1          32    1105001 Solve 1
 15.2           87          279           5       11102      55943 Mesh Setup
 13.2        0.315          241           1          32     241765 Solve 0
  7.8           80          144       38785       38785          4 MPI_Allreduce()
  7.0           69          128           6       43386      21491 DualSpaceSetUp
  6.2            1          114           4          54      28536 PCSetUp
  6.0           12          110           2         892      55407 PCSetUp_GAMG+
  3.9           70           70           1           0      70888 MPI_Init_thread()
  3.7           68           68       41747           0          2 MPI Collective Sync
  3.6            8           66           4        3536      16548 SNESFunctionEval
  2.6           45           48         171         171        281 MPI_Bcast()
  1.9           34           34        7836           0          4 MPI_Barrier()
  1.8        0.567           33           2          68      16912  GAMG Coarsen
```

TAU pprof result

Summary

- Demonstrate performance benefits from Smart-PETSc
- Matrix-vector multiplication kernel, a finite-difference application and a finite-element application
- Discussed some enhancements that have potential to further increase application performance
- Enhancements targeting GPUs will be the next set of features
- X-ScaleSolutions will be happy to interact with potential customers/collaborators

Thank You!

Sreevatsa (Sreev) Anantharamu

s.anantharamu@x-scalesolutions.com

contactus@x-scalesolutions.com

X-ScaleSolutions

http://x-scalesolutions.com/