Orchestrating **Hierarchical Algorithms** for **Complex Scientific Applications** on **OLCF Systems** with the **PaRSEC Runtime**

Qinglei Cao
Department of Computer Science

2025 OLCF User Meeting

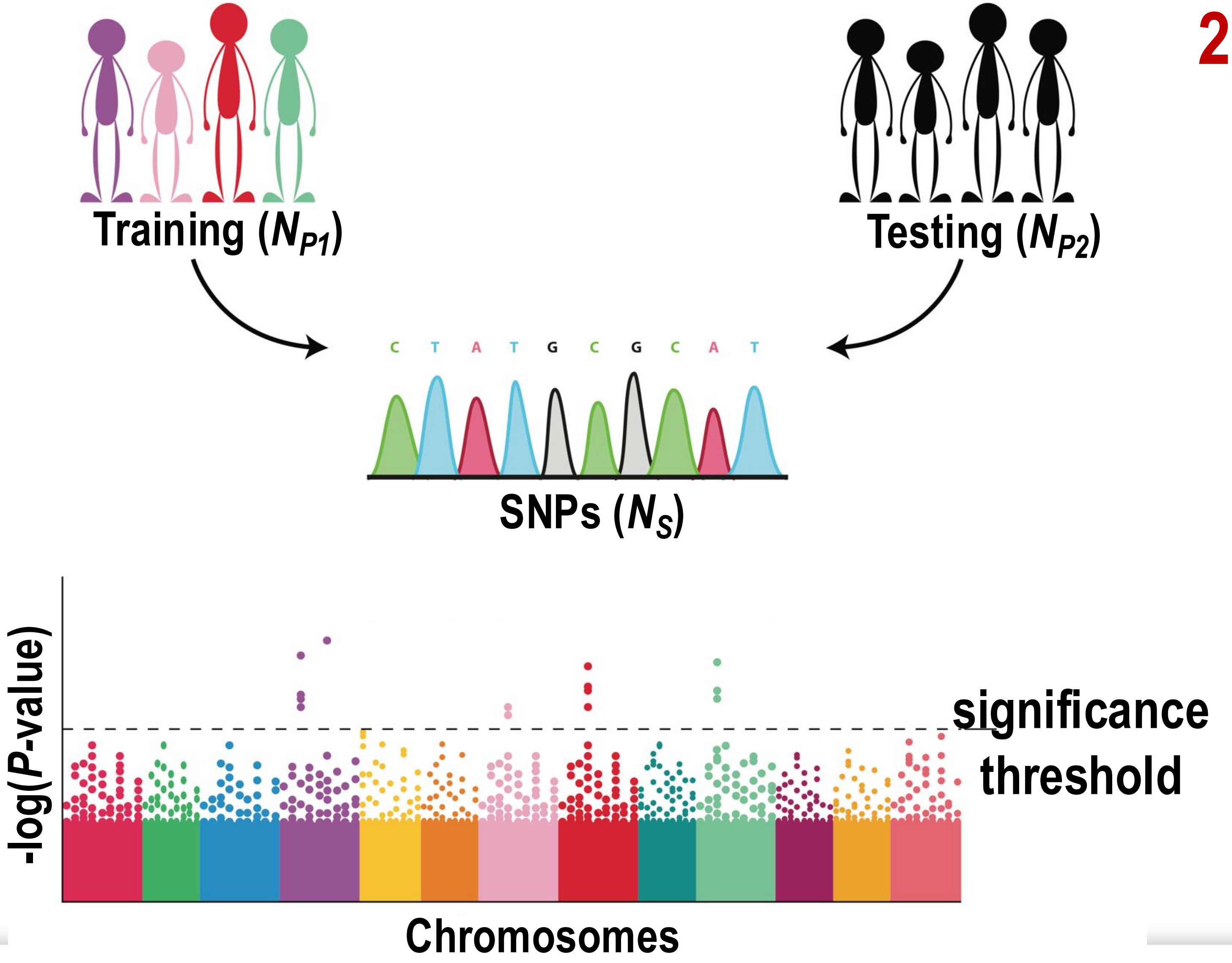
SAINT LOUIS UNIVERSITY™
EST. 1818

Thank You for OLCF's Supports

- CSC312: "2.3.1.09 STPM11-ParSEC: Distributed Tasking for Exascale "
- CLI180: "High-Resolution E3SM Land Model on GPUs"
- CLI188: "Saving PetaBytes in Earth System Model Outputs using Stochastic Approximations ”
- CLI194: "Saving PetaBytes in Earth System Model Outputs using Stochastic Approximations"
- CSC416: "Geostatistical Modeling and Prediction In Three Precisions"
- CSC574: "ICL DisCo"
- CSC612: "Optimizing PaRSEC on Frontier Supercomputer through Matrix Computations in Climate and Weather Prediction"
- CSC665: "Optimizing PaRSEC on Frontier Supercomputer"

Many Thanks to Cara Kennedy, Misty Abston, Lisa Mulig, and many others!!!

Our Journey One: Genome-Wide Association Studies

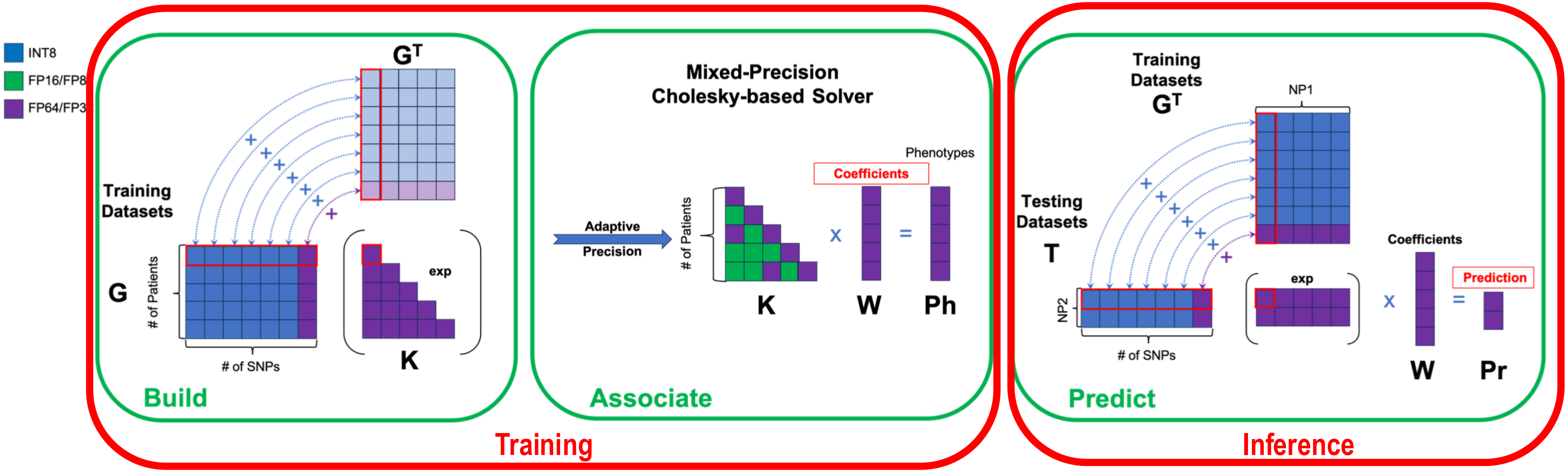


2024 Gordon Bell Prize Finalist

Paper

* SNP = single nucleotide polymorphism

Our Journey One: Genome-Wide Association Studies



Leveraging the INT8 / FP8 / FP16 / FP32 / FP64 KRR-based multivariate GWAS for genetic epistasis.

SAINT LOUIS UNIVERSITY

Our Journal Two: Exascale Climate Emulator

- Developed and validated own climate emulator
  - emulates up to 54.5 million spatial locations across the globe with spatial resolution of 0.034° (3.5 km) at an hourly resolution for 35 years (1988-2022)
- Addressed resolution limitations of existing emulators
  - compresses 2D data on sphere with fast SHTs
  - filters high frequency noise
  - democratizes climate realizations (workstations)
  - plays to architectural strengths (dense matrices)
  - lowers storage barrier

2024 Gordon Bell Prize for Climate Modelling

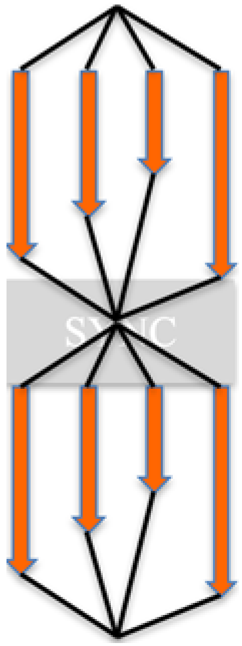
Motivations for Mixed Precision

- Many matrices arising in applications have blocks of relatively small norm and can be replaced with reduced precision.
- Computational: faster time to solution
  - ✓ lower energy consumption and higher performance, especially by exploiting heterogeneity

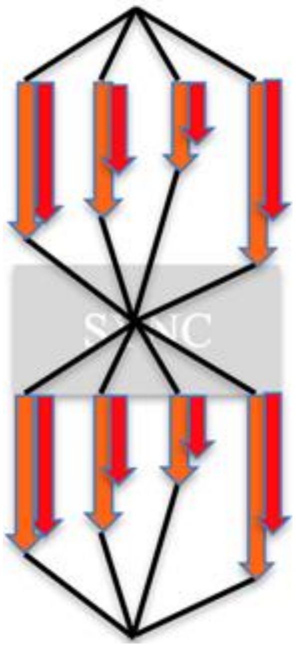
Peak Performance in TF/s	V100 NVLink	A100 NVLink	H100 SXM
FP64	7.8	9.7	34
FP32	15.7	19.5	67
FP64 Tensor Core	16x	19.5	67
TF32 Tensor Core		156 16x	494.7 15x
FP16 Tensor Core	125	312	989.4

- Mixed precision algorithms have a long history, e.g., iterative refinement (1963, Wilkinson), where multiple copies of the matrix are kept in different precisions for different purposes.
- There are many such new algorithms; see Higham & Mary, Mixed precision algorithms in numerical linear algebra, Acta Numerica (2022); up to 5 precisions!

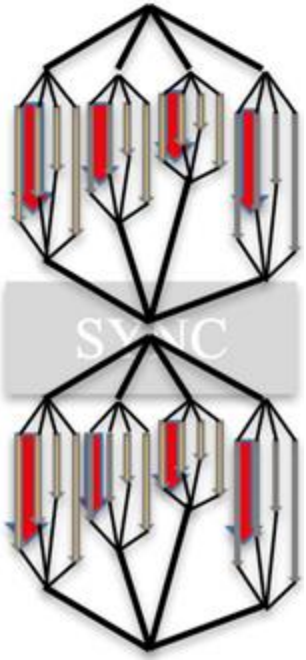
Programming paradigms



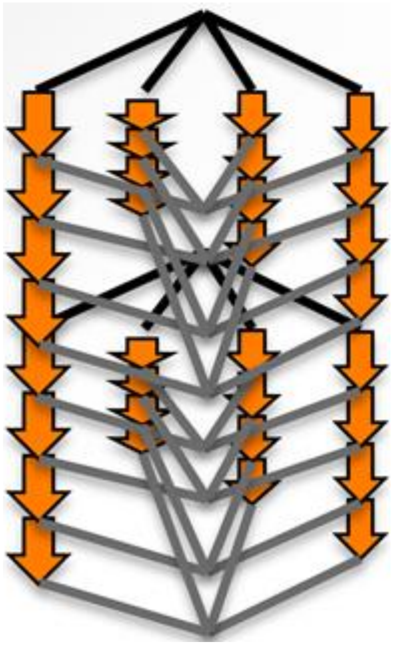
BSP and early message passing



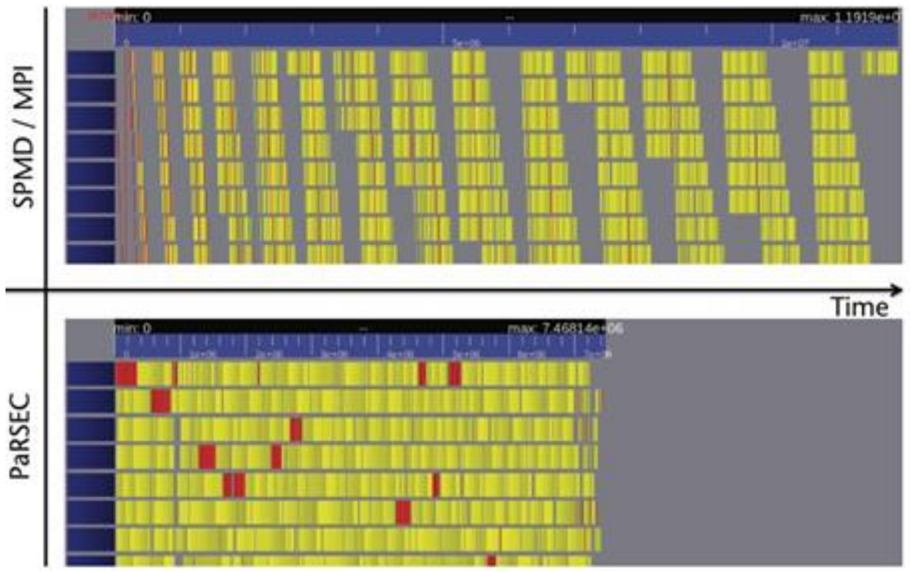
MPI + X



MPI + X + Y



Task-base runtime



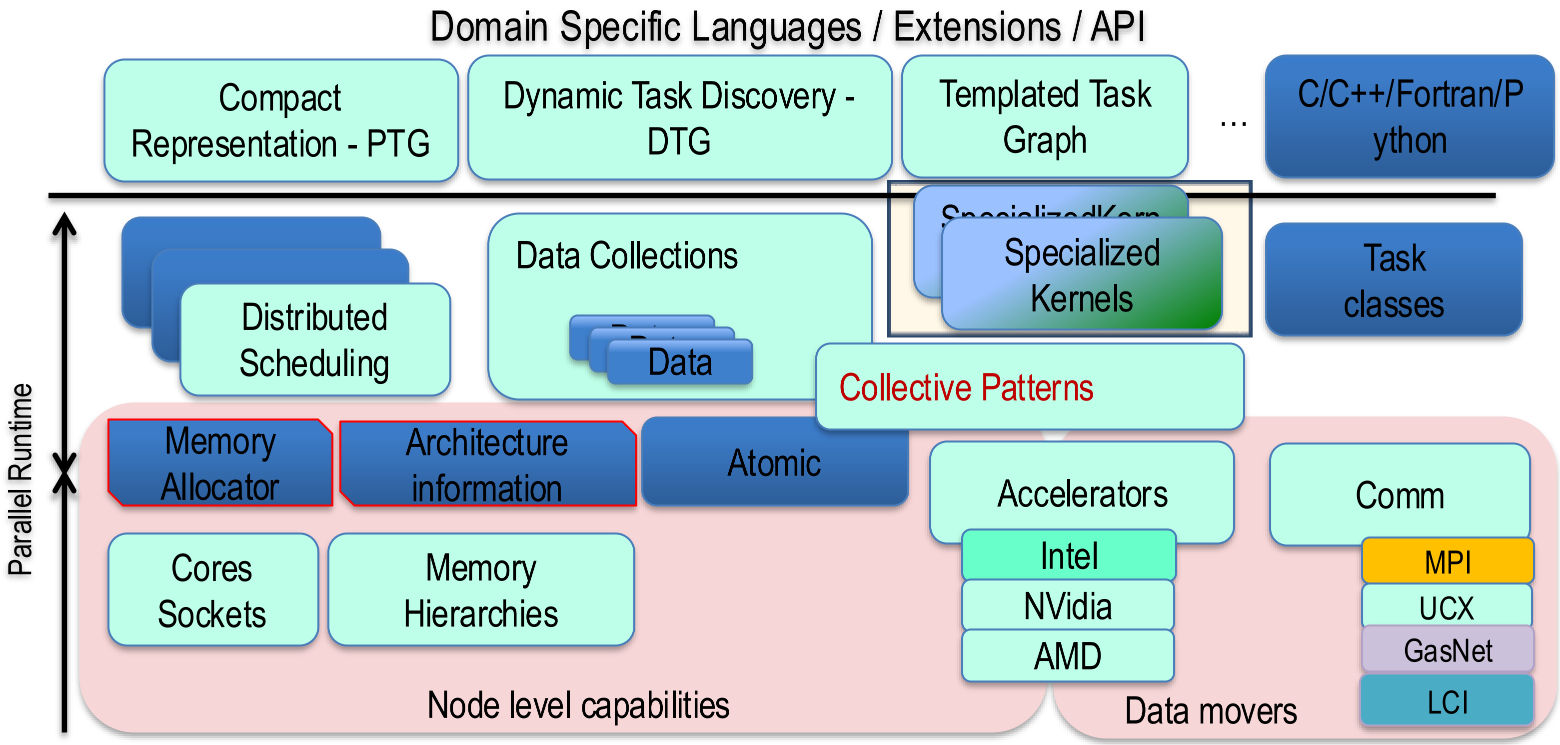
Comparison of execution traces for the same algorithm using the single-program, multiple data message passing interface (SPMD/ MPI) programming model and the dataflow model

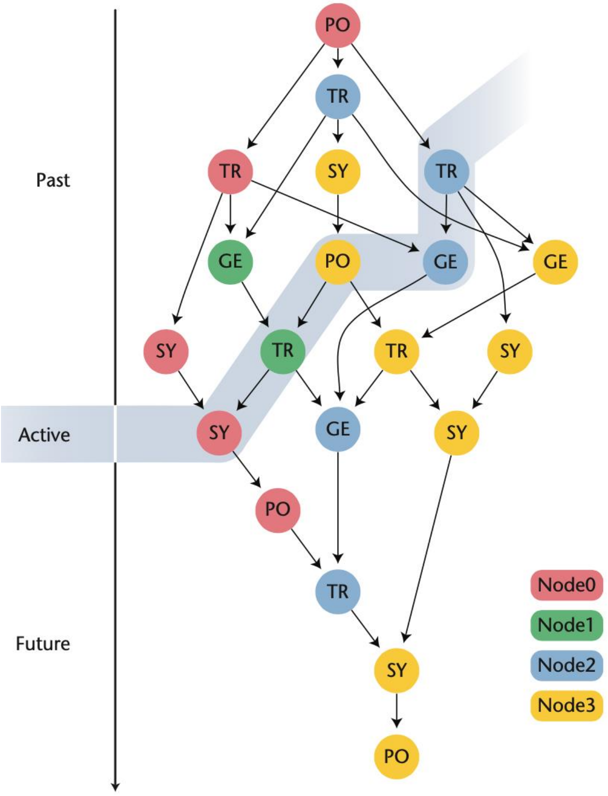
- Harder than sequential programming
- Users need to express parallelism with minimum synchronization points
- Managing shared memory, distributed memory and heterogeneous architectures

Many task-based runtime: OmpSs, OpenMP, StarPU, Legion, **PaRSEC**
...

PaRSEC

A generic runtime system for asynchronous, architecture aware scheduling of fine-grained tasks on distributed many-core heterogeneous architectures.

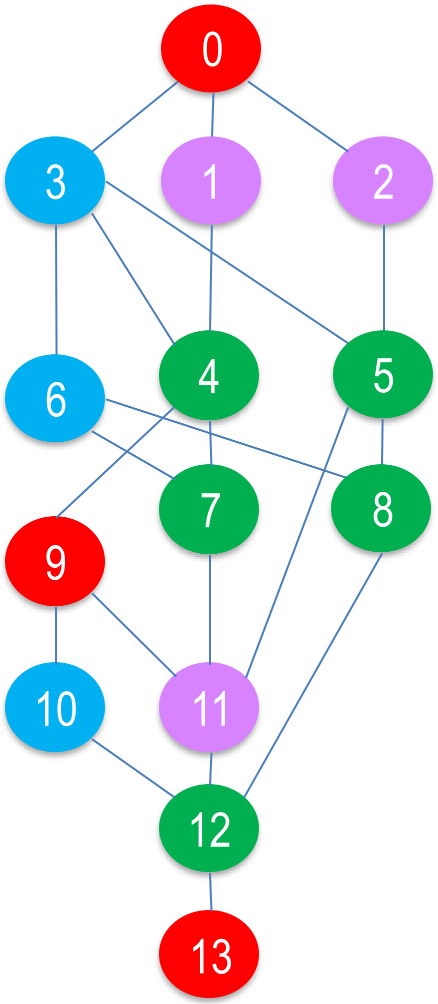


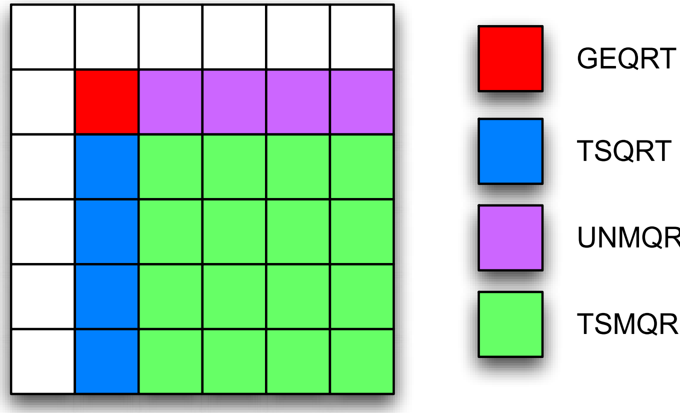


Directed acyclic graph (DAG)

Dynamic Task Discovery (DTD)

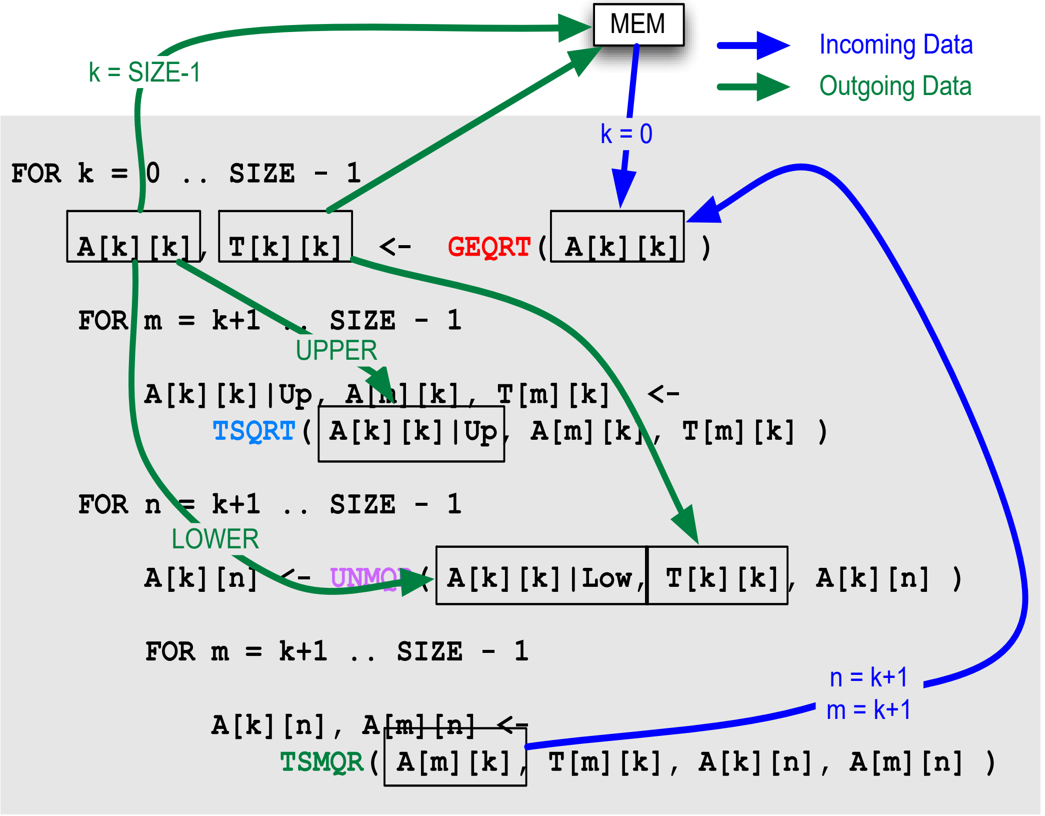
```
for( k = 0; k < SIZE; k++ ) {
    parsec_insert_task( "GEQRT",
              DATA_OF(A, k, k), INOUT|AFFINITY,
              DATA_OF(T, k, k), OUTPUT|TILE_RECT)

    for( n = k+1; n < SIZE; n++ )
     parsec_insert_task( "UNMQR",
               DATA_OF(A, k, k), INPUT|TILE_L,
               DATA_OF(T, k, k), INPUT|TILE_RECT,
               DATA_OF(A, k, n), INOUT|AFFINITY)

    for( m = k+1; m < SIZE; m++ ) {
     parsec_insert_task( "TSQRT",
               DATA_OF(A, k, k), INOUT|TILE_U,
               DATA_OF(A, m, k), INOUT|AFFINITY,
               DATA_OF(T, m, k), OUTPUT|TILE_RECT)

     for( n = k+1; n < SIZE; n++ ) {
         parsec_insert_task( "TSMQR",
                   DATA_OF(A, k, n), INOUT,
                   DATA_OF(A, m, n), INOUT|AFFINITY,
                   DATA_OF(A, m, k), INPUT,
                   DATA_OF(T, m, k), INPUT|TILE_RECT)
     }
    }
}
```

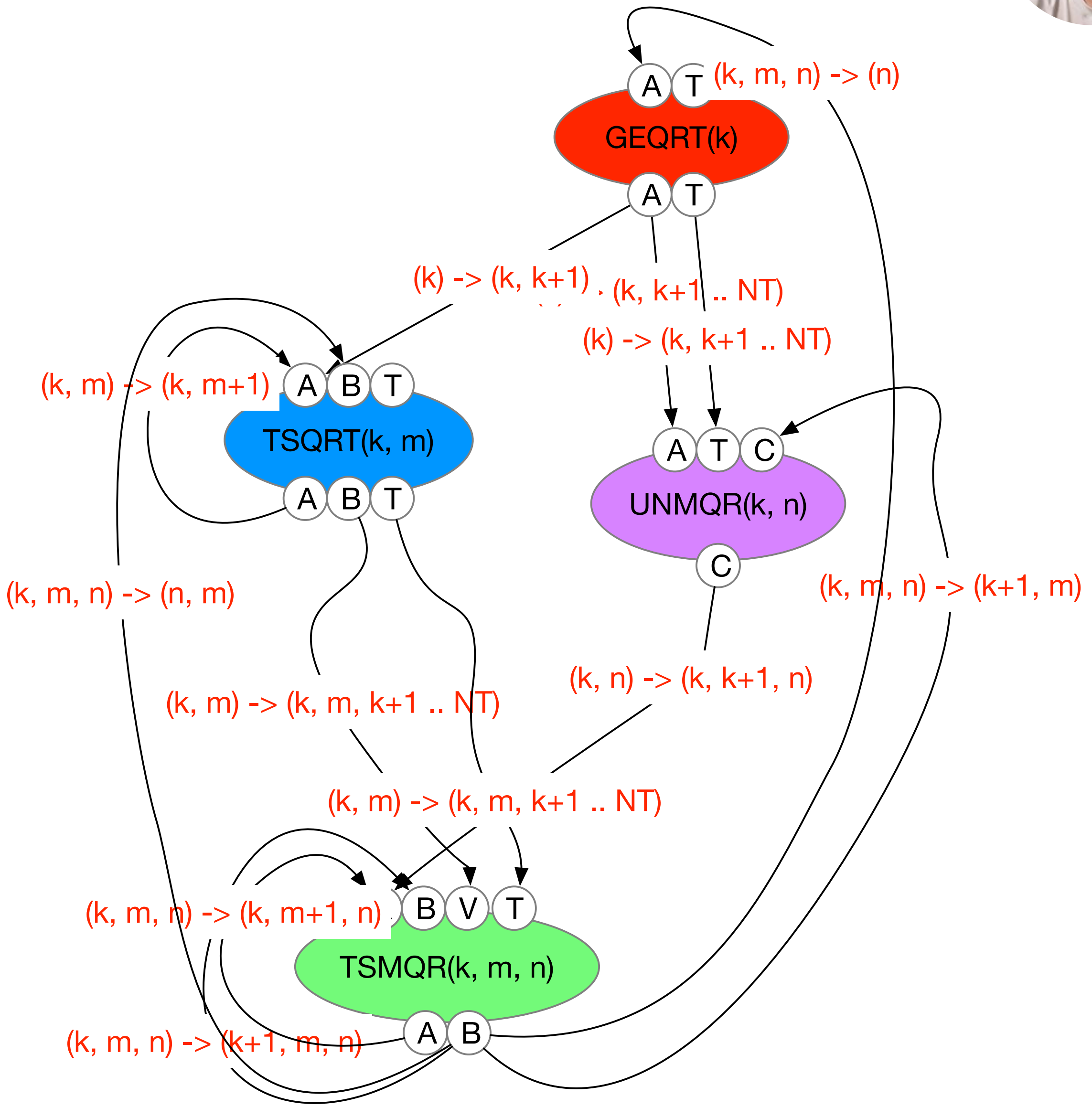




SAINT LOUIS UNIVERSITY

DSL: Parameterized Task Graph (PTG)

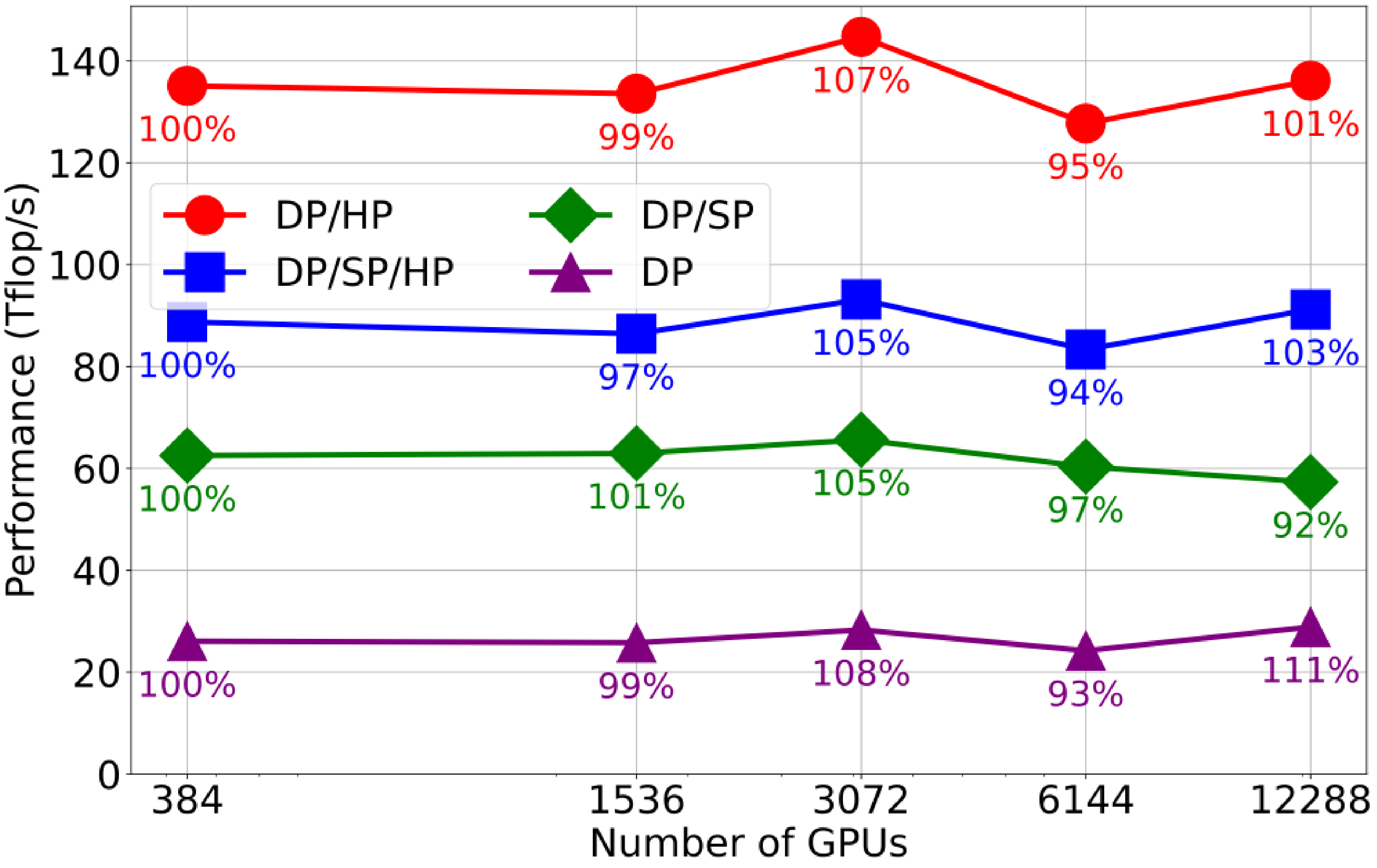




SAINT LOUIS UNIVERSITY

The Portable Software Stack

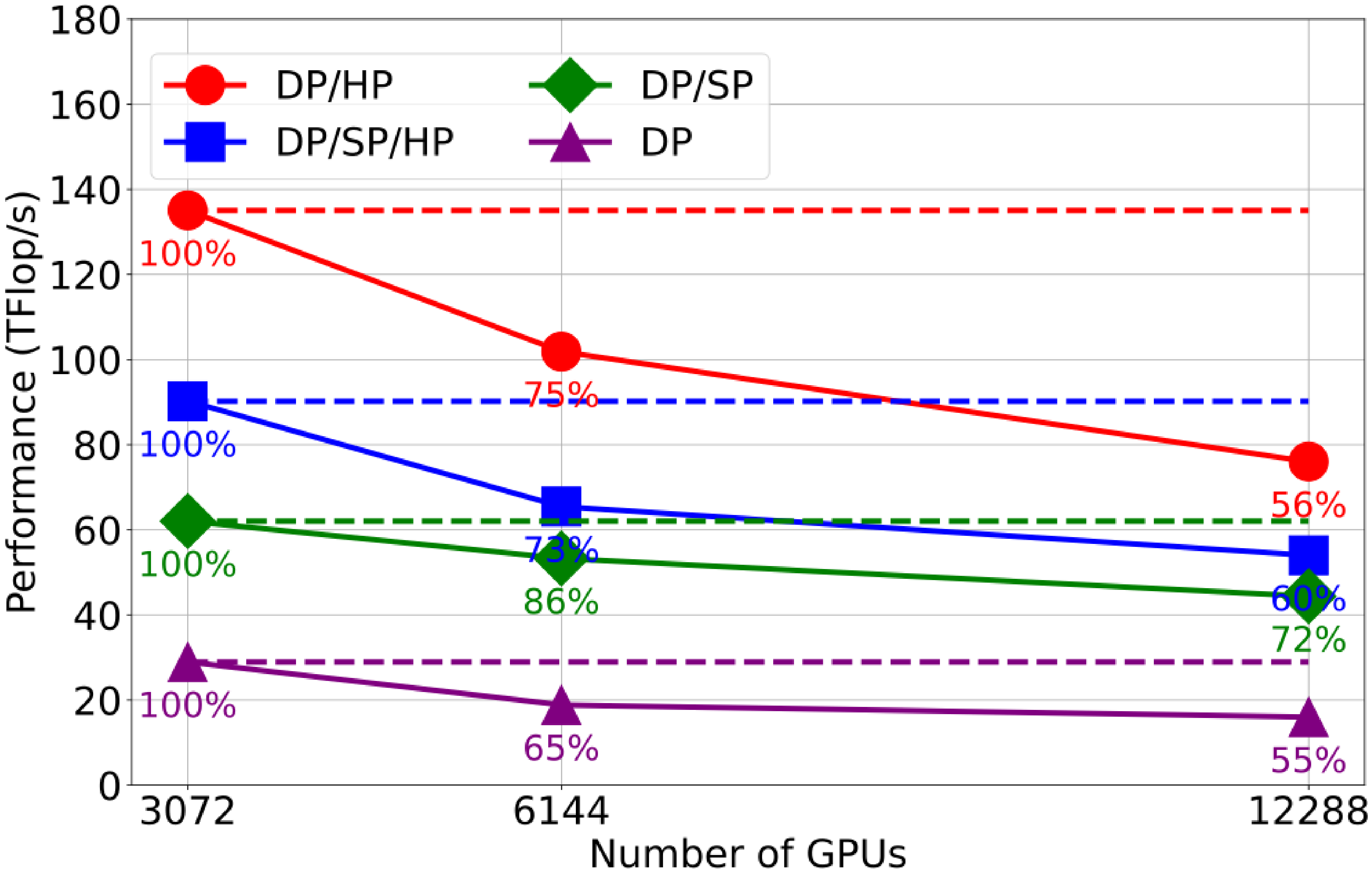
Weak Scalability on 12,288 GPUs of Summit



- **3 mixed-precision variants**
- Up to **2,048 nodes** of Summit **(12,288 NVIDIA V100 GPUs)**
- **Excellent weak scaling efficiency**

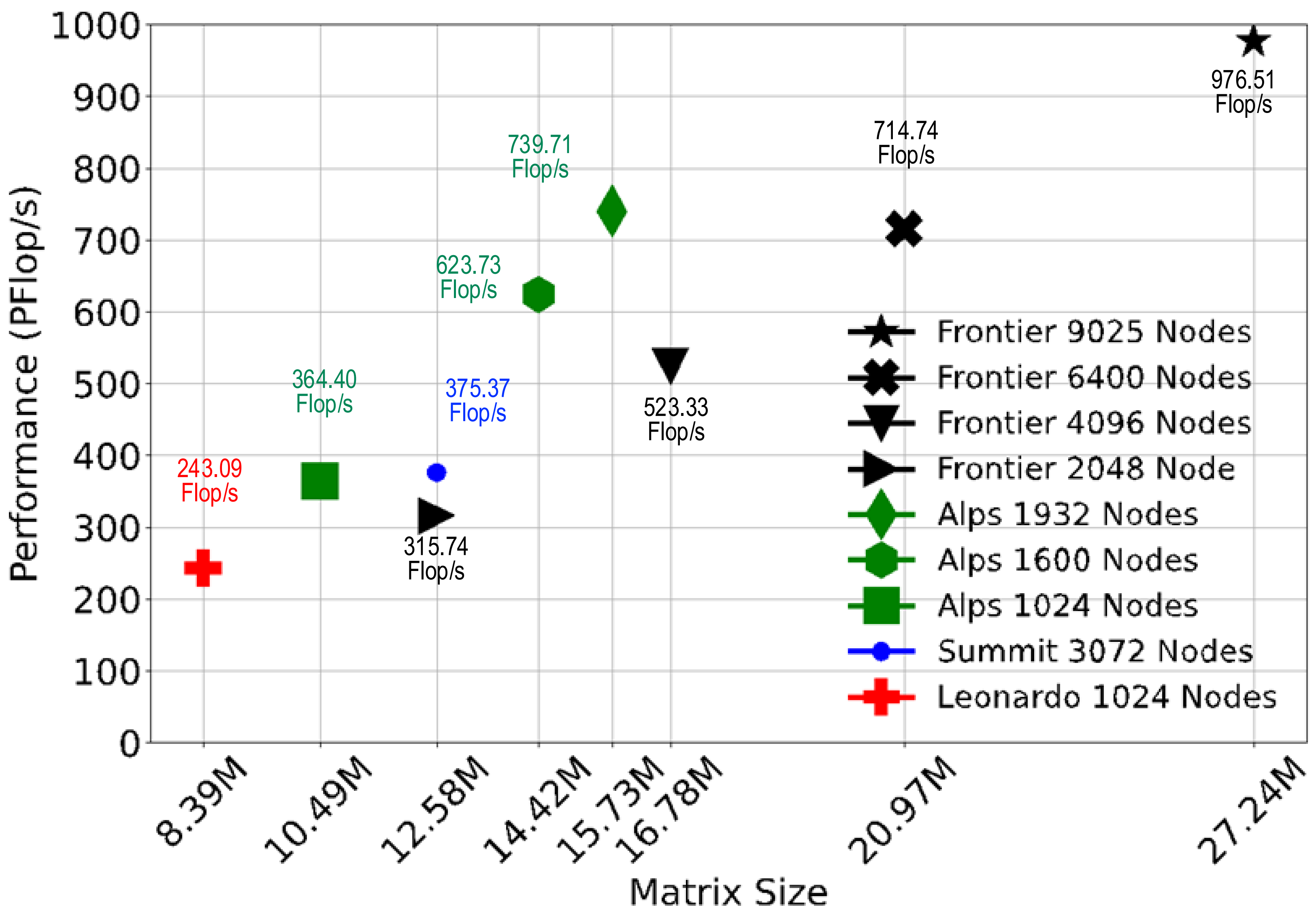
SAINT LOUIS UNIVERSITY

Strong Scalability on 12,288 GPUs of Summit



- **3 mixed-precision variants**
- Up to **2,048 nodes** of Summit **(12,288 NVIDIA V100 GPUs)**
- Up to **72% strong scaling efficiency with 12,288 GPUs**

Exascale runs



Performance of largest runs on Summit, Leonardo, Alps, and Frontier; with additional run-up points on Alps and Frontier, all using the DP/HP precision variant.

Summary

- **Innovations**: Introduces mixed-precision computation strategies optimized for different GPU generations, supported by the PaRSEC runtime system, to efficiently balance computation, communication, and memory footprint
- **Performance**: Supports many GPU-based systems as well as CPU-based systems. It has demonstrated significant computational performance across multiple exascale platforms:
  - 0.976 EFlop/s on 9,025 nodes of Frontier
  - 0.375 EFlop/s on 3,072 nodes of Summit
- **Impact**: Enables domain scientists to next-generation's applications with significantly reduced computational cost

Thank You!

Exascale Genomics

Exascale Climate Emulator

qinglei.cao@slu.edu