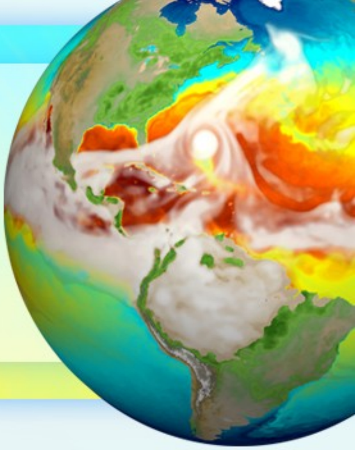


Task Placement for Reduced Communication Costs

J. Austin Ellis and Karen Devine

Sandia National Laboratories

Albuquerque, NM

E3SM All Hands Meeting

20 November 2019

E3SM Energy Exascale Earth System Model

Sandia National Laboratories is a multi-mission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, In., for the U.S. Department of Energys National Nuclear Security Administration under contract DE-NA0003525. This paper describes objective technical results and analysis. Any subjective views or opinions that might be expressed in the paper do not necessarily represent the views of the U.S. Department of Energy or the United States Government. SAND2019-14216 C

U.S. DEPARTMENT OF ENERGY

Dragonfly Task Placement Background

Objective: Minimize *distance* messages must travel by “mapping” frequently communicating MPI tasks to nearby nodes in allocation.

Contributions:

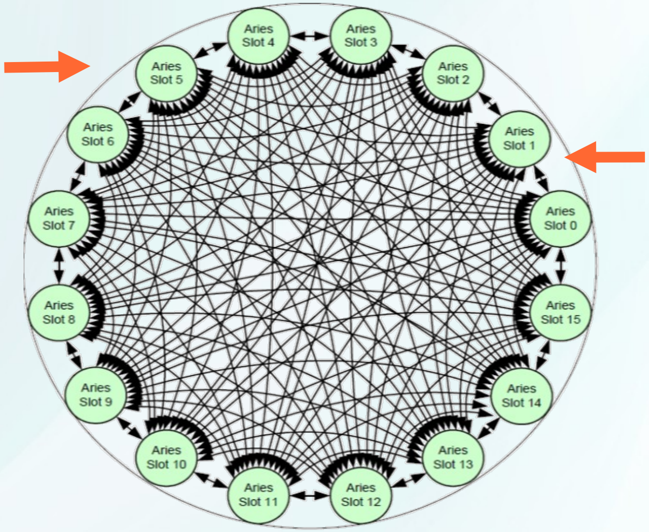
New dragonfly task placement algorithm inside Trilinos’ Zoltan2 package.

- Use high dimensional coordinate transformation to represent all-to-all connections
- Use coordinate stretching to match bandwidths and common congestion.

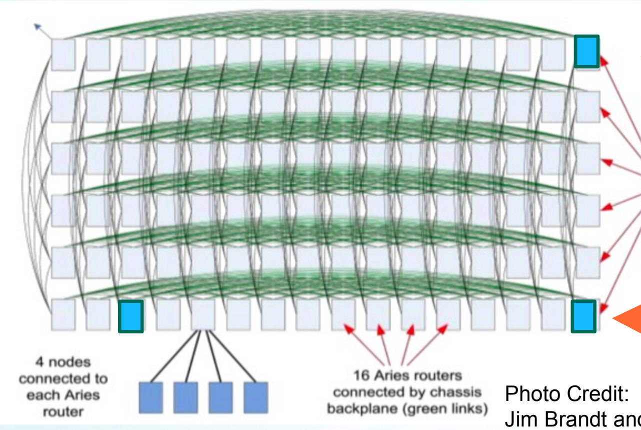
Added new capabilities to Trilinos Zoltan2’s Multi-Jagged(MJ) geometric coordinate partitioner.

- Will compute a nonuniform partitioning based on a user provided distribution
- Ensures the “group” dimension is fully partitioned first

All-to-all between “groups”



First partition highest level fully

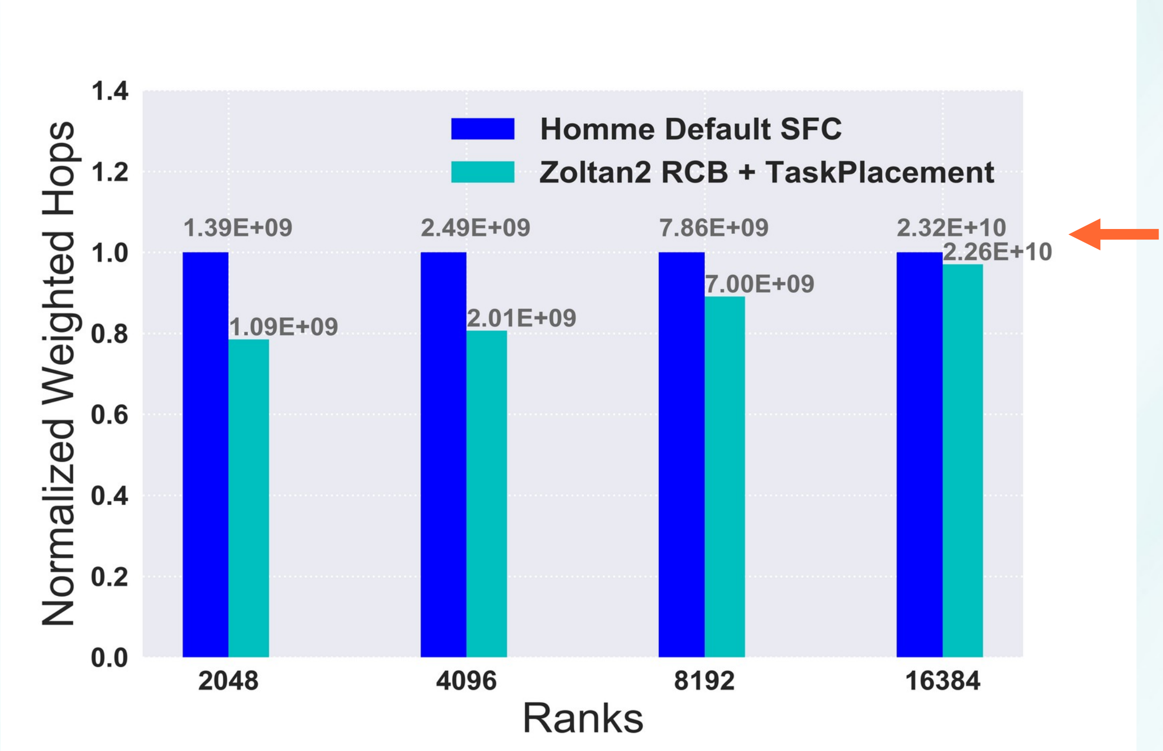

6 chassis connected by cables (black links) to form a two-cabinet group
4 nodes connected to each Aries router
16 Aries routers connected by chassis backplane (green links)

All-to-all in rows and cols

MPI messages travel “hops” through the network

Photo Credit:
Jim Brandt and Ann Gentile, SNL

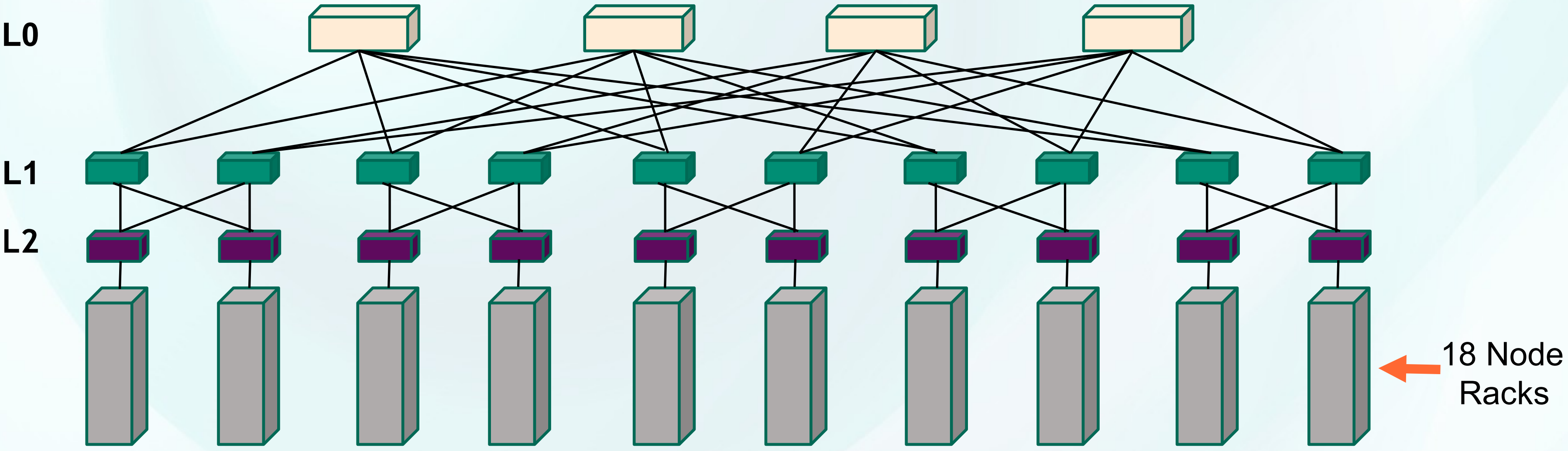
Preliminary Numerical Results from Theta



Near strong scaling limit for ne120

- **Test Case**: ALCF's Theta machine, HOMME v1 ne120 benchmark, 16 ranks per node
- Geometric task placement increases task locality in the network, decreasing distances messages must travel
- Weighted hops decreases by **up to 20%**
- Observed decreased runtime variability (50% on Cori, harder to quantify on Theta)

Future Work: Fat Tree Task Placement on Summit



Row Col Node

- Obtain Summit node floor location using gethostname(): **[A-H][01-36]n[01-18]**
- Transform floor coordinates to network **switch** “neighborhoods”
  - Ex. All nodes in **A01**-**A18** connected in a 3-hop neighborhood, **A19**-**A36** are a separate 3-hop neighborhood
- Use recursive 2-level nonuniform MJ partitioning (L1 → L2)
- Targeting full system scale runs for HOMMEXX

Thank you!

E3SM
Energy Exascale
Earth System Model

FASTMATH