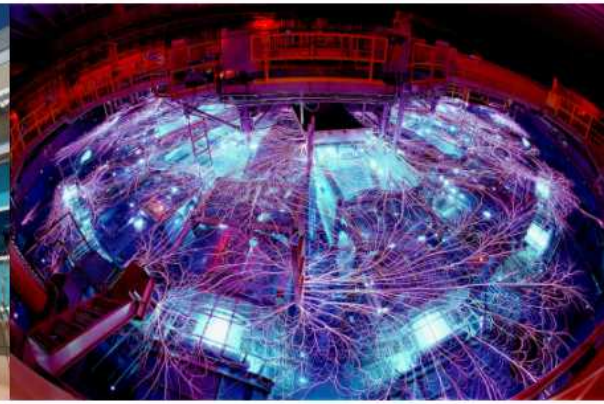
Exceptional service in the national interest

Sandia National Laboratories

SAND2019-3050PE





Reducing E3SM Communication through Task Mapping

J. Austin Ellis & Karen D. Devine

19 March 2019

U.S. DEPARTMENT OF ENERGY

oratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, neywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract

Task Mapping Background

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

Extreme-scale systems:

- Allocations may be sparse and spread far across the network
- Communication messages can travel long routes
- Network links may become congested by competing traffic

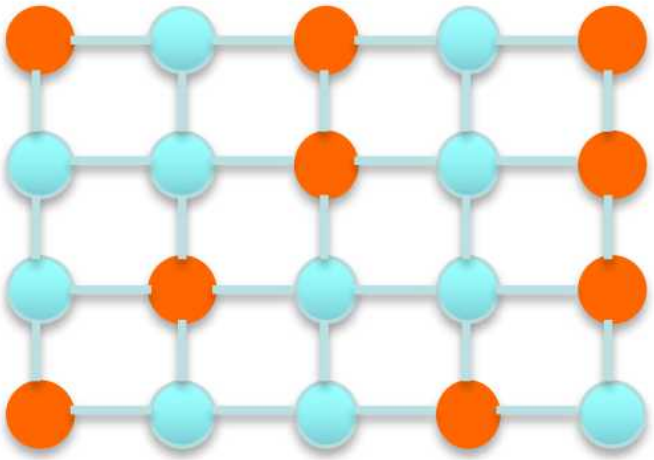


Figure: Non-contiguous node allocation in a mesh network.

Previous Results I

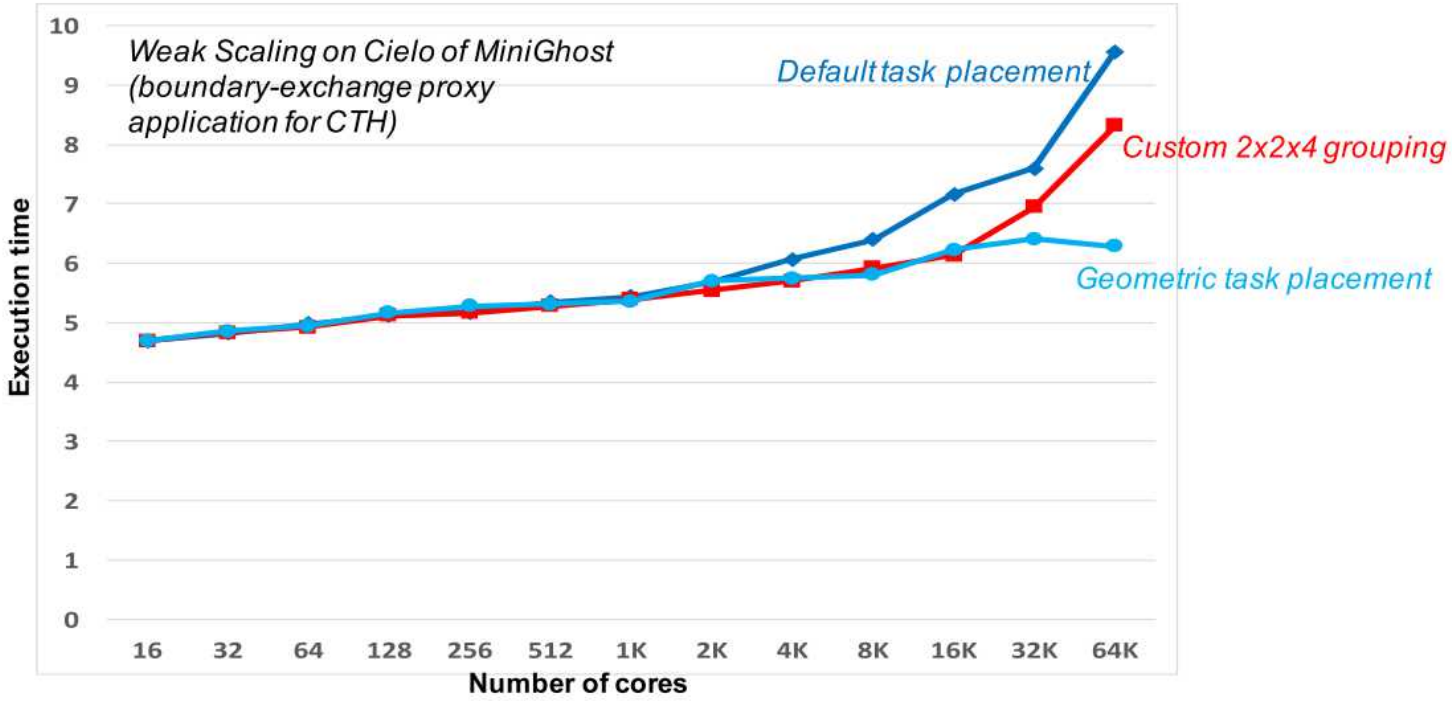


Figure: Scaling study of task mapping for the miniapp MiniGhost on Cielo (torus).

Previous Results II

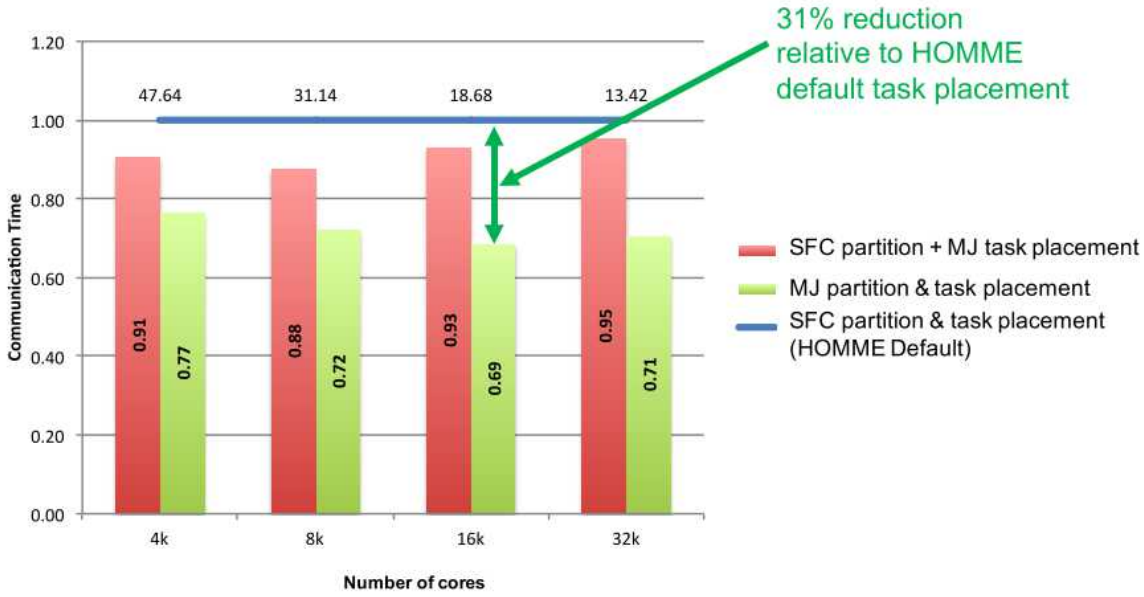


Figure: Communication reduction in HOMME through task mapping on Mira (contiguous torus).

- Reduced HOMME communication by 18% on Titan (torus) with 86k cores [Deveci et al., 2019]

Zoltan(2) and Partitioning [Boman et al., 2012]

Zoltan(2):

- Trilinos package for partitioning, load balancing, ordering, coloring, and other graph/combinatorial algorithms

Task Mapping Method:

- Represent both the processor allocation and the application communication as graphs or set of coordinates
- Partition both representations using Zoltan
- Map application parts to processor parts for reduced communication

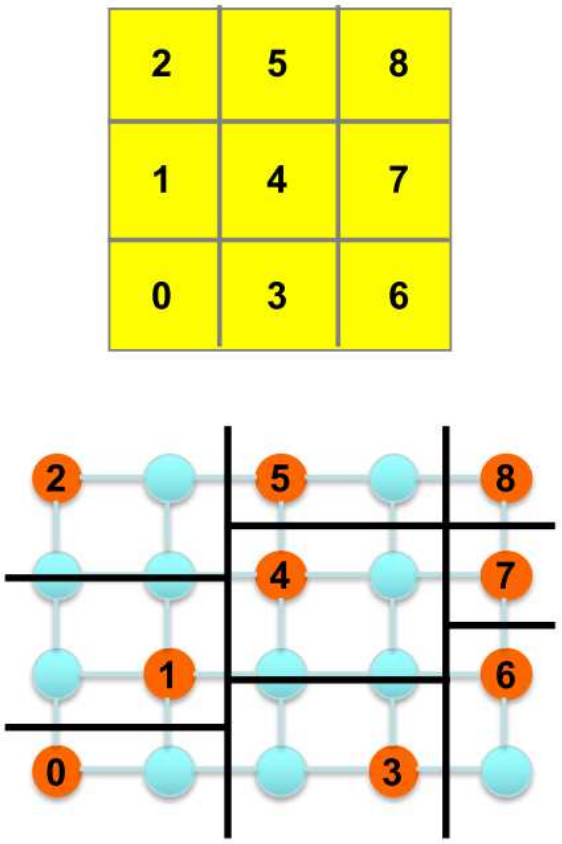


Figure: Partitioning of tasks (above) and processor allocation (below).

Cray Aries Interconnect (Dragonfly) and Coordinate Representation

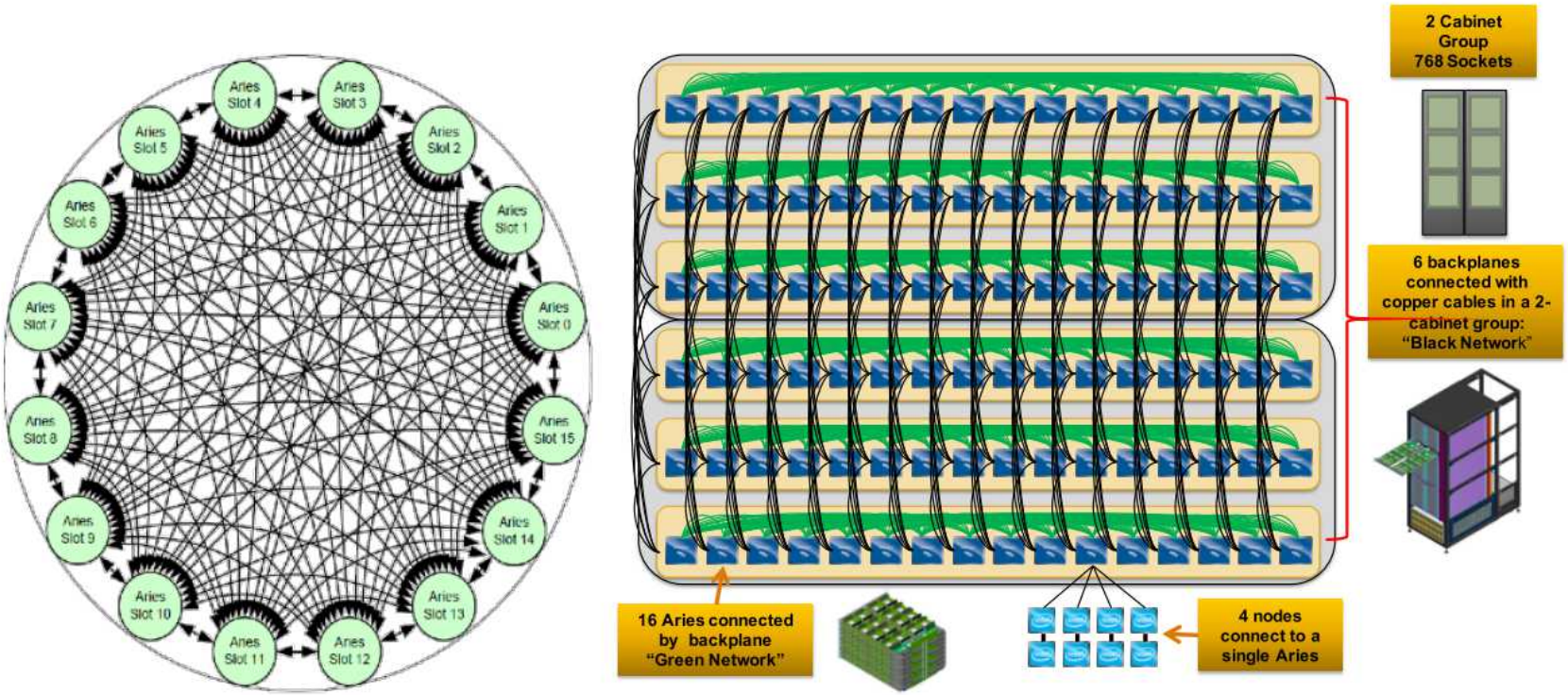


Figure: Group all-to-all **(x)** and intergroup connection **(y, z)** of dragonfly network. *Image Credit: Jeff Brooks, Cray Inc.*

E3SM Task Mapping

Sandia National Laboratories

Coupled compset method:

- Use AChax [Roth, 2018] tool (ORNL) to recover application communication graph (during "short" independent E3SM run)
- During full E3SM run:
  - Load application graph
  - Obtain and transform `rcalib` network coordinates
  - Partition both representations and perform mapping
  - Reorder global communicator with mapping solution

Advantages:

- Application and configuration agnostic
- Diagnostic application run can be done offline
- Non-invasive

References

Boman, E. G., Devine, K. D., Leung, V. J., Rajamanickam, S., Riesen, L. A., Deveci, M., and Catalyurek, U. (2012).

Zoltan2: Next-generation combinatorial toolkit.

Technical Report SAND2012-9373C, Sandia National Laboratories.

Deveci, M., Devine, K. D., Pedretti, K., Taylor, M. A., Rajamanickam, S., and Catalyurek, U. V. (2019).

Geometric Mapping of Tasks to Processors on Computers with Mesh or Torus Networks.

to appear IEEE Trans. Parallel and Distributed Systems.

Roth, P. C. (2018).

Scalable, automated characterization of parallel application communication behavior.

In *2018 Scalable Tools Workshop.*