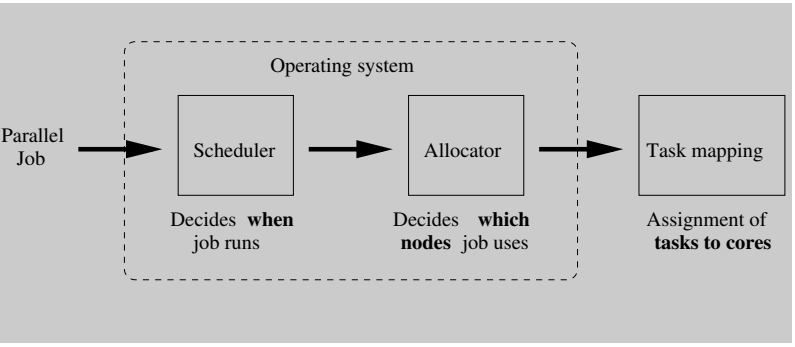
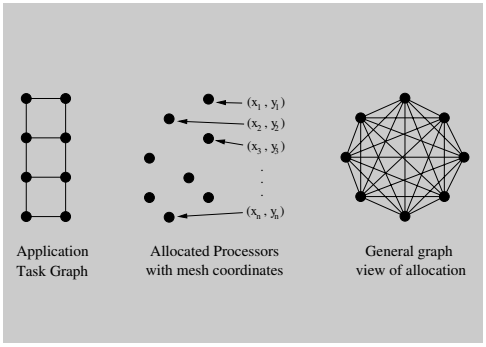
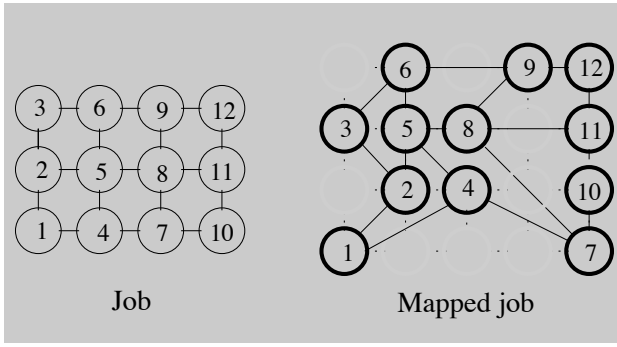
Exceptional service in the national interest

Sandia National Laboratories







Task Mapping Stencil Computations for Non-Contiguous Allocations

V. Leung (SNL), D. Bunde, J. Ebbers, S. Feer,

N. Price, Z. Rhodes, and M. Swank (Knox)

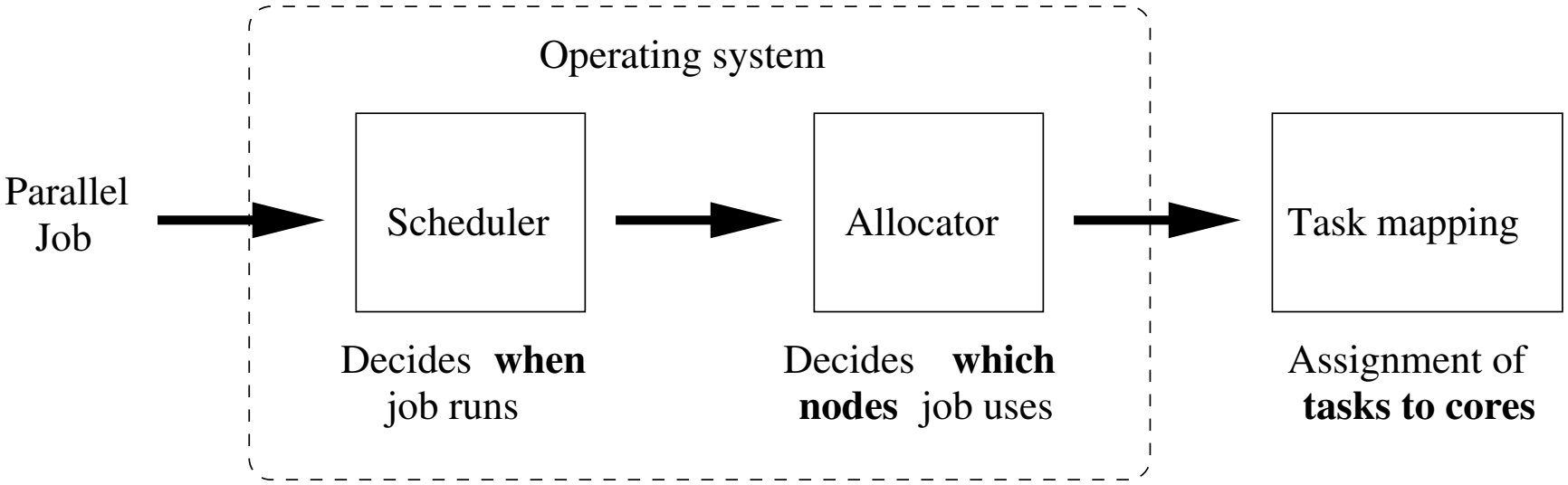
U.S. DEPARTMENT OF ENERGY

NNSA National Nuclear Security Administration

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

Parallel Resource Management Pipeline

Sandia National Laboratories

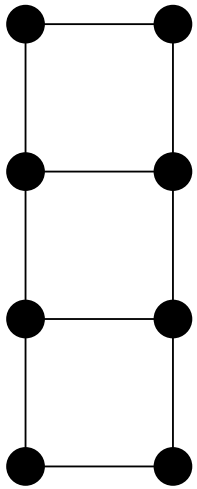


Task mapping

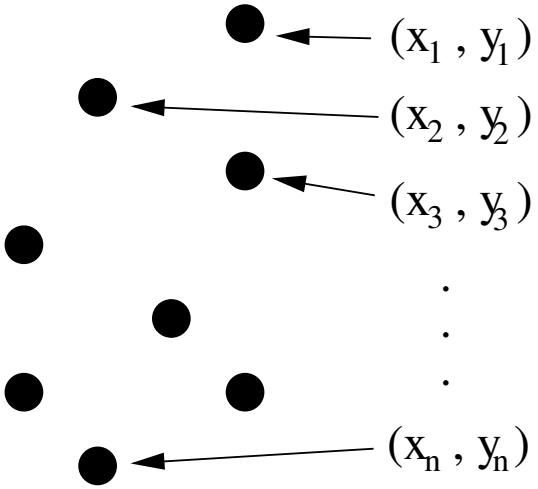
Sandia National Laboratories

- Long history [Bokhari, 1981]
- Less important in mid-1980s with wormhole routing
  - Message latency independent of size
- Recent resurgence
  - Almasi et al. 2004
  - Gygi et al. 2006 (application exhibited 1.64 times speedup)
  - Bhatelé et al. 2010 (contiguous problem)
  - Hoefler and Snir 2011 (general graph problem NP-Complete)
  - Barrett et al. 2012 (non-contiguous problem)
  - Deveci, et al. 2014 (continues this work)
- Contention for limited bandwidth
  - Processors continue improving faster than networks
  - Processor counts in state of the art HPC systems continue to grow

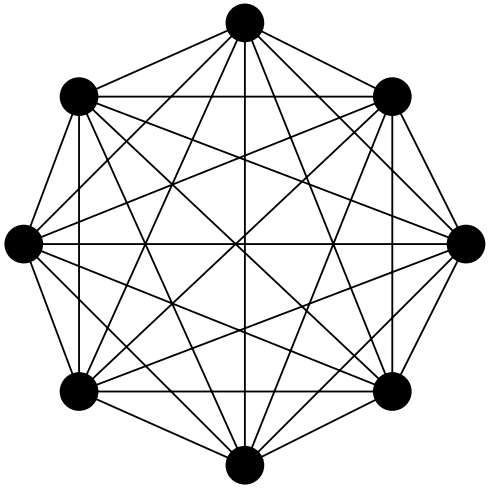
General view of task mapping



Application
Task Graph



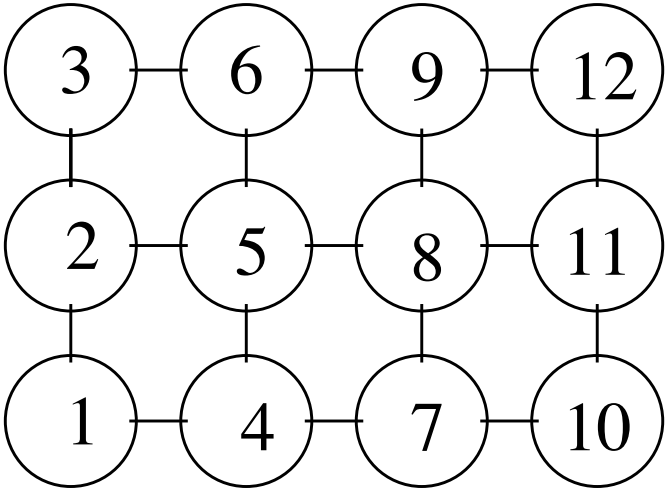
Allocated Processors
with mesh coordinates



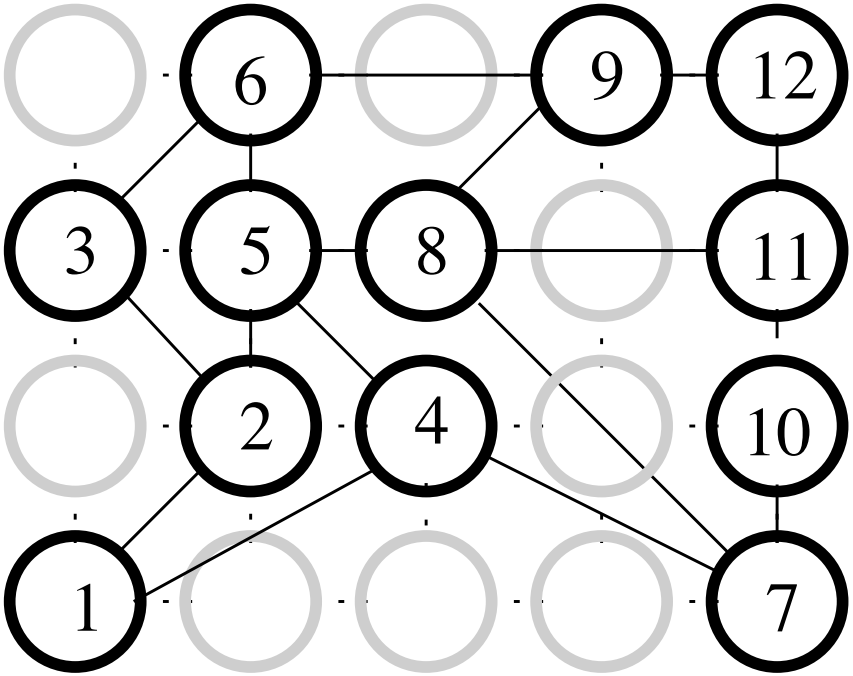
General graph
view of allocation

Possible mapping of a 4x3 job onto a 5x4 machine

Sandia National Laboratories



Job



Mapped job

This presentation

Sandia National Laboratories

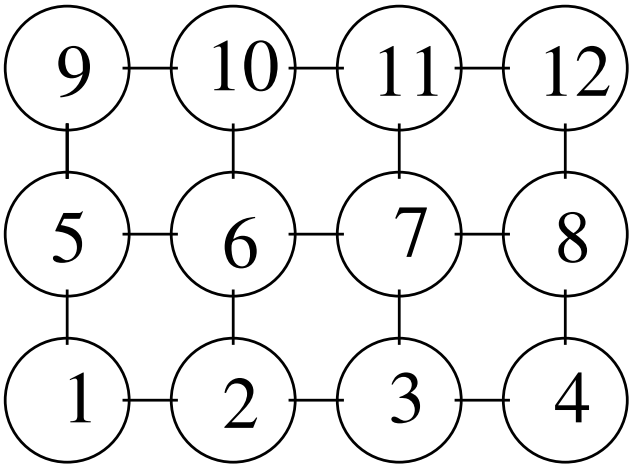
- Jobs with structured communication patterns and non-contiguous allocations
  - Jobs that communicate in a regular 3D nearest neighbor pattern
  - 3D mesh allocation with XYZ routing
- One of our algorithms greatly outperforms the others on Cielo
  - Recursively divides both task graph and set of allocated processors
- Our experiments show that average number of hops between communicating tasks is strongly correlated with running time
- Use of simulations to evaluate the algorithms on a variety of scenarios
- Discussion of future work

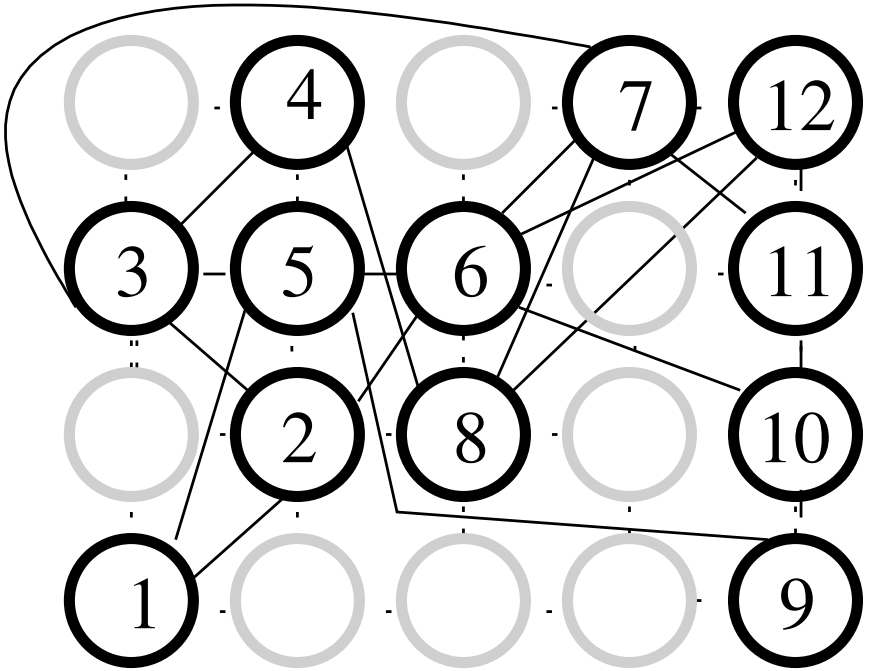
Algorithms

Sandia National Laboratories

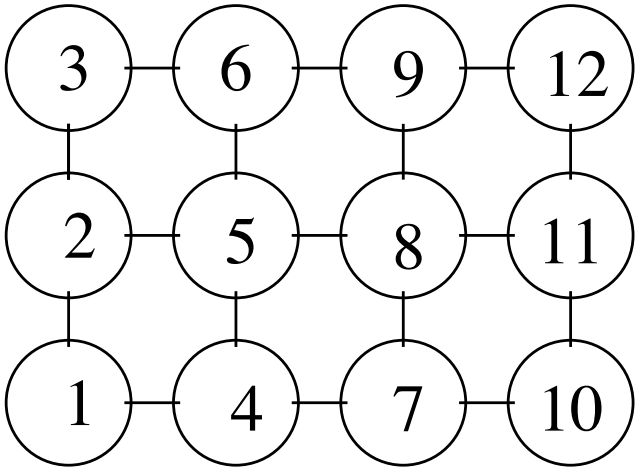
- Variations on baseline algorithm
  - Baseline
  - Barrett et al. 2012 - Grouping
  - Column major, Row major, Ordered (best of)
- Variations on contiguous algorithms
  - Corner
  - All corners
- Some greedy algorithms
  - Overlay
  - Two Way Overlay
- Recursive Coordinate Bisection (RCB)
- Rotations

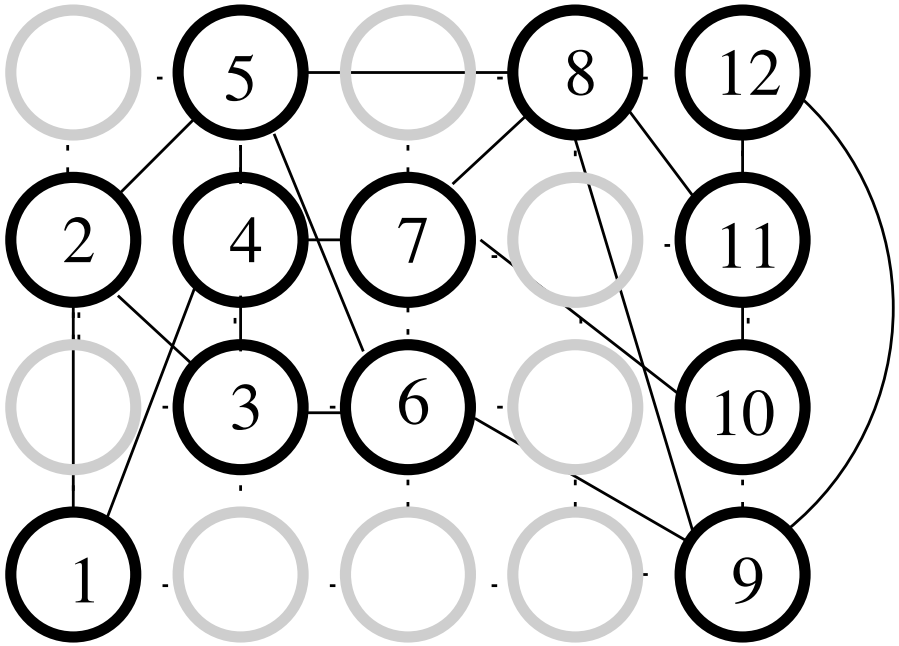
Baseline Mapping



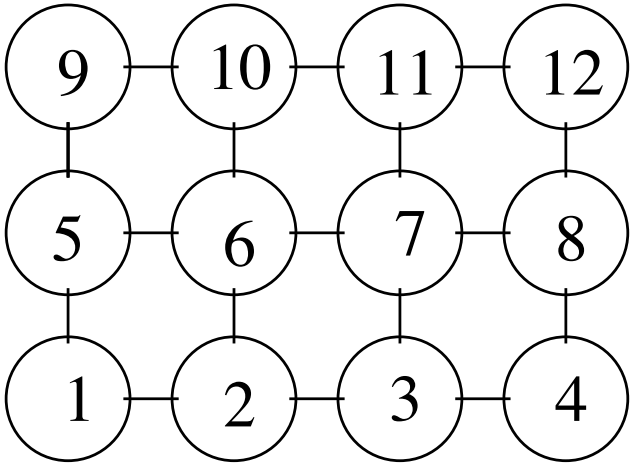


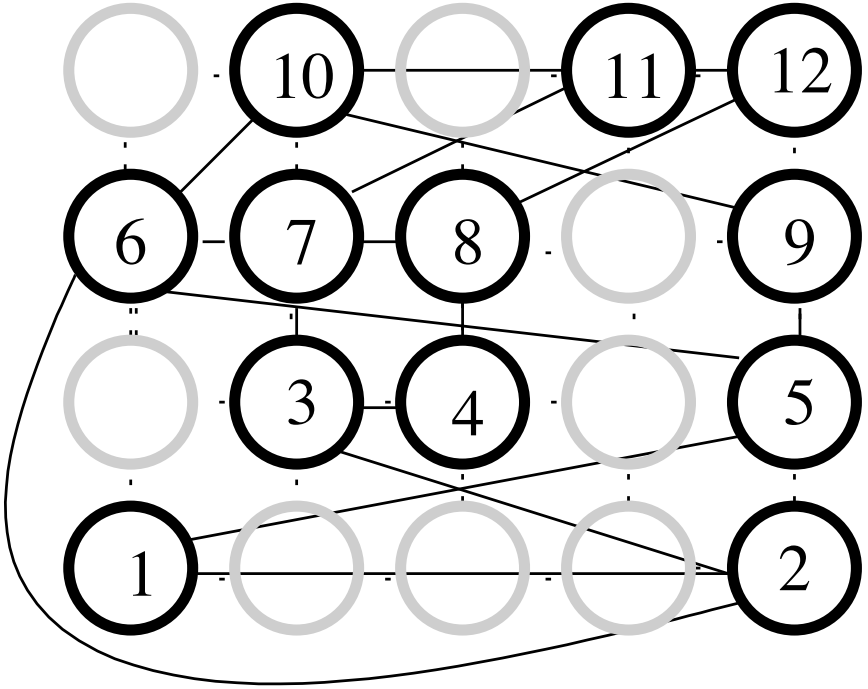
Mapping by Column major



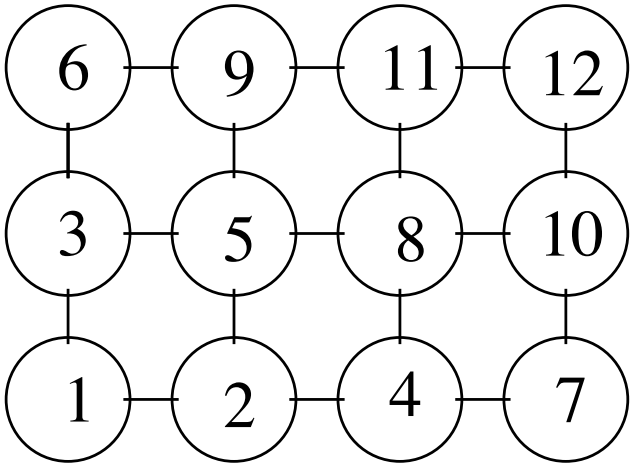


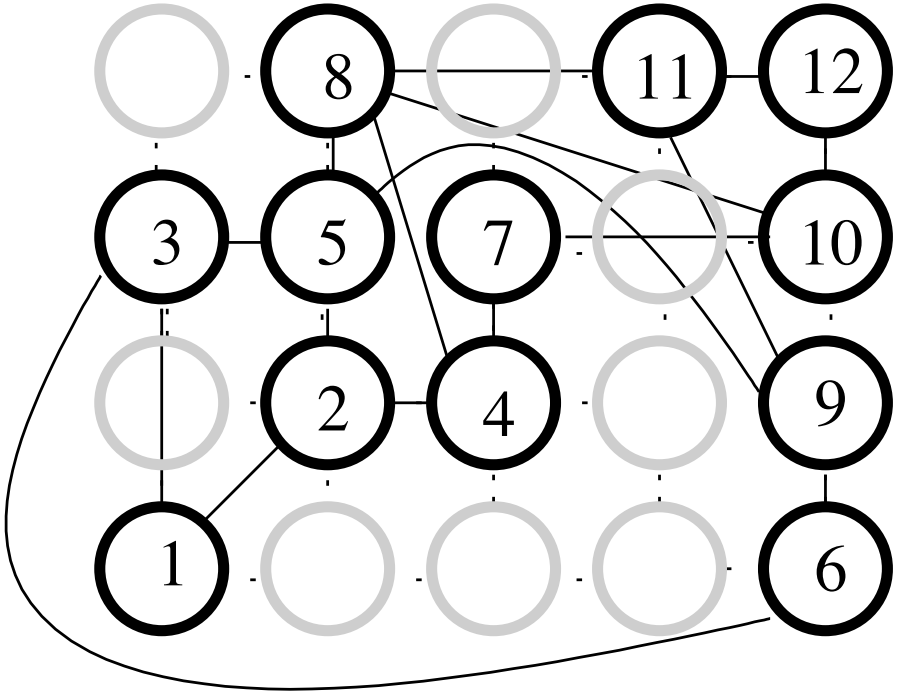
Mapping by Row major



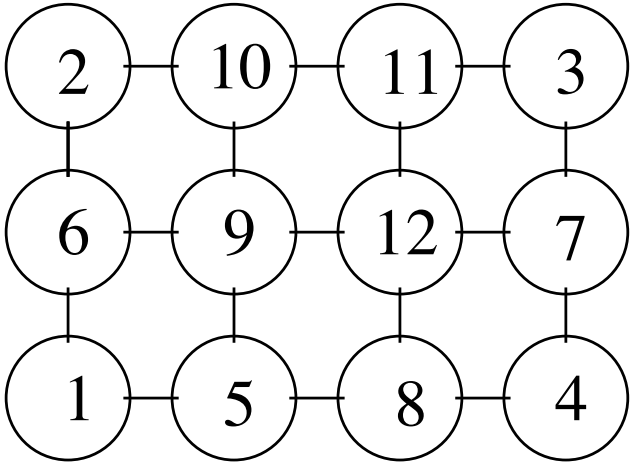


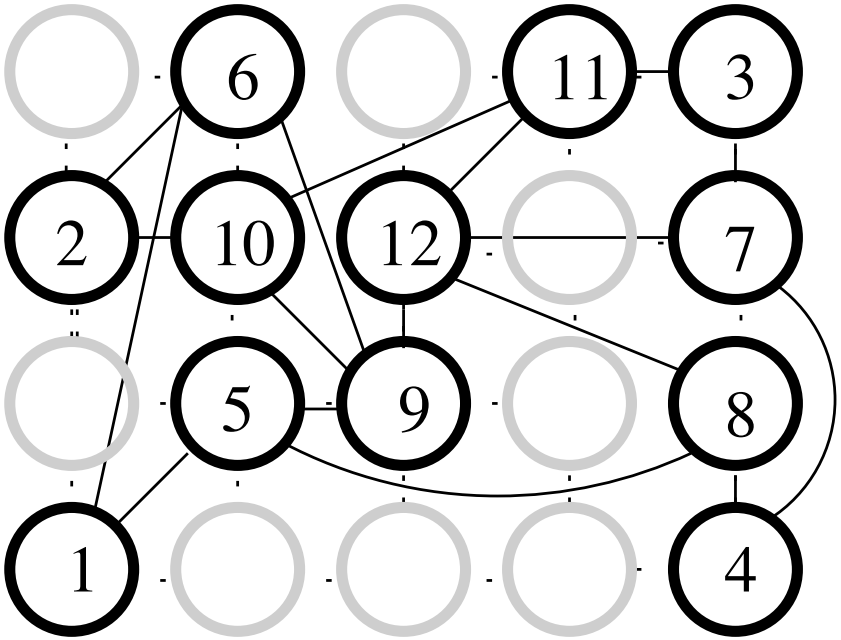
Mapping by Corner



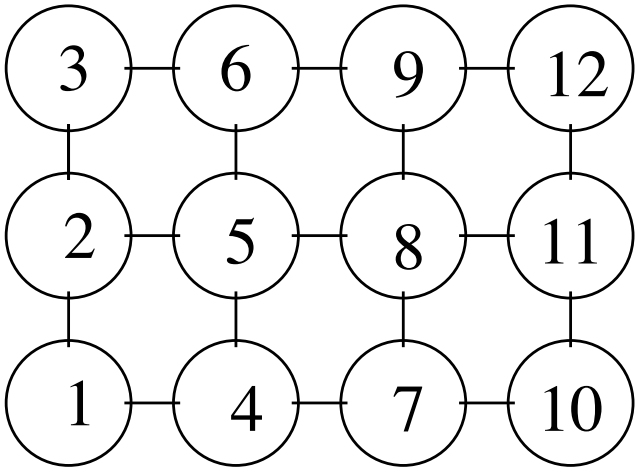


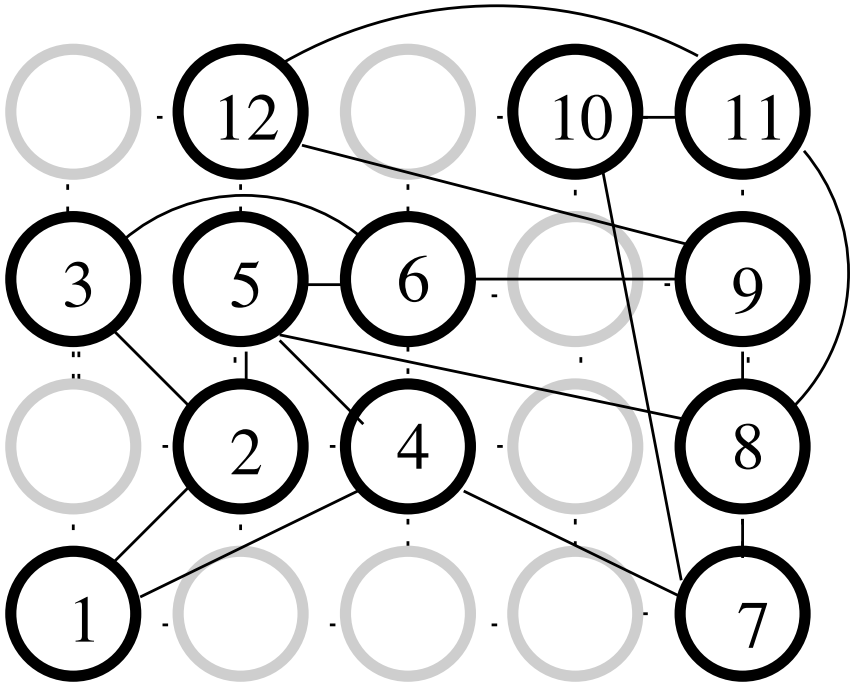
Mapping by All corners



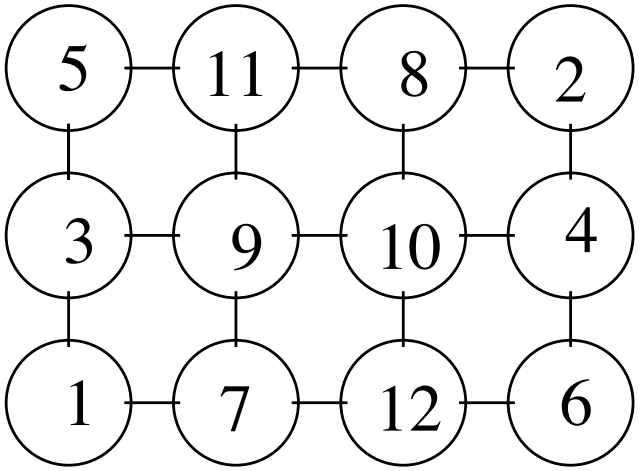


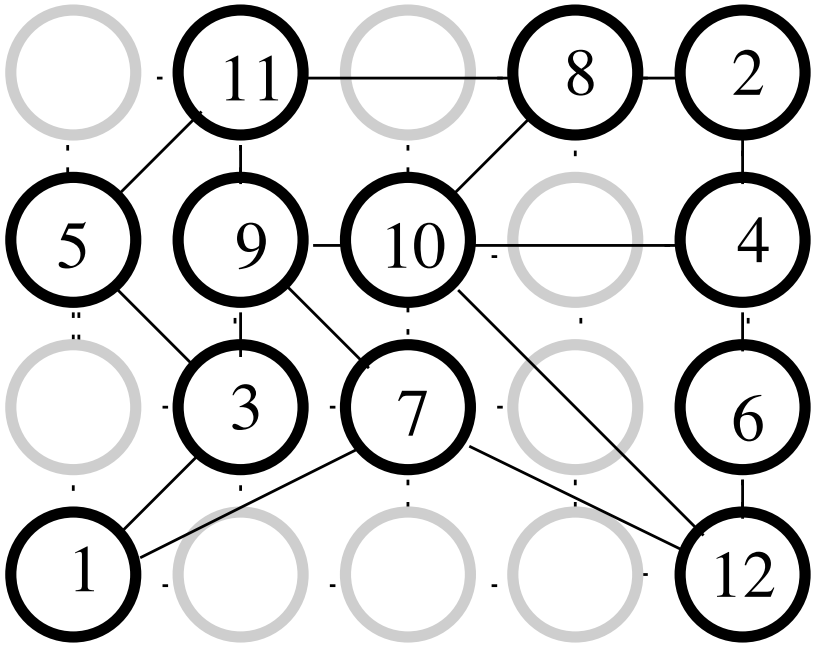
Mapping by Overlay



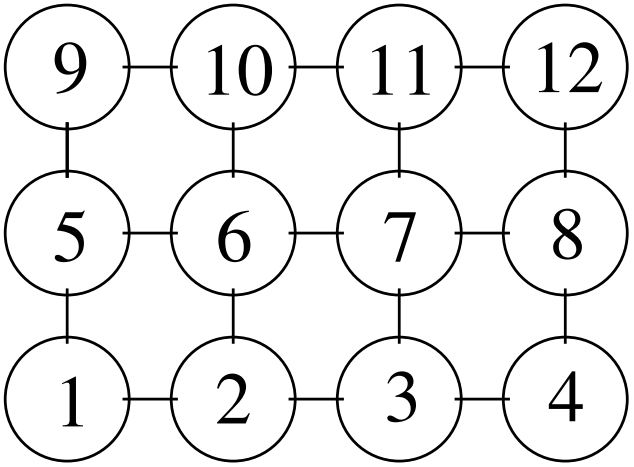


Mapping by Two Way Overlay

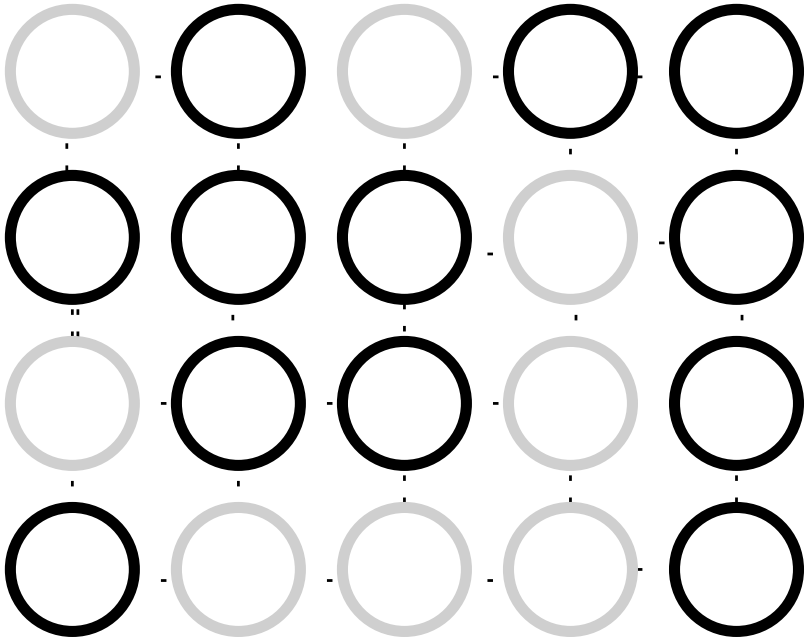




Using recursion for task mapping

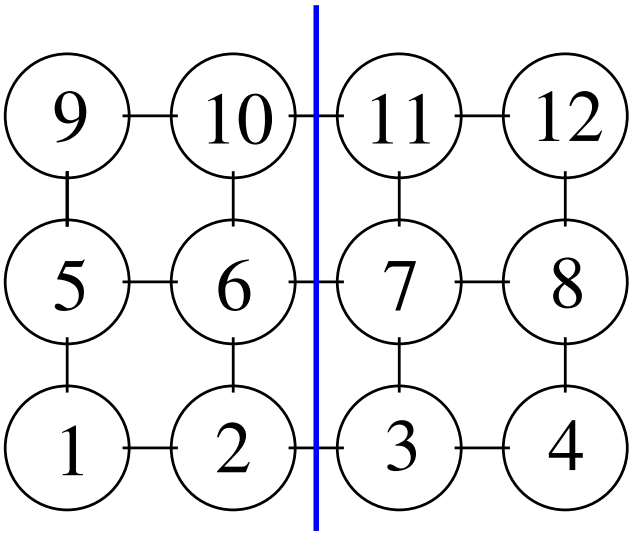


Job

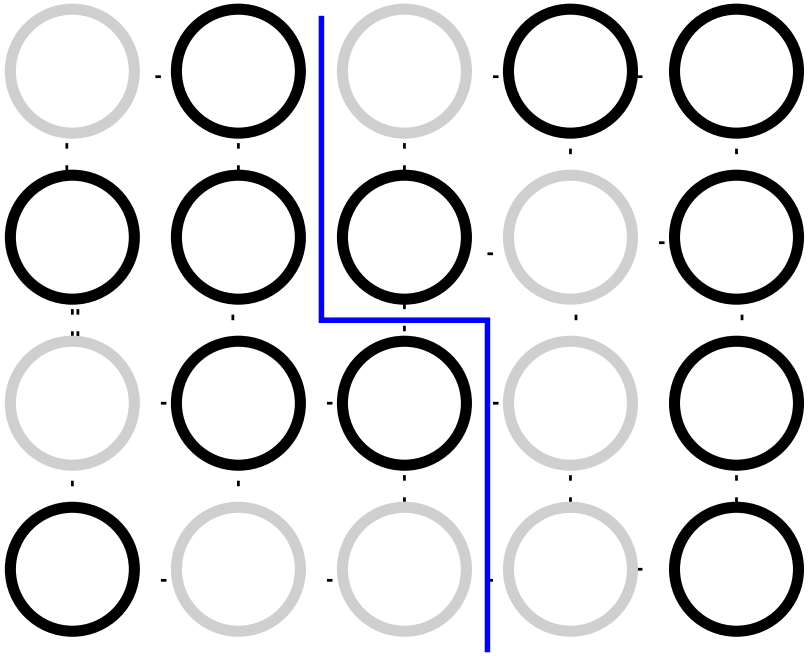


Machine

Using recursion for task mapping

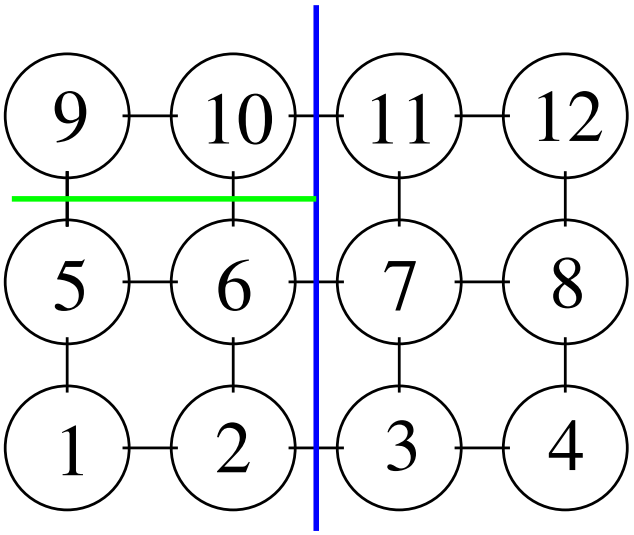


Job

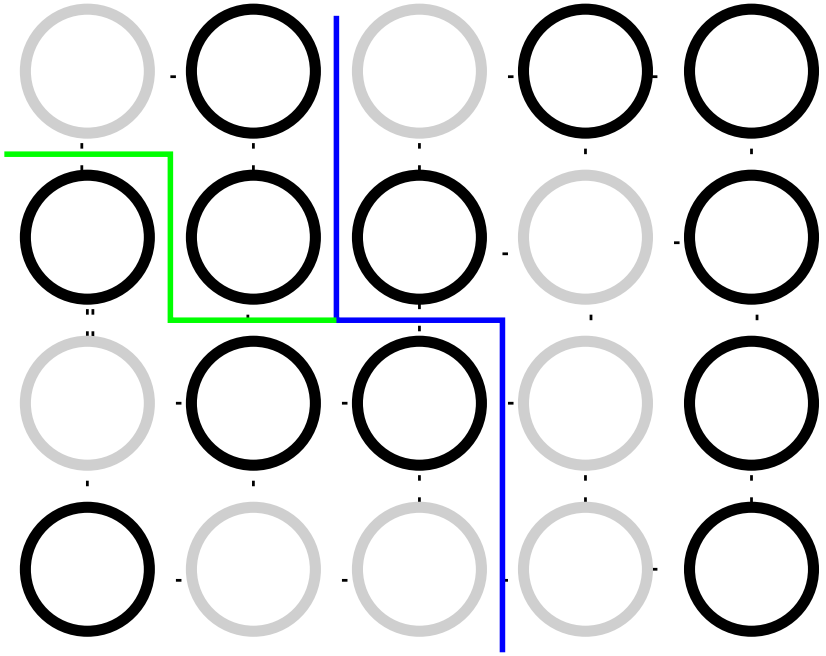


Machine

Using recursion for task mapping



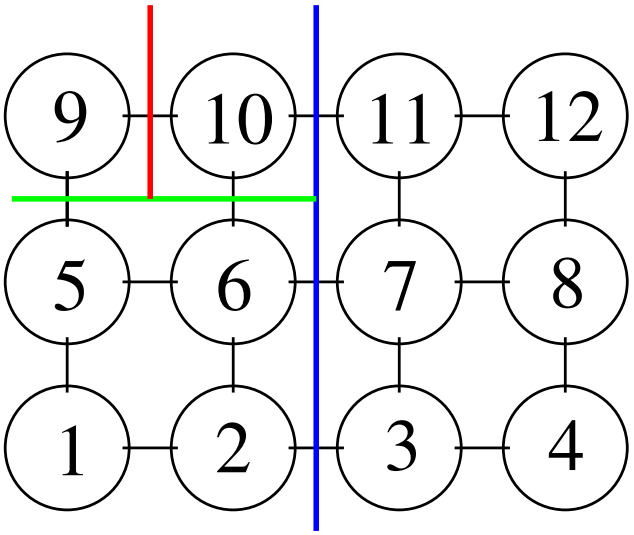
Job



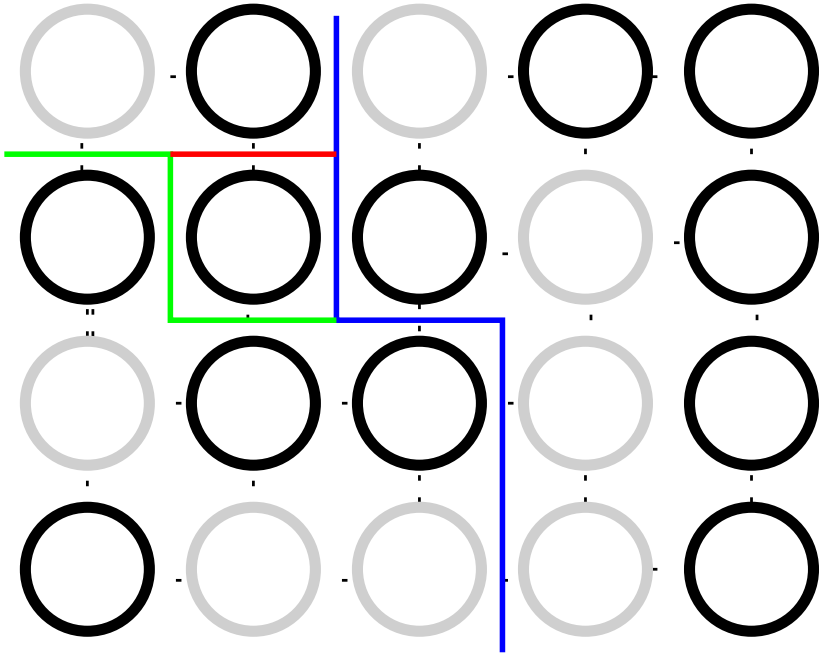
Machine

Using recursion for task mapping

Sandia National Laboratories

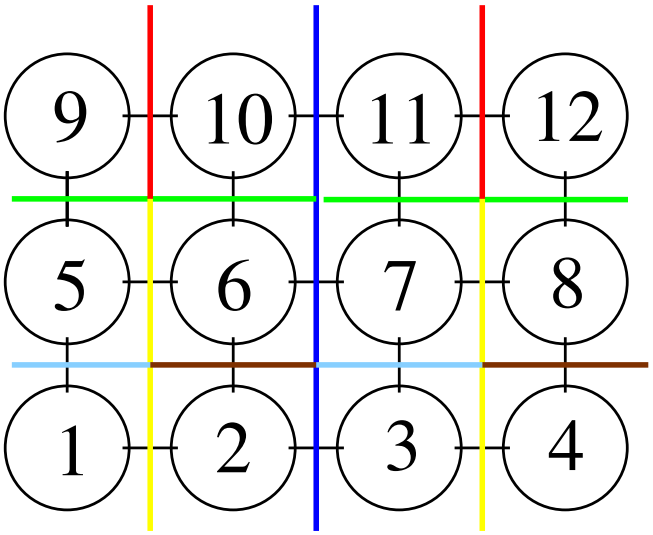


Job

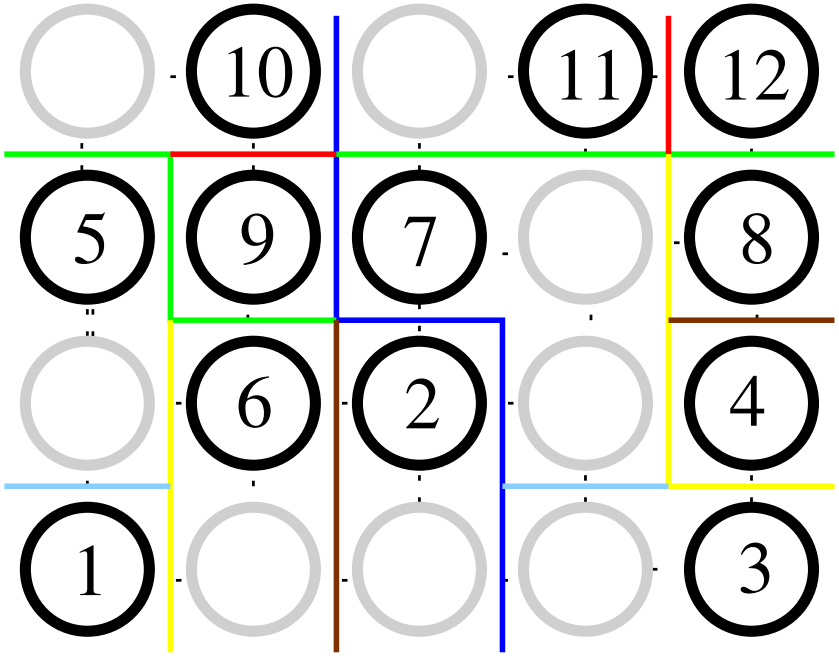


Machine

Using recursion for task mapping

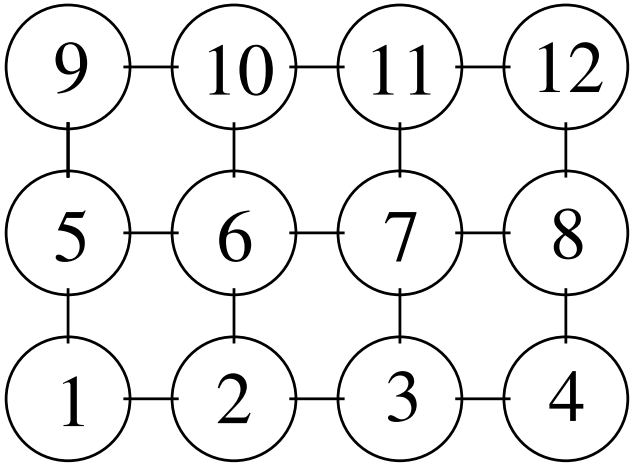


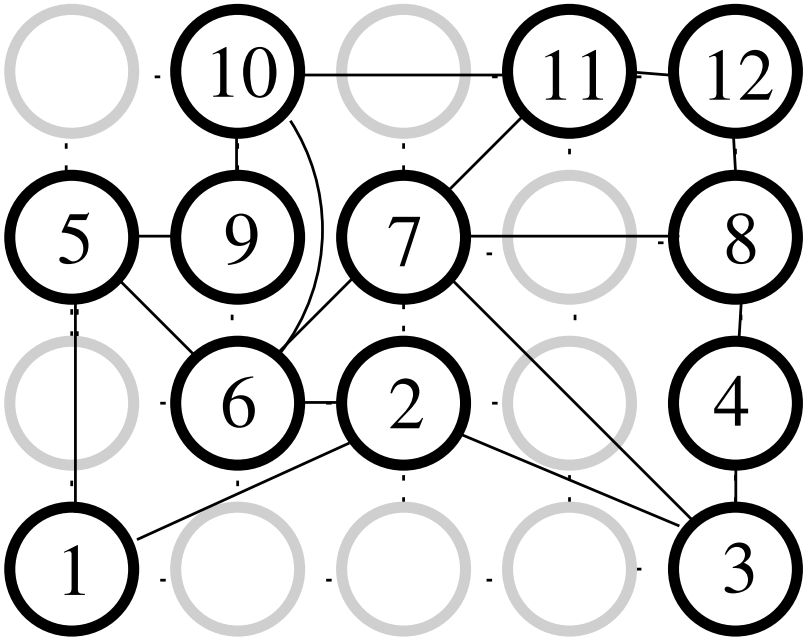
Job



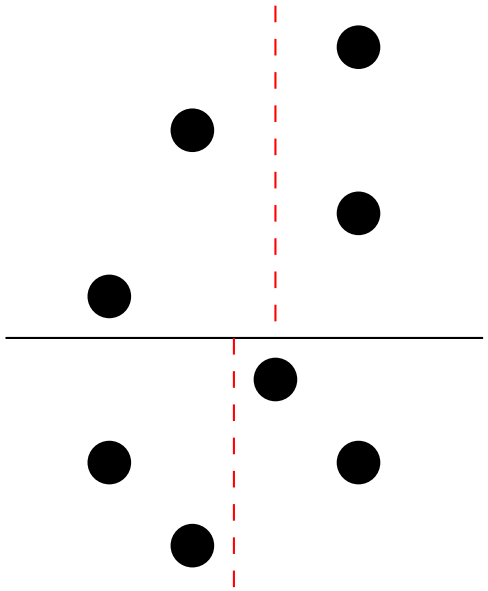
Machine

Mapping by RCB

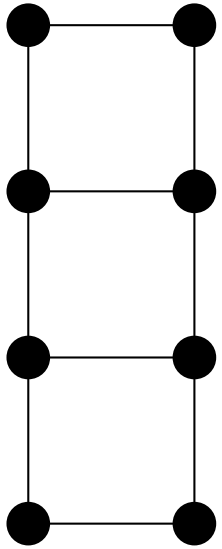




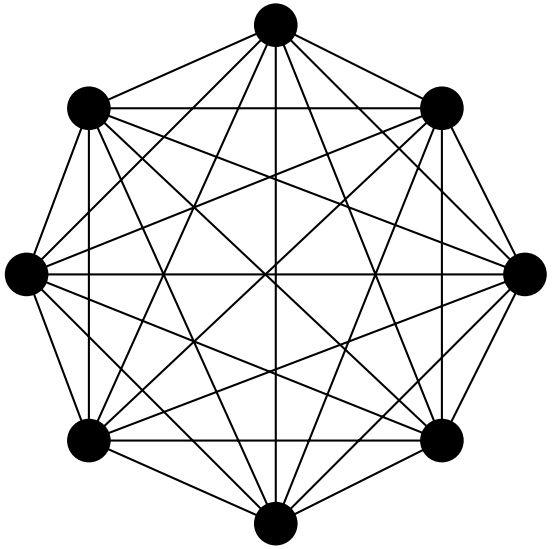
RCB versus General View



Allocated Processors
w/ 2 levels of RCB cuts



Application
Task Graph



General graph
view of allocation

Experiments

Sandia National Laboratories

- Los Alamos National Laboratory Cielo machine, Cray XE6
  - 143,104 compute cores in 8,944 compute nodes, dual AMD Opteron 6136 eight-core “Magny-Cours” socket G34 running at 2.4 GHz
  - 272 service nodes, AMD Opteron 2427 six-core “Istanbul” socket F running at 2.2 GHz
  - Gemini 3D torus in 16x12x24 (XYZ) topology, 2 compute nodes (sockets) per Gemini, 6.57x4.38x4.38 (XYZ) TB/s bi-section bandwidth
  - As of November 2013, number 26 on top 500 list
- Application used was miniGhost
  - Boundary exchange using stencil computations in scientific parallel computing, bulk-synchronous message passing code modeled on CTH
- Set of experiments consists of miniGhost runs for various numbers of total cores and cores per MPI rank

Job dimensions (total v per rank) XYZ

Sandia National Laboratories

Cores	16	8	4	2	1
16	1, 1,1	1, 2, 1	2, 2, 1	2, 4, 1	2, 4, 2
32	1, 2,1	1, 2, 2	2, 2, 2	2, 4, 2	2, 4, 4
64	1, 4,1	1, 4, 2	2, 4, 2	2, 8, 2	2, 8, 4
128	2, 4,1	2, 4, 2	4, 4, 2	4, 8, 2	4, 8, 4
256	2, 4,2	2, 4, 4	4, 4, 4	4, 8, 4	4, 8, 8
512	2, 8,2	2, 8, 4	4, 8, 4	4,16, 4	4,16, 8
1K	4, 8,2	4, 8, 4	8, 8, 4	8,16, 4	8,16, 8
2K	4, 8,4	4, 8, 8	8, 8, 8	8,16, 8	8,16,16
4K	4,16,4	4,16, 8	8,16, 8	8,32, 8	8,32,16
8K	8,16,4	8,16, 8	16,16, 8	16,32, 8	16,32,16
16K	8,16,8	8,16,16	16,16,16	16,32,16	16,32,32
32K	8,32,8	8,32,16	16,32,16	16,64,16	16,64,32
64K	16,32,8	16,32,16	32,32,16	32,64,16	32,64,32

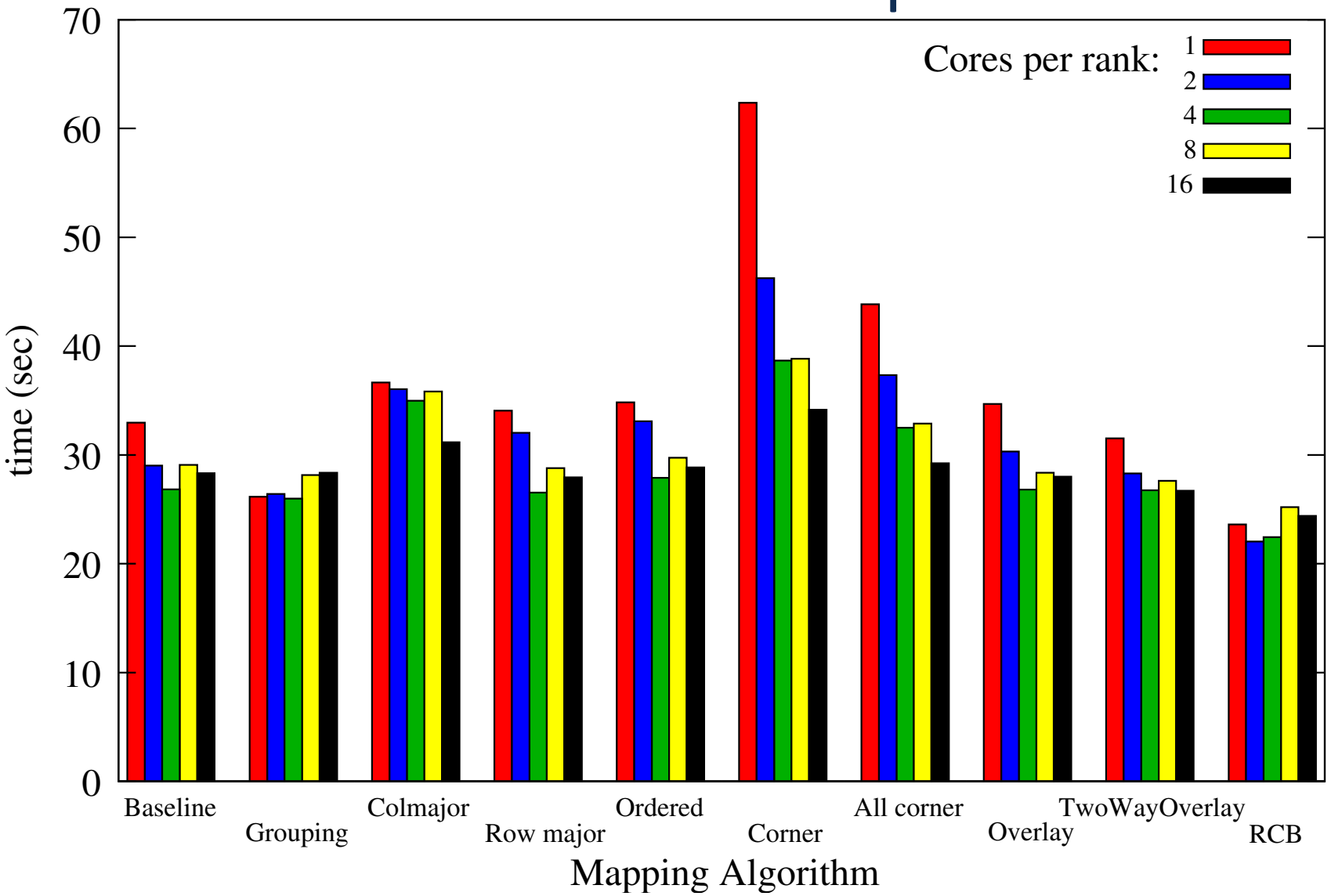
Experiments

Sandia National Laboratories

- All jobs in set of experiments submitted at same time
  - Due to system load
    - 1st set took almost 2 weeks to get on and off Cielo
    - 4 more sets took less than a day each
- For given number of cores, single script (allocation) used
  - 10 task mapping algorithms run for each core per rank
  - Entire set of 10 run one job after another to minimize experimental variances other than cores per rank and task mapping algorithms
- Task mapping algorithm selected early in application
  - All task mapping algorithms implemented in similar manner
  - Differences in running time for task mapping algorithms insignificant
  - Output includes total time, communication time as percentage of total (~30%), and average hops between neighboring ranks in application

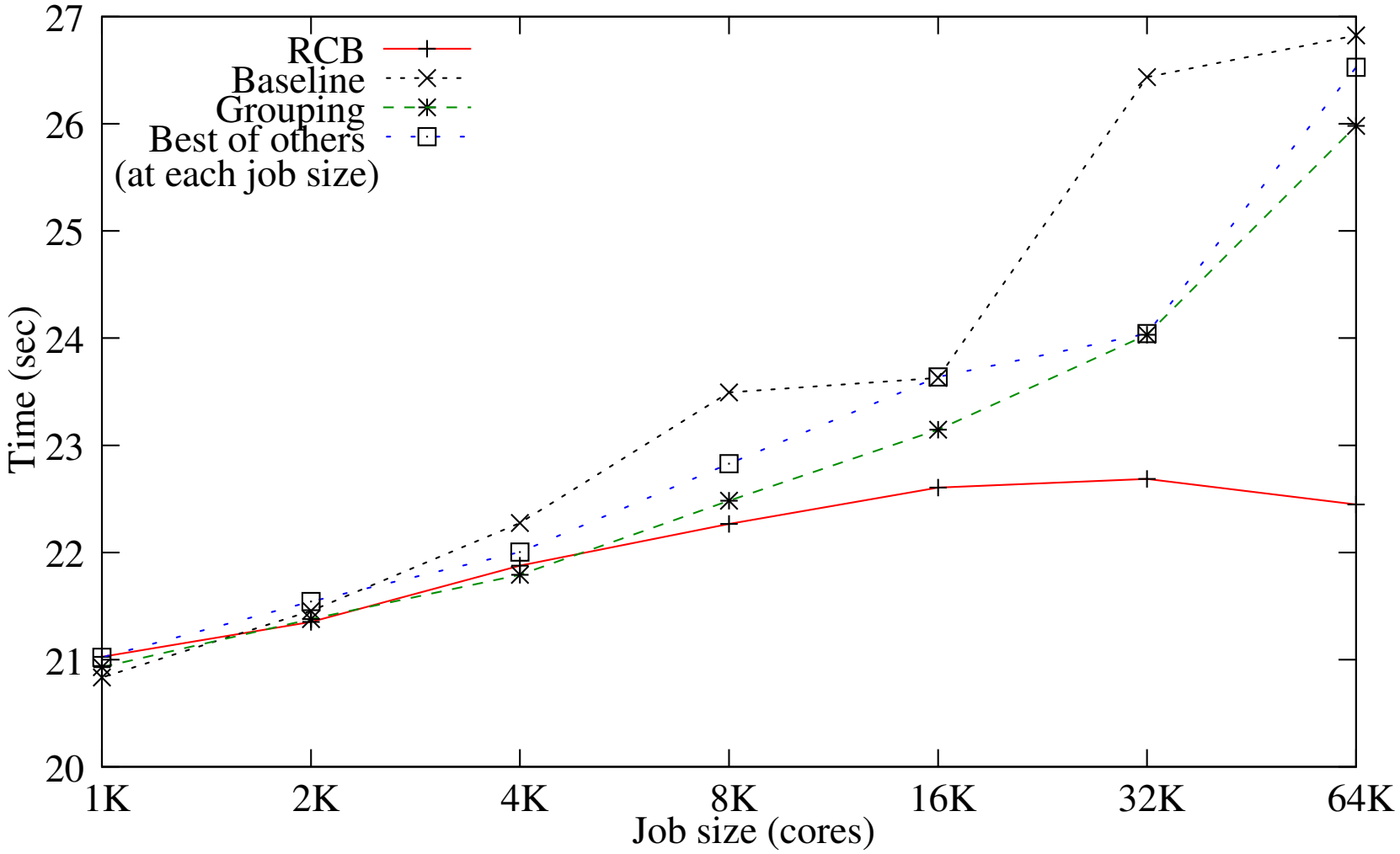
Average running time for 64K-core job as a function of number of cores per rank

Sandia National Laboratories



Weak Scaling for 4 Cores per MPI Rank

Sandia National Laboratories

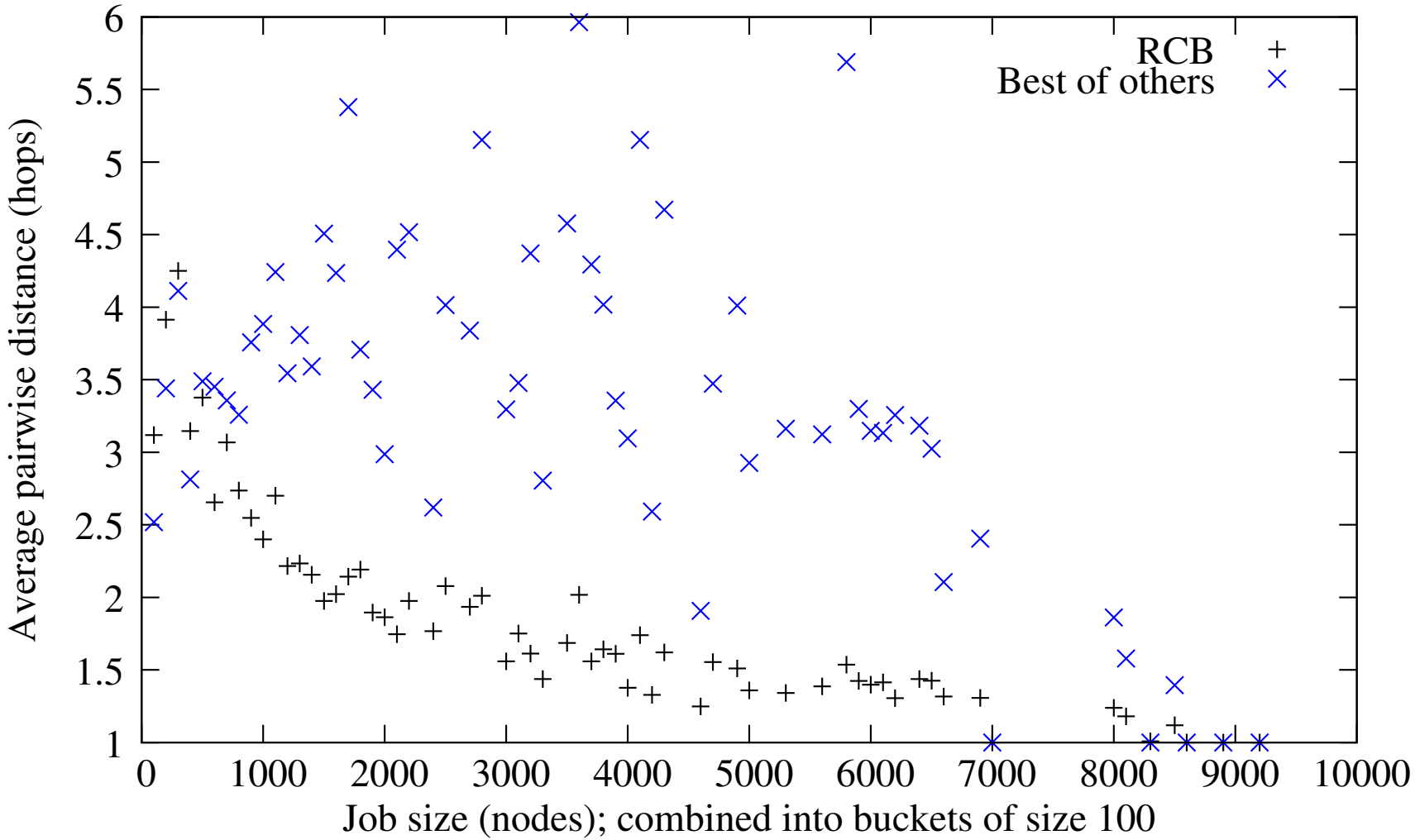


Running time correlations with hop metrics for trace-based simulations

Sandia National Laboratories

- Since time on large systems is scarce, we want to identify metrics to judge mapping quality in simulations
  - Average number of hops ≈ hops-bytes (Bhatelé et al. 2010)
  - Maximum hops
  - Variance on average number of hops
- Trace-based simulations of more varied scenarios (PWA)
  - Job arrival time, size, running time, and (in many cases) time estimate
  - On machine
    - schedule (EASY),
    - allocate (snake best fit [Lo et al. 1997 and Leung et al. 2002]), and
    - map (no Baseline and Grouping mappers)
  - Summary of trace used in simulations
    - Log name: LLNL-Atlas-2006-2.1-cln, Machine: 96x96, # jobs used: 12,474

Simulation results



Next

Sandia National Laboratories

- Transfer RCB mapper into a library so that other programs can easily adopt it
  - Zoltan2
- Investigating other communications patterns
  - With extensions of RCB being a natural place to start
- Investigate additional metrics
  - Theoretical congestion
  - Measured congestion