Task Mapping for Non-contiguous Allocations

David P. Bunde[1], Johnathan Ebbers[1], Stefan P. Feer[2], Vitus J. Leung[3],
Nickolas W. Price[1], Zachary D. Rhodes[4], Matthew Swank[1]

[1]Knox College
{dbunde,jebbers,
nprice, mswank}@knox.edu

[2]3M Health Information Systems
sfeer@mmm.com

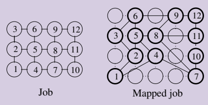
[3]Sandia National Laboratories
vjleung@sandia.gov

[4]Allstate Corporation
rhodesz87@gmail.com

Task Mapping

- Assign job tasks to allocated nodes
- Close communicating tasks use fewer links
  - Preserves bandwidth, lessening interference between jobs (Applications have seen a 1.64-times speedup from improved mapping [4])
  - Potentially saves power

We map <u>common job communication pattern</u> (stencil) onto <u>practical allocations for Cray X</u> (non-contiguous mesh nodes)



Previous work either
- Assumes contiguous allocation (e.g. [6])
  - Not useful w/ non-contiguous allocation (e.g. on Cray XE)
- States problem as graph embedding (e.g. [5])
  - Very general, but loses geometric information (coords)

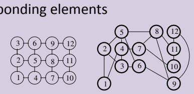
Algorithms

Rotations
- Rotate job so relative order of dimension lengths matches dimensions of bounding box of nodes
- Applied to all algorithms except Baseline and Grouping

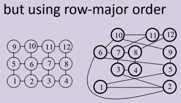
Baseline (provided by ALPS and Moab)
- Number cores in alloc. order and tasks in row-major order
- Map corresponding elements

Grouping [1]
- Baseline after grouping tasks into 2x2x4 blocks

ColMajor
- Number tasks and cores in column-major order
- Map corresponding elements

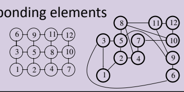


RowMajor
- As ColMajor, but using row-major order

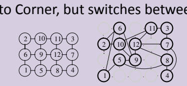


Ordered
- Try different linear orders (inc. ColMajor and RowMajor)
- Take whichever gives lowest average hops

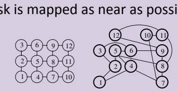
Corner (generalizes "Expand from corner" [3] w/ rotations)
- Number tasks and cores based on distance from (0,0)
- Map corresponding elements



AllCorners (generalizes "Corners to center" [3] w/ rotations)
- Similar to Corner, but switches between corners

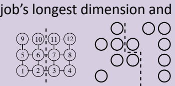


Overlay
- Identifies desired location for each task
- Lower left corner task desires lower left corner core
- Other tasks relative to this
- Each task is mapped as near as possible to desired location

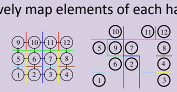


TwoWayOverlay
- Similar to Overlay, but switches between opposite corners

RCB
- Bisects job's longest dimension and bisect cores w/ same one



- Recursively map elements of each half



- [5] also has bisection heuristic, but with general graphs
  - Loses geometric relationship between recursion levels so doesn't perform as well in our setting

Experiments

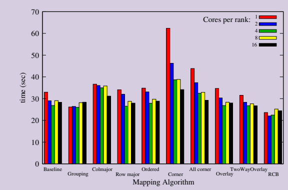
Machine: Cielo – Cray XE6
- 8,944 compute nodes (2 sockets/node, 8 cores/socket)
- Gemini 3D torus in 16x12x24 (XYZ) topology
  - 6.47x4.38x4.38 (XYZ) TB/s bisection bandwidth
- #22 on June 2013 Top500 list

Job configurations
- 16-64K cores, with rectangular prism job shape
- Between 1 and 16 cores per MPI rank (MPI + OpenMP)

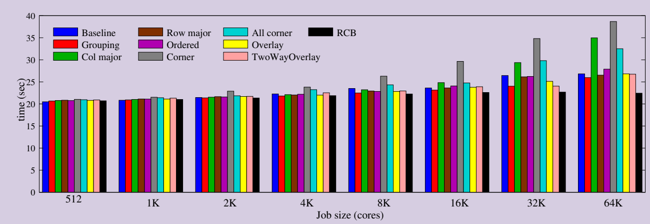
Application: miniGhost "mini application" [2]
- Boundary exchange using stencil computation
- Modeled after CTH, a multi-material, large deformation, strong shock wave, solid mechanics code
- Gives similar scaling as CTH, including relative performance improvement for remapping

Results averaged over 5 runs

Number of cores per MPI rank
- Tuning option in miniGhost (changes number of ranks)
  - More ranks means more messages, but smaller
  - More ranks also slows collective operations
- Worst mappers favor 16 cores/rank
  - Fewer messages helps network-bound runs
- Better mappers perform best at intermediate values
  - We focus on 4 cores/rank; also favored in previous work [1]



Time for 64K-core job as a function of cores/rank



Time as a function of job size (4 cores/rank)

Overall results
- RCB is consistently the best mapper
  - 4 cores/rank: 16% ave. improvement at 64K (12.8-18.6% per run)
  - Better w/ other cores/rank: 24.1% at 2 and 28.4% at 1
  - Also most consistent: Baseline running time had multi-second std. deviations while RCB always ≤ 0.6 seconds
- Best overall performance: RCB with 2 cores/rank (Best in 17 of 20 trials at 8K+ cores (5 runs of 4 sizes))

Rotations and correlations to enable future simulations
- Rotation step improves ~60% of runs (ave. benefit 1-6%)
- Average hops between communicating tasks correlates with running time (Spearman rank test; significance level 0.05 or better for 1K+ core jobs)
  - Small jobs affected by hop between sockets at node
  - Allowed trace-based simulations that confirmed relative quality of mapping algorithms

References

[1] R. Barrett et al. Navigating an evolutionary fast path to exascale. In Proc. 3rd Intern. Workshop Performance Modeling, Benchmarking and Simulation of High Performance Computing Systems (PMBS), 2012.
[2] R.F. Barrett, C.T. Vaughan, and M.A. Heroux. MiniGhost: A miniApp for exploring boundary exchange strategies using stencil computations in scientific parallel computing. Technical Report SAND2011-5294832, Sandia National Laboratories, 2011.
[3] A. Bhatelé, G.R. Gupta, L.V. Kalé, and I.-H. Chung. Automated mapping of regular communication graphs on mesh interconnects. In Proc. Intern. Conf. High Performance Computing (HiPC), 2010.
[4] F. Gygi et al. Large-scale electronic structure calculations of high-Z metals on the BlueGene/L platform. In Proc. 2006 ACM/IEEE Conf. on Supercomputing (SC), 2006.
[5] T. Hoefler and M. Snir. Generic topology mapping strategies for large-scale parallel architectures. In Proc. 25rd ACM Intern. Conf. Supercomputing (ICS), 2011.
[6] H. Yu, I-H. Chung, and J. Moreira. Topology mapping for Blue Gene/L supercomputer. In Proc. 2006 ACM/IEEE Conf. on Supercomputing (SC), 2006.

We thank Courtney Vaughan and Kevin Hastings for helpful discussions. D.P. Bunde, J. Ebbers, S.P. Feer, N.W. Price, Z.D. Rhodes, and M. Swank were partially supported by contract 899808 from Sandia National Laboratories. Z.D. Rhodes also acknowledges support from a Post-baccalaureate Fellowship from Knox College. The work was done while S.P. Feer and Z.D. Rhodes were students at Knox College.
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.

KNOX COLLEGE   Sandia National Laboratories   U.S. DEPARTMENT OF ENERGY