Optimizing task layout on the Blue Gene/L supercomputer

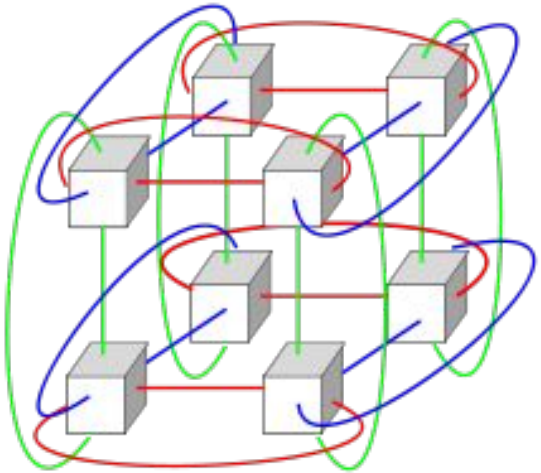
Paper Authors: G. Bhanot A. Gara P. Heidelberger E. Lawless J. C. Sexton R. Walkup
Presentation: Daniel Nichols

Paper Background

- Published in *IBM Journal of Research and Development* in 2005
- Authors from IBM Thomas J. Watson Research Center
- 21 Citations and 4 Patent Citations

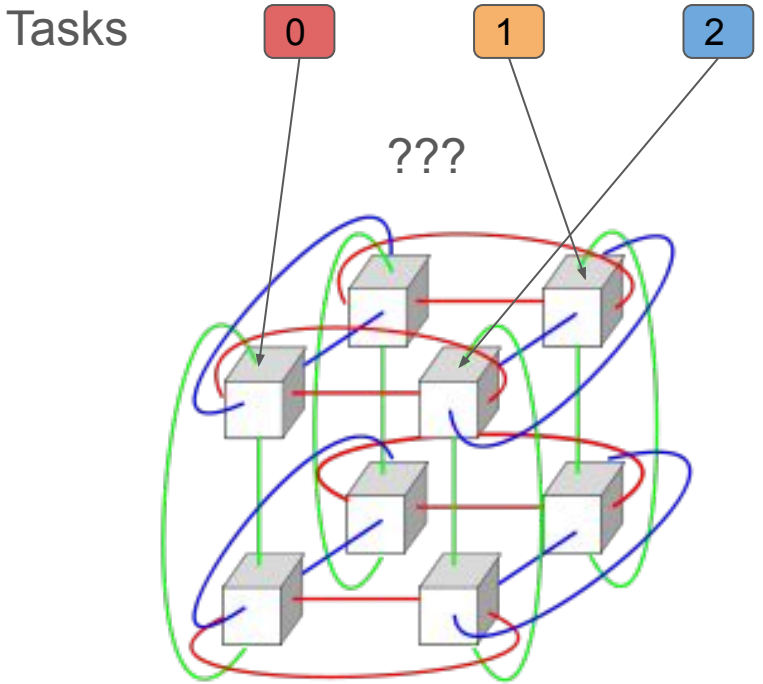
Blue Gene/L Overview

- Nearest neighbor network between nodes arrange in 3D torus
- Global collective network
- In addition to Barrier, I/O, and Control networks
- Typical node
  - 512 Mb memory
  - 2 CPUs
  - 2.8 Gflops peak
  - 6 1.4 Gb/s network connections



Problem Setup

- How tasks are mapped to nodes can greatly impact total communication time
- Problem: Find an optimal mapping
- Solution Overview
  - Model communication cost of mapping analytically
  - Minimize analytical model



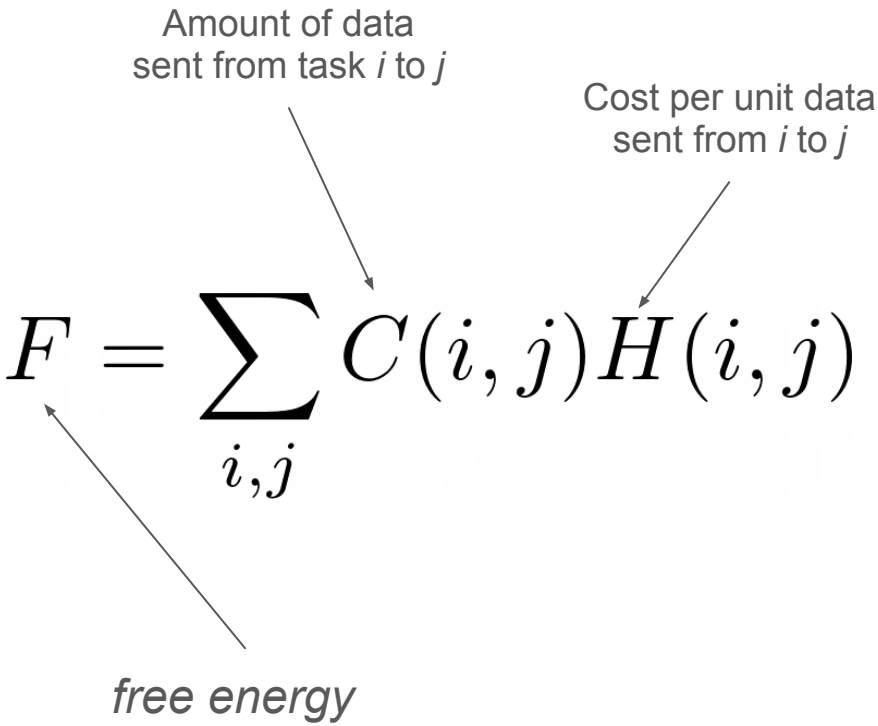
Modelling Cost

- Cost of sending message
  - Even for fixed problem *L* and *B* will vary at runtime
- Good cost model should
  - Increase as path lengths (and therefore *L*) increase. Also contention becomes more likely
  - Decrease if multiple paths are available (B increases)

$$t = \max\left(t_r, t_s + L + \frac{S}{B}\right)$$

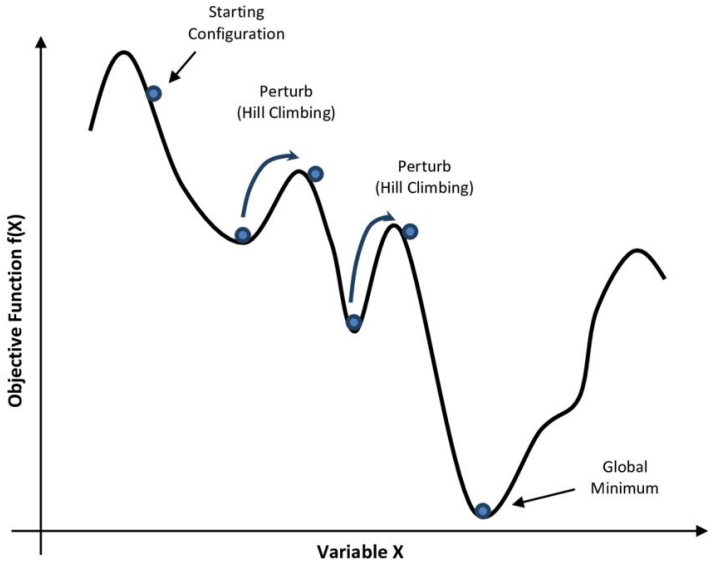
Modelling Cost

- C(i,j) is fixed by what application is running
- H(i,j) depends on mapping
  - Optimization variable
- Authors make assumption H is independent of message size and only dependent on distance
  - Set H(i,j) to number of hops



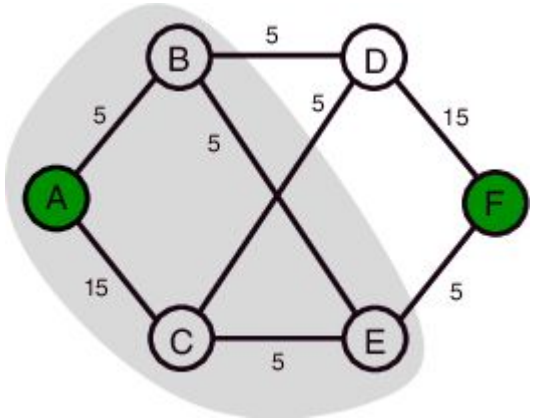
Optimization

- Minimized with Simulated Annealing (SA)
  - Iteratively try to improve mapping by making random changes
  - Construct a Markov chain $M_i,...,M_n$
  - Use Metropolis algorithm to find M_i configurations
- Choosing initial configuration
  - Uses heuristic which tries to reduce number of tasks per node



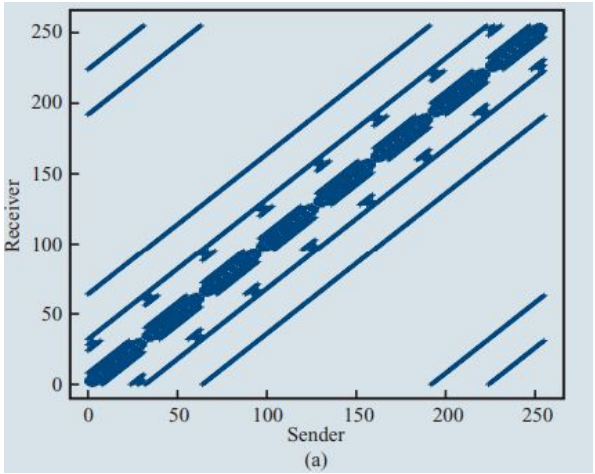
Improvements

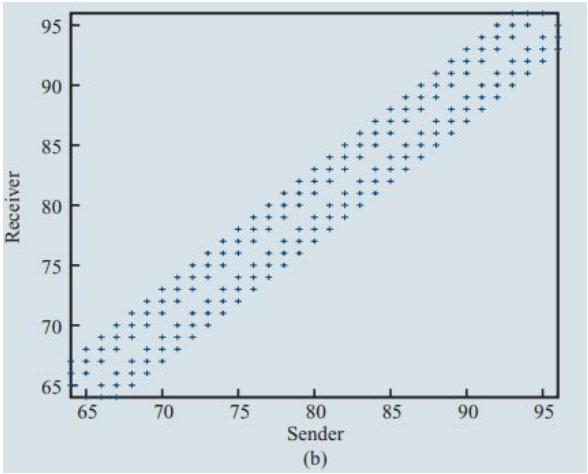
- Normal SA is too slow for large node counts
- Observation: C(i,j) is often sparse
  - Can be represented as a graph
- Split C(i,j) into subgraphs using minimum cuts (in METIS)
- Find optimal mappings for each subgraph

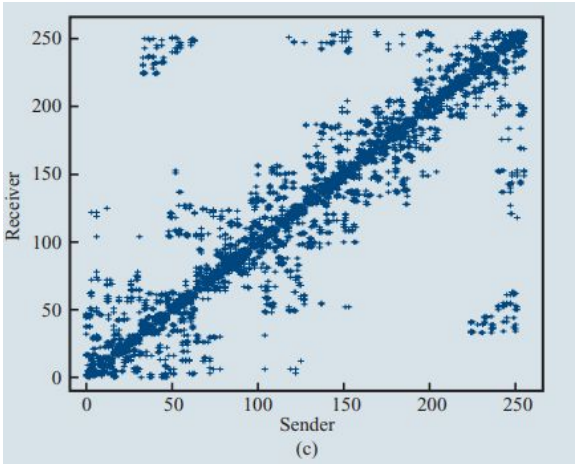


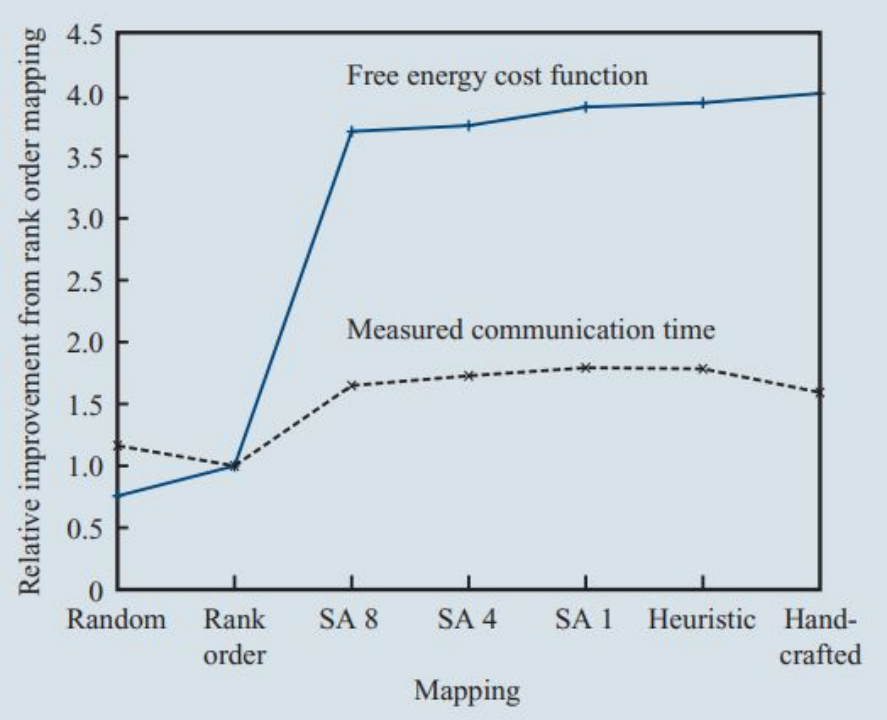
Experiments

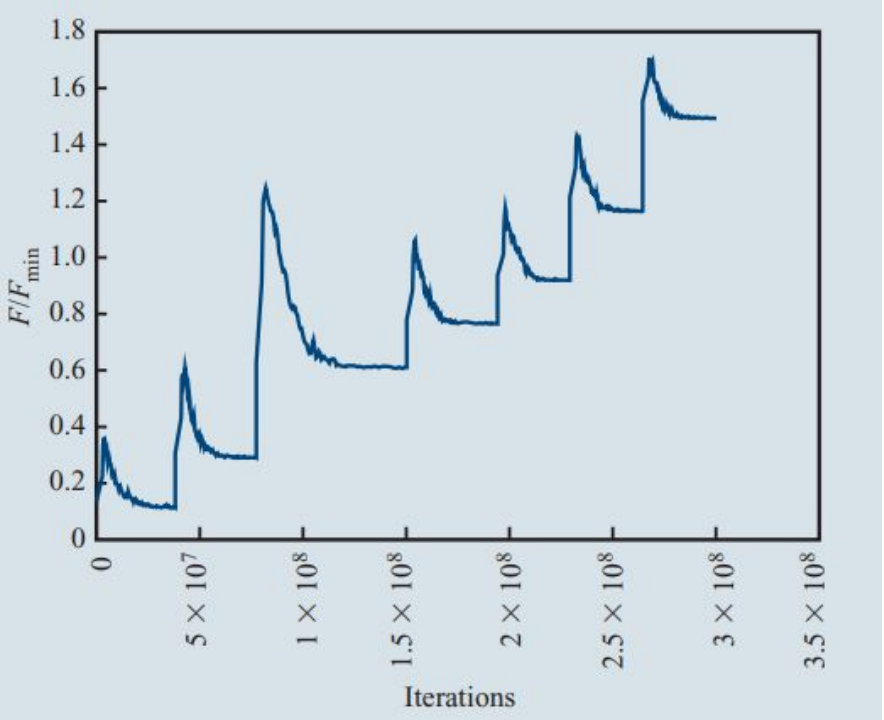
- 2 contrived
  - Cubic nearest-neighbor regular communication pattern
- 2 apps
  - SAGE -- SAIC adaptive grid Eulerian hydrodynamics application
    - Regular communication
  - UMT2000
    - Photon transport simulation
    - Irregular communication


Receiver
Sender
(a)


Receiver
Sender
(b)


Receiver
Sender
(c)


Relative improvement from rank order mapping
4.5
4.0
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0
Free energy cost function
Measured communication time
Random
Rank order
SA 8
SA 4
SA 1
Heuristic
Hand-crafted
Mapping


F/F_{min}
1.8
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0
0
5×10^7
1×10^8
1.5×10^8
2×10^8
2.5×10^8
3×10^8
3.5×10^8
Iterations

Optimizing task layout on the Blue Gene/L supercomputer

Paper Authors: G. Bhanot A. Gara P. Heidelberger E. Lawless J. C. Sexton R. Walkup
Presentation: Daniel Nichols