SAND2015-7740C

CCR
Center for Computing Research

Enabling Tractable Exploration of the Performance of Adaptive Mesh Refinement

Courtenay T. Vaughan and Richard F. Barrett
Center for Computing Research
Sandia National Laboratories

Workshop on Representative Applications
at IEEE Clusters 2015

September 7, 2015

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.

Sandia National Laboratories

CCR
Center for Computing Research

Motivation

- An implementation of AMR in a mini-app to explore issues with AMR on parallel machines
  - added complexity of AMR bookkeeping
  - refinement frequency
  - load balancing strategies (frequency and methods)
  - effects of indirection
  - effects of block size
  - communication strategies
  - OpenMP strategies (future work)
  - task parallel programming models (future work)

Sandia National Laboratories

CCR
Center for Computing Research

miniAMR

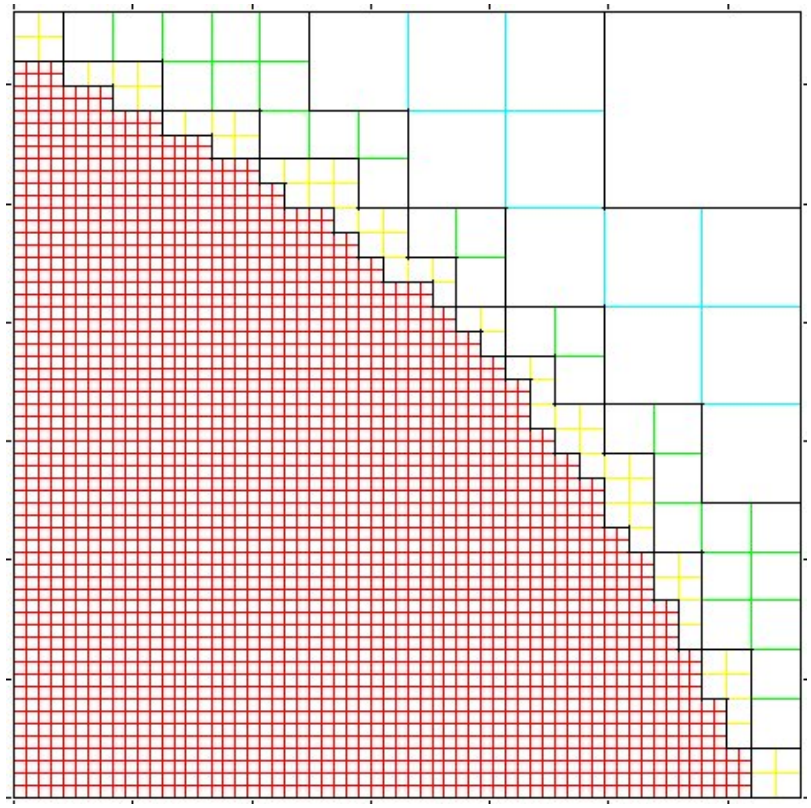
- AMR version of miniGhost – written in C
- Part of the Mantevo Suite (mantevo.org)
- Same finite volume calculation as miniGhost
  - no real physics and little fake physics, but kernel could be easily modified
- Many smaller blocks per processor
- Similar communication strategy to miniGhost
  - May have more communication partners due to the block structure
- Needs load balancing
  - One area may refine while rest does not
- More complicated bookkeeping
  - each block has between 6 and 24 neighbors and parent

Sandia National Laboratories

Details of AMR

- All blocks have same number of cells
- Blocks can only have “1 to 1” or “2 to 1” ratio with neighbors
- Area refine determined by moving shapes through mesh

2D slice of 3D mesh with sphere



CCR
Center for Computing Research

AMR Details Continued

- Initial mesh is a unit cube
- Each processor has an initial number of blocks at the lowest refinement level
- Initially the processors are arranged in a npx x npy x npz grid with position determined by an RCB (Recursive Coordinate Bisection) ordering
- Refinement is controlled by objects that move through the mesh and can change size
- Typical problems for AMR applications will have 4 to 7 levels of refinement

Sandia National Laboratories

CCR
Center for Computing Research

Structure of miniAMR

```
for some number of timesteps {
  for some number of stages {
    communicate ghost values between blocks
    perform stencil calculation on arrays
    if time for checksums
      perform checksum calculations and compare
  }
  if time for refinement
    refine mesh
}
```

Sandia National Laboratories

CCR
Center for Computing Research

Communication

- For each direction, each rank maintains a list of its block's faces that need to be communicated to adjacent ranks
  - ordered by rank
- Communication step for one direction
  - Post receives
  - Pack messages and do sends
  - Do on-rank communication of faces via memory copy
  - Complete receives and unpack messages

Sandia National Laboratories

CCR
Center for Computing Research

Refinement

- When a block is refined, it is replaced by 8 blocks (2 x 2 x 2) each being half the physical size in each direction, but with the same number of cells
- The original block's communications in the lists are revised to reflect the new blocks
- A parent block is created to replace the original block
  - Stays on rank where created during load balancing
- Coarsening is done similarly except that all eight blocks need to be on the same rank as the parent before they can be consolidated

Sandia National Laboratories

CCR
Center for Computing Research

Refinement (continued)

- Marking blocks for refinement is done by levels starting with the most refined blocks
  - Refining a block can cause its neighbors to refine or prevent them from unrefining
- After each level is marked then the results are communicated and then the next level can be marked
- Blocks that are marked to be refined are refined, changes to the mesh are communicated, and then any blocks that need to be consolidated are

Sandia National Laboratories

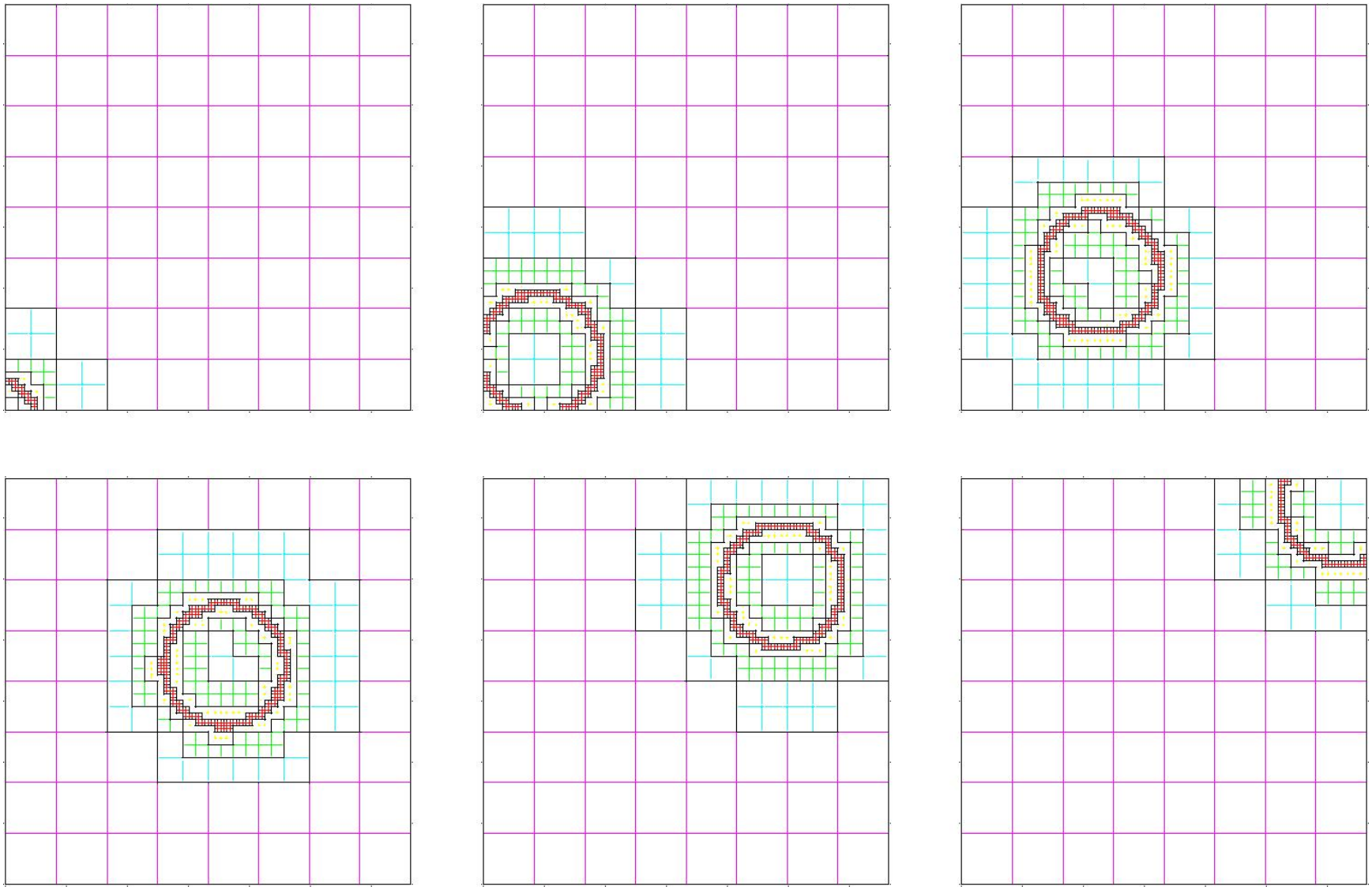
CCR
Center for Computing Research

Load Balancing

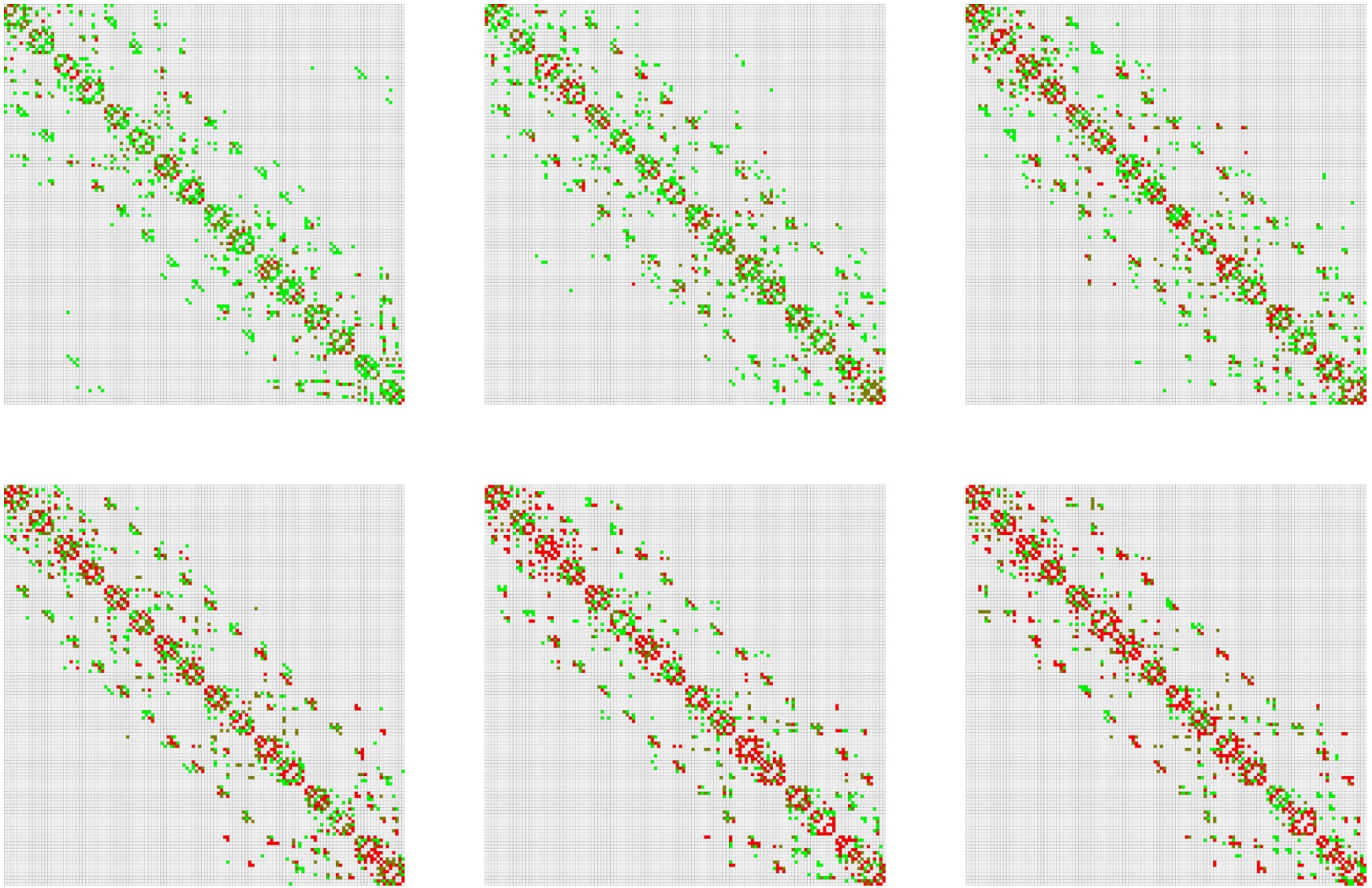
- After blocks are refined (or unrefined), then load balancing is done
- We use Recursive Coordinate Bisection (RCB) with the directions fixed during initialization
  - This keeps data movement down
- At each step, a group or ranks and associated blocks are divided into some number of sets and then the process is repeated for each set
- Since block locations in a direction are limited, we represent the centers by an integer, and determining the cut can be done by binning the centers

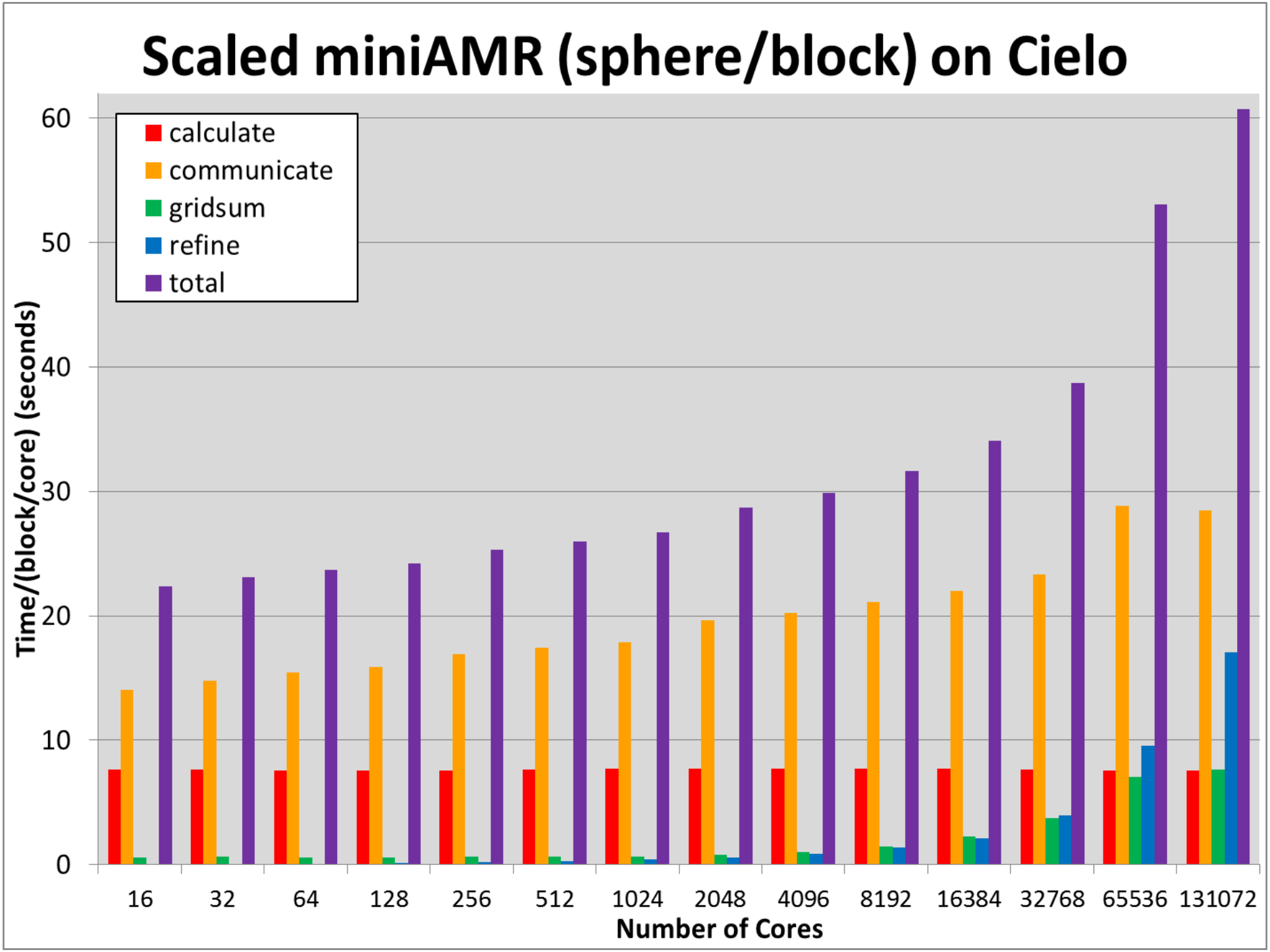
Sandia National Laboratories

miniAMR – block structure – hollow sphere



Ghost Value Communication Pattern Changes




Scaled miniAMR (sphere/block) on Cielo
calculate
communicate
gridsum
refine
total
Time/(block/core) (seconds)
0
10
20
30
40
50
60
16
32
64
128
256
512
1024
2048
4096
8192
16384
32768
65536
131072
Number of Cores

CCR
Center for Computing Research

Comments on Scaled Speedup Curve

- Communication dominates the time and is increasing gradually
  - Includes time to communicate boundary information on blocks on the same core (30.6% of communication time on 128 cores)
- Calculation time is a consistent amount of time per block
  - If completely refined, then the 128 core problem would have 524288 blocks instead of 18168 and the calculation time would be 218 seconds instead of 7.6 seconds
- The refinement and gridsum times both are increasing gradually
- These reflect tradeoffs that AMR makes to allow problems to be run in less time on fewer nodes

Sandia National Laboratories

CCR
Center for Computing Research

CTH

- Three-dimensional shock hydrodynamics code
- In AMR mode, each processor has a number of smaller blocks and typically sends more smaller messages
  - Communication pattern changes during run
- During each timestep, there are several stages, each of which has a ghost value exchange, and some number of collective operations
  - For problems we are using, there are 17 boundary exchanges and 62 collectives per timestep

Sandia National Laboratories

CCR
Center for Computing Research

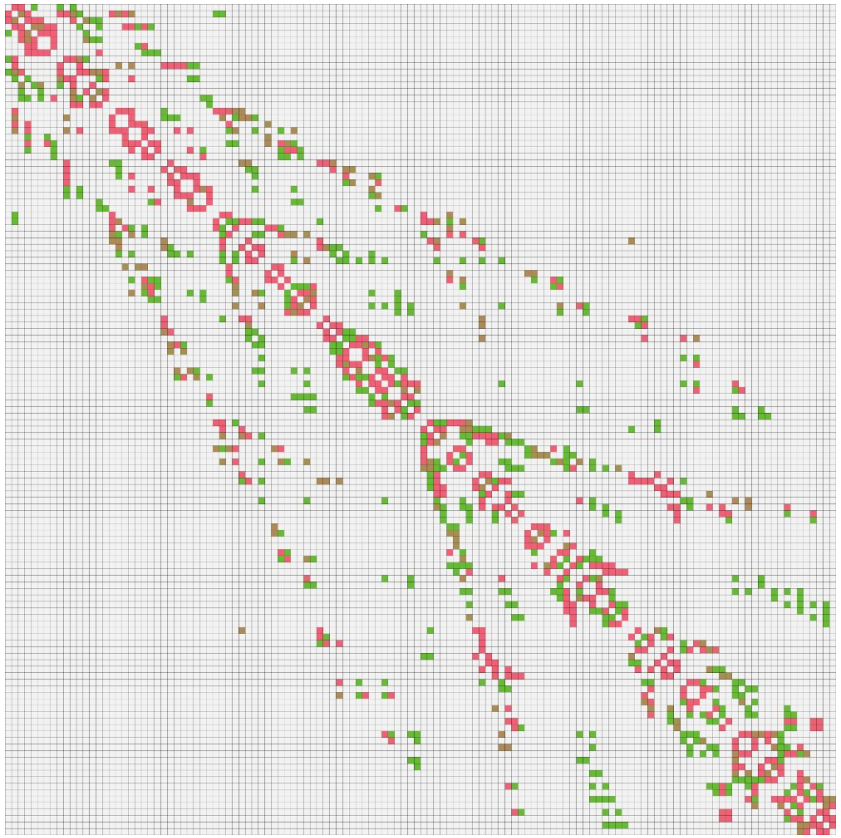
Comparison with CTH

- Run Sphere hits Block problem on 128 cores
- CTH problem is a sphere that hits a block at an oblique angle and produces a shock wave
  - modeled in miniAMR as a deforming spheroid with an expanding hemisphere to represent the shock
- CTH averages 140.9 blocks/core over the run
  - average core has 16.3 messages per communication stage that average 261 KB
- miniAMR averages 141.9 blocks/core over the run
  - average core has 18.4 messages per communication stage that average 224 KB

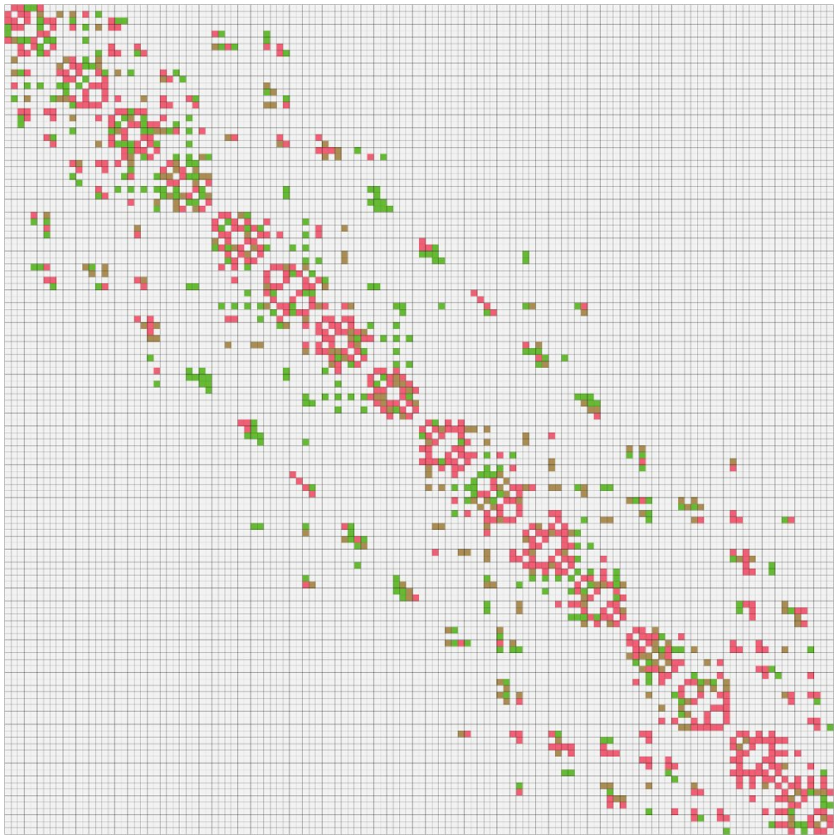
Sandia National Laboratories

Communication Matrices (sphere hits block)

CTH

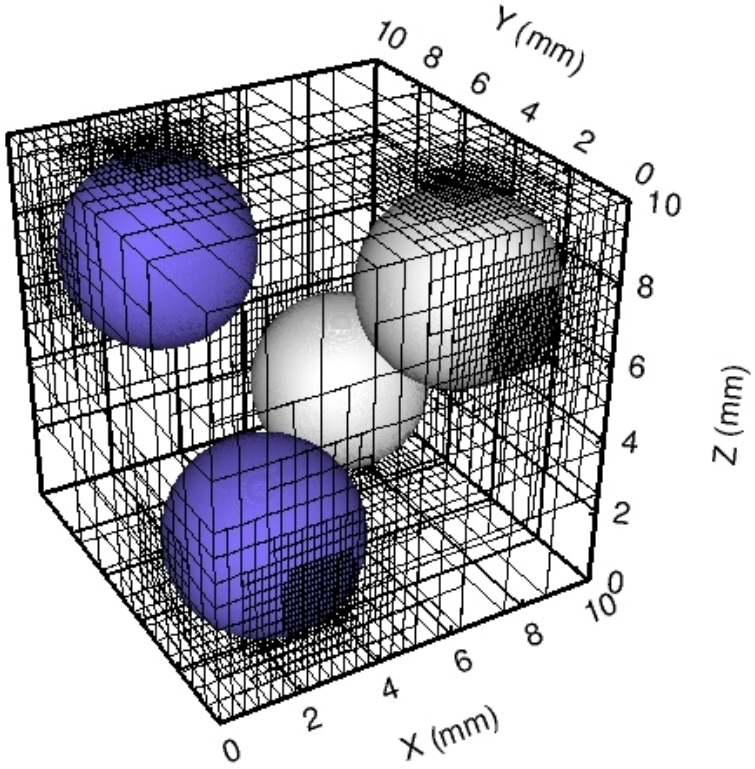


miniAMR



Comparison with CTH (Four Spheres)

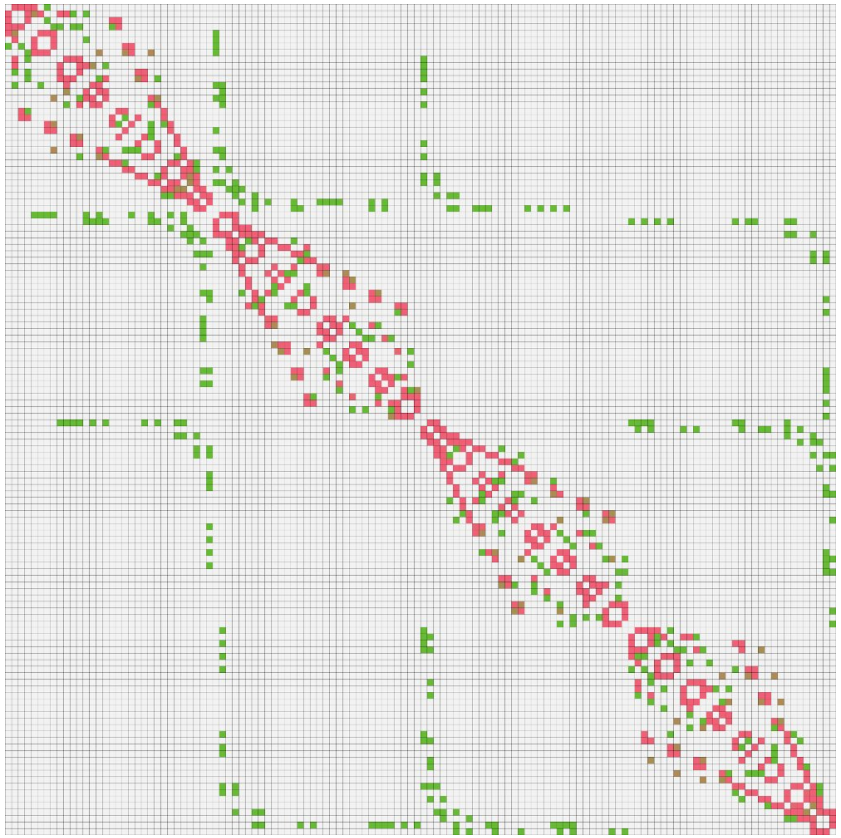
- Run on 128 cores
- CTH
  - 685.8 blocks/rank
  - 18.4 messages
  - 503 KB average
- miniAMR
  - 669.3 blocks/rank
  - 17.3 messages
  - 593 KB average



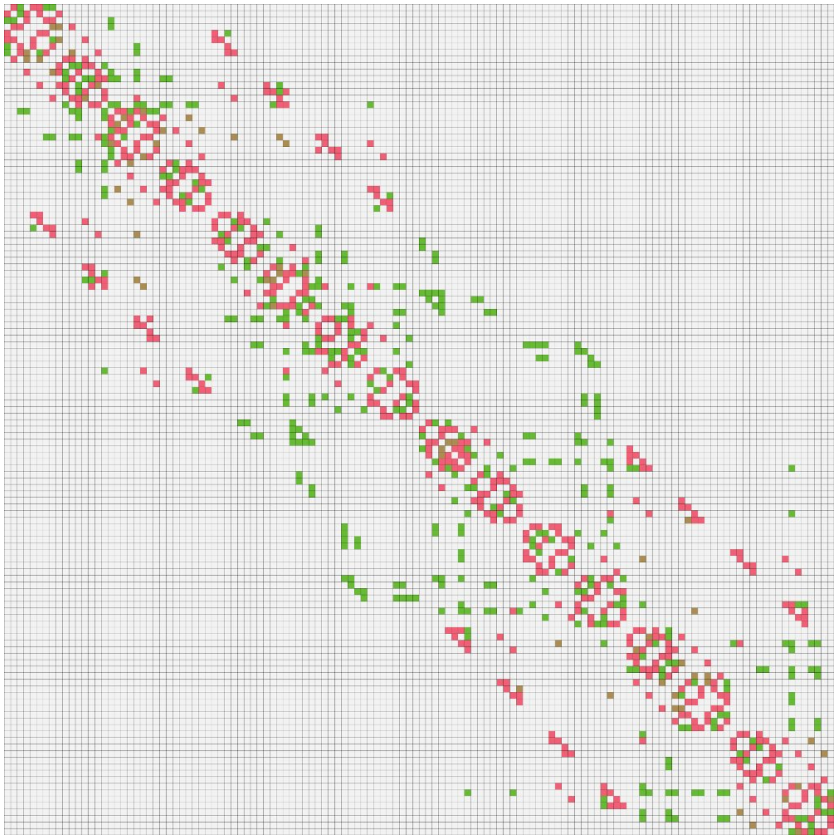
Communication Matrices
Four Spheres

CCR
Center for Computing Research

CTH



miniAMR



Sandia National Laboratories

CCR
Center for Computing Research

Communication Differences

- Communication patterns are dependent on the load balancing after refinement
- Three differences between CTH and miniAMR
  - For CTH when a cut is made and there are ties, those blocks are assigned in a random fashion, while miniAMR blocks are assigned based on their position in the cut plane
  - CTH limits the number of blocks that can be moved at any timestep, while miniAMR has no limit
  - CTH allows the cut directions in RCB to be determined when the cuts are made, while these are fixed for miniAMR at initialization

Sandia National Laboratories

CCR
Center for Computing Research

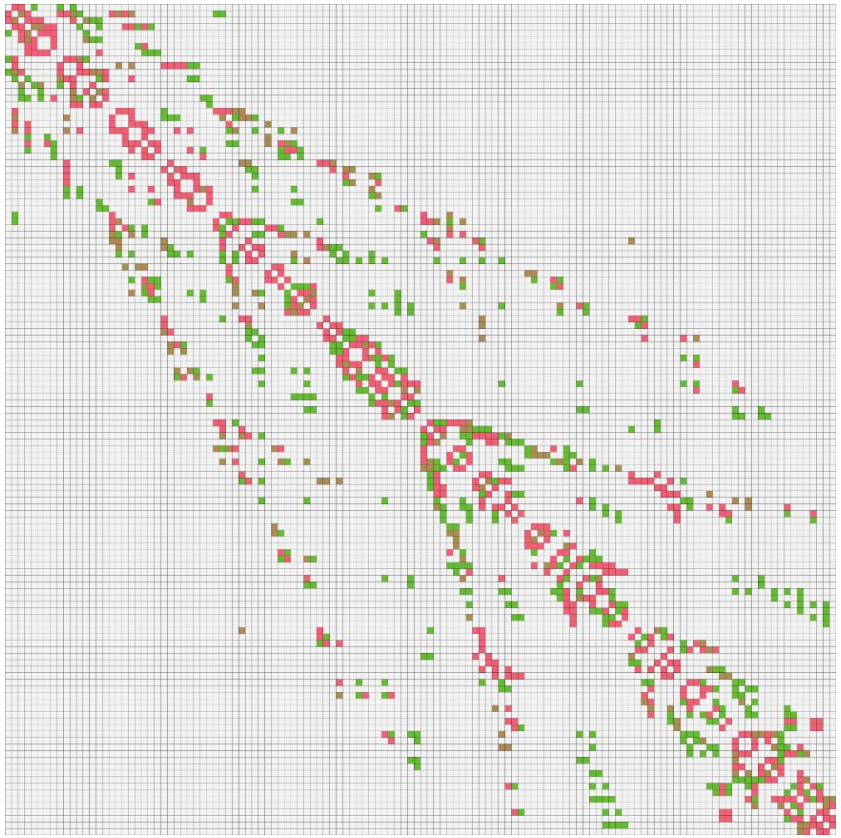
Modifications to miniAMR load balancing

- Modified miniAMR load balancing to mimic that of CTH
- For Four Spheres problem, the number of blocks moved increased by a factor of 8 and the refinement time tripled
- In addition, the communication time increased by 14% due to the number of messages and size increasing

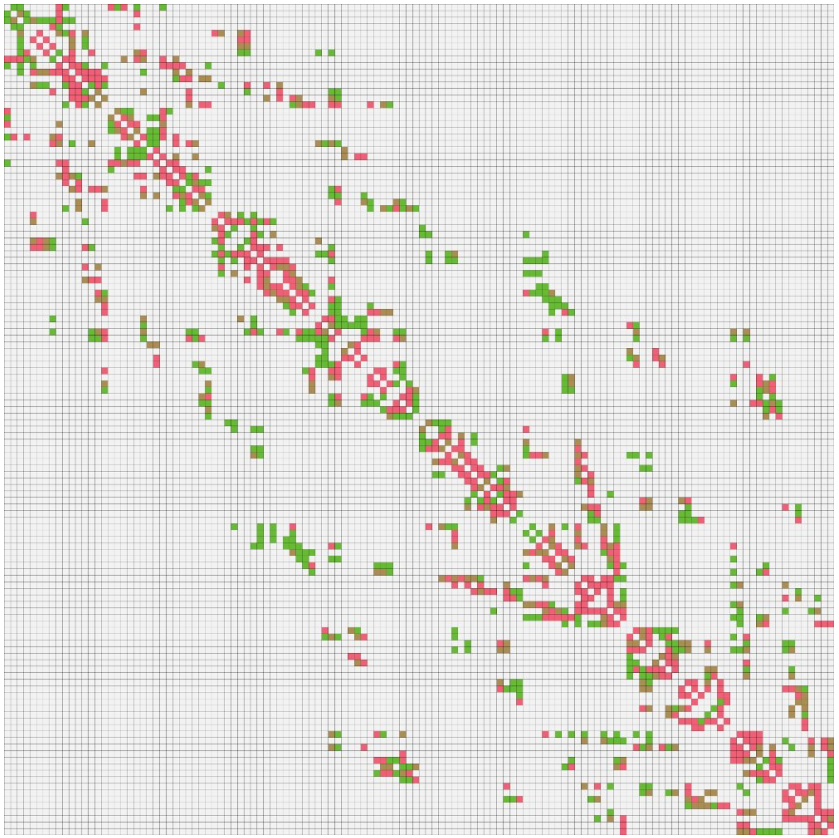
Sandia National Laboratories

Communication Matrices
Sphere hits block

CTH



modified miniAMR

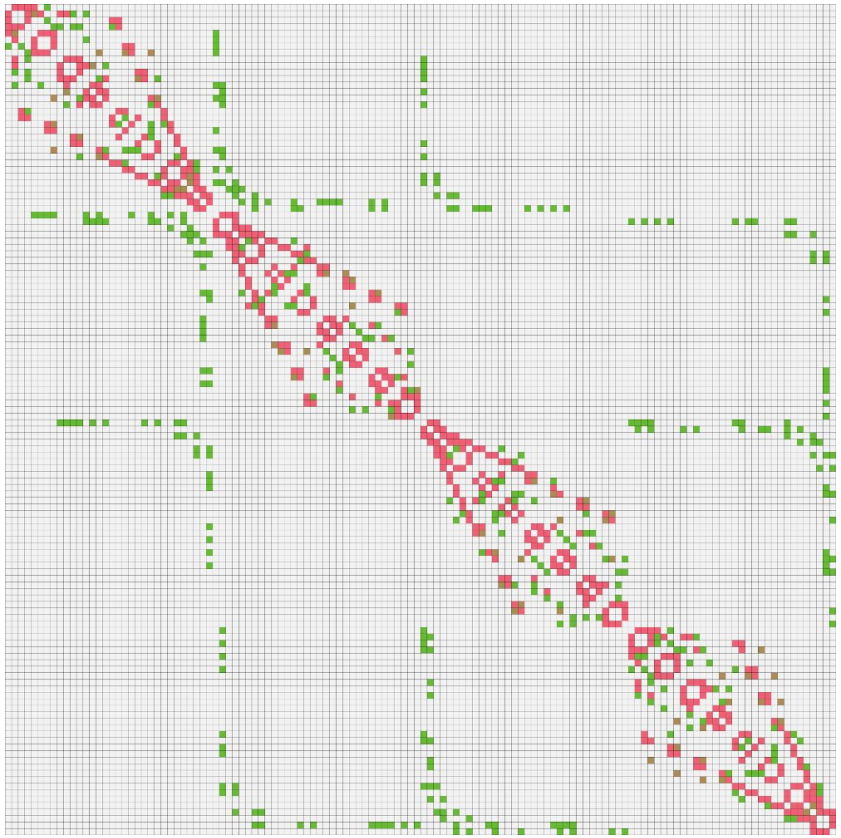


CCR
Center for Computing Research

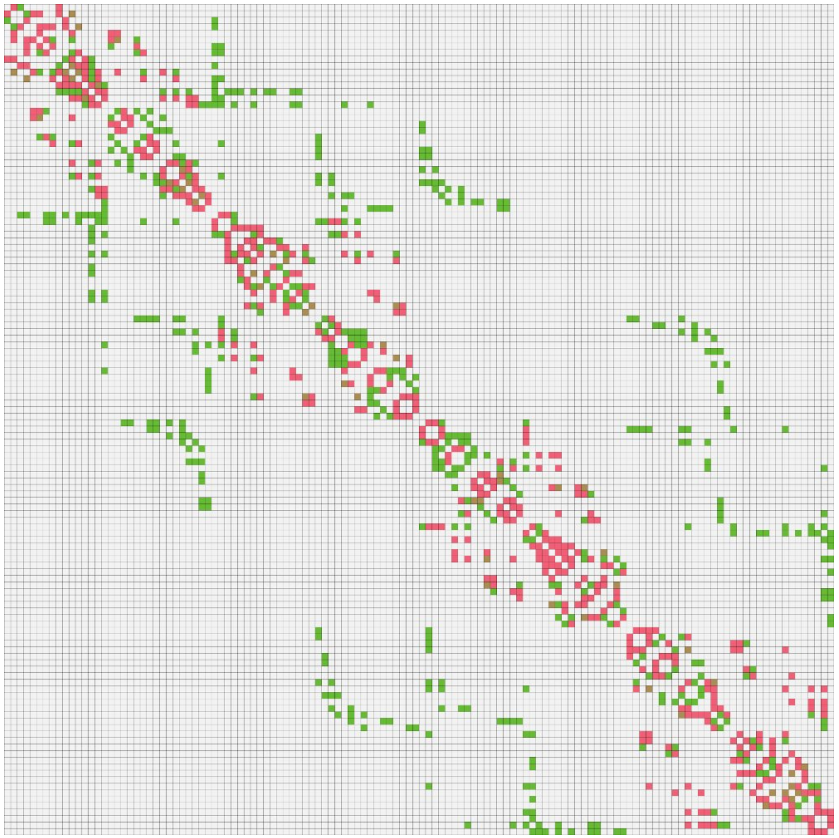
Sandia National Laboratories

Communication Matrices
Four Spheres

CTH



modified miniAMR



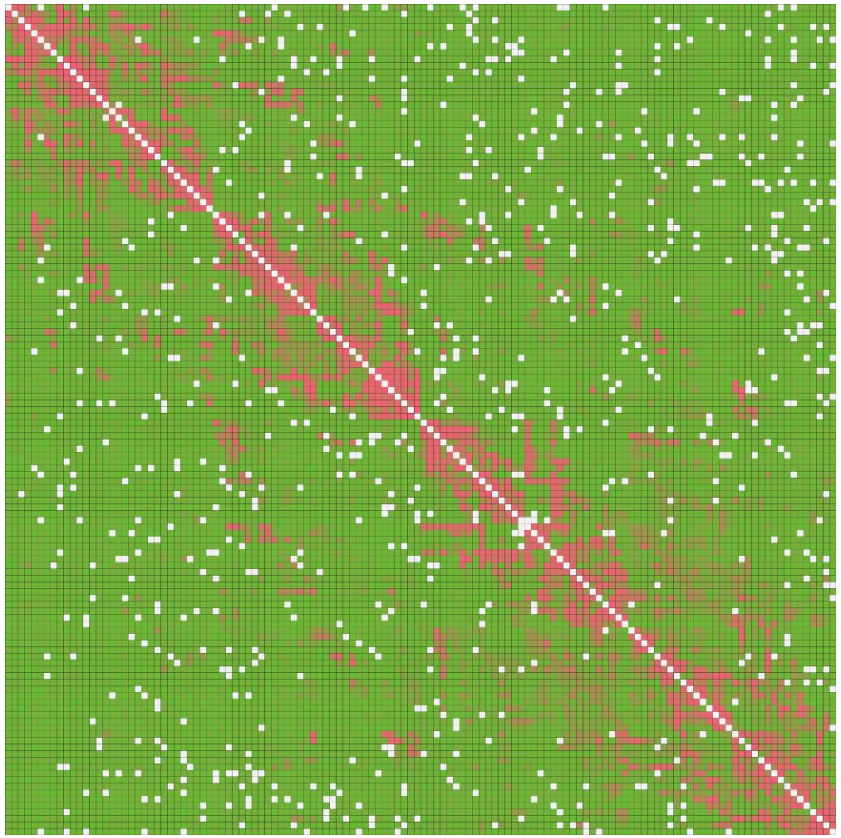
CCR
Center for Computing Research

Sandia National Laboratories

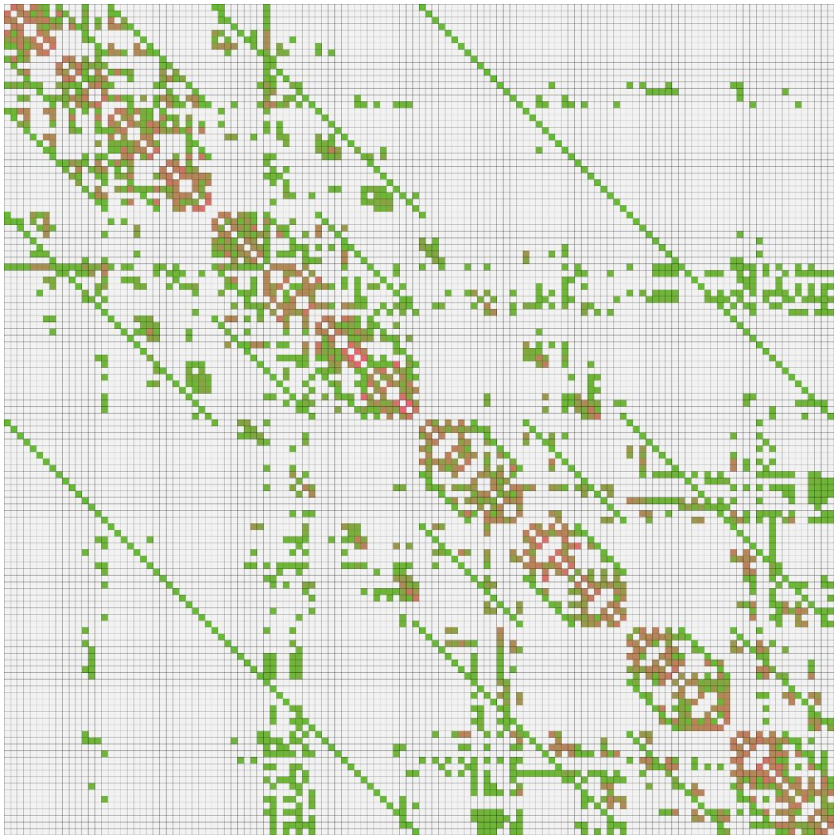
Communication for Refinement Step
Sphere hits Block

CCR
Center for Computing Research

CTH



miniAMR



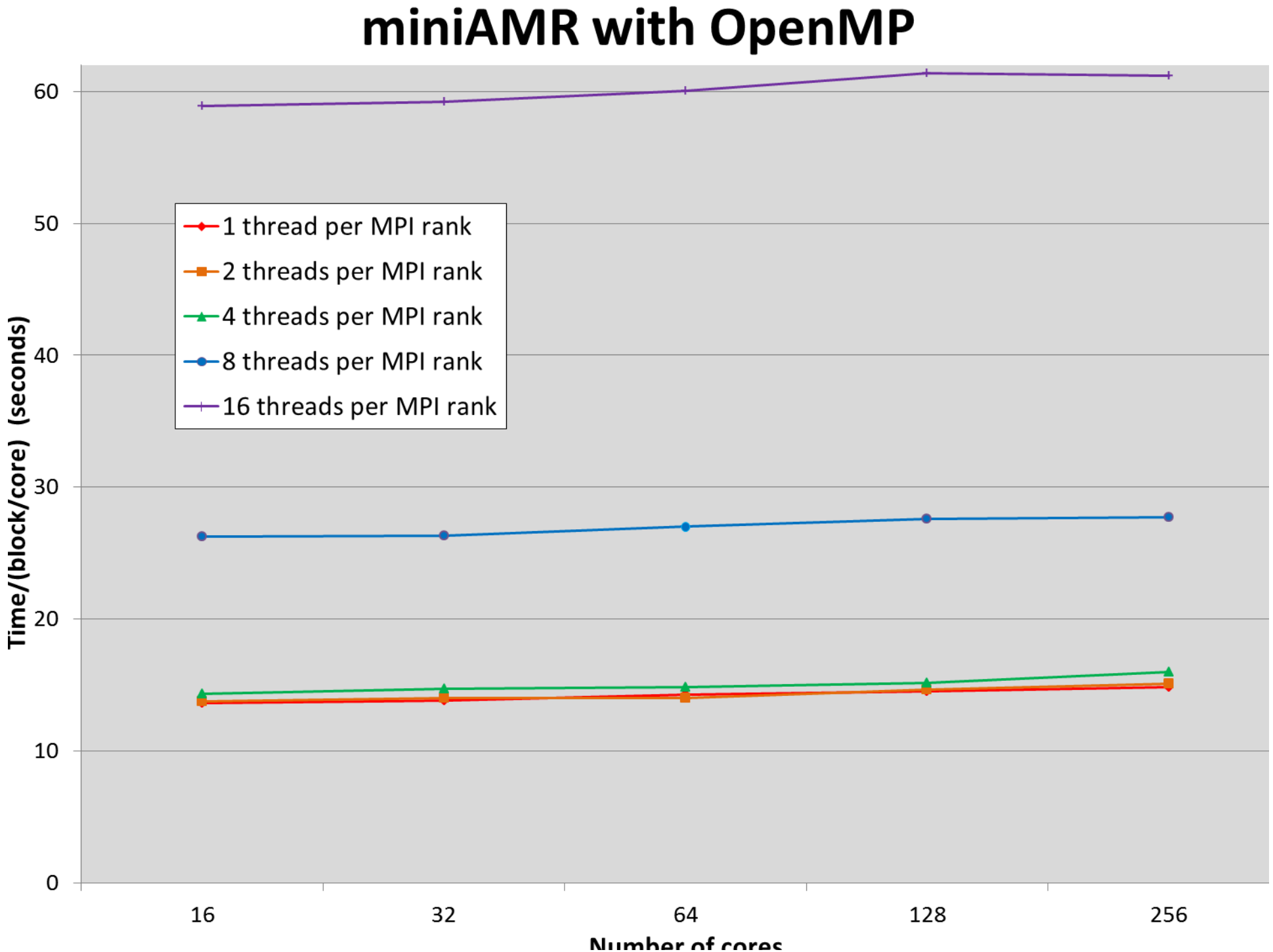
Sandia National Laboratories

CCR
Center for Computing Research

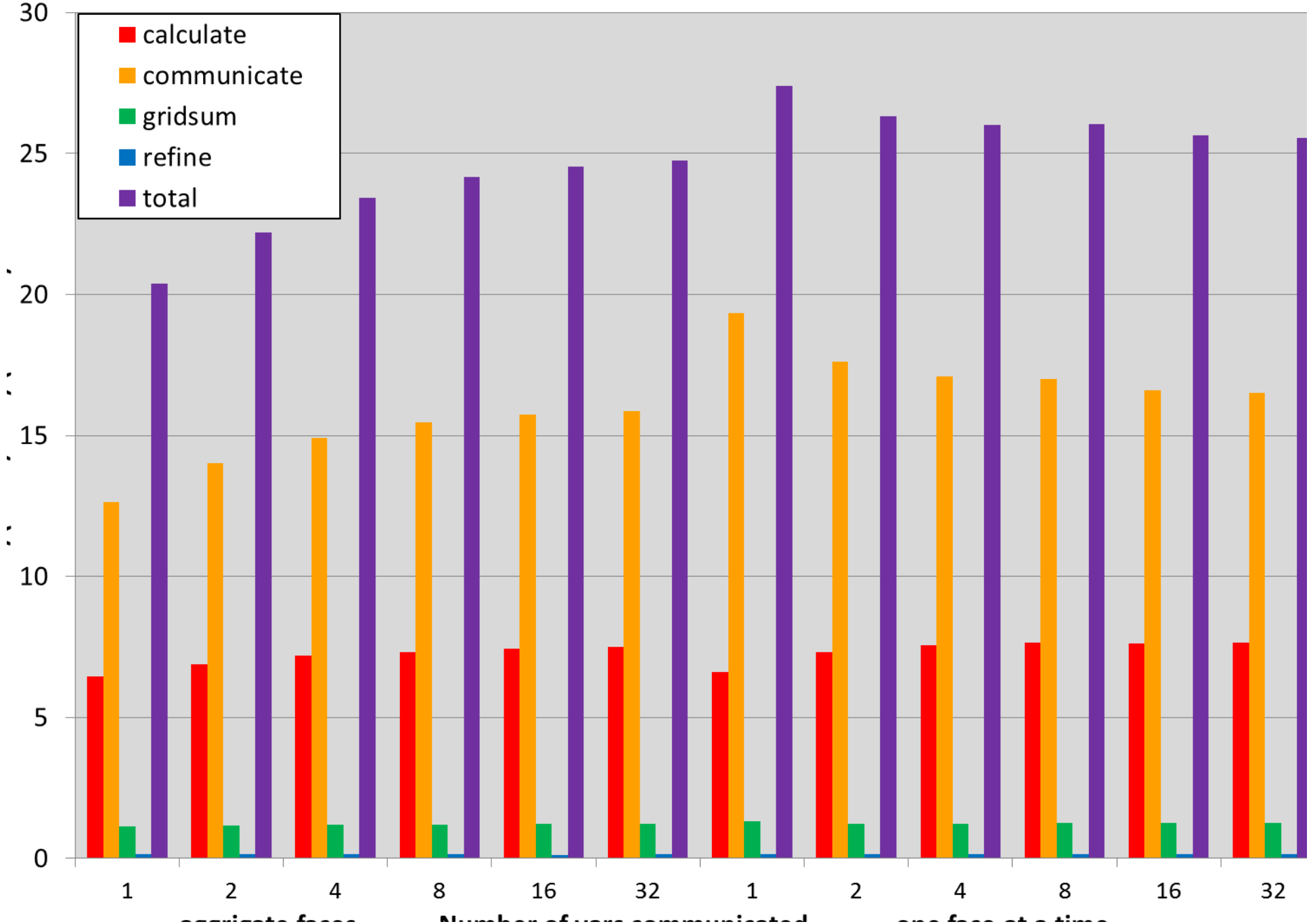
Refinement Step Differences

- Refinement step communication has the regular communication pattern embedded in it since information about what blocks are being refined has to be passed to neighboring blocks
- Diagonal lines in miniAMR matrix is communication for load balancing
- Large amount of communication for CTH is communication with parent blocks since CTH load balances those parent blocks
- CTH uses 34 times as many messages and communicates 54 times as much information for refinement than does miniAMR

Sandia National Laboratories


miniAMR with OpenMP
60
50
40
30
20
10
0
Time/(block/core) (seconds)
1 thread per MPI rank
2 threads per MPI rank
4 threads per MPI rank
8 threads per MPI rank
16 threads per MPI rank
16
32
64
128
256
Number of cores

miniAMR Sphere/Block Problem on 128 cores



CCR
Center for Computing Research

Conclusions and Future Directions

- miniAMR can be fairly representative of the communication portion of CTH in AMR mode
  - We have explained the differences in the codes
- We are planning to use what we have learned from miniAMR to improve CTH
- We are planning to improve the OpenMP implementation of miniAMR
- We are working on a task-parallel version of miniAMR
- We are working on other changes to miniAMR to look at varying workloads among blocks

Sandia National Laboratories