Scalability Challenges for Massively Parallel AMR Applications

Terry J. Ligocki
tjligocki@lbl.gov

Co-authors:
Brian Van Straalen, John Shalf,
Noel Keen, Woo-Sun Yang

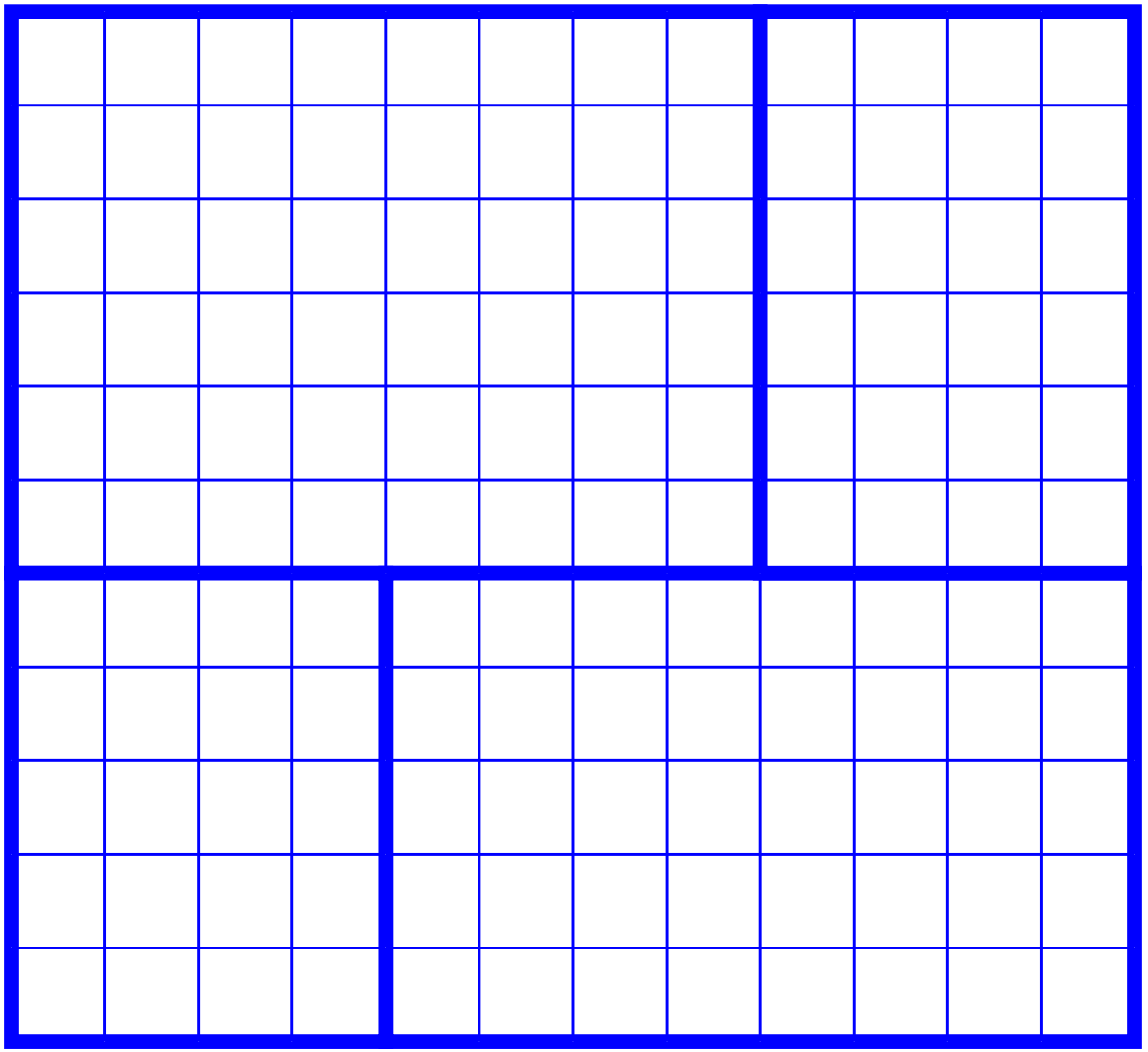
ANAG and NERSC
Lawrence Berkeley National Laboratory
Berkeley, CA, USA

IPDPS 2009

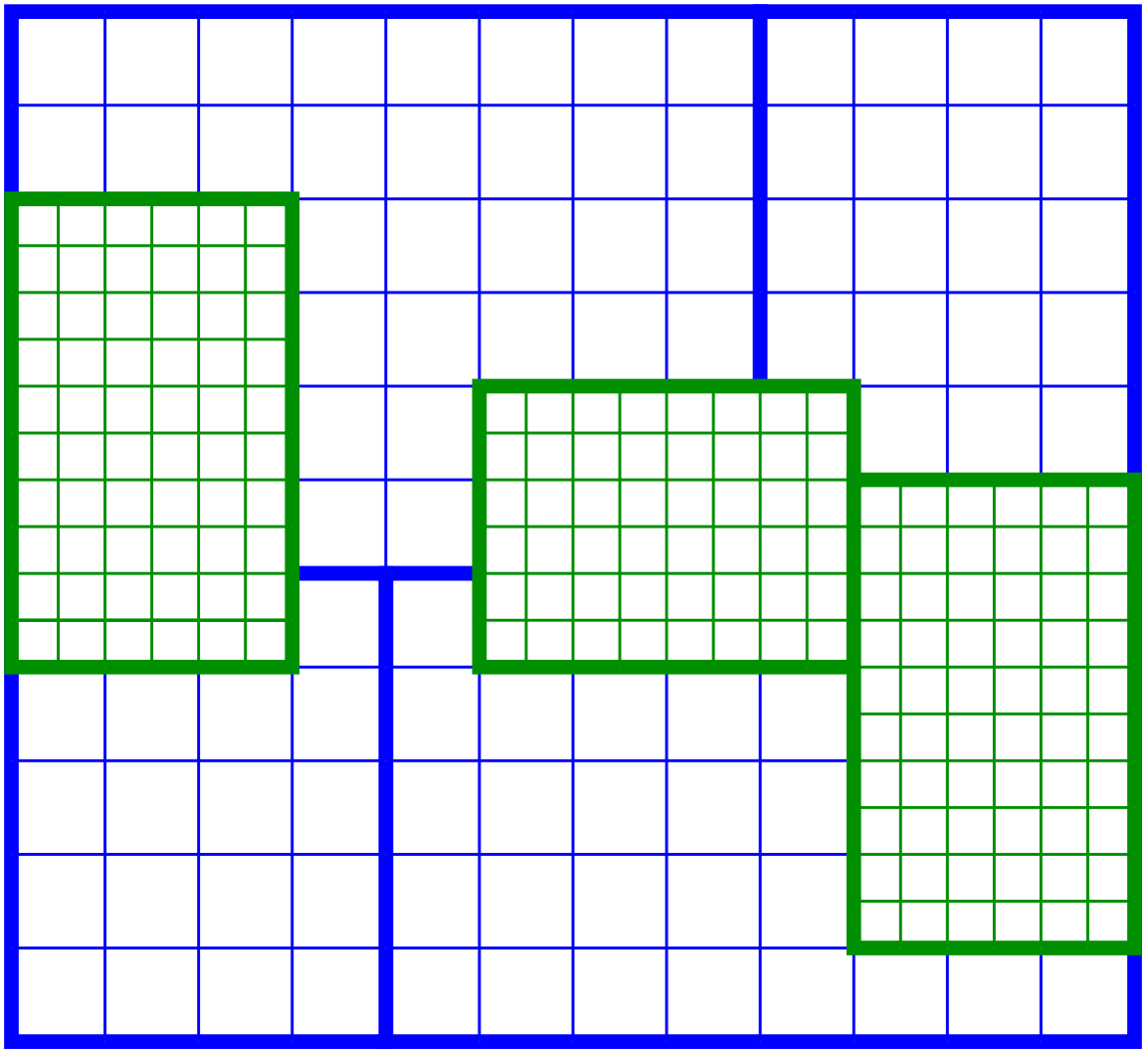
Brief Story

- Block structured AMR library and applications
- Benchmarks to test parallel performance
- Timers for performance measurements
- Ran on several Cray XT supercomputers
- Most weak scaling issues straightforward
- One unusual problem
- Six people and six months to explain and correct
- Weak scaling to thousands of processors
- Need better tools for quantifying and understanding complex systems interactions

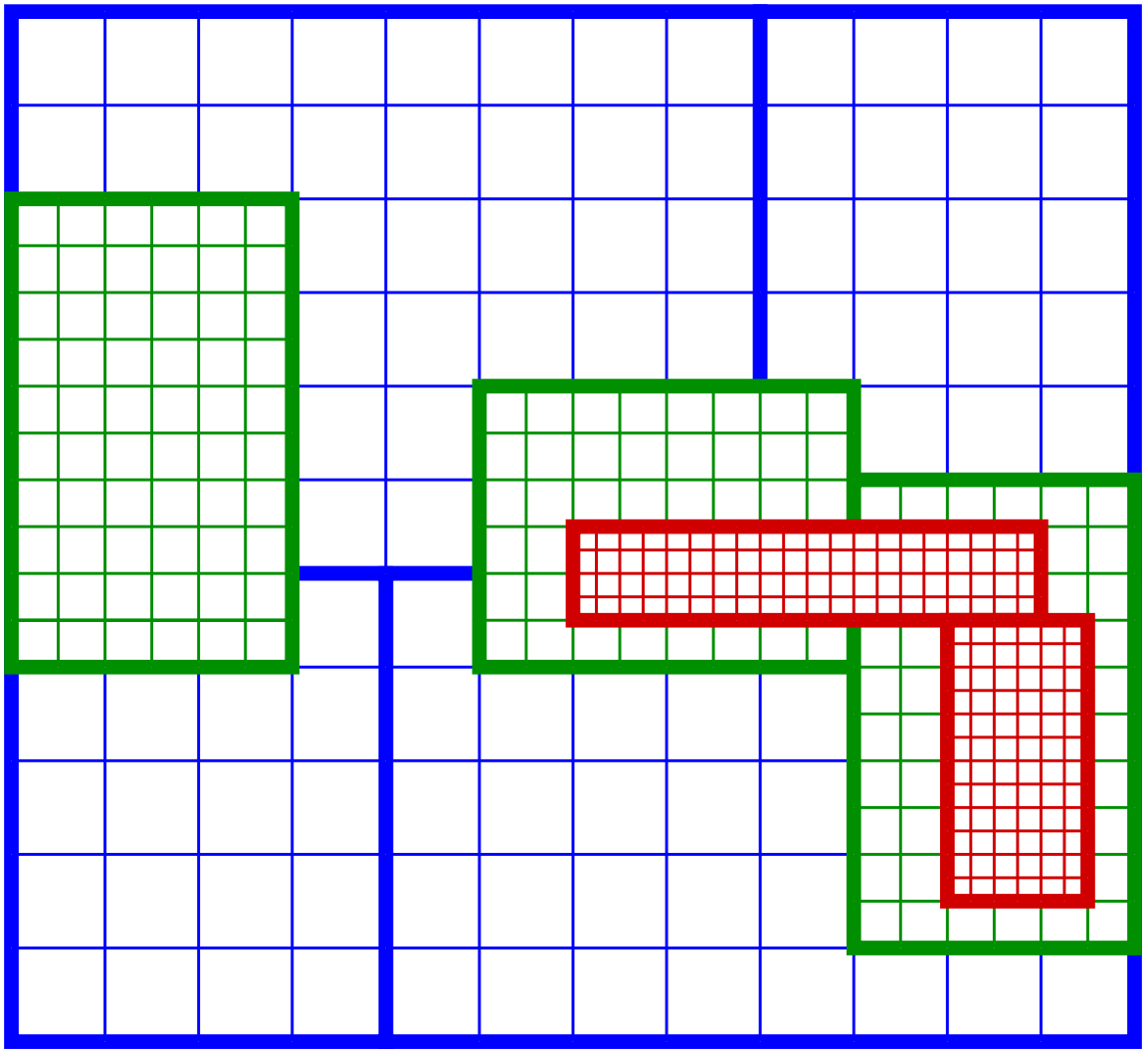
Block Structured AMR



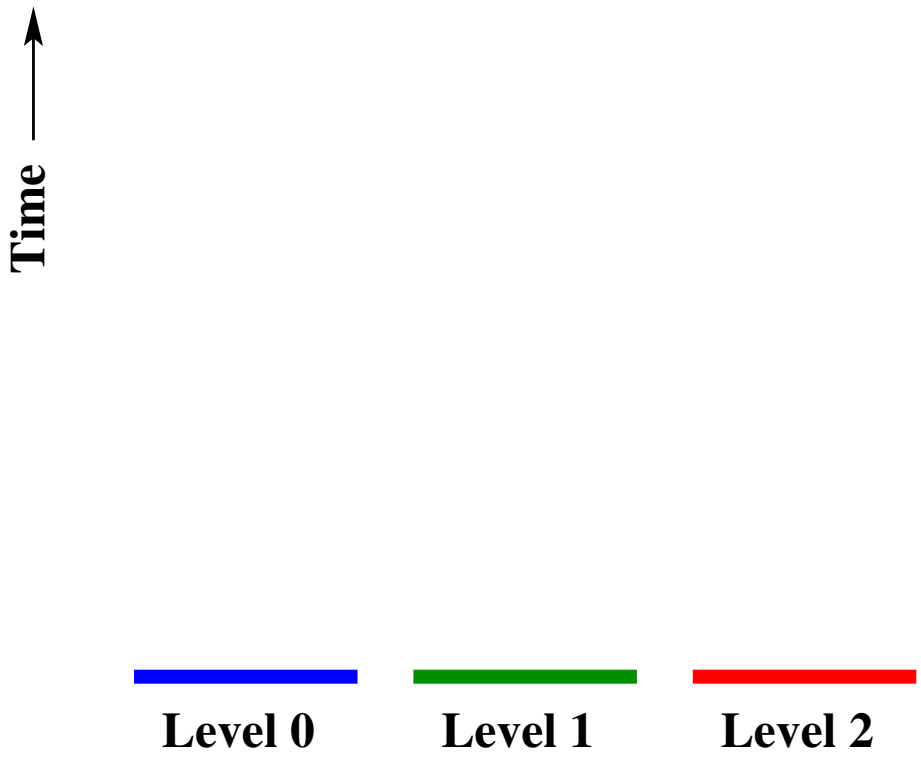
Block Structured AMR



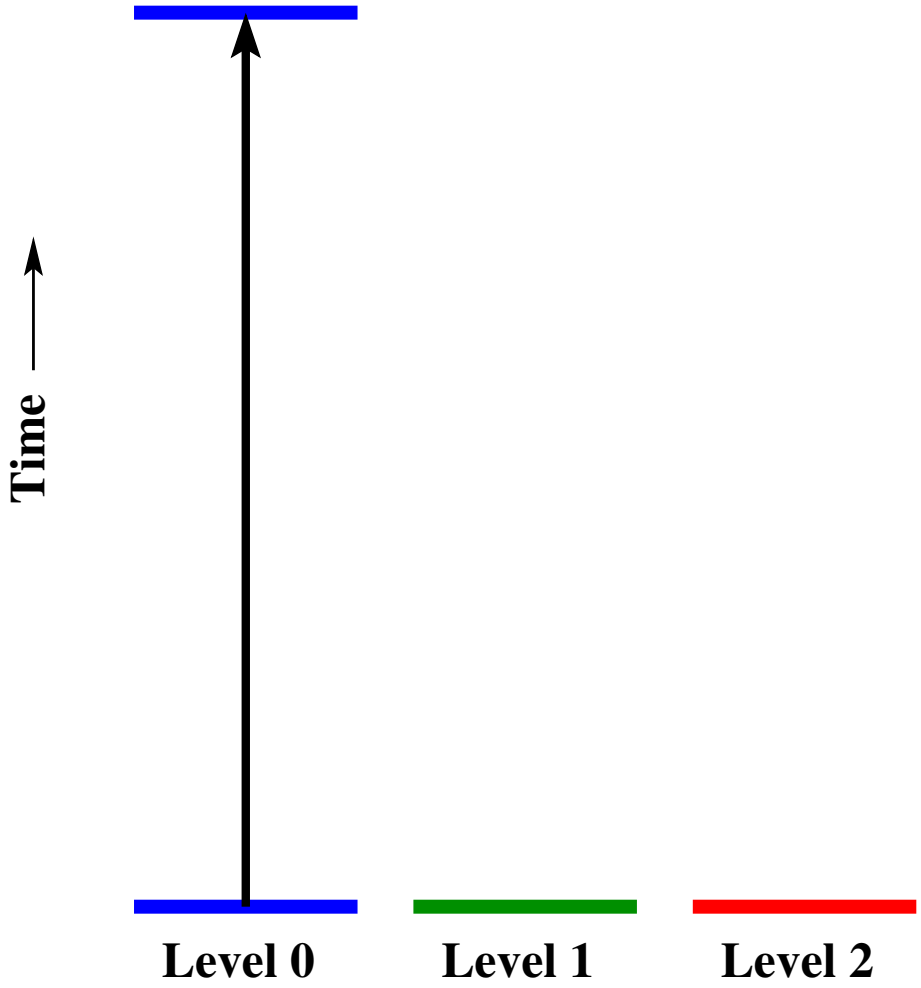
Block Structured AMR



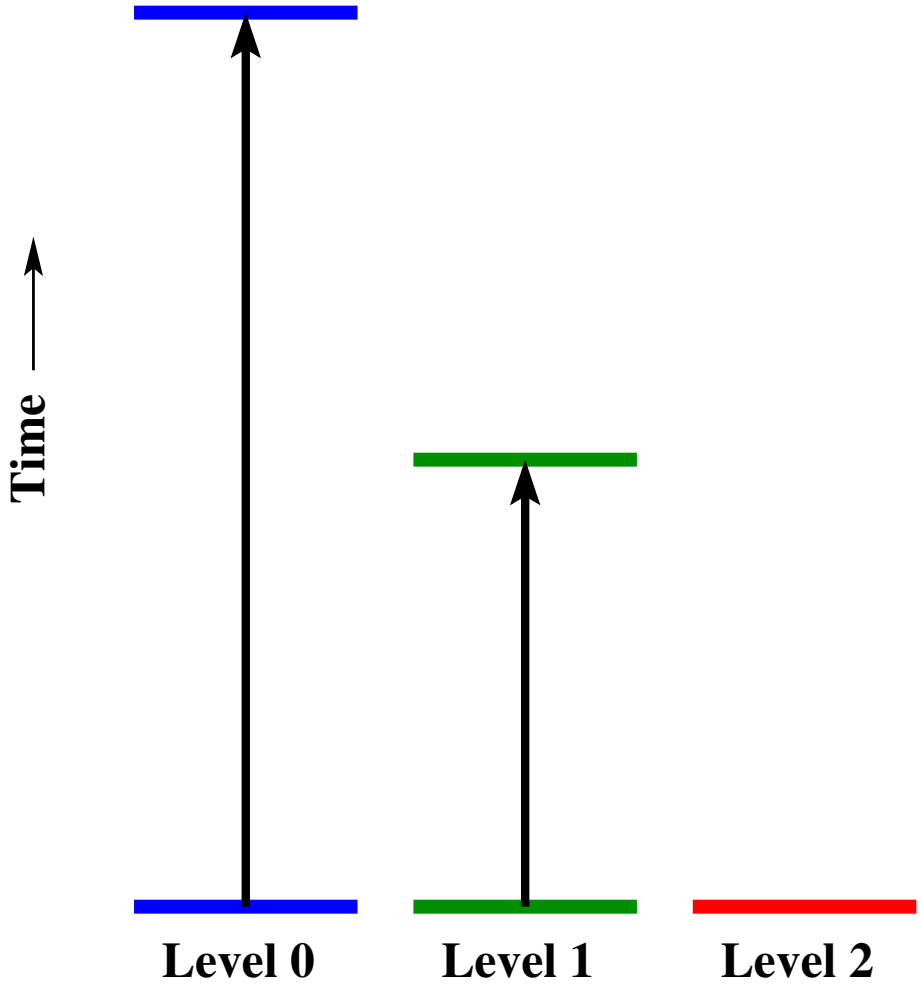
Time Advance



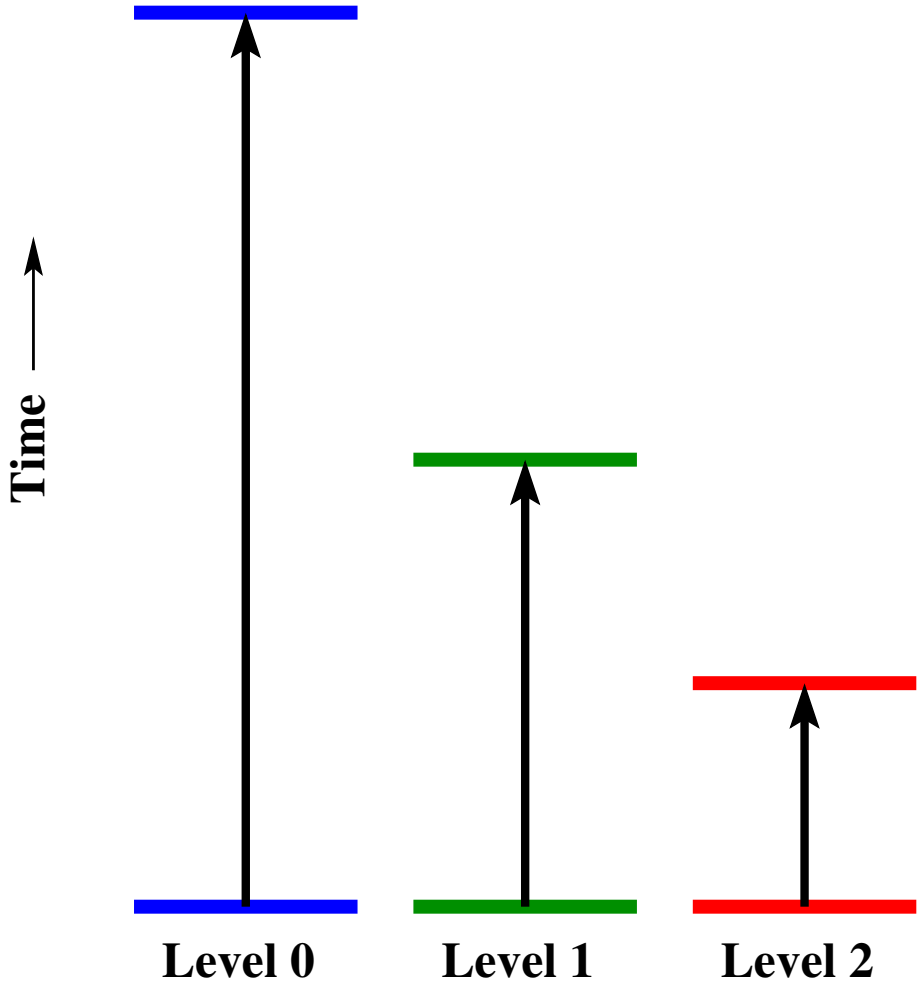
Time Advance



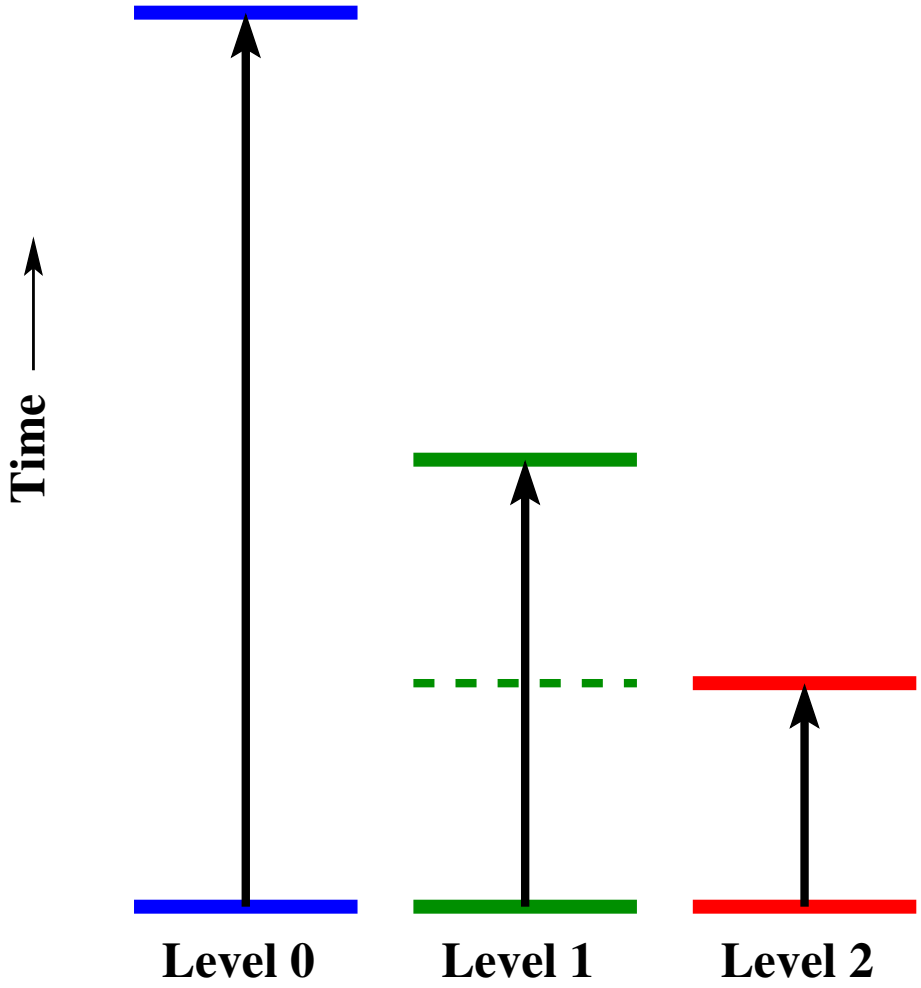
Time Advance



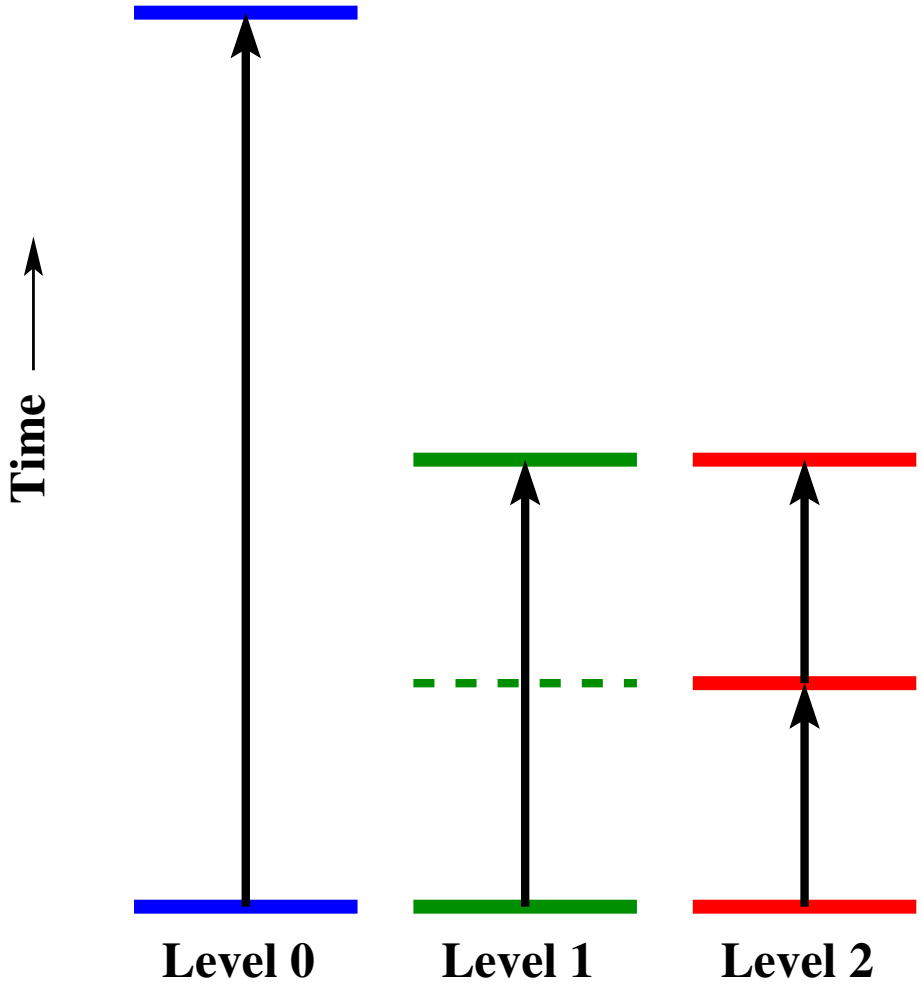
Time Advance



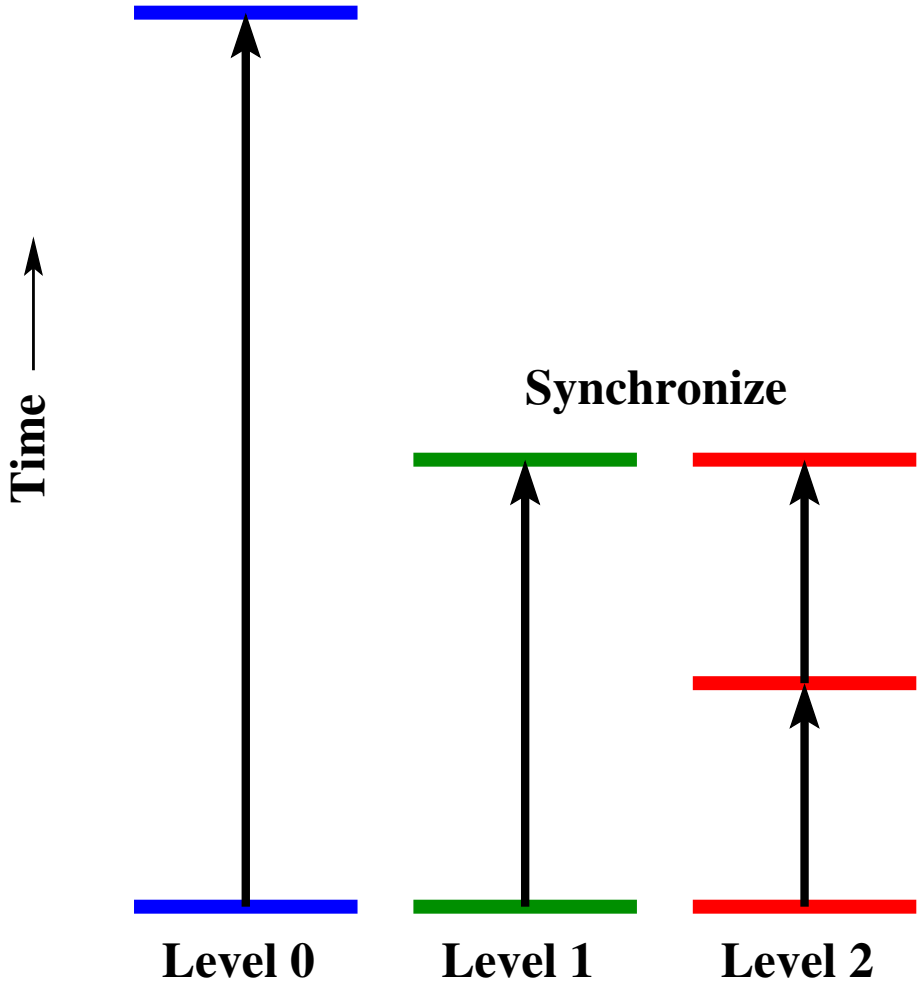
Time Advance



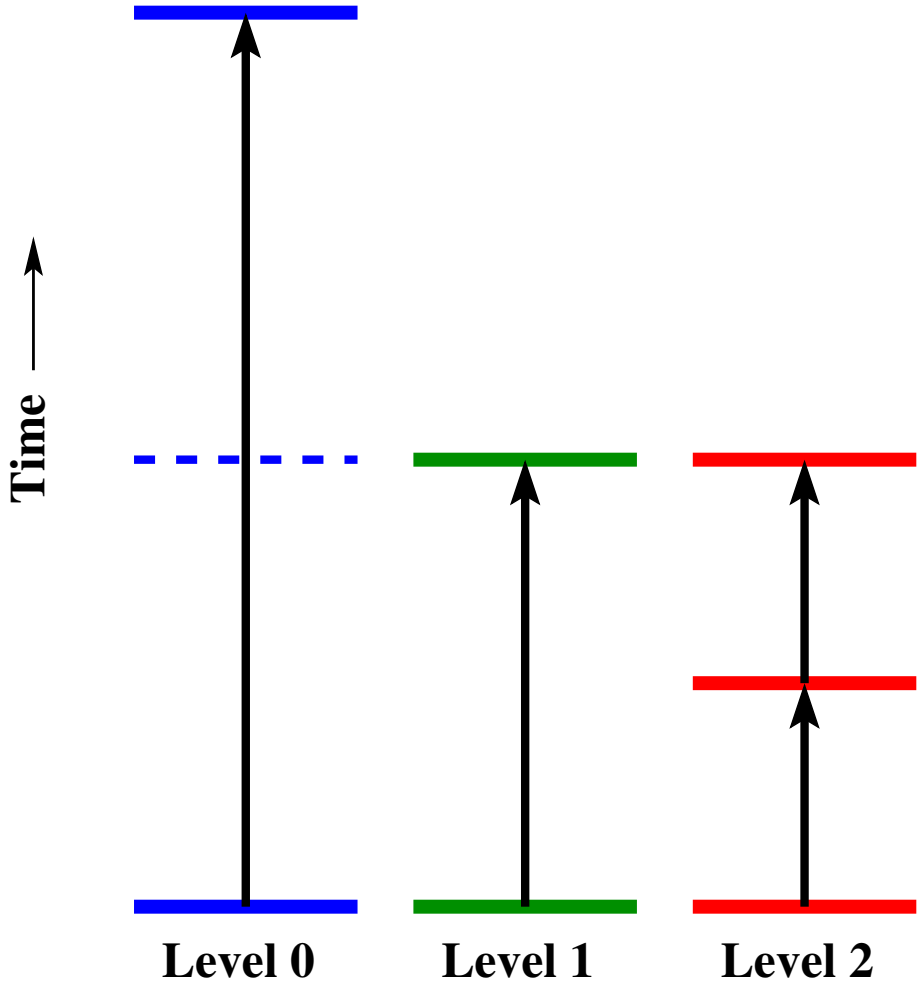
Time Advance



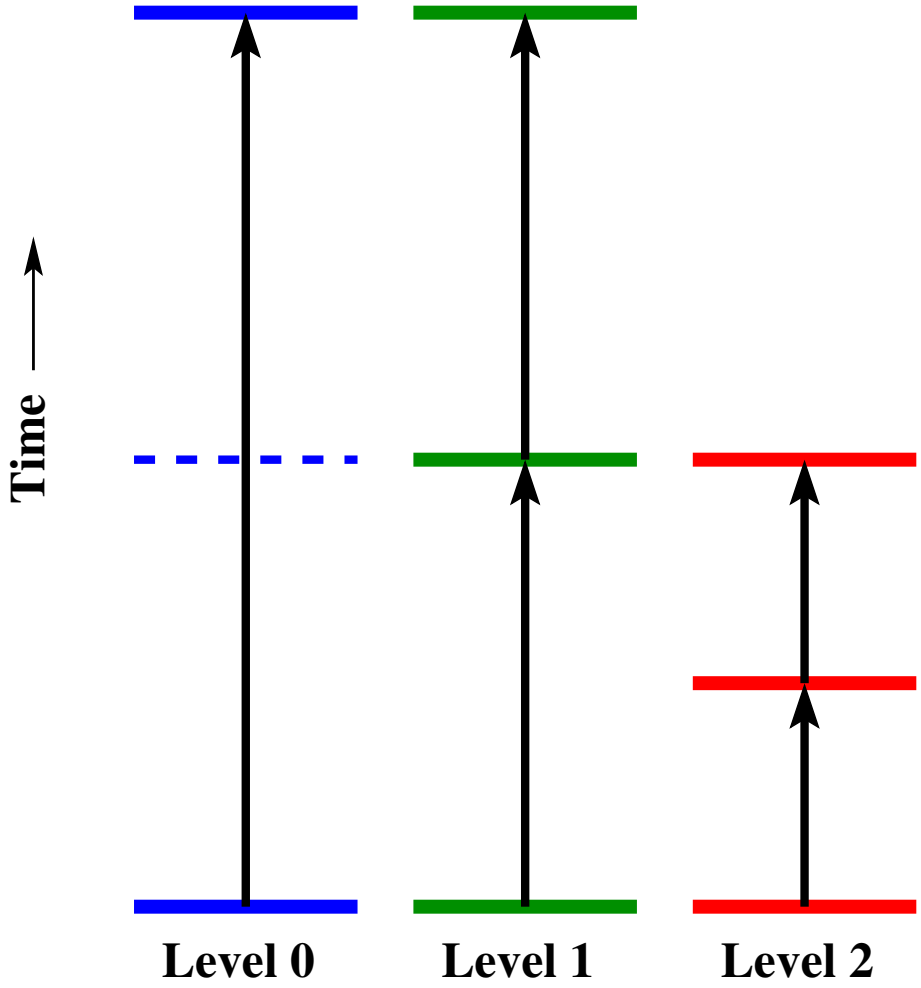
Time Advance



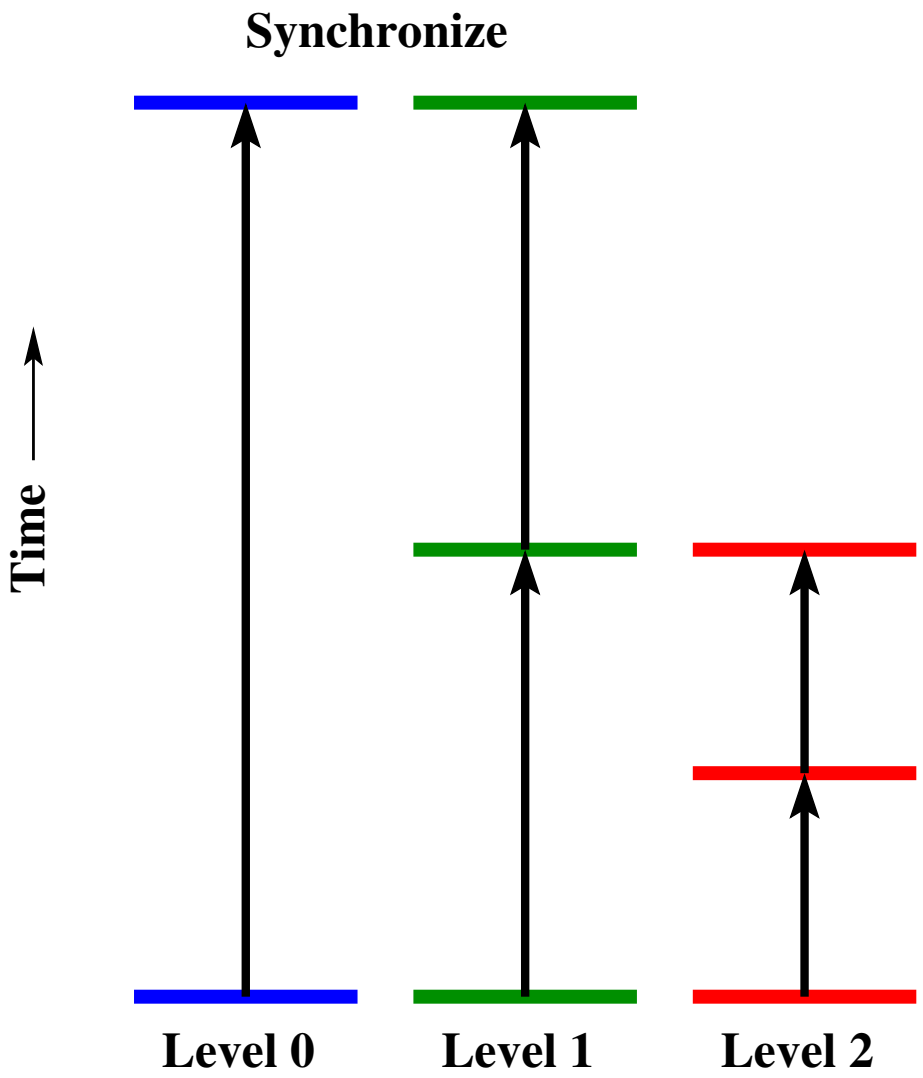
Time Advance



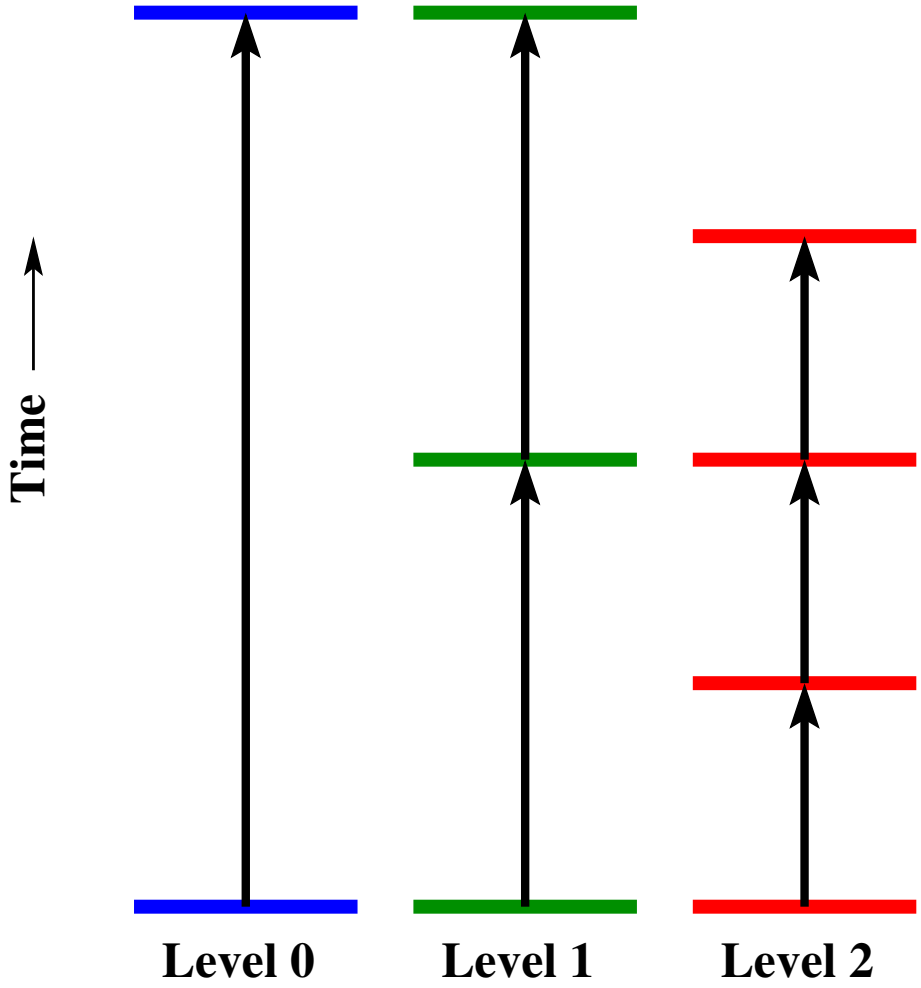
Time Advance



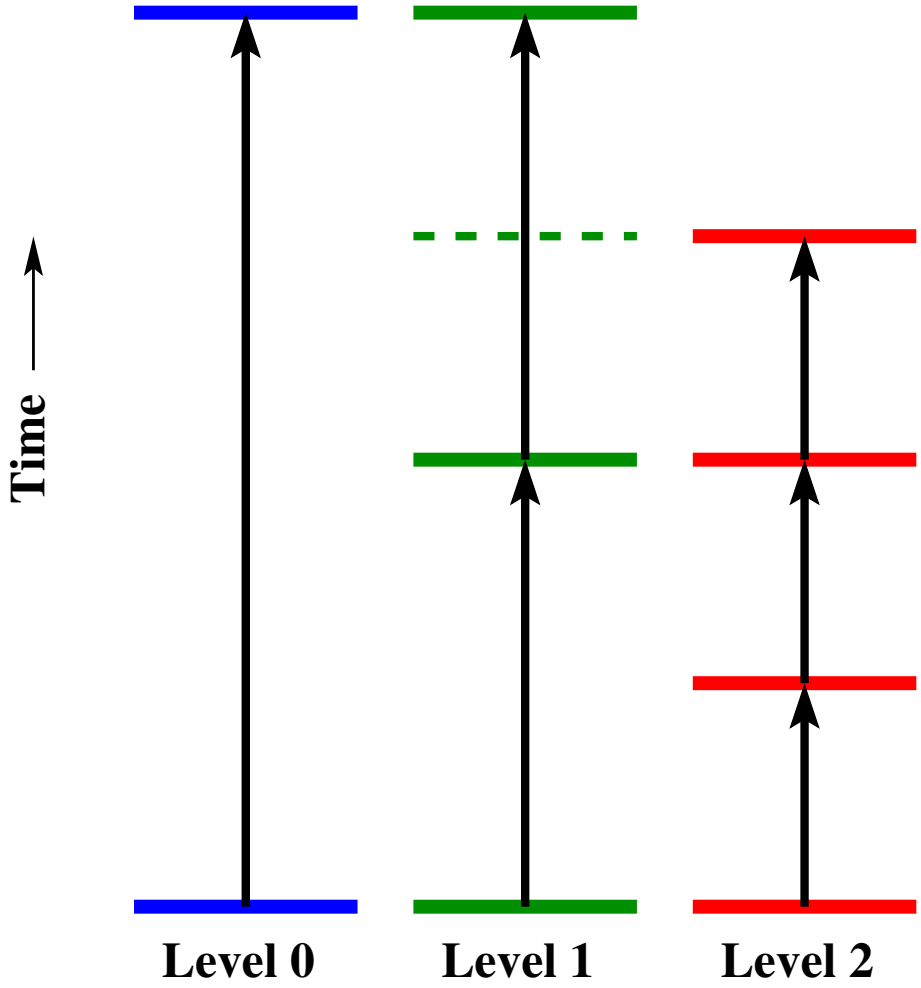
Time Advance



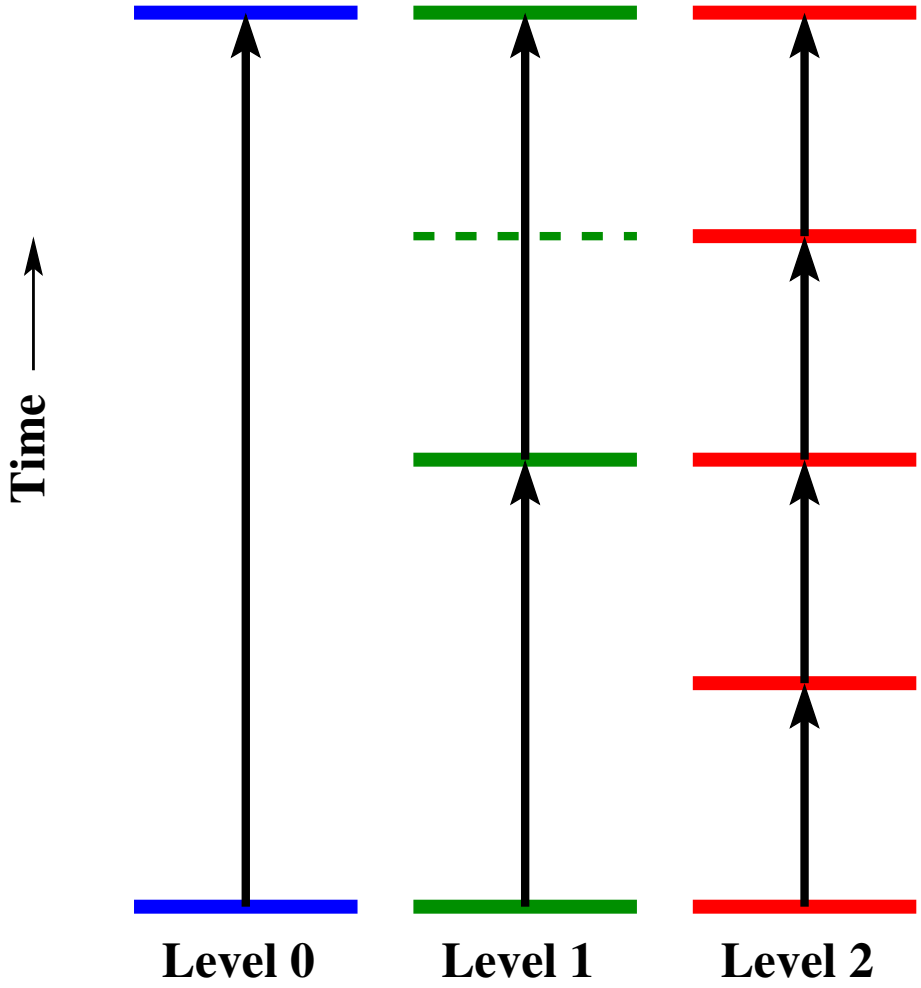
Time Advance



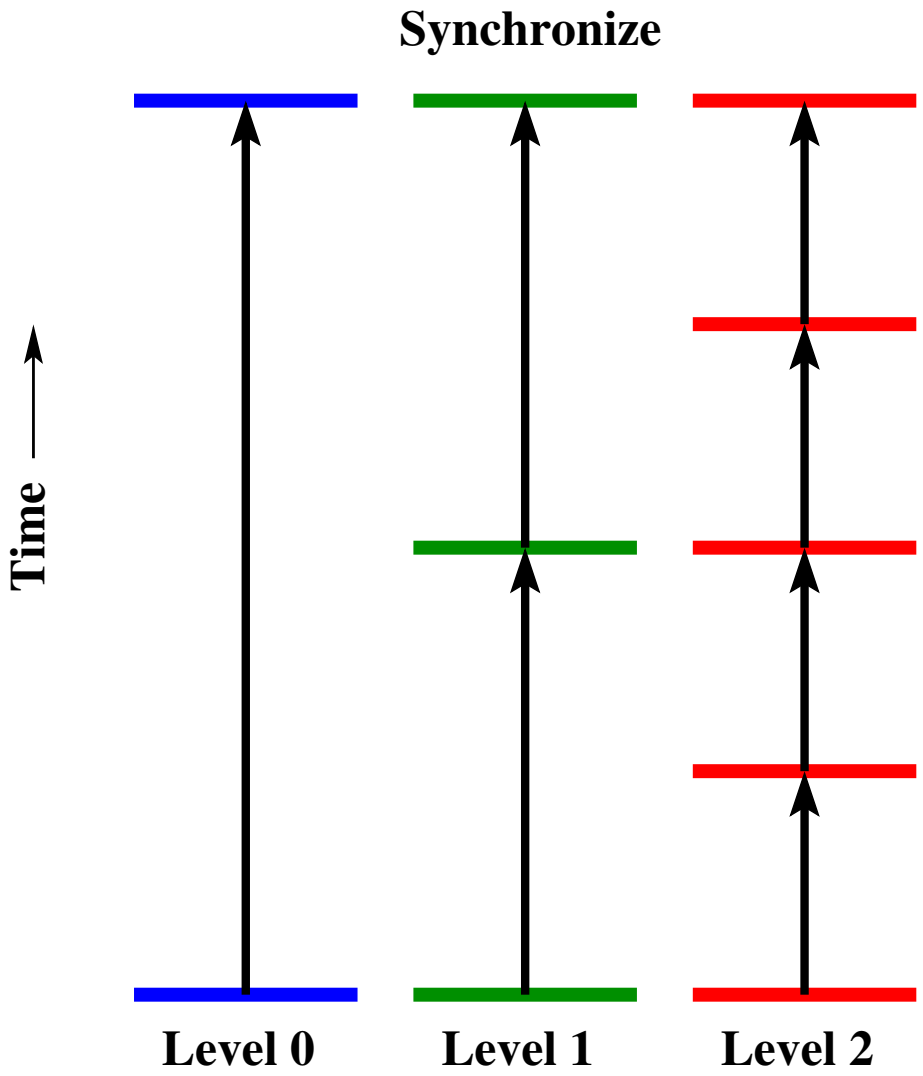
Time Advance



Time Advance



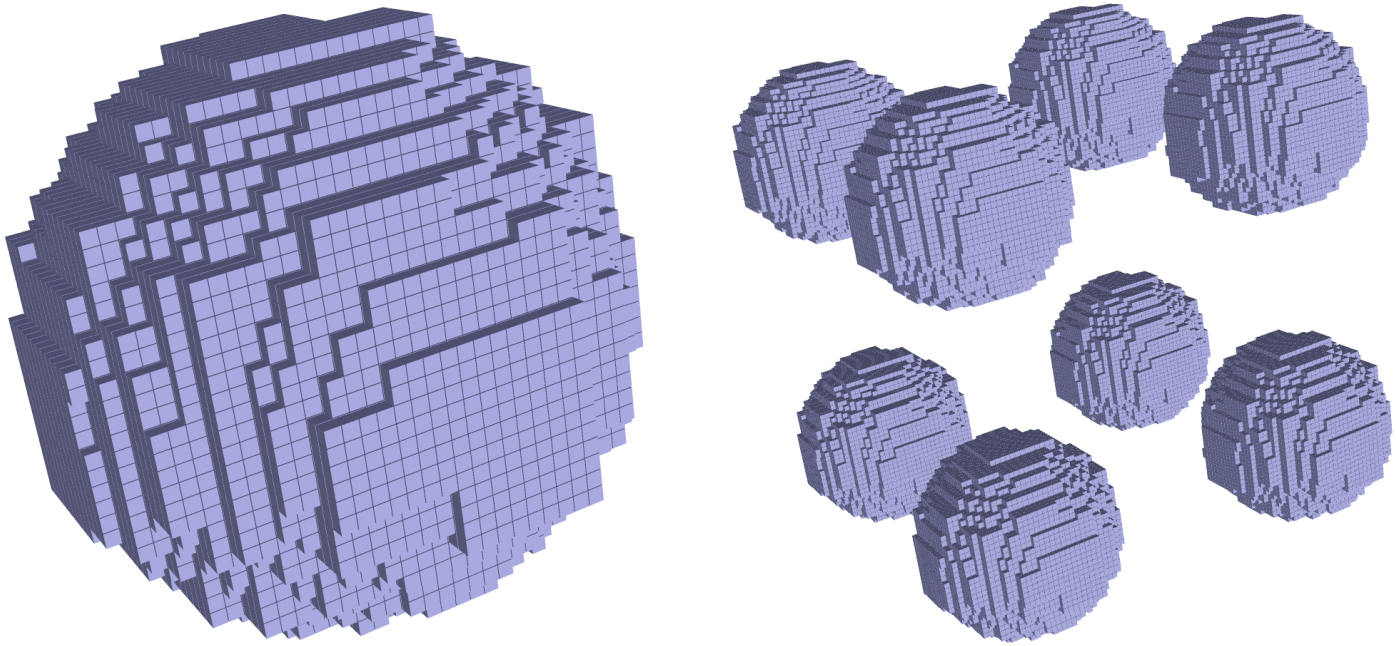
Time Advance



Benchmark

- **Chombo - C++/Fortran77 library with timers**
- **Hyperbolic PDE Solver - Gas Dynamics**
- **Spherical explosion in 3D**
- **Three levels of AMR, 4:1 refinement ratio**
- **Time steps: 1 coarse, 4 intermediate, 16 fine**
- **Static grids all 16^3**
- **6000 flops/grid-point/time-step**
- **124 million grid points on 128 processors**

Weak Scaling using Replication



Experimental Testbed

- Hardware - Cray XT4 (XT3)
  - Dual-core, 2.6 GHz AMD Opteron processors
  - DDR2-667 (DDR1-266) MHz memory - 7 (3.5) GB/s aggregate memory bandwidth per core
  - Cray SeaStar 2.1 ASIC interconnect, 6.4 GB/s bidirectional HyperTransport, 3D torus topology
- Operating Systems
  - Catamount: Specialized micro-kernel OS developed at Sandia
  - CNL: A lightweight kernel based on the Linux OS

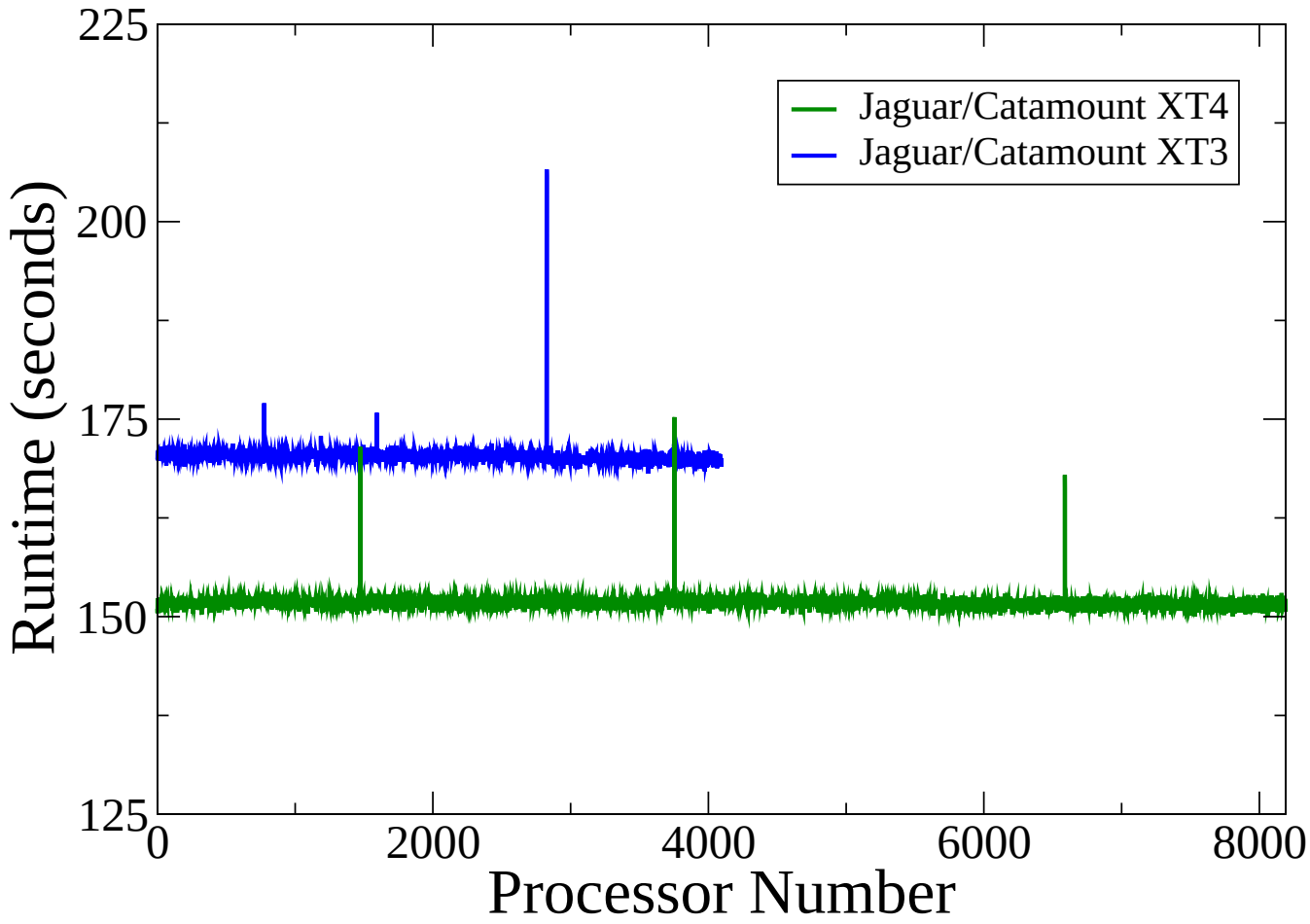
Optimizing for Scalability

- **Improving communication locality**
- **Using O(N) metadata management algorithms**
- **Optimizing coarse-fine boundary calculations**

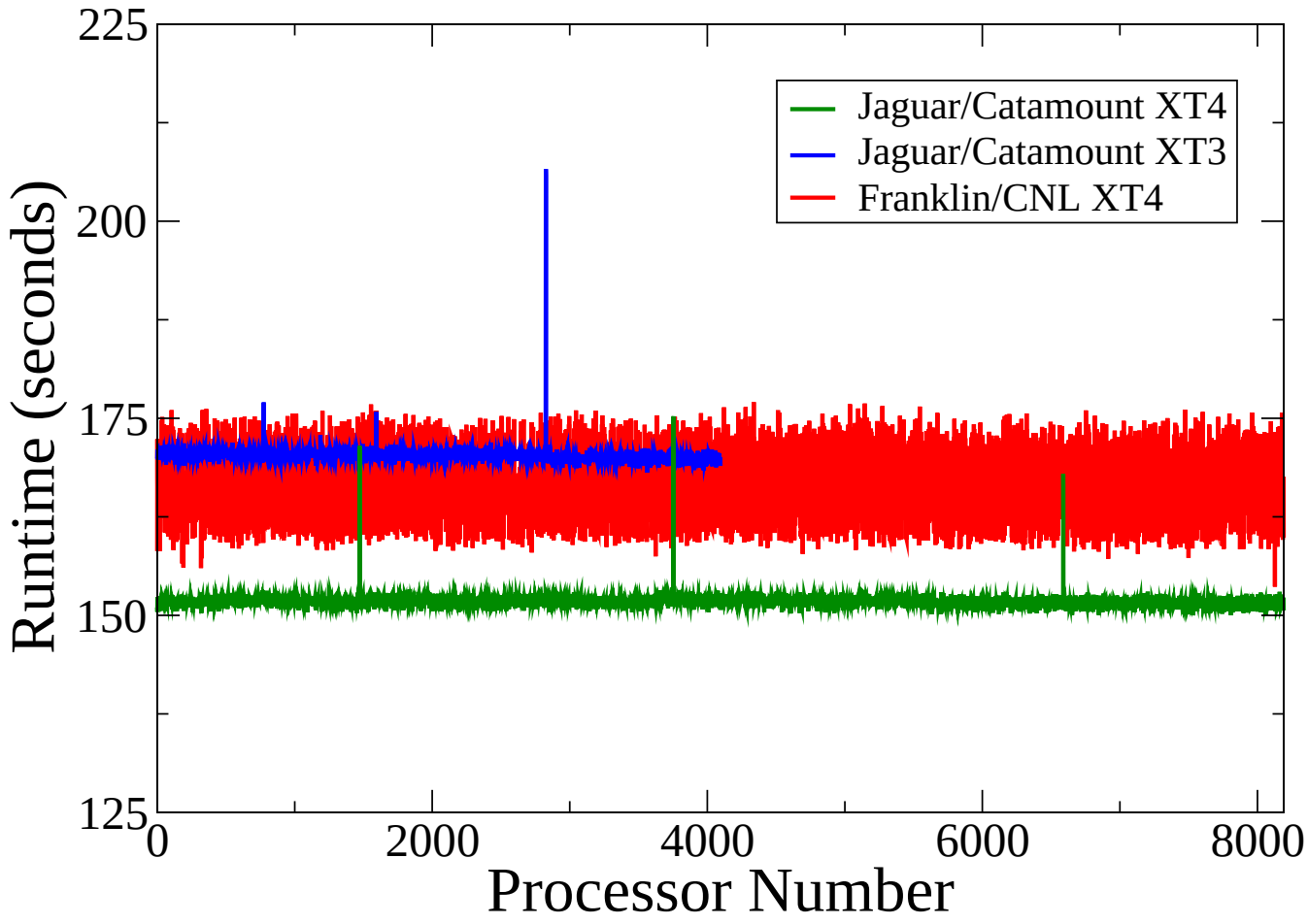
Mysterious Behavior - The Game is Afoot

- **Three phases of the core computation:**
  1. **Fill ghost cells from other grids**
  2. **Time advance the computation (no communication or I/O)**
  3. **Fill buffers used for time synchronization between AMR levels**
- **First mystery - Load imbalances on Jaguar running Catamount**
- **Detailed measurements pointed to (2) above**

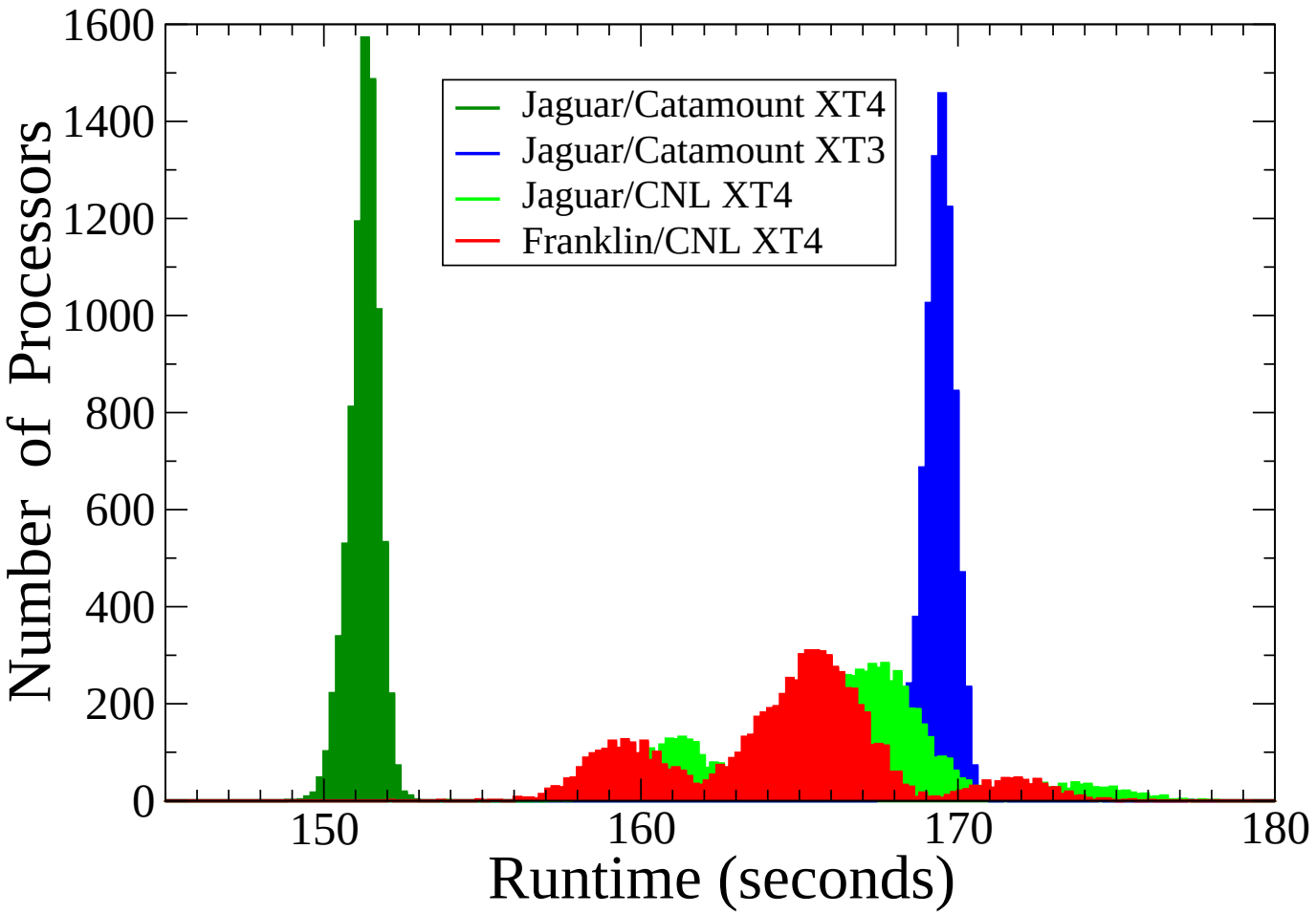
Mysterious Behavior - Jaguar running Catamount



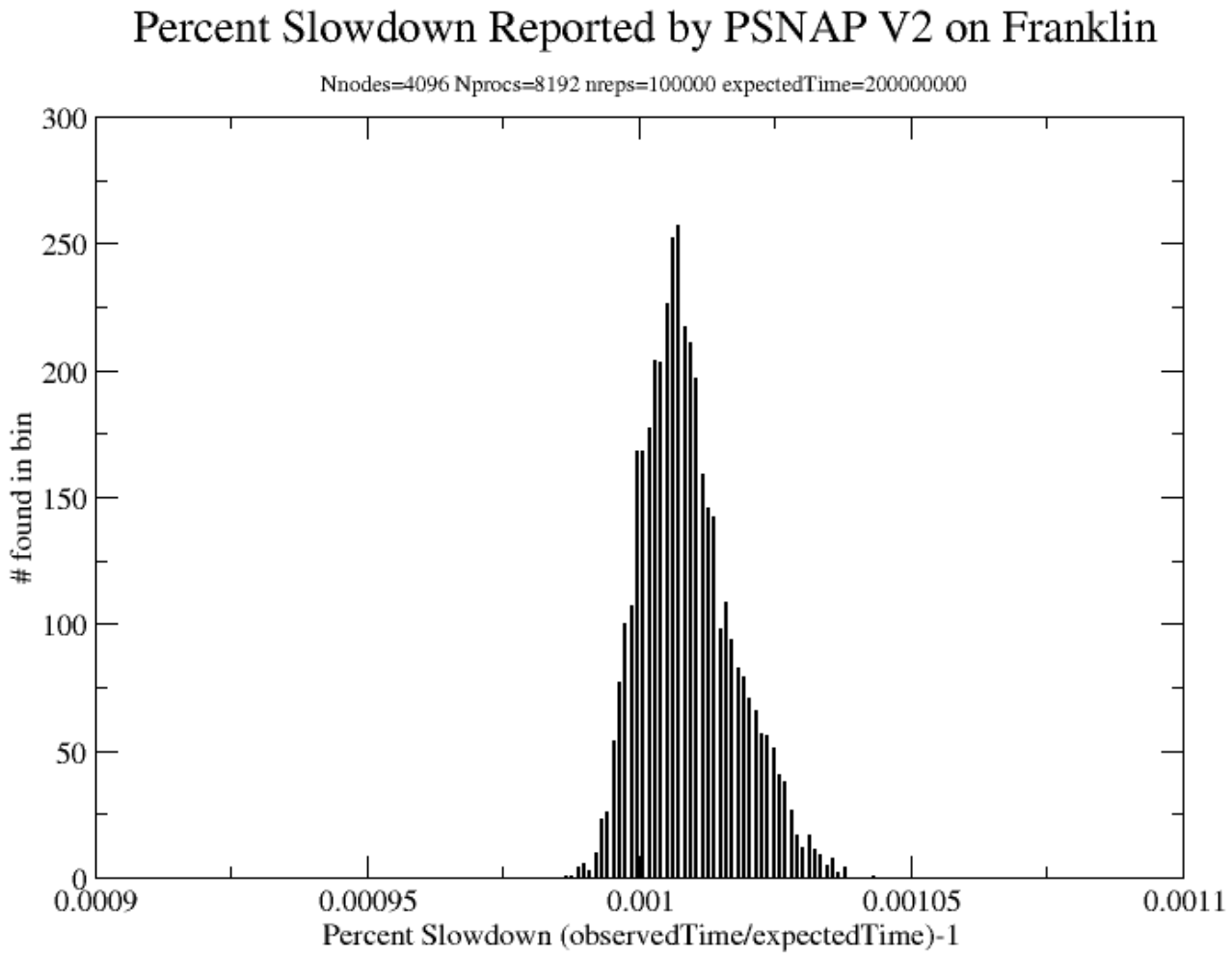
Mysterious Behavior - Jaguar running Catamount



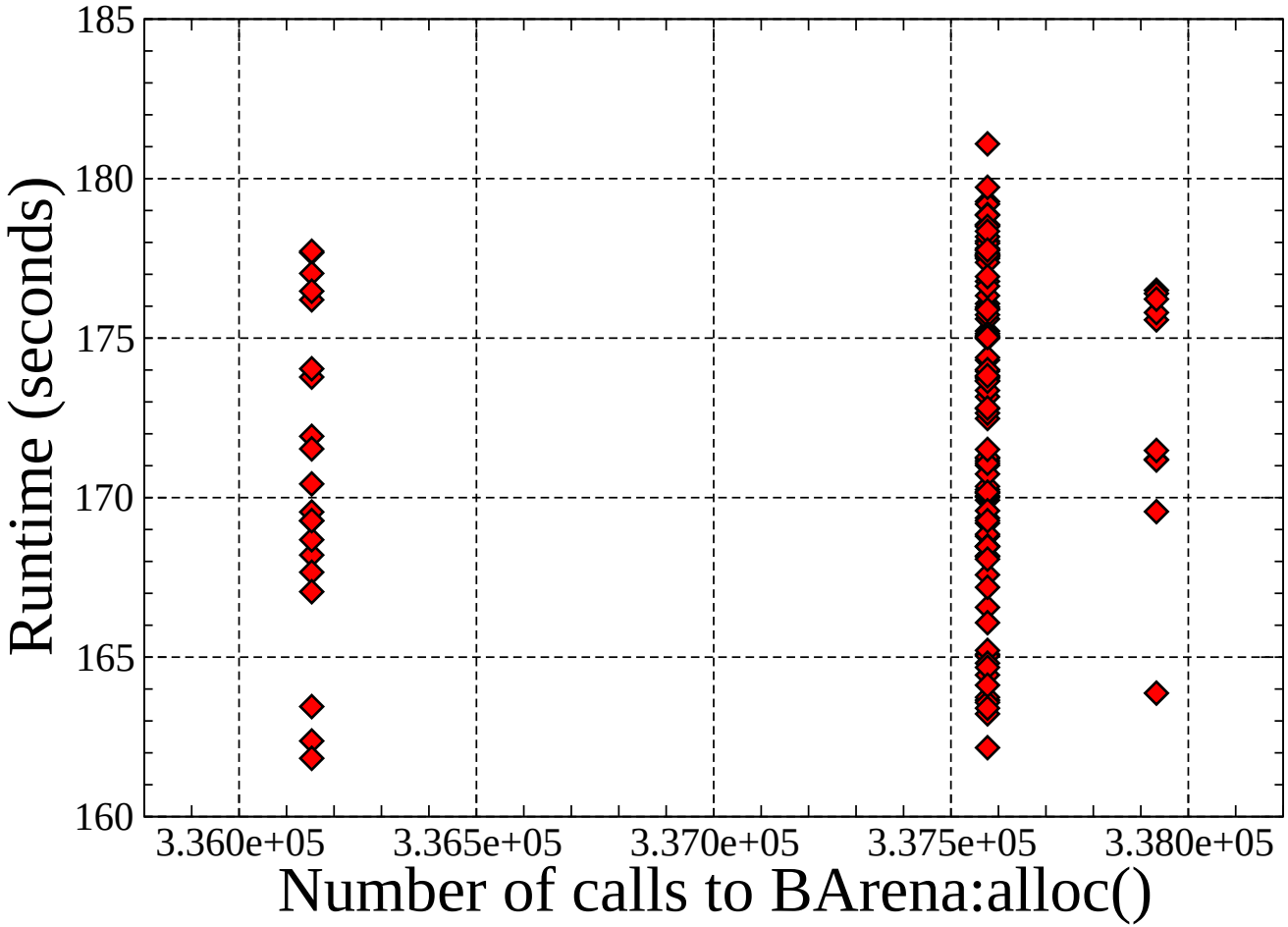
Mysterious Behavior - Franklin running CNL



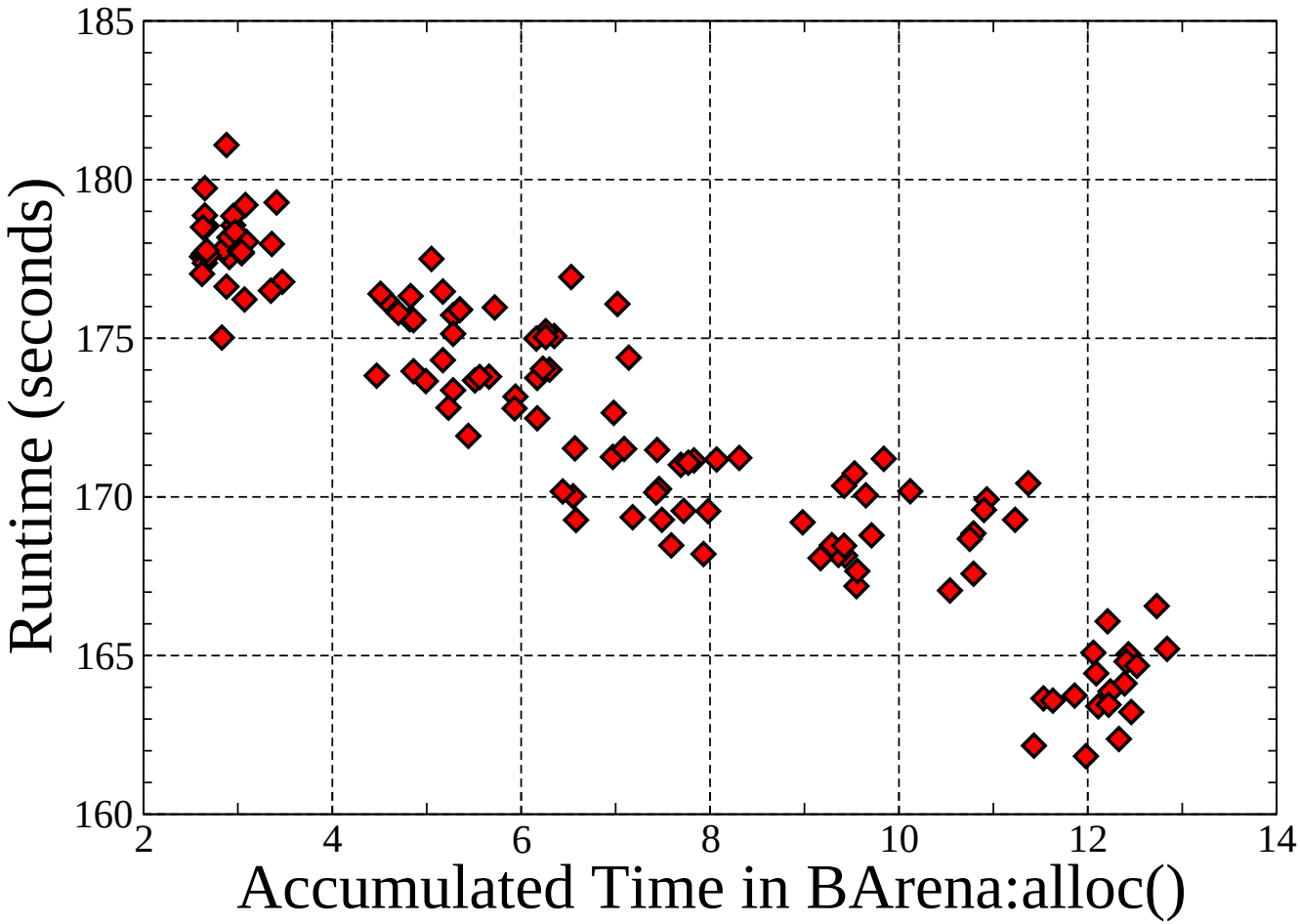
Mysterious Behavior - Looking for Clues



Mysterious Behavior - Looking for Clues



Mysterious Behavior - Looking for Clues

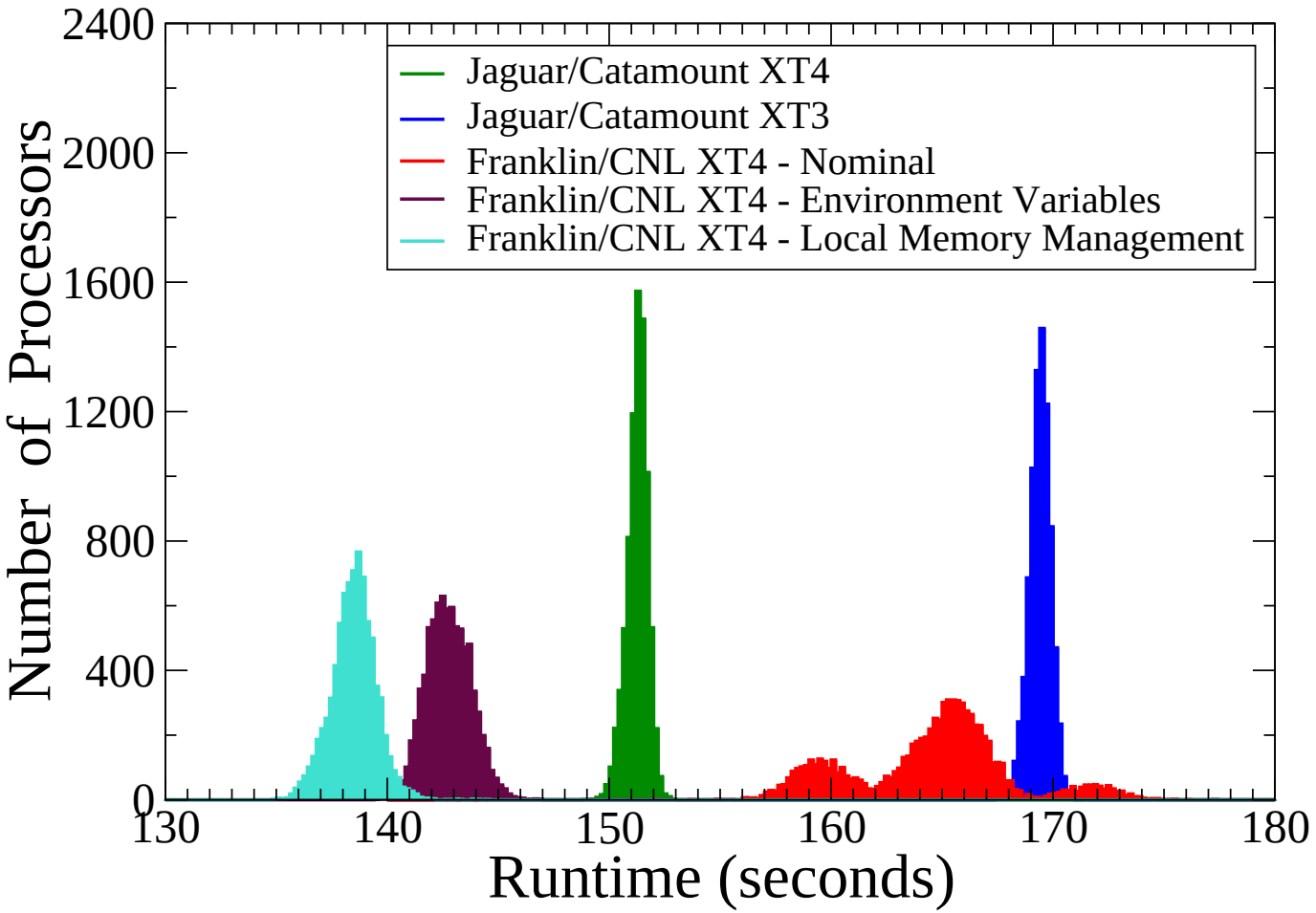


Mysterious Behavior - Hypothesis and Experiment

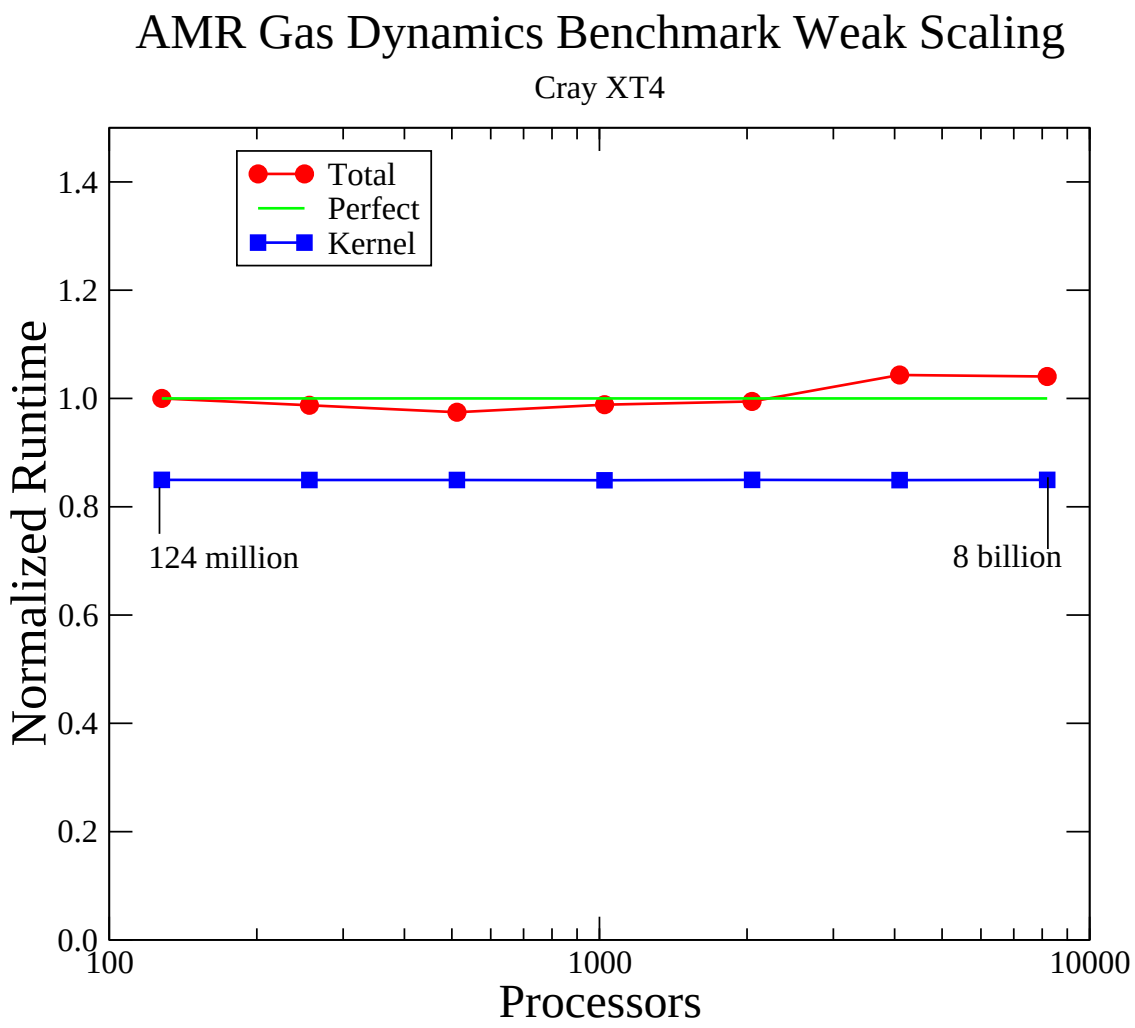
- **Hypothesis 1: Memory allocation heuristics**
- **Experiment 1: Change system memory allocation strategies**
- **Result 1: 3-14 seconds went to 25-26 seconds**

- **Hypothesis 2: Reduction of efficiency of data layout in the heap**
- **Experiment 2: Have Chombo manage its own heap**
- **Result 2: Variation decreased by a factor of 3 overall**

Mysterious Behavior - Solved!



Weak Scaling using Replication



Conclusions

- **Scaling a block structured AMR code for solving hyperbolic PDE to thousands of processors was straightforward overall**
- **The additional problems encountered are not specific to this application and will probably affect many codes with complex behavior**
- **There need to be better benchmarks and diagnostic tools for large HPC systems running complex codes**

Ongoing and Future Work

- **Scaling of elliptic and parabolic PDE solvers (done)**
- **Scaling of I/O of block structured AMR data (in process)**
- **Scaling of meta-data and meta-computations**
- **Scaling of other pieces including initialization, restart, regridding**

Acknowledgments and Thanks

- **Steve Luzmoor of Cray Inc.**
- **Patrick Worley and the PEAC INCITE grant**
- **Supported by the Office of Advanced Scientific Computing Research in the Department of Energy under Contract DE-AC02-05CH11231**