Accelerating Software for Extreme Scale Computing

Thomas Sterling
Professor of Informatics and Computing
Chief Scientist and Associate Director
Center for Research in Extreme Scale Technologies (CREST)
Pervasive Technology Institute at Indiana University
Fellow, Sandia National Laboratories CSRI

May 8, 2012
K ("KEI") Computer
Technology Demands new Response

Figure courtesy of Kunle Olukotun, Lance Hammond, Herb Sutter, and Burton Smith
Top 500 System Architectures – But there’s more!
Game Changer – Runtime System

• Runtime system
  – is: ephemeral, dedicated to and exists only with an application
  – is not: the OS, persistent and dedicated to the hardware system

• Moves us from *static* to *dynamic* operational regime
  – Exploits situational awareness for causality-driven adaptation
  – Guided-missile with continuous course correction rather than a fired projectile with fixed-trajectory

• Based on foundational assumption
  – More computational work will yield reduced time and lower power
  – Untapped system resources to be harvested
  – Opportunities for enhanced efficiencies discovered only in flight
  – New methods of control to deliver superior scalability
Efficiency & Scalability Challenges

• Starvation
  – Insufficiency of concurrency of work
  – Impacts scalability and latency hiding
  – Effects programmability

• Latency
  – Time measured distance for remote access and services
  – Impacts efficiency

• Overhead
  – Critical time additional work to manage tasks & resources
  – Impacts efficiency and granularity for scalability

• Waiting for contention resolution
  – Delays due to simultaneous access requests to shared physical or logical resources
Concepts

• Motivated by dynamic directed graphs
  – STEM
  – Knowledge management and understanding

• Split-phase transactions
  – Avoid blocking

• Message-driven computation
  – Move work to data
  – Parcels and Percolation

• Constraint-based synchronization
  – Declarative criteria for work
  – Event driven

• Data-directed execution
  – Merger of flow control and data structure

• Shared name space
ParalleX Execution Model
ParalleX

- Communicates by parcels (an advanced form of active messages) that moves both work and control state to the data as well as conventional asynchronous gathers of data to the work
ParalleX

• Supports Local Control Objects (LCOs) for lightweight synchronization to eliminate global barriers and manage asynchronous compound actions to reduce overhead and expose additional parallelism.

Note: Thread 1 is suspended only if the results from locality 2 are not readily available. If results are available Thread 1 continues to complete execution.
Constraint-based Synchronization

• Supports Dynamic-Adaptive Task Scheduling
• Declarative Semantics for Continuation of Execution
  – Defines conditions for work to be performed
  – Not imperative code by user
• Establishes Criteria for Task Instantiation
• Supports DAG flow control representation
• Examples:
  – Dataflow
  – Futures
HPX Runtime System

• Establishes a global address space that is active in the sense that a virtually addressed object may migrate across nodes without having to change its address
AdapEve	
  
  Mesh	
  
  Refinement

• Why Adaptive Mesh Refinement?
• The AMR dataflow in ParalleX
• The impact of granularity
• 3-D Test problem: comparisons between MPI based publicly available AMR toolkits and ParalleX AMR
• Impact of nedmalloc, concur, tcmalloc, and jmalloc
Constraint based Synchronization
Wallclock time ratio MPI/HPX

(Depending on levels of AMR refinement - LoR, pollux.cct.lsu.edu, 32 cores)
Application: Adaptive Mesh Refinement (AMR) for Astrophysics simulations

- ParalleX based AMR removes all global computation barriers, including the timestep barrier (so not all points have to reach the same timestep in order to proceed computing)
Finite Temperature Equation of State

• Putting in the right nuclear physics (polytropes don’t have the right compactness).
• To do anything with neutrinos, you need a temperature: neutrino cooling, dynamics of hypernova
• To do anything with radiation, you need a physical temperature
• This is a spring-board for new astrophysics and microphysics
Equations of State come in tables

• Not practical to do EOS calculation in place; it’s best to use a look-up table
• The table covers a lot of physics for you
• In high energy astrophysics:
  Temperatures vary from 0 to > 100 MeV
  Proton fraction changes from 0 to 0.6
  Density varies across 10 orders of magnitude
• Finer grid tables are better for accuracy
• Tables need to cover a wide variety of conditions: black hole formation, neutron star mergers, nucleosynthesis
MHDe Code Tests

- All C++
- PPM Reconstruction with HLLE Numerical Flux
- AMR
- Equations described in http://arxiv.org/abs/gr-qc/0605102
Preliminary Radial Pulsation Frequencies
Conclusions

• Supercomputing demands innovative response to technology challenges and application opportunities

• HPC is entering Phase VI; needs a new model of computation
  – Attack starvation, latency, overhead, & waiting for contention (SLOW)
  – Dynamic adaptive resource management & task scheduling
  – Dynamic graph-based applications for knowledge management (AI)

• Constraint-based synchronization represents an experimental step
  – Dynamic, overlap/multiphase message-driven execution

• Large scale runtime experiments required to guide progress
  – Application driven
  – Stimulate work in Architecture and Programming Models
  – ParalleX provides an experimental model with HPX implementation