Exascale Computing and the Role of Codesign

Sudip Dosanjh

Co-director ACES and IAA
Served on DOE's Exascale Initiative Steering Committee
Group Leader for Computer Science Research

Sandia National Laboratories
DOE MISSION & SCIENCE NEEDS
DOE mission imperatives require simulation and analysis for policy and decision making

- **Climate Change**: Understanding, mitigating and adapting to the effects of global warming
  - Sea level rise
  - Severe weather
  - Regional climate change
  - Geologic carbon sequestration

- **Energy**: Reducing U.S. reliance on foreign energy sources and reducing the carbon footprint of energy production
  - Reducing time and cost of reactor design and deployment
  - Improving the efficiency of combustion energy systems

- **National Nuclear Security**: Maintaining a safe, secure and reliable nuclear stockpile
  - Stockpile certification
  - Predictive scientific challenges
  - Real-time evaluation of urban nuclear detonation

Accomplishing these missions requires exascale resources.
Exascale simulation will enable fundamental advances in basic science

- High Energy & Nuclear Physics
  - Dark-energy and dark matter
  - Fundamentals of fission fusion reactions
- Facility and experimental design
  - Effective design of accelerators
  - Probes of dark energy and dark matter
  - ITER shot planning and device control
- Materials / Chemistry
  - Predictive multi-scale materials modeling: observation to control
  - Effective, commercial technologies in renewable energy, catalysts, batteries and combustion
- Life Sciences
  - Better biofuels
  - Sequence to structure to function

These breakthrough scientific discoveries and facilities require exascale applications and resources.
Exascale resources are required for predictive climate simulation

- Finer resolution
  - Provide regional details
- Higher realism, more complexity
  - Add “new” science
    - Biogeochemistry
    - Ice-sheets
  - Up-grade to “better” science
    - Better cloud processes
    - Dynamics land surface
- Scenario replication, ensembles
  - Range of model variability
- Time scale of simulation
  - Long-term implications

Adapted from Climate Model Development Breakout Background
Bill Collins and Dave Bader, Co-Chairs

Ocean chlorophyll from an eddy-resolving simulation with ocean ecosystems included

It is essential that computing power be increased substantially (by a factor of 1000), and scientific and technical capacity be increased (by at least a factor of 10) to produce weather and climate information of sufficient skill to facilitate regional adaptations to climate variability and change.

World Modeling Summit for Climate Prediction, May, 2008
Product development times must be accelerated to meet energy goals.

Conversion to CO₂ Neutral Infrastructure

Three Product Development Cycles

Full Market Transition

R&D
Simulation for product engineering will evolve from mean effects to predictive

**Current CFD tools**
- Reynolds-Averaged Navier-Stokes
- Calculate mean effects of turbulence
- Turbulent combustion submodels calibrated over narrow range
- DNS and LES for science calculations at standard pressures

**Future CFD tools**
- Improved math models for more accurate RANS simulations
- LES with detailed chemistry, complex geometry, high pressures, and multiphase transport as we achieve exascale computing
- DNS for submodel development
- Alternative fuel combustion models
National Nuclear Security

- U.S. Stockpile must remain safe, secure and reliable without nuclear testing
  - Annual certification
  - Directed Stockpile Work
  - Life Extension Programs
- A predictive simulation capability is essential to achieving this mission
  - Integrated design capability
  - Resolution of remaining unknowns
    - Energy balance
    - Boost
    - Si radiation damage
    - Secondary performance
  - Uncertainty Quantification
  - Experimental campaigns provide critical data for V&V (NIF, DARHT, MaRIE)
- Effective exascale resources are necessary for prediction and quantification of uncertainty
TECHNOLOGY NEEDS
Concurrency is one key ingredient in getting to exaflop/sec

and power, resiliency, programming models, memory bandwidth, I/O, …

Increased parallelism allowed a 1000-fold increase in performance while the clock speed increased by a factor of 40.
Many-core chip architectures are the future

The shift toward increasing parallelism is not a triumphant stride forward based on breakthroughs in novel software and architectures for parallelism … instead it is actually a retreat from even greater challenges that thwart efficient silicon implementation of traditional uniprocessor architectures.

*Kurt Keutzer*
What are critical exascale technology investments?

• **System power** is a first class constraint on exascale system performance and effectiveness.
• **Memory** is an important component of meeting exascale power and applications goals.
• **Programming model.** Early investment in several efforts to decide in 2013 on exascale programming model, allowing exemplar applications effective access to 2015 system for both mission and science.
• **Investment in exascale processor design** to achieve an exascale-like system in 2015.
• **Operating System strategy for exascale** is critical for node performance at scale and for efficient support of new programming models and run time systems.
• **Reliability and resiliency are critical at this** scale and require applications neutral movement of the file system (for check pointing, in particular) closer to the running apps.
• **HPC co-design strategy and implementation** requires a set of a hierarchical performance models and simulators as well as commitment from apps, software and architecture communities.
## Potential System Architecture Targets

<table>
<thead>
<tr>
<th>System attributes</th>
<th>2010</th>
<th>“2015”</th>
<th>“2018”</th>
</tr>
</thead>
<tbody>
<tr>
<td>System peak</td>
<td>2 Peta</td>
<td>200 Petaflop/sec</td>
<td>1 Exaflop/sec</td>
</tr>
<tr>
<td>Power</td>
<td>6 MW</td>
<td>15 MW</td>
<td>20 MW</td>
</tr>
<tr>
<td>System memory</td>
<td>0.3 PB</td>
<td>5 PB</td>
<td>32-64 PB</td>
</tr>
<tr>
<td>Node performance</td>
<td>125 GF</td>
<td>0.5 TF</td>
<td>7 TF</td>
</tr>
<tr>
<td>Node memory BW</td>
<td>25 GB/s</td>
<td>0.1 TB/sec</td>
<td>1 TB/sec</td>
</tr>
<tr>
<td>Node concurrency</td>
<td>12</td>
<td>O(100)</td>
<td>O(1,000)</td>
</tr>
<tr>
<td>System size (nodes)</td>
<td>18,700</td>
<td>50,000</td>
<td>5,000</td>
</tr>
<tr>
<td>Total Node Interconnect BW</td>
<td>1.5 GB/s</td>
<td>20 GB/sec</td>
<td>200 GB/sec</td>
</tr>
<tr>
<td>MTTI</td>
<td>days</td>
<td>O(1 day)</td>
<td>O(1 day)</td>
</tr>
</tbody>
</table>
The high level system design may be similar to petascale systems

- New interconnect topologies
- Optical interconnect

- 10x – 100x more nodes
- MPI scaling & fault tolerance
- Different types of nodes
- NVRAM on nodes

- Mass storage far removed from application data
“The Energy and Power Challenge is the most pervasive ... and has its roots in the inability of the [study] group to project any combination of currently mature technologies that will deliver sufficiently powerful systems in any class at the desired levels.”

DARPA IPTO exascale technology challenge report
Memory bandwidth and memory sizes will be >> less effective without R&D

• Primary needs are
  ▪ Increase in bandwidth (concurrency can be used to mask latency, viz. Little’s Law)
  ▪ Lower power consumption
  ▪ Lower cost (to enable affordable capacity)
• Stacking on die enable improved bandwidth and lower power consumption
• Modest improvements in latency
• Commodity memory interface standards are not pushing bandwidth enough
Investments in memory technology mitigate risk of narrowed application scope.
Cost of Memory Capacity for two different potential memory Densities

- Memory density is doubling every three years; processor logic, every two
  - Project 8Gigabit DIMMs in 2018
  - 16Gigabit if technology acceleration

- Storage costs are dropping gradually compared to logic costs
  - Industry assumption is $1.80/memory chip is median commodity cost

Cost in $M (8 gigabit modules)
Cost in $M (16 Gigabit modules)
1/2 of $200M system

Petabytes of Memory

$500,000
$400,000
$300,000
$200,000
$100,000
$0,000

16
32
64
128
256

Sandia National Laboratories
Need solutions for decreased reliability and a new model for resiliency

• **Barriers**
  • System components, complexity increasing
  • Silent error rates increasing
  • Reduced job progress due to fault recovery if we use existing checkpoint/restart

• **Technical Focus Areas**
  • Local recovery and migration
  • Development of a standard fault model and better understanding of types/rates of faults
  • Improved hardware and software reliability
    • Greater integration across entire stack
  • Fault resilient algorithms and applications

• **Technical Gap**
  • Maintaining today’s MTTI given 10x - 100X increase in sockets will require:
    10X improvement in hardware reliability
    10X in system software reliability, and
    10X improvement due to local recovery and migration as well as research in fault resilient applications

  ![Checkpoint Restart to Node Local Storage]

  **Taxonomy of errors (h/w or s/w)**
  • **Hard errors**: permanent errors which cause system to hang or crash
  • **Soft errors**: transient errors, either correctable or short term failure
  • **Silent errors**: undetected errors either permanent or transient. *Concern is that simulation data or calculation have been corrupted and no error reported.*

  ![Need storage solution to fill this gap]
Programming models and environments require early investment.

- **Barriers:** Delivering a large-scale scientific instrument that is productive and fast.
  - $O(1B)$ way parallelism in Exascale system
  - $O(1K)$ way parallelism in a processor chip
    - Massive lightweight cores for low power
    - Some “full-feature” cores lead to heterogeneity
  - Data movement costs power and time
    - Software-managed memory (local store)
  - Programming for resilience
  - Science goals require complex codes

- **Technology Investments**
  - Extend inter-node models for scalability and resilience, e.g., MPI, PGAS (includes HPCS)
  - Develop intra-node models for concurrency, hierarchy, and heterogeneity by adapting current scientific ones (e.g., OpenMP) or leveraging from other domains (e.g., CUDA, OpenCL)
  - Develop common low level runtime for portability and to enable higher level models

- **Technical Gap:**
  - No portable model for variety of on-chip parallelism methods or new memory hierarchies
  - Goal: Hundreds of applications on the Exascale architecture; Tens running at scale

---

CO-DESIGN
Co-design expands the feasible solution space to allow better solutions.

Application driven:
Find the best technology to run this code. 
Sub-optimal

Technology driven:
Fit your application to this technology. 
Sub-optimal.

Now, we must expand the co-design space to find better solutions:
• new applications & algorithms,
• better technology and performance.

Application
Model
Algorithms
Code
Technology
⊕ architecture
⊕ programming model
⊕ resilience
⊕ power
Hardware/Software co-design is a mature field in embedded computing

- Design of an integrated system that contains hardware and software
- Focus on embedded systems (cell phones, appliances, engines, controllers, etc.)
- Concurrent development of hardware and software
  - Interactions and tradeoffs
  - Partitioning is a focus
  - Must satisfy real-time and/or other performance/energy metrics/constraints
Original DOD Standard for HW/SW co-development had shortcomings
Lockheed Martin Co-design Methodology
Why has co-design not been used more extensively in HPC?

- Leveraging of COTs technology
  - Almost all leadership systems have some custom components but HPC has benefited from the ability to leverage commercial technology
- HPC applications are very complex
  - May contain a million of lines of code
- ~15-20 years of architectural and programming model stability
  - Bulk synchronous processing + explicit message passing
- Lack of Adequate Simulation Tools
  - Often use Byte to Flop ratios and Excel spreadsheets
  - Industry simulation tools are proprietary

However, there are some HPC co-design examples and there are useful tools.
Basic performance modeling

CTH is DoD’s most used code

Basic CTH Model

\[ T = E(\kappa, \phi)N^3 + C(\lambda + \tau kN^2) + S(\gamma \log(P)) + \text{Limbal} \]

- \( T \) is the execution time per time step
- \( N \) is size of an edge of a processor’s subdomain
- \( C \) and \( S \) are number of exchanges and collectives
- \( P \) is the number of processors
- \( k \) is the number of variables in an exchange
- \( \lambda \) and \( \tau \) are latency and transfer cost
- \( \gamma \) is the cost of one stage of collective
- \( E(\kappa, \phi) \) is the calculation time per cell
- \( \text{Limbal} \) is a new term representing effects of load imbalance

Limitations:
- Very simple architectural model
- Tuning parameters
- Need a new model when you change the application
SST Simulation Project

- Parallel
- Parallel Discrete Event core with conservative optimization over MPI
- Holistic
- Integrated Tech. Models for power
- McPAT, Sim-Panalyzer
- Multiscale
- Detailed and simple models for processor, network, and memory

- Current Release (2.0) at http://www.cs.sandia.gov/sst/
- Includes parallel simulation core, configuration, power models, basic network and processor models, and interface to detailed memory model
SST simulations have quantified the impact of the Memory Wall

- Most of DOE’s Applications (e.g., climate, fusion, shock physics, …) spend most of their instructions accessing memory or doing integer computations, not floating point
- Additionally, most integer computations are computing memory Addresses
- Advanced development efforts are focused on accelerating memory subsystem performance for both scientific and informatics applications
SST is providing architectural insights to algorithms developers

- Input: SST Trace for SpMV.
- Lots of instruction stream data.
- Model: Use restricted $sin^2$ function to mark start/finish of each instruction.
- Use FFTs to analyze behavior.

Trace fragment from SpMV inner loop

<table>
<thead>
<tr>
<th>$j$</th>
<th>$I_j$</th>
<th>issue</th>
<th>complete</th>
<th>$\kappa$</th>
</tr>
</thead>
<tbody>
<tr>
<td>59</td>
<td>bc</td>
<td>737</td>
<td>741</td>
<td>4</td>
</tr>
<tr>
<td>60</td>
<td>lwz</td>
<td>738</td>
<td>744</td>
<td>6</td>
</tr>
<tr>
<td>61</td>
<td>lfd</td>
<td>740</td>
<td>746</td>
<td>6</td>
</tr>
<tr>
<td>62</td>
<td>addi</td>
<td>742</td>
<td>746</td>
<td>4</td>
</tr>
<tr>
<td>63</td>
<td>addi</td>
<td>742</td>
<td>746</td>
<td>4</td>
</tr>
<tr>
<td>64</td>
<td>rlwinm</td>
<td>743</td>
<td>746</td>
<td>3</td>
</tr>
<tr>
<td>65</td>
<td>lfdx</td>
<td>744</td>
<td>850</td>
<td>106</td>
</tr>
<tr>
<td>66</td>
<td>fnadd</td>
<td>849</td>
<td>854</td>
<td>5</td>
</tr>
<tr>
<td>67</td>
<td>bc</td>
<td>850</td>
<td>854</td>
<td>4</td>
</tr>
<tr>
<td>68</td>
<td>lwz</td>
<td>851</td>
<td>857</td>
<td>6</td>
</tr>
<tr>
<td>69</td>
<td>lfd</td>
<td>853</td>
<td>859</td>
<td>6</td>
</tr>
<tr>
<td>70</td>
<td>addi</td>
<td>855</td>
<td>859</td>
<td>4</td>
</tr>
<tr>
<td>71</td>
<td>addi</td>
<td>855</td>
<td>859</td>
<td>4</td>
</tr>
<tr>
<td>72</td>
<td>rlwinm</td>
<td>856</td>
<td>859</td>
<td>3</td>
</tr>
<tr>
<td>73</td>
<td>lfdx</td>
<td>857</td>
<td>886</td>
<td>29</td>
</tr>
<tr>
<td>74</td>
<td>fnadd</td>
<td>885</td>
<td>890</td>
<td>5</td>
</tr>
<tr>
<td>75</td>
<td>bc</td>
<td>886</td>
<td>890</td>
<td>4</td>
</tr>
<tr>
<td>76</td>
<td>lwz</td>
<td>887</td>
<td>893</td>
<td>6</td>
</tr>
<tr>
<td>77</td>
<td>lfd</td>
<td>889</td>
<td>895</td>
<td>6</td>
</tr>
<tr>
<td>78</td>
<td>addi</td>
<td>891</td>
<td>895</td>
<td>4</td>
</tr>
<tr>
<td>79</td>
<td>addi</td>
<td>891</td>
<td>895</td>
<td>4</td>
</tr>
<tr>
<td>80</td>
<td>rlwinm</td>
<td>892</td>
<td>895</td>
<td>3</td>
</tr>
<tr>
<td>81</td>
<td>lfdx</td>
<td>893</td>
<td>899</td>
<td>6</td>
</tr>
<tr>
<td>82</td>
<td>fnadd</td>
<td>898</td>
<td>903</td>
<td>5</td>
</tr>
<tr>
<td>83</td>
<td>bc</td>
<td>899</td>
<td>903</td>
<td>4</td>
</tr>
</tbody>
</table>
Need to define HPC co-design methodology

• Could range from discussions between architecture, software and application groups to tight collaboration centered on the co-simulation of hardware and applications

• Opportunity to influence future architectures
  ▪ Cores/node, threads/core, scheduling width/thread
  ▪ Logic in memory subsystem
  ▪ Interconnect performance

• HPC community must work together to define the next programming model