A SYSTEMS ARCHITECTURE FOR IMBEDDING PREDICTIVE CLIMATE SIMULATION IN THE DIGITAL ECONOMY

Mark Seager
Intel Fellow, CTO for the HPC Ecosystem
RackScale Design Group

Presented to iCAS14, Annecy France
September 11, 2017
Scientific simulation at scale is qualitatively different. We do not yet understand the full implications of this technological development.

By being on the leading edge at the 10 teraFLOP/s level with a toehold into the 100 teraFLOP/s → petaFLOP/s computing era, we are fundamentally changing the nature and scope of the scientific method.

♦ Edsger Dijkstra: “A quantitative difference is also a qualitative difference, if the quantitative difference is greater than an order of magnitude.

♦ A quantitative example in transportation
  ◆ 1 Mi/Hr is the speed of a baby crawling
  ◆ 10 Mi/Hr is the speed of a top marathon runner
  ◆ 100 Mi/Hr is the speed of a fast automobile
  ◆ 1,000 Mi/Hr is the speed of a fast jet

♦ Qualitative ramifications of this transportation example
  ◆ Driving allows people to go to places they could not reach on foot.
  ◆ Flying allows people to go to places they could not reach in time.
Simulation has become the critical integrating element between theory and experiment

Predictive simulation ENABLES

- Detailed predictive assessment of complex models for overarching physical problems
- Design of experiments
- Impact assessment of policy choices
- Elimination of costly physical prototypes

Predictive simulation REQUIRES

- Verification and validation of complex models (experiment)
- Development of science based models (theory)
- Databases of physical properties and catalogues of scientific data
- Petascale simulation environments

Revolution in the making: BlueGene/L at LLNL
The telescope was one of the central instruments of what has been called the Scientific Revolution of the seventeenth century. It revealed hitherto unsuspected phenomena in the heavens and had a profound influence on the controversy between followers of the traditional geocentric astronomy and cosmology and those who favored the heliocentric system of Copernicus. It was the first extension of one of man's senses, and demonstrated that ordinary observers could see things that the great Aristotle had not dreamed of. It therefore helped shift authority in the observation of nature from men to instruments. In short, it was the prototype of modern scientific instruments.
Industrial Revolution: A period of unprecedented technological and economic development.
innovation ACROSS ALL INDUSTRIES

“DIGITAL FUSION”
Blending of Traditional and Digital Business Models

TRADITIONAL BUSINESS

DIGITAL BUSINESS

Smart Agriculture
3D Printing in Healthcare
Wearables (Industrial/Lifestyle)
Robotic Surgery
Autonomous Cars
Human Augmentation
A FLOOD OF DATA IS COMING

<table>
<thead>
<tr>
<th>By 2020...</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AVG Internet User</td>
<td>~1.5 GB</td>
<td>* TRAFFIC PER DAY</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smart Hospital</td>
<td>3,000 GB</td>
<td>* PER DAY</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Autonomous Automobile</td>
<td>4,000 GB</td>
<td>* PER CAR... EACH</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Airplane Data</td>
<td>40,000 GB*</td>
<td>* PER DAY</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Smart Factory</td>
<td>1,000,000 GB*</td>
<td>* PER DAY</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Source: Intel Investor Day Feb 9, 2017**

* Sources:

Data Center traffic is DOUBLING every 18 months**
INTEL GIVING SIGHT TO MACHINES IN THE AI AGE
THROUGH INNOVATION & ACQUISITION

- Machine & Deep Learning
- Cognitive Computing
- Programmable Solutions
- Intel RealSense Depth Sensing
- Movidius Computer Vision & Deep Learning
- Computer Vision, Tools & Standards
- Collision Avoidance & Navigation
Other emerging predictive simulation usage models are driven by coupled Big Data, Computational Science and AI.

**Government & Research**
- The Human Brain Project
- Budget: Euro 1 Billion
- "My goal is simple. It is complete understanding of the universe, why it is as it is, and why it exists at all"
- Stephen Hawking

**Commerce & Industry**
- Better Products
- Digital Twin
- Faster Time to Market
- Reduced R&D
- New Business Models
- Data Services

**New Users & New Uses**
- Genomics
- Clinical Information
- From diagnosis to personalized treatments quickly

**Basic Science**
- **Business Transformation**
- **Data-Driven Discovery**

**Transform data into useful knowledge**

Source: IDC Worldwide Technical Computing Server 2013-2017 Forecast; Other brands, names, and images are the property of their respective owners.

Other brands, names and images may be claimed as the property of others.
Climate models are complex in a number of dimensions
Climate Models have to accurately simulate multiple-physical phenomena at wide range of scales

Multi-physics, Multi-Scale

Climate Application is a large set of tightly coupled Science Packages
Climate Model Characteristics

Very large, complex codes and workflows
- >1m LOC common
- Runtime measured in months for “Simulated Years Per Day”
- Multiple components coupled in parallel
  - Amdahl's Law & load balancing critical

Data movement critical limiting factor
- Memory Bandwidth limited
- High speed, Low Latency MPI communications
- Moderate IO, but critical for analysis and visualization
Knowledge Gaps in Physics-based models

Inadequate understanding of key climatic processes

- Cloud physics, precipitation extremes, etc.

Parameterized approximations for handling knowledge gaps lead to overly complex models

- Poor predictive performance
- High uncertainty
- Difficult to explain/interpret
Where we are now.

Where next gen models need to be in < 10 years time.
Why Do Black-box ML Methods Fail?

Scientific problems are often under-constrained

- Small number of samples, large number of variables
- High-quality climate observations only available for the recent past (40 to 100 years)
- Standard methods for assessing and ensuring generalizability of machine learning models break down
- Huge number of samples is critical to success of methods such as deep learning

Advances in data science methods are needed to accelerate scientific discovery

---

Big Data Needs a Big Theory to Go with It

- Geoffrey West 2013
Pilot 1: Predictive Models for Pre-Clinical Screening

Machine Learning Based Predictive Models

- Transcriptional Phenotyping
- Functional genomics
- Feature Engineering, Cross Validation, Scalable Compute on CORAL

Uncertainty and Optimal Experiment Design

- UQ Analysis, Model Selection, Model Improvement, Proposed Experiments

Hypotheses Formation and Mixed Modeling

- Feature Importance Mining
- Biological Interaction Network Modeling
- Integration of Mechanistic, Statistical and Inferential Modeling
HOW DEEP LEARNING CAN PLAY A ROLE
HYBRID MODELS ARE NEEDED IN CANCER RESEARCH
Applications of Machine Learning in Climate

1. Evaluation and refinement of models
   - Evaluate models based on their ability to capture key climatic processes (e.g., teleconnections) that are extracted using ML algorithms
   - Significantly speed up model refinement cycle by providing quick diagnostics

2. Design of hybrid-physics-data models
   - Replace/enhance individual components of physics-based models to address knowledge gaps and improve predictive performance
Intel® Scalable System Framework
One Framework for Multiple Complex Workflows

Small Clusters Through Peta and Exascale Compute and Data-Centric Computing
Standards-Based Programmability
IA and HPC Ecosystem Enabling
On-Premise and Cloud-Based
SSF: Enabling Configurability & Scalability
from components to racks to clusters

SSF Path To Exascale

SSF for Scalable Clusters

- Xeon or Xeon-Phi – based on workload needs
- Compute flexibly aggregated
- Lowest latency compute to compute interconnect

- I/O Topologies for best performance
- Configurable I/O bandwidth director switch
- Burst buffer to decouple storage from I/O
CPU-Fabric Integration with the Intel® Omni-Path Architecture

KEY VALUE VECTORS

✓ Performance
✓ Density
✓ Cost
✓ Power
✓ Reliability

Tighter Integration

Multi-chip Package Integration

Intel® OPA HFI Card

Intel® OPA

Future Generations
Additional integration, improvements, and features

Future Intel® Xeon® processor (14nm)

Intel® Xeon Phi™ processor

Next generation Intel® Xeon® Phi™ coprocessor

Future Intel® Xeon® processor

Next generation Intel® Xeon® processor

Intel® Xeon® processor E5-2600 v3

Intel® Xeon® processor
**Tighter System-Level Integration**

**Innovative Memory-Storage Hierarchy**

<table>
<thead>
<tr>
<th>Today</th>
<th>Future</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Processor</strong></td>
<td><strong>Compute</strong></td>
</tr>
<tr>
<td>Caches</td>
<td>Caches</td>
</tr>
<tr>
<td>Local memory is now faster &amp; in processor package</td>
<td></td>
</tr>
<tr>
<td><strong>Memory Bus</strong></td>
<td></td>
</tr>
<tr>
<td>Local Memory</td>
<td>Intel® DIMMs based on 3D XPoint™ Technology</td>
</tr>
<tr>
<td>Much larger memory capacities keep data in local memory</td>
<td></td>
</tr>
<tr>
<td><strong>Compute Node</strong></td>
<td></td>
</tr>
<tr>
<td>Local Storage</td>
<td>Intel® Optane™ Technology SSDs</td>
</tr>
<tr>
<td>I/O Node storage moves to compute node</td>
<td></td>
</tr>
<tr>
<td><strong>I/O Node</strong></td>
<td></td>
</tr>
<tr>
<td>SSD Storage</td>
<td>Burst Buffer Node with Intel® Optane™ Technology SSDs</td>
</tr>
<tr>
<td>Some remote data moves onto I/O node</td>
<td></td>
</tr>
<tr>
<td><strong>Remote Storage</strong></td>
<td></td>
</tr>
<tr>
<td>Parallel File System (Hard Drive Storage)</td>
<td>Parallel File System (Hard Drive Storage)</td>
</tr>
</tbody>
</table>

*cache, memory or hybrid mode*
NAND Flash and 3DXPoint Enable New solutions for Data Intensive IO bottleneck

3D MLC and TLC NAND

- Disrupts HDD and block IO software
- Improves BW 10x and IOPs 1,000x, NVMe IO interface for Climate Warm Data Tier

3D XPoint™ Technology

- New memory class storage devices
- Improves BW 1,000x, 100,000x IOPs, endurance 100x and memory interface for Climate Warm Hot Tier
Intel® Software Solutions

Intel® Software Defined Visualization

**Low Cost**
No Dedicated Viz Cluster

**Excellent Performance**
Less Data Movement, I/O
Invest Power, Space, Budget in Greater Compute Capability

**High Fidelity**
Work with Larger Data Sets – Not Constrained by GPU Memory

HPC System Software Stack

**An Open Community Effort**
Broad Range of Ecosystem Partners
Open Source Availability

**Benefits the Entire HPC Ecosystem**
Accelerate Application Development
Turnkey to Customizable

Intel® Parallel Studio

**Faster Code**
Boost Application Performance on Current and Next-Gen CPUs

**Create Code Faster**
Utilizing a Toolset that Simplifies Creating Fast and Reliable Parallel Code

Open Software Available Today!
Bringing Your Data into Focus
Intel-Supported Software Defined Visualization (SDVis)

Embree
- CPU Optimized Ray Tracing Algorithms
- ‘Tool kit’ for Building Ray Tracing Apps
- Broadly Adopted by 3rd Party ISVs
- Web Site: http://embree.github.io

OSPRay¹
- Rendering Engine Based on Embree
- Library and API Designed to Ease the Creation of Visualization Software
- Web Site: http://ospray.org

OpenSWR¹
- Rasterization Visualization on CPUs
- Good Enough to Replace HW GPU
- Supports ParaView, Visit, VTK
- Web Site: http://openswr.org

¹ Currently available in alpha
New storage paradigm for data intensive systems

SSF Enables HPC+HPDA workloads

- System components can be configured to match workload requirements
- Enables new access methodologies (DAOS) to create new generation applications
- Incremental improvements to Lustre to provide enhanced performance for existing applications

Distributed Asynchronous Object Storage

HOT
Crystal Ridge
~200 TB/s
~15 PB

WARM
3D NAND
~10-50 TB/s
~150 PB

COLD
SSD+HDD
~1-2 TB/s
~250 PB

Load/Store
Get/put
DAOS-M
Extreme BW+LL

Get/put
DAOS
LNET
High BW+IOPS

DAOS
LNET
High file create

Applications
Tools
- Top-level APIs
- DAOS-Cache/Teiring
- DAOS-Sharding/Resilience
- DAOS-Persistent Memory