Extreme Scale Computing Advances & Challenges in PIC Simulations

William M. Tang
Princeton University
Princeton, NJ  USA

High Performance Computing (HPC) 2014
From Clouds & Big Data to Exascale & Beyond

An International Advanced Research Workshop
Cetraro, Italy

July 7-11, 2014
Performance Development of HPC Over the Last 20 Years from Top 500 Supercomputers Worldwide

Note: This has a storage capacity rivaling the text-based content of a major research library.
Applications Impact ➔ Actual value of extreme Scale HPC to scientific domain applications & industry

• Practical Considerations: achieving “buy-in” from general scientific community ➔ need to:
  - Distinguish between “voracious” (more of same - just bigger & faster) vs. “transformational” (achievement of major new levels of scientific understanding)
  - Improve on experimental validation together with verification & uncertainty quantification to enhance realistic predictive capability of both hypothesis-driven and big-data-driven statistical approaches
  - New software engineering tools & environments to enable improved “time to solution” -- without big tax on improvements of science in targeted applications

• Associated Extreme Scale Computing Challenges: As hardware performance & storage capacity increases through many orders of magnitude,
  ➔ Produce advances featuring balanced combination of memory bandwidth, interconnect performance, computational performance, & reliability

■ Hardware complexity: Heterogenous multicore (e.g., gpu+cpu ➔ “Titan”; mic+cpu ➔ “TH-2”)

■ Software challenges: New operating systems, I/O and file systems, and coding/algorithmic & solver advances in volatile environment of vastly increased computer architecture complexity that demands rewriting code focused on data movement over arithmetic ➔ innovative deployable software!

• Applications Imperative: “Accountability” aspect
  ➔ Need to articulate what impactful scientific and mission advances have been enabled by the rapid progress from terascale to petascale to today’s multi-petascale HPC capabilities?
Advanced Scientific HPC Codes --- “a measure of the state of understanding of natural and engineered systems”

- Problem with Mathematical Model?
  - Theory (Mathematical Model)
  - Applied Mathematics (Basic Algorithms)
  - Computational Physics (Scientific Codes)
  - Computer Science (System Software)

- Problem with Computational Method?
  - Computational Predictions
    - Agree* w/ Experiments?
      - Yes
        - Speed/Efficiency?
          - Inadequate
          - Adequate
            - “Performance” Loop*
      - No
        - “V&V + UQ” Loop*

*Comparisons: “empirical/extrapolation” trends; sensitivity studies; detailed structure (spectra, correlation functions, …)

Use the New Tool for Scientific Discovery
(Repeat cycle as new phenomena encountered)

*Modern “co-design” Challenges: low memory/core; locality; latency; …..
ITER Goal: Demonstration of the Scientific and Technological Feasibility of Fusion Power

- **ITER** is an ~$25B facility located in France & involving 7 governments representing over half of world’s population. 
  
  ➔ **dramatic next-step for Magnetic Fusion Energy (MFE)** producing a sustained burning plasma
  
  -- Today: 10 MW(th) for 1 second with gain ~1  
  -- ITER: 500 MW(th) for >400 seconds with gain >10

- **“DEMO”** will be demonstration fusion reactor after ITER
  
  -- 2500 MW(th) continuous with gain >25, in a device of similar size and field as ITER

- Ongoing R&D programs worldwide [experiments, theory, computation, and technology] essential to provide growing knowledge base for ITER operation targeted for ~ 2020

➔ **Realistic HPC-enabled simulations required to cost-effectively plan, “steer,” & harvest key information from expensive (~$1M/long-pulse) ITER shots**
Microturbulence in Fusion Plasmas – Mission Importance: *Fusion reactor size & cost determined by balance between loss processes & self-heating rates*

- **“Scientific Discovery”** - Transition to favorable scaling of confinement produced in simulations for ITER-size plasmas
  - $a/\rho_i = 400$ (JET, largest present lab experiment) through
  - $a/\rho_i = 1000$ (ITER, ignition experiment)

- **Multi-TF simulations** using GTC global PIC code [Z. Lin, et al, 2002] deployed a billion particles, 125M spatial grid points; 7000 time steps $\Rightarrow$ *1st ITER-scale simulation with ion gyroradius resolution*

- BUT, **compelling understanding** of plasma size scaling demands *higher physics fidelity* requiring *much greater computational resources + new algorithms & modern diagnostics for VV&UQ*

  $\Rightarrow$ Excellent Scalability of Global PIC Codes on modern HPC platforms enables much greater resolution/physics fidelity to improve understanding

$\Rightarrow$ Progress from current DOE INCITE Projects on LCF’s & from ongoing G8 Fusion Exascale Project on major international facilities

$\Rightarrow$ BUT - further improvements for efficient usage of current LCF’s demands code re-write featuring modern CS/AM methods addressing locality, low-memory-per-core, …..
Particle Simulation of the Boltzmann-Maxwell System

- The Boltzmann equation (Nonlinear PDE in Lagrangian coordinates):
  \[
  \frac{dF}{dt} = \frac{\partial F}{\partial t} + \mathbf{v} \cdot \frac{\partial F}{\partial \mathbf{x}} + \left( \mathbf{E} + \frac{1}{c} \mathbf{v} \times \mathbf{B} \right) \cdot \frac{\partial F}{\partial \mathbf{v}} = C(F).
  \]

- “Particle Pushing” (Linear ODE’s)
  \[
  \frac{dx_j}{dt} = v_j, \quad \frac{dv_j}{dt} = \frac{q}{m} \left( \mathbf{E} + \frac{1}{c} \mathbf{v} \times \mathbf{B} \right)_{x_j}.
  \]

- Klimontovich-Dupree representation,
  \[
  F = \sum_{j=1}^{N} \delta(\mathbf{x} - \mathbf{x}_j) \delta(\mathbf{v} - \mathbf{v}_j),
  \]

- Poisson’s Equation: (Linear PDE in Eulerian coordinates (lab frame))
  \[
  \nabla^2 \phi = -4\pi \sum_{\alpha} q_{\alpha} \sum_{j=1}^{N} \delta(\mathbf{x} - \mathbf{x}_{\alpha j})
  \]

- Ampere’s Law and Faraday’s Law  [Linear PDE’s in Eulerian coordinates (lab frame)]
Basic Particle-in-Cell Method

- Charged particles sample distribution function
- Interactions occur on a grid with the forces determined by gradient of electrostatic potential (calculated from deposited charges)
- *Grid resolution dictated by Debye length (“finite-sized” particles) up to gyro-radius scale*

**Specific PIC Operations:**

- "SCATTER", or deposit, charges as “nearest neighbors” on the grid
- Solve Poisson Equation for potential
- “GATHER” forces (gradient of potential) on each particle
- Move particles *(PUSH)*
- Repeat…
New Physics Insights on Fusion Confinement Scaling Enabled by Computing at Extreme Scale
DOE INCITE Project on “Kinetic Simulations of Fusion Energy Dynamics @ Extreme Scale”

Objectives

• Develop modern software capable of using low memory supercomputers to carry out high physics fidelity first principles simulations of multiscale tokamak plasmas for magnetic fusion energy (MFE)

• Fusion Physics & HPC Challenges:
  → Key decade-long MFE estimates of confinement scaling with device size (“Bohm to Gyro-bohm” trend) need much higher resolution to be realistic/reliable.
  → Major algorithmic advances needed for MFE global PIC codes to effectively engage computing at extreme scale.

Impact

• Understanding the physics governing MFE confinement scaling → one of highest priority research areas for success of next-step burning plasma experiments (e.g., ITER)

• GTC-Princeton (“GTC-P”) makes efficient use of DoE’s LCF’s to carry out ITER scale simulations with unprecedented resolution in phase-space & time.

Accomplishments

• Production-run simulations of turbulence dynamics governing confinement physics for large-scale MFE plasmas (e.g., ITER) have been successfully carried out for the first time with very high phase-space resolution and long temporal duration.

• Co-design interdisciplinary research has now produced “GTC-P” – a modern HPC fusion energy science code that enables efficient use of multi-petascale capabilities on world-class CPU systems such as the IBM BG-Q “Mira” @ ALCF & “Sequoia” @ LLNL to deliver important new scientific insights.
BG-Q Performance: Weak Scaling Results

- Mira @ ANL & Sequoia @ LLNL
- C-Version of GTC-P Global GK PIC Code: 200 ppc resolution
- Plasma system size increases from A to D with D being ITER

Bei Wang (Princeton U.) & S. Ethier (PPPL)

*NNSA’s Sequoia (16.3 PF)
- excellent scaling to all 1,572,864 processor cores (capable of pushing over 130B particles)
- hybrid MPI+OpenMP in “GTC-P C” took full advantage of highly multi-threaded nodes and large scalable interconnect in BG-Q
K-Computer Performance: Weak Scaling Results

- **Fujitsu-K Computer @ RIKEN AICS, Kobe, Japan**
- **C-Version of GTC-P Global GK PIC Code: 200 ppc resolution**
- **Plasma system size increases from A to D with D being ITER**
# Performance Evaluation Platforms

<table>
<thead>
<tr>
<th>Systems</th>
<th>IBM BG/Q</th>
<th>Cray XK7 (Titan)</th>
<th>Cray XC 30 (Piz Daint)</th>
<th>NVIDIA Kepler</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chips per node</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Cores per chip</td>
<td>16</td>
<td>8</td>
<td>8</td>
<td>14 (SMX’s)</td>
</tr>
<tr>
<td>Interconnect</td>
<td>Custom 5D Torus</td>
<td>Gemini 3D Torus</td>
<td>Aries Dragonfly</td>
<td>-</td>
</tr>
<tr>
<td>Core</td>
<td>IBM A2</td>
<td>AMD Opteron 6274 (Interlagos)</td>
<td>Intel Xeon E5-2670 (SNBe)</td>
<td>K20x</td>
</tr>
<tr>
<td>Clock (GHz)</td>
<td>1.6</td>
<td>2.2</td>
<td>2.6</td>
<td>0.732</td>
</tr>
<tr>
<td>Cores per chip</td>
<td>16</td>
<td>8</td>
<td>8</td>
<td>14 (SMX’s)</td>
</tr>
<tr>
<td>Last-level Cache</td>
<td>32 MB</td>
<td>8 MB</td>
<td>20 MB</td>
<td>1.5 MB</td>
</tr>
<tr>
<td>DP GFlop/s per chip</td>
<td>204.8</td>
<td>281.6</td>
<td>166.4</td>
<td>1311</td>
</tr>
<tr>
<td>STREAM GB/s per node</td>
<td>28</td>
<td>?</td>
<td>38</td>
<td>171</td>
</tr>
</tbody>
</table>
Weak Scaling of GTC-P (GPU-version) on Heterogenous (GPU/CPU) "Titan" and "Piz Daint"

- The number of particles per cell is 100
- GTC-P GPU obtains 1.7x speed up

*Same code for all cases → Performance difference solely due to hardware/system software*
Recent GTC-P weak scaling results from “Stampede”

<table>
<thead>
<tr>
<th>Wall-clock time for pushing ion particles 100 time steps (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Problem</td>
</tr>
<tr>
<td>Nodes</td>
</tr>
<tr>
<td>2CPU</td>
</tr>
<tr>
<td>2CPU+1MIC</td>
</tr>
<tr>
<td>2CPU+2MIC</td>
</tr>
</tbody>
</table>

- B100 means “B-size problem with 100 ppc resolution;” Number of toroidal domains set at 32 for all problems; 1 MPI/16 OpenMP threads on the host, 1 MPI/240 OpenMP threads on each MIC.
- 512 nodes GTC-P simulation on “Stampede” targeted next.
Current Ongoing Investigations on “Stampede”

**Goals:**

– Improve intra-node communication between the host and the MICs to reduce overhead in the MPI Scatter operation in GTC-P

– Improve inter-node communication between MIC’s (for particle shift operation)
GTC-P Performance Studies on Heterogeneous (MIC/CPU) TH-2 System

- GTC-P ran successfully on up to 2048 nodes (host only) of TH-2 (Sept. 2013)
- In preparation for continuation of this collaboration, we engaged NSF’s “Stampede” (MIC/CPU) System [Oct.’13 to present] in developing a MIC version of GTC-P (for symmetric mode operation).
- **Stampede Results**: for 1MIC per node, obtain up to 1.46x speed up compared with CPU-only version of GTC-P
- “Lesson Learned” from running Intel MPI benchmark – via measuring latency & bandwidth of MPI communication for CPU to CPU; CPU to MIC; and MIC to MIC → Can optimize bandwidth between host and MIC’s by tuning GTC-P in accordance with optimal MPI communication pattern in “Stampede.”
  - plan to use same optimization approach for TH-2 studies using possible TH-2 benchmark data
Collaborative Studies with TH-2

• Measure MPI bandwidth between CPU to CPU, CPU to MIC and MIC to MIC on TH-2 using the Intel MPI benchmark

• **Test GTC-P MIC performance (symmetric mode) on TH-2**
  – Weak scaling performance: starting from A100 problem size on 224 TH-2 nodes, and ultimately with D100 (ITER) problem size on 14336 nodes
  – Deployment of 1MIC, 2MIC’s and 3MIC’s respectively for these weak scaling performance studies
Comments on HPC Extreme Scale Challenges

• Need more “demo-apps” that deploy innovative algorithms within modern codes that deliver new scientific insights on world-class systems – (e.g, BG/Q, K-Computer, Sequoia & Titan, Piz Daint, Stampede, TH-2)

Example from Fusion application domain: “Scientific Discovery in Fusion Plasma Turbulence Simulations @ Extreme Scale;” W. Tang, B. Wang, S. Ethier, to be published (Sept. 2014) in special issue on leadership computing in Computing in Science & Engineering (CiSE)

• Excellent performance scaling & “time-to-solution” have been achieved on top homogeneous architecture systems → still to be demonstrated on top heterogeneous GPU/CPU and GPU/MIC platforms

• Demonstration domain applications that deliver new science needed to help provide comparative performance studies on top supercomputing systems with “time to solution” as a viable metric.

→ Need algorithmic advances enabled by Applied Mathematics – in an interdisciplinary environment together with Computer Science & Domain Applications
GEOMETRIC HAMILTONIAN APPROACH TO SOLVING GENERALIZED VLASOV-MAXWELL EQUATIONS

Hamiltonian → Lagrangian → Action → Variational Optimization → Discretized Symplectic Orbits for Particle Motion

I. “Ultrahigh Performance 3-Dimensional Electromagnetic Relativistic Kinetic Plasma Simulation

Basic foundation for symplectic integration of particle orbits in electromagnetic fields without frequency ordering constraints
Foundational approach for present-day simulations of laser-plasma interactions on modern supercomputing systems
Limited applicability with respect to size of simulation region and geometric complexity

II. “Geometric Gyrokinetic Theory for Edge Plasmas”
Basic foundation for symplectic integration of particle orbits in electromagnetic low-frequency plasma following GK ordering
Still outstanding challenge: Address reformulation of non-local Poisson Equations structure for electromagnetic field solve
Summary: Challenges in Moving toward Exascale

• **Locality:** Need to develop mathematical algorithms able to deal with data locality
  -- due to physical limitations, moving data between, and even within, modern microchips is more time-consuming than performing computations!
  -- scientific codes often use data structures that are easy to implement quickly but limit flexibility and scalability in the long run

• **Advanced Architectures:** Need to deploy innovative algorithms within modern science codes on low memory per node architectures – (e.g., BG/Q, Fujitsu-K, Titan, & Tianhe-2)
  -- multi-threading within nodes, maximizing locality while minimizing communications
  → Substantive results achieved with GTC-P PIC code on IBM BG/Q (homogeneous architecture); good progress on hybrid (heterogeneous) CPU-GPU & CPU-MIC systems

• **Advanced Algorithms:** Need to develop Geometric Hamiltonian approaches most capable of ensuring locality of calculations and symplectic features
  -- Local EM field solve needed to complement existing local particle dynamics solve for Gyrokinetics
  *(Meanwhile, focus on deployment of fastest solvers (FMM, etc.)*
US/EU Statistical Disruption Studies on JET
W. M. Tang (Princeton University/PPPL)

• Situation Analysis:
  – The most critical of all problems facing magnetic fusion energy development is the need to avoid/mitigate large-scale major disruptions in tokamaks.
  -- The most advanced conventional “hypothesis-driven” MHD codes are currently still far away from producing the timely predictive capability needed for disruption avoidance in JET (Joint European Torus)—only experiment that achieved near “break-even” fusion energy production.

• Proposed “Big Data” Project: Use of large-data-driven statistical predictions for the occurrence of disruptions in JET

  – Based on new statistical machine-learning techniques developed in the Computer Science/Applied Math community in the U.S.

  – Use powerful hardware at the ORNL Leadership Class Facility for needed large-scale “data-mining” analysis of JET data

• Current Status:
  → JET has expressed serious willingness to provide access to their large disruption-relevant multi-dimensional data base that has yet to be analyzed.
  → Excellent opportunity for G8 NuFuSE Project to possibly leverage this important emerging “Big-Data Discovery” project on a problem of great importance for Fusion Energy futures.
Fusion Data Mining Diagram

JET Tokamak

Temperature  Density  Soft X-ray  ECEI

Data Streaming

Experimental Data Repository

On-line Multi-streaming Over WAN

Off-line Data Transfer

Predictive Model

Large-scale Multi-dimensional Data Mining Applications

Feature Extraction  DA Optimizer  Ensemble Model

Model/Classifier Update  Model/Classifier Deployment  Model/Classifier Creation

Stream Data Processing  Off-line Data Access

Local Storage

JET Site

ORNL/PPPL