Present and Future Computing Requirements for Materials Genomics

Kristin Persson, Anubhav Jain, Dan Gunter (LBNL), Gerbrand Ceder (MIT), Geoffroy Hautier (UCL), Mark Asta (UCB), Alan Dozier (UKY), Shyue-Ping Ong (UCSD)

...3.5K others

NERSC BES Requirements for 2017
October 8-9, 2013
Gaithersburg, MD
Materials Genomics:

Kristin Persson (LBNL) Gerbrand Ceder (MIT)

Problem: Materials discovery is crucial to a sustainable energy future (PV, batteries, permanent magnets, cement, thermo-electrics, catalysis, etc.) but time from lab to commercialization is 15 years.

Goal: Cut that time in half w/ simulation science.

Progress: Dedicated computing @ NERSC in 2012 35K materials computed, 6 apps, ~3K data users, progress in energy storage...
Online tools are popular with thousands of users. Many publications and new users in 2013. Four (known) drive-by publications in 2012.
MatGen Progress

- > 3,500 registered users
- > 18,000 materials records downloaded last 3 months through RESTful interface

**Companies:** Toyota, Sony, Bosch, 3M, Honda, Samsung, LG Chem, Dow Chemicals, GE Global Research, Applied Materials, Energizer, Advanced Materials, General Motors, Corning, DuPont, Nippon Steel, L’Oreal USA, Caterpillar, HP, Unilever, Lockheed Martin, Texas Instruments, Ford, Bose, Sigma-Aldrich, Siemens, Raytheon, Umicore, Seagate

**Battery startups:** Envia systems, Nanoexa, Pellion, Sion Power, Planar Energy Devices, Phostech, PolyPlus
2. Computational Strategy Questions

• We approach this problem computationally at a high level by ...
  • Simulation surveys guided by organized informatic interfaces
  • Data-driven science web applications layered on top
  • Read more here: https://materialsproject.org/escience

• The codes we use are ...
  • VASP, ABINIT, FEFF, more coming: Zeo++, BerkeleyGW, Qespresso

• These codes are characterized by these algorithms: ...
  • DFT still dominates $O(N^3)$, more electronic structure emerging
  • TD-DFT $O(N^3-N^6)$, GW/BSE $O(N^6)$, QMC $O(N^3)$ (big prefactor)

• Our biggest computational challenges are ...
  • Workflow, workflow, software integration, high throughput

• Our parallel scaling is limited by ...
  • Strong scaling, system size (defects, alloys, etc. grow system size)

• Changes in strategy for 2017
  • Same strategy. New targets: Electrolyte genome. Aq. Synthesis.
  • Better APIs for data and simulation, new apps
**Supercomputing Resources**

**Input processing & transformations**

- pymatgen
  - Robust materials analysis
- Custodian
  - Self-healing error recovery
- Fireworks
  - Smart workflow management

**Workflow Manager**

**Post-processing and error-checking**

**Analysis**

**Web apps**

**Materials API**
Current HPC Usage

Hours used in 2012-2013:

- 2012 (first year) 3.5M hours used, 15K materials / year
- 2013 20M hours requested, electrolyte genome
- ~6 months downtime scaling workflow → FireWorks

Machines currently used:

- Hopper – small memory convergence runs
- Mendel - 64 Gb/node Mendel vs 32 GB/node Hopper

Necessary software, services or infrastructure:

- Workflow and queues for HTC. Data analytics and visualization, collaboration tools, web interfaces, federated authentication services, and gateway support will be needed.
4. HPC Requirements for 2017

• Compute hours needed (in units of Hopper hours)
  • $O(10M)$ hours in 2013 $\rightarrow O(200M)$ hours in 2017
• Changes to parallel concurrency, run time, number of runs per year
  • number of runs will go way up with Fireworks, HTC queuing
  • constant data per material run 200MB-20GB
• Changes to data read/written
  • web platform where diverse & massive data sets can meet and federate
  • nosql and API data traffic will grow (interest in sharded mongo)
• Changes to memory needed per ( core | node | globally )
  • 128 GB/ node would be useful (high NBAND calcs)
• Changes to necessary software, services or infrastructure
  • fast access to 1PB nosql dataset (web access, analytics, ML)
  • persistent processes for service operation
5. Strategies for New Architectures

- Our strategy for running on new many-core technologies (GPUs or MIC) is ...
  - Leverage community code work with GPUs
- Adapting to many core by ...
  - App readiness & keeping on-node where possible
- To be successful on many-core systems we will need help with ...
  - Memory footprint. Not all codes can run on hopper now.
6. Summary of MatGen Computing and Data

- Computations “on demand” from community input now delivers actionable insight about new materials through the web.
- Software and workflow are rate limiters to increased cycle usage.
- Materials genomics will require further software co-design
  - Scalable key-value stores, HTC queuing, web data APIs
- Larger allocation will be needed for electrolyte genome.
  - Cycles often not the rate limiting step though.

- NERSC Data Pilot added value
  - Advanced architectures for MGI data challenges (SSD, mongo, testbeds)
  - Attention of data experts to workflow, database, formats, and analytics
  - ML for deeper understanding of workflow (conductor/insulator wall times)

- Community discussion needed on data policy, “data users”, how to measure publications and impact from research data availability.
MatProj Computing Specifics

Growing allocation w/ growing systems sizes

- Defects for functional electronics
- Nanoparticle synthesis
- Stability in aqueous environments

Predicting aqueous synthesis products

- More atoms
- More bands
- More exp. data

Corrosion

Dissolution

Interstitial radius (Å)
MatProj Data Specifics

• Materials Project leverages NoSQL DB technology:
  – Allows for frequent changes based on new knowledge and capabilities
  – Lets domain experts gain more control of their data representations
  – Path to scalability for larger and more distributed data sets

• Materials Design: Private sandboxes – especially interesting for Industry
  – Leverages public data and capabilities but keeps data private
  – Flexible web interfaces to all data make this possible

• Data incorporation from different sources
  – CO₂ capture data, exp. data, etc. – tremendous challenge
  – Standard, generic import with provenance already defined
  – Defining robust, general validation that verifies data integrity at the database and the scientific level

• Data analytics services: Fast/deep queries on entire MatProj data
Synergistic efforts in C.S.

• Workflow research (DOE/ASCR)
  – co-PIs: Gunter, Ramakrishnan
  – Making workflows easier to write & better with data

• LBNL ALS LDRD, COSMO LDRD in HTC
  – PI: Craig Tull, ACS; co-PIs: Gunter, Ramakrishnan
  – End-to-end analysis pipeline for lightsource data
  – ✓ one end of the pipe is NERSC, "next to" MP DBs

• Research Data management research (DOE/ASCR)
  – co-PI: L. Ramakrishnan
  – Better coord. of elastic computation and data
10X Demand from Simulation to Synthesis

• You can only manufacture what is stable in synthesis. Differences in surface energy stabilize polymorphs that are metastable in the bulk.
• Bigger Simulations. Requires modeling of size and chemistry dependent interactions on the surface as a function of nanoparticle size and environment solution to achieve rational synthesis for inorganic materials in solution
• 10x compute challenge from crystals, dedicated computational resources

Simulations showing how the crystalline phase is stabilized by water absorption

From Zhang et al, Nature 424, 2003

Navrotsky, Geochem. Trans. 4 (6), 2003


Stable Pt nanoparticles as function of water environment conditions
ICSD enumerated structures (e.g., MOFs) structure predictor experimental data PV, defects, x-ray data …

Computational Survey Drivers

Compute Engines

Production HPC w/ VASP, Zeo++, BerkeleyGW, PARATEC, QEspresso

Hands-off Expertise Automation (A)
High Throughput Computing (A)
Big Data Materials Ontologies (B)
Probabilistic Data Management (B)
Anomaly Detection (C)
Consistency / V&V (C)
Query Language / Web Gateway (D)
Experimental validation is key(!)

(C) functional electronic materials phosphors & scintillators carbon storage critical materials…

The materials genome

HPC Storage, & Data Analytics

Community Web Gateway
## MatProj Active User Patterns

<table>
<thead>
<tr>
<th>Number of distinct users that….</th>
<th>Past month</th>
<th>3 months</th>
<th>Year</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logged in to the site</td>
<td>~280</td>
<td>~600</td>
<td>~2050</td>
</tr>
<tr>
<td>Queried the Materials Explorer</td>
<td>180</td>
<td>350</td>
<td>N/A</td>
</tr>
<tr>
<td>Made a Phase Diagram</td>
<td>120</td>
<td>270</td>
<td>810</td>
</tr>
<tr>
<td>Used Moogle</td>
<td>100</td>
<td>220</td>
<td>not launched</td>
</tr>
<tr>
<td>Used the Crystal Toolkit to design materials</td>
<td>80</td>
<td>170</td>
<td>not launched</td>
</tr>
<tr>
<td>Searched for battery materials</td>
<td>50</td>
<td>130</td>
<td>390</td>
</tr>
<tr>
<td>Predicted a structure</td>
<td>30</td>
<td>90</td>
<td>390</td>
</tr>
<tr>
<td>Calculated a reaction energy</td>
<td>30</td>
<td>80</td>
<td>290</td>
</tr>
<tr>
<td>Used the REST interface to get data on a large scale</td>
<td>20</td>
<td>30</td>
<td>not launched</td>
</tr>
</tbody>
</table>