Quantum Chemistry (QC) on GPUs
Feb. 2, 2017
Overview of Life & Material Accelerated Apps

**MD:** All key codes are GPU-accelerated

- Great multi-GPU performance
- Focus on dense (up to 16) GPU nodes &/or large # of GPU nodes
- ACEMD*, AMBER (PMEMD)*, BAND, CHARMM, DESMOND, ESPResSo, Folding@Home, GPUgrid.net, GROMACS, HALMD, HOOMD-Blue*, LAMMPS, Lattice Microbes*, mdcore, MELD, miniMD, NAMD, OpenMM, PolyFTS, SOP-GPU* & more

**QC:** All key codes are ported or optimizing

- Focus on using GPU-accelerated math libraries, OpenACC directives
- GPU-accelerated and available today:
  - ABINIT, ACES III, ADF, BigDFT, CP2K, GAMESS, GAMESS-UK, GPAW, LATTE, LSDalton, LSMS, MOLCAS, MOPAC2012, NWChem, OCTOPUS*, PETot, QUICK, Q-Chem, QMCPack, Quantum Espresso/PWscf, QUICK, TeraChem*
- Active GPU acceleration projects:
  - CASTEP, GAMESS, Gaussian, ONETEP, Quantum Supercharger Library*, VASP & more

green* = application where >90% of the workload is on GPU
## MD vs. QC on GPUs

<table>
<thead>
<tr>
<th>“Classical” Molecular Dynamics</th>
<th>Quantum Chemistry (MO, PW, DFT, Semi-Emp)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulates positions of atoms over time; chemical-biological or chemical-material behaviors</td>
<td>Calculates electronic properties; ground state, excited states, spectral properties, making/breaking bonds, physical properties</td>
</tr>
<tr>
<td>Forces calculated from simple empirical formulas (bond rearrangement generally forbidden)</td>
<td>Forces derived from electron wave function (bond rearrangement OK, e.g., bond energies)</td>
</tr>
<tr>
<td>Up to millions of atoms</td>
<td>Up to a few thousand atoms</td>
</tr>
<tr>
<td>Solvent included without difficulty</td>
<td>Generally in a vacuum but if needed, solvent treated classically (QM/MM) or using implicit methods</td>
</tr>
<tr>
<td>Single precision dominated</td>
<td>Double precision is important</td>
</tr>
<tr>
<td>Uses cuBLAS, cuFFT, CUDA</td>
<td>Uses cuBLAS, cuFFT, OpenACC</td>
</tr>
<tr>
<td>Geforce (Accademics), Tesla (Servers)</td>
<td>Tesla recommended</td>
</tr>
<tr>
<td>ECC off</td>
<td>ECC on</td>
</tr>
</tbody>
</table>
Accelerating Discoveries

Using a supercomputer powered by the Tesla Platform with over 3,000 Tesla accelerators, University of Illinois scientists performed the first all-atom simulation of the HIV virus and discovered the chemical structure of its capsid — “the perfect target for fighting the infection.”

Without gpu, the supercomputer would need to be 5x larger for similar performance.
<table>
<thead>
<tr>
<th>GPU-Accelerated Quantum Chemistry Apps</th>
<th>Green Lettering Indicates Performance Slides Included</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abinit</td>
<td>Gaussian</td>
</tr>
<tr>
<td>ACES III</td>
<td>GPAW</td>
</tr>
<tr>
<td>ADF</td>
<td>LATTE</td>
</tr>
<tr>
<td>BigDFT</td>
<td>LSDalton</td>
</tr>
<tr>
<td>CP2K</td>
<td>MOLCAS</td>
</tr>
<tr>
<td>GAMESS-US</td>
<td>Mopac2012</td>
</tr>
<tr>
<td>NWChem</td>
<td></td>
</tr>
<tr>
<td>Quantum SuperCharger Library</td>
<td></td>
</tr>
<tr>
<td>RMG</td>
<td></td>
</tr>
<tr>
<td>TeraChem</td>
<td></td>
</tr>
<tr>
<td>UNM</td>
<td></td>
</tr>
<tr>
<td>VASP</td>
<td></td>
</tr>
<tr>
<td>WL-LSMS</td>
<td></td>
</tr>
</tbody>
</table>

GPU Perf compared against dual multi-core x86 CPU socket.
ABINIT
ABINIT on GPUs

Speed in the parallel version:

- For ground-state calculations, GPUs can be used. This is based on CUDA+MAGMA

- For ground-state calculations, the wavelet part of ABINIT (which is BigDFT) is also very well parallelized: MPI band parallelism, combined with GPUs
Multiscale Modelling Methods for Applications in Materials Science  CECAM JÜLICH, GERMANY

Introduction to Electronic Structure Calculations with BigDFT

Thierry Deutsch, Damien Caliste, Luigi Genovese

L_Sim - CEA Grenoble

17 September 2013
BigDFT version 1.7: capabilities

- Free, surface and periodic boundary conditions
- Geometry optimizations (with constraints)
- Born-Oppenheimer Molecular Dynamics
- Saddle point searches (Nudged-Elastic Band Method)
- Vibrations
- External electric fields
- Unoccupied KS orbitals
- Collinear and Non-collinear magnetism
- All XC functionals of the ABINIT package
- Hybrid functionals
- Empirical van der Waals interactions (many flavors)
- Also available within the ABINIT package
BigDFT version 1.7: capabilities

- Free, surface and periodic boundary conditions
- Geometry optimizations (with constraints)
- Born-Oppenheimer Molecular Dynamics
- Saddle point searches (Nudged-Elastic Band Method)
- Vibrations
- External electric fields
- Unoccupied KS orbitals
- Collinear and Non-collinear magnetism
- All XC functionals of the ABINIT package
- Hybrid functionals
- Empirical van der Waals interactions (many flavors)
- Also available within the ABINIT package
GPU-ported operations in BigDFT (double precision)

Convolutions (OpenCL rewritten)

GPU speedups between 10 and 20 can be obtained for different sections

Linear algebra (CUBLAS library)
The interfacing with CUBLAS is immediate, with considerable speedups
BigDFT code on Hybrid architectures

BigDFT code can run on hybrid CPU/GPU supercomputers.
In multi-GPU environments, double precision calculations

No Hot-spot operations
Different code sections can be ported on GPU
up to 20x speedup for some operations,
7x for the full parallel code

Courtesy of BigDFT team @ CEA
Hands on

See


- First runs with BigDFT
- Basis-set convergence
- Acceleration example on different platforms:

  Kohn-Sham DFT Operation with GPU acceleration
Gaussian
Joint collaboration between Gaussian, NVDA and PGI for GPU acceleration:

No such press release exists for Intel MIC or AMD GPUs

Mike Frisch quote from press release:

“Calculations using Gaussian are limited primarily by the available computing resources,” said Dr. Michael Frisch, president of Gaussian, Inc. “By coordinating the development of hardware, compiler technology and application software among the three companies, the new application will bring the speed and cost-effectiveness of GPUs to the challenging problems and applications that Gaussian’s customers need to address.”
Excerpts from...

ENABLING THE ELECTRONIC STRUCTURE PROGRAM GAUSSIAN ON GPGPUS USING OPENACC

Roberto Gomperts (NVIDIA), Michael Frisch (Gaussian, Inc.), Giovanni Scalmani (Gaussian, Inc.), Brent Leback (NVIDIA/PGI)
PREVIOUSLY
Earlier Presentations

GRC Poster 2012

ACS Spring 2014


WATOC Fall 2014

Full presentation available

GTC Spring 2016 (this full recording at http://mygtc.gputechconf.com/quicklink/4r1305r; requires registration)
TOPICS

Gaussian: Design Guidelines, Parallelism and Memory Model

Implementation: Top-Down/Bottom-Up

OpenACC: Extensions, Hints & Tricks

Early Performance

Closing Remarks
GAUSSIAN

A Computational Chemistry Package that provides state-of-the-art capabilities for electronic structure modeling

Gaussian 09 is licensed for a wide variety of computer systems

All versions of Gaussian 09 contain virtually every scientific/modeling feature, and none imposes any artificial limitations on calculations other than computational resources and time constraints

Researchers use Gaussian to, among others, study molecules and reactions; predict and interpret spectra; explore thermochemistry, photochemistry and other excited states; include solvent effects, and many more
DESIGN GUIDELINES

General

Establish a Framework for the GPU-enabling of Gaussian

Code Maintainability (Code Unification)

Leverage Existing code/algorithms, including Parallelism and Memory Model

Simplifies Resolving Problems

Simplifies Improvement on existing code

Simplifies Adding New Code
DESIGN GUIDELINES

Accelerate Gaussian for Relevant and Appropriate Theories and Methods

Relevant: many users of Gaussian

Appropriate: time consuming and good mapping to GPUs

Resource Utilization

Ensure efficient use of all available Computational Resources

   CPU cores and memory

   Available GPUs and memory
CURRENT STATUS

**Single Node**

**Implemented**

- Energies for Closed and Open Shell HF and DFT (less than a handful of XC-functionals missing)
- First derivatives for the same as above
- Second derivatives for the same as above

**Using only**

- OpenACC
- CUDA library calls (BLAS)
IMPLEMENTATION MODEL

Application Code

GPU
Compute-Intensive Functions
Small Fraction of the Code
Large Fraction of Execution time

CPU
Rest of Sequential CPU Code
GAUSSIAN PARALLELISM MODEL

CPU Cluster

Linda

CPU Node

GPU

OpenMP

OpenACC
GAUSSIAN: MEMORY MODEL
CLOSING REMARKS

Significant Progress has been made in enabling Gaussian on GPUs with OpenACC

OpenACC is increasingly becoming more versatile

Significant work lies ahead to improve performance

Expand feature set:

- PBC, Solvation, MP2, ONIOM, triples-Corrections
ACKNOWLEDGEMENTS

Development is taking place with:

Hewlett-Packard (HP) Series SL2500 Servers (Intel® Xeon® E5-2680 v2 (2.8GHz/10-core/25MB/8.0GT-s QPI/115W, DDR3-1866)

NVIDIA® Tesla® GPUs (K40 and later)

PGI Accelerator Compilers (16.x) with OpenACC (2.5 standard)
Increase Performance with Kepler

Running GPAW 10258

The **blue nodes** contain 1x E5-2687W CPU (8 Cores per CPU).

The **green nodes** contain 1x E5-2687W CPU (8 Cores per CPU) and 1x or 2x NVIDIA K20X for the GPU.
Increase Performance with Kepler

Running GPAW 10258

The **blue nodes** contain 1x E5-2687W CPU (8 Cores per CPU).

The **green nodes** contain 1x E5-2687W CPUs (8 Cores per CPU) and 2x NVIDIA K20 or K20X for the GPU.
Increase Performance with Kepler

Running **GPAW** 10258

The **blue nodes** contain 2x E5-2687W CPUs (8 Cores per CPU).

The **green nodes** contain 2x E5-2687W CPUs (8 Cores per CPU) and 2x NVIDIA K20 or K20X for the GPU.
Multi-GPU Accelerated Large Scale Electronic Structure Calculations

Samuli Hakala
COMP Centre of Excellence
Department of Applied Physics
Aalto University School of Science
Email: samuli.hakala@aalto.fi

GPU Technology Conference, March 2013
GPAW

- Density Functional Theory (DFT) program package for electronic structure calculations
- Time-Dependent Density Functional Theory (TDDFT) is implemented in the linear response and time propagation schemes
- Can use real-space grids, atom centered basis functions or plane waves
- Random Phase Approximation (RPA) also available
- Scales to thousands of cores and suitable for large scale calculations
- Open Source software licensed under GPL

Ground state DFT calculation of 561 Au atom cluster on Blue Gene/P.
LibXC on GPUs

- A reusable library of >250 exchange-correlation functionals
- Used by 15 different codes (Abinit, GPAW, BigDFT, etc.)
- Can be a performance bottleneck for small systems
- Can “clone” existing functionals for GPU use with fairly minimal changes to existing LibXC code and parallelizes well over grid points
- More information:
  - [https://confluence.slac.stanford.edu/display/SUNCAT/libxc+on+GPUs](https://confluence.slac.stanford.edu/display/SUNCAT/libxc+on+GPUs)
- Work by Lin Li, Jun Yan, Christopher O’Grady (Stanford/SLAC)

<table>
<thead>
<tr>
<th>Functional</th>
<th>Type</th>
<th>Speedup ((GPU+CPU)/CPU)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PW, PW Mode, OB PW, PW RPA</td>
<td>LDA Correlation</td>
<td>23, 23, 23, 37</td>
</tr>
<tr>
<td>PBE, PBE sol, xPBE, PBE JRGX, RGE₂, APBE</td>
<td>GGA Correlation</td>
<td>56, 58, 58, 58, 58, 58</td>
</tr>
<tr>
<td>RPBE</td>
<td>GGA Exchange</td>
<td>95</td>
</tr>
<tr>
<td>TPSS</td>
<td>MGGA Exchange</td>
<td>51</td>
</tr>
</tbody>
</table>
# Ground State Performance

## Bulk Silicon
- 95 atoms with periodic boundary conditions, 380 bands and 1 k-point. Grid size: 56x56x80.
- Time is in seconds per one SCF iteration.
- Intel Xeon X5650, NVIDIA Tesla M2070

## Fullerene
- C60 molecule with 240 valence electrons. Grid size: 84x84x84
- Intel Xeon X5650, NVIDIA Tesla M2070

<table>
<thead>
<tr>
<th></th>
<th>Si95</th>
<th>C60</th>
</tr>
</thead>
<tbody>
<tr>
<td>Poisson Solver</td>
<td>1.8</td>
<td>13</td>
</tr>
<tr>
<td>Orthonormalization</td>
<td>23</td>
<td>11</td>
</tr>
<tr>
<td>Precondition</td>
<td>9.4</td>
<td>16</td>
</tr>
<tr>
<td>RMM-DIIS other</td>
<td>32</td>
<td>8.1</td>
</tr>
<tr>
<td>Subspace Diag</td>
<td>23</td>
<td>22</td>
</tr>
<tr>
<td>Other</td>
<td>2.7</td>
<td>3.5</td>
</tr>
<tr>
<td><strong>Total (SCF-Iter)</strong></td>
<td><strong>93</strong></td>
<td><strong>76</strong></td>
</tr>
<tr>
<td><strong>S-Up</strong></td>
<td><strong>14</strong></td>
<td><strong>20</strong></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>GPU</th>
<th>GPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S-Up</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>CPU</th>
<th>GPU</th>
</tr>
</thead>
<tbody>
<tr>
<td>%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>S-Up</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Multi-GPU Parallelization

- Parallelization is done with MPI
- Multiple GPUs can be used by domain decomposition or parallelization over k-points or spins
- Domain decomposition for the stencil operations involves exchanging boundary regions between neighboring nodes
- Communications between nodes require data movement: device memory $\rightarrow$ host memory $\rightarrow$ destinations node host memory $\rightarrow$ destinations node device memory.
- Overlaps receives, sends and computations in the middle part of the grid, BUT this causes issues with small grids
  - Small grids: Synchronous transfers
  - Medium grids: Asynchronous transfers
  - Large grids: Overlap calculations and asynchronous transfers
  - Combine of several wave functions and boundary regions into few large transfers
Weak Scalability (Carbon)

- The size of a carbon nanotube and the number of MPI tasks are varied from 80 atoms (240 states) to 320 atoms (1280 states) and 1 task to 12 tasks.
- Comparison between equal number of GPUs and CPU cores.
- CPU: Intel Xeon X5650  GPU: NVIDIA Tesla M2070
- Calculations performed on Vuori cluster at CSC
Strong Scalability

- Bulk silicon with 1151 atoms with periodic boundary conditions, 4604 bands and 1 k-point in the Brillouin zone.
- The number of GPUs is increased from 64 to 256.
- Grid size: 164x164x108
- Speed-up comparison to 64 GPUs.
- NVIDIA Tesla M2090
- Calculations performed on CURIE cluster in France at GENCI/CEA
Weak Scalability (Silicon)

- The size of bulk silicon system and the number of MPI tasks are varied from 383 atoms (1532 bands) to 2046 atoms (8188 bands) and 8 task to 256 tasks with periodic boundary conditions.
- The largest system requires about 1.3TB of memory for calculations.
- CPU: Intel Xeon E5640 GPU: NVIDIA Tesla M2090
Random Phase Approximation

GPAW Random Phase Approximation (RPA) code:
- 6000 lines of python, 1000 lines of C/CUDA (and re-uses many GPAW functions)
- Better than DFT for correlated materials, but more computationally expensive
- Useful for oxides, Van der Waals systems, etc.

GPU Techniques:
- Use BLAS3 “zherk” instead of BLAS2 “zher”
- Batch FFTs
- GPU kernels parallelized over atoms/bands/projector-functions
- No thunking: all calculations on GPU

Preliminary ((GPU+CPU)/CPU) speedup for 202-electron N₂-on-Ru: 30x

Work by Jun Yan, Lin Li, Christopher O’Grady (Stanford/SLAC)
Summary

- We have accelerated the most numerically intensive parts of ground state DFT calculations.
- Overall speed-ups in our tests varied from 8.8 to 19 depending on system size.
- Our multi-GPU implementation scales well even on large hybrid clusters.
- Code is available at GPAW Subversion repository.
- Acknowledgements to CSC and PRACE for computing resources.

NWChem
NWChem 6.3 Release with GPU Acceleration

- Addresses large complex and challenging molecular-scale scientific problems in the areas of catalysis, materials, geochemistry and biochemistry on highly scalable, parallel computing platforms to obtain the fastest time-to-solution.

- Researchers can for the first time be able to perform large scale coupled cluster with perturbative triples calculations utilizing the NVIDIA GPU technology. A highly scalable multi-reference coupled cluster capability will also be available in NWChem 6.3.

- The software, released under the Educational Community License 2.0, can be downloaded from the NWChem website at www.nwchem-sw.org
NWChem - Speedup of the non-iterative calculation for various configurations/tile sizes

System: cluster consisting of dual-socket nodes constructed from:

- 8-core AMD Interlagos processors
- 64 GB of memory
- Tesla M2090 (Fermi) GPUs

The nodes are connected using a high-performance QDR Infiniband interconnect

Courtesy of Kowolski, K., Bhaskaran-Nair, at al @ PNNL, JCTC (submitted)
Kepler, Faster Performance (NWChem)

Performance improves by **2x** with one GPU and by **3.1x** with 2 GPUs.
Quantum Espresso 5.4.0
December 2016
AUSURF112 on K80s

Running Quantum Espresso version 5.4.0

The blue node contains Dual Intel Xeon E5-2690 v4@2.6GHz [3.5GHz Turbo] (Broadwell) CPUs

The green node contains Dual Intel Xeon E5-2690 v4@2.6GHz [3.5GHz Turbo] (Broadwell) CPUs + Tesla K80 (autoboost) GPUs

1 Broadwell node

1 node + 4x K80 per node

*Lower is better

606.00

528.20

1.1X
AUSURF1120 on P100s PCIe

Running Quantum Espresso version 5.4.0

The blue node contains Dual Intel Xeon E5-2690 v4@2.6GHz [3.5GHz Turbo] (Broadwell) CPUs

The green nodes contain Dual Intel Xeon E5-2690 v4@2.6GHz [3.5GHz Turbo] (Broadwell) CPUs + Tesla P100 PCIe GPUs
TeraChem 1.5K

Speaker, Date
TeraChem 1.5K; TripCage on Tesla K40s & IVB CPUs
(Total Processing Time in Seconds)
TERACHEM 1.5K; TRIPCAGE ON TESLA K40S & HASWELL CPUS

TeraChem 1.5K; TripCage on Tesla K40s & Haswell CPUs
(Total Processing Time in Seconds)
TERACHEM 1.5K; TRIPCAGE ON TESLA K80S & IVB CPUS

(Total Processing Time in Seconds)
TERACHEM 1.5K; TRIPCAGE ON TESLA K80S & HASWELL CPUS

TeraChem 1.5K; TripCage on Tesla K80s & Haswell CPUs
(Total Processing Time in Seconds)

- 2 x Xeon E5-2698 v3@2.30GHz + 1 x Tesla K80 board (1 node)
- 2 x Xeon E5-2698 v3@2.30GHz + 2 x Tesla K80 boards (1 node)
- 2 x Xeon E5-2698 v3@2.30GHz + 4 x Tesla K80 boards (1 node)
<table>
<thead>
<tr>
<th>Configuration</th>
<th>Total Processing Time (in Seconds)</th>
</tr>
</thead>
<tbody>
<tr>
<td>2 x Xeon E5-2697 <a href="mailto:v2@2.70GHz">v2@2.70GHz</a> + 1 x Tesla K40@875Mhz (1 node)</td>
<td>~11,500</td>
</tr>
<tr>
<td>2 x Xeon E5-2697 <a href="mailto:v2@2.70GHz">v2@2.70GHz</a> + 2 x Tesla K40@875Mhz (1 node)</td>
<td>~6,500</td>
</tr>
<tr>
<td>2 x Xeon E5-2697 <a href="mailto:v2@2.70GHz">v2@2.70GHz</a> + 4 x Tesla K40@875Mhz (1 node)</td>
<td>~3,500</td>
</tr>
<tr>
<td>2 x Xeon E5-2697 <a href="mailto:v2@2.70GHz">v2@2.70GHz</a> + 8 x Tesla K40@875Mhz (1 node)</td>
<td>~2,500</td>
</tr>
</tbody>
</table>
TERACHEM 1.5K; BPTI ON TESLA K80S & IVB CPUS

TeraChem 1.5K; BPTI on Tesla K80s & IVB CPUs
(Total Processing Time in Seconds)
Terachem 1.5K; BPTI on Tesla K40s & Haswell CPUs

(Total Processing Time in Seconds)
TeraChem 1.5K; BPTI on Tesla K80s & Haswell CPUs

(Total Processing Time in Seconds)
TeraChem
Supercomputer Speeds on GPUs

Time for SCF Step

- TeraChem running on 8 C2050s on 1 node
- NWChem running on 4096 Quad Core CPUs in the Chinook Supercomputer
- Giant Fullerene C240 Molecule

Similar performance from just a handful of GPUs
Dollars spent on **GPUs** do 500x more science than those spent on CPUs.
Kepler’s Even Better

Kepler performs 2x faster than Tesla

TeraChem running on C2050 and K20C

First graph is of BLYP/G-31(d)
Second is B3LYP/6-31G(d)
Interface on K80s

The blue node contains Dual Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs

The green nodes contain Dual Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs + Tesla K80 (autoboost) GPUs

- 1x K80 is paired with Single Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell)

Interface between a platinum slab Pt(111) (108 atoms) and liquid water (120 water molecules) (468 ions)

ALGO = Fast (Davidson + RMM-DIIS)

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Time (seconds)</th>
<th>Speedup</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 Broadwell node</td>
<td>0.00171</td>
<td></td>
</tr>
<tr>
<td>1 node + 1x K80 per node</td>
<td>0.00173</td>
<td>1.0X</td>
</tr>
<tr>
<td>1 node + 2x K80 per node</td>
<td>0.00238</td>
<td>1.4X</td>
</tr>
<tr>
<td>1 node + 4x K80 per node</td>
<td>0.00317</td>
<td>1.9X</td>
</tr>
</tbody>
</table>
Interface on P100s PCIe

Running VASP version 5.4.1

The **blue node** contains Dual Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs

The **green nodes** contain Dual Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs + Tesla P100 PCIe GPUs

- 1x P100 PCIe is paired with Single Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell)

Interface between a platinum slab Pt(111) (108 atoms) and liquid water (120 water molecules) (468 ions)

1256 bands
762048 plane waves
ALGO = Fast (Davidson + RMM-DIIS)
Interface on P100s SXM2

Running VASP version 5.4.1

The blue node contains Dual Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs

The green nodes contain Dual Intel Xeon E5-2698 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs + Tesla P100 SXM2 GPUs

- 1x P100 SXM2 is paired with Single Intel Xeon E5-2698 v4@2.2GHz [3.6GHz Turbo] (Broadwell)

Interface between a platinum slab Pt(111) (108 atoms) and liquid water (120 water molecules) (468 ions)

1256 bands
762048 plane waves
ALGO = Fast (Davidson + RMM-DIIS)
Silica IFPEN on K80s

Running VASP version 5.4.1

The blue node contains Dual Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs

The green nodes contain Dual Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs + Tesla K80 (autoboost) GPUs

- 1x K80 is paired with Single Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell)

240 ions, cristobalite (high) bulk
720 bands
? plane waves
ALGO = Very Fast (RMM-DIIS)
Running VASP version 5.4.1

The **blue node** contains Dual Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs

The **green nodes** contain Dual Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs + Tesla P100 PCIe GPUs

- 1x P100 PCIe is paired with Single Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell)

240 ions, cristobalite (high) bulk
720 bands
? plane waves
ALGO = Very Fast (RMM-DIIS)
Silica IFPEN on P100s SXM2

Running VASP version 5.4.1

The blue node contains Dual Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs

The green nodes contain Dual Intel Xeon E5-2698 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs + Tesla P100 SXM2 GPUs

- 1x P100 SXM2 is paired with Single Intel Xeon E5-2698 v4@2.2GHz [3.6GHz Turbo] (Broadwell)

240 ions, cristobalite (high) bulk
720 bands
? plane waves
ALGO = Very Fast (RMM-DIIS)

<table>
<thead>
<tr>
<th>1 Broadwell node</th>
<th>1 node + 1x P100 SXM2 per node</th>
<th>1 node + 2x P100 SXM2 per node</th>
<th>1 node + 4x P100 SXM2 per node</th>
<th>1 node + 8x P100 SXM2 per node</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.00273</td>
<td>0.00352</td>
<td>0.00475</td>
<td>0.00616</td>
<td>0.00692</td>
</tr>
</tbody>
</table>
Si-Huge on K80s

Running VASP version 5.4.1

The blue node contains Dual Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs

The green nodes contain Dual Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs + Tesla K80 (autoboost) GPUs

- 1x K80 is paired with Single Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell)

512 Si atoms
1282 bands
864000 Plane Waves
Algo = Normal (blocked Davidson)
Running VASP version 5.4.1

The **blue node** contains Dual Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs

The **green nodes** contain Dual Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs + Tesla P100 PCIe GPUs

- 1x P100 PCIe is paired with Single Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell)

<table>
<thead>
<tr>
<th>Configuration</th>
<th>1 Broadwell node</th>
<th>1 node + 1x P100 PCIe per node</th>
<th>1 node + 2x P100 PCIe per node</th>
<th>1 node + 4x P100 PCIe per node</th>
<th>1 node + 8x P100 PCIe per node</th>
</tr>
</thead>
<tbody>
<tr>
<td>Performance</td>
<td>1.8X</td>
<td>2.3X</td>
<td>3.1X</td>
<td>3.9X</td>
<td></td>
</tr>
<tr>
<td>Time (s/100,000)</td>
<td>0.00019</td>
<td>0.00034</td>
<td>0.00044</td>
<td>0.00058</td>
<td>0.00074</td>
</tr>
</tbody>
</table>

512 Si atoms
1282 bands
864000 Plane Waves
Algo = Normal (blocked Davidson)
Si-Huge on P100s SXM2

Running VASP version 5.4.1

The blue node contains Dual Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs

The green nodes contain Dual Intel Xeon E5-2698 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs + Tesla P100 SXM2 GPUs

- 1x P100 SXM2 is paired with Single Intel Xeon E5-2698 v4@2.2GHz [3.6GHz Turbo] (Broadwell)

512 Si atoms
1282 bands
864000 Plane Waves
Algo = Normal (blocked Davidson)
Supported Systems on K80s

Running VASP version 5.4.1

The blue node contains Dual Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs

The green nodes contain Dual Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs + Tesla K80 (autoboost) GPUs

- 1x K80 is paired with Single Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell)

267 ions
788 bands
762048 plane waves

ALGO = Fast (Davidson + RMM-DIIS)
Supported Systems on P100s PCIe

Running VASP version 5.4.1

The **blue node** contains Dual Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs

The **green nodes** contain Dual Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs + Tesla P100 PCIe GPUs

- 1x P100 PCIe is paired with Single Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell)

**ALGO = Fast (Davidson + RMM-DIIS)**

- 267 ions
- 788 bands
- 762048 plane waves
Supported Systems on P100s SXM2

- **1x P100 SXM2** is paired with Single Intel Xeon E5-2698 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs
- **1 node + 1x P100 SXM2 per node**
- **1 node + 2x P100 SXM2 per node**
- **1 node + 4x P100 SXM2 per node**
- **1 node + 8x P100 SXM2 per node**

**Running VASP version 5.4.1**

- The **blue node** contains Dual Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs
- The **green nodes** contain Dual Intel Xeon E5-2698 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs + Tesla P100 SXM2 GPUs

- **1x P100 SXM2** is paired with Single Intel Xeon E5-2698 v4@2.2GHz [3.6GHz Turbo] (Broadwell)

**Performance Metrics**

- **267 ions**
- **788 bands**
- **762048 plane waves**

**ALGO = Fast (Davidson + RMM-DIIS)**
NiAl-MD on K80s

Running VASP version 5.4.1

The blue node contains Dual Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs

The green nodes contain Dual Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs + Tesla K80 (autoboost) GPUs

- 1x K80 is paired with Single Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell)

500 ions
3200 bands
729000 plane waves
ALGO = Fast (Davidson + RMM-DIIS)
NiAl-MD on P100s PCIe

Running VASP version 5.4.1

The blue node contains Dual Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs

The green nodes contain Dual Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs + Tesla P100 PCIe GPUs

- 1x P100 PCIe is paired with Single Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell)

500 ions
3200 bands
729000 plane waves
ALGO = Fast (Davidson + RMM-DIIS)
NiAl-MD on P100s SXM2

Running VASP version 5.4.1

The blue node contains Dual Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs

The green nodes contain Dual Intel Xeon E5-2698 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs + Tesla P100 SXM2 GPUs

- 1x P100 SXM2 is paired with Single Intel Xeon E5-2698 v4@2.2GHz [3.6GHz Turbo] (Broadwell)

500 ions
3200 bands
729000 plane waves
ALGO = Fast (Davidson + RMM-DIIS)
LiZnO on P100s PCIe

Running VASP version 5.4.1

The blue node contains Dual Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs

The green nodes contain Dual Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs + Tesla P100 PCIe GPUs

500 ions
3200 bands
729000 plane waves
ALGO = Fast (Davidson + RMM-DIIS)

<table>
<thead>
<tr>
<th></th>
<th>1 Broadwell node</th>
<th>1 node + 2x P100 PCIe per node</th>
<th>1 node + 4x P100 PCIe per node</th>
</tr>
</thead>
<tbody>
<tr>
<td>Time (seconds)</td>
<td>0.00106</td>
<td>0.00137</td>
<td>0.00153</td>
</tr>
</tbody>
</table>
LiZnO on P100s SXM2

Running VASP version 5.4.1

The **blue node** contains Dual Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs

The **green nodes** contain Dual Intel Xeon E5-2698 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs + Tesla P100 SXM2 GPUs

- 1x P100 SXM2 is paired with Single Intel Xeon E5-2698 v4@2.2GHz [3.6GHz Turbo] (Broadwell)

500 ions
3200 bands
729000 plane waves

**ALGO = Fast (Davidson + RMM-DIIS)**
**B.hR105 on P100s PCIe**

Running VASP version 5.4.1

The blue node contains Dual Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs

The green nodes contain Dual Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs + Tesla P100 PCIe GPUs

- 1x P100 PCIe is paired with Single Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell)

105 Boron atoms (B-rhombohedral structure)  
216 bands  
110592 plane waves  
Hybrid Functional with blocked Davicson (ALGO=Normal)  
LHFCALC=.True. (Exact Exchange)
Running **VASP** version 5.4.1

The **blue node** contains Dual Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs

The **green nodes** contain Dual Intel Xeon E5-2698 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs + Tesla P100 SXM2 GPUs

- 1x P100 SXM2 is paired with Single Intel Xeon E5-2698 v4@2.2GHz [3.6GHz Turbo] (Broadwell)

105 Boron atoms (**B-rhombohedral structure**)  
216 bands  
110592 plane waves  
Hybrid Functional with blocked Davicson (ALGO=Normal)  
**LHFCALC=.True. (Exact Exchange)**
B.aP107 on P100s PCIe

Running VASP version 5.4.1

The blue node contains Dual Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs

The green nodes contain Dual Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs + Tesla P100 PCIe GPUs

- 1x P100 PCIe is paired with Single Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell)

107 Boron atoms (symmetry broken 107-atom β′ variant)
216 bands
110592 plane waves

Hybrid functional calculation (exact exchange) with blocked Davidson. No KPoint parallelization.

Hybrid Functional with blocked Davidson (ALGO=Normal)

LHFCALC=.True. (Exact Exchange)
Running VASP version 5.4.1

The blue node contains Dual Intel Xeon E5-2699 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs

The green nodes contain Dual Intel Xeon E5-2698 v4@2.2GHz [3.6GHz Turbo] (Broadwell) CPUs + Tesla P100 SXM2 GPUs

- 1x P100 SXM2 is paired with Single Intel Xeon E5-2698 v4@2.2GHz [3.6GHz Turbo] (Broadwell)

107 Boron atoms (symmetry broken 107-atom $\beta'$ variant)

216 bands
110592 plane waves

Hybrid functional calculation (exact exchange) with blocked Davidson. No KPoint parallelization.

Hybrid Functional with blocked Davidson (ALGO=Normal)

LHFCALC=.True. (Exact Exchange)
Quantum Chemistry (QC) on GPUs
Dec, 19, 2016