

Linearly Scaling Three Dimensional Fragment Method for Large Scale Electronic Structure Calculations

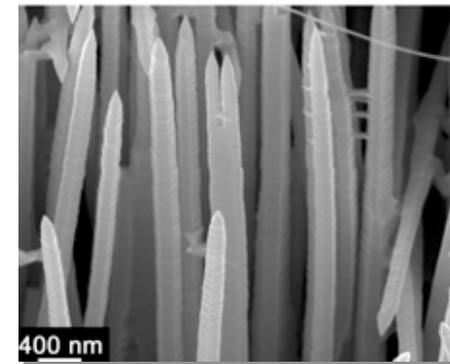
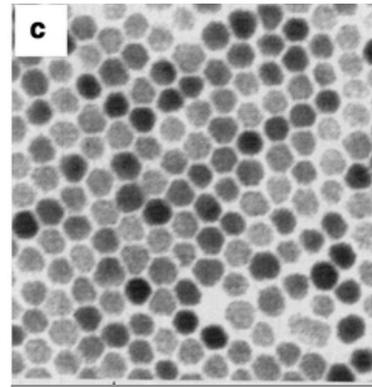
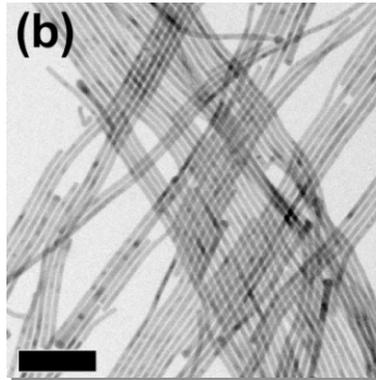
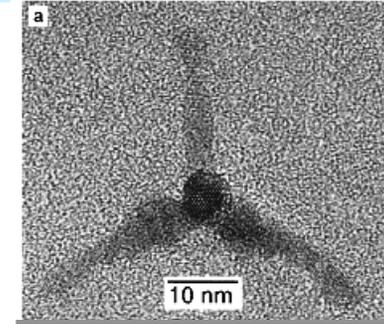
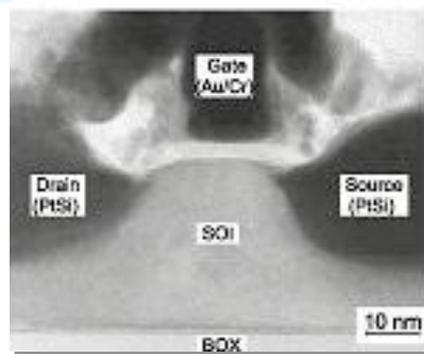
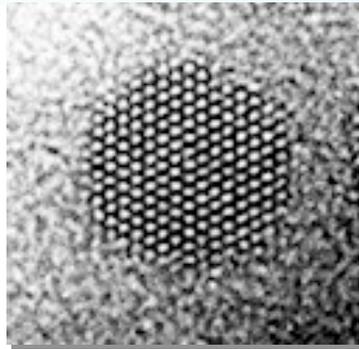
Lin-Wang Wang^{1,2}, Byounggak Lee¹, Zhengji Zhao², Hongzhang Shan^{1,2}, Juan Meza¹, David Bailey¹, Erich Strohmaier^{1,2}

¹)Computational Research Division

²)National Energy Research Scientific Computing Center (NERSC)
Lawrence Berkeley National Laboratory

US Department of Energy, Office of Science
Basic Energy Sciences and Advanced Scientific Computing Research

Nanostructures have wide applications including: solar cells, biological tags, electronics devices



- ❖ Different electronic structures than bulk materials
- ❖ 1,000 ~ 100,000 atom systems are too large for direct $O(N^3)$ *ab initio* calculations
- ❖ $O(N)$ computational methods are required
- ❖ Parallel supercomputers critical for the solution of these systems

Why are quantum mechanical calculations so computationally expensive?

$$\left[-\frac{1}{2}\nabla^2 + V_{tot}(r)\right]\psi_i(r) = \varepsilon_i\psi_i(r)$$

- ❖ If the size of the system is N :
- ❖ N coefficients to describe one wavefunction $\psi_i(r)$
- ❖ $i = 1, \dots, M$ wavefunctions $\psi_i(r)$, M is proportional to N .
- ❖ Orthogonalization: $\int \psi_i(r)\psi_j^*(r)d^3r$, M^2 wavefunction pairs, each with N coefficients: $N*M^2$, i.e N^3 scaling.

The repeated calculation of these orthogonal wavefunctions make the computation expensive, $O(N^3)$. For large systems, an $O(N)$ method is critical

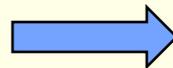
Previous Work on Linear Scaling DFT methods

- ❖ Three main approaches:
 - Localized orbital method
 - Truncated density matrix method
 - Divide-and-conquer method
- ❖ Some current methods include:
 - Parallel SIESTA (atomic orbitals, not for large parallelization)
 - Many quantum chemistry codes (truncated D-matrix, Gaussian basis, not for large parallelization)
 - ONETEP (M. Payne, PW to local orbitals, then truncated D-matrix)
 - CONQUEST (D. Bowler, UCL, localized orbital)
- ❖ Most of these use localized orbital or truncated-D matrix
- ❖ None of them scales to tens of thousands of processors

Linearly Scaling 3 Dimensional Fragment method (LS3DF)

- ❖ A novel divide and conquer scheme with a new approach for patching the fragments together
- ❖ No spatial partition functions needed
- ❖ Uses overlapping positive and negative fragments
- ❖ New approach minimizes artificial boundary effects

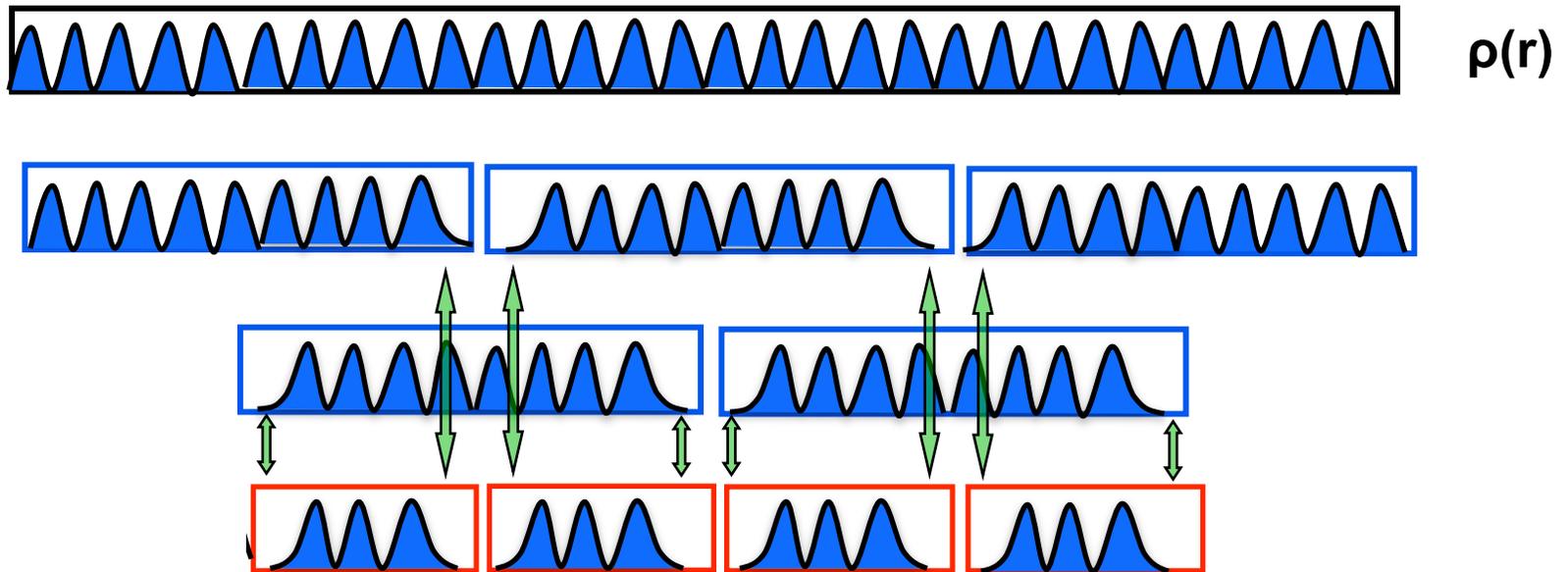
divide-and-conquer method



$O(N)$ scaling

Massively parallelizable

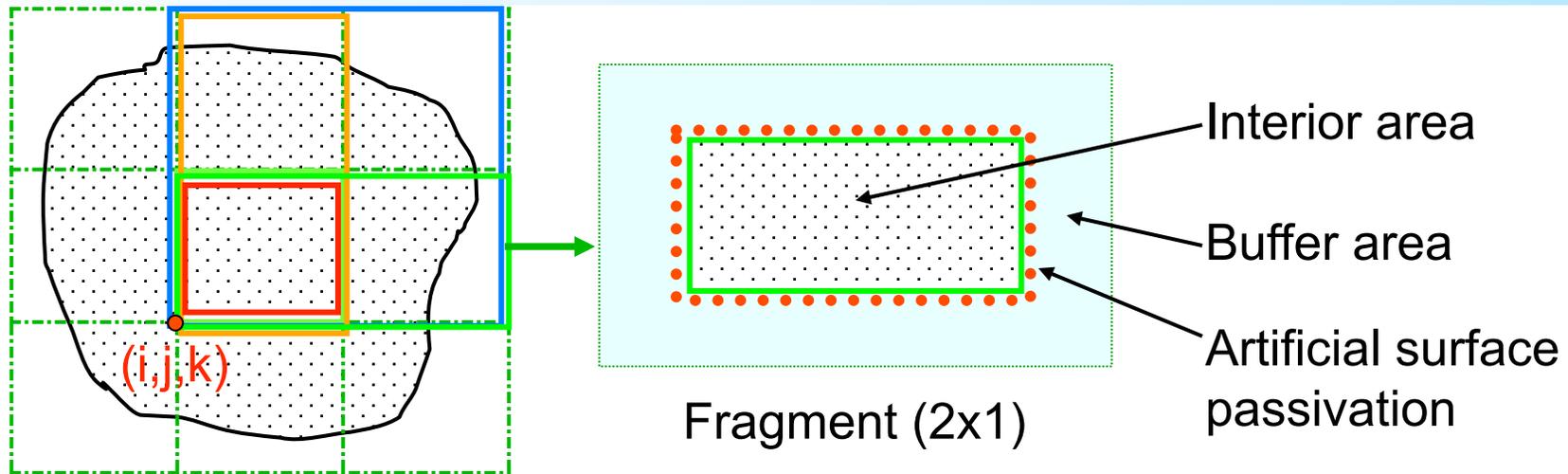
LS3DF: 1D Example



$$\text{Total} = \sum_F \{ \boxed{}_F - \boxed{}_F \}$$

Phys. Rev. B 77, 165113 (2008); J. Phys: Cond. Matt. 20, 294203 (2008)

Similar procedure extends to 2 and 3D



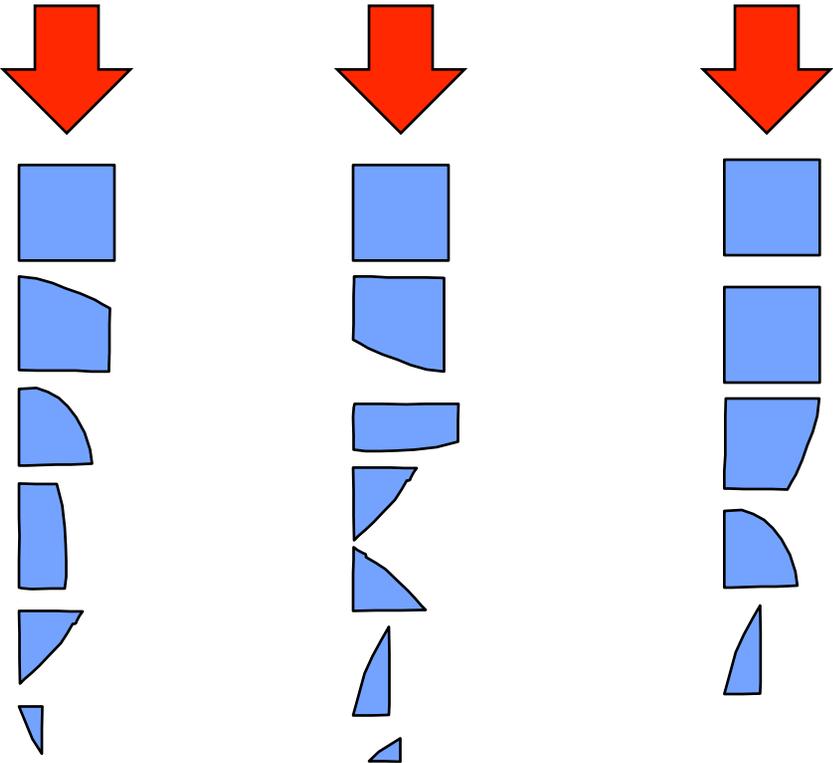
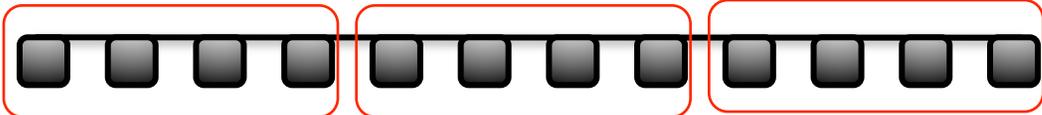
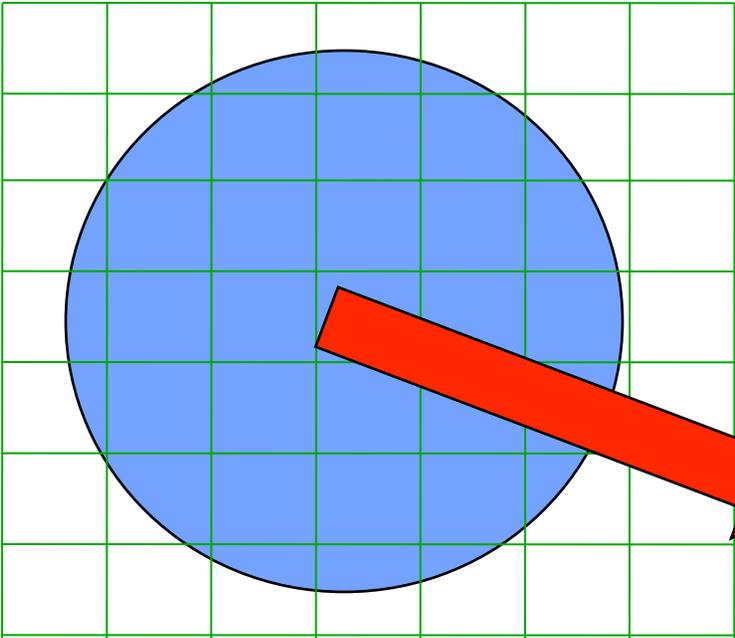
Total = $\Sigma_F \{$

$\} F - F + F F$

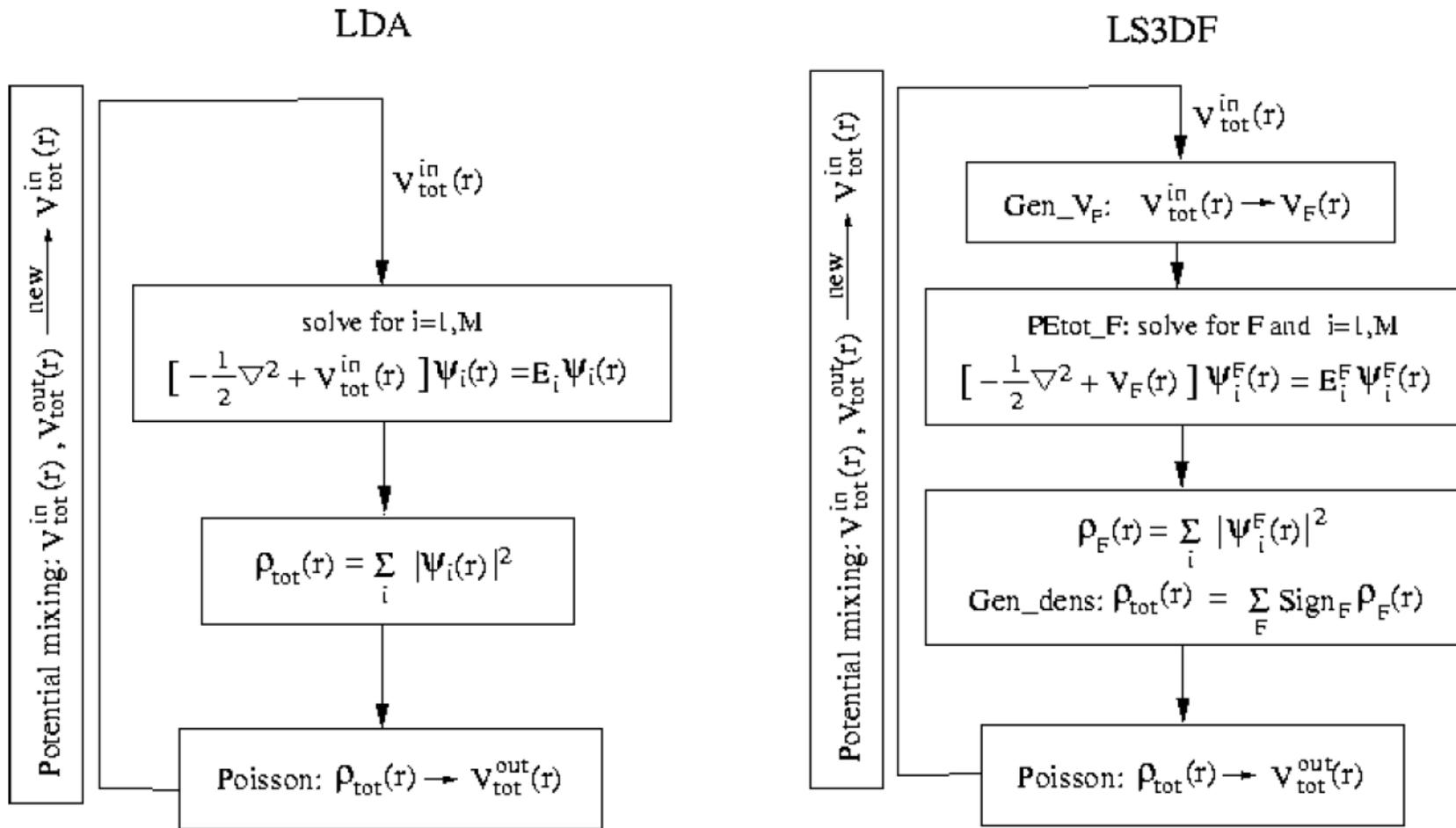
Boundary effects are (nearly) cancelled out between the fragments

$$System = \sum_{i,j,k} \{ F_{222} + F_{211} + F_{121} + F_{112} - F_{221} - F_{212} - F_{122} - F_{111} \}$$

Schematics for LS3DF calculation



Flow chart for LS3DF method



Based on the plane wave PEtot code: <http://hpcrd.lbl.gov/~linwang/PEtot/PEtot.html>

LS3DF Accuracy is determined by fragment size

- ❖ A comparison to direct LDA calculation, with an 8 atom 1x1x1 fragment size division:
 - The total energy error: 3 MeV/atom \sim 0.1 kcal/mol
 - Charge density difference: 0.2%
 - Better than other numerical uncertainties (e.g. PW cut off, pseudopotential)
- ❖ Atomic force difference: 10^{-5} a.u.
 - Smaller than the typical stopping criterion for atomic relaxation
- ❖ Other properties:
 - The dipole moment error: 1.3×10^{-3} Debye/atom, 5%
 - Smaller than other numerical errors

For most practical purposes, the LS3DF is the same as direct LDA

Some details on the LS3DF divide and conquer scheme

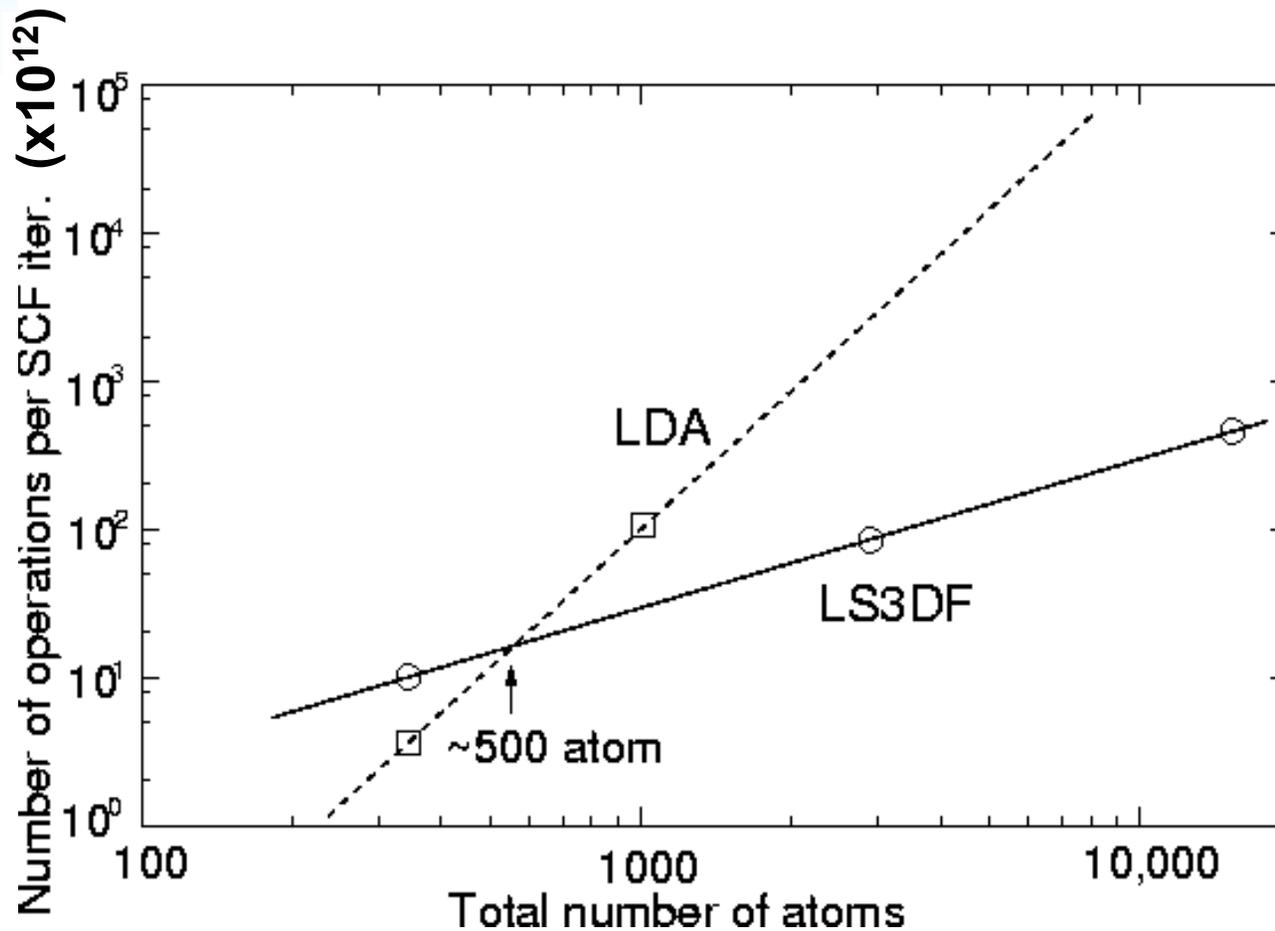
- ❖ Variational formalism, sound mathematics
- ❖ The division into fragments is done automatically, based on atom's spatial locations
- ❖ Typical large fragments (2x2x2) have ~100 atoms and the small fragments (1x1x1) have ~ 20 atoms
- ❖ Processors are divided into M groups, each with N_p processors.
 - N_p is usually set to 16 – 128 cores
 - M is between 100 and 10,000
- ❖ Each processor group is assigned N_f fragments, according to estimated computing times, load balance within 10%.
 - N_f is typically between 8 and 100

Overview of computational effort in LS3DF

- ❖ Most time consuming part of LS3DF calculation is for the fragment wavefunctions
 - Modified from the stand alone PEtot code
 - Uses planewave pseudopotential (like VASP, Qbox)
 - All-band algorithm takes advantage of BLAS3
- ❖ 2-level parallelization:
 - q-space (Fourier space)
 - band index (i in $\psi_i(r)$)
- ❖ PEtot efficiency > 50% for large systems (e.g, more than 500 atoms), 30-40% for our fragments.

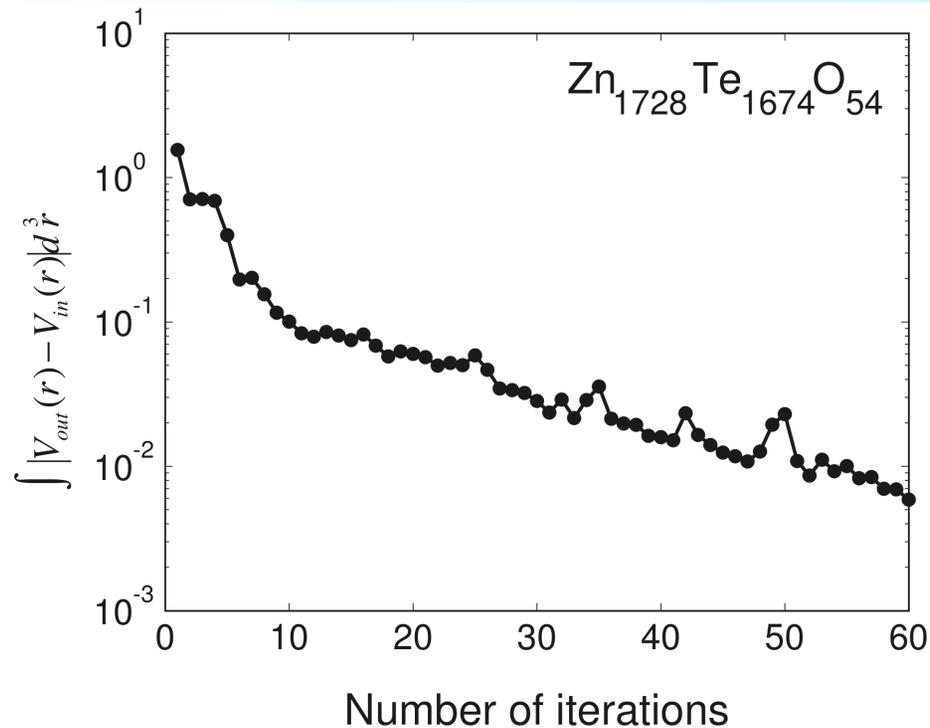
PEtot code: <http://hpcrd.lbl.gov/~linwang/PEtot/PEtot.html>

Operation counts

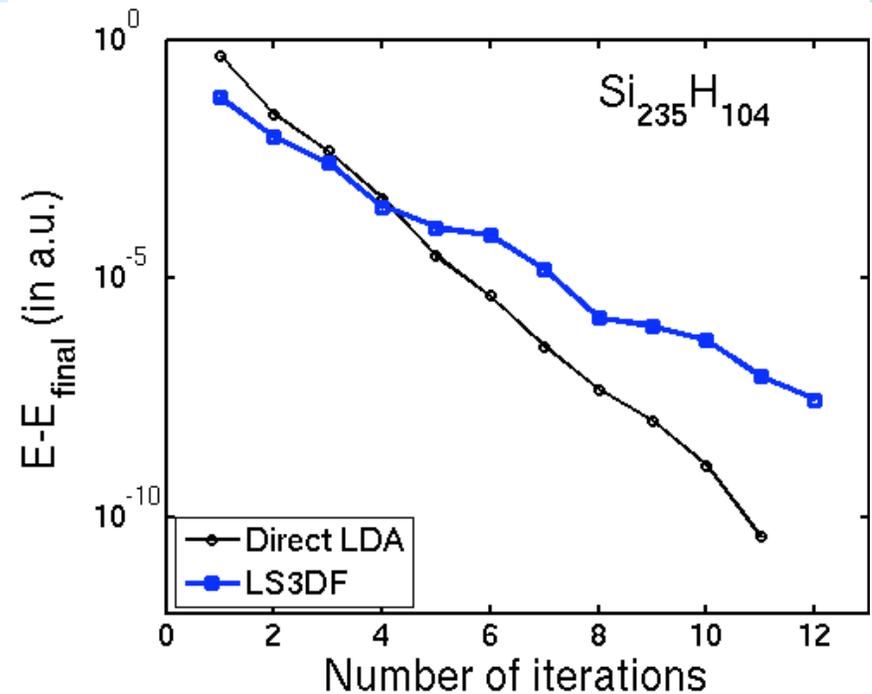


- ❖ Cross over with direct LDA method [PEtot] is 500 atoms.
- ❖ Similar to other $O(N)$ methods.

Selfconsistent convergence of LS3DF



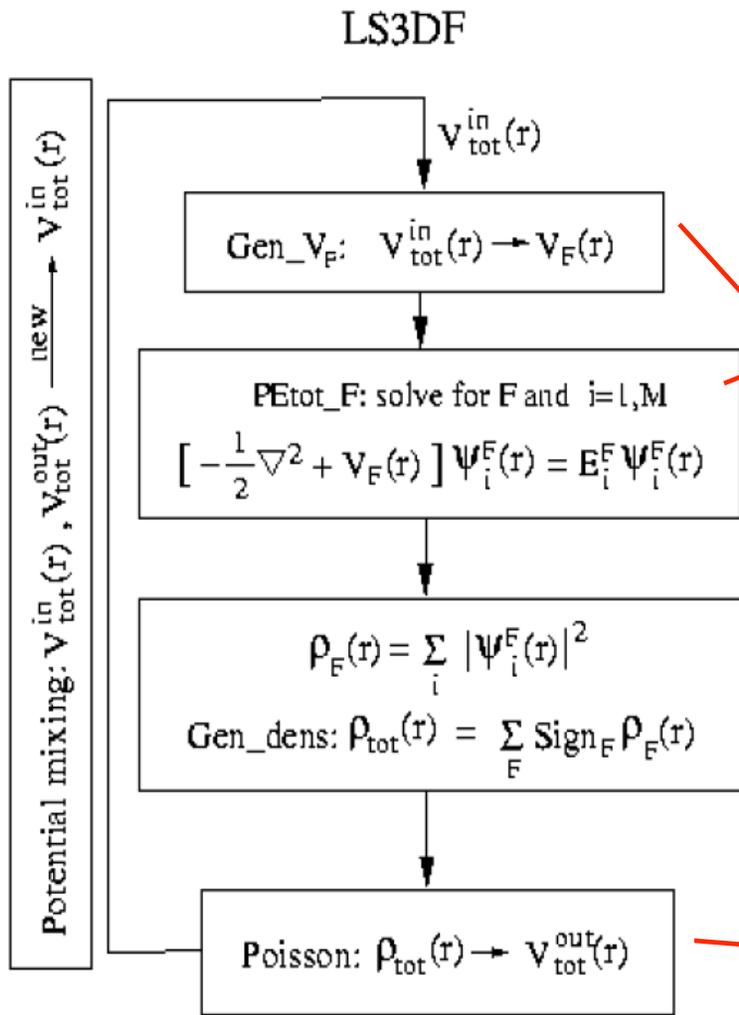
Measured by potential



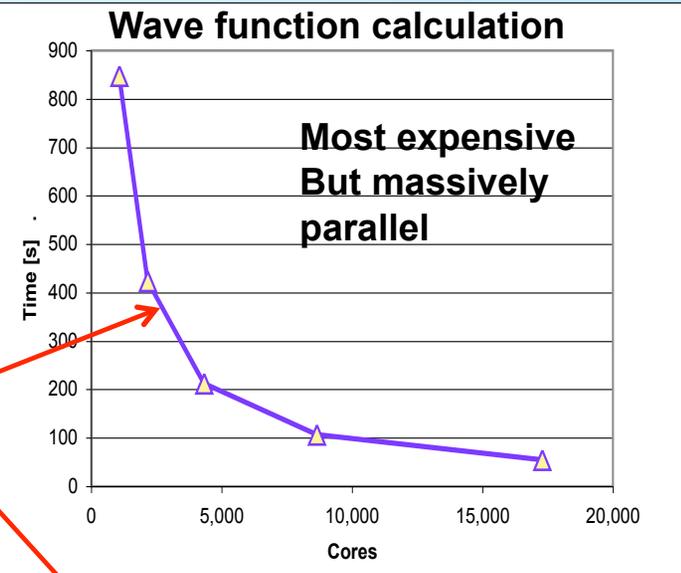
Measured by total energy

- ❖ SCF convergence of LS3DF is similar to direct LDA method
- ❖ It doesn't have the SCF problem some other $O(N)$ methods have

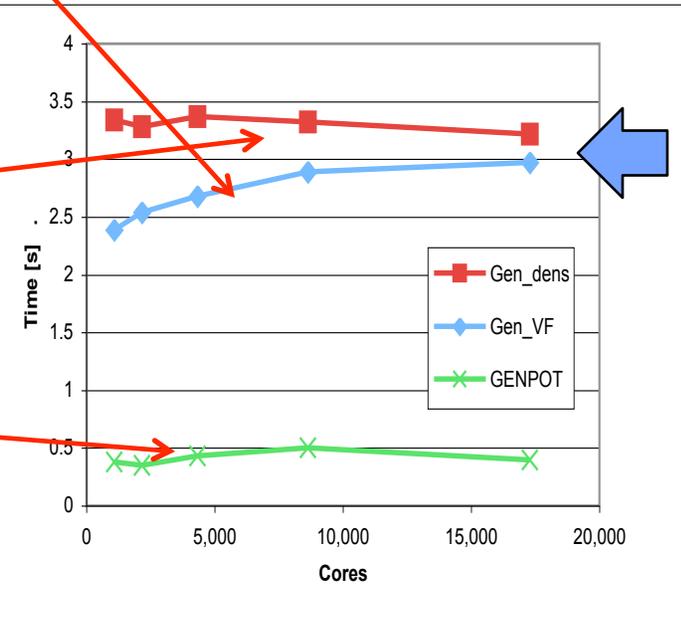
The performance of LS3DF method (strong scaling, NERSC Franklin)



Time (second)

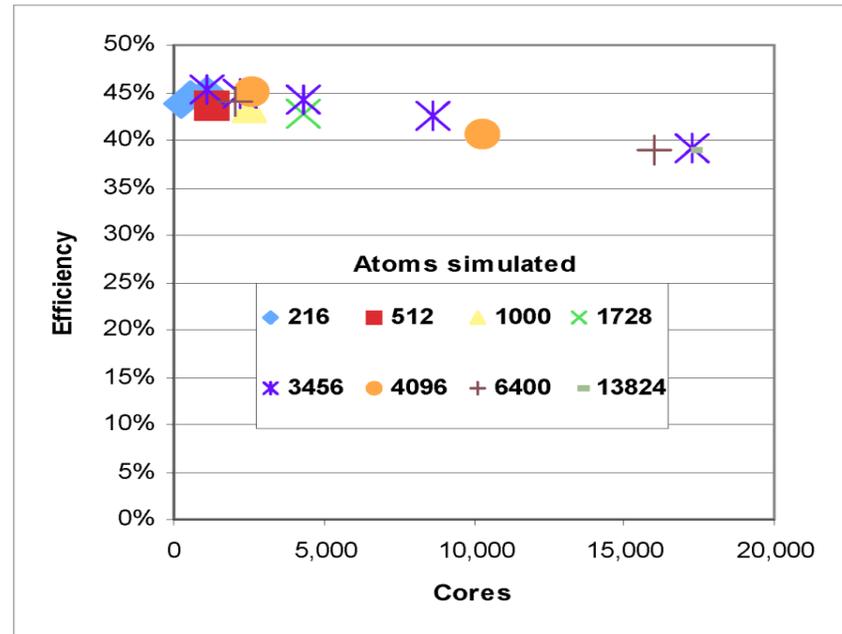
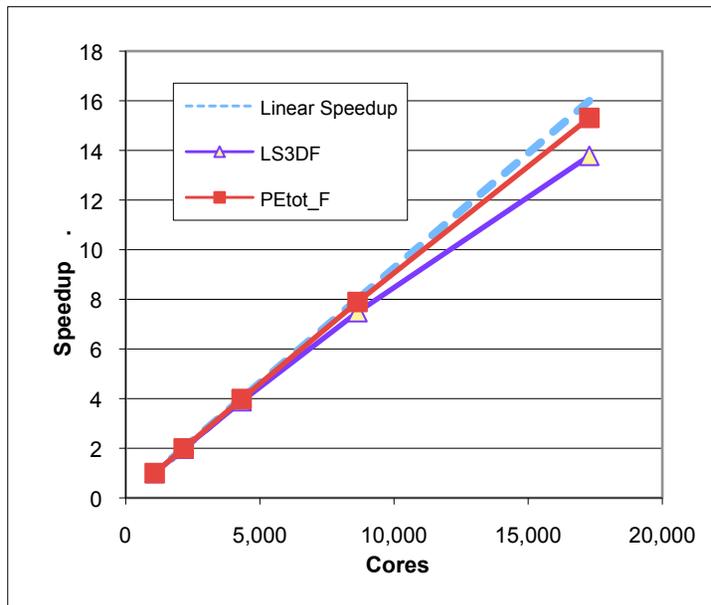


Time (second)



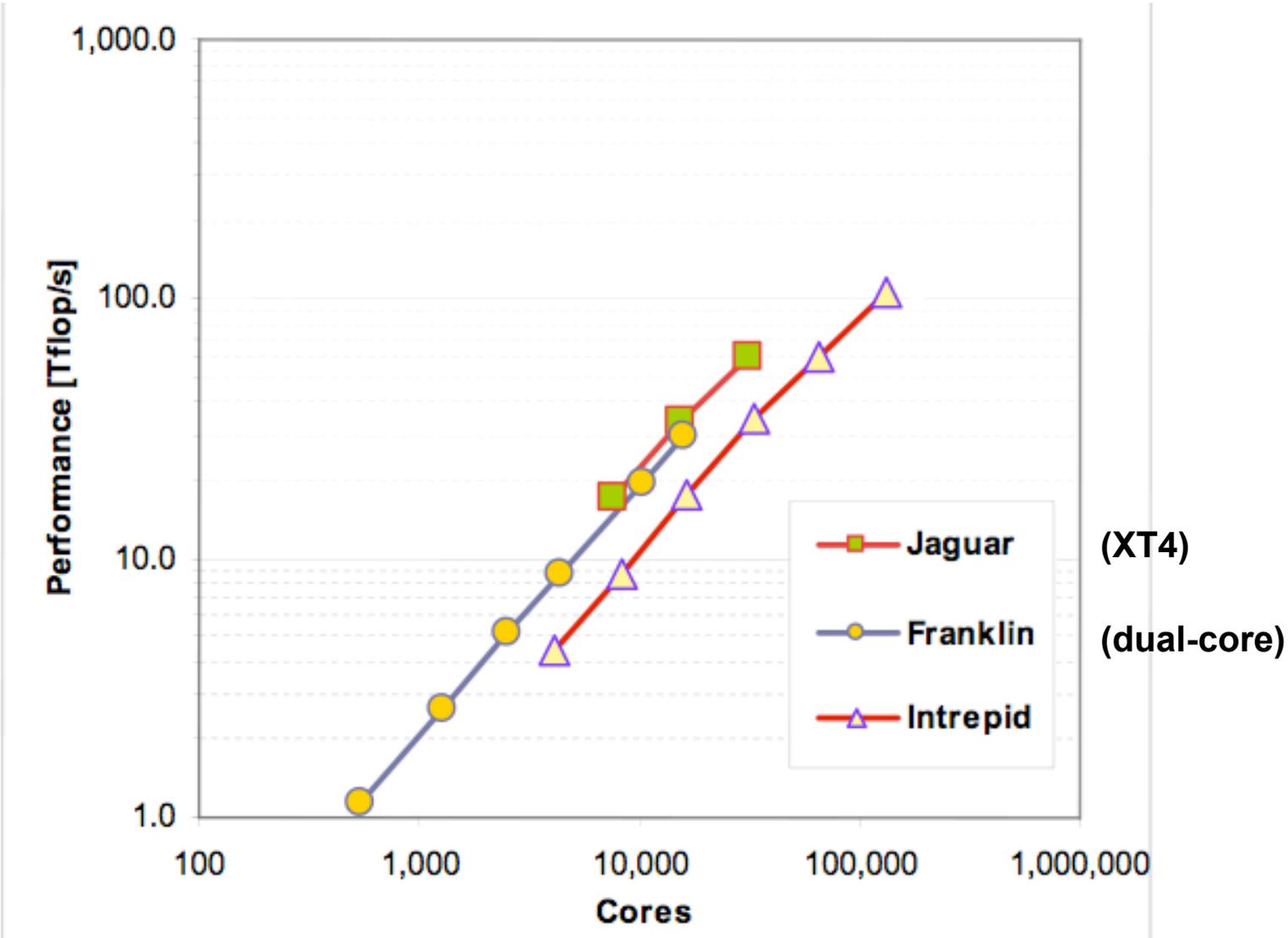
data movement

NERSC Franklin results (strong scaling)



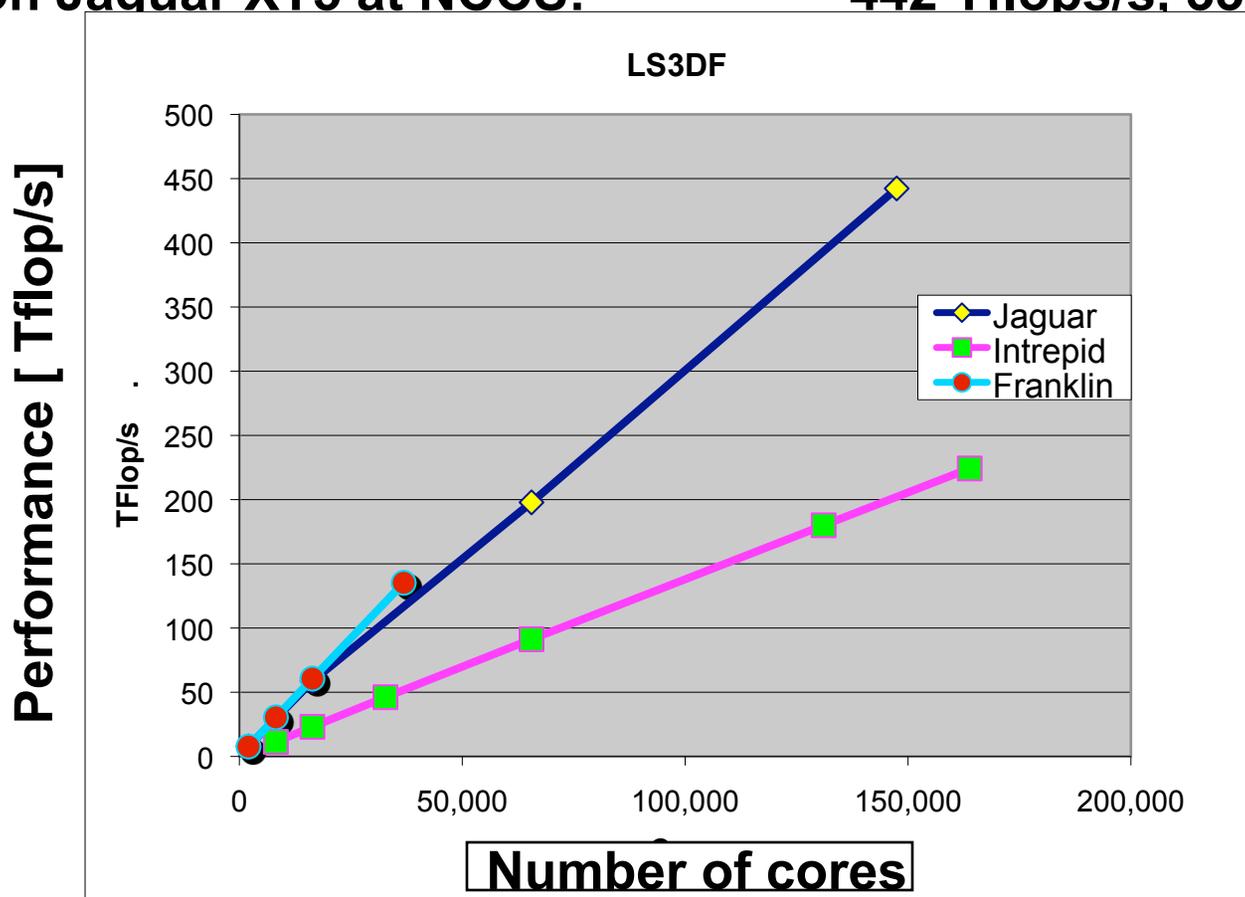
- ❖ 3456 atom system, 17280 cores:
 - one min. per SCF iteration, one hour for a converged result
- ❖ 13824 atom system, 17280 cores,
 - 3-4 min. per SCF iteration, 3 hours for a converged result
- ❖ LS3DF is 400 times faster than Petot on the 13824 atom system

Near perfect speedup across a wide variety of systems (weak scaling)



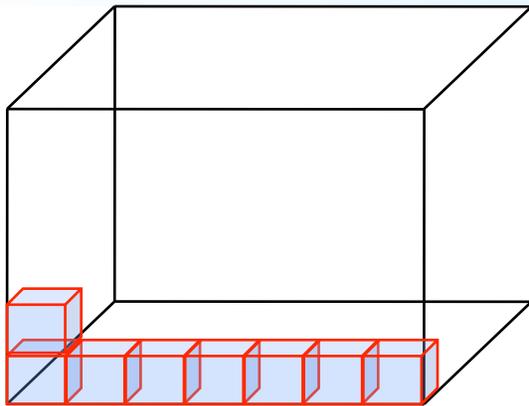
ZnTeO alloy weak scaling calculations

- First large scale run on Franklin at NERSC: 135 Tflops/s, 40% efficiency
- Subsequent runs on Intrepid at ALCF: 224 Tflops/s, 40% efficiency
- Final runs on Jaguar XT5 at NCCS: 442 Tflops/s, 33% efficiency



Note: Ecut = 60Ryd with *d* states, up to 36864 atoms

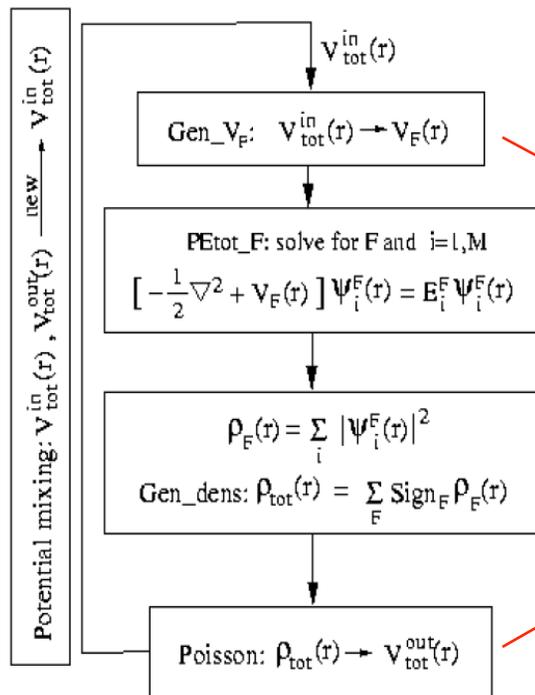
Node mapping and performance on BlueGene/P



Map all the groups into identical compact cubes, for good intra-group FFT communication, and inter-group load balance.

Time: 50% inside group FFT
50% inside group DGEM

LS3DF



Times on diff. parts of the code (sec)

| core | 8,192 | 32,768 | 163,840 |
|----------|-------|--------|---------|
| atom | 512 | 2048 | 10,240 |
| gen_VF | 0.08 | 0.08 | 0.23 |
| PEtot_F | 69.30 | 68.81 | 69.87 |
| gen_dens | 0.08 | 0.14 | 0.37 |
| Poisson | 0.12 | 0.22 | 0.76 |

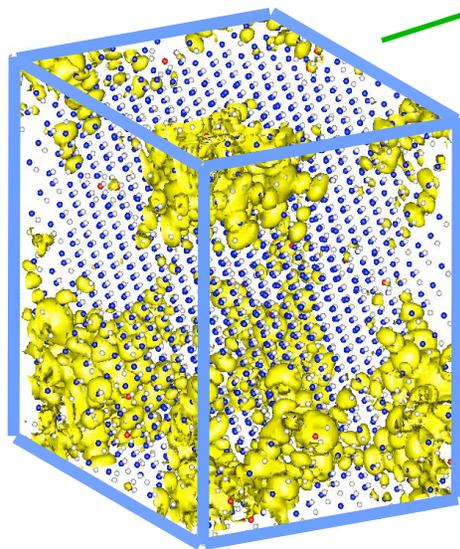
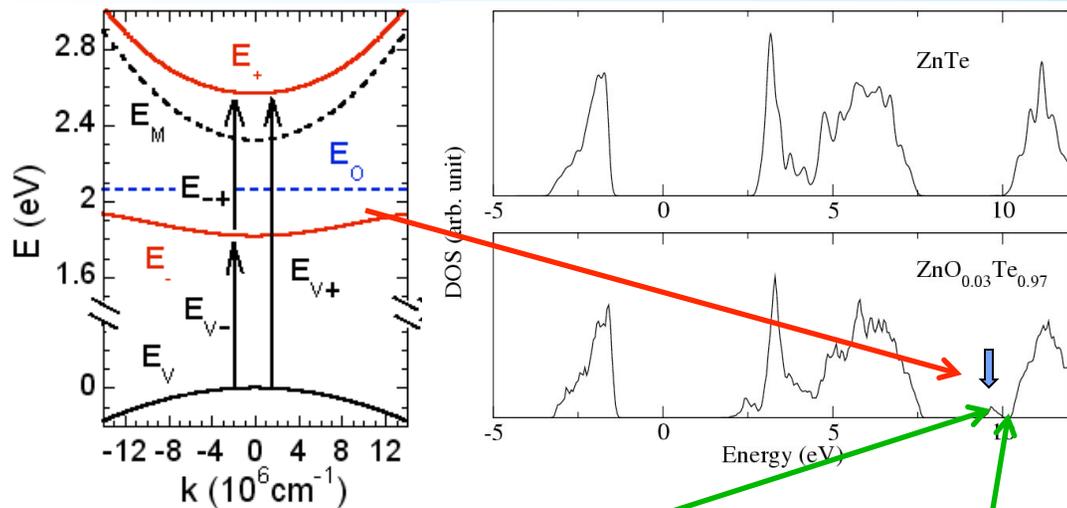
Perfect weak scaling

System Performance Summary

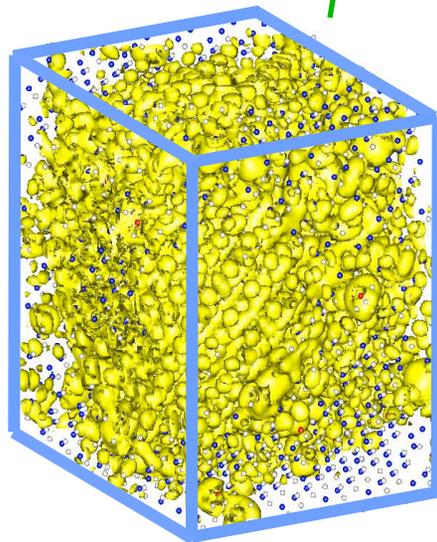


- ❖ 135 Tflops/s on 36,864 processors of the quad-core Cray XT4 Franklin at NERSC, 40% efficiency
- ❖ 224 Tflops/s on 163,840 processors of the BlueGene/P Intrepid at ALCF, 40% efficiency
- ❖ 442 Tflops/s on 147,456 processors of the Cray XT5 Jaguar at NCCS, 33% efficiency

Can one use an intermediate state to improve solar cell efficiency?



Highest O induced state

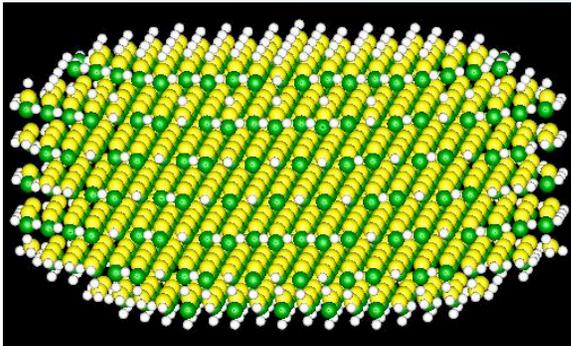


ZnTe bottom of cond. band state

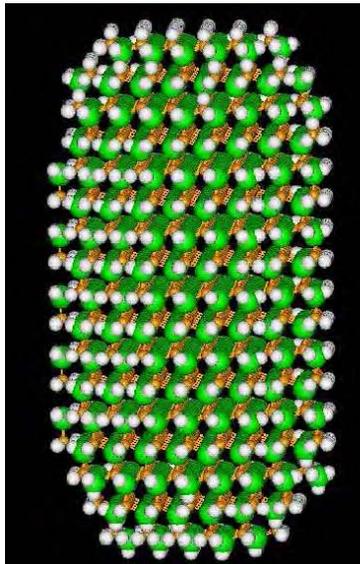
- ❖ Single band material theoretical PV efficiency is 30%
- ❖ With an intermediate state, the PV efficiency could be 60%
- ❖ One proposed material ZnTe:O
 - Is there really a gap?
 - Is there oscillator strength?
- ❖ LS3DF calculation for 3500 atom 3% O alloy [one hour on 17,000 cores of Franklin]
- ❖ Yes, there is a gap, and O induced states are very localized.

INCITE project, NERSC, NCCS.

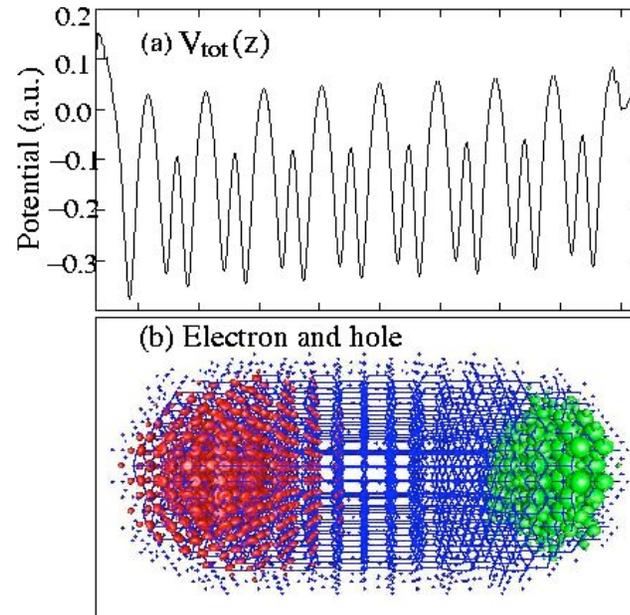
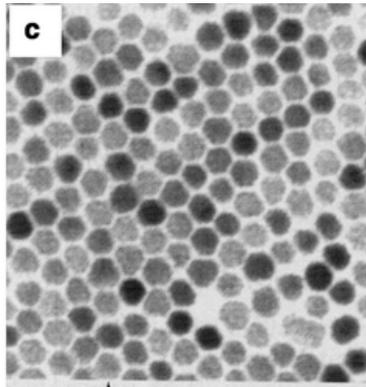
LS3DF computations yield dipole moments of nanorods and the effects on electrons



P = 30.3 Debye



P=73.3 Debye



$\text{Cd}_{714}\text{Se}_{724}$

WZ

- ❖ Equal volume nanorods can have different dipole moments
- ❖ The inequality comes from shape dependent self-screening
- ❖ Dipole moments depend on bulk and surface contributions
- ❖ Dipole moments can significantly change the electron and hole wave functions

INCITE project at NCCS and NERSC

COMPUTATIONAL RESEARCH DIVISION



Summary and Conclusions

- ❖ LS3DF scales linearly to over 160,000 processors. It reached 440 Tflop/s. It runs on different platforms with little retuning.
- ❖ For practical purposes, the numerical results are the same as a direct DFT based on an $O(N^3)$ algorithm, but at only $O(N)$ computational costs.
- ❖ LS3DF can be used to compute electronic structures for $>10,000$ atom systems with total energy and forces in 1-2 hours. It can be 1000 times faster than $O(N^3)$ direct DFT calculations.
- ❖ Enables us to yield new scientific results predicting the efficiency of a proposed new solar cell material.

Acknowledgements

- ❖ **National Energy Scientific Computing Center (NERSC)**
- ❖ **National Center for Computational Sciences (NCCS)**
(Jeff Larkin at Cray Inc)
- ❖ **Argonne Leadership Computing Facility (ALCF)**
(Katherine M Riley, William Scullin)
- ❖ **Innovative and Novel Computational Impact on Theory and Experiment (INCITE)**
- ❖ **SciDAC/PERI (Performance Engineering Research Institute)**
- ❖ **DOE/SC/Basic Energy Science (BES)**
DOE/SC/Advanced Scientific Computing Research (ASCR)

LS3DF Team



Lin-Wang Wang



Zhengji Zhao



Byounghak Lee



HongZhang Shan



Juan Meza



Erich Strohmaier

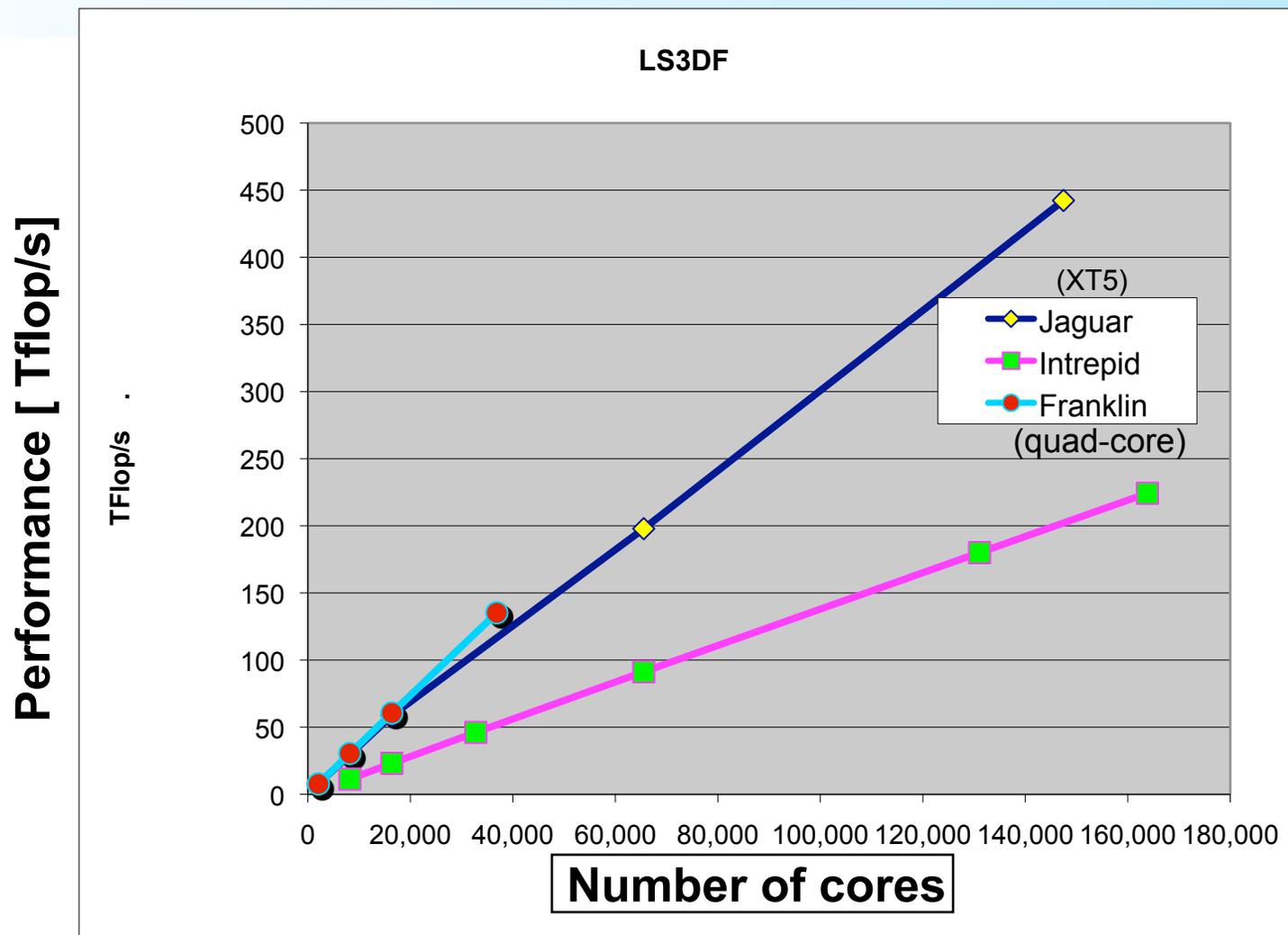


David Bailey

Backup Slides



ZnTeO alloy calculations (Ecut=60Ryd, with d states, up to 36864 atoms), weak scaling



Variational formalism of LS3DF

Original DFT formula

$$E_{tot} = \sum_i \int \psi_i^*(r) \left[-\frac{1}{2} \nabla^2 \right] \psi_i(r) dr + \int V_{ion}(r) \rho_{tot}(r) dr + \frac{1}{2} \int \frac{\rho_{tot}(r) \rho_{tot}(r')}{|r-r'|} dr dr' + \int \varepsilon_{xc}(\rho_{tot}(r)) dr$$

LS3DF formula

$$E_{tot} = \sum_F \alpha_F \sum_i \int \psi_{F,i}^*(r) \left[-\frac{1}{2} \nabla^2 \right] \psi_{F,i}(r) dr + \sum_F \alpha_F \int \Delta V_F(r) \rho_F(r) dr + \int V_{ion}(r) \rho_{tot}(r) dr + \frac{1}{2} \int \frac{\rho_{tot}(r) \rho_{tot}(r')}{|r-r'|} dr dr' + \int \varepsilon_{xc}(\rho_{tot}(r)) dr$$

❖ The fragment wave function $\psi_{F,i}(r)$ is defined within each Ω_F .

$$\int_{\Omega_F} \psi_{F,i}^*(r) \psi_{F,j}(r) dr = \delta_{i,j}. \quad \rho_F(r) = \sum_i |\psi_{F,i}(r)|^2 \quad \rho_{tot}(r) = \sum_F \alpha_F \rho_F(r),$$

Variational formalism of LS3DF

❖ Kohn-Sham equation of original DFT (N^3):

$$\left[-\frac{1}{2}\nabla^2 + V_{tot}(r)\right]\psi_i(r) = \varepsilon_i\psi_i(r)$$

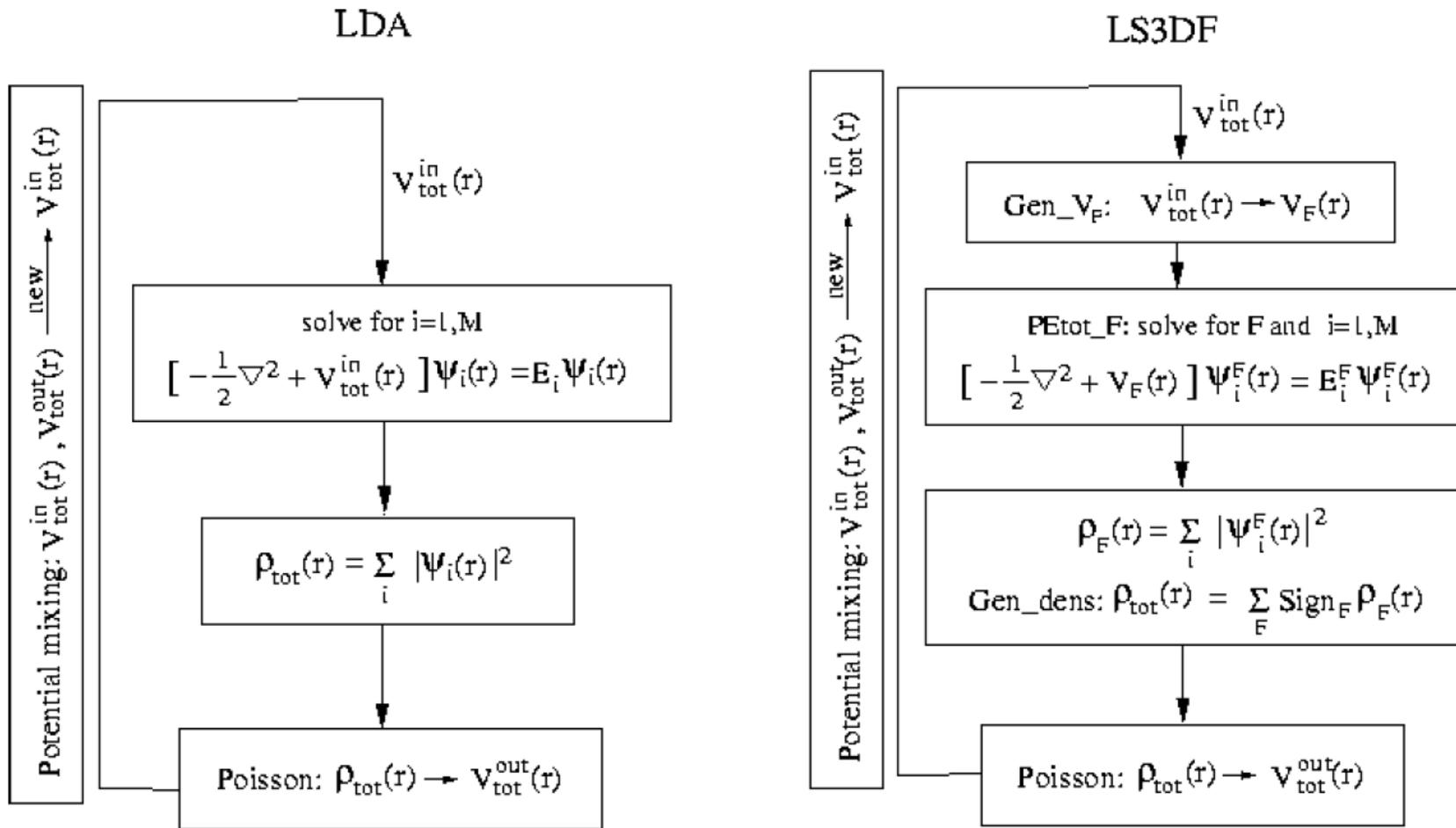
❖ Kohn-Sham equation of LS3DF :

$$\left[-\frac{1}{2}\nabla^2 + V_{tot}(r) + \Delta V_F(r)\right]\psi_{F,i}(r) = \varepsilon_{F,i}\psi_{F,i}(r) \quad \text{for } r \in \Omega_F$$

Where, $V_{tot}(r)$: usual LDA total potential calculated from $\rho_{tot}(r)$

$\Delta V_F(r)$: surface passivation potential

Flow chart for LS3DF method



Based on the plane wave PEtot code: <http://hpcrd.lbl.gov/~linwang/PEtot/PEtot.html>