SAND2014-15117PE

SNAP: Strong Scaling High Fidelity Molecular Dynamics Simulations on Leadership-Class Computing Platforms

Christian Trott
Simon Hammond
Aidan Thompson

Sandia National Laboratories is a multi-program laboratory managed and operated by Sandia Corporation, a wholly owned subsidiary of Lockheed Martin Corporation, for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-AC04-94AL85000.

The case for High Fidelity Simulations

- Most scientific relevant simulations: 10k – 10M atoms
- But typical MD simulations need ~50k atoms per GPU/Xeon Phi to be effective

=> Bad scaling beyond 10-100 nodes

- Use increased compute capabilities for higher accuracy

=> Close the gap to DFT

Sandia National Laboratories

SNAP Potential

- DFT like accuracy with lower cost and linear scaling
- Potential complexity can be increased systematically
- Parameters through fitting to DFT data
- Currently we have Silicon and Tantalum, working on InP
- Characteristics
  - 500k flops per interaction; 12M per atom
  - 300 kB temporary data per particle with calculations in flight
  - 30 kB temporary data per interaction with calculations in flight
  - Working on 27 interactions in parallel: ***1MB temporary data per atom***

Available Parallelism

```
Compute_SNAP() {
  for i in natoms() {                                   <- 100k atoms
    u_i = calc_U(i)
    Z_i = calc_Z(i,u_i)
    for j in neighbors(i) {                             <- 15.5 neighbors
      ∇_j u = calc_dUdR(i, j, u_i)
      ∇_j B^i = calc_dBdR(i, j, u_i, Z_i, ∇_j u)
      F_ij = −β · ∇_j B^i
      F_i += −F_ij; F_j += F_ij
} } }
```

Sets	per System	per Atom	per Interaction
i	1.0×10^{5}	N/A	N/A
i, j	1.6×10^{6}	15.5	N/A
$i, j, \eta_1, \eta_2, \eta$	1.0×10^{8}	992	64
$i, j, \eta_1, \eta_2, \eta, \mu, \mu'$	2.1×10^{9}	32,528	1,388
$i, j, \eta_1, \eta_2, \eta, \mu, \mu', \mu_1, \mu_1'$	2.4×10^{10}	2.4×10^{5}	15,183

```
Function Calc_dBdR(i, j) {
  for (η, η1, η2) in GetBispectrumIndices() {           <- 64 Coefficients
    ∇_j B_{η1,η2,η} = 0
    for (μ = 0; μ ≤ η; μ++ ) {                          <- 1-8 Iterations
      for (μ' = 0; μ' ≤ η; μ'++) {                      <- 1-8 Iterations
        ∇_j Z^{μ,μ'}_{η1,η2,η} = 0
        for (μ1 = max(0, μ + (η1 − η2 − η)/2);          <- 1-8 Iterations
             μ1 ≤ min(η1, μ + (η1 + η2 − η)/2); μ1++) {
          μ2 = μ − μ1
          for (μ1' = max(0, μ' + (η1 − η2 − η)/2);      <- 1-8 Iterations
               μ1' ≤ min(η1, μ' + (η1 + η2 − η)/2); μ1'++) {
        μ2' = μ' − μ1'
        ∇_j Z^{μ,μ'}_{η1,η2,η} += H^{η,μ,μ'}_{η1,μ1,μ1',η2,μ2,μ2'} (u^{η1}_{μ1,μ1'} ∇_j u^{η2}_{μ2,μ2'} + u^{η2}_{μ2,μ2'} ∇_j u^{η1}_{μ1,μ1'})
        } }
        ∇_j B_{η1,η2,η} += (u^{η}_{μ,μ'})* ∇_j Z^{μ,μ'}_{η1,η2,η} + Z^{μ,μ'}_{η1,η2,η} (∇_j u^{η}_{μ,μ'})*
} } } }
```

Target machines

- **Commodity CPUs**: Chama Intel Sandy Bridge, SNL
  - # 100 in Top 500 (Nov 2013)
  - 2 Sockets / 16 Cores per Node
  - Infiniband Interconnect
- **IBM Blue Gene Q**: Sequoia and Vulcan, LLNL
  - # 3 and #9 in Top 500 (Nov 2013)
  - 16 Cores / 64 Threads per Node
- **(NVIDIA) GPUs**: Titan, ORNL
  - # 2 in Top 500 (Nov 2013)
  - 16 Cores Opteron + NVIDIA Kepler K20x

System	Nodes	Cores	Threads	Power/Node	Perf./Node
Chama	1,230	19,680	39,360	368.8 W	332.8 GFLOP/s
Sequoia/Vulcan	122,880	2.0×10^6	7.9×10^6	80.26 W	204.8 GFLOP/s
Titan	18,688	5.0×10^7	5.4×10^8	439.3 W	1450.8 GFLOP/s

Within LAMMPS load-balancing limited to moving plains

Need local load-balancing to enable as little as 1 atom per node

Sandia National Laboratories

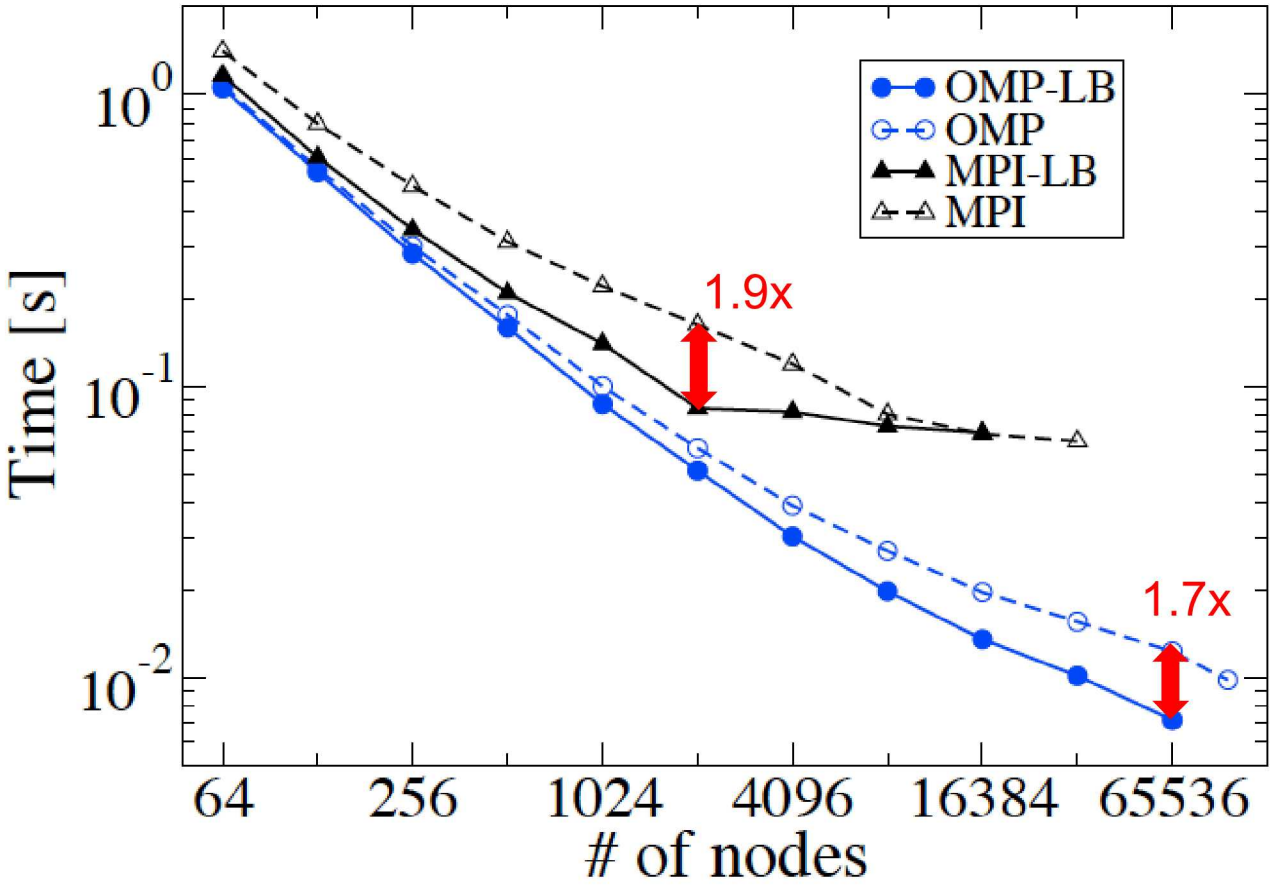
Micro Load-Balancing

Micro Load-Balancer:

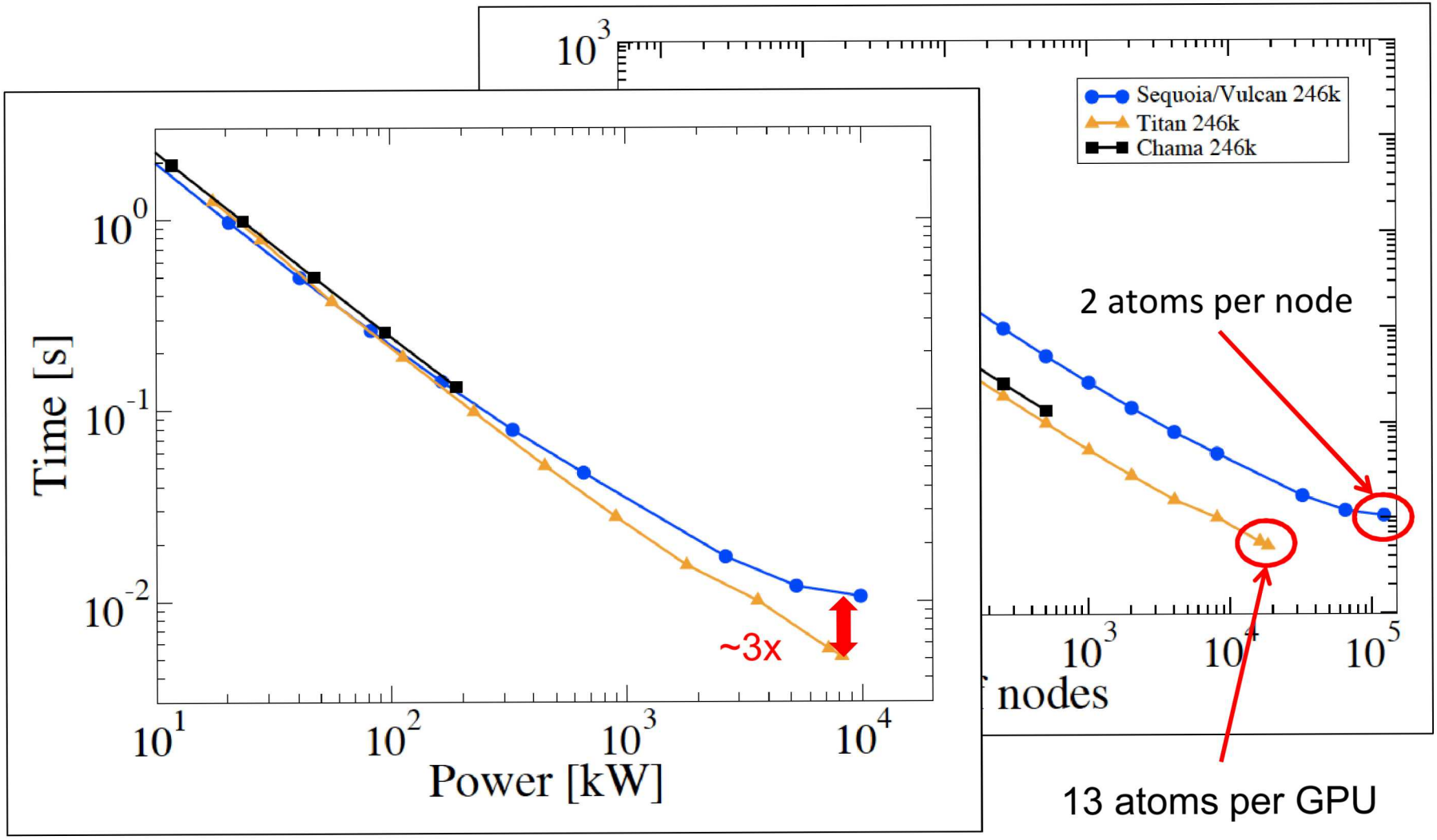
- increase communication cutoff by domain size
- reassign force calculation to neighboring process

Micro Load-Balancing 65k atoms BG/Q

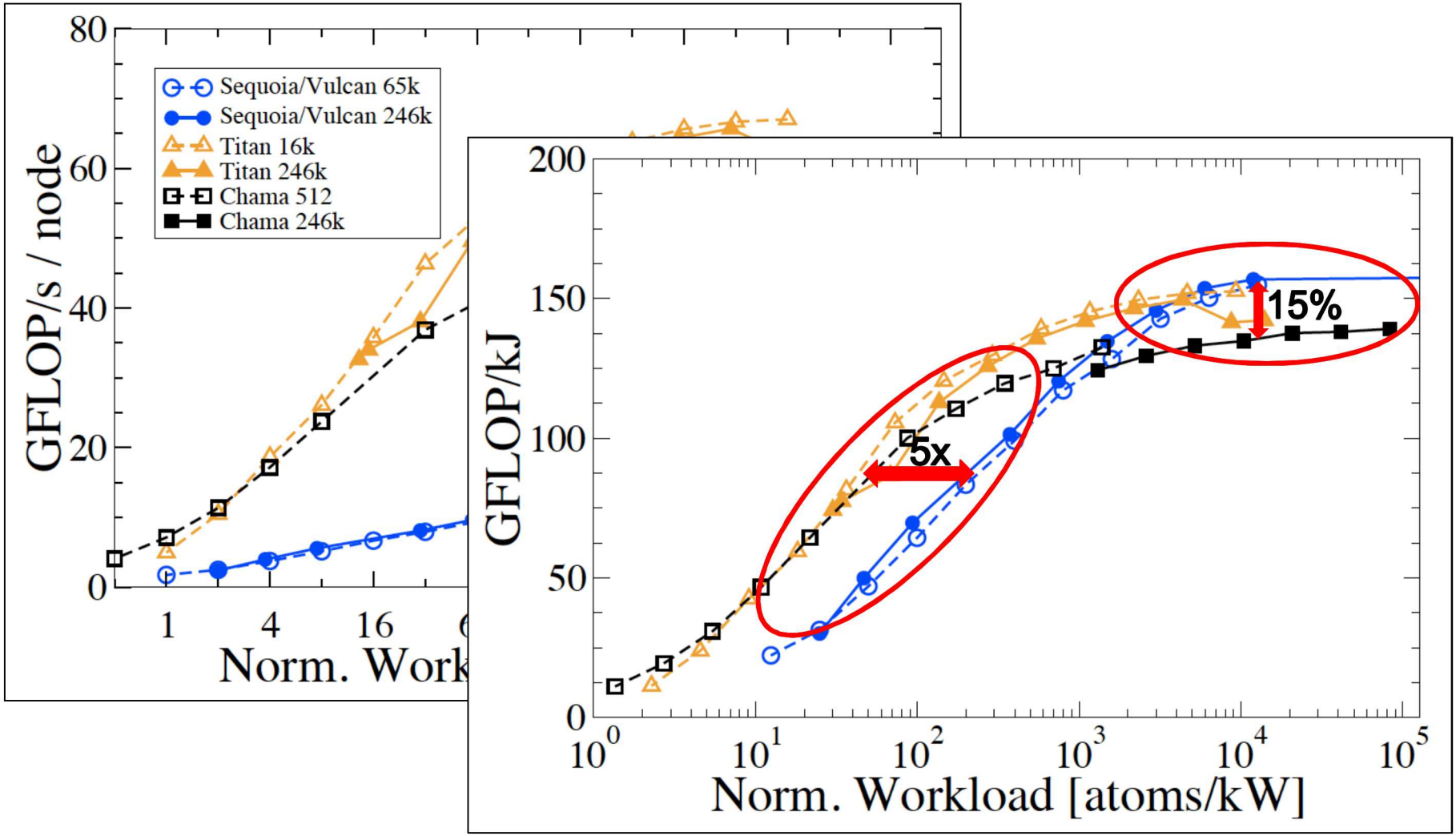
MPI – only (64 ranks per node), MPI+OMP (2x32 per node)



Strong Scaling I – Time per Step



Strong Scaling II – Normalized Performance

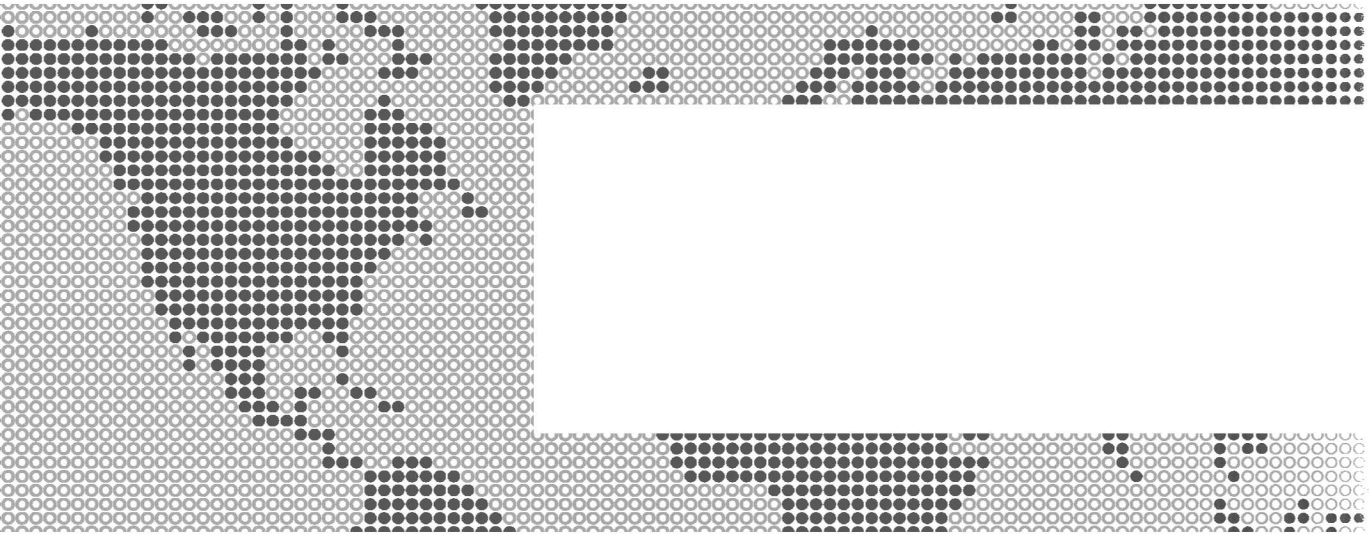


Sandia National Laboratories

Conclusions

- High Fidelity Molecular Dynamics Simulations can use PetaScale HPC systems efficiently
- MPI – only is not enough for extreme strongscaling
- Micro load-balancing can help at very low workloads
- All system architectures exhibit similar power efficiency at large workloads
- For small workloads fewer, more powerful nodes are more efficient than many smaller nodes

=> Current trend of increasing node performance is good



Questions and further discussion: crtrott@sandia.gov