[BUILD A CELL WORKSHOP 5 – MIT – IN SILICO GROUP NOTES]
[PLEASE DON'T EDIT HISTORICAL RECORD BELOW]

- **In silico cell:** Software for modeling and simulation of synthetic cell metabolism, transport, signaling and circuits
    - Expected outcome (Aug 2018): list of available codes and protocols, list of needs for inputs and wishes for outputs, start dedicated Git. Put together a v0.1 simulation of a synthetic cell that has whatever functionality we can muster.

Lead: Drew Endy

- Group Members: Akshay Maheshwari, Miro Gasparek, Francis Lee, Christian, Bryan Bartley

Questions:

a) **How much computer power is needed to model the entire cell with reasonably reliable physics?** How much power is needed if we are willing to compromise?

Methods: We did a back-of-the-envelope calculation for how much computational power is needed to represent a bacterium for 1 s or for approx. 20 min. at either Brownian or Stokes dynamics, using the rough estimates of the molecular counts, simulation timescales, and computing power available.

The calculations will be

Conclusion: If we use the quantum computer with around 50 qubits, we can simulate a cell in real time at the level of the Brownian Dynamics for one second (2E10 CPU seconds), we need 20-fold improvement in power to get real-time cytoplasm simulation.

(IBM: 8 qubit??)

(Google: 72 qubit)

	Brownian Dynamics	Stokes Dynamics	
~ 1sec	2E4 CPU months	2E7 CPU months	
~ 1200sec	2E7 CPU months	2E10 CPU months	

b) What is the real-time whole cell model good for?

i) Francis - WCM and any models that we make are more compelling when they're validated, tested and improved. Our efforts should align tightly with the interlab and planning groups.

Reference Papers:

- Feig, Michael, and Yuji Sugita. "Whole-Cell Models and Simulations in Molecular Detail." *Annual review of cell and developmental biology* 35 (2019).
- Goldberg, Arthur P., et al. "Emerging whole-cell modeling principles and methods." *Current opinion in biotechnology* 51 (2018): 97-102.

Modeling group: How much computer power is needed to model the entire cell with reasonably reliable physics in reasonable time?

Contact: Akshay Maheshwari (akshaym@stanford.edu)

Real time simulation of the cytoplasm of an entire cell is a grand challenge in building a cell. In the worst case, simulations with accuracy at the scale of molecular physics will be needed to generate useful prediction for whole cell behavior (e.g., ODE-only modeling may not be enough). **Below we produce back-of-the-envelope estimates for the minimal compute needed** for this worst-case scenario, with a given assumption that computation distribution is tractable.

Small prokaryotic organisms such as *Mycoplasma* and *E. coli* have on order 1 to 10 million proteins per cell [Schmidt et al. 2016; Milo 2013]. Physically faithful representation of all these proteins (not accounting for other molecules in the cell, e.g., lipids, metabolites, and nucleic acids) requires, in the most detailed case, atomistic molecular dynamics simulations, and in the most coarse-grained case, brownian dynamics of proteins approximated as spheres. Stokesian dynamics simulations that capture n-body interactions between proteins (e.g., hydrodynamic or electrostatic) bridge these two physical modeling regimes.

Based on benchmarks using the Smoldyn Brownian Dynamics simulation software package we estimate that simulating the Brownian motion of ~$2e3$ proteins (in a representative ~1/1000 scale ~100nm sided voxel of cytoplasm) for 1 second takes approximately 2e4 seconds using AWS T3.nano (3.1 GHz Intel Xeon Platinum 8175, 2 vCPUs, 500 MB RAM), i.e. $4e4$ cpu-seconds. Approximating linear scaling of simulation time with number of molecules simulated, simulating a whole cell with ~$2e6$ proteins would then roughly take $2e7$ seconds and $4e7$ cpu-seconds. This corresponds to **~ 3.7e17 FLOPS needed to simulate an entire *E. coli* in real time** ($3.1e9$ cycles/second * 3 flops/cycle *$4e7$ cpu-seconds)

Molecular dynamics simulations using modern Monte Carlo techniques can push simulations of 1-10M particles (estimated number of proteins in a cell) in 160,000 vCPU hours across 10,000 machines with 64GB (for 1 time step or 1 nanosecond) [Kmiecik,Jung]. We can estimate the GFLOPs executed for this computation to be ~$6e15$ FLOPS.

Assuming this computation is distributed across the cloud (e.g. - vCPUs on AWS), we can estimate the cost of cloud resources that support real time whole cell models to be around $7000 per time step using m5.4xlarge EC2 instance. To simulate a whole cell division (roughly 24 minutes for *E. coli*), we would spend around **$1e16 USD on AWS** for essentially $5e5$ CPU/s of compute resources, given the current AWS EC2 pricing model.

In 2015, there were approximately 2e20 FLOPS of computational power available worldwide [link]. As a more practical estimate, we note that the fastest computer system as of 2018, the **Folding@home** distributed computing system, which leveraged computation worldwide, had a computational power of **1.35e17 FLOPS**. **These estimates imply that it may currently be possible to simulate a whole-cell in real-time** (with sub-volume represented in parallel) using the most coarse-grained physical representation assuming we are able to efficiently distribute the computation.

Future work needed:

1. Evaluate computation needed for simulating physics/modeling cells with varying levels of accuracy (e.g., Stokesian dynamics or ODE-based modeling, e.g. whole cell model) and analyze the trade-off between the accuracy and the computational costs.
2. Consider how other forms of computation (e.g., quantum computing) may be leveraged.
3. Propose the meaningful practical applications of the whole-cell modelling in the context of building a synthetic cell?
4. Set up infrastructure to leverage such levels of computation?
5. Cross check with estimates from Illinois
6. Add quantum estimates

References

1. Milo, Ron. "What is the total number of protein molecules per cell volume? A call to rethink some published values." *Bioessays* 35.12 (2013): 1050-1055.
2. Kmiecik, Sebastian, et al. "Coarse-grained protein models and their applications." *Chemical reviews* 116.14 (2016): 7898-7936.
3. Jung, Jaewoon, et al. "GENESIS: a hybrid-parallel and multi-scale molecular dynamics simulator with enhanced sampling algorithms for biomolecular and cellular simulations." *Wiley Interdisciplinary Reviews: Computational Molecular Science*5.4 (2015): 310-323.

GFlops = (**CPU** speed in GHz) x (number of **CPU** cores) x (**CPU**instruction per cycle) x (number of **CPUs** per node)

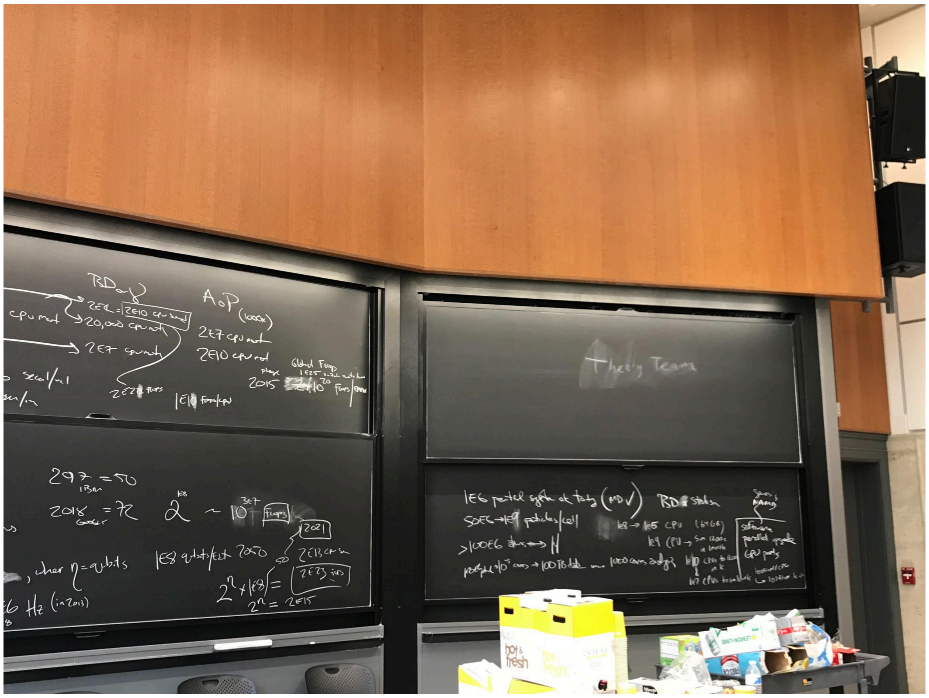
CPU speed in GHz = 3.1 (3.1 GHz Intel Xeon® Platinum 8175 processors (m4.xlarge))

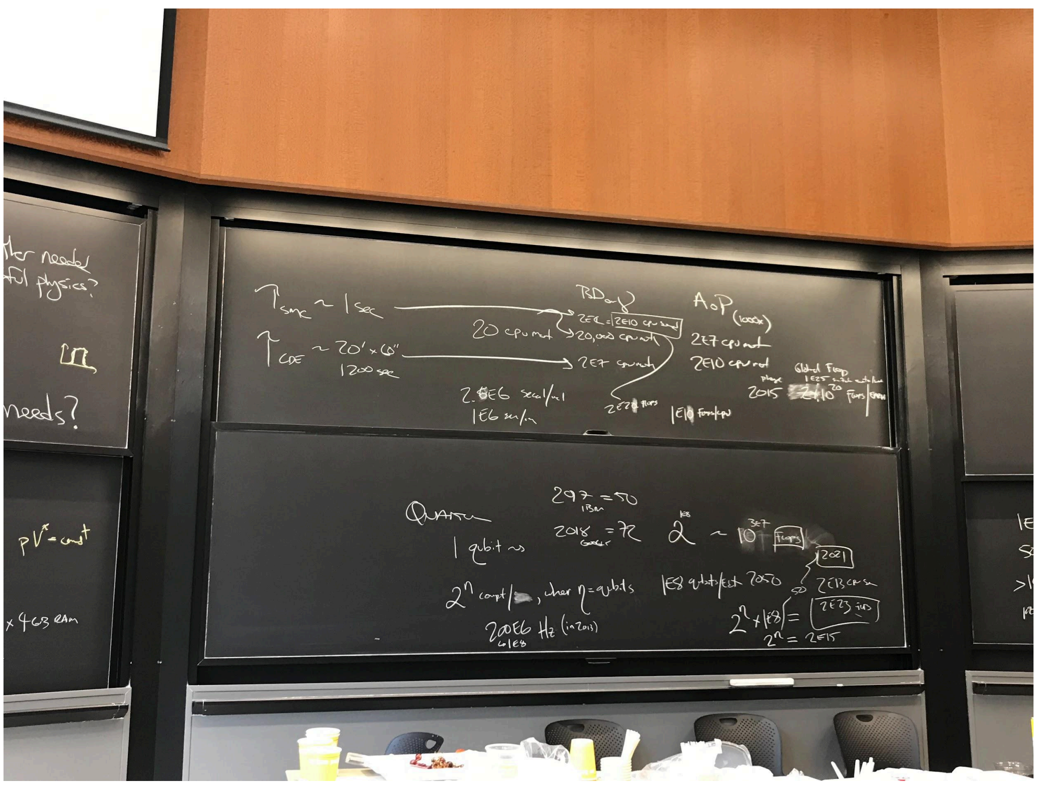
Number of CPU cores = 16

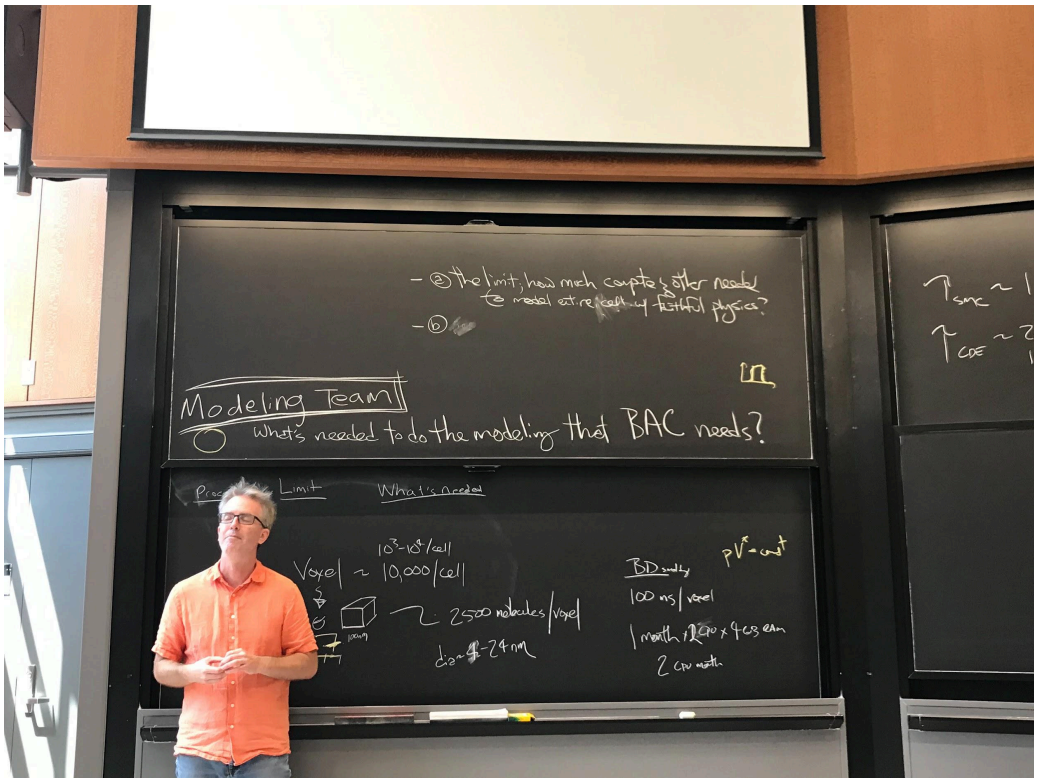
CPU instruction per cycle

(note: we assume computation is distributable)

Figures below: some details of the computation...




AoP (1000x)
2E7 cpu mod
2E10 cpu mod
1E6 sec/n
2018 = 72
2015
2021
2^n = 2E15


Modeling Team
What's needed to do the modeling that BAC needs?
Limit
What's needed
Voxel ~ 10,000/cell
2500 molecules/voxel
100 ms/voxel
2 CPU month