Using High Performance Computers for Modeling, Simulation, and Analysis

David Meyer
Statistical Solutions LLC
831-233-8438
David.statistical.solutions@gmail.com
www.statisticalsolutionsllc.com

David.statistical.solutions@gmail.com
Bottom Line (My Problem)

Limited budgets, work backlogs, & real-world needs incentivize:
1. An understanding *ahead of time* what the system will do
2. Experimentation
3. Multiple approaches/solutions
4. Some sort of optimization

**Modeling, simulation, and analysis (MS & A) software can address these, but:**
1. Typically with extensive scenario run-time
2. Overwhelming amounts of (indirect) data
3. Some analyses (optimization, stochastic modeling, sensitivity analysis, etc.,) requires many runs
4. Deliverables and students are calendar based, not event based
Meyer’s Theorem

When given choices of hardware and software approaches to solving problems, we always select what we have or are familiar with over what is optimal.

Corollary:
The more choices you have, the more likely it is that you pick the wrong one.
There is a Fork in the Road!

- The real choice is Windows vs. UNIX (Mac runs on a Unixish OS)
The Goal

Selection of the best combination of:

1. Software
2. Hardware
3. Operating System
4. The degrees of freedom you have on the above

How is “Best” defined?

1. Fastest?
2. Most reliable?
3. Easiest to use?
4. Cheapest?
5. etc
Operating Systems

1. Why Unix?
   a) Why it does batch processing!
   b) It is your fastest standard OS
   c) Been around for a long time, good documentation, very stable...
   d) Some batch processing/data mining/data processing programs are written explicitly for UNIX/LINUX
   e) If you want to run on an HPC you have to use UNIX
   f) There may be 100 UNIX commands
      i. I probably know 15
      ii. Use 5

2. Why Windows?
   a) Because Bill Gates rules the world
   b) Like it or not, this is a Windows dominated world, and much software is written only for Windows
   c) It is also perhaps what you are most comfortable with
   d) You can do batch processing on Windows platforms too

3. Maybe you have no choice. XXXX is the only tool, which means you are stuck with a Windows OS
Things I Thought...

Hardware

1. Its GAMS vs. Excel, or JMP vs MATLAB or...
2. Actually, the decision point occurs earlier: Are your approaches hardware limited?
3. Does your problem require parallel processing to solve?
   a) Good desktop, but if real speed boost is needed:
   b) Cluster – possibly influences your software choice
   c) Supercomputer – influences your choice of software, robust
4. What are your constraints?
   a) Heck will freeze over before we buy another computer around here
   b) I have access to a variety of platforms
   c) I have access to a supercomputer
   d) There are hybrid solutions...
Other Considerations

1. Clusters and supercomputers
   a) What are the considerations of a cluster?
      i. Can make Unix or Windows based
      ii. Can get a lot of tech support as part of the purchase
      iii. You may become this person

   “Only person who knows Unix and/or clusters” dude
b) So what is a cluster anyway?
   i. A cluster is essentially many computers all wired together
   ii. 1 Master blade that farms out the work
   iii. Many slaves that run the program and solve the problem
   iv. Most likely do not divvy up a single problem
   v. Speed comes from running multiple instances of the same problem (all sensitivity analysis done simultaneously)
   vi. Your laptop is likely as fast as a single cluster blade
   vii. Rack mounted, too many blades impose cooling or weight issues
c) Just how “super” is a super computer?
   i. Guaranteed to not run a Windows program
   ii. Might not run a LINUX program or even a UNIX program
   iii. Not all that different than a cluster, just bigger, needs more support, OS designed from the beginning for special use
   iv. Some software isn’t designed to be parallelized or can’t be parallelized, but you can do the pleasingly parallelized problems that clusters can do
   v. A notional supercomputer may have 4600 nodes, 16 cores per node. ~36000 potential simultaneous instances
   vi. It’s not yours – so you have to share!
      a. Admin can take months
      b. You get put in a que that is job size dependent
      c. If your estimate is off...
      d. Their priority is level loading, not getting you in and out
      e. They do have a limited budget for licenses, but,

d) Many Venders have a 1 node/1 license policy!
Other Options

e) So what is cloud computing?
   i. Something some organizations are reluctant to use because of the sensitivity of data to hacking/theft
   ii. A way to expand your OS, software, hardware options
   iii. A service that may allow for payment for use instead of just payment
   iv. A means to foist all that horrible IT “stuff’ on to someone else
   v. A great way to incur recurring costs
   vi. Something that limits your software choices once selected

f) Keep in mind that programming is programming

g) Stack Overflow!
Summary

Hardware/software/OS decisions are influenced by:

1. Cost
2. Need for parallelization (speed)
3. Capability of software
4. Comfort with and accessibility to software
5. Cost/business models
6. Code efficiency

It’s the probabilistic output, stupid!
Questions?