

The Effects of Clouds & Accelerators on the HPC Production Process

Presentation for Society of High Performance Computing Professionals

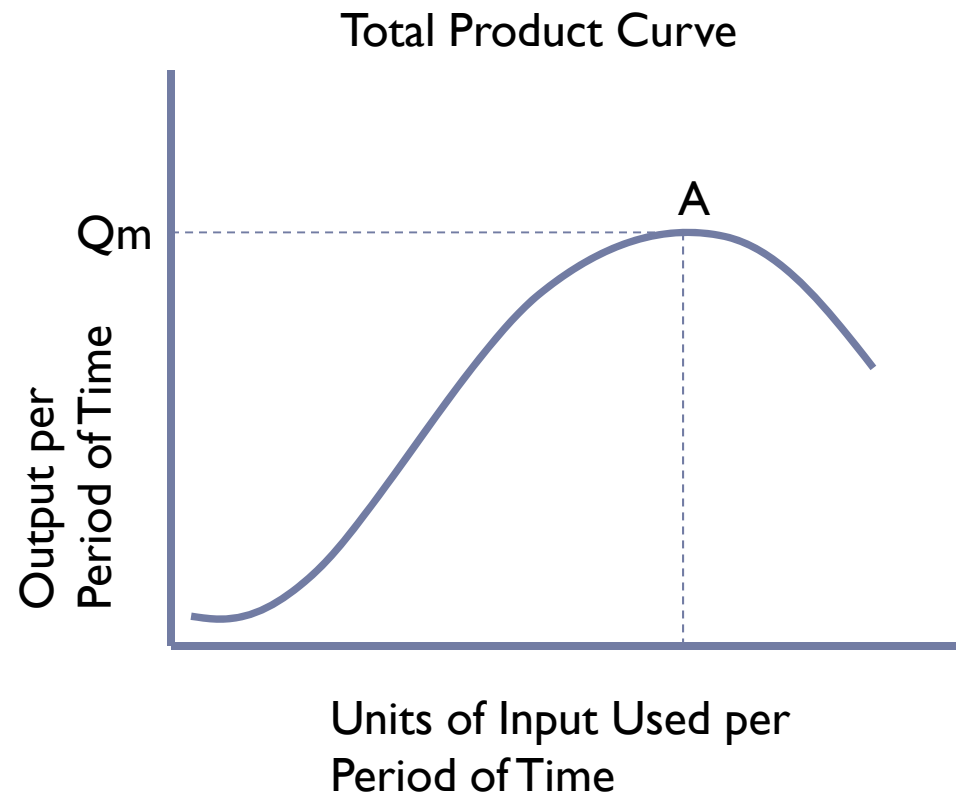
Steve Hebert, Nimbix LLC

Topic Outline

- ▶ Definition Production Process
- ▶ Production Inputs in HPC Datacenters
- ▶ Accelerators
- ▶ Cloud
- ▶ Barriers to Adoption
- ▶ Conclusion

Production

- ▶ The economic process of converting inputs to outputs
- ▶ Production uses resources to produce goods for exchange
- ▶ Measured as the rate of output per period of time

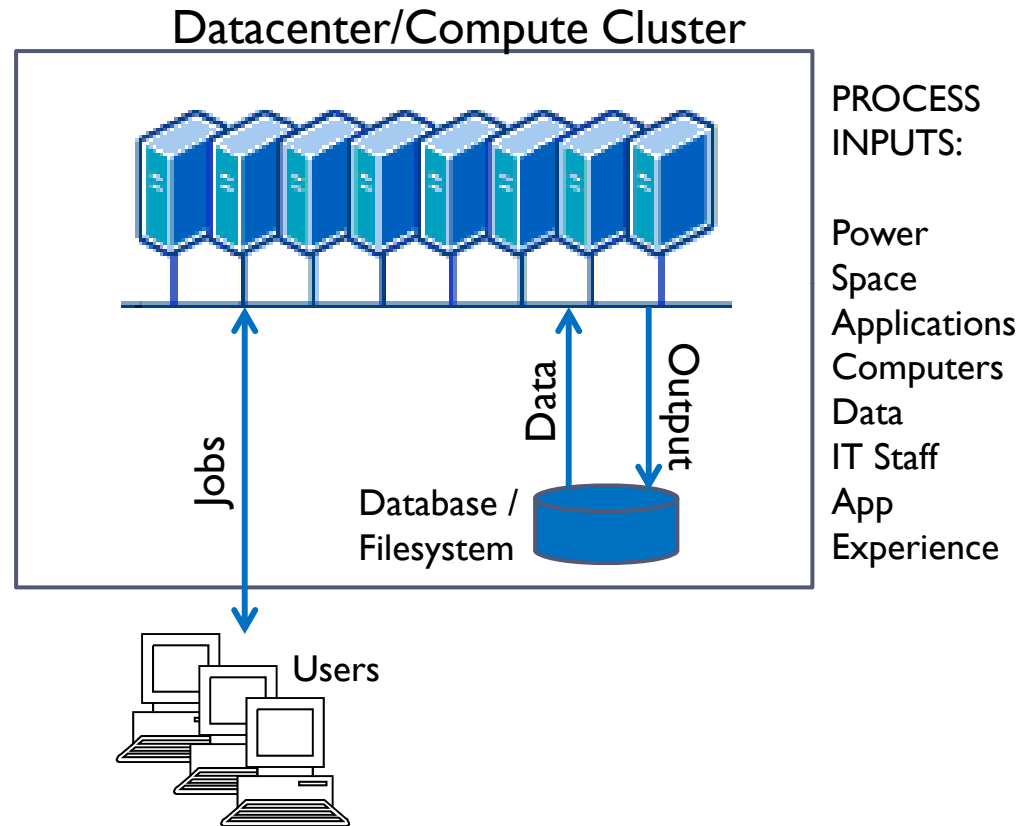


Factors of Production

- ▶ Resources used in the production process
 - ▶ In economics: Land, Labor, Capital Stock, Human Capital
- ▶ Fixed Factors of Production – Factors that are not easily changed over the short run
- ▶ Variable Factors of Production – Factors whose usage rate can be changed easily
 - ▶ Over the long run, all factors can be changed

The HPC Production Process

- ▶ Methods of combining inputs: Technology
- ▶ What are process goals? Reduce Costs? Increase productivity? Increase throughput?
- ▶ Assumption – In HPC implementations, many factors are fixed over the short run

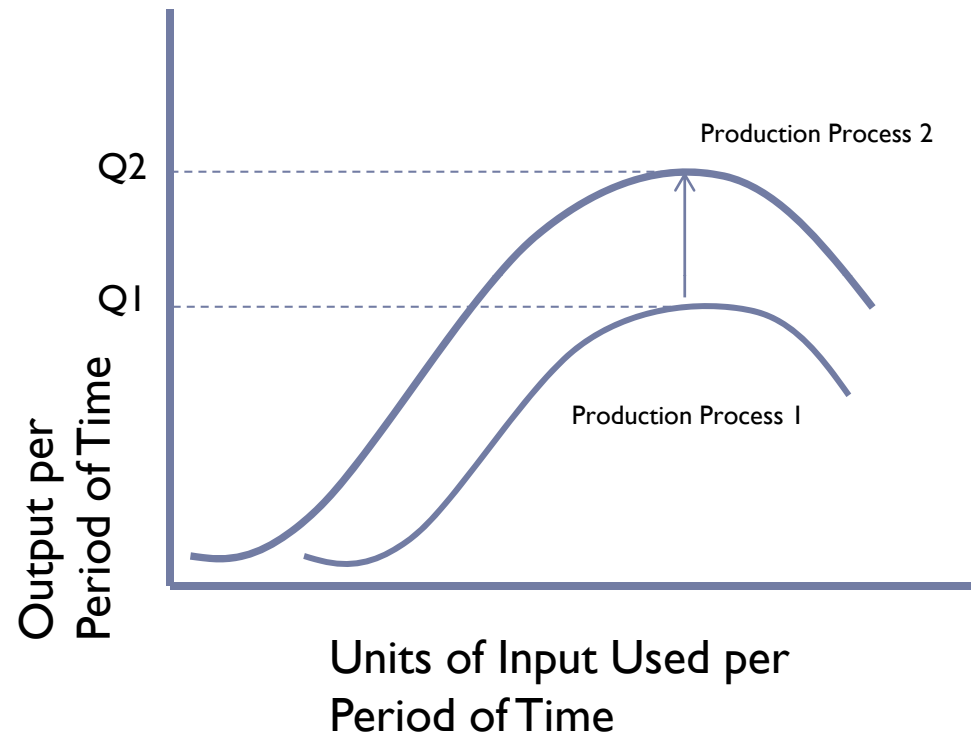


HPC Production Process Inputs

- ▶ Applications (Machinery)
 - ▶ Technology Costs: Internal Development / Licensed
- ▶ Computer Systems (Machinery)
- ▶ Data (Raw Material)
 - ▶ Workload Inputs
- ▶ Datacenter space (Land)
- ▶ HPC/Application Expertise (Human Capital)
- ▶ IT Staff (Labor)
- ▶ Power (Raw Material)

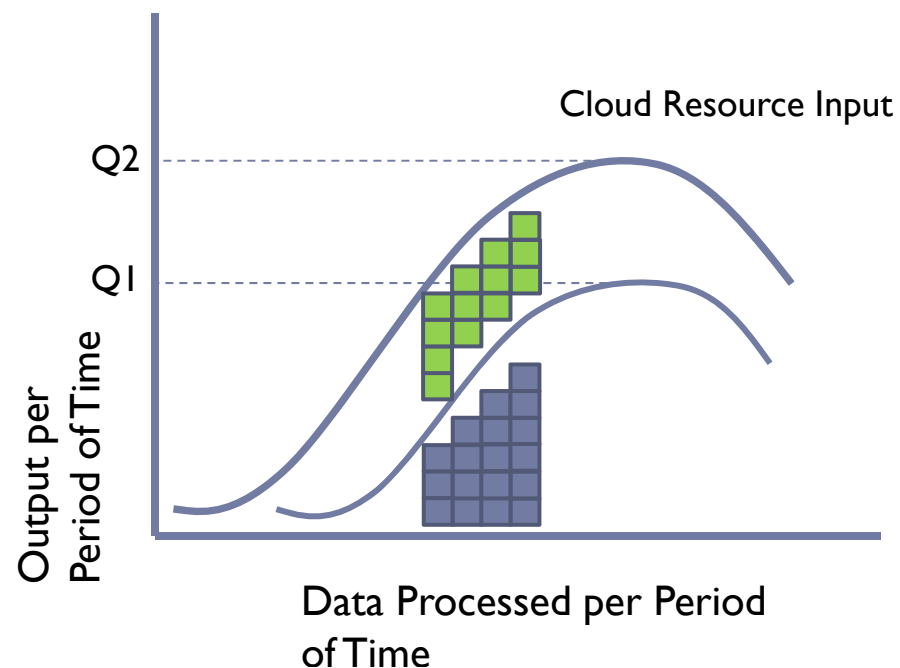
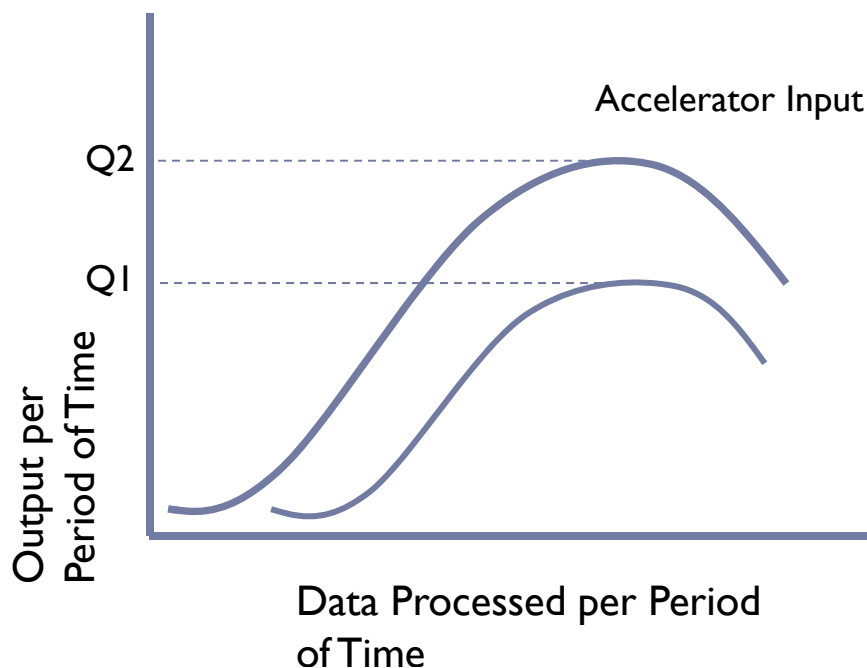
The Challenge of Increasing Production Output

- ▶ **Scale-out requires investment in:**
 - ▶ Datacenter/Colo Space
 - ▶ Capital Equipment
 - ▶ Power/Cooling
- ▶ **Deployment of new technologies**
 - ▶ New Algorithms
 - ▶ Application Investment
 - ▶ Network Infrastructure
 - ▶ Computer Architecture



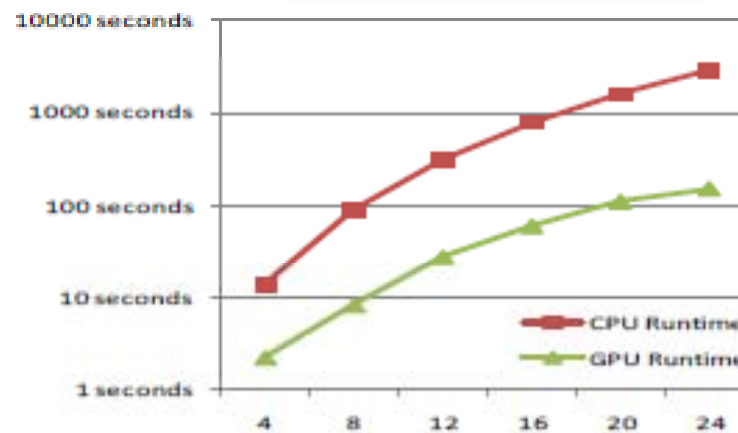
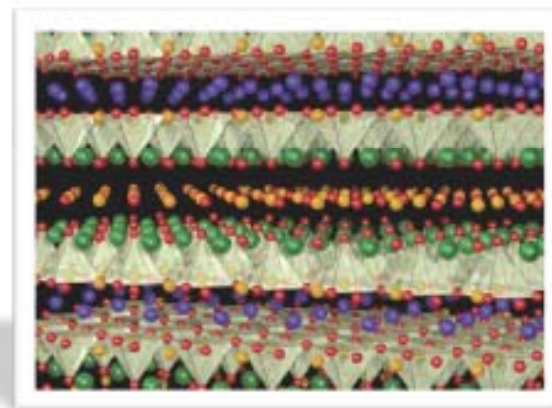
Accelerators & Cloud Resources as Inputs

- ▶ **Assumptions:**
 - ▶ Accelerated platforms reduce runtimes
 - ▶ Bandwidth sufficient for data transport to cloud resources
- ▶ **Additional Input Cost Considerations**
 - ▶ Lower energy requirements & less physical space



Accelerator Technology Decreases Runtimes

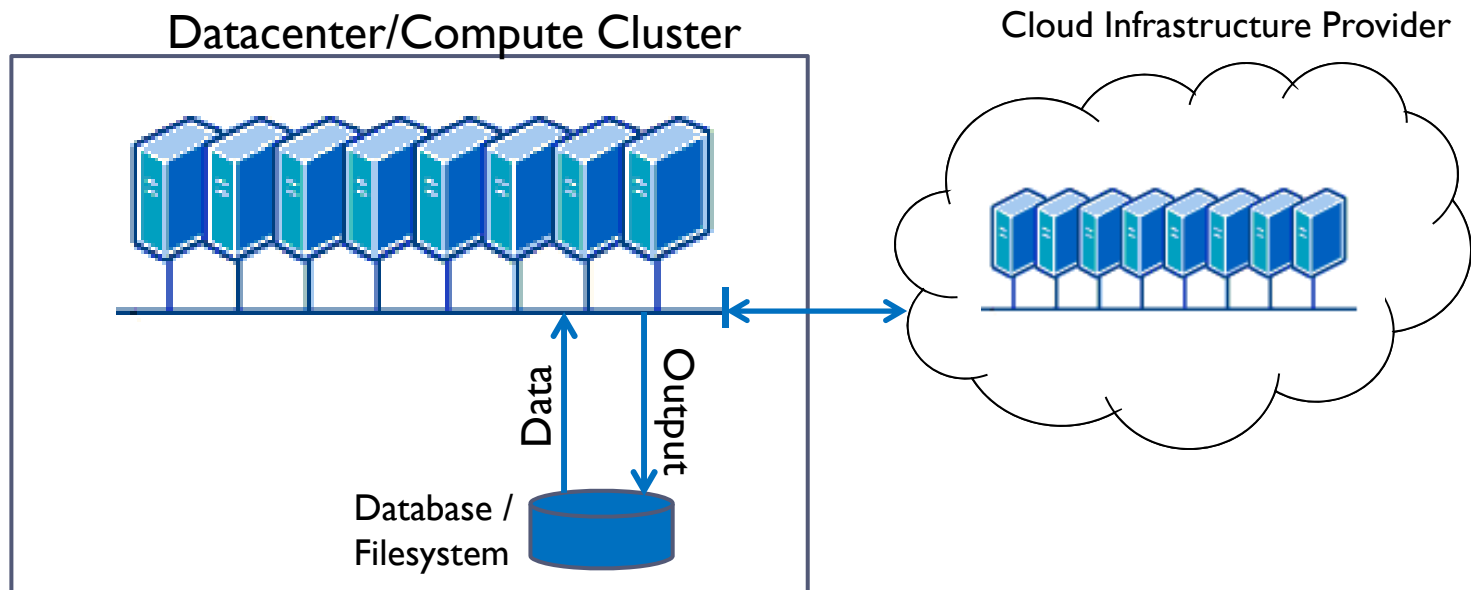
- ▶ Quantum Monte Carlo simulation
 - ▶ High-temperature superconductivity and other materials science
 - ▶ 2008 Gordon Bell Prize
- ▶ GPU acceleration speedup of 19x in main QMC Update routine
 - ▶ Single precision for CPU and GPU
 - ▶ Required detailed accuracy study and mixed precision port of app
- ▶ Full parallel app is 5x faster, start to finish on a GPU-enabled cluster on Tesla T10



GPU study: J.S. Meredith, G. Alvarez, T.A. Maier, T.C. Schulthess, J.S. Vetter, "Accuracy and Performance of Graphics Processors: A Quantum Monte Carlo Application Case Study", *Parallel Comput.*, 35(3):151-63, 2009.
Accuracy study: G. Alvarez, M.S. Summers, D.E. Maxwell, M. Eisenbach, J.S. Meredith, J. M. Larkin, J. Levesque, T. A. Maier, P.R.C. Kent, E.F. D'Azevedo, T.C. Schulthess, "New algorithm to enable 400+ TFlop/s sustained performance in simulations of disorder effects in high-Tc superconductors", *SuperComputing*, 2008. [Gordon Bell Prize winner]

Cloud Technology Impacts Short-Run Fixed Factors

- ▶ Scenario: Requirement for workload volume exceeds maximum compute capacity
 - ▶ Datacenter and computer infrastructure a “fixed input” in the short run
- ▶ External cloud resources become option for extending compute capacity



Barriers to Adoption

- ▶ **Availability of accelerated application source code**
 - ▶ Requires additional investment in code development or application acquisition
- ▶ **Capital for accelerator hardware**
 - ▶ Assessing technology risk / cost
- ▶ **Data transport costs & bandwidth constraints in/out of datacenter**
- ▶ **Data security**
- ▶ **Workload management / scheduling**
 - ▶ Difficult for both heterogeneous HW and cloud implementations

Overcoming Barriers

- ▶ **Cloud / on-demand business models may enable experimentation with processes**
 - ▶ Lower the risk to evaluate state of the art hardware
 - ▶ Help develop appropriate security measures
 - ▶ Smooth out infrastructure investment
 - ▶ Benchmark alternative technologies
- ▶ **New tools for heterogeneous computing rapidly evolving**
- ▶ **Creative carrier solutions may overcome bandwidth constraints?**

Evaluating Process Inputs

- ▶ HPC Production Process: Quantum Monte Carlo Simulation, QMC Update Routine

	CPU / Single Machine	CPU+GPU / Single Machine	Internal + External Cloud
Qmax (results/ hour)	1	19	Q_{USER}
Equipment	a_1	a_2	0
Depreciation	b_1	b_2	0
Power/Cooling/Rent	c_1	c_2	0
Other Variable Rate	0	0	d_3
Cost / Result	r_1	r_2	r_3

- ▶ Build functions to evaluate process solutions or combinations of inputs

Conclusion & Summary

- ▶ Applying economic principles provides insights on evaluating HPC processing options
- ▶ Each HPC production process is unique
 - ▶ Business needs should drive definition of process
- ▶ Accelerators & cloud resources most certainly impact a production process
- ▶ Technology adoption rates will vary based on barriers and how organizations choose to address them