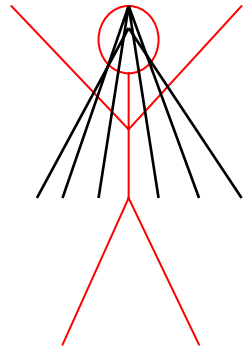

Computer Science Meets Science and Engineering

Jeannette M. Wing

Assistant Director for Computer and Information Science and Engineering, NSF
and
President's Professor of Computer Science, CMU

HEC FSIO R&D Workshop, NSF
August 6, 2007

Outline

- Two Comments on NSF
- Super Data Cluster
- Questions for You

Back to Basics : Transformative Research

- NSF is about basic science and engineering.
  - Preserve CISE core.
- It's all about good ideas and good people.
- It's about "high risk" long term impact.
  - Impact may be far in the future.
  - Impact is long-lasting (that is real science).
  - Impact can create new economies and change societal behavior.
  - ⊘ Say "No" to incrementalism!
  - Promote new, emerging areas of computing.

CDI: Cyber-Enabled Discovery and Innovation

- Computational Thinking for scientists and engineers

- Paradigm shift
  - Yesterday: metal tools (transistors and wires)
  - Today: mental tools (abstractions and methods)
    - "Algorithms" is becoming a household word, e.g., NY Times, Forbes magazine, Harvard Business Review, …

- It's a partnership.
  - To advance BOTH computer science and the other science/engineering discipline.
  - Computer scientist says "I have an incomplete solution that might help you solve your problem." In working on X's problem, new computer science is invented and new science is discovered.

- $52M cross-directorate, $20M CISE
  - Look for the solicitation in the fall

A New Kind of Computing: Focus on Data, Not Control

- Motivation: "~~Data~~ Knowledge is Gold."

  - Computer science
    - The web
  - Scientists and engineers are collecting and generating mountains of data
    - Astronomy: Sloan Digital Sky has 215 million unique objects and growing
    - Biology: Protein Data Base has 41,687 protein structures and growing (only 1% of known)
    - Geophysics: From the Earth's Surface to the Sun and From the Inner Core to the Surface
    - Engineering: Boeing 777 (no wind tunnel)
    - Medicine…
  - Beyond science
    - Art, architecture, and history: Digitization of all museum pieces, historical landmarks, ancient relics

A New Kind of Computing: Super Data Cluster

Super Data Cluster: Massive numbers of processors, each with fast access to massive amounts of data, providing fast interactive response time to end users.

- Google, Yahoo!, Microsoft, Amazon, HP, IBM, Lucas Films, etc.
  "data centers" or "data clusters"
- Randy Bryant's "data-intensive super computing" (DISC) Manifesto
  http://www.cs.cmu.edu/~bryant/pubdir/cmu-cs-07-128.pdf

Scientific Interests

Foundational

- Algorithms (spectral graph analysis a la PageRank)
- Programming languages (massively parallel and ultra reliable a la MapReduce and Dryad)
- A New Computing Platform: not a pc, not a supercomputer, not a distributed system, not a network

Systems and Software

- Self-*: self-configuring, self-managing, self-tuning, self-diagnosing, self-healing, self-repair
- Power: Google and the Columbia River, consumption equal to a metropolitan area
- Software needed to program, operate, and manage apps, e.g., "cluster O/S"

Applications

- Beyond search and web-crawling!
- CS: natural language learning, data-driven graphics and animation, SAT solvers
- CDI: scientific and engineering
- Beyond science and engineering: CreativeIT

Questions for You

- If I were a scientist currently using a high-end computer
  - Would I benefit from a Super Data Cluster?
    - If so, how? If you were a computer scientist, how would you answer my questions:
      - Do I need to rewrite my app? How do I organize my data to better exploit a data cluster as opposed to a supercomputer? Is it more reliable? Is it cheaper to use? Is it easier to share my data and computational results with other scientists?
      - Are there some things I can do on a Data Cluster that I could not do at all on a HEC? Are they really one and the same?
    - If not, why not?

- If I were a computer scientist
  - Could I use an HEC to do my data-intensive computing (e.g., search, NLP)?
    - [Parallel the questions above]

- One answer
  - It's all in the software!
    - e.g., PageRank, MapReduce, GFS, BigTable
  - It's also in the operations & maintenance.
  - It's all in the software!

Questions for You

- As a computer scientist, ask
  - What are the fundamental and practical limits of high-end computing?
    - Technology: Teragrid today, what tomorrow? Petabytes today, exabytes tomorrow?
    - Applications: What other applications besides scientific and engineering, besides massive simulations, besides numerical, etc. can exploit the kinds of high-end computing systems you build?
    - Science: What are the right performance metrics? MTTF of what? What are the right notions of correctness? Data integrity of what when? What properties can you design for, test, verify, and guarantee?

Think science, not just technology and applications.

Thank you!