Some Reflections on Advanced Geocomputations and the Data Deluge

J. A. Rod Blais
Dept. of Geomatics Engineering
Pacific Institute for the Mathematical Sciences
University of Calgary, Calgary, AB
www.ucalgary.ca/~blais
blais@ucalgary.ca

Outline

- Introduction
- Paradigms in Applied Scientific Research
- Evolution of Computing and Applications
- High Performance Computational Science (HPCS)
- Data Intensive Scientific Computing (DISC)
- Cloud Computing and Applications
- Examples, Challenges and Future Trends
- Concluding Remarks

Paradigms in
Applied Scientific Research

- **Traditionally, concept oriented and experimentally tested**
- **Currently, evolving experimental and simulated data driven**
- **Also, often multidisciplinary and multifaceted group projects**
- **In other words, more systems orientation with web collaboration**
- **Experimental / simulated data are becoming cheap and accessible**
- **Distributed data and metadata are often practically unmovable**
- **Sensor webs and computing clouds are changing the landscape**
- **Over 90% of web users look at less than 10% of the pages**

Evolution of Advanced Computing

- **Megaflops to Petaflops, i.e. 10^6 to 10^{15} floating point operations / sec.**
- **Megabytes to Petabytes, i.e. 10^6 to 10^{15} bytes (i.e. 8 binary bits)**
- **Single Core CPUs to Multi- and Many-Core CPUs**
- **Practical limitations:**
    - **TB HDs are common but I/O speeds imply max. ~100 TBs**
    - **CPUs have memory, // instructions and power limitations**
    - **Random-Access files have max. volume I/O limitations**

► **CPUs to GPUs with CUDA code development**

► **The strategic future: The push to Exascale [Turek, 2009]**

From CPUs to GPUs

- 2001: NVIDIA GeForce 3 series of GPUs was a breakthrough
- Coding limited to OpenGL and DirectX for rendering graphics
- Principal applications for games on personal computers
- Computational applications not really feasible for GPUs
- 2006: NVIDIA released GPUs with CUDA Architecture
- 2007: CUDA C, …, under MS Win 7, …, changed things
- CUDA development toolkits available from NVIDIA
- Up to ~10x improvements for general computations
- Up to ~100x improvements for parallel computations

High Performance Computational Science (HPCS)

- **Computations in terms of numerics, symbolics, graphics, …**
- **Languages: Fortran, C, C++, Matlab, Mathematica, Maple, …**
- **Numerical libraries IMSL, LinPack, BLAS, etc.**
- **Parallel programming with MPI, MPV, etc.**
- **GPU coding with OpenGL, DirectX, CUDA C, …**
- **GRID computer systems using the Internet and the WWW**

► Complications with Petascale to Exascale [Turek, 2009]

Data Intensive Scientific Computing (DISC)

- Challenges in Google, Yahoo, Facebook, Amazon, …
- High-level storage and accessibility from anywhere
- Distributed computer systems of unprecedented scale
- Computational capacity for searching and retrieving data

► MapReduce software framework for parallel computations on multi-PB datasets on clusters of computers (from Google)

Cloud Computing and HPC Applications

- **Computing as a service goes back to Computer Centres of the 70s**
- **Google and WWW often seen as prototypes of Grid Computing**
- **Cloud means ubiquitous computer access not only for HPC**
- **Scalable, Elastic, Shared, Metered by Use, Internet based, …**
- **Virtualization and Direct Access → HPC Cloud Computing**
- **Adaptive HPC available everywhere (www.adaptivecomputing.com)**

Astronomy Examples

- **Sloan Digital Sky Survey (SDSS)** www.sdss.org

  Tens of TBs of well-calibrated and well-documented data

- **Dark Energy Survey** www.darkenergysurvey.org

  CCD mosaics in older telescopes (Chile)

- **Large Synoptic Survey Telescope (LSST)** www.lsst.org

  Imaging every 15 sec. => 100 PBs in a decade

- **Virtual observatories** http://usvao.org **and** www.ivoa.net
- **Square Kilometer Array (SKA)** www.skatelescope.org

  Radio telescopes for interferometry => 960 PBs / day (raw data)

Physics / Biology Examples

- **In PARTICLE PHYSICS, Large Hadron Collider (LHC)**

  *http://**lhc**.web.cern.ch* *and* **http://atlas.ch**

  ~ 15 PBs of data annually (over ~ 15 years)

  or ~ 1.7 million dual-layer DVDs per year

- **In BIOLOGY, genomic data are increasingly used for DNA**
- **GenBank at NCBI, the DDBJ of Japan and the European EMBL**

  http://www.ncbi.nlm.nih.gov/genbank/

- **GenBank db of nucleotide sequences (~ 3.4x10^{11} bases in 2011)**
- **Neural activity signal analysis in EEG, MEG, ECoG and MRI**

  Nonlinear series of (Time × Channels × Sampling Rate) data

Climate / Geomatics Examples

- **In CLIMATE, the Coupled Model Intercomparison Project**

  CMPI3 in 2004 ~ 35 TBs of data (used by IPCC)

  CMPI5 in 2013 ~ tens of PBs of data (replicated at many centers)

  www-pcmdi.llnl.gov/projects/cmip/ and www.wcrp-climate.org

- **In NASA's , Modern Era Retrospective-analysis for Research and Applications (MERRA)** http://gmao.gsfc.nasa.gov/merra/

  Data holdings ~ 75.5 TBs (1979 – Feb. 2011)

- **Open Topography with LIDAR data (**www.opentopography.org/**)**

- **Sensor webs for geomatics and environmental research**

  http://*sensorweb*.*geomatics.ucalgary.ca/*

  http://www.lifeunderyourfeet.org

Simulation Examples

- Gravitational N-body simulations [Szalay in CISE, 2011]:
  Millenium Simulations with 10 billion particles
  (done in four time steps requiring ~ 70 TBs over ~ 10 days)
- Remote Interactive Visualization and Analysis (RIVA) at JPL …
  Mars and other planetary applications
  Earthquake research mostly in California
  www.openchannelsoftware.com
- Interactive Visual Analysis of simulated hurricanes

  http://vis.computer.org/vis2004contest
- Geopotential simulations for ~ 5 km resolution on Earth surface
- Monte Carlo simulations for terrain corrections using LIDAR data

Virtual Computing Lab & Web Collaboration

- **IBM Cloud Academy 2009**

  www.ibm.com/solutions/education/cloudacademy

- **North Carolina State University (NCSU) Cloud Education Cloud**

  for general instruction, research and administration
  https://cwiki.apache.org/VCL/

- **Universities of Lethbridge & Alberta Cybera Education Cloud**

  www.cybera.ca and www.ecampusalberta.ca

Concluding Remarks

- The landscape has changed and only more changes can be expected
- Collecting and simulating data are becoming easier and easier …
- Sensor webs, simulations, etc. can bring about a DATA DELUGE!
- Data visualization and multimedia renderings often help in analysis
- In geocomputations as …., ‘what is not scalable is not sustainable’
- Cloud based developments such as VCL are most promising
- Other things on the horizon: nanostructures, quantum computations

P.S.- See Big Data Fact Sheet in U.S. gov. (www.whitehouse.gov/OSTP)

WHAT WOULD YOU DO WITH ALL THIS DATA?

Mathematics and statistics provide the tools to understand ever-increasing amounts of data. To learn more, visit the Mathematics Awareness Month website and enter for a chance to win an iTunes gift card at ***www.mathaware.org***.

Mathematics, Statistics, and the Data Deluge

MATHEMATICS AWARENESS MONTH

Sponsored by the Joint Policy Board for Mathematics—American Mathematical Society, American Statistical Association, Mathematical Association of America, Society for Industrial and Applied Mathematics