SoI MSc Project
Computing the best answer you can afford

Proposer Malcolm Atkinson

Context

- The Global Digital Revolution
- A Cornucopia of Data
- Accessing and processing data costs
    - We will hit the power wall or ...
- Change our habits
    - Restrict our questions to those we can afford

Hard to Change

- Users don’t learn new ways of working
- Existing systems and software profligate

CS to the Rescue

- Frameworks that work within a budget

- Energy efficient hardware architectures

- Energy efficient software architectures

Partition the Problem

- User wants to calculate some function F(D)
  - ag(f(proj(d_i))) where $d_i \ni$ D
  - Logically over all of the data
  - To limit costs only on a sample of data
- Smart framework seeks a good sample
  - with minimum disk and network transfers

Research question

- Can we take advantage of locality
  - data on disk on local node
  - data in same disk transfer
- and still produce a good approximation?

Two approaches

- Given D is already distributed across nodes and disks
  - study ways of sampling with knowledge of the distribution
- Given a set of anticipated functions to compute
  - study ways of distributing D over nodes and disks to make economic sampling feasible