Defining Data Intensive Computing

SOS10

March, 2006

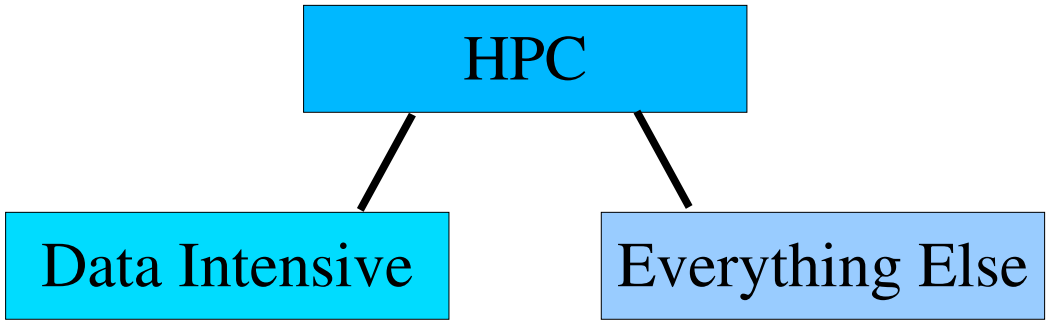
Maui (*yeah!*)

What is “Data Intensive”?

- Can we define it as a set of applications?
- Aren't all “big” applications “data intensive”?
- What is NOT data intensive?

Can we agree that ...

- Data Intensive applications are a subset of High Performance Computing applications?



- So what we need to do is differentiate data intensive applications from those HPC applications that are NOT data intensive.

HPC Applications

- Those applications that require the use of extreme architectures and/or technologies to meet their performance demands.

- Performance demands derive from extreme computation, together with large amounts of data

The HPC Metric

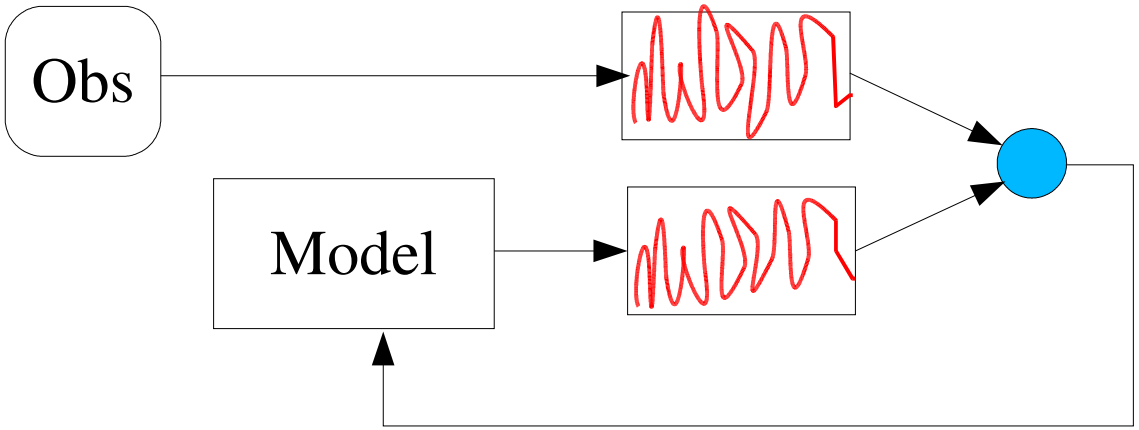
- FLOPS – this has been the single most important metric for most of the history of HPC
- I/O – is only an issue in that it supports FLOPS
- Thus the historical HPC applications
  - Aerospace
  - Ocean/Atmosphere
  - Galaxy Formation
  - Computational Chemistry
  - Structural Biology

Today's HPC Applications

- Newer applications don't have as much FP
  - Genomics, bioinformatics
  - Visualization (computation done on-card now)
- Some have very little computation – almost data processing apps with huge data sets
  - Data mining
  - Homeland security
  - Digital libraries

Closed-Loop Simulation

- Simulate a mathematical model
  - Verify against observation data
  - Close loop by integrating to correct model



Pure Storage / Analysis

- Applications that store continuous output from high data-rate instruments
  - Telescopes
  - Satellites
  - High-energy instruments
- Data analysis, not computation, is the focus

So what IS Data Intensive Computing?

- Does it have to include computation?
- Does it have to include FP computation?
- How much data before its data intensive?
- How important is data intensive to HPC?

Questions for the Panel

- What do you see as the driving applications for parallel file systems and high performance I/O over the next 5 to 10 years? Does this represent a change over the last decade?
- What do these applications need that current parallel files systems don't provide? Do we need a shift in thinking about high performance I/O?
- Should we expect the same file systems to support applications that have both significant computation and I/O as well as those with only significant I/O?
- The file systems from the 60's and 70's, designed for data processing, were quite different from those of the 80's and 90's, designed for interactive workstations. Have we lost any good ideas from those older systems that we should be reconsidering now?