

Patterns of Scientific Computation

a speculative discussion

Clark Gaylord

Virginia Tech Transportation Institute

cgaylord@vt.edu

Problem statement

- How do we advise researchers about how to approach their science (and its computational needs)?
- How do we plan our computational and communications infrastructure to support science?

Infrastructure planning

- We get involved in debates about how to best establish our HPC/HPN resources
- But how do we map these to our research?
- Does the science we engage in become a driver for our infrastructure or are we simply building and hoping they will come?

Why patterns?

- Scientific researchers are not experts in computation
- Computational researchers are not experts in science
- A taxonomy provides a common framework for our language

High Performance?

- What do we mean by “high performance”?
- Computational and communication resources that are beyond those normally achievable by individual desktop workstations or stand-alone servers in typical enterprise environments.

High Performance Computing

- What requires HPC?
 - Computational complexity
 - Scale of the datasets being stored or analyzed
- Virtually all approaches to high performance computing harness some method of parallelism

Taxonomies of computation

- Flynn's taxonomy of computer architectures:
 - Single Instruction Single Data
 - Single Instruction Multiple Data
 - Multiple Instruction Single Data
 - Multiple Instruction Multiple Data
- or, by extension, of parallel computation:
 - Single Program Multiple Data
 - Multiple Program Multiple Data

High Performance Networking

- Simply a matter of scale?
 - Up to certain scale (10Mbps?), use commodity Internet
 - When suitable, use R&E IP network (<1Gbps?)
 - When necessary, get a lambda
- Do we move the data to the computation or the computation to the data?

Data management

- How do we transport data?
- Where do we house data?
- Where do we analyze data?
 - What is the role of “MPP” database systems?

Taxonomies of Computational Science

- Computational workload
 - Data are small (e.g. boundary conditions, parameters), but algorithms are complex
- Monte Carlo simulation
 - Data themselves are meaningless
 - Many computations across multiple parameterizations
- Large data collection
 - Very large raw datasets are gathered and retained for long periods of time
 - Sometimes the raw data need only be archived

Portfolios

- By establishing our portfolio of science and what its infrastructure requirements are, we can better assess our strengths and weaknesses relative to the business of accomplishing science
- Adequately meeting the scientific challenges of a major research institute requires a comparable portfolio of computation infrastructure