Where HPC & Big Data Intersect (HPC Data Analysis Software)

Bruce Hendrickson

Computational Sciences & Math Group

Sandia National Labs, Albuquerque

NNSA National Nuclear Security Administration

Sandia National Laboratories

What is in Scope?

- What is “Big Data Analytics”?
  - SQL Queries?
  - Knowledge discovery?
  - Human-in-the-loop?

- What is “HPC”?
  - Map-Reduce?
  - Shared-Memory?
  - Trans-petascale machines?

Does Big Data Really Need HPC?

- **Lots of talk about "convergence" between big compute and big data**
  - **Comforting, self-serving conclusion**
    - Big compute generates and is needed to analyze big data
    - Networking and memory performance are critical to both
    - Etc.

- **If this is true, why haven't we sold lots of supercomputers to support data analytics!?**

The Search for El Dorado

- **Why use expensive machines when cheap ones suffice?**
  - **Answers must be very valuable**
  - **Response times must be fast, OR**
  - **Analysis is complex (== not amenable to map-reduce)**
- **Limited number of possible consumers**
  - **Wall Street (quants & high-speed traders)**
  - **National security community**
- **Limited number of possible applications**
  - **Graph analytics?**

Reasons to Avoid Using HPC

- **Getting data onto an HPC platform is painful**
  - **Must be able to amortize cost over many analyses**
  - **Or must generate data on the machine**
- **HPC networks weren't designed for analysis tasks**
  - **Need support for fast injection, small messages, many outstanding requests**
- **Software is hard (aka expensive)**
  - **Ecosystem of HPC analysis software barely exists**
  - **Is need persistent enough to justify development costs?**

HPC Data Analysis Software

- **Mostly non-existent**
- **Only real niche – analyzing data generated by HPC**
  - **Even we wouldn't choose to do analysis in situ if we could avoid it, but given poor bandwidth, any alternative would be even worse!**

Ask not what HPC can do for big data …

… but ask what big data can do for HPC!

National Nuclear Security Administration

Sandia National Laboratories

Backup Slides (spoken to the next day)

What I *Really* Believe

Bruce Hendrickson

Computational Sciences & Math Group

Sandia National Labs, Albuquerque

HPC & Big Data Analytics

- **Today's HPC platforms are not cost-effective for most big data challenges**
  - **Over-provisioned processors, under-provisioned I/O system**
  - **Network, programming model, usage model & software ecosystem optimized for scientific workloads**
- **But this is the *wrong* question!**
- **Needs at the component level have strong synergies**
  - **Smarter memories**
  - **Improved power efficiency & management**
  - **Better networks**
  - **More flexible & productive programming models**

Future Opportunities

- Co-investment in solving common component problems
- Potential leverage of each-other's software stacks
- Exchange of ideas and best-practices
- Machines built out of common components
- Enriching HPC via new approaches to parallelism