eScience in the Cloud: Possibilities and Challenges

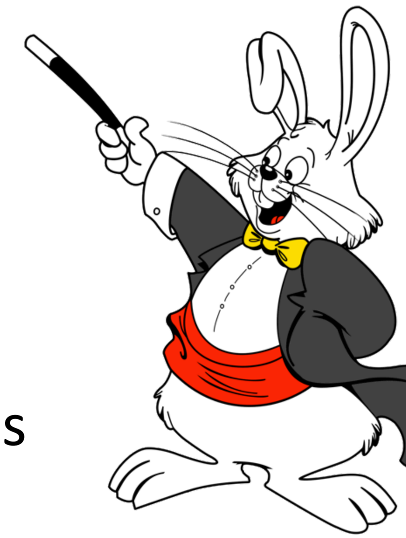
My Perspectives

Philip Papadopoulos

University of California, San Diego

There are always…possibilities

- #1 Possibility : "cloud" computing brings the possibility of computing and data that is available <u>instantaneously</u> and at "<u>infinite</u>" scale
  - A large fraction of scientific computing is compatible
  - But Not ALL science fits on this infrastructure



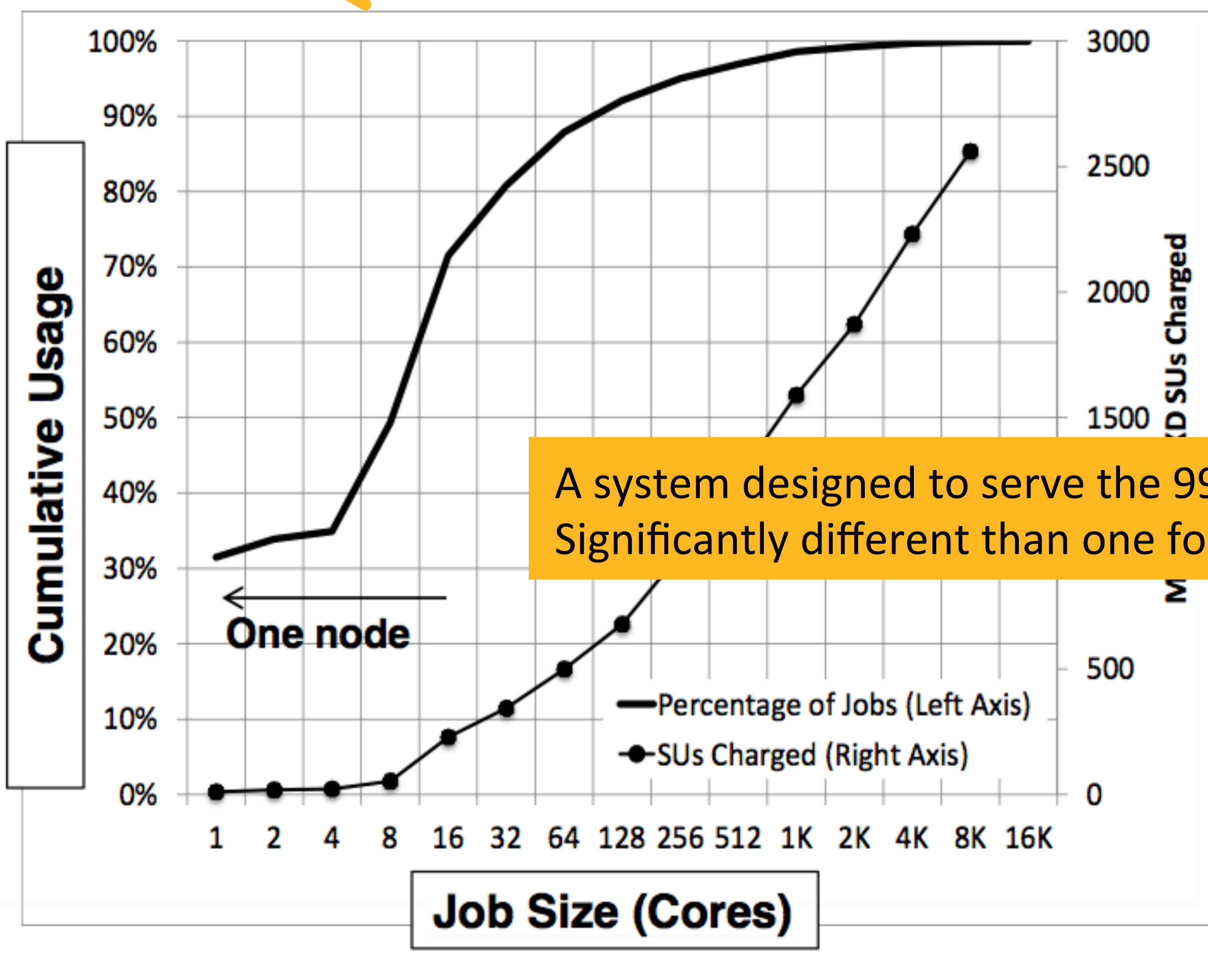
The Cloud is NOT r

- #1 Challenge:
  - Underlying HW tech moves faster than scaling of applications
    - portends a move towards High Throughput Computing
- #1a Challenge
  - Matching your data to computing

Data extracted from NSF's XDMoD data ➔ changing the way big (academic) systems are being built

- **99% of jobs run on NSF's HPC resources <u>in 2012</u> used <2,048 cores**
- **And consumed >50% of the total core-hours across NSF resources**

One rack of Comet – Fit > 90% of Apps



A system designed to serve the 99% is Significantly different than one for the 1%

tional Science Foundation, USA

Scalable Computing → HTC

Core counts on a single System are growing faster than the size of individual computing jobs

- A single comet node (24 cores) could handle 10% of all jobs in XSEDE
- Today’s tech 2016: A single node (96 cores) could handle 20% of all jobs
- Not too distant: A 512-core node would handle about 50% of all jobs (1 rack of comet is 1722)

➔ in the not too distant future. Most <u>individual</u> scientific computing jobs will be able to run on a single machine

➔ Science inquiry will run many individual jobs to answer a single question. (so-called high throughput computing)

- Unknown: Will commercial cloud providers enable users to have access to full nodes for a reasonable cost?

Getting your data in the right place

Imagine a scientific "big data" run in the not-too-distant future that runs 10K simulations at 128-cores/simulation.

What kind of data challenges does one envision?

- Interfaces to data must be tuned to be "high performance".
- ➔ implies larger data chunks retrieved per data query.
- ➔ applications that must attach to many different data sources won't scale. These will run in the cloud, but not efficiently

Does one need data aggregators to support "cloud-scale" computing?

Q: How can international cooperation help accelerate adoption or demonstration of big data and cloud computing solutions in e-science?

- What are <u>some</u> known sources of Big Data?
  - Google has exabytes (but we can't access most of it)
  - Twitter allows tweet archive download (larger data chunks)
  - Traditional simulation output: e.g., KNMI Climate Explorer
  - Larger data aggregators: NIH Sequence Databases, Cancer Genomics Data, Protein Databank, iDigBio. (all sit behind custom web-service APIs)
- What about data captured in the field and put on the network by 100's or 1000's of labs around the world?
  - **If you build data aggregators (to facilitate analysis), how do you retain data ownership and attribution to the originator?**
  - This, is by nature, an international activity

Summary

- For many branches of science, high-throughput computing starts to take center stage
- HTC → many single NODE jobs (instead of many single core)
- Challenge: Getting your Data to your computing
  - This is not new, but it is made more difficult by cloud-computing infrastructure
  - Data aggregation may be a practical approach to matching disparate data sources with many HTC jobs.