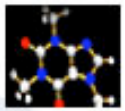
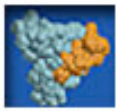
SciDAC
Scientific Discovery through Advanced Computing

Office of Science

Petascale Data Storage Workshop, PDSW08

Rewarding the Public Release of Valuable Data and Resources

Garth Gibson

Carnegie Mellon University and Panasas Inc.

SciDAC Petascale Data Storage Institute (PDSI)

www.pdsi-scidac.org

w/ LANL (Gary Grider), LBNL (William Kramer), SNL (Lee Ward),
ORNL (Phil Roth), PNNL (Evan Felix),
UCSC (Darrell Long), U.Mich (Peter Honeyman)

Carnegie Mellon
Parallel Data Laboratory

pdsi

Bolstering the Data Collection Ecosystem

- Claim1: science is better with data
  - DSN06: asking for fixed MTTI is not == to getting it
  - Google05: 1B words + 1K nodes
    - First qualitative Arabic translation for NIST
  - Hubble, LHC, LSST ... quarks, quasars, dark stuff

  - Science w/ big data “beats” science w/o big data

Bolstering the Data Collection Ecosystem

- Claim1: science is better with data
- Claim2: gathering data is a royal pain
  - Traces (cpu, mem, IO) often a decade old
  - Competitive advantage/marketing embarrassment
  - Lawyers and lawsuits
  - Never transparent, not easy to document
  - Costly to be bigger, more transparent, approved
  - Huge outputs to be distributed

  - Takes fortitude & character to be a data gatherer

Bolstering the Data Collection Ecosystem

- Claim1: science is better with data
- Claim2: gathering data is a royal pain
- Claim3: reward is paper on results from data
  - Not the data release
  - The surprising result extracted from data
  - No reward if getting results not done by gatherer
  - No reward if public download gets to paper first

Bolstering the Data Collection Ecosystem

- Claim1: science is better with data
- Claim2: gathering data is a royal pain
- Claim3: reward is paper on results from data
- Claim4: demotivates continuous collection
  - Finding new results less likely first year after paper
  - Much more likely if systems 100x faster (10 years)
  - Leads to once a decade data collection
    - The current students don't remember the pain
  - Not the best style of data collection
  - Slows down data-led understanding of systems

Bolstering the Data Collection Ecosystem

- Claim1: science is better with data
- Claim2: gathering data is a royal pain
- Claim3: reward is paper on results from data
- Claim4: demotivates continuous collection
- Claim5: no review process for data release
  - Current don't "peer review" a data release
  - A collection paper has novel collection techniques
  - Want "this data collection is best-in-class"

Bolstering the Data Collection Ecosystem

- Claim1: science is better with data
- Claim2: gathering data is a royal pain
- Claim3: reward is paper on results from data
- Claim4: demotivates continuous collection
- Claim5: no review process for data release
- Claim6: confs reluctant to give “paper status”
  - “Bias” paper review for “data release papers” ?
  - Rejects “strong” papers from timely publication
  - Non-competitive selection not good for promotion

Bolstering the Data Collection Ecosystem

- Claim1: science is better with data
- Claim2: gathering data is a royal pain
- Claim3: reward is paper on results from data
- Claim4: demotivates continuous collection
- Claim5: no review process for data release
- Claim6: confs reluctant to give “paper status”
- What makes one release better than another?
  - Bigger? Harder to get? Better documentation?
  - Fidelity = closeness to what really happens?
  - Coverage = contains the info that will be needed?

Bolstering the Data Collection Ecosystem

- Claim1: science is better with data
- Claim2: gathering data is a royal pain
- Claim3: reward is paper on results from data
- Claim4: demotivates continuous collection
- Claim5: no review process for data release
- Claim6: confs reluctant to give “paper status”
- What makes one release better than another?
  - Data size, obstacles, docs, fidelity, coverage ….
- Action: Vet a compelling review process
  - It takes a community to raise a strong discipline