Support for Science: Understanding the Researcher

Eli Dart, Network Engineer
ESnet Network Engineering Group

Researcher Support BoF
Summer 2012 Joint Techs, Stanford, CA
July 16, 2012
Outline

The importance of sociology

Rough classification of science collaborations

Implications for effective support
Sociology

Different collaborations are different (yes, I know you know that)

The important part is that while some collaborations have high performance networking and systems skills within them, most do not

• In some cases they get lucky (e.g. grad student with sysadmin experience)
• It is hard to depend on the presence of network and systems people with performance experience
• The reality is that these skills are not considered to be a core component of an R&E system/network administrator’s skill set

The reasons for different capabilities have to do with the human scale of the collaboration
Science Community Classifications

- **Low** Scientists per collaboration
- **High** Number of collaborations

**Small data instrument science**
(e.g. Light Source users, Nanoscience centers, Microscopy)

**Supercomputer simulations**
(e.g. Climate, Fusion, Life Sciences)

**Large data instrument science**
(e.g. HEP, NP)

A few large collaborations have their own internal software and networking groups.

- **Approximate data set size**
  - 10GB
  - 100GB
  - 1TB
  - 10TB
  - 100TB
  - 1PB
  - 10PB
  - 100PB
Collaborations Have Different Capabilities

Science collaborations of sufficient human scale can adapt to increased data scale better than smaller collaborations

- LHC has made significant investments in time and expertise; other collaborations do not have capital or similar organization
- This shows – the LHC experiments are highly capable users of the network
- Able to reap the scientific benefits of data scale

Smaller-scale science collaborations need help

- These collaborations are unable to bootstrap the necessary expertise
- Alternative structures must exist for smaller collaborations to import expertise that they cannot develop internally
Common Denominator – Data Mobility

Data produced at one facility, analyzed elsewhere

- Scientist has allocation at facility A, data at facility B
- Transactional and workflow issues
  - Experiment, data collection, analysis, results, interpretation, action
  - Short duty cycle workflows between distant facilities

The inability to move data hinders science

- Instruments are run at lower resolution so data sets are tractable
- Grad students often assigned to data movement rather than research

Large data movement doesn’t happen by accident, requires:

- Properly tuned system and network, default settings do not work
- Combination of networks, systems, tools infrastructure must work together cohesively
Implications

We cannot assume that our constituents will attain or preserve high performance skillsets on their own

This means we have to get the architecture right

• #include <science_dmz.h>
• There is an element of self-defense in this – if you can help them get things done right the first time, it’s less work to fix it when it breaks

We also have to understand what they are trying to do – big picture

• One of my first questions is always some flavor of this, unless I already know who I’m dealing with
• Workflow is important, and understanding it can tell you where the issues might be
Questions?

Thanks!

Eli Dart - dart@es.net
http://www.es.net/
http://fasterdata.es.net/