U.S. DEPARTMENT OF ENERGY | Office of Science

Workflow Science: Moving from Tool generation to Discovery

Richard Carlson
Richard.carlson@science.doe.gov

Symposium of the Center for network and Storage Enabled Collaborative Computational Science

May 18-19, 2017

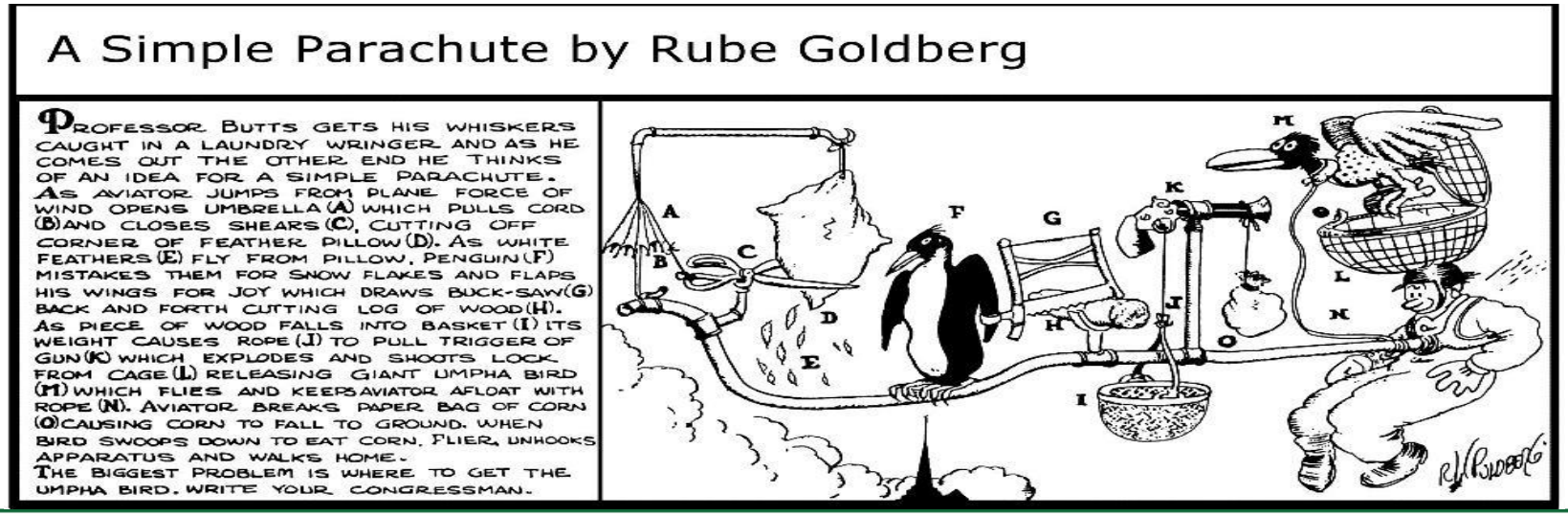
‘Spoilers’

- **The Advanced Scientific Computing Research (ASCR) program office is still developing it’s strategic goals for a scientific workflow program**
  - Funding levels or specific program directions are still being discussed inside ASCR
  - A future workflow program will cross-cut several ASCR programs and SC program offices

- **A broader mix of science communities will use a wider variety of powerful tools and services to make discoveries**
  - A future ASCR workflow program must support this goal
  - Computationally intensive and Data intensive workflows are both part of this future Distributed Computing Ecosystem

Assumptions

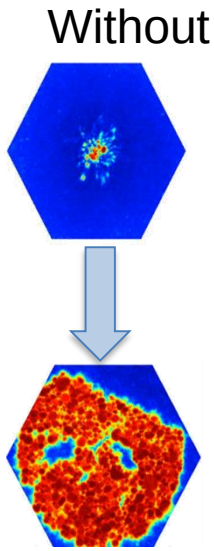
- **Scientific Workflows are an enabling technology to support science in the 21st century**
- **Workflow scientists need to develop the models and experiments that scientifically prove the correct operation of workflows and infrastructure components**
- **Simulation and Modeling work expands, not replaces, historical ModSim activities**



Scientific Computing Yesterday & Tomorrow

- **Computationally intensive codes generate more data and adopt in-situ analysis methodologies**
- **Community based instruments generate more data requiring near-time processing**
- **Multi-user science collaborations with internal specialists providing support community derived data analysis and visualization services**

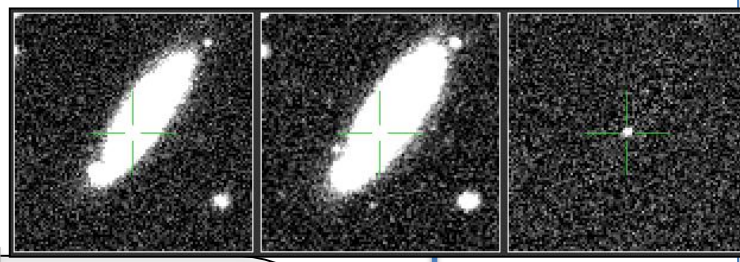
Impact of using immediate assessment of sample alignment in a near-field, high-energy diffraction microscopy experiment



- **Externally operated experimental instruments generate more data requiring near real-time processing**
- **Single scientist collaborating with facilities staff using externally developed data analysis and visualization services**

Palomar Transit Factory

100 TBs of Reference Imaging



Palomar 48" Telescope

HPWREN Microwave Relay

SDSC to ESNET

NERSC Data Transfer Node

Networking Data Transfer

500 GB/night

Computing – I/O

Astrometric Solution

Reference Image Creation

Image Processing / Detrending

Image Subtraction

Nightly Image Stacking

Star/Asteroid Rejection

Transient Candidate

Real-Bogus ML Screening

Heavy DB Access

Real-Time Trigger

40 Minutes

1.5B objects in DB

Scanning Page

Wake Me Up – Real Time Trigger

Publish to Web

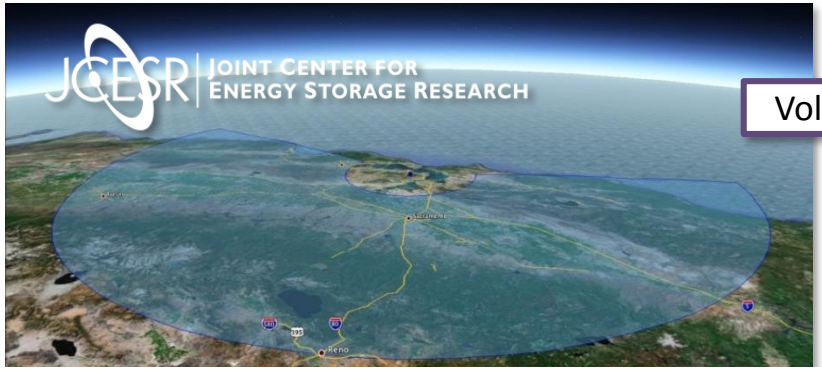
Outside Database for Triggers

Web UI Marshal

Outside Telescope Follow-up

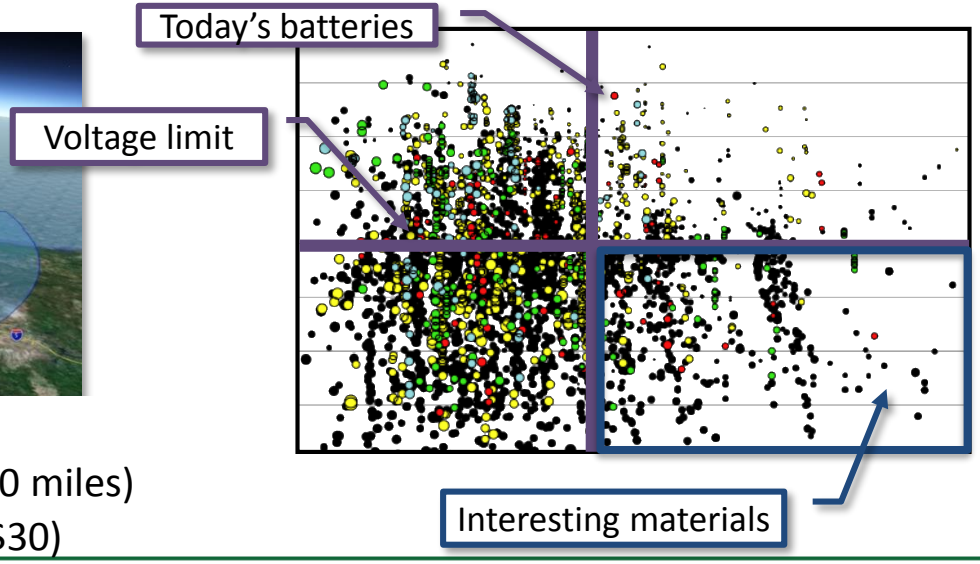
Computationally Intensive - Materials Genome

Computing 1000× today	Data services for industry and science
• Key to DOE's Energy Storage Hub • Tens of thousands of simulations used to screen potential materials • Need more simulations and fidelity for new classes of materials, studies in extreme environments, etc.	• Results from tens of thousands of simulations web-searchable • Materials Project launched in October 2012, now has >3,000 registered users • Increase U.S. competitiveness; cut in half 18 year time from discovery to market



By 2018:

- Increase energy density (70 miles → 350 miles)
- Reduce battery cost per mile ($150 → $30)



Collaboratively Intensive – Material Structures

Contribute to knowledge base

Knowledge base
Past experiments; simulations; literature; expert knowledge

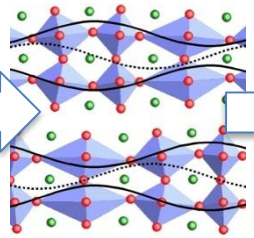
Select experiments (mins—hours)

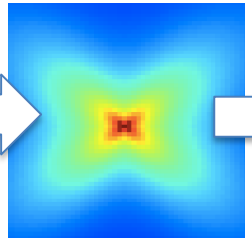
Simulations driven by experiments (mins—days)

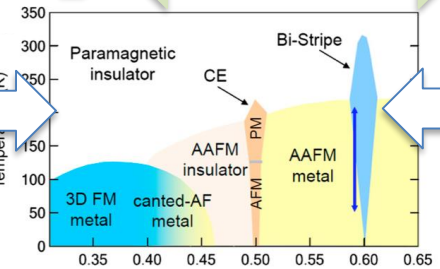
Knowledge-driven decision making

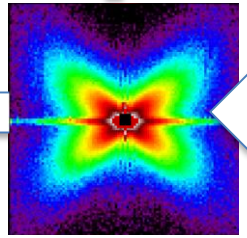
Detect errors (secs—mins)

La 60%
Sr 40%









Material composition

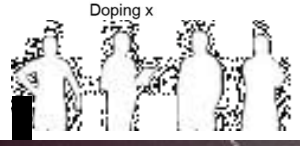
Simulated structure

Simulated scattering

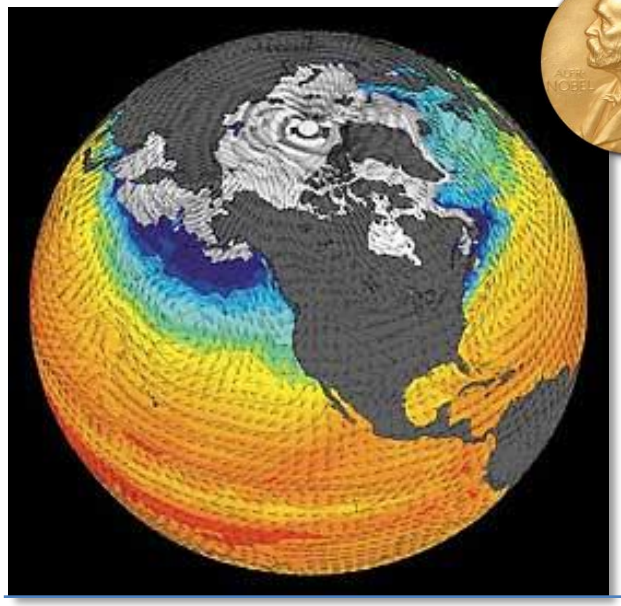
Experimental scattering

Sample

Evolutionary optimization



Computationally Intensive - Climate change analysis



Simulations

- Cloud resolution, quantifying uncertainty, understanding tipping points, etc., will drive climate to exascale platforms
- New math, models, and systems support will be needed

Extreme data

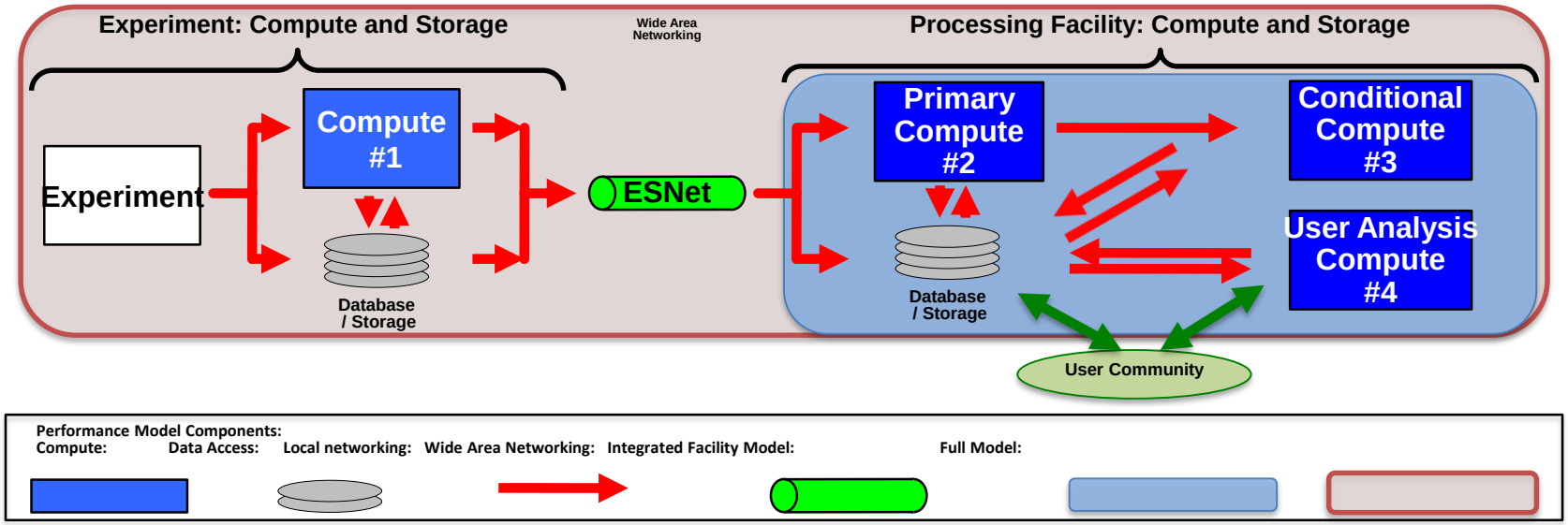
- "Reanalysis" projects need 100× more computing to analyze observations
- Machine learning and other analytics are needed today for petabyte data sets
- Combined simulation/observation will empower policy makers and scientists

Scientific Workflows to the Rescue

- **A scientific workflow system is a specialized form of a workflow management system designed specifically to compose and execute a series of computational or data manipulation steps, or workflow, in a scientific application.**
  - Automate the steps used to complete the science task
  - Automate the transfer of data between tasks
  - Automate the collection of provenance and other meta-data to allow understanding and scientific reproducibility

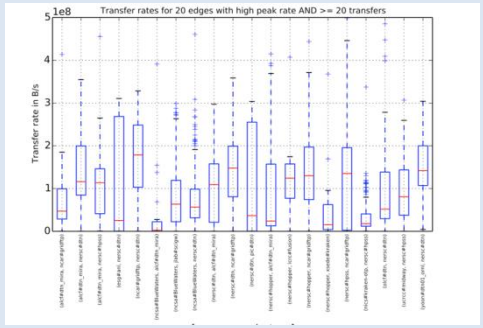
Workflow Tools

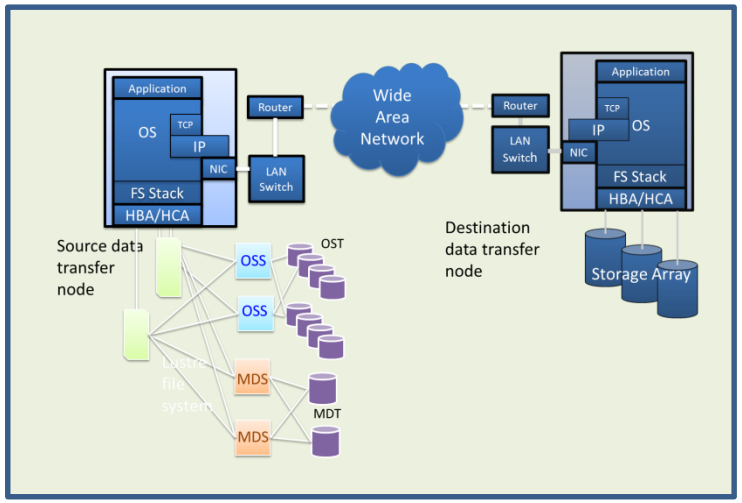
- **Research is required to develop/enhance workflow tools and services**
  - Allow scientists to automate the science data collection process
  - Require standards and agreements to move data, metadata, and instructions between workflow tasks
  - Need to evolve to deal with emerging in-situ and real-time experimental data



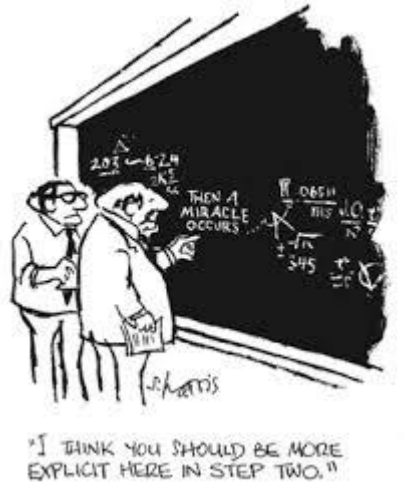
The 3 legs of Scientific Discovery

Collect Experimental data

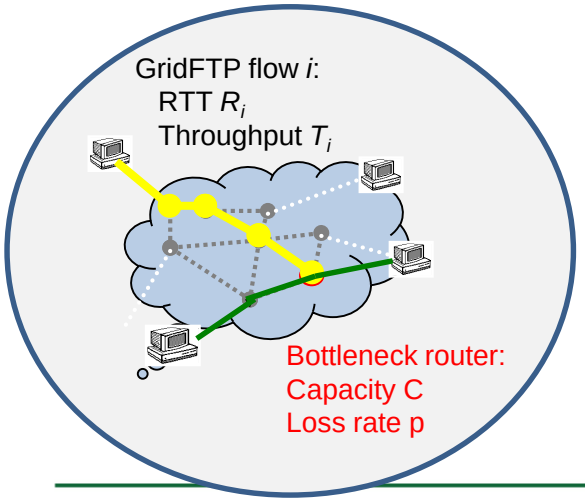




Theory and Math to create predictions and test hypothesis



Computational models provide deep insight



Make File Transfer Times Predictable

Workflow Science

- **Research is required to develop the science of workflows to fully understand how workflows operate**
  - Did the workflow operate as expected?
  - Did the infrastructure (computer, instrument, network, storage) operate as expected?
  - Can the data or metadata be trusted?
  - Is the experiment repeatable?

Experimental Workflow Scientist

- **Design and execute large complex experiments to validate the function of scientific workflows and the facility resources (computers, storage devices, networks, instruments) used by these workflows**
  - What data needs to be collected and how will that task be accomplished?
  - Where will the data be stored and analyzed?
  - What data analysis tools need to be used and/or created?
  - How can the experiment be made repeatable?
  - What are the procedures to align results with theory and simulations?

Computational Workflow Scientist

- **Design and execute large computationally intensive simulations and models that describe the workflow behavior.**
  - What model and simulation tools does the community need to use and/or create?
  - What kind of computational resource is needed to complete the simulation in a timely manner?
  - How can the simulation be scaled up to include more details?
  - What is the process to validate and refine the models and simulations to align with experimental results?

Theoretical Workflow Scientist

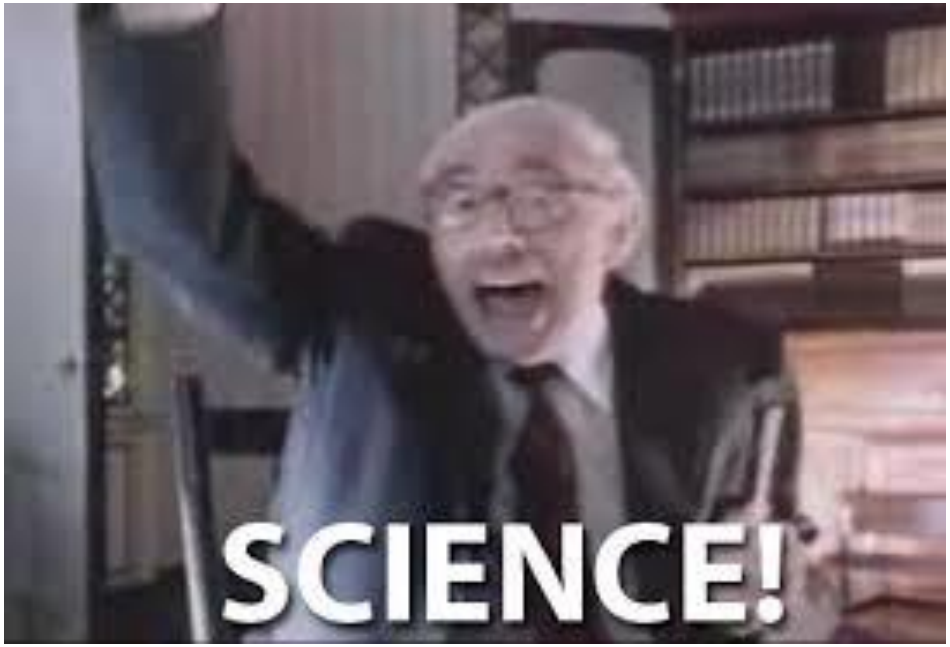
- **Develop and execute theoretical models that describe the expected behavior of workflows.**
  - Are there fundamental behaviors that should be possible at all times?
  - How do implementation details (both hardware and software) of specific resources impact the workflow behavior?
  - What types of predictions are possible?
  - What process can you use to translate these theories into experiments and/or simulations?

Challenges for Workflow Scientists

- **Problem space is large and complex**
  - Spans multiple disciplines and communities
- **Some tools and services exist, but no coherent solution**
  - Simulation tools
  - Experimental results
- **Need for both basic and applied research**
  - Need for adoption by science communities

Opportunities for Workflow Scientists

- **Create new knowledge that explains observed behavior of scientific workflows**
- **Create new theories and experiments that can validate behavior models**
- **Create new models, or model components that predict workflow behavior**



http://escience2017.org.nz/programme/workshops/
International Workshop on Workflow Science (WoWS 2017)

Conclusions

- **There is a strong need for workflow scientists**
  - Specialist in designing experiments to validate tool operation and data collection
  - Specialist in developing models and simulation environments
  - Specialist in developing new, testable, theories of workflow behavior
- **Workflow scientists must work with Domain scientists to**
  - Co-design methods, tools, and algorithms that will automate the application workflow
  - Refine workflow science experiments and simulations to capture sufficient details of applications and facility hardware (computers, storage, clusters, networks, instruments, …)