

Beyond Experiment Control: Experiment Workflow

Robert Ricci

University of Utah,
2nd GENI Engineering Conference
March 4, 2008



The Point of My Talk

- It's clear we need, and can do, experiment control
- But maybe we can and should go farther than that
 - *Services to help manage not just the tasks of GENI experimentation, but the process*
- I'm up here to raise questions, not give answers
 - *Should we consider these things to be in our scope?*
 - *I'm not taking positions one way or the other*
 - *But the questions do come from somewhere: years of experience with Emulab experimenters*



A Laboratory Is Not Enough (?)

- Experiment control is great for managing the laboratory
- ...but not the experimentation *process*
 - *plan / distribute / configure / instrument / init / execute / monitor / stop / collect / analyze / archive / revise / package / publish / repeat*
- A big problem for large-scale activities
 - *Lots of collaborators*
 - *A big laboratory*
 - *Many experiments*
 - *Mountains of data collected*
- Strongly related to scientific workflow
 - *Formally capturing the process of experimentation*
 - *Link together tasks into explicit dependency and dataflow graphs*
- → Can we produce better research process through better services?



Provenance: Where the heck did this data come from?

- We generate a lot of data running experiments
 - *Which run did this data come from?*
 - *Which data set did I use to make this graph?*
 - *Which version of the code was I running?*
 - *Which set of components was I using?*
 - *What were the exact series of steps I ran?*
- Should there be a GENI service to help with this?
 - *Help users explicitly manage experimental runs, so they can connect data generated to the environment it was generated in*
 - *How much of the environment can we really capture?*
 - *Another way of thinking about it: version control for experiments*
 - *Storage, long term archiving?*
- Clearly interacts with GIMS
- Other scientific disciplines have looked at this



Repeatable and Replayable Research

- Not just trace back source of data, but play it again
- Should *some* experiments on GENI be repeatable?
 - *Clearly, not all can be, but probably, some could be*
- Maybe a lower bar: replayable?
 - *Run the same experiment under similar conditions, but without expectation of exact same results*
- Repeatability (or replayability)
 - *... is more than just having the same substrate*
 - *Need to capture as much of the environment as possible*
 - *Similar to capturing for provenance, but the bar is higher*
- Repeating it yourself, or enabling others to repeat it
 - *Packaging and archiving play a big role here*
- Maybe there ought to be a service to manage and/or store all this for researchers?