

I am attaching a diagram I threw together with some high-level thoughts on many of the topics we have covered in our discussions. I'll use this write-up to summarize some of those thoughts with the goal of us getting closer to finding my dissertation topic.

I started with "Cloud Computing" at the central substrate (left-center of the diagram). From there I branched out in two different directions. The first path is what I am calling "Clouds for Research". This path looks at research directions where the cloud is used to enable research. For example, there seems to be research within domain specific areas (e.g., Bioinformatics) on how to use clouds.

This is the area where I see having a framework for benchmarking and reproducibility is very useful. There are areas of CS research where example/public datasets have been stored (e.g., UC Irvine's Machine Learning Repository - <http://archive.ics.uci.edu/ml/>). There are other domains (e.g., bioinformatics) where public datasets are stored on-line. It seems natural to think of using the cloud to analyze these public data. I think having an algorithm repository where researchers can compare an algorithm on a dataset and can have a testbed to run the algorithm on the dataset seems like a great use for Cloud Computing.

The machine learning example above is a particular use case (as we’ve discussed possibly a particular chapter to a dissertation). If I take it up a level, it seems like having a framework/platform to answer questions like: "how do I know which algorithm to use"; "how does my algorithm compare to this other researcher in my field"; etc. This experiment framework must be able to provide the researcher with a way to measure/compare. It also seems to fit with your comments of it better have a good user interface or be easy to use otherwise it becomes too cumbersome for people and thus they will probably not use it.

We’ve discussed items of what I would call “proof-of-concept to scale” where a researcher could start with running locally on his/her laptop to running the same item within a cloud at scale. This is similar to what I have done with VirtualBox locally to then run in FutureGrid or Amazon.

Also, we touched upon the use of clouds for educational purposes. For example for CS courses, the students can actually build their project in the cloud. Reproducibility also seems to fit here. I envision a professor being able to run all of his/her students’ projects within the cloud. For example, the professor could run the projects on a new dataset (one the students didn’t test against to see if the project actually works as well as benchmark each of the projects to capture the performance.

Another specific topic discussed in this area was the use of tools like Twister or Spark, etc. to perform iterative Map/Reduce. This seems to fit with the Machine Learning example above as well where we can use a framework to deploy, compare, and benchmark on a cloud a particular domain of computer science research.

Overall, this branch of cloud computing is interesting to me. This is where I started when I initially signed up for a FutureGrid account. I was looking into to various unsupervised learning algorithms to see which would be the best for my particular use case. I believe it fits well with the thought that there will be many “data scientists” who don’t have the depth of computer science expertise. Can we build platforms or “lower the barrier” for these data scientists to solve problems that are becoming larger and more challenging in their particular domain.

One of the other branches I started in the diagram heads down the path of looking at Cloud Computing from the systems view. I labeled the block “Extending Clouds” only because many of our discussion topics have been investigating where the systems level of clouds can be extended.

There are many research areas of investigating federated clouds, or investigating where grids have been and what can be learned from this decade of research and applied to clouds. We discussed how better user interfaces could help, especially for students who are not computer science majors.

Additionally, we have discussed computing-testbeds-as-a-service and software-defined systems. This research direction seems to move into an intersection with networking research areas where now computing, storage, and networking requirements for an application can all be defined by an API. Other areas such as federated clouds (or sky computing) are being investigated.

Provisioning is another interesting area. FutureGrid provides a unique opportunity where we have hardware at our disposal that other researchers do not. We have the ability to look at bare metal vs. virtualized research areas.

Overall, my interest in cloud computing started with the thread of using clouds to enable research. The use of clouds and the reproducibility paper I submitted to CloudCom 2011 moved me away from Melanie Wu’s research. I wonder if working further on Twister in the cloud would have better aligned with Judy’s research, but we never went this direction and we parted ways on our research work. In some aspects I feel like I have lost a year and a half because of my transition between advisors. I want to make sure this doesn’t happen again. I don’t think I have time in the university’s mind to have this happen again.

If I were to put a statement around a dissertation topic in this “cloud enabling research” area then I would start with something like:

* “A Cloud Computing Platform for xxx Research Enablement” where xxx could be “Machine Learning” or “Database Systems”
* “xxx Research as a Service”
* “An Experimentation Platform using Cloud Computing for Enabling xxx Research”

Again, I am not tied to these areas. If you have visions for where you want to take FutureGrid (and FutureGrid II) then let’s look for a dissertation topic within that space.