Eric Paulz
ECE 4930-004
September 1, 2018

Paper Review: Blue Waters

Summary
This paper presents the results from the first ever in-depth failure study of a sustained petaflop system (Blue Waters at the University of Illinois). The main takeaways are that software is a much larger contributor to failures, both minor and system-wide, than hardware. The hardware showed high resiliency, being able to work through failures and correct itself in many cases. One of the main challenges the researchers faced was analyzing human-generated failure reports, which included natural language aspects and required manual review.

Strengths
I found it helpful that the authors gave such a detailed description of the system (hardware, software, capabilities, etc.) so that readers who are not familiar with it have a better understanding before diving into the data. They also made it clear why they believe their findings are important for the field of large-scale computing. Often you will read a paper presenting lots of impressive data and so forth, but it's all for nothing if you don't have a sense of why any of it actually matters. Figure 2 gives us some lovely visualizations of the data collected about failures. This is essential when presenting lots of technical information because very few readers will have a full understanding based solely on the text.

Shortcomings
Though I thought it was good that they included so much technical detail about the system, I can see how someone who doesn't care much for those details would quickly get lost, confused, or just bored. Also, a few of the tables (V, VII, VIII) are kind of difficult to grasp immediately.

Improvements
I found it difficult to maintain focus throughout the entire paper. I think that they could have left out some of the extremely minor technical details without losing any integrity… for the sake of making it a less cumbersome read. Also, they presented a lot of data about what causes the system to fail, but didn't really present any potential solutions to the problems. I realize that wasn't the goal of the paper, but it would be cool to see a section where they address some potential fixes.

Question(s)
How might the findings from this study translate to an exaflop system if/when one is created? Will the source of failures and downtime be similar, or will the increased order of magnitude produce an entirely different set of issues?