Making Resilience a Reality through a Resilience Consortium

James Elliott

LACSS 2008

Workshop on Resiliency for Petascale HPC

XCR group

LOUISIANA TECH UNIVERSITY

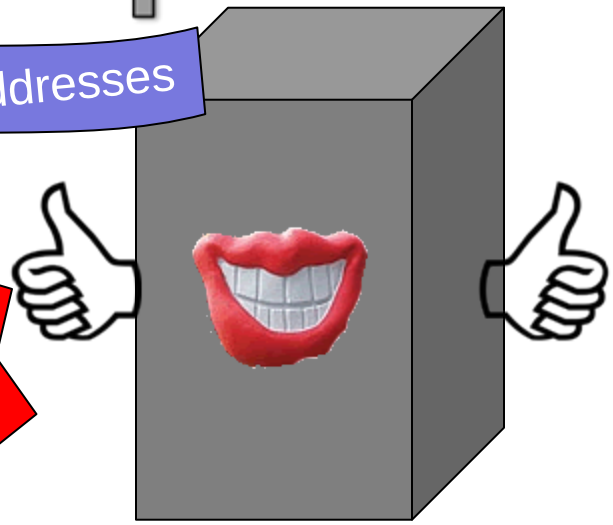
Overview of Researchers

- Advising Professors:
  - Dr. Box Leangsuksun
  - Dr. Mihaela Paun
- Student Researchers
  - James Elliott, Clayton Chandler, Narate Taerat, Nichamon Naksinehaboon
- Collaborators
  - ORNL Team

What is wrong with this picture?

Super Supercomputer!!1!

We Ship to Government addresses



MTBF 100,000!

20 Petaflops on Linpack!

BUY NOW!

VISA MasterCard AMERICAN EXPRESS DISCOVER NOVUS

PayPal

$9,999,999!*
*after mail-in-rebate.

TXCR group

LOUISIANA TECH UNIVERSITY

Wait it doesn't work!

My app keeps hanging!

I thought the time between failure was 100,000 hours?

I'm sorry, we only consider failures to be when the system completely loses power.

Thank You, have a nice day.

What you say matters

With resilience interest growing, now is the time to lay the proper semantic and terminology ground work.

Overview

- Current research
  - Some examples of the research we do
  - What are the challenges we face
- How to approach these challenges
- How a Resilience Consortium can benefit the entire HPC community

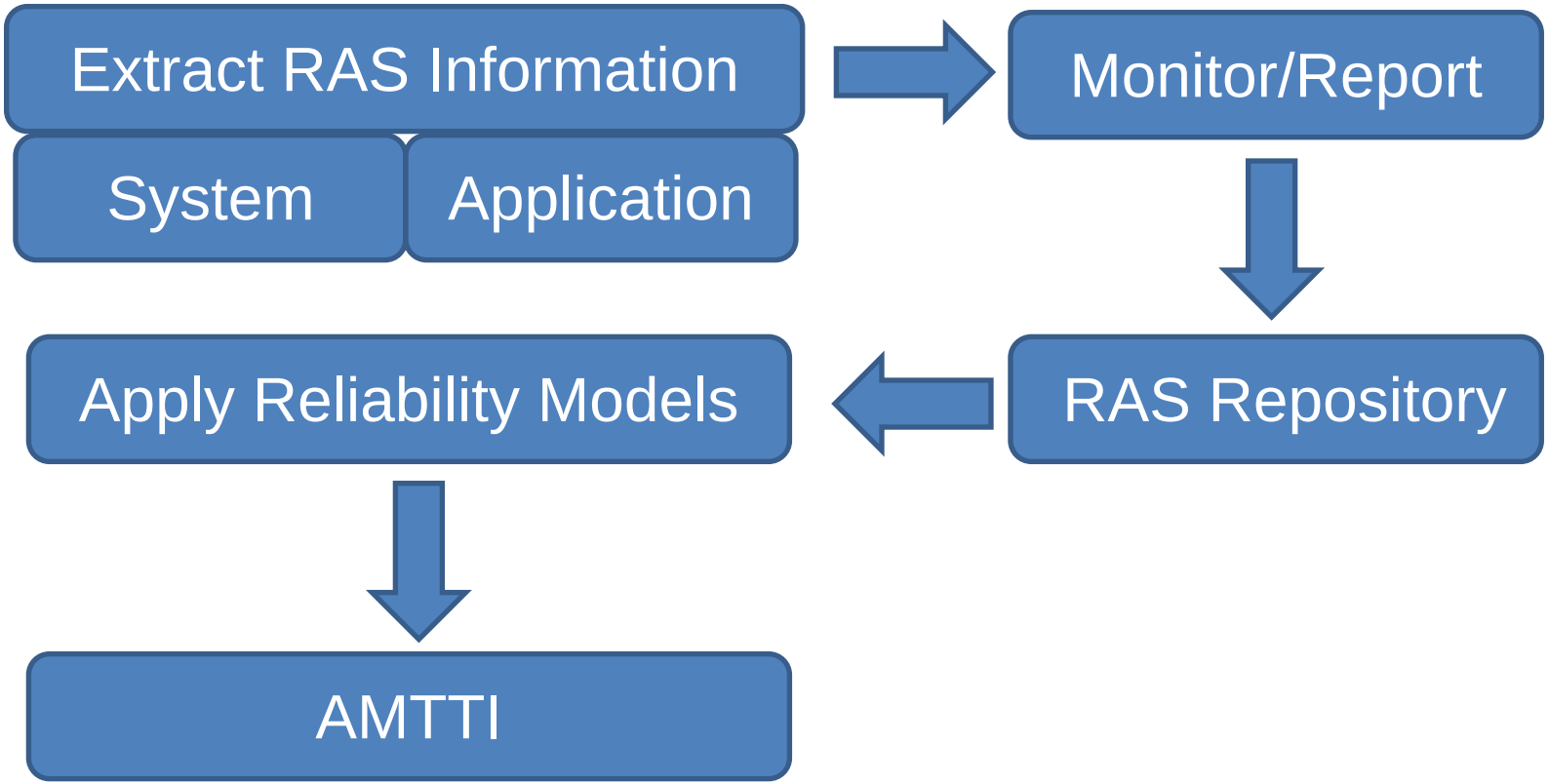
Current Research

- Overarching goal:
  - Increase the Application Mean Time to Interrupt (AMTTI)
- Is that the real goal?
  - No!
  - The point is to get more productive work from a system.

Current Research

- 4 primary areas
  - Logfile Analysis
  - Failure Prediction
  - Application Instrumentation
  - Reliability Modeling

Current Research



Logfile Analysis

- System log messages may identify failures
- Goal
  - Be able to obtain failure information from system logs (failure being both system and application level)
- Primary Challenge
  - Need application health information present in the logfile
  - Currently must make too many assumption. Verify results?

Failure Prediction

- Identify failure from log files
  - RAS support in the logs is needed
- Identify additional information to be recorded alongside the logged events
  - IPMI sensors, /proc/stat, etc.
- Causality analysis (Bayesian analysis) on logged events and additional info.
- Attempt to model the system using a Bayesian network model

Application Instrumentation

- Need more information than is currently present about running applications.
- Goal
  - Gather Resilience-pertinent information from currently-processing applications.
- Challenges
  - Need metrics that express application information
    - How do you express that an application is making productive work? A probability? Simple ‘yes’ or ‘no’?

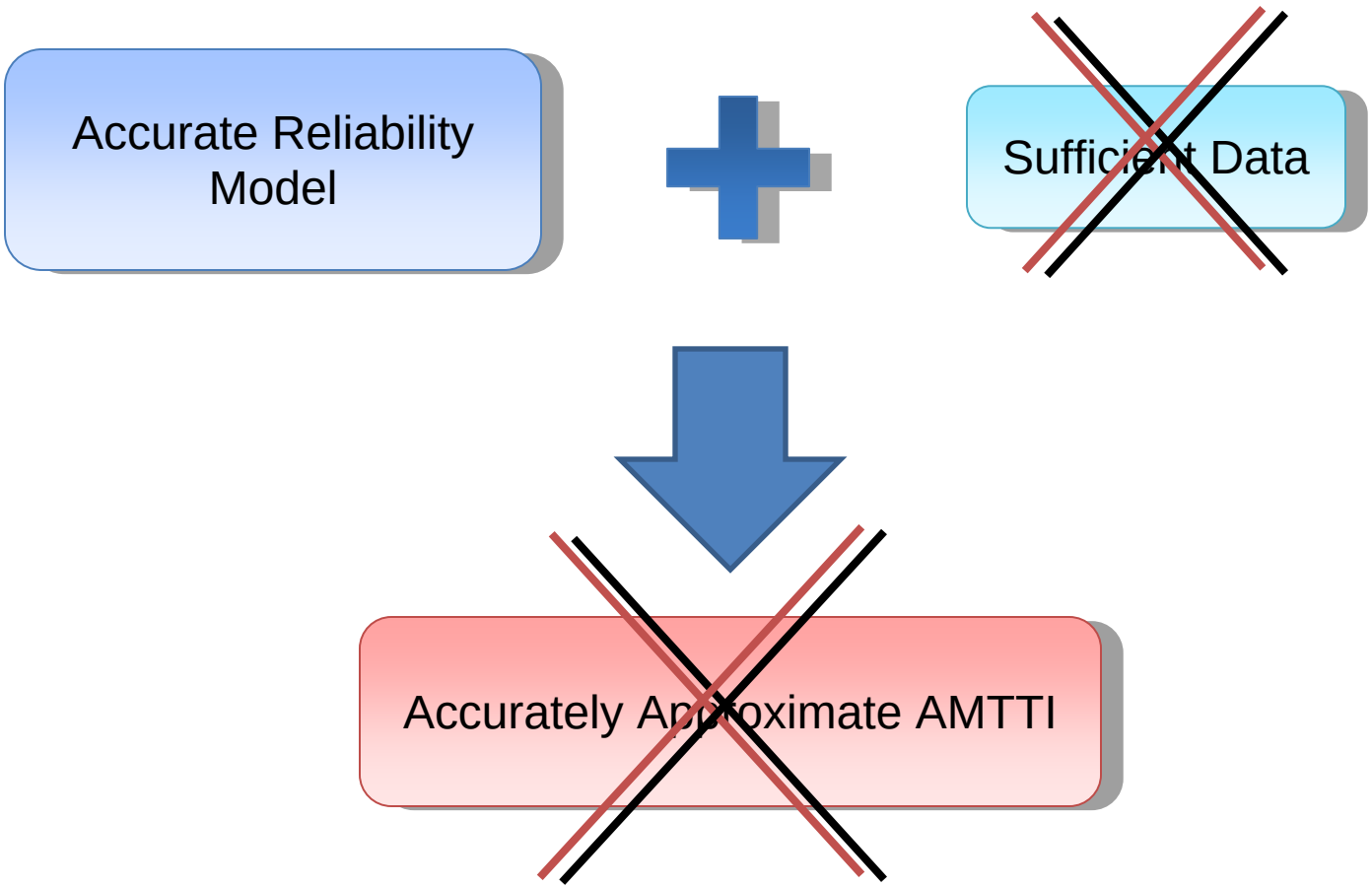
Reliability Modelling

- Represent reliability behavior of both system and application
  - Approximate AMTTI
  - Utilize to increase application productivity
    - Schedule applications onto nodes based on node MTTI
- Main Challenges
  - Need accurate application data
    - Hard/impossible to tell when application failed from logs. OS may not report either.

Reliability Modelling

- Main Challenges
  - AMTTI is useful, but what matters is how much productive work is accomplished.
  - Need to define a way(s) to measure and express application resilience information.

Reliability Modelling



Problem Summary

- How to measure application productivity.
- How to express application productivity.
- What information already available is useful for resilience.
- What information would be really nice to have.
- Historical data is very useful, but logfiles do not currently contain the type of application data that is needed.

Need for a Resilience Consortium

- Many questions are actively being researched by experts.
- Graduate Students are dumb! (or so I heard)
  - A set of guide lines will ensure that resilience research is conducted and presented in such a way that audiences unfamiliar with resilience will clearly see the goals of resilience (and why it is so important)

Resilience Consortium

- Goal
  - Form a working group to define standards and create a knowledge repository.
- Main Portal (wiki)
  - http://resilience.latech.edu
    - Controlled Access
    - Launched April 2008

Resilience Consortium

- Current Tasks
  - Improve communication among members
    - Determine the best technique for prolonged communication (email, forum, blog)
  - Form a knowledge repository for members to share files

Resilience Consortium

- Current Tasks (cont)
  - Plan for Resilience 2009 Conference
  - Define a common cause and find a champion.
  - Publish initial work by years end.

Conclusion

- Terminology and Semantics must be agreed upon.
- Some basic guidelines for conducting resilience research (point of view)
- Plan into ‘09

- http://resilience.latech.edu
- http://cfdr.usenix.org/