U.S. DEPARTMENT OF ENERGY | Office of Science

Report on Inter-Agency Resiliency Workshop

Lucy Nowell

ASCAC Meeting

Wednesday March 28, 2012

Inter-Agency Workshop on Resilience at Extreme Scale

- **Catonsville MD**
- **Feb 21-24, 2012**
- **Organized by John Daly on behalf of DoD, DOE/ASCR and DOE/NNSA**
- **35 attendees representing DOD, DOE/OS, DOE/NNSA, ISI,**
  Vendors: IBM, Intel,
  Academia: UMN, UMD, Rice, IU, Utah

Conference website: https://stone.umd.edu

Define Research & Development in Resilience

Basic Research	Knowledge of Error Types and Rates Prediction
R&D	Detection Notification Recovery
r&D	Integration with whole SW stack and API
Mostly Development & adoption	Develop small set of new capabilities (next slide) Explore transactional model for faults

Four flexible, powerful capabilities could enable resilience in Apps thru 2015

- **Capabilities: task can specify**
  1. **Persistent state storage**
  2. **Ability to recover this state (even neighbors)**
  3. **Get notification (what, when, where fault happened)**
  4. **log_messages(start, stop, delete)**

- **"What happened" examples:**
  - **Lost data**
  - **Lost resources**
  - **Lost capability eg. degradation [BW, proc speed, memory]**

  - **Persistent state storage for predictive apps is assumed "local". For non-predictive apps the storage can be "anywhere"**

Resilience Gap Analysis from Resilience Report

Top 10 Priorities

1. Existing fault tolerance techniques (global checkpoint/global restart) will be impractical at Exascale. Local checkpoint techniques for saving and restoring state need to be developed into practical solutions that are near-term, minimizing app changes

2. There is no standard fault model, nor standard fault test suite or metrics to stress resilience solutions and compare them fairly. (needed to develop solutions)
3. Errors, fault root causes, and propagation are not well understood (HW/SW solutions)
4. Understand rate of silent errors and need for increased V&V (need HW/SW solutions)

5. No resilience in the programming models. MPI, does not offer a paradigm for resilient programming. A failure of a single task often leads to the killing of the entire application.
6. System software is not fault tolerant nor fault aware and are not designed to confine errors/faults, to avoid or limit their propagation, or to recover from them when possible.
7. There is no communication or coordination between the layers of the software stack in error/fault detection and management, nor coordination for preventive or corrective actions. Requires crosscutting solutions.
8. Present Applications are not fault tolerant nor fault aware
9. No effective fault prediction due to #3 and #4
10. In face of continuous failure, the system needs to have continuous repair capability

Last day began Sketching out Resilience Roadmap

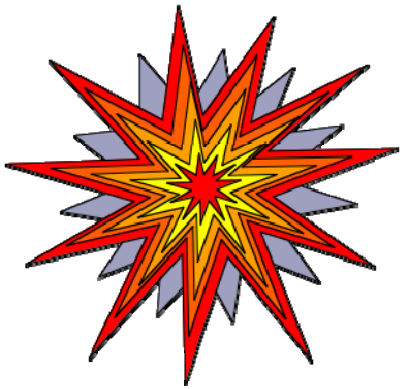
Danger curves ahead.

Roadmap is still a work in progress. Present state can be found on the workshop website: https://stone.umd.edu

Current State of the Roadmap

Resilience Workshop

- **In planning stages**
- **Tentative date is June 6, 2012**
- **Washington, D.C. area**
- **Target Participants: Representatives from DOE National Labs plus two industry people.**
- **Focus: Triage the ideas from the Catonsville workshop to identify the research topics that are strategic for DOE, with emphasis on near term priorities.**



Thank you!

(Special thanks to Al Geist of ORNL for help with these slides.)

Lucy Nowell

lucy.nowell@science.doe.gov