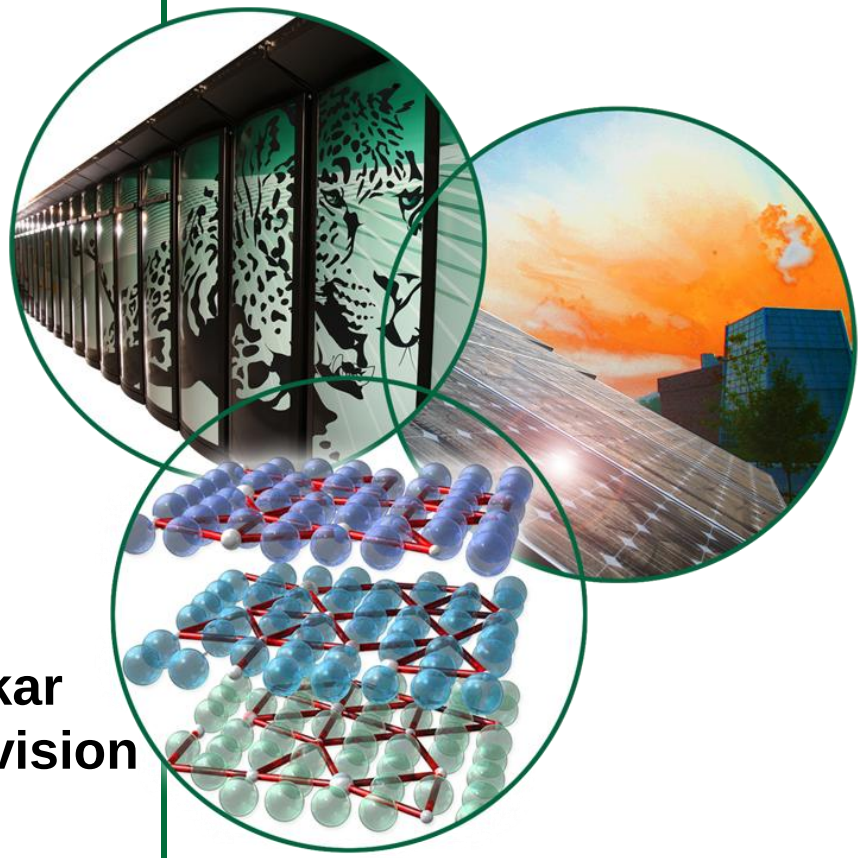
High Performance Computing Resilience Testbed



Héctor J. Machín Machín
University of Turabo
Research Alliance in Math and Science

Thomas Naughton, Dr. Geoffroy Vallée,
Dr. Stephen L. Scott, and Anand Tikotekar
Computer Science and Mathematics Division

August 2009

U.S. DEPARTMENT OF ENERGY

OAK RIDGE NATIONAL LABORATORY
MANAGED BY UT-BATTELLE FOR THE DEPARTMENT OF ENERGY

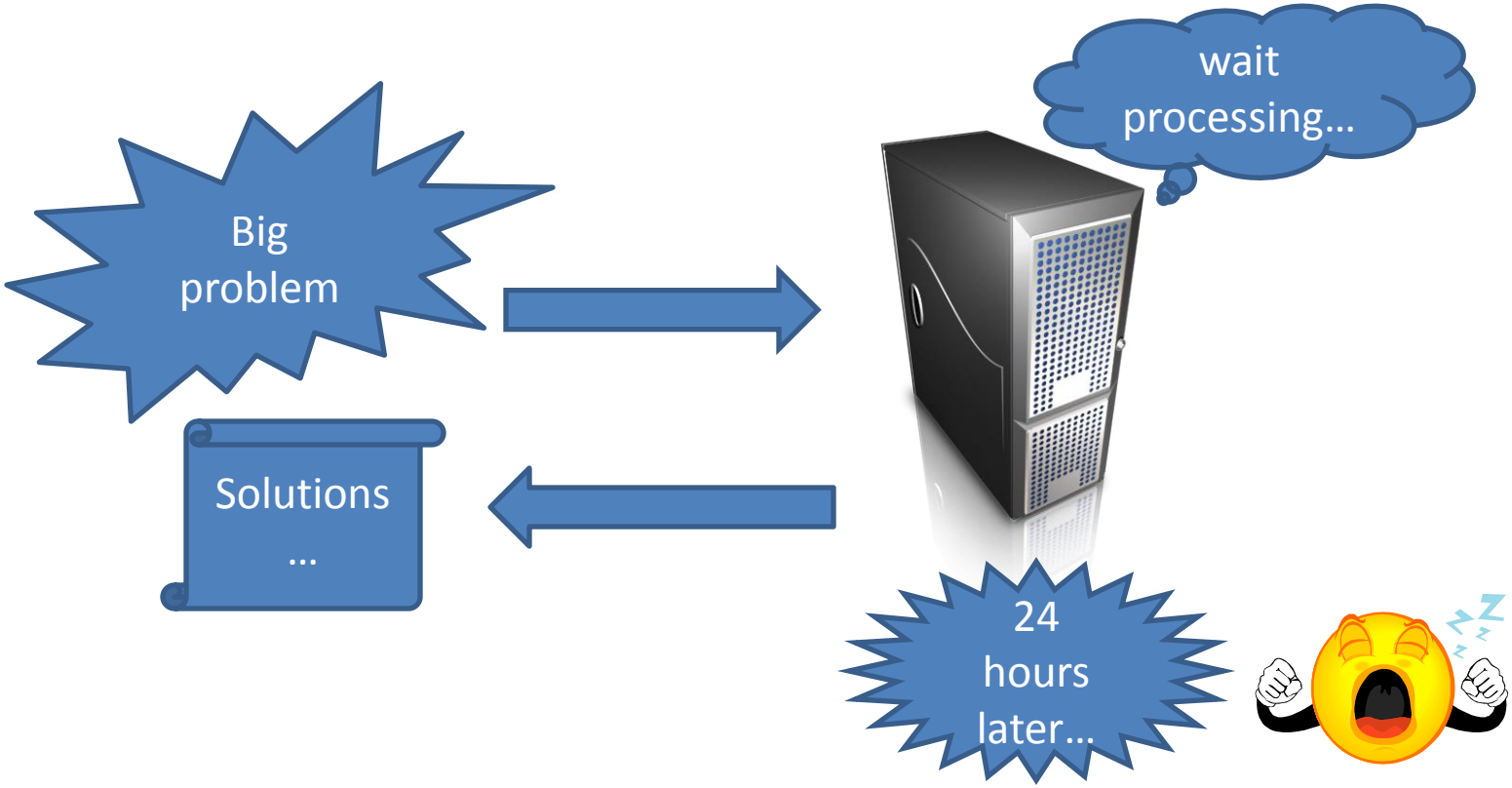
Outline

- **Background**
  - Fault tolerance - checkpoints
- **Research objectives**
- **Resilience testbed**
  - Methodology
- **Experiments guide**
- **Results**
- **Future work**

OAK RIDGE National Laboratory

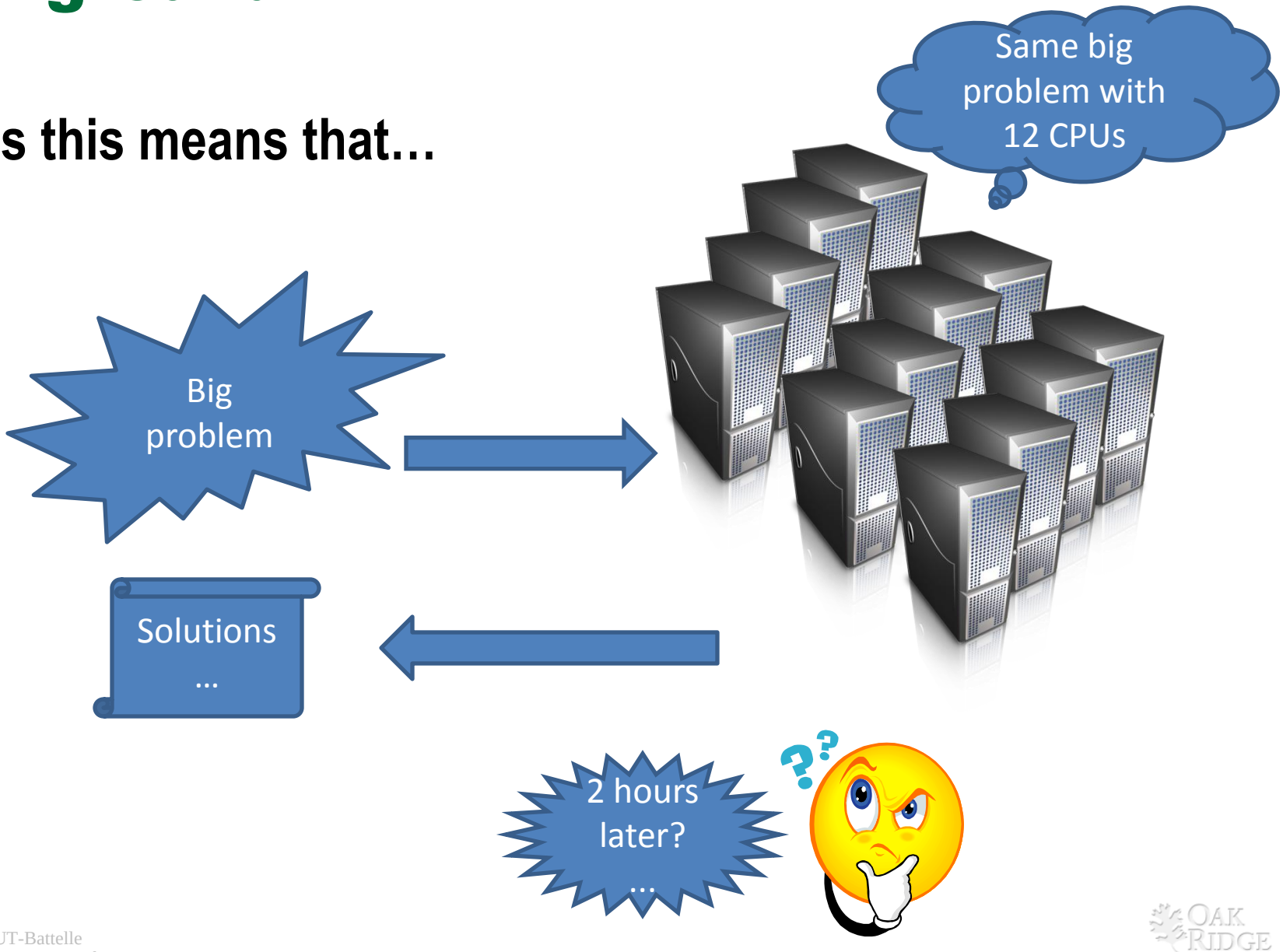
Background

- **Computers have helped us to solve many problems**



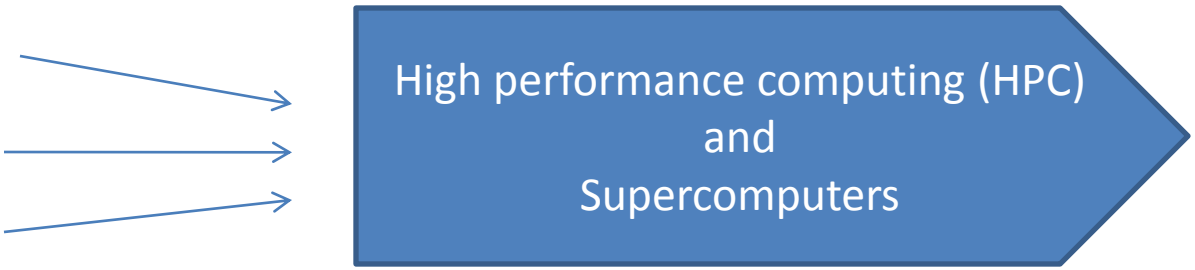
Background

- **Does this means that…**



Background

- **Answer: no…**
  - Limiting case and can only be approached…
- **Depends on many factors**
  - Can the problem be expressed in a full parallelized program?
  - Ideally: Amdahl's Law
    - The serial portion of a program will be the limiting factor in how much you can speed up the execution of the program using multiple processors
  - Overhead will be introduced by parallelizing the code
- **Solutions**
  - Better algorithms
  - Faster computers
  - More computers

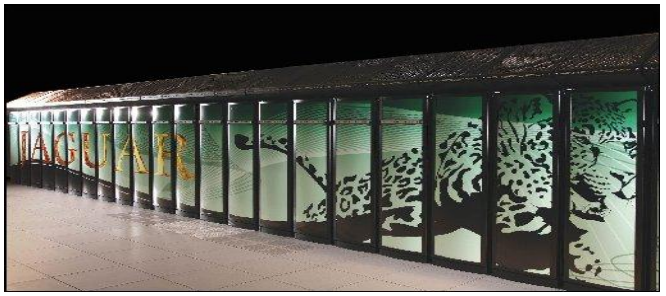


OAK RIDGE National Laboratory

Background

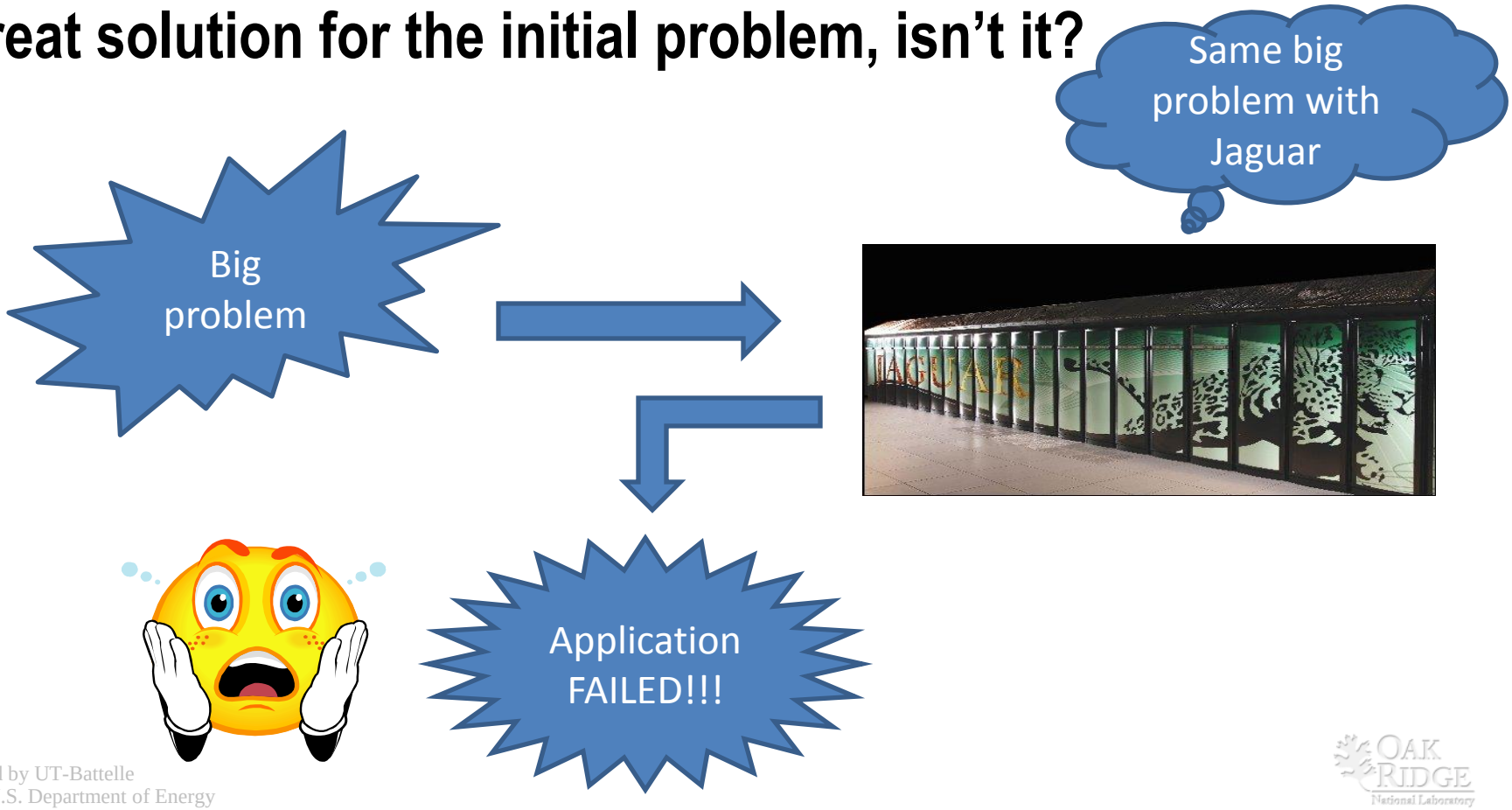
- **Supercomputers have reached petascale computing power!!!**
  - Petaflops – one quadrillion floating point operations per second
    - 1,000 trillion operations per second
  - Fastest supercomputers: †
    - Los Alamos National Laboratoy: Roadrunner
      - 1.105 petaflops
      - 3,456 nodes
      - 129,600 cores
    - Oak Ridge National Laboratory: Jaguar
      - 1.059 petaflops
      - 18,688 nodes
      - 150,152 cores

† http://www.top500.org



Background

- **These systems will continue to be upgraded…**
  - Increasing computational power
- **Great solution for the initial problem, isn't it?**

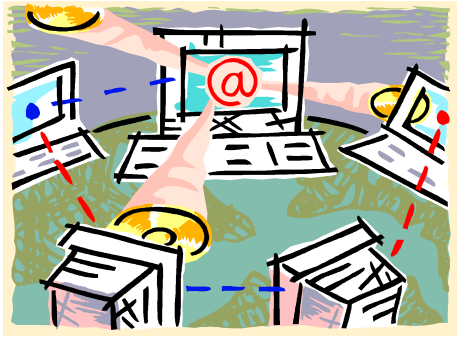


OAK RIDGE National Laboratory

Background

- **Increase in probability of failures in HPC systems**
  - Systems increase in size
  - Complex scientific computing applications
- **Software and hardware failures**
  - Increases application interruption
  - Decreases productive use of resources
  - Wastes allocation hours (costs)

What can be done about it?....



OAK RIDGE National Laboratory

Fault tolerance method: checkpoints

- **Checkpoints: fault tolerance in HPC systems**
  - Application periodically stops useful work
  - Writes checkpoints to disk
- **Transition to petascale computing** †
  - Multicore size doubling each 30 months
  - System utilization drops to zero by 2013
  - 100% time only writing checkpoints

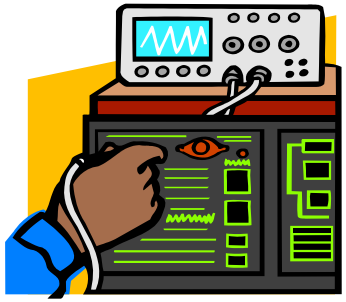
† "*Understanding Failures in Petascale Computers*" - B. Schroeder, G. Gibson

OAK RIDGE National Laboratory

Research objectives: resilience testbed

Fault injection techniques

Testing application



Analyzers for failure evaluation

Data gathering/monitoring

Target System

OAK RIDGE National Laboratory

Resilience testbed

- **Research on fault injection techniques**
  - Register bit flips
  - Network congestion
  - Kill an instance of the application on one node
  - Overload

OAK RIDGE National Laboratory

Resilience testbed

- **Research on data gathering/monitoring**
  - Ganglia
    - Monitors supercomputers
    - Stores data in round robin databases (RRD)
  - Ovis 1.1
    - Reads the RRD
    - Creates an output file with numeric format of the RRD

OAK RIDGE National Laboratory

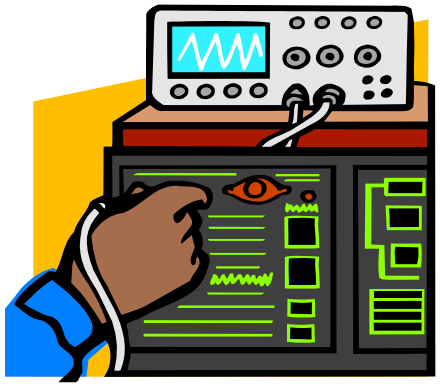
Resilience testbed

- **Research on testing application**
  - HPCC parallel benchmarks
    - Tests that examine the performance of HPC architectures
  - Cbench – benchmark framework tool
    - Building
    - Running
    - Analyzing output

OAK RIDGE National Laboratory

Resilience testbed

- **Research on analyzers for failure evaluation**
  - "*Detecting Anomalies for High Performance Computing Resilience*" S. Quiñones, E. Avilez, Y. Camacho (FAST interns, summer 2009)
  - R - software environment for statistical computing and graphics
  - Ggobi - project and analyze multivariate data



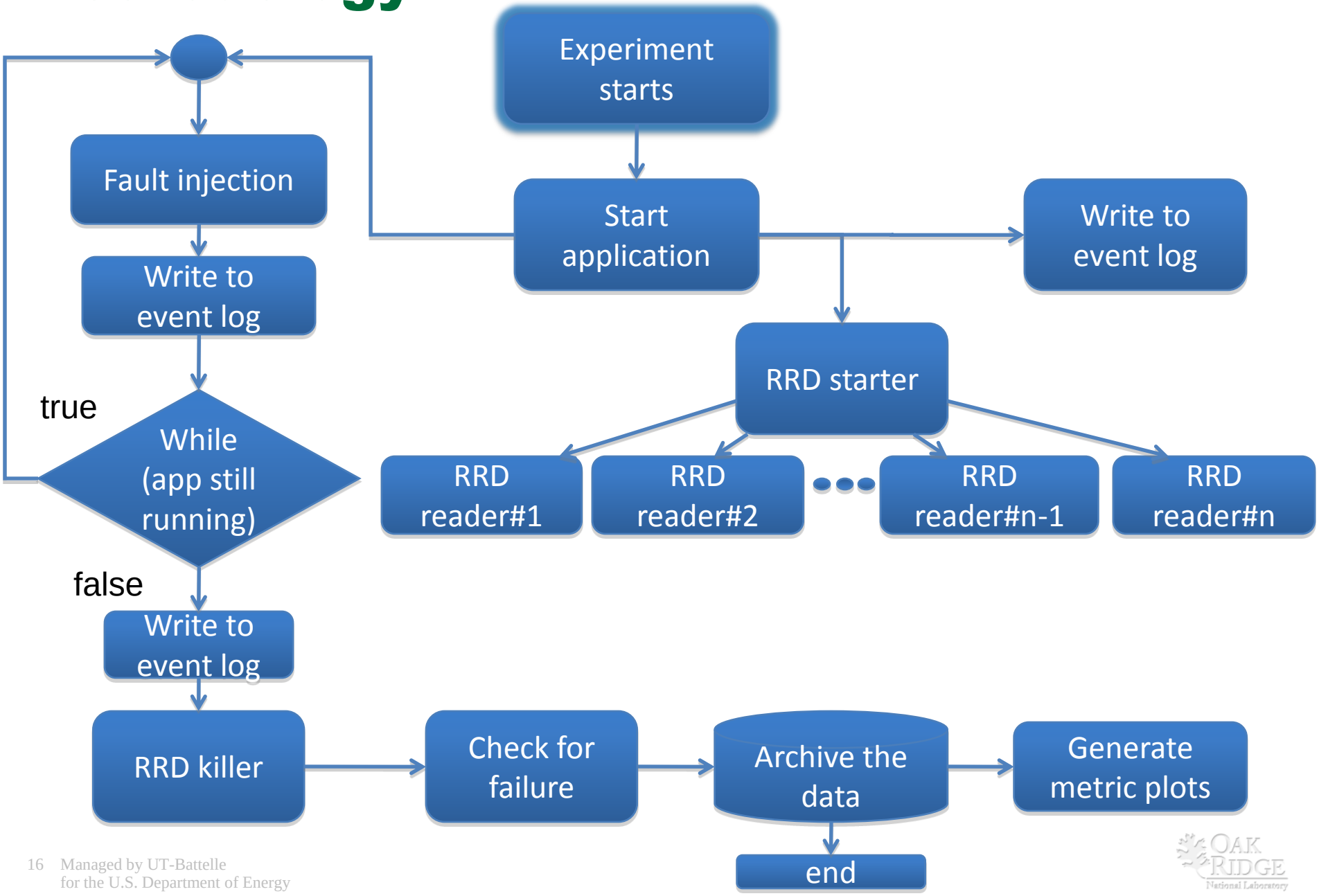
OAK RIDGE National Laboratory

Target system: XTORC cluster

- **64 Nodes -** 1 processor per node
- **Pentium IV Processors**
  - L2/L3 cache: 256KB
  - Speed: 1695.87 MHz
  - **Node memory: 768MB**
- **100 Mbps ethernet**

OAK RIDGE National Laboratory

Methodology



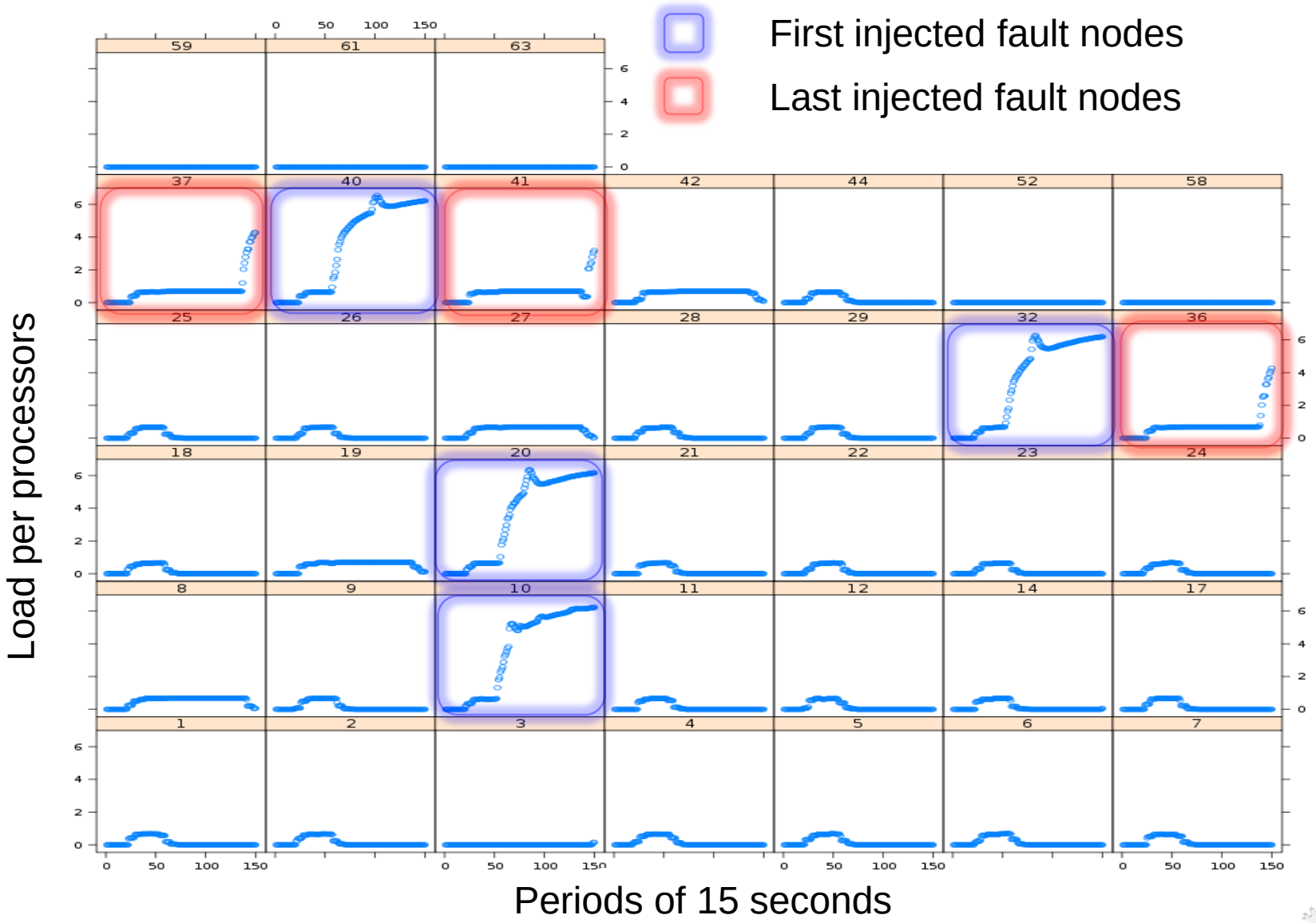
OAK RIDGE National Laboratory

Experiments Guide

1. sleep12hrs - gather data without any workload
2. hpcc_onenode - ran hpcc for one node
3. hpcc_allnodes - ran hpcc for all nodes
4. kill_fault - killed an instance of hpcc for one node
5. network_fault - ran infinite pings on some nodes
6. overload_fault - ran more than one instances of hpcc

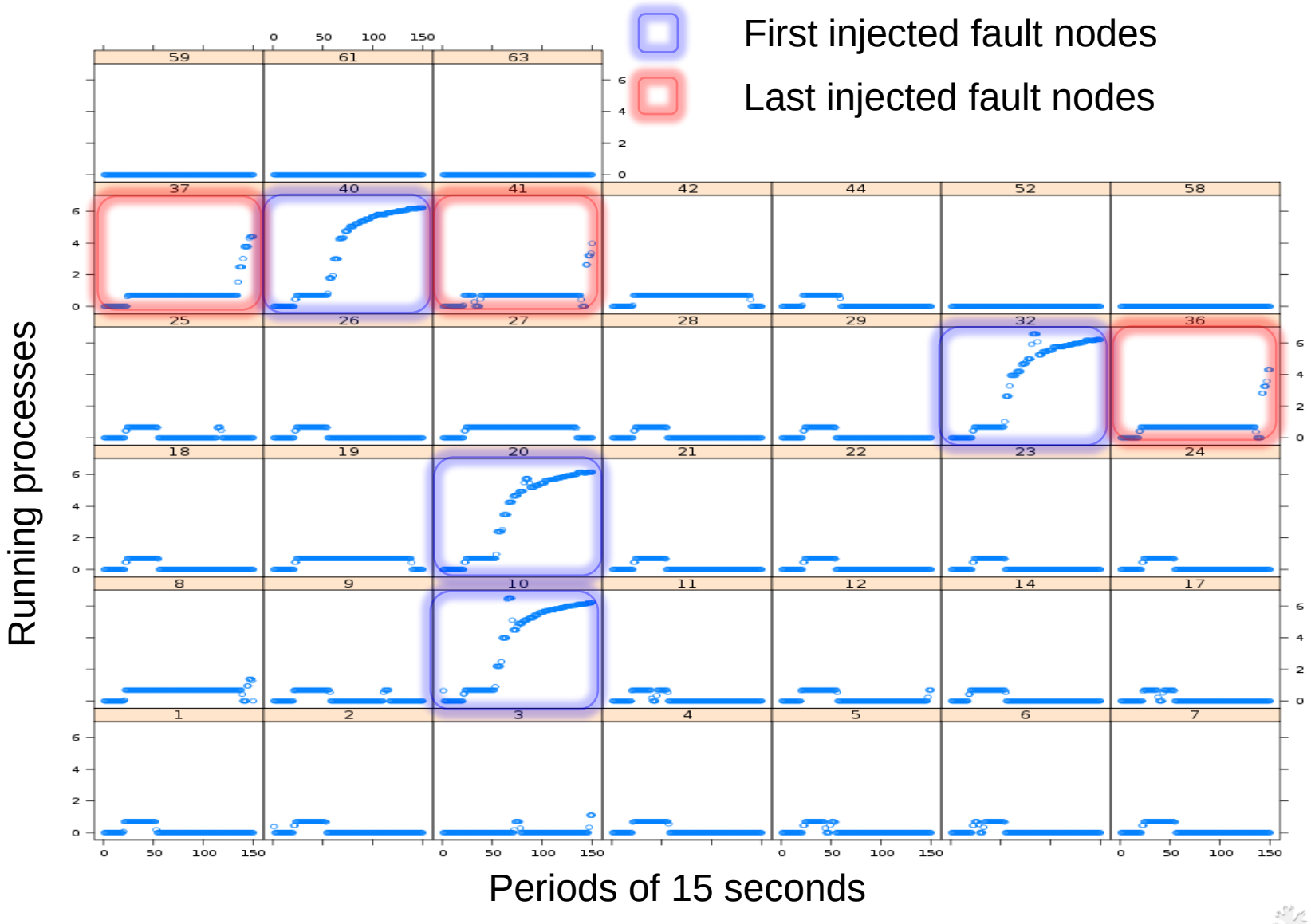
OAK RIDGE National Laboratory

Results: network_fault experiment



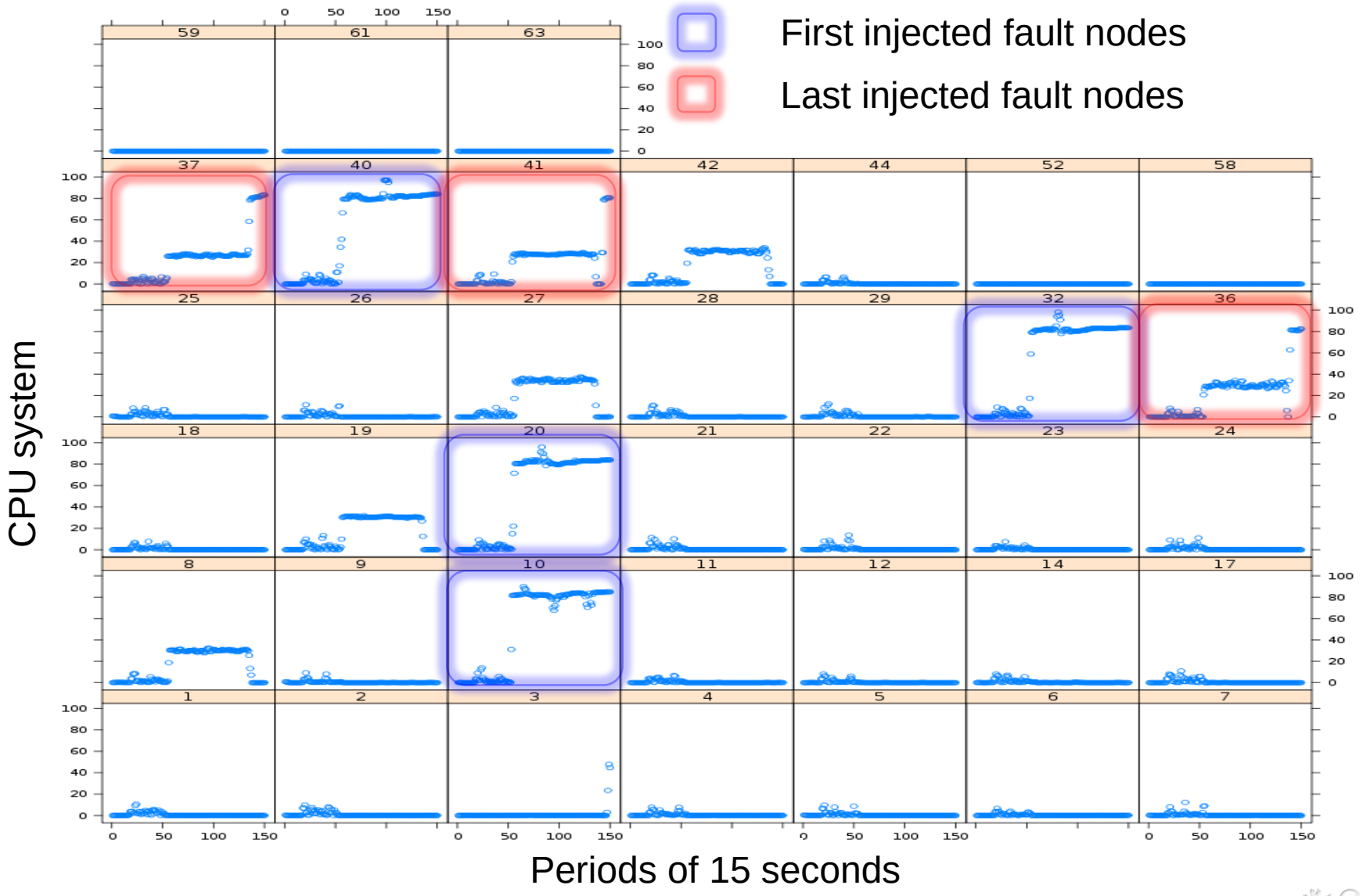
OAK RIDGE National Laboratory

Results: network_fault experiment



OAK RIDGE National Laboratory

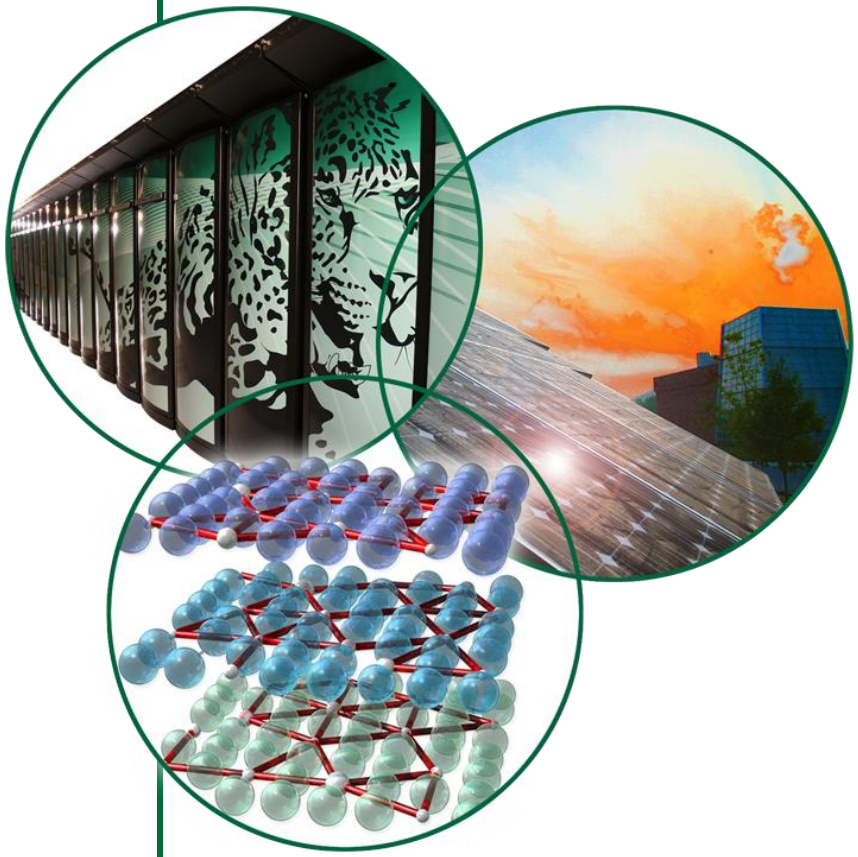
Results: network_fault experiment



OAK RIDGE National Laboratory

Future work

- **Explore more fault injection techniques**
- **Explore the integrity of the fault injected**
- **Extend research on more complex HPC systems**
- **Complete package**
  - Integrate to OSCAR (Open Source Cluster Application Resources)
  - Continued for Capstone project next semester

OAK RIDGE National Laboratory

Questions?