Improving Fault Tolerance for Extreme Scale Systems

Eduardo Berrocal
eberroca@iit.edu
Illinois Institute of Technology, Chicago, IL

SC16 Doctoral Showcase

Content

- Hard Error Prediction
- Soft Error Detection

Hard Error Detection

- Motivation
  - More faults arising from hardware components
  - Traditional methods may be rendered useless
  - New data = new opportunities

Hard Error Detection

- Mira Supercomputer
  - IBM BlueGene/Q
  - Top 5 (as of June'14)
  - 10 petaflops
  - 48 racks (48k nodes)
  - Environmental Sensors
    - Temperature, coolant flow and pressure, fan speed, voltage and current

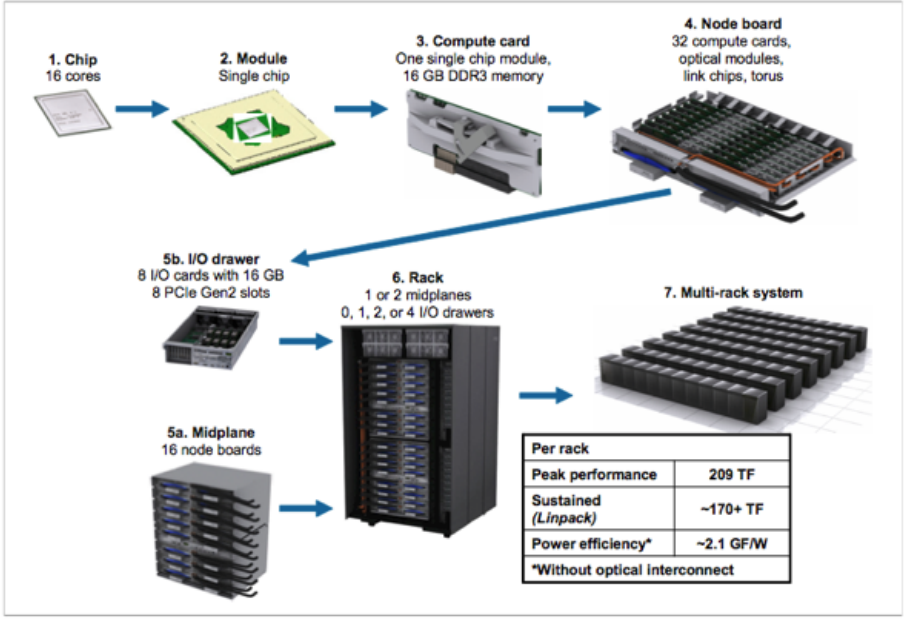


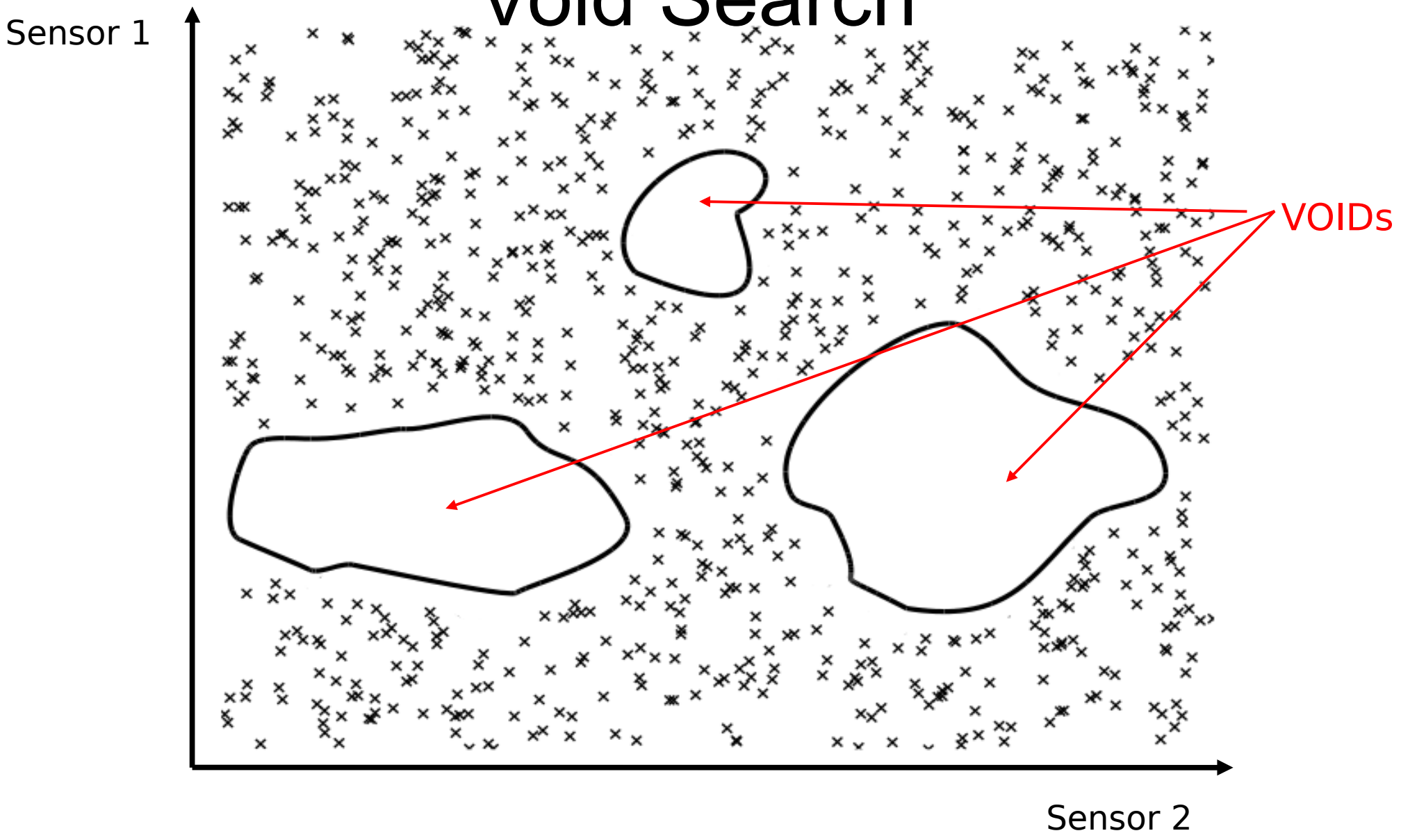
Figure 1-2 Blue Gene/Q hardware overview

Environmental Information

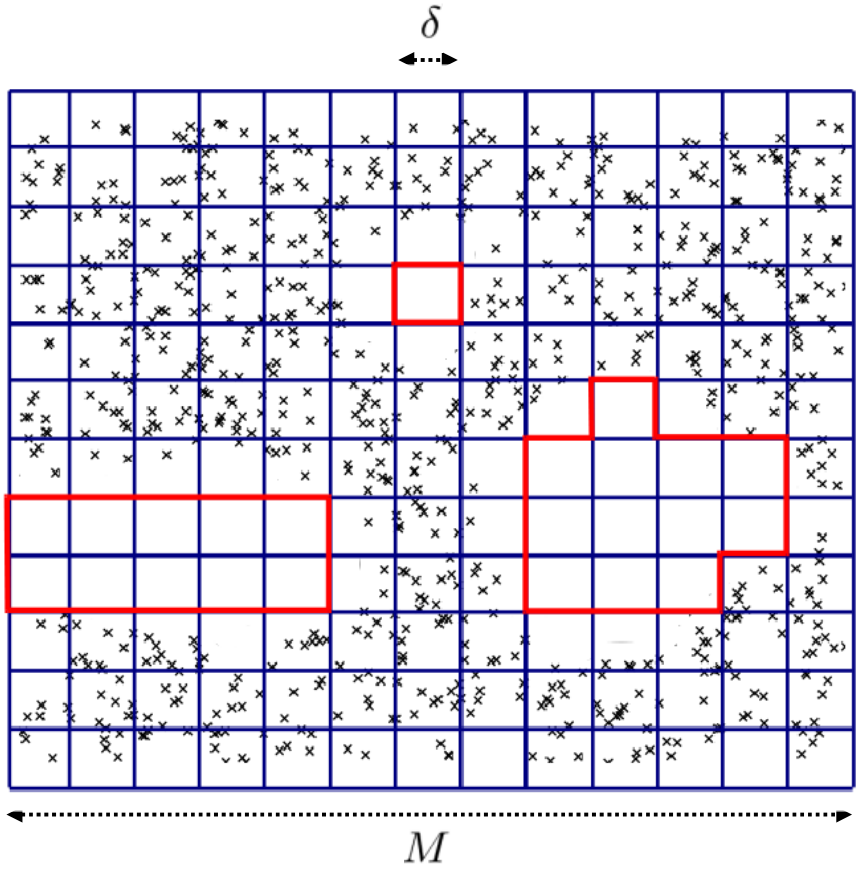
location	time	inletFlowRate	coolantPressure	ambientHumidity	…
R1A-L	2012-09-01 00:03:03	2680	3949	3144	…
R0F-L	2012-09-01 00:03:03	2417	4049	4718	…
R10-L	2012-09-01 00:38:10	2673	3985	3329	…
R1F-L	2012-09-01 00:53:14	2515	4073	4250	…
R04-L	2012-09-01 00:53:14	2503	4031	3083	…

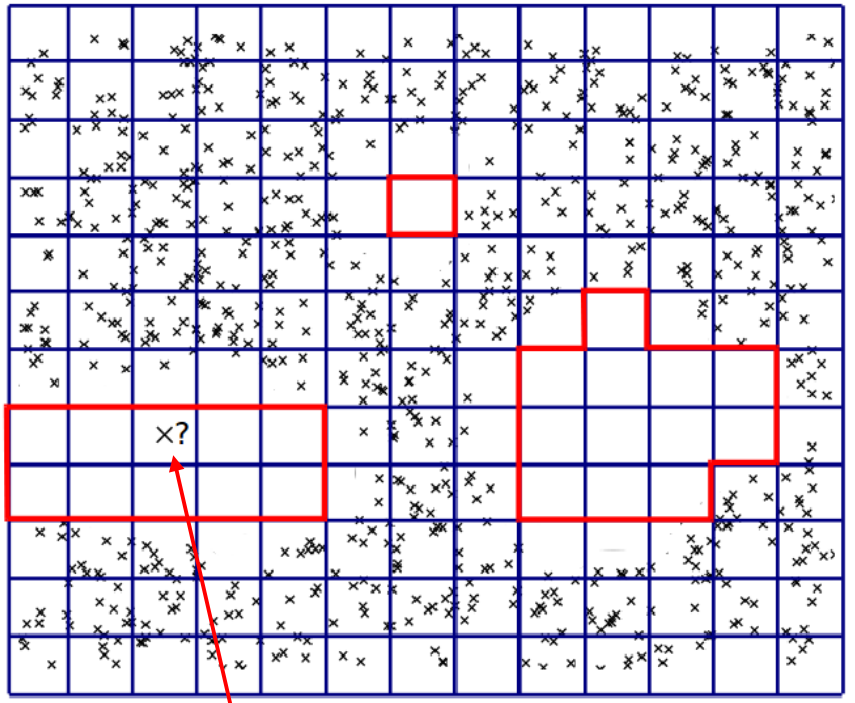
180 sensors per compute card (node)

Void Search

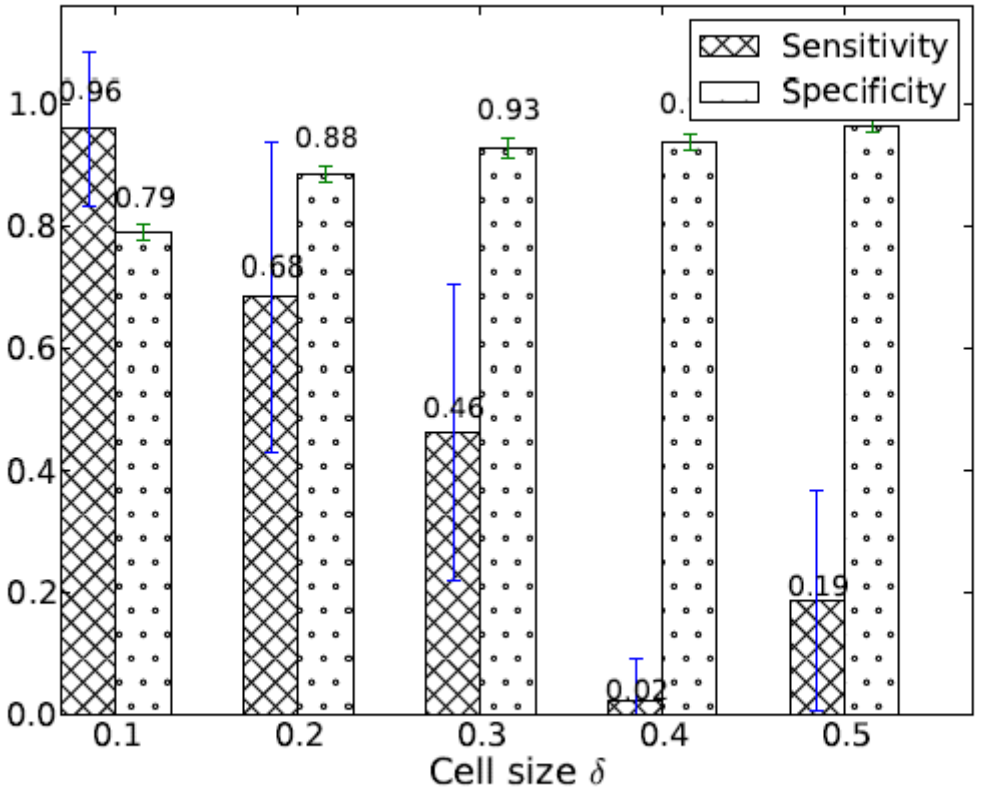


Algorithm

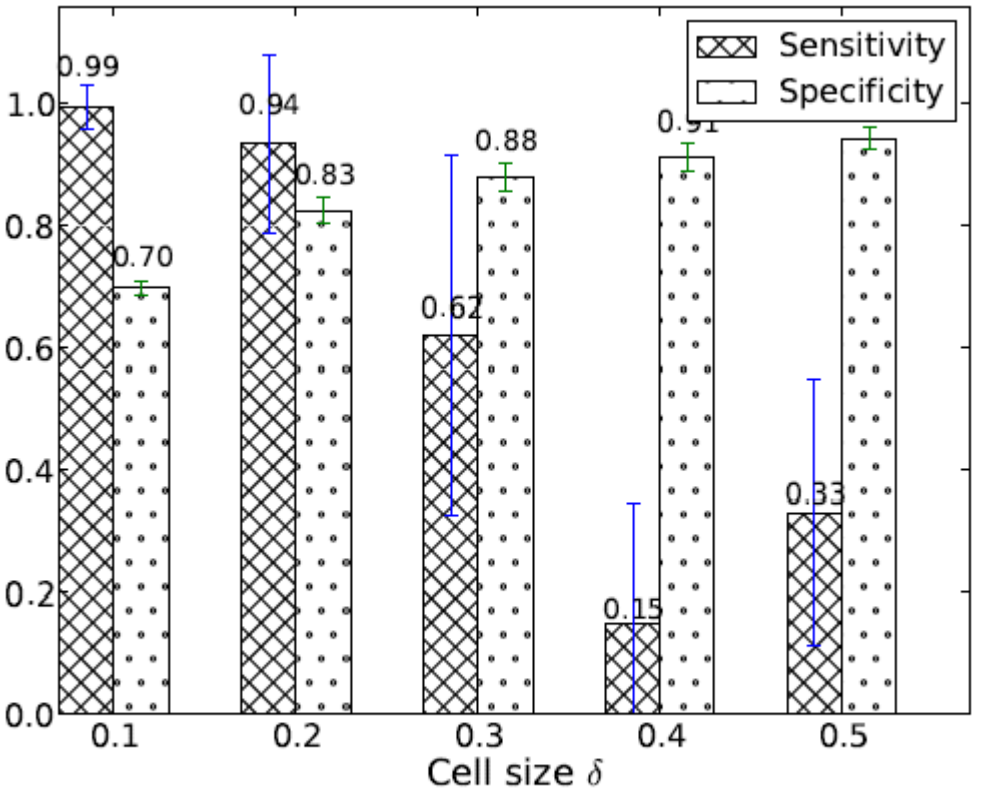




Evaluation



(a) $n = 2$

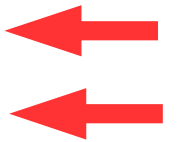


(b) $n = 3$

Evaluation

Table 2.2: Comparative study of our VS based algorithm with other detection algorithms.

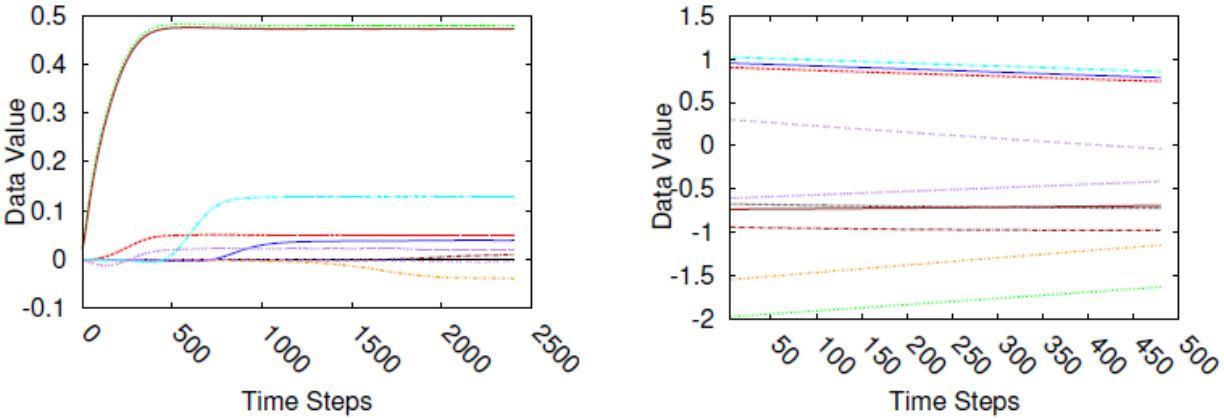
	Sensitivity	Specificity	S-measure
NB	0.493	0.857	0.625
SVM-RBF ($\mu = 0.1, \gamma = 10^{-3}$)	1	0.458	0.628
ANN	0	1	0
K-MEANS ($k = 2$)	1	0.516	0.681
VS ($\delta = 0.2, n = 3$)	0.935	0.825	0.877
VS ($\delta = 0.1, n = 2$)	0.959	0.791	0.867



- What is Silent Data Corruption?
  - SDC is caused by *Soft Errors* undetected by hardware.
  - ***Soft Error**: An unintended change in the state of an electronic device that alters the information that it stores without destroying its functionality, e.g. a bit flip caused by a cosmic-ray-induced neutron[1].*

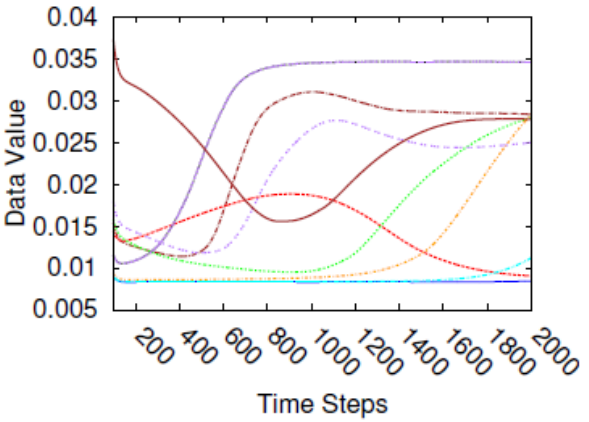
1. Hengartner et. al., “Evaluating Experiments for Estimating the Bit Failure Cross-Section of Semiconductors Using a Colored Spectrum Neutro Beam”, Technometrics, 2008.

2. http://cacm.acm.org/news/169482-addressing-the-threat-of-silent-data-corruption/fulltext

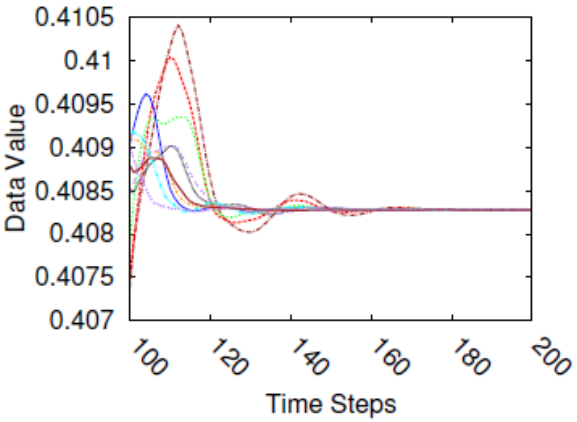


(a) Nek5000-Vortex

(b) Nek5000-Eddy



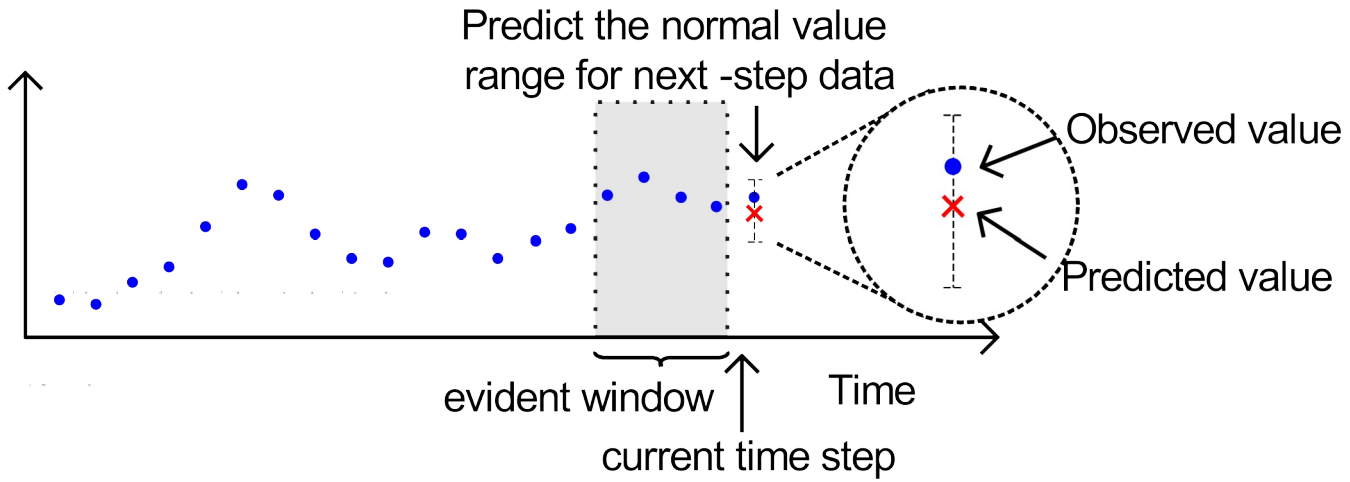
(c) FLASH-Sedov

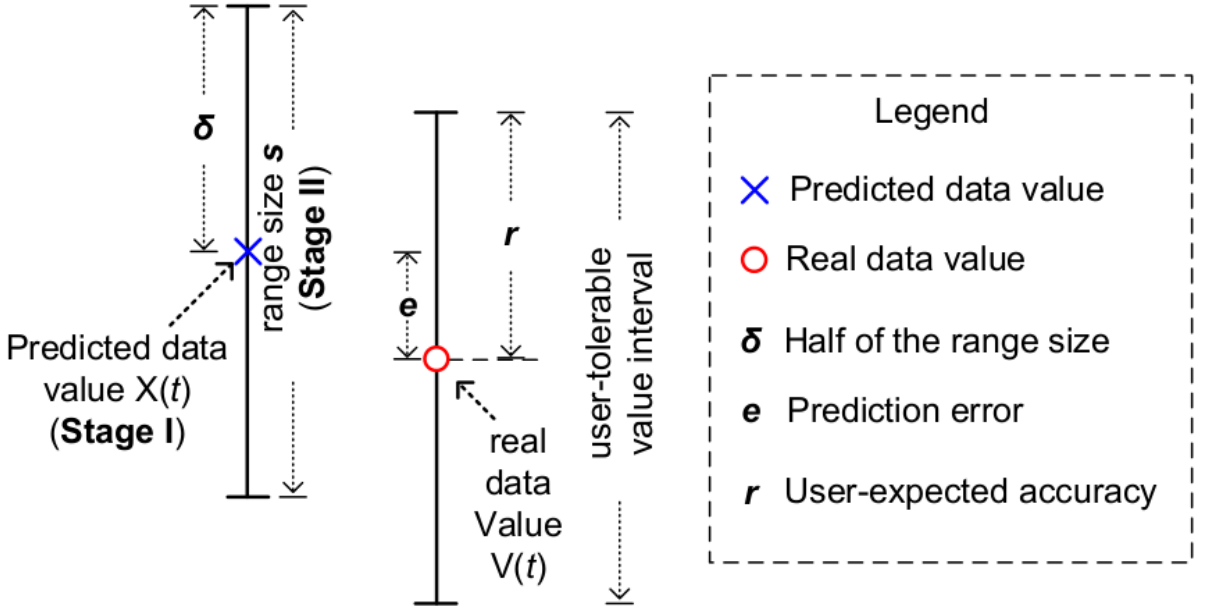


(d) FLASH-Sod

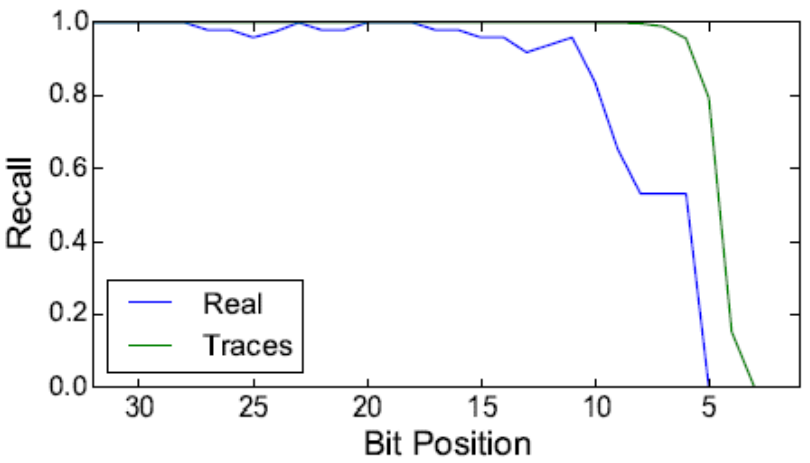
Figure 3.1: Smoothness of numerical simulations from real-world HPC application datasets.

One step ahead prediction and detection model

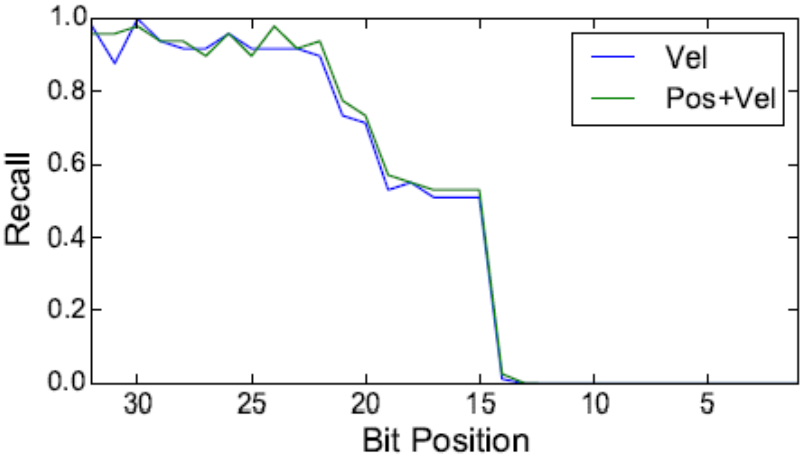




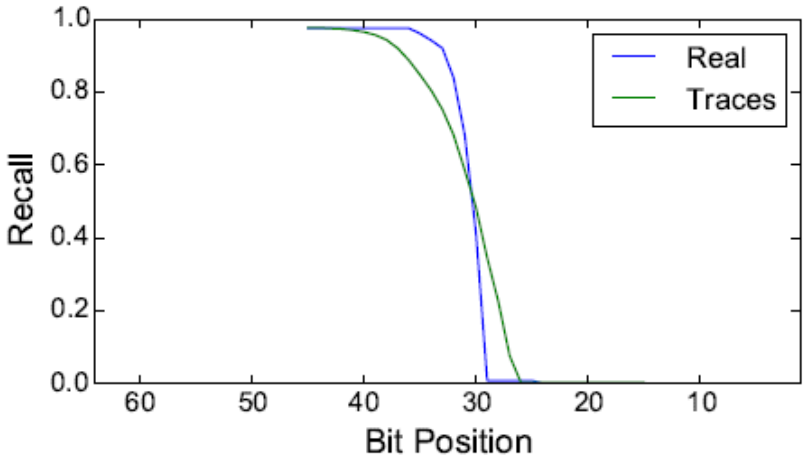
Some results...



(a) HACC (particles' position)

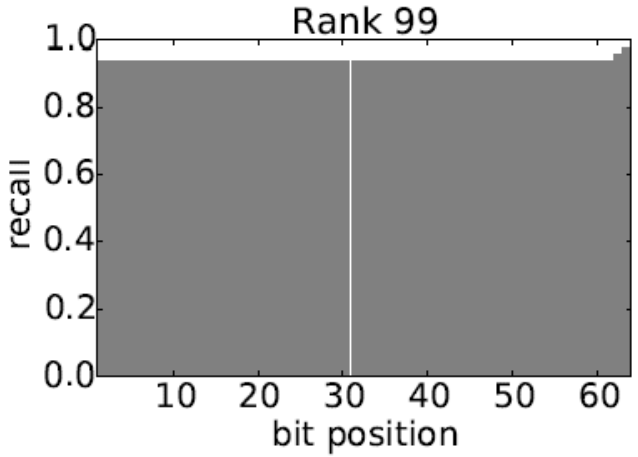


(b) HACC (velocity)

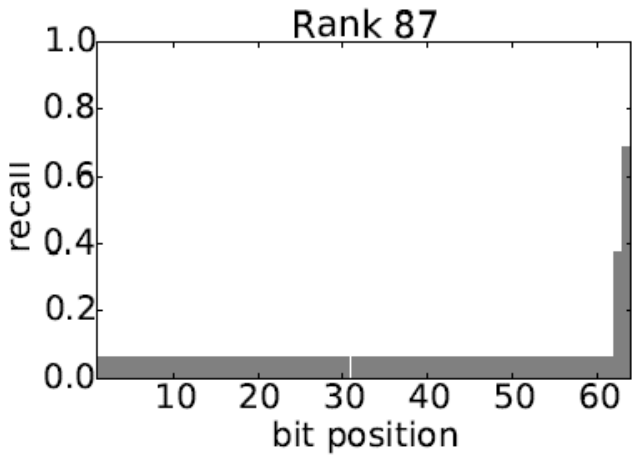


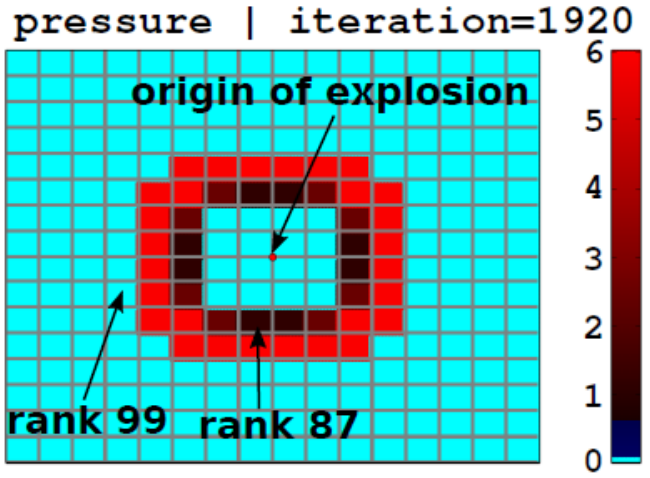
(c) Nek5000 (vortex)

Always smooth?



(a)



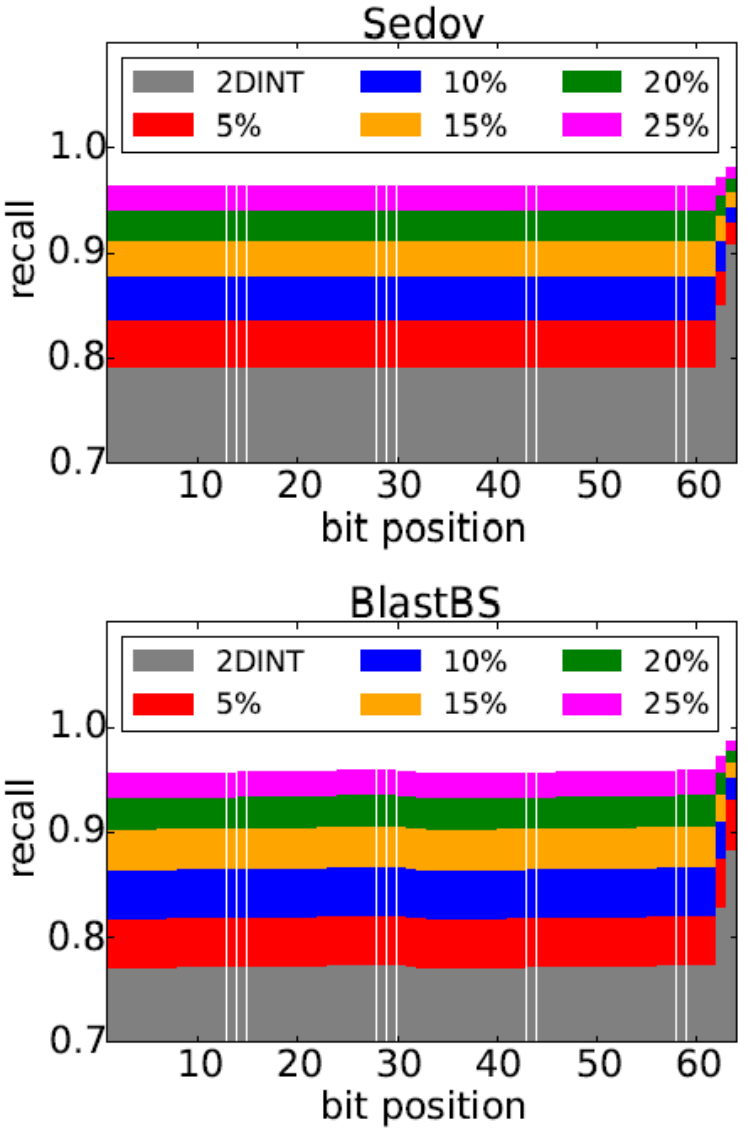


(c)

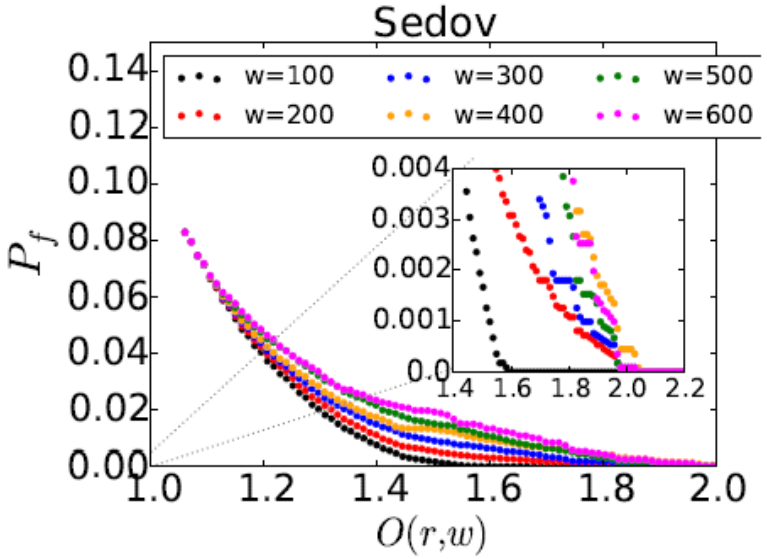
Adaptive Method

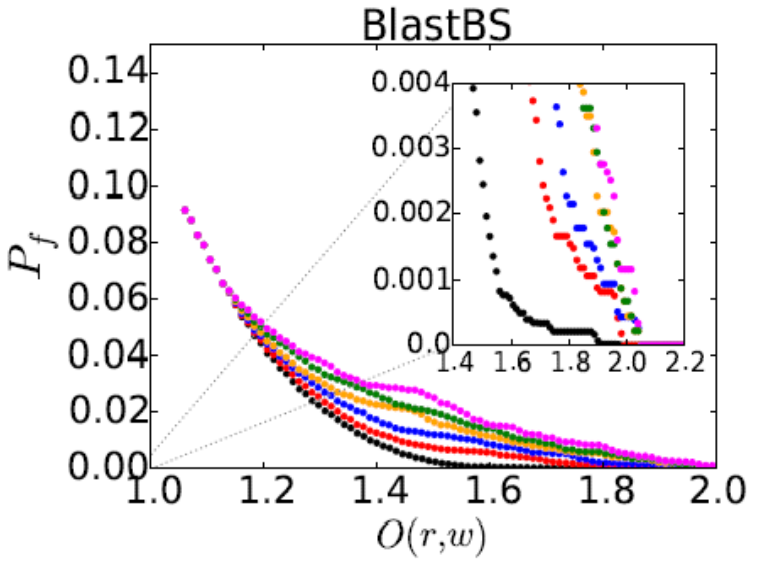
- Full replication (2x, 3x, ...)
  - Deterministic only
  - 100% recall & precision
  - Problem: too expensive
- Data-Analytic-Based
  - Lightweight
  - Requires data too always be smooth
- **Solution**
  - **Combining both**

Evaluation (recall)



Evaluation (Overhead)





Evaluation

Table 4.1: Detection recall and overhead for DAB-only detectors, 2x replication, and the adaptive solution. In the latter, two cases are shown corresponding to two protection levels: 97% and 99.999% recall, respectively.

	Sedov		BlastBS	
	Overhead	*Recall*	*Overhead*	*Recall*
DAB-only	6%	92%	6%	91%
Duplication	110%	100%	110%	100%
Adaptive (case 1)	25%	97%	26%	97%
Adaptive (case 2)	52%	99.999%	56%	99.999%

Future Work

- Elastic replication bugdet
- More applications
- Comprehensive understanding of how much application's variables can be protected

Thanks !