What to do with unpredicted failures?

Slim Bouguerra, Ana Gainaru and Franck Cappello

Joint Lab workshop November 2013-UIUC
Source code: scale_up.c

- Number_of_cores ++ ; // (several Millions)
- Die_shrinking++; // Next generation Xeon Phi on 14 nm.
- Assert(Power < 20 Megawatts); // can not afford the bill
Source code: scale_up.c

- Number_of_cores ++; // (several Millions)
- Die_shrinking++; // Next generation Xeon Phi on 14 nm.
- Assert(Power < 20 Megawatts); // can not afford the bill

IBM's Sequoia

1.25 failure per day

Failure Isn’t An Option, It’s a Certainty !!
Motivations

Main Motivation

Effective and efficient combination between proactive and preventive fault tolerance strategies.
Motivations

Main Motivation
Effective and efficient combination between proactive and preventive fault tolerance strategies.

Target problem
Checkpoint interval selection problem.

Decision:
- Perform or not proactive checkpoint?
- Optimal interval for preventive
Motivations

Main Motivation
Effective and efficient combination between proactive and preventive fault tolerance strategies.

Target problem
Checkpoint interval selection problem.

Objective
Advanced models to shape the relation between the occurrences of failures and the failure prediction mechanisms in HPC.
Problem description

1. Investigating the failure prediction transformations.

Input Failure data:
- Failure rate
- Distribution
- Correlation

Prediction tool

Unpredicted failures:
- Failure rate
- Distribution
- Correlation

Predicted failures:
- Failure rate
- Distribution
- Correlation
Problem description

1. Investigating the failure prediction transformations.

Input Failure data:
- Failure rate
- Distribution
- Correlation

What is the Transformation function??

Prediction tool

Unpredicted failures:
- Failure rate
- Distribution
- Correlation

Predicted failures:
- Failure rate
- Distribution
- Correlation
1. Investigating the failure prediction transformations.

2. How to deal with the unpredicted and predicted failures.

- **Input Failure data:**
  - Failure rate
  - Distribution
  - Correlation

- **Prediction tool**

- **Unpredicted failures:**
  - Failure rate
  - Distribution
  - Correlation

- **Predicted failures:**
  - Failure rate
  - Distribution
  - Correlation
The Results

1. The failure prediction mechanism is scaling filter.

Input Failure data: ✦ Weibull(scale, shape)

Prediction tool

Unpredicted failures: ✦ Weibull(a*scale, shape)

Predicted failures: ✦ Weibull((1-a)*scale, shape)
The Results

1. The failure prediction mechanism is scaling filter.
2. Correlation between failures isn’t bad news and it helps to improve the recall.
The Results

1. The failure prediction mechanism is scaling filter.
2. Correlation between failures isn't bad news and it helps to improve the recall.
3. The failure prediction mechanism catches the noise (correlations) in data (Easier to infer mathematical models).
The Results

1. The failure prediction mechanism is scaling filter.
2. Correlation between failures isn’t bad news and it helps to improve the recall.
3. The failure prediction mechanism catches the noise (correlations) in data (Easier to infer mathematical models).
4. Combing proactive and preventive checkpointing leads to an improvement of 12% to 30% of the amount of useful work.
Outline

1. Failure prediction terminology and concepts
2. Data source and characteristics
3. Modeling and fitting methodology
4. Study case
5. Conclusion and future work
Let’s remember ELSA

Blue Waters

Log DB

ELSA
Let’s remember ELSA
Online failure prediction terminology

Terminology

- True positive alert (correct prediction)
- False positive alert (misleading prediction)
- False negative alert (unpredicted failure)
Online failure prediction terminology

**Terminology**
- True positive alert (correct prediction)
- False positive alert (misleading prediction)
- False negative alert (unpredicted failure)

**Metric**
- Recall:
  \[
  \frac{\text{#True positive}}{\text{#True positive + #False negative}}
  \]
- Precision:
  \[
  \frac{\text{#True positive}}{\text{#True positive + #False positive}}
  \]
Proactive and preventive fault tolerance

Prediction is feasible

- ELSA: Signal analysis with data mining:
  - 90% precision and 45% recall.
  - At least 10 seconds of lead-time.
  - Failure location is provided.
Proactive and preventive fault tolerance

Prediction is feasible
- ELSA: Signal analysis with data mining:
  - 90% precision and 45% recall.
  - At least 10 seconds of lead-time.
  - Failure location is provided.

Fast checkpointing strategies exist
- FTI (Fault Tolerance Interface):
  - Capable of taking a checkpoint in 5s for 1GB memory.
  - Multi-level checkpoint with 8% overhead.
Outline

1. Failure prediction terminology and concepts
2. Data source and characteristics
3. Modeling and fitting methodology
4. Study case
5. Conclusion and future work
Data characteristics

- 22 High performance computing systems from Los Alamos National Lab.
  - Different architectures and sizes.
  - 433,490 per system.
  - MTBF, 13 to 215 hours.
  - Failures are manually annotated.

BlueGene/L at Lawrence Livermore National Lab.
- 128K PowerPC 440 processors.
- 4,747,963 events.
- MTBF 24h.
- Anomaly detection technique.
Data characteristics

- 22 High performance computing systems from Los Alamos National Lab.
  - Different architectures and sizes.
  - 433,490 per system.
  - MTBF, 13 to 215 hours.
  - Failures are manually annotated.

- BlueGene/L at Lawrence Livermore National Lab.
  - 128K PowerPc 440 processors.
  - 4,747,963 events.
  - MTBF 24h.
  - Anomaly detection technique.
Failure prediction characteristics

22 HPC systems

- Hardware: 62%
- Software: 23%
- Network: 17%
- Memory: 22%
- Midplan switches: 4%
- Nods cards: 16%
- APP IO: 25%
- Misc: 15%
- Facilities: 2%
- Human: 1%
- Miscellaneous: 10%

**BG/L**
Failure prediction characteristics

22 HPC systems

Hardware 62%
Software 23%
Human 2%
Network 2%
Facilities 2%
Misc 10%

BG/L

22 HPC systems

Facilities 38%
Hardware 45%
Human 42%
Network 41%
Software 23%
Misc 15%

BG/L

Nods cards 61%
Midplan switches 41%
Memory 45%
Network 15%
APP IO 62%

Recall and precision of failures

Resilience and reliability of HPC systems

Joint Lab workshop November 2013-UIUC
Outline

1. Failure prediction terminology and concepts
2. Data source and characteristics
3. Modeling and fitting methodology
4. Study case
5. Conclusion and future work
Methodology: Randomness Test

Method:
- Runs test
- Runs up/down test
- Autocorrelation function test (ACF)
## Randomness tests output

**Input data** → **Randomness tests** → **Non-iid Data set** → **iid Data set** → **Failure intervals** → **False negative Intervals**

### Table: Randomness tests P-values

<table>
<thead>
<tr>
<th>System name</th>
<th>Failures</th>
<th>Randomness tests P-values</th>
<th>iid</th>
<th>False negative</th>
</tr>
</thead>
<tbody>
<tr>
<td>BlueGene/L</td>
<td>235</td>
<td>0.11</td>
<td>0.17</td>
<td>129</td>
</tr>
<tr>
<td>LANL Sys 2</td>
<td>1951</td>
<td>0.01</td>
<td>0.17</td>
<td>1177</td>
</tr>
<tr>
<td>LANL Sys 4</td>
<td>294</td>
<td>0.08</td>
<td>0.73</td>
<td>151</td>
</tr>
<tr>
<td>LANL Sys 6</td>
<td>298</td>
<td>0.75</td>
<td>0.42</td>
<td>163</td>
</tr>
<tr>
<td>LANL Sys 9</td>
<td>304</td>
<td>0.51</td>
<td>0.95</td>
<td>158</td>
</tr>
<tr>
<td>LANL Sys 10</td>
<td>63</td>
<td>1.00</td>
<td>0.88</td>
<td>32</td>
</tr>
<tr>
<td>LANL Sys 11</td>
<td>266</td>
<td>0.30</td>
<td>0.03</td>
<td>270</td>
</tr>
<tr>
<td>LANL Sys 12</td>
<td>255</td>
<td>0.01</td>
<td>0.23</td>
<td>172</td>
</tr>
<tr>
<td>LANL Sys 13</td>
<td>194</td>
<td>0.01</td>
<td>0.23</td>
<td>172</td>
</tr>
<tr>
<td>LANL Sys 14</td>
<td>120</td>
<td>0.22</td>
<td>0.72</td>
<td>122</td>
</tr>
<tr>
<td>LANL Sys 15</td>
<td>53</td>
<td>0.01</td>
<td>0.56</td>
<td>154</td>
</tr>
<tr>
<td>LANL Sys 16</td>
<td>304</td>
<td>0.01</td>
<td>0.19</td>
<td>154</td>
</tr>
<tr>
<td>LANL Sys 18</td>
<td>3917</td>
<td>0.04</td>
<td>0.74</td>
<td>123</td>
</tr>
<tr>
<td>LANL Sys 19</td>
<td>3235</td>
<td>0.06</td>
<td>0.36</td>
<td>75</td>
</tr>
<tr>
<td>LANL Sys 20</td>
<td>2400</td>
<td>0.01</td>
<td>0.87</td>
<td>32</td>
</tr>
<tr>
<td>LANL Sys 21</td>
<td>105</td>
<td>0.04</td>
<td>0.98</td>
<td>159</td>
</tr>
<tr>
<td>LANL Sys 22</td>
<td>17</td>
<td>0.01</td>
<td>0.14</td>
<td>1310</td>
</tr>
<tr>
<td>LANL Sys 23</td>
<td>226</td>
<td>0.02</td>
<td>0.01</td>
<td>76</td>
</tr>
<tr>
<td>LANL Sys 24</td>
<td>23</td>
<td>0.32</td>
<td>0.41</td>
<td>129</td>
</tr>
</tbody>
</table>

**Not enough lines**

**slim.bouguerra@gmail.com (INRIA)**

Resilience and reliability of HPC systems

Joint Lab workshop November 2013-UIUC
### Randomness tests output

#### Table: Randomness tests P-values

<table>
<thead>
<tr>
<th>System name</th>
<th>Failures</th>
<th>False negative</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># lines</td>
<td>Runs test</td>
</tr>
<tr>
<td>Blue Gene/L</td>
<td>235</td>
<td>0.11</td>
</tr>
<tr>
<td>LANL Sys 2</td>
<td>1951</td>
<td>0.01</td>
</tr>
<tr>
<td>LANL Sys 3</td>
<td>294</td>
<td>0.08</td>
</tr>
<tr>
<td>LANL Sys 4</td>
<td>298</td>
<td>0.75</td>
</tr>
<tr>
<td>LANL Sys 5</td>
<td>304</td>
<td>0.51</td>
</tr>
<tr>
<td>LANL Sys 6</td>
<td>63</td>
<td>1.00</td>
</tr>
<tr>
<td>LANL Sys 8</td>
<td>436</td>
<td>0.30</td>
</tr>
<tr>
<td>LANL Sys 9</td>
<td>279</td>
<td>0.01</td>
</tr>
<tr>
<td>LANL Sys 10</td>
<td>234</td>
<td>0.22</td>
</tr>
<tr>
<td>LANL Sys 11</td>
<td>266</td>
<td>0.01</td>
</tr>
<tr>
<td>LANL Sys 12</td>
<td>255</td>
<td>0.01</td>
</tr>
<tr>
<td>LANL Sys 13</td>
<td>194</td>
<td>0.04</td>
</tr>
<tr>
<td>LANL Sys 14</td>
<td>120</td>
<td>0.06</td>
</tr>
<tr>
<td>LANL Sys 15</td>
<td>53</td>
<td>0.01</td>
</tr>
<tr>
<td>LANL Sys 16</td>
<td>245</td>
<td>0.04</td>
</tr>
<tr>
<td>LANL Sys 18</td>
<td>3917</td>
<td>0.01</td>
</tr>
<tr>
<td>LANL Sys 19</td>
<td>3235</td>
<td>0.03</td>
</tr>
<tr>
<td>LANL Sys 20</td>
<td>2400</td>
<td>0.01</td>
</tr>
<tr>
<td>LANL Sys 21</td>
<td>105</td>
<td>0.02</td>
</tr>
<tr>
<td>LANL Sys 22</td>
<td>17</td>
<td>not enough</td>
</tr>
<tr>
<td>LANL Sys 23</td>
<td>226</td>
<td>0.32</td>
</tr>
<tr>
<td>LANL Sys 24</td>
<td>23</td>
<td>not enough</td>
</tr>
</tbody>
</table>

---

**slim.bouguerra@gmail.com** (INRIA)  
**Resilience and reliability of HPC systems**  
**Joint Lab workshop November 2013-UIUC**
between inter-arrival time of failures in the LANL traces. Fig-

data and in the online phase it produces a warning if statistical

failures, for example temporal correlations from the training

for each of the analyzed systems. Statistical based methods

enced us to implement a statistical based prediction method,

that describes the inter arrival time between false negative

the failure prediction mechanism that catches the failures with

the optimal interval between checkpoints. One solution could

failures distribution could be estimated, thus there is no way to

We note that for the systems with non true randomness no

a high correlation of order one like for example the system

observation for system 10, again the recall is high and it has

has also the smallest amount of correlations. We have the same

system. Techniques like the one in [6] can only be used for

specifically that the statistical based prediction technique is

technique and so are creating a large number of false positives,

between failures and as a consequence the prediction technique

recall value for the LANL systems.

The results are closely related to the amount of correlation

The lack of correlation between figure 3 and V-A2, influ-

3 the autocorrelation values using the first lag. Horizontal lines

non randomness on the prediction process, we show in figure

3 the autocorrelation values using the first lag. Horizontal lines

the failures intervals overpass the confidence interval. This

Fitting methodology : 

modeling and fitting methodology

runs test

Up/Down test

#lines

Autocorrelation Failures

Autocorrelation False negative

Failures 95% confidence interval

False negative 95% confidence interval

Blue Gene/L Sys 2

Sys 2

Sys 3

Sys 4

Sys 5

Sys 6

Sys 7

Sys 8

Sys 9

Sys 10

Sys 11

Sys 12

Sys 13

Sys 14

Sys 15

Sys 16

Sys 17

Sys 18

Sys 19

Sys 20

Sys 21

Sys 22

Sys 23
The recall ratio is related to the amount of data and in the online phase it produces a warning if statistical tests output 

<table>
<thead>
<tr>
<th>System name</th>
<th>LANL Sys 2</th>
<th>LANL Sys 3</th>
<th>LANL Sys 4</th>
<th>LANL Sys 5</th>
<th>LANL Sys 6</th>
<th>LANL Sys 7</th>
<th>LANL Sys 8</th>
<th>LANL Sys 9</th>
<th>LANL Sys 10</th>
<th>LANL Sys 11</th>
<th>LANL Sys 12</th>
<th>LANL Sys 13</th>
<th>LANL Sys 14</th>
<th>LANL Sys 15</th>
<th>LANL Sys 16</th>
</tr>
</thead>
<tbody>
<tr>
<td>Model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**B. Distribution fitting**

We first investigate the statistical model and the false negative parameter $r$. We note that the parameters $\alpha$ and $\beta$ are the closest fit to the available data. For $\beta$ we have the highest recall and a small correlation value. This confirms also previous studies [14], [28].

**Randomness tests output**

Fig. 3. First Lag autocorrelation coefficients

**TABLE:**

<table>
<thead>
<tr>
<th>System name</th>
<th>Sys 2</th>
<th>Sys 3</th>
<th>Sys 4</th>
<th>Sys 5</th>
<th>Sys 6</th>
<th>Sys 7</th>
<th>Sys 8</th>
<th>Sys 9</th>
<th>Sys 10</th>
<th>Sys 11</th>
<th>Sys 12</th>
<th>Sys 13</th>
<th>Sys 14</th>
<th>Sys 15</th>
<th>Sys 16</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>2400</td>
<td>1951</td>
<td>226</td>
<td>53</td>
<td>436</td>
<td>234</td>
<td>266</td>
<td>32</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>$\beta$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.6</td>
<td>0.74</td>
<td>0.56</td>
<td>0.72</td>
<td>0.53</td>
<td>0.48</td>
<td>0.92</td>
</tr>
<tr>
<td>$\text{CV}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Negative log likelihood value produced by the MLE to rank**

**VI reports the fitting results concerning the first set of data**

**Randomness test for false negative intervals. We use the second**

**VI the best fitted distribution for the data concerning the**

**exponential case or strictly decreasing failure rate since all the**

**variables, its CV should be close to one as well. For**

**weibull are the closest fit to the available data.**

**Randomness on the prediction process, we show in figure**

**that can be used to describe formally the inter arrival time**

**inter-arrival time of failures in the LANL traces. Fig-**

**failure events and then using the obtained characteristics for**

**periodic patterns and correlation we can infer statistical models**

**failures distribution could be estimated, thus there is no way to**

**model to describe mathematically the arrival time of failures.**

**we can still get high recall even if the trace does not contain**

**important finding. The recall ratio is related to the amount**

**expect that the bigger the size is the higher is the correlation**

**that those clusters have the same hardware type denoted by the**

**truly randomly distributed. The important finding is, thanks to**

**truly randomness behavior lead to a false negative samples**

**use fault tolerance strategies like [5], [33], [1], [21] to compute**

**six that has the highest recall and a small correlation value.**

**observation for system 10, again the recall is high and it has**

**has also the smallest amount of correlations. We have the same**

**different computing systems. This figure point out another**

**important finding. The recall ratio is related to the amount**

**of data and in the online phase it produces a warning if statistical**

**technique and so are creating a large number of false positives,**

**false correlations that are later used for false predictions. For**

**decreasing the overall precision. This is the case of system**

**have a strong correlation are misinterpret by the prediction**

**recall value for the LANL systems.**

**Escape this problem, we decided to only keep very strong**

**systems that have string correlations and as a results the gain**

**this is not true. In fact the smallest clusters 9, 10 and 12**

**behavior are relevant to be used to find a good probability**

**IV the best fitted distribution for the data concerning the**

**among this set of systems the correlation is**

**Kolmogorov-Smirnov [24] test and the standard probability-**

**synthetic sample. Literature describes dozens of goodness-of-**

**most common methodology is to first select a set of candi-**

**Kolmogorov-Smirnov [24] test and the standard probability-**

**distribution function. In fact as it can been seen in table**

**weibull are the closest fit to the available data.**

**Randomness tests output**

**TABLE:**

<table>
<thead>
<tr>
<th>System name</th>
<th>Sys 2</th>
<th>Sys 3</th>
<th>Sys 4</th>
<th>Sys 5</th>
<th>Sys 6</th>
<th>Sys 7</th>
<th>Sys 8</th>
<th>Sys 9</th>
<th>Sys 10</th>
<th>Sys 11</th>
<th>Sys 12</th>
<th>Sys 13</th>
<th>Sys 14</th>
<th>Sys 15</th>
<th>Sys 16</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\alpha$</td>
<td>2400</td>
<td>1951</td>
<td>226</td>
<td>53</td>
<td>436</td>
<td>234</td>
<td>266</td>
<td>32</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>$\beta$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.6</td>
<td>0.74</td>
<td>0.56</td>
<td>0.72</td>
<td>0.53</td>
<td>0.48</td>
<td>0.92</td>
</tr>
<tr>
<td>$\text{CV}$</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Methodology: Fitting

Method:
- Maximum Likelihood Estimation (MLE)

Target Distributions: Exponential, Weibull, Log-normal and Gamma.
Fitting output

Table: Statistical Fitting false negative random

<table>
<thead>
<tr>
<th>System name</th>
<th>False negative</th>
<th>Mean</th>
<th>CV</th>
<th>Best Fit</th>
<th>KS</th>
</tr>
</thead>
<tbody>
<tr>
<td>LANL Sys 8</td>
<td>weibull a = 401499 b = 0.767798</td>
<td>7859.6</td>
<td>1.4</td>
<td></td>
<td>0.74</td>
</tr>
<tr>
<td>LANL Sys 10</td>
<td>weibull a = 318087 b = 0.647838</td>
<td>8247.0</td>
<td>3.6</td>
<td></td>
<td>0.29</td>
</tr>
<tr>
<td>LANL Sys 11</td>
<td>weibull a = 232647 b = 0.609348</td>
<td>6353.5</td>
<td>3.0</td>
<td></td>
<td>0.61</td>
</tr>
<tr>
<td>LANL Sys 13</td>
<td>lognormal µ = 11.5257 σ = 1.87004</td>
<td>8164.3</td>
<td>3.9</td>
<td></td>
<td>0.14</td>
</tr>
<tr>
<td>LANL Sys 14</td>
<td>weibull a = 391931 b = 0.559039</td>
<td>11351.0</td>
<td>2.5</td>
<td></td>
<td>0.77</td>
</tr>
<tr>
<td>LANL Sys 15</td>
<td>exponential µ = 728203</td>
<td>12136.7</td>
<td>1.2</td>
<td></td>
<td>0.17</td>
</tr>
<tr>
<td>LANL Sys 16</td>
<td>weibull a = 182624 b = 0.810939</td>
<td>3430.6</td>
<td>1.3</td>
<td></td>
<td>0.69</td>
</tr>
<tr>
<td>LANL Sys 18</td>
<td>lognormal µ = 10.1123 σ = 1.28677</td>
<td>818.6</td>
<td>1.5</td>
<td></td>
<td>0.37</td>
</tr>
<tr>
<td>LANL Sys 19</td>
<td>exponential µ = 29000.5</td>
<td>863.6</td>
<td>1.4</td>
<td></td>
<td>0.18</td>
</tr>
<tr>
<td>LANL Sys 21</td>
<td>lognormal µ = 10.6382 σ = 1.46402</td>
<td>1986.9</td>
<td>2.3</td>
<td></td>
<td>0.85</td>
</tr>
</tbody>
</table>
### Fitting output

Table: Statistical fitting all random (fitting parameters scale are in seconds)

<table>
<thead>
<tr>
<th>System name</th>
<th>Failures Mean</th>
<th>CV</th>
<th>Best Fit</th>
<th>KS</th>
<th>False negative Mean</th>
<th>CV</th>
<th>Best Fit</th>
<th>KS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue Gene/L</td>
<td>1040.5</td>
<td>0.92</td>
<td>exponential $\mu = 62431.3$</td>
<td>0.10</td>
<td>1888.1</td>
<td>1.10</td>
<td>exponential $\mu = 113289$</td>
<td>0.79</td>
</tr>
<tr>
<td>LANL Sys 3</td>
<td>3595.1</td>
<td>1.1</td>
<td>exponential $\mu = 215705$</td>
<td>0.98</td>
<td>6559.0</td>
<td>1.1</td>
<td>exponential $\mu = 393538$</td>
<td>0.70</td>
</tr>
<tr>
<td>LANL Sys 4</td>
<td>3409.1</td>
<td>1.1</td>
<td>exponential $\mu = 204544$</td>
<td>0.77</td>
<td>6187.0</td>
<td>1.1</td>
<td>exponential $\mu = 371218$</td>
<td>0.99</td>
</tr>
<tr>
<td>LANL Sys 5</td>
<td>3294.5</td>
<td>1.1</td>
<td>exponential $\mu = 197671$</td>
<td>0.95</td>
<td>6377.9</td>
<td>1.2</td>
<td>exponential $\mu = 382671$</td>
<td>0.35</td>
</tr>
<tr>
<td>LANL Sys 6</td>
<td>16796.7</td>
<td>0.9</td>
<td>exponential $\mu = 1007800$</td>
<td>0.81</td>
<td>31878.2</td>
<td>1.1</td>
<td>exponential $\mu = 1912690$</td>
<td>0.99</td>
</tr>
<tr>
<td>LANL Sys 23</td>
<td>9288.2</td>
<td>1.3</td>
<td>weibull $a = 509380 \ b = 0.846905$</td>
<td>0.97</td>
<td>16272.3</td>
<td>1.2</td>
<td>weibull $a = 895274 \ b = 0.851258$</td>
<td>0.98</td>
</tr>
</tbody>
</table>

Table: Statistical Fitting false negative random

<table>
<thead>
<tr>
<th>System name</th>
<th>False negative Mean</th>
<th>CV</th>
<th>Best Fit</th>
<th>KS</th>
</tr>
</thead>
<tbody>
<tr>
<td>LANL Sys 8</td>
<td>7859.6</td>
<td>1.4</td>
<td>weibull $a = 401499 \ b = 0.767798$</td>
<td>0.74</td>
</tr>
<tr>
<td>LANL Sys 10</td>
<td>8247.0</td>
<td>3.6</td>
<td>weibull $a = 318087 \ b = 0.647838$</td>
<td>0.29</td>
</tr>
<tr>
<td>LANL Sys 11</td>
<td>6353.5</td>
<td>3.0</td>
<td>weibull $a = 232647 \ b = 0.609348$</td>
<td>0.61</td>
</tr>
<tr>
<td>LANL Sys 13</td>
<td>8164.3</td>
<td>3.9</td>
<td>lognormal $\mu = 11.5257 \ \sigma = 1.87004$</td>
<td>0.14</td>
</tr>
<tr>
<td>LANL Sys 14</td>
<td>11351.0</td>
<td>2.5</td>
<td>weibull $a = 391931 \ b = 0.559039$</td>
<td>0.77</td>
</tr>
<tr>
<td>LANL Sys 15</td>
<td>12136.7</td>
<td>1.2</td>
<td>exponential $\mu = 728203$</td>
<td>0.17</td>
</tr>
<tr>
<td>LANL Sys 16</td>
<td>3430.6</td>
<td>1.3</td>
<td>weibull $a = 182624 \ b = 0.810939$</td>
<td>0.69</td>
</tr>
<tr>
<td>LANL Sys 18</td>
<td>818.6</td>
<td>1.5</td>
<td>lognormal $\mu = 10.1123 \ \sigma = 1.28677$</td>
<td>0.37</td>
</tr>
<tr>
<td>LANL Sys 19</td>
<td>863.6</td>
<td>1.4</td>
<td>exponential $\mu = 29000.5$</td>
<td>0.18</td>
</tr>
<tr>
<td>LANL Sys 21</td>
<td>1986.9</td>
<td>2.3</td>
<td>lognormal $\mu = 10.6382 \ \sigma = 1.46402$</td>
<td>0.85</td>
</tr>
</tbody>
</table>
Methodology: Goodness of fit

Method:
- Kolmogorov-Smirnov test
- Probability-Probability plot (PP-plot).
### Goodness of fit outputs

Distribution for failure
- Intervals + false negative
- false negative only

Data set
- iid

#### Distribution for failure

<table>
<thead>
<tr>
<th>System name</th>
<th>Failures Mean</th>
<th>CV</th>
<th>Best Fit</th>
<th>KS</th>
<th>False negative Mean</th>
<th>CV</th>
<th>Best Fit</th>
<th>KS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blue Gene/L</td>
<td>1040.5</td>
<td>0.92</td>
<td>exponential $\mu = 62431.3$</td>
<td>0.10</td>
<td>1888.1</td>
<td>1.10</td>
<td>exponential $\mu = 113289$</td>
<td>0.79</td>
</tr>
<tr>
<td>LANL Sys 3</td>
<td>3595.1</td>
<td>1.1</td>
<td>exponential $\mu = 215705$</td>
<td>0.98</td>
<td>6559.0</td>
<td>1.10</td>
<td>exponential $\mu = 393538$</td>
<td>0.70</td>
</tr>
<tr>
<td>LANL Sys 4</td>
<td>3409.1</td>
<td>1.1</td>
<td>exponential $\mu = 204544$</td>
<td>0.77</td>
<td>6187.0</td>
<td>1.10</td>
<td>exponential $\mu = 371218$</td>
<td>0.99</td>
</tr>
<tr>
<td>LANL Sys 5</td>
<td>3294.5</td>
<td>1.1</td>
<td>exponential $\mu = 197671$</td>
<td>0.95</td>
<td>6377.9</td>
<td>1.2</td>
<td>exponential $\mu = 382671$</td>
<td>0.35</td>
</tr>
<tr>
<td>LANL Sys 6</td>
<td>16796.7</td>
<td>0.9</td>
<td>exponential $\mu = 1007800$</td>
<td>0.81</td>
<td>31878.2</td>
<td>1.1</td>
<td>exponential $\mu = 1912690$</td>
<td>0.99</td>
</tr>
<tr>
<td>LANL Sys 23</td>
<td>9288.2</td>
<td>1.3</td>
<td>weibull $a = 509380$ $b = 0.846905$</td>
<td>0.97</td>
<td>16272.3</td>
<td>1.2</td>
<td>weibull $a = 895274$ $b = 0.851258$</td>
<td>0.98</td>
</tr>
</tbody>
</table>

#### Joint Lab workshop June 2013

<table>
<thead>
<tr>
<th>System name</th>
<th>False negative Mean</th>
<th>CV</th>
<th>Best Fit</th>
<th>KS</th>
</tr>
</thead>
<tbody>
<tr>
<td>LANL Sys 8</td>
<td>182624</td>
<td>0.81</td>
<td>weibull $a = 182624$ $b = 0.810939$</td>
<td>0.97</td>
</tr>
<tr>
<td>LANL Sys 10</td>
<td>391931</td>
<td>0.56</td>
<td>weibull $a = 391931$ $b = 0.559039$</td>
<td>0.81</td>
</tr>
<tr>
<td>LANL Sys 11</td>
<td>232647</td>
<td>0.61</td>
<td>weibull $a = 232647$ $b = 0.609348$</td>
<td>0.77</td>
</tr>
<tr>
<td>LANL Sys 12</td>
<td>401499</td>
<td>0.77</td>
<td>weibull $a = 401499$ $b = 0.767798$</td>
<td>0.74</td>
</tr>
<tr>
<td>LANL Sys 13</td>
<td>12136.7</td>
<td>1.2</td>
<td>exponential $\mu = 12136.7$</td>
<td>0.17</td>
</tr>
<tr>
<td>LANL Sys 14</td>
<td>7859.6</td>
<td>1.4</td>
<td>weibull $a = 12136.7$ $b = 0.767798$</td>
<td>0.69</td>
</tr>
<tr>
<td>LANL Sys 15</td>
<td>8247.0</td>
<td>3.6</td>
<td>weibull $a = 318087$ $b = 0.647838$</td>
<td>0.77</td>
</tr>
<tr>
<td>LANL Sys 16</td>
<td>6353.5</td>
<td>3.0</td>
<td>weibull $a = 318087$ $b = 0.647838$</td>
<td>0.77</td>
</tr>
<tr>
<td>LANL Sys 17</td>
<td>8164.3</td>
<td>3.9</td>
<td>lognormal $\mu = 11.5257$ $\sigma = 1.87004$</td>
<td>0.14</td>
</tr>
<tr>
<td>LANL Sys 18</td>
<td>11351.0</td>
<td>2.5</td>
<td>weibull $a = 318087$ $b = 0.647838$</td>
<td>0.77</td>
</tr>
<tr>
<td>LANL Sys 19</td>
<td>12136.7</td>
<td>1.2</td>
<td>exponential $\mu = 12136.7$</td>
<td>0.17</td>
</tr>
<tr>
<td>LANL Sys 20</td>
<td>8247.0</td>
<td>3.6</td>
<td>weibull $a = 318087$ $b = 0.647838$</td>
<td>0.77</td>
</tr>
<tr>
<td>LANL Sys 21</td>
<td>836.3</td>
<td>1.4</td>
<td>exponential $\mu = 29000.5$</td>
<td>0.18</td>
</tr>
<tr>
<td>LANL Sys 22</td>
<td>1988.9</td>
<td>2.3</td>
<td>lognormal $\mu = 10.6382$ $\sigma = 1.46402$</td>
<td>0.85</td>
</tr>
</tbody>
</table>

---

**Resilience and reliability of HPC systems**

**Joint Lab workshop November 2013-UIUC**
Goodness of fit outputs

Data set

System name | Failures | False negative
--- | --- | ---
| Mean | CV | Best Fit | Mean | CV | Best Fit | KS
Blue Gene/L | 1040.5 | 0.92 | exponential $\mu = 62431.3$ | 0.10 | 1888.1 | 1.10 | exponential $\mu = 113289$ | 0.79
LANL Sys 3 | 3595.1 | 1.1 | exponential $\mu = 215705$ | 0.98 | 6559.0 | 1.1 | exponential $\mu = 393538$ | 0.70
LANL Sys 4 | 3409.1 | 1.1 | exponential $\mu = 204544$ | 0.77 | 6187.0 | 1.1 | exponential $\mu = 371218$ | 0.99
LANL Sys 5 | 3294.5 | 1.1 | exponential $\mu = 197671$ | 0.95 | 6377.9 | 1.2 | exponential $\mu = 382671$ | 0.99
LANL Sys 6 | 16796.7 | 0.9 | exponential $\mu = 1007800$ | 0.81 | 31878.2 | 1.1 | exponential $\mu = 1912690$ | 0.99
LANL Sys 23 | 9288.2 | 1.3 | weibull $a = 509380$ $b = 0.846905$ | 0.97 | 16272.3 | 1.2 | weibull $a = 895374$ $b = 0.851258$ | 0.98

System name | False negative
--- | ---
| Mean | CV | Best Fit | KS
LANL Sys 8 | 31859.6 | 1.4 | weibull $a = 401499$ $b = 0.767798$ | 0.74
LANL Sys 10 | 8247.0 | 3.6 | weibull $a = 318087$ $b = 0.647838$ | 0.29
LANL Sys 11 | 6353.5 | 3.0 | weibull $a = 232647$ $b = 0.609348$ | 0.61
LANL Sys 13 | 8164.3 | 3.9 | lognormal $\mu = 11.5257$ $\sigma = 1.87004$ | 0.14
LANL Sys 14 | 11351.0 | 2.5 | weibull $a = 391931$ $b = 0.559039$ | 0.77
LANL Sys 15 | 12136.7 | 1.2 | exponential $\mu = 1912690$ | 0.17
LANL Sys 16 | 3430.6 | 1.3 | weibull $a = 182624$ $b = 0.810939$ | 0.69
LANL Sys 18 | 818.6 | 1.5 | lognormal $\mu = 10.1123$ $\sigma = 1.28677$ | 0.37
LANL Sys 19 | 863.6 | 1.4 | exponential $\mu = 204544$ | 0.18
LANL Sys 21 | 1988.9 | 2.3 | lognormal $\mu = 10.6382$ $\sigma = 1.46402$ | 0.85

Resilience and reliability of HPC systems
Outline

1. Failure prediction terminology and concepts
2. Data source and characteristics
3. Modeling and fitting methodology
4. Study case
5. Conclusion and future work
Mathematical Modeling: proposed combination

Preventive checkpoints

Prediction

$t_a$

Execution Time

$W_p = p(R + c_l + l_c^2) + p c_t^2$

$W_{np} = p(R + t_a + l)$

The proactive action is performed if $W_p < W_{np} < p c_t^2 / p$.

slim.bouguerra@gmail.com (INRIA)
Mathematical Modeling: proposed combination

**Preventive checkpoints**

**Prediction**

\[ W_p = p(R + c_2 + l_c^2) + p c_2^2 \]

**To ignore**

\[ W_{np} = p(R + t_a + l) \]

The proactive action is performed if

\[ W_p \leq W_{np} \]

\[ \frac{p c_2^2}{p} \leq t_a \]

Failure happens with a probability \( p \)
Mathematical Modeling (Proactive decision)

The proactive action is performed iff

\[ W_p \leq W_{np} \equiv \bar{p}c_2/p \leq t_a \]
Mathematical Modeling

Preventive period

- Unpredicted failures are randomly distributed with a mean $\mu$.
- The preventive checkpoint cost $c_1$. 
Mathematical Modeling

Preventive period

- Unpredicted failures are randomly distributed with a mean $\mu$.
- The preventive checkpoint cost $c_1$.

The first order approximation of the interval between preventive checkpoints:

$$\sqrt{2\mu c_1}$$
Simulation results

System 19 LANL actual failures data and prediction.

- More than 3,000 failures and 1,700 unpredicted failures.
- 45% recall and 90% precision.
Simulation results

System 19 LANL actual failures data and prediction.

- More than 3,000 failures and 1,700 unpredicted failures.
- 45% recall and 90% precision.

13% of improvement which is the theoretical peak for such configuration.
Outline

1. Failure prediction terminology and concepts
2. Data source and characteristics
3. Modeling and fitting methodology
4. Study case
5. Conclusion and future work
Conclusion

- Classification based on the randomness tests (iid vs non-iid)
Conclusion

- Classification based on the randomness tests (iid vs non-iid)
- Most of the available failure traces are not random (Can not be used to infer probability distributions)
Conclusion

- Classification based on the randomness tests (iid vs non-iid)
- Most of the available failure traces are not random (Can not be used to infer probability distributions)
- Failure prediction mechanism catches the non-randomness and correlation.
Conclusion

- Classification based on the randomness tests (iid vs non-iid)
- Most of the available failure traces are not random (Can not be used to infer probability distributions)
- Failure prediction mechanism catches the non-randomness and correlation.
- Failure prediction mechanism acts as a scale function and it affects only the scale parameter.
Conclusion

- Classification based on the randomness tests (iid vs non-iid)
- Most of the available failure traces are not random (Can not be used to infer probability distributions)
- Failure prediction mechanism catches the non-randomness and correlation.
- Failure prediction mechanism acts as a scale function and it affects only the scale parameter.
- The peak of correlation on the initial traces has an important impact on the prediction results, specifically on the recall value.
Future Work

- Analyze more deeply the set of systems with a high correlation like system 2 or 20 and isolate sources of non-randomness.
Future Work

- Analyze more deeply the set of systems with a high correlation like system 2 or 20 and isolate sources of non-randomness.
- Investigate if a cross-correlation of different time scale has an impact of the prediction mechanism.
Future Work

- Analyze more deeply the set of systems with a high correlation like system 2 or 20 and isolate sources of non-randomness.
- Investigate if a cross-correlation of different time scale has an impact of the prediction mechanism.
- Manage the tradeoff between the precision and the recall.
Questions?