Challenges in predicting failures on the Blue Waters system

Ana Gainaru

Franck Cappello, Marc Snir, Bill Kramer

Goal

- Online failure prediction
  - On Blue Waters

So far

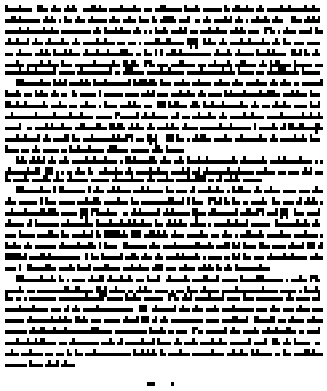
- Data mining
  - Different correlation methods
    - Best result: 80% precision and 70% recall
- Signal analysis
  - Used for anomaly detection
    - 87% precision and 51% recall (IPDPS12)
- Hybrid
  - ELSA: Signal analysis with data mining
    - 90% precision and 45% recall (SC12)
    - At least 10 seconds delay

Failure prediction

- Training
  - Between 3-5 months
- 3 training phases: Parameters influence results

Failure prediction

- HELO



List of templates

* errors detected and corrected

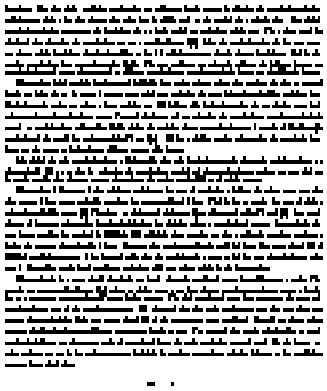
node card * is not fully functional

…

vm: killing process %s n+

Failure prediction

- HELO



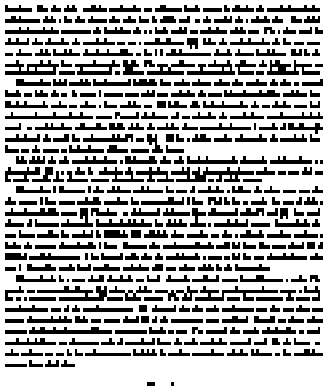
Parameter

- **Cluster goodness**
- High values

10 errors detected and corrected
2 errors detected and corrected

Failure prediction

- HELO



Parameter

- **Cluster goodness**
- Low values

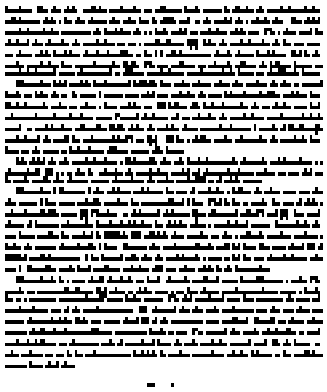
vm: * process %s n+

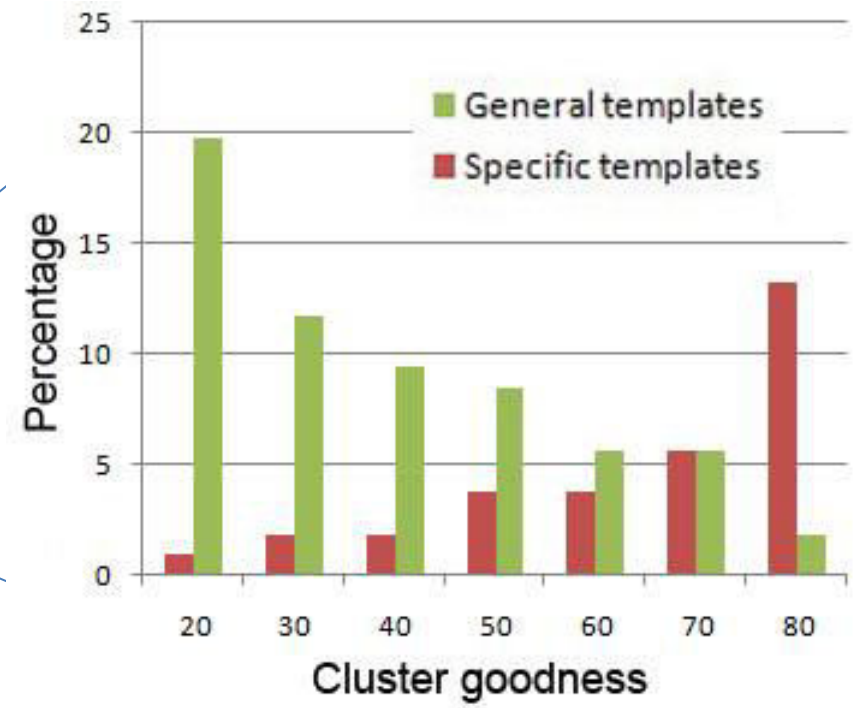
vm: killing process %s n+

vm: starting process %s n+

Failure prediction

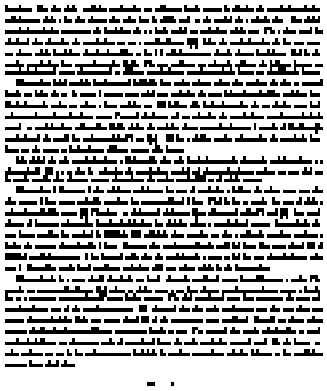
- HELO

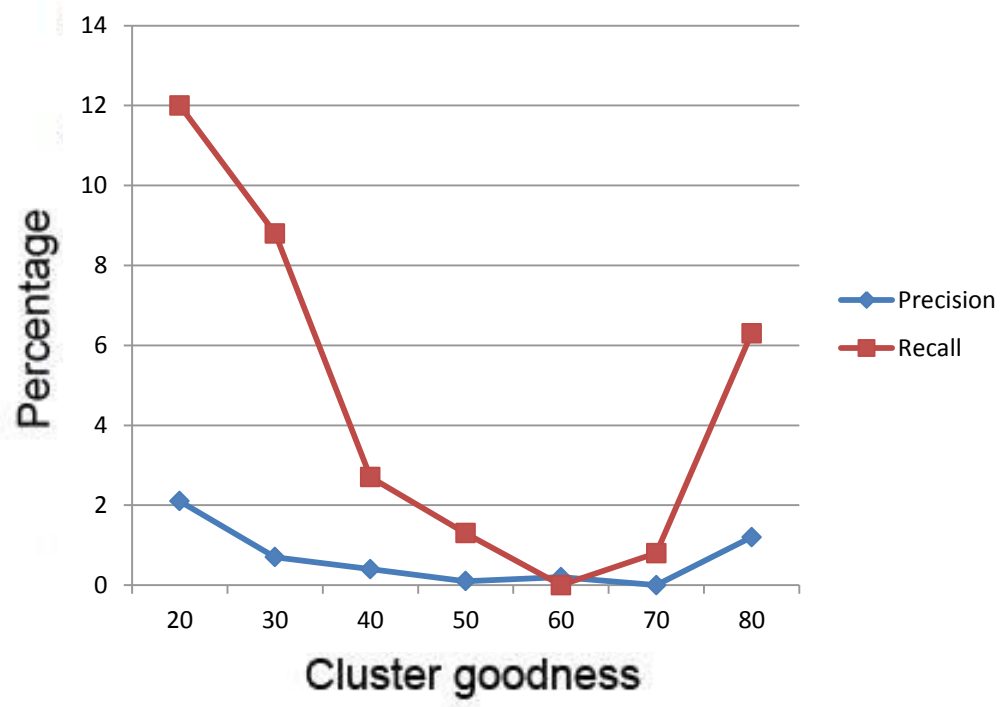




Failure prediction

- HELO



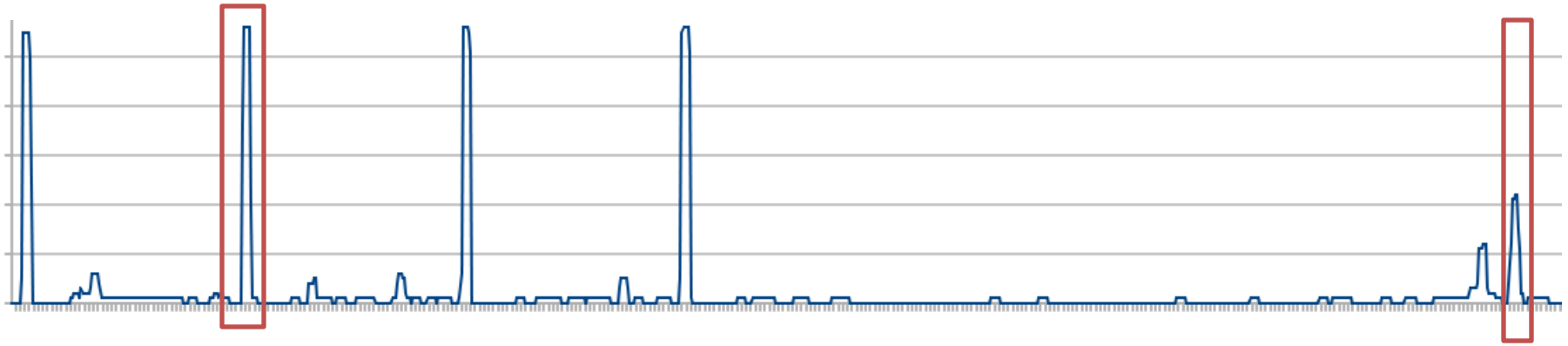


Failure prediction

- ELSA

Parameter

- **Anomaly sensibility**

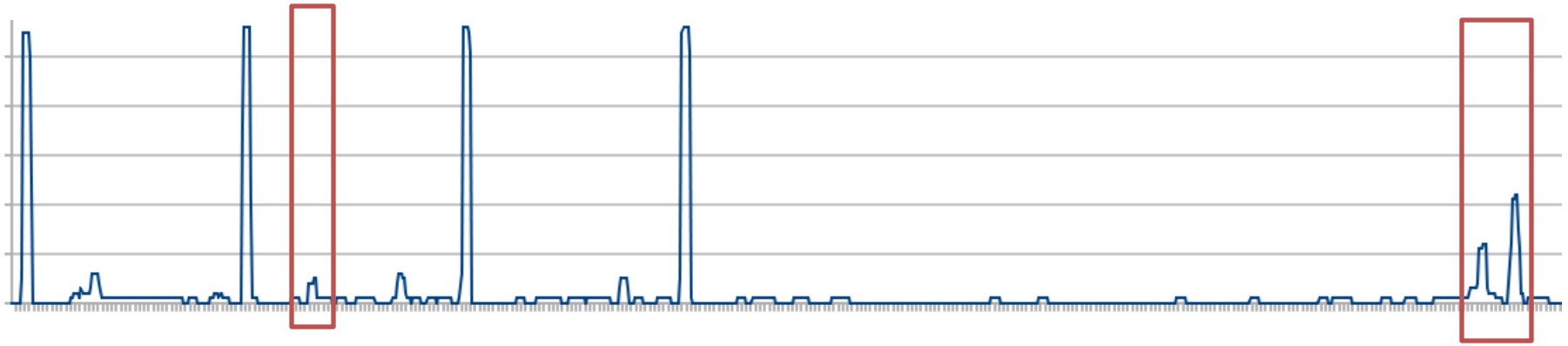


Failure prediction

- ELSA

Parameter

- **Anomaly sensibility**

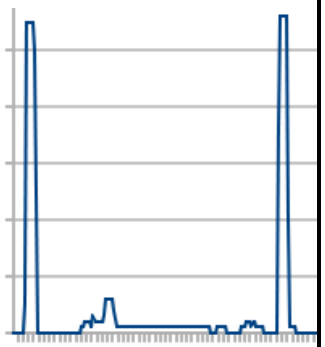


Failure prediction

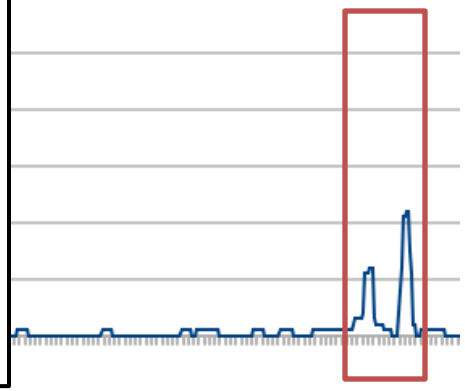
- ELSA

Parameter

- **Anomaly sensibility**



- Recall varies with 15-18%
- Precision varies with ~10%
- For too many false neg/pos
  - No correlation parameter gives good results

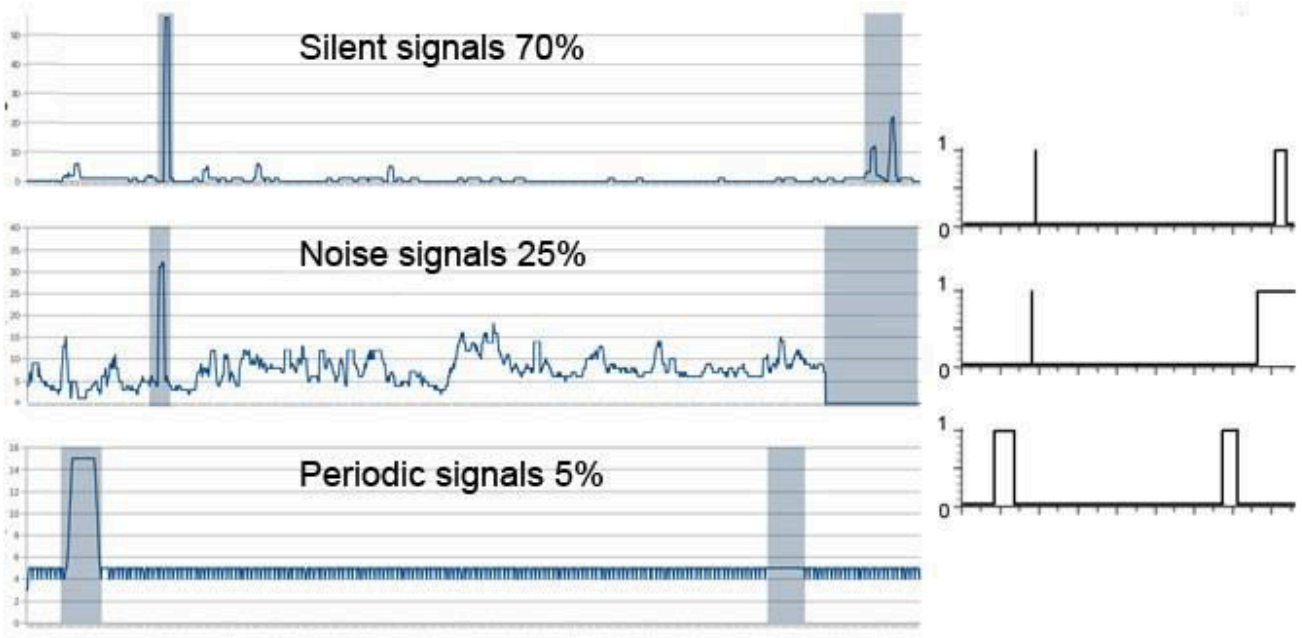


Failure prediction

- ELSA

Parameter

- **Correlation parameters**

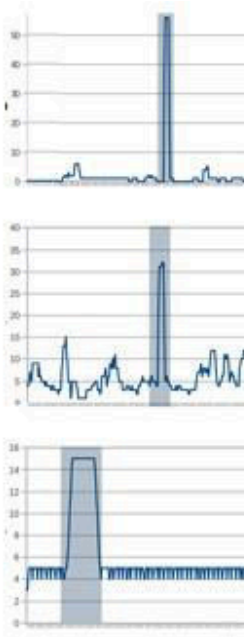


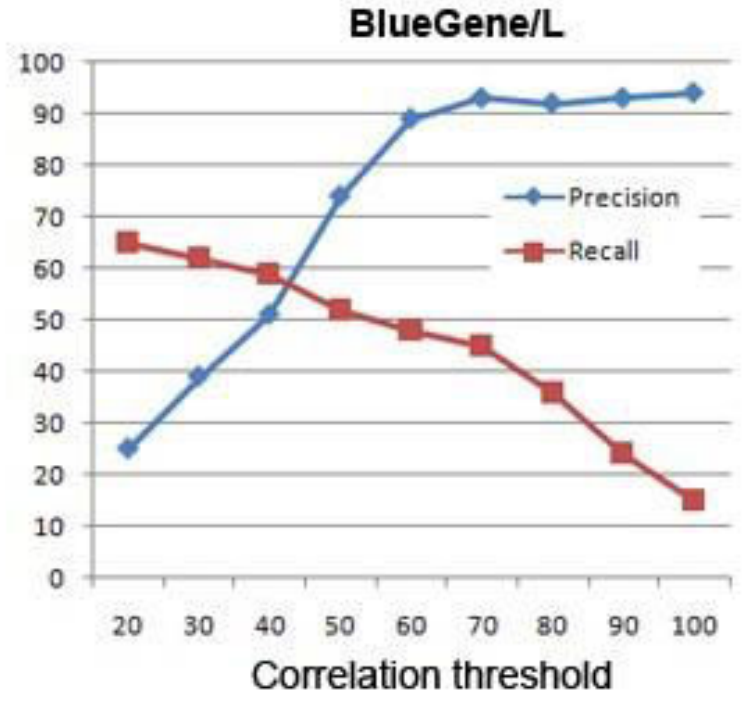
Failure prediction

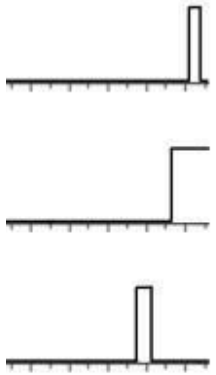
- ELSA

Parameter

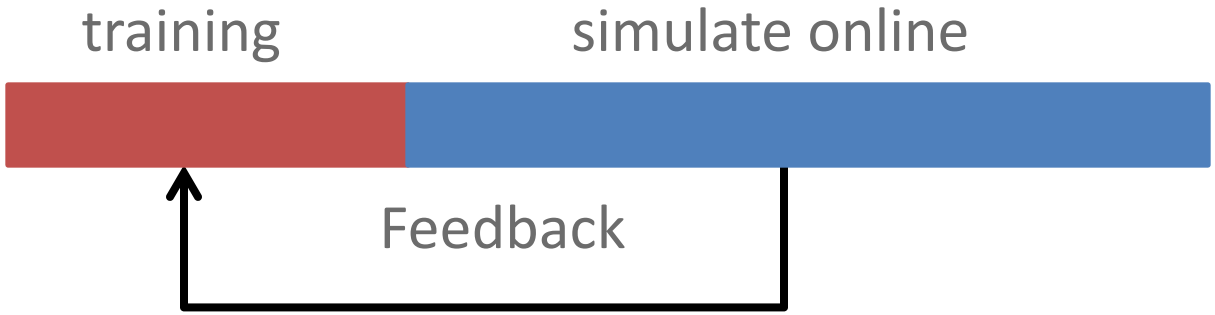
- **Correlation parameters**





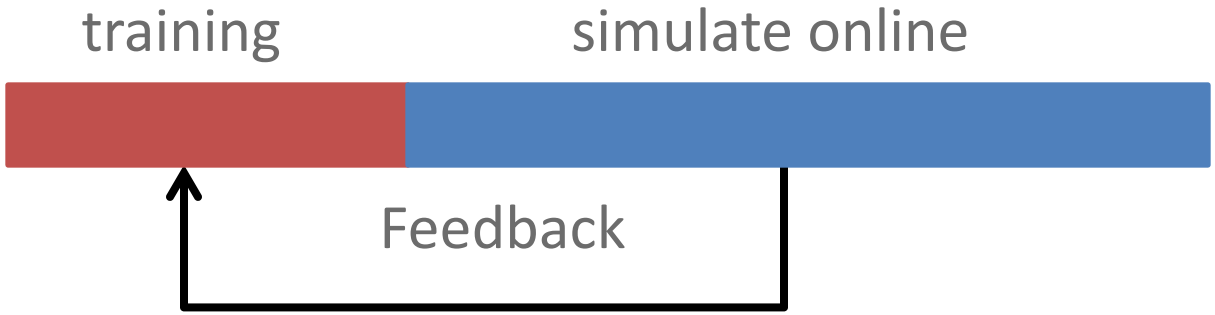


Failure prediction

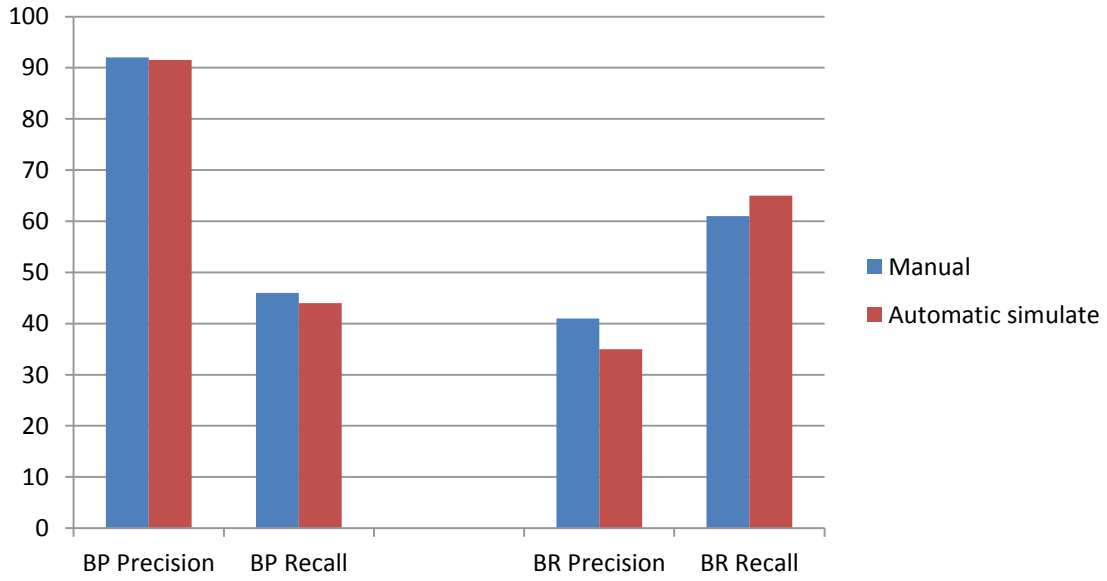


- Simulate online
  - Redo the analysis based on the results
- Automatic tuning of the parameters
  - To get max precision/recall/given relation
  - Brute force

Failure prediction

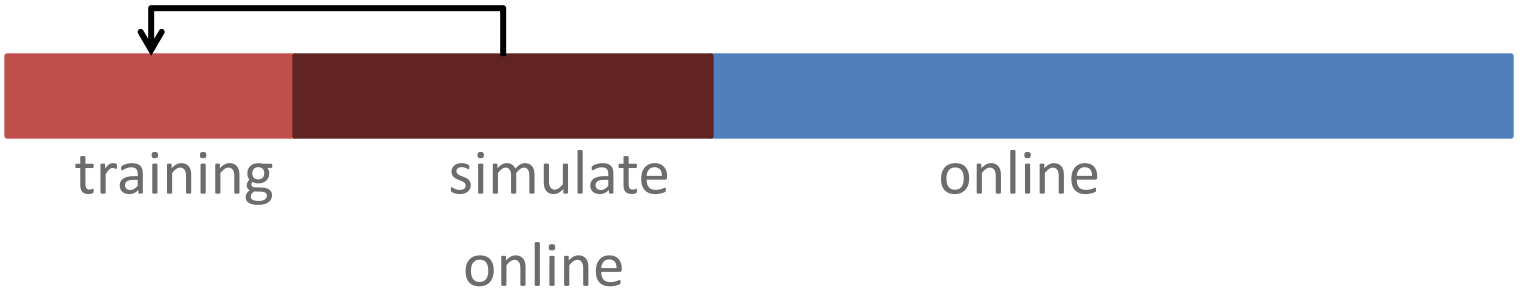


- Simul
  - Red
- Autor
  - To g
  - Bru



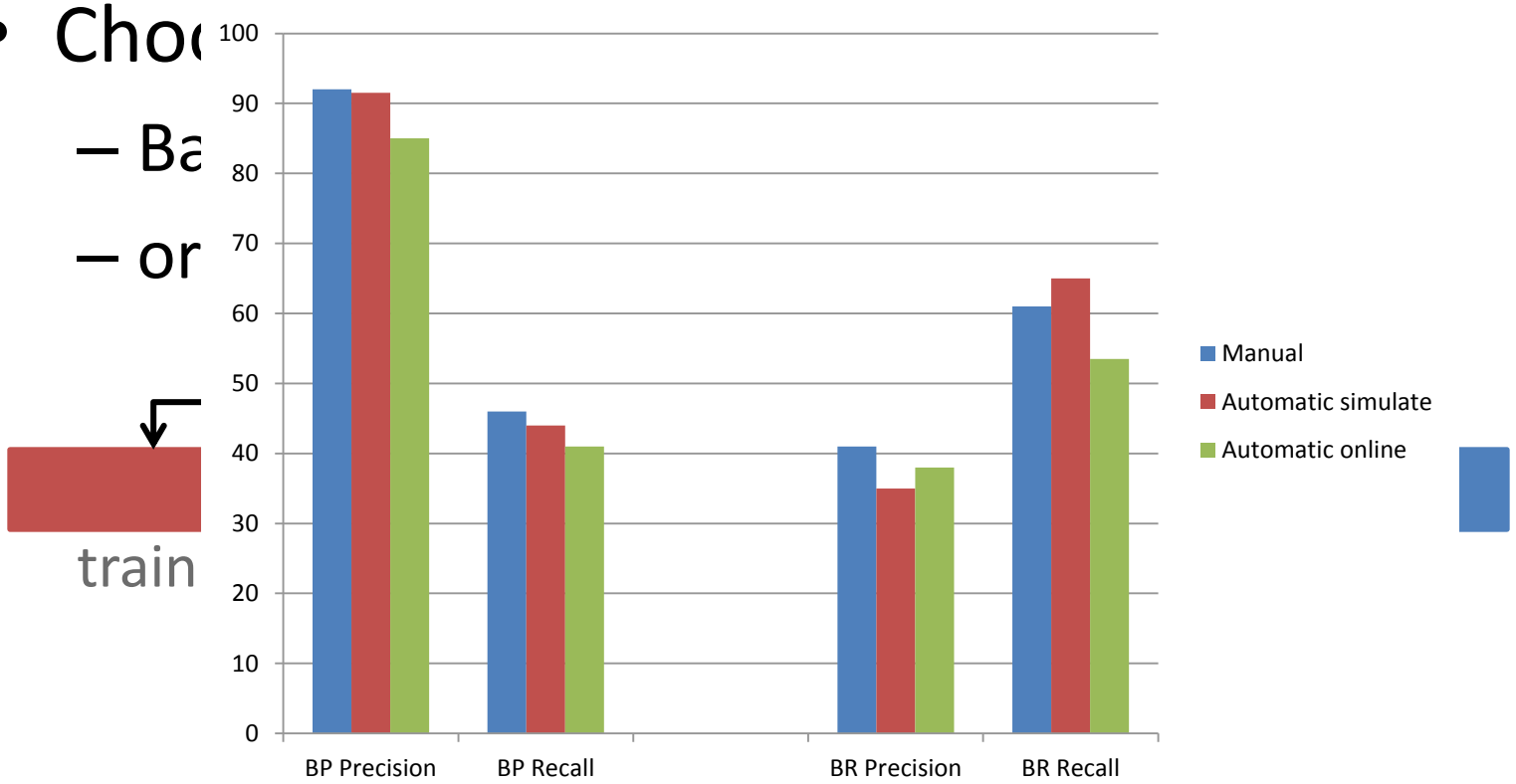
Online prediction

- Choosing parameters
  - Based on experience with other systems
  - or



Online prediction

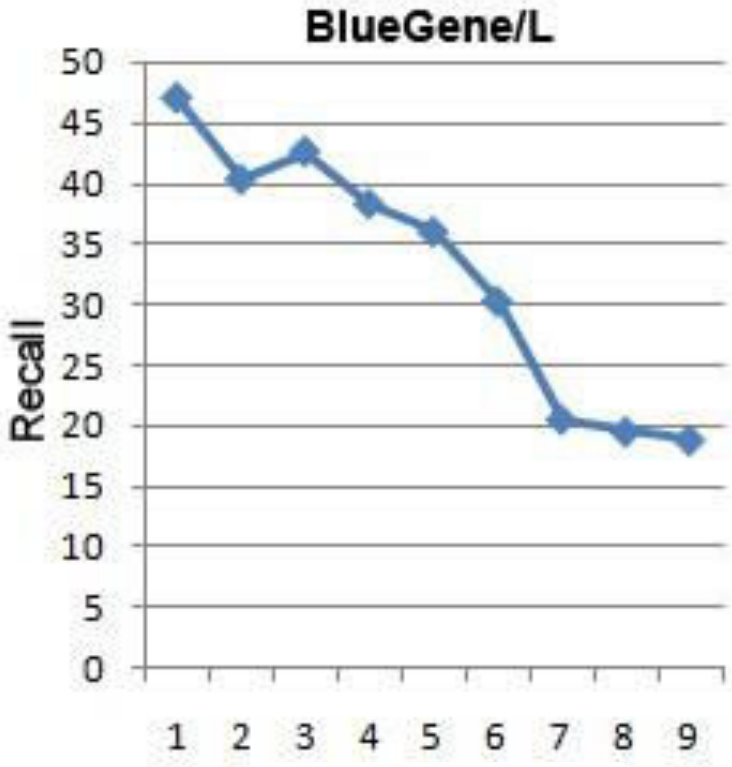
- Choo
  - Ba
  - or



train

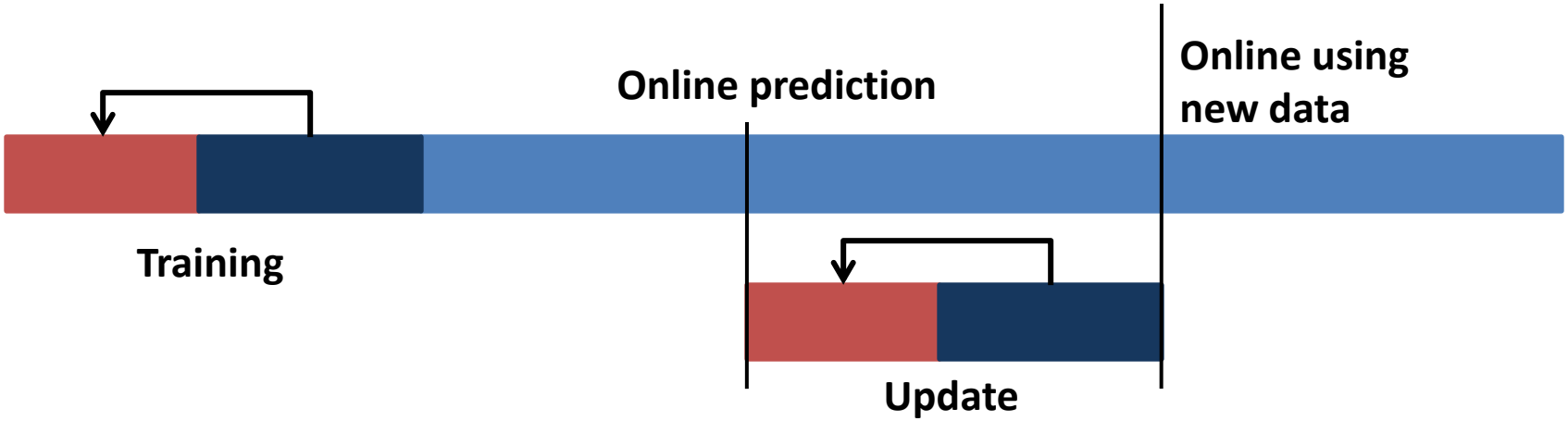
Online prediction

- Recall on different months



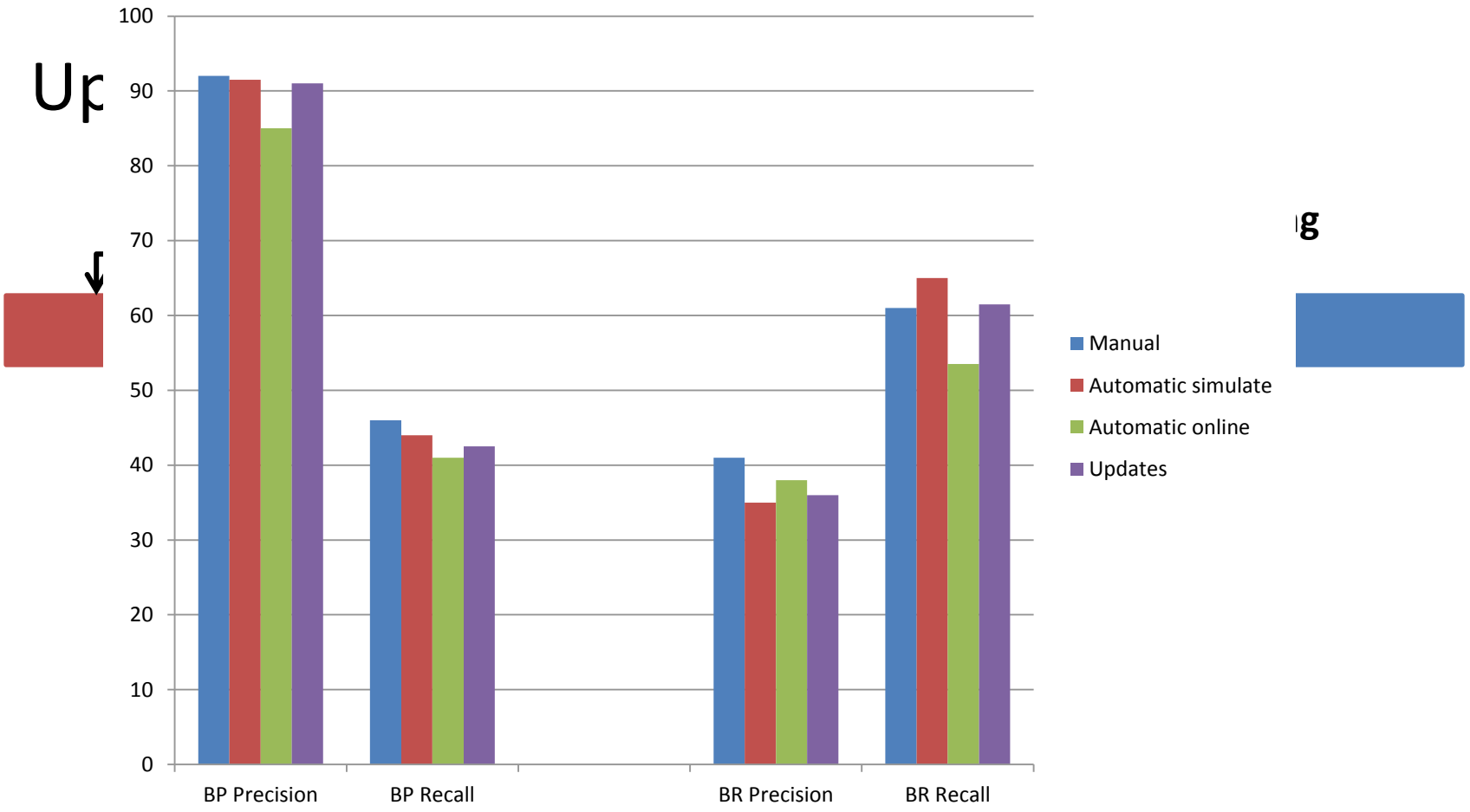
Online prediction

- Updates



Online prediction

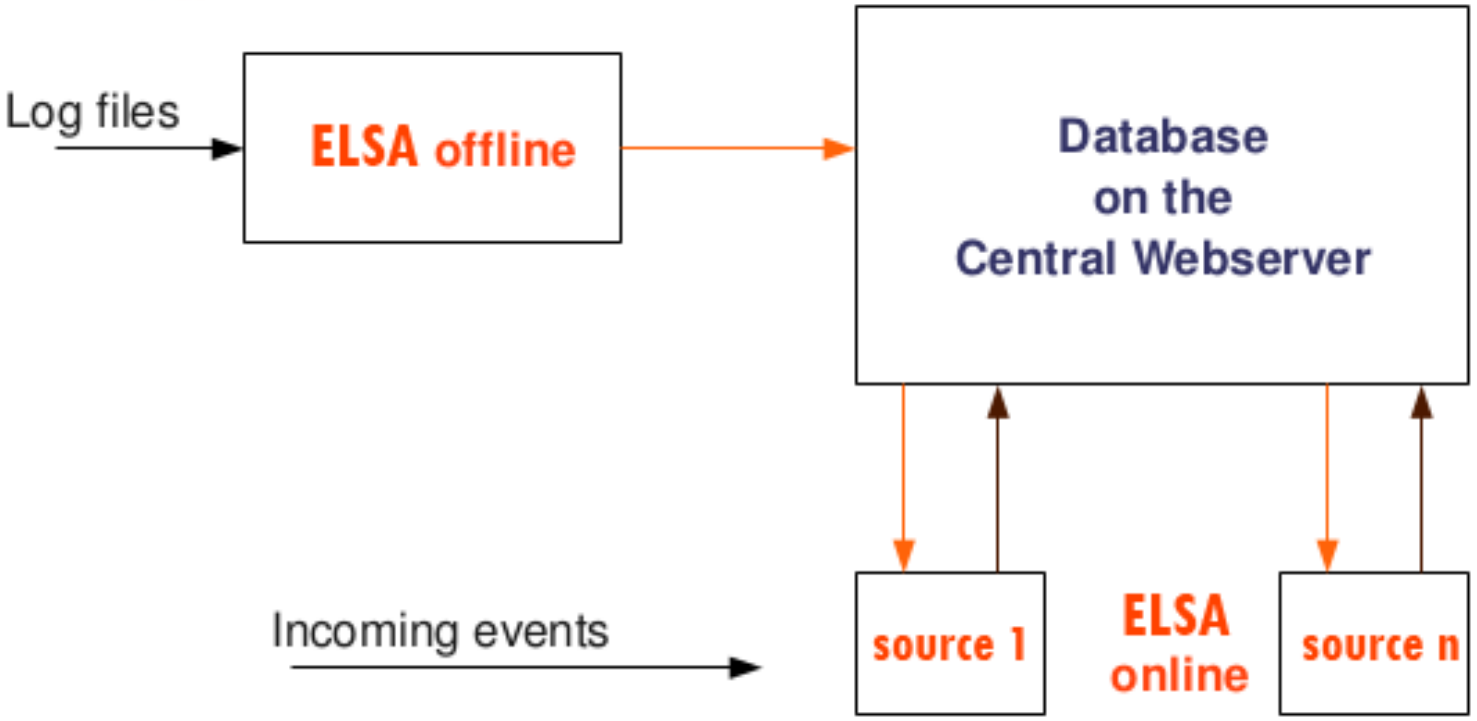
- Up



Blue Waters

- Event logs are gathered on a separate machine
  - System logs
  - HPSS (High Performance Storage System)
  - Sonexion storage system
  - Moab job scheduler
  - ESMS (Data system manager)
- Failure logs
  - Annotated by Cray
  - Templates identified by NCSA

Blue Waters


Log files
ELSA offline
Database
on the
Central Webserver
Incoming events
source 1
ELSA
online
source n

Blue Waters

- Some numbers

Table 1: Frequency of Special Characters

Source	Events/Day	Total Event Types
Syslog	8GB (50mil events)	3,852
HPSS	1MB (900,000 events)	358
Sonexion	3.5GB (10mil events)	3,112
Moab	500 MB (15mil events)	725
ESMS	3GB (12mil events)	2,452
System	Events/Day	Total Event Types
BlueGene/L	5.76MB (25,000 events)	186
BlueGene/P	8.12MB (120,000 events)	252
Mercury	152.4MB (1.5mil events)	563

Blue Waters

- Some numbers

n – no events t – no templates c – no correlations	HELO	Anomaly detection	Correlation extraction / identification
Complexity	Offline: O(n*logn) Online: O(t)	Offline: alg*O(t) Online: alg	Offline: O(t*n*logn) Online: O(c*avg_len(c))
BGL	~10min	~30min	~8h
Blue Waters	~2h	~5h	>24h

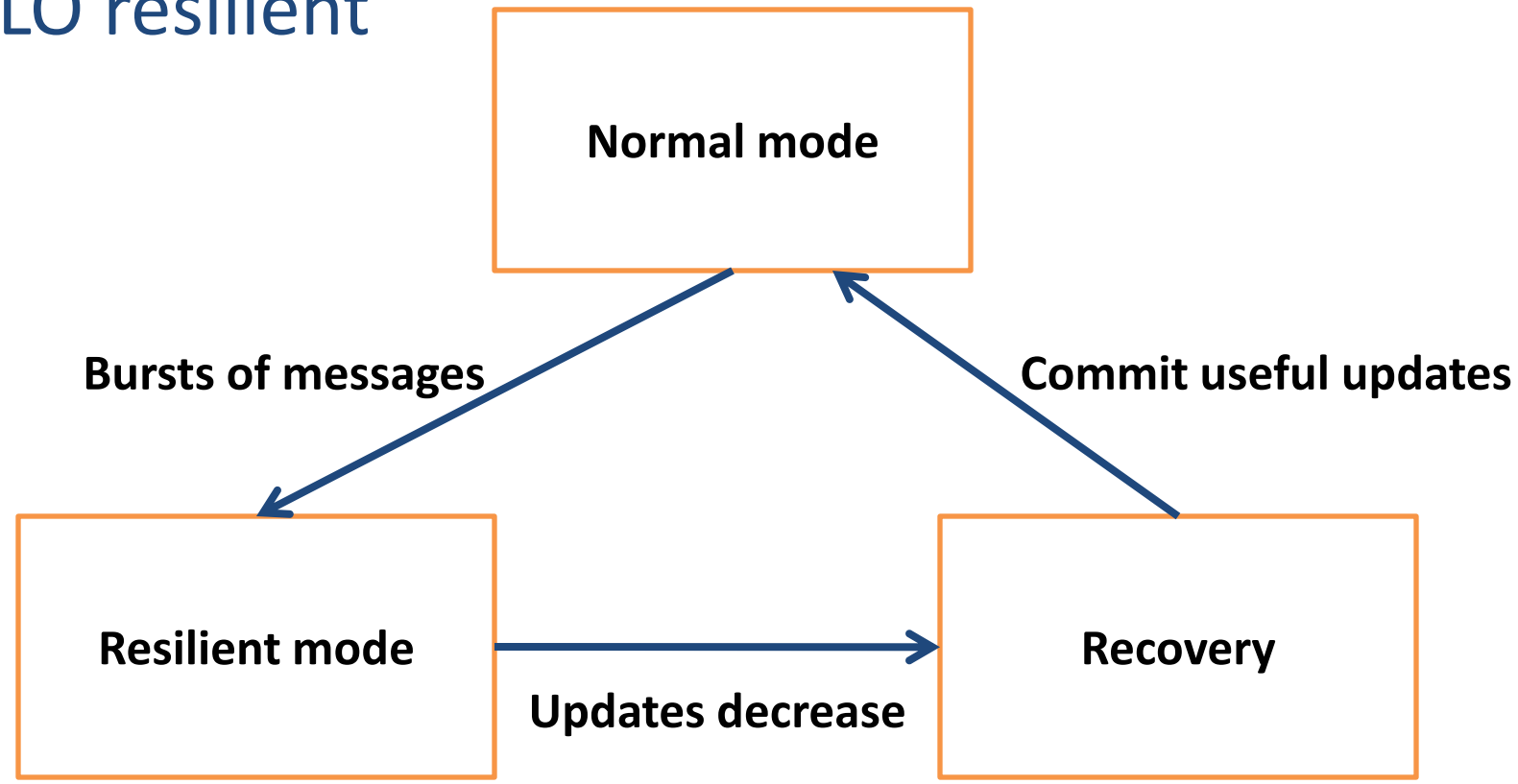
- Execution in parallel
  - Still getting delays

Blue Waters

- First problem
  - Protect ELSA from junk messages
  - HELO resilient

Blue Waters

- HELO resilient



Blue Waters

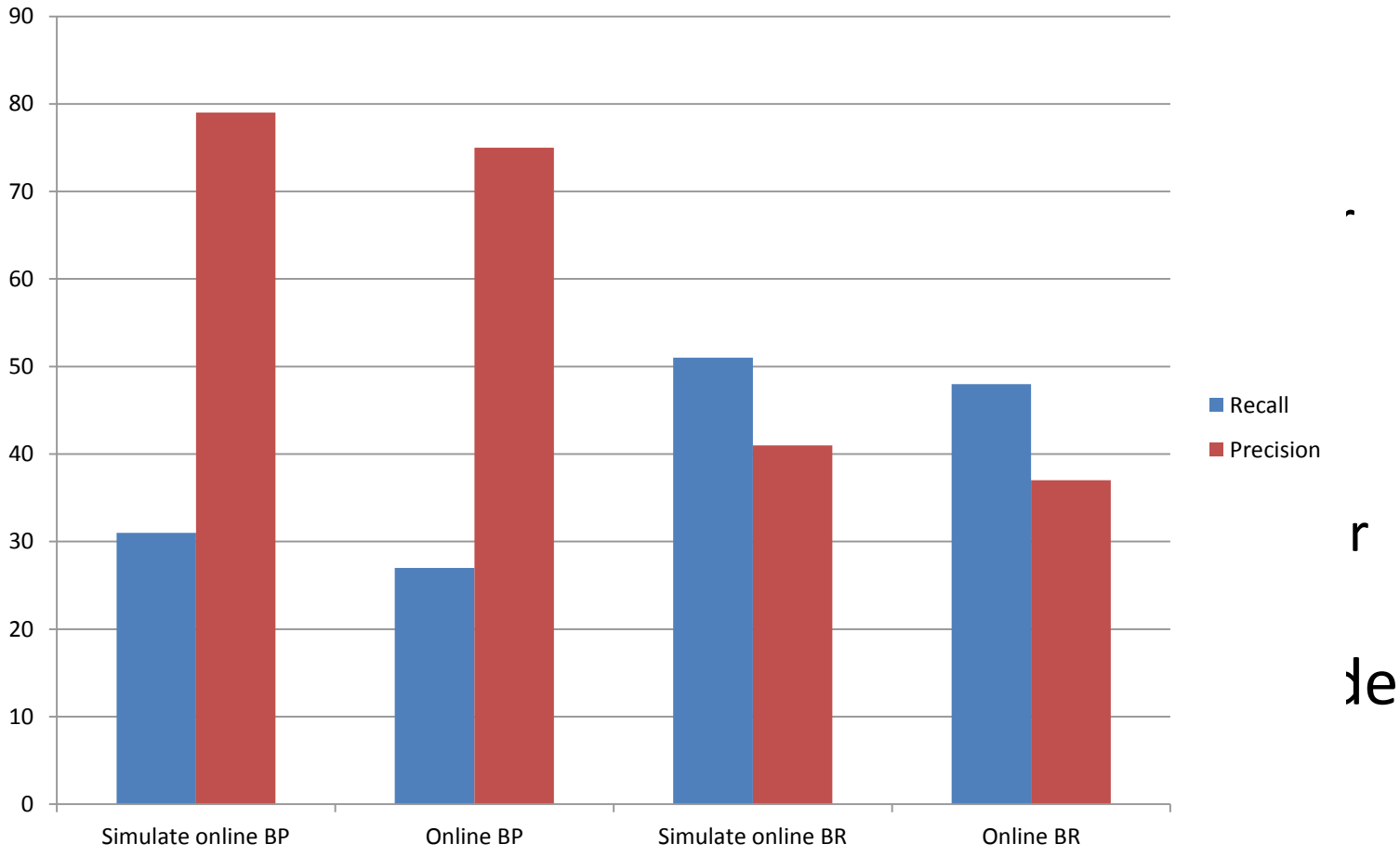
- Failure identification
  - Templates identified by NCSA
  - Cray records events of interest
    - Cross correlate between each other and with templates
  - 30% failures - no template match
    - DIMM errors do not correlate with any
    - Represent 25% of failure occurrences in April 2013

Blue Waters

- Correlations
  - Offline: Decreasing the execution time
    - Hierarchical algorithm (fast less reliable and slower more accurate)
    - Focus on the identified problems
  - Online: Only 42% of the type of failures
    - Are part of any correlation that gives a delay grater then 10 seconds
  - 13.5% of the failures affect more than one node
    - Random set of locations

Blue Waters

- (


90
80
70
60
50
40
30
20
10
0
Simulate online BP
Online BP
Simulate online BR
Online BR
Recall
Precision

Conclusion

- What we want to achieve
  - Online failure prediction
  - Seems feasible on small systems
- Challenges with Blue Waters
  - Large execution time (in progress)
  - Sensible to log perturbations (done)
  - Some anomalies do not present any symptoms
  - Failures are not part of any correlation

Future work

- Difference between small systems and Blue Waters
  - Anomaly detection
  - Correlation extraction
- Increase the current results by a combination
  - Different predictors

Collaboration directions

1) Combining prediction with other solutions
   - Collaboration with UIUC / INRIA
2) Using ELSA for root cause analysis
   - Collaboration with NCSA / ANL (also Sandia)
3) Understanding failures in HPC: precursor detectors
   - Collaboration with NCSA / ANL

Additional Q&A

Thank you

Ana Gainaru

againaru@illinois.edu