Multi-layer provenance framework for job distributed system

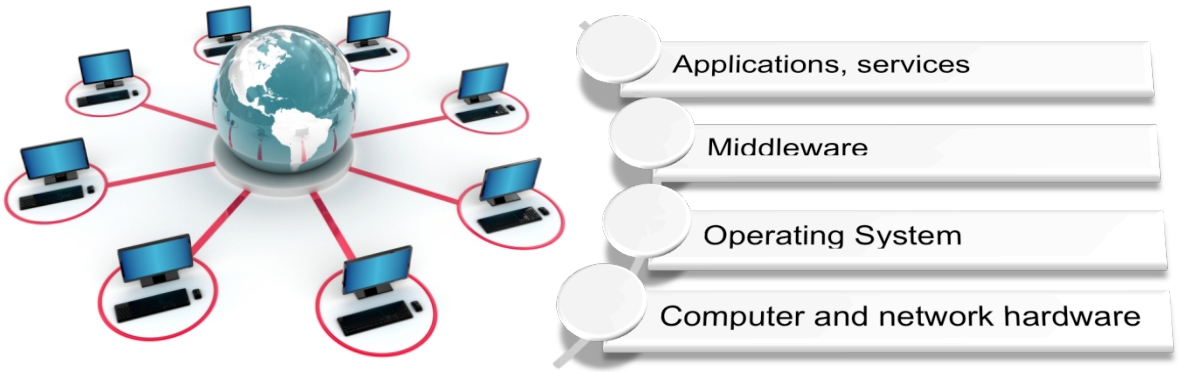
Quan Zhou
quzhou@indiana.edu
Indiana University Bloomington

May 5, 2016

DATA TO INSIGHT CENTER
INDIANA UNIVERSITY
Pervasive Technology Institute

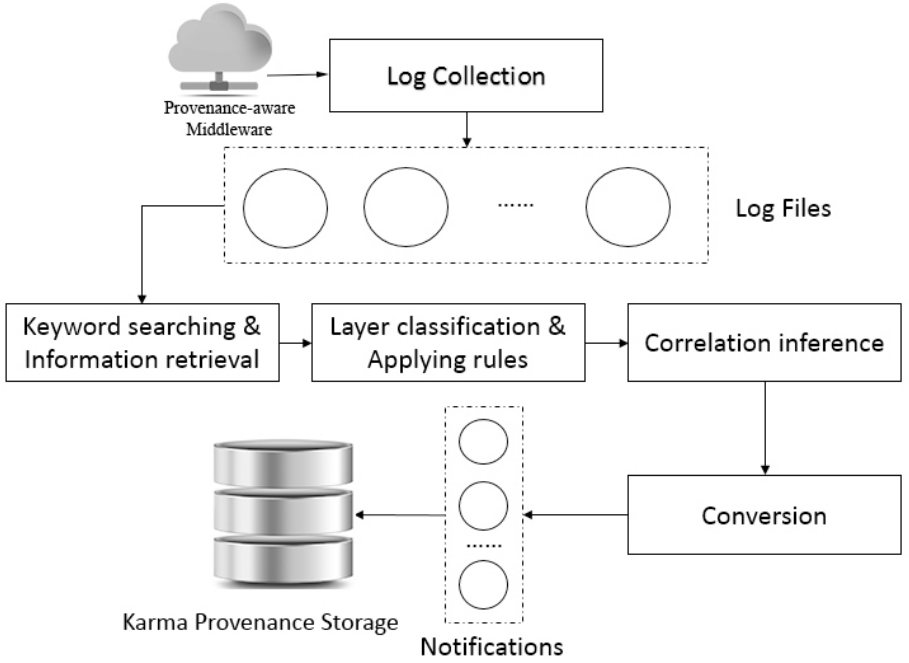
Motivation

- Global information is expensive in distributed environment
  - Multi-layer software stacks and distributed knowledge
  - Massive overhead to infer correlation between events or activities in varied layers
- Enable provenance-based failure tracing and data lineage tracking
  - Data management requirements in data driven scientific workflows
  - Failure tracing and diagnose requirements in highly distributed system
- Examine the quality and usefulness of provenance that can be collected by middleware and platform provenance only



Methodology

- Provenance-aware middleware
  - Middleware wrapper script
- Provenance adaptor
  - Log collection
  - Keyword searching
  - Information retrieval
- Correlation inference engine
  - Timestamp similarity
  - Data flow similarity
- Provenance storage
  - Raw data to provenance conversion
  - Karma Provenance Repository



Application Cases

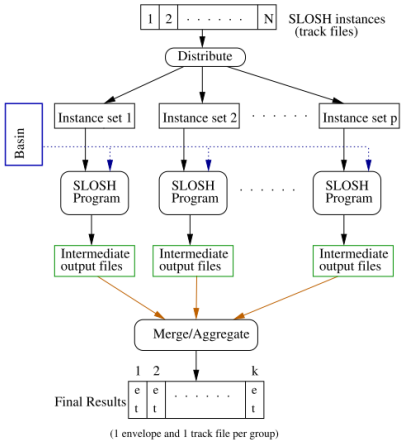


Fig. 1 the Sea, Lake and Overland Surge from Hurricanes (SLOSH) model By NOAA

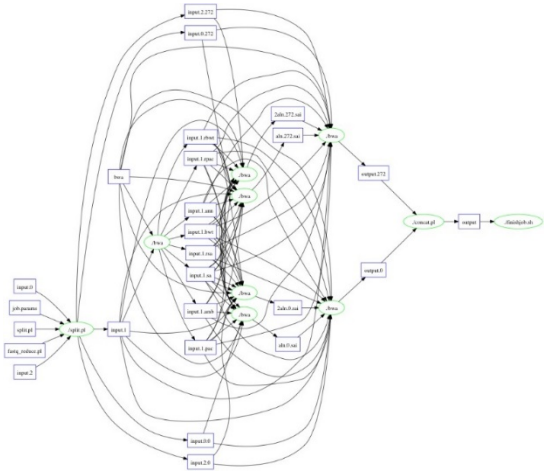


Fig. 2 BWA workflow (Input size: 6.5 GB; Used workers: 10)

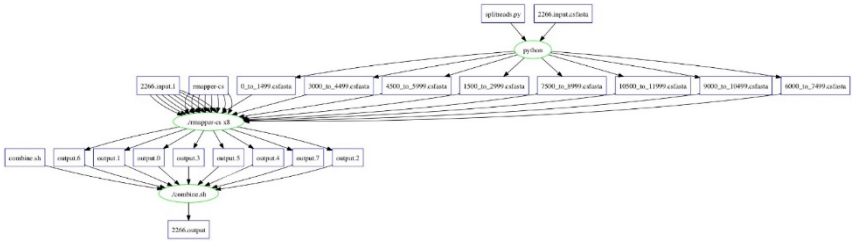


Fig. 3 BLAST workflow (Input size: 8.4 GB; Used workers: 10)

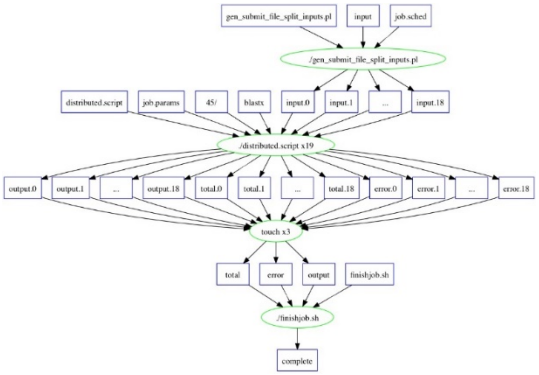
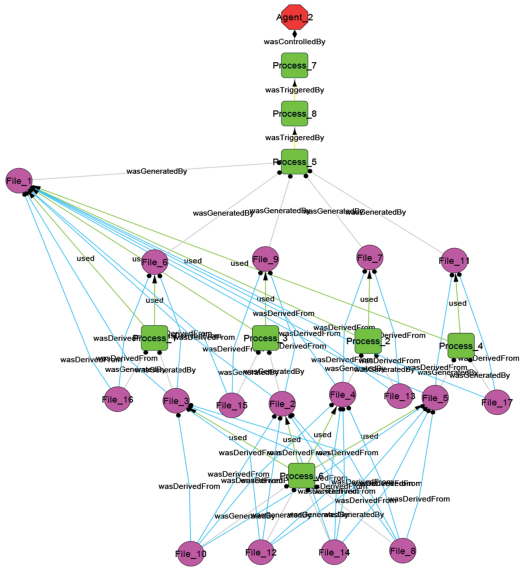


Fig. 4 Shrimp workflow (Input size:675 MB; Used workers: 10)

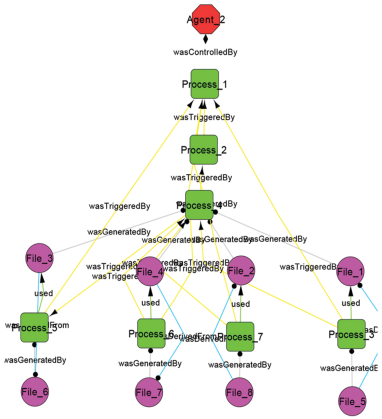
Evaluation

- Multi-layer provenance events correlation inference

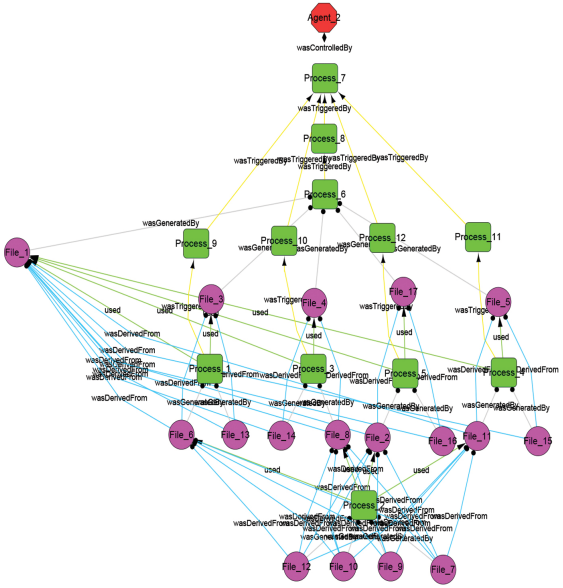
Application

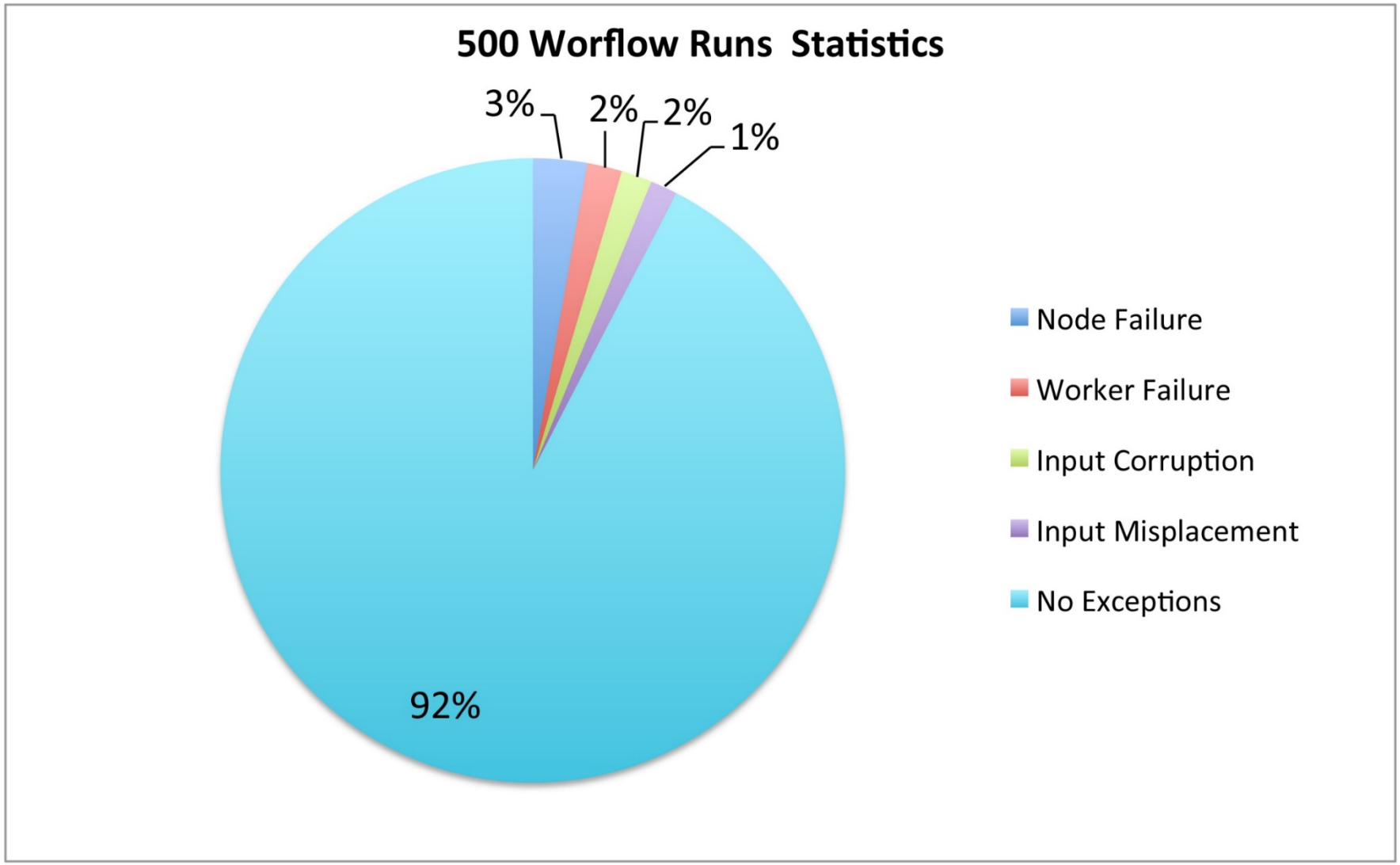


Middleware



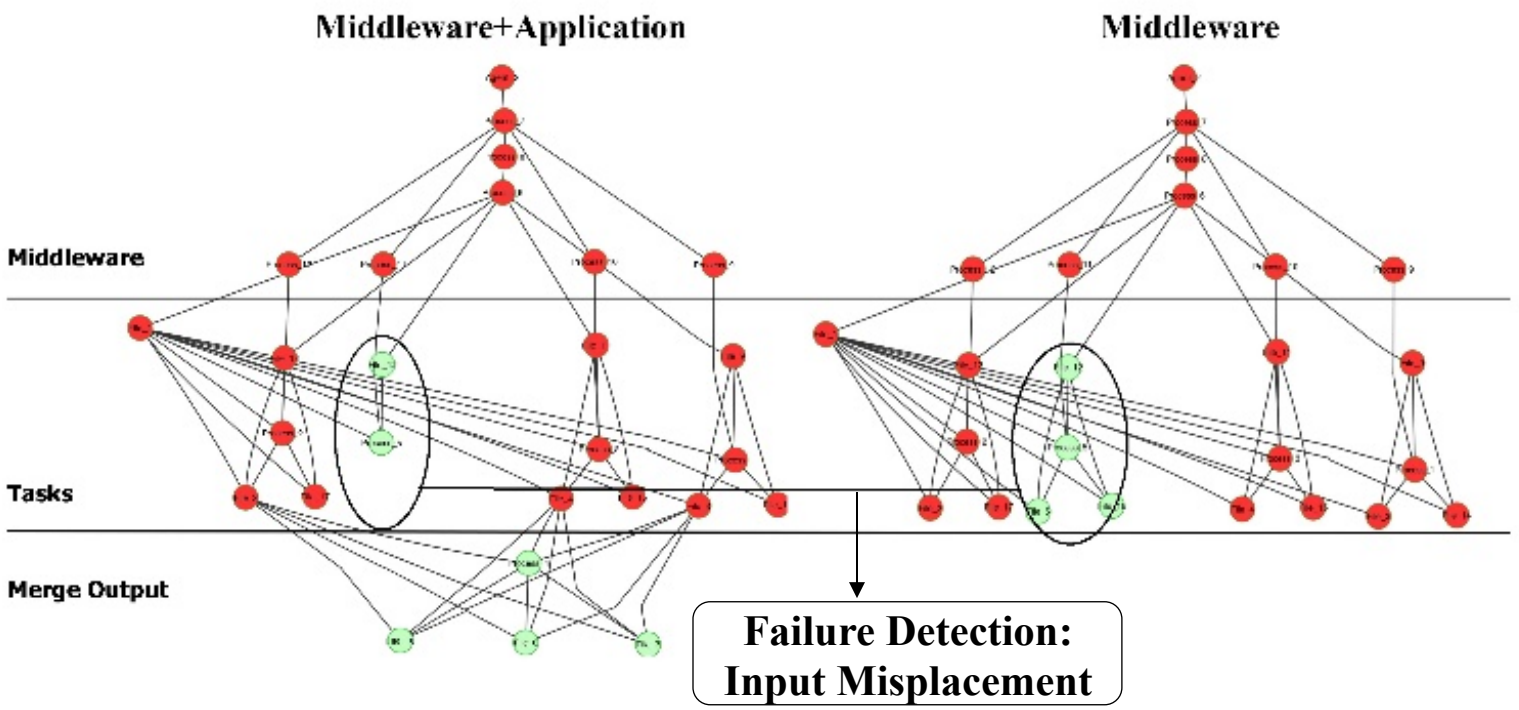
Middleware+Application




500 Worflow Runs Statistics
3%
2%
2%
1%
92%
Node Failure
Worker Failure
Input Corruption
Input Misplacement
No Exceptions

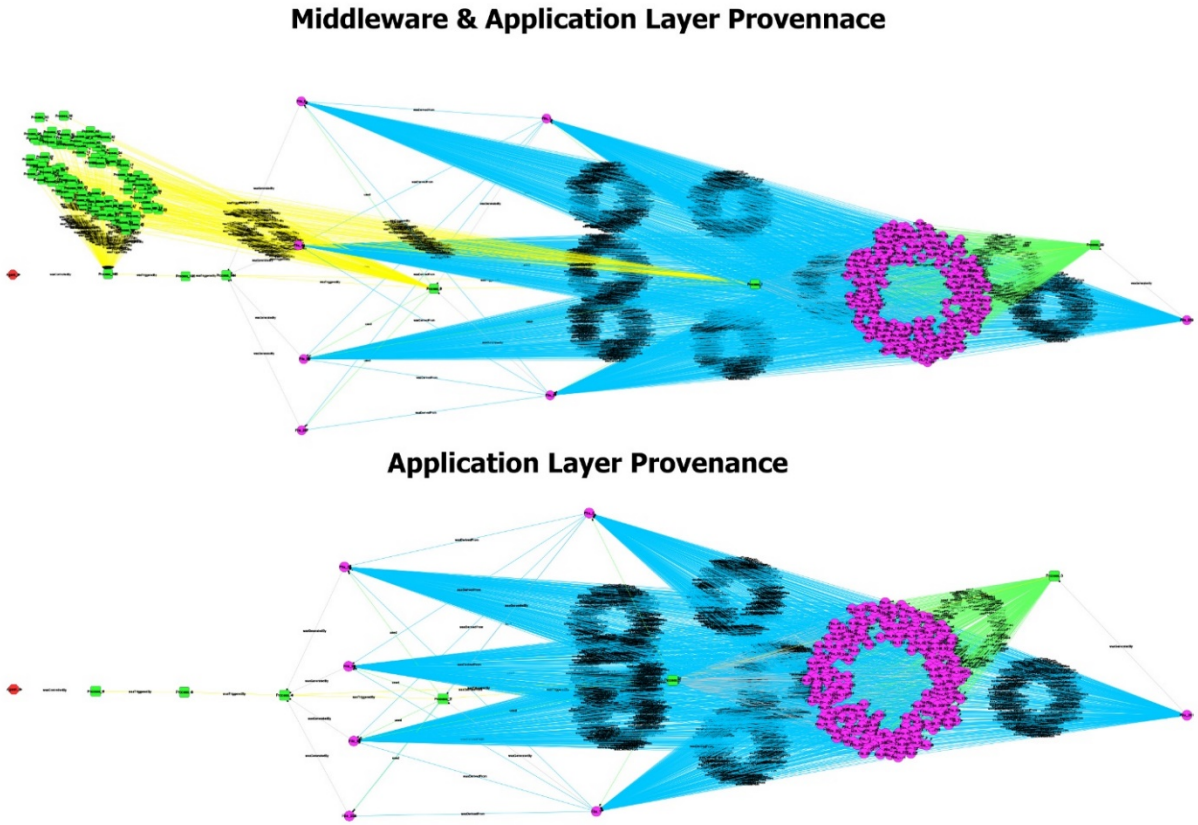
Evaluation

- Provenance-based failure tracing and data lineage tracking



Evaluation

- Generalizing multi-layer provenance framework



Ongoing and Future Work to handle BigData in distributed environment

- Quality metrics for provenance data quality assessment
- Adding more configurability and portability for provenance identification and collection mechanism
- Seeking more application cases for generalization

Acknowledgements

This work is funded in part by the National Science Foundation OCI 1148359. We thank Craig Mattocks of the National Hurricane Center, Miami, Florida for his expertise with the Sea, Lake and Overland Surges from Hurricanes (SLOSH) model and Doug Thain of Notre Dame for sharing datasets that run on WorkQueue.