Failure Analysis of Jobs in Compute Clouds: A Google Cluster Case Study

Xin Chen and **Karthik Pattabiraman**

University of British Columbia (UBC)

Charng-Da Lu, Unaffiliated

Cloud Computing

Compute Clouds

amazon webservices™ EC2

Data & Storage Clouds

amazon webservices™ S3

Failures are a fact of life – applications need to be resilient to failures

Pervious Studies on Failures

- System Failures
  - HPC [Martino et al., DSN 14’], [El-Sayed et al., DSN 13’]
  - Cloud hardware reliability [Vishwanath et al., SoCC 10’]

- Application Failures
  - Hadoop [Kavulya et al., CCGrid 10’], [Ren et al., IISWC 12’]

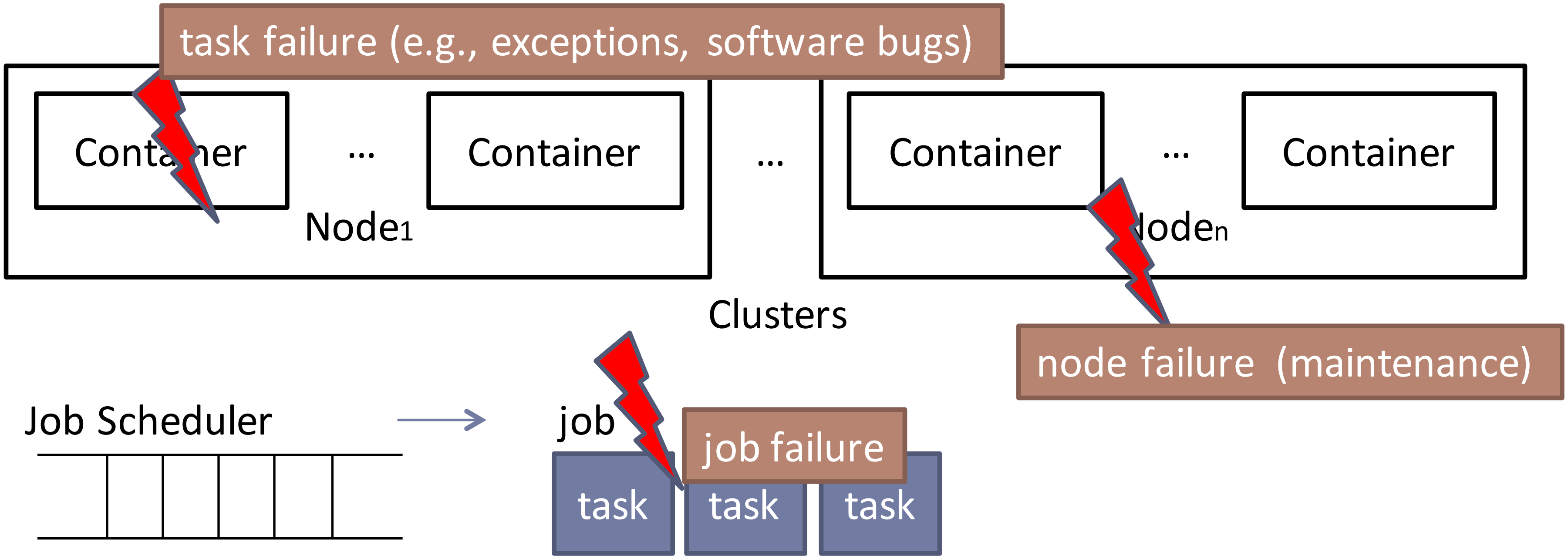
No prior application failure study on a generic production cloud with heterogeneous workloads

Dataset used in our paper [ISSRE'14]

- **Google cluster workload traces [Wilkes2012]**
  - Originally released for job scheduling studies
  - Publicly available, open-source license
  - One month data on production cluster of 1,2500 nodes
  - Includes both failure data and periodic resource usage data

- **Hides important information such as nature of jobs, users, spatial locations of tasks etc. for privacy reasons**
  - Root causes of failures is not provided – no ground truth
  - **Standard disclaimer: Correlation is NOT causation**

Google Clusters: Failures

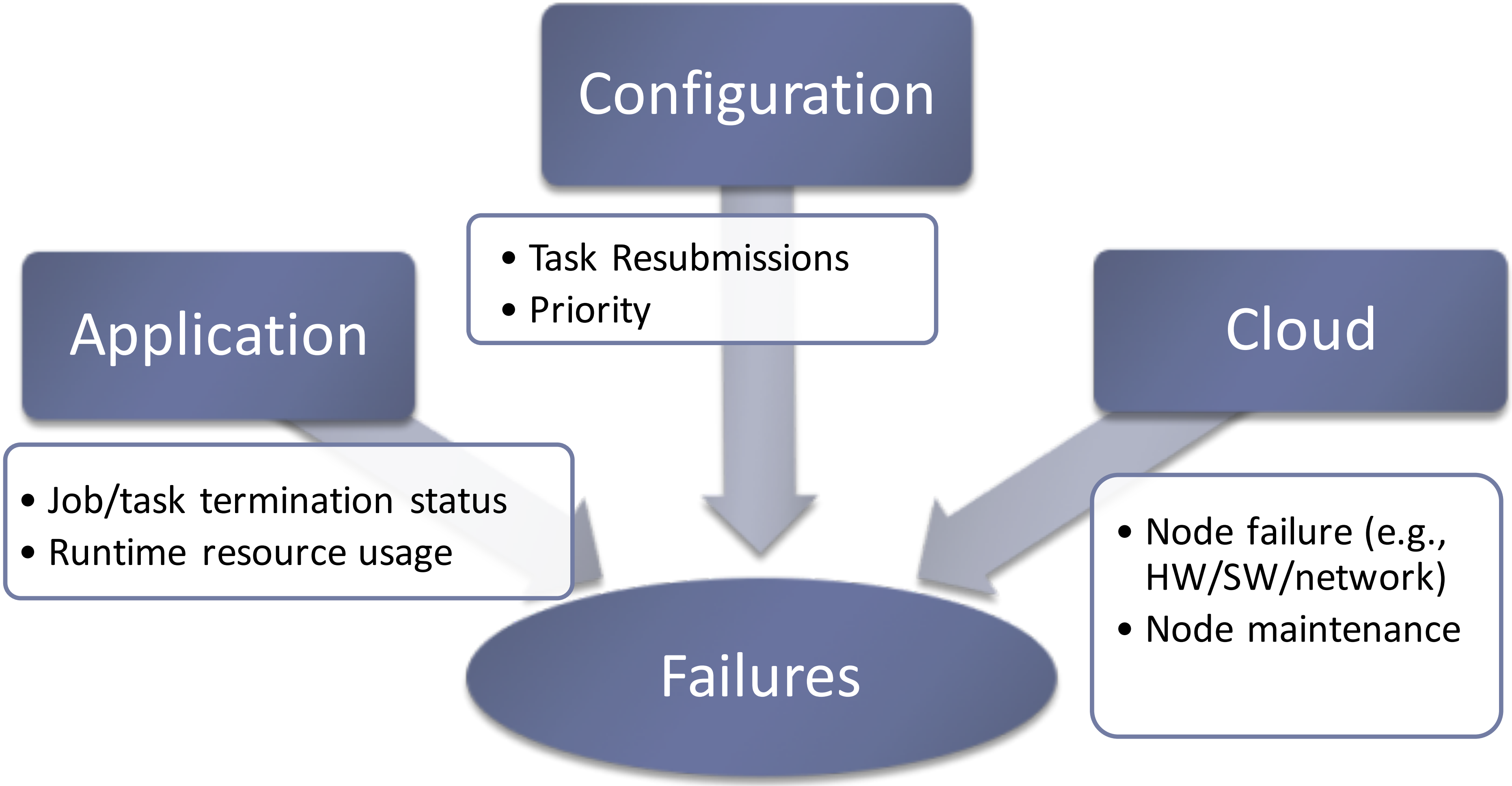


- Production jobs
- Batch jobs

- Around 680 users
- 670,000 jobs
- 48 million tasks
- 12,500 nodes for 1 month

- An average of 14.6 jobs fail in an hour > 10,000 job failures
- Failed jobs constitute about 1.5% of the total jobs (670,000)

Factors leading to Cloud Application Failures



Configuration

- Task Resubmissions
- Priority

Application

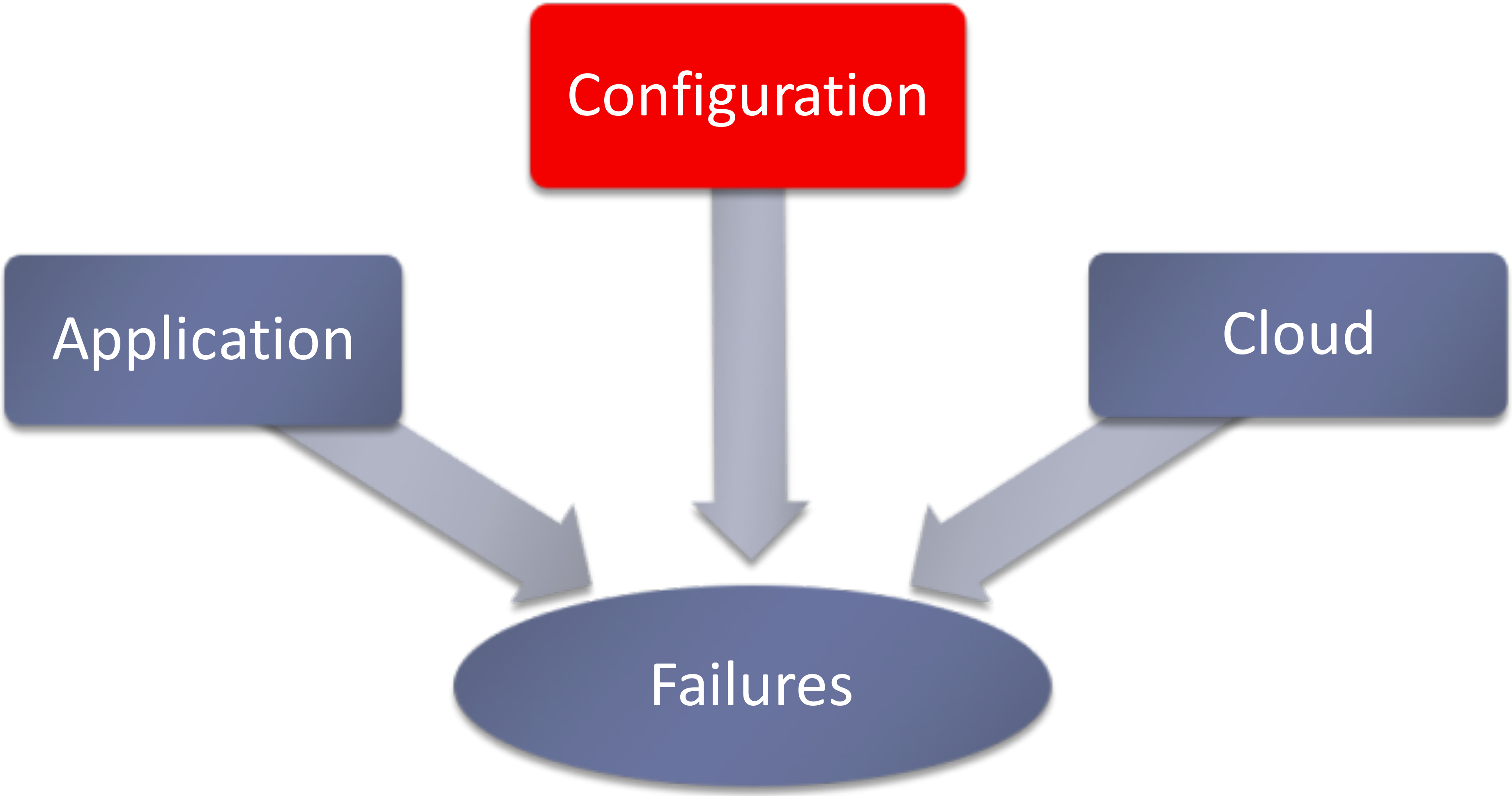
- Job/task termination status
- Runtime resource usage

Cloud

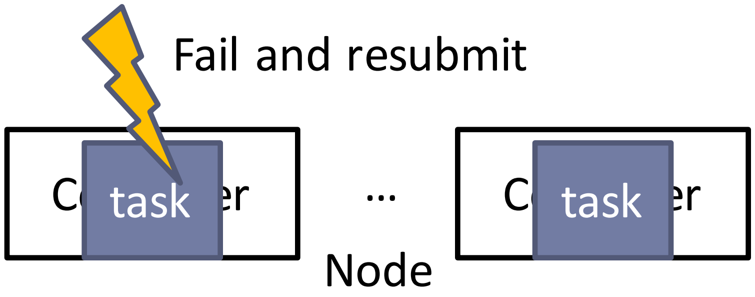
- Node failure (e.g., HW/SW/network)
- Node maintenance

Failures

Factors leading to Cloud Application Failures



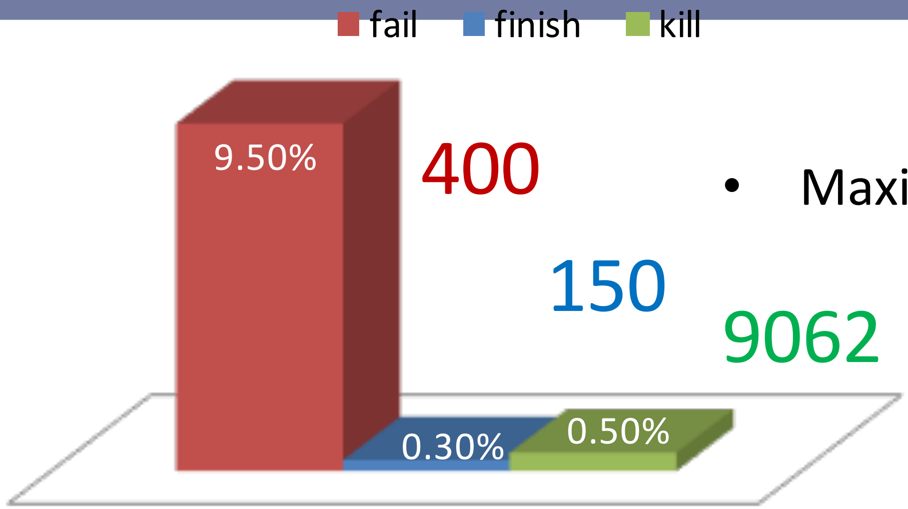
Configuration Factor: Task Resubmissions


Fail and resubmit

- Task resubmission

Frequent task resubmissions may waste resources and energy, particularly in failed and killed jobs.

bs



400

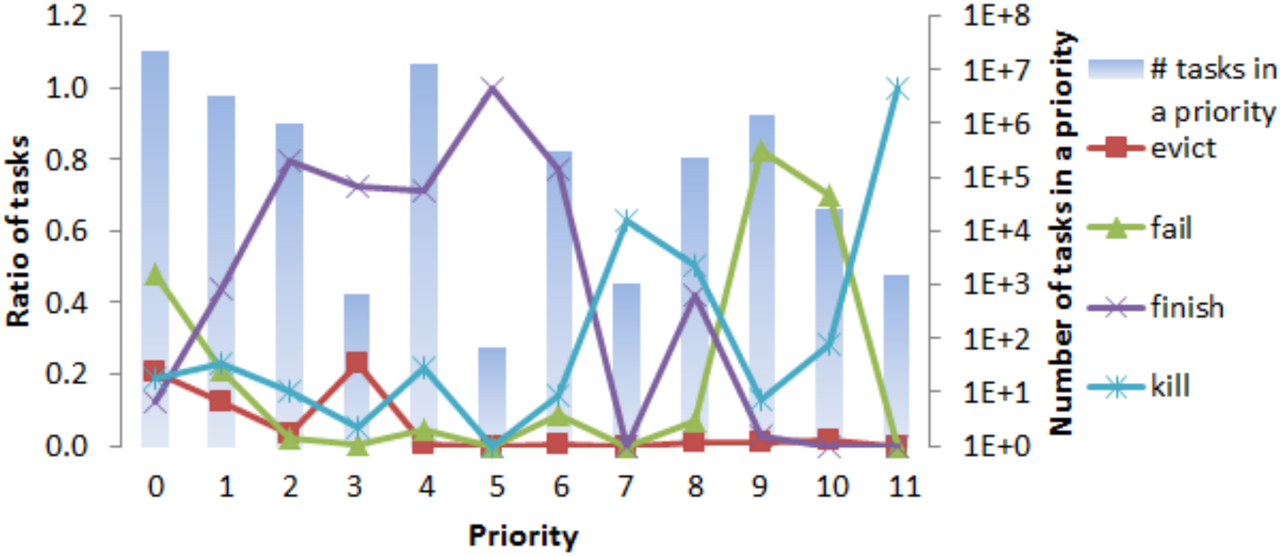
150

9062

- Maximum resubmissions

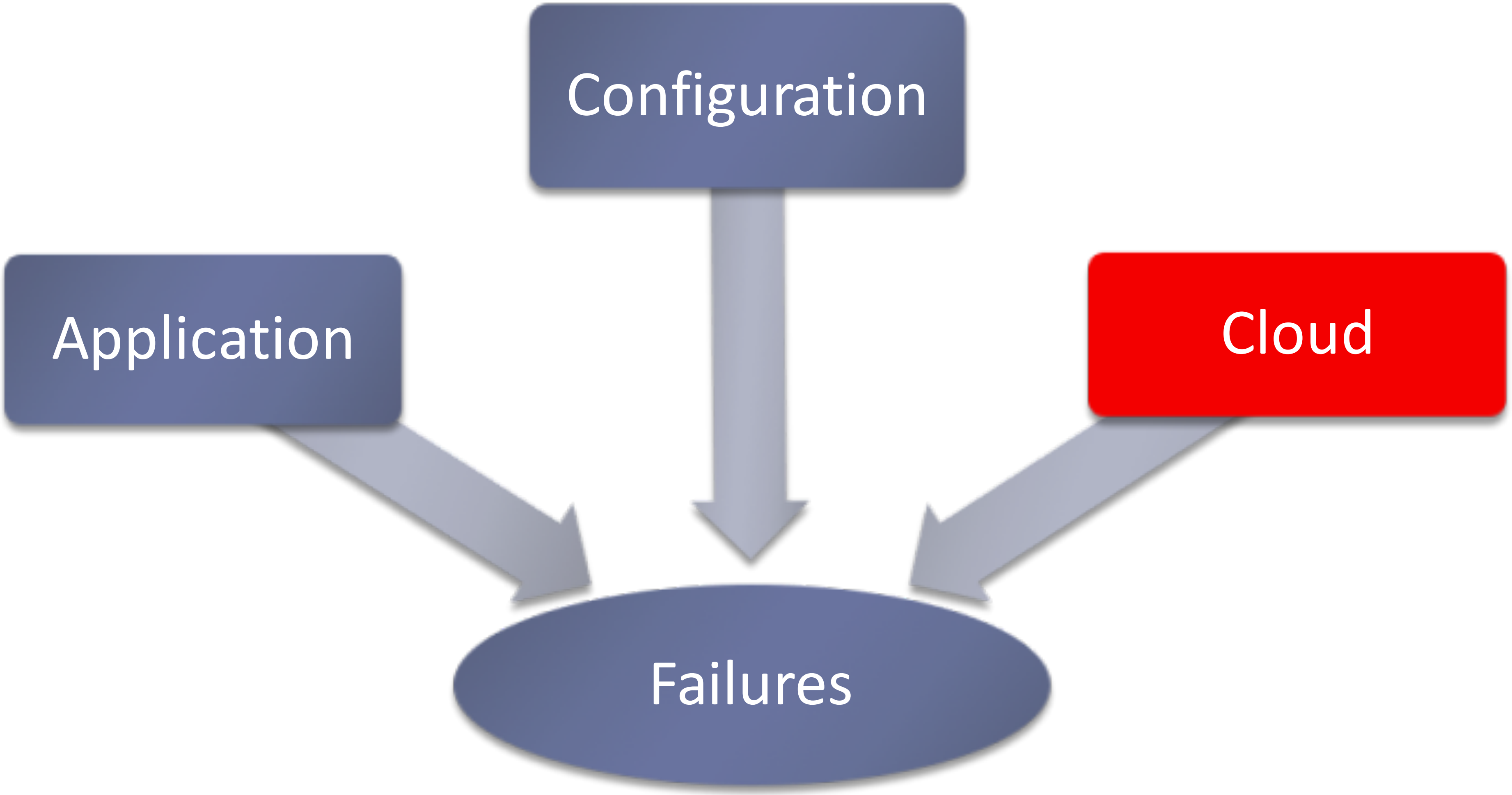
Configuration Factor: Priority

Priority determines the nodes assigned to the task.

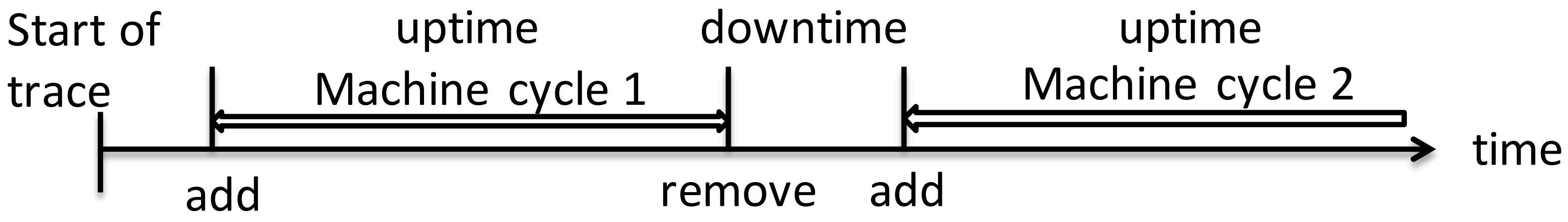


- Low-priority and high-priority jobs experience high failure rates
  - Result holds even when disregarding resubmissions
  - Can be used in failure prediction

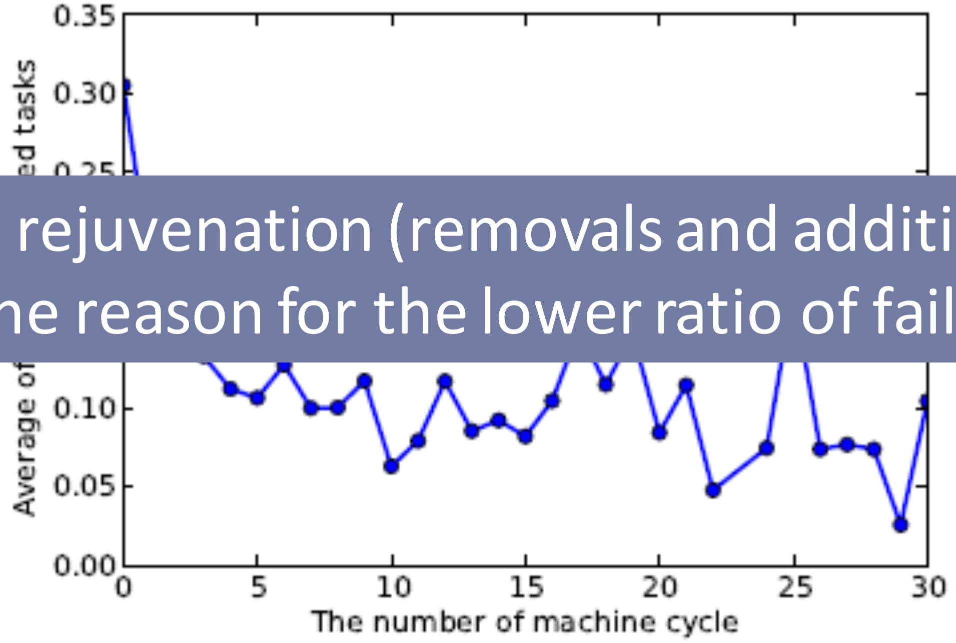
Factors leading to Cloud Application Failures



Cloud Factor: Node Removal and Addition

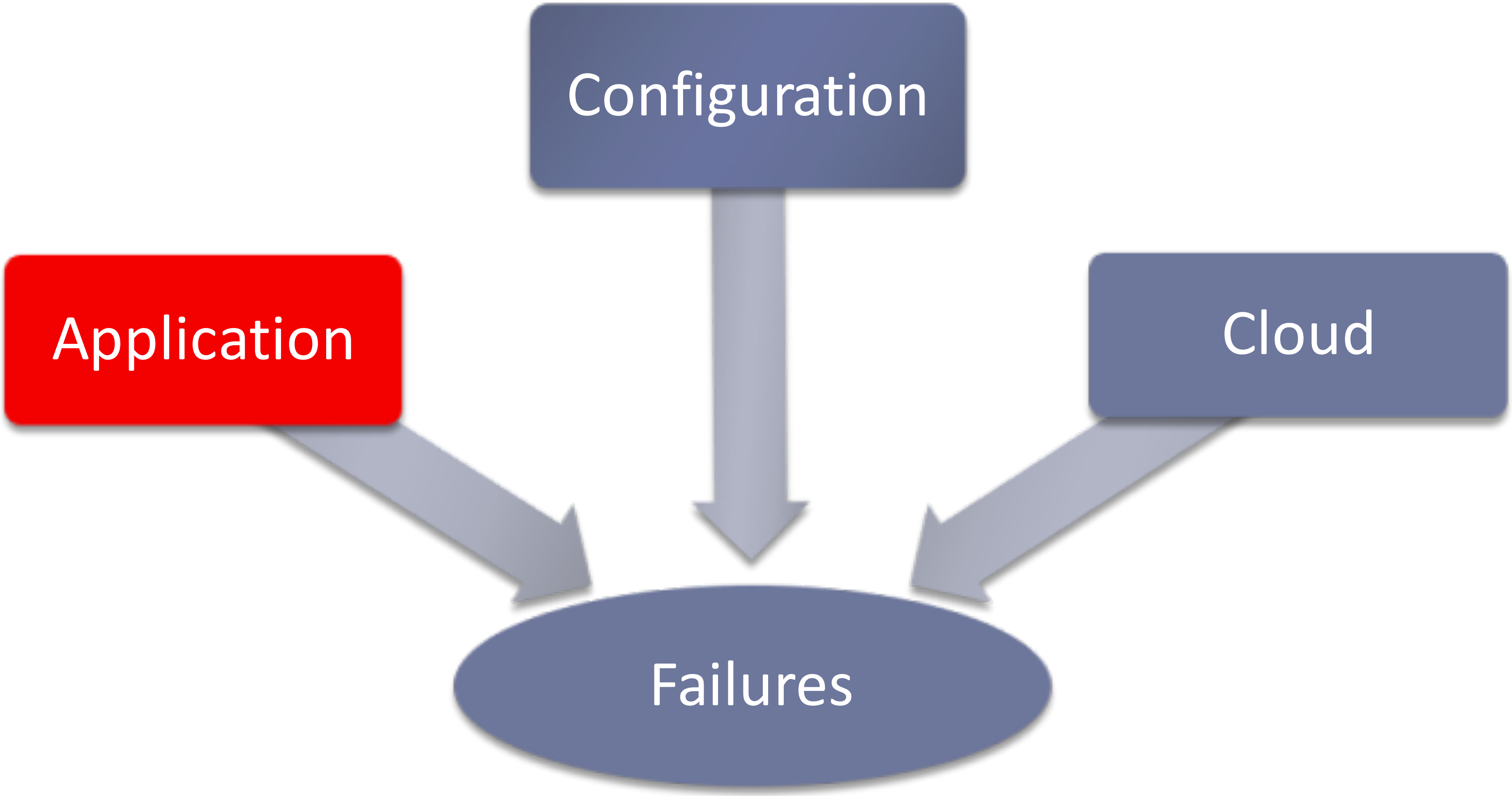


- Average of failed task ratio VS number of machine cycles



Machine rejuvenation (removals and additions) may be the reason for the lower ratio of failures

Factors leading to Cloud Application Failures



Configuration

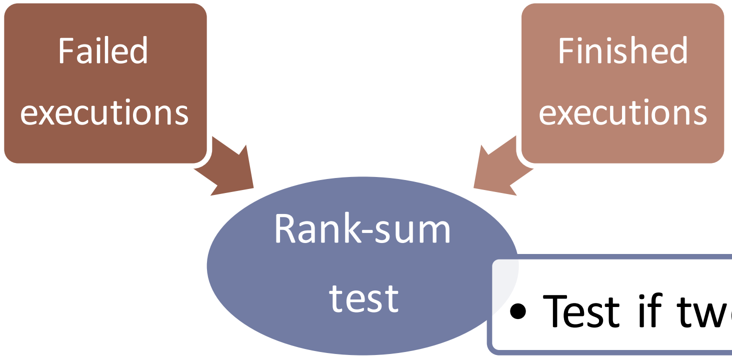
Application

Cloud

Failures

Application Factor: Resource Usage

- Distinctions in the task resource usages



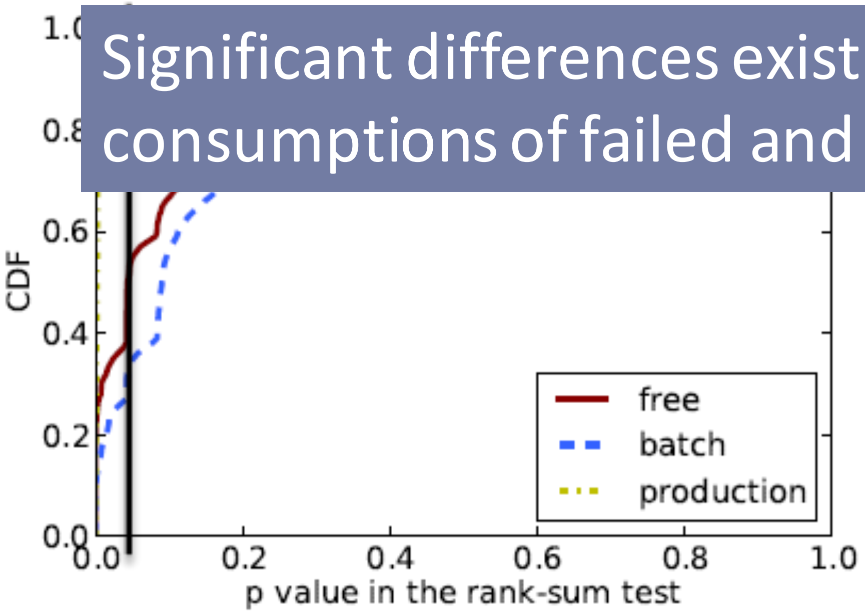
Failed executions

Finished executions

Rank-sum test

- Test if two samples significantly differ

- CPU usage



Significant differences exist between the resource consumptions of failed and finished tasks of same job

0.05

- Batch: 34.8%
- Production: 93.2%

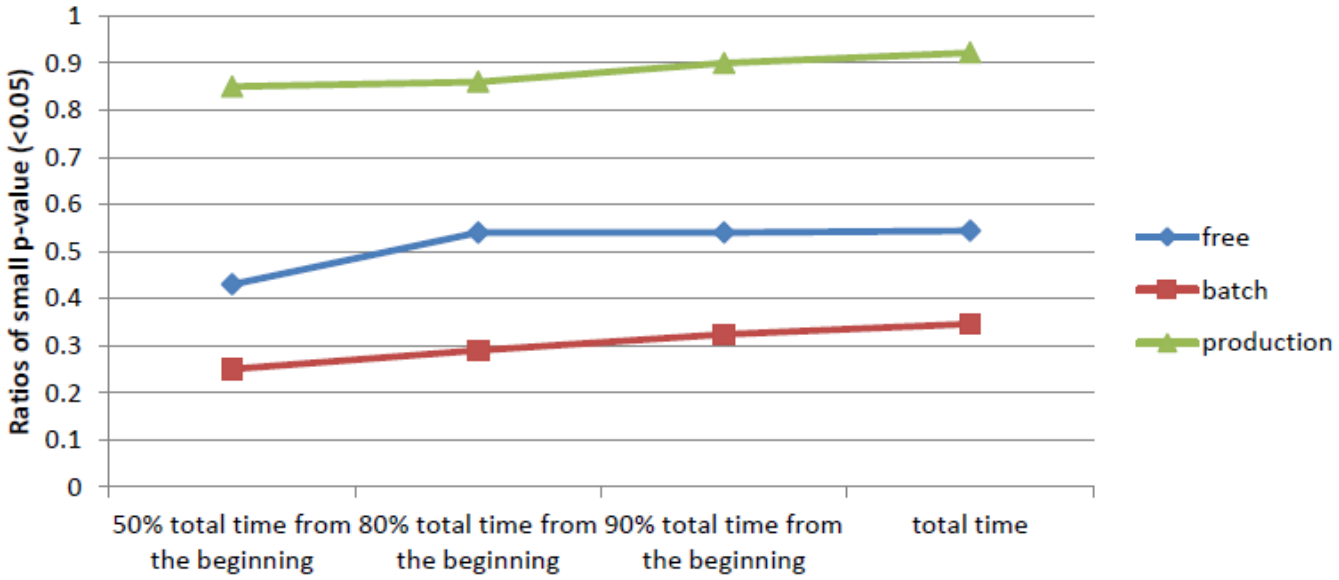
* Free: low priority batch

Application Factor: Early Failure Manifestation

- Differences between failed and finished executions manifest much earlier than the job's termination

Rank-Sum test

- Test if two samples significantly differ



- Resource consumption differences are significant even halfway into the job

Summary of Findings

- **Job failures**
  - High number of task resubmissions in failed jobs
  - Both low and high priority jobs - 3 times as many failures
  - Node maintenance and update improve overall reliability

- **Differences in resource consumption exist between failed and finished jobs**
  - Differences manifest even halfway into a long job's execution

Failure Analysis of Jobs in Compute Clusters: A Google Cluster Case Study. Xin Chen, Charng-da Lu and Karthik Pattabiraman, ISSRE 2014.