PANIC: Modeling Application Performance over Virtualized Resources

Ioannis Giannakopoulos

Computing Systems Laboratory, School of ECE
National Technical University of Athens, Greece
ggian@cslab.ece.ntua.gr

International Conference on Cloud Engineering (IC2E) 2015
Tempe, AZ, USA

Presentation Overview

A really cloudy world...

Cloud deployments are gaining ground against traditional computing.

- Reduce administrative costs
- Pay-as-you-go billing
- Elasticity

Achieving elasticity entails knowledge about the application!!

Application models

Application models: estimate the application performance under different *conditions*:

- Utilized Resources
- Application Level configuration
- Application Load

The *application model* expresses how is performance affected when the parameters change their values.

Challenges

How can I extract an application model?
Obvious solution: capture the application performance for all the parameters' values (and their combinations).
Assumption: *Deploying the same configuration will result in (approx) the same performance*
Not practical for complex applications:

- Deployment time
- Deployment cost

Problem formulation

Web Application example:

- Web Server: {1,2,4,8} cores and {1,2,4,8,16}G of RAM.
- Database Server: {1,2,4,8,16}G of RAM and disk {HDD/SSD}.

$|\{1,2,4,8\}| \cdot |\{1,2,4,8,16\}| \cdot |\{1,2,4,8,16\}| \cdot |\{HDD, SSD\}| = 200$

Deployment space:

$$D = d_1 \times d_2 \times .. \times d_n$$

Application Load:

- discretized (if continuous)
- increases D's dimensionality

Performance Function

Application performance:

$$p : D \to P$$

Approximating p:

- $\widehat{p} : D \to P$
- sample D and deploy application
- obtain performance point for sample
- function approximation approaches
- objective: keep $|p - \widehat{p}|$ for all $d \in D$ minimum

Function Approximation

Require: application A, deployment space D, models M
Ensure: model m

```
while not termination_condition do
    p ← SAMPLE(D)
    d ← DEPLOY(A, p)
    for m ∈ M do
        m.train_incrementaly(p,d)
    end for
end while
return best_model(M)
```

Sampling

Samples: knowledge about the application performance

- Static sampling
  - Uniform sampling
  - Random sampling
- Adaptive sampling
  - Greedy Adaptive Sampling

Greedy Adaptive Sampling algorithm

Require: input domain D, chosen samples L, number K
Ensure: sample s

```
if |L| < K then
    s = borderPoint(D)
else
    max = 0
    for all t1 ∈ L do
        for all t2 ∈ L do
            a = find_midpoint(t1, t2, D)
            if |t1 − t2| > max and a ∉ L then
                max = |t1 − t2|
                s = a
            end if
        end for
    end for
end if
return s
```

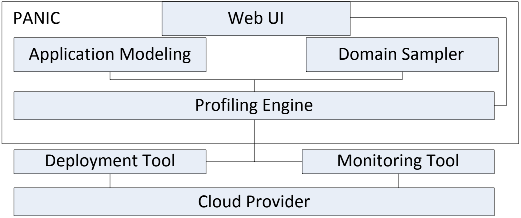
Models

Models:

- Regression techniques
- Classification
- WEKA framework
- Multiple instances trained concurrently

PANIC Architecture

PANIC: **P**rofiling **A**pplicatio**n**s **I**n the **C**loud



- Profiling Engine synchronizes the system's workflows.
- Deployment Tool is generic and works in multiple cloud providers.
- Ganglia: application level metrics are reported as custom metrics.

Applications

Demo applications (deployed over ~okeanos):

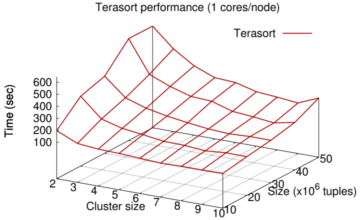
- TeraSort (Hadoop Application)
- PageRank (Hama Application)
- SSSP (Hama Application)

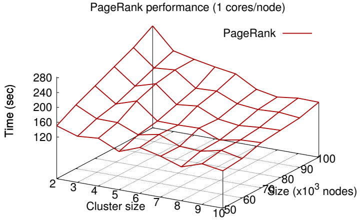
Dimension	Values	
Nodes	2, 3, 4, 5, 6, 7, 8, 9, 10	
Cores/node	1, 2, 4	
Dataset size	Terasort (Millions of Key Values)	10, 20, 30, 40, 50
	PageRank (Thousands of Nodes)	50, 60, 70, 80, 90, 100
	SSSP (Thousands of Nodes)	50, 100, 200, 300, 400, 500

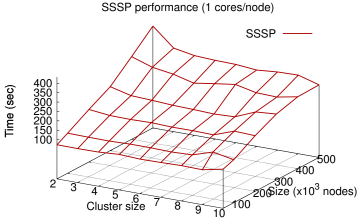
Accuracy metrics: $R^2 = 1 - \frac{\sum_i (y_i - f_i)^2}{\sum_i (y_i - \bar{y})^2}$ and $MAE = \frac{1}{n} \sum_i |f_i - y_i|$.

Performance metric: execution time

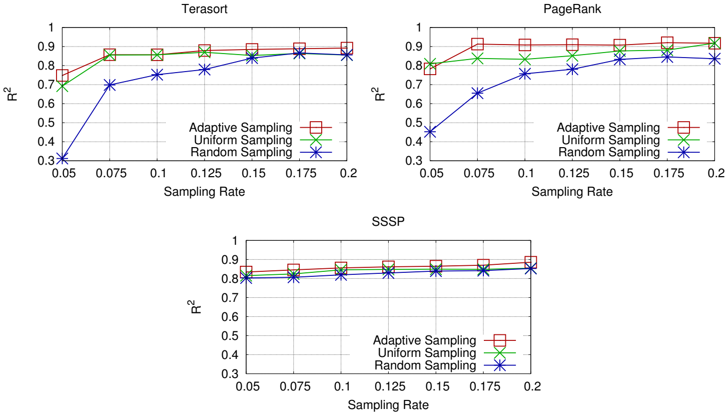
Raw performance



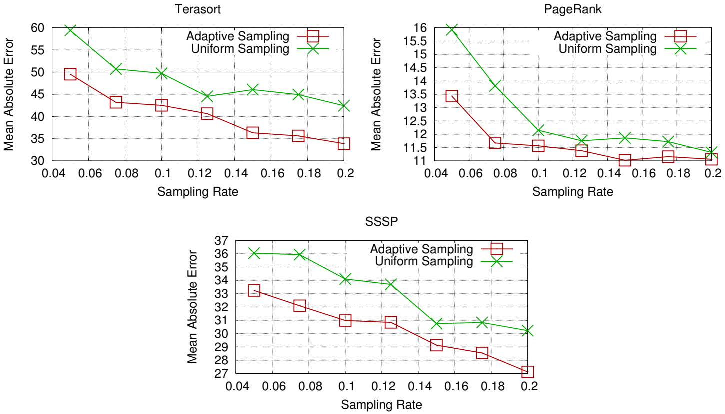




R^2 vs Sampling Rate



Mean Absolute Error vs Sampling Rate



Conclusions

In this paper

- Proposed a generic profiling approach applicable to any cloud application
- Proposed the Greedy Adaptive Sampling algorithm which bases its functionality in identifying the steepest points of p
- Achieved acceptable accuracy when only 10% of D was investigated

Thank you!

Questions?