Resource Central

Cortez et al., 2017

Motivation

Improve resource management in large cloud platforms.

🤔 But what exactly does this mean?

- Balancing of disk IOPS loads
- Preventing physical resource exhaustion in oversubscribed servers
- Reducing cluster migration
- Prioritizing interactive workloads (over delay-insensitive ones)
- etc.

Key Insights

1. If we **knew** certain characteristics of VMs, we could improve resource management in cloud platforms.
    a. Average CPU utilization
    b. P95 max CPU utilization
    c. Deployment size in # of VMs
    d. Deployment size in # of cores
    e. Lifetime
    f. Workload class

🤔 What other VM metrics might be useful to know?

2. Empirical evidence shows that certain VM characteristics are fairly consistent over multiple lifetimes (so they can be **predicted**).
    a. These characteristics are especially consistent on a per-subscription basis

🤔 Are there any VM metrics that *can't* be predicted?

Architecture

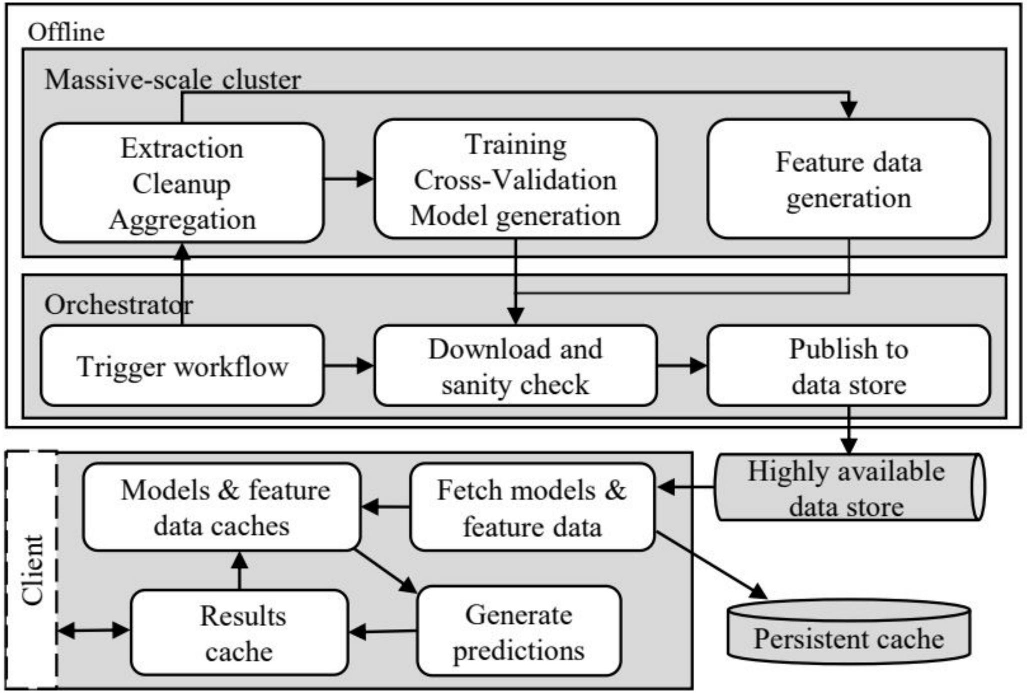
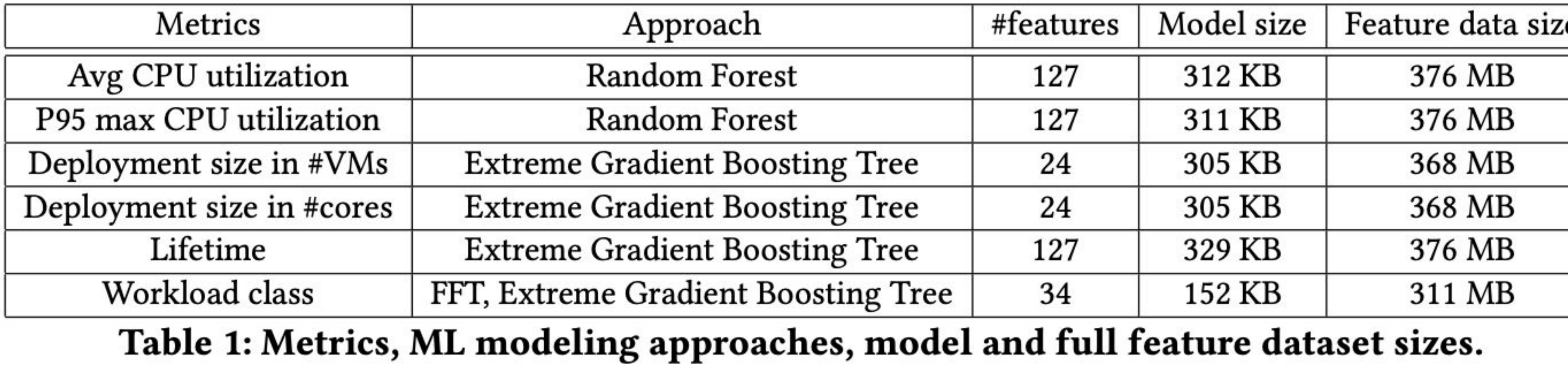


Figure 9: RC architecture (pull version).

Discussion: Consider the offline-online split pictured in this system diagram. What aspects of it do you agree or disagree with? Why?

Discussion: Would you add or remove any components? Why or why not?

Modeling



Metrics	Approach	#features	Model size	Feature data siz
Avg CPU utilization	Random Forest	127	312 KB	376 MB
P95 max CPU utilization	Random Forest	127	311 KB	376 MB
Deployment size in #VMs	Extreme Gradient Boosting Tree	24	305 KB	368 MB
Deployment size in #cores	Extreme Gradient Boosting Tree	24	305 KB	368 MB
Lifetime	Extreme Gradient Boosting Tree	127	329 KB	376 MB
Workload class	FFT, Extreme Gradient Boosting Tree	34	152 KB	311 MB

Table 1: Metrics, ML modeling approaches, model and full feature dataset sizes.

Discussion: What are some strengths and weaknesses of their modeling approach? Aspects to consider include fidelity, performance, and ability to handle outliers.

Discussion: How exactly can the prediction results inform scheduling systems to improve resource utilization?

Novelties & Strengths

- Performed a careful analysis of production VM workloads of a real cloud provider (i.e., Azure) and made this data public.
- Provided convincing evidence that certain VM characteristics can be predicted with high fidelity.
- Created a performant and robust prediction system (i.e., Resource Central).

🤔 What figure did you find most illuminating and why?

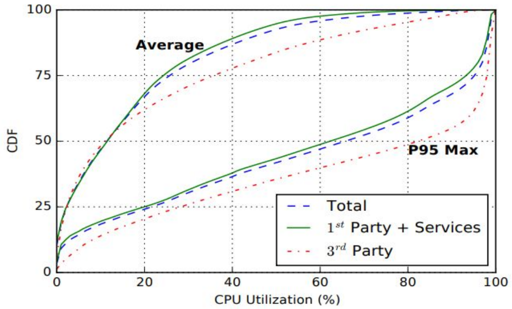


Figure 1: Average and P95 of max CPU utilizations.

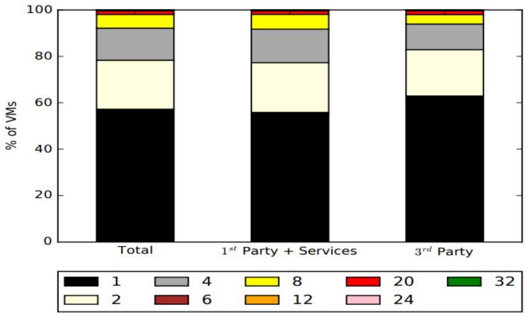


Figure 2: Number of virtual CPU cores per VM.

Critique & Weaknesses

- IOPS analysis should differentiate between reads and writes, as the former is far cheaper than the latter.
- No graph or figure for the evaluation of how RC-informed scheduling (§6.2) can improve CPU utilization over baseline. Also, the cost of having >100% utilization is unclear, so it is difficult to evaluate the scheduling policies.
- Too much discussion of result caching, which is neither novel nor interesting.
- Interarrival times are remarked on and said to be easily modeled (§3.7), but are not used by Resource Central for predictions, even for VM Lifetime.

Discussion: What other deficiencies are present in the approach of the paper? What about in the way it was presented?

Thanks for your participation!