# Serverless Boom or Bust?
# An Analysis of Economic Incentives

Xiayue Charles Lin, Joseph E. Gonzalez, Joseph M. Hellerstein
*UC Berkeley*

HotCloud 2020 · Monday, July 13

# An Economic Model for Serverless

- Serverless: pay for consumption instead of capacity
- In broad strokes, when is serverless advantageous?

Why an economic model?

- Serverless is exciting, but still in its relative infancy - provisioned servers are far from being replaced
- Inform research and build intuition:
  - Which parts of the design space are economically sensible?
  - Which directions have transformative potential?

# Questions we want to reason about

- Gauging how “compelling” arbitrary improvements are
  - Suppose a new paper shows “some technique can reduce straggler latency by 2x for serverless application Y”
  - What does that mean? Is it game changing? Does this enable the previously infeasible?
  - What if cloud vendors change serverless prices in the future? Instead of redoing benchmarks, intuitively reason about whether serverless Y makes sense fundamentally or temporarily

- Informing design decisions
  - Autoscaling policies
  - Pricing Quality of Service

# Developing the model

- The constraint: for any serverless product to be viable,
  **both the provider and the customer must prefer** it to a serverful option

- For the provider, we assume profit to be the most important
  - Serverless product should bring in at least as much revenue as if the resources were spent on a serverful product instead
  - We consider utilization ratio and price ratio (for any particular vendor and product)

$$c > \frac{p_r}{p_s}$$

  - “Resource underutilization from serverless
    must be compensated by higher product price”

# Developing the model

- The customer also faces an analogous price-to-utilization tradeoff:
  the premium they pay for serverless
  must be worth the time they would waste if they provisioned a serverful product

$$\frac{p_s}{p_r} < \frac{t_r}{t_s}$$

# Developing the model

- Another decision factor: the customer may also find serverless less “useful”
  - **Specialized hardware** requirements?
  - **Quality of service** requirements?
  - Transition cost, operational concerns, lock-in concerns…

# Developing the model

- Another decision factor: the customer may also find serverless less “useful”
  - **Specialized hardware** requirements?
  - **Quality of service** requirements?
  - Transition cost, operational concerns, lock-in concerns…

- Can model this as a binary variable, but might as well make it continuous:

$$\frac{p_s}{p_r} < \frac{t_r}{t_s} \cdot \frac{v_s}{v_r}$$

(teal term represents relative usefulness
of serverless product over serverful)

# Developing the model

- Combining the provider and customer models:

$$c^{-1} < \frac{p_s}{p_r} < \frac{t_r}{t_s} \cdot \frac{v_s}{v_r}$$

- On the two ends: **how much better providers are at using resources** than individual customers, and **how useful serverless products are**
- Price ratio serves as a public bound for these otherwise opaque terms
- For brevity, we will denote the customer characteristics as $\alpha$

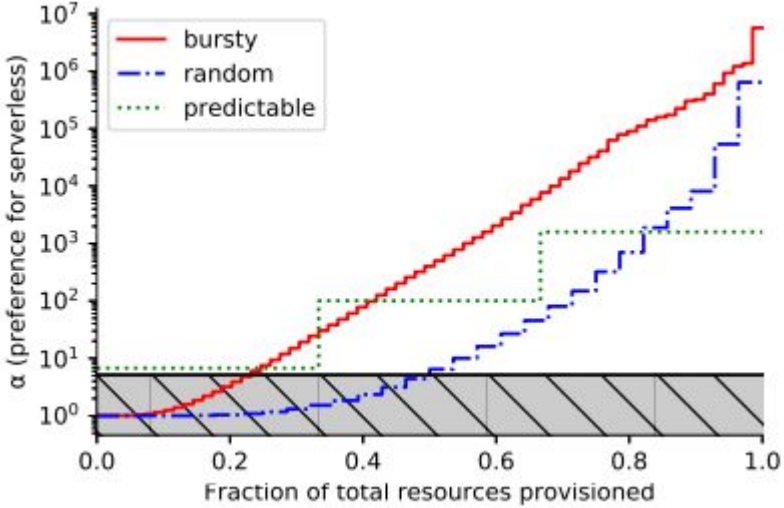
# Classifying customers

Individual customers (and use cases) have different characteristics, and thus different levels of alpha. All levels of alpha fall into **one of three categories**:

- alpha < 1
  - No amount of utilization or price improvements will help them; **more useful serverless products are required**
- 1 < alpha < $c^{-1}$
  - These customers prefer consumption-based versus capacity-based pricing if possible, but **providers cannot profitably serve them yet**
- alpha > $c^{-1}$
  - Providers can profitably provide serverless products to these customers

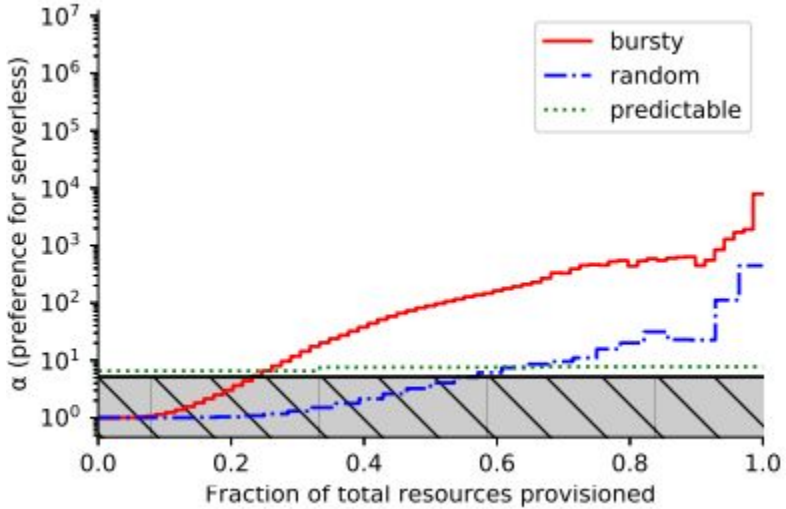
# Examining Autoscaling

- Increasing granularity from customers to their individual provisioned resources (e.g. individual VMs)
- Simulation: A customer provisions for peak to serve a generic job queue



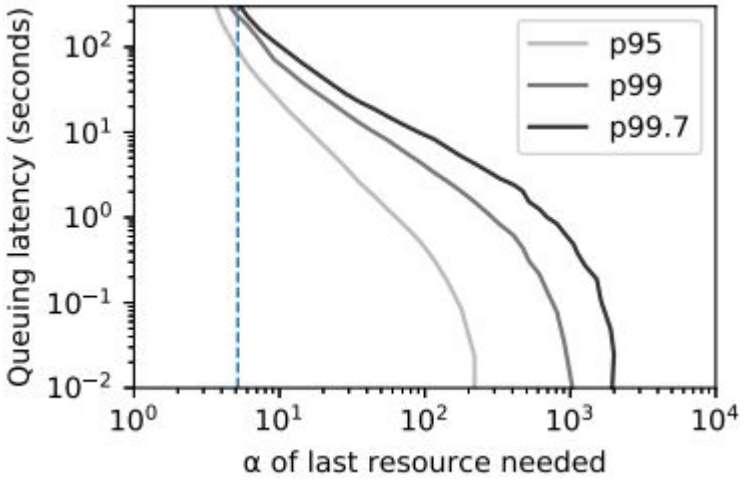
# Examining Autoscaling

- Oracle provisioning (1-minute windows) does not substantially change today's breakeven points for preferring serverless:



# Examining quality of service

- Customer might not know peak, or deliberately underprovision anyway, which incurs queuing latency
- Alpha of last VM needed to reduce queuing latency below a p(xx) target



# Conclusions

- Serverless systems that are price-competitive with serverful designs are to be expected, and we will inevitably see more of these - especially as specialized hardware enters serverless

- We should explore a mix of provisioned capacity and pay-for-consumption (“**hybrid serverless designs**”)

- We should **think consciously about incentives and tradeoffs** to consumers when designing policies for new serverless systems

# Thank you!

Charles Lin · charles.lin@berkeley.edu
Joseph Gonzalez · jegonzal@berkeley.edu
Joseph Hellerstein · hellerstein@berkeley.edu