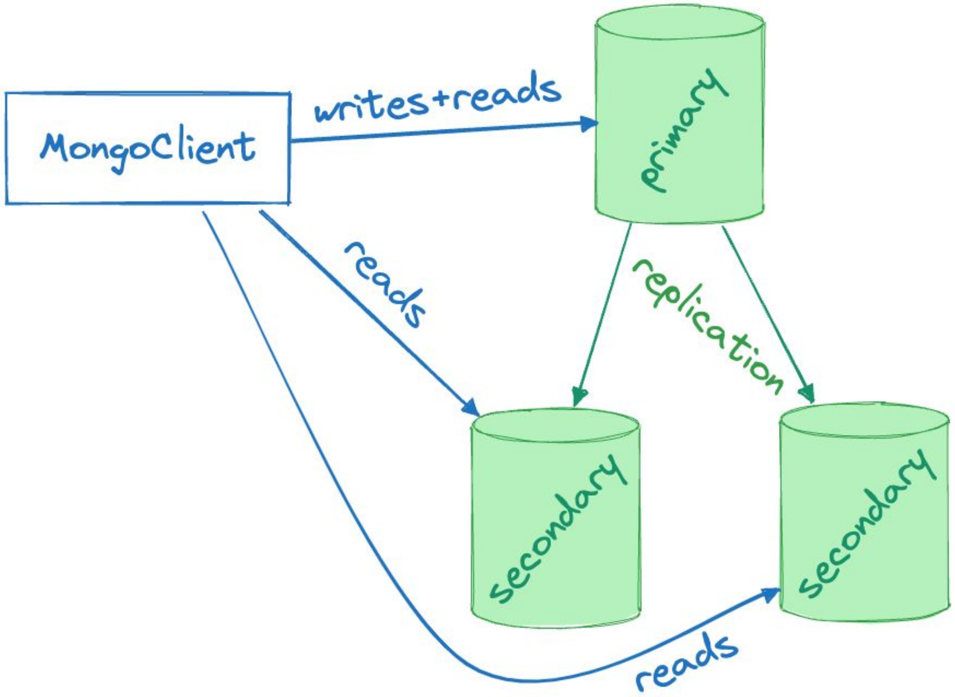
Predictive Scaling in MongoDB Atlas, an Experiment

Matthieu Humeau and A. Jesse Jiryu Davis
Data Council 2024

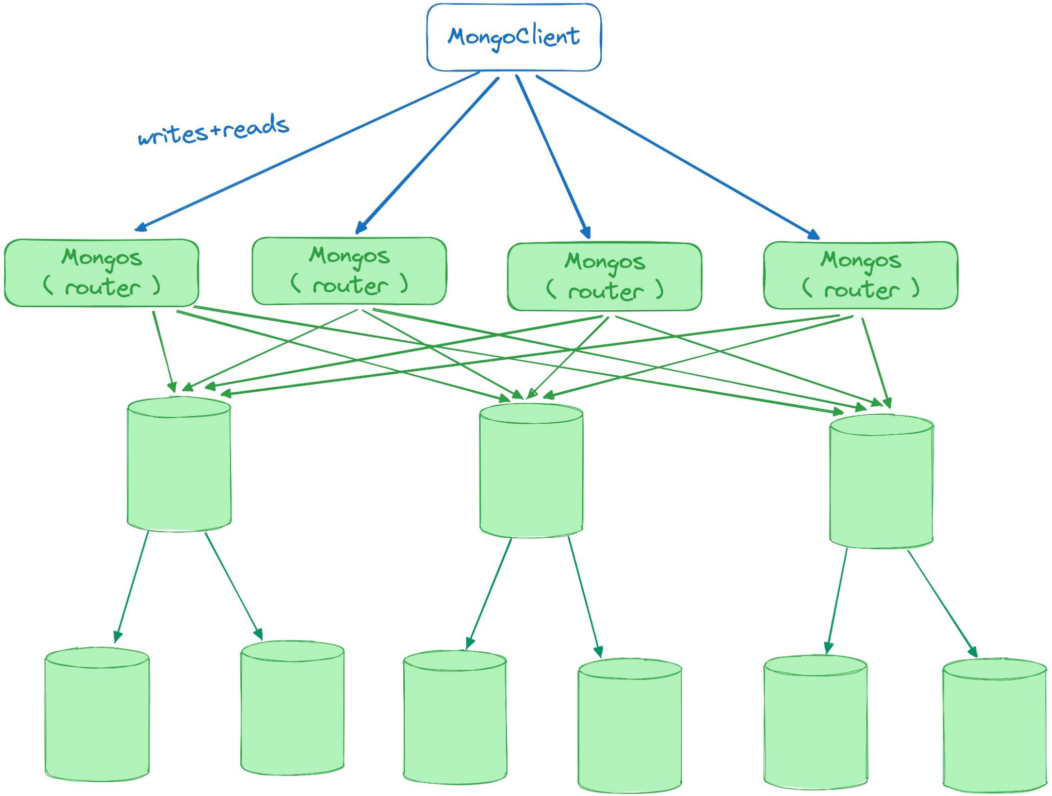
MongoDB®

- NoSQL, document database
- MongoDB Query Language
- High consistency, high availability, ACID transactions

MongoDB replica set

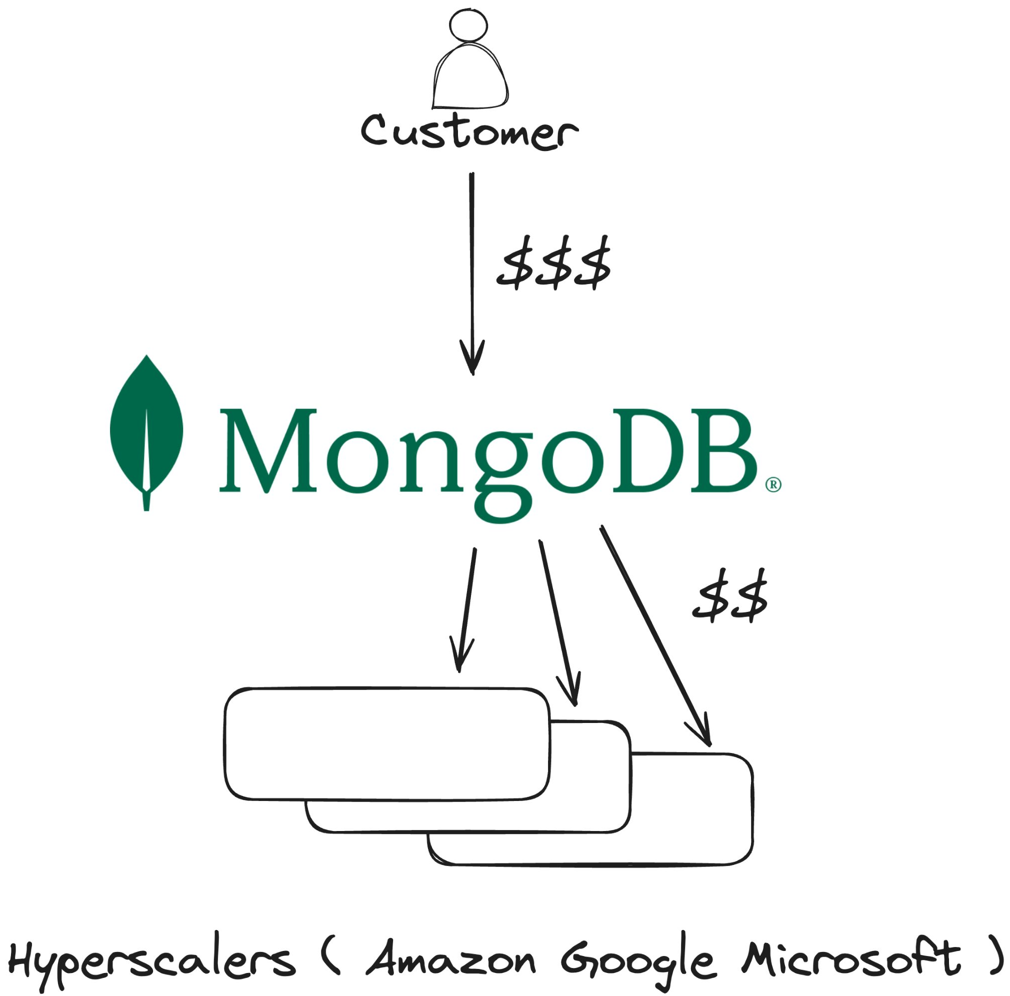


MongoDB sharded cluster



MongoDB Atlas

- "Developer Data Platform": DBaaS and much more
- Multi-region, multi-cloud
- MongoDB's cloud is actually the hyperscalers' clouds


Customer
$$$
MongoDB
$$
Hyperscalers (Amazon Google Microsoft)

MongoDB Atlas tiers

Cluster Tier	Storage	RAM	vCPUs	Base Price
M10	10 GB	2 GB	2 vCPUs	$0.08/hr
M20	20 GB	4 GB	2 vCPUs	$0.20/hr
M30	40 GB	8 GB	2 vCPUs	$0.54/hr
M40	80 GB	16 GB	4 vCPUs	$1.04/hr
M50	160 GB	32 GB	8 vCPUs	$2.00/hr
M60	320 GB	64 GB	16 vCPUs	$3.95/hr

MongoDB Atlas vertical scaling

Take a secondary offline.

Detach its network storage.

Restart it with a different server size.

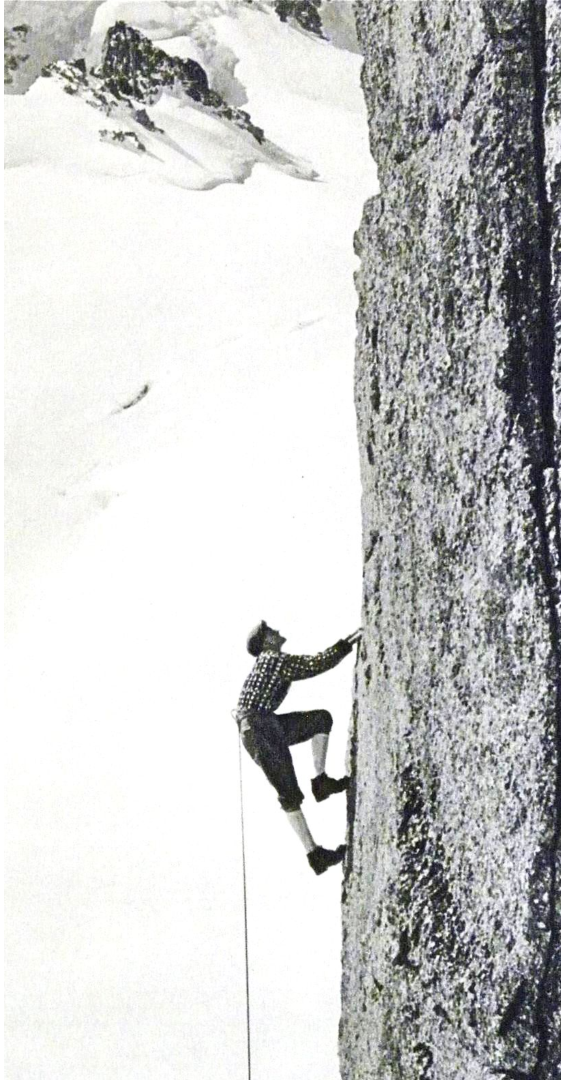
Reattach storage.

Wait for it to catch up to the primary.

Scale the other secondary likewise.

Step down the primary and scale it.

Scaling takes ~15 minutes



Auto-scaling Today

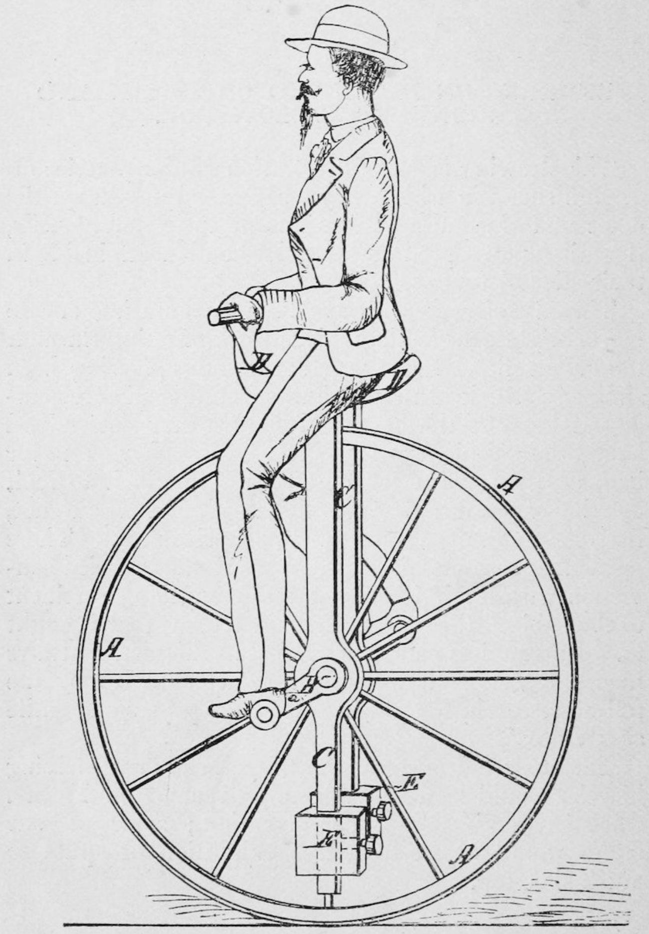
We scale infrequently and **reactively**:

- scale up by one tier after 1 hour of overload,

- scale down by one tier after 24 hours of underload.

Clusters can be over/underloaded for long periods!

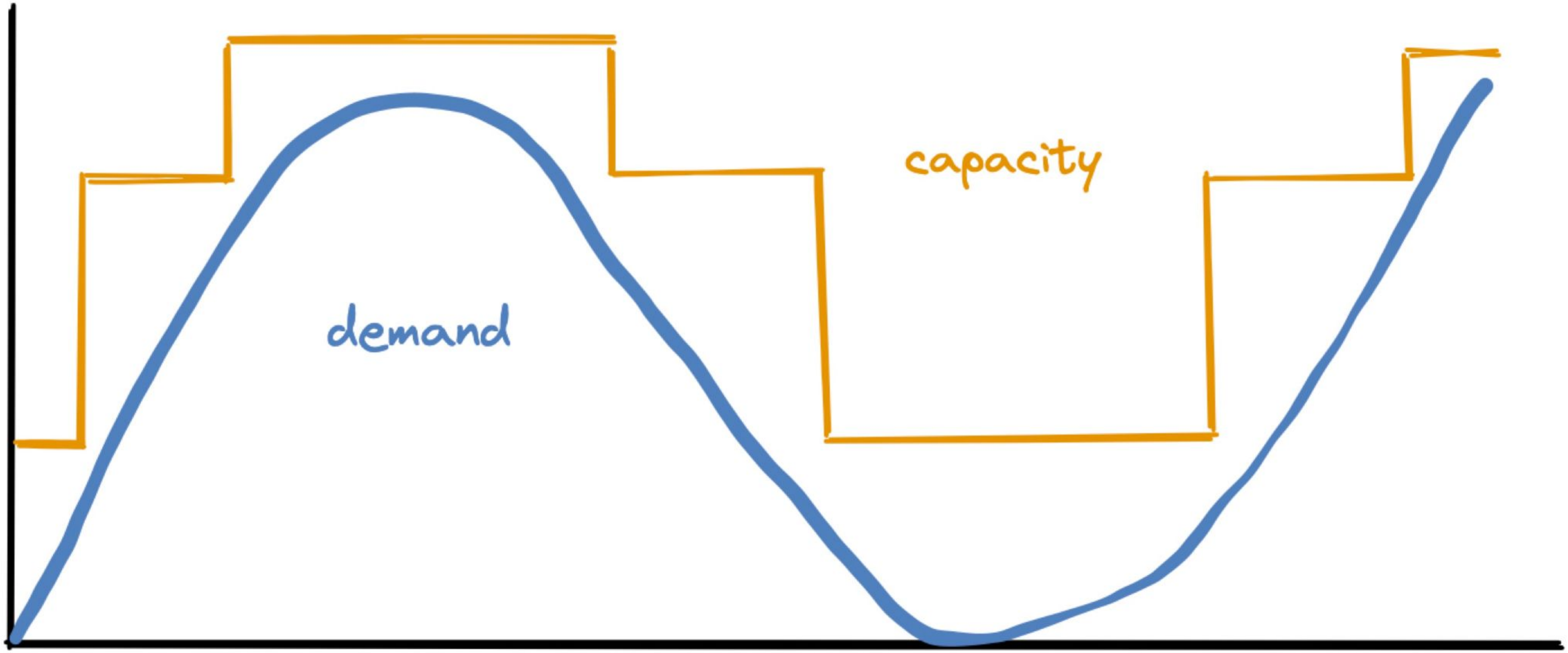
Clusters are **already** overloaded
when they start scaling up.



T. W. Ward, of New York. Velocipede. No. 88,683. Patented April 6, 1869.


IDEAL FUTURE

Ideal Future



Ideal Future

Forecast each cluster's resource needs.

Scale a cluster up **before** it's overloaded.

Scale it down **as soon as** it's underloaded.

Scale **directly** to the right size, **skipping** intermediate tiers.

Predictive Scaling Experiment

We keep servers' performance metrics in a data warehouse, 1-minute intervals.

We chose 10,000 clusters, analyzed their 2023 history.

Split the history into a training period and a testing period.

Trained models to forecast the clusters' demand and CPU utilization.

Guessed how a predictive scaler would've performed during testing period, compared to the reactive scaler that was running at that time.

Predictive Scaling Experiment

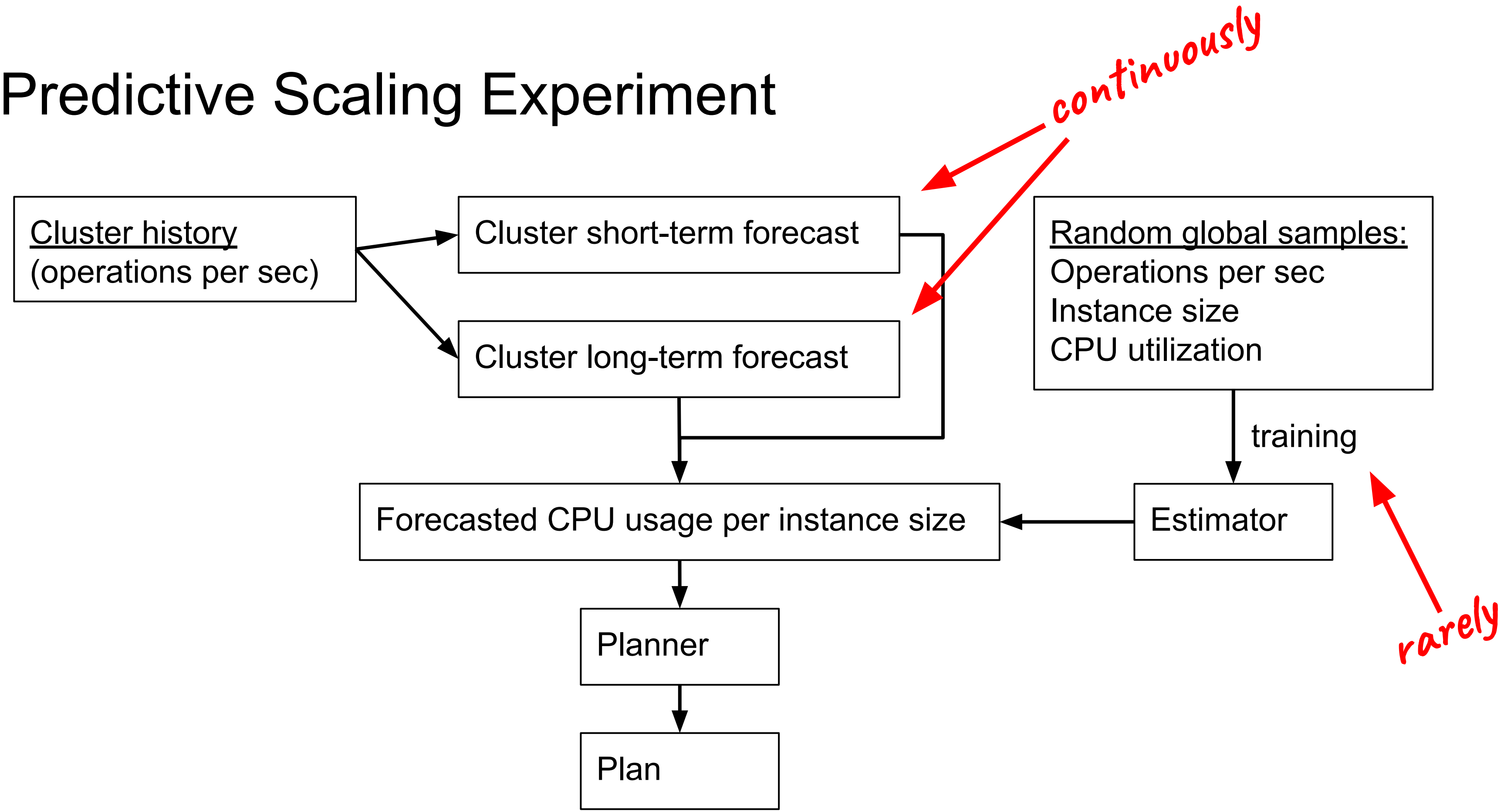
Three components:

Forecaster: forecasts each cluster's future workload.

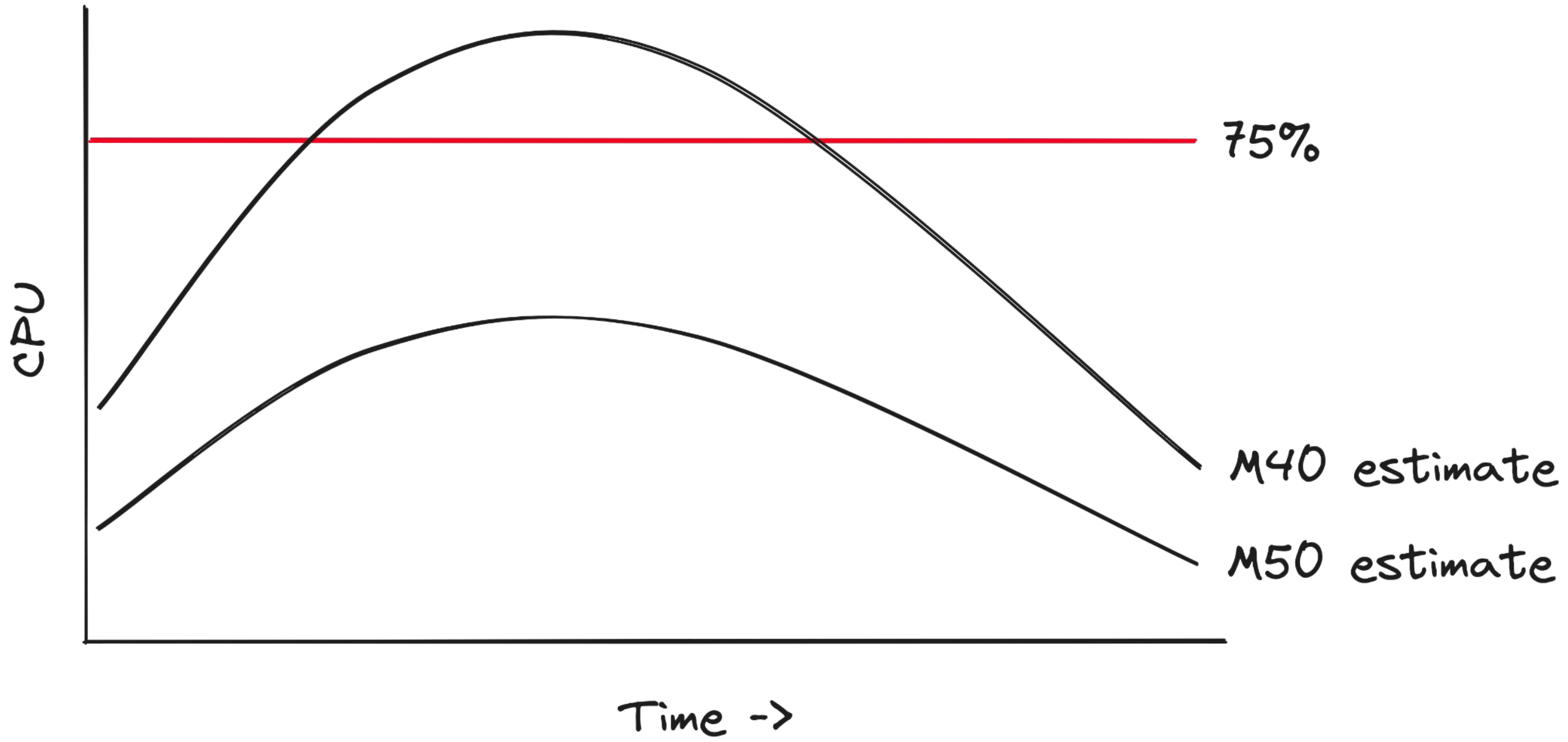
Estimator: estimates CPU% for any workload, any instance size.

Planner: chooses cheapest instance that satisfies forecasted demand.

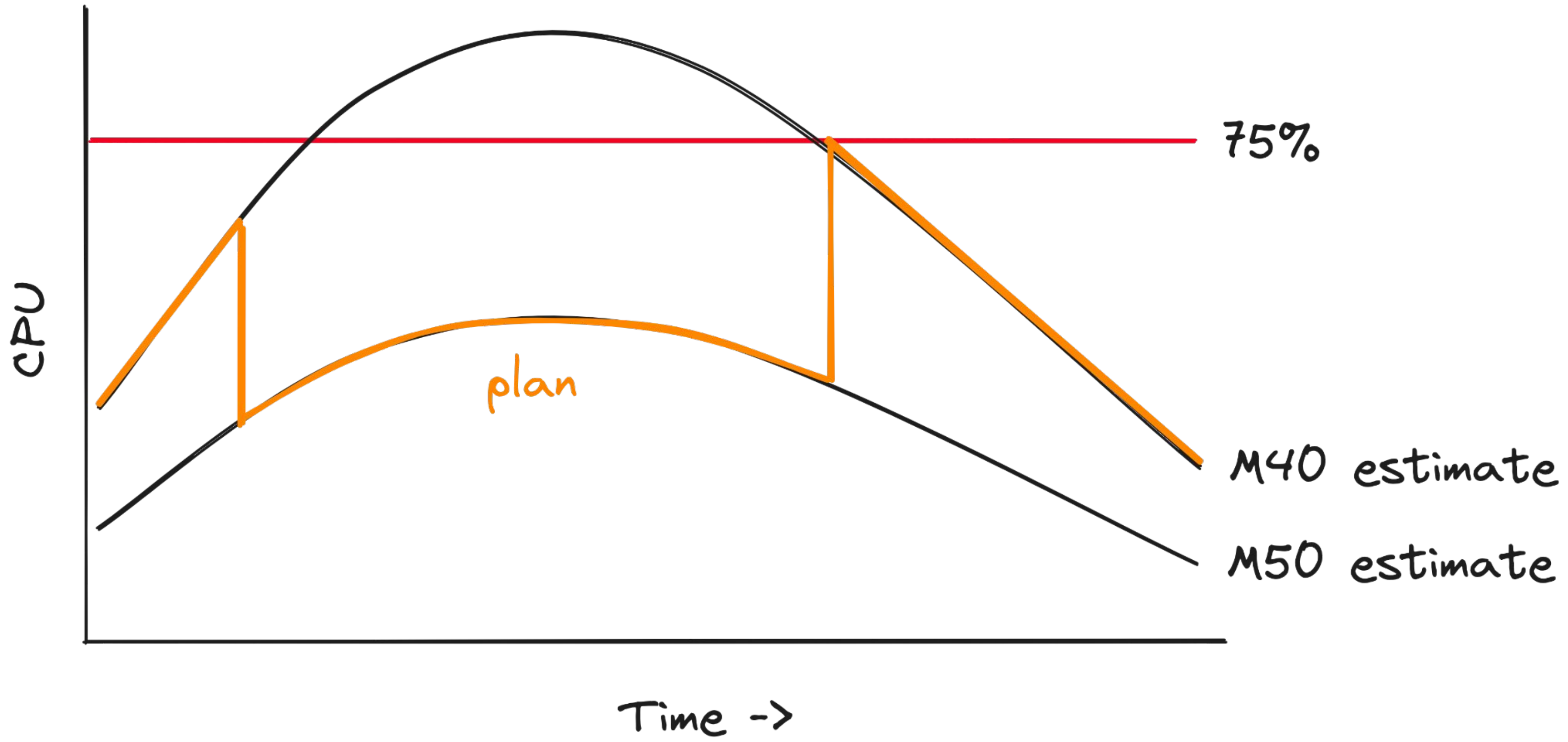
Predictive Scaling Experiment



Predictive Scaling Experiment: Planner



Predictive Scaling Experiment: Planner



Predictive Scaling Experiment: Forecaster (Long-Term)

Forecast customer-driven metrics, such as:

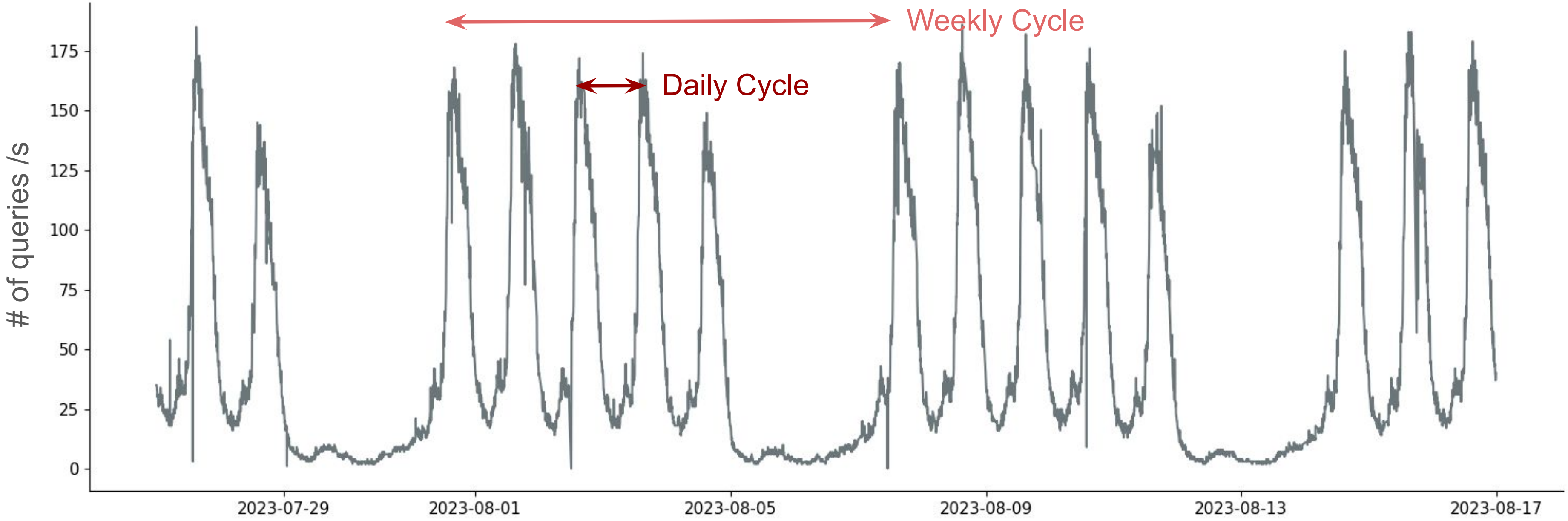
- Queries/s
- Connections
- Query complexity represented by number of items scanned by the solver
- DB size

These metrics are **independent*** from the instance size and the state of the cluster

**(almost)*

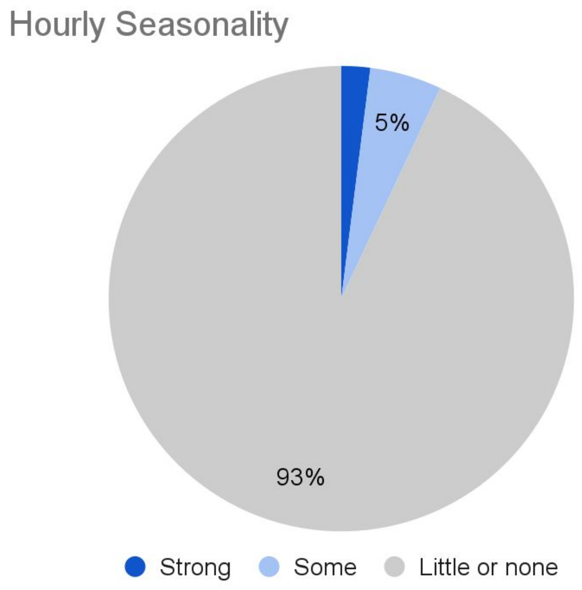
Predictive Scaling Experiment: Forecaster (Long-Term)

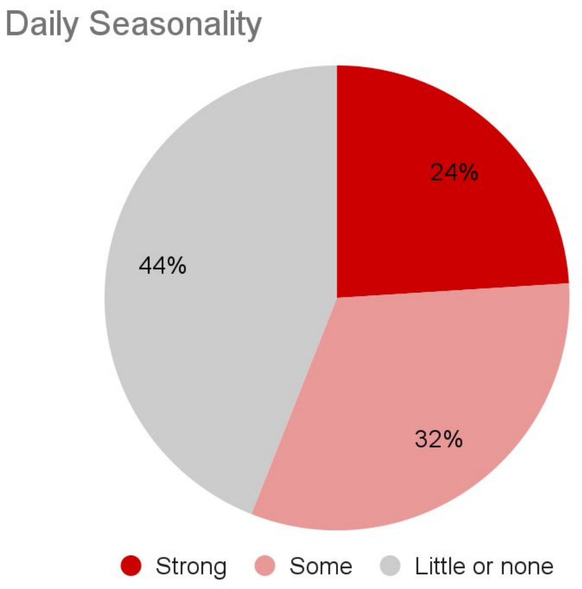
Predicting seasonal variations in demand

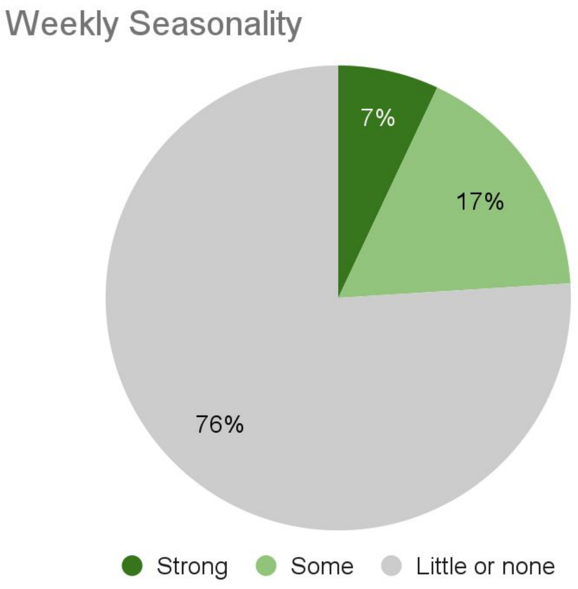


Predictive Scaling Experiment: Forecaster (Long-Term)

How often is demand seasonal?



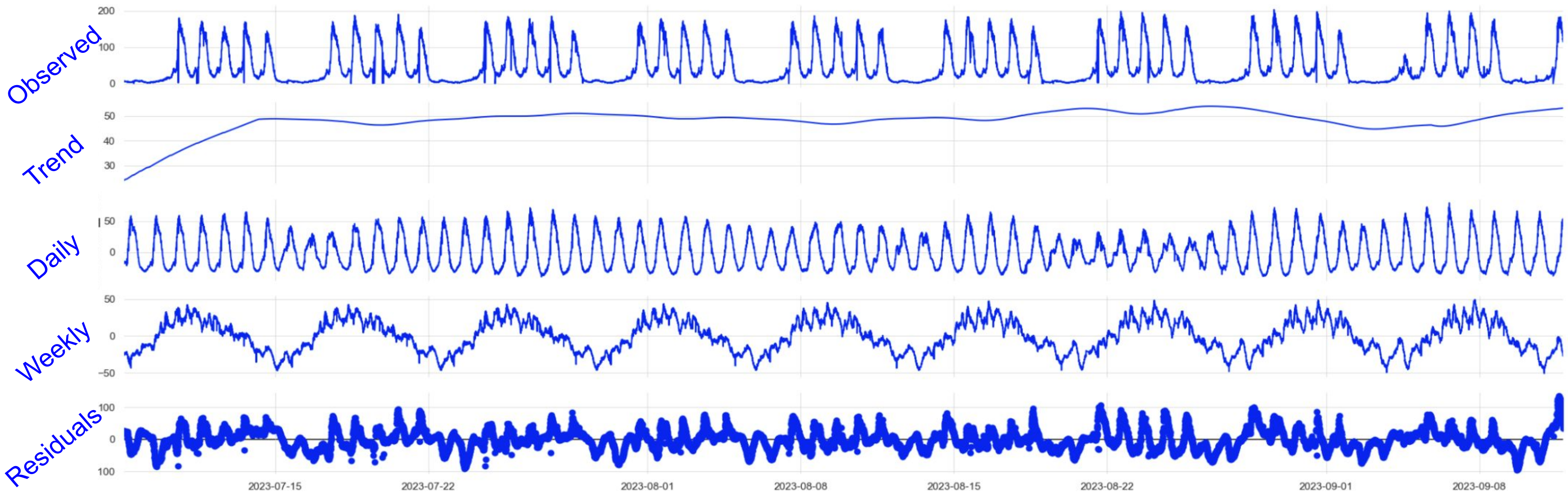




Predictive Scaling Experiment: Forecaster (Long-Term)

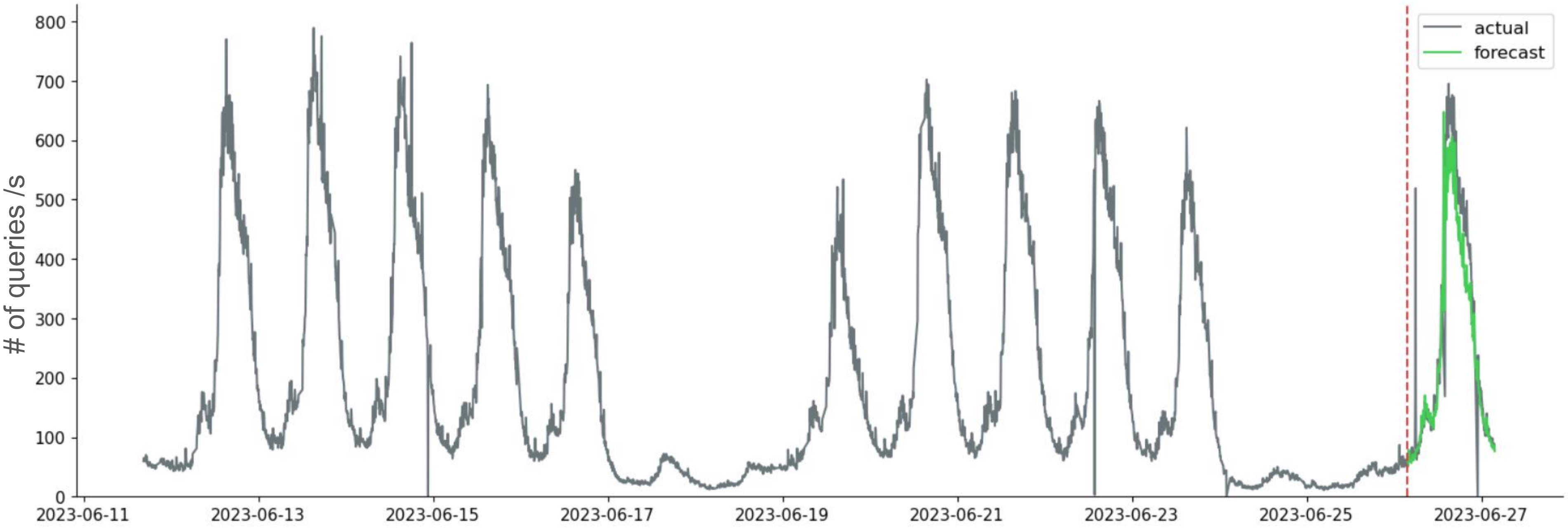
Best Forecasting Model:

- Multi-Season Trend decomposition using LOESS (MSTL)
- + ARIMA forecast of residuals



Predictive Scaling Experiment: Forecaster (Long-Term)

Forecast example



Predictive Scaling Experiment: Forecaster (Long-Term)

How accurate is it?

- Median MAPE (Mean Abs. Perc. Error)

	Seasonal Clusters	Non-seasonal Clusters
Connections	3%	50%
Query Rate	19%	71%
Scanned objects Rate	27%	186%

Not usable

=> Self-censoring mechanism based on forecast confidence

Predictive Scaling Experiment: Forecaster (Long-Term)

How accurate is it?

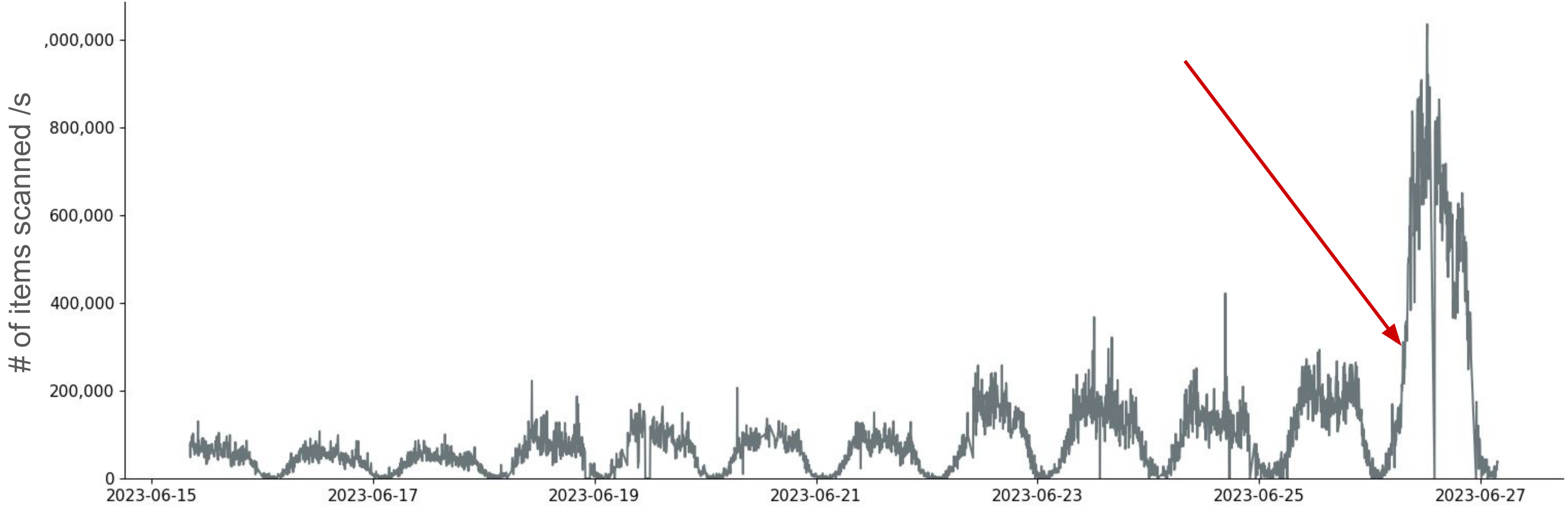
- Median MAPE (Mean Abs. Perc. Error)

	Seasonal Clusters	**Non-seasonal Clusters**
Connections	3%	50%
Query Rate	19%	71%
Scanned objects Rate	27%	186%

Not usable

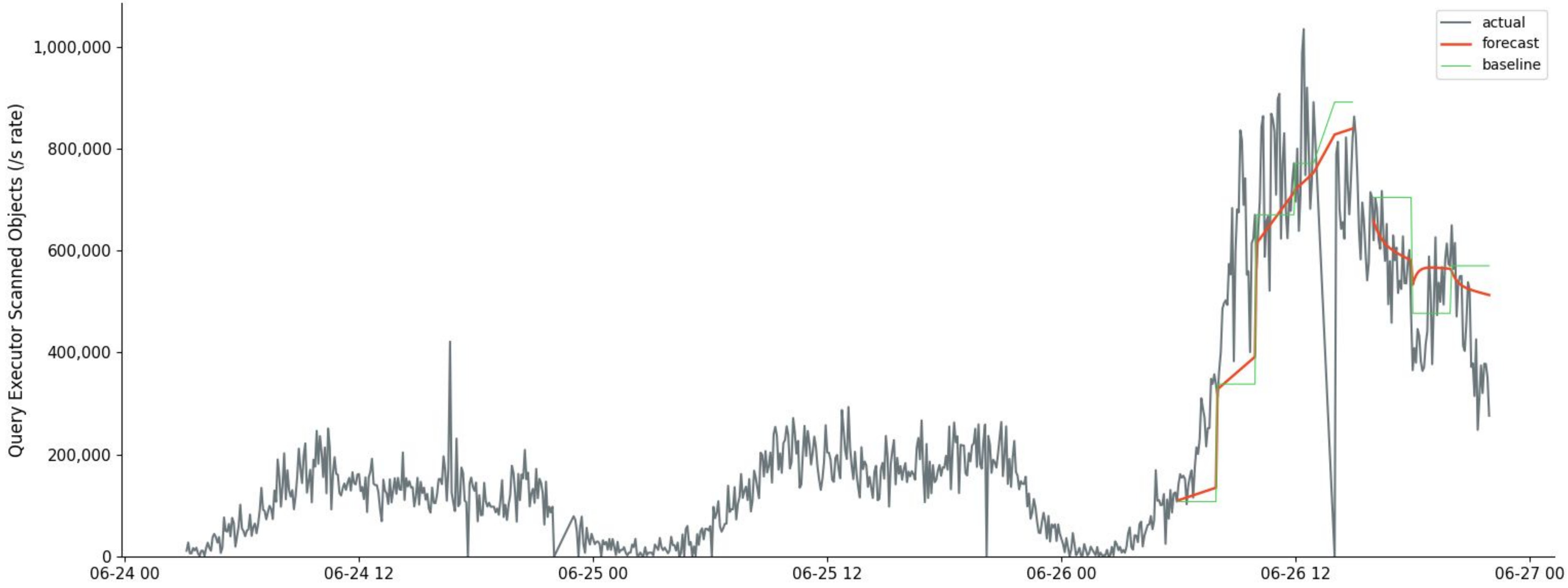
Predictive Scaling Experiment: Forecaster (Short-Term)

What about unexpected (= non-seasonal) changes in demand?

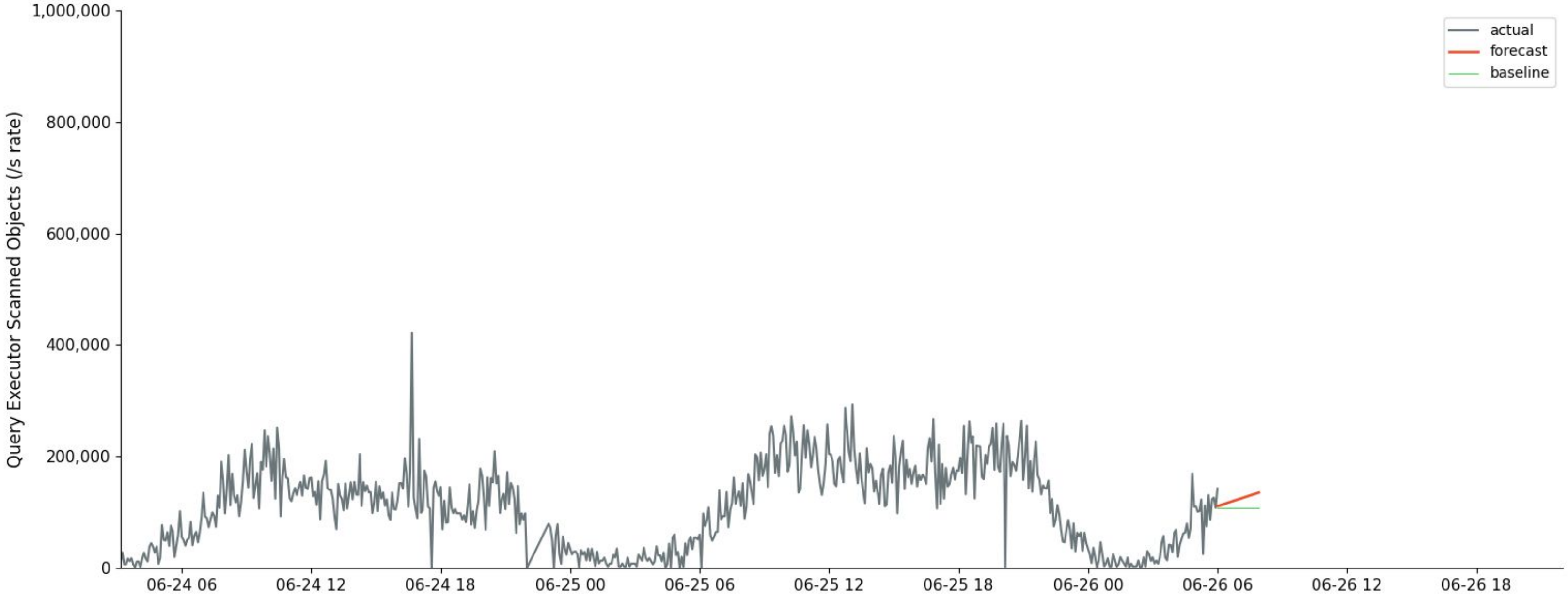


Predictive Scaling Experiment: Forecaster (Short-Term)

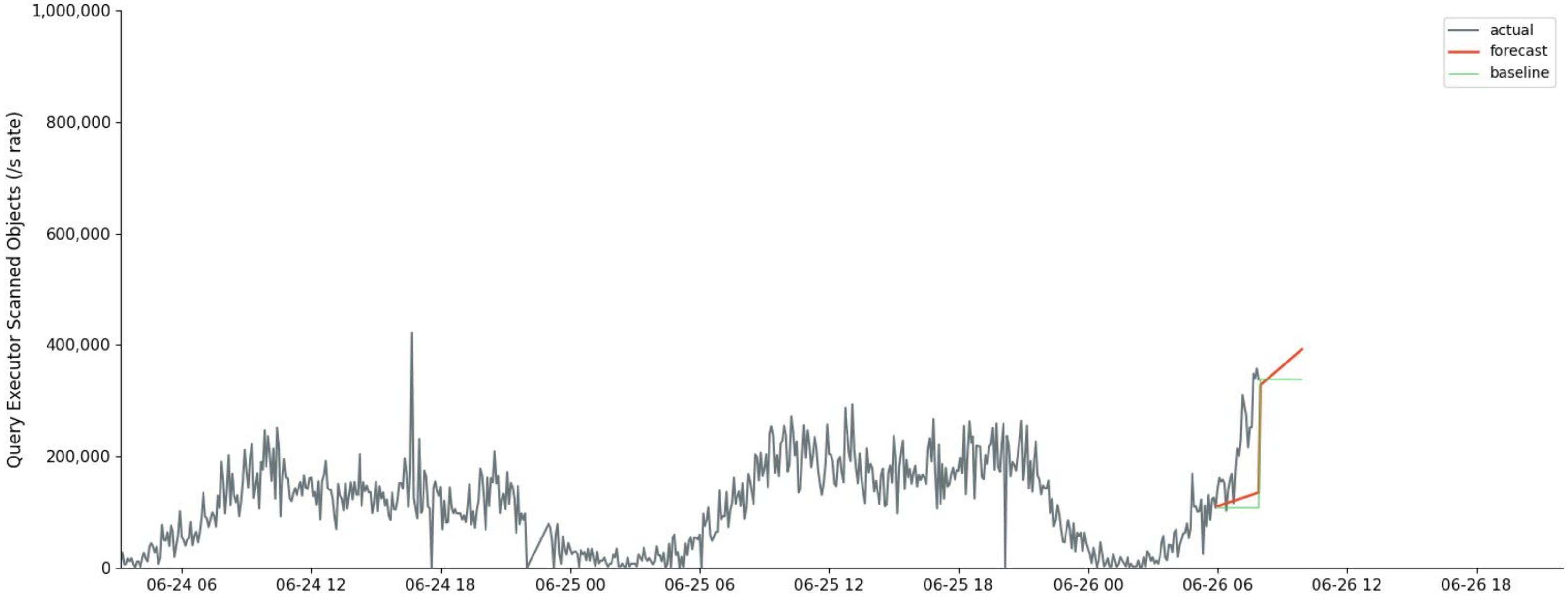
Approximation of local trends for near future



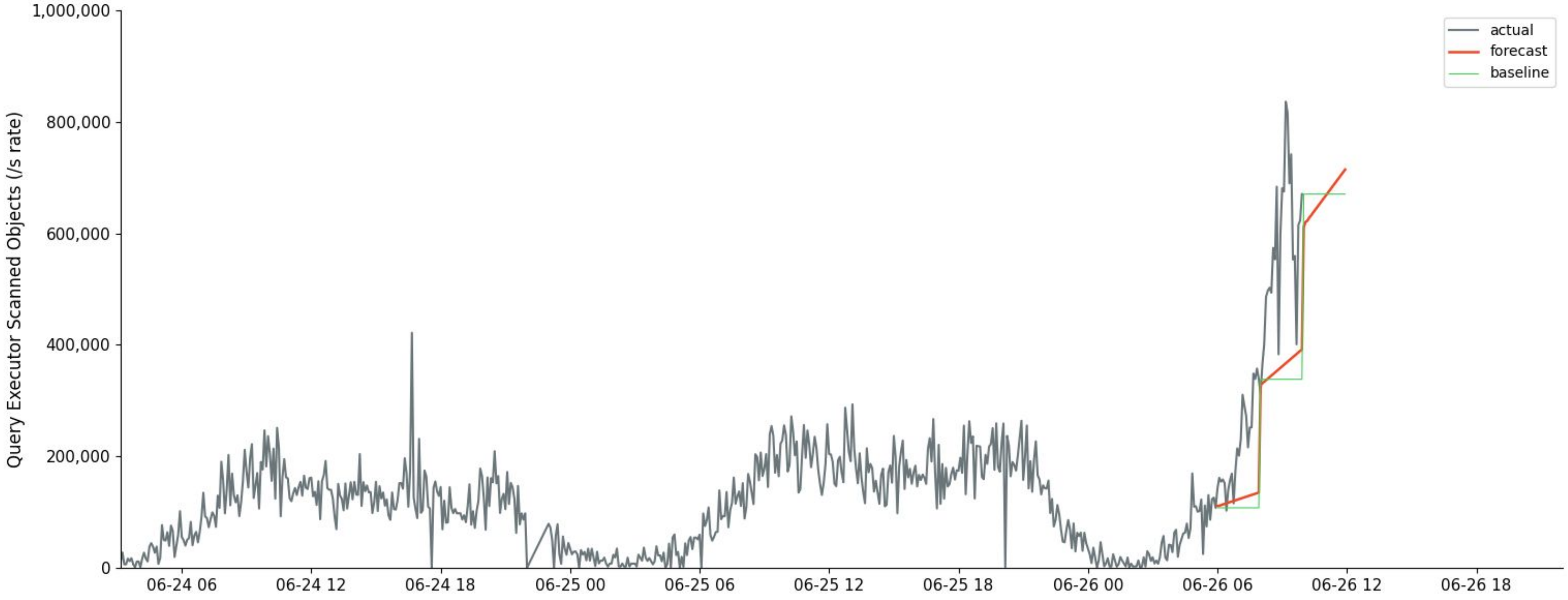
Predictive Scaling Experiment: Forecaster (Short-Term)



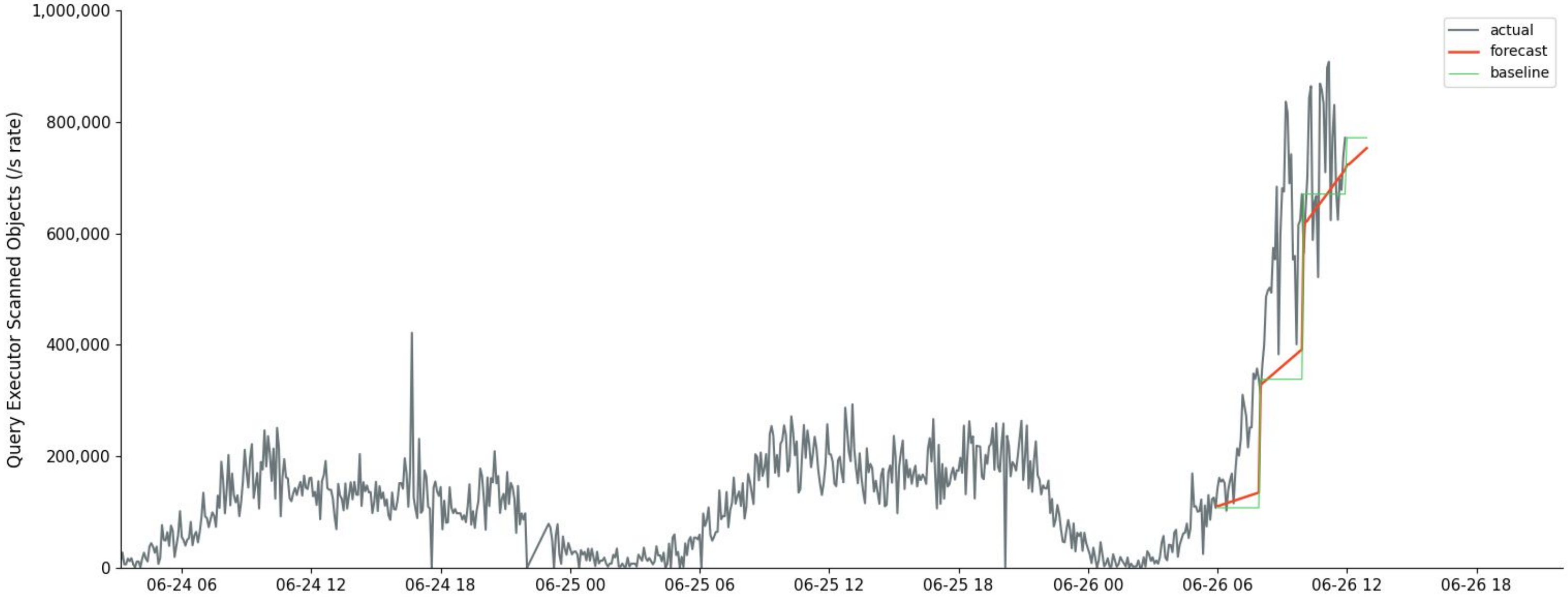
Predictive Scaling Experiment: Forecaster (Short-Term)



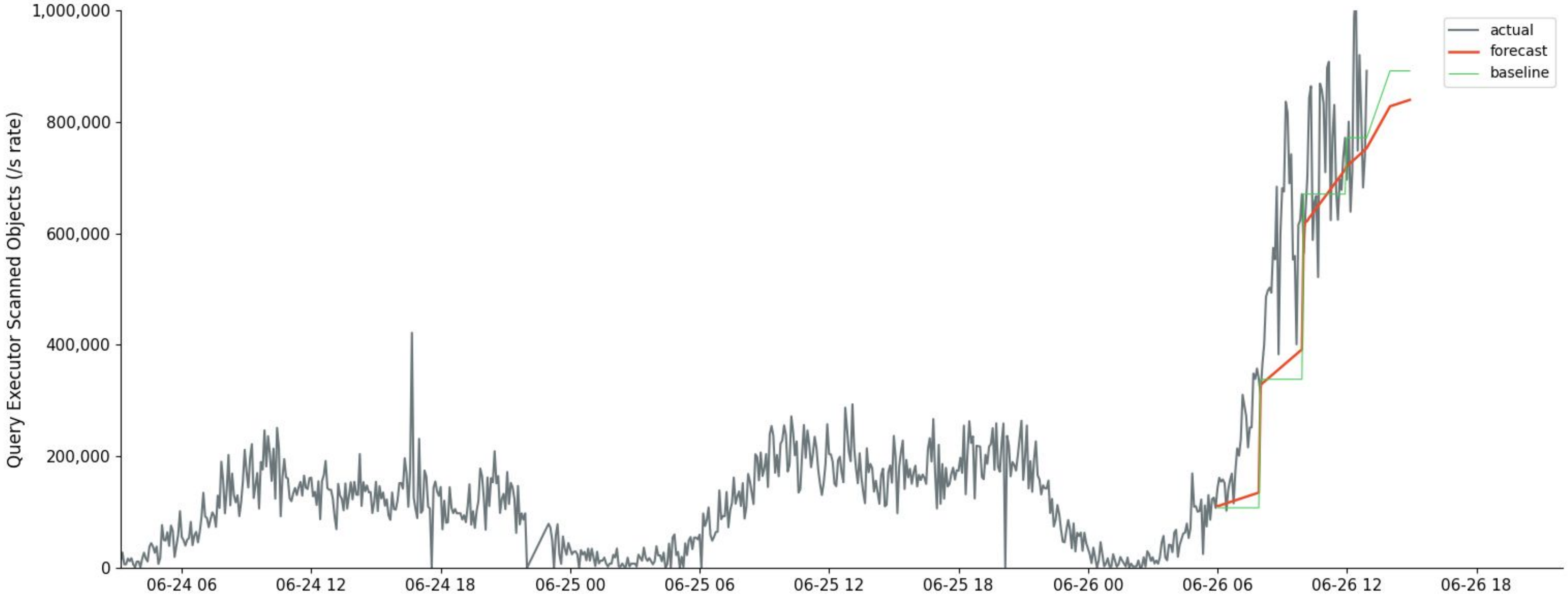
Predictive Scaling Experiment: Forecaster (Short-Term)



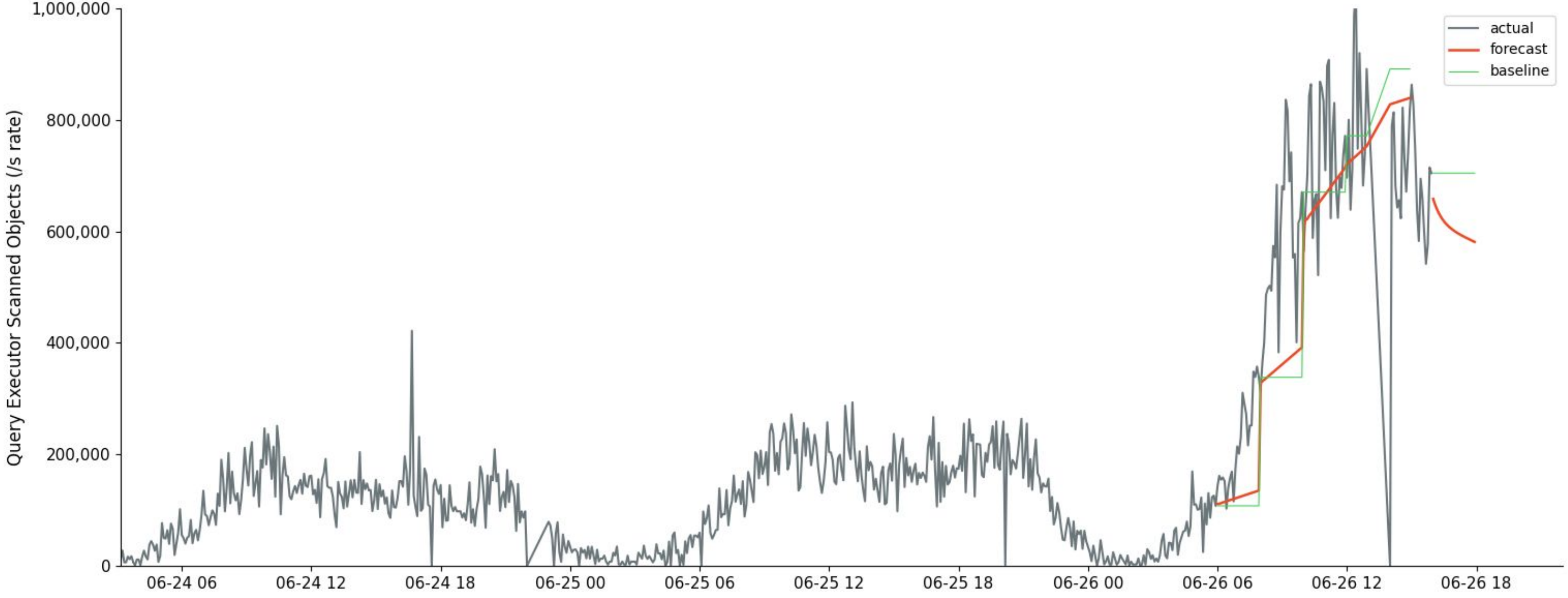
Predictive Scaling Experiment: Forecaster (Short-Term)



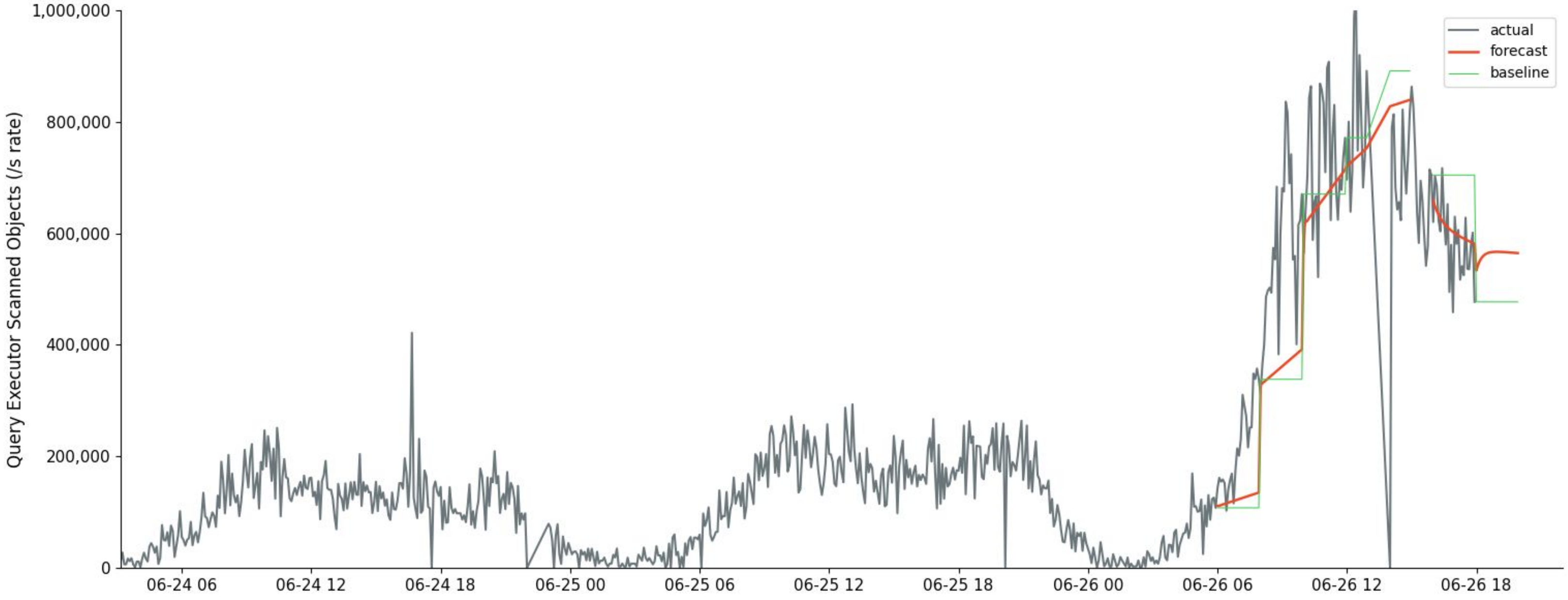
Predictive Scaling Experiment: Forecaster (Short-Term)



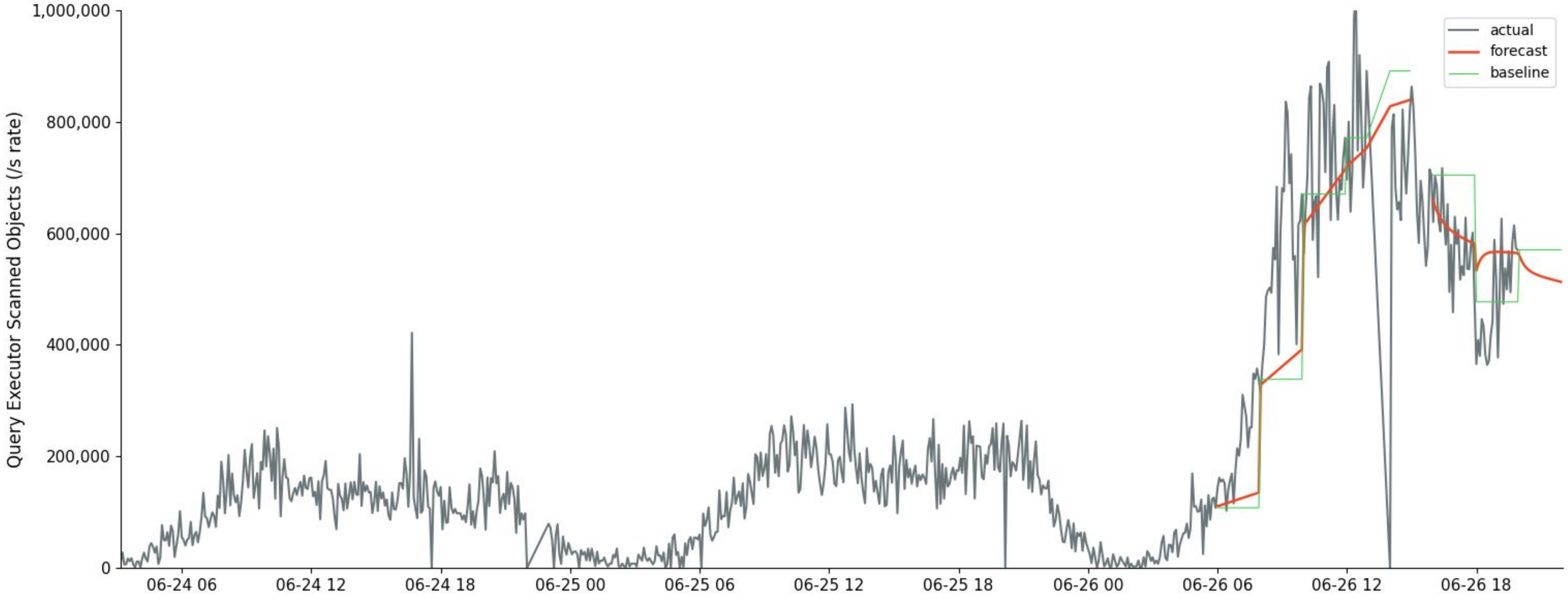
Predictive Scaling Experiment: Forecaster (Short-Term)



Predictive Scaling Experiment: Forecaster (Short-Term)



Predictive Scaling Experiment: Forecaster (Short-Term)



Predictive Scaling Experiment: Forecaster (Short-Term)

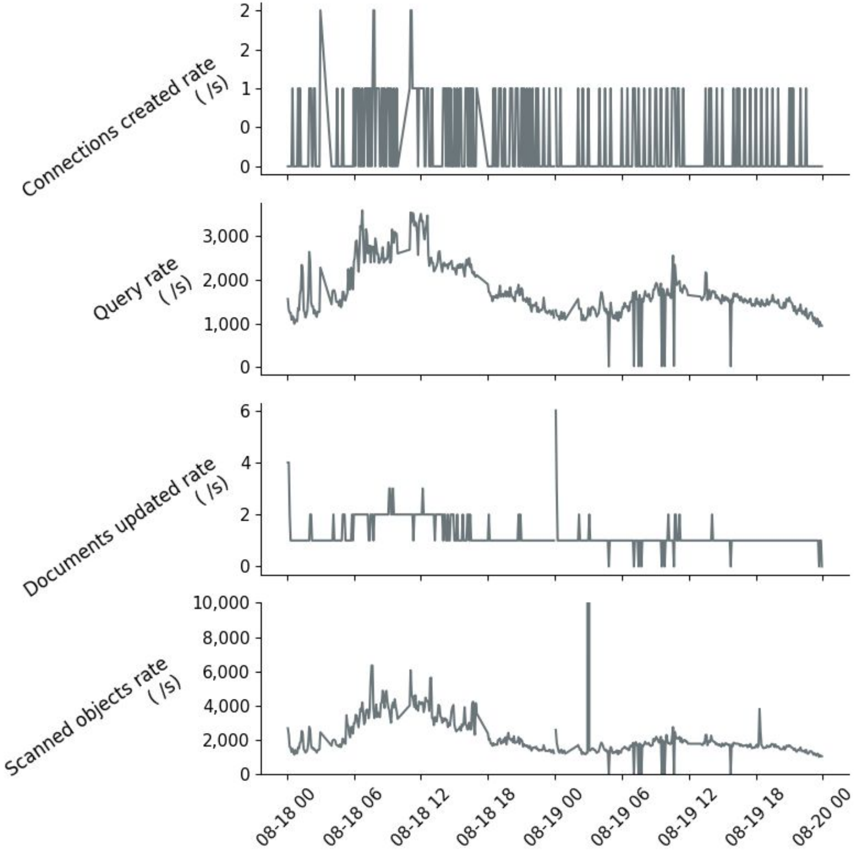
We compared to naive approach: future will look like the last observation

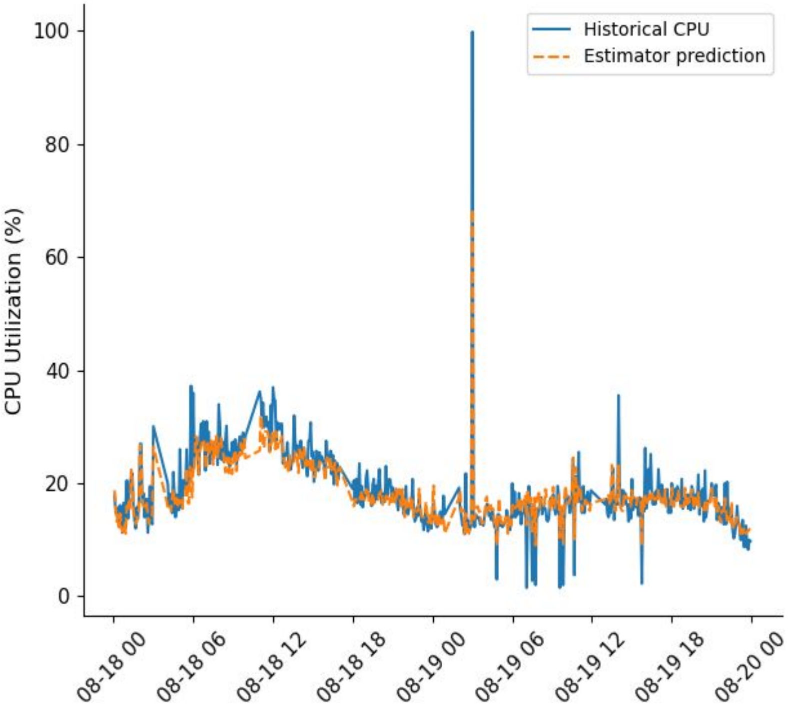
Local trend approach beats it 68% of the time (29% reduction in error)

Predictive Scaling Experiment: Estimator

Predictive Scaling Experiment: Estimator

Predictive Scaling Experiment: Estimator





Predictive Scaling Experiment: Estimator

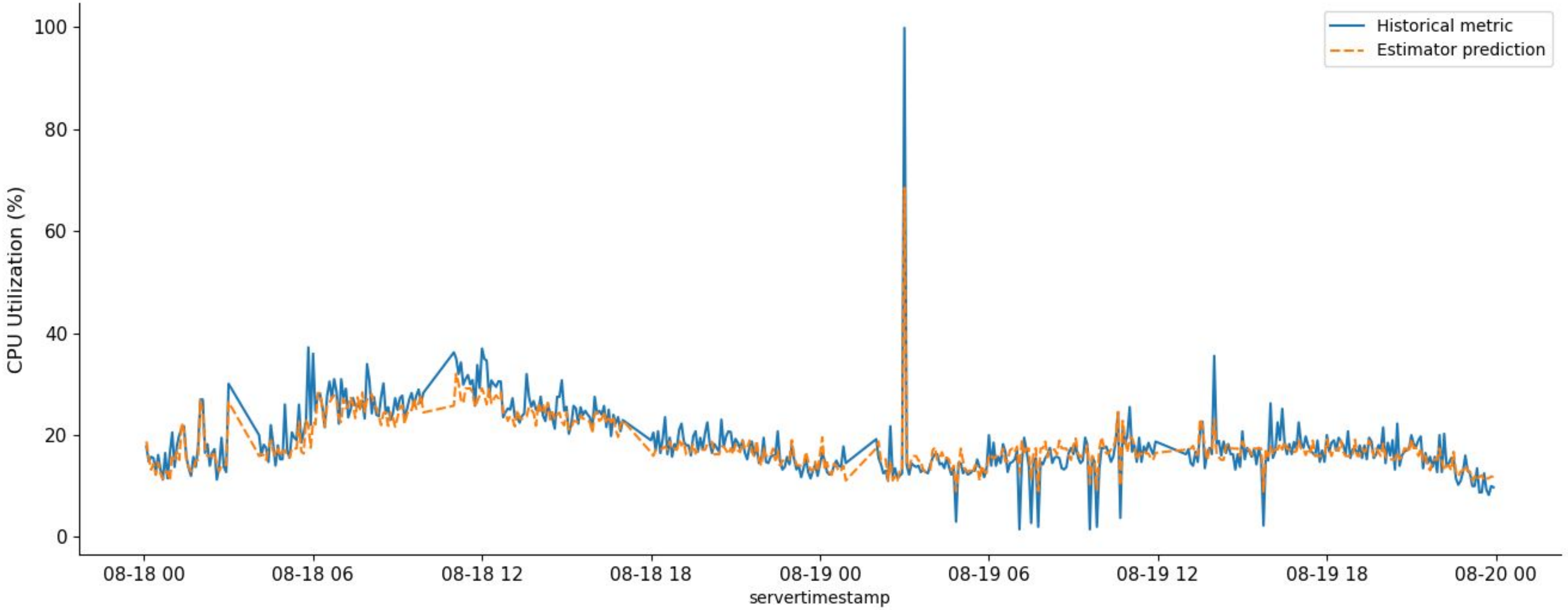
In ML terms: regression problem to predict CPU utilization

Model:

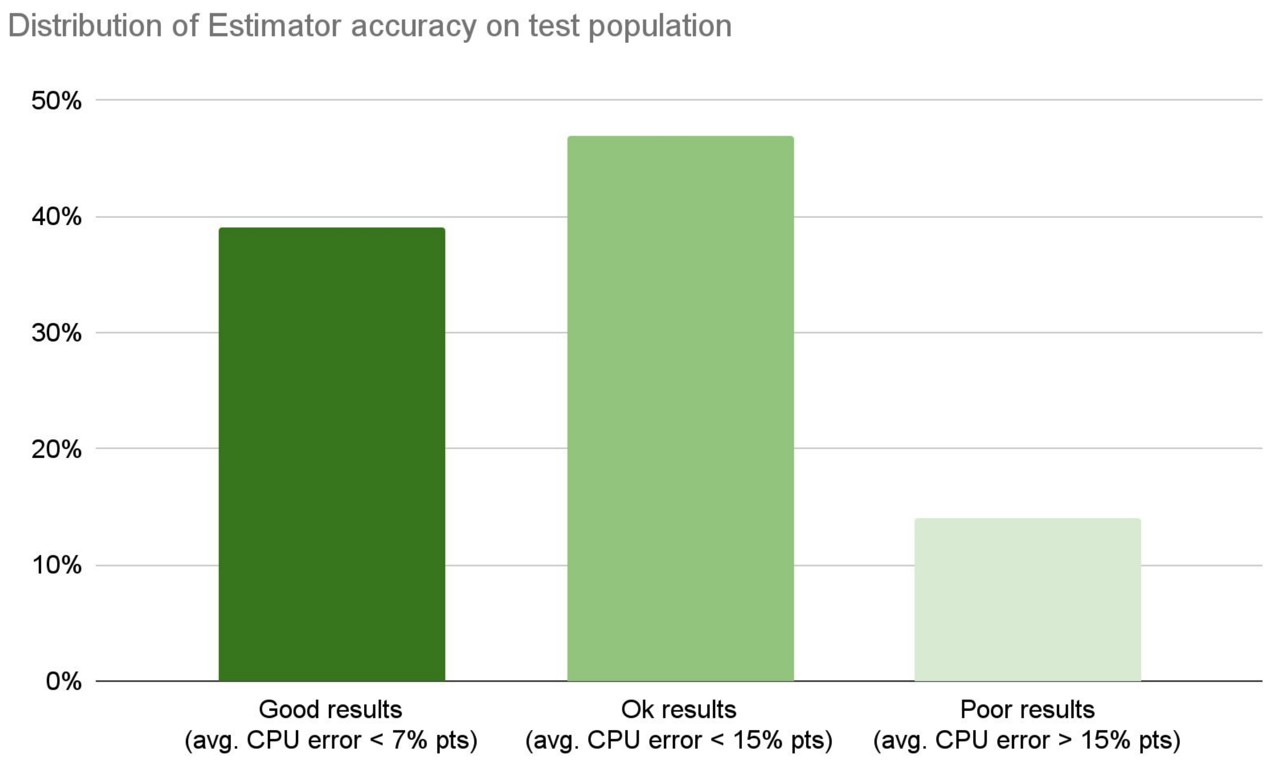
- Gradient Boosted Trees
- Trained on 25M records with 20 features

Predictive Scaling Experiment: Estimator

Example of Estimator result:



Predictive Scaling Experiment: Estimator



Predictive Scaling Experiment: Conclusion

Putting it all together

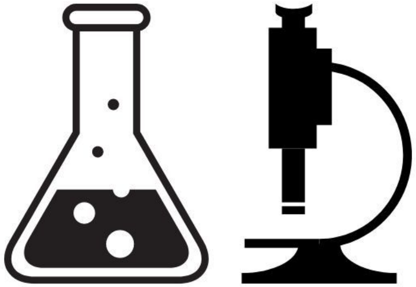
	Predictive auto-scaler	Reactive auto-scaler
Avg. distance from 75% util. target	18.6%	32.1%
Avg. under-utilization	18.3%	28.3%
Avg. over-utilization	0.4%	3.8%

Avg. Estimated $ cost savings: $0.09 per cluster/h

Future work

On addressing the Estimator’s shortcomings:

- Additional data about hardware spec
- Model memory with CPU
- Add query pattern data



Run live simulations to validate accuracy results

Product Release

Goal is to integrate with current reactive scaler

Unknown release date for now. Beta release to come…

Further Info

- This work inspired by Rebecca Taft's PhD thesis:
  emptysqua.re/blog/e-store-and-p-store

- Also interesting:
  "Is Machine Learning Necessary for Cloud Resource Usage Forecasting?"
  ACM Symposium on Cloud Computing 2023

- MongoDB Atlas: mongodb.com/atlas

- Jesse's Twitter: @jessejiryudavis
  (Matthieu abstains from Twitter)

Join us for Q & A in Room 115