Delayed-Dynamic-Selective (DDS) Prediction for Reducing Extreme Tail Latency in Web Search

Saehoon Kim[1], Yuxiong He[2], Seung-wong Hwang[1], Sameh Elnikety[2], Seungjin Choi[1]

[1] Department of Computer Science and Engineering, POSTECH
[2] Microsoft Research Redmond

POSTECH Pohang University of Science and Technology
Microsoft Research

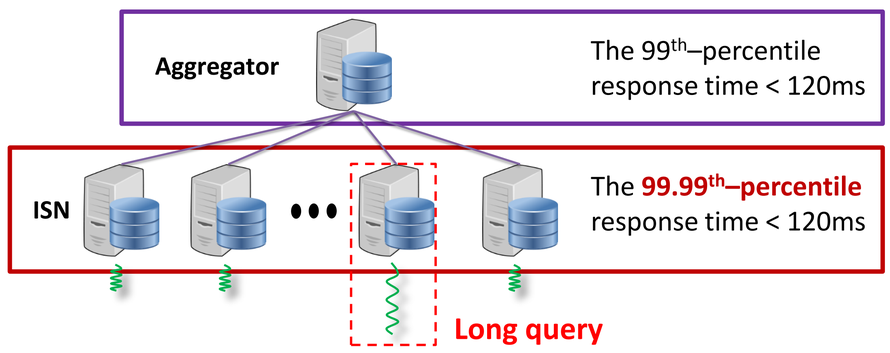
Motivation

- Reduce tail latency (high-percentile latency) of user queries, e.g., 99th percentile
- Reduce **extreme tail latency** at each index server, e.g., 99.99th percentile

Contribution

- **Delayed-Dynamic-Selective (DDS)** prediction: identify long(-running) queries with high accuracy
- **DDS Parallelization: use DDS to parallelize index servers for reducing extreme tail latency**

Why Extreme Tail Latency?



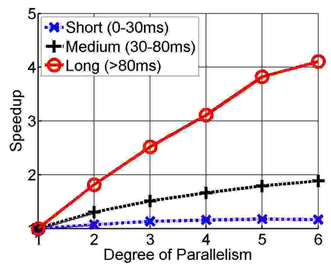
Reducing Tail Latency by Parallelization

Opportunity

Breakdown	Latency
Network	4.26 ms
Queueing	0.15 ms
I/O	4.70 ms
CPU	194.95 ms

1. Available idle cores
2. CPU-intensive workloads

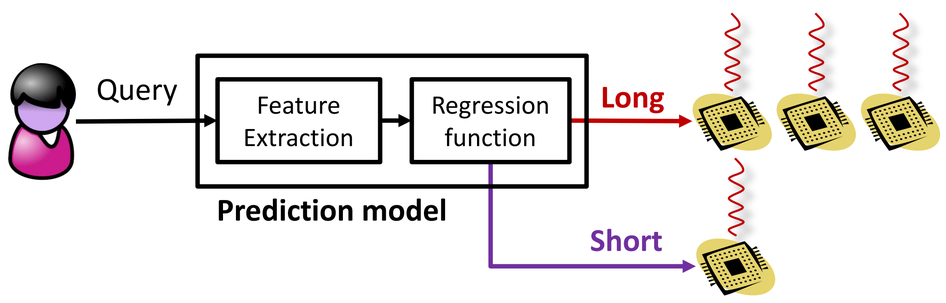
Challenges



1. Parallelizing all queries (inefficient)
2. Parallelizing short queries (no speed up)

PREDictive Parallelization [SIGIR'14]

Parallelize the predicted long queries only



Requirements

1. 99th tail latency at aggregator <= 120ms
2. Reduce 99.99th tail latency at each ISN <= 120ms

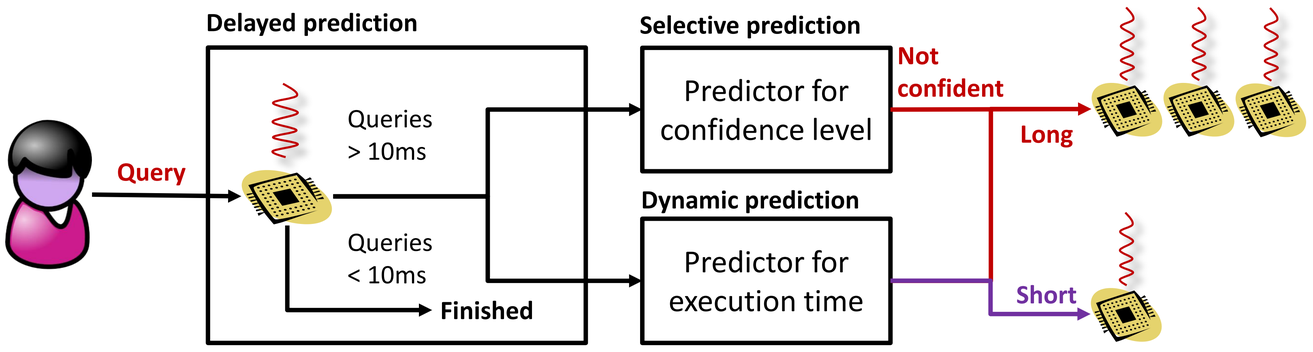
	Recall	Precision
Requirements	>= 98.9%	Should be high
Reason	To optimize 99.99th tail latency	Less queries to be parallelized

Limitation of PRED

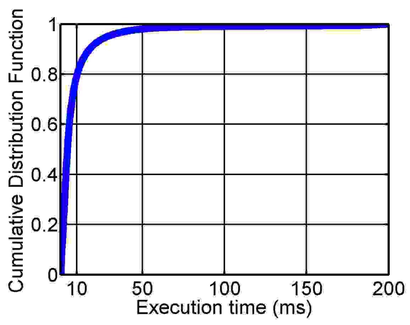
PRED cannot effectively reduce 99.99th tail latency

θ	Recall	Precision
100ms	0.601	0.789
20ms	0.905	0.098
10ms	0.952	0.037
2.3ms	0.989	0.011

DDS (Delayed-Dynamic-Selective) Prediction



Delayed prediction

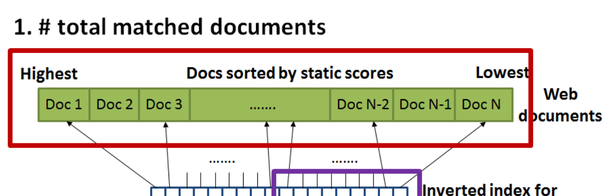


- Complete many short queries sequentially
- Collect **dynamic features**

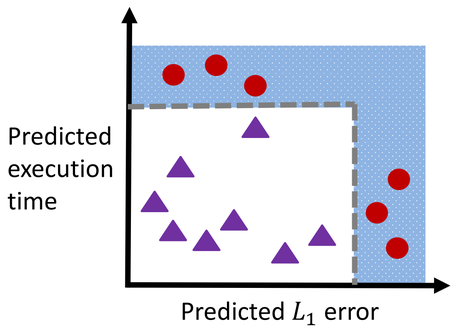
Dynamic features

Collected at query **runtime**

1. NumEstMatchDoc := $\frac{\#\ current\ matched\ docs}{\#\ processed\ docs}$
2. Statistics of the dynamic score distribution



Selective prediction



- Parallelize the unpredictable queries
- Parallel query if
  - ✓ Predicted execution time > α
  - ✓ Predicted L_1 error > β

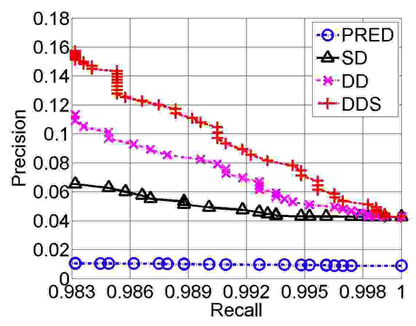
DDS Parallelization

Importance of dynamic features

Feature	Importance
NumEstMatchDoc	1
MinDynScore	0.7075
MinIDF	0.2767
VarIDF	0.2730
MaxDynScore	0.2662

- Top-5 feature importance by boosted regression tree
- NumEstMachDoc helps **to predict # total matched doc**
- DynScore helps **to predict early termination**

Predictor accuracy



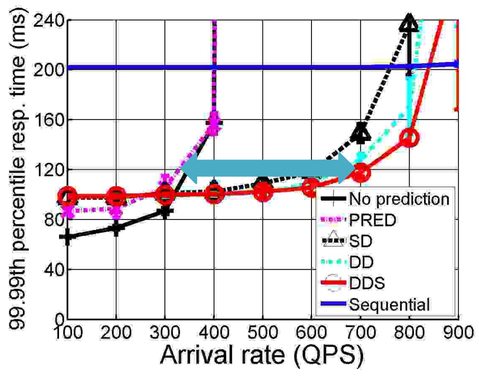
- Baseline: PRED
- 957% precision improvement at 98.9% recall over PRED

Simulation results on tail latency reduction

Baseline *S*
✓Prediction before running a query
✓Parallelize the long query

Proposed DDS
✓Run a query for **10ms sequentially**
✓Parallelizes the predicted **long or unpredictable queries**

Response time at index server



Response time at aggregator