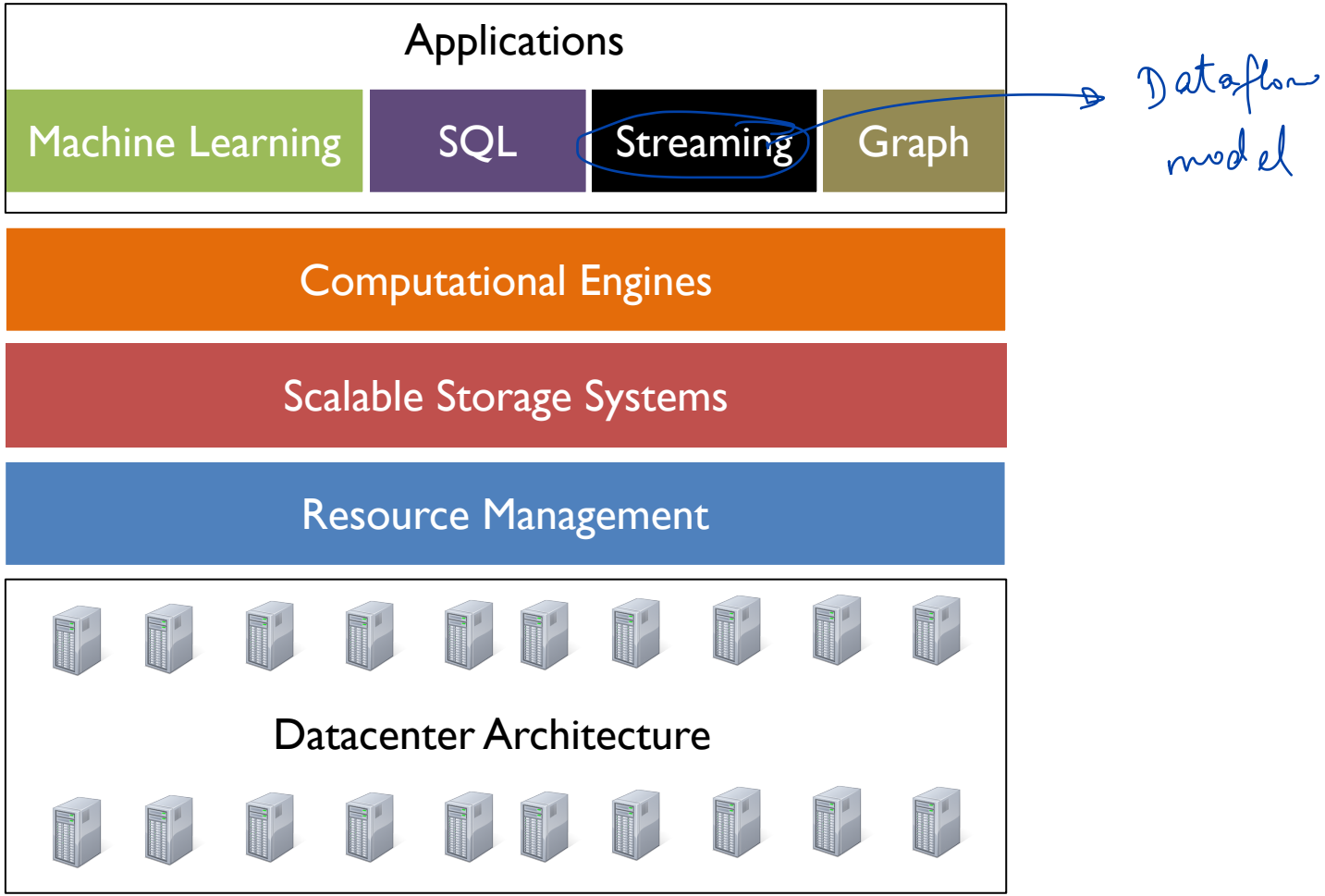
CS 744: NAIAD

Shivaram Venkataraman

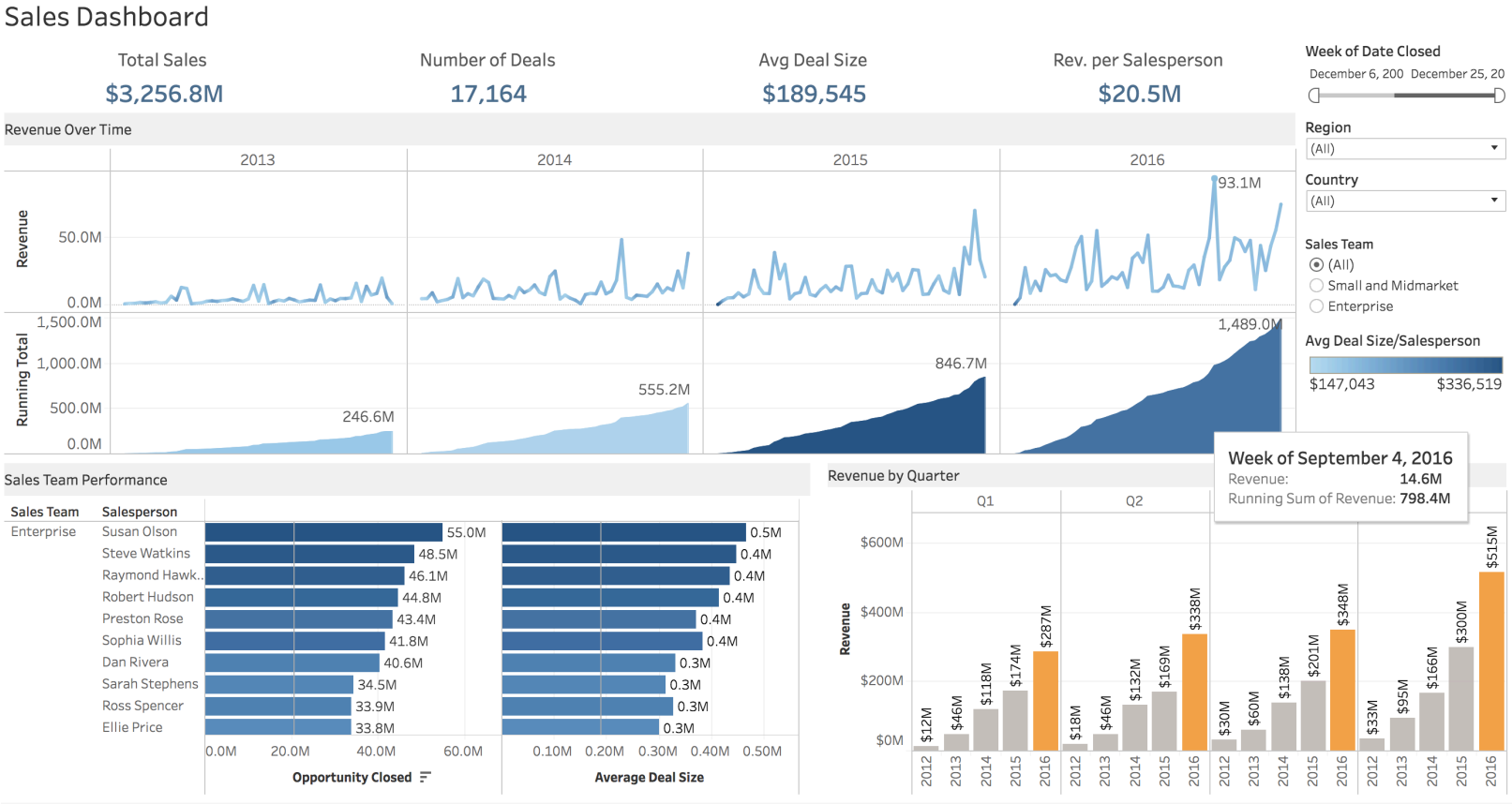
Fall 2019

ADMINISTRIVIA

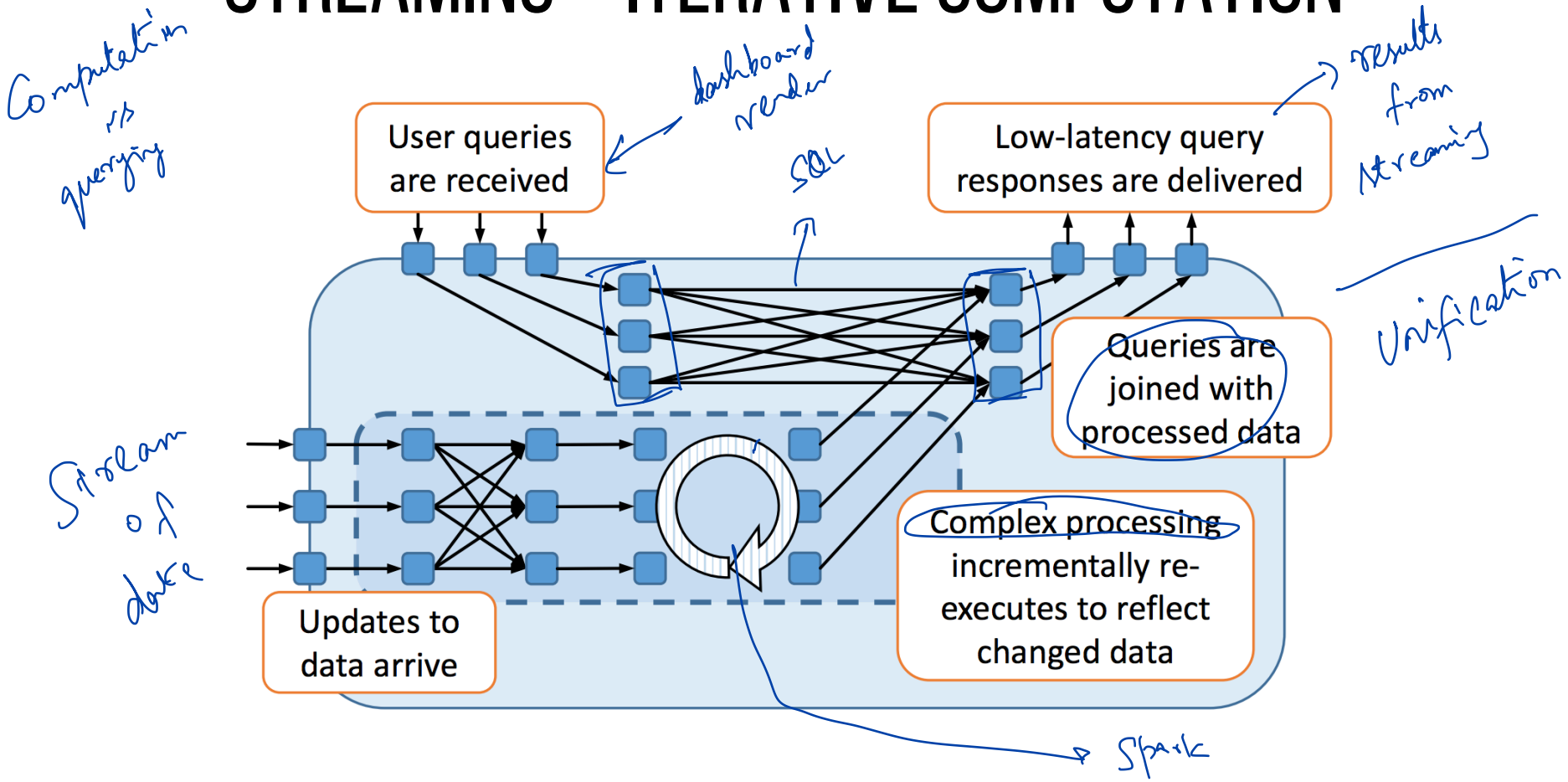
- Course Project Proposal feedback
- Midterm grades
- Checkins?


Applications
Machine Learning
SQL
Streaming
Graph
Dataflow model
Computational Engines
Scalable Storage Systems
Resource Management
Datacenter Architecture

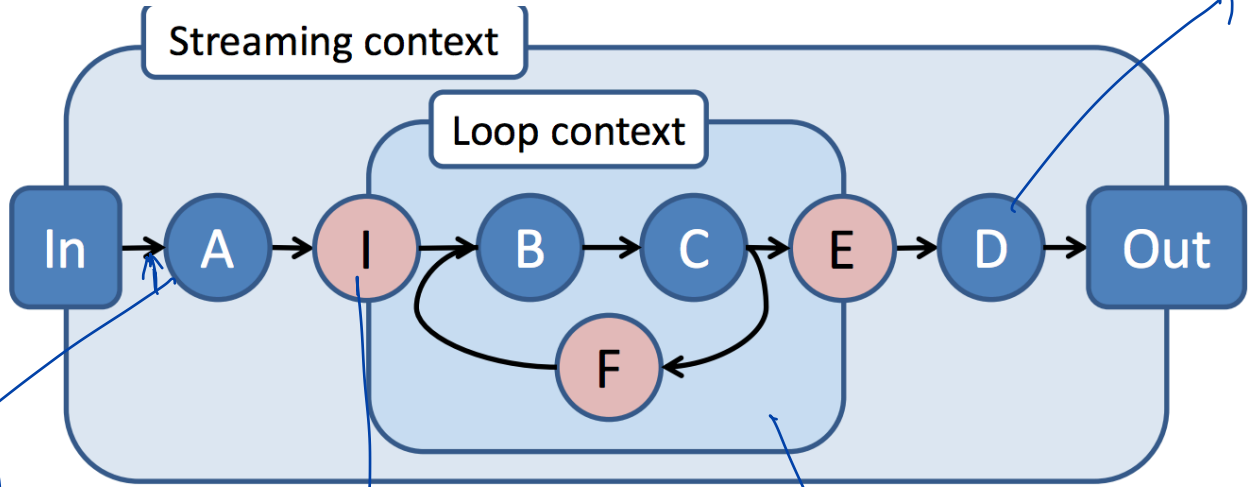
DASHBOARDS


Sales Dashboard
Total Sales
$3,256.8M
Number of Deals
17,164
Avg Deal Size
$189,545
Rev. per Salesperson
$20.5M
Week of Date Closed
December 6, 200 December 25, 20
Region
(All)
Country
(All)
Sales Team
(All)
Small and Midmarket
Enterprise
Avg Deal Size/Salesperson
$147,043
$336,519
Revenue Over Time
2013
2014
2015
2016
Revenue
50.0M
0.0M
93.1M
Running Total
1,500.0M
1,000.0M
500.0M
0.0M
246.6M
555.2M
846.7M
1,489.0M
Week of September 4, 2016
Revenue: 14.6M
Running Sum of Revenue: 798.4M
Sales Team Performance
Sales Team
Salesperson
Enterprise
Susan Olson 55.0M 0.5M
Steve Watkins 48.5M 0.4M
Raymond Hawk.. 46.1M 0.4M
Robert Hudson 44.8M 0.4M
Preston Rose 43.4M 0.4M
Sophia Willis 41.8M 0.4M
Dan Rivera 40.6M 0.3M
Sarah Stephens 34.5M 0.3M
Ross Spencer 33.9M 0.3M
Ellie Price 33.8M 0.3M
0.0M 20.0M 40.0M 60.0M
Opportunity Closed
0.10M 0.20M 0.30M 0.40M 0.50M
Average Deal Size
Revenue by Quarter
Q1
Q2
Revenue
$600M
$400M
$200M
$0M
$12M $46M $118M $174M $287M
$18M $46M $132M $169M $338M
$30M $60M $138M $201M $348M
$33M $95M $166M $300M $515M
2012 2013 2014 2015 2016

STREAMING + ITERATIVE COMPUTATION



TIMELY DATAFLOW

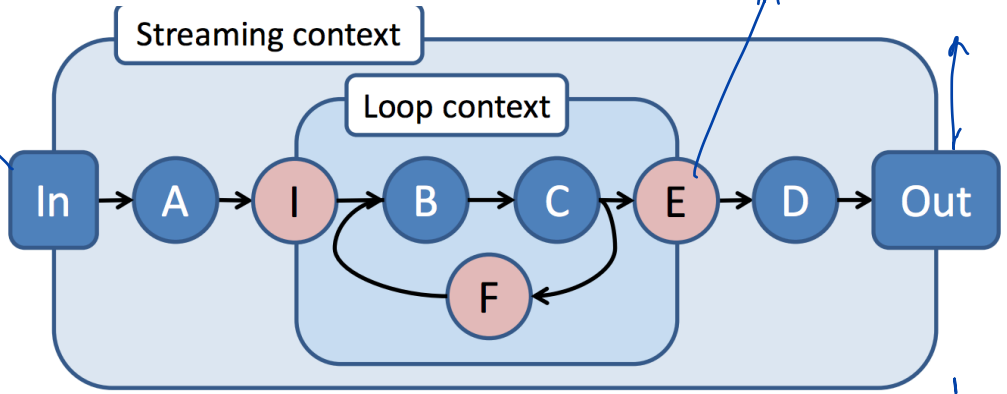


TIMELY DATAFLOW

for (i)
for (j)
for (p)
→ k : 3

epoch start

loop is over



Timestamp : $(\underbrace{e \in \mathbb{N}}_{\text{epoch}}, \underbrace{\langle c_1, \ldots, c_k \rangle \in \mathbb{N}^k}_{\text{loop counters}})$

1

(0, <1, 1>) t_1

(0, <2>) t_2 → t_1

loop restarts

Vertex	Input timestamp	Output timestamp
Ingress	$(e, \langle c_1, \ldots, c_k \rangle)$	$(e, \langle c_1, \ldots, c_k, 0 \rangle)$
Egress	$(e, \langle c_1, \ldots, c_k, c_{k+1} \rangle)$	$(e, \langle c_1, \ldots c_k \rangle)$
Feedback	$(e, \langle c_1, \ldots, c_k \rangle)$	$(e, \langle c_1, \ldots c_k + 1 \rangle)$

New k+1 loop 0 iter

→ Inc loop counter

VERTEX API

Receiving Messages

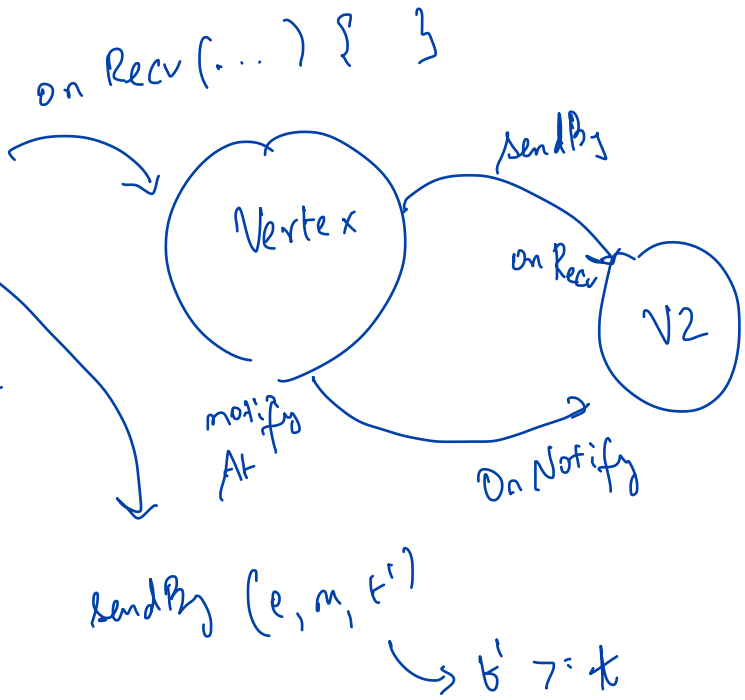
v.OnRecv(e : Edge, m : Msg, t : Time)

v.OnNotify(t : Timestamp)

Sending Messages

this.SendBy(e : Edge, m : Msg, t : Time)

this.NotifyAt(t : Timestamp)



"onRecv ∀ t' ≤ t

Not specifying any vertex

IMPLEMENTING TIMELY DATAFLOW

Need to track when it is safe to notify

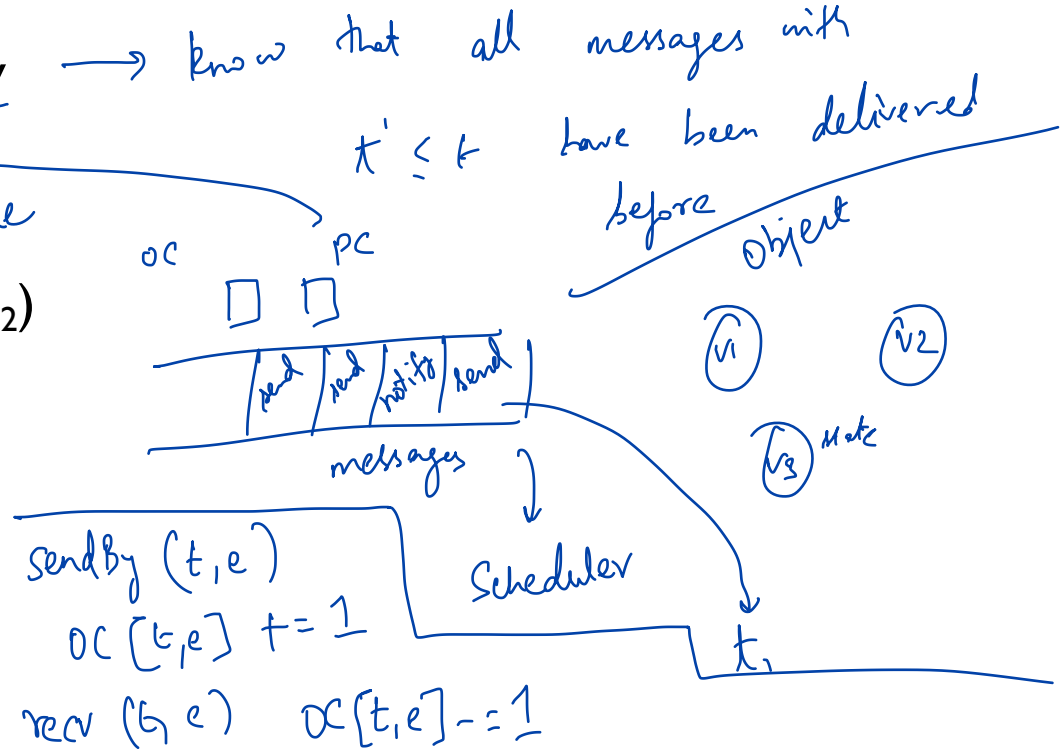
Path Summary

Check if (t_1,l_1) could-result-in (t_2,l_2)

Scheduler

Occurrence and Precursor count

Precursor count = 0 → Frontier



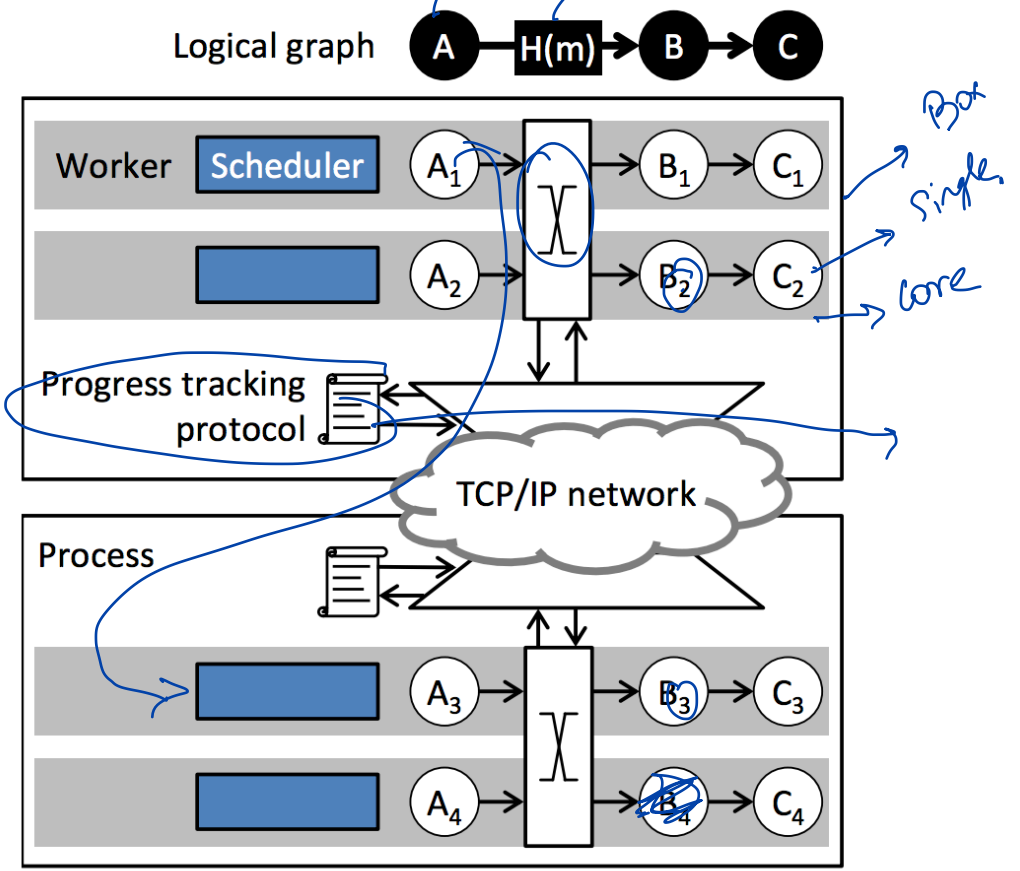
ARCHITECHTURE

Workers communicate using Shared Queue

Batch messages delivered

Account for cycles

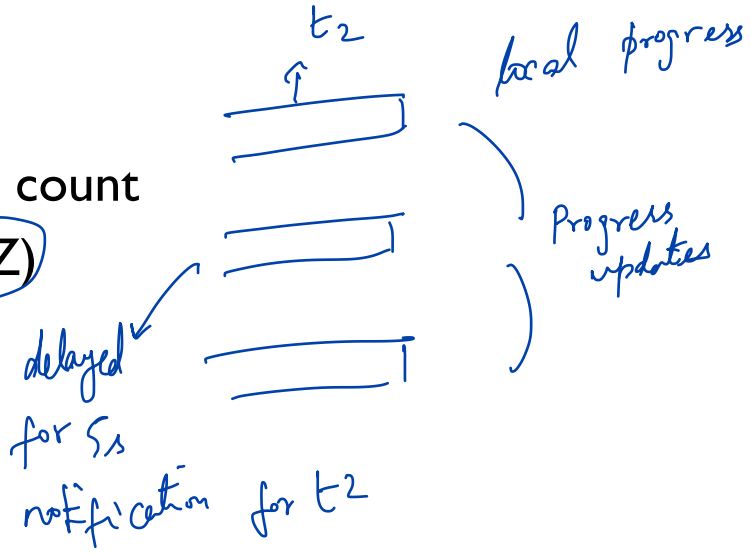
Vertex single threaded



DISTRIBUTED PROGRESS TRACKING

Broadcast-based approach

- Maintain local precursor count, occurrence count
- Send progress update (p ∈ Pointstamp, δ∈ Z)
- Local frontier tracks global frontier



Optimizations

- Batch updates and broadcast
- Use projected timestamps from logical graph

vertices
Stateful

FAULT TOLERANCE

Checkpoint

- Log data as computation goes on
- Write a full checkpoint on demand

Pause worker threads

Flush message queues OnRecv

Recovery time could be large

failures are rare?

Restore

- Reset all workers to checkpoint
- Reconstruct state

Resume execution

Driver → Centralized point of contention

→ Simplifies

MICRO STRAGGLERS

What is different from stragglers in MapReduce?

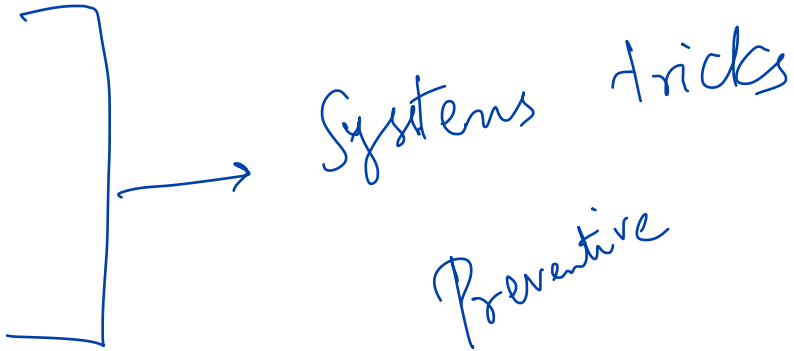
Stateful

Sources of stragglers

Network

Concurrency

Garbage Collection



DIFFERENTIAL DATAFLOW

```
// 1a. Define input stages for the dataflow.
var input = controller.NewInput<string>();
// 1b. Define the timely dataflow graph.
// Here, we use LINQ to implement MapReduce.
var result = input.SelectMany(y => map(y))
                  .GroupBy(y => key(y),
                    (k, vs) => reduce(k, vs));
// 1c. Define output callbacks for each epoch
result.Subscribe(result => { ... });
// 2. Supply input data to the query.
input.OnNext(/* 1st epoch data */);
input.OnCompleted();
```

SUMMARY

Stream processing → Increasingly important workload trend

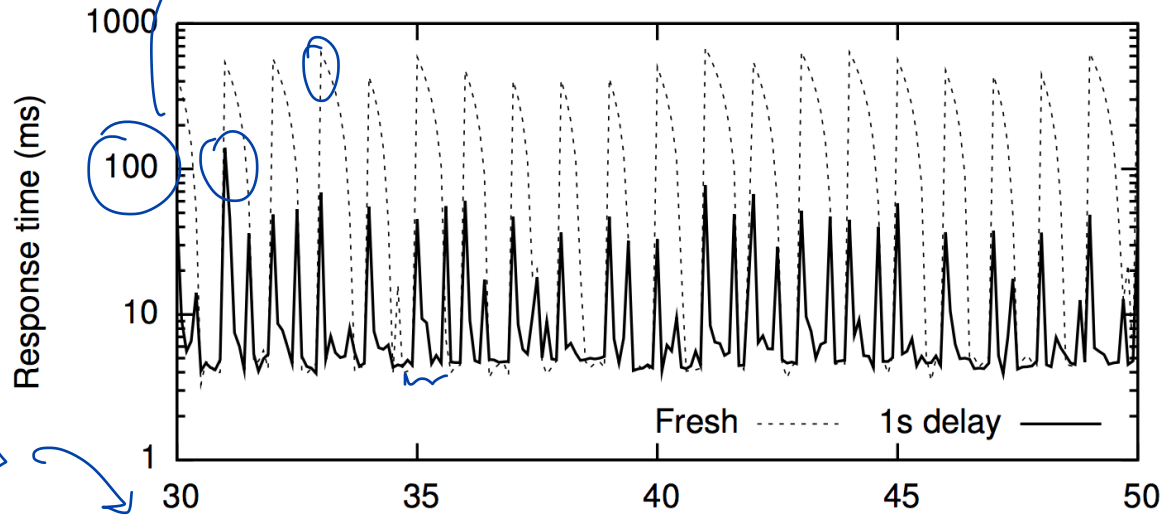
Timely dataflow: Principled approach to model batch, streaming together

Vertex message model

- Compute frontier
- Distributed progress tracking

DISCUSSION

https://forms.gle/v3YsW1HvnqsxCuPu5


1000
100
10
1
Response time (ms)
30
35
40
45
50
Time from start of trace (s)
Fresh
1s delay

What are some example scenarios discussed in the dataflow paper that are NOT a good fit for implementation using Naiad?

Stale updates → watermark

Triggering →

Good but not perfect is fine!

Consider you are implementing a micro-batch streaming API on top of Apache Spark. What are some of the bottlenecks/challenges you might have in building such a system?