The Power of Choice in Data-Aware Cluster Scheduling

Shivaram Venkataaraman\textsuperscript{1}, Aurojit Panda\textsuperscript{1}
Ganesh Ananthanarayanan\textsuperscript{2}, Michael Franklin\textsuperscript{1}, Ion Stoica\textsuperscript{1}

\textsuperscript{1}UC Berkeley, \textsuperscript{2}Microsoft Research
Trends: Big Data

Data grows faster than Moore’s Law

[Kathy Yelick LBNL, VLDBJ 2012, Dhruba Borthakur]
Trends: Big Data

Facebook Hive cluster
Last 4 years:
data growth 2500x
queries/day 60x

Microsoft Scope Cluster
“The number of daily jobs has doubled every six months for the past two years.”

[Kathy Yelick LBNL, VLDBJ 2012, Dhruba Borthakur]
Trends: Low Latency

- 2004: MapReduce batch job (10 min)
- 2009: Hive (1 min)
- 2010: Dremel (10s)
- 2012: In-memory Spark (2s)
Big Data or Low Latency?

SQL Query: 2.5 TB on 100 machines

> 15 minutes
1 - 5 Minutes
< 10s
Sampling
Applications

Approximate Query Processing
  blinkdb, presto, minitable

Machine learning algorithms
  stochastic gradient, coordinate descent
Choices

N

Any K
Sampling \( \rightarrow \) Smaller Inputs + Choice
Example

$N = 4$

$K = 2$

$M$

$M$

$R$
Available (N) = 2

Required (K) = 2

Available Data

Running

Unavailable Data

Busy

Rack

Time

Existing
Available (N) = 4

Required (K) = 2

Choice-Aware
Available (N) = 4

Launched (M) = 3

Required (K) = 2

Choice-Aware

Available Data

Running

Busy

Rack

Time
KMN Scheduler

- How much can KMN improve locality
- Propagate benefits across stages
- Handling stragglers
Job $\rightarrow$ DAG

KMN Scheduler

One-to-One

Many-to-One
One-to-One Stages

Locality

Disk ~ 100MB/s
Network ~ 10 Gbps (~1GB/s)
Memory ~ 50GB/s
KMN Locality

\[ \binom{N}{K} \] Choices

Any K

N

1 2 3 4

1 2 2 3 3 4

1 3 2 4 1 4
Locality, $K=100$

K – Number of blocks chosen
N – Number of blocks available

KMN significantly improves locality
Many-to-One Stages

KMN Scheduler
Many-to-One Stage

15 transfers
Many-To-One Transfers
Bottleneck Link

Link with Max. transfers

Cross Rack Data Skew

Maximum transfers

Minimum transfers

\[ \frac{6}{2} = 3 \]
Facebook Trace

Cross Rack Data Skew

Maximum transfers

Minimum transfers

CDF

Cross Rack Data Skew

<50 tasks
50-150 tasks
>150 tasks
Power of Choice

Load balancing: balls and bins

Insight: Run extra tasks ($M > K$)

Cross Rack Data Skew = 3
Power of Choice

Technique:
Spread out choice of K tasks to reduce skew

M = 7, K = 5
Cross Rack Data Skew = 2
Handling Stragglers

M_1, M_2, M_3, M_4, M_5, M_6, M_7

Stragglers

Rack

Cross-Rack

Data Skew

Time
// Create Spark RDD
file = sc.textFile("tpc-h.data")

// Select a 10% sample using KMN
sample = file.blockSample(0.1)

// RDD operations
sample.map { li =>
    (li.linestatus, li.quantity)
}.collect()
Also in the paper

User-defined sampling functions
Placing reduce tasks
Killing extra tasks
Evaluation

Facebook traces replay
Long DAGs (Stochastic Gradient Descent)
SQL queries from Conviva
Reducer placement
Varying Utilization

Baseline: Use a pre-selected random sample
Setup: 100 m2.4xlarge EC2 machines, 60GB RAM/mc
Facebook Overall

Job Completion Time (s) vs. Job Size

Baseline vs. KMN-M/K=1.05

Job Size: 0-10, 11-100, >100

Job Completion Time (s): 0-50
Cross Rack Skew

Shuffle Stage Time (seconds)

Cross Rack Skew

>8

4-to-8

<=4

Baseline

KMN-M/K=1.0

KMN-M/K=1.05

KMN-M/K=1.1
How many extra tasks?

50 - 150 tasks

> 150 tasks
KMN: How many stages?

Stochastic Gradient Descent

Aggregate1

Aggregate2

Aggregate3

Gradient
KMN: How many stages?

<table>
<thead>
<tr>
<th>KMN Stages</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gradient</td>
<td>15.27</td>
</tr>
</tbody>
</table>
KMN: How many stages?

<table>
<thead>
<tr>
<th>KMN Stages</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gradient</td>
<td>15.27</td>
</tr>
<tr>
<td>Gradient + Agg1</td>
<td>12.72</td>
</tr>
</tbody>
</table>
KMN: How many stages?

<table>
<thead>
<tr>
<th>KMN Stages</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gradient</td>
<td>15.27</td>
</tr>
<tr>
<td>Gradient + Agg1</td>
<td>12.72</td>
</tr>
<tr>
<td>Gradient + Agg2</td>
<td>11.79</td>
</tr>
</tbody>
</table>

Diagram:
- Gradient
  - Aggregate1
  - Aggregate2
  - Aggregate3
KMN: How many stages?

---

**Diagram:**
- Gradient → Aggregate1
- Aggregate1 → Aggregate2
- Aggregate2 → Aggregate3

**Table:**

<table>
<thead>
<tr>
<th>KMN Stages</th>
<th>Time (s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gradient</td>
<td>15.27</td>
</tr>
<tr>
<td>Gradient + Agg1</td>
<td>12.72</td>
</tr>
<tr>
<td>Gradient + Agg2</td>
<td>11.79</td>
</tr>
<tr>
<td>Gradient + Agg3</td>
<td>12.09</td>
</tr>
</tbody>
</table>
Related Work

Power of Choice
- Power-of-Two choices [TPDS’01]
- Sparrow [SOSP’13]

Improving Cluster Scheduling
- Quincy [SOSP’09]
- alsched [SOCC’12]
- Dolly [NSDI’13]
KMN Scheduler

Emerging applications: ML algorithms, AQP
Improves locality, Balances network transfers