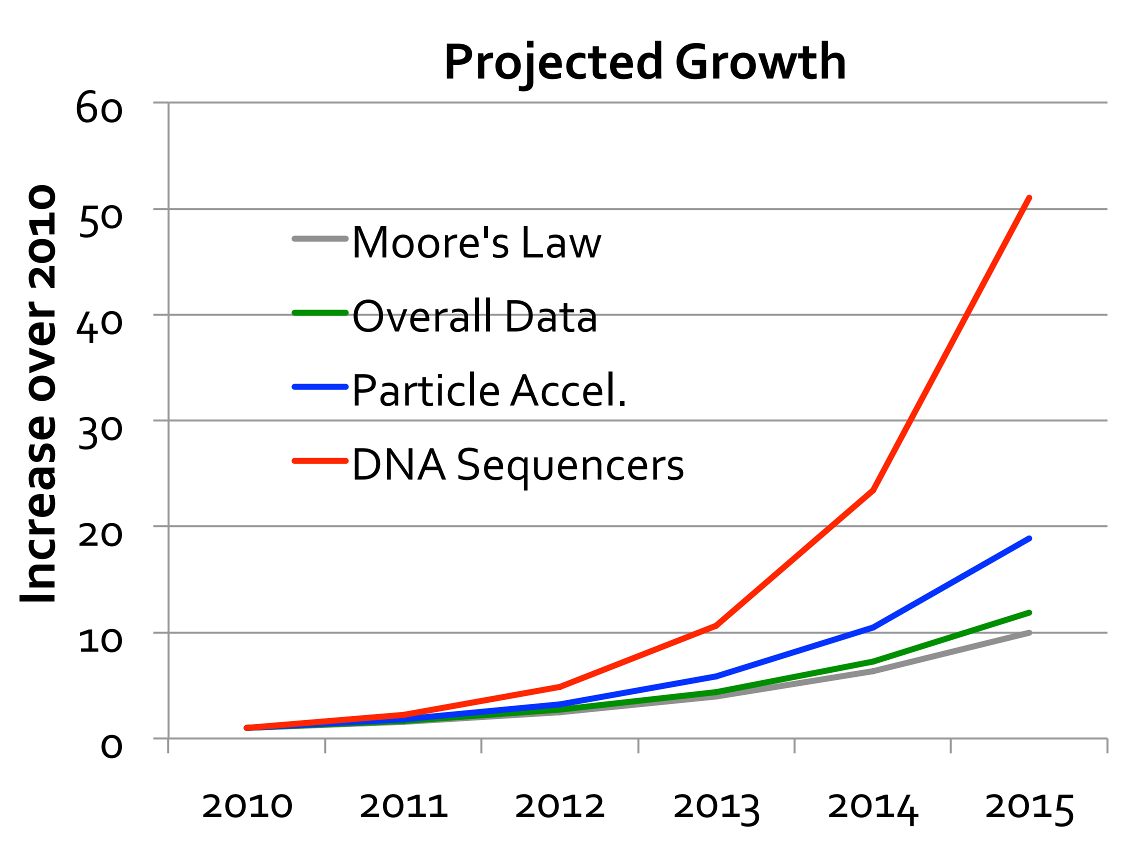
The Power of Choice in Data-Aware Cluster Scheduling

Shivaram Venkataraman[1] , Aurojit Panda[1]
Ganesh Ananthanarayanan[2], Michael Franklin[1], Ion Stoica[1]

[1]UC Berkeley, [2]Microsoft Research

amplab

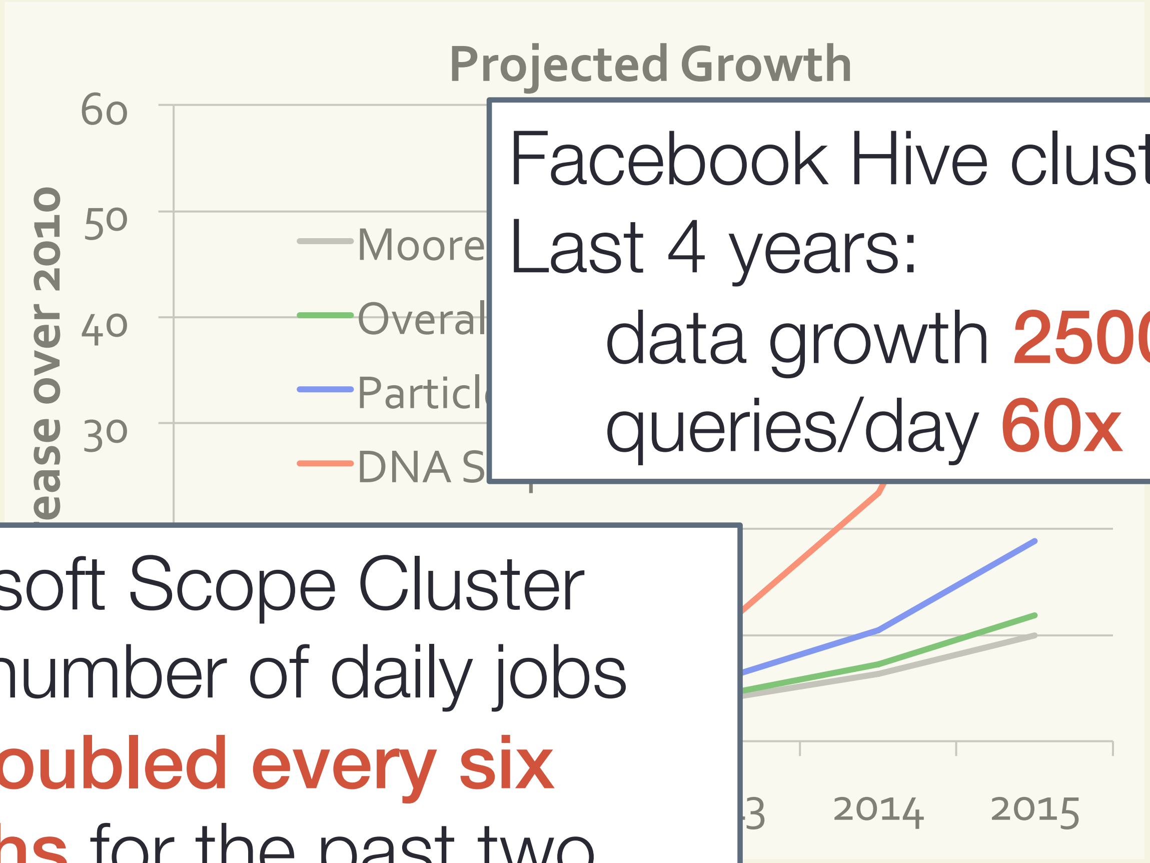
Trends: Big Data



Data grows faster than Moore’s Law

[Kathy Yelick LBNL, VLDBJ 2012, Dhruba Borthakur]

Trends: Big Data

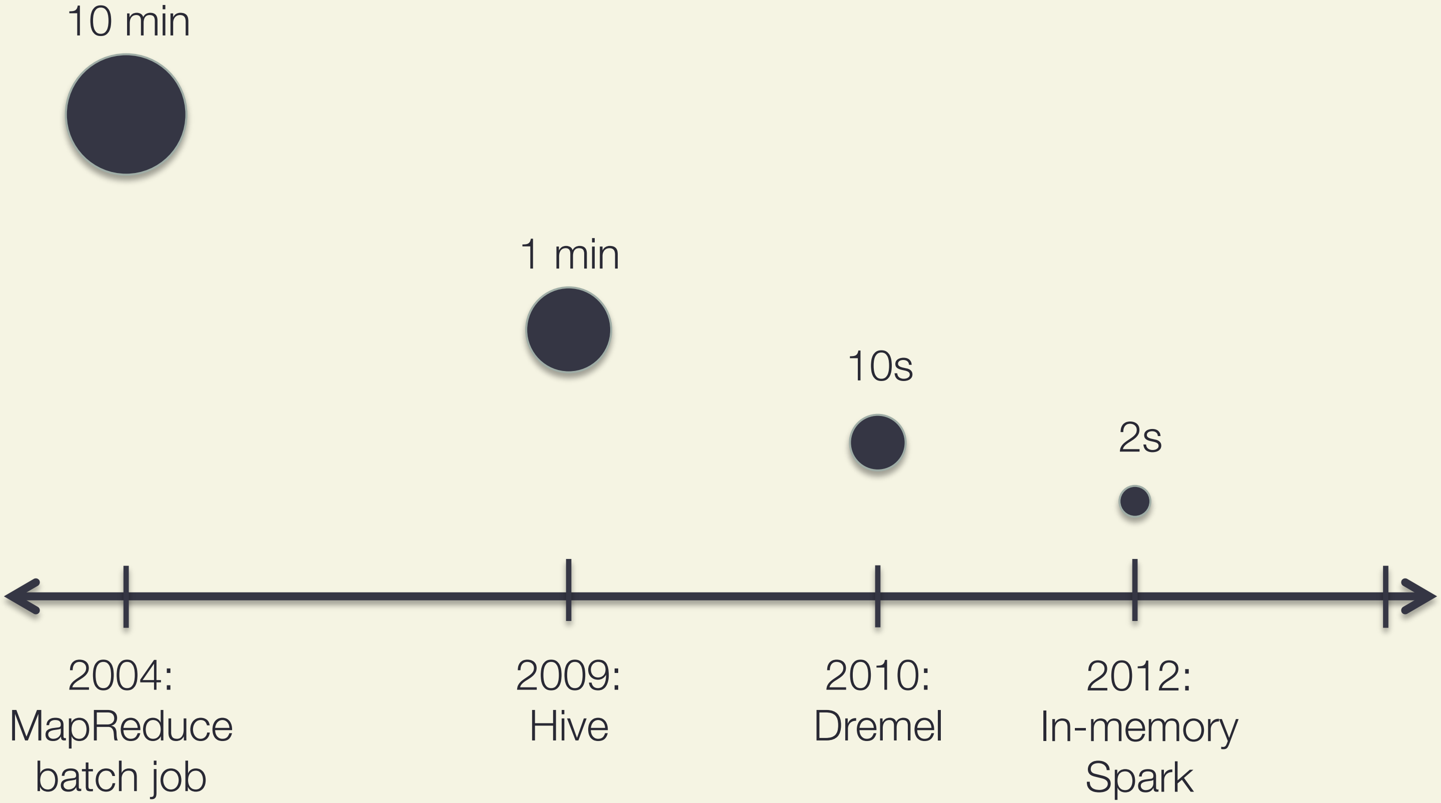


Facebook Hive cluster
Last 4 years:
data growth **2500x**
queries/day **60x**

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



Big Data or Low Latency ?

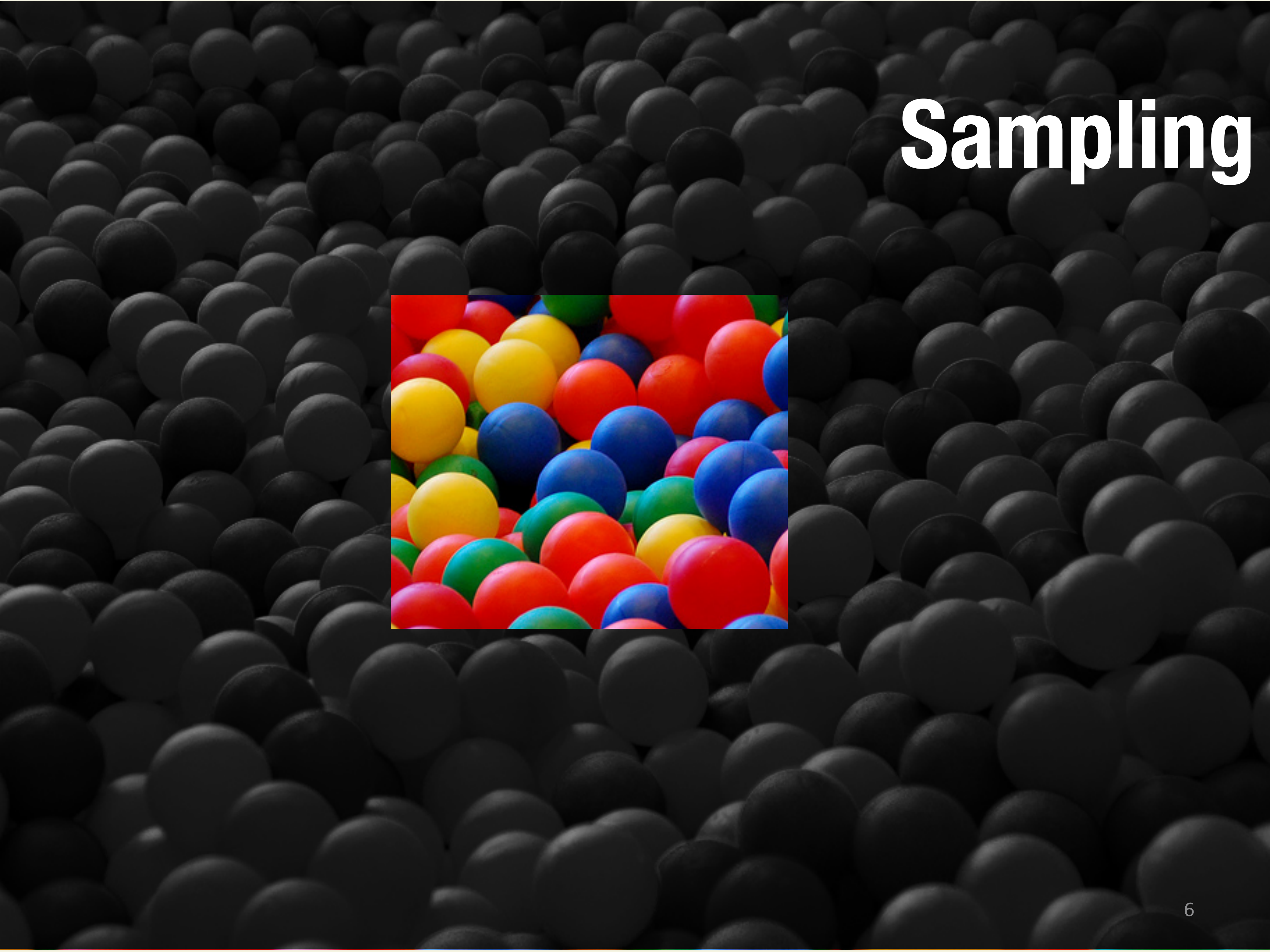
SQL Query : 2.5 TB on 100 machines

?

> 15 minutes

1 - 5 Minutes

< 10s



Sampling

Applications

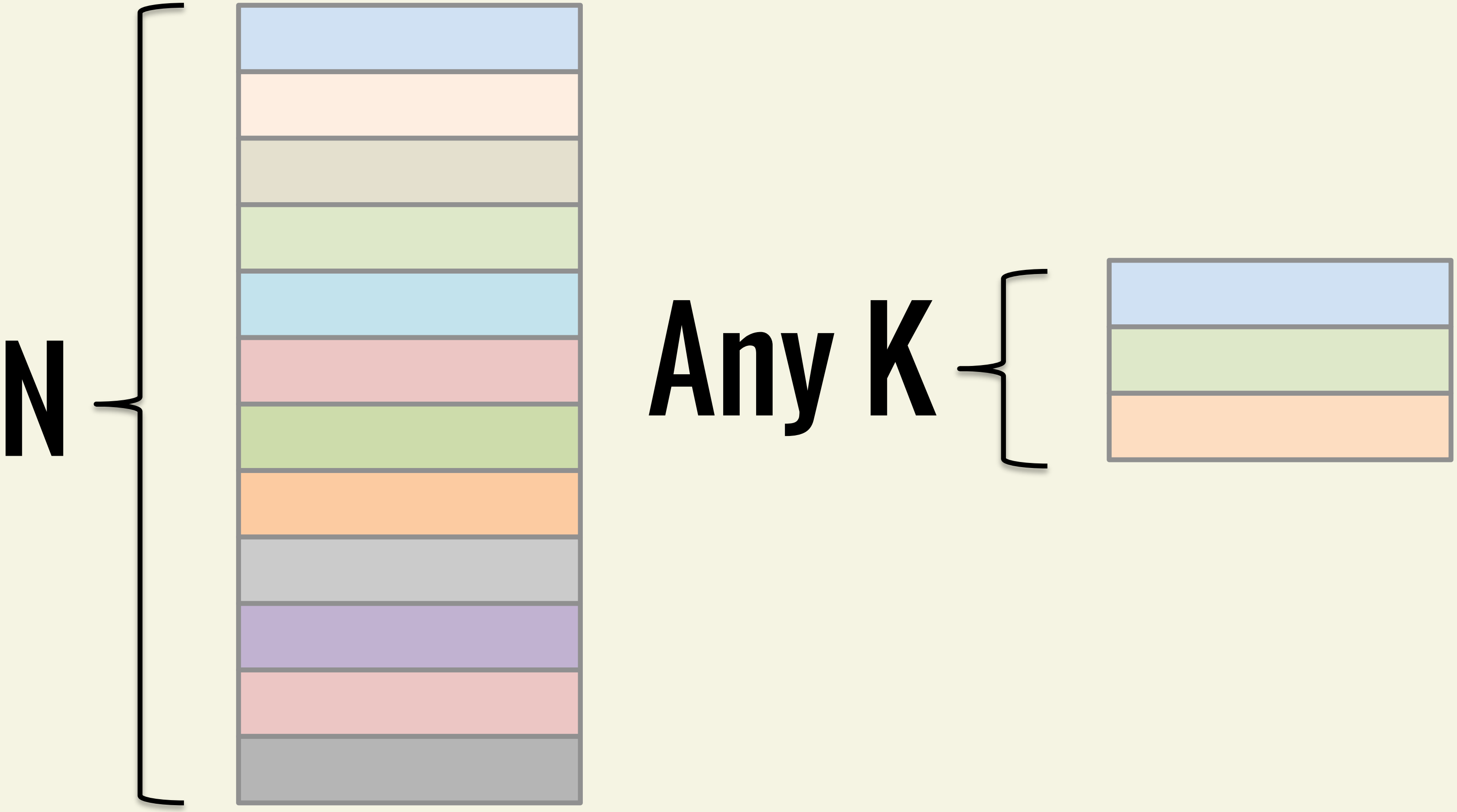
Approximate Query Processing

- blinkdb, presto, minitable

Machine learning algorithms

- stochastic gradient, coordinate descent

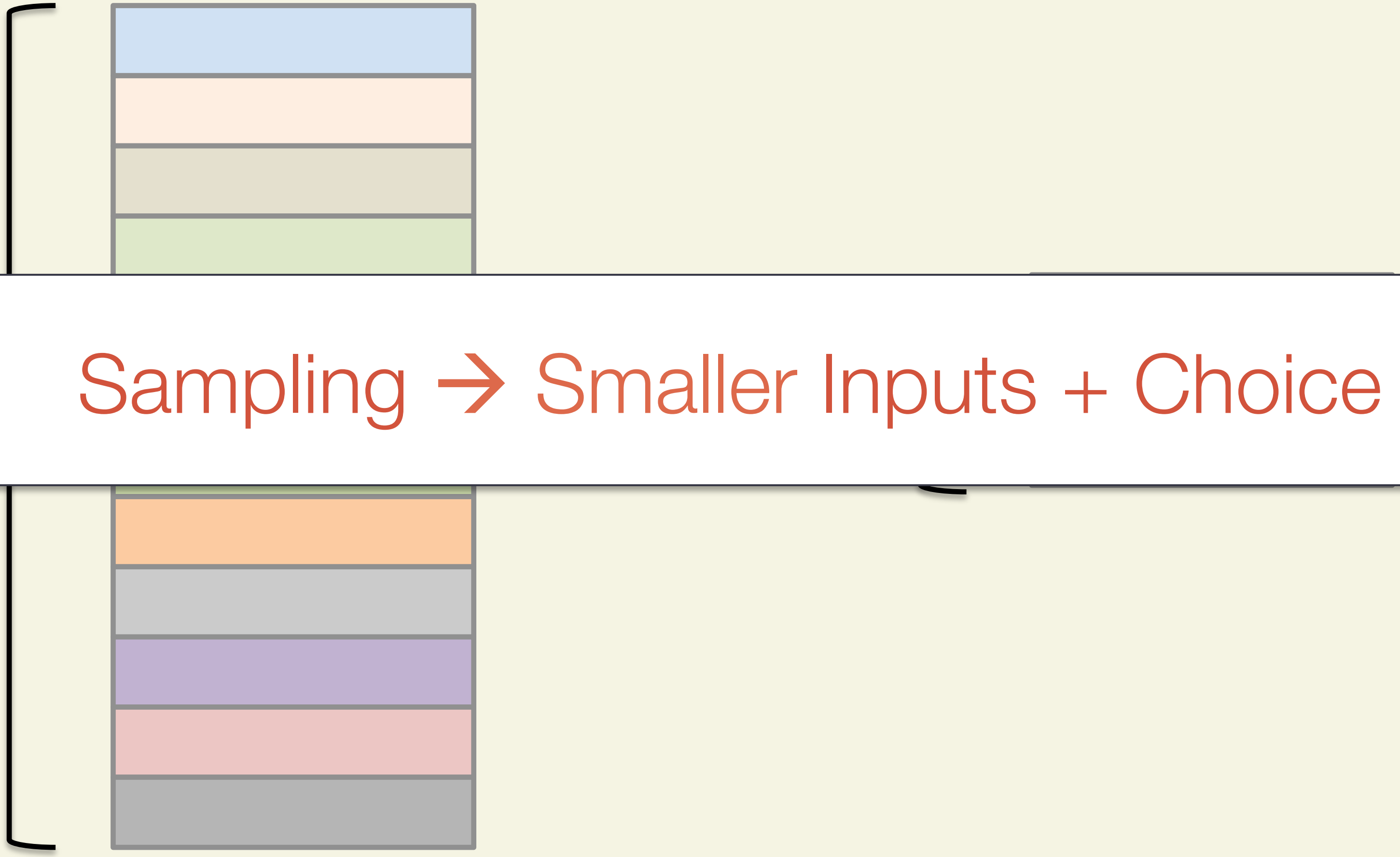
Choices



N

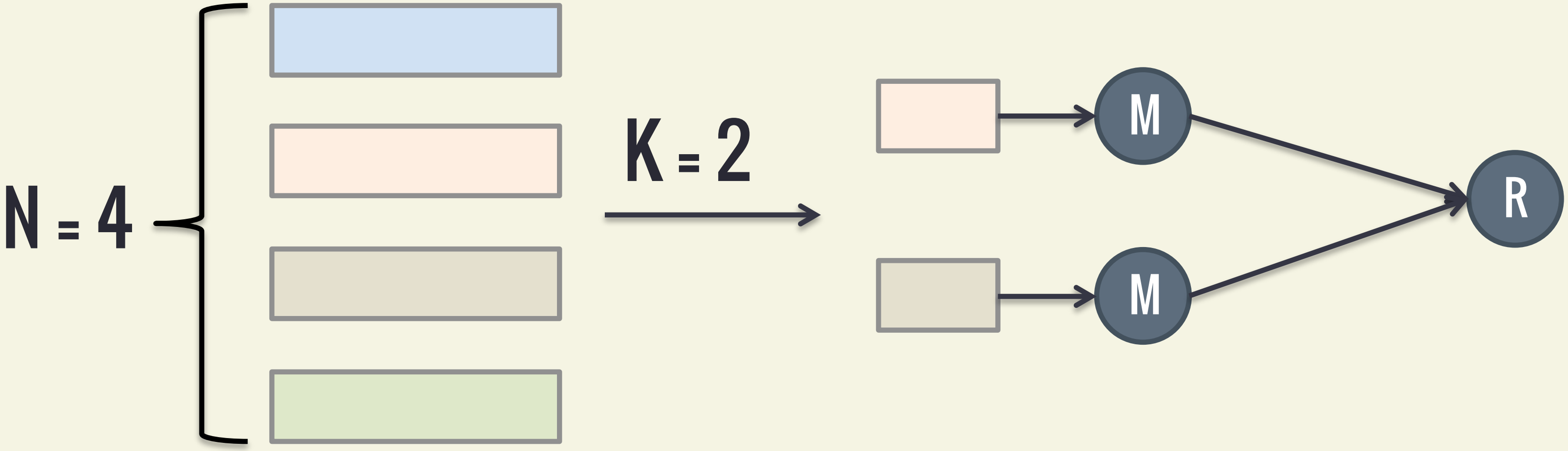
Any K

Choices

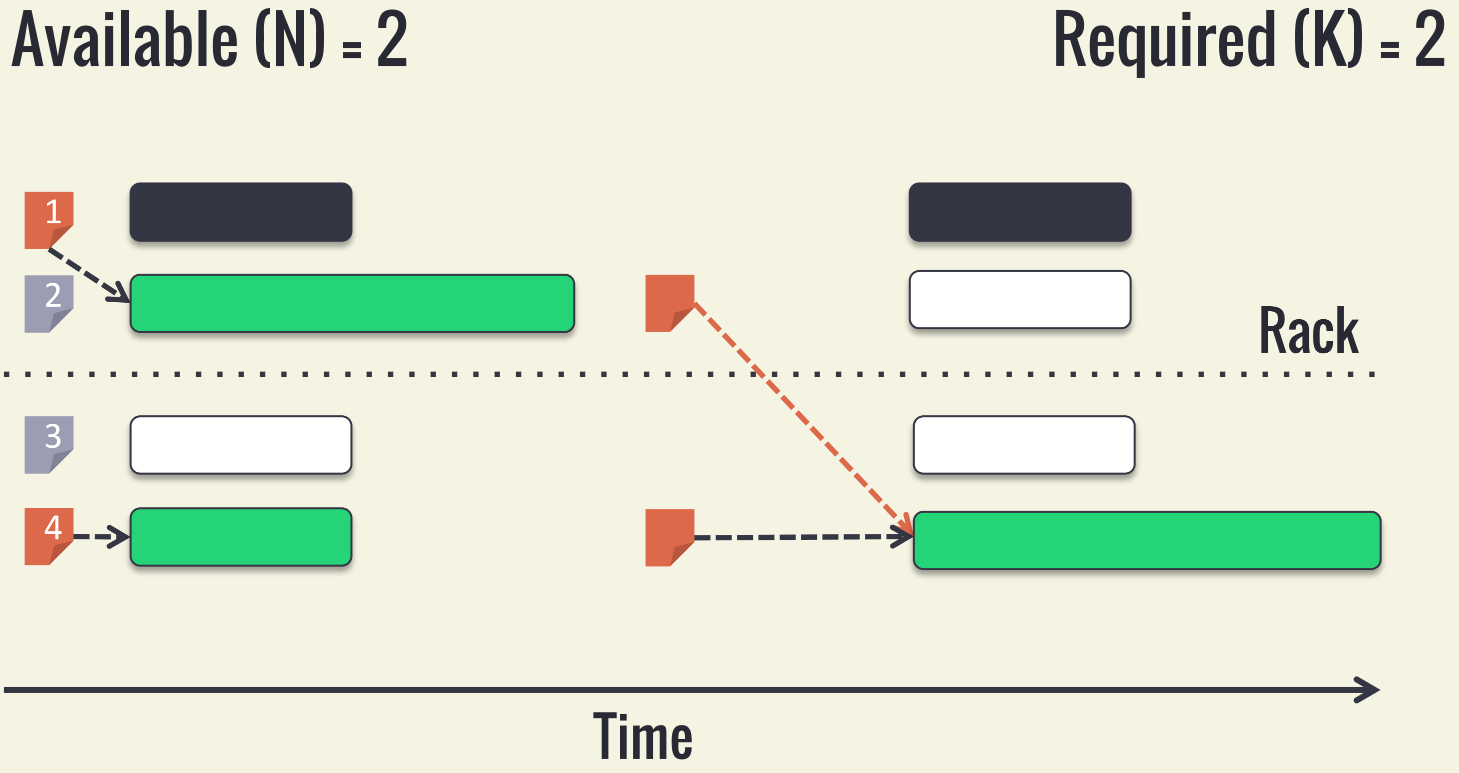


Sampling → Smaller Inputs + Choice

Example



Existing



🟥 Available Data

⬜ Unavailable Data

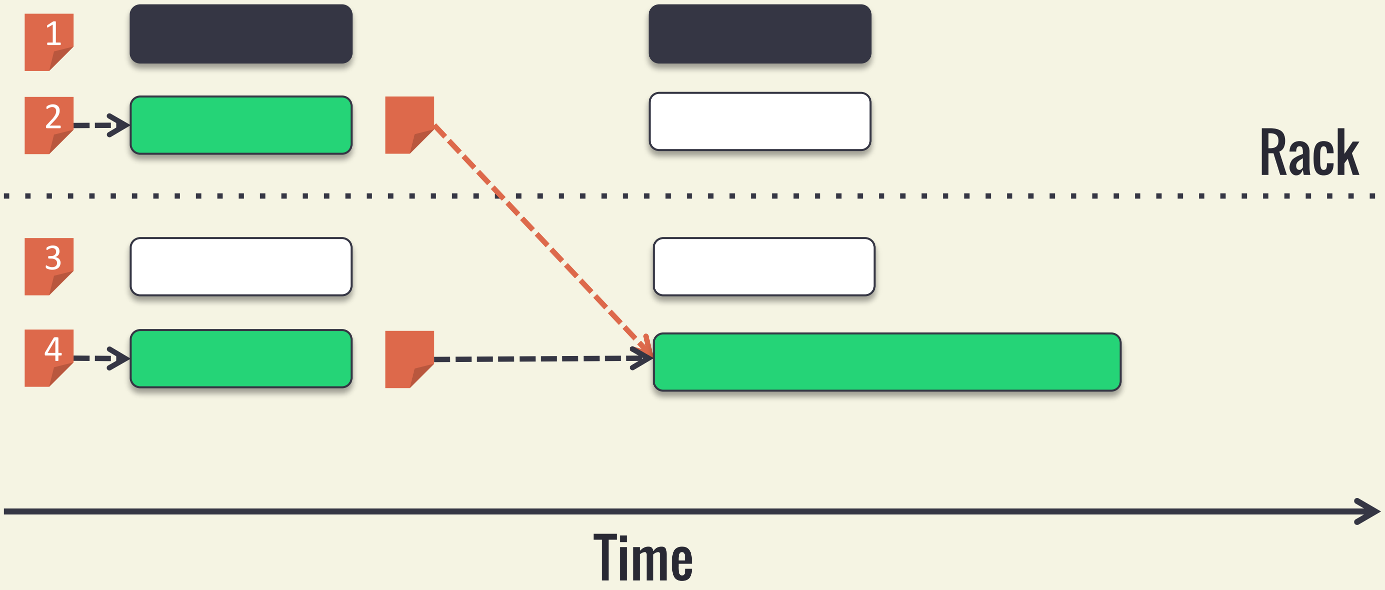
🟩 Running

⬛ Busy

Choice-Aware

Available (N) = 4

Required (K) = 2



1

2

3

4

Rack

Time

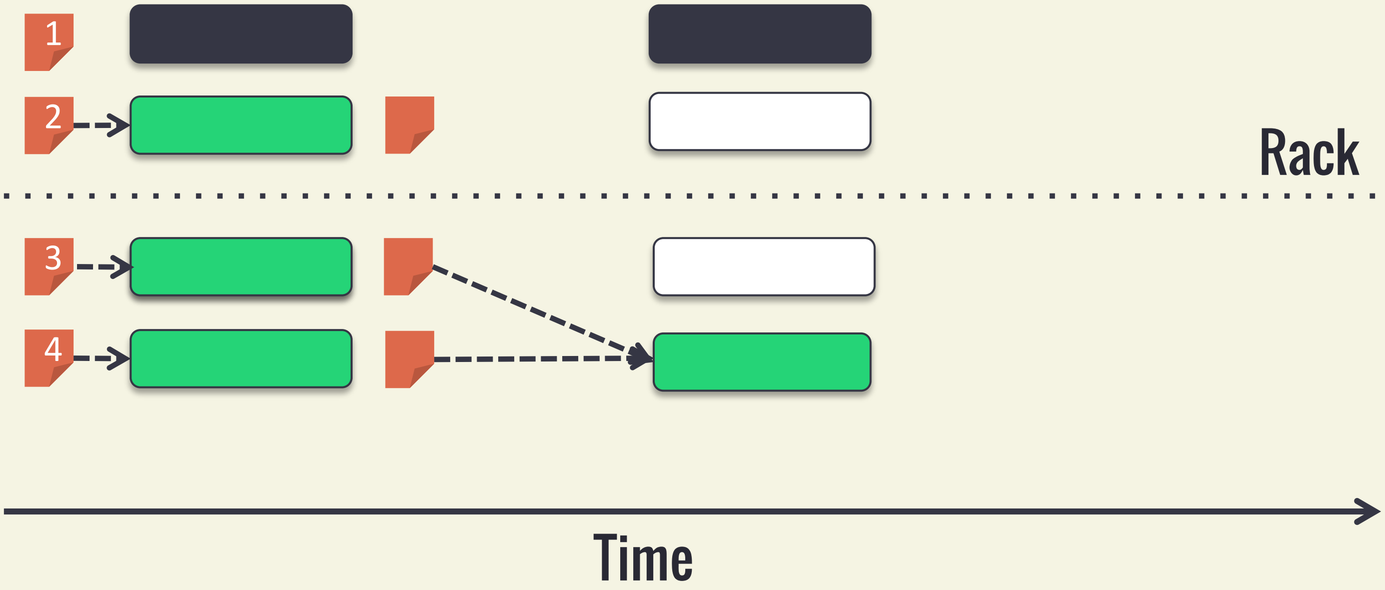
▶ Available Data

🟩 Running

■ Busy

Choice-Aware

Available (N) = 4 Launched (M) = 3 Required (K) = 2



▰ Available Data

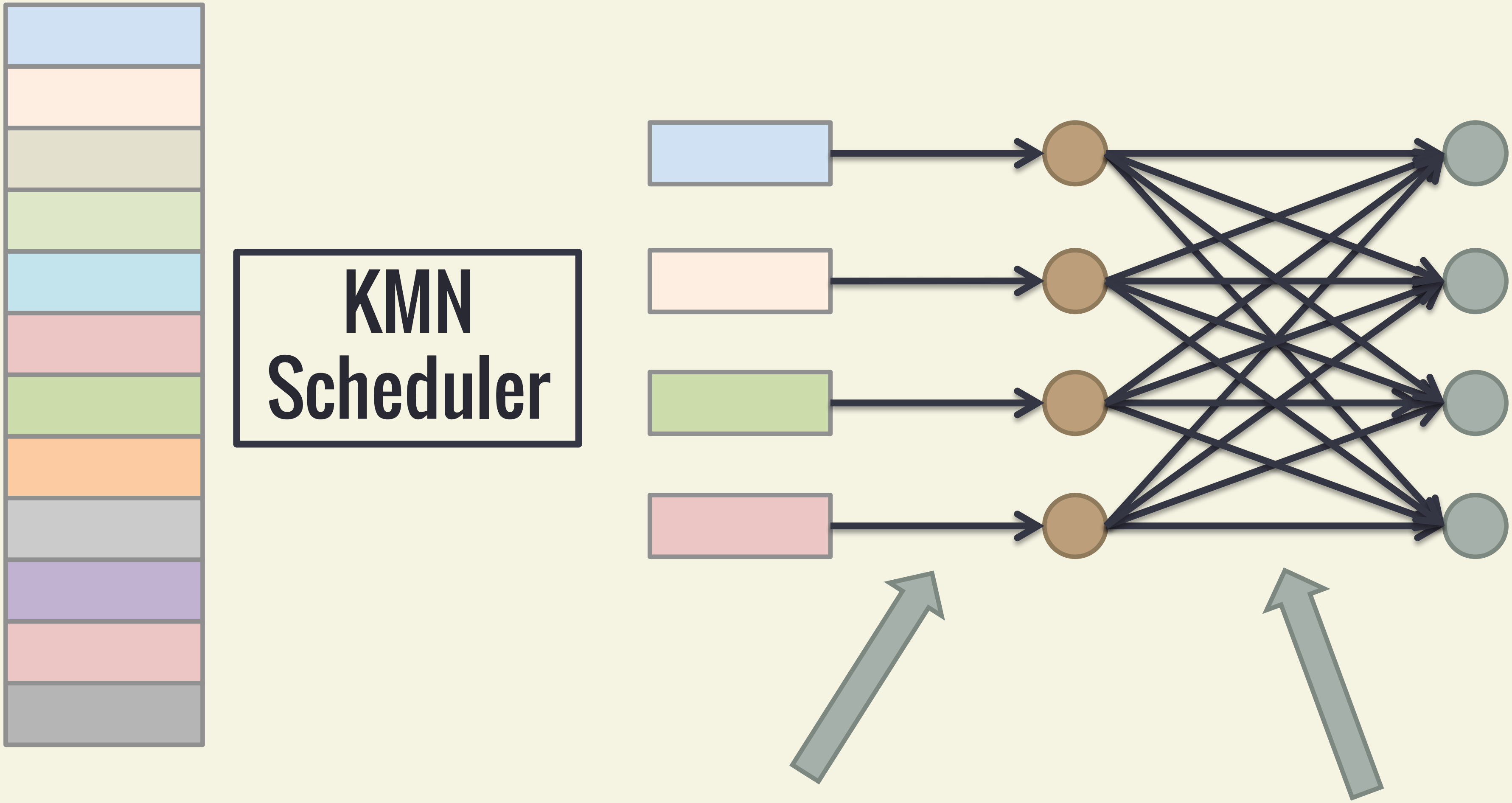
🟩 Running

⬛ Busy

KMN Scheduler

- How much can KMN improve locality
- Propagate benefits across stages
- Handling stragglers

Job → DAG

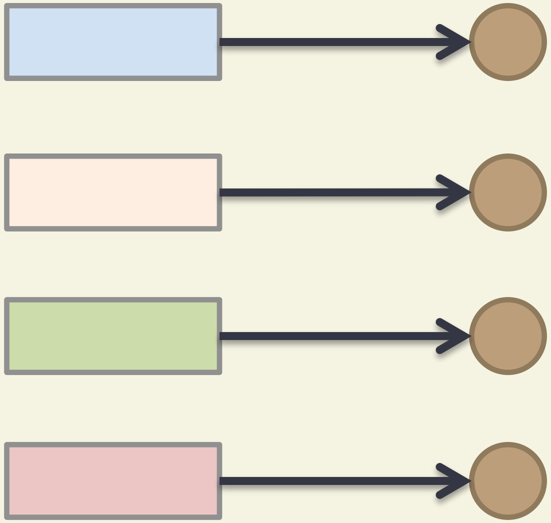


KMN Scheduler

One-to-One

Many-to-One

One-to-One Stages



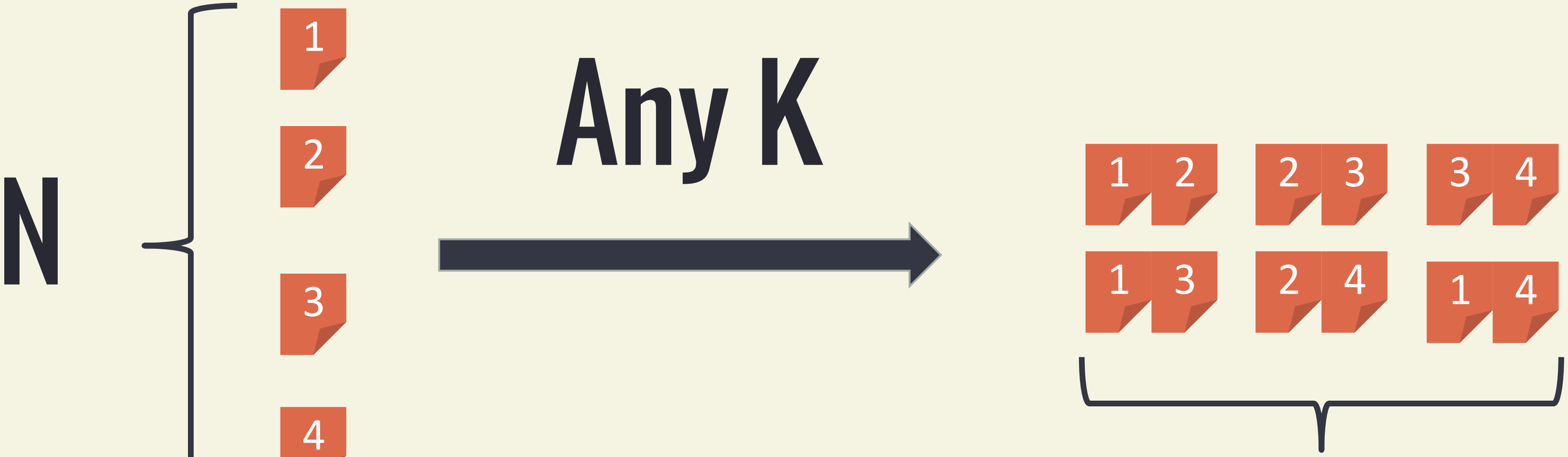
Locality

Disk ~ 100MB/s

Network ~ 10 Gbps (~1GB/s)

Memory ~ 50GB/s

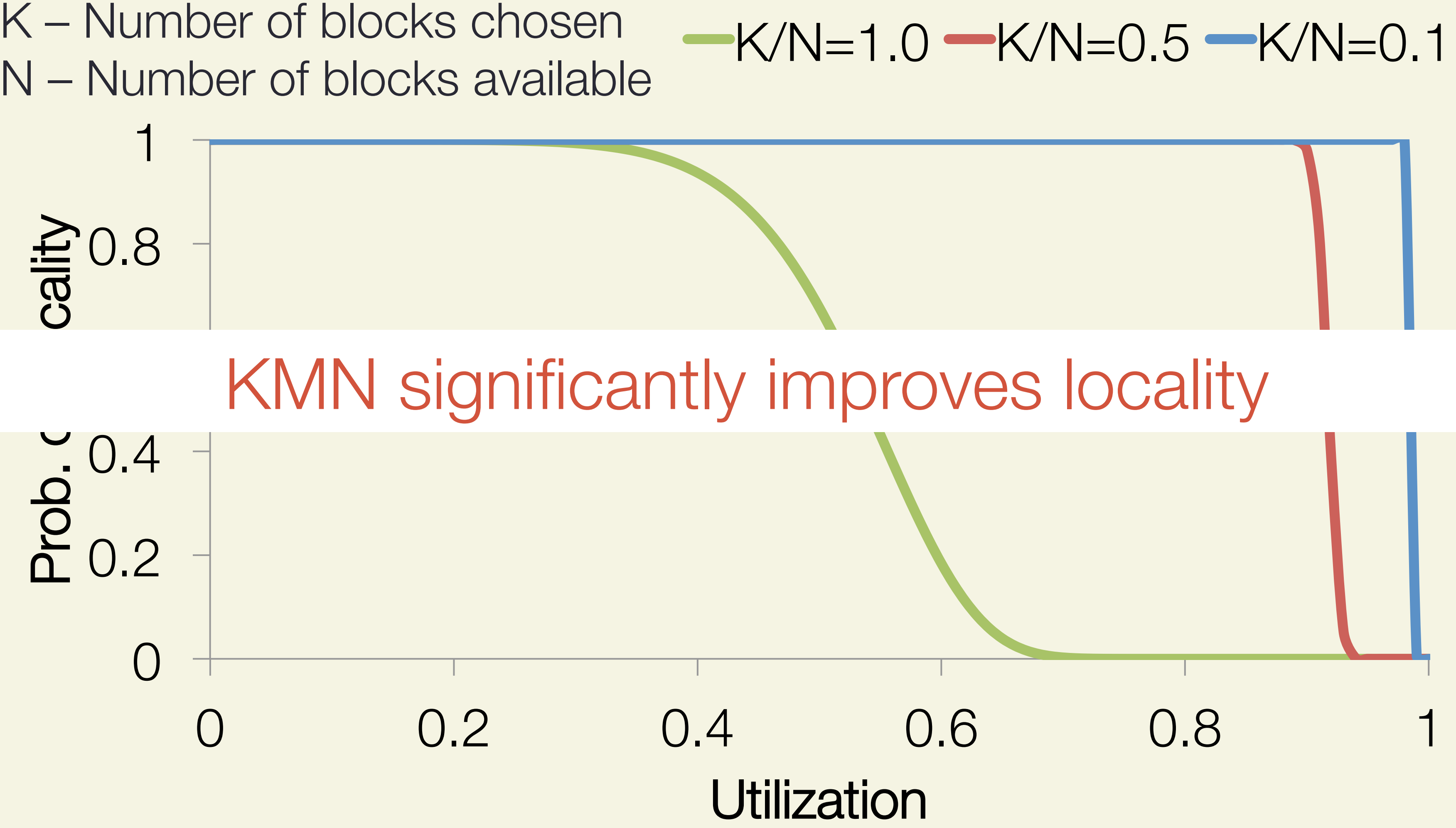
KMN Locality



$\binom{N}{K}$ Choices

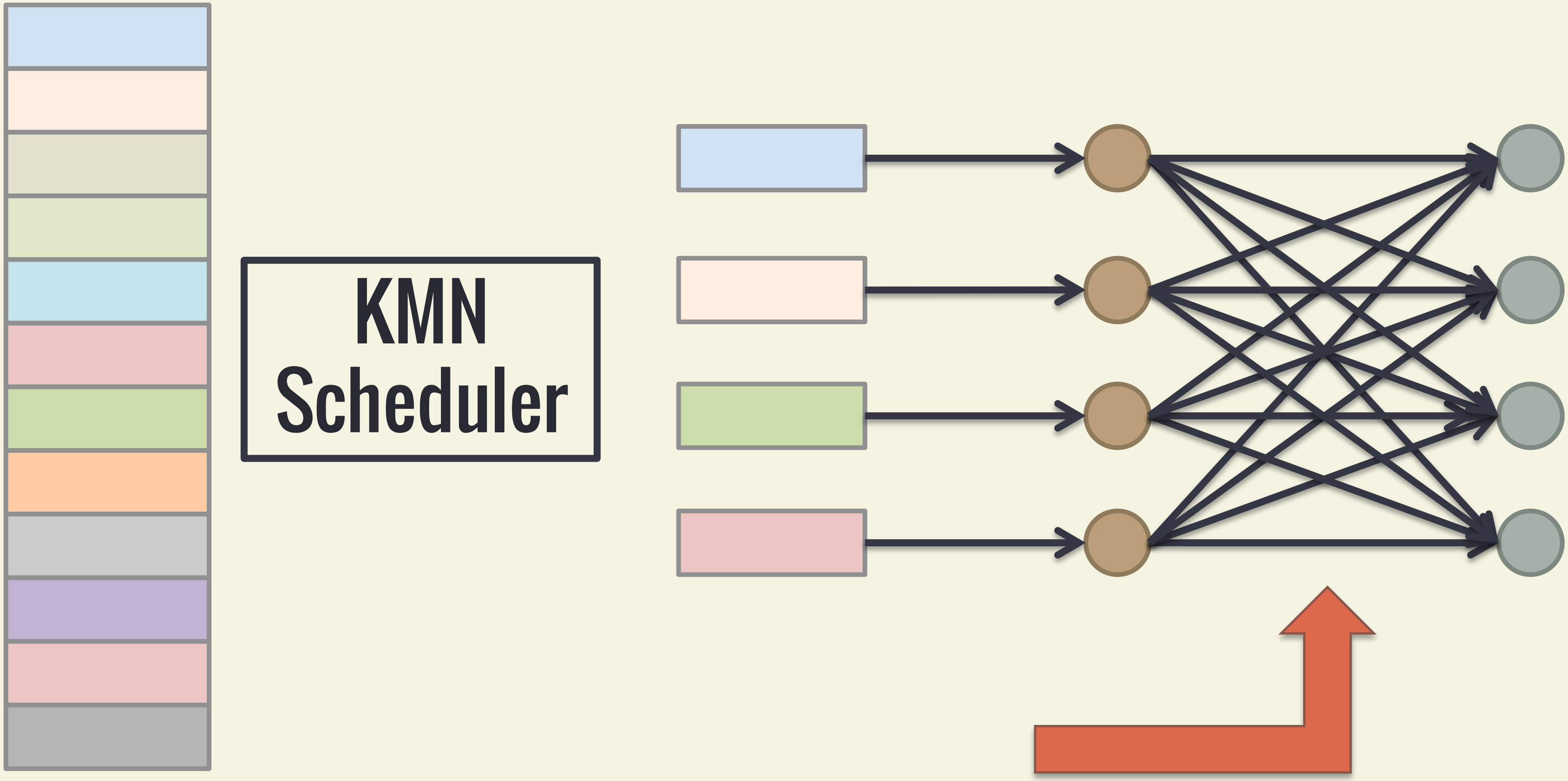
Locality, K=100

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

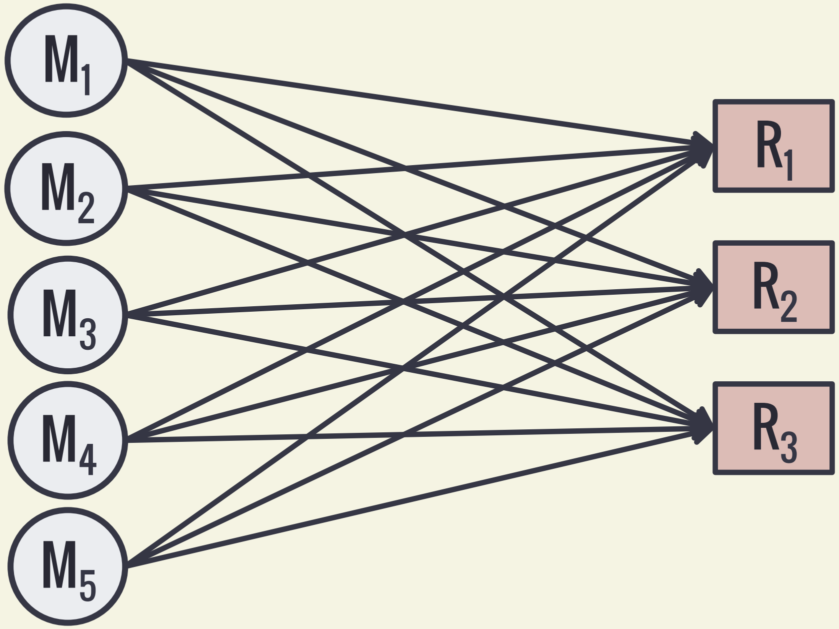


KMN significantly improves locality

Many-to-One Stages

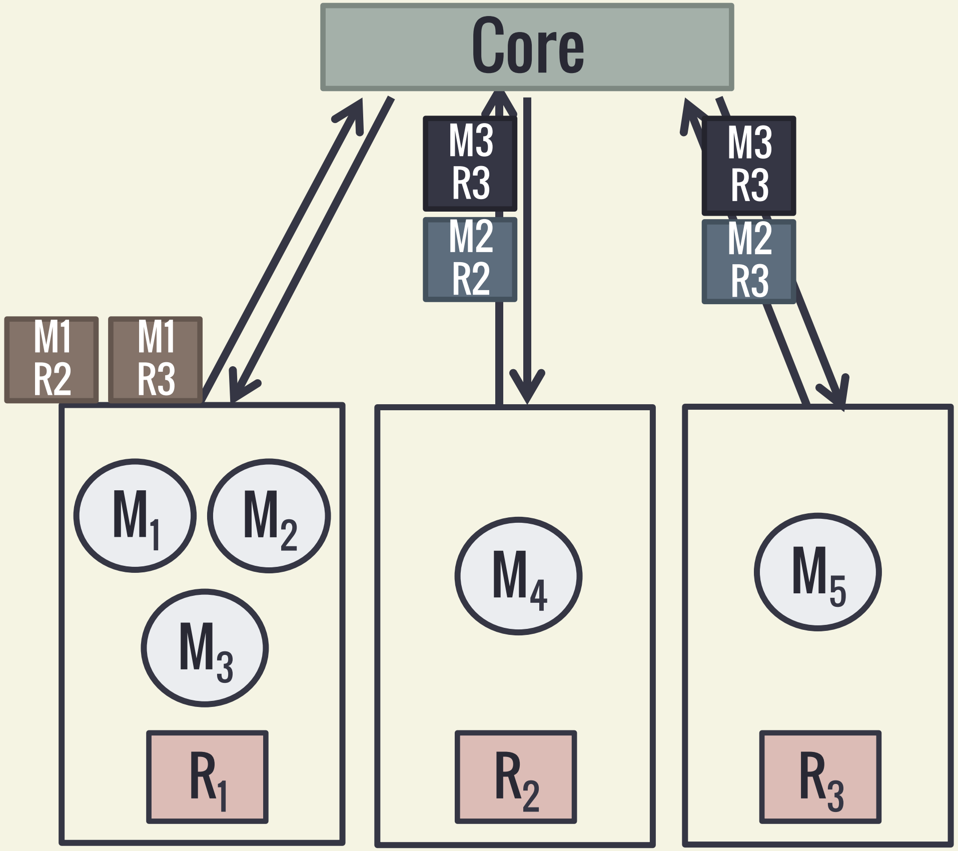


Many-to-One Stage

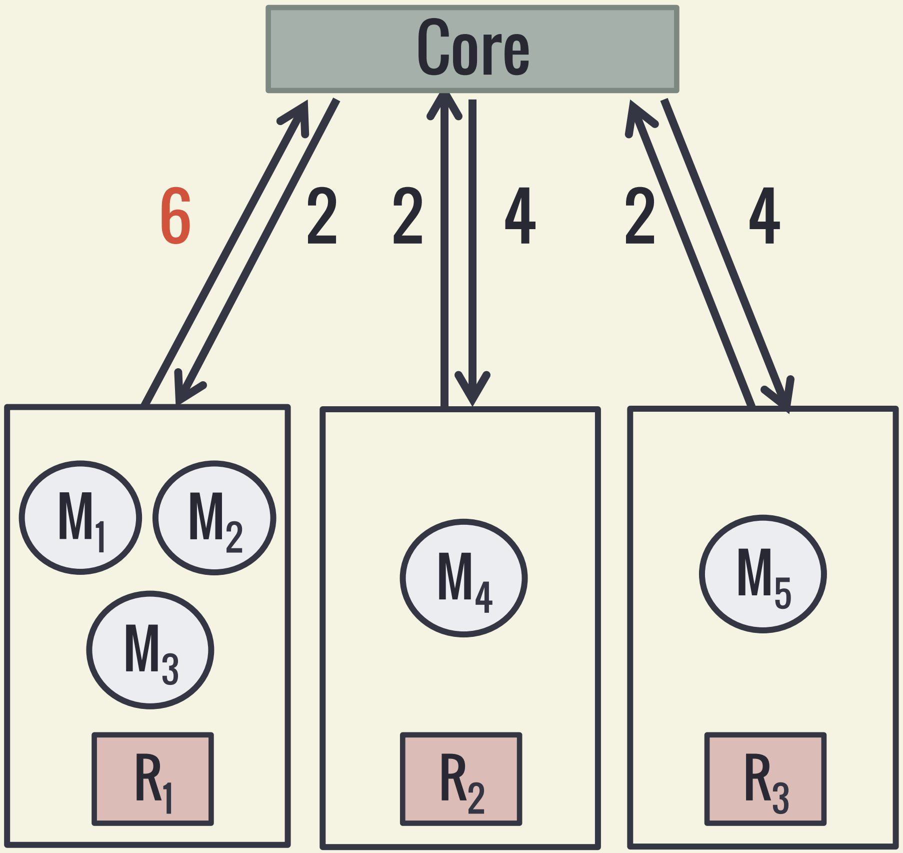


15 transfers

Many-To-One Transfers



Bottleneck Link



Bottleneck Link

Link with Max. transfers

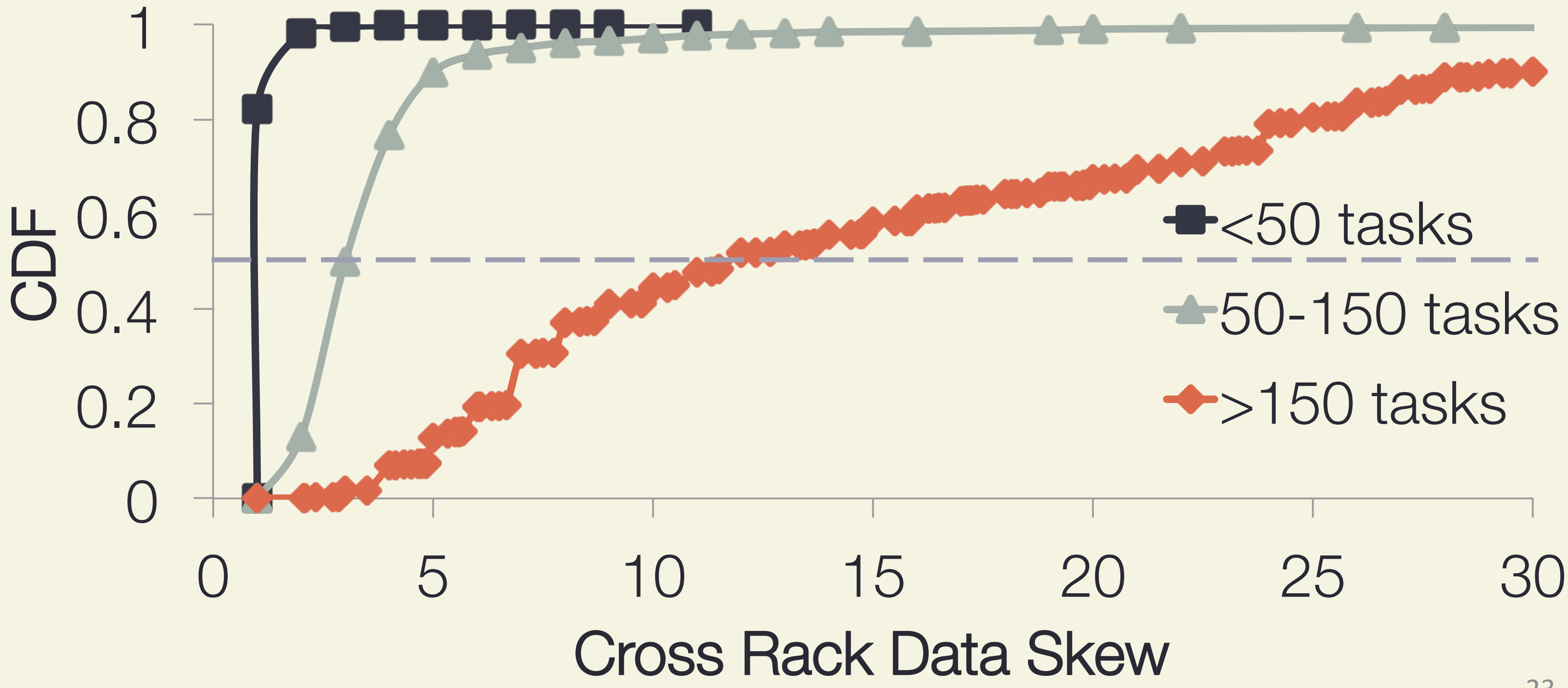
Cross Rack Data Skew

$$\frac{\text{Maximum transfers}}{\text{Minimum transfers}}$$

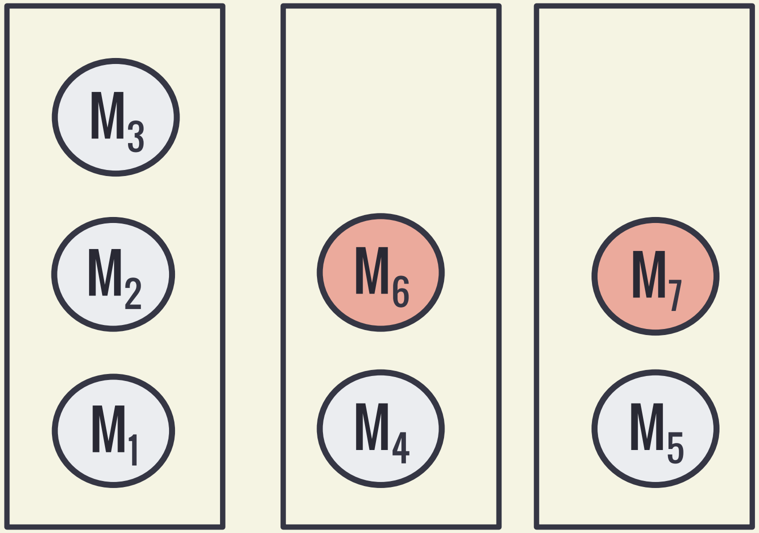
$$= \frac{6}{2} = 3$$

Facebook Trace

Cross Rack Data Skew $\dfrac{\text{Maximum transfers}}{\text{Minimum transfers}}$



Power of Choice

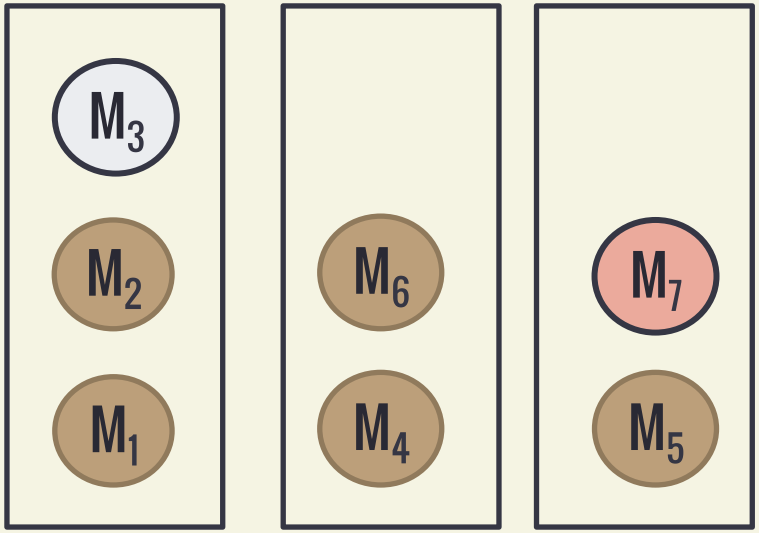


Cross Rack Data Skew = 3

Load balancing: balls and bins

Insight: Run extra tasks (M > K)

Power of Choice

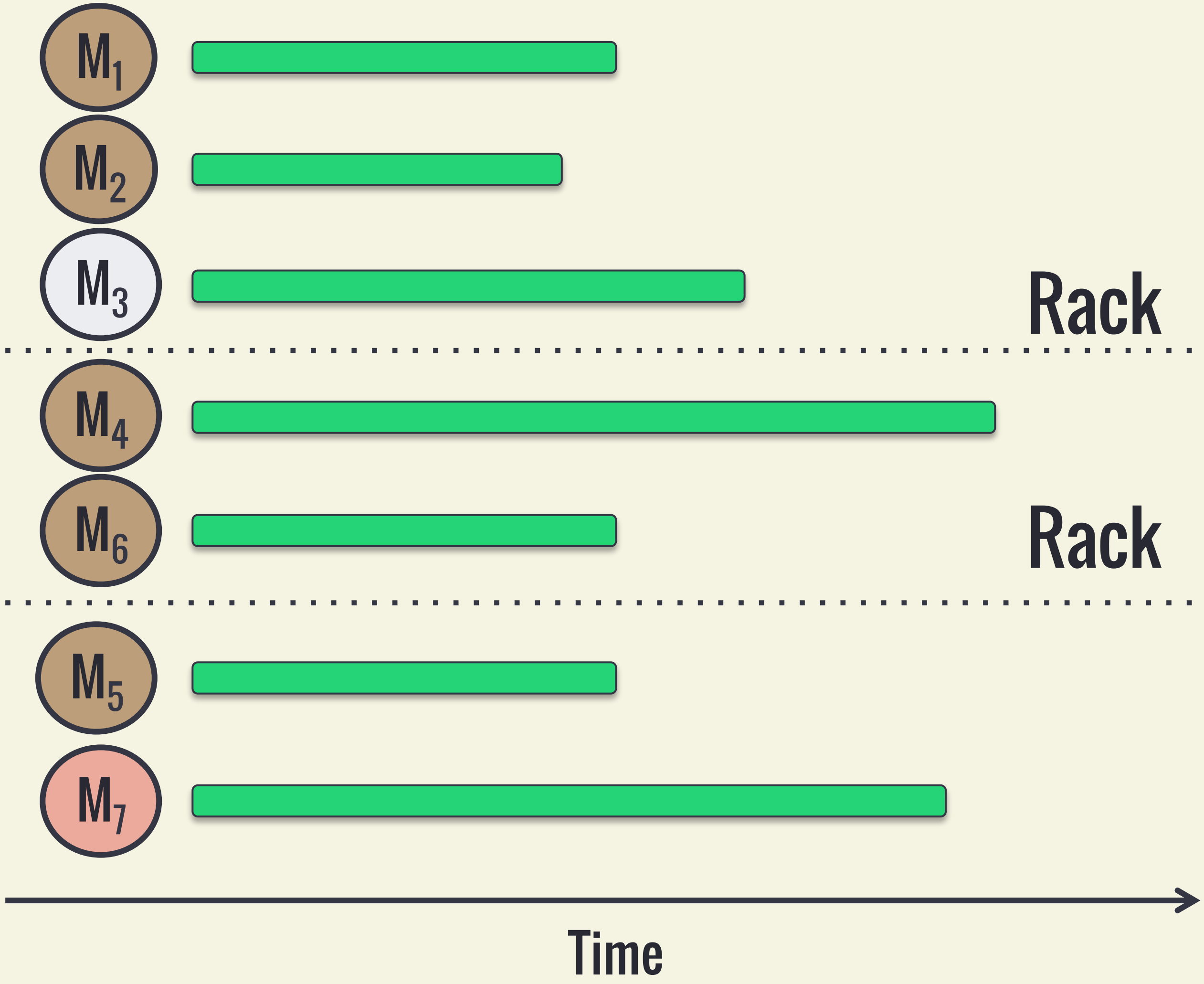


M = 7, K = 5
Cross Rack Data Skew = 2

Technique:

Spread out choice of K tasks to reduce skew

Handling Stragglers



Stragglers

vs.

Cross-Rack Data Skew

Using KMN

```
// 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

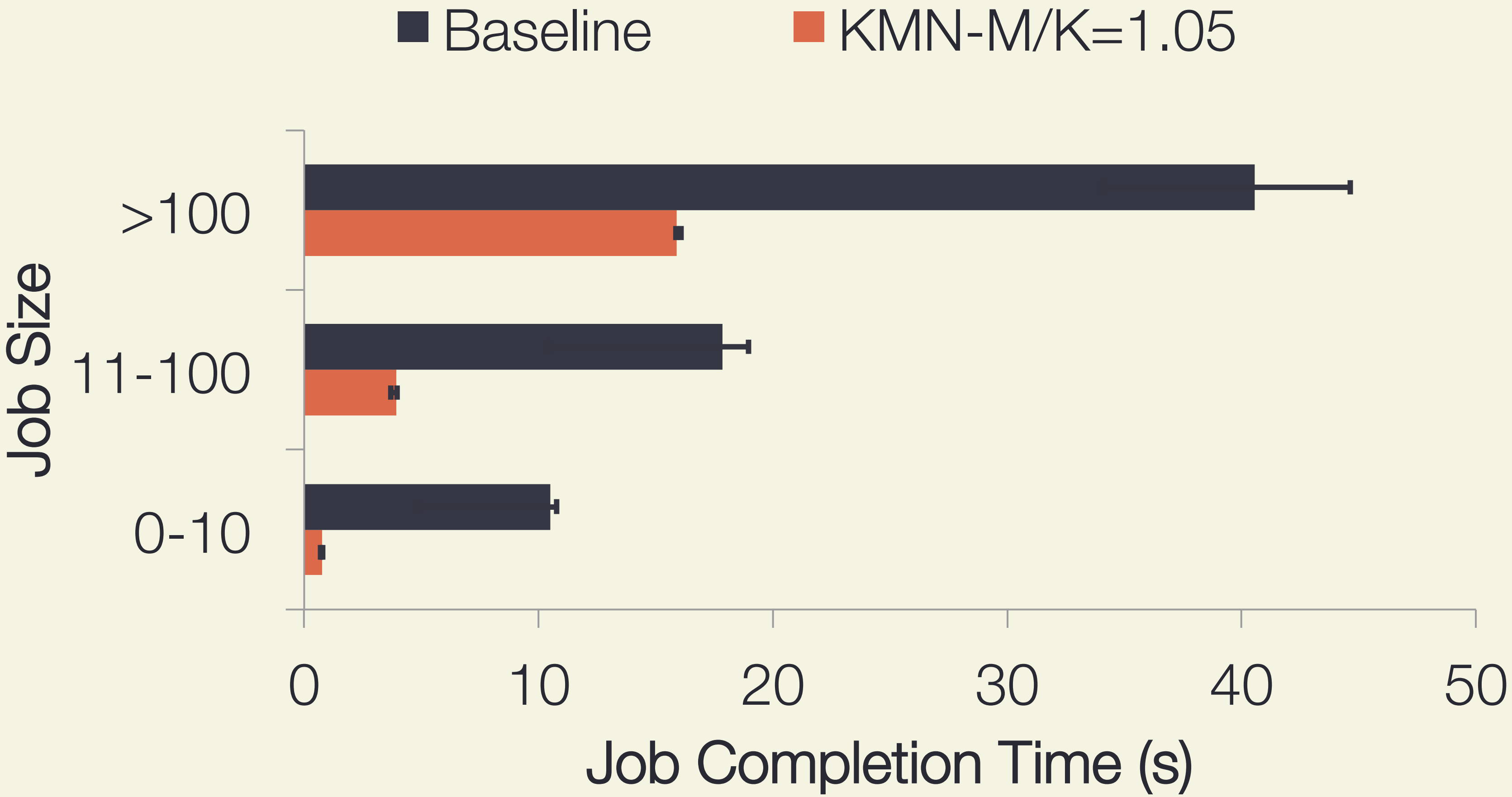
Reducer placement

Varying Utilization

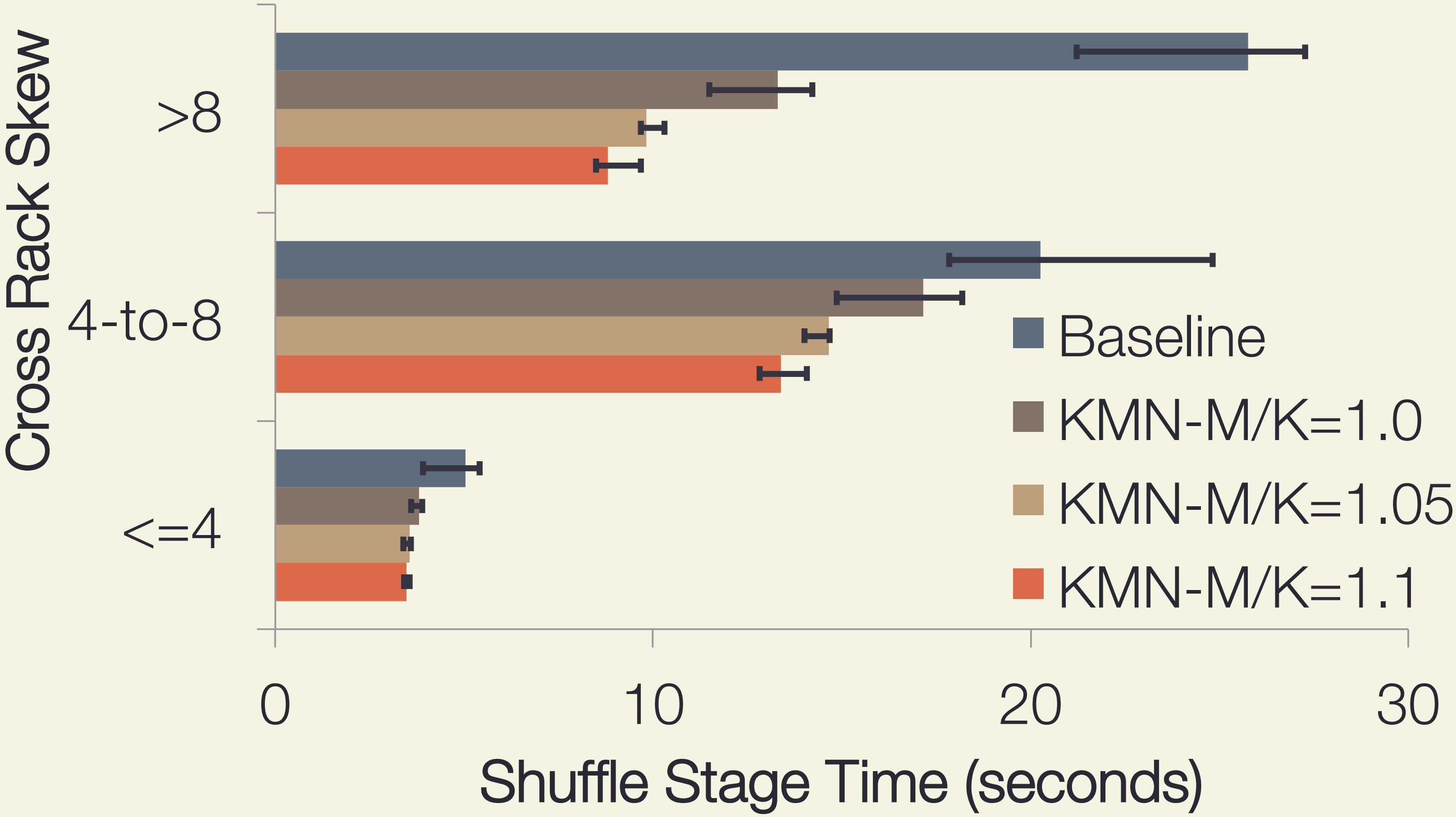
Baseline: Use a pre-selected random sample

Setup: 100 m2.4xlarge EC2 machines, 60GB RAM/mc

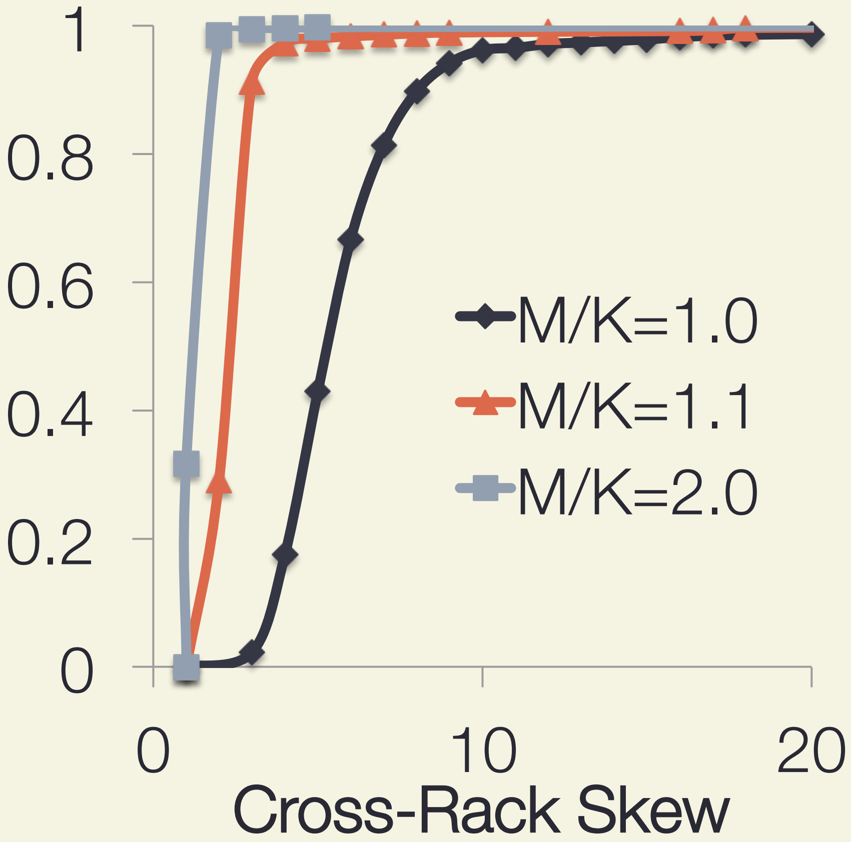
Facebook Overall



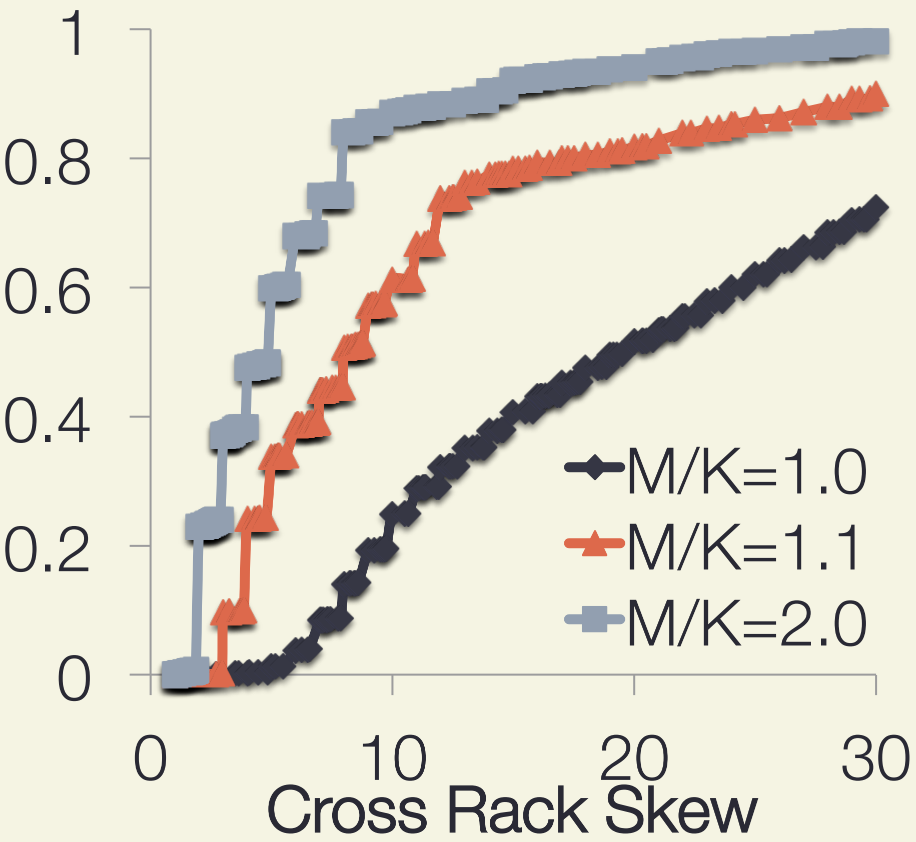
Cross Rack Skew



How many extra tasks ?

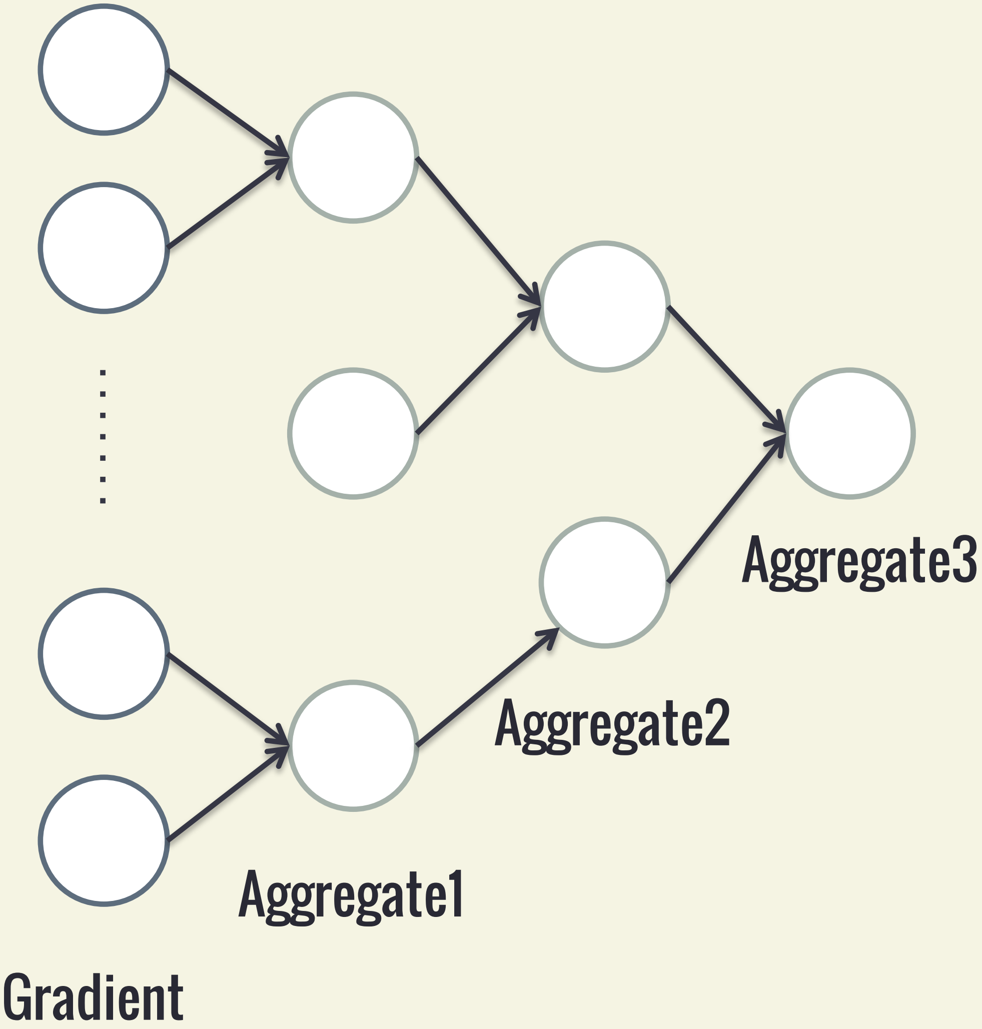


50 - 150 tasks



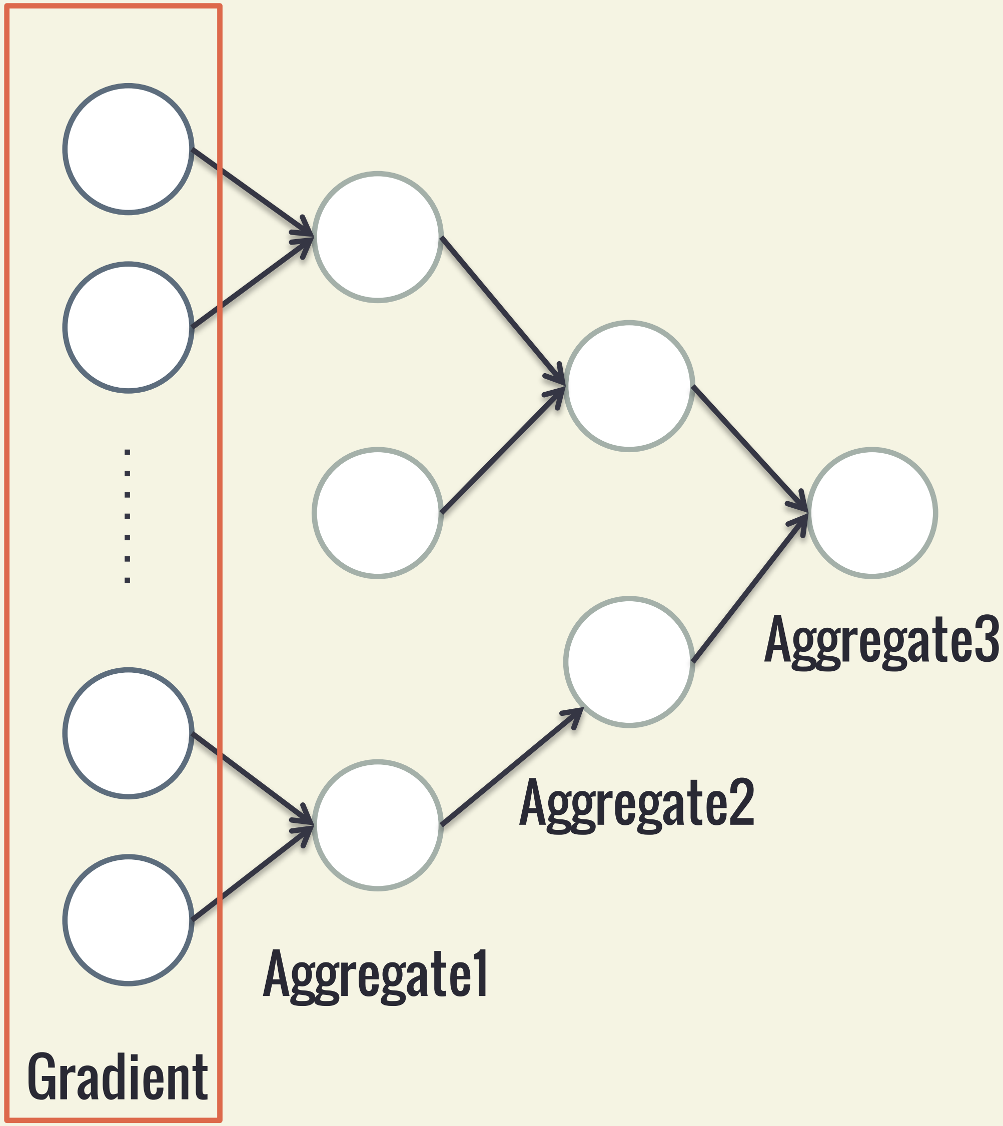
> 150 tasks

KMN: How many stages ?



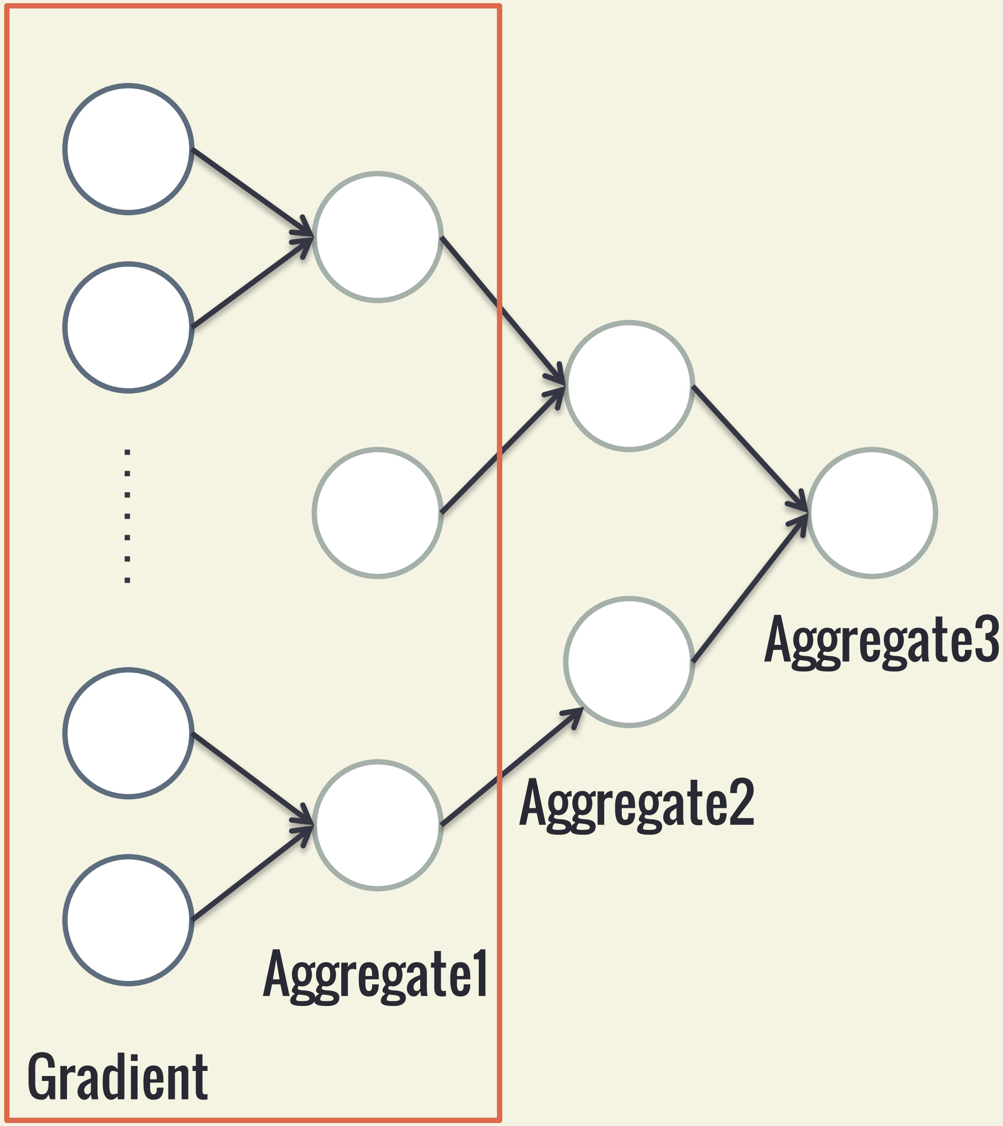
Stochastic
Gradient
Descent

KMN: How many stages ?



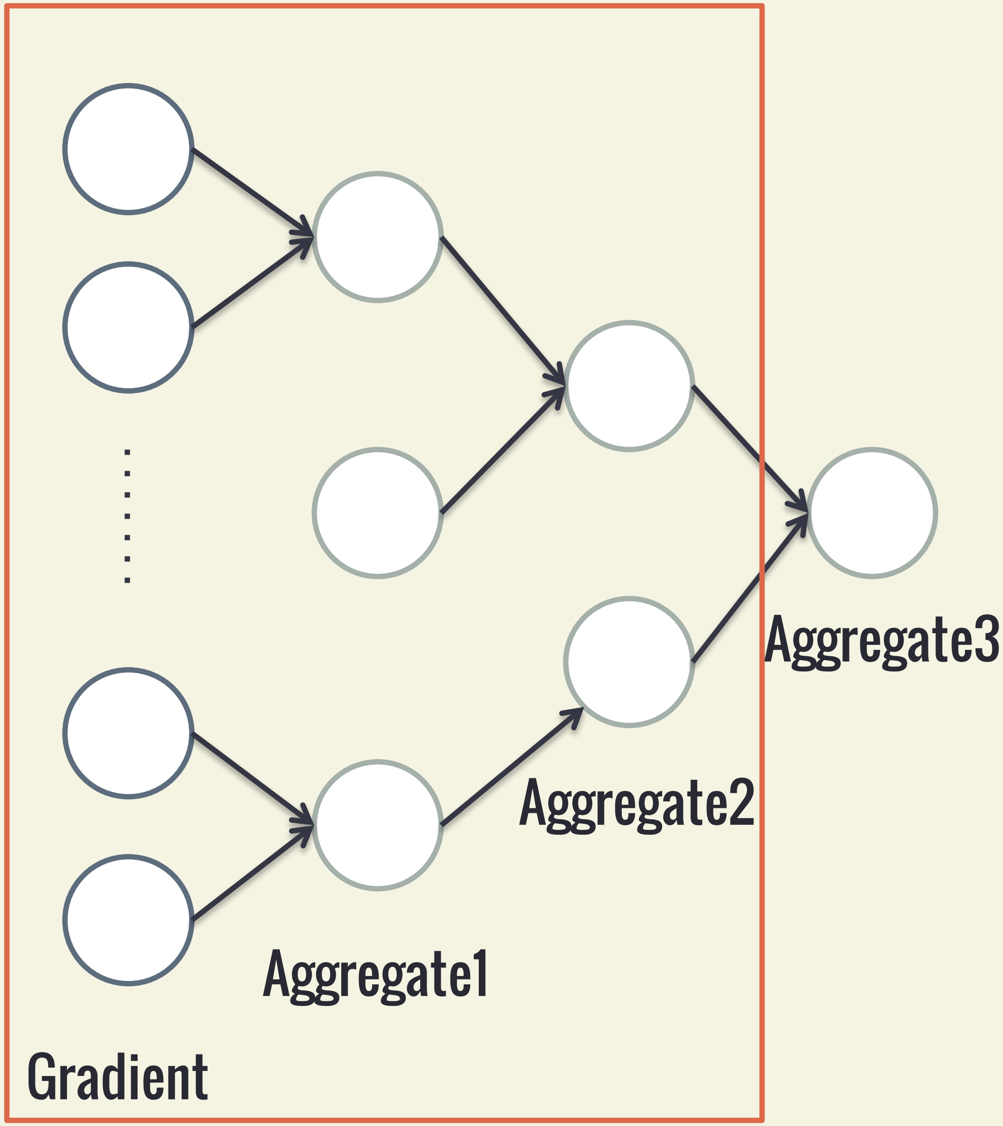
KMN Stages	Time (s)
Gradient	15.27

KMN: How many stages ?



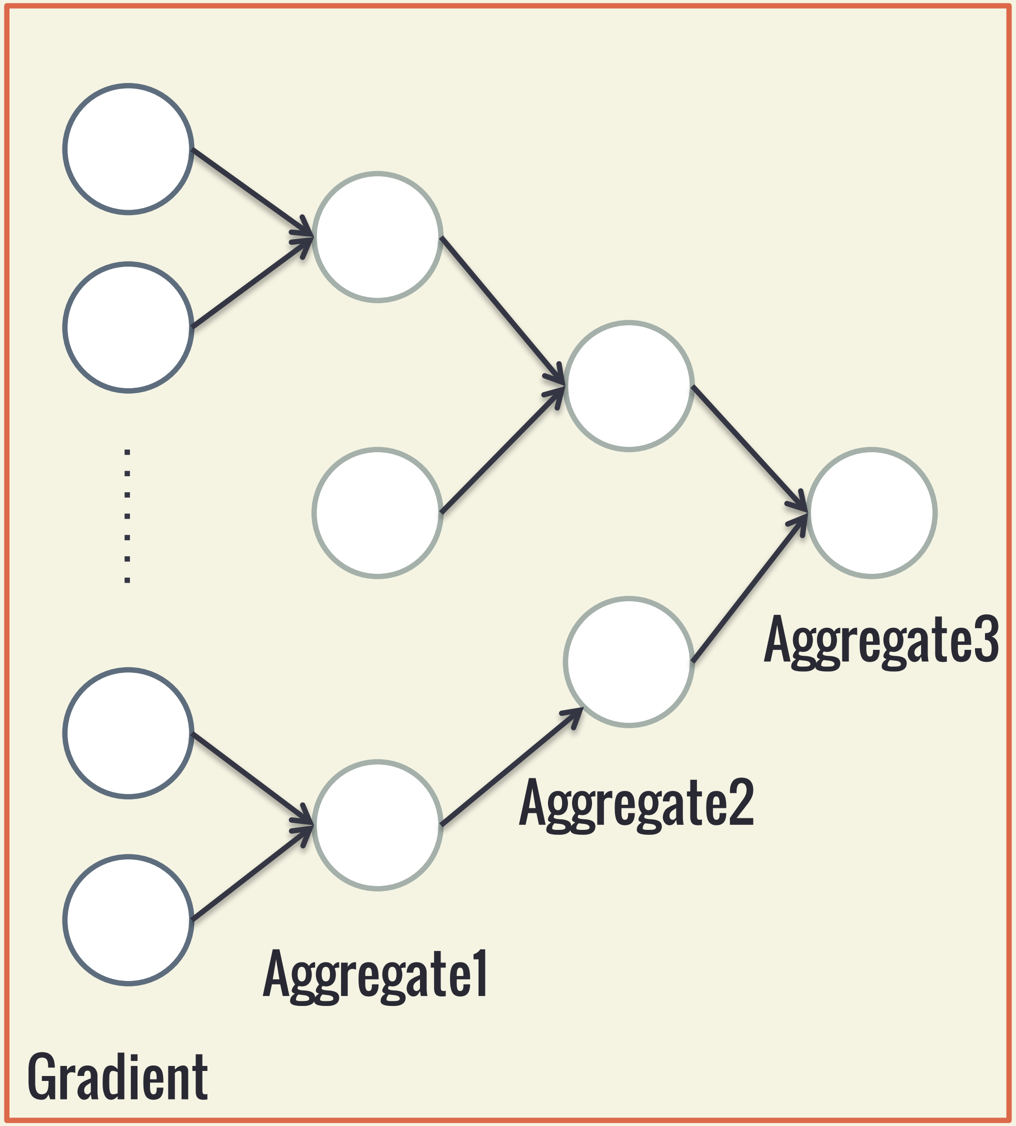
KMN Stages	Time (s)
Gradient	15.27
Gradient + Agg1	12.72

KMN: How many stages ?



KMN Stages	Time (s)
Gradient	15.27
Gradient + Agg1	12.72
Gradient + Agg2	11.79

KMN: How many stages ?



KMN Stages	Time (s)
Gradient	15.27
Gradient + Agg1	12.72
Gradient + Agg2	11.79
Gradient + Agg3	12.09

Related Work

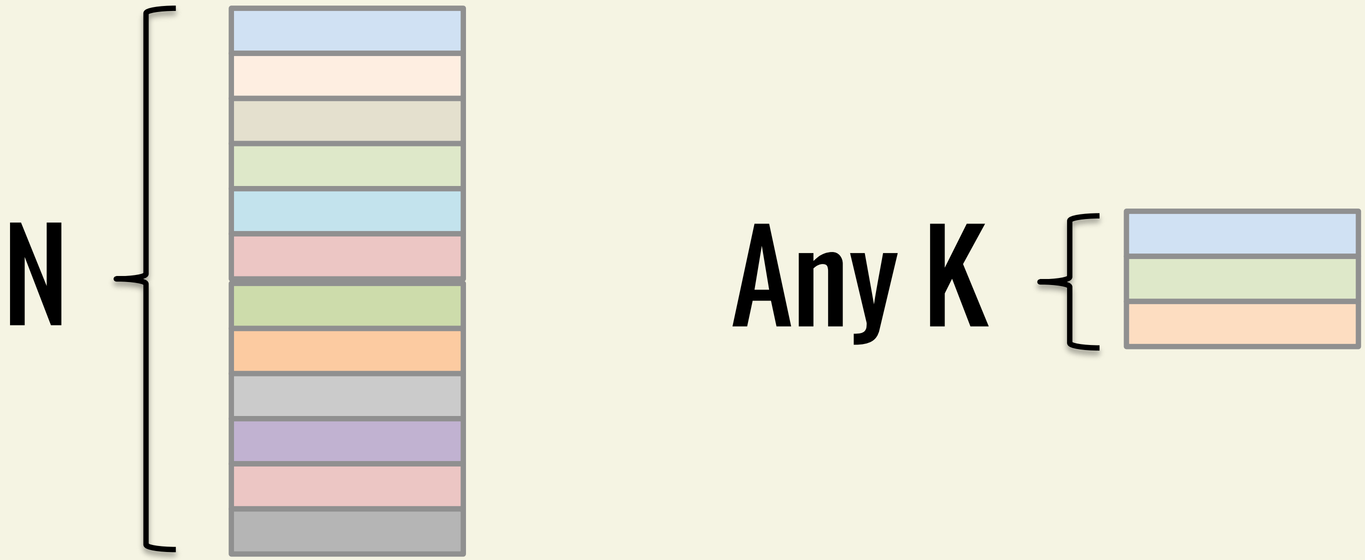
Power of Choice

- 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