

Skews in a Join Operation

- If there are skews in parallel processing,
 - We cannot obtain enough scalability
 - Speed-up is restricted by the slowest PE

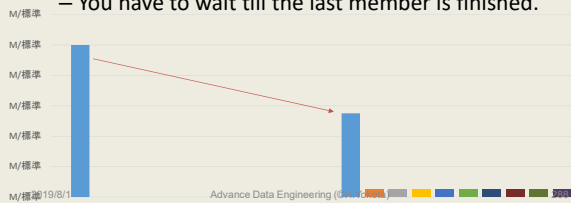
2019/8/1

Advance Data Engineering (©H.Yokota)

287

Image of skews

- Suppose you have 100 jobs and 10 members
- If the jobs are evenly distributed to 9 members (5 jobs for each), but one member takes 55 jobs, then speed up is less than twice.
 - You have to wait till the last member is finished.



Skews in a Join Operation

- If there are skews in parallel processing,
 - We cannot obtain enough scalability
 - Speed-up is restricted by the slowest PE
- Consider GRACE Hash Join
 - There are several reasons for skews
- Assumptions
 - Selection Operations are executed before the Join Operation
 - Results can be output from each PE

2019/8/1

Advance Data Engineering (©H.Yokota)

289

Types of Skews in a Join Operation

- Tuple Placement Skew
 - Tuple distribution skew before starting the query
- Selectivity Skew
 - Skew in the results of the selection before join
- Redistribution Skew
 - Bucket size skew in distribution phase of join operation
- Join Product Skew
 - Skew in the results of join phase

2019/8/1

Advance Data Engineering (©H.Yokota)

290

Handling of Skews in a Join

- Tuple Placement Skew:
 - Adjustment of tuple placement
 - Round-Robin partitioning, Hash partitioning, others
- Selectivity Skew / Redistribution Skew:
 - Fine Bucket Method (will be described soon)
- Join Product Skew:
 - Dynamic bucket allocation / output tuple allocation
- Focus on the Fine Bucket Method

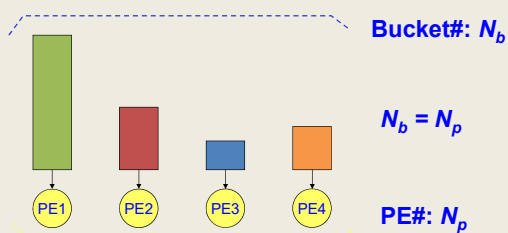
2019/8/1

Advance Data Engineering (©H.Yokota)

291

Fine Bucket Method (1)

- If the number of PEs N_p is equal to the number of buckets N_b
 - Skews cannot be removed with any placement strategies



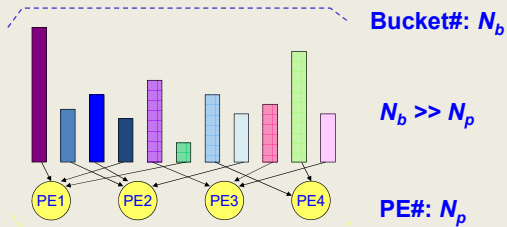
2019/8/1

Advance Data Engineering (©H.Yokota)

292

Fine Bucket Method (2)

- Make the number of buckets N_b quite larger than the number of PEs N_p



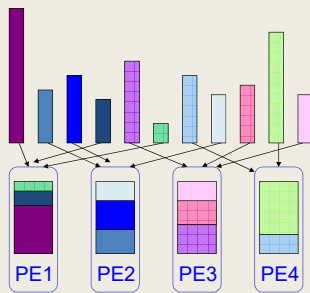
2019/8/1

Advance Data Engineering (©H. Yokota)

293

Fine Bucket Method (3)

- Goal: task size in each PE becomes equivalent



2019/8/1

Advance Data Engineering (©H. Yokota)

294

A Bucket Allocation Strategy

- LPT (Longest Processing Time) First Strategy
 - Heuristics for Minimum Make Span
- Spreading Bucket Method
 - Calculation of bucket size and plan making
 - Distribute buckets to all PEs and make a plan in one of them
 - Merit of Spreading Bucket
 - There is no data concentration in a particular PE + Disk
 - Distribution of fine bucket in each module is similar
 - It is easy to obtain statistics information

2019/8/1

Advance Data Engineering (©H. Yokota)

295

Rotational Bucket Collection (1)

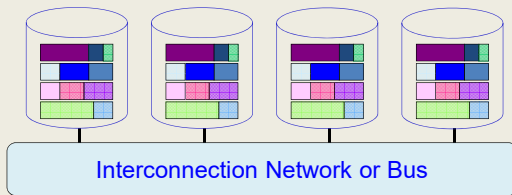
- Routing without congestion during collecting buckets
- Cluster fine buckets into equalized task group by LPT First
 - Distribute N_p subtask group into N_p PEs
- i -th PE PE_i ($1 \leq i \leq N_p$)
 - Read i -th subtask group from $((i + j) - 2) \bmod N_p + 1$ module in j -th step

2019/8/1

Advance Data Engineering (©H.Yokota)

296

Rotational Bucket Collection (2)

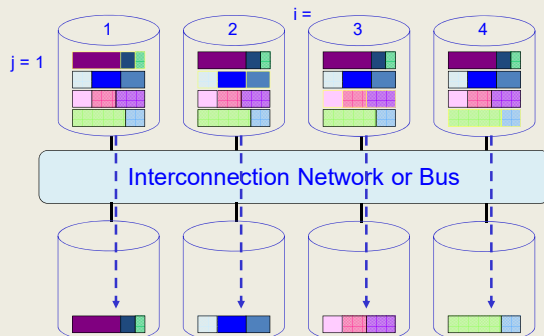


2019/8/1

Advance Data Engineering (©H.Yokota)

297

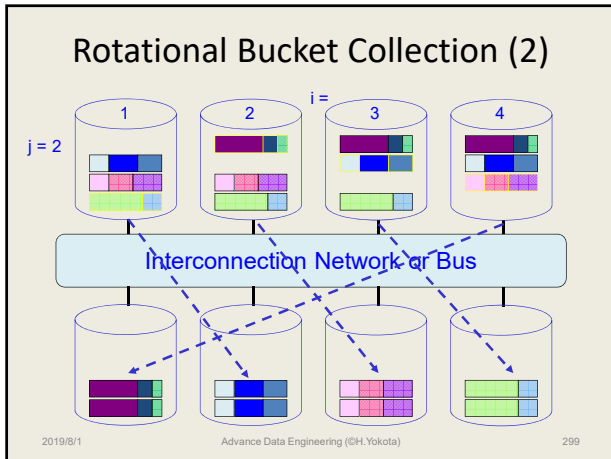
Rotational Bucket Collection (2)

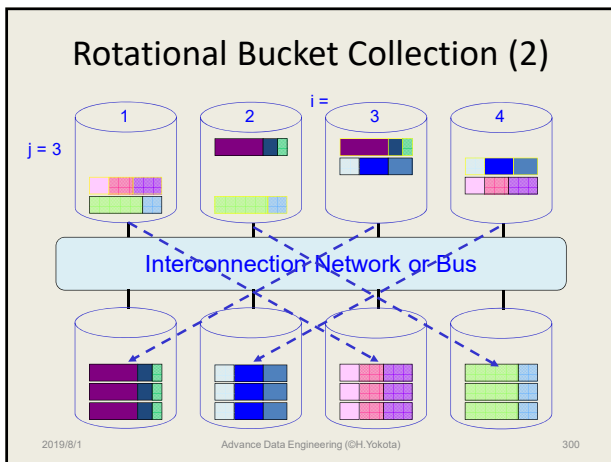


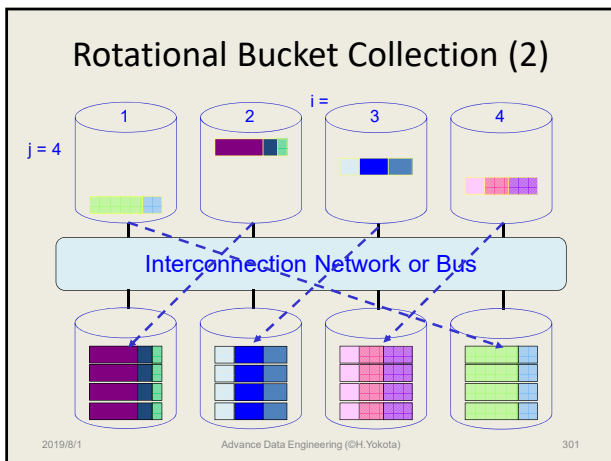
2019/8/1

Advance Data Engineering (©H.Yokota)

298







Process Flow of Fine Buckets

1. All tuples are hashed into N_b buckets, where $N_p \ll N_b$
2. Applying the Spreading-Bucket Method
3. Make task groups by the LPT First Scheduling
4. Applying the Rotational Bucket Collection
5. Do Join operation in each node

2019/8/1

Advance Data Engineering (©H.Yokota)

302

Costs of Fine Buckets

1. I/O for 1 PE during Hash:
 $2 \times (|R| + |S|) / N_p$
2. The Spreading-Bucket Method can be done on-the-fly
3. Data collection for scheduling can be overlap on the I/O
4. I/O for collecting task groups:
 $2 \times (|R| + |S|) / N_p$
5. I/O for Join operation:
 $(|R| + |S|) / N_p$

2019/8/1

Advance Data Engineering (©H.Yokota)

303

Comparison on costs of Fine Buckets

- Let the maximum skew $\alpha\%$
 - When N_p is infinity, $\alpha\%$ for sequential execution time
 - Thus, the execution time: $\alpha / 100 \times 3 \times (|R| + |S|)$
- When we adopt the Fine Bucket Method with the Spreading Bucket Method
 - Total I/O Cost: $5 \times (|R| + |S|) / N_p$
- A Rough Comparison

2019/8/1

Advance Data Engineering (©H.Yokota)

304

Combination of Methods

- Tuple placement Method
 - For Tuple Placement Skews
- Fine Bucket Method
 - For Selectivity / Redistribution Skews
- Dynamic bucket allocation / output tuple allocation Method
 - Join Product Skews
- Each method is independent
 - Combine these methods

2019/8/1

Advance Data Engineering (©H.Yokota)

305

Parallelize Hybrid Hash Join

- Prepare a corresponding Hash Table in each PE
 - It is difficult to build the Hash Table during read disk, because data is fragmented to all disks
 - Build the Hash Table while writing data into the disk in the Phase 1
 - It can reduce time of Phase 2
- However, we can not apply the fine bucket method

2019/8/1

Advance Data Engineering (©H.Yokota)

306

Assignment 12

- In actual situations, it is hard to make the load distribution completely even by the LPT First Strategy.
 - a. Consider the condition of α for the case in which the Fine Bucket Method with the Spreading Bucket Method is effective for $N_p = 100$, when we assume that the maximum skew remains $\beta\%$ (difference between the longest and shortest execution time is $\beta\%$ of the sequential execution time) after applying the Fine Bucket Method.
 - b. Consider approaches to make β smaller.

2019/8/1

Advance Data Engineering (©H.Yokota)

307
