Carnegie Mellon University

13 Query Execution – Part II

Intro to Database Systems
15-445/15-645
Fall 2020

AP Andy Pavlo
Computer Science
Carnegie Mellon University

ADMINISTRIVIA

Homework #3 is due Sun Oct 18th @ 11:59pm

Mid-Term Exam is Wed Oct 21st

→ Morning Session: 9:00am ET

→ Afternoon Session: 3:20pm ET

Project #2 is due Sun Oct 25th @ 11:59pm

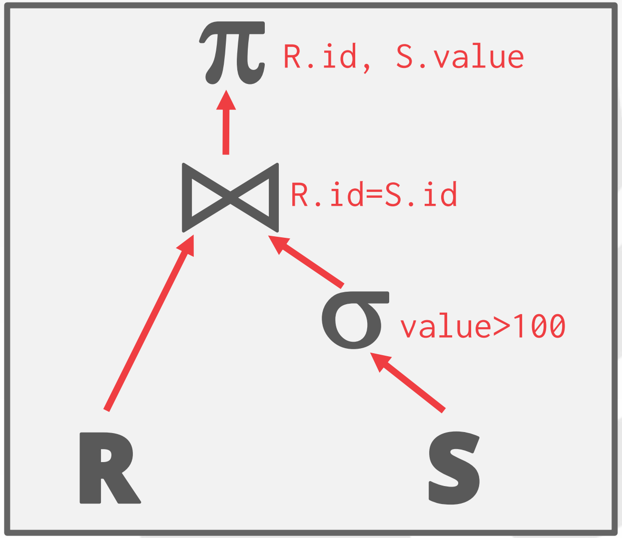
QUERY EXECUTION

We discussed last class how to compose operators together to execute a query plan.

We assumed that the queries execute with a single worker (e.g., thread).

We now need to talk about how to execute with multiple workers…

```
SELECT R.id, S.cdate
  FROM R JOIN S
    ON R.id = S.id
 WHERE S.value > 100
```



WHY CARE ABOUT PARALLEL EXECUTION?

Increased performance.
→ Throughput
→ Latency

Increased responsiveness and availability.

Potentially lower ***total cost of ownership*** (TCO).

PARALLEL VS. DISTRIBUTED

Database is spread out across multiple **resources** to improve different aspects of the DBMS.

Appears as a single database instance to the application.

→ SQL query for a single-resource DBMS should generate same result on a parallel or distributed DBMS.

PARALLEL VS. DISTRIBUTED

Parallel DBMSs:

→ Resources are physically close to each other.
→ Resources communicate with high-speed interconnect.
→ Communication is assumed to cheap and reliable.

Distributed DBMSs:

→ Resources can be far from each other.
→ Resources communicate using slow(er) interconnect.
→ Communication cost and problems cannot be ignored.

TODAY'S AGENDA

Process Models

Execution Parallelism

I/O Parallelism

PROCESS MODEL

A DBMS's **process model** defines how the system is architected to support concurrent requests from a multi-user application.

A **worker** is the DBMS component that is responsible for executing tasks on behalf of the client and returning the results.

PROCESS MODELS

Approach #1: Process per DBMS Worker

Approach #2: Process Pool

Approach #3: Thread per DBMS Worker

PROCESS PER WORKER

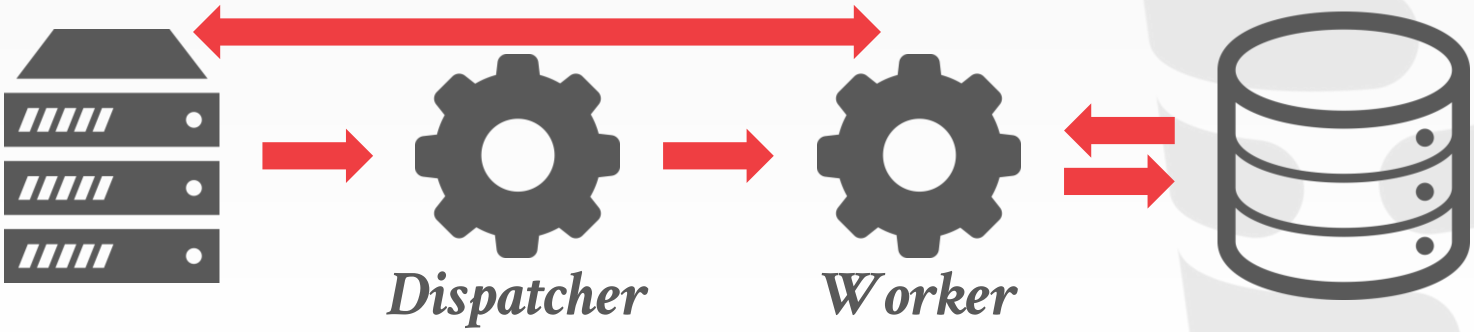
Each worker is a separate OS process.

→ Relies on OS scheduler.

→ Use shared-memory for global data structures.

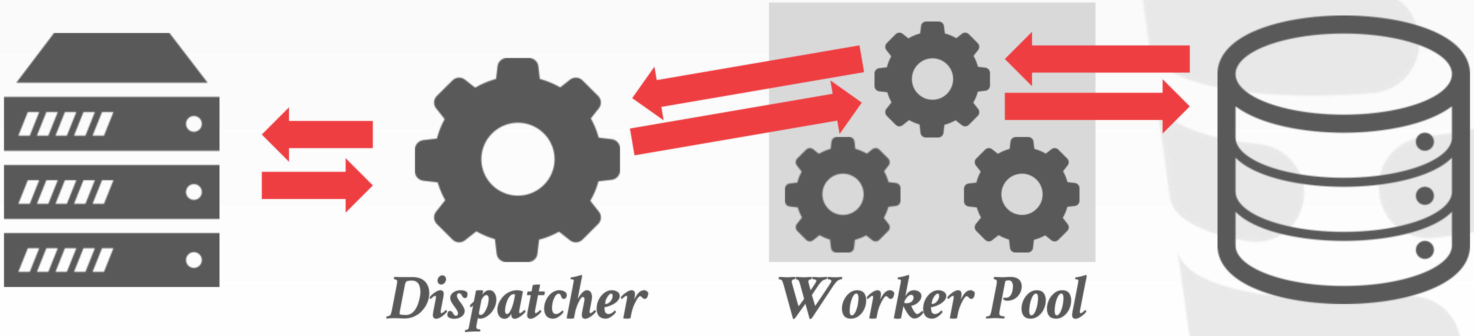
→ A process crash doesn't take down entire system.

→ Examples: IBM DB2, Postgres, Oracle



PROCESS POOL

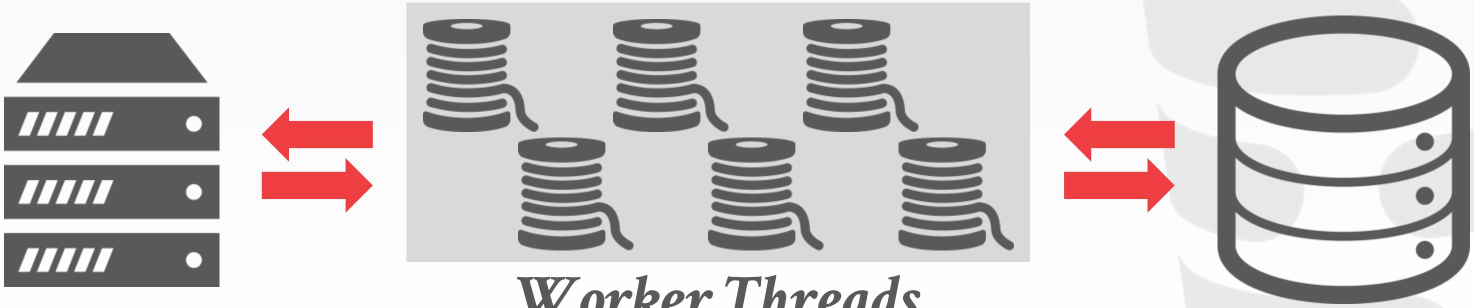
A worker uses any process that is free in a pool
→ Still relies on OS scheduler and shared memory.
→ Bad for CPU cache locality.
→ Examples: IBM DB2, Postgres (2015)



THREAD PER WORKER

Single process with multiple worker threads.

→ DBMS manages its own scheduling.
→ May or may not use a dispatcher thread.
→ Thread crash (may) kill the entire system.
→ Examples: IBM DB2, MSSQL, MySQL, Oracle (2014)



Worker Threads

PROCESS MODELS

Using a multi-threaded architecture has several advantages:

→ Less overhead per context switch.
→ Do not have to manage shared memory.

The thread per worker model does **not** mean that the DBMS supports intra-query parallelism.

Andy is not aware of any new DBMS from last 10 years that doesn't use threads unless they are Redis or Postgres forks.

SCHEDULING

For each query plan, the DBMS decides where, when, and how to execute it.

→ How many tasks should it use?
→ How many CPU cores should it use?
→ What CPU core should the tasks execute on?
→ Where should a task store its output?

The DBMS ***always*** knows more than the OS.

SQL SERVER – SQLOS

SQLOS is a user-level OS layer that runs inside of the DBMS and manages provisioned hardware resources.

→ Determines which tasks are scheduled onto which threads.

→ Also manages I/O scheduling and higher-level concepts like logical database locks.

Non-preemptive thread scheduling through instrumented DBMS code.

SQL

SQLOS is a us
the DBMS an
resources.

→ Determines
threads.

→ Also manag
like logical

Non-preemp
instrumente

Join Extra Crunch
Login
Search
Startups
Videos
Audio
Newsletters
Extra Crunch
Advertise
Events
—
More

Transportation
Apple
Tesla
Security

How Microsoft brought SQL Server to Linux

Frederic Lardinois @fredericl / 12:00 pm EDT • July 17, 2017

Comment

Back in 2016, when **Microsoft** announced that SQL Server would soon run on Linux, the news came as a major surprise to users and pundits alike. Over the course of the last year, Microsoft's support for Linux (and open source in general), has come into clearer focus and the company's mission now seems to be all about bringing its tools to wherever its users are.

The company today launched the first release candidate of SQL Server 2017, which will be the first version to run on Windows, Linux and in Docker containers. The Docker container alone has already seen more than 1 million pulls, so there can be no doubt that there is a lot of interest in this new version. And while there are plenty of new features and speed improvements in this new version, the fact that SQL Server 2017 supports Linux remains one of the most interesting aspects of this release.

Ahead of today's announcement, I talked to Rohan Kumar, the general manager of Microsoft's Database Systems group, to get a bit more info about the history of this project and how his team managed to bring an extremely complex piece of software like SQL Server to Linux. Kumar, who has been at Microsoft for more than 18 years, noted that his team noticed many enterprises were starting to use SQL Server for their mission-critical workloads. But at the same time, they were also working in mixed environments that included both Windows Server and Linux. For many of these businesses, not being able to run their database of choice on Linux became a friction point.

"Talking to enterprises, it became clear that doing this was necessary," Kumar said. "We were forcing customers to use Windows as their platform of choice." In another incarnation of Microsoft, that probably would've been seen as something positive, but the company's strategy today is quite different.

SQL SERVER – SQLOS

SQLOS quantum is 4 ms but the scheduler cannot enforce that.

DBMS developers must add explicit yield calls in various locations in the source code.

```
SELECT * FROM A WHERE A.val = ?
```

Approximate Plan

```
for t in range(table.num_tuples):
  tuple = get_tuple(table, t)
  if eval(predicate, tuple, params):
    emit(tuple)
```

SQL SERVER – SQLOS

SQLOS quantum is 4 ms but the scheduler cannot enforce that.

DBMS developers must add explicit yield calls in various locations in the source code.

```
SELECT * FROM A WHERE A.val = ?
```

```
last = now()
for t in range(table.num_tuples):
  tuple = get_tuple(table, t)
  if eval(predicate, tuple, params):
    emit(tuple)
  if now() - last > 4ms:
    yield
    last = now()
```

INTER- VS. INTRA-QUERY PARALLELISM

Inter-Query: Different queries are executed concurrently.

→ Increases throughput & reduces latency.

Intra-Query: Execute the operations of a single query in parallel.

→ Decreases latency for long-running queries.

INTER-QUERY PARALLELISM

Improve overall performance by allowing multiple queries to execute simultaneously.

If queries are read-only, then this requires little coordination between queries.

If multiple queries are updating the database at the same time, then this is hard to do correctly…

Lecture 16

INTRA-QUERY PARALLELISM

Improve the performance of a single query by executing its operators in parallel.

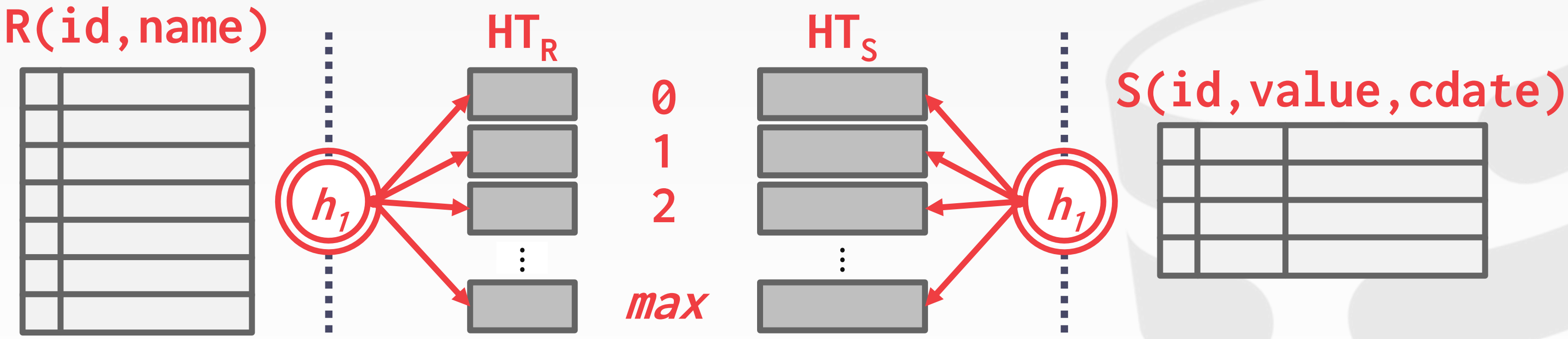
Think of organization of operators in terms of a ***producer/consumer*** paradigm.

There are parallel algorithms for every relational operator.

→ Can either have multiple threads access centralized data structures or use partitioning to divide work up.

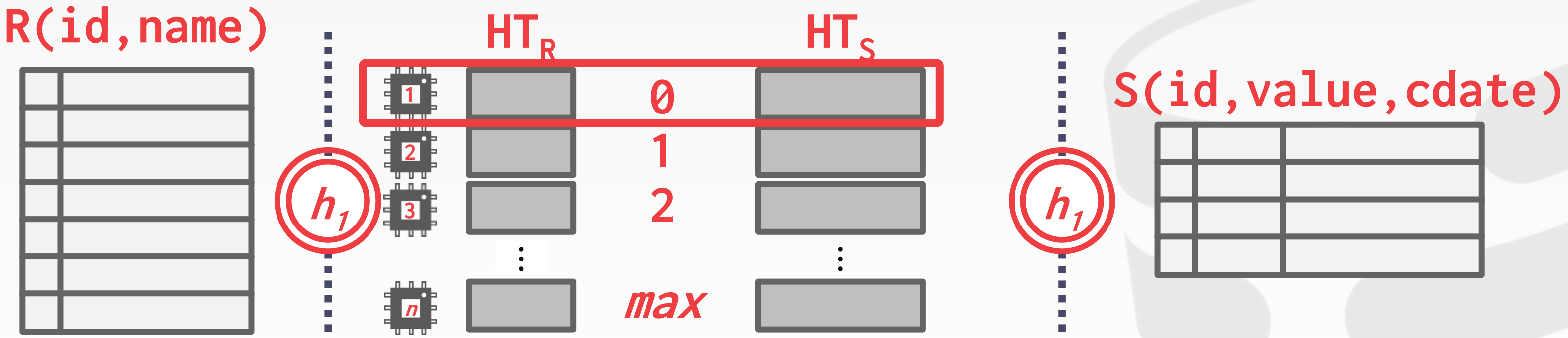
PARALLEL GRACE HASH JOIN

Use a separate worker to perform the join for each level of buckets for **R** and **S** after partitioning.



PARALLEL GRACE HASH JOIN

Use a separate worker to perform the join for each level of buckets for **R** and **S** after partitioning.



INTRA-QUERY PARALLELISM

Approach #1: Intra-Operator (Horizontal)

Approach #2: Inter-Operator (Vertical)

Approach #3: Bushy

INTRA-OPERATOR PARALLELISM

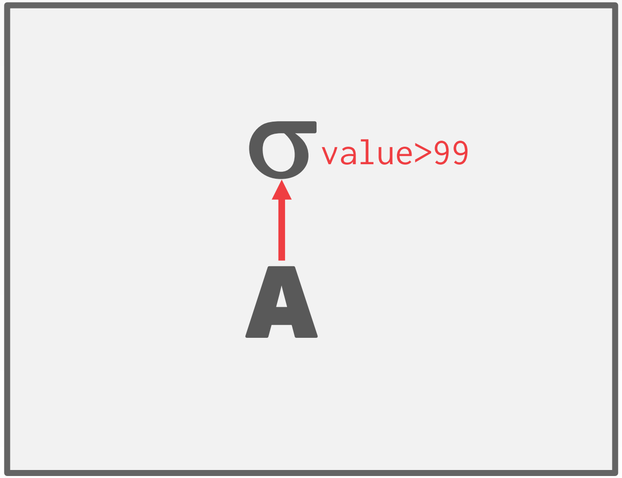
Approach #1: Intra-Operator (Horizontal)

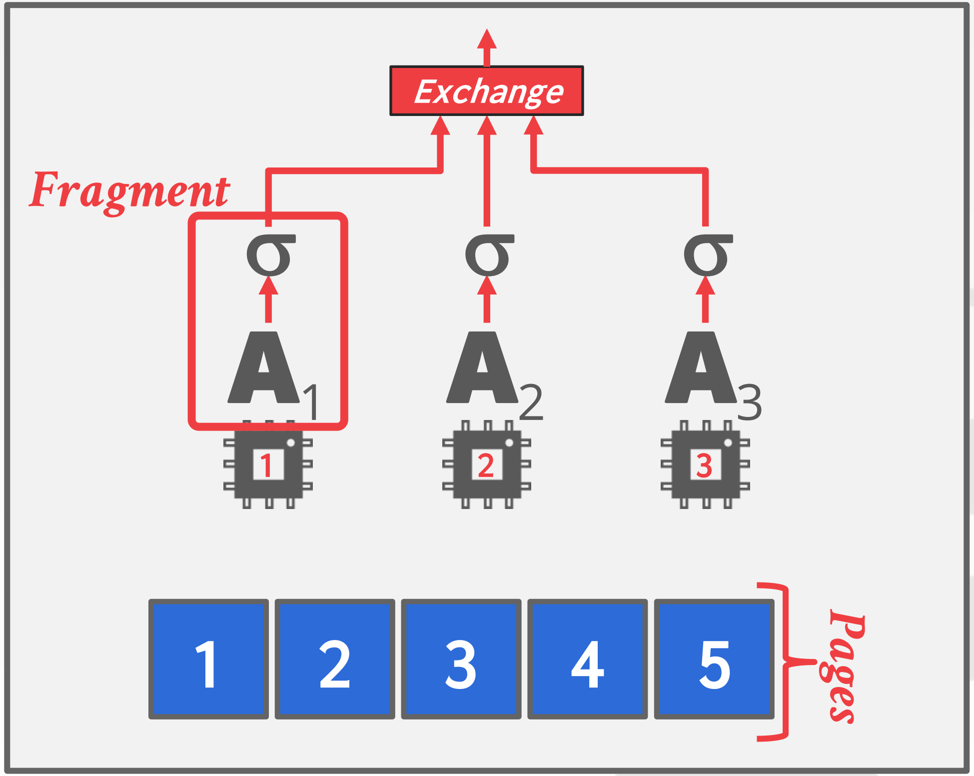
→ Decompose operators into independent **fragments** that perform the same function on different subsets of data.

The DBMS inserts an **exchange** operator into the query plan to coalesce/split results from multiple children/parent operators.

INTRA-OPERATOR PARALLELISM

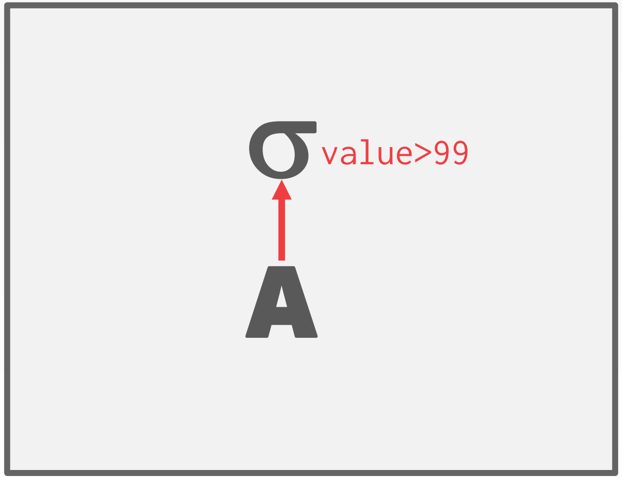
```
SELECT * FROM A
 WHERE A.value > 99
```

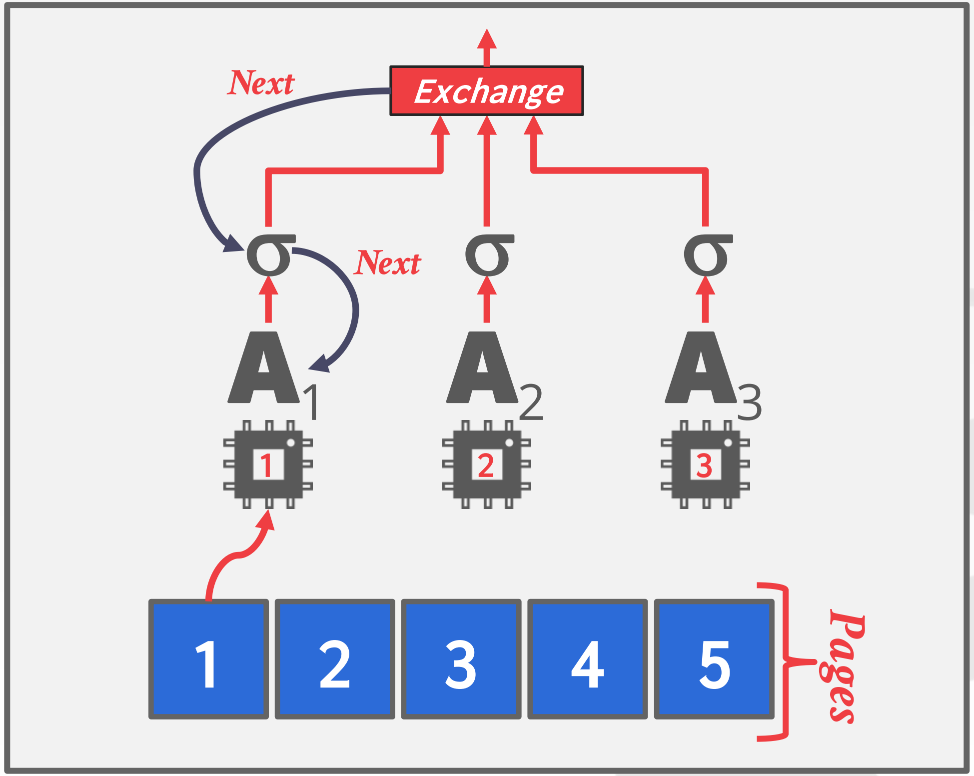




INTRA-OPERATOR PARALLELISM

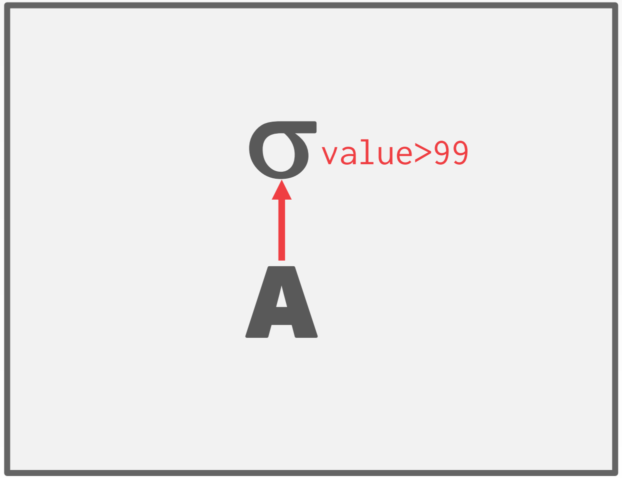
```
SELECT * FROM A
 WHERE A.value > 99
```

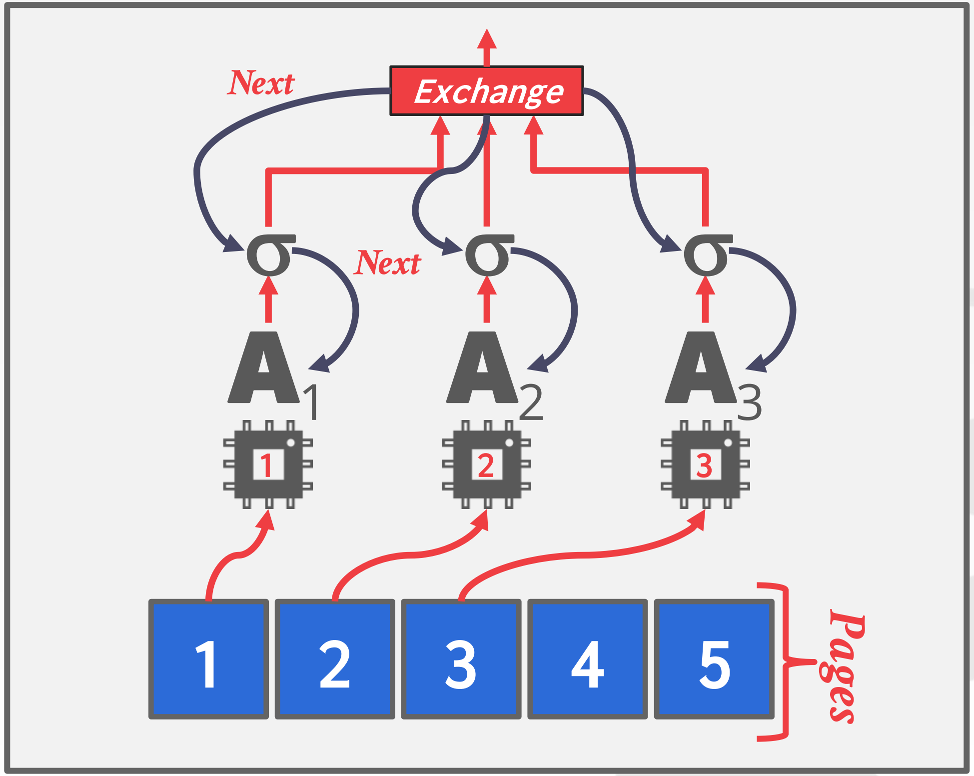




INTRA-OPERATOR PARALLELISM

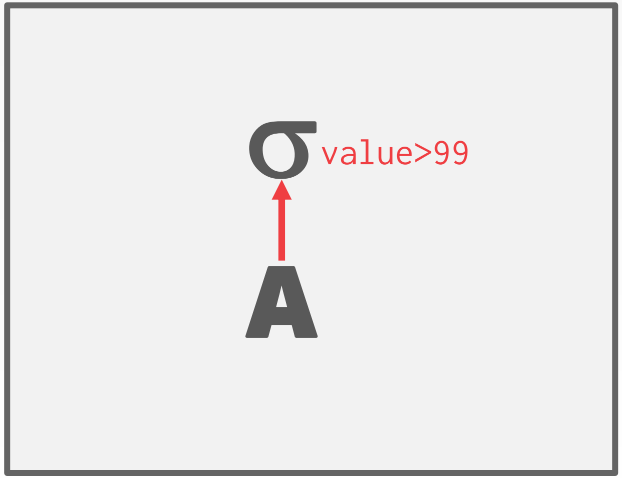
```
SELECT * FROM A
 WHERE A.value > 99
```

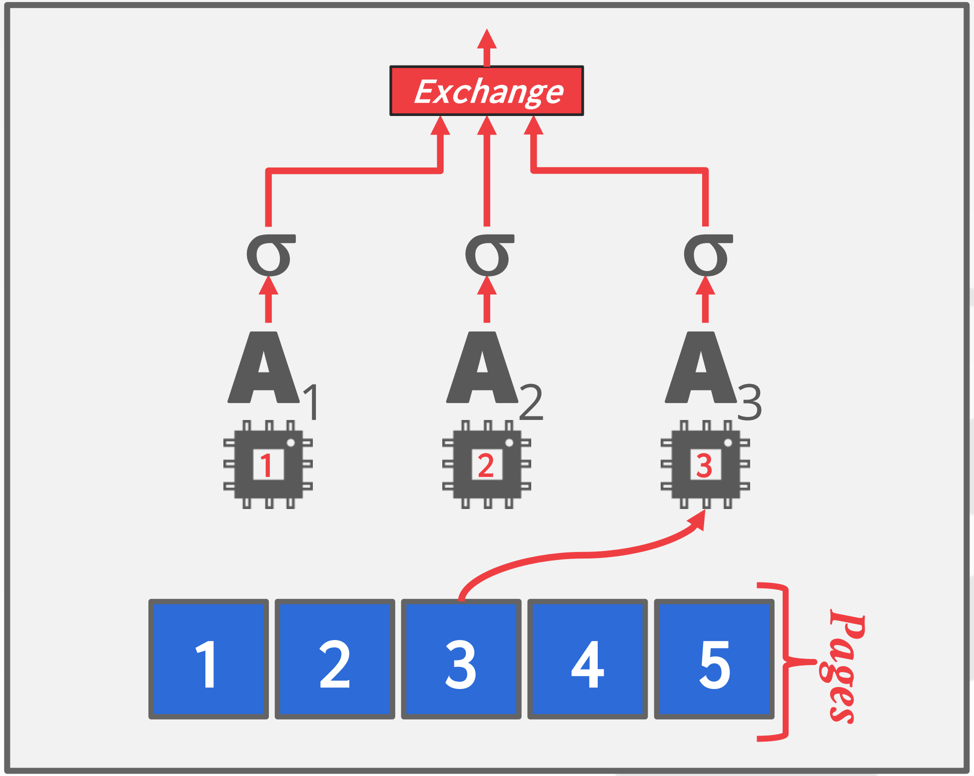




INTRA-OPERATOR PARALLELISM

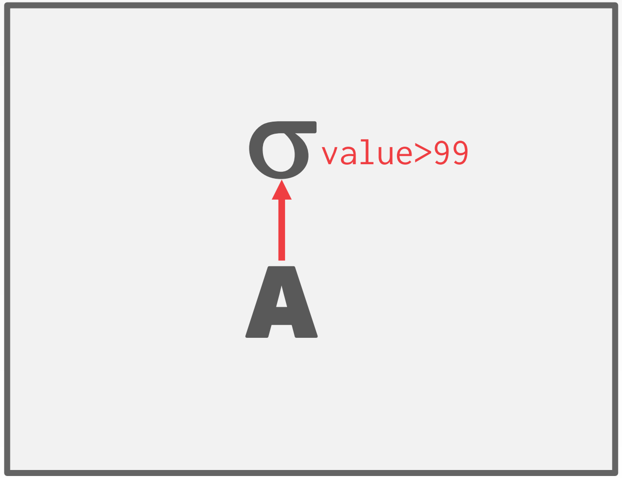
```
SELECT * FROM A
 WHERE A.value > 99
```

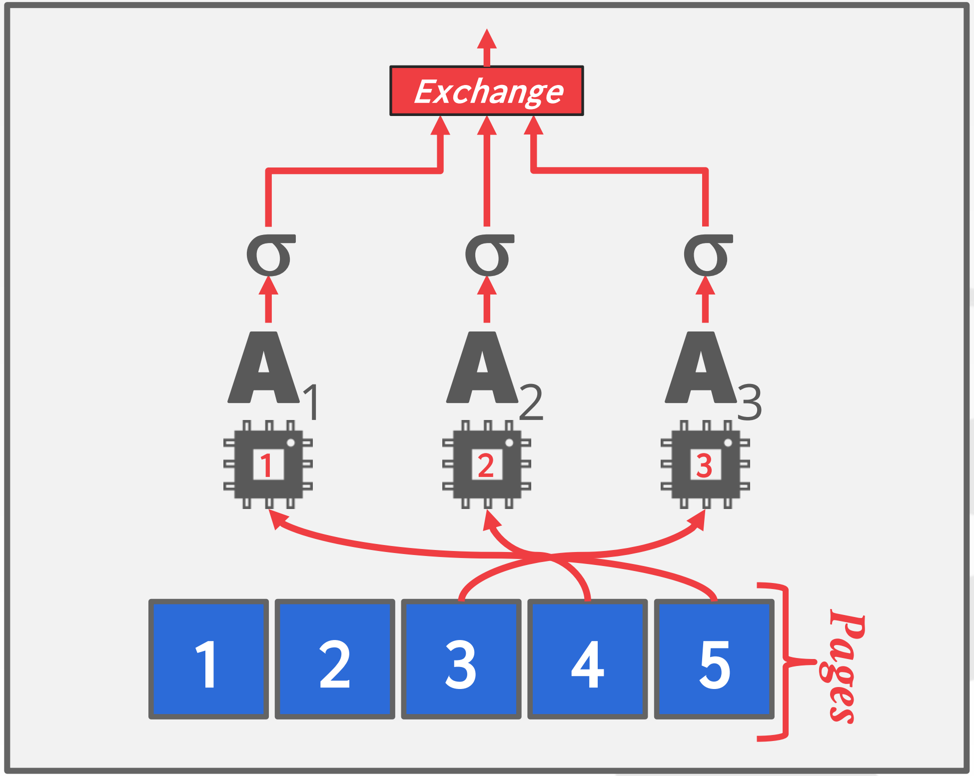




INTRA-OPERATOR PARALLELISM

```
SELECT * FROM A
 WHERE A.value > 99
```

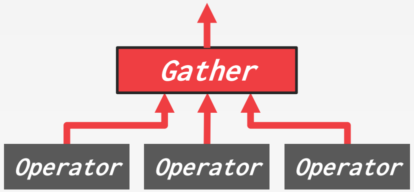




EXCHANGE OPERATOR

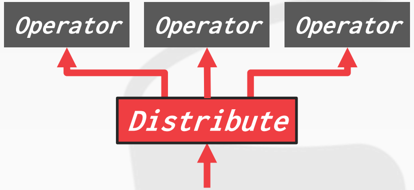
Exchange Type #1 – Gather

→ Combine the results from multiple workers into a single output stream.



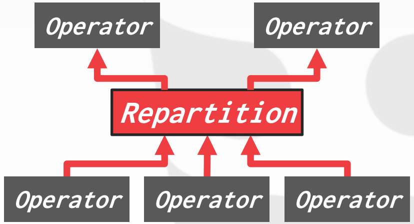
Exchange Type #2 – Distribute

→ Split a single input stream into multiple output streams.



Exchange Type #3 – Repartition

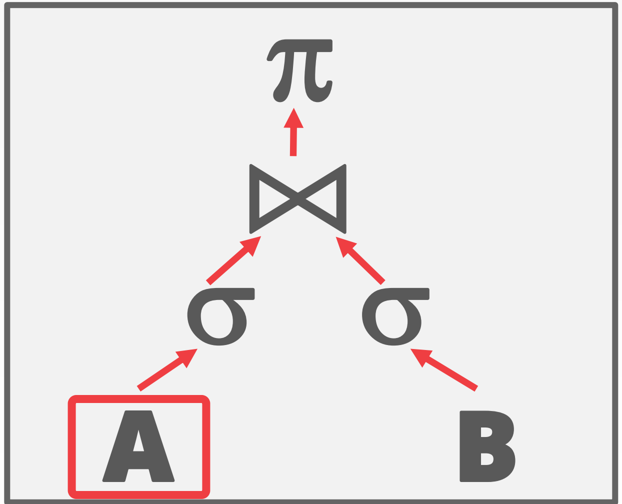
→ Shuffle multiple input streams across multiple output streams.

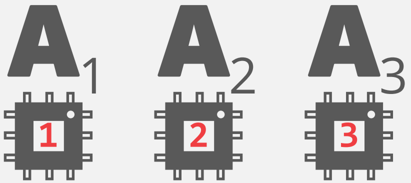


Source: Craig Freedman

INTRA-OPERATOR PARALLELISM

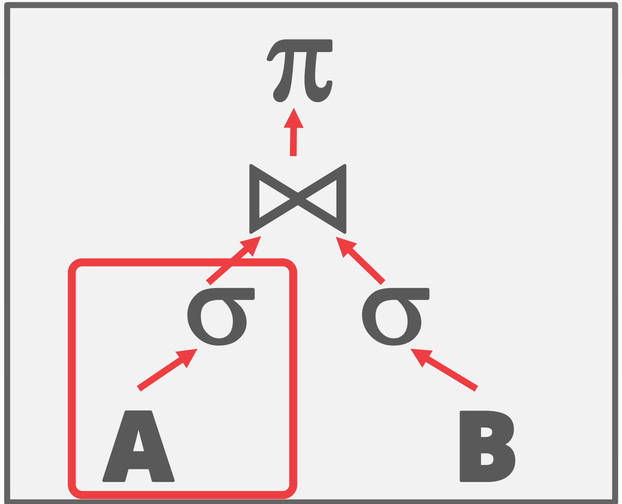
```
SELECT A.id, B.value
  FROM A JOIN B
    ON A.id = B.id
 WHERE A.value < 99
   AND B.value > 100
```

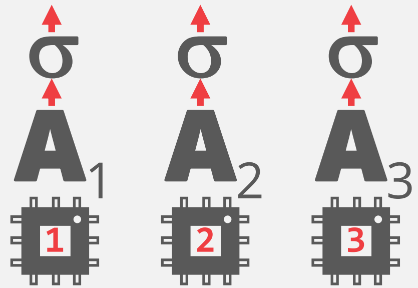




INTRA-OPERATOR PARALLELISM

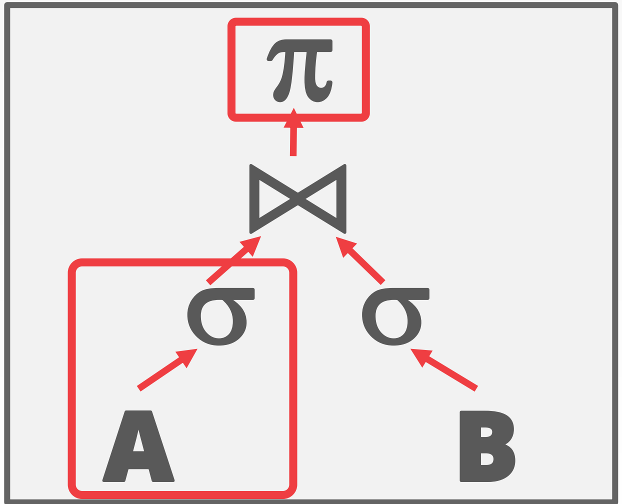
```
SELECT A.id, B.value
  FROM A JOIN B
    ON A.id = B.id
 WHERE A.value < 99
   AND B.value > 100
```

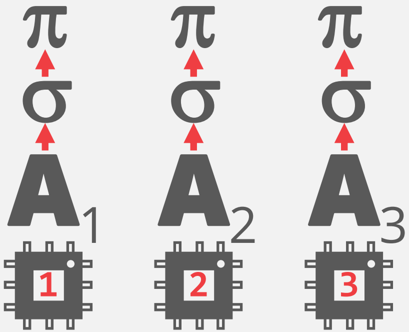




INTRA-OPERATOR PARALLELISM

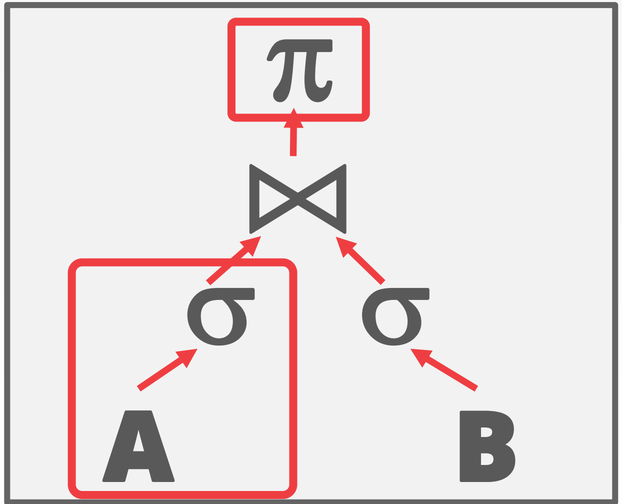
```
SELECT A.id, B.value
  FROM A JOIN B
    ON A.id = B.id
 WHERE A.value < 99
   AND B.value > 100
```

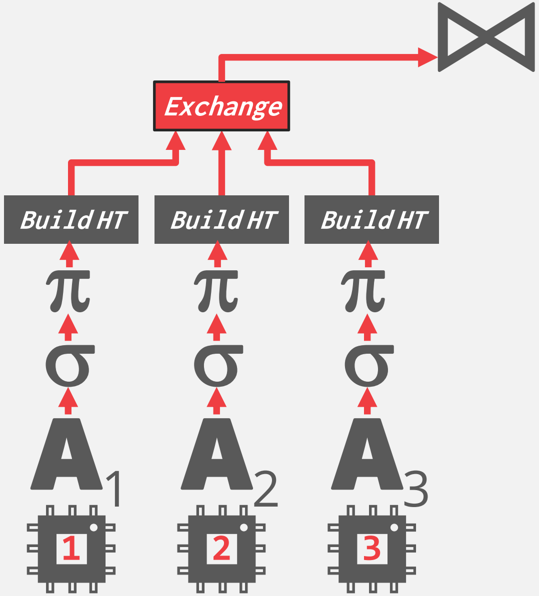




INTRA-OPERATOR PARALLELISM

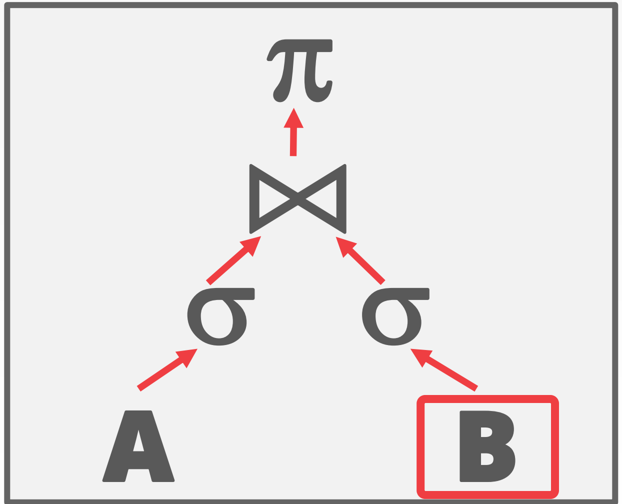
```
SELECT A.id, B.value
  FROM A JOIN B
    ON A.id = B.id
 WHERE A.value < 99
   AND B.value > 100
```

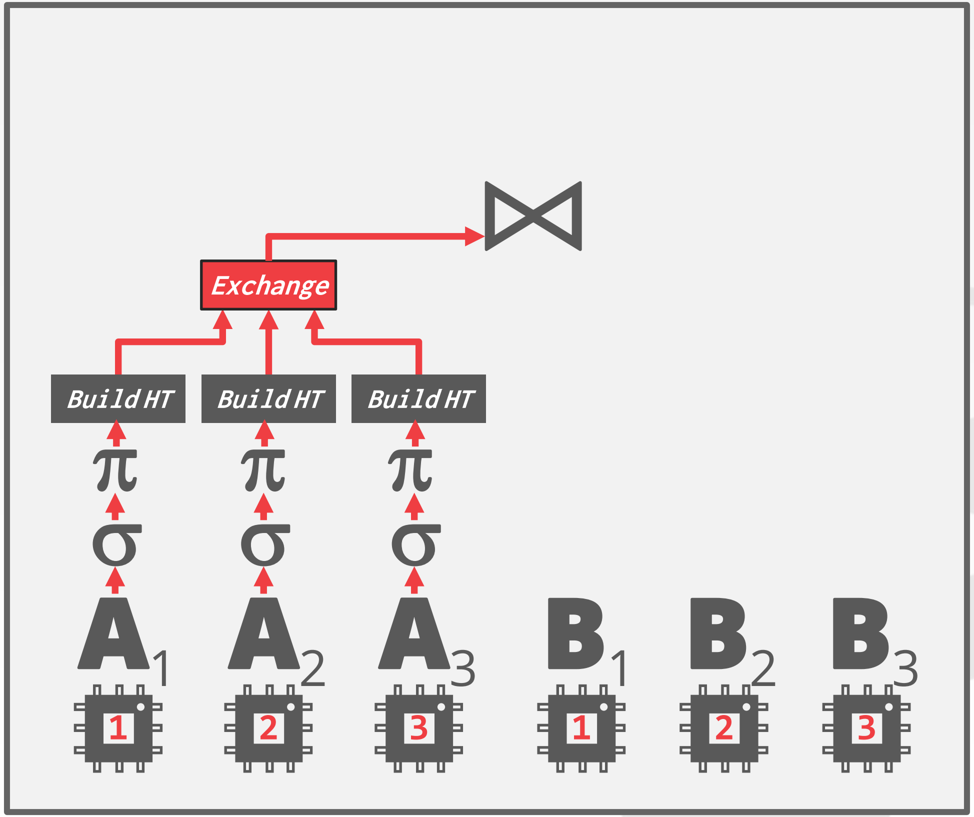




INTRA-OPERATOR PARALLELISM

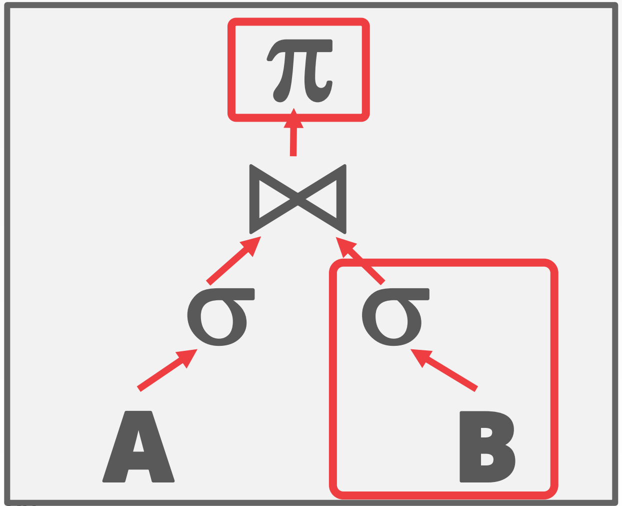
```
SELECT A.id, B.value
  FROM A JOIN B
    ON A.id = B.id
 WHERE A.value < 99
   AND B.value > 100
```

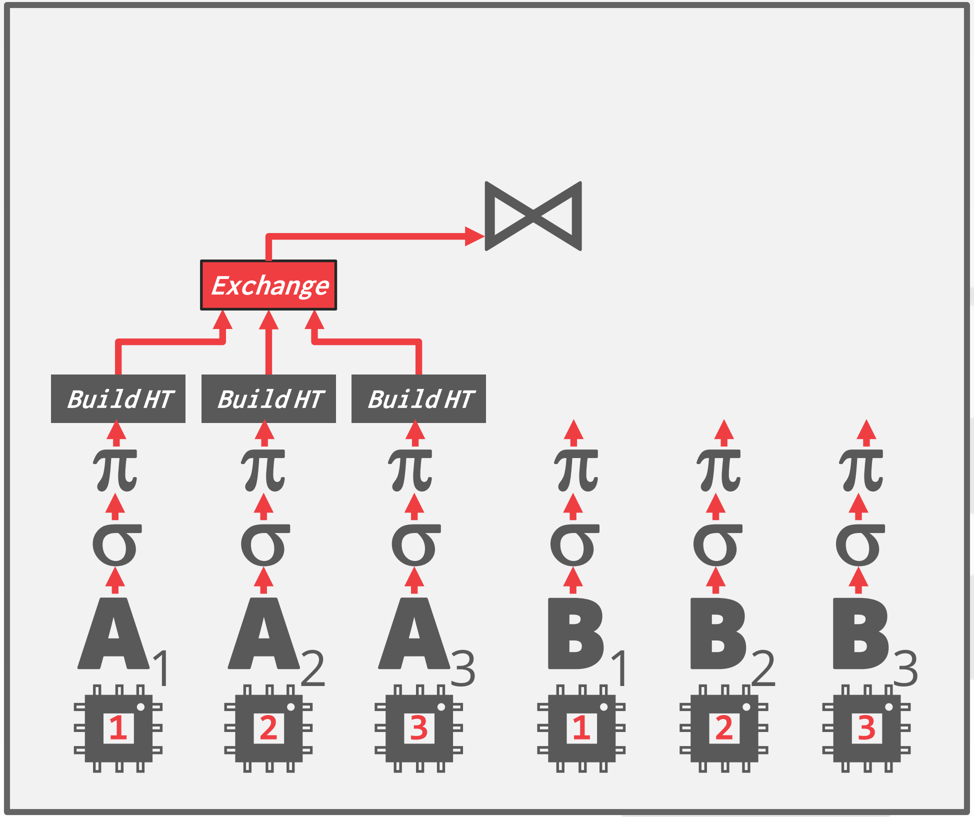




INTRA-OPERATOR PARALLELISM

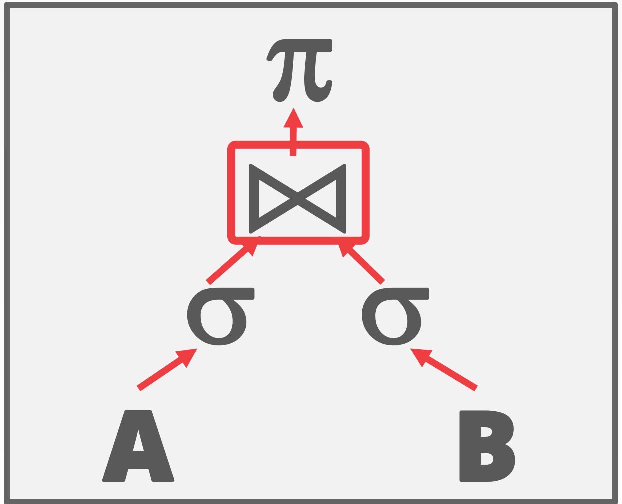
```
SELECT A.id, B.value
  FROM A JOIN B
    ON A.id = B.id
 WHERE A.value < 99
   AND B.value > 100
```

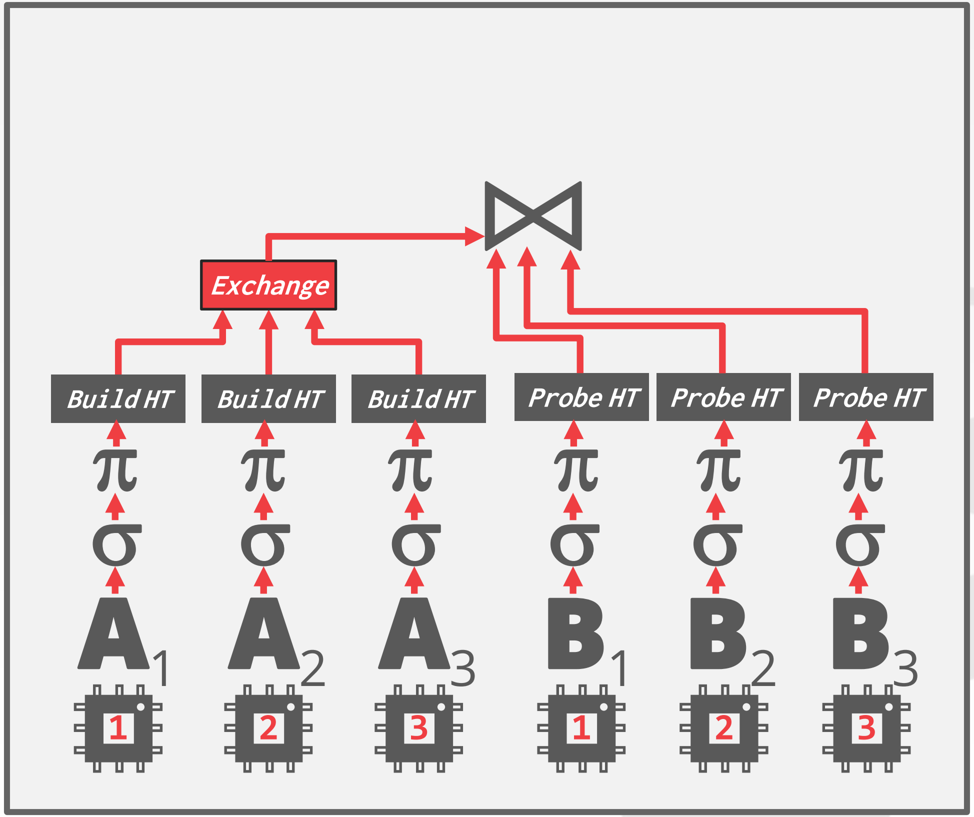




INTRA-OPERATOR PARALLELISM

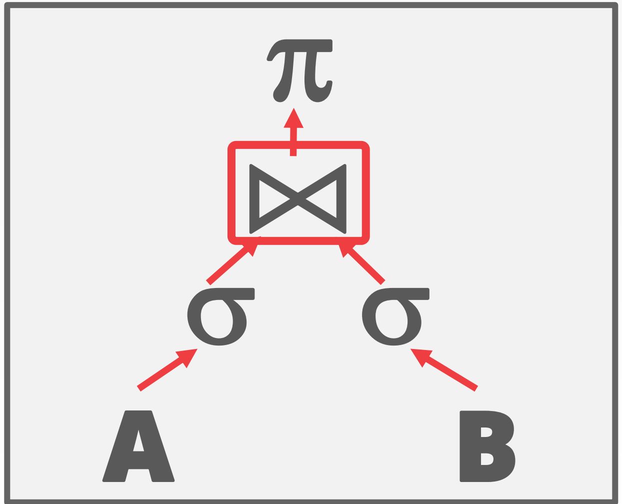
```
SELECT A.id, B.value
  FROM A JOIN B
    ON A.id = B.id
 WHERE A.value < 99
   AND B.value > 100
```

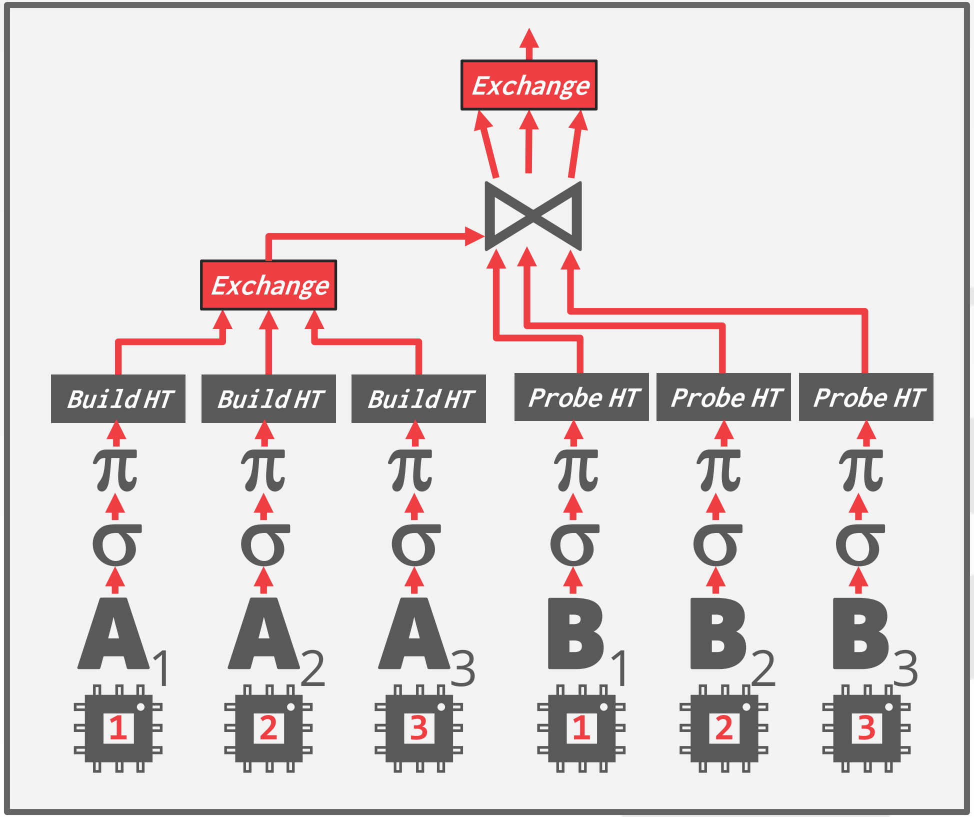




INTRA-OPERATOR PARALLELISM

```
SELECT A.id, B.value
  FROM A JOIN B
    ON A.id = B.id
 WHERE A.value < 99
   AND B.value > 100
```





INTER-OPERATOR PARALLELISM

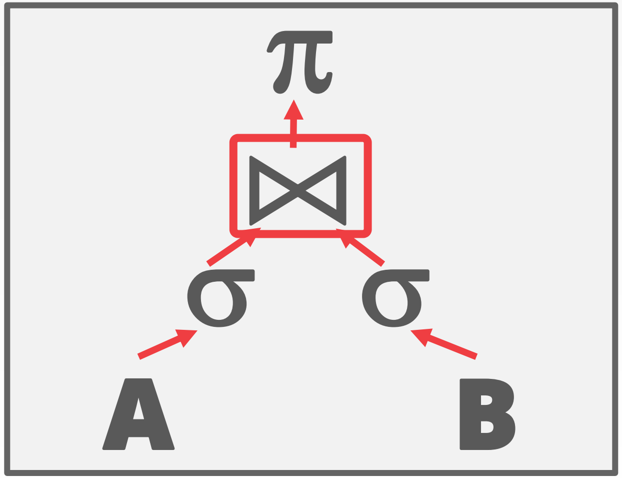
Approach #2: Inter-Operator (Vertical)

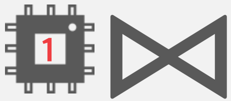
→ Operations are overlapped in order to pipeline data from one stage to the next without materialization.
→ Workers execute multiple operators from different segments of a query plan at the same time.
→ Still need exchange operators to combine intermediate results from segments.

Also called **<u>pipelined parallelism</u>**.

INTER-OPERATOR PARALLELISM

```
SELECT A.id, B.value
  FROM A JOIN B
    ON A.id = B.id
 WHERE A.value < 99
   AND B.value > 100
```

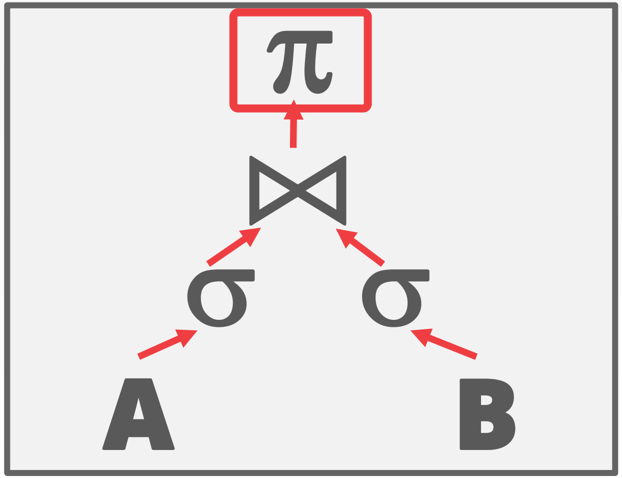


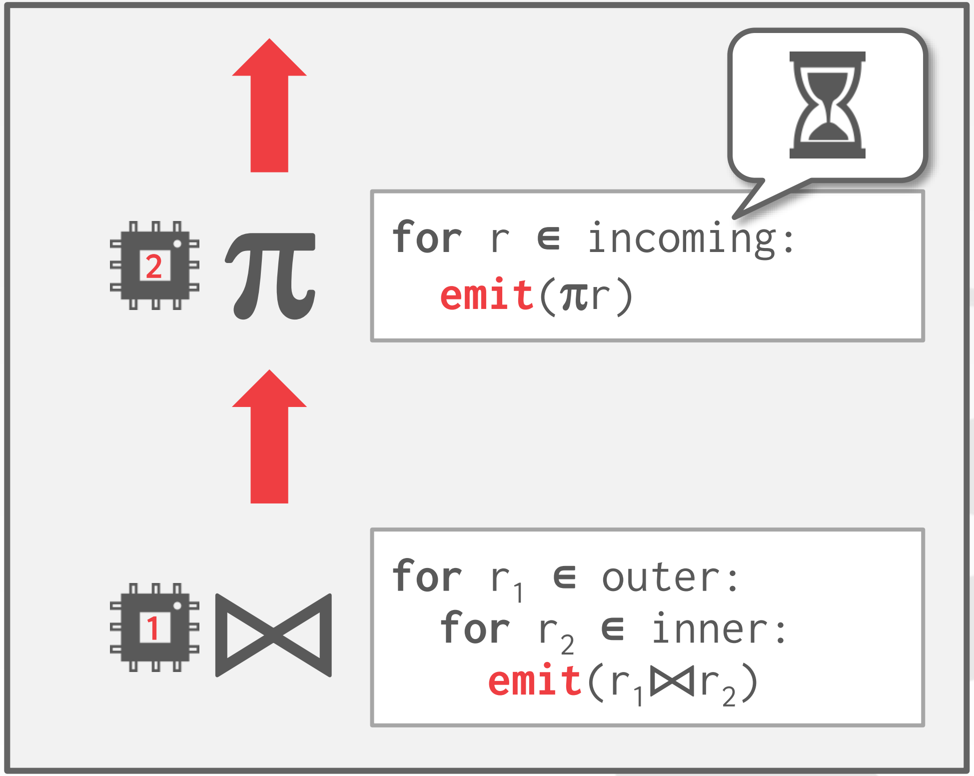


for r_1 ∈ outer:
  for r_2 ∈ inner:
    emit(r_1⋈r_2)

INTER-OPERATOR PARALLELISM

```
SELECT A.id, B.value
  FROM A JOIN B
    ON A.id = B.id
 WHERE A.value < 99
   AND B.value > 100
```





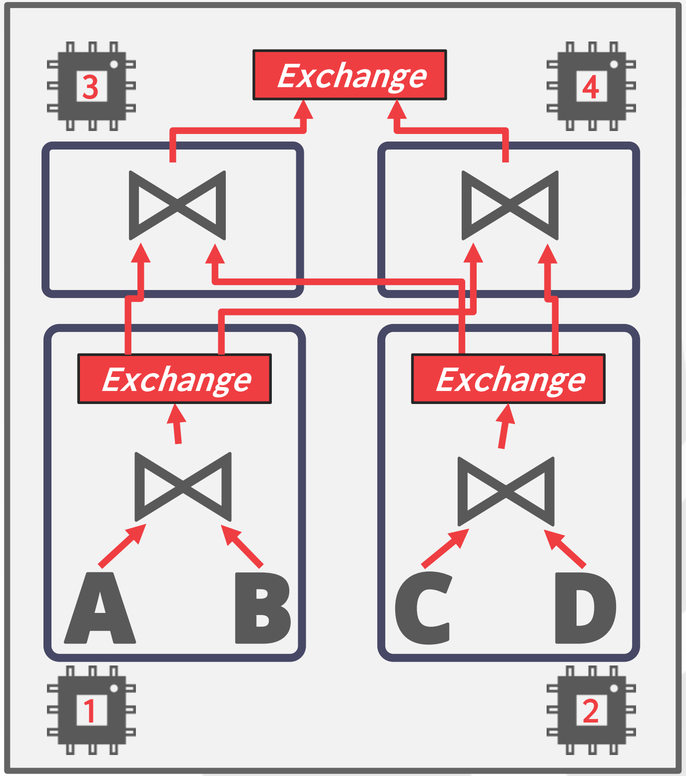
BUSHY PARALLELISM

Approach #3: Bushy Parallelism

→ Extension of inter-operator parallelism where workers execute multiple operators from different segments of a query plan at the same time.

→ Still need exchange operators to combine intermediate results from segments.

```
SELECT *
  FROM A JOIN B JOIN C JOIN D
```



OBSERVATION

Using additional processes/threads to execute queries in parallel won't help if the disk is always the main bottleneck.

→ Can make things worse if each worker is reading different segments of disk.

I/O PARALLELISM

Split the DBMS installation across multiple storage devices.

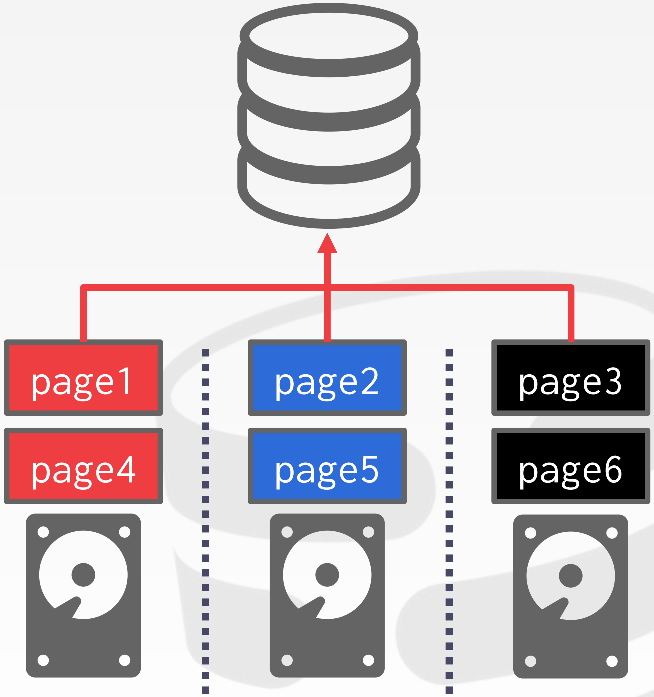
→ Multiple Disks per Database
→ One Database per Disk
→ One Relation per Disk
→ Split Relation across Multiple Disks

MULTI-DISK PARALLELISM

Configure OS/hardware to store the DBMS's files across multiple storage devices.

→ Storage Appliances
→ RAID Configuration

This is transparent to the DBMS.

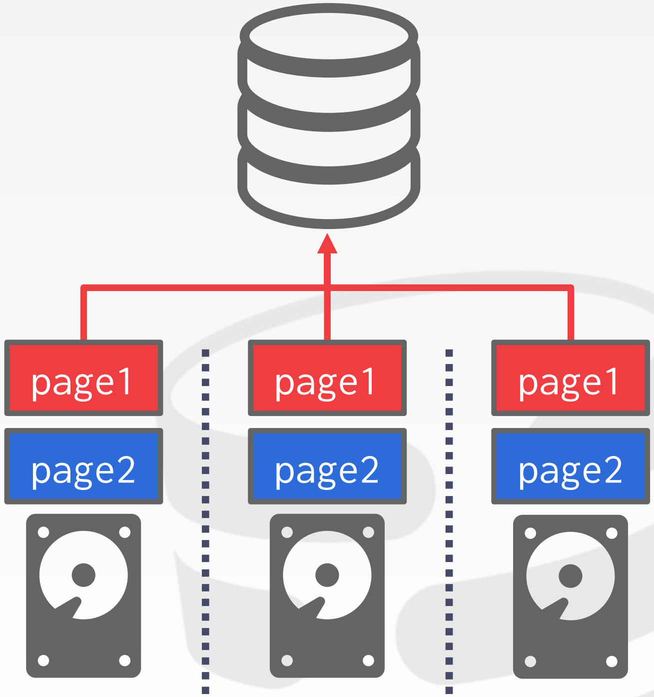


RAID 0 (Stripping)

MULTI-DISK PARALLELISM

Configure OS/hardware to store the DBMS's files across multiple storage devices.
→ Storage Appliances
→ RAID Configuration

This is transparent to the DBMS.



RAID 1 (Mirroring)

DATABASE PARTITIONING

Some DBMSs allow you specify the disk location of each individual database.

→ The buffer pool manager maps a page to a disk location.

This is also easy to do at the filesystem level if the DBMS stores each database in a separate directory.

→ The DBMS recovery log file might still be shared if transactions can update multiple databases.

PARTITIONING

Split single logical table into disjoint physical segments that are stored/managed separately.

Ideally partitioning is transparent to the application.

→ The application accesses logical tables and does not care how things are stored.

VERTICAL PARTITIONING

Store a table's attributes in a separate location (e.g., file, disk volume).

Must store tuple information to reconstruct the original record.

```
CREATE TABLE foo (
  attr1 INT,
  attr2 INT,
  attr3 INT,
  attr4 TEXT
);
```

Tuple#1	attr1	attr2	attr3	attr4
Tuple#2	attr1	attr2	attr3	attr4
Tuple#3	attr1	attr2	attr3	attr4
Tuple#4	attr1	attr2	attr3	attr4

VERTICAL PARTITIONING

Store a table's attributes in a separate location (e.g., file, disk volume).

Must store tuple information to reconstruct the original record.

```
CREATE TABLE foo (
  attr1 INT,
  attr2 INT,
  attr3 INT,
  attr4 TEXT
);
```

Partition #1

Tuple#1	attr1	attr2	attr3
Tuple#2	attr1	attr2	attr3
Tuple#3	attr1	attr2	attr3
Tuple#4	attr1	attr2	attr3

Partition #2

Tuple#1	attr4
Tuple#2	attr4
Tuple#3	attr4
Tuple#4	attr4

HORIZONTAL PARTITIONING

Divide the tuples of a table up into disjoint segments based on some partitioning key.

→ Hash Partitioning
→ Range Partitioning
→ Predicate Partitioning

```
CREATE TABLE foo (
  attr1 INT,
  attr2 INT,
  attr3 INT,
  attr4 TEXT
);
```

Tuple#1	attr1	attr2	attr3	attr4
Tuple#2	attr1	attr2	attr3	attr4
Tuple#3	attr1	attr2	attr3	attr4
Tuple#4	attr1	attr2	attr3	attr4

HORIZONTAL PARTITIONING

Divide the tuples of a table up into disjoint segments based on some partitioning key.

→ Hash Partitioning
→ Range Partitioning
→ Predicate Partitioning

```
CREATE TABLE foo (
  attr1 INT,
  attr2 INT,
  attr3 INT,
  attr4 TEXT
);
```

Partition #1

Tuple#1	attr1	attr2	attr3	attr4
Tuple#2	attr1	attr2	attr3	attr4

Partition #2

Tuple#3	attr1	attr2	attr3	attr4
Tuple#4	attr1	attr2	attr3	attr4

CONCLUSION

Parallel execution is important.

(Almost) every DBMS support this.

This is hard to get right.

→ Coordination Overhead
→ Scheduling
→ Concurrency Issues
→ Resource Contention

MIDTERM EXAM

Who: You

What: Midterm Exam

Where: Gradescope

When: Wed Oct 21st (Two Sessions)

Why: https://youtu.be/EDRsQQ6Onnw

https://15445.courses.cs.cmu.edu/fall2020/midterm-guide.html

MIDTERM EXAM

Two Exam Sessions:

→ Session #1: Wed Oct 21^{st} @ 9:00am ET
→ Session #2: Wed Oct 21^{st} @ 3:20pm ET
→ *I will email you to confirm your session.*

Exam will be available on Gradescope.

Please email Andy if you need special accommodations.

MIDTERM EXAM

Exam covers all lecture material up to today (inclusive).

Open book/notes/calculator.

You are not required to turn on your video during the video.

We will answer clarification questions via OHQ.

RELATIONAL MODEL

Integrity Constraints

Relation Algebra

SQL

Basic operations:
→ SELECT / INSERT / UPDATE / DELETE
→ WHERE predicates
→ Output control

More complex operations:
→ Joins
→ Aggregates
→ Common Table Expressions

STORAGE

Buffer Management Policies
→ LRU / MRU / CLOCK

On-Disk File Organization
→ Heaps
→ Linked Lists

Page Layout
→ Slotted Pages
→ Log-Structured

HASHING

Static Hashing
→ Linear Probing
→ Robin Hood
→ Cuckoo Hashing

Dynamic Hashing
→ Extendible Hashing
→ Linear Hashing

TREE INDEXES

B+Tree

→ Insertions / Deletions

→ Splits / Merges

→ Difference with B-Tree

→ Latch Crabbing / Coupling

Radix Trees

SORTING

Two-way External Merge Sort

General External Merge Sort

Cost to sort different data sets with different number of buffers.

JOINS

Nested Loop Variants

Sort-Merge

Hash

Execution costs under different conditions.

QUERY PROCESSING

Processing Models

→ Advantages / Disadvantages

Parallel Execution

→ Inter- vs. Intra-Operator Parallelism

NEXT CLASS

Query Planning & Optimization