Lecture #06

Carnegie Mellon University

ADVANCED DATABASE SYSTEMS

Query Execution

Andy Pavlo // 15-721 // Spring 2023

LAST CLASS

Last two lectures were about minimize the amount of data that the DBMS processes when executing sequential scans.

We are now going to start discussing ways to improve the DBMS's query execution performance.

SEQUENTIAL SCAN OPTIMIZATIONS

Data Prefetching / Scan Sharing

Task Parallelization / Multi-threading

Clustering / Sorting

Late Materialization

Materialized Views / Result Caching

Data Skipping

Data Parallelization / Vectorization

Code Specialization / Compilation

EXECUTION OPTIMIZATION

DBMS engineering is an orchestration of a bunch of optimizations that seek to make full use of hardware. There is not a single technique that is more important than others.

Andy's Unscientific Top-3 Optimizations:

→ Data Parallelization (Vectorization)
→ Task Parallelization (Multi-threading)
→ Code Specialization (Compilation)

OPTIMIZATION GOALS

Approach #1: Reduce Instruction Count
→ Use fewer instructions to do the same amount of work.

Approach #2: Reduce Cycles per Instruction
→ Execute more CPU instructions in fewer cycles.

Approach #3: Parallelize Execution
→ Use multiple threads to compute each query in parallel.

OPTIMIZATION GOALS

Approach #1
→ Use fewer in

Approach #2
→ Execute mor

Approach #
→ Use multipl

From: Linus Torvalds <torvalds@linux-foundation.org>
To: Arnd Bergmann <arnd@kernel.org>, "Jason A. Donenfeld" <Jason@zx2c4.com>
CC: Linux Kernel Mailing List <linux-kernel@vger.kernel.org>
Subject: Re: [PATCH v2 07/13] asm-generic: unaligned always use struct helpers
Date: Tue, 18 May 2021 06:12:03 -1000 [thread overview]
Message-ID: <CAHk-=wjuoGyxDhAF8SsrTkN0-YfCx7E6jUN3ikC_tn2AKWTTsA@mail.gmail.com> (raw)
In-Reply-To: <CAK8P3a3hbts4k+rrfnE8Z78ezCaME0UVgwqkdLW5NOps2YHUQQ@mail.gmail.com>

On Tue, May 18, 2021 at 5:42 AM Arnd Bergmann <arnd@kernel.org> wrote:
>
> To be on the safe side, we could pass -fno-tree-loop-vectorize along
> with -O3 on the affected gcc versions, or use a bigger hammer
> (not use -O3 at all, always set -fno-tree-loop-vectorize, ...).

I personally think -O3 in general is unsafe.

It has historically been horribly buggy. It's gotten better, but this case clearly shows that "gotten better" really isn't that high of a bar.

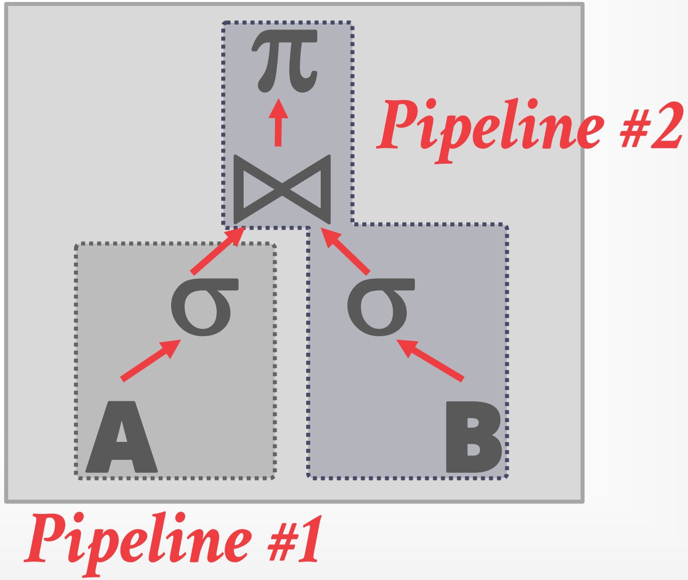
QUERY EXECUTION

A query plan is a DAG of **operators**.

An **operator instance** is an invocation of an operator on a unique segment of data.

A **task** is a sequence of one or more operator instances (also sometimes referred to as a **pipeline**).

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



TODAY'S AGENDA

MonetDB/X100 Analysis
Processing Models
Parallel Execution

MONETDB/X100 (2005)

Low-level analysis of execution bottlenecks for in-memory DBMSs on OLAP workloads.

→ Show how DBMS are designed incorrectly for modern CPU architectures.

Based on these findings, they proposed a new DBMS called MonetDB/X100.

→ Renamed to Vectorwise and acquired by Actian in 2010.

→ Rebranded as Vector and Avalanche.

MONETDB/X100: HYPER-PIPELINING QUERY EXECUTION
CIDR 2005

CPU OVERVIEW

CPUs organize instructions into **pipeline stages**.

The goal is to keep all parts of the processor busy at each cycle by masking delays from instructions that cannot complete in a single cycle.

Super-scalar CPUs support multiple pipelines.

→ Execute multiple instructions in parallel in a single cycle if they are independent (**out-of-order** execution).

Everything is fast until there is a mistake…

DBMS / CPU PROBLEMS

Problem #1: Dependencies

→ If one instruction depends on another instruction, then it cannot be pushed immediately into the same pipeline.

Problem #2: Branch Prediction

→ The CPU tries to predict what branch the program will take and fill in the pipeline with its instructions.

→ If it gets it wrong, it must throw away any speculative work and flush the pipeline.

BRANCH MISPREDICTION

Because of long pipelines, CPUs will speculatively execute branches. This potentially hides the long stalls between dependent instructions.

The most executed branching code in a DBMS is the filter operation during a sequential scan.
But this is (nearly) impossible to predict correctly.

BRAN

Because of lon
execute branc
stalls between

The most exe
the filter oper
But this is (ne

cppreference.com

Create account | Search

Page | Discussion | View | Edit | History

C++ | C++ language | Declarations | Attributes

C++ attribute: likely, unlikely (since C++20)

Allow the compiler to optimize for the case where paths of execution including that statement are more or less likely than any alternative path of execution that does not include such a statement

Syntax

`[[likely]]`	(1)
`[[unlikely]]`	(2)

Explanation

These attributes may be applied to labels and statements (other than declaration-statements). They may not be simultaneously applied to the same label or statement.

1) Applies to a statement to allow the compiler to optimize for the case where paths of execution including that statement are more likely than any alternative path of execution that does not include such a statement.

2) Applies to a statement to allow the compiler to optimize for the case where paths of execution including that statement are less likely than any alternative path of execution that does not include such a statement.

A path of execution is deemed to include a label if and only if it contains a jump to that label:

```
int f(int i) {
    switch(i) {
    case 1: [[fallthrough]];
    [[likely]] case 2: return 1;
    }
    return 2;
}
```

`i == 2` is considered more likely than any other value of `i`, but the `[[likely]]` has no effect on the `i == 1` case even though it falls through the `case 2:` label.

Example

This section is incomplete
Reason: no example

cppreference.com

Create account

Search

View Edit History

ly (since C++20)

ere paths of execution including that statement are more or less likely
es not include such a statement

atements (other than declaration-statements). They may not be
ement.

er to optimize for the case where paths of execution including that
ive path of execution that does not include such a statement.

er to optimize for the case where paths of execution including that
e path of execution that does not include such a statement.

and only if it contains a jump to that label:

lue of i, but the [[likely]] has no effect on the i == 1 case

Don't use the [[likely]] or [[unlikely]] attributes

Posted on 2020-08-27 by Aaron Ballman

C++20 introduced the likelihood attributes `[[likely]]` and `[[unlikely]]` as a way for a programmer to give an optimization hint to their implementation that a given code path is more or less likely to be taken. On its face, this seems like a great set of attributes because you can give hints to the optimizer in a way that is hopefully understood by all implementations and will result in faster performance. What's not to love?

The attribute is specified to appertain to arbitrary statements or labels with the recommended practice "to optimize for the case where paths of execution including it are arbitrarily more likely|unlikely than any alternative path of execution that does not include such an attribute on a statement or label." Pop quiz, what does this code do?

```
if (something) {
  [[likely]];
  [[unlikely]];
  foo(something);
}
```

Sorry, but the answer key for this quiz is currently unavailable. However, one rule you should follow about how to use these attributes is: never allow both attributes to appear in the same path of execution. Lest you think, "but who would write such bad code?", consider this reasonable-looking-but-probably-very-unfortunate code:

SELECTION SCANS

```
SELECT * FROM table
 WHERE key > $(low)
   AND key < $(high)
```

Source: Bogdan Raducanu

SELECTION SCANS

Scalar (Branching)

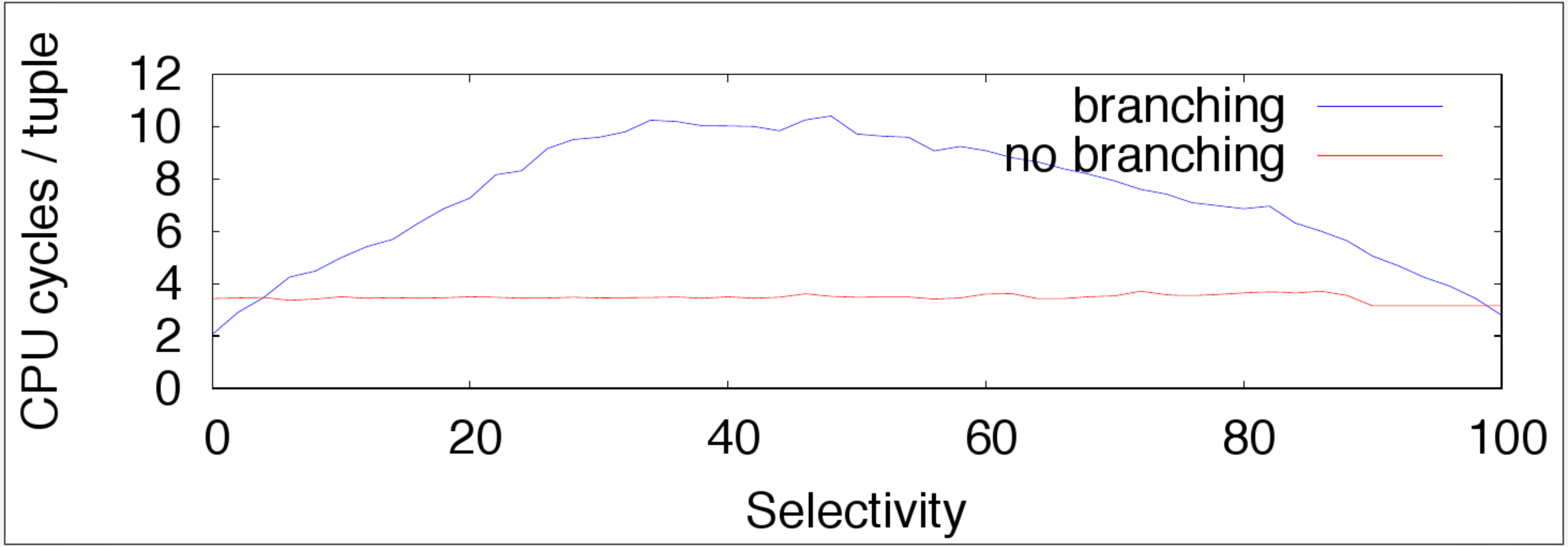
```
i = 0
for t in table:
  key = t.key
  if (key>low) && (key<high):
    copy(t, output[i])
    i = i + 1
```

Scalar (Branchless)

```
i = 0
for t in table:
  copy(t, output[i])
  key = t.key
  delta = (key>low ? 1 : 0) &
          (key<high ? 1 : 0)
  i = i + delta
```

Source: Bogdan Raducanu

SELECTION SCANS



Source: Bogdan Raducanu

EXCESSIVE INSTRUCTIONS

The DBMS needs to support different data types, so it must check a values type before it performs any operation on that value.

→ This is usually implemented as giant switch statements.
→ Also creates more branches that can be difficult for the CPU to predict reliably.

Example: Postgres' addition for **NUMERIC** types.

EXCESSIV

The DBMS needs to
it must check a value
operation on that va

→ This is usually imple
→ Also creates more br
CPU to predict reliab

Example: Postgres' a

```c
/* ----------
 * add_var() -
 *
 *  Full version of add functionality on variable level (handling signs).
 *  result might point to one of the operands too without danger.
 * ----------
 */
int
PGTYPESnumeric_add(numeric *var1, numeric *var2, numeric *result)
{
    /*
     * Decide on the signs of the two variables what to do
     */
    if (var1->sign == NUMERIC_POS)
    {
        if (var2->sign == NUMERIC_POS)
        {
            /*
             * Both are positive result = +(ABS(var1) + ABS(var2))
             */
            if (add_abs(var1, var2, result) != 0)
                return -1;
            result->sign = NUMERIC_POS;
        }
        else
        {
            /*
             * var1 is positive, var2 is negative Must compare absolute values
             */
            switch (cmp_abs(var1, var2))
            {
                case 0:
                    /* ----------
                     * ABS(var1) == ABS(var2)
                     * result = ZERO
                     * ----------
                     */
                    zero_var(result);
                    result->rscale = Max(var1->rscale, var2->rscale);
                    result->dscale = Max(var1->dscale, var2->dscale);
                    break;

                case 1:
                    /* ----------
                     * ABS(var1) > ABS(var2)
                     * result = +(ABS(var1) - ABS(var2))
                     * ----------
                     */
                    if (sub_abs(var1, var2, result) != 0)
                        return -1;
                    result->sign = NUMERIC_POS;
                    break;

                case -1:
                    /* ----------
```

PROCESSING MODEL

A DBMS's **processing model** defines how the system executes a query plan.

→ Different trade-offs for workloads (OLTP vs. OLAP).

Approach #1: Iterator Model

Approach #2: Materialization Model

Approach #3: Vectorized / Batch Model

ITERATOR MODEL

Each query plan operator implements a **next** function.

→ On each invocation, the operator returns either a single tuple or a null marker if there are no more tuples.
→ The operator implements a loop that calls next on its children to retrieve their tuples and then process them.

Also called **Volcano** or **Pipeline** Model.

ITERATOR MODEL

Control Flow →
Data Flow →

Next()

```
for t in child.Next():
  emit(projection(t))
```

Next()

```
for t1 in left.Next():
  buildHashTable(t1)
for t2 in right.Next():
  if probe(t2): emit(t1⋈t2)
```

Next()

```
for t in child.Next():
  if evalPred(t): emit(t)
```

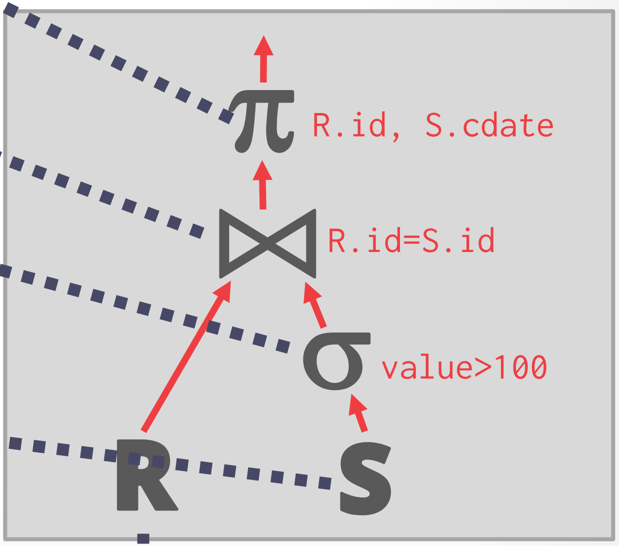
Next()

```
for t in R:
  emit(t)
```

Next()

```
for t in S:
  emit(t)
```

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



ITERATOR MODEL

Control Flow →
Data Flow →

1

```
for t in child.Next():
  emit(projection(t))
```

2

```
for t₁ in left.Next():
  buildHashTable(t₁)
for t₂ in right.Next():
  if probe(t₂): emit(t₁⋈t₂)
```

Single Tuple

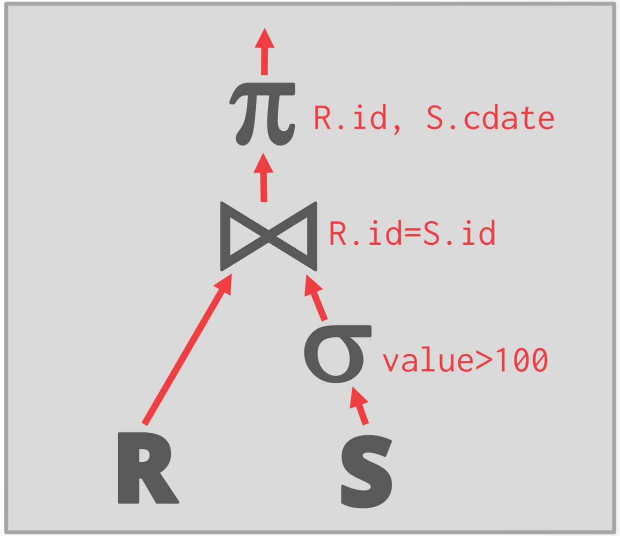
3

```
for t in R:
  emit(t)
```

```
for t in child.Next():
  if evalPred(t): emit(t)
```

```
for t in S:
  emit(t)
```

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



ITERATOR MODEL

Control Flow →
Data Flow →

1
```
for t in child.Next():
  emit(projection(t))
```

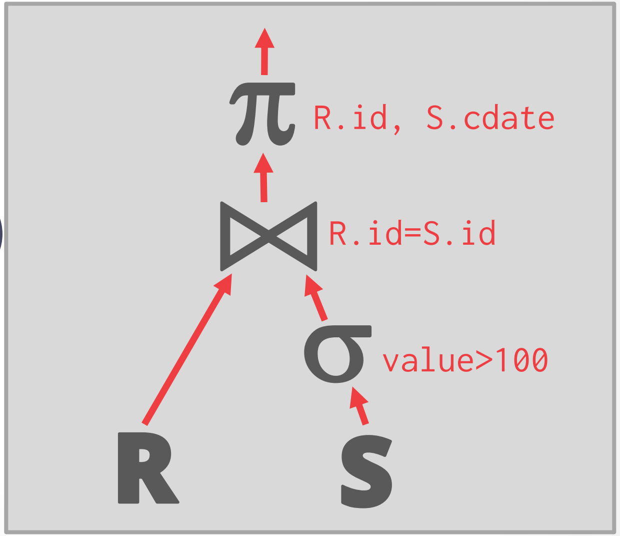
2
```
for t₁ in left.Next():
  buildHashTable(t₁)
for t₂ in right.Next():
  if probe(t₂): emit(t₁⋈t₂)
```

3
```
for t in R:
  emit(t)
```

4
```
for t in child.Next():
  if evalPred(t): emit(t)
```

5
```
for t in S:
  emit(t)
```

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



ITERATOR MODEL

This is used in almost every DBMS. Allows for tuple **pipelining**.

Some operators must block until their children emit all their tuples.
→ Joins, Subqueries, Order By

Output control works easily with this approach.

MATERIALIZATION MODEL

Each operator processes its input all at once and then emits its output all at once.

→ The operator "materializes" its output as a single result.
→ The DBMS can push down hints (e.g., **LIMIT**) to avoid scanning too many tuples.
→ Can send either a materialized row or a single column.

The output can be either whole tuples (NSM) or subsets of columns (DSM).

MATERIALIZATION MODEL

Control Flow →
Data Flow →

1
```
out = [ ]
for t in child.Output():
  out.add(projection(t))
return out
```

2
```
out = [ ]
for t₁ in left.Output():
  buildHashTable(t₁)
for t₂ in right.Output():
  if probe(t₂): out.add(t₁⋈t₂)
return out
```

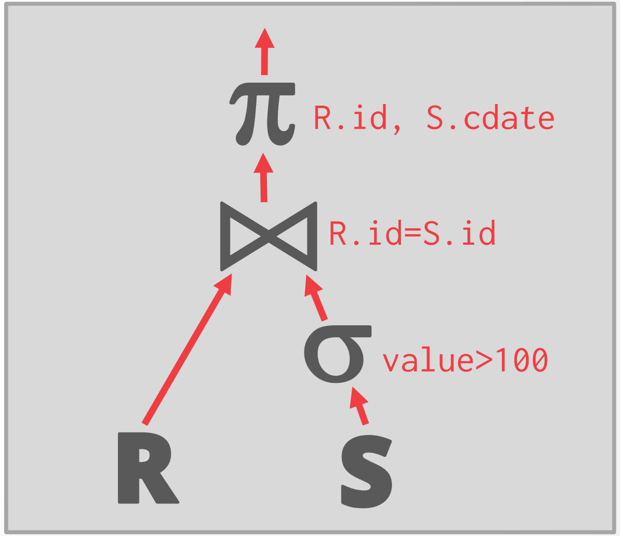
```
out = [ ]
for t in child.Output():
  if evalPred(t): out.add(t)
return out
```

All Tuples

3
```
out = [ ]
for t in R:
  out.add(t)
return out
```

```
out = [ ]
for t in S:
  out.add(t)
return out
```

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



MATERIALIZATION MODEL

Control Flow →
Data Flow →

1
```
out = [ ]
for t in child.Output():
  out.add(projection(t))
return out
```

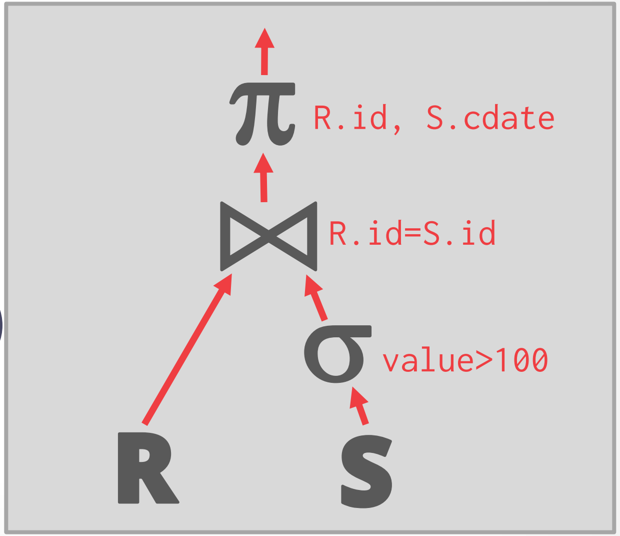
2
```
out = [ ]
for t1 in left.Output():
  buildHashTable(t1)
for t2 in right.Output():
  if probe(t2): out.add(t1⋈t2)
return out
```

3
```
out = [ ]
for t in R:
  out.add(t)
return out
```

4
```
out = [ ]
for t in child.Output():
  if evalPred(t): out.add(t)
return out
```

5
```
out = [ ]
for t in S:
  out.add(t)
return out
```

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



MATERIALIZATION MODEL

Control Flow →
Data Flow →

1

```
out = [ ]
for t in child.Output():
  out.add(projection(t))
return out
```

2

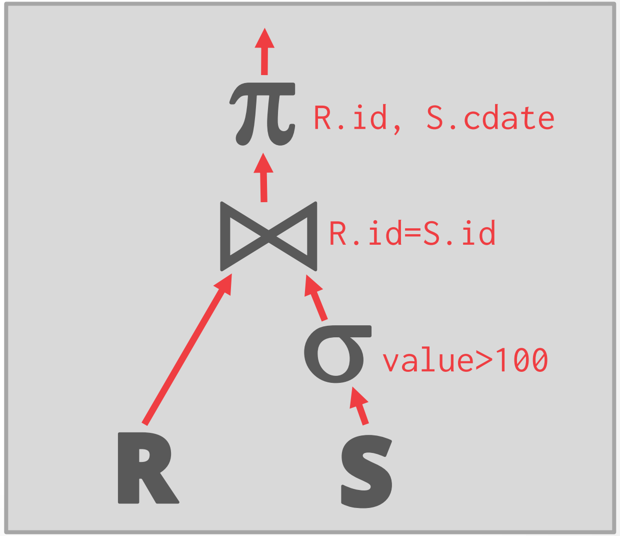
```
out = [ ]
for t₁ in left.Output():
  buildHashTable(t₁)
for t₂ in right.Output():
  if probe(t₂): out.add(t₁⋈t₂)
return out
```

```
out = [ ]
for t in S:
  if evalPred(t): out.add(t)
return out
```

3

```
out = [ ]
for t in R:
  out.add(t)
return out
```

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



MATERIALIZATION MODEL

Better for OLTP workloads because queries only access a small number of tuples at a time.

→ Lower execution / coordination overhead.
→ Fewer function calls.

Not good for OLAP queries with large intermediate results.

VECTORIZATION MODEL

Like the Iterator Model where each operator implements a **next** function, but…

Each operator emits a **batch** of tuples instead of a single tuple.

→ The operator's internal loop processes multiple tuples at a time.

→ The size of the batch can vary based on hardware or query properties.

VECTORIZATION MODEL

Control Flow →
Data Flow →

```
out = [ ]
for t in child.Next():
  out.add(projection(t))
  if |out|>n: emit(out)
```

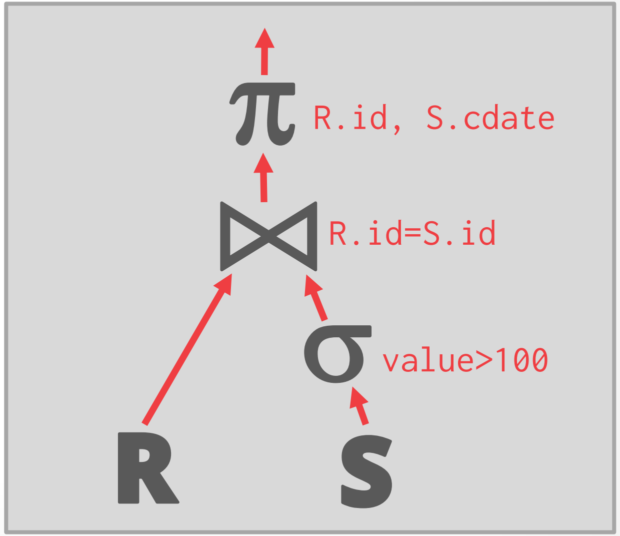
```
out = [ ]
for t1 in left.Next():
  buildHashTable(t1)
for t2 in right.Next():
  if probe(t2): out.add(t1⋈t2)
  if |out|>n: emit(out)
```

```
out = [ ]
for t in child.Next():
  if evalPred(t): out.add(t)
  if |out|>n: emit(out)
```

```
out = [ ]
for t in R:
  out.add(t)
  if |out|>n: emit(out)
```

```
out = [ ]
for t in S:
  out.add(t)
  if |out|>n: emit(out)
```

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



VECTORIZATION MODEL

Control Flow →
Data Flow →

```
out = [ ]
for t in child.Next():
  out.add(projection(t))
  if |out|>n: emit(out)
```

1

```
out = [ ]
for t1 in left.Next():
  buildHashTable(t1)
for t2 in right.Next():
  if probe(t2): out.add(t1⋈t2)
  if |out|>n: emit(out)
```

2

```
out = [ ]
for t in child.Next():
  if evalPred(t): out.add(t)
  if |out|>n: emit(out)
```

4

```
out = [ ]
for t in R:
  out.add(t)
  if |out|>n: emit(out)
```

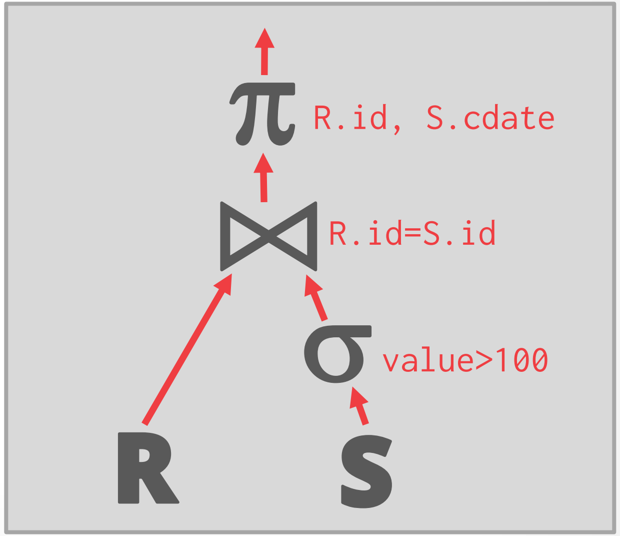
3

Tuple Batch

```
out = [ ]
for t in S:
  out.add(t)
  if |out|>n: emit(out)
```

5

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



VECTORIZATION MODEL

Ideal for OLAP queries because it greatly reduces the number of invocations per operator.

Allows for operators to more easily use vectorized (SIMD) instructions to process batches of tuples.

OBSERVATION

In the previous examples, the DBMS was starting at the root of the query plan and pulling data up from leaf operators.

This is the how most DBMSs implement their execution engine.

PLAN PROCESSING DIRECTION

Approach #1: Top-to-Bottom (Pull)

→ Start with the root and "pull" data up from its children.
→ Tuples are always passed with function calls.

Approach #2: Bottom-to-Top (Push)

→ Start with leaf nodes and "push" data to their parents.
→ We will see this later in HyPer and Peloton ROF.

HyPer
DuckDB
snowflake

PLAN PROCESSING DIRECTION

Approach #1: To

→ Start with the ro

→ Tuples are alway

Approach #2: Bo

→ Start with leaf no

→ We will see this l

er

DB

ake

Move to push-based execution model #1583

Closed **Mytherin** opened this issue on Apr 8, 2021 · 2 comments

Mytherin commented on Apr 8, 2021 • edited Collaborator ···

Currently our execution model operates in a pull-based volcano-like fashion. That means that an operator exposes a `GetChunk` function that fetches a result chunk from the operator. The operator will, in turn, fetch result chunks from its children using this same interface until it reaches a source node (e.g. a base table scan or a parquet file) which can actually emit files, after which execution resumes.

A simple example of such an operator is the projection:

```
void PhysicalProjection::GetChunkInternal(ExecutionContext &context, DataChunk &chunk, PhysicalOperatorState *s
        auto state = reinterpret_cast<PhysicalProjectionState *>(state_p);

        // get the next chunk from the child
        children[0]->GetChunk(context, state->child_chunk, state->child_state.get());
        if (state->child_chunk.size() == 0) {
                return;
        }

        state->executor.Execute(state->child_chunk, chunk);
}
```

This works semi-elegantly and has generally served us well. However, now that we have introduced pipeline parallelism the model is beginning to show cracks. In the pipeline parallelism model, we no longer want to have the behavior of "pulling from the root node". Instead, we want to execute pipelines separately.

The way this is done right now is a semi-hacky solution on top of this model. If we have a pipeline (e.g. a hash table build), we pull from the child node of that hash table using `GetChunk`, and then call `Sink` with the result of this. Partitioning is done by writing partition information to the thread-local `ExecutionContext` object, and using that in the source node to determine the desired partitioning. For example, here is how this is done in the TableScan:

```
// table scan
auto &task = context.task;
// check if there is any parallel state to fetch
state.parallel_state = nullptr;
auto task_info = task.task_info.find(this);
```

PUSH-BASED ITERATOR MODEL

Control Flow

Data Flow

1

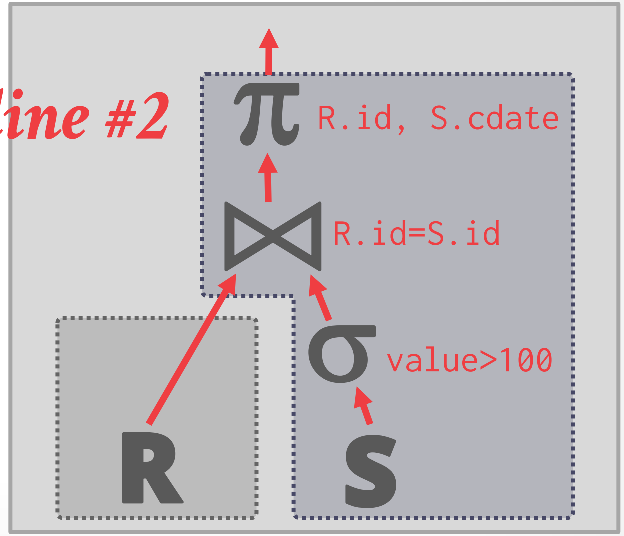
```
for t1 in R:
  buildHashTable(t1)
```

2

```
for t2 in S:
  if evalPred(t):
    if probeHashTable(t2):
      emit(projection(t1⋈t2))
```

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

Pipeline #2



Pipeline #1

PLAN PROCESSING DIRECTION

Approach #1: Top-to-Bottom (Pull)

→ Easy to control output via **LIMIT**.
→ Parent operator blocks until its child returns with a tuple.
→ Additional overhead because operators' **next** functions are implemented as virtual functions.
→ Branching costs on each **next** invocation.

Approach #2: Bottom-to-Top (Push)

→ Allows for tighter control of caches/registers in pipelines.
→ Difficult to control output via LIMIT.
→ Difficult to implement Sort-Merge Join.

PUSH VS. PULL-BASED LOOP FUSION IN QUERY ENGINES
ARXIV 2016

TODAY'S AGENDA

~~MonetDB/X100 Analysis~~

~~Processing Models~~

Parallel Execution

PARALLEL EXECUTION

The DBMS executes multiple tasks simultaneously to improve hardware utilization.

→ Active tasks do <u>not</u> need to belong to the same query.

Approach #1: Inter-Query Parallelism

Approach #2: Intra-Query Parallelism

INTER-QUERY PARALLELISM

Improve overall performance by allowing multiple queries to execute simultaneously.

→ Most DBMSs use a simple first-come, first-served policy.

OLAP queries have parallelizable and non-parallelizable phases. The goal is to always keep all cores active.

We will discuss scheduling queries and multiplexing tasks on cores in the next lecture.

INTRA-QUERY PARALLELISM

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

Approach #1: Intra-Operator (Horizontal)
Approach #2: Inter-Operator (Vertical)

These techniques are not mutually exclusive.

There are parallel algorithms for every relational operator.

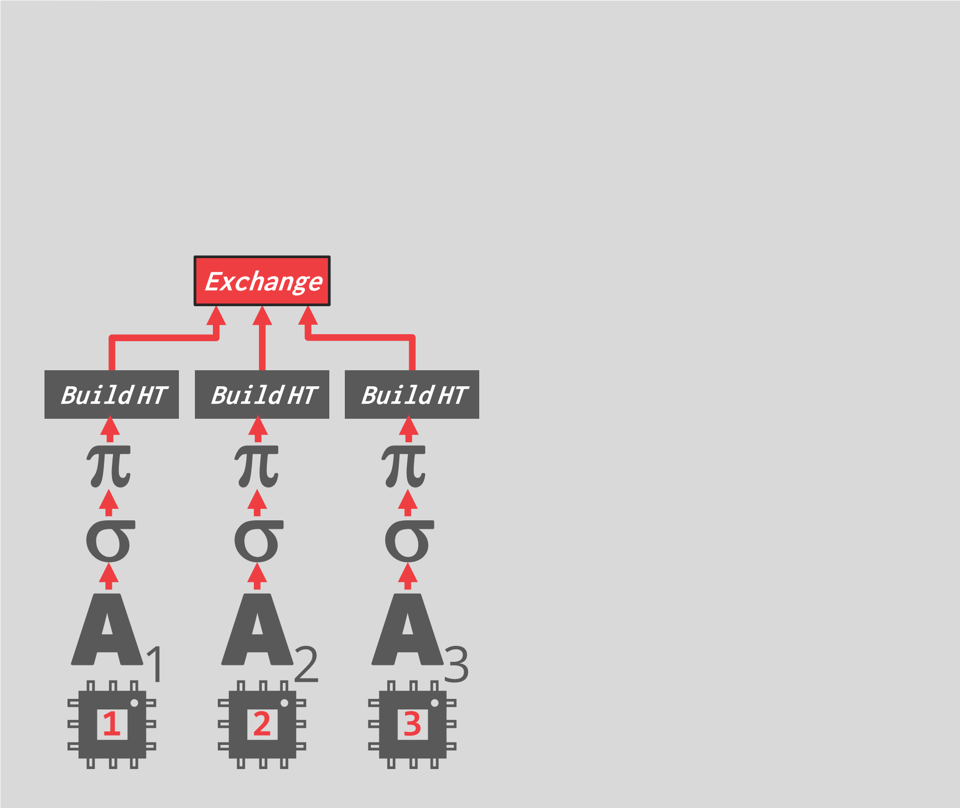
INTRA-OPERATOR PARALLELISM

Approach #1: Intra-Operator (Horizontal)

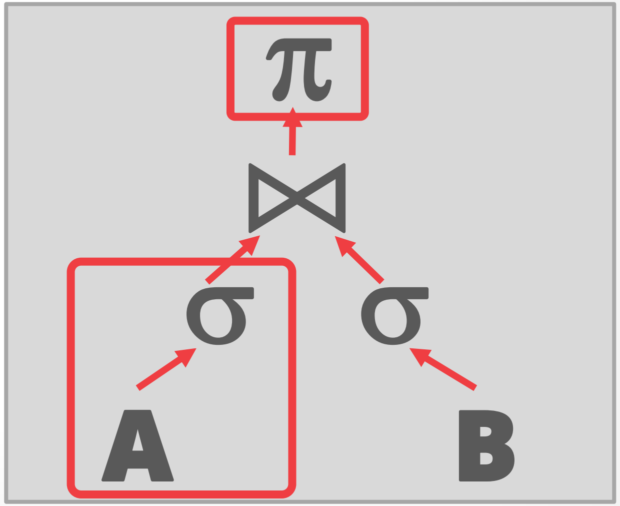
→ Operators are decomposed into independent instances that perform the same function on different subsets of data.

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

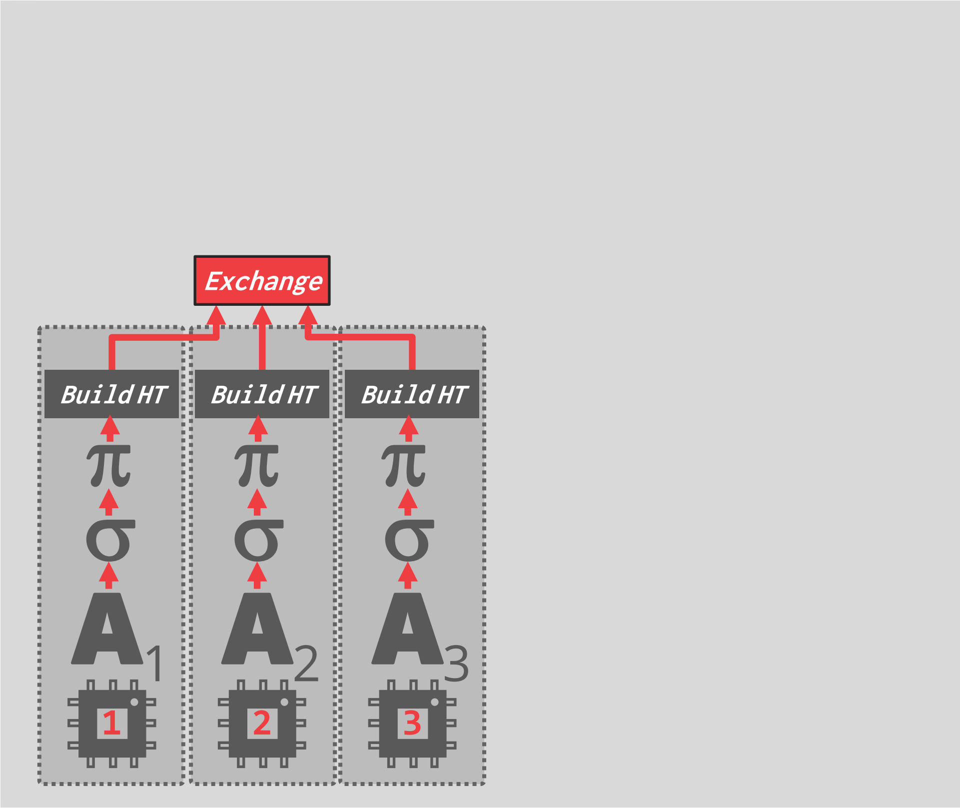
INTRA-OPERATOR PARALLELISM



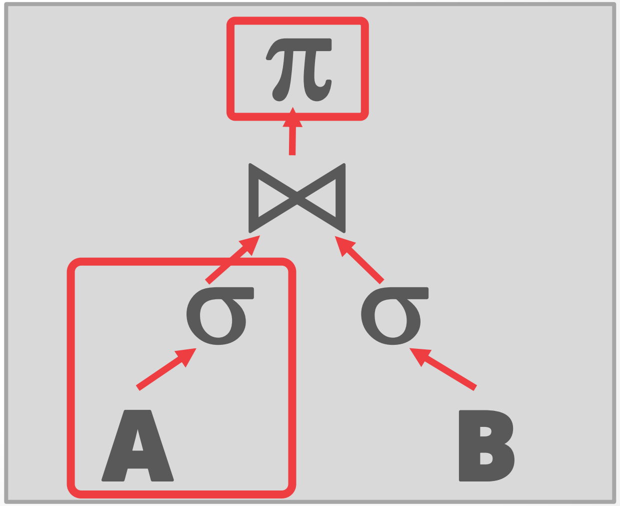

```
SELECT A.id, B.value
  FROM A JOIN B
    ON A.id = B.id
 WHERE A.value < 99
   AND B.value > 100
```



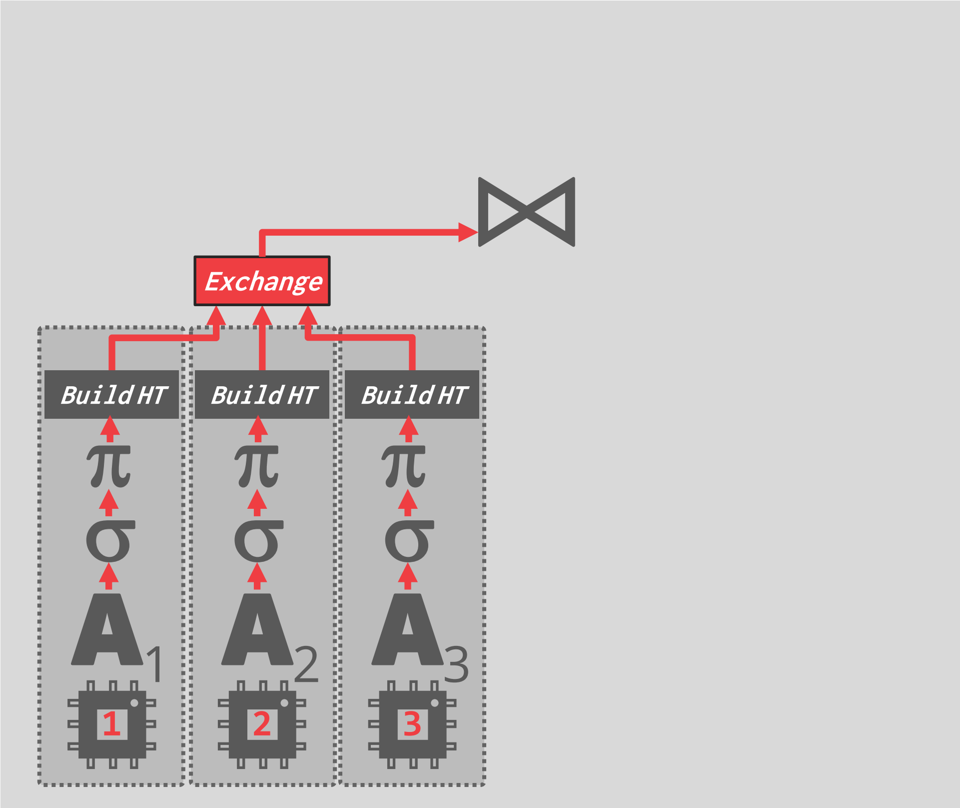
INTRA-OPERATOR PARALLELISM



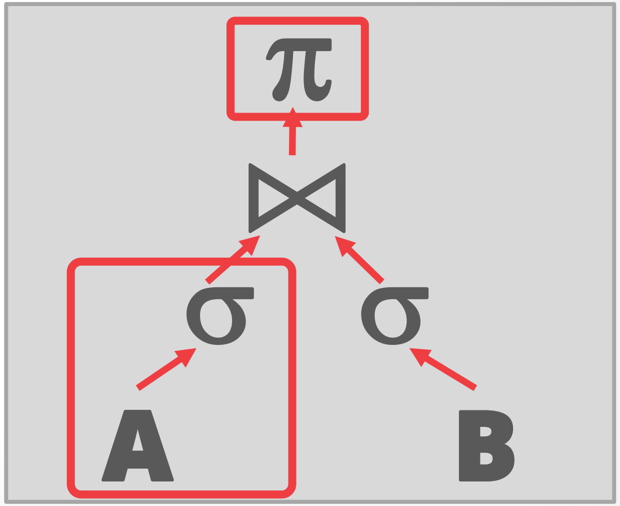

```
SELECT A.id, B.value
  FROM A JOIN B
    ON A.id = B.id
 WHERE A.value < 99
   AND B.value > 100
```



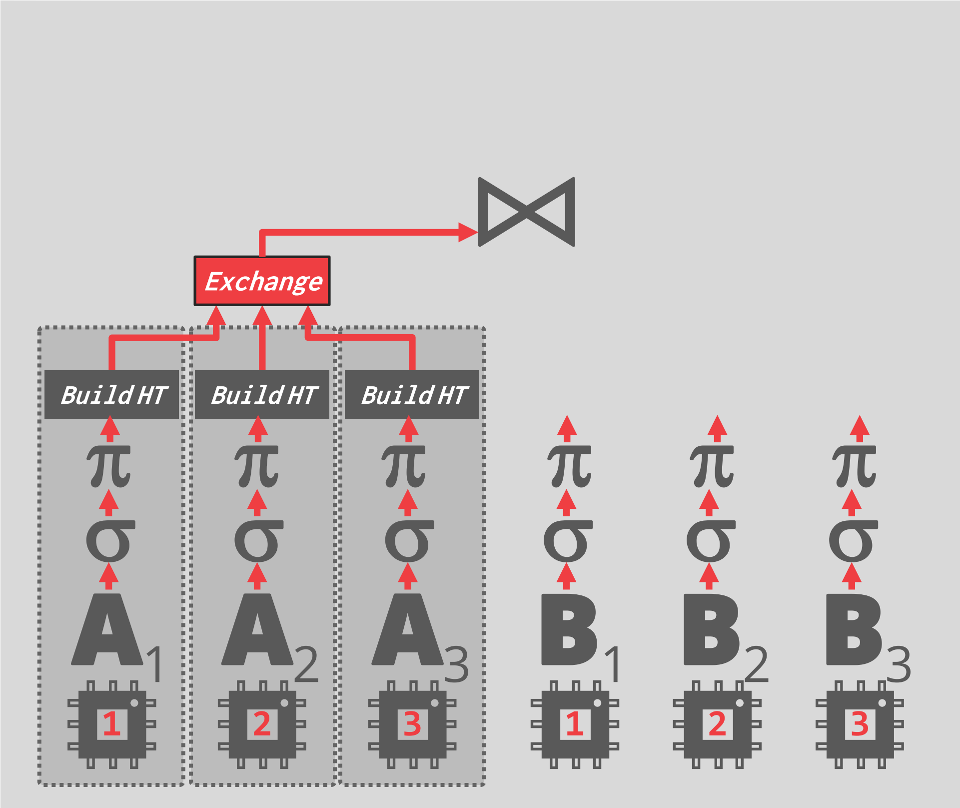
INTRA-OPERATOR PARALLELISM



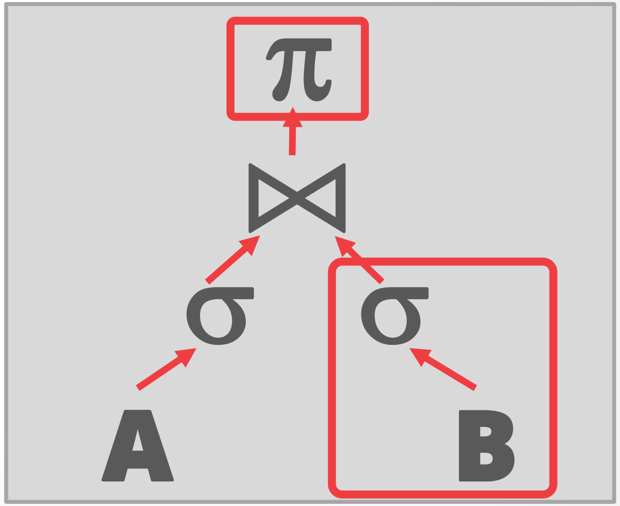

```sql
SELECT A.id, B.value
  FROM A JOIN B
    ON A.id = B.id
 WHERE A.value < 99
   AND B.value > 100
```



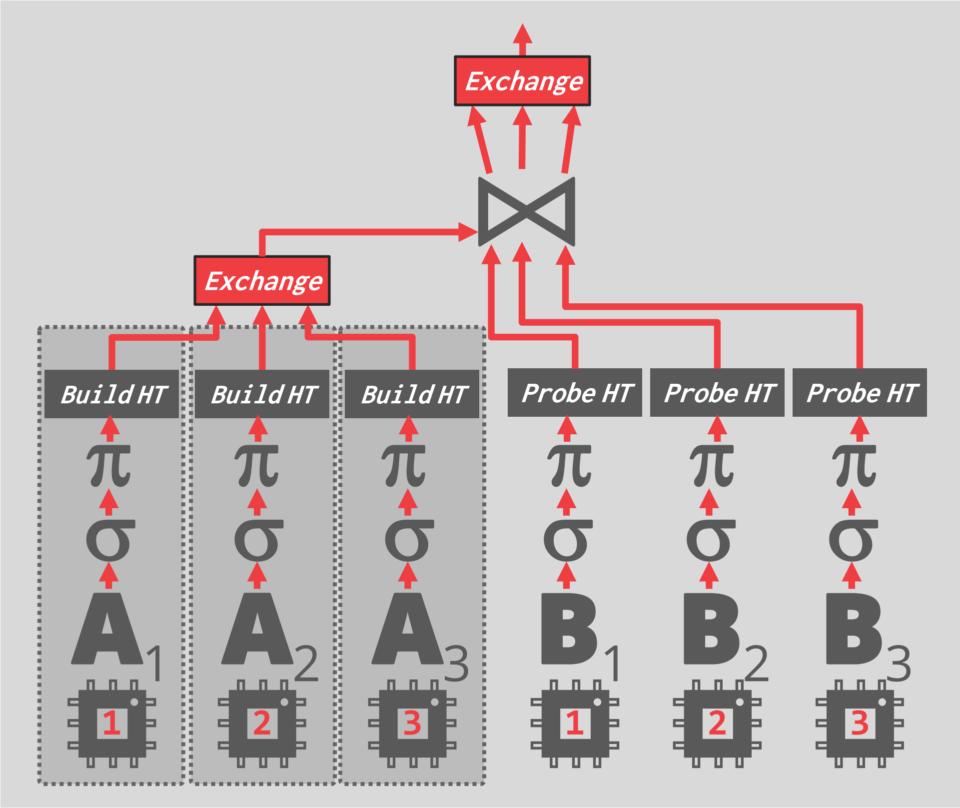
INTRA-OPERATOR PARALLELISM



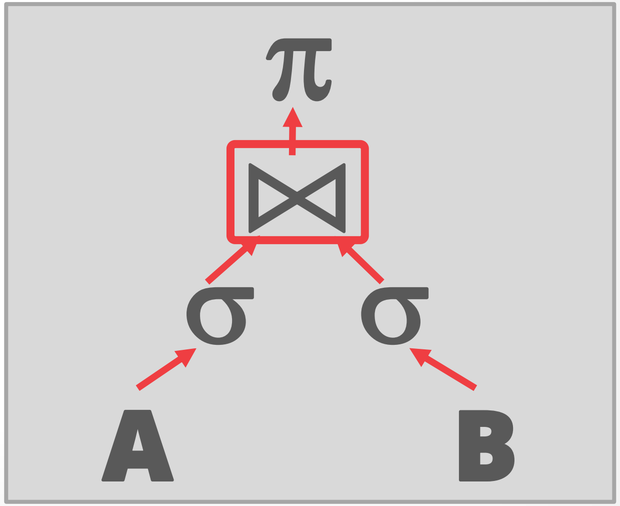

```
SELECT A.id, B.value
  FROM A JOIN B
    ON A.id = B.id
 WHERE A.value < 99
   AND B.value > 100
```



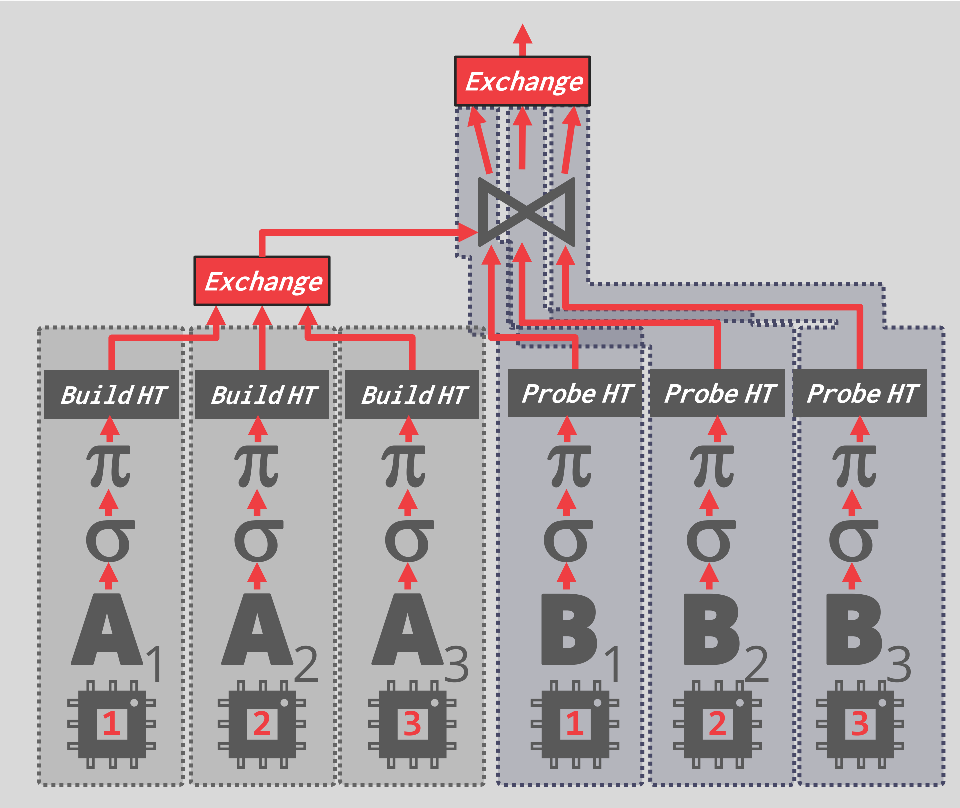
INTRA-OPERATOR PARALLELISM



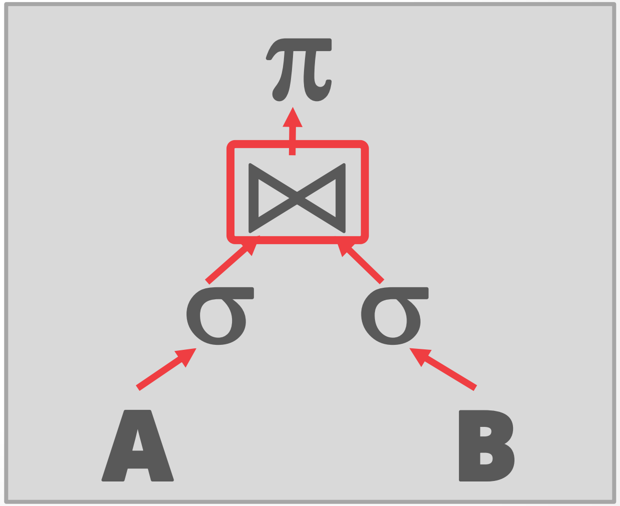

```sql
SELECT A.id, B.value
  FROM A JOIN B
    ON A.id = B.id
 WHERE A.value < 99
   AND B.value > 100
```



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
```

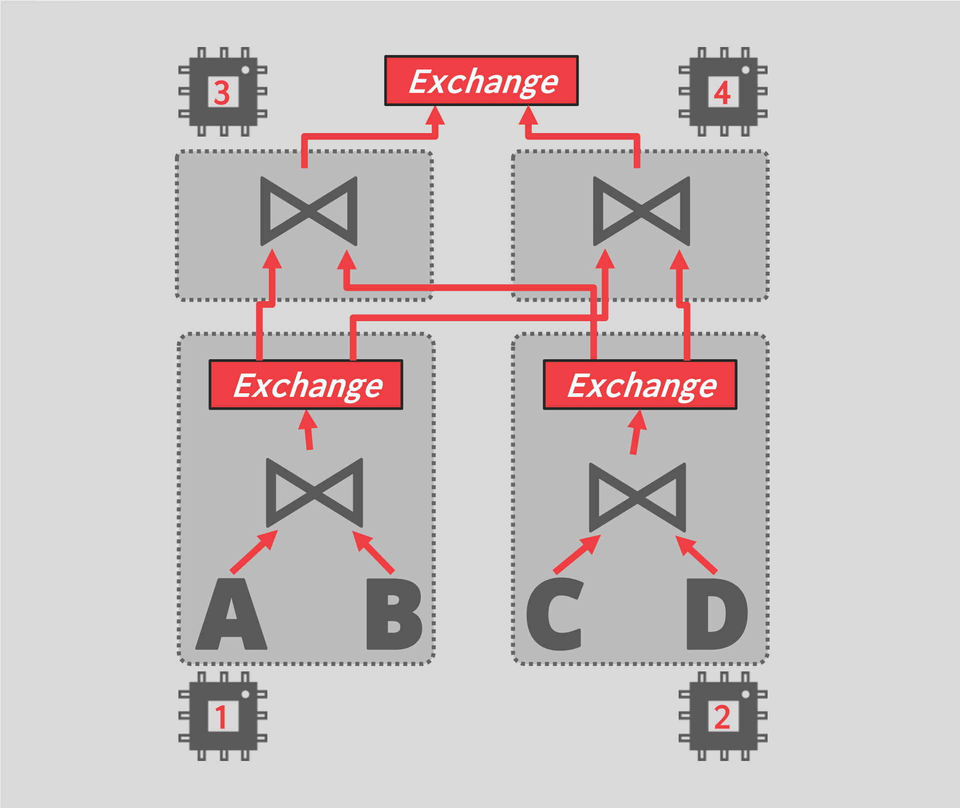


INTER-OPERATOR PARALLELISM

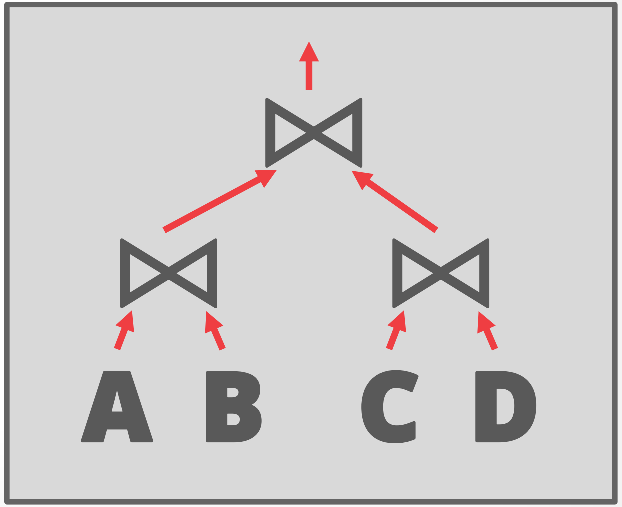
Approach #2: Inter-Operator (Vertical)
→ Operations are overlapped in order to pipeline data from one stage to the next without materialization.
→ Workers execute multiple operators from different segments of a query plan at the same time.
→ Still need exchange operators to combine intermediate results from segments.

Also called **pipelined parallelism**.

INTER-OPERATOR PARALLELISM




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



PARTING THOUGHTS

The easiest way to implement something is not going to always produce the most efficient execution strategy for modern CPUs.

We will see that vectorized / bottom-up execution will be the better way to execute OLAP queries.

NEXT CLASS

Query Task Scheduling