

Scaling Inter-procedural Dataflow Analysis on the Cloud

Static Analysis and Software Verification

Simone Colli

2025-12-10

- 1 Introduction
- 2 Background
- 3 Methods
- 4 Results
- 5 Conclusion

Definition: Dataflow analysis is a static technique used to gather information about the possible set of values calculated at various points in a program.

Beyond compiler optimizations, dataflow analysis is widely used in:

- Bug detection.
- Security analysis.

Applying inter-procedural dataflow analysis to large, real-world software presents major challenges:

- 1 **Memory explosion:** Analysis tracks information at every program point. For precise analysis (e.g., context-sensitive alias analysis), the state size is huge.
- 2 **Compute intensity:** Flow-sensitive analysis requires updating facts via transfer functions for every statement until convergence (Fixpoint).

Theoretical foundation

Background

The framework relies on the standard Monotone Dataflow Analysis Framework.

Mathematical model

- **Lattice** L : Represents the domain of abstract values.
- **Transfer function** $f: L \rightarrow L$. Models the effect of statements (e.g., *GEN/KILL* sets).
- **Merge operator** \sqcup (Join) or \sqcap (Meet): Combines information from predecessors.

The goal is to compute the **least fixpoint** (lfp) iteratively.

Worklist algorithm

Background

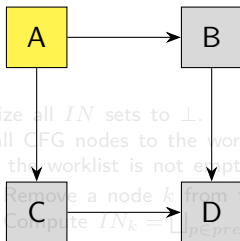
The standard sequential worklist algorithm iteratively updates dataflow facts:

- ① Initialize all IN sets to \perp .
- ② Add all CFG nodes to the worklist.
- ③ While the worklist is not empty:
 - Remove a node k from the worklist.
 - Compute $IN_k = \bigsqcup_{p \in preds(k)} OUT_p$.
 - Compute $OUT_k = f_k(IN_k)$.
 - If OUT_k changed, add successors of k to the worklist.

This continues until no more changes occur (fixpoint reached).

Worklist animation: Step 1

Background



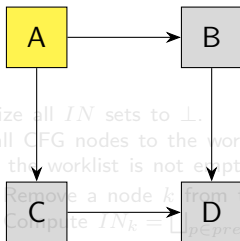
- Initialize all IN sets to \perp .

- 1: Initialize all IN sets to \perp .
- 2: Add all CFG nodes to the worklist.
- 3: While the worklist is not empty:
 - Remove a node k from the worklist.
 - Compute $IN_k = \bigcup_{p \in preds(k)} OUT_p$.
 - Compute $OUT_k = f_k(IN_k)$.
 - If OUT_k changed, add successors of k to the worklist.
- 4: This continues until no more changes occur (fixpoint reached).

Worklist: {A, B, C, D}

Worklist animation: Step 1

Background



- 1: Initialize all IN sets to \perp .
- 2: Add all CFG nodes to the worklist.
- 3: While the worklist is not empty:

- Remove a node k from the worklist.
- Compute $IN_k = \bigcup_{p \in preds(k)} OUT_p$.
- Compute $OUT_k = f_k(IN_k)$.
- If OUT_k changed, add successors of k to the worklist.

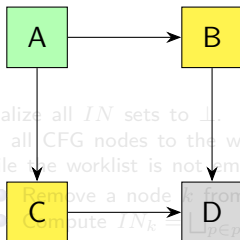
- 4: This continues until no more changes occur (fixpoint reached).

- Initialize all IN sets to \perp .
- Add all CFG nodes to the worklist.

Worklist: {A, B, C, D}

Worklist animation: Step 2

Background



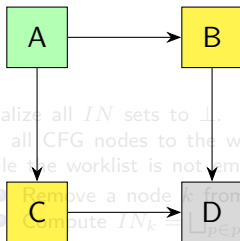
- Remove A from worklist.

- 1: Initialize all IN sets to \perp .
- 2: Add all CFG nodes to the worklist.
- 3: While the worklist is not empty:
 - Remove a node k from the worklist.
 - Compute $IN_k = \bigcup_{p \in preds(k)} OUT_p$.
 - Compute $OUT_k = f_k(IN_k)$.
 - If OUT_k changed, add successors of k to the worklist.
- 4: This continues until no more changes occur (fixpoint reached).

Worklist: {B, C, D}

Worklist animation: Step 2

Background



- 1: Initialize all IN sets to \perp .
- 2: Add all CFG nodes to the worklist.
- 3: While the worklist is not empty:

- Remove a node k from the worklist.
- Compute $IN_k = \bigcup_{p \in preds(k)} OUT_p$.
- Compute $OUT_k = f_k(IN_k)$.
- If OUT_k changed, add successors of k to the worklist.

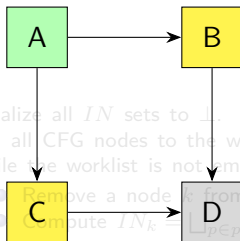
- 4: This continues until no more changes occur (fixpoint reached).

- Remove A from worklist.
- Compute IN_A (no predecessors).

Worklist: {B, C, D}

Worklist animation: Step 2

Background



- 1: Initialize all IN sets to \perp .
- 2: Add all CFG nodes to the worklist.
- 3: While the worklist is not empty:

- Remove a node k from the worklist.
- Compute $IN_k = \bigcup_{p \in preds(k)} OUT_p$.
- Compute $OUT_k = f_k(IN_k)$.
- If OUT_k changed, add successors of k to the worklist.

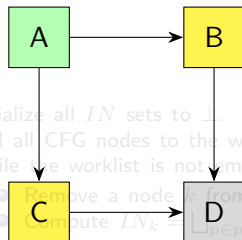
- 4: This continues until no more changes occur (fixpoint reached).

- Remove A from worklist.
- Compute IN_A (no predecessors).
- Compute $OUT_A = f_A(IN_A)$.

Worklist: {B, C, D}

Worklist animation: Step 2

Background



- 1: Initialize all IN sets to \perp .
- 2: Add all CFG nodes to the worklist.
- 3: While the worklist is not empty:

- Remove a node k from the worklist.
- Compute $IN_k = \bigcup_{p \in preds(k)} OUT_p$.
- Compute $OUT_k = f_k(IN_k)$.
- If OUT_k changed, add successors of k to the worklist.

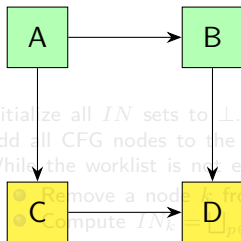
- 4: This continues until no more changes occur (fixpoint reached).

- Remove A from worklist.
- Compute IN_A (no predecessors).
- Compute $OUT_A = f_A(IN_A)$.
- OUT_A changed, add B and C to worklist.

Worklist: {B, C, D}

Worklist animation: Step 3

Background



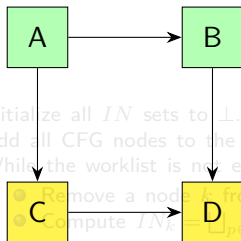
- Remove B from worklist.

- 1: Initialize all IN sets to \perp .
- 2: Add all CFG nodes to the worklist.
- 3: While the worklist is not empty:
 - Remove a node k from the worklist.
 - Compute $IN_k = \bigcup_{p \in preds(k)} OUT_p$.
 - Compute $OUT_k = J_k(IN_k)$.
 - If OUT_k changed, add successors of k to the worklist.
- 4: This continues until no more changes occur (fixpoint reached).

Worklist: {C, D}

Worklist animation: Step 3

Background



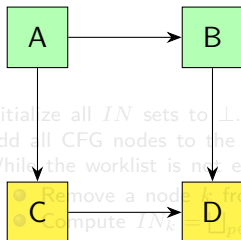
- Remove B from worklist.
- Compute $IN_B = OUT_A$.

- 1: Initialize all IN sets to \perp .
- 2: Add all CFG nodes to the worklist.
- 3: While the worklist is not empty:
 - Remove a node k from the worklist.
 - Compute $IN_k = \bigcap_{p \in preds(k)} OUT_p$.
 - Compute $OUT_k = j_k(IN_k)$.
 - If OUT_k changed, add successors of k to the worklist.
- 4: This continues until no more changes occur (fixpoint reached).

Worklist: {C, D}

Worklist animation: Step 3

Background



- 1: Initialize all IN sets to \perp .
- 2: Add all CFG nodes to the worklist.
- 3: While the worklist is not empty:

• Remove a node k from the worklist.

• Compute $IN_k = \bigcup_{p \in preds(k)} OUT_p$.

• Compute $OUT_k = f_k(IN_k)$.

• If OUT_k changed, add successors of k to the worklist.

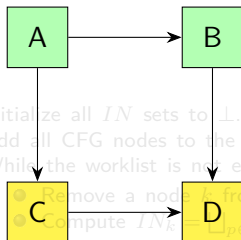
- 4: This continues until no more changes occur (fixpoint reached).

- Remove B from worklist.
- Compute $IN_B = OUT_A$.
- Compute $OUT_B = f_B(IN_B)$.

Worklist: {C, D}

Worklist animation: Step 3

Background



- 1: Initialize all IN sets to \perp .
- 2: Add all CFG nodes to the worklist.
- 3: While the worklist is not empty:

- Remove a node k from the worklist.

- Compute $IN_k = \bigcup_{p \in preds(k)} OUT_p$.

- Compute $OUT_k = f_k(IN_k)$.

- If OUT_k changed, add successors of k to the worklist.

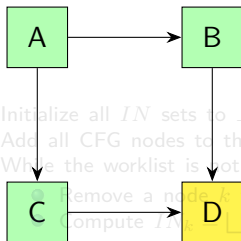
- 4: This continues until no more changes occur (fixpoint reached).

- Remove B from worklist.
- Compute $IN_B = OUT_A$.
- Compute $OUT_B = f_B(IN_B)$.
- OUT_B changed, add D to worklist.

Worklist: {C, D}

Worklist animation: Step 4

Background



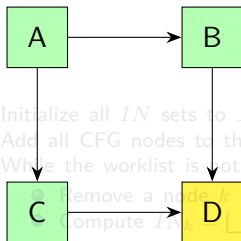
- Remove C from worklist.

- 1: Initialize all IN sets to \perp .
- 2: Add all CFG nodes to the worklist.
- 3: While the worklist is not empty:
 - Remove a node k from the worklist.
 - Compute $IN_k = \bigcup_{p \in preds(k)} OUT_p$.
 - Compute $OUT_k = f_k(IN_k)$.
 - If OUT_k changed, add successors of k to the worklist.
- 4: This continues until no more changes occur (fixpoint reached).

Worklist: {D}

Worklist animation: Step 4

Background



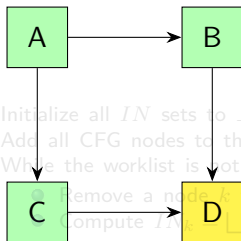
- Remove C from worklist.
- Compute $IN_C = OUT_A$.

- 1: Initialize all IN sets to \perp .
- 2: Add all CFG nodes to the worklist.
- 3: While the worklist is not empty:
 - Remove a node k from the worklist.
 - Compute $IN_k = \bigcup_{p \in preds(k)} OUT_p$.
 - Compute $OUT_k = f_k(IN_k)$.
 - If OUT_k changed, add successors of k to the worklist.
- 4: This continues until no more changes occur (fixpoint reached).

Worklist: {D}

Worklist animation: Step 4

Background



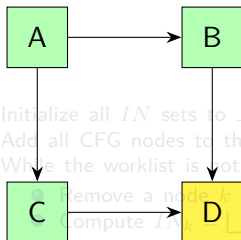
- 1: Initialize all IN sets to \perp .
- 2: Add all CFG nodes to the worklist.
- 3: While the worklist is not empty:
 - Remove a node k from the worklist.
 - Compute $IN_k = \bigcup_{p \in preds(k)} OUT_p$.
 - Compute $OUT_k = f_k(IN_k)$.
 - If OUT_k changed, add successors of k to the worklist.
- 4: This continues until no more changes occur (fixpoint reached).

- Remove C from worklist.
- Compute $IN_C = OUT_A$.
- Compute $OUT_C = f_C(IN_C)$.

Worklist: {D}

Worklist animation: Step 4

Background



- 1: Initialize all IN sets to \perp .
- 2: Add all CFG nodes to the worklist.
- 3: While the worklist is not empty:

• Remove a node k from the worklist.

• Compute $IN_k = \bigcup_{p \in preds(k)} OUT_p$.

• Compute $OUT_k = f_k(IN_k)$.

• If OUT_k changed, add successors of k to the worklist.

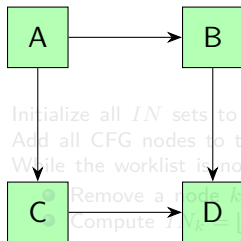
- 4: This continues until no more changes occur (fixpoint reached).

- Remove C from worklist.
- Compute $IN_C = OUT_A$.
- Compute $OUT_C = f_C(IN_C)$.
- OUT_C changed, add D to worklist (already there).

Worklist: {D}

Worklist animation: Step 5

Background



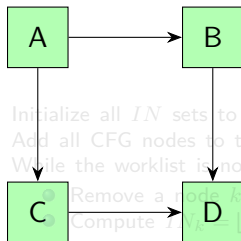
- Remove D from worklist.

- 1: Initialize all IN sets to \perp .
- 2: Add all CFG nodes to the worklist.
- 3: While the worklist is not empty:
 - Remove a node k from the worklist.
 - Compute $IN_k = \bigcup_{p \in preds(k)} OUT_p$.
 - Compute $OUT_k = f_k(IN_k)$.
 - If OUT_k changed, add successors of k to the worklist.
- 4: This continues until no more changes occur (fixpoint reached).

Worklist: {}

Worklist animation: Step 5

Background



1: Initialize all IN sets to \perp .

2: Add all CFG nodes to the worklist.

3: While the worklist is not empty:

- Remove a node k from the worklist.

- Compute $IN_k = \bigsqcup_{p \in preds(k)} OUT_p$.

- Compute $OUT_k = f_k(IN_k)$.

- If OUT_k changed, add successors of k to the worklist.

4: This continues until no more changes occur (fixpoint reached).

- Remove D from worklist.

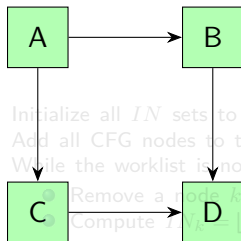
- Compute

$$IN_D = OUT_B \sqcup OUT_C.$$

Worklist: {}

Worklist animation: Step 5

Background



1: Initialize all IN sets to \perp .

2: Add all CFG nodes to the worklist.

3: While the worklist is not empty:

• Remove a node k from the worklist.

• Compute $IN_k = \bigsqcup_{p \in preds(k)} OUT_p$.

• Compute $OUT_k = f_k(IN_k)$.

• If OUT_k changed, add successors of k to the worklist.

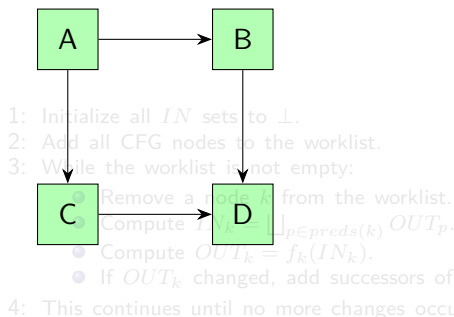
4: This continues until no more changes occur (fixpoint reached).

- Remove D from worklist.
- Compute $IN_D = OUT_B \sqcup OUT_C$.
- Compute $OUT_D = f_D(IN_D)$.

Worklist: {}

Worklist animation: Step 5

Background

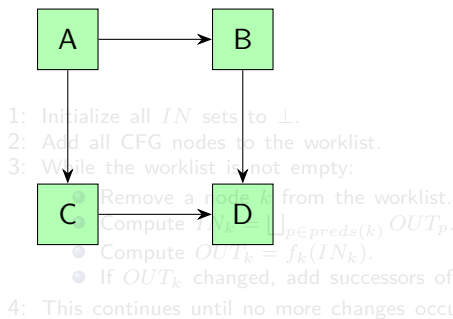


- Remove D from worklist.
- Compute $IN_D = OUT_B \sqcup OUT_C$.
- Compute $OUT_D = f_D(IN_D)$.
- OUT_D changed, add successors of D to worklist (none).

Worklist: {}

Worklist animation: Step 5

Background



- Remove D from worklist.
- Compute $IN_D = OUT_B \sqcup OUT_C$.
- Compute $OUT_D = f_D(IN_D)$.
- OUT_D changed, add successors of D to worklist (none).
- Worklist is empty, fixpoint reached.

Worklist: {}

Vertex-centric graph processing

Background

To distribute the analysis, the paper adopts the **vertex-centric model**.

The **GAS** model:

- **Gather**: A vertex collects data (messages) from neighbors (predecessors in CFG).
- **Apply**: The vertex updates its internal state (Dataflow fact) using the transfer function.
- **Scatter**: If the state changes, the vertex activates its neighbors (successors) and sends new messages.

Mapping:

- Graph vertices \rightarrow CFG basic blocks/statements.
- Vertex value \rightarrow Abstract state (*IN* and *OUT* sets).

Vertex-centric graph processing algorithm

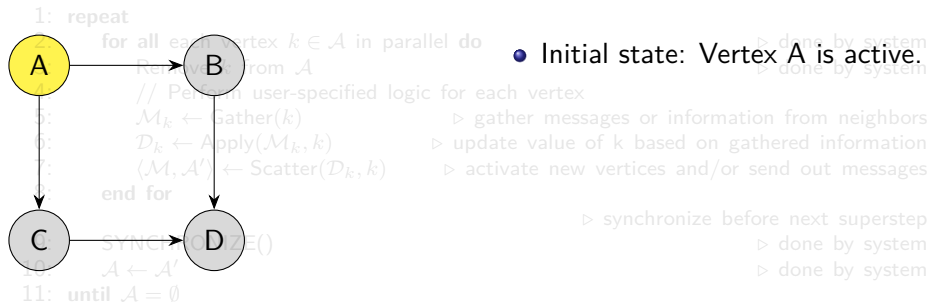
Background

Algorithm 1 The vertex-centric GAS model algorithm.

```
1: Data:  $\mathcal{A}$ , the set of active vertices during processing
2: repeat
3:   for all each vertex  $k \in \mathcal{A}$  in parallel do                                ▷ done by system
4:     Remove  $k$  from  $\mathcal{A}$                                                          ▷ done by system
5:     // Perform user-specified logic for each vertex
6:      $\mathcal{M}_k \leftarrow \text{Gather}(k)$                                              ▷ gather messages or information from neighbors
7:      $\mathcal{D}_k \leftarrow \text{Apply}(\mathcal{M}_k, k)$                                        ▷ update value of  $k$  based on gathered information
8:      $\langle \mathcal{M}, \mathcal{A}' \rangle \leftarrow \text{Scatter}(\mathcal{D}_k, k)$                        ▷ activate new vertices and/or send out messages
9:   end for                                                                    ▷ synchronize before next superstep
10:  SYNCHRONIZE()                                                            ▷ done by system
11:   $\mathcal{A} \leftarrow \mathcal{A}'$                                                     ▷ done by system
12: until  $\mathcal{A} = \emptyset$ 
```

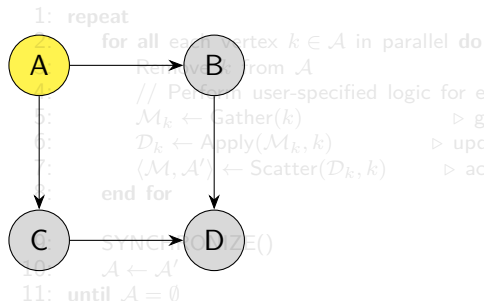
GAS animation: superstep 1

Background



GAS animation: superstep 1

Background



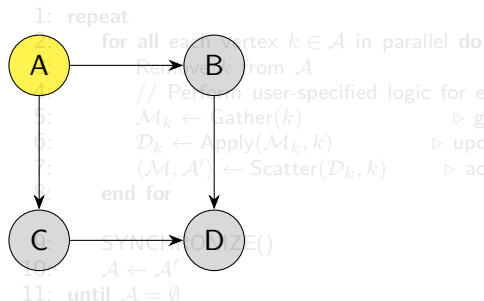
- Initial state: Vertex A is active.

- A Gathers from neighbors (none).

- ▷ done by system
- ▷ done by system
- ▷ gather messages or information from neighbors
- ▷ update vertices based on gathered information
- ▷ activate new vertices and/or send out messages
- ▷ synchronize before next superstep
- ▷ done by system
- ▷ done by system

GAS animation: superstep 1

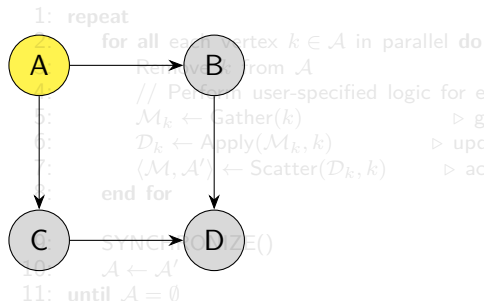
Background



- Initial state: Vertex A is active.
- A Gathers from neighbors (none).
- A Applies function, value changes.

GAS animation: superstep 1

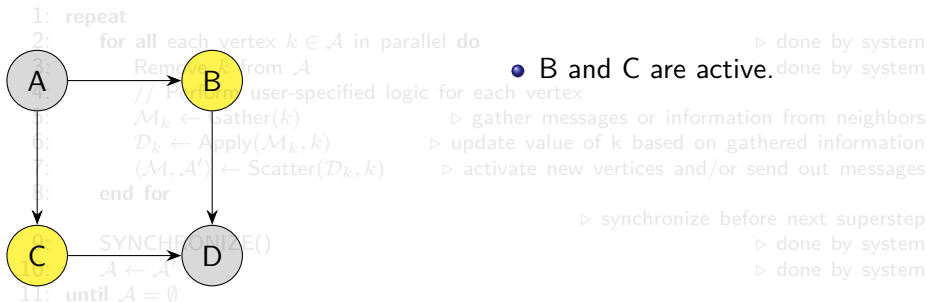
Background



- Initial state: Vertex A is active.
- A Gathers from neighbors (none).
- A Applies function, value changes.
- A Scatters to B and C, activating them.

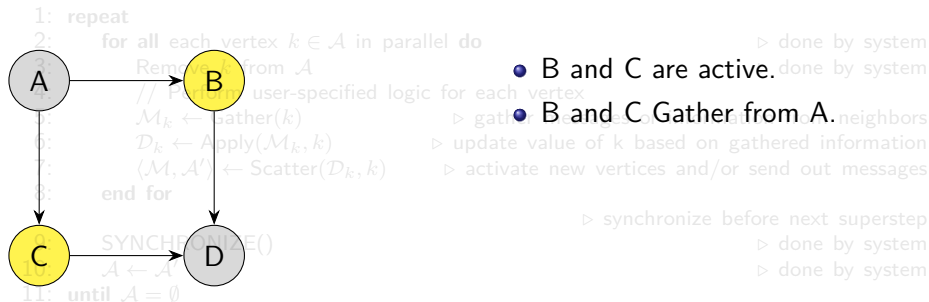
GAS animation: superstep 2

Background



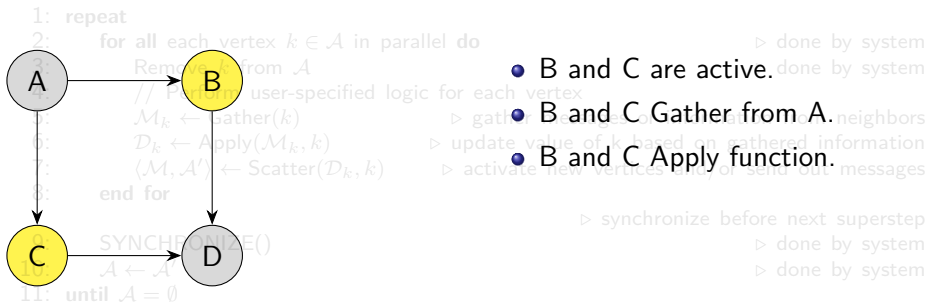
GAS animation: superstep 2

Background



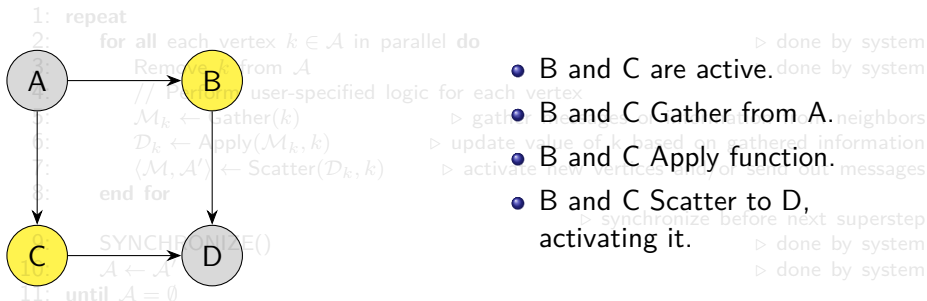
GAS animation: superstep 2

Background



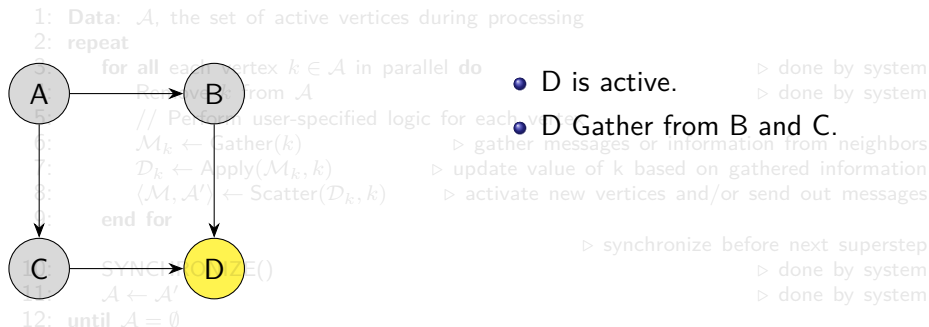
GAS animation: superstep 2

Background



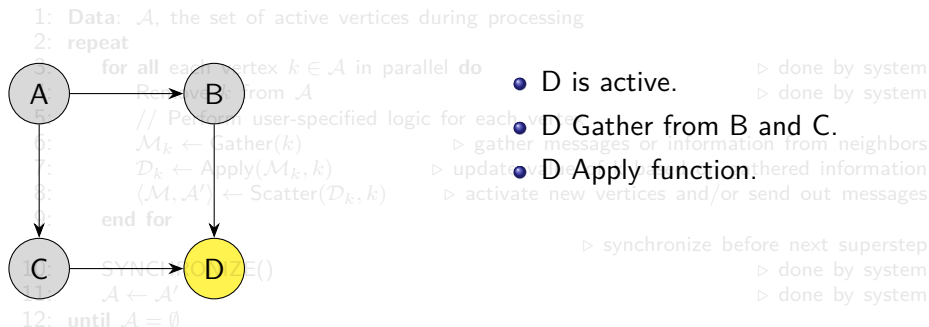
GAS animation: superstep 3

Background



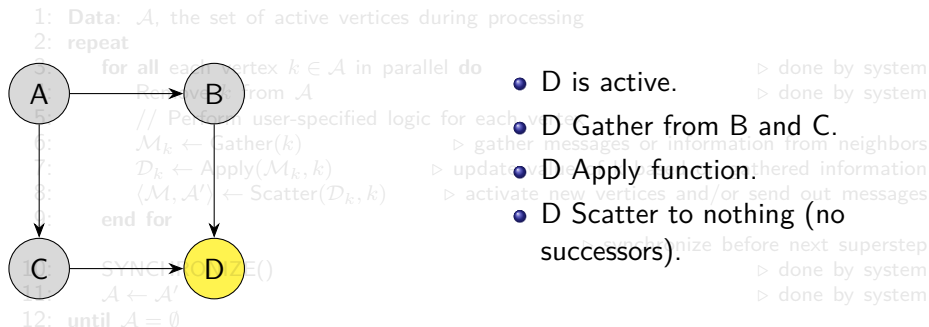
GAS animation: superstep 3

Background



GAS animation: superstep 3

Background



Inspired by the GAS model, the paper designs the distributed vertex-centric worklist algorithm.

The paper proposes 2 frameworks:

- BigDataflow.
- BigDataflow-incremental.

Each in a classic and an optimized version:

- BigDataflow-classic.
- BigDataflow.
- BigDataflow-incremental.
- BigDataflow-incremental optimized.

System implementation

Methods

BigDataflow is built on top of mature distributed systems to handle large-scale computation and state persistence.

Apache Giraph (Computation)

- **Role:** Executes the distributed worklist algorithms.
- Functionality:
 - **Partitions** the massive CFG across cluster nodes.
 - Manages **synchronization** (supersteps) and message passing between workers.

Redis (State storage)

- **Role:** Enabler for **incremental analysis**.
- Functionality:
 - Stores **whole-program analysis results**.
 - Allows low-latency **query** (for impact analysis) and **update** of facts.

BigDataflow-classic algorithm

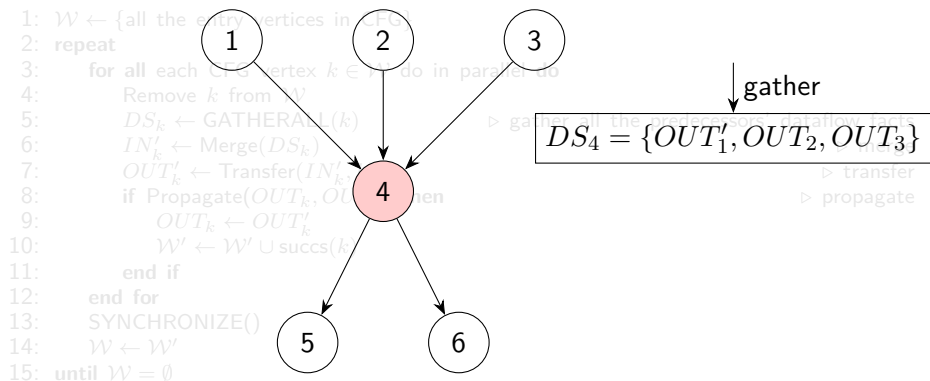
Methods

Algorithm 2 The BigDataflow-classic algorithm.

```
1: Data:  $\mathcal{W}$ , the list of all active vertices during analysis;  $DS_k : \{OUT_p | p \in preds(k)\}$  a set  
   containing all the dataflow facts of  $k$ 's predecessors  
2:  $\mathcal{W} \leftarrow \{\text{all the entry vertices in CFG}\}$   
3: repeat  
4:   for all each CFG vertex  $k \in \mathcal{W}$  do in parallel do  
5:     Remove  $k$  from  $\mathcal{W}$   
6:      $DS_k \leftarrow \text{GATHERALL}(k)$  ▷ gather all the predecessors' dataflow facts  
7:      $IN'_k \leftarrow \text{Merge}(DS_k)$  ▷ merge  
8:      $OUT'_k \leftarrow \text{Transfer}(IN'_k, k)$  ▷ transfer  
9:     if  $\text{Propagate}(OUT_k, OUT'_k)$  then ▷ propagate  
10:       $OUT_k \leftarrow OUT'_k$   
11:       $\mathcal{W}' \leftarrow \mathcal{W}' \cup \text{succs}(k)$   
12:     end if  
13:   end for  
14:    $\text{SYNCHRONIZE}()$   
15:    $\mathcal{W} \leftarrow \mathcal{W}'$   
16: until  $\mathcal{W} = \emptyset$ 
```

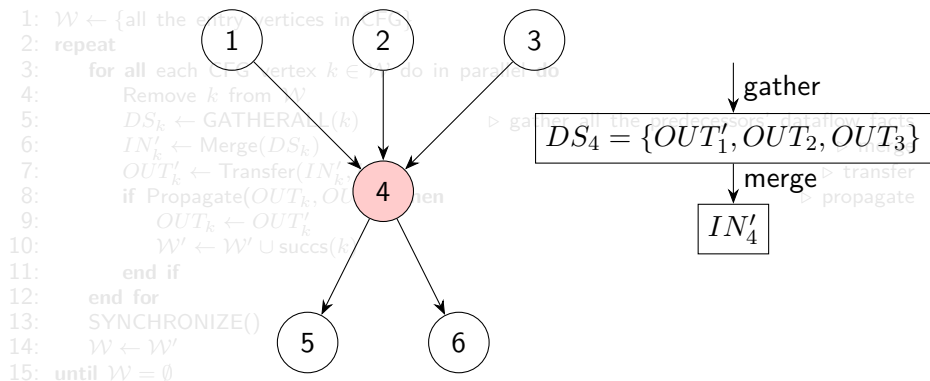
BigDataflow-classic animation

Methods



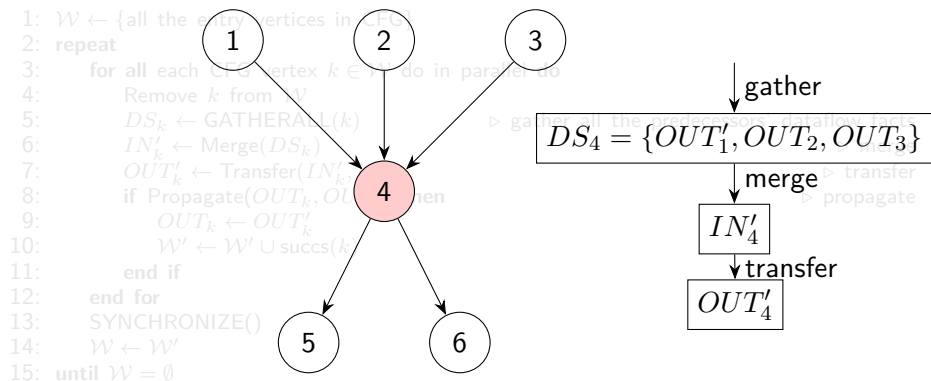
BigDataflow-classic animation

Methods



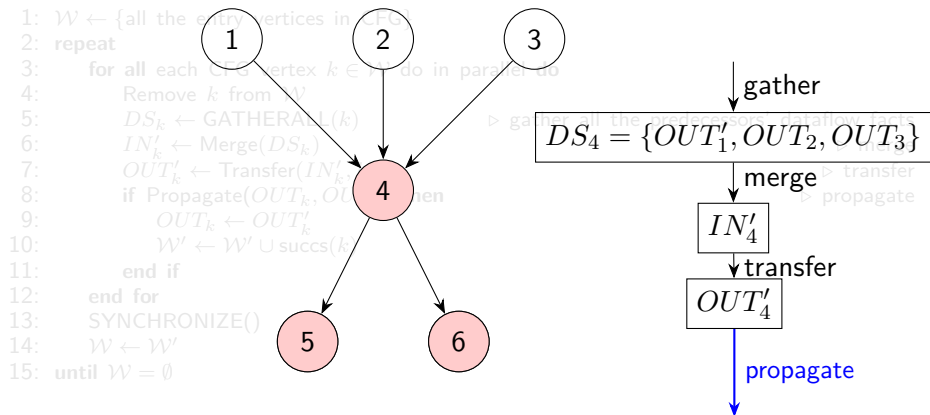
BigDataflow-classic animation

Methods



BigDataflow-classic animation

Methods



BigDataflow-classic limitations

Methods

```
1:  $\mathcal{W} \leftarrow \{\text{all the entry vertices in CFG}\}$ 
2: repeat
3:   for  $\mathcal{W}$  in parallel
4:     Despite that the algorithm succeeds in leveraging large-
5:     scale distributed computing resources to accelerate dataflow facts
6:     dataflow analysis, it suffers from poor scalability.
7:   end for
8:   if  $\text{Propagate}(\mathcal{W}, \mathcal{W}', \text{CFG})$  then
9:     Each vertex gathers all predecessor facts in ev-
10:    ery superstep, leading to redundant data trans-
11:    mission and high memory consumption.
12:   end for
13:   SYNCHRONIZE()
14:    $\mathcal{W} \leftarrow \mathcal{W}'$ 
15: until  $\mathcal{W} = \emptyset$ 
```

BigDataflow optimized

Methods

The optimized algorithm gathers only the **updated** dataflow facts from predecessors in the previous superstep.

This approach is achieved via a push-based message passing mechanism

BigDataflow optimized algorithm

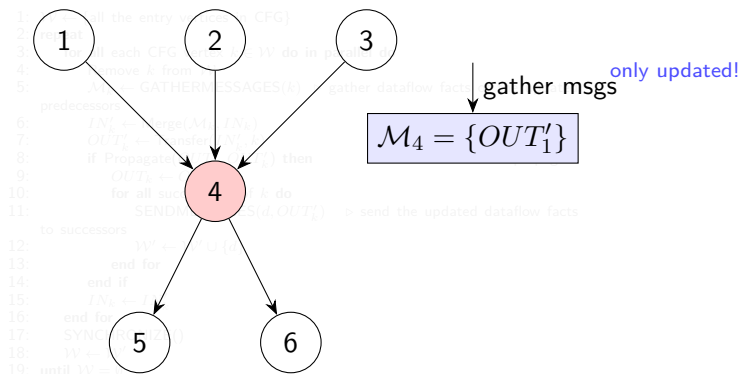
Methods

Algorithm 3 The BigDataflow optimized algorithm.

```
1:  $\mathcal{W} \leftarrow \{\text{all the entry vertices in CFG}\}$ 
2: repeat
3:   for all each CFG vertex  $k \in \mathcal{W}$  do in parallel do
4:     Remove  $k$  from  $\mathcal{W}$ 
5:      $\mathcal{M}_k \leftarrow \text{GATHERMESSAGES}(k)$  ▷ gather only updated dataflow facts
6:      $IN'_k \leftarrow \text{Merge}(\mathcal{M}_k, IN_k)$  ▷ merge
7:      $OUT'_k \leftarrow \text{Transfer}(IN'_k, k)$  ▷ transfer
8:     if  $\text{Propagate}(OUT_k, OUT'_k)$  then ▷ propagate
9:        $OUT_k \leftarrow OUT'_k$ 
10:    for all successor  $d$  of  $k$  do
11:       $\text{SENDMESSAGES}(d, OUT'_k)$  ▷ send the updated dataflow facts
12:       $\mathcal{W}' \leftarrow \mathcal{W}' \cup \{d\}$ 
13:    end for
14:  end if
15:   $IN_k \leftarrow IN'_k$ 
16: end for
17:  $\text{SYNCHRONIZE}()$ 
18:  $\mathcal{W} \leftarrow \mathcal{W}'$ 
19: until  $\mathcal{W} = \emptyset$ 
```

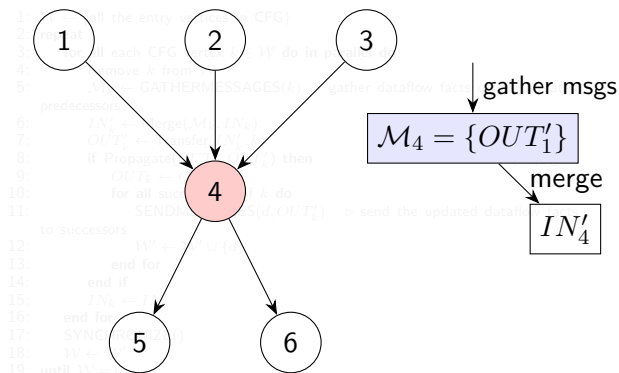
BigDataflow optimized animation

Methods



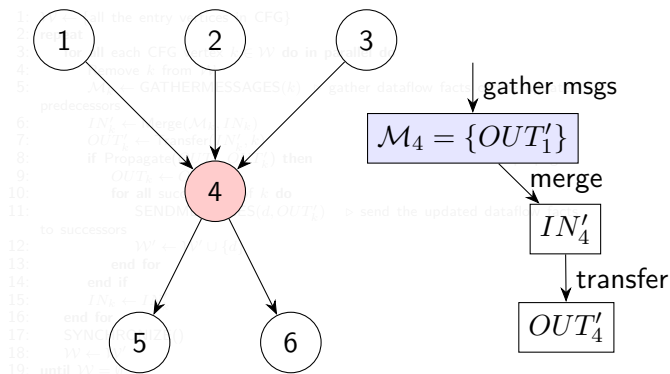
BigDataflow optimized animation

Methods



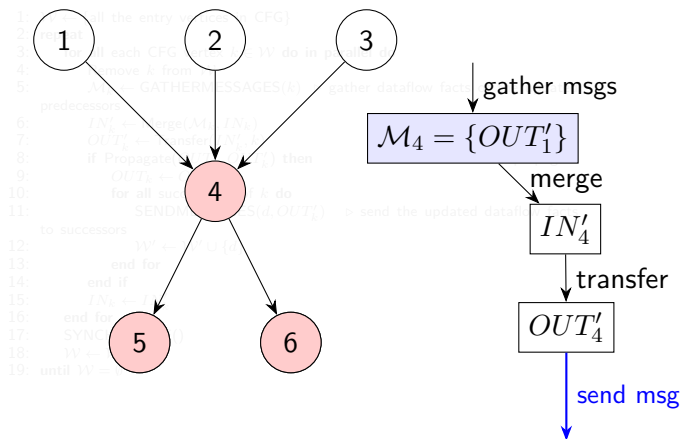
BigDataflow optimized animation

Methods



BigDataflow optimized animation

Methods



BigDataflow-incremental

Methods

Real-world software changes frequently, especially in open-source projects. Re-analyzing from scratch is time-consuming and expensive.

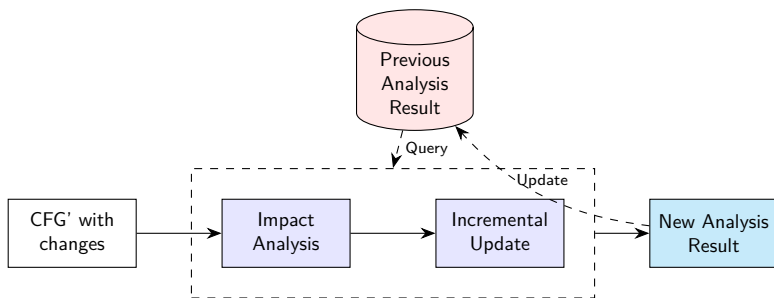


Figure: Workflow of Distributed Incremental Dataflow Analysis.

Workflow of BigDataflow-incremental

Methods

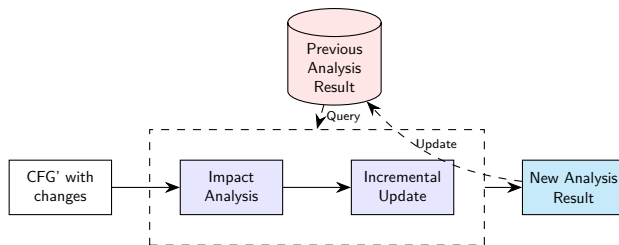


Figure: Workflow of Distributed Incremental Dataflow Analysis.

Given a **CFG with a batch of updates**, the framework:

- Performs **impact analysis** to identify affected nodes.
- Performs **incremental update**:
 - Identification of affected nodes.
 - Extraction of affected sub-CFGs.
 - Re-analyzing only the affected sub-CFGs.
- Produces the **new analysis result**.

The distributed incremental analysis reuses the existing analysis result and only performs local analysis on the updated parts to achieve efficiency.

Atomic Changes: The framework handles three categories of changes on the CFG:

- **Addition:**

- Added edge between existing nodes.
- Added source node and edge.
- Added destination node.

- **Deletion:**

- Deleted edge.
- Deleted source/destination node.

- **Change:**

- Changed source node (updates transfer function).
- Changed destination node.

These atomic changes trigger the **Impact Analysis** to identify the initial set of affected nodes.

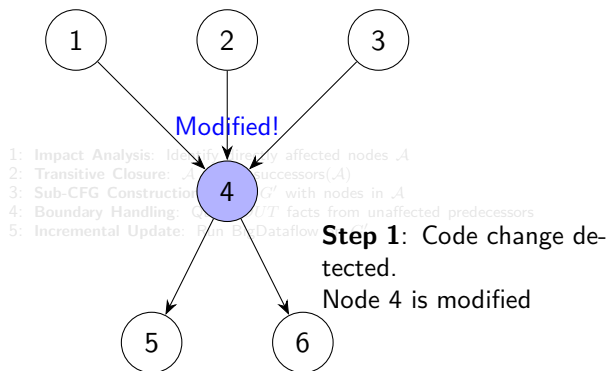
BigDataflow-incremental algorithm steps

Methods

- 1: **Impact Analysis:** Identify directly affected nodes \mathcal{A} based on atomic changes.
- 2: **Transitive Closure:** $\mathcal{A} \leftarrow \mathcal{A} \cup \text{successors}(\mathcal{A})$.
- 3: **Sub-CFG Construction:** Build G' with nodes in \mathcal{A} and connecting edges.
- 4: **Initialization:**
- 5: For $k \in \mathcal{A}_{add_only}$: reuse previous results (Optimized).
- 6: For others: reset to \perp or \top .
- 7: **Boundary Handling:** Query OUT facts from unaffected predecessors.
- 8: **Incremental Update:** Run BigDataflow on G' .

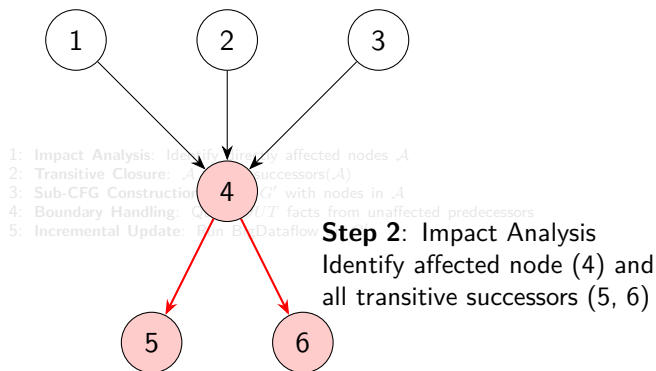
BigDataflow-incremental algorithm

Methods



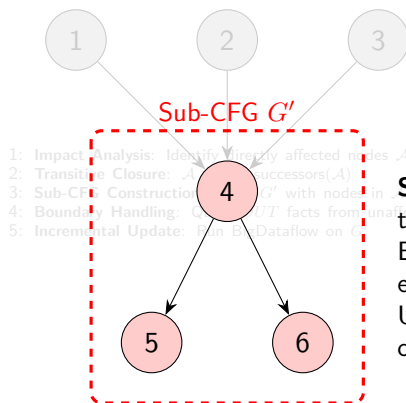
BigDataflow-incremental algorithm

Methods



BigDataflow-incremental algorithm

Methods



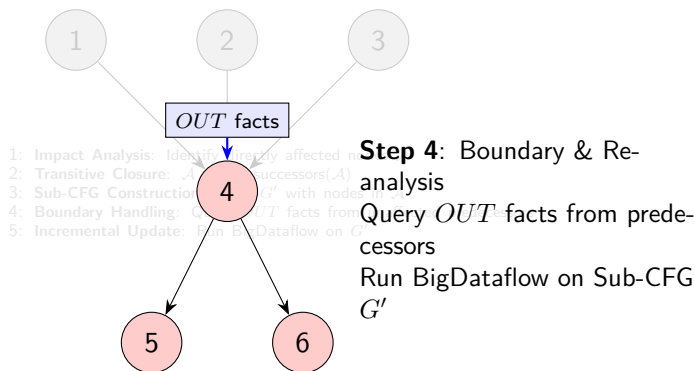
Step 3: Sub-CFG Construction

Extract affected nodes and edges

Unaffected nodes are grayed out

BigDataflow-incremental algorithm

Methods



The evaluation aims to answer the following questions:

- Q1: What is the overall performance of BigDataflow given a rich set of distributed computing resources?
- Q2: How does BigDataflow perform compared with other competitive analysis systems/tools?
- Q3: What about the performance of BigDataflow given the varying numbers of cores and resources?
- Q4: How about the performance of BigDataflow in the mode of incremental analysis?

Evaluation subjects

Results

Table: Subjects for evaluation

| Subject | Version | #LoC | #Functions | Description |
|------------|---------|-------|------------|------------------|
| Linux | 5.2 | 17.5M | 565K | operating system |
| Firefox | 67.0 | 7.9M | 770K | web browser |
| PostgreSQL | 12.2 | 1.0M | 30K | database system |
| OpenSSL | 1.1.1 | 519K | 12K | TLS protocol |
| Httpd | 2.4.39 | 196K | 6K | web server |

Where #LoC is the number of lines of code and #Functions is the number of functions.

Reference tools and hardware

Results

Table: Tools and hardware

| Tool | Hardware | Description |
|---------------------------------------|---|---|
| BigDataflow-classic (Baseline) | Cluster: 125 nodes (8 vCPUs, 64GB RAM) | Classic distributed algorithm with full predecessor gathering |
| BigDataflow (Proposed) | Cluster: 125 nodes (8 vCPUs, 64GB RAM) | Optimized with delta updates and message passing |
| Chianina (State-of-the-Art) | Single server (104 vCPUs, 768GB RAM, 1TB SSD) | Parallel single-machine system with out-of-core support |

Evaluation analysis

Results

To test the performance and scalability of BigDataflow, two notoriously “heavy” dataflow analyses were chosen:

Context-sensitive alias analysis

Determines which pointers can point to the same memory location, considering the calling context of functions. It is computationally expensive because tracking every possible call path leads to an exponential explosion in the state space, resulting in massive memory consumption and long computation times.

Instruction cache analysis

Predicts whether a machine instruction is likely to be in the CPU’s instruction cache. It is memory-intensive as it requires modeling the cache behavior for all possible execution paths. The number of states can be enormous for large programs with complex control flow.

Alias analysis

Results

| Subject | Tool | #PAliases | #Workers/#Part | #PMem | Time | Cost (\$) |
|------------|---------------------|-----------|----------------|--------|-------|-----------|
| Linux | BigDataflow | 12.5B | 350 | 3.5T | 16.7m | 11.1 |
| | BigDataflow-classic | - | 350 | x | x | x |
| | Chianina | - | 4 | 453.4G | 17.4h | 110.4 |
| Firefox | BigDataflow | 11.5B | 140 | 1.2T | 16.5m | 4.4 |
| | BigDataflow-classic | - | 140 | x | x | x |
| | Chianina | - | 4 | 131.6G | 5.3h | 33.6 |
| PostgreSQL | BigDataflow | 727.0M | 50 | 329.7G | 2.8m | 0.3 |
| | BigDataflow-classic | - | 50 | 330.7G | 4.9m | 0.5 |
| | Chianina | - | 1 | 61.9G | 50.4m | 5.3 |
| OpenSSL | BigDataflow | 734.8M | 30 | 285.3G | 3.5m | 0.2 |
| | BigDataflow-classic | - | 30 | 329.8G | 6.8m | 0.4 |
| | Chianina | - | 1 | 43.2G | 35.4m | 3.7 |
| Httpd | BigDataflow | 183.1M | 10 | 119.9G | 2.8m | 0.1 |
| | BigDataflow-classic | - | 10 | 137.9G | 4.0m | 0.1 |
| | Chianina | - | 1 | 14.2G | 11.2m | 1.2 |

Where #PAliases is the number of pointer alias pairs discovered.

Instruction cache analysis

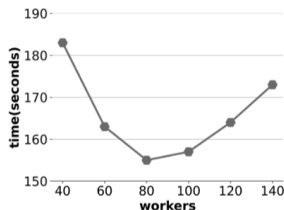
Results

| Subject | Tool | #BCached | #Workers/#Part | #PMem | Time | Cost (\$) |
|------------|---------------------|----------|----------------|--------|-------|-----------|
| Linux | BigDataflow | 21.5B | 500 | 5.6T | 44.4m | 42.0 |
| | BigDataflow-classic | - | 500 | x | x | x |
| | Chianina | - | 4 | 555.4G | 9.4h | 59.6 |
| Firefox | BigDataflow | 15.8B | 400 | 4.4T | 39.0m | 29.5 |
| | BigDataflow-classic | - | 400 | x | x | x |
| | Chianina | - | 4 | 351.5G | 7.2h | 45.7 |
| PostgreSQL | BigDataflow | 1.4B | 180 | 1.1T | 3.2m | 1.1 |
| | BigDataflow-classic | - | 180 | 1.1T | 6.5m | 2.2 |
| | Chianina | - | 1 | 115.3G | 38.1m | 4.0 |
| OpenSSL | BigDataflow | 2.8B | 180 | 1.3T | 6.9m | 2.3 |
| | BigDataflow-classic | - | 180 | 1.5T | 13.4m | 4.6 |
| | Chianina | - | 1 | 227.6G | 1.7h | 10.8 |
| Httpd | BigDataflow | 782.0M | 100 | 684.3G | 3.0m | 0.6 |
| | BigDataflow-classic | - | 100 | 781.3G | 4.6m | 0.9 |
| | Chianina | - | 1 | 58.3G | 18.5m | 2.0 |

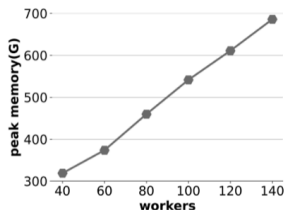
Where #BCached is the number of blocks that may be cached.

Scalability

Results



(a) time

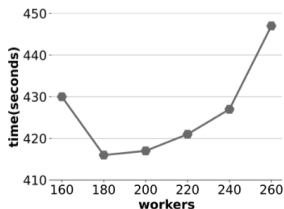


(b) peak memory

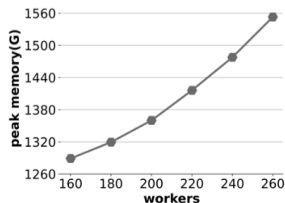
Figure: Scalability of BigDataflow on OpenSSL subject for alias analysis.

Scalability

Results



(a) time



(b) peak memory

Figure: Scalability of BigDataflow on OpenSSL subject for cache analysis.

Incremental analysis: Httpd-3days

Results

| Subject | BigDataflow-whole | | | | BigDataflow-incremental | | | |
|---------|-------------------|---------|----------|--------|-------------------------|---------|----------|--------|
| | Cost | Time | #Workers | #PMem | Cost | Time | #Workers | #PMem |
| NU10 | \$0.643 | 3.4mins | 100 | 626.5G | \$0.038 | 1.0mins | 20 | 90.6G |
| NU9 | / | / | / | / | / | / | / | / |
| NU8 | \$0.662 | 3.5mins | 100 | 623.2G | \$0.129 | 1.7mins | 40 | 206.2G |
| NU7 | / | / | / | / | / | / | / | / |
| NU6 | \$0.681 | 3.6mins | 100 | 626.5G | \$0.030 | 0.8mins | 20 | 75.6G |
| NU5 | / | / | / | / | / | / | / | / |
| NU4 | / | / | / | / | / | / | / | / |
| NU3 | / | / | / | / | / | / | / | / |
| NU2 | / | / | / | / | / | / | / | / |
| NU1 | / | / | / | / | / | / | / | / |

Where NU1 to NU10 are 10 consecutive updates to the Httpd codebase over 3 days.

Conclusion

BigDataflow addresses the scalability challenge of interprocedural dataflow analysis by leveraging cloud resources.

This work presents three key contributions:

- **Distributed Framework:** Adopts a vertex-centric graph model to distribute heavy computations across cluster nodes.
- **Algorithmic Innovation:**
 - Optimized worklist: Prunes redundant data transmission via delta updates.
 - Incremental analysis: Efficiently handles code updates, reducing cost and time by orders of magnitude.
- **Usability:** Provides high-level APIs, allowing users to implement complex client analyses (e.g., Alias, Cache) easily.

BigDataflow transforms “intractable” analysis tasks (e.g., analyzing 17.5M LoC Linux Kernel) from hours to minutes, making deep software verification more viable for modern CI/CD pipelines.

Thank you for your attention