Programmable Multi-Dimensional Table Filters for Line Rate Network Functions

Vishal Shrivastav
Purdue University
Evolution of Programmable Data Plane Hardware*

* Not an exhaustive list
Evolution of Programmable Data Plane Hardware*

2013

**RMT** [SIGCOMM’13]
*Match-Action Tables*

2016

**Domino** [SIGCOMM’16]
*Stateful Atoms*

* Not an exhaustive list
Evolution of Programmable Data Plane Hardware*

* Not an exhaustive list
Evolution of Programmable Data Plane Hardware*

2013
- RMT [SIGCOMM’13]
  - Match-Action Tables

2016
- Domino [SIGCOMM’16]
  - Stateful Atoms
- PIFO [SIGCOMM’16]
  - Push-In-First-Out Queue

2019
- FlowBlaze [NSDI’19]
  - Finite State Machine

* Not an exhaustive list
Evolution of Programmable Data Plane Hardware*

2013
- RMT [SIGCOMM’13]
  - Match-Action Tables

2016
- Domino [SIGCOMM’16]
  - Stateful Atoms
- PIFO [SIGCOMM’16]
  - Push-In-First-Out Queue

2019
- FlowBlaze [NSDI’19]
  - Finite State Machine
- PIEO [SIGCOMM’19]
  - Push-In-Extract-Out Queue

* Not an exhaustive list
Evolution of Programmable Data Plane Hardware*

2013
- RMT [SIGCOMM’13]
  Match-Action Tables

2016
- Domino [SIGCOMM’16]
  Stateful Atoms
- PIFO [SIGCOMM’16]
  Push-In-First-Out Queue

2019
- FlowBlaze [NSDI’19]
  Finite State Machine
- PIEO [SIGCOMM’19]
  Push-In-Extract-Out Queue

2022
- Taurus [ASPLOS’22]
  Map-Reduce

* Not an exhaustive list
Evolution of Programmable Data Plane Hardware*

2013
- RMT
  - [SIGCOMM’13] Match-Action Tables

2016
- Domino
  - [SIGCOMM’16] Stateful Atoms
- PIFO
  - [SIGCOMM’16] Push-In-First-Out Queue

2019
- FlowBlaze
  - [NSDI’19] Finite State Machine
- PIEO
  - [SIGCOMM’19] Push-In-Extract-Out Queue

2022
- Taurus
  - [ASPLOS’22] Map-Reduce
- Thanos
  - [SIGCOMM’22] Multi-Dimensional Table Filtering

* Not an exhaustive list
Data Plane Filtering is Prevalent

Performance-aware Routing
Filter paths with \( \text{delay} < d \) and \( \text{utilization} < u \).
Choose a \textit{random} path from the filtered set.

Multi-dimensional Filtering
# Data Plane Filtering is Prevalent

## Attributes

<table>
<thead>
<tr>
<th>Paths</th>
<th>Congestion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Performance-aware Routing
Filter paths with delay < d and utilization < u. Choose a random path from the filtered set.

### Congestion-aware Load Balancing
Filter path with minimum congestion.

---

# Multi-dimensional Filtering
Data Plane Filtering is Prevalent

<table>
<thead>
<tr>
<th>Ports</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Queuing</td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Performance-aware Routing**
Filter paths with delay < d and utilization < u. Choose a random path from the filtered set.

**Congestion-aware Load Balancing**
Filter path with minimum congestion or
Filter d random egress ports. Choose the least queued port from those d ports.

**Multi-dimensional Filtering**
Data Plane Filtering is Prevalent

Performance-aware Routing
Filter paths with delay < d and utilization < u. Choose a random path from the filtered set.

Congestion-aware Load Balancing
Filter path with minimum congestion or Filter d random egress ports. Choose the least queued port from those d ports.

Resource-aware L4 Load Balancing
Filter servers with avail mem > m and avail bw > b. From the filtered set, choose server with least cpu utilization.

Multi-dimensional Filtering

<table>
<thead>
<tr>
<th>Servers</th>
<th>Attributes</th>
<th>mem</th>
<th>bw</th>
<th>cpu</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Data Plane Filtering is Prevalent

<table>
<thead>
<tr>
<th>Ports</th>
<th>Attributes</th>
<th>Rate</th>
</tr>
</thead>
</table>

**Performance-aware Routing**
Filter paths with delay < d and utilization < u. Choose a random path from the filtered set.

**Congestion-aware Load Balancing**
Filter path with minimum congestion or
Filter d random egress ports. Choose the least queued port from those d ports.

**Resource-aware L4 Load Balancing**
Filter servers with avail mem > m and avail bw > b. From the filtered set, choose server with least cpu utilization.

**Data Plane Diagnosis**
Filter switch ports with packet rate > t.

**Multi-dimensional Filtering**
Data Plane Filtering is Prevalent

<table>
<thead>
<tr>
<th>Paths</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>A (0/1)</td>
</tr>
<tr>
<td></td>
<td>B (0/1)</td>
</tr>
</tbody>
</table>

**Performance-aware Routing**
Filter paths with delay < d and utilization < u. Choose a random path from the filtered set.

**Congestion-aware Load Balancing**
Filter path with minimum congestion or Filter d random egress ports. Choose the least queued port from those d ports.

**Resource-aware L4 Load Balancing**
Filter servers with avail mem > m and avail bw > b. From the filtered set, choose server with least cpu utilization.

**Data Plane Diagnosis**
Filter switch ports with packet rate > t.

**Policy Compliance**
From all available paths, filter the paths not carrying tenant A’s or B’s traffic. Choose a path at random from the filtered paths to route a new flow from tenant C.

**Multi-dimensional Filtering**
Data Plane Filtering is Prevalent

<table>
<thead>
<tr>
<th>Resources</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Performance-aware Routing**
Filter paths with delay < d and utilization < u. Choose a random path from the filtered set

**Congestion-aware Load Balancing**
Filter path with minimum congestion or
Filter d random egress ports. Choose the least queued port from those d ports

**Resource-aware L4 Load Balancing**
Filter servers with avail mem > m and avail bw > b. From the filtered set, choose server with least cpu utilization

**Data Plane Diagnosis**
Filter switch ports with packet rate > t

**Policy Compliance**
From all available paths, filter the paths not carrying tenant A's or B's traffic. Choose a path at random from the filtered paths to route a new flow from tenant C.

**Chained Multi-dimensional Filtering**
## Data Plane Filtering is Prevalent

### Performance-aware Routing
Filter paths with `delay < d` and `utilization < u`. Choose a random path from the filtered set.

### Congestion-aware Load Balancing
Filter path with minimum congestion or
Filter `d` random egress ports. Choose the least queued port from those `d` ports.

### Resource-aware L4 Load Balancing
Filter servers with `avail mem > m` and `avail bw > b`. From the filtered set, choose the server with least CPU utilization.

### Data Plane Diagnosis
Filter switch ports with packet rate `> t`.

### Attributes

<table>
<thead>
<tr>
<th>Resources</th>
<th>Attributes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Policy Compliance
From all available paths, filter the paths not carrying tenant A’s or B’s traffic. Choose a path at random from the filtered paths to route a new flow from tenant C.

### Chained Multi-dimensional Filtering at Line Rate
State-of-the-Art

Current Programmable Switch Processing Pipeline

Match-Action

Does not support line rate multi-dimensional filtering
Thanos
Thanos Switch Processing Pipeline

Match-Action

Multi-dimensional Table

Filter Processing Pipeline

Programmable Filter Module
Filter Abstractions and Primitives
Abstractions and Primitives

Unary Filter Processing Unit (UFPU)

UFPU: table $\rightarrow$ table
Abstractions and Primitives

Unary Filter Processing Unit (UFPU)

UFPU: table \rightarrow table

\text{random( )}

Filter a load balancing server at \text{random}

Attributes (attr)

<table>
<thead>
<tr>
<th>ID</th>
<th>mem</th>
<th>bw</th>
<th>cpu</th>
<th># conn</th>
</tr>
</thead>
</table>

Servers
Abstractions and Primitives

Unary Filter Processing Unit (UFPU)

UFPU: table $\rightarrow$ table

\{ random( )
\}
predicate(attr relop X)

Filter load balancing servers with avail mem $> X$

Attributes (attr)

<table>
<thead>
<tr>
<th>ID</th>
<th>mem</th>
<th>bw</th>
<th>cpu</th>
<th># conn</th>
</tr>
</thead>
</table>

Servers

UFPU

mem $> X$
Abstractions and Primitives

Unary Filter Processing Unit (UFPU)

UFPU: table $\rightarrow$ table

\{
  \text{random( )},
  \text{predicate(attr relop X)},
  \text{weighted-round-robin(attr)}
\}

Filter load balancing servers in a \textit{round robin} manner weighted by \textit{avail bw}

Attributes (attr)

<table>
<thead>
<tr>
<th>ID</th>
<th>mem</th>
<th>bw</th>
<th>cpu</th>
<th># conn</th>
</tr>
</thead>
</table>

Servers

UFPU

w-rr(bw)
Abstractions and Primitives

Unary Filter Processing Unit (UFPU)

UFPU: table $\rightarrow$ table

- random($\cdot$)
- predicate(attr $\textit{relop}$ X)
- weighted-round-robin(attr)
- min/max(attr)

Filter load balancing server with $\textit{min}$ cpu utilization

<table>
<thead>
<tr>
<th>ID</th>
<th>mem</th>
<th>bw</th>
<th>cpu</th>
<th># conn</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ID</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ID</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Abstractions and Primitives

Unary Filter Processing Unit (UFPU)

UFPU: table $\rightarrow$ table

\[
\begin{aligned}
\text{random( )} \\
\text{predicate(attr relop X)} \\
\text{weighted-round-robin(attr)} \\
\text{min/max(attr)}
\end{aligned}
\]

Filter **N least cpu utilized** load balancing servers

or

Filter **N random** load balancing servers

**Attributes (attr)**

<table>
<thead>
<tr>
<th>ID</th>
<th>mem</th>
<th>bw</th>
<th>cpu</th>
<th># conn</th>
</tr>
</thead>
</table>

A chain of **N UFPU**s

\[
I \quad I - O_1 \quad O_1 \quad O_2 \quad O_{N-1} \quad O_N = O
\]
Abstractions and Primitives

Unary Filter Processing Unit (UFPU)

UFPU: table → table

\{ random( )
  predicate(attr relop X)
  weighted-round-robin(attr)
  min/max(attr) \}

Filter N least cpu utilized load balancing servers
or
Filter N random load balancing servers

Attributes (attr)

<table>
<thead>
<tr>
<th>ID</th>
<th>mem</th>
<th>bw</th>
<th>cpu</th>
<th># conn</th>
</tr>
</thead>
</table>

A chain of N UFPU}s
Unary Filter Processing Unit (UFPU)

UFPU: table $\rightarrow$ table

\{ random( )
  predicate(attr relop X)
  weighted-round-robin(attr)
  min/max(attr) \}

Filter **N least cpu utilized** load balancing servers

or

Filter **N random** load balancing servers

Attributes (attr)

<table>
<thead>
<tr>
<th>ID</th>
<th>mem</th>
<th>bw</th>
<th>cpu</th>
<th># conn</th>
</tr>
</thead>
</table>

Servers

\[ I \rightarrow K-UFPU \rightarrow O \]
Abstractions and Primitives

Unary Filter Processing Unit (UFPU)

UFPU: table $\rightarrow$ table

Random processing:
- random()
- predicate(attr relop X)
- weighted-round-robin(attr)
- min/max(attr)

Attributes (attr)

Servers:

Filter $N$ least cpu utilized load balancing servers

or

Filter $N$ random load balancing servers

K-UFPU comprises $N$ UFPUs and adds a new configurable parameter $K$

K specifications:
- $K=N$, min(cpu)
- $K=N$, random()

(by setting $K=1$, K-UFPU reduces to UFPU)
Abstractions and Primitives

Unary Filter Processing Unit (UFPU)

UFPU: table → table

\{ random() \\
predicate(attr \textit{rel}op \textit{X}) \\
weighted-round-robin(attr) \\
min/max(attr) \}

Filter \textbf{N least cpu utilized} load balancing servers \\
or \\
Filter \textbf{N random} load balancing servers

Attributes (\textit{attr})

<table>
<thead>
<tr>
<th>ID</th>
<th>mem</th>
<th>bw</th>
<th>cpu</th>
<th># conn</th>
</tr>
</thead>
</table>

K-UFPU comprises N UFPU\(\textit{s} \) and adds a new configurable parameter \(K\)

\(K=\text{N, min(cpu)}\) \\
or \\
\(K=\text{N, random()}\)

We use \textbf{K-UFPU} (instead of UFPU) as the basic computing unit

K specifies the length of chain (max \(N\)) (by setting \(K=1\), K-UFPU reduces to UFPU)
Abstractions and Primitives

Unary Filter Processing Unit (UFPU)

UFPU: table \(\rightarrow\) table

\{ random() \\
predicate(attr \(relop\) X) \\
weighted-round-robin(attr) \\
min/max(attr) \}

<table>
<thead>
<tr>
<th>Attributes (attr)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ID</td>
</tr>
</tbody>
</table>

Binary Filter Processing Unit (BFPU)

BFPU: table, table \(\rightarrow\) table

\{ union() \\
intersection() \\
difference() \}

Servers
Abstractions and Primitives

Unary Filter Processing Unit (UFPU)

UFPU: table → table

- random()
- predicate(attr \text{ relop} X)
- weighted-round-robin(attr)
- min/max(attr)

Binary Filter Processing Unit (BFPU)

BFPU: table, table → table

- union()
- intersection()
- difference()

Filter all load balancing servers with avail mem > X and avail bw > Y
Abstractions and Primitives

Unary Filter Processing Unit (UFPU)

UFPU: table → table

- random()
- predicate(attr relop X)
- weighted-round-robin(attr)
- min/max(attr)

Binary Filter Processing Unit (BFPU)

BFPU: table, table → table

- union()
- intersection()
- difference()

Filter all load balancing servers with avail mem > X and avail bw > Y and # conn < Z
**Abstractions and Primitives**

**Unary Filter Processing Unit (UFPU)**

UFPU: table $\rightarrow$ table

- random( )
- predicate(attr $\text{relop}$ X)
- weighted-round-robin(attr)
- min/max(attr)

**Binary Filter Processing Unit (BFPU)**

BFPU: table, table $\rightarrow$ table

- union( )
- intersection( )
- difference( )

Filter all load balancing servers with \textit{avail mem} $>$ X \textbf{and} \textit{avail bw} $>$ Y \textbf{and} \textit{# conn} $<$ Z

Filter \textit{m} load balancing servers at \textit{random}

**Attributes (attr)**

<table>
<thead>
<tr>
<th>ID</th>
<th>mem</th>
<th>bw</th>
<th>cpu</th>
<th># conn</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- \textbf{K-UFPU} with $K=1$, \textit{mem} $>$ X
- \textbf{BFPU} with \textbf{intersect}
- \textbf{K-UFPU} with $K=1$, \textit{bw} $>$ Y
- \textbf{BFPU} with \textbf{intersect}
- \textbf{K-UFPU} with $K=1$, \textit{# conn} $<$ Z
- \textbf{BFPU} with \textbf{intersect}
- \textbf{K-UFPU} with $K=m$, random()
Abstractions and Primitives

Unary Filter Processing Unit (UFPU)
UFPU: table $\rightarrow$ table
- random($\cdot$)
- predicate(attr $\ relateop X$)
- weighted-round-robin(attr)
- min/max(attr)

Binary Filter Processing Unit (BFPU)
BFPU: table, table $\rightarrow$ table
- union($\cdot$)
- intersection($\cdot$)
- difference($\cdot$)

Filter all load balancing servers with $\text{avail mem} > X \ \text{and} \ \text{avail bw} > Y \ \text{and} \ # \ \text{conn} < Z$
Filter $m$ load balancing servers at random:
Filter load balancing server with $\text{min cpu util}$

Attributes (attr)

<table>
<thead>
<tr>
<th>ID</th>
<th>mem</th>
<th>bw</th>
<th>cpu</th>
<th># conn</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Servers
## Abstractions and Primitives

### Unary Filter Processing Unit (UFPU)

- **UFPU**: table $\rightarrow$ table
- random()
- predicate(attr $relop$ X)
- weighted-round-robin(attr)
- min/max(attr)

### Binary Filter Processing Unit (BFPU)

- **BFPU**: table, table $\rightarrow$ table
- union()
- intersection()
- difference()

### A 5-stage serial chain filter pipeline

(outputs from stage i are inputs to stage i+1)

<table>
<thead>
<tr>
<th>Stage 1</th>
<th>Stage 2</th>
<th>Stage 3</th>
<th>Stage 4</th>
<th>Stage 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-UFPU</td>
<td>BFPU</td>
<td>BFPU</td>
<td>K-UFPU</td>
<td>K-UFPU</td>
</tr>
</tbody>
</table>

### Attributes (attr)

<table>
<thead>
<tr>
<th>ID</th>
<th>mem</th>
<th>bw</th>
<th>cpu</th>
<th># conn</th>
</tr>
</thead>
</table>

### Diagram

- **Stage 1**: K-UFPU
- **Stage 2**: BFPU
- **Stage 3**: BFPU
- **Stage 4**: K-UFPU
- **Stage 5**: K-UFPU
From Abstractions to Hardware Design
Two Hardware Components

Attributes (attr)

<table>
<thead>
<tr>
<th>ID</th>
<th>mem</th>
<th>bw</th>
<th>cpu</th>
<th># conn</th>
</tr>
</thead>
</table>

Servers

Stage 1
- K-UFPU

Stage 2
- BFPU
- K-UFPU

Stage 3
- BFPU
- K-UFPU

Stage 4
- K-UFPU

Stage 5
- K-UFPU
# Hardware Component # 1

## 1. Multi-dimensional Table
## 2. Programmable Filter Pipeline
How to design an efficient data structure for a multi-dimensional relational table?

Should allow line rate read, write, update!
Multi-Dimensional Table Data Structure

Limitations of Classic Data Structures

No universal data structure

• Range trees / B-Trees for range filtering
• Heap for min/max filtering
• Disjoint Set Data Structure for set operations
• Either compromise on performance of certain operations or pay the cost of maintaining multiple data structures over the same data

Hierarchical Structure

• Fundamental O(log(N)) latency
• Hard to pipeline

Multi-Dimensional Table Data Structure

Our Solution:

Sorted Multidimensional Bidirectional Map (SMBM)
Multi-Dimensional Table Data Structure

Our Solution:
Sorted Multidimensional Bidirectional Map (SMBM)

<table>
<thead>
<tr>
<th>ID</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>19</td>
</tr>
<tr>
<td>3</td>
<td>22</td>
<td>19</td>
</tr>
<tr>
<td>4</td>
<td>15</td>
<td>8</td>
</tr>
</tbody>
</table>

Example Table
# SMBM Property # 1

**Our Solution:**

**Sorted Multidimensional Bidirectional Map (SMBM)**

Store each dimension as *flat* list of flip-flops

Allows parallel access and processing

<table>
<thead>
<tr>
<th>ID</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>19</td>
</tr>
<tr>
<td>3</td>
<td>22</td>
<td>19</td>
</tr>
<tr>
<td>4</td>
<td>15</td>
<td>8</td>
</tr>
</tbody>
</table>

Example Table

![Example Table](image-url)

<table>
<thead>
<tr>
<th>ID</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>15</td>
<td>19</td>
</tr>
<tr>
<td>4</td>
<td>22</td>
<td>19</td>
</tr>
</tbody>
</table>

Stored as
**SMBM Property # 2**

**Our Solution:**
Sorted Multidimensional Bidirectional Map (SMBM)

- Each list is kept sorted.
- Allows fast max/min filter operations.

### Example Table

<table>
<thead>
<tr>
<th>ID</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>19</td>
</tr>
<tr>
<td>3</td>
<td>22</td>
<td>19</td>
</tr>
<tr>
<td>4</td>
<td>15</td>
<td>8</td>
</tr>
</tbody>
</table>

Stored as:

- **ID**
  - Sorted: 1, 2, 3, 4
- **X**
  - Sorted: 4, 15, 15, 22
- **Y**
  - Sorted: 6, 8, 19, 19
SMBM Property # 3

Our Solution: Sorted Multidimensional Bidirectional Map (SMBM)

Each dimension is stored as an independent list. Allows parallel filter operations on multiple dimensions.
SMBM Property # 4

Our Solution:
Sorted Multidimensional Bidirectional Map (SMBM)

Bidirectional mapping between ID and attributes

Example Table

<table>
<thead>
<tr>
<th>ID</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>19</td>
</tr>
<tr>
<td>3</td>
<td>22</td>
<td>19</td>
</tr>
<tr>
<td>4</td>
<td>15</td>
<td>8</td>
</tr>
</tbody>
</table>
Our Solution:
Sorted Multidimensional Bidirectional Map (SMBM)

Example Table

<table>
<thead>
<tr>
<th>ID</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>19</td>
</tr>
<tr>
<td>3</td>
<td>22</td>
<td>19</td>
</tr>
<tr>
<td>4</td>
<td>15</td>
<td>8</td>
</tr>
</tbody>
</table>

Bidirectional mapping between ID and attributes
Forward map keeps track of which attributes belong to which ID
SMBM Property # 4

Our Solution:
Sorted Multidimensional Bidirectional Map (SMBM)

Bidirectional mapping between ID and attributes
Forward map keeps track of which attributes belong to which ID
Reverse map allows fast mapping of filtered attributes to their respective IDs

Example Table

<table>
<thead>
<tr>
<th>ID</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>19</td>
</tr>
<tr>
<td>3</td>
<td>22</td>
<td>19</td>
</tr>
<tr>
<td>4</td>
<td>15</td>
<td>8</td>
</tr>
</tbody>
</table>

Stored as

<table>
<thead>
<tr>
<th>ID</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>15</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>15</td>
<td>19</td>
</tr>
<tr>
<td>4</td>
<td>22</td>
<td>19</td>
</tr>
</tbody>
</table>
SMBM Performance

Our Solution:
Sorted Multidimensional Bidirectional Map (SMBM)

Example Table

<table>
<thead>
<tr>
<th>ID</th>
<th>X</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>15</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>19</td>
</tr>
<tr>
<td>3</td>
<td>22</td>
<td>19</td>
</tr>
<tr>
<td>4</td>
<td>15</td>
<td>8</td>
</tr>
</tbody>
</table>

Stored as

Can read the entire data structure in parallel in 1 clock cycle

Add and Delete can be issued every clock cycle with a latency of 2 cycles
How to design a **fully reconfigurable** and **fast** filter pipeline?

Can express any arbitrary chain of filter operation with chain length \( \leq k \) on the \( n \) input lines

Runs at line rate
Filter Pipeline Layout

Stage 1

Stage 2

Stage k

Input lines

n

Output lines

n
Reconfigurable Pipeline

Sufficient condition for full reconfigurability

In each stage:
1. Ability to apply any filter operation to an input (or pair of input) line
2. Ability to connect the output of a filter operation to any output line
Single Pipeline Stage Design

Naive design with both K-UFPUs and BFPUs

n X 2n Crossbar

K-UFPU
K-UFPU
K-UFPU
K-UFPU
BFPU
BFPU

n/2 X n Crossbar

High overhead

n = 4 Input lines

n = 4 Output lines
Single Pipeline Stage Design

Our design with both K-UFPUs and BFPUs

Cell

Input lines

Output lines

n = 4

n X n Crossbar
Apply a UFPU op on Input 1 and connect to Output 3.
Example Configuration

Apply a UFPU op on Input 1 and connect to Output 3

n = 4
Input lines

n = 4
Output lines

n X n
Crossbar
Example Configuration

Apply a BFPU op on Input 1 and 3 connect to Output 2

Input lines

Output lines

n X n Crossbar

Output lines

Input lines

n = 4

n = 4
Example Configuration

Apply a BFPU op on Input 1 and 3 and connect to Output 2

n = 4
Input lines

n = 4
Output lines

n X n
Crossbar

Cell

no-op
BFPU

no-op
BFPU

K-UFPUS
BFPU

K-UFPUS
BFPU
Filter Pipeline Performance

• Filter pipeline can process a new filter request every clock cycle

• Our implementation runs at clock speeds in excess of 1 GHz
  • 1 GHz is the typical clock speed of today’s switches

• However, scalability is limited to a few 1000s of table entries
  • …beyond that the clock speed falls below 1 GHz
  • Still sufficient for many applications where table entries include network paths, switch ports, servers in a cluster, etc.
Application Performance

Load balancing client requests

**Policy 1.** Select a server uniformly at random.

**Policy 2.** Select a server uniformly at random from the set of servers with CPU utilization $< X$ and available memory $> Y$ and available bandwidth $> Z$. If the filtered set is empty, select a server uniformly at random from the entire set.

Response time for Policy 2 normalized w.r.t. Policy 1

1.7x—1.3x better response time for 70% of client requests
Application Performance

In-network caching of relational graph filter queries

Response time with caching normalized w.r.t. no caching

4x—2.8x better response time for cached filter queries
Summary

• **Thanos** extends programmable switches with the ability to do programmable line rate filtering over a multi-dimensional table

• Use cases include performance-aware routing, load balancing, network diagnosis, security, firewall…

• The design runs in excess of 1 GHz clock speed and scales to thousands of table entries

• Evaluations show up to 1.7x improvement in performance of key network functions, such as routing and load balancing

• Overall, advances the current state of in-network computing
  • enables **richer** algorithms for existing in-network applications
  • enables **new** in-network applications, e.g., caching of relational queries
Thank you!