

Dr. Top- k : Delegate-Centric Top- k on GPUs

Anil Gaihre Da Zheng Scott Weitze Lingda Li Shuaiwen Leon Song
Caiwen Ding Xiaoye S Li Hang Liu



THE UNIVERSITY OF
SYDNEY



November 16, 2021



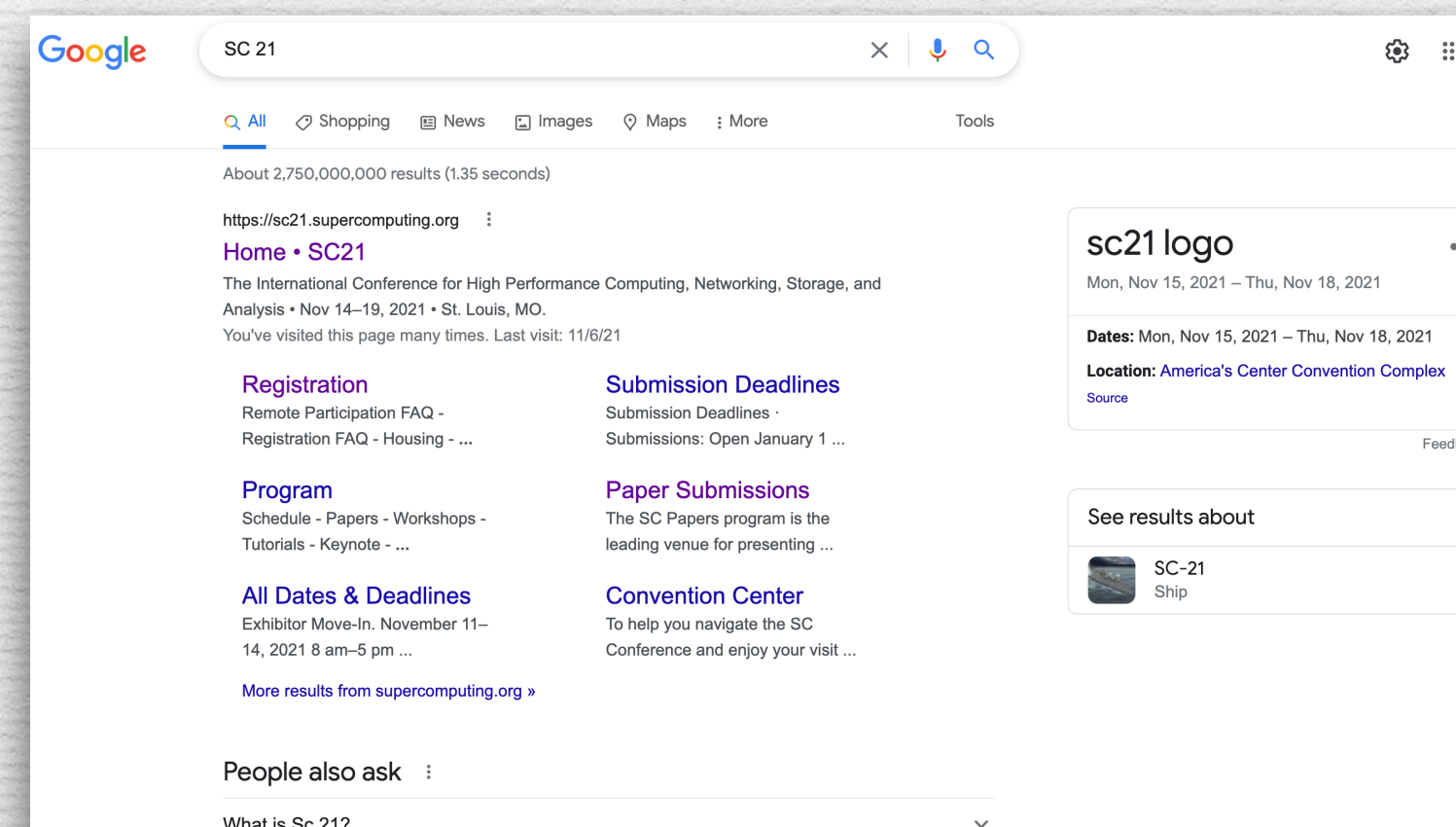
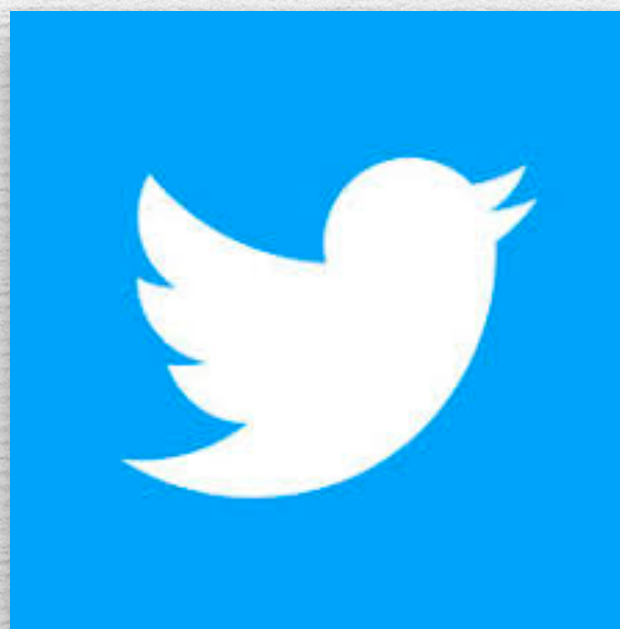
Outline

- ❖ Top- k Background
- ❖ State-of-the-art methods
 - ✓ Sort and choose
 - ✓ Top- k methods: Bucket, Bitonic top- k
- ❖ Dr.Top- k Workflow
- ❖ Dr.Top- k : Maximum Delegate Top- k
- ❖ Dr.Top- k : Optimizations
- ❖ Evaluations: CUDA platform
- ❖ Conclusion

Top- k computation is everywhere!

Given a sequence of data, selecting the **top K number of elements** out it.

Applications



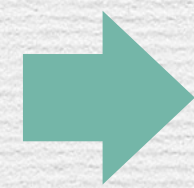
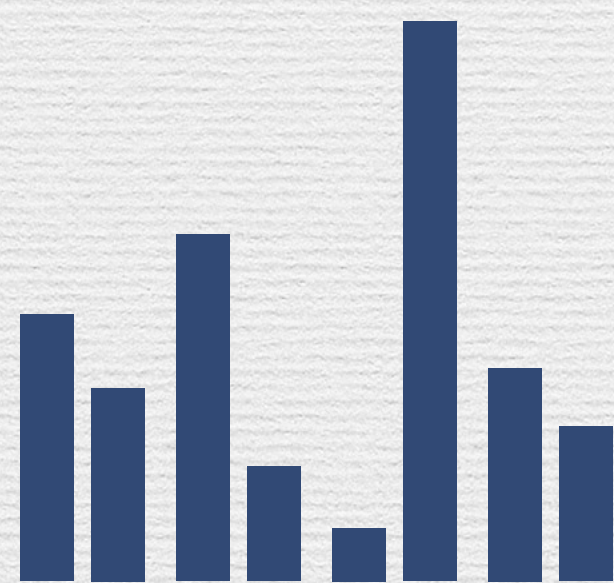
...

State-of-the-art GPUs Designs

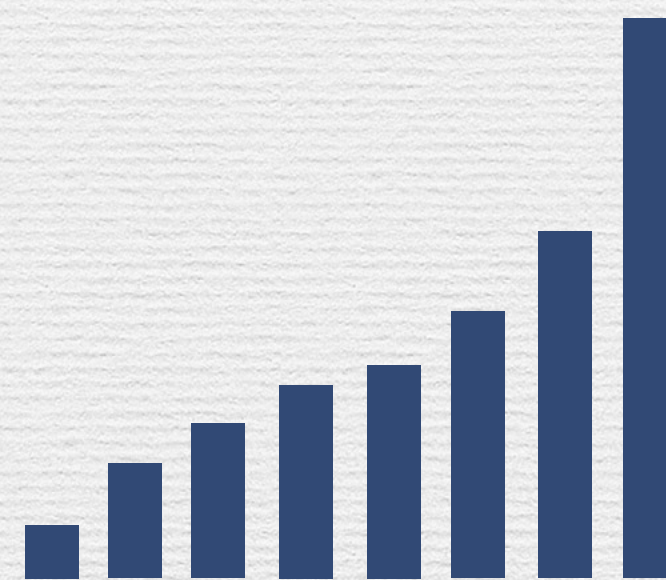
Compute top-2

- ♦ Existing solution #1: Sort and choose

Question: Find the top 2 largest numbers from the sequence below.



Sorting

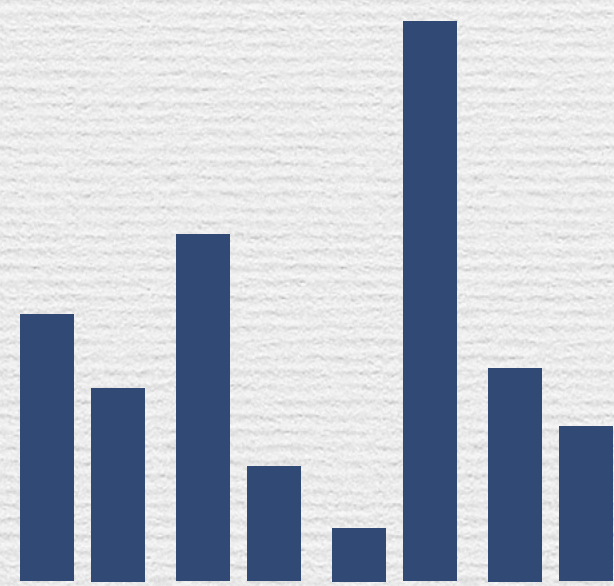


State-of-the-art GPUs Designs

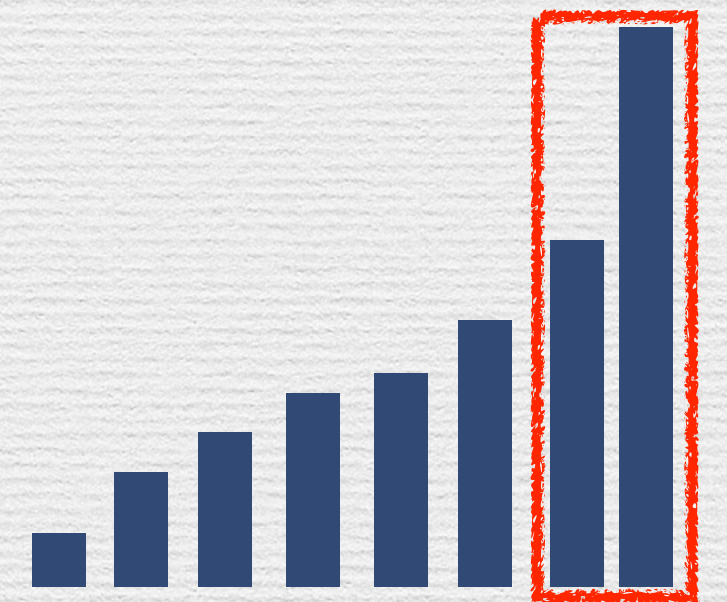
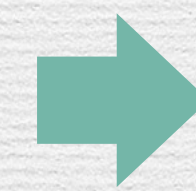
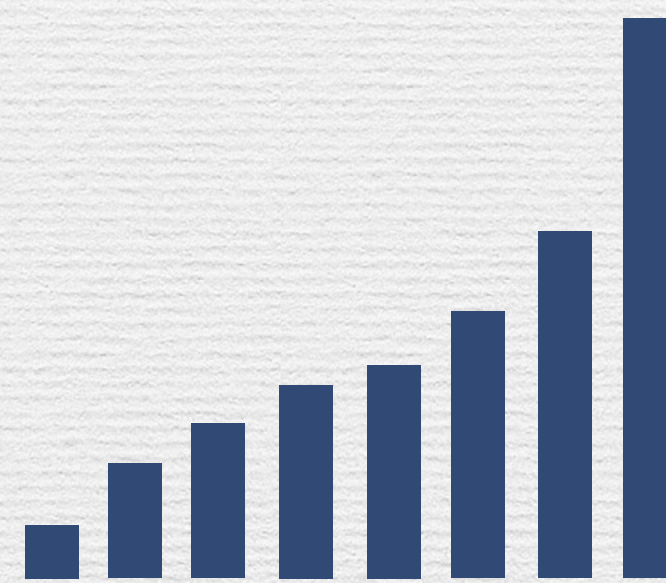
Compute top-2

- ♦ Existing solution #1: Sort and choose

Question: Find the top 2 largest numbers from the sequence below.



→
Sorting



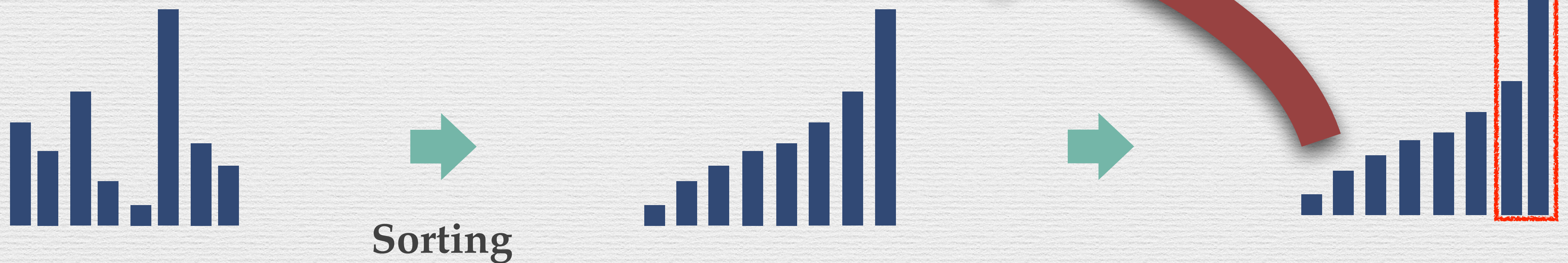
State-of-the-art GPUs Designs

Compute top-2

- ♦ Existing solution #1: Sort and choose

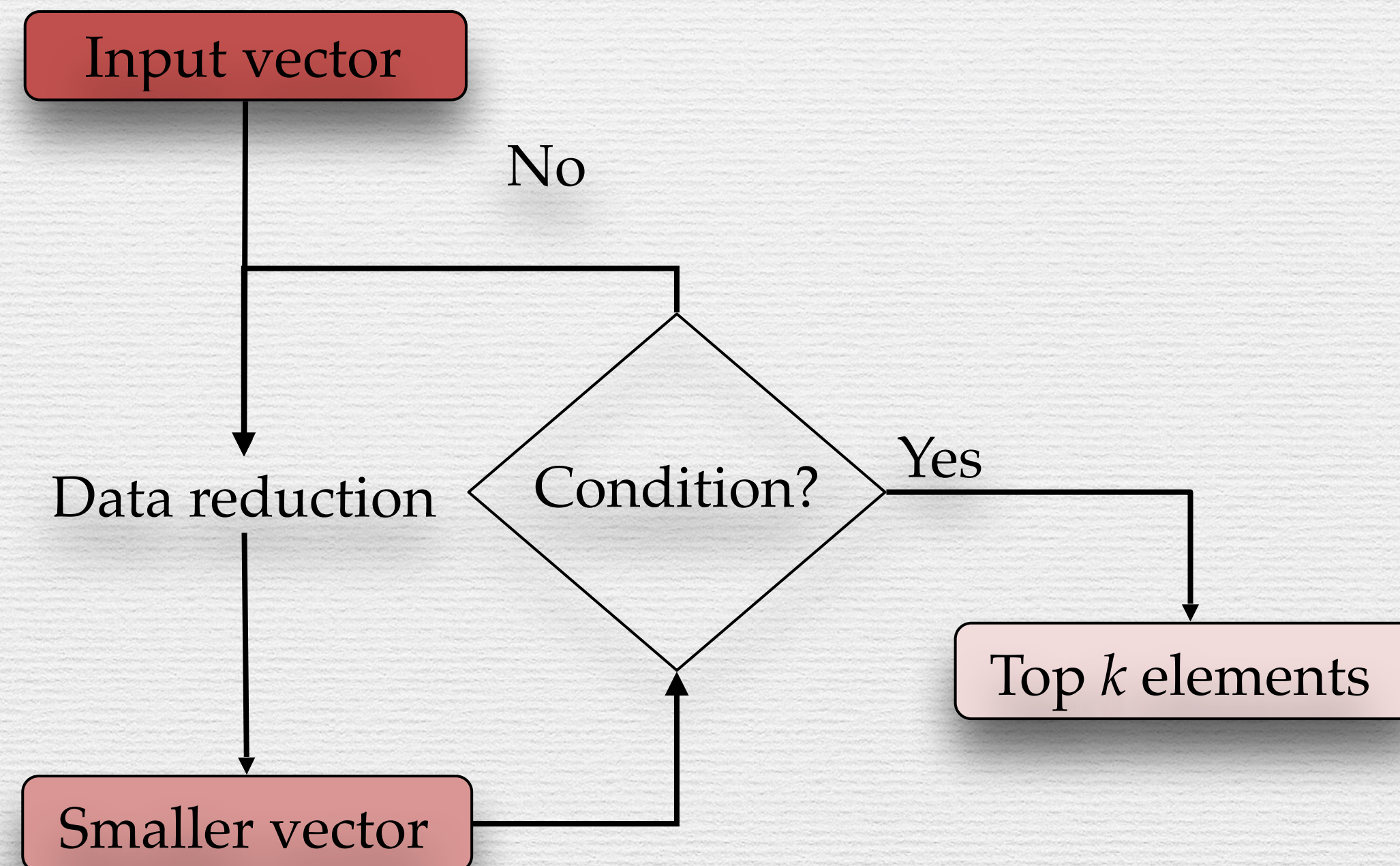
Question: Find the top 2 largest numbers from the sequence below.

Does unnecessary work!



State-of-the-art GPUs Designs

- Existing solution #2: Top-k algorithms

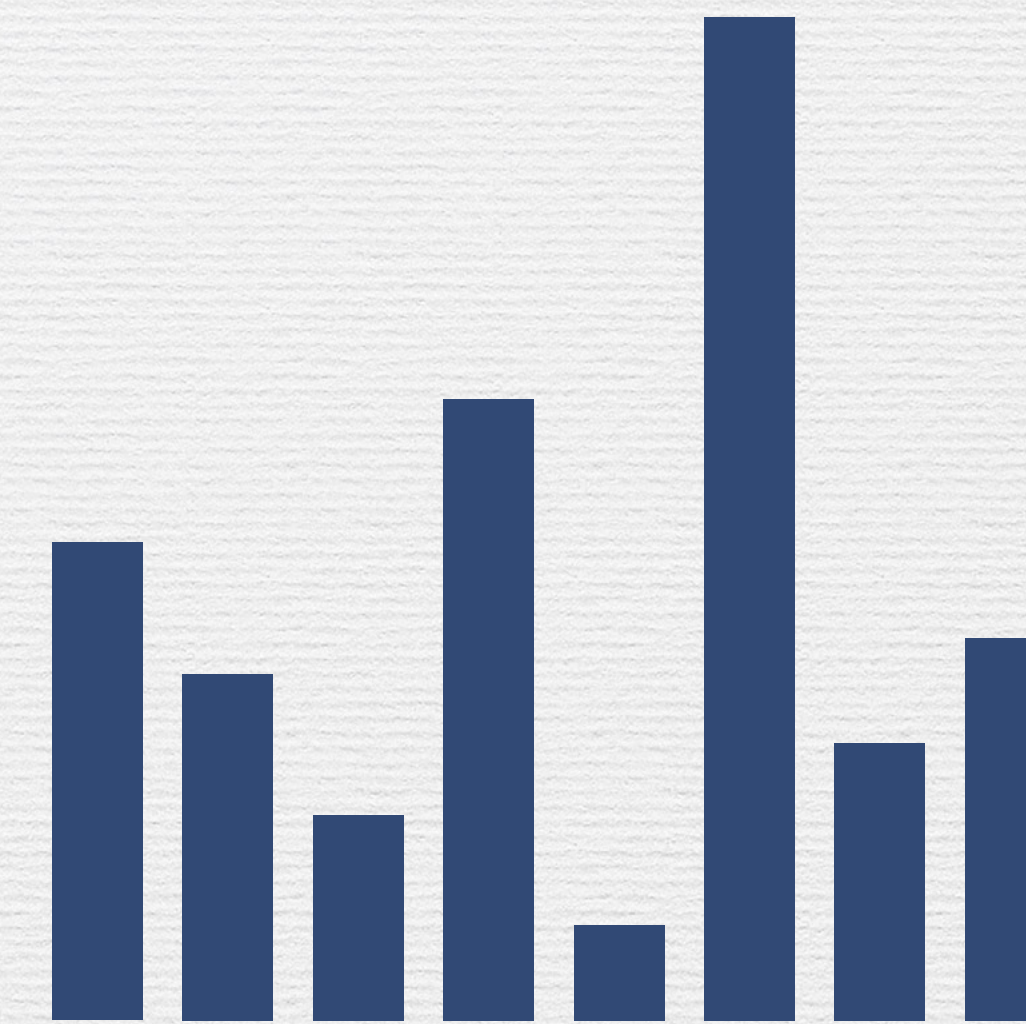
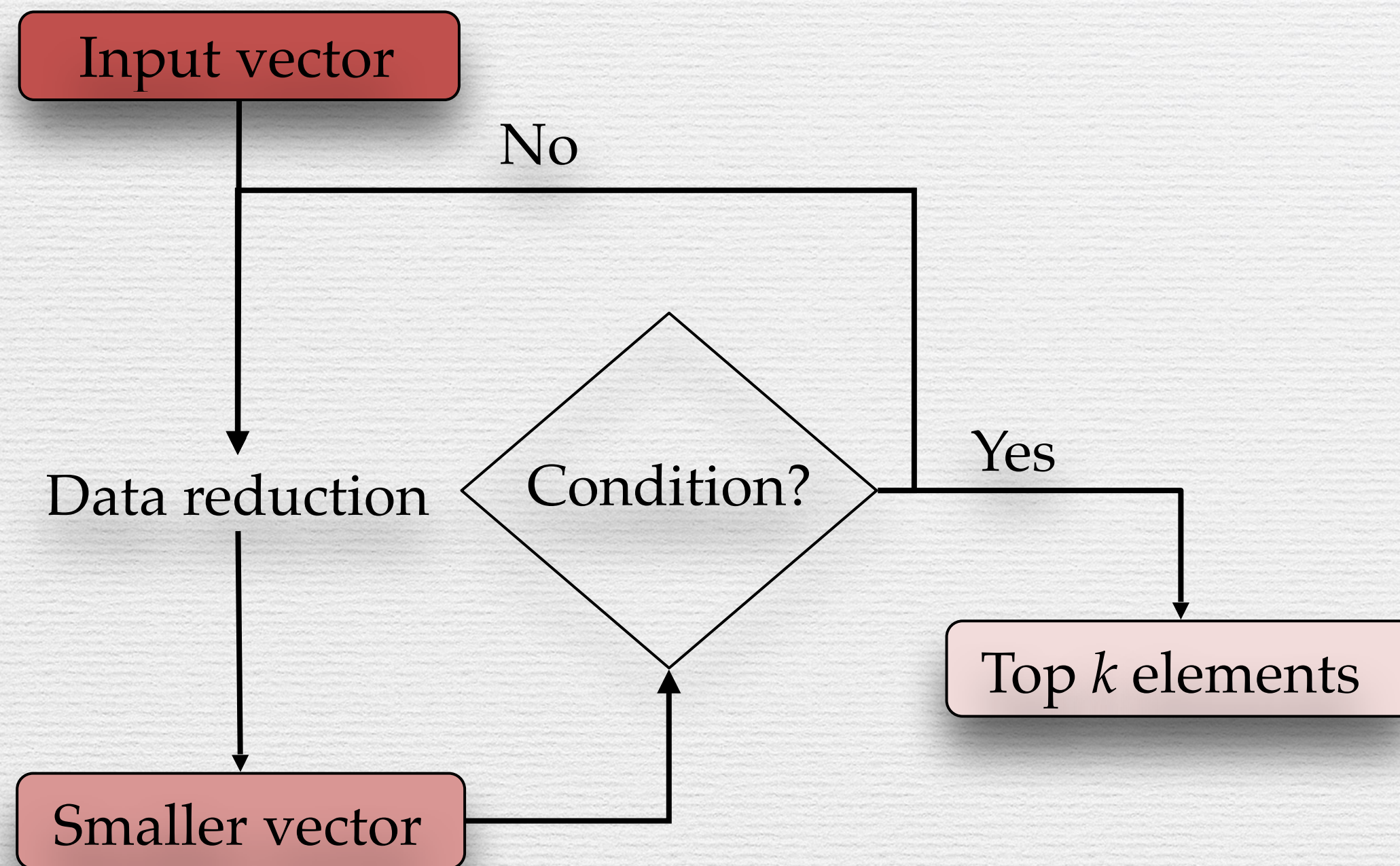


Bucket/radix/bitonic top-k workflow

State-of-the-art GPUs Designs

Bitonic Top- k

Compute top-2



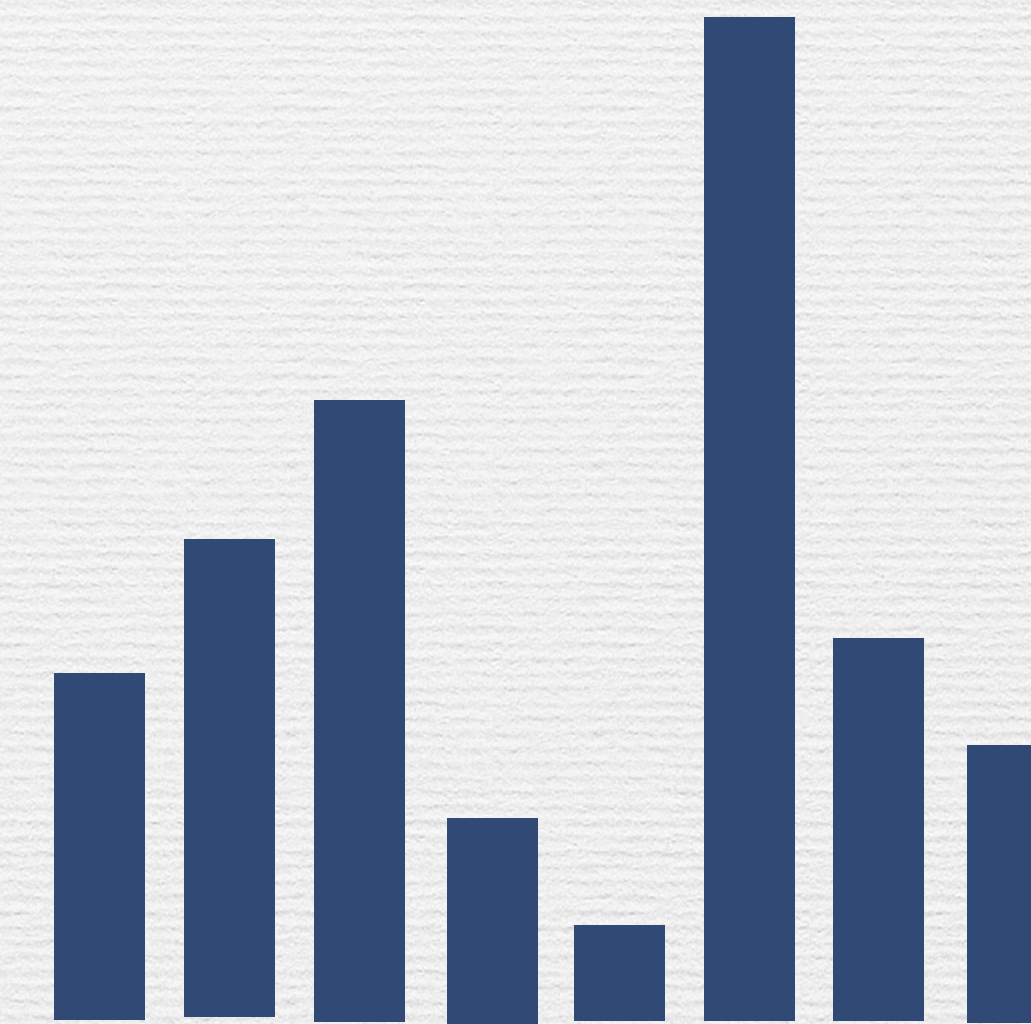
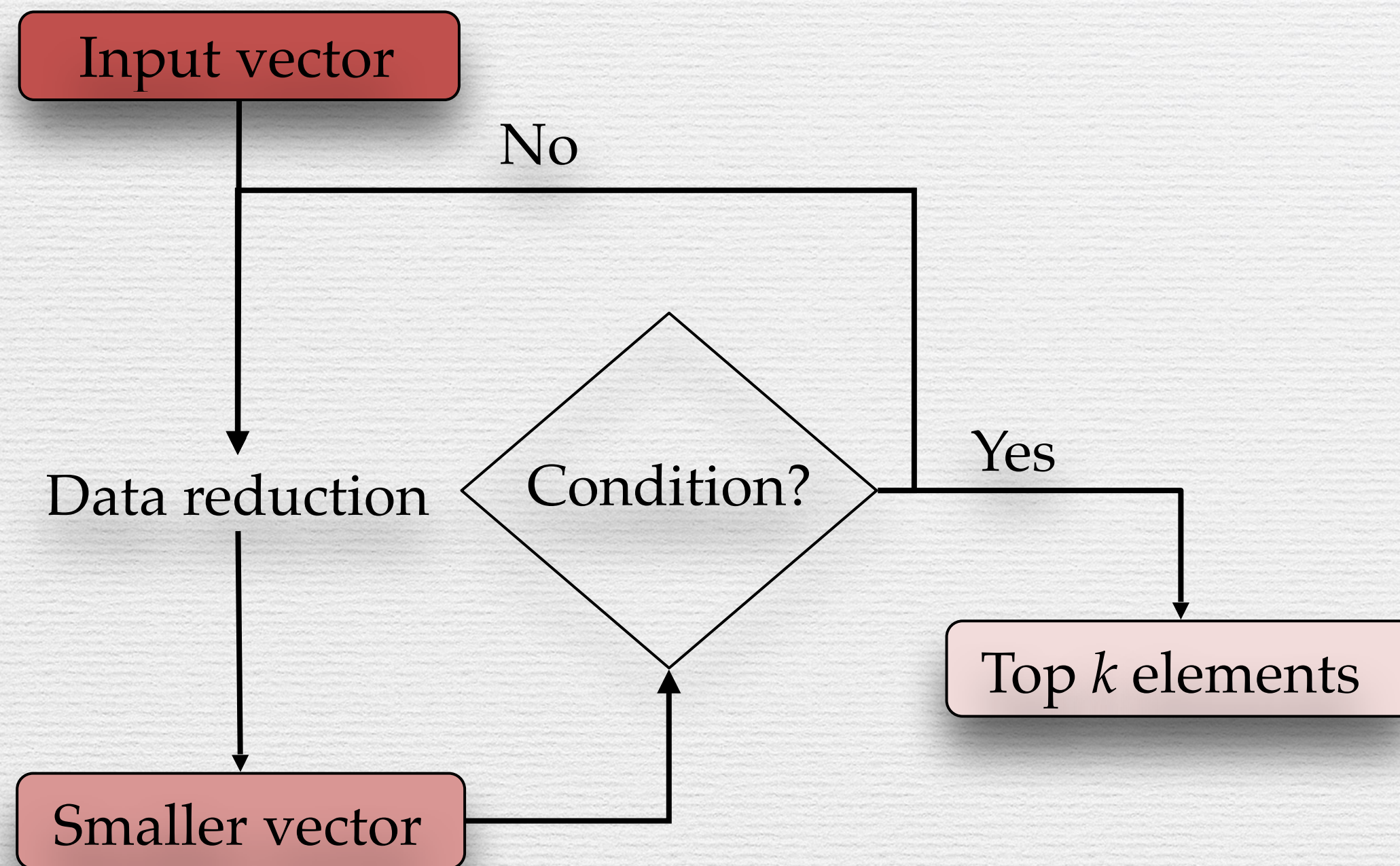
Iteration 1

Bucket/radix/bitonic top- k workflow

State-of-the-art GPUs Designs

Bitonic Top- k

Compute top-2



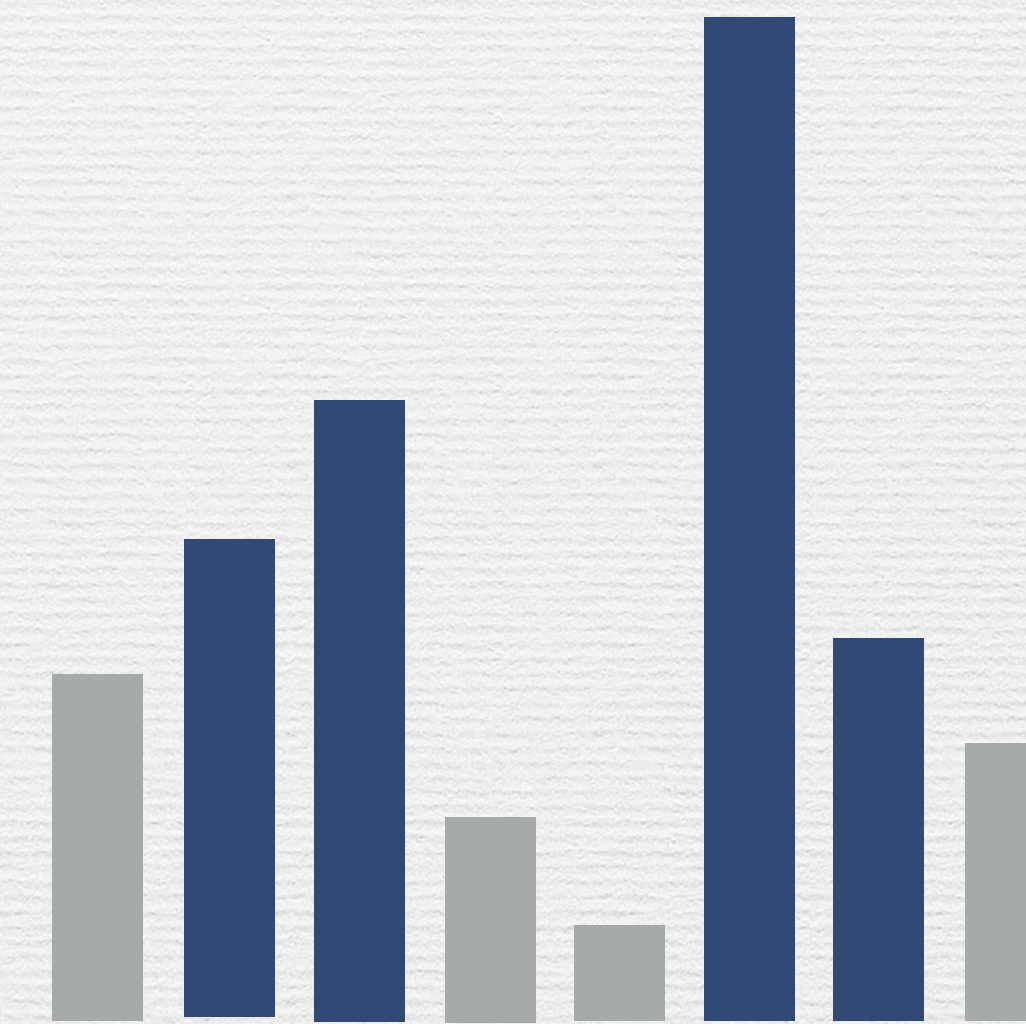
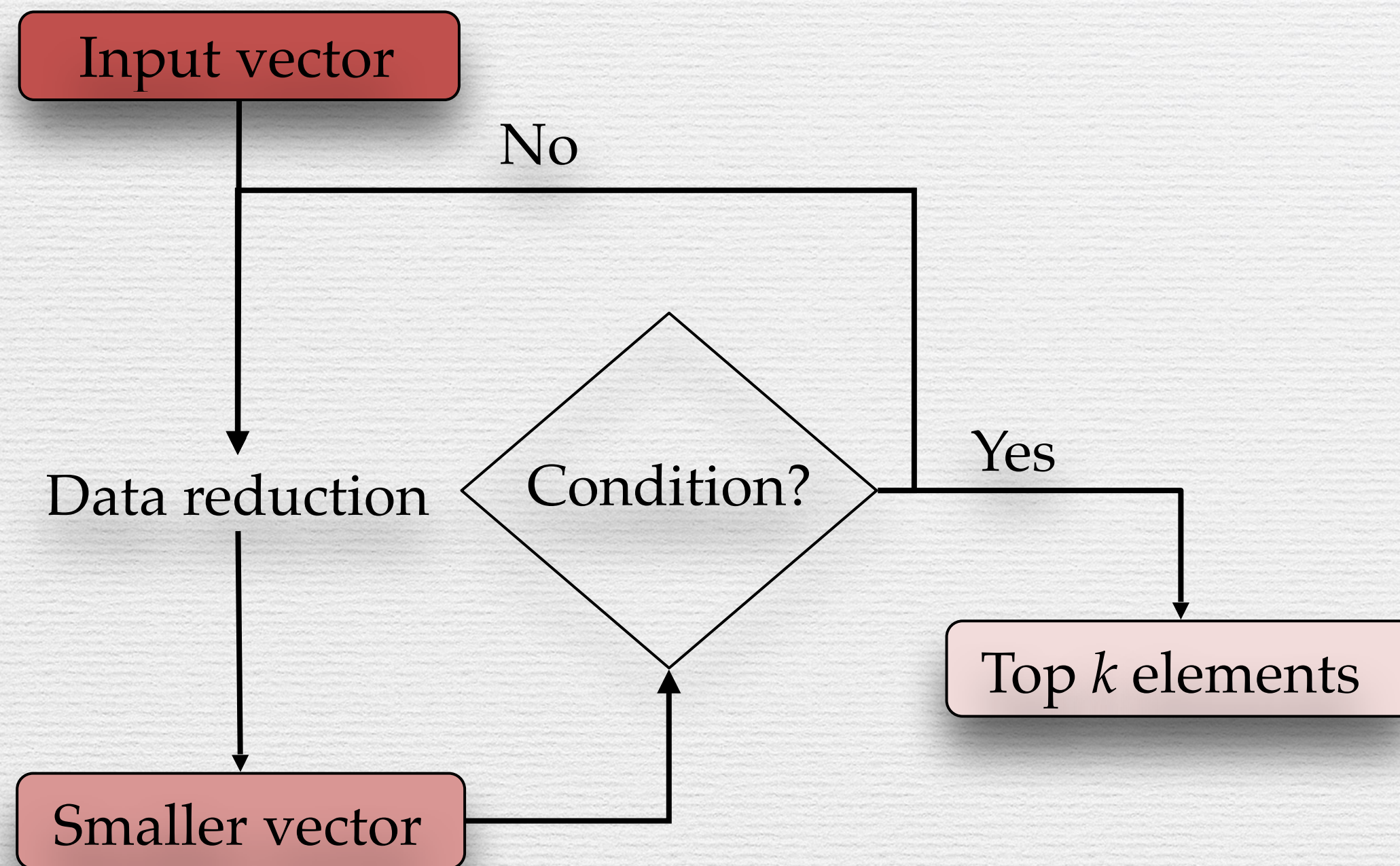
Iteration 1

Bucket/radix/bitonic top- k workflow

State-of-the-art GPUs Designs

Bitonic Top- k

Compute top-2



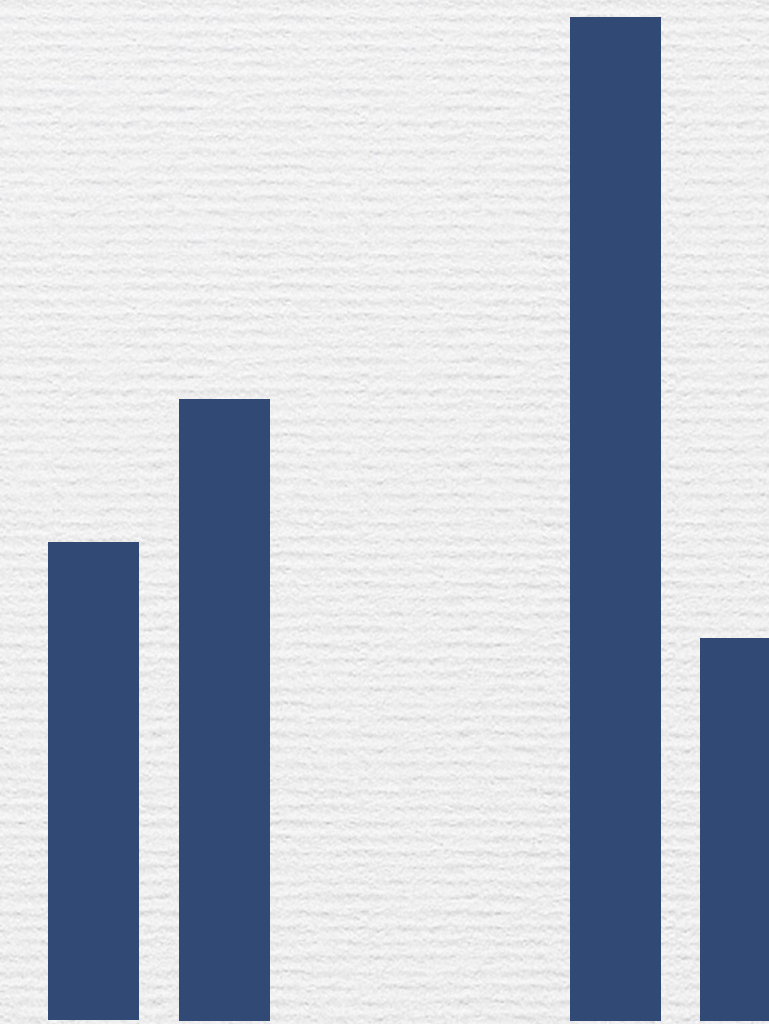
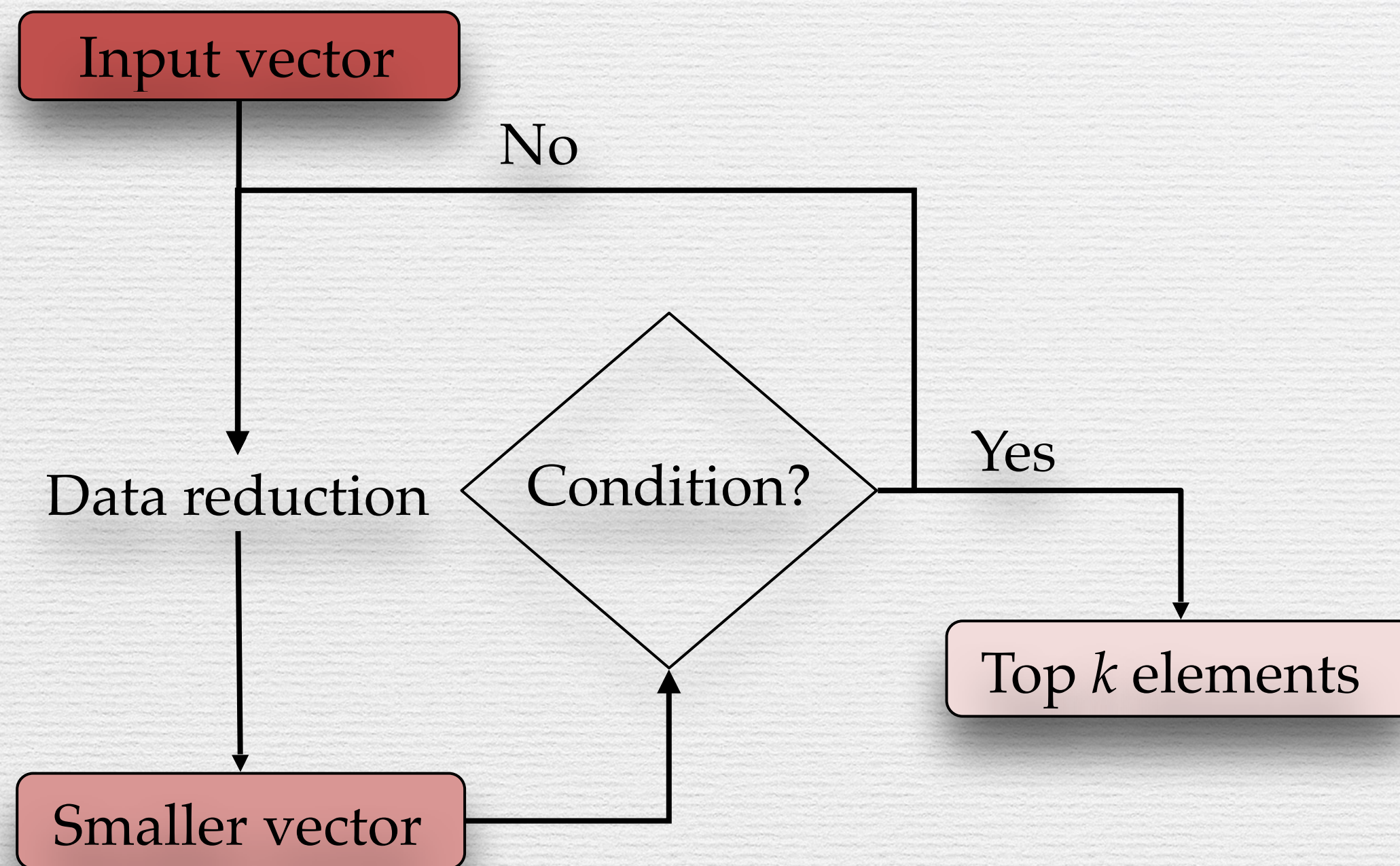
Iteration 1

Bucket/radix/bitonic top- k workflow

State-of-the-art GPUs Designs

Bitonic Top- k

Compute top-2



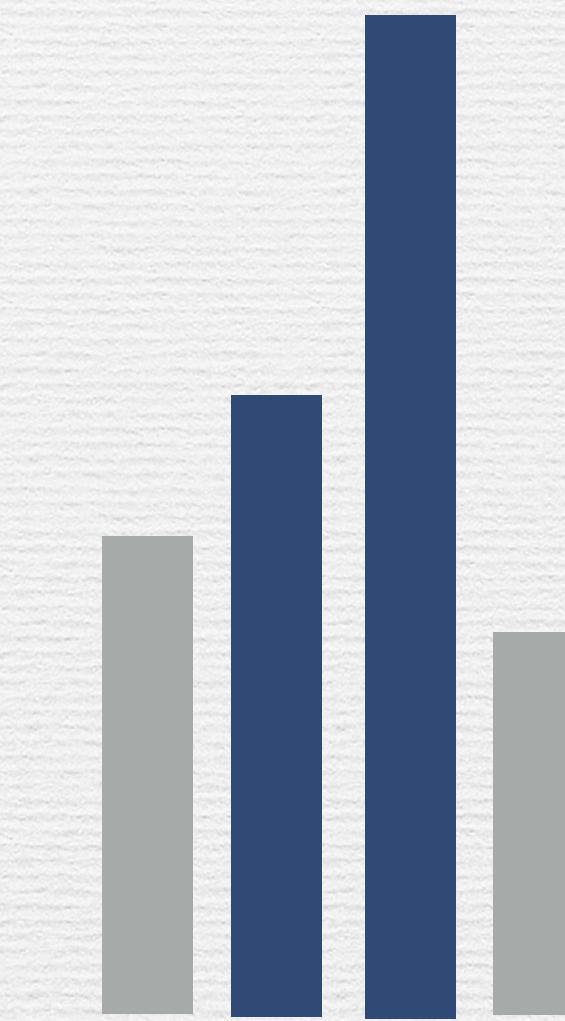
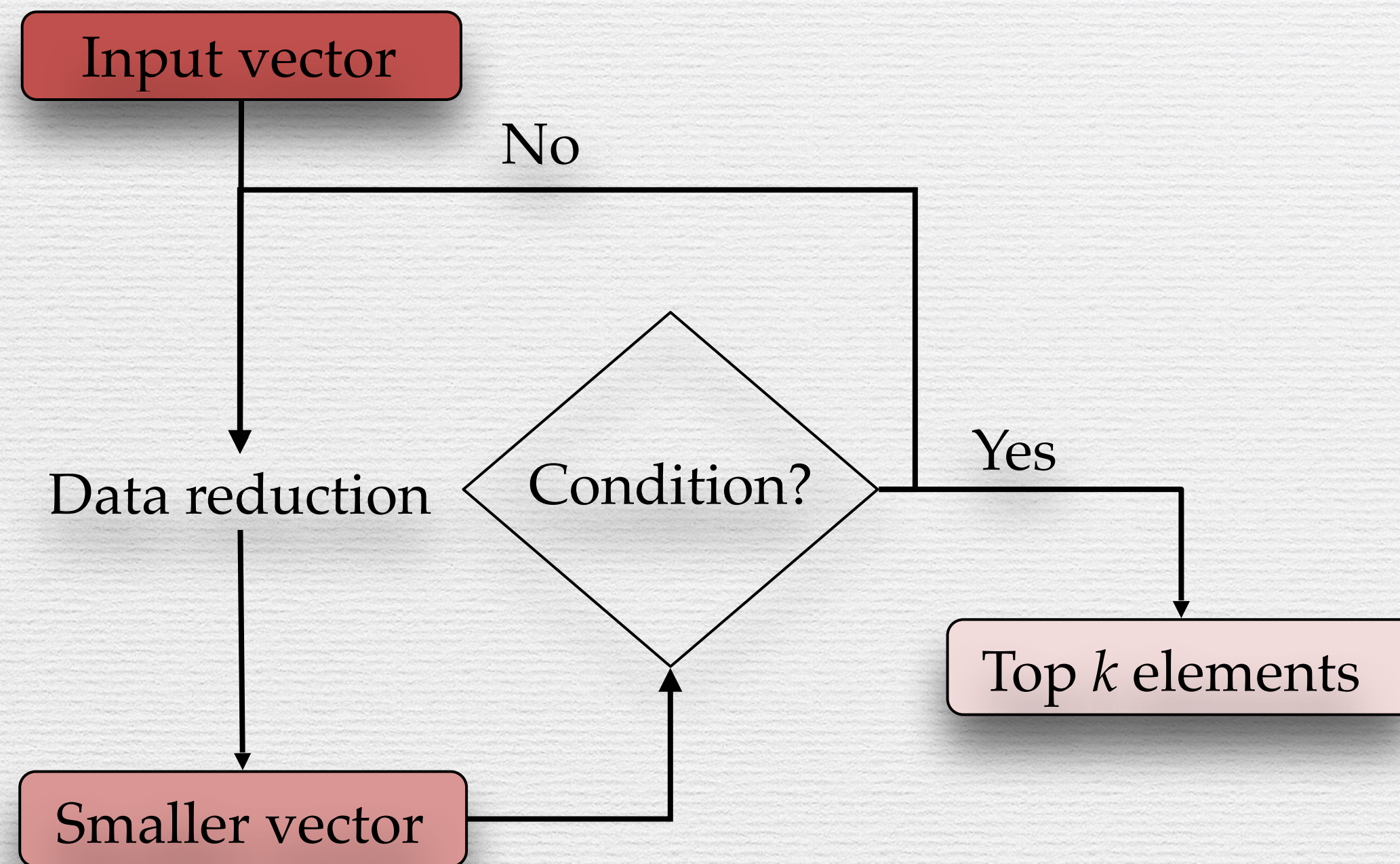
Iteration 1

Bucket/radix/bitonic top- k workflow

State-of-the-art GPUs Designs

Bitonic Top- k

Compute top-2



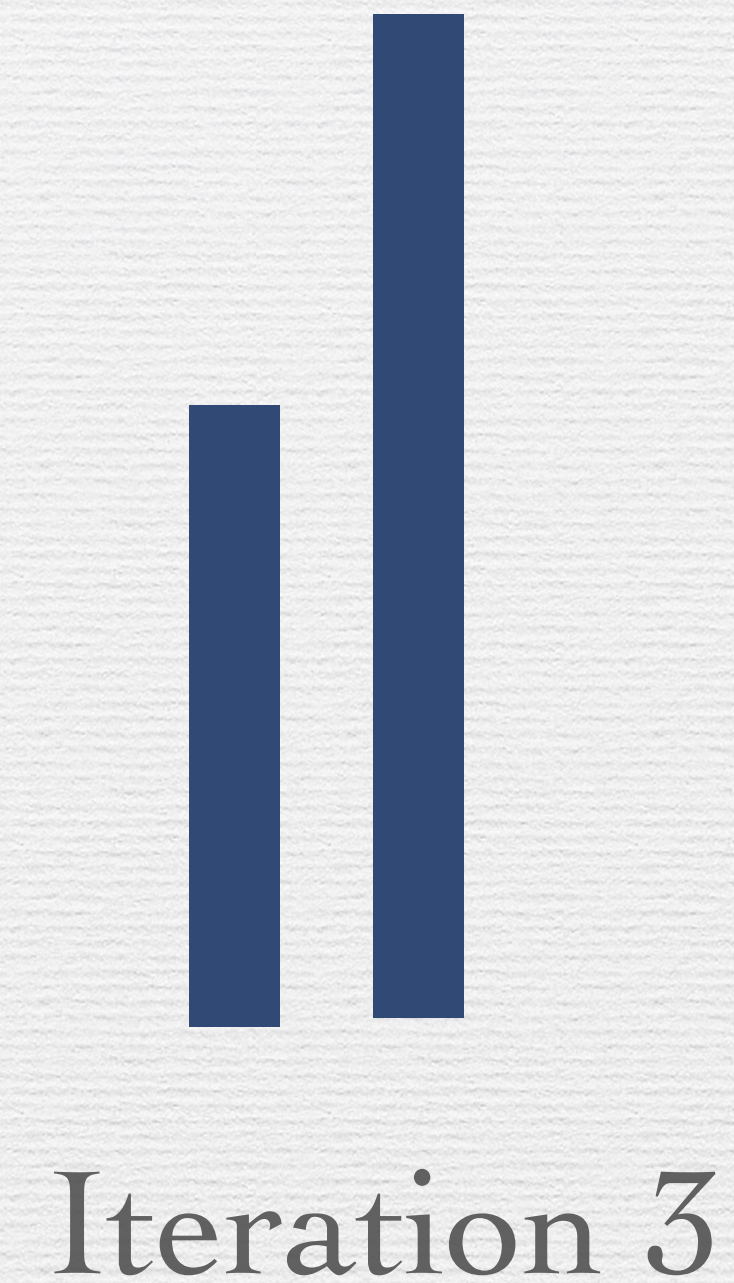
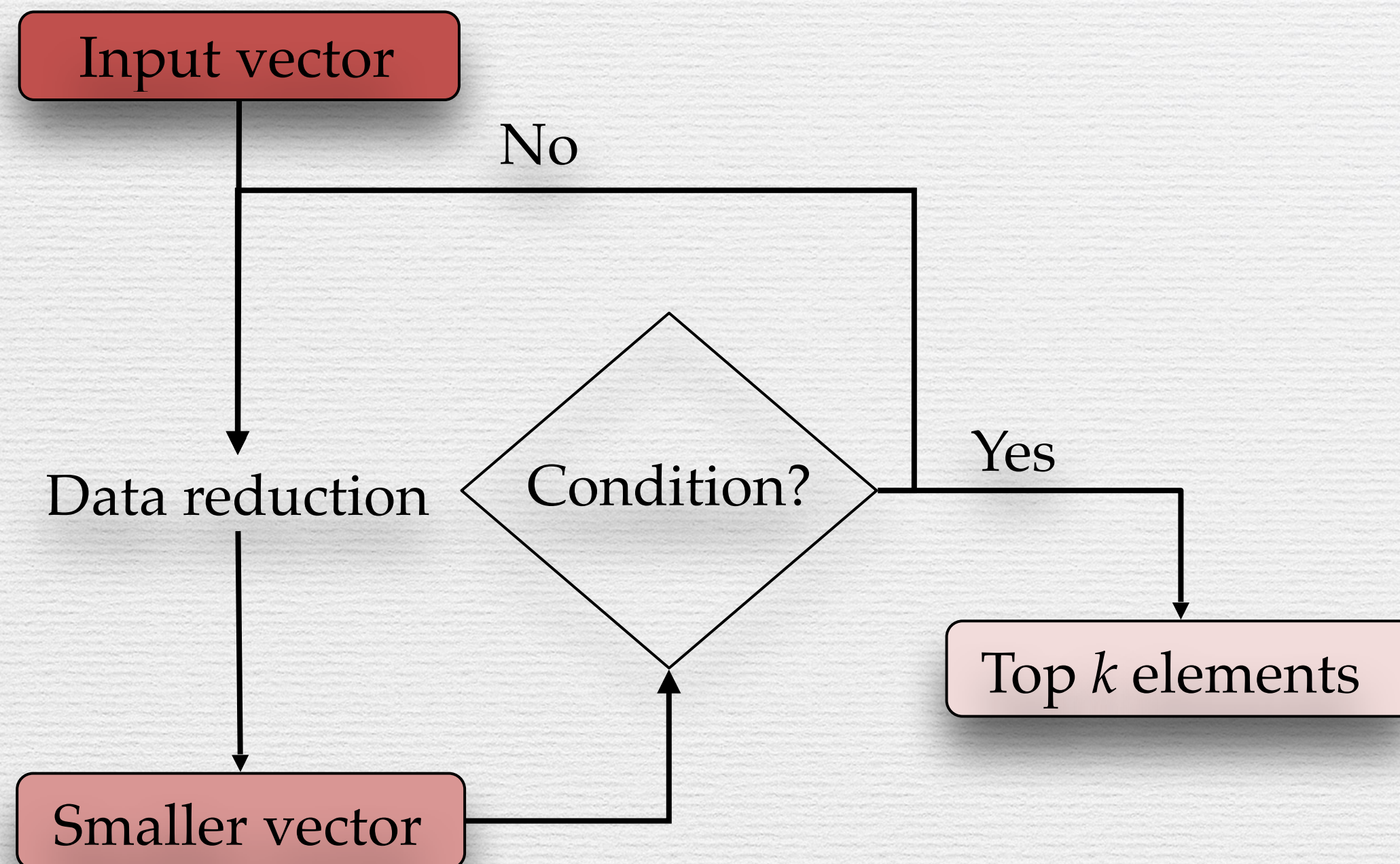
Iteration 2

Bucket/radix/bitonic top- k workflow

State-of-the-art GPUs Designs

Bitonic Top- k

Compute top-2

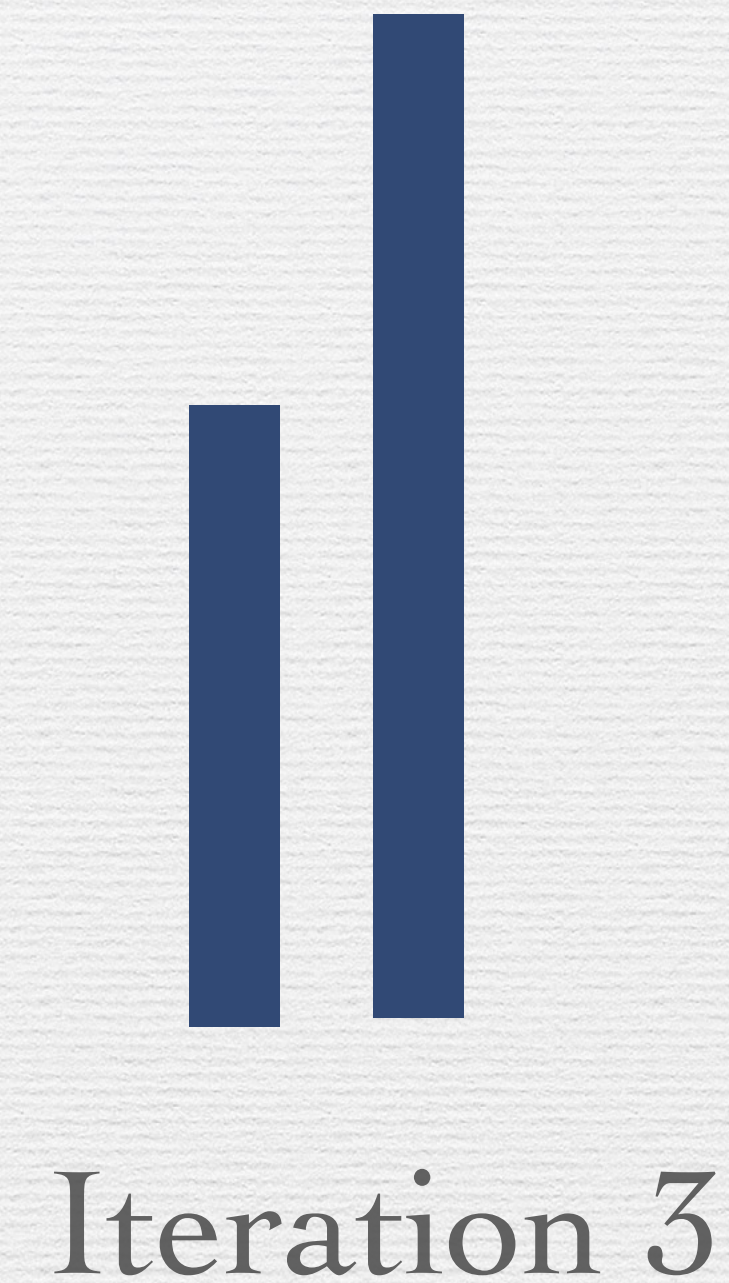
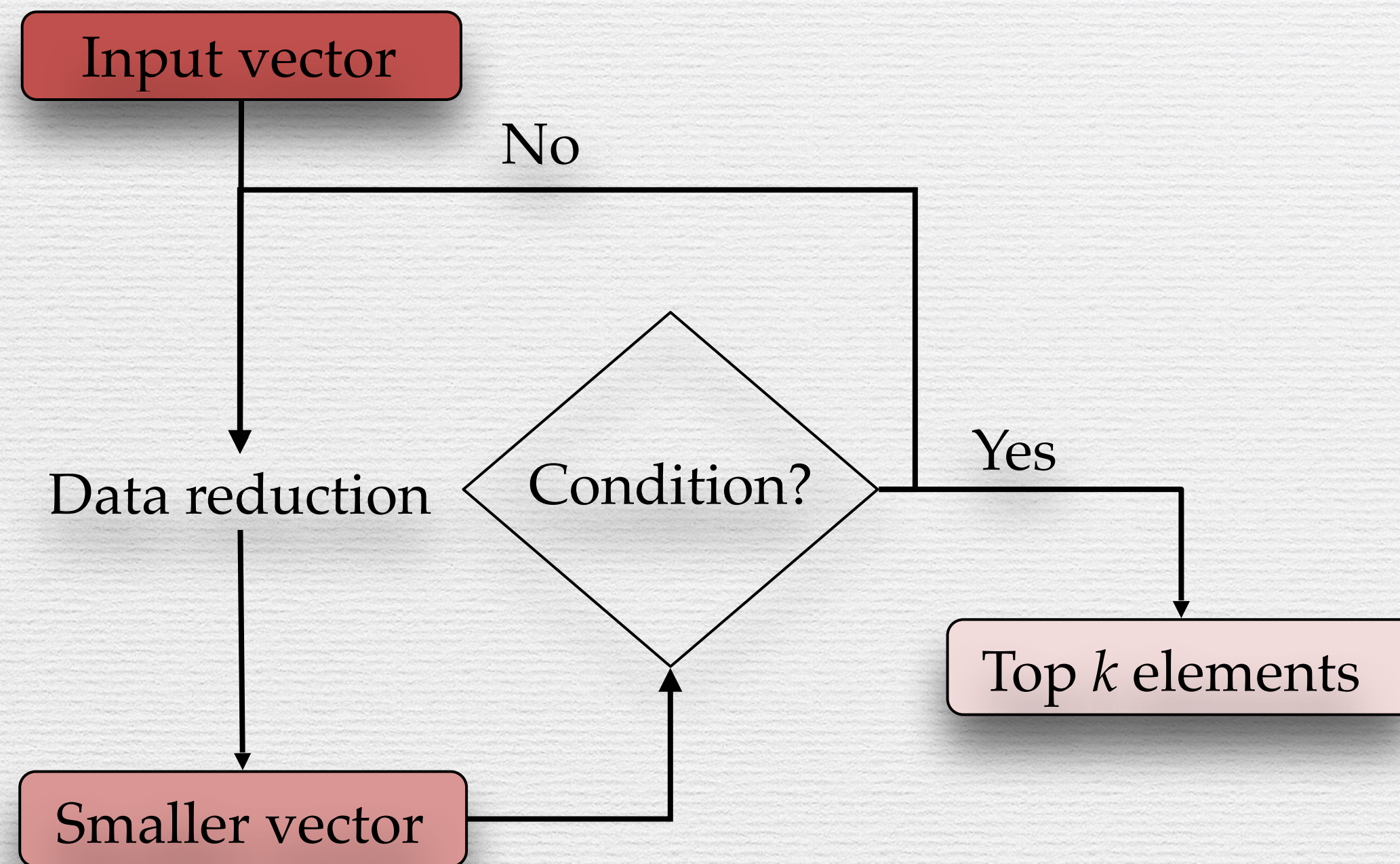


Bucket/radix/bitonic top- k workflow

State-of-the-art GPUs Designs

Bitonic Top- k

Compute top-2



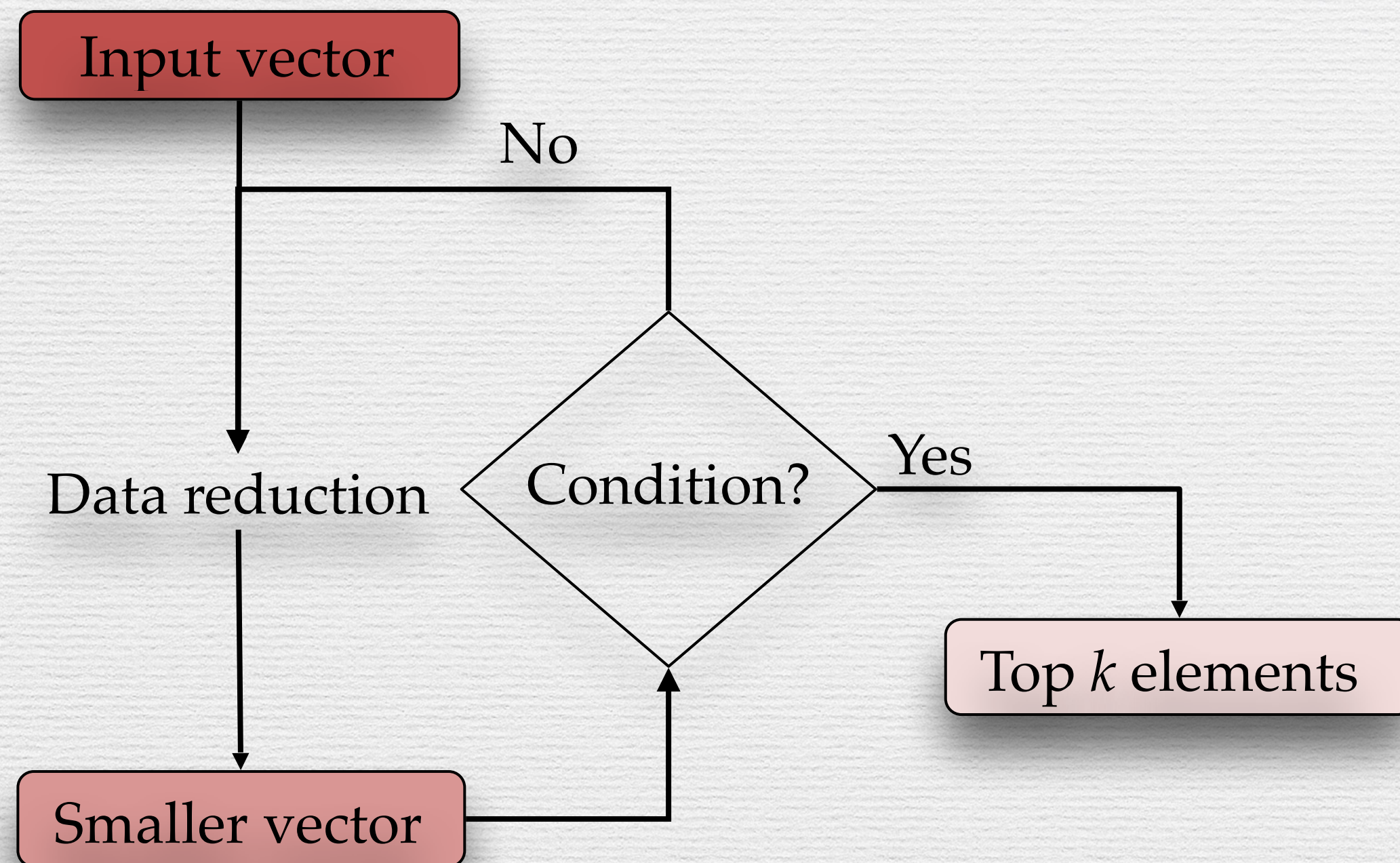
Bucket/radix/bitonic top- k workflow

Workload reduction is too low!

State-of-the-art GPUs Designs

Bucket Top- k

Compute top-2



Bucket/radix/bitonic top- k workflow

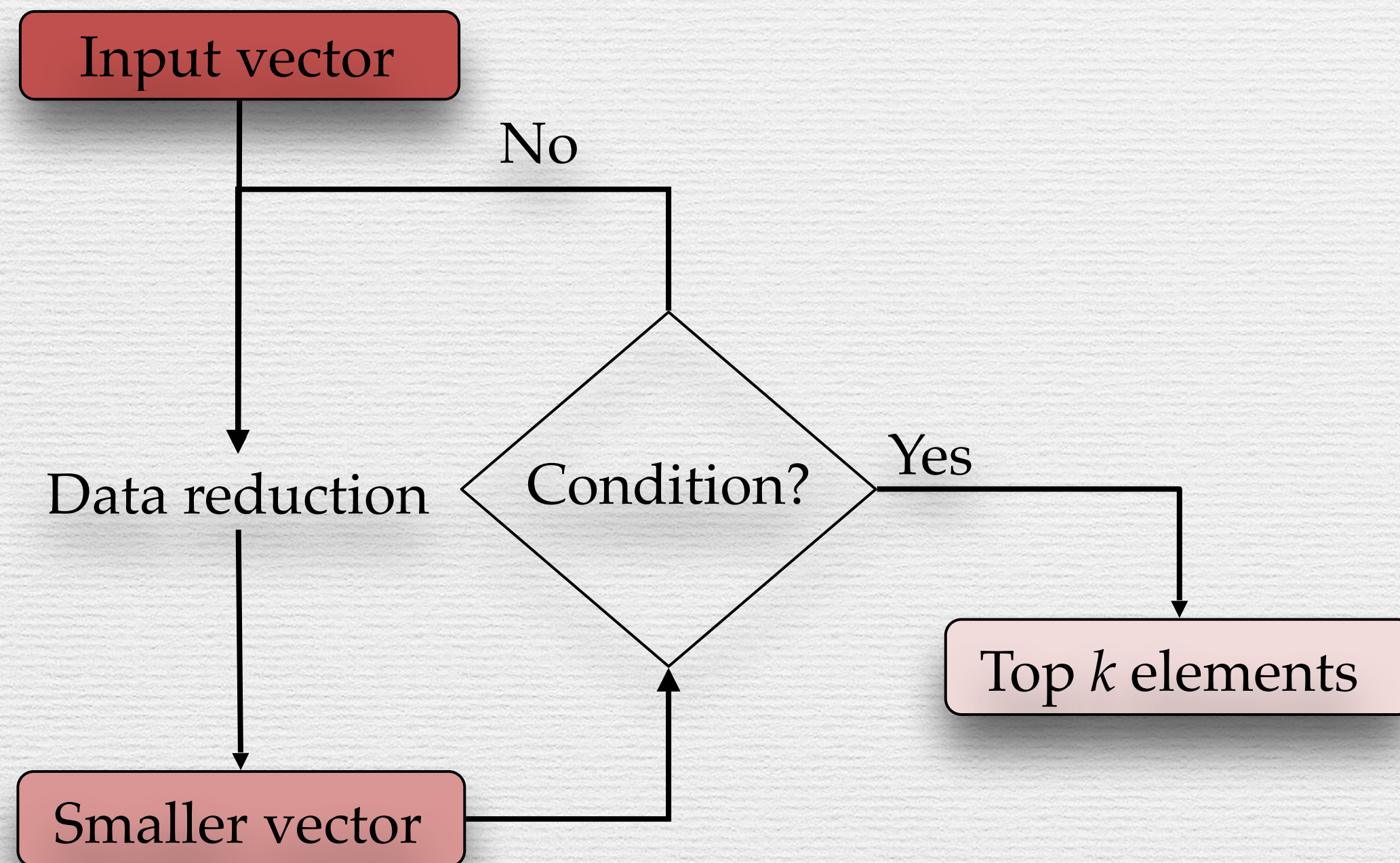
State-of-the-art GPUs Designs

Bucket Top- k

Compute top-2

Input vector

2001	101	1323	3012	212	1132	2310	2313	3000	3010	1002	3210	1020	333	2321	2003
------	-----	------	------	-----	------	------	------	------	------	------	------	------	-----	------	------



Bucket/radix/bitonic top- k workflow

State-of-the-art GPUs Designs

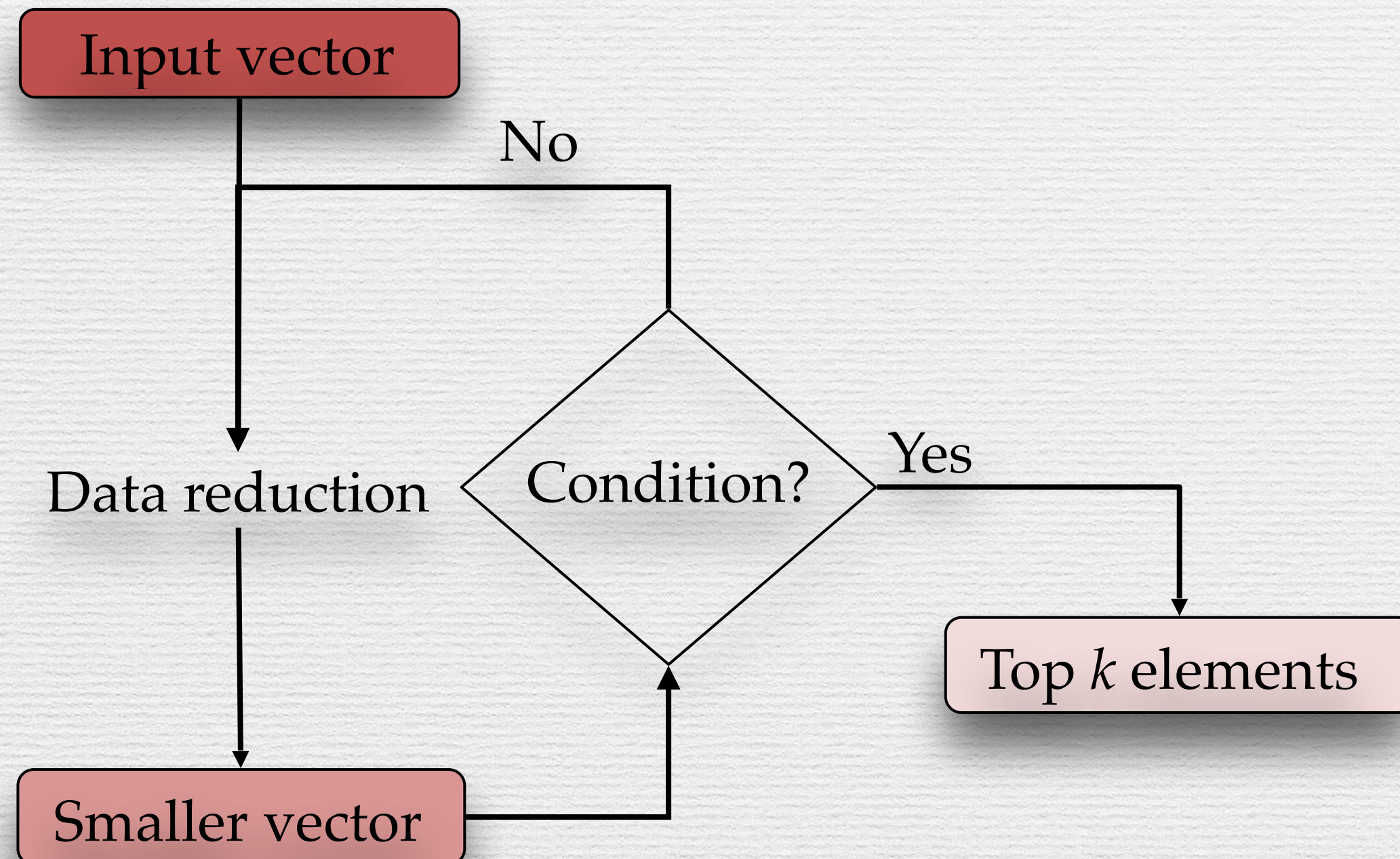
Bucket Top- k

Compute top-2

Input vector

2001	101	1323	3012	212	1132	2310	2313	3000	3010	1002	3210	1020	333	2321	2003
------	-----	------	------	-----	------	------	------	------	------	------	------	------	-----	------	------

Iteration 1			
[101, 878.25)	[878.25, 1655.5)	[1655.5, 2432.75)	[2432.75 3210]
			



Bucket/radix/bitonic top- k workflow

Bucket top-2 computation for an input vector with 16 elements.

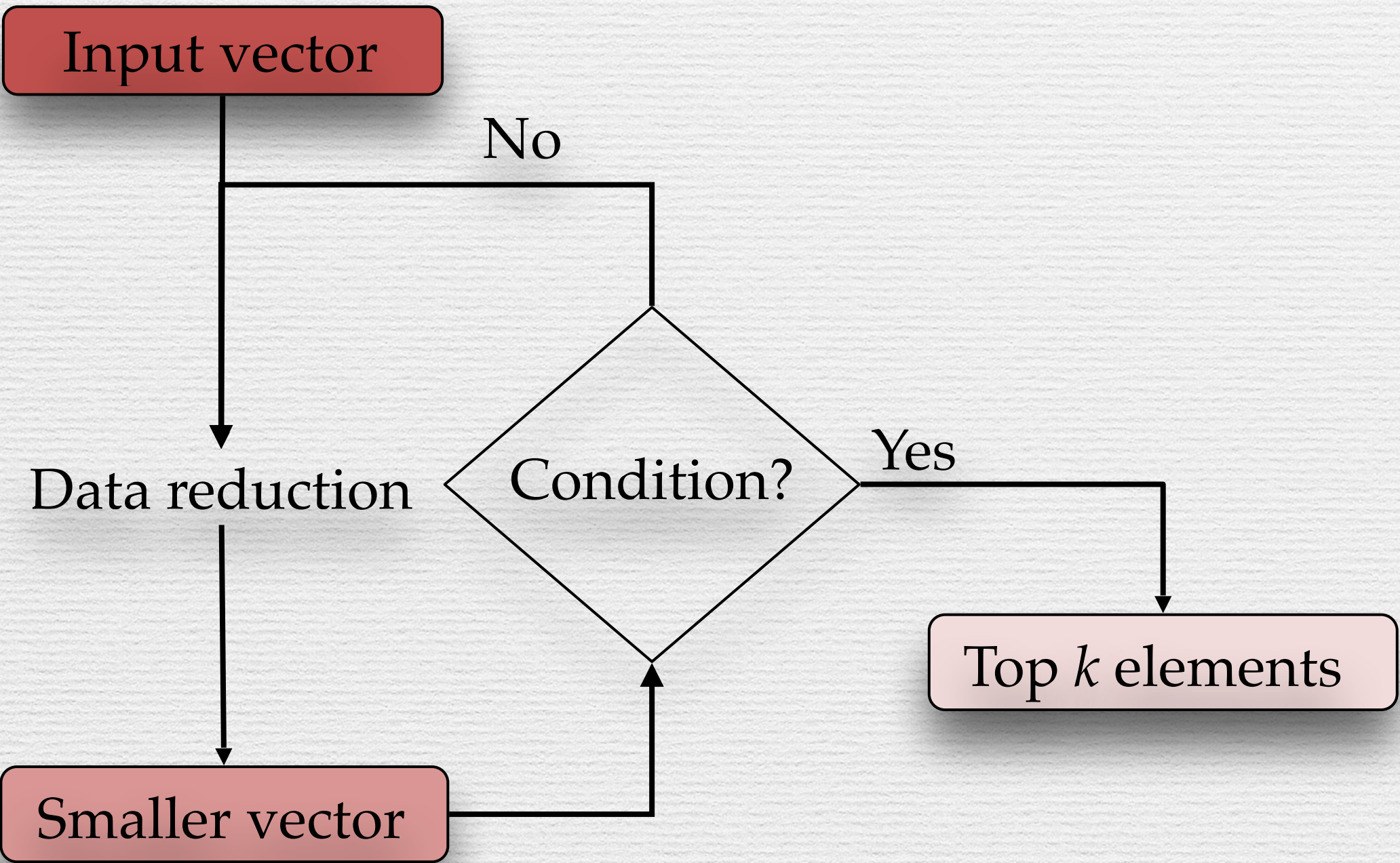
State-of-the-art GPUs Designs

Bucket Top- k

Compute top-2

Input vector

2001	101	1323	3012	212	1132	2310	2313	3000	3010	1002	3210	1020	333	2321	2003
------	-----	------	------	-----	------	------	------	------	------	------	------	------	-----	------	------



Bucket/radix/bitonic top- k workflow

Iteration 1			
[101, 878.25)	[878.25, 1655.5)	[1655.5, 2432.75)	[2432.75, 3210]
			
101	1323	2001	3012
212	1132	2310	3000
333	1002	2313	3010
		2321	3210

Bucket top-2 computation for an input vector with 16 elements. The highlighted bucket is of interest at each iteration.

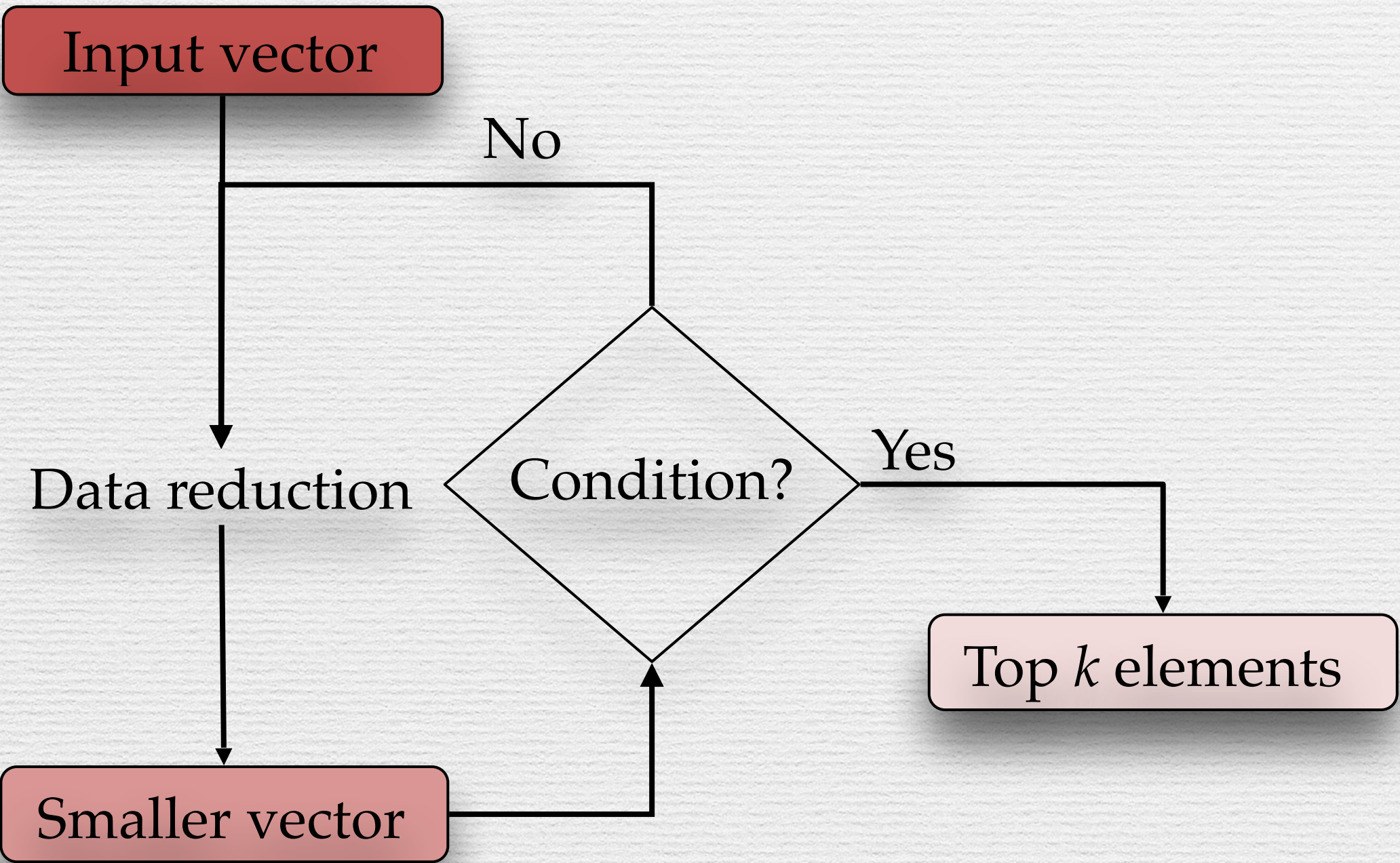
State-of-the-art GPUs Designs

Bucket Top- k

Compute top-2

Input vector

2001	101	1323	3012	212	1132	2310	2313	3000	3010	1002	3210	1020	333	2321	2003
------	-----	------	------	-----	------	------	------	------	------	------	------	------	-----	------	------



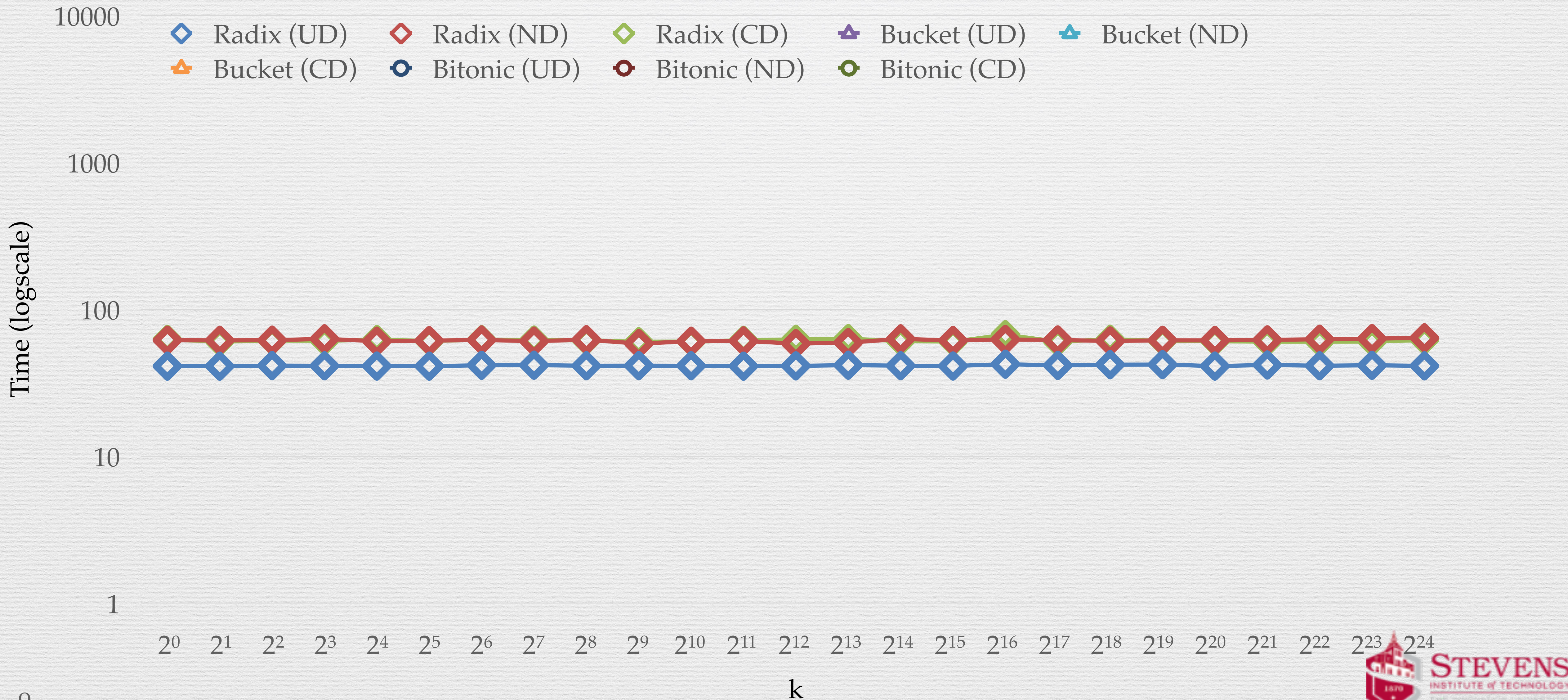
Bucket/radix/bitonic top- k workflow

Iteration 1			
[101, 878.25)	[878.25, 1655.5)	[1655.5, 2432.75)	[2432.75, 3210]
			
101	1323	2001	3012
212	1132	2310	3000
333	1002	2313	3010
		2321	3210

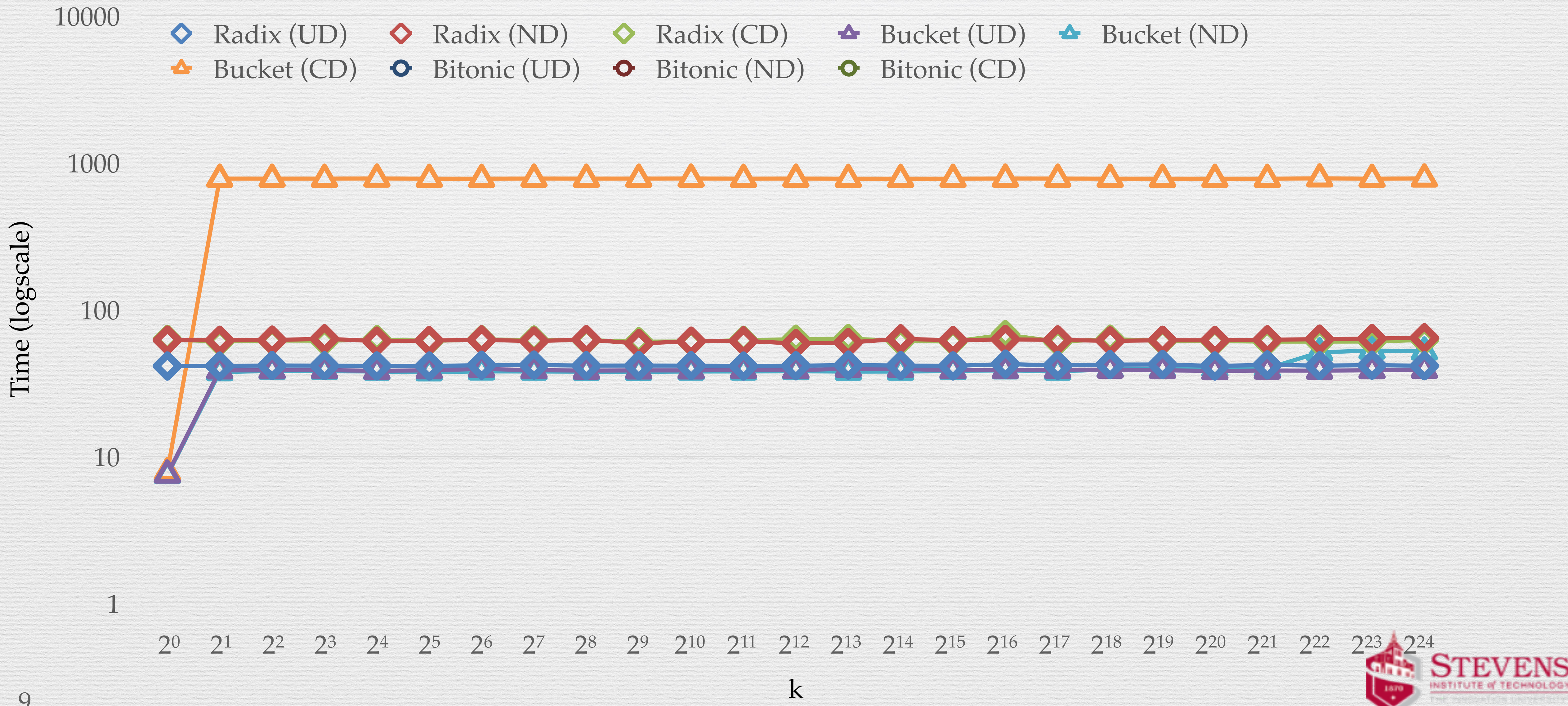
Bucket top-2 computation for an input vector with 16 elements. The highlighted bucket is of interest at each iteration.

[8]. Alabi, T., Blanchard, J.D., Gordon, B. and Steinbach, R. Fast k-Selection Algorithms for Graphics Processing Units. *Journal of Experimental Algorithmics (JEA)*, 2012.

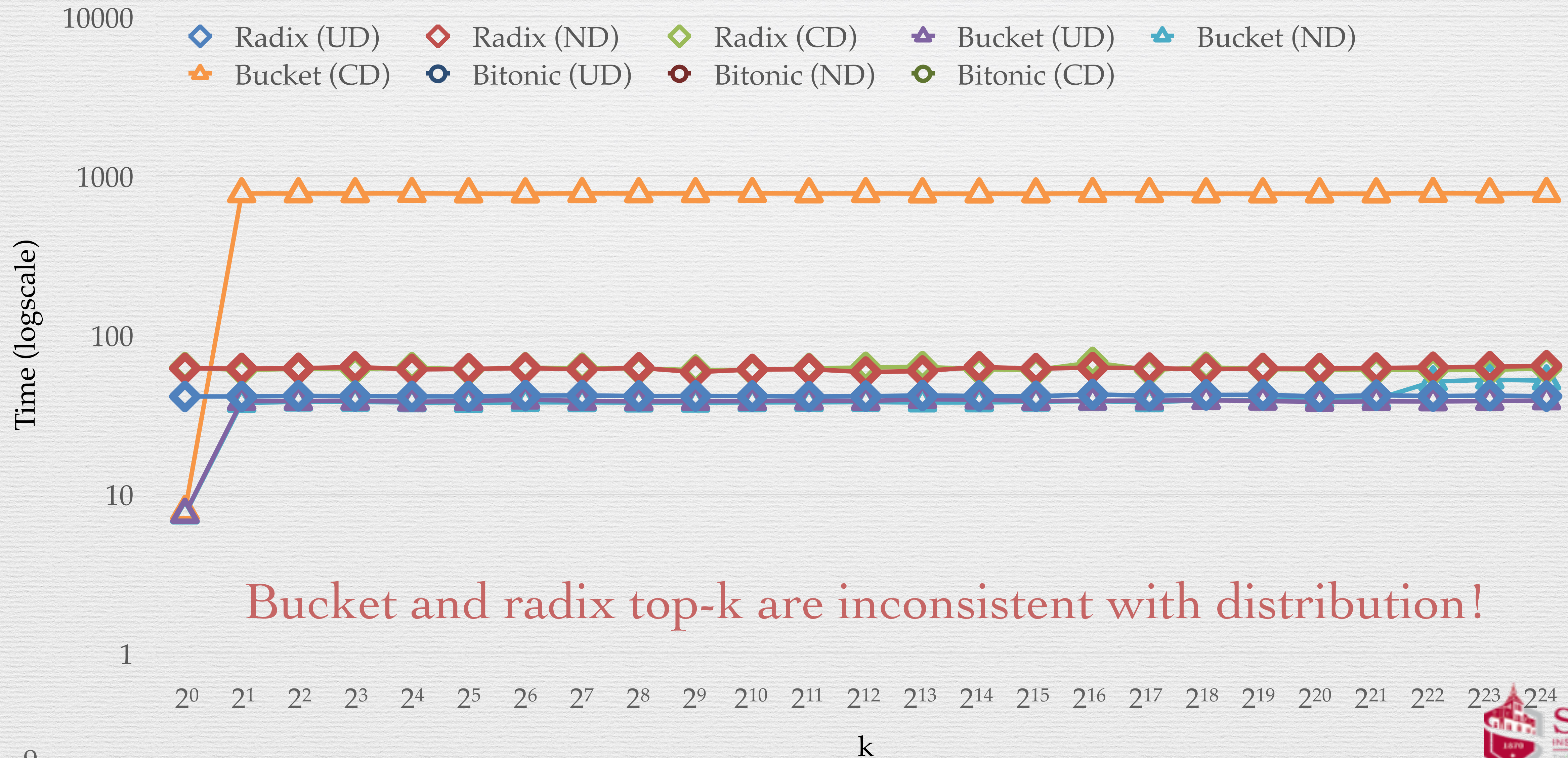
Inconsistency In State-Of-The-Art



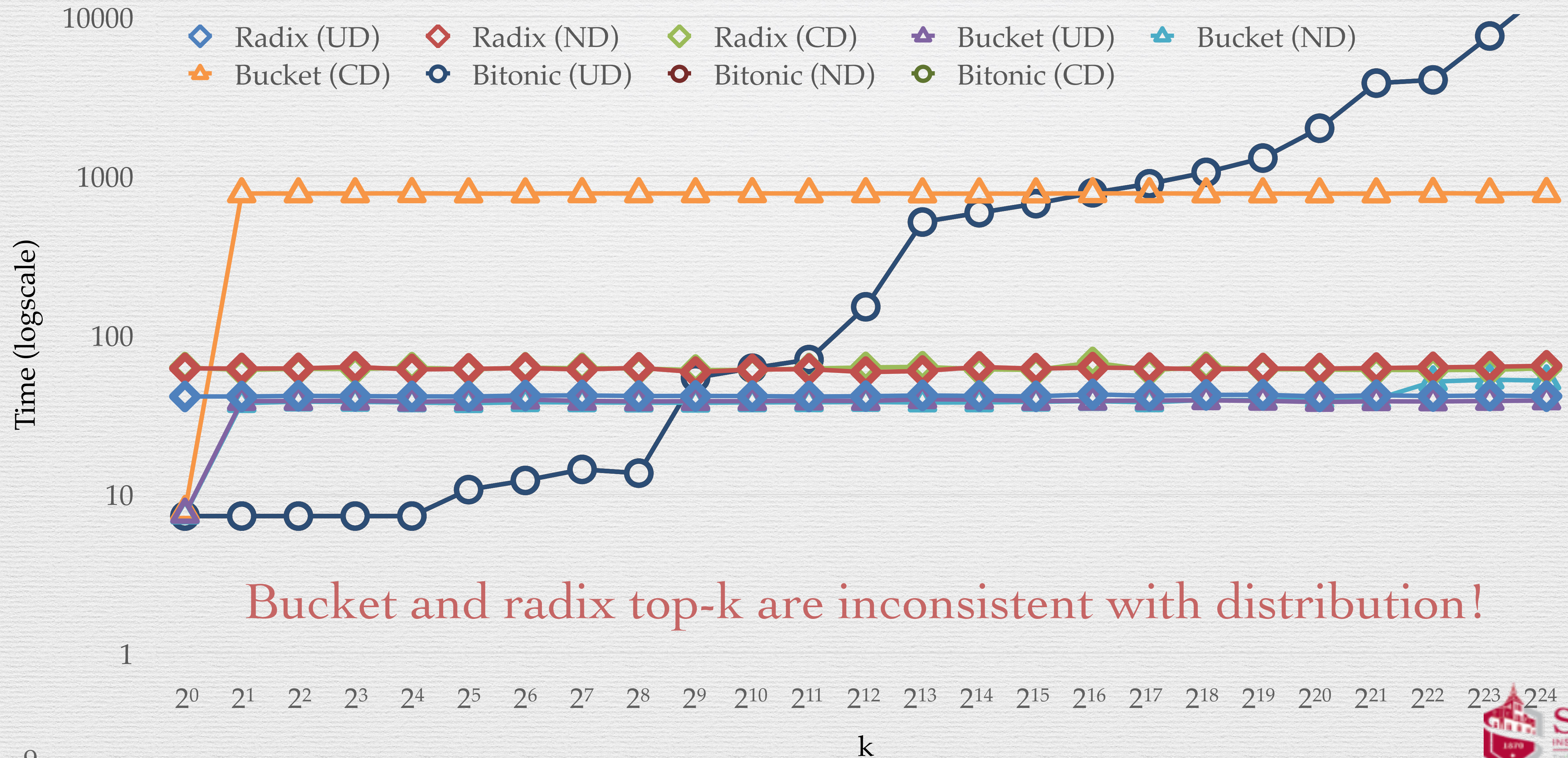
Inconsistency In State-Of-The-Art



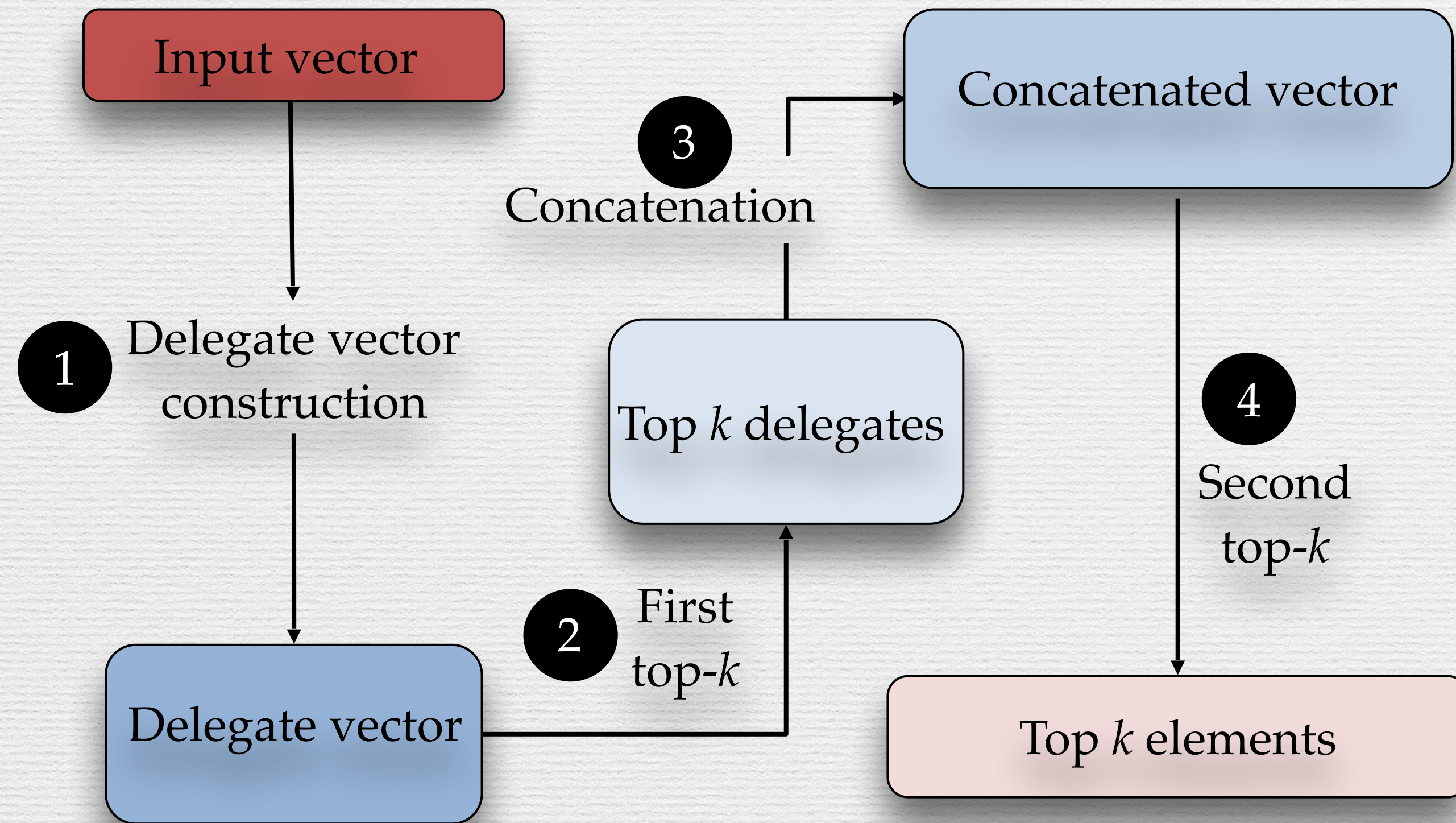
Inconsistency In State-Of-The-Art



Inconsistency In State-Of-The-Art

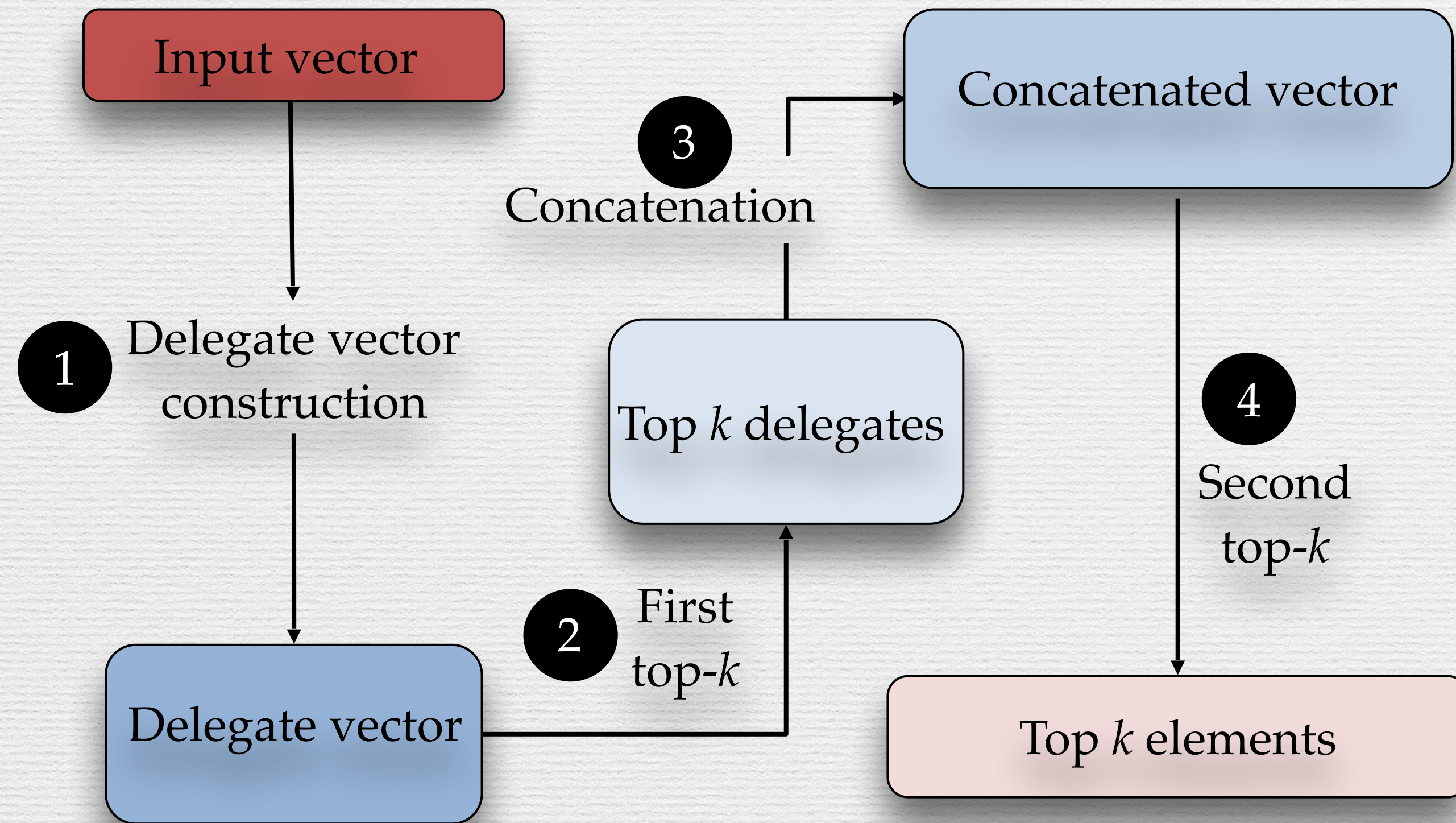


Dr.Top- k Workflow

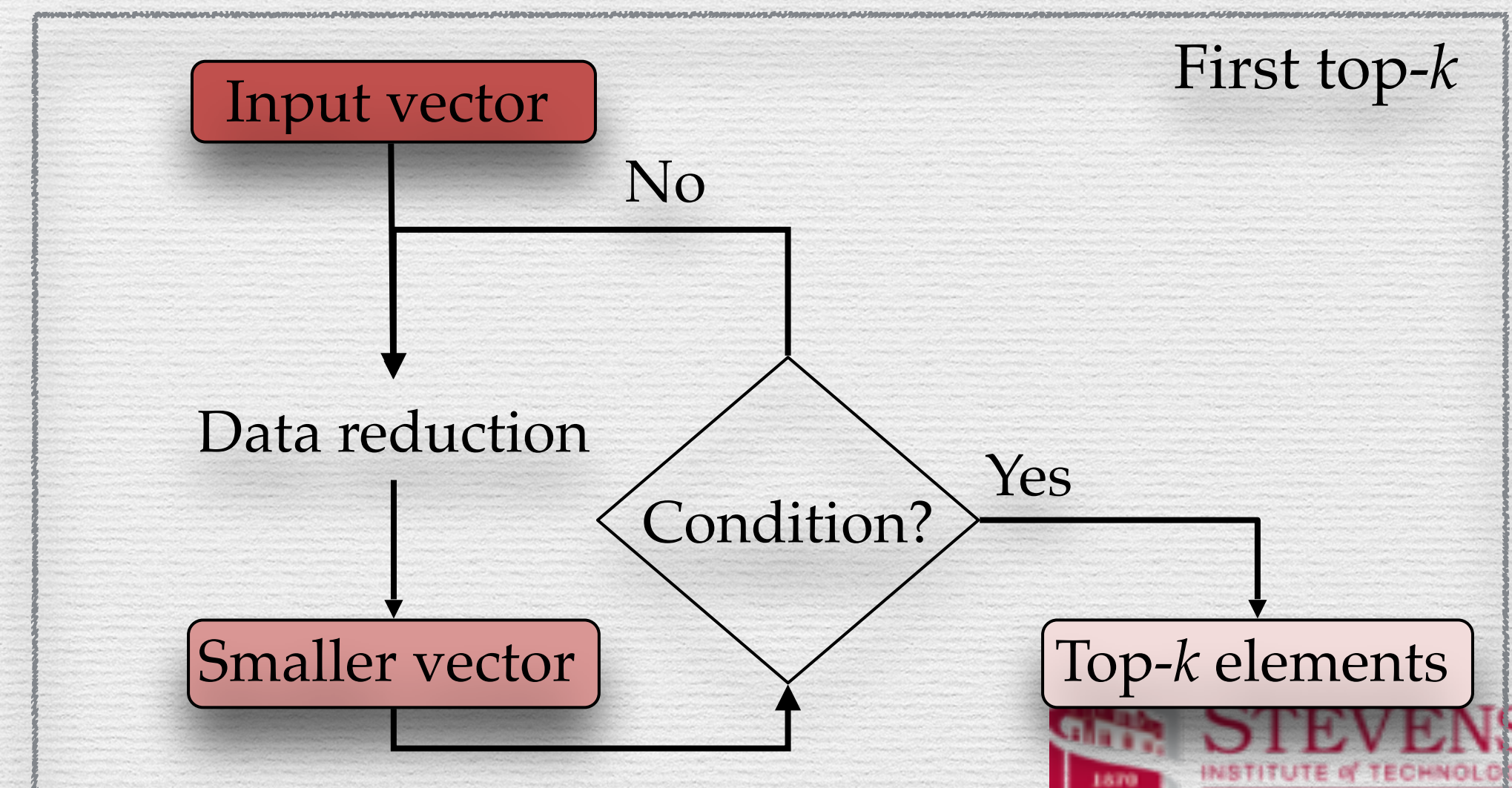
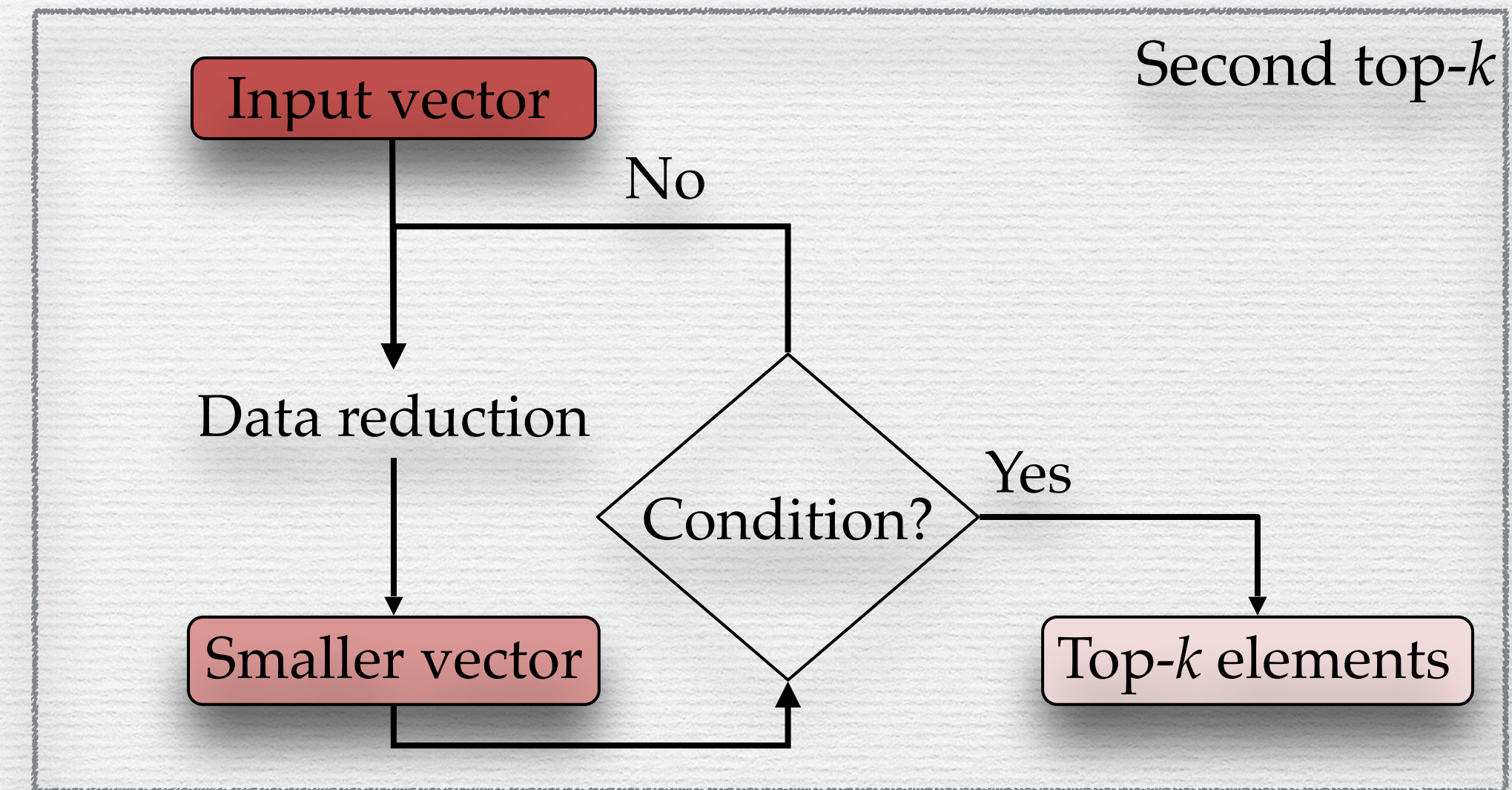


Dr.Top- k Workflow

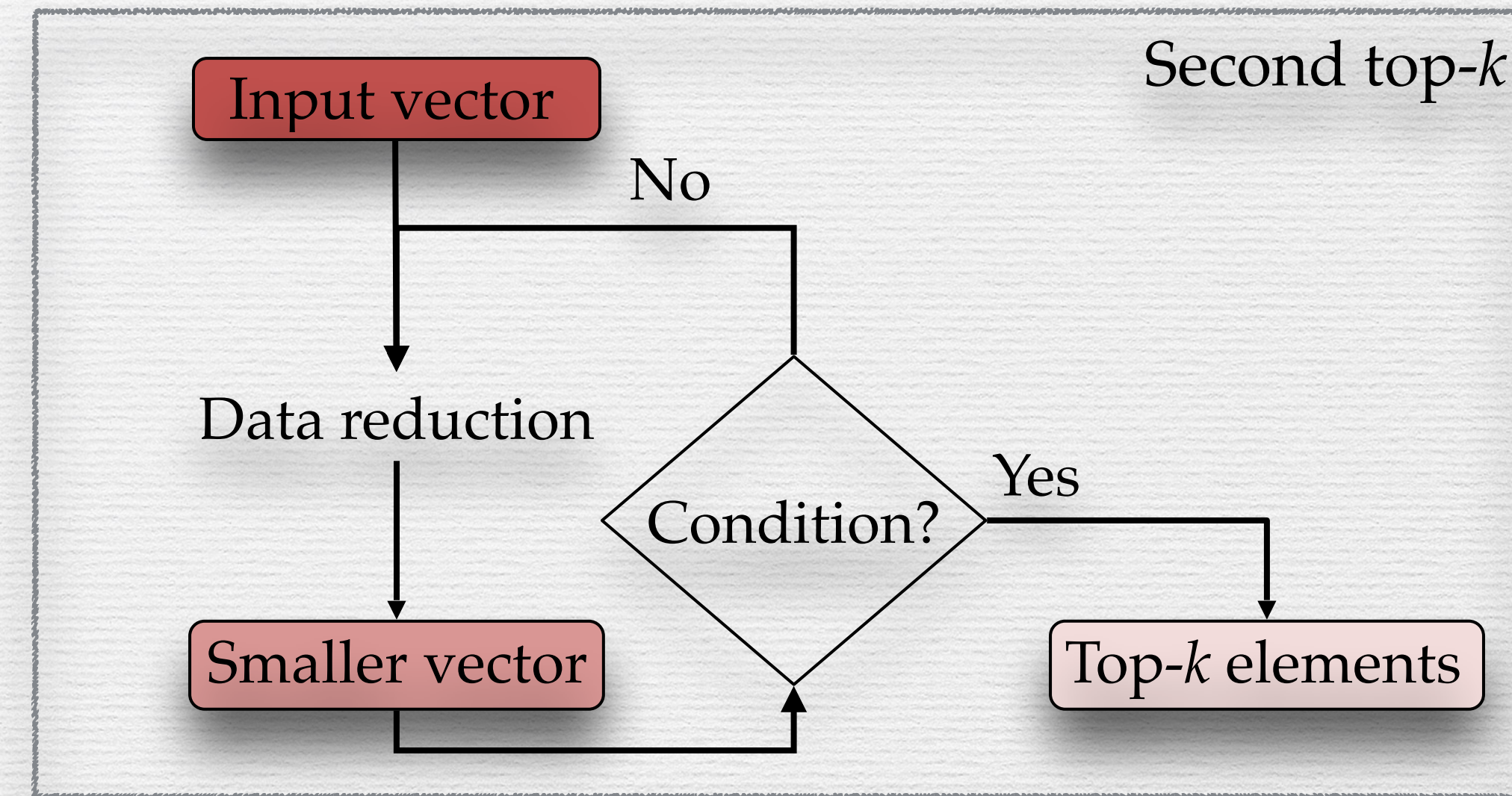
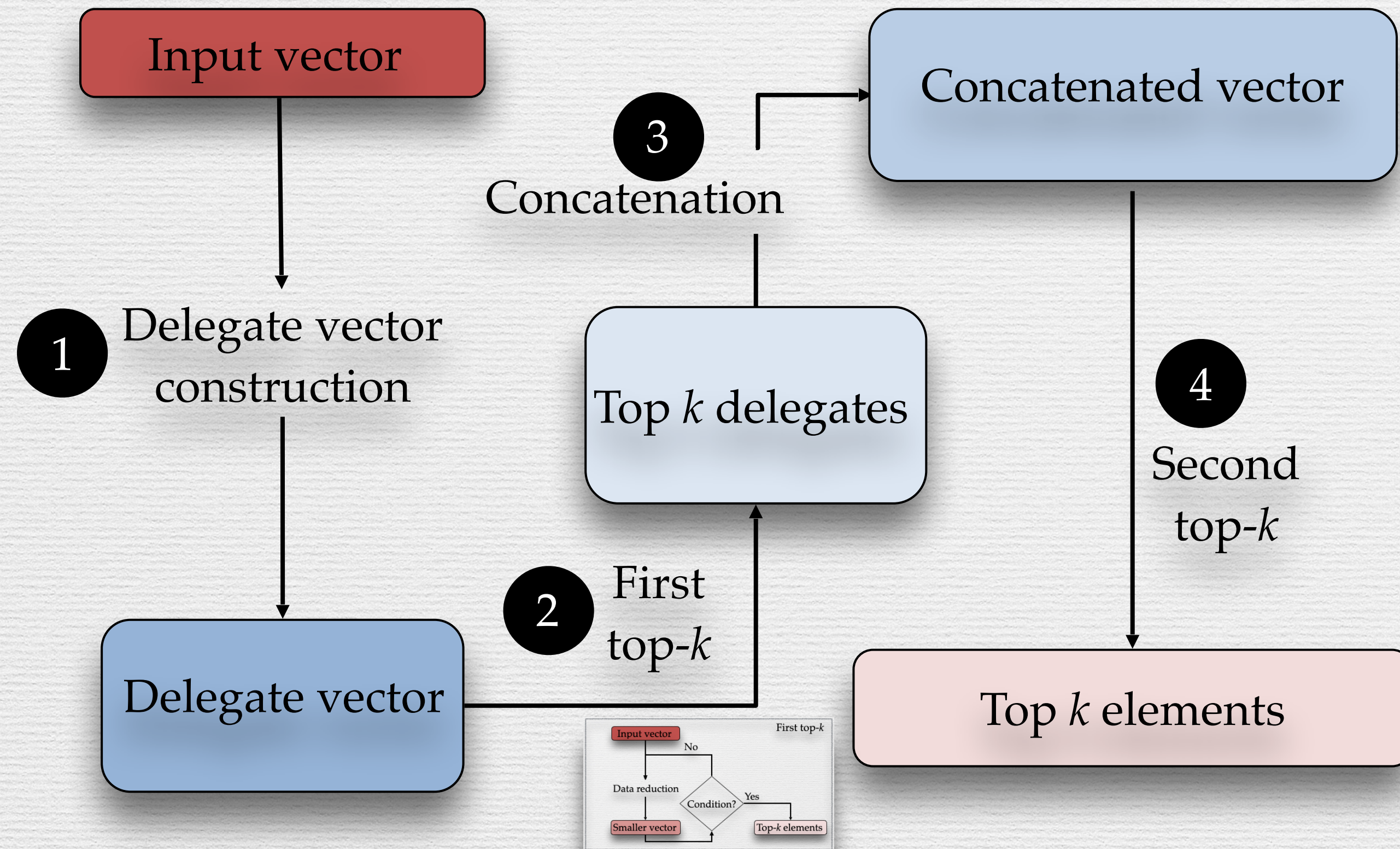
Dr.Top- k Workflow



Dr.Top- k Workflow

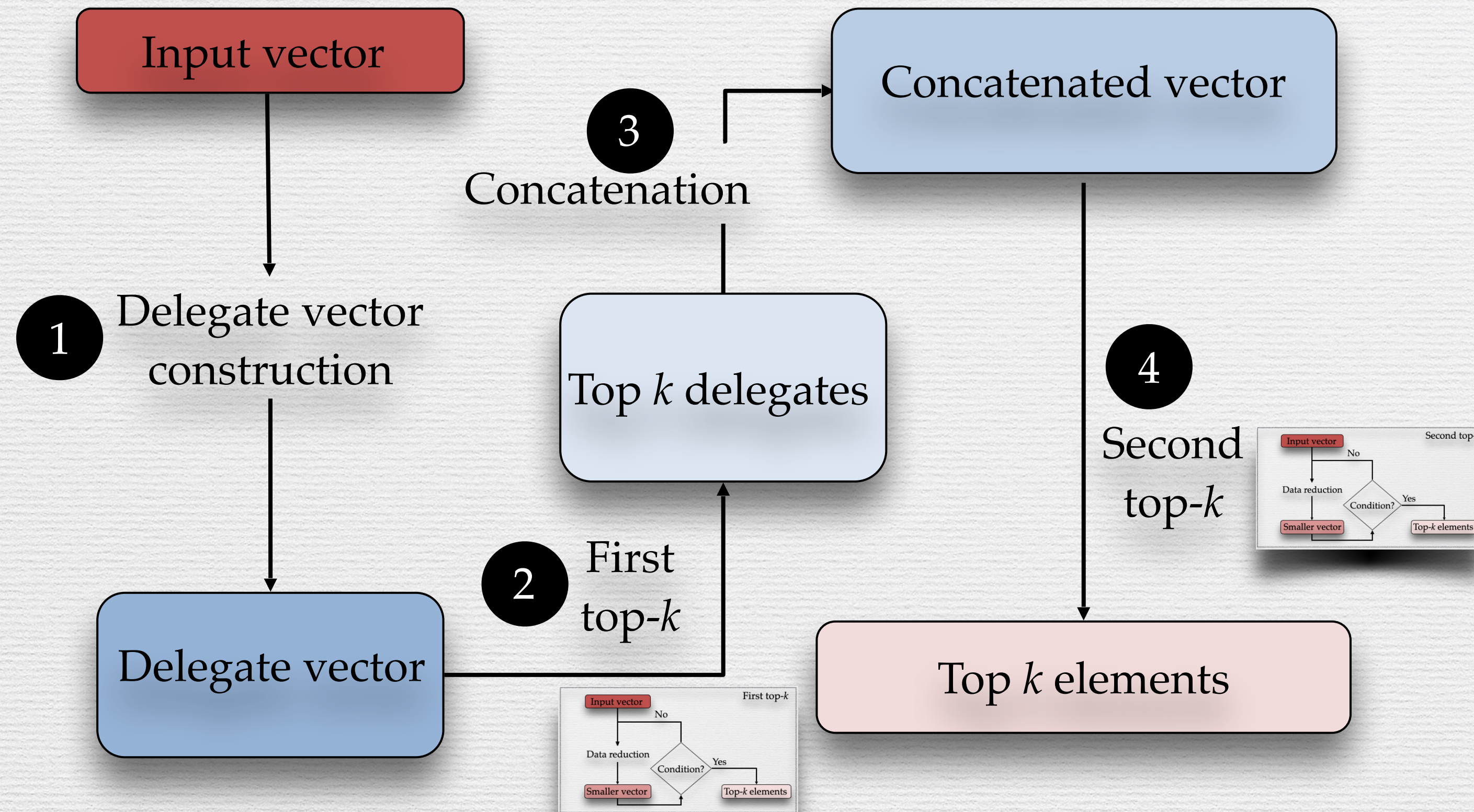


Dr.Top- k Workflow



Dr.Top- k Workflow

Dr.Top- k Workflow



Dr.Top- k Workflow

Dr. Top- k : Maximum Delegate Top- k

Compute top-2

Input vector

2001	101	1323	3012	212	1132	2310	2313	3000	3010	1002	3210	1020	333	2321	2003
------	-----	------	------	-----	------	------	------	------	------	------	------	------	-----	------	------

Dr. Top- k : Maximum Delegate Top- k

Compute top-2

Input vector

2001	101	1323	3012	212	1132	2310	2313	3000	3010	1002	3210	1020	333	2321	2003
------	-----	------	------	-----	------	------	------	------	------	------	------	------	-----	------	------

Subrange 0

Subrange 1

Subrange 2

Subrange 3

2001	101	1323	3012	212	1132	2310	2313	3000	3010	1002	3210	1020	333	2321	2003
------	-----	------	------	-----	------	------	------	------	------	------	------	------	-----	------	------

Input vector

Dr. Top- k : Maximum Delegate Top- k

Compute top-2

Input vector

2001	101	1323	3012	212	1132	2310	2313	3000	3010	1002	3210	1020	333	2321	2003
------	-----	------	------	-----	------	------	------	------	------	------	------	------	-----	------	------

Subrange 0

Subrange 1

Subrange 2

Subrange 3

2001	101	1323	3012	212	1132	2310	2313	3000	3010	1002	3210	1020	333	2321	2003
------	-----	------	------	-----	------	------	------	------	------	------	------	------	-----	------	------

Input vector



1

Delegate vector
construction



Delegate vector

Dr. Top- k : Maximum Delegate Top- k

Compute top-2

Input vector

2001	101	1323	3012	212	1132	2310	2313	3000	3010	1002	3210	1020	333	2321	2003
------	-----	------	------	-----	------	------	------	------	------	------	------	------	-----	------	------

Subrange 0

Subrange 1

Subrange 2

Subrange 3

2001	101	1323	3012	212	1132	2310	2313	3000	3010	1002	3210	1020	333	2321	2003
------	-----	------	------	-----	------	------	------	------	------	------	------	------	-----	------	------

Delegate vector

3012	2313	3210	2321
------	------	------	------

Input vector

1

Delegate vector
construction

Delegate vector

Dr. Top- k : Maximum Delegate Top- k

Compute top-2

Input vector

2001	101	1323	3012	212	1132	2310	2313	3000	3010	1002	3210	1020	333	2321	2003
------	-----	------	------	-----	------	------	------	------	------	------	------	------	-----	------	------

Subrange 0

Subrange 1

Subrange 2

Subrange 3

2001	101	1323	3012	212	1132	2310	2313	3000	3010	1002	3210	1020	333	2321	2003
------	-----	------	------	-----	------	------	------	------	------	------	------	------	-----	------	------

Input vector

Delegate vector

3012	2313	3210	2321
------	------	------	------

Top-2 delegates

3012	3210
------	------

1

Delegate vector
construction

Top k delegates

2

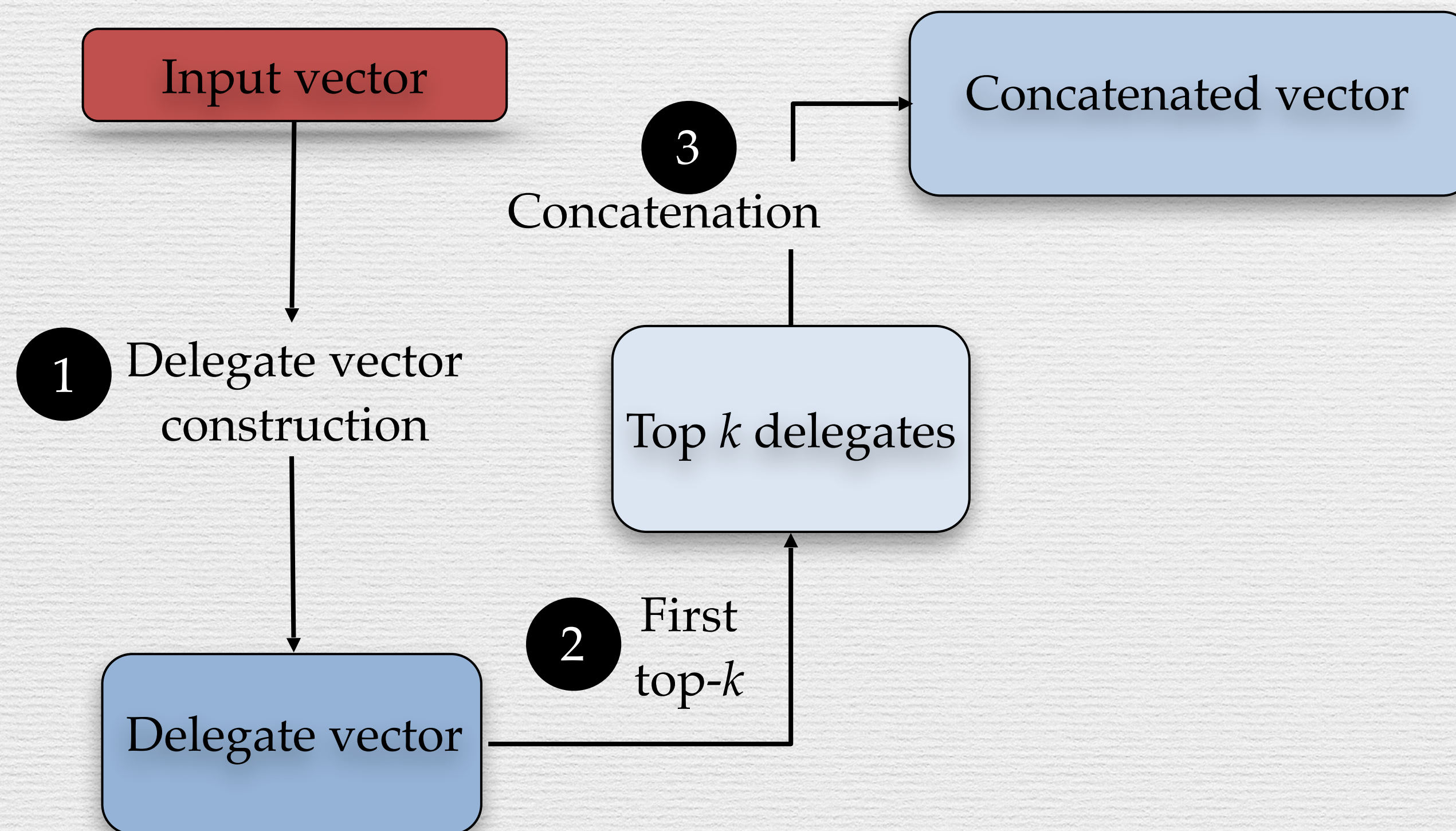
First
top- k

Delegate vector

Dr. Top- k : Maximum Delegate Top- k

Compute top-2

Input vector	2001	101	1323	3012	212	1132	2310	2313	3000	3010	1002	3210	1020	333	2321	2003
	Subrange 0				Subrange 1				Subrange 2				Subrange 3			
	2001	101	1323	3012	212	1132	2310	2313	3000	3010	1002	3210	1020	333	2321	2003



Delegate vector

3012	2313	3210	2321
------	------	------	------

Top-2 delegates

3012	3210
------	------

Dr. Top- k : Maximum Delegate Top- k

Compute top-2

Input vector

2001	101	1323	3012	212	1132	2310	2313	3000	3010	1002	3210	1020	333	2321	2003
------	-----	------	------	-----	------	------	------	------	------	------	------	------	-----	------	------

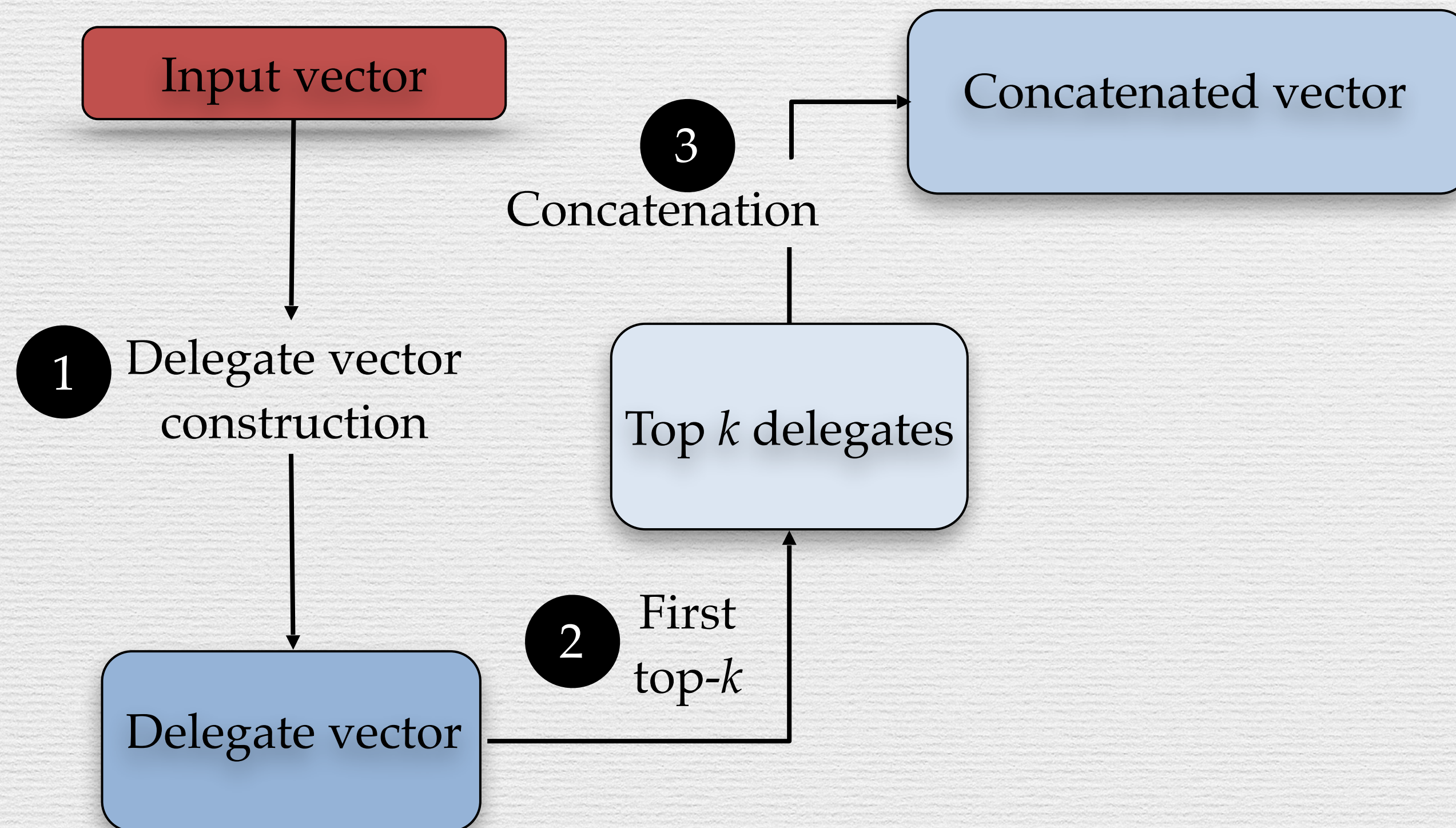
Subrange 0

Subrange 1

Subrange 2

Subrange 3

2001	101	1323	3012	212	1132	2310	2313	3000	3010	1002	3210	1020	333	2321	2003
------	-----	------	------	-----	------	------	------	------	------	------	------	------	-----	------	------



Delegate vector

3012	2313	3210	2321
------	------	------	------

Top-2 delegates

3012	3210
------	------

2001	101	1323	3012	3000	3010	1002	3210
------	-----	------	------	------	------	------	------

Concatenated vector

Dr. Top- k : Maximum Delegate Top- k

Compute top-2

Input vector

2001	101	1323	3012	212	1132	2310	2313	3000	3010	1002	3210	1020	333	2321	2003
------	-----	------	------	-----	------	------	------	------	------	------	------	------	-----	------	------

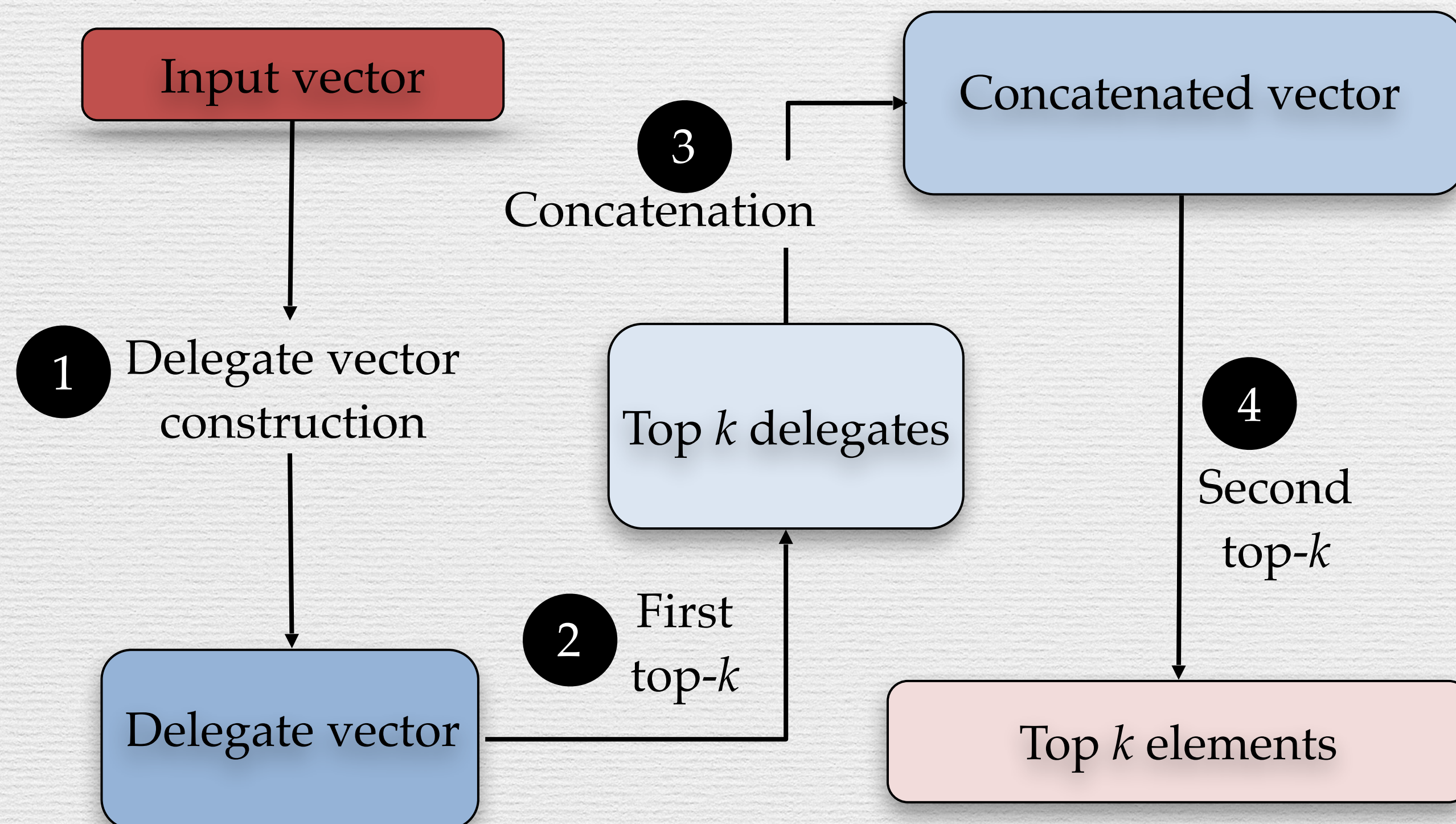
Subrange 0

Subrange 1

Subrange 2

Subrange 3

2001	101	1323	3012	212	1132	2310	2313	3000	3010	1002	3210	1020	333	2321	2003
------	-----	------	------	-----	------	------	------	------	------	------	------	------	-----	------	------



Delegate vector

3012	2313	3210	2321
------	------	------	------

Top-2 delegates

3012	3210
------	------

2001	101	1323	3012	3000	3010	1002	3210
------	-----	------	------	------	------	------	------

Concatenated vector

Top-2 elements

3012	3210
------	------

Dr. Top- k : Importance of Subrange Size

Compute top-2

Input vector

2001	101	1323	3012	212	1132	2310	2313	3000	3010	1002	3210	1020	333	2321	2003
Subrange 0				Subrange 1				Subrange 2				Subrange 3			
2001	101	1323	3012	212	1132	2310	2313	3000	3010	1002	3210	1020	333	2321	2003

Delegate vector

3012	2313	3210	2321
------	------	------	------

Top-2 delegates

3012	3210
------	------

2001	101	1323	3012	3000	3010	1002	3210
------	-----	------	------	------	------	------	------

Concatenated vector

Top-2 elements

3012	3210
------	------

Dr. Top- k : Importance of Subrange Size

Compute top-2

Input vector

2001	101	1323	3012	212	1132	2310	2313	3000	3010	1002	3210	1020	333	2321	2003
------	-----	------	------	-----	------	------	------	------	------	------	------	------	-----	------	------

Subrange 0

Subrange 1

Subrange 2

Subrange 3

2001	101	1323	3012	212	1132	2310	2313	3000	3010	1002	3210	1020	333	2321	2003
------	-----	------	------	-----	------	------	------	------	------	------	------	------	-----	------	------

|V|

Delegate vector

3012	2313	3210	2321
------	------	------	------

Top-2 delegates

3012	3210
------	------

2001	101	1323	3012	3000	3010	1002	3210
------	-----	------	------	------	------	------	------

Concatenated vector

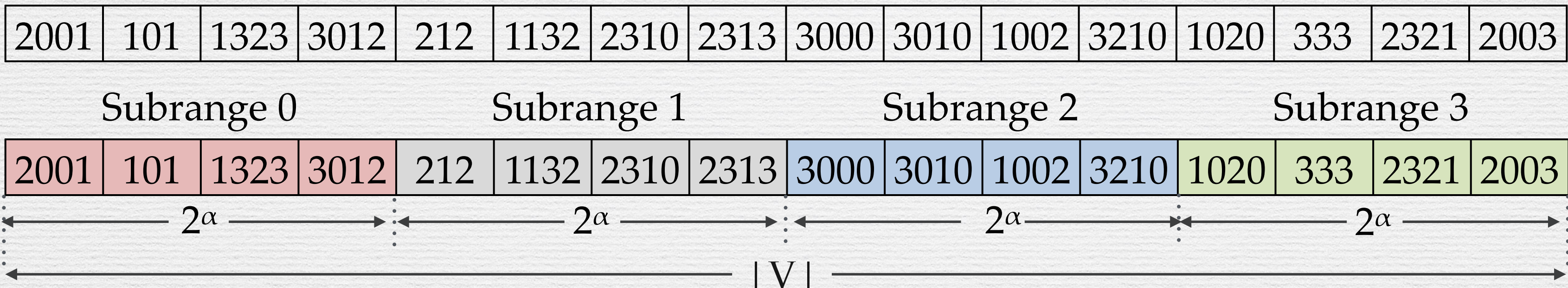
Top-2 elements

3012	3210
------	------

Dr. Top- k : Importance of Subrange Size

Compute top-2

Input vector



Delegate vector

3012	2313	3210	2321
------	------	------	------

Top-2 delegates

3012	3210
------	------

2001	101	1323	3012	3000	3010	1002	3210
------	-----	------	------	------	------	------	------

Concatenated vector

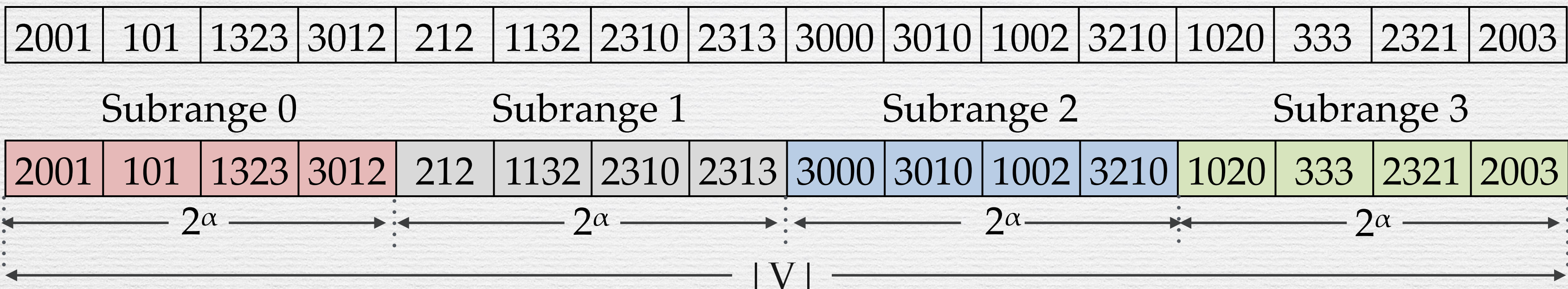
Top-2 elements

3012	3210
------	------

Dr. Top- k : Importance of Subrange Size

Compute top-2

Input vector



Delegate vector

3012	2313	3210	2321
------	------	------	------

Top-2 delegates

3012	3210
------	------

2001	101	1323	3012	3000	3010	1002	3210
------	-----	------	------	------	------	------	------

Concatenated vector

Top-2 elements

3012	3210
------	------

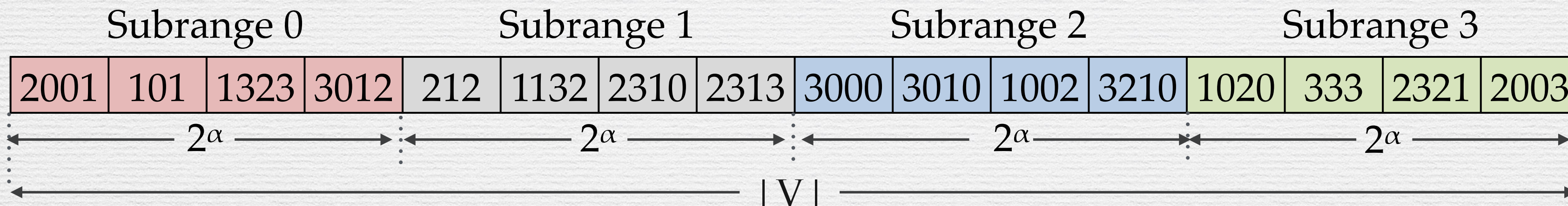
$$\text{Total top-}k\text{ workload} = \frac{|V|}{2^\alpha} + k * 2^\alpha$$

Dr. Top- k : Importance of Subrange Size

Compute top-2

Input vector

2001	101	1323	3012	212	1132	2310	2313	3000	3010	1002	3210	1020	333	2321	2003
------	-----	------	------	-----	------	------	------	------	------	------	------	------	-----	------	------



Delegate vector

3012	2313	3210	2321
------	------	------	------

Top-2 delegates

3012	3210
------	------

Concatenated vector

2001	101	1323	3012	3000	3010	1002	3210
------	-----	------	------	------	------	------	------

Top-2 elements

3012	3210
------	------

$$\text{Total top-}k\text{ workload} = \frac{|V|}{2^\alpha} + k * 2^\alpha$$

Dr. Top- k : Importance of Subrange Size

Compute top-2

Input vector

2001	101	1323	3012	212	1132	2310	2313	3000	3010	1002	3210	1020	333	2321	2003
------	-----	------	------	-----	------	------	------	------	------	------	------	------	-----	------	------

Subrange 0

Subrange 1

Subrange 2

Subrange 3

2001	101	1323	3012	212	1132	2310	2313	3000	3010	1002	3210	1020	333	2321	2003
------	-----	------	------	-----	------	------	------	------	------	------	------	------	-----	------	------

2^α

2^α

2^α

2^α

$|V|$

Delegate vector

3012	2313	3210	2321
------	------	------	------

Top-2 delegates

3012	3210
------	------

$$\text{Total top-}k \text{ workload} = \frac{|V|}{2^\alpha} + k * 2^\alpha$$

2001	101	1323	3012	3000	3010	1002	3210
------	-----	------	------	------	------	------	------

Concatenated vector

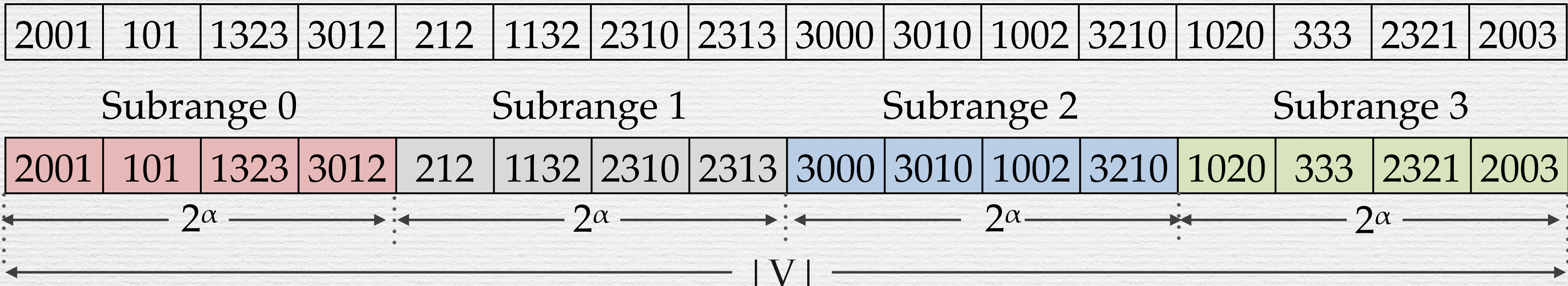
Top-2 elements

3012	3210
------	------

Dr. Top- k : Importance of Subrange Size

Compute top-2

Input vector



Delegate vector

3012	2313	3210	2321
------	------	------	------

Subrange size determines workload of top- k !

Top-2 delegates

3012	3210
------	------

$$\text{Total top-}k\text{ workload} = \frac{|V|}{2^\alpha} + k * 2^\alpha$$

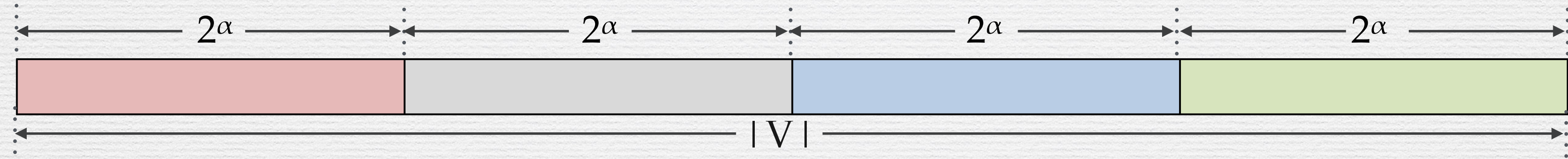
2001	101	1323	3012	3000	3010	1002	3210
------	-----	------	------	------	------	------	------

Concatenated vector

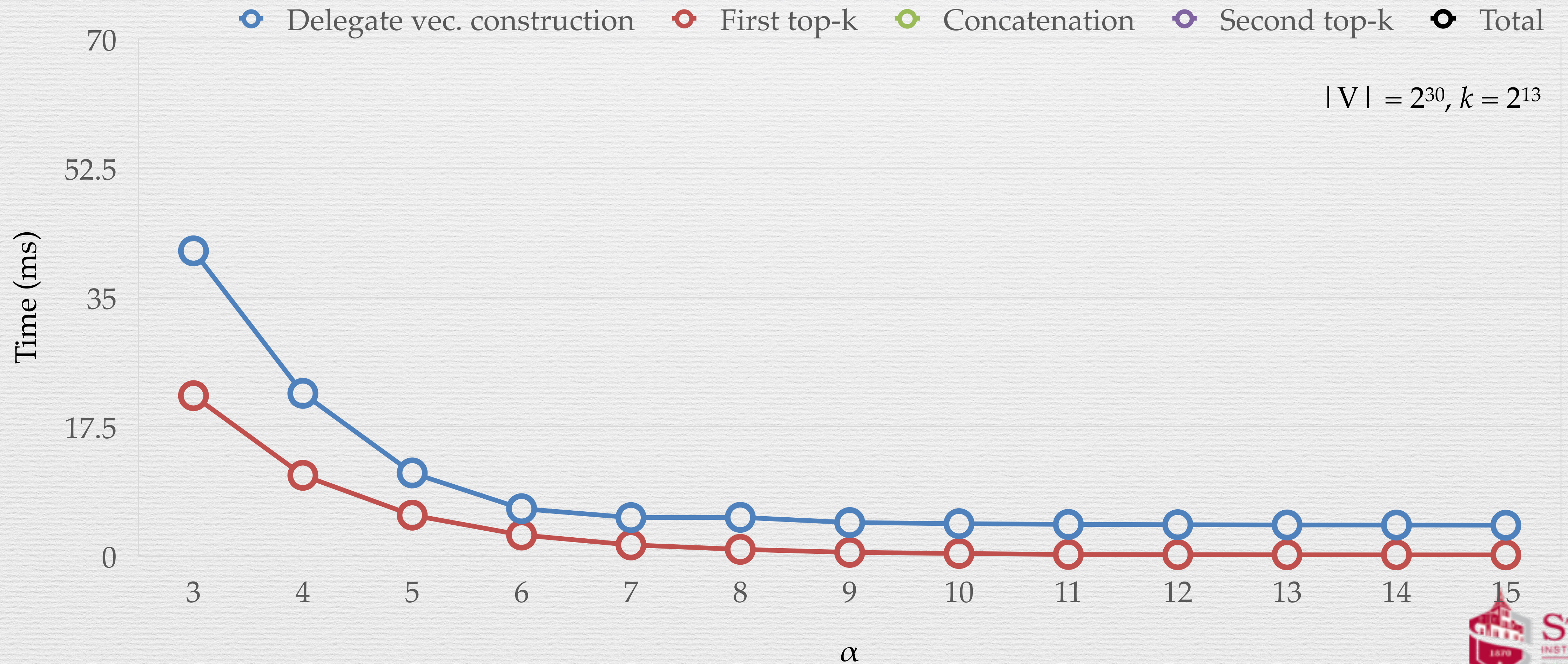
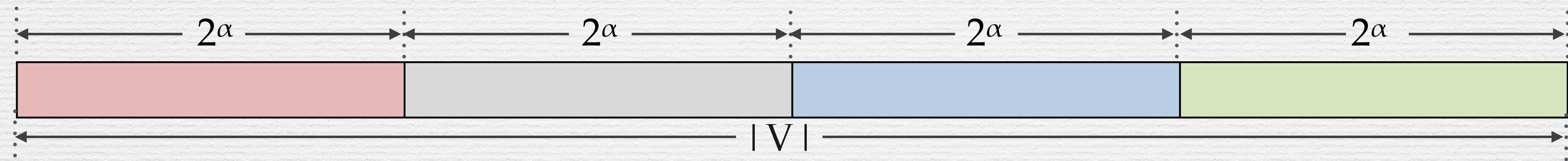
Top-2 elements

3012	3210
------	------

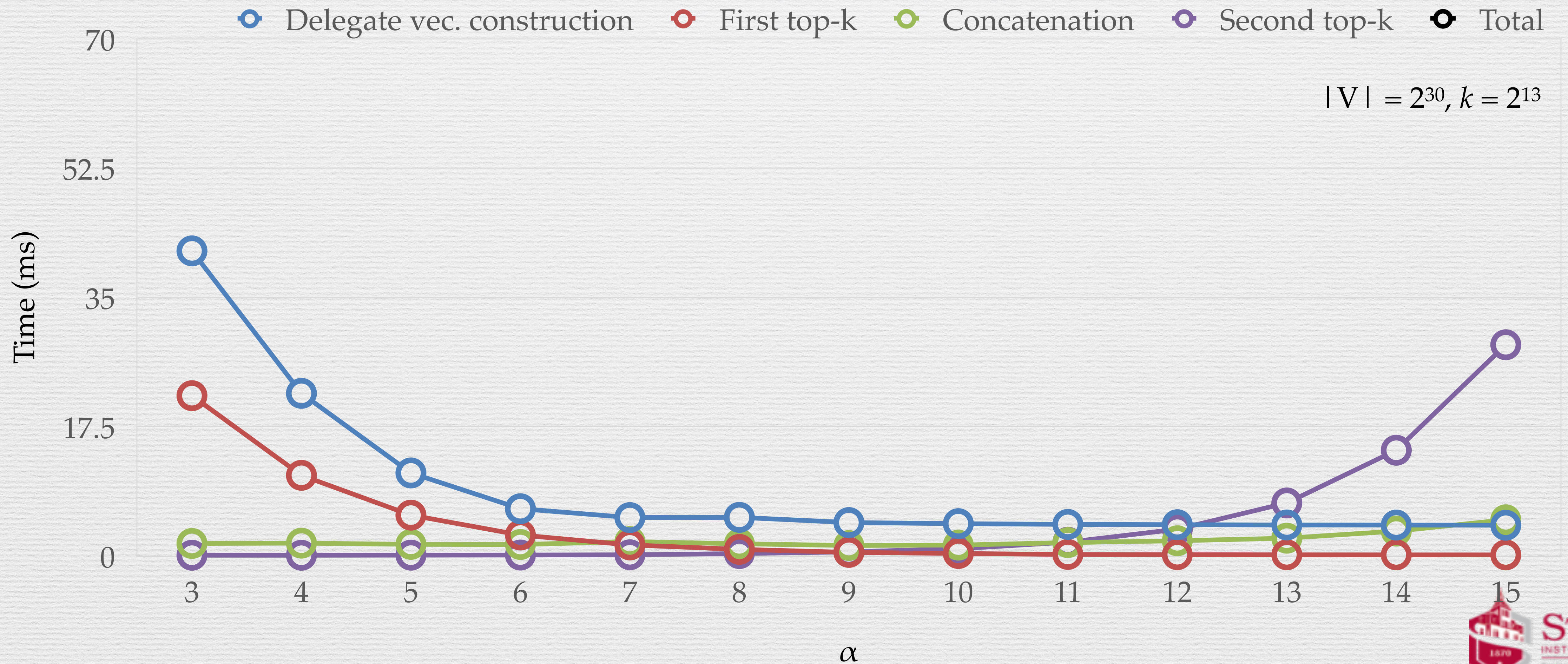
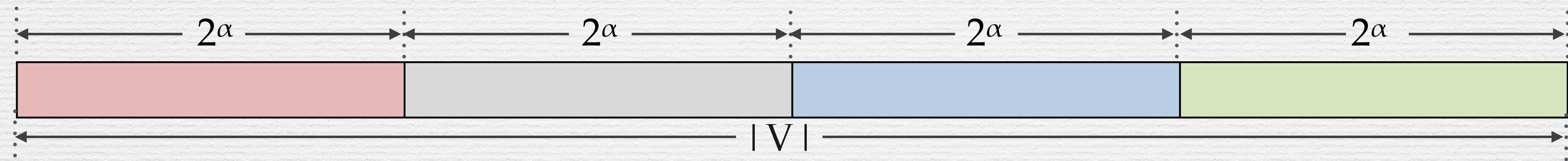
Dr.Top- k : α Tuning



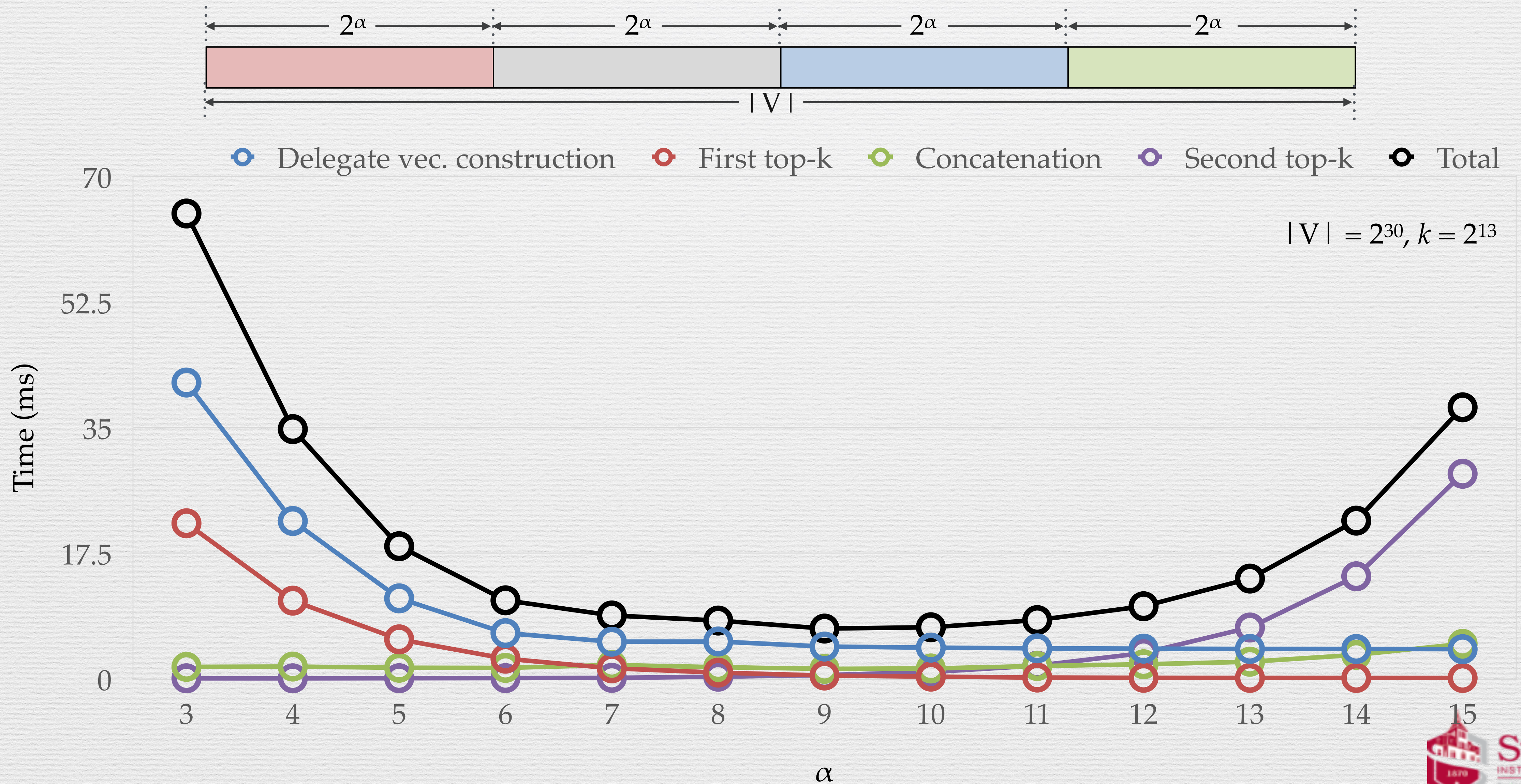
Dr.Top- k : α Tuning



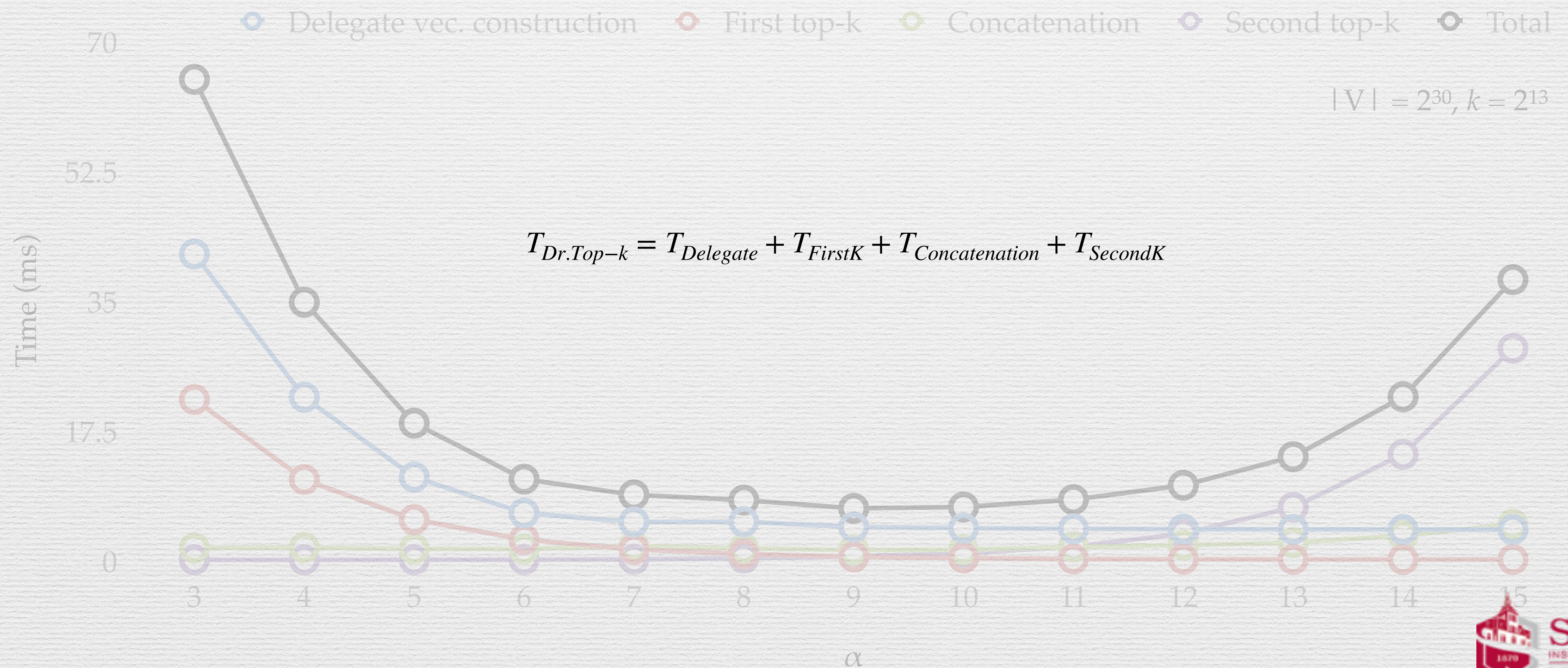
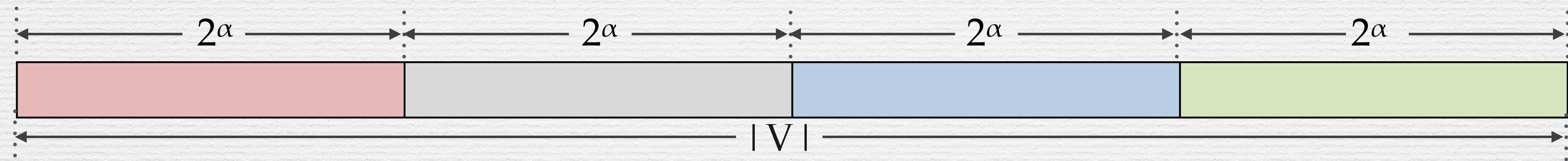
Dr.Top- k : α Tuning



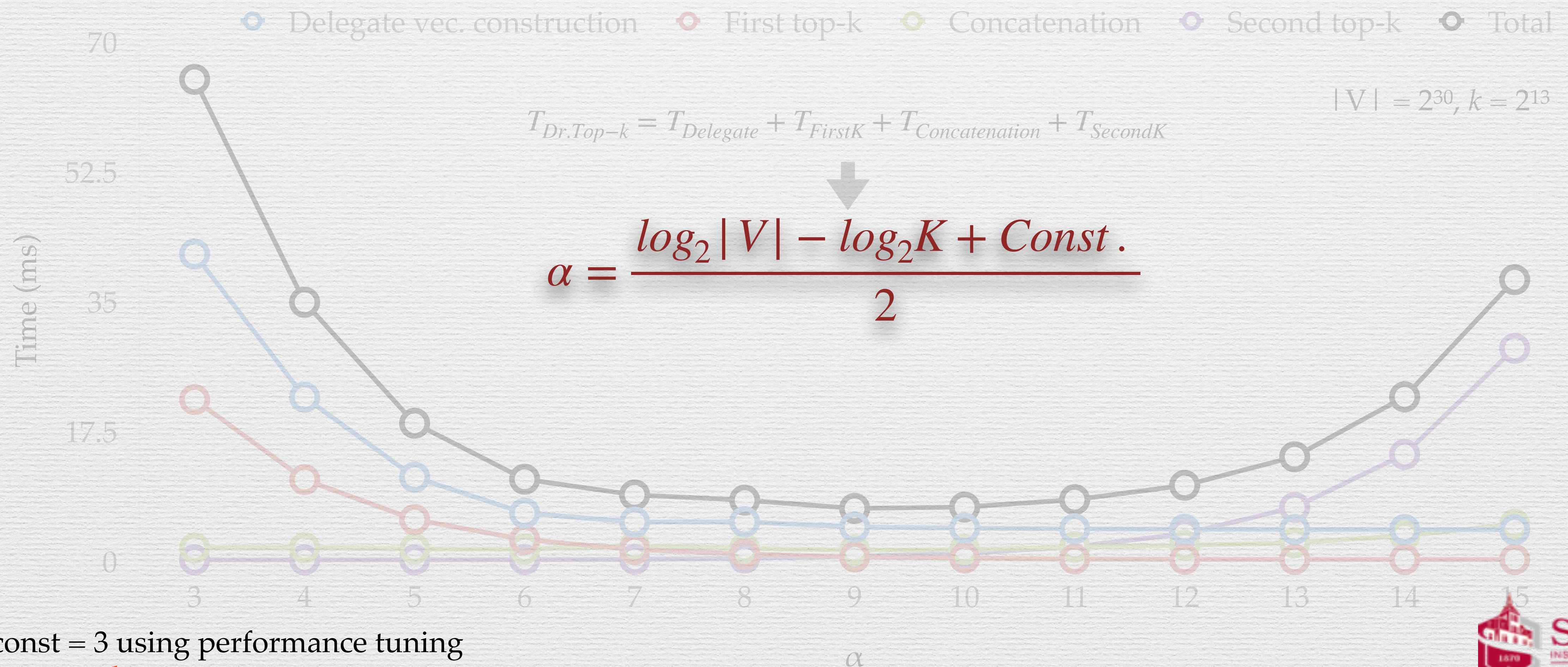
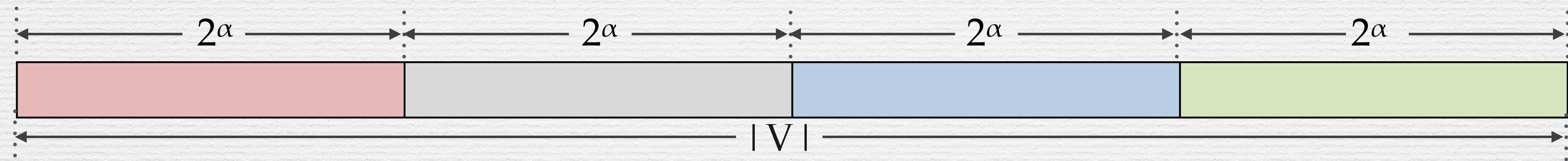
Dr.Top- k : α Tuning



Dr.Top- k : α Tuning

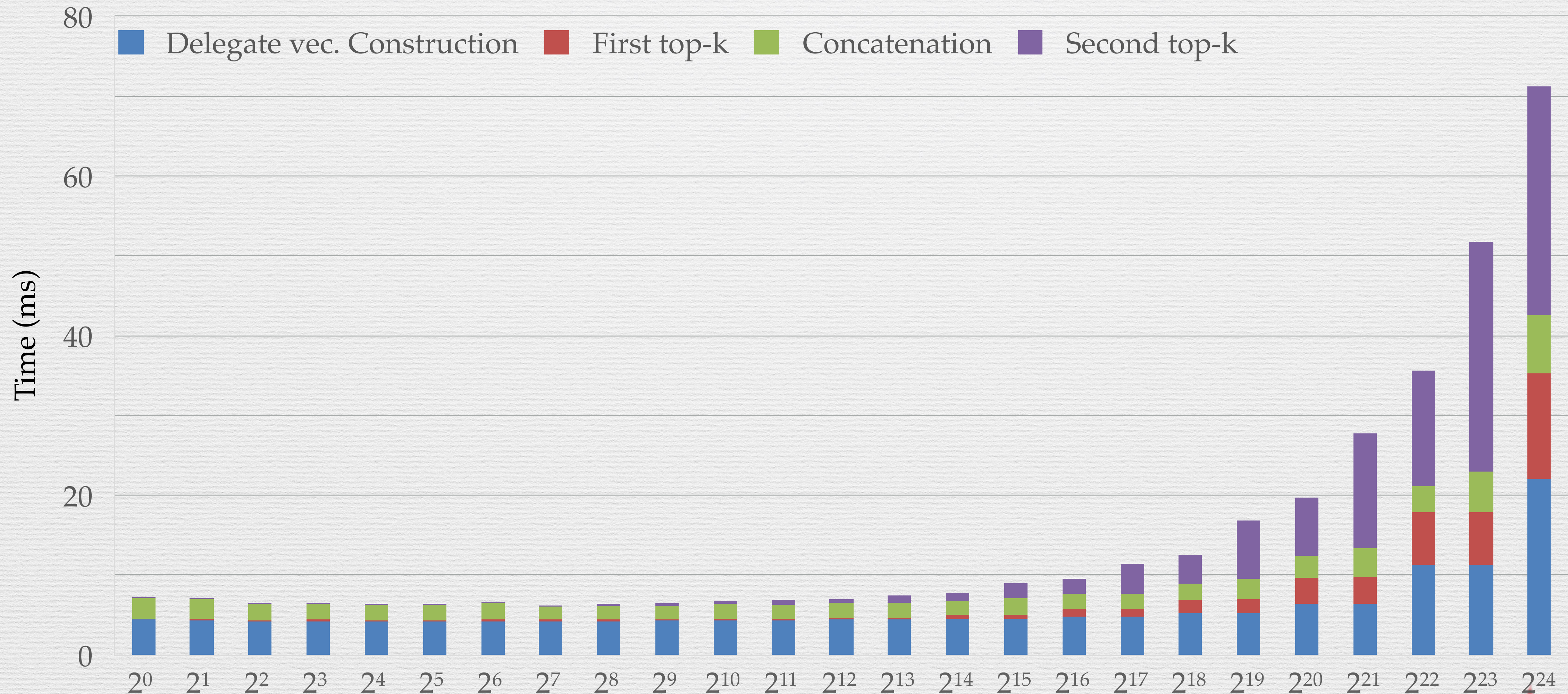


Dr.Top- k : α Tuning



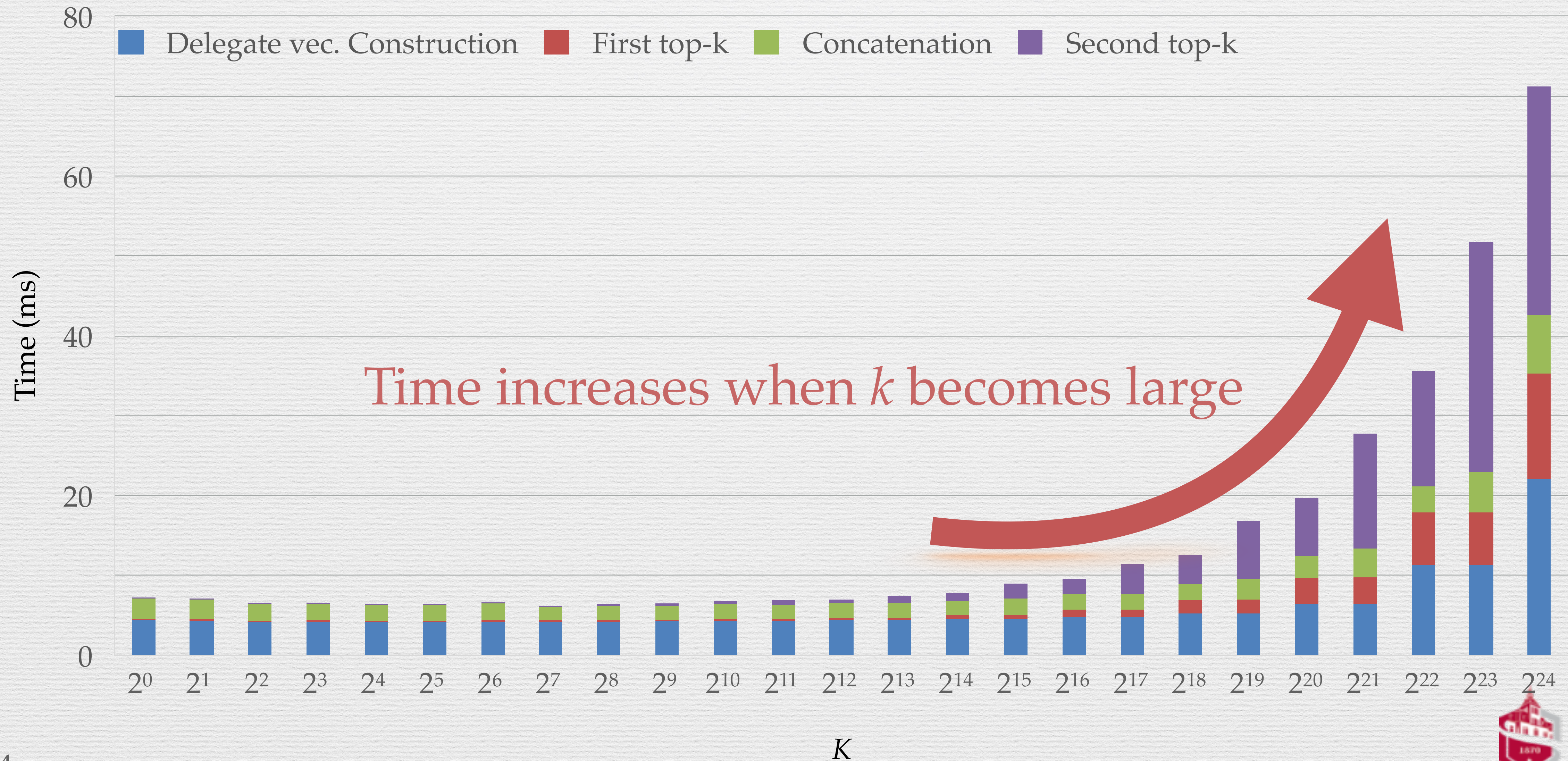
Dr. Top- k : Maximum Delegate Top- k

Max delegate: Time breakdown of Dr.Top- k



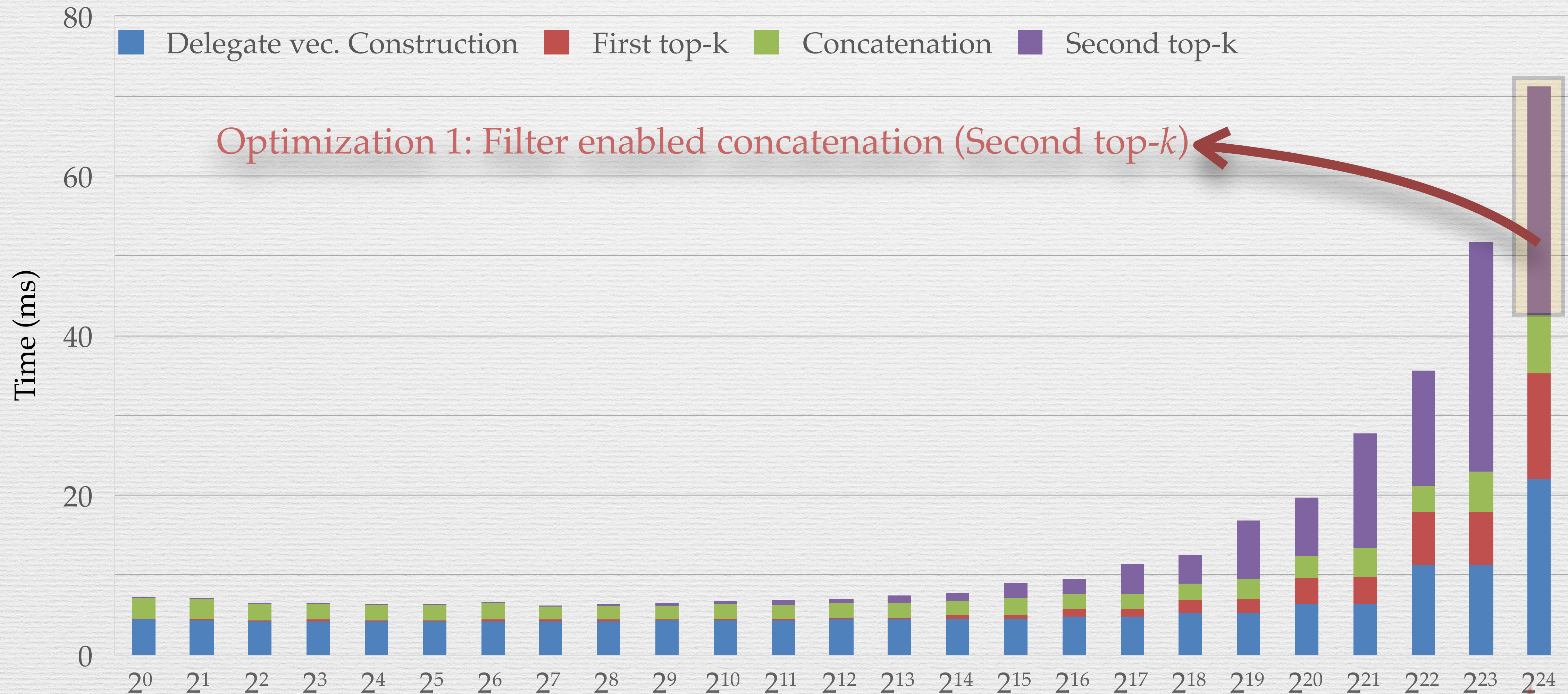
Dr. Top- k : Maximum Delegate Top- k

Max delegate: Time breakdown of Dr.Top- k



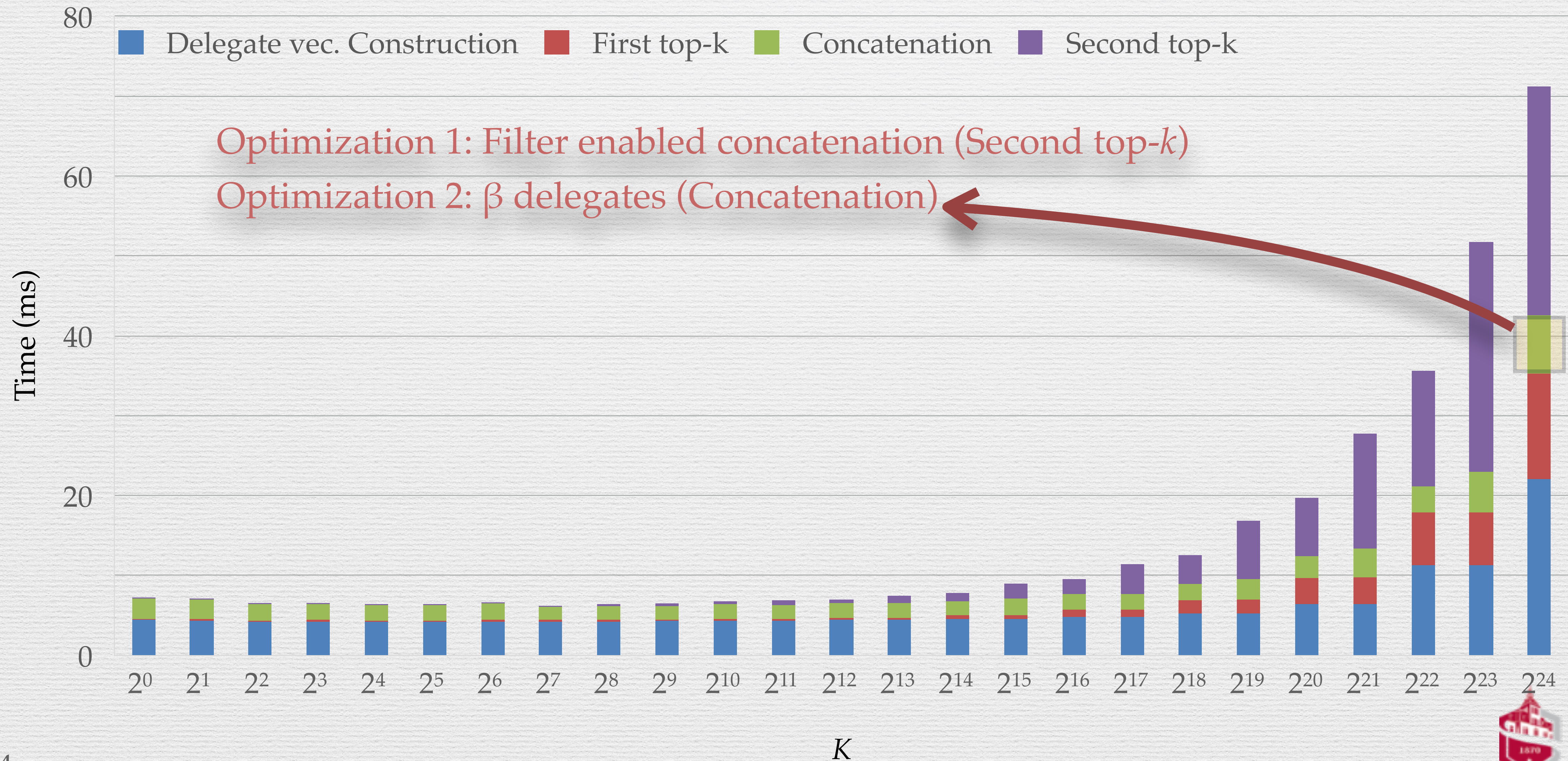
Dr. Top- k : Maximum Delegate Top- k

Max delegate: Time breakdown of Dr.Top- k



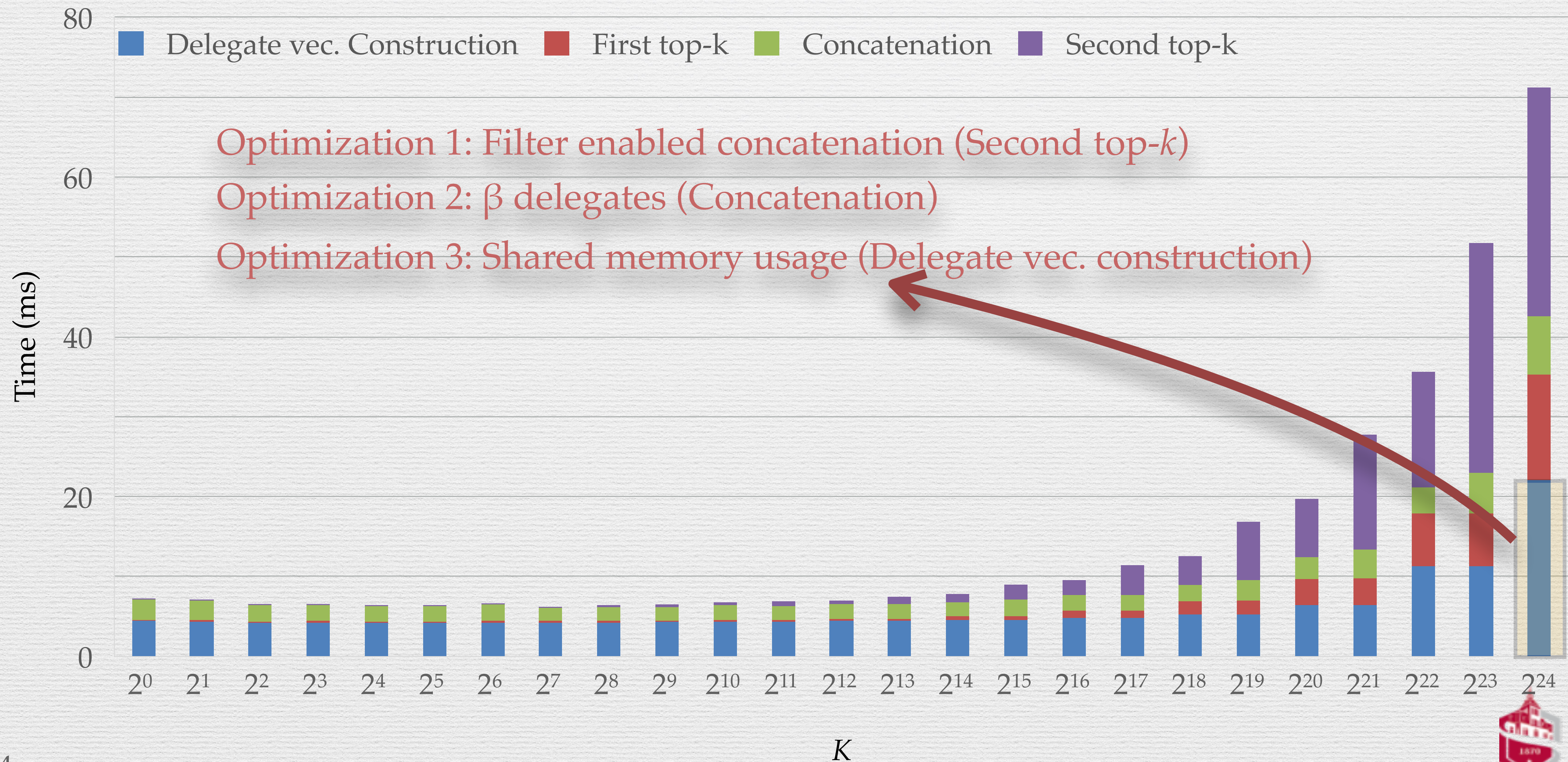
Dr. Top- k : Maximum Delegate Top- k

Max delegate: Time breakdown of Dr.Top- k



Dr. Top- k : Maximum Delegate Top- k

Max delegate: Time breakdown of Dr.Top- k



Dr. Top- k : Filter Enabled Concatenation

Compute top-2

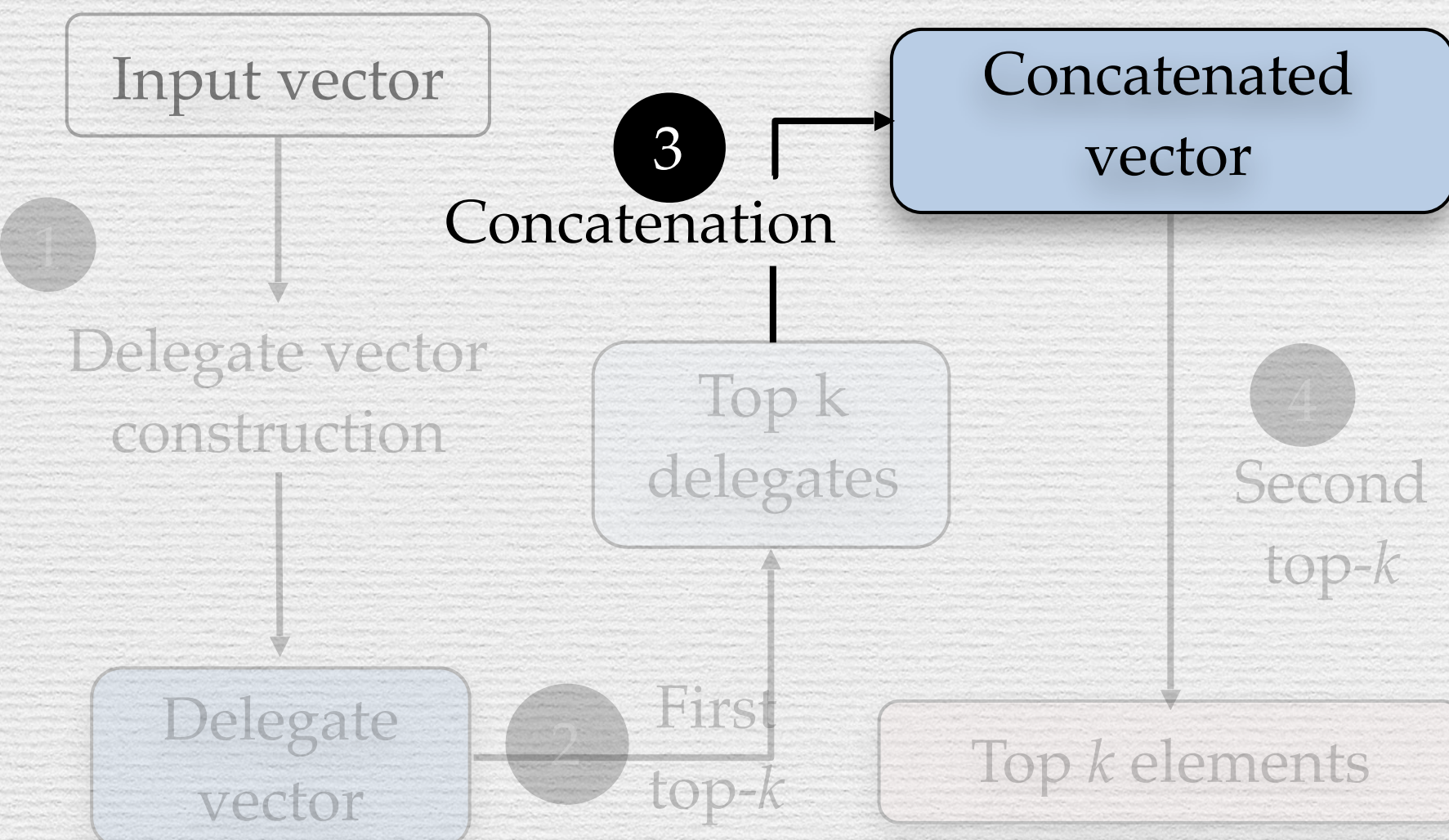
	Subrange 0				Subrange 1				Subrange 2				Subrange 3			
Input vector	2001	101	1323	3012	212	1132	2310	2313	3000	3010	1002	3210	1020	333	2321	2003

Delegate vector

3012	2313	3210	2321
------	------	------	------

Top-2 delegates

3012	3210
------	------

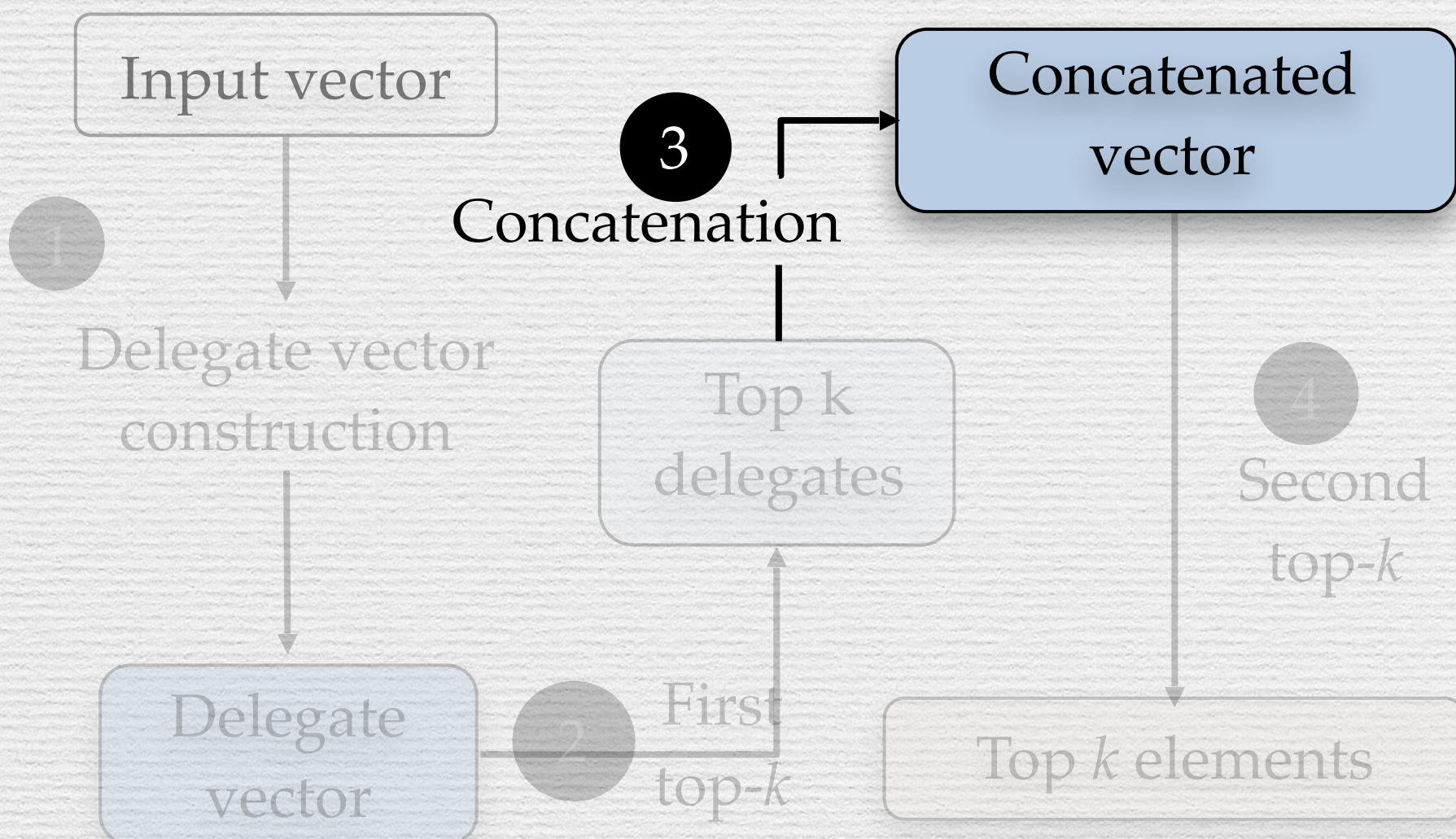


Changes for optimization

Dr. Top- k : Filter Enabled Concatenation

Compute top-2

	Subrange 0				Subrange 1				Subrange 2				Subrange 3			
Input vector	2001	101	1323	3012	212	1132	2310	2313	3000	3010	1002	3210	1020	333	2321	2003



Changes for optimization



Delegate vector

3012	2313	3210	2321
------	------	------	------

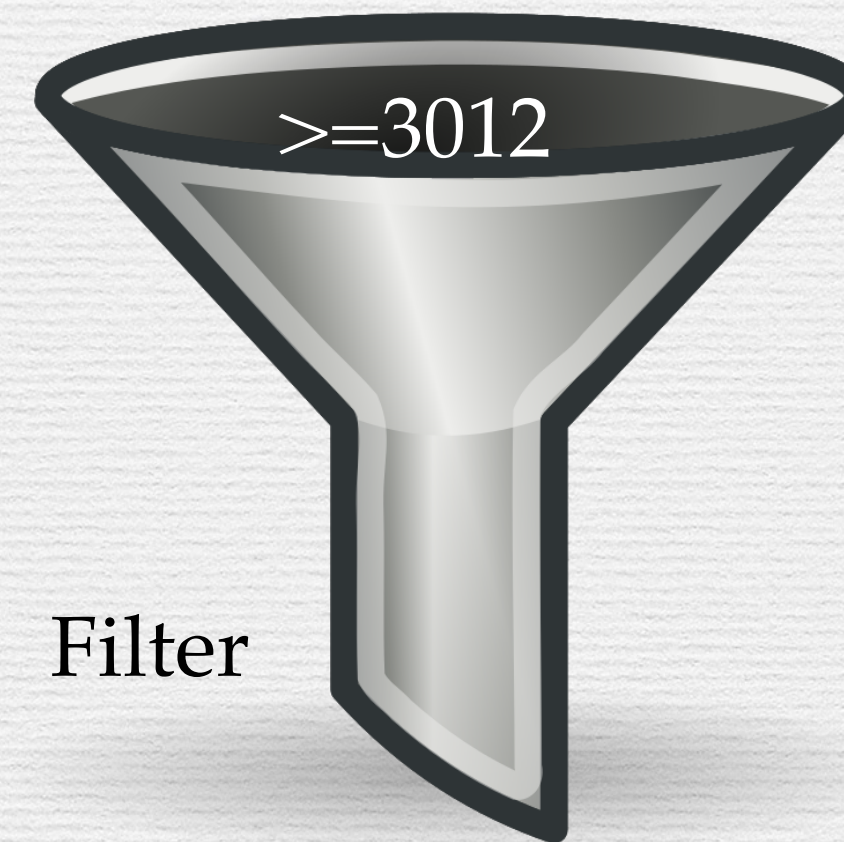
Top-2 delegates

3012	3210
------	------

Dr. Top- k : Filter Enabled Concatenation

Compute top-2

	Subrange 0				Subrange 1				Subrange 2				Subrange 3			
Input vector	2001	101	1323	3012	212	1132	2310	2313	3000	3010	1002	3210	1020	333	2321	2003

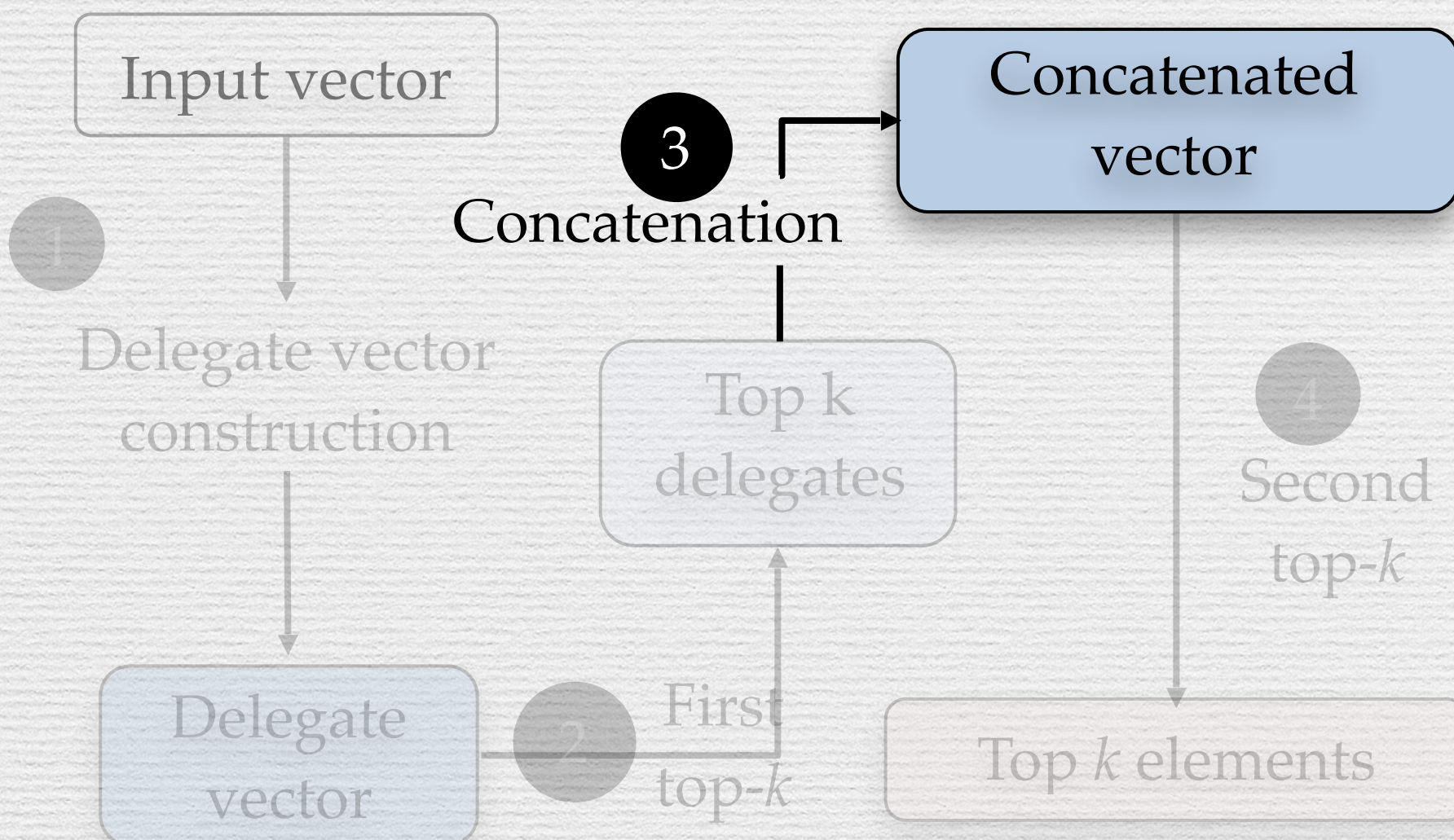


Delegate vector

3012	2313	3210	2321
------	------	------	------

Top-2 delegates

3012	3210
------	------

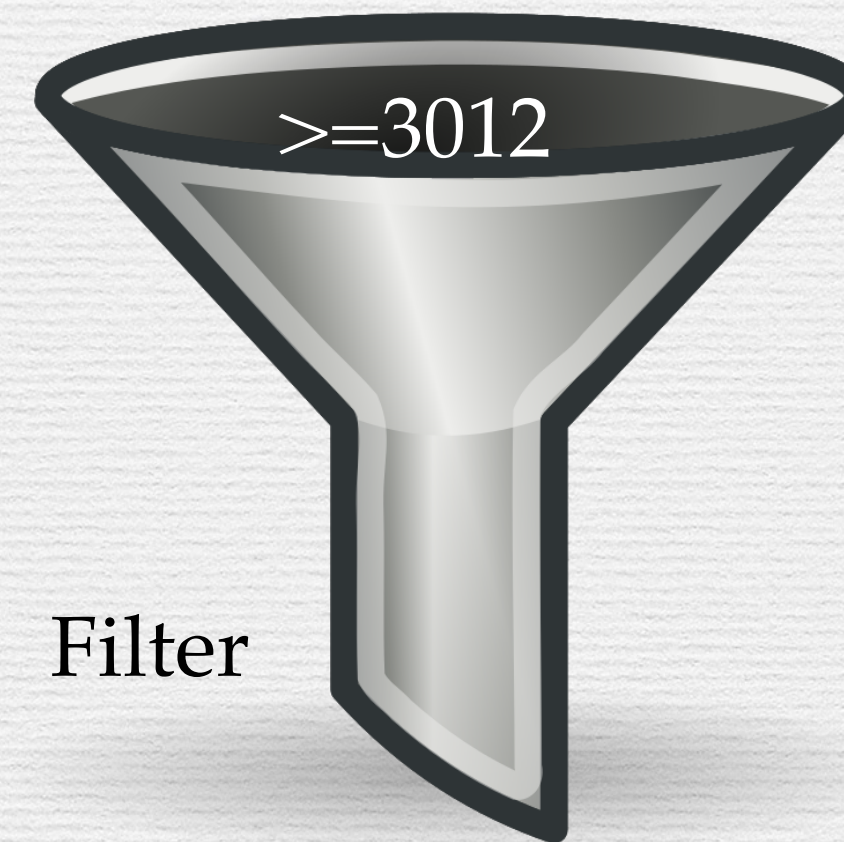


Changes for optimization

Dr. Top- k : Filter Enabled Concatenation

Compute top-2

	Subrange 0				Subrange 1				Subrange 2				Subrange 3			
Input vector	2001	101	1323	3012	212	1132	2310	2313	3000	3010	1002	3210	1020	333	2321	2003



Delegate vector

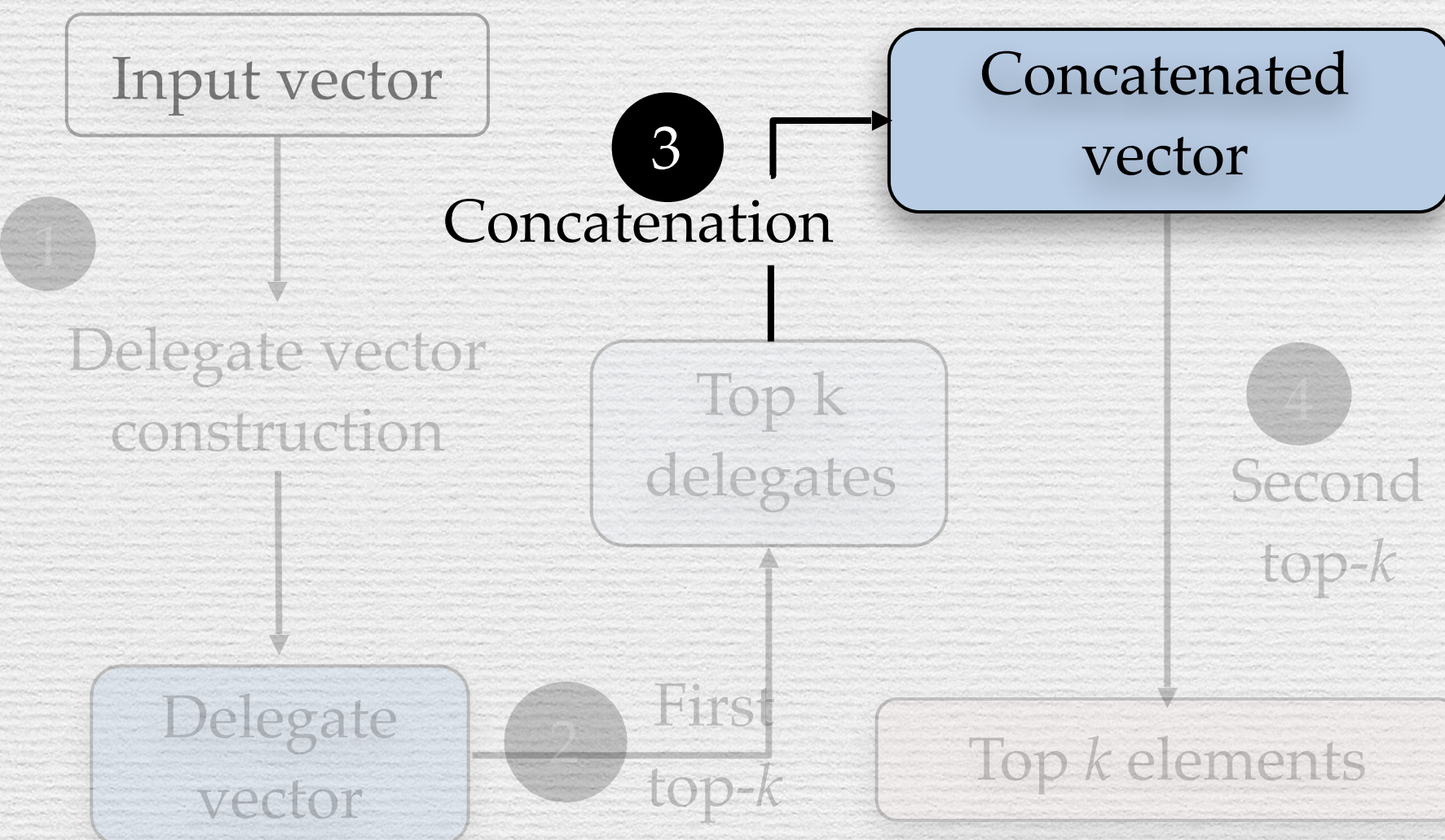
3012	2313	3210	2321
------	------	------	------

Top-2 delegates

3012	3210
------	------

Concatenated vector

3012	3210
------	------

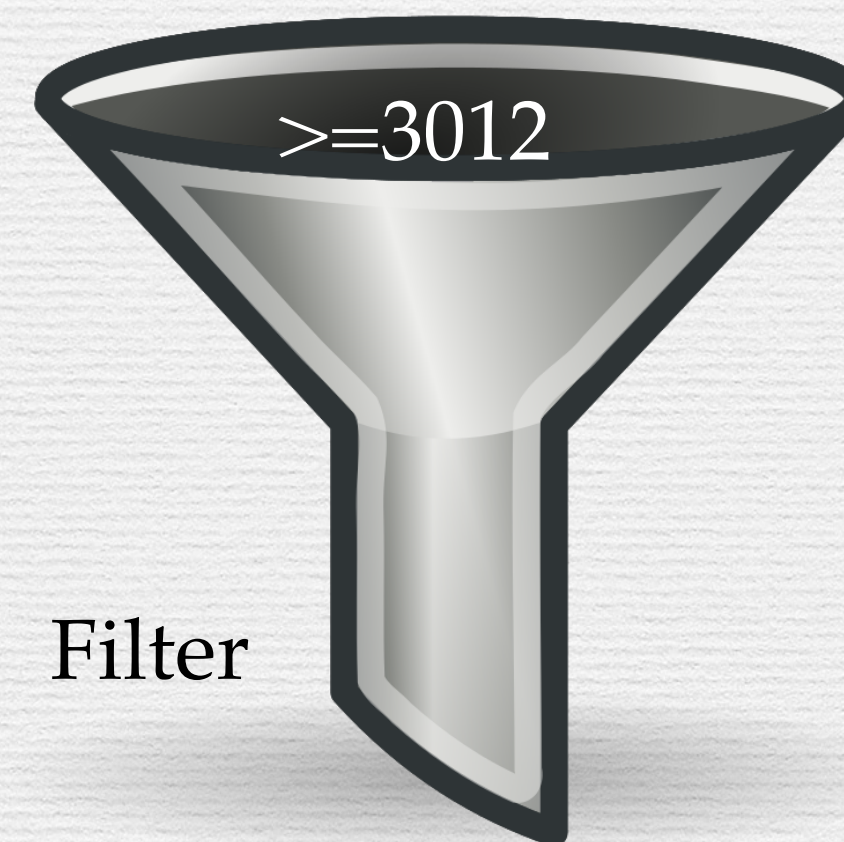


Changes for optimization

Dr. Top- k : Filter Enabled Concatenation

Compute top-2

	Subrange 0				Subrange 1				Subrange 2				Subrange 3			
Input vector	2001	101	1323	3012	212	1132	2310	2313	3000	3010	1002	3210	1020	333	2321	2003



Delegate vector

3012	2313	3210	2321
------	------	------	------

Top-2 delegates

3012	3210
------	------

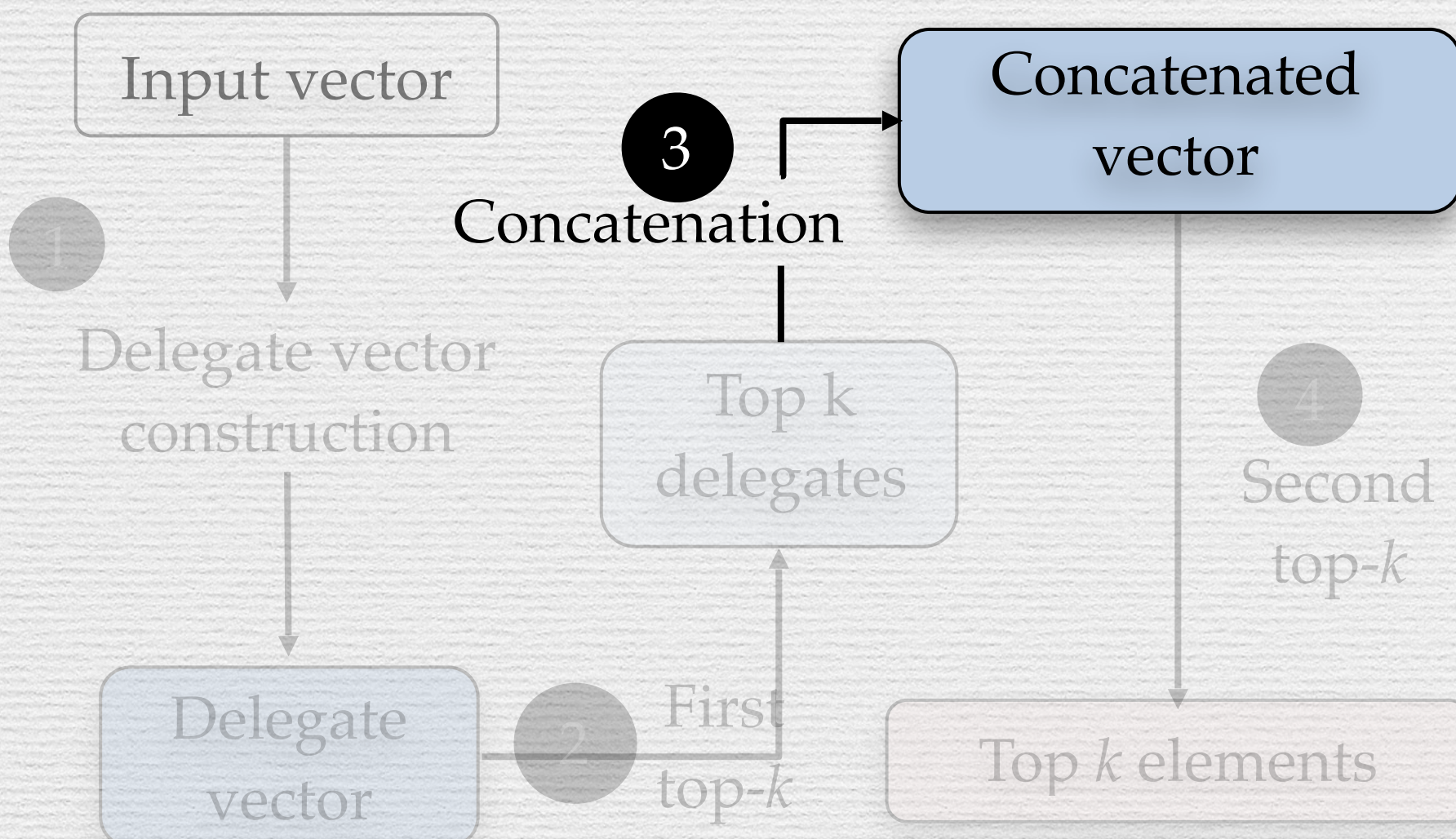
Concatenated vector

3012	3210
------	------

Top-2 elements

3012	3210
------	------

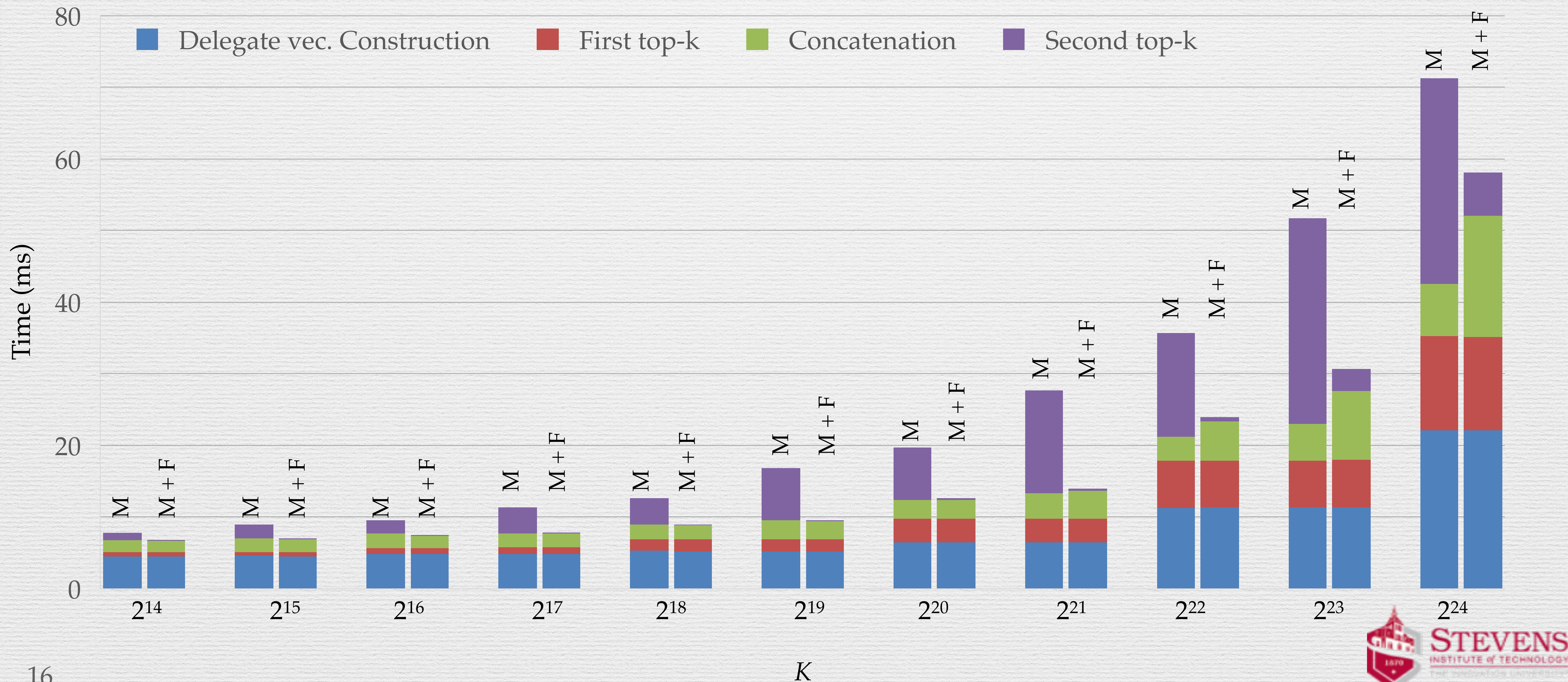
Reduces workload for second top- k .



Changes for optimization

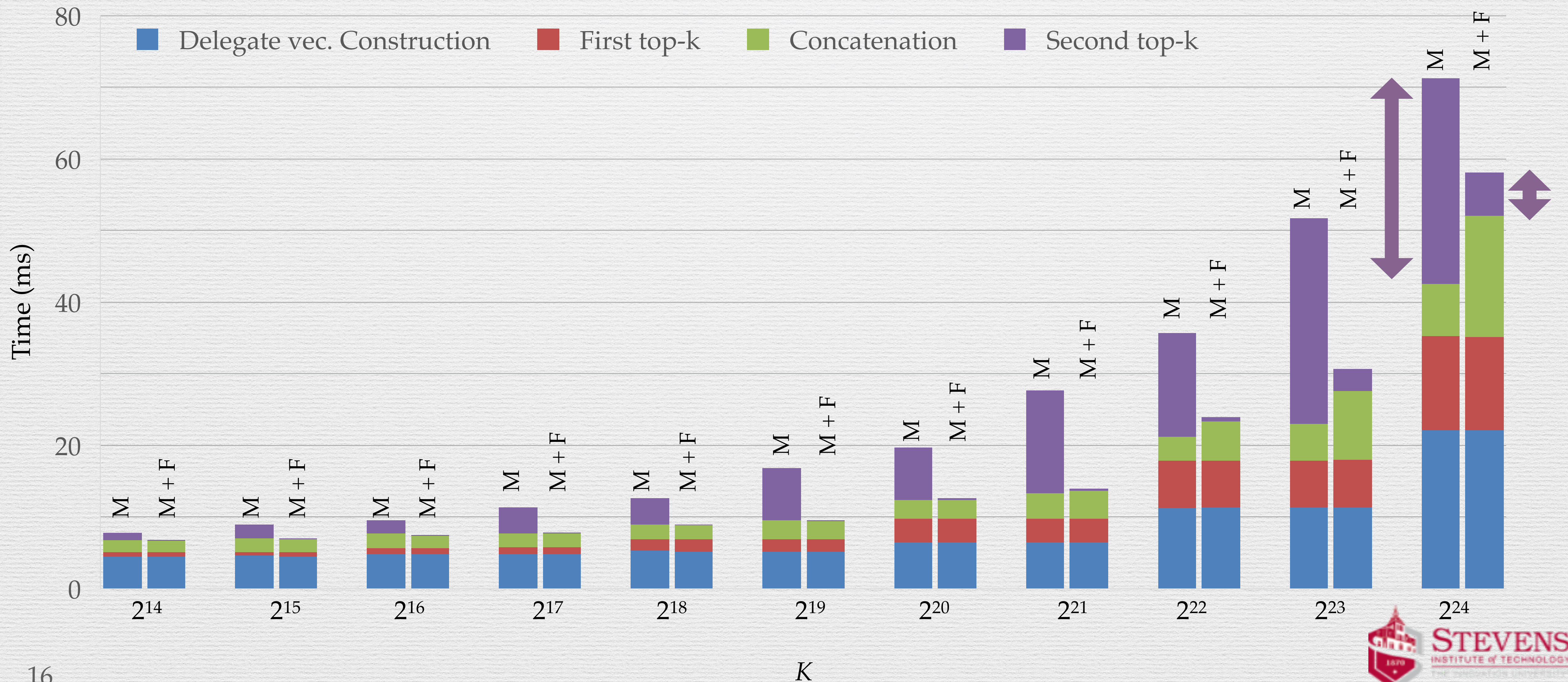
Dr. Top- k : Filter Enabled Concatenation

Max delegate (M) + Filter (F): Time breakdown of Dr.Top-k



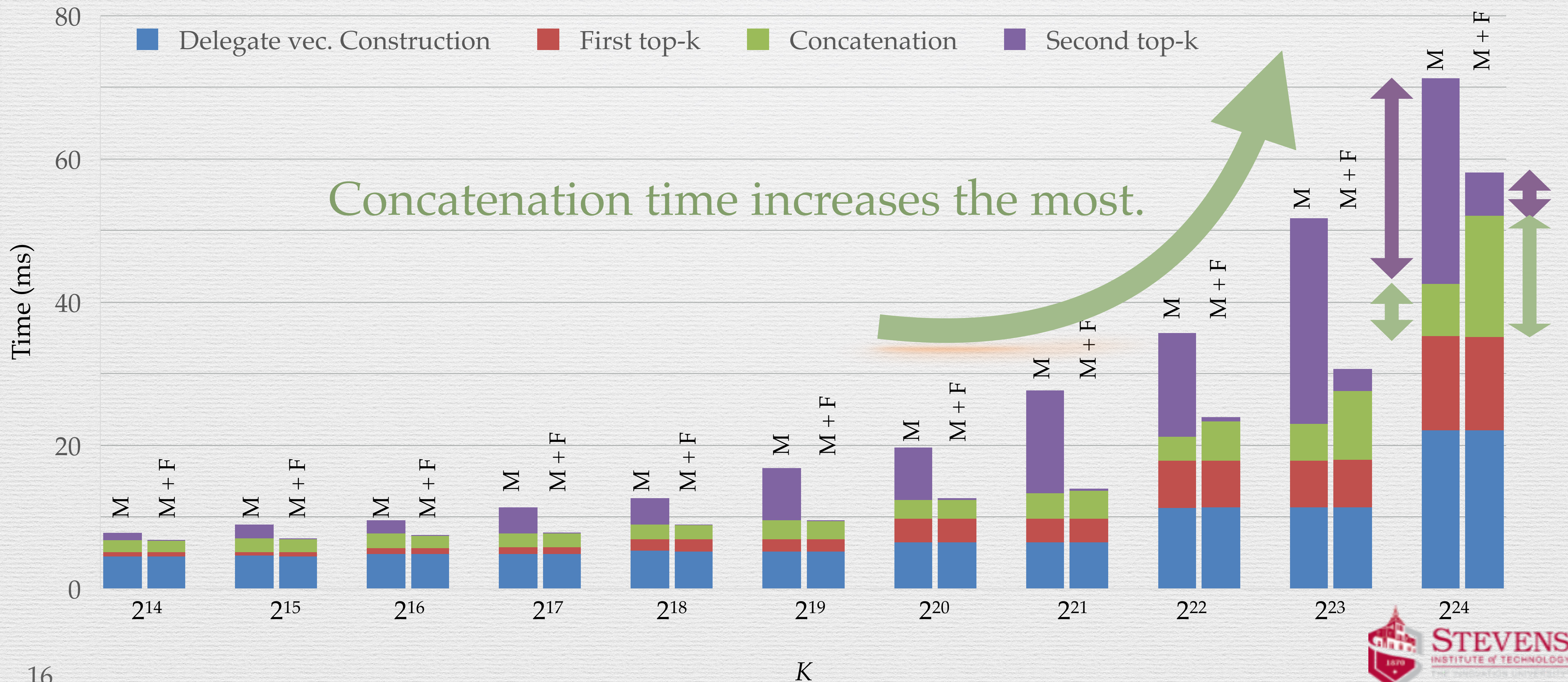
Dr. Top- k : Filter Enabled Concatenation

Max delegate (M) + Filter (F): Time breakdown of Dr.Top-k



Dr. Top- k : Filter Enabled Concatenation

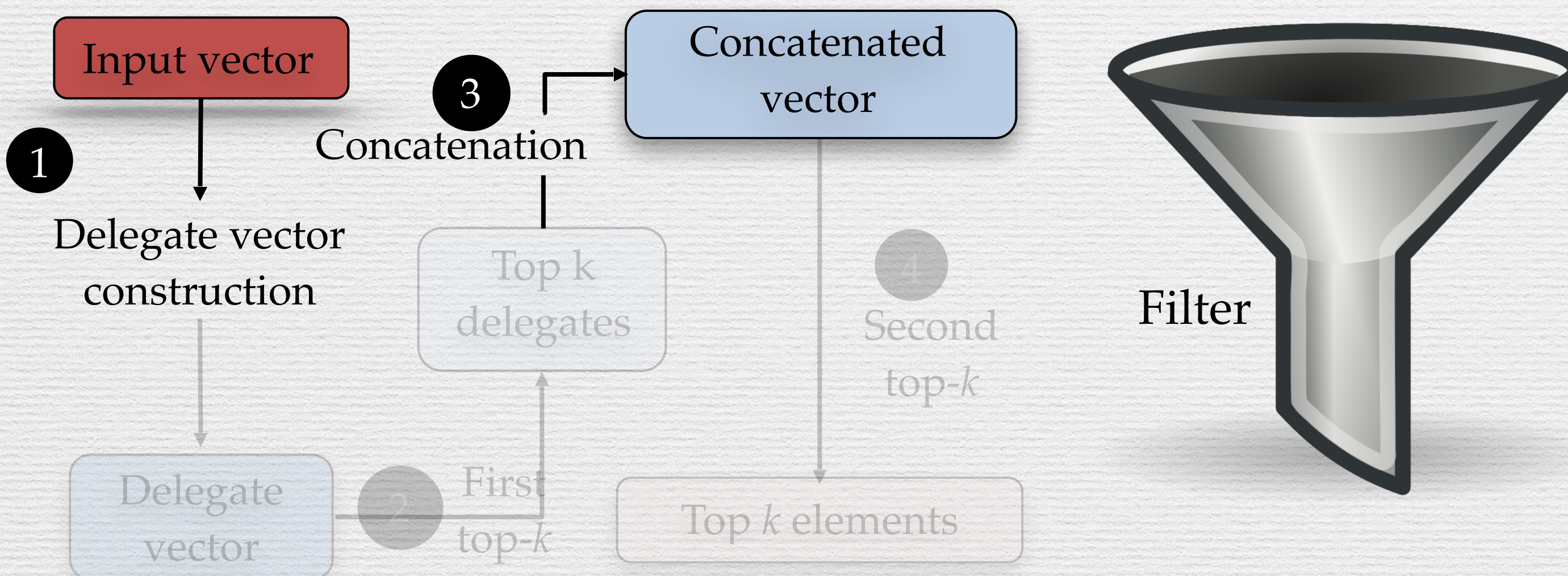
Max delegate (M) + Filter (F): Time breakdown of Dr.Top-k



Dr. Top- k : β Delegate Top- k

Compute top-3

	Subrange 0				Subrange 1				Subrange 2				Subrange 3			
Input vector	2001	101	1323	3012	212	1132	2310	2313	3000	3010	1002	3210	1020	333	2321	2003

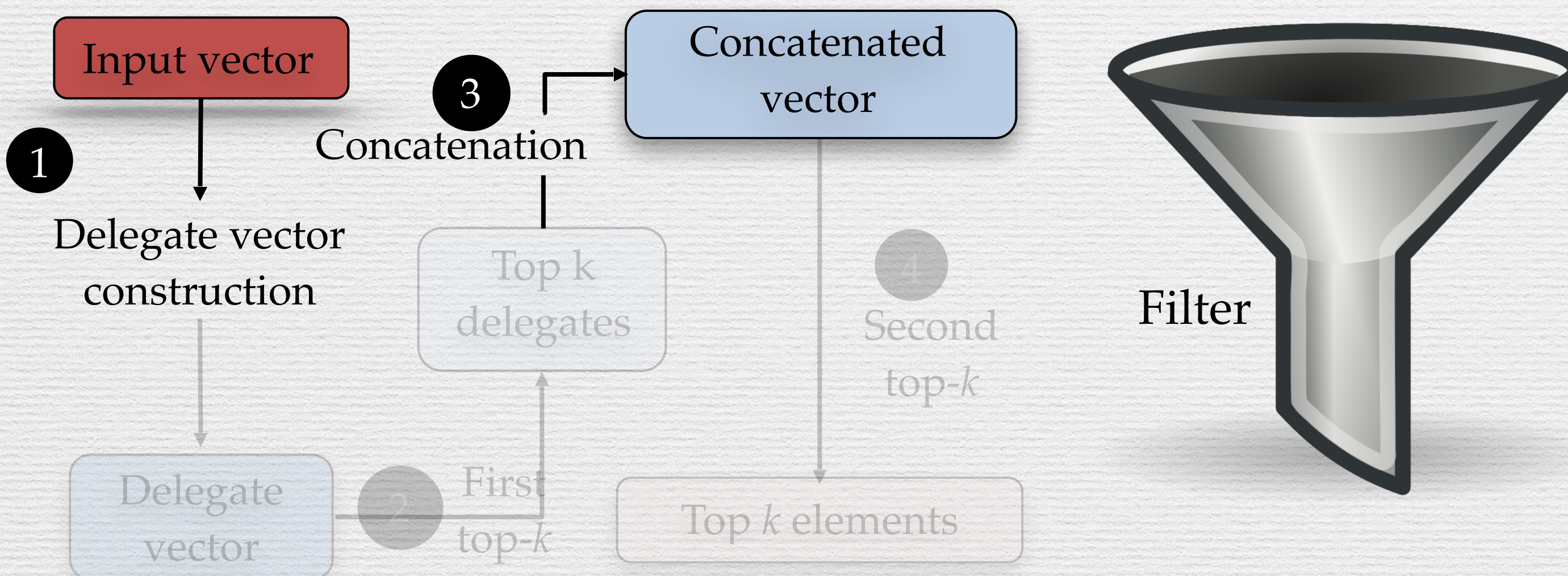


Changes for optimization

Dr. Top- k : β Delegate Top- k

Compute top-3

	Subrange 0				Subrange 1				Subrange 2				Subrange 3			
Input vector	2001	101	1323	3012	212	1132	2310	2313	3000	3010	1002	3210	1020	333	2321	2003

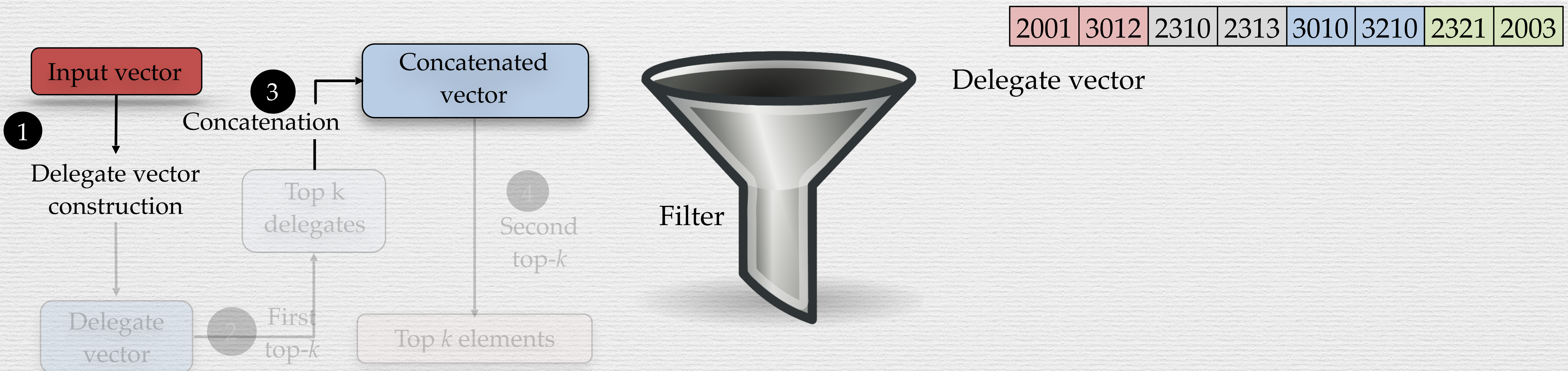


Changes for optimization

Dr. Top- k : β Delegate Top- k

Compute top-3

	Subrange 0				Subrange 1				Subrange 2				Subrange 3			
Input vector	2001	101	1323	3012	212	1132	2310	2313	3000	3010	1002	3210	1020	333	2321	2003

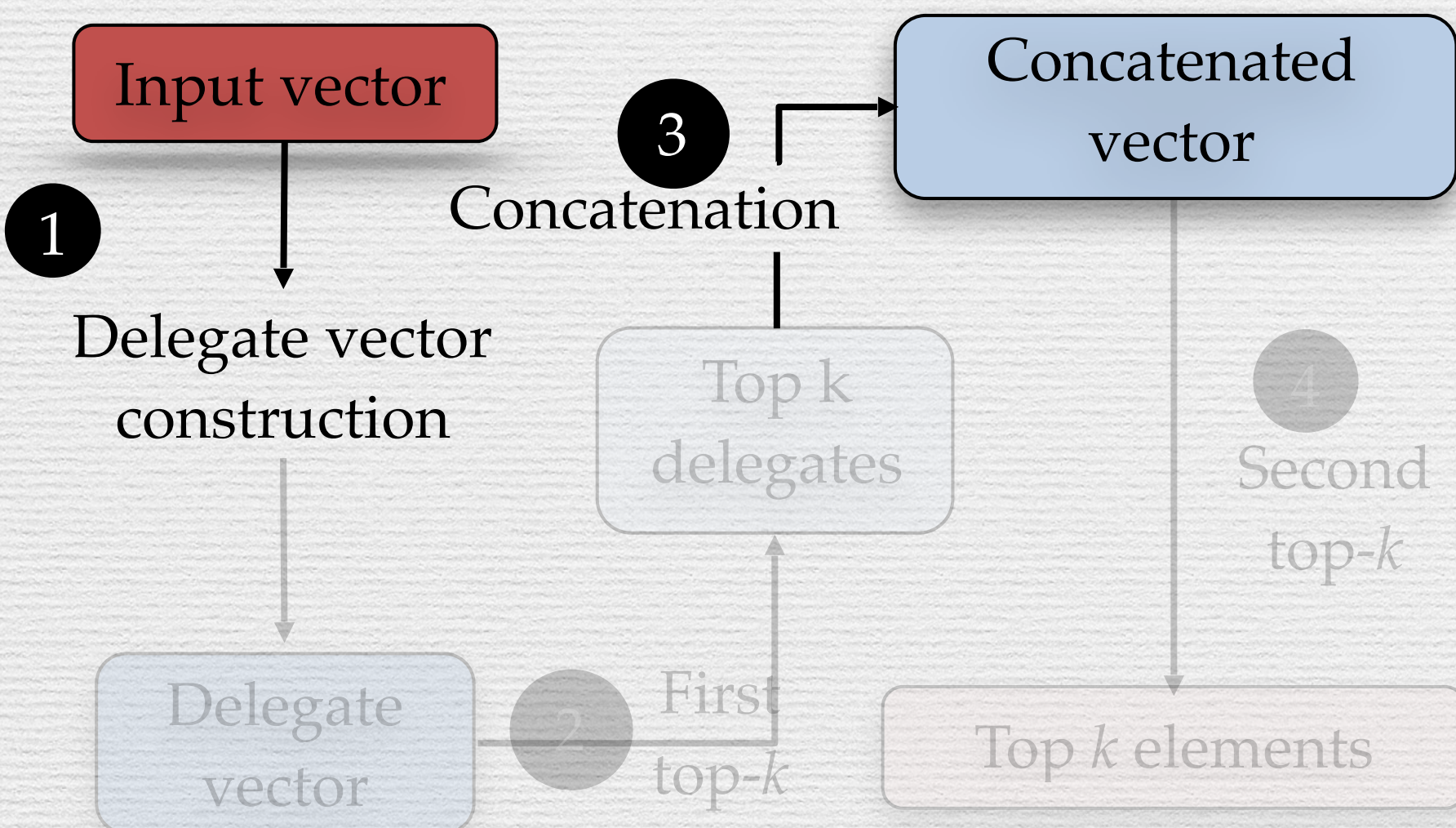


Changes for optimization

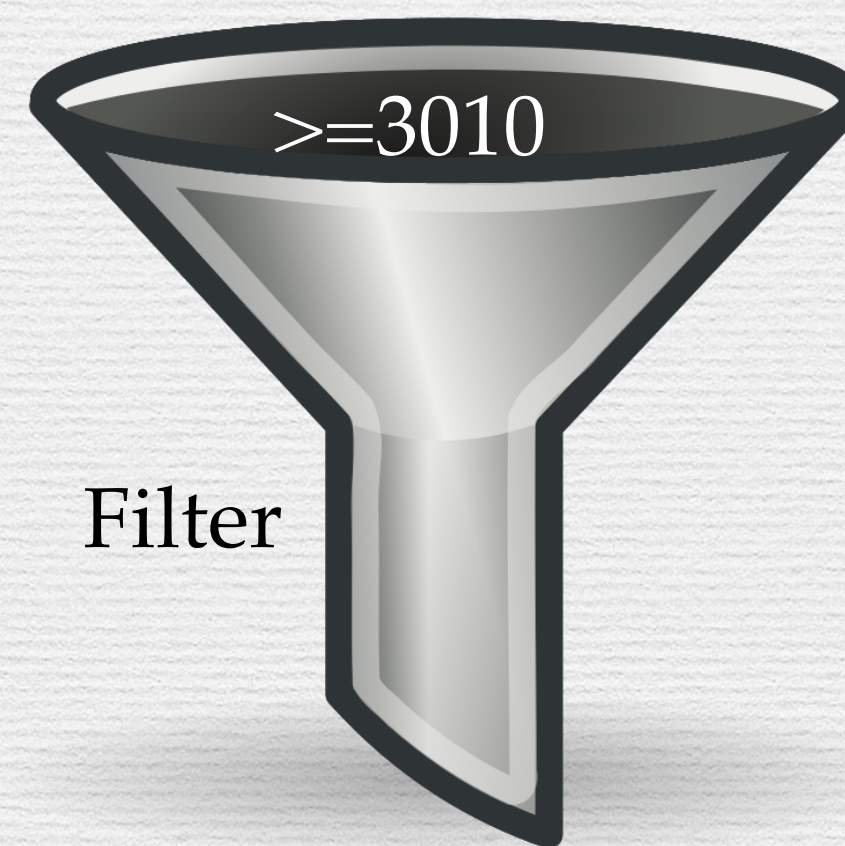
Dr. Top- k : β Delegate Top- k

Compute top-3

	Subrange 0				Subrange 1				Subrange 2				Subrange 3			
Input vector	2001	101	1323	3012	212	1132	2310	2313	3000	3010	1002	3210	1020	333	2321	2003



Changes for optimization



Filter

2001	3012	2310	2313	3010	3210	2321	2003
------	------	------	------	------	------	------	------

Delegate vector

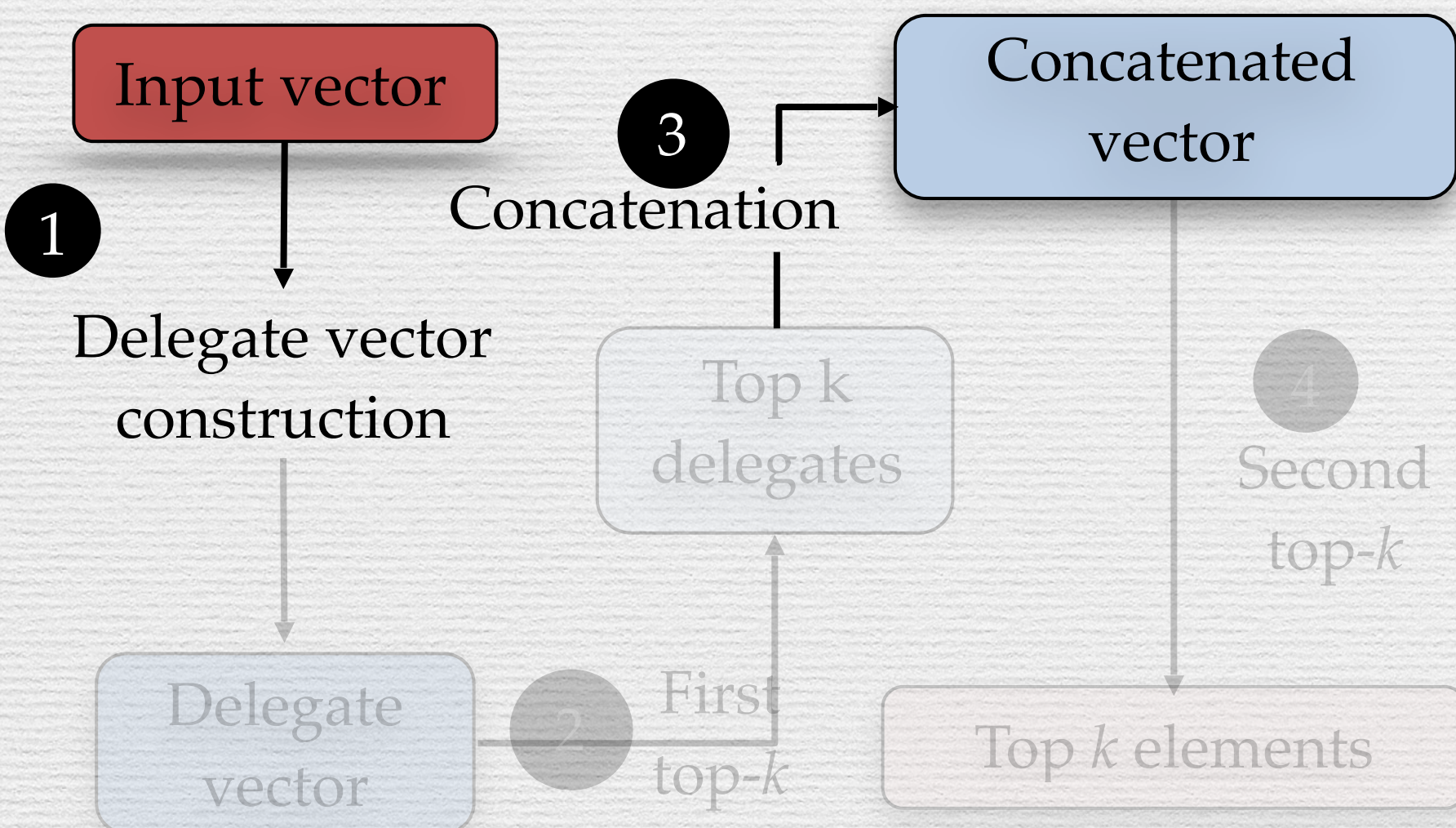
Top-3 delegates

3012	3010	3210
------	------	------

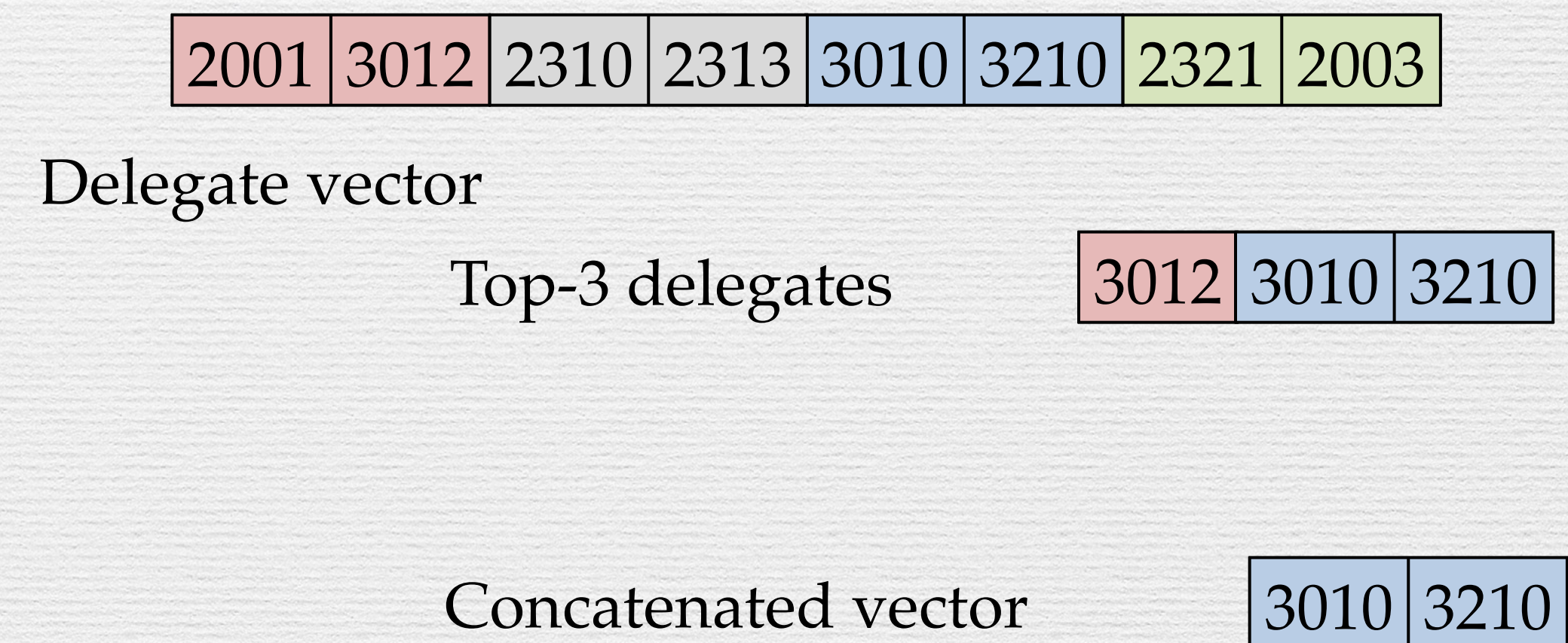
Dr. Top- k : β Delegate Top- k

Compute top-3

	Subrange 0				Subrange 1				Subrange 2				Subrange 3			
Input vector	2001	101	1323	3012	212	1132	2310	2313	3000	3010	1002	3210	1020	333	2321	2003



Changes for optimization

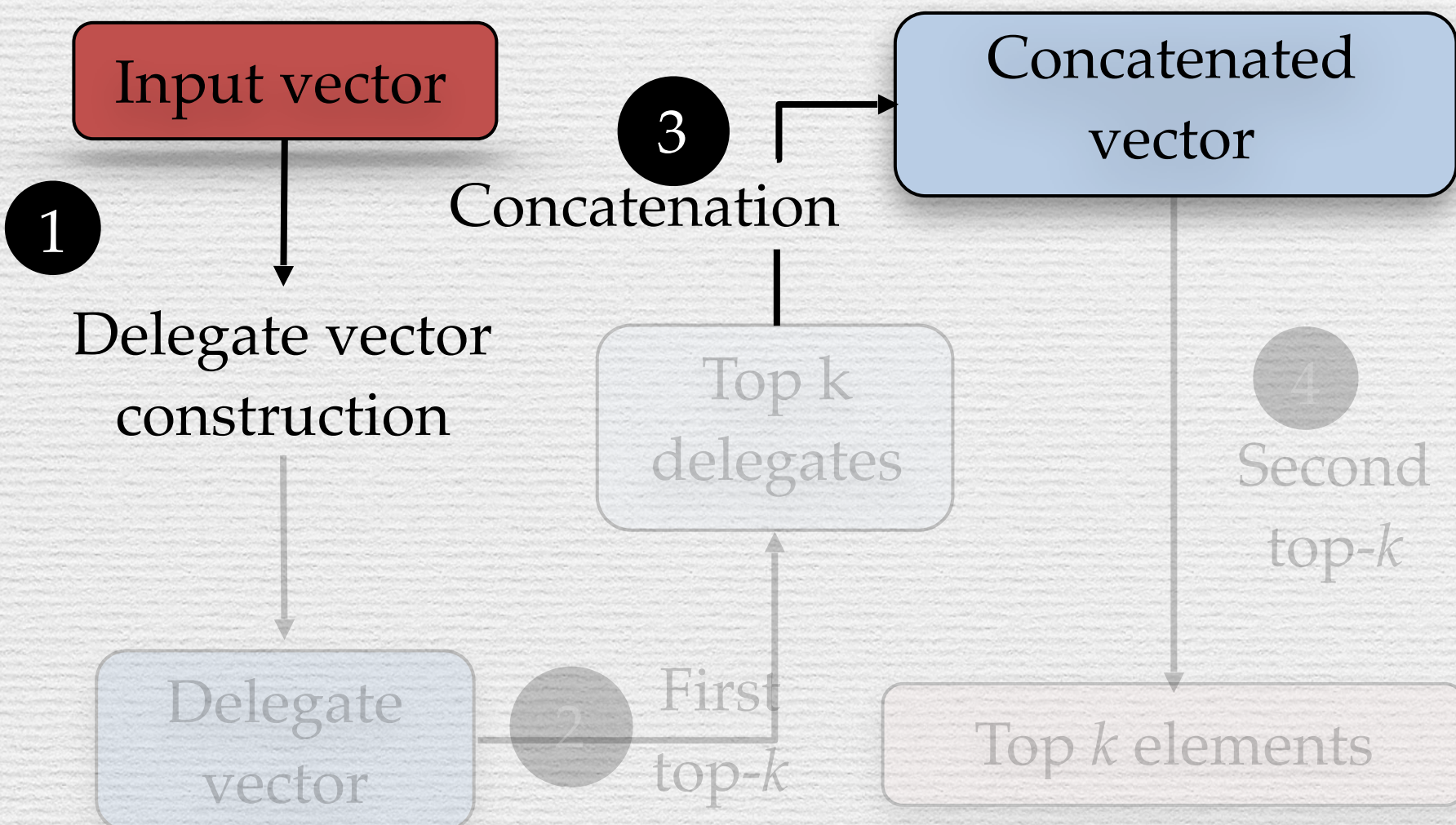


Dr. Top- k : β Delegate Top- k

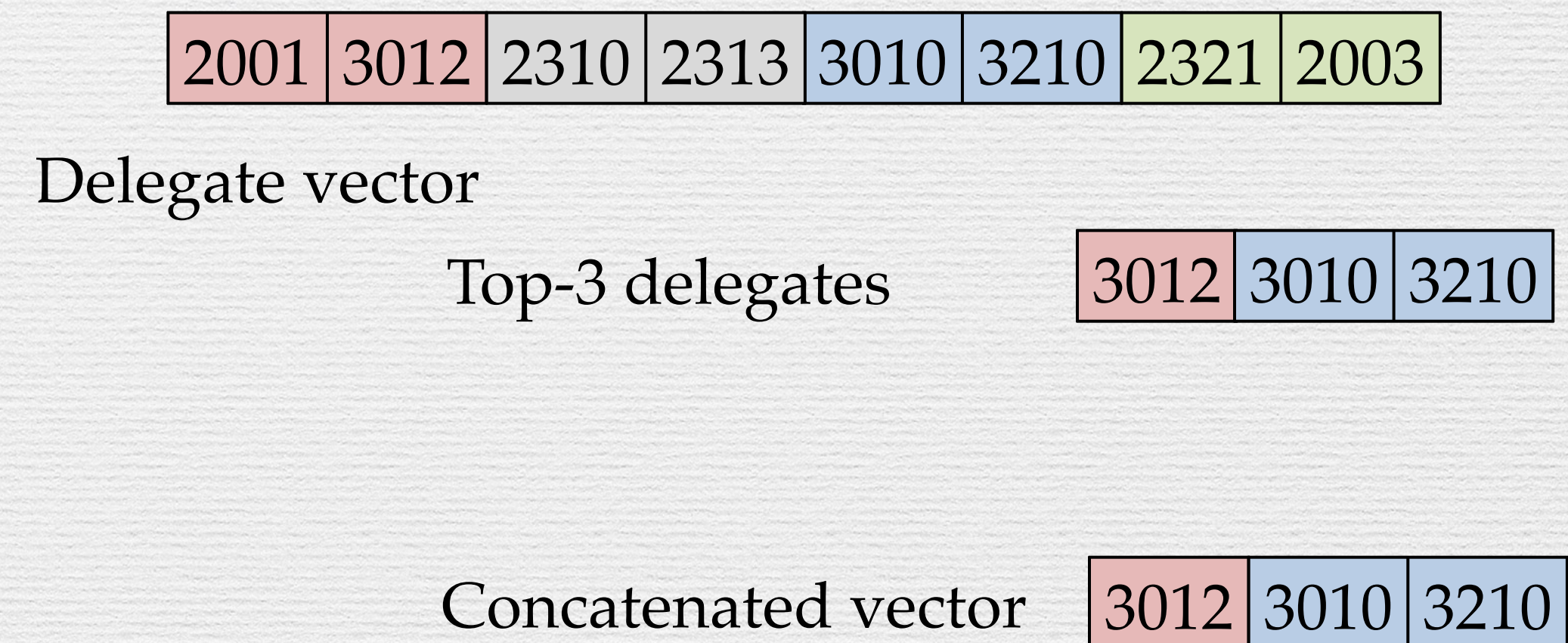
Compute top-3

Input vector

Subrange 0				Subrange 1				Subrange 2				Subrange 3			
2001	101	1323	3012	212	1132	2310	2313	3000	3010	1002	3210	1020	333	2321	2003



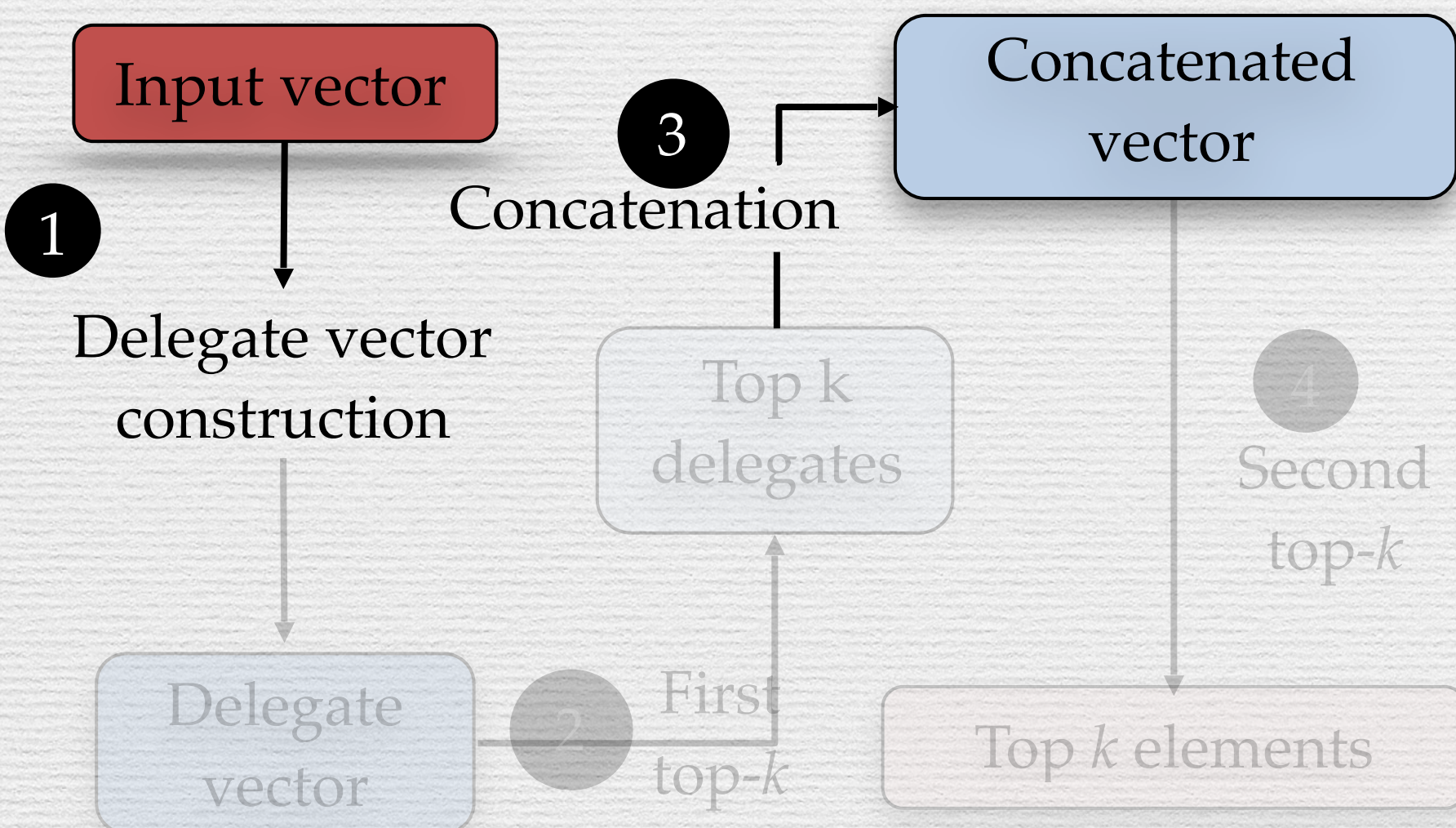
Changes for optimization



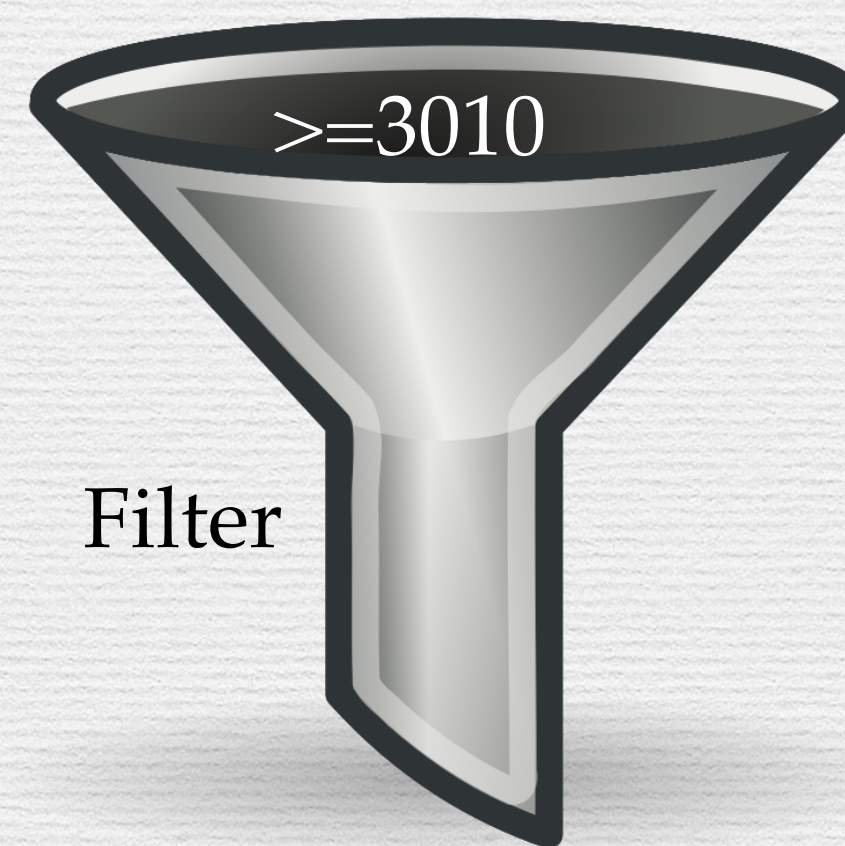
Dr. Top- k : β Delegate Top- k

Compute top-3

	Subrange 0				Subrange 1				Subrange 2				Subrange 3			
Input vector	2001	101	1323	3012	212	1132	2310	2313	3000	3010	1002	3210	1020	333	2321	2003



Changes for optimization



2001	3012	2310	2313	3010	3210	2321	2003
------	------	------	------	------	------	------	------

Delegate vector

Top-3 delegates

3012	3010	3210
------	------	------

Concatenated vector

3012	3010	3210
------	------	------

Final top-3

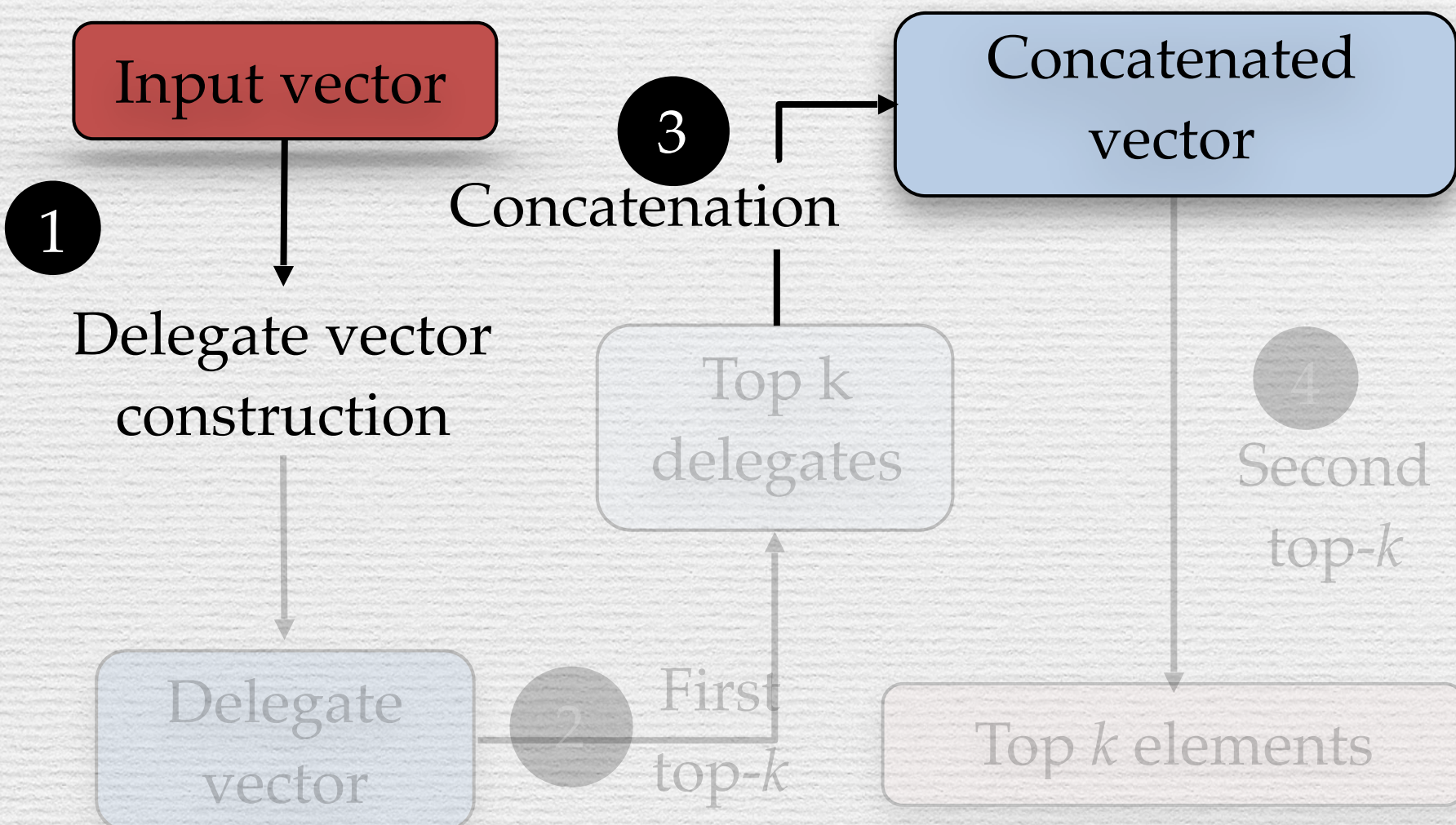
3012	3010	3210
------	------	------

Dr. Top- k : β Delegate Top- k

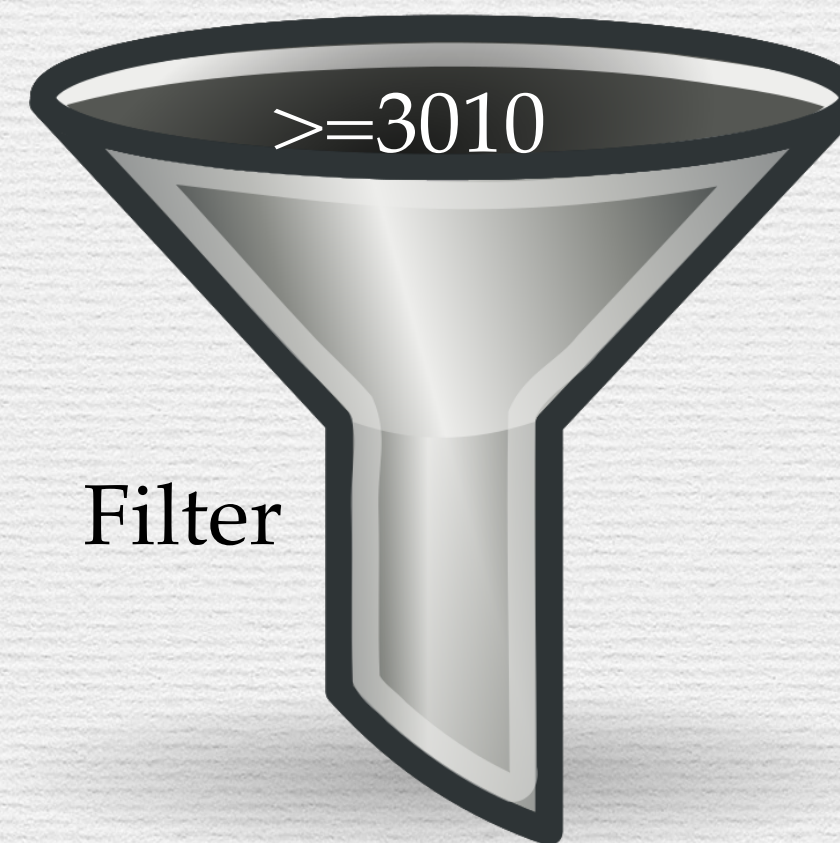
Compute top-3

Input vector

Subrange 0				Subrange 1				Subrange 2				Subrange 3			
2001	101	1323	3012	212	1132	2310	2313	3000	3010	1002	3210	1020	333	2321	2003



Changes for optimization



Delegate vector

2001	3012	2310	2313	3010	3210	2321	2003
------	------	------	------	------	------	------	------

Top-3 delegates

3012	3010	3210
------	------	------

Concatenated vector

3012	3010	3210
------	------	------

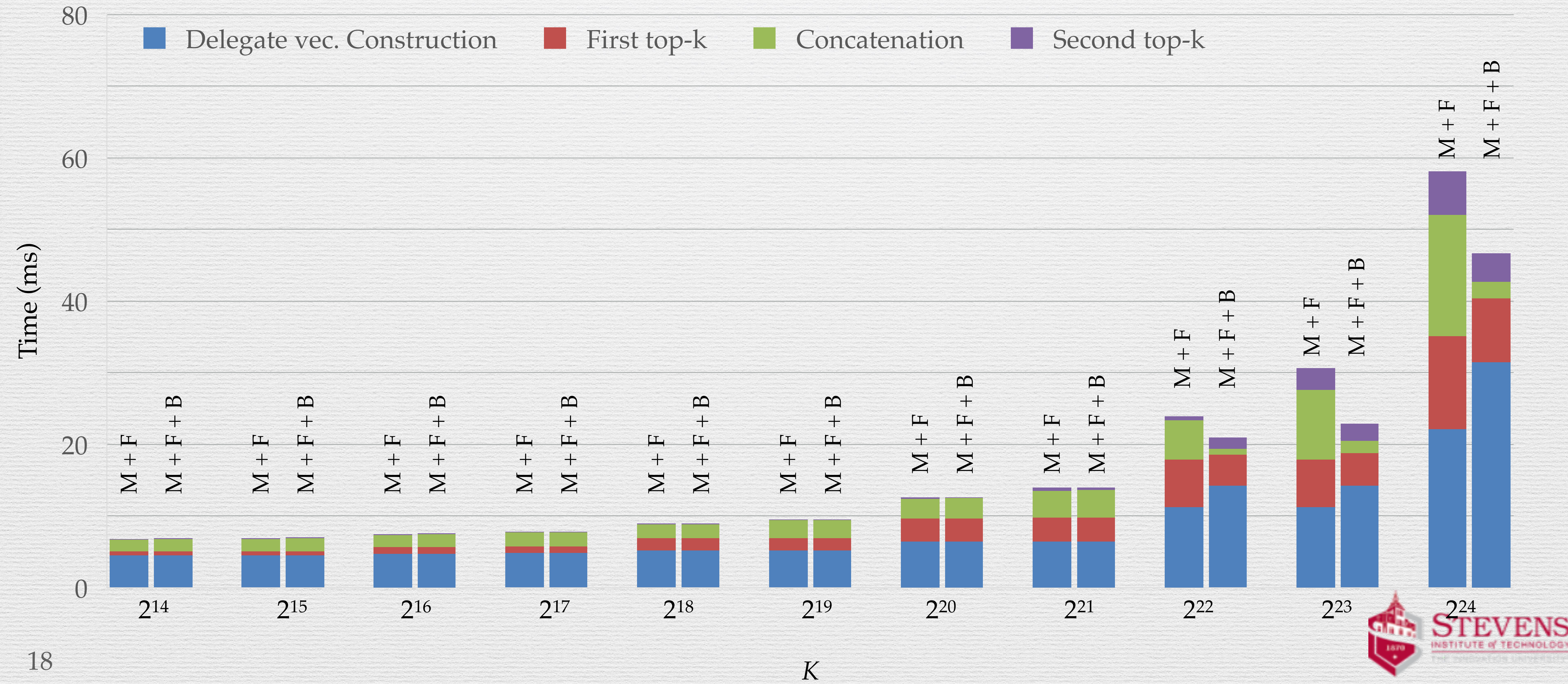
Final top-3

3012	3010	3210
------	------	------

Reduces workload for both the concatenation and second top- k .

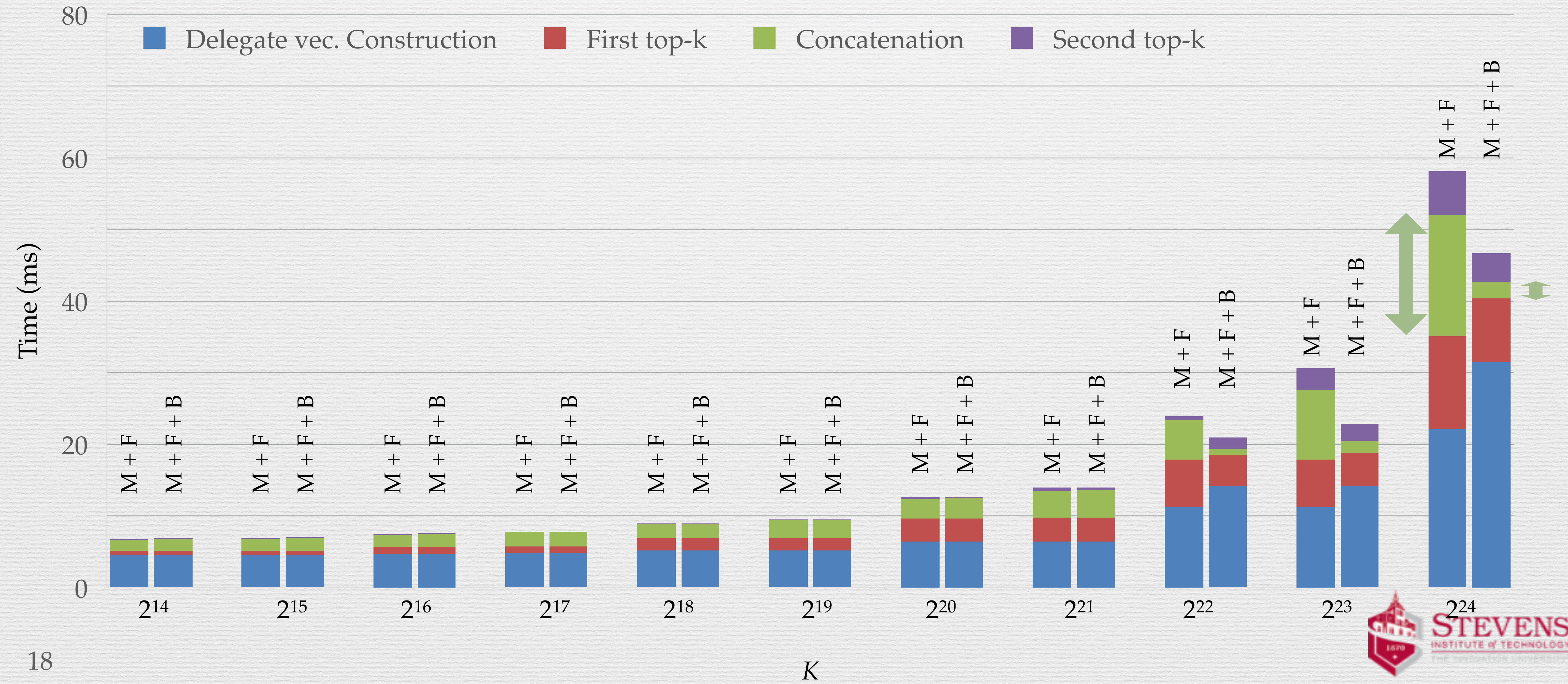
Dr.Top- k With β Delegate Optimization

Max delegate (M) + Filter (F) + β (B): Time breakdown of Dr.Top- k



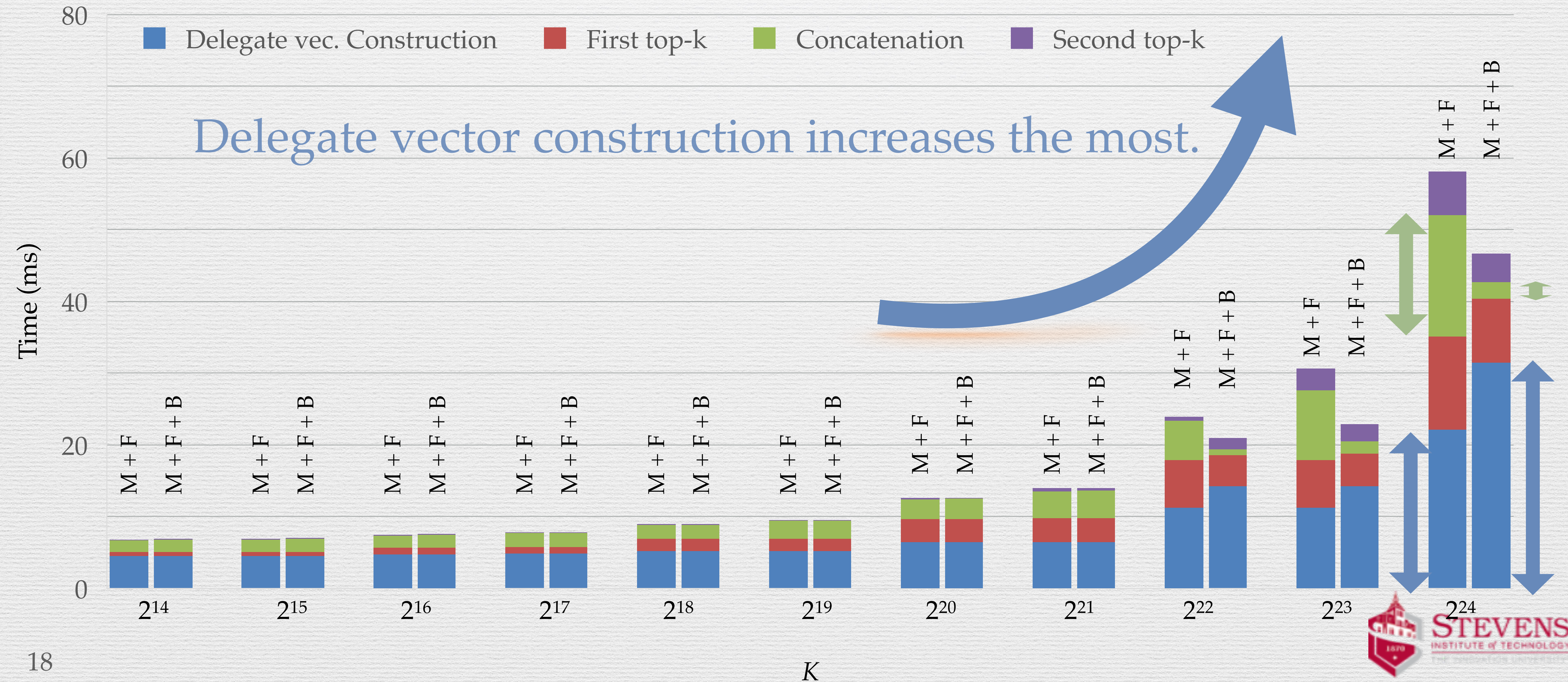
Dr.Top- k With β Delegate Optimization

Max delegate (M) + Filter (F) + β (B): Time breakdown of Dr.Top- k

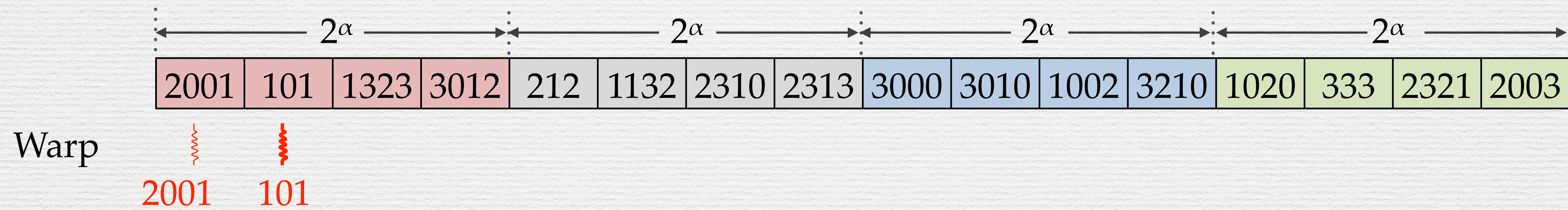


Dr.Top- k With β Delegate Optimization

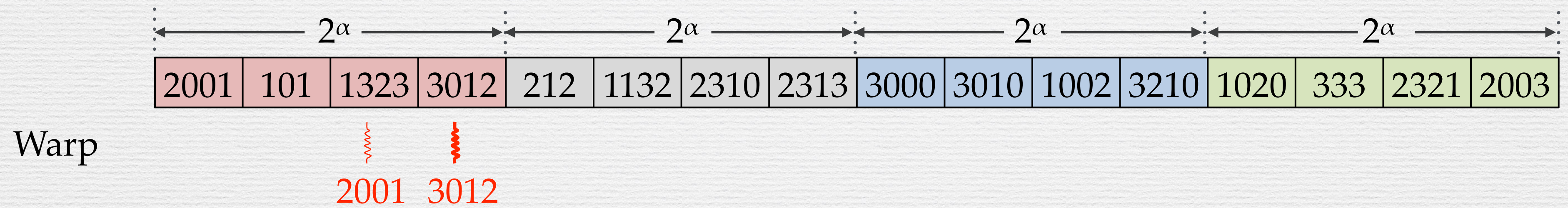
Max delegate (M) + Filter (F) + β (B): Time breakdown of Dr.Top- k



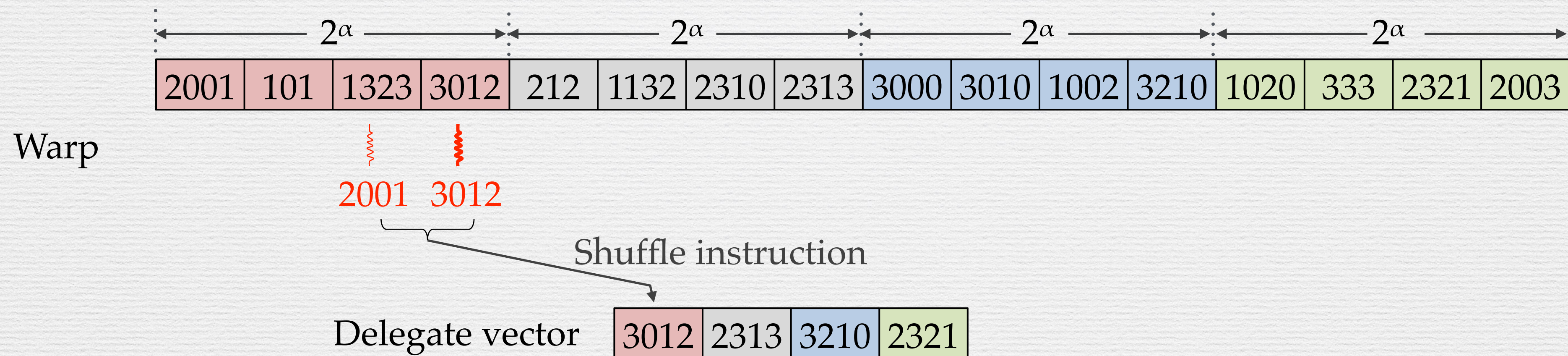
Dr.Top- k : Delegate Vec. Construction Optimization



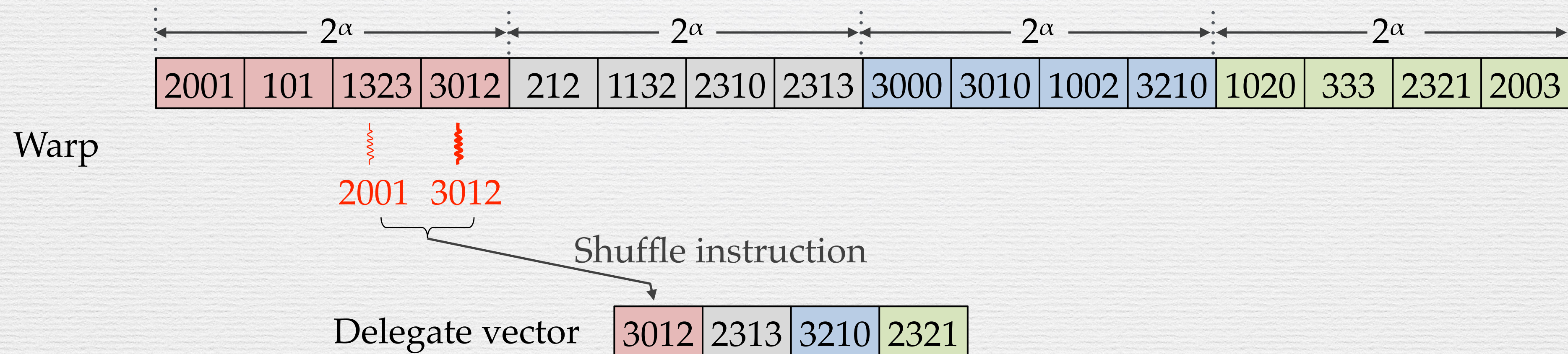
Dr.Top- k : Delegate Vec. Construction Optimization



Dr.Top- k : Delegate Vec. Construction Optimization

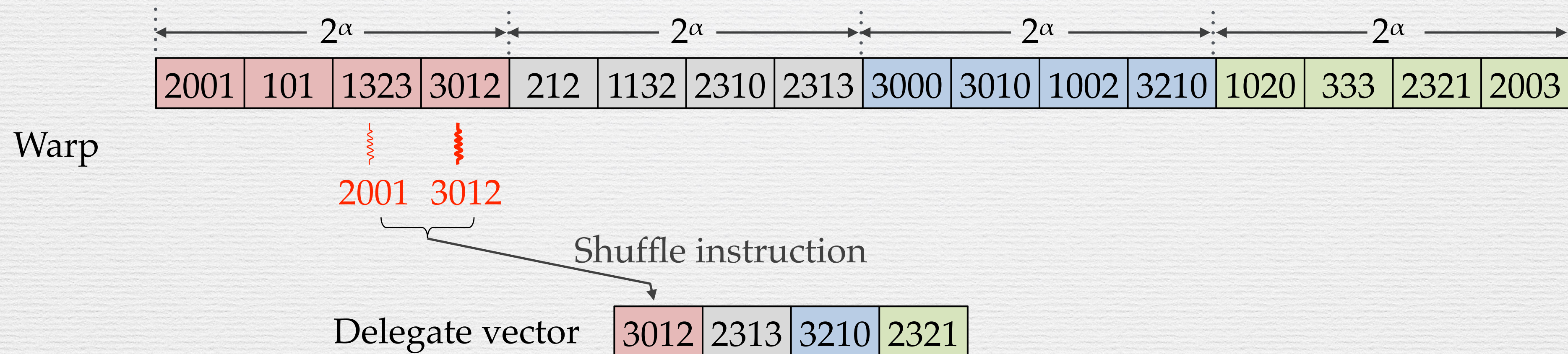


Dr.Top- k : Delegate Vec. Construction Optimization



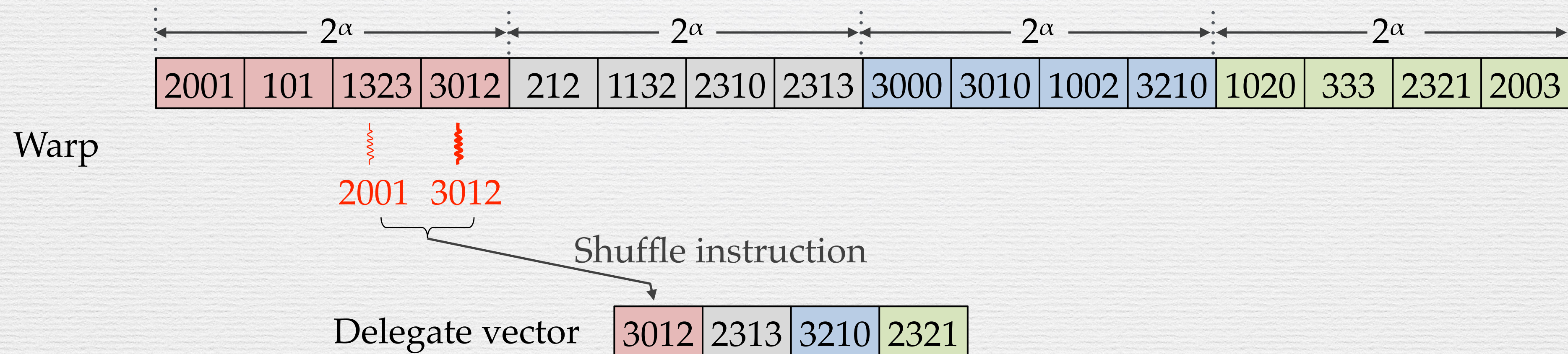
$$\alpha = \frac{\log_2 |V| - \log_2 K + \text{Const.}}{2}$$

Dr.Top- k : Delegate Vec. Construction Optimization



$$\alpha = \frac{\log_2 |V| - \log_2 K + \text{Const.}}{2}$$

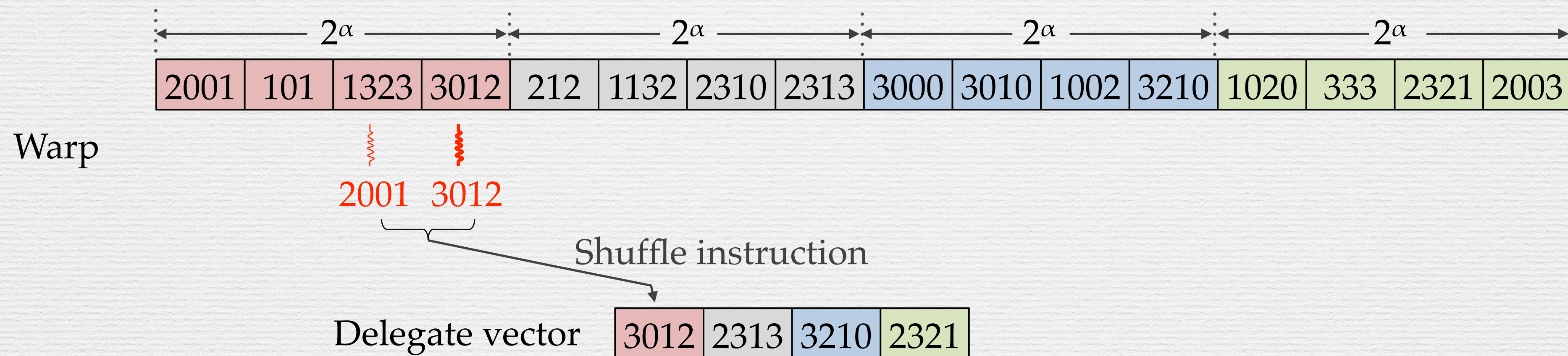
Dr.Top- k : Delegate Vec. Construction Optimization



$$\alpha = \frac{\log_2 |V| - \log_2 K + \text{Const.}}{2}$$

Subrange size (2^α) is small

Dr.Top- k : Delegate Vec. Construction Optimization



$$\alpha = \frac{\log_2 |V| - \log_2 K + \text{Const.}}{2}$$

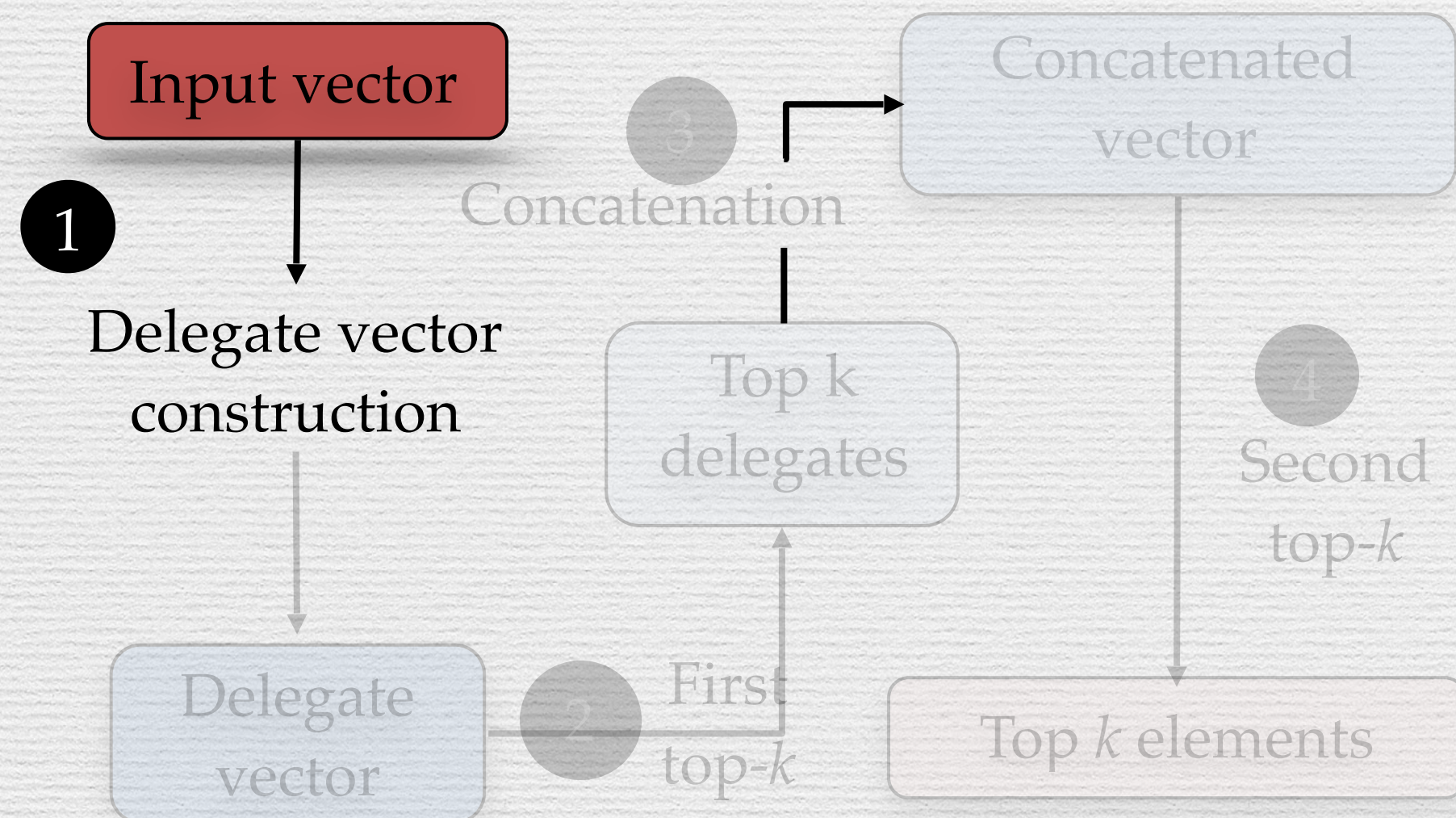
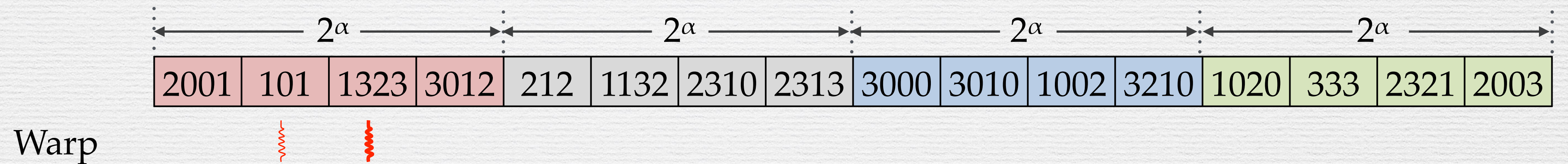
Subrange size (2^α) is small

Problems:

- 1. Failure to saturate GPU warp.
- 2. Overwhelming number of shuffle instruction.

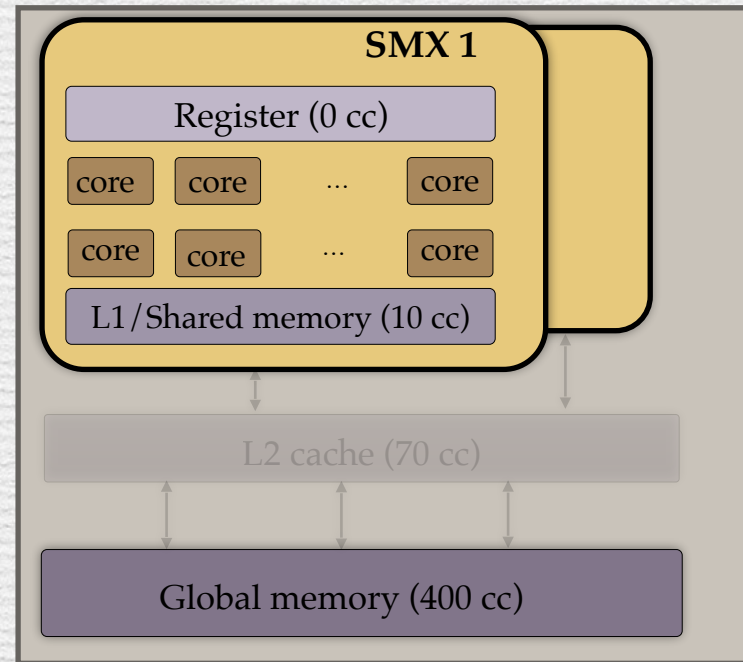
Poor Delegate Vector Construction

Dr.Top- k : Delegate Vec. Construction Optimization

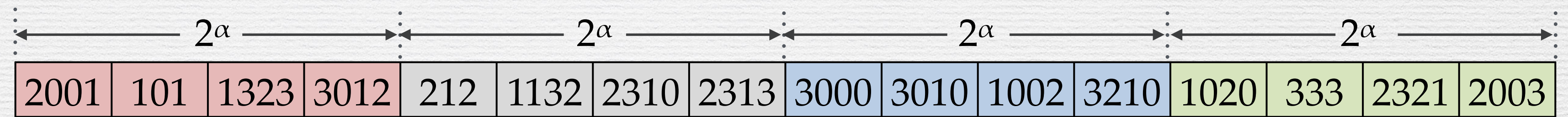


Changes for optimization

Dr.Top- k : Delegate Vec. Construction Optimization

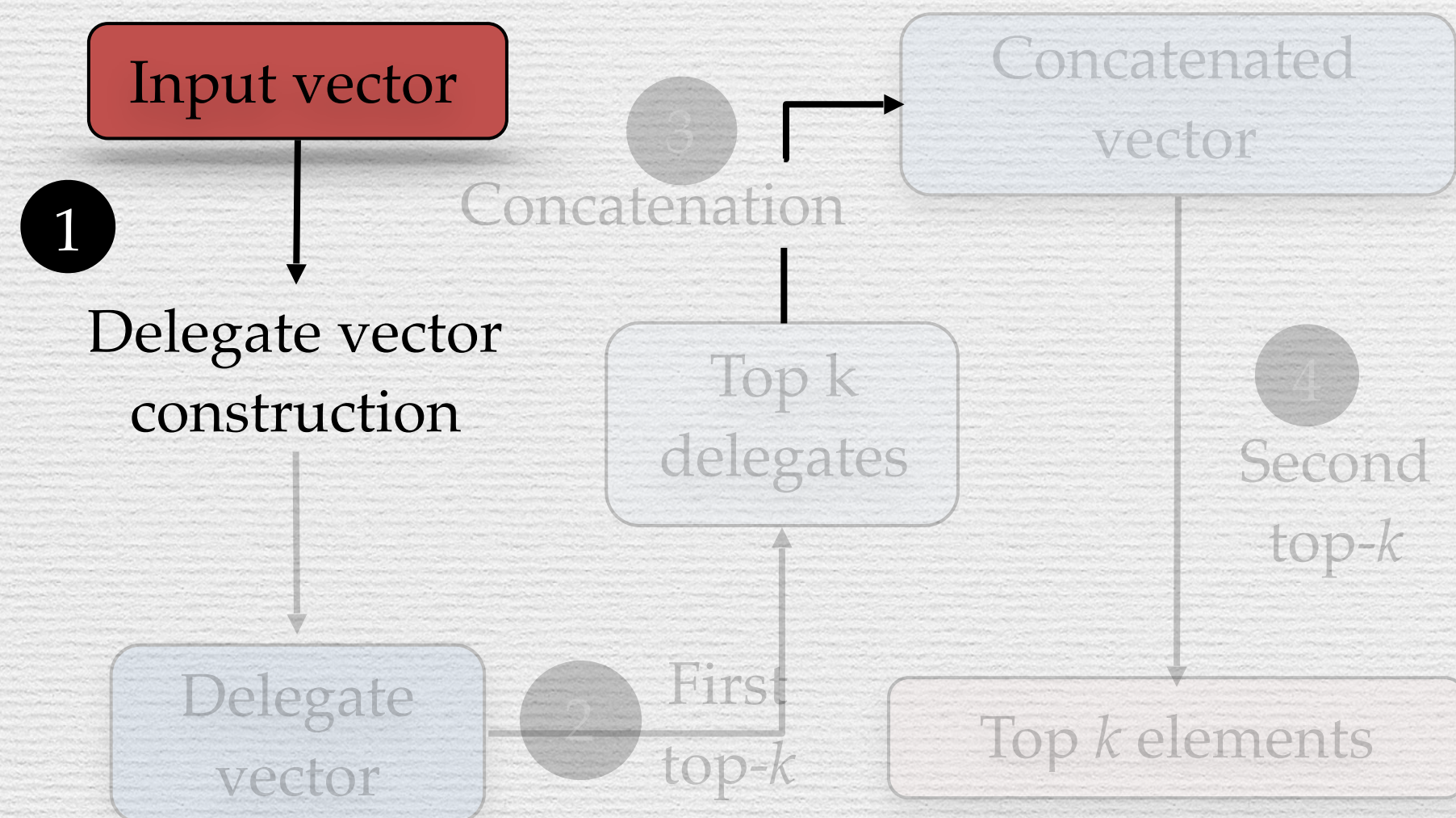
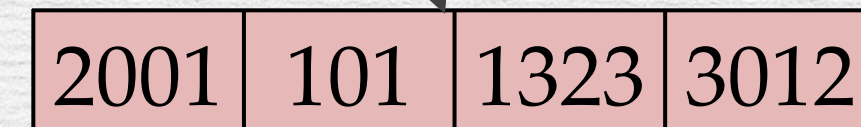


Warp



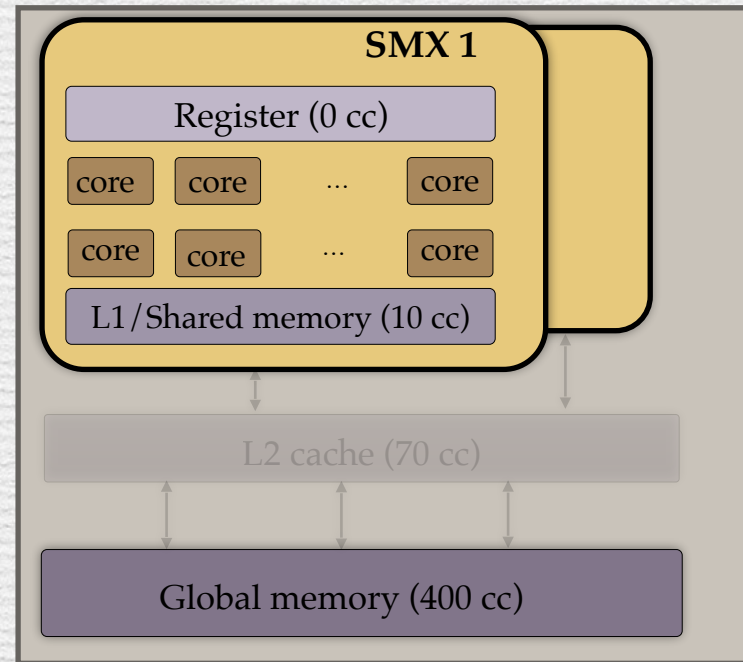
Warp collaborative data loading

Shared memory

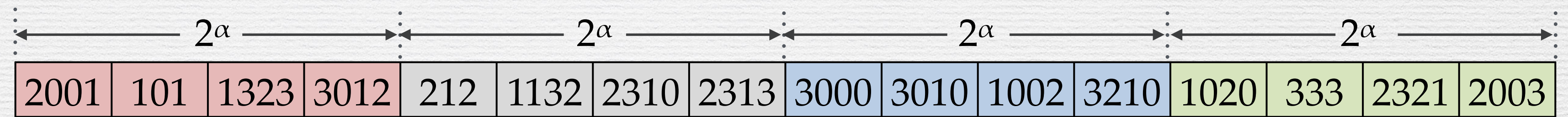


Changes for optimization

Dr.Top- k : Delegate Vec. Construction Optimization

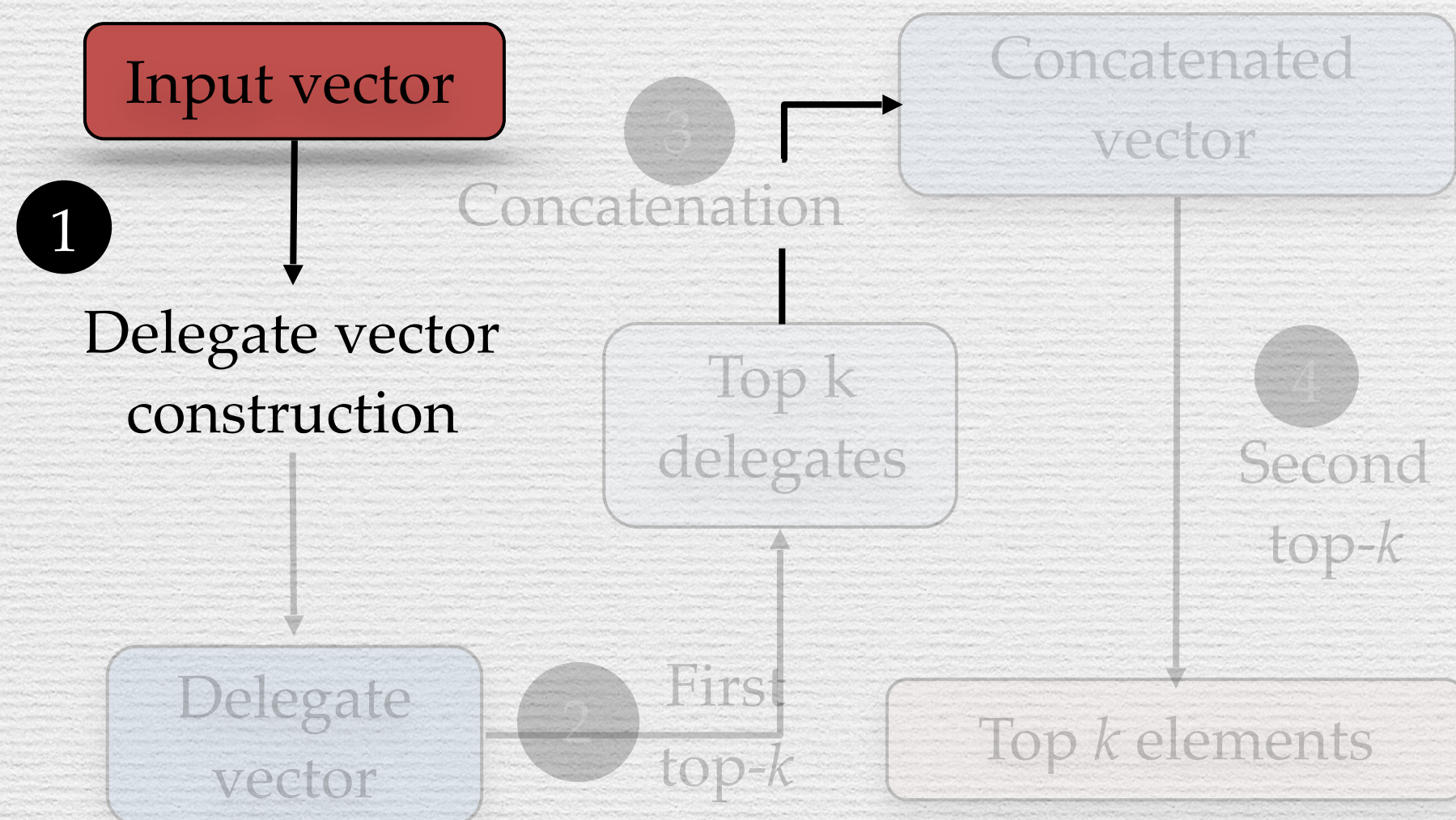
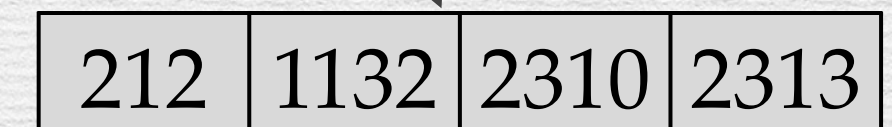
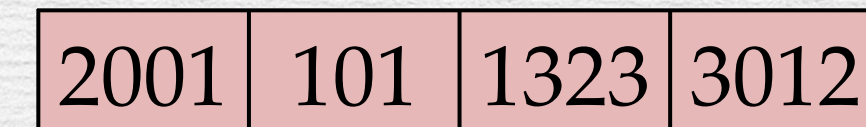


Warp



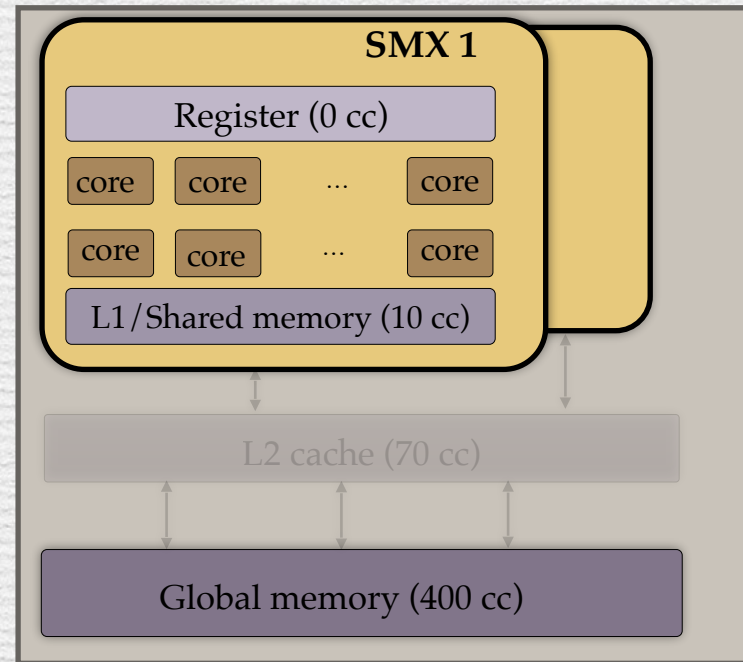
Warp collaborative data loading

Shared memory

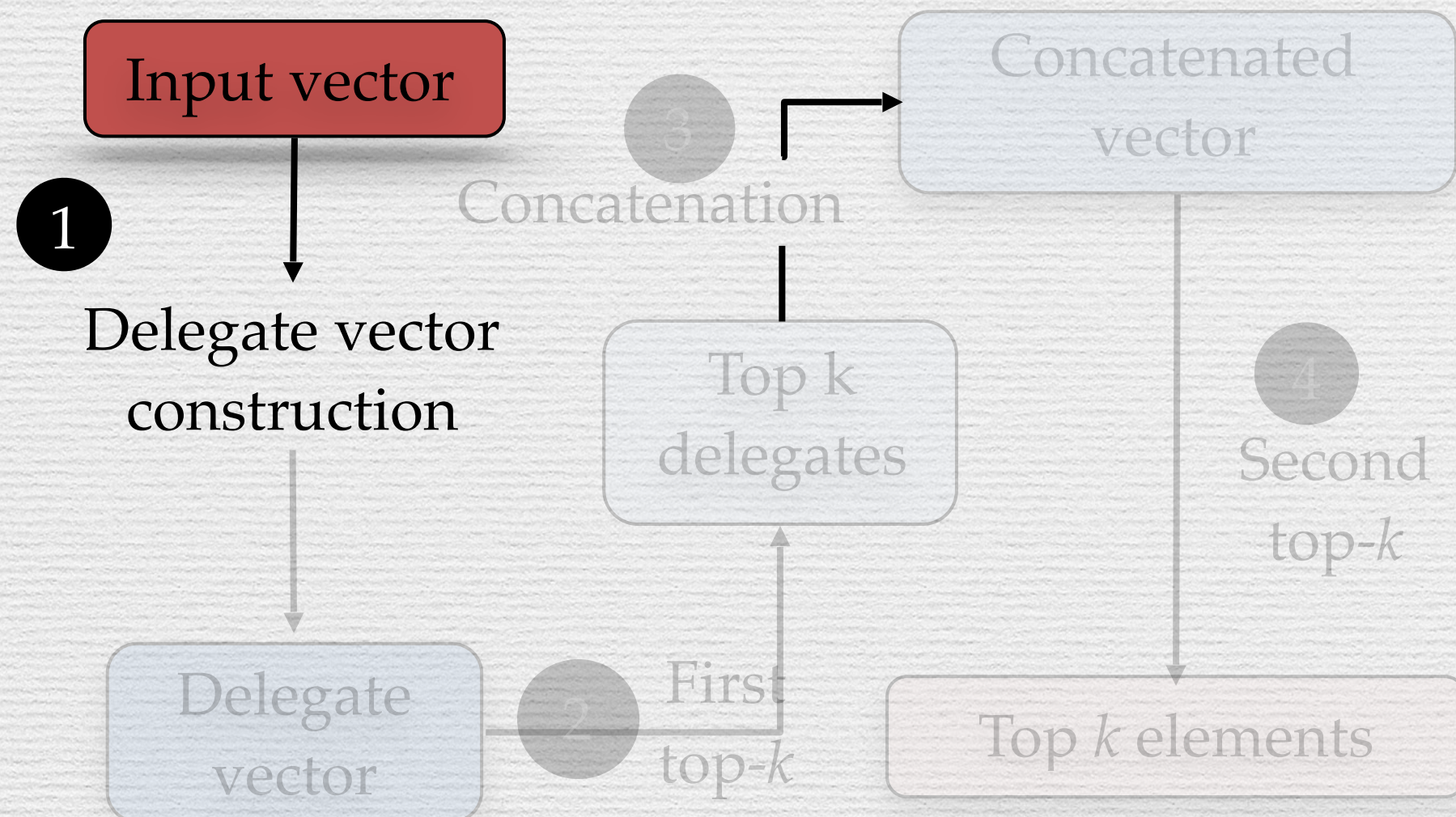
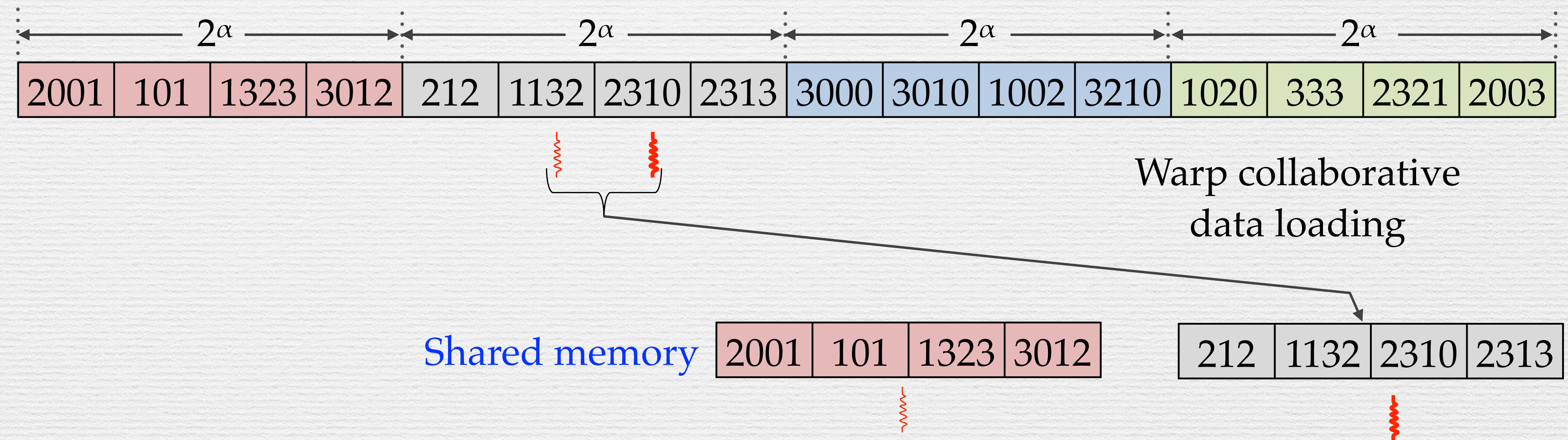


Changes for optimization

Dr.Top- k : Delegate Vec. Construction Optimization

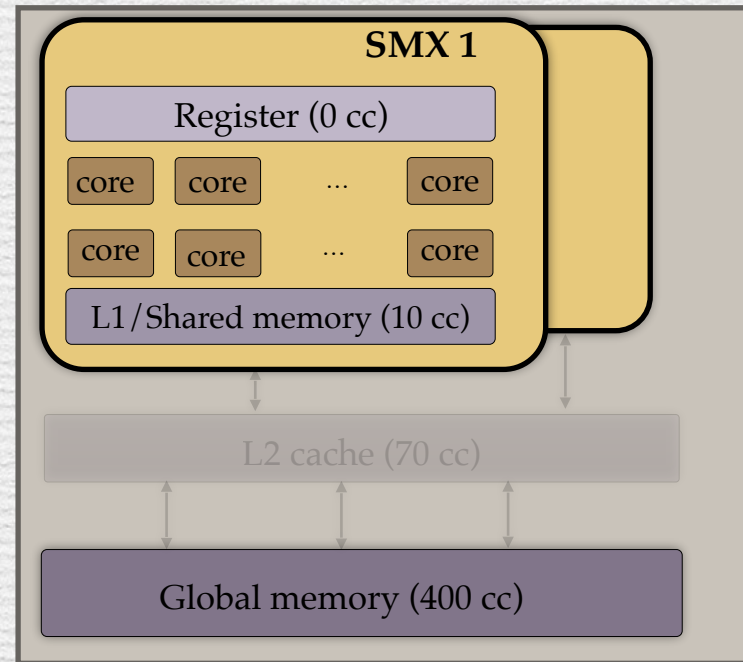


Warp

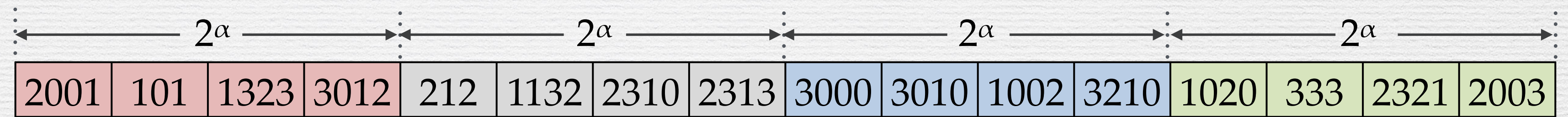


Changes for optimization

Dr.Top- k : Delegate Vec. Construction Optimization

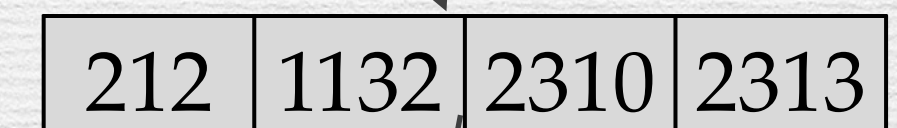
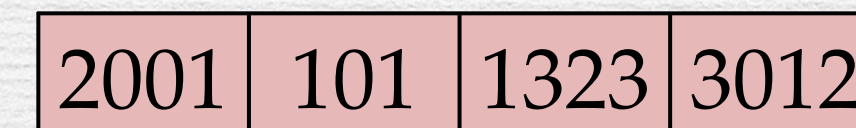


Warp



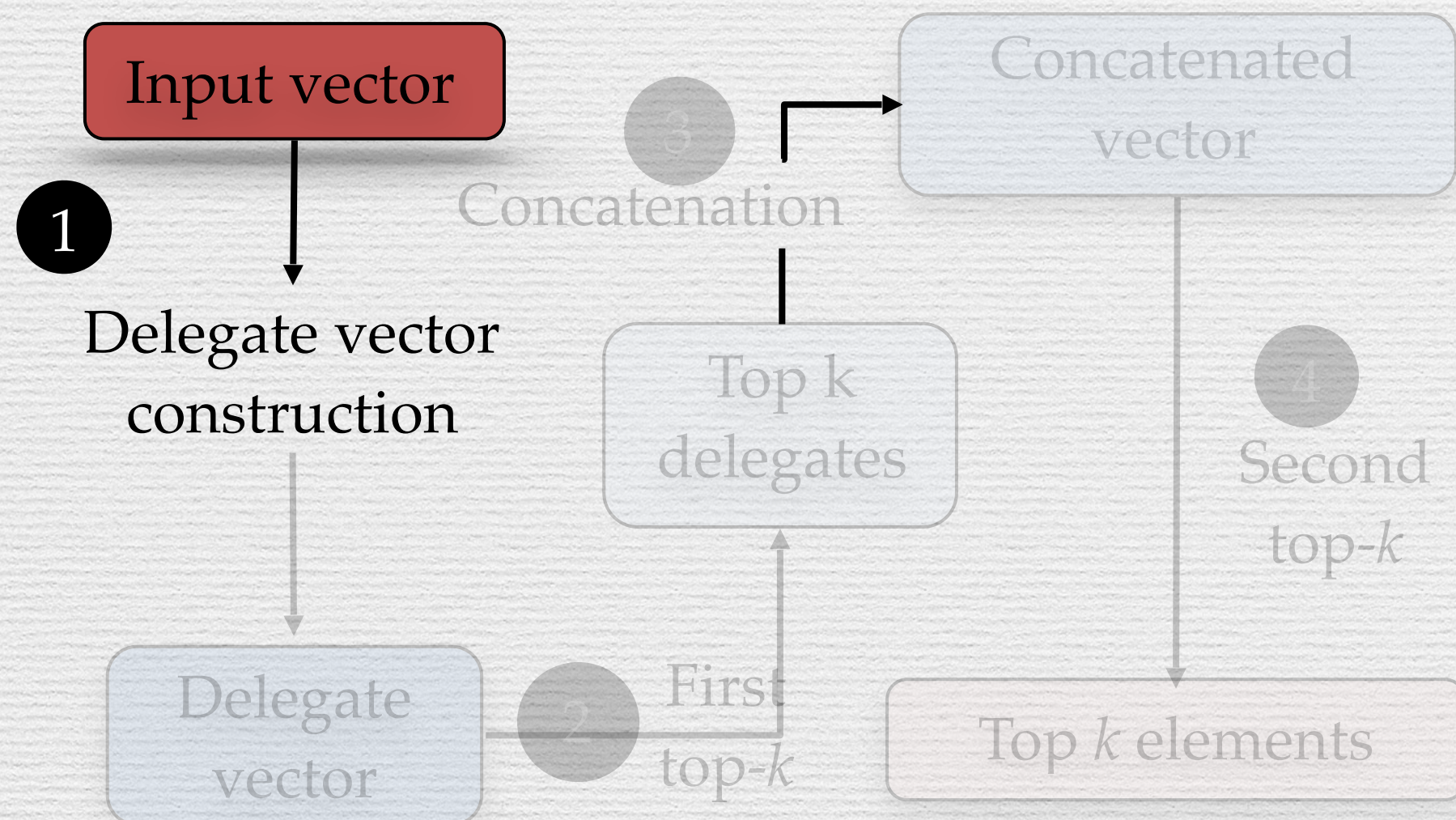
Warp collaborative data loading

Shared memory



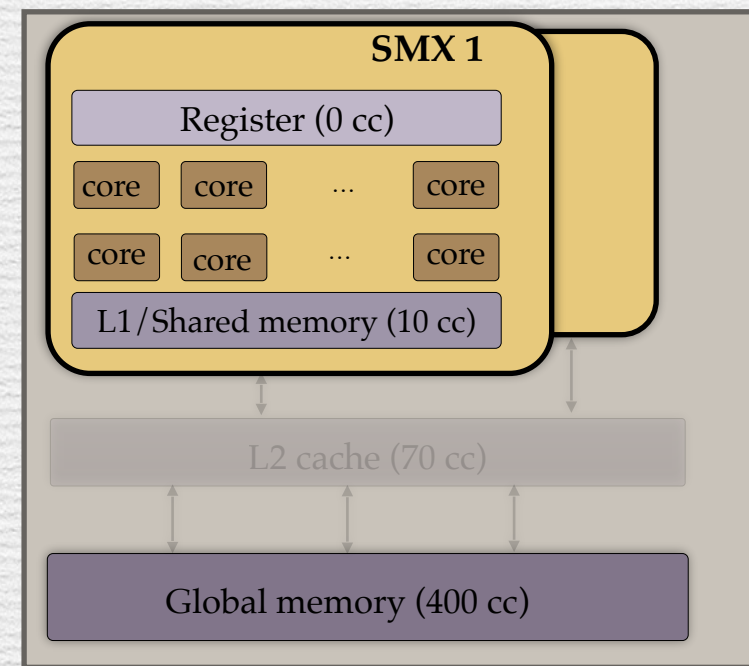
Thread based delegate vector construction.

Delegate vector

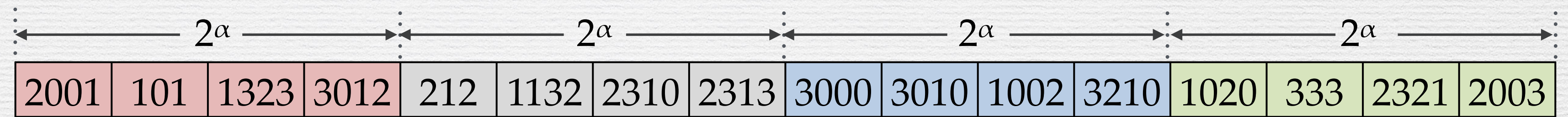


Changes for optimization

Dr.Top- k : Delegate Vec. Construction Optimization

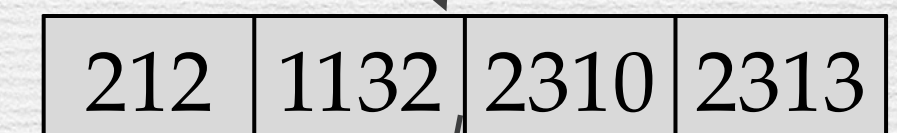
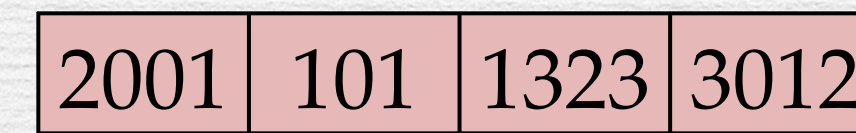


Warp



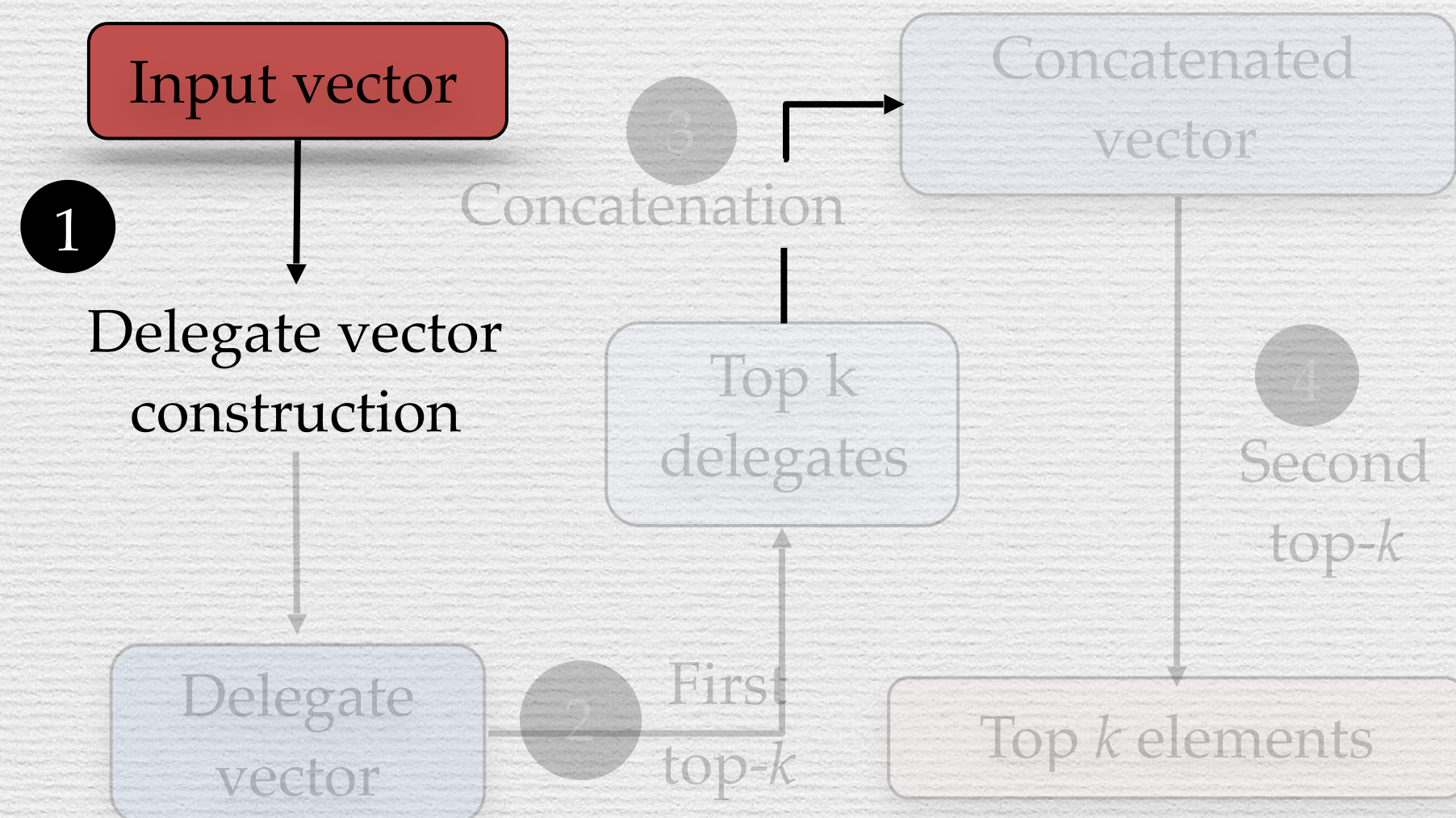
Warp collaborative data loading

Shared memory



Thread based delegate vector construction.

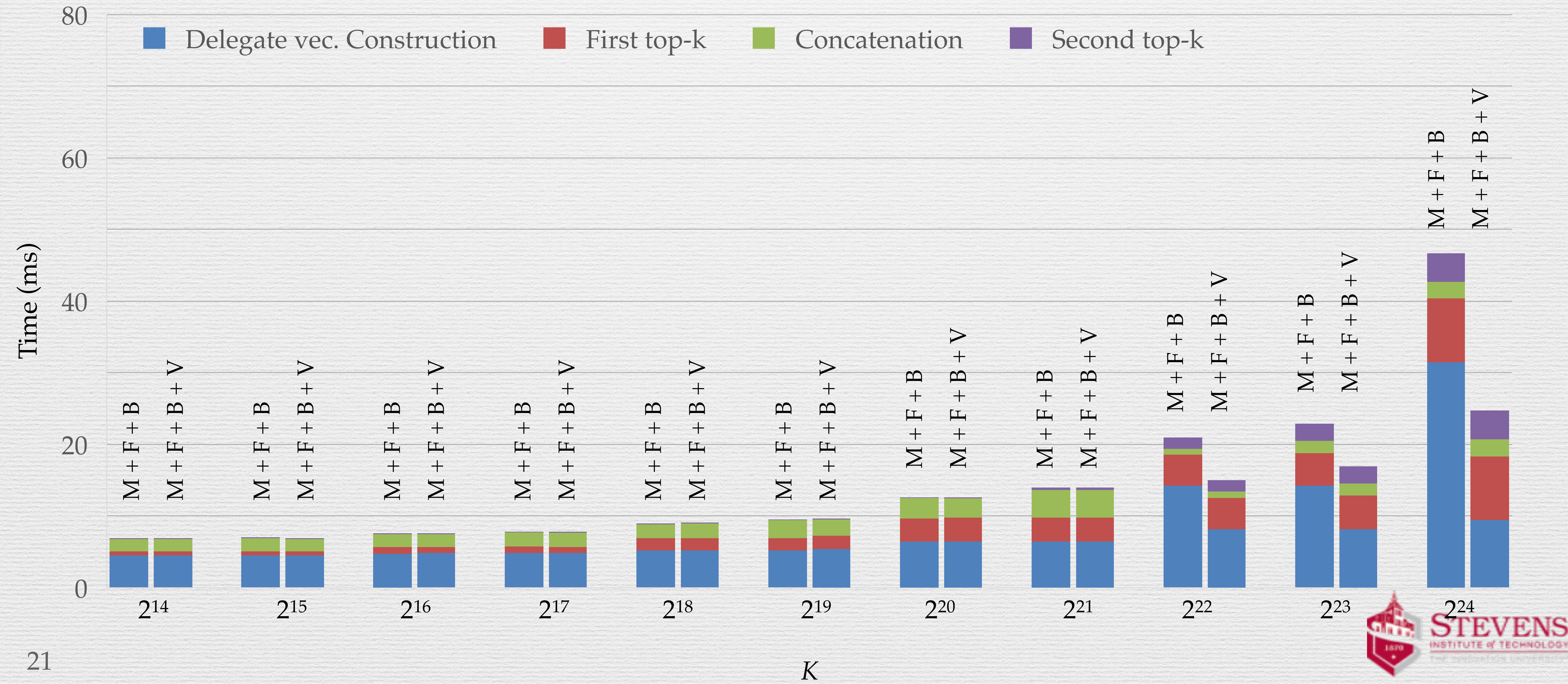
Delegate vector



Observed more than 2x speedup in delegate vector construction!

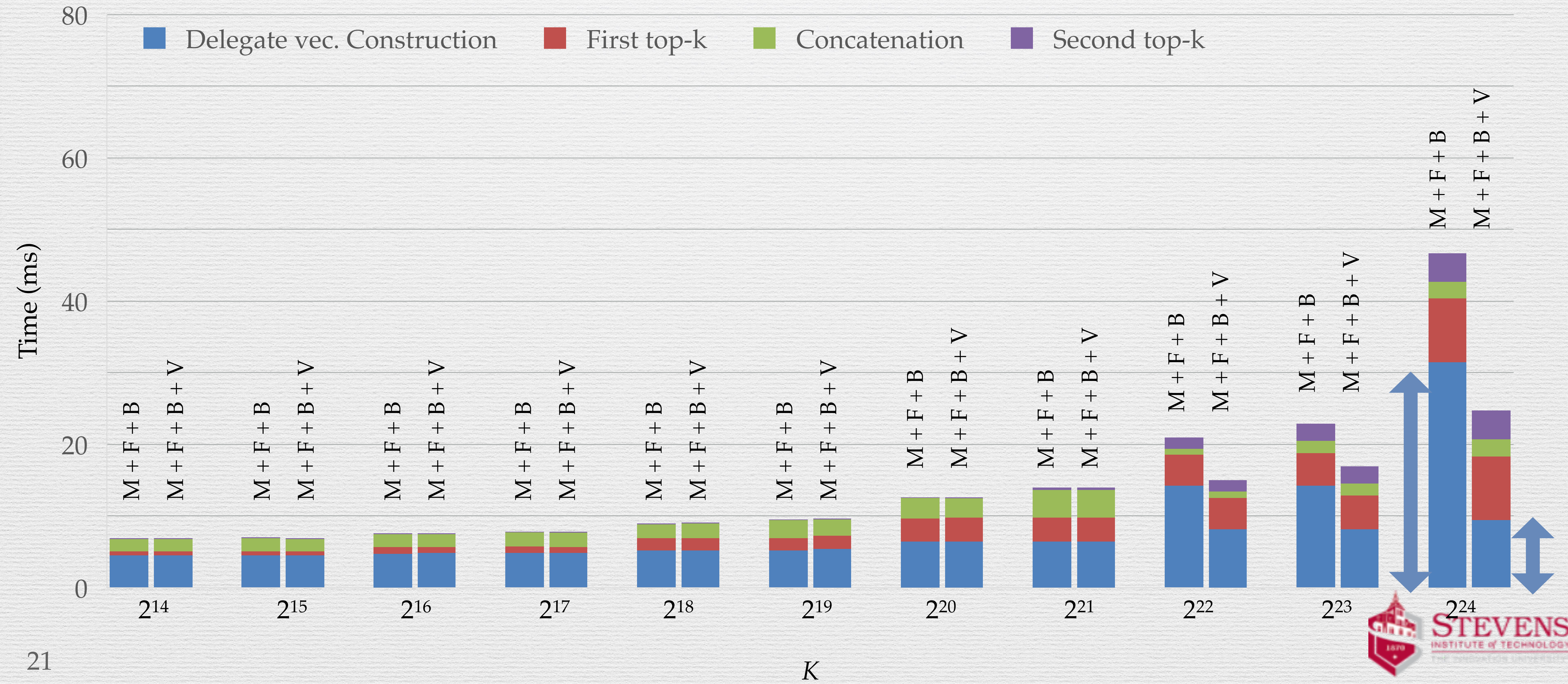
Dr.Top- k : Delegate Vec. Construction Optimization

Max delegate (M) + Filter (F) + Beta (B) + vec. opt. (V): Time breakdown of Dr.Top- k

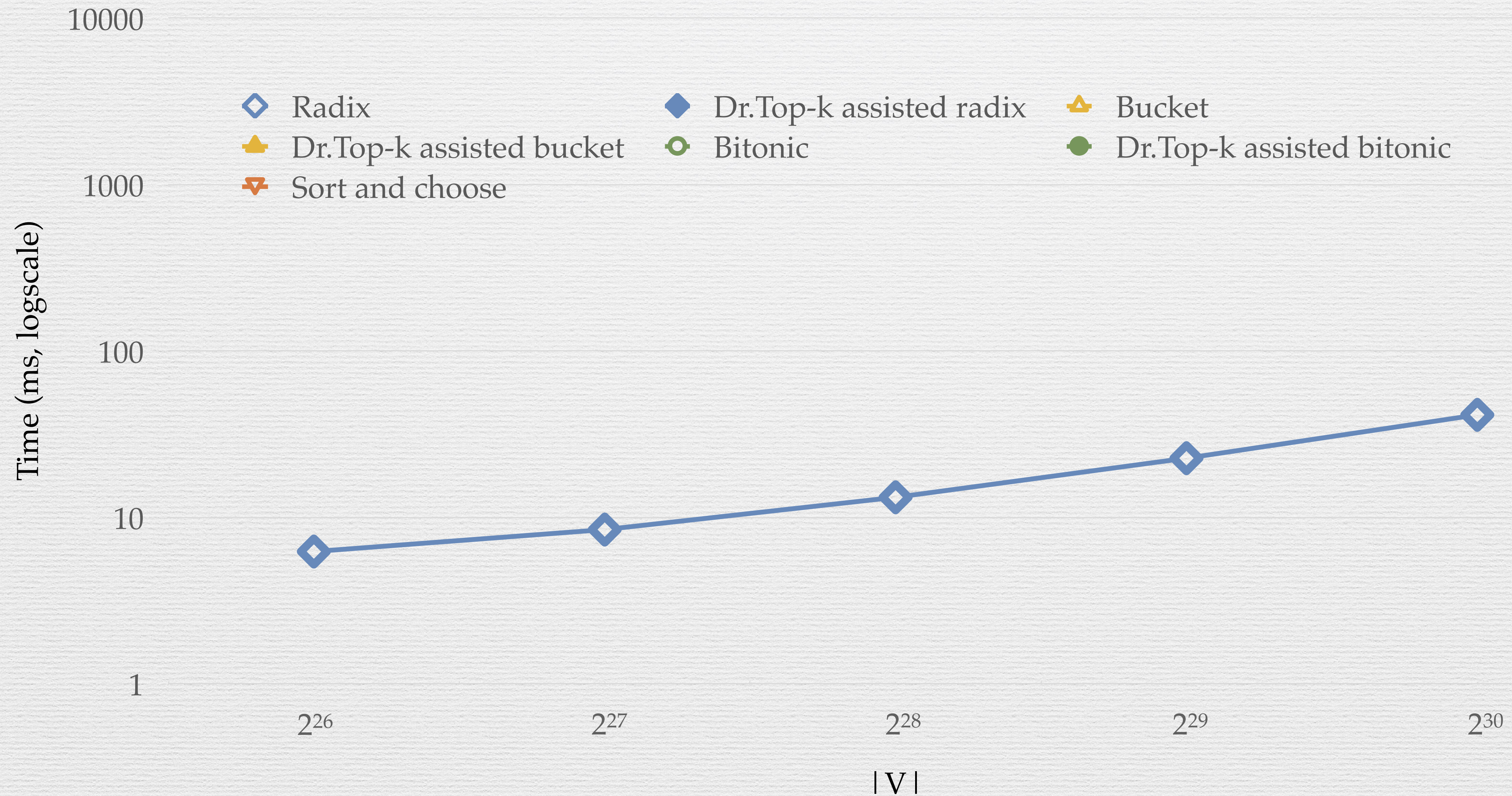


Dr.Top- k : Delegate Vec. Construction Optimization

Max delegate (M) + Filter (F) + Beta (B) + vec. opt. (V): Time breakdown of Dr.Top- k

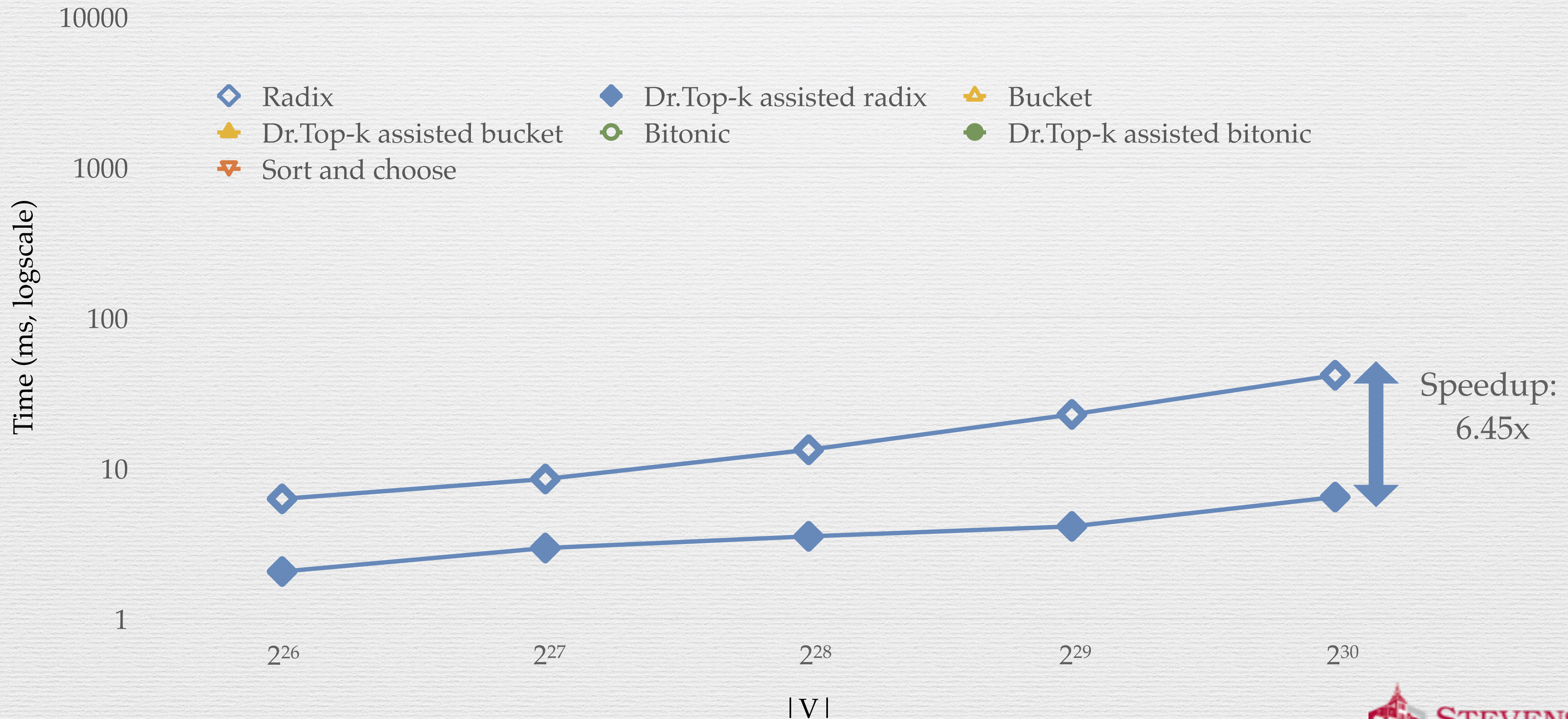


Speedup: Dr.Top- k Vs. State-of-the-art



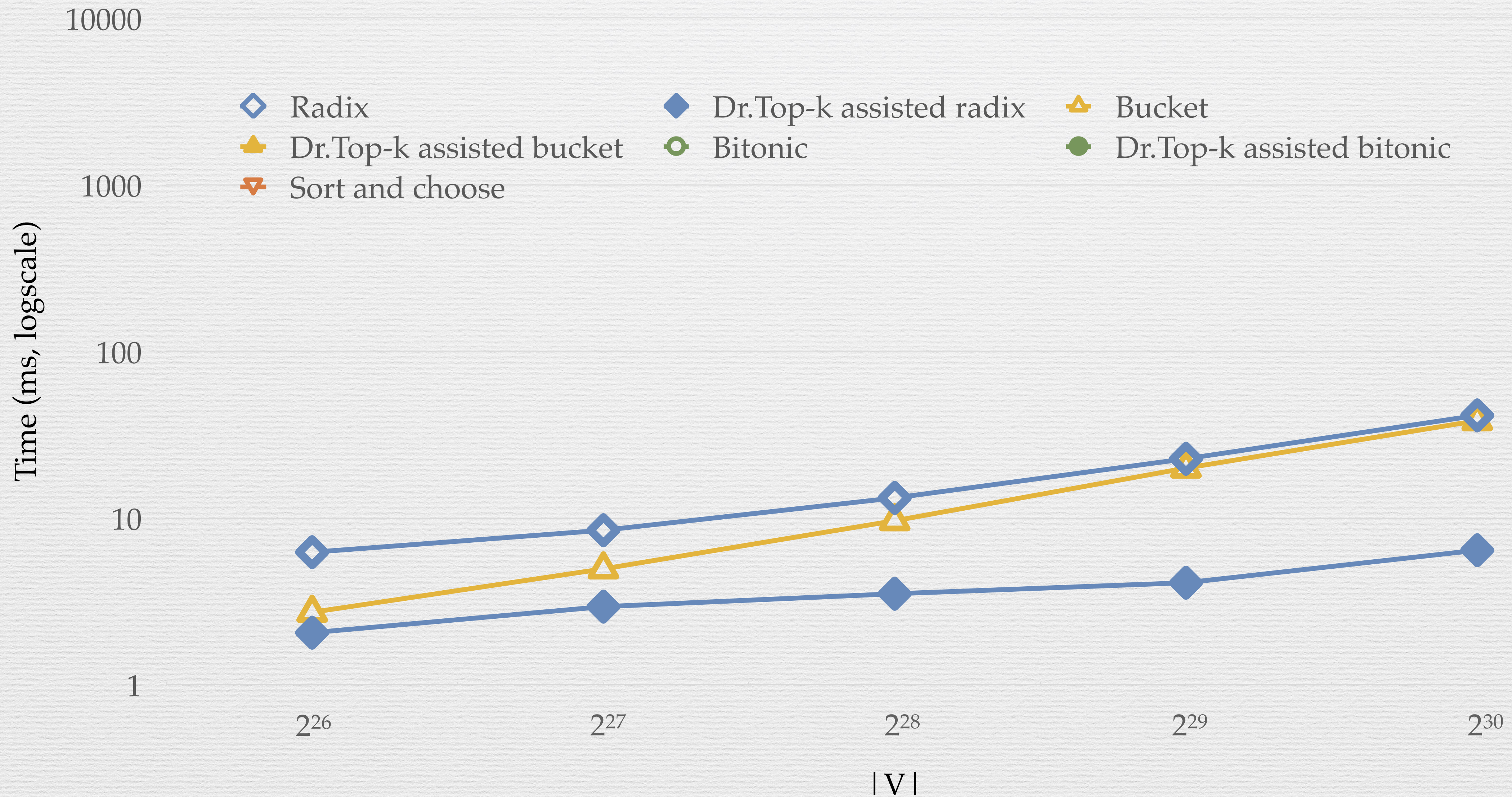
Performance with varying $|V|$ and $k = 2^{10}$

Speedup: Dr.Top- k Vs. State-of-the-art



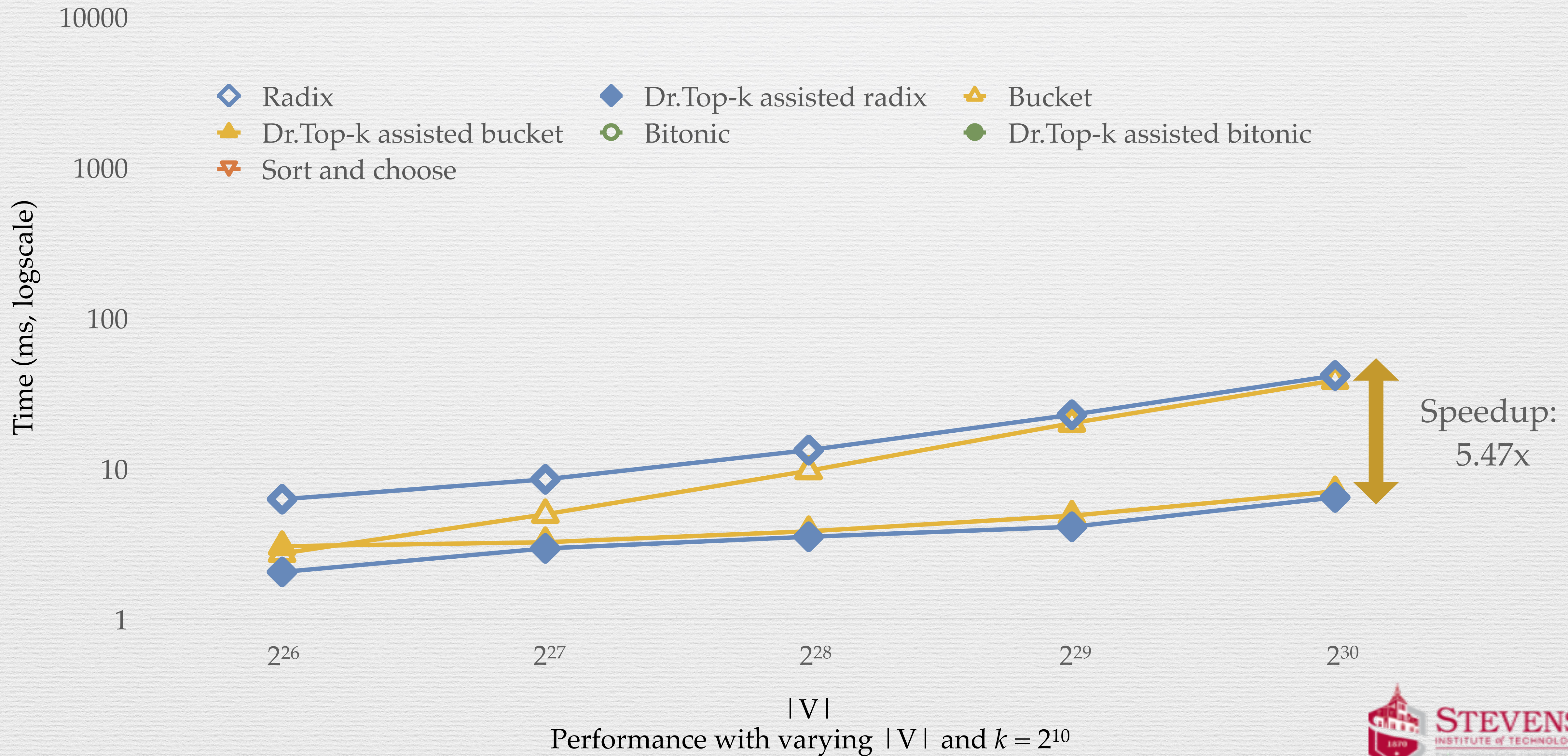
Performance with varying $|V|$ and $k = 2^{10}$

Speedup: Dr.Top- k Vs. State-of-the-art

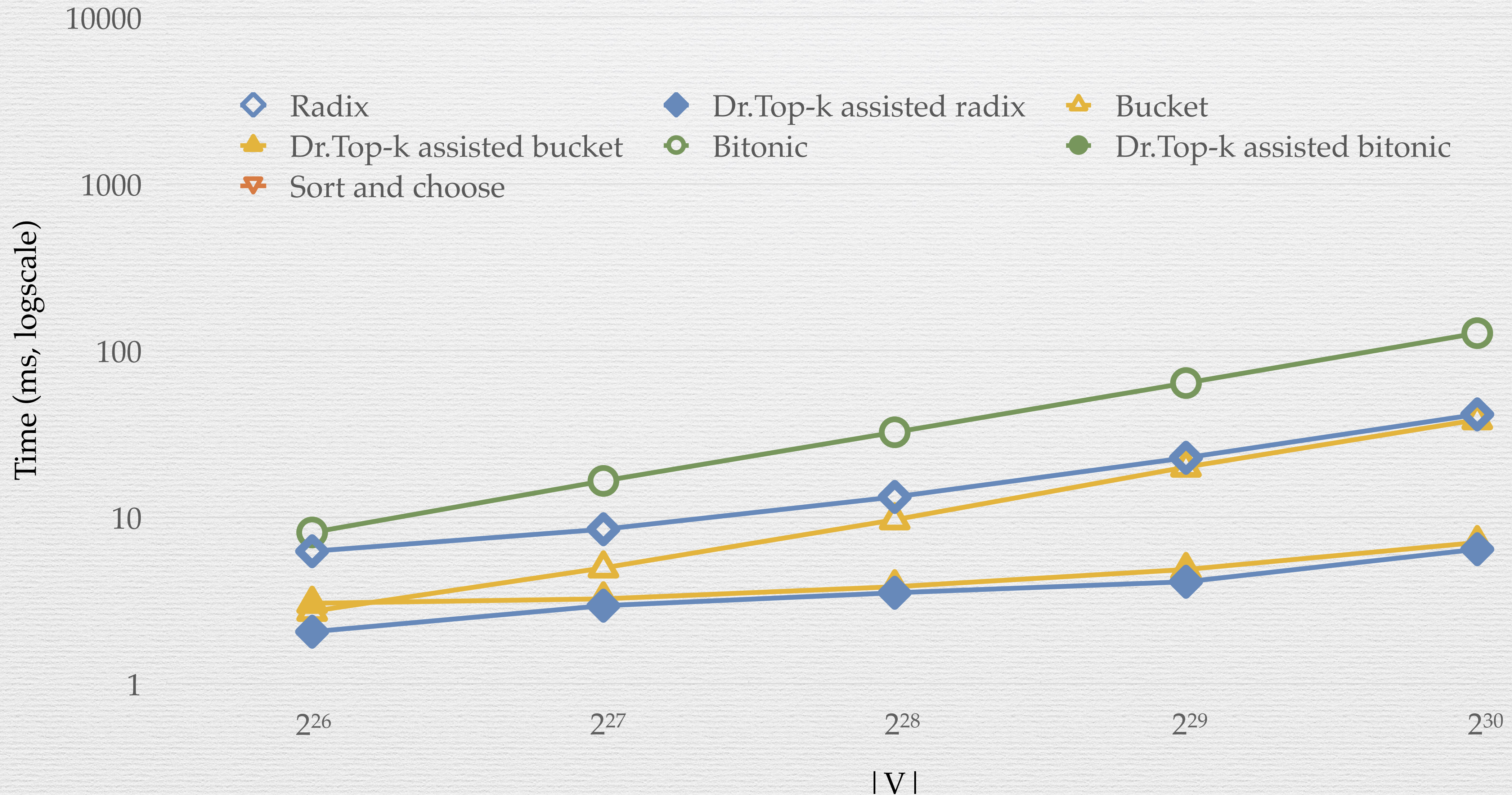


Performance with varying $|V|$ and $k = 2^{10}$

Speedup: Dr.Top- k Vs. State-of-the-art

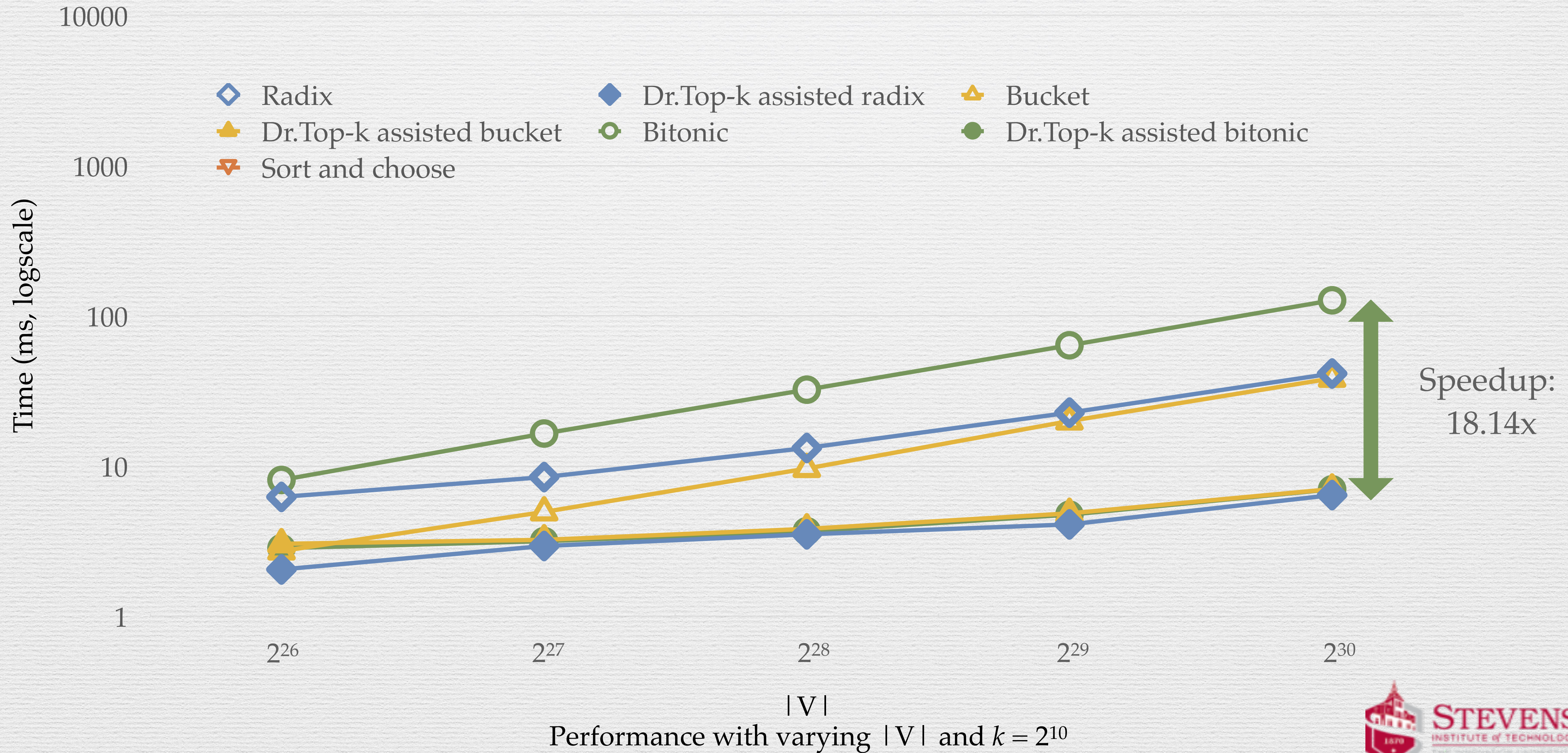


Speedup: Dr.Top- k Vs. State-of-the-art

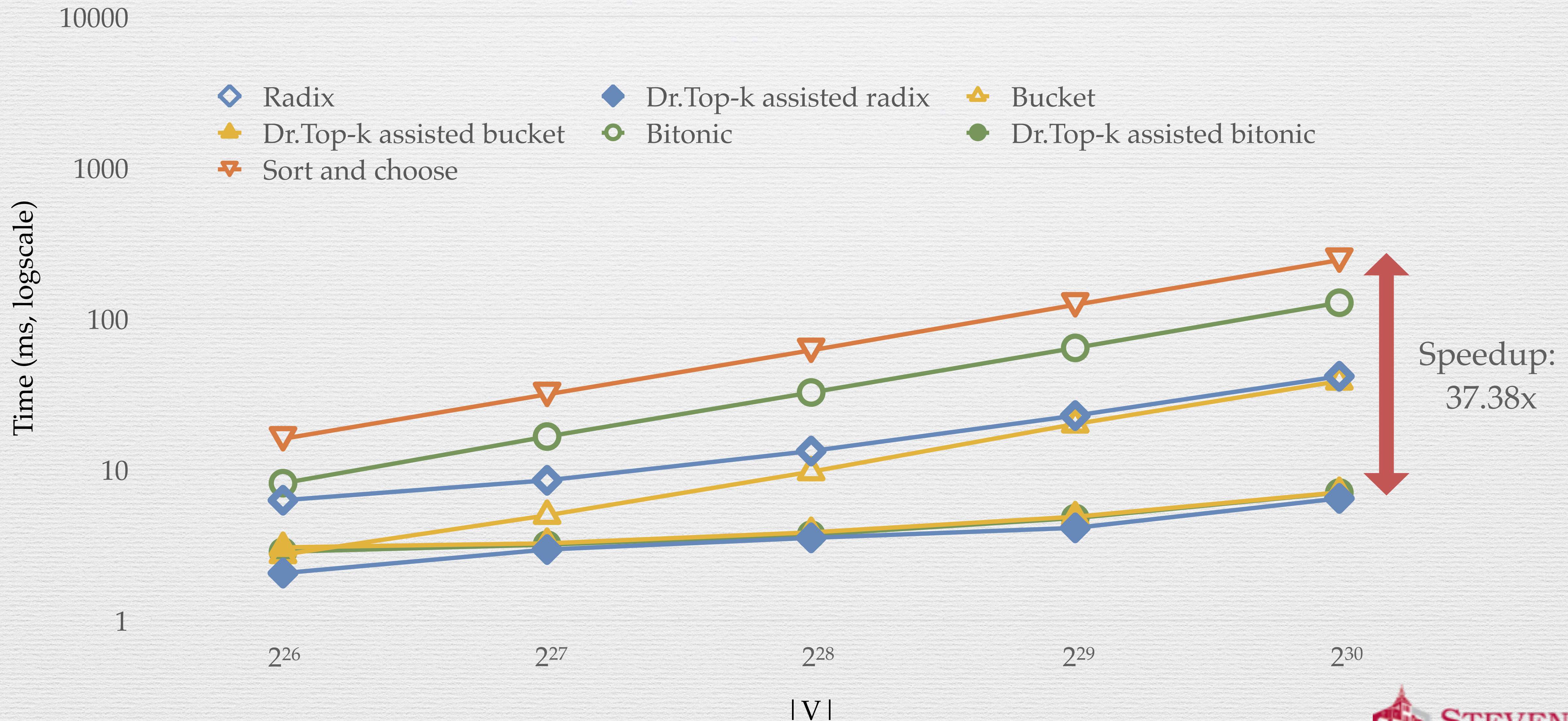


Performance with varying $|V|$ and $k = 2^{10}$

Speedup: Dr.Top- k Vs. State-of-the-art



Speedup: Dr.Top- k Vs. State-of-the-art



Conclusion

- ◆ Dr.Top-k introduces a delegate-centric concept to help tremendously reduce the workload for top-k computations.
- ◆ We deploy our project on distributed GPUs to handle extreme large input vectors along with various system optimizations.
- ◆ Dr.Top-k consistently outperforms the state-of-the-art methods for a wide range of k.

OpenSource: <https://github.com/Anil-Gaihre/DrTopKSC>

Acknowledgements



U.S. DEPARTMENT OF
ENERGY

Office of Science



Australian Government

Australian Research Council



National Nuclear Security Administration

Thank You!
Any Questions?

agaihre@stevens.edu
hliu77@stevens.edu

OpenSource: <https://github.com/Anil-Gaihre/DrTopKSC>