

# SCUD: Scalable Counting of Unique Data



Navendu Jain, Dmitry Kit, Prince Mahajan, Praveen Yalagandula, Mike Dahlin, and Yin Zhang  
 Laboratory for Advanced Systems Research, The University of Texas at Austin



## Problem Description

Providing support for complicated queries for distributed data-intensive applications

Example Applications:

- Top-k users in a storage system grouped by activity type
  - Put – store information
  - Get – retrieve information

Estimated Network Size		IS nodes
Top 10 Gets by client:		
Client IP	Number of Gets	
127.0.0.1	1152	
128.83.144.30	243	
128.83.120.172	168	
128.83.120.245	147	
128.83.120.138	120	
128.83.120.21	90	
128.83.144.43	75	
128.83.130.11	48	
128.83.120.114	27	
128.83.144.241	12	
Top 10 Puts by client:		
Client IP	Number of Puts	
127.0.0.1	1100	
128.83.120.138	150	
128.83.144.30	144	
128.83.144.241	120	

Other Applications

Network Monitoring:

- Top-k flows by bytes and packets
- Aggregate MAX/MIN/AVG. incoming flows in an organization

CDN:

- Counting the # of unique clients in the system

Telematic Applications

- Identifying roads which have traffic heavier than some threshold

## Challenges

### 1. Scalability

- High volume data streams
- Highly distributed data

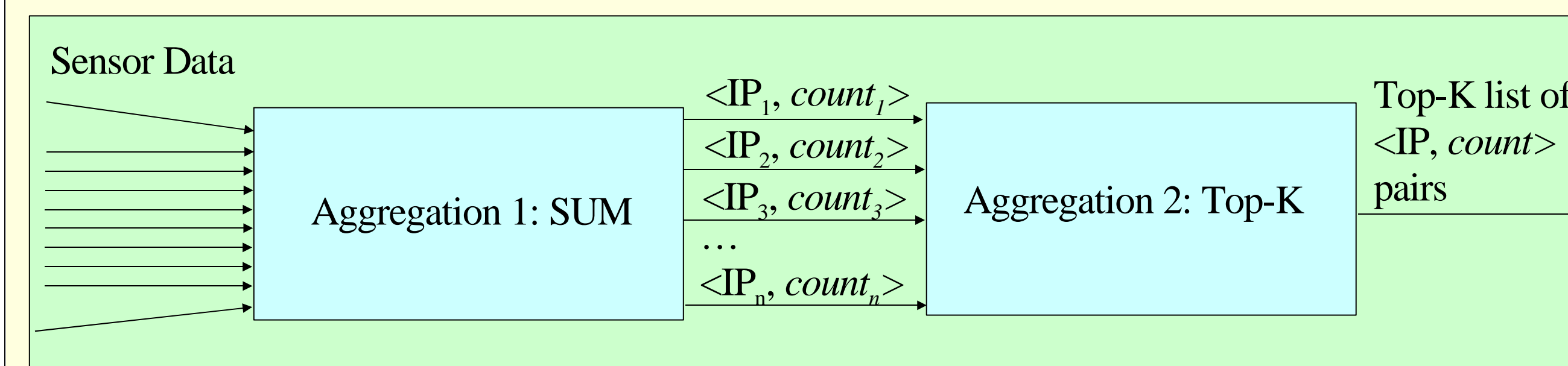
### 2. Centralized

1. High B/W cost
2. High Processing cost
3. High Latency

## Approach

Push query processing into the network

Chaining of Aggregation functions



1. Use one aggregation function to get basic information regarding each data item

- # of Client accesses
- Bandwidth consumption by a client
- ...

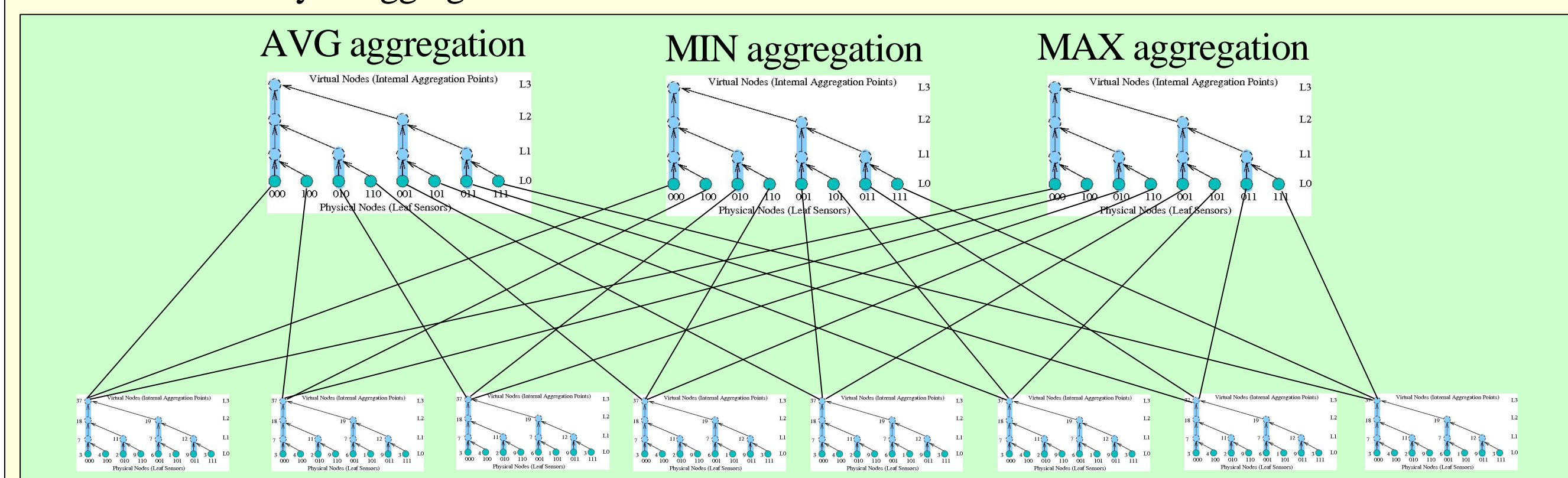
2. Feed the results of first aggregation into the second aggregation to obtain higher level information

1. Top-k
2. Max/Min/Avg
3. ...

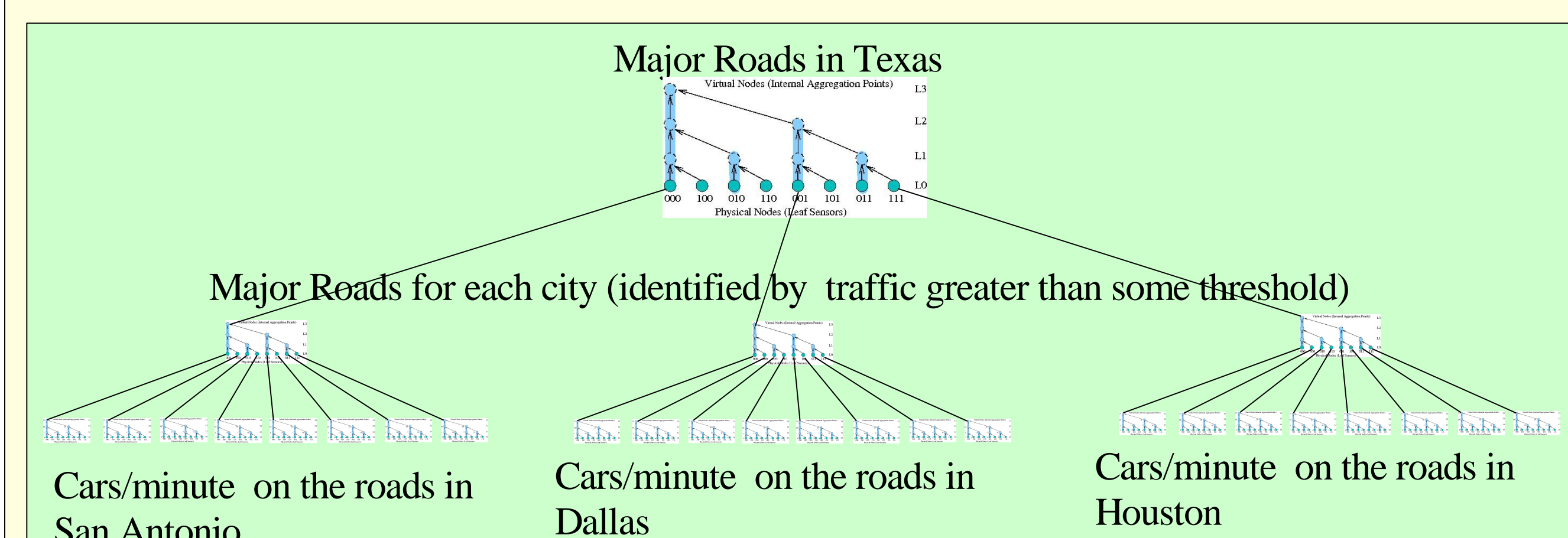
## Our Contribution

### 1. Scalability

- Each aggregation is built on a scalable system called SDIMS
- Reusability of aggregated data



- Powerful way to build hierarchical aggregations



## Demo

Web Demo

- Perform two operations on the system put/get
- Can Specify the address from which requests originate
- View the top-10 list of clients who performed a put or a get.

Details:

- Uses Bamboo DHT as the storage system
- When bamboo receives a put or a get request it notifies SDIMS with the new access count (old\_count+1)
- When this information is fully aggregated SDIMS contacts bamboo
- Bamboo updates its local top-10 list
- If there is a change this list is inserted into SDIMS again under a different aggregation
- The lists from each root is combined to form the final top-10 list

Location:

- [http://z.cs.utexas.edu/users/dkit/bamboo\\_test.php](http://z.cs.utexas.edu/users/dkit/bamboo_test.php)

## Further Information

URL: <http://www.cs.utexas.edu/users/ypraveen/sdims>

Email: [ypraveen@cs.utexas.edu](mailto:ypraveen@cs.utexas.edu)