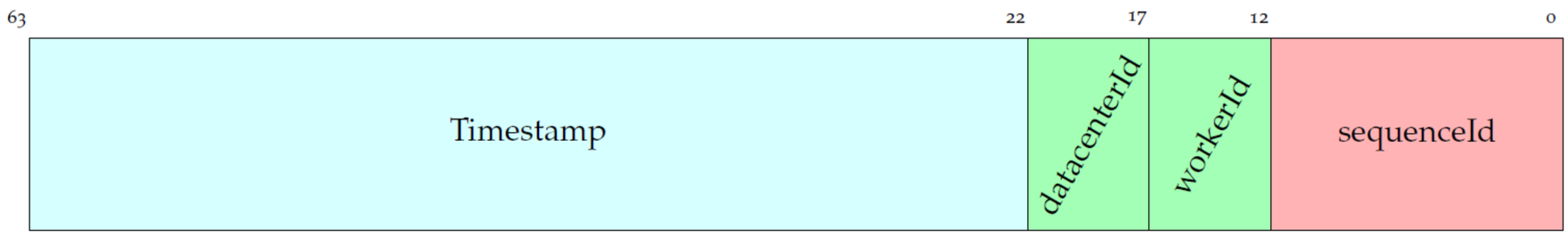
Collecting More Tweets From Twitter API

A case study of Twitter's Snowflake algorithm for generating unique IDs

- **Snowflake** announced in June 2010:
- *Generate ids in the scale of 10,000 per second*
- *Generate ids in a distributed and uncoordinated approach*
- *Generate ids so they are roughly sortable by time*



- **Twitter API statues/sample**:
- 1% of all tweets on the timeline

- **Analyzed 10 Million tweet IDs**
- On average 3320 tweets per minute

- **Twitter API statuses/lookup**
- Check 120,000 tweet ids every 15 minutes

- **Can we collect all the tweets on Twitter?**
- For 90% sample 2,500 API keys

- ***Let's Share our Keys!***

sequence_id	Percentage	Cumulative Percentage
0	**59.574%**	59.574%
1	24.163%	83.737%
2	7.861%	91.598%
5	2.551%	94.149%
3	2.157%	96.306%
6	1.646%	97.952%
7	0.695%	98.647%
4	0.649%	**99.297%**

	DatacenterId + workerId	Percentage	Cumulative Percentage
1	363	4.6%	4.6%
2	365	4.6%	9.2%
3	375	4.6%	13.8%
4	336	4.6%	18.4%
…	…	…	…
19	35	0.5%	89.0%
20	36	0.5%	89.5%
21	33	0.5%	**90.0%**

Internet Technologies and Systems
Seyed Ali Alhosseini, Christoph Meinel

E-Mail: seyedali.alhosseni@hpi.de
Tel.: 033155094836
Web: https://www.hpi.uni-potsdam.de
Prof.-Dr.-Helmert-Str. 2-3 14482 Potsdam

HPI Hasso Plattner Institut