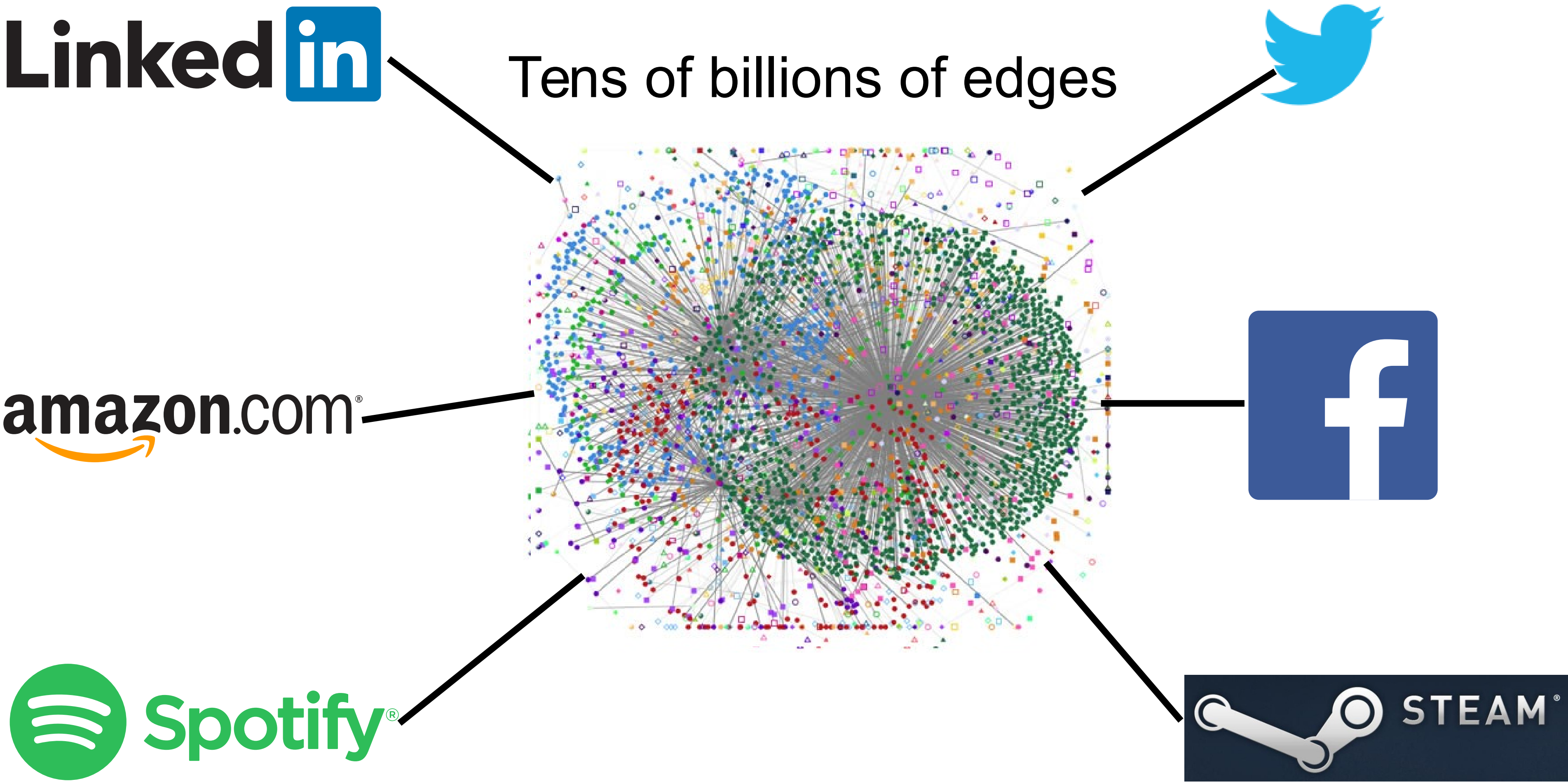
Graphalytics:
A Big Data Benchmark for Graph-Processing Platforms

Mihai Capotă, Tim Hegeman, Alexandru Iosup,
Arnau Prat-Pérez, Orri Erling, Peter Boncz

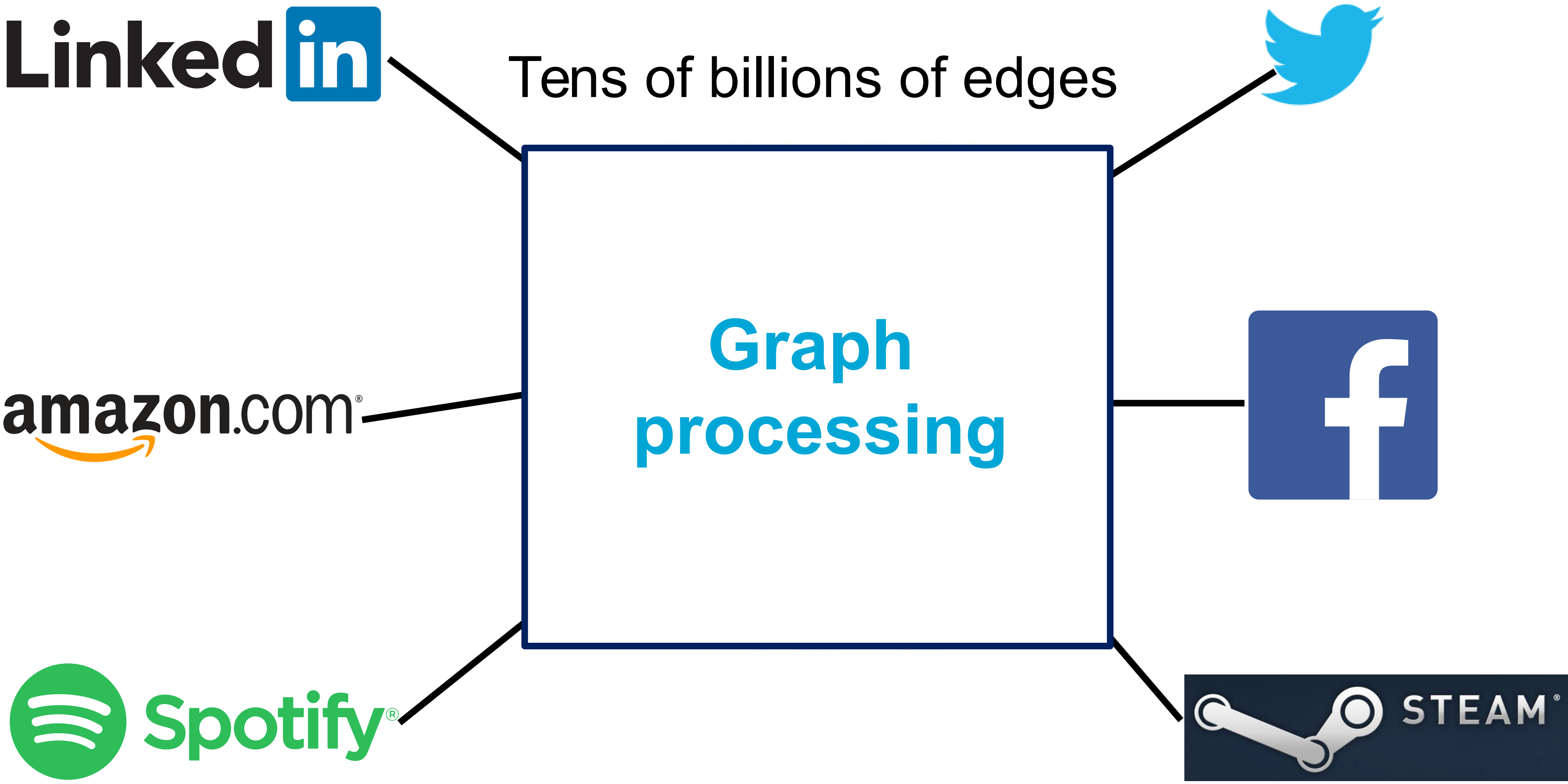
Delft University of Technology
Universitat Politècnica de Catalunya
OpenLink Software
CWI

The data deluge: large-scale graphs



Tens of billions of edges

The data deluge: large-scale graphs



Tens of billions of edges

Graph processing

Platform diversity

hadoop

Oracle Labs
PGX

Neo4j

APACHE
GIRAPH

GraphLab

Spark
GraphX

Yet another benchmark?

Yet another benchmark?

- Lack of benchmarks for generic graph processing platforms

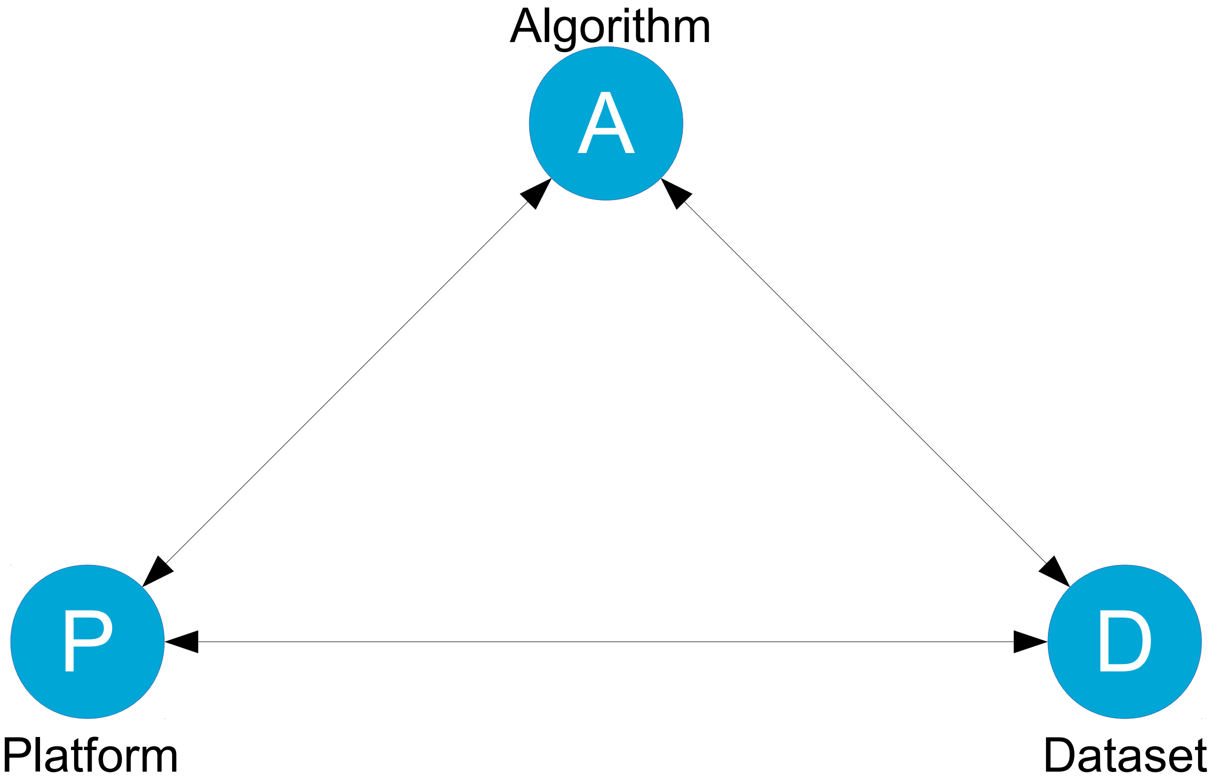
Yet another benchmark?

- Lack of benchmarks for generic graph processing platforms
- Graph500
  - BFS
  - Kroneker graph

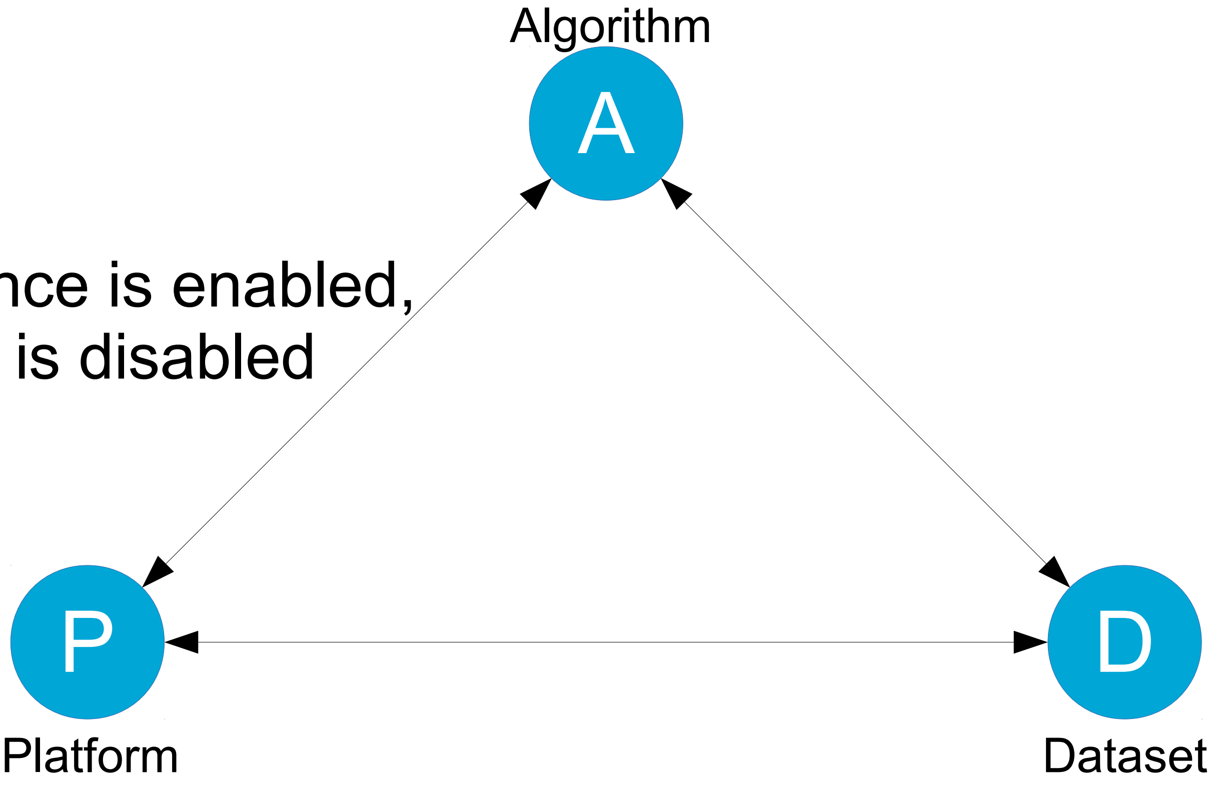
Yet another benchmark?

- Lack of benchmarks for generic graph processing platforms
- Graph500
  - BFS
  - Kroneker graph
- Several academic studies
  - Specific to graph or RDF databases
  - Ad hoc setup, difficult to extend

P-A-D triangle

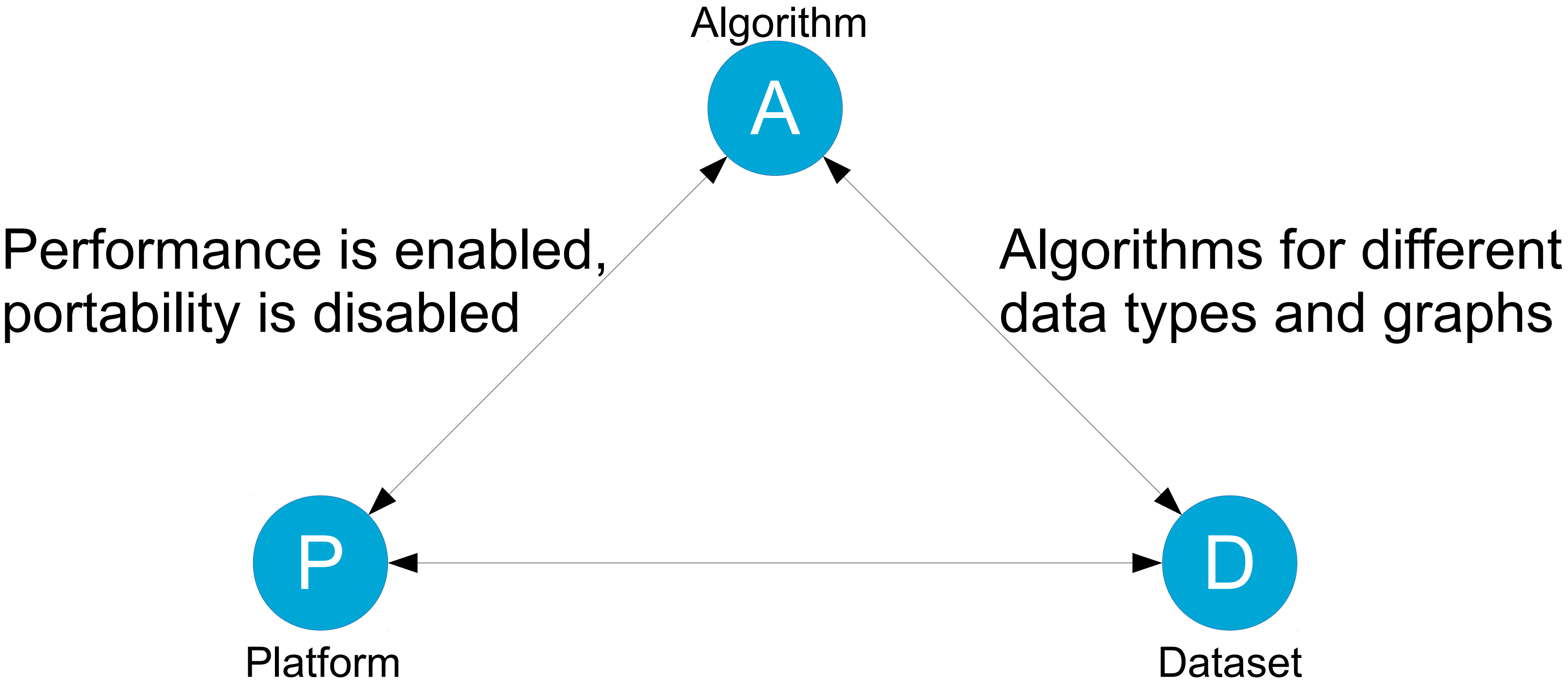


P-A-D triangle



Performance is enabled, portability is disabled

P-A-D triangle



Algorithm

A

Performance is enabled, portability is disabled

Algorithms for different data types and graphs

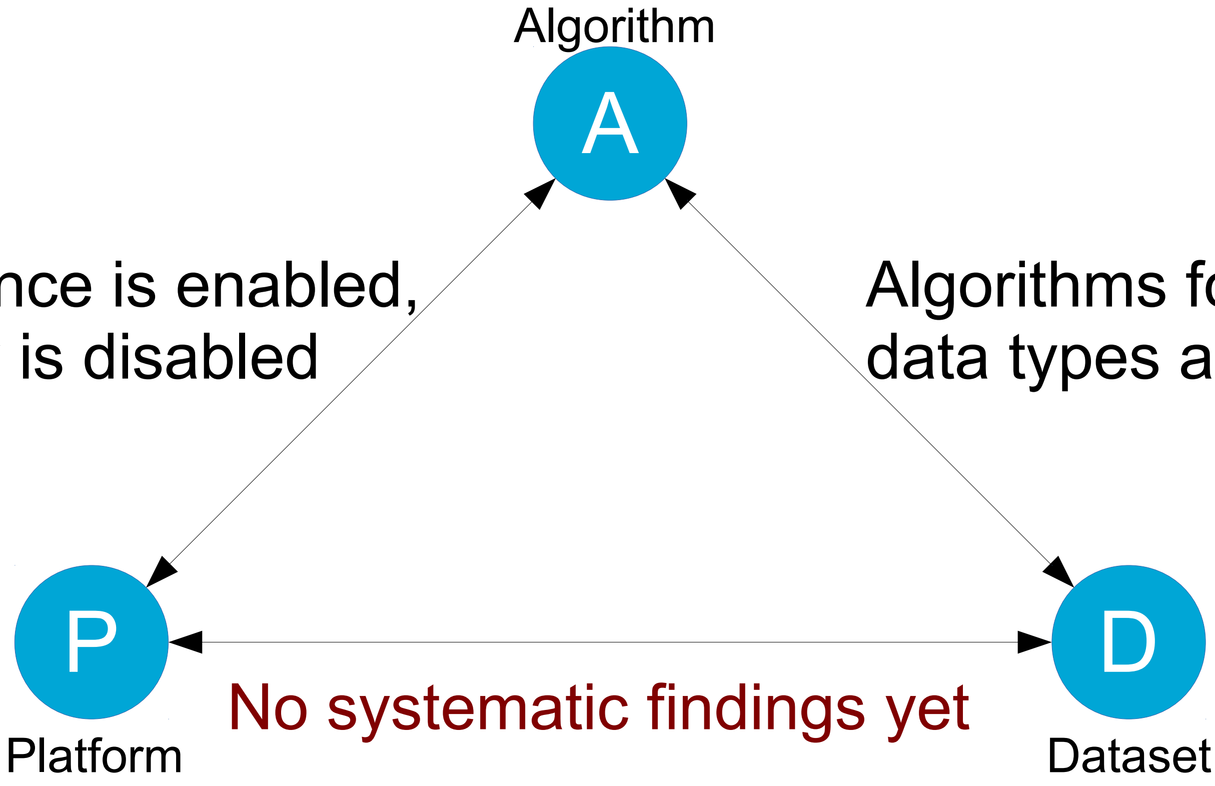
P

Platform

D

Dataset

P-A-D triangle

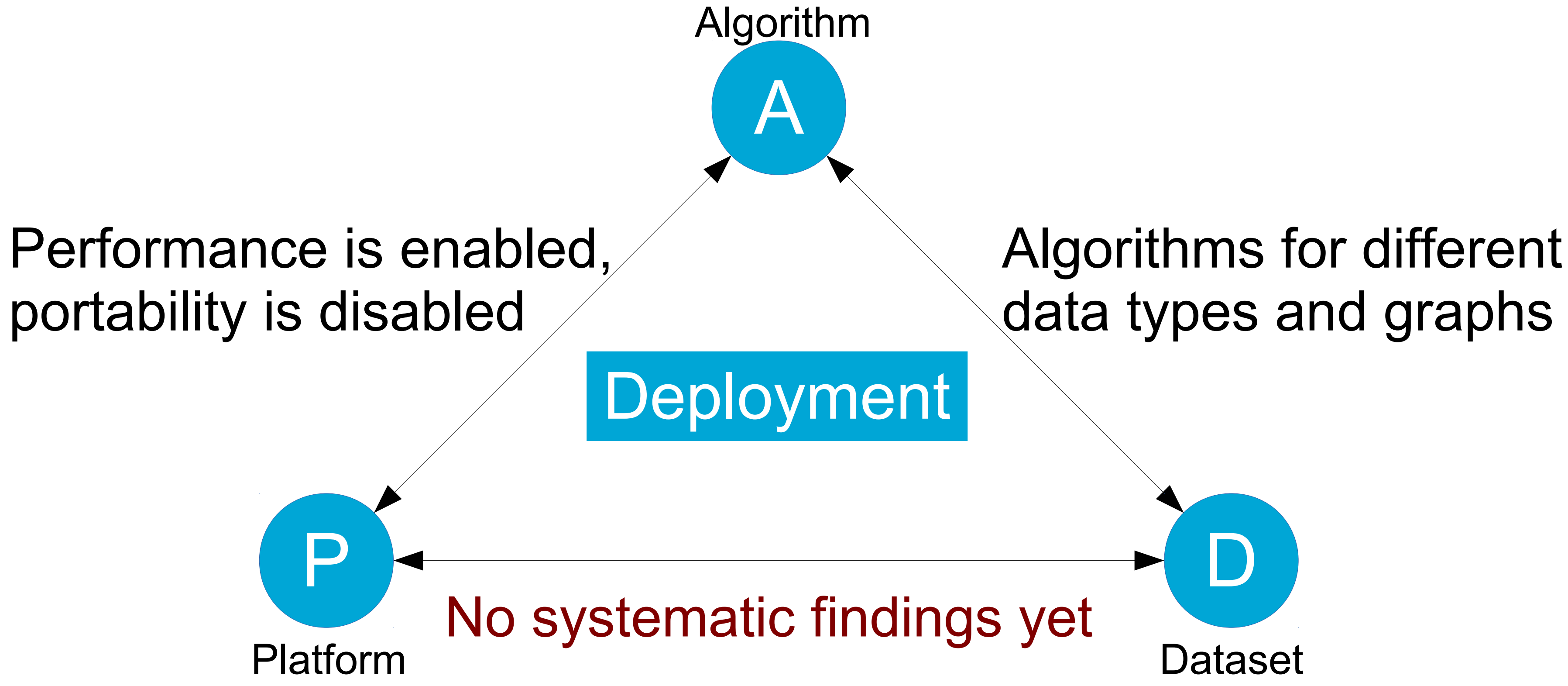


Performance is enabled, portability is disabled

Algorithms for different data types and graphs

No systematic findings yet

P-A-D triangle



Graphalytics

Graphalytics

- The first comprehensive benchmark for big data graph-processing platforms

Graphalytics

- The first comprehensive benchmark for big data graph-processing platforms
  - Advanced benchmark harness

Graphalytics

- The first comprehensive benchmark for big data graph-processing platforms
  - Advanced benchmark harness
  - Choke-point analysis

Graphalytics

- The first comprehensive benchmark for big data graph-processing platforms
  - Advanced benchmark harness
  - Choke-point analysis
  - Realistic graph generator

Graphalytics

- The first comprehensive benchmark for big data graph-processing platforms
  - Advanced benchmark harness
  - Choke-point analysis
  - Realistic graph generator

LDBC The graph & RDF benchmark reference

spec Research

Graphalytics

- The first comprehensive benchmark for big data graph-processing platforms
  - Advanced benchmark harness
  - Choke-point analysis
  - Realistic graph generator

LDBC The graph & RDF benchmark reference

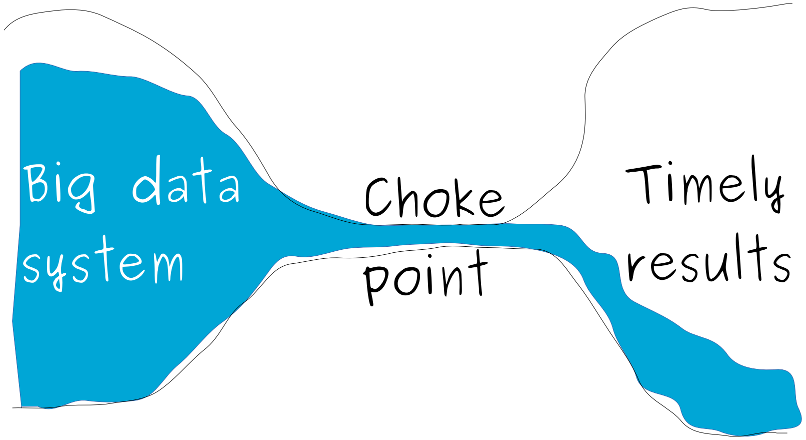
spec Research

- Co-sponsored by Oracle

Choke points

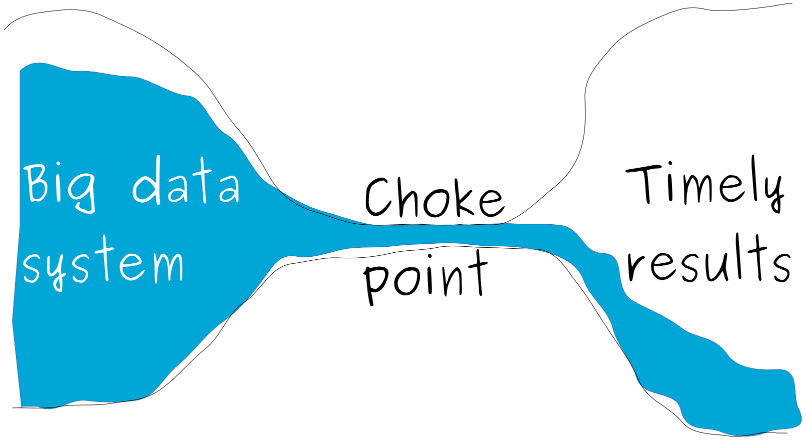
Choke points

- Technological challenges for platforms



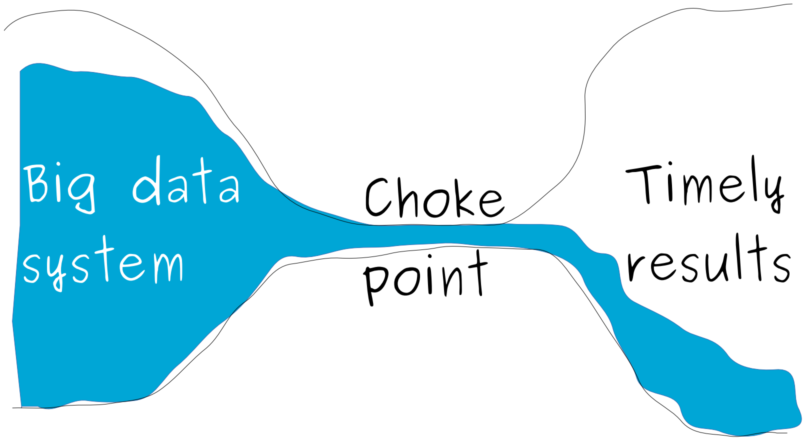
Choke points

- Technological challenges for platforms
- Identified by experts



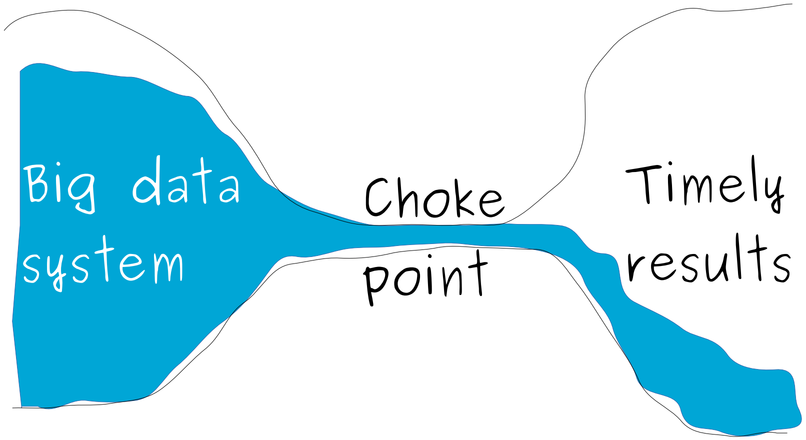
Choke points

- Technological challenges for platforms
- Identified by experts
- Real-world scenarios are enhanced to stress choke points
  - Prevent tunnel vision
  - Advance state of the art



Choke points

- Technological challenges for platforms
- Identified by experts
- Real-world scenarios are enhanced to stress choke points
  - Prevent tunnel vision
  - Advance state of the art



Erling et al., The LDBC Social Network Benchmark: Interactive Workload , SIGMOD 2015

Examples of choke points

Examples of choke points

- Network utilization

Examples of choke points

- Network utilization
- Memory footprint

Examples of choke points

- Network utilization
- Memory footprint
- Access locality

Examples of choke points

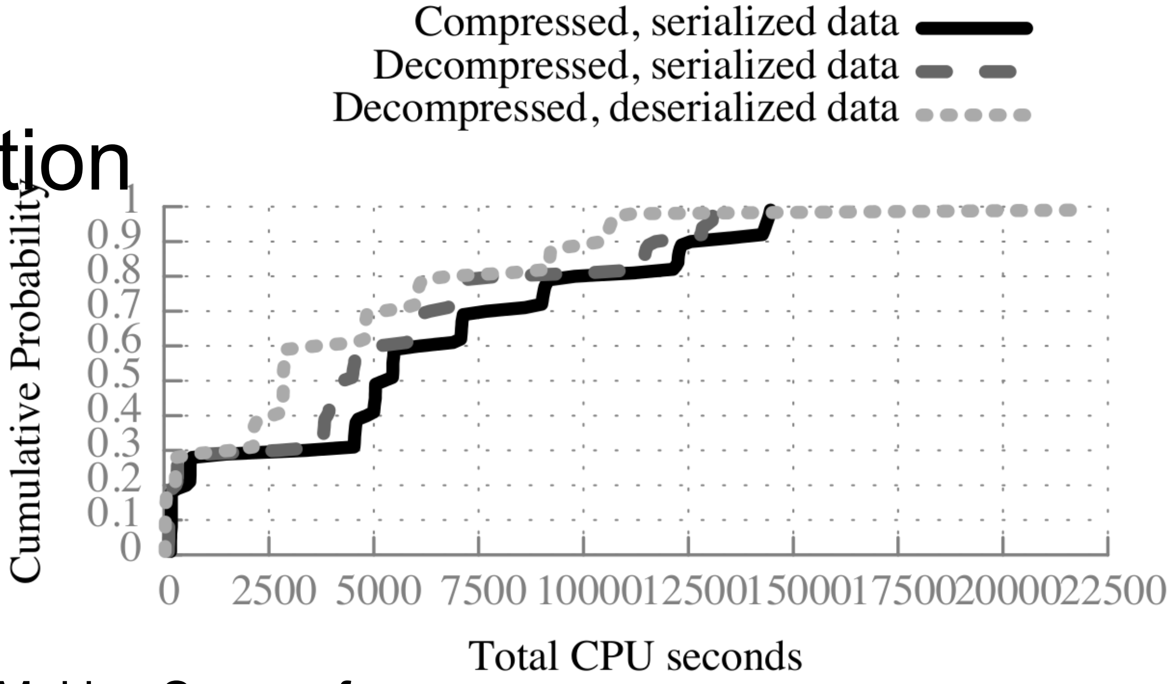
- Network utilization
- Memory footprint
- Access locality
- Skewed execution

Examples of choke points

- Network utilization
- Memory footprint
- Access locality
- Skewed execution
- CPU?

Examples of choke points

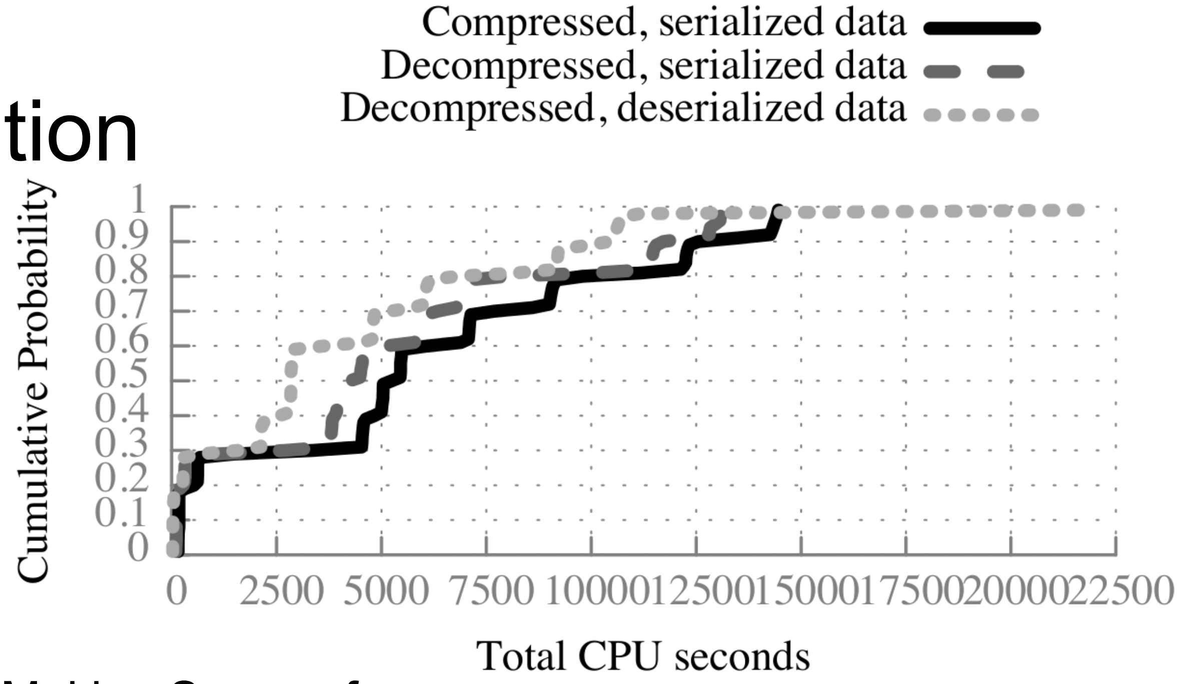
- Network utilization
- Memory footprint
- Access locality
- Skewed execution
- CPU?



Ousterhout et al., “Making Sense of Performance in Data Analytics Frameworks”, NSDI 2015

Examples of choke points

- Network utilization
- Memory footprint
- Access locality
- Skewed execution
- CPU?
- Others?



Ousterhout et al., "Making Sense of Performance in Data Analytics Frameworks", NSDI 2015

Realistic graph generator

Realistic graph generator

- LDBC Datagen
  - Synthetic social network similar to Facebook

Realistic graph generator

LDBC The graph & RDF benchmark reference

- LDBC Datagen
  - Synthetic social network similar to Facebook
- Graphalytics enhancements
  - Multiple degree distributions
    - Zeta and geometric implemented

Realistic graph generator

- LDBC Datagen
  - Synthetic social network similar to Facebook
- Graphalytics enhancements
  - Multiple degree distributions
    - Zeta and geometric implemented
  - Other structural characteristics
    - Clustering coefficient
    - Assortativity

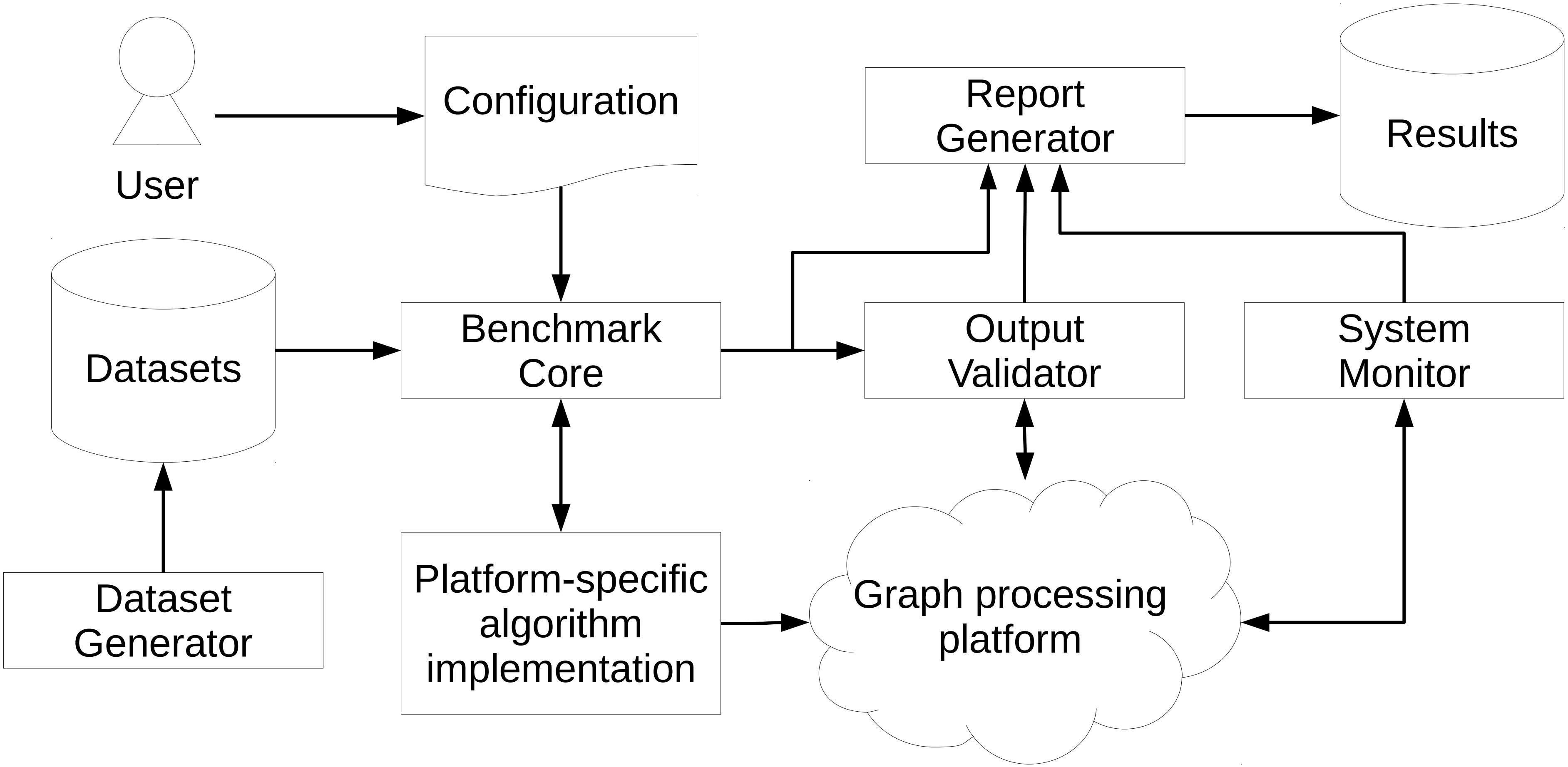
Realistic graph generator

- LDBC Datagen
  - Synthetic social network similar to Facebook
- Graphalytics enhancements
  - Multiple degree distributions
    - Zeta and geometric implemented
  - Other structural characteristics
    - Clustering coefficient
    - Assortativity

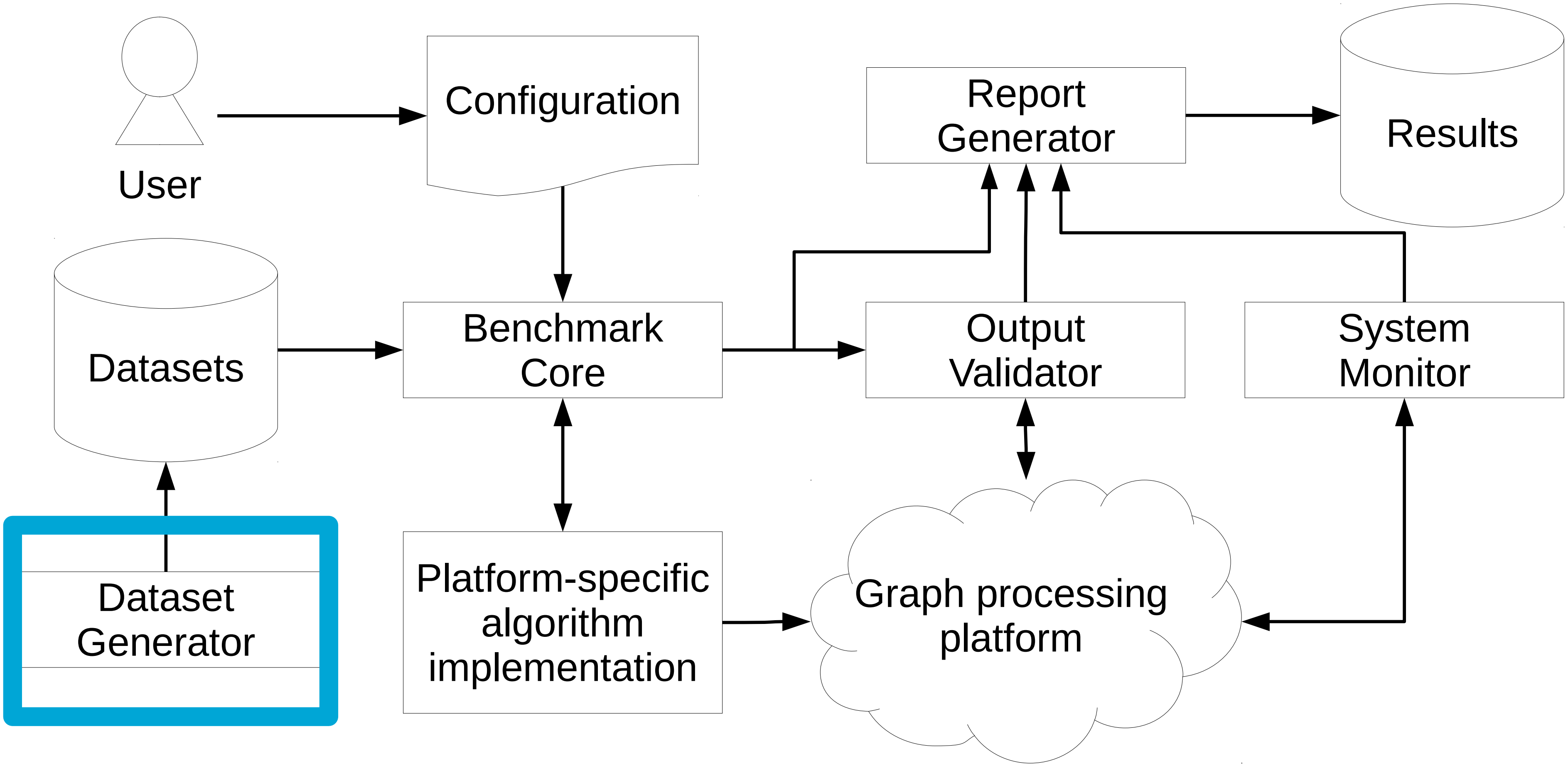
LDBC The graph & RDF benchmark reference

Erling et al., The LDBC Social Network Benchmark: Interactive Workload , SIGMOD 2015

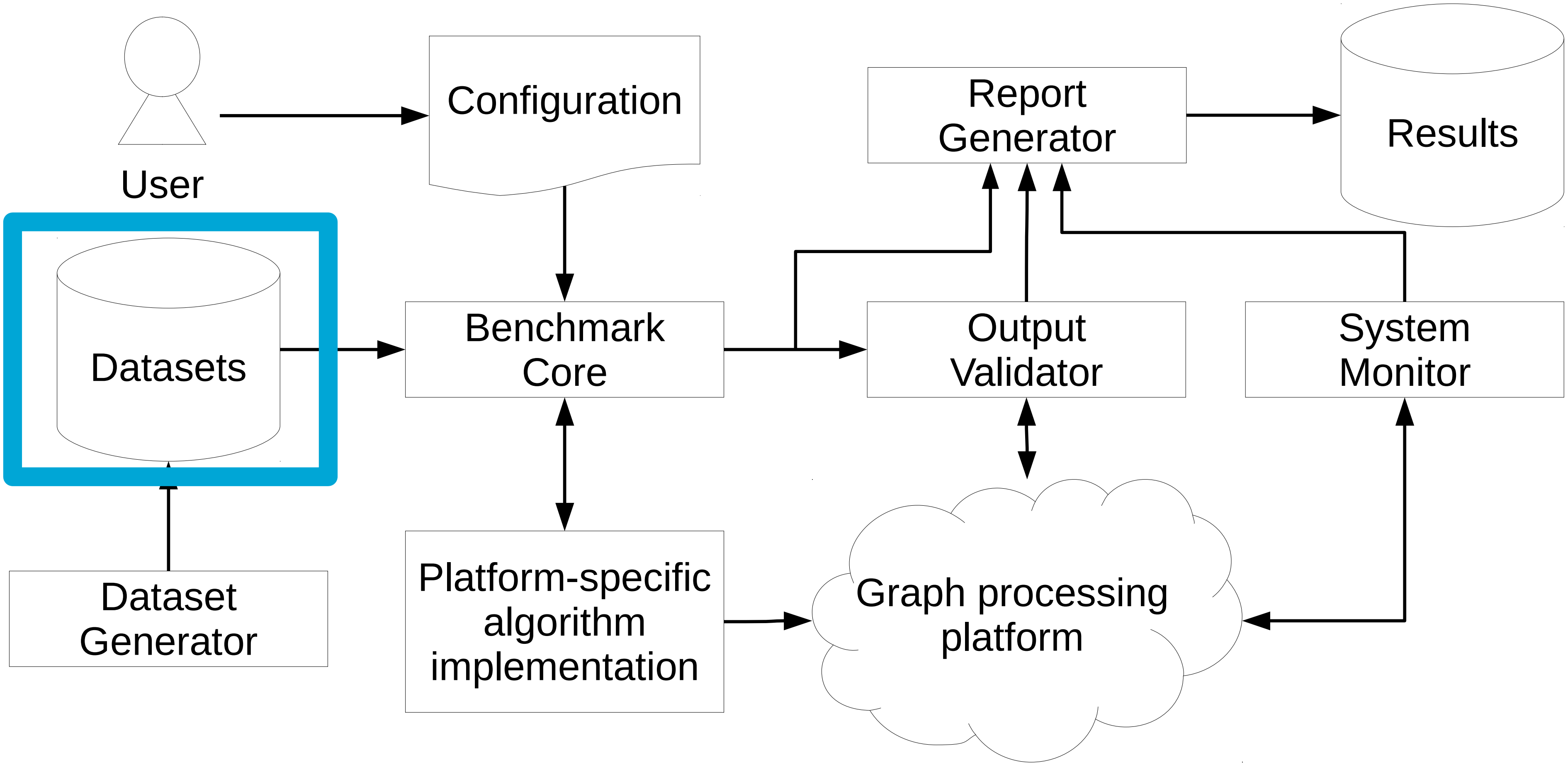
Advanced benchmark harness



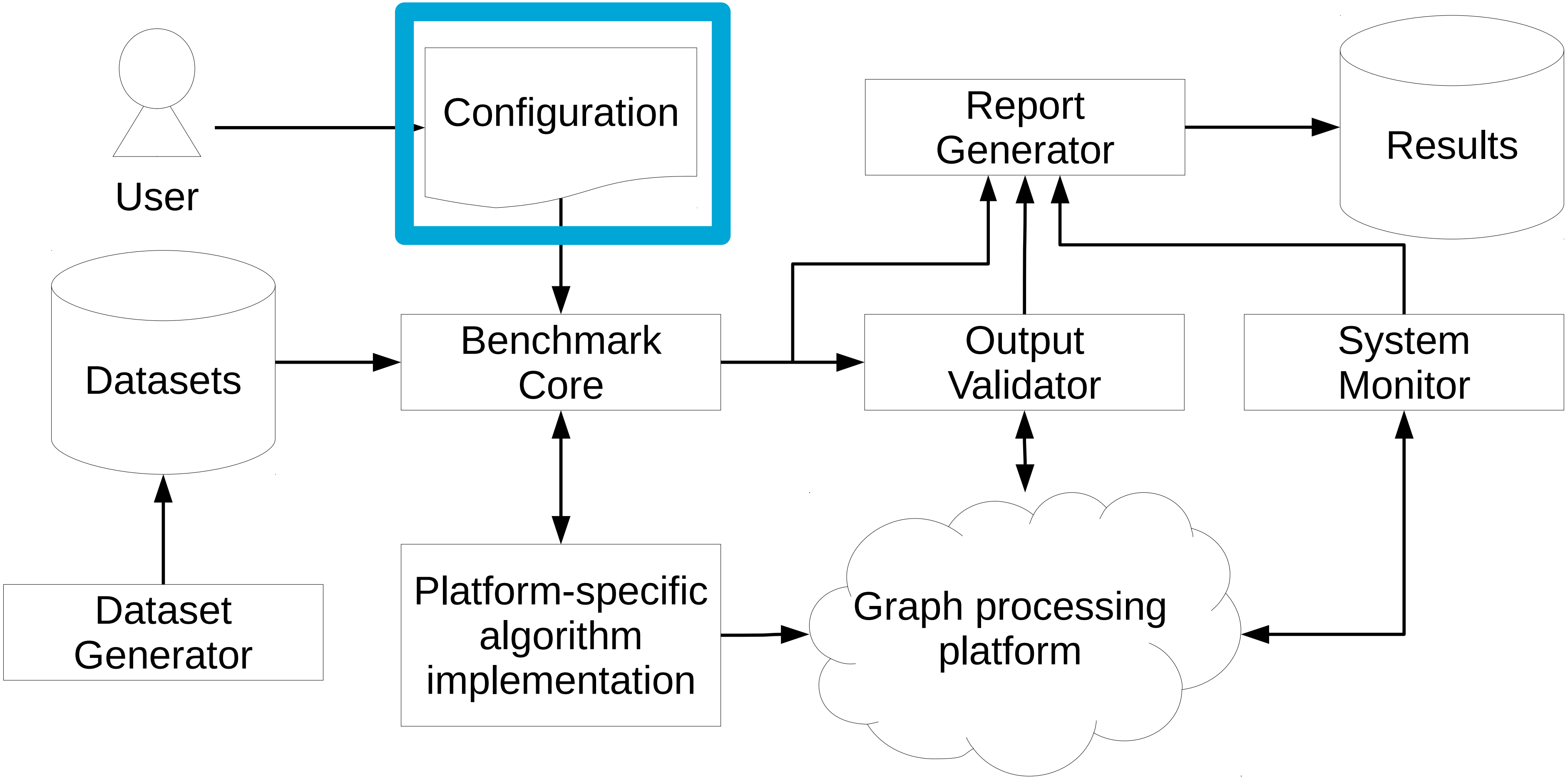
Advanced benchmark harness



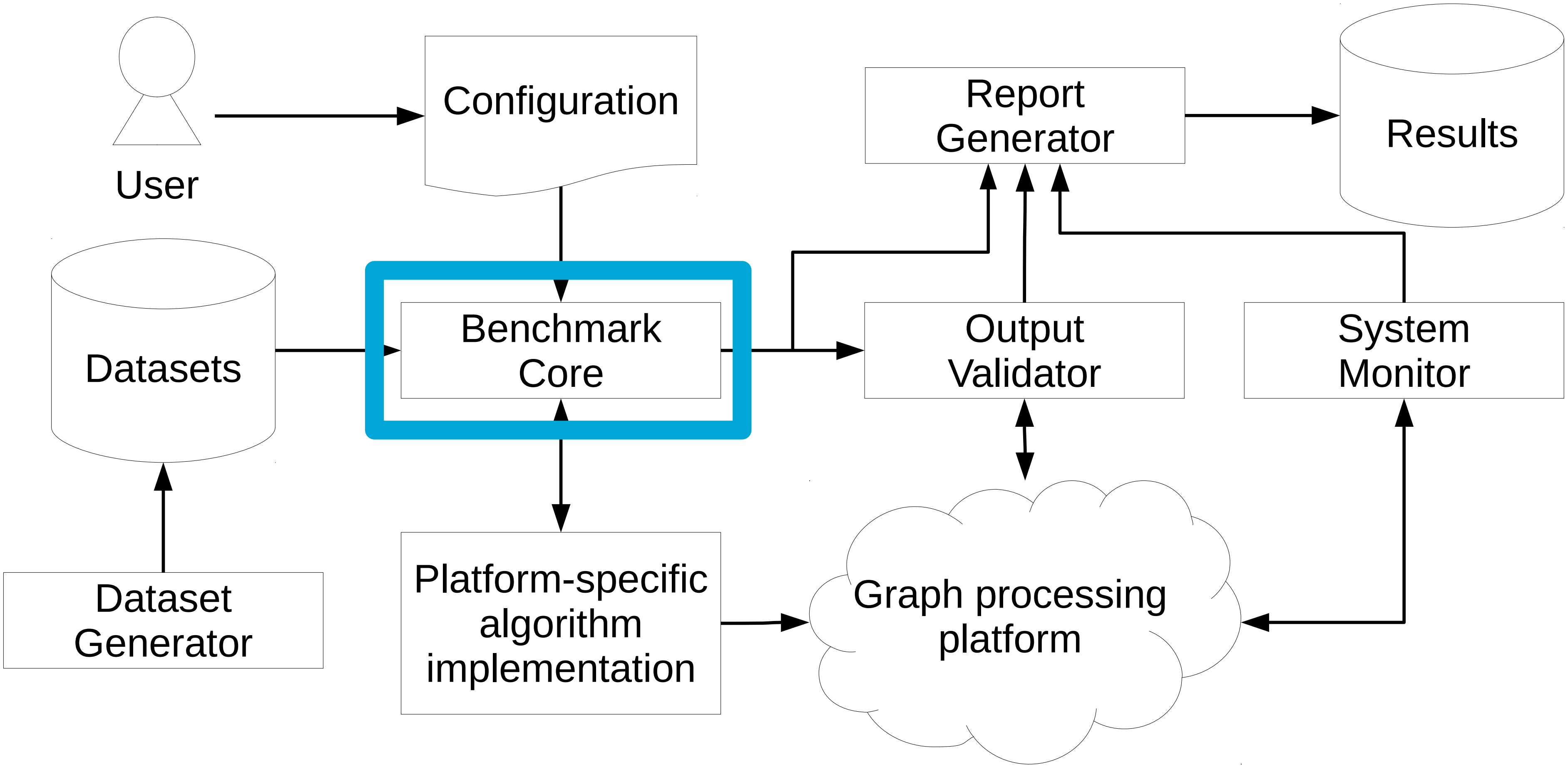
Advanced benchmark harness



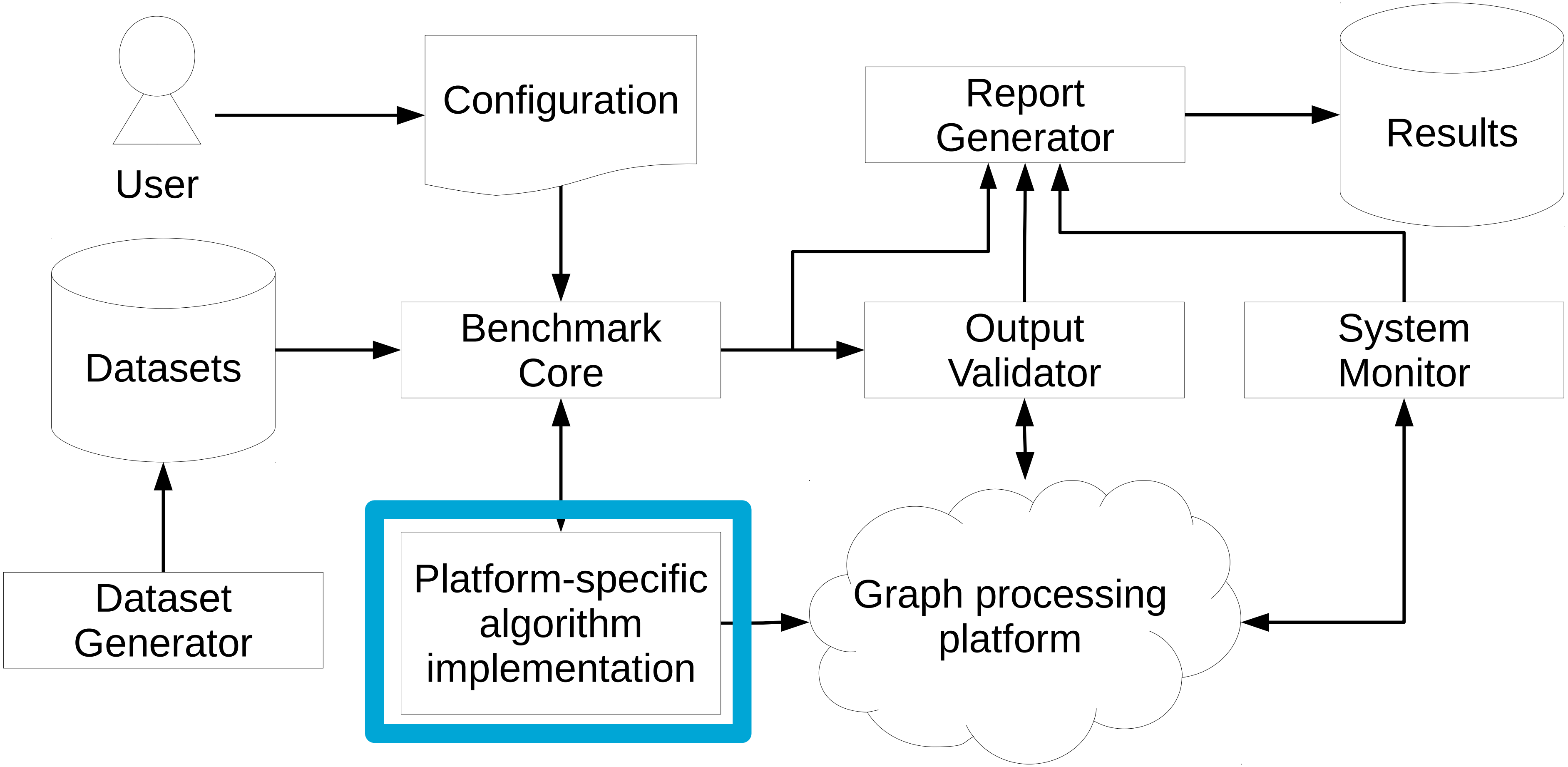
Advanced benchmark harness



Advanced benchmark harness

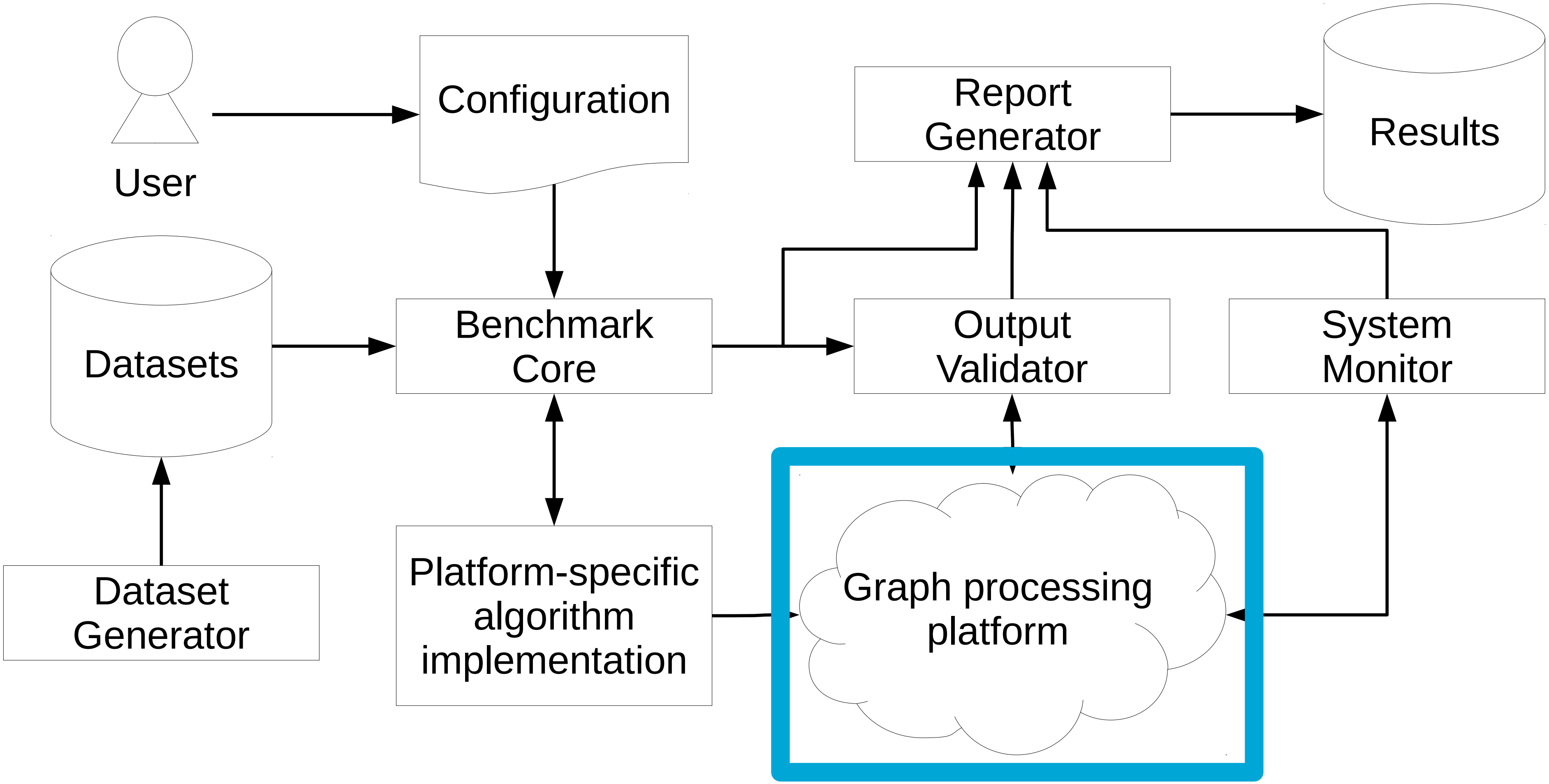


Advanced benchmark harness

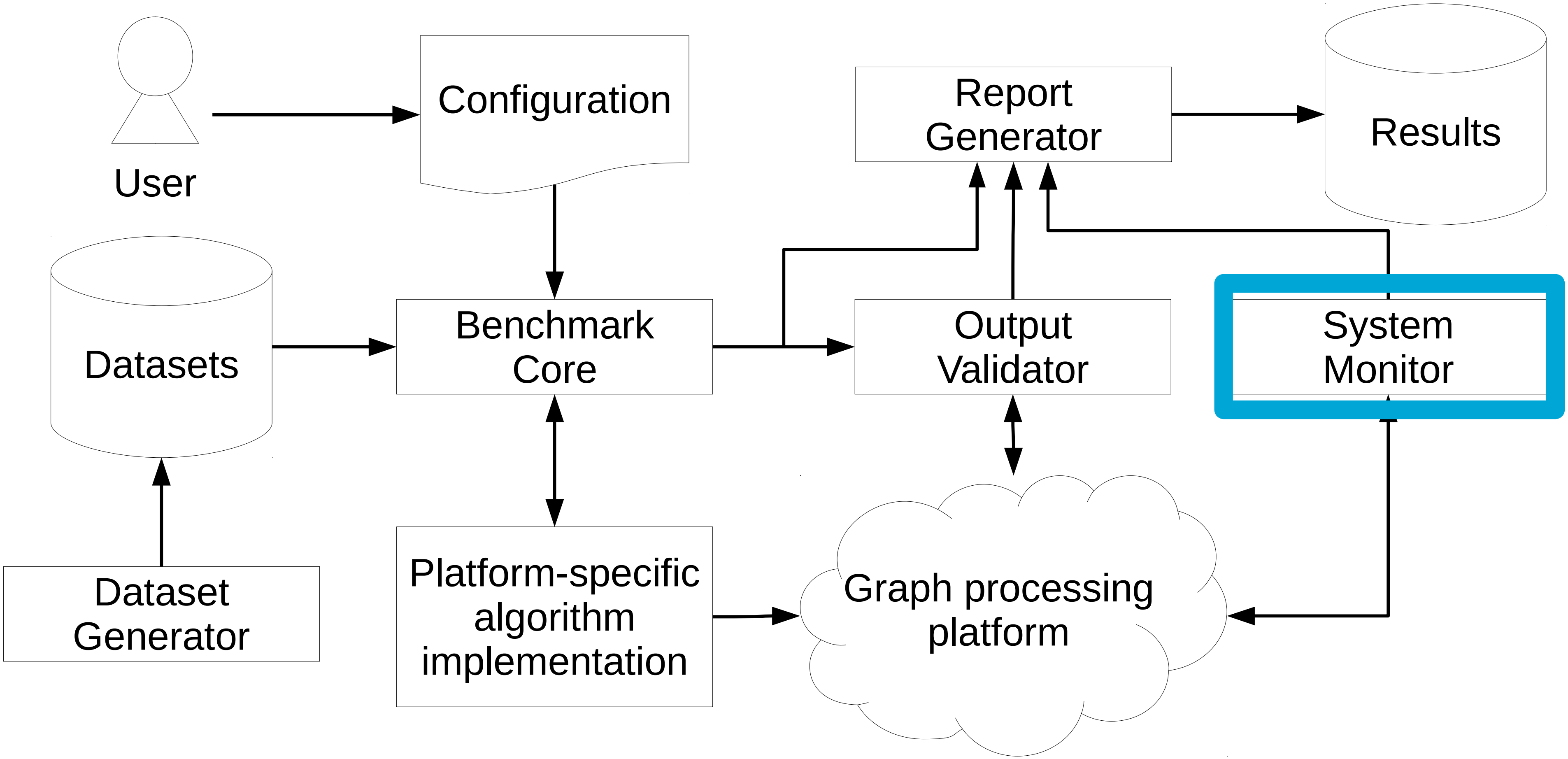


TUDelft

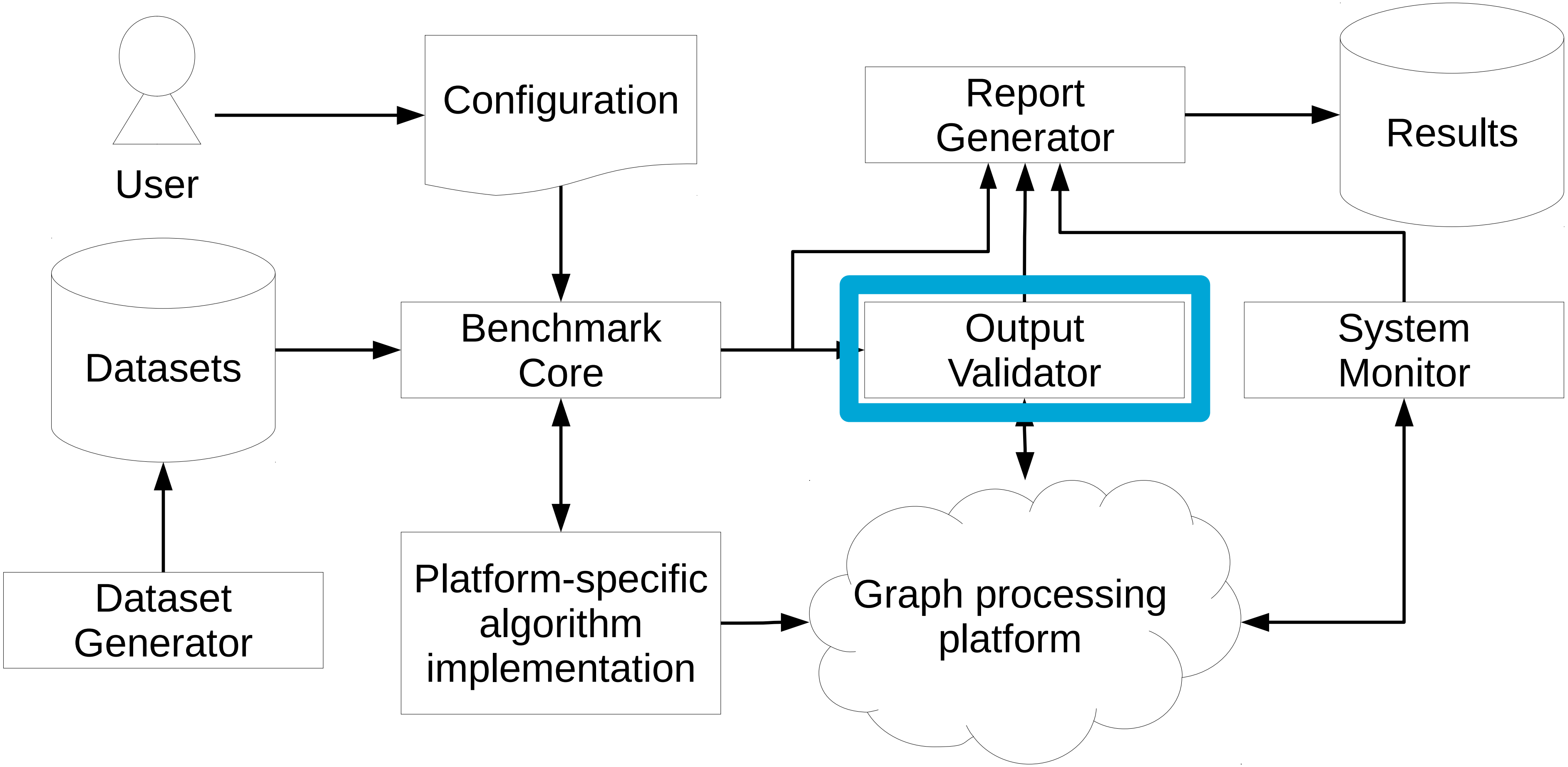
Advanced benchmark harness



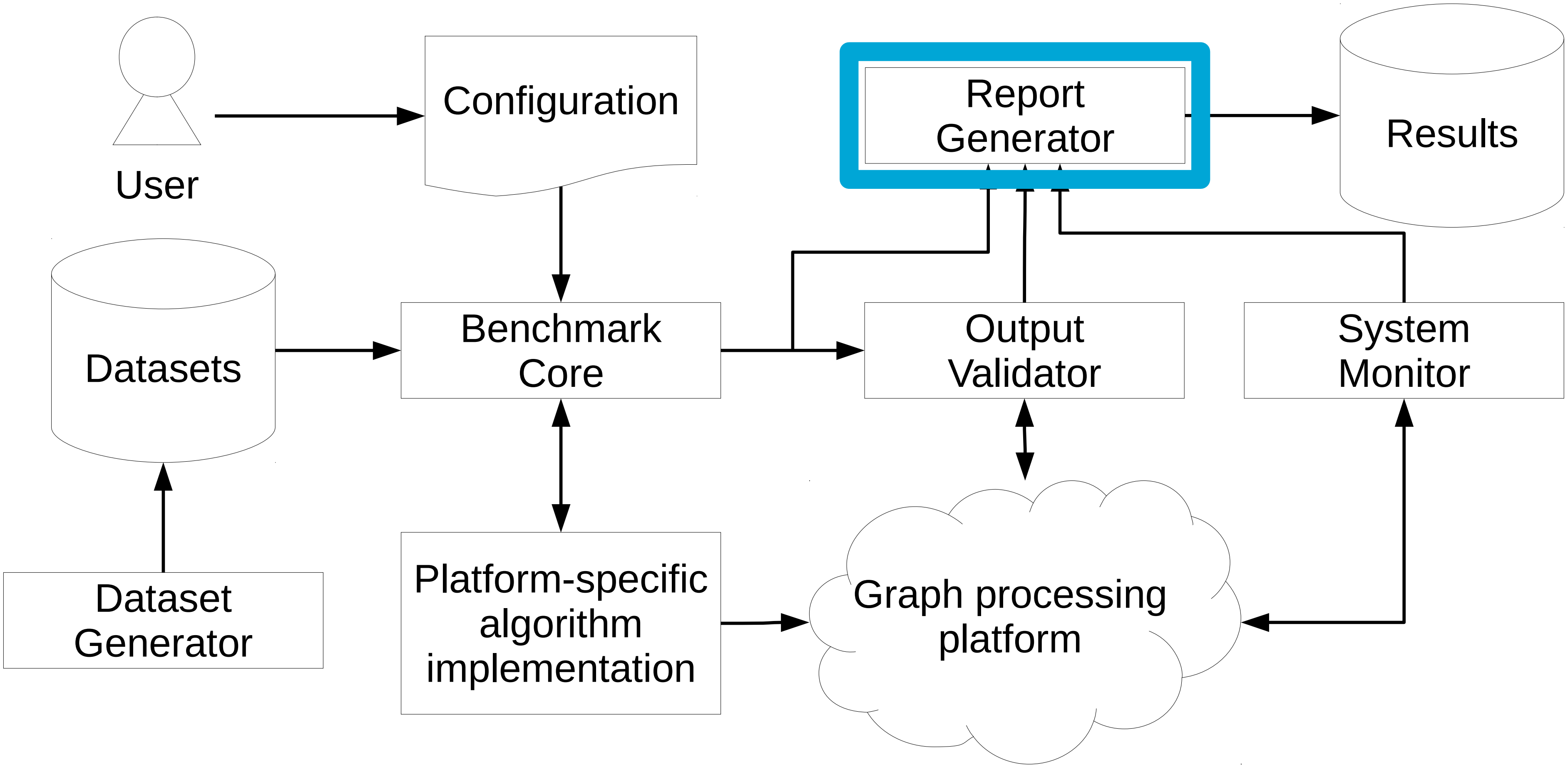
Advanced benchmark harness



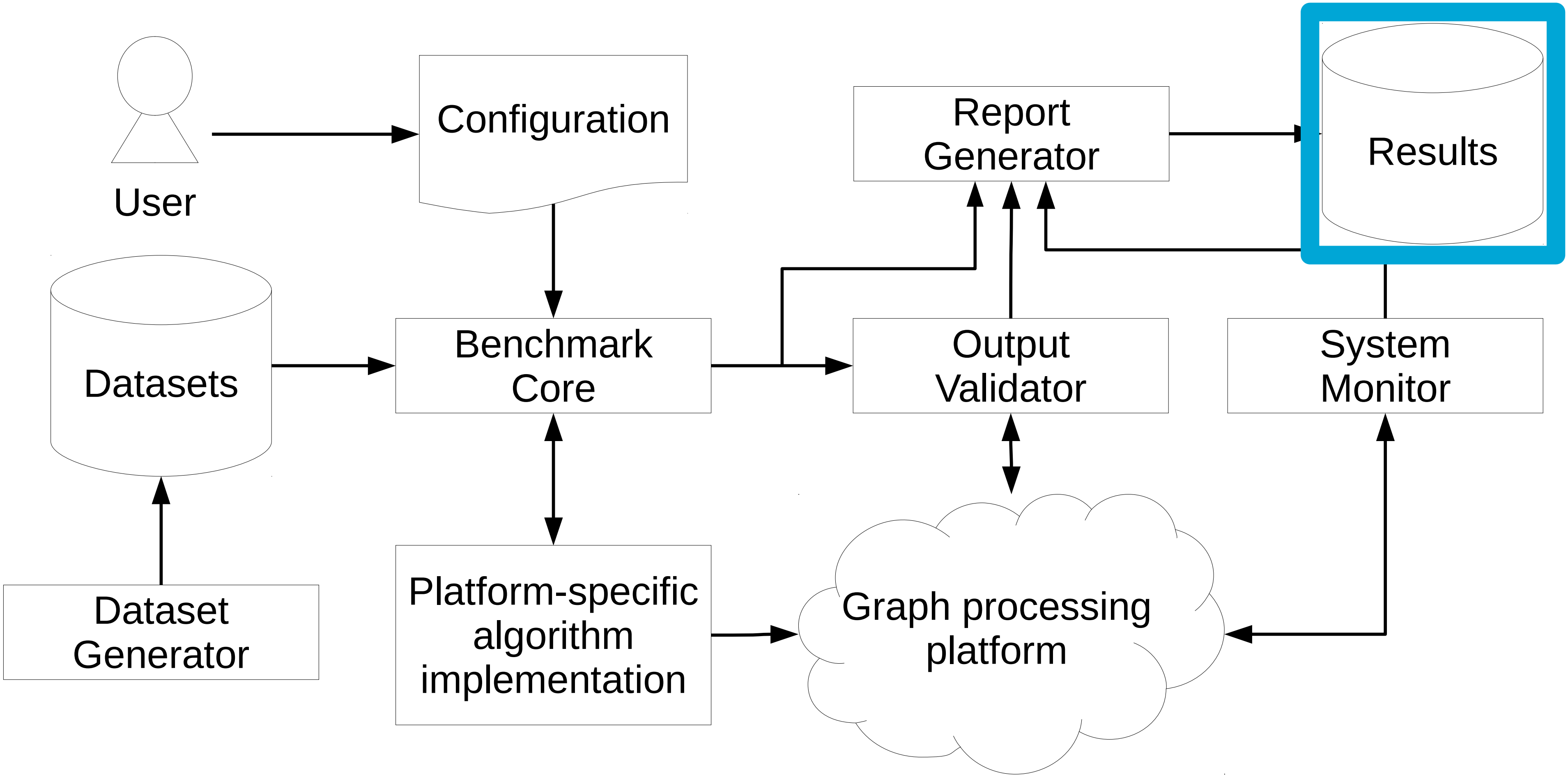
Advanced benchmark harness



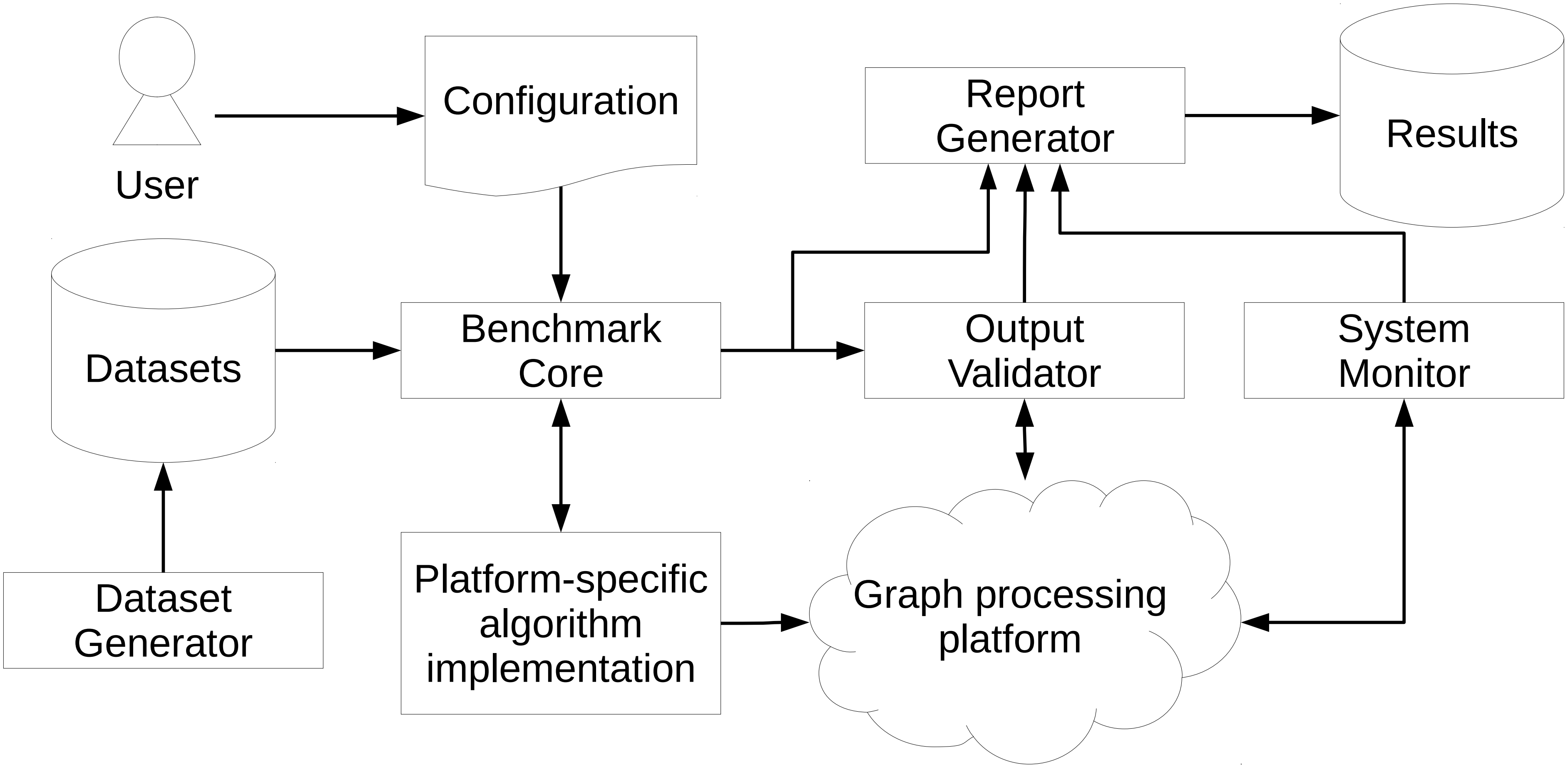
Advanced benchmark harness



Advanced benchmark harness



Advanced benchmark harness



Supported algorithms

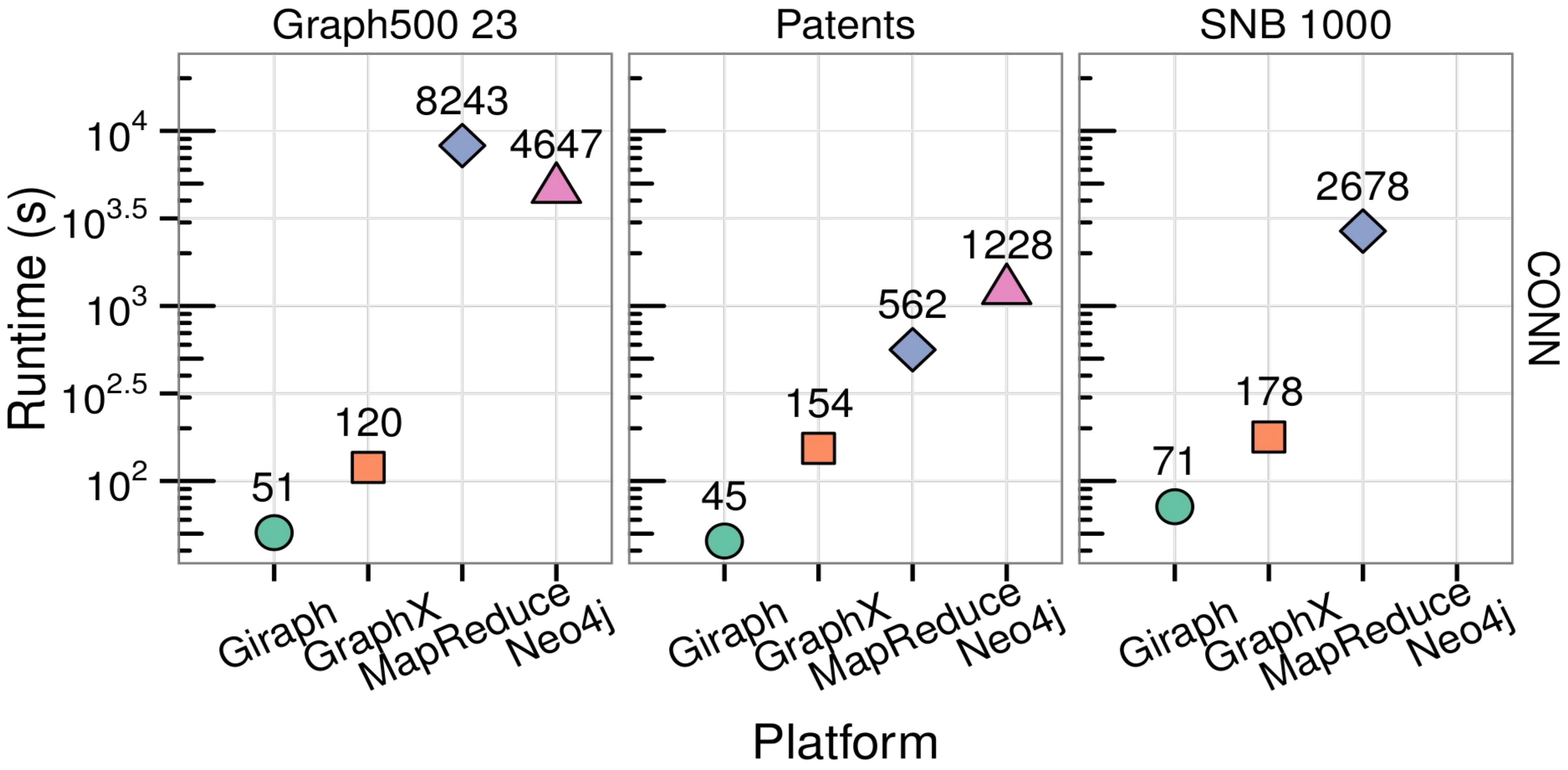
ID	Algorithm	Class	Use (%)
BFS	Breadth-first search	Traversal	46
STATS	Local clustering coefficient	Statistics	16
CONN	Weakly connected components	Connected components	13
CD	Label propagation	Community detection	5
EVO	Forest fire evolution	Evolution	4

Guo et al., How Well Do Graph-Processing Platforms Perform? An Empirical Performance Evaluation and Analysis, IPDPS 2014

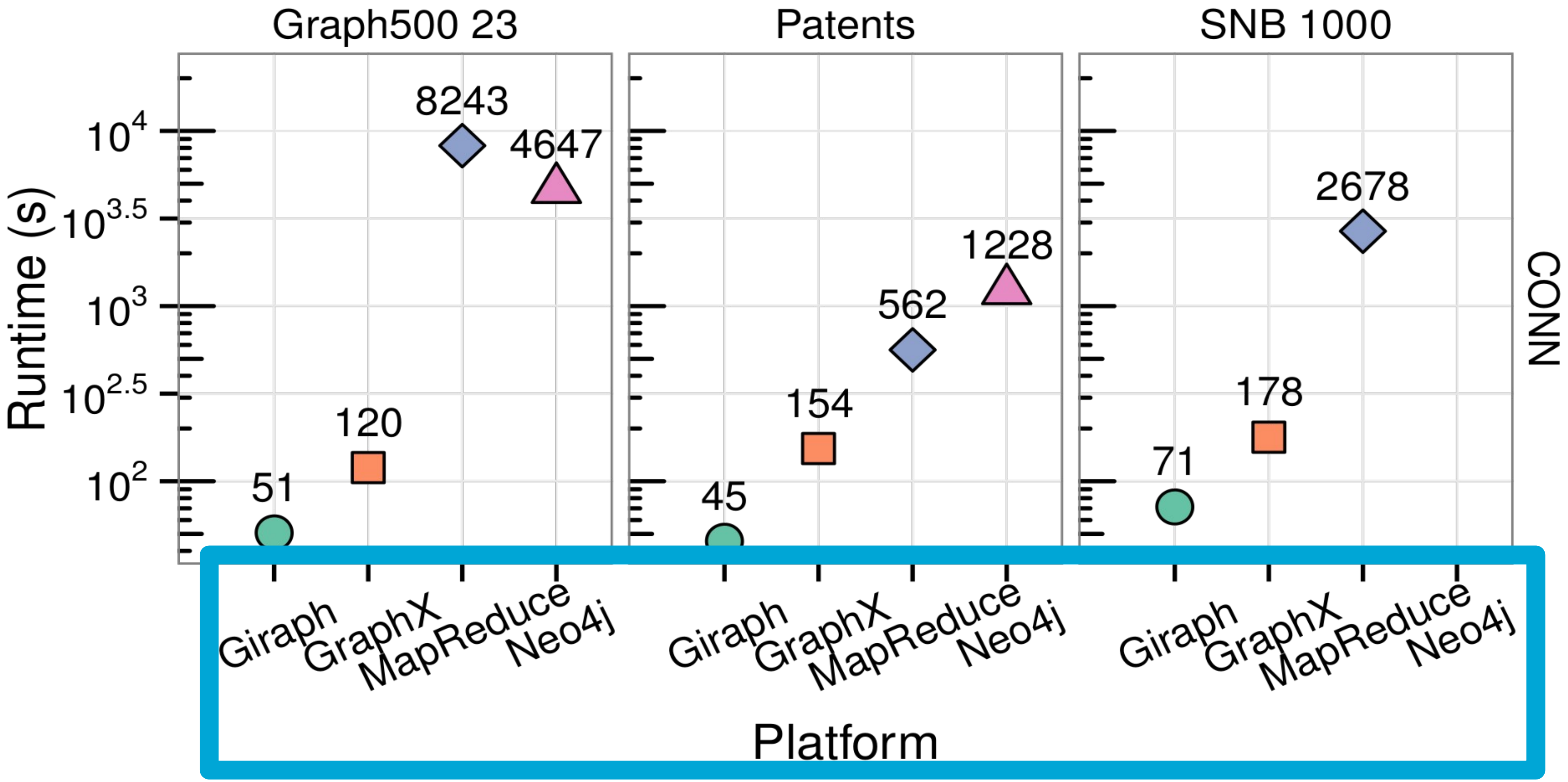
Experimental setup

- DAS-4 cluster
  - Typical big data setup
  - 11 nodes, 24 GiB RAM, 2 x 8-core Xeon E5620
  - 1 Gbit/s Ethernet
- Single machine
  - HPC-like setup
  - 192 GiB RAM, 2 x 8-core Xeon E5-2450 v2

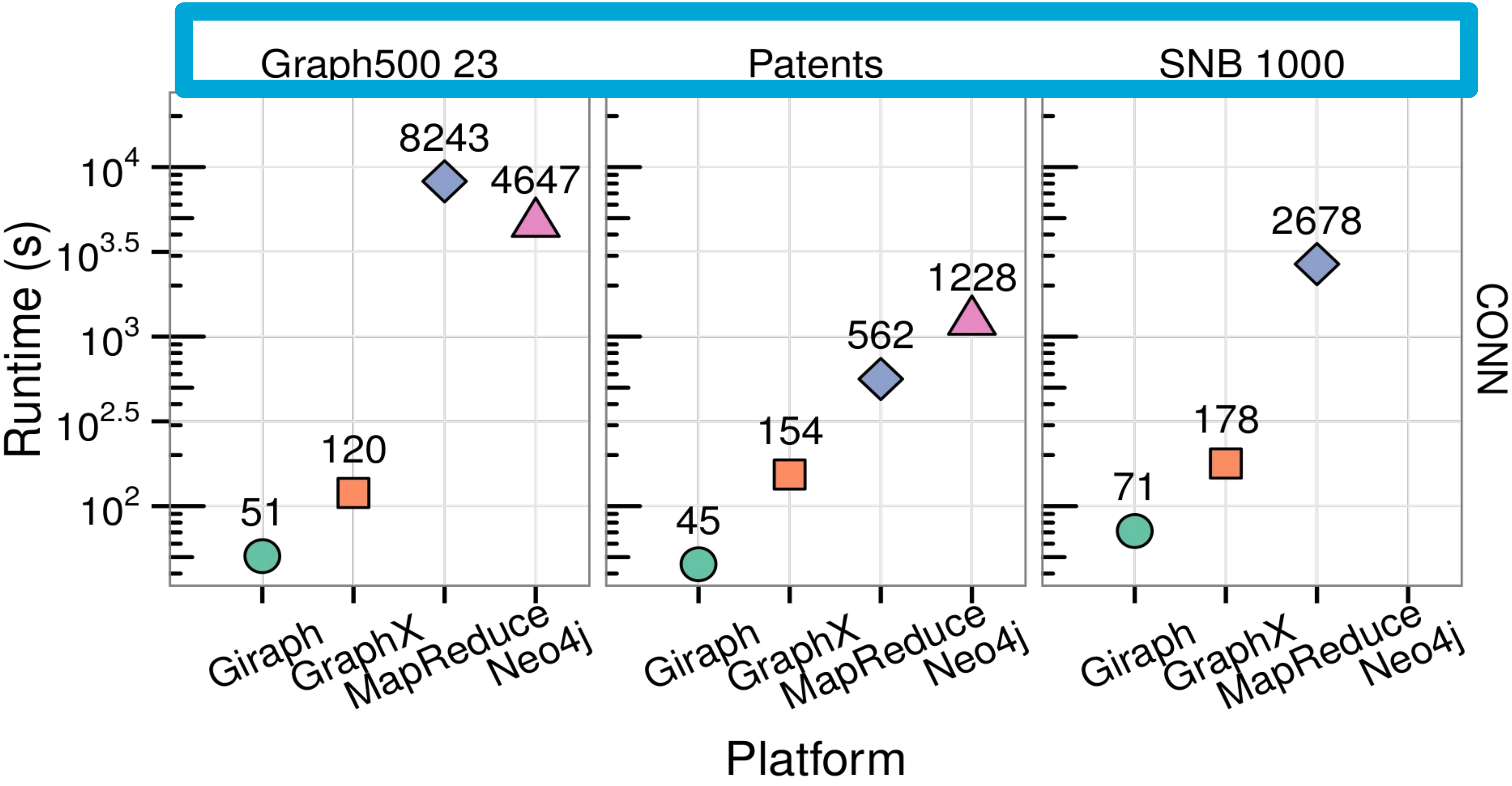
Runtime


Graph500 23
Patents
SNB 1000
Runtime (s)
10^4
$10^{3.5}$
10^3
$10^{2.5}$
10^2
8243
4647
120
51
1228
562
154
45
2678
178
71
CONN
Giraph
GraphX
MapReduce
Neo4j
Platform

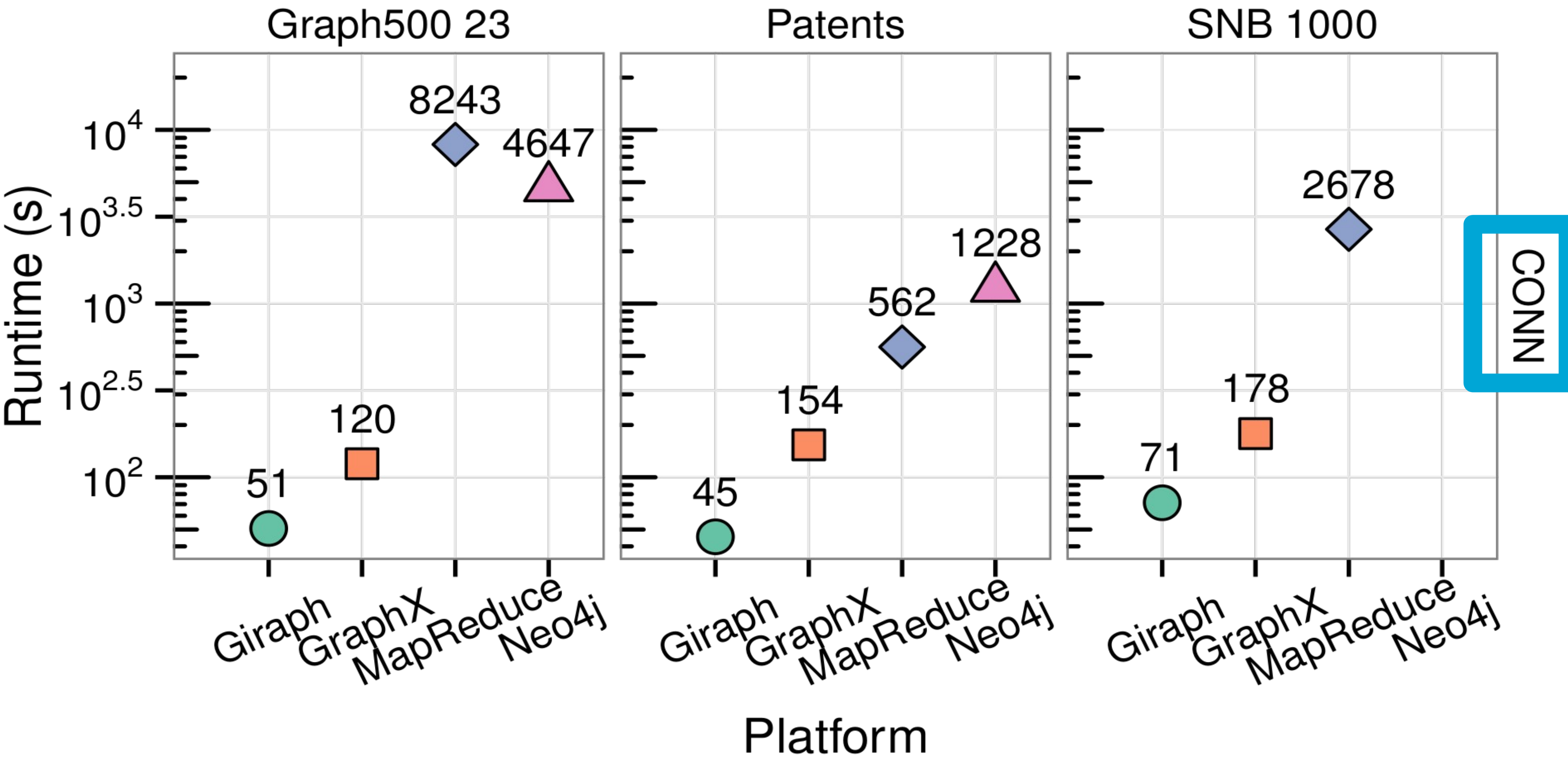
Runtime


Graph500 23
Patents
SNB 1000
Runtime (s)
10^4
$10^{3.5}$
10^3
$10^{2.5}$
10^2
51
120
8243
4647
45
154
562
1228
71
178
2678
Giraph
GraphX
MapReduce
Neo4j
Platform
CONN

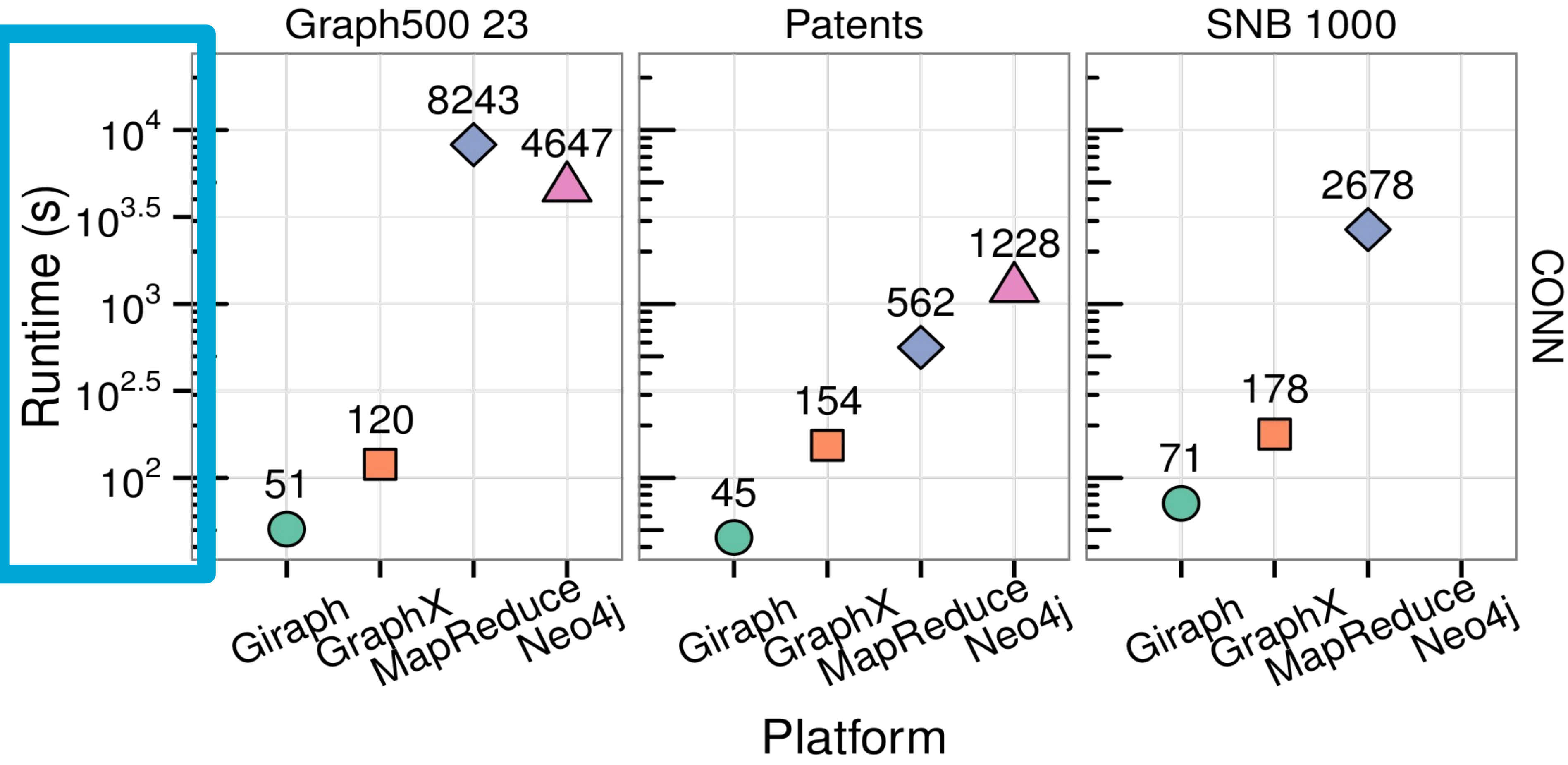
Runtime


Graph500 23
Patents
SNB 1000
Runtime (s)
10^4
$10^{3.5}$
10^3
$10^{2.5}$
10^2
8243
4647
120
51
1228
562
154
45
2678
178
71
CONN
Giraph
GraphX
MapReduce
Neo4j
Platform

Runtime

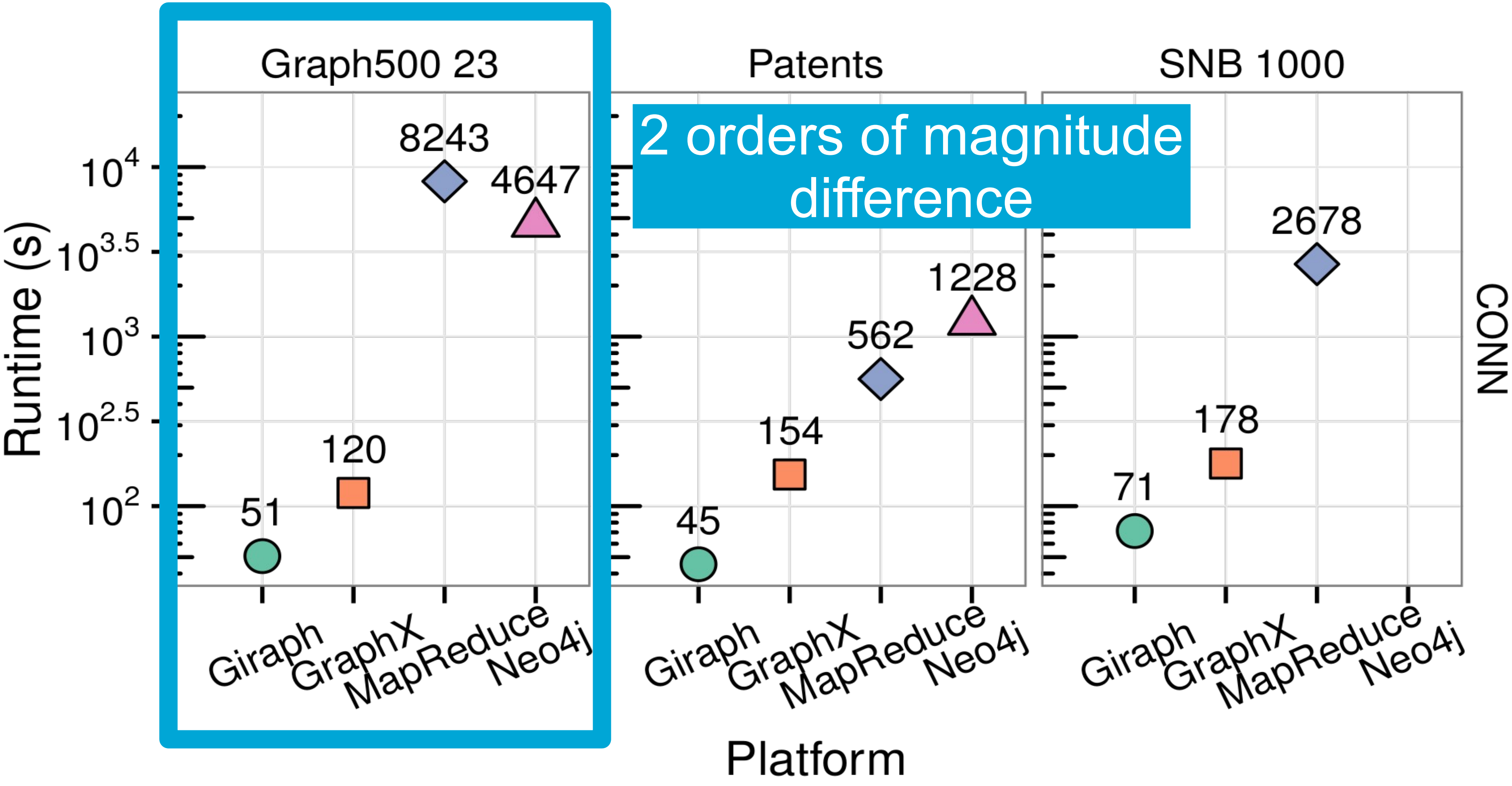

Graph500 23
Patents
SNB 1000
Runtime (s)
10^4
$10^{3.5}$
10^3
$10^{2.5}$
10^2
51
120
8243
4647
45
154
562
1228
71
178
2678
Giraph
GraphX
MapReduce
Neo4j
Giraph
GraphX
MapReduce
Neo4j
Giraph
GraphX
MapReduce
Neo4j
Platform
CONN

Runtime



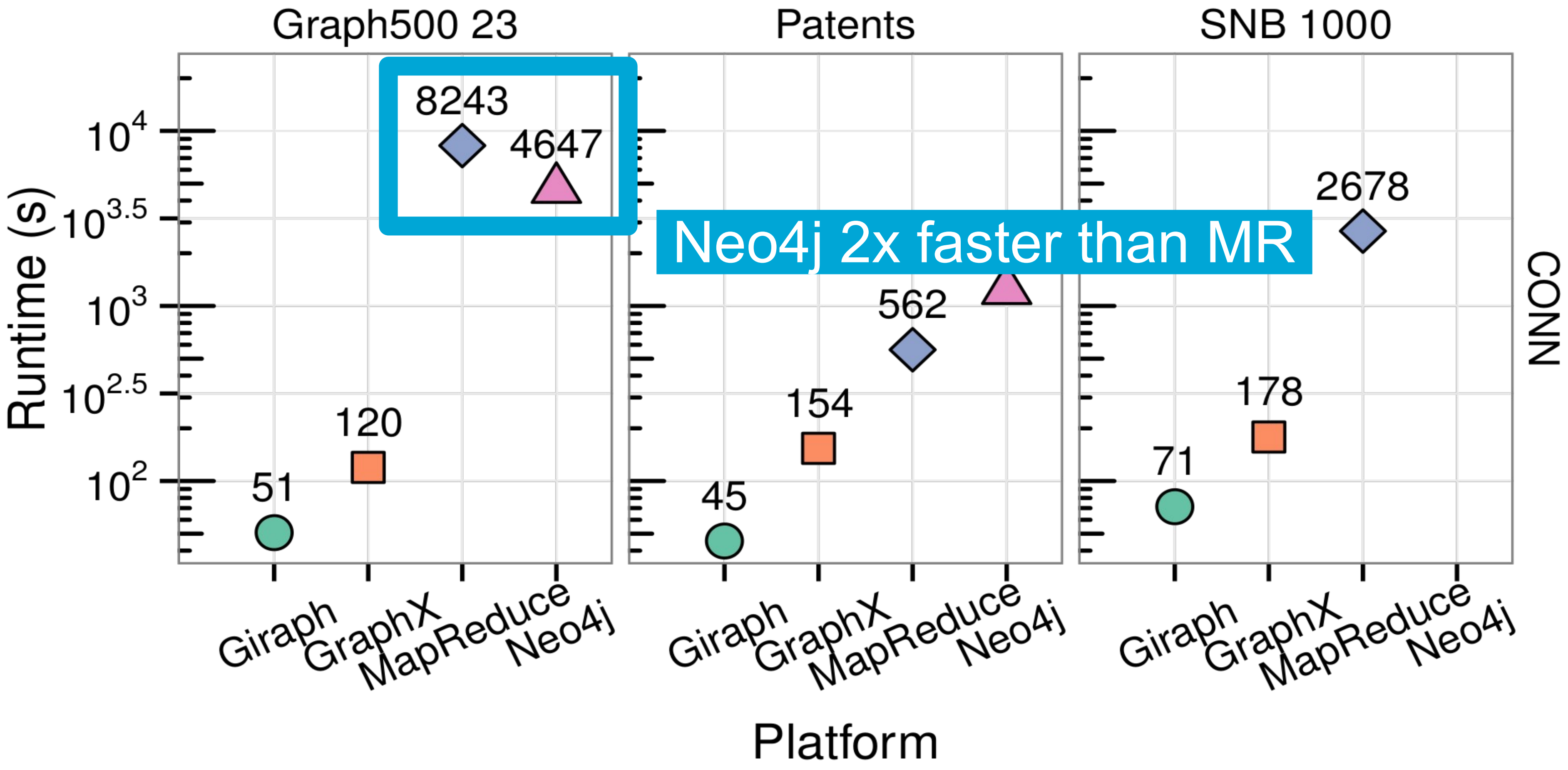
Graph500 23
Patents
SNB 1000
Runtime (s)
10^4
$10^{3.5}$
10^3
$10^{2.5}$
10^2
51
120
8243
4647
45
154
562
1228
71
178
2678
CONN
Giraph
GraphX
MapReduce
Neo4j
Platform

Runtime



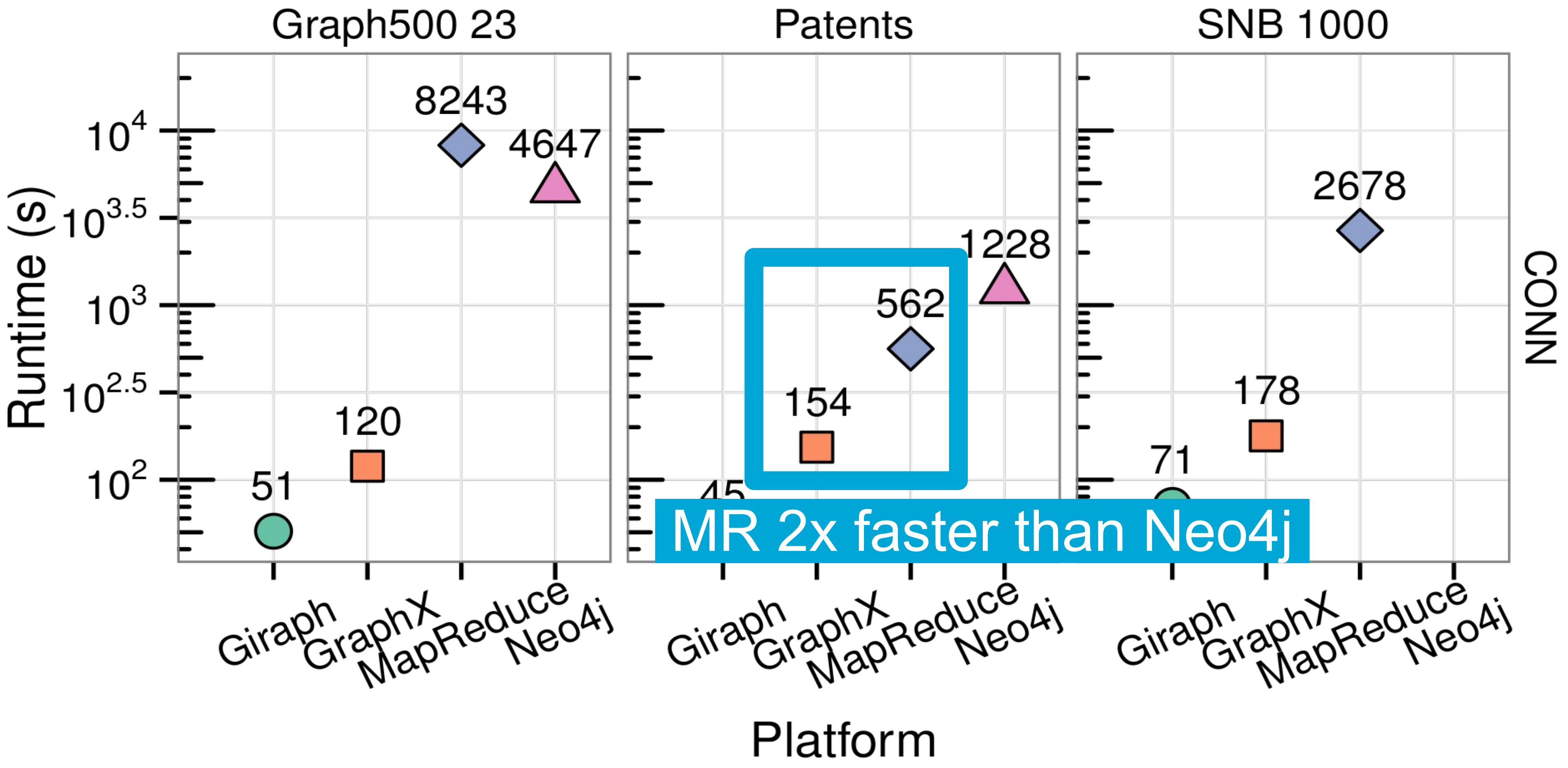
2 orders of magnitude difference

Runtime



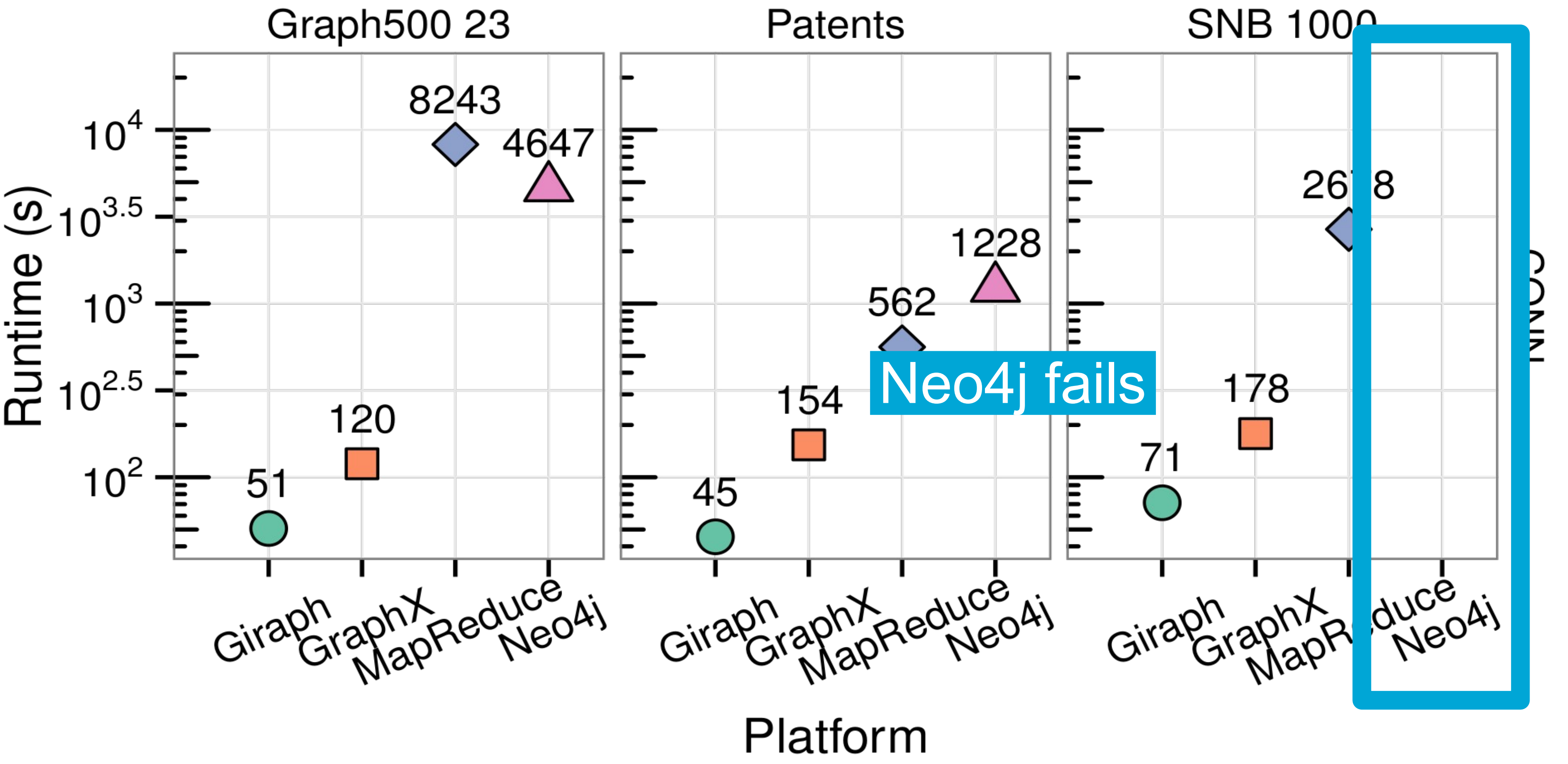
Neo4j 2x faster than MR

Runtime



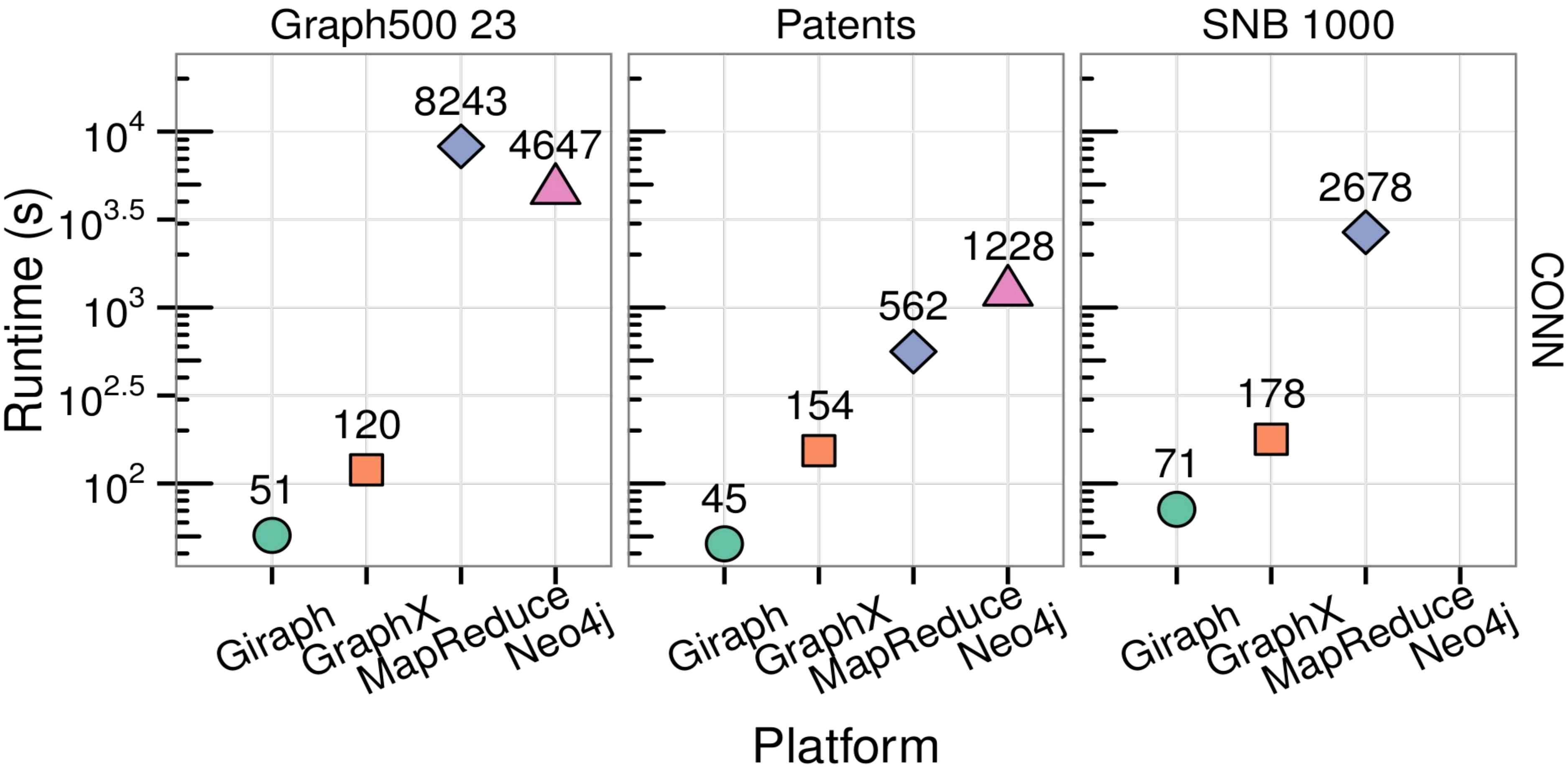
MR 2x faster than Neo4j

Runtime



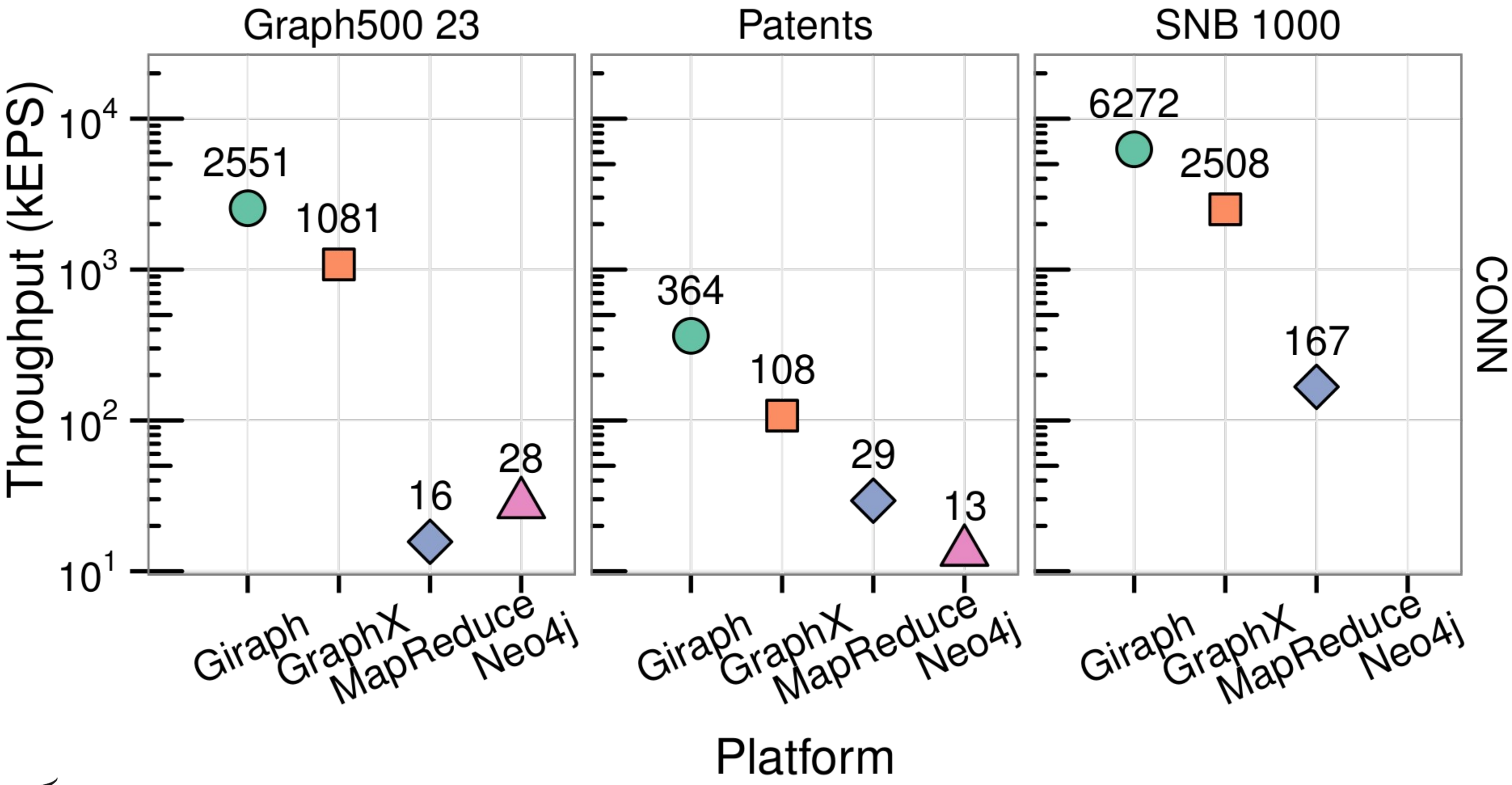
Neo4j fails

Runtime



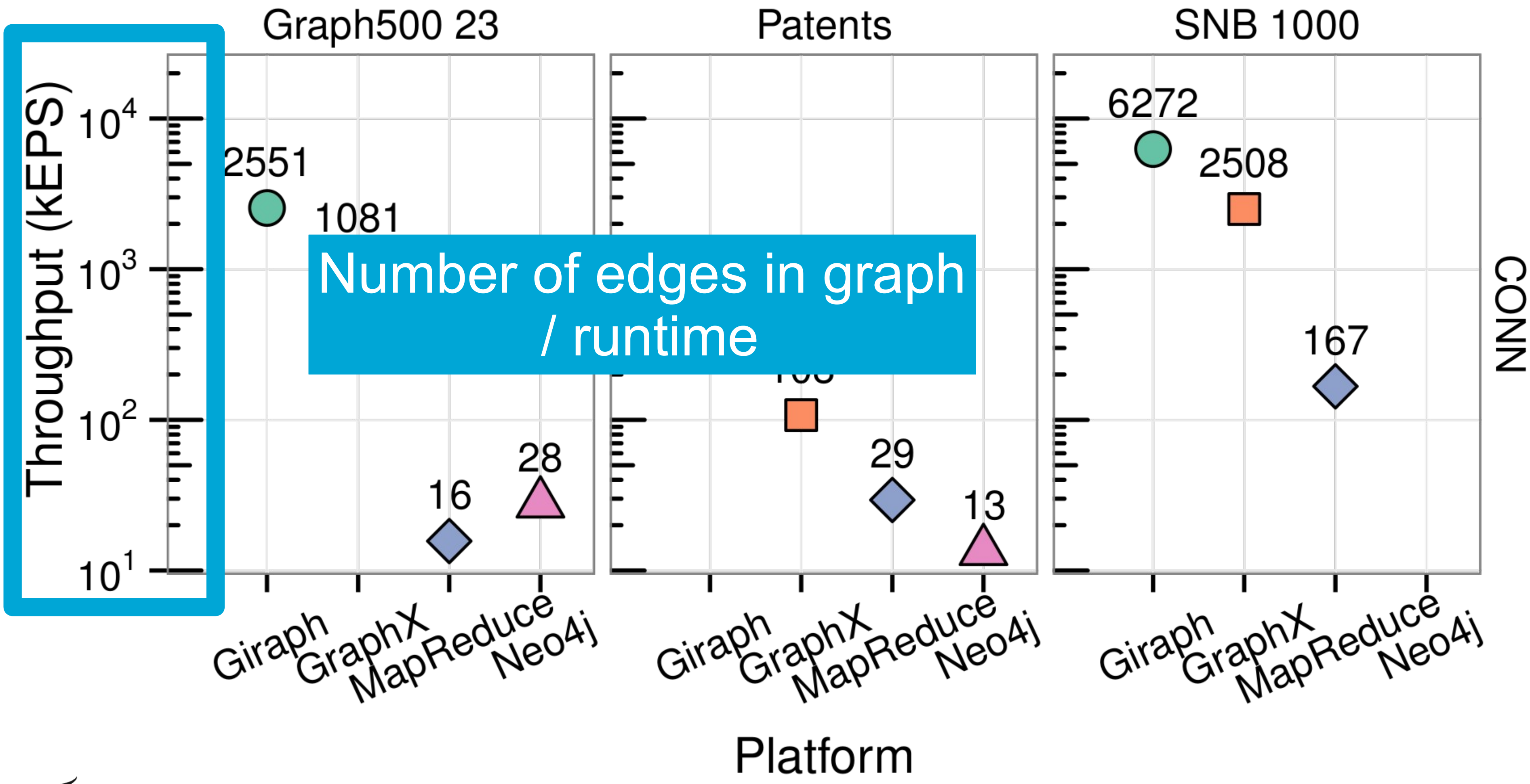
Graph500 23
Patents
SNB 1000
Runtime (s)
10^4
$10^{3.5}$
10^3
$10^{2.5}$
10^2
51
120
8243
4647
45
154
562
1228
71
178
2678
CONN
Giraph
GraphX
MapReduce
Neo4j
Platform

Edge-normalized performance



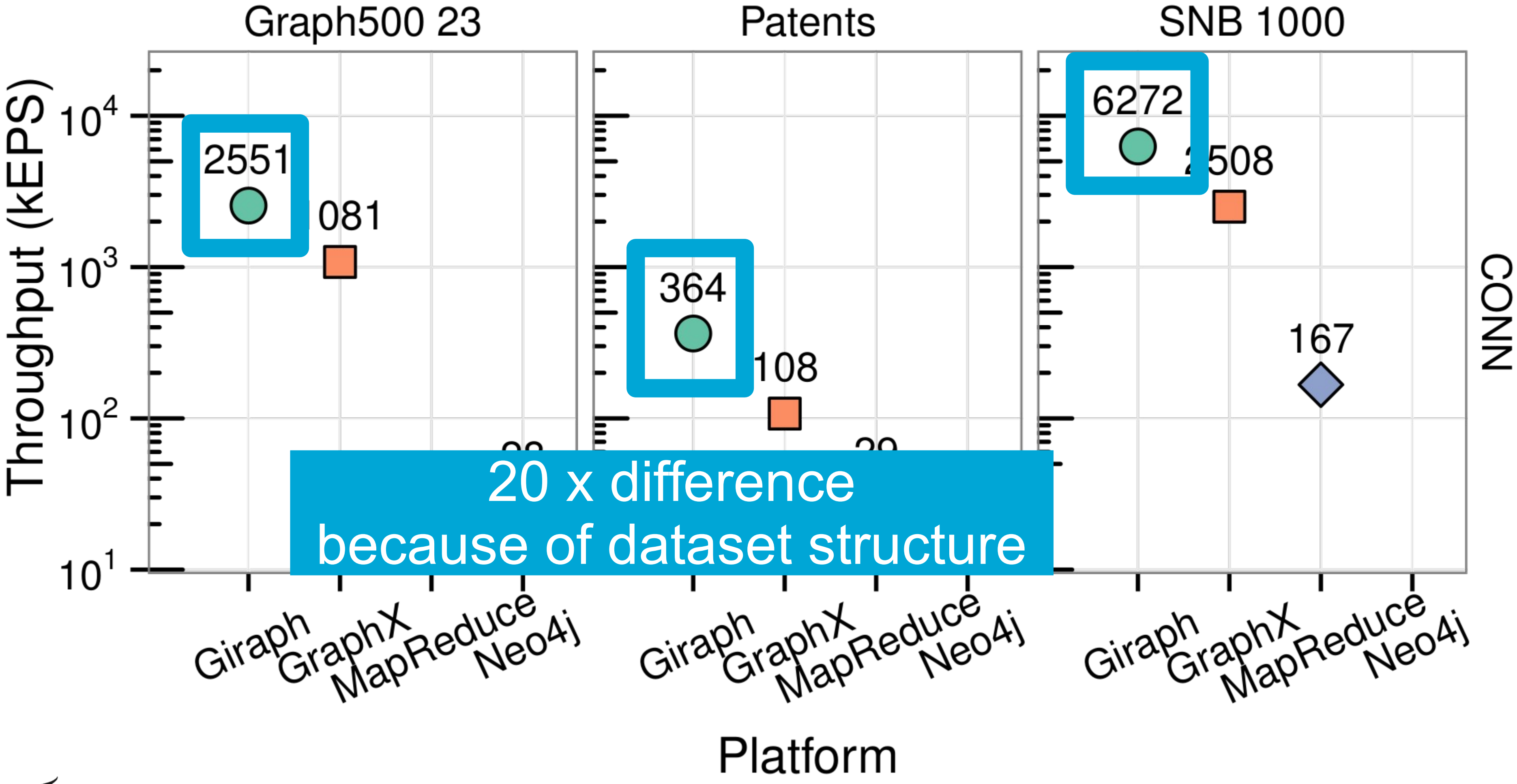
TUDelft

Edge-normalized performance



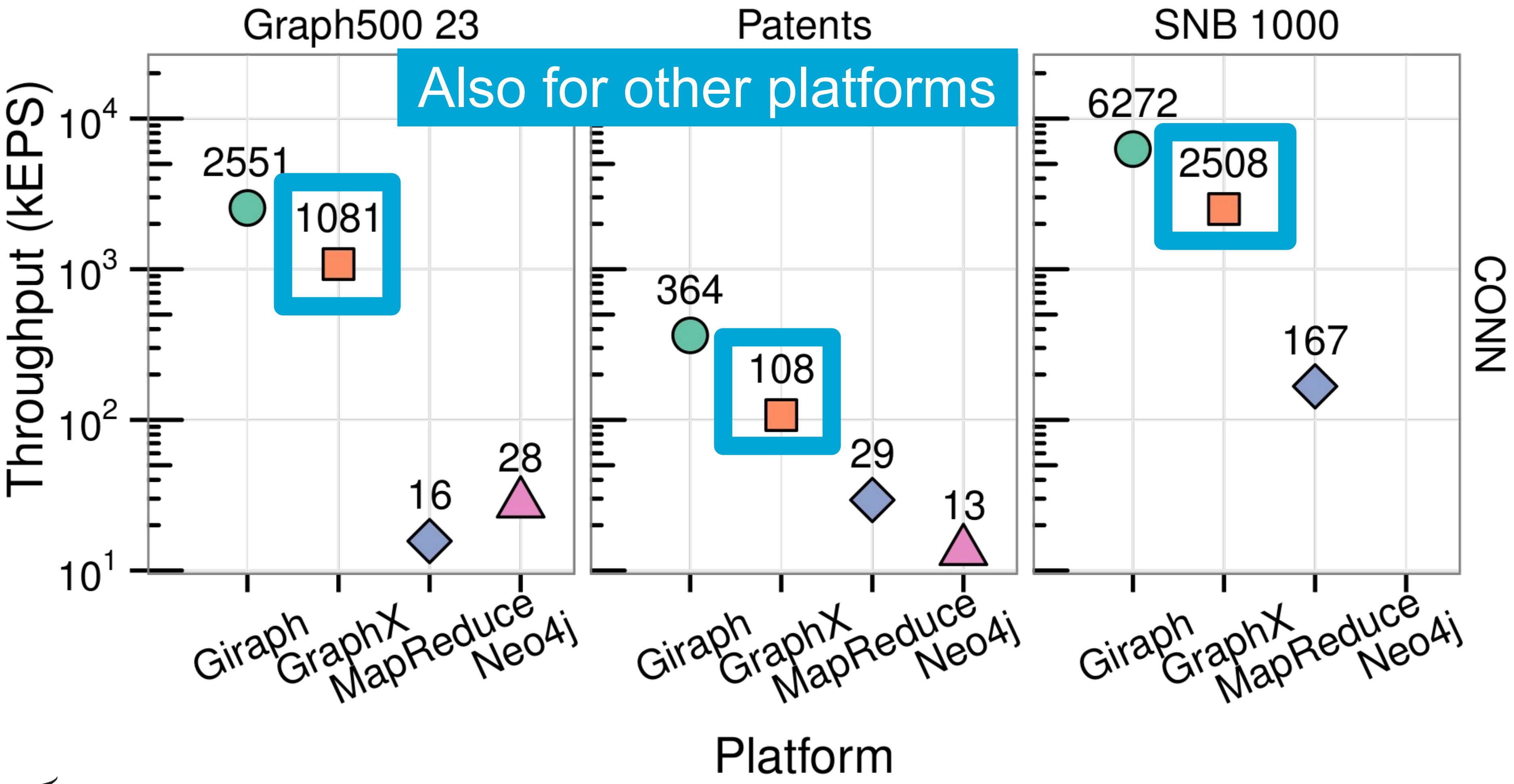
Number of edges in graph / runtime

Edge-normalized performance



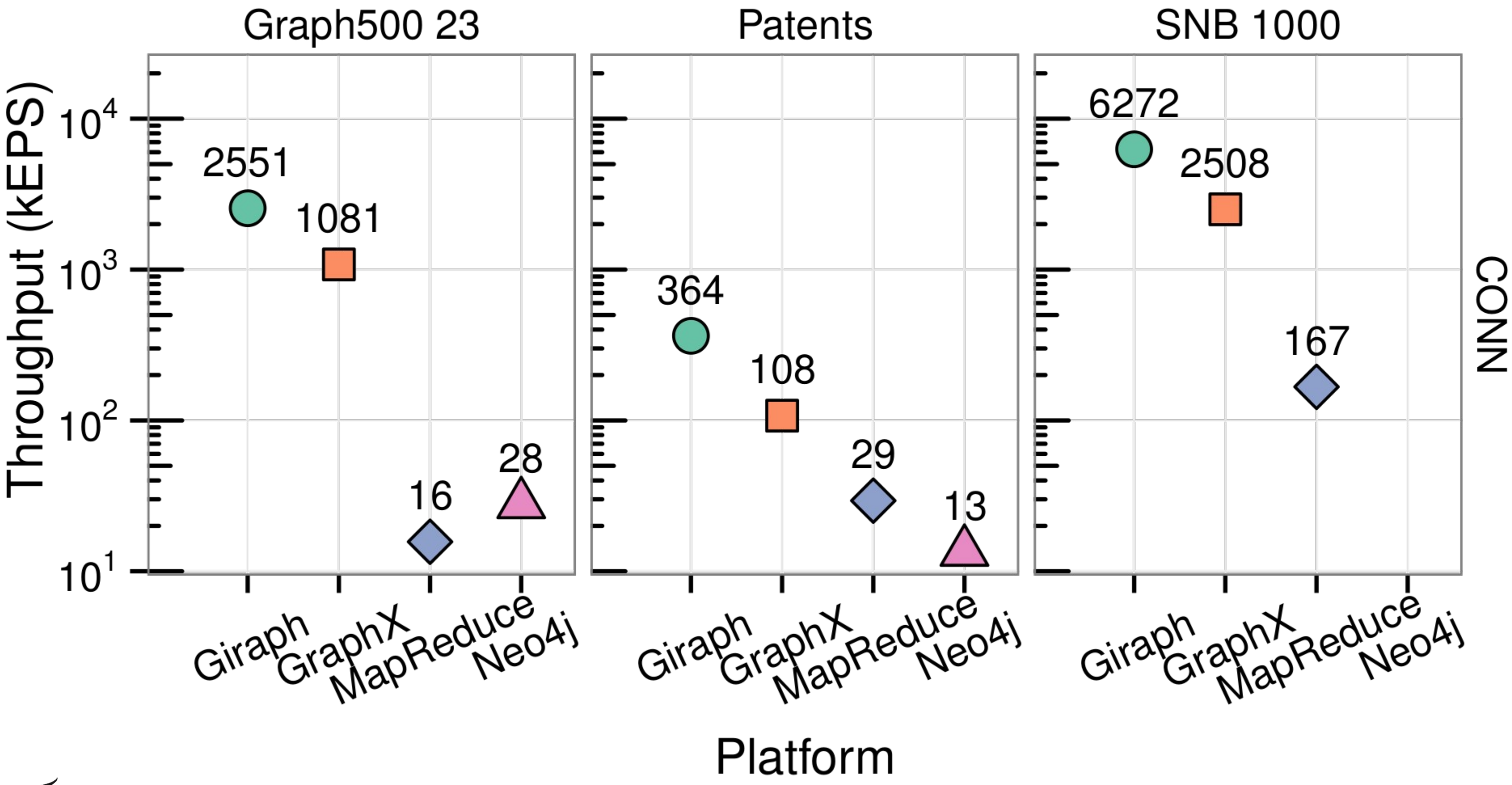
20 x difference
because of dataset structure

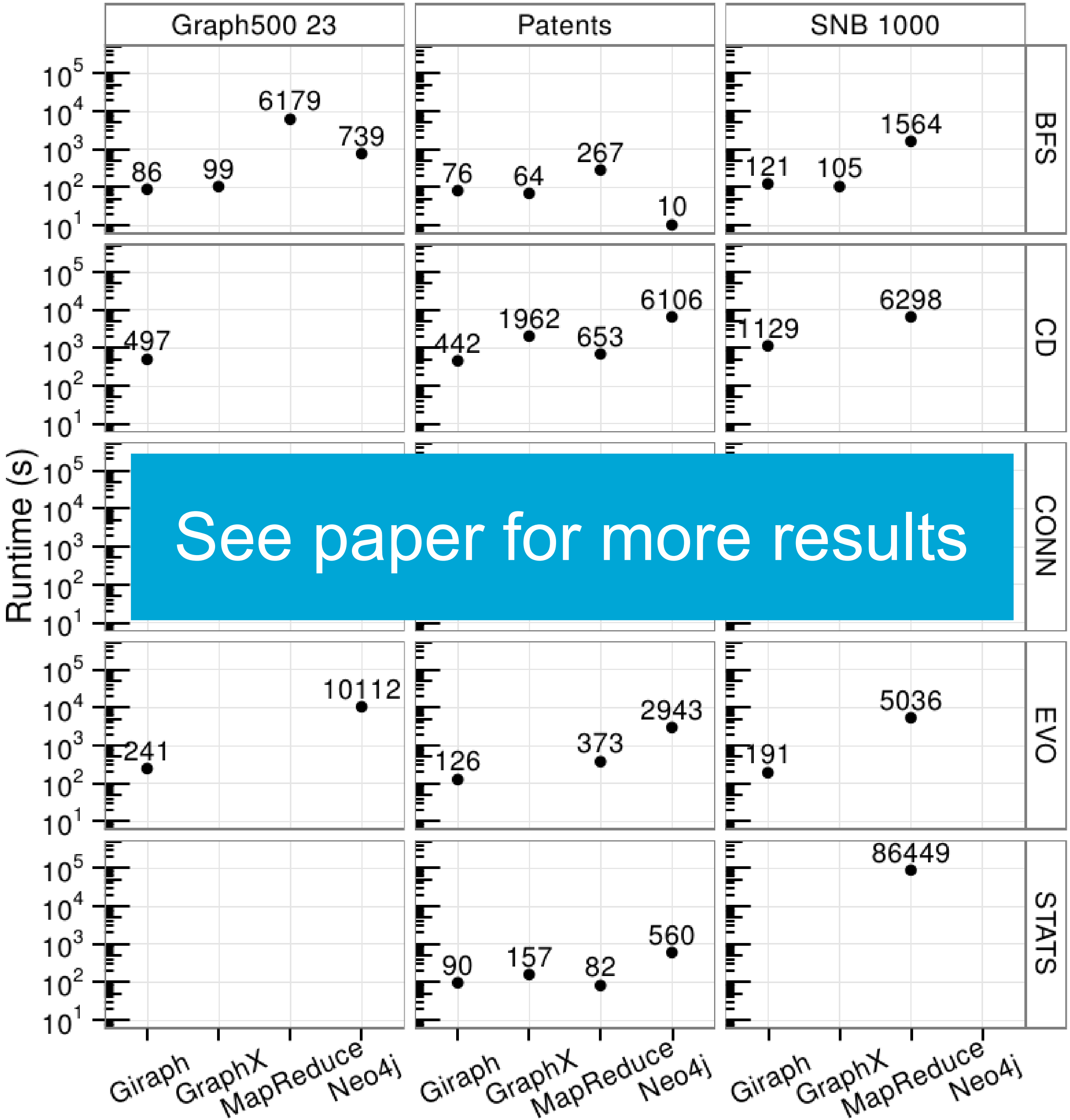
Edge-normalized performance



Also for other platforms

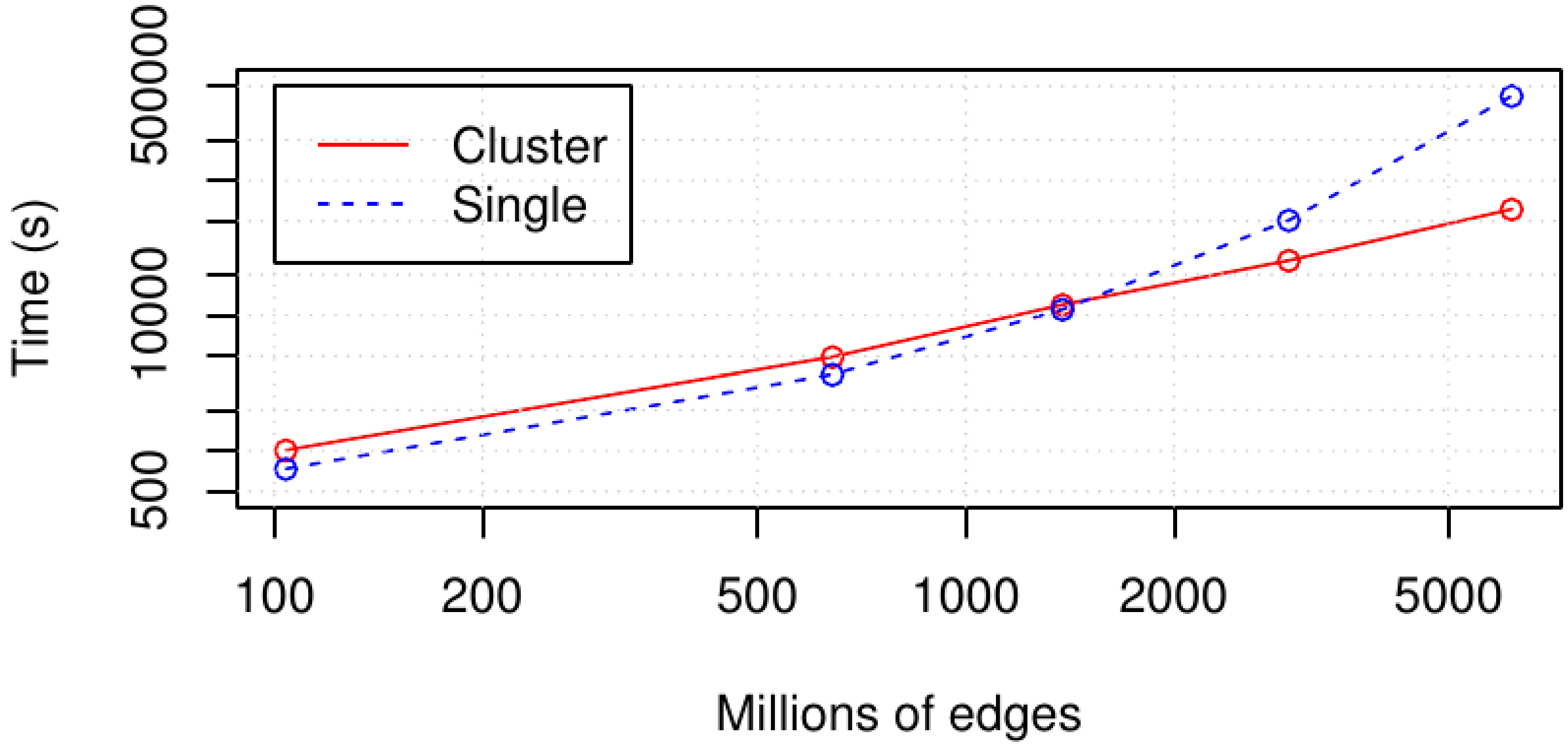
Edge-normalized performance





See paper for more results

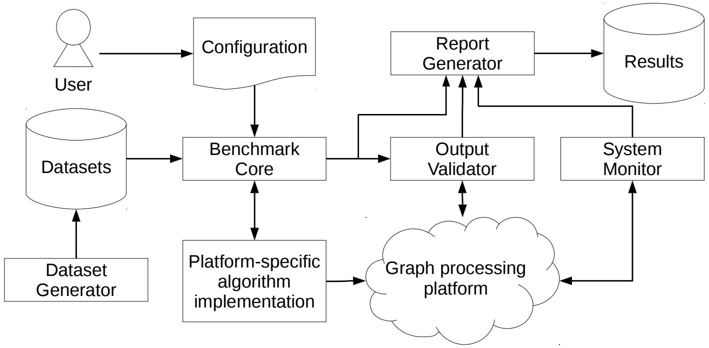
Datagen scalability



Conclusion

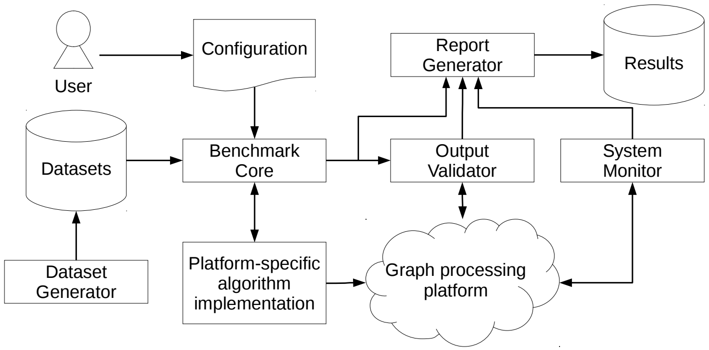
Conclusion

- Use Graphalytics to:
  - Compare
  - Tune
  - Re-design



Conclusion

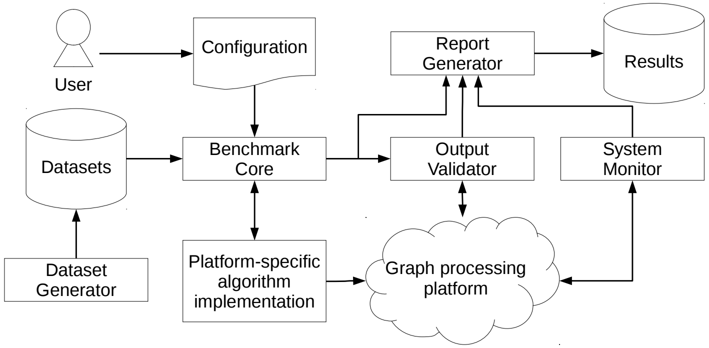
- Use Graphalytics to:
  - Compare
  - Tune
  - Re-design



- Open source (Apache License 2.0)
  - Contribute an implementation for your platform

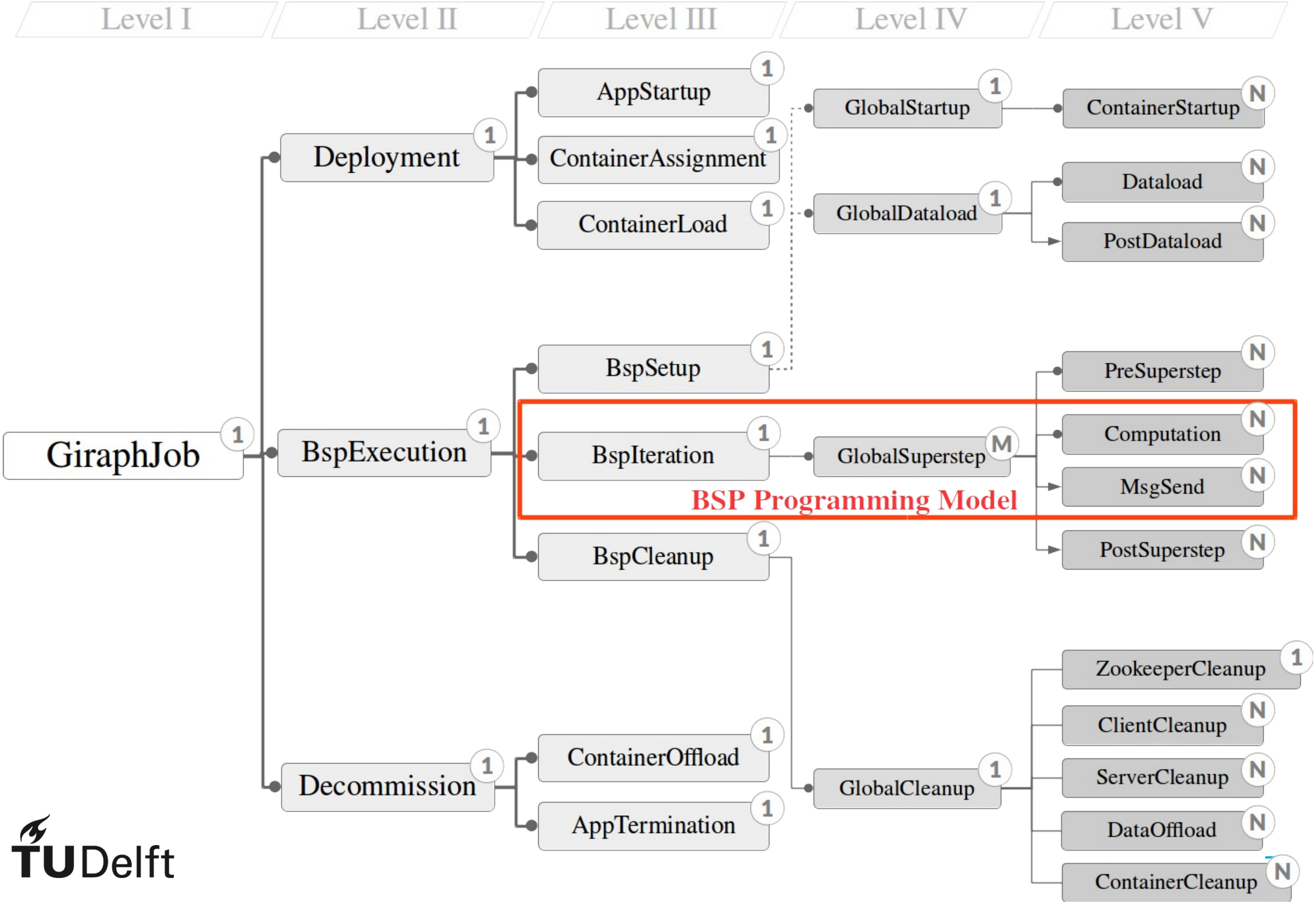
Conclusion

- Use Graphalytics to:
  - Compare
  - Tune
  - Re-design



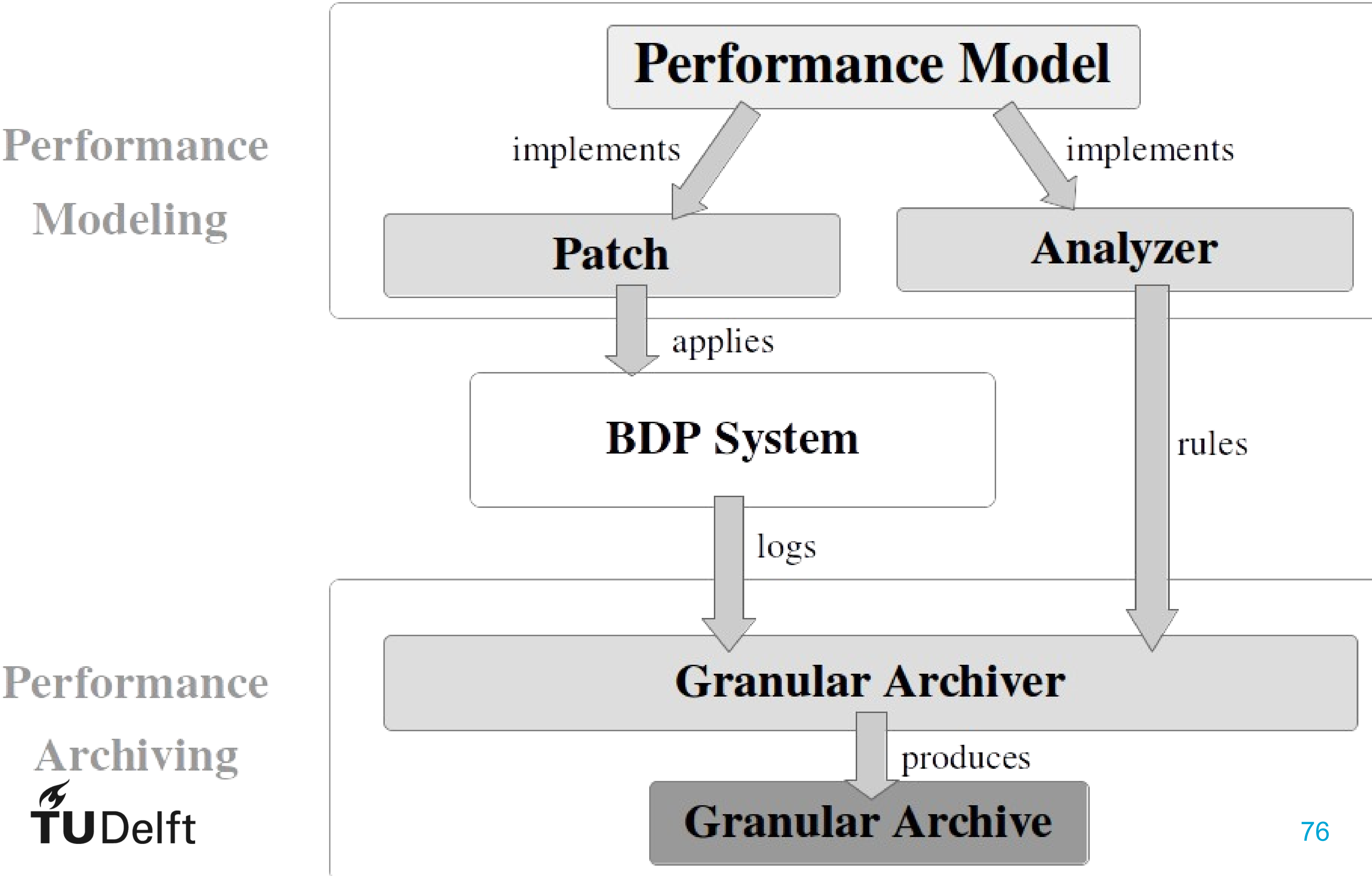
- Open source (Apache License 2.0)
  - Contribute an implementation for your platform
- SPEC RG standardization coming

Performance evaluation – Granular

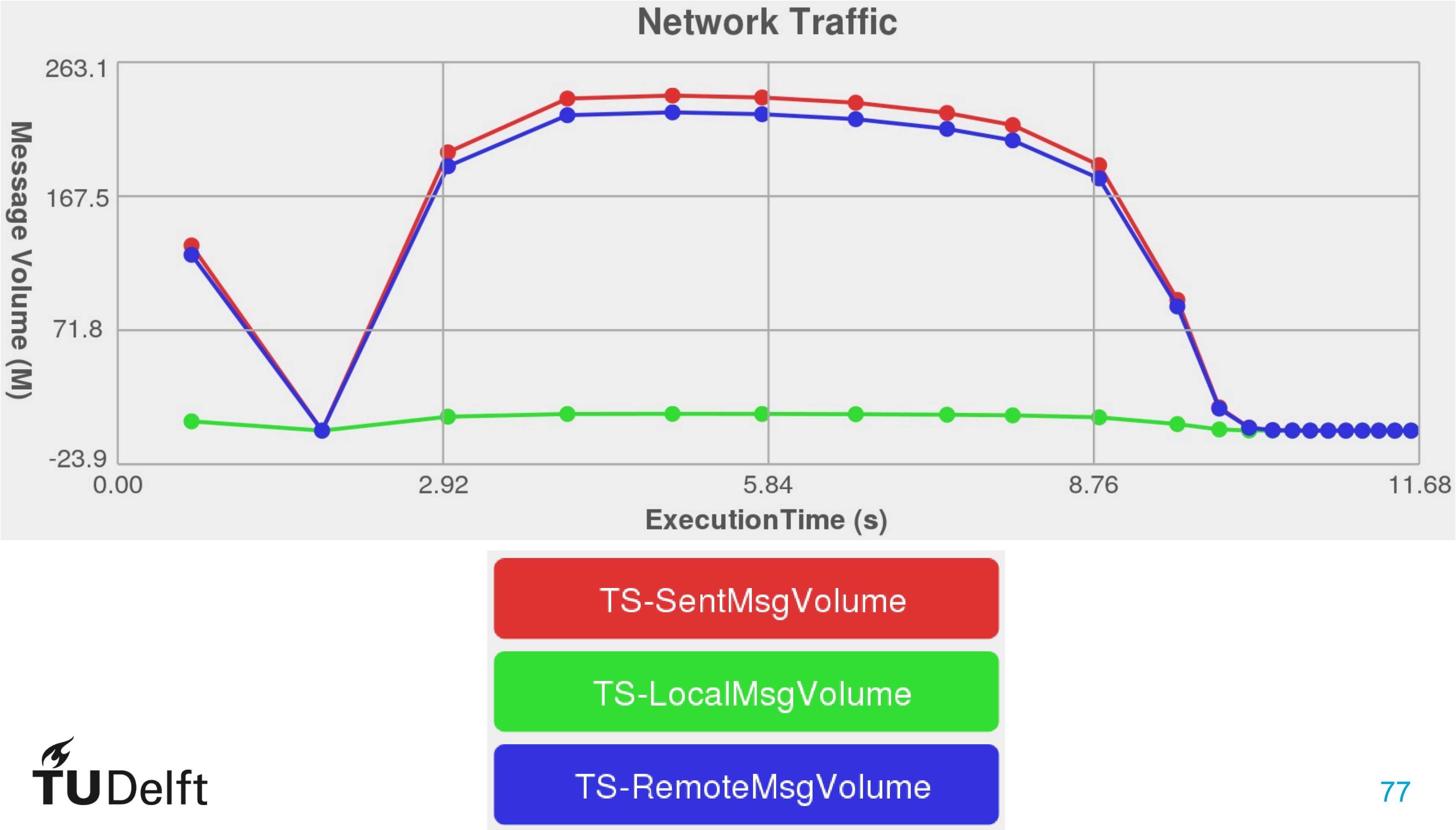


TUDelft

Deploying Granular



Granular results



Thank you!

mihai@mihaic.ro

graphalytics.ewi.tudelft.nl

GitHub

github.com/tudelft-atlarge/graphalytics