

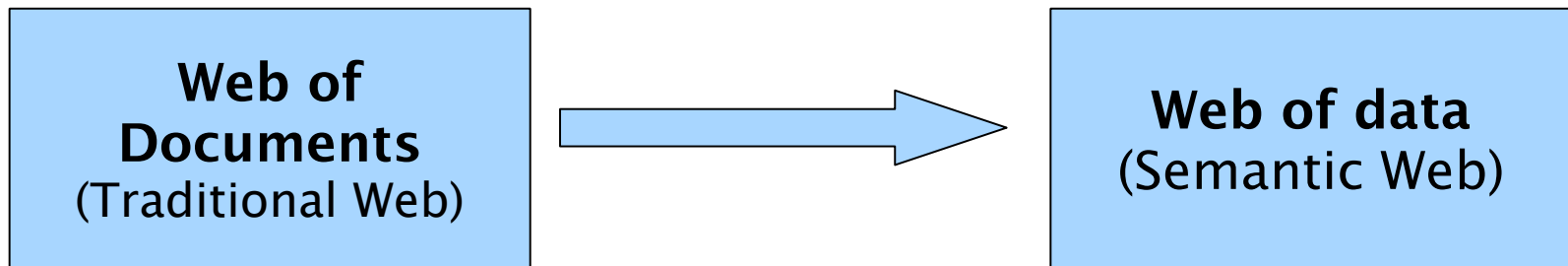
WEBPIE: A WEB-SCALE PARALLEL INFERENCE ENGINE

Jacopo Urbani, Spyros Kotoulas, Jason Maassen, Niels Drost,
Frank Seinstra, Frank van Harmelen, Henri Bal

Department of Computer Science
Vrije Universiteit Amsterdam

The Semantic Web

- The Semantic Web is an extension of the current Web where the semantics is defined
- Basically the idea is to move from



The Semantic Web



- In Semantic Web the data is written in RDF

The Semantic Web

- In Semantic Web the data is written in RDF

```
<http://www.vu.nl> <rdf:type> <http://university.com>
```

The Semantic Web

- In Semantic Web the data is written in RDF

```
<http://www.vu.nl> <rdf:type> <http://university.com>
```

- Machines can apply rules and derive new statements. We call it reasoning

The Semantic Web

- In Semantic Web the data is written in RDF

```
<http://www.vu.nl> <rdf:type> <http://university.com>
```

- Machines can apply rules and derive new statements. We call it reasoning

Input:

```
<Jacopo> <type> <Student>
```

```
<Student> <subclass> <Person>
```

Rule to apply:

if a type B and B subclass C **then** a type C

Output:

```
<Jacopo> <type> <Person>
```

The Semantic Web



Advantages:

- able to combine data from different documents

The Semantic Web

Advantages:

- able to combine data from different documents
- answer to complex queries
 - “list of the first ten peaks in Europe”
 - “find all publications on grid computing since 1995”

The Semantic Web

Advantages:

- able to combine data from different documents
- answer to complex queries
 - “list of the first ten peaks in Europe”
 - “find all publications on grid computing since 1995”
- return also derived information
 - ask for persons → the system returns also the students

Web scale reasoning

- Size of the Semantic Web
 - March 2009: 4.4 Billions
 - Sept. 2009: 7.7 Billions
 - Jan. 2010: 13.1 Billions
 - Now: ?!?
- Input size → need for **parallelization**
- Explosive growth → need for **scalability**

Web scale reasoning

- Size of the Semantic Web
 - March 2009: 4.4 Billion
 - Sept. 2009: 7.7 Billion
 - Jan. 2010: 13.1 Billion
 - Now: ?!?
- Input size → need for **parallelization**
- Explosive growth → need for **scalability**



**Solution:
WebPIE!!**

WebPIE – what is it?

WebPIE is a MapReduce distributed reasoner that works on a Web scale

Features:

- High performance
- High scalability

WebPIE – what is it?

WebPIE is a MapReduce distributed reasoner that works on a Web scale

Features:

- High performance
- High scalability

60 times faster!!!

**Reason on the
entire Semantic Web**

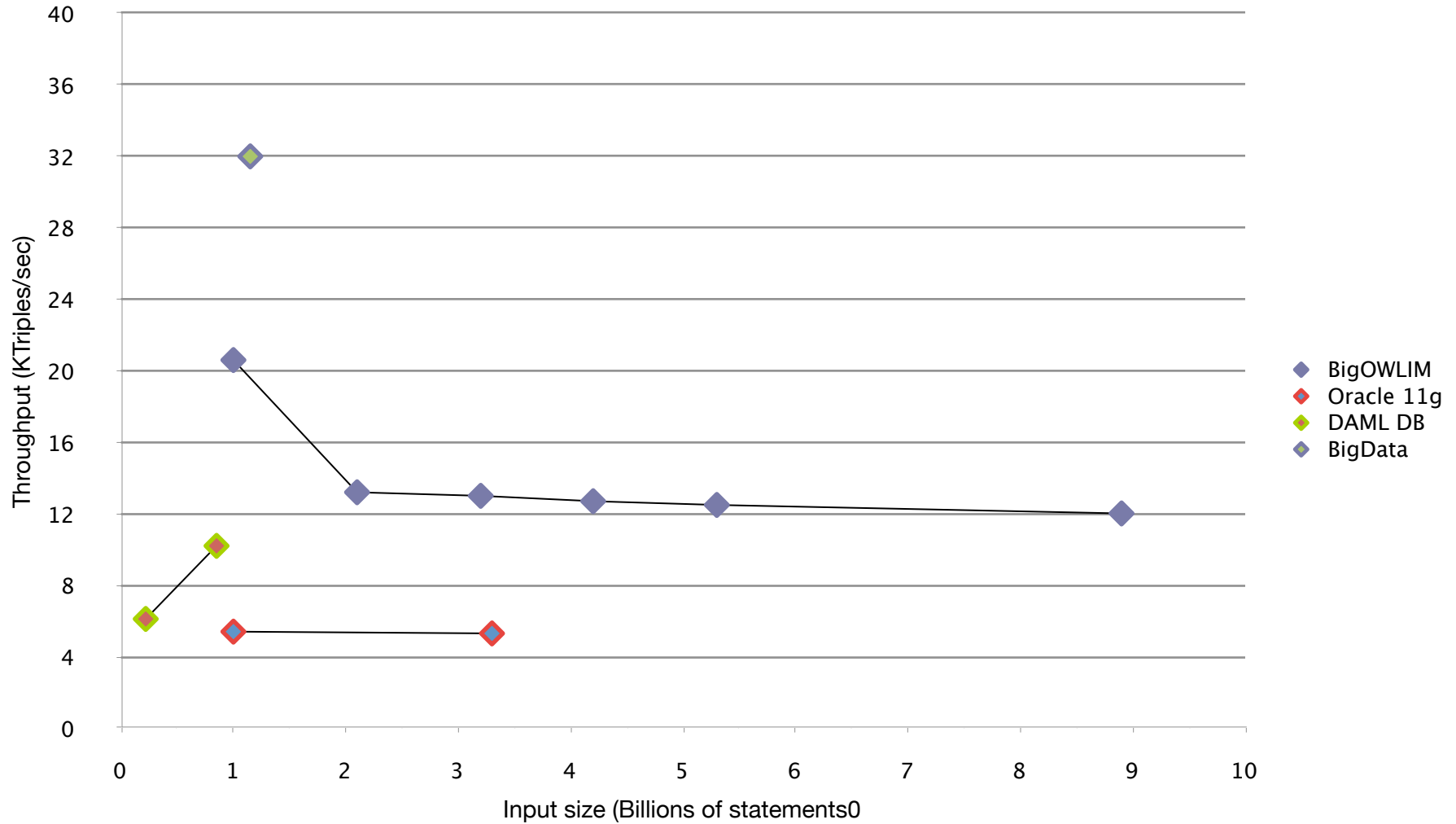
WebPIE – implementation

- Straightforward MapReduce is slower than sequential program
 - Load balancing
 - Duplicate derivations
 - etc.
- WebPIE introduces novel techniques that, combined, solve all the issues
(for details, see the papers)

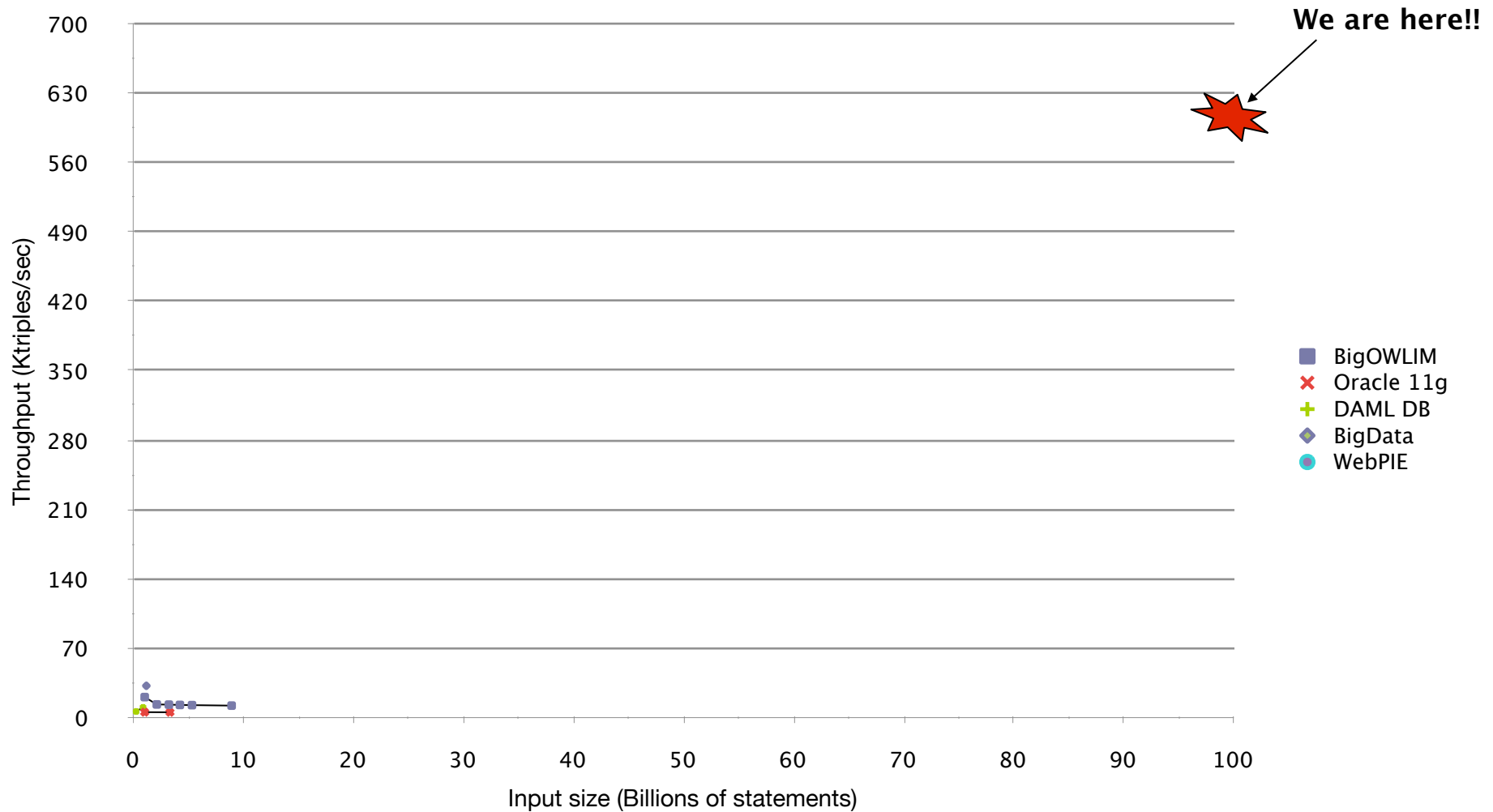
WebPIE – implementation

- Written in Java, uses Hadoop (0.20.2)
- Run on cluster or on the Amazon cloud
- Tests performed at the DAS3 cluster
<http://www.cs.vu.nl/das3>
- Code, tutorial, etc. available at
<http://www.cs.vu.nl/webpie>

Performance



Performance



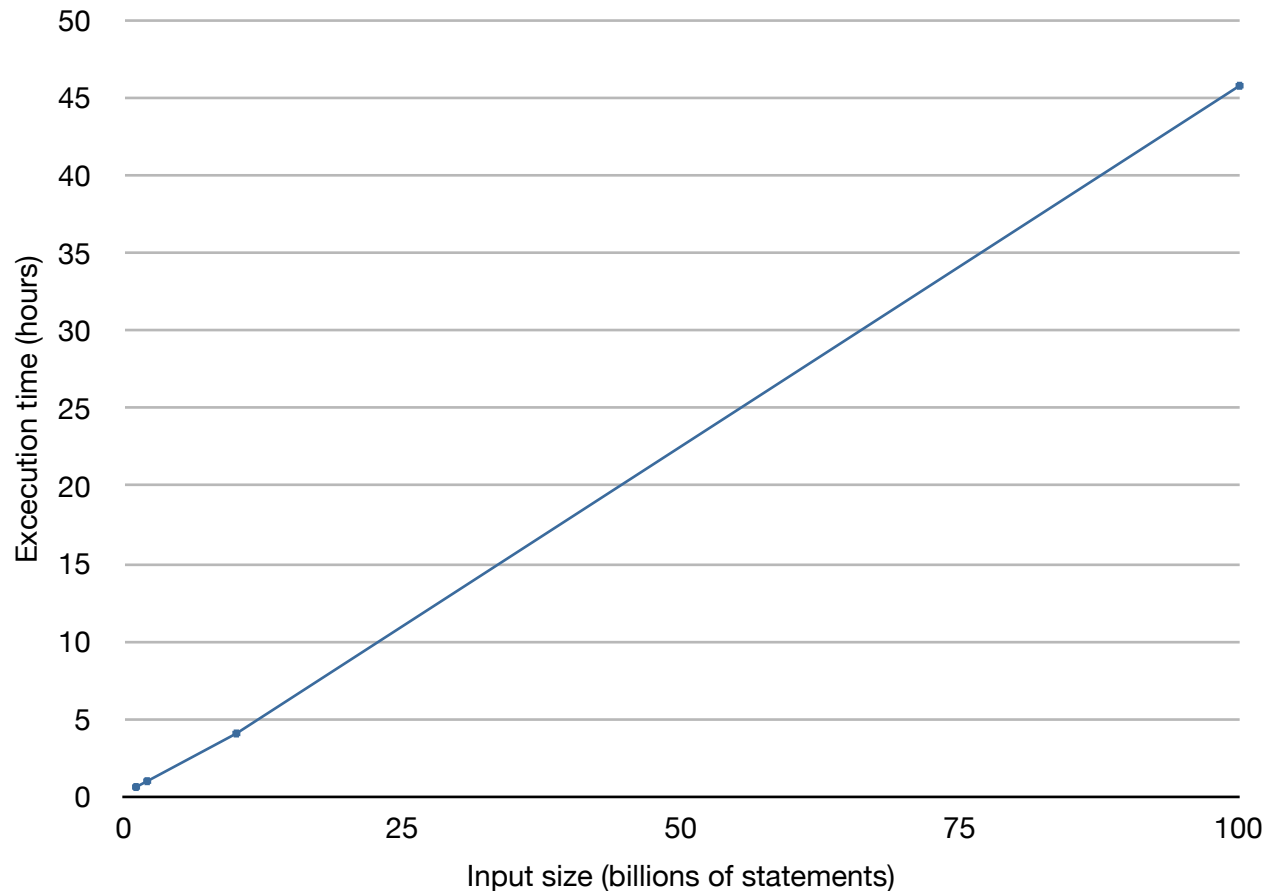
Performance

- Tested on different datasets
 - different input size
 - different input complexity
- In all cases the performance is better than best technique

Dataset	Input size	Output size	Exec. time
LUBM	1 Billion	0.5 Billion	0.6 hours
Uniprot	1.5 Billion	2.0 Billion	6.1 hours
LDSR	0.9 Billion	0.9 Billion	3.5 hours

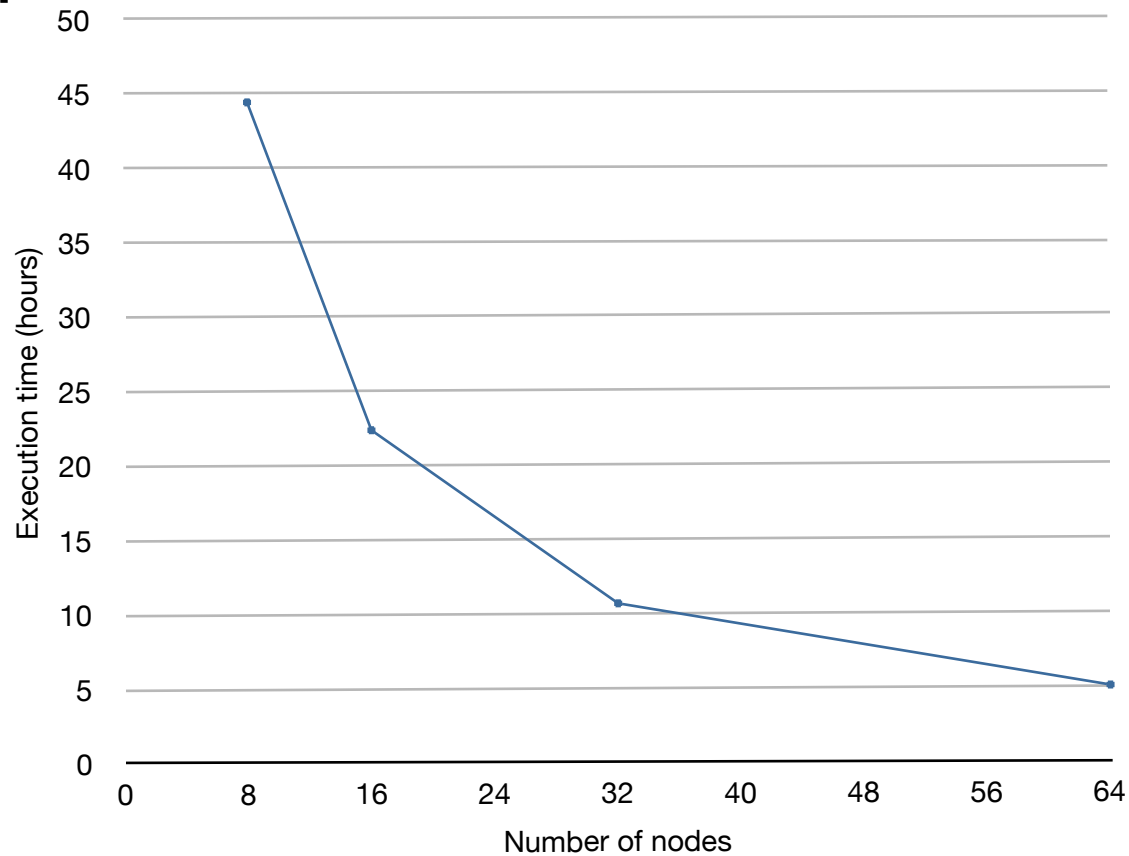
Scalability

- Scalability on the input size using 64 nodes



Scalability

- Scalability on the number of nodes
(input: 10 billions statements)



Conclusions

- Vastly outperforms current state of the art
 - one order of magnitude input size
 - Throughput between 5 and 60 times higher

“WebPIE makes reasoning over the entire Semantic Web possible”

Demo

