MaRVIN: a distributed platform for massive RDF processing

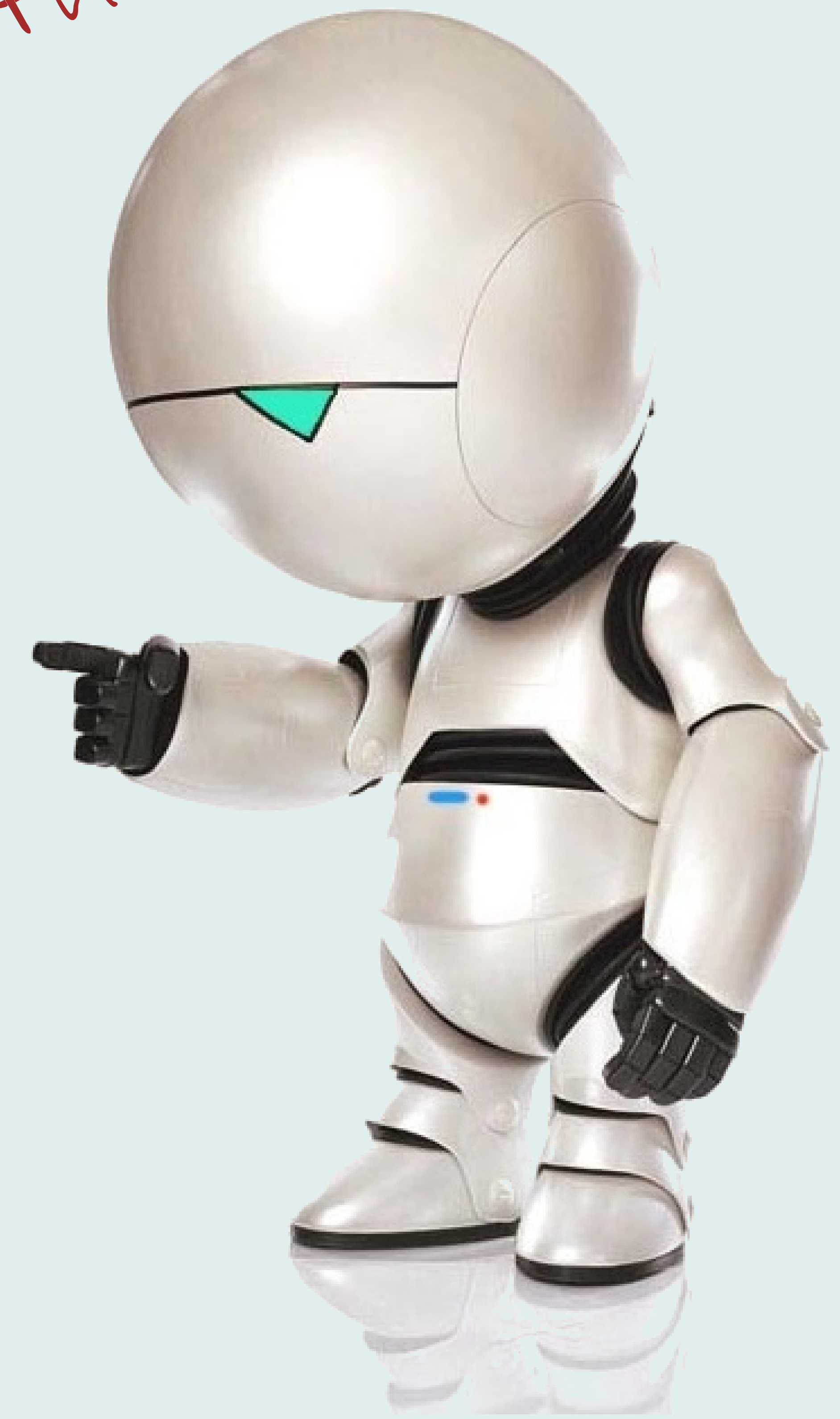
George Anadiotis, Spyros Kotoulas, Eyal Oren, Ronny Siebes, Frank van Harmelen
Niels Drost, Roelof Kemp, Jason Maassen, Frank J. Seinstra, Henri E. Bal
Vrije Universiteit Amsterdam

MaRVIN is:
a platform for processing lots of RDF data
(now: computing RDFS/OWL closure)

MaRVIN scales by:
distributing computation over many nodes
approximate (sound but incomplete) reasoning
anytime convergence (more complete over time)

MaRVIN runs on:
in principle: any grid, using Ibis middleware
currently: the DAS-3 distributed supercomputer (300 nodes)
soon: a wide-area a peer-to-peer network

"a brain the size of a planet"



Main loop: divide-conquer-and-swap

1. *divide*: split input data in chunks
2. *conquer*: each node:

   reads some chunks,

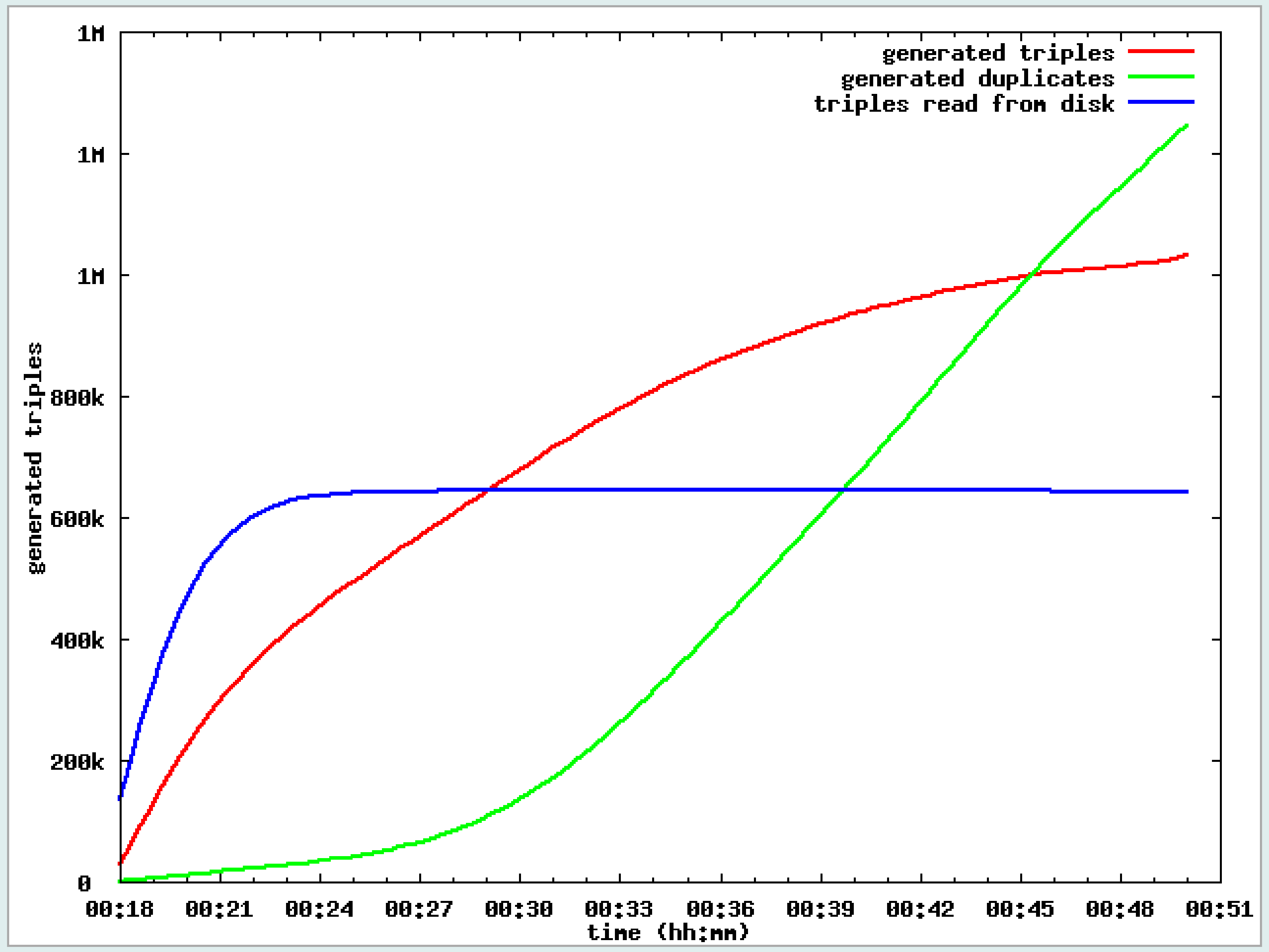
   DO: computes closure.
3. *swap*: each node:

   removes all triples:

   sends some to central storage,

   sends other to some peer

repeat 2-3 ad infinitum



anytime incremental results

Currently:

- running on DAS-3, a five-cluster grid system
- max. 271 machines, 791 cores (2.4Ghz, 4Gb RAM)
- suffering from growing pains
- reading data @100ktps/min/node (1B in 1hr, on 100 nodes)
- producing data @15-25ktps/min/node

Closure on the dataset (computed) just one example:
Infrastructure for experiments over massive RDF data

Questions:

- network overhead vs benefit?
- scalability (nodes and data)?
- output quality (anytime behaviour)?
- routing policy?

- modular architecture:
  - change initial data distribution
  - change functionality of node
  - change routing policy
  - ...
- real-time logging, visualisation, analysis

EXPERIMENT AND EVALUATE
A TOOL FOR THE RESEARCH COMMUNITY

MaRVIN: a distributed platform for massive RDF processing

contact: Eyal Oren
http://larkc.eu/marvin

vrije Universiteit amsterdam