Middleware for Big Data processing: Test results

I. Gankevich V. Gaiduchok Yu. Tipikin
V. Korkhov A. Degtyarev A. Bogdanov

NEC’15

Plan

Theorising

Our approach

Test results

Philosophising

Plan

Theorising

Our approach

Test results

Philosophising

Definition

The data is considered *big*, if its pre- and post-processing[1] time is much larger than processing time.

Big data does not always mean big volume.

- Tightly-coupled data is big.
- High-volume data is big.
- Semi-structured data is big.

Edge cases.

- OpenFOAM: $t_{pre} + t_{post} \approx t_{proc}$ (not so big data)

[1] general I/O, decompressing, decoding, filtering etc.

Data metrics

Approach: Try to be as close as possible to the edge case (i.e. decrease pre/post time).

$$t_{read} \xrightarrow{\text{no. of replicas} \to \infty} 0,$$
$$t_{write} \xrightarrow{\text{no. of chunks} \to \infty} 0.$$

No. of replicas/chunks:

- Capped by physical constraints (total no. of nodes, max. no. of nodes per job etc.)
- Should be per-dataset configurable.
- Should be dynamic.

Resilience to failures

availability

partition tolerance

consistency

Resilience to failures

availability

partition tolerance

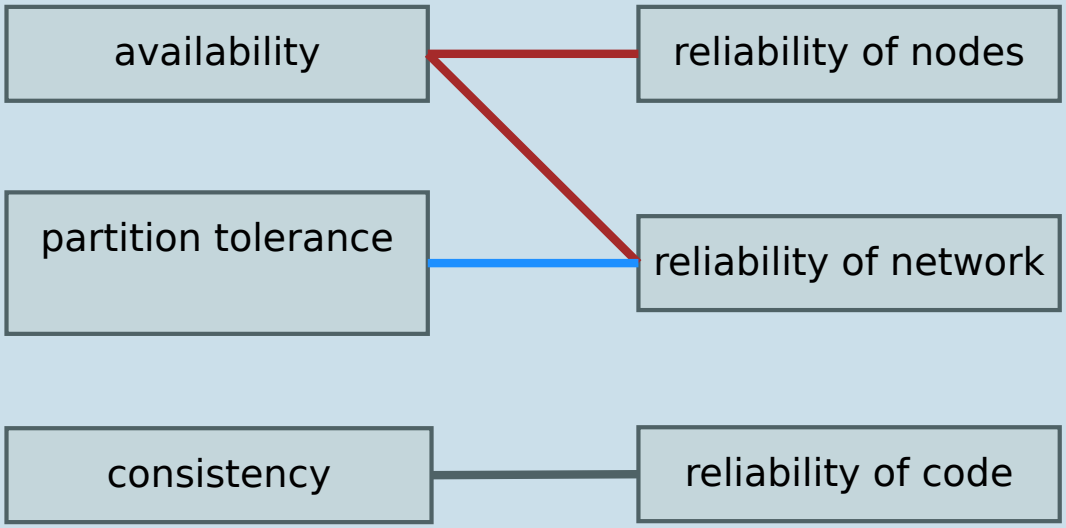
consistency

reliability of nodes

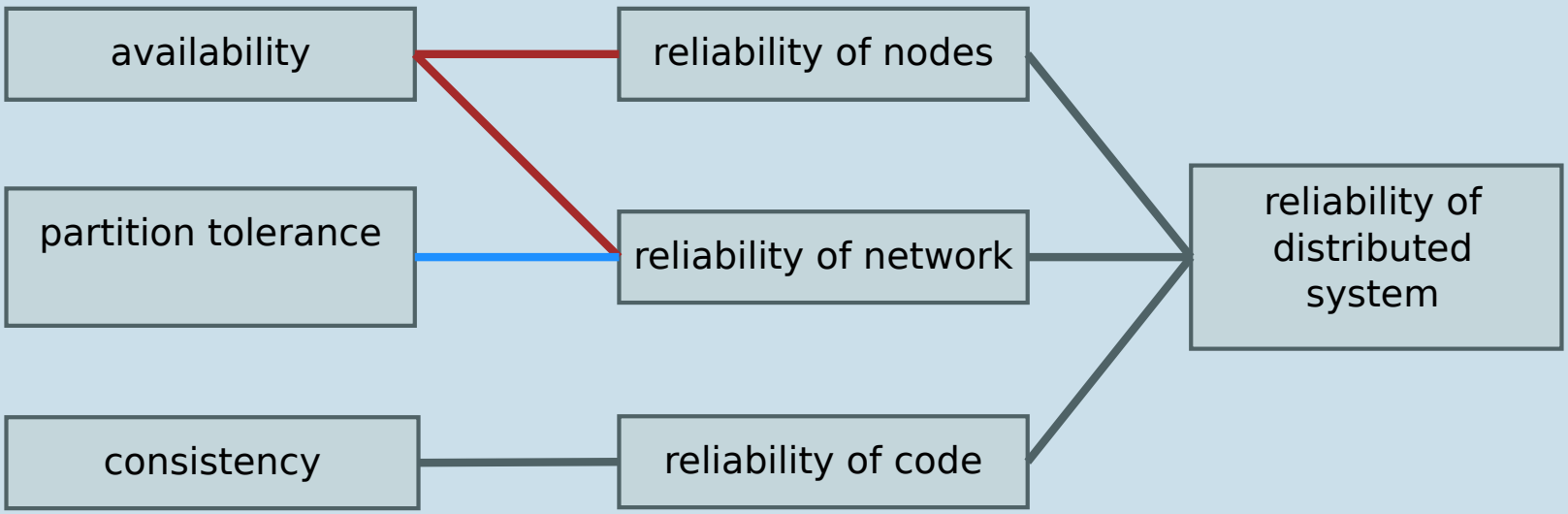
reliability of network

reliability of code

Resilience to failures



Resilience to failures



Resilience to failures

Reliability	How to improve	How to measure
of nodes	replication	$1 - \alpha^n$
of network	hierarchy	$1 - \beta^m$
of code	dist. transactions	$1 - \alpha n' - \beta^{m'}$

α — probability of a node failure.
β — probability of a network link failure.
n — amount of redundant nodes, $n \geq 1$.
m — amount of redundant network links, $m \geq 1$.
m', n' — no. of nodes and net. links participating in a distributed transaction.

Plan

Theorising

Our approach

Test results

Philosophising

The API

Everything is a micro-kernel — a unit of work which *always* binds to the compute node where the data is stored.

- Homogeneous API (no message objects).
- Micro-kernels can communicate locally when sent to the same compute node.
- When nodes with all replicas are full, new replicas may be created.
- Create subordinate kernels to parallelise the programme.
- Event-driven design: callbacks for data processing, collecting results from subordinates, reading/writing.

The implementation

A lightweight Linux service.

- Portable C++ programme (8130 SLOC).
- Basically a scheduler that allows applications to interact with a whole cluster via C++ API.
- An application for determining where file replicas are stored.
- An application for auto-discovery and building virtual tree of healthy nodes [1].
- An application for exposing basic web interface (not finished).

[1] I. Gankevich, Yu. Tipikin, and V. Gaiduchok. 2015. Subordination: Cluster management without distributed consensus. In *International Conference on High Performance Computing Simulation (HPCS)*. 639–642.

Plan

Theorising

Our approach

Test results

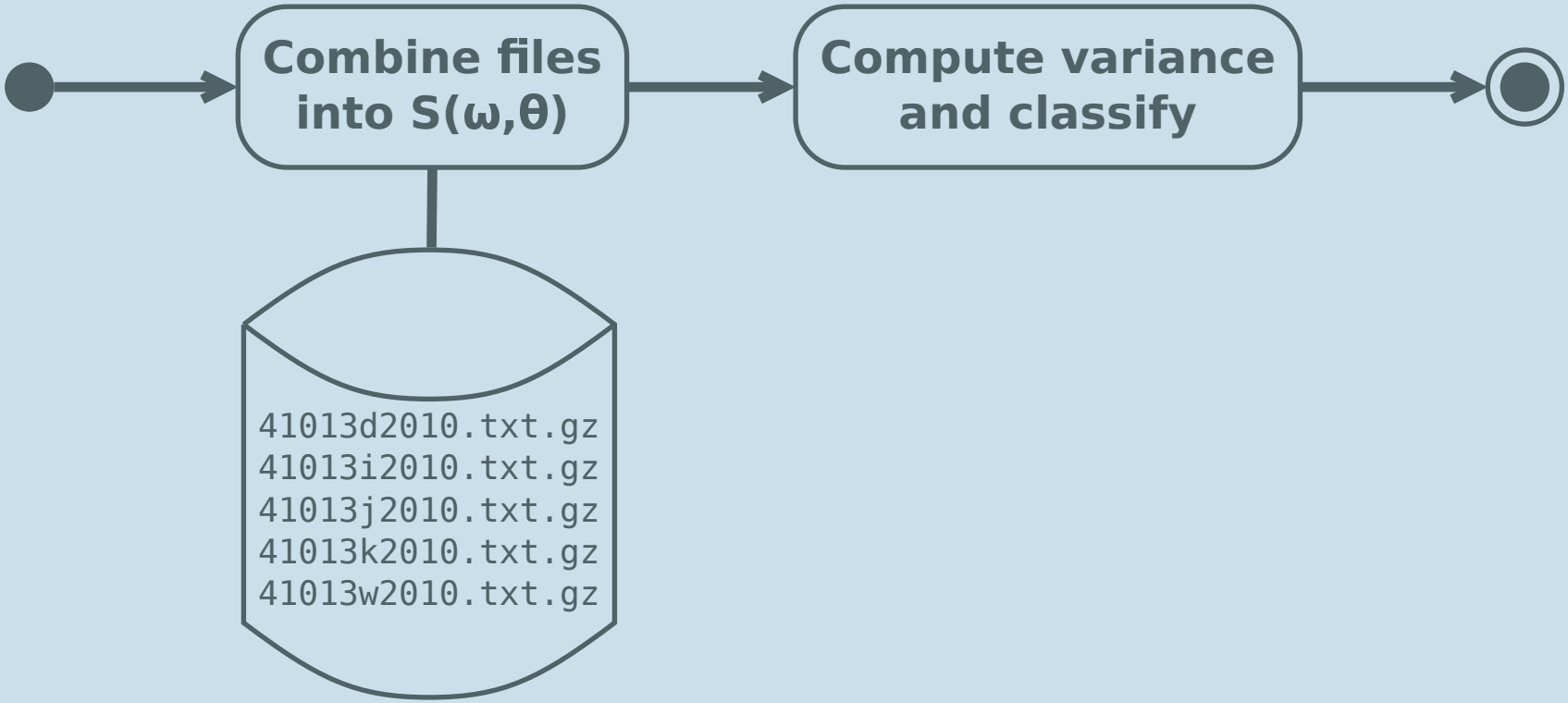
Philosophising

NDBC dataset

Dataset size	144MB
Dataset size (uncompressed)	770MB
Number of wave stations	24
Time span	3 years (2010–2012)
Total number of spectra	445422

- High 1:5 compression ratio,
- the spectra are stored using 5 variables,
- text-based file format.

The algorithm



Reconstruction formula:

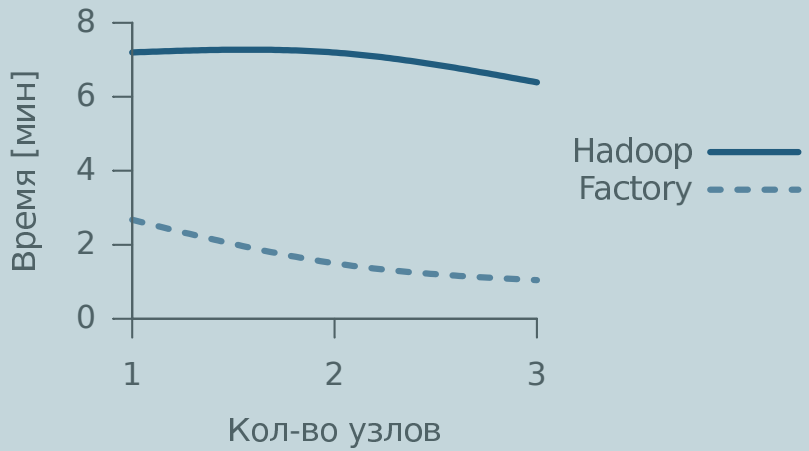
$$S(\omega,\theta) = \frac{1}{\pi}\left[\frac{1}{2} + r_1\cos(\theta-\alpha_1) + r_2\cos(2(\theta-\alpha_2))\right] S_0(\omega).$$

Comparison to Hadoop

Setup

Hadoop version	2.3.0
Hadoop nodes	3
RAM (GB)	4
CPU	Intel Q9650
No. of cores	4
Core freq. (GHz)	3.0
OS	Debian 7.5

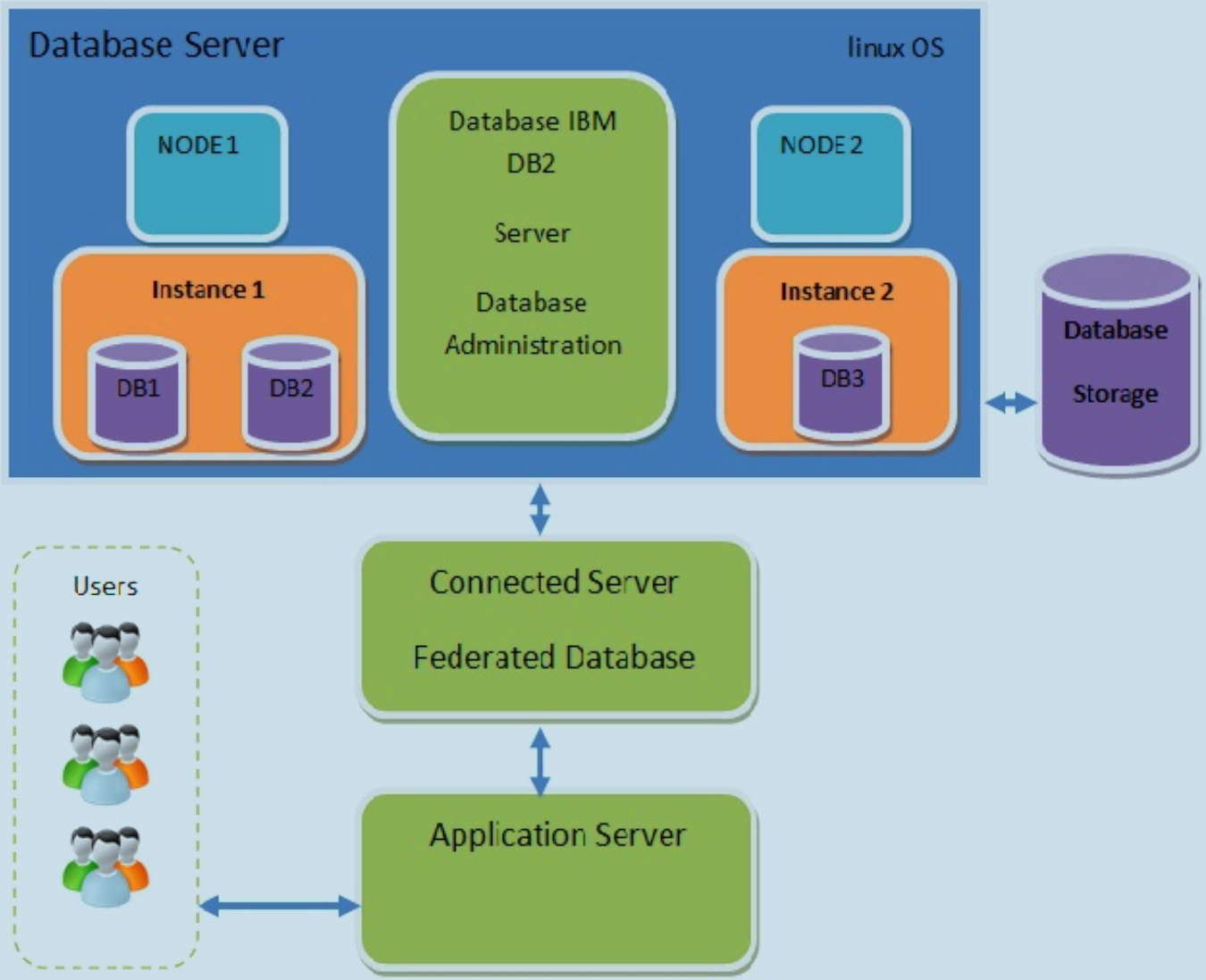
Performance



Hadoop $\approx$ 1000 spec./sec.
Factory $\approx$ 7000 spec./sec.

Reference: I. Gankevich, A. Degtyarev Efficient processing and classification of wave energy spectrum data with a distributed pipeline. *Computer Research and Modeling* 7, 3 (2015), 517–520.

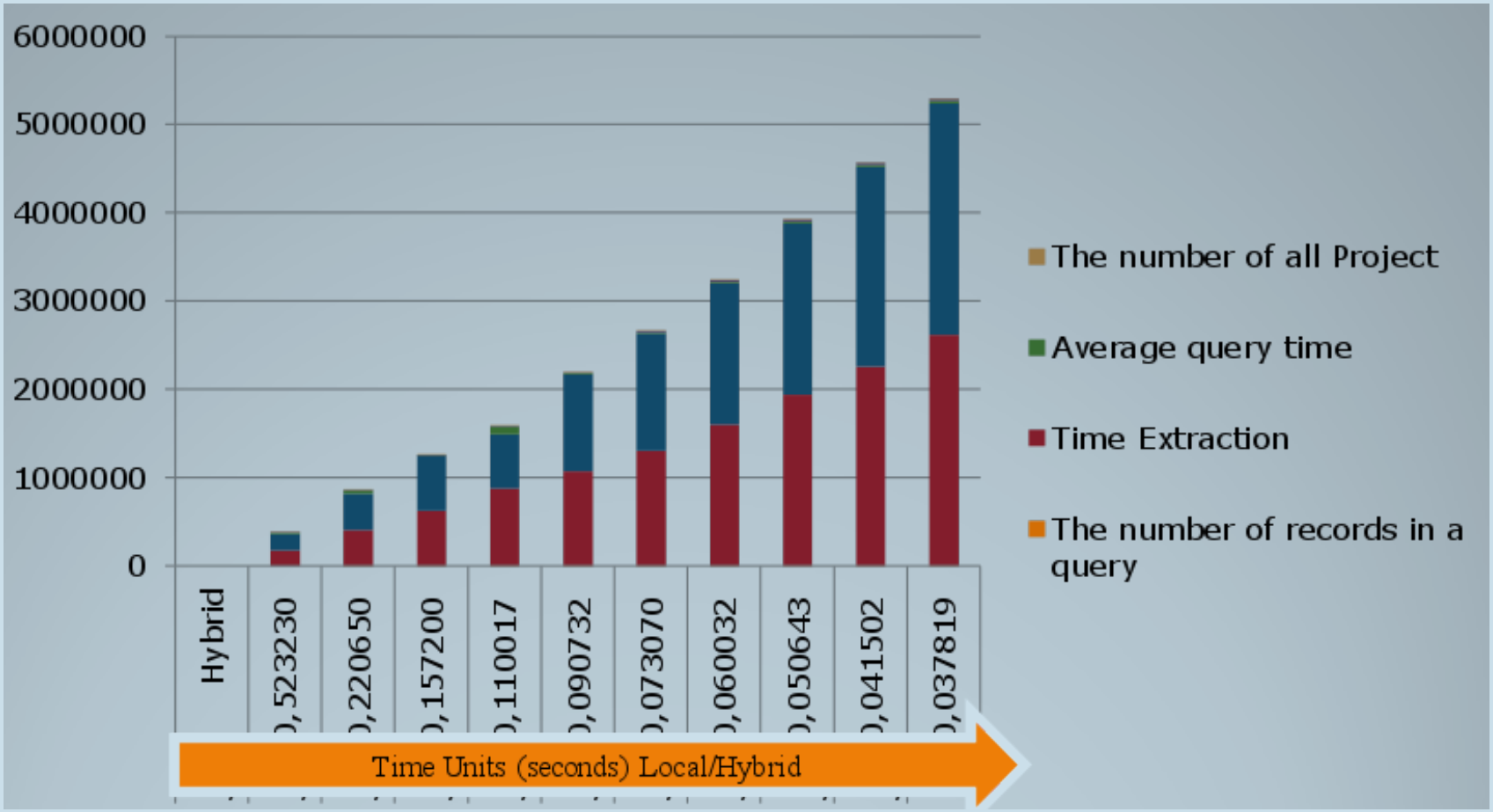
Database test setup



Database test results

Record of s.		# of	Extraction time		Avg. query time		# of
local	hybrid	rec.	local	hybrid	local	hybrid	proj.
0.53	0.52	1	178	183	1.80	1.83	100
0.23	0.22	2	406	412	4.09	4.12	200
0.14	0.16	3	622	627	6.24	6.27	300
0.10	0.11	4	881	884	8.82	8.84	400
0.09	0.09	5	1075	1099	10.96	10.99	500
0.07	0.07	6	1306	1330	13.06	13.08	600
0.06	0.06	7	1600	1604	16.01	16.04	700
0.05	0.05	8	1941	1944	19.41	19.44	800
0.04	0.04	9	2263	2265	22.62	22.65	900
0.03	0.03	10	2620	2622	26.20	26.22	1000

Database test results



Plan

Theorising

Our approach

Test results

Philosophising

Scott McNealy:

Network is a computer.

Scott McNealy:
Network is a computer.

Anonymous:
Then data is network traffic!

Thank you for attention!