The 8th International Conference "Distributed Computing and GRID-Technologies in Science and Education"

Building Up Intelligible Parallel Computing World

Prof. Vladimir Voevodin

Deputy Director, Research Computing Center, MSU

Head of Department on Supercomputers and Quantum Informatics, CMC, MSU





MSU Petascale Facilities: Supercomputers Lomonosov-2 and Lomonosov







MSU Supercomputing Center today:

Users: 2955 Projects: 880

MSU Faculties / Institutes : 21
Institutes of RAS : 95
Russian Universities: 102

Supercomputer Technologies
Are Everywhere...

MSU Supercomputer Platforms







Technological background:
Intel Xeon 4/6/10/12... cores
SMP node on 128 cores
Intel Xeon Phi (KNL)
NVIDIA 2070 / 2090 / K40 / P100
IBM Power 8 / IBM Blue Gene/P
Memory per node: from 12GB up to 2TB

Diversity is great...

Existing methodologies to compare computing platforms (Top500, Graph500, HPCG)



High Performance Linpack
Benchmark
Top500.org



BFS & SSSP Graph Benchmarks Graph500.org



High Performance Conjugate Gradients
Benchmark
hpcg-benchmark.org



Well-known theoretical potential of algorithms



Well-described and available community experience

Existing methodologies to compare computing platforms (Top500, Graph500, HPCG)



High Performance Linpack
Benchmark
Top500.org



BFS & SSSP Graph Benchmarks Graph500.org



High Performance Conjugate Gradients
Benchmark
hpcg-benchmark.org



Well-known theoretical potential of algorithms



Well-described and available community experience

General methodology to compare computing platforms (using any algorithm)





Well-known theoretical potential of algorithms



Well-described and available community experience

Well-known theoretical potential of algorithms

How can we describe theoretical potential and implementation details of any algorithm?

What is a description of an algorithm?

Description of algorithms

(What properties of algorithms should be included in the description?)

For positive definite Hermitian matrices (symmetric matrices in the real case), we use the decomposition $A = LL^*$, where L is the lower triangular matrix G General Description ion $A = U^*U$, where U is the upper triangular matrix G. These forms of the Cholesky decomposition are equivalent in the sense of the amount of arithmetic operations and are different in the sense of data representation.

implementation of formulae obtained uniquely for the elements of the matrix I, from

Input data: a symmetric positive definite matrix A whose elements are denoted by a_{ij}). features.

Output data: the lower triangular matrix L whose elements are denoted by l_{ij}).

The Cholesky algorithm can be represented in the for

$$l_{11} = \sqrt{a_{11}}$$
, Mathematical Description

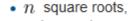
$$l_{j1} = \frac{a_{j1}}{l_{11}}, \quad j \in [2, n],$$

$$l_{ii} = \sqrt{a_{ii} - \sum_{p=1}^{i-1} l_{ip}^2}, \quad i \in [2, n],$$

$$\frac{2}{6} \text{ multiplications and } \frac{n^3 - n}{6} \text{ additions (subtractions): the main}$$

$$l_{ji} = \left(a_{ji} - \sum_{p=1}^{i-1} l_{ip} l_{jp}\right) / l_{ii}, \quad i \in [2, n-1], j \in [i+1, n].$$

The following number of operations should be perf matrix of order η , using a serial version of the Chd



•
$$\frac{n(n-1)}{2}$$
 divisiona,

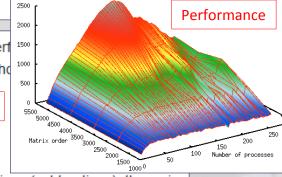
Serial Complexity

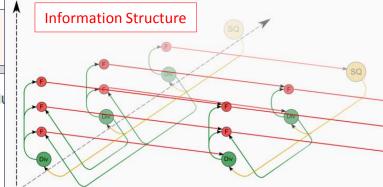
amount of computational work

A computational kernel of its serial version can be composed of $\frac{n(n-1)}{2}$ dot produ

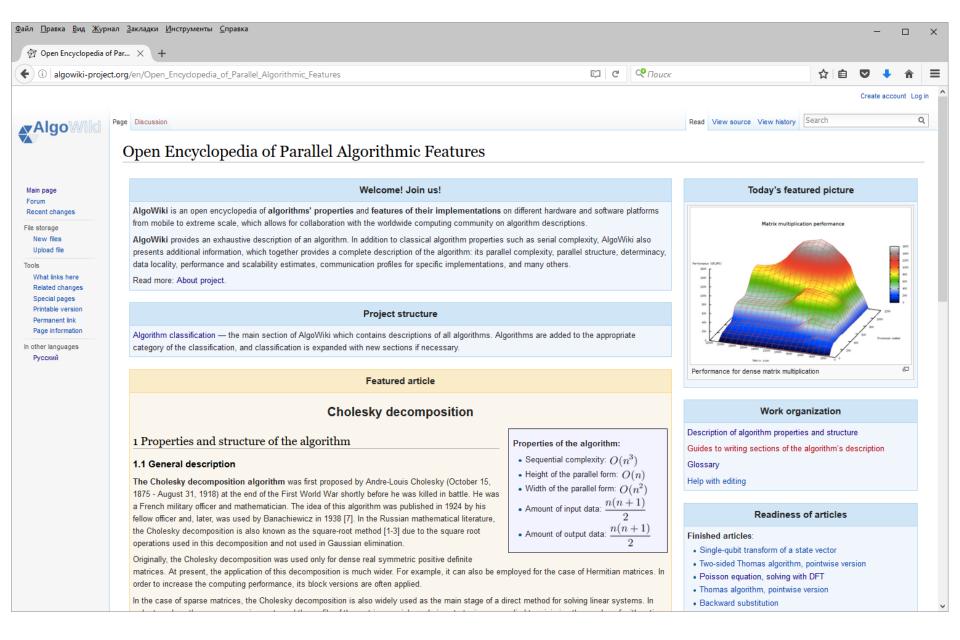
$$\sum_{p=1}^{i-1} l_{ip} l_{jp}.$$

Computational Kernel



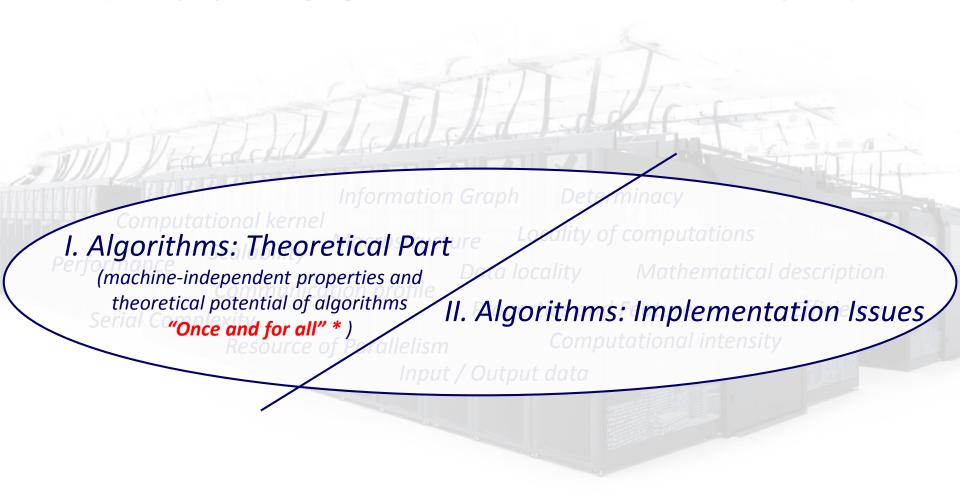






Description of algorithms

(What properties of algorithms should be included in the description?)



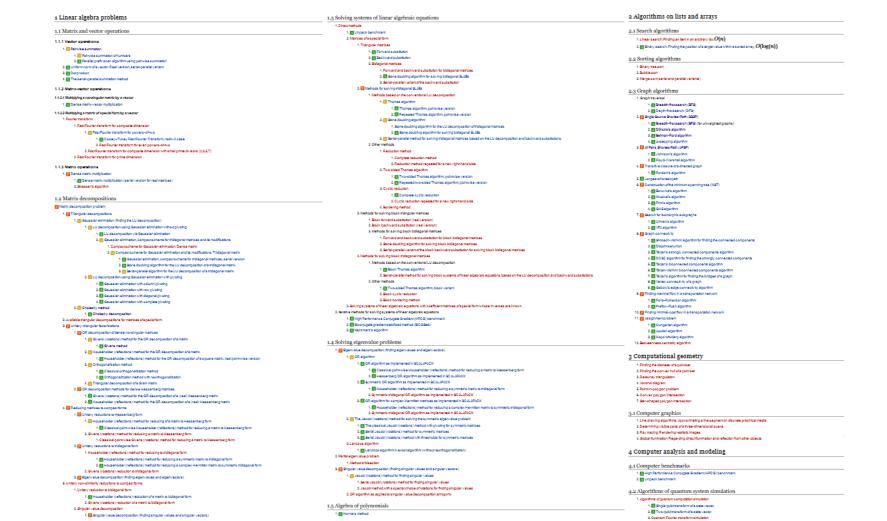
^{*} Changes in architectures require changes in implementations, not in algorithms!

Unified description of algorithms in AlgoWiki (predefined unified structure)

- 1 Properties and structure of the algorithm
 - 1.1 General description of the algorithm
 - 1.2 Mathematical description of the algorithm
 - 1.3 Computational kernel of the algorithm
 - 1.4 Macro structure of the algorithm
 - 1.5 Implementation scheme of the serial algorithm
 - 1.6 Serial complexity of the algorithm
 - 1.7 Information graph
 - 1.8 Parallelization resource of the algorithm
 - 1.9 Input and output data of the algorithm
 - 1.10 Properties of the algorithm
- 2 Software implementation of the algorithm
 - 2.1 Implementation peculiarities of the serial algorithm
 - 2.2 Locality of data and computations
 - 2.3 Possible methods and considerations for parallel implementation of the algorithm
 - 2.4 Scalability of the algorithm and its implementations
 - 2.5 Dynamic characteristics and efficiency of the algorithm implementation
 - 2.6 Conclusions for different classes of computer architecture
 - 2.7 Existing implementations of the algorithm
- 3 References

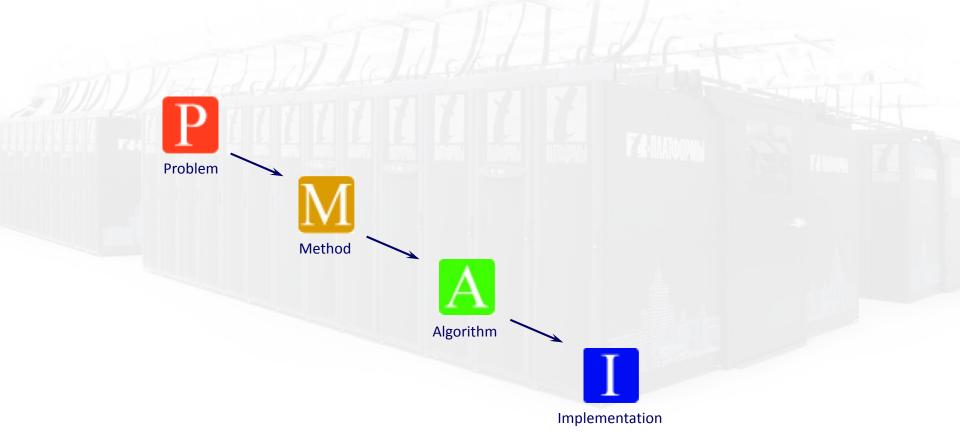


AlgoWiki: algorithm classification (tree structure)

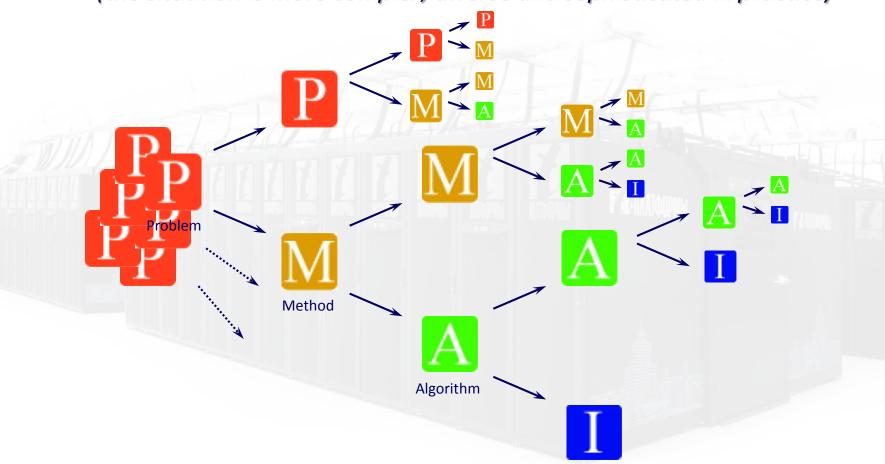


AlgoWiki: Problem - Method - Algorithm - Implementation

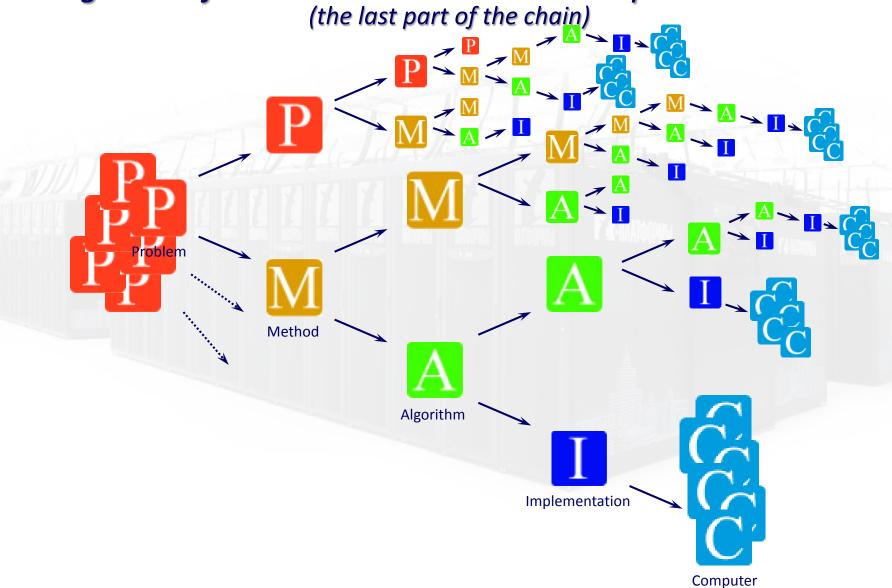
(What do we have for each algorithm in AlgoWiki? An exhaustive description of the chain)

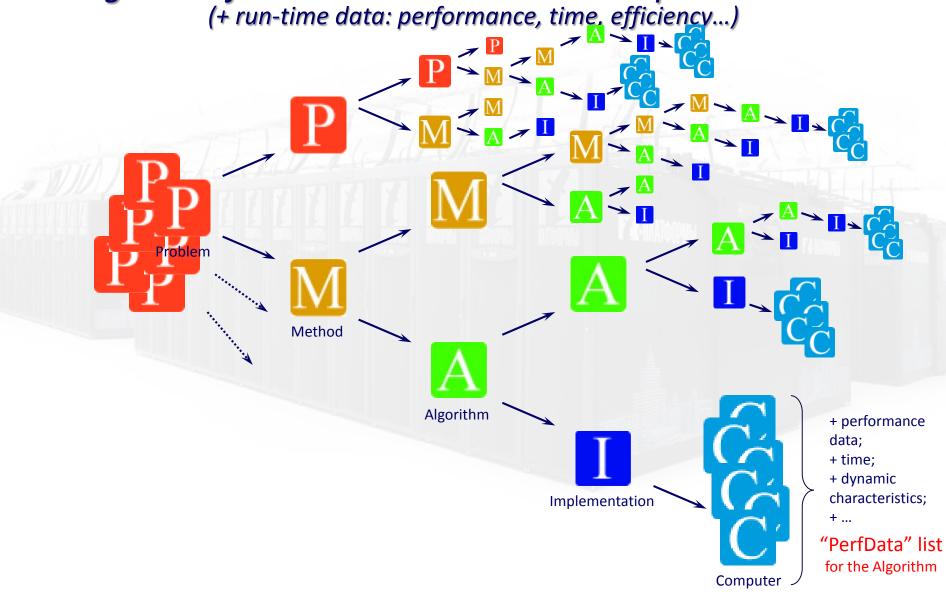


AlgoWiki: Problem – Method – Algorithm – Implementation (the situation is more complex, diverse and sophisticated in practice)



Implementation





(+ run-time data: performance, time, efficiency...) **Problem** Method Algorithm + performance data; + time; + dynamic characteristics; Implementation "PerfData" list for the Algorithm Computer

(+ run-time data: performance, time, efficiency... = PerfData) **Problem** Method Algorithm + performance data; + time; + dynamic characteristics; Implementation + ... "PerfData" list

for the Algorithm

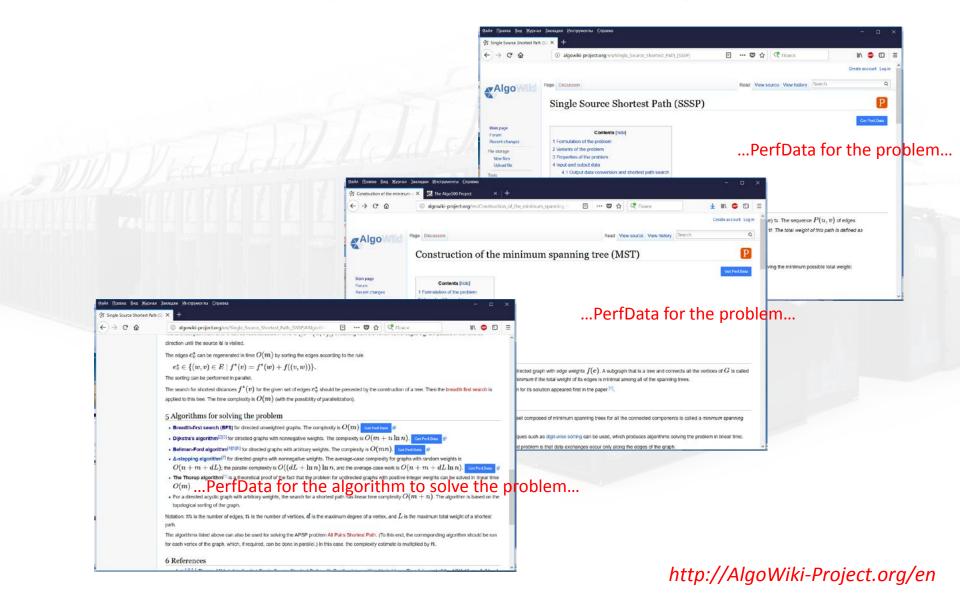
Computer

How Does It Work?

What Can We Obtain from AlgoWiki?



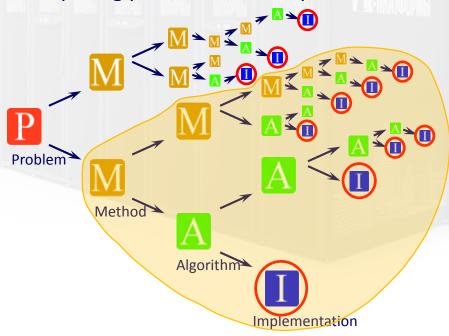
Retrieve Performance Data from AlgoWiki...



"PerfData" list on "Strongly Connected Components" (Method = Forward-Backward)



- compare different algorithms, implementations and computing platforms for the problem and method;

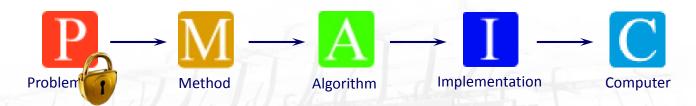


"PerfData" list on "Strongly Connected Components" (Method = Forward-Backward)

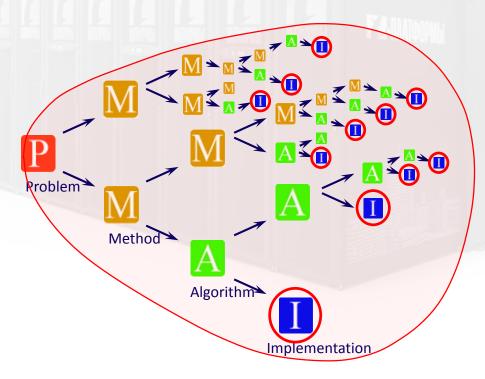
Rating	Method	Implementation	Platform	MTEPS	GraphType	GraphSize
1	Forward-Backward	RCC for GPU	Lomonosov-2 (P100)	634,00	RMAT	2^20
2	Forward-Backward	RCC for GPU	Lomonosov-2 (P100)	620,00	RMAT	2^21
3	Forward-Backward	RCC for CPU	Lomonosov-2	564,00	RMAT	2^24
4	Forward-Backward	RCC for GPU	Lomonosov-2 (P100)	544,00	RMAT	2^22
5	Forward-Backward	RCC for GPU	Lomonosov-2 (P100)	528,00	RMAT	2^23
6	Forward-Backward	RCC for CPU	Lomonosov-2	498,00	RMAT	2^26
7	Forward-Backward	RCC for CPU	Lomonosov-2	497,00	RMAT	2^25
8	Forward-Backward	RCC for CPU	Lomonosov-2	486,00	RMAT	2^27
9	Forward-Backward	RCC for GPU	Lomonosov-2 (P100)	456,00	RMAT	2^25
10	Forward-Backward	RCC for GPU	Lomonosov-2 (P100)	453,00	RMAT	2^24
11	Forward-Backward	RCC for CPU	Lomonosov-2	452,00	RMAT	2^22
12	Forward-Backward	RCC for CPU	Lomonosov-2	440,24	SSCA-2	2^21
13	Forward-Backward	RCC for CPU	Lomonosov-2	432,00	RMAT	2^23
14	Forward-Backward	RCC for CPU	Lomonosov-2	426,00	RMAT	2^21
15	Forward-Backward	RCC for GPU	Lomonosov-2 (P100)	426,00	RMAT	2^26
16	Forward-Backward	RCC for CPU	Lomonosov-2	418,00	RMAT	2^20
17	Forward-Backward	PBGL MPI	IBM BlueGene/P	232,86	RMAT	2^20
18	Forward-Backward	RCC for GPU	Lomonosov-2	195,00	RMAT	2^18
19	Forward-Backward	PBGL MPI	Lomonosov	91,07	RMAT	2^21
20	Forward-Backward	RCC for CPU	Lomonosov-2	55,44	RMAT	2^18
21	Forward-Backward	RCC for CPU	IBM Regatta	53,60	SSCA-2	2^18
22	Forward-Backward	PBGL MPI	IBM BlueGene/P	45,75	RMAT	2^20
23	Forward-Backward	RCC for GPU	Lomonosov	44,78	RMAT	2^16
24	Forward-Backward	RCC for CPU	Lomonosov	42,00	RMAT	2^22
25	Forward-Backward	RCC for CPU	Lomonosov	41,00	RMAT	2^20
26	Forward-Backward	RCC for CPU	IBM Regatta	36,90	RMAT	2^18
27	Forward-Backward	RCC for CPU	Lomonosov	32,54	RMAT	2^20
28	Forward-Backward	PBGL MPI	IBM BlueGene/P	13,39	SSCA-2	2^16
29	Forward-Backward	PBGL MPI	IBM BlueGene/P	13,12	SSCA-2	2^18
30	Forward-Backward	RCC for CPU	Lomonosov	10,05	SSCA-2	2^20
31	Forward-Backward	RCC for CPU	Lomonosov	9,20	SSCA-2	2^18
32	Forward-Backward	RCC for CPU	Lomonosov	8.30	SSCA-2	2^20

"PerfData" list on "Strongly Connected Components"

(various methods, algorithms, implementations, computers)



compare different ways of solving the problem;



"PerfData" list on "Strongly Connected Components" (various methods, algorithms, implementations, computers)

Rating	Method	Implementation	Platform	MTEPS	GraphType	GraphSize
1	Shiloach-Vishkin	Ligra	Lomonosov-2	1307,00	RMAT	2^26
2	Shiloach-Vishkin	Ligra	Lomonosov-2	986,00	RMAT	2^23
3	Shiloach-Vishkin	Ligra	Lomonosov-2	947,00	RMAT	2^22
4	Shiloach-Vishkin	Ligra	Lomonosov-2	894,00	RMAT	2^24
5	Shiloach-Vishkin	Ligra	Lomonosov-2	864,00	RMAT	2^25
6	Shiloach-Vishkin	Ligra	Lomonosov-2	830,00	RMAT	2^20
7	Shiloach-Vishkin	Ligra	Lomonosov-2	782,00	RMAT	2^21
8	Forward-Backward	RCC for GPU	Lomonosov-2 (P100)	634,00	RMAT	2^20
9	Forward-Backward	RCC for GPU	Lomonosov-2 (P100)	620,00	RMAT	2^21
10	Forward-Backward	RCC for CPU	Lomonosov-2	564,00	RMAT	2^24
11	Shiloach-Vishkin	GAP	Lomonosov-2	547,00	RMAT	2^20
12	Forward-Backward	RCC for GPU	Lomonosov-2 (P100)	544,00	RMAT	2^22
13	Forward-Backward	RCC for GPU	Lomonosov-2 (P100)	528,00	RMAT	2^23
14	Forward-Backward	RCC for CPU	Lomonosov-2	498,00	RMAT	2^26
15	Forward-Backward	RCC for CPU	Lomonosov-2	497,00	RMAT	2^25
16	Forward-Backward	RCC for CPU	Lomonosov-2	486,00	RMAT	2^27
17	Shiloach-Vishkin	GAP	Lomonosov-2	480,00	RMAT	2^22
18	Forward-Backward	RCC for GPU	Lomonosov-2 (P100)	456,00	RMAT	2^25
19	Forward-Backward	RCC for GPU	Lomonosov-2 (P100)	453,00	RMAT	2^24
20	Forward-Backward	RCC for CPU	Lomonosov-2	452,00	RMAT	2^22
21	Forward-Backward	RCC for CPU	Lomonosov-2	440,24	SSCA-2	2^21
22	Forward-Backward	RCC for CPU	Lomonosov-2	432,00	RMAT	2^23
23	Forward-Backward	RCC for CPU	Lomonosov-2	426,00	RMAT	2^21
24	Forward-Backward	RCC for GPU	Lomonosov-2 (P100)	426,00	RMAT	2^26
25	Forward-Backward	RCC for CPU	Lomonosov-2	418,00	RMAT	2^20
26	Shiloach-Vishkin	GAP	Lomonosov-2	387,00	RMAT	2^23
27	Shiloach-Vishkin	GAP	Lomonosov-2	335,00	RMAT	2^21
28	Forward-Backward	PBGL MPI	IBM BlueGene/P	232,86	RMAT	2^20
29	Shiloach-Vishkin	GAP	Lomonosov-2	231,00	RMAT	2^24
30	Forward-Backward	RCC for GPU	Lomonosov-2	195,00	RMAT	2^18
31	Shiloach-Vishkin	GAP	Lomonosov-2	180,00	RMAT	2^25
32	Shiloach-Vishkin	GAP	Lomonosov-2	174 00	RMAT	2^26

Data are submitted by different people...

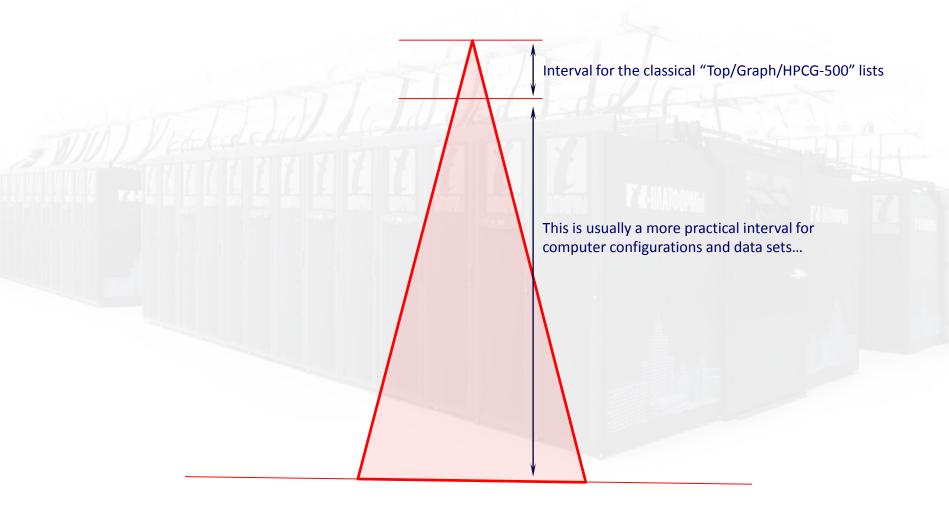
What does it mean to submit performance data to AlgoWiki?

Path to the algorithm in AlgoWiki 1 Implementation details 2 Computer platform 3 Input data 4 Performance data 5

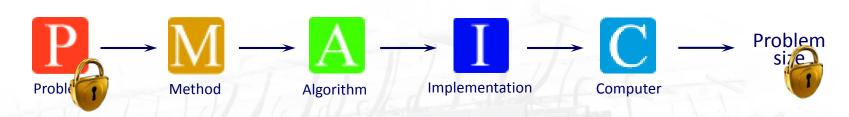
- 1. Information on the full chain: $P \rightarrow M \rightarrow A \rightarrow I \rightarrow C$ (a path in AlgoWiki)
- 2. Source code or reference to a package. Compiler, compilation options and flags, libraries used, makefile, launch parameters...
- 3. Configuration of the computing system with all necessary details.
 - Max / Medium / Small configurations.
- 4. Description of input data (matrices (dense, sparse, symmetric, ...), graphs (RMAT, SSCA-2...), sizes, ...)
 - Max / Medium / Small sizes.
 - Execution attributes (block size, processor matrix...)
- 5. Performance data:
 - Time, Flops, MTEPS, Gflops/Watt, efficiency (R_{max}/R_{peak}), ...

Important: we need to keep all data and details to be able to reproduce submitted performance results.

Max / Medium / Small – What does it mean? (computer configurations, data sets)



"PerfData" list on "Single Source Shortest Path" (How to solve the problem of the given size?)



Method	Implementation	Computing Platform	MTEPS	GraphType	GraphSize
Bellman-Ford	RCC for GPU	Lomonosov	1309,0	SSCA-2	2^20
Delta-Stepping	GAP	Lomonosov-2	512,0	RMAT	2^20
Bellman-Ford	Ligra	Lomonosov-2	511,0	RMAT	2^20
Bellman-Ford	RCC for GPU	Lomonosov	452,9	SSCA-2	2^20
Bellman-Ford	RCC for CPU	Lomonosov-2	418,0	RMAT	2^20
Bellman-Ford	Graph500 MPI	Lomonosov	350,0	RMAT	2^20
Bellman-Ford	RCC for CPU	Lomonosov	204,1	RMAT	2^20
Bellman–Ford	RCC for CPU	Lomonosov	183,5	SSCA-2	2^20
Dijkstra's	PBGL MPI	Cluster / "Angara" interconnect	150,0	SSCA-2	2^20
Bellman–Ford	Graph500 MPI	Lomonosov	120,0	RMAT	2^20
Bellman-Ford	Graph500 MPI	Lomonosov	18,0	SSCA-2	2^20
Dijkstra's	PBGL MPI	IBM BlueGene/P	8,9	SSCA-2	2^20
Delta-Stepping	PBGL MPI	IBM BlueGene/P	3,8	SSCA-2	2^20
Delta-Stepping	PBGL MPI	IBM BlueGene/P	1,3	RMAT	2^20
Dijkstra's	PBGL MPI	IBM BlueGene/P	0,6	RMAT	2^20

"PerfData": the most efficient graph applications (Where is this computer platform effective?)



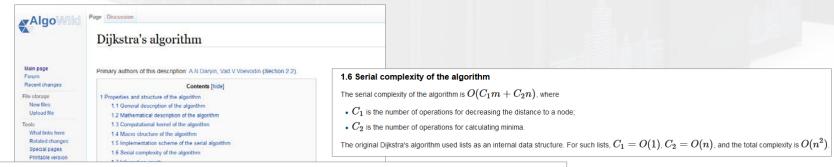
Problem	Algorithm	Implementation	Platform	MTEPS	GraphType	GraphSize
Breadth-first search	BFS	RCC for GPU	Lomonosov-2 (NVIDIA P100)	11061	RMAT	2^22
Breadth-first search	BFS	GAP	Lomonosov-2 (NVIDIA K40)	5350	RMAT	2^22
Breadth-first search	BFS	Ligra	Lomonosov-2 (NVIDIA K40)	4168	RMAT	2^22
Minimum Spanning Tree	Boruvka's	RCC for GPU	Lomonosov-2 (NVIDIA P100)	3793	RMAT	2^26
Single Source Shortest Path	Bellman-Ford	RCC for GPU	Lomonosov (NVIDIA 2090)	1309,0	SSCA-2	2^20
Strongly Connected Components	Shiloach-Vishkin	Ligra	Lomonosov-2 (x86)	1307,00	RMAT	2^26
Single Source Shortest Path	Bellman-Ford	Ligra	Lomonosov-2 (x86)	1035,0	RMAT	2^21
Single Source Shortest Path	Delta-Stepping	PBGL MPI	Cluster / "Angara" net	809,5	SSCA-2	2^21
Page Rank	Page Rank	Nvidia nvGraph	Lomonosov-2 (NVIDIA K40)	753	RMAT	2^18
Breadth-first search	BFS	RCC for CPU	NEC SX-ACE	715	RMAT	2^22
Strongly Connected Components	Forward-Backward	RCC for GPU	Lomonosov-2 (NVIDIA P100)	620,00	RMAT	2^21
Single Source Shortest Path	Delta-Stepping	GAP	Lomonosov-2 (x86)	616,0	RMAT	2^21
Strongly Connected Components	Forward-Backward	RCC for CPU	Lomonosov-2 (x86)	564,00	RMAT	2^24
Strongly Connected Components	Shiloach-Vishkin	GAP	Lomonosov-2 (x86)	547,00	RMAT	2^20
Minimum Spanning Tree	Boruvka's	RCC for GPU	Lomonosov (NVIDIA 2090)	204,441	SSCA-2	2^14
Page Rank	Page Rank	GAP	Lomonosov-2 (NVIDIA K40)	172	RMAT	2^24
Breadth-first search	BFS	PBGL MPI	IBM BlueGene/P	167	SSCA-2	2^22
Breadth-first search	BFS	PBGL MPI	Lomonosov (x86)	155	SSCA-2	2^22
Minimum Spanning Tree	Boruvka's	RCC for CPU	Intel Xeon Phi (KNL)	25,16	SSCA-2	2^21
Single Source Shortest Path	Dijkstra's	PBGL MPI	IBM BlueGene/P	8,9	SSCA-2	2^20

AlgoWiki: an easy step back to analyze "PerfData" (theoretical potential of algorithms is described in AlgoWiki)

SSSP problem:

Method	Implementation	Computing Platform	MTEPS	GraphType	GraphSize	
Bellman-Ford	RCC for CPU	Lomonosov-2	418,0	RMAT	2^20	
Bellman-Ford	Graph500 MPI	Lomonosov	350,0	RMAT	2^20	
Bellman-Ford	RCC for CPU	Lomonosov-2	204,1	RMAT	2^20	
Dijkstra's PBGL MPI		Cluster / "Angara" interconnect	150,0	SSCA-2	2^20	
Delta-Stepping	PBGL MPI	Lomonosov	124,1	SSCA-2	2^21	
Bellman-Ford	Graph500 MPI	Lomonosov	120,0	RMAT	2^20	
Dijkstra's	PBGL MPI	IBM BlueGene/P	8,9	SSCA-2	2^20	
Dijkstra's	PBGL MPI	Lomonosov	5,3	SSCA-2	2^21	
Delta-Stepping	PBGL MPI	IBM BlueGene/P	3.8	SSCA-2	2^20	



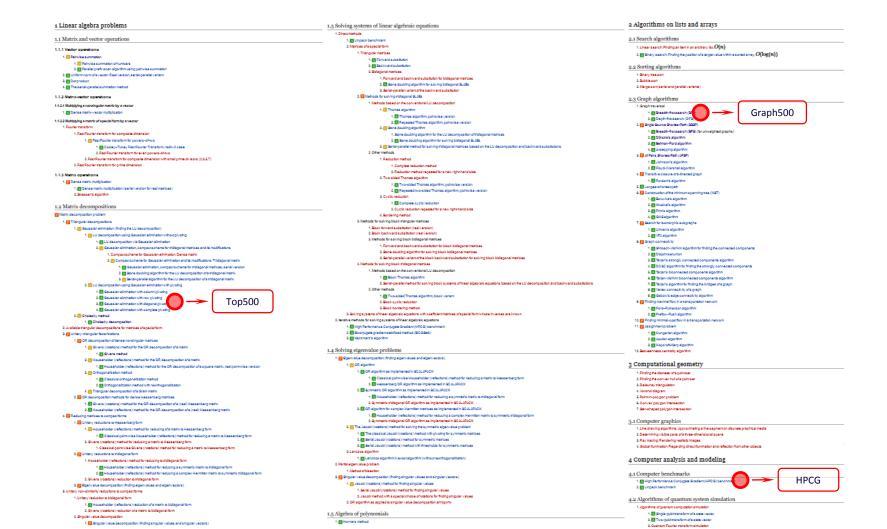


1.8 Parallelization resource of the algorithm

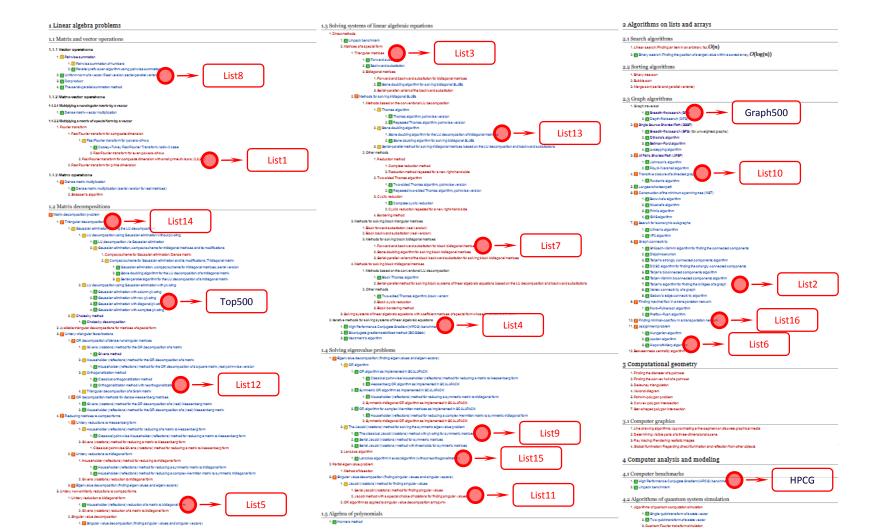
Dijkstra's algorithm admits an efficient parallelization $^{[3]}$ its average execution time is $O(n^{1/3} \ln n)$, and the computational complexity is $O(n \ln n + m)$.

The algorithm of Δ-stepping can be regarded as a parallel version of Dijkstra's algorithm.

Algorithm classification and computer comparison (What have we had up to now? Three points on the entire set of algorithms)



Algorithm classification and computer comparison (What can AlgoWiki add to the standard three benchmarks/lists?)



General methodology to compare computing platforms (using any algorithm)

problems for evaluation of computer platforms



