International Advanced Research Workshop
on High Performance Computing
from Clouds and Big Data to Exascale and Beyond

Medical practice: diagnostics, treatment and surgery in supercomputing centers

Prof. Vladimir V. Voevodin
Moscow State University
voevodin@parallel.ru

July, 11, 2014, Cetraro, Italy

Why are they together?

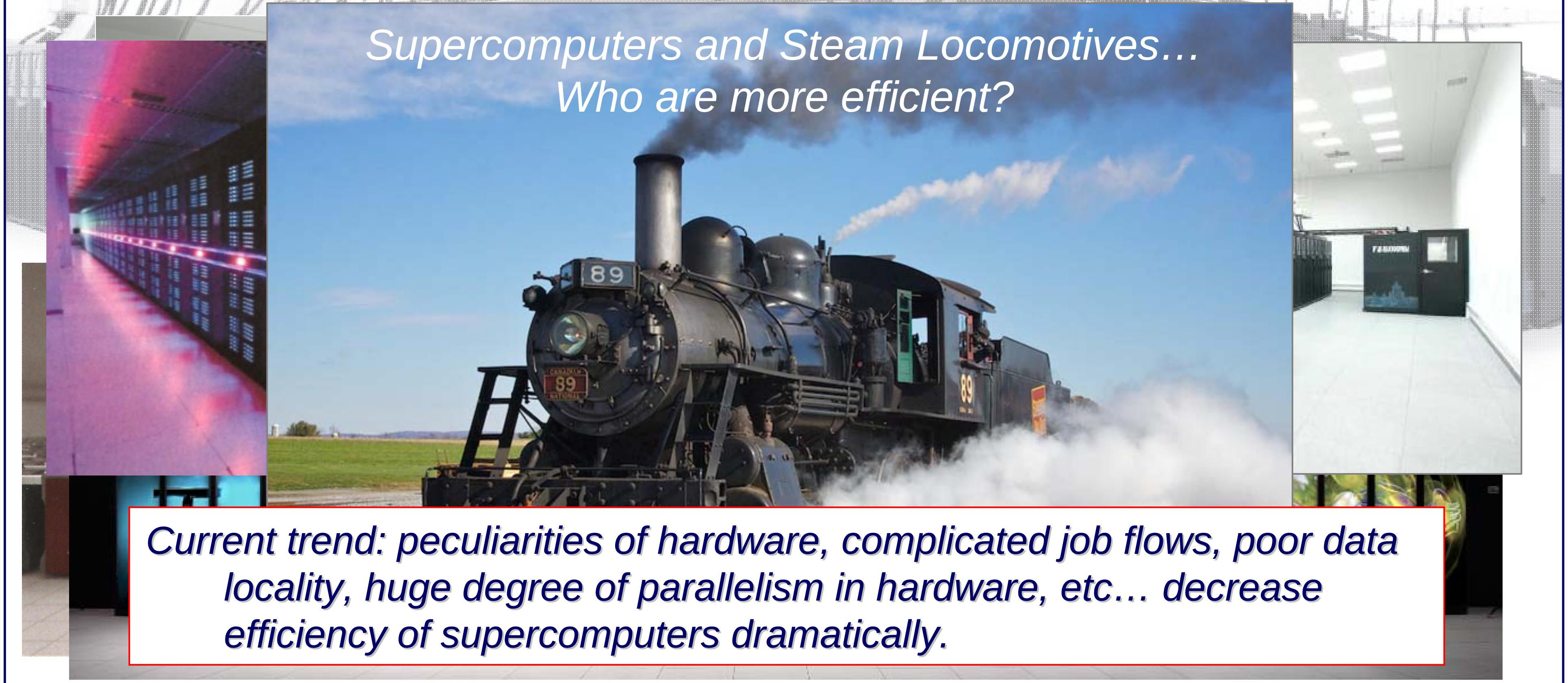
Efficiency of Supercomputing Centers

1 Pflop/s system… What do we expect?

useful

*1Pflop * 60sec * 60min * 24hours * 365days = 31,5* ***ZettaFlop (10^{21})*** *per year*

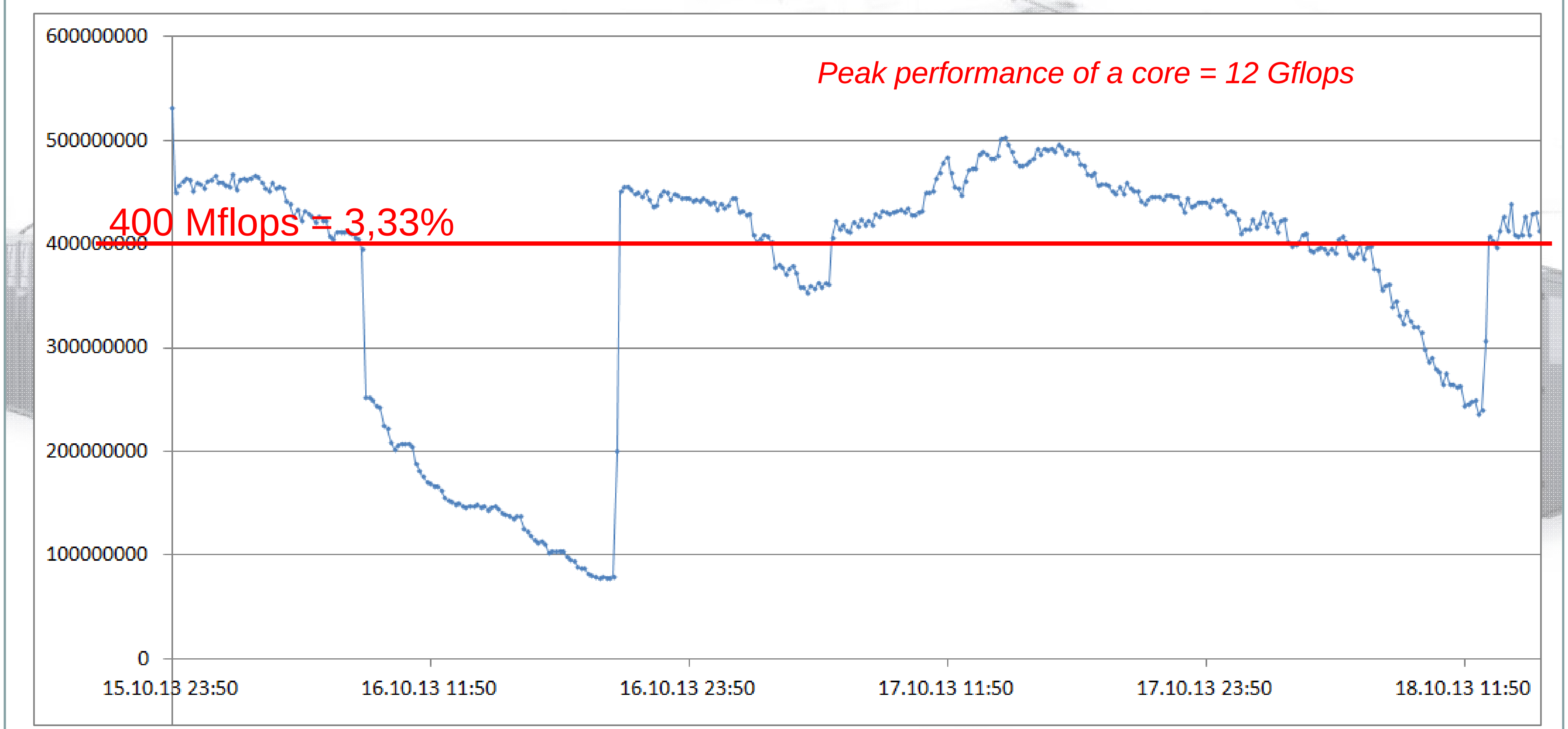
What is in reality?   A small, small, small fraction…



Supercomputers and Steam Locomotives…
Who are more efficient?

Current trend: peculiarities of hardware, complicated job flows, poor data locality, huge degree of parallelism in hardware, etc… decrease efficiency of supercomputers dramatically.

Efficiency of Supercomputing Centers
(straightforward approach)



Peak performance of a core = 12 Gflops

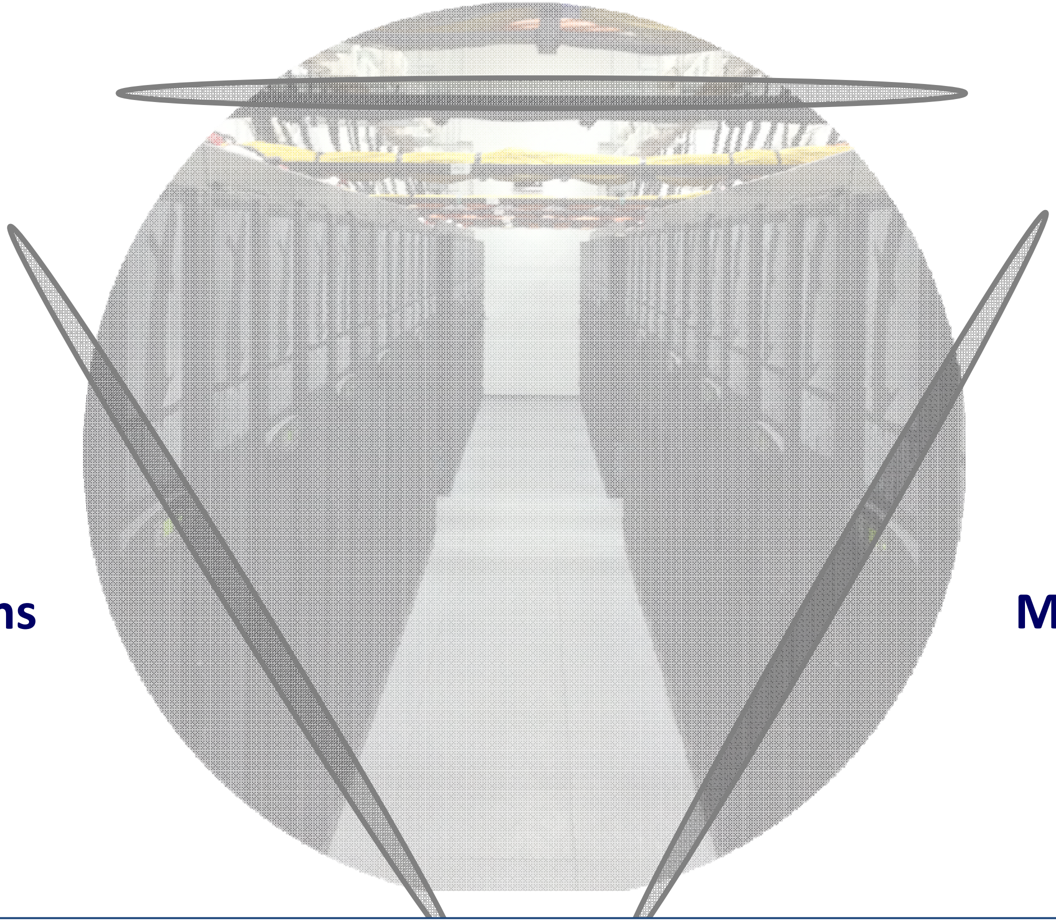
400 Mflops = 3,33%

Average performance (one core) of “Chebyshev” supercomputer for 3 days

Efficiency of Supercomputing Centers

Where are sources of efficiency losses?

Who is interested in efficiency of supercomputing centers?



Users

SysAdmins

Management

Users, Management, SysAdmins: work at different scope, have different rights, make different decisions.

What is efficiency of supercomputing centers?

Users *– efficiency in solving their problems, sometimes efficiency of apps*

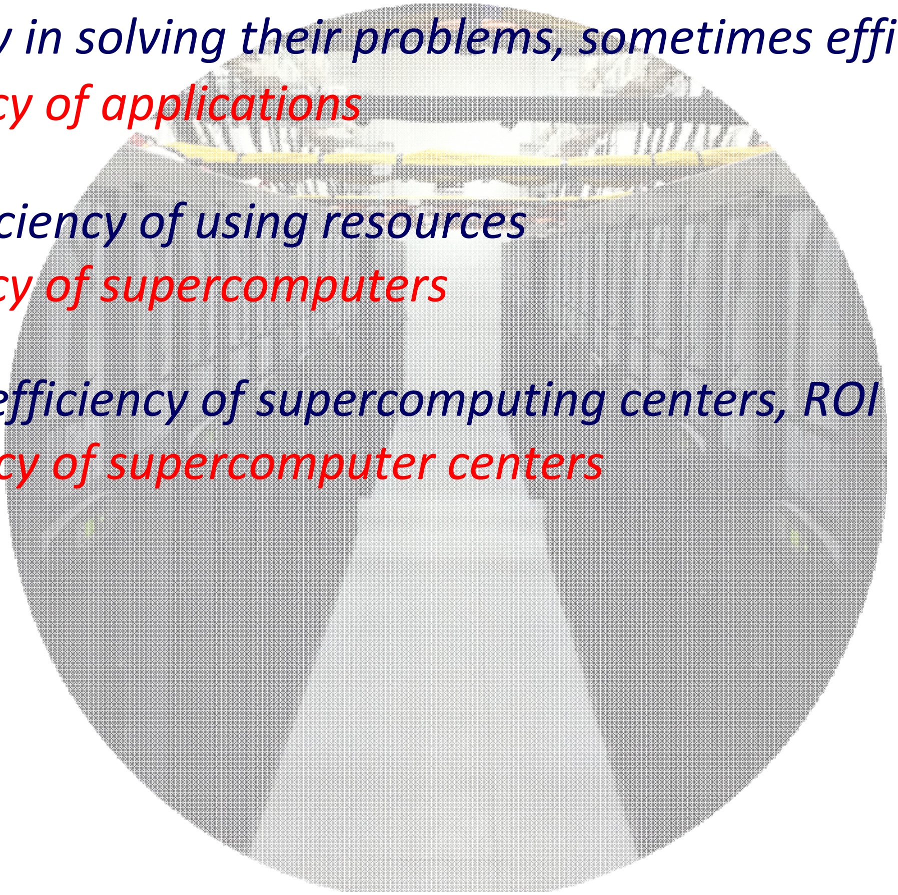
Efficiency of applications

SysAdmins *– efficiency of using resources*

Efficiency of supercomputers

Management *– efficiency of supercomputing centers, ROI*

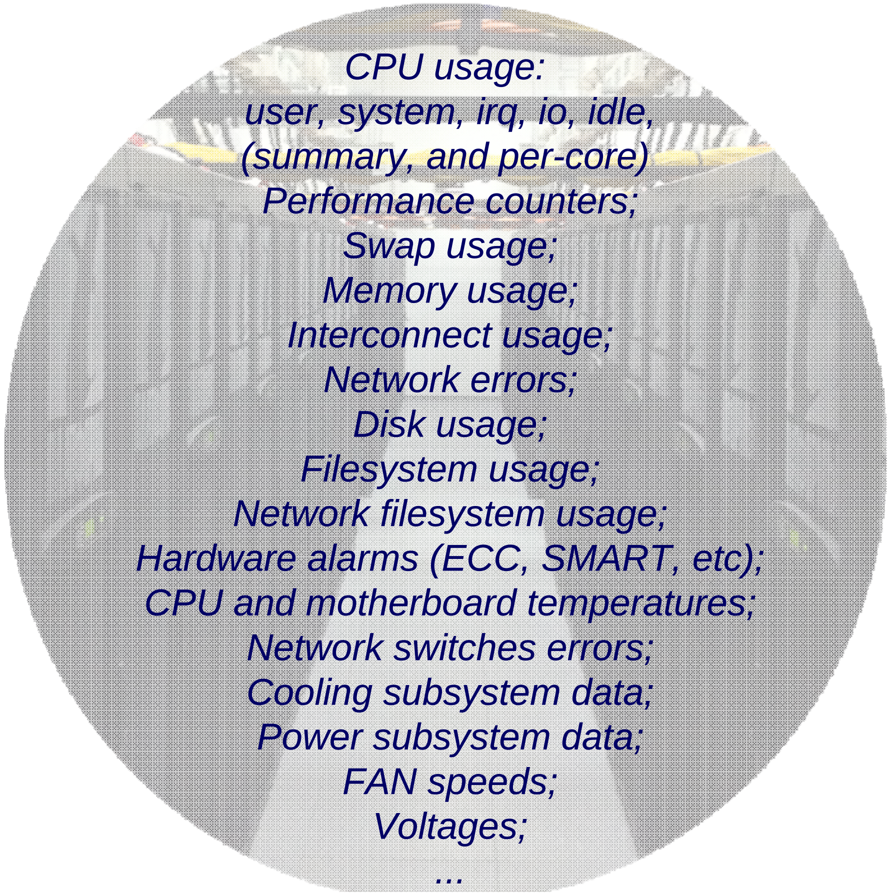
Efficiency of supercomputer centers



Users, Management, SysAdmins: work at different scope, have different rights, make different decisions.

Efficiency of Supercomputing Centers

(system-level view)



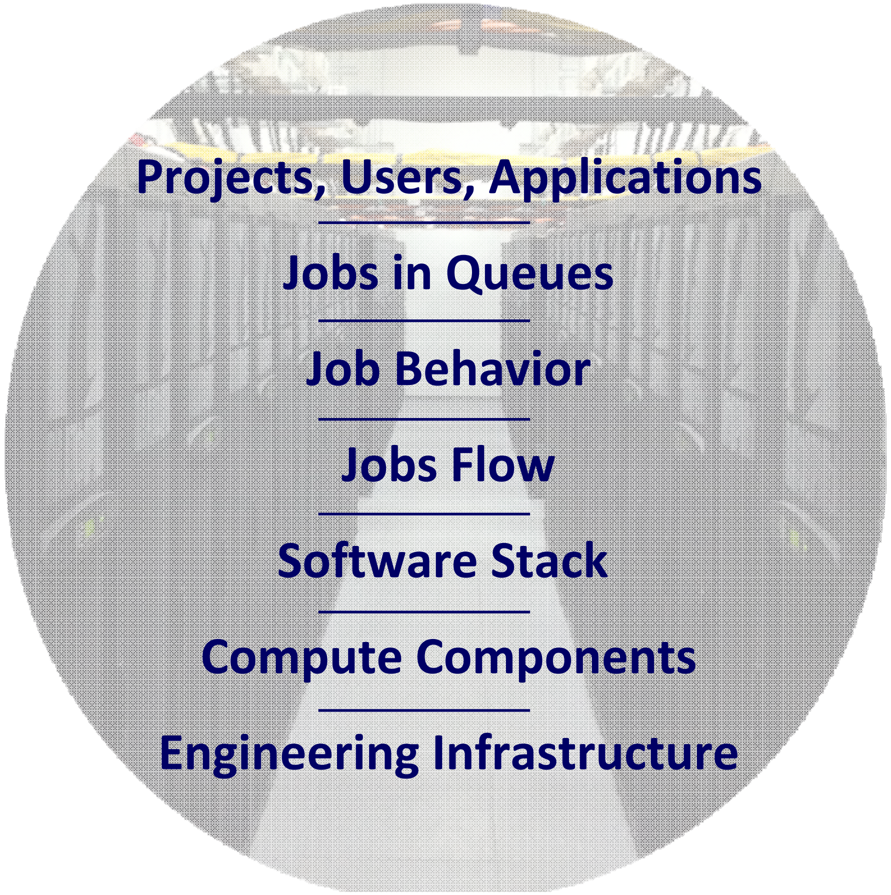
CPU usage:
user, system, irq, io, idle,
(summary, and per-core)
Performance counters;
Swap usage;
Memory usage;
Interconnect usage;
Network errors;
Disk usage;
Filesystem usage;
Network filesystem usage;
Hardware alarms (ECC, SMART, etc);
CPU and motherboard temperatures;
Network switches errors;
Cooling subsystem data;
Power subsystem data;
FAN speeds;
Voltages;
...

Sources of efficiency losses can be everywhere…

We must be able to detect and show ***not symptoms but the root causes*** *of efficiency degradation.*

Efficiency of Supercomputing Centers

(SC Center-level view)



Projects, Users, Applications

Jobs in Queues

Job Behavior

Jobs Flow

Software Stack

Compute Components

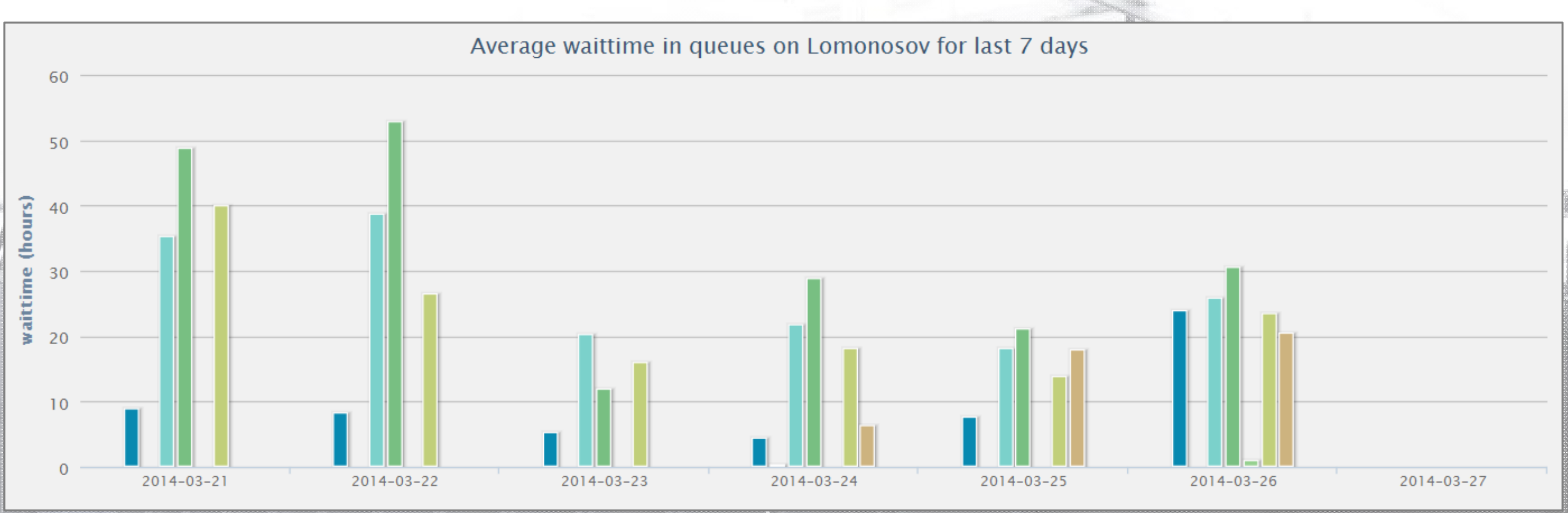
Engineering Infrastructure

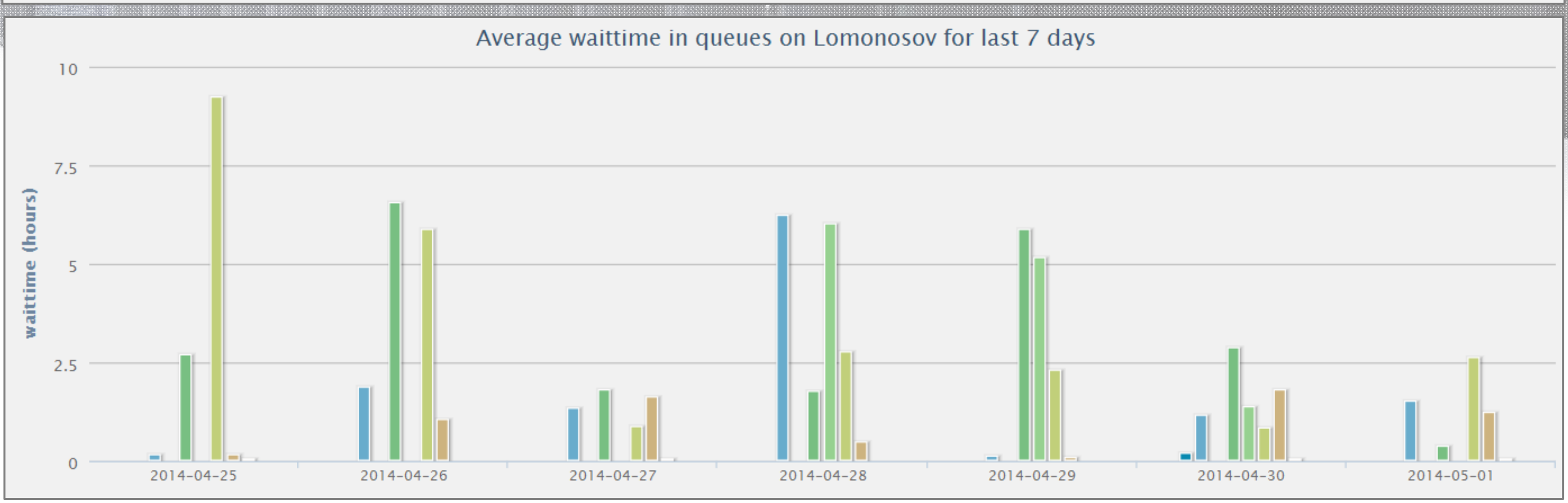
Sources of efficiency losses can be everywhere…

We must be able to detect and show ***not symptoms but the root causes*** *of efficiency degradation.*

Efficiency of Supercomputing Centers

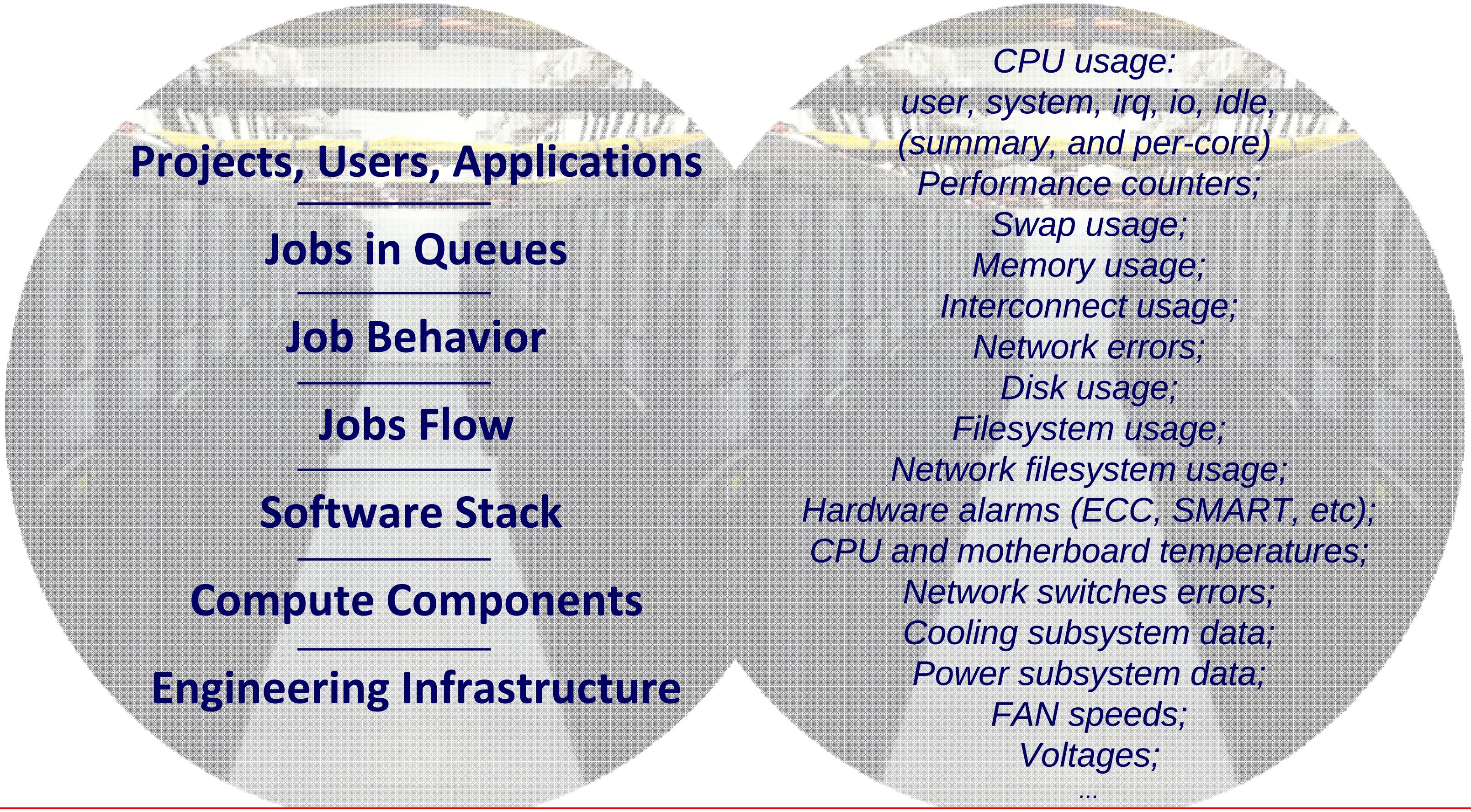
(users, quotas and queues)





Efficiency of Supercomputing Centers

(three target groups + system level + SC Center level)



Projects, Users, Applications

Jobs in Queues

Job Behavior

Jobs Flow

Software Stack

Compute Components

Engineering Infrastructure

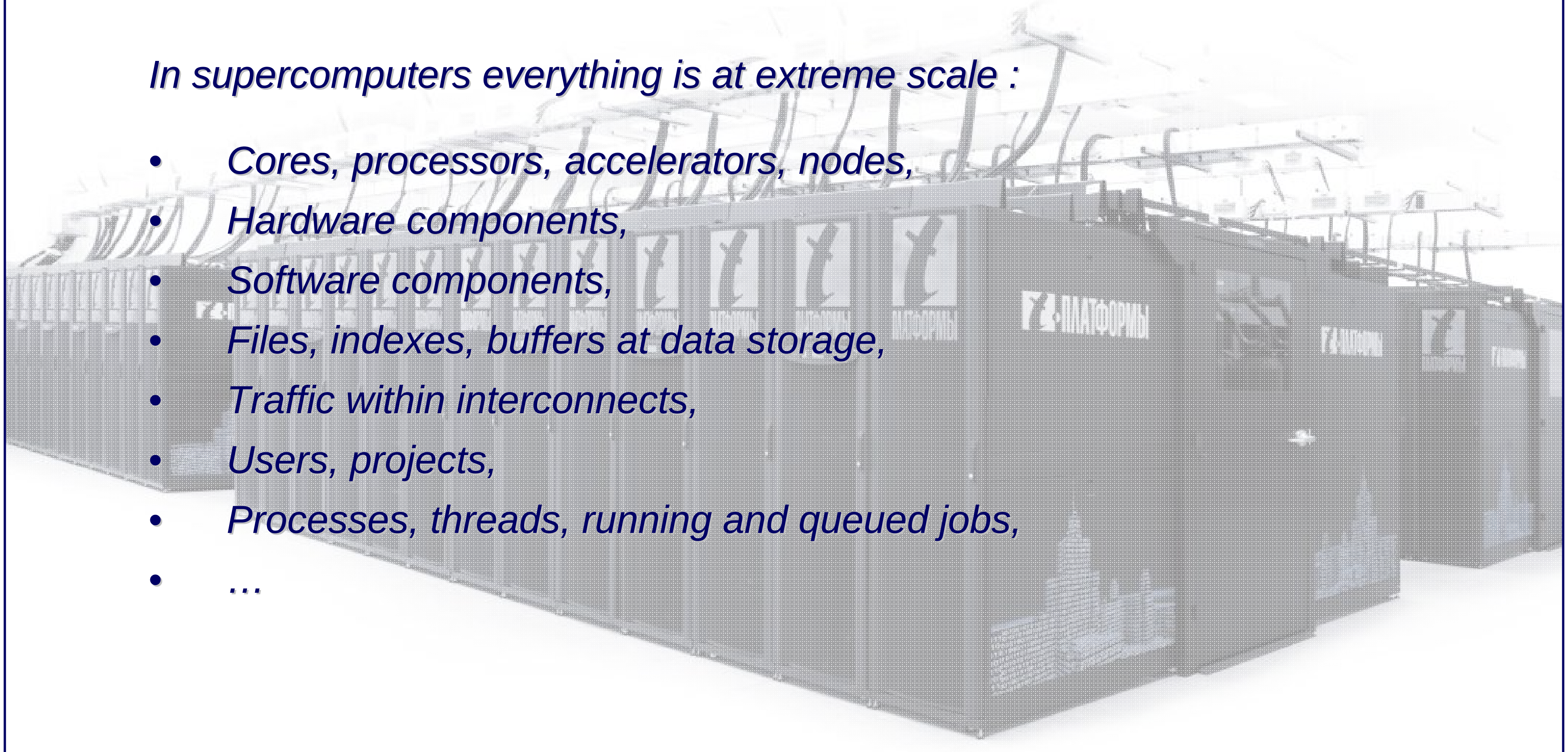
CPU usage:
user, system, irq, io, idle,
(summary, and per-core)
Performance counters;
Swap usage;
Memory usage;
Interconnect usage;
Network errors;
Disk usage;
Filesystem usage;
Network filesystem usage;
Hardware alarms (ECC, SMART, etc);
CPU and motherboard temperatures;
Network switches errors;
Cooling subsystem data;
Power subsystem data;
FAN speeds;
Voltages;
...

Current trend: too sophisticated structure of supercomputers has led to loss of control over full understanding (knowledge) of their behavior.
Our goal is the total control over HW/SW and applications.

What is a 10-petaflops supercomputer today?

- *High price,*
- *High power consumption,*
- *Diversity of applications,*
- *High degree of parallelism,*
- *Large numbers are everywhere,*

Large Numbers in Supercomputers
(large now, huge very soon)

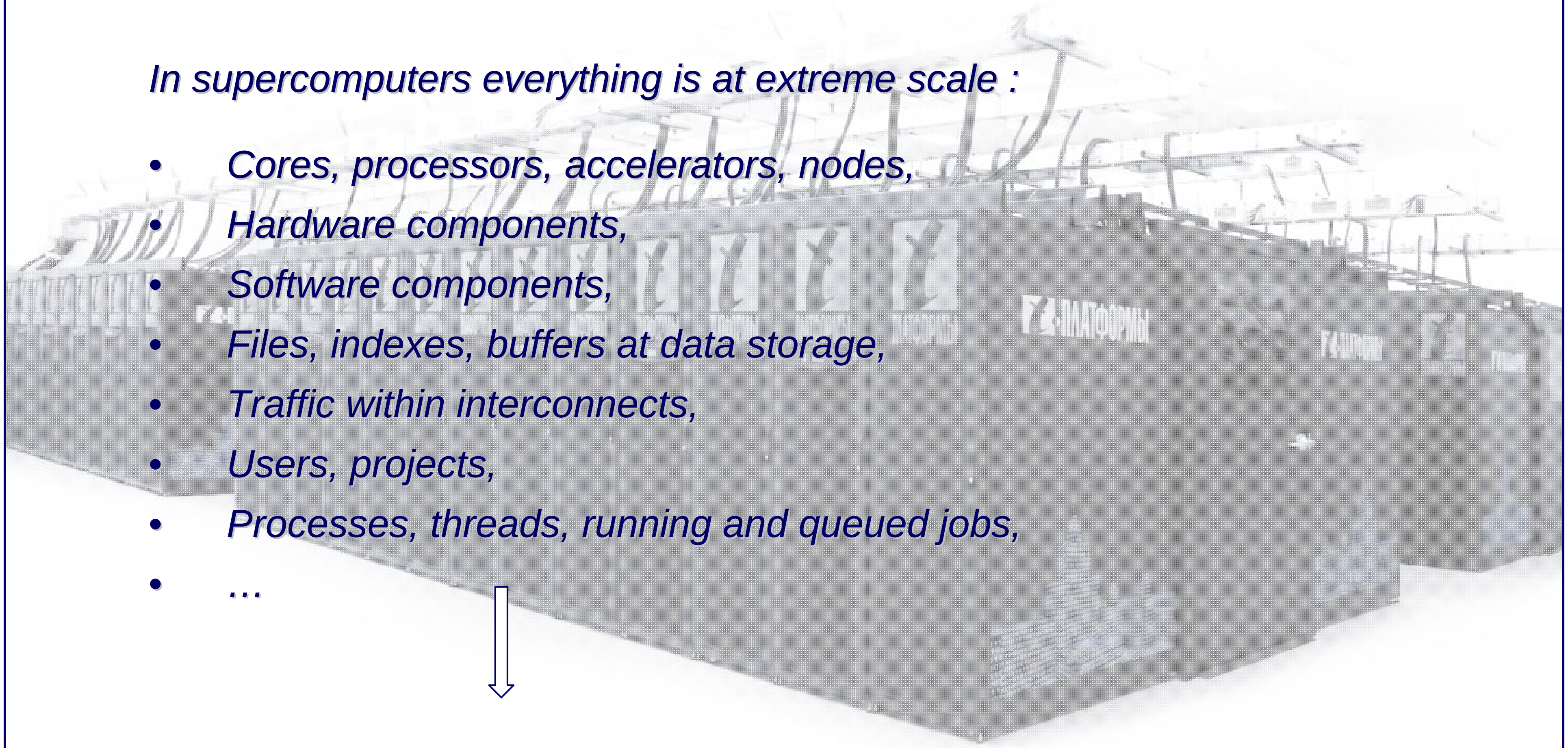


In supercomputers everything is at extreme scale :

- *Cores, processors, accelerators, nodes,*
- *Hardware components,*
- *Software components,*
- *Files, indexes, buffers at data storage,*
- *Traffic within interconnects,*
- *Users, projects,*
- *Processes, threads, running and queued jobs,*
- *…*

Current trend: all these numbers grow extremely fast!

Large Numbers in Supercomputers
(large now, huge very soon)



In supercomputers everything is at extreme scale :

- *Cores, processors, accelerators, nodes,*
- *Hardware components,*
- *Software components,*
- *Files, indexes, buffers at data storage,*
- *Traffic within interconnects,*
- *Users, projects,*
- *Processes, threads, running and queued jobs,*
- *…*

⇩

It’s impossible to predict/describe state of a supercomputer…

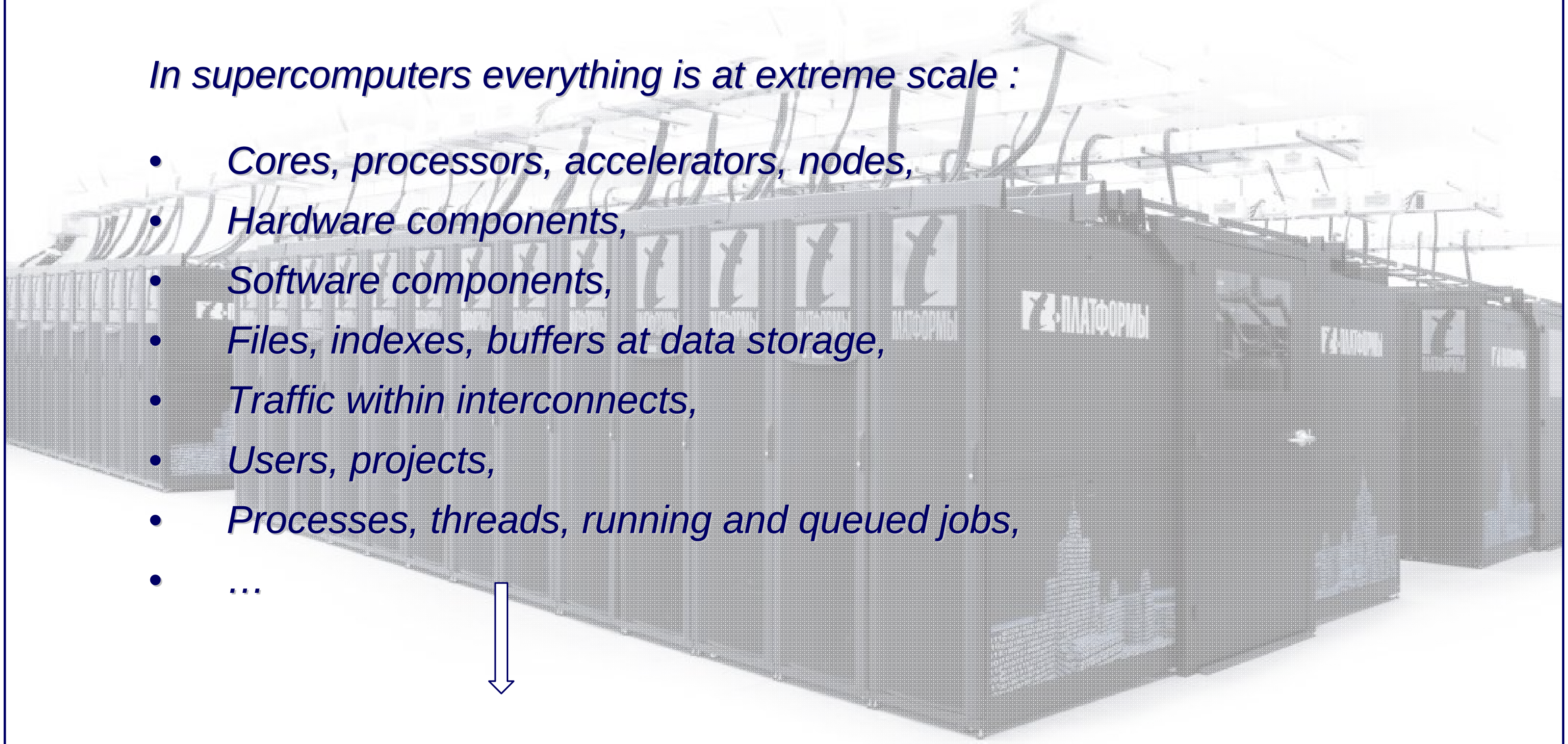
We have almost lost control…

Nuclear Power Stations

(total control)

Large Numbers in Supercomputers

(large now, huge very soon)



In supercomputers everything is at extreme scale :

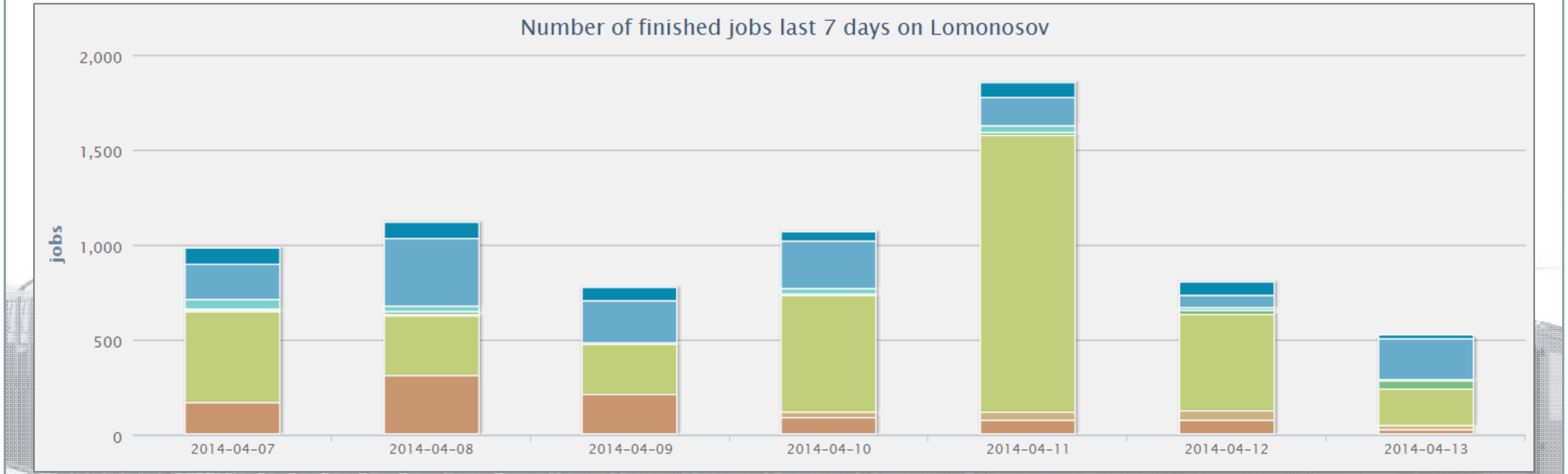
- *Cores, processors, accelerators, nodes,*
- *Hardware components,*
- *Software components,*
- *Files, indexes, buffers at data storage,*
- *Traffic within interconnects,*
- *Users, projects,*
- *Processes, threads, running and queued jobs,*
- *…*

↓

It's impossible to predict/describe state of a supercomputer…

We have almost lost control… Do we need to keep control over supercomputers?

Total control: cost of delay…



Supercomputer “Lomonosov”:

- about 1000 completed jobs per day,
- approx. 200 running jobs all the time,

if a job scheduler hangs/dies, a half of the supercomputer will be idle in 2-3 hours.

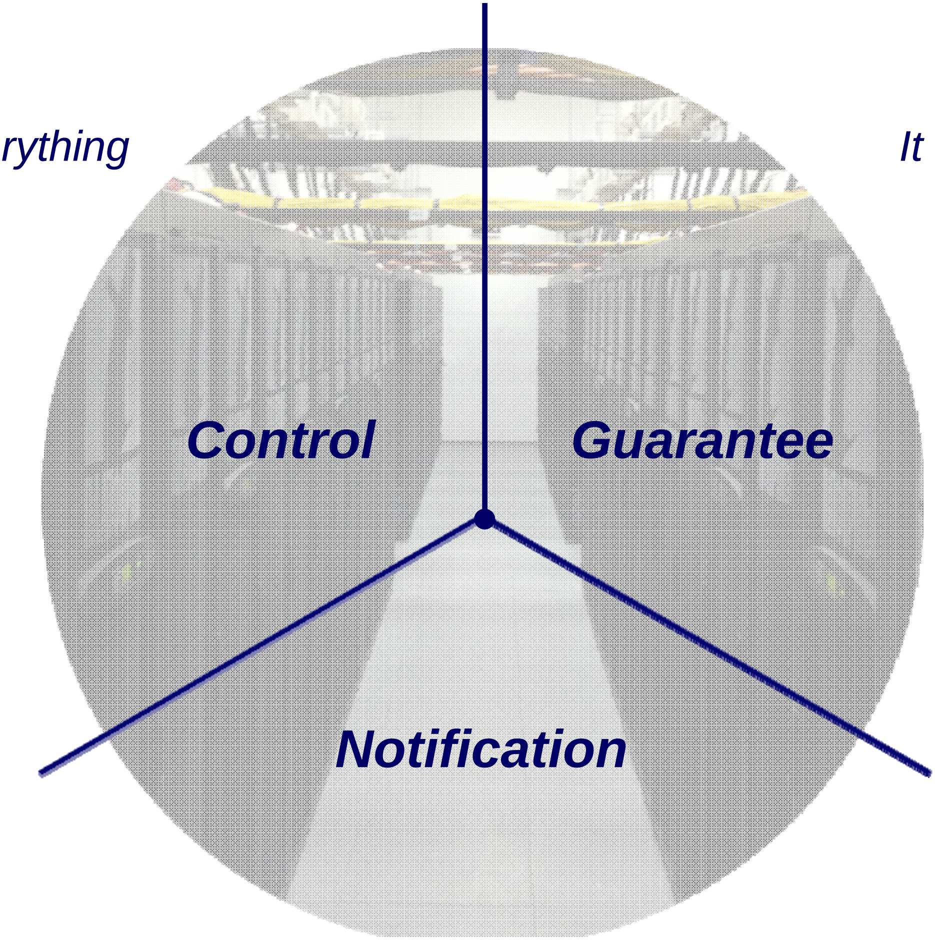
We need to keep control over supercomputers!

Current trend: the cost of delay with a proper reaction grows permanently.

Supercomputers: three parts of efficiency

*1st part.
We must control everything
what is necessary to
control efficiency
permanently.*

*2nd part.
It behaves like we expect,
coincidence between
theory and practice.
Guarantee.*



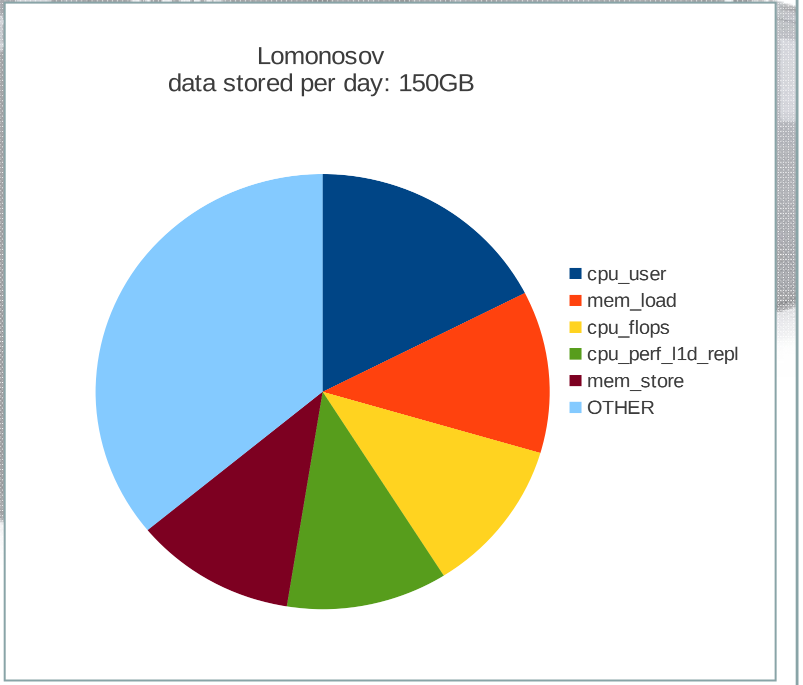
3rd part. We must know (be notified) about everything on time.

Monitoring System for Supercomputers

(1st part: control)

Monitoring system, requirements:

- *we need to know: what, where, when.*
- *scalability: millions of compute nodes, dozens sensors per node,*
- *low overheads: CPU, disks, interconnects (1% and less),*
- *frequency: a few seconds and less,*
- *easily reconfigurable and expandable,*
- *portable across platforms,*
- *active and passive modes.*

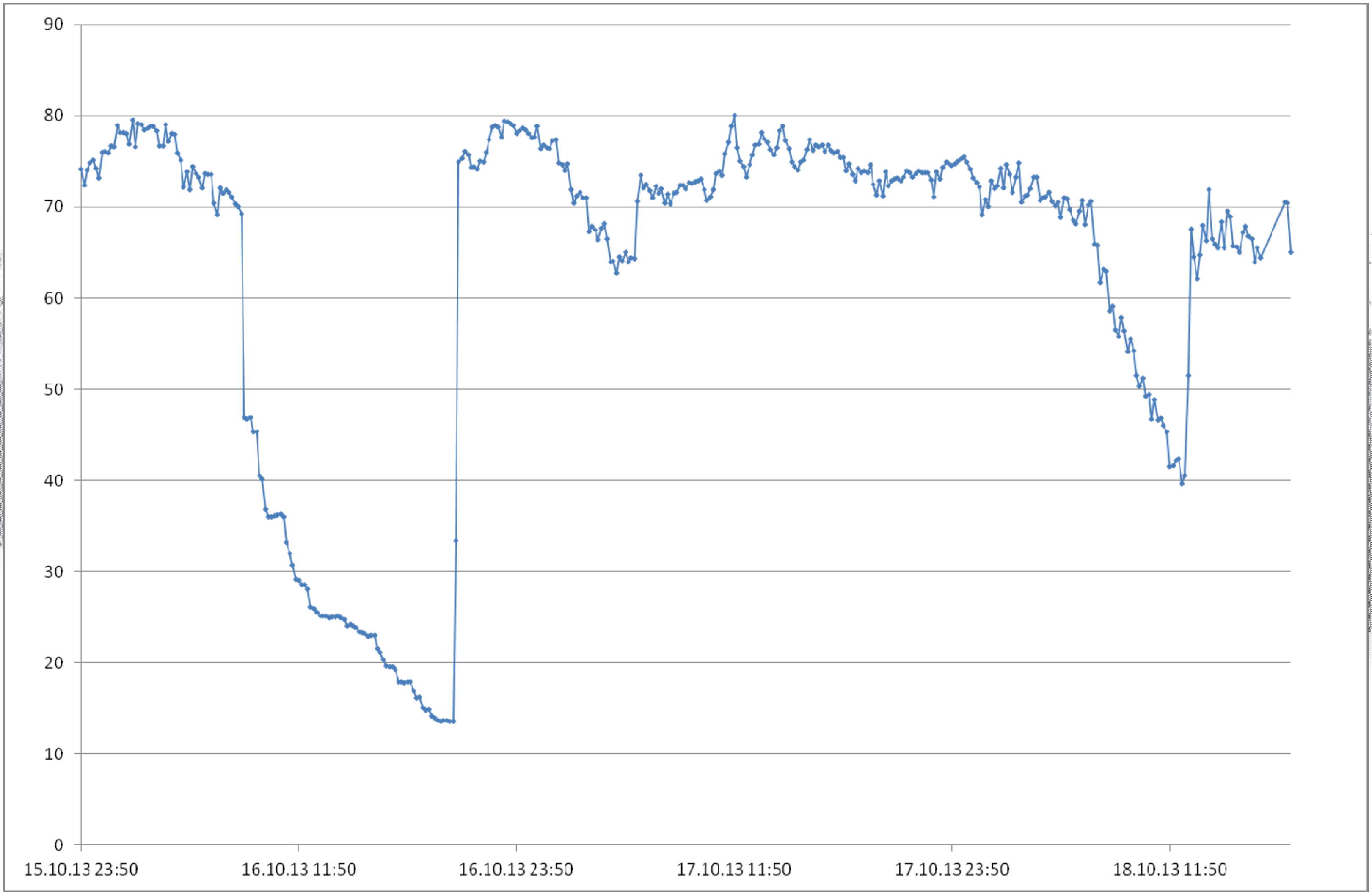


Aggressive filtering of data!

Current trend: monitoring will be an integral part of all future complex HW&SW systems.

Efficiency of supercomputing centers

(1st part: control. Integral characteristics)

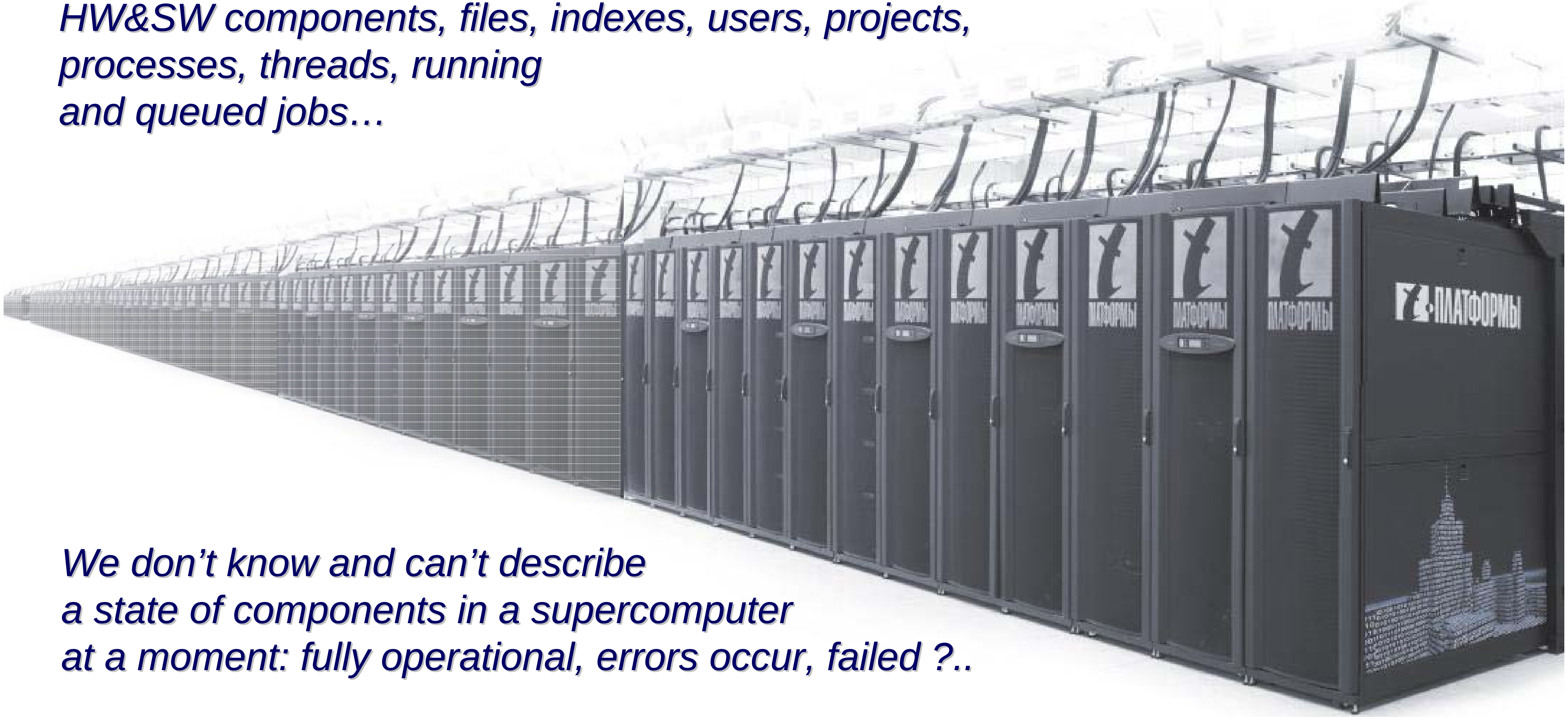


Average CPU Load of “Chebyshev” supercomputer for 3 days

Guarantee, Predictability and Autonomous Life of Supercomputers

(2nd part: guarantee)

Large numbers in supercomputers*: cores, processors, accelerators, nodes, HW&SW components, files, indexes, users, projects, processes, threads, running and queued jobs…*



We don't know and can't describe a state of components in a supercomputer at a moment: fully operational, errors occur, failed ?..

Guarantee, Predictability and Autonomous Life of Supercomputers

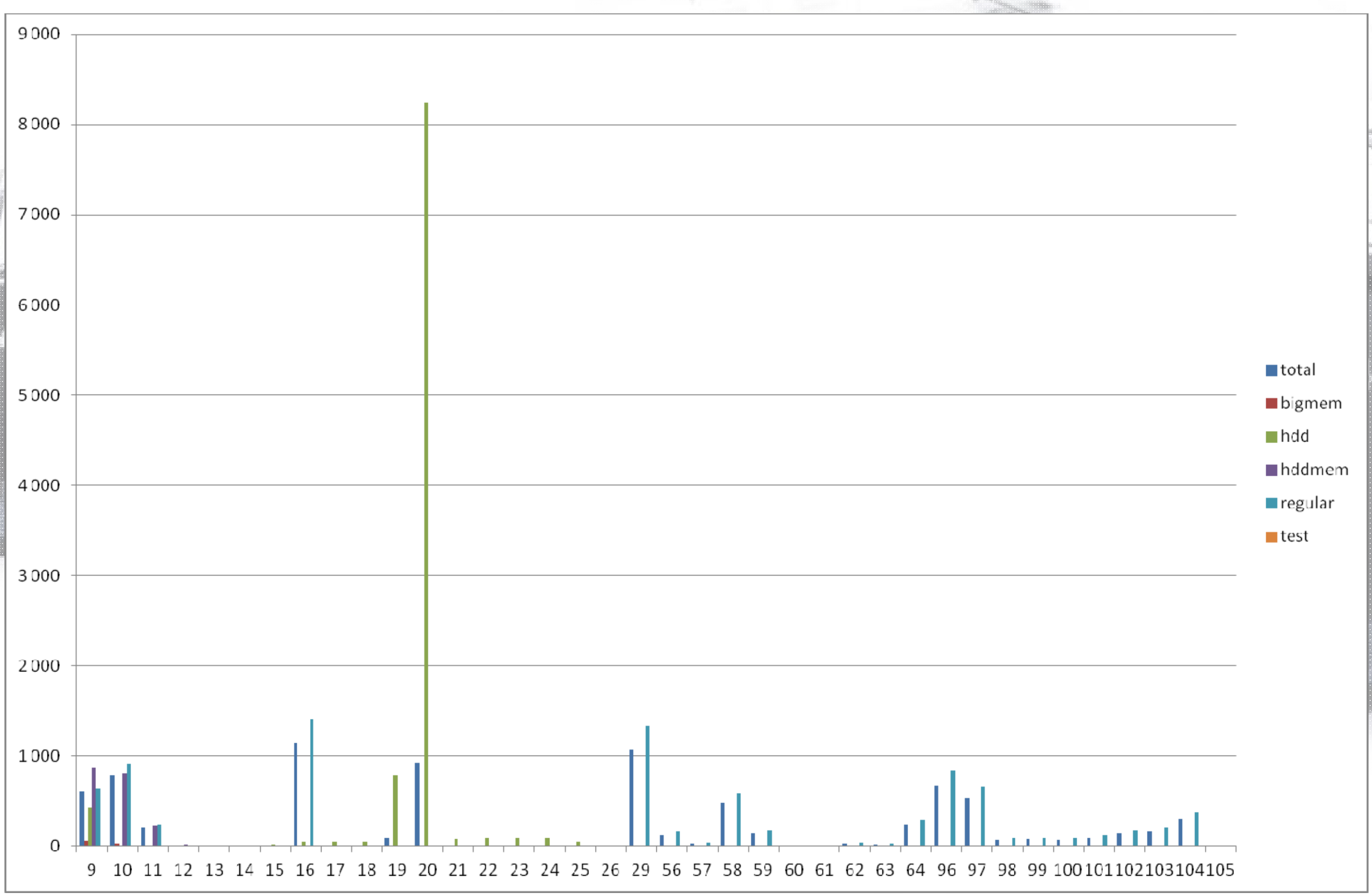
(2nd part: guarantee)

What is now? We hope a HW/SW component works until we get an evidence that it has failed.

What do we need?

We need a guarantee:
if something goes wrong inside a
supercomputer we shall be notified immediately.

Distribution of LoadAVG for 3 days
(2nd part: guarantee)



LoadAVG: an average number of processes which are ready for execution.

Control over everything!

Guarantee, Predictability and Autonomous Life of Supercomputers

(2nd part: guarantee)

What is now? We hope a component works until we get an evidence that it has failed.

What do we need?

Our expectations = Reality

We need a guarantee:
if something goes wrong inside a
supercomputer we shall be notified immediately.

We want a system behaves in a way we expect it should behave.

Guarantee, Predictability and Autonomous Life of Supercomputers

(2nd part: guarantee)

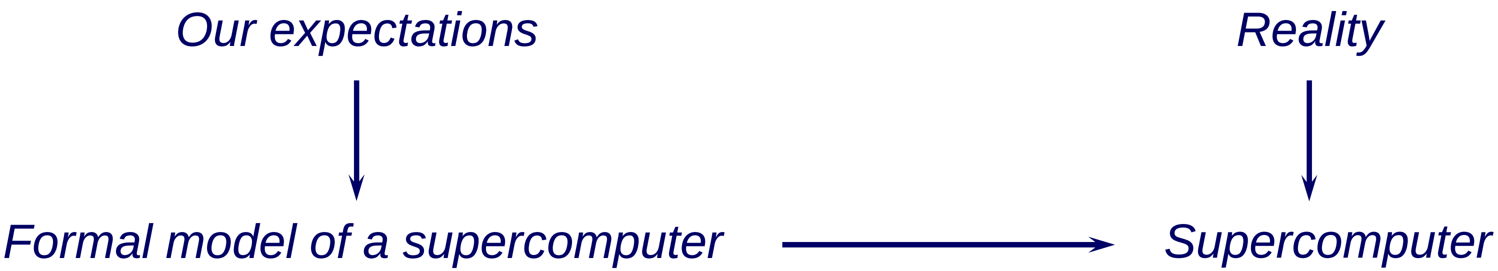
If discrepancy occurs between our expectations and supercomputer behavior we need to know immediately about it.
But…

Supercomputer is huge, we can’t control it to a full extent anymore.
But…

Supercomputer can do it itself (instead of us), if we explain
what “our expectations” are.

Guarantee, Predictability and Autonomous Life of Supercomputers

(2nd part: guarantee)



Our expectations

Reality

Formal model of a supercomputer

Supercomputer

Supercomputers should be autonomous in self-control.

Moreover:
The larger a supercomputer, the more autonomous it should be.

Guarantee, Predictability and Autonomous Life of Supercomputers

(2nd part: guarantee)

How it can be done?

- *Total monitoring of hardware and software components, engineering infrastructure;*
- *As a guarantee of “our expectations = reality”:*
  - *a formal model of supercomputers (a graph),*
  - *a set of formal rules,*

as a basis for an Autonomous life and control of MSU supercomputers:

- “Chebyshev”, 60 Tflops, 625 CPUs:

In its model: 9113 nodes, 24906 edges, 150 rules, 100 reactions;

- “Lomonosov”, 1.7 Pflops, 12K CPUs, 2K GPU:

In a model: 400K+ nodes.

Initial deployment, Detection of faults, critical and emergency situations, Turning off minimum amount of hardware, Self diagnostics, Previous accidents, etc. are done according to a model and rules.

Current trend: many decisions about control over HW&SW of supercomputers must be taken automatically.

A concept of “situation screen”: requirements

(3rd part: notification)

Visualization of all components of supercomputers:

- hardware: a computational part.
- hardware: engineering infrastructure.
- software stack.
- dynamics of applications.
- jobs flows.
- users.

The total control over supercomputer.

Extreme level of parallelism.

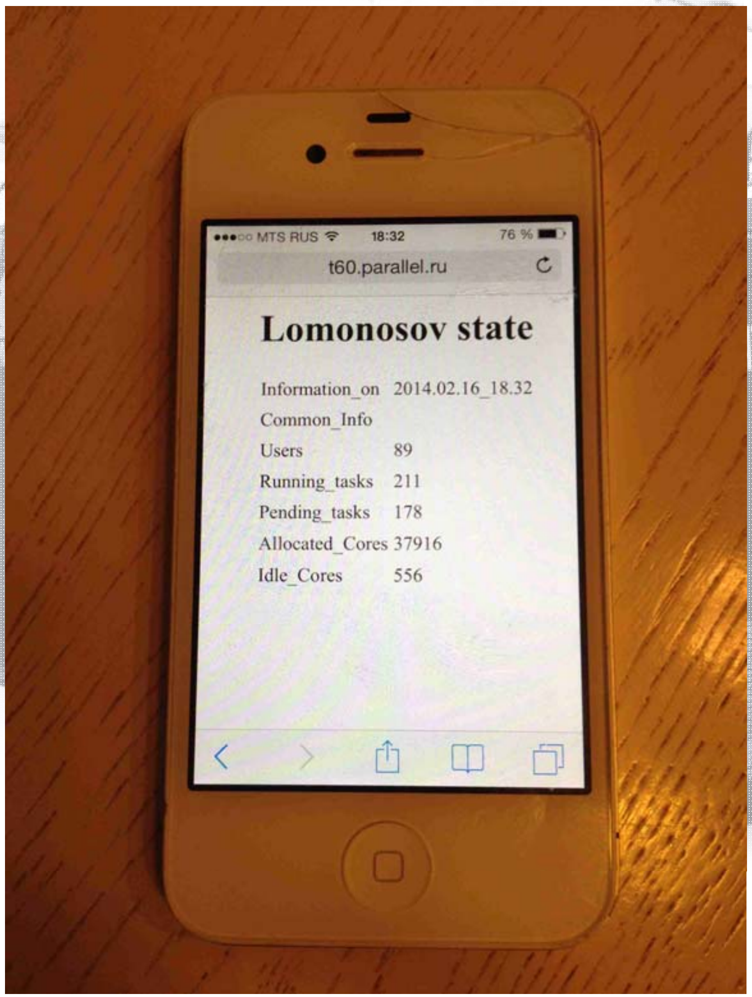
Low overheads.

General and specific views.

Openness to external data sources.

Three target groups in supercomputers centers.

Situation screen: a mobile option

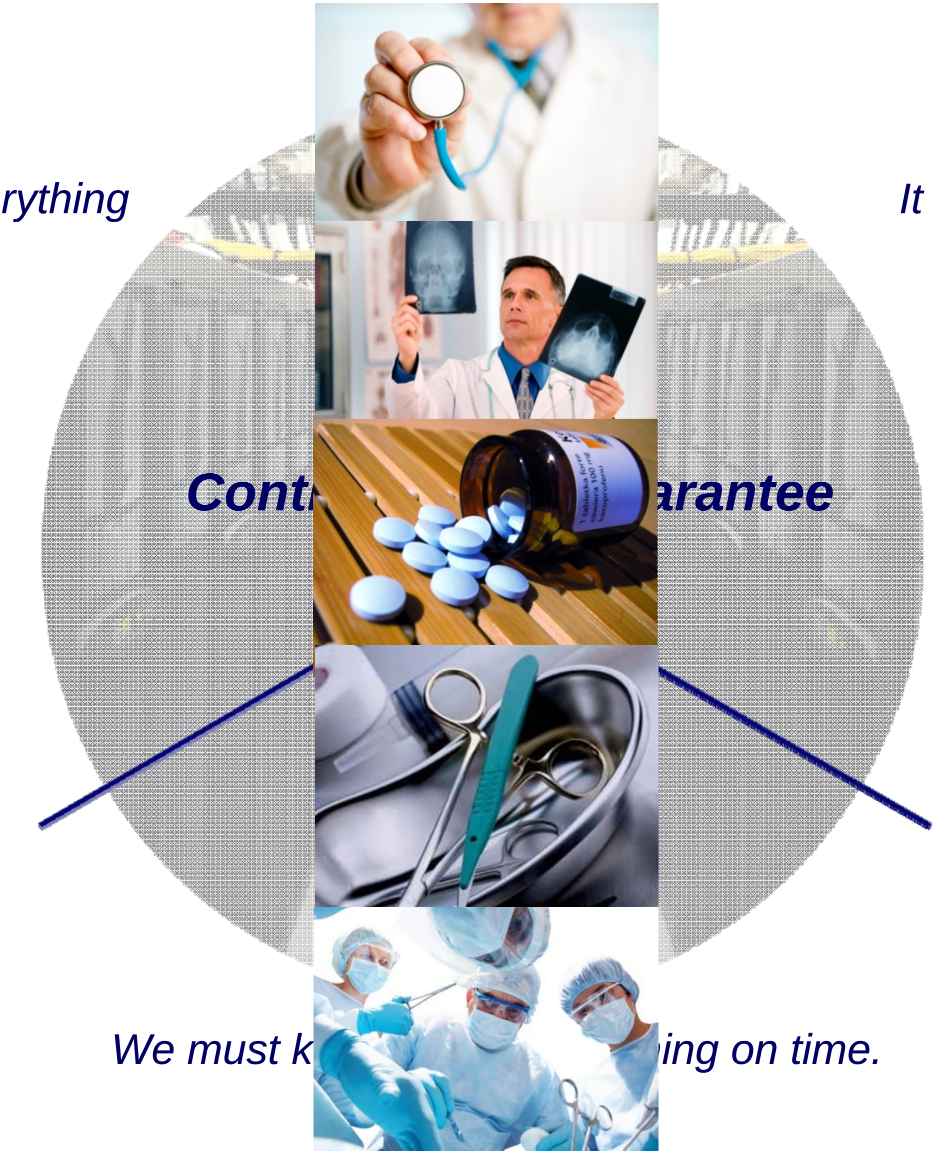


Lomonosov state

Information_on	2014.02.16_18.32
Common_Info	
Users	89
Running_tasks	211
Pending_tasks	178
Allocated_Cores	37916
Idle_Cores	556

Supercomputers: three parts of efficiency

We must control everything what is necessary to control efficiency permanently.



Cont

arantee

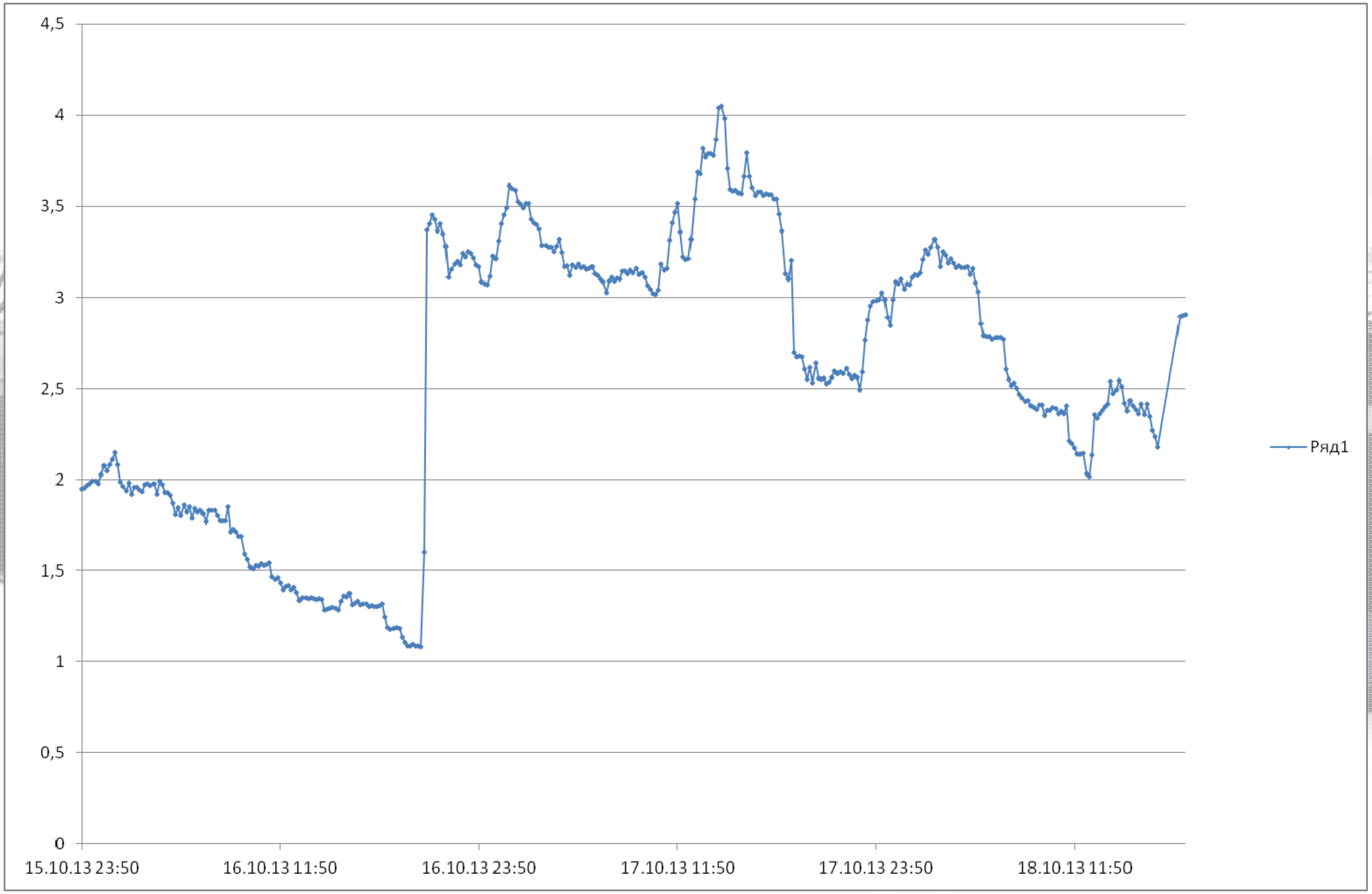
It behaves like we expect, coincidence between theory and practice. Guarantee.

We must k

ing on time.

Efficiency of supercomputing centers

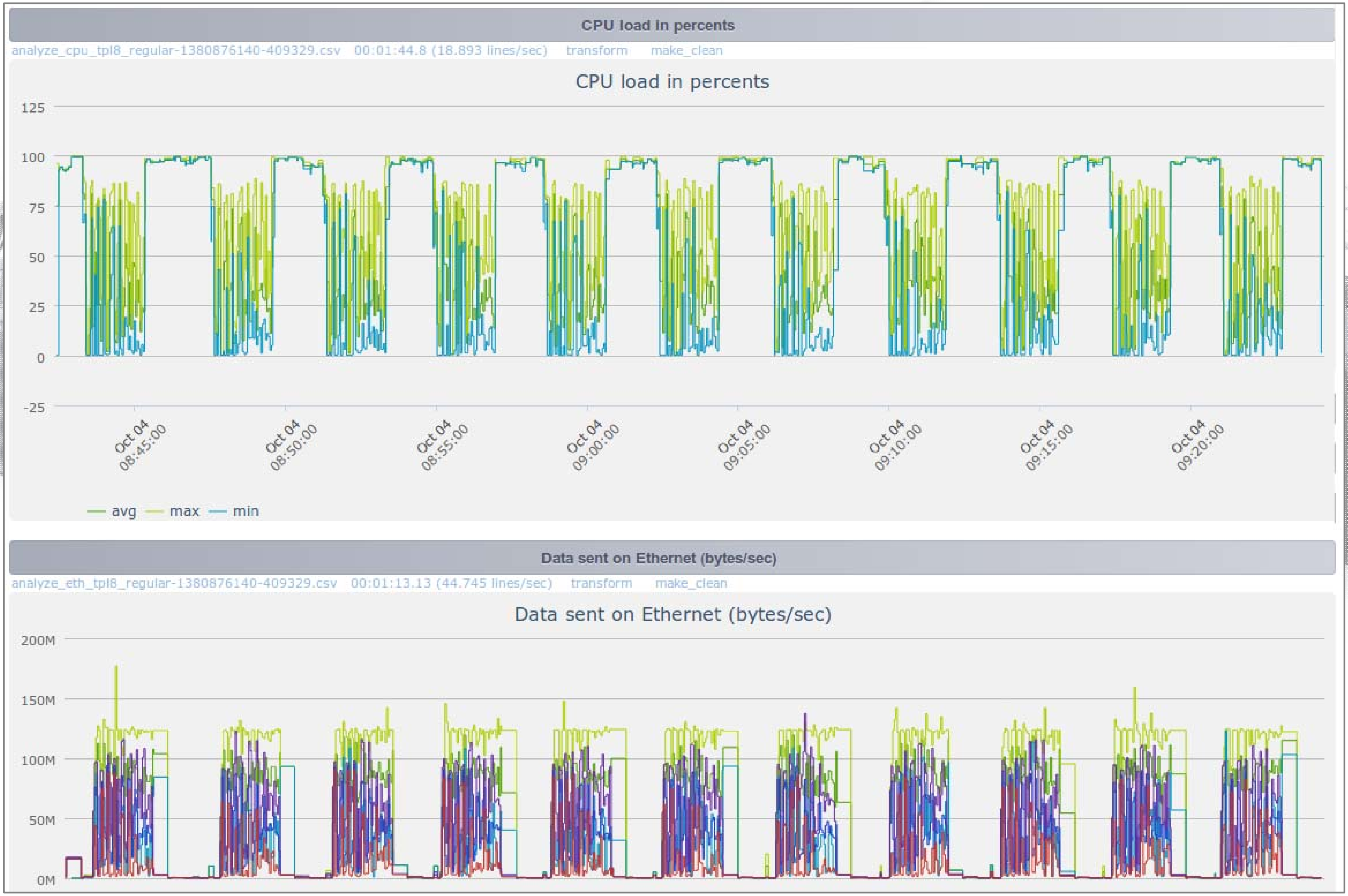
(Integral characteristics)



Average LoadAVG for nodes of “Chebyshev” supercomputer for 3 days

Fine analysis of supercomputing applications efficiency

(control over everything!)



CPU load in percents

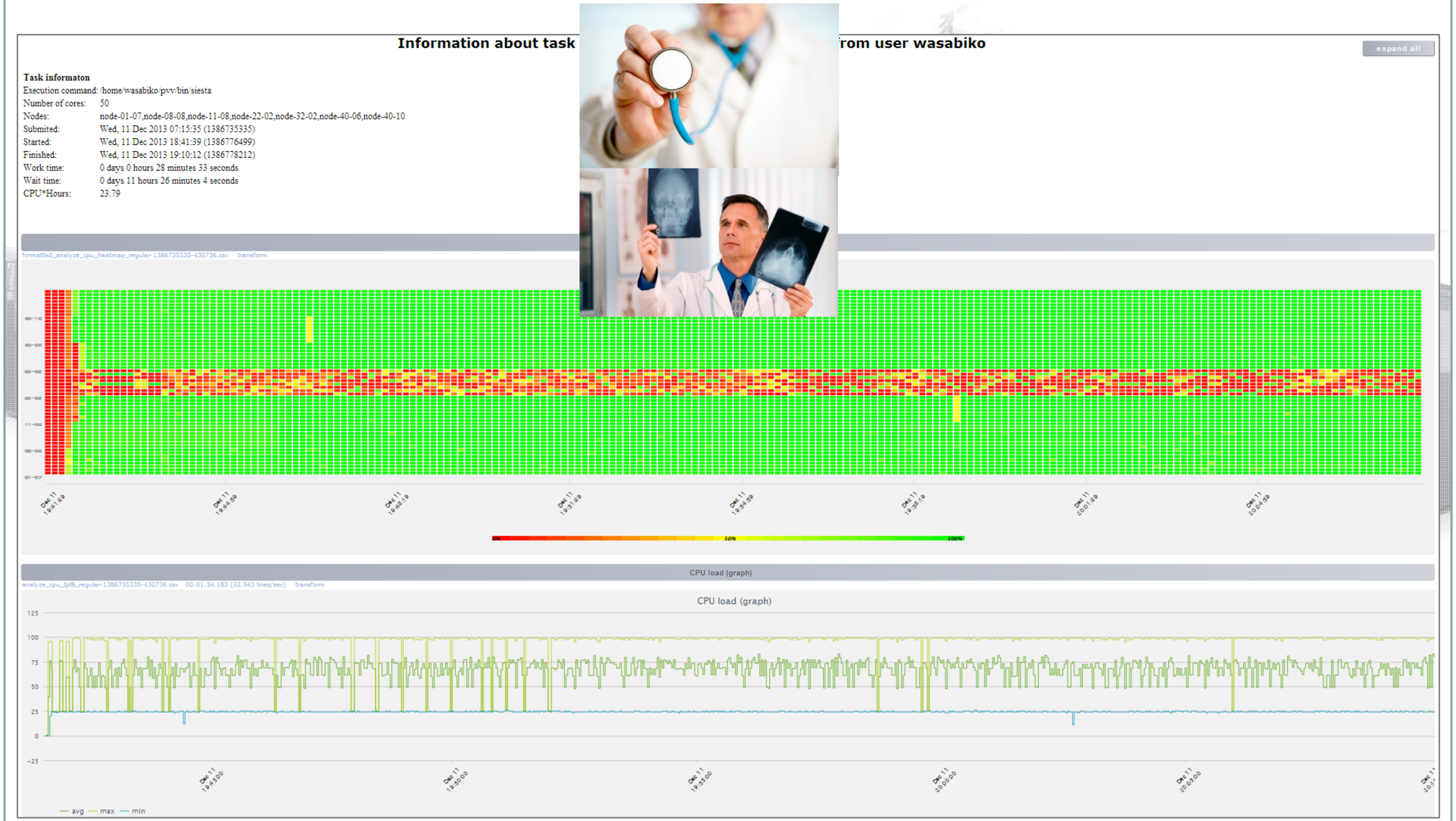
analyze_cpu_tpl8_regular-1380876140-409329.csv 00:01:44.8 (18.893 lines/sec) transform make_clean

Data sent on Ethernet (bytes/sec)

analyze_eth_tpl8_regular-1380876140-409329.csv 00:01:13.13 (44.745 lines/sec) transform make_clean

Fine analysis of supercomputing applications efficiency

(total control)



Information about task ... rom user wasabiko

Task informaton

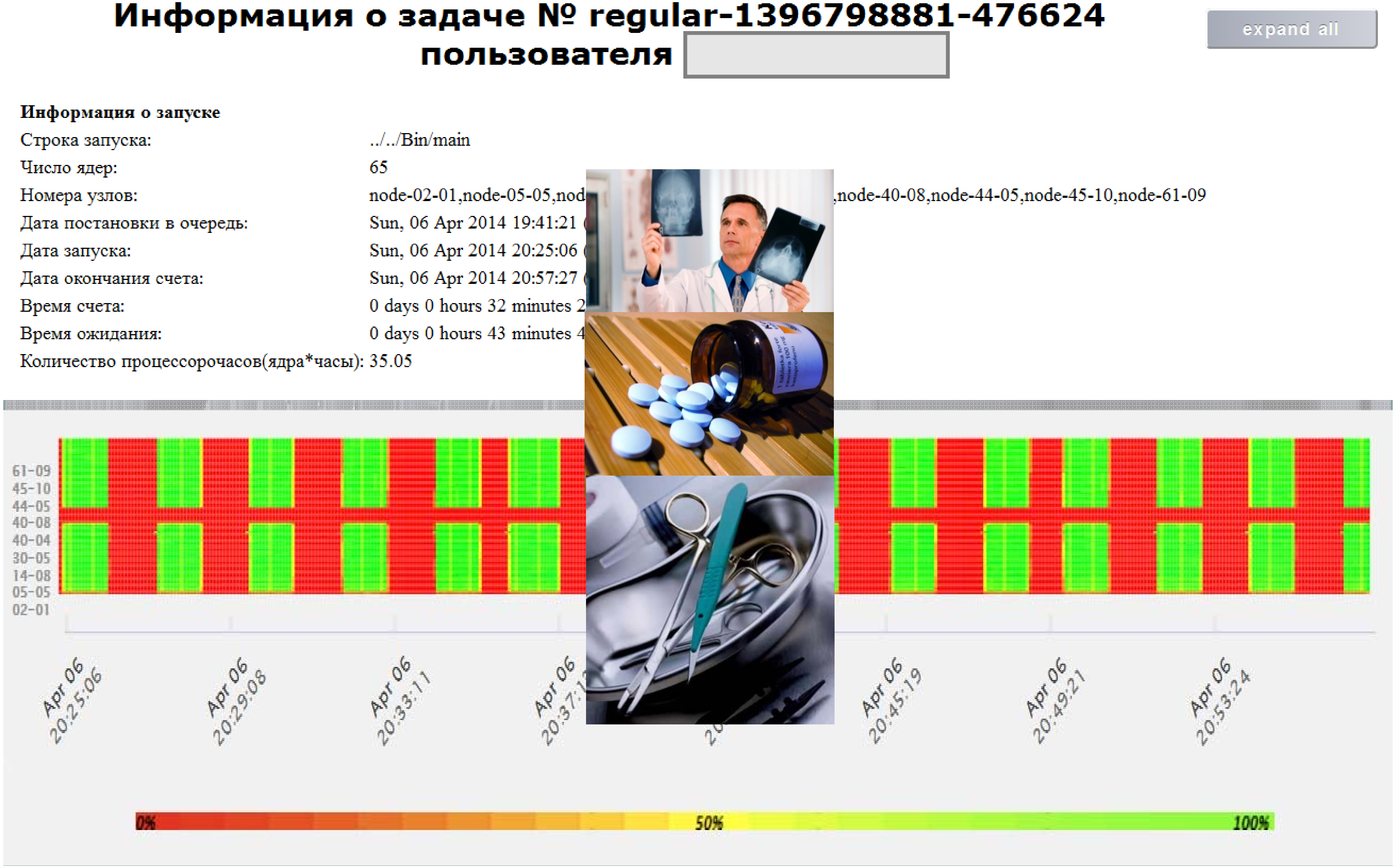
Execution command:	/home/wasabiko/pvv/bin/siesta
Number of cores:	50
Nodes:	node-01-07,node-08-08,node-11-08,node-22-02,node-32-02,node-40-06,node-40-10
Submited:	Wed, 11 Dec 2013 07:15:35 (1386735335)
Started:	Wed, 11 Dec 2013 18:41:39 (1386776499)
Finished:	Wed, 11 Dec 2013 19:10:12 (1386778212)
Work time:	0 days 0 hours 28 minutes 33 seconds
Wait time:	0 days 11 hours 26 minutes 4 seconds
CPU*Hours:	23.79

formatted_analyze_cpu_heatmap_regular-1386735335-430736.csv transform

analyze_cpu_tpl8_regular-1386735335-430736.csv 00:01:54.183 (22.543 lines/sec) transform

Fine analysis of supercomputing applications efficiency

(total control)



Информация о задаче № regular-1396798881-476624 пользователя [redacted]

Информация о запуске

Строка запуска: ../../Bin/main

Число ядер: 65

Номера узлов: node-02-01,node-05-05,nod ,node-40-08,node-44-05,node-45-10,node-61-09

Дата постановки в очередь: Sun, 06 Apr 2014 19:41:21 (

Дата запуска: Sun, 06 Apr 2014 20:25:06 (

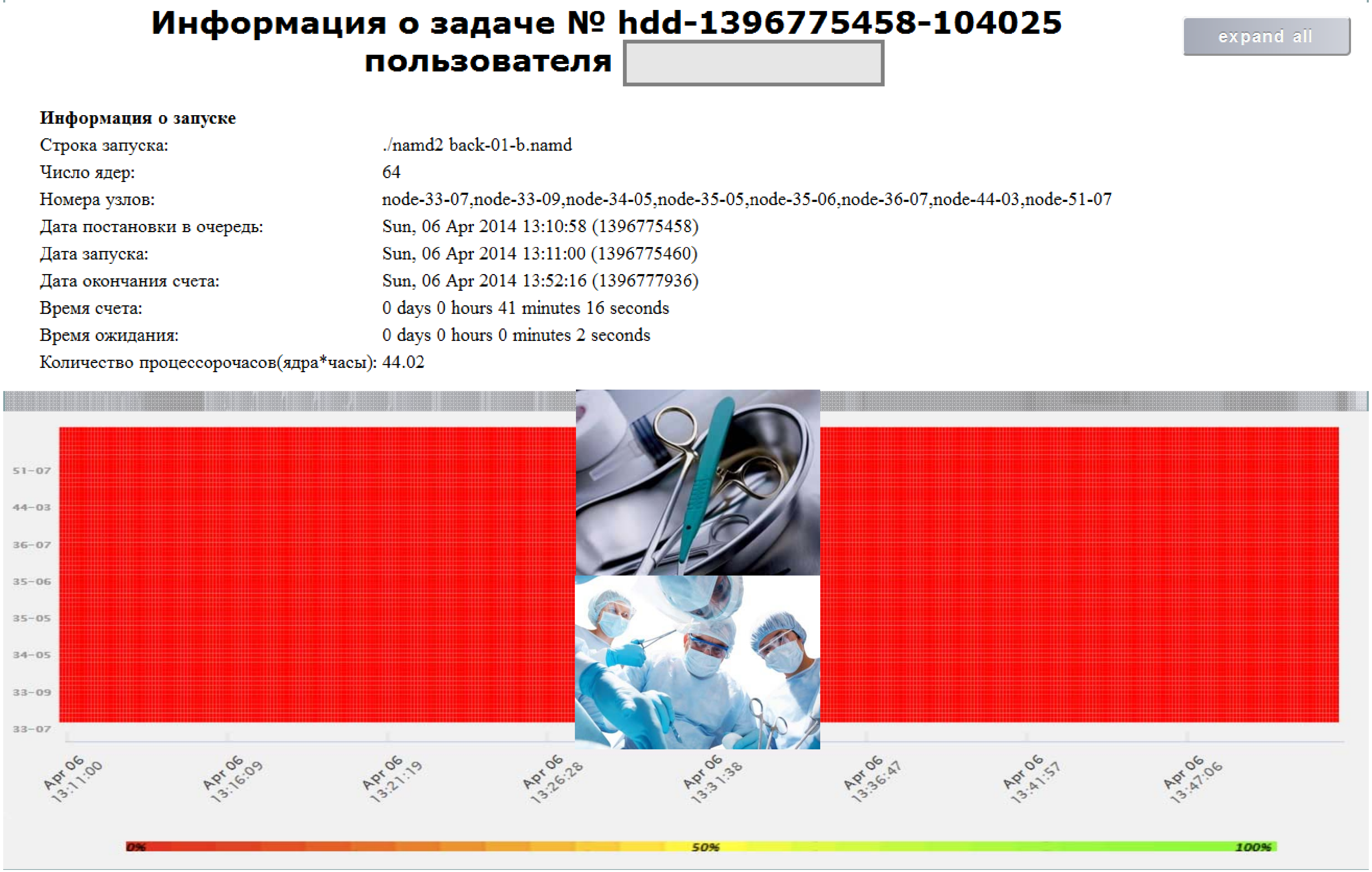
Дата окончания счета: Sun, 06 Apr 2014 20:57:27 (

Время счета: 0 days 0 hours 32 minutes 2

Время ожидания: 0 days 0 hours 43 minutes 4

Количество процессорочасов(ядра*часы): 35.05

Supercomputing applications: symptoms of losses



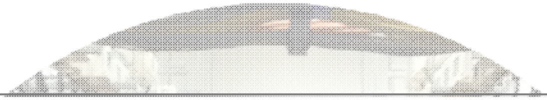
Информация о задаче № hdd-1396775458-104025 пользователя [redacted]

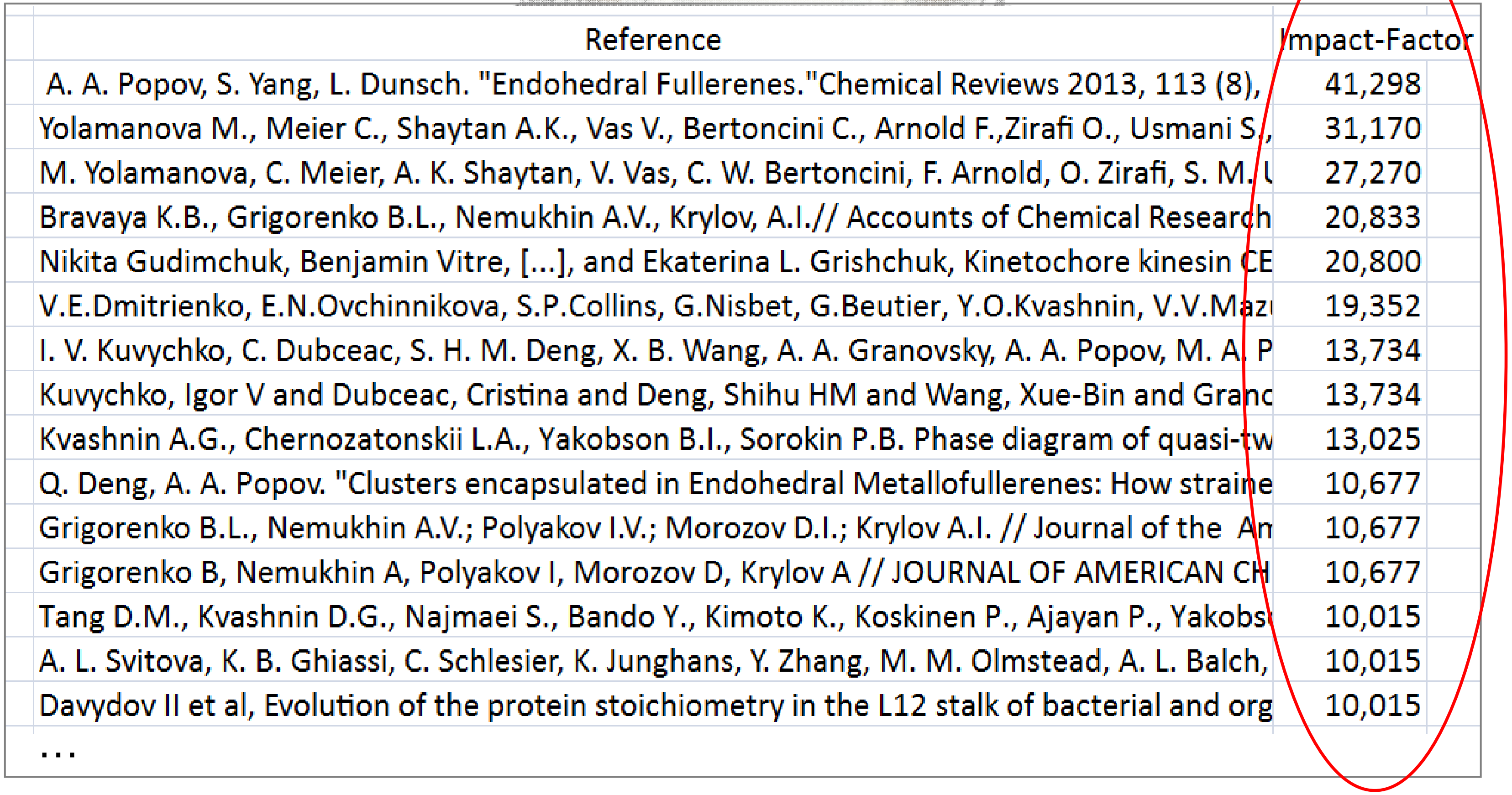
Информация о запуске

Строка запуска:	./namd2 back-01-b.namd
Число ядер:	64
Номера узлов:	node-33-07,node-33-09,node-34-05,node-35-05,node-35-06,node-36-07,node-44-03,node-51-07
Дата постановки в очередь:	Sun, 06 Apr 2014 13:10:58 (1396775458)
Дата запуска:	Sun, 06 Apr 2014 13:11:00 (1396775460)
Дата окончания счета:	Sun, 06 Apr 2014 13:52:16 (1396777936)
Время счета:	0 days 0 hours 41 minutes 16 seconds
Время ожидания:	0 days 0 hours 0 minutes 2 seconds
Количество процессорочасов(ядра*часы):	44.02

Efficiency of supercomputing centers

(what is efficiency?)





Reference	Impact-Factor
A. A. Popov, S. Yang, L. Dunsch. "Endohedral Fullerenes."Chemical Reviews 2013, 113 (8),	41,298
Yolamanova M., Meier C., Shaytan A.K., Vas V., Bertoncini C., Arnold F.,Zirafi O., Usmani S.,	31,170
M. Yolamanova, C. Meier, A. K. Shaytan, V. Vas, C. W. Bertoncini, F. Arnold, O. Zirafi, S. M. U	27,270
Bravaya K.B., Grigorenko B.L., Nemukhin A.V., Krylov, A.I.// Accounts of Chemical Research	20,833
Nikita Gudimchuk, Benjamin Vitre, [...], and Ekaterina L. Grishchuk, Kinetochore kinesin CE	20,800
V.E.Dmitrienko, E.N.Ovchinnikova, S.P.Collins, G.Nisbet, G.Beutier, Y.O.Kvashnin, V.V.Mazu	19,352
I. V. Kuvychko, C. Dubceac, S. H. M. Deng, X. B. Wang, A. A. Granovsky, A. A. Popov, M. A. P	13,734
Kuvychko, Igor V and Dubceac, Cristina and Deng, Shihu HM and Wang, Xue-Bin and Granc	13,734
Kvashnin A.G., Chernozatonskii L.A., Yakobson B.I., Sorokin P.B. Phase diagram of quasi-tw	13,025
Q. Deng, A. A. Popov. "Clusters encapsulated in Endohedral Metallofullerenes: How straine	10,677
Grigorenko B.L., Nemukhin A.V.; Polyakov I.V.; Morozov D.I.; Krylov A.I. // Journal of the Am	10,677
Grigorenko B, Nemukhin A, Polyakov I, Morozov D, Krylov A // JOURNAL OF AMERICAN CH	10,677
Tang D.M., Kvashnin D.G., Najmaei S., Bando Y., Kimoto K., Koskinen P., Ajayan P., Yakobs	10,015
A. L. Svitova, K. B. Ghiassi, C. Schlesier, K. Junghans, Y. Zhang, M. M. Olmstead, A. L. Balch,	10,015
Davydov II et al, Evolution of the protein stoichiometry in the L12 stalk of bacterial and org	10,015
...	

International Advanced Research Workshop
on High Performance Computing
from Clouds and Big Data to Exascale and Beyond

Thank you!

July, 11, 2014, Cetraro, Italy