



JobDigest – Detailed System Monitoring-Based Supercomputer Application Behavior Analysis

Dmitry Nikitenko, Alexander Antonov, Pavel Shvets, Sergey Sobolev, Konstantin Stefanov, Vadim Voevodin, Vladimir Voevodin, Sergey Zhumatiy

> Research Computing Center Lomonosov Moscow State University

26.09.2017 Moscow, Russia



Moscow State HPC Facilities





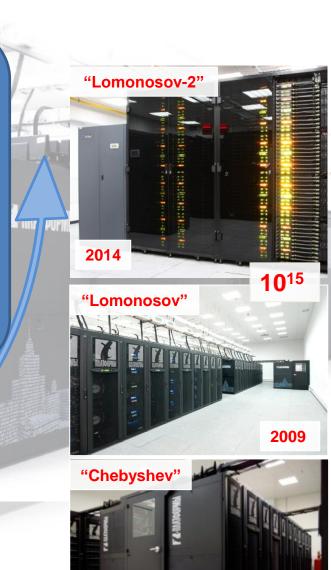
"BESM-6"



MSU HPC Center 2017 Q3

- "Lomonosov-2"
 - 2,9 PF peak
 - 42688 cores
- "Lomonosov"
 - 1,7 PF peak
 - 82468 cores
- Some 10x TF systems
- Advanced Infrastructure



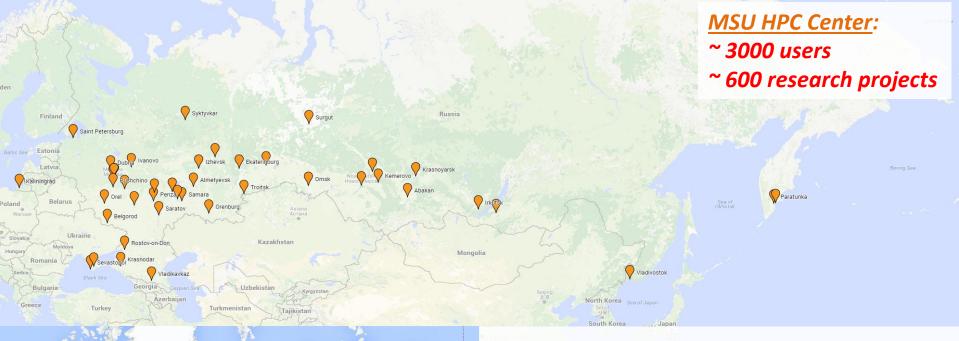


2008

10¹³



Scientific projects of MSU HPC Center The Largest HPC Center in Russia



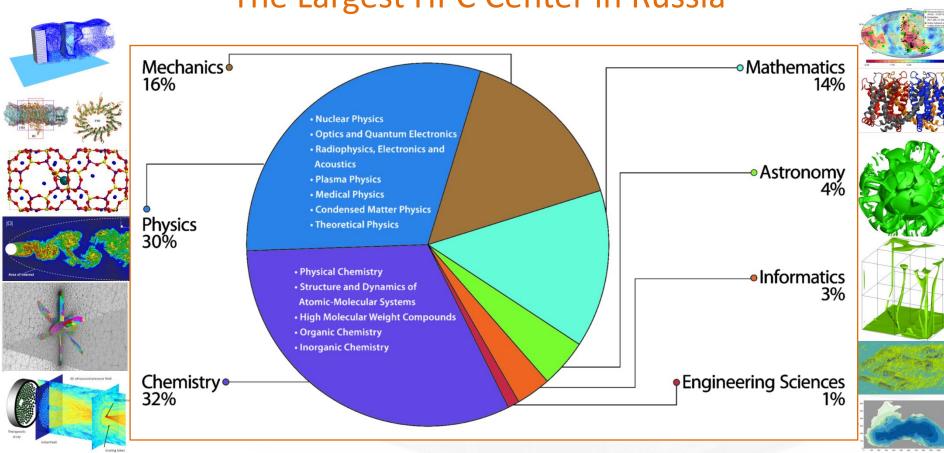


• Scientific groups from 350+ organizations and institutions allover Russia (~50 cities)

 Collaboration of leading teams and specialists all over the world:
 ~150 collaborative projects with international scientific groups from over 90 locations worldwide



User Applications of MSU HPC Center The Largest HPC Center in Russia

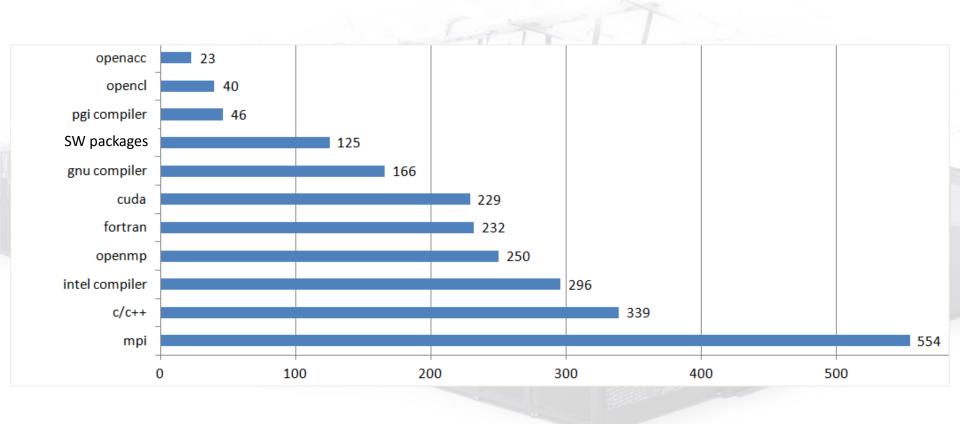


TOTAL DIVERSITY:

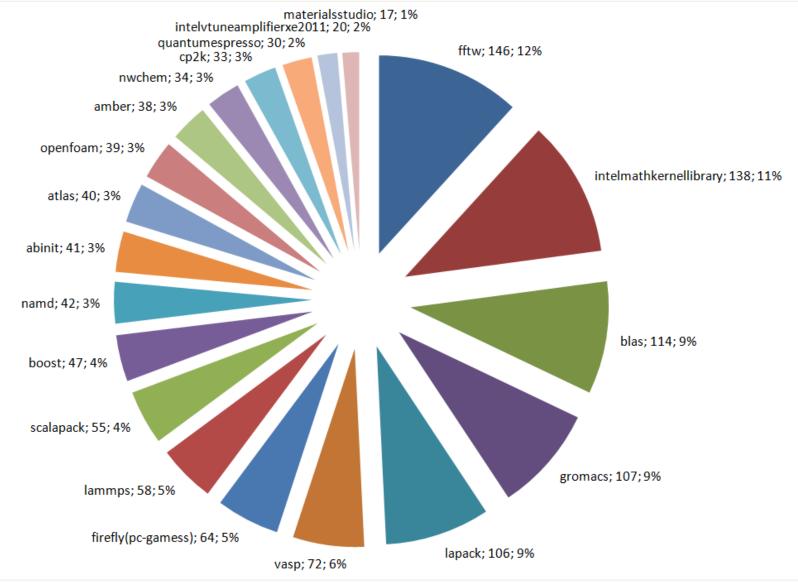
- Programming Languages & Technologies
- Algorithms & Implementations
- Hardware & Software Envolved



Computing Facilities of MSU Diversity of parallel programming technologies in use



Computing Facilities of MSU Diversity of software in use



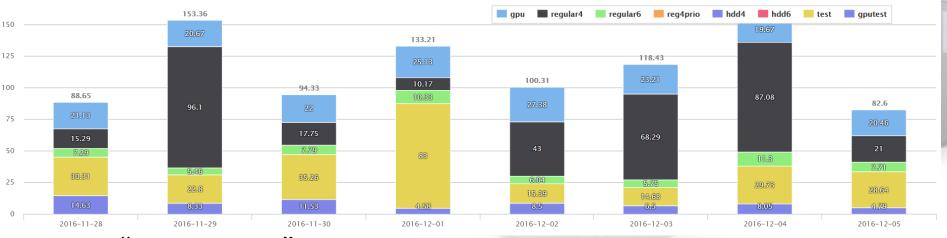


Operability and productivity

Monitoring and automated reactions needed

Idle system: expensive Inefficient load: expensive Late decisions: expensive dangerous





Recent "Lomonosov" stats: 100-150 simultaneously run jobs, over 1000 jobs per day



Operability and productivity

Everything plays role and contributes. Total monitoring.

It is imperative to permanently keep track on all components that influence efficiency of large-scale system output

- Computing HW: nodes, CPUs, memory stack, disks and storage hierarchy, networks, etc.
- Infrastructure HW: much more fault-tolerance critical. It includes cooling system: chillers, heat exchangers, air conditioners; piping, pumps; a set of components of the power system in conjunction with an uninterruptible power supply; fire safety systems and smoke removal; access control; etc.
- Whole SW stack: OS, package and license usage rates and peculiarities, etc.
- Job scheduling and queuing details from various points of view.
- Job execution details: dynamics and peculiarities of resource utilization by any and every executed user job.
- Users. Peculiarities of resource utilization, using available SW, cheating, etc.



Total Monitoring

Scopes of analysis

Time period:

• Past

post-mortem analysis: performance analysis, event processing efficiency

7 & INATODAK

Present

quick and immediate reactions, present state screening

Future

prediction and planning

Role-specific viewpoints and interests:

- User
 - Regular User
 - Project Manager
- HPC Center
 - Administrator
 - Expert / Supervisor / Reviewer
 - System Holder



Total Monitoring

Data Sources

 Recent events and HW/SW state streaming data for performance analysis & quick reactions (monitoring systems: DiMMon, Collectd, Nagios, Zabbix, etc.)

P 4. INUTIONIS

- Compute nodes
- Infrastructure
- Historic coarse-grained data (DataBases, CSV, JSON, logs, etc.)
 - Resource Managers (SLURM, CLEO, etc.)
 - Resilience Systems
 - User/Account Management Systems



Total monitoring of HPC systems

All-round and Holistic Approach practiced at MSU

OctoShell

HPC Facility Management System: Accounting, Research Project Management, Helpdesk, etc.

OctoTron

Reliability of HPC Center Functioning: Rules/Reactions on Occurring Events Based on Formal Supercomputer Model

OctoStat

Periodical Statistics on User Activity, HPC System Load, Queue Structures, etc.

OctoScreen

Show Everything We Need to See and Know in Various Ways of Presentation

JobDigest

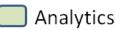
Detailed Analysis of Application Behavior, Resource Utilization Dynamics and Job Categorization

DiMMon

On-the-Fly Reconfigurable **Di**stributed **M**odular **Mon**itoring system

🔵 Data Source

Shell / Access for Regular Users





Approach Design Principles

Key feature - availability of sufficient info on every job

- General information on every finished job must be available, including data from resource manager and average rate of resource utilization (integral job characteristics) obtained from the monitoring system.
- The used monitoring system (sensor set, polling frequency and saved data coarsening) must be configured to grant availability of job profiles for all finished jobs right after execution with no resource-intensive post mortem operations.
- There must be means for **job marking and categorization** based on job characteristics in manual and automatic modes.
- Job information access restrictions must meet workflow regulations, supporting various scopes of analysis: regular user, research project manager, system administrator, etc.
- Flexible configuration, supporting diverse data sources.



General Job Characteristics

Key feature - availability of sufficient info on every job

General Job Information

- general job queue structure analysis (OctoStat)
- stats on user/workgroup resource utilization (OctoStat, OctoShell)
- based on resource manager data / logs

Dynamical Job Characteristics

- represent the dynamics of resource utilization during application execution
- application execution resource utilization dynamics analysis, finding abnormal program behavior (JobDigest)
- based on resource manager data + system monitoring data

Integral Job Characteristics

- represent average resource utilization during application execution
- Resource utilization analysis for job collections (user/workgroup, partition, set of nodes, etc.)
- based on resource manager data + system monitoring data

System monitoring:

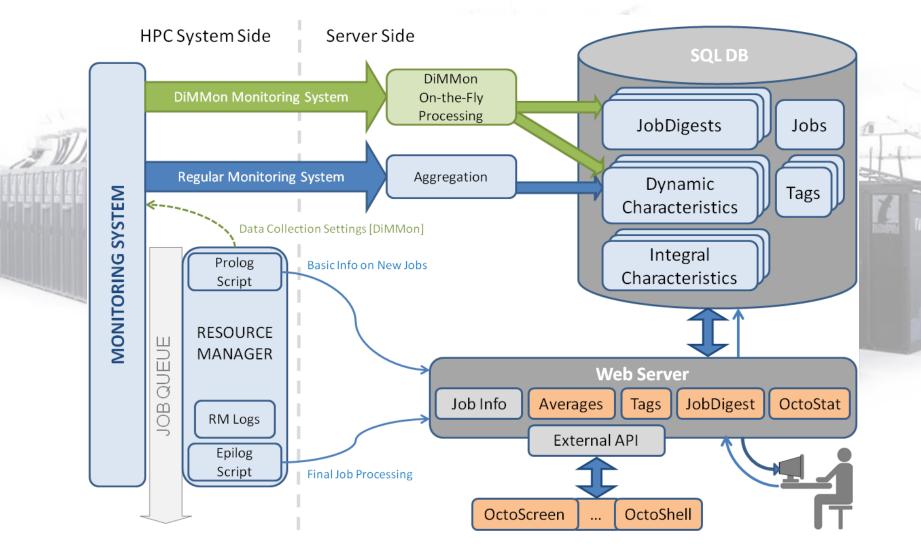
- CPU user
- Flops
- LoadAverage
- Interconnect usage
- L1 replacements
- LLC misses
- Memory access rate
- GPU load
- etc.

Resource Manager:

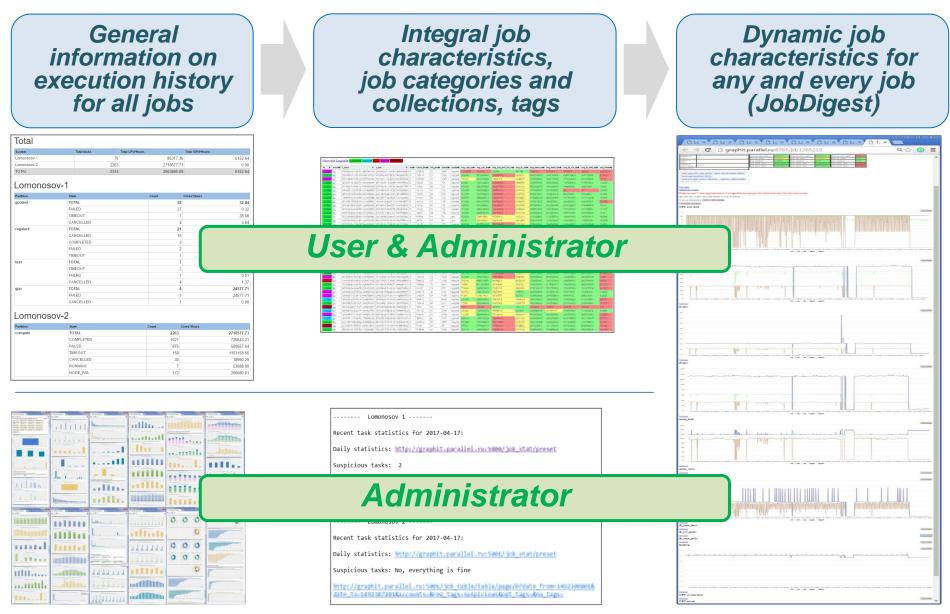
- ID
- Owner
- Status (finished, cancelled, etc.)
- Timestamps (queued, started, finished)
- Command line (to parse for details)
- Partition
- Node set allocated
- Allocated number of cores
- Execution time and CPUh

Approach Design Principles

General workflow based on system monitoring data analysis



Job and Queue Structure Analysis Hierarchy



26.09.2017



Job History and Summary

Available for every user in personal account web page

Ability groups:

• Regular user

- personal accounts only
- Project manager
 - all accounts of the workgroup
- Expert
- Administrator

Project:	per-auerasa	2014		
		регутиция фу ем высокопр		
Project: Pyran	 Figmetti oxo 2013 	s dimension	8 5	8
UniCheck all Start date 2016-02-07 End date 2017-04-17 Show				

Total

System	Total tasks	Total CPU*Hours	Total GPU*Hours
Lomonosov-1	70	85317.36	6152.64
Lomonosov-2	2263	2718577.73	0.00
TOTAL	2333	2803895.08	6152.64

Lomonosov-1

Partition	State	Count	Cores*Hours
gputest	TOTAL	38	32.84
	FAILED	27	0.32
	TIMEOUT	7	28.68
	CANCELLED	4	3.84
regular4	TOTAL	21	60697.23
	CANCELLED	15	35631.08
	COMPLETED	3	7559.38
	FAILED	2	6582.61
	TIMEOUT	1	10924.16
test	TOTAL	7	9.59
	TIMEOUT	2	8.21
	FAILED	1	0.01
	CANCELLED	4	1.37
gpu	TOTAL	4	24577.71
	FAILED	3	24577.71
	CANCELLED	1	0.00

Lomonosov-2

Partition	State	Count	Cores*Hours
compute	TOTAL	2263	2718577.73
	COMPLETED	1021	726643.21
	FAILED	875	500657.64
	TIMEOUT	158	1153158.56
	CANCELLED	30	18950.29
	RUNNING	7	53088.00
	NODE_FAIL	172	266080.01



Workgroup Job Collection Details

Research project jobs summary

TIME PERIOD	2016-01-01 18:47:05 2016-12-09 18:47:05
LOGINS	shvets vadim coctic serg
TOTAL CPU TIME (CPUh)	805.79
USER JOBS	178
COMPLETED	167
CANCELLED	4
TIMEOUT	0
FAILED	7
NODE_FAIL	0

"Lomonosov-2": CPU load, GPU load, LA, IB MPI, IB FS

									N N	/			/	
id	account \$	t_start \$	t_end \$	state \$	cores_hours \$	num_cores 🖨	duration 🖨	partition \$	cpu_user \$	gpu_load ♦	loadavg 🖨	ib_rcv_data_mpi	ib_rcv_data_fs	\$
122141	<u>shvets</u>	2016-12-09 17:48:13	2016-12-09 18:33:51	COMPLETED	42.59	56	45.63	compute	46.26	35.73	13.96	1030631127.20	1030724468.14	
122062	vadim	2016-12-09 15:50:32	2016-12-09 17:50:51	COMPLETING	28.07	14	120.32	compute	0.04	0.00	0.03	0.00	276.98	
122064	vadim	2016-12-09 15:50:33	2016-12-09 17:50:51	COMPLETING	28.07	14	120.30	compute	0.11	0.00	0.02	0.00	279.83	
122084	shvets	2016-12-09 16:58:45	2016-12-09 17:38:18	COMPLETING	9.23	14	39.55	compute	0.04	0.00	0.00	0.00	187.13	
<u>122049</u>	<u>shvets</u>	2016-12-09 15:45:11	2016-12-09 16:23:57	COMPLETED	36.18	56	38.77	compute	45.87	25.99	13.98	939884640.36	925061313.08	
<u>122072</u>	vadim	2016-12-09 15:58:46	2016-12-09 16:09:24	COMPLETED	2.48	14	10.63	compute	18.24	0.00	13.98	0.00	485.15	
122073	vadim	2016-12-09 15:58:47	2016-12-09 16:08:56	COMPLETED	2.37	14	10.15	compute	19.45	0.00	16.03	0.00	6328.39	
122070	vadim	2016-12-09 15:52:10	2016-12-09 16:05:58	COMPLETED	3.22	14	13.80	compute	11.80	99.00	8.56	0.00	457.38	
122067	<u>vadim</u>	2016-12-09 15:58:45	2016-12-09 16:05:51	COMPLETED	3.31	28	7.10	compute	38.35	0.00	15.57	152445207.65	133954687.73	
122071	<u>vadim</u>	2016-12-09 15:55:01	2016-12-09 16:05:33	COMPLETED	2.46	14	10.53	compute	19.16	0.00	12.30	0.00	494.26	
122063	vadim	2016-12-09 15:52:05	2016-12-09 15:59:56	COMPLETED	7.33	56	7.85	compute	47.72	0.00	17.33	187262462.26	187675724.51	
122066	vadim	2016-12-09 15:52:06	2016-12-09 15:59:54	COMPLETED	3.64	28	7.80	compute	24.66	99.12	11.44	164043411.16	166637443.66	
122069	vadim	2016-12-09 15:52:08	2016-12-09 15:59:20	COMPLETED	3.36	28	7.20	compute	45.34	99.00	19.73	152720283.82	147813432.90	
122068	<u>vadim</u>	2016-12-09 15:52:07	2016-12-09 15:59:14	COMPLETED	3.32	28	7.12	compute	39.08	99.29	18.35	226869299.78	229201610.44	
122065	<u>vadim</u>	2016-12-09 15:50:34	2016-12-09 15:58:37	COMPLETED	7.51	56	8.05	compute	50.04	0.00	21.64	167065425.74	167085160.13	

Highlighting statuses (left column) and average resource utilization (right colored block)



Job Details

Information on a job collection

	t Cstart 0	t_end 0	state co	res_houds	req_prods	duration	partition	avg_cpu_us#r	avg_cpu_flogts	avg_cpu_perf_Hd_repi	avg_lic_misk	avg_mem_load	avg_mem_store	avg_ib_rcv_date	avg_ib_xmit_date	avg_loadaw
and the	· 19978-550-14-14-526.35.1989529497101	Streamwer, managed and ser	1 109	98.90	192	éges	Anicade .	LASHERD.	1011308	K20796#	8512.789	A VALUE A	States.	Helmist	THEY BU	S Santage
111 a	STRATES CONTRACTOR STRATES	10716-009-016-01509-016-01646-002-97678-0	3 174	PR 64-	\$	234	anguine.	Maries)	15x5x445x60	1800495	19,0,000	Water-death	wig offers	41005.00010	4140 spins	12.22.5
and and a	SPITUS IS IN SEARCHARD	structure tracky strate topologi	7. 124	bet dal.	4	1,210	-station to	No. Cast	anicity states	100421918	- managed to	nandana	ANANCI SAMPLE	225344000.0	Street Bank	14.150
100	Benefit in the set of	WEATHING CONTRACTOR	3 14	6-94 6-94	44	÷32.	100010376	el tre	Manager	-specification	HIGHNAN	CTTS BLACK	ARACTORNEL	-minangles.	au heist miles 2	Section.
11 A.	Line segurity representation	peter a consideration.	17 14	ang.		14	helipida	and the second	SERIES	This Tak	-	all and the second	1022-002600	electrolytically	JULDPHILDISE	tistar!
2005	244-06-012036-06-04-004204	Purchasing to 10 spectra manager	F 144	STATISTICS IN THE OWNER	Sec.	Contract of the local division of the local	veticies		MARCHER	Independ	VERTON	211-199005	Calibration .	30	15	14-26.31
100	2015-District Of Tax 25 Char Lebens	find on the property of the subscript	C. N.	1991 C	11	545	Pango		Alexandration of	DOMESION OF	visting-	1100000000	aballion 5	Relativity-1	and a consider	Sharing .
		muche stag waget suggesting.		A data		18	ADCENT		MARGONALL.	Lawrence .	PONST	APRICE A	Printerant	1517-0499520174-	24121234	VERTER
110 .00		water and the second state should be		e ja	12	3			descent fit	eligipetterit	TRUENA N	services.	Tales Conford u	Kithefall Anto	Autor Sciles	74.55608
	the second s	problem the start of recommendation	and the second second		16M	- AL	Mariase		Sectore States	1745-BACK	America	STATES A	Distriction	There albert	annessentian.	28:515
	 A second sec second second sec	met die nat in 19. other blaink		100	and the second	-161	Insulard		99990 98498	attelete.	Shidow S	minet-crin d	alaren alar	and indiana.	COST INTERIO	in sites
		2110-06-10-20-2-20-144456-2000		and the second second	Serie - Chief	NT.	10148		DESCRIPTION OF	CHARGE BOAT	Weights	COLUMN SCO.	anter Statute	Annagha	Sentingeners	10.45
and the second		JS-1-RE-MACHINERIC A-44 (RE-ADA)	1	STR		24		AR GET M	Niettwiltig	41.568.4	CARLES	DOM NOWS	A STREET	-	-572-675-01-2	19-2648
		desidenti 13 de transmistérics		a participante de la composición de la La composición de la c	12	na starter.	3Portante	THE ALL		TORYTH &	Mysilla	SO/GOUDING	instated a	diameters.	VERSIEGEN	-110/16.
100	when a comparison of the data when a service of the			and the second		100	1250	261 PM	altreatures of	and constitution of the second	470364		Catalana	20134267.0	Souther Marter	14940
and the second		Standard and State which Prove Early			Stranger and		Jetalija-	100000000000000000000000000000000000000	Contraction of the local sectors of the local secto	Parts Allen	Street of Articlastics and	17.9603688.8				
112 (d		and the standard standa		180		揷	考生	NADIA.	Mandalania	3/44/8455	Harton .	Assemble of	** detribute	32951345-6K	城市市中学新 动	SANERA.
107		stuardade is the suit of indiana.		6座	10	51	100	教保治	31596-0590-26	Ar politica	MAGENE	Alexandra d	(1)-19-365次	383465503	封空后的中国。以	12 Yest
ALL VILLE	and share the sherry regeneration bulls, reserves	Standard at State considerable			12.00 million (1997)	10	-490.04	御 林道	9000073994-b	- English	465704	WHEN PROVED IN	2914margaro	2010/10/2013	2000-200-022	导导性
111 St.	and the second sec	Stratefices , age to wrapplishe			3A	g*5	gen.	\$1.23%	105-108-2	\$95546-n	国际的文学	September 8	新新的 生物的这	42400740636	etupitx@da	4卷535支
2014 99		· 清洁的一语 化和内 124年间的新生		190	51	and the second	39494	1770/58		-4%2.05	305 A	#19第1 年 的早時	STREEPINGEN	Statestar.	(desided))	\$3808A
2270		Talanceren (E.F.3) (1944) abarros			11	1997	2014	ALC:NO	ALL AND ADD	2212日48	garrais i	-alitherations	SCHWERE'S	Managerington	white an apply	和副創業
2222 (124)		的传导性关系的行行性情望 的一		10 Yal	2	3945	1997	Name.	非常的新闻	3799:04	Circlest.	Esternetica.	米勒东海峡军兵	president-	24147084	ANDING
100 B2	Maay dente to dente 1946 vandere	We could be the up of the end through	a . 16	W.	22	7,4	dia -	astated.	7.3 MidEEST	showing.	-HERERY MA	ing and a state	*****	8-464-900ch	2.2000年1月2日	193664
2222 74		"如何不必必要的",不会这个人不能。他们却将他		34	21	读	(LPA)	in syn	的影响和教	diffeditional	20456.50	70110999836	Landberglagets	33644694	2141220451	96.80959
224 a.	mentional and a complete addition	Application of all their care of the sec-	2 12	24	W.	42	20	AN 8622	4425449504	angezang a	特的编辑	学说学生制的教育	· 法由和法律性	2400 and a fail of	対応ない意味	at they
Me.	approximite the second s	WE LOOK TO AN INCOME THE REAL PROPERTY OF	4 14	19.35	01j	39:27	-mapping-	44,035.4	新动物的动	194166-3	1055 mg	NIND MARRIES	programme and	影响和自	KNOLOGE &	12.20
400 (c)	frage entry servicements p	and and shares intermet	3 40	2.44	18.04 257	4)¥-	140.984	distances	141726 BL2565	17385/AHK	53454	Maximina -	100000000	A87933003.4	· 4.1%366331.0x	34-342
100	sources of place to considering the	· 通知教育的 [15] 中国 [12] 中国 [12] 中国 [12] 中国	1 20	19,191	34	*10:FG	-Mariana	48.1155	SPACE TRANS	四/他同時在多	graphete.	编码的编码	1346-144A	ACCESSION OF THE OWNER.	Sector address	深行神经
Sx.	A particular of the second second second	an water and an even and the second	6 Pag	1454	12	NUT	odploaff.	2月-12285	errection.	N/ARGO (NUCH	134199.7	aversident	2014年1月10日	123010-002-2	SERAMORA.	Syran .
10 100	Subsection to Hat Disposition	UNITED STOF STORE , AREASSON	e : 16	98.92	12	194255	register8	247.94	and the second	pinealage. A	Starph.	(Casabbingo)	taniense.	Anterest and a second s	34007905443	Direkt?
211 2.	2014 Research of the state of the	Northgia 28 (Reads) Tradepier 171	\$ 11	asian .	32	34.01	nactional.	Ini filie	And Address of	santateans	No. Property	Westeriosy.	Triestoriana.	-Manmanicky	463/042/03	11 12 10
1015 20	Top: Soliga to aparts the provide	plate products bar for I prochadas.	8 Las	AUR .	12	135	-aguint	an Jagit	VARASIZATE	Philesipe.	Saleatra	335310002/2	STREPHOND &	1100000048	Antonio a	a start
13		16-54 Add parts apie SHEND		0.5	33.	ine		-16 gT-(-	transferration	der istere	>>3434.8.	BRETTINGATED	Lingdom A.G.	519641-052	0845388-381-05	Tripple
200		(industrial of the second of the second of the		05.49	C. Sciencing room SEA	and -	11821760	4.005	VISING AT	normation	141025	Gouldenande b	Q 5500 1015	agarethe a	AND A REAL	NAMES OF
ALS STREET	and yours or define - phone but may reduct building allow re-	STREET BELLEVILLE	statute and re-	AR OF	1. S	2:00	Mariate		Response	en423/16.3	1610003	The strate	LONGROUPS.	an analysis of	Etangolita is	-
THE REAL	and in the second set of the second sec	Sexultar states and the		50.78	1	1395		Autoria	Anternational	esian zigea	and the	minakasi	Louistinge a	Pla Ponecka	29/2004/03/g 1	
- 10 March		with the Carter and a state of the		an a	S. Caldenna	-1	NADOWS	STAY	Less feet	and water	3741-19	10 march	125424	Nave No.	- Andrews	10.00
A 200 - 100 - 100		Service and the Lord American		14.55	de la composition de la compos	and the		Thursd	17294622-5	MARRIA	6,2452	104204445	Last make	MERCONNE	Tr schätusti 5	COUNT
-	and the second s	WHICH IN TERMOLTS ADDRESS	the Lot Distance	144	14	4)22	NOT THEY	11.1948	Manufacture o-	- Eligano	2050415	TENANDIN S	schelabilite	Juanti 4 alburts	-declaration	22200
A 12		Consulta di deservisi antesis consultaria		RTA C	10	8-75	of June W		-Josefferda	Proto-state	26.400	N-Balakas-V	Stratists.	pustradice.	allers toward	Strates.
THE R. C.		Light of the residence of the second second		10.77	44	John	10,000		Contraction in		JOHLAN E	NEW AND AND	Linester		SOF YMERIC	AL LONG
and the second	and a second second day and so in the second second and			12		and a second				- ACUCARDINE	and the second second			4月1日3月9年		
100	The shallow alteration with a the st	Springer and the state of the s	F	18	ișt	P	report.	An other	-254973969624F	中非常的	610062	國的自由國主	福台、安静的新闻	443.4651国际公	的中国的国际	海湖市

"Lomonosov" system average resource utilization rate (left to right):

- CPU user
- Flops
- L1 replacements
- LLC misses
- Memory read&write
- IB send&receive
- LoadAverage

Highlighting statuses (left column) and average resource utilization (right colored block)

Categories of similar jobs by combination of criteria and tagging:

- Formal criteria (mostly thresholds)
- Machine learning



auto_low_gpu_load + partition_GPU

🗅 Lor	nonosov tasks tab	× 🗋 Lomonosov ta	asks tab 🗙 🕻	ງ Lomono	sov tasks t	ab ×	🕒 Lom	nonc	osov tas	sks tab 🗙 🗋	Lomonosov task	s tab 🗙 🚬					l		×
$\leftrightarrow \rightarrow$	୯ ୍																	0	Ξ
	onosov ta																		•
WHERE pa	rtition NOT IN ('tes	st', 'gputest') AND t_e	nd - t_start >	600 ORDER	BY t_end DE:	SC LIMI	T 10000												
auto	_low_gpu_load ×	partition_GPU ×	AUof																
Atleas	t on																		
None o	f																		
query Short tab	leLong table																		2
id 🗢				cores_houts	num_cores						avg_cpu_perf_l1d_re ¢ l							avg_gpu_load	۰
	Mart Aught JSS	2016-03- 2016-03- 12 15 13:47:45 13:47:59	COMPLETING	576.03	8	4320	gpu	45.	677	52823000.0	5276600.0	1176800.0	464239000.0	183886000.0	32454100.0	32414200.0	7.96475	0.0	
LIDALLO	NIN, MARINE	2016-03- 14 02:29:37 2016-03- 15 11:49:59	COMPLETING	266.72	8	2000	gpu	44.	148	380629000.0	19568800.0	3780880.0	293508000.0	72241400.0	95725400.0	95811900.0	7.88789	0.0	
	antennet al la	2016-03- 14 02:30:07 2016-03- 15 10:37:53	COMPLETED	2313.32	72	1927	gpu	42.	242	379755000.0	17409200.0	3784420.0	271429000.0	71041500.0	153444000.0	153431000.0	7.44805	0.0	
128233	researches, 1914	2016-03- 12 13:20:04 08:45:08	COMPLETED	1078.68	16	4)45	gpu	44.	52	99837600.0	10123900.0	2729550.0	402283000.0	132364000.0	0.0	0.0	7.98006	0.0	
SEASER S	Adam.	2016-03- 09 12 14:42:29 13:58:30	TIMEOUT	570.14	8	4276	gpu	44.	851	66777600.0	6628040.0	2826120.0	401902000.0	161723000.0	0.0	0.0	7.82719	0.0	
Santon fin		2016-03- 09 12 14:42:29 13:58:30	TIMEOUT	570.14	8	276	gpu	45.	1539	66973300.0	6674920.0	2890210.0	375740000.0	148305000.0	0.0	0.0	7.84421	0.0	
102020		2016-03- 09 12 14:42:29 13:58:30	TIMEOUT	570.14	8	4276	gpu	44.	635	66907300.0	6554740.0	2838920.0	385120000.0	153089000.0	0.0	0.0	7.8385	0.0	
20098033	enen az altan - Neti nervele ekilekin gr	2016-03- 09 12 09:37:54 09:38:15	COMPLETING	576.05	8	4320	gpu	45.	987	222342000.0	5308890.0	1479480.0	221179000.0	101954000.0	0.0	0.0	8.00101	0.0	
Sand He		2016-03- 08 11 21:21:26 21:01:49	TIMEOUT	9174.15	128	4300	gpu	45.	477	113253000.0	17185400.0	324663.0	484449000.0	124135000.0	64174500.0	64145800.0	7.99672	0.0	
10200	-94. P. A. 2015	2016-03- 2016-03-	COMPLETED	732.18	16	2745	apu	45	267	105411000 0	10468700 0	2621000.0	417612000 0	136514000.0	72875000.0	72864400.0	8 00139	0.0	₽

26.09.2017



Single core jobs not in test partitions

ery ort tabi	le Long table		S	ingle	e n	00	e		CP	U_us	ser					Loa	
¢	account	♦ t_start	♦ t_end	\$ state \$	cores_ho	🚓 num_co	s duratic	partition	avg_cpu_us¢	avg_cpu_flops \$	avg_cpu_perf_l1d_re#l	avg_llc_miss \$	avg_mem_load \$	avg_mem_store \$	avg_ib_rcv_dat&	avg_ib_xmit_da	avg_loadav
38308	ideaberina_fIIM	2016-02-04 06:25:13	2016-02-04 07:51:49	COMPLETED	11.55	8	86	regular4	6.52	8979330.0	183216.0	118759.0	37346000.0	22979700.0	1732.93	1967.65	1.00467
<u>36734</u>	makhans	2016-02-03 05:54:57	2016-02-03 07:58:22	FAILED	16.46	8	123	regular4	6.07174	28813500.0	1132060.0	101059.0	30475200.0	13338400.0	1719400.0	4526380.0	0.994348
<u>37308</u>	passo	2016-02-03 00:59:36	2016-02-03 04:24:47	COMPLETED	27.36	8	205	hdd4	0.0923077	1498.76	30833.7	1890.64	594213.0	245048.0	80744.6	79506.6	0.994103
<u>37190</u>	Markets, III	2016-02-02 22:01:55	2016-02-02 22:22:17	COMPLETED	2.72	8	20	regular4	10.27	8689710.0	166478.0	124506.0	37970300.0	23214700.0	2142.9	2049.93	1.0
37093	Herbering, 1114	2016-02-02 18:53:35	2016-02-02 19:54:02	TIMEOUT	8.06	8	60	regular4	7.03	8864480.0	153599.0	96245.3	37148900.0	22867400.0	0.0	0.0	1.001
37055	Mandowska, 1934	2016-02-02 16:58:39	2016-02-02 17:33:38	FAILED	4.66	8	34	regular4	6.12	9650980.0	56315.4	20220.9	36315500.0	22835900.0	0.0	0.0	0.99
35368	saaiharos	2016-01-28 14:51:53	2016-01-31 14:52:04	TIMEOUT	576.02	8	4320	regular4	6.19784	7540280.0	287333.0	32770.1	8116410.0	4075520.0	50941.7	2798470.0	0.994907
<u>36081</u>	simbleme	2016-01-31 02:07:46	2016-01-31 05:33:30	COMPLETED	27.43	8	205	regular4	0.0489744	25378.4	177844.0	46266.7	<u>195</u> 80800.0	12530700.0	54673600.0	844684.0	0.900513
<u>30559</u>	summers, 211	2016-01-24 08:28:31	2016-01-27 08:28:33	COMPLETING	576.00	8	4320		Up [•]	to 3 d	lays lo	ng!	20100.0	5824880.0	2890.48	2911.17	0.98058
30224	makhan	2016-01-23	2016-01-25 05:30:34	COMPLETED	354.14	8	2000	Togunar	0.20101	20001100.0	1100000.0	10100.0	20063900.0	12936200.0	5563.94	5926180.0	1.00758
30057	andahas	2016-01-23 03:22:49	2016-01-24 18:48:44	COMPLETING	315.46	8	2365	regular4	6.08654	311.465	6483.29	409.6	114895000.0	25839200.0	1704.98	1907.76	1.01085
30597	46,25283	2016-01-24 13:02:31	2016-01-24 13:28:52	FAILED	3.51	8	26	hdd4	0.253333	116354.0	2007430.0	539.793	51266300.0	32041400.0	3184.72	2307.22	0.986667
230097	hoptanos	2016-01-23	2016-01-23 11:09:05	COMPLETING	36.58	8	274	regular4	0.0424528	310.48	8636.37	420.936	39045500.0	6020430.0	1760.95	1911.44	1.00528
30221	makhang	2016-01-23	2016-01-23 09:46:39	FAILED	4.28	8	32	regular4	5.418	22515000.0	1447870.0	83310.4	33935100.0	16352300.0	7107.31	149630.0	1.002
2 <u>30145</u>	sakou	2016-01-22 21:52:12	2016-01-23 02:52:32	TIMEOUT	40.04	8	300	regular4	0.825517	113100.0	1877160.0	7024.57	49070900.0	30596800.0	46279200.0	133681.0	1.01707
30198	pote	2016-01-23 00:16:42	2016-01-23 00:53:46	COMPLETED	4.94	8	37	hdd4	0.904	51709.4	2285050.0	3309.51	42122700.0	24885900.0	25130.5	2207980.0	1.048
30194	passo	2016-01-23 00:07:12	2016-01-23 00:49:11	COMPLETED	5.60	8	41	hdd4	0.923333	49520.7	2304560.0	4253.29	42744700.0	25238900.0	23307.5	2071330.0	1.03
<u>29822</u>	alice	2016-01-22 13:03:29	2016-01-22	TIMEOUT	13.37	8	100	regular4	0.639444	118970.0	1941090.0	5203.37	50921000.0	31777900.0	54200400.0	107747.0	1.00778
<u>29238</u>	ala, 252651	2016-01-21	2016-01-21	FAILED	21.54	8	161	regular4	0.462903	377364.0	1954350.0	1676.62	49992900.0	31221100.0	3537700.0	32939.6	1.01516

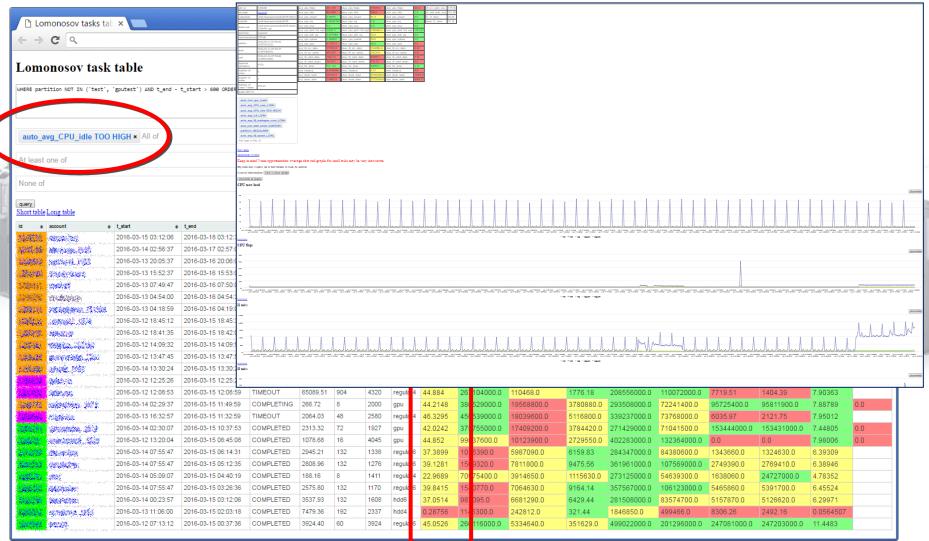


auto_avg_CPU_idle TOO HIGH (CPU idle > 25%)

→ C Q																		0
																		•
monosov tasl	table																	
partition NOT IN ('test'	, 'gputest') AND t_end -	t_start > 600 ORDER BY	t_end DESC LIMIT	100														
to_avg_CPU_idle TOC	O HIGH × All of																	
east one of																		
le of																		
table Long table																		
	 t_start 2016-03-15 03:12:06 	t_end d 2016-03-18 03:12:33	state completing	cores_hourse 864.09	num_cores 12	duration 4320	partitio hdd6		g_cpu_users av	169.0	avg_cpu_perf_11d_repl 888596.0	avg_lic_miss • 2666.27	avg_mem_load • 49532300.0	avg_mem_store e		avg_lb_xmlt_data 2034470.0		avg_gpu_i
	2016-03-15 03:12:06	2016-03-17 02:57:00	COMPLETING	576.05	8	4320	regula		02955 34 3.4453 41	10200.0	9313360.0	2893730.0	49532300.0	103160000.0	2774010.0 42518500.0	42116400.0	0.996068	
AN MARKANA SAL	2016-03-13 20:05:37	2016-03-16 20:06:00	COMPLETING	576.05	8	4320	regula		1.455 13	857000.0	3378370.0	340711.0	681893000.0	173124000.0	219350000.0	220950000.0	7.28103	
and mathematic (1935) And Annalysiana	2016-03-13 15:52:37	2016-03-16 15:53:00	COMPLETING	864.08	12	4320	regula		14252 46	07.3	22455.5	1899.62	115720000.0	40758600.0	0.0	0.0	1.01347	
and realized	2016-03-13 07:49:47	2016-03-16 07:50:00	COMPLETING	576.03	8	4320	regula		107533 48	389	12641.5	521.114	282382.0	124403.0	117628.0	1868.38	0.0188986	
TEACHER	2016-03-13 04:54:00	2016-03-16 04:54:30	COMPLETING	576.07	8	4320	regula		56161 82		318784.0	29658.8	8596570.0	4180510.0	0.0	0.0	1.06509	
in rearing and Action	2016-03-13 04:18:59	2016-03-16 04:19:00	COMPLETING	576.00	8	4320	regula		5.9642 15	471000.0	4021940.0	362244.0	625586000.0	144334000.0	199369000.0	199058000.0	7.98615	
an annaith 1914	2016-03-12 18:45:12	2016-03-15 18:45:30	COMPLETING	864.06	12	4320	regula		5.8264 25	199000.0	4975520.0	320847.0	529639000.0	204627000.0	212641000.0	212267000.0	11.876	
NE MERCEN	2016-03-12 18:41:35	2016-03-15 18:42:00	COMPLETING	576.06	8	4320	regula		3.7856 11	749000.0	7060340.0	793086.0	95724300.0	60647700.0	0.0	0.0	1.98051	
A THE LA PACE	2016-03-12 14:09:32	2016-03-15 14:09:59	COMPLETING	576.06	8	4320	hdd4	46	6.0686 14	67000.0	1406590.0	64333.8	591282000.0	261165000.0	1403850.0	1369240.0	7.96508	
C BANTHE THE	2016-03-12 13:47:45	2016-03-15 13:47:59	COMPLETING	576.03	8	4320	gpu	45	5.7677 52	23000.0	5276600.0	1176800.0	464239000.0	183886000.0	32454100.0	32414200.0	7.96475	0.0
ate share that	2016-03-14 13:30:24	2016-03-15 13:30:29	COMPLETING	192.01	8	1440	regula	4 45	5.04 89	19400.0	27519400.0	1007380.0	544654000.0	144182000.0	0.0	0.0	7.86042	
and the second second se	2016-03-12 12:25:26	2016-03-15 12:25:29	TIMEOUT	65088.75	904	4320	regula	4 45	5.6274 26	841000.0	97567.5	1800.79	211683000.0	111954000.0	13622.0	1843.85	7.94228	
in the second	2016-03-12 12:08:53	2016-03-15 12:08:59	TIMEOUT	65089.51	904	4320	regula	4 44	4.884 26	104000.0	110468.0	1776.18	208556000.0	110072000.0	7719.51	1404.39	7.90363	
lite universities suite	2016-03-14 02:29:37	2016-03-15 11:49:59	COMPLETING	266.72	8	2000	gpu	44	4.2148 38	629000.0	19568800.0	3780880.0	293508000.0	72241400.0	95725400.0	95811900.0	7.88789	0.0
nde riederice	2016-03-13 16:32:57	2016-03-15 11:32:59	TIMEOUT	2064.03	48	2580	regula	4 46	6.3295 45	539000.0	18039600.0	5116800.0	339237000.0	73768000.0	6035.97	2121.75	7.95012	
al second dif	2016-03-14 02:30:07	2016-03-15 10:37:53	COMPLETED	2313.32	72	1927	gpu		2.0242 37	755000.0	17409200.0	3784420.0	271429000.0	71041500.0	153444000.0	153431000.0	7.44805	0.0
de senaphen, stic	2016-03-12 13:20:04	2016-03-15 08:45:08	COMPLETED	1078.68	16	4045	gpu		4.852 99	37600.0	10123900.0	2729550.0	402283000.0	132364000.0		0.0	7.98006	0.0
the state of the s	2016-03-14 07:55:47	2016-03-15 06:14:31	COMPLETED	2945.21	132	1338	regula		7.3899 10	6390.0	5987090.0	6159.83	284347000.0	84380600.0	1343660.0	1324630.0	6.39309	
And the second s	2016-03-14 07:55:47	2016-03-15 05:12:35	COMPLETED	2808.96	132	1276	regula		9.1281 18	9320.0	7811800.0	9475.56	361961000.0	107569000.0	2749390.0	2769410.0	6.38946	
and a strength and the state of the strength of the strength of the state of the strength of the strengt of the strength of the strength of the strength of th	2016-03-14 05:09:07	2016-03-15 04:40:19	COMPLETED	188.16	8	1411	regula		2.9689 70	75400.0	3914650.0	1115630.0	273125000.0	54639300.0	1638060.0	24727000.0	4.78352	
	2016-03-14 07:55:47	2016-03-15 03:26:36	COMPLETED	2575.80	132	1170	regula	-	9.8415 18	0770.0	7064630.0	9164.14	357567000.0	106123000.0	5465860.0	5391700.0	6.45524	
												6429.44	281508000.0	83574700.0	5157870.0	5126620.0	6.29971	
in anaratan Maratatan Maratatan	2016-03-14 00:23:57	2016-03-15 03:12:06 2016-03-15 02:03:18	COMPLETED	3537.93 7479.36	132	1608 2337	hdd6 hdd4		7.0514 98 28756 11	095.0 5300.0	6681290.0 242812.0	321.44	1846850.0	499466.0	8306.26	2492.16	0.0564507	



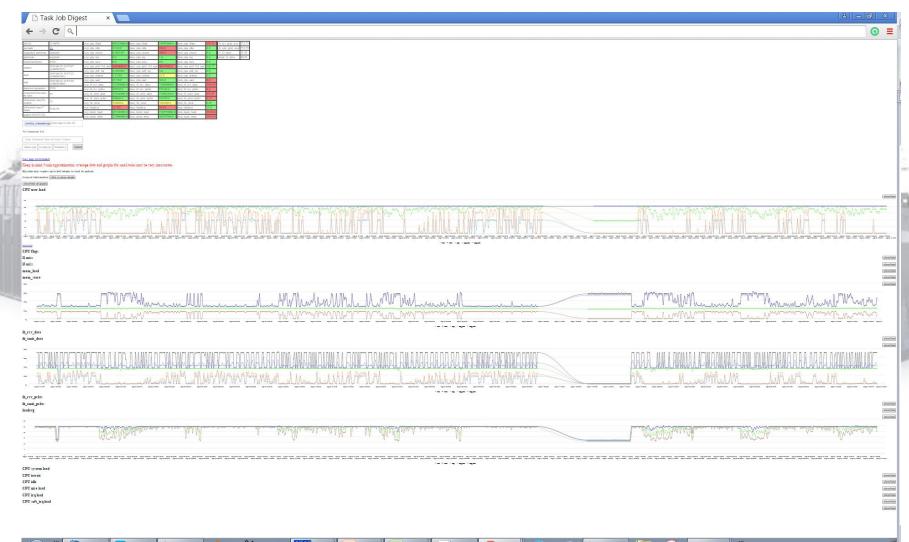
auto_avg_CPU_idle TOO HIGH (CPU idle > 25%)





JobDigest

Analysis of job behavior and resource utilization dynamics

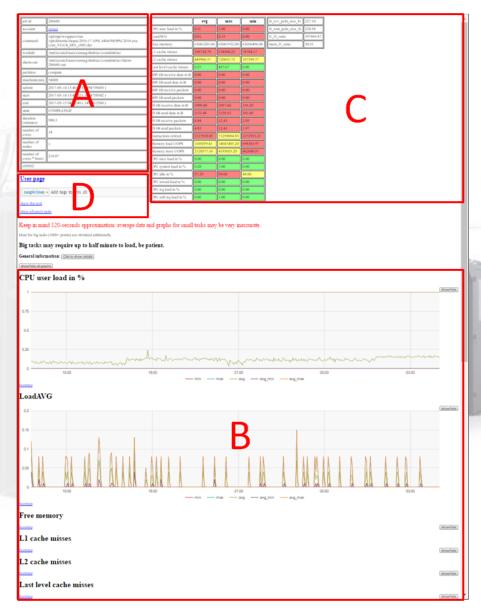




The JobDigest Report

JobDigest report blocks: A – General job information, B – Dynamic job characteristics, C – Integral job characteristics,

D – Tags and job categories.





JobDigest

Dynamics of resource utilization execution

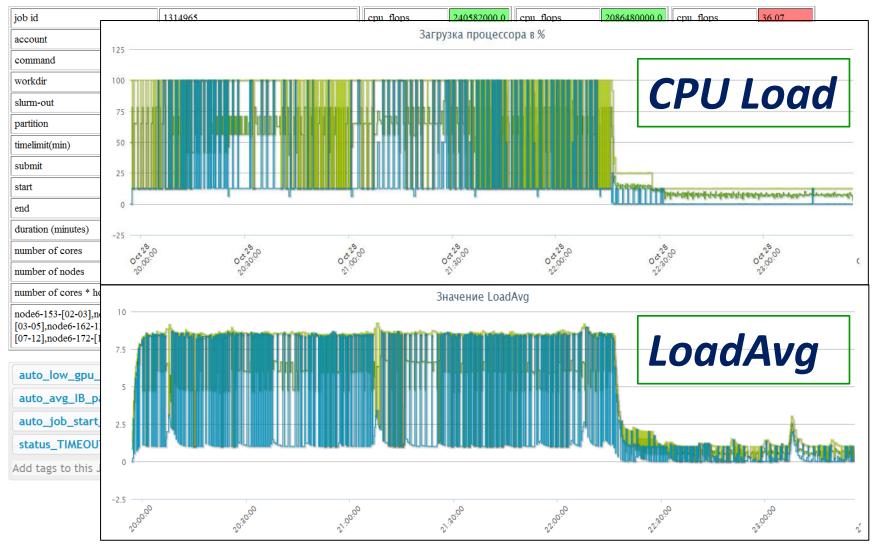
		1					
job id	1314965	cpu_flops	240582000.0	cpu_flops	2086480000.0	cpu_flops	36.07
account		cpu_idle	47.5283	cpu_idle	100.0	cpu_idle	0.0
command	ffxp	cpu_iowait	0.0178268	cpu_iowait	86.0	cpu_iowait	0.0
workdir	unknown	cpu_irq	0.000436416	cpu_irq	1.0	cpu_irq	0.0
slurm-out	unknown/slurm-1314965.out	cpu_nice	0.0	cpu_nice	0.0	cpu_nice	0.0
partition	gpu	cpu_perf_l1d_repl	2661700.0	cpu_perf_l1d_repl	28549900.0	cpu_perf_l1d_repl	35.91
timelimit(min)	4320	cpu_soft_irq	0.32768	cpu_soft_irq	6.0	cpu_soft_irq	0.0
submit	2016-07-09 01:18:40 (1468016320)	cpu_system	1.21602	cpu_system	100.0	cpu_system	0.0
start	2016-07-12 22:45:07 (1468352707)	cpu user	28.4004	cpu user	100.0	cpu user	0.0
end	2016-07-15 22:46:10 (1468611970)	gpu_load	0.0	gpu_load	0.0	gpu_load	0.0
duration (minutes)	4321	gpu_mem_load	0.0	gpu_mem_load	0.0	gpu_mem_load	0.0
number of cores	168	gpu_mem_usage	10658800.0	gpu_mem_usage	10858500.0	gpu_mem_usage	10334200.0
number of nodes	21	ib_rcv_data	4700.33	ib_rcv_data	16399600.0	ib_rcv_data	0.0
number of cores * hours	12098.94	ib_rcv_pckts	15.3161	ib_rcv_pckts	608.17	ib_rcv_pckts	0.0
node6-153-[02-03],node6-154-[0	2*	ib_xmit_data	1879.58	ib_xmit_data	129574.0	ib_xmit_data	0.0
[03-05],node6-162-11,node6-164 [07-12],node6-172-[12-13]	4-16,node6-165-[01-03,15],node6-170-	ib_xmit_pckts	8.23421	ib_xmit_pckts	600.87	ib_xmit_pckts	0.0
		llc_miss	102829.0	llc_miss	1668000.0	llc_miss	4.69
auto_low_gpu_load ×	$auto_avg_CPU_idle TOO HIGH \times$	loadavg	4.8232	loadavg	14.88	loadavg	0.0
auto_avg_IB_packages_	num_LOW ×	mem_load	377220000.0	mem_load	1639010000.0	mem_load	8287.4
auto_job_start_script_C	uto_job_start_script_CUSTOM ×partition_GPU ×		149185000.0	mem_store	634774000.0	mem_store	2581.3
status_TIMEOUT × auto	o_avg_IB_speed_LOW ×						
Add tags to this JD		In In	tegra	ıl job cł	naraci	teristics	5

Resource manager data



JobDigest

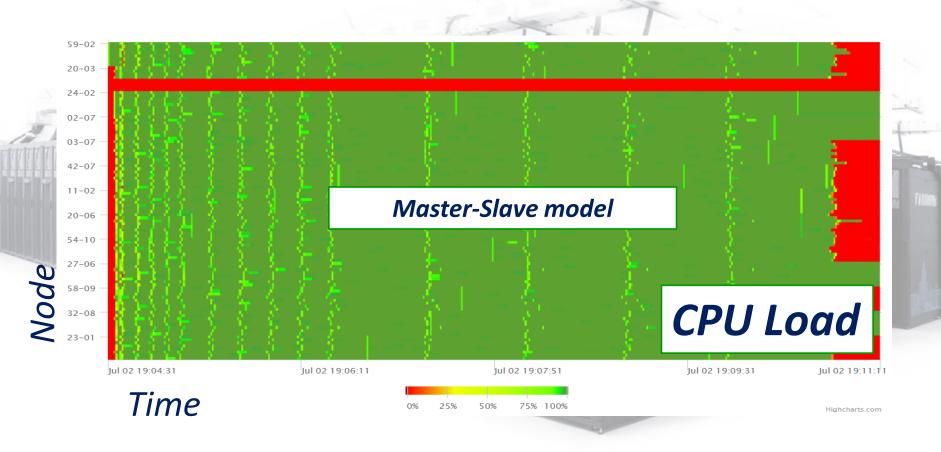
Dynamics of resource utilization execution







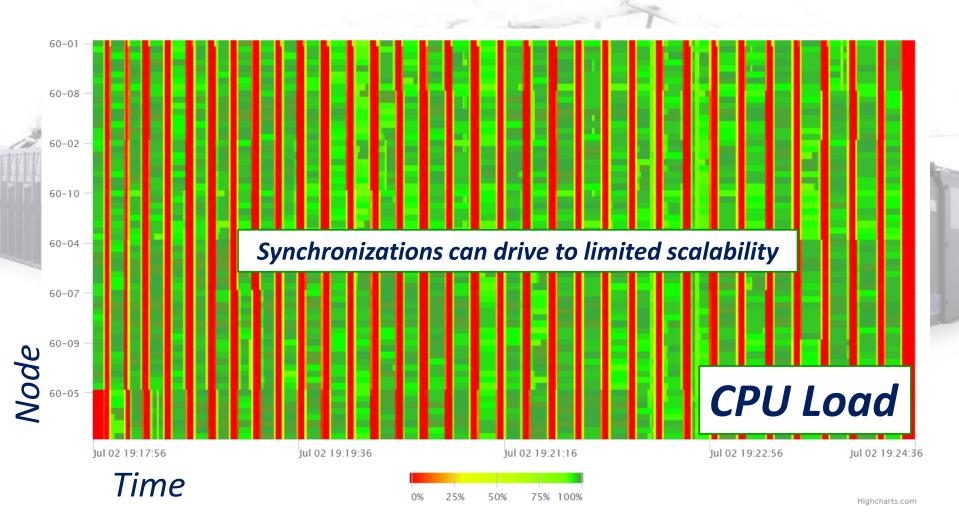
Dynamics of resource utilization execution





JobDigest

Dynamics of resource utilization execution

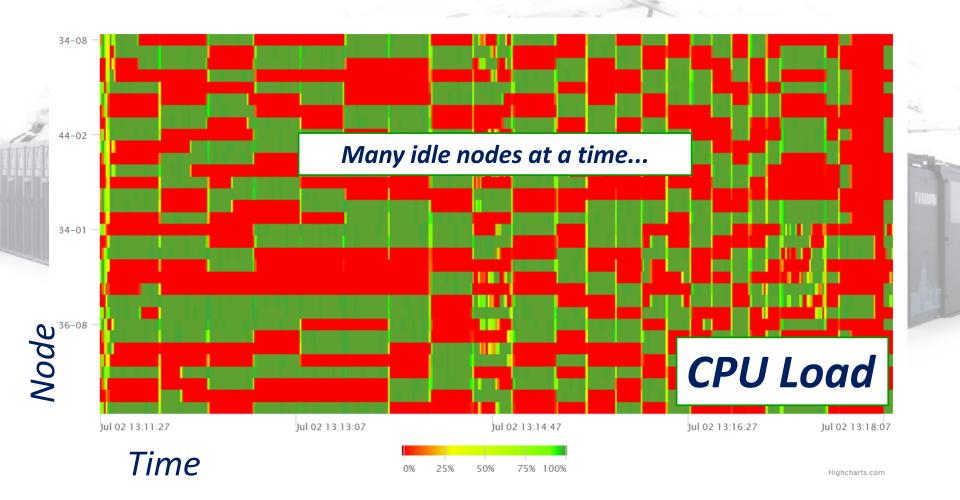


26.09.2017



JobDigest

Dynamics of resource utilization execution



MSU/Uppsala application analysis

Joint project on apps analysis under STINT support

• Three different applications from Uppsala:

- "Chunks and Tasks" programming model. Evaluation of previously developed code at Uppsala.
- 2) Parallel block preconditioned inner-outer solvers with application to Glacial Isostatic Adjustment (GIA). Extensive usage of third party libraries.
- 3) The Fast Multipole Method (FMM). Self-written code at Uppsala.
- Methodology
 - Providing MSU's Lomonosov-2 system as a testbed
 - Scalability and overall resource utilization analysis with the tools developed by MSU based on system monitoring
 - Further diagnostics with other tools (Scalasca, Valgrind, mpiP)

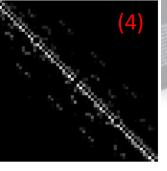
MSU/Uppsala application analysis

Parallel block preconditioned inner-outer solvers with application to Glacial Isostatic Adjustment (GIA)

- 1) JobDigest demonstrated low memory locality on the last phase of program execution the solver
- 2) The solver demonstrates worse scalability among all application phases
- 3) Profiling of a single thread execution with Valgrind shows, that most time is spent in library call to epetra_dcrsmv_, which is internal function for sparse matrix - vector multiplication routine from Trilinos
- 4) The communication pattern shows that most interactions are seen between neighbor processes



				And a local distance of the local distance o		
Processes	Assembling	Grid generator	Setup DOFs	Setup preconditioner	Solve	Total
1	479.00	64.80	253.00	27.40	962.00	1,790.00
2	245.00	36.50	285.00	16.20	548.00	1,130.00
4	124.00	19.10	224.00	8.53	302.00	678.00
8	61.80	9.93	103.00	4.74	208.00	387.00
16	31.10	5.39	51.60	(Z) 2.87	105.00	197.00
32	16.20	3.93	27.20	2.44	77.40	128.00
64	9.34	2.57	16.10	2.81	57.70	89.70
128	4.83	2.03	12.70	3.12	62.90	86.50
256	3.27	1.95	11.30	7.27	96.90	123.00



Transmitted data



Taking care of resource utilization and productivity, take care of your users!



Russian Supercomputing Days 2017 International Supercomputing Conference Moscow, Russia

