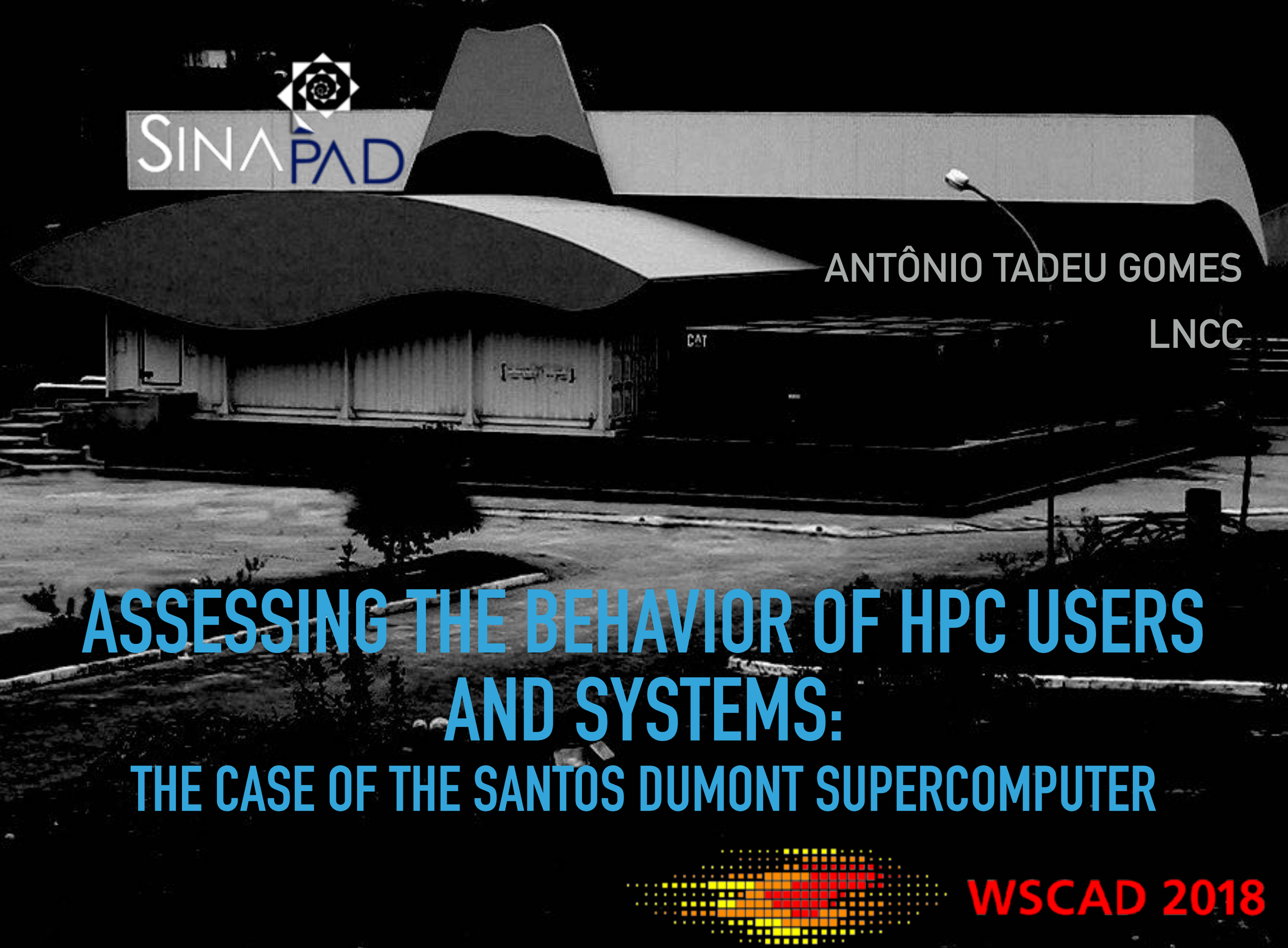


ANTÔNIO TADEU GOMES

LNCC

ASSESSING THE BEHAVIOR OF HPC USERS AND SYSTEMS:

THE CASE OF THE SANTOS DUMONT SUPERCOMPUTER

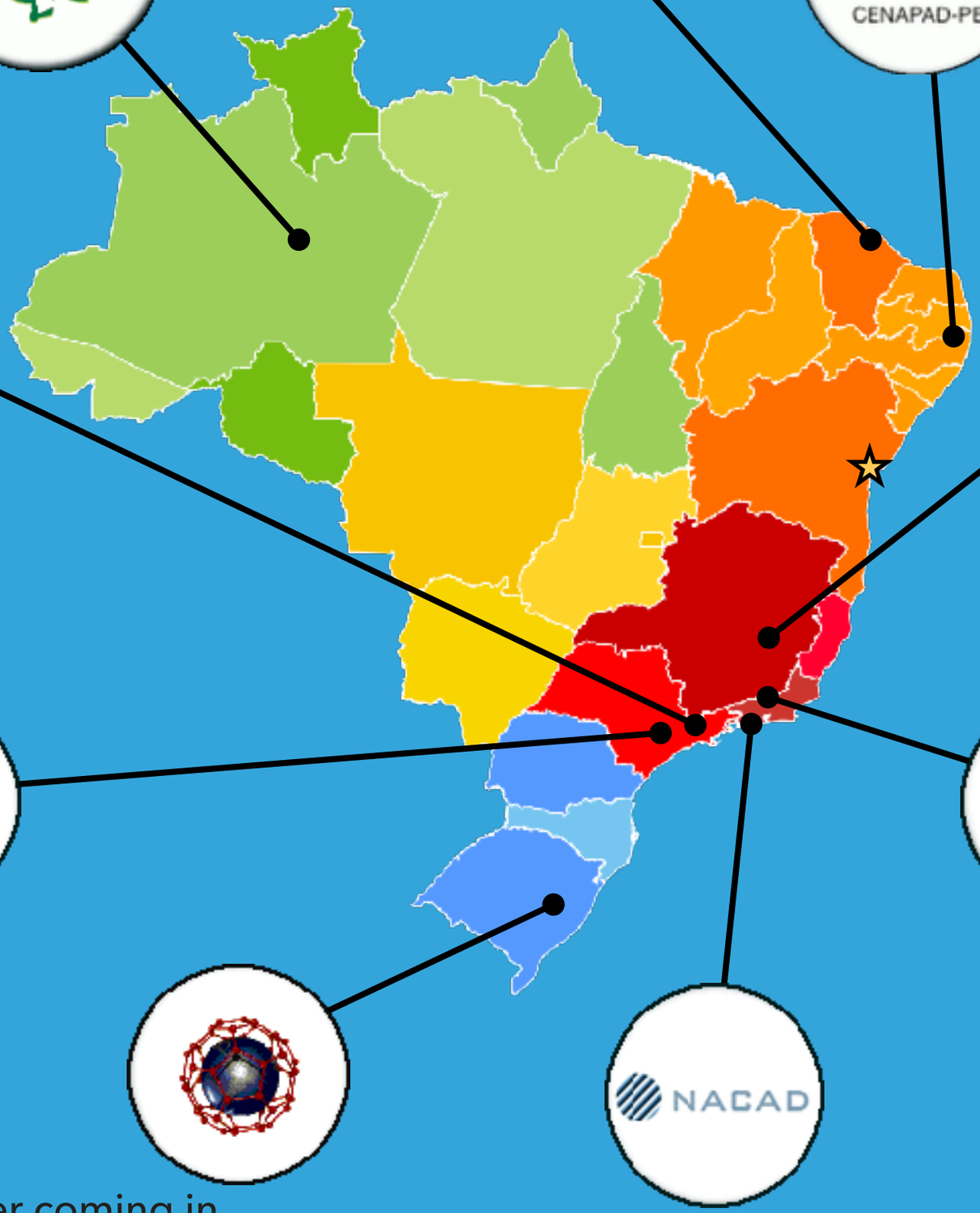
WSCAD 2018

9 CENTERS

- SERVICE PROVISIONING

- DEVELOPMENT (E.G. SCIENCE GATEWAYS)

- TRAINING



SINAPAD

★ New center coming in...

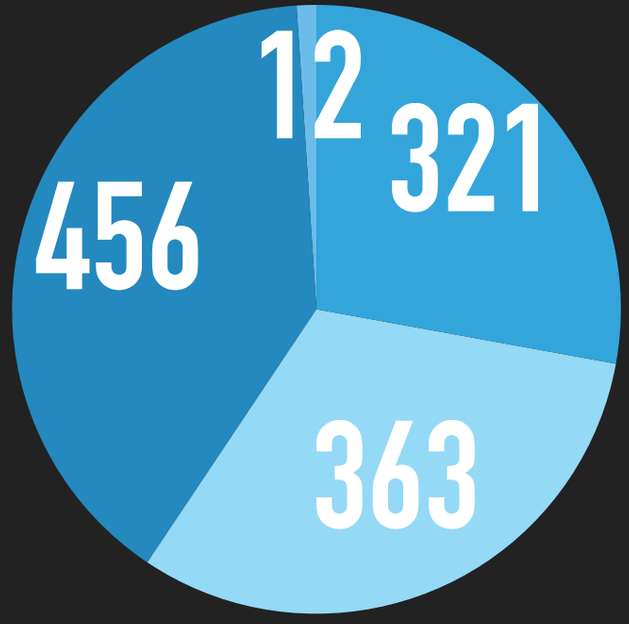
LNCC

LNCC

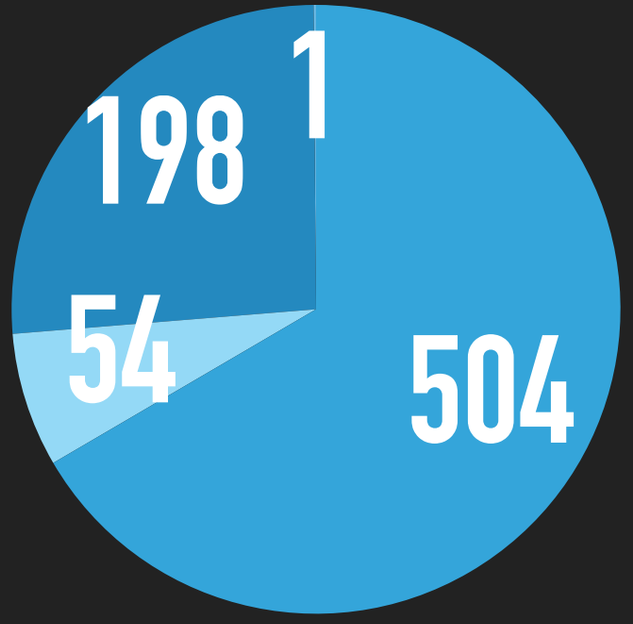
The Santos Dumont petaflopic facility

CONFIGURATION

- ~1.1 PFlops computing capability
- 756 nodes with various configurations: CPUs, GPGPUs, MICs, SHMEM
- ~1.7 PBytes Lustre storage; Infiniband interconnection
- Linux OS; Slurm resource manager



- B710 CPU
- B715 CPU+MIC
- B715 CPU+GPGPU
- Mesca2



SANTOS DUMONT: STATISTICS

3 OPEN CALLS
(PROJECTS FROM 1ST CALL ENDING THIS YEAR; FROM 3RD CALL BEGINNING THIS YEAR)

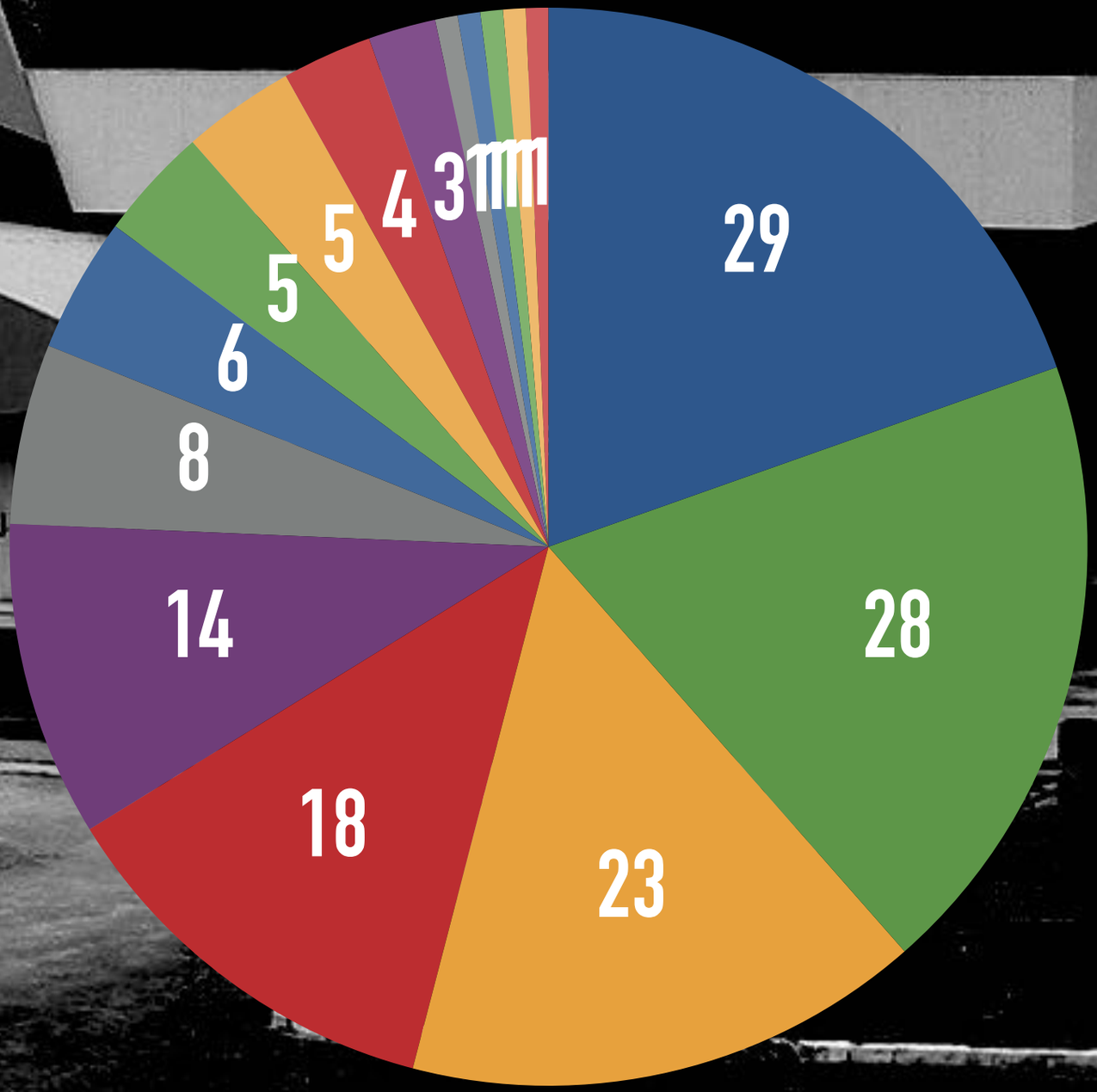
100+ PROJECTS IMPLEMENTED (PEER-REVIEWED)

~550 USERS

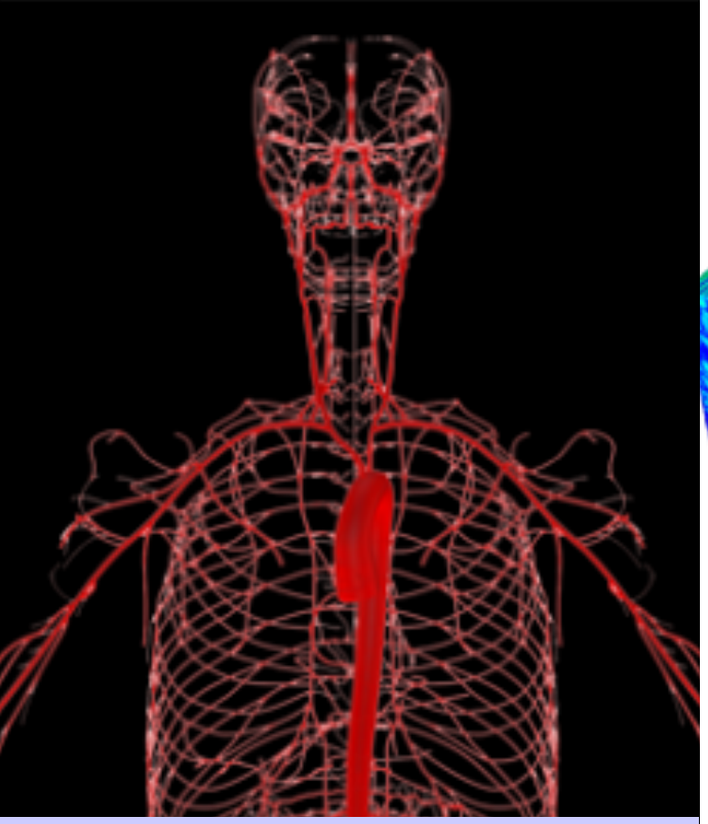
140.000+ JOBS AND 260.000.000+ SERVICE UNITS SINCE AUG/2016

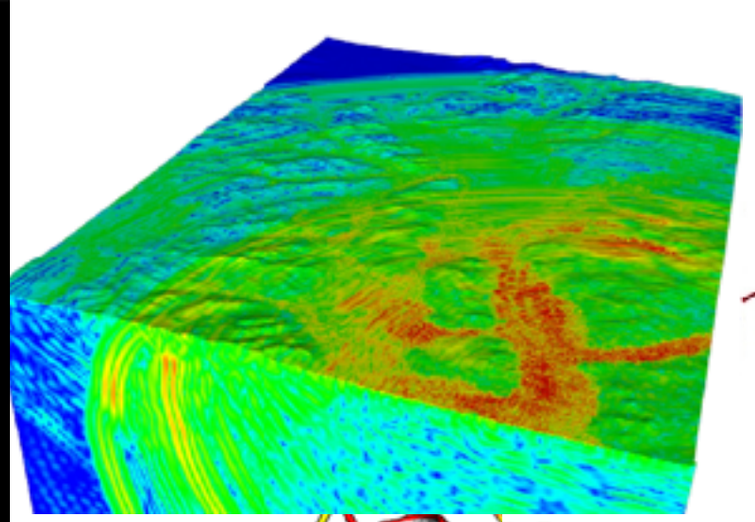
260+ TERABYTES STORED

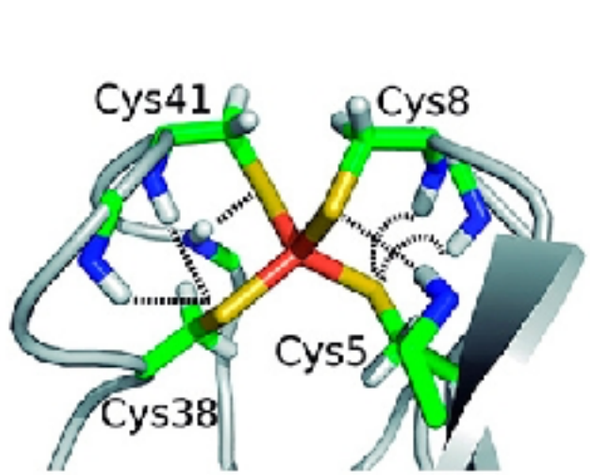
15 AREAS

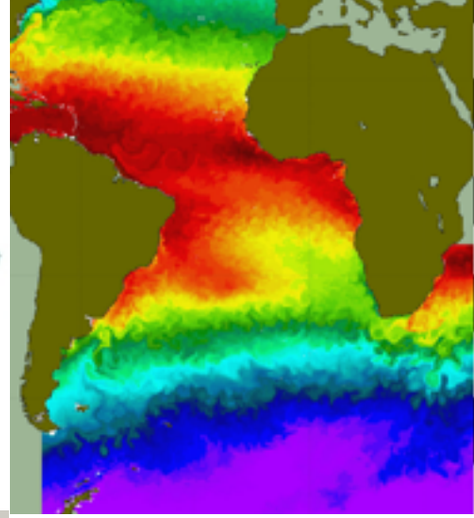


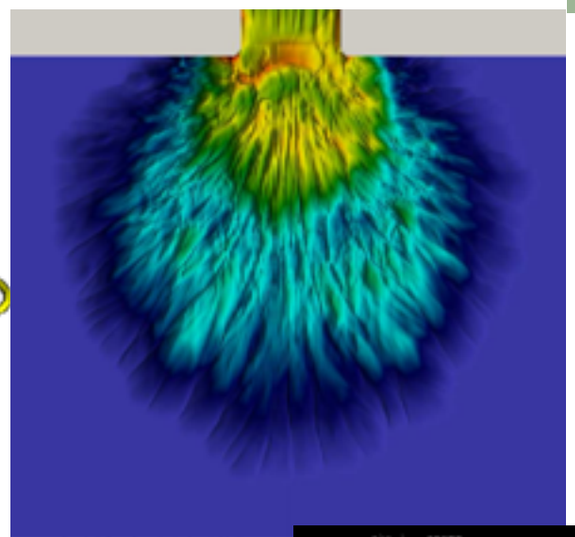
- Chemistry
- Computer Science
- Material sciences
- Biodiversity
- Physics
- Geosciences
- Maths
- Linguistics
- Engineering
- Astronomy
- Climate&Weather
- Pharmacy
- Biology
- Health
- Agriculture
- Social Sciences

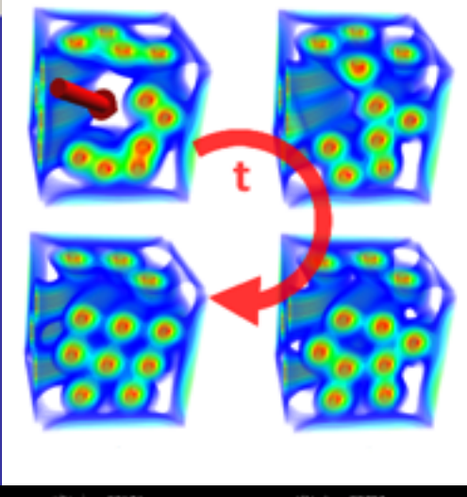


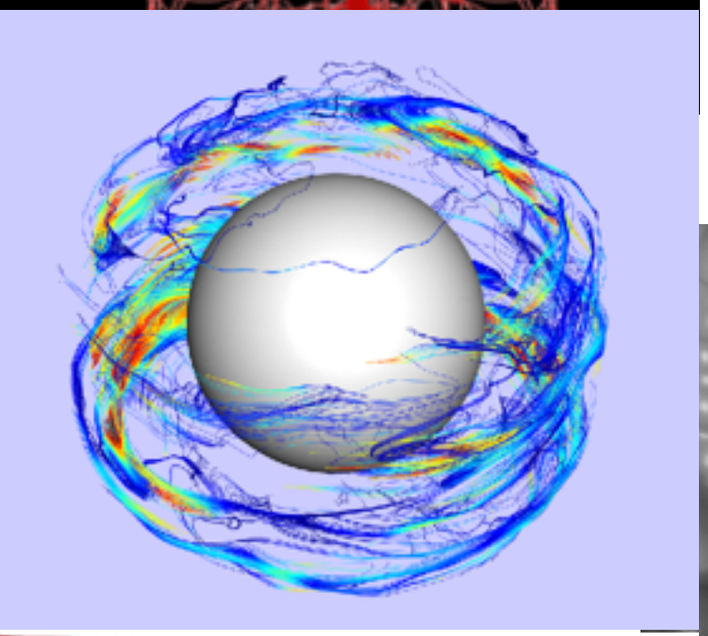


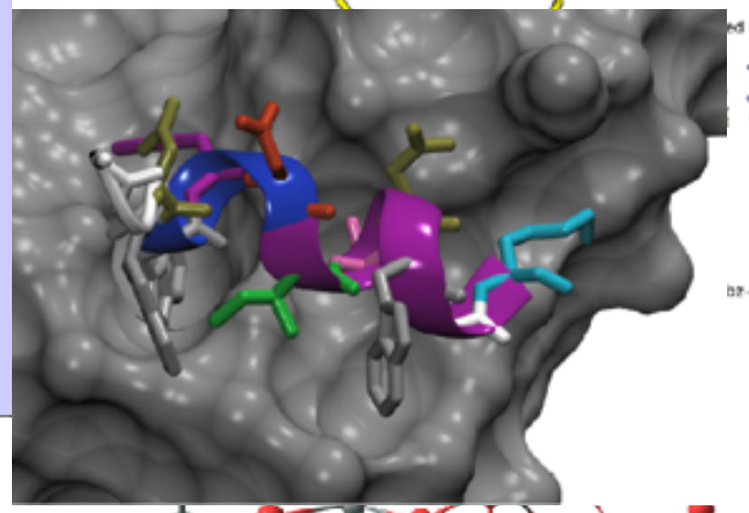

Cys41
Cys8
Cys5
Cys38

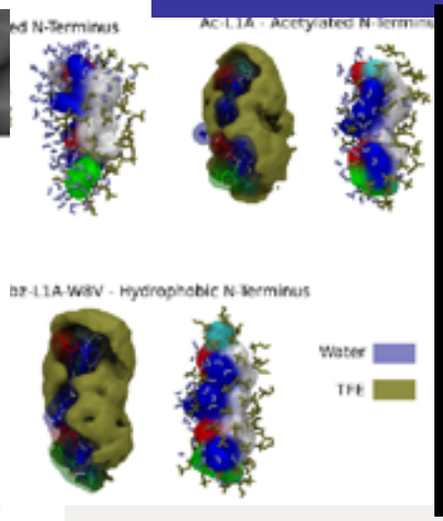


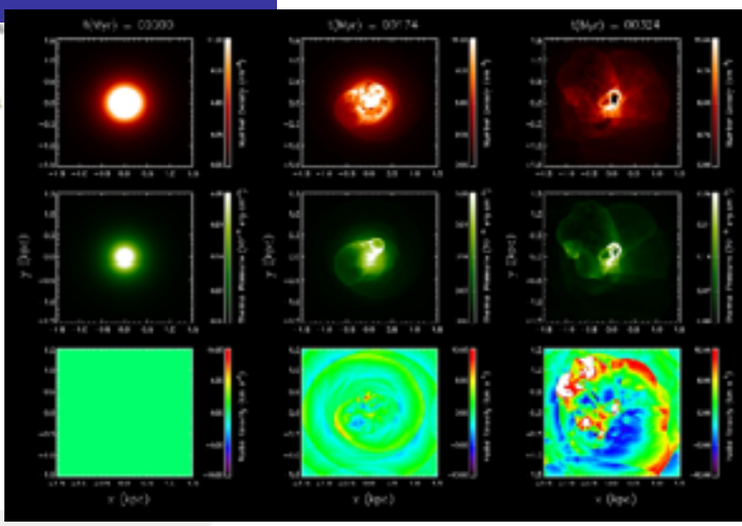


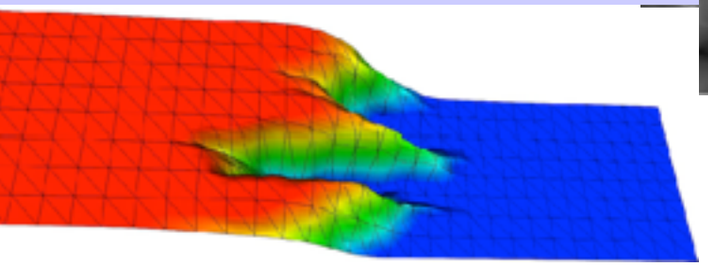

t

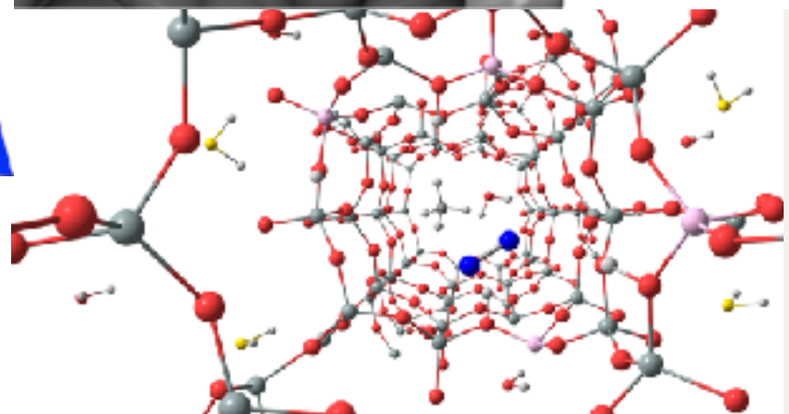


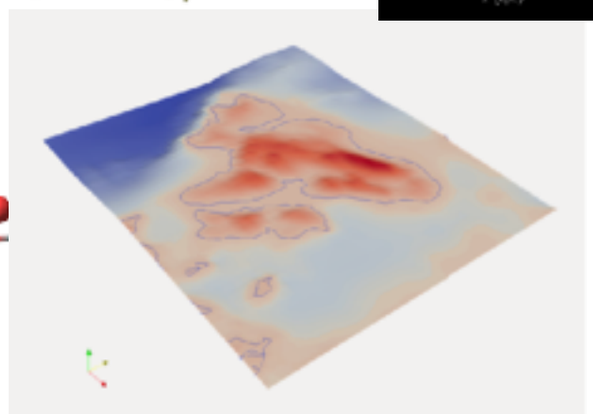


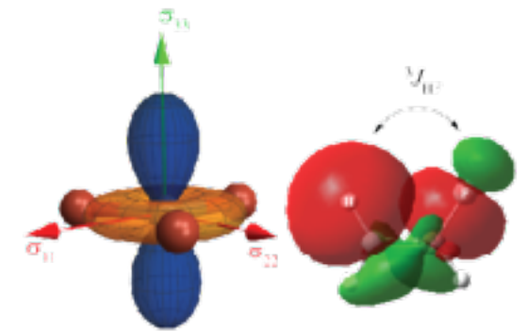

ed N-Terminus
Ac-L1A - Acetylated N-termin
b2-L1A-W8V - Hydrophobic N-terminus
Water
TFE

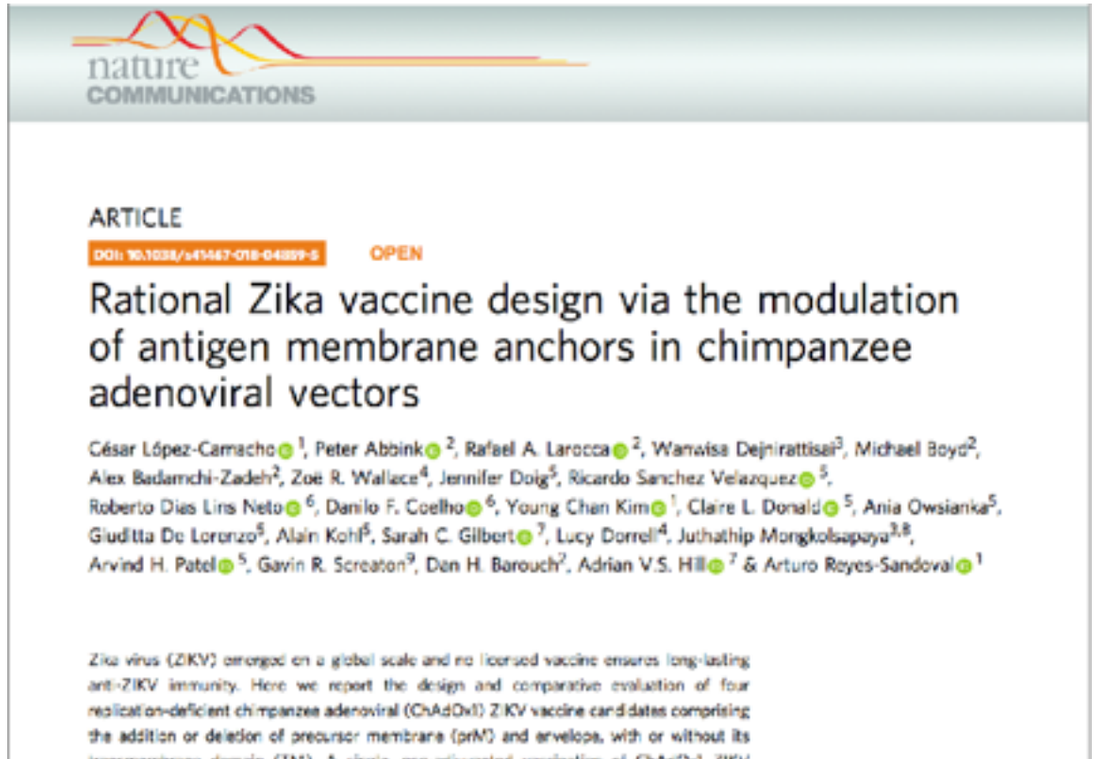












Ano	Produção bibliográfica								Projetos			Produção técnica e de inovação		
	APP	AAP	LC	TAC	DMA	DMD	TDA	TDD	OPB	PP	PDT	PAT	PCSR	OPT
2017	58	8	1	67	7	6	7	10	23	9	1	1	9	40
2016	19	0	0	12	4	2	4	1	0	12	0	0	0	15
2015	0	0	0	0	0	0	5	0	0	5	0	0	0	0
2014	0	0	0	0	0	0	8	0	0	1	0	0	0	0
2013	0	0	0	0	0	0	0	0	0	1	0	0	0	0

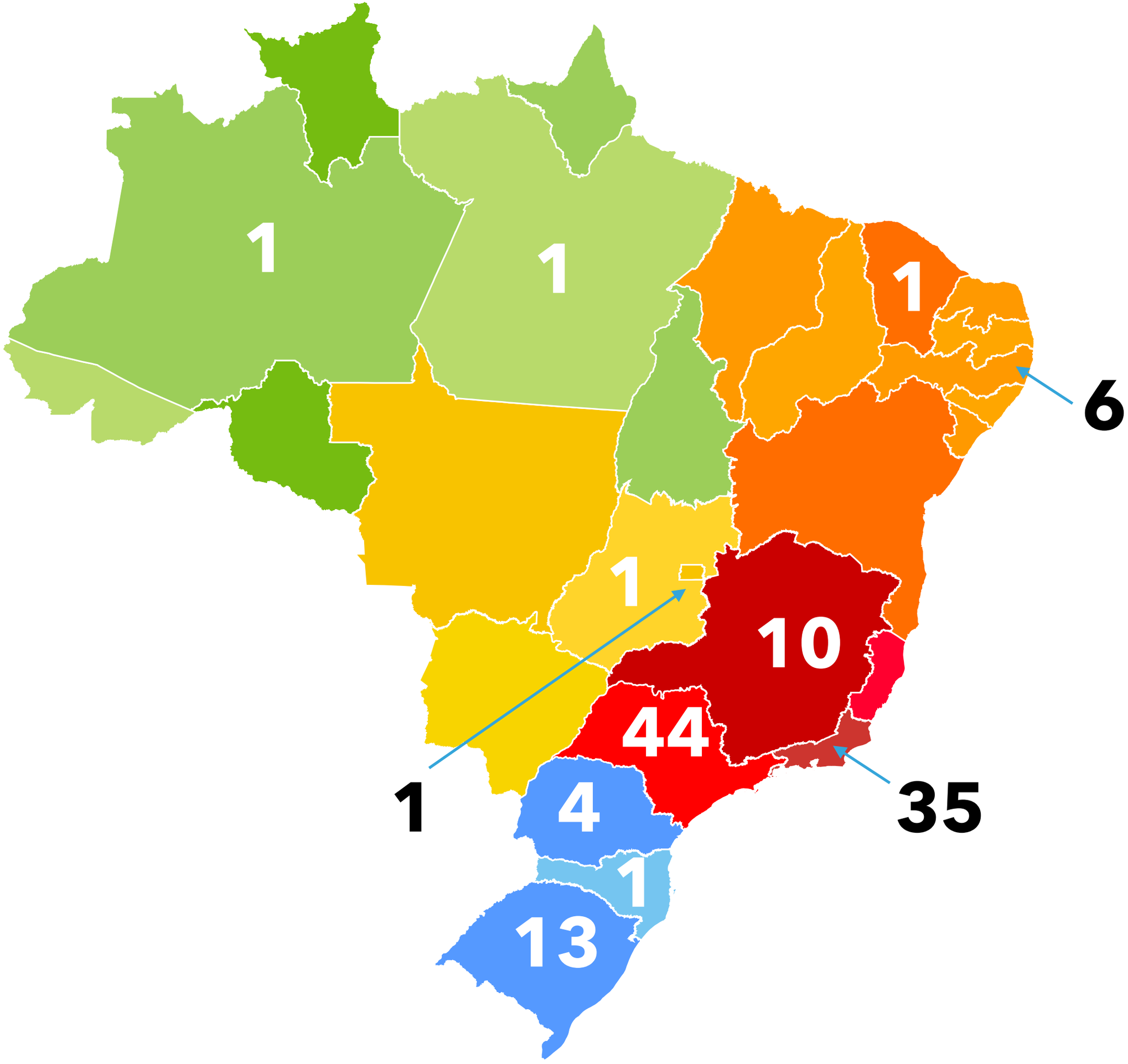
Legenda:

Produção bibliográfica
APP Artigos completos publicados em periódicos
AAP Artigos aceitos para publicação
LC Livros e capítulos
TAC Trabalhos publicados em anais de congressos
DMA Dissertações de mestrado em andamento
DMD Dissertações de mestrado defendidas
TDA Teses de doutorado em andamento
TDD Teses de doutorado defendidas
OPB Outras produções bibliográficas

Projetos
PP Projetos de pesquisa financiados
PDT Projeto de desenvolvimento tecnológico

Produção técnica e de inovação
PAT Patentes
PCSR Programas de computador sem registro
OPT Outras produções técnicas

+100 PROJECTS IN SDUMONT



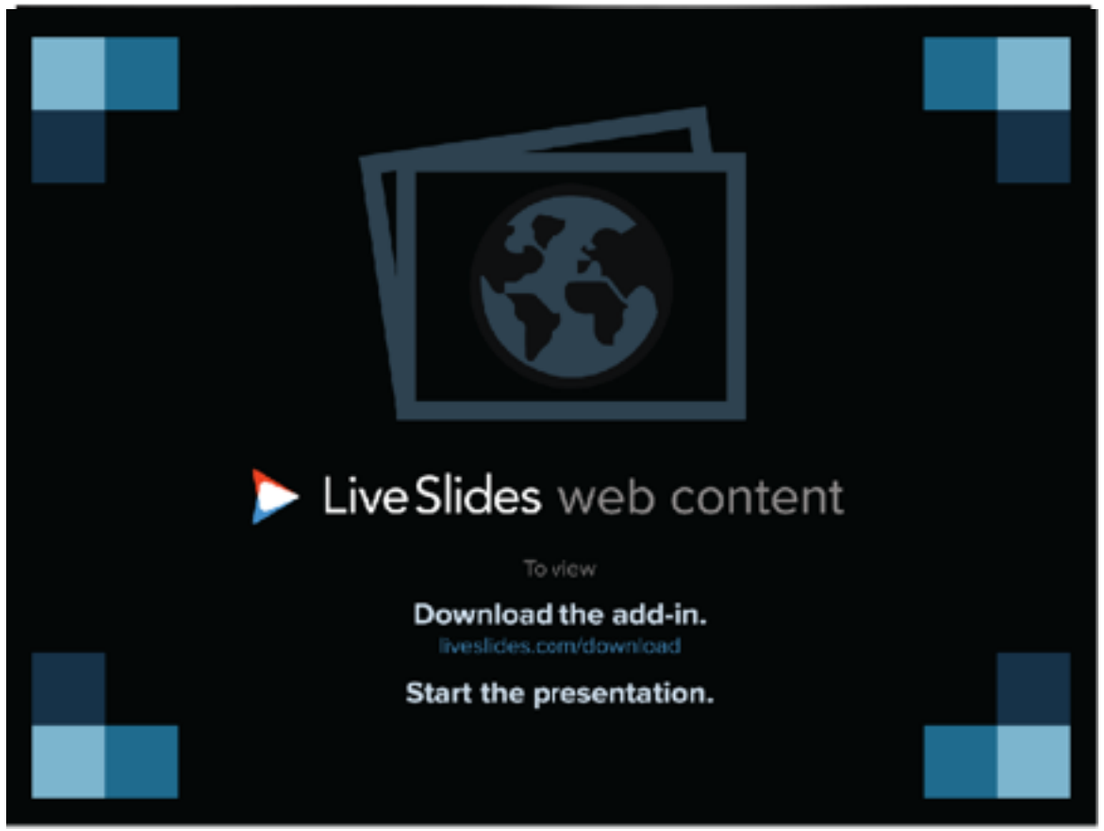
IS THIS CAPACITY USED EFFICIENTLY?

USERS/DEVELOPERS READINESS FOR SUPERCOMPUTING

- (./configure && make) and go for it!
- Not just a matter of **coding or not coding**:
  "*Yeah, my gromacs 3.0.4 compiled!*"
- **New methods** (mathematical and computational) to the rescue?
  "*Hmmm, not sure it will work…*"
  - Don’t blame them
- At LNCC/SDumont a **parallelization and optimization** group does exist
  - Problem of **scale**…

USERS READINESS FOR TIME-SHARING SYSTEMS

- "1963 Timesharing: A Solution to Computer Bottlenecks”

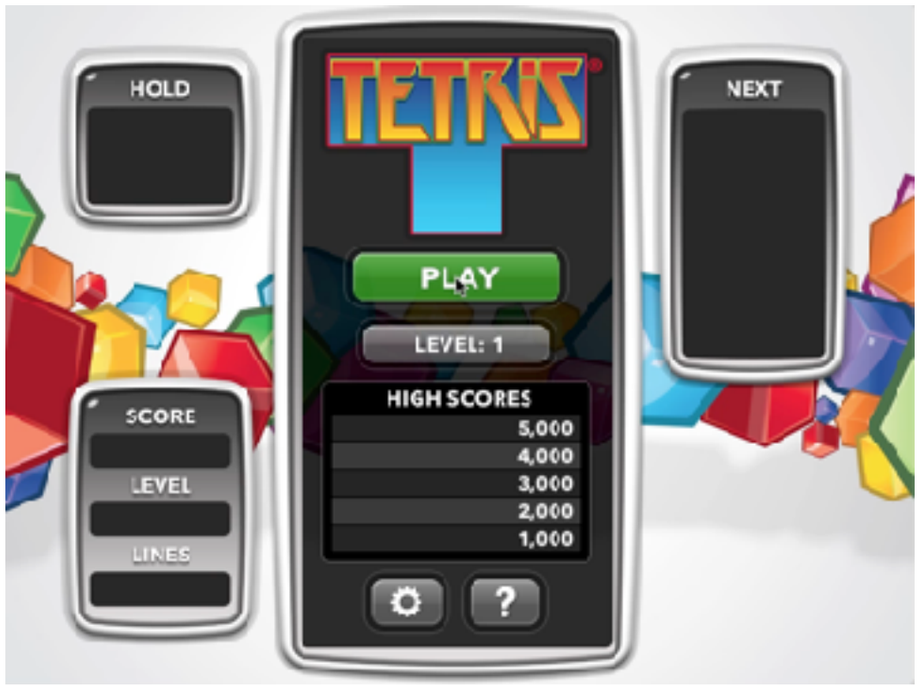


https://youtu.be/Q07PhW5sCEk

USERS READINESS FOR TIME-SHARING SYSTEMS

- "1963 Timesharing: A Solution to Computer Bottlenecks"
- Today it's more like a Tetris game

https://youtu.be/Q07PhW5sCEk



- Concept of **job geometry**

THE USERS' AND JOBS' BEHAVIOR

- Analysis using Slurm accounting facility
  - "Exclusive mode", Default time estimation = max W.C.T.

Partition	Max W.C.T (hours)	Max # cores	Max # executing jobs per user	Max # enqueued jobs per user
cpu	48	1200	4	24
nvidia	48	1200	4	24
phi	48	1200	4	24
mesca2	48	240	1	6
cpu_dev	2	480	1	1
nvidia_dev	2	480	1	1
phi_dev	2	480	1	1
cpu_scal	18	3072	1	8
nvidia_scal	18	3072	1	8
cpu_long	744	240	1	1
nvidia_long	744	240	1	1

THE JOBS' BEHAVIOR

- Overall statistics from Aug/2016 to May/2018
  - Job status

Status	Total number of jobs	% of total
COMPLETED	77147	53,55 %
FAILED	30847	21,41 %
CANCELLED	25197	17,49 %
TIMED-OUT	10809	7,50 %
NODE FAILURE	53	0,04 %

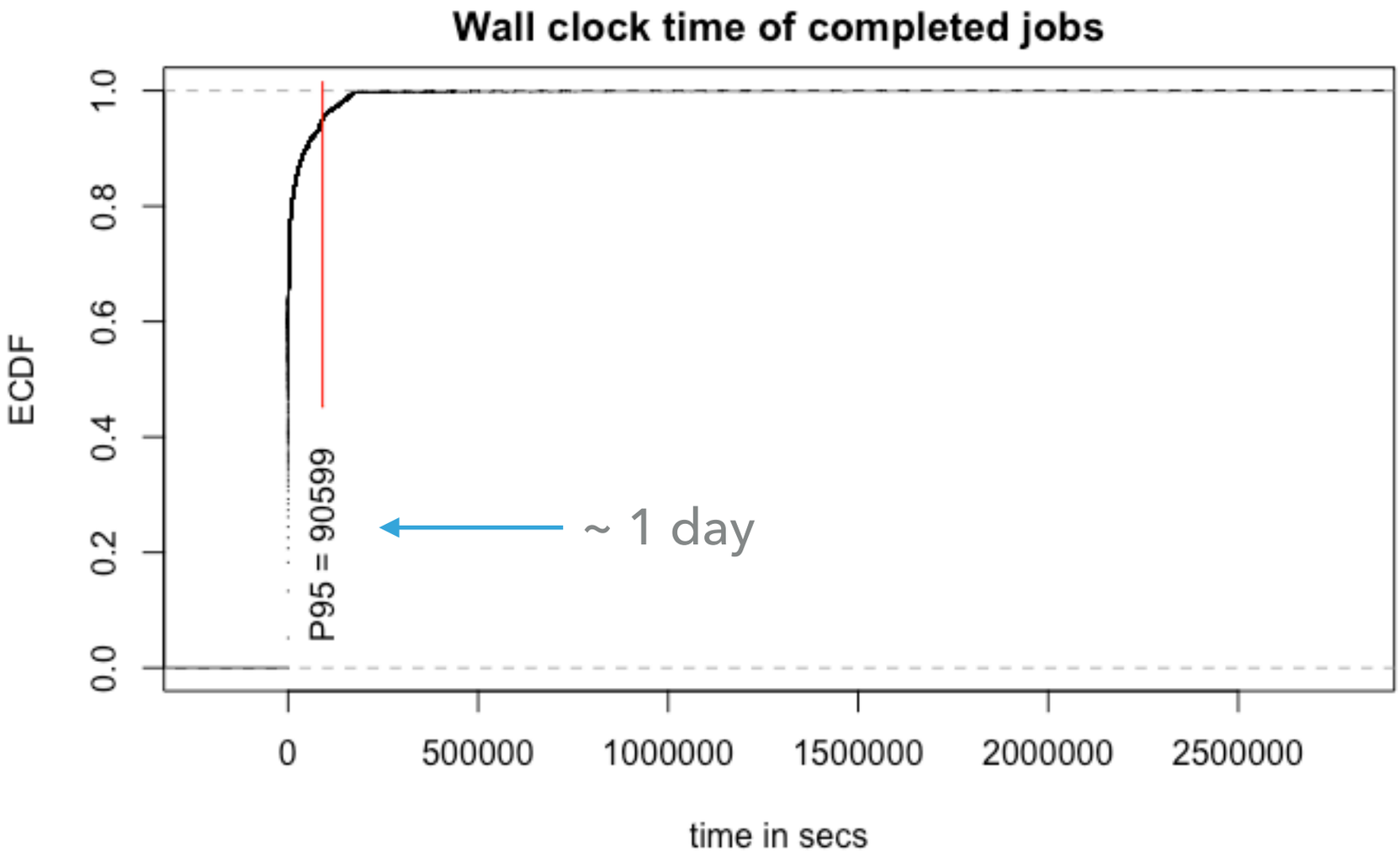
THE JOBS' BEHAVIOR (CONTINUED)

- Overall statistics from Aug/2016 to May/2018
  - Percentage of completed jobs in each partition

Partition name	Total number of jobs	% of total
cpu	34856	49,89 %
cpu_dev	21858	31,29 %
nvidia	9049	12,95 %
nvidia_dev	2115	3,03 %
mesca2	776	1,11 %
cpu_long	608	0,87 %
cpu_scal	467	0,67 %
nvidia_long	68	0,10 %
nvidia_scal	68	0,10 %

THE JOBS' BEHAVIOR (CONTINUED)

- **Wall-clock time** statistics from Aug/2016 to May/2018



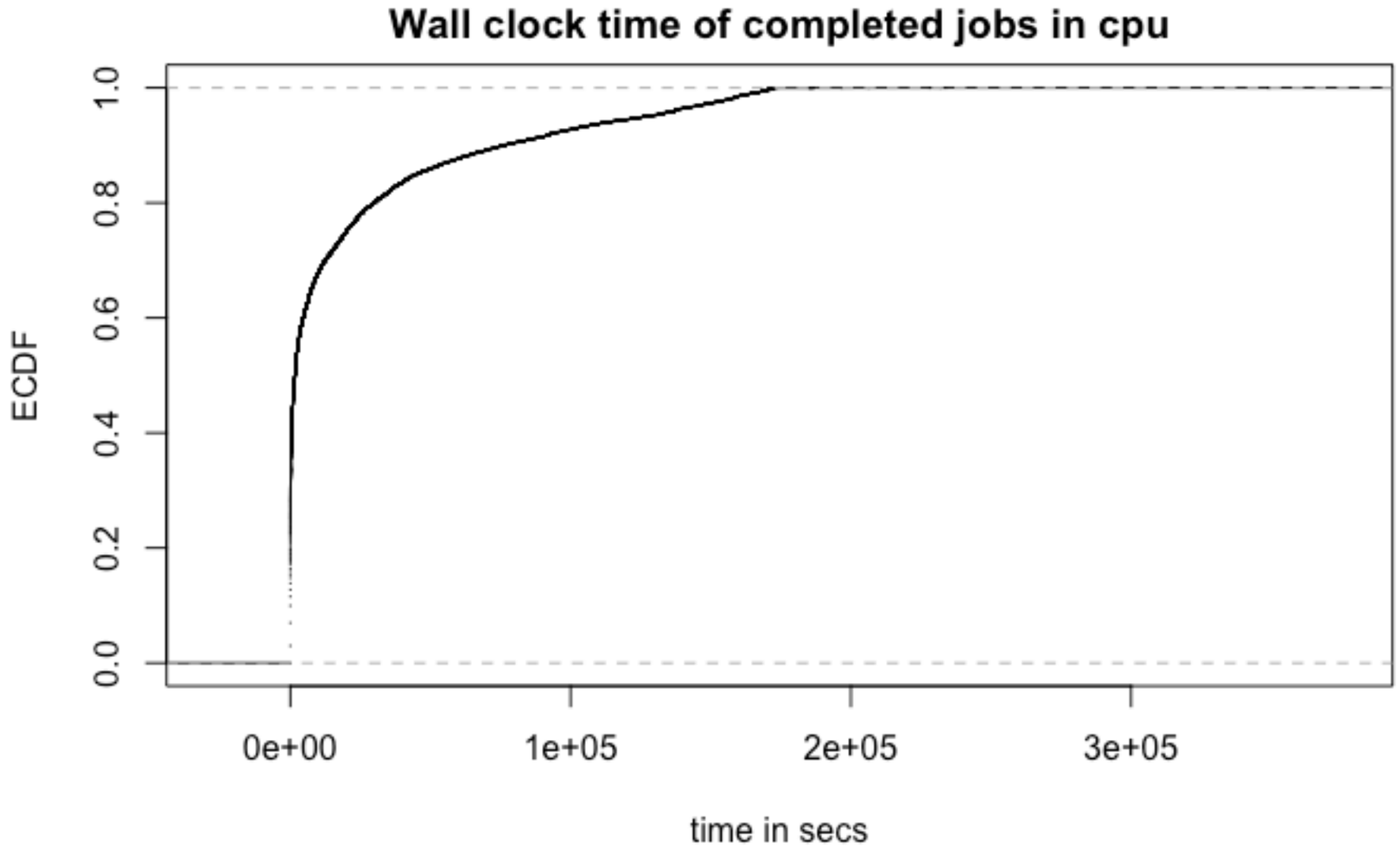
Quartile	
0 %	0
25 %	6
50 %	95
75 %	4224
100 %	2584100

~ 1 hour

~ 30 days

THE JOBS' BEHAVIOR (CONTINUED)

- Wall-clock time statistics from Aug/2016 to May/2018



Quartile	
0 %	0
25 %	45
50 %	1666
75 %	19972
100 %	172800

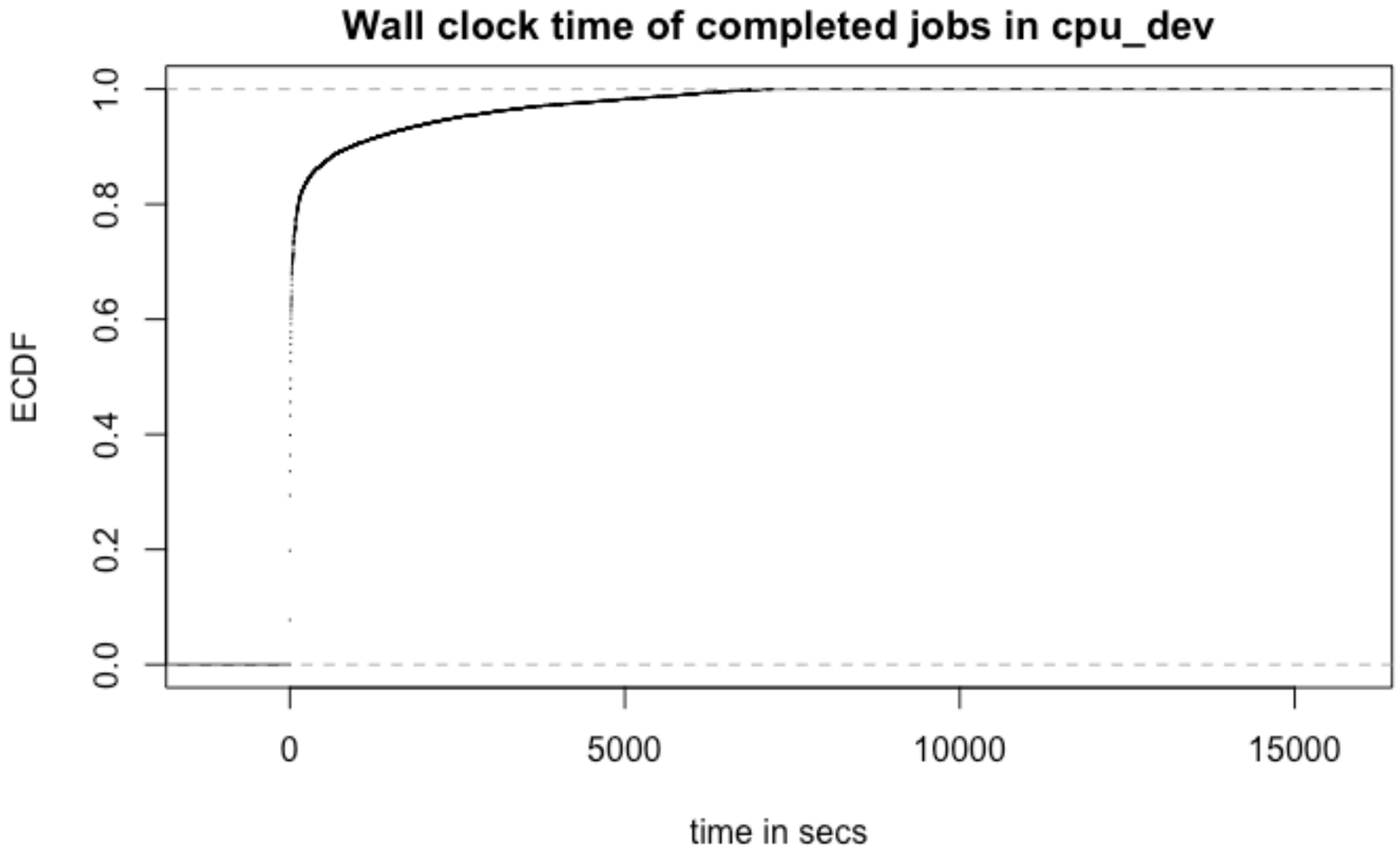
~ 26 min

~ 6 hours

48 hours (max)

THE JOBS' BEHAVIOR (CONTINUED)

- Wall-clock time statistics from Aug/2016 to May/2018

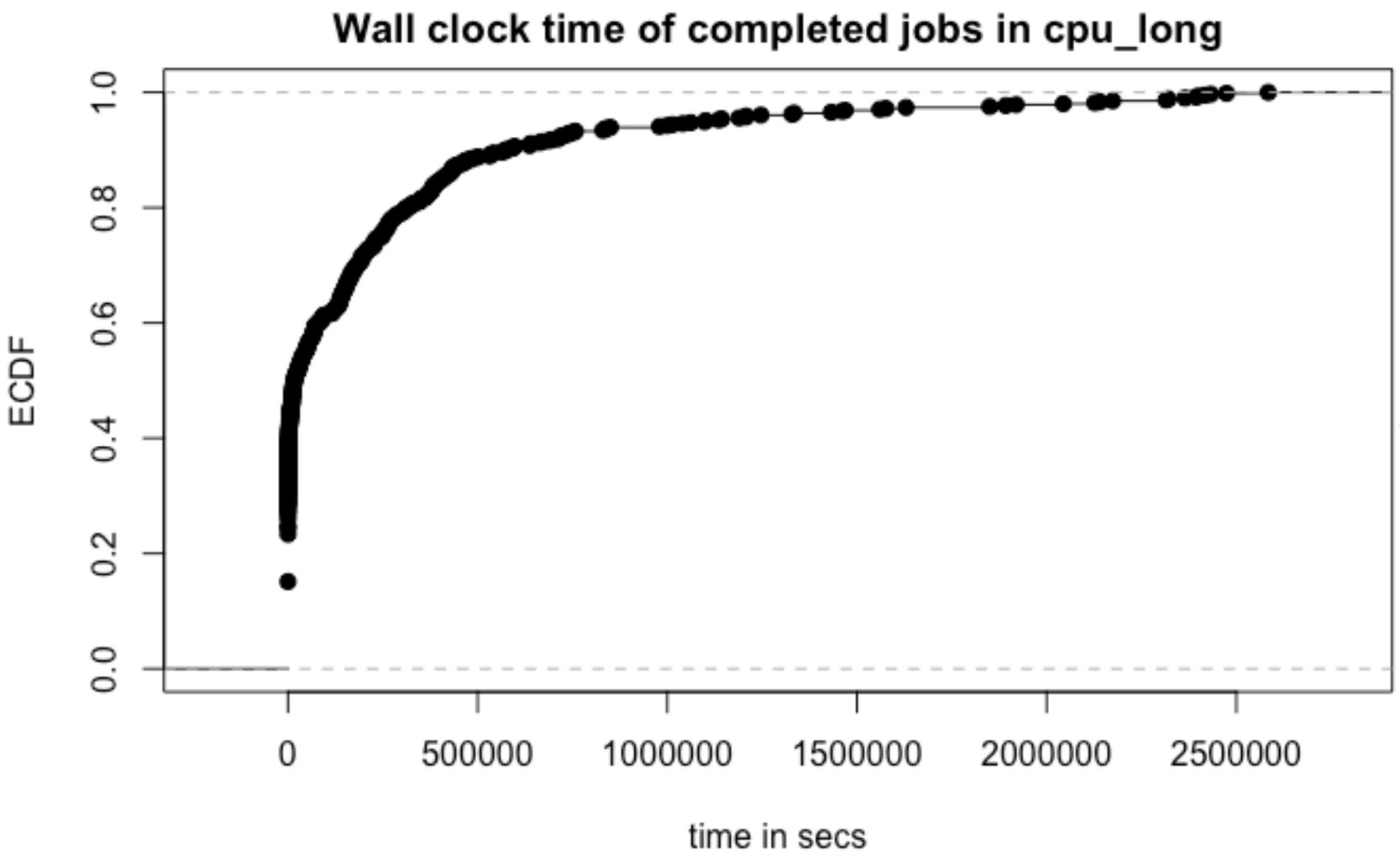


Quartile	
0 %	0
25 %	2
50 %	10
75 %	69
100 %	7200

2 hours (max)

THE JOBS' BEHAVIOR (CONTINUED)

- Wall-clock time statistics from Aug/2016 to May/2018



~ 48 hours (!!!)

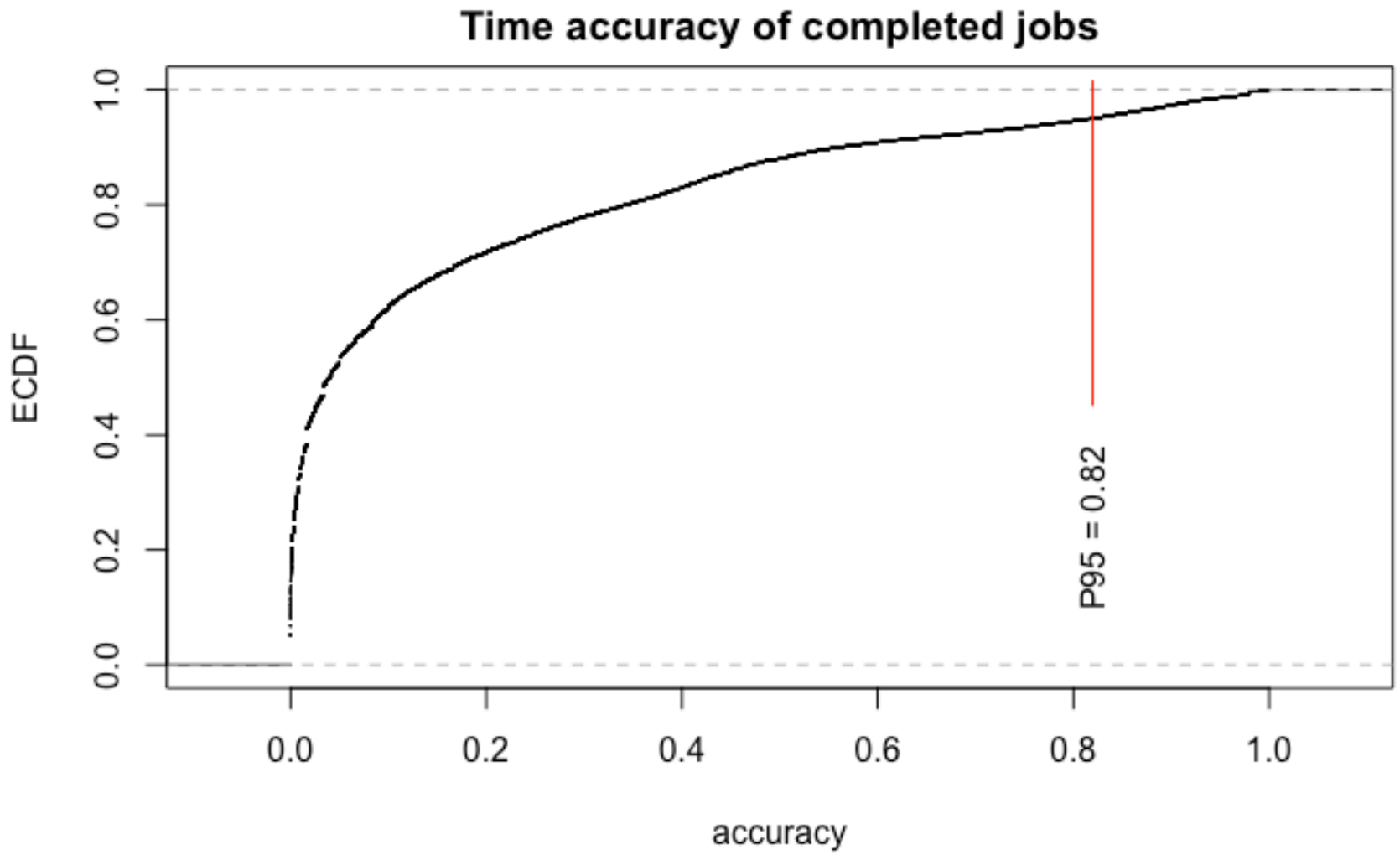
< 7 days (!!!)

~ 30 days

Decile	
0 %	0
10 %	0
20 %	1
30 %	39
40 %	2034
50 %	16944
60 %	84208
70 %	183507
80 %	319424
90 %	579019
100 %	2584100

THE USERS' BEHAVIOR

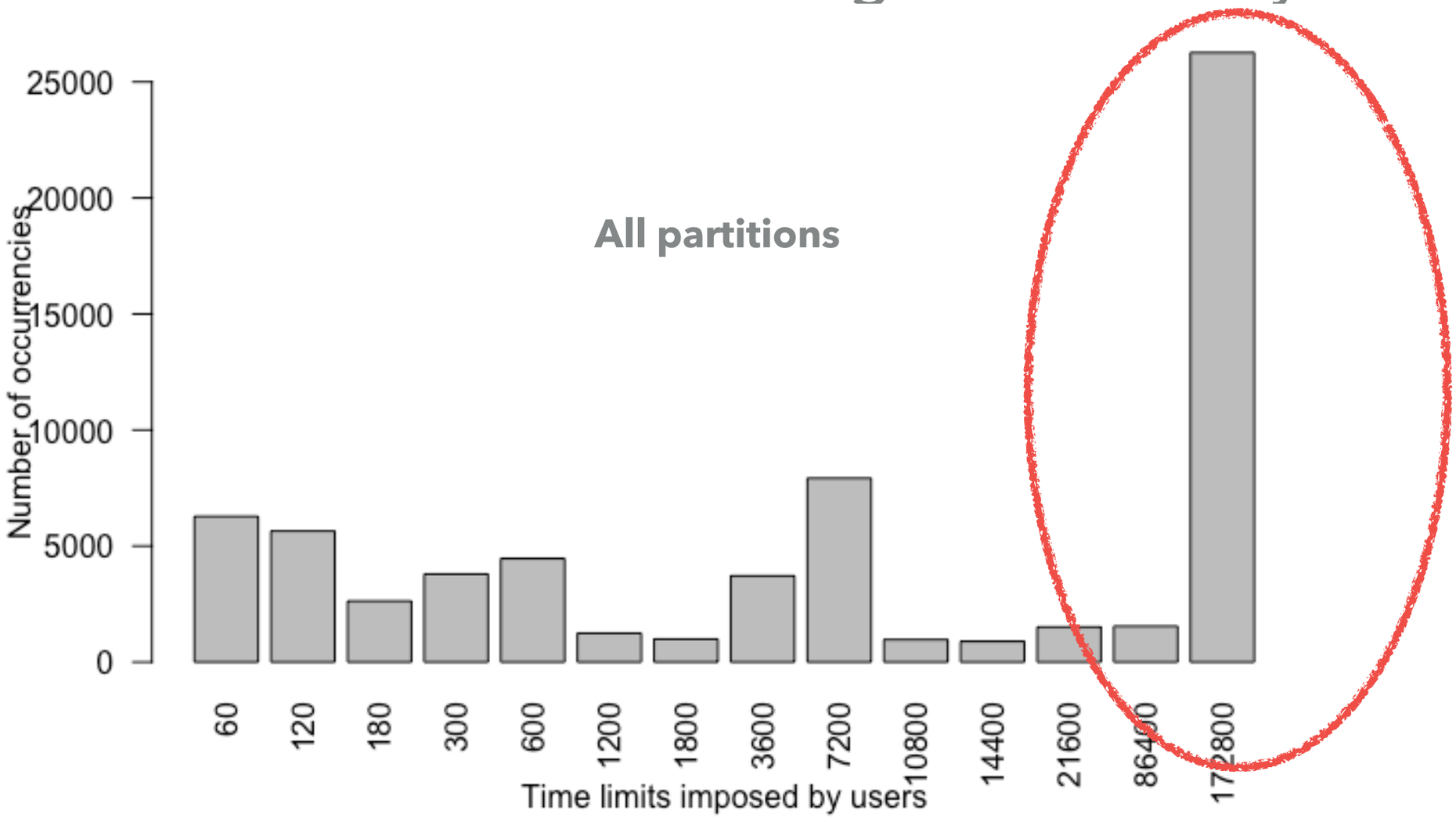
- **Estimated time** statistics from Aug/2016 to May/2018



Quartile	
0 %	0,00
25 %	0,00
50 %	0,04
75 %	0,25
100 %	1,00

THE USERS' BEHAVIOR (CONTINUED)

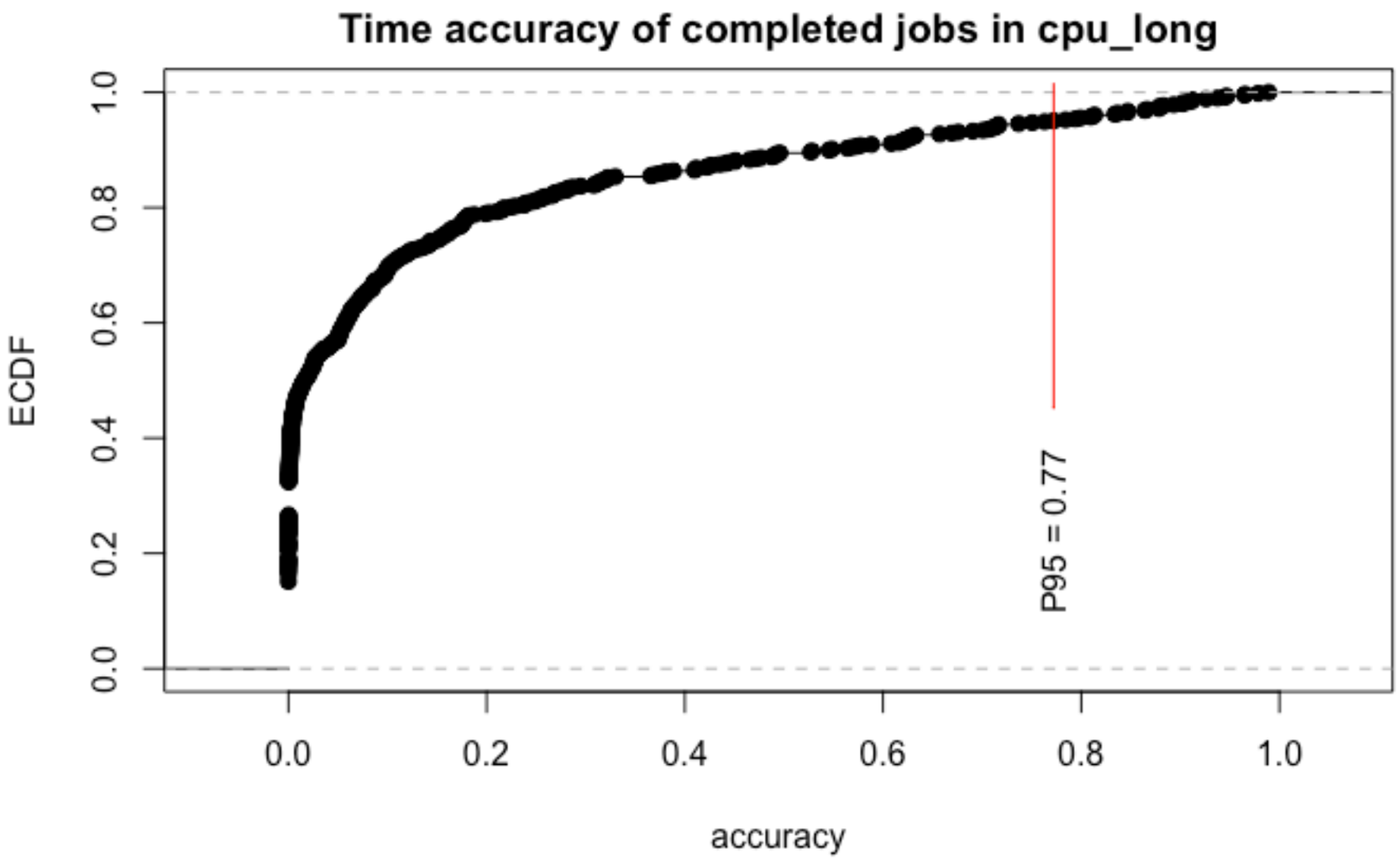
- Estimated time statistics from Aug/2016 to May/2018



* only those with more than 500 occurrences

THE USERS' BEHAVIOR (CONTINUED)

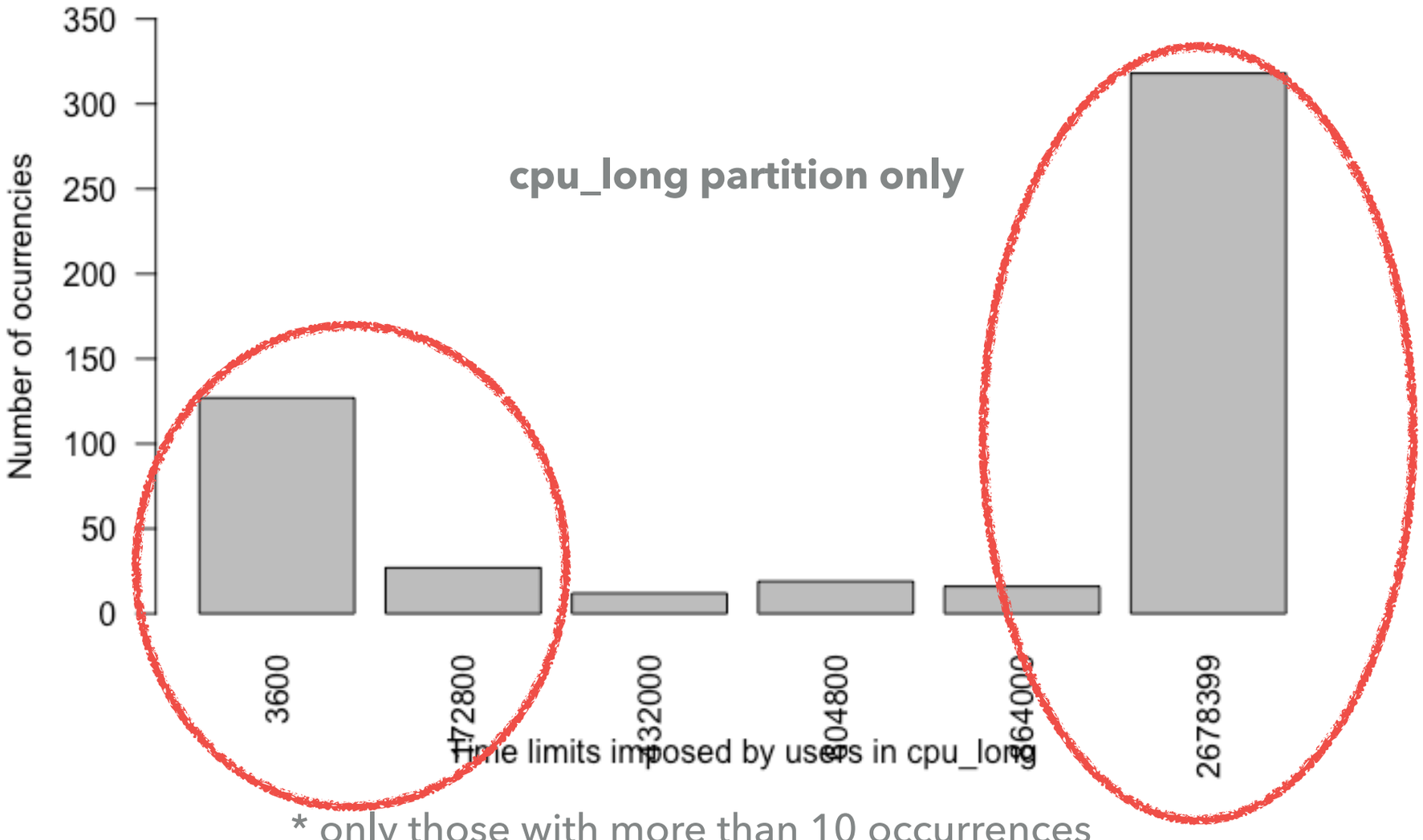
- Estimated time statistics from Aug/2016 to May/2018



	Decile
0 %	0,000000
10 %	0,000000
20 %	0,000006
30 %	0,000278
40 %	0,002451
50 %	0,016416
60 %	0,057696
70 %	0,103508
80 %	0,224877
90 %	0,548517
100 %	0,989172

THE USERS' BEHAVIOR (CONTINUED)

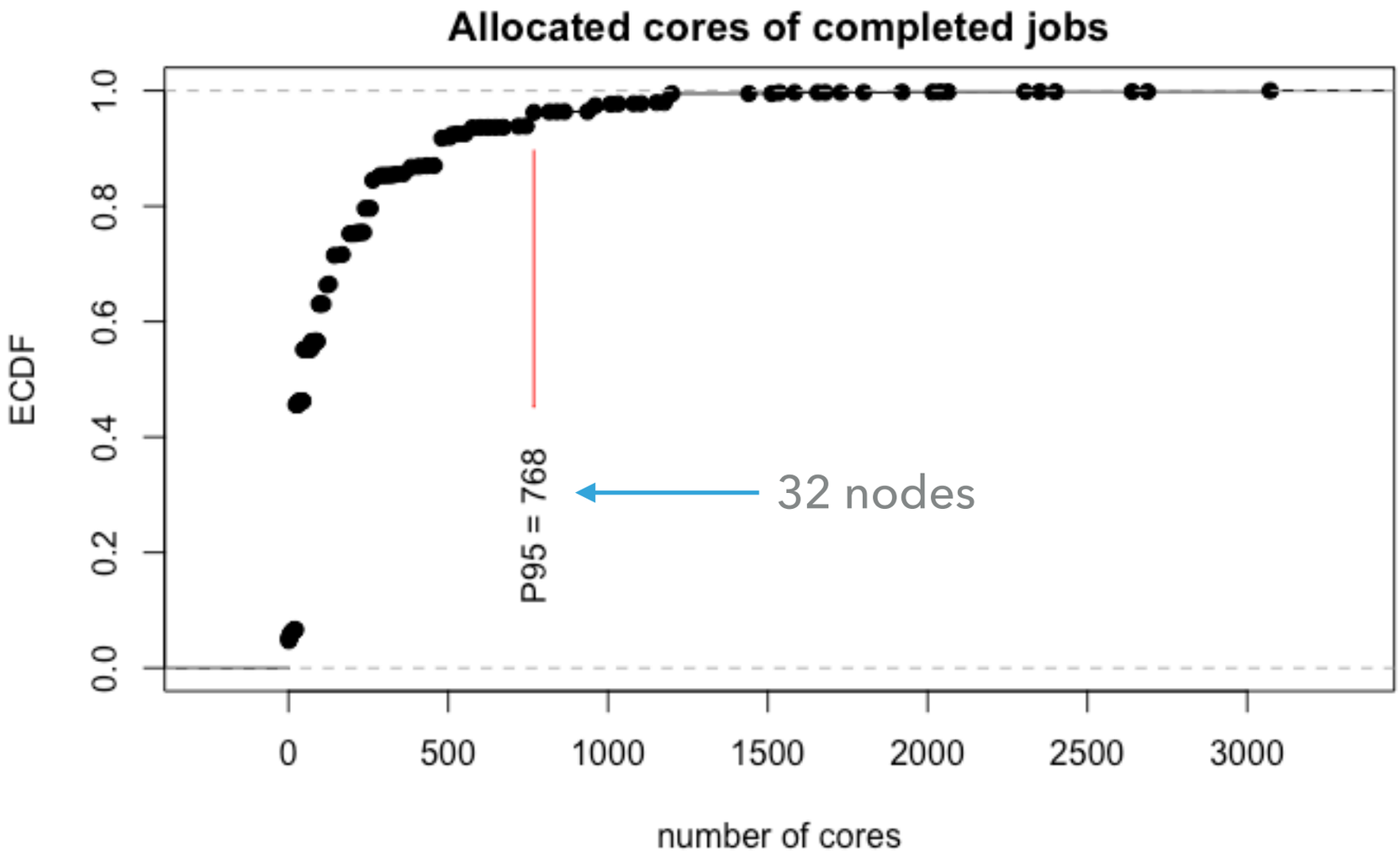
- Estimated time statistics from Aug/2016 to May/2018



* only those with more than 10 occurrences

THE USERS' BEHAVIOR (CONTINUED)

- **Core allocation** statistics from Aug/2016 to May/2018

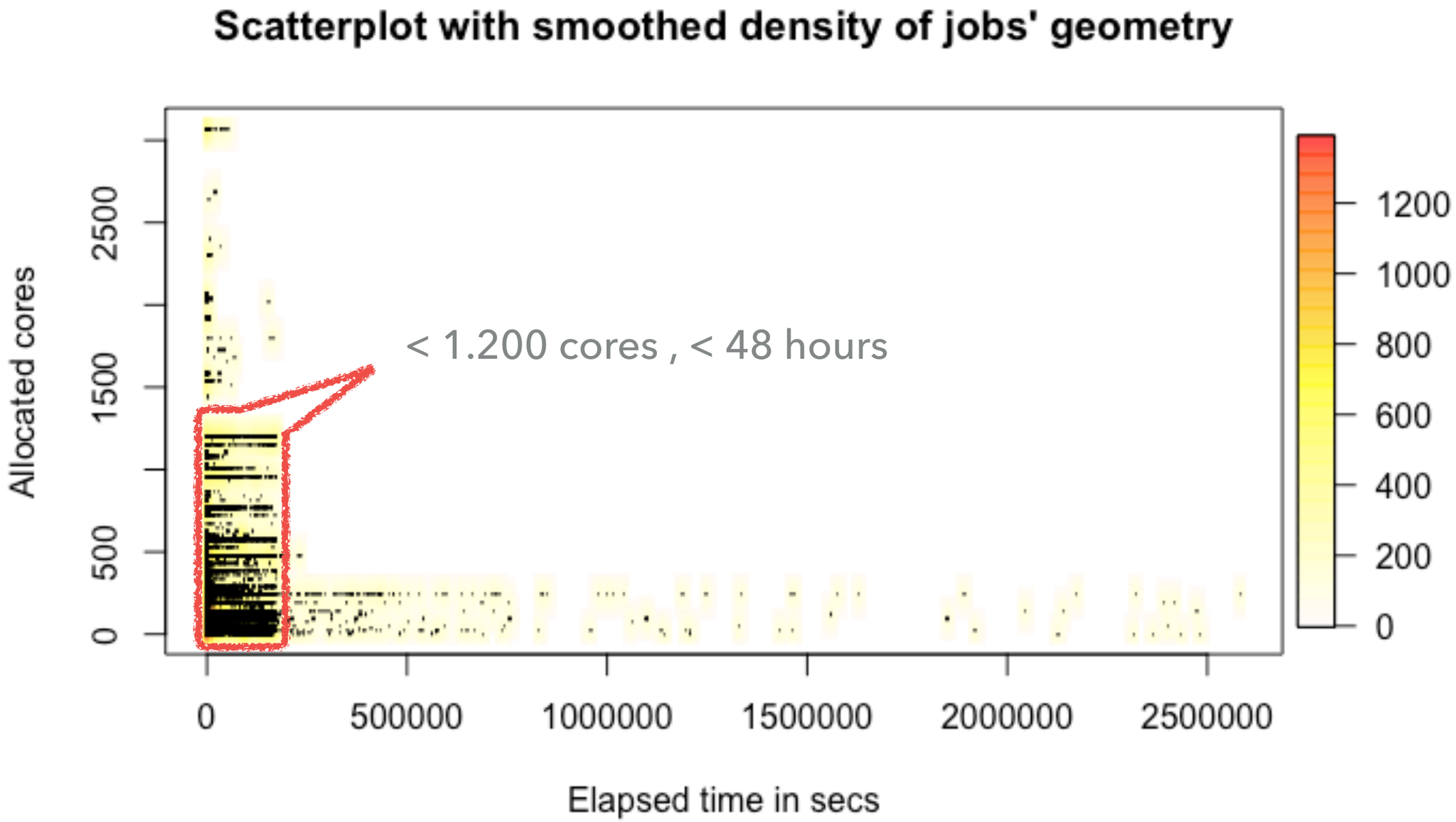


Serial jobs?

Quartile	
0 %	1
25 %	24
50 %	48
75 %	192
100 %	3072

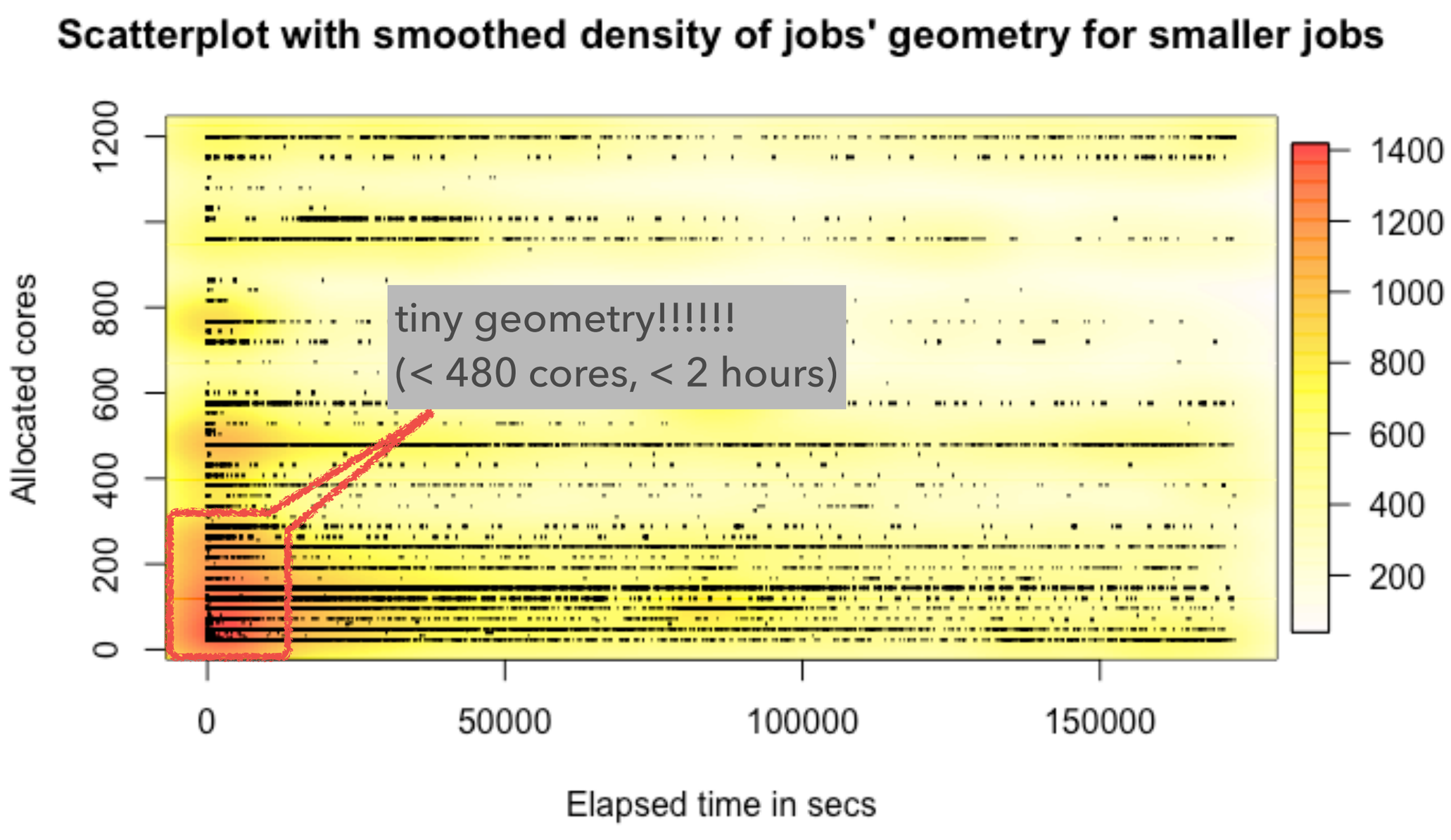
THE USERS' VERSUS JOBS' BEHAVIOR

- **Job geometry** statistics from Aug/2016 to May/2018



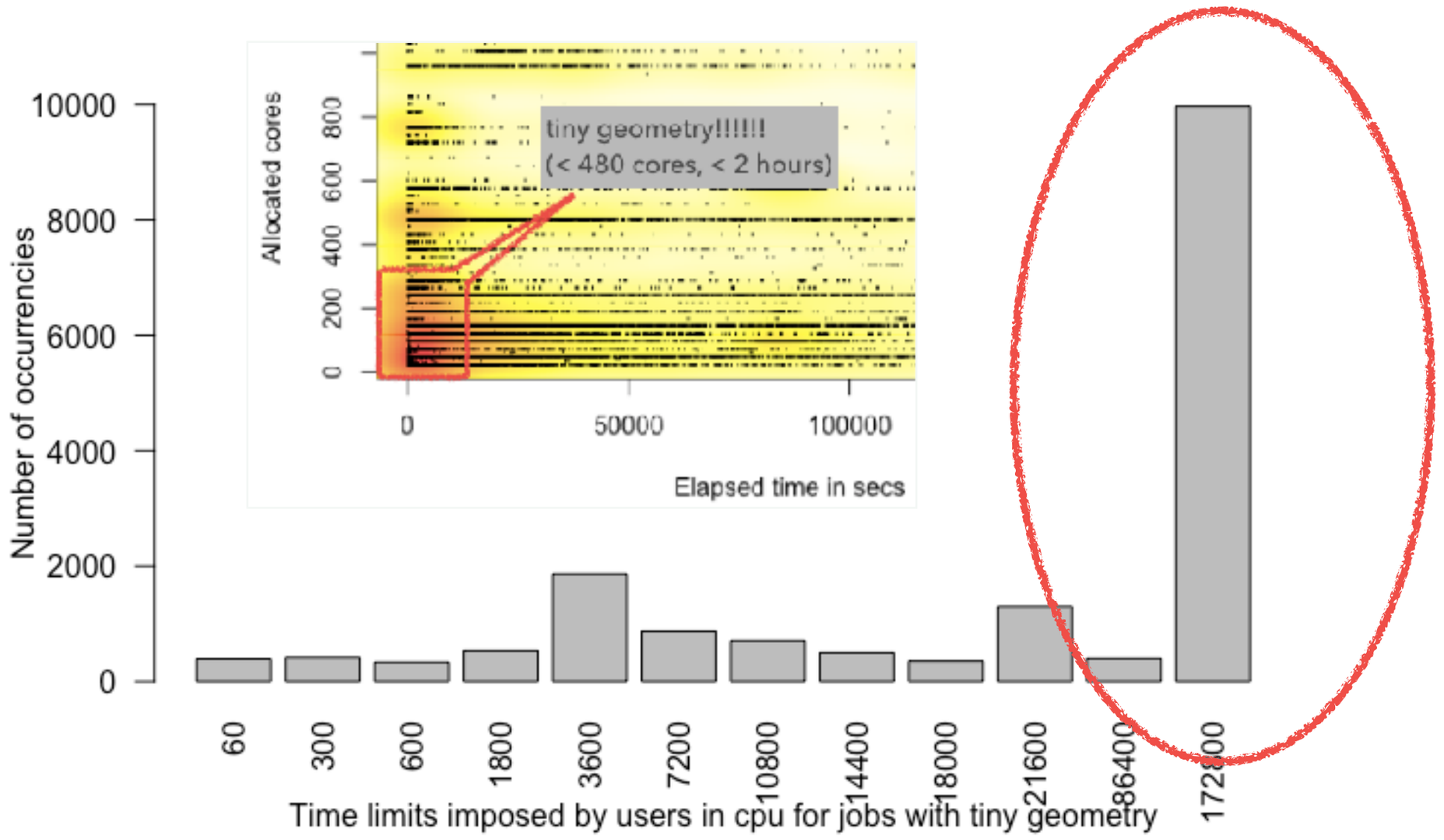
THE USERS' VERSUS JOBS' BEHAVIOR (CONTINUED)

- Job geometry statistics from Aug/2016 to May/2018



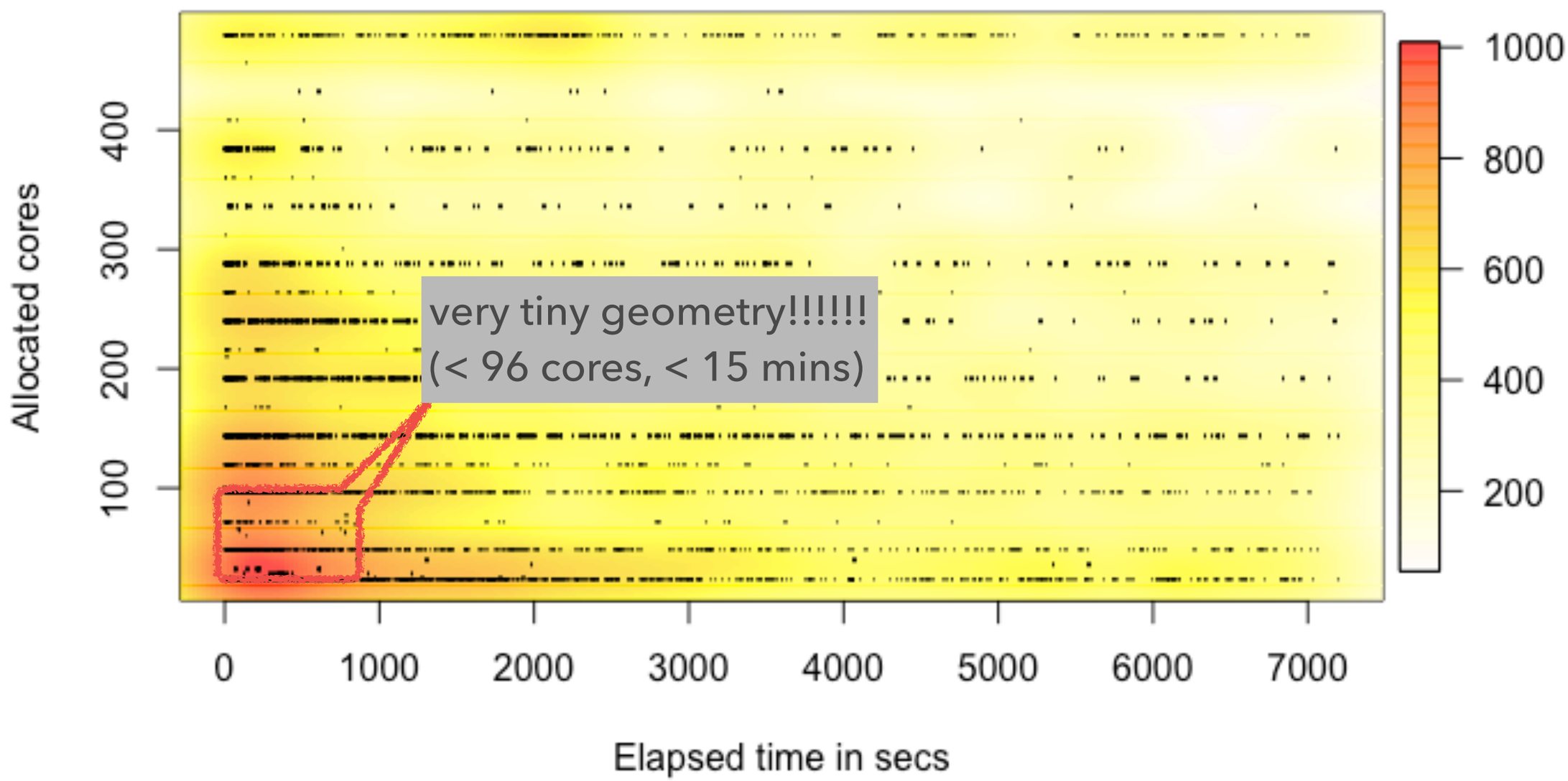
(BACK TO) THE USERS' BEHAVIOR

- **Estimated time** statistics from Aug/2016 to May/2018



THE USERS' BEHAVIOR (CONTINUED)

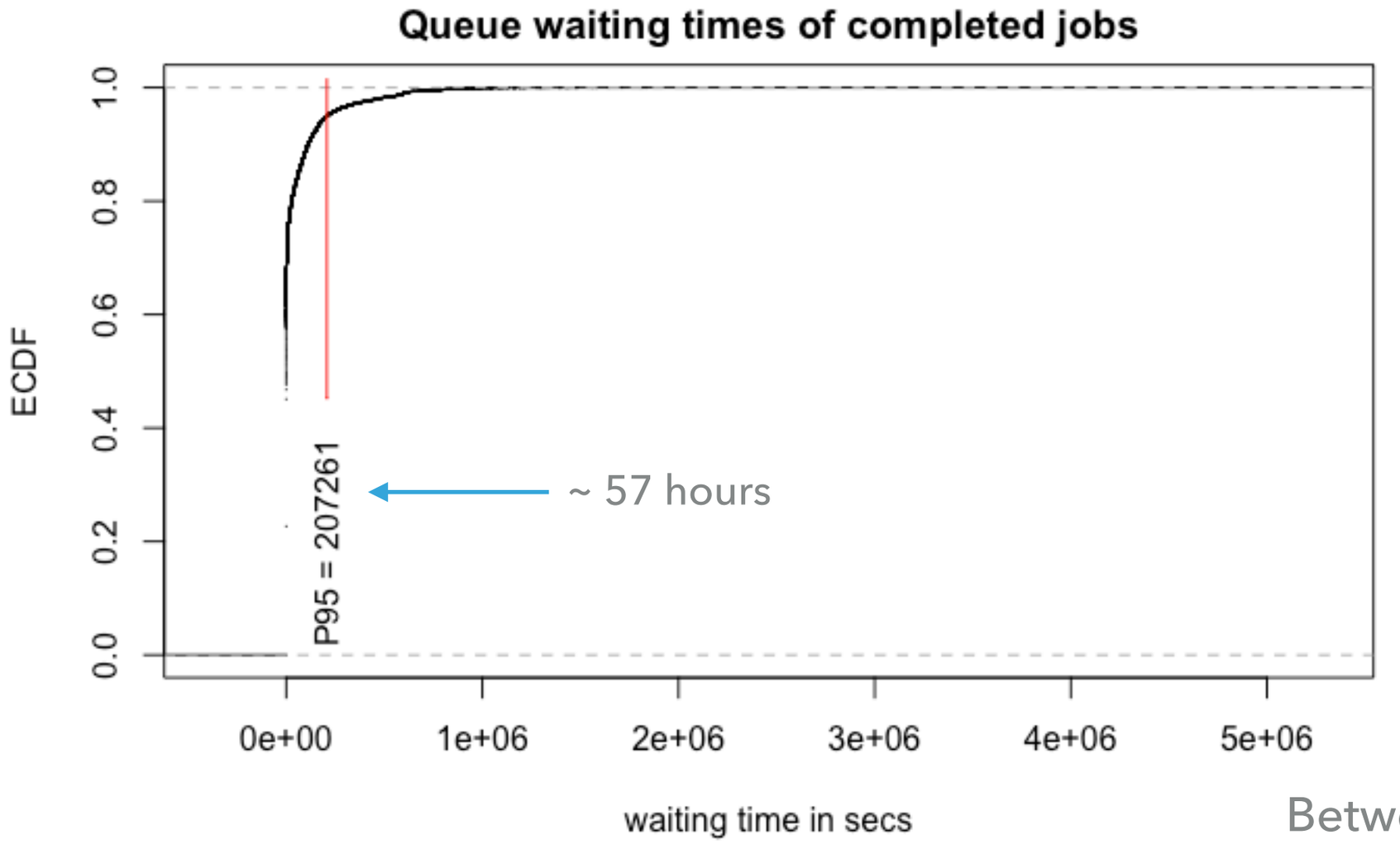
- Estimated time statistics from Aug/2016 to May/2018



BUT WHY SHOULD USERS BOTHER?

THE SYSTEMS' BEHAVIOR

- **Queue waiting time** statistics from Aug/2016 to May/2018

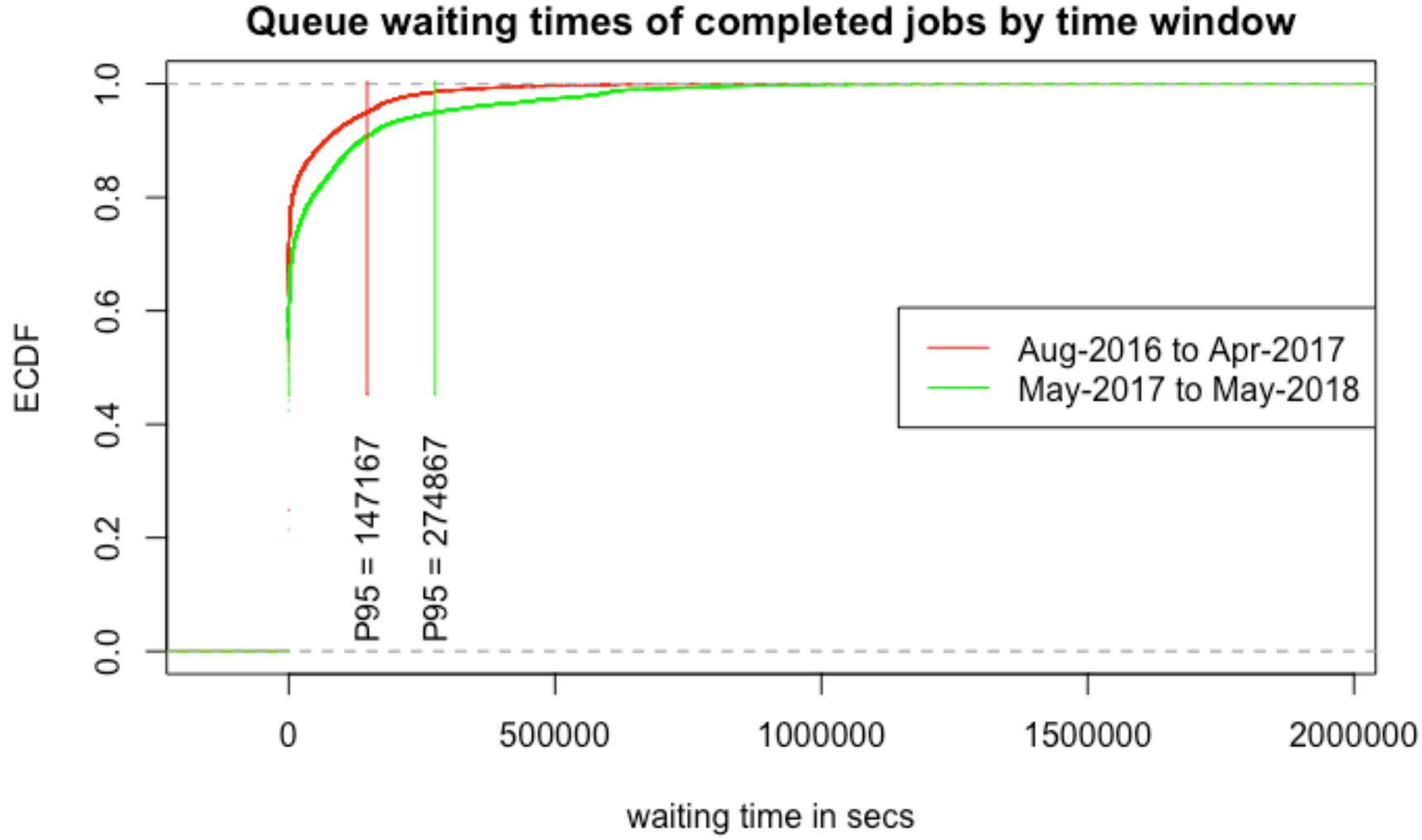


Decile	
0 %	0
10 %	0
20 %	0
30 %	1
40 %	1
50 %	10
60 %	111
70 %	2554
80 %	28055
90 %	112358
100 %	4920842

Between
1 and 23 days!

THE SYSTEMS' BEHAVIOR (CONTINUED)

- Split statistics from Aug/2016 to Apr/2017 (after 1st call) and from May/2017 to May/2018 (after 2nd call)



Design by Harry Movie Art

CAN WE HELP?

REVISITING THE SCHEDULING POLICIES

Partition	Max W.C.T (hours)	Min # cores	Max # cores	Max # executing jobs per user	Max # enqueued jobs per user
cpu	48	**504**	1200	4	24
nvidia	48	**504**	1200	4	24
phi	48	**504**	1200	4	24
mesca2	48	**1**	240	1	6
cpu_dev	~~2~~ **0,3**	**24**	~~480~~ **96**	1	1
nvidia_dev	~~2~~ **0,3**	**24**	~~480~~ **96**	1	1
phi_dev	~~2~~ **0,3**	**24**	~~480~~ **96**	1	1
cpu_scal	18	**1224**	3072	1	8
nvidia_scal	18	**1224**	3072	1	8
cpu_long	744	**24**	240	1	1
nvidia_long	744	**24**	240	1	1
cpu_small	**2**	**24**	**480**	**4**	**24**
nvidia_small	**2**	**24**	**480**	**4**	**24**

REVISITING THE SCHEDULING POLICIES (CONTINUED)

- "Non-exclusive mode" for mesca2 partition
- Default time estimation = 1/2 max W.C.T.

Entered in operation in June/2018

THE JOBS' BEHAVIOR

- Overall statistics from Jun/2018 to Sep/2018
  - Percentage of completed jobs in each partition

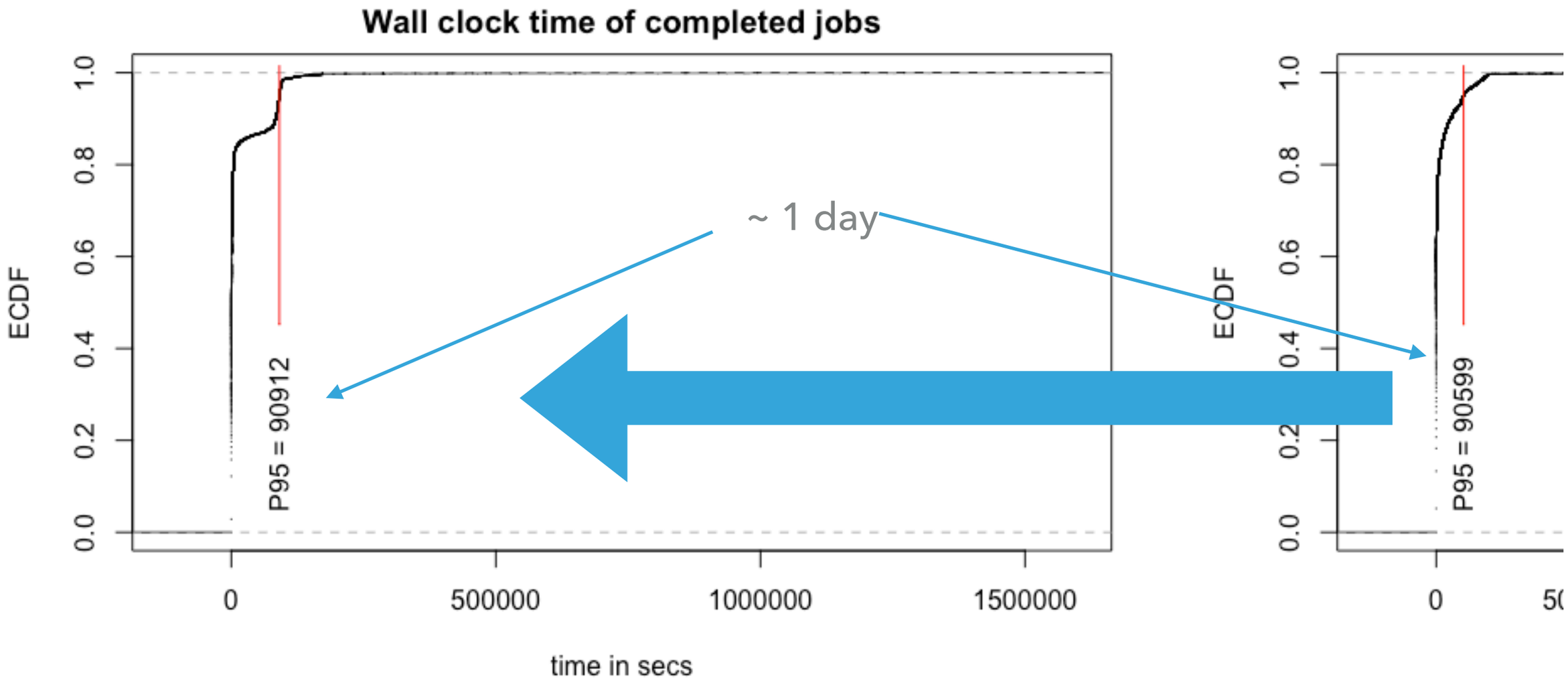
Partition name	Total number of jobs	% of total
cpu	34856	49,89 %
cpu_dev	21858	31,29 %

...

Partition name	Total number of jobs	% of total
cpu_small	11204	55 %
cpu_dev	4621	23 %
cpu	1606	8 %
nvidia_dev	1009	5 %
nvidia_small	878	4 %
nvidia_long	286	1 %
nvidia	270	1 %
cpu_long	182	1 %
mesca2	142	1 %
cpu_scal	22	0 %
nvidia_scal	17	0 %

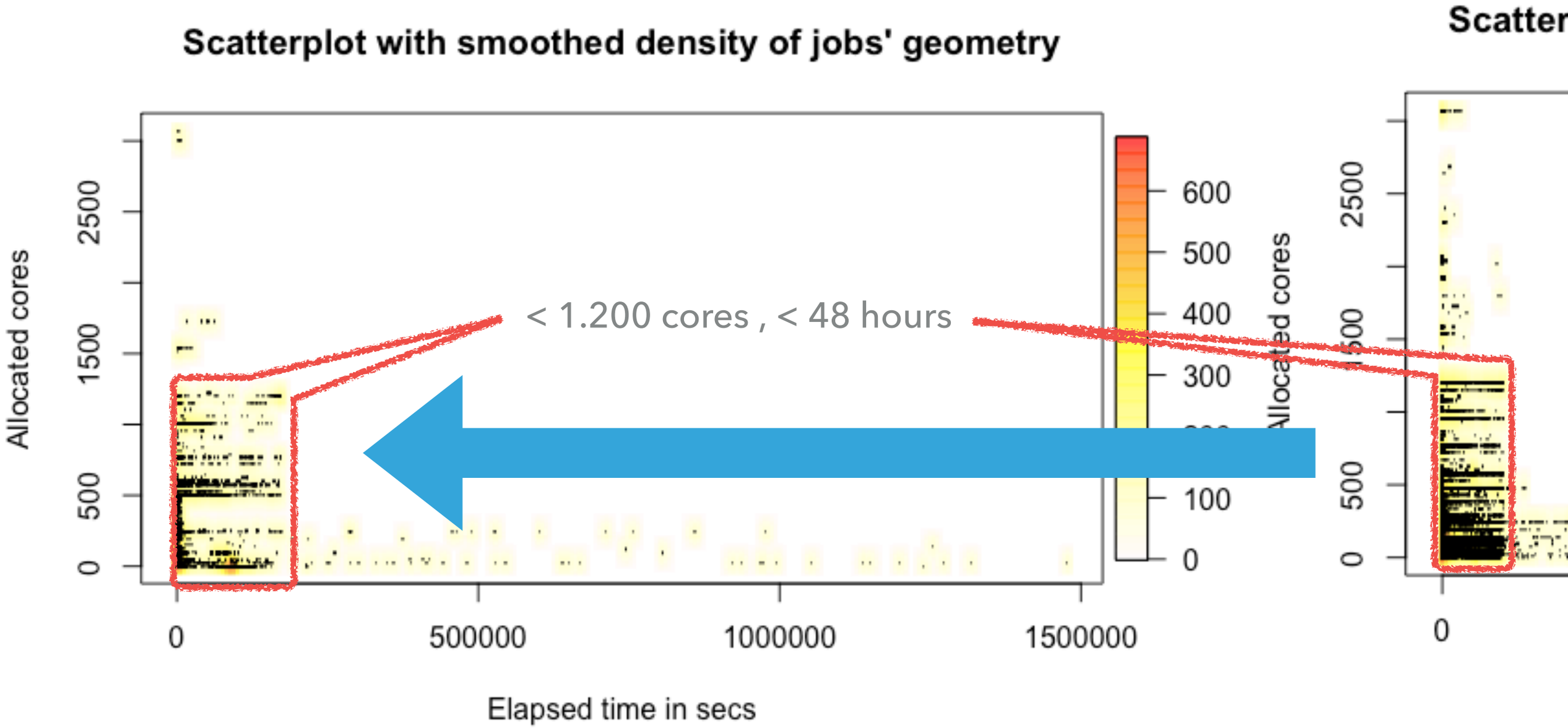
THE JOBS' BEHAVIOR (CONTINUED)

- **Wall-clock time** statistics from Jun/2018 to Sep/2018



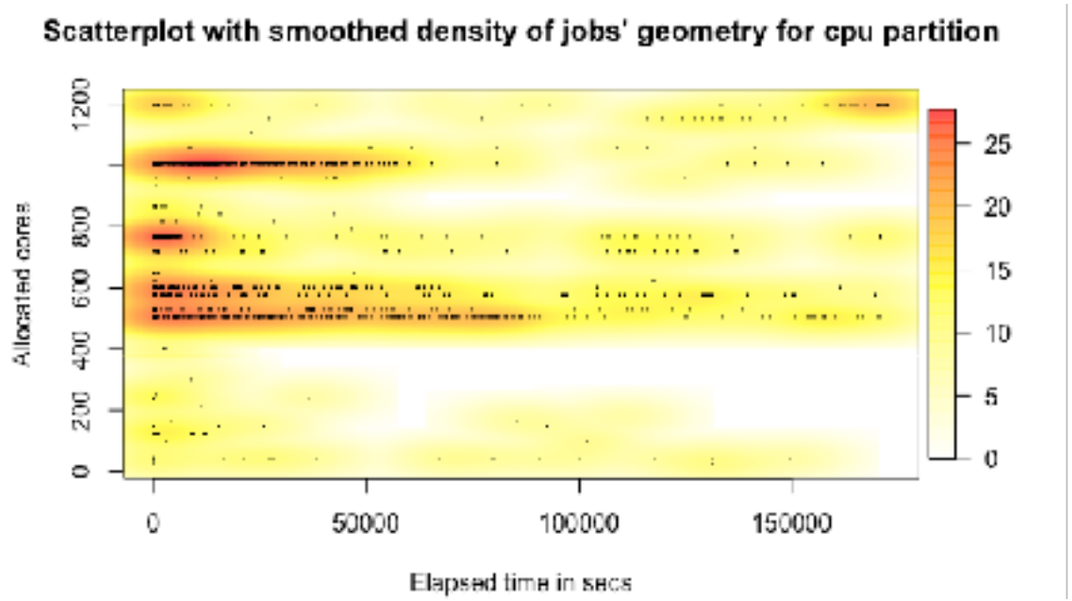
THE USERS' VERSUS JOBS' BEHAVIOR

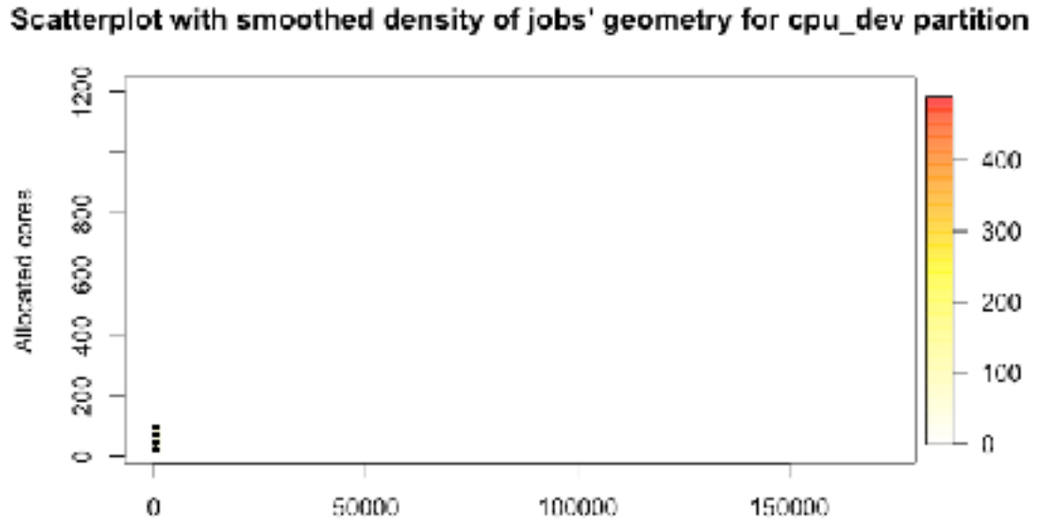
- **Job geometry** statistics from Jun/2018 to Sep/2018

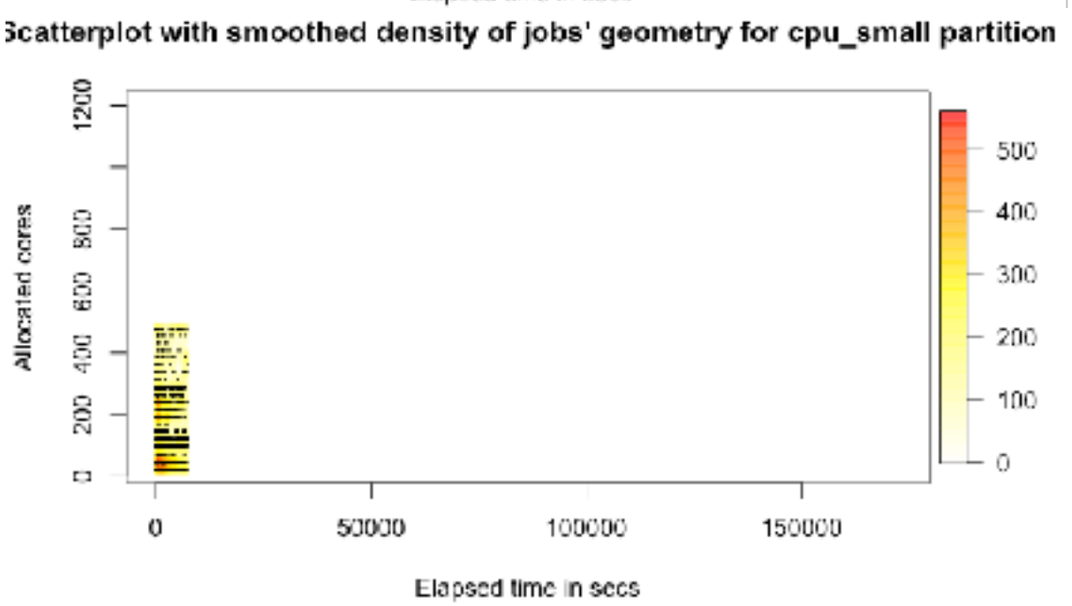


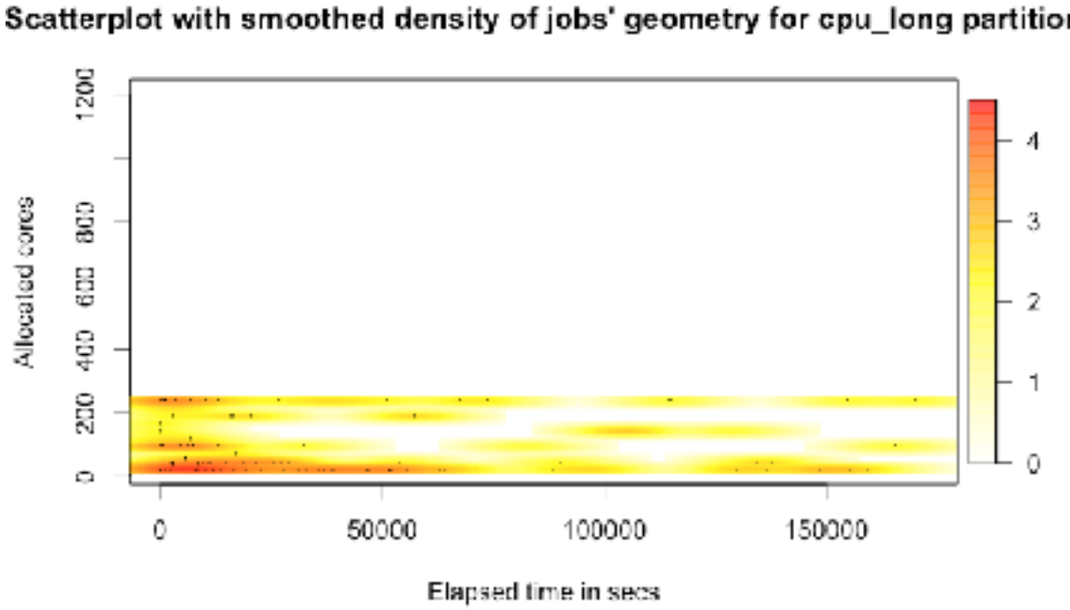
THE USERS' VERSUS JOBS' BEHAVIOR (CONTINUED)

- **Job geometry** statistics from Jun/2018 to Sep/2018



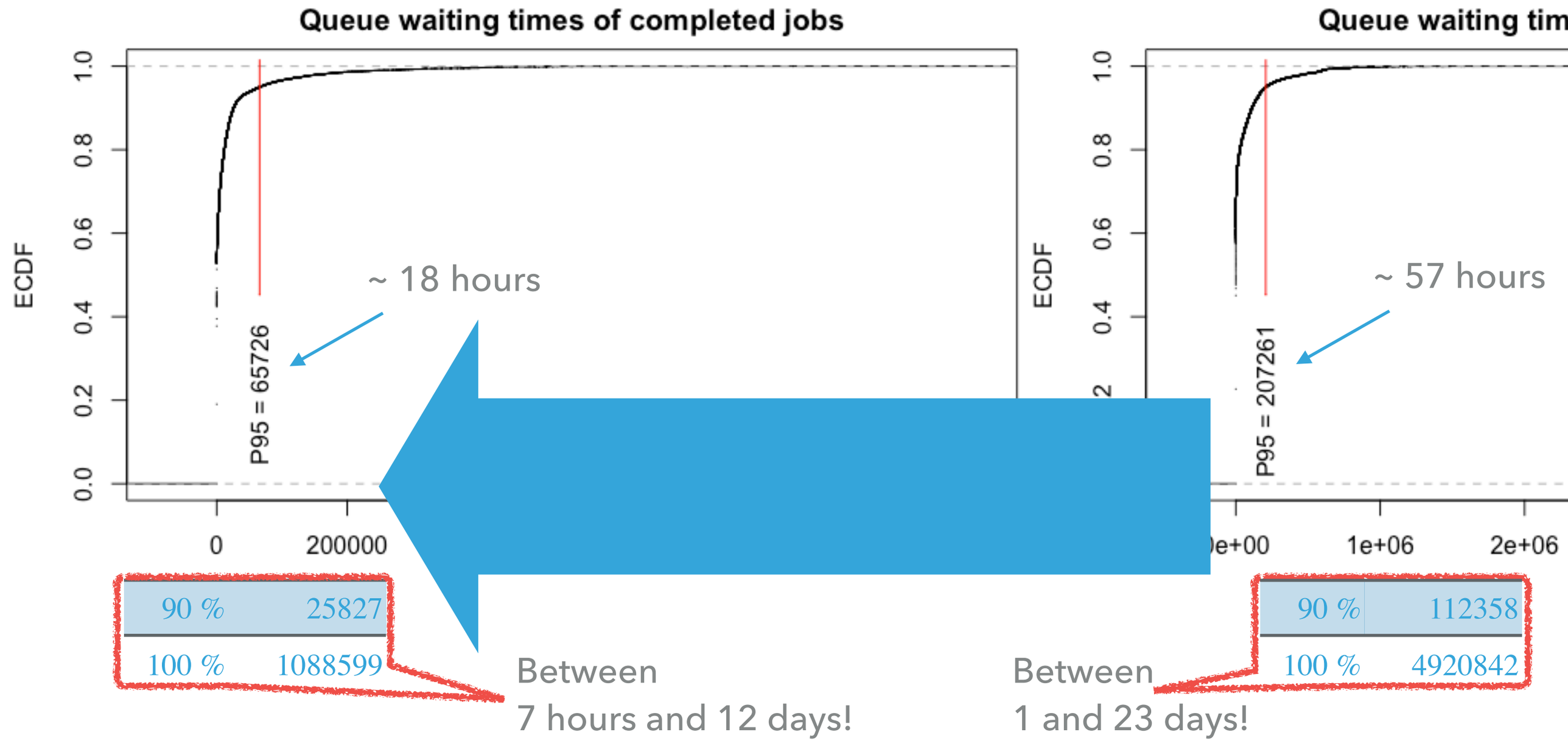






THE SYSTEMS' BEHAVIOR

- **Queue waiting time** statistics from Jun/2018 to Sep/2018



SUMMARY AND OUTLOOK

THE SINAPAD EXPERIENCE

- Demand is **clear**, updating is **flaky**
- Mismatch between **policy and action**
  - SINAPAD formal establishment

X

modus operandi of funding agencies

James Green on January 25, 2016 at 10:00 am

THE SDUMONT EXPERIENCE

- **Gap** between CSE researchers/ technologists and the application researchers is still huge
  - Efforts do exist (e.g. **HPC4e** project) but are not the norm
- Keeping the system operating the **best as possible** is a daunting task:
  - **Recommendation** systems
  - **Self-tuning** policies
  - Again, CSE researchers to the rescue!

Laboratório Nacional de Computação Científica

Finep
INOVAÇÃO E PESQUISA

Ministério da Ciência, Tecnologia e Inovação

HTTP://WWW.LNCC.BR

HTTP://SDUMONT.LNCC.BR

HTTPS://WWW.FACEBOOK.COM/SISTEMA-NACIONAL-DE-PROCESSAMENTO-DE-ALTO-DESEMPENHO-SINAPAD-135321166533790

THANK YOU! OBRIGADO!

WSCAD 2018