

Analyzing the EGEE production grid workload: application to jobs submission optimization

Diane Lingrand, Johan Montagnat, Janusz Martyniak and David Colling

London, UK

Imperial College
London

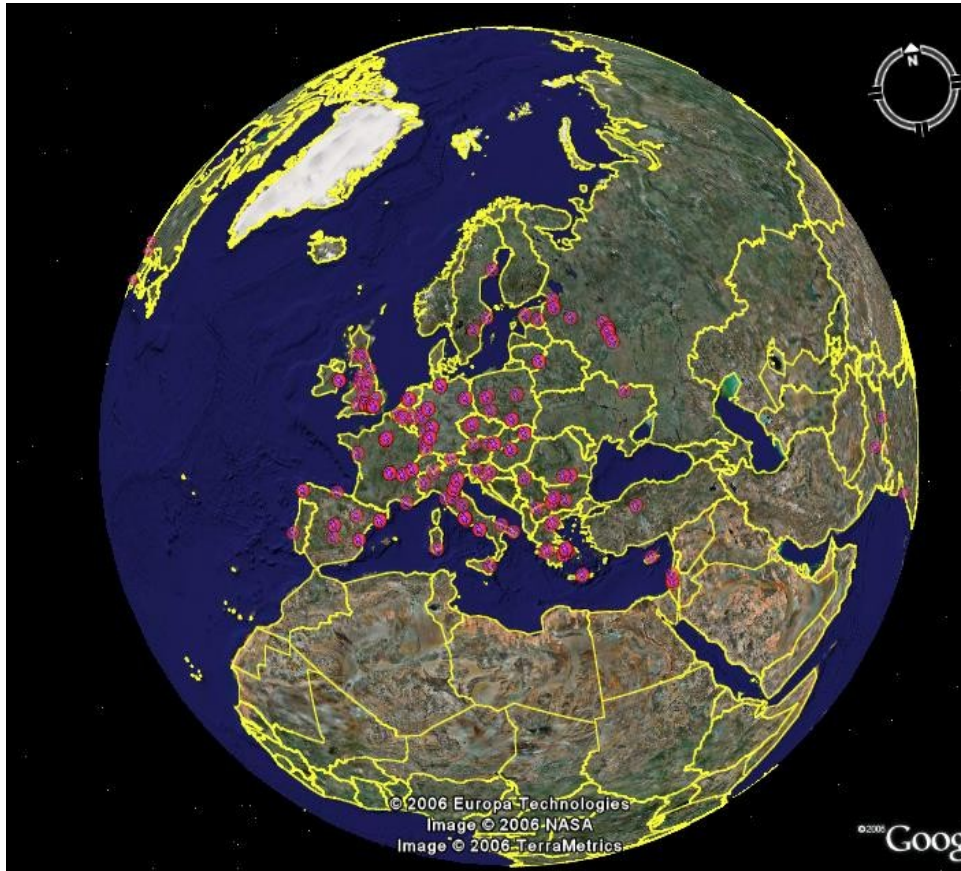
modalis

I3S, Sophia Antipolis, France



The EGEE production grid

<http://www.eu-egee.org>



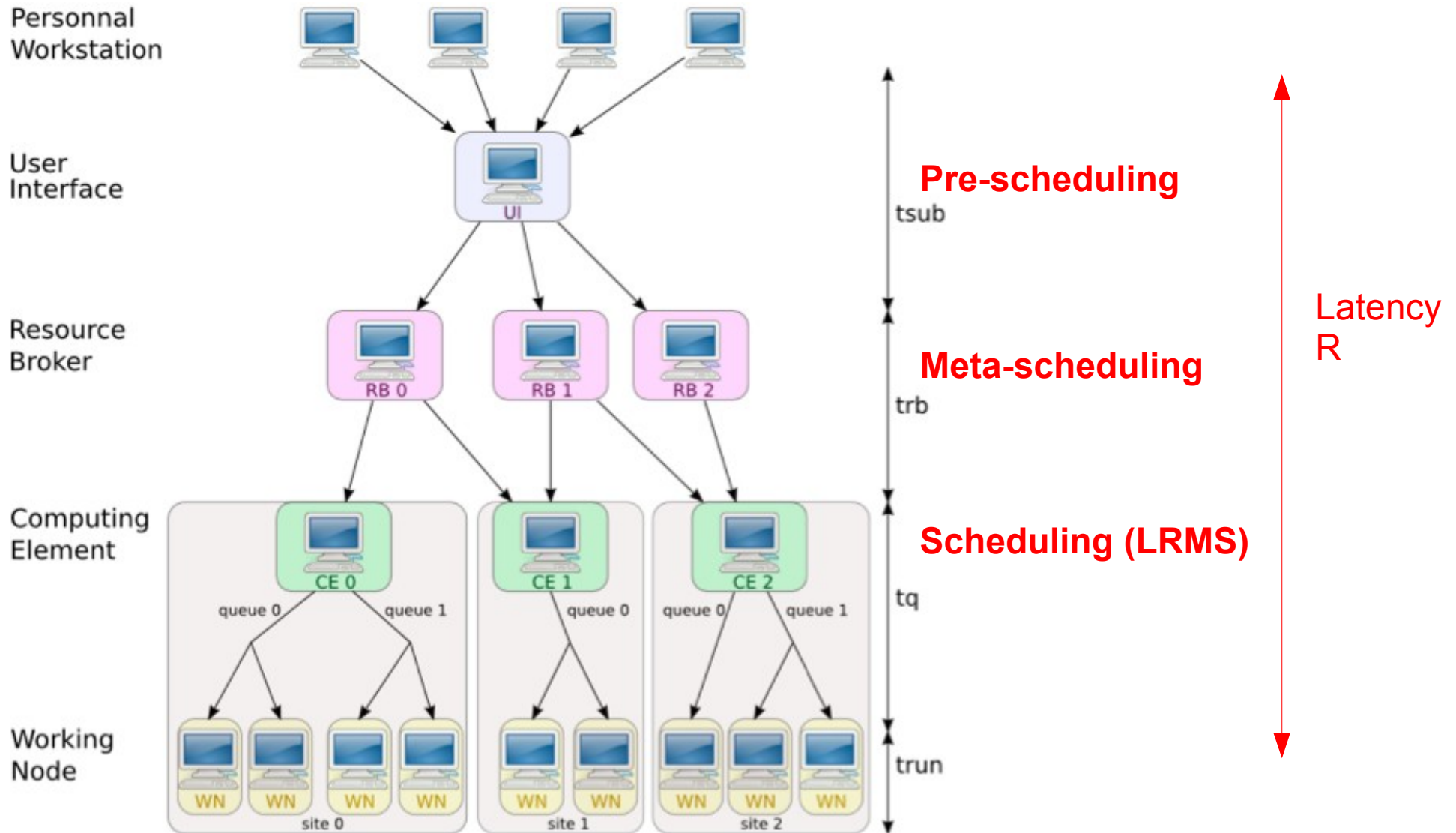
Huge computing power and data storage facility:

- > 100,000 CPUs
- > 300 computing centers worldwide
- > 300,000 jobs/day
- > 9,000 registered users

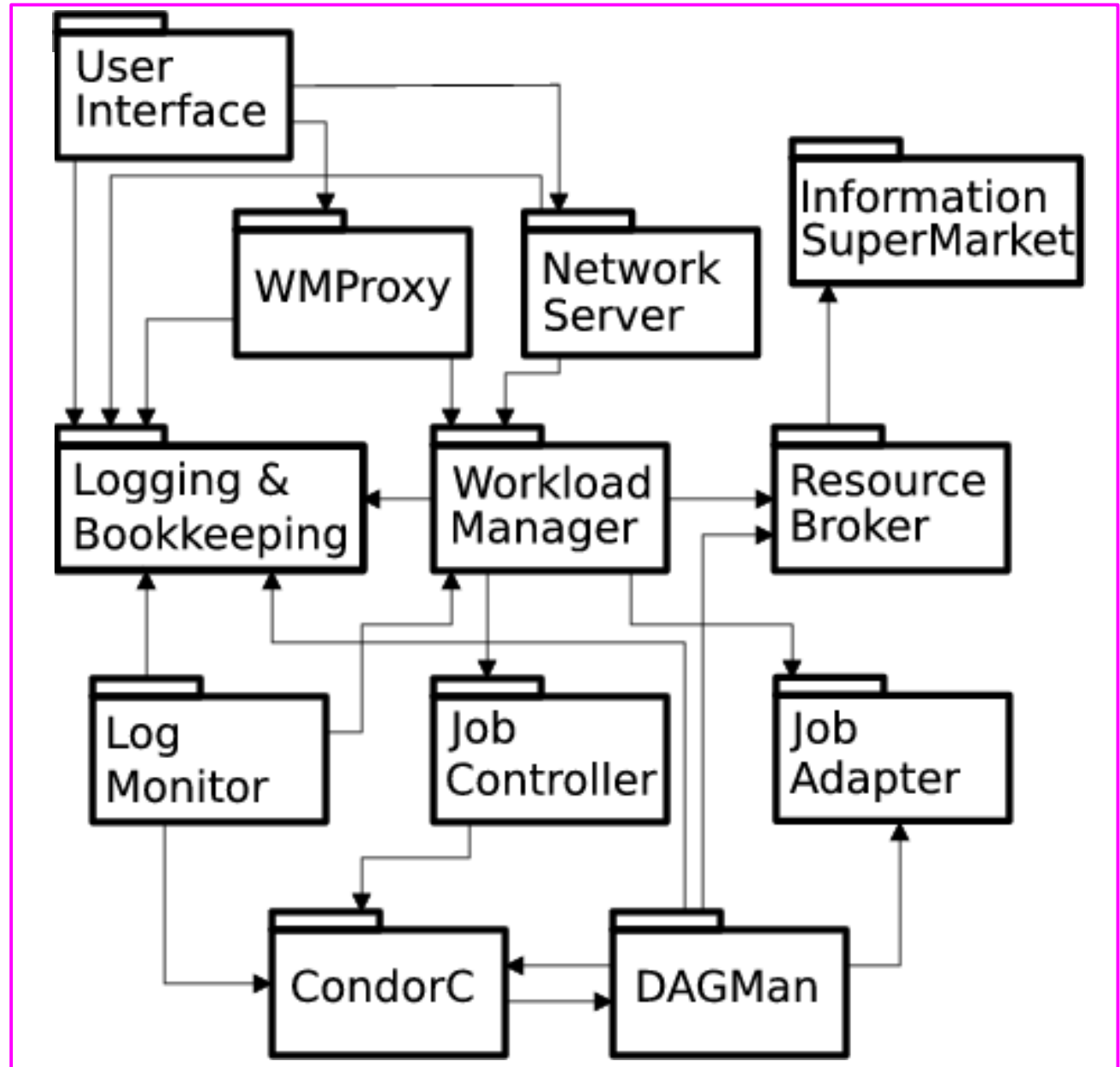
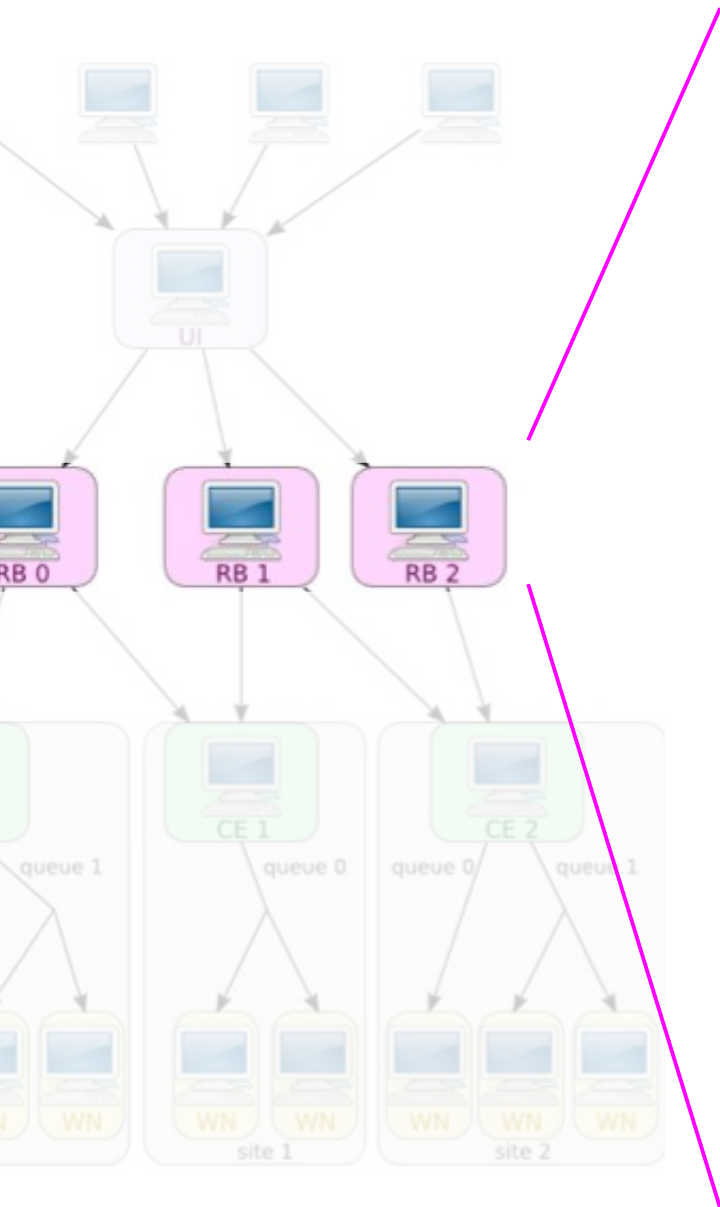
- Drawbacks

- Behavior hardly understood
- Many faults, difficult to track
- Variability, performances difficult to control

System overview: job management

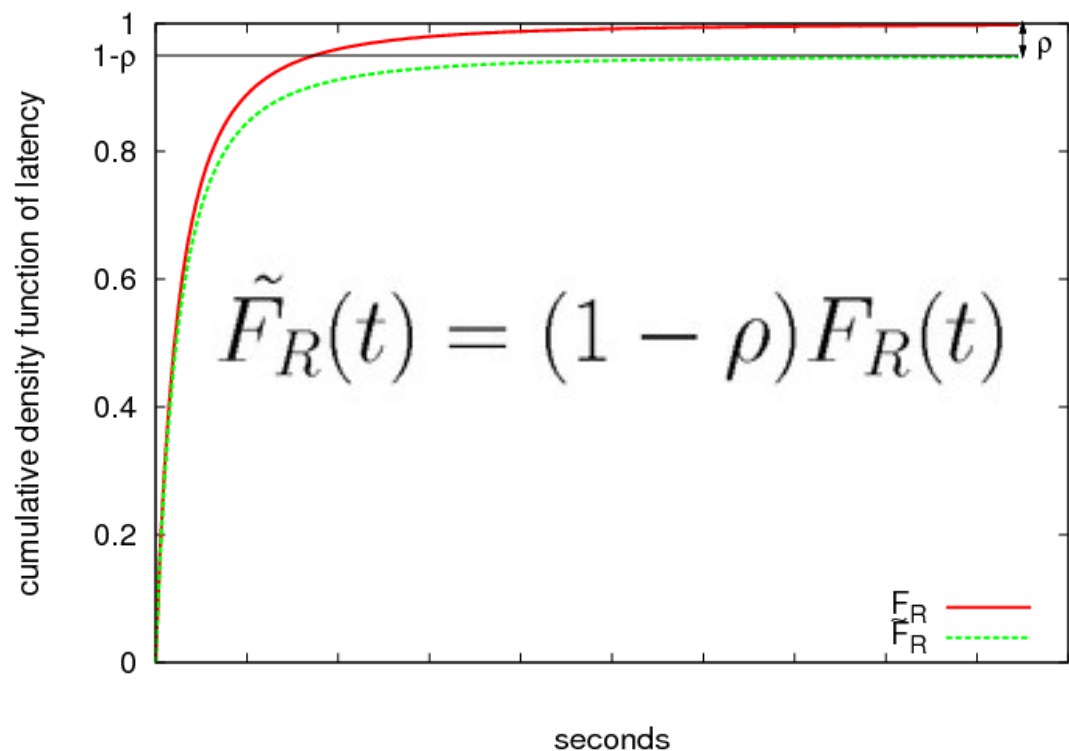


Refining the model: Workload Management System (WMS)



Probabilistic model

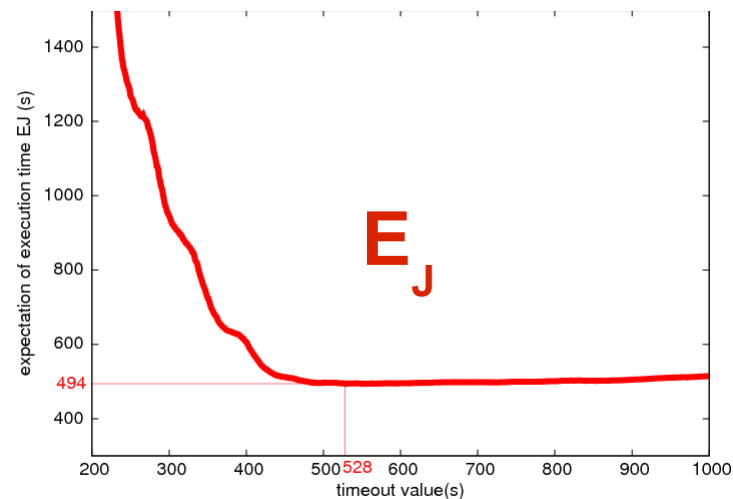
- Jobs are either “successful” or “outliers”
 - ρ as the ratio of outliers (measured)
 - latency R for successful jobs (distribution estimated from system traces)
- Latency distribution
 - Probabilistic model
 - Heavy-tailed p.d.f.



Timeout optimization

- Latency expectation minimization [CCGrid'07]:

$$E_J(t_\infty) = \frac{1}{\tilde{F}_R(t_\infty)} \int_0^{t_\infty} (1 - \tilde{F}_R(u)) du$$



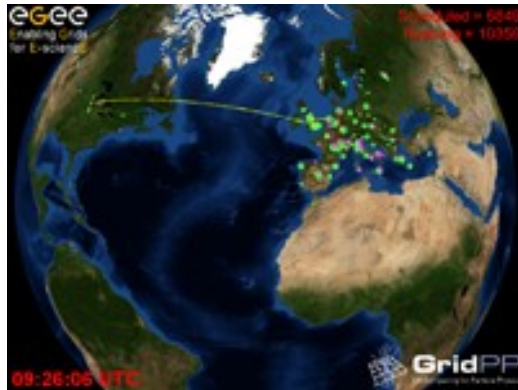
- Main facts

- Timeout $< \infty \Leftrightarrow \tilde{F}_R$ is heavy-tailed (proven)
- Better overestimate the timeout (observed)

Analyzed data

- Recorded by the Real Time Monitor

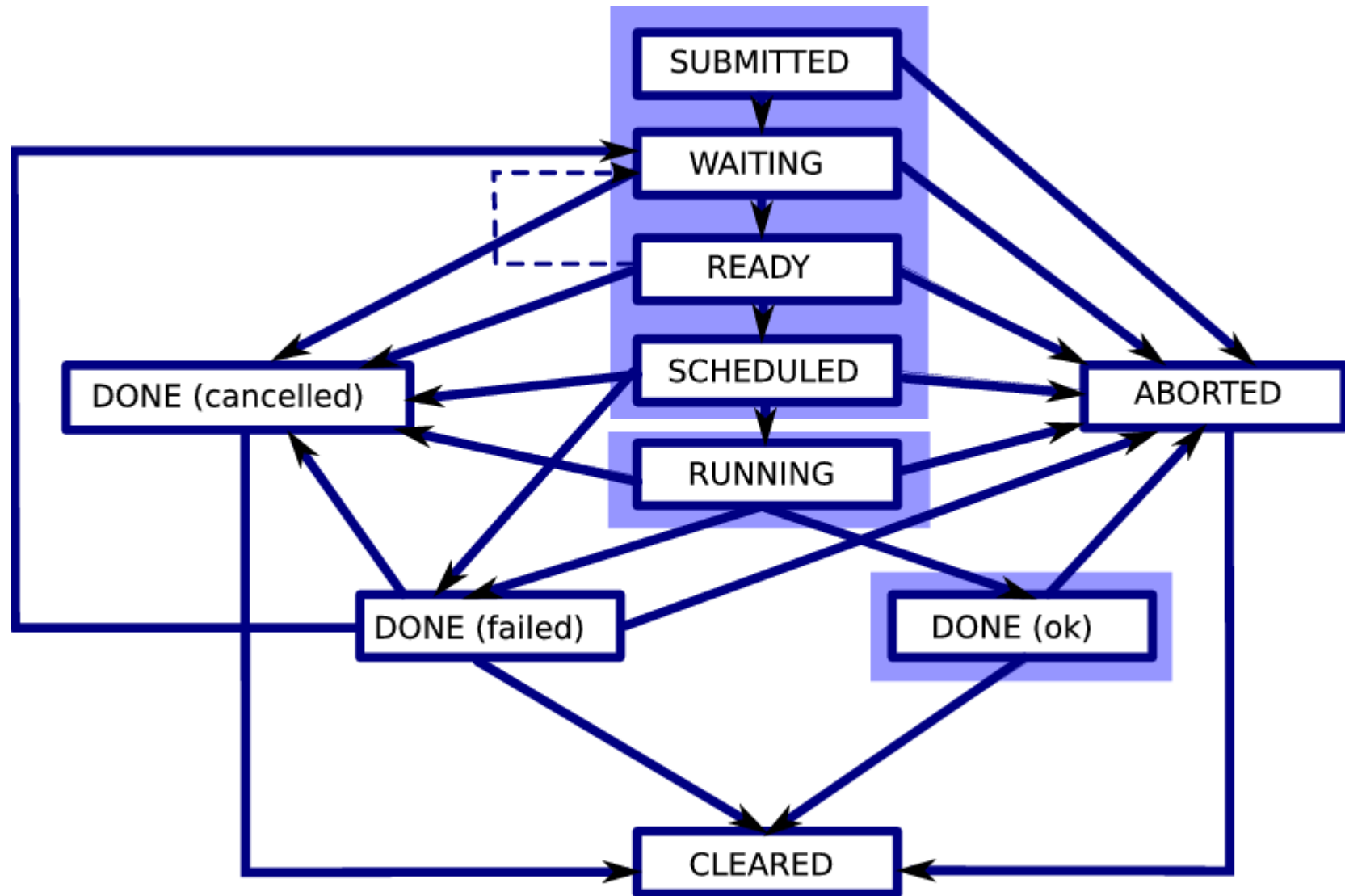
<http://gridportal.hep.ph.ic.ac.uk/rtm/>



- 33,419,946 jobs traces (Sept. 05 to June 07)

- Hosts for UI, RB, CE, VO
- epoch for registration on UI, accepted on the network server, matched to a target CE, accepted, transferred to a CE, started running, job completed
- *type* and *final_reason*
 - e.g. REGISTERED-DONE-RAN-CLEARED, “Job term. successfully”
 - e.g. REGISTERED-DONE and “Aborted by user”

Job's life cycle on EGEE



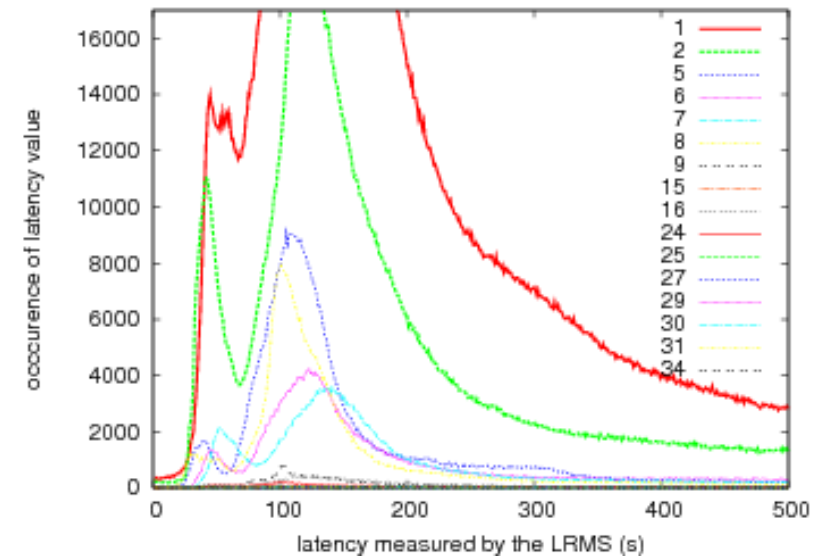
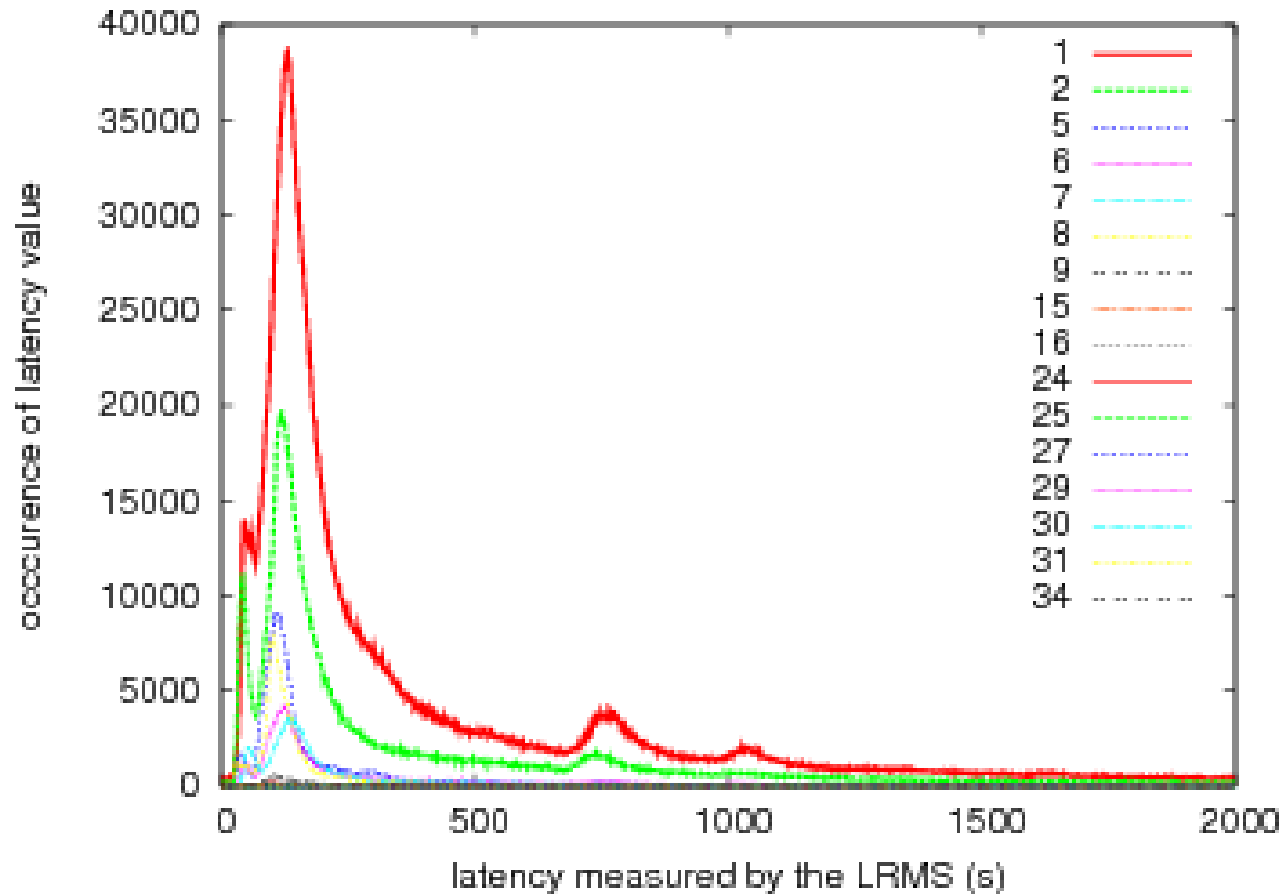
Classification w.r.t *type* and *final_reason*

- 315 different cases
 - some are very frequent
 - some are very rare
 - ⇒ limitation to the 37 most frequent cases (99.4% of the total number of jobs)
- classification into
 - successful jobs (64.8 %)
 - failed jobs (ratio $\varphi = 19.1$ %)
 - outliers (ratio $\rho = 16.1$ %)

case	type and final_reason	occurrences	%	class
1	RDRC Job terminated successfully	11,563,331	34.6%	R (9,999,928)
2	RDR Job terminated successfully	5,639,638	16.9%	R (5,035,776)
3	RA Job RetryCount (0) hit	3,838,380	11.5%	outlier
4	RA Cannot plan : BrokerHelper : no compa	3,422,319	10.2%	F
5	RDRC -	1,299,235	3.89%	R (1,138,324)
6	RDRC There were some warnings : some file	1,004,800	3.01%	R (884,932)
7	RDR There were some warnings : some file	911,500	2.73%	R (813,405)
8	RDR -	877,229	2.62%	R (750,473)
9	RD Aborted by user	863,094	2.58%	R (875,89)
10	RA Job RetryCount (3) hit	582,152	1.74%	outlier
11	RA -	557,055	1.67%	F
12	RA Job proxy is expired.	495,519	1.48%	F
13	RA cannot retrieve previous matches fo	322,839	0.97%	F
14	RAR Job proxy is expired.	267,890	0.80%	F
15	Una -	235,458	0.70%	R (10,632)
16	RDR Aborted by user	188,421	0.56%	R (15,3479)
17	RA Job RetryCount (1) hit	165,231	0.49%	outlier
18	UA Error during proxy renewal registra	149,095	0.45%	F
19	RA Unable to receive	115,867	0.35%	F
20	RE -	109,089	0.33%	F
21	RA Cannot plan : BrokerHelper : Problems	89,553	0.27%	F
22	UA Unable to receive	70,215	0.21%	F
23	RA Job RetryCount (2) hit	63,595	0.19%	outlier
24	RnaR -	56,044	0.17%	R (53,055)
25	RD -	45,400	0.14%	R (2,091)
26	RAC cannot retrieve previous matches fo	35,887	0.11%	F
27	UDR Job terminated successfully	31,722	0.09%	R (236)
28	RAR -	26,268	0.08%	F
29	RDRC There were some warnings : some outp	25,868	0.08%	R (20,341)
30	RDR There were some warnings : some outp	23,178	0.07%	R (18,315)
31	RRR -	22,983	0.07%	R (19561)
32	RA Submission to condor failed.	22341	0.07%	F
33	RA Job RetryCount (5) hit	22,260	0.07%	outlier
34	RT Job successfully submitted to Globu	18,972	0.06%	R (3,768)
35	RT unavailable	18,065	0.05%	F
36	RA Job RetryCount (7) hit	17,328	0.05%	outlier
37	RA hit job shallow retry count (0)	16,863	0.05%	outlier

The 37 cases

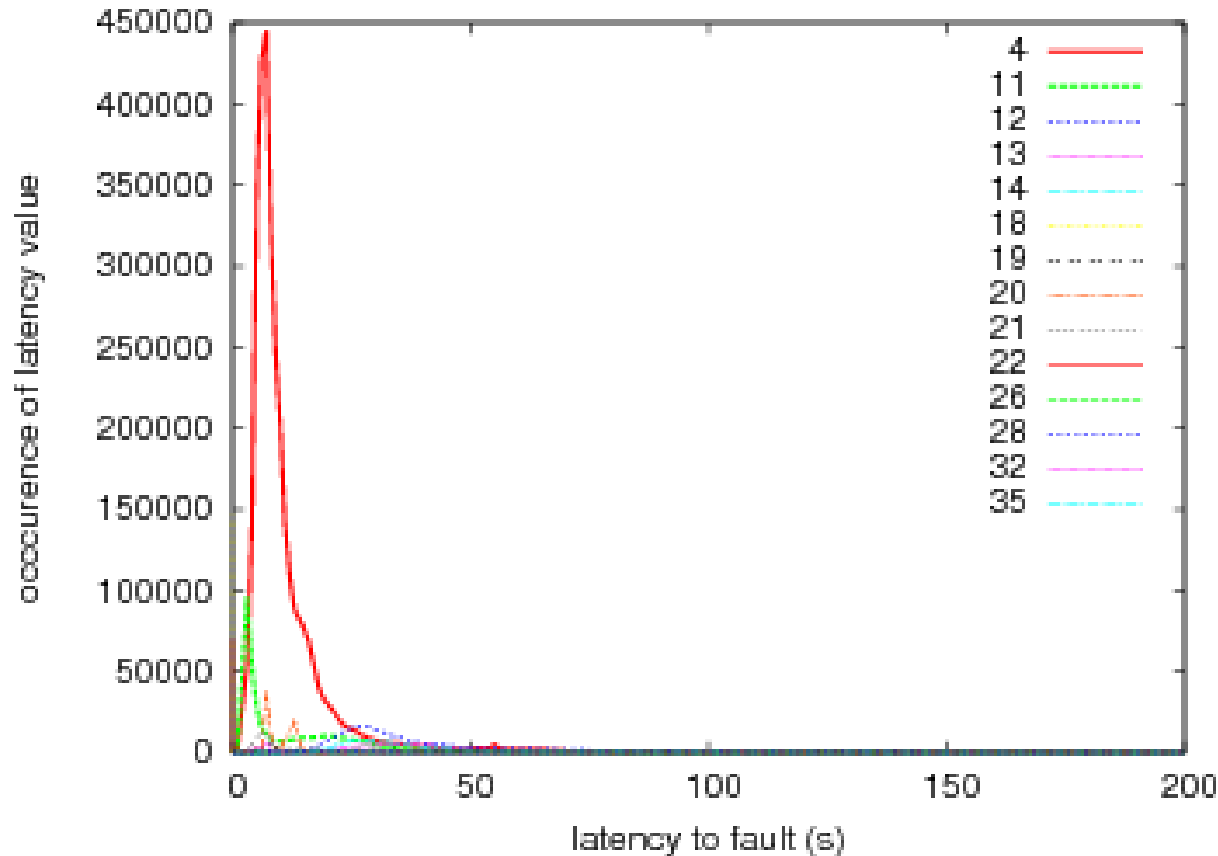
Successful jobs



- Heavy-tailed distribution (r.v. R)

RDR - Job term. succ.
 RDR - Job term. succ.
 RDRC -
 RDRC -
 RDR - warnings
 RDRC - warnings

Failed jobs

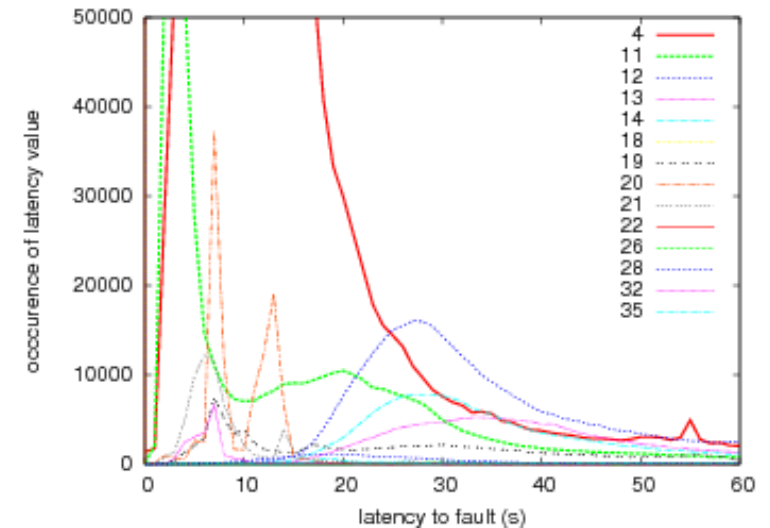


RA - No compatible resource

RA -

RA - Job proxy expired

RA - cannot retrieve previous matches

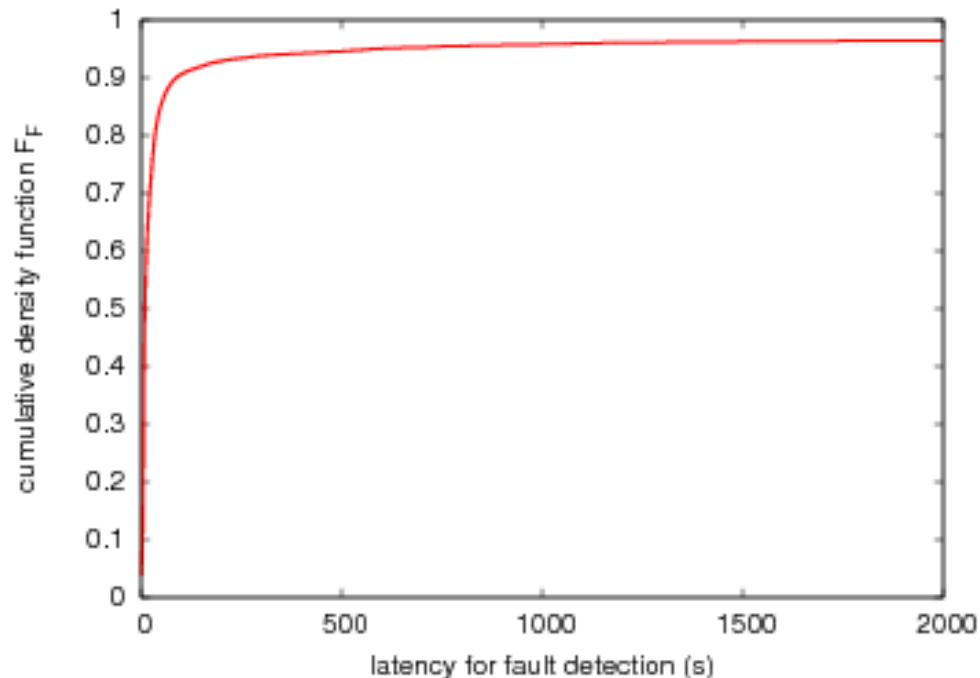


- Quite grid-specific

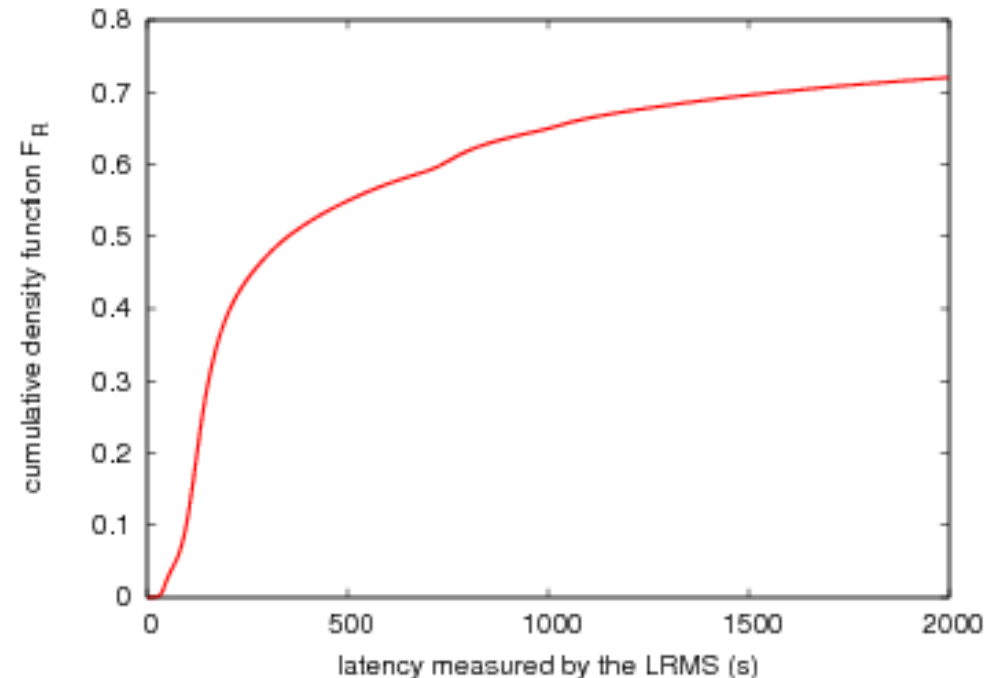
- Grid reliability is an order of magnitude below homogeneous / single-site clusters

Taking faults into account

- Cumulative density function of successful jobs: F_R



- Cumulative density function of failed jobs: F_F

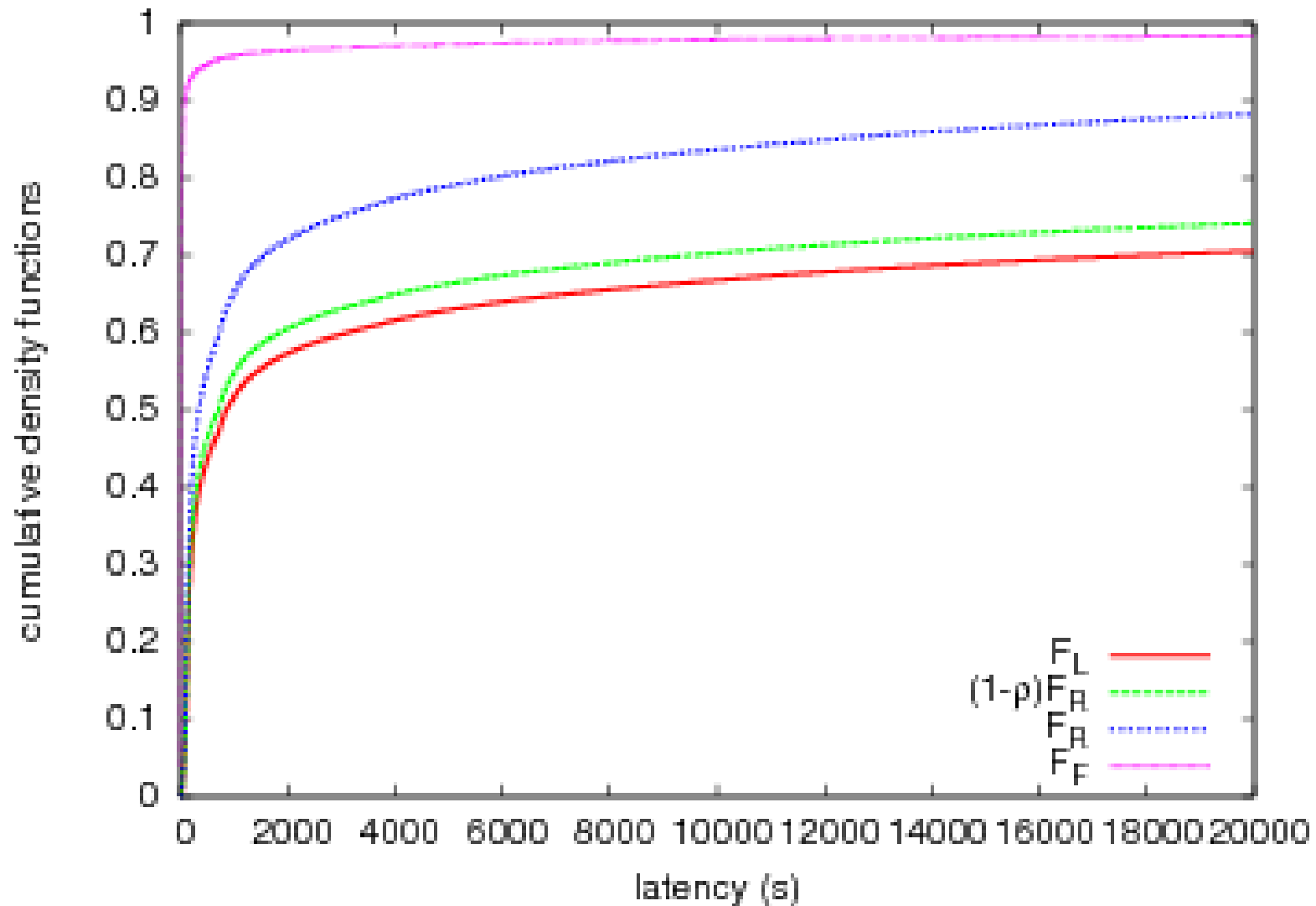


Dealing with failed jobs

- After a fault, the job is resubmitted leading to:
 - a failed job again (resubmitted again) (ratio ϕ)
 - a successful job (ratio $1-\rho-\phi$)
 - an outlier (ratio ρ)
- Total latency \mathbf{L} of cdf \mathbf{F}_L (computed recursively):

$$\begin{aligned}F_L(0) &= 0 \\F_L(1) &= \frac{1 - \rho - \phi}{1 - \phi f_F(0)} F_R(1) \\F_L(t > 1) &= \frac{1}{1 - \phi f_F(0)} \left[(1 - \rho - \phi) F_R(t) + \phi \sum_{u=1}^{t-1} f_F(t - u) F_L(u) \right]\end{aligned}$$

Experiment: computation of F_L



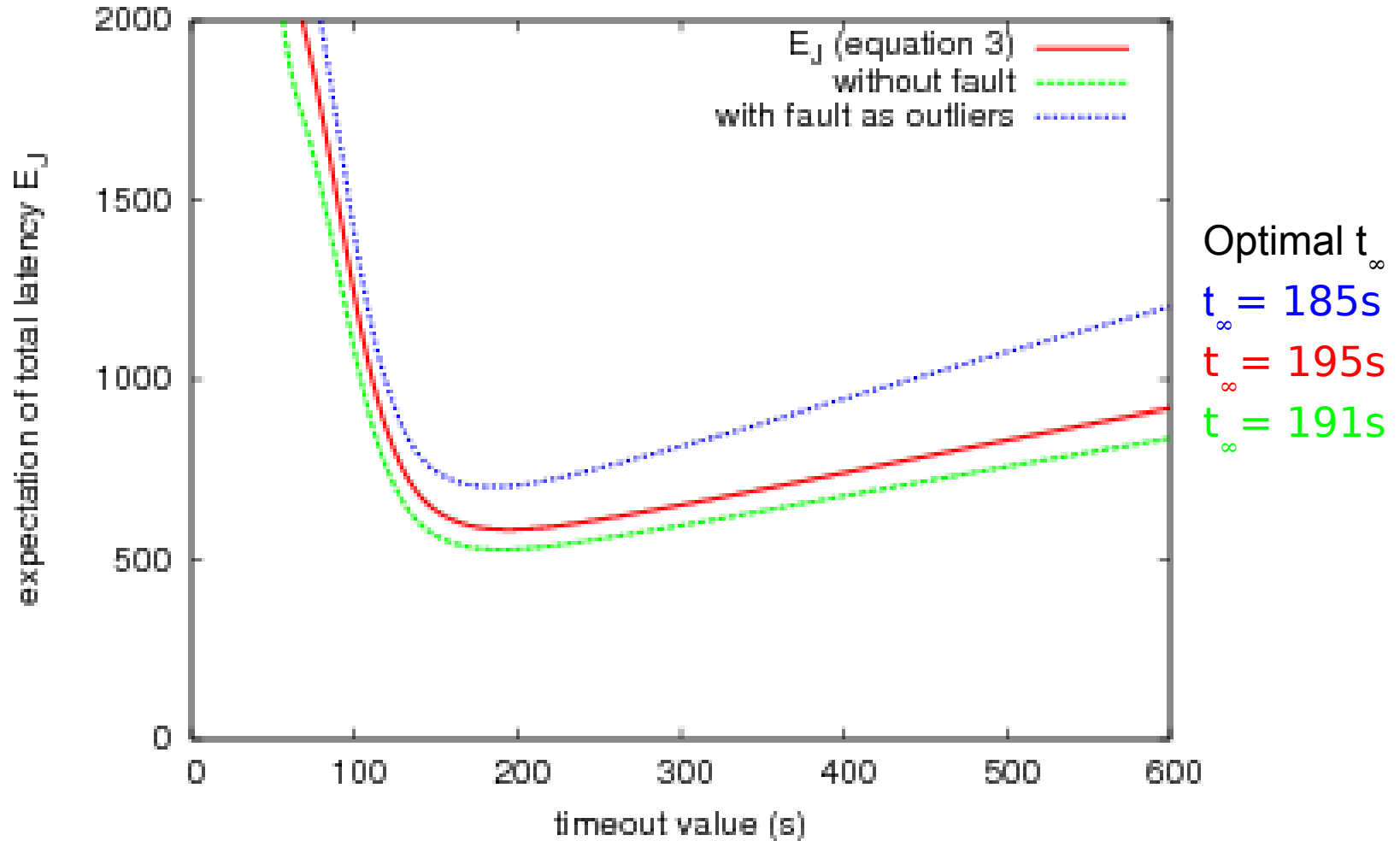
Timeout optimization: now including faults

- Revisiting resubmission with faults:

$$E_J(t_\infty) = \frac{1}{F_L(t_\infty)} \int_0^{t_\infty} (1 - F_L(u)) du$$

Timeout optimization including faults

E_J estimation
 $E_J = 704s$
 $E_J = 584s$
 $t_\infty = 529s$



Reducing the number of cases

- Influence on modeling and optimization

nb. of cases	with faults (F_L)		without faults (\tilde{F}_R)	
	opt. t_∞	min. E_J	opt. t_∞	min. E_J
37	195s	584s	191s	529s
30	194s	584s	191s	529s
20	195s	577s	191s	524s
10	192s	558s	189s	530s
4	199s	606s	197s	570s

- No influence on optimal timeout
- Significant influence on E_J

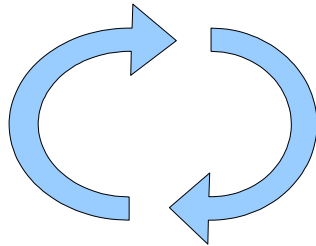
Conclusion

- Probabilistic model taking into account successful jobs, outliers and failures
- Extensive exploitation of real usage traces from a production grid
- Gives insights on the WMS behavior

Perspectives

- Grid simulation

- “coarse-grain” simulation based on the behavioral model



- Model validation

- Model accuracy & model extensions

- Live exploitation to optimize performances taking into account the grid workload in real-time

- Study various resubmission / multisubmission strategies