Crowdsourcing GUI Tests

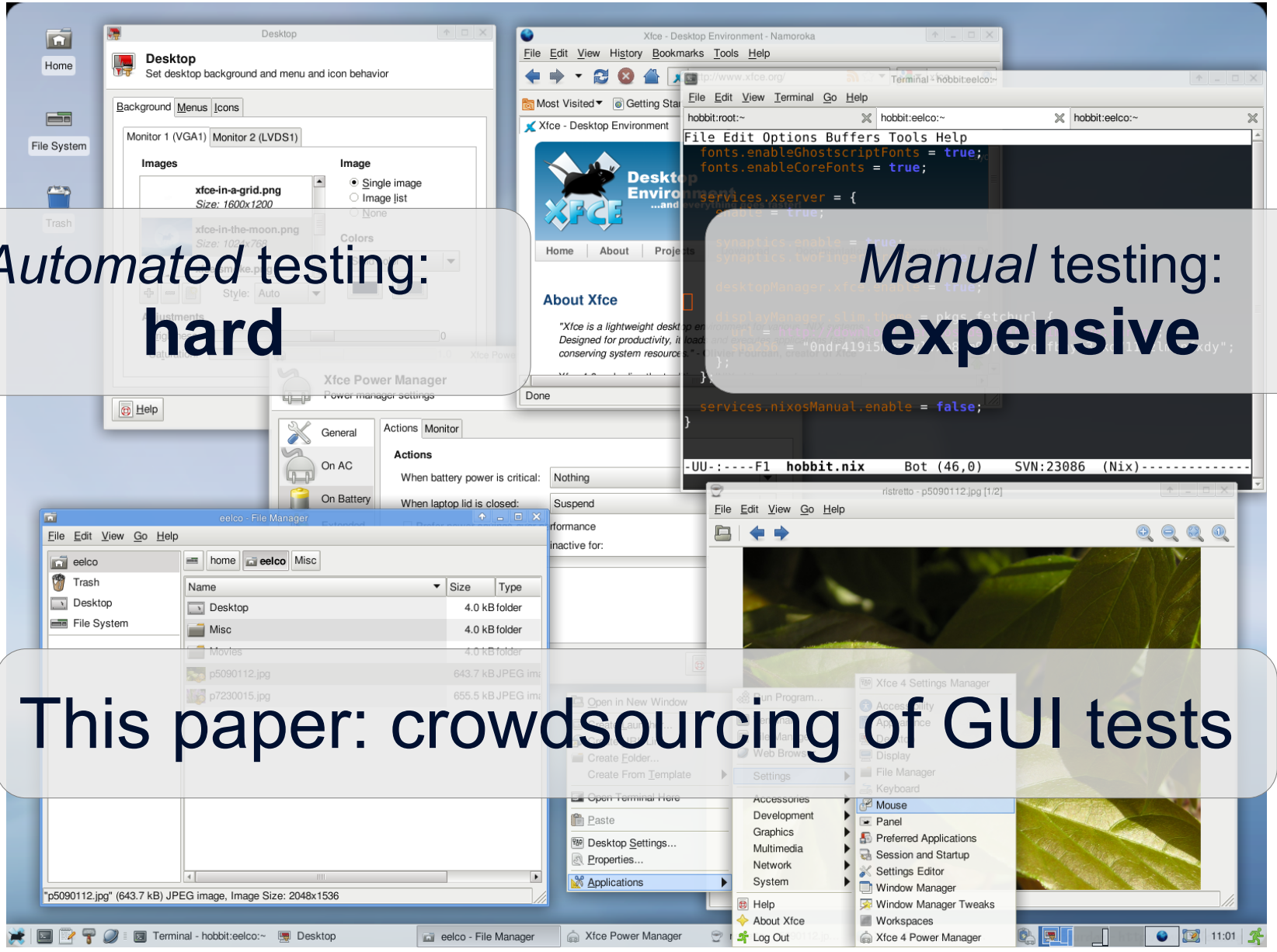
Eelco Dolstra
LogicBlox, Inc.

Raynor Vliegendhart
Delft University of Technology

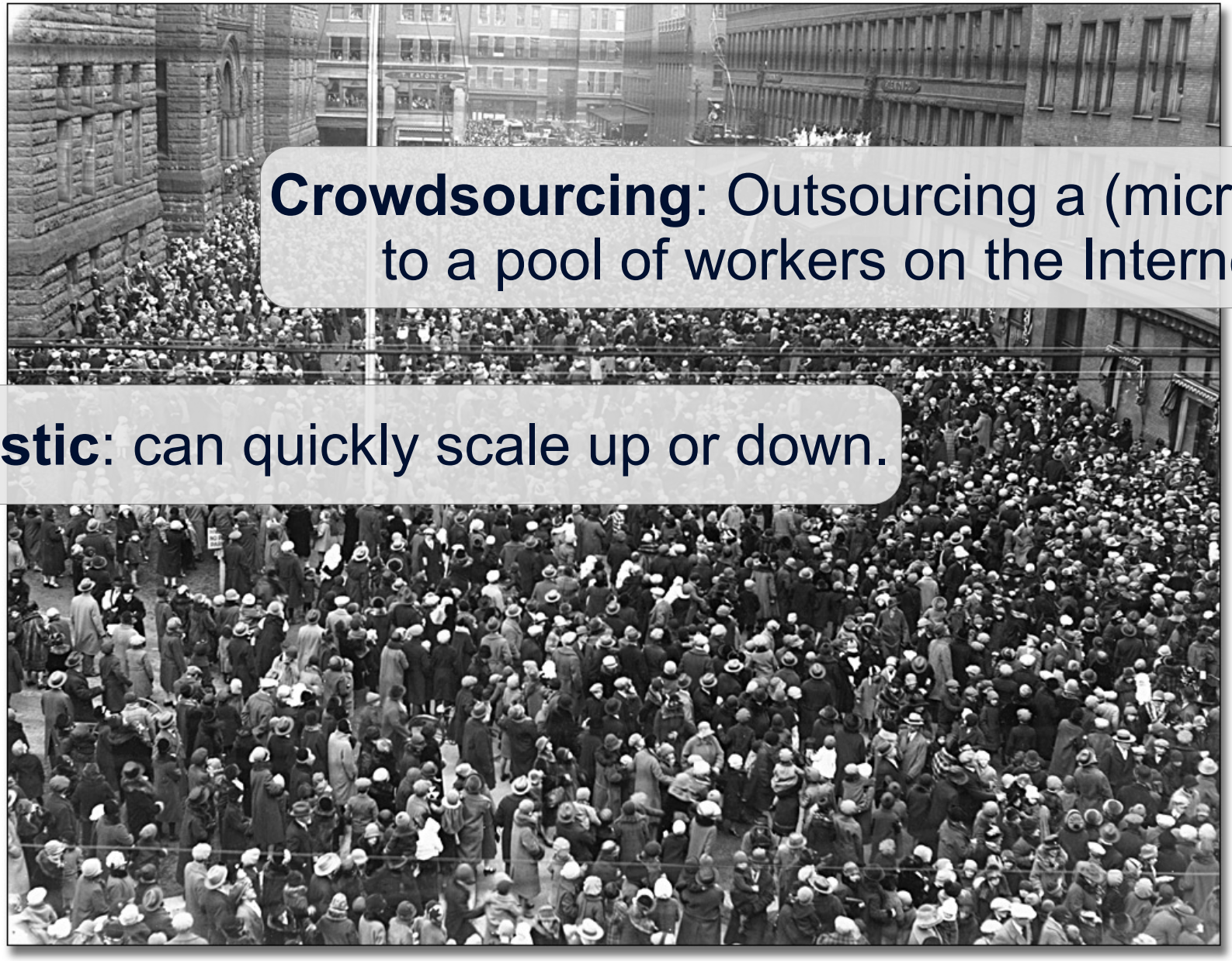
Johan Pouwelse
Delft University of Technology

ICST 2013, Luxembourg
21 March 2013

GUI testing


Automated testing:
hard
Manual testing:
expensive
This paper: crowdsourcing of GUI tests

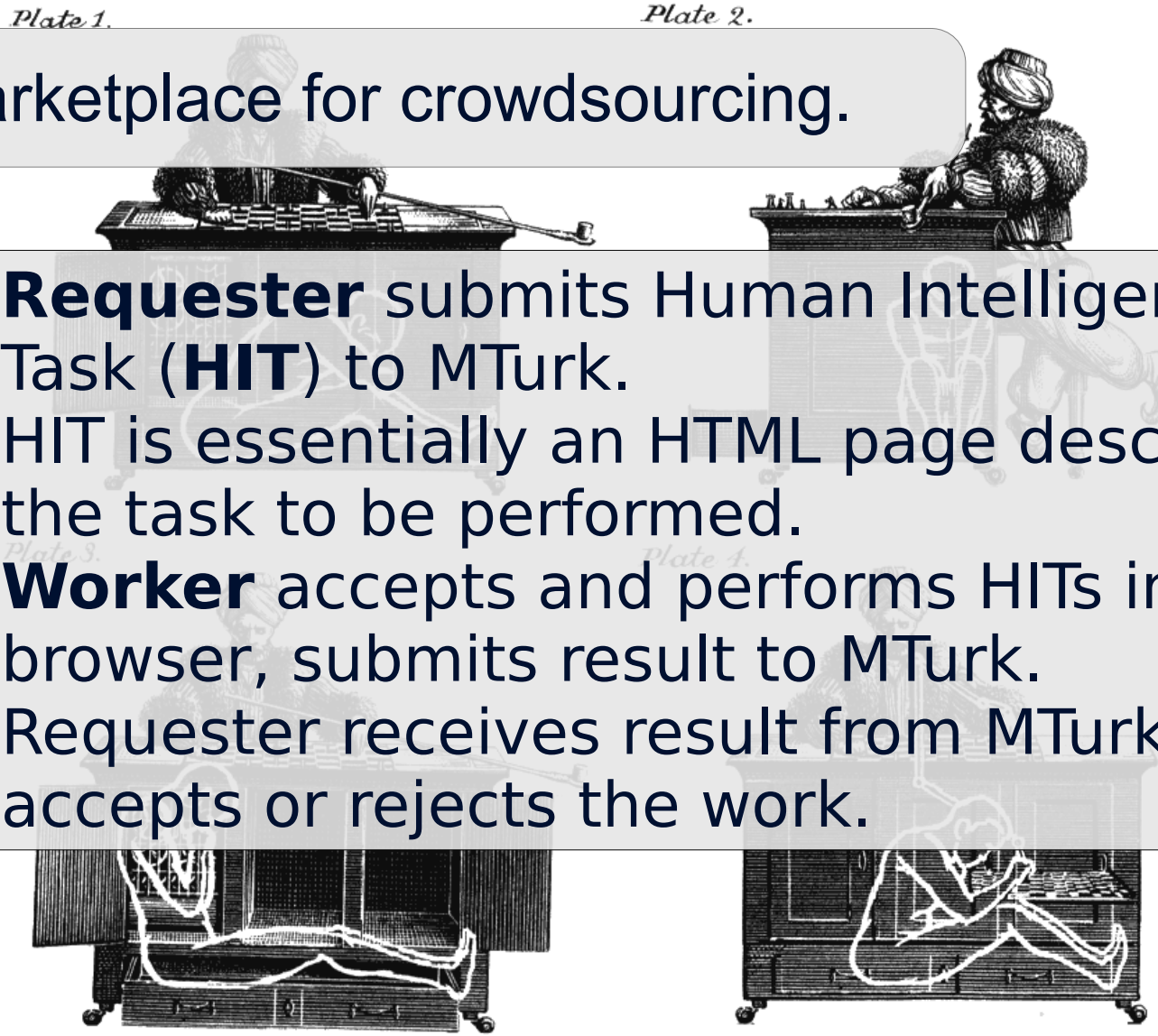
Crowdsourcing



Crowdsourcing: Outsourcing a (micro)task to a pool of workers on the Internet.

Elastic: can quickly scale up or down.

Amazon Mechanical Turk



Marketplace for crowdsourcing.

- **Requester** submits Human Intelligence Task (**HIT**) to MTurk.
- HIT is essentially an HTML page describing the task to be performed.
- **Worker** accepts and performs HITs in web browser, submits result to MTurk.
- Requester receives result from MTurk and accepts or rejects the work.

HIT example

amazonmechanical turk beta Artificial Artificial Intelligence

Eelco Dolstra | Account Settings | Sign Out | Help

Your Account | HITs | Qualifications

216,797 HITs available now

All HITs | HITs Available To You | HITs Assigned To You

Find HITs containing that pay at least $ 0.00 for which you are qualified require Master Qualification GO

Timer: 00:00:00 of 15 minutes

Want to work on this HIT? Accept HIT

Want to see other HITs? Skip HIT

Total Earned: $0.36
Total HITs Submitted: 7

Version 2: Identify the exact number of people depicted in an image

Requester: Corbis Holdings, Inc **Reward:** $0.02 per HIT **HITs Available:** 7702 **Duration:** 15 minutes

Qualifications Required: HIT approval rate (%) is greater than 80

Identify the number of people or human-like figures displayed in the image below.

Instructions:

- Evaluate each image and identify if it depicts exactly one person **(1)**, exactly two people **(2)**, exactly three people **(3)**, exactly four people **(4)**, exactly five people (to 10 people **(6-10)** or 11 or more people **(11+)**
- Along with actual human beings, also count any human-like figures such as drawings, stick figures, statues, mannequins, robots or figurines.
- Body parts such as hands or feet should be counted as a person.
- Select **"No People"** if the image depicts no people or human-like figures.

- 1 person
- 2 people
- 3 people
- 4 people
- 5 people
- 6-10 people
- 11+ people
- No people

- 1 person
- 2 people
- 3 people
- 4 people
- 5 people
- 6-10 people
- 11+ people
- No people

Crowdsourcing GUI tests

- We want to crowdsource the task of (regression) testing GUIs.
- E.g. on every commit, create a HIT that asks workers to test the GUI.
- Main requirement: worker should not need to install anything; everything should run in the browser.

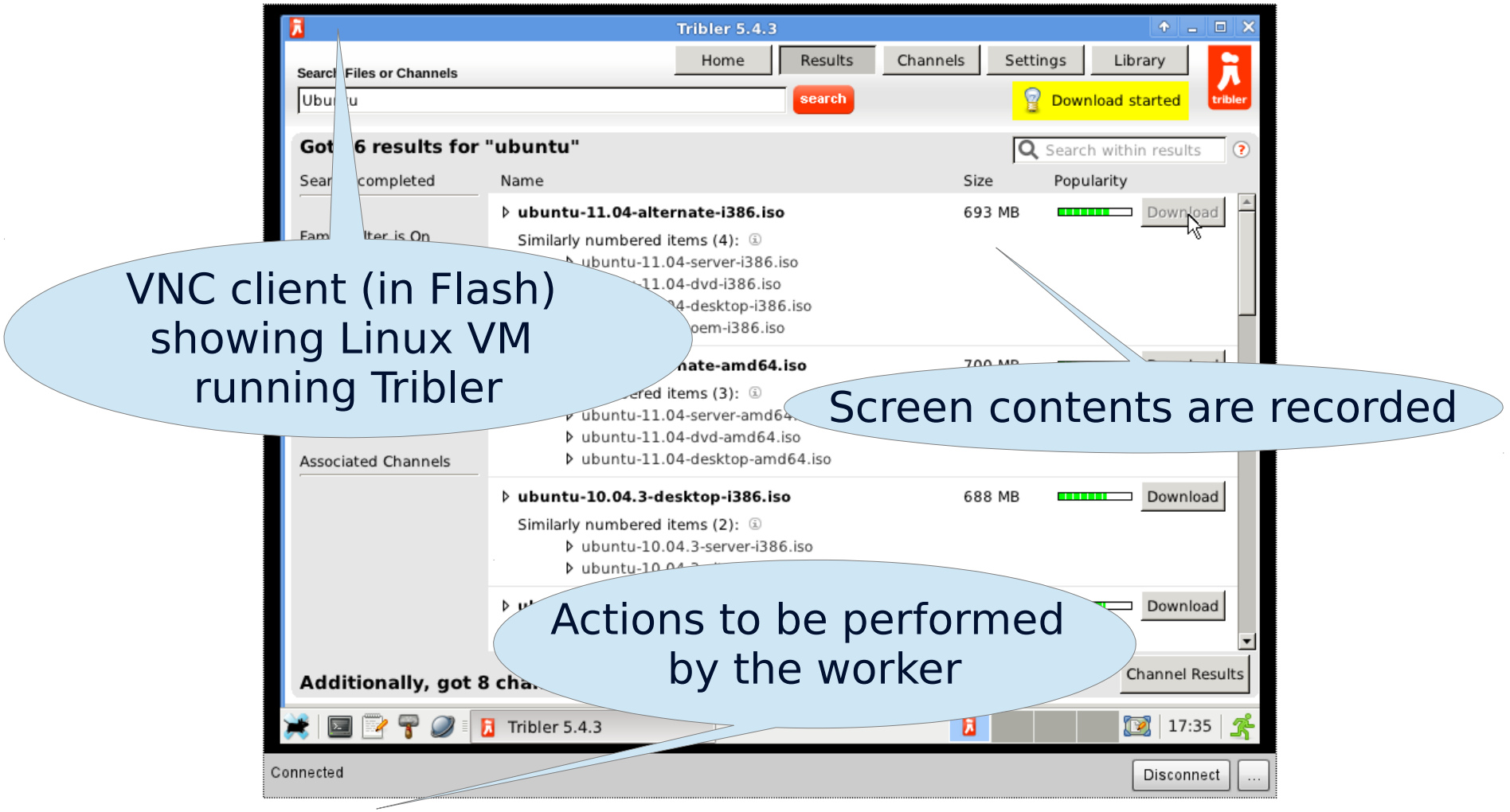
Example: Tribler

- Tribler is a fully decentralized peer-to-peer filesharing program.
- Regression testing task:
  - Search for files
  - Download a file
  - Use channels
  - Check family filter effectiveness

Test a Graphical User Interface

The goal of this task is to perform a list of actions to test software. Below you see the display of a computer running some software. The task is **to perform the following steps *precisely* and report whether they succeed**. If you don't succeed in any step, **report what went wrong** in the form at the bottom.

Virtual machine display



Step 3 / 8: Click on the **Download** button next to the top result. This should start the download.

Did you succeed? ○Yes ○No

HIT #6

Info

Status:	**Reviewable**
Result:	✗ (4 passed, 6 failed, 0 unreliable)
Description:	Basic Tribler download test (release-5.3.x, r21071)
Created at:	2011-08-29 09:49:36
Runtime:	0 h 24 m 39 s
Task ID:	`tribler-test` (XML)
HIT ID:	20Y5TYMNCWYB49GM5YJH111JAE1K0A
Reward:	$0.15
Resolution:	1024 x 768
Assignments submitted / requested:	10 / 10
Assignments returned or abandoned:	1
Average duration:	386.0 s ($1.40 hourly wage)

Assignments

[Show answers]

Res	Acc	Sta	Assignment ID	Where	Submitted at	Duration
✗	✓	✓	20AUUU000QKKA4JV5EJ2Y268PUTVVA	GB	2011-08-29 09:59:10	176 s
✗	✓	✗	21ALKH1Q6SVQ8H9LZWL6P50BKSG1K6	IN	2011-08-29 09:59:21	123 s
✗	?	✓	29YR70G5R98D8J48RDNSWTNY1W0RB7		2011-08-29 10:02:04	54 s
✓	✓	✓	2I6R9Y8TNKIMZSNIXL8S60M0EWRV84	IN	2011-08-29 10:02:05	374 s
✗	✓	✓	2Z0PJWKUDD8XMB43GFQ0F6UTDMQAL5	IN	2011-08-29 10:03:10	388 s
✗	✓	✓	2VW2J1LY58B727NDVM6KLD16NVPR6M	IN	2011-08-29 10:05:08	270 s
✓	✓	✓	28U4E6AUBXHWSH8SYF0VFZWDG1UX1Q	IN	2011-08-29 10:06:37	251 s
✓	✓	✓	2R8YQ4J3NAB6BFNSSUPYMT1T8VAQDC	IN	2011-08-29 10:11:42	907 s
✗	✓	✓	2K49KSSZVXX2Q4677QH02GNILHKM48	IN	2011-08-29 10:12:29	699 s
✓	✓	✓	280TNKIMZSM5KFI5FUD0TGVFSYKCZT	IN	2011-08-29 10:14:15	618 s

Assignment 2Z0PJWKUDD8XMB43GFQ0F6UT

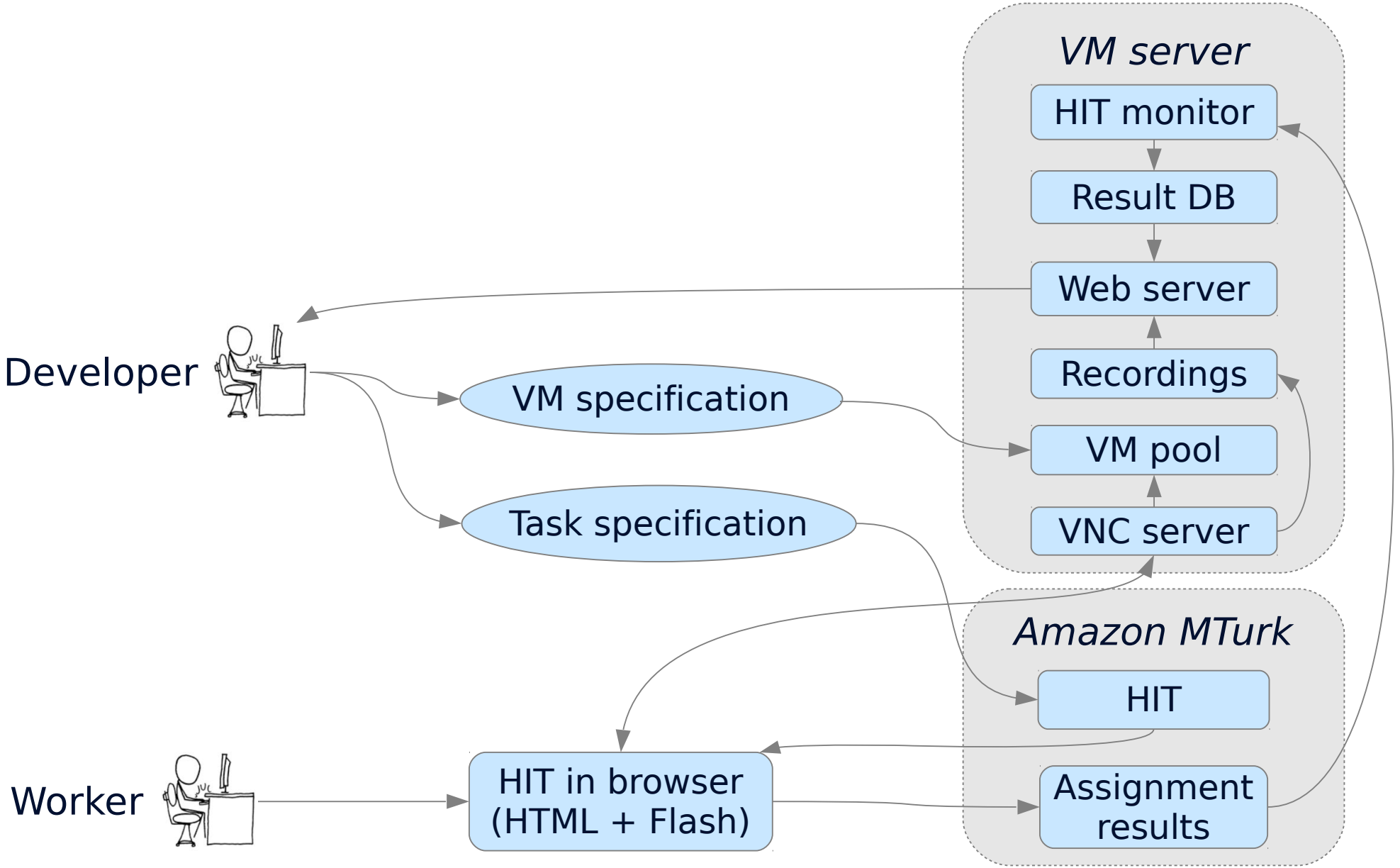
Info

Test result:	✘ **Failed** *(2 out of 7 steps failed)*
Acceptance check:	✔ **Passed** *(log)*
Status:	✔ **Approved**
Worker ID:	*(redacted)*
Location:	Vaniyambadi, Tamil Nadu, India
Accepted at:	2011-08-29 09:56:42
Submitted at:	2011-08-29 10:03:10
Duration:	388 s
Recordings:	Video #1 (379.0 s, 14.73 MiB)

Answers

Question ID	Answer
step1	yes
step2	yes
step3	yes
step4	no
step5	no
step6	yes
step7	yes
offensive_words	ubuntu server black swan alternate
comments	The download process didn't start it still remains wait state

Implementation



Task description example

```
<task reward="0.15" assignments="10">
  <steps>
    <step onFailGoTo="end">
      <question>Do you see a window named
        "Tribler"?</question>
    </step>

    <step onFailGoTo="channels">
      <action>
        In the search box, type <strong>Ubuntu</strong>
        and press enter.  Wait a few seconds.
      </action>
      <question>Do results appear?</question>
    </step>
    ...
  </steps>
</task>
```

VM specifications

- Virtual machines are instantiated automatically from a **declarative specification** of the desired configuration of the entire machine.
- Based on NixOS, a Linux distribution with a declarative configuration model.
- Previously used for automated system tests (ISSRE'10).

NixOS

VM specification example

```
machine =
  { config, pkgs, ... }:

  let tribler = ...; in

  { require = [ ./common-xfce.nix ];

    environment.systemPackages = [ tribler ];
  };

prepare =
  ''
    $machine->execute("su - alice tribler &");
    $machine->waitForWindow(qr/Tribler/);
  '';
```

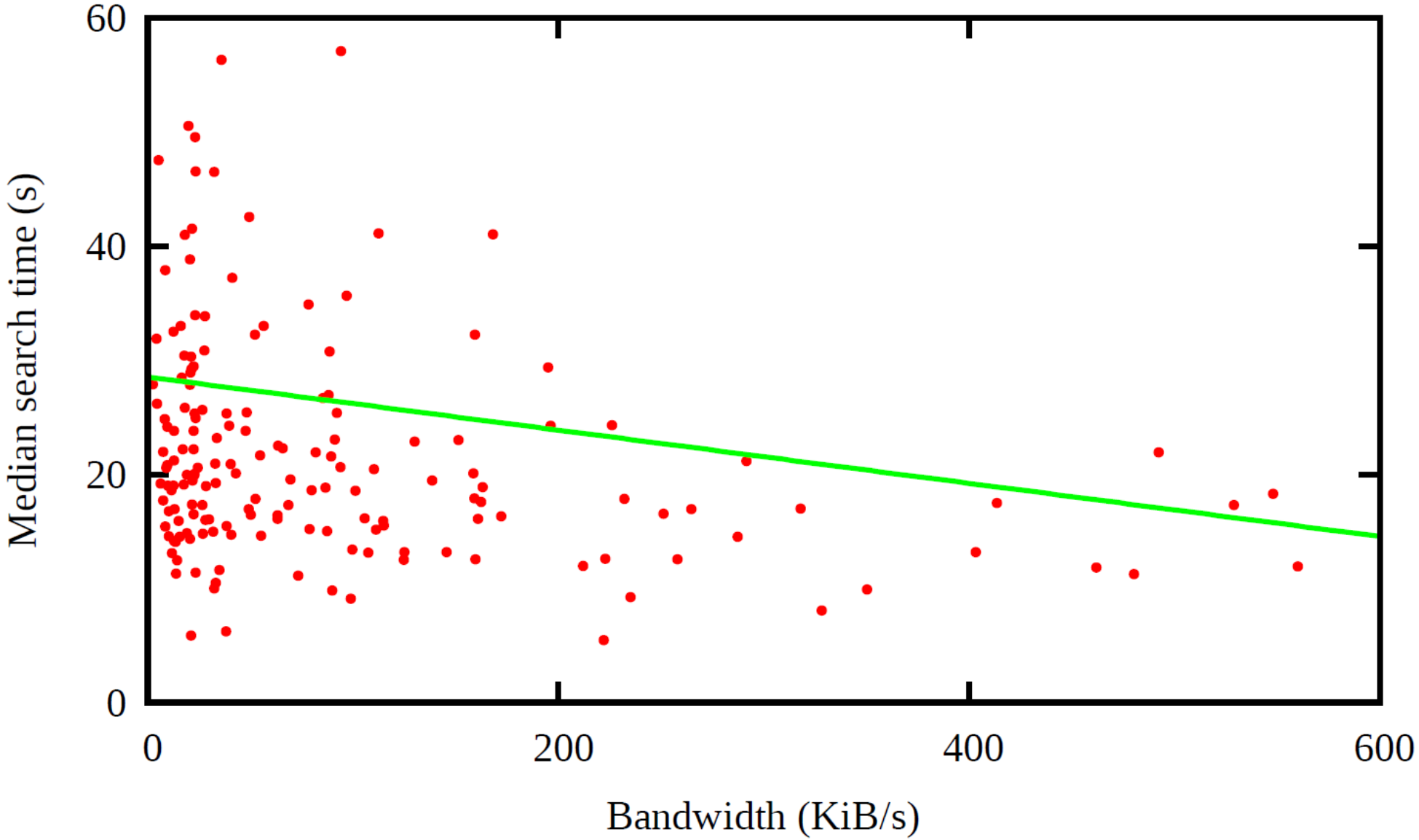
Experimental results

- Regression test cases
  - Tribler
  - KDE USB stick mounting
  - KDE logout/login
  - Xfce editor test

Experimental results

- **RQ1**: Are workers *technically* able to perform the tasks?
  - E.g. if they all have horrible latency, it's not going to work.
  - **Yes**.

Country	Workers	Assign-ments	Median speed (KiB/s)	Mean ping (ms)
India	247	490	33.7	329
United States	42	49	200.3	202
United Kingdom	11	28	535.1	52
Pakistan	8	9	24.6	299
Romania	7	14	468.0	25
(27 countries omitted)				
Total	398	700	48.0	260


60
40
20
0
Median search time (s)
0
200
400
600
Bandwidth (KiB/s)

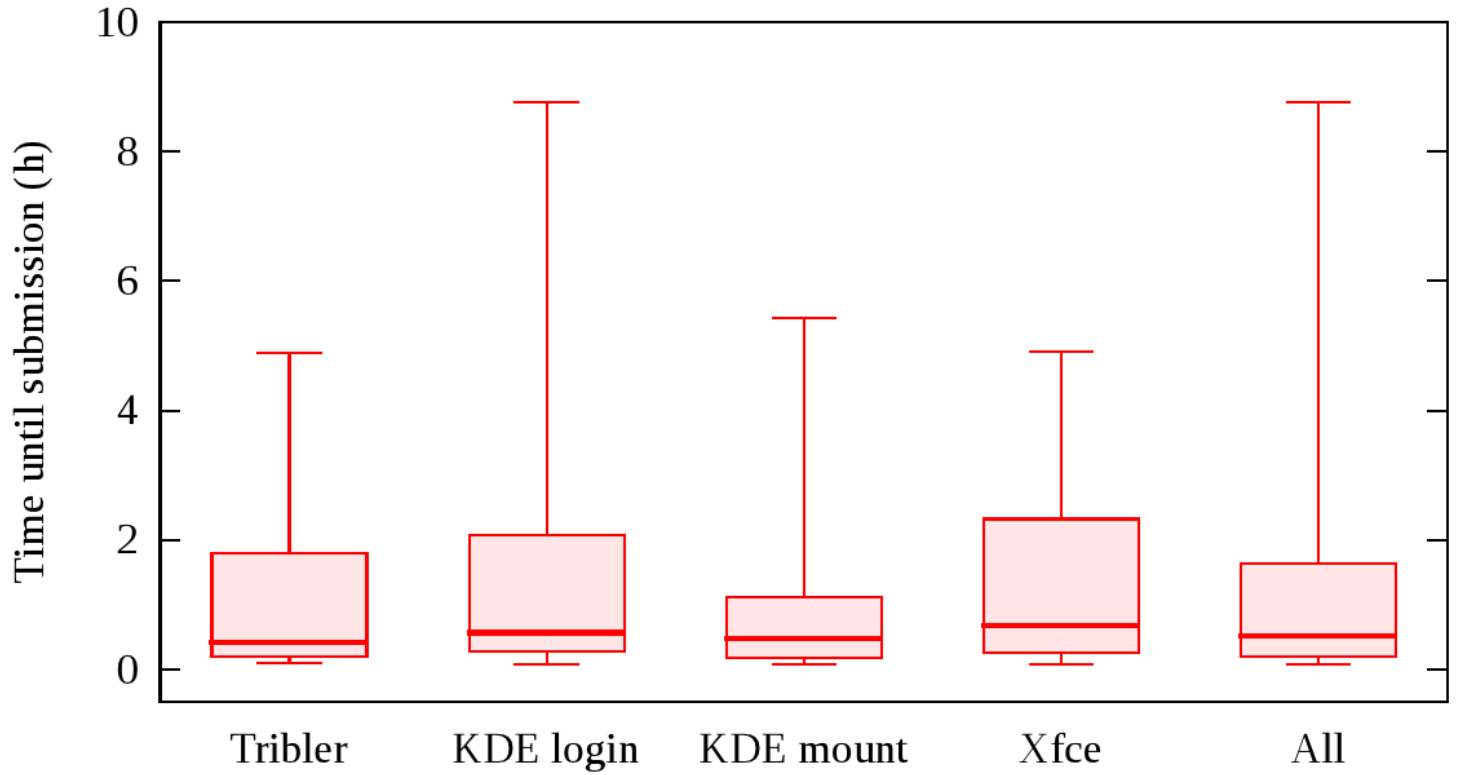
Experimental results

- **RQ2**: Is crowdsourcing a feasible approach for continuous testing?
  - This requires workers to be sufficient correct.
  - **Qualified yes**.

	Tribler	KDE login	KDE mount	Xfce
Reward	$0.15	$0.10	$0.10	$0.10
# Hits	14	10	11	10
Average runtime	2.0 h	3.6 h	2.0 h	2.1 h
# Submitted	145	100	115	100
# Abandoned	9	9	11	7
# Workers	112	86	94	85
Median duration	314.0 s	327.5 s	240.0 s	246.5 s
Hourly rate	$1.72	$1.10	$1.50	$1.46
% Correct	66.9%	77.0%	68.7%	82.0%
% Tech. issues	5.5%	6.0%	5.2%	3.0%
% Misunderstood	2.1%	6.0%	13.9%	2.0%
% Fraud	3.4%	4.0%	2.6%	7.0%

Experimental results

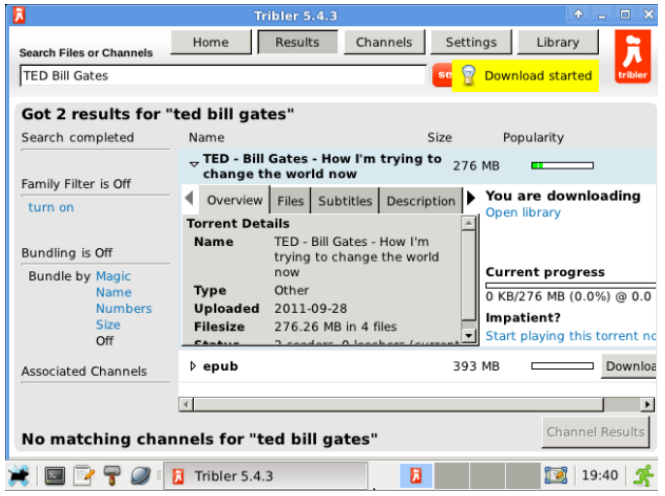
- **RQ3**: How long do crowdsourced GUI tests take?
  - I.e. what's the average runtime of a HIT?



Usability experiments

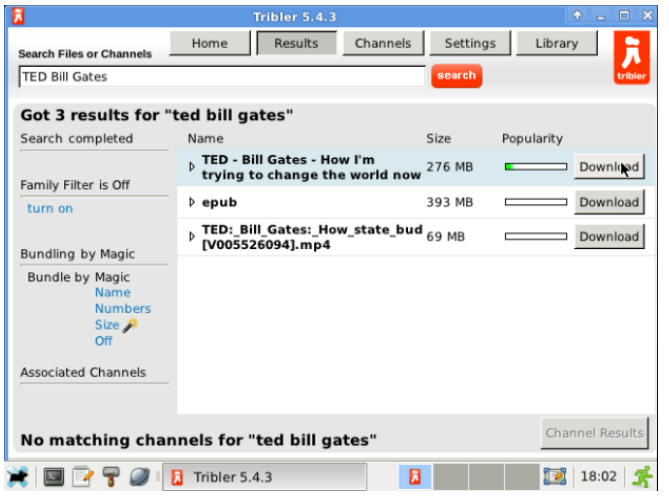
- **RQ4**: Is crowdsourcing feasible for usability experiments?
- Tribler A/B test: Does the experimental new search interface work better than the old one?

"no bundling"



vs.

"bundling"



- **Yes**. (Experiment cost: 100 x $0.25 = $25)

Conclusions

- We have developed a system for crowdsourcing of GUI tests.
- Experiments demonstrate technical feasibility.
- More work needed on task design and worker qualification.
- **Questions?**