Time is Money:
A Report from the POPL 2007 Chairman

Matthias Felleisen
College of Computer Science, Northeastern University, Boston, MA 02115

November 7, 2006

1 POPL and the Nature of Conferences

POPL, like many of ACM's flagship conferences, is a highly selective forum for innovative papers on programming languages. In any typical year, the program committee (PC) receives between 150 to 200 submissions. A large fraction of those papers are interesting in many regards and would appeal to the typical POPL audience. In the end, however, the PC picks around 15% for presentation at the conference and inclusion in the proceedings, and that's that.

Unfortunately, most traditional paper selection mechanisms for conferences do not reflect that conference minutes are scarce and that therefore, the selection process is as much about taste—the personal taste of committee members and the collective taste of the committee—as it is about soundness and innovation. When the POPL/SIGPLAN steering committee asked me to serve as program chairman for POPL 2007, I agreed to take on the task on the condition that I would be allowed to implement two innovations that address this scarcity:

1. a selection system that allocates a fixed number of votes to every PC member; and
2. a category of short papers.

With this short paper, I am reporting on the two innovations and how they worked out for the POPL 2007 paper selection process.

2 Measuring Enthusiasm

Scarcity of conference minutes means that when a PC picks one paper for inclusion, it simultaneously rejects a number of papers that are equally qualified in a technical sense. I believe that a good paper selection system for a conference should constantly remind the members of the PC of this fact, starting from the very moment when the PC chairman assigns them a bundle of submissions for review.

To emphasize this scarcity, I introduced a voting system that allocated every PC member a number of votes based on the number of assigned reviews. [1] Specifically, after the bidding stage, every PC member got six votes per assigned paper. This averaged out to 180 votes per PC member. I also let them know that this number of votes empowered them to get two full papers or four short papers into a program, regardless

[1] I called these votes "minutes" to remind the PC members of the role that they played in the selection process. For several members, this naming convention wasn't helpful. My successors may wish to test a different name.

of how other PC members voted. Thus, in the extreme, the PC as a whole could choose 80 short papers or 40 long papers, though I didn't anticipate such extreme allocations, and the PC didn't let me down.

I then asked the PC members to distribute these votes according to their enthusiasm for a paper. The suggestion was to use a balanced measure of soundness, relevance, POPL-appropriateness, and their taste for problems. Naturally, as they allocated votes to papers, their pool of votes shrank, and thus they had to realize that being on the PC is about making choices.

A few PC members expressed frustration with the system. It was new, they had no experience, and it definitely demanded making choices. Others happily distributed their votes in near-uniform distributions over the papers that they liked. Many however allocated reasonably large blocks of votes to individual papers, making their intentions clear to their co-evaluators. In general, the PC members "smeared" their votes in a more uniform manner than I had expected but favorites emerged rather quickly and the pool of eligible papers shrank rapidly.

One week before the deadline I assessed the status of the vote distribution and set two thresholds. I informed the PC that a paper needed at least 20 votes (averaged over the number of reviewers) for discussion in the group of long papers and at least six votes (averaged over the number of reviewers) for inclusion in the group of short papers. The latter consisted of short submissions and long submissions whose authors had given us permission to accept their submission as a short paper. I chose those thresholds in response to the actual distribution of votes and in such a manner that accepting all of the eligible papers in this pool would be a long but still acceptable POPL program.

Many PC members responded with a flurry of activity, exchanging messages and arguing about papers via email. They cc'ed me often on their emails and, depending on the situation, I encouraged them to exchange their opinions via modifications of their evaluations.[2]

Approximately one day before the meeting I shut down the server (as announced) and ranked the papers according to the votes they had received. In total, 58 papers had made the discussion list out of 200: 28 on the list of potential long papers and 30 on the list of short papers.

3 Making Decisions

The plan of my scheme was to discuss the papers in descending order of enthusiasm (number of votes):

1. all long submissions that had gathered enough votes to be acceptable as long papers;
2. all short submissions that had gathered enough votes to be acceptable as short papers;
3. all qualified PC submissions;
4. all long submissions that had gathered enough votes to be acceptable as short papers.

For each paper, I requested that the PC members who had read the paper spoke first. Others could ask questions after the informed members had finished their presentation. These questions could challenge the technical content; the evaluation of the technical contribution; the expertise of the evaluator and even their "POPL taste," though all within reason.

My goal of the discussion was to probe whether the responsible PC members would change their minds. If it became clear that this wasn't happening, the paper would be accepted, and we would move on. If the discussion changed a PC member's mind about a paper and the PC member removed points, the

[2] The submission server automatically sent email to co-evaluators of a paper if one evaluator changed the score or the write-up.

paper would drop in ranking and would be re-discussed later. If too many points were removed, the paper would drop out. I also made it clear that we would repeat this process until we had used up the official "POPL budget" of 900 conference minutes or nobody wanted to discuss any more papers.

"White Balling" If a single PC member wanted to spend a third of his votes (or more) on a paper and the discussion couldn't change his mind, the paper was accepted. Dick Gabriel has dubbed this mechanism "white balling" as opposed to the common practice of black-balling a submission, i.e., bringing forth potentially negative aspects of the work. ▮

3.1 Results

We picked 23 long papers and 13 short papers, which means that in principle we used 845 minutes of the 900 allowable minutes.

After we had decided to accept the last paper on the list, I asked people to take a five-minute break and to reflect on the meeting. If they were unhappy they should speak up and demand the discussion of an accepted or rejected paper or be quiet forever. One person brought up the last paper again and we discussed it for another few minutes. This brief discussion represented the entire extent of the PC's expressed unhappiness with the paper selection.

The PC meeting lasted 10.5 hours, 8 hours on the first day and 2.5 hours on the second one, meaning the PC spent considerable less time in the physical meeting than comparable programming language conferences with far fewer submissions.

4 Evaluation

When I designed the voting system, I wanted to eliminate several annoying aspects of the decision making process and improve some others. Naturally, any decision making process will suffer from human elements, and no matter what we do, we cannot change human nature. A few safe guards, however, can avoid the worst flaws of the processes. The first subsection presents my diagnosis of the existing decision systems for program committees. The second and third subsections are evaluations of what went well and what went wrong, together with ideas on how to improve the system.

The two evaluation represents my personal evaluation, plus a few comments from PC members. After the PC meeting was over, I solicited written opinions from the PC. Six of them sent me emails: one with complaints, two with critical remarks, and three with praise. I have incorporated their ideas and suggestions as appropriate and thank them anonymously.

4.1 Analysis of Existing Decision Systems

First, in my personal experience with past PCs, it is common that members who haven't read the submission under discussion make the decision or at least heavily influence it. This influence can take many forms: technical questions; positive or negative opinions about a sub-area or worse the author; non-technical facts about authors; abstract discussions concerning the balance of the program; and so on.

On one hand, this kind of decision making exploits the collective wisdom and taste of a committee. After all, the committee as a whole represents the current and intended target community of a conference and, if the PC members dislike a technically correct paper, then perhaps the conference isn't a good home for the paper. On the other hand, it is deeply unfair that a person without knowledge about a submission has more influence than someone who has actually taken the time to read the paper and study its contents. While the committee as a whole should have the right to question the readers of a paper and to present opinions based on the presented summaries or quick scans of the abstract and introduction, the ultimate decision should rest with the readers and nobody else.

Second, an associated problem is the "farming out" of papers to sub-reviewers. Numerous PC members hand papers to acquaintances with a matching background, local colleagues, PhD students and even undergraduates. When these subreviews arrive, they are upload, sometimes without a glance; worse, the assigned PC members don't spend enough time on the papers themselves or don't read them at all. Although it is natural to ask experts for a second opinion, PC members ought to incorporate such opinions into their own evaluations and should not substitute them for their own. Unfortunately, the latter is much more common than authors suspect. During the meeting PC members often read these "secondary" evaluations aloud and then admit that they haven't read the paper. The committee as a whole is left without a true opinion and certainly a submission without a knowledgeable champion in the room, meaning that the decision is again left to non-readers.

Third, traditional PC meetings prefer papers with a focus on narrow technical problems that allow "perfect" solutions over papers on problems that have "muddled" statements because of their real-world connection or over papers with novel ideas whose implications have not been explored in all possible ways.[3] Non-readers bring down the latter papers with long series of minor, negative questions and suggestions. They praise the former kind of papers because they are safe, solid advances on some well-stated problem—even if they are boring and unrelated to reality.

Fourth, conferences should provide good feedback to authors independently of whether the papers are accepted or rejected. Ideally, the discussion that produces the decision should become a part of the feedback, at least as much as possible. In traditional PC meetings, opinions are formed during oral discussions. Even though PC members typically get a week to revise their evaluations, few do and even fewer report the discussions in sufficiently informative detail. The authors are left with haphazard, unproductive feedback.

4.2 Things that Went Well

Although I implemented the voting system in a flawed manner, it worked well in three regards:

1. By design, the system prefers the opinion of readers over non-readers. As the numbers show, the evaluators took cues from the discussions of the papers and withdrew their support on several occasions. On some occasions, the committee as a whole decided that some papers weren't "POPL material." In others, a discussion teased out major flaws. On some rare occasions, it also came to light that submissions were only minor advances over existing papers and would be better off in journal papers.

2. In the same vein, PC members felt free to advocate papers with interesting ideas rather than perfect solutions for technical problems. Several times during the meeting the PC agreed to the value of a paper but heavily disagreed on the potential. Some of these papers didn't have a problem statement per se, but presented an interesting idea. Others had a problem statement and a simple solution but it was also clear that a lot of additional research would be needed to turn this idea into something relevant or solid. In every case, the decision rested with the evaluators and, I believe, the POPL '07 program shows they were willing to take some risks.

[3] Usually, PCs also prefer well-written papers but this dimension is orthogonal to the three choices, though it is much easier to write a well-written paper about a narrow, technical result than about a broad problem from the real world.

3. During the last week, I repeatedly pushed the PC to argue with each other by revising their reviews, not just their votes. This happened in many cases and should have produced somewhat more informative reviews than traditional voting systems.

   Overall, I was impressed by the length and detail of many of the reviews. I hadn't seen such volumes of feedback as an author or as a PC member. Several experienced authors (of accepted and rejected submissions) confirmed this impression in writing and in personal conversations (at ICFP and OOPSLA)

   I had also heard from other PC chairmen about the vast email volume of complaints from rejected (and sometimes accepted) authors. Personally, I received 11 complaints overall. Five authors requested their "scores" meaning an ABCD or numeric quality score. Four wanted to clarify mistaken opinions in reviews. Finally two authors complained about unprofessionalism in their reviews; one was correct and I forwarded the opinion to the responsible PC member; one was wrong and I re-explained the questionable statement to the author.

   Naturally, none of my subjective evaluations can confirm (or reject) that the voting system improved the reviews. In the best case, we could prove that the reviewers edited their reviews in substantial ways *before* the PC meeting, but this would only confirm a built-in feature of the system; it would not demonstrate anything about the quality of the reviews themselves.

In addition, the system also shortened the PC meeting. Altogether we spent 10.5 hours discussing papers, significantly less than conferences of comparable size: eight hours on the first day (including lunch), and two and a half hours on the second. With a bit of more self-discipline within the PC, the meeting could have been wrapped up after one day.[4]

The voting system could not completely prevent the problem of reviewers farming out papers and not having read their assigned papers. Although I made it clear from the beginning that a score needed the personal backing of a PC member, we still had a small number of cases where a PC member had simply translated a score from a subreviewer into a vote without further attention to the paper. This is naturally disappointing but I can't think of a mechanism to overcome this form of irresponsibility.

4.3 Things that Went Wrong

In implementing the voting system, I made two mistakes. First, I re-opened the submission server on the first day of the PC meeting because people realized how serious I was about the voting system. They wanted to reallocate points to papers that they thought should be discussed. This action significantly disturbed the process and had the PC members worried about their votes rather than the discussion. In hind-sight I should have made it clear that the PC would have eight weeks to read the papers and to vote, and that I would not allow any changes in the vote allocation once the meeting had started.

Second, "white balling" requires too much work from the PC chairman. Although the program doesn't contain too many "white balls," a program chairman may have difficulties implementing a carefully balanced policy. I had personally read 180 of the 200 submissions, and 57 of the 58 discussed papers. I could make some argument for every one of the "white balls" and felt therefore comfortable including them in the program. Having said this, however, reading 90% of the submissions for a large conference is a huge time-consuming effort, and not every PC chairman may have the time to do so. A more efficient mechanism is to allocate additional PC members to "white balls" during the last week so that the

[4] One PC member from the periphery of POPL thought that the meeting had gone extremely well compared to meetings in his own area.

PC has additional information and opinions on such cases. Indeed, a chairman who uses the same voting policy may even require that at least some small share of votes (say 10%) for a paper come from a second member.

A Note of Objection (The Law of Small Numbers): A number of people who have heard of my voting scheme asked me whether I wasn't worried that one PC member could be assigned an unusually large number of "good" papers while another person would have to review an unusually large number of "bad" papers. Although such assignments are definitely possible, I believe such suggestions are indicative of the unwillingness of researchers to make choices and to admit to the fact that conference programs reflect taste as much as raw quality. In the end, my voting system at least guarantees that readers have more influence on decisions than non-readers, which definitely appeals to my sense of fairness.

5 Short Papers

The primary purpose of a conference publication is to present ideas to peers and to get feedback from peers. After presenting several related papers at conferences, an author (or an author team) ought to use the feedback from conferences to write an archival and refereed journal paper. A conference plays its role properly if it accommodates as many papers as possible while assuring an adequate rate and quantity of feedback.

As the organizer or co-organizer of small workshops over the past couple of years, I have experimented with variable-length presentations. In every instance the workshop participants told me how much they enjoyed these short, inspiring presentations as changes from the regular, stodgy programs.

In this spirit, POPL 2007 introduces the category of short papers and presentations because I firmly believe that established mainstream conferences can benefit from such talks and papers, too. The PC chose two kinds of short papers: those that present cool ideas and those that introduce niffty tools. On occasion, a paper is really about "cool tools" both of theoretical as well as practical nature. At the same time, all of these short papers present work that satisfies the same quality standards as regular papers and presentations.

What these short papers demonstrate is that an idea or a tool is presentable with a few pages and in a few minutes. Both short presentations as well as short papers suffice to get feedback from the conference attendees. This feedback, in turn, can motivate future research directions, or it can help with the arrangement of the material for a journal submission.

The POPL conference has at least two benefits, too. First, it can cover more ground with short papers than with just regular papers. Second, while short presentations require more preparation than long presentations, they can also bring a light, refreshing tone to otherwise boring programs.

6 Conclusion

The paper presented the two major innovations concerning the paper selection process of POPL 2007. Even though I didn't exactly implement the market-oriented voting system as I had intended to, my observations suggest that it can play a significant role in improving the quality and the speed of the process. As far as short papers are concerned, the POPL 2007 attendees will have ample opportunity to evaluate the idea and to tell my successor(s) whether to continue, expand, reduce or even abandon the new category.