WOWCS scribe notes

Colin Dixon (University of Washington, ckd@cs.washington.edu)
Fred Douglis (IBM Research, fdouglis@us.ibm.com)
Geoff Kuenning (Harvey Mudd College, geoff@cs.hmc.edu)
Jane-Ellen Long (USENIX Association, jel@usenix.org)
Jeffrey C. Mogul (HP Labs, Palo Alto, Jeff.Mogul@hp.com)

May 20, 2008

Contents

1 Introduction

The Workshop on Organizing Workshops, Conferences, and Symposia for Computer Systems (WOWCS), was held on April 15, 2008, co-located with the 5th USENIX Symposium on Networked Systems Design and Implementation (NSDI 2008) in San Francisco. The goal of WOWCS was to bring together conference organizers (past, present, and future) and other interested people to discuss the issues they confront. Roughly 25 people attended the workshop, including several people via speakerphone.

WOWCS was sponsored by USENIX, with additional support from HP Labs, IBM Research, and Microsoft Research.

The WOWCS papers and presentations are available at `http://www.usenix.org/events/wowcs08/tech/`. However, the workshop was designed to encourage a lot of discussion, covering many points not in the formal papers.

This document is a rather detailed summary of the discussions at WOWCS. It is based on scribe notes from Colin Dixon, Fred Douglis, and Geoff Kuenning, with some additional notes from Jane-Ellen Long. Jeff Mogul merged these notes and did some light editing.

Please do not assume that anything here is an accurate, exact quote. Real-time note-taking is at best an unreliable process, and more errors might have been introduced in the process of merging and editing the notes.

This is an initial draft and has not been reviewed by the participants for accuracy! Please do not circulate! Please send corrections ASAP to jeff.mogul@hp.com.

2 Session 1: Issues Within the Scope of a Single PC

Session Chair: Richard Draves, Microsoft Research

2.1 Best Practices for the Care and Feeding of a Program Committee, and Other Thoughts on Conference Organization

Author: Fred Douglis, IBM T.J. Watson Research Center

Summary of Fred's presentation:

- Fred has chaired Usenix '98, USITS '99, other conferences.
- End goal is a Wiki with a guide and ideas, and best practices. Not a lot has happened on Wiki so far. (The Wiki is available at `http://wiki.usenix.org/bin/view/Main/Conference/CollectedWisdom`; anyone can self-register and contribute material or updates.)
- Problem 1: Bad reviewers. Not everyone is a good PC member. Some never do anything, some do crummy reviews, some are too argumentative, dominate the rest of the PC, or have other personality issues. But how to find who's a known bad reviewer? Past experience, word of mouth? People rarely ask. How about reviewer database? Fred has had experiences where undesirable PC members showed up again on PCs he was on. Database has obvious problems with people not liking what's said.

- Problem 2: inexperienced reviewers. Always want to bring in new blood (how about Google Scholar script to scour for repeat authors who haven't served?). But not everybody knows what to expect or is cut out for it. So avoid taking lots of people you don't know. Set expectations early; have multiple deadlines to force people to miss them early; help with calibration (e.g. publish average scores of reviewers to compare to peers).

- Problem 3: PC Composition. Avoid inbreeding, overlap from year to year, excessive institutional overlap. Established conferences should have lots of people who've published. Reward participation, bring in people from the community.

- Bells & Whistles: options include rebuttals, reviewer ratings. Fred had recent rebuttal experience; he likes it and thinks it might help. In a submission, he was asked to rate reviewers, but he was cautious because it was before results came back, and was worried about the risk that a non-blind reviewer rating could get back to a reviewer. Also, authors do not have much basis to calibrate on.

- Running a conference: different sponsoring organizations can increase or decrease the amount of work (e.g., Usenix is a pleasure). Doing it alone is risky. For new conferences, it's better to be swamped than miss targets. Publicity is really critical. Scheduling needs to avoid conflicts; problematic because other conferences can do things without telling you. Consider impact of rejected papers being submitted to other conferences.

Q & A and comments: Ken Birman: Somebody once tried to bribe me. I wonder if anyone else has had that happen? –cut off for later discussion by Jeff–

Greg Minshall: you saw author side of rebuttals. What's it like from PC side?

Carla Ellis: I was on a PC w/rebuttals. The unfortunate thing is that they didn't seem to get read or paid attention to. The PC chair needs to make sure they get read and actually dealt with: that the PC member who is talking about a paper needs to have read the rebuttal and present it to the PC.

Eddie Kohler: Rebuttals are common in PL community, etc. A rebuttal can kill a paper, but in general it only makes a difference if there's a disagreement on the PC. A case I heard of was where somebody disliked the tone and thought the paper overclaimed; the rebuttal capitulated and said author would fix it. Mostly rebuttals make no difference. An example of a rebuttal killing a paper was when it was criticized for inaccuracy and rebuttal defended with wrong argument.

Jeff Mogul: I've heard of a case where a rebuttal cleared up a misunderstanding and got the paper in. I suspect that generally rebuttals will save 1-2 papers per conference (about 5%)

Fred Douglis: PC members aren't always right. Rebuttals help with PC members misinterpreting things.

Rich Draves: In my experience, out of 30 papers rebuttals made a difference on a couple.

Rich Draves: Authors complain that they don't have time to do a good rebuttal.

Ken Birman: Students, in rebuttals, have shot down their papers by aggressively attacking "bad reviews", which isn't the point. Conferences are polluted by really bad reviews, rebuttals can fix egregious errors. E.g., "the fourth review was not about my paper." The Wiki should point out that rebuttals will be read by reviewers, so don't make them angry!

Colin Dixon: SIGGRAPH has a good tutorial for writing rebuttals.

Eddie Kohler: We do have bad reviews and that's not going to change. If rebuttals are necessary then there's probably already something wrong. The art of learning from a review is important.

Phokion Kolaitis: The practice in PODS is that the draft PC is sent to the executive committee, which includes the last 3 PODS chairs, and this helps weed out bad PC members.

Fred Douglis: The problem is that bad PC members are forgotten in a few years because of how people rotate through PCs. We need longer institutional memory.

Ken Birman: There have been people who were flagrantly bad reviewers, but sometimes people have good reasons for underperformance, so this reputation/blacklist system could be really harmful. I would worry about this. You have to be careful not to be unfair. Sometimes there are short-term emergencies that make people flake out; you don't want to blacklist people for the long term based on that.

Fred Douglis: I think that this needs to not be automated and should contain PC-chair annotations to avoid this kind of mistake.

Ken Birman: Wouldn't it look bad if they'd been on the PC and dropped off?

Fred Douglis: I think it's happened. But I don't really have any good answers here [2, 3].

2.2 What Ought a Program Committee to Do?

Author: Mark Allman, International Computer Science Institute

Summary of Mark's presentation:

- Old world: The Internet is our sandbox. New world: it's crucial infrastructure. That has practical implications, including questions about appropriateness of experiments and measurements. I want to talk about etiquette.
- PCs have been starting to ask these questions, for example: When does measurement traffic become an attack? How should we treat public measurement data? What are rules for things like Planetlab? How should we treat users? It means different things to different people. One example involved a paper based on measurements from a packet tap was put on their laptops. [Greg Minshall: Was there consent or knowledge by the users? Answer: the paper didn't say.]
- A PC's job is to accept or reject papers. Should ethical considerations apply? The easy approach is to just judge technical contributions, and let community police itself. But PCs are in unique position to take action, to avoid spiraling into an ends-justify-means culture, and to avoid getting bad reputation due to shady experimental techniques.
- If the PC is the enforcer, should it stop at just rejection? Should there be investigation, sanction, or reporting? Treat it like plagiarism? Have a database of previous reviews?
- What about the Institutional Review Board (IRB) mechanism? Seems useful, but if an IRB approves do we automatically take it for a conference? What about thorny things that don't involve human subjects?

Q & A and comments: John Wilkes: Explicitly make the methodology visible in the publication; make it a first-order concept, like other parts of an experimental setup. Require authors to disclose their methodologies in their submissions.

Mark's response: Yes, I think that helps.

John Wilkes: Warnings about plagiarism are popping up in CFPs. Does ethics belong there too? We need to give people fair warning, e.g. in the CFP, of what we expect.

Fred Douglis: I like the idea of requiring publishing of methodology. I think having PC members vetting it has issues. [Someone, apparently not Fred, mentions Godwin's law, defined (per Wikipedia): "As a Usenet discussion grows longer, the probability of a comparison involving Nazis or Hitler approaches one." [4]] An example: a paper that collected network data for characterizing connectivity, before and after the bombing of Bosnia. One PC member quit over it; the paper was published but there was a lot of controversy.

Geoff Kuenning: What do other communities do here? That seems like a good place to get guidance.

Mark's response: They have practices that we don't: The IRBs are set up for them, not us. They have community-wide sets of norms, which they've spend a long time coming to a consensus over. It's pretty clear that we don't.

Geoff Kuenning: We need to set up these guidelines and ways to deal with doing the research anyway. Will IRBs have the expertise to evaluate computer-systems studies? They mostly deal with biological and psychological experiments, not packet traces.

Ken Birman: The medical community has had all of these ethical scandals that have come out. We don't check for these things and it's hard because we don't really publish our data, experiments, etc. Mark Weiser proposed that experiments be rerunnable. But Google can say "I did proprietary things, and this is the new standard for how to do things." Should we have a reproducibility standard? Do you think that fraud is occurring in our community?

Mark Allman: I'm not going to stand here and say fraud is occurring in our community. [Tom Anderson: why not!?] In an ideal world I'd like all the data and experimental setups, and be able to compare techniques among researchers. But we've gotten some really nice insights from Google, AT&T, etc. I'd hate to lose those because they weren't willing to release data.

Jeff: The last time you renewed with ACM, you agreed to the ACM ethics code, `http://www.acm.org/about/code-of-ethics`, which covers things like user privacy. Also, traces will dry up very quickly if we punish people for releasing them. An example is the AOL traces that got deanonymized; I don't think AOL will ever publish a trace again. HP is very nervous about the traces we're releasing. It might be a mistake to push the "you must release your data" requirement.

Mark Allman: The deanonymization that someone did of traces that LBL had released has caused at least one person to back off from releasing traces.

Jon Crowcroft: Publishing de-anonymized traces in Europe is a crime. Sorry, don't need any ethics committee at all - just a prison.

Phokion Kolaitis: SIGMOD is using, for the first time, an experimental validation procedure, where the experiments can be re-run to verify the experiments of the accepted papers. This is optional, and will essentially be an asterisk next to the good papers saying results were verified. Many companies are not participating. It's controversial and it might not continue. The lawyers are nervous.

Rich Draves: One last question is in the security space, with the whole class of "how to attack X" papers. Is that bad?

2.3 Program Committee Meetings Considered Harmful

Author: Robbert van Renesse, Cornell University

Summary of Robbert's presentation:

- Bad things about PC meetings: unproductive dynamics (waves of negativity, overly assertive PC members); skewed representation (too much travel for one-day meeting); environmental problems from travelling, especially air flight.
- There might be better ways to spend time and money involved in PC meeting. [Jeff: the speakerphone today is costing us about the same as one air flight.]
- Alternatives: phone PC meetings are tiring. How about 2-3 PC chairs scoring the papers, using online discussions for controversial papers? Then chairs will select highest-ranked papers, with one chance for PC members to disagree with selection. Maybe only the chairs could meet physically.
- Robbert described a simulation experiment to see if PC meetings helped choose a good set of papers. He pointed out that someone had asked if this was a joke, and he said not entirely, but it was somewhat

of a joke. Sample: 150 submissions, 25 papers to accept, 100 experiments run on a simulator. Results: compared to simply averaging the reviewer scores, voting (as simulated) doubles the error rate for rejection, so ranking may be the better determinant.

Q & A and comments: Tom Anderson: Somebody has to be first to throw out the PC meeting.

Fred Douglis: I like the approach and I think you're on the right track. The use of Heavy/Light PCs is a good idea. I like reducing the number of PC attendees, but don't cut down too far. Dropping the heavy PC down to the 2-3 chairs is bad. You need diverse vocal views at the meeting itself. I know of a case where the second-ranked paper was rejected because all three reviewers liked it, but at the PC meeting somebody else pointed out the flaws.

Robbert: you need more than three reviews, and reviewer expertise needs to be well matched to the papers. I think the all papers that get in should be seeing 6ish reviews. and that will help. I doubt that having more people discussing a paper at a PC meeting outweighs the drawbacks of holding meetings.

Ken Birman: You could run the experiment for real for the next OSDI, by doing both systems in parallel for one conference. But it's the PC chair's job to generate the best possible program. That means maybe giving papers to specific reviewers because you know they'll have a critical eye. There are two possible roles for PC chairs. (1) The chair stands off to the side as a referee, or (2) the chair should guide the program, even if you have to manipulate things a little to do that. I believe in a strongly interventional PC chair up to the PC meeting, but relatively quiet afterwards.

Robbert: I think it's important NOT to go down to a single chair. The 2-3 chairs should come up with a reasonable ranking function. A small number of co-chairs helps with the problem of a single power-hungry guy, and avoids at least appearance of favoritism. I also think there needs to be some diversity among reviewers for a given paper. You want expertise, but you also want some breadth. A good rule for PCs as well: get diversity!

Colin Dixon: Do you get most of the benefit of a PC with only 2-3 people?

Mothy Roscoe: I liked all the things you said, but I think I disagree with the whole framework. You point out all the problems that occur at the meeting, but it's not clear why you don't address the problems rather than throw the baby out with the bathwater. I'm in favor of strong chairing, in a facilitative role. If somebody tries to dominate things, the chair should shut them up. Database committees don't have PC meetings. [Phokion Kolaitis: Not true; SIGMOD has a meeting.] My experience is that people don't focus when they're not at a meeting, and you don't see everything. Nobody on the PC has a clear view of the entire field. I don't like very large PCs. I've been at meetings where many papers don't get discussed, and I felt disconnected. I like having a feel of seeing the whole paper selection and all discussions.

Eddie Kohler: I served on an electronic-only meeting and it was kind of nice. I think the meetings could probably go. I thought that people did get involved in papers. It did disconnect me from the conference. Start it with a conference where people are already engaged; you wouldn't want to start a new conference this way, but it would be OK for SOSP or OSDI.

2.4 Paper Rating vs. Paper Ranking

Author: John R. Douceur, Microsoft Research

Note: John sometimes goes by "JD."

Summary of John's presentation:

- The current approach to reviewing conference papers has two steps. First is rating papers. That's an implicit comparison against the conference's quality bar. Step 2 is the PC meeting, and papers are considered one at a time by PC. Each paper is separately judged as being below or above the quality bar.

- Problem 1: quality of accepted papers varies year to year. So PC member from a strong year will have a different quality bar, and ratings aren't consistent.
- Problem 2: separate consideration of papers. PC process gradually sets up consensus about quality bar. That is unfair to papers that have already been judged. (cf. Brehm's phenomenon[1].) Early decisions become psychologically entrenched.
- Proposed alternative: when reviewers assess, don't ask for accept/reject rating. Instead, ask for ranking against other papers. Then, at PC meeting have two phases. First, generate overall ranking of all papers as a partial order. (The method is undefined, but JD claims it's not too important.) Note that you don't need a full overall order. It's primarily at the boundary where you need to set a quality bar; you might want to assign additional reviews there. Second, choose the quality bar setting. If there's a cycle, avoid cutting the cycle. If there's a gap in opinions, that's a good place to cut.

during talk: People pointing out that putting many rankings together isn't necessarily tractable. Put bar on one side or the other of cycles.

Q & A and comments: John Wilkes: We've done this at HP for pay raises. One rule for combining rankings is that you can't change another person's order – only they can. This requires the ability to do pair-wise comparisons across different rankings, of course and so it only works if the people involved have a large enough set of papers to rank them effectively. Also, there have been times when I've gotten half a dozen papers, and all were below the threshold. So this means you have to have a higher reviewing load. [JD: more than the standard 15?] Yes, I asked for 25-30. I really like this idea of ranking papers!

Tom Anderson: Great paper! Having tried to do this for SIGCOMM 2006. We collected rankings and tried to use it additionally. It floundered for two reasons. First, I didn't have help from Eddie's code. Second, I found it hard to come to an agreement about the overall ranking – it was unclear how to go from individual rankings to global order.

JD's response: The original order might not matter much; you just need to get the ballpark right.

Tom Anderson: With a big PC, there's always a conflict of interest. That makes it almost impossible logistically to have a discussion of pair-wise comparisons.

Fred Douglis: I agree that you're going to need to review more papers to get this ranking, but the larger number of reviews would have helped the rating anyway. The more papers you review, the more internal consistency you have. With a sample that big, past bias goes away. If your sample is too small, you might wind up with everything near the bar. Or you might have several people saying "this is the best of my pile," and have the paper migrate to the top even though it's not all that good. It's an interesting idea, but I don't buy it.

JD's response: I don't get how the small sample size hurts. You can even rank only 5 papers. Isn't that much better than rating too?

Fred's response: How do you integrate "best of bad lot" vs. "worst of good lot" (from two different reviews of the same paper)?

Robbert van Renesse: When I go to Google and I do a search it gives me a ranking of pages based on PageRank. I'd really like to see multiple columns of rankings; I don't think there is one good ranking system for all queries. For PCs it's the same, I don't think there's only one ranking metric for papers. [JD: PC chairs should tell the reviewers what they're looking for.] But how about if we made a few different ranking functions. For example, "we want the 10 best technical papers and the 5 most original."

Jeff Mogul: for OSDI 2006, Brian Bershad and I didn't ask reviewers to give an overall rating. They rated technical, novelty, and presentation quality. Then we created an overall ranking based on a linear combination of those scores. We ended up using 1 part tech, 1 part novelty, 0.5 part presentation. That let the PC chairs pick the weighting later, and removed the "making the cut" decision from the reviewers. I

think the scores for papers that arrive at the PC meeting are seriously flawed anyway, and the ranking they imply is essentially useless. Making accept/reject decisions should be done as late in the process as possible. The real question is at what point do you start making the accept/reject decision, not ranking/rating.

John Wilkes: I think the ranking system is really not intended to generate overall ranking; instead it's really to get equivalence classes – i.e., groups of papers that are all approximately the same goodness. The final goal is to create one large equivalence class for "accept."

Fred Douglis: What about outside reviewers?

JD: There are three questions. First, what if a reviewer wants an outside opinion? Second, what if the PC asks? Third, what about professors who delegate everything? The first problem I think is just fine; pull that opinion into the ranking. Outside advice is fine. It's hard to get a single expert onto the PC, and you're basically screwed to fit that into the framework. The second: I don't have a good way to deal with that. The third: I think that's a feature rather than a bug. If you delegate, you need to come up with a way to do a ranking. That's your problem.

3 Session 2: Issues Beyond the Scope of a Single PC

Session chair: Greg Minshall, University of Washington

3.1 Overcoming Challenges of Maturity

Author: Ken Birman, Cornell University

Summary of Ken's presentation:

- Systems conferences are approaching crisis point: overwhelming submission counts, hard to get PC to deal with heavy load.
- Problem feeds on itself because students are trying to get published, think conferences are a dice roll. Conferences roll dice because of first round; Ken believes that the PC members don't get involved until second round (except for the most dedicated members). Many reviews are farmed out to first-year students whose opinion weighs as much as a PC member's. The PC does much better job on the papers that survive into the next round.
- While on the NSDI PC, he found a couple of unreviewed papers, and one of them was one of best papers accepted. (There were several PC members who didn't do some of their reviews.) This happens with a lot of papers.
- What about letting past conference attendees do the first round of reviewing, using social networking with help from HotCRP or Google? We could get reviews from highly qualified people.
- Other ideas are in his paper: For example, offer more ways to publish, like medical conferences. (These need to be citable on CV.)

Q & A and comments: Phokion Kolaitis: Thank you for looking at extended reviewers. How many external reviewers do you use in your community? I saw the data for a recent DB conference that had 160 papers submitted, 22 PC members, 4 reviews per paper (PC submissions had 5-6 reviews each); there were 160 external reviewers, half I didn't recognize.

Ken Birman: There's no reason to believe there are any PC reviews in the first round. Lots of papers get rejected based on two random reviews. When there are 300 papers, even the chair doesn't read all.

Greg Minshall: I like the idea of getting a social network to do the reviews. But one complaint is that multiple reviewers raise fears of getting ideas stolen: the more people that see it, the more people who might be able to steal my work.

Ken's response: Their papers are getting read by random people anyway, they need to deal with it.

Fred Douglis: Should we be looking at approaches used for open patent reviewing?

Ken's response: Yes, and how about a game-theoretic approach, where you bring in the reviews with the paper? This would avoid the initial noise. The BAR (Byzantine Altruistic Rational) model addresses the cheating problem.

Mark Allman: The review form is the same regardless of round, and no distinction is made between the rounds of reviews.

Ken's response: Yes, we do that, it's probably not right.

Ken Birman: At SOSP, what happened was that the stack was overwhelming so you put it off, then panicked and assigned it to a student. Only a few of us read everything.

Jon Crowcroft (via email ... this email apparently was read during the next talk, but the relevant discussion seems to belong in this part of the summary): There's a pro and a con for farming out reviews: Authors of rejected papers have the right to know that the number of people that saw the paper is limited to those bound by some expectation of confidentiality (indeed, patent law expects this) so no unknown reviews. On the other hand, PhD students have a reasonable expectation that they get trained in all aspects of research including reviewing, and need real, as well as simulated, reviewing experiences...

Robbert van Renesse: I often merge student reviews into my own because their reviews are often overly negative. Sometimes I forward on the reviews on with both names, but sometimes I go back to my students and explain why I'm not using their review.

Mark Allman: [missed]

Ken Birman: We don't use sensible ranking. We just average all the scores and put them into a meaningless histogram and trust it.

S. Keshav (on the phone): What's the medical model? What is the standard for publication? If you can publish anything (i.e., things of any length), how do you define a good paper?

Ken's response: The paper can be judged by whether it gets cited.

Tom Anderson: NIPS (Neural Information Processing Systems) does a lot of what you've suggested. For example, they eliminate paper lengths and accept more papers than will be presented. The AI and vision conferences also do this. It doesn't seem to be a problem. But if you accept more, you'll get more submissions. Infocom gets 1000 submissions. So it's possible that you'll expand your reviewing load.

Ken Birman: You're the chair, you can pick what you accept. Don't have to take garbage. I don't think the random PC members are doing good quality control.

Tom Anderson: People will just submit more letters/WIP/short papers and the review load will still skyrocket.

John Wilkes: I raised the idea of short papers for Eurosys, and there was a lot of controversy.

3.2 Thoughts on How to Improve Reviews

Author: Paul Francis, Cornell University; Presented by Robbert Van Renesse

Summary of Paul's presentation:

- Reviews are of poor quality because PC members are overworked. Both ideas in this presentation focus on just this problem.
- Idea one: let PC members ask authors (via the review software) simple questions like "where in the paper can I find this?" Authors can answer only with page/column/line numbers.

- Proposal one: For papers that are rejected by one conference and then submitted to another, pass the reviews on to next PC. That is, create a repository for blind reviews of rejected papers. Authors can revise the paper and resubmit. Then reviewers can do blind review, then check for sanity, or can reduce work by just checking to see if author did reasonable changes.

Q & A and comments: Robbert van Renesse: I'm not sure I like these ideas myself. I think authors should have the opportunity to forget history. But maybe we could make this system optional.

Fred Douglis: I agree, there's a grey area of "this isn't really the same paper" and so it's hard to make it required. What if the author claims it's not really the same paper? Also, authors will cherry-pick which reviews to forward. Could we move toward "accept with revisions", as periodicals already do? We've already moved close to that with shepherding. If you're going to allow fixing, you want to go back to the same reviewers, not bounce it from person to person with attached reviews. Add a month to the schedule to allow authors to submit a new paper in response to reviews.

Ken Birman: People have said a lot about lazy PC members today. Is there really a phenomenon of PC members doing a bad job of reviewing, or is it farming out to others? My experience is that PC members are actually pretty good about this, they at least read all the papers they review even though they might not do the review.

Robbert's response: Many conferences are moving toward the heavy/light PCs. I've been on two-level-PC conferences. The first level comes with strong encouragement NOT to farm work out.

Gün Sirer: For the next NSDI, I've considered using this idea (carrying reviews forward). The situation is really dire, because on the second or third iteration of the paper, there's somebody in the room who has seen it before and slams it because the authors didn't fix problem X. That often sinks a paper. As an author, you often get back a biased review that completely trashes the paper. Occasional jerk reviews punishing you for all time seems bad. It would be nice to not have that negative weight carried forever. My basic sampling is that 60-80% of people are against it.

Ken Birman: All they need to do is change the title.

Gün Sirer: But there should be a penalty for trying to game the system like that.

Robbert van Renesse: I think it needs to be optional; there are so many tricks you can use to get around the system that you can't make it mandatory.

(Ken Birman, Gün Sirer and Robbert van Renesse argue about whether it should be mandatory, and if not participating in the optional thing, will that hurt you?)

John Wilkes: About overload, maybe we need to be more aggressive about limits (i.e., not 14 pages in 9-point type). If you can't get your idea across in 10 pages, you probably don't get it. Regarding prior reviews, it seems useful to be able to draw on prior submissions and reviews. Just some archive of old submissions, and some kind of automated matching of current submissions with old ones in similarity. The PC should take into consideration that the paper might have been very heavily revised.

Greg Minshall: Especially if you could add rebuttals into that archive.

Tom Anderson: Maybe we should declare that submission equals publishing; then you could catch multiple submissions, plagiarism, reviews and rebuttals, etc. It's all public record. You have to be willing to stand by things before they're reviewed. We should declare that you can't submit without putting up a Web version. Also, if reviews are signed, there's a danger of people being cautious, and self-editing their comments in the public reviews. But if reviews were at least visible to other reviewers (at later PCs), that might cause people to be more careful about bad reviews.

Fred Douglis: I know of at least two cases where a rejected paper was resubmitted to the same conference and won best paper (presumably after revision). Saying that it's a reiteration might harm that kind of thing. There is a huge difference between "here's new work" and "here's work that's rehashed and resubmitted." There will be a huge bias against the resubmitted ones.

3.3 Scaling Internet Research Publication Processes to Internet Scale

authors: Jon Crowcroft, University of Cambridge; S. Keshav, University of Waterloo; Nick McKeown (presenter), Stanford University

Summary of their presentation:

- Think about the problem as game theory, and design the incentives explicitly. Any system can be gamed, so can this proposal, but let's experiment.
- Problems: increase in number of papers without matching increase in number of reviewers. This leads to skimpy reviews, declining paper quality, favoritism, overly negative reviews (a systems tradition).
- The "Game" is for authors to get published or feedback without work; reviewers want to minimize work and stay in the "club"; PC chairs want a good conference.
- Suggestions: publish author's acceptance rate for conference. Publish titles & authors of rejected submissions. Give best-reviewer award and public acknowledgment. Or create virtual economy, where reviewers get tokens (1 per review), need three tokens to submit. Accepted papers are credited 1-3 tokens. Some mechanism to bootstrap new authors with a few tokens.
- To improve reviewing, authors should rate reviews, with results circulated to PC. Or use pseudonyms and publish reviews, or even be open. But we have no ideas about how to stop favoring famous people; public reviews might worsen the problem.
- Too much of process is hidden; we don't exploit the Web. So let's make all papers and reviews public and signed (maybe using pseudonyms).

Q & A and comments: Jeff Mogul: Awards for "best presentation" go to the few people who know how to give good presentations. The reason for these awards is to bring up the quality of the poorer presentations, not to raise the quality of the best ones, but the people who need improvement the most know they will never win the award, so the awards provide no incentive for them to try. Maybe recognizing "most improved" rather than "best" is a good way to design the incentives.

Nick McKeown: We need a blend of mechanisms for sure.

Eddie Kohler: I love the idea of publishing papers and putting them all on the web. The word "incentive" I also love. Adding incentives to any process that currently relies on altruism can destroy systems by creating an expectation of payback. (Anecdote about how people who would eagerly do something for free will refuse to do it if you offer them an insultingly low fee.) A lot of things rely on peer pressure, which is already at play, so how do you expect your approach to change that?

Nick McKeown: Peer pressure is just another word for "I want respect from the people in the room." It's clearly a factor, using it requires some more effort and I think it's healthy and you need to get the wording right to do that. Most people want to do the right thing.

Phone voice: If you're already part of the community, then none of these mechanisms really apply, but this is about the margin.

Eddie Kohler: I'm one of the younger people in the room. I feel like I started off writing good reviews, but I've followed the guidance of my elders, which has led me to write shorter reviews of less quality.

Mark Allman: I'm not sure if the idea of pseudonyms for publishing reviews work, since they will eventually be decoded.

Nick McKeown: Yes, it's not clear how to make that work. Wouldn't it be nice though?

Mark: I've co-authored with famous people and I got a review back recently saying "my score is a little high on this because I was giving the authors the benefit of the doubt based on their reputation."

Gün Sirer: It's hard to criticize famous people

Ken Birman: Personal attacks based on fame are a problem. Opening reviews will make that worse.

Geoff Kuenning: Speaking as somebody on the edges of the Club (not quite totally ingrained in the community), people on the edges of club would love to write more reviews, but simply don't exist (in the eyes of PC chairs choosing reviewers). New people and people who are at nowhere universities get ignored despite having good brains.

Gün Sirer: I really liked the ideas including the token economy. I agree that we should make the process as transparent as possible. Some people on the PC voluntarily unblind themselves, this can also sadly unblind the other people or reduce their anonymity. The best and worst papers I've read come from famous people. PCs should head to one extreme or the other: completely blind or completely unblind.

Ken Birman: I think that people are talking about helping friends, but there's also the problem of real attacks. Unblinding would be horrible because it could create animosity.

Nick McKeown: I think we're actually too polite, and need more open criticism and debate even against famous people. Valuable comments don't get said. Unmasking may make people more polite. But I believe this won't really change the culture.

3.4 Towards a Model of Computer Systems Research

Author: Thomas Anderson, University of Washington

Summary of Tom's presentation:

- Tom graphed paper scores with standard deviations for several conferences. There is enormous variation, and a lot of the score differences are statistically meaningless.
- For SIGCOMM '06, they tried to manage randomness. Had large PC split between "light" and "heavy". Light plus heavy binned into accept, marginal, reject, minimizing reviews. Added reviews for papers with high variance or at the margin. Heavy PC pre-accepted half the papers, then discussed the rest.
- The problem is that it's (perhaps) a Zipf distribution of quality. But authors see a square wave (accept or reject). At the margin, the differences are minor.
- Tom graphed citation distributions for SOSP '03 and '05. It's easy to see that there's a Zipf-like distribution. [Gün Sirer: Note that citation distribution doesn't necessarily match the PC's evaluation.] Tom: yes, but the PC shouldn't be choosing for citability anyway.
- Need incentives to make effort reflect merit of idea.
- Reward should be a continuous function. Publish paper rank and error bars. Maybe redo after some time has elapsed.

Q & A and comments: *(During talk)*

Eddie Kohler: Why Zipf and not exponential?

Tom Anderson: It's unclear. There must be some value in the papers at the margin of good conferences so it's heavy tailed as far as it matters. Also, citation counts are, in fact, Zipf.

Greg Minshall: These two curves for the SOSP citations track almost identically. That's impressive.

Tom Anderson: I think it's in part happenstance, but Zipf curves tend to be stable.

(After talk)

Robbert van Renesse: Your error-bar graphs for the PC ratings of submissions don't look Zipf-ish. Are the rankings Zipf-distributed? I'd say that the bottom 50% of submissions don't belong in the conference. I played with setting the true value of papers to Zipf in my simulation, then you might as well give up because nothing gets resolved.

Tom Anderson: Yes, Stefan Savage has pointed out that the bottom 50% in a conference basically don't get cited at all. Citations are Zipf for the top 10-20% of papers, and then it's not. Keep in mind those 10-20% are the papers we're talking about here at the top conferences. It's clearly not a step function; that's important.

Gün Sirer: We do sometimes re-rank papers. At the last two NSDIs the PC ranked papers. Then at the meeting people put down chips to vote for the papers they liked best. They had a limited number of chips. It's interesting that the two ranking systems produce very different results.

Tom Anderson: There's a bunch of work on voting theory and I don't really know it that well, but there's a notion of saying how much you want to accept something rather than just whether you do or don't. But I think we've overestimated our error bars because our system is trying to force people to choose to accept or reject. Also, the Lake Wobegon Effect makes all authors think their papers are high on the curve. Then when they get on the PC, they think that since they got rejected, they should set the bar higher than their paper. That makes the bar creep up.

Krishna Gummadi: How do you like the idea of publishing all correct papers and then publishing a ranking. Only reject "wrong" papers.

Tom Anderson: I don't think we can start there, but we might be able to get there. First we need to get people to understand rankings and how they work before we totally jump into another boat.

Jeff Mogul: Reviews have two functions. We've been discussing reviews as deciding what gets in. But there's also a paper-improving function of reviewing. The PC members need to think about whether they're willing to invest effort into shepherding a particular paper. When people are deciding papers at the margin, the decision can be made based on what people want to shepherd more than what has more or less merit.

Nick McKeown: I think your point about randomness in reviewing is spot on and I think we could really use more data to look at things more closely. Regarding the step function, do you think that graded acceptance would help? Different kinds of acceptance (with an A, B, C, D, etc.). What do you think works and doesn't work?

Tom Anderson: We're taking too radical an approach by being binary. Awards help, but they're just a start. Also, in practice the PC has trouble evaluating both long and short papers effectively at the same time. It's hard to switch back and forth.

Tom: I want to thank the PC chairs who provided the data used in my paper.

4 Session 3: Review-Management Software

Session chair: Greg Minshall

4.1 Banal: Because Format Checking Is So Trite

Author: Geoffrey M. Voelker, University of California, San Diego

Summary of Geoff's paper:

- Banal is a format-checker for PDF documents. It deduces how the document was formatted, and checks this against a specification. Why?
  - Avoid "phone home" mechanism in Acroread. Recent versions of Acroread have a Javascript feature that can be used to call home whenever you read a paper!

  - Ensure anonymity rules.
  - Might also be used to mine bibliographies.
  - Time is precious.

- Banal uses pdftohtml to check dimensions, font sizes, etc., to ensure compliance.
- It's in HotCRP and EDAS, and has been used in over 800 EDAS-supported conferences.
- Deep question: what are the goals of our community? Trivial restrictions on authors, or practicalities of publishing costs and community time?

Q & A and comments: Phokion Kolaitis: What's the % of violators?

Jeff Mogul: At the strictest checking level, a third of the OSDI 2006 submissions flunked. After careful analysis 6 papers were actually kicked out. For SOSP 2007, there were three violators (some for anonymity) but none were kicked out for that. OSDI 2006 did not integrate banal into the submission form (it was too new for that) and some authors were upset that they didn't get immediate feedback about their format.

Geoff Voelker: Now it's part of the submission process. It'll tell you if you have violations, but won't prevent submission.

Greg Minshall: Why not just limit word count?

Jane-Ellen Long: Illustrations matter, so word count isn't perfect. The highly illustrated papers are the abusers.

Ken Birman: This is backwards. We need to find a way to let people submit long papers with video, demos, etc. We should accept 40-page papers, and let people write extended abstracts, then have the whole long paper to be the "official" version. It's an electronic world, and this paper that clogs our shelves isn't relevant.

Robbert van Renesse (in Paul Francis mode): I don't like reading bad papers, but PCs force me to do that. I want to be able to read three pages and reject the paper after that point. I can still write some constructive comments about the first three pages.

Fred Douglis: I think we need to review what people actually publish. We can't review one thing (extended abstract) and let them publish another thing. Conferences that take extended abstracts tend to be low-quality conferences.

Ken Birman: We could put that burden on the shepherd.

Mothy Roscoe: Has a shepherd ever rejected a paper?

Consensus that nobody has. Discussion about a paper which was nearly rejected and somebody had to step in to negotiate a truce.

4.2 Hot Crap!

Author: Eddie Kohler, University of California, Los Angeles

Summary of Eddie's paper (note that most of the discussion took place during Eddie's presentation, and is shown inline here as much as possible):

- Lots of other review-software systems, all have flaws. START and EasyChair are worth considering. Start geared toward submitters. "START submission page is easily found with search engines, so people may find it through serendipity." Easychair popular, minimal, powerful, fast. 1024 conferences using EasyChair in 2008, some signed up to 2011. EasyChair is geared toward virtual PC conferences. [Audience comment: Eddie is known for fixing any flaws within minutes.]
- HotCRP UI design principles: reduce modes, prefer search. Any listing of papers appears by search page.

- Conferences that ask for "exciting but flawed" papers are kidding themselves.
- Reviewers vary widely in self-assessment of their expertise, but the information is still useful.
- Making score scales too precise, as did NSDI '08 with 7 levels, is a mistake. The "Identify the Champion" scale (Google for it) is good. The three-level expertise scale works well (I'm expert, I'm knowledgeable, I'm not expert).
- Eddie thinks binning is good, because you don't want to have to make arbitrary choices. [Greg Minshall?: ID the Champion has the drawback of putting papers into accept/reject slots early. I like the NSDI binning. Jeff Mogul: But the percentiles imply certain numbers of papers in each bin. Tom Anderson: I've found that people put things in the middle categories; they're reluctant to choose the extremes. Ken Birman: These things don't work too well for external reviewers who aren't seeing lots of papers.]
- A number of people have asked for limiting functionality or forcing people to behave; for example, forcing the binning to match the official percentiles. [Jeff Mogul: I had a conference where I got a lot of bad papers, so even though my reviewer average was higher than the other reviewers, I only reviewed 1 paper in the top 20 by overall ranking. Clearly the distribution of paper quality among reviewers isn't uniform. Fred Douglis: So you need to correlate things with the scores other reviewers gave to those papers.]
- The n best-ranked papers aren't necessarily the best conference. Is the goal to accept the best papers or to produce the best program for people attending the conference? [Tom Anderson: Papers may be valid and correct but not interesting. The poster session can be more interesting than the main program.]
- Eddie is against 14-page papers; he sees lots of 5-page introductions, and doesn't want to read 5 pages of motivation. [Ken Birman: But there are some papers that are inherently long. Jeff Mogul: Ken, why don't you start a long-paper conference? or should this really be TOCS?]
- Eddie showed a long list of discriminators extracted from HotNets-V. (see `http://www.read.cs.ucla.edu/hotnets5/forward.pdf`) Some were pretty funny, such as writing in Word (likely reject) vs. TeX (likely accept). A submission PDF 100K bytes or smaller was 42% likely to get in, 500K or larger was 14% likely.
- HotCRP was written to allow both anonymous and non-anonymous submissions and reviews. Based on HotNets-V (2005) results, submitting anonymously was a way to get rejected. 7% of anonymous submissions got in, vs. 25% of non-anonymous submissions. (But the top-ranked paper in the workshop was anonymous.) Eddie thinks that this is an argument against double-blind reviewing. [Geoff Voelker: I think it says the reverse; it says people think being non-anonymous helps them. Tom Anderson: Especially because your evidence says that next time everybody will be non-anonymous. The reality is that there is bias, it's huge, even though people aren't trying to be biased. Also, lots of blind submissions are unblindable (maybe over half).]

Q & A and comments: *Most of this discussion actually took place during the presentation.*

General discussion about Microsoft's CMT application. Seems to be some consensus on not liking it much.

Discussion about ConfMan; it seems to have been abandoned. (Jeff Mogul has an email from the author of ConfMan, Pål Halvorsen, who says he thinks it was one of the first such systems when it started in 1997.

The last major extensions were for ACM MM 2003, and since a lot of the functionality is now available in other systems, he no longer supports ConfMan, as it was just a hobby project.)

Nick McKeown: What do you mean about novelty not helping much?

Eddie's response: for HotNets '05, novelty didn't actually correlate with acceptance as well as technical merit did.

Ken Birman: What about ranking vs. rating?

Eddie Kohler: Papers are likely to be binned instead of pure ranking. Doing away with the bins might actually lose information.

Tom Anderson: In practice, everyone puts everything in the marginal accept and marginal reject bins. We should combine ranking with "identify the champion." I like this idea.

Ken Birman: We're back in this mode where everyone's a PC member, and I think that half the reviews are coming from outside the PC.

Eddie Kohler: I've been surprised how much I've been asked to limit functionality (in HotCRP) and punish people more.

Lots of discussion about normalizing reviews. Dealing with different set points, as well as dealing with discrepancy of sets of papers that have varying average quality.

Someone: PC chairs have broad view, are committed to conference as a whole. Example of paper that was accepted by chairs over objection of PC at PC meeting. ("strong chair" model)

5 Moderated Discussion/Debate/Flaming on Proposals That Go Beyond the Scope of a Single PC

Moderator: Tom Anderson, University of Washington

5.1 List of possible discussion topics

The group brainstormed to create a list of things to discuss:

- Ethical issues
  - Require data/code to be made public? Validation?
  - How do we know how much fraud is happening?
  - Problematic PC members / PC history
- Encouraging better reviews
  - Community reviewing
  - Permanent record of reviews, public/signed reviews
  - Rate the reviewer / timeliness of PC members
- Encouraging better papers
  - No review without publication
  - Publish ranking
  - Charge for submission/publication
  - Avoid permanent PC members
- Leveraging the web

  - Require multiple versions of paper (abstract + longer)
  - Allow papers of any length?
  - Publish all valid papers (present subset?) Two levels of acceptance?
  - Publication as living process
  - PC member database
  - Signup list for people who want to be reviewers/on PCs
- What organization would own the community DBs (of reviewers and/or reviews)?
- Avoiding the appearance of bias
- Rebuttals: good, bad, ugly
- How do we make the PC meetings more productive?
  - Not hold one?
  - Is shepherding helpful?
- Single vs. double vs. not blind
- Do PCs favor PC papers?
- How do you keep people from being listed as authors when they had no appropriate role? [Ken Birman: I'm concerned about people padding CVs by putting all their students on every paper.]
- Selection/review criteria for workshops vs. conferences vs. journals [Jeff Mogul: SIGCOMM has discussed this a lot, has a formal policy that includes how to deal with this during blind reviewing.]
- Identifying problematic PC members / database of PC history; new blood for PCs.
- Role of PC chair: strong chair vs. chair as mechanical tool/facilitator
- Better citation and impact indices for computer systems (or computer science in general)

We did not discuss all of the items on this list.

5.2 Single vs. Double Blind

Ken Birman: This is actually about avoiding the appearance of bias. At SOSP I was harried by Europeans who thought SOSP was biased against Europeans and therefore we should allocate slots for them. Obviously they wanted double blind, but one wanted unblinded reviews but blinded submissions.

Mothy Roscoe: It's worth publicly airing these silly suggestions so that people realize that we're not going to accept them.

Ken Birman: The broader question is to convince ourselves that there isn't room for bias. I've had experience with PCs that clearly *were* biased. The blinding question is all about bias.

Mothy Roscoe: I think PCs are biased, and I'm not sure how we fix it. But some of the problem is that Europeans don't understand the process and thus think their failure to get in is a result of bias.

Greg Minshall: No matter what the blinding, somebody makes accusations of bias. Double blinding seems to be trying to avoid the appearance of bias, but it doesn't seem to help.

Fred Douglis: The database community discussed this and published it in TODS. Why not follow them?

Jeff Mogul: They decided to double-blind, right?

Phokion Kolaitis: Not exactly. On the TODS editorial board, about 1/3 were in favor, 1/3 against, and 1/3 indifferent. We're doing double-blind right now in SIGMOD. The other two DB conferences (VLDB and PODS) are single-blind.

Mark Allman: I take being on a double-blind committee as a reminder to try to not figure out who wrote it. A benefit to double blinding is that it helps to remind the PC that we should try to be academically honest. Maybe that, in and of itself, is useful.

Ken Birman: We polled the SOSP audience in Brighton, and they voted to keep it double-blind even though OSDI is unblinded.

Tom Anderson: But it was close.

Jeff Mogul: I was on the next SOSP PC. We discussed the question via email before the CFP was finished. It seemed like we had the votes to change the rules, but we weren't sure we had the authority, so we went with the status quo.

Tom Anderson: We haven't talked about why changing it might be good. Sometimes bias is reasonable because knowledge of who's involved is useful in evaluating papers. That's not necessarily true with math theorems. But in systems, some people write "this has been done" when they mean "this will have been done." Single-blinding can serve as a counterweight to that. Some reputation and name recognition is good for systems. I'm in favor of a mixed approach, where some conferences are blind and some aren't, so people have a choice of the system. It's a bad idea to give choice within one conference, as HotNets '05 demonstrates; only low-status people chose to be anonymous.

Eddie Kohler: I don't think that's how it worked. It wasn't only low-status people who chose to be anonymous. But I'm not sure about what would happen in the longer term.

Tom Anderson: I think it would develop that way. In the DB community, somebody did the comparison study that we might do between SOSP and OSDI. The NSDI steering committee asked whether there was a difference in the effect. The DB paper did that, and came to the conclusion that there wasn't a difference, but a rebuttal paper said they analyzed the data wrong. In any case, there's a perception of bias.

Phokion Kolaitis: The AMS visited double-blinding and rejected it. The physicists had author-selected double-blinding and dropped it because it was little-used. Social sciences use it more heavily. ACM doesn't have a global policy.

Tom Anderson: There are a large number of people that think we are biased and will no matter what we do and we need to deal with that no matter what.

Eddie Kohler: If this population exists, then is there any point in addressing them at all? Is there any point to trying to satisfy the people who think there's bias?

Mark Allman: Are we biased? How are we biased? How biased are we? Can we collect some data? For example, data about what the perceptions are? If we can answer these questions than we can go off and maybe deal with that.

Rich Draves: Isn't it the appearance, not the actual bias?

Mark Allman: Then is there data we can use to show that there isn't bias, is that useful?

Group saying no, the data doesn't exist.

Gün Sirer: The truth is that European systems papers are of lower quality, and are rightfully rejected more often.

Tom Anderson: There's a corollary to Zipf's law that says that authors who publish, get published more. That leads to the conclusion that there will be a natural perception that there's a bias toward the ones who publish. Simply increasing participation might be a cure. (That is, increasing the number of slots for papers.)

John Wilkes: I prefer double blind, only slightly, because it makes it a bit easier for people outside the community to get in, which is good for the community, too. I think it might be better to spend our time looking at PC member bias rather than mechanism bias. Maybe we should try to get data about whether PC members are biased.

Fred Douglis: Something close to Eddie's chart, where papers reviewed by X had Y% acceptance rates. If there was a huge gap over a large sample, it might come out. Also, there's the "appearance of impropriety" problem, and double-blind helps that. Also, when a large organization (company) sponsors a conference, and then somebody from that company does a keynote, it looks bad.

Mothy Roscoe: I've run into somebody demanding to have a paper accepted because they were giving the keynote. Regarding bias, maybe we're the wrong people to discuss whether there's bias. This room is full of people who've had an SOSP paper [show of many hands]. We've all gotten SOSP papers, so we are hugely biased. If we believe this is a problem, we should find the people who think there is bias and we should talk to them about it. I think this is the wrong set of people to be discussing it.

5.3 Publication vs. Submission

Tom Anderson: Next topic! What about publishing everything? Publish before submission vs. publish later (just before the conference). But how do you force a revised version? And there is the issue of when to publish to not hurt industrial authors.

Greg Minshall: Just post the submitted version.

Gün Sirer: How does this lead to better papers?

Greg Minshall: It leads to fewer bad papers.

Tom Anderson: This is a slight incentive to make people stand by their work. If you're unwilling to attach your name to work, then it's probably not quite ready. This might get you around this. In the past, there was a consensus that only great papers would get submitted to SOSP. That's gone.

Geoff Kuenning: How does this show up on a CV? Now it's a prior publication that can't be resubmitted.

Fred Douglis: You put something out there, John Wilkes gave the example of something (which ultimately was cited 100 times) which was rejected 5 times, and then accepted. That's going to make his paper look really bad; you run the risk of having to cut things off early. It also breaks blind submission and exposes things that shouldn't be exposed. There are patent problems; there's a difference between showing the paper to a committee of 20-30 bound by confidentiality and publishing it on the Web.

John Wilkes: I completely agree with Tom that subtle incentives are the way to go, but this is NOT a subtle incentive. The good people will have confidence; the others will fear ridicule and plagiarism. It will also get papers labeled as losers; prior judgements can continue to affect things. Both of these things are bad. Reviewing a really bad paper is cheap; it's the ones in the middle that take time, and those are the ones we still want to receive.

Jeff Mogul: The goal is to increase the floor of the quality of submissions. But this would be a big change that would upset the apple cart of career advancement in both academia and the corporate world. I think it would be a big controversy in the community. I wonder if we can do this in a bit more in an evolutionary way. Even as it stands, career evaluation in systems is hard because we focus on conferences.

Eddie Kohler: I am not a fan of publishing before review, but I really like the version idea that if you submit, you'll get your paper published, guaranteed, but as late as possible.

Jeff Mogul: The publication date has to be the conference date or it screws up patent timelines.

Eddie Kohler: Definitely. But the lawyers have to understand that the submission might result in publication.

Tom Anderson: Some fields don't have a history of confidential review. We do, and so it's expected. Some fields, like math, have a life cycle for articles that includes preliminary versions. In math and physics, early versions of papers go in ArXiv; later ones get published. Offer it as a service to put ideas up with a date.

Unknown: Why tie this to conferences? If we give any benefit to submitted papers we will get a million submissions for any value at all.

Geoff Kuenning: If you guarantee publication, it will generate more submissions, not fewer, especially from China.

Rich Draves: I don't see that. It wouldn't help tenure cases to have rejections on your CV.

Fred Douglis: If you're trying to discourage *bad* papers, don't publish *all* papers. You should just warn people that you're going to decide when the submission of a paper is so egregious that it deserves to be recognized as such, in order to use social pressure to keep down the "Hail Mary" attempts. Don't "out" the reasonable ones (and don't even out the unreasonable ones until the official publication date).

Eddie Kohler: The thing which I really like about this mechanism is that it doesn't require we do this for only one reason. It might reduce bad papers, it will get ideas out there with your name on it, and it is unbiased.

Tom Anderson: Another nice thing is that it gives us a way to evaluate program committees. If the rejects are public, people can see whether the decisions are good. Currently we have no way to judge PCs.

5.4 Reviewing

Ken Birman: Reviewing scores are coming from people who've seen widely varying information. The scores are extremely noisy, and we don't know how to deal with noise. We've talked about bias, and the bias from prior reviews. But the ranking is a bias too. We don't have a concrete suggestion for how to deal with all the noise in the ratings.

Tom Anderson: My proposal was to publish rank and variance. For example, this paper was between 5th and 100th. If we published rankings, we could publish the error bars as well.

Ken Birman: We need to discard the non-PC reviews at some point. Too much noise in them.

Rich Draves: I'd like to see a repository of reviews which stay anonymous (by reviewer) but are attachable to the paper.

Jeff Mogul: It's wise to change things one at a time when experimenting. Coupling this with outing of reviewers would cause confusion. Should it just be a repository, or should the authors be allowed to respond?

Robbert van Renesse: That's what I was going to say. I think it would be good for authors to be able to submit previous reviews signed by the previous PC, along with their rebuttal and a discussion of how they have responded, because it would help them if they took the reviews seriously.

Ken Birman: But they shouldn't be forced to haul along a review from somebody clearly biased.

Jeff Mogul: Or rebut it.

Robbert van Renesse: Maybe they should have to submit the original paper and a diff.

Tom Anderson: Ken, didn't TOCS do that?

Ken Birman: Yes. I don't understand why conferences don't use those methods.

John Douceur: If we had a non-anonymous log of reviewers and reviews, this would encourage people to write good reviews. You say it happens, that people go for revenge, but I don't know of an exception. Are there really evil people?

Ken Birman: I have examples, but I'm not outing them because it matters and it happens. There are people who try to retaliate, try to get people denied tenure, that sort of thing.

John Douceur: Do you think I'm foolish for signing my reviews?

Ken Birman: Yes.

Somebody mentions that JD isn't fighting for a tenure case.

Eddie Kohler: For somebody who writes a lot of reviews (200 a year), it's hard to digest a lot of reviews. How about if HotCRP allows the PC to do an Amazon thing: "This review was helpful/not helpful." The results would be exposed to the PC, and potentially to the chair of a future PC.

Mark Allman: I guess I can see where exposing the reviews would be useful in picking good PC members. I'm just not convinced that carrying along the baggage of old reviews with each paper is going to help.

If you want to rebut a review, do it in the paper, in the introduction or the motivation.

Eddie Kohler: I don't think that this carrying along reviews will happen.

Mark Allman (or Jeff Mogul?): Maybe we should allow authors an extra page or two to talk about previous reviews, common misconceptions, etc.

John Wilkes: If the review points out a flaw, don't you want to know if it was fixed?

Mark Allman: I don't care if it was broken before; I just care if it's correct now. That's my job.

John Douceur: A database of reviews would enable people to evaluate reviewers.

Fred Douglis: A public review database has to be tied with making papers public, and thus it's a big step and I don't think we're there yet. We can't put reviews out there unless the papers are public, because the reviews reveal something about the paper.

Tom Anderson: An interesting angle is that clearly the Web has caused us to move toward an evolutionary process for work. We seem to do all of this stuff for journals, but we do none of it for conferences. Journal papers are considered as "here's the final version of the paper." So we carry the entire history along with it. Everything is context. But it seems like we don't have that case for conference papers, and conferences are the terminal publication mechanism for a lot of papers. I'm not sure how to implement it, though. There are a lot of issues with controlling access, preserving anonymity, etc.

Rich Draves: I don't think this would be that hard to do. At the end of every PC, you get a signed blob of your reviews. You can opt to submit a signed blob.

Tom Anderson: If you're going to make if voluntary, why would anybody do this? If the author had an opportunity to reply to a review, especially if there was PC overlap, he might prefer that.

Rich Draves: For me, as a PC member, I would see a paper that had addressed reviews in a more positive light.

Tom Anderson: There is some validity to the notion of saying "if you saw this before, here's what's changed."

Sanjay Agrawal: In SIGMOD/VLDB they have a few tens of "rollover" papers. The author has the option of rolling them over to VLDB. The reviews are made available to the PC of the second conference.

Mark Allman: If I get this blob, what are my obligations as a PC member? Do I have to look at it?

Several: No

Gün Sirer: You wouldn't get a chance until after the PC meeting.

Fred Douglis: Rollover papers are a great idea. For USITS '99 we had a rollover arrangement with SIGCOMM because of overlapping deadlines. Two papers took advantage; both were rejected from SIGCOMM and accepted to USITS.

Gün Sirer: I heard from Adrian Perrig that the vast majority of security bugs found by reviewers are bogus. They are either trivially fixable or not there at all in the first place. In this way, there are lots of reviews which hit little points. I've often seen reviewers pick on things that weren't core, and the author decided it wasn't worth fixing. That brings up rebuttals, do we need them?

5.5 Rebuttals

Colin Dixon: Students view SIGGRAPH rebuttals that are more than 2-3 sentences as excruciating. You wind up trying to put the entire content of the paper into a couple of pages. It's really horrible.

Ken Birman: Maybe people don't understand the purpose of rebuttals. The point is to give a chance to just say "reviewer B is just plain wrong."

Colin Dixon: If it's only helping papers once or twice a conference, maybe it's not worth it to invent a mechanism here?

Tom Anderson: Maybe the solution is several conferences per year.

Fred Douglis: Colin said that the long rebuttals are the hard ones. So maybe rebuttal size limits are the solution.

Rich Draves: Some people try to slip new work into a rebuttal.

Eddie Kohler: For ASPLOS, you only have a couple of days. My friends just learned to write rebuttals. Did we conclude that rebuttals are fine?

Tom Anderson: Yes, except that they cause students to go insane.

Ken Birman: We need to explain the mechanism to the students.

Greg Minshall: Part of the mechanism is that the authors need to see the reviews. Why not have process transparency, such as reporting when reviewers get assigned, when reviews come in, etc.?

Eddie Kohler: People would sit on the HotCRP refresh button.

Fred Douglis: ManuscriptCentral does that, but it makes people bug me way too much after short periods of inactivity.

Jeff: As far as catching egregious errors, such as missing reviews or reviews submitted against the wrong paper, it seems like there must be better solutions.

5.6 Ethics Stuff

Eddie Kohler: Crappy papers are unethical.

Ken Birman: Cornell has been seeing applicants with very long CVs where a lot of the papers have 8-10 authors. My worry is that the standard for being an author has dropped. We're dropping our standard for "least authorship contribution." People have padded CVs with papers when they didn't really have any contribution to the papers. I'd like to see a published policy of what's acceptable.

Phokion Kolaitis: I've seen a resume of a graduating student who already had more than 50 co-authors. It is difficult to determine whether this person was a real contributor.

Greg Minshall: There are groups in biochemistry where all people in the research group are authors of all papers.

Jeff Mogul: Is this a real problem? If it is, would publishing a policy change it? I don't think so.

Fred Douglis: Yes, I think it would help. If we give guidelines, people will understand what's authorship and what's an acknowledgment. I've downgraded a fellowship applicant on this issue. I agree with Ken. I think a policy and making people check a box on the submission form saying they are really an author would help. In the end it hurts because people think it's just not possible to be an author of so many papers, and judge such candidates harshly.

Eddie Kohler: If that's true, we're already solving the problem. I don't think this is a big problem. I'm fine with a policy, but I don't know what good it will do. I don't think we should invent it here.

Jane-Ellen Long: I've recently been seeing people added to the author list after acceptance.

Jeff Mogul: I've had it happen that an intern did enormous work after submission to get the paper publishable. But it's rare, and I think it's happening often.

Mark Allman: I don't think we need to say "this is what an author is," but some discussion of the implications would be useful. The consumers of the CVs are being misled.

Mothy Roscoe: At my university they just circulated a set of guidelines including a few pages about what it should mean to be an author. There are resources outside for us to look at. Mostly the goal appears to be to stop people for getting the authorship simply for funding it or being the professor.

Tom Anderson: Sometimes people in physics or the medical field get authorship just because they own the lab. Not doing the work, not writing the grant, just owning the lab.

Mothy Roscoe: In systems, sometimes we build really big things that require quite a few people. People are complaining about incredibly narrow SOSP papers. I think we have to be careful about going too far down the path of shrinking author lists.

Rich Draves: Speaking as a consumer of CVs, you think about whether the student was a lead author or being just some random person. If their name is buried in the author list, it hurts them. We should tell them that. I think students should make sure they have good reasons for stuff on their CVs.

Ken Birman: My students get anxious about authorship. Sometimes they feel like they should include everybody, others want to claim sole control. I've had students come to me with a finished paper with my name on it, and I told them to remove me and just do an acknowledgment.

Greg Minshall: I would be somewhat offended by having to sign a thing or a submission checkbox asserting that I am an author according to some official policy handed down from ACM.

Jeff Mogul: As a consumer of CVs, as soon as there are more than two students on the paper, what I do is read the reference letter. That's what matters. The length of the author list isn't what counts. If the letter doesn't reference their contribution, I basically ignore that paper.

John Wilkes: This has nothing to do with conferences, and everything to do with teaching people to not put their names on everything. As program committees, we should be accepting papers that lead to good programs, not teaching authorship techniques.

Mothy Roscoe: I just remembered that there's a very strong section in the ETH guidelines that the paper should end with a section listing every author and what their contribution was. I found this kind of shocking, I'm not sure we want that but it has been suggested.

Fred Douglis: This workshop isn't just about PCs, it's about organizing conferences. The checkbox might be going too far, but I'd like to see a policy statement. This is similar to issues about who is an inventor on patents.

6 Moderated Discussion on Concrete Things We Can Do

Moderator: Jeff Mogul, HP Labs

List of possible things to discuss:

- Database of PC members and their service
  - Non-controversial at simplest level
  - Reviewer recognition (best reviewer, most improved, etc.)
- Improving citation indices
- Living papers

Some of the discussions in this session got rather tangled up; these have been separated into distinct sections for ease of reading.

6.1 Living papers

Jane-Ellen Long: Suggestion from USENIX board: living papers. All papers from USENIX-sponsored events would be in corpus; authors could submit other stuff related to their papers, and other people could submit links to their own papers. Anyone could comment, with higher weighting for USENIX members. Papers would get rated; gives more info on which papers are valuable. People could link to papers. It would be a way to develop reviewing credentials for the reviewer database; People could volunteer to serve as reviewer, using reviews from this site to jumpstart and validate their credentials.

Jeff Mogul: In the USENIX system, if you could do the "I found this review helpful" thing without being able to cheat, that could be useful.

Jane-Ellen Long: Is this a good idea? Would it be useful? It's all about the reviews/comments being good. It's all about less work, not more.

Ken Birman: I think this is a topic change. This workshop is about a publication resource, which could be very valuable but not directly relevant to organizing conferences. I think it's off-topic. It's also USENIX-specific. [Others disagree about whether this is off-topic.]

Jane-Ellen Long: What if it were non-USENIX specific?

Ken Birman: That would be better.

Eddie Kohler : Think about it this way: here's a resource related to publishing papers that's going to happen anyway. Could it be used to improve the situation?

Jeff Mogul: I don't want to get into a huge scope argument, but what's our goal here?

Fred Douglis: It would be useful to PCs.

6.2 Database of reviewers

Jane-Ellen Long: What would be helpful in attacking the problems we've discussed today?

Ken Birman: A database of people in the community and what they're working on would be very helpful in finding PC members.

[Rebecca Isaacs made a similar suggestion during an earlier session: set up a database of people who want to serve on PCs, but have not done this before. The database entries could include "these are my credentials", "this is why I want to serve", etc.]

Tom Anderson: I don't view this as related to what we currently do. Also, don't just count number of citations as being that important. But overall, it would be valuable. Encourage community knowledge about what we're doing.

Rich Draves: I'd like it for organizing committees. Being able to pull up candidates, where they've served, where they've published, would be helpful. Currently PC chairs have to redo this year after year. Robbert and I made a big Excel spreadsheet when picking the next OSDI PC.

Robbert van Renesse: If a reviewer gets a lot of "this review was helpful," that doesn't prove they're a good reviewer.

Jeff Mogul: It would just be a first cut; then you'd need to read the reviews.

Eddie Kohler: I'd love to see all rejected papers and their reviews go into this database. Put a checkbox on the submission form so that authors can agree to this.

Phokion Kolaitis: Keeping track of quality of reviewers shouldn't be controversial. A database of past PC members is done in some communities, like PODS. But info on quality of reviewers is too explosive.

John Wilkes: When putting a PC together, information from prior chairs is enormously helpful. But the most useful stuff is the "don't touch this person with a 10-foot pole" and you don't want that in the DB. And a database of behaviour would have to age over time.

6.3 Discussion of alternate publishing models

Phokion Kolaitis: The VLDB board is considering a new proposal for a VLBD e-journal with an editorial board. People can submit all year, at VLDB length. An editorial board does quick reviews (8-10 weeks). Once a year, a PC is formed that would see papers and reviews and decide which would appear in VLDB as conference. No proceedings; invite journal papers for presentation rather than the other way around. Third, VLDB is going to create the VLDB E-Journal.

Greg Minshall: I don't think the VLDB approach is going to work. You're just torturing people all year long.

Tom Anderson: And you'll still get 800 submissions at the last minute.

Greg Minshall: Having an editorial board doesn't scale. My radical model is that you don't have an editorial board, you just have comments. Then at the end the comments are used as the filter. Because we're not scaling the review process to match the submission increase.

Tom Anderson: One of my students submits to biology journals and they get much faster reviews.

Eddie Kohler: How long are the papers? A lot of the sciences have very short papers.

Tom Anderson: True, review delay seems related to paper length.

Rebecca Isaacs: I'm not sure the community will do a good job of reviewing. Nobody's going to read a paper by a complete unknown.

Eddie Kohler: The nice thing about this proposal is that it doesn't have to have a particular meaning. We can run it as an experiment in parallel with existing systems, and then figure out how to use it.

Tom Anderson: The WOWCS PC made me read papers because my original didn't have citations. There's a paper in the economics literature that looks at the rate at which famous economists publish. There's been a drop because the famous people have the ability to get people to pay attention – they can publish books and op-ed columns – [Jeff Mogul: and they have tenure] and so they don't bother publishing in journals. Entire departments are famous enough to do this.

Jeff Mogul: They're evading peer review.

Tom Anderson: True. But conferences have a role in recognizing non-famous people. They asked a question of the citation benefit of being in a top conference. We might want to do that.

Mark Allman: What's the incentive to use this thing? I could issue a tech report.

Eddie Kohler : There's a check box in the submission form...

Jeff Mogul: We know there are things that aren't worth a conference slot but deserve to be seen.

Mark Allman: So the PC should do that?

Jeff Mogul: Or the author, at the PC's suggestion.

Mark Allman: But everybody could do that with a tech report.

Fred Douglis: Getting published in SOSP, etc. means somebody has vetted things. The published archive is unreviewed stuff, so it won't have much impact.

Ken Birman: [missed some] The Web is very faddish. But having random people blog on papers is driven by the issue of the moment. We expect people on PCs to have maturity and be familiar with classical papers. I'd worry that the type of reviews and citations wouldn't be what we're about.

6.4 Short papers

John Wilkes: Short papers. I floated having a short-paper track at Eurosys and got mixed reactions. Some people thought having a short paper on the CV as if it were a full paper was a bad thing. But you need to have it in the proceedings to get travel money. Short papers are sometimes a sort of consolation prize from the PC. It's hard to get one PC to review both short and long papers.

Eddie Kohler: At IMC and IMW, it works really well. They mix the papers very deeply in the program; there's no separate short-paper track. The PC can ask you to write a short paper when you submitted a long one. There were 2-, 6- and 10-to-12-page papers. The other thing I'd like to talk about is position papers vs. short papers. They are two different things.

Tom Anderson: SIGCOMM tried position papers and had trouble distinguishing them. IMC tried "this is a paper and it takes only 6 pages." I'd argue that you should have a separate conference (or name) but run it together.

Jeff Mogul: Isn't that the current SIGCOMM model?

Ken Birman: At ICDCS they do it in the sessions and have workshops on the side as well.

Fred Douglis: At Middleware we're targeting it at a different model, primarily industrial. They're half-length (10 instead of 20 pages) and go into the ACM Digital Library, but not the proceedings.

Ken Birman: We could just offer short-paper submissions as well as long ones, and trust the community to understand that 6 pages aren't the same.

Mark Allman: It's worked well at IMC. I don't think the PC is giving consolation prizes. They do offer a short vs. long paper sometimes, but I think it's been a bit weird. Sometimes it's easy to trim, but sometimes we get 10 pages and it's not clear whether to shorten or lengthen it. The big problem is when two people butt heads in a program committee over acceptance, and it gets resolved by a short paper. Both people win

because the anti guy kept the long paper out and the pro guy got the paper in, but it's not clear it's the right decision.

Robbert van Renesse: We always feel obligated to fill the 14 pages. But 6 pages of good stuff isn't necessarily inferior.

Jeff Mogul: Right, some HotOS papers are cited more often than many SOSP papers.

Jane-Ellen Long: Encourage people to publish papers of "appropriate length." Don't feel obliged to write a certain number of pages.

Tom Anderson: The 14-page limit might be hurting the field. I've been in situations where I had to compromise the quality of the paper to make it fit. That's especially true in networking where the principal journal has the same page limit as the principal conference.

Jeff Mogul: That's a bug, but it's in the journal page limit.

Tom Anderson: We talked earlier about having multiple versions of papers, with no page limit for the longer version.

Jeff Mogul: There was a time when a really good SOSP paper didn't show up in the proceedings because it was passed directly to TOCS.

Ken Birman: But the TOCS papers were often no longer than the SOSP version, because of time constraints. I agree with the concern that cutting things to 14 pages can hurt things.

Jeff Mogul: How about if the reviewers expressed a value per page idea? E.g., "this paper is worth 16 pages; that one could fit in 12." There would be page budgeting problems, but I think we could hit the average.

Jane-Ellen Long: Sometimes there are papers under the maximum size. We might encourage that at CFP time.

Eddie Kohler: I feel differently because I can't come up with a paper where I thought, "Every page is packed with information, I want to reject this paper." Instead I see "If you hadn't babbled for six pages, you'd have room for the results."

Tom Anderson: I've seen a paper rejected (SOSP '93) because critical information got cut to meet the page limit. By the way, the economics community has studied the length of introductions, and they've gone from 1 to 4 pages over 25 years.

John Wilkes: perhaps authors could add a note of "if I had two more pages, this is what I'd add."

Greg Minshall: I worry that you might give extra pages to a bad author because the omission was glaring, while a good author might have made the absence of material unnoticeable.

Jane-Ellen Long: Essentially everybody fills the page limit. Is Eddie right that people expand to the limit?

Jeff Mogul: Poll: how many expand their papers to fit the limit? [everybody is a shrinker – must cut their papers down.]

Greg Minshall, Eddie Kohler: That's not meaningful information.

Ken Birman: Why do we value journal versions? They can cover the topic at full length. Conference length limits mean that sometimes people split one paper into two conferences. I still think there should be an option to have a full version online. But it should still be shepherded.

Eddie Kohler: You're really asking for a more functional connection between conferences and journals. I don't want to see a longer conference page limit; conferences are about 25-minute presentations.

Jeff Mogul: The point of the debate isn't whether the PC should review infinite-length papers but whether they should review what eventually gets published.

Geoff Voelker: Enhancing journal/conference cooperation doesn't reduce work, it just moves it.

Greg Minshall: I suggest people do 6-page extended abstracts, which is what we would review, and we assume that the shepherds work to produce good 12- or 14-page papers. Since there's inter-shepherd variation, you use multiple shepherds. At the PC meeting, you pick the clear accepts and clear rejects, and in the middle you ask who's willing to shepherd.

Rich Draves: You can already publish a long paper as a TR, and include a link in the conference paper.

Jane-Ellen Long: One thing people do to cut the paper to the allotted length is to say "This material can be found online." Sometimes it might be better to have that material in the paper itself. Greg's suggestion of extended abstracts would let the shepherds decide what would be a well-written longer paper.

Geoff Voelker: I'd worry that you can game the system because you can submit more 6-page papers rather than better ones. Also, it would be biased toward people who are extremely good at presenting ideas, and maybe favor the people who don't follow through.

Ken Birman: You could require them to submit both the 6-page version and the full one, so that you could check that the work actually was done.

Jeff Mogul: Some of the older USENIX annual conferences reviewed only extended abstracts. I thought it was a disaster because nobody knew how to write a good extended abstract. But requiring the writer to give both 6-page and 14-page versions would get around this.

Rich Draves: In the worst case you have to read 20 pages.

Jeff Mogul, Robbert van Renesse: It doubles the work for the authors because it's just as hard to write the short one.

John Douceur: I liked Eddie's 25-minute point. But I can't present 14 pages in 25 minutes, so I wind up leaving a lot of stuff out and pointing to the paper.

Geoff Voelker: The question is whether after the presentation you want to do more work. Right now, after the presentation you're done, and you don't continue work on the project.

Tom Anderson: Another model would be to require the accepted authors to write a 6-page synopsis of the paper. That doesn't help the reviewing process but it helps dissemination. Right now we have too many 14-page papers to wade through. It would be saying, "The PC looked at the long version and validated the paper, but here's the essence."

Greg Minshall: I was surprised by Tom's suggestion because my whole goal is to reduce the PC's work.

Tom Anderson: I'm trying to reduce the work on the readers. There are 1000 papers published in networking.

Robbert van Renesse: Where did the 14-page limit come from?

Tom Anderson: It used to be 12, and the font size was absurdly small. USENIX said let's increase the font size, and ACM followed.

Robbert van Renesse: Where did the 12-page limit come from?

[Nobody seems to know]

Jeff Mogul: Can we just reject everything in IEEE format?

[Laughter]

References

[1] J. W. Brehm. Post-decision changes in the desirability of alternatives. *Journal of Abnormal and Social Psychology*, 52:384–389, 1956.

[2] Fred Douglis. Collective Wisdom: A Modest Proposal to Improve Peer Review, Part 1. *Internet Computing, IEEE*, 11(5):3–6, Sept.-Oct. 2007.

[3] Fred Douglis. Collective Wisdom: A Modest Proposal to Improve Peer Review, Part 2. *Internet Computing, IEEE*, 11(6):3–5, Nov.-Dec. 2007.

[4] Wikipedia. Godwin's law. `http://en.wikipedia.org/wiki/Godwin's_law`.