WOWCS '08: Workshop on Organizing Workshops, Conferences, and Symposia for Computer Systems

San Francisco, CA

April 15, 2008

Summarized by Jeffrey C. Mogul (Jeff.Mogul@hp.com)

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. Several participants (Colin Dixon, Fred Douglis, and Geoff Kuenning, with some additional help from Jane-Ellen Long) volunteered to be "scribes" and recorded the discussions in detail. The detailed scribe notes are available from http://www.usenix.org/events/wowcs08/tech/.

Program committee chairs, past and present, should also look at the Wiki created as part of WOWCS: http://wiki.usenix.org/bin/view/Main/Conference/CollectedWisdom. This Wiki is an attempt to create a community handbook for program committee chairs and other conference organizers, and we encourage your contributions.

This article is a shorter summary of WOWCS, for which I have relied mostly on the more detailed notes, and partly on my own recollections. Summarizing necessarily involves deciding what to leave out; I tried to avoid injecting too many of my own biases about what was important, but I did not feel obliged to be perfectly objective.

OVERVIEW

Readers who have been on conference program committees (PCs) are probably already aware of the topics discussed at WOWCS, and the primary goal of the workshop was to provide a forum for these discussions. However, we also wanted the results of the workshop (the papers, scribe notes, and this summary) to be useful to people who are writing papers and would like to know how PCs work, and to people expecting to serve on a PC for the first time.

Computer systems fields (such as operating systems and networking) place a lot more emphasis on conference papers than most other disciplines, especially relative to journal papers. PC members and chairs consequently grapple with a variety of recurring issues, including:

Real or perceived bias: Do PC members consciously or unconsciously favor some authors or institutions? What can be done to assure authors that they are being treated fairly?

Making good decisions without overloading the PC: Given the large number of papers submitted to good conferences, the need to generate many careful reviews per paper, and the difficulty of making good decisions in over-large PCs, what techniques lead to good decisions without excessive reviewer workloads?

Finding good PC members: PC chairs need to recruit PC members who will provide the necessary range of expertise, who will work well with other PC members, and who will actually do the work on time and well. In addition, we value institutional and demographic diversity, and we want to bring new people into the process whenever possible.

"Open" models for publication: The traditional model for selecting useful and "correct" conference publications uses anonymous reviews and confidential reviews, and only the accepted papers ever become public. People have proposed opening up some or all aspects of this process, often in conjunction with social-networking techniques to define what is useful and correct.

Privacy, secrecy, and other ethical issues: Program committees must deal with a variety of social problems, somewhat independent of the technical merits of papers being reviewed.

Lack of institutional memory: Being a PC chair takes a lot of time, and most people do it too rarely to become experts. PC chairs often spend a lot of time re-learning what their predecessors knew how to do. The community needs to convert this scattered folklore into explicit advice and documentation.

WOWCS addressed all of these issues and several others.

ISSUES WITHIN THE SCOPE OF A SINGLE PC

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

Fred Douglis, IBM T.J. Watson Research Center

Fred's experiences come from chairing USENIX '98, USITS '99, and other conferences. He primarily addressed the problems of how to run a PC, describing what has worked for him in the past and some challenges for the future.

Problem 1: Some reviewers and PC members don't do their jobs, do things badly, or behave badly at the PC meeting. PC chairs want to choose good PC members, but it's hard to get the information. One has to rely on word of mouth. One could imagine a database rating the performance of past reviewers, but there are obvious social problems with this.

Problem 2: Inexperienced reviewers. We always want to bring in new blood. but not everybody knows what to expect or is cut out for the job. PC chairs generally avoid taking on too many people they don't know. It's also important to set expectations early; to have multiple deadlines to force people to miss them early; and to help with calibration (e.g. publish average scores of reviewers so that they can be compared against their peers).

Problem 3: PC chairs need to manage the composition of their PCs to avoid inbreeding, too much overlap from year to year, or excessive institutional overlap. Established conferences should have lots of people who have strong publication records. Fred would like to find ways to reward participation and to bring in people from the community.

Regarding the reviewing process per se, Fred discussed options, including author-provided rebuttals and reviewer ratings. He likes rebuttals. He was not so sure about having authors rating reviewers, including the possibility of retaliation by a grumpy reviewer and the difficulty for authors to calibrate their ratings.

Fred pointed out that different sponsoring organizations can increase or decrease the amount of work (e.g., working with USENIX is a pleasure). Doing it alone is risky. For new conferences, it's better to be swamped with papers, and publicity is really critical. It's sometimes hard to avoid scheduling conflicts, because other conferences can do things without telling you.

Fred also set up a Wiki for PC chairs, basically a guidebook with ideas, best practices, and discussions of policy issues. (The Wiki is available at http://wiki.usenix.org/bin/view/Main/Conference/CollectedWisdom; anyone can self-register and contribute material or updates.)

During the comments, several people discussed the value of rebuttals. Carla Ellis pointed out that the PC chair should insist that the PC member who leads a paper's discussion should have read the rebuttal and should present it to the PC. Eddie Kohler remarked that a rebuttal can kill a paper, but in general it only makes a difference if there's a disagreement on the PC. Ken Birman observed that students, in rebuttals, have shot down their papers by aggressively attacking "bad reviews," which isn't the point. Colin Dixon reported that SIGGRAPH has a good tutorial for writing rebuttals. In general, people agreed that rebuttals do save some papers, but only rarely.

Regarding bad reviewers, Phokion Kolaitis reported that the PODS executive committee (including the past three PC chairs) reviews the draft list of PC members, which helps weed out bad ones. Fred observed that we need longer institutional memory. But generally people agreed that an actual blacklist could be harmful and unfair.

What Ought a Program Committee to Do?

Mark Allman, International Computer Science Institute

Since the Internet is crucial infrastructure, that raises questions about the appropriateness of experiments and measurements in papers. For example: When does measurement traffic become an attack? How should we treat public measurement data? What are rules for things such as Planetlab? How should we treat users—should we always get their consent when gathering data about them?

A PC's job is to accept or reject papers; should it also apply ethical considerations? PCs are in a unique position to take action, to avoid spiraling into an ends-justify-the-means culture. But if the PC is the enforcer, should this role stop at rejection, or should there be investigation, sanction, or reporting (as with plagiarism)?

Mark wondered whether the Institutional Review Board (IRB) mechanism would help. Should a PC always defer to an IRB? What about thorny things that don't involve human subjects?

During the comments, John Wilkes suggested (to general agreement) that authors be required to disclose their methodologies in their submissions. Geoff Kuenning asked whether we could learn from other communities; Mark replied that they have community-wide sets of norms, which they've spend a long time coming to a consensus over, and it's pretty clear that we don't have them yet. Geoff wondered whether IRBs would have the necessary expertise. Jeff Mogul pointed out that the ACM code of ethics covers things such as user privacy.

Several people discussed whether authors should be required to release their data in order to make experiments truly repeatable. Mark and others pointed out that this could conflict with privacy concerns, and we're probably better off getting the insights from companies who would clam up if we insisted on releasing that data. Jon Crowcroft pointed out that publishing de-anonymized traces in Europe is a crime.

Program Committee Meetings Considered Harmful

Robbert van Renesse, Cornell University

Robbert asserted that PC meetings are bad because of unproductive social dynamics, and because they aren't worth harming the environment with extra air flights. Phone-based PC meetings aren't much better. How about having two or three PC chairs meeting in person to score the papers based on the review, using online discussions for controversial papers, and offering a chance for PC members to disagree with the selection? He described a simulation experiment to see whether PC meetings helped in choosing 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. The results suggest that voting as is done in PC meetings leads to more errors than his proposal.

During the discussion, people generally agreed that PC meetings have their problems, but that (per Mothy Roscoe) Robbert might be throwing the baby out with the bathwater. Fred Douglis argued that the diversity of voices at PC meetings is important, especially for controversial papers or those that suffered from bad reviewers; Robbert asserted that having enough reviews ("sixish") would be sufficient, as long as there was sufficient diversity among the reviewers for each paper.

Mothy argued that some PC members don't focus very well except at the meeting, and Eddie Kohler agreed that although serving on an electronic-only PC was "kind of nice," it did disconnect him somewhat from the conference.

Mothy doesn't like very large PCs. Fred praised the recent practice of heavy plus light PCs.

Ken Birman discussed two possible roles for PC chairs: (1) the chair as referee, or (2) the chair who guides the program, even if this latter role requires a little manipulation of things. He believes the PC chair should be strongly interventional up to the PC meeting, but relatively quiet afterward.

Paper Rating vs. Paper Ranking

John R. Douceur, Microsoft Research

John described the current approach to reviewing conference papers as having two steps: (Step 1) rating the papers by making an implicit comparison against the conference's quality bar, and (step 2) at the PC meeting, considering papers one at a time, to judge whether it is below or above the bar. He asserts this leads to several problems, including inconsistent ratings owing to inconsistent reviewer beliefs about where the bar is. Also, during the PC meeting, the bar moves as a result of the discussions, which could be unfair. John cited research asserting that early decisions could become psychologically entrenched.

He proposed an alternative: Instead of asking reviewers for an accept/reject rating, ask for ranking against other papers. The PC meeting then has two phases: Phase 1 entails generating overall rankings of all papers as a partial order. In phase 2, one chooses 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.

In the discussion, John Wilkes reported that HP has sometimes done this for setting pay raises, and the process works but requires some additional rules. It probably also would require a higher reviewer load, so that each reviewer sees a large enough set of papers to create a proper ranking. Tom Anderson reported that they tried this for SIGCOMM 2006, and it floundered partly from lack of software support. He also said it was unclear how to go from individual rankings to a global order. Also, 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 suggested it would be hard to do this with external reviewers.

Robbert van Renesse suggested that it might be better to have multiple rank-orderings: For example, "we want the 10 best technical papers and the 5 most original."

Jeff Mogul reported that, for OSDI 2006, reviewers were not asked for an overall rating, but rated technical, novelty, and presentation quality. The PC chairs created an overall ranking based on a linear combination of those scores, which removed the "making the cut" decision from the reviewers. He thinks the scores for papers that arrive at the PC meeting are seriously flawed, and the ranking they imply is essentially useless, and that making accept/reject decisions should be done as late in the process as possible.

John Wilkes pointed out that the purpose of a ranking system is really to get equivalence classes (groups of papers that are all approximately the same goodness) and finally arrive at one large equivalence class for "accept."

ISSUES BEYOND THE SCOPE OF A SINGLE PC

Overcoming Challenges of Maturity

Ken Birman, Cornell University

Ken asserted that systems conferences are approaching a crisis point: Overwhelming submission counts make it hard to get a PC to deal with the heavy load. He thinks PC members consequently do a mediocre job on first-round reviews, often farming these out to inexperienced students, and this turns the paper-submission process into a "dice roll." As an alternative, he suggests that we could let past conference attendees do the first round of reviewing, using social networking techniques. This could get reviews from highly qualified people. Ken also suggested allowing papers of any length but reviewing them based on a shorter version also prepared by the authors.

During the discussion, Ken asserted that there is no reason to believe there are any PC reviews in the first round and that lots of papers get rejected based on two random reviews. When there are 300 papers, even the chair doesn't read them all.

Greg Minshall liked the idea of getting a social network to do the reviews, but he predicted that people would worry about having their ideas stolen. Fred Douglis asked whether we should be looking at approaches used for open patent reviewing. Ken agreed that could be useful, as well as a game-theoretic approach, where you bring in the reviews with the paper. Mark Allman pointed out that review forms are the same regardless of round, and Ken agreed that this is probably not right.

Jon Crowcroft observed, regarding farming out reviews, that authors of rejected papers have the right to expect (per patent law) limits on the number of people who see their papers, but PhD students need to be trained in all aspects of research including reviewing and need real, as well as simulated, reviewing experiences. Robbert van Renesse said that he often merges student reviews into his own, because their reviews are often overly negative.

S. Keshav wondered how, if you can publish papers of any length, do you define a good paper? Ken responded that a paper can be judged by whether it gets cited.

Tom Anderson pointed out that the NIPS (Neural Information Processing Systems) conferences do a lot of what Ken suggested; they eliminate paper lengths and accept more papers than will be presented. But it's possible that this will expand the reviewing load.

Thoughts on How to Improve Reviews

Paul Francis, Cornell University

Paul could not attend, so Robbert Van Renesse presented his paper. Paul observed that reviews are of poor quality because PC members are overworked, and he suggested two solutions.

Idea: Let PC members ask authors (via the review software) simple questions such as "Where in the paper can I find this?" Authors could answer only with page/column/line numbers.

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

Robbert wasn't sure he liked these ideas himself, and he thinks authors should have the opportunity to forget history, but perhaps making this system optional would work. Fred Douglis agreed, pointing out that deciding what is the "same paper" could lead to confusion, arguments, or attempts to game the system. Fred suggested moving toward "accept with revisions," as periodicals already do, and we almost do with shepherding, but this would require adding time to the pre-conference schedule.

Gün Sirer said that, for the next NSDI, he considered using this idea (carrying reviews forward), but is worried that a "jerk review" could stick with a paper and repeatedly poison future PCs against the paper.

John Wilkes suggested addressing the PC overload by being more aggressive about length limits; if you can't get your idea across in 10 pages, you probably don't get it. He thought an archive of old reviews and submissions would be valuable; Greg Minshall suggests that authors should be able to add rebuttals into that archive.

Tom Anderson proposed declaring that submission equals publishing (all on the public record); then we could catch multiple submissions, plagiarism, reviews and rebuttals, etc. Fred Douglis was concerned that this would harm the occasional cases where a paper really did improve after a rejection and would create a huge bias against the resubmitted papers.

Scaling Internet Research Publication Processes to Internet Scale

Jon Crowcroft, University of Cambridge; S. Keshav, University of Waterloo; Nick McKeown, Stanford University

Nick McKeown presented this paper, which proposed that we think about the problems afflicting conferences (an increasing number of papers, without a matching increase in the number of people willing to write good reviews) as a game-theory problem. Then we can try to design the incentives explicitly. (This is an experiment, since any incentive system can be gamed.) In their view, the "Game" for authors is to get published or feedback without contributing additional work; for reviewers to minimize their workload and stay in the "club"; and for PC chairs to get a good conference. Obviously this oversimplifies, since some authors and reviewers might be altruistic, too.

Their proposals included the creation of a virtual economy, where writing a review gets you one token, submitting a paper costs you three tokens, and you get a partial rebate if your paper is accepted (there would be some way to bootstrap new authors into the economy). As non-economic disincentives for submitting too many papers, they suggested publishing each author's acceptance rate for conferences, or even publishing the titles and authors of rejected submissions. For reviewer incentives, they suggested giving out best-reviewer awards. This could be driven by having authors rate the reviews.

The paper also discussed more radical approaches, such as making all papers and reviews public and signed (maybe using pseudonyms).

During the discussion, Jeff Mogul suggested that "best presentation" awards at conferences don't actually help, because the people who most need improvement don't expect they could win. So maybe recognizing "most improved" reviewers rather than "best" is a good way to design the incentives.

Eddie Kohler worried that adding incentives to any process that currently relies on altruism can destroy systems by creating an expectation of payback. Since we already rely on peer pressure, could incentives undermine that? Nick suggested we could find ways to improve how we use peer pressure (phrased in terms of gaining respect from peers).

Regarding publishing all submissions and reviews, some people were positive, but worried that pseudonyms would quickly be decoded. This led to a discussion about whether papers from famous people are currently being favored, and whether making reviews public could lead to more animosity and personal attacks, as well as people helping their friends.

Geoff Kuenning observed that people on the edges of the Club (new people or people at teaching-oriented colleges) would love to write more reviews, but they simply don't exist in the eyes of PC chairs choosing reviewers.

Towards a Model of Computer Systems Research

Thomas Anderson, University of Washington

Tom collected actual review scores from several past conferences and showed a few graphs. For example, differences in scores are for the most part statistically meaningless, especially near the margin between "accept" and "reject." Based on citation counts for SOSP '03 and '05, which is one plausible measure, the quality of the papers is perhaps a Zipf distribution, yet the authors see a square wave (accept or reject). The problem with this reward function is that it might not provide enough incentive for authors to generate good papers; instead, they keep submitting a paper to more conferences until the noise in the reviewing process lifts it above the threshold. He suggested trying to make the reward function continuous, such as publishing paper's rank and error bars; maybe papers need to be re-scored after several years of hindsight.

When Tom co-chaired SIGCOMM '06, they tried to manage the randomness. They used a heavy/light PC model and added more reviews for papers with high variance or at the margin.

During the discussion, Robbert van Renesse questioned whether the reviewer scores were Zipf-distributed. Tom replied (based on a citation forwarded by Stefan Savage) that the bottom 50% of papers in a conference basically don't get cited at all. The citation distribution is Zipfian for the top 10%–20% of papers, and then it's not.

Gün Sirer described how, at the last two NSDI PC meetings, the PC first ranked papers, then people were given a limited number of chips to vote for the papers they liked best. The two ranking systems produce very different results. Tom responded that our system is trying to force people to choose to accept or reject, rather than to quantify the merit of each paper.

Krishna Gummadi asked Tom how he liked the idea of publishing all correct papers along with a ranking, rejecting only "wrong" papers. Tom thought this might be a good long-term goal, but not before we get people to understand rankings. Nick McKeown wondered whether publicly grading the papers we accept would help.

Jeff Mogul suggested that reviews function to decide not only what papers get in but also which papers are worth further PC investment for shepherding. Decisions for marginal papers might depend on whether someone wants to shepherd them.

REVIEW-MANAGEMENT SOFTWARE

Banal: Because Format Checking Is So Trite

Geoffrey M. Voelker, University of California, San Diego

Geoff described banal, a format-checker for PDF documents. It deduces how the document was formatted and checks this against a specification. Banal also has other features, such as detecting attempts to discover reviewer names via a feature of Acroread. Banal is now included in HotCRP and has been used in over 800 EDAS-supported conferences.

During the discussion, Phokion Kolaitis asked how frequently authors violate format rules. Jeff Mogul reported that one-third of the OSDI '06 submissions flunked the stricter check, but only six were kicked out for egregious violations. Greg Minshall asked why we don't just limit the word count; Jane-Ellen Long replied that illustrations matter, so word count isn't perfect.

Ken Birman argued that this is backward: We should be encouraging really long submissions and reviewing on extended abstracts. We shouldn't be counting pages in an electronic world. Fred Douglis countered that we can't review one thing (extended abstract) and let them publish another thing. Conferences that do that tend to be of low quality.

Hot Crap!

Eddie Kohler, University of California, Los Angeles

Eddie described the HotCRP review system, including how its design has been influenced by his preferences about how reviewing should work. (A lot of the discussion reflected differences of opinion about these preferences.) The HotCRP user interface follows two principles: avoid modes, and prefer search (all the ways of listing papers are forms of "search").

Eddie showed a long list of discriminators extracted from HotNets-V. Some were pretty funny, such as writing in Word (likely reject) vs. TeX (likely accept). A submitted PDF of under 100K bytes was 42% likely to get in, but those over 500K were only 14% likely.

HotCRP allows both anonymous and nonanonymous submissions and reviews. Based on HotNets-V (2005) results, submitting anonymously was a way to get rejected: Only 7% of anonymous submissions got in vs. 25% of nonanonymous submissions. Eddie thinks that this is an argument against double-blind reviewing. Geoff Voelker and Tom Anderson drew the opposite conclusion, that some people expect an advantage from having their names known.

There was some discussion of the "Identify the Champion" model for running a PC, which people seemed to like. There was also lots of discussion about normalizing review scores, given reviewers with different set points or the situation in which some reviewers just get assigned worse batches of papers than others.

MODERATED DISCUSSION/DEBATE/FLAMING ON PROPOSALS THAT GO BEYOND THE SCOPE OF A SINGLE PC

Tom Anderson moderated this session. The group began by brainstorming to create a list of things to discuss (which is shown in the detailed scribe notes).

Single-Blind Reviewing (SBR) vs. Double-Blind Reviewing (DBR)
Ken Birman observed that this is actually about avoiding the appearance of bias and convincing ourselves that there isn't any room for bias. Ken and Mothy Roscoe both reported experience with biased PCs; Mothy is not sure how we fix it. One problem is that some people don't understand the process and thus think their failure to get in is a result of bias. Greg Minshall remarked that people still make accusations of bias even with DBR; it doesn't seem to help.

Phokion Kolaitis reported that the database community grappled with DBR. On the TODS editorial board, about one-third were in favor, one-third against, and one-third indifferent, so the decision ended with DBR being used. The SIGMOD conference uses DBR, but the other two DB conferences (VLDB and PODS) use SBR. (There have been several SIGMOD newsletter papers comparing data on the two sets of papers, but they came to opposite conclusions about the merits of DBR.) Ken Birman polled the SOSP audience in Brighton, which narrowly voted to keep SOSP DBR even though OSDI is SBR.

Mark Allman thinks DBR has the benefit of reminding PC members that they should try to be academically honest. However, Tom Anderson argued that sometimes bias is reasonable because knowledge of who's involved is useful in evaluating systems papers, where some authors aren't always honest about what they have actually done (i.e., reputations can help). Tom favors having some conferences SBR and some DBR, so authors have some choice. He observed that many people think systems PCs are biased regardless of what we do; Eddie Kohler asked, if so, what is the point in trying to satisfy them?

Mark Allman asked whether we should collect some data on whether and how we are biased, and how people perceive the level of bias, before trying to deal with the problem.

John Wilkes slightly prefers DBR because it helps outsiders break in, which is good for the community, but he suggested spending efforts on reducing PC-member bias rather than "mechanism bias" and suggested collecting data about individual biases. Fred Douglis suggested that huge biases in per-reviewer acceptance rates might come out.

Mothy Roscoe pointed out that most WOWCS attendees were successful SOSP authors, so the people who really perceived the bias were probably not in the room. We need to find those people and talk with them, not just talk among the insiders.

Publication vs. Submission
Tom Anderson asked whether we should publish all submissions, and, if so, how do you force authors to improve on their submissions? Gün Sirer asked how would this lead to better papers; Greg Minshall argued that it leads to fewer bad papers. Tom believes that people would have an incentive not to attach their names to not-quite-ready papers.

Geoff Kuenning asked how this would show up on a CV, and whether authors could then resubmit these "publications" to another conference. Fred Douglis worried that it would discourage papers that needed a few rounds of reviewing to become really good, that it breaks DBR, and that it creates problems for corporate researchers who might need to file patents.

John Wilkes agreed with Tom that subtle incentives are the way to go, but this is *not* a subtle incentive; in particular, we need to be willing to invest time into the papers in the middle between "really good" and "really bad." Jeff Mogul predicted that this change would create controversy by upsetting the apple cart of career advancement in both academia and corporations, already pressured because we focus on conferences instead of journals. Is there a more evolutionary approach? Fred Douglis suggested just "outing" the really egregious submissions to create some social pressure against them.

Tom Anderson observed that some fields don't have our history of confidential review. In math and physics, early versions of papers go in ArXiv; later ones get published.

Eddie Kohler likes the proposed mechanism because it might reduce bad papers, it will get ideas out there with your name on it, and it is unbiased. Tom Anderson pointed out that, because rejections are public, it gives people a way to evaluate PCs; currently we have no way to do that.

Review Repositories

Ken Birman observed that reviewer scores are extremely noisy, and we don't know how to deal with noise; this can lead to perceptions of bias when people see the results. Tom Anderson pointed out that he proposed publishing rank and variance values.

Rich Draves wants a repository of reviews that stay anonymous (by reviewer) but are attachable to the paper. Jeff Mogul and Robbert van Renesse suggested allowing authors to rebut reviews in the repository and, when resubmitting, to include a discussion of how they responded. Jeff and Ken agreed that authors should not be forever haunted by reviews that are clearly biased.

Robbert suggested requiring authors to submit the original paper and a diff; Ken pointed out that TOCS does that and he doesn't understand why conferences don't use those methods.

Again there was some discussion of the benefits and curses of having signed, public reviews.

Eddie Kohler suggested that HotCRP could allow PC members to rate other reviews as "helpful" or "not helpful." The results would be exposed to the PC and, potentially, to the chair of a future PC.

Mark Allman argued against carrying along the baggage of old reviews with each paper. If authors want to rebut a review, they should do it in the paper. Maybe we should allow authors an extra page or two, to talk about previous reviews, common misconceptions, etc.

Tom Anderson observed that the Web has moved us toward evolutionary processes. We seem to do all of this stuff for journals, but none of it for conferences. A journal paper is considered as the final version, so everything is context. But we don't have context for conference papers, and conferences are the terminal publication for a lot of papers. But implementation raises a lot of issues with controlling access, preserving anonymity, etc.

There was a general discussion of the merits and challenges of whether and how to carry reviews around for resubmitted papers. Some people view positively a paper that explains how it has fixed previously reported problems. Sanjay Agrawal reported that SIGMOD has a few "rollover" papers, where the author has the option of rolling them over to VLDB. The reviews are made available to the PC of the second conference. Fred Douglis reported on successful use of rollover papers from SIGCOMM '99 to USITS.

Rebuttals

Gün Sirer has heard from Adrian Perrig that the vast majority of security bugs found by reviewers are bogus (either trivially fixable or not actually bugs). So do we need to allow rebuttals?

Colin Dixon said that students view SIGGRAPH rebuttals as excruciating attempts to put the entire content of the paper into a couple of pages. Ken Birman observed that this misses the point of rebuttals, which is just to give a chance to say "reviewer B is just plain wrong." We need to explain the mechanism to the students.

Authorship Ethics

Ken Birman was worried about applicants with very long CVs where a lot of the papers have 8–10 authors and that the standard for "least authorship contribution" has dropped. He wants a published policy of what's acceptable.

Greg Minshall observed that there are groups in biochemistry where all people in the research group are authors of all papers (or, as Tom Anderson points out, because they own the lab but didn't do any other work on the paper.) Jeff Mogul wondered whether this is a real problem and whether publishing a policy change it. Fred Douglis thought it would indeed help to provide guidelines distinguishing between "authorship" and something that should be in the acknowledgments. Mothy Roscoe reported that ETH has guidelines including a few pages about what it should mean to be an author. Papers should end with a section listing every author and what their contribution was. He found this kind of shocking and is not sure we want it.

Mothy Roscoe pointed out that, in systems, sometimes we build really big things requiring many people. Since people also complain about the narrowness of some SOSP papers, we should be careful not to push too far on shrinking author lists.

There was general discussion about whether the consumers of CVs should be policing this problem (by questioning whether papers with long author lists add information to their hiring decisions), not PCs.

MODERATED DISCUSSION ON CONCRETE THINGS WE CAN DO

Jeff Mogul moderated the final session.

Living Papers

Jane-Ellen Long brought up a suggestion from the USENIX Board of Directors: "living papers." All papers from USENIX-sponsored events would be in the corpus; authors could submit other stuff related to their papers, and other people could submit comments and links to their own papers. Anyone can rate papers (with higher weighting for USENIX members), which could give more information on which papers are valuable. Also, people who want to develop their reviewing credentials could use their reviews from this site to get started.

Jeff Mogul suggested incorporating Eddie Kohler's proposed "I found this review helpful" mechanism.

People seemed to like the idea, especially if it were not necessarily specific to USENIX events.

Database of Reviewers

Picking up on an earlier suggestion from Rebecca Isaacs, to set up a database of people who want to serve on PCs but have not done this before, Ken Birman asked for a database of people in the community and what they're working on; this would be very helpful when finding PC members.

Tom Anderson and Rich Draves both thought it would be valuable. Rich pointed out that he and Robbert had to make a big spreadsheet when picking the next OSDI PC. Eddie Kohler wants all rejected papers and their reviews to go into this database, if the authors consent in advance.

Generally, people like the idea of tracking past PC membership in one place, but they felt that negative information (especially) about reviewers cannot go into a database.

Discussion of Alternate Publishing Models

Phokion Kolaitis reported that the VLDB board is considering a new proposal for a VLDB e-journal. People can submit all year, at VLDB length. An editorial board does quick (8–10 weeks) reviews. Once a year, a PC is formed to decide which papers would appear in the VLDB conference, based on these reviews. The conference would not have its own proceedings.

Greg Minshall didn't think the VLDB approach would work; it would just torture people all year long. Tom Anderson predicted there would still be 800 submissions at the last minute.

Greg suggested that having an editorial board doesn't scale, so get rid of it and just use comments as the filter. Rebecca Isaacs wasn't sure the community will do a good job of reviewing or that people would read a paper by a complete unknown.

Eddie Kohler observed that this could be run as an experiment in parallel with existing systems, and then we could figure out how to use it.

Tom Anderson described a paper in the economics literature looking at the rate at which famous economists publish. The famous people publish less often because they have other ways (books, op-ed columns, etc.) to get people to pay attention. Conferences are still good for recognizing non-famous people.

Jeff Mogul observed that there are things that aren't worth a conference slot but deserve to be seen. Mark Allman said that authors could get their work out as tech reports. Fred Douglis countered that getting published in SOSP, etc. means somebody has vetted things. A published archive is unreviewed stuff, so it won't have much impact. Ken Birman pointed out that we expect people on PCs to have maturity and be familiar with classical papers, and he worried that this would be lost.

Short Papers

John Wilkes 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.

Eddie Kohler believes that at IMC/IMW, short papers work really well; Mark Allman agreed. IMC mixes the short 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.

Eddie thinks that position papers and short papers are two different things. Tom Anderson said that SIGCOMM tried position papers and had trouble evaluating them. Jeff Mogul pointed out that the current SIGCOMM model is to put position papers in separate workshops on the side; Ken Birman said that ICDCS has short papers in both the main conference and in workshops. Fred Douglis reported that Middleware has some half-length papers, primarily industrial, that go into the ACM Digital Library but not the proceedings.

Robbert van Renesse observed that 6 pages of good stuff isn't necessarily inferior to 14 pages. Jane-Ellen Long suggested encouraging people to publish papers of "appropriate length"; they should not feel obliged to write a certain number of pages. Tom Anderson thought the 14-page limit might be hurting the field; paper quality can be compromised by trying to fit into this limit. Ken Birman agreed. That's especially true in networking, where the principal journal has the same page limit as the principal conference. Jeff Mogul suggested that this bug is in the journal page limit and wondered whether reviewers could express a value-per-page rating (e.g., "This paper is worth 16 pages; that one could fit in 12."). There would be page budgeting problems, but he thinks we could hit the average.

Eddie Kohler couldn't come up with an example of a paper where he thought, "Every page is packed with information; I want to reject this paper." Instead he thought, "If you hadn't babbled for six pages, you'd have room for the results."

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

Jane-Ellen Long wondered whether people expand their papers to fit the limit, but a straw poll of the room revealed that we all struggle to cut the papers down.

Ken Birman asked why 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. He still thinks there should be an option to have a full version online, but it should be shepherded.

Eddie Kohler didn't want to see longer conference papers; conferences are about 25-minute presentations. Conferences and journals have different functions.

Jeff Mogul argued that the debate isn't whether the PC should review infinite-length papers but whether they should review what eventually gets published. Greg Minshall suggested reviewing 6-page extended abstracts, then using shepherds to produce good 12- or 14-page papers. Geoff Voelker worried that you could 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 suggested requiring authors to submit both the 6-page version and the full one, so that you could check that the work actually was done.

Jeff Mogul remembered older USENIX annual conferences in which only extended abstracts were reviewed; he 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. Tom Anderson suggested requiring the accepted authors to write 6-page synopses of their papers. That doesn't help the reviewing process but it helps dissemination, because 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 was surprised by Tom's suggestion, because his goal is to reduce the PC's work. Tom said he was trying to reduce the work on the readers. There are 1000 papers published in networking each year.

The session ended with a discussion of where the 14-page limit came from; it came from letting people use readable fonts on what used to be a 12-page page limit. Where that came from, nobody knew.