HotNets-I: Workshop Summary

David Wetherall and Larry Peterson
Program Co-Chairs

Introduction

The first HotNets workshop was held on October 28 and 29, 2002 at Princeton University. Twenty-five papers were presented over a day and a half in a single-track, discussion-oriented format. Attendance was kept to 65 people to encourage interaction among the participants. The participants were drawn mostly from presenters and authors, plus the organizing committee and 10 students generously supported by an NSF grant.

We organized HotNets with specific goals in mind. We wanted to foster "new idea" papers on early-stage, creative networking research, especially those discussing architectural or high-level design decisions. We also wanted to provide a forum for reflective discussions on the research needs and direction of the broad networking community. The need for such a venue had become steadily apparent to us. Top-tier venues such as SIGCOMM have become highly selective (1 in 12 in 2002!) with the side-effect of discouraging risky papers. Part of the difficulty is that "new idea" papers can be challenging to evaluate because judgment and wisdom are often more important than detailed simulation and measurement. Moreover, when papers cover new ground, there is often little agreement about the proper set of assumptions and importance of problems. Because few existing venues are suited to "new idea" papers, few papers of this kind are published. This is unfortunate, since good architecture and design-oriented papers are often important to the future of networking and have a different impact than purely technical results.

We are not the first to recognize this problem. The CSTB's *Looking Over the Fence at Networks* report[1] in 2001 highlights the "ossification" of the Internet and networking research, and calls for networking research to "more aggressively seek to develop new ideas and approaches." The SIGCOMM Technical Advisory Committee report[2] in 2001 puts forward proposals for re-invigorating the SIGCOMM conference, including complementary workshops. We hope that venues such as HotNets can play a role in stimulating innovative new networking research by providing the community feedback needed to help early-stage, creative work mature into solid conference papers.

The response to HotNets in this role has been positive from its inception. We loosely modeled HotNets on HotOS and the SIGOPS European Workshop, successful workshops in the systems community. We focused on short (6 page) position papers that argued a thoughtful point-of-view rather than full papers that included new results, gathered a small and tightly-knit Program Committee, and emphasized breadth and diversity of topics. The PC included a mix of experience and active researchers in the areas of sensor networks, wireless, and operating systems, as well as the more traditional SIGCOMM topics. Our Call For Papers was even broader.

We received 127 submissions, many more than we had anticipated, across a fairly broad range of topics. We accepted 25 papers, which is quite selective for a workshop. Interestingly, almost half of these papers were based on fresh SIGCOMM rejects! (We leave it to the reader to guess which ones.) This suggests to us that there is a strong demand for a position paper outlet in the networking community. At the workshop itself, we reserved fully half of the formal time for discussion, making do with 15-minute paper presentations and discussing groups of papers together where we could. The size of the workshop, around 60 people, was judged by most to have worked well at providing a welcoming atmosphere for discussion.

Summaries of the workshop discussion are contained in the body of this report. Several overall themes can be seen to emerge. A number of papers led to a discussion of the research process itself: our models and simulation practices, along with testbeds that seek to change the way we do research (and technology transfer). There seemed to be much scope for improvement in these areas, and we note that the NSF has recently held a workshop and announced the creation of a new program in Networking Research Testbeds.

Other clusters of papers were accepted in the areas of peer-to-peer, sensor networks and network and protocol architecture. Papers in these areas mostly explored approaches quite different than the norm. Some of the P2P papers advocated combining information retrieval techniques with P2P structures to enable search. Papers on sensor networks helped to lift us out of an Internet-centric view of networking. The set of papers on architectural issues ranged from the use of overlays to work toward QoS, to the technologies in the future Internet core. The remaining papers constituted a mix of material, little of which shows up in most networking conferences: economic models of networks, router design, network management, and the use of techniques like epidemics in network protocols. This broad set of topics and revolutionary rather than evolutionary papers were very much

what we sought to accept when we first planned HotNets.

There are many people we would like to thank for taking HotNets from an idea to a reality. The other members of the Program Committee (Deborah Estrin, Stefan Savage, Srini Seshan, Scott Shenker, Ion Stoica, Amin Vahdat, and John Wroclawski) contributed an enormous amount of time and energy (far more than they anticipated!) to select a lively program. We asked each PC member to read roughly half of the 127 submissions, including all the likely candidates for acceptance. We were fortunate that Chris Edmonson-Yurkanan took on the role of General Chair. She brought much experience to the task, and this ensured that all of our plans unfolded smoothly. Both the paper reviewing process and the workshop itself ran smoothly too, and we thank Tammo Spalink and Melissa Lawson for much behind-the-scenes "heavy lifting" that made this possible. ACM SIGCOMM, especially Craig Partridge, provided encouragement and lent its sponsorship to the workshop. The NSF, and in particular Ty Znati, helped us by providing a grant that included student travel stipends. We are also indebted to our other sponsor, Intel Research, for key support that enabled a small workshop to remain financially viable. And we wish to thank all of our scribes, listed at the end of this report, for the discussion summaries that form the body of this report.

Planning for HotNets-II is well underway, and we are pleased that Scott Shenker and Hari Balakrishnan have agreed to be PC co-chairs. We hope that HotNets will grow into a successful venue that works in synergy with leading networking conferences by providing feedback and exposure to early-stage work, helping it to grow into mature research, and we would be delighted if HotNets papers help to break new ground for conference publications. How well HotNets succeeds in these respects, however, depends on your support and participation.

Session 1: Architecture

David Wetherall chaired the first session, which included three talks describing architectural proposals. The first, by Ion Stoica, argued that we have failed to deploy QoS in today's Internet, at least partly due to the fact that ISPs have little incentive to change IP. However, with the success of many overlay networks, a natural question to ask is whether overlays can be used to provide QoS functionality. His answer is yes, at least in some cases. Stoica described the architecture of OverQoS, where overlay nodes are deployed in different Autonomous Systems (ASes) within the Internet, and virtual links between entry and exit nodes are used to carry a bundle of application data packets from multiple transport flows across different sources and destinations. The key idea is to use a controlled-loss virtual link (CLVL) abstraction that aggregates a set of flows along a virtual link into a bundle, uses FEC or ARQ mechanism to protect bundle traffic against losses, and redistributes bandwidth and loss across flows within a bundle at the entry node.

During the question period, Dan Rubenstein asked if OverQoS is TCP-Friendly? Stoica responed that the goal of OverQos is not about being TCP friendly, but rather to control losses. Stefan Savage asked if the CLVL abstraction is point-to-point and if there are any routing choices? Stoica said this issue is future work, and that there are no easy answers. Another participant asked if OverQoS supports fair sharing among flows, to which Stoica responded that CLVL is an abstraction; there's freedom in how one manages flows within a bundle.

Bob Braden presented the second talk, which advocated an alternative to the traditional layered architecture: a so called *role-based architecture* (RBA). The idea is motivated by problems with layered architecture; e.g., layer violations, sub-layer proliferation, feature interactions, and middlebox's breaking the end-to-end model. Many of the problems seem to be related to traditional layering. To provide better clarity, more generality and in-band signaling with data flow is need. Braden suggested that we could change the way we think about protocols by giving up layers and using RBA, in which functional building blocks are called roles, packet headers include an arbitrary collection of sub-headers or role-specific-headers (RSHs), and header processing needs new rules for ordering and access control.

Craig Partridge asked if RBA just creates a new name space, which is a known pain. How can we ensure that there are no fights over names? Braden responded by saying that roles are static, so a hierarchical name space should suffice. David Wetherall asked if everyone would need to be aware of roles, as opposed to today, where IP is the single common protocol. Braden replied that roles do not control routing; routing is done as today. Roles are purely local (i.e., are known only inside a node). Larry Peterson asked if roles are significantly different than IPv6 extension headers. John Wroclawski responded that in addition to supporting out-of-order processing, RBA also introduces a problem of role discovery, which is a significant difference from IPv6. Jay Lepreau asked if there are security consequences of out-of-order processing. The answer is yes. Finally, there was a discussion about whether role interaction is a problem, with no consensus reached.

Richard Gold presented the final talk of the session, which began with the observation that there are many forces working against the Internet as a black box, including CDN plumbers, RON do-it-yourselfers, DDoS firefighters, and P2P people. This is because end nodes may know more than BGP (e.g., RON), and ISPs may know more than BGP (e.g., CDN). His main argument is that IP is too direct. The Internet needs an additional level of indirection that allows one to allocate a remote network pointer, and then direct traffic to that pointer, which forwards the traffic to its ultimate destination. He advocated the idea of *network pointers* (NP), which live below IP, thereby making IP an overlay on top of network pointers. NPs can be dynamically allocated and also have processing associated with them. Gold spent the rest of his talk walking through several scenarios that might benefit from network pointers.

Craig Partridge observed that all the papers in this session are creating name spaces, there is fighting over name spaces already, and this is the problem. Braden responded that IANA would assign role IDs and that perhaps there would be a hierarchical name space. Gold replied that NP names are purely local—there is no global name space problem. Gold was also asked about dangling pointers, to which he

replied that they would be handled by higher levels, in an application-specific way. He remarked that NPs can be hidden or explicit (in terms of discovery).

Session 2: Models

Amin Vahdat chaired the second session, which discussed the models we use in networking research. Sally Floyd presented the first paper, which argued that incorrect models constitute a major problem for network research. She provided examples of unrealistic assumptions that have led researchers to wrong conclusions. To make Internet models as simple as possible but not simpler, Floyd proposed application-specific modeling, where only those aspects that are relevant to the examined problem are represented precisely.

The talked provoked a long discussion. Tom Anderson suggested that we need better protocols rather than better models. Floyd responded by saying we need both, the point being that wrong models lead to wrong conclusions about protocols. Stefan Savage then asked if it is the right approach to come up with a model first and to research and design protocols later? Floyd replied that meaningful research requires correct models. Craig Partridge asked how we can get the correct measurement data needed for model validation? Floyd suggested that we start by telling network providers what we need, although we do not know all the relevant parameters right now.

Timothy Roscoe argued that by using an application-specific model, we potentially miss the overall impact of the application on the network. Floyd replied that application-specific models do not narrow the focus. They still represent the complete system, although only relevant aspects are modeled precisely. Ion Stoica wondered if the focus on the right experimental setups creates a hidden danger of not accounting for future developments? Floyd noted that "designing for a benchmark" is bad, but that the analysis of designs should substantiate their merits. David Wetherall suggested that each design should be tested for its sensitivity to all relevant parameters? Floyd agreed, but noted that researchers rely on intuition and their own mental model of what is relevant. It is important to learn which parameters are relevant. Finally, Eddie Kohler and Jay Lepreau both observed that the infrastructure they are building (see next section) attempts to provide such support.

In the second talk of the session, Konstantinos Psounis presented a new methodology for performing large-scale simulations: (1) take a sample of the network flows, (2) feed the sample into a scaled-down version of the network, and (3) extrapolate from the performance of the scaled-down network to that of the original.

During the follow-on discussion, Psounis was asked if in addition to predicting average statistics, his approach could predict the distribution tail, to which he replied that they predict the whole distribution including the tail, but that this can take long time. Taieb Znati then asked if the method is applicable to weird distributions? Psounis responded that it was not, but one can sample on the session level that exhibits a suitable distribution. David Wetherall asked how much the approach can scale down simulations without degrading the prediction accuracy, for example, with errors less than 5%? Psounis pointed out that he had presented two cases. In the first case, scaling-down is unlimited. In the second case, scaling-down by a factor of 50 preserves the precision. Psounis explained that the authors have not looked in general at topology-related problems, which can impose a limit on scaling.

During the final talk, David Alderson suggested a new approach for generating a topology for network modeling and design. The approach formulates topology generation as an optimization problem that considers economic tradeoffs and technical constraints faced by a single ISP in its network design. Solving the problem provides realistic annotated topologies for different levels of ISP hierarchies.

During the discussion, Alderson was asked if his approach considers time as a factor. He replied that the evolutionary growth of the Internet has to be addressed, and that the formulation of his optimization problem can account for legacy issues and requirements of incremental deployment. In response to a question about how the authors know they are right, co-author Walter Willinger responded by saying that to validate the results they need to examine various Internet topologies. Finally, Timothy Roscoe asked a general question about the differences between the agendas of ISPs and the researchers at HotNets. The consensus was that ISPs think the researchers are too theoretical, even though we can tell them many useful things.

Session 3: Infrastructure

John Wroclawski chaired a panel on infrastructure for network research. Eddie Kohler began by describing XORP, an open extensible router platform. Kohler argued that network research is increasingly divorced from reality, with a growing gap between research and practice. Trying new ideas in commercial routers would help narrow this gap, but unfortunately, router vendors are reluctant to deploy new services if there is no immediate monetary return. Even when router APIs are exposed, the basic router functionality is still immutable. XORP makes it possible to experiment with new protocols on routers. It currently includes open APIs for WWW, RTP, SIP, and an open implementation of TCP SACK, and IGMPv3. The code is likely to be released with a BSD-like license.

Stefan Savage commented that XORP's robustness definition (it shouldn't crash) may not be correct. Kohler said that this was a fair point. In this case, robustness means that the forwarding path doesn't crash. Larry Peterson asked about performance robustness; what about a service that runs too long? Kohler replied that the robustness definition is different for user code vs. the forwarding path, and for or the forwarding path, robustness will likely be provided through code review.

The second panelist, Jay Lepreau, argued that research on wireless and mobile communications, as well as on sensor networks, can greatly benefit from emulation and testbeds. He then described how his group is extending Emulab to include a wireless space/time-shared testbed. The idea is to give PDAs and laptops to students, and either attach sensors to public buses or lay them out in large (empty) hangers.

Ratul Mahajan asked about the privacy of students. Lepreau responded that there must be appropriate guidelines to protect privacy. Deborah Estrin asked how hangars can be used for research on sensor networks. What is the input to the sensors? This brings up the issue of the line between reality and simulation when sharing. Craig Partridge noted that there has to be some input (e.g., a moving tank) and Brad Karp asked about interference of external factors, and how they might be estimated/controlled remotely. The consensus was that reproducibility is always a concern in mobile experiments.

Finally, Larry Peterson described the PlanetLab overlay testbed, which is designed to support emerging research into geographically distributed services and network measurement. PlanetLab's goal is to have 1000 viewpoints into the Internet, including both edge sites and network crossroads (e.g., co-location centers). Peterson argued that PlanetLab has the potential to evolve into a service-oriented network architecture in which a common set of base services provide the foundation for high-level services.

During a general discussion period, Peterson was asked what control sites have over the installed PlanetLab nodes. His response was that the main knob is the output bandwidth of the interface card — that is, the percentage of host infrastructure a PlanetLab node is allowed to utilize. Stefan Savage asked why the presenters for the session chose to specialize. Why don't you play together? Kohler responded by saying that his group doesn't have code yet and are focused on routers. Peterson responded by saying that performance is at the bottom of their list of concerns right now, but that eventually PlanetLab will need to move to the physical path (rather than the proxy path) and so the extensible router infrastructure will become useful. He also expects a growing synergy between emulation and PlanetLab's real-world Testbed. In response to a question about the diversity of node placement in PlanetLab, Peterson said the distribution was currently a bit Internet2-centric but that nodes will be going into co-lo centers soon. He also said that PlanetLab would like to see individuals at member sites take machines home and connect them to PlanetLab by cable modems and DSL. Brad Karp directed a question to Peterson and Lepreau, noting that if these projects are successful, researchers will have access to the largest measurement projects in history. Are there tools to make sense of all the data? Peterson responded by saying that building an instrumentation service for PlanetLab is one of their top priorities. Werner Vogels commented that Cornell gave laptops to students, but the real problem was to track them, as it's expensive to put access points everywhere.

Session 4: Routing

Stefan Savage chaired a session on routing. Timothy Roscoe gave the first talk, and began with the observation that routing and firewalls are separate processes in the Internet. He then proposed a new perspective, called *controlled networking*, in which routing and filtering are unified operations. With controlled networking, the presence of every packet is precisely assured and every packet flow is explicitly "white listed". Specifically, a predicate describes what can be seen and what is allowed. He claimed that there is no need to change routers, end-systems, or IP itself to achieve this.

During the discussion that followed, he was asked if there could be high-level languages to specify policy? Roscoe responded that predicates are fairly low level, and probably suitable for being implemented in hardware. Several people then questioned the expressiveness of the system: for example, if it would allow one to stripe packets over multiple paths. Roscoe said such things might not be captured in this model, although the functionality they provide is pretty basic.

Dapeng Zhu gave the second talk, pointing out the weakness of the current interdomain routing protocol (BGP), specifically that it is vulnerable to a single misconfigured router. This is because meaningful filtering/verification of routing updates is not done due to scalability concerns. BGP takes 3 minutes to fail over on average (15-30 min in reality), which prevents mission critical applications from using the Internet (or at least the current IP-level recovery mechanism). He stated that the fundamental problem of BGP routing is that every BGP router needs to have a consistent view of the world, and then proposed that we use a new interdomain routing protocol called Feedback Based Routing (FBR), which basically separates static information (topology) from dynamic information (quality of routes).

During the discussion, Stefan Savage noted that since ISPs try to get traffic off their own network as soon as possible, they have no incentive to adopt the proposed scheme. Zhu responded that it is up to the ISPs, and how they want to tradeoff between cheaper with less traffic on their own networks and more expensive but more robust alternatives.

The next talk, by Venkata Padmanabhan, started with the observation that networks are vulnerable to malfunctioning routers that may compromise routing or misroute packets. Similar problems are also present in other scenarios, such as wireless and p2p networks. He then argued that existing techniques to deal with the problem include flooding link state routing information (unscalable), authenticated routing advertisements (doesn't guard against compromised routers), and central repository (ISPs don't share policy information). Moreover, all these techniques try to protect routing; they don't verify forwarding. He proposed secure traceroute as a technique to detect faulty or malicious routers. The approach assumes single-path routing, and verifies the origin and contents of data. It uses keys for hop-by-hop verification, and identifies faulty links and routers.

Padmanabham was asked what can be done after discovering a bad link. He said that packets should be routed around it if possible. He was also asked if secure traceroute could prevent routers from faking links or routers behind it. Padmanbham said no, but one could apply secure traceroute persistently to discover the problem.

In the final talk of the session, Douglas De Couto reported his experiences with routing in ad hoc networks. He noted that the min-hop-count routing metric is well understood, and the alternatives are complex than simply assuming that hop-count is the most important metric, even in wireless networks where link quality is bimodal. He showed simulation results demonstrating that DSDV is not opti-

mal, and listed some reasons for this being the case, the bottom-line being that the intuition for wired networks is wrong for wireless networks. On the other hand, he argued that a reasonable alternative—maximizing bottleneck bandwidth—doesn't work in practice. Instead, the key goals should be to maximize spectrum usage by minimizing packet transmissions.

During the question period, it was noted that the design space is very large, and that other parameters such as delay should also be considered. De Couto said that these metrics optimize throughput. Ion Stoica asked what additional factors are specific to wireless? De Couto said that the variation in link quality is important, and things generally happen faster; e.g., there are external factors, such as doors opening, elevators moving, and so on.

Session 5: P2P/Overlay Networks

Ion Stoica chaired a session on peer-to-peer (P2P) and overlay networks. Chunqiang Tang gave the first talk, and began by making the following observations: (1) search engines index less than 16% of the web, (2) the Internet is growing by a factor of 3 each year, and (3) Google is growing by a factor of 2.5 each year. The upshot is that search engines only index a modest fraction of the Internet and the growth curve of search engines is not keeping up with the growth curve of the Internet as a whole. He then described pSearch, which has the goal of providing a scalable information retrieval (IR) infrastructure using P2P. He also argued that current approaches are deficient. For example, existing P2P systems are either not scalable, not accurate, or both. Distributed Hash Table (DHT) systems are scalable but do not directly support full-text search.

During the discussion, Tang was asked about an A/V searcher. He said that has been work at IBM to represent A/V files as vectors so this approach should still work.

The next speaker, Edith Cohen, argued that peer-to-peer overlays need a versatile and scalable search mechanism with two features: scope and support for partial-match queries. While centralized solutions provide both features, existing decentralized schemes offer at most one. Cohen proposed associative overlays, a new architecture that supports both partial matches and search for rare items. Associative overlays couple the overlay topology with the search strategy.

During the question period, Lakshminarayanan Subramanian asked why we can't use a web search engine instead? Cohen said that we are looking for a decentralized solution. Eugene Ng asked why we can't use structured overlays, to which Cohen replied that structured overlays rely on distributed hash tables that support only exact matches. We would like to support imprecise queries, like in Google. Chunqiang Tang added that unstructured overlays are also easer to maintain, and that data is growing faster than centralized architectures can cope with. Cohen also noted that centralized approaches also have legal, political, and economic problems. Craig Partridge asked if there is a fault-tolerance tradeoff between centralized and decentralized solutions? Marcel Waldvogel responded that it is easier to mount attacks on decentralized schemes, and Timothy Roscoe commented that P2P services can be easily disrupted and shut down. Cohen responded that P2P services can be made robust.

In the next talk, Marcel Waldvogel describes MITHOS, a scalable p2p overlay network infrastructure for data location that supports low latency routing and fast forwarding using minimum information. After reviewing possible structures for data location—e.g., a DHT, d-dimensional (ir)regular mesh, rectangular tile, distributed tries—he argued that one would like to have a geographic layout in which less routing information is necessary, a rough estimate of relative distances is possible, and even third-parties can figure the distance. His approach was to assign Cartesian coordinate as the id for each node, and do a quadrant-based routing. Since a node's id cannot be known a priori, the proposed approach determines the id during join phase, bootstrapping from some known member.

During the question period, Eugene Ng wanted to know what Waldvogel meant by the possibility that MITHOS could possibly replace IP routing. Waldvogel replied that he had ambitiously stated that as a possibility.

In the last talk of the session, Mayank Bawa observed that the metrics we normally use to evaluate Application Level Multicast (ALM) include stress (duplicate packets on same physical link), delays in routing along an overlay, and increased usage of network resources. Then he argued that existing metrics applied to ALMs are incomplete, and that one should also account for transience of peers and its impact on connectivity of multicast sessions. For example, peers show herd behavior in unstable P2P networks. In order to achieve good end-application performance on such an unstable infrastructure, one must mask peer transience by keeping the topology connected, and by continuing application interrupted. He proposed an *infrastructural peering* layer below applications, between the end-application and data-transfer layer, that is, between RTSP and TCP. This layer maintains state to decouple the two sessions and serves as place holder for primitives for resource discovery and policies for maintaining topology.

Craig Partridge asked why adding a layer helps? Bawa responded by saying that peering functions are difficult and put a heavy burden on application developers. Thus, the peering layer separates concerns: someone provides peering functionality, and all applications use this functionality. Bob Braden suggested that an alternative conclusion to draw was that that RTSP provides insufficient functionality. Venkata Padmanabhan asked, since peering in media streaming is application-specific, what would serve as common useful functionality? Bawa replied that it could be the specification of topological policies; e.g., vertex-disjoint or edge-disjoint paths.

During an extended discussion of the entire session David Wetherall said that the speakers made a strong case for doing IR using P2P, so why don't we throw out DHT and always do this. What is the downside? Someone wanted to know if unstructured overlays make hard-to-find items become *really* hard to find? Cohen commented that DHTs would only work with structured queries, and Tang said that unstructured overlays are easy to maintain and that he is

trying to find a middle ground between structured (DHT) and unstructured systems. Someone asked why not use a centralized service? Cohen pointed out that a centralized service like Google has legal issues to consider. Tang said that a decentralized approach may be better at keeping pace with the growth of the Internet and that it may have better fault tolerant attributes. There was then a long discussion about whether distributed architectures are really better than centralized ones. Stefan Savage concluded that the design space has not been fully explored. We have been restricted to napster, gnutella, and DHTs. In reality, machines have different capabilities, and we should leverage that. Timothy Roscoe objected to the argument that P2P systems have better fault tolerance, since "Civil Attacks" are a very inexpensive and efficient way to bring down a P2P network.

Session 6: Network/Protocol Design Issues

Deborah Estrin chaired a session on network and protocol design issues. Hui Zhang presented the first talk, which argued several points: (1) IP is yet to conquer voice and public data networks; (2) today we have SONET at the edge, and WDM in the middle of the network; and (3) SONET has had a faster growth spurt than the Internet. Zhang then question various IP myths—such as efficiency, simplicity, robustness and low cost—and conclude that IP is a good service layered network (e.g., enterprise voice) but that the core of the network will remain circuit switched.

During the question period, David Wetherall wanted to know what Zhang was advocating. Zhang responded that he is asking a question, more than providing an answer. Where do we want IP to be? Should it takeover the service layer or be the core. Zhang's position is that the former is more realistic. A general discussion about the economics of IP-based networks versus circuit-based networks ensued, with little agreement. Zhang then went back to his original argument, which is that IP has to be driven by enterprise level services; it will take over from the edge. John Wroclawski remarked that this is what already happens, and that Zhang must be advocating something else? Zhang replied that IP's success is in its service model, and a design that deals with heterogeneity, but it is not a complete success in deployment. Timothy Roscoe supported this position by noting that anyone in the business of making money is not using IP in the wide area right now.

The next speaker, Cyriel Minkenberg, shared his experiences working on a multi-terabit-per-second router (2-3 Tb/s) targeted at the OEM market, and comprising either 256 OC-192 or 64 OC-768 links. The design uses a single stage, electronic (not optical) switch. He pointed out that most of what he is presenting is well known to the switch/router design community. He then gave a practical perspective on how most well-known techniques don't work beyond a terabit. For example, he reported that power (and not gate count) is the limiting factor with respect to the number of chips and boards in a system. He also remarked that packaging options are constrained for a variety of reasons, including building codes, form factors, industry standards, power budget (cooling), and serviceability.

Stefan Savage asked that if power and packaging are the main concerns, then why is circuit switching simpler. What are the differences between the design and implementation of circuit and packet switched designs? Minkenberg said that optical switches are well suited to circuit switching. The drawback is that one needs core optics.

In the next talk, Tom Anderson motivated the need for robustness in protocol design. He started with a list of configuration and protocol related problems seen in the Internet, and argued how they did more damage than the Baltimore fire or the WTC attack. He claimed that better protocols are possible, and without too much cost. He backed up this position by giving several examples of major incidents in the Internet, including how BGP handles errors (which results in cascades), how the TCP connection establishment protocol leaves the door open for SYN flooding, and how TCP's fast recovery algorithm is open to attack. He concluded by arguing for a better design methodology—not just software engineering, but better designs—and presented a set of guidelines.

During the discussion period, a questioner pointed out that the guidelines were too general, and designers should have been following them all along. Anderson argued that this clearly has not been the case. Someone then asked if IP-based protocols are less robust? Anderson disagreed; others such as ATM have not been tested.

Werner Vogels presented the final talk of the session. He said the starting point for their work was their troubles with reliable multicast—bigger groups lead to bigger problems (chances of a member being unavailable increase), and the throughput went down under stress. At that point, the authors moved to designing robust distributed systems based on epidemics, which provide probabilistic guarantees, asynchronous communication patterns, and robustness to message loss. He said that even though the science is well understood, epidemic-based systems require significant engineering to work.

Deborah Estrin asked if the size of the system was a problem? Vogel said that communication is scalable, but you need to know complete membership. Local state is on the order of number of nodes, and this is a problem; i.e., how do you select a small subgroup an still provide global guarantees?

During an extended session-wide discussion, a questioner asked Hui Zhang what the utilization of circuit switched networks was. He said roughly 40%. At this point a question was raised as to how this is counted, specifically how is reserved bandwidth that is not used counted. Someone else commented that packet switched networks have to be over-provisioned because of highly dynamic traffic patterns. Tom Anderson blamed TCP for this. Badri Nath asked Zhang what his view of the world was? Zhang said that he is only pointing out facts that the IP community has been ignoring. Brad Karp noted that Cyriel Minkenberg questioned some assumptions, and argued for eliminating some features to bring down the cost of routers. He asked if there are any other assumptions that might be appropriate to discard, such as in-order delivery. Minkenberg responded positively to that suggestion. Someone asked Tom Anderson

about whether following his guidelines would make protocols more complex. Anderson said that there was a strong case for simplicity, but some of our notions of simplicity, such as "circuit switching is complex" are wrong. He also noted that most vulnerabilities he pointed out could be attributed to simplifying assumptions.

Session 7: Sensor/Ad Hoc Networks

Larry Peterson chaired the final session on sensor and ad hoc networks. Sylvia Ratnasamy gave the first talk, and began by explaining that the context of her work is very large networks (a million or more nodes), where each node has local memory, processing, and short-range communication. The goal is to find a scalable and energy efficient method of data retrieval. She pointed out that the main difference between her work and previous work is that her work focuses on data centric *storage* (DCS). The idea is to store data in the sensornet by name, similar to DHTs. Without systematically placing specific data on specific nodes, queries can only be made by flooding the network, with replies returned by the reverse path. With DCS, however, event information is shipped unicast to the specific node assigned to store that event. As a result, queries do not require flooding and can be made more efficient.

During the question period, someone asked what about catastrophic failures of a node that contains important event information? Ratnasamy said one can replicate the data. In response to a question about other fault tolerant issues, she said you can store data along the retrieval path.

Deepak Ganesan gave the second talk on a new data handling architecture for resource constrained sensor networks. The system, called DIMENSIONS, uses wavelet compression to provide better support for observing, analyzing and querying distributed sensor data at multiple resolutions, as well as exploiting spatio-temporal correlations. Typical examples include micro-climate monitoring. The goal of the system is to provide for flexible spatio-temporal querying and mining of sensor data with the ability to drill down on details, while preserving the capability for long-term data mining within the constraints of node storage. These goals can be achieved by exploiting redundancy of the data, the rarity of interesting features, the scale of sensor networks, and the low cost of approximate queries.

Ganesan was asked how he deals with bad sensors. He responded that wavelets are reasonably good for these things, and that you will lose some information anyway. In response to a question about what fraction of the sensors can fail, co-author Deborah Estrin said the work is too recent to say for sure. When asked about how good wavelets are at summarizing, Ganesan said they have two useful features: lossy compression and the fact that they handle general, rather than specific features. He also said they are good for many types of data, especially images, sequences, and time/frequency series.

Jeremy Elson began his talk by questioning if time synchronization really matters. For the Internet, the answer is "sometimes". However, for sensornets, time synchronization is usually a fundamental requirement because they often make time-dependent measurements of physical events. He next questioned whether the problem has already been solved by NTP, 802.11 sync, GPS, WWVB, or high-stability oscillators. He pointed out that the major difference between the Internet and sensornets is the assumption about energy. This means that sending and listening for packets, or operating the CPU is not free. Elson also pointed out that NTP relies on other infrastructure such as GPS to supply out-of-band time information, and that GPS doesn't work indoors or on Mars. GPS is also expensive and rather big. In response to this problem, his approach was to develop a palette of methods, which each application using the most appropriate method. For example, many times one need not worry about having a single global time reference. Instead, each node just needs to know how its clock is related to neighboring nodes. As another example, one could use "post-facto" synchronization: start the system unsynchronized, use an interesting event recorded by all the nodes as a reference, and later line up the timescales to this event. The benefit is that it avoids wasting energy performing the synchronization if it isn't necessary for the operation of the experiment.

Craig Partridge asked what the government uses for underwater time synchronization of their detection networks. Elson didn't know. Partridge then asked how one reconciles a multi-solution approach with the idea that extra solutions add complexity. Elson responded that sensornets can't afford the cost of general solutions.

Badri Nath presented the final talk, arguing that source-addressed routing won't scale for sensornets because the headers will get too big. By describing the routing direction as a mathematical curve and choosing next-hops based on whatever node is closest to the curve in the desired routing direction, one can keep the header small. Part of the reason this can work is that nodes in sensornets often follow some physical topology (e.g., along a river bank) so that "direction" has a well-defined physical meaning. An additional benefit to this approach is that because specific nodes are no longer named, this routing is better able to tolerate node failures. Nath then presented an example of "spoke flooding" where packets are sent along rays from a source node. With only 40% of the communication that standard approaches us, 80% of the nodes can be reached.

When asked if it is possible to reach a particular node, Nath responded by saying yes, if it is within one hop of the trajectory. If not, the node is not reached and it is considered a routing failure.

During a session-wide discussion, David Wetherall commented that these solutions all seem very application-specific. Nath said that his initial domain is a sensor net where the nodes are cars. Elson's domain is environmental monitoring. Traditionally this monitoring is very coarse grained. The scientists he works with are excited at the potential of getting very fine grained data. Craig Partridge observed that it is common to create lots of point solutions in a new space, and wondered if something will gel out, or will there be several solutions? Elson didn't know, but said that by their nature, sensornets are very application-specific. He allowed for the possibility that we can reuse techniques and generalize. In response to a question about the average battery life, Ganesan said months (days if operating at full throttle), and that the goal is one year. Elson pointed out that one could deploy a heterogeneous set of nodes. Some with "bigger" batteries, and that you then set up a hierarchy similar to a memory cache hierarchy.

Acknowledgments

This material was collected by the dedicated efforts of the HotNets I scribe team: Sergey Gorinsky, Scott C. Karlin, Ratul Mahajan, Akihiro Nakao, Dragos Niculescu, Tammo Spalink, and Limin Wang.

References

[1] Computer Science and Telecommunications Board, US National Research Council. *Looking Over the Fence at Networks: A Neighbor's View of Networking Research*. National Academy Press, 2001. `http://www.cstb.org/pub_lookingover.html`

[2] SIGCOMM Technical Advisory Committee. Improving Sigcomm: A few straw proposals, July 2001. `http://www.acm.org/sigcomm/admin/July2001RepFinal.pdf`