Microsoft Research India Summer School on Networking 2009

Okay. There have been no updates to this blog since the first post. I can explain. I spent the last 2 weeks in IISc, Bangalore, at the Microsoft Research India Summer School on Networking, 2009. In a single word, the experience was awesome. But, since I have the luxury of space and time, I am not going to constrain myself to just that.

I went into the Summer School hoping that it would provide me some clarity about what I want to do my PhD in. The choice is roughly between a core specialization like Networking or a more "general" area like Applied Mathematics or Operations Research. I am afraid that I have really not made much progress in that direction yet.

To begin with, I must thank IISc and Microsoft Research India (MSRI) and, in particular, Ashwini, Venkat Padmanabhan, Ram Ramjee, Anurag Kumar, Ranjita Bhagwan etc for organizing this wonderful event and giving me an opportunity to attend the same. They were very helpful and systematic and there was hardly any cause for complaint. Kudos to them all!

Now, let me give you a summary of all the lectures that we had. Naturally, there is a bias towards the things I found interesting or worth thinking about.

Preparatory Lectures

We started with a series of preparatory lectures on the first day.

Anurag Kumar (IISc) gave an introduction about the purpose of the lecture series and what we should try to learn from this. He reminisced about how, when he entered the field of networking in '88, there were at most 15-20 researchers in the entire country working in the area. Infocomm ('94) had not yet started and Transactions on Networking ('84) had just begun. They even had to *encourage* people to use email. The mails were collected all day but all uploads and downloads of mail would only be done every night (as the call rates were cheaper then). Indeed, it has been quite a long and eventful journey to this day and age of the information super-highway.
The most important (and possibly controversial) point he made was this :
As we are embarking on our 40-year career paths now, it is time to decide whether networking is the right area to go into. When AK joined IITK back in the '70s, the hot topic of research at the time was power engineering. Since then, research in the area has matured, gotten much colder, and now, has become just an infrastructure facility (taken for granted?). He expects that the same is going to happen to communication networks too - that they will become as essential and as ubiquitous as power and much less of an active research area. He clarified later that he only meant this to be the case for those interested in system modeling and such, not systems building and designing. He thinks that only about 5 years of active research in this area remains. AK suggested that we look at how the ideas that we use in networking, like those that will be presented to us over the course of the next two weeks, can be applied to other areas such as smart grids, biological systems etc. The three dominant areas of research in the coming years are going to be energy, environment and health.

Going back to AK's claim about networking, is this really true? Will we really have exhausted all the interesting research problems in networking in the coming 5 years? If so, this is not an area to go into for a PhD right now. If I am going to be applying principles from this field to other areas for the rest of my career, I might as well wait and watch how the field evolves, while doing a PhD in something more general, like Operations Research (OR) or Applied Mathematics (AM). I am seriously considering looking at Networking from the OR or AM perspective (in a Business School

or Math Department) rather than going into an EECS department for this. Would this be a good idea?

The next lecture was by **Venkat Padmanabhan** (MSRI). He gave a quick introduction to the common protocols and other basics in networking. It's remarkable that he managed to compress the content of an entire undergraduate course on networking to a 2-hour lecture.

This was followed by a lecture by **Ram Ramjee** (MSRI) on wireless networks. A few key points that I'd like to think about:
1. Why is there an inverse square dependence on frequency in the Friis Free Space Formula?
2. Maximum bitrate in 802.11g is about 54Mbps in theory. But, for a single TCP session, because of all the collision avoidance mechanism, the effective throughput after the overhead is only about 20-30Mbps. RTS-CTS further drops this to about 10Mbps. That is extremely lossy. It naturally makes you wonder if the right area to be doing research in, assuming you want to improve throughput, is in the networking and collision avoidance aspects or in the PHY layer.
Is this the best that we can do? RR believes that a throughput of 1Gbps+ should be "easy" to get, and the real question is if we can go beyond that.
3. The idea of dual-tone carrier sensing. I have to read about this. Basically, a pilot signal on a different tone lets me transmit at full throughput while there is no-one else around. We could also detect collisions in the pilot signal channel to allocate transmit times according to the number of contenders. Is this a workable idea?
4. Power Management (PSM) - This is quite similar to sleep-wake cycling. Also talked to Vishnu Navda (MSRI) who is working on a system that uses input from the network to do PSM on a smartphone. Is all the theoretical work in this area already done? Or is there new stuff?

The final lecture of the day was by **Rajesh Sundaresan** (IISc) rather humorously titled "All you wanted to know about wireless PHY layer but were afraid to ask". He covered a wide spectrum of PHY layer topics, from the basics (modulation) to the hottest research topics of today (MIMO). I won't go into the details here, but it suffices to say that I am looking forward to doing the Wireless Communication course next sem with Prof David Koilpillai.

Broadband Wireless Technologies - Bhaskar Ramamurthi (IITM)
From the next day onwards, we had talks on more specific technical aspects of networking. BR gave a wonderful lecture on PHY layer stuff in 4G. Of course, BR was his usual self - giving us the intuition for things as complicated as CDMA and OFDM. It made me wish I could attend another one of his courses! Content-wise, it was too technical for most people to understand much, especially since the crowd had a heavy CS bias. As for me, I really enjoyed the lecture and understood the basic idea in most of what he said. But I am really looking forward to seeing all the mathematics behind this and think more deeply about it. Luckily, there's a whole semester of that coming up next!

Overlays - Ant Rowstron (MSR-Cambridge)
Ant gave a series of 4 lectures on Overlays, an area that he has pioneered research in.
The first two lectures were quite basic, just laying the foundations by telling us what an overlay is, and then describing the famous Pastry system, "a scalable, distributed object location and routing substrate for wide-area P2P applications" [1].
An overlay is basically a set of selected nodes in a network (such as the users of Gnutella on the Internet) that form a virtual network that is overlaid on the physical network. Typically, this is used by an application to provide greater control to the application designers over the lower-layer functionalities like routing. A typical examples are Gnutella's (and Azureus'?) P2P network that uses

this concept to decentralize information or content so that the costs, both legal and economic, of maintaining a central server are avoided. Akamai and Skype also use overlay networks to do their own routing, directing traffic to their nearest servers to reduce latency or balance load. Overlays can either be Unstructured, like Gnutella's is, or structured, such as the overlays that are created using Pastry, Chord (Hari Balakrishnan et al at MIT) or CAN. The flooding or random walk based queryig in unstructured networks is inelegant and (I'm guessing) inefficient. Pastry tackles this by using a neat idea of a very large id-space of say, 128 bit integers. Both nodeIds and content keys are picked randomly from this large space (by hashing). Each key is managed by its root node, the node with an id closest to the key. Given a key, a node can easily route its query in this id-space using a greedy algorithm involving prefix matching in O(log n) steps. Content dissemination in the overlay network can be done efficiently using multicast trees like in Scribe [2] and striping multiple multicast trees like in SplitStream [3].
The next two lectures were more about how to apply concepts from overlays to other networking problems. According to Ant, there has been so much work creating new overlay topologies for content dissemination that any more research in the area is highly unlikely to be fruitful. Instead, consider Virtual Ring Routing [4] as an example, a wireless routing system that is inspired by structured overlays. Or the vehicle-to-vehicle routing system "PVRP" by G Pau, P Lutterotti and Ant. I don't think the latter has been published yet, but it's an interesting application of overlays along with the innovative use of map data. This is the kind of research in overlays that is actually useful today.

Apart from the lectures, there were also quite a few discussion sessions. In one of these, he said that he doesn't trust simulations much and always tests a prototype before publishing his work. I think that both simulations and prototypes have their own places. While prototypes tell us about how the system is affected by aspects of reality that we may not have anticipated or may have otherwise ignored, simulations let us see how the solution scales and allows us to "explore the parameter space" more completely.

I'd like to mention that Ant gave a very nice list of references which I'm still in the process of exploring to find interesting problems.

[1] "Pastry: Scalable, decentralized object location and routing for large-scale peer-to-peer systems." Antony Rowstron and Peter Druschel.
[2] "Scribe: A large-scale and decentralized application-level multicast infrastructure." Miguel Castro, Peter Druschel, Anne
Marie Kermarrec and Antony Rowstron.
[3] "SplitStream : High-bandwidth multicast in a cooperative environment". Miguel Castro, Peter Druschel, Anne-Marie Kermarrec, A. Nandi,
Antony Rowstron and A. Singh.
[4] "Virtual Ring Routing : Network Routing Inspired by DHTs". Matthew Caesar, Miguel Castro, Edmund Nightingale, Greg O'Shea, Antony Rowstron.

Network Intrusion Detection - Vern Paxson (UCB, ICSI)
This was a first look at security for most people in the audience including myself. Vern gave a series of very enlightening lectures that covered various aspects of network-based detection of attacks. The idea here is to tap some link(s) of the network which most (if not all) of the data in the network goes through and analyze the packets for troublesome patterns. Even though there are many reasons why an end-host based attack detection may be preferable and more effective than a network-based approach, its chief appeal is that it is cheap. As a beneficial side-effect, a lot of insight into the network's general use can also be gathered from this detection.

In his first lecture, Vern told us about the different styles of intrusion detection: those based on signatures, anomaly-detection, and specifications. A variant of the specification-based approach is the behavioral approach. It was quite clear that Vern came heavily on the side of this last approach when he designed and built his NIDS called Bro (keeping in mind the threat to privacy, he gave an Orwellian touch in the name!). The way it works is to tap the packet stream from the network, filter it down using libpcap (packet capture library also used in Wireshark, I think), extract "events" from the large volume of packet data, and finally use policies and past data (state) to decide and execute appropriate action, based on the events reported. The problem with this traditional approach is that with increasingly sophisticated attacks, the gain from using libpcap has diminished over time to practically nothing today. In his next lecture, Vern illustrated how a large amount of state for each connection becomes necessary to detect all attacks (with few false negatives and false positives) once the adversary starts to actively evade detection by the NIDS. First, the adversary could insert "holes" in TCP streams from distributed hosts to force a buffer overflow at the NIDS. Vern showed that with a large memory (512MB), it is possible to prevent an attack from even 20,000 zombies at a time by maintaining 25KB of state for each connection. He seemed to be quite excited about this particular result that he showed in the form of a graph. I thought it was perfectly obvious that with 25kB/connection and 512MB of memory, you can detect 20,000 zombies without affecting benign connections. Maybe I missed something. It turns out that even this full TCP reassembly may not be enough. This attack was quite interesting. The attacker sends a lot of packets with the same sequence number but with different TTL (time-to-live) values in such a way that all of them pass through the NIDS (thus overwhelming it) but only one reaches the end-host (so no ambiguity at the end). There are many other such attacks that use various idiosyncarsies of the end-host in handling ambiguities. Since the NIDS cannot be expected to know and understand all of these, the most promising approach is to "normalize" the data, that is, scrub out all the ambiguities in the data before forwarding it. Naturally, this adversely impacts latency and requires a *lot* of state. But maybe that is the price of security!

There was a change of pace in the next lecture. Vern talked about the increased scanning activity in this decade and how the simple metric of "failure ratio" [1] (i.e. fraction of connection attempts resulting in failure) can lead to very effective detection of scanners. As an aside, I found [2] to be a very interesting read. It basically extends the work in [1] using another metric that also makes intuitive sense. Finally Vern told us about the latest sucurity threat in the form of botnets. In particular, he told us about the Storm botnet, how it was controlled, how it made money, how the security researchers managed to learn about its working and the legal challenges that are faced by officials trying to thwart the botnet menace (inadequate cybercrime laws). Vern also told us about how they used the spam campaign that Storm was raging to collect data about how lucrative spam is. His methodology and findings were extremely interesting. I think this was from [3], but I haven't actually read it yet, so I can't be sure. Quick facts: they make about $80 to $800 per million spam messages (depending on type of spam) and the Storm botmaster is estimated to have made about $3.5 million a year (though probably not for a whole year). BTW, what happened to the Storm botmaster is still uncertain. Rumour has it that he was arrested (possibly in Russia).

Overall, even though this was very interesting reading and listening, I don't think the area is "mathematically sound" enough for me to be interested in it. But if I see more work like [1,2], I would gladly change my mind!

[1] "Fast Portscan Detection Using Sequential Hypothesis Testing". Jung, Paxson, Berger and Balakrishnan.
[2] "On the Adaptive Real-Time Detection of Fast-Propagating Network Worms". Jaeyeon Jung, Rodolfo A. Milito, and Vern Paxson
[3] "An Inquiry into the Nature and Causes of the Wealth of Internet Miscreants". Jason Franklin,

Vern Paxson, Adrian Perrig, Stefan Savage.

Network Algorithmics - George Varghese (UCSD)
The title above is also the title of George's book (which he was kind enough to give us a copy of the first chapter of!). The term algorithmics here is meant to mean that it's more than just about algorithms - it's about solving the problems like processing bottlenecks in the network using a systems approach along with algorithmic thinking. The key thing that he tried to teach us (since we are mostly used to thinking up algorithms to solve problems) is that algorithms are rarely the best solution in most problems that come up in his field. Usually, a combination of clever ideas, along with some principles that he introduced us to, are much better solutions to most problems. George's style was clearly one of a teacher showing his students the tricks of his trade and his enthusiasm was quite infectious.

In his first lecture, he showed us what he meant by network algorithmics by walking us through a single example problem. Successive solutions enchanted us with new ideas and principles. In the next lecture, he showed us abstract models of hardware that the "software guys" like us can use to be effective in designing systems that address various issues. The third lecture was all about 10 principles that he finds particularly useful as part of a "systems approach" to coming up with efficient solutions to problems. In his final lecture, George taught us how to build a fast router: it turns out to be just a matter of building 4 key components. These solve four problems : longest matching prefix search, packet classification, switchingc and router QoS. He took us from the naive or simplistic solutions for each of these problems, all the way to the state of the art today, some of which were his own ideas.

Wireless Networks - Victor Bahl (MSR – Redmond)
Victor gave a series of three lectures overviewing different "hot" areas of wireless networking research. These lectures were rather non-technical in nature as the intent was just to introduce us to these areas and summarise the main research problems (along with some relevant literature) in them.
The tutorial on Wireless Mesh Networks explored a lot of different areas in the field while also leaving out some other important research results (in, for example, PHY techniques, routing, power control and security). A wireless mesh network is a peer-to-peer multi-hop wireless network where the nodes cooperate with each other in routing packets. Typically, a mesh network is envisaged as a way to make broadband access available to the masses. Wireless links make for cheap last-miles, particularly when deployed in the form of wireless mesh network where the nodes in the community cooperate with each other and bear the cost of maintenance. An advantage here is that a lot of services that have only community-wide relevance (such as local news, billboards etc) need only stay within the community. Mesh networks can also be formed between the devices at home or in an office in order to reduce the "wire mess", while also being cheaper and more convenient. Finally, an application of mesh networks is as an emergency response network in case of a flood, earthquake etc. For a number of reasons, the current IEEE 802.11 protocols are not suited for use in mesh networking. New approaches to MAC, routing, channel assignment etc have been proposed in literature. Victor gave a summary of some of these along with a large set of references to start working in wireless mesh network research.

He then talked about Wireless Network Management. Apparently, it is this hot new area of research that has come out of the headache of maintaining heterogenous large-scale networks. The problem is particularly compounded in the wireless case since wireless channels and RF propogation have an

unpredictable nature. A number of solutions were highlighted. Some of these had been implemented and tested in MSR Redmond by Victor's group and he told us about some of his experiences.

Victor used his final lecture to describe the current efforts in White Space Networking (his term for Cognitive Networks). Much of this talk was about policy matters : white space has been in the tech news for too long for me to say much here. Again, Victor showed us his "systems" bias by explaining the results from some deployments by his group. He also gave a lot of references to follow up. I am yet to dig into it, though.

BGP and Interdomain Routing - Jennifer Rexford (Princeton)
The Internet is actually a collection of a large number of autonomous networks, with cooperation between some, competition between some. Each autonomous network has its own objectives, policies and relationships with other networks. What allows the individual network operators to exert their policy preferences is the Border Gateway Protocol. Jennifer started with the basics of internet routing and address allocation, then took us through an introduction to BGP, the various problems that it tries to solve and how it manages to solve them (or doesn't). It was amply clear that BGP was more like a "hack", with a lot of vulnerabilities and convergence issues, and we aren't even sure it can guarantee to do (correctly) what it is supposed to. Various alternatives to BGP as well as modifications to the existing version were introduced, along with their respective pros and cons. In all, the lecture series was quite interesting and informative. Significant research problems do exist in various aspects of BGP's performance evaluation and improvement. Some that I found particularly interesting include multi-path routing, proof of convergence of BGP, security in control plane (BGP) as well as in data plane etc. I think a mathematical formalization of the "rules of the game" could help us build a reasonably good model of BGP. This would require tools from game theory, optimization theory, graph theory, etc. A reference that goes in this direction is [1] (I think; I haven't read it yet).

A very interesting part of the lecture was an analysis of how Pakistan Telecom brought down YouTube for 2 hours in February 2008 with an attempt at censorship that went terribly (read embarassingly) wrong. It will take a while to explain this, so I'll leave it for later.

[1] BGP stability without global coordination. Jennifer Rexford, Lixan Gou.

Modeling Wireless LAN - Anurag Kumar (IISc)
Anurag Kumar's lectures were much more mathematical in nature than the other talks in the Summer School. He gave us some background on discrete event stochastic processes, starting from the axioms of probability and going all the way to renewal theory.
He then showed us how to model WLANs using these tools. The simplicity of the model (relative to reality, that is) was in stark contrast to the accuracy of the results. The equations predicted the performance of WLAN to a very high degree of accuracy. Unfortunately, this field has been mined a lot for exciting research problems and it might not be worthwhile to look at new problems here. But the tools that were used here, along with an understanding of the modeling process, can probably be applied fruitfully to many other areas of research.

Data Centre Networking - Balaji Prabhakar (Stanford)
In data centres today, the common networking protocols used are Infiniband and FiberChannel. But there is now a move towards Ethernet in this space. The problem is that the kind of reliability that

these protocols provide cannot be provided by TCP over Ethernet, considering the small buffer sizes in Ethernet switches. Ethernet has a lot of non-TCP traffic too, making a native, hardware-based (for high-speed operation) congestion control protocol a necessity. Balaji started with an introduction to data centre networking and how it is different from the usual networking that we are all used to. He then showed us an analytical model (using Mean Field theory) for TCP-RED which matched experimental results quite closely. But, as expected, with large RTTs, TCP-RED becomes highly unstable and shows oscillatory behaviour. Next, he explained how his Quantized Congestion Notification (QCN) congestion Ethernet control protocol works. The key idea here is the "averaging principle" that his group came up with. After each congestion message arrives, the transmit rate sees a multiplicative decrease, as usual. But the transmitter "marks" its original rate as the Target Rate and increases its rate slowly to the target rate in steps. Simulation results show that QCN remains very stable even with long delays in the control loop. It turns out that QCN (as also BIC-TCP) is just the averaging principle applied to TCP. Applying AP to RCP gives a more stable (under large RTTs) version of RCP. So, in general, AP can be thought of as a control theoretic tool that, when applied to some existing control law, leads to greater stability under large delays (at the cost of responsiveness). It is also seen that with shallow buffers (like in Ethernet switches), QCN shows very high throughput even for large delays, as opposed to TCP by reducing the variance in the sending rate.

In the next lecture, Balaji talked about network measurement. He showed us a very neat "counter braid" structure for counting the number of packets from each flow at a router. The key idea is that the packet counts for the flows follow a Pareto Distribution (most flows are "mice" with very few packets while the rest are "elephants" with large numbers of packets, no middle ground). Using multiple hash functions, a reversible mapping from the flows to the counters (memory locations) is made. Note that the mapping is not one to one. Each memory location may be mapped to by multiple flows! Then, a "message passing decoder" can be used to get back the number of packets in the flows. What was really interesting was that they used "Density Evolution" from coding theory to show that there is a threshold value of number of counters (for a given large number of flows) such that it is possible to get arbitrarily small error probability (of not decoding the flow counts correctly) using at least this number of counters, while there is a positive proportion of flows that cannot be decoded correctly with fewer than this threshold number of counters.

After these two topics, Balaji talked about some of his recent work in "Societal Networks" which started out as a hobby after he was frustrated by Bangalore traffic. The idea is to device incentive mechanisms to motivate people not to commute during the high congestion hours, rather than forcing the cars off the roads like the approaches taken in Singapore, Brazil etc. An interesting fact is that the total cost (in terms of fuel and time) of congestion in the US is $80 billion a year and about 3 billion gallons of fuel is wasted in this way every year (that's 6 days' worth of fuel)! Balaji's group consider the "right to congest" as a tradable commodity by charging congestors while paying decongestors. Additionally, an interesting game theoretic result tells us that in games with low stakes, players tend to be more risk seeking. So, the way to pay decongestors is not to give them all equal parts of the small amount that would be collected from the congestors, but instead to give out prizes in a "lottery" system. This idea was implemented by Infosys (a prominent IT firm with about 240 buses plying every day for employees in Bangalore). The incentive mechanism seems to have worked really well. I'm looking forward to reading their publication on this subject.

Some interesting discussion sessions were also part of the Summer School. A common topic that all the speakers talked about was which areas of networking they thought were "hot".
Here's a list :
Data centres, cloud comuting, multi-core, virtualization (of network resources), Web2.0 services, mobile systems (applications, privacy etc), cognitive radios and networking, network neutrality

measurement, home networking (whitespace, cognitive networks, opportunistic networking), network management, sensor networks (distributed signal processing), optimization and game theoretic modeling of protocols, incentive mechanisms to ensure sharing of resources, fountain codes and message passing algorithm for compression in the network, protocols for the Ethernet and network coding.

George Varghese and Ant Rowstron gave us a list of some of the best of conferences in this area. These are :
Networks (SIGCOMM, NSDI, INFOCOMM, CONEXT),
Mobility (MOBICOM, MOBISYS),
Performance (SIGMETRICS, Internet Measurement Conference),
Systems (SOSP, OSDI),
Sensors (SENSYS).

The common tools that could be useful for researhers in the field include :
Operating Systems (Linux Kernel), hardware (NetFPGA), Distributed systems (Emulab, Modelnet, PlanetLab), Simulations (NS-2), Wireless (Orbit-lab).

A rather silly discussion was about whether modeling or prototyping should be done first. Victor Bahl (prototypes) and Anurag Kumar (models) took opposite sides on the issue. None of the arguments were made very well. I think ultimately the system design process is an iterative one, where we move from model to better model, prototype to better prototype. Gaurav Raina (visiting faculty at IITM) pointed out that TCP is an excellent example of both these approaches. TCP was built in 1983 without any theoretical modeling. This worked well from 1983 to 1988, at which point it was on the verge of collapse. Van Jacobson's algorithm (with theoretical justification) saved the Internet at the time and really good models of TCP stability and fairness have come about only recently. It seems like TCP may not be the best answer to most problems in congestion control and these models give us insight into how to modify TCP. Clearly, both models and prototypes have their place in the research process.

There were enlightening discussions on network neutrality, various issues in network security, the difficulties of network measurements (especially latency measurement in microseconds and large-scale topography of the Internet), the changing nature of Internet traffic (volume and content) etc.

Overall, the Summer School was an awesome experience that opened my eyes to many different avenues of research in networking. I now have the task of going through this large volume of literature to think about interesting problems to work on. Looking forward to it!