# Turing Switches - Turing Machines for All-Optical Internet Routing - Part 2

Jon Crowcroft

October 13, 2002

**Abstract**

This is an EPSRC responsive mode Proposal to carry out basic long term research into the architectures for programmable all-optical Internet routers.

We are revisiting some of the fundamental tenets of computer science to carry out this work, and so it is necessarily highly speculative.

Currently, the processing elements in all-electronic routers are typically fairly conventional von-Neumann architecture computers with processors that have large, complex instruction sets (even RISC is relatively complex compared with the actual requirements for packet processing) and Random Access Memory.

As the need for speed increases, first this architecture (see Nick McKeown's excellent talk in reference [6]), and then the classical computing hardware components, and finally, electronics cease to be able to keep up.

At this time, optical device technology is making great strides, and we see the availability of gates, as well as a plethora of invention in providing buffering mechanisms.

However, a critical problem we foresee is the ability to re-program devices for different packet processing functions such as classification and scheduling. This proposal is aimed at researching one direction for adding optical domain programmability.

## 1 Scientific and Technological Rationale and Relevance

### 1.1 Purpose - Turing Switching

The goal of this project is to revisit basic computer science and the architectural design of computers for switching - we refer to our approach as *Turing Switching*[1], as we will argue that by analogy with the original *Turing Machine*, a packet switch control system need not be a von Neumann architecture, or anything closely resembling it.

We wish to provide a programmable optical packet switch, which has the flexibility of a software system, and the performance of an optical system. This is against current trends towards circuit (or at least label switching and wavelength or wave-band) switching.

The motivation is to retain advantages of:

- an evolvable network, due to programmability;
- fully distributed control, with commensurate advantages of faster than RTT (round trip time) time-scale adaption to traffic and outage conditions;
- continuity with Internet models and management.

[1] This term was recently unearthed by Tom Scott `<tscott@vedatel.com>` and reported in a personal communication with the proposer.

## 1.2 History

To date the OSI program and other research programs have rightly concentrated on optical systems to support relatively simple data structures, and opto-electronic systems where programming is needed, where the software (code/execution) resides largely on the electronics and physics side.

We have examined the available publications from the first four phases of the program, and note that the engagement with computer science has been quite modest (around two departments/projects), and feel that there is now a moment to open up the approaches. Thus, the goal here is to engage more fully with computer science, starting from fundamental (and quite mature) principles, to see whether programmability can be applied in a novel way on a pure optical, packet switch.

Packet switches (especially IP routers) have been a major success in the last two decades because of programmability, far more than because of performance. This latter failing is becoming a major obstacle to better Internet performance, but we do not wish to compromise the former in fixing this!

## 1.3 Approach

Packet processing requires a nearly (but not quite) full Turing machine, but it does not need a Von Neumann Architecture (complex CPU(s) and random access store for program and data).

Instead we can start from first principles:

1. revisit Turing machines with multiple tapes and heads, and
2. re-examine at programming models for systems with delay line memory;
3. survey at available current and near future devices (e.g. SOAs, TOAD, etc).
    - For example, a single XOR gate is sufficient to build a general processor: an existence proof at BT's Martlesham research labs was carried out (personal communication from John Midwinter) and elsewhere where ATM VCI lookup and swapping was achieved; more recent work by Blumenthal and Dorren (see references) has gone much further in this direction and has achieved full optical only label switching.
    - fast tunable parallel optical delay lines (c.f. work at Princeton), and other techniques for more flexible program and packet and configuratrion (forwarding table prefix) store.
4. We will look at IP packet processing as an exemplar of data communications problem:

    The typical modern IP router looks at around 40 bytes, updating usually only a couple of fields. The first 20 bytes contain routing and type of service detail including destination address prefix matching, source address lookup for RPM for multicast, and mobile (IP in IP tunnels). The second 20 bytes typically hold information that is used for flow identification, which is used to select finer grain type of service, and is also important for firewall, and traceback functions in the event of security problems (e.g. flash crowd or DDOS).
5. System design should include program language and compilers (like other specialised areas such as VLSI and logic languages and DSP and signal processing languages) - tools would be developed to check designs on a potentially realized hardware platform;

Additional outputs could include:

- a more ideal packet network, seeing how IP must evolve, if it is to fit programmable optical network devices, but retain sensible costs;
- IP packet processing only contains finite iteration (not general recursion), and such programs are *provable* (not just checkable). In fact the programs can typically be transformed into Finite State Automata, although we will concentrate on massive scale instruction re-scheduling as an approach;

- We will also look at how this drives more optical device research with Marconi;
- We can look at hybrids of this with circuit, burst and lambda switched networks such as the new generalized MPLS architecture.

### 1.4 Assumptions

There are some assumptions that we are making about the type of (minimal) optical gates, and their costs, e.g. Semi-conductor Optical Amplifiers, or FTGRs. We will discuss this with photonics and optical device experts.

### 1.5 Evaluation

In the final evaluation of our work, we will know if we have succeeded or failed based on whether an affordable solution in terms of expensive (in power, space and price) gates, versus how simple we can make the packet processing programs without them being very long in instruction length and execution time (an instruction miss or stall cycle could be a delay line loop!). There are a variety of interesting problems caused by resulting process time rather than queueing time jitter, and potential for packet re-ordering (this is not fatal in IP networks)!

For realistic packet header complexity, we evaluate program complexity for: not just longest prefix match, but also DS code-point mapping, and TTL and checksum processing, but also say NAT, port NAT, traceback log/hash, etc. and compare to current approaches.

## 2 Relevance to Beneficiaries

The market in high speed switches and routers is growing, even in the current climate (Cisco's last quarter was still 20% up on the previous year). There are clear UK benefits (e.g. for Marconi) in integrating packet switching with UK strengths in photonics.

The project will be of some interest also in the area of optical backplane design, and the approach includes potential further research areas for other novel computing approaches (e.g. gene sequencing computers, such as the proposed IBM "Blue Gene") where classical von Neumann architectures are sub-optimal in any case.

The area of code re-writing and instruction scheduling approach we are advocating is also currently an important topic in VLIW (very long instruction word) processor design (the Crusoe and next generation of Intel processors are related to this category, and the trend is general). Thus although we are taking an extreme approach, it is clear that some ideas are common.

## 3 Dissemination

Working with Marconi, it is clear how they may be able to exploit the work directly. Also, Cambridge University has a corporate liaison system which is able to spin out work and has a track record here.

Given the architectural and *systems oriented* approach that we are proposing, we would expect to publish work in IEEE Infocom conference or ACM SIGCOMM, as well as the IEEE/ACM Transactions on Networks archival journal.

Later in the research, it is possible that developments might lead to possible publications in European Conference on Optical Communication, and possibly the Journal of Lightwave Technology or IEEE Technology Photonic Letters as appropriate.

Any Intellectual property generated will be protected in the normal manner to allow effective exploitation.

# 4 Work Programme

Typical IP header processing in a conventional processor takes around 100 instructions, using a small percentage of the actual opcodes avaialble in a conventional RISC or CISC Processor (e.g. MIPS in Cisco, Intel in Linux routers)

If we define an processor with a very few opcodes, and a very simple isntruction fetch and execute cycle based on optimum programming, together with instruction scheduler, we can estimate the number of instructions for covnetional IPv4 on an 8 instruction, 1 bit operand, 2 operand, machine on the order of several thousand. This is offset by the speed of the gates (on the order of 5ns) and opcode fetch times.

The main task is to define this architecture in detail, and to develop an assembler, and instruction scheduler for it. We will compare this with FSA approaches too. The output from the work is the architecture, and its performance on typical real internet packet and routing table traces.

A subsidiary task is to look at a variety of possible designs for accelerator/co-processors (e.g. parallel ternary cams etc), which might be built from more exotic optical hardware as it is available. Most importantly, we need a clean architectural system design to integate these with the main processor architecture.

The programme will this naturally develop in the following stages:

| Work Package | Name | Duration |
|---|---|---|
| 1 | Optical Devices Survey | 6 months |
| 2 | Packet Processor and Buffer Design | 6 months |
| 3 | Assembler/Simulator | 6 months |
| 4 | Scheduler | 6 months |
| 5 | Survey of Co-Processor Options | 6 months |
| 6 | Cost/Performance | 6 months |

The output of each stage is a report which will be submitted for publication, culminating in a thesis.

# 5 Management and Resources

The PI will be supervising the work, at the normal level, but also putting in a fraction of his time working on pieces of the problem with the student and with collaborators.

The student will have access to other staff from the University of Cambridge, both from the Computer Laboratory and from the Laboratory for Communications Engineering.

## 5.1 Staff

We are mainly asking for 1 Research Studentship for three years. This is a preliminary investigation, which we hope, if succesful, will lead to further proposals in due time.

## 5.2 Equipment

The simulation/modelling work described above is potentially quite computationally expensive. We would like to use some of the publicly available Internet traces to drive some of the performance work, and this will require very large amounts of storage (particularly as these are ongoing projects creating more storage needs as time goes by) hence we have specified a very high end machine for this work.

A PC with a server-class motherboard (probably Tyan Thunder K7 with two 1.2GHz Athlon MP processors), three four-channel IDE RAID controllers (something like Adaptec ATA RAID 2400A) running RAID 5, and twelve 100GB IDE disks, and a fast SCSI disk as the system/workarea disk.

The massive storage requirements here are also because we want to log many many simulation runs. We also need very fast performance to make sure that we can analyse the output from the simulations in a reasonable time frame.

Hardware costs:

| 100GB disks: £260 each | £3120 |
|---|---|
| 36GB U160 SCSI disk | £435 |
| RAID cards: £325 each | £975 |
| Motherboard (Tyan S2462) | £525 (integrated SCSI, LAN, and VGA) |
| CPUs: £185 each | £370 |
| RAM: 1GB Registered DDR | £160 |
| Server Case: | £400 |
| 550W power supply: £163 | |
| | |
| Total: | £6148 |

A gigabit ethernet card (these motherboards have 64bit PCI, so can drive them properly) - that's about another £175.

Another reasonable choice for motherboard would be the Tyan Thunder HEsl S2567 which takes Pentium-III processors instead of AMD Athlon processors. The downside is that Pentium-III processors are slower than Athlons, and that it takes 133MHz SDRAM as opposed to 266MHz DDR SDRAM. The downside of the S2462 is that the largest DDR Registered DIMMS you can get are 256MB, limiting the Athlon solution to 1GB in total. 133MHz Registered SDRAM is currently available in 1GB DIMMS, so for an extra 650 you could put 4GB of RAM in the system.

We looked for Pentium 4 systems, but there don't appear to be any server motherboards that use them, which is odd. We would not want to use Rambus memory anyway - too expensive for no significant gain over DDR.

### 5.3 Justification of Resources

### 5.4 Collaboration

As the PI has just moved to the Computer Lab, he has only just established links with the new Marconi research staff here. However, these early links promise a lot.

As an ex member of UCL, I has collaborated in the past with people in the photonics group in UCL EE. I hope to continue informally meeting with them, and other members of EPSRC OSI program projects. If possible, we will make these links more formal, since there is also significant Marconi fudned work at UCL EE, and therefore common interest and minimal Intellectual Property problems.

## 6 References

## References


[1] `http://www.darpa/mil/ito/psum2000/H647-0.html` DARPA Sponsored WDM Optical Label Swapping Project, Professor Dan Blumenthal, University of California, Santa Barbara.

[2] "All-Optical Label Swapping Networks and Technologies," D. J. Blumenthal, B. E. Olsson, G. Rossi, T. Dimmick, L. Rau, M. Masanovic, O. A. Lavrova, R. Doshi, O. Jerphagnon, J. E. Bowers, V. Kaman, L. A. Coldren, J. Barton, IEEE, Journal of Lightwave Technology, Special Issue on Optical Networks, 18 (12), pp. 2058-2075, December, 2000.

[3] "The l -Scheduler: A Multiwavelength Scheduling Switch," J. P. Lang, E. Varvarigos, D. J. Blumenthal,IEEE Photonic Technology Letters, 18 (8), 1049-1062,August, 2000.

[4] "All-optical Header Erasure and Penalty-free Rewriting in a Fiber-based high-speed Wavelength Converter," P. Ohlen, B. E. Olsson, D. J. Blumenthal,IEEE Photonic Technology Letters,12 (6), 663-665, June, 2000

[5] "WDM to OTDM Multiplexing using an Ultra-fast All-Optical Wavelength Converter," B. E. Olsson, L. Rau, D. J. Blumenthal, IEEE Photonics Technology Letters. 13 (9), September, (2001)

[6] `http://tiny-tera.stanford.edu/ nickm/talks/"High Performance Routers"`, Talk at IEE, London UK. October 18th, 2001 Professor Nick McKeown, Stanford University (designer of the Cisco 12000 router family, and Tiny Tera Router)

[7] M.T. Hill All-optical flip-flop based on coupled laser diodes Microwave and Optical Technology Letters, vol. 25, no 3, May 2000, pp.157-159

[8] E.C. Mos Optical neural network by use of laser diode longitudinal modes Ph.D. Thesis. Promotoren: prof.ir. G.D. Khoe, prof.dr.ir. W.M.C. van Bokhoven; Copromotor: dr. J.J.H.B. Schleipen Eindhoven, ISBN 90-386-1650-3, 1999, pp. 1-130. [ECO-20 a1]

[9] E.E.E. Frietman, G.D. Khoe, M. Shimoji and R.E. Crosbie Optical backplanes: portal to the 21st century (invited) Proc. First East Asian Conference on Light-wave Systems, Lasers & Optoelectronics (LisLO'99), Kuala Lumpur, Malaysia, 16-18 March 1999, ed. Proc. K.S. Low (chairman), 1999, pp. 19-35. [ECO-20 b4]

[10] E.E.E. Frietman, G.D. Khoe, M. Shimoji and R.E. Crosbie Optoelectronic processing and networking in MPP's: a gate to the future Proc. 1999 Summer Computer Simulation Conference, 11-15 July 1999, ISBN 1-56555-173-7, ed. M.S. Obaidat, A. Nisanci, and B. Sadoun, 1999, pp. 627-633. [ECO-20 b4]

[11] E.E.E. Frietman, F. Zhao and G.D. Khoe A prototype for optical interconnection in massively parallel processing and its physical and optical modelling J. Opt. A: Pure Appl. Opt. 1, 1999, pp. 290-294. [ECO-20 b3]