EoI for a *Network of Excellence*:

Switch Architecture (SWARC)

version 2 – 20 Sep. 2002

Thematic Priority: 1.1.2.ii (Communication, computing and software technologies)
Other Thematic Priority: 1.1.2.iii (Components and microsystems)

Manolis Katevenis
Foundation for Research & Technology – Hellas (FORTH)
ICS-FORTH, Vassilika Vouton, P.O. Box 1385, Heraklion, Crete, GR-711.10 Greece
katevenis@ics.forth.gr ; +30 (810) 39.16.64 (tel), 39.16.61 (fax)

ABSTRACT:
SWARC will address the architecture and design of the switching and routing infrastructure for all kinds of networks and digital systems of the future, including:

- electronic & optical switching and their seamless integration;
- routers and switches for high-performance networks;
- switches for WAN, MAN, LAN, storage area, system area, embedded system networks, (multi-) processor-memory interconnects, and networks-on-a-chip.

SWARC aims to show how to cost-effectively build this vital switching infrastructure for future communication networks, at all of the above scales, striving to unify the concepts and the architectures, and thus to allow the reuse of design in this wide and growing spectrum of technologies and applications.

1. Need & Relevance:

1.1 Switches: a Basic Infrastructure and a Growing Market

The way computers are used has moved from desktop computing to Internet computing. As a consequence, users critically depend on the availability of high-bandwidth IP routers and large servers (e.g., Web search engines), which are usually based on clusters and require fast interconnects among processors, and among them and the storage subsystem.

Communication networks to access those servers are wireless –where necessary for convenience– or wired (copper or fiber) –wherever possible, for high performance. Wired networks are based on switches and routers, which enable high-throughput communication by supporting massive parallelism. Routers form the basic infrastructure for all IP networks. Switches are an essential component of every router, whether in a house, in a small office / home office (SOHO), or in the Internet backbone. To be in a leading position in the networked world of tomorrow, *Europe needs to be highly competitive in switching technology*, cost-performance-wise.

New markets for switches are opening up, currently and in the next few years. LAN's migrated from bus-based to switch-based. Computer I/O is increasingly switch-based (SAN - system / storage area networks). Buses in high performance computing servers are going to be replaced by switches; several standards activities are already under way addressing this trend: RapidIO, PCI-Express (formerly 3GIO), Hypertransport, and others. Embedded systems are increasingly based on large numbers of processors, interconnected through a network. Systems-on-a-chip (SoC) go to networks-on-a-chip to get the performance that buses cannot offer. The volume of these markets will be substantially higher than telco switches; in order to compete in these marketplaces, Europe needs to develop *leading switch technologies.*

1.2 Open Questions and Unifying Concepts

Contemporary switch architectures vary widely, evolve rapidly, and do not yet meet a number of objective goals, especially if one considers the above wide spectrum of application domains. If we use an analogy to the wide spectrum of processor architectures before the mid-eighties, contemporary switch architectures are still in their "pre-RISC" stage: *the "RISC architecture" for switches still remains to be found*, and that is a central task of SWARC.

In addition, we must search to discover the *unifying concepts* for the switches in all of the above scales, from internet WAN's down to the chip-network level. Discovering these will allow reuse of design and great cost savings. SWARC's mandate is to achieve this too.

1.2.1 Need for Lower Cost at Increasing Throughput

Switches and routers, today, especially at the gigabit level, are much more expensive than computers. Optical or electronic switches with a large number of fast ports, appropriate for WDM, are extremely expensive. A fully deployed router with 32 ports at OC-192 (10Gbps) per port will cost around 300 K to 1 M Euro, depending on configuration, options and vendor. By all means, lower-cost switching technology is urgently needed, and this must happen while aggregate throughput keeps climbing at astounding rates. In the same way supercomputers moved from expensive proprietary vector architectures to massive parallelism based on commodity microprocessors, in the future we expect high-performance switches and routers to be based on multistage switch fabrics built out of commodity switches (e.g., switches for PCI-express), thus taking advantage of sales volume to drastically reduce costs.

1.2.2 Need for Integrated Opto-Electronic Switching

Will switches be all-optical or all-electronic? The all-optical router requires two essential pieces of technology: an optical switching element (the optical equivalent of a transistor), and an optical memory --routers require routing tables and buffer memories of many Megabytes. For the optical switch elements, substantial advances have been made over the past two decades, although nowhere near the cost/performance level of CMOS transistors. However, for optical memories, the only practical technology is still a fiber loop, which has limited storage (a few thousand bits) and only serial access. Consequently, optical routers today are all circuit switches, or hybrids using a technique of queuing up packets in a burst to be sent over the circuit (burst switching).

The world has decided on IP, and therefore packet routers will always be needed. If a packet router can be built for a certain speed using electronics only, it will be cheaper than the all-optical switch or router, as long as the above reasons hold. As a result, the two technologies will co-exist for at least the next decade, if not longer. Let us use another analogy: in the early eighties, when optical storage appeared as a commercial alternative to magnetic storage, there were forecasts that the former would totally displace the latter, because optical storage promised so much higher density than what magnetic storage was offering at the time. But of course, magnetic storage kept improving, like optical storage did too. Today, two decades later, both technologies thrive and co-exist, each with its own strengths and weaknesses, each for its own applications.

In conclusion, Europe needs to invest in both technologies, and especially in their seamless integration.

1.2.3 Need for Architectural Improvements

Besides electronic versus optical switching, a number of other central open issues remain to be decided, in switch architecture. How to build a single stage switch with limited ports (up to 64) is fairly well understood today. However, for telco applications higher numbers of ports will be needed, and bus replacement applications will typically require multi-stage arrangements with limited bandwidth links. For large numbers of ports, multistage is the only solution. Today, multistage packet switching is still an unsolved problem when compounded with high utilization, throughput, and Quality of Service (QoS). Other open research questions include fixed size cells versus variable length packets, reserved capacity versus bandwidth-on-demand, flow and congestion control, flow isolation versus aggregation, queueing structures, buffer management, packet dropping versus backpressure, scheduling algorithms and QoS.

In chip-to-chip interconnect, while improvements in integration scale provide designers with a huge number of transistors, the availability of increasingly higher link bandwidth is leading to deeply pipelined links, which significantly complicates flow and congestion control. Also, power consumption and thermal dissipation

are becoming increasingly important, especially for embedded systems. Therefore, new switch architectures that address these constraints are needed.

1.3 SWARC's Mandate

The above discussion established the need and relevance of a coordinated research effort on all aspects of switching and routing. While many companies or projects work on switching or routing among other things, they do so in the context of the particular application sector that each is in --WAN, LAN, server, embedded, or chips. Thus, they fail to seek, find, or exploit the unifying concepts in all of these areas. The same was true with past ESPRIT/ACTS/IST projects. SWARC aims to resolve this scattered nature of research in switching; its mandate will be:

> To address the architecture and design of the switching and routing infrastructure for all kinds of networks and digital systems of the future, including:
>
> - electronic & optical switching and their seamless integration;
> - routers and switches for high-performance networks;
> - switches for WAN, MAN, LAN, storage area, system area, embedded system networks, (multi-)processor-memory interconnects, and networks-on-a-chip.
>
> SWARC aims to show how to cost-effectively build this vital switching infrastructure for future communication networks, at all of the above scales, striving to unify the concepts and the architectures, and thus to allow the reuse of design in this wide and growing spectrum of technologies and applications.

2. Excellence:

Although Europe is not the leader in switching and routing today (USA is first), Europe is in a quite good (second) position. We estimate that the USA has on the order of ten Academic/Research Centers of Excellence[1] and a few dozen companies[2], large or SME (start-up), working in switch architectures. Europe has a considerable mass of resources and expertise in switch architectures, and if these are suitably mobilized and coordinated by the proposed Network of Excellence, they can bring Europe to a leading position.

2.1 Existing European Expertise

An approximate, evolving list of existing European Expertise in switch architecture follows:

2.1.1 Academic Research:

- **FORTH**, Greece (Inst. of Computer Sci.: pioneered round-robin and max-min fairness (IEEE JSAC Oct. 87), weighted round-robin (IEEE JSAC Oct. 91), advanced architectures for backpressure within switching fabrics with the "ATLAS I" switch chip (1995-98) - http://archvlsi.ics.forth.gr/sw_arch/) (Inst. of Electronic Str. & Lasers: exper. in optical gating, 4-wave mixing in semicon. optical amplifiers, mode locked fibre lasers).
- **Univ. Politecnica de Valencia**, Spain (Dept. of Computer Eng.: pioneered the design of deadlock-free adaptive routing for wormhole networks, used today by Cray T3E (Hot Interconnects 97), Compaq Alpha 21364 (IEEE Micro Jan. 02), IBM BlueGene/L e.a., efficient deadlock detection in wormhole networks (HPCA 98), dynamic reconfiguration protocols for clusters/SANs (HPCA 00), fast hardware support for QoS with the MMR router (HPCA 99), congestion control with only local information (IPDPS 00), routing algorithms for irregular topologies (IEEE TPDS Nov 00), efficient routing for InfiniBand (ICPP 01); authored the most popular book on interconnection networks - http://www.gap.upv.es/people/jduato/english.html) (Dept. of Communications: work in optical-switching related IST projects). Cooperating Centers: Universidad de Castilla-La Mancha, Spain; Universidad de Murcia, Spain.

---

1 Stanford, Bell Labs, Washington Univ. at Saint Louis, CMU, etc.

2 Lucent/Agere, PMC-Sierra/Abrizio, Vitesse, IBM, AMCC/MMC, Pluris, Riverstone, Silicon Access, Mellanox, CISCO, Juniper, Avici, SUN, Compaq/DEC, SGI, etc.

- **Politecnico di Torino**, Italy (exp. in switching, routing, networking) - www.tlc-networks.polito.it
- **University College London (UCL)**, UK (optical packet switching) - http://www.ee.ucl.ac.uk/~ong/
- **CERN**, Switzerland (contrib. in 1355-networking, ethernet switching) - http://hsi.web.cern.ch/HSI/dshs/
- **SIMULA Research Lab**, Norway (contrib. in local and dynamic routing reconfiguration in SAN's, load balancing, and irregular topology routing).

2.1.2 Industrial Research:

- **Alcatel** (leading worldwide maker of switches, routers, e.a.) - http://www.alcatel.com/telecom/mbd/keytech
- **Ericsson** (leading worldwide maker of switches, routers, e.a.). See: http://www.ericsson.com/technology
- **IBM Zurich Research Lab**, Switzerland (inventors of the PRIZMA architecture, marketed as "PowerPRS", a scalable switch used in IBM products and –more recently, on the OEM switch market– used / planned by various vendors in almost every electronic switching application) - http://www.zurich.ibm.com/cs/index.html
- **Lucent Technologies/Bell Labs,** EMEA (leading worldwide maker of switches, routers, e.a.) - http://www.lucent.com/ and http://www.bell-labs.com
- **Siemens** (leading worldwide maker of switches, routers, e.a.) - http://w4.siemens.de/ct/en/technologies
- **Marconi,** UK (leading worldwide maker of switches, routers, e.a.) - http://www.marconi.com/html/solutions/bbrs.htm
- **SwitchCore**, Sweden (leading worldwide maker of switch chips) - http://www.switchcore.com/products/technology/
- **Quadrics**, UK (designers of the highest-speed clustering interconnect available today, used by Los Alamos National Lab, USA, and selected as the interconnect for the 30 Teraflops supercomputer (ASCI Q) to be installed at LANL) - http://www.quadrics.com/
- **BATM** Advanced Communications, Israel - http://www.batm.com
- **Nokia**, Finland - http://www.nokia.com/aboutnokia/inbrief/nrc.html

2.2 Network of Excellence Participation

Initial participation in the NoE must be based on proven past excellence, as documented by leading research and publications or cutting-edge prototypes and products, and interest and promise for continued work of the same calliber. When the network starts operating, it will strive to spread excellence throughout Europe, and new members will be gradually admitted, while it will keep reviewing the progress of old members. We are in contact with the organizations listed in section 2.1, above; the following organizations have already declared their interest in participating:

- *Academic Research:* FORTH, Univ. Politecnica de Valencia, Politecnico di Torino**,** CERN, SIMULA Res. Lab.
- *Industrial Research:* Alcatel[3] SEL AG, Ericsson Sweden, IBM Zurich Research Lab, Lucent Technologies/Bell Labs EMEA.

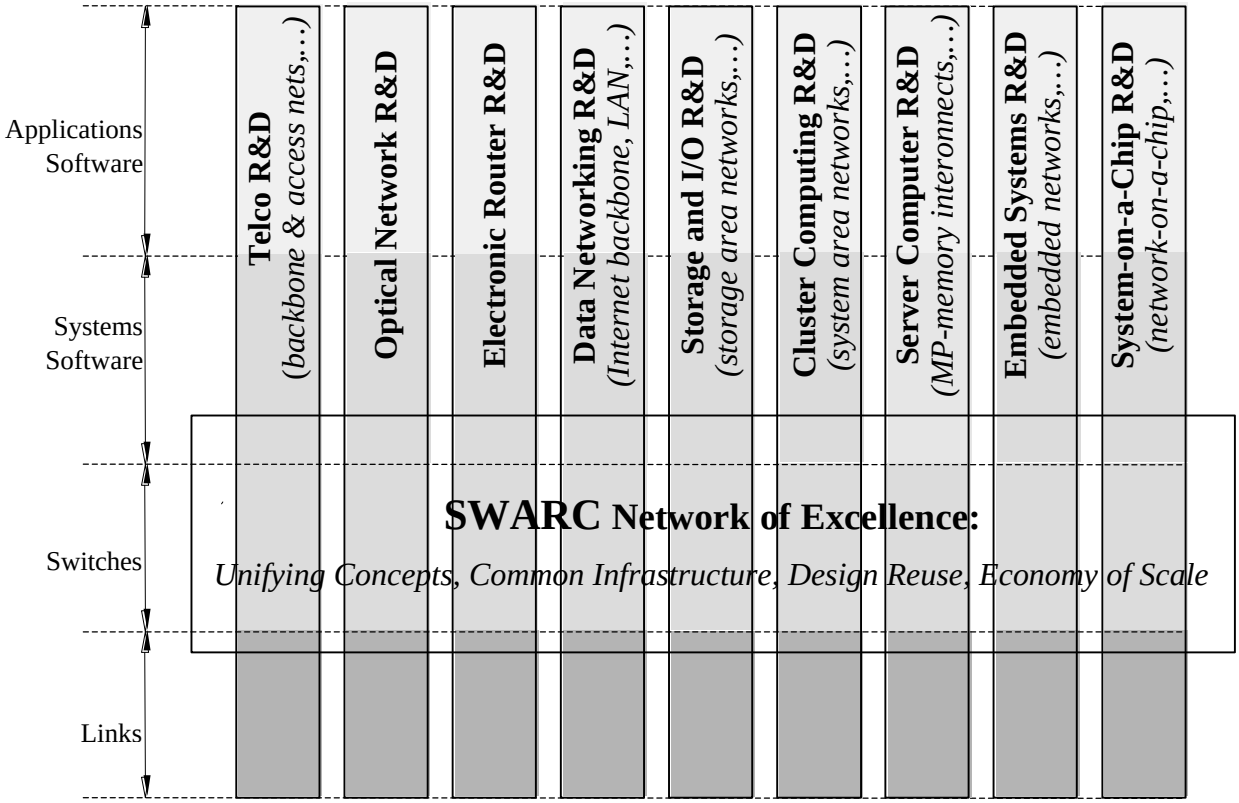
Given the international nature of contemporary research, where openness and exchange of ideas is of vital importance for progress, we believe that this Network of Excellence has to include a few, selected top-ranking participants from the rest of the world. We have contacted the following two, and they are potentially interested to participate, once the details for extra-European participation have been finalized:

- Stanford University, USA - prof. Nick McKeown - http://klamath.stanford.edu/~nickm/
- Bell Laboratories, USA - Dr. D. Stiliadis, Dr. T.V. Lakshman - http://www.bell-labs.com/user/stiliadi/

[3] We refer to the EoI for the IP "Flexible, high quality end-to-end networks (FlexNet)" for interaction and collaboration: for example, in the figure of section 3 (next page), FlexNet deals with several of the vertical bars, while SWARC is the horizontal bar shown.

3. Integration and Structuring Effect:

Many companies, research centers, and consortia, in Europe and elsewhere, as well as past ESPRIT/ACTS/IST projects, have worked or are working on switching or routing. However, they do so in the context of the particular market or scientific sector that each of them is in, as scematically illustrated by the vertical boxes in the following figure.



These existing and emerging areas of application of switches are perceived as separate, independent, and largely unrelated, with different tradeoffs in each. Yet, all of them are switches, and a closer look will reveal intriguing similarities and fundamental concepts in common. As the field of switch architecture matures, little by little, these unifying concepts will start appearing, just like the many and vastly different processor architectures of the seventies (supercomputers, mainframes, minis, micros, high-level-language architectures) later converged to two architectures (microprocessor, DSP), and these will probably soon converge to a single merged organization.

To make this integration of ideas happen in Europe, as the horizontal box shows in the above figure, and to make and keep European companies *competitive* in these evolving and growing markets, a Network of Excellence with a *unifying view* like the one proposed here is needed.