RZ 3087 (#93133) 12/14/98
Computer Science/Mathematics 5 pages

Research Report

A Single-Chip Lossless 16 × 16 Switch Fabric with 28 Gb/s Throughput

Michel Colmant and Ronald P. Luijten

IBM Research Division
Zurich Research Laboratory
8803 Rüschlikon
Switzerland

A Single-Chip Lossless 16 × 16 Switch Fabric with 28 Gb/s Throughput

Michel Colmant and Ronald P. Luijten

IBM Research Division, Zurich Research Laboratory, 8803 Rüschlikon, Switzerland

Abstract

A new switch module was developed based on the shared-memory PRIZMA switch architecture. It provides the highest throughput in the industry for a single-chip lossless switch fabric with integrated control. It employs 440 Mb/s serial I/O technology in a 0.25 μm L_{eff} CMOS process to offer a 16 × 16 switch module with 28.4 Gb/s aggregate throughput. Its scalability allows the deployment of switch fabrics of 400 Gb/s throughput and greater.

Introduction and Architecture

PRIZMA-E is the second-generation switch chip in PRIZMA[1] architecture. This chip features the highest throughput in the industry for a single-chip, shared-memory, lossless packet switch fabric with integrated control. It addresses the ATM (Asynchronous Transfer Mode) and IP (Internet Protocol) switch and router market.

The PRIZMA-E chip provides a switch fabric with 16 input ports and 16 output ports, each running at a throughput of 1.77 Gb/s. It is a self-routing module that integrates data buffering and control to achieve a 28.4 Gb/s aggregate throughput. Two identical chips can be operated in parallel (speed expansion) and thus increase the port speed to 3.52 Gb/s and the throughput to 56.8 Gb/s.

The on-chip queuing memory provides 1024 locations of 20 bytes. Along with the control section, it allows 256 packets of 64 to 80 bytes to be buffered with one chip, and 512 packets with two chips, each one receiving half of the packet. This packet size range accommodates in one switch packet the 53-byte ATM cell and the 64-byte minimum Ethernet packet size.

The architecture is based on the established scalable fixed-length packet switch PRIZMA architecture (1). Whereas other switch architectures aim at providing performance by combining input queuing with a crossbar and scheduler (2), the PRIZMA architecture achieves performance with a self-routing, shared- output queuing structure. Moreover, it can be implemented as a single-chip solution, whereas a crossbar and scheduler approach usually requires two or more different chips. The PRIZMA-E chip architecture is shown in Fig. 1. Incoming packets are buffered in a shared memory, while the storage addresses are processed by the control section and organized into output queues. Packets of each flow are transmitted in the order in which they were received, thus maintaining packet sequence which is an ATM requirement.

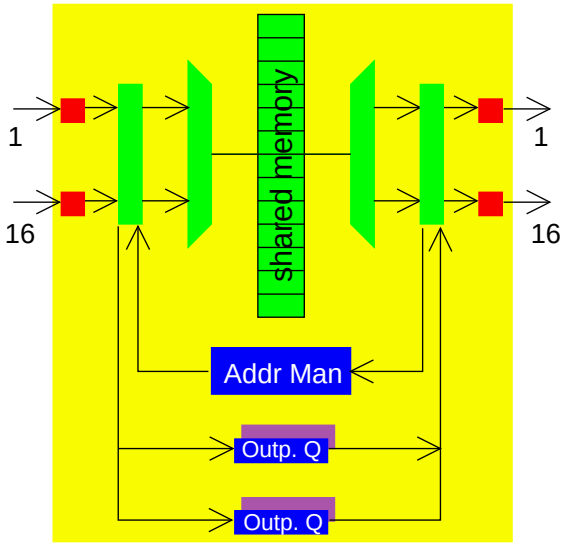

Fig. 1: PRIZMA-E architecture block diagram.

[1] Packetized Routing In Zurich's Modular Approach

A key feature of this architecture is its scalability. To construct large fabrics, multiple identical chips can be combined to increase both the port speed (speed expansion) and the number of ports in a single stage, as described in (1) and (3). This second-generation chip also adds QoS (Quality of Service) support by maintaining traffic priorities for buffer allocation and packet transmission. PRIZMA-E also supports multicast as defined in the original architecture.

Tools and Methodology

This chip was designed in a standard-cell methodology using industry standard tools from synthesis to physical design to reduce costs and time to market. The entire logic was coded using behavioral VHDL and mapped to the standard-cell library using high-level and logic-synthesis tools. The only custom-designed macros were the 440 Mb/s receivers and drivers, which are described in the section below.

I/O Technology

A main challenge in designing switch chips is getting the data in and out of the module, while minimizing the number of module pins. Even though packages with thousands of pins are available, such solutions are prohibitively expensive. Therefore, it is important to optimize the bandwidth utilization of pins. Whereas (4) and (5) propose solutions for Gb/s interfaces, a robust 440 Mb/s serial interface was developed, the drivers and receivers of which use a differential JEDEC (HSTL) electrical interface (6).

Each switch port is composed of four receive and four transmit subports in order to provide 1.77 Gb/s full duplex. Each transmit subport serializes a four-bit stream at 110 MHz into a single-bit stream at 440 MHz, whereas each receive subport deserializes one 440-MHz stream into a four-bit stream at 110 MHz. The chip uses a total of 64 times 440 MHz drivers and 64 times 440 MHz receivers. The driver provides an HSTL differential signal across a 50-Ohm controlled impedance line terminated by 50 Ohm resistors with decoupling capacitor. The high-speed interconnect and low power consumption allowed by this driver are fundamental to the operation of 64 times 440-MHz pairs. The receiver operates on a differential signal centered around 0.75 V, with a 0.5 V swing after circuit card attenuation. It also provides a loss of signal information for diagnostic purposes. Fig. 2 shows high-speed link measurement results in a test setup running at 500 MHz.

All switch fabric modules are required to run on the same reference clock. This implies that only bit-phase alignment has to be performed by the receive macro. A bit-phase alignment control algorithm is shared by all receive macros. This concept of a shared controller greatly reduces the amount of silicon area otherwise required by dedicated phase-alignment macros. This shared-control algorithm provides dynamic phase alignment to compensate for slow variations due to temperature changes.

Whereas initial phase alignment requires a training sequence, dynamic adjustment operates on user packet data to minimize line overhead and achieve maximum throughput.

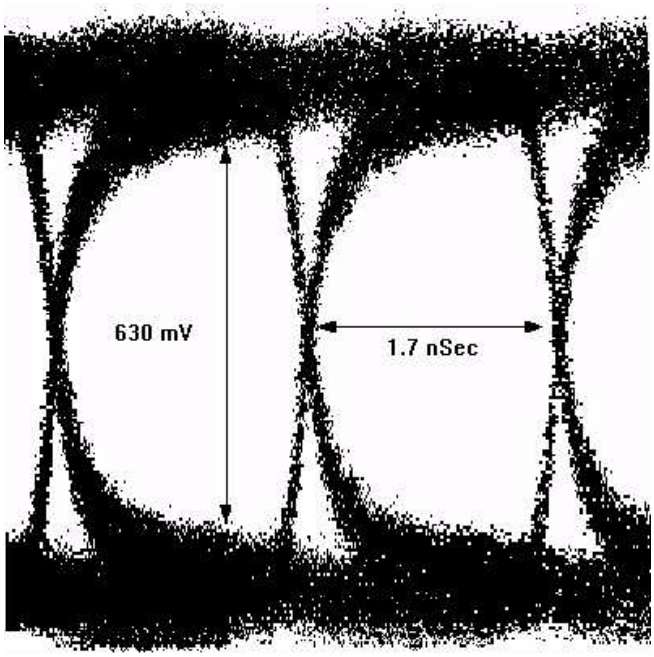

Fig. 2: Eye diagram of pseudo-random data at receiver input.

CMOS Process and Packaging

PRIZMA-E is manufactured in a 0.32-μm CMOS process, the characteristics of which are listed in Table 1. It uses a 624 CCGA (Ceramic Column Grid Array) pin package, which provides dual power supply, 3.3 V for the core logic and 1.5 V for the serial I/O driver and receivers. Table 2 summarizes various chip statistics.

Table 1: CMOS Process Characteristics

Channel Length	0.32 μm drawn, 0.25 μm eff.
Supply Voltage	+3.3 V ±0.3 V
Wiring Levels	4 level of metal
Power Dissipation	0.18 μW/MHz/gate @ 3.3 V
Gate Delay	145 ps (2NAND)
Cell Size	51.84 μm^2
Junction Temp. Range	0ºC to 100ºC

Table 2: Chip Statistics

Chip Size	10.4 mm × 10.4 mm
Transistor Count	3.9 million
Package	624 CCGA dual power supply
Internal Frequency Range	100 MHz to 111.1 MHz
Power Dissipation	1.5 V supply: 1.5 Watt 3.3 V supply: 14 Watt
Total Number of SRAM bits	280,576
Total Number of Standard Cells	1,207,844
Total Number of Latches	45,583

Memory Organization

This implementation of PRIZMA was performed using a shared bus approach. Its original architecture is based on input and output routers as described in (1). This simplification, done to reduce chip size and cost, was achieved by means of on-chip fast SRAM. The shared memory is composed of fourty 128 rows by 32- bit RAMs. Similarly, the 16 output queues are implemented using SRAM.

Chip Layout and Physical Design

Owing to the large image size and wiring density, logic placement and physical design were among the main challenges of this project. Prior to generating the clock trees, placement was performed in multiple iterations to achieve the desired performance. Fig. 3 presents a layout of the chip. The receive macros are placed on the outer left-hand half of the chip, and the transmit macros are placed on the outer right-hand half. This allowed data to flow from left to right. The 40 shared- memory RAMs are organized into two banks and placed in the middle section to facilitate data bus routing and wiring. The sixteen output queue RAMs are placed on the right-hand side close to the output ports.

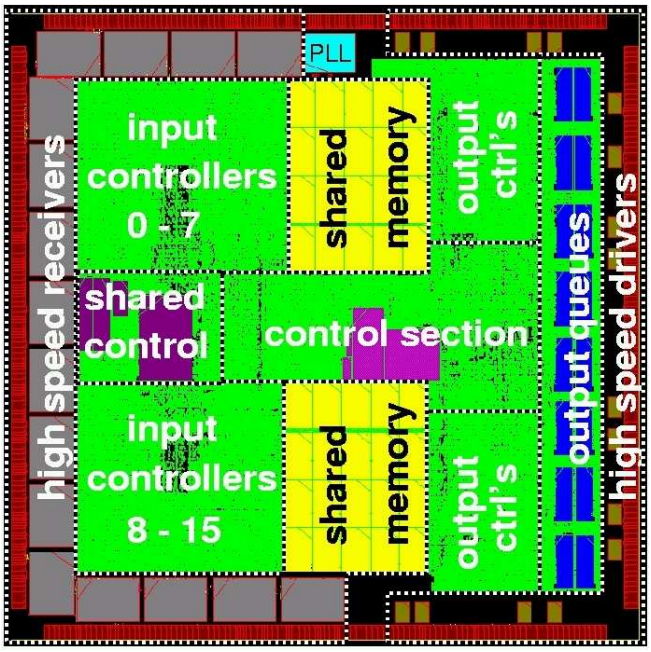

Fig. 3: PRIZMA-E layout.

An internal PLL allows alignment of the clock at the end of the clock distribution tree with a system reference clock. This ensures that a common clock phase is used for all serial macros and that chips in speed expansion can communicate via a synchronous bus. The PLL distributes a 440 MHz clock to all receive and transmit macros via a custom-designed clock tree. This high-speed clock is also divided by four and distributed to the entire core logic, which runs at 110 MHz. The main clock tree distributed this clock to 37,000 LSSD (Level Sensitive Scan Design) latches. Owing to the project schedule and the capacity of the tools used, a clock tree with a skew of 650 ps was generated. The performance of the clock distribution is the main limitation in the maximum internal frequency of the chip.

Testability

Wafer and module test are available via the LSSD test structure. They enable static verification of the internal logic and of embedded macros, such as SRAMs and PLL. The total stuck-at fault test coverage is 99.91%.

The internal logic can further be verified at full clock speed with an internal BIST (Built-In Self-Test) engine. All SRAMs are also testable with a built-in ABIST (Array Built-In Self-Test). Finally, the module is IEEE 1149.1 (JTAG) standard compliant, including the high-speed differential I/Os.

Conclusion

With this new implementation of PRIZMA, we have shown the feasibility of a fully integrated scalable switch chip with an aggregate throughput of 28.4 Gb/s. Receiver and transmitter macros running at 440 Mb/s each were developed to provide the desired high port data rate with a minimum of I/O pins. This allowed the use of commonly available and economical packaging. Using the PRIZMA architecture scaling techniques (1), switch fabrics with an aggregate throughput of 400 Gb/s and greater can be constructed. Future challenges to single-switch chip solutions will be to optimize the data bandwidth per pin (4,5), thus allowing even greater throughputs, and to optimize the internal shared memory structure to provide higher performance.

References

(1) W.E. Denzel, A.P.J. Engbersen, and I. Iliadis, "A flexible shared-buffer switch for ATM at Gb/s rates," *Computer Networks and ISDN Systems,* Vol. 27, pp. 611-624, January 1995.

(2) N. McKeown, M. Izzard, A. Mekkittikul, B. Ellersick, and M. Horowitz, "The Tiny Tera: A Packet Switch Core," *Hot Interconnects V,* Stanford University, August 1996.

(3) PRIZMA press releases and further information available under http://www.lagaude.ibm.com/prizma/.

(4) K. K.-Y. Chang, W. Ellersick, S.-T. Chuang, S. Sidiropoulos, M. Horowitz, and N. McKeown,"A 2 Gb/s Asymmetric Serial Link for High-Bandwidth Packet Switches," *Hot Interconnects VI*, Stanford University, August 1997.

(5) Dally, W., M. Lee, F. An, and J. Poulton, "A Small Low-Power High-Speed Link," *IEEE/LEOS Ninth Workshop on Inter- connections within High-Speed Digital Systems*, May 18, 1998.

(6) Electronic Industries Association / JEDEC STANDARD *"High Speed Transceiver Logic (HSTL). A 1.5 V Output Buffer Supply Voltage Based Interface Standard for Digital Integrated Circuits"*, EIA/JESD8-6, August 1995