8.2 A 51mW 1.6GHz On-Chip Network for Low-Power Heterogeneous SoC Platform

Kangmin Lee, Se-Joong Lee, Sung-Eun Kim, Hye-Mi Choi, Donghyun Kim, Sunyoung Kim, Min-Wuk Lee, Hoi-Jun Yoo

KAIST, Daejeon, Korea

SoC design is composed of two major parts; the design of computing cores and their communication architecture. As the die sizes and the number of subsystems on a chip increase, power consumed by the interconnection structures, including clocks, takes significant portion of the overall power-budget. This calls for techniques to reduce the energy consumed in on-chip communication while satisfying quality of service (QoS) requirements such as bandwidth, latency, or reliability. Recently, on-chip networks (OCN) have been studied actively to address these communication problems [1-4], but their implementations are not energy-efficient so far [1][3]. In this paper, we report successful implementation of a 51mW 1.6GHz hierarchical star-connected on-chip network supporting 11.2GB/s bandwidth with various low-power circuit techniques.

The star-topology guarantees constant and minimum switch hop counts between every communicating IP. However, 1-level flat star-topology [1] as shown in Fig. 8.2.1a results in a number of capacitive global wires that may cause long latency and large power dissipation. Figure 8.2.1b shows a hierarchical star-connected SoC which is composed of several clusters of tightly-connected IPs for their communication locality. Intra-cluster local links provide high-bandwidth with shorter latency and less energy consumption, and inter-cluster global links show higher link utilization by link-sharing. Figure 8.2.2 shows the OCN-based SoC platform applicable to low-power mobile devices [5]. The OCN has two separate networks: forward networks and backward networks that configure the Master-to-Slave path and Slave-to-Master path, respectively [1]. To reduce the area of OCN, 100MHz packets are serialized by Up-Sampler with a 1.6GHz network clock before transmission and then deserialized by Down-Sampler upon arrival. To deserialize a packet without a globally synchronized clock, a strobe signal is transmitted together with the packet. The strobe and the packet experience the same wire-delay without skew. A forward network packet consists of 32b address, 32b data, and 16b header fields while a backward one does not have the address field. The packet header generated by a network interface contains routing information, a type of burst length, a read/write command, an acknowledgment request, and a QoS level.

The global link connecting clusters in the 2nd level star-topology is usually several millimeters long. By using overdrivers [6], clocked sense-amplifiers and twisted differential signaling, packets are transmitted reliably with less than 600mV swing. The sizes of a tranceiver and the overdrive voltage are chosen to obtain a 200mV separation at the receiver end as shown in Fig. 8.2.3. A 5mm global link of 1.6μm wire-pitch can carry a packet at 1.6GHz with 320ps wire-delay and consumes 35pJ/packet (= 0.35pJ/bit). In contrary, a full-swing link consumes up to 3x more power and additional area of repeaters.

A crossbar switch for intra-cluster packets performs buffer-less cut-through switching to minimize packet latency and to reduce its area and power consumption. A round-robin scheduling of the switch ensures fairness and starvation-freedom to OCN. A highly-modulated Mux-Tree based implementation of the round-robin scheduler performs O(logN) scheduling latency, where N is the number of ports in the switch. An n x n crossbar fabric comprises n^2 crosspoint junctions which contain NMOS pass-transistors. In a conventional crossbar fabric, each input driver wastes power to charge two long wires (horizontal and vertical) and 2n transistor-junction-capacitors. Figure 8.2.4 shows a crossbar partial activation technique (CPAT). By splitting the crossbar fabric into 4x4 tiles, input and output wires can be divided into four. A gated input driver at each tile is activated only when the scheduler grants the access to the tile. The output signal does not propagate to other tiles to reduce the power consumption on the vertical wire. A 43% power saving is obtained in a 16x16 crossbar switch fabric with a negligible area overhead.

The on-chip serialization reduces the OCN area significantly [1]. However, it increases signal transitions on a link since it removes the temporal locality between adjacent packets. A proposed serialized low-energy transmission (SILENT) coding technique of Fig. 8.2.5 decreases the number of transitions on a wire by using the temporal locality between packets. The encoder generates '1' only when there is difference between a current packet and a previous packet before it is serialized. The decoder then uses this encoded packet to reconstruct the original input, using its previously stored packet. The power saving of 13.4% is obtained while a multimedia application, a 3D graphics vector in this case [5], is running on the OCN. The additional power consumption due to the encoder/decoder is only 0.4mW. The SILENT coding can also be on and off by software to optimize the power consumption in each application.

A programmable power management unit provides four clocks with PLL; 1.6GHz for the OCN, 800MHz for schedulers, 100MHz for processors, and 50MHz for peripherals. Those clock frequencies are scalable by software for power-mode control and also for optimal operation of each application. The globally asynchronous and locally synchronous (GALS) approach reduces power consumption and simplifies the design with no need for the global clock skew minimization.

The low-power OCN-based SoC platform is implemented using a 0.18μm CMOS technology and its die area takes 25mm^2. The 1.6GHz hierarchical star-connected OCN provides 11.2GB/s aggregated bandwidth (1.6GHz x 8b/link x 7) and consumes 51mW at 1.6V. The OCN is applied to low-power SoC platform integrating heterogeneous intellectual properties (IPs) such as two 32b RISC, an FPGA (8x8 array of logic blocks), an off-chip gateway [1], two 8kB SRAM, and peripheral circuits. Figure 8.2.6 shows the chip micrograph and Fig. 8.2.7 summarizes total power reduction. It is demonstrated that this OCN can provide a platform for the rapid and successful implementation of portable multimedia SoC [5].

References:

[1] Se-Joong Lee et al., "An 800MHz Star-Connected On-Chip Network for Application to Systems on a Chip," *ISSCC Dig. Tech. Papers*, pp. 468-469, Feb. 2003

[2] M. Sgroi et al., "Addressing the System-on-a-Chip Interconnection Woes Through Communication-Based Design," *Proc. of the DAC*, pp. 667-672, Jun. 2001

[3] Michael B. Taylor et al., "A 16-Issue Multiple-Program-Counter Microprocessor with Point-to-Point Scalar Operand Network," *ISSCC Dig. of Tech. Papers*, pp. 170-171, Feb. 2003

[4] Kangmin Lee et al., "A Distributed On-Chip Crossbar Switch Scheduler for On-Chip Networks," *Proc. of CICC*, pp. 671-674, May 2003

[5] Ramchan Woo et al., "A 210mW Graphics LSI Implementing Full 3D Pipeline with 264Mtexels/s Texturing for Mobile Multimedia Applications," *ISSCC Dig. Tech. Papers*, pp. 44-45, Feb. 2003

[6] Ron Ho et al., "Efficient On-Chip Global Interconnects," *Symp. of VLSI Circuits Dig.*, pp. 271-274, Jun. 2003

 0-7803-8267-6/04

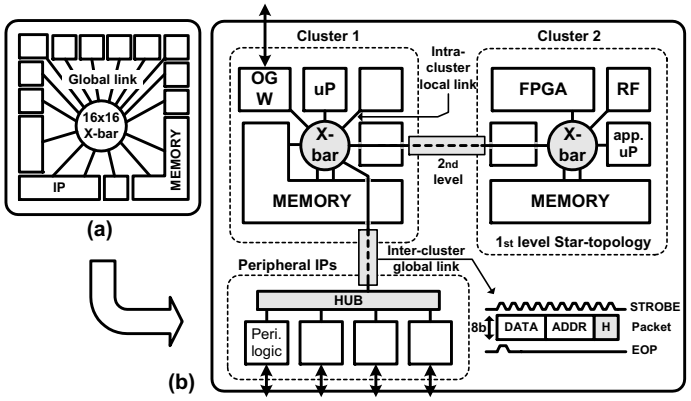

Figure 8.2.1: (a) 1-level flat star-topology and (b) 2-level hierarchical star-topology.

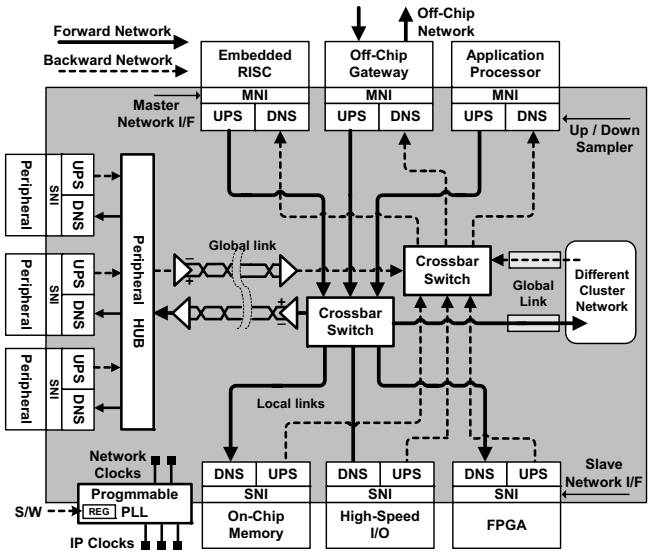

Figure 8.2.2: On-Chip Network based SoC platform.

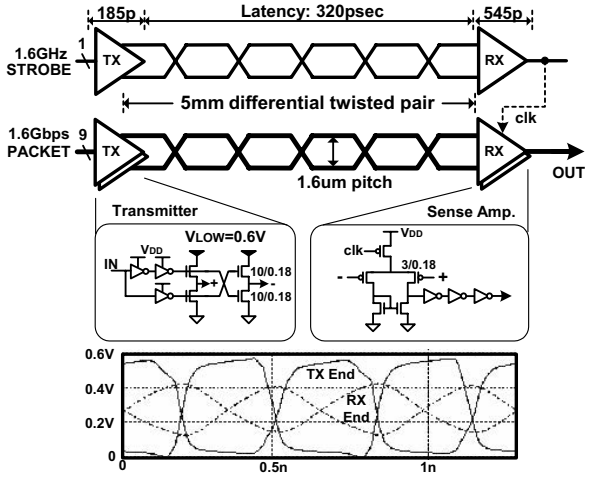

Figure 8.2.3: Low-swing signaling on a global link, Tranceiver and STROBE signal waveforms.

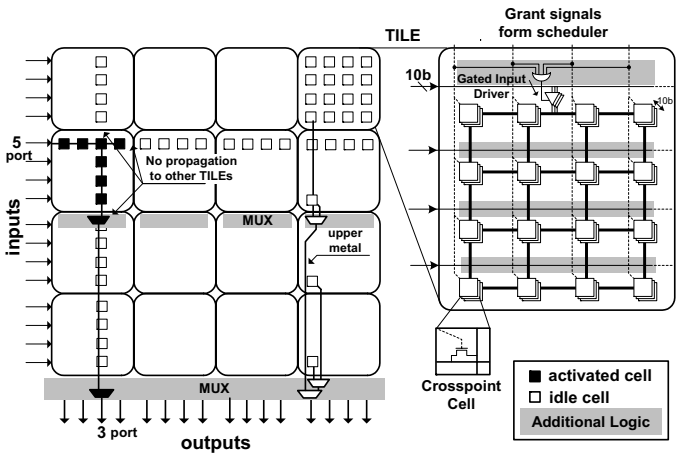

Figure 8.2.4: Crossbar partial activation technique.

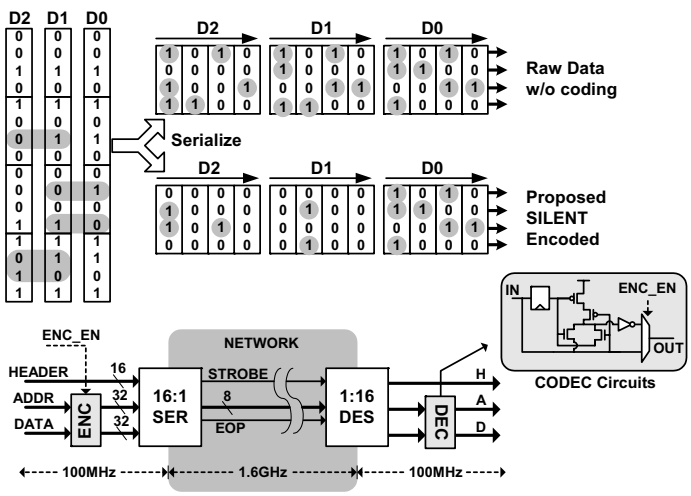

Figure 8.2.5: Serialized low-energy transmission coding.

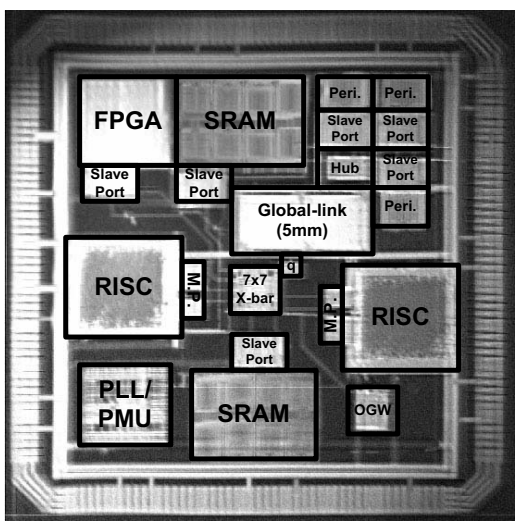

Figure 8.2.6: Die micrograph.

 0-7803-8267-6/04

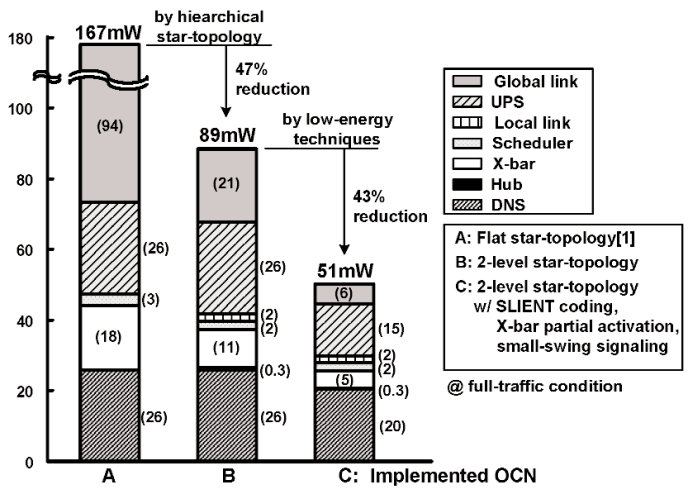

Figure 8.2.7: Overall power reduction.

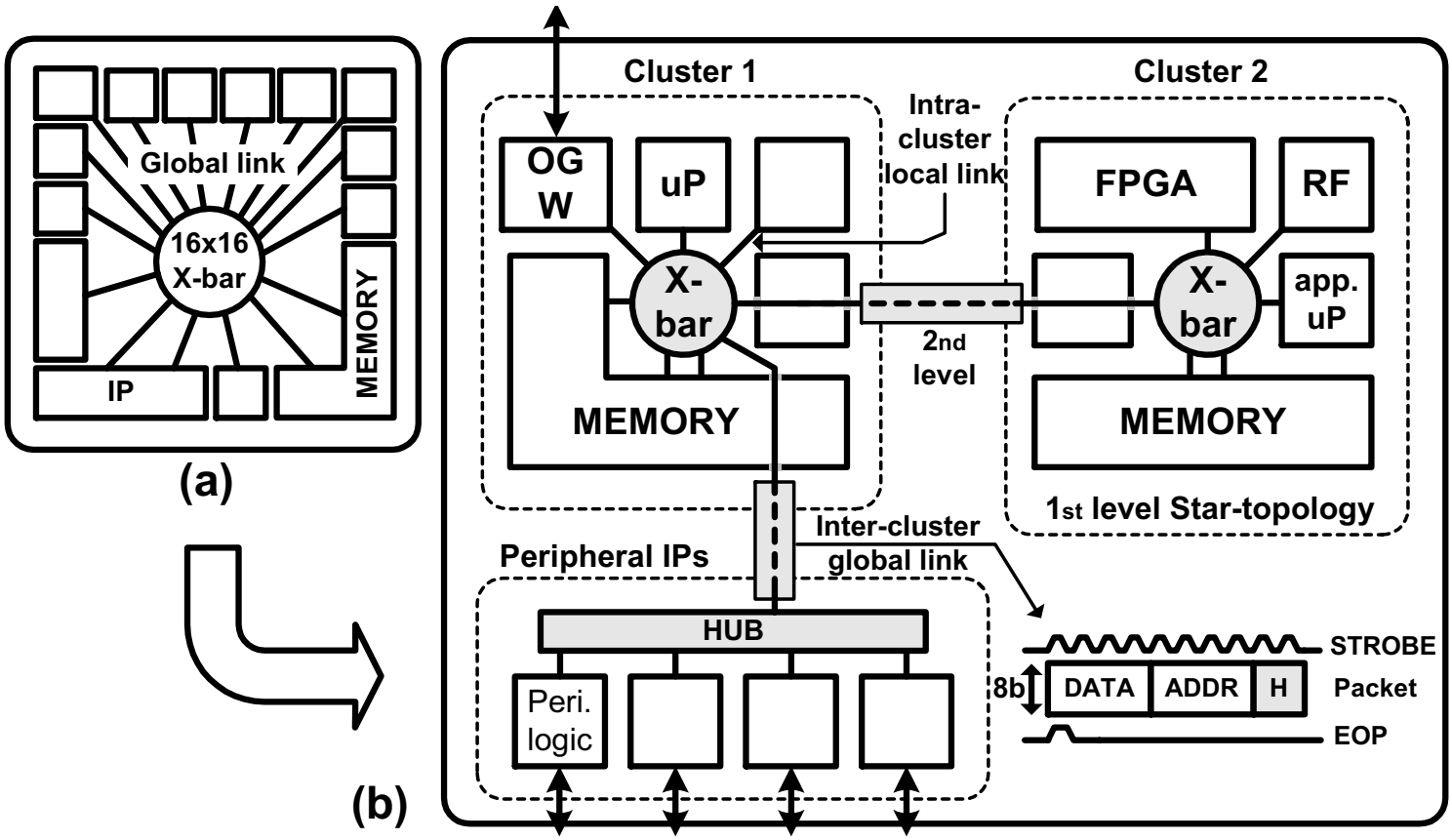

Figure 8.2.1: (a) 1-level flat star-topology and (b) 2-level hierarchical star-topology.

0-7803-8267-6/04

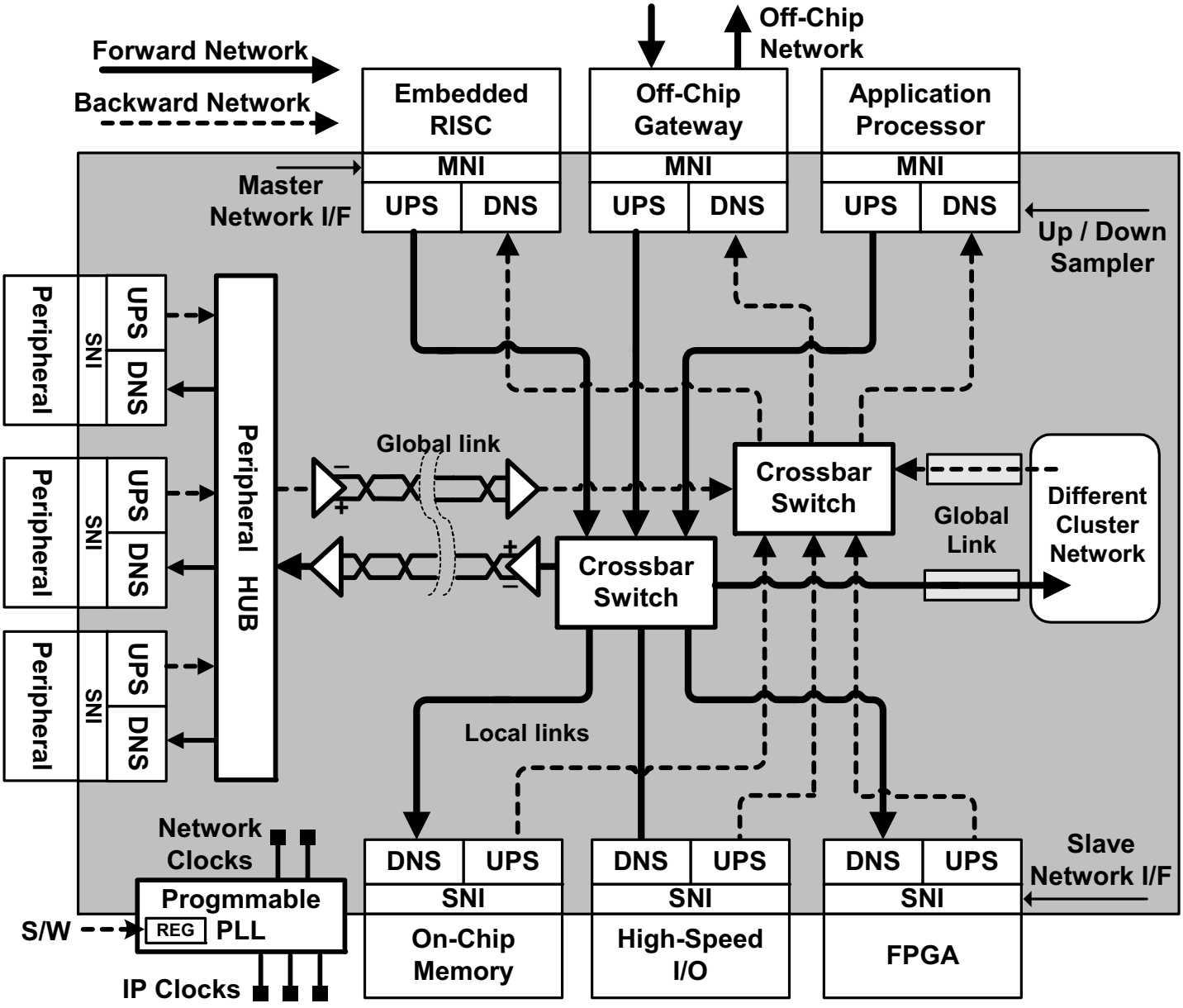

Figure 8.2.2: On-Chip Network based SoC platform.

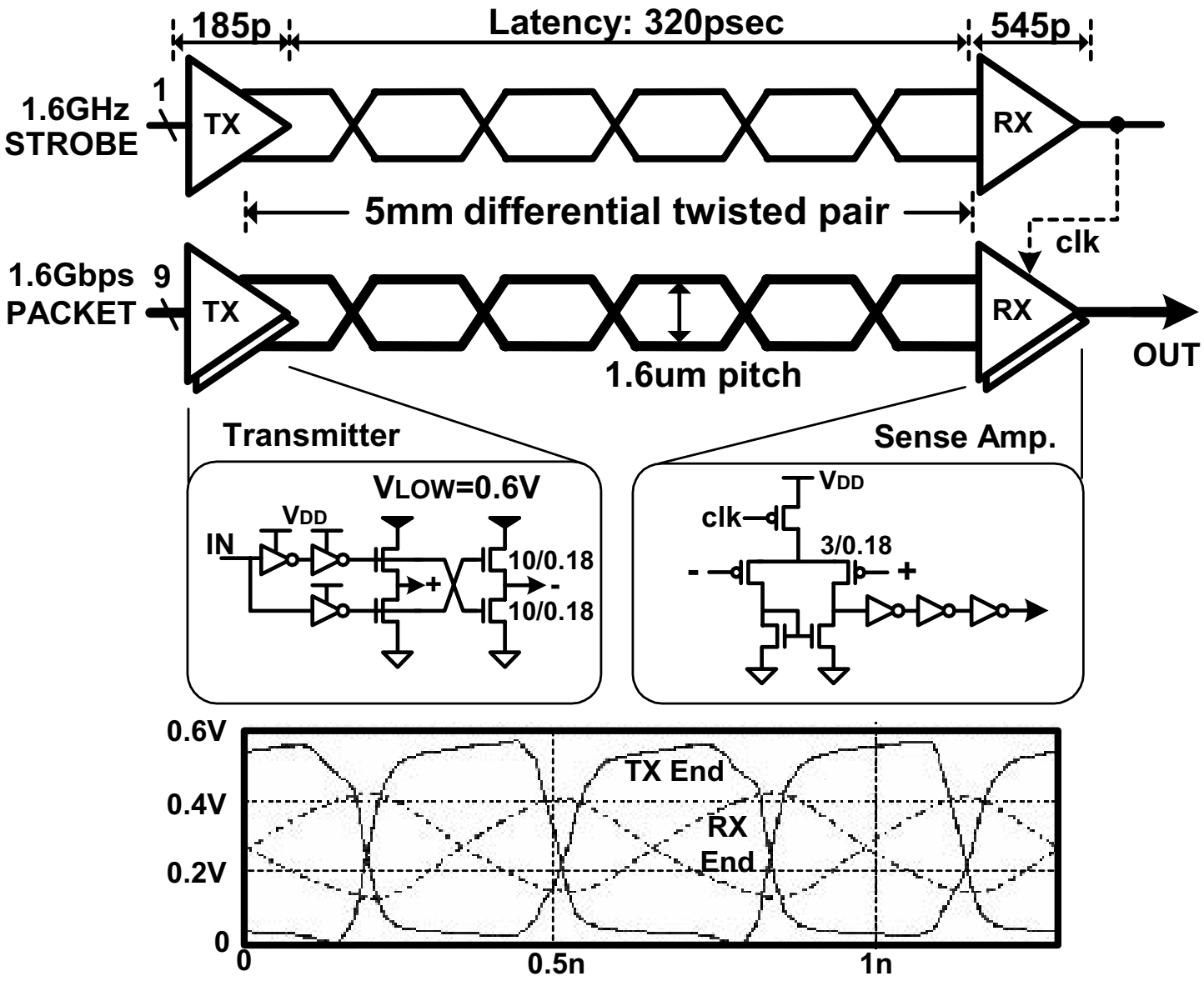

Figure 8.2.3: Low-swing signaling on a global link, Tranceiver and STROBE signal waveforms.

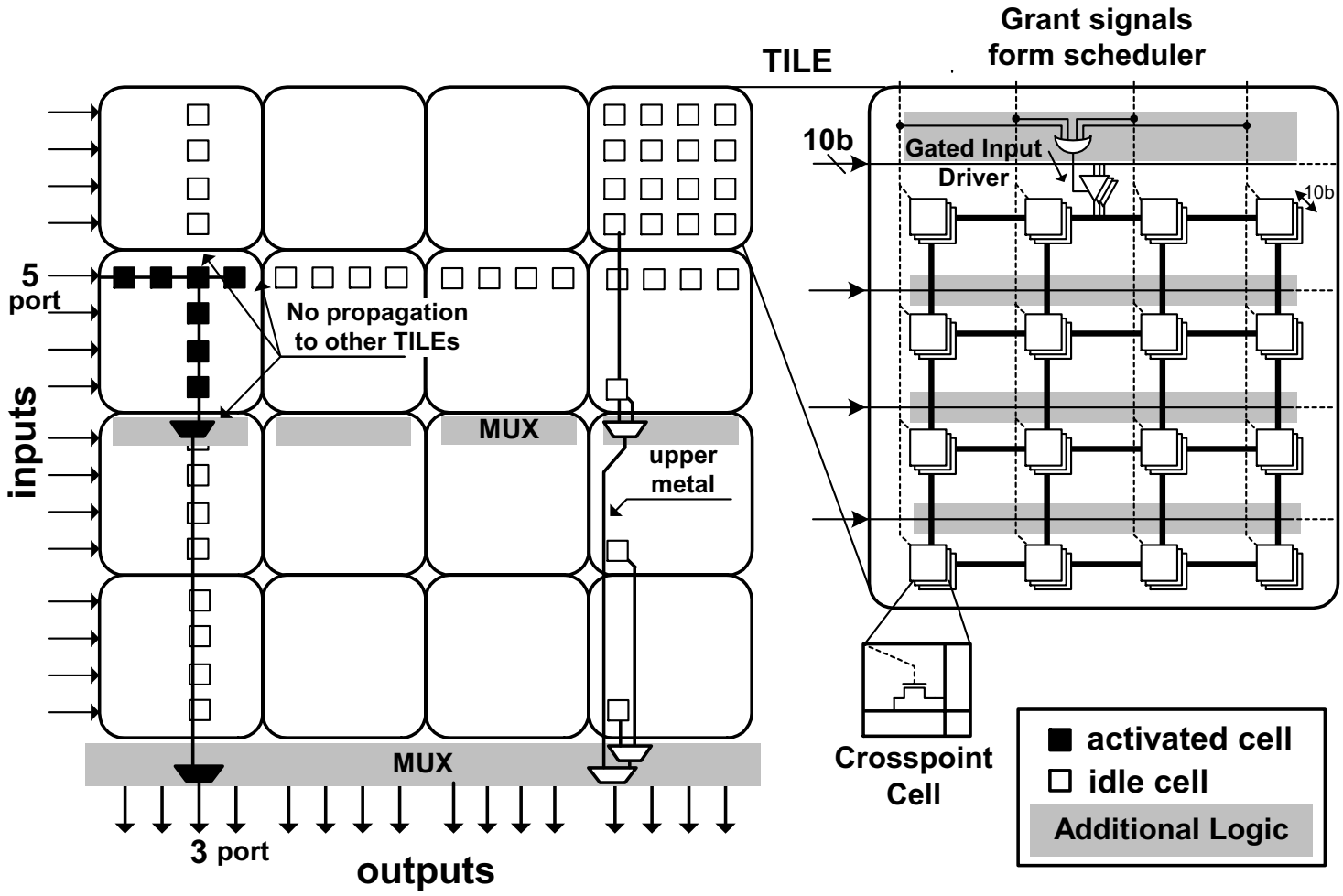

Figure 8.2.4: Crossbar partial activation technique.

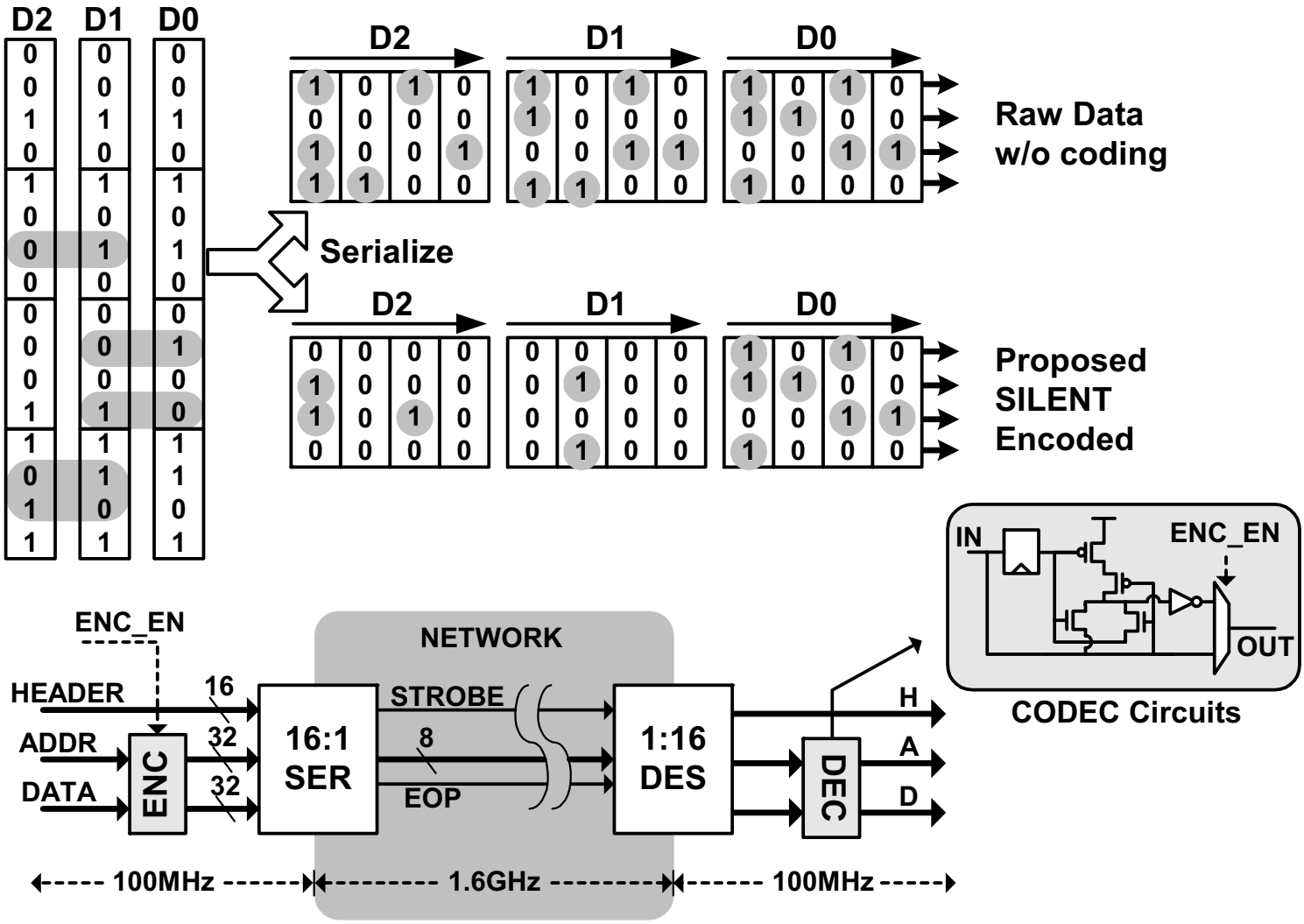

Figure 8.2.5: Serialized low-energy transmission coding.

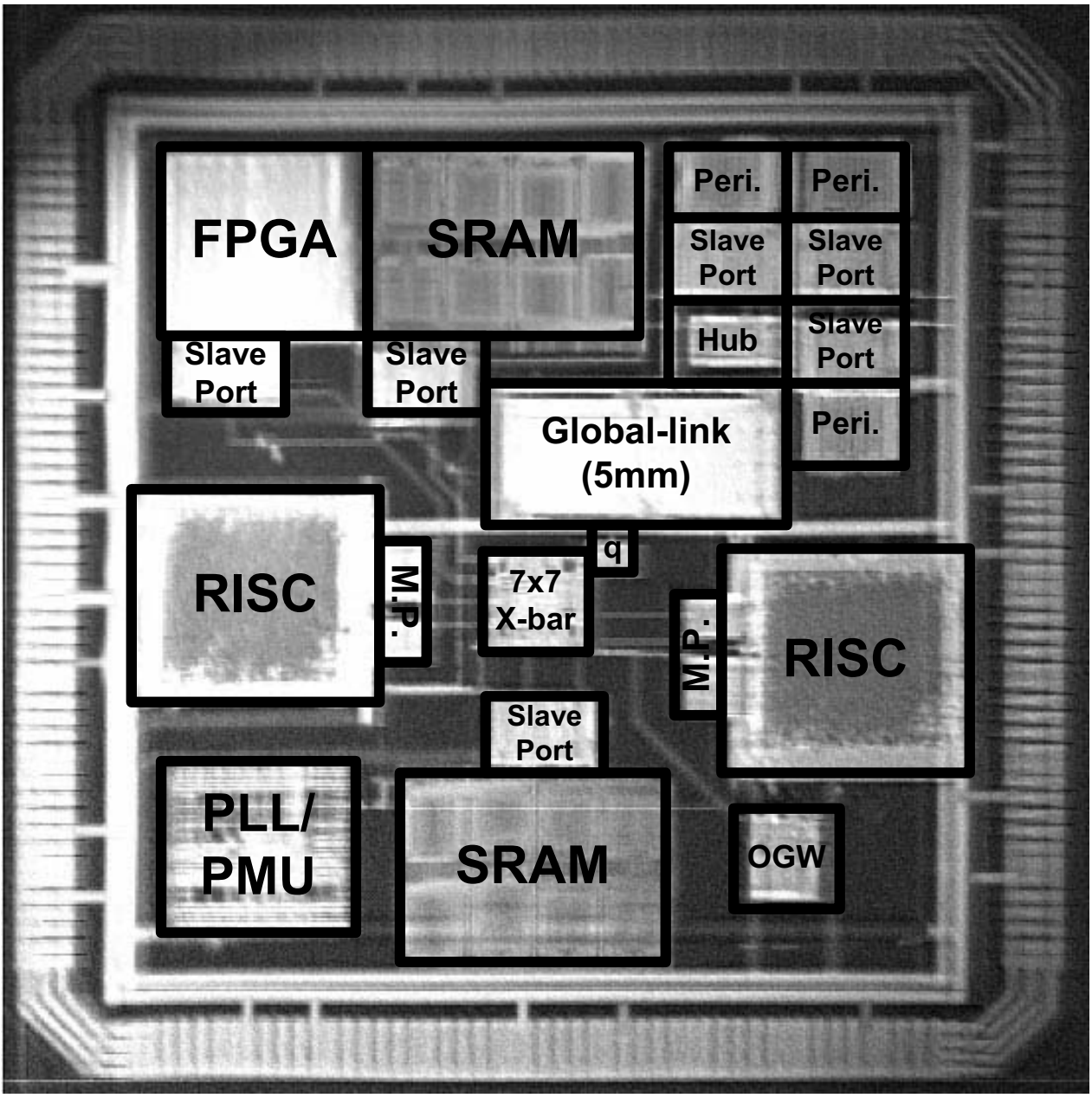

Figure 8.2.6: Die micrograph.

0-7803-8267-6/04

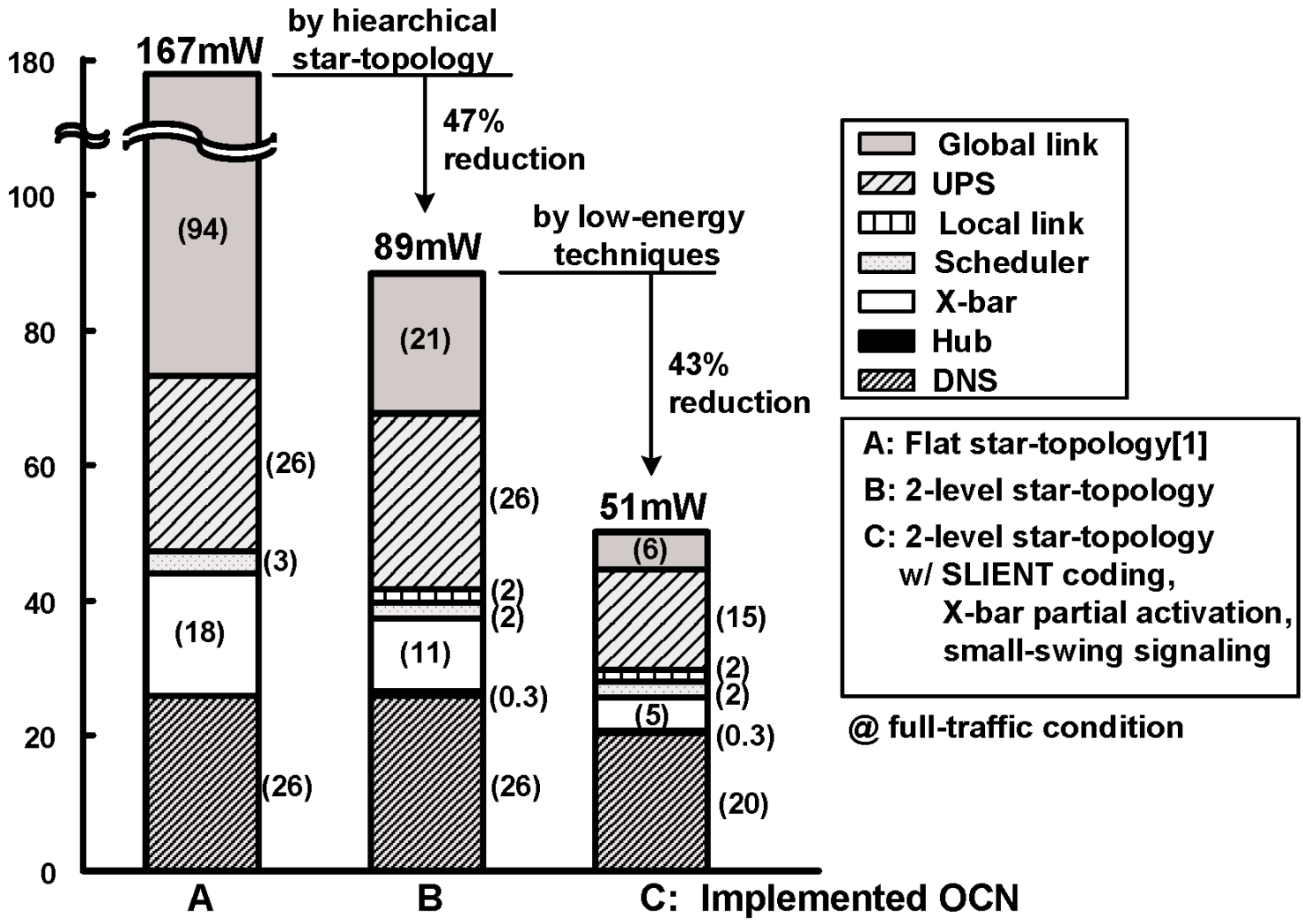

Figure 8.2.7: Overall power reduction.

0-7803-8267-6/04