4.1 22nm Next-Generation IBM System z Microprocessor

James Warnock[1], Brian Curran[2], John Badar[3], Gregory Fredeman[2], Donald Plass[2], Yuen Chan[2], Sean Carey[2], Gerard Salem[4], Friedrich Schroeder[5], Frank Malgioglio[2], Guenter Mayer[5], Christopher Berry[2], Michael Wood[2], Yiu-Hing Chan[2], Mark Mayo[2], John Isakson[3], Charudhattan Nagarajan[6], Tobias Werner[5], Leon Sigal[7], Ricardo Nigaglioni[3], Mark Cichanowski[3], Jeffrey Zitz[8], Matthew Ziegler[7], Tim Bronson[3], Gerald Strevig[3], Daniel Dreps[3], Ruchir Puri[7], Douglas Malone[2], Dieter Wendel[5], Pak-Kin Mak[2], Michael Blake[2]

[1]IBM Systems and Technology, Yorktown Heights, NY,
[2]IBM Systems and Technology, Poughkeepsie, NY,
[3]IBM Systems and Technology, Austin, TX,
[4]IBM Systems and Technology, Williston, VT,
[5]IBM Systems and Technology, Boeblingen, Germany,
[6]IBM Systems and Technology, Bangalore, India,
[7]IBM Research, Yorktown Heights, NY,
[8]IBM Systems and Technology, Hopewell Junction, NY

The next-generation System z design introduces a new microprocessor chip (CP) and a system controller chip (SC) aimed at providing a substantial boost to maximum system capacity and performance compared to the previous zEC12 design in 32nm [1,2]. As shown in the die photo, the CP chip includes 8 high-frequency processor cores, 64MB of eDRAM L3 cache, interface IOs ("XBUS") to connect to two other processor chips and the L4 cache chip, along with memory interfaces, 2 PCIe Gen3 interfaces, and an I/O bus controller (GX). The design is implemented on a 678 mm^2 die with 4.0 billion transistors and 17 levels of metal interconnect in IBM's high-performance 22nm high-κ CMOS SOI technology [3]. The SC chip is also a 678 mm^2 die, with 7.1 billion transistors, running at half the clock frequency of the CP chip, in the same 22nm technology, but with 15 levels of metal. It provides 480 MB of eDRAM L4 cache, an increase of more than 2× from zEC12 [1,2], and contains an 18 MB eDRAM L4 directory, along with multi-processor cache control/coherency logic to manage inter-processor and system-level communications. Both the CP and SC chips incorporate significant logical, physical, and electrical design innovations.

Systems are built from configurable nodes of tightly-coupled CP and SC chips, each packaged on single-chip modules (Fig. 4.1.1). This structure provides improved flexibility and modularity compared to the multi-chip modules used previously. All high-speed node-to-node and drawer-to-drawer communication is through the SC chip using micro-controllers to manage the flow. Each SC chip contains over 440 of these micro controllers along with a series of wide multiplexers to manage the traffic. Both the CP and SC chips support high levels of I/O bandwidth, with about 5Tb/s total bandwidth for each CP or SC chip, running at speeds of up to 5Gb/s (single-ended) and 9.6Gb/s (differential).

The CP chip adopted a unique floorplan configuration, driven by the width of the cores, which were too wide to fit four across on the die. This floorplan created significant logical and physical complexities in the L3 design, but careful engineering prevented these issues from having any meaningful impact on latency or bandwidth of the L3. The entire L3 and all 8 cores are covered with a single large "mega-mesh" clock domain, maximizing on-chip bus bandwidth. The unified mega-mesh design enables double-pumping of many on-chip buses for wider effective bandwidth, and eliminates any mesh-to-mesh timing margins in critical core-to-L3 timing paths.

The CP processor core design, shown in Fig. 4.1.2, improves upon the zEC12 processor [4] with two vector execution units, significantly higher instruction-per-cycle throughput, and a new SMT2 micro-architecture supporting simultaneous execution of two threads. The microprocessor core features a wide superscalar, out-of-order pipeline that can sustain an instruction fetch, decode, dispatch and completion rate of six CISC instructions per cycle. The instruction execution path is predicted by multi-level branch direction and target prediction logic. Complex (CISC) instructions are cracked at decode into two or more simpler RISC micro-ops. Instructions and micro-ops are issued out of program order from an instruction issue queue to multiple RISC-like execution units. The super-scalar design can sustain an issue and execution rate of ten instructions or micro-ops per cycle: two instructions of load/store type, four fixed-point (integer) instructions, two floating point or vector instructions and two branch instructions. These advances yield large performance gains in legacy online transaction processing and business analytics workloads.

Another significant innovation is the use of eDRAM more pervasively across the chip, not only for the L3 cache and memory control unit (MCU), but also within the processor core itself, for the L2 and BTB (Branch Target Buffer) caches. In the case of the L2, eDRAM enabled a doubling of the cache size to (2+2) MB (Instruction + Data). Measured in terms of processor clock cycles, the average L2 latency was kept the same as for the previous product, providing a significant system performance boost. To provide high bandwidth and reduced latency, a dual-bank interleaved design was chosen with a new two-level bitline hierarchy, shorter signal lines, and a high-speed late select path (LATESEL to DOUT), as shown in Fig. 4.1.3. The new local bitline sense amplifier uses a lightly loaded power-gated inverter, which is singled-ended and avoids the overhead of a reference cell. The primary sense amp outputs are connected directly to the 8-way read data mux that drives the data line (RDL) to the I/O block, thereby eliminating the delay of an intermediate subarray bitline and secondary sense amp. Critical to the L2 application is not only the sub-ns overall latency that was achieved, but also the late select access time that limits the response time from a directory hit. The one-hot LATESEL inputs bypass input latches and fan out to the subarray read data muxes to provide a rapid response. For the BTB application, an even higher bandwidth is provided by page mode reads, whereby the outputs of 8 local sense amps are read serially by sequencing the 8 LATESEL inputs every 2 clocks.

The System z CP also featured several high-performance SRAM innovations. Fast read latency to support single-cycle access was essential for all core array applications such as the L1 caches and the Directory/TLB look-up function. Selective threshold voltage tuning in conjunction with highly optimized dynamic circuits, custom tapered-wire solutions, and skewed buffers were utilized to reduce critical path delay. Array redundancy was eliminated to speed up access time wherever the yield impact was deemed to be acceptable. This was crucial for the D-, I-cache, and Set Predict arrays to allow the timing closure of the Load Store and Instruction Cache units. In addition to speed optimization, a major design focus of the core SRAMs was power take down. Fig. 4.1.4 shows a novel bit-column R/W circuit (Local Sense/Write) with a 2-to-1 bit selection function to minimize read global bitline (GBL) and bit write control (WRT) switching power. Relative to a conventional design using a global bit decode scheme, the new Local Sense/Write topology cuts down the number of GBL and WRT lines by half, thus saving column metal usage, circuit area, and power. This design is applied to the L2 Directory and Branch Prediction arrays, lowering macro read/write power dissipation by ~25%.

As with previous System z products, reliability was a key design focus. New methodologies were developed to analyze the thermal aspects of the design at a variety of length scales from the gate level, all the way up to the chip level. To avoid micro hot-spots at the individual gate level, caused by device self-heating, high switching factor nets were identified during functional simulation. Gates driving these nets had their maximum output load capacitances reduced, and were spaced apart from other gates driving such nets in order to avoid excessive heating. In addition the design was broken into small tiles, with total current through low-level power vias calculated, looking for local regions of high power density. Power dissipation was then rolled up to the chip level, for a detailed thermal analysis including package and system effects. The results of such an analysis are shown in Fig. 4.1.5, where the cores are clearly visible given their elevated thermal profile.

Fig. 4.1.6 shows the measured CP chip performance, plotting minimum operating voltage at fixed frequency (5.2GHz) as a function of relative process speed. The chip functional exerciser uses all cores, along with the L3 and ex-core logic, and matches closely the frequency capability measured by LBIST. These observations validate the ability of the design to operate in the 5GHz regime, which, along with the micro-architectural enhancements, provides significant system performance improvement for this next-generation design.

The authors wish to thank the whole System z team and the IBM technology development and manufacturing teams for their contributions to the success of this project.

References:

[1] J. Warnock, *et al.*, "5.5GHz System z Microprocessor and Multi-Chip Module," *ISSCC Dig. Tech. Papers*, pp. 46-47, Feb. 2013.
[2] J. Warnock, *et al.*, "Circuit and Physical Design of the zEnterprise™ EC12 Microprocessor Chips and Multi-Chip Module," *IEEE J. Solid-State Circuits*, vol. 49, no. 1, pp. 9-18, 2014.
[3] S. Narasimha, *et al.*, "22nm High-Performance SOI Technology Featuring Dual-Embedded Stressors, Epi-Plate High-K Deep-Trench Embedded DRAM and Self-Aligned Via 15LM BEOL," *IEDM Dig. Tech. Papers*, p. 3.3.1-3.3.4, 2012.
[4] C.-L. Shum, "IBM zNext: the 3rd Generation High Frequency Microprocessor Chip," *Hot Chips*, 2012.

 978-1-4799-6224-2/15/$31.00

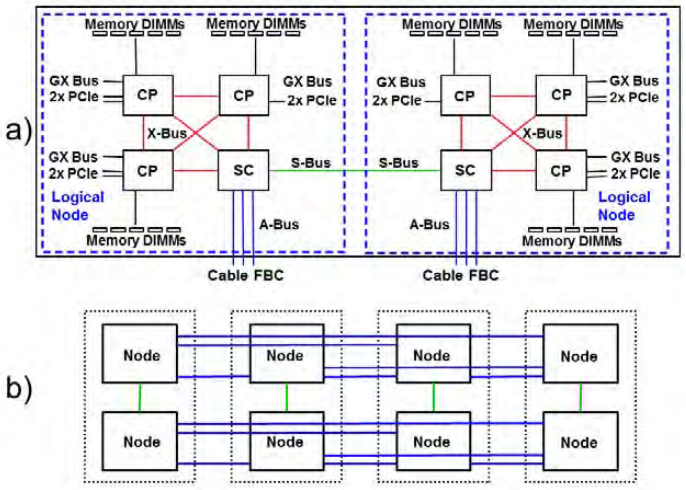

Figure 4.1.1: System structure, for maximum size configuration. a) Drawer, with 2 nodes. b) 4-drawer system.

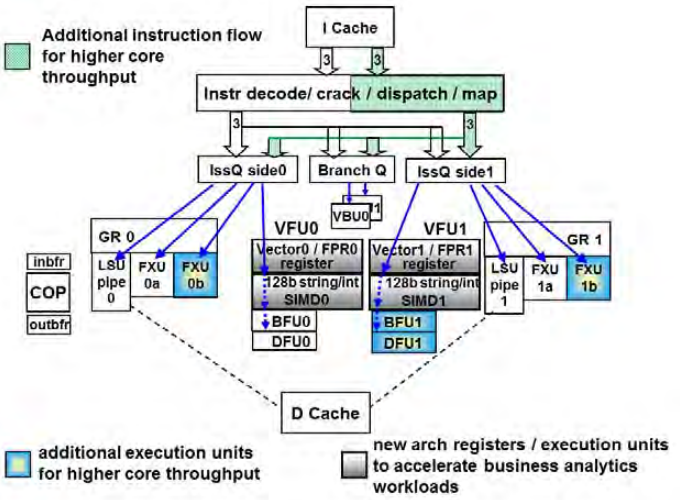

Figure 4.1.2: Processor core pipeline, with comparison to zEC12 design.

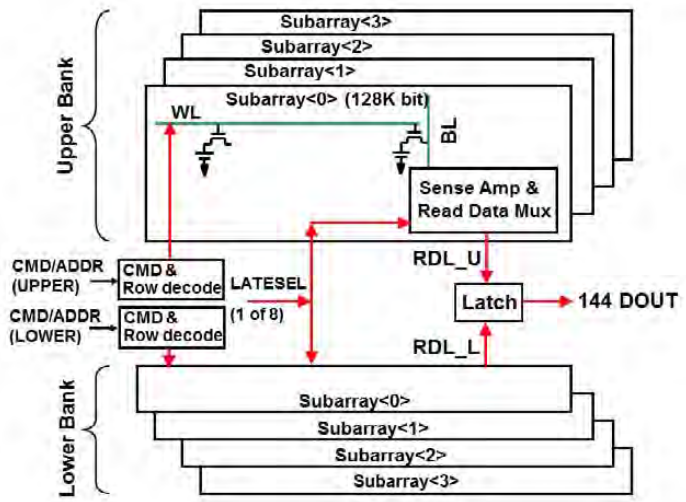

Figure 4.1.3: eDRAM structure for L2. Each cache (L2I or L2D) contains 16 of the above macros.

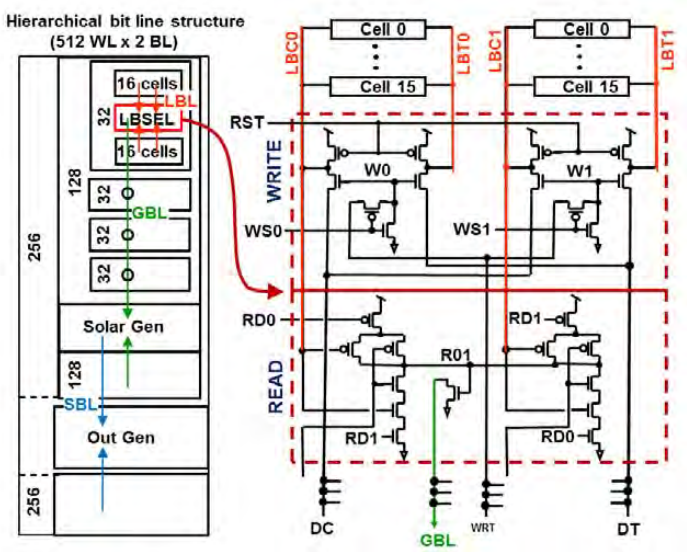

Figure 4.1.4: High-performance SRAM structure.

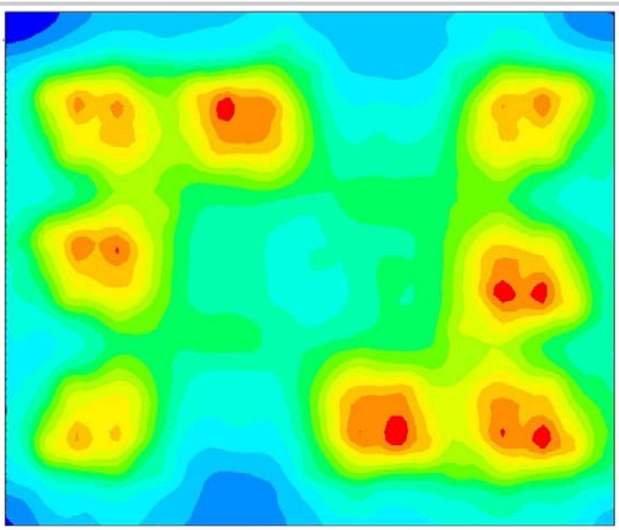

Figure 4.1.5: Chip thermal analysis with a high-power workload (above TDP). Minimum-to-maximum temperature differential is about 27°C.

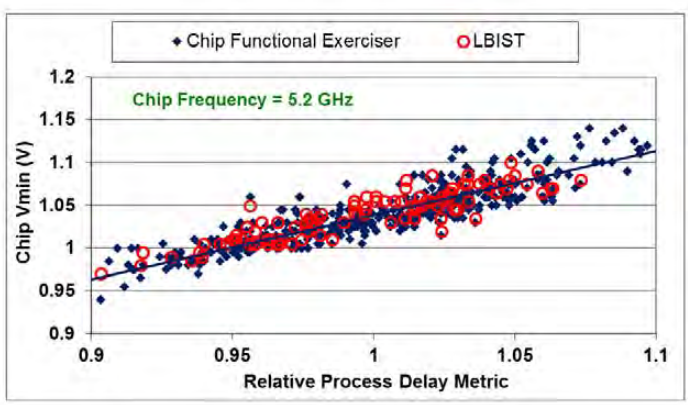

Figure 4.1.6: High-frequency chip Vmin.

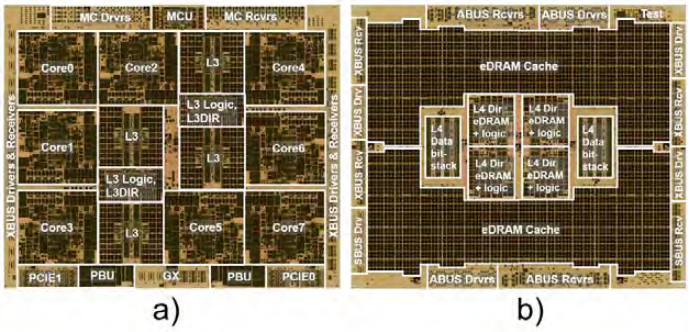

Figure 4.1.7: Die photos. a) CP chip. b) SC chip.

978-1-4799-6224-2/15/$31.00