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 (CP) and a system controller chip (SC) aimed at the previous generation design. The CP chip adopts a unique floorplan configuration, driven by the width of the cores, which were too wide to fit four across the die. This floorplan created significant logical and physical complexity 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 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 56Gb/s (single-ended) and 96Gb/s (differential).

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 being distributed over the yield impact was deemed to be acceptable. This was crucial for the D-lcache, 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 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%.

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:
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.