Broad social and economic trends are challenging the semiconductor industry to adapt to the next generation of systems and services. Computing continues to migrate from desktops and notebooks to mobile devices and the cloud. Social media and multimedia growth will drive further consumer demand for wireless bandwidth. The emerging machine-to-machine communication network known as the Internet of Things (IoT) will place unprecedented demands on chip designs to deliver connectivity, reliability, and security with extremely low power consumption. Moore’s law continues to enable unparalleled levels of integration for processors, field-programmable gate arrays (FPGAs), and systems on chip (SoCs). All four trends were in evidence at the 27th Hot Chips Symposium on High Performance Chips (Hot Chips 27), held in August 2015 at the Flint Center in Cupertino, CA. This article presents highlights of new developments in silicon that align with these trends and offers a glimpse into the emerging technologies that will define the coming decade, focusing on a few designs and keynote addresses not covered by the technical articles in this special issue.

**Silicon for Cloud Computing**

The datacenters and enterprise servers that provide the infrastructure for today’s cloud-based and back-office business services depend on energy-efficient chips and high-performance networking. The Oracle Sonoma processor is a low-cost multicore Sparc processor targeting enterprise workloads. The processor integrates eight fourth-generation Sparc cores and is implemented in 20-nm CMOS with 13 levels of metallization. The cores are dynamically threaded, supporting one to eight threads with dual-issue out-of-order (OoO) execution. Each core includes support for cryptography in silicon. User-level instructions support a range of encryption and checksum standards, such as the Advanced Encryption Standard (AES), the Data Encryption Standard (DES), and the message-digest algorithm MD5. The hardware support provides security and transparent encryption that extends across the software stack.

The cores are organized into two four-core clusters. Cores have private 16-Kbyte level-1 (L1) instruction and data caches and shared 256-Kbyte level-2 (L2) instruction caches. Two cores share writeback 256-Kbyte L2 data caches. All L2 caches support 500 Gbytes per second (GBps) throughput. The shared level-3 (L3) cache is divided into 8-Mbyte partitions to reduce latency and improve performance because hardware accelerators and the integrated PCI Express (PCIe) and Infiniband direct memory access controllers can directly allocate lines into designated L3 partitions. Two integrated DDR4 memory controllers support four direct-attached DDR4-2133/2400 channels with up to two dual inline memory modules (DIMMs) per channel, enabling up to 1 Tbyte of memory per socket. Two Infiniband links (28 GBps bidirectional bandwidth) and 2 PCIe links (32 GBps) provide networking interfaces. Four coherence links (128 GBps) provide connections to other Sonoma processors.

On-chip power estimators in each core and L3 cache estimate dynamic power at 250-ns intervals and provide these estimates to an on-chip power management controller that performs dynamic voltage and frequency scaling (DVFS) according to software-defined policies. The fine-grained power-management features simplify system design and lower thermal design power. The processor includes a hardware database accelerator for business analytics that supports multiple operations on in-memory database vectors. This lets the software perform analytics at the system memory bandwidth. The Sonoma processor offers 1.3 to 2.6 times performance over the previous-generation Sparc T5."
L2 cache with directory-based cache coherence, in addition to the eight 4-issue (OoO) Xiaomi cores. Panels are connected with a 2D mesh network on chip (NoC). Network latency is three cycles per hop using Y-X dimension-ordered routing. Network bandwidth is 384 Gbps. Each panel also connects to an external cache and memory chip that integrates a 16-Mbyte L3 cache and provides two DDR3-800 memory interfaces (25.6 Gbps memory bandwidth). A proprietary parallel interface between the Mars CPU chip and the external cache chip decreases latency by sacrificing increased pin count compared to a serializer/deserializer interface. The interface has a separate write data/command channel (6.4 Gbps effective bandwidth) and read data channel (12.8 Gbps effective bandwidth). Mars is implemented in 28-nm CMOS running at 0.9 V in the core and supporting 1.8 V I/O. The chip consumes 120 W at 2.0 GHz in 640 mm² area.

Deep Learning and Social Media
Image recognition and speech recognition are essential functions of today’s mobile devices and social media platforms such as Facebook. Traditional machine learning involves training algorithms to recognize objects by first using a manually created feature extraction block to pull out salient distinguishing characteristics of the object and then classifying the features as representing a particular object with a trainable classifier block. Deep learning uses many hierarchical feature representations (each of which are trained) before the final classifier stage. Convolutional neural networks use filters and subsampling at each layer of the hierarchy. Because convolution is a dot-product operation, it can be implemented using basic multiply-accumulate functional units.

The recent development of large image databases (for example, the ImageNet dataset has 1.2 million training samples distributed across 1,000 categories) for training and the availability of fast and programmable general-purpose GPUs has greatly accelerated the use of convolutional neural networks (CNNs) in image recognition. Development of special-purpose hardware using FPGAs or ASICs is also proceeding rapidly, with many large companies and start-ups investing in accelerators for CNNs, especially with the goal of using them in embedded applications. Massive parallelism allows Facebook to process each of the 600 million images uploaded each day through two separate networks, one for image tagging and recognition and a second for facial recognition.

FPGAs coupled with servers in the datacenter could emerge as a significant competitor to GPUs. Although FPGAs have one-fifth the performance of GPUs on CNNs, their better energy efficiency could allow them to scale up to match GPU performance while staying within power constraints. Machine learning researcher Yan LeCun, a keynote speaker at the 2015 Hot Chips Symposium, asserts that soon CNNs will be everywhere, and not just in the cloud, but in self-driving cars, smart cameras, toys, and robots.

Multimedia and Graphics
Gaming performance continues to demand new innovations in GPU design and implementation, forcing today’s graphics systems to evolve into new form factors. AMD’s next-generation GPU, the Radeon R9 Fury, shows how packaging and the chip-to-chip interconnect can be leveraged to balance power consumption of memory and logic chips, a key enabler of small-form-factor PCs. The GPU is a multichip module comprising a central graphics-processor chip mounted on a silicon interposer and surrounded by die-stacked, high-bandwidth DRAM. The multichip module is mounted on a conventional ball grid array package substrate. The core processor implemented in 28-nm CMOS and configured for gaming performance contains 64 computation units and can deliver 4,096 operations per cycle.

In addition to several multimedia accelerators, the die also integrates a 2-Mbyte L2 cache and eight memory controllers to interface with the high-bandwidth memory (HBM). Each HBM die is a DRAM specially built for an interposer, with low power consumption and ultra-wide bus width. Die stacking is achieved with through-silicon vias (TSVs) and microbump connections between DRAM chips. TSVs and microbumps also connect the graphics processor to the interposer. The result is that the HBM implements a 1,024-bit-wide bus (compared to 32-bit-wide buses for GDDR5). Although the 500-MHz HBM clock frequency is 3.5 times lower than GDDR5, the extreme parallelism in the bus enables memory bandwidth of greater than 100 Gbps per memory stack. The HBM runs at 1.3 V instead of 1.5 V for GDDR5, further reducing dynamic power. Dissipating 175 W in a closed-loop liquid cooling solution for a graphics card, the GPU achieves more than 8 Tflops and 54 frames per second rendering on average (43 fps minimum) on a 4K ultra-high-definition (UHD) display.

The Road to 5G
Wireless has made a performance leap approximately every decade, from the advent of digital voice communications (2G) in the 1990s to the development of mobile broadband (3G) in the 2000s to today’s faster broadband (4G). The vision for 5G, the next-generation of mobile broadband communications, encompasses building on traditional growth in bandwidth while adding reliable, low-latency communication for mission-critical applications and scaling down to connect low-cost, low-power devices. According to another Hot Chips 2015 keynote speaker, Matt Rob, the CTO of Qualcomm, the 5G ecosystem will provide scalable wireless connectivity across multiple dimensions. To do this, 5G services will use licensed spectrum and opportunistically use unlicensed spectrum. First, 5G will support low (less than 1 ms) end-to-end latency and multi-Gbps throughput for mobile broadband services such as UHD video and virtual reality. Second, in addition to maintaining low latency, 5G will support high-reliability channels with low packet loss rates and multiple redundant links for
fault tolerance. These features are essential to support medical, industrial automation, smart grid, and vehicle-to-vehicle/vehicle-to-infrastructure communications. Finally, 5G must scale down to connect low-cost, low-power devices that have modest bandwidth needs but may need to communicate over long distances, such as the wearables, sensors, and actuators that will implement the smart homes, smart cities, and smart infrastructure of the future.

**Internet of Things**

Much of the IT industry is focused on the IoT as the next wave of computing technology. The number of connected devices worldwide is projected to reach 50 billion by 2020 and possibly 1 trillion wireless sensors by 2025. Even if wireless sensors had a 10-year operating lifetime, 1 trillion devices would require 275 million battery replacements or recharges per day. Energy harvesting from environmental sources such as indoor solar, thermal gradients, incident RF signals, or vibrations yields tens of μW/cm², necessitating two to three orders of magnitude improvement in power consumption. University projects have demonstrated a wearable electrocardiogram chip that transmits RF updates every 3 to 5 seconds and consumes 19 μW active power, and a wireless activity monitor that interfaces to a three-axis accelerometer and consumes 6.5 μW. The key contributions of these designs and similar chips in industry and academia is to combine subthreshold digital circuit operation and low-power RF circuits using system-level integration to improve energy efficiency. The tradeoffs involved result in “sweet spots” of between 1 m and 4 km for the RF range, between 1 Kbit per second (Kbps) and 1 Mbit per second for the data rate, and from hundreds of kHz to tens of MHz for the processor clock frequency.

Two chips from the start-up PsiKick exemplify these tradeoffs. The first is a self-powered wakeup radio chip that can be used to activate an existing IoT device with a conventional, mW-level RF chipset triggered by receiving an RF-addressing signal. The wakeup radio consumes approximately 500 nW during active mode and therefore needs no duty cycling to reduce average power consumption. Therefore, it can be retrofitted onto existing IoT devices to reduce the system power from the mW range to the less than 1 μW regime. A 32-kHz crystal oscillator provides the wakeup radio’s clock. An integrated power management unit (PMU) can harvest energy from a photovoltaic or thermoelectric generator and includes an inductive boost converter to raise the harvested voltage to 5 V. The energy harvesting subsystem can interface to a rechargeable battery with a range of 0.8 to 5 V, has integrated maximum power point tracking, and can boost a minimum input voltage of 30 mV. The cold-start voltage for the boost converter is less than 400 mV. The PMU’s voltage regulator subsystem is a single-inductor, multiple-output (SIMO) DC–DC converter that outputs 0.5, 1.0, and 2.5 V supplies while consuming 350 nA active current. The wakeup radio receiver can target the 433 MHz, 915 MHz, and 2.4 GHz industrial-scientific-medical (ISM) bands with the particular channel selected by a passive off-chip antenna matching network. The RF signal is rectified and passed through a comparator with automatic comparator threshold control to reject interferers. Once the comparator limits the input signal to swing rail-to-rail, a digital correlator matches the received code with the wakeup radio’s identifier. If the codes match, the output wakeup signal is driven off-chip to alert the rest of the system. The wakeup radio supports an 8.192 Kbps bit rate at 2.4 GHz and consumes 103 nW for −40 dBm sensitivity (corresponding to a 1 m range) and 235 nW for −56 dBm sensitivity (6.3 m range). The energy-harvesting boost converter achieves 97 percent peak efficiency, whereas the SIMO DC–DC converter achieves 80 percent and can deliver up to 50 mW to the load.

The second chip is an SoC designed for IoT sensing applications. The chip adds an analog front end, analog-to-digital converter, time stamping, and a digital interface for external sensors. The second chips expand the digital section of the wakeup chip to include a full microcontroller core, more digital I/O, and, eventually, hardware accelerators. The SoC’s power target is 20 μW.

“Always-on” IoT devices that constantly watch or listen for visual or auditory inputs present a significant challenge for chip designers, because the power consumption for capturing and transmitting such signals is on the order of hundreds of mW. A test chip in 14-nm CMOS uses several key features to drive the power consumption for such applications to the single-mW range for speech recognition and to tens of mW for image recognition. A vision-processing engine supports low-power imaging by enabling aggressive image sensor power gating, autoexposure processing assisted by light detection, and intraframe and data-analysis-driven encoding. Multiple image-processing applications (such as sign detection and recognition, face detection, character recognition, and gesture recognition) are implemented using preprocessing blocks tailored to the application, but they share a single optimized shifted neural network processor for classification. Power consumption is minimized by executing each block of the image processing pipeline for each application on different computational units in the SoC. For example, the gesture recognition pipeline implemented at 2 fps has a maximum response time of 200 ms while consuming several mW down to below 1 mW. The always-on speech-processing pipeline implements voice activity detection, keyword recognition, and command and control recognition on the chip while higher-level natural language processing occurs in the cloud. Voice activity detection and keyword recognition consume 1 to 1.5 mW and 1.9 to 2.5 mW (including the microphone power), respectively. Running audio/video capture, hand gesture recognition, and speech processing consumes approximately 22 mW total power.
Moore's Law Continues
The transition from 22-nm tri-gate technology to the 14-nm process node drives new designs for both FPGAs and mobile SoCs. The Stratix 10 FPGA aims to double the performance of the previous generation while reducing power by up to 70 percent. The FPGA is a system in package (SiP) with a core digital FPGA die connected to multiple transceiver chips through a package substrate die-to-die interconnect technology called the Embedded Multi-Die Interconnect Bridge. The individual dies also connect to package balls through the same substrate. This approach’s advantages include reduced packaging complexity compared to a full interposer and no reticle limits that would be imposed by a silicon interposer (as in the AMD Radeon 9 Fury) and decoupling analog transceiver development from the digital FPGA fabric. New variants can be created by replacing the transceiver dice with other components, such as memory or ASICs.

Modern FPGA configuration requires significant additional functionality. The traditional approach uses a shift register to address and load the configuration RAM under the control of a finite-state machine. Today’s FPGA configuration subsystem supports features such as encryption and decryption, bitstream compression, data redundancy, authentication, and partial-configuration management. The Stratix 10 configuration is managed in software by a secure device manager that communicates through a configuration NoC with local sector managers that configure the sectors within the FPGA fabric. A hard-macro 1.5-GHz quad-core ARM Cortex A53 application processor is integrated with the configuration subsystem, sharing peripherals. A cache-coherence unit maintains coherency between the processor and FPGA accelerators. Clocks are software routable, allowing the creation of local clock domains and more efficient use of global clock routing. Software-routed clocks also permit active skew management.

The FPGA core fabric building blocks consist of lookup-table-based adaptive logic modules; 1-GHz, 20-Kbit RAMs for data forwarding; and 1-GHz digital signal processor (DSP) multiply-accumulate units. The FPGA can provide 10 Tflops of IEEE 754–compliant floating-point performance. To double the performance, the FPGA fabric was rearchitected to include an optional register in each routing mux and building block input. Therefore, the insertion of an additional pipeline stage does not require consumption of a configurable logic block and its lookup-table delay overhead (eliminating wasted logic blocks and increasing the FPGA’s ability to host designs with high flip-flop to lookup-table ratios).

The routing mux registers also eliminate wasted routing that would have been required to connect to an available register for each added pipeline stage. The place-and-route algorithm chooses the optimal register to enable, and critical paths can be retired for delay minimization simply by pushing or pulling register placement along the path. Expanded use of voltage domains and power management enables power reduction (high-performance blocks run at 800 to 940 mV $V_{DD}$, whereas low-power blocks operate from 850 mV down to 800 mV). Memory and DSP blocks can be power gated for further power reductions.

SoCs for mobile devices must satisfy increasing demand for performance and battery life. The Atom-based SoC product family (code-named Cherry Trail) is the first chip manufactured in the Intel 14-nm tri-gate SoC process. The chip is 25 percent smaller, while integrating 30 percent more transistors and delivering more than twice the graphics performance of the previous generation fabricated in 22 nm. Two dual-core CPU modules, each with a 1-Mbyte L2 cache and max Turbo Mode clock frequency of 2.4 GHz, execute general-purpose code. The microarchitecture includes larger branch predictor arrays and enhanced out-of-order execution functions through a larger reorder window, deeper reservation stations, deeper store buffers, and the ability to handle increased load misses in flight.

A larger data TLB and targeted floating-point execution improvements combine with the preceding modifications to increase instructions per clock (IPC). New instructions are added to support cryptography and message authentication and to provide hardware assists for security. Other new instructions improve performance by accelerating multimedia and streaming performance and string and text processing of large datasets. The integrated graphics and media processor supports various graphics standards. Each execution unit in the graphics processor runs seven hardware threads, with 128 32-byte registers per thread. Execution units also include a simultaneous multithreading instruction dispatcher, a branch and messaging unit, and two floating-point units. Eight execution units are aggregated with a thread dispatcher, instruction cache, and texture/image sampler unit into a subslice. The subslices each support 64 bytes per cycle read bandwidth. Two subslices are combined with a 384-Kbyte L3 data cache with 64-byte cache lines and a shared local memory (64 Kbytes allocated per subslice). The graphics processor delivers four times the computing and pixel throughput and two times the texture throughput of the previous generation.

A dedicated media fixed-function unit supports standard video encoding and decoding such as H.264. Isolated power domains and power gating at the graphics/media partition, per graphics subslice, and per execution unit doubles performance-per-watt energy efficiency over the previous generation. A separate display processor can support up to three displays. An integrated sensor hub provides “always on, always sensing” functionality similar to the IoT test chip described earlier, in addition to acquiring sensor data, the hub can combine data from individual sensors to create a more complex virtual sensor directly addressable by firmware or the operating system. Low-power operation is achieved through clock and power-gating subsystems in the hub and the ability to turn sensors off. The sensor hub can operate...
independently when the host platform is shut off as well.

The chip contains a total of nine power rails, including logic rails at 0.75 to 1.2 V, a RAM rail at 1.15 V, and standard 1.8 and 3.3 V I/O rails. The chip is divided into power islands implemented in clusters at the physical partition level. Each cluster contains up to two power-gated islands and one always-on power island. More than 40 clusters on the chip provide fine-grained power gating. DVFS is applied to the following areas:

- the CPU cores;
- the graphics/media processor;
- the camera imaging processor, in which the imaging pipeline frequency is determined by the use case and the pipeline frequency in turn drives the voltage level selection; and
- the display processor, in which the display resolution determines the pipeline frequency, which in turn determines the display processor voltage level.

DVFS is also applied to the off-chip DRAM through the two integrated memory controllers. The Cherry Trail SoC delivers comparable performance to the previous generation chip on Windows

---

Table 1. Other Hot Chips 27 presentations and posters.

<table>
<thead>
<tr>
<th>Presentation</th>
<th>Author(s)</th>
<th>Affiliation</th>
</tr>
</thead>
<tbody>
<tr>
<td>PULP: A Parallel Ultra Low Power Platform for Next Generation IoT Applications</td>
<td>Davide Rossi et al.</td>
<td>University of Bologna, ETH Zurich, and ST Microelectronics</td>
</tr>
<tr>
<td>Architecture of the V6x Hexagon DSP for Mobile Imaging and Always-On Applications</td>
<td>Lucian Codrescu et al.</td>
<td>Qualcomm</td>
</tr>
<tr>
<td>A Scalable Heterogeneous Multicore Architecture for ADAS</td>
<td>Zoran Nikolic et al.</td>
<td>Texas Instruments</td>
</tr>
<tr>
<td>Revisiting DSP Acceleration with the Kalray MPPA Manycore Processor</td>
<td>Benoit Dupont de Dinechin</td>
<td>Kalray</td>
</tr>
<tr>
<td>LS2085/8A Freescale’s New QorIQ Layerscape Communications Processor</td>
<td>Zheng (John) Xu and Sam Fuller</td>
<td>Freescale</td>
</tr>
<tr>
<td>I/O Virtualization and System Acceleration in POWER8</td>
<td>Michael Gschwind</td>
<td>IBM</td>
</tr>
<tr>
<td>MIAOW—An Open Source GPGPU</td>
<td>Raghuraman Balasubramanian et al.</td>
<td>University of Wisconsin–Madison</td>
</tr>
<tr>
<td>The ARM Mali-T880 Mobile GPU</td>
<td>Ian Bratt</td>
<td>ARM</td>
</tr>
<tr>
<td>Professional H.265/HEVC Encoder LSI Toward High-Quality 4K/8K Broadcast Infrastructure</td>
<td>Hiroe Iwasaki et al.</td>
<td>NTT</td>
</tr>
<tr>
<td>Ultra-Low-Light CMOS Biosensor Helps Tackle Infectious Diseases</td>
<td>Zhimin Ding</td>
<td>Anitoa Systems</td>
</tr>
<tr>
<td>Five-Speed PHY Enables 2.5 Gbps and 5 Gbps</td>
<td>Ramin Shirani and Ramin Farjadrad</td>
<td>Aquantia</td>
</tr>
<tr>
<td>Ethernet Rates Over Legacy Copper Cables</td>
<td>Dheemanth Nagaraj and Chris Gianos</td>
<td>Intel</td>
</tr>
<tr>
<td>Intel Xeon Processor D: The First Xeon Processor Optimized for Dense Solutions</td>
<td>Adrian Tang et al.</td>
<td>Columbia University</td>
</tr>
<tr>
<td>Anti-Virus in Silicon</td>
<td>Koji Yamazaki et al.</td>
<td>NTT Labs</td>
</tr>
<tr>
<td>Lagopus FPGA—A Reprogrammable Data Plane for High-Performance Software SDN Switches</td>
<td>Sami Saab et al.</td>
<td>Cryptography Research</td>
</tr>
<tr>
<td>Why Microarchitecture Matters for Successful Security Library Implementations</td>
<td>John Lockwood</td>
<td>Algo-Logic</td>
</tr>
<tr>
<td>Comparison of Key/Value Store (KVS) in Software and Programmable Hardware</td>
<td>David Roberts et al.</td>
<td>AMD Research</td>
</tr>
<tr>
<td>NMI: A New Memory Interface for Cellphones to Supercomputers</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Flexible Video Processing Platform for 8K UHD TV</td>
<td>Sukjin Kim et al.</td>
<td>Samsung</td>
</tr>
<tr>
<td>A Low-Power and Real-Time Augmented Reality Processor for the Next Generation Smart Glasses Under 100-Cycle Thread Migration Latency in a Single-ISA Heterogeneous Multicore Processor</td>
<td>Gyeonghoon Kim and Hoi-Jun Yoo</td>
<td>Korea Advanced Institute of Science and Technology</td>
</tr>
</tbody>
</table>

- the display processor, in which the display resolution determines the pipeline frequency, which in turn determines the display processor voltage level.

DVFS is also applied to the off-chip DRAM through the two integrated memory controllers. The Cherry Trail SoC delivers comparable performance to the previous generation chip on Windows.
desktop and web applications and 1.4 to 2 times better performance on gaming. Video conversion is performed 60 percent faster, and systems built around the SoC deliver longer battery life with a smaller battery.

Silicon technology continues to evolve as die stacking, tri-gate field-effect transistors, multichip modules based on silicon interposers, and multichip SiPs become more widespread. Traditional chips such as multicore processors, FPGAs, and GPUs continue to set new benchmarks in performance and energy efficiency. Emerging applications such as deep learning, 5G mobile, and the IoT offer new opportunities to leverage silicon and will drive the next generation of chip innovations. Although Moore’s law continues to bring ever-increasing levels of integration, its inevitable end will result in more innovation in architectures, circuits, and applications. Table 1 lists other presentations and posters from the conference, highlighting the diversity of challenges being addressed by silicon solutions today. From the perspective of Hot Chips, the future indeed looks bright.

References
2. Oracle’s SPARC T5-2, SPARC T5-4, SPARC T5-8, and SPARC T5-1B Server Architecture, white paper, Oracle, Feb. 2014.

Rajeevan Amirtharajah is a professor in the Electrical and Computer Engineering Department at the University of California, Davis. He has a PhD in electrical engineering from the Massachusetts Institute of Technology. Contact him at ramirtha@ucdavis.edu.

Behnam Robatmili is the CTO and cofounder of AxStream. He has a PhD in computer sciences from the University of Texas at Austin. Contact him at behnam@axstream.io.