A Low-Power Integrated x86–64 and Graphics Processor for Mobile Computing Devices

Denis Foley, *Member, IEEE*, Pankaj Bansal, *Member, IEEE*, Don Cherepacha, Robert Wasmuth, Aswin Gunasekar, Srinivasa Gutta, and Ajay Naini

***Abstract*—The first AMD Fusion™ accelerated processing unit (APU), code-named "Zacate," incorporates a pair of Bobcat x86 processors, a 1 MB L2 cache, an AMD Radeon™ 6310 DirectX®11 GPU with 80 stream processors, a media accelerator, an integrated NorthBridge (NB), integrated DisplayPort, LVDS, and VGA display interfaces, a PCIe® Gen1 or Gen2 I/O interface, and a single 64-bit memory channel at up to DDR3-1066 on a single die implemented in a 40 nm bulk CMOS process.**

***Index Terms*—AMD fusion, APU, Bobcat, integrated graphics, low-power, Zacate.**

I. Introduction

THE AMD Fusion™ accelerated processing unit (APU) code-named Zacate is implemented in a 40 nm CMOS bulk process. The design features the new synthesizable low-power Bobcat x64–64 core and integrated AMD Radeon™ graphics. This paper shares details on the architecture, technology, functional units, justification for the AMD Fusion approach, and details on power savings techniques, power gating, and clocking. Some performance data is shared to provide context for the power profile of the device. This paper is divided into 13 sections including this introduction, technical topics, performance, conclusion, and references.

II. Architecture

The AMD Fusion APU code-named Zacate shown in Fig. 1 incorporates two low-power Bobcat x86 processors, each with a dedicated 512 KB L2 cache, an AMD Radeon™ 6310 DirectX®11 graphics processing unit (GPU), and AMD's Universal Video Decoder (UVD) media acceleration engine. Memory access to a single 64-bit DDR3-1066 memory channel is controlled through an integrated NorthBridge (NB). Zacate supports a four-lane PCIe interface to the AMD Fusion Controller Hub (FCH) and a four-lane PCIe interface to an external GPU if desired. The PCIe links are capable of running at either 2.5 Gb/s Gen1 rate or 5 Gb/s Gen2 rate, and are capable of switching between modes depending on the current allowed power state. Two unique display output streams can be presented over any two of the following: 1) a DisplayPort (DP1.1a) interface; 2) a combination low-voltage differential signaling (LVDS)/DP1.1a port; or 3) an integrated video graphics array (VGA) DAC. A digital frequency synthesizer (DFS) block supplies the CPU core, graphics, multi-media, display, I/O, and NB clocks.

The AMD Fusion architecture implements a very efficient form of unified memory architecture (UMA) in which a portion of system memory is reserved as graphics frame buffer memory. The graphics memory controller (GMC) arbitrates between graphics, video, and display memory requests and presents a well-ordered stream of system memory transactions through the NB over dedicated 256-bit-wide read and write busses. These GMC requests bypass all NB coherency mechanisms, allowing for fast direct access to memory and exposing all of the available memory bandwidth (8.53 GB/s).

III. Technology

Zacate is implemented in a 40 nm bulk CMOS process. Ten metal layers are used, the top two of which are redistribution layer (RDL). The die area is 75 mm^2. Each Bobcat core is 4.9 mm^2, and each L2 is 3.1 mm^2. Each core has approximately 7 nF of on-die capacitance. Excluding PHYs, the graphics, I/O, and multi-media blocks occupy 35 mm^2. There is approximately 60 nF of on-die capacitance associated with this logic. The die is packaged in a 19 mm $\times$19 mm ball-grid array (BGA) package with 413 0.8 mm balls. The package substrate is a 2-2-2 layup. 10 nF of VDDNB package capacitance and 3.3 μF of VDD capacitance are supported. Fig. 2 shows a Zacate die photograph with functional units labeled.

IV. Bobcat Core

The Bobcat core shown in Fig. 3 is a brand-new x86 design making its debut in Zacate. The core features a decoder capable of decoding two complex operations (COPs) per cycle. It supports the AMD64 64-bit ISA. The execution engine supports full out-of-order (OoO) execution, and the load/store engine can execute loads and stores out of order. The design features a high-performance floating-point unit and an advanced branch predictor. Streaming SIMD extensions including SSE1, SSE2, SSE3, SSSE3, SSE4A, and 128-bit mis-aligned data-type extensions are also supported. The design features 32 KB L1 caches and a dedicated 512 KB L2 cache. The design runs at 1.6 GHz, with the L2 running at half that rate. Bobcat supports core power gating.

Manuscript received April 26, 2011; revised June 22, 2011; accepted July 05, 2011. Date of publication October 19, 2011; date of current version December 23, 2011. This paper was approved by Guest Editor Alice Wang.

D. Foley is with Advanced Micro Devices (AMD), Inc., Boxborough, MA 01719 USA (e-mail: denis.foley@amd.com).

P. Bansal, S. Gutta, and A. Naini are with Advanced Micro Devices (AMD), Hyderabad 500034 A.P., India.

D. Cherepacha is with Advanced Micro Devices (AMD), Oakville, Markham, ON, Canada L6H 6T5.

R. Wasmuth and A. Gunasekar are with Advanced Micro Devices (AMD), Austin, TX 78735 USA.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/JSSC.2011.2167776

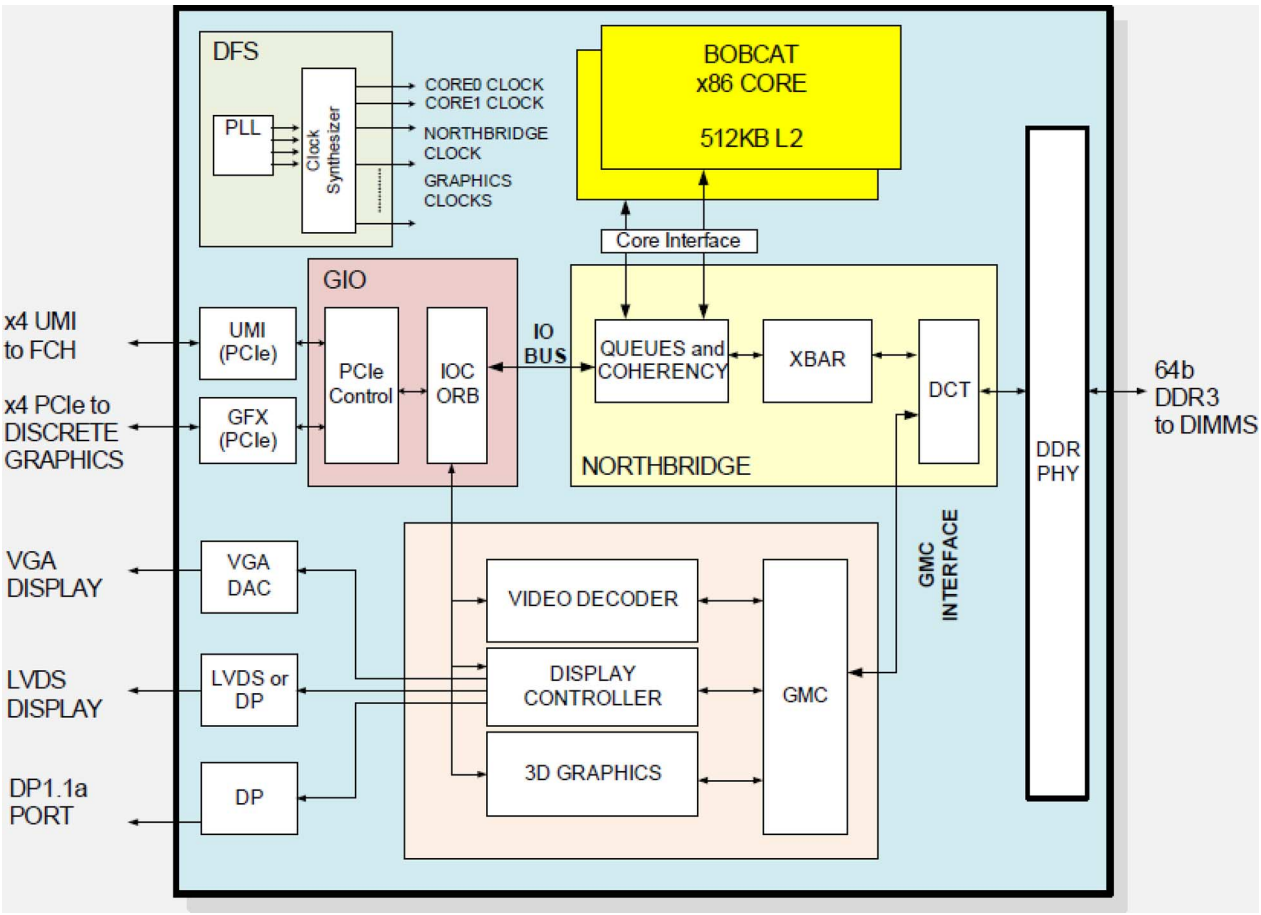


Fig. 1. Zacate block diagram.

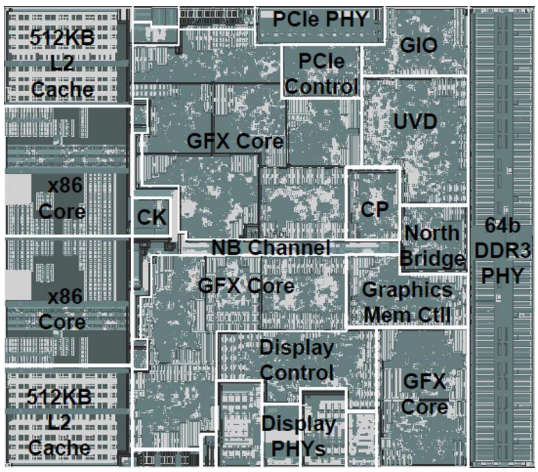


Fig. 2. Zacate die shot with labeled functional units.

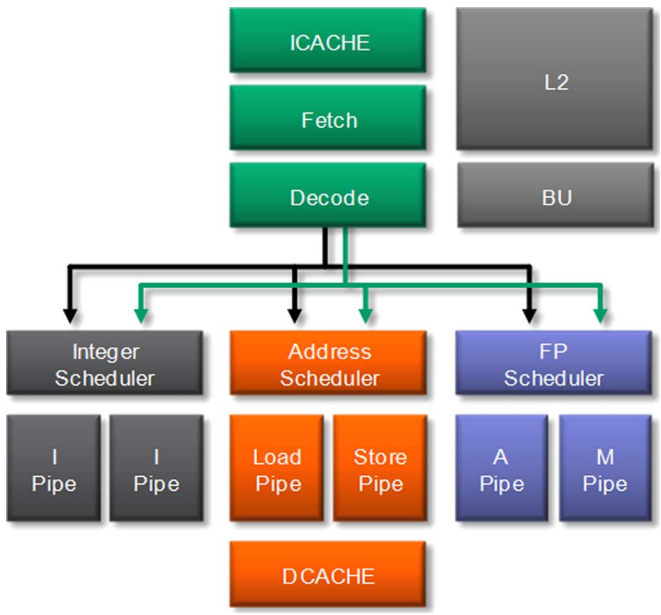


Fig. 3. Bobcat low-power core.

V. Graphics and Multi-Media

Zacate contains a small, power-efficient AMD Radeon™ HD 6310 GPU that is sized to take full advantage of the 8.53 GB/s of memory bandwidth provided by the 64-bit channel of DDR3-1066 memory. The GPU comprises graphics, video, audio, and display capabilities similar to a discrete graphics card. The APU uses a UMA in which the GPU's frame buffer is implemented by reserving a section of system memory. Typical memory allocation for the frame buffer is 384 MB. Additional GPU memory can be allocated dynamically from system memory.

TABLE I
BLU-RAY PLAYBACK

	UVD Enabled (HW-accelerated Decode)	UVD Disabled (SW Decode)
***Avatar* BD Video Playback**	**Smooth**	**Dropped Frames**
Ave. CPU Utilization	**31%**	**98%**

The DirectX 11-compliant graphics core contains 80 stream processing units, eight texture units, 16 Z/Stencil ROP units, and four color ROP units. The computational units are organized as a 2-SIMD x 8 array of vector processors in which each vector processor consists of five stream processing units. The unified shader architecture allows a flexible set of programs to be executed. At 492 MHz, the core's peak rate of 78.7 GFLOPs provides sufficient compute capability for gaming or compute applications enabled by DirectCompute and OpenCL.

Utilizing the AMD's latest video decoder, UVD3, Zacate provides uncompromised HD video playback, including many advanced video quality processing algorithms. UVD3 provides hardware acceleration for decode of H.264, VC-1, MPEG-2, and DivX/Xvid video streams. This facilitates low-power 1080p/1080i video Blu-ray playback. Table I shows measured CPU utilization when playing the *Avatar* Blu-ray disc (BD) with UVD acceleration enabled and disabled. Without acceleration, the CPU cores are fully utilized and unable to play the video without dropping video frames. With UVD acceleration, the CPU load is reduced to approximately 31%, facilitating a smooth video experience. The UVD also offloads from the CPU the decode of a variety of on-line video content, including Adobe® Flash® video.

Zacate provides two independent display interfaces that can natively support VGA, LVDS, and DisplayPort. HDMI and DVI can be supported with additional system-level components. High-definition audio is supported, including Dolby® TrueHD and DTS®-HD Master Audio.

The GIO block, shown in Fig. 1, performs system connectivity functions to route host and DMA traffic between the CPU cores, system memory, internal devices, and external devices. It contains a root complex supporting two four-lane PCIe Gen1 or Gen2 links. One link serves as the unified media interface (UMI) to the FCH. The second link supports discrete graphics attachments.

VI. FUSION BASICS

The traditional model of a processor chip (with integrated NB) coupled with an integrated graphics processor has a number of shortfalls. The high-speed PHY coupling the two processors (shown in red in Fig. 4) occupies significant area and consumes power. The power associated with the PHY can exceed 1 W during media playback. Additionally, the link may present a bandwidth bottleneck. When the two dies are integrated, a wide (256 bits in each direction) data path from the graphics memory controller to the NB is added, allowing for full access to system memory from the GMC. This path provides GMC clients with a low latency path to non-snooped regions of system memory, reducing the minimum read latency by up to 40%. Compared

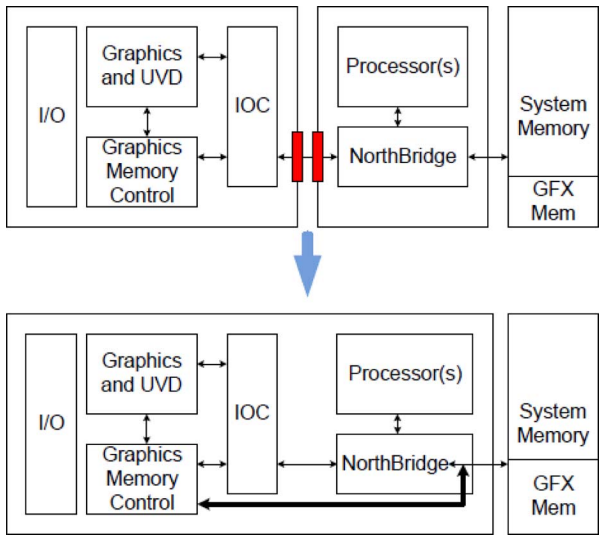


Fig. 4. AMD fusion advantage.

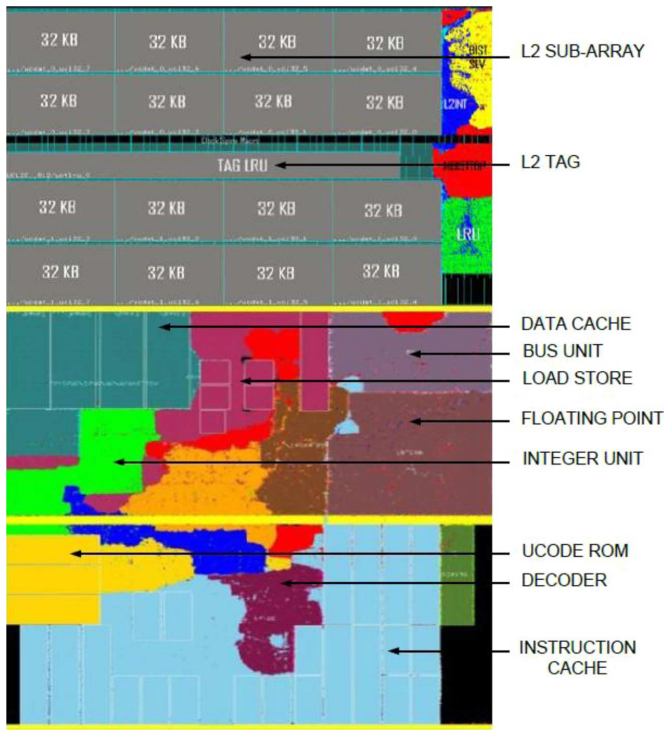


Fig. 5. Bobcat floorplan.

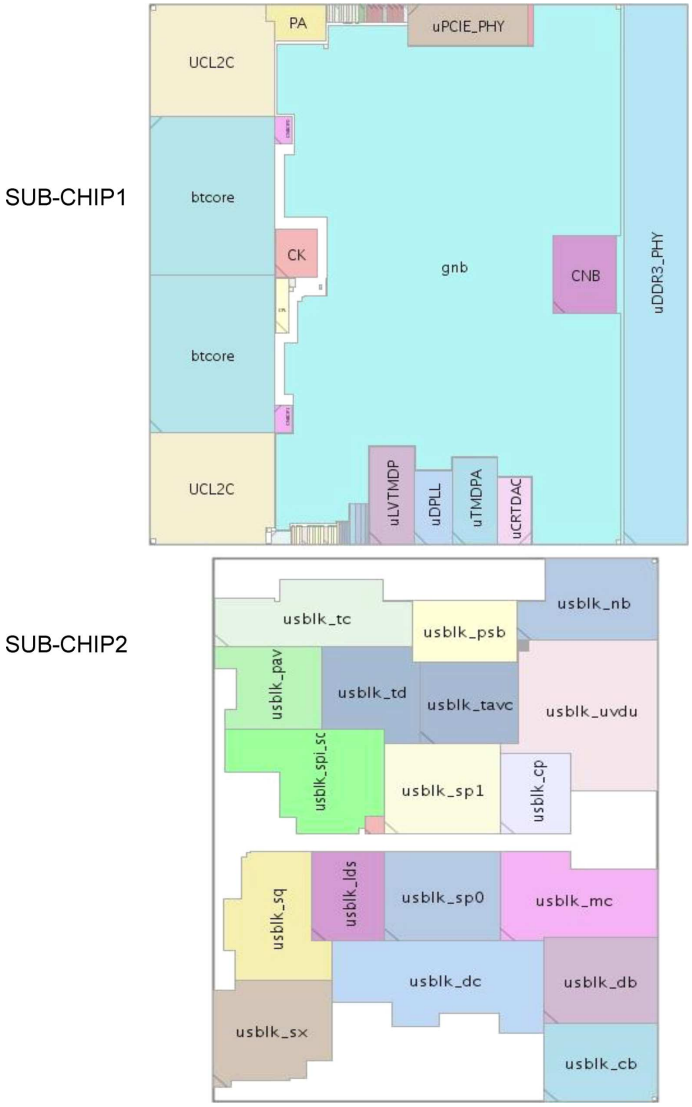
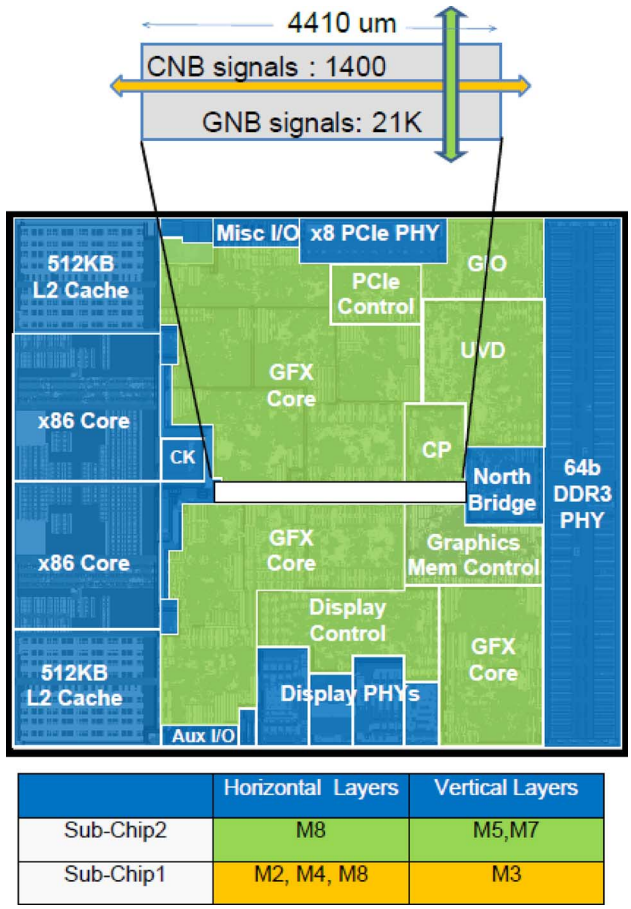


Fig. 6. Sub-chip1 and Sub-chip2 partitioning.



	Horizontal Layers	Vertical Layers
Sub-Chip2	M8	M5,M7
Sub-Chip1	M2, M4, M8	M3

Fig. 7. Routing allocation in channel for Sub-chips.

to two-chip solutions, use of the on-die integrated GPU significantly reduces memory latency, improves request ordering, and reduces area and power.

VII. Implementation

The AMD Fusion APU—including the x86 Bobcat cores—is synthesized from RTL and implemented using standard ASIC-style synthesis auto-place and route (SAPR) flows. The die shown in Fig. 2 has more than 450 million transistors.

The Bobcat core is implemented as a single physical entity consisting of 1.1 million instances and uses 35 instances of seven custom memory macros. The GPU core logical RTL is partitioned into multiple physical entities with varying numbers of standard cell instances and memory macros generated using standard memory compilers. The GPU sub-chip uses feed-through repeater bundles to connect the interfaces between different physical entities and has minimal grout space at the boundaries.

The Bobcat floorplan shown in Fig. 5 shows the various sub-blocks in the core as placement regions. L2 sub-array, tag array, and caches occupy most of the top and bottom of the floorplan, while the bus unit and floating-point unit are placement regions placed using various placement optimization techniques. The typical amoeba placement for various other sub-blocks in the floorplan shows the placement as achieved through careful floorplanning and tool optimization.

The SOC was implemented in two sub-chips, and a two-level hierarchical floorplanning and partitioning approach was used. Sub-chip1 consisted of the Bobcat core cluster, NB, and all I/O PHYs. Sub-chip2 consisted of the graphics core and multi-media and I/O control, collectively referred to as the GNB (Fig. 6).

Sub-chip1 and Sub-chip2 were further floorplanned to create an overlay channel between the Bobcat cores and DDR PHY through GNB sub-chip. Thus GNB Sub-chip2 could be independently designed and integrated with Sub-chip1 in SOC without causing any routing and integration issues. As shown in Fig. 7, specific metal layers were allocated to Sub-chip1 and Sub-chip2 for their respective signal interactions in horizontal and vertical directions.

The CPU and GPU core operate on separate voltage supplies and support dynamic voltage and frequency scaling (DVFS) to optimize power consumption. The design is optimized for power and performance at discrete process and voltage points. The CPU is optimized for power and performance at 1.2 V and 0.8 V. Timing optimization at high voltage secures the performance of the CPU, while timing at low voltage secures the minimum

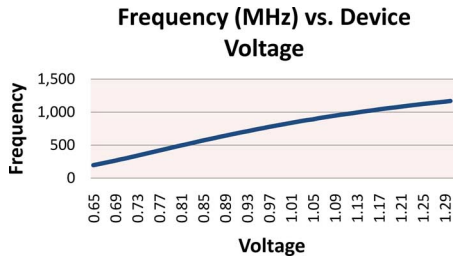


Fig. 8. Frequency vs. voltage for a representative GFX (SVT) path.

Instance Count	
Macro Instances	225
Memory Instances	866
Sequential Instances	2,473,912
Logic Instances	8,080,771
Inv/Buf instances	3,089,581
Total Instance count	**13,645,355**
Memory Bits	
Total Memory Bits in GPU	10,559,376
Total Memory Bits in Core+L2	12,230,872
Total Memory Bits in SOC	**22,790,248**
Transistor Count	
Memory Transistors	144,200,024
Standard cells (HVT, LVT, SVT)	305,330,039
IO Transistors	1,828,283
Total Transistors in SOC	**451,358,346**

Fig. 9. SOC statistics.

operating frequency and ensures that the best performance/watt is provided at the low frequency. The GPU is optimized at 1.0 V, 0.9 V, and 0.8 V. Incremental functionality or feature sets can be enabled at progressively higher voltages. The GPU design is closed at multiple points to ensure that the incremental capability is matched across the design elements and is optimized for performance/watt at all operating points. Fig. 8 shows how frequency of a representative GPU path—a mix of device and wire delays using SVT devices—varies with device voltage.

Standard cell logic transistors comprise 2/3 of the total 451 million transistors, while memories and macros cover the remaining 1/3. The dual Bobcat core and L2 account for 12 Mb of the 23 Mb on-chip memory.

Extensive powers saving techniques were used, including VT swapping across the GPU design and the Bobcat core. The design has transistors fabricated in various threshold voltages and lengths to facilitate performance/leakage trade-offs. At 105°C, LVT devices are approximately 4x leakier than SVT devices, which are in turn approximately 3.5x leakier than HVT devices. As shown in Fig. 10, approximately 56% of the logic transistors are HVT, and approximately 42% are SVT. To limit leakage power, less than 2% are LVT devices, and those are used only in critical paths. The VT distribution contains the regular-and different-length variants for the same VT devices to further reduce power leakage without compromising performance/area.

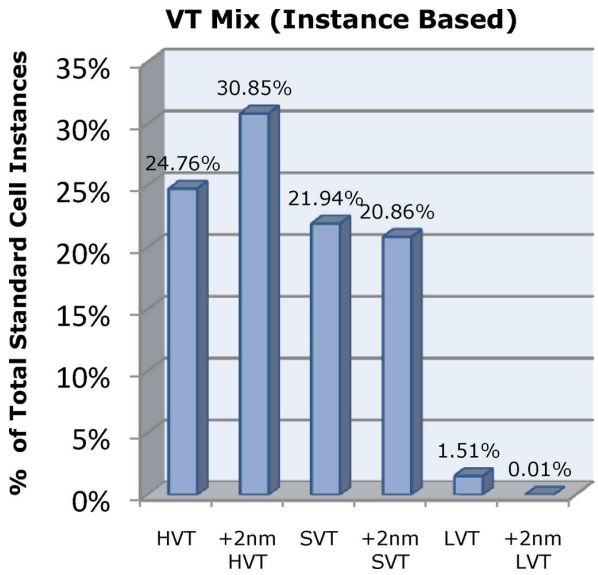


Fig. 10. Transistor threshold voltage mix.

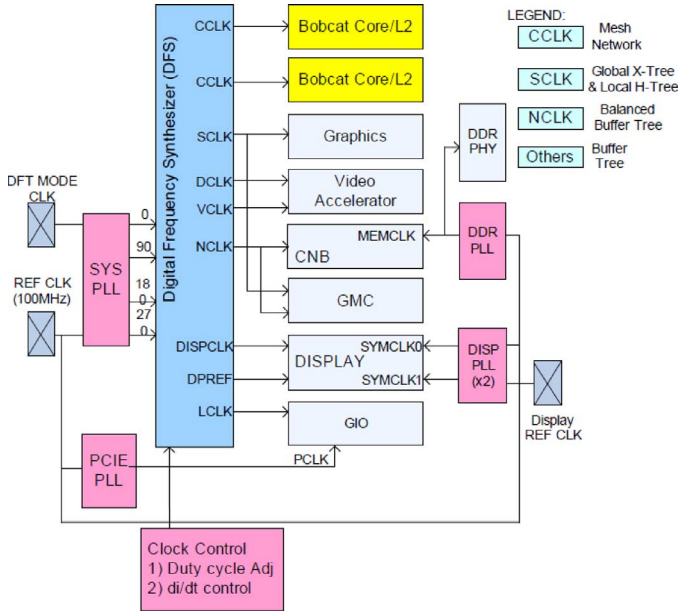


Fig. 11. DFS clock generation.

VIII. Clocking

The design has 16 functional, 10 scan, and 11 debug mode clocks. A digital frequency synthesizer (DFS) is used to generate nine functional clocks used by the CPUs, NB, and GPU (Fig. 11). The system PLL provides four phase-offset references to the DFS, which combines the phases to generate the required clock frequencies.

As shown in Fig. 12, ClkEn_A[3:0] is a 4-bit phase-enable value that is presented to the DFS every VCO period. Each of the bits corresponds to one of the VCO output phases. If a phase is enabled in a particular VCO period (shaded in blue in Fig. 12), that phase is combined with other enabled phases to generate the output clock. By controlling a repeating phase-enable sequence, different clock frequencies can be generated easily. Because the

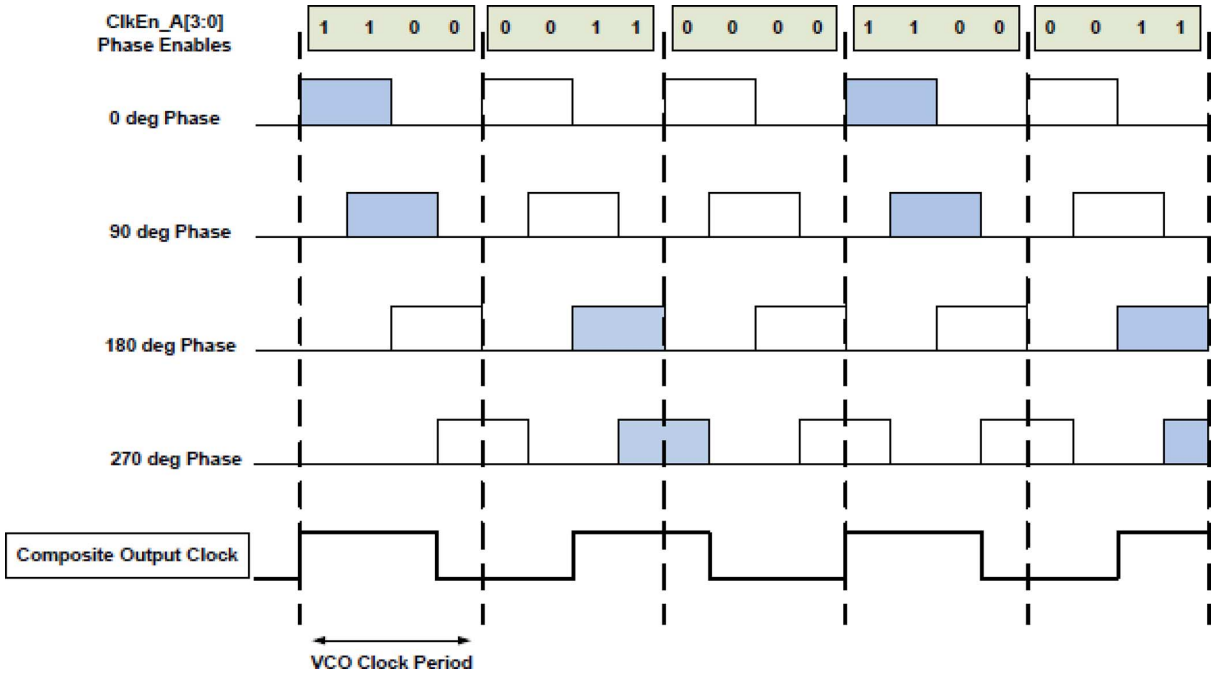


Fig. 12. Generating clocks using DFS.

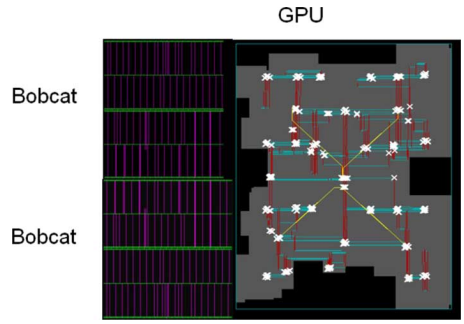


Fig. 13. Clock delivery.

phase-enable sequence can be changed at any time, clock generation is not interrupted on frequency transitions. This clock generation mechanism builds on a similar mechanism used in AMD Turion™ processors, but adds more VCO phases for finer frequency control.

The high-speed CPU clock is distributed as a mesh network with six to eight levels of local buffers. The GPU clock (SCLK) is implemented with X-tree and H-tree topologies feeding into the clock tree synthesis (CTS) branches inside a physical tile. All other clocks are implemented as balanced buffer trees from DFS to the physical tile. The NB and GPU clocks are balanced globally with respect to each other within a half-cycle of the fastest clock using programmable delay buffers pre-placed beside the global trees.

With such an implementation, Zacate achieves a global skew of 38 ps and 50 ps in CPU and GPU clock distribution, respectively. Fig. 13 shows the two distinctive clock tree structures as deployed for top-level clock distribution in the two sub-chips. A 5% clock jitter target was used for the synthesized clocks.

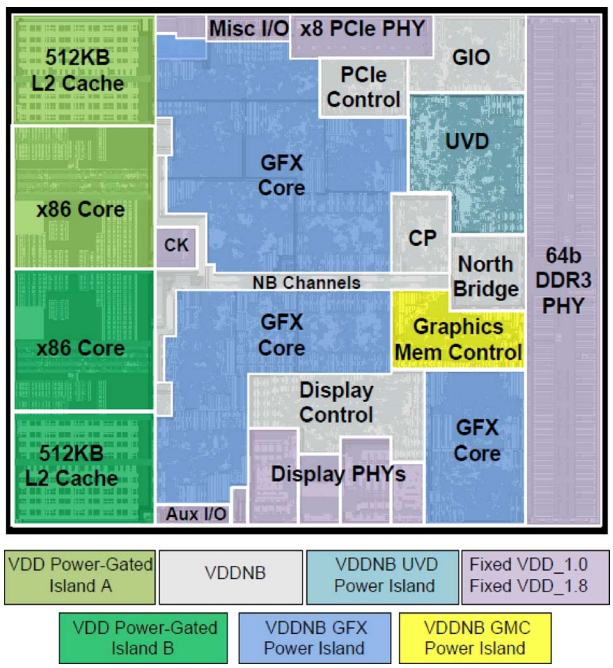


Fig. 14. Zacate power domains.

IX. Design for Power

Zacate's various power domains are shown in Fig. 14. Each Bobcat core and its L2 cache together share a power island (Fig. 15). A single variable VDD rail supplies both core power islands. Core clocks can be varied independently and VDD (CPU voltage) is selected to be the lowest voltage required to support the highest of the selected core clocks. Bobcat supports Core C6 (CC6) power state. If the core is idle, its caches can be flushed, its state saved to memory, and power to the core island is gated off. If both cores are idle and power-gated, the VDD

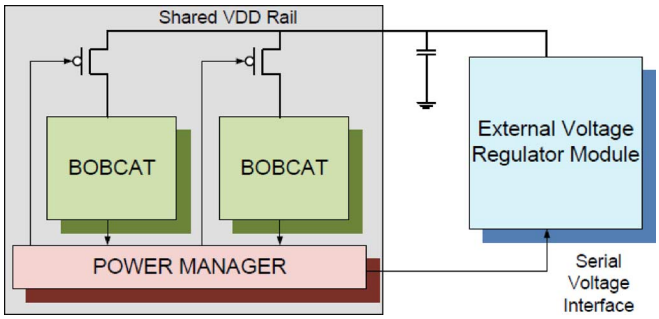


Fig. 15. Bobcat power gating.

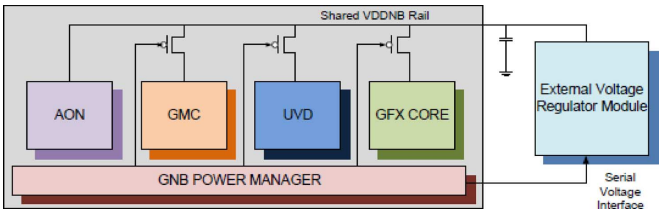


Fig. 16. VDDNB power gating.

power rail can be further reduced to eliminate leakage in even the power-gating structures. This state is called Package C6 (PC6). In PC6, voltage can be lowered to 0 V, but with frequent entry/exit to PC6, VDD is typically set at 0.45 V to minimize voltage ramp time and energy wasted in charging/discharging motherboard capacitance.

GPU and the NB share a variable VDDNB rail as shown in Fig. 16. GPU and NB clocks are varied independently based on activity. VDDNB is selected based on the highest voltage requested. The GPU supports several power islands, allowing for driver-controlled power gating of the UVD video acceleration engine and independent dynamic power gating of the GFX core and the GMC. The display controller supports a static-screen refresh-stutter mode in which the controller requests data periodically from memory; between requests, the memory is kept in self-refresh and the NB enters a low-power state. By stuttering the display requests to memory, the time spent in self-refresh for the memory can be maximized, the time that clocks are running in the NB can be minimized and startup penalties (PLL power on and lock, DLL lock time) can be amortized over a larger memory transfer. Typical stutter efficiency (time spent in self-refresh during static screen rendering) is greater than 94%. APU MM07 power varies by approximately 130 mW for every 10% change in stutter efficiency.

The combination of power savings techniques yields an APU with an average MobileMark® (MM07) power consumption of less than 1.8 W.

X. Power Gating Implementation

On-die power gating is implemented using a leakage-optimized HVT PFET header switch to isolate the VDD (or VDDNB) supplies. The header uses parallel stacked PFET transistors with separate enables for each PFET. This is in contrast to other x86 designs [1], [2] in which VSS isolation is

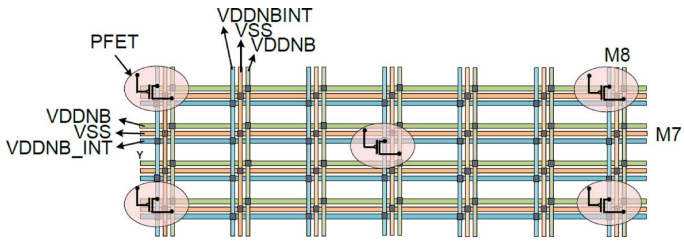


Fig. 17. VDDNB global power grid implementation.

used. The Bobcat core uses a dense grid implemented in M9 and M10 to satisfy the high current density in the core and to reduce IR drop impact on frequency. The total PFET width for a core and its L2 is about 1.0 m. The corresponding number of power-gating PFETs is approximately 30,000.

VDDNB power grid is implemented in M7 and M8 with power-gater PFETs placed on a checkerboard pattern (Fig. 17). Power gating for compiled memories is arranged at the top and bottom of the memory to avoid interference with the memory arrays. Tall memories implement a double row of PFETs at the top and bottom, while shorter memories implement a single row at the top and bottom. This is illustrated in Fig. 18, which is a close-up of a GPU tile showing the checkerboard pattern for logic and the single or double rows of PFETs for compiled memories.

Each PFET header, as shown in Fig. 19, is made from two FETS with separate enables. A smaller WAKE FET is used to initiate current flow into the grid on enable. The controls for the WAKE FETs are daisy-chained with a return path to the power controller. Once the WAKE FETs have been enabled, large RUN FETs can be turned on to provide a robust low-on-resistance connection to the power grid. The controls for the RUN FETs are similarly daisy-chained; once all the RUN FETs have been enabled, the power to the gated region is completely restored. This WAKE/RUN sequence is critical to avoid in-rush current spikes as the grid is enabled. The WAKE/RUN timing is programmable, allowing for post-silicon tuning of the delays in enabling the grid.

The VDDNB grid is designed to drive a current density of 0.5 A/mm^2 and achieve a 2% static IR drop. The total gate width of GPU headers is 1.93 m and resistance is 0.6 mΩ. Fig. 20 shows the simulated voltage drop distribution across PFET headers used in the design. The switches were uniformly distributed in the design through pre-placement, and the voltage drop varies across switches depending on density and logic in the region.

Fig. 21 shows four photon-emission captures of the Zacate die. All these captures are on a tester with clocks disabled, so the current flow is purely leakage. Capture duration was 30 seconds. Captures 1, 2, and 3 have VDD = 1.225 V and VDDNB = 0.875 V; capture 4 has VDD = 0.875 V, and VDDNB = 0.875 V. In capture 1, all functional units are enabled. In capture 2, CPU0 is power-gated off. In capture 3, CPU1 is gated off. In capture 4, all power-gated areas are off. The effect of power gating is immediately obvious. Capture 4 also highlights areas that are not gated off – the always on (AON) blocks. The small active area adjacent to the gated core in capture 2 and capture 3 is

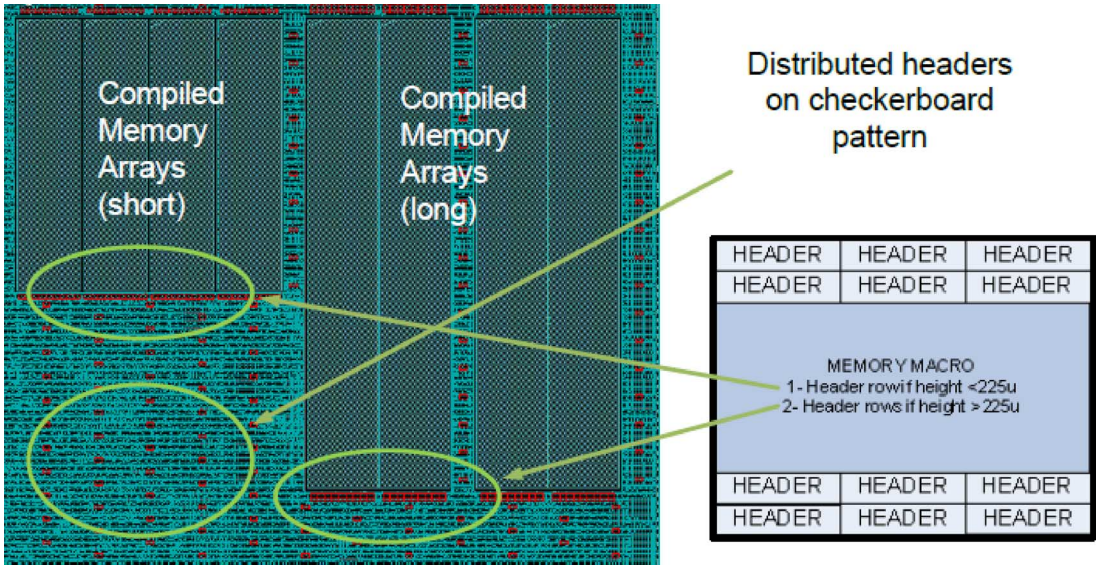


Fig. 18. Header distribution in typical GPU tile.

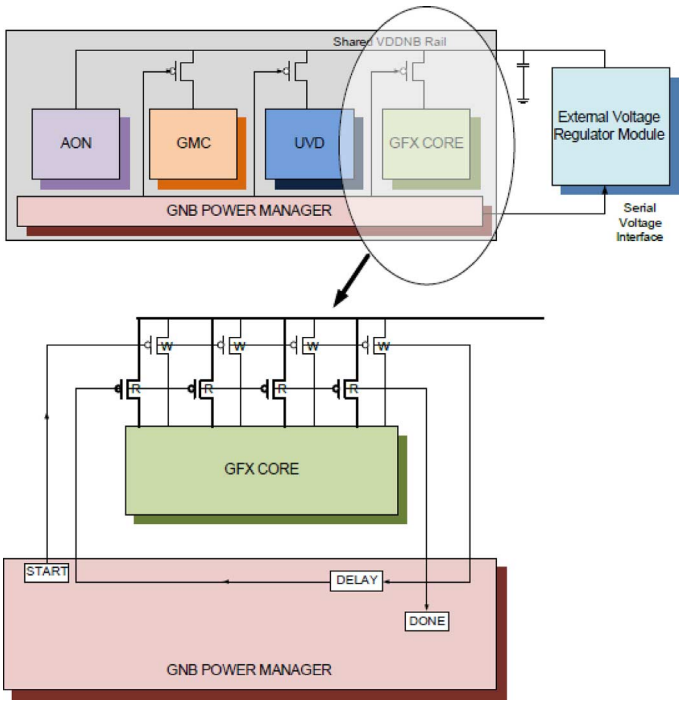


Fig. 19. WAKE/RUN power gater control.

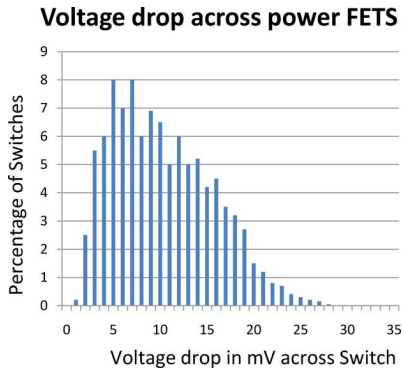


Fig. 20. Voltage across VDDNB PFETs.

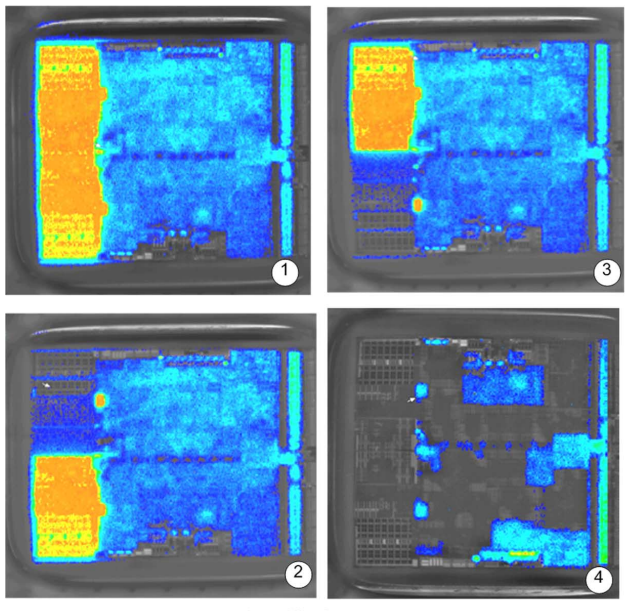


Fig. 21. Meridian photon-emission power-gating analysis.

interface logic between the core and the NB channel that was not power-gated.

XI. Bobcat Flop

Two flop types were used in the CPU core: a conventional master-slave flop, and a Bobcat flop (BT flop) shown in Fig. 22—a three-stage, pre-charged, asymmetric, MUXD flop. The BT flop is much faster than the master-slave flop, but is also bigger. The BT flop was used only on critical paths, thereby realizing its speed advantages while increasing the core area only a small amount.

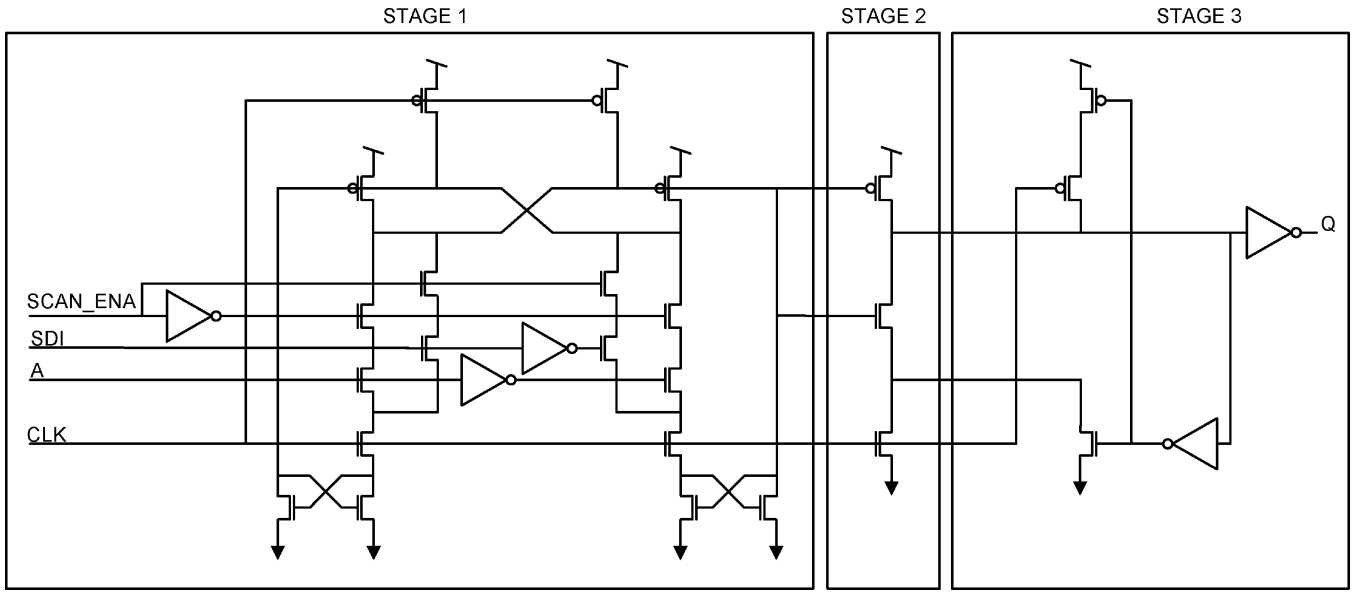


Fig. 22. Bobcat high-speed flop.

TABLE II
IMPACT ON AREA AND FREQUENCY OF BT FLOP

	MS Flop%	BT Flop%	Area Growth%	Fmax Improvement%
No BT Flops	100%	0	0	0
With BT Flops	87.4%	12.6%	2.06%	7%

The BT flop operates similarly to a sense amp flop [3]. While clock is low, both NFET stacks of Stage 1 are pre-charged. The half tri-state of Stage 2 is, therefore, disabled, and Q is determined by the active latch of Stage 3. When clock rises, Stage 1 is evaluated, Stage 2 is enabled, and Stage 3 feedback is disabled.

The BT flop is faster than a sense amp flop for two reasons: stack order in Stage 1, and size skewing afforded by asymmetry of the transitions. In Stage 1, cross-coupled NFETs are at the bottom of the stack, and pre-discharged when clock is low. Last-arriving data is at the top of stack. Regarding the second advantage, the BT flop is nominally three gate delays if the left NFET stack of Stage 1 discharges, but only two gate delays if the right stack discharges. This is exploited in the beta ratios of each stage so rising and falling transitions have nearly equal clock-to-q delays.

A majority of critical timing paths in processor designs are limited by slow setup, clock-to-q, or rising/falling edge, but not all of them simultaneously. A few experiments validated the hypothesis that SAPR tools could take advantage of flops designed to be the best in any given performance metric.

A flop library development effort focused on using such a flop as the base style and building specialist variants of the same to maximize frequency. A close review of the top critical paths revealed important criteria such as output edge rate, fast rising/fast falling, clock cycle extension, or clock pin buffering/unbuffering to favor fast clock-to-q or setup, or slow buffer to further favor setup. Other criteria included increasing the range of drive strengths and adding combinational functions such as NAND and NOR. The resultant library contained about 100 flop variants to optimize for the various combinations of these features.

TABLE III
GPU CAPABILITIES

GPU Clock Speed	492 MHz
ALU Compute	78.7 GFLOPs
Geometry Rate	123 MTri/s
Texture Rate	3.9 GTex/s
Color Fill Rate	2.0 GPix/s
WEI – Gaming Graphics	5.4

Performance is summarized in Table II. BT flops have longer hold time and higher dynamic power than master-slave flops, but when their usage was limited to only critical paths, frequency improvement for the CPU core was 7%.

XII. PERFORMANCE

Bobcat is an AMD64 $\times$86 ISA core targeting low-cost/area and high power efficiency while maintaining comparable single-thread performance to mainstream client processors such as the

TABLE IV
ZACATE GPU PERFORMANCE RELATIVE TO RADEON 4270

[2]	AMD Radeon HD 4270M	Zacate
3DMark06	1785	2015
3DMark Vantage – Entry	E2690	E3300
3DMark Vantage – Performance	P305	P690
3DMark 11 - Entry	DX10 GPU – no score	E450

[2] 3D-Mark configuration: AMD Mobility Radeon™ HD 4270M running at 592 MHz. AMD Guam reference platform with four-core Champlain processors (engineering samples) running at 2.3 GHz. RS880M/SB800. 4GB, 128-bit DDR3-1333. Resolution: 1280x1024 default, 1280x1024 performance, 1024x768 entry.
Integrated AMD Radeon™ HD6310 at 492 MHz, two-Core Bobcat core running at 1.6 GHz. SB810. 4GB 64-bit DDR3-1066. All games were run at 1024x768 resolution.

1.6 GHz AMD Turion Neo X2 L625, derived from the first-generation AMD Opteron™ core [4]. Although many applications are now multi-threaded and can take advantage of higher performance of Zacate's two cores, numerous user-sensitive operations are dependent on the performance of a single thread.

Central to Bobcat providing high single-thread performance was the choice of OoO execution with a two-wide decode/retire scheme. The fully OoO design with deep speculation enables energy saving by keeping the pipeline filled while reducing dynamic and static energy spent on stalled cycles as well as significantly increasing IPC relative to in-order designs used in other contemporary small cores. To reduce power, several micro-architectural approaches were used, including minimizing over-provisioning [5], reducing mis-speculation, and minimizing data transfers and data structure accesses. The two-wide design, relative to the three-wide design in the comparison machine, minimized over-provisioning in terms of execution units, register file ports, etc., but with some loss of performance, some of which was recovered through other optimizations. To reduce mis-speculation and improve performance, a sophisticated branch prediction structure was developed that utilized 32 k two-way L1 instruction cache with a return stack and indirect dynamic and advanced conditional branch predictors with a high-capacity 512/8 entry (4 k/2 m pages) instruction translation buffer (I-TLB). A physical register file is used both by the integer and floating-point units to reduce data transfers. The floating-point unit (FP) has a two execution stacks capable of two single-precision (SP) single-instruction/multiple-data (SIMD) adds and two SP SIMD multiplies per cycle. The load-store unit (LS) is fully OoO with a hazard predictor to minimize mis-speculations, a 32 k eight-way data cache with a 40/8 entry (4 k/2 m) and a 512/64(4 k/2 m) entry Level 2 data translation buffer (L2DTLB), and an eight-stream data pre-fetcher. The instruction and data cache share an ECC-protected 16-way 512 KB per-core private L2 cache, clocked at half the core rate to reduce power.

Fig. 23 illustrates single-thread performance comparable (∼0.9x) to the AMD Turion L625 based on SPECint2006 [6]. This is achieved in less than half the process scaled area and dynamic power of the AMD Turion L625 core. SPECint2006 is a benchmark that includes a broad range of compute-intensive integer applications that stress different aspects of the processor, from branch prediction to the cache hierarchy.

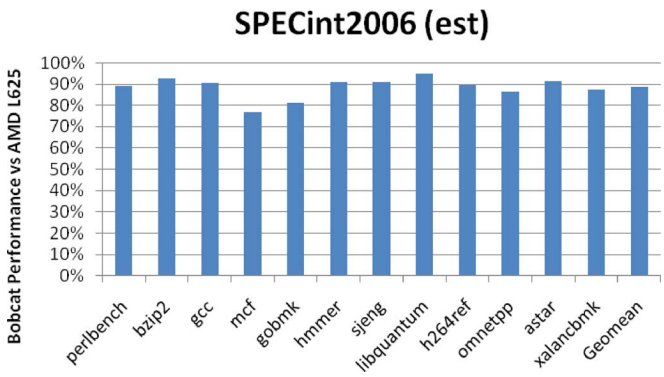


Fig. 23. Bobcat performance relative to AMD turion L625.

Table III specifies the basic performance metrics for the graphics core. The low-latency, high-bandwidth GMC memory interface increases performance per watt by allowing a higher percentage of the power to be used for computation rather than for moving large amounts of data across high-power chip-to-chip interconnects. For example, the power consumed by the HyperTransport™ link between the CPU and the AMD 880 G chipset was more than 1 W during periods of high GPU memory activity. With the AMD Fusion architecture, most of this interconnect power is saved.

The advanced, multi-level memory arbitration units resolve performance issues with previous integrated graphics processors (IGP) by enabling the GPU to achieve memory efficiency similar to a discrete GPU without compromising CPU performance. GPU memory accesses are sent in DRAM efficient streams of reads and writes to avoid memory access penalties while CPU accesses are optimized for low read latency. These improvements yield a 17% increase in Zacate UMA memory efficiency compared to its two-chip predecessor.[1]

Table IV illustrates the gaming capabilities of Zacate using the industry-standard 3DMark® benchmarks. The low-cost,

[1]Memory efficiency was measured using the Win7 Experience Index (WEI) graphics memory benchmark using system populated with one 64-bit DDR3-1066 SODIMM.

TABLE V
3D GAME PERFORMANCE

Benchmark	Quality Settings	Frame Rate
Call of Duty 4 – Strike (DX9)	Medium	29.3 fps
Borderlands (DX10)	Medium	27.2 fps
HAWX (DX10)	Low	38.2 fps
World in Conflict (DX10)	Low	44.0 fps

power-efficient Zacate APU outperforms the mainstream mobile AMD Radeon™ 4270 M GPU in the AMD 880 G chipset. In 3DMark Vantage – Performance, Zacate is 2.26 times faster. Table V shows benchmarking results using several popular games.

In addition to providing computational resources for 3D graphics operations, the 80 stream processing units may also be used to speed up many general-purpose floating-point-intensive applications. The OpenCL implementation of the floating-point-intensive Mandelbrot set generator in SiSoftware Sandra 2011 multi-media benchmark [7] illustrates this capability, yielding a 5x improvement in performance of the benchmark running on the GPU versus a multi-threaded, SIMD-vectorized implementation on the CPU cores. This result is consistent with the GPU-to-CPU peak FLOPS ratio of 6 to 1 (78.7 GFLOPs versus 12.8 GFLOPs).

XIII. CONCLUSION

The AMD Fusion Zacate APU brings together x86 processors, NB, I/O, multi-media acceleration, and compelling graphics processing capability on a single die. Using the low-power techniques described in this document, the part provides an excellent balance of performance and long battery life.

REFERENCES

[1] R. Jotwani *et al.*, "An x86–64 core implemented in 32 nm SOI CMOS," in *2010 IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 7–11, 2010, pp. 106–107.

[2] T. Fischer *et al.*, "Design solutions for the bulldozer 32 nm SOI 2-core processor module in an 8-core CPU," in *2011 IEEE Int. Solid-State Circuits Conf. (ISSCC) Dig. Tech. Papers*, Feb. 2011, pp. 78–79.

[3] J. Montanaro *et al.*, "A 160-MHz, 32-b, 0.5-W CMOS RISC microprocessor," *IEEE J. Solid-State Circuits*, vol. 31, no. 11, pp. 1703–1714, Nov. 1996.

[4] C. N. Keltcher, K. J. McGrath, A. Ahmed, and P. Conway, "The AMD Opteron processor for multiprocessor servers," *IEEE Micro*, vol. 23, no. 2, pp. 66–76, Mar.-Apr. 2003.

[5] K. Natarajan, H. Hanson, S. W. Keckler, C. R. Moore, and D. Burger, "Microprocessor pipeline energy analysis," in *Proc. 2003 IEEE Int. Symp. Low Power Electronics and Design (ISLPED)*, Aug. 25–27, 2003, pp. 282–287, Session 12.

[6] SPEC CPU 2006 Benchmark Suite, Standard Performance Evaluation Corp. (SPEC). [Online]. Available: http://www.spec.org/benchmarks.html, Reported results are estimated because measurements were made on pre-production hardware.

[7] SiSoftware: Sandra 2011. [Online]. Available: http://www.sisoftware.net/

Denis Foley (M'06) received the B.Eng.(Elect.) degree from University College Cork, Ireland, in 1983.

He is a Senior Fellow at Advanced Micro Devices, Inc. He has more than 28 years of experience in the computing industry. At Digital Equipment and Compaq, he was a design lead in the Alpha server group. At Hewlett-Packard, he was the implementation lead for a high-end Itanium server. At ATI, he was the chip lead for a game console cost-down before turning his attention to low-power design, first as the system architect for a licensable 3D GPU core, then as the system architect for ATI's Imageon family of application processors for hand-held devices. With AMD's acquisition of ATI, he moved into the SOC architect role for a number of AMD's low-power designs, including the recently announced AMD Fusion Zacate and Ontario APUs.

Pankaj Bansal (M'11) received the B.Tech. (E.E.) degree from the Indian Institute of Technology, Delhi, India, in 1997.

Since then, he has worked on various domains in microarchitecture, logic design, verification, and physical/circuit design/analysis at companies such as Intel, Freescale, Centillium, and Beceem. Prior to joining AMD in 2009, he was SOC lead/manager for 3G baseband ICs at Freescale. Since joining AMD in 2009, he has been responsible for multiple discrete GPU parts. Most recently, he has been leading design of the low-power AMD Fusion APUs from AMD's India Design Center.

Don Cherepacha received B.A.Sc. and M.A.Sc. degrees from the University of Toronto, Toronto, ON, Canada, in 1991 and 1994, respectively.

He is currently a Fellow at AMD in Markham, Ontario, and has more than 18 years of ASIC architecture, design, and verification experience. He joined LSI Logic in 1993 to work on PC and server chipsets as well as PCI and USB core development. He moved to Cogency Semiconductor, Inc. in 1998, where he led the ASIC architecture and development of that company's first HomePlug powerline networking ICs. He then joined ATI/AMD in 2004 and has worked in the areas of IGP/AMD Fusion architecture and performance for a range of chipsets and APUs.

Robert Wasmuth received the B.S.E.E. degree from the University of Texas at Austin in 1984.

He then joined IBM working on logic synthesis, circuit design tools, and performance modeling for the POWER series of microprocessors. He did performance analysis and modeling for various designs, including a network processor, a web-based transaction system, and a VOIP switch at different start-ups in Austin. After joining AMD in 2002, he has been involved in various aspects of x86 design, most recently leading the performance and power modeling team for the Bobcat core and SOC.

Aswin Gunasekar received the B.S. degree in electrical engineering from the University of Madras, India, in 2001, and the M.S. degree in computer engineering from the University of Texas, Austin, in 2003.

He joined AMD in 2004, and has worked in a variety of design roles that include I/Os, SRAMs, and high-performance standard cell libraries. He currently leads design methodology for a low-power family of microprocessors. He has four pending patents on processor circuits.

Srinivasa Gutta received the B.E. (Electronics and Communication) degree from Osmania University, Hyderabad, India, and the M.S.E.E. from the University of Texas at San Antonio.

He has 18 years of experience in SOC design, verification, and architecture. He started his career at CommQuest Technologies (acquired by IBM) designing voice coder DSPs for GSM baseband chips. He then joined Lucent/Agere Systems and spent 13 years working as SOC lead/architect on various communication chips, including the world's first PCI modem, K56Flex modem, and ADSL-Lite modems for client and central office. He also worked as lead engineer on second- and third-generation Sirius satellite radio chipset and mobile application processors in the Agere Mobility division. At AMD, he led the design of the Imageon audio processor and AMD Fusion Zacate/Ontario APU from AMD's India Design Center.

Ajay Naini received the M.S. degree from Mississippi State University.

He is a Senior Director at AMD. He has 25 years of CPU design experience and worked at Motorola, Cyrix, and HaL Computer Systems in various capacities before joining AMD 10 years ago. He was one of the lead engineers on the K8 processor family design at AMD, which delivered the first 64-bit x86 microprocessor and the first dual-core processor. Most recently, he was the Project Director for the AMD Fusion Zacate/Ontario design. He has more than 10 patents and several publications in floating-point and microprocessor design.