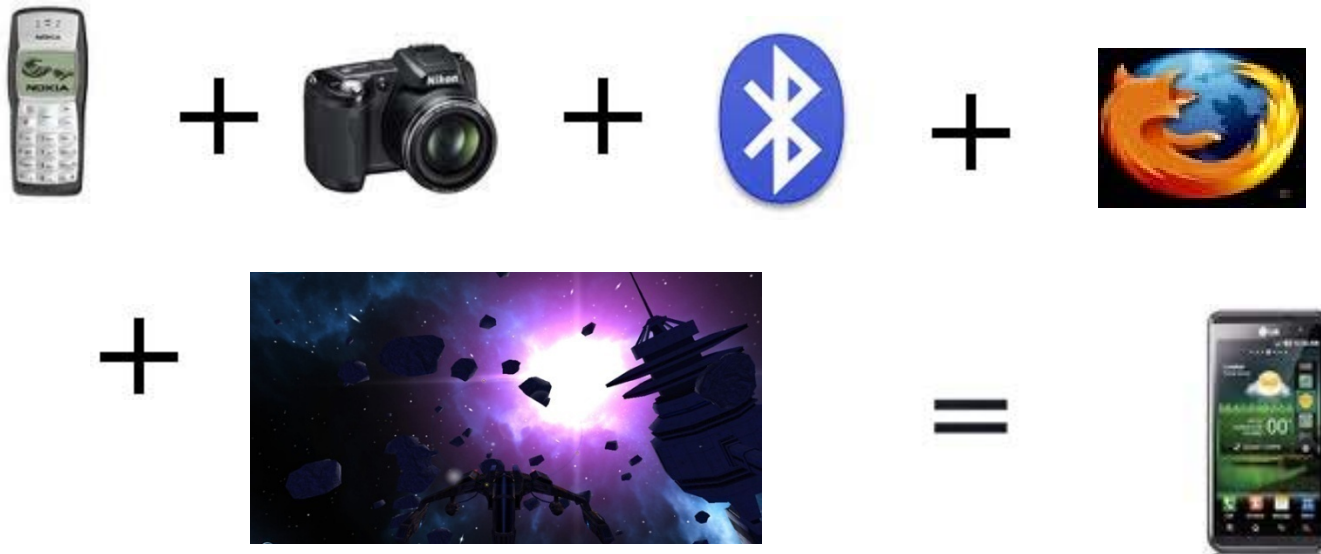


Abhishek Rawat
Avinash Kalyanaraman

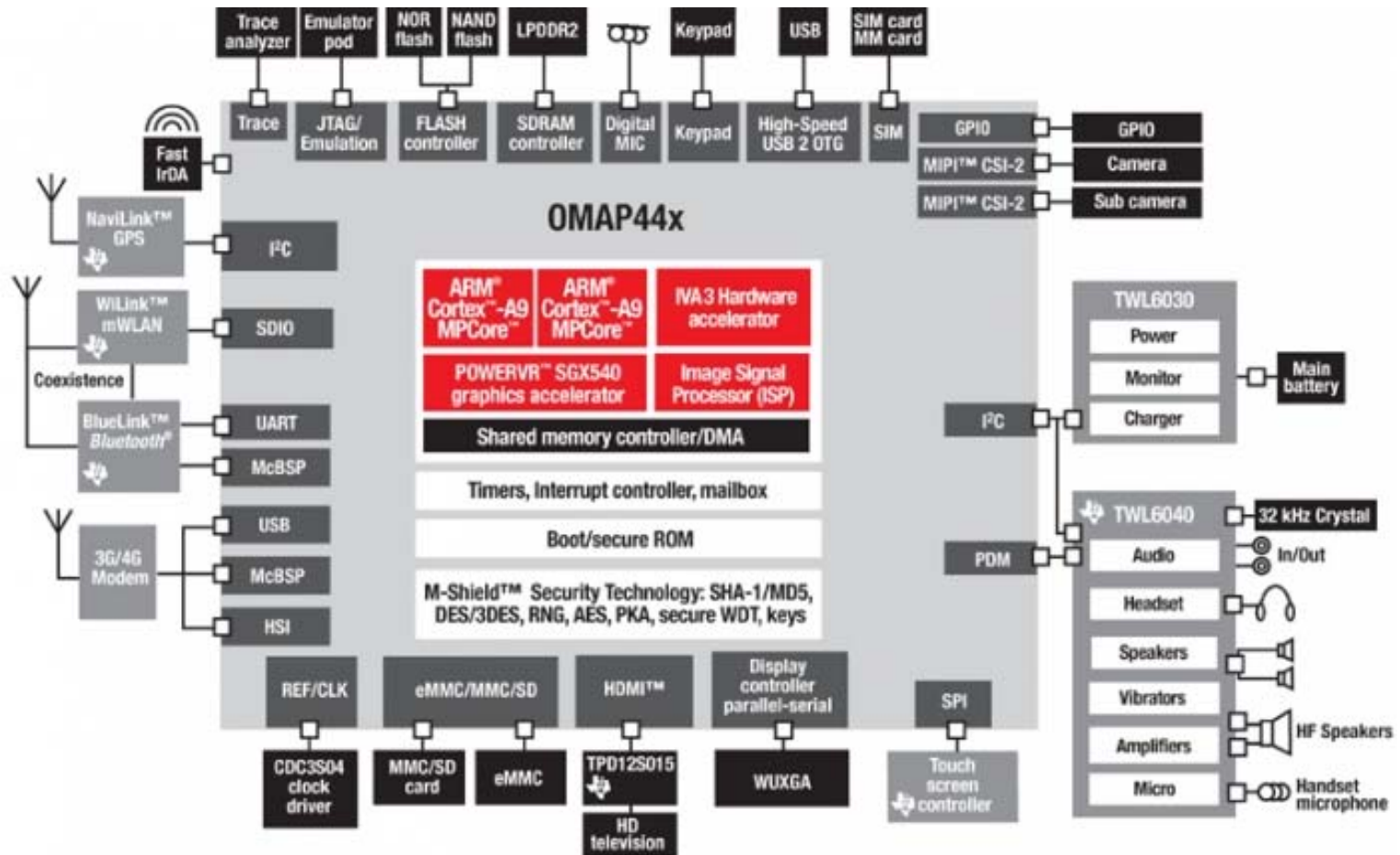
- Mobile Devices today –
Telephony + PC-like functions +
Multimedia + Gaming





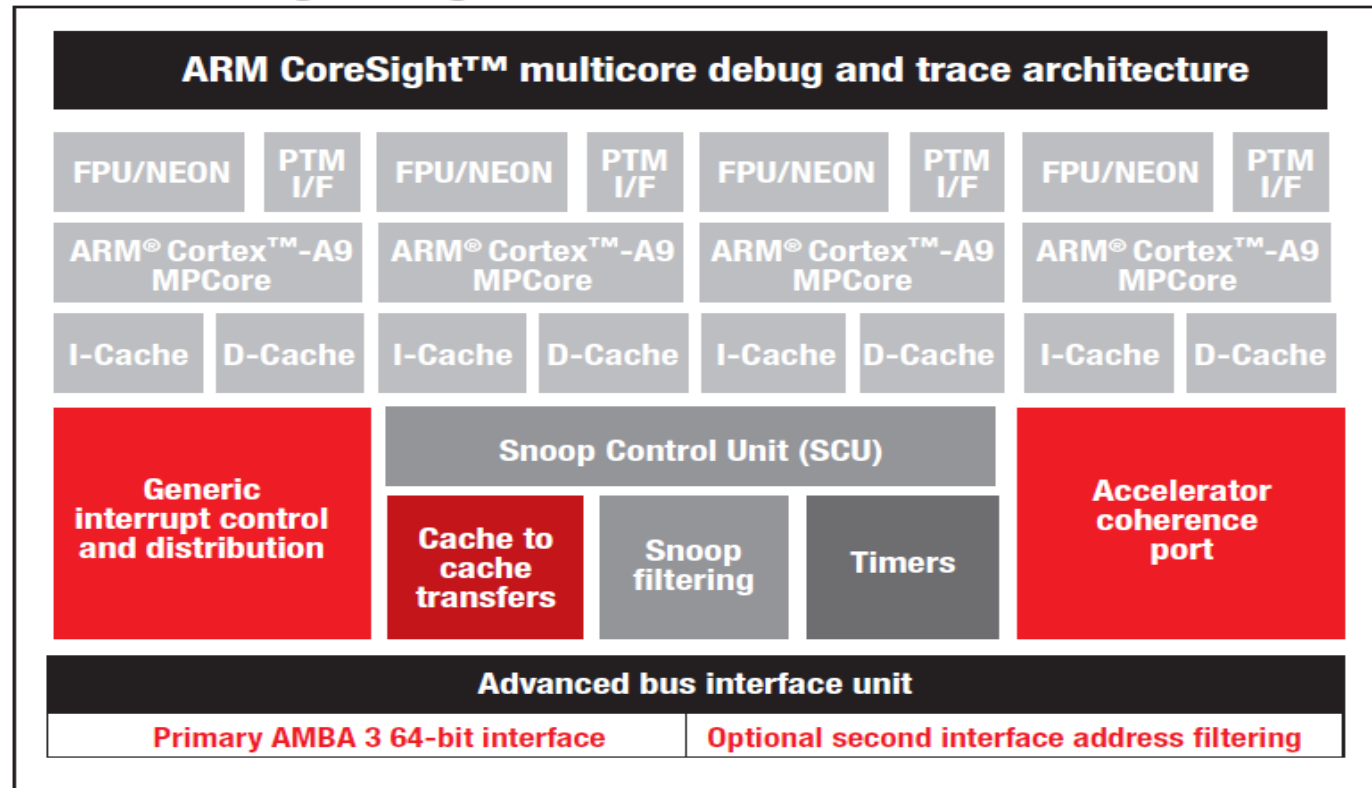
- **Challenge:**
 - Small form-factor (Size)
 - Providing high-performance with all-day battery life.
 - Cost

OMAP 44x System Diagram



Source: TI

Processing Engine #1: ARM CORTEX-A9



According to NVIDIA and TI: ARM Cortex A9 has 20% greater IPC than ARM Cortex A8

Source: TI



Need for SMP

- Inability of current uncore solutions to meet the processing demand of applications.
- Diminishing returns of large cores
 - Increased Si complexity with low performance gain
 - Complex design and validation process
 - Increased power consumption
 - Transition similar to PC
- Care must be taken to ensure additional cores do not negatively affect low power consumption.

Browsers

The screenshot shows a Mozilla Firefox browser window displaying the CNN.com 'Topics' page. The browser's address bar shows the URL <http://www.cnn.com/topics/>. The page content includes the CNN logo, navigation links (HOME, WORLD, U.S., POLITICS, CRIME, ENTERTAINMENT, HEALTH, TECH, TRAVEL, LIFE), and a 'Topics' section with various news items like 'Generation Islam', 'Black in America', 'Summer 1969', 'Michael Jackson', 'Iran Election Fallout', 'Money & Main St.', 'Remembered', 'CNN Heroes', 'The Sotomayor Vote', and 'The 44th President'. Overlaid on the browser are three system monitoring windows from the System Monitor application. The top window, titled 'CPU History', shows a line graph of CPU usage over 60 seconds, with CPU1 at 56.9% and CPU2 at 30.4%. The middle window, also titled 'CPU History', shows a similar graph with the same CPU usage percentages. The bottom window, titled 'Memory and Swap History', shows a line graph of memory usage over 60 seconds, with Memory at 191.8 MiB (43.6% of 439.6 MiB) and Swap at 0 bytes (0.0% of 0 bytes). Below the memory graph is a 'Network History' window showing a line graph of network activity over 60 seconds, with a 'Receiving' section showing 996 bytes/s and a 'Sending' section showing 0 bytes/s. The total received is 14.2 MiB and the total sent is 1008.7 KiB.

Source: NVIDIA

Gaming Engines

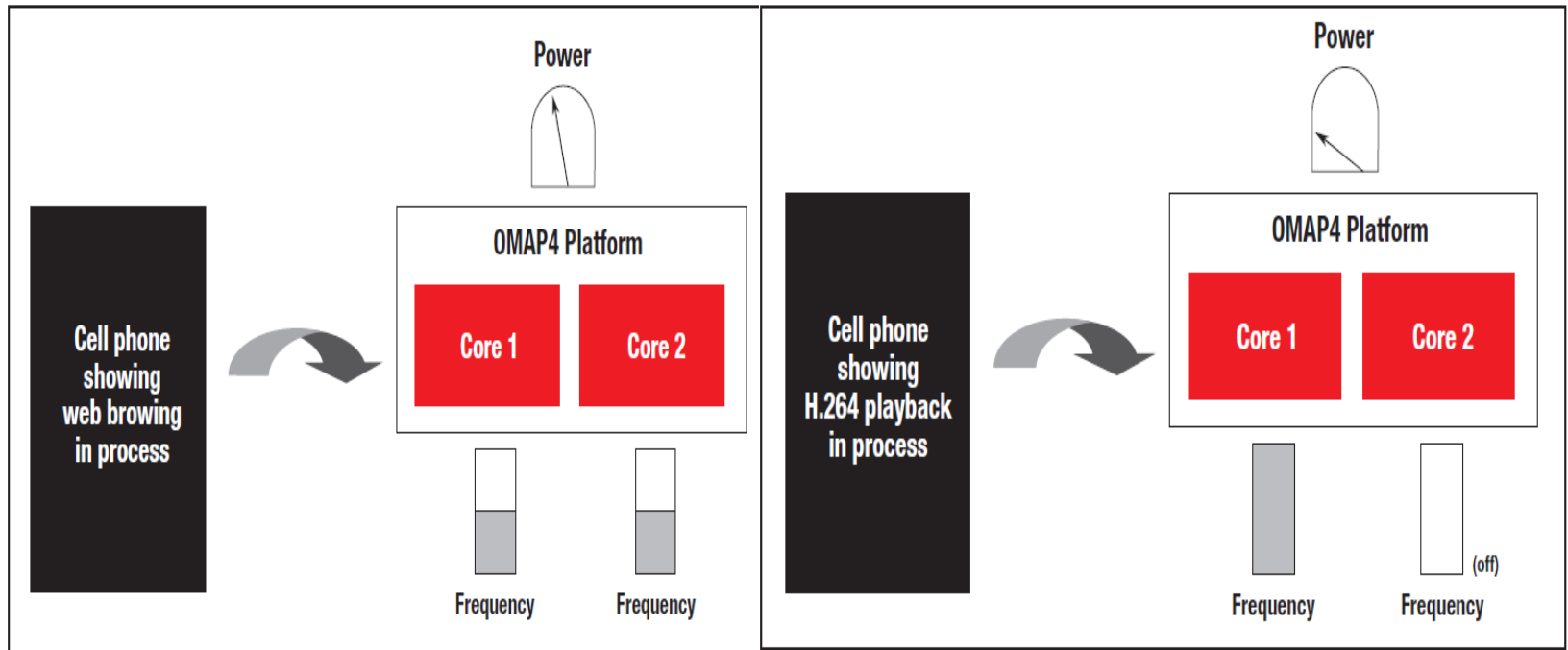
Game/Engine	Number of Threads
Unreal Engine 3	4+
Id Tech 5	6+
Frostbite	14
Civilization 5	12
Mafia 2	4
Crysis	8
Uncharted 2	8
Killzone 2	8+

The number of mobile gamers in the US alone is expected to increase by 50% from 2010 to 2014, touching 100million by 2014 *.


*: <http://www.mobilemarketer.com/cms/news/research/7096.html>

Source: NVIDIA

SMP at work



Source: TI



Processing Engine #2: IVA 3 Hardware Accelerator

- Permits full HD video record and playback at 30 frames per second
- Hardwired codecs for mainstream codecs like H.264 HP, MPEG4 ASP, MPEG-2 MP etc to deliver high performance at low power levels.
 - Few codecs are hardwired. Exact number unknown.
- Programmable DSP provides flexibility for adding future codecs to the chip without replacement.
- No mention about power consumption

Tegra-2 consumes 400mW for decoding full HD H.264 video at 20Mbps.



Processing Engine #3: Image Signal Processor

- Goal:
 - Deliver digital-SLR like performance
 - 20MP still-image capture support at one second shot-to-shot delay.
- Support image processing features like digital anti-aliasing, auto-focus, digital zoom, noise-filtering etc.
- IVA and ISP seem to be accelerators from TI.
 - TI did not prefix these products with the owner of IP or mark them to be trademarks of someone else, in their whitepapers.



Processing Engine #4: POWERVR SGX540 GPU

- Why hardware accelerated graphics on phones ?
 - Mobile devices today are increasingly being defined by the type of UI they offer.
 - UIs influence sales. Intuitive touch screens.
 - Support for advanced Gaming.
- Therefore h/w accelerated graphics = paramount importance.
- >4x the sustained performance against the previous SGX530 core
- Uses a technique called Tile-based deferred Rendering
- No mention about the frequency of operation
 - Anandtech.com speculates it to be ~300MHz



Memory Management

- Why ?
 - The processing engines result in an increase in processing bandwidth.
 - How to supply data at high bandwidth for processing ?
 - Users would want to use the system without experiencing system hangs.
 - Inability to deliver full HD video at 30fps with in-sync audio. Insufficient bandwidth results in out-of-sync audio or lower fps, when the user can start detecting inconsistency in the video.
 - Browsing would no longer give a PC-like experience.



Memory Management continued

- High bandwidth memory interfacing
 - 2-channel LPDDR2 Memory Interface
 - OMAP 4430 provides >4x memory bandwidth over OMAP 3630.
 - Tegra-2 provides single-channel 32-bit LPDDR2 memory interface.
- Video Subsystem:
 - High bandwidth internal L2 memory to reduce traffic to DDR
 - Compression of motion estimation search window luma (brightness) data.
 - Motion estimation search window management
 - Motion compensation bounding box to combine loads into one larger load.



Memory Management continued

- Image Subsystem
 - "Fine bandwidth-control for memory-to-memory transfer delivering 200MP/s throughput at 200MHz"
 - High Performance DMA Engine



Memory Management continued

- **POWERVR SGX540 GPU:**
 - Tile based deferred rendering
 - conserves both power and memory bandwidth- two precious commodities on smartphones.
 - Given the lack of serious 3D gaming, much less geometry heavy titles on smartphones today, the tile based approach makes a lot of sense.
 - 128-bit internal memory bus.
- **Tegra-2 :**
 - Immediate-mode rendering with hidden surface removal
 - Similar approach to their PC GPUs



Memory Management continued

- “Improved” SDRAM Controller improves performance
 - Re-orders commands to maximize overall memory usage
 - Delayed Writes to limit read-to-write or write-to-read transitions
 - Support for single request/multiple data transactions to effectively increase look-ahead FIFO depth and allow more effective re-ordering of commands



SmartReflex Power and Performance Management Technologies

- Goal : reduce power consumption and optimize performance
- Motivation
 - Battery capacity cannot keep pace with the exciting new functionality - video, mobile digital TV, high-fidelity audio, 3-D video games, digital photography etc



SmartReflex Power and Performance Management Technologies

- Motivation continued
 - Smaller, more highly integrated components are needed to fit more functionality into compact new form factors
 - Leakage current exponentially increases with process scaling
 - With shrinking process technology, the extra on-chip memory added to support new applications becomes prime source of standby leakage current
 - Power dissipation translates to heat which is
 - discomfort to users,
 - negative effects on performance
 - eroding the reliability and durability of the system itself



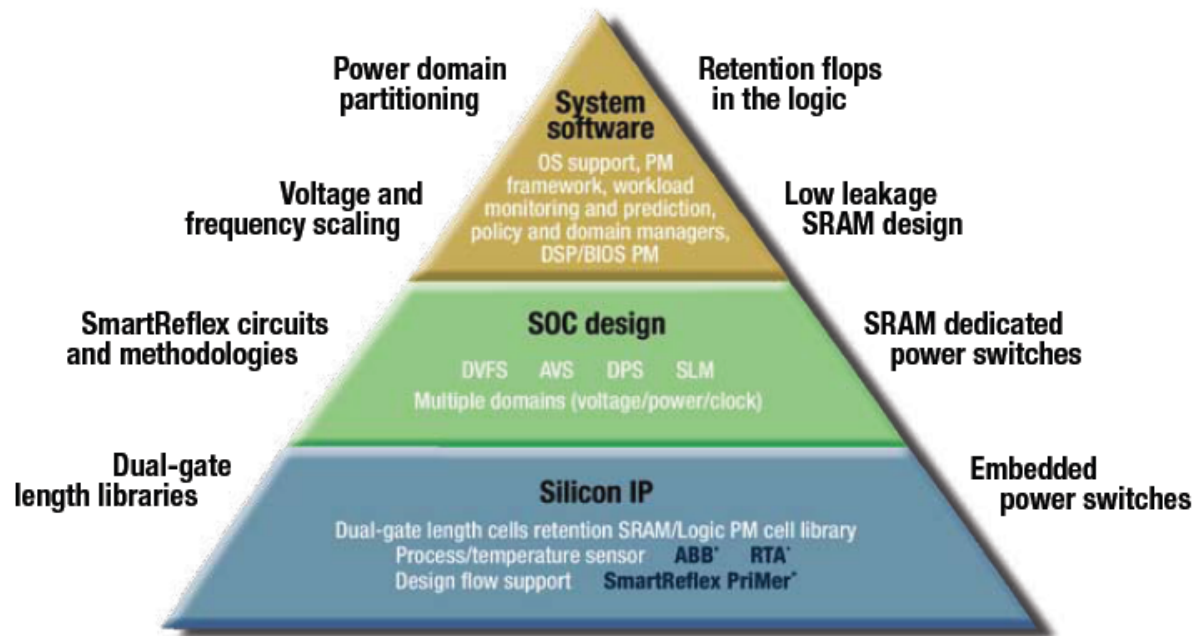
SmartReflex Power and Performance Management Technologies

- Approach
 - Holistic power management approach: Process technology, hardware, system-on-a-chip and software
- Key Benefits and Features
 - Dynamically adjusts transistor performance versus leakage
 - Dynamically lowers voltages for idle memory banks
 - Automates the application of SmartReflex technologies in the design process
 - Adaptive silicon circuit design and software
 - Adaptively adjust voltage, frequency and power based on device activity, modes of operation and temperature for maximum power reduction
 - Add new multimedia applications without sacrificing standby time, talk time or battery life

SmartReflex Power and Performance Management Technologies

TI SmartReflex™ 2 Technologies

Aggressive power management involving all system components - silicon technology, SoC design and software
Reductions in both active and static power



*New SmartReflex 2 Technologies

Source: TI

SmartReflex Power and Performance Management Technologies

- Silicon IP

- Retention SRAM and logic: SRAM and logic retention cells support dynamic power switching without state loss, lowering voltage and reducing leakage
- Dual gate lengths: longer for lower leakage and shorter for higher performance
- Power management cell library:
 - Switching, isolation and voltage shifting support multiple domains in SOC implementation.
 - Multiple domains => functional blocks can be powered down or put into a standby power mode when they are not active
- Process and temperature sensor: adapts voltage dynamically in response to silicon processes and temperature variations

SmartReflex Power and Performance Management Technologies

- Silicon IP continued
 - Design flow support: Complete, nonintrusive support for easily integrating SmartReflex technologies
 - Adaptive body biasing: modulates transistor bias voltage dynamically in order to optimize switching speed versus leakage. FBB and RBB together are called ABB
 - SmartReflex PriMer: Automates the implementation of SmartReflex power management techniques in the design and provides a UPF-compliant specification
 - SRAM retention till access (RTA): Reduces leakage while retaining contents in SRAM arrays by lowering the voltage on idle memory banks

SmartReflex Power and Performance Management Technologies

- SOC architecture and design technologies
 - **Adaptive voltage scaling:** Maintains high performance while minimizing voltage based on silicon process and temperature
 - **Dynamic Power Switching:** Dynamically switches between power modes based on system activity to reduce leakage power
 - **Dynamic voltage and frequency scaling:** Dynamically adjusts voltage and frequency to adapt to the performance required
 - **Multiple domains (voltage/power/clock):** Enables distinct physical domains for granular power/performance management by software
 - **Standby leakage management (SLM):** maintains lowest standby power mode compatible with required system responsiveness to reduce leakage power



SmartReflex Power and Performance Management Technologies

- System software
 - OS support : Provides an open environment for blending with operating systems and supports Symbian and Linux
 - Software power management framework: Intelligent control for power and performance management that is transparent to application programs and legacy code – Monitors system activity and not just processor activity
 - Workload monitoring and prediction: Determines system performance needs used to make intelligent power and performance management decisions
 - Policy and domain managers: Dynamically controls the system, providing the performance needed at the lowest power
 - DSP/BIOS software kernel foundation : power and performance management software for DSPs

Memory Interface Comparison

2010 Application Processor Comparison	
	Memory Interface
Apple A4	32-bit LPDDR1/LPDDR2 (?)
Intel Atom Z600	32-bit LPDDR1
TI OMAP 3430	32-bit LPDDR1
TI OMAP 4430	2 x 32-bit LPDDR2
NVIDIA Tegra 2	32-bit LPDDR2
Qualcomm Snapdragon QSD8250	32-bit LPDDR1

- A decade ago, at 400MHz, 32-bit LPDDR1 was our PC's memory bandwidth
- Intel's claim : Most of the smartphone workloads are compute-bound and not memory bandwidth bound. So the reduction in memory bandwidth isn't going to be an issue.
- Nvidia's claim: Narrower memory bus with more efficient arbitration logic is the best balance for power/performance at the 40nm process node

Source: Anandtech.com

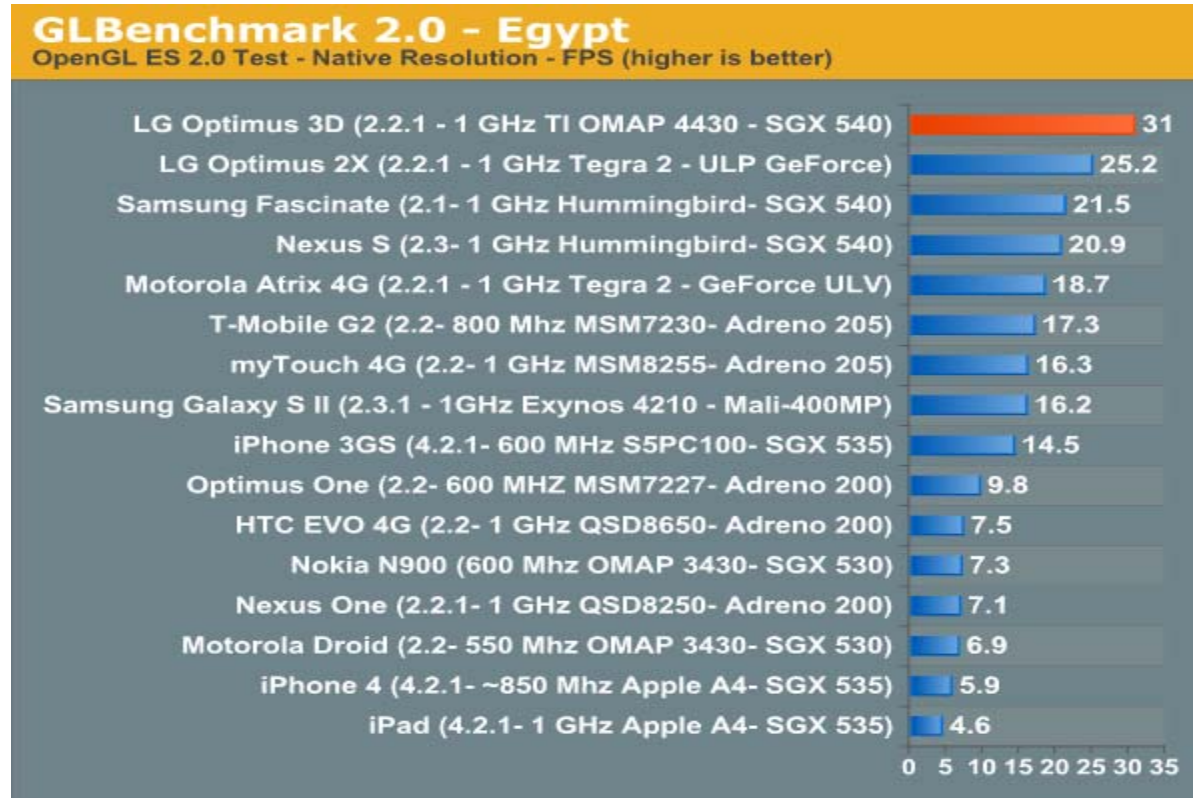
EFFECT of ARM NEON SIMD ENGINE

Minimum Instruction Latencies (Single Precision)						
Instruction	FADD	FSUB	FMUL	FMAC	FFDIV	FSQRT
ARM Cortex A8 (FPU)	9 cycles	9 cycles	10 cycles	18 cycles	20 cycles	19 cycles
ARM Cortex A9 (FPU)	4 cycles	4 cycles	5 cycles	8 cycles	15 cycles	17 cycles
ARM Cortex A8 (NEON)	1 cycle	1 cycle	1 cycle	1 cycle	N/A	N/A
ARM Cortex A9 (MPE/NEON)	1 cycle	1 cycle	1 cycle	1 cycle	10 cycles	13 cycles

- A9's FPUs are pipelined. Yet lower performance, compared to NEON MPE.
- Nvidia's claim: MPEs incur a 30% die penalty for a performance improvement that impacts only a minimal amount of code. They admit that they might integrate the SIMD Engine sometime later, but not now.

Source: Anandtech.com

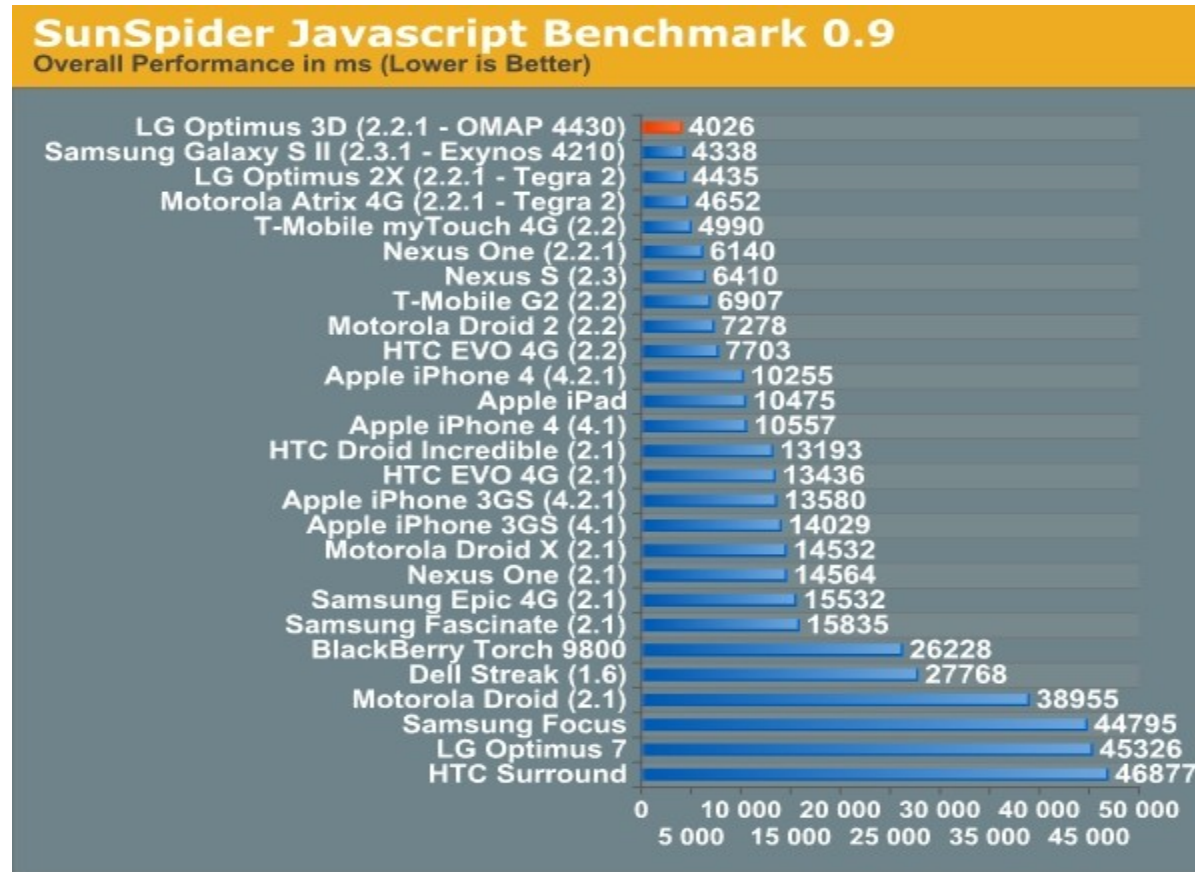
Graphics Benchmark



- LG Optimus 3D leads the pack.
- Could be benefited by the dual-channel LPDDR2 Memory Interface
- **Note:** LG Optimus3D and Samsung Galaxy S II are not final products, yet.

Source:
AnandTech.com

Browser benchmark



Source: AnandTech.com

References

- *Brian Carlson and Steve Jahnke, “Leveraging the Benefits of Symmetric Multiprocessing (SMP) in Mobile Devices”*
- *Scott Eisenhart and Robert Tolbert, “Designing for the Use Case: Using the OMAP™ 4 platform to overcome the challenges with integrating multiple applications”*
- *OMAP Mobile Applications Platform <http://www.ti.com/lit/swpt034>*
- *Brian Carlson and Bill Giolma, “SmartReflex™ Power and Performance Management Technologies”*
- *Bringing High End Graphics to Handheld Devices, NVIDIA whitepaper.*
- *Anandtech.com reviews*
 - *[TI OMAP4 vs Tegra-2](#)*
 - *[Memory Interface Comparison](#)*
 - *[Effect of ARM Neon SIMD Engine vs ARM FPU](#)*



Thank You!



Backup Slides



SmartReflex PriMer

- A user-friendly tool that simplifies design, reduces development time, and improves verification without compromising the flexibility needed to meet different product requirements
- It generates a UPF compliant specification, complete with power management features that include RTL descriptions of power domain insertion and protocol control, FBB/RBB and SRAM power management controllers, and a full suite of power management verification and assertion checks.



DVFS

- DVFS:
 - Clock rates and voltages are lowered in software depending on the performance required by the application
 - Software selects pre-defined processor operating performance points (OPPs), which include a voltage that ensures the processor runs at minimum frequency to meet system's processing requirements
 - OPPs are also predefined for interconnects and peripherals in processors
 - Corresponding to the given OPP, software sends control signals to external regulators in order to set the minimum voltage

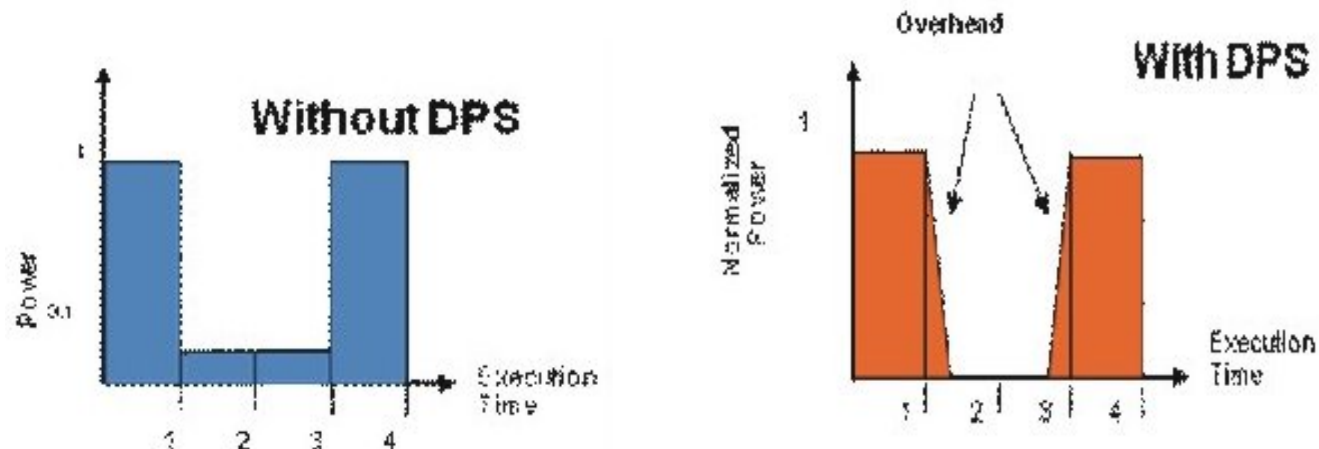


AVS

- Based on variations that come up during chip manufacturing as well as during a device's operational lifetime
- In contrast to DVFS all processors do not have same operating performance points (OPPs)
- Hot devices can achieve a given frequency with a lower voltage than can cold devices
- The processor senses its own performance level and adjusts voltage supply accordingly
- Dedicated on chip Voltage Scaling hardware implements a feedback loop, which does not require processor intervention, that dynamically optimizes voltage levels to account for variations in process results, temperature and silicon degradation
- Software sets up the AVS hardware for each OPP, and the control also sends commands to external voltage regulators, to lower the appropriate regulator's output in incremental steps until the processor just exceeds target-frequency

DPS

- DPS determines when a device has completed its current computational tasks and, if it's not needed at the moment, then puts the device into a low power state. On wakeup the processor returns to its normal state in a matter of microseconds





SLM

- Static power consumption management or passive power management
 - Keeping an idle system in a power efficient state until more processing is required
 - This uses SLM – which relies on several low power modes from standby to power off
 - Standby mode – device retains internal memory and logic, whereas in device-off mode all system states are saved in external memory
 - Wakeup time is faster than cold boot as program is already loaded in internal/external memory and no OS starting overhead

TBDR

- In order to render, the display is split into rectangular sections in a grid pattern. Each section is known as a tile. Associated with each tile is a list of the triangles that visibly overlap that tile. Each tile is rendered in turn to produce the final image.
- As the polygon generating program feeds triangles to the PowerVR (driver), it stores them in memory in a triangle strip. Polygon rendering is not performed until all polygon information has been collated for the current frame. Furthermore, the expensive operations of texturing and shading of pixels (or fragments) is delayed, whenever possible, until the surface visible at a pixel is determined — hence rendering is deferred.

TBDR continued

- In the early days of the PC GPU race deferred renderers were quite competitive. As geometry complexity in games increased, ATI and NVIDIA's immediate mode rendering + hidden surface removal proved to be the better option. Given the lack of serious 3D gaming, much less geometry heavy titles on smartphones today the tile based approach makes a lot of sense.
- Tile based renderers conserve both power and memory bandwidth. The rendering is limited to one tile at a time, the whole tile can be in fast onchip memory, which is flushed to video memory before processing the next tile. Under normal circumstances, each tile is visited just once per frame.