MultiCore DSP vs GPUs

Nissim Saban
Regional Application Manager
Texas Instruments
Nissim_Saban@ti.com
MultiCore DSP vs GPUs

- Today’s system developer’s needs
- Multicore DSP Architecture
- Shared GPU and DSP features
- Differences highlighted
- Summary
Today’s system developer’s needs

- **Processing power** (floating point)
  - Systems that use GPU/DSP need high processing power, usually floating point operations.
  - Today’s DSP/GPU can provide many GFLOPS

- **Connectivity**
  - Fast data movements (bandwidth)
    - GPU uses PCIe
    - DSP has PCIe, SRIO (20Gb/s), GMII (1Gb/s), HYPER LINK (50Gb/s)
  - Common Standards compliance (PCIe, SRIO, LAN, USB)

- **High level programming tools**
  - Ease of use and reuse (OPENCL, OPENMP, CUDA......)
  - Easy Platform migration

- **Size**

- **Price** $ 

The next slides discuss other parameters as well
Specialization (app oriented acceleration)

• When a specific well defined task is needed usually the silicon contains HW accelerators
• 2 examples follow in the next slides

• The HW accelerator has API that the controlling DSP can use to offload it’s computations
• **The accelerators are equivalent to many additional GFLOPS**
• Typical accelerators: Video Compress/Decompress, Communication protocols, FFT, Encryption............
App oriented acceleration - HD Video Transcode, Encode & Decode

1GHz - TMS320DM6467T Processor

Features
Core
- ARM926EJ-S™ (MPU) at 500 MHz
- TMS320C64x+™ DSP Core at 1 GHz

Memory
- ARM: 16K I-Cache, 8K D-Cache, 32K TCM RAM, 8K Boot ROM
- DSP: 32KB L1P Cache/SRAM, 32KB L1D Cache/SRAM, 128KB L2 Cache/SRAM, 64K Boot ROM

Video Encoder/Decoder Capabilities
- Real-Time HD-HD Transcoding Up to 1080p
  - Multi-format (mf) HD to mf HD or mf SD
  - Up to 2x real time for HD-to-SD Transcoder
  - Real-time HD-HD transcoding for PVR
- Video Encode and Decode
  - HD 1080p30 H.264 BP encode
  - HD 1080p60 H.264 BP decode
  - Dual HD 1080i60/p30 MPEG-2 MP@HL decoding
  - Simultaneous 720p30 H.264 BP encode and 1080p30 BP/HP decode
  - Simultaneous 720p60 H.264 BP encode and BP decode

Benefits
- Scalable video engine building on high-performance C64x+ media DSP, low-cost local controllers, and rich suite of multi-format video accelerators

Applications
- Transcoding (HD-HD, HD-SD) HD-Video Conferencing, HD- IP
- Set-Top Boxes, Digital Media Adapters, Video Surveillance, Medical Imaging

bak
Nyquist (C6670) with accelerators

- SoC for Baseband
  - 4x Performance of Faraday
  - RTOS, Linux and GCC
- Fixed/Floating C66x™ CorePac
  - Eight C66x DSP cores @ > 1.0 GHz
  - 1 MB Local L2
  - >128 GMAC, >64 GFLOP
  - 2.0 MB shared memory
- Navigator
  - Make Multi-core easy
  - Packet Infrastructure
- Wireless Acceleration for Layer 1 & Layer 2
  - Enable 2 cells on Single Chip
  - Chip architecture tailor made for 3G/4G packet data
- Network Processing Engine
  - IP Fast Path in hardware
- Low Power Consumption
  - Adaptive Voltage Scaling
  - Various and many Power saving techniques
- Hyperlink
  - Expansion port
  - Transparent to Software
- Terabit Switch
  - Reduces bottlenecks and unleashes full performance of the chip
- Multicore Debugging
Power (cooling, reliability....)

- The DSP is general purpose and intended to be used in embedded system also
- Power is a major concern today to system designers and has many impacts on price, reliability
- DSPs are available in industrial temperature range

- Next slide shows a Power/Performance GPP/GPU/DSP comparison example
GeForce GTX 580

Note: The below specifications represent this GPU as incorporated into NVIDIA's reference graphics card design. Graphics card specifications may vary by Add-in-card manufacturer. Please refer to the Add-in-card manufacturers' website for actual shipping specifications.

**GPU Engine Specs:**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>CUDA Cores</td>
<td>512</td>
</tr>
<tr>
<td>Graphics Clock (MHz)</td>
<td>772 MHz</td>
</tr>
<tr>
<td>Processor Clock (MHz)</td>
<td>1544 MHz</td>
</tr>
<tr>
<td>Texture Fill Rate (billion/sec)</td>
<td>49.4</td>
</tr>
</tbody>
</table>

**Memory Specs:**

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Memory Clock (MHz)</td>
<td>2004</td>
</tr>
<tr>
<td>Standard Memory Config</td>
<td>1536 MB GDDR5</td>
</tr>
<tr>
<td>Memory Interface Width</td>
<td>384-bit</td>
</tr>
<tr>
<td>Memory Bandwidth (GB/sec)</td>
<td>192.4</td>
</tr>
</tbody>
</table>

**Feature Support:**

**GeForce GTX 580**

### Standard Graphics Card Dimensions:

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>4.376 inches (111 mm)</td>
</tr>
<tr>
<td>Length</td>
<td>10.5 inches (267 mm)</td>
</tr>
<tr>
<td>Width</td>
<td>Dual-Slot</td>
</tr>
</tbody>
</table>

### Thermal and Power Specs:

<table>
<thead>
<tr>
<th>Specification</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Maximum GPU Temperature (in C)</td>
<td>97 C</td>
</tr>
<tr>
<td>Graphics Card Power (W)</td>
<td>244 W</td>
</tr>
<tr>
<td>Minimum Recommended System Power (W)</td>
<td>600 W</td>
</tr>
<tr>
<td>Supplementary Power Connectors</td>
<td>One 6-pin and One 8-pin</td>
</tr>
</tbody>
</table>

1. A GeForce GTX 580 GPU must be paired with another GeForce GTX 580 GPU (graphics card manufacturer can be different). SLI requires sufficient system cooling and a compatible power supply. Visit [www.slizone.com](http://www.slizone.com) for more information and a listing of SLI-Certified components.
2. NVIDIA 3D Vision Surround require two or more graphics cards in NVIDIA SLI configuration, 3D Vision glasses and three matching 3D...
## FFT benchmark comparison with GPP/GPU

<table>
<thead>
<tr>
<th>Platform</th>
<th>Effective Time to complete 1024 point complex to complex FFT (single precision), μs</th>
<th>Power (Watts)</th>
<th>Energy per FFT (μJ)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GPU: nVidia Tesla C2070</td>
<td>0.16</td>
<td>225</td>
<td>36</td>
</tr>
<tr>
<td>GPU: nVidia Tesla C1060</td>
<td>0.3</td>
<td>188</td>
<td>56.4</td>
</tr>
<tr>
<td>GPP: Intel Xeon Core Duo @ 3 GHz¹</td>
<td>1.8</td>
<td>95</td>
<td>171</td>
</tr>
<tr>
<td>GPP: Intel Nehalam Quad Core @ 3.2 GHz¹</td>
<td>1.2</td>
<td>130</td>
<td>156</td>
</tr>
<tr>
<td>DSP: TI C6678 @ 1.2 GHz¹</td>
<td>0.86</td>
<td>10</td>
<td>8.6</td>
</tr>
</tbody>
</table>

¹For GPP and DSP, we show the effective time of completion assuming that multiple FFTs are run in parallel to the multiple cores available

- **When performance is distributed over power**
  - DSP >4x better than GPU
  - DSP >18x better than GPP

### GPP/GPU information sources:
Real time systems
(HW connectivity , controlled response timing.....)

• The typical system we are discussing has very fast inputs and outputs
• The system has to meet real time constraint as frame rate , communications , latency
• DSP operating system and HW architecture enables real time reliable response
• Following slides show typical connections of GPU and DSP in real world
• The DSP is designed to connect easily to FPGA and other HW inputs/outputs
GPU environment
DSP environment

Please note connectivity to FPGA and to external buses
DSP Architecture

• The next slide shows the world of embedded processing today 16,32 bits, Accelerators, Multicores
• A slide with the internal architecture of the Multicore DSP follows
• The DSP has 8 functional units that can run 8 instructions per cycle
• In term of “CUDA cores” each DSP has 16 “cores”
• In C6678 the HYPER LINK enables to connect DSPs in pairs so you can get 32 cores @1.2GHZ
## Embedded processing portfolio

### TI Embedded Processors

<table>
<thead>
<tr>
<th>Microcontrollers (MCUs)</th>
<th>ARM®-Based Processors</th>
<th>Digital Signal Processors (DSPs)</th>
</tr>
</thead>
<tbody>
<tr>
<td>16-bit ultra-low power MCUs</td>
<td>32-bit real-time MCUs</td>
<td>32-bit ARM Cortex™-M3 MCUs</td>
</tr>
<tr>
<td>ARM Cortex-A8 &amp; ARM9™ MPUs</td>
<td>DSP DSP+ARM</td>
<td>Multi-core DSP</td>
</tr>
<tr>
<td>Ultra Low power DSP</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### MSP430™
- **Up to 25 MHz**
- Flash: 1 KB to 256 KB
- Analog I/O, ADC, LCD, USB, RF
- Measurement, Sensing, General Purpose
- $0.25 to $9.00

### C2000™
- **40MHz to 300 MHz**
- Flash, RAM: 16 KB to 512 KB
- PWM, ADC, CAN, SPI, I²C
- $1.50 to $20.00

### Stellaris®
- **Up to 100 MHz**
- Flash: 8 KB to 256 KB
- USB, ENET, MAC+PHY CAN, ADC, PWM, SPI
- Connectivity, Security, Motion Control, HMI, Industrial Automation
- $1.00 to $8.00

### Sitara™
- **375MHz to >1GHz**
- Cache, RAM, ROM
- USB, CAN, SATA, SPI, PCIe, EMAC
- Industrial automation, POS & portable data terminals
- $5.00 to $25.00

### C6000™
- **320 GMACS**
- Cache RAM, ROM
- USB, ENET, PCIe, SATA, SPI
- Floating/ Fixed Point
- $4.00 to $200.00

### OMAP™
- **300MHz to >1Ghz +Accelerator**
- Cache RAM, ROM
- USB, ENET, PCIe, SATA, SPI
- Industrial automation, POS & portable data terminals
- $40.00 to $200.00

### C5000™
- **Up to 300 MHz +Accelerator**
- Up to 320KB RAM
- Up to 128KB ROM
- USB, ADC, McBSP, SPI, I²C
- Audio, Voice
- Medical, Biometrics
- $3.00 to $10.00

### Software & Dev. Tools

- MPUs – Microprocessors

---

[Image: Texas Instruments]
TMX320C6678/4/2 processors are ideal for

Applications such as

- Mission Critical
  - Military
  - Avionics
  - Public Safety
- Medical Imaging
  - Ultrasound
  - Endoscopy
  - MRI / CT Scan
  - Emerging modalities
- Test & Automation
  - Wafer/LCD inspection
  - Niche printing/scanning
- Emerging Video

Key Advantages

- Support for fixed & floating point processing
- Focused security & network coprocessors
- Multiple high speed peripherals
- Enhanced memory architecture with large on chip memory
- New multicore navigator for fast data transfers
- Switch fabric with 2 terabits of bandwidth
- New multicore focused debug and performance profiling capabilities
Multicore High Performance DSP

**C66x Core**
Next generation Fixed / Floating-Point DSP core with clock speeds ranging from 1GHz – 1.25GHz and Up to 8 core options

**Memory Architecture**
Up to 8MB of combined memory with an enhanced memory architecture that allows fast on- and off-chip memory access including a DDR3-1600MHz (64-bit) interface (8GB of addressable space).

**Multicore Navigator**
Data transfer engine that is architected to move data between various system elements without using any CPU overhead so maximum system efficiency is achieved

**Network Co-Processor and Accelerators**
A cost effective implementation to off-load the wireless and secure networking functions from the DSP

**TeraNet**
Switch fabric that has 2 Terabits of bandwidth which allows maximum data transfer between system components to realize full system entitlement

**Peripherals and I/O Interfaces**
High bandwidth peripherals that operate independently (NOT Shared) allowing simultaneous data transfer to prevent bottle necks - featuring:
- RapidIO v2.1 – 4lanes @ 5Gbps with 1x and 4x support
- PCIe x2 – 2lanes, running independently of RapidIO

**HyperLink**
Ultra high-speed low latency serial interface that connects to other DSPs and FPGAs in the systems
Application SW

The obstacle to get fast performance?

• In real life the applications can be easy to parallel or they may demand serial processing
• GPU merit is the high parallelism
• DSP merit is the flexibility

• The next slides show examples of App and the tools embedded in DSP for inter processor and tasks communication that are crucial to real world Apps
Master / Slave Processing Model

- Many small independent threads that fit easily within the processing resources of a single core
- High parallelism
Data Flow Processing Model

- One or more tasks to be mapped to each core.
- Synchronization using message passing between cores.
- Data passed between cores using shared memory or DMA transfers.

- See next slide for inter communication option in the DSP.
Architecture challenges

IPC = inter processor communication

Data Flow Diagram Example
Many communication interactions needed
# Summery

## DSP
- Power
- Real time
- Stand alone (no CPU)
- Versatility (general purpose)
- Embedded use

## GPU
- Extreme Performance
- Special design (graphics)
- Parallelism (if application allows)
- High level programming

When designing a system some key points above can help select between a GPU and a DSP.

- Price/Performance of either system choice is unique to the application and needs to be discussed individually.

- The application’s needs will drive the choice.