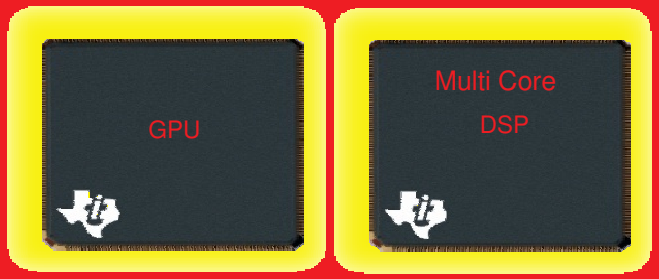
MultiCore DSP vs GPUs

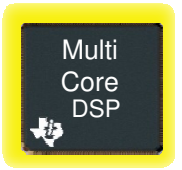
Nissim Saban
Regional Application Manager
Texas Instruments
Nissim_Saban@ti.com

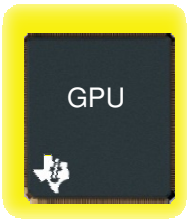


Texas Instruments

MultiCore DSP vs GPUs

- Today's system developer's needs
- Multicore DSP Architechture
- Shared GPU and DSP features
- Differences highlighted
- Summery





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

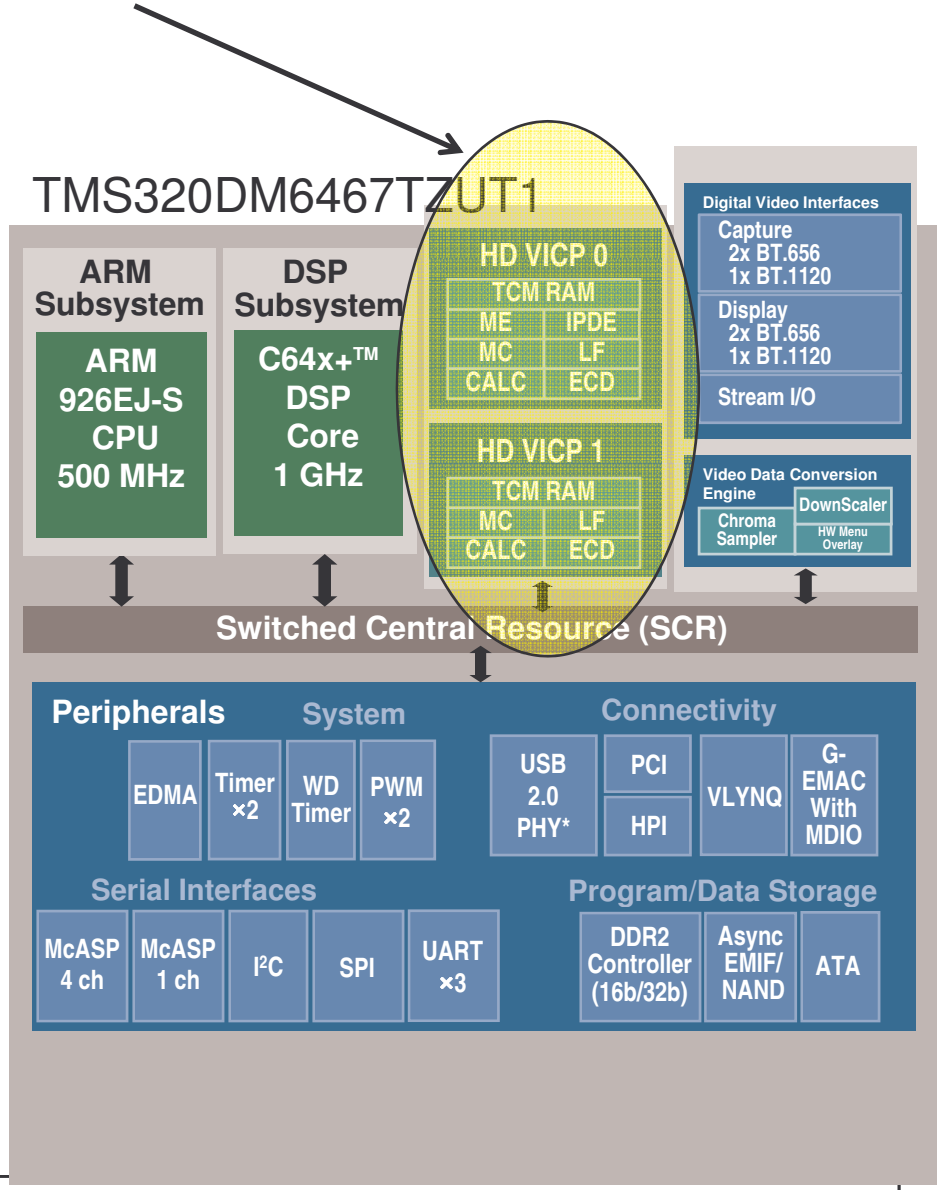
Core
- ARM926EJ-S™ (MPU) at 500 MHz
- TMS320C64x+™ DSP Core at 1 GHz

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 2× 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

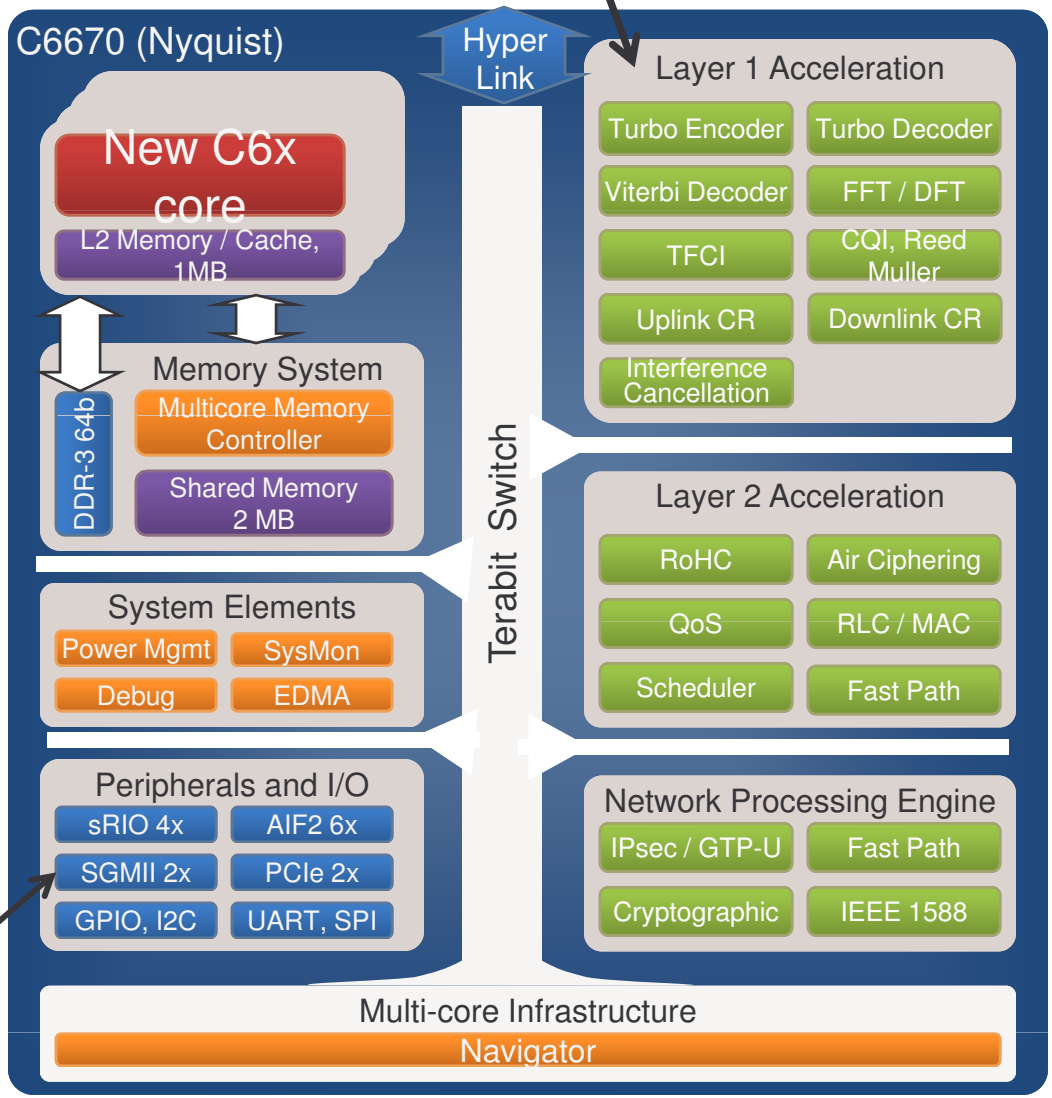


bak

Texas Instruments

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



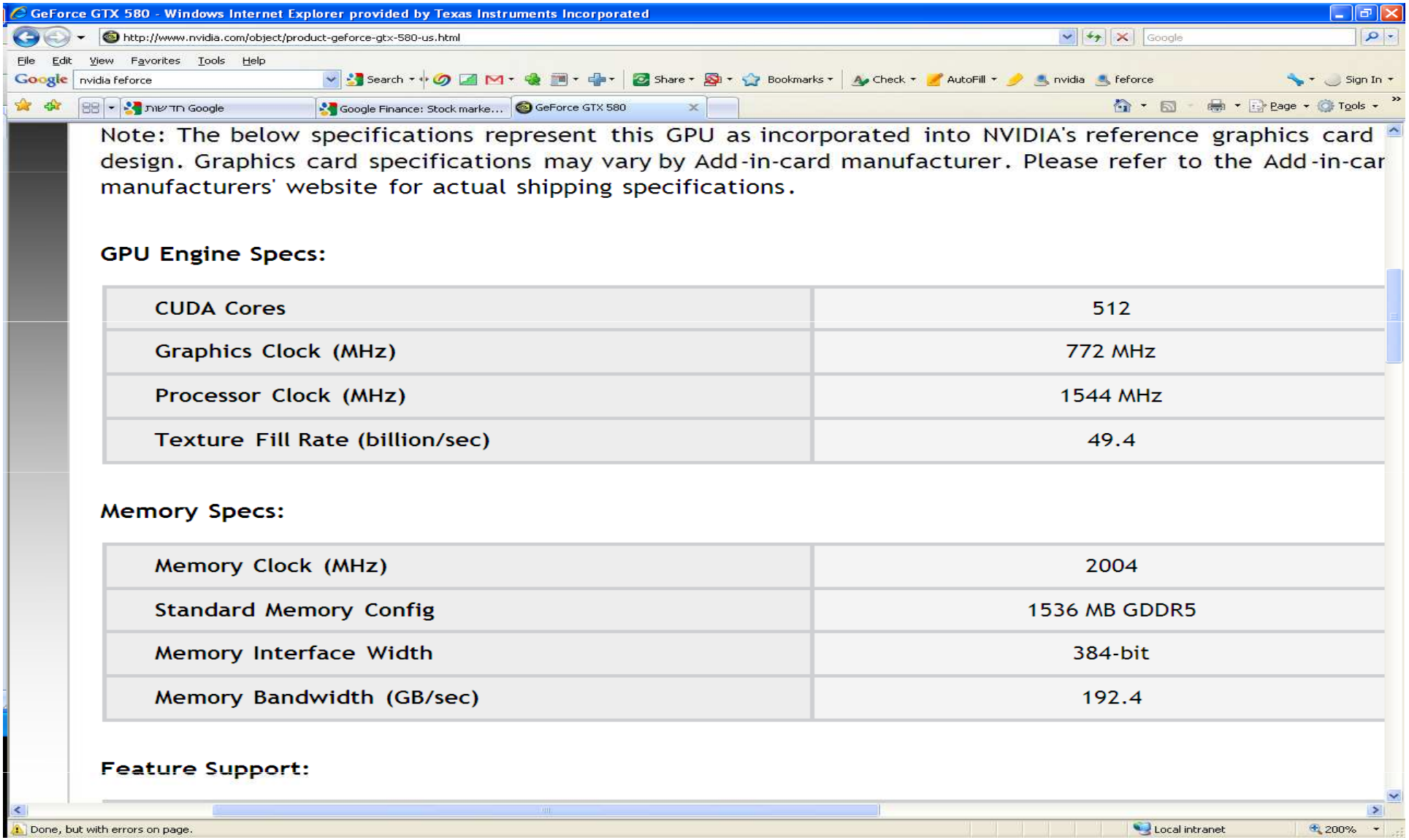
connectivity

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-car manufacturers' website for actual shipping specifications.

GPU Engine Specs:

CUDA Cores	512
Graphics Clock (MHz)	772 MHz
Processor Clock (MHz)	1544 MHz
Texture Fill Rate (billion/sec)	49.4

Memory Specs:

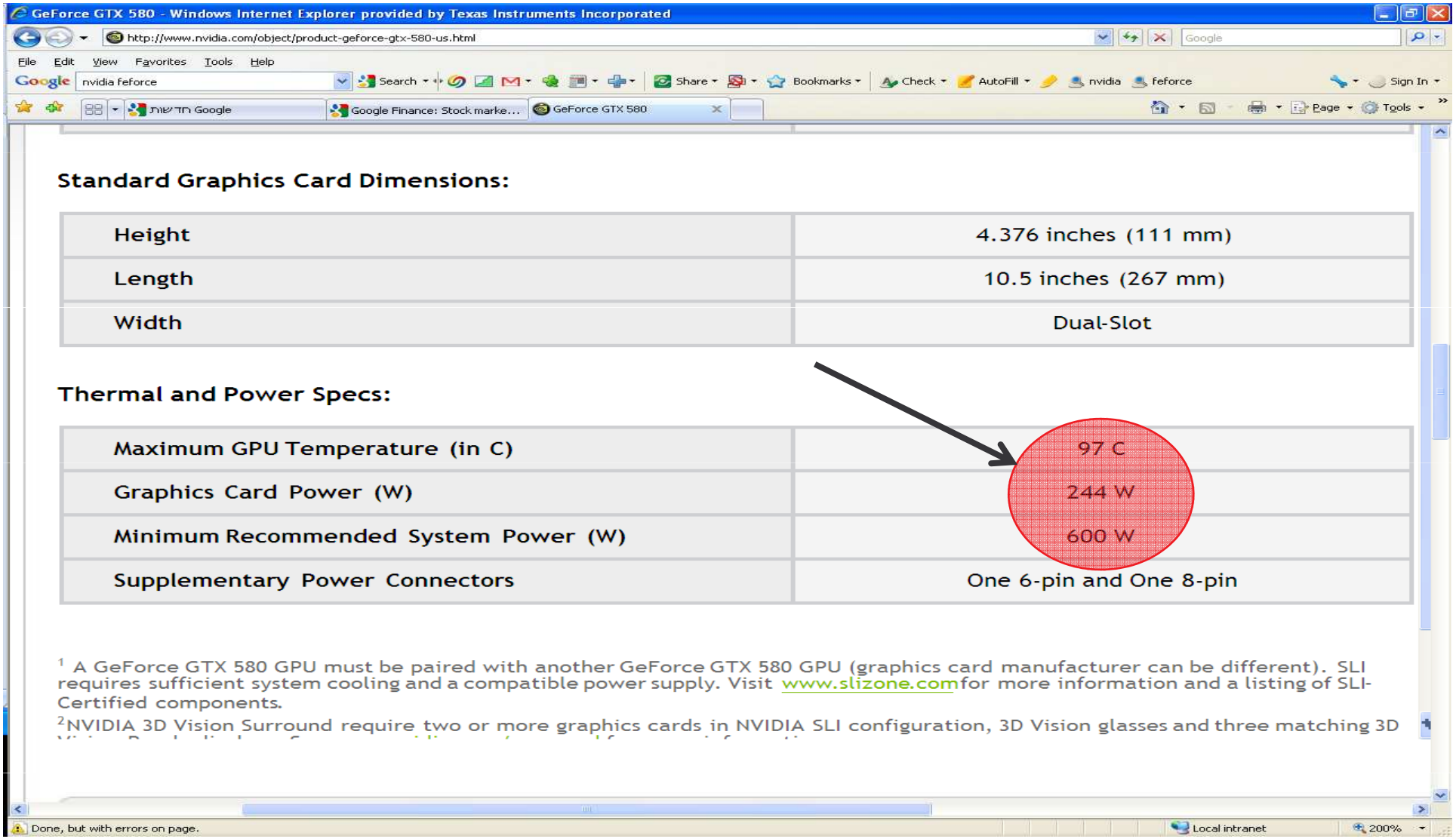
Memory Clock (MHz)	2004
Standard Memory Config	1536 MB GDDR5
Memory Interface Width	384-bit
Memory Bandwidth (GB/sec)	192.4

Feature Support:

Source : http://www.nvidia.com/object/product-geforce-gtx-580-us.html

Texas Instruments

GeForce GTX 580



Standard Graphics Card Dimensions:

Height	4.376 inches (111 mm)
Length	10.5 inches (267 mm)
Width	Dual-Slot

Thermal and Power Specs:

Maximum GPU Temperature (in C)	97 C
Graphics Card Power (W)	244 W
Minimum Recommended System Power (W)	600 W
Supplementary Power Connectors	One 6-pin and One 8-pin

[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 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

Platform	Effective Time to complete 1024 point complex to complex FFT (single precision), µs	Power (Watts)	Energy per FFT (µJ)
GPU: nVidia Tesla C2070	0.16	225	36
GPU: nVidia Tesla C1060	0.3	188	56.4
GPP: Intel Xeon Core Duo @ 3 GHz[1]	1.8	95	171
GPP: Intel Nehalam Quad Core @ 3.2 GHz[1]	1.2	130	156
DSP: TI C6678 @ 1.2 GHz[1]	0.86	10	8.6

[1]For 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

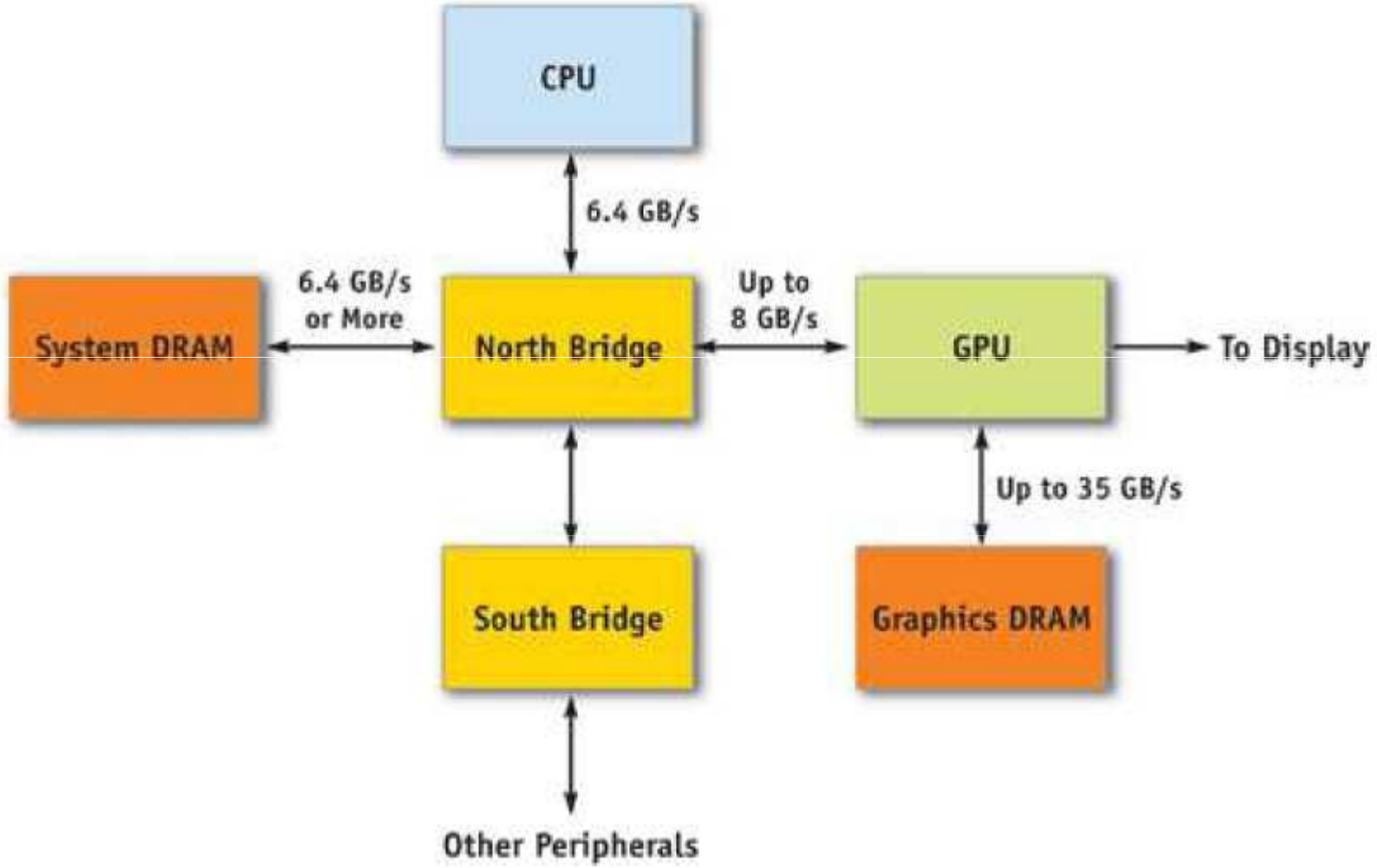
GPP/GPU information sources:

- *http://www.microway.com/pdfs/TeslaC2050-Fermi-Performance.pdf*
- *http://www.fftw.org/speed/CoreDuo-3.0GHz-icc/*

Real time systems
(HW connectivity ,controlled response timing…..)

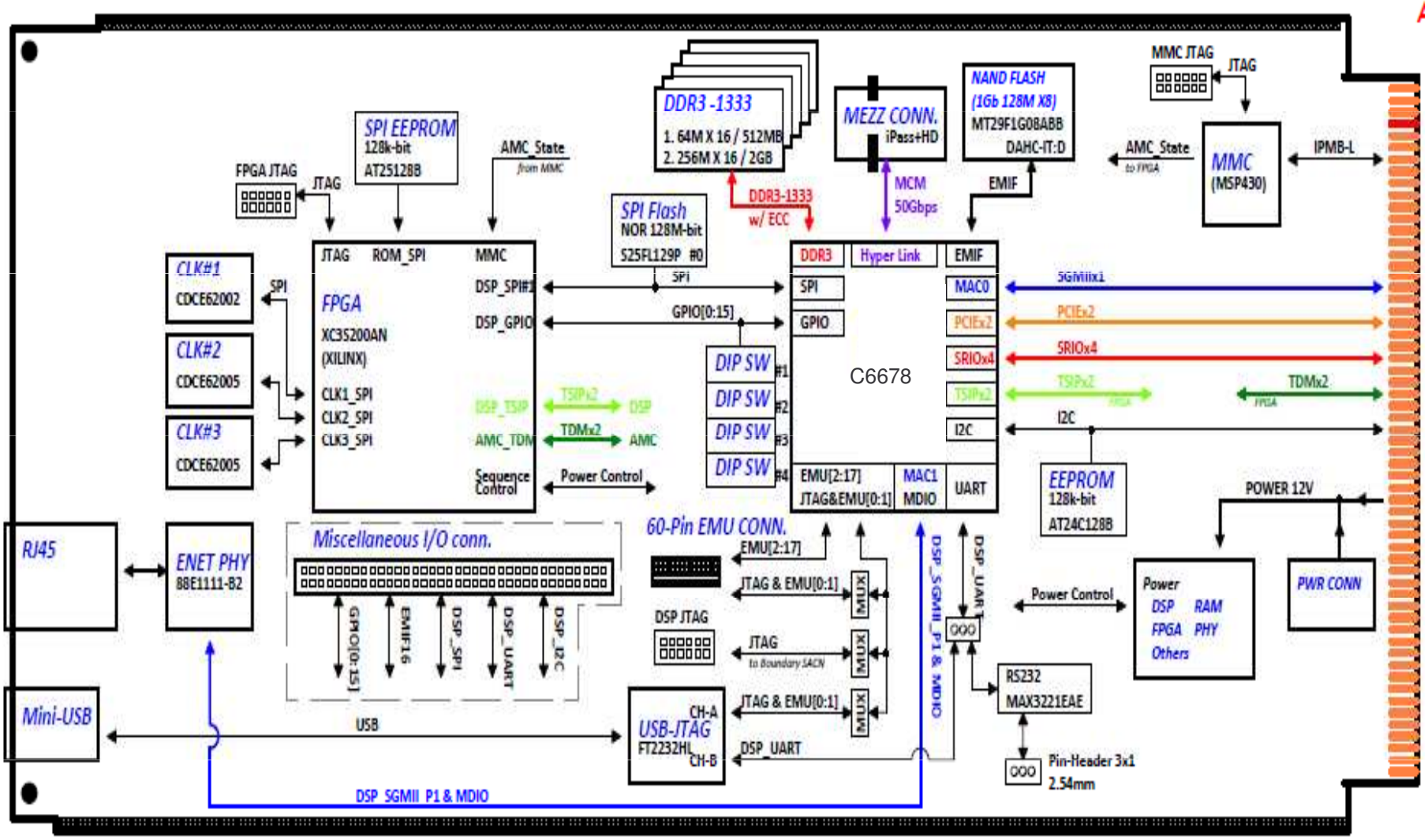
- The typical system we are discussing has very fast inputs and outputs
- The system has to meet real time constraint as frame rate , communications , latency
- DSP operating system and HW architecture enables real time reliable response
- Following slides show typical connections of GPU and DSP in real 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 architechture 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

Microcontrollers (MCUs)			ARM®-Based Processors		Digital Signal Processors (DSPs)	
16-bit ultra-low power MCUs	32-bit real-time MCUs	32-bit ARM Cortex™-M3 MCUs	ARM Cortex-A8 & ARM9™ MPUs	DSP DSP+ARM	Multi-core DSP	Ultra Low power DSP
MSP430™	C2000™ Delfino™ Piccolo™	Stellaris® ARM Cortex-M3	Sitara™ ARM Cortex-A8 & ARM9	C6000™ DaVinci™ video processors OMAP™	C6000™	C5000™
Up to 25 MHz	40MHz to 300 MHz	Up to 100 MHz	375MHz to >1GHz	300MHz to >1Ghz +Accelerator	320 GMACS	Up to 300 MHz +Accelerator
Flash 1 KB to 256 KB	Flash, RAM 16 KB to 512 KB	Flash 8 KB to 256 KB	Cache, RAM, ROM	Cache RAM, ROM	Cache RAM, ROM	Up to 320KB RAM Up to 128KB ROM
Analog I/O, ADC LCD, USB, RF	PWM, ADC, CAN, SPI, I²C	USB, ENET MAC+PHY CAN, ADC, PWM, SPI	USB, CAN, SATA, SPI, PCIe, EMAC	USB, ENET, PCIe, SATA, SPI	SRIO, EMAC DMA, PCIe	USB, ADC McBSP, SPI, I²C
Measurement, Sensing, General Purpose	Motor Control, Digital Power, Lighting, Ren. Energy	Connectivity,Security, Motion Control, HMI, Industrial Automation	Industrial automation, POS & portable data terminals	Floating/Fixed Point Video, Audio, Voice, Security, Conferencing	Telecom test & meas, media gateways, base stations	Audio, Voice Medical, Biometrics
$0.25 to $9.00	$1.50 to $20.00	$1.00 to $8.00	$5.00 to $25.00	$5.00 to $200.00	$40.00 to $200.00	$3.00 to $10.00





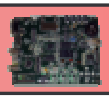






Texas Instruments Developer Network


Software & Dev. Tools

Texas Instruments expressDSP

MPUs – Microprocessors

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







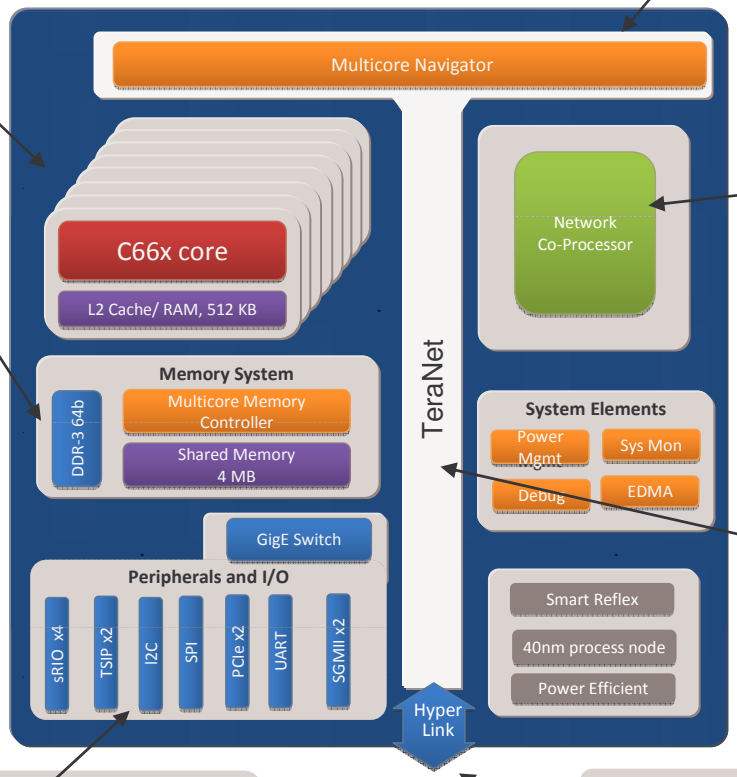
Key Advantages

- **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

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

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).

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
- PCIex2 – 2lanes, running independently of RapidIO

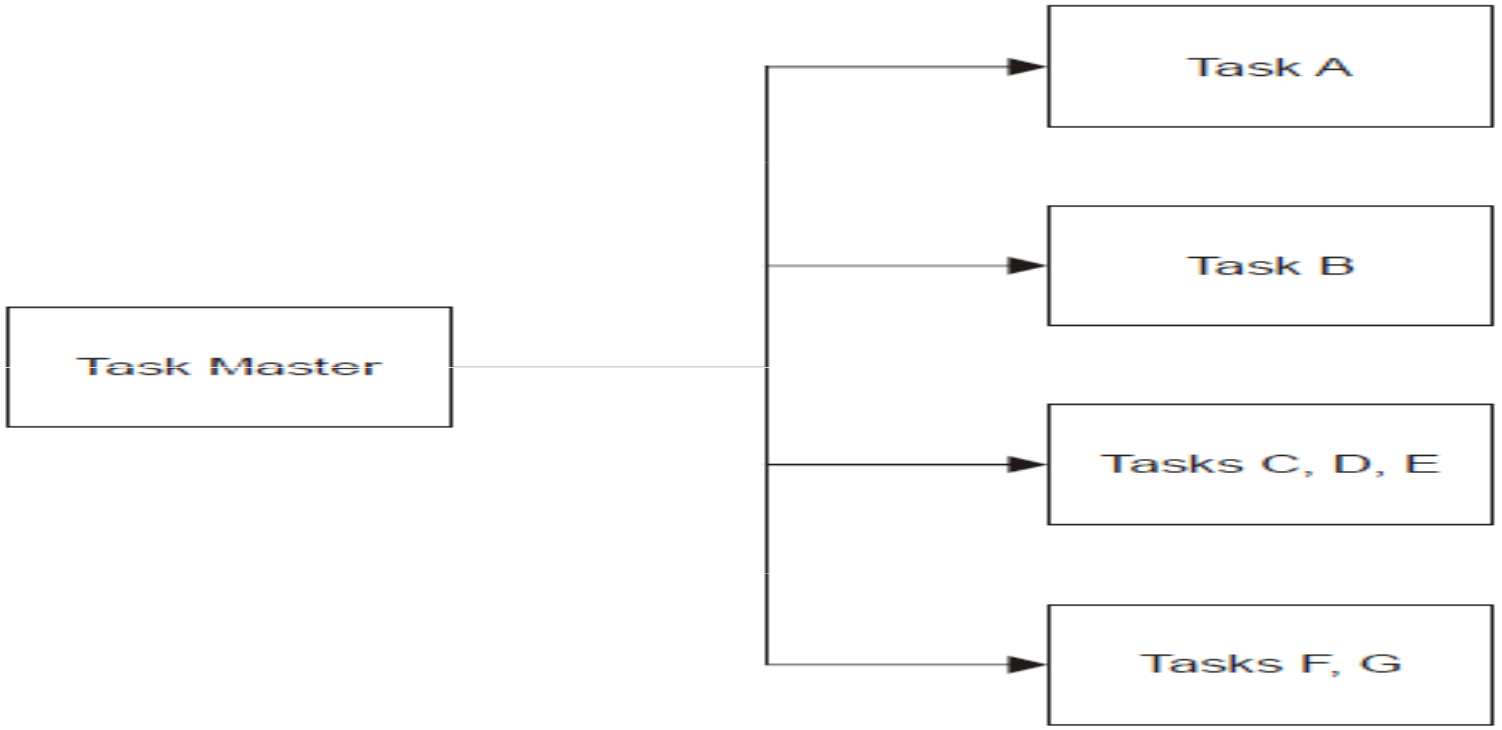
HyperLink
Ultra high-speed low latency serial interface that connects to other DSPs and FPGAs in the systems

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

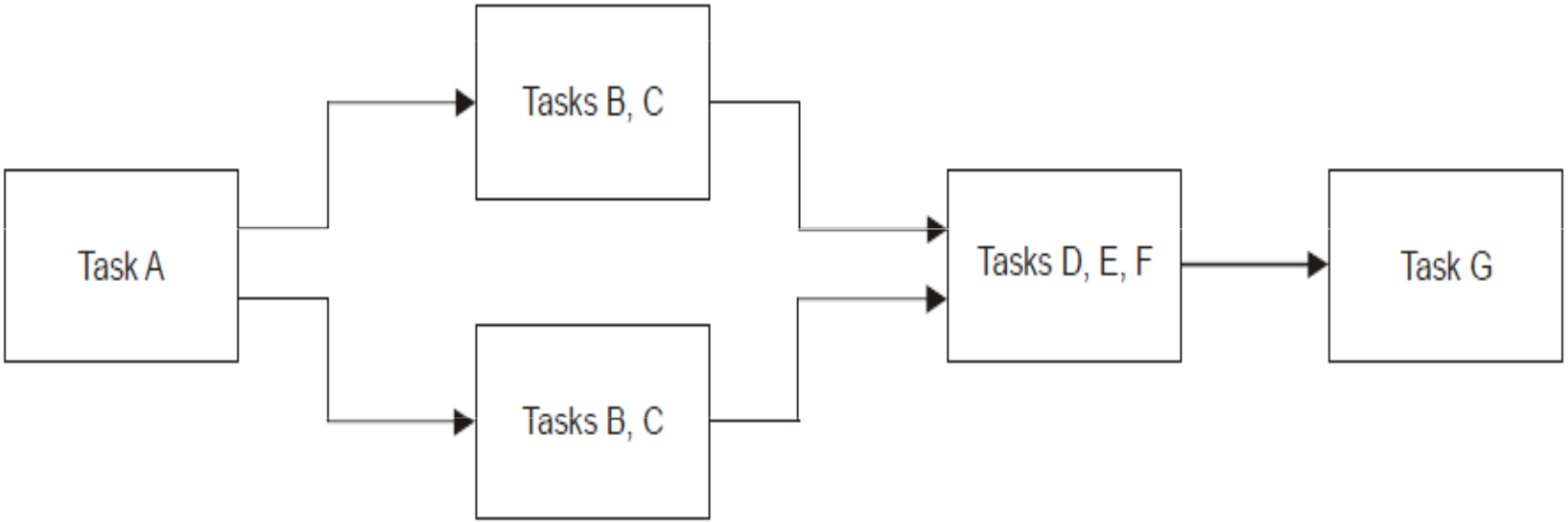
- The next slides show examples of App and the tools embedded in DSP for inter processor and tasks communication that are crucial to real world Apps

Master / Slave Processing Model



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

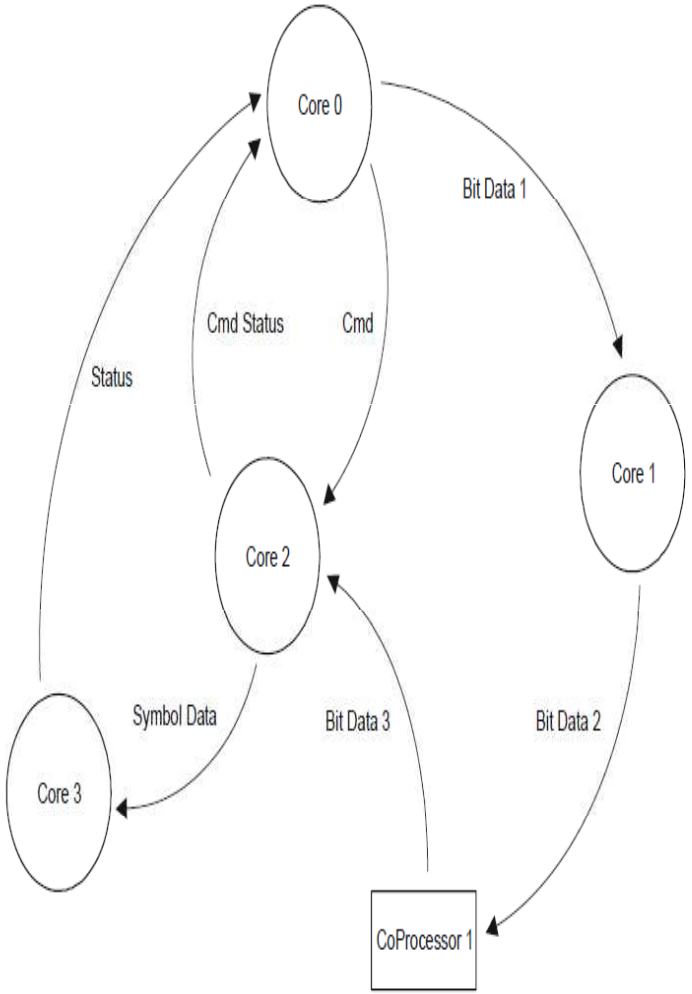
Data Flow Processing Model



- 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

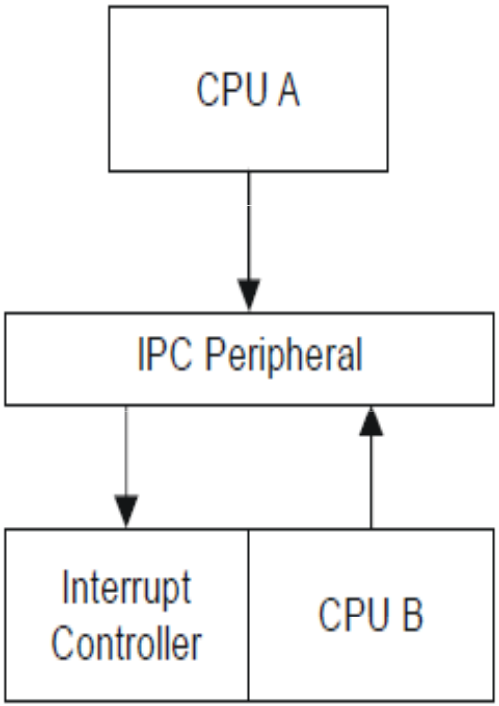
- See next slide for inter communication option in the DSP

Architecture challenges

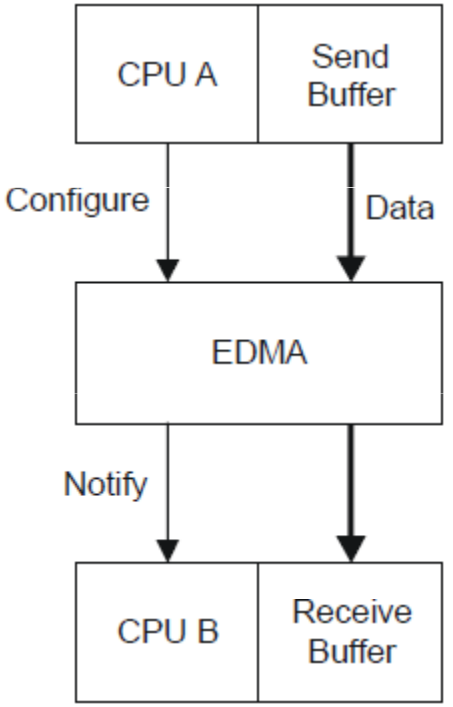


Data Flow Diagram Example
Many communication interactions needed

IPC = inter processor communication



Direct Signaling



Indirect Signaling

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