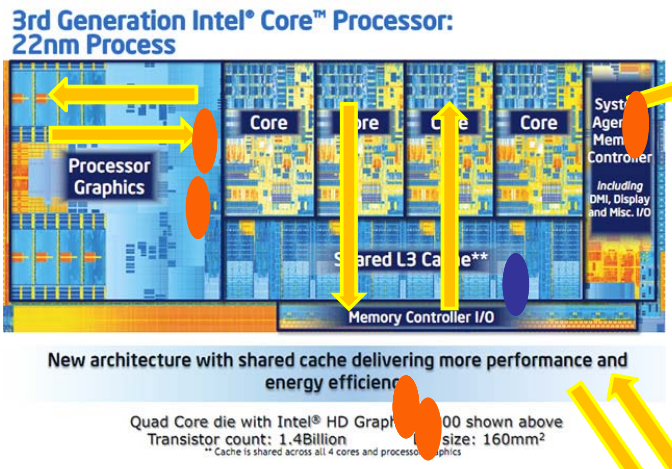
Microprocessor Technology:
Heterogeneous Multi-Core Processors

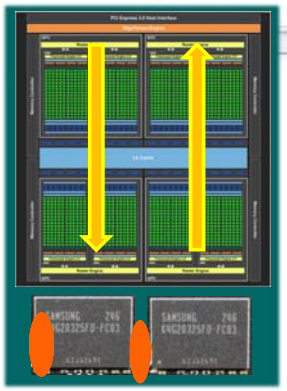
Won Woo Ro

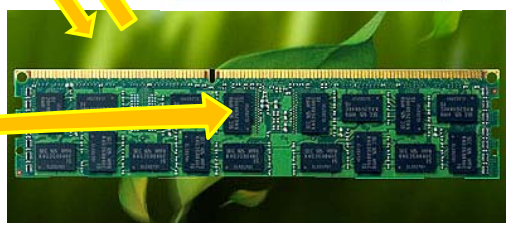
esCaL

Embedded Systems and Computer Architecture Lab.

Computer Systems







esCaL
Embedded Systems and Computer Architecture Lab.

Processor-Memory Systems

Cloud and Mobile Computing

2010 ~

Processor: Heterogeneous Multi-Core/AP
Memory: SSD/PRAM/MRAM/LPDDR/Wide I/O

Internet and Desktop Computing

2000 ~

Processor: Multi-Core/GPU
Memory: DDR/Cache

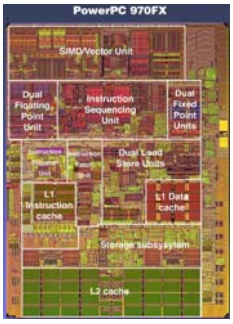
1990 ~

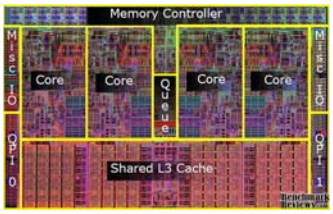
Processor: High Clock Rate, ILP
Memory: SDRAM

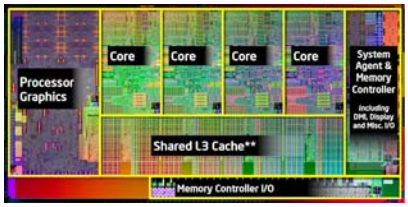
Server Client Computing

esCaL
Embedded Systems and Computer Architecture Lab.

The History of Breaking Walls







High Clock Rate
ILP & Cache

Multi-Core

Heterogeneous
Multi-Core

Thermal/ Power

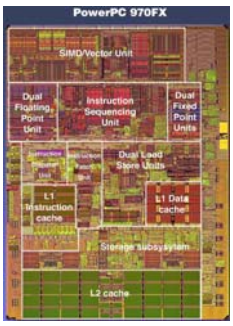
Dark Silicon

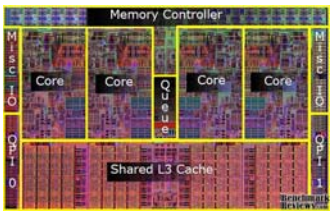
TLP

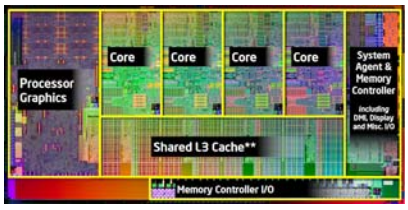
Efficiency
(energy/speed)

esCaL
Embedded Systems and Computer Architecture Lab.

Software Performance







Sequential Programing → Multitasking Parallel Programming → OpenCL / CUDA

Thermal/ Power

Dark Silicon

TLP

Efficiency (energy/speed)

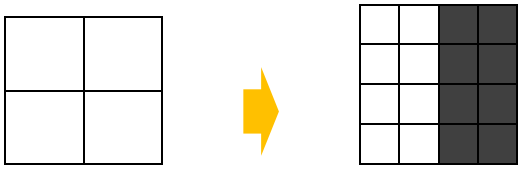
esCaL
Embedded Systems and Computer Architecture Lab.

DARK SILICON

esCaL
Embedded Systems and Computer Architecture Lab.

Dark Silicon

Dark Silicon: 최대 성능을 낼 때, 칩 내에서 그 성능에 기여하지 못하는 실리콘 면적



65nm, 4 Cores, 1.8 GHz

32nm, 8 Cores, >=1.8 GHz

Multi-Core Processor

프로그램 실행 시, 사용되지 않는 코어들의 면적 합

코어 스케일링 진행됨에 따라 Dark silicon 면적 증가

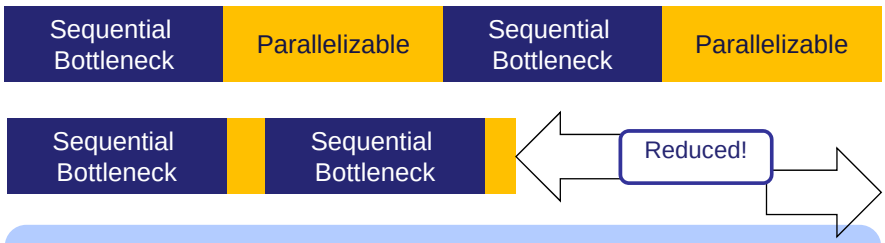
주요 원인	관련 이론
Parallelism Limit	Amdahl's Law
Power Limit	End of Dennard scaling

* Esmaeilzadeh et al., "Dark Silicon and the End of Multicore Scaling"

esCaL Embedded Systems and Computer Architecture Lab.

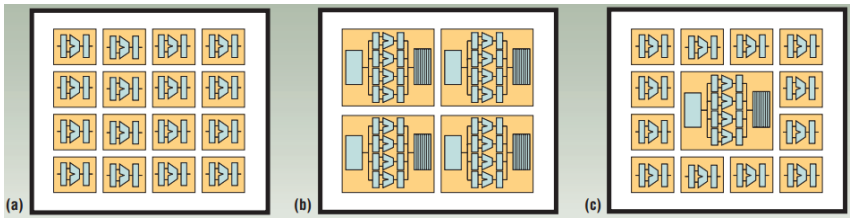
Parallelism의 제한

Amdahl's Law*: 순차 실행 시 또는 제한된 병렬 실행 시, 나머지 코어 → Dark silicon



Dark silicon 줄이기 위해 무조건 각 코어의 크기를 증가?

Software의 parallelism이 높으면 Dark silicon 감소



동일 자원으로 구성 가능한 멀티코어 예

* Hill, M. D. et al., Amdahl's Law in the Multicore Era

esCaL Embedded Systems and Computer Architecture Lab.

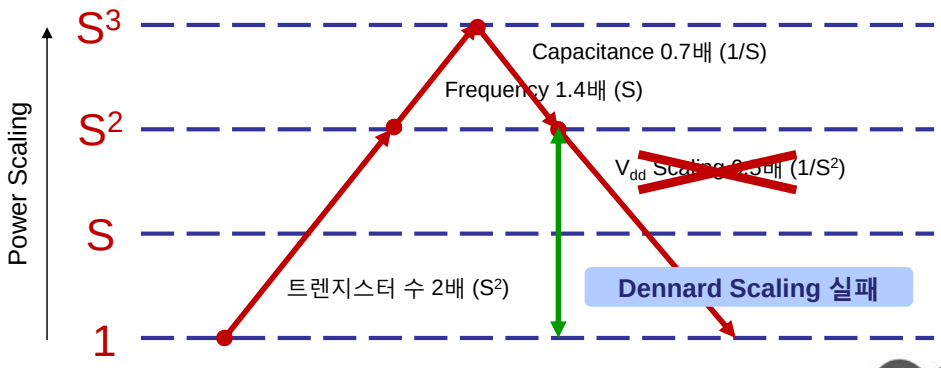
Power의 제한

Moore's Law:
매 18 개월 마다 Transistor 개수 2배씩 증가, Dennard scaling과 함께 multicore 스케일링의 기반 법칙

Dennard Scaling의 실패*

Transistor 수 2배 증가 → Technology 0.7배로 감소
(S : Technology scaling factor = 1/0.7 ≒ 1.4)

$$P_{total} = n \times \alpha C V^2 f$$



* Michael B. Taylor, "Is Dark Silicon Useful?: Harnessing the Four Horsemen of the Coming Dark Silicon Apocalypse"

esCaL
Embedded Systems and Computer Architecture Lab.

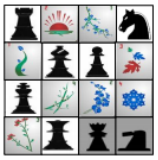
Dark Silicon 해결 위한 연구 방향

1. **칩 크기 축소**
   - 비용 문제 및 전력 밀도가 높아짐에 따라 온도 증가 문제

2. **Dim Silicon**
   - Core scaling시 전력 제한 극복 위해 under-clocking 방식 사용
   - 순간적으로 clock frequency 올려 성능 증폭

   예) Turbo Boost 2.0 (Intel), Computational Sprinting (HPCA'12), big.LITTLE Core (ARM)

3. **Specialized Hardware**
   - 특수 목적 유닛을 내장하여 필요 시 켜서 사용
   - 예) Greendroid, Heterogeneous Processor



4. **새로운 device 개발**
   - MOSFET의 한계를 극복할 수 있는 새로운 device 개발
   - Dennard scaling의 실패 근본 원인인 leakage 문제 해결

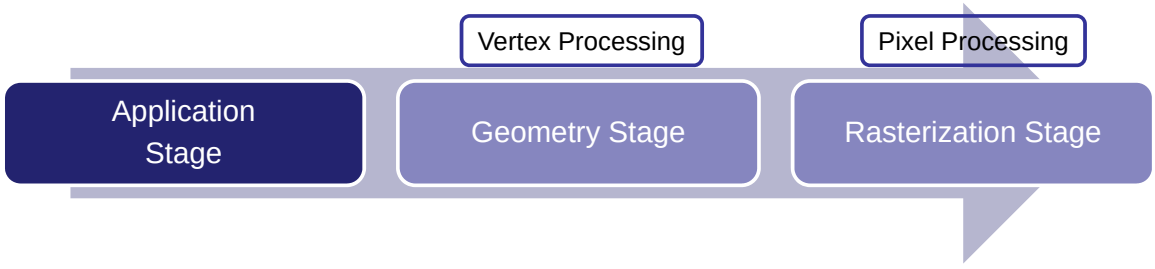
* Michael B. Taylor, "Is Dark Silicon Useful?: Harnessing the Four Horsemen of the Coming Dark Silicon Apocalypse"

esCaL
Embedded Systems and Computer Architecture Lab.

GPU

esCaL
Embedded Systems and Computer Architecture Lab.

GPU – *Graphics only before…*



A **graphics processing unit** (**GPU**), also occasionally called **visual processing unit** (**VPU**), is a specialized **electronic circuit** designed to rapidly manipulate and alter memory to accelerate the building of images in a **frame buffer** intended for output to a display.

- from Wiki

esCaL
Embedded Systems and Computer Architecture Lab.

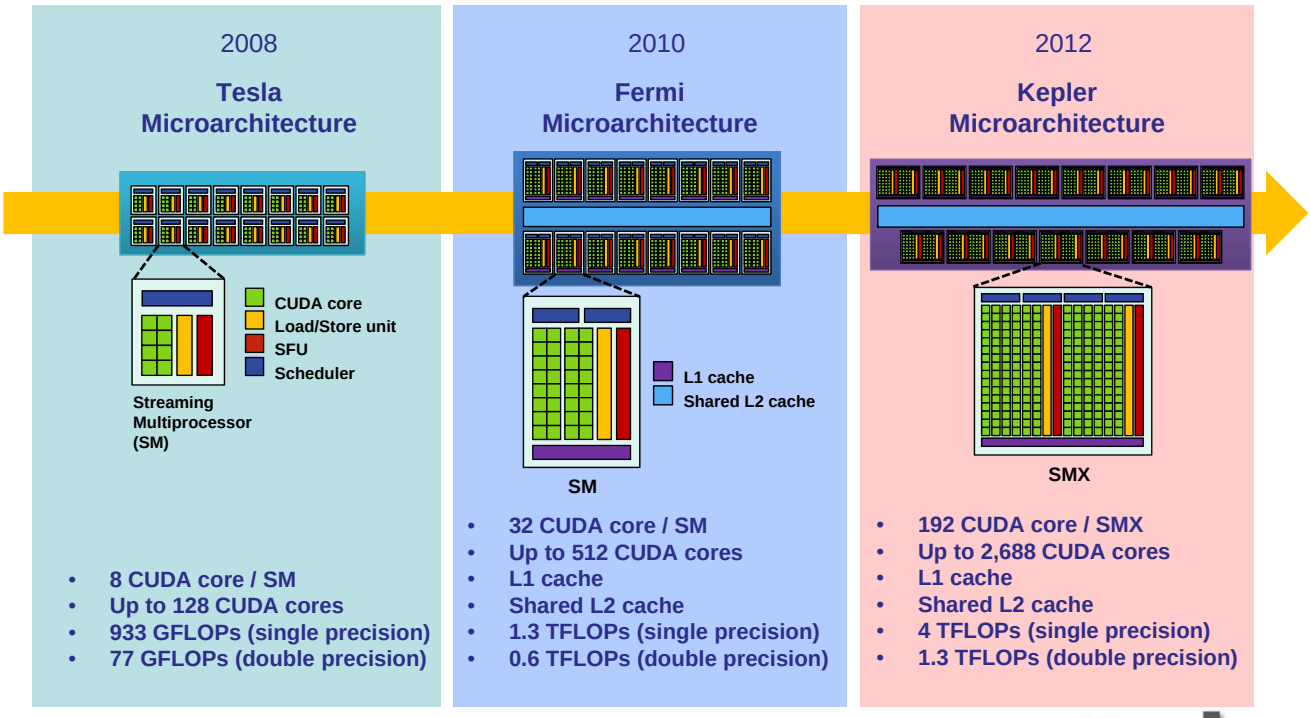
GPU – *GPGPU now!*

- **General-Purpose Computation on GPU**
  - Unified shader architecture의 도입, programmability의 증대
  - 연산 성능의 비약적인 발전
  - 3D graphics 만이 아닌, 일반 범용 연산 분야에서의 사용 시작

The utilization of a graphics processing unit (GPU), which typically handles computation only for computer graphics, to perform **computation in applications traditionally handled by the central processing unit** (CPU).

- from Wiki

NVIDIA GPUs

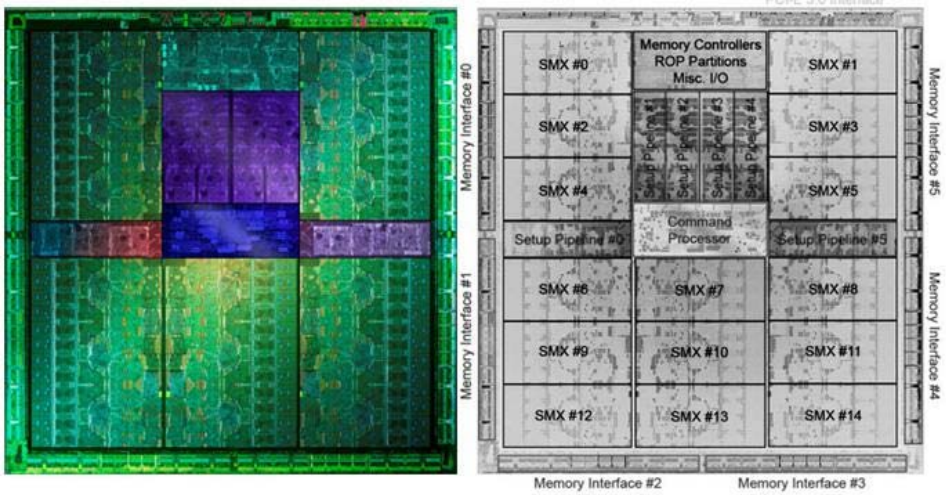


Fermi Architectures

esCaL
Embedded Systems and Computer Architecture Lab.

Kepler GPU Architecture (GK110)

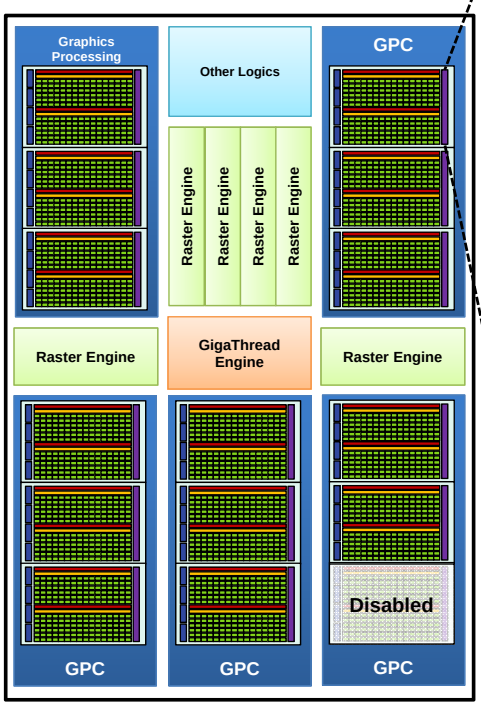
- **28nm, 561mm², 7,080 million Transistors**
- **Core 837MHz**
- **GDDR5, 6GHz, 384-bit, 288.4GB/s bandwidth**
- **TDP 250 watts**



esCaL
Embedded Systems and Computer Architecture Lab.

SMX: 192 CUDA Cores

14 SMX * 192 CUDA cores = 2,688 Execution units



esCaL
Embedded Systems and Computer Architecture Lab.

GPU Architecture 특징

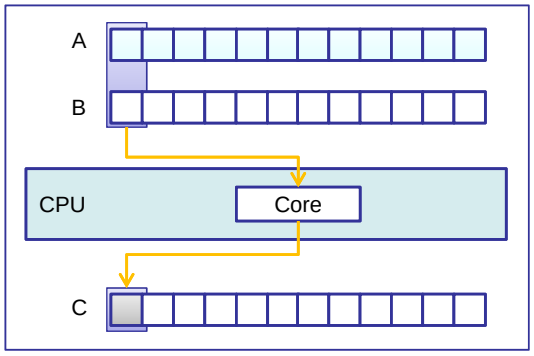
- **3D graphics에서의 픽셀(pixel)당 연산을 고속처리하기 위한 다수의 ALU**
  - 일반적인 CPU의 수십~수백 배의 processor core 내장
- **단순화된 control logic**
  - 분기 예측, 비순차적 실행 지원을 위해 control logic이 매우 비대한 CPU와 달리 매우 단순
- **작은 cache memory, 고용량 register file**
  - 최소한의 cache만을 갖는 대신, 모든 thread에 대하여 register를 독립적으로 할당, context switching latency가 거의 없음
- **고속/광대역폭의 DRAM**
  - 소용량의 cache memory에 의한 성능 저하 극복
    (NVIDIA Fermi GPU: 192.4GB/s vs. Intel Core i7 CPU : 25.6GB/s)

esCaL
Embedded Systems and Computer Architecture Lab.

CPU vs. GPU: Vector Addition

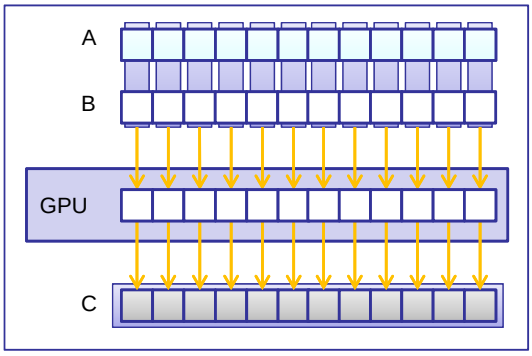
CPU

```
for(int i=0; i<N ; i++){
    c[i] = a[i] + b[i];
}
```



GPU

```
c[threadIdx.x] = a[threadIdx.x] + b[threadIdx.x];
```

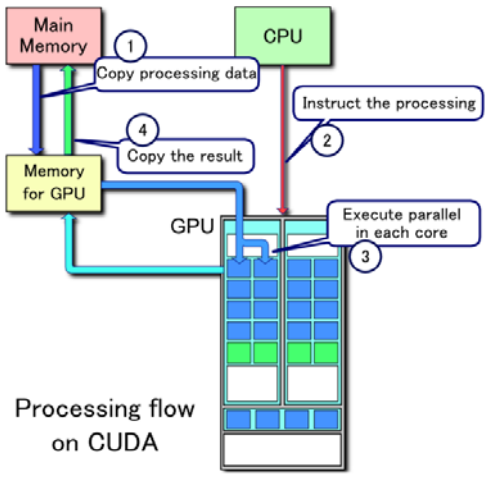


다수의 processor core를 사용, 대량의 데이터에 대한 연산을 병렬처리

esCaL
Embedded Systems and Computer Architecture Lab.

GPGPU Programming Model: CUDA

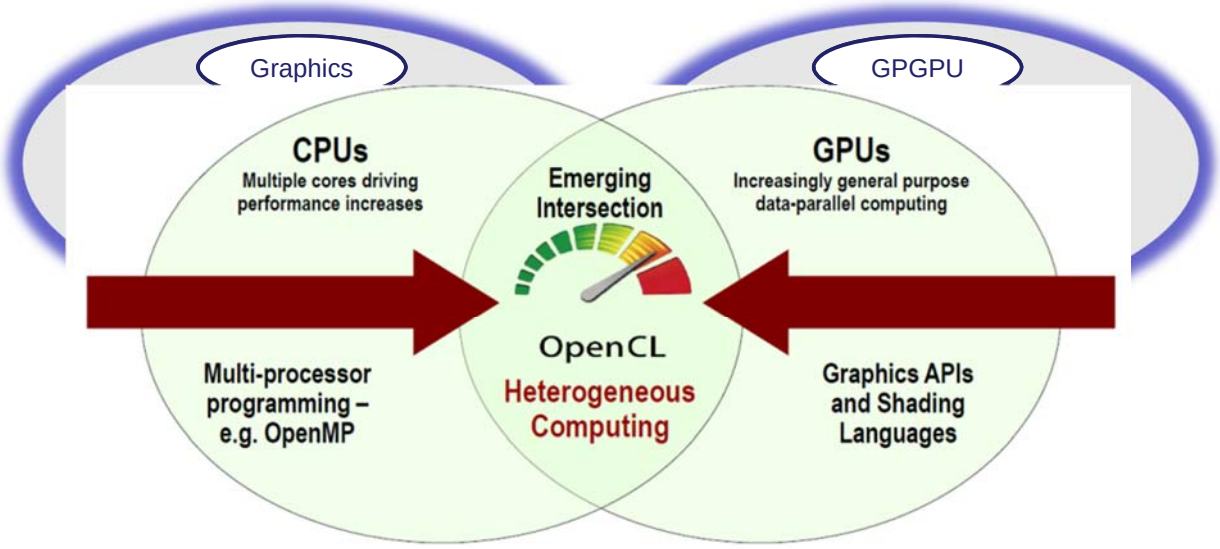
- **NVIDIA GPU의 general purpose processing을 위한 GPU programming model (GPU를 CPU의 연산 보조 프로세서로 활용)**
- **CPU에서 응용 프로그램을 실행, 프로그래머가 지정한 데이터를 GPU로 전송하여 처리, 결과를 CPU에서 활용하는 구조**
- **Many core 프로세서인 GPGPU의 구조적 특징을 활용하기 위해 매우 많은 수의 thread를 사용**
- **각각의 GPU thread들이 SIMT (Single-Instruction, Multiple-Thread) 구조로 동작, 각기 할당 받은 작업을 수행**



Processing flow on CUDA

esCaL
Embedded Systems and Computer Architecture Lab.

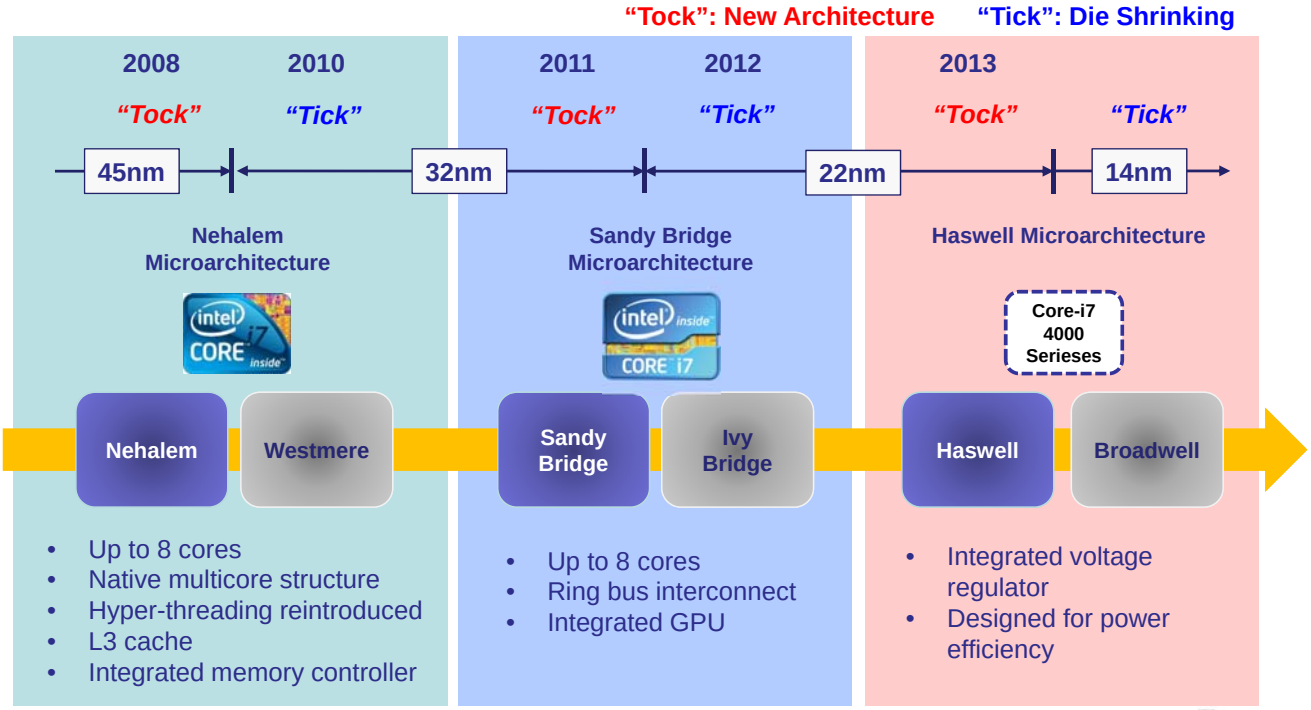
GPU API



INTEL

esCaL
Embedded Systems and Computer Architecture Lab.

Intel Processor: for Desktop/Server



esCaL
Embedded Systems and Computer Architecture Lab.

Intel MIC

Larrabee Microarchitecture
(canceled)

2006~2010

Larrabee

- Designed for 3D graphics / GPGPU
- P54C Pentium based x86 cores
- 512-bit vector processing unit
- Up to 1TFLOPS (single precision FP)

Intel Xeon Phi Family

Intel Mani Integrated Core Architecture
(Intel MIC)

2010	2012	(Planed)
Knight Ferry	**Knight Corner**	**Knight Landing**
• MIC prototype • 45nm • 32 core • 4 thread/core • 2GB RAM • Up to 750GFLOPS (single precision)	• Commercial product • 22nm • 60 core • 4 thread/core • 8GB RAM • Up to 1TFLOPS (double precision FP)	• 2nd generation MIC

Designed for High Performance Computing

esCaL
Embedded Systems and Computer Architecture Lab.

HETEROGENEOUS SYSTEM ARCHITECTURE

esCaL
Embedded Systems and Computer Architecture Lab.

Limitations of Current CPU+GPU Architecture

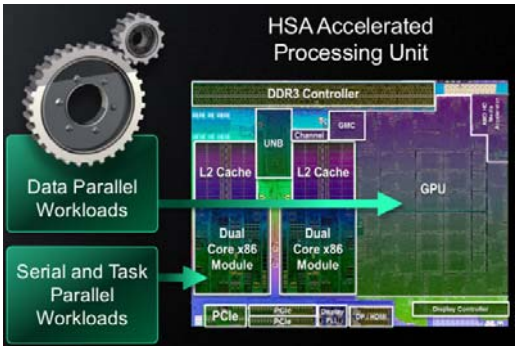
- **Separated Memory Address Space**
  - CPU 및 GPU가 사용하는 memory address space가 서로 분리되어 있음
  - CPU-GPU 간 데이터 전송, memory 할당 등을 모두 programmer가 처리하여야 함
  - GPU의 memory address space가 CPU 대비 작음
  - 대규모의 memory를 사용하는 application의 경우 GPU로 port하기가 어려운 경우가 많음
- **Programming Model**
  - OpenCL 혹은 DirectCompute 등의 general-purpose programming API는 CPU와 GPU를 모두 사용 가능
  - 그러나 각 프로세서 개체 간 데이터 전송 및 제어 과정을 전적으로 programmer에게 의존
  - CPU/GPU를 동시에 효율적으로 제어하기 어렵고 과정이 복잡함
  - GPU측에서 task/thread 생성 및 제어가 매우 제한적

esCaL
Embedded Systems and Computer Architecture Lab.

Heterogeneous System Architecture (HSA)*

AMD가 제안하는 차세대 Computer Architecture

HSA APU = LCU + TCU
Latency Compute Unit (기존 CPU 개념) - Serial/Task parallel workload 처리
Throughput Compute Unit (기존 GPU 개념) - Data parallel workload 처리



Unified Memory Address

HMMU (HSA-specific Memory Management Unit) 적용: LCU와 TCU 간 memory address space를 통합

기존 CPU-GPU와 같은 memory copy 등이 필요 없이, 기존 programming 방법과 같이 pointer 전달로 memory 접근 가능

* Lisa T. Su, “Architecting the Future through Heterogeneous Computing”, ISSCC 2013

esCaL
Embedded Systems and Computer Architecture Lab.

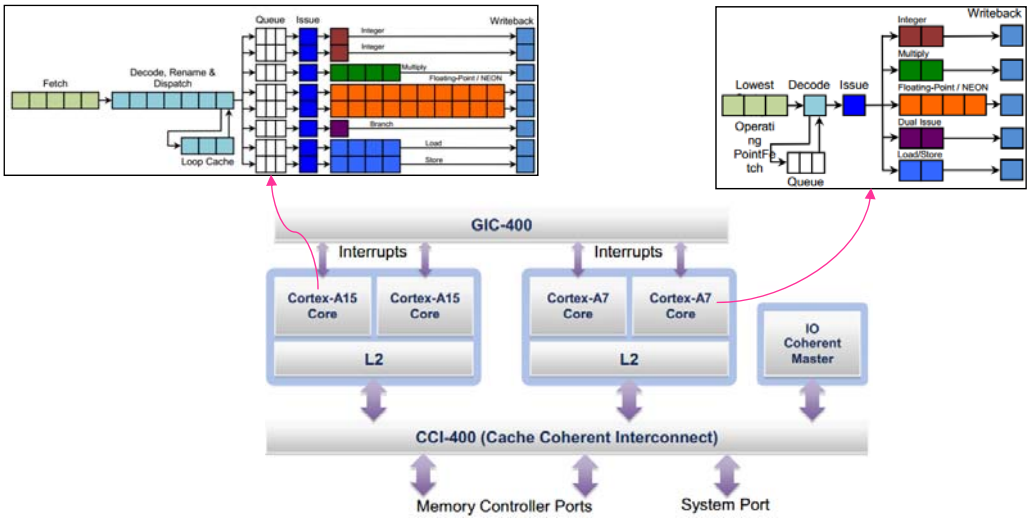
ARM’S BIG.LITTLE ARCHITECTURE

esCaL
Embedded Systems and Computer Architecture Lab.

ARM's big.LITTLE Processor*

* ARM, "Big.LLITTLE Processing with ARM Cortex™-A15 & Cortex-A7" White Paper

- 구조
  - 동일한 ISA를 가진 서로 다른 성능의 두 코어가 그룹을 이루는 아키텍처
  - 고성능/고전력 코어 (Big, Out-of-Order) 와 저성능/저전력 (Small, Inorder) 코어로 구성
  - 하나의 Workload는 한 그룹에 할당 → 두 코어 중 하나에서만 실행 가능



Coretex-A15-CCI-Coretex-A7 System

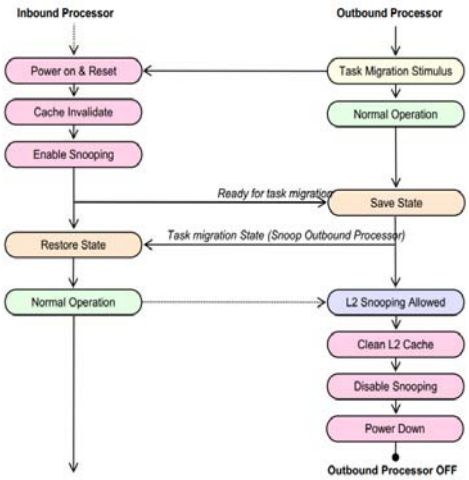
esCaL
Embedded Systems and Computer Architecture Lab.

Task Assignment

- **Performance vs. Power**
  - DVFS를 적용하였을 때의 performance vs. power 그래프
  - 빗금 친 영역 : 동일한 성능을 가질 때의 절약되는 power
- **Task migration**
  - 상호간의 snooping을 허용함으로 인해 overhead 줄임
  - 어느 시점에서 cache snooping 종료 → 전력 절약 모드

Performance vs. Power 그래프

Task migration 과정



esCaL
Embedded Systems and Computer Architecture Lab.

MOBILE

esCaL
Embedded Systems and Computer Architecture Lab.

GPU Architecture Trend

- **Mobile GPU**
  - Desktop PC의 그래픽 경험을 mobile device에서도 동일하게 경험할 수 있도록 함을 목표
  - Mobile 환경에 적합한 저전력을 유지하면서 높은 3D 그래픽 성능 및 media 처리 성능을 얻는 것에 주력
  - 과거 discrete GPU에서 지원하던 graphics API(DirectX, OpenGL 등)에 대한 지원 강화 중
- **Discrete GPU**
  - 현실에 가까운 고수준의 3D graphics 실시간 렌더링을 위한 높은 처리 성능 획득에 주력
  - Graphics 처리 뿐만 아닌 high-performance computing을 위한 구조적 변화
  - 대규모 연산 유닛 및 고속 메모리 사용에 따른 전력/발열 문제로 performance per watt 가 중요 이슈

esCaL
Embedded Systems and Computer Architecture Lab.

Trends: High-Performance Mobile CPU + GPU

Mobile GPUs

PowerVR SGXMP
Multi-core
60M tri/s, 400M pix/s

Mali T450
Quad core
104M tri/s 3.8G pix/s
...

Innovation

Latest Mobile GPUs

Mali T6xx (ARM)
Double-precision floating point
5x performance improvement

PowerVR 6 (Imagination Tech.)
Tile Based Deferred Rendering
Cluster GPU architecture

Adreno GPU (Qualcomm)
128 shader cores
2x performance improvement
…

Influence

Desktop GPUs

GeForce GTX Titan
2688 cores, 4.5T(single)-1.3T(double) FLOPS

Radeon HD 7970
2048 cores, 4.3T(single)-1.1T(double) FLOPS

Future trends for high performance mobile applications:
Parallel computing on heterogeneous MP-SoC with GPGPU

esCaL
Embedded Systems and Computer Architecture Lab.

Commercial Phones

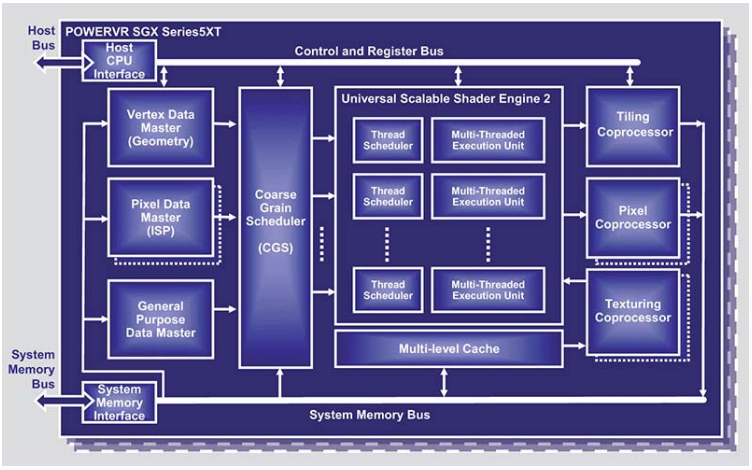
	Samsung Galaxy Note 2	Samsung Galaxy S3	Samsung Galaxy S4	Apple iPhone 5	HTC One X +	LG Optimums G	LG Optimums G pro	Nokia Lumia 920	Motorola Droid Razr Maxx HD	Sony Xperia T
										
Release Date	Aug, 2012	May, 2012	April, 2013	Sep, 2012	Oct, 2012	Oct, 2012	April, 2013	Sep, 2012	Sep, 2012	Sep, 2012
Application Processor	Exynos 4412	Exynos 4412	Exynos 5410	Apple A6	Nvidia Tegra 3 AP37	Qualcomm S4 APQ8064	Qualcomm APQ 8964T	Qualcomm MSM8960	Qualcomm MSM8960	Qualcomm MSM8260A
Technology	32nm	32nm	28nm	32nm	40nm	28nm	28nm	28nm	28nm	28nm
CPU	Quad core Cortex-A9 @1.6GHz	Quad core Cortex-A9 @1.6GHz	Octa core Cortex-A9&15 @1.2&1.6GHz	Dual core Cortex-A15 @1.2GHz	Quad core Cortex-A9 @1.7GHz	Quad core Cortex-A9 @1.5GHz	Quad core Cortex-A9 @1.7GHz	Dual core Cortex-A9 @1.5GHz	Dual core Cortex-A9 @1.5GHz	Dual core Cortex-A9 @1.5GHz
GPU	**QUAD core Mali-400MP @533MHz**	**QUAD core Mali-400MP**	**Triple core PowerVR SGX844MP3**	**Triple core PowerVR SGX543MP3**	**ULP GeForce @520MHz**	**Adreno 320**	**Adreno 320**	**Adreno 225**	**Adreno 225**	**Adreno 225**

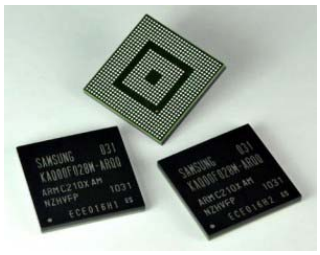
esCaL
Embedded Systems and Computer Architecture Lab.

Imagination

임베디드/모바일 GPU코어 세계시장 1위 업체

PowerVR: Imagination의 mobile GPU 브랜드로서 mobile AP 시장에서 높은 점유율을 가짐
(Apple A4, A5, A6 및 Samsung Hummingbird/Exynos5-Octa 등 다양한 제품에 적용)





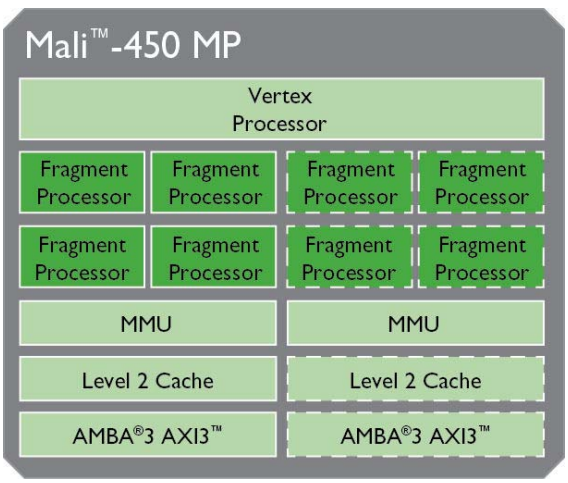
PowerVR SGX GPGPU 코어 아키텍쳐

esCaL
Embedded Systems and Computer Architecture Lab.

ARM

Mali: 노르웨이의 Falanx를 ARM이 인수하여 확보한 GPU 브랜드명

Graphics 전용 Mali400 series 및 high-end 스마트 기기 및 태블릿 시장을 위한 Mali-600 series 출시, Samsung Exynos 4/5에 적용



Mali-450MP 코어 아키텍쳐

esCaL
Embedded Systems and Computer Architecture Lab.

Qualcomm

Adreno: Qualcomm의 mobile GPU 브랜드

Adreno320은 가장 최근에 출시된 unified shader구조의 GPU코어로 dual 메모리 채널 및 OpenCL 1.1을 지원 (Snapdragon600, Snapdragon800에 적용)

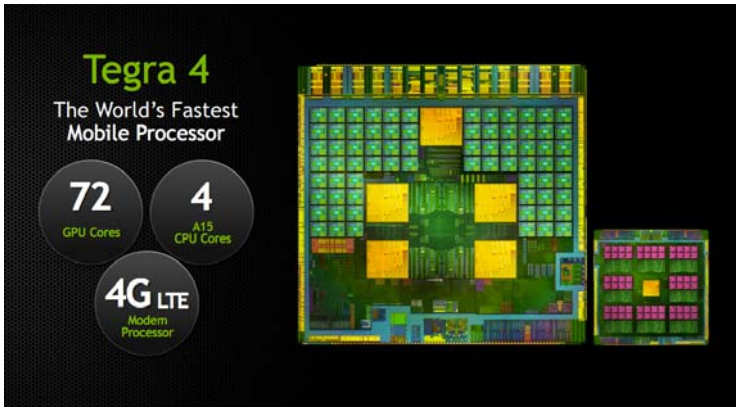
Qualcomm Adreno GPU Roadmap

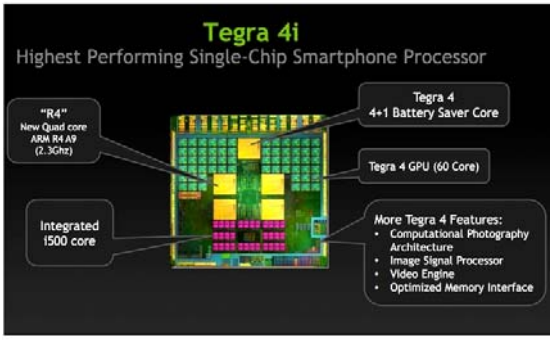
esCaL
Embedded Systems and Computer Architecture Lab.

NVIDIA

Mobile GPU

자사의 Tegra series로 mobile AP 시장 공략 중, Tegra 및 이에 포함되는 GPU인 ULP



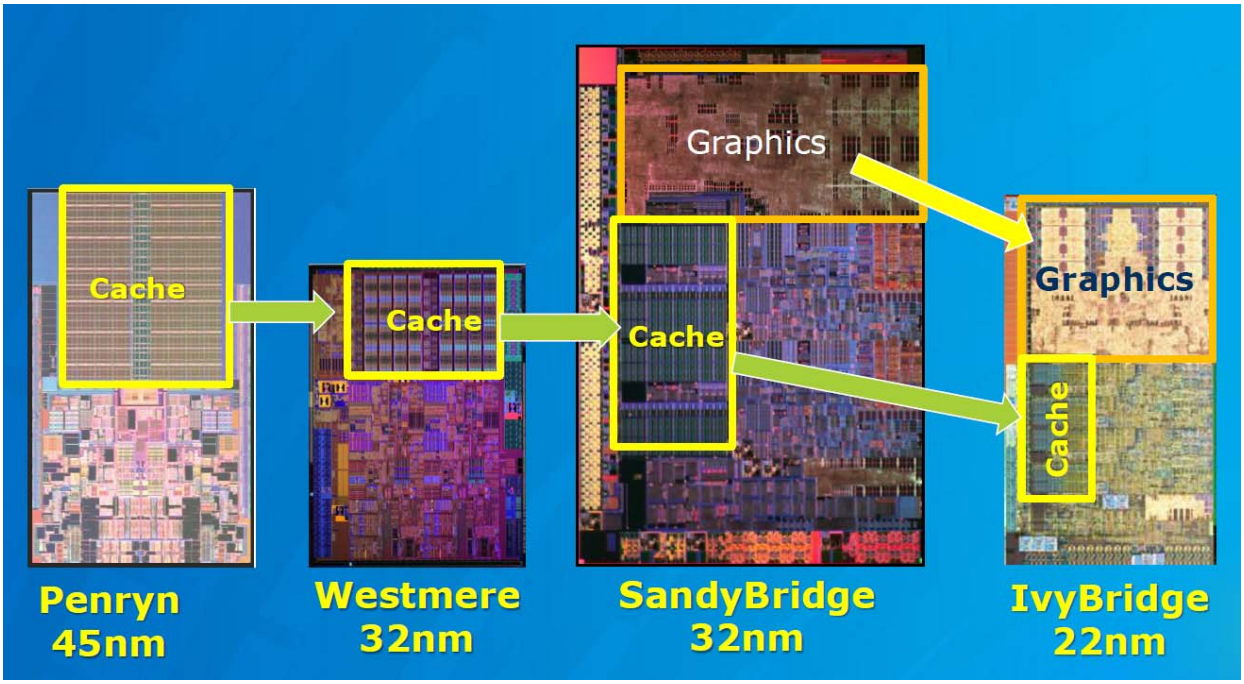


esCaL
Embedded Systems and Computer Architecture Lab.

CONCLUSION

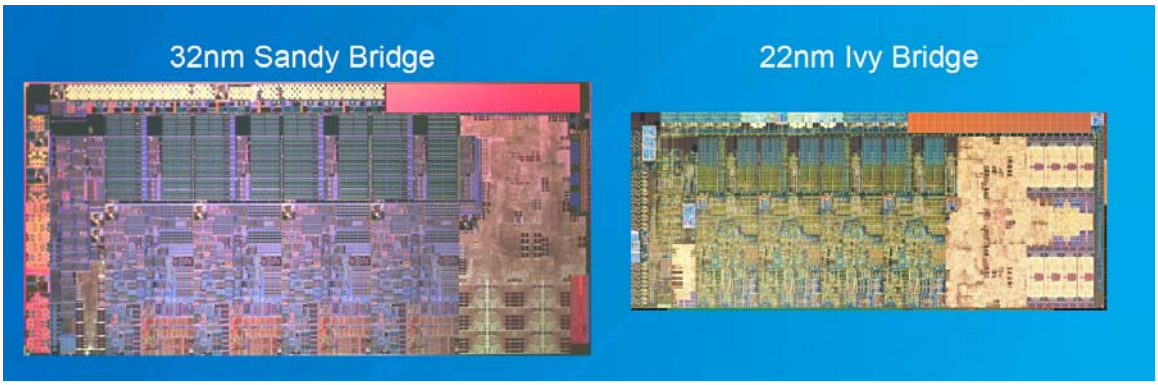
esCaL
Embedded Systems and Computer Architecture Lab.

Intel Processors



esCaL
Embedded Systems and Computer Architecture Lab.

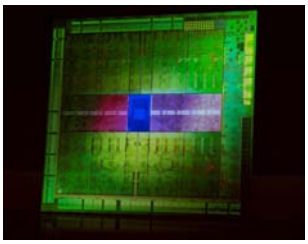
Heterogeneous…



	Sandy Bridge	Ivy Bridge
Die Size	212 mm^2	160 mm^2
Total Transistors	1.16 B	1.40 B
Core Transistors	79.4 M	80.4 M

CaL
nd Computer Architecture Lab.

Commercial Processors

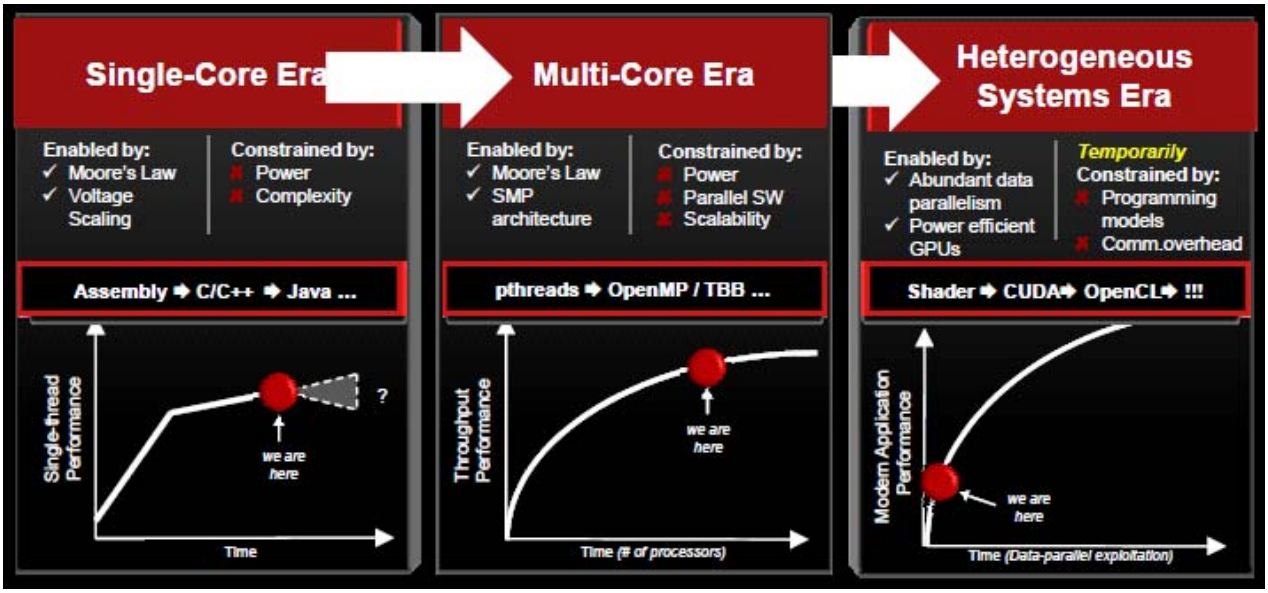


intel inside
Xeon Phi

	Sandy Bridge (Core i7 3970x)	Kepler (GK110)	Knights Corner (Xeon Phi SE10x)
Die Size	435 mm^2	551 mm^2	350 mm^2
Total Transistors	2.3 B	7.1 B	5 B
No of Cores	6 CPU Cores	2,688 CUDA Cores	61 Cores
TDP	150 W	250 W	300 W

b.

So…



- **Cache and memory**

esCaL
Embedded Systems and Computer Architecture Lab.

Q & A

Thank you!

esCaL
Embedded Systems and Computer Architecture Lab.