# TALLER DE PROGRAMACIÓN SOBRE GPUS

**Facultad de Informática – Universidad Nacional de La Plata**

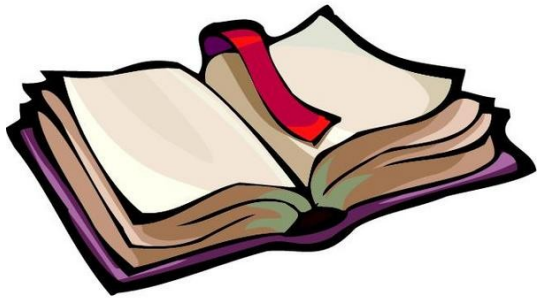

Dr. Adrián Pousa

## Arquitecturas GPU

# Agenda



I. *Arquitectura GPU: Introducción*
II. *GPUs de arquitectura fija y evolución hacia arquitecturas unificadas*
III. *Arquitecturas Nvidia y su evolución*
IV. *Arquitecturas ATI-AMD*
V. *Otras arquitecturas Manycore*
    I. *Intel Xeon PHI*
    II. *Pezy-SC*
    III. *Systema Sunway*
VI. *Rankings*
VII. *GPUs provistas por la cátedra*



# Agenda



I. **Arquitectura GPU: Introducción**

II. GPUs de arquitectura fija y evolución hacia arquitecturas unificadas

III. Arquitecturas Nvidia y su evolución

IV. Arquitecturas ATI-AMD

V. Otras arquitecturas Manycore

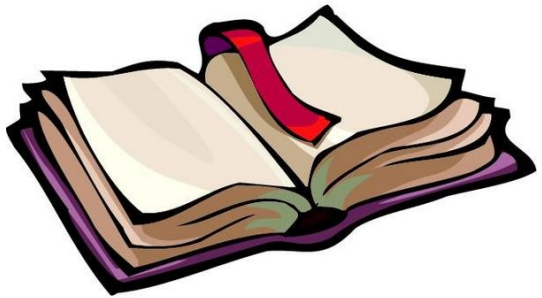
   I. Intel Xeon PHI

   II. Pezy-SC

   III. Systema Sunway

VI. Rankings

VII. GPUs provistas por la cátedra



# GPUs



- Surgen con el fin de liberar a la CPU del proceso de renderizado (a partir de un modelo con determinadas características generar una imagen).

- Actualmente agregan mayor funcionalidad:
  - Mip Mapping
  - Z-Buffering
  - Texturados
  - Antialiasing
  - Etc…

- Evolucionan a partir de la industria de videojuegos.

# GPUs



- La idea es optimizar el **throughput** de muchos hilos (threads) ejecutando en paralelo - ocuparlos la mayor parte del tiempo.

- La mayor parte de la arquitectura esta orientada a cómputo.

- Mayor ancho de banda de memoria que CPU (la CPU debe optimizar el ancho de banda de memoria para SO, aplicaciones, E/S):
  - Nvidia 590: 327 GB/s
  - Nvidia Titan V (Volta) 653 GB/s
  - Xeon E7: 102 GB/s

# GPU: Arquitecturas



- Dos empresas lideran el mercado de GPUs:
  - Nvidia:
    - GPU serie Geforce para gráficos de escritorio
    - GPU serie Tesla y Quadro para HPC
  - AMD (ex ATI Tecnologies):
    - GPU Radeon, Evergreen, Northern Island, Southern Island para gráficos de escritorio
    - GPU Flipper, Xenos, Hollywood para consolas
    - GPU serie Fire para HPC

# Agenda



I. *Arquitectura GPU: Introducción*

II. ***GPUs de arquitectura fija y evolución hacia arquitecturas unificadas***

III. *Arquitecturas Nvidia y su evolución*

IV. *Arquitecturas ATI-AMD*

V. *Otras arquitecturas Manycore*

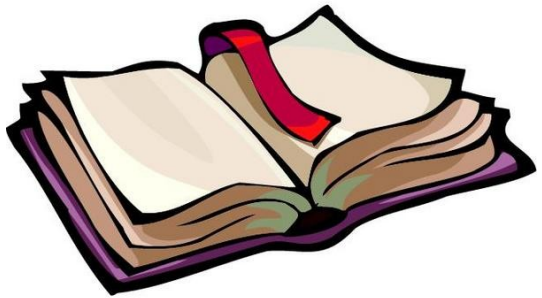
   I. *Intel Xeon PHI*

   II. *Pezy-SC*

   III. *Systema Sunway*

VI. *Rankings*

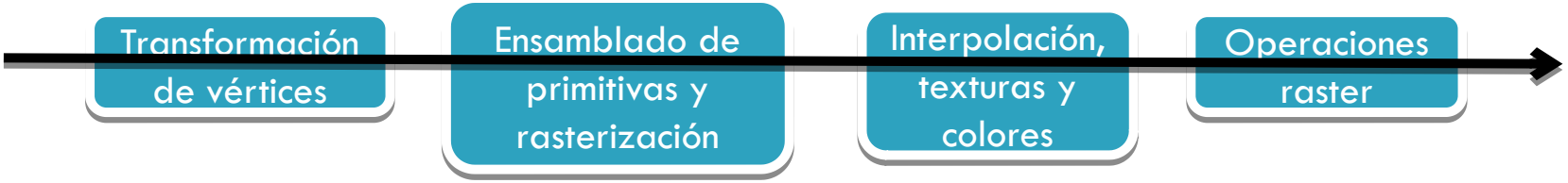
VII. *GPUs provistas por la cátedra*



# GPUs: Evolución



- Primeras GPU - Pipeline Gráfico:
  - Recibe una imagen 3D y da como resultado una imagen 2D
  - 4 etapas ejecutadas en paralelo

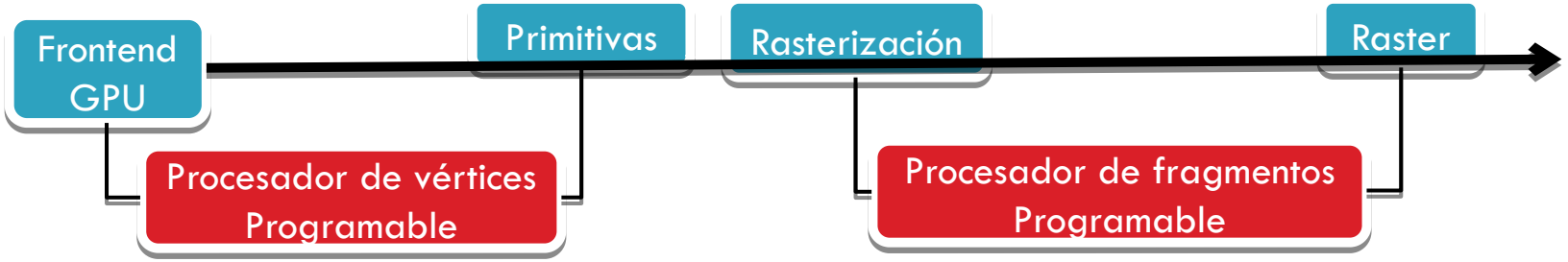



  - Realizado por hardware y con recursos dedicados
    - En CPU debe hacerse por software y todos los recursos para un estado.
  - Limitados en el tamaño del problema y en operaciones complejas (sombras e iluminación)

# GPUs: Evolución



- Primeras GPU – Modelo shader (Geforce 6 y 7):
  - Soluciona las limitaciones con procesadores dedicados y programables:
    - Shader de Vértices (geometría)
    - Shader de Fragmentos (pixeles)




  - Problema de desbalance de carga entre procesadores:
    - Aplicaciones con mucha carga geométrica ocupa procesador de vértices desaprovechando del procesador de fragmentos (ociosos)

# GPUs: Evolución



- Modelos anteriores de arquitectura fija:
  - Hardware dedicado a cada función
  - Subutilización de procesadores

- Arquitecturas unificadas (a partir de Nvidia G80 – ATI R500):
  - Elimina la división de shader de vértices y fragmentos
  - Cada unidad de procesamiento puede realizar las dos funciones
  - Las unidades de procesamiento se llaman Stream Processors (**SM**)
  - Se eliminan las partes específicas del pipeline
  - Una unidad de procesamiento puede realizar todas las operaciones según la exigencia
  - Soluciona el problema de desequilibrio de carga (Autoequilibrio de carga)

# Agenda



I. *Arquitectura GPU: Introducción*

II. *GPUs de arquitectura fija y evolución hacia arquitecturas unificadas*

III. ***Arquitecturas Nvidia y su evolución***

IV. *Arquitecturas ATI-AMD*

V. *Otras arquitecturas Manycore*

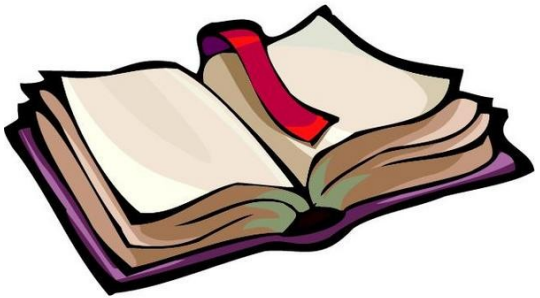
   I. *Intel Xeon PHI*

   II. *Pezy-SC*

   III. *Systema Sunway*

VI. *Rankings*

VII. *GPUs provistas por la cátedra*



# GPUs: Evolución NVidia



| G80 | GT200 | Fermi | Kepler | Maxwell | Pascal | Volta | Turing |
|---|---|---|---|---|---|---|---|
| 2006 | 2008 | 2010 | 2012 | 2014 | 2016 | 2017 | 2018 |

# GPUs: Evolución NVidia



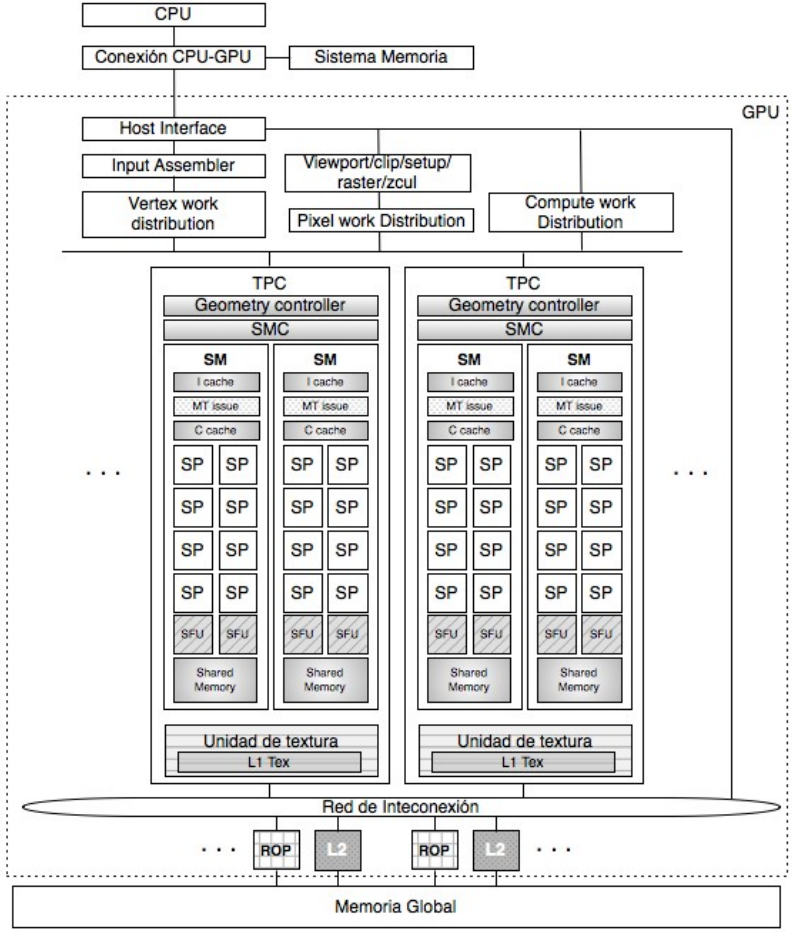
- Arquitectura G80:
  - Reemplaza los pipelines por un procesador unificado
  - Introduce un procesador escalar (**SP**) de threads
  - Presenta el modelo STMD
  - Introduce la memoria compartida y la sincronización por barreras entre threads
  - Soporte en lenguaje C

| G80 | GT200 | Fermi | Kepler | Maxwell | Pascal | Volta | Turing |
|---|---|---|---|---|---|---|---|
| 2006 | 2008 | 2010 | 2012 | 2014 | 2016 | 2017 | 2018 |

# GPUs: Evolución NVic



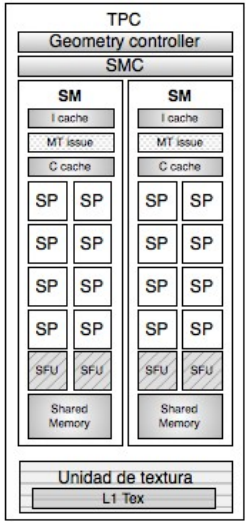
- Arquitectura G80:
  - Arreglo de procesadores de *streaming (SPA)*
  - 8 Unidades de procesamiento independientes
    - Texture/Processor Cluster (**TPC**)
  - Compute work distribution:
    - Distribuye threads a TPC
  - Sistema de memoria:
    - Memoria Global (DRAM)
    - Procesadores Raster de función fija (ROP)
    - Acceso mediante una red de interconexión
    - Acceden todos los TPC



# GPUs: Evolución NVidia



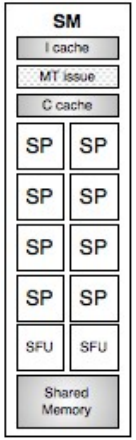
- Arquitectura G80:
  - Cada TPC:
    - Cache L1
    - Unidades de acceso a texturas
    - Array de Stream Processors (**SM**):
      - 2 por TPC
    - Planificador de threads



# GPUs: Evolución NVidia



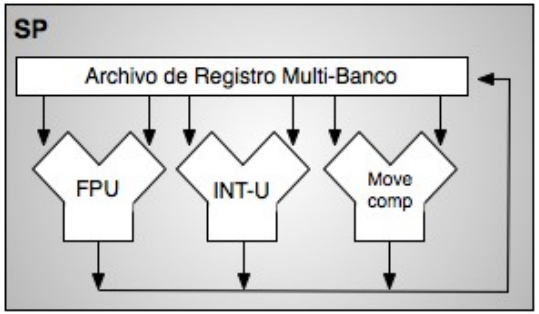
- Arquitectura G80:
  - Cada SM:
    - 8 Scalar Processors (**SP**)
    - Cache de instrucciones (**I cache**) compartida por los SP
    - Cache de solo lectura (**C cache**) compartida por los SP
    - 2 Unidades de funciones especiales (SFU)
      - Sqrt, Sin, Cos, Log etc
      - Si las SFU están ocupadas el planificador ejecuta otras sentencias
    - Unidad MT que envía instrucciones a los SP y SFU
    - Memoria compartida (**shared memory**) por los SP



# GPUs: Evolución NVidia



- Arquitectura G80:
  - Cada SP:
    - Operaciones matemáticas
    - Unidad de punto flotante (FPU)
    - Unidad de punto fijo (INT-U)
    - Movimientos de datos hacia y desde la memoria

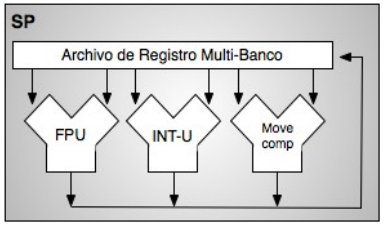


# GPUs: Evolución NVidia

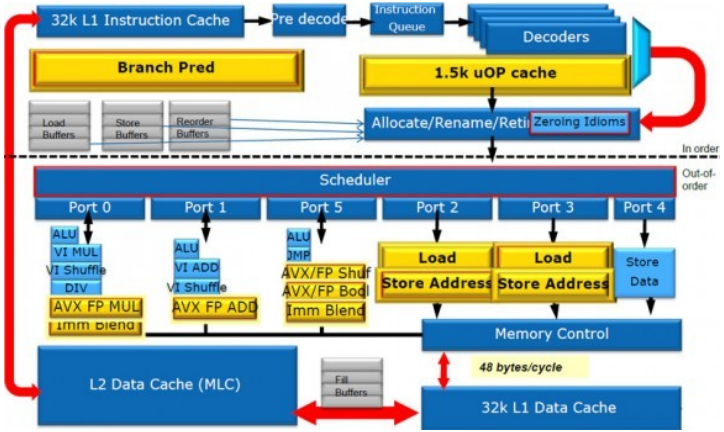


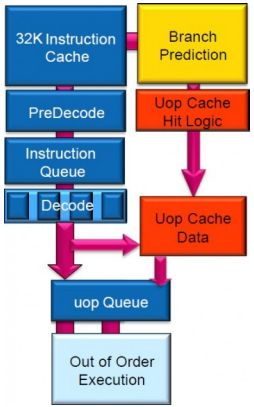
- Si comparamos un SP con una arquitectura CPU Intel Sandy Bridge…

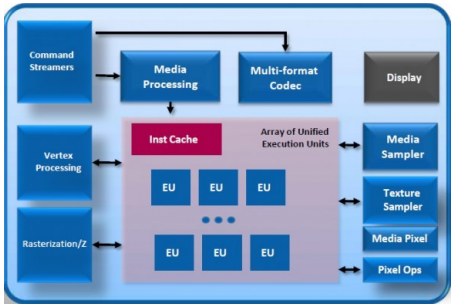
**GPU Core**




**CPU Core**



# GPUs: Evolución NVidia



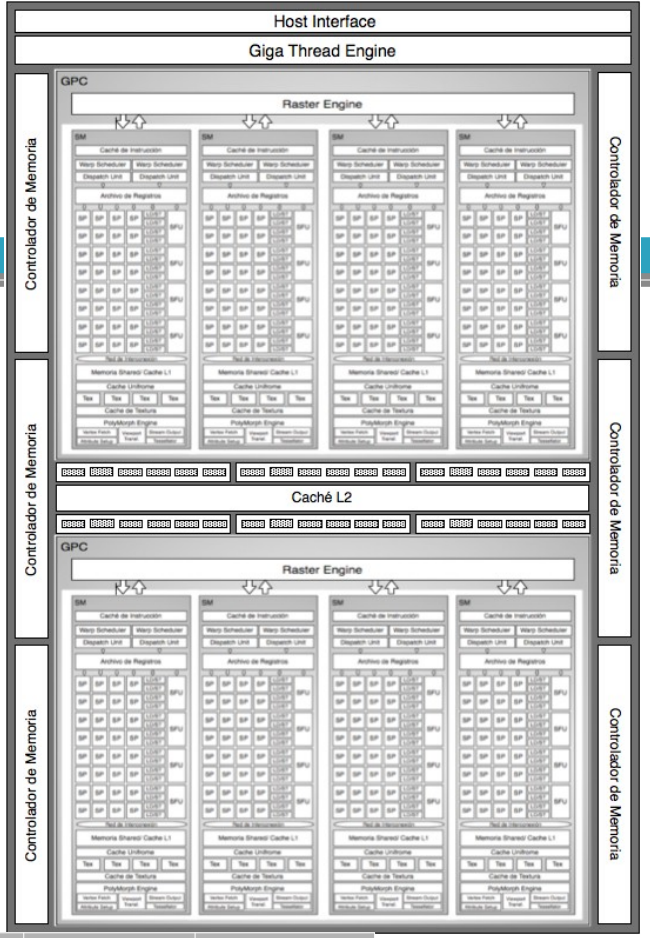
- Arquitectura GT200 (Geforce, Quadro y Tesla):
  - Incremento en el número de SPs
  - Mayor capacidad en los registros
  - Hardware adicional para acceso eficiente a memoria (**Coalescencia**)
  - Soporte para punto flotante en doble precisión
  - Soporte para mayor número de thread.
  - Memoria compartida de mayor tamaño (64Kb) y configurable:
    - 16KB de memoria compartida y 48KB de cache L1
    - 48KB de memoria compartida y 16KB de cache L1

| G80 | GT200 | Fermi | Kepler | Maxwell | Pascal | Volta | Turing |
|---|---|---|---|---|---|---|---|
| 2006 | 2008 | 2010 | 2012 | 2014 | 2016 | 2017 | 2018 |

# GPUs: Evolución NVidia



- Arquitectura GF100 (Fermi):
  - Reemplaza TPC por *Graphics Processing Clusters*(**GPC**)
  - Hasta 4 GPC (dependiente del modelo)
  - Los SMs comparten una cache de nivel 2
  - 6 controladores de memoria

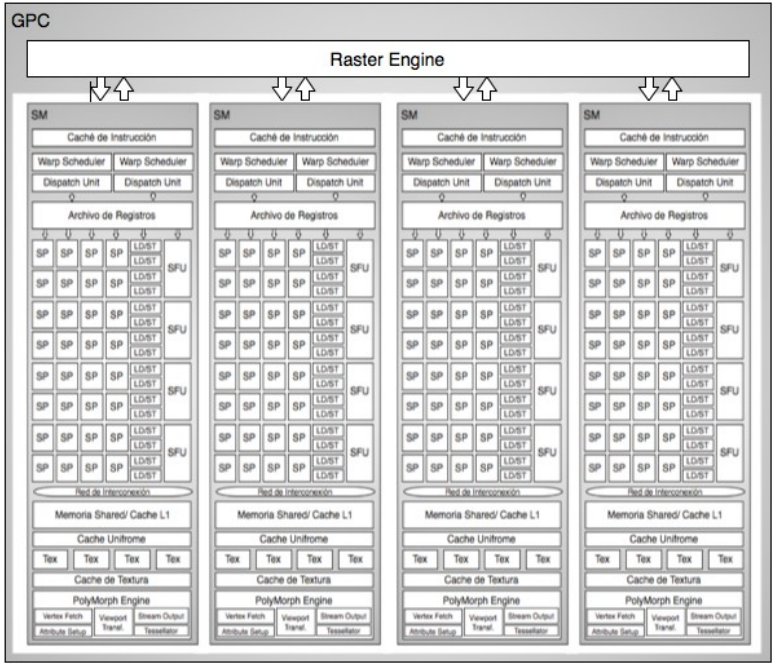



| G80 | GT200 | Fermi | Kepler | Maxwell | Pascal | Volta | Turing |
|---|---|---|---|---|---|---|---|
| 2006 | 2008 | 2010 | 2012 | 2014 | 2016 | 2017 | 2018 |

# GPUs: Evolución NVidia



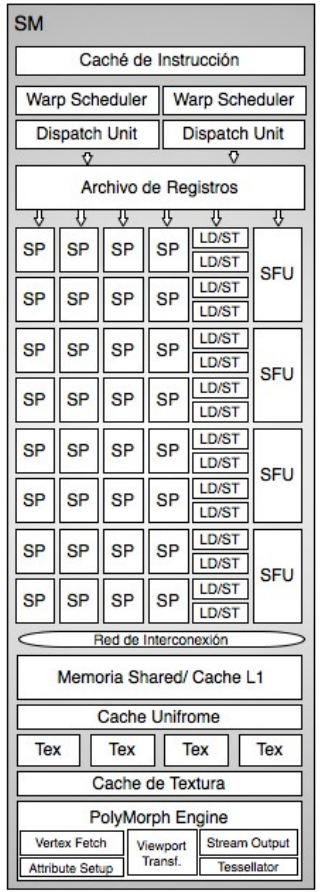
- Arquitectura GF100 (Fermi):
  - GPC:
    - Raster engine
    - Hasta 4 SMs



# GPUs: Evolución NVidia



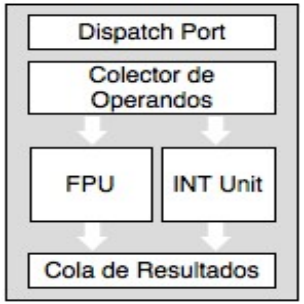
- Arquitectura GF100 (Fermi):
  - Cada SM tiene 32 SPs (2 bloques de 16 SPs)
  - Los SMs comparten una cache de nivel 2 (768KB)
  - Unidades de Load/Store:
    - Escriben o leen direcciones de 16 threads por ciclo en cache o en DRAM
  - Incremento en el número de SFUs
  - Hilos administrados en warps (32 hilos):
    - 2 scheduler de warps (permite ejecutar 2 warp concurrentemente – no en 64 bits)
    - 2 dispatchers de instrucciones



# GPUs: Evolución NVidia



- Arquitectura GF100 (Fermi):
  - Cada SP:
    - Una unidad de punto flotante (FPU), soporta simple y doble precisión
    - Una unidad aritmético lógica (ALU), soporta nuevas operaciones:
      - Shift
      - Bit-inversion
      - Compare
      - Convert
      - Etc…



# GPUs: Evolución NVidia



- Arquitectura GF100 (Fermi):
  - Planificador – Administrador de threads:
    - La CPU envía comandos a la GPU que recibe la unidad GigaThread
    - La unidad GigaThread envía bloques de threads a varios SMs
    - Cada SM administra los bloques en warps (32 threads)
    - El planificador dual de warps selecciona 2 warps y da una instrucción de cada warp a 16SPs, 16 unidades Load/Store o 4 SFU
    - La mayoría de las instrucciones pueden ser lanzadas dual:
      - Se pueden lanzar 2 operaciones enteras, Load/Store y SFU en concurrente
      - Las instrucciones de doble precisión no pueden ser lanzadas concurrentemente
  - **Multithreading por hardware**: cambios de contextos por hardware nativo.

# GPUs: Evolución NVidia



- Arquitectura GK110 (Kepler):
  - SM renombrado a SMX:
    - 192 CUDA cores (SPs)
    - 32 Special Function Units (SFU)
    - 32 Load/Store units (LD/ST)
    - 4 warp scheduler
  - Anidamiento de kernel sin interacción con la CPU
  - Creación dinámica de threads sin interacción con la CPU
  - Kernels pueden crear otros kernels
  - Hyper-Q: permite conexiones entre cuda streams y procesos MPI

| G80 | GT200 | Fermi | Kepler | Maxwell | Pascal | Volta | Turing |
|---|---|---|---|---|---|---|---|
| 2006 | 2008 | 2010 | 2012 | 2014 | 2016 | 2017 | 2018 |

# GPUs: Evolución NVidia

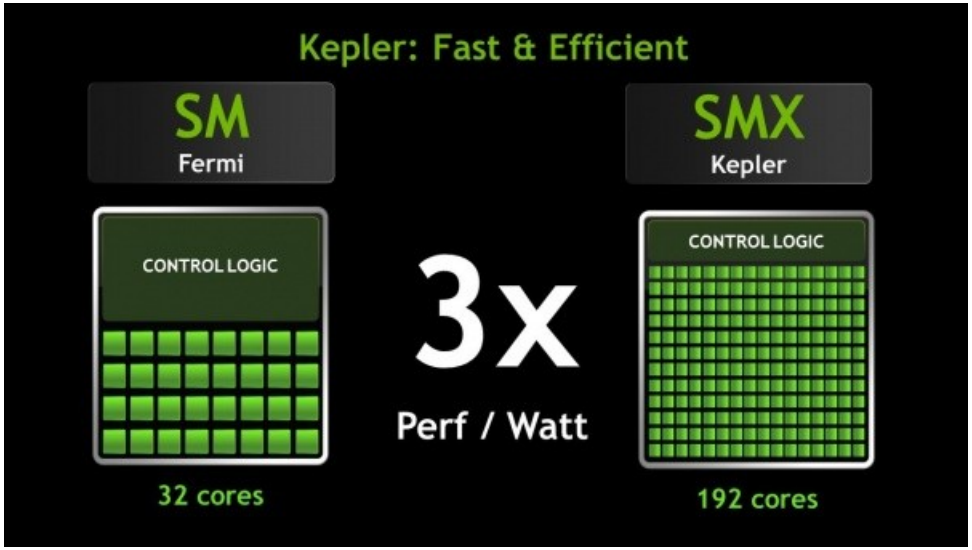


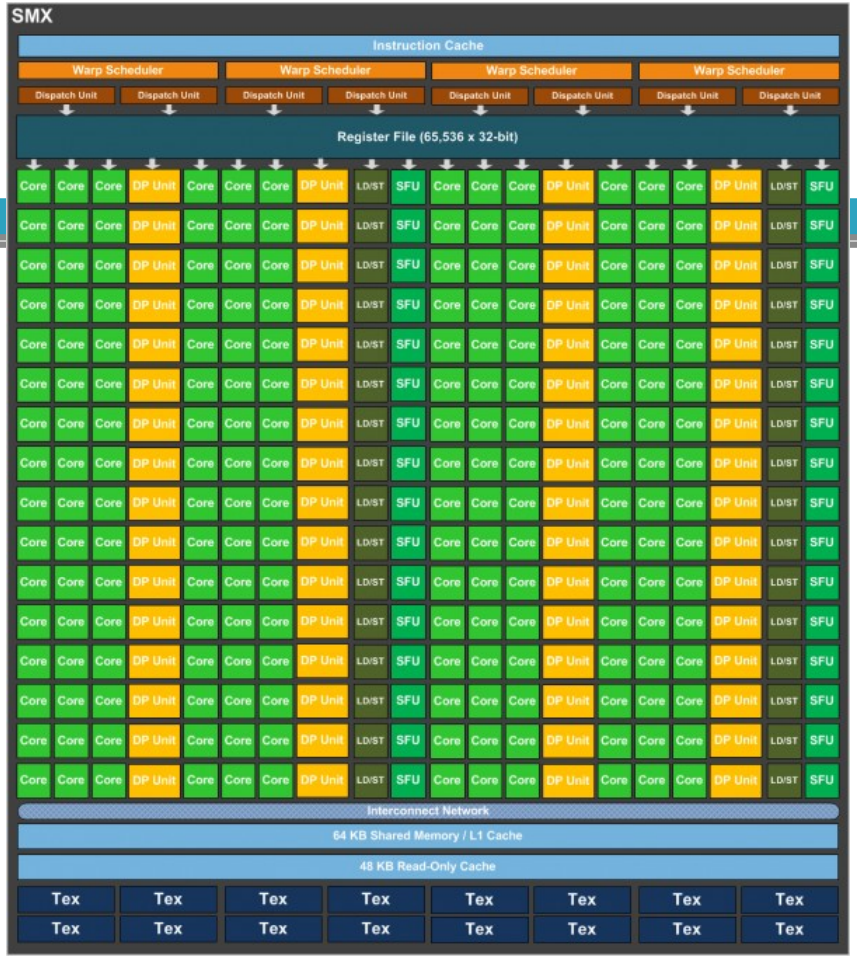
- Arquitectura GK110 (Kepler) :
  - Hasta 8 SMX (1536 SPs)
    - Hasta 2,08 Gflops
  - Modelo 690 2x8 SMX ( 3072 SPs)
    - Hasta 2 x 2,08 Gflops
  - Hasta 6 Gb de Memoria GDDR5
  - Entre 4 y 18 Gflops/Watt
  - GPU Boost (análogo turboboost)

# GPUs: Evolución NVi



- Arquitectura GK110 (Kepler):



# GPUs: Evolución NVidia



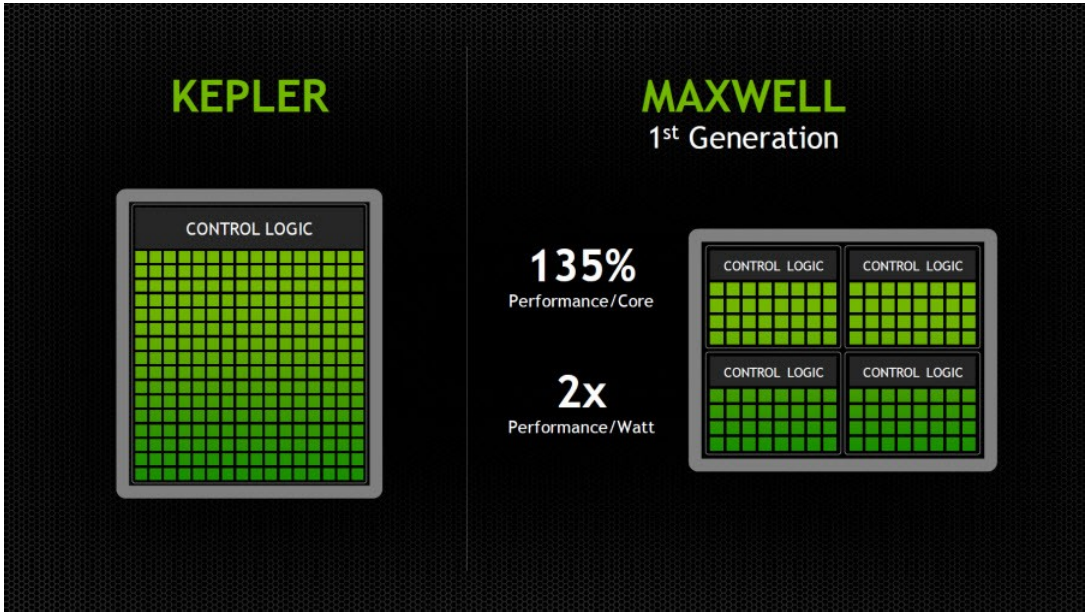
- Arquitectura GM107 (Maxwell):
  - Reordenamiento de los cores para mejoras en cuanto al ahorro de energía:
    - 60Watts contra 250Watts de la GTX 480
    - 30GFlops/Watt contra 15GFlop/Watts de las anteriores
  - SMX renombrado a SMM.
  - Cada SMM dividido en 4 Control Logic, cada Control Logic:
    - 32 CUDA cores (SPs)
    - 8 Special Function Units (SFU)
    - 8 Load/Store units (LD/ST)
    - 4 warp scheduler
  - Se integra con CPUs ARM (Karma)
  - Incremento de caché L2 – Reducción de las necesidades de ancho de banda

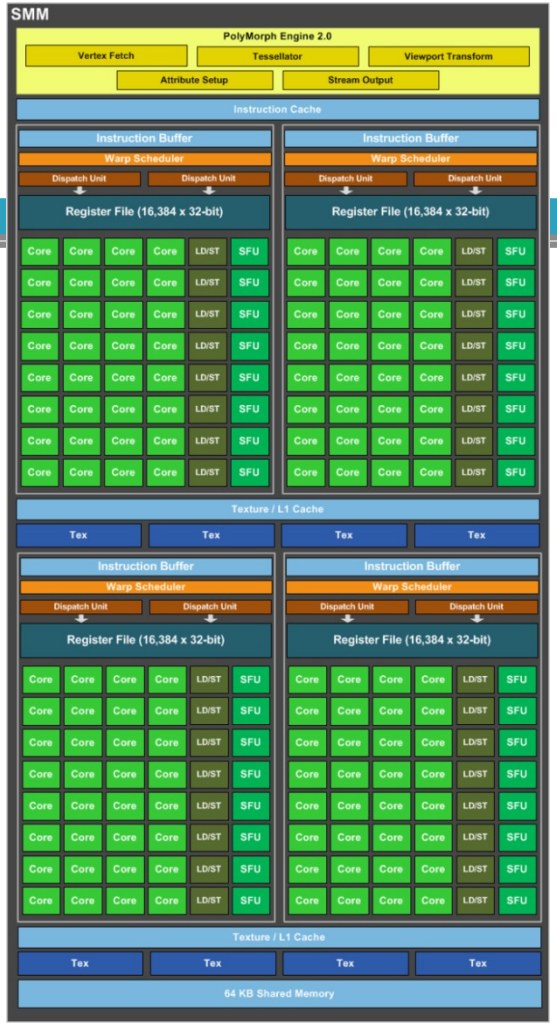
| G80 | GT200 | Fermi | Kepler | Maxwell | Pascal | Volta | Turing |
|---|---|---|---|---|---|---|---|
| 2006 | 2008 | 2010 | 2012 | 2014 | 2016 | 2017 | 2018 |

# GPUs: Evolución NVidia



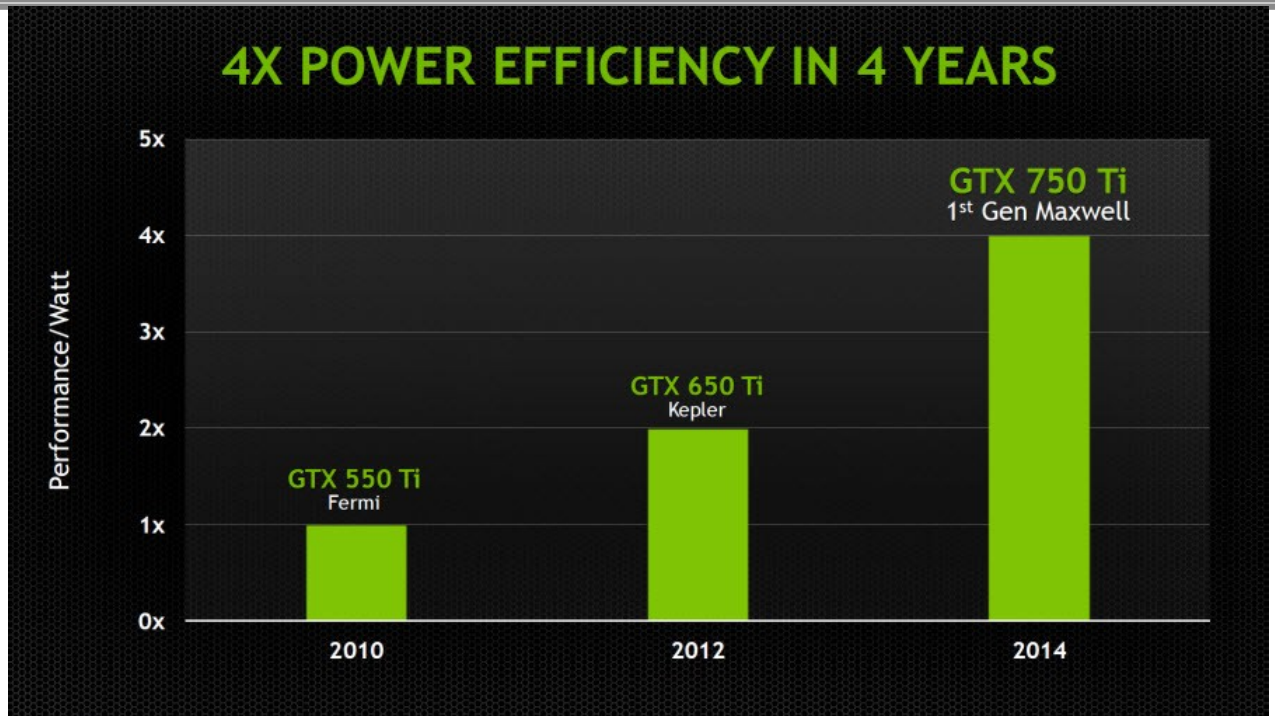
- Arquitectura GM107 (Maxwell):



# GPUs: Evolución NVidia

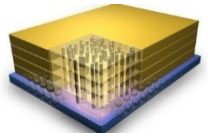




# GPUs: Evolución NVidia



- Arquitectura GP100 (Pascal):
  - Memoria HBM2 (**High Bandwidth Memory**): Sólo algunos modelos
  - 3D Stacked Memory (2,5D): La memoria permite ser "Apilada" alcanzando velocidades de 720GB/s

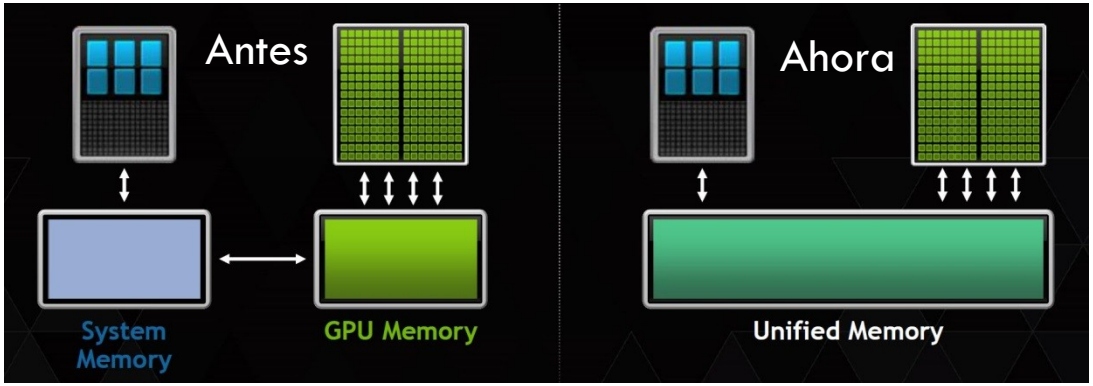


| G80 | GT200 | Fermi | Kepler | Maxwell | Pascal | Volta | Turing |
|---|---|---|---|---|---|---|---|
| 2006 | 2008 | 2010 | 2012 | 2014 | 2016 | 2017 | 2018 |

# GPUs: Evolución NVidia



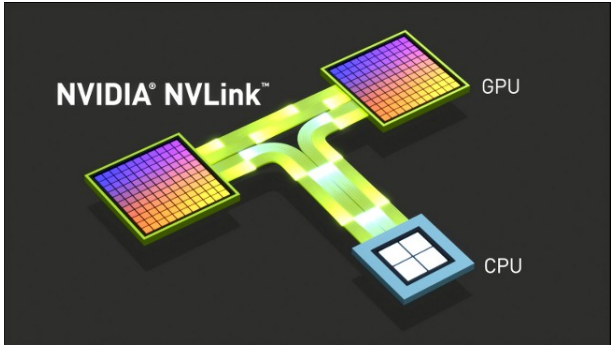
- Arquitectura GP100 (Pascal):
  - Memoria Unificada:



# GPUs: Evolución NVidia



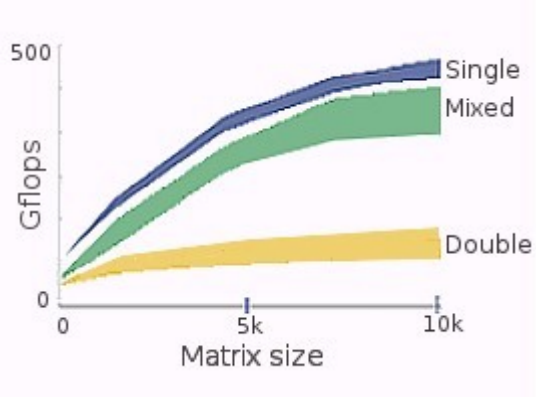
- Arquitectura GP100 (Pascal):
  - Nvlink (NVidia NVLINK high-speed interconect):
    - Interconexión con alto ancho de banda y eficiente energéticamente para comunicación ultra-rápida entre la CPU y la GPU
    - De 5 a 12 veces más rápido que interconexiones PCIe Gen3



# GPUs: Evolución NVidia



- Arquitectura GP100 (Pascal):
  - Mejoras en cálculo de doble precisión
  - **Mixed-Precision:** método que usa diferente precisión en un cálculo
  - Uso típico: mezcla de operaciones de simple y doble precisión
  - Ejemplo: double(a) + double( float(b) + float(c) )
  - Limitaciones anteriores con doble precisión
    - Doble precisión no soportada hasta GT200
    - A partir de Fermi pero con bajo rendimiento



# GPUs: Evolución NVidia



- SM:
  - Mayor número de SMs:
    - 60 (activos 56).
  - Menos SPs (Cuda cores) por SM:
    - 64
  - Por SM, mayor porción de:
    - Memoria compartida
    - Registros

# GPUs: Evolución NVidia



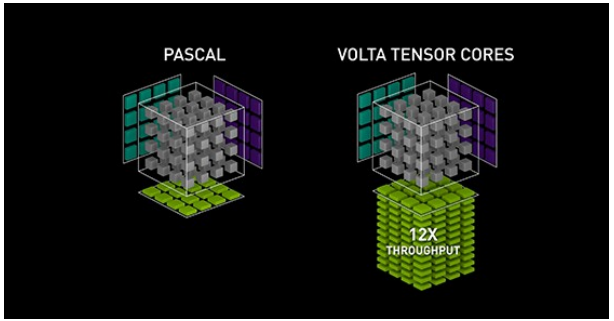
**Total cores**

**56 * 64 = 3584**

# GPUs: Evolución NVidia



- Arquitectura GV100 - 400 (Volta):
  - **NVLink 2.0:** Actualización del bus NVLink. Velocidades mayores al PCI Express (300GB/s).
  - **HBM2:** Mejora en la tecnología de la memoria (anchos de banda hasta 900GB/s)
  - **Tensor cores:** elementos de cómputo diseñados para multiplicar matrices FP16 de 4x4, y permiten también la acumulación de una tercera matrix FP16 o FP32 (Formato en coma flotante de simple precisión). Introducidos para acelerar el entrenamiento de redes neuronales (**Deep learning**).

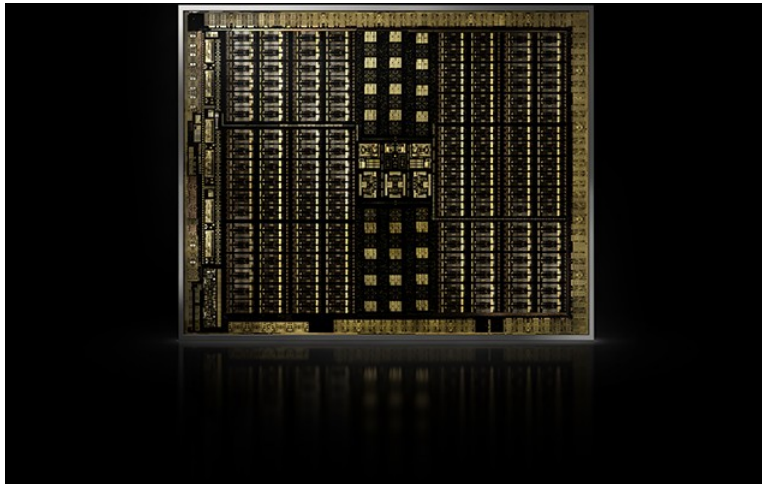



| G80 | GT200 | Fermi | Kepler | Maxwell | Pascal | Volta | Turing |
|---|---|---|---|---|---|---|---|
| 2006 | 2008 | 2010 | 2012 | 2014 | 2016 | 2017 | 2018 |

# GPUs: Evolución NVidia



- Nvidia **RTX** (Turing) (septiembre de 2018)
- RTX 2080 TI
  - **Sps:** 4352
  - **SM:** 32
  - **Clock:** 1.3Ghz boostclock1.5Ghz/1.6Ghz
  - **Memoria GDDR6:** 11GB



| G80 | GT200 | Fermi | Kepler | Maxwell | Pascal | Volta | Turing |
|---|---|---|---|---|---|---|---|
| 2006 | 2008 | 2010 | 2012 | 2014 | 2016 | 2017 | 2018 |

# Agenda



I. *Arquitectura GPU: Introducción*

II. *GPUs de arquitectura fija y evolución hacia arquitecturas unificadas*

III. *Arquitecturas Nvidia y su evolución*

IV. ***Arquitecturas ATI-AMD***

V. *Otras arquitecturas Manycore*

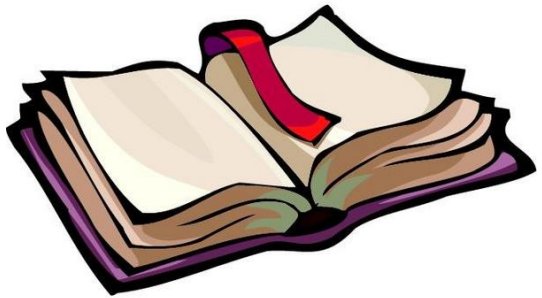
   I. *Intel Xeon PHI*

   II. *Pezy-SC*

   III. *Systema Sunway*

VI. *Rankings*

VII. *GPUs provistas por la cátedra*



# GPUs: Evolución ATI



- ATI fundada en 1985 y comprada por AMD en 2006:
  - 1987: tarjetas graficas EGA Wonder y VESA Wonder
  - 1989: ayuda en el standard VESA
  - 1991 – 1994: Aceleradoras gráficas Mach8, Mach32, Mach64
  - 1996 - 1998: Linea aceleradores 3D Rage, Rage II, Rage Pro, Rage 128 GL
  - 1999 : Rage Movility (Primera tarjeta para portátiles), Rage 128 Pro
  - 2000 : Chip R100 inicio de la linea Radeon
  - 2001 – 2003: R200 y R300, procesadores para consolas (Xbox 360 y Wii)
  - 2004: R400 y R480 sobre PCIexpress
  - 2005: R500 primera en usar shaders unificados

# GPUs: Evolución ATI



- 2006: Comprada por AMD
- 2007: Revision RV670 del chip R600
- 2008: RV770 competencia de Nvidia Geforce 8800
- 2009: RV800 y RV870 gran éxito por el retraso de la arquitectura Nvidia Fermi

- Actualidad: Tecnologia ATI STREAM: tecnologias hardware y software que permiten que la CPU y la GPU trabajen en conjunto en programación de proposito general

# GPUs: Arquitectura ATI



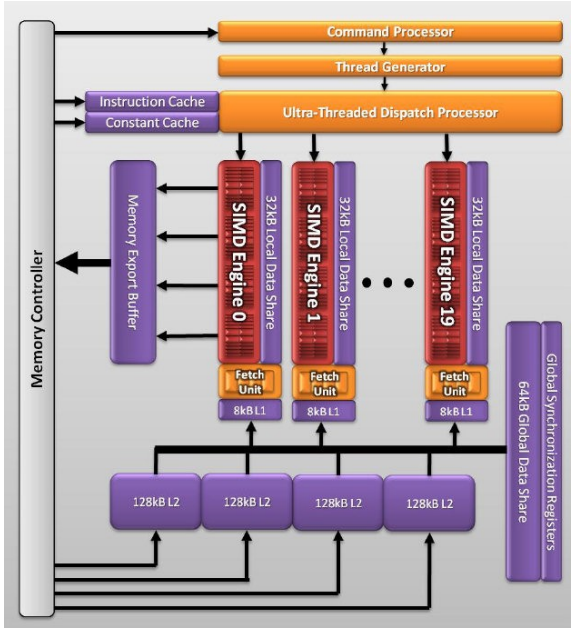
- Tres niveles de procesamiento (uno más que Nvidia):
  - Posee varias **Computing Units** (CU) o SIMD Engines
  - Cada computing unit contiene varios **Stream Cores** (SC)
  - Cada Stream Core posee 5 **Processing Elements** (PE)

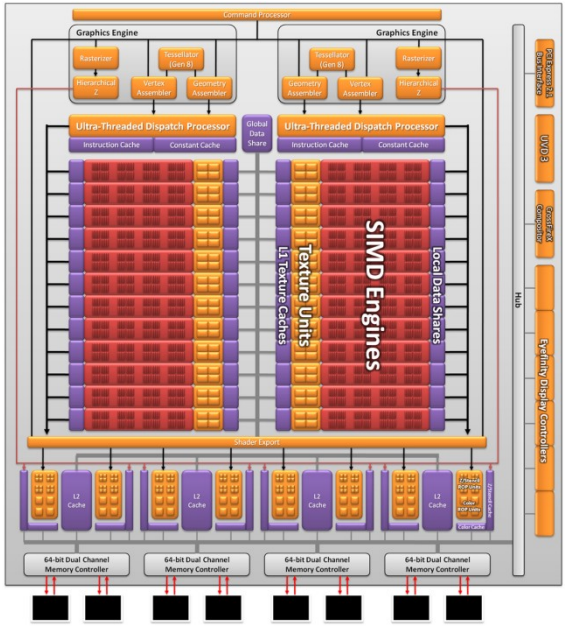
- Cada arquitectura tiene diferentes cantidades de CU y de SCs por CU.

# GPUs: Arquitectura ATI



- **Computing Units** (CU) o SIMD Engines.

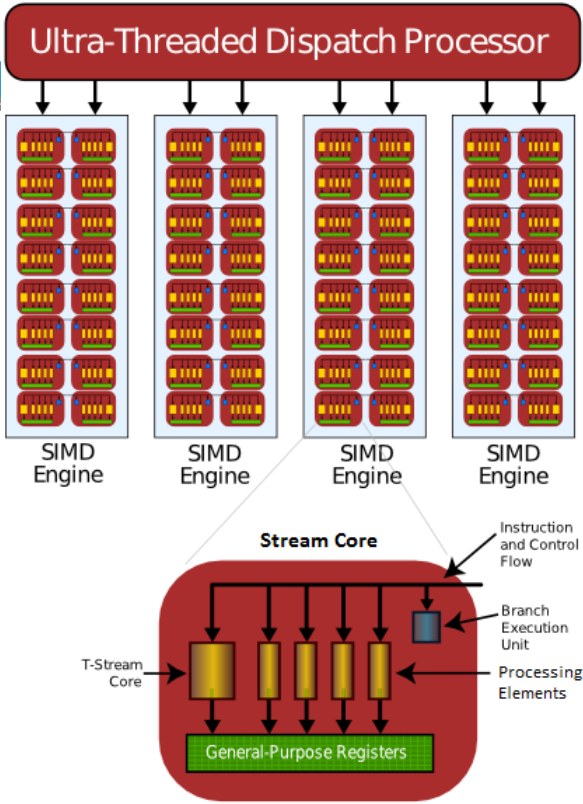


# GPUs: Arquitectura ATI



- **Stream Cores** (SC) y **Processing Elements** (PE).
  - 4 PE puede realizar operaciones:
    - Escalares
    - Punto Flotante
  - El 5to PE realiza operaciones especiales (sin, cos, log, etc).
  - Las operaciones de punto flotante requieren 2 o mas PE.
  - Las operaciones en doble precision necesitan 4 PE.



# GPUs: Arquitectura ATI



- Planificador:
  - Diferentes cargas de trabajo son asignadas a las diferentes SIMD Engines
  - Cada work item (o hilo) es asignado a los SC
  - Se asignan hasta 4 work items por SC
  - En una arquitectura con 16 SC por SIMD Engine, 64 works items son ejecutados juntos
  - El grupo de 64 works items ejecutados juntos se los llama workfronts
  - El numero de work items deberia ser multiplo de workfronts

# Agenda



I. *Arquitectura GPU: Introducción*

II. *GPUs de arquitectura fija y evolución hacia arquitecturas unificadas*

III. *Arquitecturas Nvidia y su evolución*

IV. *Arquitecturas ATI-AMD*

V. ***Otras arquitecturas Manycore***

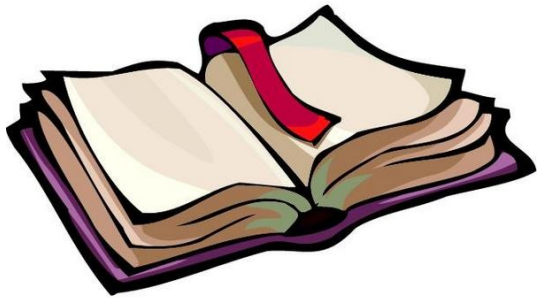
   I. ***Intel Xeon PHI***

   II. ***Pezy-SC***

   III. ***Systema Sunway***

VI. *Rankings*

VII. *GPUs provistas por la cátedra*



# GPU: Arquitecturas Manycore



- A partir del concepto de GPGPU se conoce a las GPUs como arquitecturas Manycore.
- Además de Nvidia y ATI-AMD surgen otras arquitecturas Manycore que no son placas gráficas:
  - Intel Xeon PHI
  - Pezy-SC (Japón)
  - Sistema Sunway TaihuLight (China)

# Agenda



I. *Arquitectura GPU: Introducción*

II. *GPUs de arquitectura fija y evolución hacia arquitecturas unificadas*

III. *Arquitecturas Nvidia y su evolución*

IV. *Arquitecturas ATI-AMD*

V. ***Otras arquitecturas Manycore***

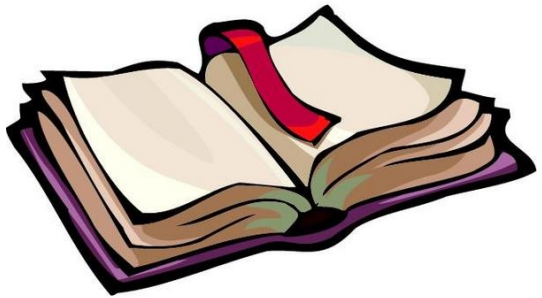
   I. ***Intel Xeon PHI***

   II. *Pezy-SC*

   III. *Systema Sunway*

VI. *Rankings*

VII. *GPUs provistas por la cátedra*



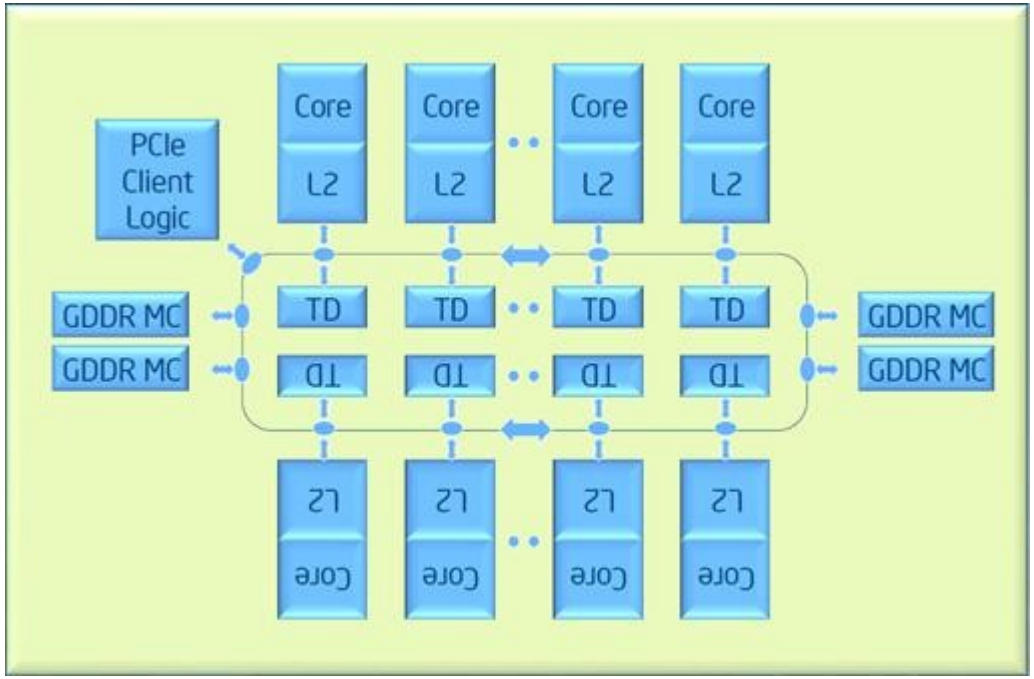
# Intel Xeon Phi



- Intel Many Integrated Core Architecture (MIC), en el mercado Xeon Phi:
  - Coprocesador basado en X86
  - No es una GPU pero se basa en el concepto de GPGPU
  - Más fácil de programar y compilar que en una GPU
    - (100% compatible X86 y X86_64)
  - Menos procesadores que las GPU pero rendimiento similar
  - Uso de herramientas existentes: OpenMP, Cilk, OpenCl etc.
  - Multithreading por hardware

# Intel Xeon Phi: 1ra Generación

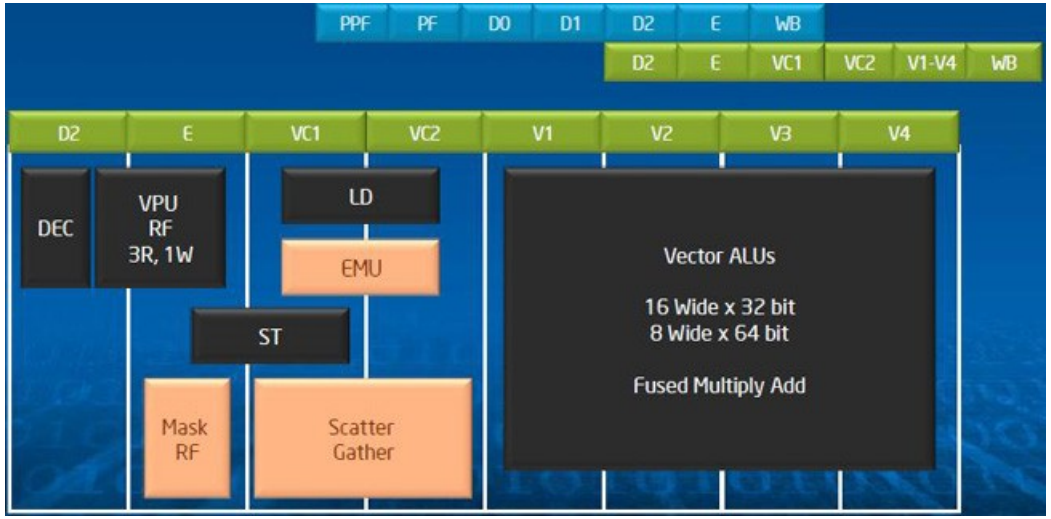





- Cores in-order 4 threads
- Cache L2 por core
- Anillo bidireccional
- 2010 – Modelo Knights Ferry (PCIe):
  - 32 cores, 1.2Ghz, 2GB DDR5
- 2011 – Modelo Knights Corner (PCIe):
  - Basado en pentium P54C
  - 60 cores, 1.2Ghz, 6-16 GB DDR5

# Intel Xeon Phi: 1ra Generación

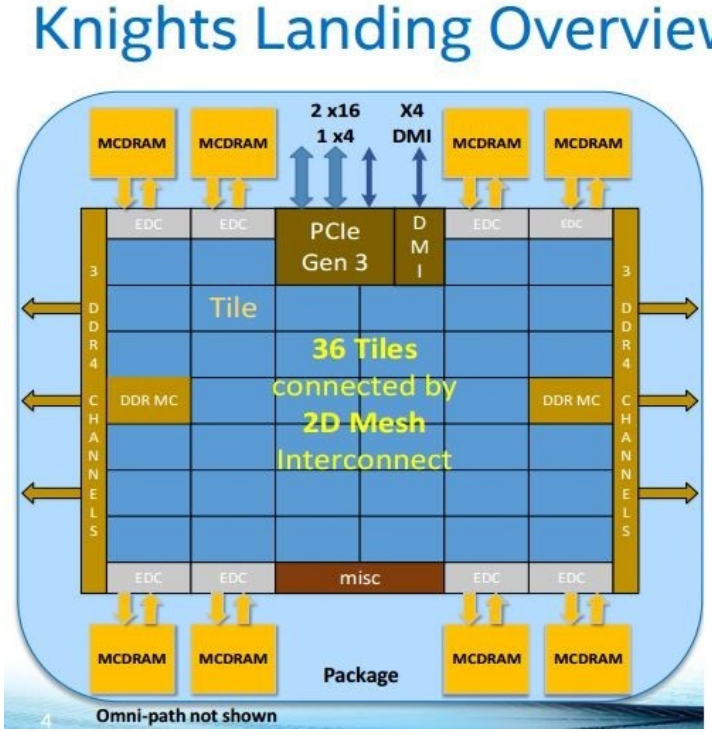





- Vector Processing Unit.
- Novel 512-bit SIMD instruction set.
- Por ciclo: 16 operaciones simple precicion (SP) u 8 en doble (DP).
- Soporta instrucciones Fused Multiply-Add (FMA) 32 SP o 16 DP de punto flotante por ciclo.

# Intel Xeon Phi: 2da Generación






- 2013 – Modelo Knights Landing:
  - PCIe Processor Host
  - Basado en el procesador Atom
  - 64 o 72 cores, 1.3–1.5 Ghz, MCDRAM
  - 256 a 278 hilos (Multithreading por hardware)

# Agenda



I. *Arquitectura GPU: Introducción*

II. *GPUs de arquitectura fija y evolución hacia arquitecturas unificadas*

III. *Arquitecturas Nvidia y su evolución*

IV. *Arquitecturas ATI-AMD*

V. ***Otras arquitecturas Manycore***

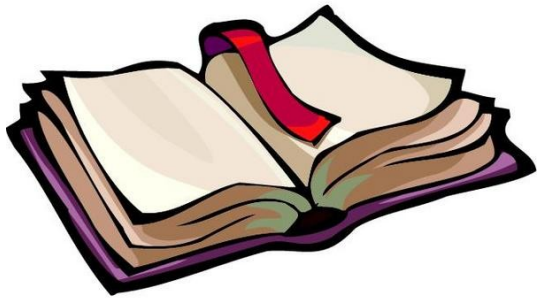
   I. *Intel Xeon PHI*

   II. ***Pezy-SC***

   III. *Systema Sunway*

VI. *Rankings*

VII. *GPUs provistas por la cátedra*

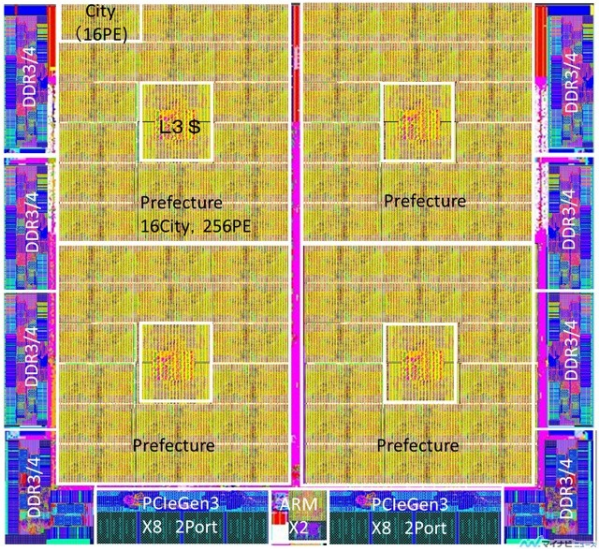


# Pezy-SC



- Pertenece a una corporacion Japonesa.

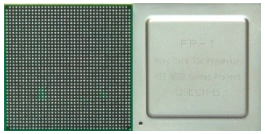
  (http://pezy.co.jp/en/products/pezy-sc.html)



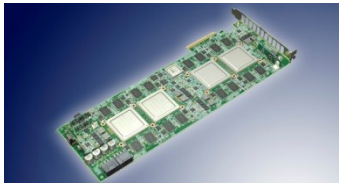
# Pezy-SC



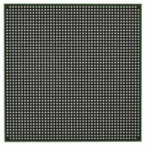
- Primera generación (2012):
  - PEZY-1 processor:
    - 512 cores
    - 666Mhz



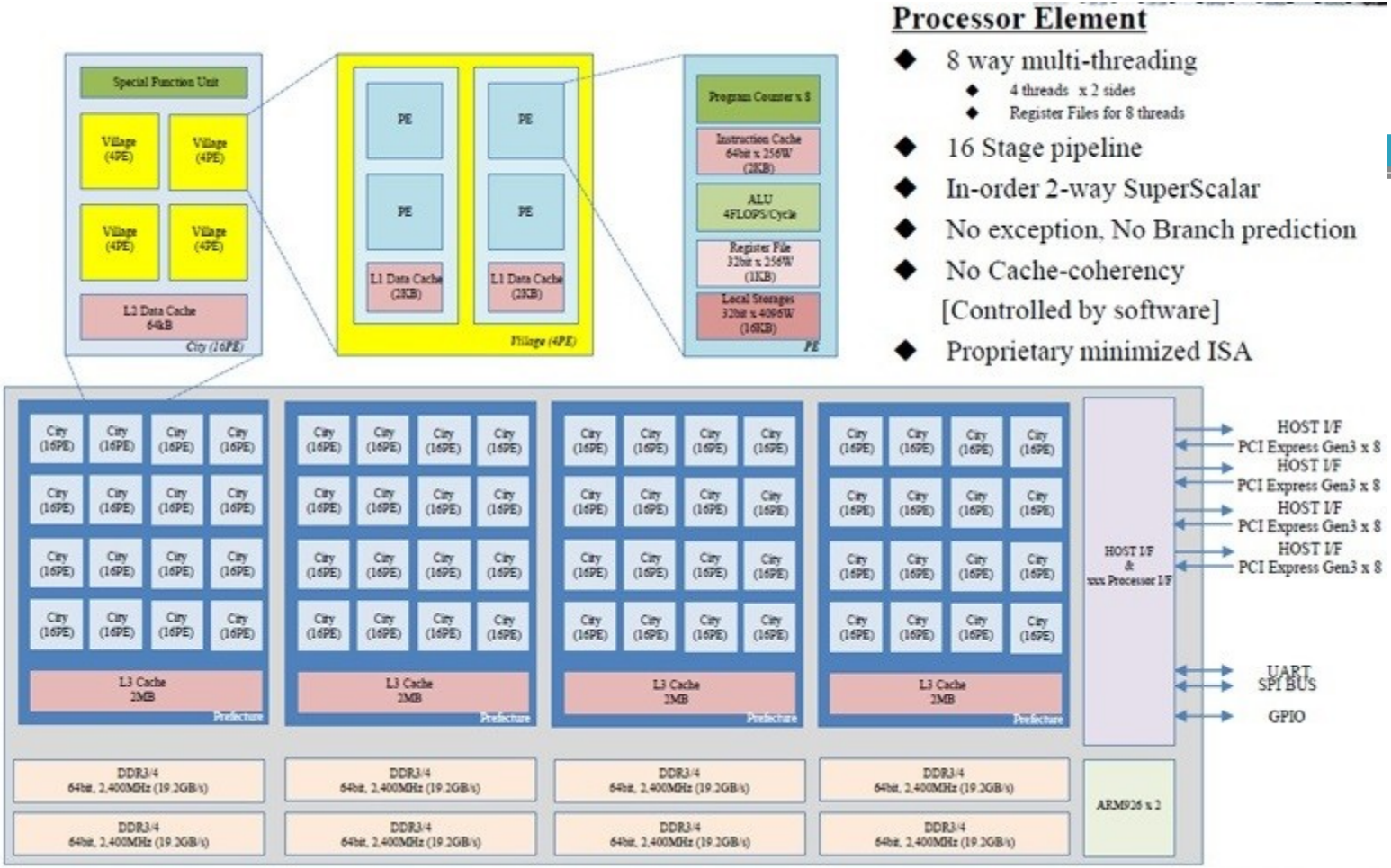
  - PEZY-1 Quad PCI Board:
    - PEZY-1 x 4 = 2048 cores



- Segunda generación (2014):
  - PEZY-SC processor:
    - 1024 cores
    - 733Mhz

  - PEZY-SC Quad PCI Board:
    - PEZY-SC x 4 = 4096 cores



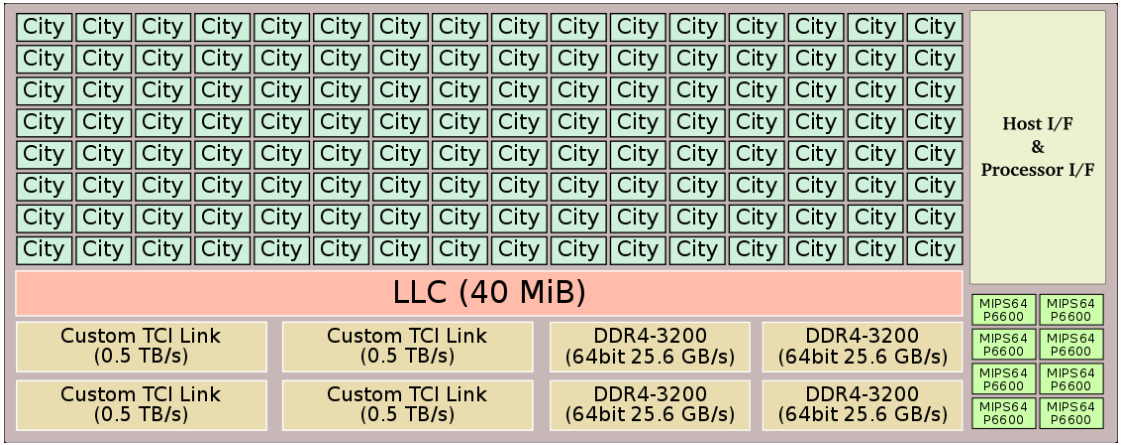

## Processor Element

- 8 way multi-threading
  - 4 threads x 2 sides
  - Register Files for 8 threads
- 16 Stage pipeline
- In-order 2-way SuperScalar
- No exception, No Branch prediction
- No Cache-coherency
  [Controlled by software]
- Proprietary minimized ISA

# Pezy-SC



- PEZY tercera generación (2017):
  - PEZY-SC2 processor:
    - 2048 cores
    - 1Ghz



# Agenda



I. *Arquitectura GPU: Introducción*

II. *GPUs de arquitectura fija y evolución hacia arquitecturas unificadas*

III. *Arquitecturas Nvidia y su evolución*

IV. *Arquitecturas ATI-AMD*

V. ***Otras arquitecturas Manycore***

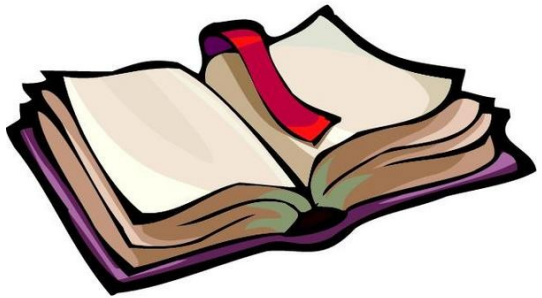
   I. *Intel Xeon PHI*

   II. *Pezy-SC*

   III. ***Systema Sunway***

VI. *Rankings*

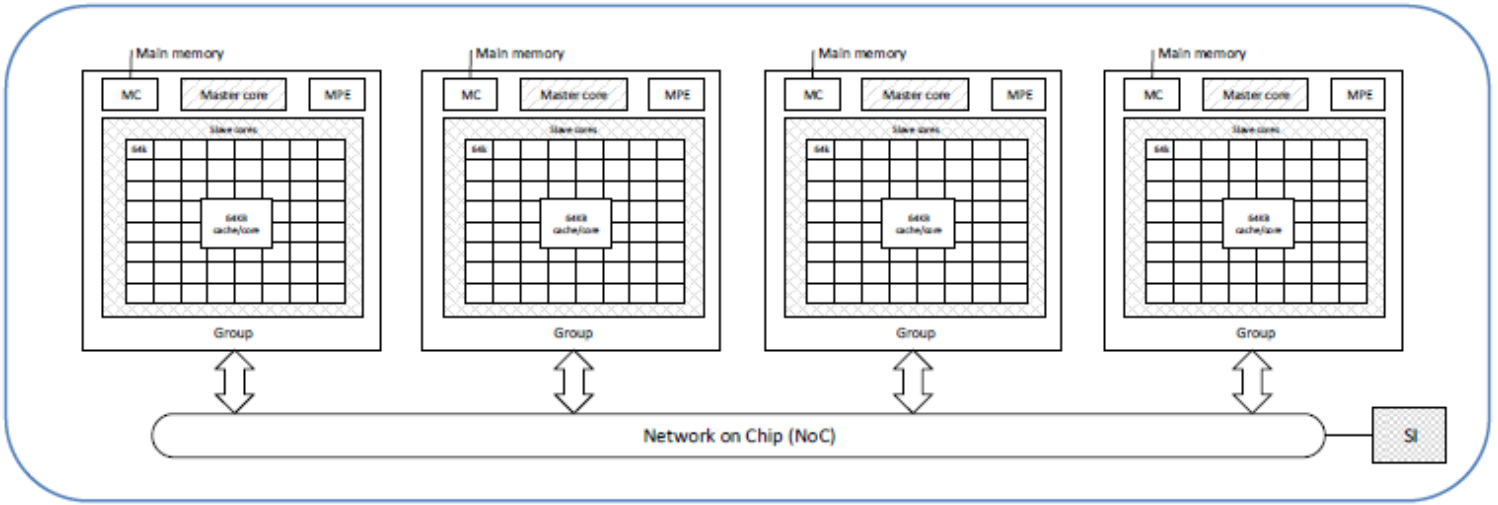
VII. *GPUs provistas por la cátedra*



# Sistema Sunway TaihuLight



- Mega sistema compuesto por 10649600 cores organizados en grupos de multi procesadores.



# Agenda



I. *Arquitectura GPU: Introducción*

II. *GPUs de arquitectura fija y evolución hacia arquitecturas unificadas*

III. *Arquitecturas Nvidia y su evolución*

IV. *Arquitecturas ATI-AMD*

V. *Otras arquitecturas Manycore*

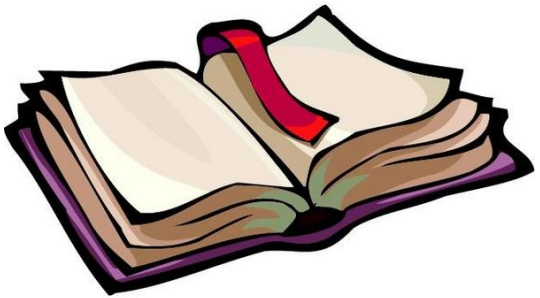
   I. *Intel Xeon PHI*

   II. *Pezy-SC*

   III. *Systema Sunway*

VI. ***Rankings***

VII. *GPUs provistas por la cátedra*



# Arquitecturas paralelas - Evolución



- Existen dos rankings, que se elaboran en junio y noviembre de cada año, que evalúan distintos aspectos de las arquitecturas paralelas:

**Top 500:** proyecto que elabora un ranking de las computadores más rápidas. Utiliza como unidad de medida el Gflops.

**Green 500:** proyecto que elabora un ranking basado en el Top 500 pero considerando la eficiencia energética. Utiliza como unidad de medida el GFlop por Watt.

# GPU: Entre los puestos en TOP 500



| | AMD-ATI | NVIDIA | XEON PHI | Pezy-SC | Sunway |
|---|---|---|---|---|---|
| Nov 2009 | 5° Cluster Radeon | - | - | - | - |
| Jun 2010 | - | 2° | - | - | - |
| Nov 2010 | - | 1° | - | - | - |
| 2011-2012 | - | Primeros 10 | - | - | - |
| Nov 2012 | - | 1° Kepler | 7° | - | - |
| Jun 2013 | - | 2° Kepler | 1° y 6° | - | |
| 2015-2017 | No | 2° - 6° Pascal y Kepler | 1° y 8° | 69° | 1° |
| Jun 2019 | No | 1°-2°, 8° y 10° Volta, 6° Pascal | 7° | No | 3° |

# GPU: Entre los puestos GREEN 500

| | AMD-ATI | NVIDIA | XEON PHI | Pezy-SC | Sunway |
|---|---|---|---|---|---|
| Nov 2009 | 8° Cluster | - | - | - | - |
| Jun 2010 | - | 4° y 8° | - | - | - |
| 2010 - 2011 | Primeros 10 | Primeros 10 | - | - | - |
| Nov 2012 | 2° FirePro | 3° y 4° Kepler | 1° | - | - |
| Jun 2013 | 4° FirePro | 1° y 2° Kepler | 3° | - | - |
| Nov 2014 | 1ª FirePro | 3° al 10° Kepler | 19° | 2° | - |
| Jun 2016 | 4° | 5°-10° Kepler | 44° | 1° y 2° | - |
| Jun 2017 | 42° | 1°-6° y 8°-10° Pascal | 18° | 7° | - |
| Nov 2017 | No | 3°, 6°-10° Volta | 18° | 1° y 3° | 20° |
| Jun 2019 | No | 1°-4°, 6°, 7°, 9°-10° Volta – 5° Pascal | 29° | No | 25° |

# Agenda



I. *Arquitectura GPU: Introducción*

II. *GPUs de arquitectura fija y evolución hacia arquitecturas unificadas*

III. *Arquitecturas Nvidia y su evolución*

IV. *Arquitecturas ATI-AMD*

V. *Otras arquitecturas Manycore*

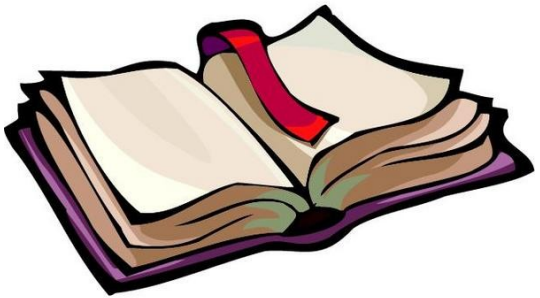
   I. *Intel Xeon PHI*

   II. *Pezy-SC*

   III. *Systema Sunway*

VI. *Rankings*

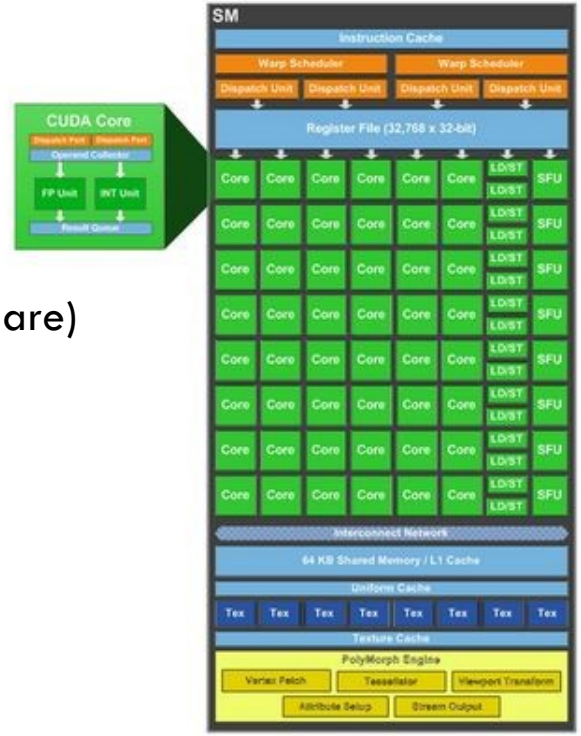
VII. ***GPUs provistas por la cátedra***



# Arquitecturas disponibles en la sala



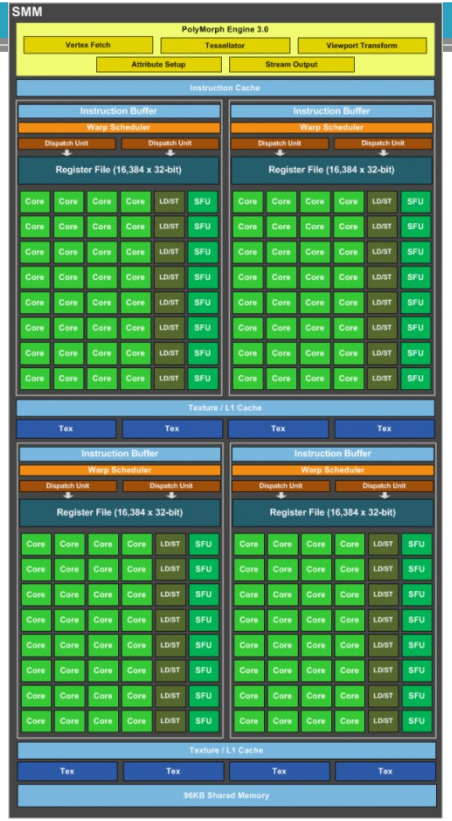
- GPU: Nvidia Geforce 560TI (Fermi):
  - Arquitectura Fermi GF114
  - 8 SMs
  - 48 SPs por SM (32 por SM mas 16 SP de spare)
  - Total 384 cores
  - 1Gb de RAM GDDR5



# Arquitecturas disponibles en la sala



- GPU: Nvidia Geforce 960 (Maxwell):
  - Arquitectura Maxwell GM206-300
  - 8 SMMs x 4 Control Logic
  - Cada control Logic 32 SPs (128SPs por SMM)
  - Total 1024 cores
  - 2Gb de RAM GDDR5



# Arquitecturas disponibles en la sala



- Cada GPU conectada por PCI-Express a un Host Intel i5 2300:
  - 4 cores fisicos
  - 8Gb de RAM
- Hosts forman un cluster de GPU conectado a 1Gbit Ethernet.