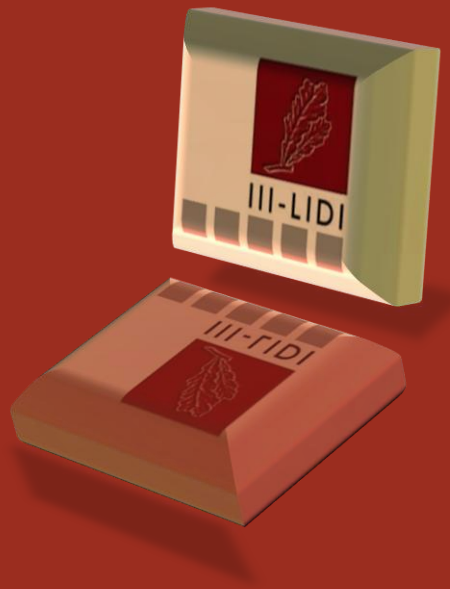
FACULTAD DE INFORMÁTICA

ANÁLISIS DE RENDIMIENTO DE UN ALGORITMO DE CRIPTOGRAFÍA SIMÉTRICA SOBRE GPU Y CLUSTER DE GPU

LIC. ADRIÁN POUSA
LIC. VICTORIA SANZ
ING. ARMANDO DE GIUSTI



HPCLatam 2013

Instituto de Investigación en Informática LIDI - Facultad de Informática - Universidad Nacional de La Plata

Objetivo

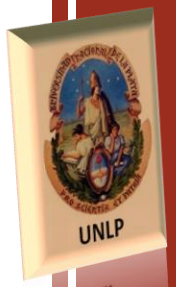
Trabajos previos:

- Análisis de Rendimiento de AES sobre GPU (cuda), Multicore (OpenMP) y Cluster de Multicores (MPI).
- Análisis de Energia sobre GPU (cuda) y Cluster de Multicores (MPI).

Se presenta un análisis de rendimiento del algoritmo de cifrado simétrico por bloques AES (Advanced Encryption Standard) sobre una GPU y un cluster de GPU.

Motivación

- En trabajos previos los requerimientos de memoria del algoritmo no superaban la memoria de la GPU.
- Disponemos de GPU de poca memoria y no expandible.
- Disponemos de un cluster de GPUs.
- Si los requerimientos de memoria del algoritmo superan la memoria de una GPU:
  - Usar una sola GPU: fragmentar los datos e invocar varias veces al kernel?
  - Usar el cluster de GPU ??? - Impacto de las comunicaciones.



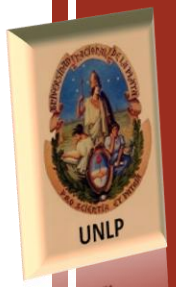
AES: Características

- Algoritmo simétrico: misma clave para cifrar y descifrar.
- Operaciones sobre bytes.
- Algoritmo por bloques :Los datos a encriptar se dividen en bloques de tamaño fijo (128 bits), cada bloque se representa como una matriz de 4x4 bytes llamada *estado.*

A1	CC	B0	11	23	F1	E2	07	A8	01	AA	CB	90	5E	34	D1

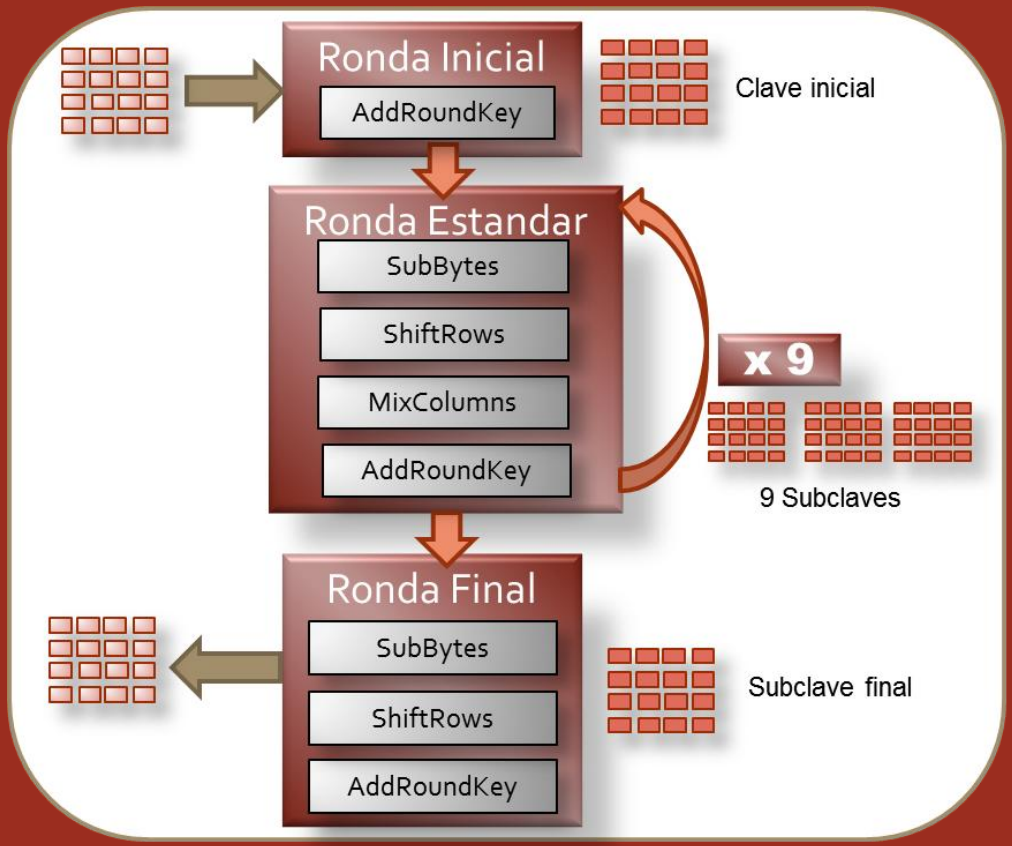
A1	23	A8	90
CC	F1	01	5E
B0	E2	AA	34
11	07	CB	D1

- *Clave inicial de 128 bits (192 bits ó 256 bits).*
- *La clave se representa igual que el estado.*
- *Se generan 10 claves (total 11 claves).*



AES: Funcionamiento

- Rondas de operaciones cifrado (descifrado orden inverso):



Arquitecturas utilizadas

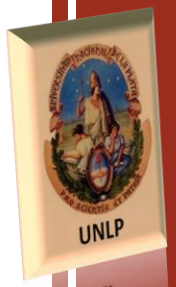
- GPU:
  - ✓ Nvidia Geforce 560TI – 384 cores – 1Gb Memoria.
  - ✓ Conectada en un máquina Intel i5 2300.
- Cluster de GPU:
  - ✓ GPU igual a la anterior.
  - ✓ Conectividad Ethernet 1Gbit.

AES - Implementaciónes

Casos de prueba de interés: la memoria utilizada por el algoritmo supera la capacidad de memoria de una GPU.

Dos implementaciones:

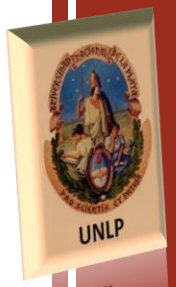
- AES-GPU: una única GPU invocando el kernel Cuda varias veces.
- AES-ClusterGPU: Distribución de datos por MPI entre los nodos y cifrado sobre la GPU de cada nodo invocando el mismo kernel Cuda.



AES - Implementación Cuda

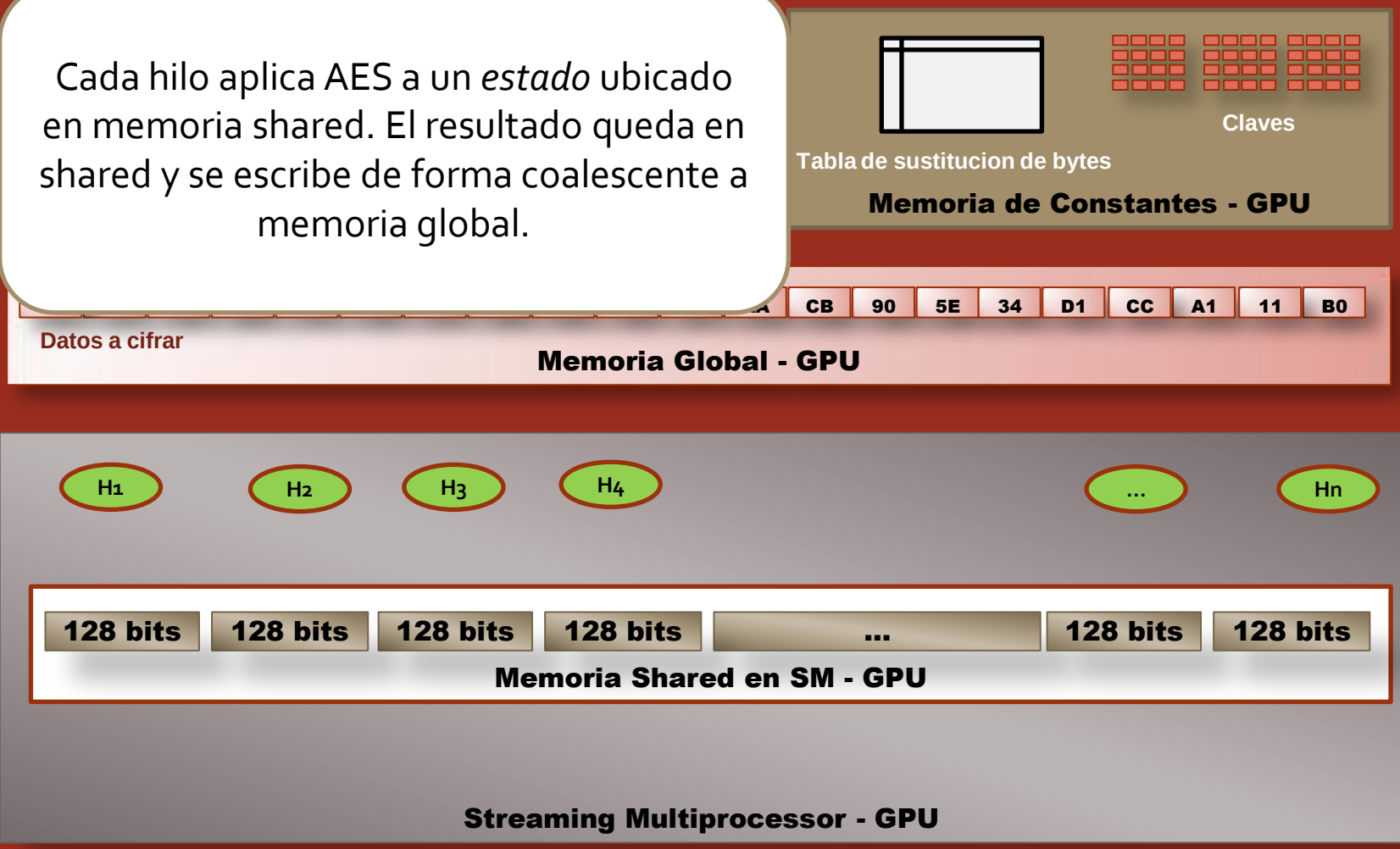
- El Host:
  - ✓ Calcula las claves.
  - ✓ Copia claves, tablas auxiliares y datos de entrada a la GPU.
  - ✓ Invoca al kernel (creando bloques de hilos).
  - ✓ Recupera de la GPU los datos cifrados.

- Cada hilo:
  - ✓ Coopera con los demás hilos del bloque para cargar los datos a cifrar a memoria *shared. (Acceso Coalescente)*
  - ✓ Cifra un *estado.*
  - ✓ Coopera con los demás hilos del bloque para copiar los datos a cifrados a memoria *global. (Acceso Coalescente)*



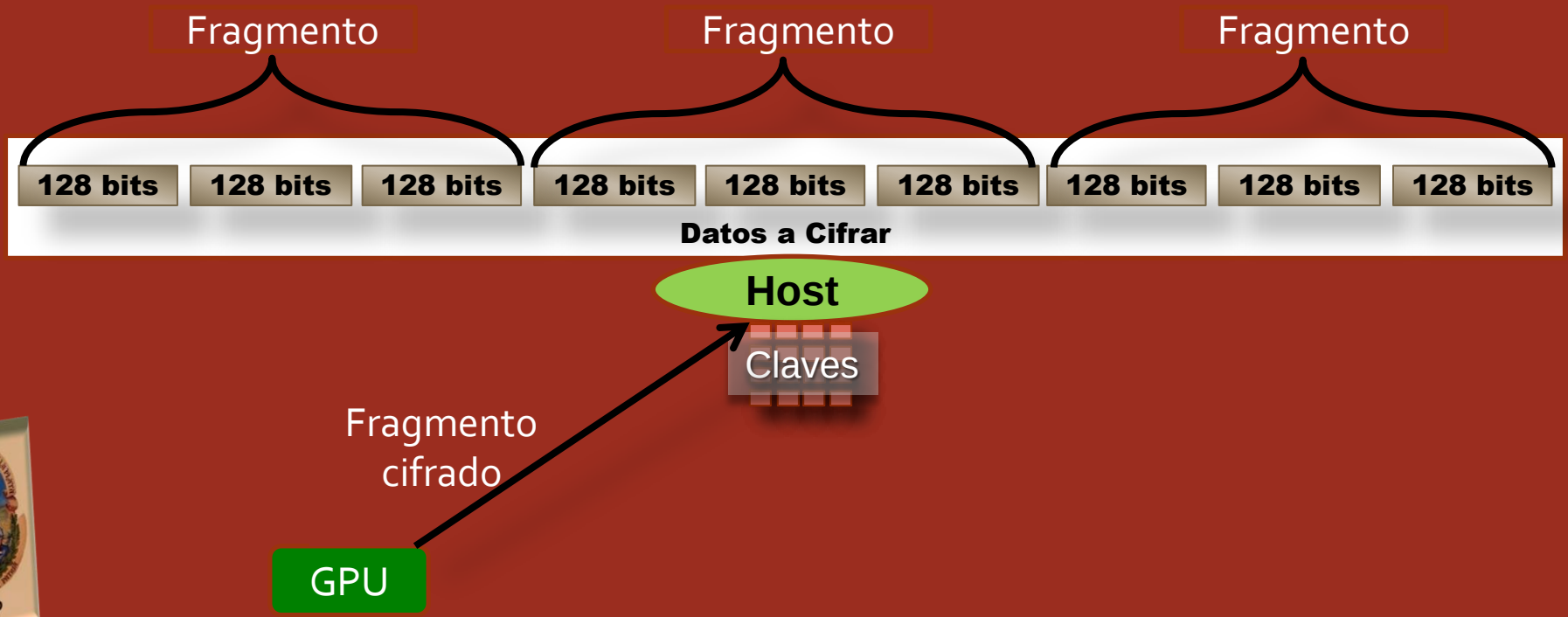
AES - Implementación Cuda

Cada hilo aplica AES a un *estado* ubicado en memoria shared. El resultado queda en shared y se escribe de forma coalescente a memoria global.



Implementación AES-GPU

- División en Fragmentos (no > a 256MB).
- Invocación al kernel para cada fragmento.



Implementación AES-ClusterGPU

- Un proceso genera las claves y la envía junto con un conjunto consecutivo de *estados* a los procesos restantes.
- Cada proceso luego invoca al kernel Cuda.

Rendimiento AES-GPU
Tamaño de Fragmentos

- Tiempo (en seg) para Fragmentos de 128MB:

Datos de entrada =>	256MB (2 fragmentos)	512MB (4 fragmentos)	1GB (8 fragmentos)
Tiempo total de copia CPU-GPU	0,048	0,095	0,18
Tiempo total de Cifrado	0,81	1,63	3,26
Tiempo total de copia GPU-CPU	0,074	0,11	0,21
Tiempo total	**0,932**	**1,835**	**3,65**

- Tiempo (en seg) para Fragmentos de 256MB:

Datos de entrada =>	256MB (1 fragmento)	512MB (2 fragmentos)	1GB (4 fragmentos)
Tiempo total de copia CPU-GPU	0,048	0,093	0,18
Tiempo total de Cifrado	0,81	1,63	3,26
Tiempo total de copia GPU-CPU	0,1	0,14	0,23
Tiempo total	**0,958**	**1,863**	**3,67**

Rendimiento AES-ClusterGPU

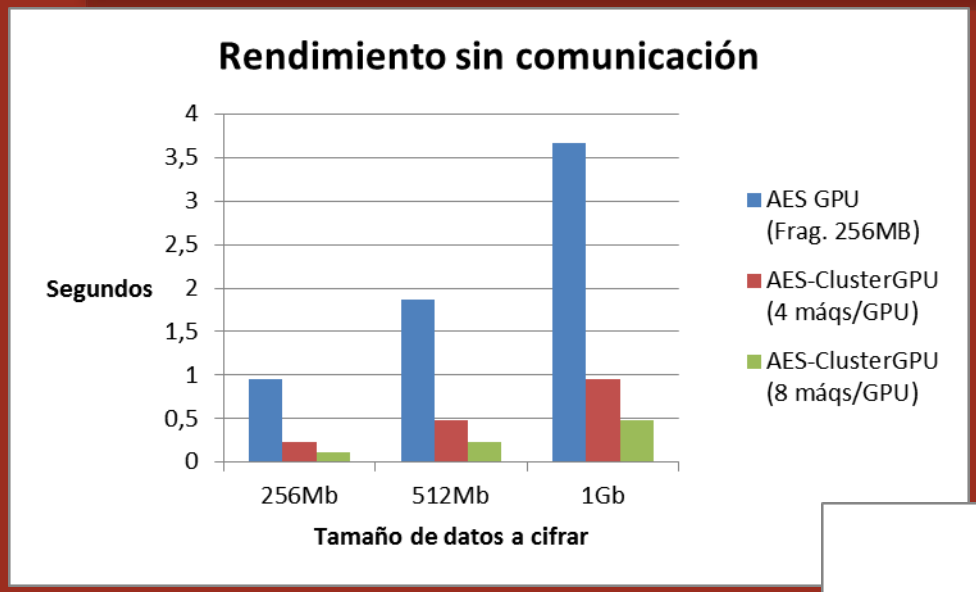
- Tiempo (en seg) para un Cluster 4 GPUs:

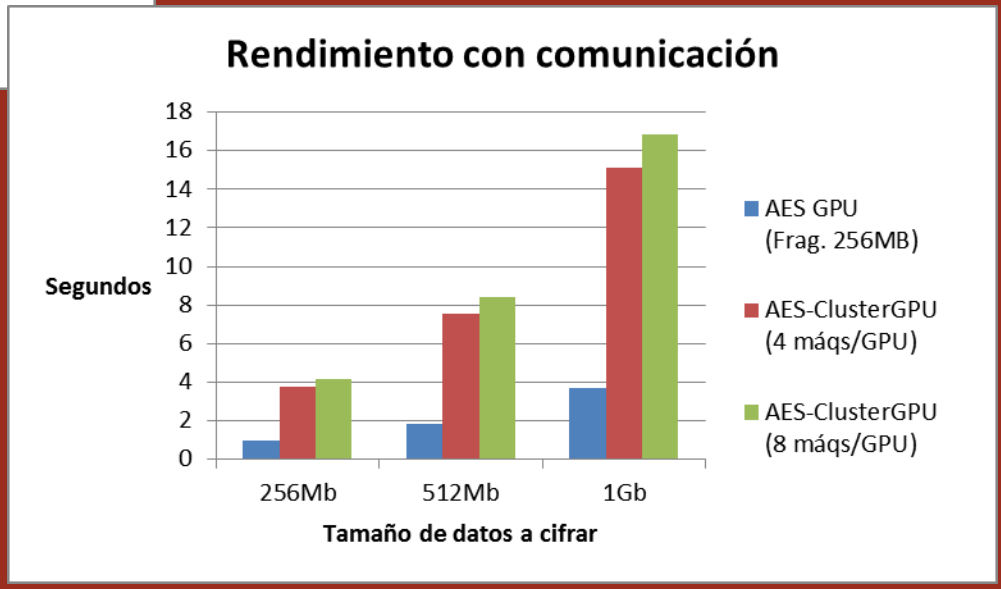
	256MB (64MB por GPU)	512MB (128MB por GPU)	1GB (256MB por GPU)
Tiempo de comunicación envío	1,76	3,53	7,07
Tiempo de copia CPU-GPU por nodo	0,012	0,024	0,048
Tiempo de Cifrado por nodo	0,2	0,4	0,81
Tiempo de copia GPU-CPU por nodo	0,025	0,051	0,1
Tiempo de comunicación recepción	1,76	3,53	7,06
Tiempo total con comunicaciones	3,757	7,535	15,088
Tiempo total sin comunicaciones	0,237	0,475	0,958

- Tiempo (en seg) para un Cluster 8 GPUs:

	256MB (32MB por GPU)	512MB (64MB por GPU)	1GB (128MB por GPU)
Tiempo de comunicación envío	2,04	4,06	8,1
Tiempo de copia CPU-GPU por nodo	0,007	0,012	0,025
Tiempo de Cifrado por nodo	0,1	0,2	0,4
Tiempo de copia GPU-CPU por nodo	0,012	0,026	0,051
Tiempo de comunicación recepción	2,03	4,14	8,28
Tiempo total con comunicaciones	4,189	8,438	16,856
Tiempo total sin comunicaciones	0,119	0,238	0,476

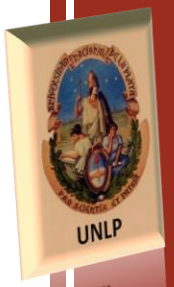
Rendimiento AES-ClusterGPU





Conclusiones y Trabajo Futuro

- Las comunicaciones en el cluster introducen un incremento importante en el tiempo total de cómputo.
- Para este algoritmo en particulary con la arquitectura que disponemos es más adecuado dividir los datos en fragmentos y cifrar cada uno utilizando una única GPU, que utilizar un cluster de GPU.
- Trabajo futuro:
  - Análisis sobre aplicaciones donde el impacto de las comunicaciones sea menor.
  - Optimizaciones de algoritmos para arquitecturas hibridas (GPU, multicores, clusters).
  - Análisis de rendimiento y energía.



Muchas Gracias ...