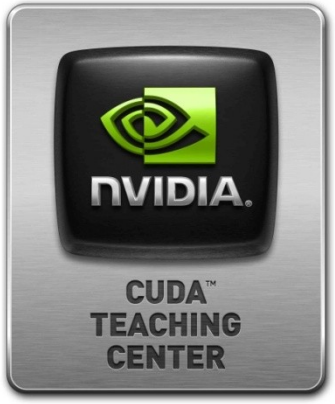
Computación de Propósito General en Unidades de Procesamiento Gráfico (GPGPU)

E. Dufrechou, P. Ezzatti y M. Pedemonte



Clase 12
ALN en GPUs

Contenido

- Motivación
- Conceptos básicos de ALN
  - Problemas
  - Tipo de datos y matrices
  - Estándares - bibliotecas
- ALN densa con GPUs
- ALN dispersa con GPUs
- Algunos trabajos en FING

Motivación

Motivación

- **Muchas aplicaciones "científicas" se basan en la resolución de problemas de ALN:**
  - **Optimización y Simulación**
  - **Computación gráfica**
  - **Control**
  - **BDs, BDs de grafos**
  - **Redes Neuronales**
- **Campo con grandes requerimientos de poder de cómputo.**
- ***(Morpheus) The Matrix is everywhere, it is all around us. Even now, in this very room.*** ***You can see it when you look out your window or when you turn on your television. You can feel it when you go to work…when you go to church…when you pay your taxes.***

Conceptos básicos de ALN

Conceptos básicos de ALN

Problemas tratados

- Operaciones con matrices y vectores
- Multiplicación matriz–vector, matriz-matriz
- Factorización LU, Cholesky, QR
- Valores y vectores propios
- Descomposición SVD

Conceptos básicos de ALN

- Tipo de datos (Punto flotante)
  - Simple – doble precisión
  - Reales complejos (simple y doble precisión)

- Tipo de matrices
  - Densas
  - Triangulares
  - De banda
  - Dispersas no estructuradas

Conceptos básicos de ALN

- **Ejemplo típico la resolución de sistemas lineales de ecuaciones ..**

Conceptos básicos de ALN

Métodos de resolución

Dos grandes familias:

Métodos directos: Llegan a la solución en un número específico de pasos, a menos de errores numéricos consiguen la solución exacta. **(fact. LU)**

Basados en diferentes operaciones básicas.

www.netlib.org/lapack/lapack-3.1.1/html/dgetrf.f.html

Métodos iterativos: Intentan aproximar la solución mediante una sucesión generada iterativamente. Si el método converge se obtiene una solución aproximada del problema, cuyo error satisface algún criterio prefijado. **(Gradiente Conjugado -GC-)**

Basados en la multiplicación matriz-vector.

Conceptos básicos de ALN

Sobre matrices densas (triangulares, banda):

Uso intensivo de estándares (bibliotecas)

- **BLAS, LAPACK**

Fuerte desarrollo de técnicas de HPC

- **BLAS (multi-hilo), LAPACK, SCALAPACK**

Conceptos básicos de ALN

BLAS (visto en más detalle en Ecosistema por cublas)

Resuelve operaciones

- **Vector-vector**
- **Matriz-vector**
- **Matriz-Matriz**

Para distinto tipo de matrices

- **Simple, doble, complejos**
- **Densas, triangulares, de banda**

Desarrollada desde los años 70s

Implementada originalmente en FORTRAN, referencias OpenBlas y MKL.

Conceptos básicos de ALN

LAPACK

- **Ofrece operaciones como:**

  Factorización LU, Cholesky, QR, etc.
- **Implementación de referencia utiliza BLAS para resolver problemas de base.**
- **Se dispone de implementaciones multi-core de la biblioteca (MKL)**

Conceptos básicos de ALN

Álgebra dispersa:

- Amplio uso de bibliotecas (MUMPS, SuperLU, PARDISO).
- Los métodos directos pueden romper la dispersión.
  - Control del problema del fill-in.
  - Resolución de sub-problemas densos.
- Los métodos iterativos son más sencillos de implementar
  - Se basan en la multiplicación matriz dispersa – vector (SpMv).
  - En ocasiones no son lo suficiente precisos.
- Gran desarrollo de los kernels básicos ...

Conceptos básicos de ALN

Álgebra dispersa:

- Multiplicación matriz dispersa vector (**SpMv, SpMM**)
  - Operación principal en las iteraciones, por ejemplo en la resolución de sistemas lineales por métodos iterativos.
- Resolución de sistemas lineales dispersos triangulares (**SpTrsv**)
  - Presente, como parte de los métodos de precondicionado, en los métodos iterativos de resolución de sistemas lineales
- Multiplicación matriz dispersa matriz dispersa (**SpGeMM**)
  - Operación que ha cobrado importancia en los últimos años con el creciente uso de operaciones de grafos.

Conceptos básicos de ALN

LINPACK (Benchmark basado en operaciones de ALN densa)

- **Factorización LU**
- **Utilizado para confeccionar el top500**
  - **Lista con las 500 computadoras más potentes del mundo.**

HPCG (Benchmark basado en operaciones de ALN dispersa)

- **SpMV para resolver CG**

graph500 (Benchmark basado en op. de ALN dispersa)

- **Búsquedas en grafos, mapeable a operaciones de ALN**

Algo de historia de ALN y GPU

Algo de historia de ALN y GPU

Trabajos pioneros

Multiplicación de matrices

- [2001, Larsen y McAllister] Trabajan con números en la precisión disponible en las GPUs de la época, 8 bits.

Resolución de sistemas lineales con métodos iterativos

- [2001, Rumpf y Strzodka] Resolución de sistemas lineales con el método de Jacobi para la resolución de ecuaciones diferenciales mediante el MEF.

Algo de historia de ALN y GPU

Multiplicación de matrices (extensión al trabajo previo)

- [2003, Moravansky] Cambio de acceso.
- [2004, Fatahalian y otros] Evalúan restricciones en las tranferencias de datos.

Resolución de sistemas lineales con métodos iterativos

- [2003, Bolz y otros] Implementan distintos métodos iterativos de resolución de sistemas lineales (multigrid, gradiente conjugado).

Factorización LU

- [2005, Galoppo y otros] Implementaciones pioneras.

Algo de historia de ALN y GPU

Bases para BLAS

- [2003, Kruger y Westermann] Primeras ideas tendientes sobre implementación de BLAS en GPU.

Con CUDA

- Se incluye CUBLAS
  - Implementación en GPU de BLAS.
  - Comienza en simple precisión, luego evoluciona ..

Algo de historia de ALN y GPU

Mejoras a CUBLAS

- [2008, Volkov y Demmel] Importantes mejoras en la operación gemm (multiplicación de matrices generales). Luego incorporado a CUBLAS.

- [2008, Barrachina y otros] Uso de padding y estrategias híbridas (CPU+GPU).

Algo de historia de ALN y GPU

Precisión mixta

- **Estrategias utilizadas en los primeros años de cálculo numérico.**

1. **Resolver un problema en una precisión (más barato) .**
2. **Refinar el resultado en otra precisión.**

LU = L U = lu(A)	SIMPLE	$O(n^3)$
x = L\(U\b)	SIMPLE	$O(n^2)$
r = b – Ax	DOBLE	$O(n^2)$
WHILE || r || not small enough		
z = L\(U\r)	SIMPLE	$O(n^2)$
x = x + z	DOBLE	$O(n^1)$
r = b – Ax	DOBLE	$O(n^2)$
END		

Algo de historia de ALN y GPU

Precisión mixta (despertares, la película, varias veces)

- [2006, Strzodka y Göddeke] Utilización de técnicas de precisión mixta.
- Diversos trabajos que permitían alcanzar doble precisión.
- Diversos trabajos que permitían acelerar el uso de doble precisión.
- Paulatinamente, se fue perdiendo el interés en el tema ..
- Desde Kepler, vuelven a crecer las diferencias entre simple y doble.
- Ahora: half precision, adaptativos ...
- [2020, H. Anzt et al.] GINKGO library

ALN densa en GPUs

ALN densa en GPUs

- Uno de los tópicos principales para GPU
  - Relacionado con computación gráfica.
  - Muchos problemas necesitan la resolución de problemas de ALN.
- Trabajos basados en CUBLAS
  - Permite adaptarse a nuevo hardware fácilmente.
  - Aportes en CUBLAS impactan en los desarrollos “automáticamente”.

ALN densa en GPUs

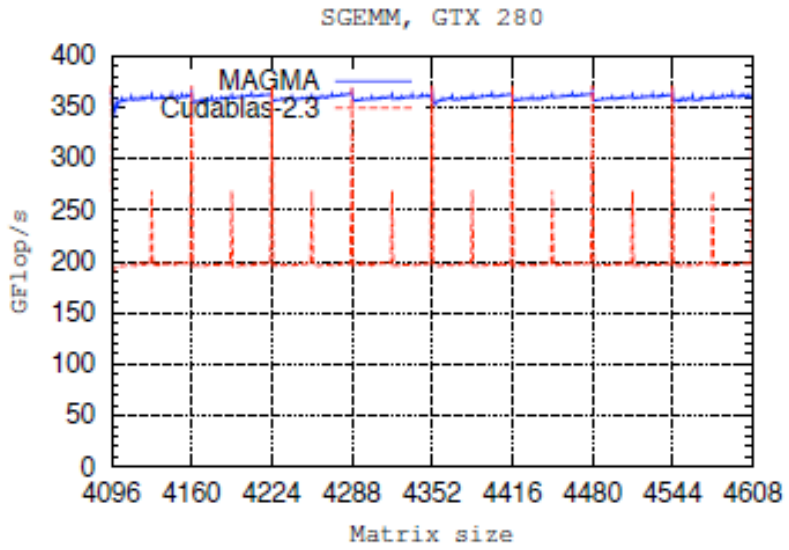
- **Procesamiento de matrices a bloques**

$$A=\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}=\begin{pmatrix} L_{11} & 0 \\ L_{21} & L_{22} \end{pmatrix}\begin{pmatrix} U_{11} & U_{12} \\ 0 & U_{22} \end{pmatrix}$$

Permite acceso regular, explotar warps, etc !!!!

- **Utilización de padding**

$$\tilde{A} = \begin{bmatrix} A & 0 \\ 0 & I_{p-n} \end{bmatrix}$$



ALN densa en GPUs

- Estrategias híbridas CPU+GPU
  - Cada etapa del problema en la mejor arquitectura.
  - Concurrencia
  - Un trozo en cada arquitectura, LINPACK (TOP500).
- Utilización de múltiples GPUs
  - Mayor poder de cómputo.
  - Mayor memoria.

ALN densa en GPUs

- Biblotecas
  - cublas
  - cusolver
  - magma
  - CULA
  - culapack

ALN dispersa en GPUs

ALN dispersa en GPUs

- Originalmente, mucho menos trabajo que para matrices densas
- Implementaciones de problemas particulares.
- Métodos directos (casi no hay trabajos)
  - Aceleración de etapas densas
- Métodos iterativos
  - Basados en la operación sparse matrix-vector multiplication (SpMv).
  - Problemas con accesos irregulares, más difíciles para GPU !!

ALN dispersa en GPUs

Cusparse

Biblioteca para trabajar con matrices dispersas

3 niveles de operaciones:

- Vector disperso – vector denso
- Matriz dispersa – vector denso
- Matriz dispersa – vectores densos (matriz)

El usuario se encarga de la reserva y transferencia de memoria

ALN dispersa en GPUs

Cusparse

Soporta diferentes tipos de datos

- **Float, Double, cuComplex, cuDoubleComplex**

Índices

- **0 (tipo C) y 1 (tipo Fortran)**

Diferentes formatos dispersos:

- **Denso, Elemental (coordinado, COO), CSR, CSC, BSR**

ALN dispersa en GPUs

Cusparse

Funciones:

- Transformaciones de tipo de dato
- Nivel 1 : axpy, doti, gthr, gthrz, roti y sctr
- Niveles 2 y 3: mv, sv, mm, sm

Resolución de sistemas lineales triangulares (2 tipos de funciones):

- Simbólicas, `cusparseSpSM_analysis`
- Numéricas, `cusparseSpSM_solve`

ALN densa en GPUs

- Bibliotecas
  - Cusparse (NVIDIA)
  - Magma-sparse (ICL - Dongarra)
  - GINKGO (KIT - Anzt)

ALN dispersa en GPUs

Operaciones – SpMv (o SpMM)

- Presente desde los trabajos pioneros
- Paralelismo a nivel de filas
- Diversos trabajos sobre el formato disperso utilizado para la operación
- Formatos particulares para mejorar la localidad de datos

ALN dispersa en GPUs

Operaciones – SpTrsv

- Dos paradigmas:
  - Por niveles (NVIDIA)
  - Sync-free (los más eficientes, adoptado luego también por NVIDIA).
- Especial desarrollo en los últimos 5-7 años.

ALN dispersa en GPUs

Operaciones – SpGeMM

- Pocos trabajos
- Operación con mayor dificultad que las anteriores (dos patrones dispersos)
- Formatos para procesar los bloques.

Algunos trabajos en FING (vintage)

Algunos trabajos en FING

Factorización de matrices generales

- Implementaciones basadas en LAPACK.
- Trabajo a bloques
- Estrategias híbridas (CPU+GPU).

$$\left(\begin{array}{c|c} A_{TL} & A_{TR} \\ \hline A_{BL} & A_{BR} \end{array}\right) \rightarrow \left(\begin{array}{c|c|c} A_{00} & A_{01} & A_{02} \\ \hline A_{10} & A_{11} & A_{12} \\ \hline A_{20} & A_{21} & A_{22} \end{array}\right)$$

where A_{11} is $b \times b$

$$\begin{array}{rcl} \left(\begin{bmatrix} A_{11} \\ A_{21} \end{bmatrix}\right) & := & \mathrm{LU}_{\mathrm{UNB}}\left(\begin{bmatrix} A_{11} \\ A_{21} \end{bmatrix}\right) \\ A_{12} & := & A_{11}^{-1} A_{12} \\ A_{22} & := & A_{22} - A_{21} A_{12} \end{array}$$

Algunos trabajos en FING

Inversión de Matrices generales

- Implementaciones basadas en LAPACK.
- Implementaciones basadas en Gauss-Jordan.
- Estrategias híbridas.
- Estrategias concurrentes.

* Using graphics processors to accelerate the computation of the matrix inverse. P. Ezzatti, E. Quintana, A. Remón, J. of Supercomputing 58(3): 429-437 (2011).

Algunos trabajos en FING

Inversión de Matrices generales

- **Implementaciones basadas en LAPACK.**

$$
\begin{array}{rcl}
\text{GPU} & \rightarrow & \text{CPU}\left(\left[\begin{array}{c} A_{11}; \\ A_{21} \end{array}\right]\right) \\
\left(\left[\begin{array}{c} A_{11} \\ A_{21} \end{array}\right]\right) & := & \text{LU}_{\text{UNB}}\left(\left[\begin{array}{c} A_{11} \\ A_{21} \end{array}\right]\right) \\
\text{CPU} & \rightarrow & \text{GPU}\left(\left[\begin{array}{c} A_{11}; \\ A_{21} \end{array}\right]\right) \\
A_{12} & := & A_{11}^{-1} A_{12} \\
A_{22} & := & A_{22} - A_{21} A_{12}
\end{array}
$$

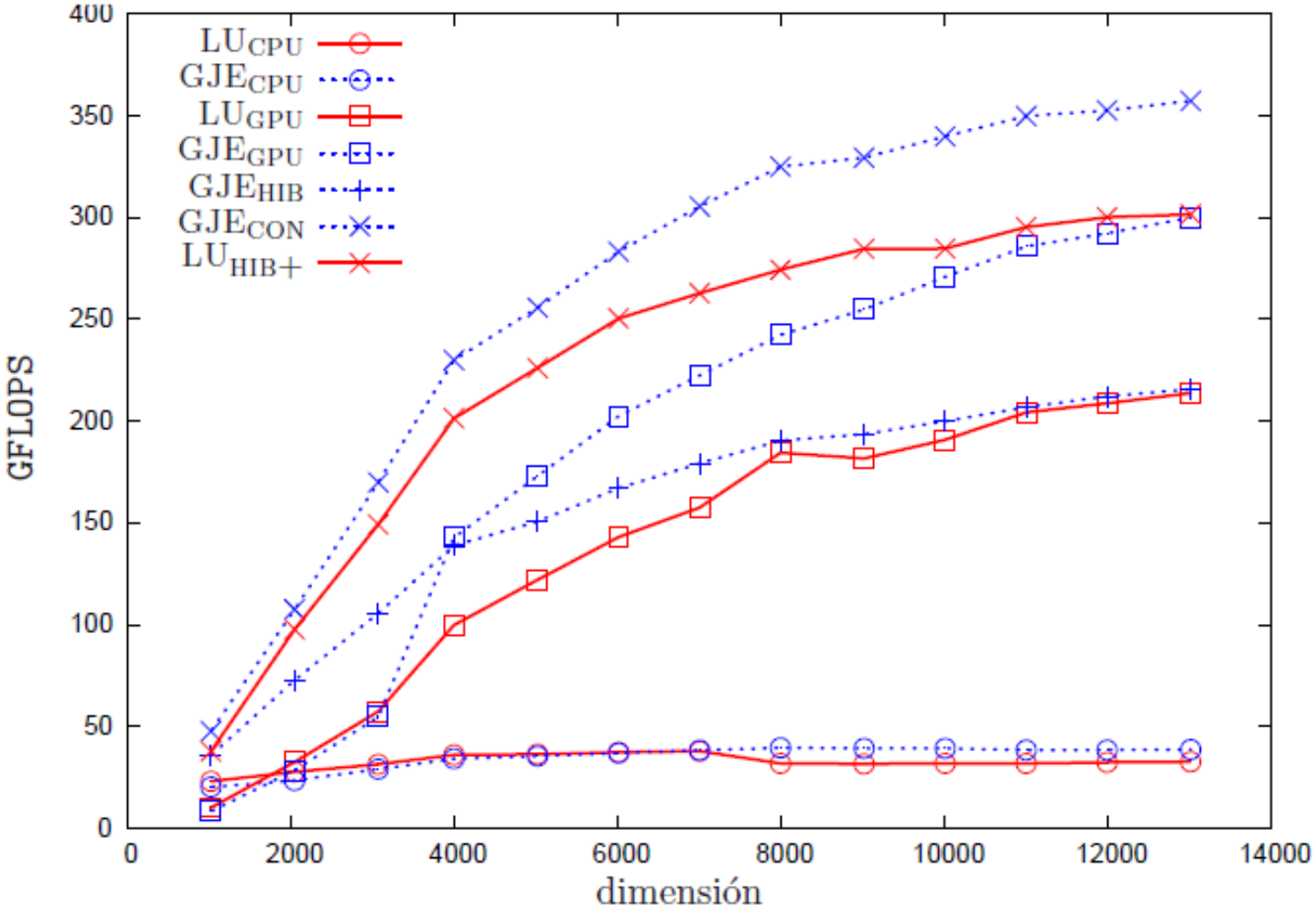
- **Implementaciones basadas en Gauss-Jordan.**
  - **Estrategias híbridas.**
  - **Estrategias concurrentes.**

Algunos trabajos en FING

$$
\begin{array}{rcl}
\left[\begin{array}{c} A_{01} \\ A_{11} \\ A_{21} \end{array} \right] & := & \mathrm{GJE}_{\mathrm{UNB}} \left(\left[\begin{array}{c} A_{01} \\ A_{11} \\ A_{21} \end{array} \right] \right) \\
A_{00} & := & A_{00} + A_{01} A_{10} \\
A_{20} & := & A_{20} + A_{21} A_{10} \\
A_{10} & := & A_{11} A_{10} \\
A_{02} & := & A_{02} + A_{01} A_{12} \\
A_{22} & := & A_{22} + A_{21} A_{12} \\
A_{12} & := & A_{11} A_{12}
\end{array}
$$

Estrategia de resolución

Algunos trabajos en FING



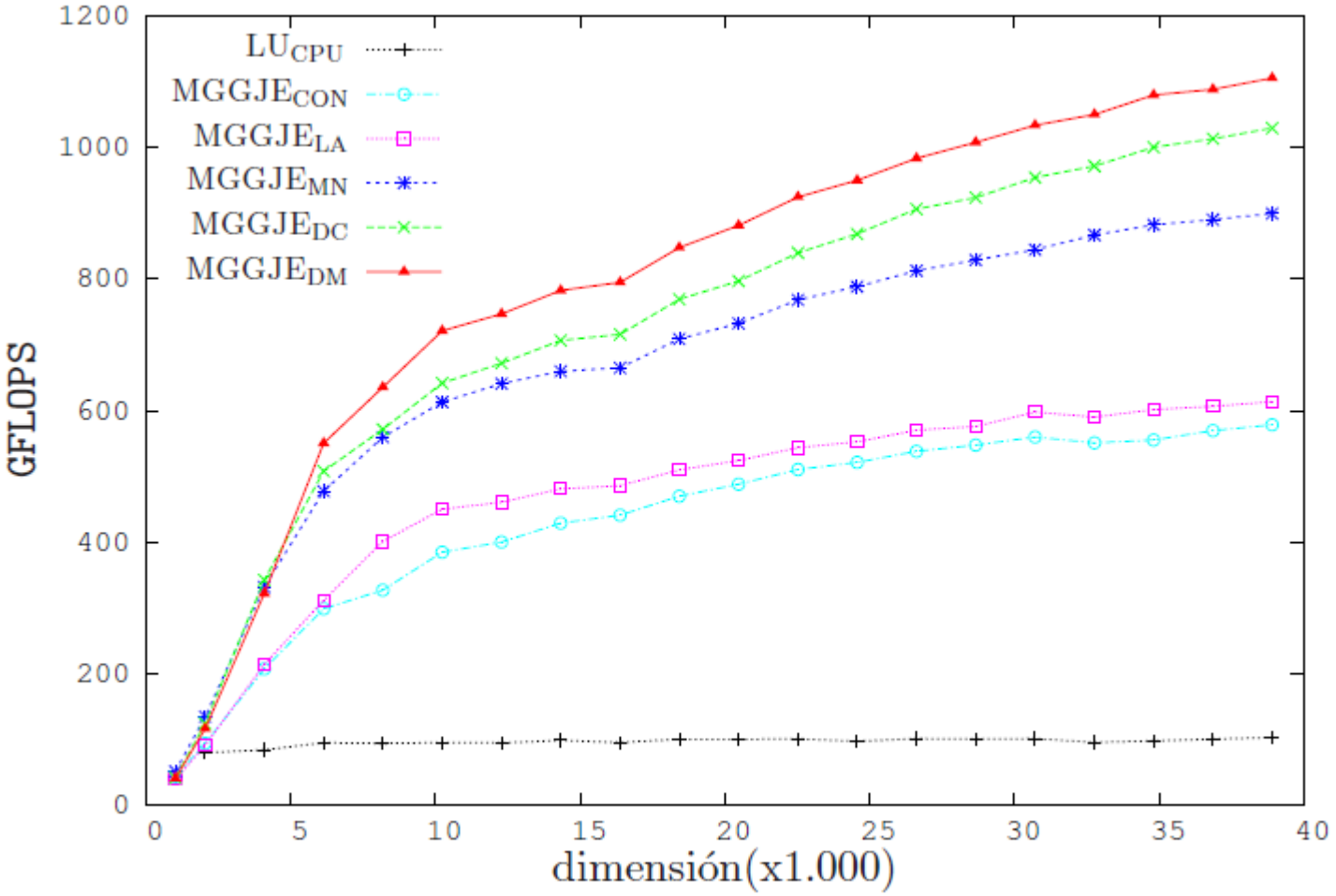
Algunos trabajos en FING

Inversión de matrices generales en múltiples GPUs

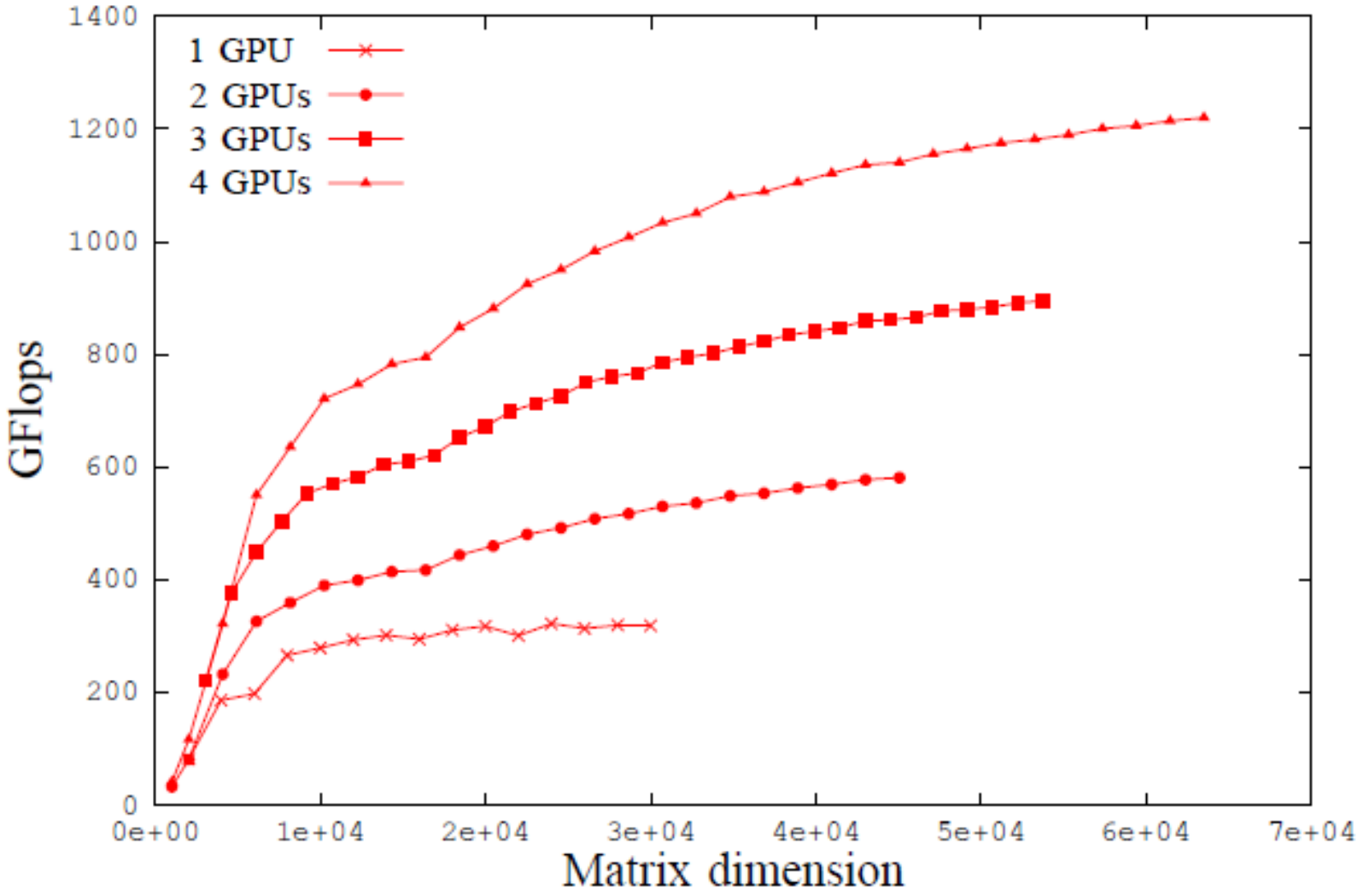
- Implementaciones basadas en Gauss-Jordan.
- Diferentes optimizaciones:
  - Estrategias híbridas
  - Concurrentes
  - Multi-bloque
  - Fusión de operaciones

* High Performance Matrix Inversion on a Multi-core Platform with Several GPUs. P. Ezzatti, E. Quintana, A. Remón, PDP 2011: 87-93.

Algunos trabajos en FING



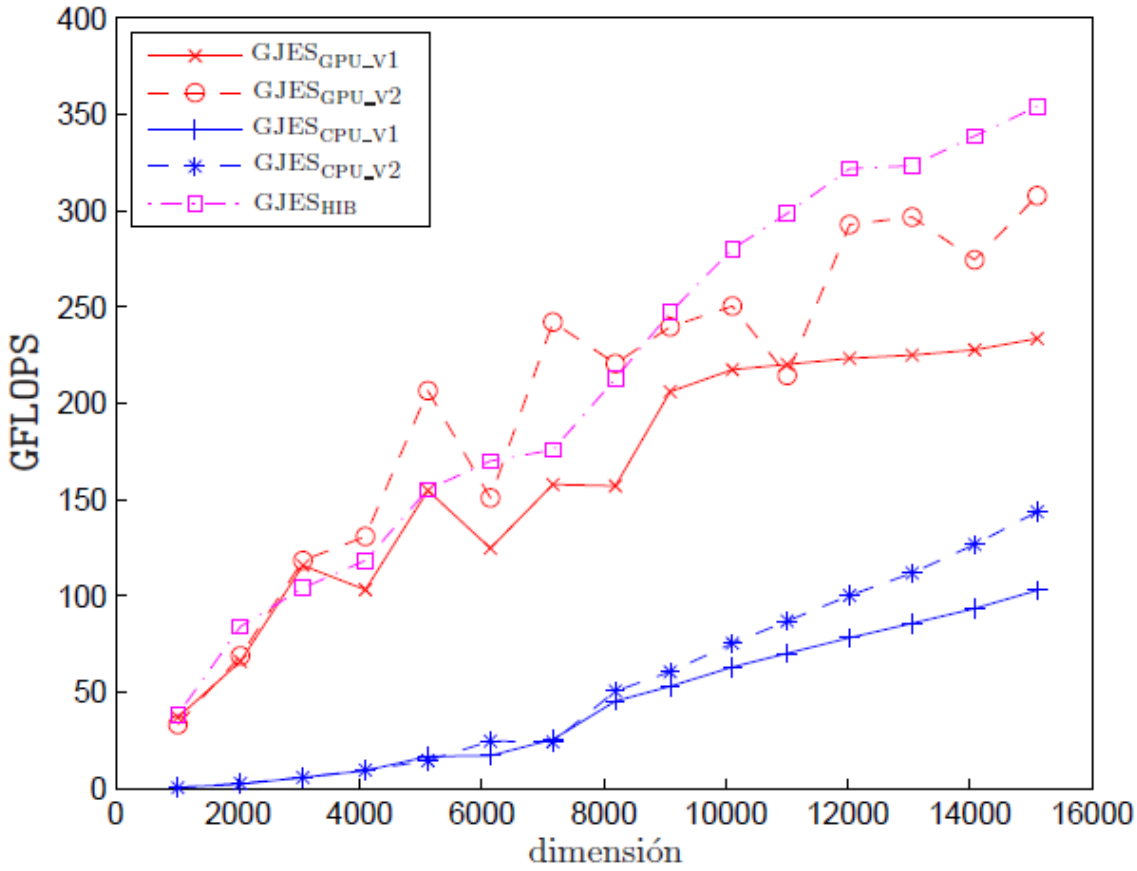
Algunos trabajos en FING



Algunos trabajos en FING

Inversión de matrices SDP

- Estrategias tradicionales con Cholesky.
- Implementaciones basadas en trabajo de [Bientinesi y otros].
- Estrategias a bloque, híbridas, concurrentes.



*** High performance matrix inversion of SPD matrices on graphics processors P. Benner, P. Ezzatti, E. S. Quintana, A. Remón. Workshop on Exploitation of Hardware Accelerators --WEHA 2011, pp. 640-646. Estambul (Turquía). 2011**

Algunos trabajos en FING

Resolución de sistemas triangulares

- **Función implementada por CuBLAS.**
- **Implementación basada en acceso por bloques.**
- **Utilizando herramienta de derivación automática de código (FLAME).**
- **Aceleración de hasta 6x.**

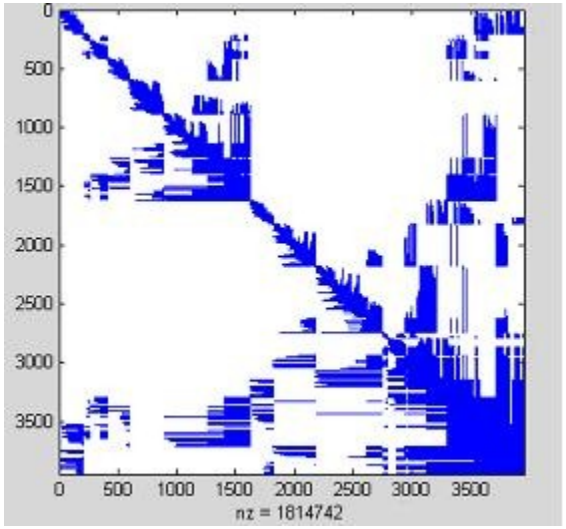
$$\left(X_0 \mid X_1 \mid X_2 \right) \left(\begin{array}{c|c|c} A_{00} & 0 & 0 \\ \hline A_{10} & A_{11} & 0 \\ \hline A_{20} & A_{21} & A_{22} \end{array} \right) = \left(B_0 \mid B_1 \mid B_2 \right);$$

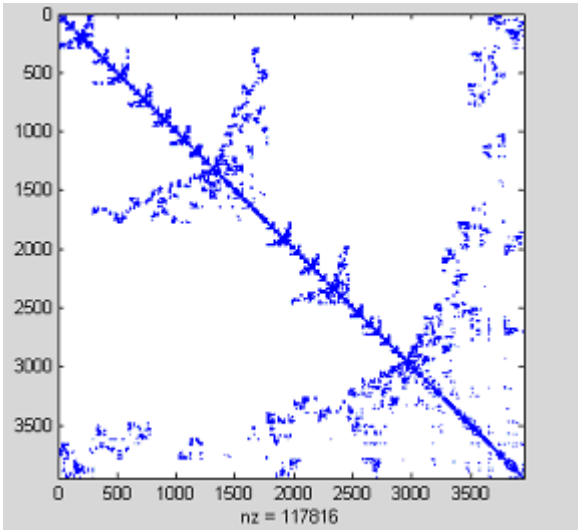
*** Resolución de sistemas triangulares en GPU. P. Ezzatti; E. S. Quintana, A. Remón, XXXI CILAMCE - IX MECOM, 2010.**

Algunos trabajos en FING

Resolución de sistemas dispersos

- A. Zinemanas.
- Métodos directos
  - Métodos multifrontales.
  - Estrategias híbridas (CPU+GPU)
- Matrices simétricas y Def. Pos.



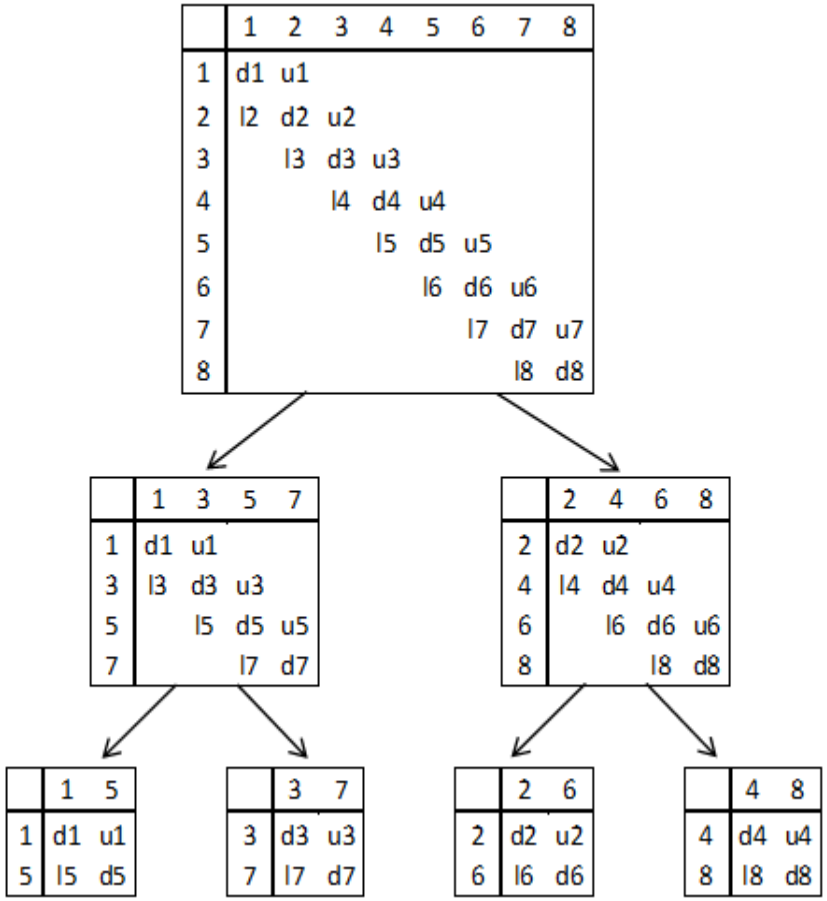


*** Towards a GPU-accelerated direct sparse solver. A. Zinemanas, P. Ezzatti, SIAM Conference on Applied Linear Algebra, 2012.**

Algunos trabajos en FING

Sistemas tridiagonales

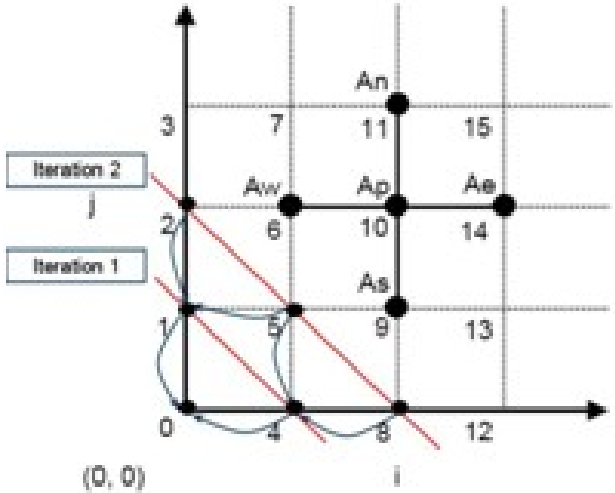
- P. Alfaro y P. Igounet
- Reducción cíclica (y flia)
- Thomas.
- Valores de speedup de 3x.
- Acceso fusionado, memoria compartida, etc.



*** A study on the implementation of tridiagonal systems solvers using a GPU.**
P. Alfaro, P. Igounet, P. Ezzatti, WSDP 2011.

Algunos trabajos en FING

Sistemas penta y hepta diagonales



- P. Igounet y P. Alfaro
- Método SIP en GPU.
  - Acceso fusionado, memoria compartida
  - Distintas transferencias
- Valores de speedup de 4x.

•A GPU implementation of the SIP method. P. Igounet, P. Alfaro, M. Pedemonte, P. Ezzatti, SCCC 2011.

Algunos trabajos en FING

Precisión mixta para el SIP en GPU

- P. Igounet y

E. Dufrechou

Grid	Version	Computation Time	Transfer Time
$64 \times 64 \times 64$	pSIP-GPU$_{simple}$	35.2	6.2
	pSIP-GPU$_{mixed}$	36.3	8.5
$128 \times 128 \times 128$	pSIP-GPU$_{simple}$	176.7	19.6
	pSIP-GPU$_{mixed}$	183.3	34.7

Caso 128 x128x128, error residual:

- en simple 7.6 $x10^{-8}$
- con p. mixta 1.1 x 10^{-11}

*** A Study on Mixed Precision Techniques for a GPU-based SIP Solver. P. Igounet, E. Dufrechou, M. Pedemonte, P. Ezzatti, 3rd Workshop on Applications for Multi-Core Architecture, 2012**

Algunos trabajos en FING

Desempeño y consumo energético

- Juan P. Silva
- Inversión de matrices en diferentes arquitectura
  - ATOM
  - XEON
  - SECO (CARMA)

*** Trading Off Performance for Power-Energy in Dense Linear Algebra Operations. P. Benner, P. Ezzatti, E. Quintana-Ortí , A. Remón, HPCLatam 2013.**

Algunos trabajos en FING

Energía

+ ARM Cortex A9

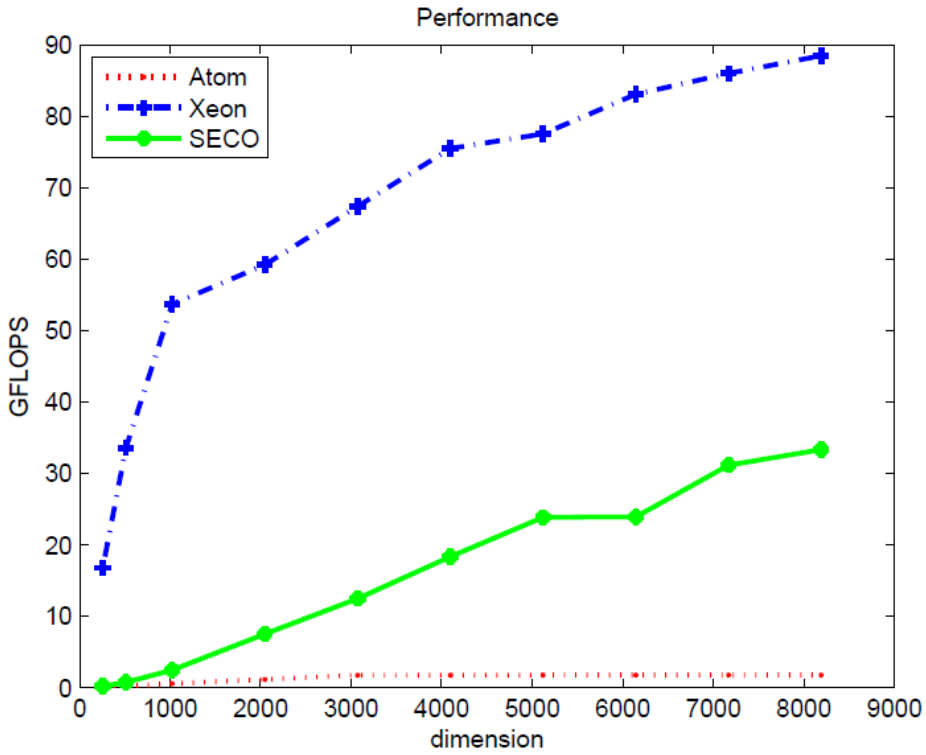
  4 cores 1.3GHz 2GB

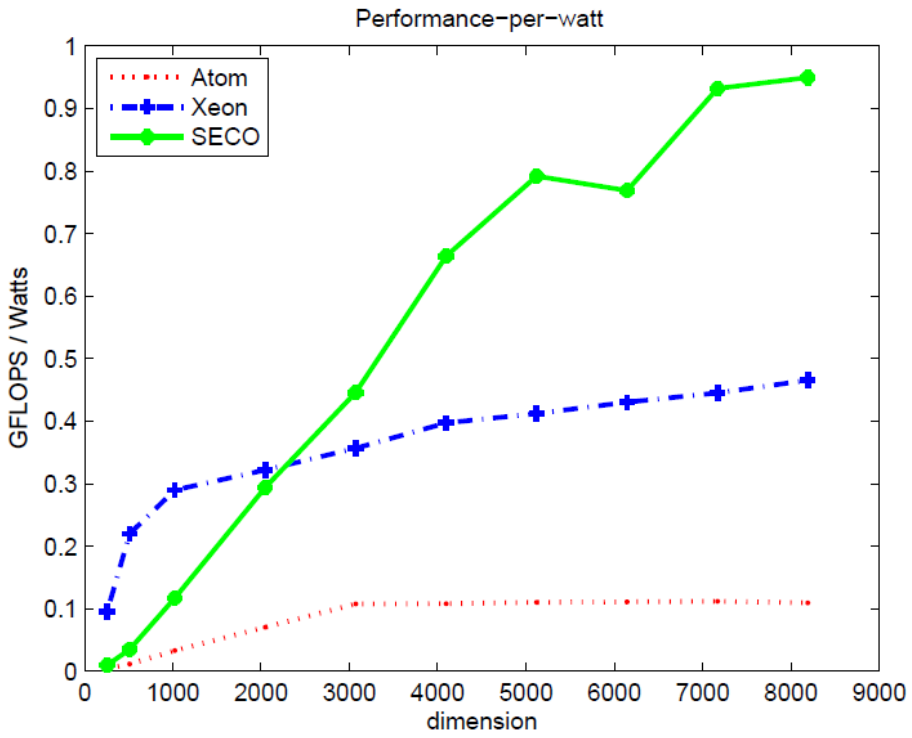
+ NVIDIA Quadro 1000M

  96 cores 1.4GHz 2GB

Algunos trabajos en FING

Energía





Algunos trabajos en FING (actuales)

Algunos trabajos en FING

Álgebra densa:

- Resolución de ecuaciones matriciales.
- Varias GPUs (distribuido)
- Estrategias híbridas.

*** Factorized solution of generalized stable Sylvester equations using many-core GPU accelerators. Benner, P., Dufrechou, E., Ezzatti, P., Gallardo, R., Quintana-Ortí, E.S. Journal of Supercomputing, 2021**

Algunos trabajos en FING

Uso de diferentes dispositivos

- Dispositivos que integran ARMs con GPUs de NVIDIA
- Coral de Google.
- Neural Compute Stick 2 de Intel (Stick 3)
- FPGAs

* Evaluation of architecture-aware optimization techniques for Convolutional Neural Networks. Marichal, R., Toyos, G., Dufrechou, E., Ezzatti, P. PDP 2023: 177-184
* Understanding the performance of elementary NLA kernels in FPGAs Favaro, F., Oliver, J.P., Dufrechou, E., Ezzatti, P. Proceedings - 2020 IEEE 34th International Parallel and Distributed Processing Symposium Workshops, IPDPSW 2020

Algunos trabajos en FING

Álgebra dispersa:

- Modelado de desempeño:
  - En operaciones como SpMv, SpTrsv y SpMM.
- Implementaciones eficientes de las operaciones
  - SpTrsv y SpMM
- Estrategias de compresión de datos.
- Estrategias Sync-free

* Selecting optimal SpMv realizations for GPUs via machine learning. Dufrechou, E., Ezzatti, P., Quintana-Ortí, E.S. International Journal of High Performance Computing Applications, 2021, 35(3), pp. 254–267

Algunos trabajos en FING

Conclusiones

- Implementación eficiente de diversas operaciones de ALN.
- Cuanto mayor es la relación cómputos vs trasferencias mayores las ganancias.
  - Estrategias de precisión mixta pueden ayudar …
- Posibilidad de acelerar la resolución de diversos problemas de computación científica.

Algunos trabajos en FING

Líneas abiertas

- Extender a múltiples GPUs varios enfoques.
- Estudiar otros problemas (p.ej. V&V propios) y sus aplicaciones.
- Evaluar el uso de la nueva arquitectura en profundidad (FPGAs, Tensor Cores, RT Cores, otras GPUs).
- Configuración automática, scheduling dinámico, estudio teórico de desempeño (RLM).
- Avanzar en desarrollos de kernels de álgebra dispersa eficientes en las diferentes plataformas.