Implementación paralela de optimizaciones computacionales del filtro DNLM para la plataforma Xeon Phi

Trabajo Final de Graduación

Manuel Zumbado Corrales

Área Académica de Ingeniería en Computadores
Tecnológico de Costa Rica

11 Junio, 2018

Contenido

Contexto

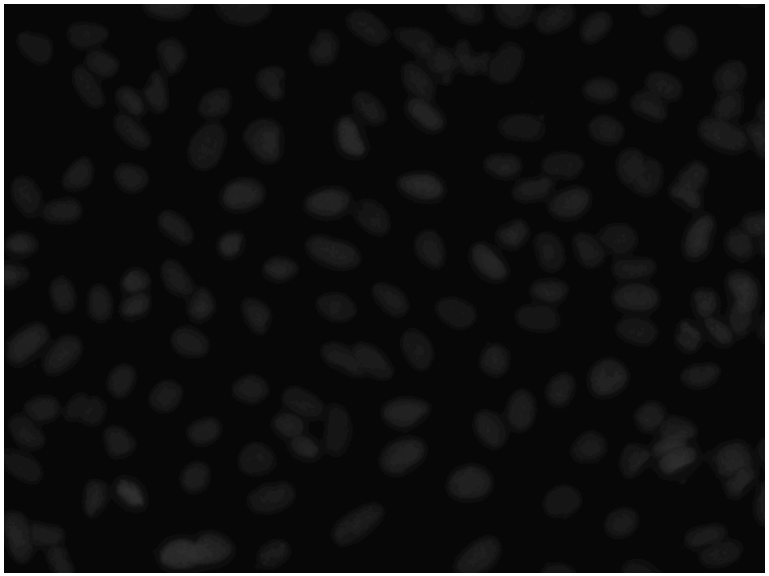
- Proyecto FEES "Análisis funcional genómico de células cancerosas por RNA de interferencia para la identificación de redes de regulación asociadas a proliferación y muerte en respuesta a quimioterapia genotóxica"
- Participan laboratorios de investigación del TEC, UCR y CeNAT

Contexto

1. Imagen de entrada

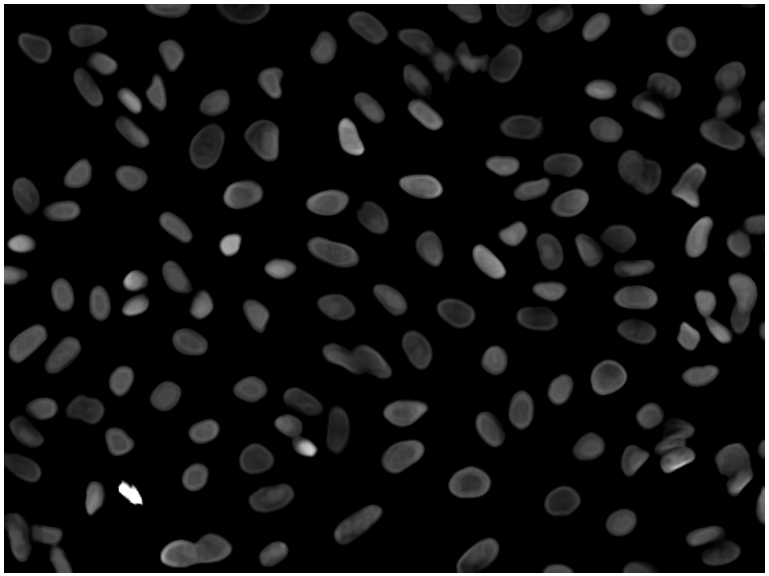
Contexto

1. Imagen de entrada
2. Ecualización de histogramas CLAHE



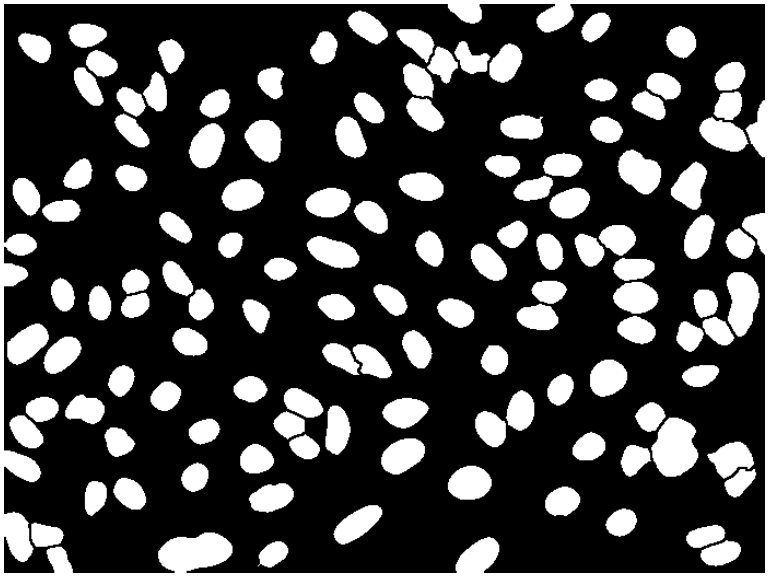
Contexto

1. Imagen de entrada
2. Ecualización de histogramas CLAHE
3. Filtro DNLM

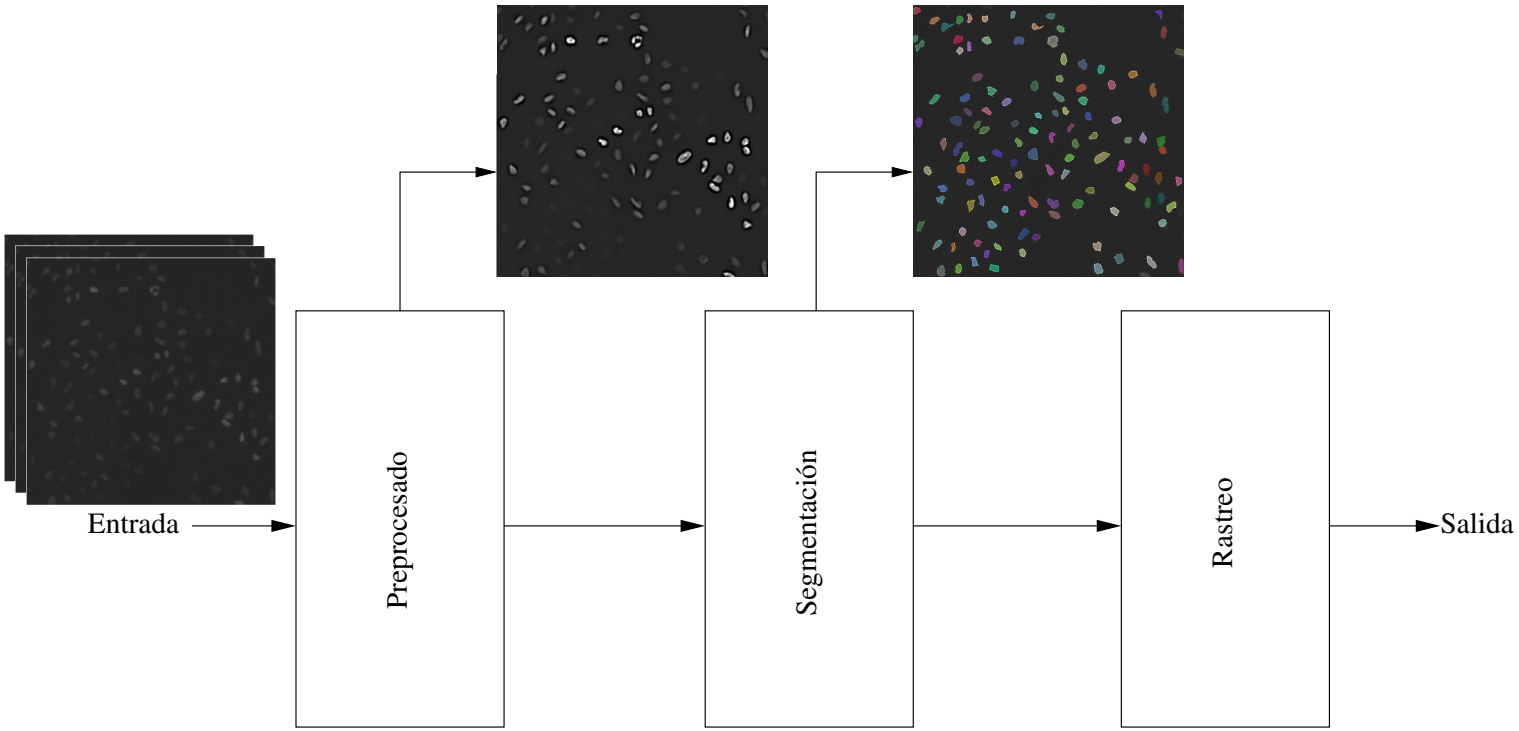


Contexto

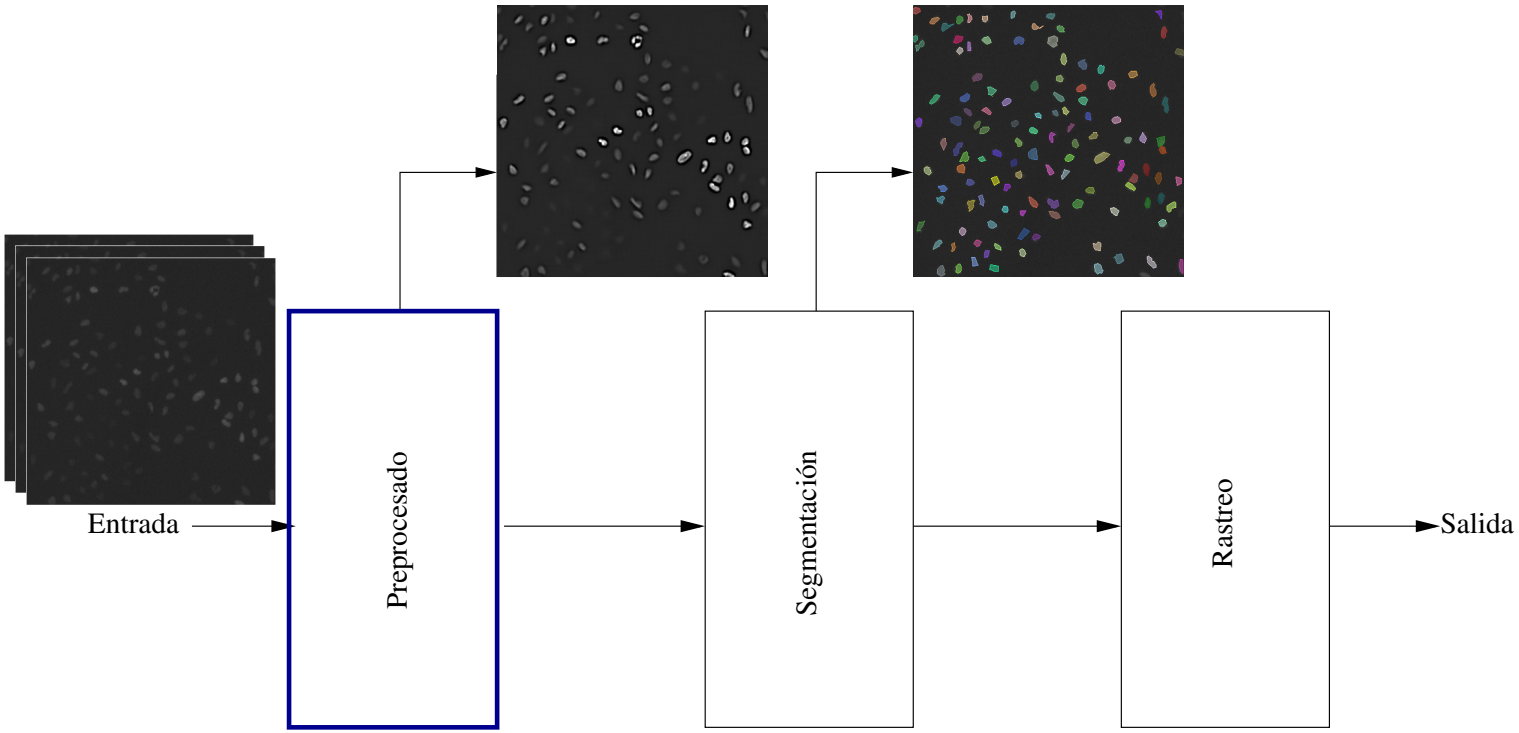
1. Imagen de entrada
2. Ecualización de histogramas CLAHE
3. Filtro DNLM
4. Segmentación



Flujo de procesamiento para videos de actividad celular



Flujo de procesamiento para videos de actividad celular



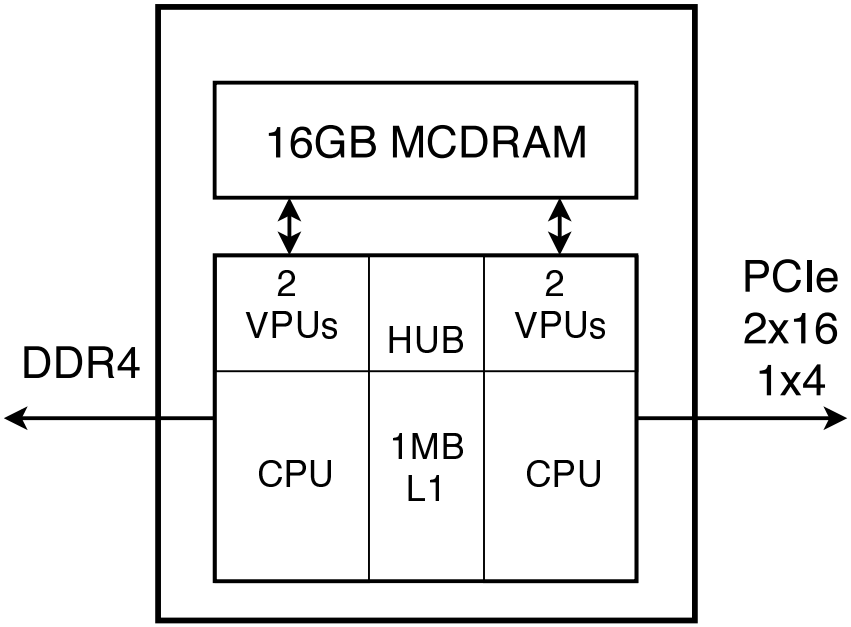
Problema

- Complejidad computacional del filtro DNLM
- Se requieren preprocesar lotes de hasta 170000 imágenes

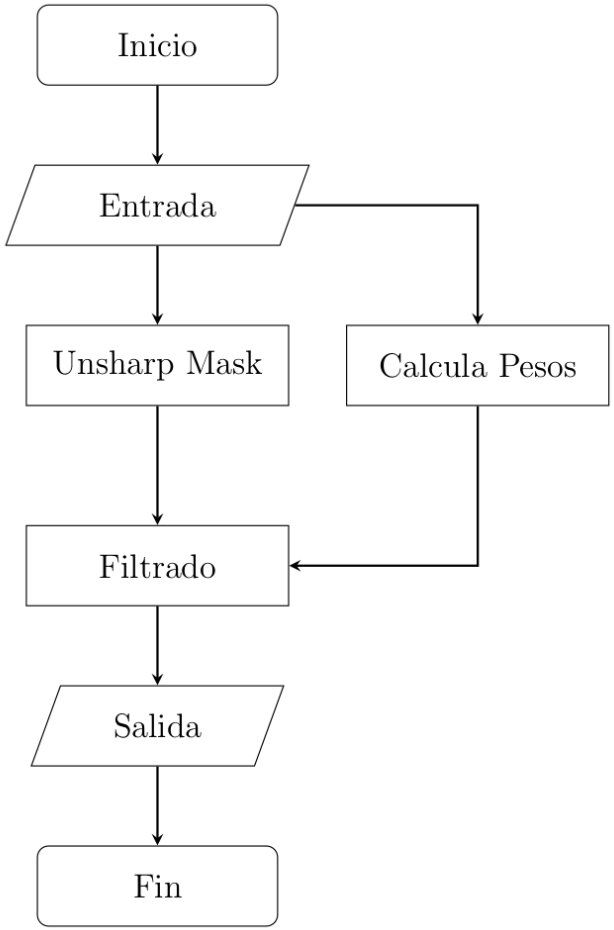
Objetivos

- Proponer al menos una optimización computacional y paralelización del filtro DNLM para la plataforma *Xeon Phi Knights Landing*
- Utilizar paralelismo en dos niveles: a nivel de tareas y a nivel del datos
- Evaluar el rendimiento de las optimizaciones computacionales paralelas

Xeon Phi KNL



Filtro DNLM


Inicio
Entrada
Unsharp Mask
Calcula Pesos
Filtrado
Salida
Fin

Implementación paralela de tres algoritmos:

- DNLM: Filtro DNLM original
- DNLM-IIFFT: Filtro DNLM con Imágenes Integrales y FFT
- DNLM-MA: Filtro DNLM con Media movil y simetría

Complejidad computacional para una imagen de entrada de N pixeles, ventana de búsqueda de S pixeles y tamaño de vecindario de W pixeles:

- DNLM: $\mathcal{O}(N \cdot S \cdot W)$
- DNLM-IIFFT: $\mathcal{O}(N \cdot S \log S)$
- DNLM-MA: $\mathcal{O}(N \cdot S \cdot \sqrt{W})$

Métricas utilizadas

- $\text{CPI}_{\text{hilo}} = \frac{\#\text{ de ticks de reloj}}{\#\text{de instrucciones}}$
- $\text{CPI}_{\text{CPU}} = \frac{\text{CPI}_{\text{hilo}}}{\text{cantidad CPU}}$
- $\text{VPU}_{\text{i}} = \frac{\#\text{ SIMD empaquetadas}}{\#\text{ SIMD empaquetadas}+\#\text{ SIMD escalares}}$

Métricas utilizadas

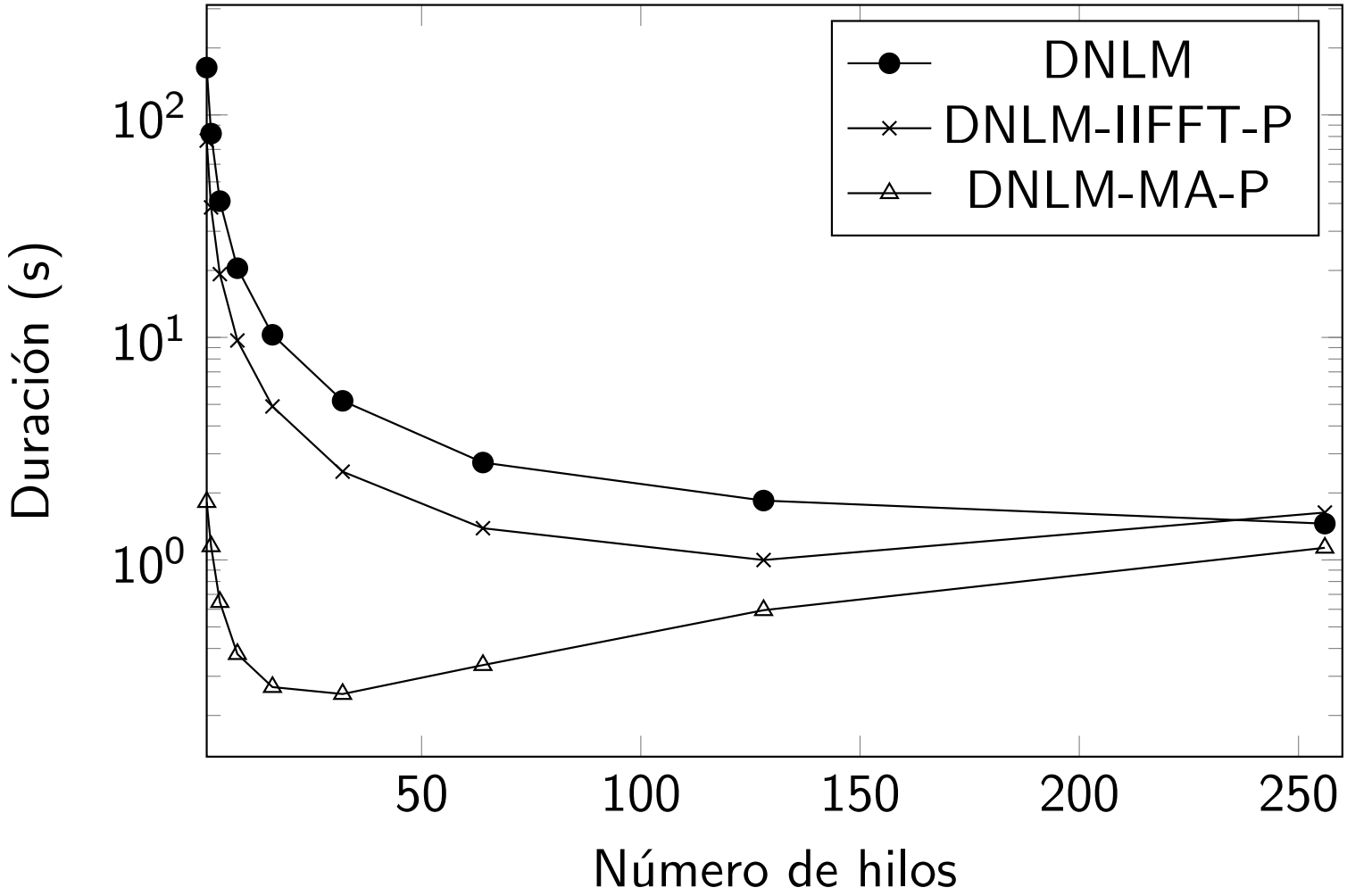
- $\mu\text{Ops}_\text{i} = \frac{\#\mu\text{Ops retiradas MS}}{\#\mu\text{Ops retiradas totales}}$
- $\text{L1}_\text{Hit} = \frac{\#\text{ operaciones de lectura} - \#\text{ fallos de lectura en L1}}{\#\text{ operaciones de lectura}}$
- $\text{L2}_\text{Hit} = \frac{\#\text{ operaciones de lectura} - \#\text{ fallos de lectura en L2}}{\#\text{ operaciones de lectura}}$

Experimento

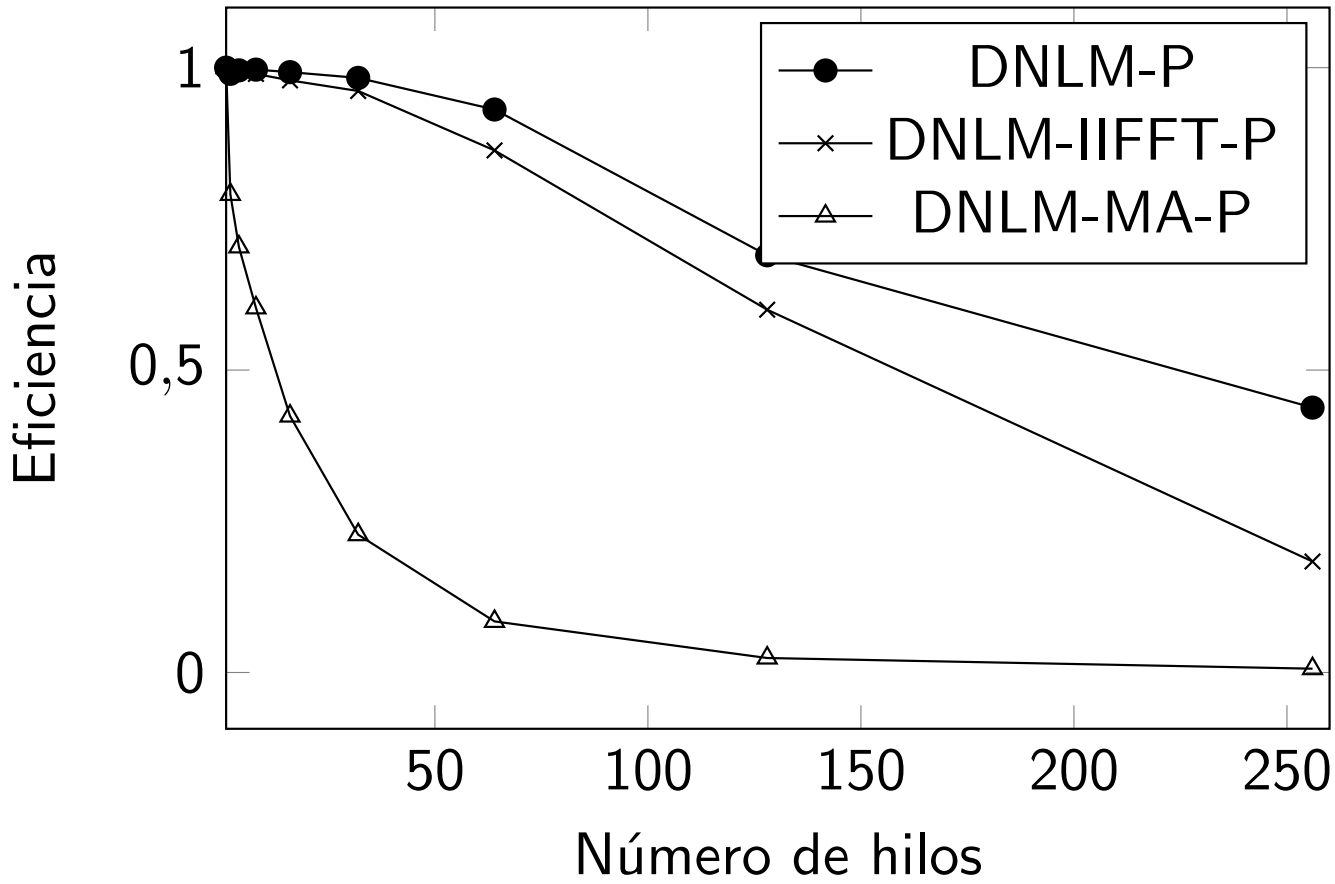
Configuración:

- Promedio de 10 ejecuciones por prueba
- Imagen de entrada de $N = 1024 \times 1024$ pixeles, con $S = 21 \times 21$ y $W = 7 \times 7$
- Se escala la cantidad de hilos de 1 a 256

Escalabilidad de optimizaciones del filtro DNLM



Eficiencia de las implementaciones paralelas del filtro DNLM



Aceleración alcanzada con las optimizaciones del filtro DNLM

Aceleración promedio de optimizaciones del filtro DNLM

Filtro	Duración [s]	Coef. de var.	Aceleración [×]
DNLM	$163,71 \pm 1,32$	$0,01$	1
DNLM-IIFFT	$76,68 \pm 0,06$	$0,00$	$2,14$
DNLM-MA	$1,81 \pm 0,01$	$0,00$	$90,44$
DNLM-P	$1,5 \pm 0,10$	$0,06$	$108,99$
DNLM-IIFFT-P	$1,00 \pm 0,00$	$0,00$	$163,92$
DNLM-MA-P	$\mathbf{0,25 \pm 0,00}$	$\mathbf{0,01}$	$\mathbf{665,63}$

Intensidad de uso de VPU

Intensidad vectorial de operaciones con un hilo de ejecución

Filtro	VPU_i
DNLM-P	1,00
DNLM-IIFFT-P	0,81
DNLM-MA-P	0,95

Ciclos por instrucción por núcleo

Cambio en CPI_{CPU} al escalar el número de hilos

Filtro	1	2	4	8	16	32	64	128	256
DNLM-P	1,87	0,94	0,47	0,23	0,12	0,06	0,03	0,02	0,02
DNLM-IIFFT-P	1,36	0,68	0,34	0,17	0,09	0,04	0,02	0,02	0,02
DNLM-MA-P	1,74	1,03	0,54	0,27	0,14	0,07	0,04	0,02	0,02

Intensidad de operaciones de microarquitectura

Cambio en intensidad de micro-operaciones al escalar el número de hilos

Filtro	1	2	4	8	16	32	64	128	256
DNLM-P	0,26	0,26	0,26	0,26	0,26	0,27	0,27	0,27	0,28
DNLM-IIFFT-P	0,05	0,05	0,05	0,05	0,06	0,07	0,09	0,11	0,21
DNLM-MA-P	0,03	0,06	0,10	0,16	0,23	0,28	0,32	0,32	0,33

Razón de aciertos de lectura en caché L1

Cambio en la razón de aciertos de lectura en caché L1 al escalar el número de hilos

Filtro	1	2	4	8	16	32	64	128	256
DNLM-P	0,99	0,99	1,00	1,00	1,00	1,00	1,00	0,99	0,99
DNLM-IIFFT-P	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99	0,99
DNLM-MA-P	0,80	0,83	0,85	0,88	0,93	0,96	0,98	0,99	0,99

Razón de aciertos de lectura en caché L2

Cambio en la razón de aciertos de lectura en caché L2 al escalar el número de hilos

Filtro	1	2	4	8	16	32	64	128	256
DNLM-P	0,99	0,99	0,97	0,96	0,96	0,95	0,94	0,97	0,95
DNLM-IIFFT-P	1,00	1,00	0,99	0,99	0,99	0,99	0,99	0,98	0,93
DNLM-MA-P	0,91	0,86	0,85	0,85	0,85	0,85	0,85	0,88	0,89

Conclusiones

- Se desarrollaron dos optimizaciones computacionales del filtro DNLM y sus paralelizaciones
- Se alcanzó un grado de vectorización de 81 % para IIFFT, 95 % para MA y 100 % para el filtro original
- La escalabilidad de los algoritmos implementados se ve afectada por el ancho de banda de memoria y aciertos en caché L2
- La mayor aceleración alcanzada (1663×) supera la mejor aceleración reportada en el estado del arte (740×)

Conclusiones

- Procesamiento de un conjunto de 170000 imágenes en unas 11 horas
- Efectiva vectorización es la clave en el escalamiento
- Eficiencia en tiempo de procesamiento permitiría su uso en CNNs
- Artículos relacionados publicados en CIARP 2017 y próximamente ICIP 2018

Trabajo futuro

- Explorar paralelización multi nodo
- Añadir *Unsharp Masking* adaptativo
- Calibración automática de parámetros

Resumen