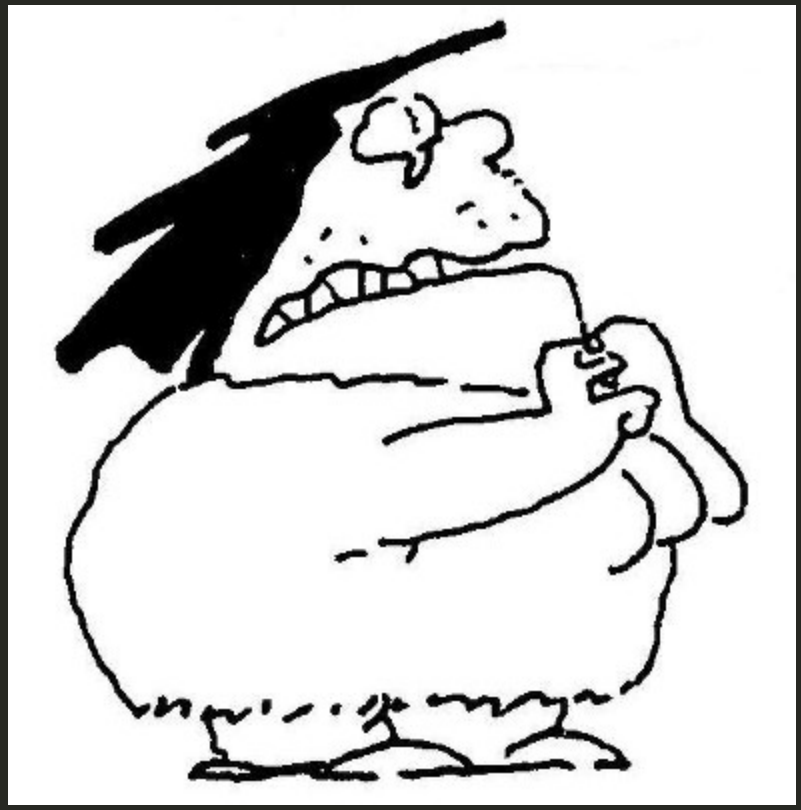
Conociendo a Eulogia

Carlos Bederián, CCAD-UNC

carlos.bederian@unc.edu.ar



Cómo llegamos hasta acá

Cristina

- Propósito general
- Retirada por eficiencia energética

Cómo llegamos hasta acá

Mendieta

- Cluster con GPUs
  - Alta performance/$
  - Alta performance/W
  - GPGPU no siempre es aplicable

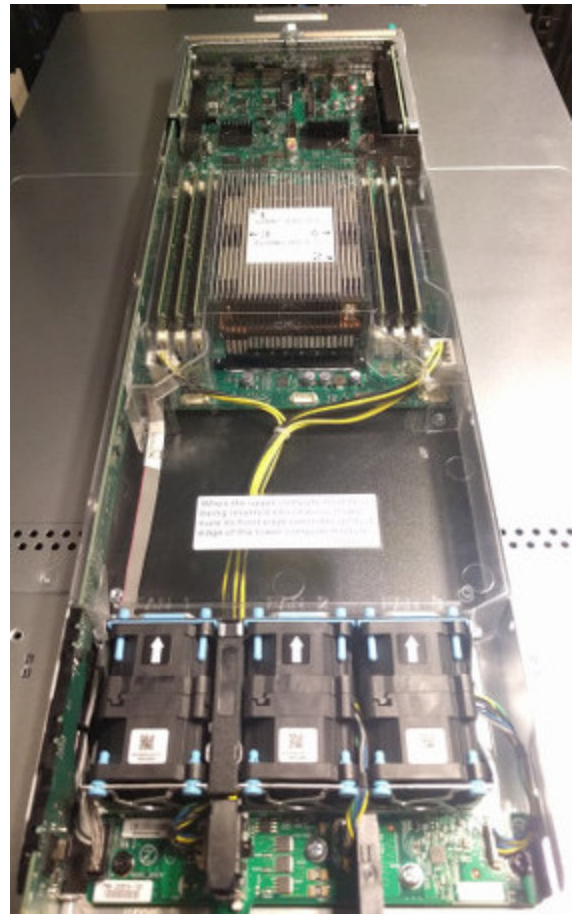
Funding

1. ~300K ARS de UNC para retirar Cristina
2. ~600K ARS subsidio SNCAD y contraparte UNC
3. ~400K ARS PMT FAMAF y recursos corrientes de CCAD (en curso)

Qué armamos con eso

Nodo Eulogia

- Intel Xeon Phi 7210
- 96GB de memoria en 6 canales DDR4-2133
- Disco de estado sólido de 240GB



Larrabee

- Proyecto de Intel para crear una GPU en software
  - Cores x86
  - Memoria coherente
- Nunca vio la luz del día como GPU
- Primer producto: Knights Corner (Xeon Phi x100)
  - Acelerador PCI Express
  - 57-61 cores basados en P54C
  - Unidades de vectores IMCI de 512 bits
  - SMT para ocultar latencia
  - Memoria GDDR5

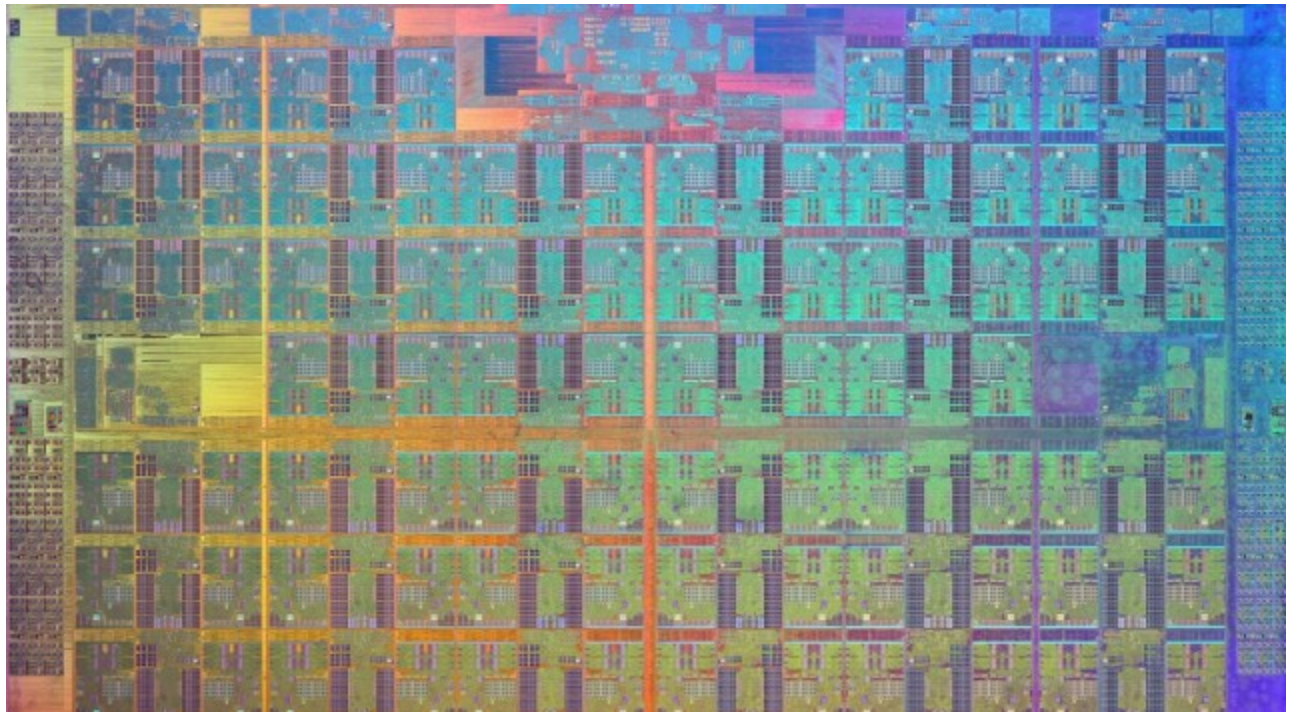
Knights Landing (Xeon Phi x200)

Solución a la mayoría de los problemas de Knights Corner

- Procesador principal, no acelerador
- Soporte X86-64 completo
  - AVX-512, evolución de IMCI que es estándar en adelante
- Memoria expandible

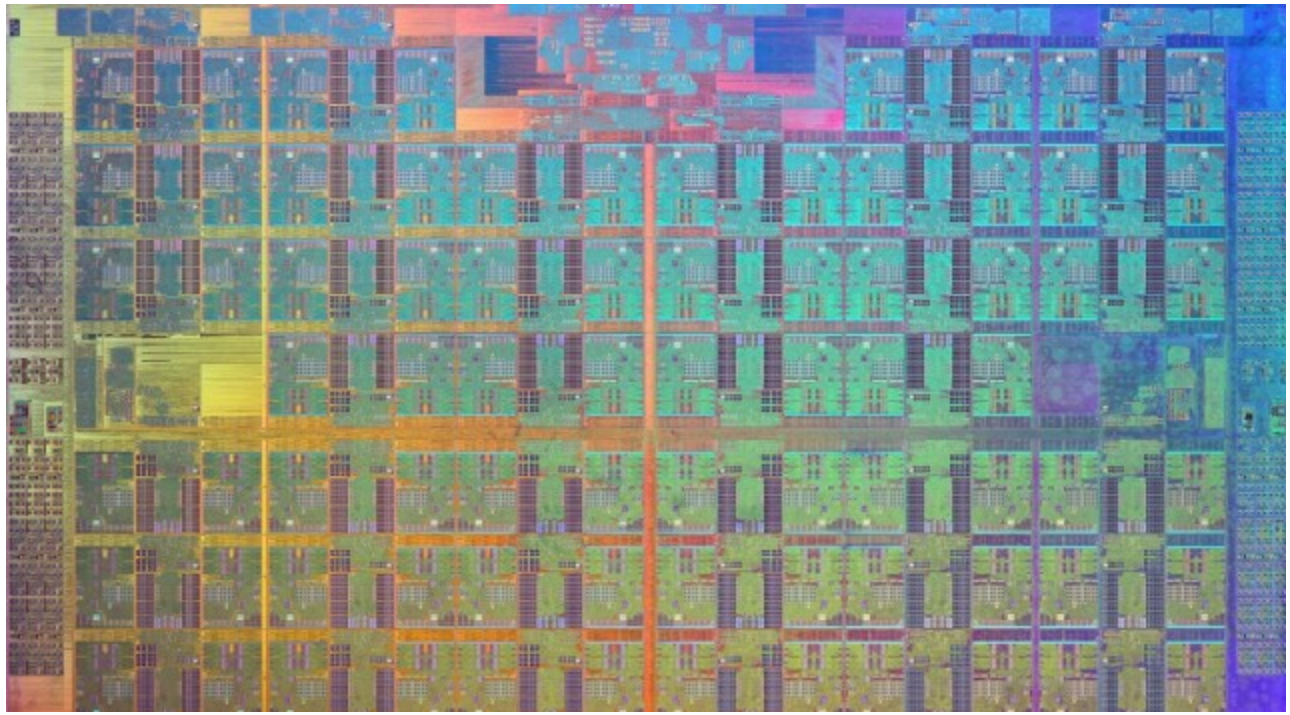
Knights Landing es...lento

- 64 a 72 cores Silvermont (Atom)
  - 1.3 a 1.5 GHz



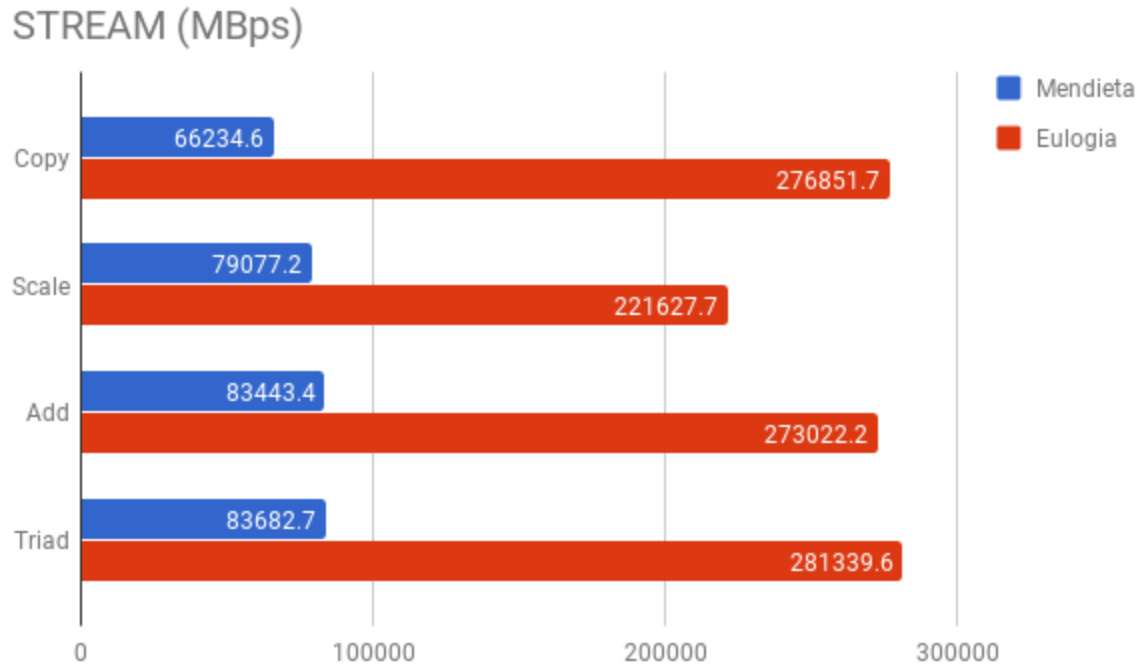
Knights Landing es trabajador

- **Dos** unidades de vectores AVX-512 de 512 bits
- 2.6 a 3.5 TFLOPS FP64



Memoria

- 6 canales de memoria DDR4: ~90GBps
- **16 GB** de memoria HMC integrada: 400-480GBps (pero mayor latencia)
  - Uso como cache o memoria adicional



Comparativa: HPL

- Nodo Mendieta con NVIDIA K20X, ~600W

```
================================================================================
T/V                N    NB     P     Q               Time                 Gflops
--------------------------------------------------------------------------------
WR10L2L2       85000  1536     1     2             292.08              1.402e+03
--------------------------------------------------------------------------------
||Ax-b||_oo/(eps*(||A||_oo*||x||_oo+||b||_oo)*N)=        0.0035472 ...... PASSED
================================================================================
```

- Nodo Eulogia, ~350W

```
================================================================================
T/V                N    NB     P     Q               Time                 Gflops
--------------------------------------------------------------------------------
WR00C2R2      102000   336     8     8             524.66              1.348e+03
--------------------------------------------------------------------------------
||Ax-b||_oo/(eps*(||A||_oo*||x||_oo+||b||_oo)*N)=        0.0012399 ...... PASSED
================================================================================
```

Comparativa: QE

Benchmark AUSURF112

- Nodo Mendieta, QE 5.4.0, 20 procesos, sin GPUs, ~300W

```
PWSCF        : 56m15.08s CPU    57m11.79s WALL
```

- Nodo Eulogia, QE 6.2.1, 64 procesos, ~350W

```
PWSCF        : 20m 2.92s CPU    20m46.55s WALL
```

Caveat emptor

- Resultados preliminares
- Hyperthreading empeora los resultados en ambos casos
  - Quantum Espresso y HPL hacen uso intensivo de MKL
  - No es el caso para todos los programas

Estado actual del cluster

- 11 nodos de cómputo, 1 actuando como cabecera
  - Adquisición de nodo cabecera en curso
- Storage compartido con el resto de los equipos del CCAD
  - Adquisición de dos servidores de storage en curso
- Instalación Infiniband en curso
  - Trabajos multi nodo
  - Mejoras en acceso a almacenamiento
  - Gracias a Intel y McAfee por sus donaciones

Aplicaciones

Software soportado actualmente

- Intel Parallel Studio XE 2018 (`icc` e `ifort`)
- Intel MKL
- OpenMPI 3.x
- Quantum Espresso

Toolchains en consideración

- GCC 7.x (8.x?)
- LLVM: clang y flang

Más aplicaciones

- Solicitarlas a soporte
- Mejor tiempo de respuesta con la llegada de nueva cabecera

Flags de compiladores

- Es imprescindible activar AVX-512 para obtener buena performance
- La nueva cabecera no soporta el mismo subconjunto de AVX-512 que los nodos
  - Intel: `-xCOMMON-AVX512 -axMIC-AVX512` en vez de `-xHost`
  - GCC: `-march=knl -mno-avx512pf -mno-avx512er` en vez de `--march=native`

Ante la duda revisar los CFLAGS del módulo del compilador

Lanzamiento de trabajos

Hay 64 cores que ocupar, con hasta 4 hilos cada uno

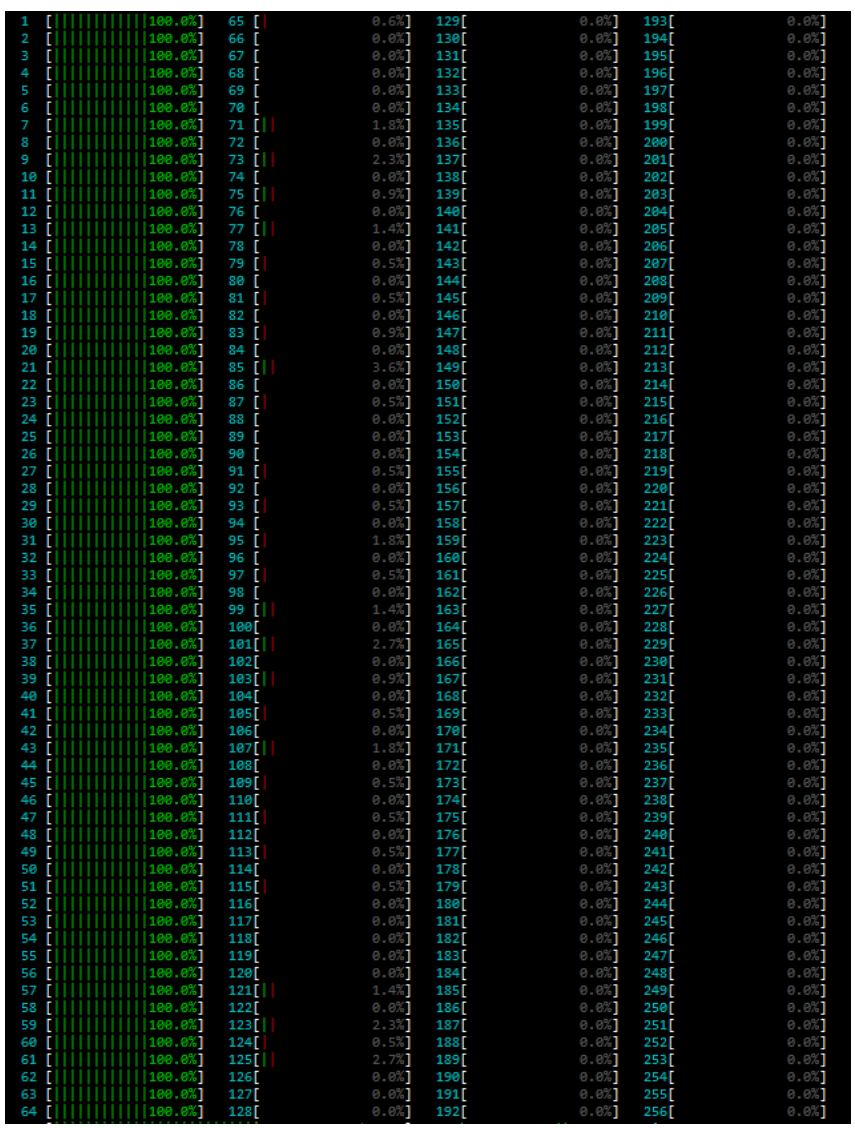
- `--ntasks` * `cpus-per-task` deberían ser 64, 128, 192 o 256
- La configuración óptima depende del programa y el problema

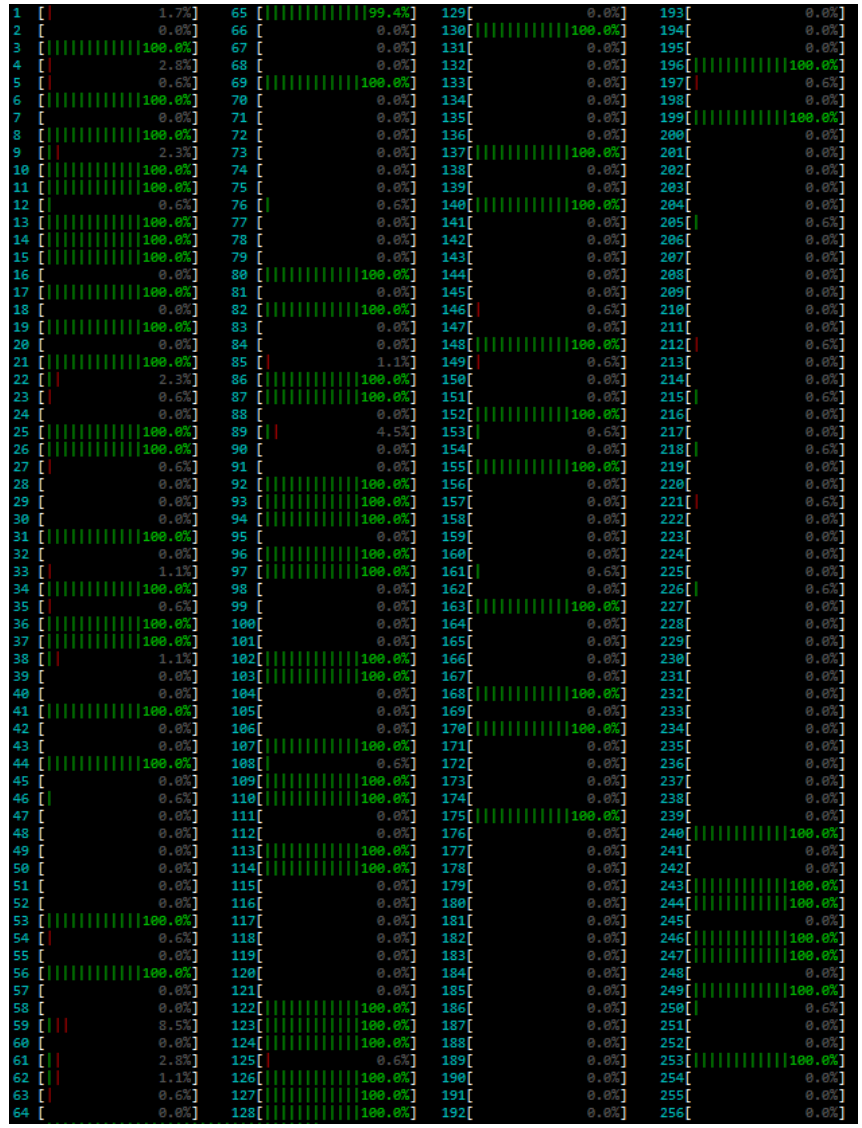
Recordatorio:

- `#SBATCH --ntasks`: Cantidad de procesos (e.g. MPI) a lanzar
- `#SBATCH --cpus-per-task`: Cantidad de hilos a reservar por proceso
  - Nota: A los hilos los lanza el programa (e.g. `OMP_NUM_THREADS`), SLURM sólo los reserva

Numeracion de cores

- Los procesadores `i,i+64,i+128,i+192` de Linux corresponden al mismo core
  - Ojo con la distribución de procesos e hilos!
- Cada fila de htop corresponde a los 4 hilos de un mismo core





Q&A