CLUSTERS BEOWULF EN MECANICA COMPUTACIONAL

por A. Yommi, N. Nigro, M. Storti and V. Sonzogni

Centro Internacional de Métodos Numéricos

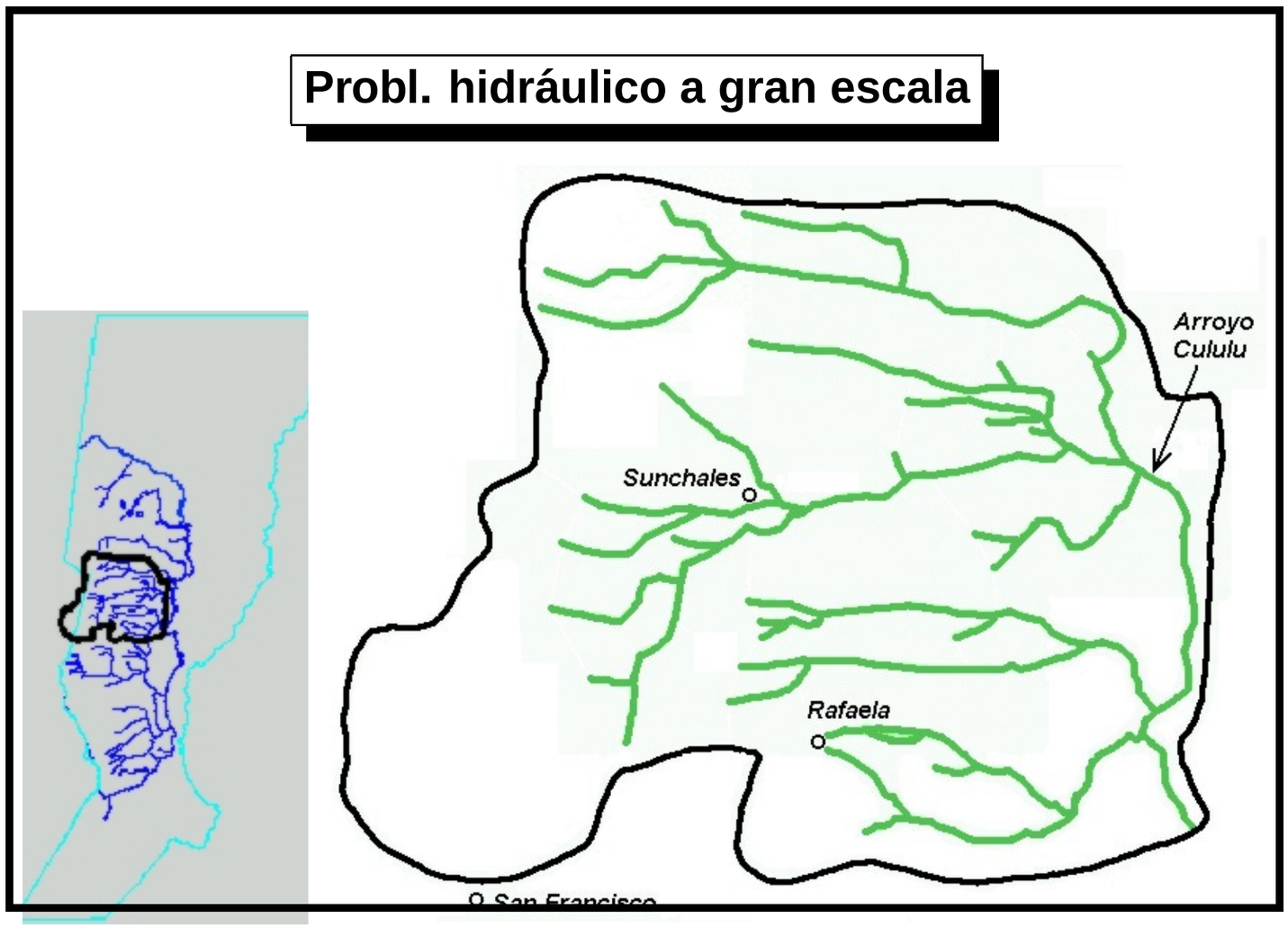
en Ingeniería - CIMEC

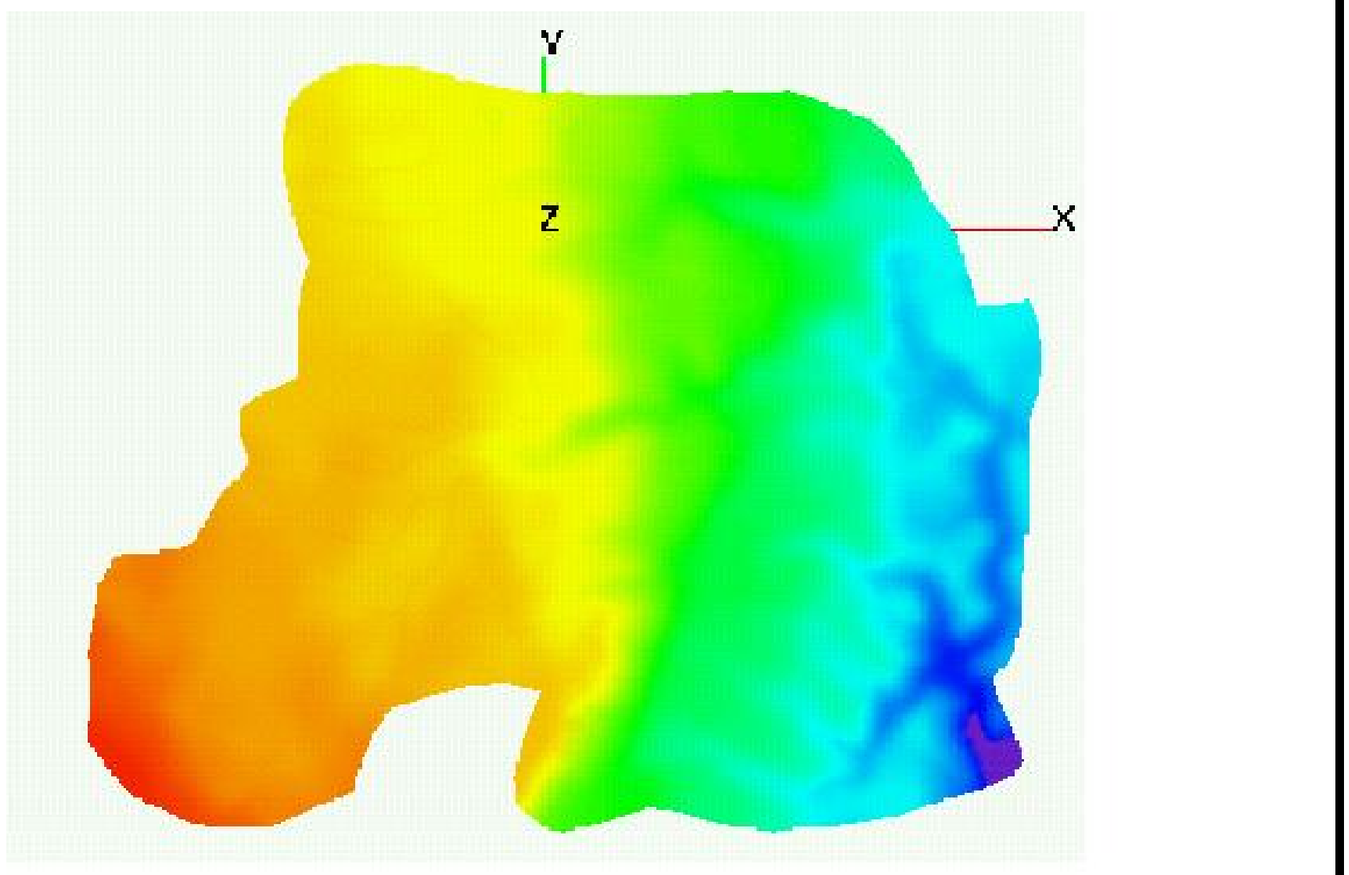
INTEC, (CONICET-UNL), Santa Fe, Argentina

<mstorti@intec.unl.edu.ar>

http://www.cimec.com.ar

2 de octubre de 2002

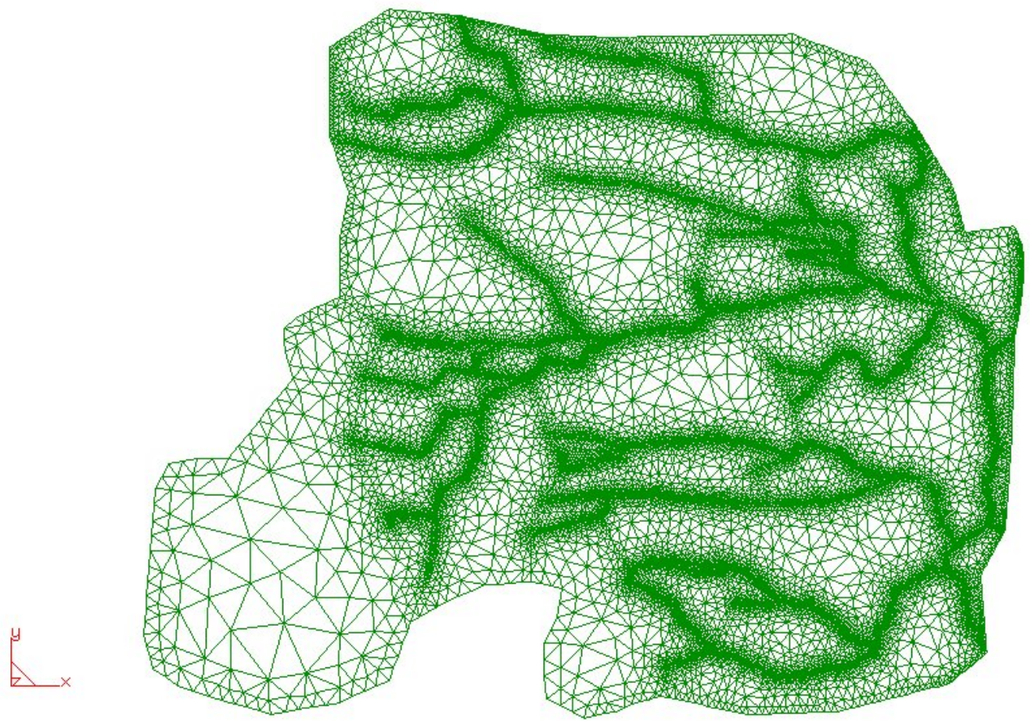

Probl. hidráulico a gran escala
Arroyo
Cululu
Sunchales
Rafaela
San Francisco


Y
Z
X

Probl. hidráulico a gran escala (cont..)

- **La diversidad de escalas de longitud y temporales lleva a problemas computacionales de gran magnitud. Tipicamente: escalas de longitud que van de 100km a 100mts en el río.**
- **Refinando localmente (FEM): mallas de entre 100,000 y 1,000,000 elementos triangulares.**
- **Problema multiacífero con varias capas por acuífero debe pensarse en un número d incǵnitas de 10-20 por nodo.**
- **De esta forma se llega fácilmente a un número de incǵnitas de entre $10^6 - 10^7$.**
- **Proyecto ANPCyT FLAGS PID 99/74 “Simulación numérica en gran escala de la interrelación entre el FLujo de Aguas Superficiales y el FLujo de AGuas Subterráneas”. Dir: Dr. Sergio Idelsohn.**

Probl. hidráulico a gran escala (cont..)



Probl. hidráulico a gran escala (cont..)

- **FEM: a) evaluación del residuo y matrices, b) resolución del problema lineal asociado. (para cada paso de tiempo)**
- **Evaluación: 0.1 sec/Kelem/capa (un acuífero procesador Pentium4).**
- **Para un número de elementos de 10^6 y 10 capas esto da un tiempo de evaluación de 1000secs$\approx$ 17 m.**
- **Este tiempo debe ser multiplicado por el número de pasos de tiempo y, si el problema es no-lineal por el número de iteraciones del lazo de Newton-Raphson. El problema de flujo en medios porosos es no-lineal por el acuífero freático y por el acoplamiento con corrientes de superficie.**
- **Si consideramos un número de pasos de tiempo en el orden de 1000 legamos a tiempos de cálculo de varios días.**

Cohete con tanque propulsante líquido.

- **Cohete de uso académico IUA. Instituto Universitario Aeronáutico Córdoba, Instituto Balseiro, financiado por CONAE, Comisión Nacional de Investigaciones Espaciales.**
- **Combustible y oxidante líquidos (anilina/acido nitríco) y estabilización por rotación.**
- **Líquidos confinados tienden a tener un efecto desestabilizante sobre la trayectoria.**
- **Velocidades de rotación O(300 rpm). Aceleraciones lineales O(15 G). Dos contenedores cilíndricos concéntricos de D=15cm, H=75cm.**
- **Predecir las fuerzas del fluido sobre el contenedor en función de la trayectoria.**

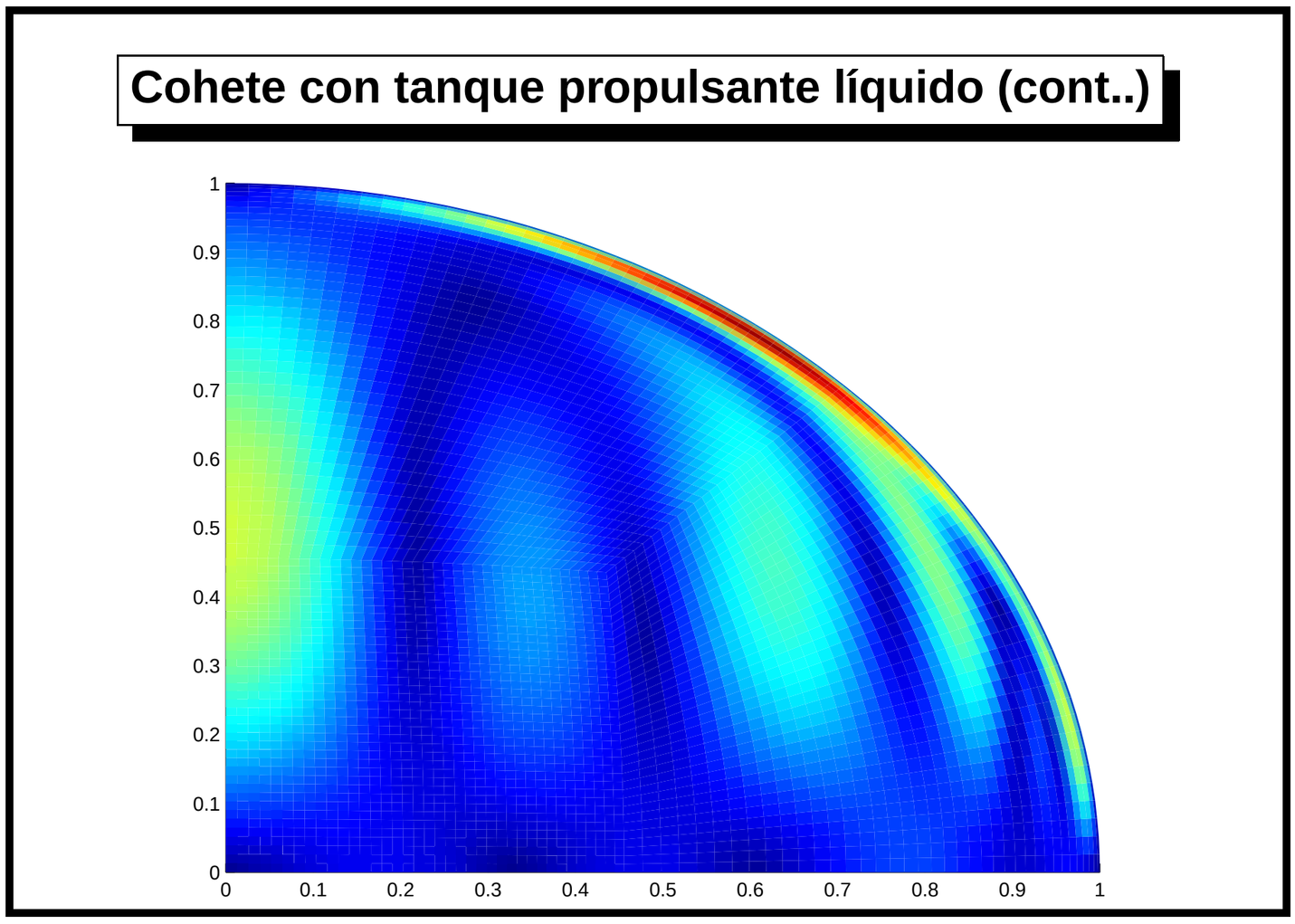


Figura 1: Ondas inerciales en una esfera rotante

Cohete con tanque propulsante líquido (cont..)

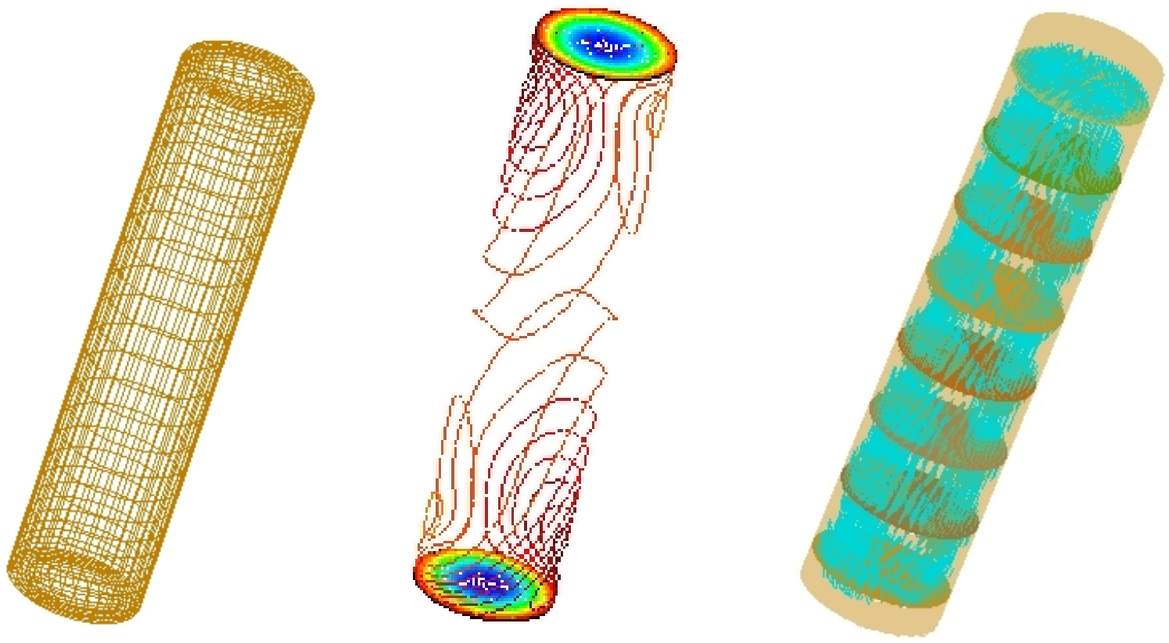


Figura 2: Ondas inerciales en una esfera rotante

Técnicas de computación de alta-performance (HPC)

- **Procesamiento Distribuido: dividir el problema en subproblemas más pequeños y resolver cada uno de los problemas en un procesador por separado.**
- **Para problemas desacoplados tiempo de cálculo en n procesadores es $T_n = T_1/n$.**
- **Problemas no desacoplados: $T_n = T_1/n + T_{\text{comm}} + T_{\text{synchro}}$.**
- **El factor de ganancia al paralelizar se llama *"factor de aceleración"* (*"speedup"*) y se define como $S_n = T_1/T_n$. Completamente desacoplado entonces $S_n = n$.**
- ***"eficiencia"* de la paralelización $\eta = S_n/n < 1$.**
- ***"Top 500"* `http://www.netlib.org`. (Todas usan procesamiento distribuido).**

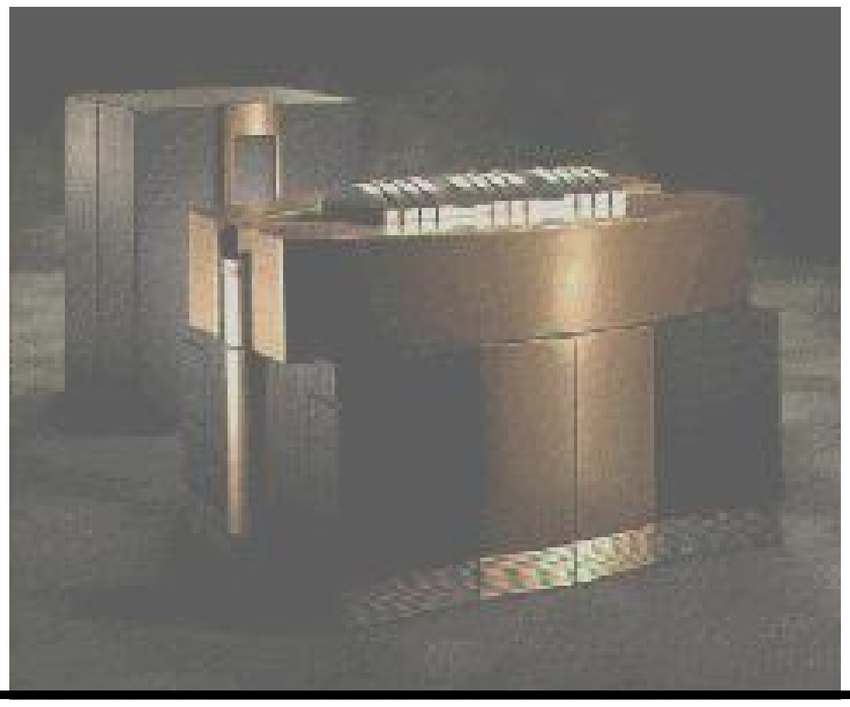
Hardware para procesamiento distribuido

- **En el extremo “superior” estan las grandes firmas que venden supercomputadoras como Cray (hoy SGI) o IBM Asci-Red con miles de procesadores. Hoy las más comunes son las SGI Origin 2000.**
- **NOW/COW: (Network/Cluster of workstations) A partir de los 80’s era común encontrar en los laboratorios de las universidades un cierto número de estaciones de trabajo (SUN/HP/DEC/SGI) y surgió la motivación de utilizar esta potencia de cálculo en corridas nocturnas. Surgieron las bibliotecas de *“Paso de Mensajes”* como PVM y MPI.**

Cray T94 en CESUP UFRGS, Porto Alegre Brasil

2 Procs. $\times$ 1.8 Gflops, Costo aprox. $ 5,000,000.

`http://www.cesup.yfrgs.br`



Clementina II, SECTIP

- **Silicon Graphics Origin 2000, 40 procs.**
- **Costo aprox. $ 3,000,000**
- **`http://www.setcip.org.ar`**

Clusters Beowulf

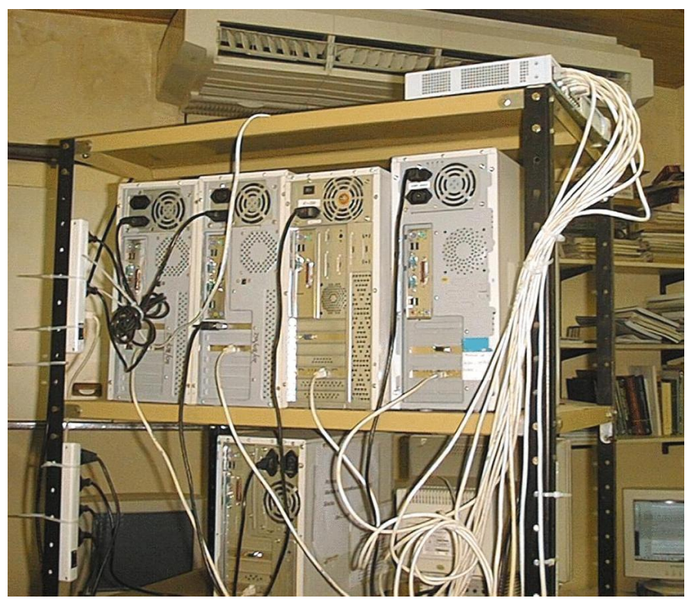
- **Con el abaratamiento de las PC's y el advenimiento de software libre surgió la posibilidad de crear clusters de PC's completamente dedicados a cálculo. Estos son los llamados clusters Beowulf.**
- **De *"How to build a Beowulf"* (Sterling, T.L. et.al., MIT Press, 1999) un *"cluster Beowulf"* es *" Un cluster the 'mass-market commodity off-the-shelf' ($\mathrm{M}^2\mathrm{COTS}$) PC's interconectadas por tecnología LAN de bajo costo corriendo un OS open source de tipo Unix y ejecutando aplicaciones en paralelo con una librería de paso de mensajes estándar en la industria."* El *"Proyecto Beowulf"* fue desarrollado originalmente en el *Goddard Space Flight Center* (GSFC). También son populares los clusters con procesadores DEC/Alpha.**

Top 500 cluster

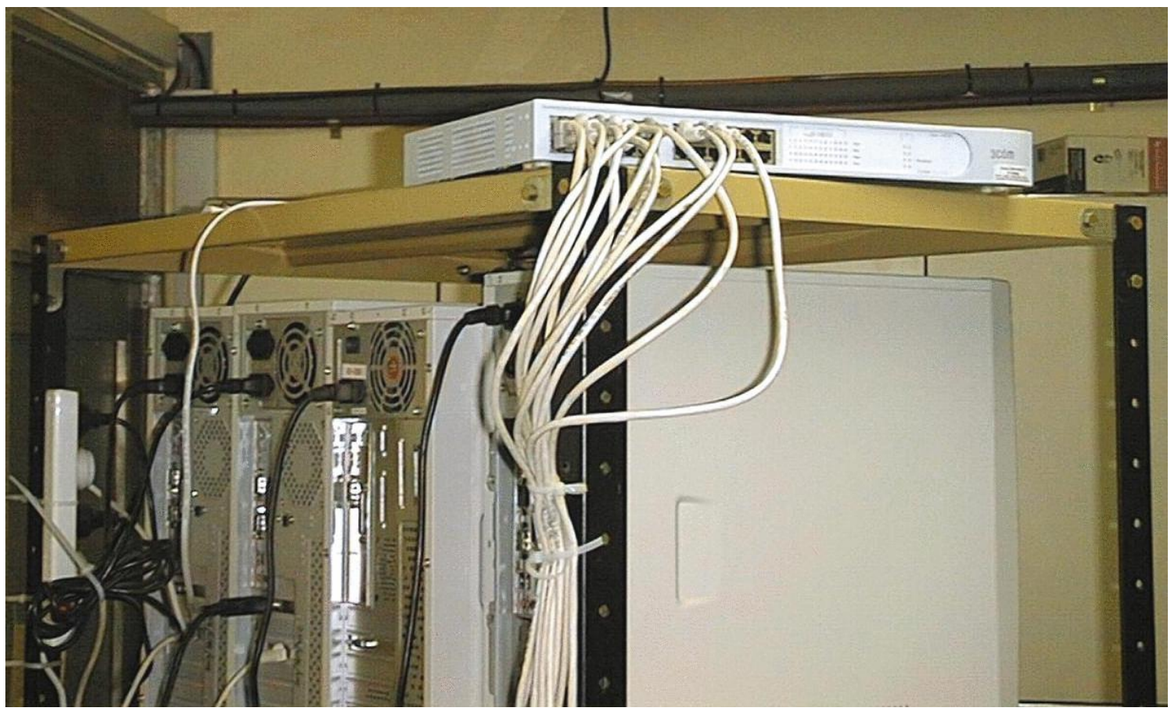
- Mantenido en `http://www.beowulf.org`
- Intitución: AIST - Computational Biology Research Center, Japan.
  Nombre: CBRC Magi System. Integrador: NEC
  # de nodos: 520. # de procesadores: 1040.
  Performance: 967.20 Gflops. Tipo de red: Myrinet
- Intitución: Real World Computing, Japan.
  Nombre: CBRC RWC SCore Cluster III. Integrador: Self-made
  # de nodos: 512. # de procesadores: 1024.
  Performance: 955.40 Gflops. Tipo de red: Myrinet 2000
- Intitución: Partnership American Museum of Natural History. USA
  Nombre: ABACUS. Integrador: ???
  # de nodos: 281. # de procesadores: 564.
  Performance: 432.13 Gflops. Tipo de red: Fast Ethernet
- Intitución: Chemnitz University of Technology. Germany
  Nombre: CLiC. Integrador: Self-made
  # de nodos: 528. # de procesadores: 528.
  Performance: 422.40 Gflops. Tipo de red: Fast Ethernet

El cluster “Geronimo” en el CIMEC

- **El CIMEC (Centro Internacional de Métodos Numéricos en Ingeniería, ubicado en Santa Fe, dependiente del CONICET) desarrolla tareas de investigación en métodos numéricos desde 1982.**
- **Desde el año 1997 se vienen desarrollando experiencias en procesamiento distribuido. Originalmente en 4 procesadores DEC/Alpha 500/333 Mhz.**
- **Desde 1999 este esfuerzo se ha orientado a los cluster de PC corriendo bajo GNU/Linux. Actualmente, el cluster cuenta con 12 procesadores Pentium 4 1.4-1.7Ghz con 512 Mb RAM (Rambus) conectado a través de un switch Fast Ethernet (100 Mbit/sec, latency=$O(100)$) 3COM SuperStack 3300.**
- **Configuración “disk-less”**



El cluster “Geronimo” en el CIMEC (cont...)



Software utilizado

- **El software instalado en el cluster está basado en la distribución RedHat 7.1.**
- **Desarrollamos un código de elementos finitos escrito en C++ llamado PETSc-FEM. Actualmente tiene +35,000 líneas de código.**
- **PETSc-FEM también compila con diferentes versiones de GCC y EGCS hasta incluso la 2.96.**
- **Como librería de paso de mensajes usamos MPI (*"Message Passing Interface"*, `http://www.mcs.anl.gov/mpi`) en su implementacion MPICH (versión 1.0.2, `http://www.mcs.anl.gov/mpich`) desarrollados en el *Argonne National Laboratory* (ANL).**

Software utilizado. (cont.)

- **En general MPI no es llamado directamente sino a través de PETSc (versión 2.1.3) *Parallel Extensible Toolkit for Scientific Computations* que es un paquete orientado a métodos numéricos en procesamiento distribuido, y permite operaciones abstractas como definir vectores y matrices distribuidos y resolver los sistemas lineales asociados.**

Software utilizado. (cont.)

- **Particionamiento de malla utilizando Metis. Este particionador de malla permite dividir el *"grafo dual"* (es decir áquel donde los elementos son vértices del grafo y los nodos son aristas que conectan los vértices) de conectividades en subdominios tratando de mantener la masa total de los vértices (esfuerzo computacional en computar los elementos) igual entre los diferentes subdominios manteniendo mínima la comunicación entre procesadores (la cual se define asignando un peso a las aristas del grafo dual). En el caso de realizar *"balance de carga"* la masa total de los vértices en cada procesador debe ser proporcional a la velocidad del procesador. (`http://www.cs.umn.edu/~metis`, `http://www.cs.umn.edu/~karypis/metis`).**

Software utilizado. (cont.)

- **Para el álgebra lineal se utiliza los paquetes estándar Lapack y Blas distribuidos por Netlib (`http://www.netlib.org/lapack/`).**
- **Contenedores abstractos de *Libretto* (`http://pobox.com/~aaronc/tech/libretto/`), Glib (http://www.gtk.org) y la C++ STL Template Library que viene con el compilador GNU gcc.**
- **Librería de matrices Newmat (`http://webnz.com/robert/`). Esta está siendo reemplazada por una liberría desarrollada por nosotros mismos llamada FastMat2.**
- **Cantidad de GNU – Open Source software como: TEX/LATEX, Emacs, Xfig, Tgif, Octave, Perl y muchos otros.**

Características de PETSc-FEM

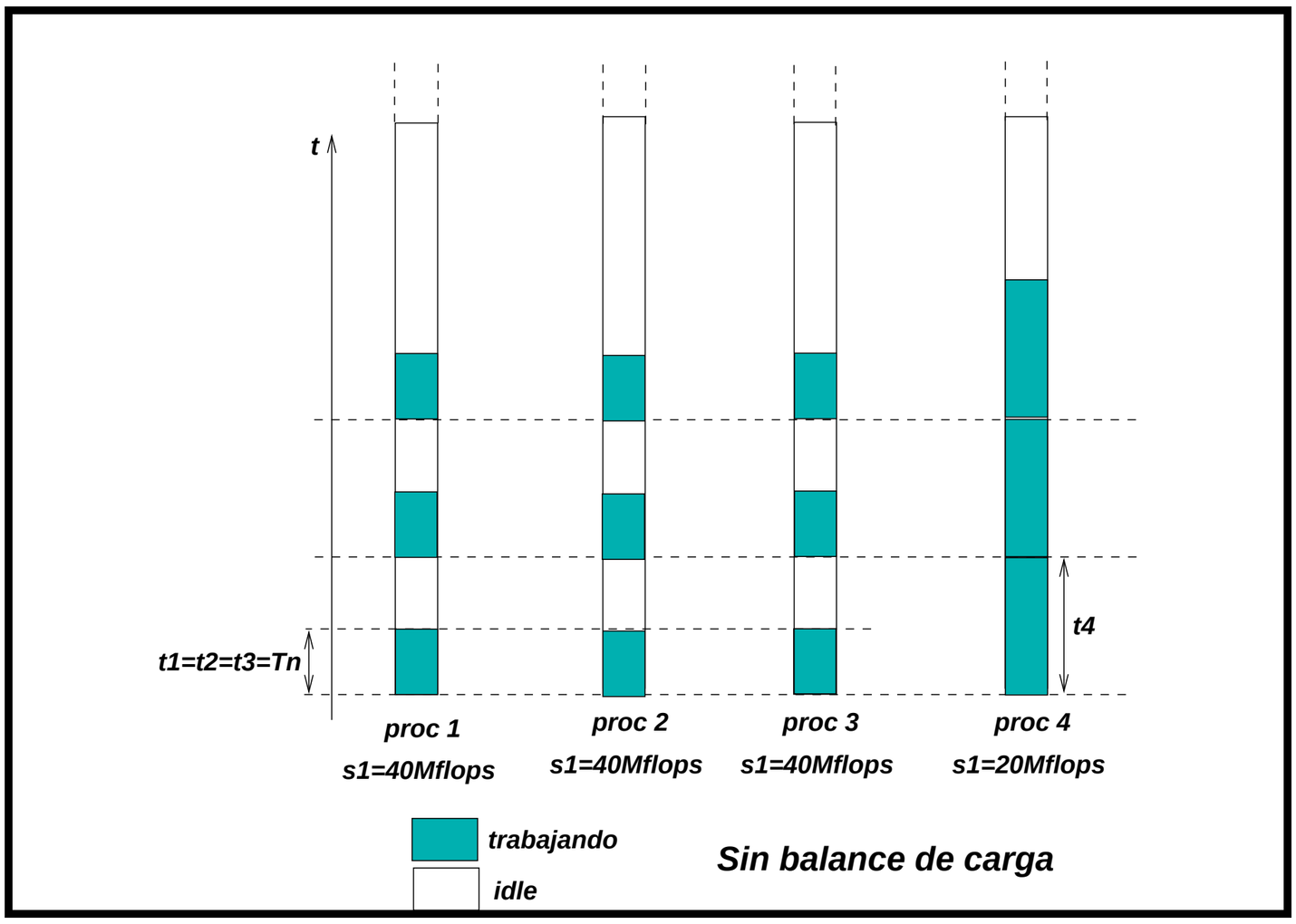
- **GPL, accesible en `http://minerva.ceride.gov.ar/petscfem`, versión actual es petscfem-beta-3.01.**
- **Programa de elementos finitos de uso general y multi-física.**
- **Procesamiento en paralelo via uso de PETSc/MPI/Metis.**
- **Resolución de sistemas lineales via el Método de Descomposición de Dominios (DDM).**
- **Escrito en C++ con una concepción OOP, con especial énfasis en la eficiencia.**
- **Muchos tipos de elementos pueden ser usados en la misma aplicación, agrupando los elementos del mismo tipo en “elemsets”.**

Características de PETSc-FEM

- **Condiciones de contorno Dirichlet/Newman/mixtas/periódicas o restricciones generales.**
- **Cálculo de jacobianos numéricos.**
- **Actualmente implementados módulos de Navier-Stokes, hidrología sub-superficial, shallow-water, Euler, sistemas advectivos generalizados y ec. de Laplace.**
- **Perfiles de sistemas de ecuaciones calculados automáticamente.**
- **Librerías de matrices rápidas usando “caches” para las direcciones de memoria.**
- **Balance de carga**

Rendimiento en clusters heterogéneos

- **Si un procesador es más rápido y se asigna la misma cantidad de tarea a todos los procesadores independientemente de su velocidad de procesamiento, cada vez que se envía una tarea a todos los procesadores, los más rapidos deben esperar a que el más lento termine, de manera que la velocidad de procesamiento es a lo sumo n veces la del procesador más lento. Esto produce un deterioro en la performance del sistema para clusters heterogéneos. El concepto de speedup mencionado anteriormente debe ser extendido a grupos heterogéneos de procesadores.**


t
t1=t2=t3=Tn
t4
proc 1
s1=40Mflops
proc 2
s1=40Mflops
proc 3
s1=40Mflops
proc 4
s1=20Mflops
trabajando
idle
Sin balance de carga

Rendimiento en clusters heterogéneos (cont.)

- **El trabajo W (un cierto número de operaciones) es dividido en n partes iguales $W_i = W/n$**
- **Cada procesador tarda un tiempo $t_i = W_i/s_i = W/ns_i$ donde s_i es la velocidad del procesador i (por ejemplo en Mflops). El hecho de que los t_i sean distintos entre si indica una pérdida de eficiencia.**
- **El tiempo total transcurrido es el mayor de los t_i, que corresponde al menor s_i:**

$$T_n = \max_i t_i = \frac{W}{n \min_i s_i} \quad (1)$$

- **El tiempo T_1 podemos tomarlo como el obtenido en el más rápido, es decir**

$$T_1 = \min t_i = \frac{W}{\max_i s_i} \quad (2)$$

Rendimiento en clusters heterogéneos (cont.)

- **El speedup resulta ser entonces**

$$S_n = \frac{T_1}{T_n} = \frac{W}{\max_i s_i} \Big/ \frac{W}{n \min_i s_i} = n \frac{\min_i s_i}{\max_i s_i} \qquad (3)$$

El *"factor de desbalance"* $\min_i s_i / \max_i s_i$ **puede llegar en nuestro caso a** 0.82 **(1.4GHz/1.7Ghz) con lo cual el speedup teórico puede llegar a bajar de** 12 **a** 9.8**, con lo cual puede llegar a convenir no introducir los procesadores mas lentos. Esto sin tener en cuenta el deterioro en el speedup por los tiempos de comunicación.**

Balance de carga

- **Si distribuimos la carga en forma proporcional a la velocidad de procesamiento**

$$W_i = W \frac{s_i}{\sum_j s_j}, \quad \sum_j W_j = W \tag{4}$$

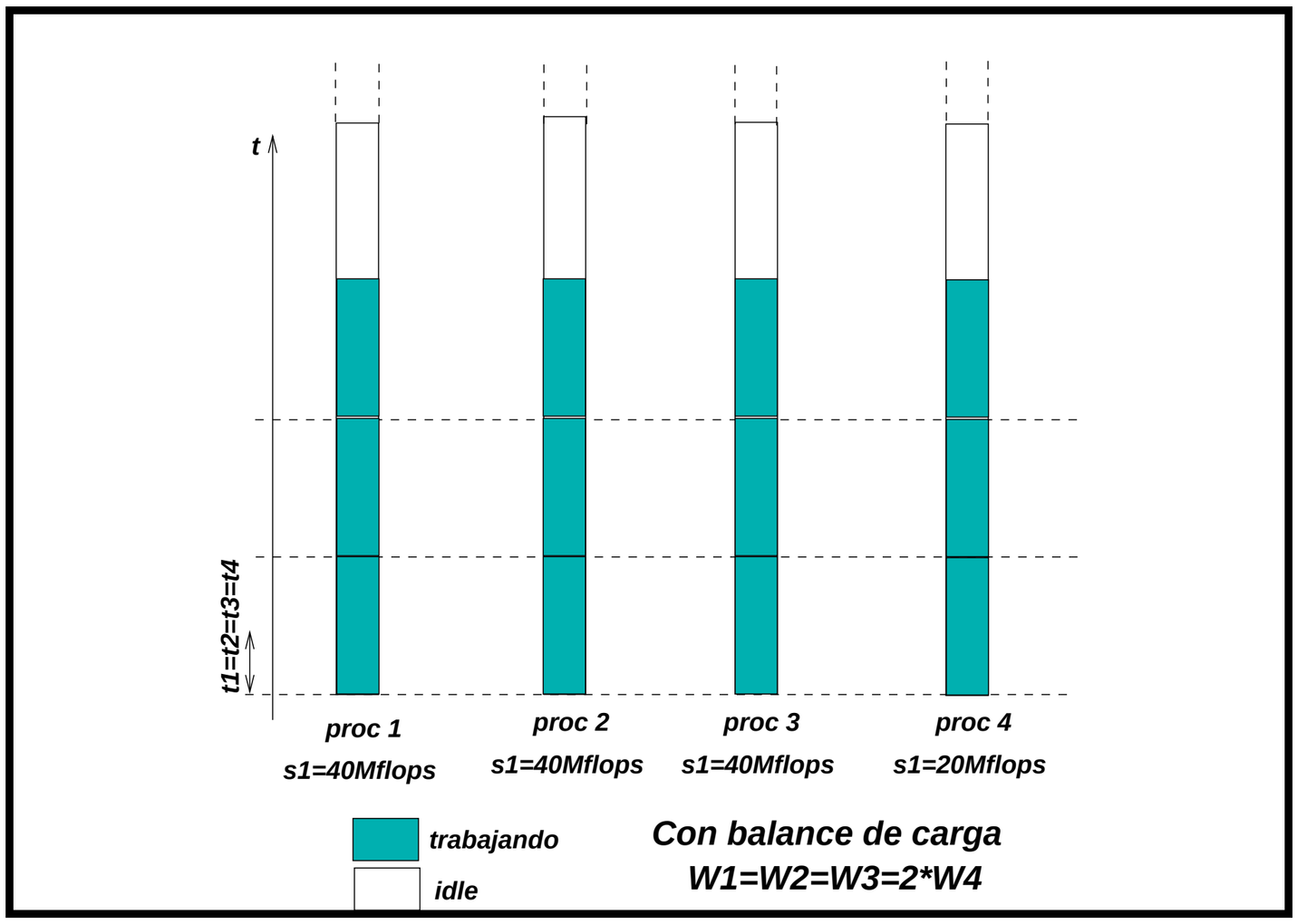
- **El tiempo necesario en cada procesador es**

$$t_i = \frac{W_i}{s_i} = \frac{W}{\sum_j s_j} \quad \textbf{(independiente de } i\textbf{!!!)} \tag{5}$$

- **El speedup ahora es**

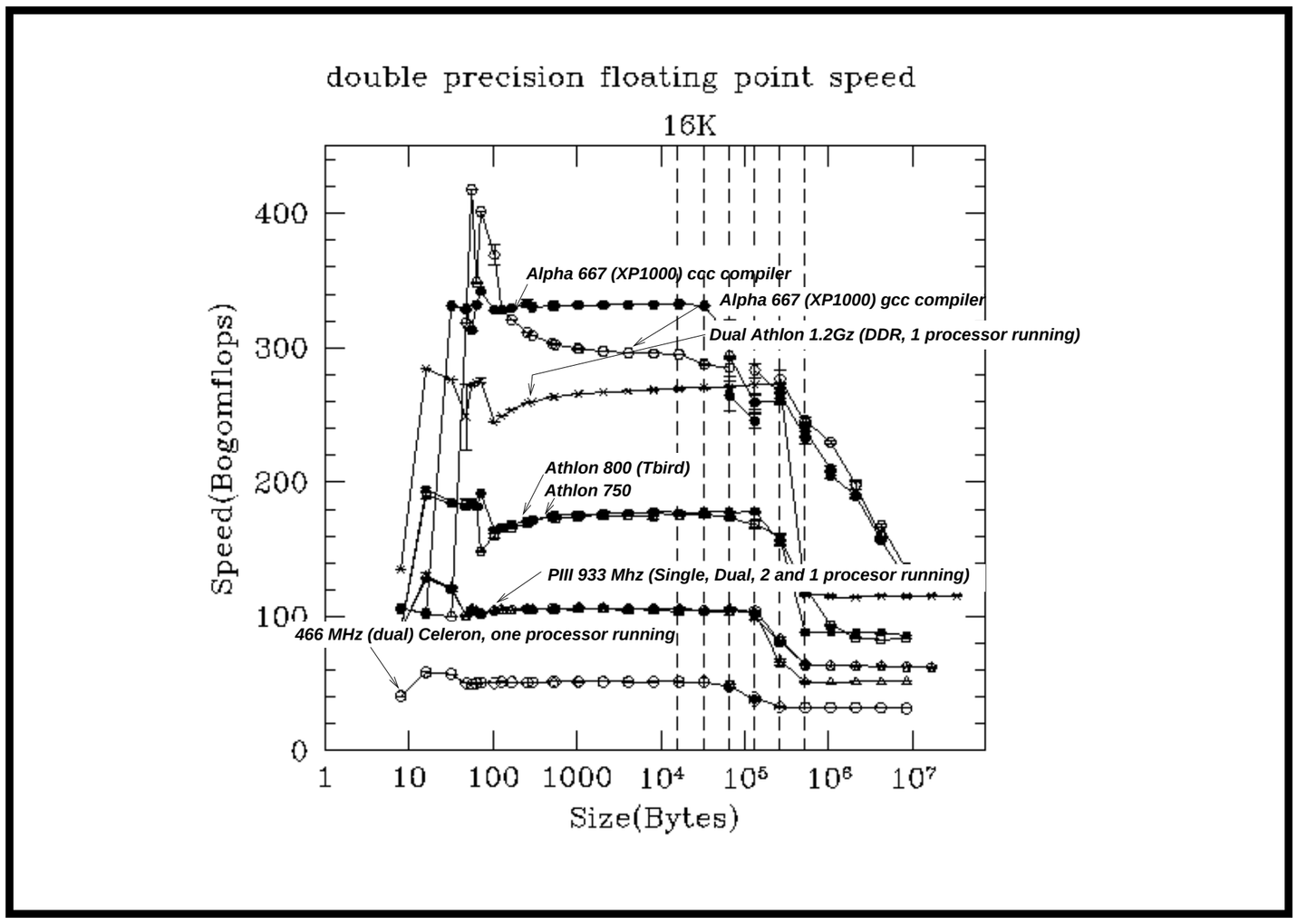
$$S_n = \frac{T_1}{T_n} = (W / \max_j s_j) / \left(\frac{W}{\sum_j s_j} \right) = \frac{\sum_j s_j}{\max_j s_j} \tag{6}$$

Este es el máximo speedup teórico alcanzable en clusters heterogéneos.


t
t1=t2=t3=t4
proc 1
s1=40Mflops
proc 2
s1=40Mflops
proc 3
s1=40Mflops
proc 4
s1=20Mflops
trabajando
idle
Con balance de carga
W1=W2=W3=2*W4

Balance de carga estático en PETSc-FEM

- **Actualmente PETSc-FEM permite balancear en forma estática la carga, esto es, una vez determinada la velocidad de los procesadores estos son leidos antes de comenzar la ejecución y la partición de la malla se realiza de manera de mantener en cada procesador la misma cantidad de “carga” (número de elementos).**
- **Desventajas del balance estático: Como tomar la velocidad del procesador?**
- **El procesador parece tener diferentes velocidades dependiendo del compromiso entre número de operaciones realizadas y cantidad de información que tiene que fluir desde la memoria al procesador.**
- **A veces el trabajo a realizar en el procesador no es el mismo para todos los elementos.**


double precision floating point speed
16K
Speed(Bogomflops)
400
300
200
100
0
Alpha 667 (XP1000) ccc compiler
Alpha 667 (XP1000) gcc compiler
Dual Athlon 1.2Gz (DDR, 1 processor running)
Athlon 800 (Tbird)
Athlon 750
PIII 933 Mhz (Single, Dual, 2 and 1 procesor running)
466 MHz (dual) Celeron, one processor running
1
10
100
1000
10⁴
10⁵
10⁶
10⁷
Size(Bytes)

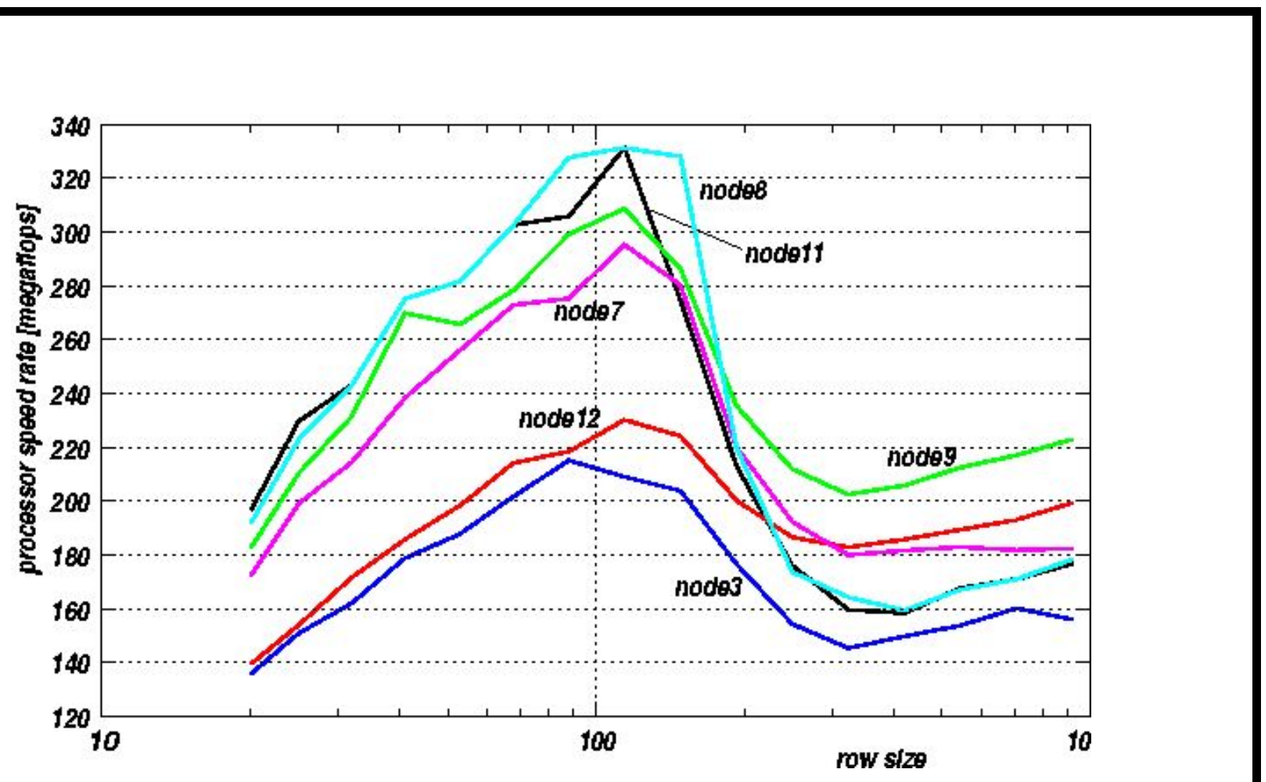


Figura 3: Performance de procesadores P4

Trabajo futuro – Kernel

- **Balance dinámico de carga**
- **Refinamiento adaptativo**
- **Desarrollo de mejores precondicionarores para el Método de Descomposición de Dominios**
- **Llamada interactiva o scripting desde Perl/Guile/otros...**

Trabajo futuro – Modulos

- **En el módulo de hidrología**
  - ▷ **Incorporar varios acuíferos y subcapas**
  - ▷ **Transporte de contaminantes acoplado con el flujo**
  - ▷ **Acoplar el modelo de shallow water con el flujo subterráneo**
  - ▷ **Soporte para refinamiento adaptativo**
- **En el módulo de Navier-Stokes**
  - ▷ **Flujo multifásico**
  - ▷ **ALE / mallas móviles**
  - ▷ **Reacciones químicas / Combustión**