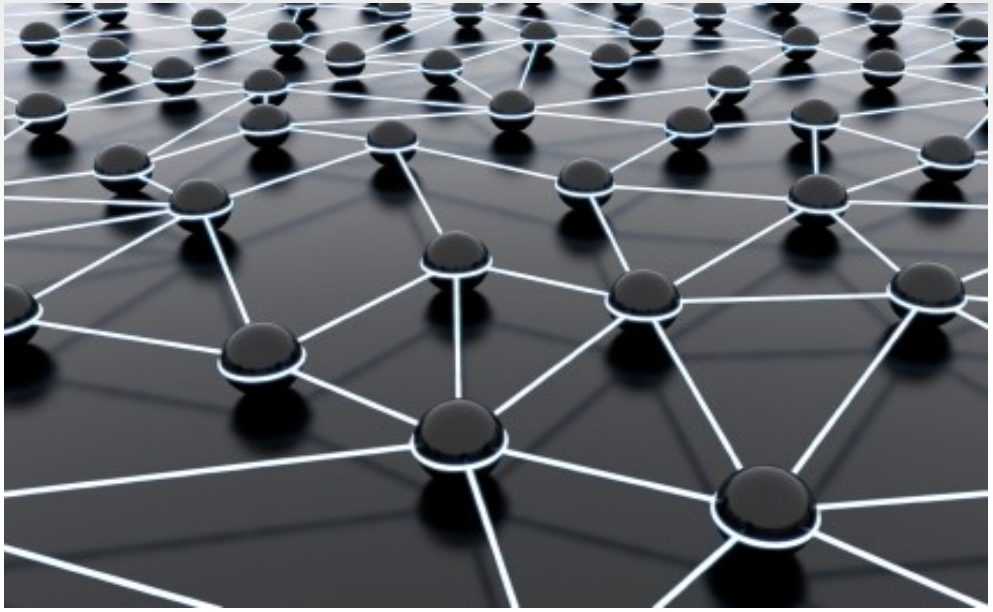
Extensió de COMP Superscalar

9 de Febrer de 2011



escola d'enginyeria

UAB
Universitat Autònoma de Barcelona

Alumne: Roger Rafanell Mas
Direcció: Porfidio Hernández Budé

Índex

- **Introducció**
- Desenvolupament del projecte
- Experiments i resultats
- Conclusions

Context – Paradigmes (1)

- **Paradigma seqüencial:**
  - Aplicacions importants programades encara en seqüencial.
  - No aprofiten infraestructures/arquitectures existents.
- **Paradigmes de programació paral·lela:**
  - L'usuari ha de modificar el codi de l'aplicació.
  - Dependència forta del llenguatge de programació.
  - Exemples: MPI, OpenMP, OpenMP+MPI, MapReduce, R, etc...

Context – Paradigmes (2)

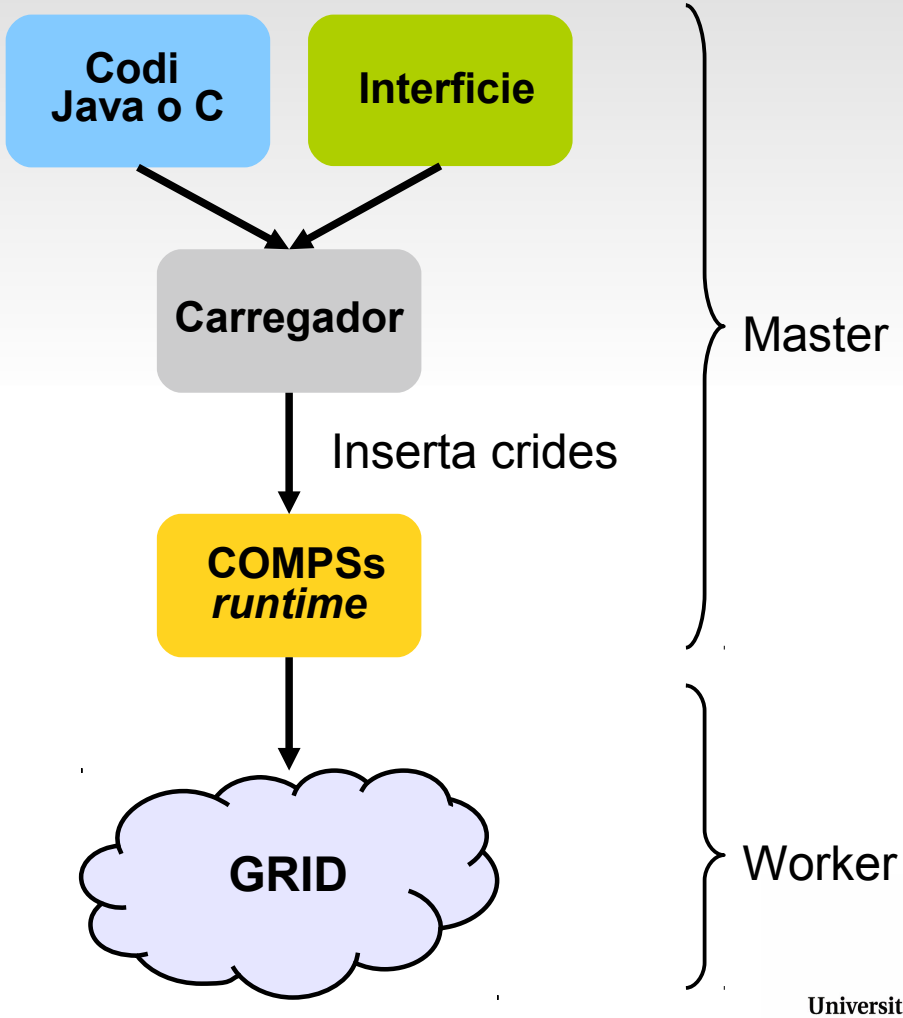
RECODIFICAR APLICACIONS

PROVOCA RESISTÈNCIA AL CANVI!

Context – Nous paradigmes

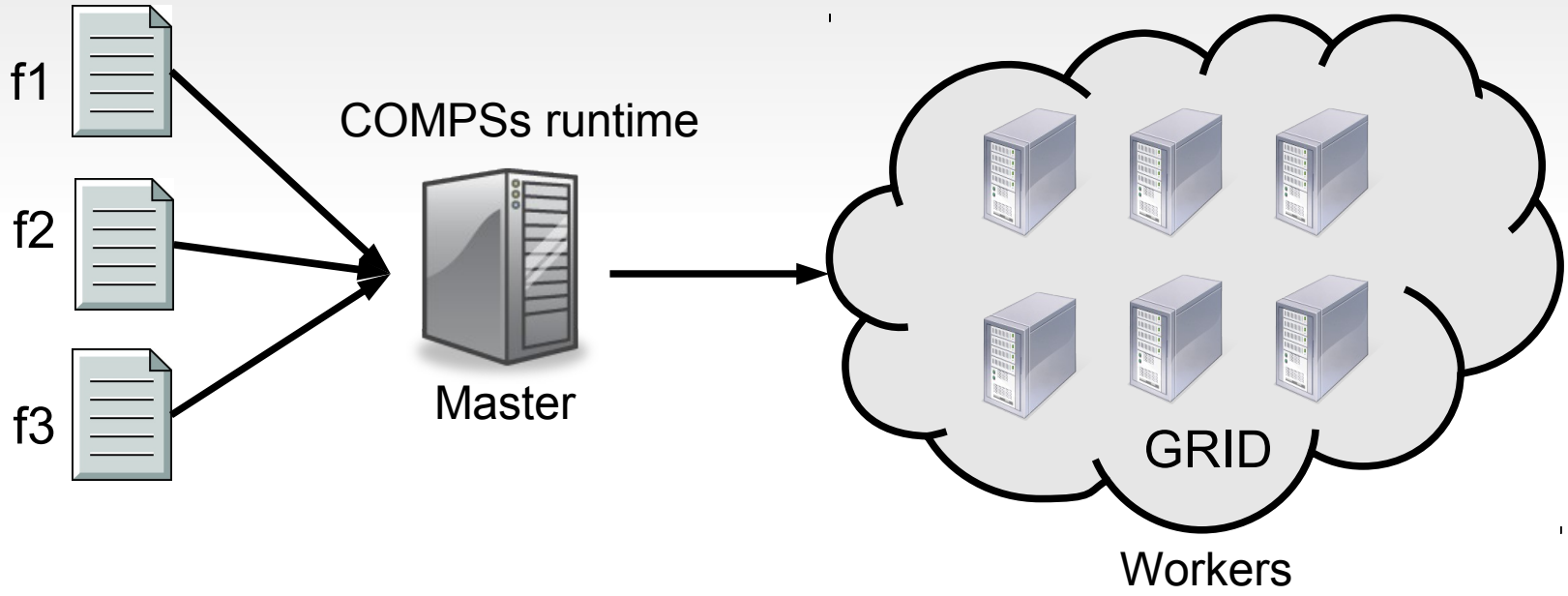
- **Programació seqüencial amb execució paral·lela**
  - Paralelització automàtica
    - Compiladors especifics (Multi-core)
    - Transparent a l'usuari
  - Paralelització semiautomàtica
    - COMP Superscalar (Grid, Cloud, Cluster)

COMP Superscalar



COMP Superscalar – Problemes (1)

- **COMPSs no té "consciència" entre execucions dels fitxers transferits.**

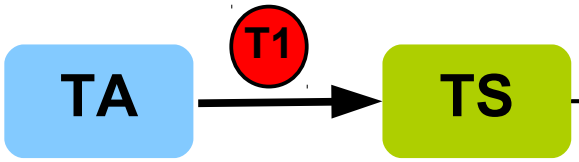


PROBLEMA → COMPSs transfereix cada vegada els fitxers al Grid.

COMP Superscalar – Problemes (2)

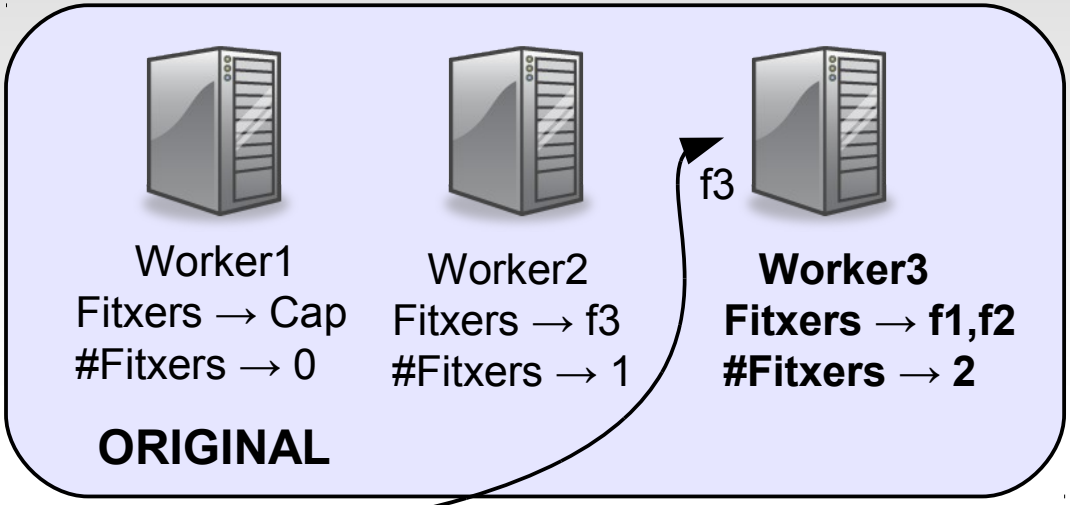
Tasca T1 (f1, f2, f3, f4)

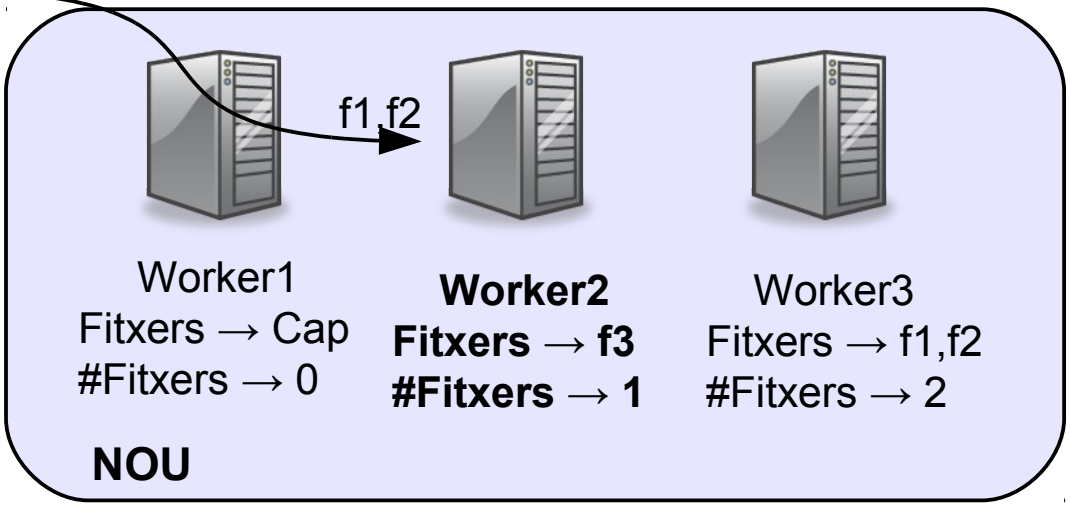
- f1 → 20MB → IN
- f2 → 200MB → IN
- f3 → 2GB → IN
- f4 → OUT



Original → f3 = 2GB
Nou → f1,f2 = 0.22GB

escola d'enginyeria





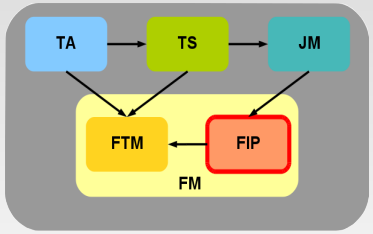
Objectius – Extensions

- **Extensions proposades:**
  - Desenvolupar un gestor de rèpliques per a minimitzar transferències.
  - Desenvolupar un planificador basat en la predicció del temps d'execució.
  - Millorar la tolerància a fallades a nivell de planificador.

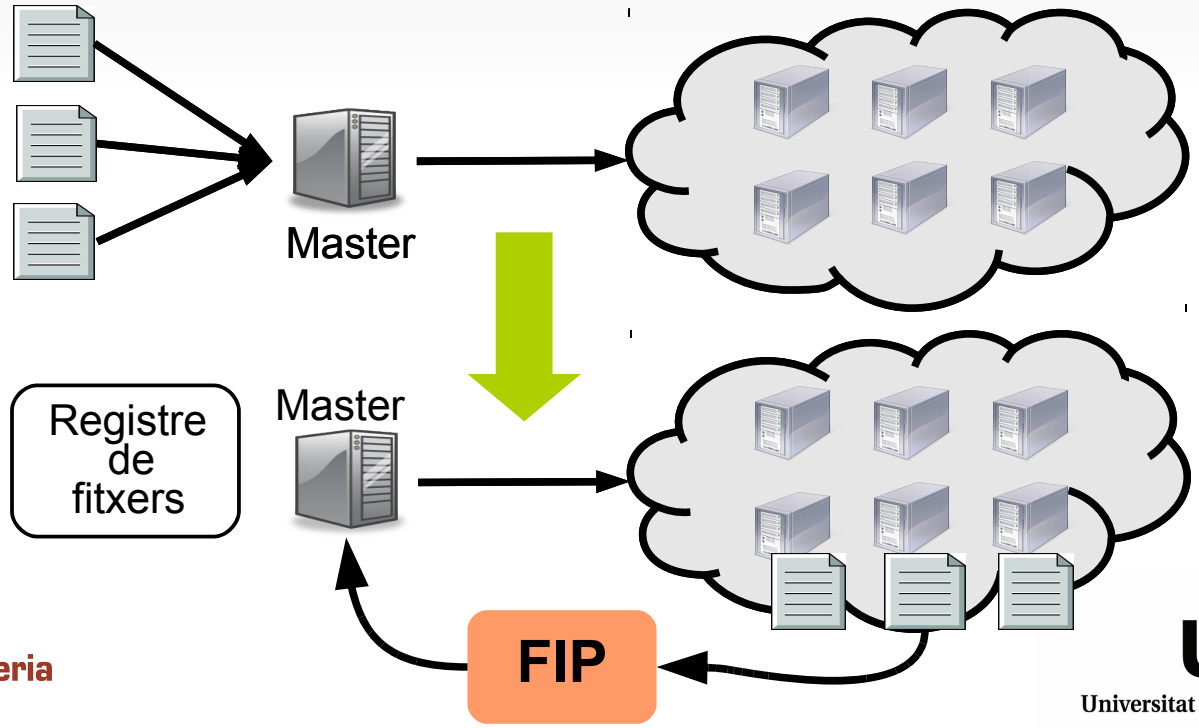
Índex

- Introducció
- **Desenvolupament del projecte**
- Experiments i resultats
- Conclusions

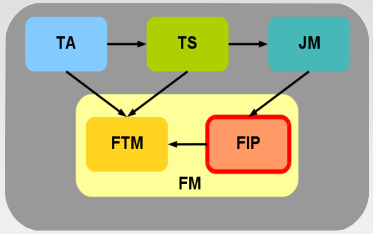
Desenvolupament – Rèpliques (1)



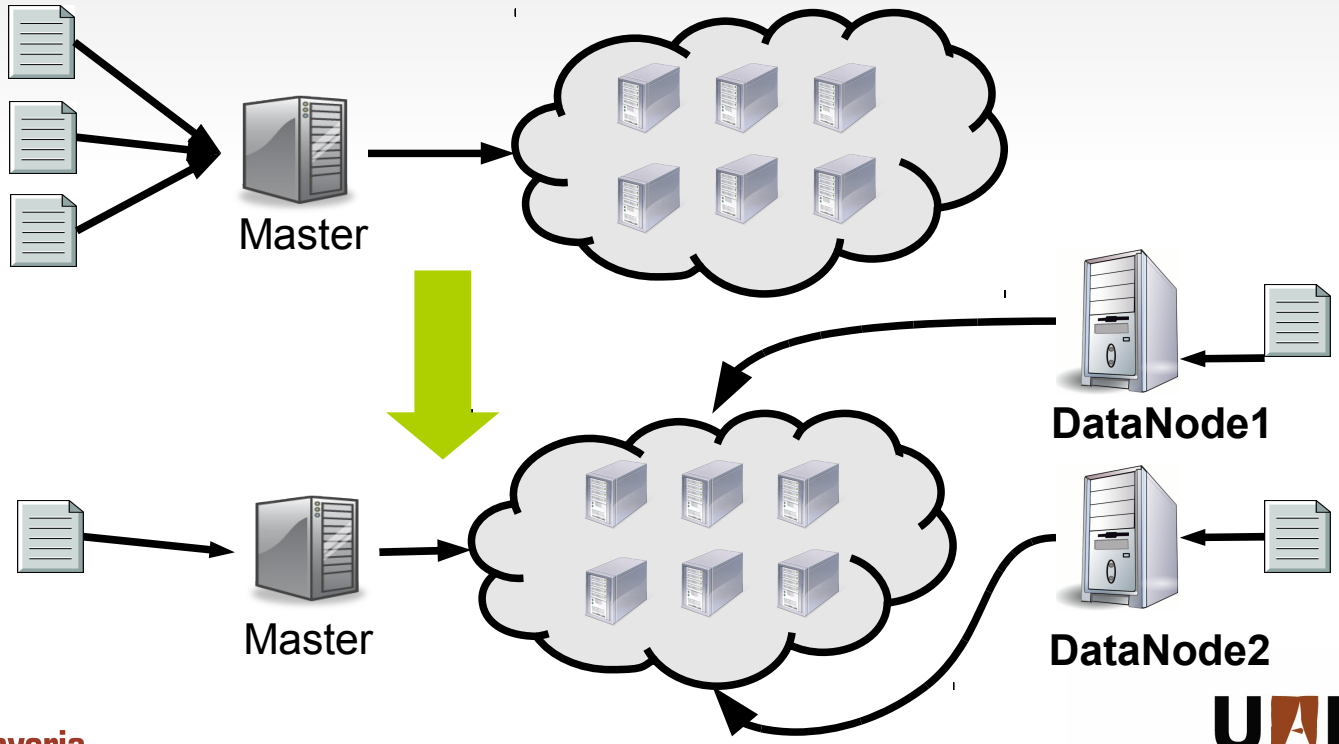
- **El gestor de rèpliques:**
  - Permet reduïr el volum de transferències de les aplicacions.
  - Els canvis en els fitxers, es controlen a través de l'última data de modificació.



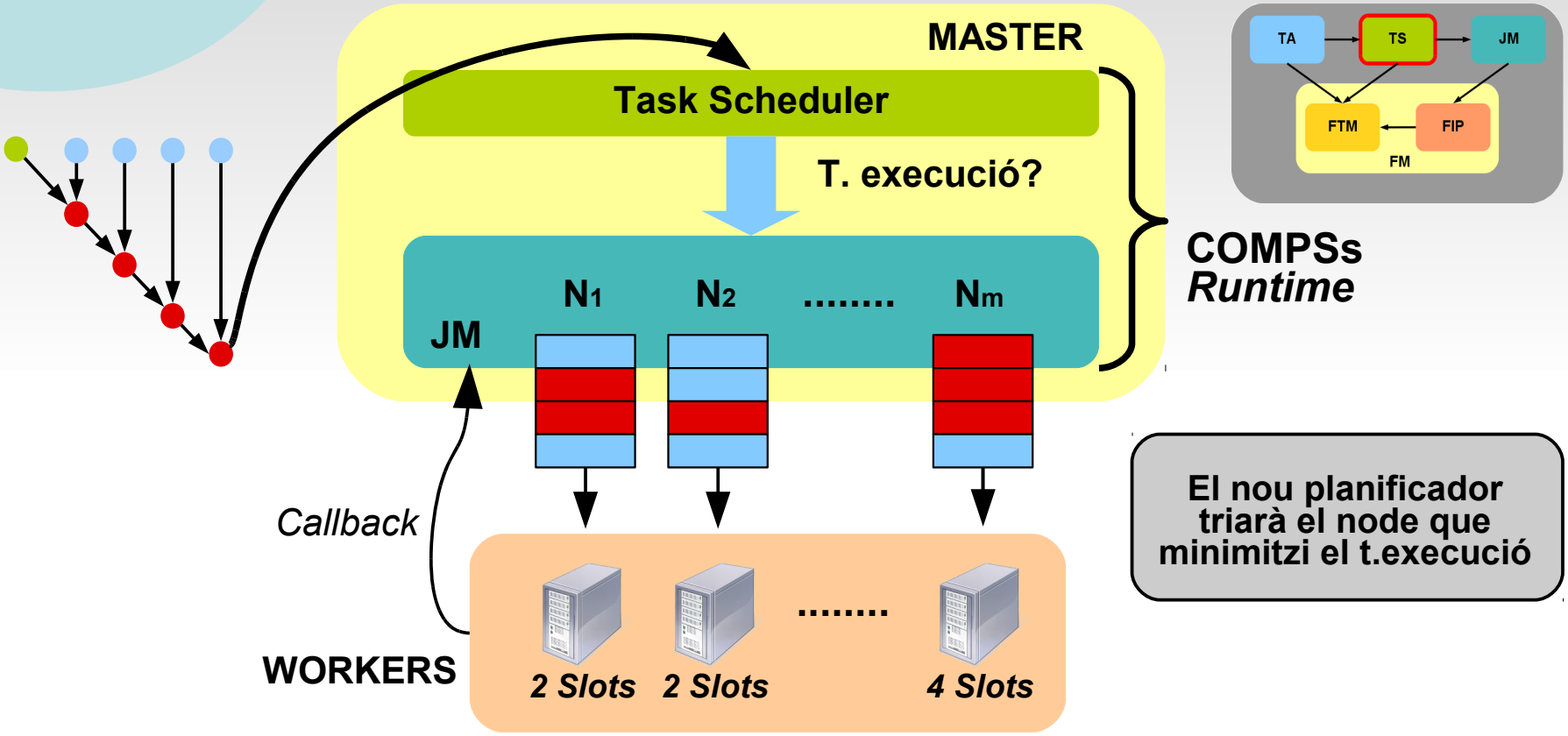
Desenvolupament – Rèpliques (2)



- **Habilitació de *DataNodes*:**
  - **DataNode:** node (no Master) d'on agafar fitxers d'entrada.



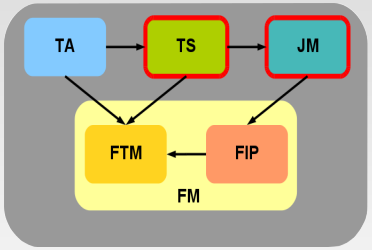
Desenvolupament – Planificador



$$ExecutionTime_r = TrfPrediction_{t,r} + WaitTimeInQueue_r + TypExecTime_{t,r}$$

NECESSITAT D'ACUMULAR HISTÒRIC!

Desenvolupament – Fiabilitat



- **La fiabilitat a la planificació:**
  - *Ranking* dels temps d'execució.
  - Ponderem segons el *ratio* de fiabilitat.
  - Ordenem de gran a petit (de millor a pitjor).

Recursos	MethodId	TP	WT	ET	AvailRate	Rank	Score
host1	1	0.11s	1504.1484s	0s	0.9	2	1.8
host2	**1**	**0.10s**	**1316.8849s**	**0s**	**1.0**	**3**	**3**
host3	1	0.15s	1254.6338s	0s	0.7	4	2.8
host4	1	0.09s	8451.267s	0s	1.0	1	1

$$RatioFiabilitat_r = \frac{FeinesOK_r}{FeinesEnviades_r} \quad on\ r = recurs$$

Problema → Exclusió permanent de nodes!

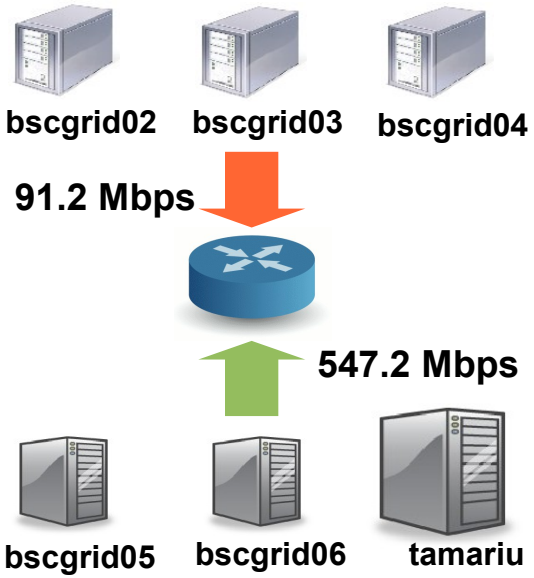
Sol·lució → Mètode de recuperació de fiabilitat

Índex

- Introducció
- Desenvolupament del projecte
- **Experiments i resultats**
- Conclusions

Entorn de proves

Recursos	CPU	Velocitat	#Cores	Memòria	Disc
bscgrid02	Intel Pentium 4 Dual Core 2MB *caché*	3.6Ghz	2	1GB	60 GB
bscgrid03	Intel Pentium 4 Dual Core 2MB *caché*	3.6Ghz	2	1GB	60GB
bscgrid04	Intel Pentium 4 Dual Core 2MB *caché*	3.6Ghz	2	1GB	60GB
bscgrid05	Intel Q9300 Core 2 Quad 3MB *caché*	2.5Ghz	4	4GB	220GB
bscgrid06	Intel Q9300 Core 2 Quad 3MB *caché*	2.5Ghz	4	4GB	220GB
tamariu	Intel Xeon E7450 12MB *caché*	2.4Ghz	4x6 = 24	47GB	550GB 320 GB



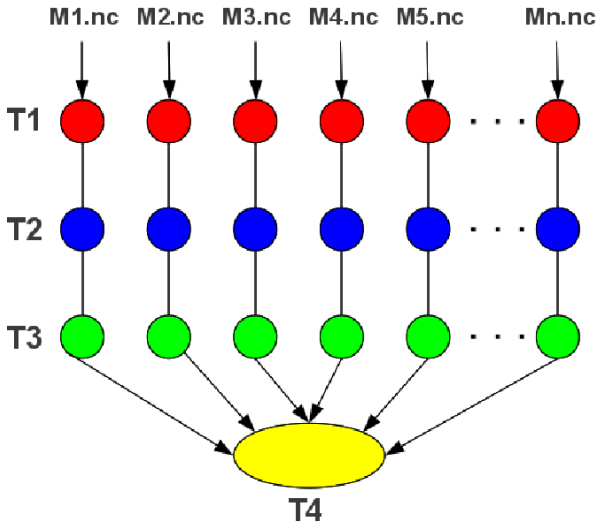
Node **tamariu** no dedicat!

escola d'enginyeria

UAB Universitat Autònoma de Barcelona

Avaluació de resultats (1)

- Anàlisi del gestor de rèpliques:
  - **Aplicació de test: JRA4**
    - Predicció de temperatures en superfícies.
    - Predicció a partir de models de temperatura.
    - Complexitat $3(N_f)+N_f \rightarrow$ **$4N_f$.**
    - Fitxers de models grans (>1GB).

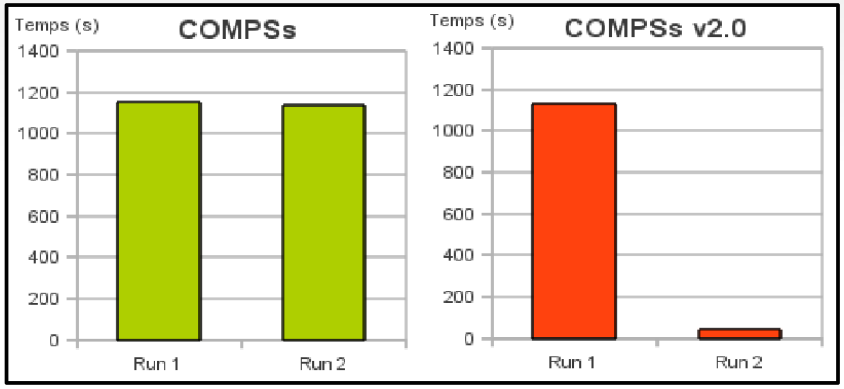


escola d'enginyeria

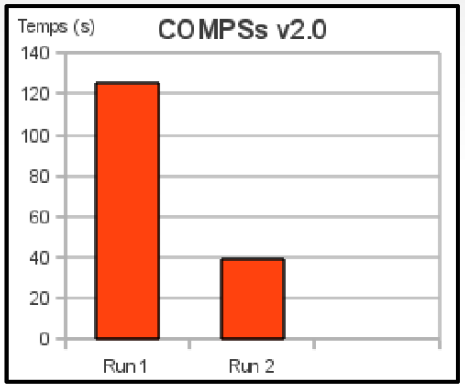
UAB Universitat Autònoma de Barcelona

Avaluació de resultats (2)

- Anàlisi del gestor de rèpliques:
  - **Anàlisi del temps d'execució → 10 Workers i 12 models d'1GB**



T.Execució → sense rèpliques vs rèpliques
(1150s i 1150s vs 1133s i 39s)



Modificació d'un fitxer
(125s vs 39s)

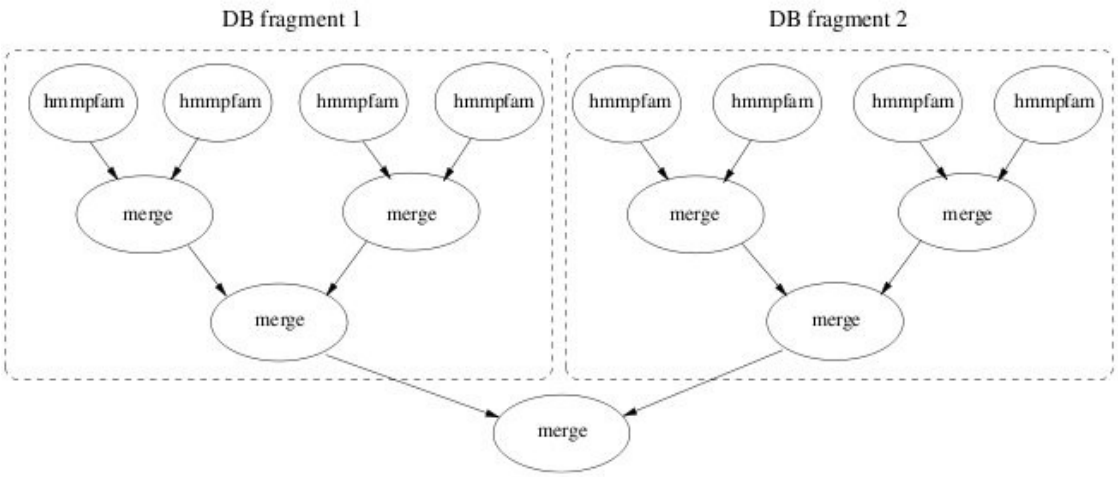
bscgrid02 → 2W
bscgrid03 → 2W
bscgrid04 → 2W
bscgrid06 → 4W

escola d'enginyeria

UAB Universitat Autònoma de Barcelona

Avaluació de resultats (3)

- Anàlisi de rendiment del nou planificador:
  - **Aplicació de test: HMMER**
    - Suite bioinformàtica per l'alineament de seqüències.
    - Detecta concidències entre cadenes de ADN.
    - Complexitat **$2(F_s+F_{db}) - 1$.**
  - **Graf que genera l'aplicació:**



Avaluació de resultats (4)

- Anàlisi de rendiment del nou planificador:
  - **Anàlisi del temps d'execució**

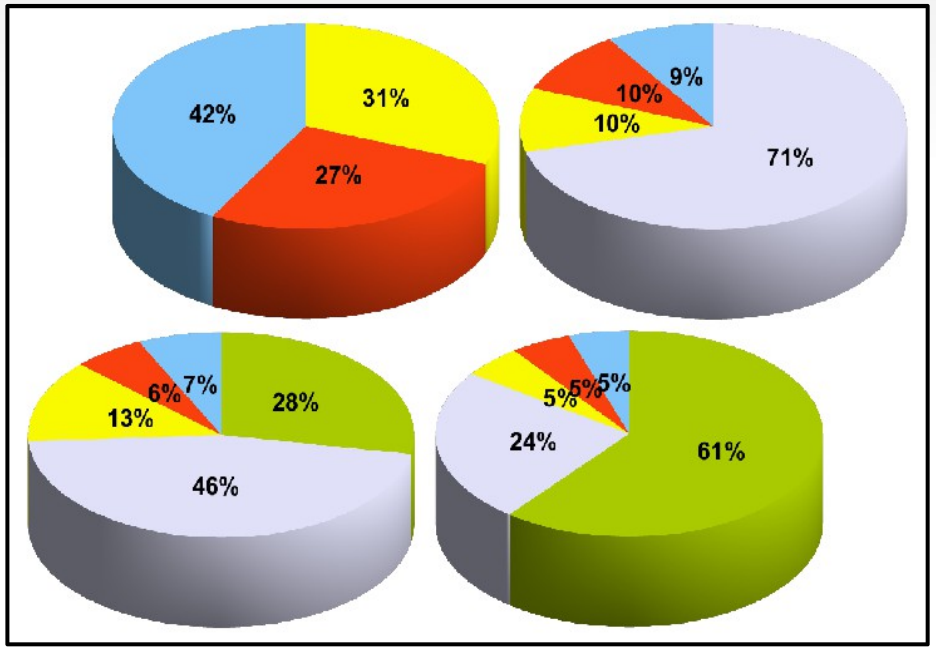
4096 S | 1279 T	COMPSs	COMPSs v2.0	Millora
2 Workers	7342.74	7762.90	-5.72%
4 Workers	3819.44	4264.4	-11.65%
6 Workers	2667.91	3059.05	-14.66%
10 Workers	1517.55	1481.44	2.38%
14 Workers	1395.19	1051.02	24.67%
18 Workers	1313.20	1021.93	22.18%
22 Workers	1250.30	1008.33	19.35%
8192 S | 2559 T	**COMPSs**	**COMPSs v2.0**	**Millora**
2 Workers	15480.51	15654.31	-1.12%
4 Workers	8058.07	8539.07	-5.97%
6 Workers	5931.41	5891.61	0.67%
10 Workers	5105.58	2867.39	43.84%
14 Workers	5007.11	2255.92	54.94%
18 Workers	4843.51	2050.09	57.67%
22 Workers	4780.31	1844.27	61.42%

El *overhead* del nou planificador perd pes respecte la millora en la planificació

2W → 02
4W → 02+03
6W → 02+03+04
10W → 02+03+04+06
14W → 02+03+04+06+t(4)
18W → 02+03+04+06+t(8)
22W → 02+03+04+06+t(12)

Avaluació de resultats (5)

- Anàlisi de rendiment del nou planificador:
  - **Balanceig de tasques**

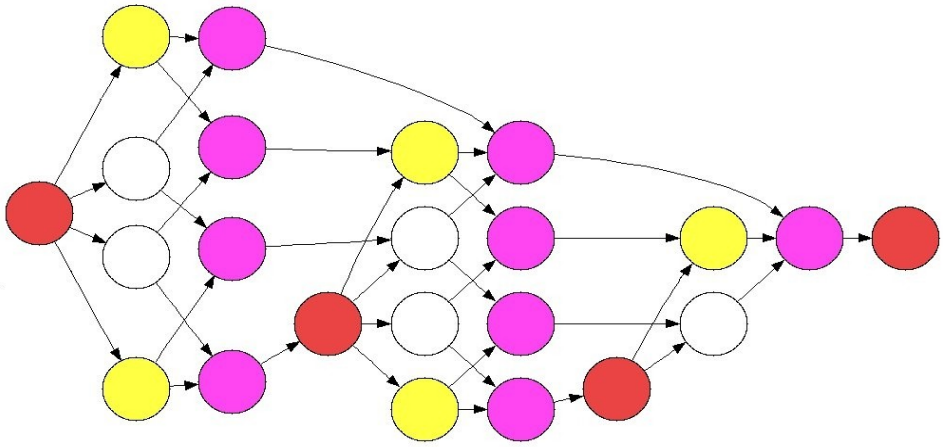


bscgrid02 → 2W
bscgrid03 → 2W
bscgrid04 → 2W
bscgrid06 → 4W
tamariu → 4W/12W

Balanceig de tasques en HMMER de 8192 seqüències amb **6,10,14 i 22** Workers

Avaluació de resultats (6)

- Anàlisi de la tolerància a fallades:
  - **Aplicació de test: SparseLU**
    - Factorització de matrius mitjançant la descomposició LU.
    - Nombre de tasques fàcilment variable → variant dimensió matriu.
    - Complexitat **O(n³)**.
  - **Graf que genera l'aplicació:**



Avaluació de resultats (7)

- Anàlisi de la tolerància a fallades:
  - **Experiment: 10 Workers i matriu de 20x20 de blocs 4x4 (6400 elements)**
    - Simulació de diversos nombres de fallades a la màquina més potent.
    - El planificador varia el nombre de tasques assignades a cada Worker.
  - **Obtenim la següent taula:**

Fiabilitat	**100%**	**90%**	**80%**	**75%**	**70%**
bscgrid02	14%	13%	54%	5%	30%
bscgrid03	12%	10%	5%	32%	51%
bscgrid04	7%	19%	3%	44%	19%
bscgrid06	**67%**	**58%**	**38%**	**20%**	**0%**
Tasques Ok	890	791	692	643	594
Fallades	0	99	198	247	296
Temps exec.	1011 s	1094 s	1512 s	1592 s	1610 s

El planificador tolera taxes d'error < del 30%

Índex

- Introducció
- Desenvolupament del projecte
- Experiments i resultats
- **Conclusions**

Conclusions del projecte

- El prototip assoleix els objectius plantejats inicialment:
  - Gestiona rèpliques de fitxers d'entrada (evitant transferències).
  - Planificador pren decisions més acurades.
  - Considera l'índex de fallades dels nodes.
  - Bon rendiment en aplicacions d'ús reiterat (històric).
- El model segueix mantenint l'arquitectura i elegància inicials.

Treball futur

- Línies d'evolució proposades:
  - Implementar política de reemplaçament de rèpliques.
  - Reduïr més el consum de memòria.
  - Obtenir la velocitat de la xarxa a través d'algun *Grid Information Service*.
  - Provar el model en un entorn més gran per trobar els límits d'escalabilitat.

Torn de qüestions

Moltes gràcies per l'atenció!

Torn de qüestions

Qüestions?

escola d'enginyeria

UAB Universitat Autònoma de Barcelona

Torn de qüestions

Annexos

Planificació

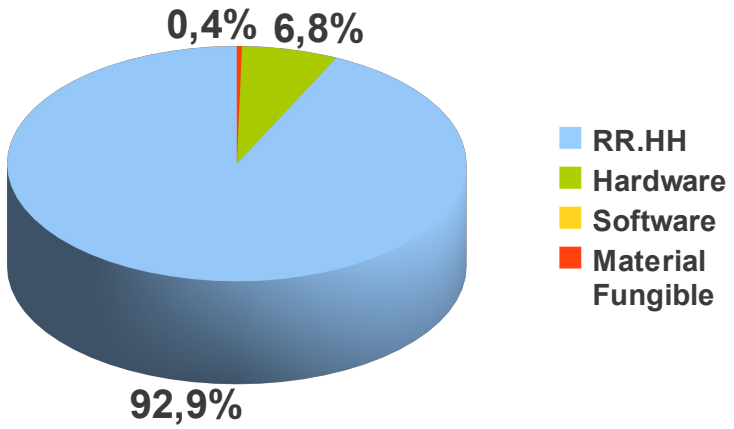
Etapa	Planificació inicial	Planificació real
Familiarització	**100**	**76**
Definició del projecte	12	12
Anàlisi i documentació inicial	32	24
Comprensió de COMPSs	56	40
Implementació	**432**	**444**
Gestió de rèpliques	56	52
Test gestió de rèpliques	36	40
Disseny del planificador	56	56
Gestió de l'històric	28	32
Tolerància a fallades	52	52
Càlcul de velocitat de xarxa	60	60
Mesura del temps d'espera en cua	48	52
Mesura del temps d'execució	28	32
Test planificador	68	68
Optimitzacions	**40**	**32**
Avaluació	**120**	**100**
Correcions finals	**24**	**32**
Memòria	**244**	**200**
Total	**960**	**890**

Reducció de 70 hores en l'elaboració → Aprox. **25 dies** abans.

Anàlisi econòmica

- Recursos humans: 50.887€
- Hardware utilitzat: 3.710€
- Material fungible: 200€
- Software utilitzat per:
  - Desenvolupar el projecte: gratuït
  - Desenvolupar la memòria: gratuït

Total: 54.797€



escola d'enginyeria

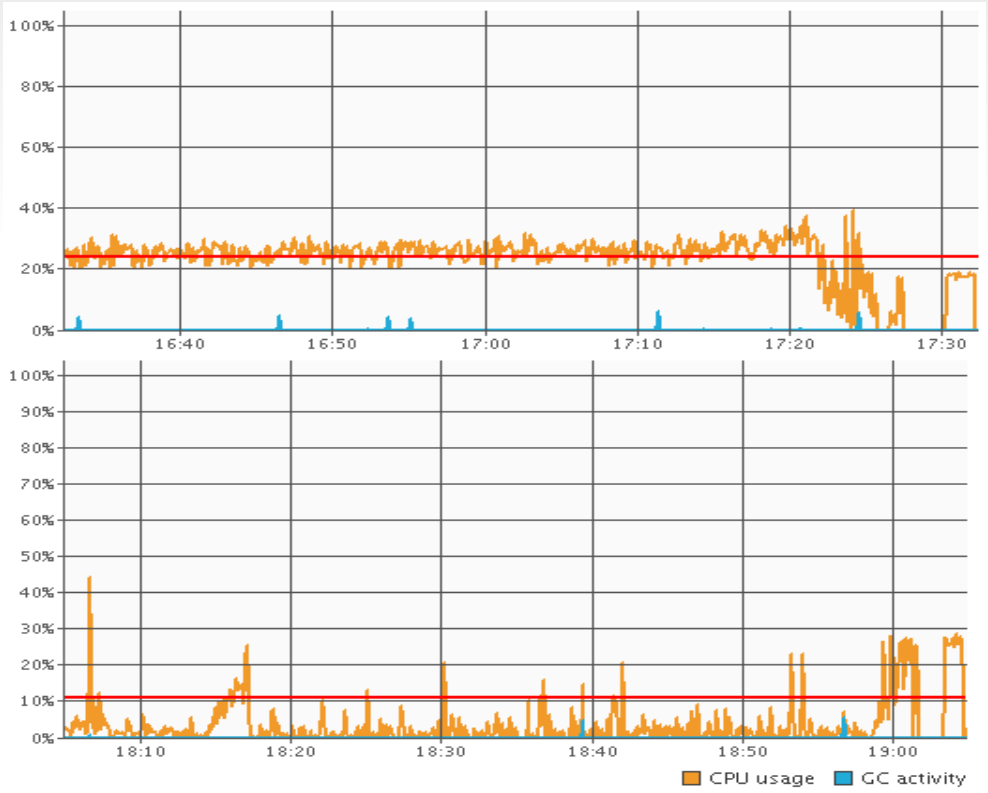
UAB Universitat Autònoma de Barcelona

Avaluació de resultats

- Anàlisi del consum de recursos:
  - **Experiment: execució amb 10 Workers de HMMER de 8192 seqüències.**
    - Experiment en ambdúes versions de COMPSs.
    - Monitorització del consum de CPU i memòria del node Master (runtime).

Avaluació de resultats

- Anàlisi del consum de recursos (CPU):



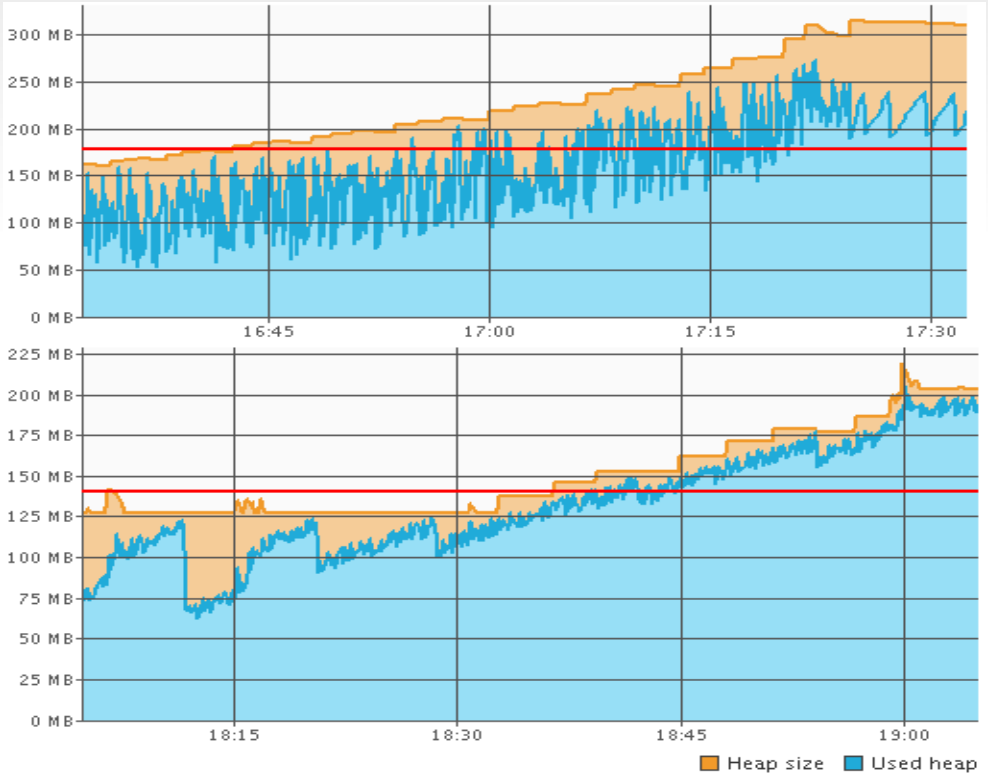
Consum mig CPU:
Original → 24%
Nou → 11%
Estalvi del 54%

escola d'enginyeria

UAB Universitat Autònoma de Barcelona

Avaluació de resultats

- Anàlisi del consum de recursos (Memòria):



Consum mig memòria:
Original → 184MB
Nou → 140.5MB
Estalvi del 24%

escola d'enginyeria

UAB Universitat Autònoma de Barcelona