Configurazioni ottimizzate per lo scheduling dei job

Stefano Dal Pra,
stefano.dalpra@cnaf.infn.it
Davide Salomoni,
davide.salomoni@cnaf.infn.it
Alessandro Italiano,
alessandro.italiano@cnaf.infn.it

DEFINIZIONE DEL PROBLEMA:

- Premessa: Il batch system LSF sceglie tra i WN (nodi) “disponibili” quello “meno carico”
  - *Disponibile*: "
  - “almeno uno slot libero” & “adeguato”
  - *Adeguato*: risorse sufficienti per il job da mandare in run
  - *Meno carico*: rispetto, es. a system load
- Si desidera condizionare la scelta del WN in base ad una o più caratteristiche C (note) del Job (es: coda|gruppo|<criterio X>).

Motivazioni

- Per alcune attività è desiderabile concentrare l'esecuzione su meno nodi possibile
  - es. exp. Auger al T1
- Ridurre il rischio che Job di un certo tipo mandino occasionalmente in blocco il nodo, causando la perdita di job “altrui”.
- Sfruttare un futuribile tag espresso dai job (CPU Intensive vs. IO intensive) – vedi report WM TEG
- Miglior supporto Job MPI

Motivazioni (2)

- WNoDeS può sfruttare queste possibilità per ottenere:
  - Virtual wn “uguali” nello stesso HV (minimo set di immagini copiate nel HV)
  - Generalizzazione di gestione di attributi dinamici non nativi al batch system
- Packing o no_packing di job che richiedono risorse comuni

Politiche di Packing

- **PACKING_RELAXED** (aggregazione):
  - Job J con le proprietà (C(J) == True) devono preferire nodi con loro simili già in esecuzione.
  - Nessuna restrizione di scelta per Job di altri tipi.
- **PACKING_EXCLUSIVE** (concentrazione):
  - Job tipo C devono preferire nodi che hanno altri job C in esecuzione (come sopra)
  - job di altro tipo devono evitare nodi con job C in esecuzione.
- **PACKING_NONE** (diffusione):
  - Job tipo C devono preferire nodi che NON hanno altri job C in esecuzione

Realizzazione (con LSF)

- Si tratta di modificare l'insieme dei nodi disponibili al momento del dispatching, introducendo dinamicamente condizioni sulla “adeguatezza” del nodo
- uso di elim (nei wn, per pubblicare “risorse”) ed esub (nel master LSF, per verificarne la presenza)
- Es. per packing relaxed:
  - bsub -R "packing_atlas == 1 || packing_atlas == 0"
- Si sfrutta la “short evaluation” delle espressioni booleane
- In ogni nodo elim pubblica un valore per atlas_packing

elim (esempio)

- [root@wn-xyz ~]# . elim_packing.sh

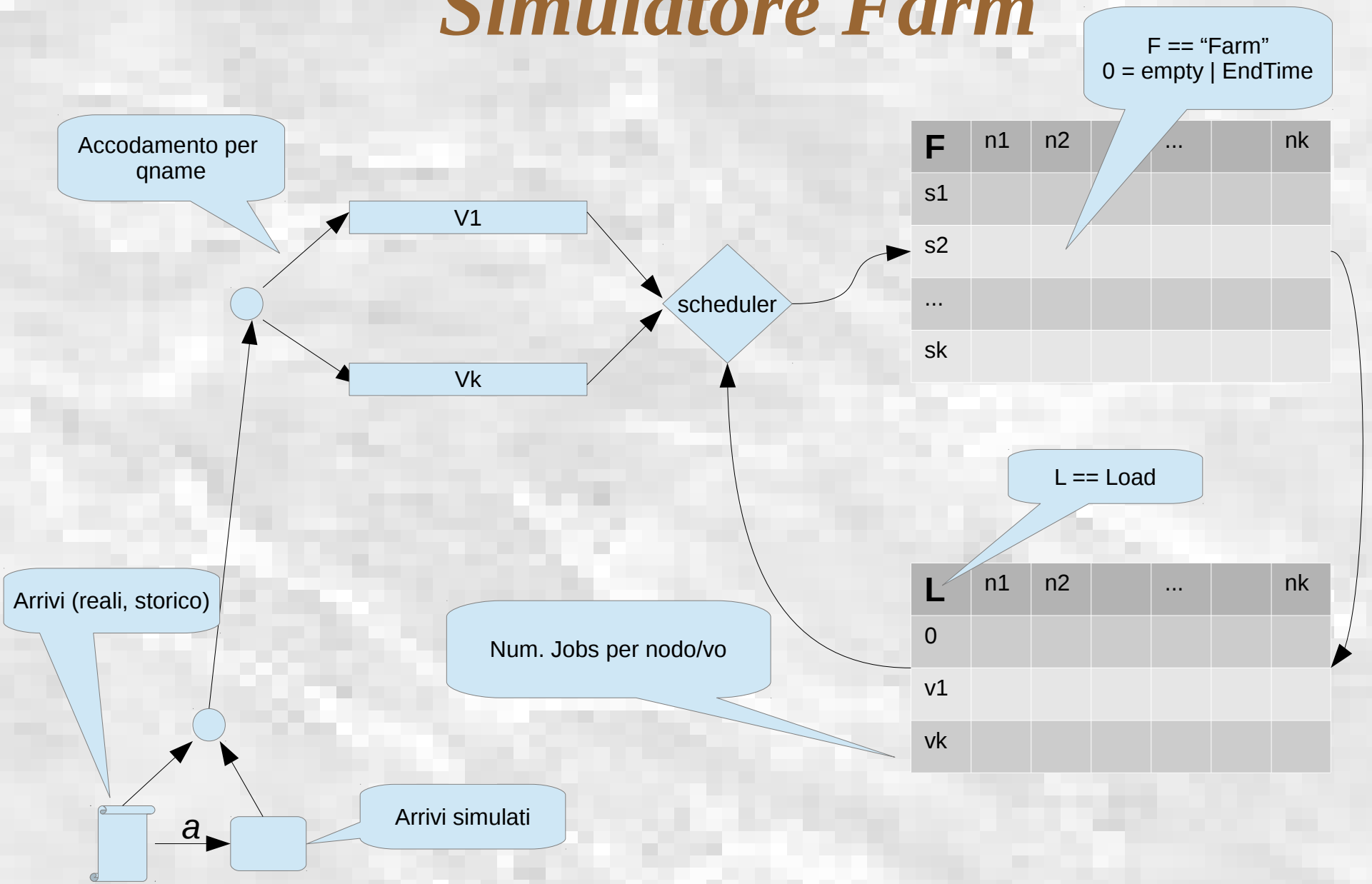
  4 2 packing_auger 3 packing_alicesgm 1 packing_superb 1 packing_atlasprd

- Ogni N secondi Elim pubblica in una linea:
  - n valore1 risorsa1 … valore_n risorsa_n
- Info ricavate via /bin/ps
  - ps -o pid --ppid `pidof sbatchd` #pid figli di sbatchd
  - ps -o group -p *pid1,...,pidn* *#gruppo di appartenenza*
- *Nb: info ricavate localmente nel WN*

Valutazione impatto sulla Farm

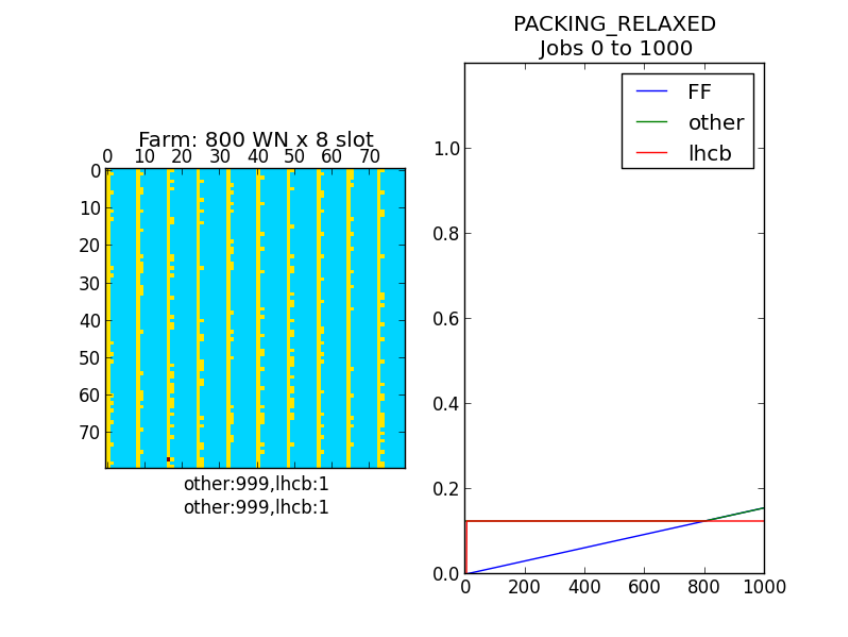
- E' stato realizzato un “semplice” simulatore, per valutare effetti e differenze tra politiche di packing.
- Ci aiutano nella valutazione due indicatori sintetici:
  - “Packing Index” (indice di concentrazione)
    - PI = Nodi_sufficienti / Nodi_usati
  - “Fill Factor”: (indice di saturazione)
    - FF = Slot in uso / Slot disponibili
- python, pylab (matplotlib, numpy)
- dati reali (tstart, tend, queuename; circa 2.5Y di storia, ~15Mrecord) o arrivi simulati (modellati su statistiche dallo storico)

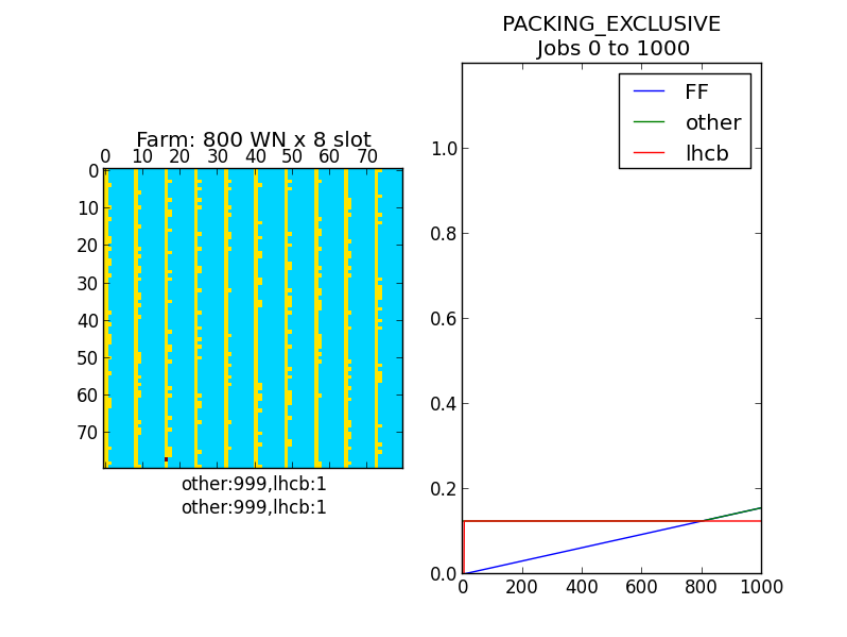
Simulatore Farm



Relaxed vs Exclusive, 1VO

Farm, FillFactor, dispersione



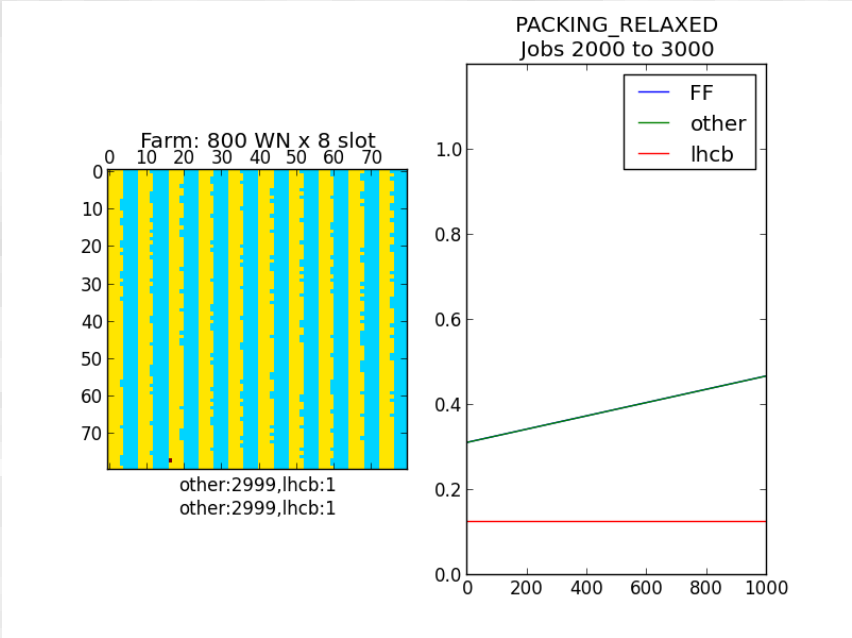


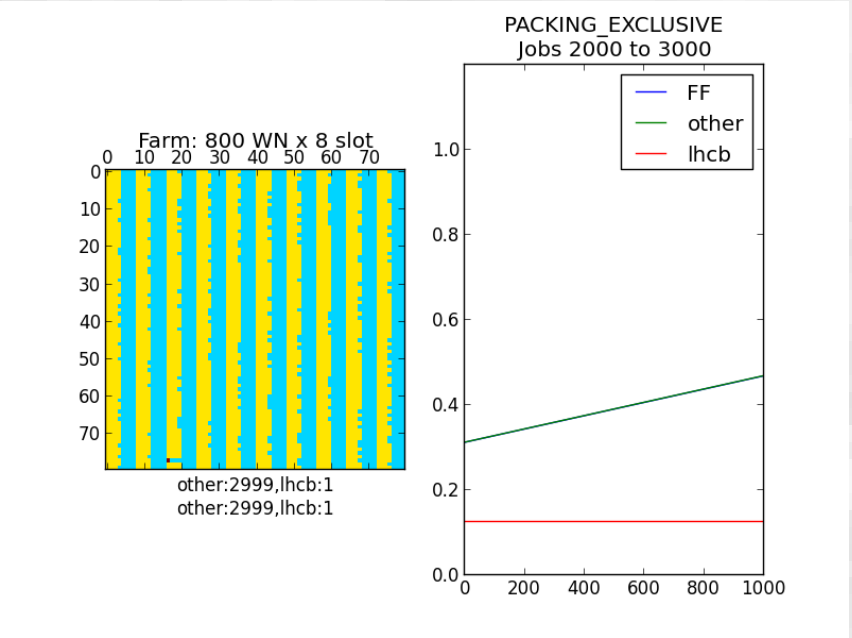
Partendo con Farm vuota Il comportamento è equivalente

FF: Fill Factor

Other: ogni job che non fa packing

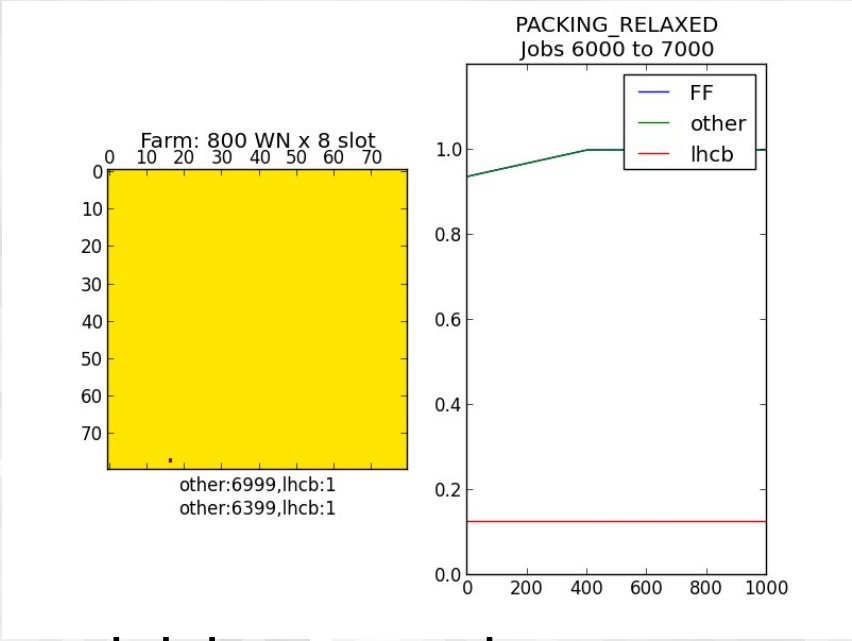
Relaxed vs Exclusive
lhcb

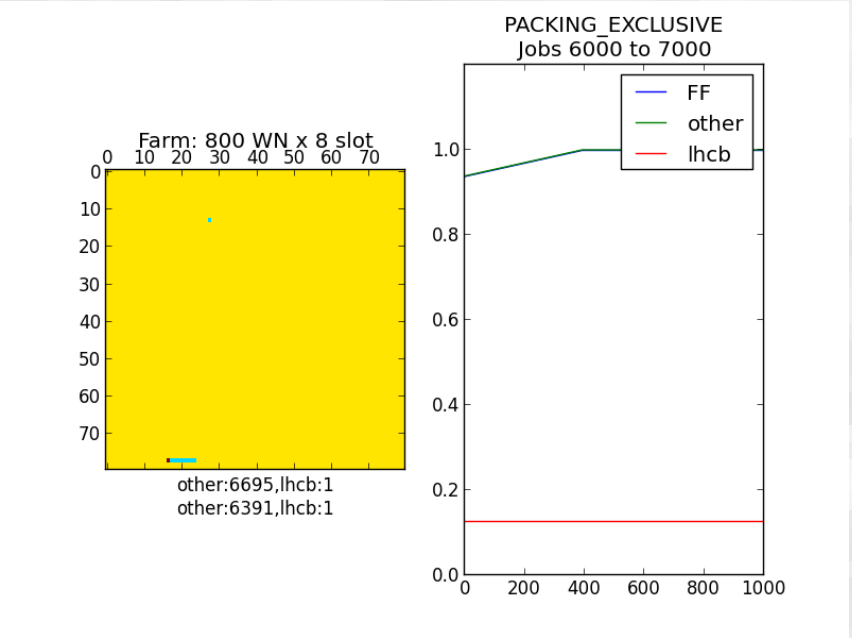




Relaxed: sul nodo con job LHCB si aggiungono altri job
Exclusive: il nodo con 1 job LHCB rimane riservato

Relaxed vs Exclusive
lhcb



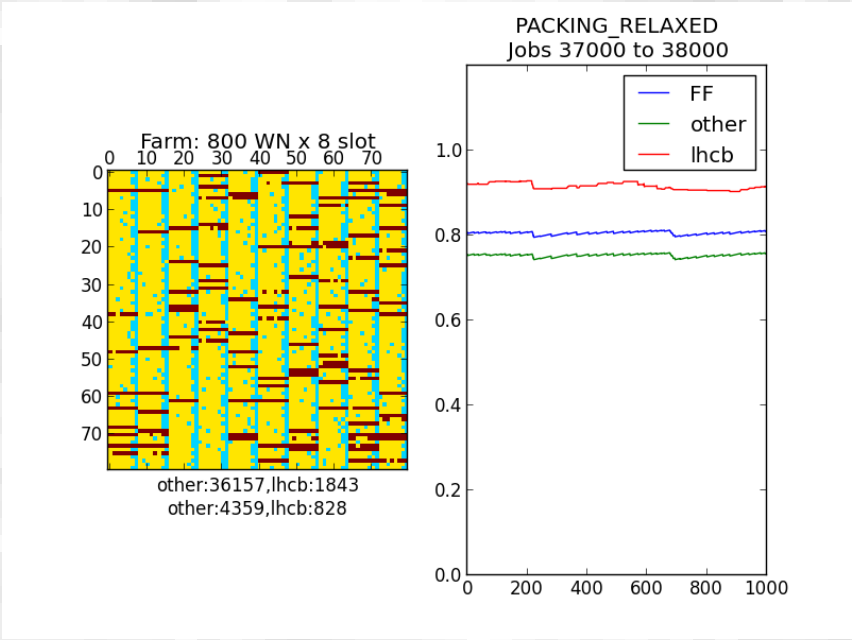


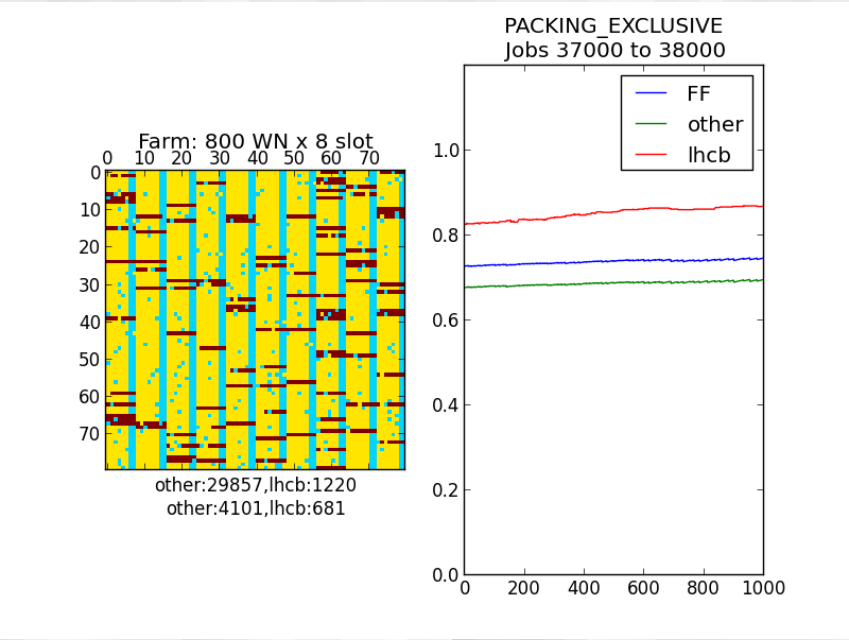
Farm, inizio saturazione:

Relaxed: Tutti gli slot occupati, un solo JP

Exclusive: il nodo con 1 job LHCB rimane riservato (abbiamo slot inutilizzati)

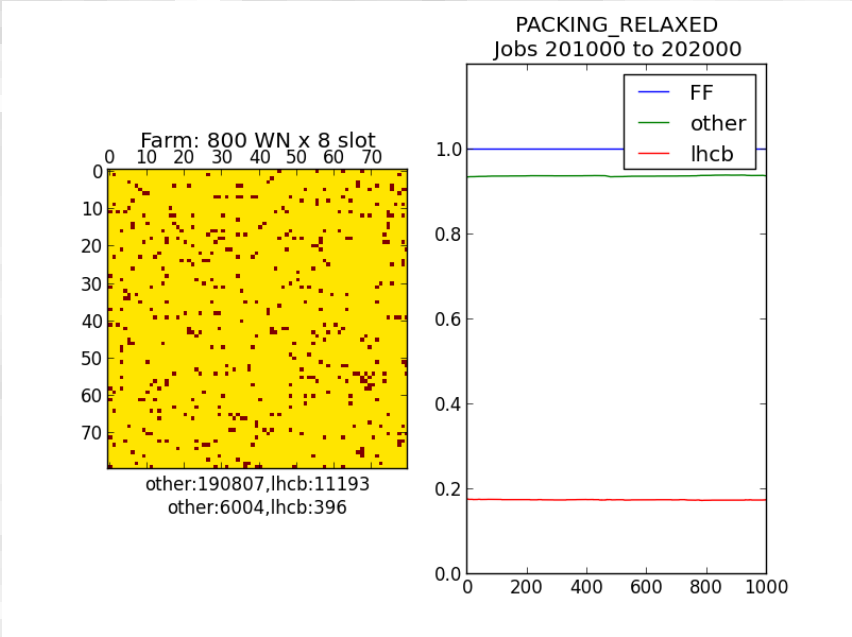
Relaxed vs Exclusive lhcb

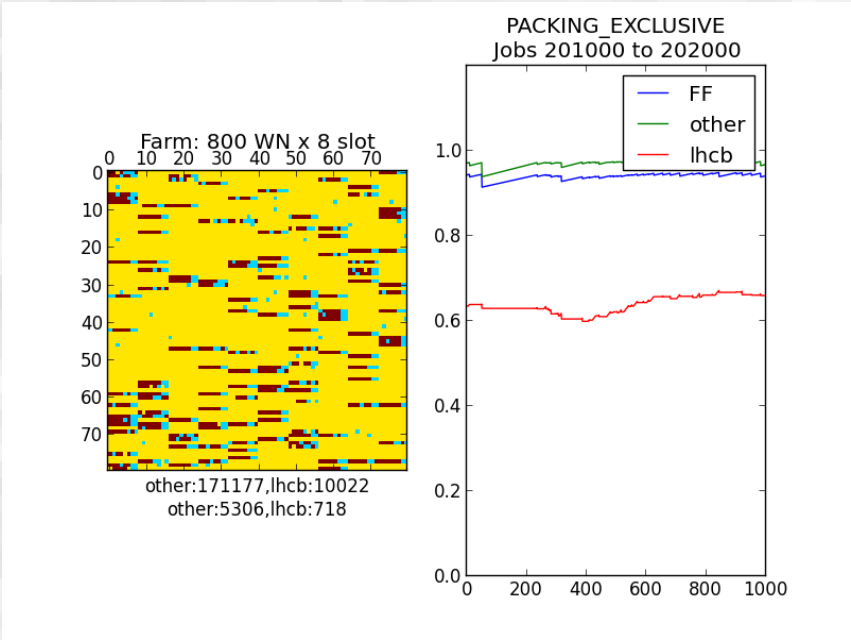




Farm, post saturazione:
Relaxed: leggermente migliore di Exclusive

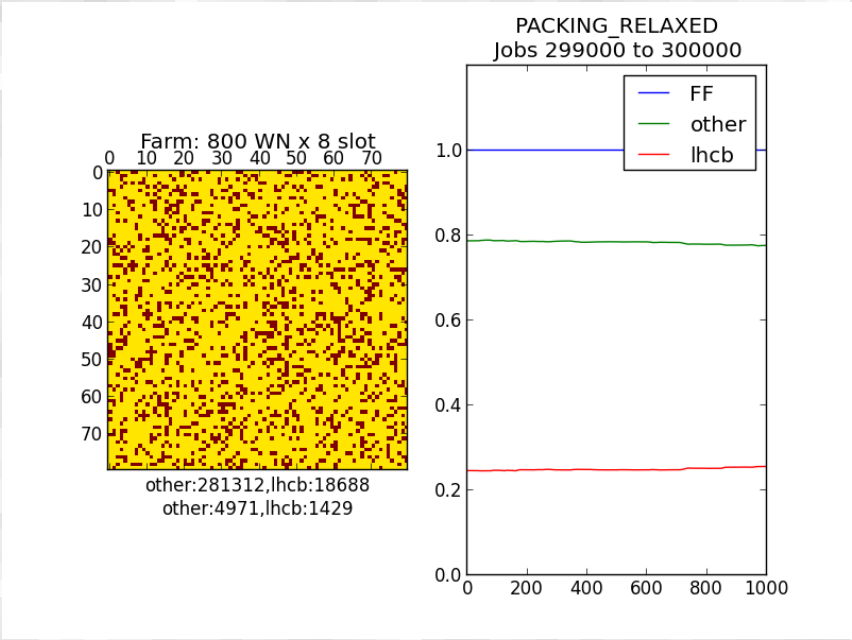
Relaxed vs Exclusive lhcb

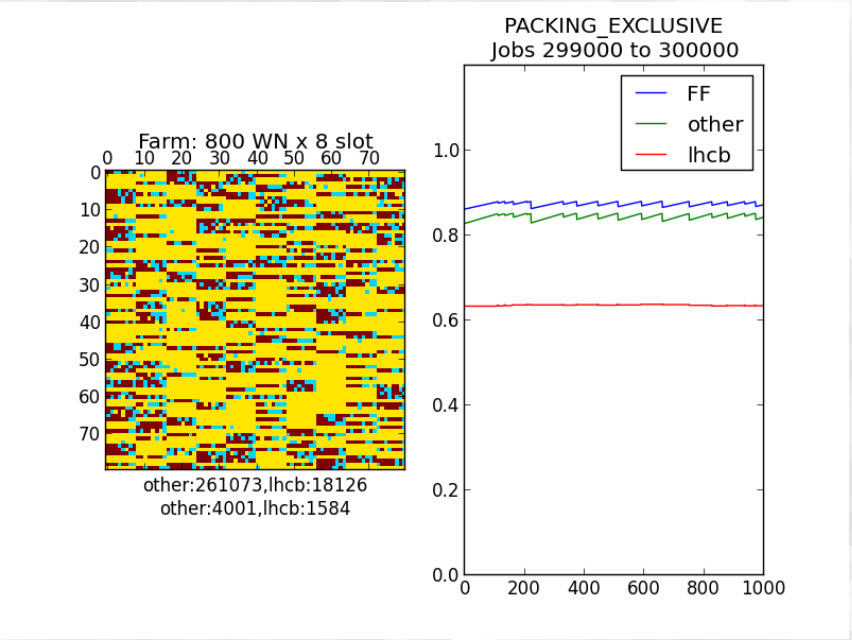




Farm, successiva saturazione:
Relaxed: JP molto dispersi
Exclusive: poca dispersione ma diversi slot, vuoti, --> Farm “piu' lenta”

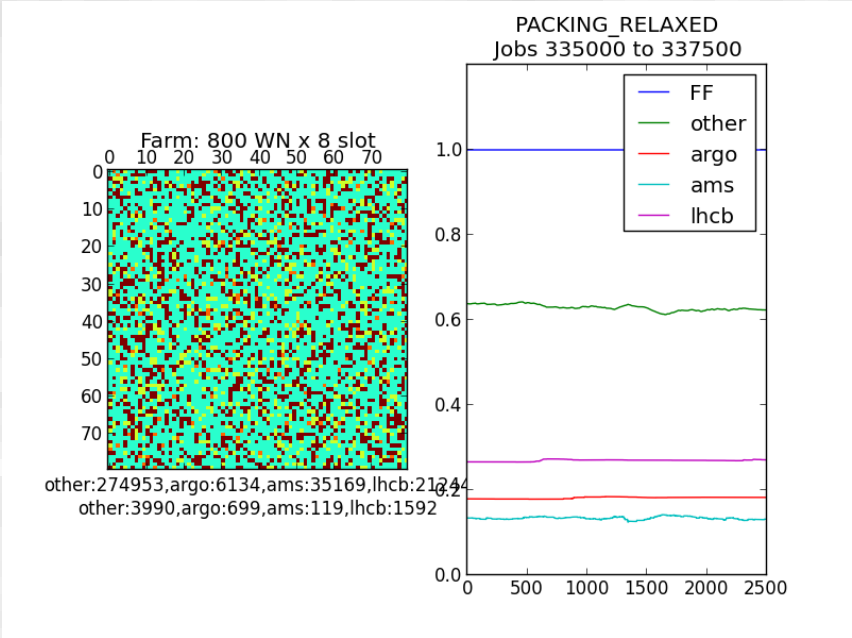
Relaxed vs Exclusive lhcb

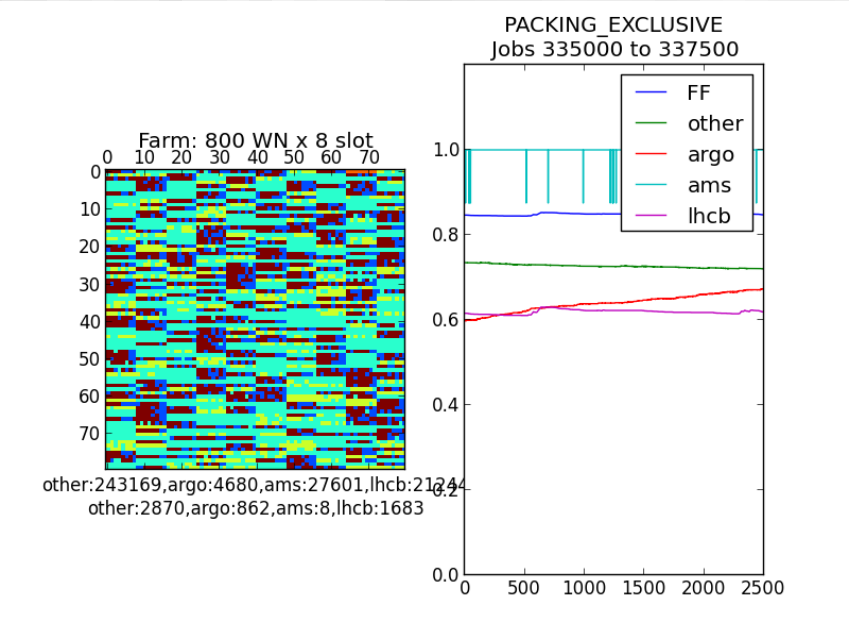




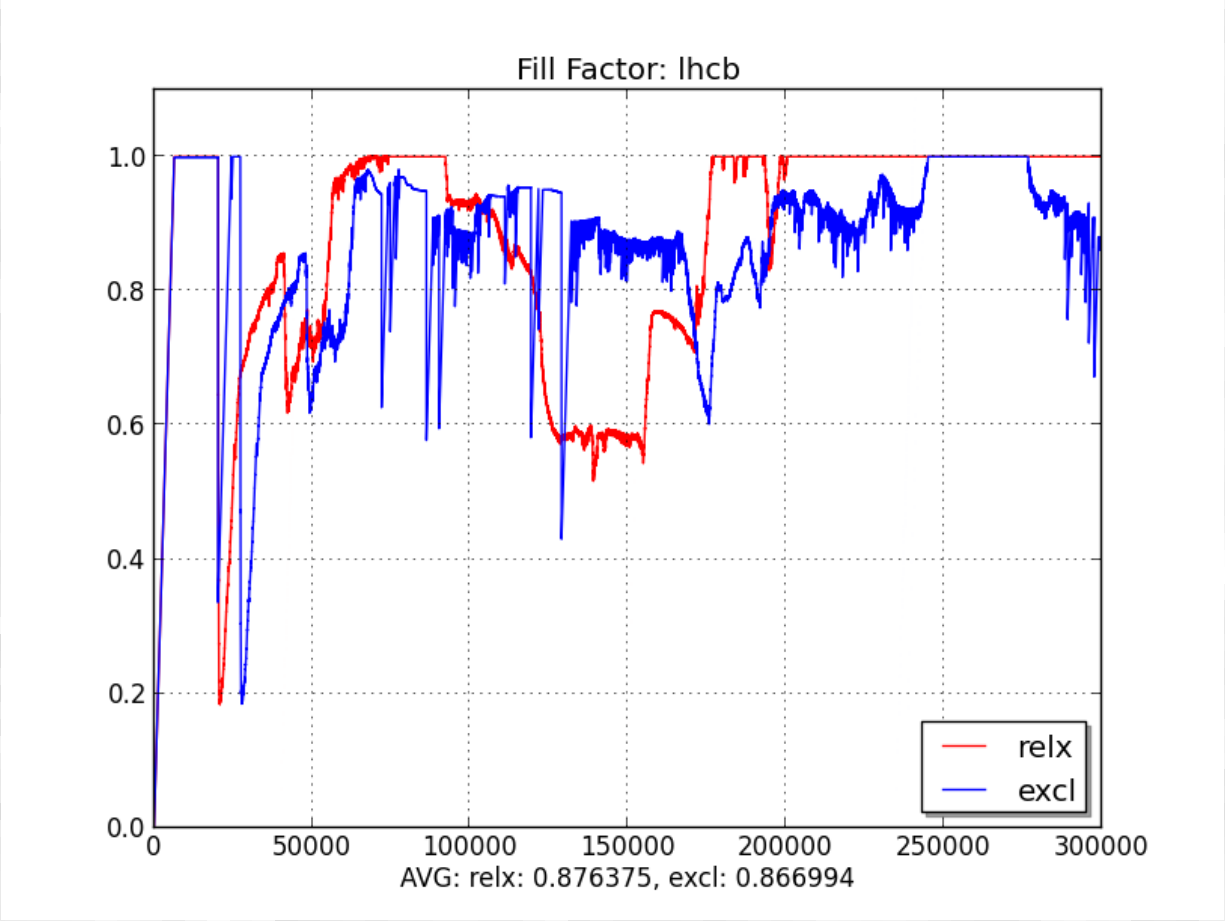
Farm, situazione “a regime”
Relaxed: Tutti gli slot occupati, un solo JP
Exclusive: il nodo con 1 job LHCB rimane riservato (abbiamo slot inutilizzati)

Relaxed vs Exclusive ams,argo,lhcb





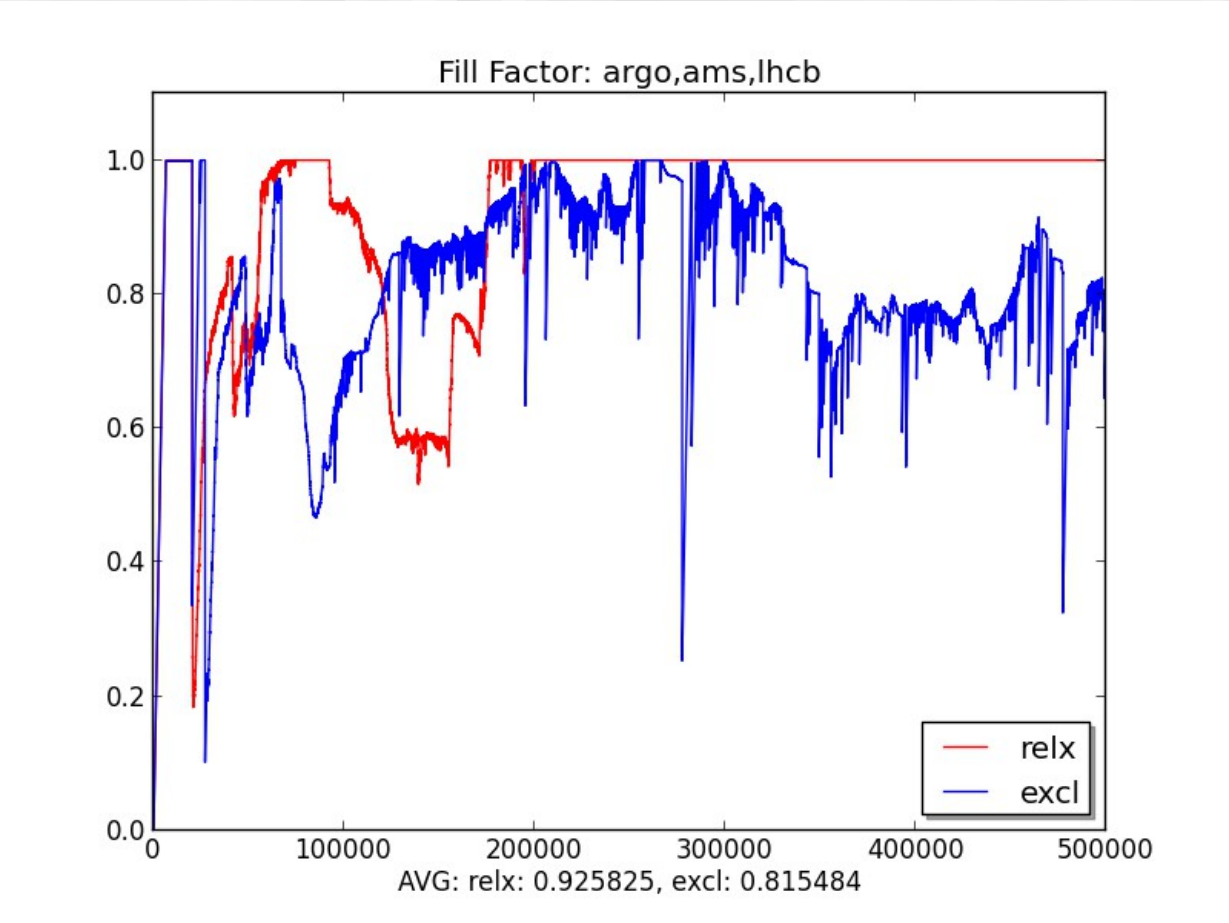
Fill Factor, lhcb



avg(ff_relx – ff_excl) = 0.0094 (~ 1%)

0.0094 x 800 * 8 = 60 ⇒ Il packing excl. “costa” 60 slot

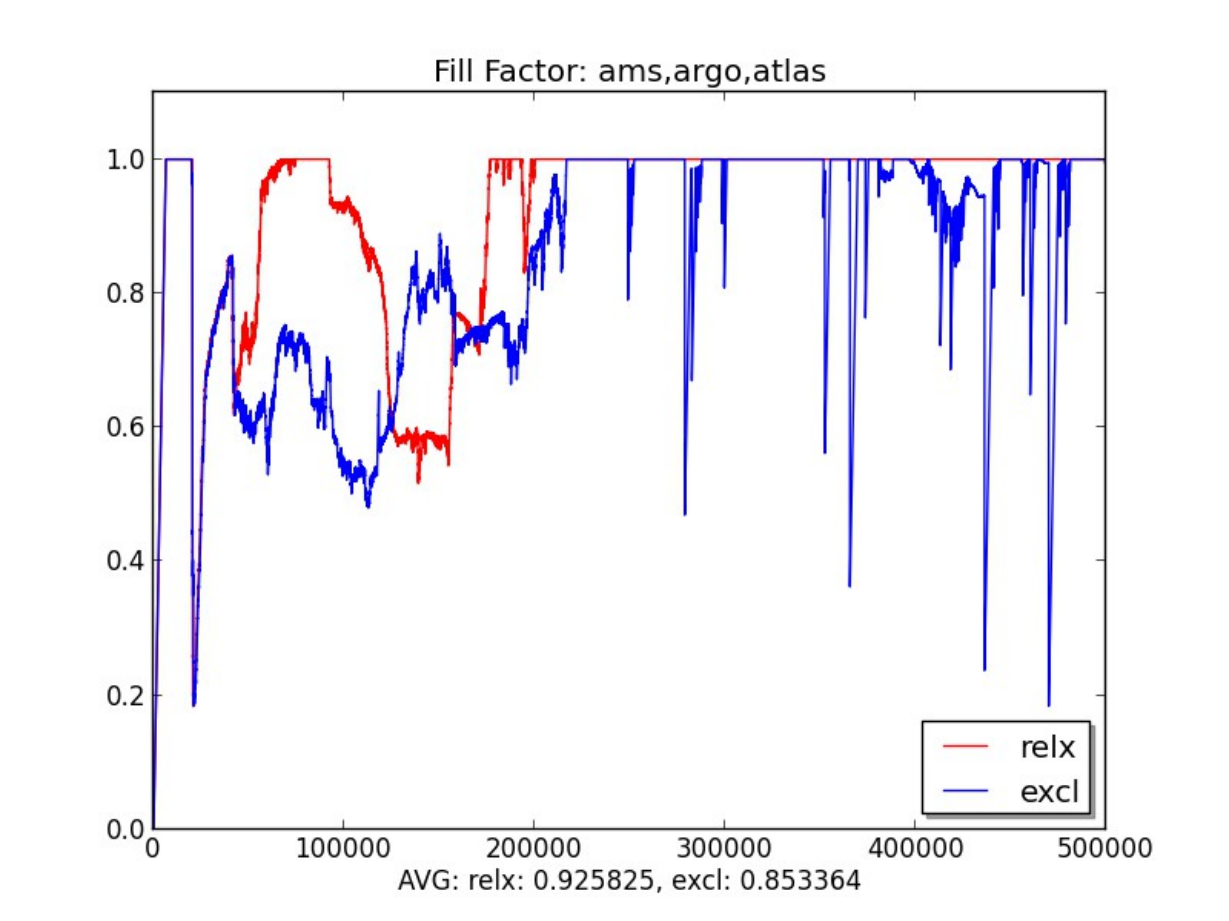
Fill Factor, ams,argo,lhcb



avg(ff_relx – ff_excl) = 0.110341 (~ 11%)

0.110341 x 800 * 8 = 706 ⇒ Il packing “costa” 700 slot

Fill Factor, ams,argo,lhcb



Nota: nella seconda metà il “costo” si riduce a 0.0438 x 800 * 8 = 280

Conclusioni

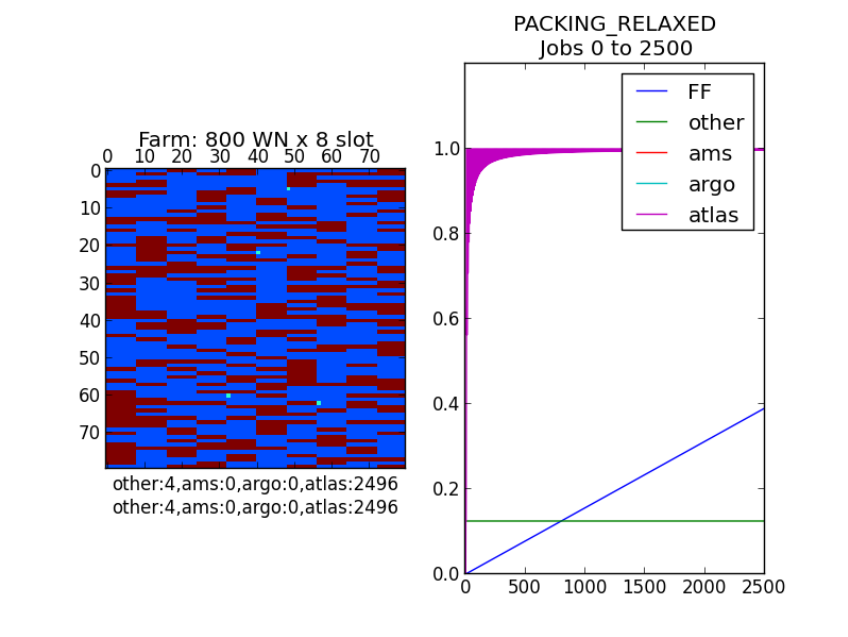
- L'ordine degli arrivi è importante:
  - in Exclusive mode può ridurre il FF, pur mantenendo buon PI.
    - Job molto lunghi possono peggiorare il FF
  - In Relaxed mode il PI si può deteriorare molto ma senza alcun effetto sul FF.
- Si può mediare tra Relaxed ed Exclusive introducendo un vincolo di “reservation Time To Live”.
- In questo modo Relaxed ed Exclusive diventano casi particolari:
  - Relaxed: TTL = 0 ; Exclusive: TTL $\rightarrow \infty$

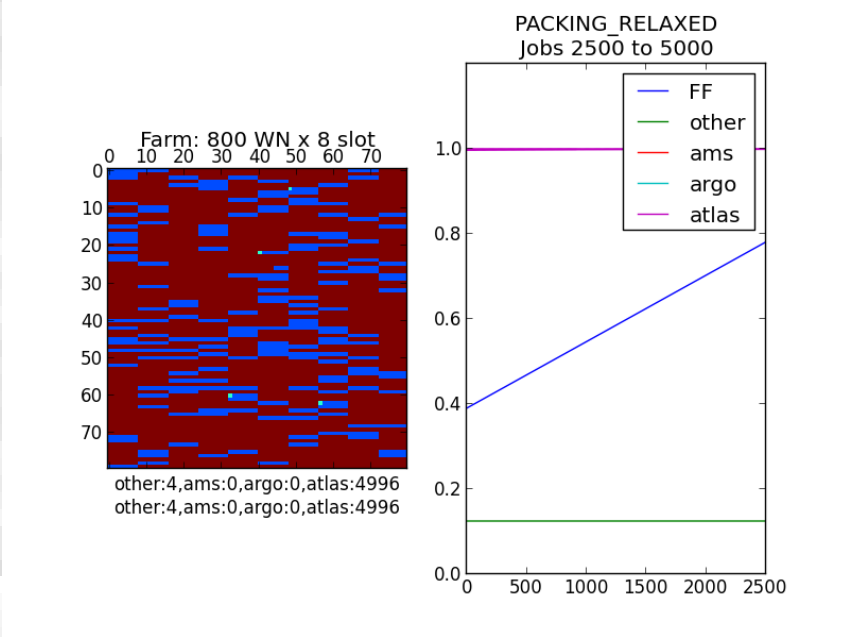
Conclusioni

- Le Packing Policies sono un caso particolare di gestione di attributi dinamici non nativi al batch system – un'area che pensiamo di sviluppare in futuro.
- Le simulazioni permettono di indagare diversi scenari:
  - Scelta della sequenza di arrivi (o da storico, o da modello statistico, o “self-made”)
  - Il simulatore mantiene una rappresentazione di stato relativamente completa, per cui anche altre grandezze possono essere tracciate e valutate (CPUTime, QueueTime etc.)

Backup slides

Relaxed





Ilpacking Relaxed può comportarsi molto bene, inizialmente (arrivi uniformemente dello stesso tipo).