Augmenter le débit d'un circuit numérique

Augmenter le débit
Sujets de ce thème

- Débit: définition et calcul
- Trois stratégies:
  - Réduire le délai
  - Utiliser le pipeline
  - Paralléliser les calculs
- Combiner les stratégies

Débit d'un circuit numérique

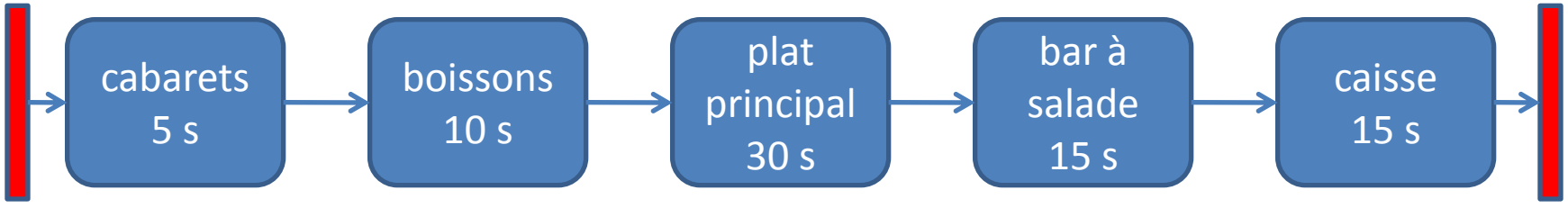
- Le *débit d'information* (*throughput*) est le nombre de résultats produits par unité de temps.
  - On suppose qu'une quantité suffisante de données est disponible à l'entrée du système pour le garder toujours actif.
  - Grand débit = meilleure performance.
  - Le débit est exprimé en nombre de résultats par seconde.
  - Une fréquence d'horloge plus élevée correspond à un débit plus grand.
  - Le parallélisme augmente le débit.

- Débit =

  fréquence d'horloge

  × #résultats produits par cycle

  × #unités de calcul en parallèle



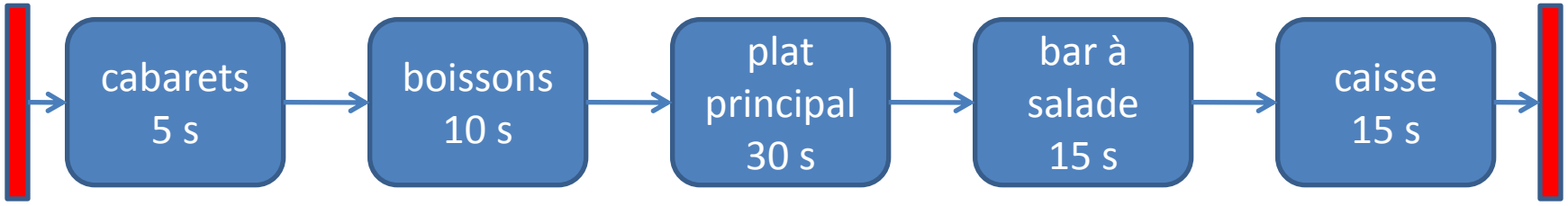
Calcul du débit d'un système

- Considérons une cafétéria avec 5 stations.
- Supposons:
  - il y a un seul client dans la ligne à la fois
  - chaque client passe par chaque station
  - un client ne peut pas prendre son cabaret tant que le client précédent n'a pas fini à la caisse

- Période = 75 s; fréquence = 13.3 mHz
- Latence:

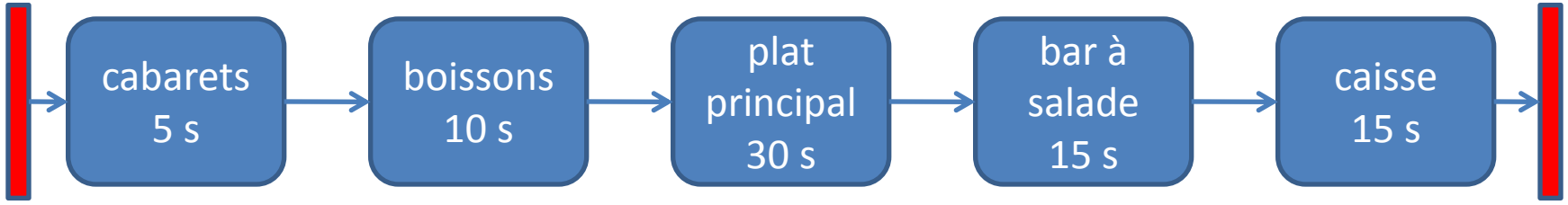
  1 cycle = 75 secondes pour servir un client
- Débit:

  1 client / 75 s × 3600 s/h = 48 clients par heure

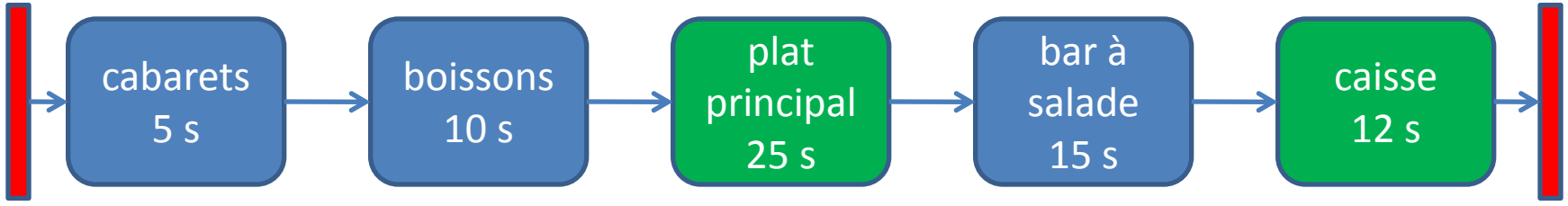


Réduire le délai sur le chemin critique

- Débit = fréquence d'horloge × #résultats produits par cycle × #unités de calcul en parallèle
- On peut augmenter le débit en augmentant la fréquence d'horloge, c'est à dire en réduisant le délai sur le chemin critique.
- Il est parfois possible de modifier une partie d'un circuit pour en réduire le délai.

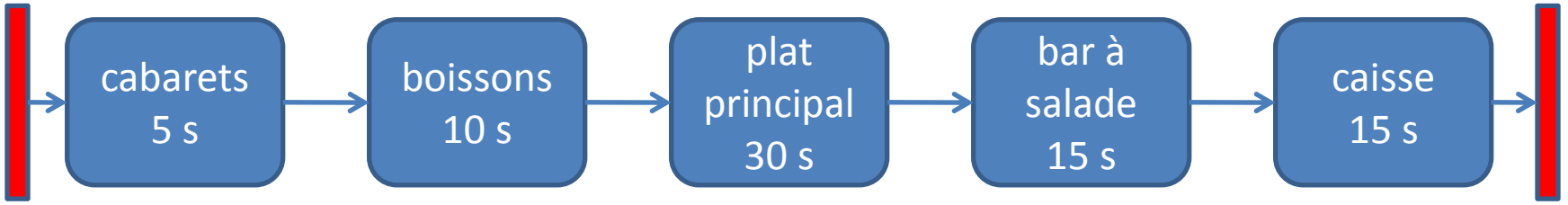


Cafétéria originale: période d'horloge 75 s, latence 1 cycle, débit 48 clients par heure

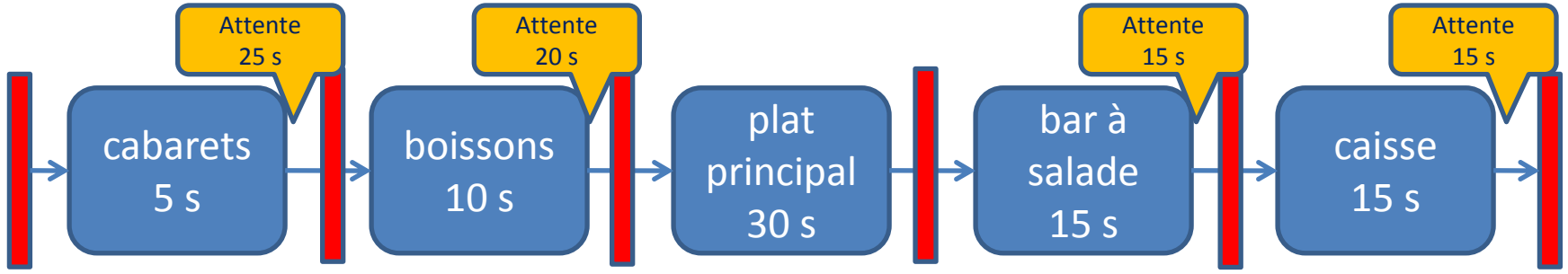


Cafétéria améliorée: période d'horloge 67 s, latence 1 cycle, débit 53.7 clients par heure

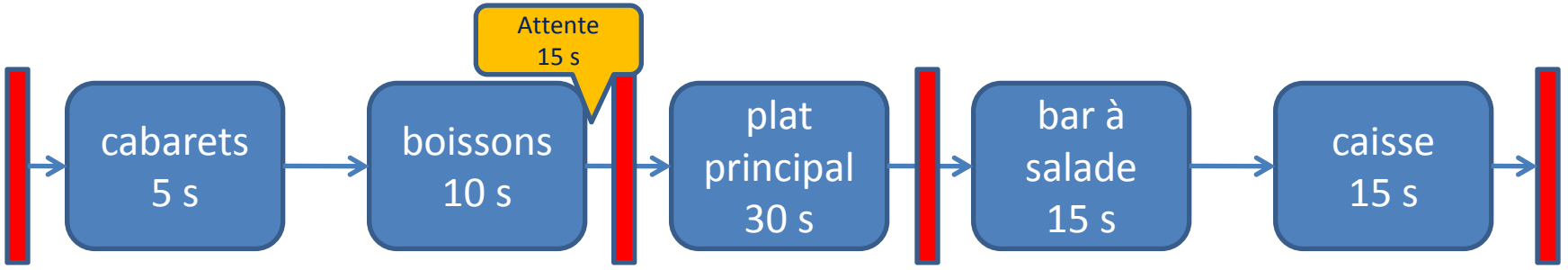
Briser le chemin critique avec des registres de pipeline



Original:
période d'horloge 75 s
latence 1 cycle = 75 s
débit 48 clients par heure



Pipeline naïf:
1 client par station
période d'horloge 30 s
latence 5 cycles = 150 s
débit 120 clients par heure

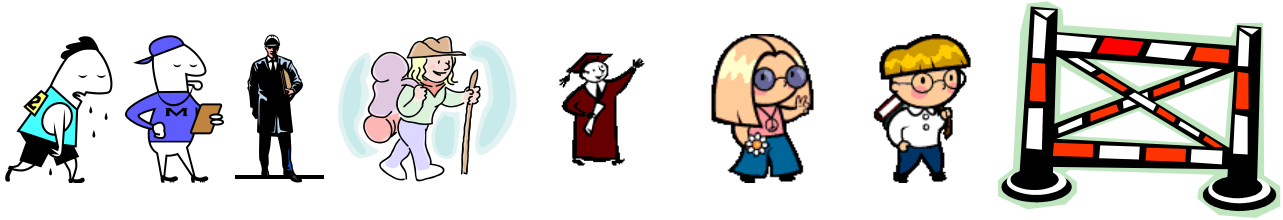


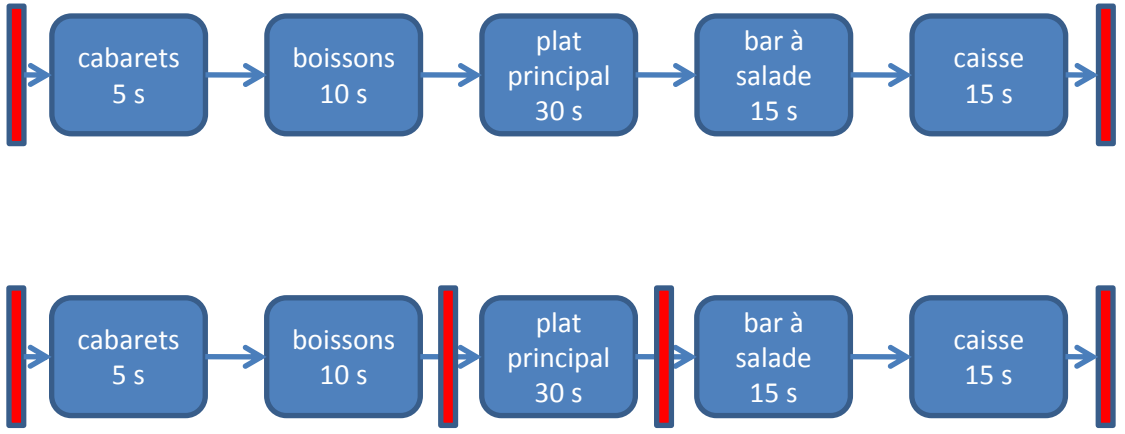
Pipeline ajusté:
1 client par station
période d'horloge 30 s
latence 3 cycles = 90 s
débit 120 clients par heure

Exercice

- En supposant une très longue file d'attente qui ouvre à 12h00:
  - À quelle heure mangera le premier client, avec et sans pipeline?
  - À quelle heure mangera le 2e client, avec et sans pipeline?
  - À quelle heure mangera le 3e client, avec et sans pipeline?
  - À quelle heure mangera le 4e client, avec et sans pipeline?

ARRÊTEZ LA VIDÉO ET FAITES L'EXERCICE!





Exercice

Pas de pipeline

heure	dans la file	manger
12:00:00		
12:01:15		
12:02:30		
12:03:45		
12:05:00		
12:06:15		
12:07:30		

Avec pipeline

heure	cabaret et boisson	plat	salade et caisse	manger
12:00:00				
12:00:30				
12:01:00				
12:01:30				 
12:02:00				 
12:02:30				 
12:03:00				 

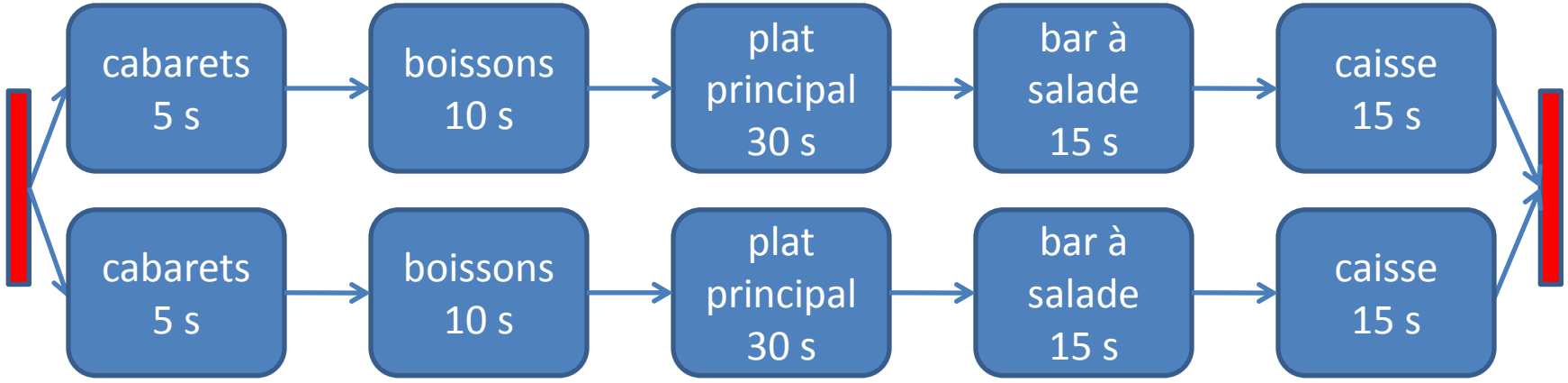
Observations sur le pipeline

- Le pipeline augmente le débit significativement:
  - de 48 à 120 clients par heure.
- Pour le premier client, le pipeline n'est pas intéressant, il perd:
  - 75 secondes dans le cas du pipeline naïf;
  - 15 secondes dans le cas du pipeline ajusté.
- Pour le 3^{e} client, le pipeline est très intéressant:
  - il pourra manger après 150 s (cinq coups d'horloge) au lieu de 225 s pour le cas sans pipeline.
- Le pipeline n'est intéressant que s'il y a suffisamment de clients pour le remplir.

- Pour plusieurs applications impliquant un flux de données, une latence supplémentaire est sans importance.
- Exemple d'un lecteur audio:
  - taux de données de 44.1 KHz
  - latence de 10 cycles d'horloge $< 230\ \mu s$
  - imperceptible pour un être humain qui vient de lancer la lecture d'un fichier de musique.
- Pour certaines applications, une latence supplémentaire est à éviter.
  - communications bidirectionnelles;
  - transactions financières en bourse;
  - recherche web.

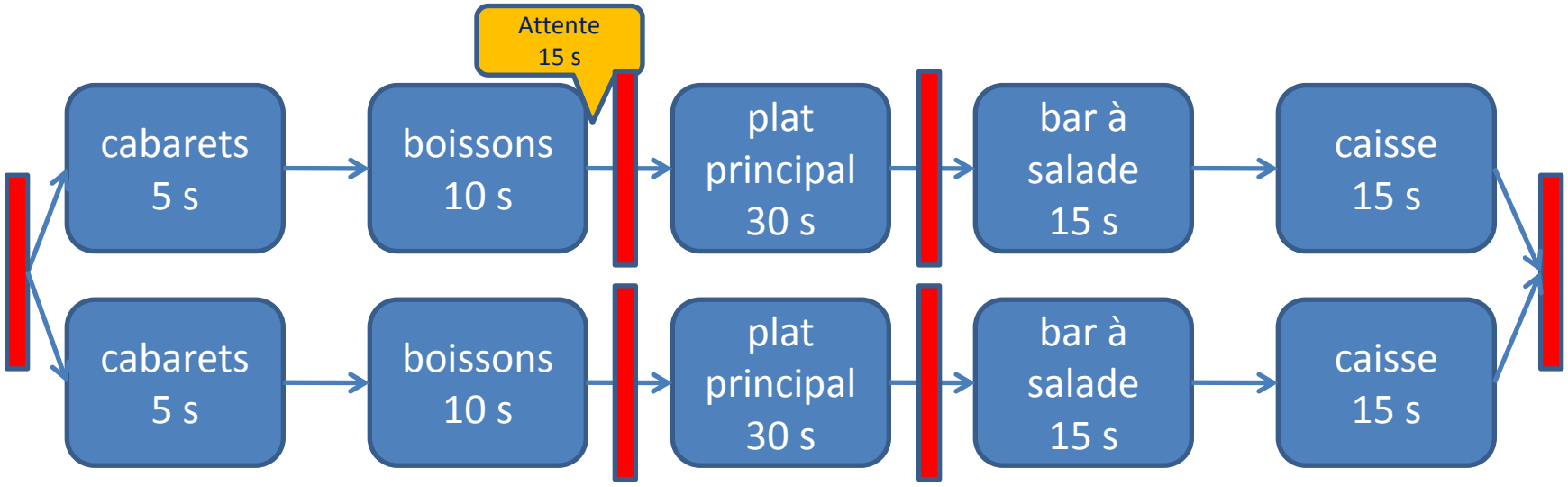
Paralléliser les calculs

- Débit = fréquence d'horloge × #résultats produits par cycle × #unités de calcul en parallèle
- Instancier plusieurs unités de traitement permet de traiter plusieurs données en parallèle et augmenter le débit.
- Exemple: pour un processeur vidéo, on peut décomposer une image en régions indépendantes et associer le traitement de chaque région à un processeur différent.



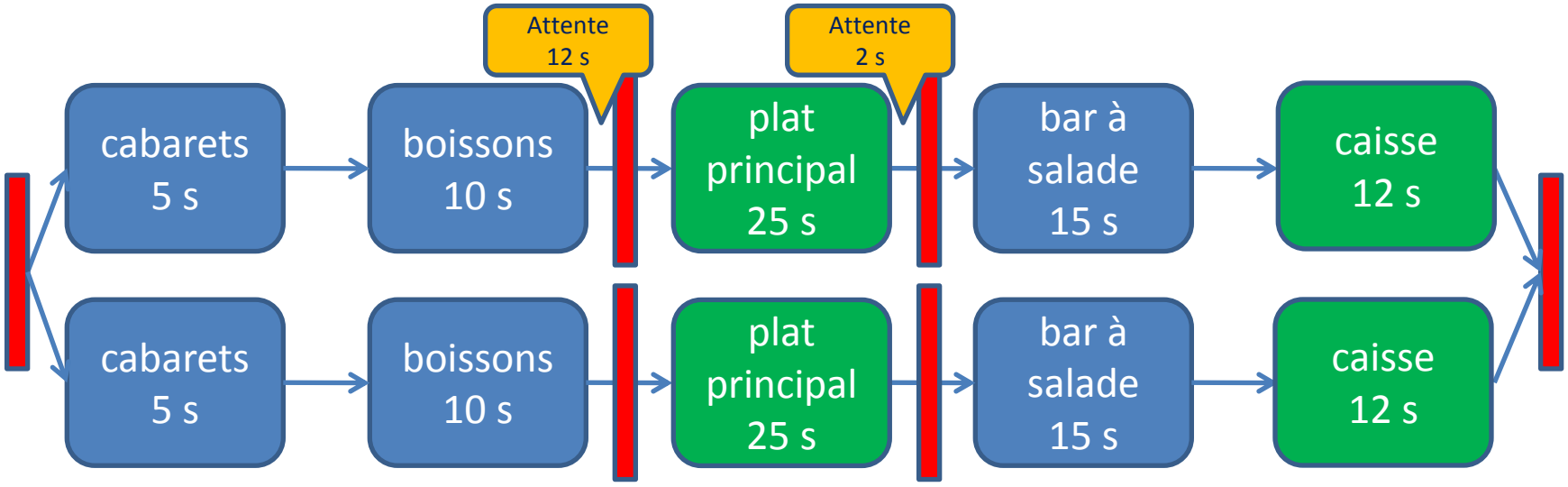
Période d'horloge 75 s, latence 1 cycle, 2 clients par cycle,
débit 96 clients/heure

Pipeline et parallélisation



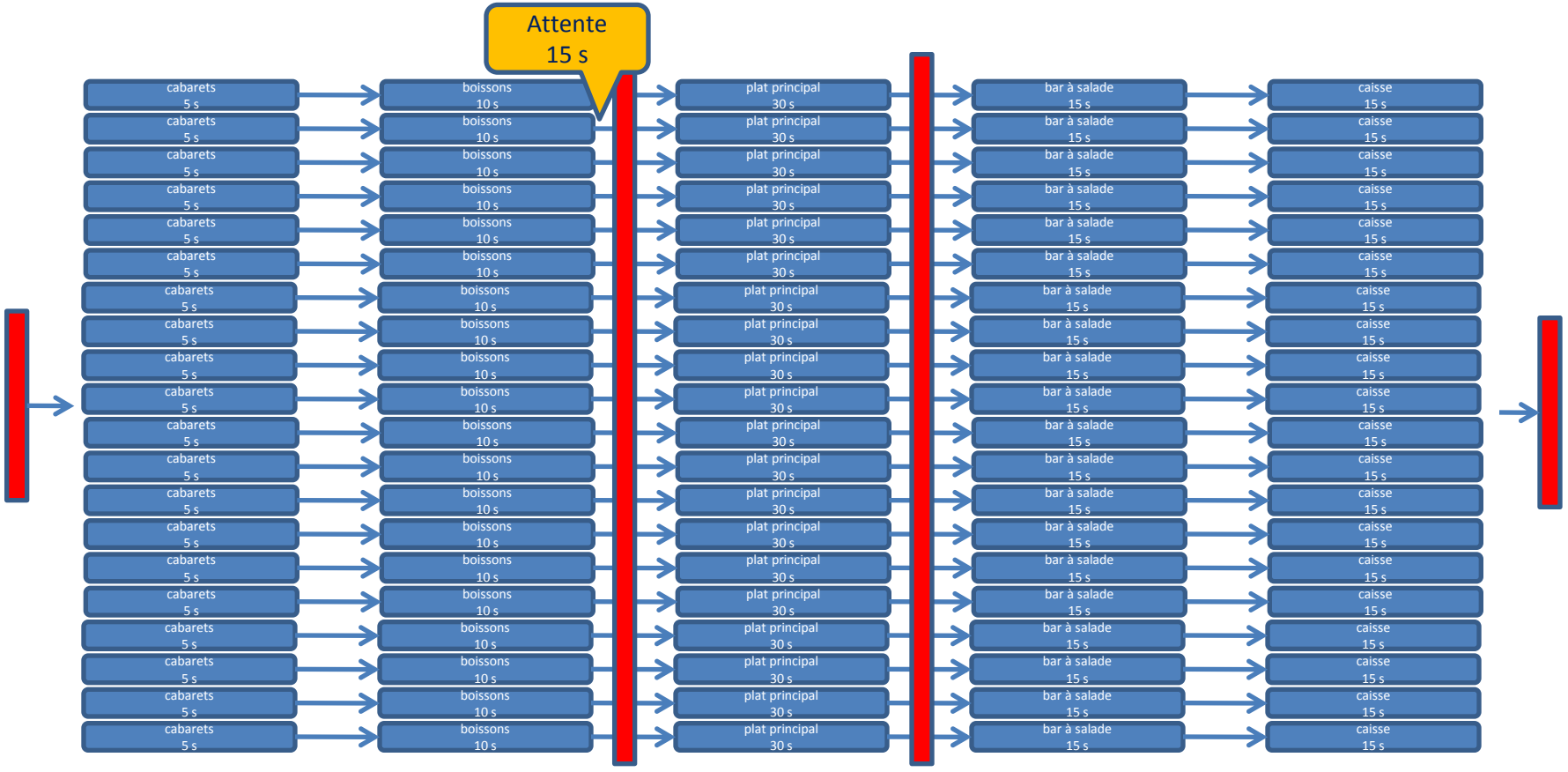
Période d'horloge 30 s, latence 3 cycles, 2 clients par cycle,
débit 240 clients/heure

Réduction du délai, pipeline et parallélisation



Période d'horloge 27 s, latence 3 cycles, 2 clients par cycle,
débit 266.7 clients/heure

Le cas FPGA extrême



Période d'horloge 30 s, latence 3 cycles, 20 clients par cycle,
débit 2400 clients par heure

Vous devriez maintenant être capable de …

- Expliquer les concepts de latence de calcul et de débit d'information, et comment ils sont reliés au délai du chemin critique, à la fréquence d'horloge et au parallélisme. (B2)
- Calculer la latence et le débit d'un circuit numérique. (B3)
- Appliquer les stratégies de la réduction du délai, du pipeline et de la parallélisation des calculs pour augmenter le débit d'un circuit numérique. (B3)

Code	Niveau (http://fr.wikipedia.org/wiki/Taxonomie_de_Bloom)
B1	Connaissance – mémoriser de l'information.
B2	Compréhension – interpréter l'information.
B3	Application – confronter les connaissances à des cas pratiques simples.
B4	Analyse – décomposer un problème, cas pratiques plus complexes.
B5	Synthèse – expression personnelle, cas pratiques plus complexes.