

Ensino de Software Pipelining e Escalonamento em GPUs com Python no Google Colab

Ricardo Ferreira, José A. Nacif Universidade Federal de Viçosa







Sumário

Python, Jupyter Notebooks e Google Colab

Software Pipelining

Escalonamento em GPU

Considerações Finais

Python

• ipython = Interatividade

• Interesse dos estudantes de Computação

Muitos exemplos, facilidades de prototipação, Chatgpt, etc....



Google Colab

• Documentação (texto+gráfica), Código, Interatividade

Acesso e Configuração fácil, Gratuito e Colaborativo

Rápido ciclo de desenvolvimento e teste



Software Pipelining

 Ferramentas: Unrolling e Escalonamento Dinâmico...porém não existem ferramentas para ensino de Software Pipelining

• Complexidade para validação e inversão do código

Desafios de balanceamento

$$v[i] = (v[i] + v[i])^2$$

Codigo	F	D	E	W	F	D	E	W
Ld f3,0(r1)	1	2	34	5	15	16	1718	19
add f3,f3,f3	2	34	567	8	16	1718	19-21	22
mult f3,f3,f3	34	567	8-12	13				
sd f3,0(r1)	567	8-12	1314	-				
Addi r1,r1,8	8-12	13	14	15				
Beq r1,r2,LOOP	13	14						

$$v[i] = (v[i] + v[i])^2$$

Codigo	F	D	E	W	F	D	E	W
Ld f3,0(r1)	1	2	34	5	15	16	1718	19
add f3,f3,f3	2	34	567	8	16	1718	19-21	22
mult f3,f3,f3	34	567	8-12	13				
sd f3,0(r1)	567	8-12	1314	-				
Addi r1,r1,8	8-12	13	14	15				
Beq r1,r2,LOOP	13	14						

$$v[i] = (v[i] + v[i])^2$$

Codigo	F	D	E	W	F	D	E	W
Ld f3,0(r1)	1	2	34	5	15	16	1718	19
add f3,f3,f3	2	34	567	8	16	1718	19-21	22
mult f3,f3,f3	34	567	8-12	13				
sd f3,0(r1)	567	8-12	1314	=				
Addi r1,r1,8	8-12	13	14	15				
Beq r1,r2,LOOP	13	14						

$$v[i] = (v[i] + v[i])^2$$

Codigo	F	D	E	W	F	D	E	W
Ld f3,0(r1)	1	2	34	5	15	16	1718	19
add f3,f3,f3	2	34	567	8	16	1718	19-21	22
mult f3,f3,f3	34	567	8-12	13				
sd f3,0(r1)	567	8-12	1314	-				
Addi r1,r1,8	8-12	13	14	15				
Beq r1,r2,LOOP	13	14						

2 ciclos para Load3 ciclos para Add5 ciclos para Mult



Loop Original

Software Pipelining

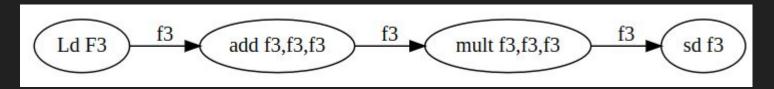
Codigo	F	D	E	W	F	D	E	W
Ld f3,0(r1)	1	2	34	5	15	16	1718	19
add f3,f3,f3	2	34	567	8	16	1718	19-21	22
mult f3,f3,f3	34	567	8-12	13				
sd f3,0(r1)	567	8-12	1314	-				
Addi r1,r1,8	8-12	13	14	15				
Beq r1,r2,L00P	13	14						

Codigo	F	D	E	w	F	D	E	W
sd f3,0(r1)	1	2	34	-	9	10	1112	58
mult f3,f4,f4	2	3	4-8	9	10	11	12-16	17
add f4,f5,f5	3	4	567	8	11	12	13-15	16
Addi r1,r1,8	4	5	6	7	12	13	14	15
Ld f5,0(r1)	5	67	89	10	13	1415	1617	18
Beq r1,r2,LOOP	67	8						

14 ciclos

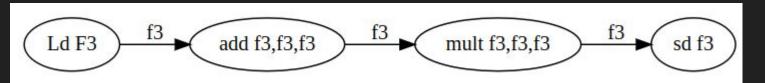
7 ciclos

Software Pipelining



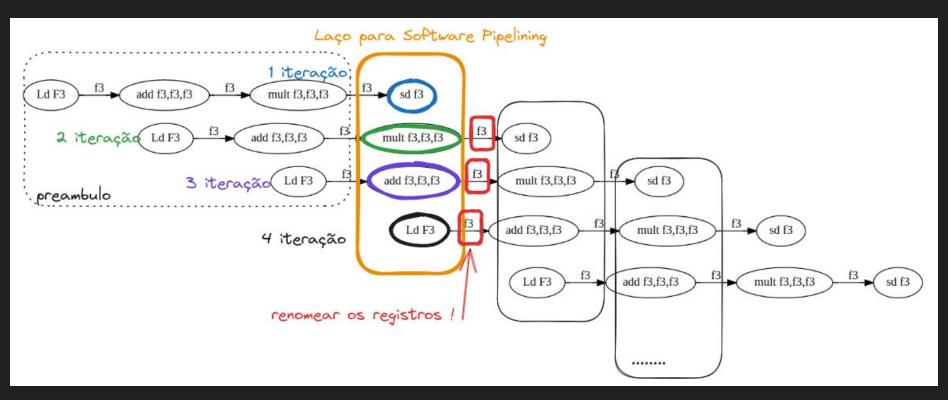
- 1. Executar em Python -> Resultado
- 2. Grafo de dependência
- 3. Renomear Registradores
- 4. Preâmbulo e Epílogo
- 5. Balancear Caminhos
- 6. Validar a execução em Python

Execução do loop original em Python



```
10 memory = [0,1,2,3,4,5,6,7,8,9,10,11,12,13,14,15,16,17,1]
11
12 for i in range(20):
     f3 = memory[i]
14
   f3 = f3+f3
  f3 = f3*f3
16
     memory[i] = f3
17
18 print(memory)
[0, 4, 16, 36, 64, 100, 144, 196, 256, 324, 400, 484, 576,
```

Construção do Software Pipelining com grafo



Original

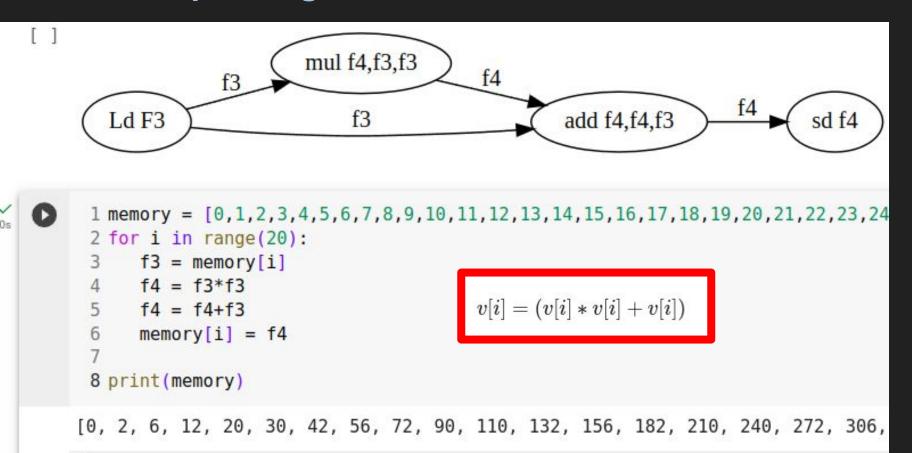
Software Pipelining



```
10 memory = [0,1,2,3,4,5,6,7,8,9,10,
11
12 for i in range(20):
13    f3 = memory[i]
14    f3 = f3+f3
15    f3 = f3*f3
16    memory[i] = f3
17
18 print(memory)
[0, 4, 16, 36, 64, 100, 144, 196, 25]
```

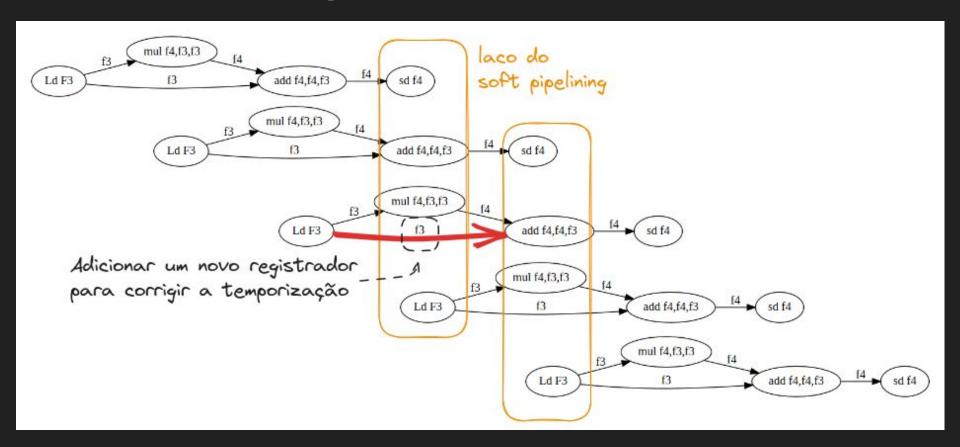
```
12 f5= memory[0]
13 f4 = f5 + f5
14 f3= f4*f4
15 f5= memory[1]
16 \, f4 = f5 + f5
17 f5= memory[2]
18 for i in range(15):
     memory[i] = f3 # sd f3
19
20
     f3 = f4*f4 \# mul f3, f4, f4
21
     f4 = f5+f5 \# add f4, f5, f5
22
      f5 = memory[i+3] # ld f5,"12"(r1)
23
24 print(memory)
[0, 4, 16, 36, 64, 100, 144, 196, 256, 324,
```





Software Pipelining com balanceamento





Original

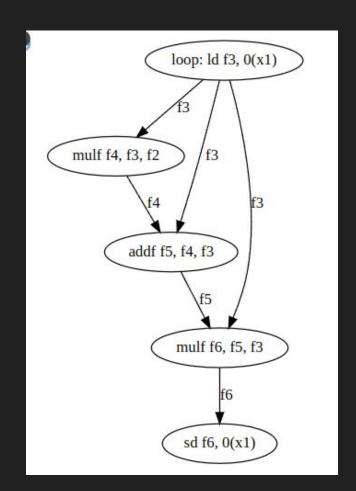
Software Pipe ining^{CAD} 2023 com balanceamento

```
2 for i in range(20):
      f3 = memory[i]
                                            10 f3 = memory[0]
     f4 = f3*f3
                                            11 \, \text{f5} = \text{f3*f3}
    f4 = f4 + f3
                                            12 f4 = f5 + f3 # 1
    memory[i] = f4
                                            13 f3 = memory[1]
                                            14 \, \text{f5} = \text{f3*f3}
 8 print(memory)
                                            15 \, \text{f6} = \, \text{f3} \, \text{\#2}
[0, 2, 6, 12, 20, 30, 42, 56, 72, 90]
                                            16 f3 = memory[2] # 3
                                            17 for i in range(20):
                                            18 \quad memory[i] = f4 \# SD
                                            19 f4 = f5+f6 \# add f4, f5, f3
                                            20 f5 = f3*f3 \# mul f5, f3, f3
                                            f6 = f3
                                            22
                                                 f3 = memory[i+3] # ld f3
                                            23
                                            24
                                            25 print (memory)
                                            [0, 2, 6, 12, 20, 30, 42, 56, 72, 90, 110, 132, 156,
```

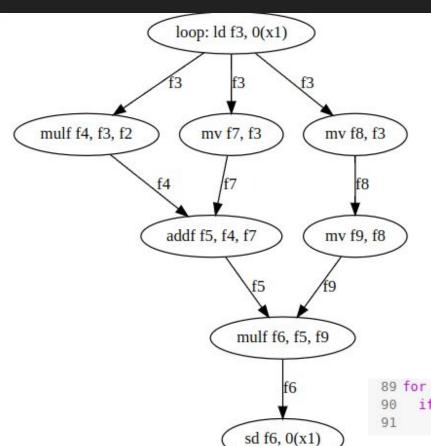
Exercício para os estudantes

```
v[i] = \left( \left( 2 * v[i] + v[i] \right) * v[i] \right)
```

```
4 x1 = 0
5 f2 = 2
6 x5 = T
7 while True:
8   f3 = v[x1] # ld f3, 0(x1)
9   f4 = f3 * f2 # mulf f4, f3, f2
10   f5 = f4 + f3 # addf f5, f4, f3
11   f6 = f5 * f3 # mulf f6, f5, f3
12   v[x1] = f6 # sd f6, 0(x1)
13   x1 = x1 + 1 # addi x1, x1, 4
14   if (x1 == x5): # bne x1, x5, loop
15   break
```



Grafo, Software Pipelining e validação

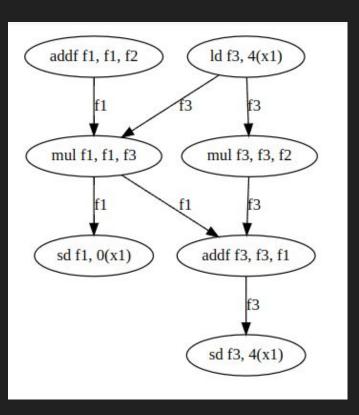


```
41 while True:
    l[x1] = f6
    f6 = f5 * f9
    f5 = f4 + f7
    f9 = f8
    f4 = f3 * f2
    f7 = f3
    f8 = f3
52
    f3 = l[x1 + 4]
54
    x1 = x1 + 1
    if (x1 == (x5 - 4)):
      break
57
```

```
89 for i in range(T):
90   if v[i] != l[i]:
91    print("Wrong calculation in software pipeline at position", i)
```

geração do grafo de dependência

```
1 instructions = (
2   'ld f1, 0(x1)',
3   'addf f1, f1, f2',
4   'ld f3, 4(x1)',
5   'mul f1, f1, f3',
6   'sd f1, 0(x1)',
7   'mul f3, f3, f2',
8   'addf f3, f3, f1',
9   'sd f3, 4(x1)'
Here the provided HTML of the provided HT
```





Versão Simplificada de uma GPU

- Warps com 4 threads
- Banco de Registradores com 64 registros
- Máximo de 32 threads por Multiprocessador
- Máximo de 4 blocos por Multiprocessador
- Máximo de 16 threads por bloco
- Instruções
 - Ld 4 ciclos cache, 20 ciclos memoria
 - Sd 4 ciclos cache, 20 ciclos memoria
 - Add e Mul 3 ciclos
- Ordem de swap dos Warps é aleatória

Planilha CSV para entrada do trace

Trecho codigo a = b + c

```
[ ] 1 #@title Trecho codigo a = b + c
     2 %writefile exemplo.csv
     3 PC; Inst; Warp; Fetch; Decode; Exec; W
     4 0; Ld r1; w1; 1; 2; 3-6; 7;
     5 1; Ld r2; w1; 2; 3; 4-7; 8;
     6 0; Ld r1; w2; 3; 4; 5-8; 9;
     7 1; Ld r2; w2; 4; 5; 6-9; 10;
     8 x; nop; ; 5
     9 x; nop; ; 6
    10 2; add r1, r1, r2; w1; 7; 8; 9-11; 12
    11 x; nop; ; 8
    12 2; add r1, r1, r2; w1; 9; 10; 11-13; 14
    13 x; nop; ; 10
    14 3; sd r1; 11; 12; 13-17
    15 x; nop; ; 12
    16 3; sd r1; 13; 14; 15-19
```

Visualização do Trace

	PC	Inst	Warp	Fetch	Decode	Exec	W	
0	0	Ld rl	w1	1	2	3-6	7.0	
1	1	Ld r2	w1	2	3	4-7	8.0	
2	0	Ld rl	w2	3	4	5-8	9.0	
3	1	Ld r2	w2	4	5	6-9	10.0	
4	X	nop		5	-	_	-	
5	X	nop		6	8	29	923	
6	2	add rl,rl,r2	w1	7	8	9-11	12.0	
7	X	nop		8	- 2	-	(57.0)	
8	2	add r1, r1, r2	w1	9	10	11-13	14.0	
9	X	nop		10	15	7.5	-	
10	3	sd r1	11	12	13-17	- 5	-	
11	Х	nop		12	-	-	+	
12	3	sd rl	13	14	15-19	-		

```
2 %%writefile exemplo.csv

3 warp; 1;2;3;4;5;6;7;8;9;0;1;2;3

4 w1; ld; ld; ; ; ; add; ; ; ; sd

5 w2; ; ; ld; ld; ; ; ; add; ; ; ; sd
```



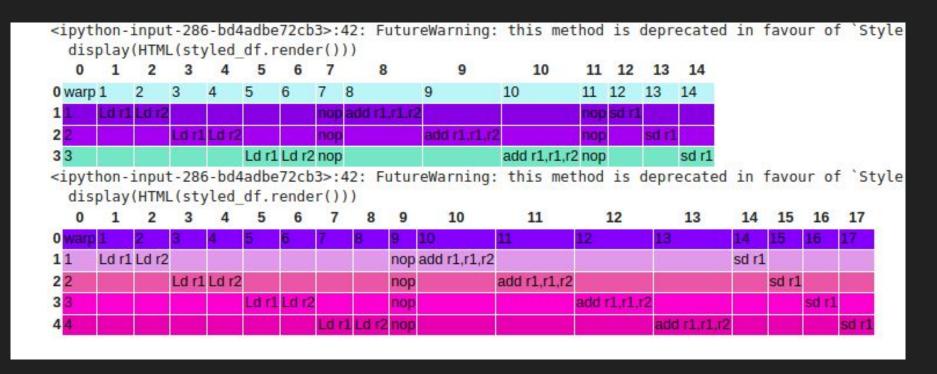
WSCAD 2023

Resultados dos Estudantes para Visualização com Tabulate

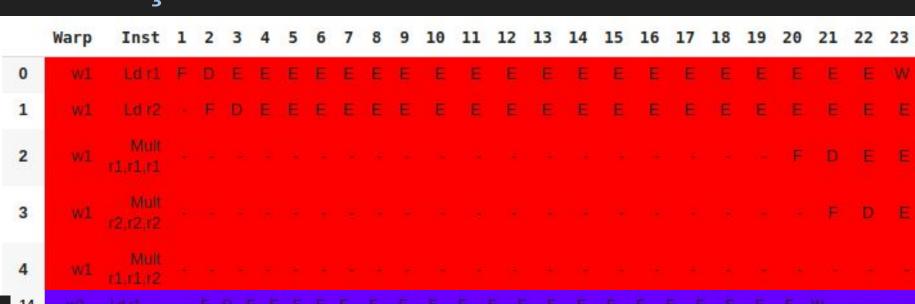
<u>Warp</u>	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
w1	Ld	Ld							add			sd			
w2			Ld	Ld						add			sd		
w3					Ld	Ld					add			sd	
w4							Ld	Ld			add				sd

WSCAD 2023

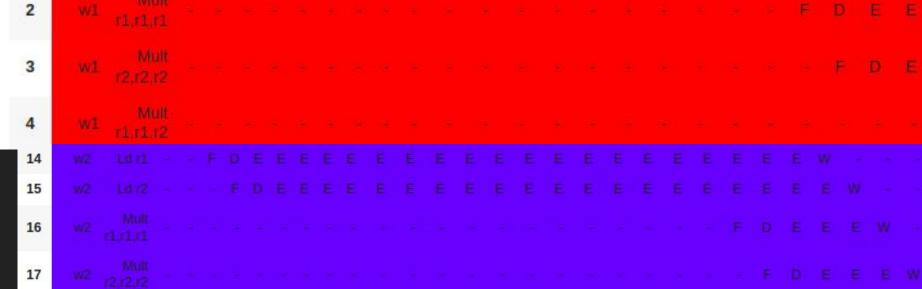
Resultados dos Estudantes para Visualização com Html



Resultados dos Estudantes para Visualização com cores

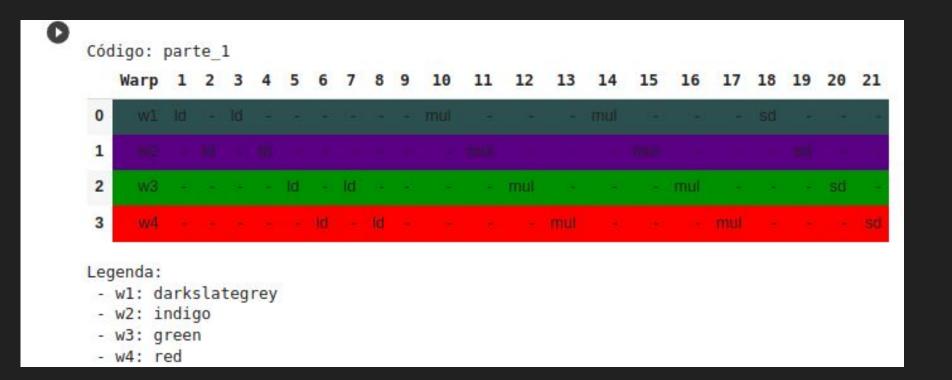


WSCAD 2023



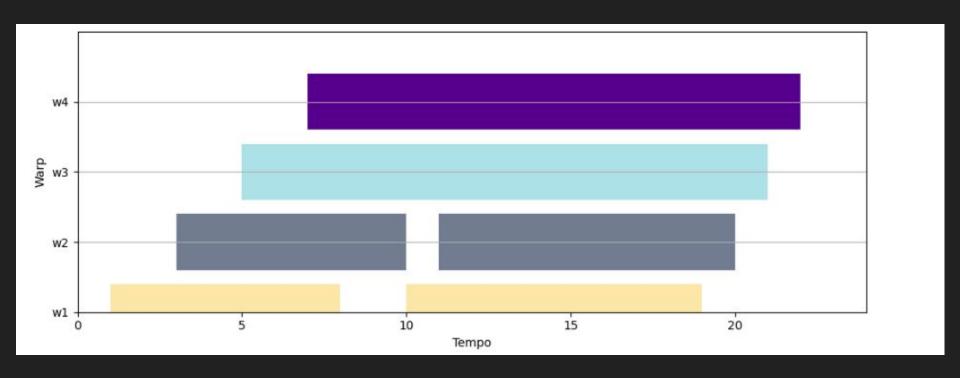
WSCAD 2023

Resultados dos Estudantes para Visualização com cores II



Resultados dos Estudantes para Visualização com barras





Considerações Finais

 Duas novas ferramentas de Ensino: Software Pipelining e Escalonamento de GPU

• Estudantes gostaram de usar Python e de aprender recursos de visualização

 Estudantes não tiveram dificuldades (comparado aos anos anteriores) para resolver questões dos temas trabalhados nas avaliações Perguntas? ricardo@ufv.br

Acknowledgments

Financial support from FAPEMIG APQ-01577-22, CNPq, and UFV. This work was also carried out with the support of the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Financing Code 001



Obrigado pela atenção e perguntas ? ricardo@ufv.br





outros links de Colab ricardo@ufv.br

Livro Lógica Digital



Minicurso Educomp 2023





outros links de Colab ricardo@ufv.br

lab. de Ideias



SBESC 2020 ESSE

